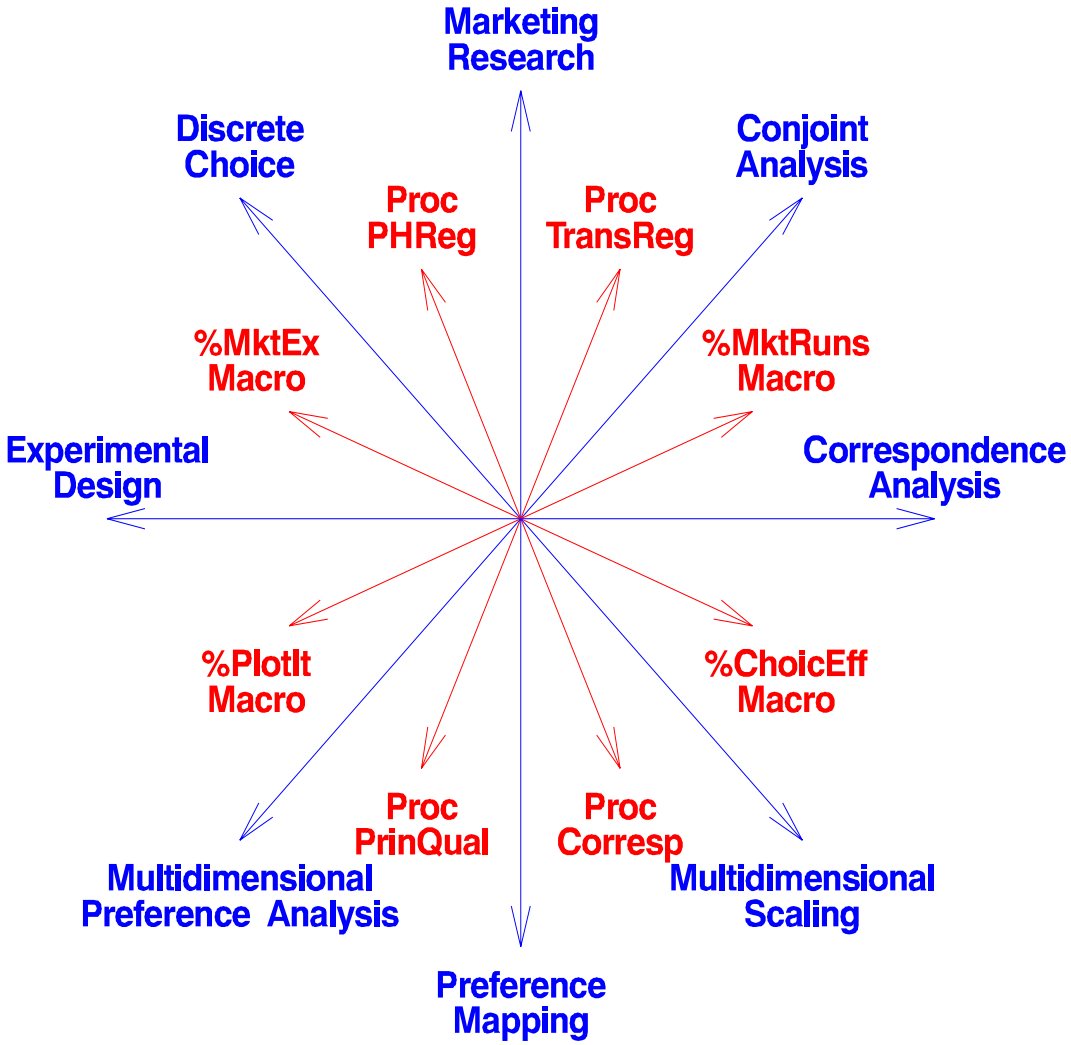
Marketing Research Methods in SAS

Experimental Design, Choice, Conjoint, and Graphical Techniques



Warren F. Kuhfeld

May 1, 2003
SAS 9.1 Edition
TS-689

Marketing Research Methods in SAS

Marketing Research Methods in SAS discusses experimental design, discrete choice, conjoint analysis, and graphical and perceptual mapping techniques. The book has grown and evolved over many years and many revisions. For example, the section on choice models grew from a two page handout written by Dave DeLong about ten years ago. This is the May 1, 2003 edition for SAS 9.1. All of the macros and code used in this book should work in SAS 8.2, 9.0, and 9.1. However, if you are not running SAS 9.1, you must get the latest macros from the web or from me. The macros shipped with SAS 8.2 and 9.0 are obsolete. I hope that this book and tool set will help you do better research, do it quickly, and do it more easily. I would like to hear what you think. Many of my examples and enhancements to the software are based on feedback from people like you. If you would like to be added to a mailing list to receive periodic email updates on SAS marketing research tools (probably no more than once every few months), email Warren.Kuhfeld@sas.com. This list will not be sold or used for any other purpose.

With this edition, **Marketing Research Methods in SAS** becomes for the first time a true book and more than its original collection of papers and handouts. For the first time, all chapters are composed in LaTeX using the same style, with a single table of contents, reference section, and index. For the first time, there is a complete set of referrals (see page ...) between different sections of the book. Several chapters are based on papers written in the early 1990's. In past editions, I tried to maintain as much of the original paper as I could. For example, the "Efficient Experimental Design with Marketing Research Applications" chapter is based on a paper that appeared in the *Journal of Marketing Research* in 1994. In previous editions, I kept all of the words in the original paper but added new footnotes to discuss new tools. In this edition, that approach started getting a bit silly. The tools have changed so much in these past nine years, so the article needed to be revised to reflect that. Similarly, the chapter on splines has been revised as well. Another change you might notice is I changed to a larger font. Going from a 10 point to an 11 point font added around 80 pages to the book, so I apologize to the trees, but I am not as young as I used to be, and even with these progressive multifocal lenses, I just can't read what I used to be able to read. Please help save trees and use the PDF files. Copies of this book are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market (reports beginning with "TS-689") or by writing Warren.Kuhfeld@sas.com. I used to distribute printed copies with each new edition. I quit doing that several editions back because the book got too big.

The following people contributed to writing portions of this book: Mark Garratt, Joel Huber, Ying So, Randy Tobias, Wayne Watson, and Klaus Zwerina. My parts could not have been written without the help of many people. I would like to thank Joel Huber, Ying So, Randy Tobias, and John Wurst. My involvement in the area of experimental design and choice modeling can be traced to several conversations with Mark Garratt in the early 1990's and then to the influence of Don Anderson, Joel Huber, Jordan Louviere, and Randy Tobias. Don Anderson, Neil Sloane, and J.C. Wang kindly helped me with some of the orthogonal designs in the `%MktEx` macro. Brad Jones helped me with coordinate exchange. Randy Tobias has been a great colleague and a huge help to me over the years in the area of experimental design, and many components of the `%MktEx` macro and other design macros are based on his ideas.

Finishing a project this size causes one to pause and reflect. I am particularly proud of this edition of the book, this version of the tools, and the `%MktEx` macro, which is the world's most comprehensive factorial designer. Reflecting back however, it is clear that I have stood on the shoulders of giants.

There are a few other people I would like to acknowledge. Without these people, I would have never been in the position to write a book such as this. From my undergraduate days at Kent State, I would like to thank Roy Lilly, Larry Melamed, my mentor Ben Newberry, and Steve Simnick. From graduate school at UNC, I would like to thank Ron Helms, Keith Muller, and especially my advisor and mentor Forrest Young. From SAS, I would like to thank Warren Sarle and all of my colleagues in SAS/STAT Research and Development. It is great to work with such a smart, talented, productive, and helpful group of people. Finally, I would like to thank my mother, my late father, and my stepfather Ed, for being so good to my Mom and for being such a wonderful grandfather to my children. I dedicate this book to my wife Peg and my children Megan and Rusty.

Warren F. Kuhfeld, Ph.D.
Manager, Multivariate Models R&D
SAS Institute Inc.
May 1, 2003

Contents

Marketing Research: Uncovering Competitive Advantages

Warren F. Kuhfeld

Abstract

SAS provides a variety of methods for analyzing marketing data including conjoint analysis, correspondence analysis, preference mapping, multidimensional preference analysis, and multidimensional scaling. These methods allow you to analyze purchasing decision trade-offs, display product positioning, and examine differences in customer preferences. They can help you gain insight into your products, your customers, and your competition. This chapter discusses these methods and their implementation in SAS.*

Introduction

Marketing research is an area of applied data analysis whose purpose is to support marketing decision making. Marketing researchers ask many questions, including:

- Who are my customers?
- Who else should be my customers?
- Who are my competitors' customers?
- Where is my product positioned relative to my competitors' products?
- Why is my product positioned there?
- How can I reposition my existing products?
- What new products should I create?
- What audience should I target for my new products?

*This is a minor modification of a paper that was presented to SUGI 17 by Warren F. Kuhfeld and to the 1992 Midwest SAS Users Group meeting by Russell D. Wolfinger. Copies of this chapter (TS-689A) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market .

Marketing researchers try to answer these questions using both standard data analysis methods, such as descriptive statistics and crosstabulations, and more specialized marketing research methods. This chapter discusses two families of specialized marketing research methods, perceptual mapping and conjoint analysis. Perceptual mapping methods produce plots that display product positioning, product preferences, and differences between customers in their product preferences. Conjoint analysis is used to investigate how consumers trade off product attributes when making a purchasing decision.

Perceptual Mapping

Perceptual mapping methods, including correspondence analysis (CA), multiple correspondence analysis (MCA), preference mapping (PREFMAP), multidimensional preference analysis (MDPREF), and multidimensional scaling (MDS), are data analysis methods that generate graphical displays from data. These methods are used to investigate relationships among products as well as individual differences in preferences for those products.[†]

CA and MCA can be used to display demographic and survey data. CA simultaneously displays in a scatter plot the row and column labels from a two-way contingency table (crosstabulation) constructed from two categorical variables. MCA simultaneously displays in a scatterplot the category labels from more than two categorical variables.

MDPREF displays products positioned by overall preference patterns. MDPREF also displays differences in how customers prefer products. MDPREF displays in a scatter plot both the row labels (products) and column labels (consumers) from a data matrix of continuous variables.

MDS is used to investigate product positioning. MDS displays a set of object labels (products) whose perceived similarity or dissimilarity has been measured.

PREFMAP is used to interpret preference patterns and help determine why products are positioned where they are. PREFMAP displays rating scale data in the same plot as an MDS or MDPREF plot. PREFMAP shows both products and product attributes in one plot.

MDPREF, PREFMAP, CA, and MCA are all similar in spirit to the biplot, so first the biplot is discussed to provide a foundation for discussing these methods.

The Biplot. A *biplot* (Gabriel, 1981) simultaneously displays the row and column labels of a data matrix in a low-dimensional (typically two-dimensional) plot. The "bi" in "biplot" refers to the *joint* display of rows and columns, not to the dimensionality of the plot. Typically, the row coordinates are plotted as points, and the column coordinates are plotted as vectors.

Consider the artificial preference data matrix in Figure 1. Consumers were asked to rate their preference for products on a 0 to 9 scale where 0 means little preference and 9 means high preference. Consumer 1's preference for Product 1 is 4. Consumer 1's most preferred product is Product 4, which has a preference of 6.

The biplot is based on the idea of a matrix decomposition. The $(n \times m)$ data matrix $\mathbf{Y}$ is decomposed into the product of an $(n \times q)$ matrix $\mathbf{A}$ and a $(q \times m)$ matrix $\mathbf{B}'$. Figure 2 shows a decomposition of the data in Figure 1.[‡] The rows of $\mathbf{A}$ are coordinates in a two-dimensional plot for the row points in

[†]Also see Kuhfeld (2003) starting on pages 661 and 675.

[‡]Figure 2 does not contain the decomposition that would be used for an actual biplot. Small integers were chosen to simplify the arithmetic.

	Consumer 1	Consumer 2	Consumer 3
Product 1	4	1	6
Product 2	4	2	4
Product 3	1	0	2
Product 4	6	2	8

Figure 1. Preference Data Matrix

$$\underset{Y}{\begin{bmatrix} 4 & 1 & 6 \\ 4 & 2 & 4 \\ 1 & 0 & 2 \\ 6 & 2 & 8 \end{bmatrix}} = \underset{A}{\begin{bmatrix} 1 & 2 \\ 2 & 0 \\ 0 & 1 \\ 2 & 2 \end{bmatrix}} \times \underset{B'}{\begin{bmatrix} 2 & 1 & 2 \\ 1 & 0 & 2 \end{bmatrix}}$$

Figure 2. Preference Data Decomposition

$\mathbf{Y}$, and the columns of $\mathbf{B}'$ are coordinates in the same two-dimensional plot for the column points in $\mathbf{Y}$. In this artificial example, the entries in $\mathbf{Y}$ are exactly reproduced by *scalar products* of coordinates. For example, the $(1,1)$ entry in $\mathbf{Y}$ is $y_{11} = a_{11} \times b_{11} + a_{12} \times b_{12} = 4 = 1 \times 2 + 2 \times 1$.

The rank of $\mathbf{Y}$ is $q \leq MIN(n, m)$. The rank of a matrix is the minimum number of dimensions that are required to represent the data without loss of information. The rank of $\mathbf{Y}$ is the full number of columns in $\mathbf{A}$ and $\mathbf{B}$. In the example, $q = 2$. When the rows of $\mathbf{A}$ and $\mathbf{B}$ are plotted in a two-dimensional scatter plot, the scalar product of the coordinates of $\mathbf{a}'_i$ and $\mathbf{b}'_j$ *exactly* equals the data value y_{ij}. This kind of scatter plot is a biplot. When $q > 2$ and the first two dimensions are plotted, then $\mathbf{AB}'$ is *approximately* equal to $\mathbf{Y}$, and the display is an *approximate biplot*.[§] The best values for $\mathbf{A}$ and $\mathbf{B}$, in terms of minimum squared error in approximating $\mathbf{Y}$, are found using a singular value decomposition (SVD).[¶] An approximate biplot is constructed by plotting the first two columns of $\mathbf{A}$ and $\mathbf{B}$.

When $q > 2$, the full geometry of the data cannot be represented in two dimensions. The first two columns of $\mathbf{A}$ and $\mathbf{B}$ provide the best approximation of the high dimensional data in two dimensions. Consider a cloud of data in the shape of an American football. The data are three dimensional. The best one dimensional representation of the data—*the first principal component*—is the line that runs from one end of the football, through the center of gravity or *centroid* and to the other end. It is the longest line that can run through the football. The second principal component also runs through the centroid and is perpendicular or *orthogonal* to the first line. It is the longest line that can be drawn through the centroid that is perpendicular to the first. If the football is a little thicker at the laces, the second principal component runs from the laces through the centroid and to the other side of the football. All of the points in the football shaped cloud can be projected into the plane of the first two principal components. The resulting scatter plot will show the approximate shape of the data. The two longest dimensions are shown, but the information in the other dimensions are lost. This is the principle behind approximate biplots. See Gabriel (1981) for more information on the biplot.

Multidimensional Preference Analysis. Multidimensional Preference Analysis (Carroll, 1972) or MDPREF is a biplot analysis for preference data. Data are collected by asking respondents to rate their preference for a set of objects—products in marketing research.

Questions that can be addressed with MDPREF analyses include: Who are my customers? Who else should be my customers? Who are my competitors' customers? Where is my product positioned relative to my competitors' products? What new products should I create? What audience should I target for my new products?

[§]In practice, the term biplot is sometimes used without qualification to refer to an approximate biplot.

[¶]SVD is sometimes referred to in the psychometric literature as an Eckart-Young (1936) decomposition. SVD is closely tied to the statistical method of principal component analysis.

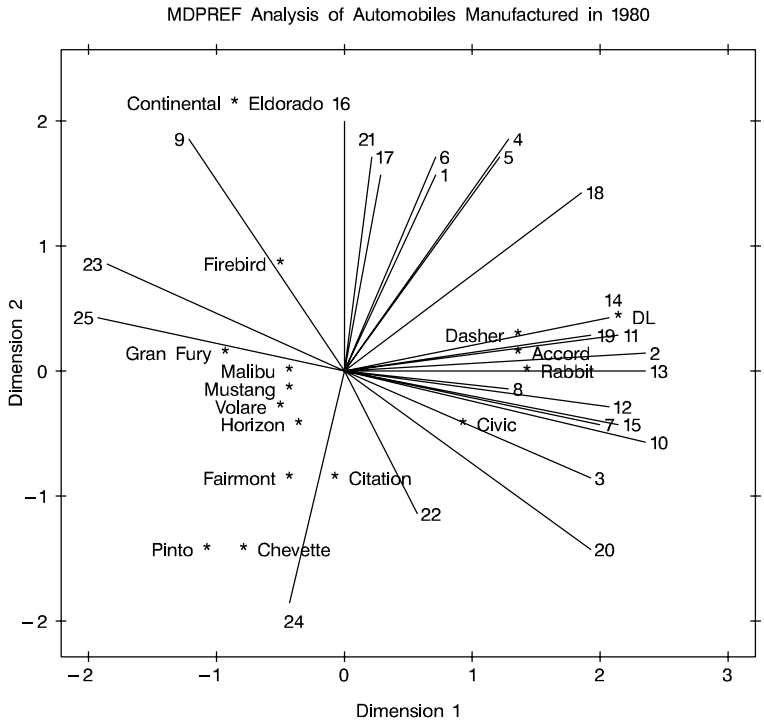


Figure 3. Multidimensional Preference Analysis

For example, consumers were asked to rate their preference for a group of automobiles on a 0 to 9 scale, where 0 means no preference and 9 means high preference. $\mathbf{Y}$ is an $(n \times m)$ matrix that contains ratings of the n products by the m consumers. Figure 3 displays an example in which 25 consumers rated their preference for 17 new (at the time) 1980 automobiles. Each consumer is a vector in the space, and each car is a point identified by an asterisk (*). Each consumer's vector points in *approximately* the direction of the cars that the consumer most preferred.

The dimensions of this plot are the first two principal components. The plot differs from a proper biplot of $\mathbf{Y}$ due to scaling factors. At one end of the plot of the first principal component are the most preferred automobiles; the least preferred automobiles are at the other end. The American cars on the average were least preferred, and the European and Japanese cars were most preferred. The second principal component is the longest dimension that is orthogonal to the first principal component. In the example, the larger cars tend to be at the top and the smaller cars tend to be at the bottom.

The automobile that projects farthest along a consumer vector is that consumer's most preferred automobile. To project a point onto a vector, draw an imaginary line through a point crossing the vector at a right angle. The point where the line crosses the vector is the *projection*. The length of this projection differs from the predicted preference, the scalar product, by a factor of the length of the consumer vector, which is constant within each consumer. Since the goal is to look at projections of points onto the vectors, the absolute length of a consumer's vector is unimportant. The relative lengths of the vectors indicate fit, with longer vectors indicating better fit. The coordinates for the endpoints of the vectors were multiplied by 2.5 to extend the vectors and create a better graphical display. The direction of the preference scale is important. The vectors point in the direction of increasing values of the data values. If the data had been ranks, with 1 the most preferred and n the least preferred, then the vectors would point in the direction of the least preferred automobiles.

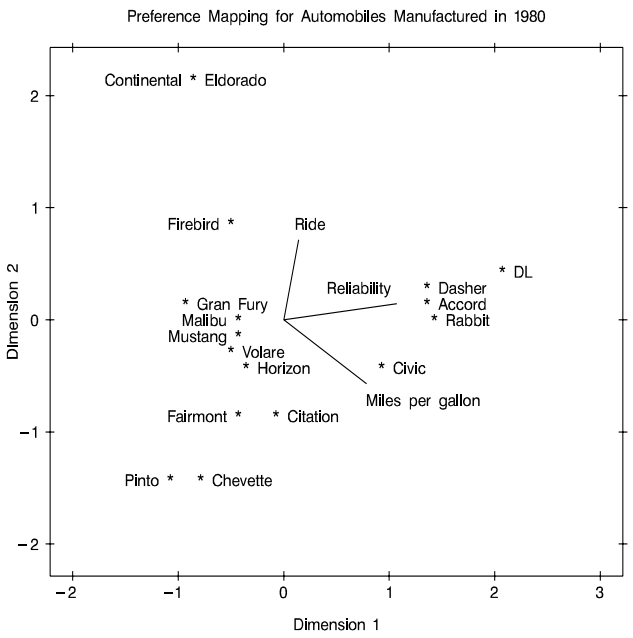


Figure 4. Preference Mapping, Vector Model

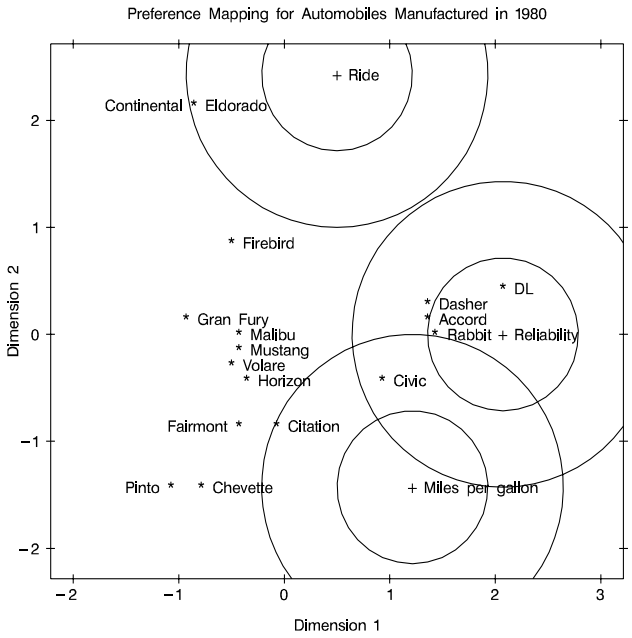


Figure 5. Preference Mapping, Ideal Point Model

Consumers 9 and 16, in the top left portion of the plot, most prefer the large American cars. Other consumers, with vectors pointing up and nearly vertical, also show this pattern of preference. There is a large cluster of consumers, from 14 through 20, who prefer the Japanese and European cars. A few consumers, most notably consumer 24, prefer the small and inexpensive American cars. There are no consumer vectors pointing through the bottom left portion of the plot between consumers 24 and 25, which suggests that the smaller American cars are generally not preferred by any of these consumers.

Some cars have a similar pattern of preference, most notably Continental and Eldorado, which share a symbol in the plot. This indicates that marketers of Continental or Eldorado may want to try to distinguish their car from the competition. Dasher, Accord, and Rabbit were rated similarly, as were Malibu, Mustang, Volare, and Horizon. Several vectors point into the open area between Continental/Eldorado and the European and Japanese cars. The vectors point away from the small American cars, so these consumers do not prefer the small American cars. What car would these consumers like? Perhaps they would like a Mercedes or BMW.

Preference Mapping. Preference mapping[‖] (Carroll, 1972) or PREFMAP plots resemble biplots, but are based on a different model. The goal in PREFMAP is to project external information into a configuration of points, such as the set of coordinates for the cars in the MDPREF example in Figure 3. The external information can aid interpretation.

Questions that can be addressed with PREFMAP analyses include: Where is my product positioned relative to my competitors' products? Why is my product positioned there? How can I reposition my existing products? What new products should I create?

[‖] Preference mapping is sometimes referred to as external unfolding.

The PREFMAP Vector Model. Figure 4 contains an example in which three attribute variables (ride, reliability, and miles per gallon) are displayed in the plot of the first two principal components of the car preference data. Each of the automobiles was rated on a 1 to 5 scale, where 1 is poor and 5 is good. The end points for the attribute vectors are obtained by projecting the attribute variables into the car space. Orthogonal projections of the car points on an attribute vector give an approximate ordering of the cars on the attribute rating. The ride vector points almost straight up, indicating that the larger cars, such as the Eldorado and Continental, have the best ride. Figure 3 shows that most consumers preferred the DL, Japanese cars, and larger American cars. Figure 4 shows that the DL and Japanese cars were rated the most reliable and have the best fuel economy. The small American cars were not rated highly on any of the three dimensions.

Figure 4 is based on the simplest version of PREFMAP—the *vector model*. The vector model operates under the assumption that some is good and more is *always* better. This model is appropriate for miles per gallon and reliability—the more miles you can travel without refueling or breaking down, the better.

The PREFMAP Ideal Point Model. The *ideal point* model differs from the vector model, in that the ideal point model does not assume that more is better, *ad infinitum*. Consider the sugar content of cake. There is an ideal amount of sugar that cake should contain—not enough sugar is not good, and too much sugar is also not good. In the cars example, the ideal number of miles per gallon and the ideal reliability are unachievable. It makes sense to consider a vector model, because the ideal point is infinitely far away. This argument is less compelling for ride; the point for a car with smooth, quiet ride may not be infinitely far away. Figure 5 shows the results of fitting an ideal point model for the three attributes. In the vector model, results are interpreted by orthogonally projecting the car points on the attribute vectors. In the ideal point model, Euclidean distances between car points and ideal points are compared. Eldorado and Continental have the best predicted ride, because they are closest to the ride ideal point. The concentric circles drawn around the ideal points help to show distances between the cars and the ideal points. The numbers of circles and their radii are arbitrary. The overall interpretations of Figures 4 and 5 are the same. All three ideal points are at the edge of the car points, which suggests the simpler vector model is sufficient for these data. The ideal point model is fit with a multiple regression model and some pre- and post-processing. The regression model uses the MDS or MDPREF coordinates as independent variables along with an additional independent variable that is the sum of squares of the coordinates. The model is a constrained *response-surface model*.

The results in Figure 5 were modified from the raw results to eliminate *anti-ideal points*. The ideal point model is a distance model. The rating data are interpreted as distances between attribute ideal points and the products. In this example, each of the automobiles was rated on these three dimensions, on a 1 to 5 scale, where 1 is poor and 5 is good. The data are the reverse of what they should be—a ride rating of 1 should mean this car is similar to a car with a good ride, and a rating of 5 should mean this car is different from a car with a good ride. So the raw coordinates must be multiplied by -1 to get ideal points. Even if the scoring had been reversed, anti-ideal points can occur. If the coefficient for the sum-of-squares variable is negative, the point is an anti-ideal point. In this example, there is the possibility of *anti-anti-ideal points*. When the coefficient for the sum-of-squares variable is negative, the two multiplications by -1 cancel, and the coordinates are ideal points. When the coefficient for the sum-of-squares variable is positive, the coordinates are multiplied by -1 to get an ideal point.

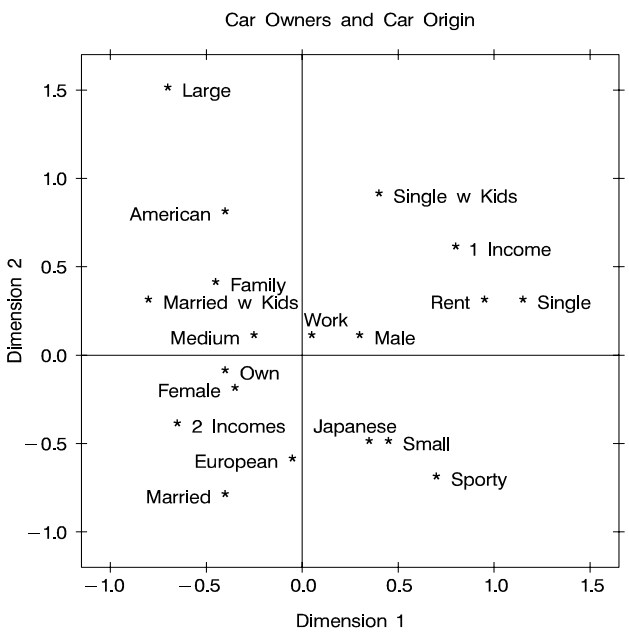


Figure 6. Multiple Correspondence Analysis

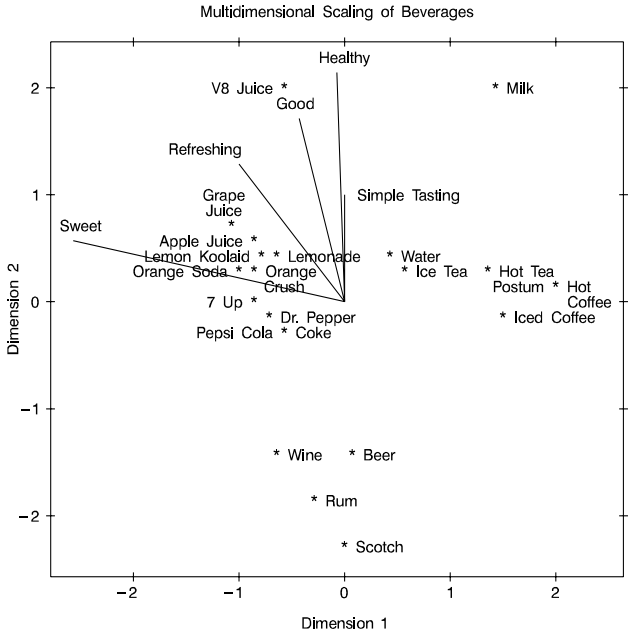


Figure 7. MDS and PREFMAP

Correspondence Analysis. Correspondence analysis (CA) is used to find a low-dimensional graphical representation of the association between rows and columns of a contingency table (crosstabulation). It graphically shows relationships between the rows and columns of a table; it graphically shows the relationships that the ordinary chi-square statistic tests. Each row and column is represented by a point in a Euclidean space determined from cell frequencies. CA is a popular data analysis method in France and Japan. In France, CA analysis was developed under the strong influence of Jean-Paul Benzécri; in Japan, under Chikio Hayashi. CA is described in Lebart, Morineau, and Warwick (1984); Greenacre (1984); Nishisato (1980); Tenenhaus and Young (1985); Gifi (1990); Greenacre and Hastie (1987); and many other sources. Hoffman and Franke (1986) provide a good introductory treatment using examples from marketing research.

Questions that can be addressed with CA and MCA include: Who are my customers? Who else should be my customers? Who are my competitors' customers? Where is my product positioned relative to my competitors' products? Why is my product positioned there? How can I reposition my existing products? What new products should I create? What audience should I target for my new products?

MCA Example. Figure 6 contains a plot of the results of a multiple correspondence analysis (MCA) of a survey of car owners. The questions included origin of the car (American, Japanese, European), size of car (small, medium, large), type of car (family, sporty, work vehicle), home ownership (owns, rents), marital/family status (single, married, single and living with children, and married living with children), and sex (male, female). The variables are all categorical.

The top-right quadrant of the plot suggests that the categories single, single with kids, one income, and renting a home are associated. Proceeding clockwise, the categories sporty, small, and Japanese are associated. In the bottom-left quadrant you can see the association between being married, owning your own home, and having two incomes. Having children is associated with owning a large American family car. Such information can be used to identify target audiences for advertisements. This interpretation is based on points being located in approximately the same direction from the origin and in approximately

the same region of the space. Distances between points are not interpretable in MCA.

Multidimensional Scaling. Multidimensional scaling (MDS) is a class of methods for estimating the coordinates of a set of objects in a space of specified dimensionality from data measuring the distances between pairs of objects (Kruskal and Wish, 1978; Schiffman, Reynolds, and Young, 1981; Young, 1987). The data for MDS consist of one or more square symmetric or asymmetric matrices of similarities or dissimilarities between objects or stimuli. Such data are also called *proximity data.* In marketing research, the objects are often products. MDS is used to investigate product positioning.

For example, consumers were asked to rate the differences between pairs of beverages. In addition, the beverages were rated on adjectives such as Good, Sweet, Healthy, Refreshing, and Simple Tasting. Figure 7 contains a plot of the beverage configuration along with attribute vectors derived through preference mapping. The alcoholic beverages are clustered at the bottom. The juices and carbonated soft drinks are clustered at the left. Grape and Apple juice are above the carbonated and sweet soft drinks and are perceived as more healthy than the other soft drinks. Perhaps sales of these drinks would increase if they were marketed as a healthy alternative to sugary soft drinks. A future analysis, after a marketing campaign, could check to see if their positions in the plot change in the healthy direction.

Water, coffee and tea drinks form a cluster at the right. V8 Juice and Milk form two clusters of one point each. Milk and V8 are perceived as the most healthy, whereas the alcoholic beverages are perceived as least healthy. The juices and carbonated soft drinks were rated as the sweetest. Pepsi and Coke are mapped to coincident points, as are Postum (a coffee substitute) and Hot Coffee. Orange Soda is near Orange Crush, and Lemon Koolaid is near Lemonade.

Geometry of the Scatter Plots. It is important that scatter plots displaying perceptual mapping information accurately portray the underlying geometry. All of the scatter plots in this chapter were created with the axes equated so that a centimeter on the y-axis represents the same data range as a centimeter on the x-axis.** *This is important.* Distances, angles between vectors, and projections are evaluated to interpret the plots. When the axes are equated, distances and angles are correctly presented in the plot. When axes are scaled independently, for example to fill the page, then the correct geometry is not presented. This important step of equating the axes is often overlooked in practice.

For MDPREF and PREFMAP, the absolute lengths of the vectors are not important since the goal is to project points on vectors, not look at scalar products of row points and column vectors. It is often necessary to change the lengths of *all* of the vectors to improve the graphical display. If all of the vectors are relatively short with end points clustered near the origin, the display will be difficult to interpret. To avoid this problem in Figure 3, *both* the x-axis and y-axis coordinates were multiplied by the same constant, 2.5, to lengthen all vectors by the same relative amount. The coordinates must *not* be scaled independently.

Conjoint Analysis

Conjoint analysis is used in marketing research to analyze consumer preferences for products and services. See Green and Rao (1971) and Green and Wind (1975) for early introductions to conjoint analysis and Green and Srinivasan (1990) for a recent review article.

**If the plot axes are not equated in this chapter, it is due to unequal distortions of the axes that occurred during the final printing process.

Conjoint analysis grew out of the area of *conjoint measurement* in mathematical psychology. In its original form, *conjoint analysis* is a main effects analysis-of-variance problem with an ordinal scale-of-measurement dependent variable. Conjoint analysis decomposes rankings or rating-scale evaluation judgments of products into components based on qualitative attributes of the products. Attributes can include price, color, guarantee, environmental impact, and so on. A numerical *utility* or *part-worth utility* value is computed for each level of each attribute. The goal is to compute utilities such that the rank ordering of the sums of each product's set of utilities is the same as the original rank ordering or violates that ordering as little as possible.

When a monotonic transformation of the judgments is requested, a *nonmetric conjoint analysis* is performed. Nonmetric conjoint analysis models are fit iteratively. When the judgments are not transformed, a *metric conjoint analysis* is performed. Metric conjoint analysis models are fit directly with ordinary least squares. When all of the attributes are nominal, the metric conjoint analysis problem is a simple main-effects ANOVA model. The attributes are the independent variables, the judgments comprise the dependent variable, and the utilities are the parameter estimates from the ANOVA model. The metric conjoint analysis model is more restrictive than the nonmetric model and will generally fit the data less well than the nonmetric model. However, this is not necessarily a disadvantage since over-fitting is less of a problem and the results should be more reproducible with the metric model.

In both metric and nonmetric conjoint analysis, the respondents are typically not asked to rate all possible combinations of the attributes. For example, with five attributes, three with three levels and two with two levels, there are $3 \times 3 \times 3 \times 2 \times 2 = 108$ possible combinations. Rating that many combinations would be difficult for consumers, so typically only a small fraction of the combinations are rated. It is still possible to compute utilities, even if not all combinations are rated. Typically, combinations are chosen from an *orthogonal array* which is a *fractional-factorial design*. In an orthogonal array, the zero/one indicator variables are uncorrelated for all pairs in which the two indicator variables are not from the same factor. The main effects are orthogonal but are confounded with interactions. These interaction effects are typically assumed to be zero.

Questions that can be addressed with conjoint analysis include: How can I reposition my existing products? What new products should I create? What audience should I target for my new products?

Consider an example in which the effects of four attributes of tea on preference were evaluated. The attributes are temperature (Hot, Warm, and Iced), sweetness (No Sugar, 1 Teaspoon, 2 Teaspoons), strength (Strong, Moderate, Weak), and lemon (With Lemon, No Lemon). There are four factors: three with three levels and one with two levels. Figure 8 contains the results.[††]

Sweetness was the most important attribute (the importance is 55.795). This consumer preferred two teaspoons of sugar over one teaspoon, and some sugar was preferred over no sugar. The second most important attribute was strength (25.067), with moderate and strong tea preferred over weak tea. This consumer's most preferred temperature was iced, and no lemon was preferred over lemon.

Software

SAS includes software that implements these methods. SAS/STAT software was used to perform the analyses for all of the examples. Perceptual mapping methods are described with more mathematical detail in Kuhfeld (2003) starting on page 675.

[††]See Kuhfeld (2003) starting on page 361 for more information on conjoint analysis. Note that the results in Figure 8 have been customized using ODS. See page 366 for more information on customizing conjoint analysis output.

Conjoint Analysis of Tea-Tasting Data

The TRANSREG Procedure

The TRANSREG Procedure Hypothesis Tests for Linear(subj2)

Univariate ANOVA Table Based on the Usual Degrees of Freedom

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	617.7222	88.24603	32.95	<.0001
Error	10	26.7778	2.67778		
Corrected Total	17	644.5000			

Root MSE	1.63639	R-Square	0.9585
Dependent Mean	12.16667	Adj R-Sq	0.9294
Coeff Var	13.44979		

Utilities Table Based on the Usual Degrees of Freedom

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	12.1667	0.38570	
Lemon: No	0.7222	0.38570	7.008
Lemon: Yes	-0.7222	0.38570	
Temperature: Hot	0.5000	0.54546	12.129
Temperature: Iced	1.0000	0.54546	
Temperature: Warm	-1.5000	0.54546	
Sweetness: No Sugar	-7.3333	0.54546	55.795
Sweetness: 1 Teaspoon	3.1667	0.54546	
Sweetness: 2 Teaspoons	4.1667	0.54546	
Strength: Moderate	1.8333	0.54546	25.067
Strength: Strong	1.5000	0.54546	
Strength: Weak	-3.3333	0.54546	

Figure 8. Conjoint Analysis

Correspondence Analysis. The SAS/STAT procedure CORRESP performs simple and multiple correspondence analysis and outputs the coordinates for plotting. Raw data or tables may be input. Supplementary classes are allowed.

Multidimensional Preference Analysis. The SAS/STAT procedure PRINQUAL performs multidimensional preference analysis and outputs the coordinates for plotting. Nonmetric MDPREF, with transformations of continuous and categorical variables, is also available.

Preference Mapping. The SAS/STAT procedure TRANSREG performs preference mapping and outputs the coordinates. Nonmetric PREFMAP, with transformations of continuous and categorical variables, is also available.

Multidimensional Scaling. The SAS/STAT procedure MDS performs multidimensional scaling and outputs the coordinates. Metric, nonmetric, two-way, and three-way models are available.

Scatter Plots. The Base SAS procedure PLOT can plot the results from these analyses and optimally position labels in the scatter plot. PROC PLOT uses an algorithm, developed by Kuhfeld (1991), that uses a heuristic approach to avoid label collisions. Labels up to 200 characters long can be plotted.

The `%PlotIt` macro, was used to create graphical scatter plots of labeled points. There are options to draw vectors to certain symbols and draw circles around other symbols. This macro is in the SAS autocall macro library. Also see Kuhfeld (2003) starting on page 661.

Conjoint Analysis. The SAS/STAT procedure TRANSREG can perform both metric and nonmetric conjoint analysis. PROC TRANSREG can handle both *holdout* observations and *simulations*. Holdouts are ranked by the consumers but are excluded from contributing to the analysis. They are used to validate the results of the study. Simulation observations are not rated by the consumers and do not contribute to the analysis. They are scored as passive observations. Simulations are *what-if* combinations. They are combinations that are entered to get a prediction of what their utility would have been if they had been rated. Conjoint analysis is described in more detail in Kuhfeld (2003) starting on page 361.

The `%MktEx` macro can generate orthogonal designs for both main-effects models and models with interactions. Nonorthogonal designs—for example, when strictly orthogonal designs require too many observations—can also be generated. Nonorthogonal designs can be used in conjoint analysis studies to minimize the number of stimuli when there are many attributes and levels. This macro is in the SAS autocall macro library and is also available free of charge on the web: http://support.sas.com/techsup/tnote/tnote_stat.html#market . Experimental design and the `%MktEx` macro are described in more detail in Kuhfeld (2003) starting on pages 39, 61, 81, 361, 479, and 546.

Other Data Analysis Methods. Other procedures that are useful for marketing research include the SAS/STAT procedures for regression, ANOVA, discriminant analysis, principal component analysis, factor analysis, categorical data analysis, covariance analysis (structural equation models), and the SAS/ETS procedures for econometrics, time series, and forcasting. Discrete choice data can be analyzed with multinomial logit models using the PHREG procedure. Discrete choice is described in more detail in Kuhfeld (2003) starting on page 81.

Conclusions

Marketing research helps you understand your customers and your competition. Correspondence analysis compactly displays survey data to aid in determining what kinds of consumers are buying your products. Multidimensional preference analysis and multidimensional scaling show product positioning, group preferences, and individual preferences. Plots from these methods may suggest how to reposition your product to appeal to a broader audience. They may also suggest new groups of customers to target. Preference mapping is used as an aid in understanding MDPREF and MDS results. PREFMAP displays product attributes in the same plot as the products. Conjoint analysis is used to investigate how consumers trade off product attributes when making a purchasing decision.

The insight gained from perceptual mapping and conjoint analysis can be a valuable asset in marketing decision making. These techniques can help you gain insight into your products, your customers, and your competition. They can give you the edge in gaining a competitive advantage.

Introducing the Market Research Analysis Application

Wayne E. Watson

Abstract

Market research focuses on assessing the preferences and choices of consumers and potential consumers. A new component of SAS/STAT software in Release 6.11 of the SAS System is an application written in SAS/AF that provides statistical and graphical techniques for market research data analysis. The application allows you to employ statistical methods such as conjoint analysis, discrete choice analysis, correspondence analysis, and multidimensional scaling through intuitive point-and-click actions.*

Conjoint Analysis

Conjoint analysis is used to evaluate consumer preference. If products are considered to be composed of attributes, conjoint analysis can be used to determine what attributes are important to product preference and what combinations of attribute levels are most preferred.

Usually, conjoint analysis is a main-effects analysis of variance of ordinally-scaled dependent variables. Preferences are used as dependent variables, and attributes are used as independent variables. Often, a monotone transformation is used with the dependent variables to fit a model with no interactions.

As an example, suppose you have four attributes that you think are related to automobile tire purchase. You want to know how important each attribute is to consumers' stated preferences for a potential tire purchase. The four attributes under investigation are

- brand name
- expected tread mileage
- purchase price
- installation cost

The attributes of brand name, tread mileage, and purchase price have three possible values and installation cost has two values. The values for each attribute are:

*For current documentation on the Market Research Application see SAS Institute Inc, *Getting Started with The Market Research Application*, Cary, NC: SAS Institute Inc., 1997, 56 pp. This paper was written and presented at SUGI 20 (1995) by Wayne E. Watson. This paper was also presented to SUGI-Korea (1995) by Warren F. Kuhfeld. Wayne Watson is a Research Statistician at SAS and wrote the Marketing Research Application which uses procedures and macros written by Warren F. Kuhfeld. Copies of this chapter (TS-689B) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market .

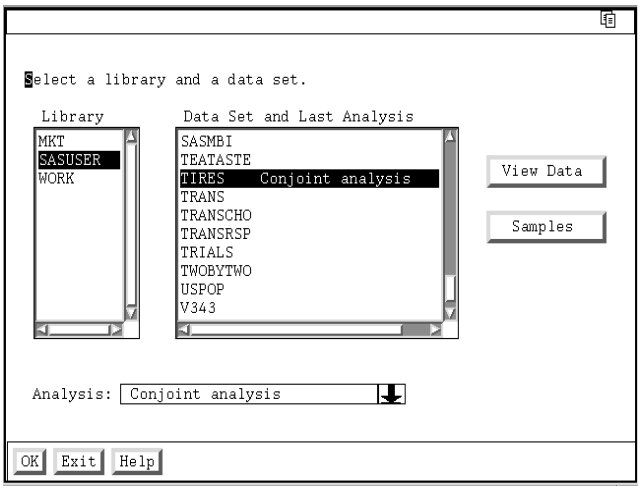


Figure 1. Selecting a Data Set and Analysis

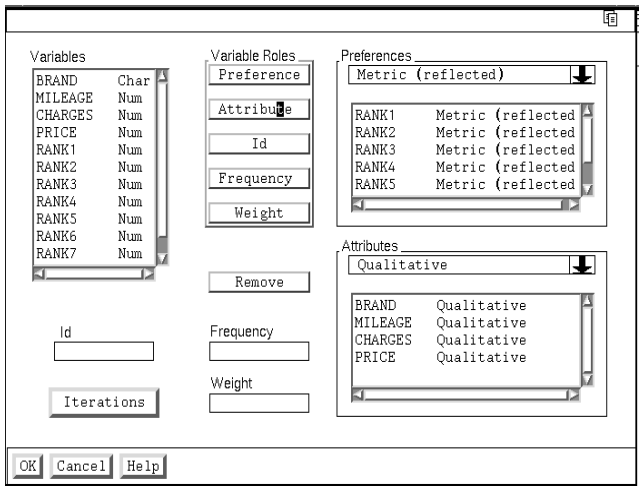


Figure 2. Conjoint Analysis Variable Selection

```
Brand:             Michelin, Goodyear, Firestone
Tread Mileage:     40,000, 60,000, 80,000
Price:             $45.00, $60.00, $75.00
Installation Cost: $0.00, $7.50
```

Seven respondents are asked to rank in order of preference 18 out of the possible 54 combinations. Although rankings are used in this example, preference ratings are frequently used in conjoint analysis.

Invoking the Application. With the data in the SAS data set, SASUSER.TIRES, you can invoke the Market Research application and perform a conjoint analysis. The application is invoked by issuing the "market" command on any command line.

Selecting a Data Set and Analysis. The first window displayed requires you to select a data set and an analysis. Because your data set is SASUSER.TIRES, select SASUSER as the library in the left-hand list box and TIRES as the data set in the right-hand list box. Then, select an analysis by clicking on the down arrow to the right of the analysis name field below the list boxes and select "Conjoint Analysis" from the displayed popup menu. See Figure 1.

View the data by pressing the View Data button and then selecting "Data values." The other selection under the View Data button,"Variable attributes," displays information about each variable.

Selecting Variables. To proceed with the analysis once you have selected a data set and an analysis, press the OK button at the bottom of the window.

The analysis requires preference and attribute variables. The preference variables are the ranks from the seven respondents and the attribute variables are the four factors. See Figure 2.

You can choose to perform a metric or a non-metric conjoint analysis; the metric analysis uses the ranks as they are, while the non-metric analysis performs a monotone transformation on the ranks. To set the measurement type for the preferences, click on the down arrow in the Preferences box at the top right of the window. Select "Metric (reflected)." "Reflected" is used because the lowest rank value, 1, corresponds to the most preferred offering. If the highest preference value corresponded to the most

preferred offering, the "Metric" selection should be used instead.

To select preference variables, select RANK1, RANK2, ... RANK7 in the Variables list box on the left side of the window, and press the Preference button in the Variable Roles box.

Likewise, you must select a measurement type for the attribute variables you want to use. The default measurement type for attributes is Qualitative, which treats the variable as a set of dummy variables with the coefficients of the dummy variables summing to 0. In this way, the utility coefficients † of each attribute sum to 0.

Use this measurement type for all four attribute variables, BRAND, MILEAGE, CHARGES, and PRICE. After selecting these four variables in the Variables list box, press the Attribute button in the Variable Roles box. Alternatively, you could use the "Continuous" measurement type for MILEAGE, CHARGES, or PRICE because these attributes are quantitative in nature.

To delete one or more of the Preference or Attribute variables, either double-click on each one in the appropriate right-hand list box or select them in any of the three list boxes and press the Remove button.

To obtain help about the window, press the Help button at the bottom of the window or click on any of the border titles on the window, for example, "Variables," "Variable Roles," "Preferences."

Once the variables have been selected, press the OK button at the bottom of the window to perform the analysis. To change the analysis, return to the Variable Selection window by pressing the Variables button on the analysis main window.

Results. The first result is a plot of the relative importance of each attribute. Relative importance is a measure of importance of the contribution of each attribute to overall preference; it is calculated by dividing the range of utilities for each attribute by the sum of all ranges and multiplying by 100.

In the example, Tire Mileage is the most important attribute with an average relative importance of 49%. The box-and-whisker plot displays the first and third quartiles as the ends of the box, the maximum and minimum as the whiskers (if they fall outside the box), and the median as a vertical bar in the interior of each box. See Figure 3.

To display a selection of additional results, press the Results button on the window. The first selection, the Utilities Table window, displays the utility coefficients for each level of an attribute for all preferences (the dependent variables). The relative importance of each attribute is displayed separately for each preference variable. This table illustrates that BRAND is the most important attribute for RANK1, the first respondent, and Michelin is the most preferred brand, because it has the highest utility coefficient value. Thus, the first respondent preferred a 80,000 mile, $45 Michelin with no installation charge.

After closing this window, you can view these results in graphical form by pressing the Results button again and selecting "Utilities plots." The plot of the Brand utilities indicates that one respondent clearly prefers Michelin while the other respondents only mildly prefer one brand over another.

To change the plot from the BRAND to the MILEAGE attribute, select MILEAGE in the list box at the right. All but one person prefer longer over shorter mileage tires, and that one prefers the 60,000 mile tire. You can examine plots for the PRICE and CHARGES attributes in the same way.

†Utility coefficients are estimates of the value or worth to a subject of each level of an attribute. The most preferred combination of attributes for a subject is the one with the attribute levels having the highest utility coefficient values for each attribute.

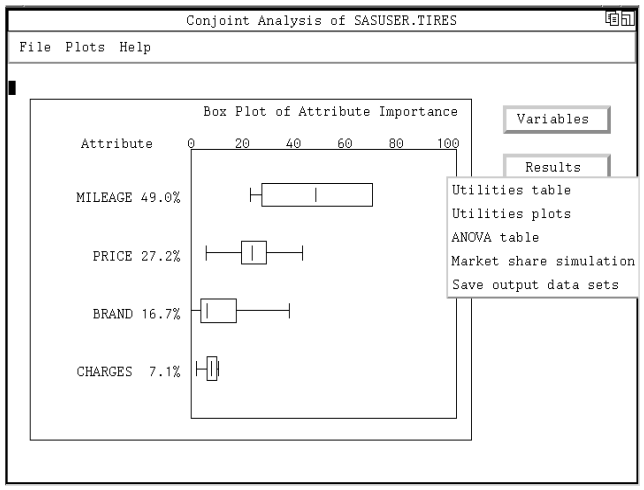


Figure 3. Plot of Relative Importance of Attributes

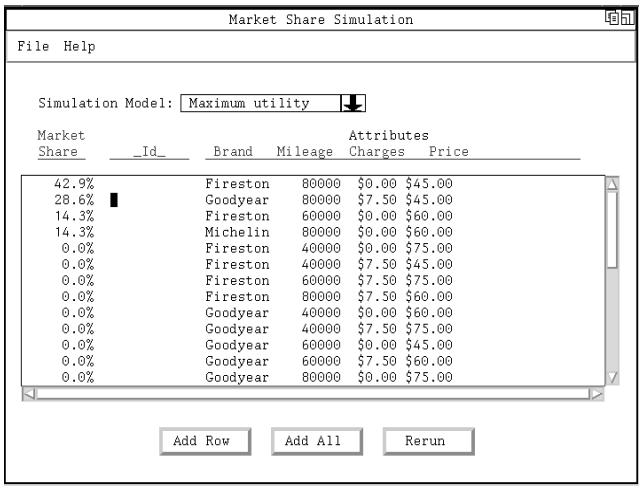


Figure 4. Estimating Market Share

Estimating Market Share. You also can calculate the expected market share for each tire purchase alternative in the sample. To do so, press the Results button and select "Market Share Simulation." The entry in the table with the largest market share is the 80,000 mile, $45 Firestone with no installation charge. It is expected to account for 42.9% of the market. The maximum utility simulation model, the default, was used to calculate the market share. You can choose from two other models: the logit model and the Bradley-Terry-Luce model. Click on the down arrow at the top of the window and select the desired model from the displayed list. See Figure 4.

Only 18 of the 54 possible tire purchase combinations were presented to the respondents. You may want to predict the expected market share of one or more of the combinations that were not present in the sample. To do so, press the Add Row button at the bottom of the window and fill in the observation in the top row of the table. Click on "-Select-" in each attribute column and select the desired level. If the observation that you create is a duplicate, a warning message is displayed. You can modify the contents of the Id column to contain a description of your own choice. After you have added some combinations, you can produce the expected market shares by pressing the Rerun button.

As an example an 80,000 mile, $45 Michelin with no installation charges would be expected to have a 64.3% market share if it was the only combination added to the original sample. Adding combinations may change the estimated market share of the other combinations.

Discrete Choice Analysis

Conjoint analysis is used to examine the preferences of consumers. The rationale for the use of preferences is that they indicate what people will choose to buy. Often in market research, the choices that consumers actually make are the behavior of interest. In these instances, it is appropriate to analyze choices directly using discrete choice analysis.

In discrete choice analysis, the respondent is presented with several choices and selects one of them. As in conjoint analysis, the factors that define the choice possibilities are called attributes. Here, they are called choice attributes to distinguish them from other factors, like demographic variables, that may be of interest but do not contribute to the definition of the choices. Each set of possible choices is called a choice set.

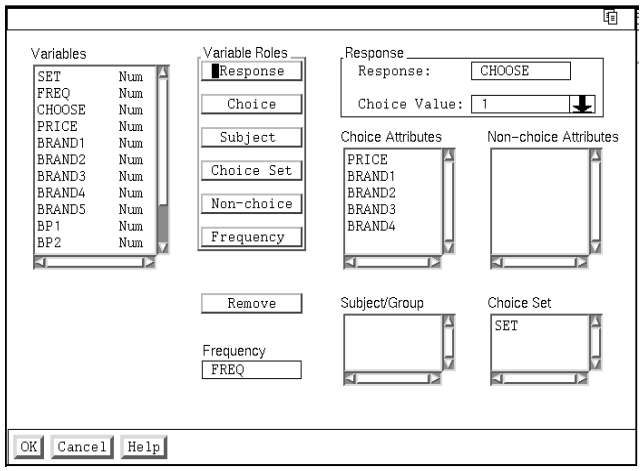


Figure 5. Discrete Choice Analysis Variable Selection

This example has choice possibilities defined by two attributes, price and brand. Five choice alternatives are presented at a time to a respondent, from which one alternative is chosen. Eight of these choice sets are presented, each one with a different set of five combinations of price and brand.

To change to a different data set or analysis, select "File → New dataset/analysis" on the main analysis window. Each time you change the data set or analysis or exit the application, you are asked if you want save the changes that you have made during the session. On the data set selection window, select the PRICE data set in the SASUSER library and then select "Discrete choice analysis." To continue, press the OK button.

With the other analyses in the application, you would be taken directly to the appropriate variable selection window. With discrete choice analysis, a supplementary window is displayed to help you determine if your data are in the appropriate form.

With discrete choice analysis, the structure of the data is important and must be in one of several layouts. After specifying if your data are contained in one or two data sets and whether a frequency variable is used, you can view the appropriate layout by pressing the Examine button. The most important requirement of the data layout is that all choice alternatives must be included, whether chosen or not.

If your data are not in the proper form, they must be rearranged before proceeding with the analysis. If your data are in the proper form, continue with the analysis by pressing the OK button. If not, press the Cancel button.

On the Variable Selection window that appears next, you must select several required variables: a response variable, some choice attribute variables, and a subject variable. Optionally, you can also choose a frequency variable and some non-choice attribute variables. If you select a frequency variable, a subject variable is not necessary.

For this example, select CHOOSE as the response variable. You also must indicate which value of the variable represents a choice. Click on the down arrow to the right of "Choice Value:" and select 1 from the list. In this example the value 1 indicates the chosen alternative and the value 0 indicates the non-chosen alternatives. See Figure 5.

Next, select PRICE and BRAND1, BRAND2, ..., BRAND4 as Choice attributes. BRAND is a nominal variable with five levels. It can be represented as four dummy-coded variables. ‡

Select FREQ as the frequency variable. The frequency variable contains the count of the number of times that a choice alternative was selected.

Because the data include more than one choice set, a Choice Set variable is needed; the choice set variable in this example is SET. After selecting the appropriate variables, press the OK button to perform the analysis.

On the analysis main window, a bar chart is displayed of the significances of each of the choice and non-choice attributes. The chart illustrates that PRICE, BRAND1, BRAND2, and BRAND4 are significant.

You can view other results by pressing the Results button and selecting "Statistics," "Choice probabilities," or "Residual plots" from the ensuing menu. Overall model fit statistics and parameter estimates for the attributes are available from the Statistics window. Probabilities for each choice alternative are available from the Choice Probabilities window. Plots of residual and predicted values are available from the Residual Plots window.

Correspondence Analysis

Categorical data are frequently encountered in the field of market research. Correspondence analysis is a technique that graphically displays relationships among the rows and columns in a contingency table. In the resulting plot there is a point for each row and each column of the table. Rows with similar patterns of counts have points that are close together, and columns with similar patterns of counts have points that are close together.

The CARS data set in the SASUSER library is used as an example (also described in the *SAS/STAT User's Guide*). The CARS data are a sample of individuals who were asked to provide information about themselves and their cars. The pertinent questions for the example are country of origin of their car and their family status.

Simple Correspondence Analysis. Simple correspondence analysis analyzes a contingency table made up of one or more column variables and one or more row variables. To select a data set on which to perform a correspondence analysis, select "File → New dataset/analysis" on the main analysis window. First, select the CARS data set, then select "Correspondence analysis" as the analysis, and then press the OK button.

This example uses raw variables instead of an existing table. The desired type of analysis (simple correspondence analysis) and data layout (raw variables) are default selections on the Variable Selection window. Select ORIGIN, the country of origin of the car, as the column variable and MARITAL, family status, as the row variable to create the desired contingency table. See Figure 6.

‡Each dummy-coded variable has the value of 1 for a different level of the attribute. In this way, each dummy-coded variable represents the presence of that level and the absence of the other levels.

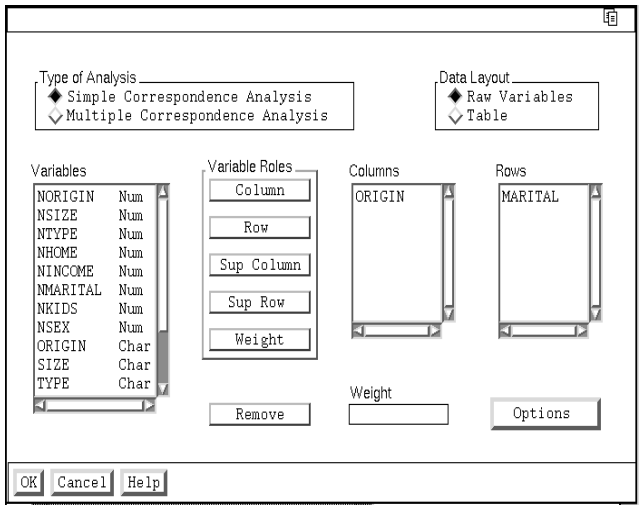


Figure 6. Simple Correspondence Analysis Variable Selection

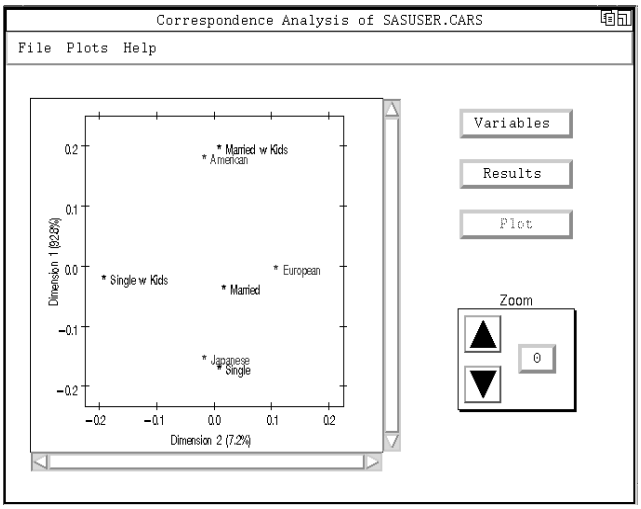


Figure 7. Correspondence Analysis Plot

Plot. The plot displays the column points and row points. The first example in the *SAS/STAT User's Guide* provides an interpretation of the plot. The interpretation has two aspects: what each dimension represents and what the relationship of the points in the dimensional space represents. An interpretation of the vertical dimension is that it represents the country of origin of the cars, with most of the influence coming from whether the car is American or Japanese. The horizontal dimension appears to represent "Single with kids" versus all of the other values. See Figure 7.

Although the row and column points are spread throughout the plot, "married" and "single" appear to be slightly more similar to each other than any of the other points. Keep in mind that distances between row and column points cannot be compared, only distances among row points and distances among column points. However, by treating the country-of-origin points as lines drawn from the 0,0 point and extending off the graph, you can see that the "Married with kids" point is closest to the American car line and the "Single" point is closest to the Japanese car line.

Plot Controls. To enlarge the plot, click on the up arrow in the zoom control box. To return the plot to its zero zoom state, click on the [0] button. If the plot is zoomed, you can move the plot left and right and up and down using the scroll bars.

Results. You can view other results by pressing the Results button and selecting "Inertia table," "Statistics," or "Frequencies." The Inertia Table window lists the singular values and inertias for all possible dimensions in the analysis. The Statistics window displays tables of statistics that aid in the interpretations of the dimensions and the points: the row and column coordinates, the partial contributions to inertia, and the squared cosines. The Frequency Table window displays observed, expected, and deviation contingency tables and row and column profiles.

Multiple Correspondence Analysis. In a multiple correspondence analysis, only column variables are used. They are used to create a Burt table [§] which is then used in the analysis.

The same data set can be used to illustrate multiple correspondence analysis. Return to the Variables Selection window by pressing the Variables button on the main analysis window. See Figure 8. Perform the following steps:

1. Remove the current column and row variables either by double-clicking on them or by selecting them and pressing the Remove button.
2. Select "Multiple Correspondence Analysis" in the Type of Analysis box in the upper left of the window.
3. Select the column variables ORIGIN, TYPE, SIZE, HOME, SEX, INCOME, and MARITAL by clicking on the ORIGIN variable and dragging through the list to the MARITAL variable, then press the Column button.
4. Press the OK button to perform the analysis.

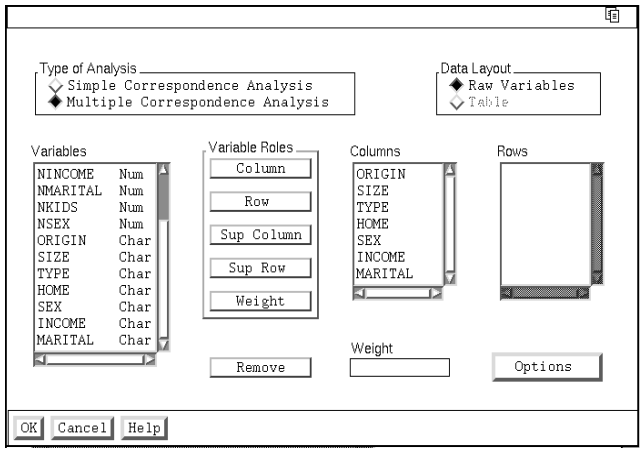


Figure 8. Multiple Correspondence Analysis Variable Selection

The distances between points of different variables can be interpreted in multiple correspondence analysis because they are all column points. However, the multiple correspondence analysis example has more dimensions (12) to interpret and examine than the single correspondence analysis example (2). The total number of dimensions can be examined in the inertia table, which is accessed from the Results button.

By default, a two-dimensional solution is computed. To request a higher dimensional solution, open the Variable Selection window, press the Options button, and select (or enter) the desired number of dimensions.

If you request a three-dimensional (or higher) solution, you can plot the dimensions two at a time by pressing the Plot button and selecting dimensions for the x axis and the y axis.

[§]A Burt table is a partitioned symmetric matrix containing all pairs of crosstabulations among a set of categorical variables. For further explanation, see the *SAS/STAT User's Guide*

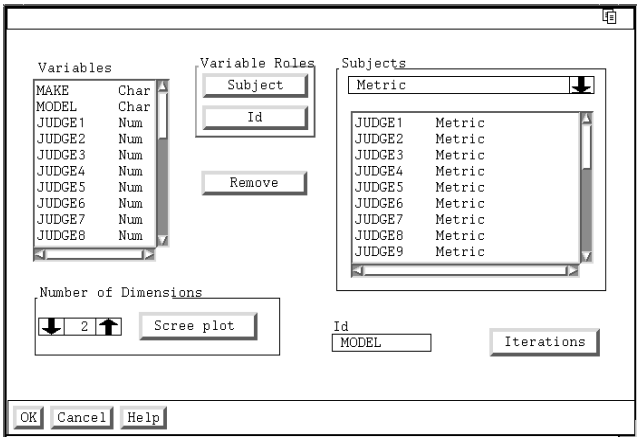


Figure 9. MDPREF Analysis Variable Selection

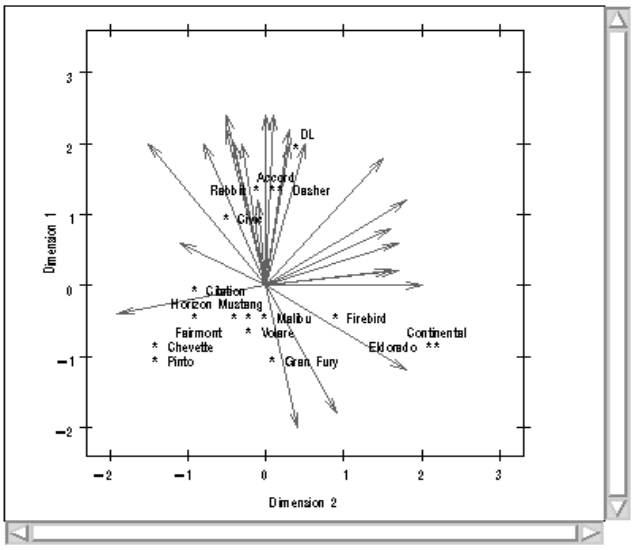


Figure 10. MDPREF Plot

Multidimensional Preference Analysis

With conjoint analysis, respondents indicate their preferences for products that are composed of attributes determined by the experimenter. Sometimes, the data of interest may be preferences of existing products for which relevant attributes are not defined for the respondent. Multidimensional preference analysis (MDPREF) is used to analyze such data.

MDPREF is a principal component analysis of a data matrix whose columns correspond to people and whose rows correspond to objects, the transpose of the usual people by objects multivariate data matrix.

The CARPREF data set in the SASUSER library is used as an example (also described in the *SAS/STAT User's Guide*. It contains data about the preferences of 25 respondents for 17 cars. The preferences are on a scale of 0 to 9 with 0 meaning a very weak preference and 9 meaning a very strong preference. Select the data set and analysis as described in the preceding examples.

As in conjoint analysis, you can choose to perform a metric or non-metric analysis. Choose the measurement type by clicking the arrow in the upper right corner of the window and selecting the desired type. Other, less frequently used, types are available under the "Other" selection. The measurement type is used for all subsequently selected Subject variables. Infrequently, subject variables with different types may be used.

For the example, use the Metric measurement type. Select the preference ratings of each respondent, JUDGE1, JUDGE2, ..., JUDGE25, as Subject variables. Also, select MODEL as the Id variable. See Figure 9.

You also can set the number of dimensions for the analysis; the default is two. A scree plot of the eigenvalues is useful in determining an appropriate number of dimensions. To display the scree plot, press the Scree Plot button. The plot illustrates that the magnitude of the eigenvalues falls off for the first two dimensions; then the plot flattens out for the third and remaining dimensions. From this graph, two dimensions appear appropriate. After closing the Scree Plot window, press the OK button to perform the analysis. See Figure 10.

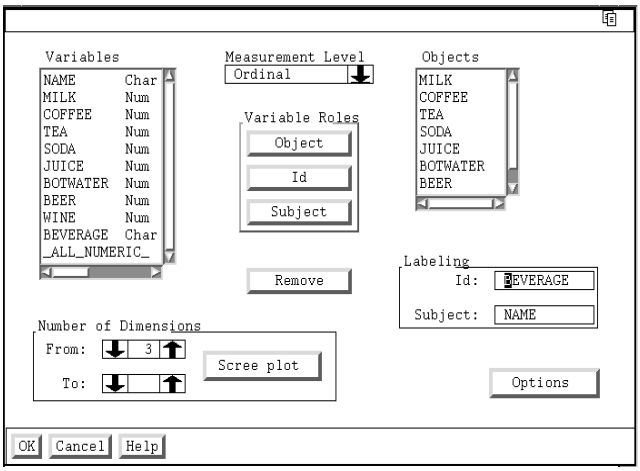


Figure 11. MDS Variable Selection

Results. The plot on the main analysis window contains points for the 17 car models and vectors for the 25 respondents. Interpretations of the two dimensions are 1) the vertical dimension separates foreign and domestic cars in the upper half and lower half, respectively, and 2) the horizontal dimension separates small cars and big cars in the left and right halves, respectively. Respondents prefer cars whose points are closest to their vector. Notice that there are a number of vectors in the upper right quadrant of the plot but there are no cars. This lack of available products to satisfy peoples' preferences indicates a possible niche to fill.

Other results are the "Initial Eigenvalue Plot," "Final Eigenvalue Plot," and "Configuration Table." The Initial Eigenvalue plot is the same as the scree plot on the Variable Selection window. The Final Eigenvalue plot is also a scree plot; it differs from the initial plot only if a measurement type other than Metric is used. The Configuration Table contains the coordinates for the car points.

Multidimensional Scaling

Multidimensional Scaling (MDS) takes subjects' judgments of either similarity or difference of pairs of items and produces a map of the perceived relationship among items.

For example, suppose you ask seven subjects to state their perceived similarity on a 1 to 7 scale for pairs of beverages, with 1 meaning very similar and 7 meaning very different. The beverages are milk, coffee, tea, soda, juice, bottled water, beer, and wine. Someone may state that their perceived similarity between coffee and tea is 3, somewhat similar, or 7, very different. There are 28 possible pairs of these eight beverages.

The data are ordered in an eight observation by eight variable matrix with one matrix (eight observations) for each subject. On the Data Set Selection window, select the BEVERAGE data set in the SASUSER library, then press the OK button. A message window informs you that MDS requires either similarity or distance data. Press the Continue button.

On the Variables Selection window, select the variables MILK, COFFEE, TEA, SODA, JUICE, BOTWATER, BEER, AND WINE as the objects. See Figure 11. It is crucial that the order of the objects is the same as their order in the rows of each matrix. In other words, from the above order, the upper left corner element in the matrix is MILK, MILK (which has a distance of zero) and the element to its

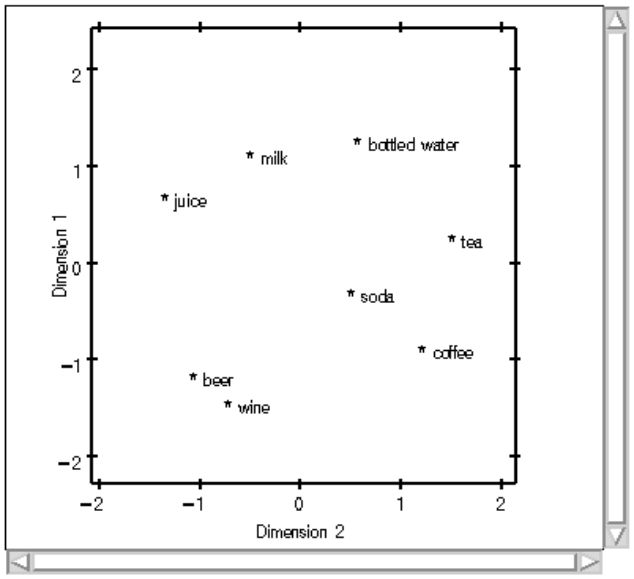


Figure 12. MDS Coordinates Plot

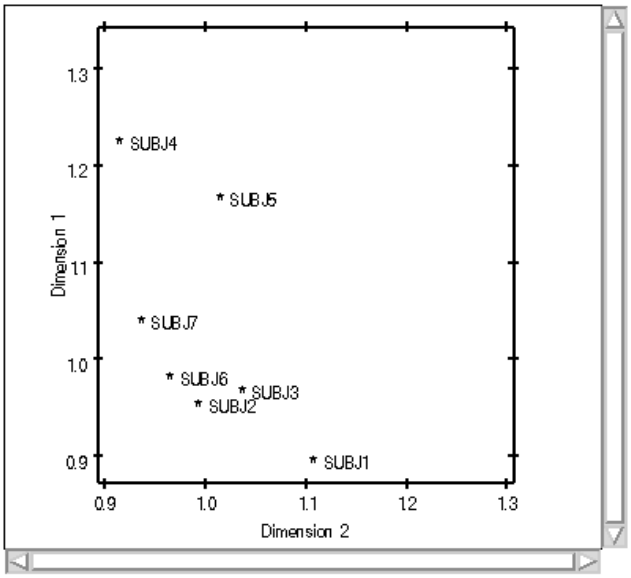


Figure 13. MDS Individual Coefficients Plot

right is MILK, COFFEE.

Also, select BEVERAGE, the beverage names, as the ID variable and NAME, the subject identifiers, as the SUBJECT variable. Because the objects are ordinally-scaled, the ordinal measurement level, the default, is appropriate for this example.

If you think that your subjects may use different perceptual schemes for judging similarity, you can choose to perform an individual differences analysis. Press the Options button and select "Individual Differences Analysis." The data are distances, the default, because larger numbers represent more difference (less similarity). If the data were similarities, you would choose the appropriate selection on the Options window. To close the options window, press the OK button.

As in correspondence analysis and MDPREF analysis, you can set the number of dimensions for the solution. With MDS you have an extra capability; you can solve for several dimensional solutions in one analysis.

Choose a three-dimensional solution by entering a "3" in the input field to the right of the "From:" label or by clicking on the up arrow to its right until the number 3 appears in the input field. As with the other dimensional analyses, a scree plot may be useful in determining the appropriate number of dimensions. You can create the plot by pressing the Scree Plot button.

To continue with the analysis, press the OK button on the Variable Selection window.

Results. As with the correspondence analysis and MDPREF plots, interpreting the MDS plot has two parts: 1) finding a reasonable interpretation for each of the plot dimensions, and 2) finding a reasonable interpretation of the relationship of the points in the plot. See Figure 12.

The presence of bottled water, milk, and juice at the top of the plot and wine, beer, and coffee at the bottom of the plot might indicate a good for you/not so good for you interpretation for the vertical dimension, Dimension 1. The horizontal dimension, Dimension 2, does not have as clear an interpretation. Try to come up with your own interpretation that would have tea, coffee, and water on one side and juice, beer, and wine on the other.

Because you requested a three-dimensional solution, two other plots can be displayed: Dimensions 1 and 3 and Dimensions 2 and 3. To change which dimensions are plotted, press the Plot button and select the desired dimensions. Also, on this window you can choose to display the coefficients of the individual differences analysis instead of the coordinates. To do so, select "Coefficient" at the bottom of the window and press the OK button.

In an individual differences analysis, there is a common perceptual map for all subjects, but different subjects have different weights for each dimension. See Figure 13. SUBJ4 is found to be highest on the vertical axis and lowest on the horizontal axis. In other words, SUBJ4 weights whatever this dimension represents more than do the other subjects and it weights whatever dimension 2 represents less than the other subjects. If the good-for-you interpretation is appropriate for Dimension 1, then it plays a larger role in SUBJ4's perceptual mapping of these beverages than it does for other subjects.

It is possible that SUBJ1, SUBJ2, SUBJ3, SUBJ6, and SUBJ7 may cluster together and SUBJ4 and SUBJ5 may be outliers. Additional subjects may sharpen this possible clustering or eliminate it. MDS is useful in market research for discovering possible perceptual perspectives used by consumers and for revealing possible market segments.

You can display other results by pressing the Results button. These results include Fit statistics, Configuration tables, Residual plots, and the Iteration history. The fit statistics are measures of how well the data fit the model. The Configuration tables contain the coordinates and, optionally, the individual difference coefficients that are used in the plots. The Residual plots allow you to assess the fit of the model graphically. The iteration history contains information about how many iterations were needed and how the criterion changed over the iterations.

Summary

Investigators in the field of market research are interested in how consumers make decisions when they choose to buy products. What attributes are important? Do all people make decisions in the same way? If not, how do they differ? What are the perceptual schemes that people use in their purchasing decisions?

The analyses described in this paper can be used with many different types of data to investigate these questions. The Market Research application makes these analyses easy to use, and it is available in Release 6.11 and subsequent releases with the SAS/STAT product.

Acknowledgements

I would like to thank Greg Goodwin, Warren Kuhfeld, Julie LaBarr, Donna Sawyer, and Maura Stokes for their thoughtful comments on this paper.

Efficient Experimental Design with Marketing Research Applications

Warren F. Kuhfeld

Randall D. Tobias

Mark Garratt

Abstract

We suggest using *D*-efficient experimental designs for conjoint and discrete-choice studies, and discuss orthogonal arrays, nonorthogonal designs, relative efficiency, and nonorthogonal design algorithms. We construct designs for a choice study with asymmetry and interactions and for a conjoint study with blocks and aggregate interactions.*

Introduction

The design of experiments is a fundamental part of marketing research. Experimental designs are required in widely used techniques such as preference-based conjoint analysis and discrete-choice studies (e.g., Carmone and Green 1981; Elrod, Louviere, and Davey 1992; Green and Wind 1975; Huber, et al. 1993; Lazari and Anderson 1994; Louviere 1991; Louviere and Woodworth 1983; Wittink and Cattin 1989). Ideally, marketing researchers prefer *orthogonal* designs. When a linear model is fit with an orthogonal design, the parameter estimates are uncorrelated, which means each estimate is independent of the other terms in the model. More importantly, orthogonality usually implies that the coefficients will have minimum variance, though we discuss exceptions to this rule. For these reasons, orthogonal designs are usually quite good. However, for many practical problems, orthogonal designs are simply not available. In those situations, *nonorthogonal* designs must be used.

*This chapter is a revision of a paper that appeared in *Journal of Marketing Research*, November, 1994, pages 545−557. Warren F. Kuhfeld is now Manager, Multivariate Models R&D, SAS. Randall D. Tobias is now Manager, Linear Models R&D, SAS. Mark Garratt was Vice President, Conway | Milliken & Associates when this paper was first published in 1994 and is now with Miller Brewing Company. The authors thank Jordan Louviere, *JMR* editor Barton Weitz, and three anonymous reviewers for their helpful comments on earlier versions of this article. Thanks to Michael Ford for the idea for the second example. The *JMR* article was based on a presentation given to the AMA Advanced Research Techniques Forum, June 14, 1993, Monterey CA.

Our primary message when this paper was published in 1994 was that marketing researchers should use *D*-efficient experimental designs. This message remains as strong as ever, but today, we have much better tools for accomplishing this than we had in 1994. Most of the revisions of the original paper are due to improvements in the tools. Our new design tool, the `%MktEx` SAS macro, is easier to use than our old tools, and it usually makes better designs. Copies of this chapter (TS-689C) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market .

Orthogonal designs are available for only a relatively small number of very specific problems. They may not be available when some combinations of factor levels are infeasible, a nonstandard number of *runs* (factor level combinations or hypothetical products) is desired, or a nonstandard model is being used, such as a model with interaction or polynomial effects. Consider the problem of designing a discrete choice study in which there are alternative specific factors, different numbers of levels within each factor, and interactions within each alternative. Orthogonal designs are not readily available for this situation, particularly when the number of runs must be limited. When an orthogonal design is not available, an alternative must be chosen − the experiment can be modified to fit some known orthogonal design, which is undesirable for obvious reasons, or a known design can be modified to fit the experiment, which may be difficult and inefficient.

Our primary purpose is to explore a third alternative, the use of optimal (or nearly optimal) designs. Such designs are typically nonorthogonal; however they are efficient in the sense that the variances and covariances of the parameter estimates are minimized. Furthermore, they are always available, even for nonstandard situations. Finding these designs usually requires the aid of a computer, but we want to emphasize that we are not advocating a black-box approach to designing experiments. Computerized design algorithms do not supplant traditional design-creation skills. Our examples show that our best designs were usually found when we used our human design skills to guide the computerized search.

First, we will summarize our main points; next, we will review some fundamentals of the design of experiments; then we will discuss computer-generated designs, a discrete-choice example, and a conjoint analysis example.

Summary of Main Points. Our goal is to explain the benefits of using computer-generated designs in marketing research. Our main points follow:

1. The goodness of an experimental design (*efficiency*) can be quantified as a function of the variances and covariances of the parameter estimates. Efficiency increases as the variances decrease. Designs should not be thought of in terms of the dichotomy between orthogonal versus nonorthogonal but rather as varying along the continuous attribute of efficiency. Some orthogonal designs are less efficient than other (orthogonal and nonorthogonal) alternatives.

2. Orthogonality is not the primary goal in design creation. It is a secondary goal, associated with the primary goal of minimizing the variances of the parameter estimates. Degree of orthogonality is an important consideration, but other factors should not be ignored.

3. For complex, nonstandard situations, computerized searches provide the only practical method of design generation for all but the most sophisticated of human designers. These situations do not have to be avoided just because it is extremely difficult to generate a good design manually.

4. The best approach to design creation is to use the computer as a tool along with traditional design skills, not as a substitute for thinking about the problem.

Background and Assumptions. We present an overview of the theory of efficient experimental design, developed for the general linear model. This topic is well known to specialists in statistical experimentation, though it is not typically taught in design classes. Then we will suggest ways in which this theory can be applied to marketing research problems.

Certain assumptions must be made before applying ordinary general linear model theory to problems in marketing research. The usual goals in linear modeling are to estimate parameters and test hypotheses about those parameters. Typically, independence and normality are assumed. In conjoint analysis, each subject rates all products and separate ordinary-least-squares analyses are run for each subject. This is not a standard general linear model; in particular, observations are not independent and normality cannot be assumed. Discrete choice models, which are nonlinear, are even further removed from the general linear model.

Marketing researchers have always made the critical assumption that designs that are good for general linear models are also good for conjoint analysis and discrete choice. We also make this assumption. Specifically, we assume the following:

1. Market share estimates computed from a conjoint analysis model using a more efficient design will be better than estimates using a less efficient design. That is, more efficient designs mean better estimates of the part-worth utilities, which lead to better estimates of product utility and market share.
2. An efficient design for a linear model is a good design for the multinomial logit (MNL) model used in discrete choice studies.

Investigating these standard assumptions is beyond the scope of this article. However, they are supported by Carson and colleagues (1994), our experiences in consumer product goods, and limited simulation results. Much more research is needed on this topic, particularly in the area of discrete choice.

Design of Experiments

Orthogonal Experimental Designs. An experimental design is a plan for running an experiment. The *factors* of an experimental design are variables that have two or more fixed values, or *levels*. Experiments are performed to study the effects of the factor levels on the dependent variable. In a conjoint or discrete-choice study, the factors are the attributes of the hypothetical products or services, and the response is preference or choice.

A simple experimental design is the *full-factorial design*, which consists of all possible combinations of the levels of the factors. For example, with five factors, two at two levels and three at three levels (denoted $2^2 3^3$), there are 108 possible combinations. In a full-factorial design, all main effects, two-way interactions, and higher-order interactions are estimable and uncorrelated. The problem with a full-factorial design is that, for most practical situations, it is too cost-prohibitive and tedious to have subjects rate all possible combinations. For this reason, researchers often use *fractional-factorial designs*, which have fewer runs than full-factorial designs. The price of having fewer runs is that some effects become confounded. Two effects are *confounded* or *aliased* when they are not distinguishable from each other.

A special type of fractional-factorial design is the *orthogonal array*, in which all estimable effects are uncorrelated. Orthogonal arrays are categorized by their *resolution*. The resolution identifies which effects, possibly including interactions, are estimable. For example, for resolution III designs, all main effects are estimable free of each other, but some of them are confounded with two-factor interactions.

For resolution V designs, all main effects and two-factor interactions are estimable free of each other. Higher resolutions require larger designs. Orthogonal arrays come in specific numbers of runs (e.g., 16, 18, 20, 24, 27, 28) for specific numbers of factors with specific numbers of levels.

Resolution III orthogonal arrays are frequently used in marketing research. The term "orthogonal array," as it is sometimes used in practice, is imprecise. It correctly refers to designs that are both orthogonal and balanced, and hence optimal. It is also imprecisely used to refer to designs that are orthogonal but not balanced, and hence potentially nonoptimal. A design is *balanced* when each level occurs equally often within each factor, which means the intercept is orthogonal to each effect. Imbalance is a generalized form of nonorthogonality, which increases the variances of the parameter estimates.

Design Efficiency. Efficiencies are measures of design goodness. Common measures of the efficiency of an $(N_D \times p)$ design matrix $\mathbf{X}$ are based on the *information matrix* $\mathbf{X}'\mathbf{X}$. The variance-covariance matrix of the vector of parameter estimates $\boldsymbol{\beta}$ in a least-squares analysis is proportional to $(\mathbf{X}'\mathbf{X})^{-1}$. An efficient design will have a "small" variance matrix, and the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}$ provide measures of its "size." Two common efficiency measures are based on the idea of "average eigenvalue" or "average variance." *A-efficiency* is a function of the arithmetic mean of the eigenvalues, which is given by $\text{trace}\,((\mathbf{X}'\mathbf{X})^{-1})/p$. *D-efficiency* is a function of the geometric mean of the eigenvalues, which is given by $|(\mathbf{X}'\mathbf{X})^{-1}|^{1/p}$. A third common efficiency measure, *G-efficiency*, is based on σ_M, the maximum standard error for prediction over the candidate set. All three of these criteria are convex functions of the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}$ and hence are usually highly correlated.

For all three criteria, if a balanced and orthogonal design exists, then it has optimum efficiency; conversely, the more efficient a design is, the more it tends toward balance and orthogonality. A design is balanced and orthogonal when $(\mathbf{X}'\mathbf{X})^{-1}$ is diagonal (for a suitably coded $\mathbf{X}$, see page 91). A design is orthogonal when the submatrix of $(\mathbf{X}'\mathbf{X})^{-1}$, excluding the row and column for the intercept, is diagonal; there may be off-diagonal nonzeros for the intercept. A design is balanced when all off-diagonal elements in the intercept row and column are zero.

These measures of efficiency can be scaled to range from 0 to 100 (for a suitably coded $\mathbf{X}$):

$$
\begin{aligned}
A\text{-efficiency} &= 100 \times \frac{1}{N_D\ \text{trace}\,((\mathbf{X}'\mathbf{X})^{-1})/p} \\
D\text{-efficiency} &= 100 \times \frac{1}{N_D\ |(\mathbf{X}'\mathbf{X})^{-1}|^{1/p}} \\
G\text{-efficiency} &= 100 \times \frac{\sqrt{p/N_D}}{\sigma_M}
\end{aligned}
$$

These efficiencies measure the goodness of the design relative to hypothetical orthogonal designs that may be far from possible, so they are not useful as absolute measures of design efficiency. Instead, they should be used relatively, to compare one design with another for the same situation. Efficiencies that are not near 100 may be perfectly satisfactory.

Figure 1 shows an optimal design in four runs for a simple example with two factors, using interval-measure scales for both. There are three candidate levels for each factor. The full-factorial design is shown by the nine asterisks, with circles around the optimal four design points. As this example shows,

Figure 1
Candidate Set and Optimal Design

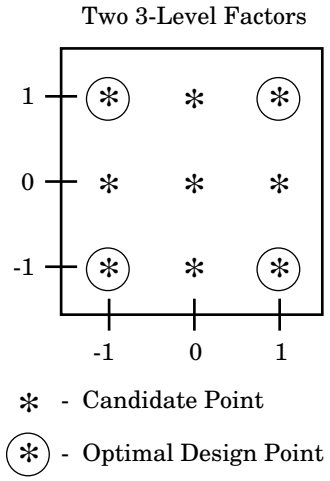


Table 1
Full-Factorial Design

Information Matrix

	Int	X1	X2	X3	-	X4	-	X5	-
Int	108	0	0	0	0	0	0	0	0
X1	0	108	0	0	0	0	0	0	0
X2	0	0	108	0	0	0	0	0	0
X3	0	0	0	108	0	0	0	0	0
-	0	0	0	0	108	0	0	0	0
X4	0	0	0	0	0	108	0	0	0
-	0	0	0	0	0	0	108	0	0
X5	0	0	0	0	0	0	0	108	0
-	0	0	0	0	0	0	0	0	108

100.0000	*D*-efficiency
100.0000	*A*-efficiency
100.0000	*G*-efficiency

efficiency tends to emphasize the corners of the design space. Interestingly, nine different sets of four points form orthogonal designs − every set of four that forms a rectangle or square. Only one of these orthogonal designs is optimal, the one in which the points are spread out as far as possible.

Computer-Generated Design Algorithms. When a suitable orthogonal design does not exist, computer-generated nonorthogonal designs can be used instead. Various algorithms exist for selecting a good set of *design points* from a set of *candidate points*. The candidate points consist of all of the factor-level combinations that can potentially be included in the design − for example the nine points in Figure 1. The number of runs, N_D, is chosen by the researcher. Unlike orthogonal arrays, N_D can be any number as long as $N_D \geq p$.[†] The algorithm searches the candidate points for a set of N_D design points that is optimal in terms of a given efficiency criterion.

It is almost never possible to list all N_D-run designs and choose *the* most efficient or optimal design, because run time is exponential in the number of candidates. For example, with $2^2 3^3$ in 18 runs, there are $108!/(18!(108-18)!) = 1.39 \times 10^{20}$ possible designs. Instead, nonexhaustive search algorithms are used to generate a small number of designs, and the most efficient one is chosen. The algorithms select points for possible inclusion or deletion, then compute rank-one or rank-two updates of some efficiency criterion. The points that most increase efficiency are added to the design. These algorithms invariably find efficient designs, but they may fail to find *the* optimal design, even for the given criterion. For this reason, we prefer to use terms like *information-efficient* and *D-efficiency* over the more common *optimal* and *D-optimal*.

There are many algorithms for generating information-efficient designs. We will begin by describing some of the simpler approaches and then proceed to the more complicated (and more reliable) algo-

[†]In fact, this restriction is not strictly necessary. So called "super-saturated" designs (Booth and Cox, 1962) have more runs than parameters. However, such designs are typically not used in marketing research. The `%MktRuns` SAS macro provides some guidance on the selection of N_D. See page 600.

rithms. Dykstra's (1971) sequential search method starts with an empty design and adds candidate points so that the chosen efficiency criterion is maximized at each step. This algorithm is fast, but it is not very reliable in finding a globally optimal design. Also, it always finds the same design (due to a lack of randomness).

The Mitchell and Miller (1970) simple exchange algorithm is a slower but more reliable method. It improves the initial design by adding a candidate point and then deleting one of the design points, stopping when the chosen criterion ceases to improve. The DETMAX algorithm of Mitchell (1974) generalizes the simple exchange method. Instead of following each addition of a point by a deletion, the algorithm makes excursions in which the size of the design may vary. These three algorithms add and delete points one at a time.

The next two algorithms add and delete points simultaneously, and for this reason, are usually more reliable for finding the truly optimal design; but because each step involves a search over all possible pairs of candidate and design points, they generally run much more slowly (by an order of magnitude). The Fedorov (1972) algorithm simultaneously adds one candidate point and deletes one design point. Cook and Nachtsheim (1980) define a modified Fedorov algorithm that finds the best candidate point to switch with each design point. The resulting procedure is generally as efficient as the simple Fedorov algorithm in finding the optimal design, but it is up to twice as fast. We extensively use one more algorithm, the coordinate exchange algorithm of Meyer and Nachtsheim (1995). This algorithm does not use a candidate set. Instead it refines an initial design by exchanging each level with every other possible level, keeping those exchanges that increase efficiency. In effect, this method uses a virtual candidate set that consists of all possible runs, even when the full-factorial candidate set is too large to generate and store.

Choice of Criterion and Algorithm. Typically, the choice of efficiency criterion is less important than the choice between manual design creation and computerized search. All of the information-efficient designs presented in this article were generated optimizing D-efficiency because it is faster to optimize than A-efficiency and because it is the standard approach. It is also possible to optimize A-efficiency, though the algorithms generally run much more slowly because the rank-one updates are more complicated with A-efficiency. G-efficiency is an interesting ancillary statistic; however, our experience suggests that attempts to maximize G-efficiency with standard algorithms do not work very well.

The candidate set search algorithms, ordered from the fastest and least reliable to the slowest and most reliable, are: sequential, simple exchange, DETMAX, and modified Fedorov. We always use the modified Fedorov and coordinate exchange algorithms even for extremely large problems; we never even try the other algorithms. For small problems in which the full factorial is no more than a few thousand runs, modified Fedorov tends to work best. For larger problems, coordinate exchange tends to be better. Our latest software, the `%MktEx` macro, tries a few iterations with both methods, then picks the best method for that problem and continues on with more iterations using just the chosen method. See Kuhfeld (2003) page 546 and all of the examples starting on page 81.

Nonlinear Models. The experimental design problem is relatively simple for linear models and much more complicated for nonlinear models. The usual goal when creating a design is to minimize some function of the variance matrix of the parameter estimates, such as the determinant. For linear models, the variance matrix is proportional to $(\mathbf{X}'\mathbf{X})^{-1}$, and so the design optimality problem is well-posed. However, for nonlinear models, such as the multinomial logit model used with discrete-choice data, the variance matrix depends on the true values of the parameters themselves. (See pages 61, 481, and

303 for more on efficient choice designs based on assumptions about the parameters.) Thus in general, there may not exist a design for a discrete-choice experiment that is always optimal. However, Carson and colleagues (1994) and our experience suggest that *D*-efficient designs work well for discrete-choice models.

Lazari and Anderson (1994) provide a catalog of designs for discrete-choice models, which are good for certain specific problems. For those specific situations, they may be as good as or better than computer-generated designs. However, for many real problems, cataloged designs cannot be used without modification, and modification can reduce efficiency. We carry their work one step further by discussing a general computerized approach to design generation.

Design Comparisons

Comparing Orthogonal Designs. All orthogonal designs are not perfectly or even equally efficient. In this section, we compare designs for $2^2 3^3$. Table 1 gives the information matrix, $\mathbf{X}'\mathbf{X}$, for a full-factorial design using an orthogonal coding. The matrix is a diagonal matrix with the number of runs on the diagonal. The three efficiency criteria are printed after the information matrix. Because this is a full-factorial design, all three criteria show that the design is 100% efficient. The variance matrix (not shown) is $(1/108)\mathbf{I} = 0.0093\mathbf{I}$.

Table 2 shows the information matrix, efficiencies, and variance matrix for a classical 18-run orthogonal design for $2^2 3^3$, Chakravarti's (1956) L_{18}, for comparison with information-efficient designs with 18 runs. (The SAS ADX menu system was used to generate the design. Tables A1 and A2 contain the factor levels and the orthogonal coding used in generating Table 2.) Note that although the factors are all orthogonal to each other, X1 is not balanced. Because of this, the main effect of X1 is estimated with a higher variance (0.063) than X2 (0.056).

The precision of the estimates of the parameters critically depends on the efficiency of the experimental design. The parameter estimates in a general linear model are always unbiased (in fact, best linear unbiased [BLUE]) no matter what design is chosen. However, all designs are not equally efficient. In fact, all orthogonal designs are not equally efficient, even when they have the same factors and the same number of runs. Efficiency criteria can be used to help choose among orthogonal designs. For example, the orthogonal design in Tables 3 and A3 (from the Green and Wind 1975 carpet cleaner example) for $2^2 3^3$ is less *D*-efficient than the Chakravarti L_{18} ($97.4166/98.6998 = 0.9870$). The Green and Wind design can be created from a 3^5 balanced orthogonal array by collapsing two of the three-level factors into two-level factors. In contrast, the Chakravarti design is created from a $2^1 3^4$ balanced orthogonal array by collapsing only one of the three-level factors into a two-level factor. The extra imbalance makes the Green and Wind design less efficient. (Note that the off-diagonal 2 in the Green and Wind information matrix does not imply that X1 and X2 are correlated. It is an artifact of the coding scheme. The off-diagonal 0 in the variance matrix shows that X1 and X2 are uncorrelated.)

Orthogonal Versus Nonorthogonal Designs. Orthogonal designs are not always more efficient than nonorthogonal designs. Tables 4 and A4 show the results for an information-efficient, main-effects-only design in 18 runs. The OPTEX procedure of SAS software was used to generate the design, using the modified Fedorov algorithm. The information-efficient design is slightly better than the classical L_{18}, in terms of the three efficiency criteria. In particular, the ratio of the *D*-efficiencies for the classical and information-efficient designs are $99.8621/98.6998 = 1.0118$. In contrast to the L_{18}, this design is

Table 2
Orthogonal Design
Information Matrix

	Int	X1	X2	X3	-	X4	-	X5	-
Int	18	6	0	0	0	0	0	0	0
X1	6	18	0	0	0	0	0	0	0
X2	0	0	18	0	0	0	0	0	0
X3	0	0	0	18	0	0	0	0	0
-	0	0	0	0	18	0	0	0	0
X4	0	0	0	0	0	18	0	0	0
-	0	0	0	0	0	0	18	0	0
X5	0	0	0	0	0	0	0	18	0
-	0	0	0	0	0	0	0	0	18

98.6998 *D*-efficiency
97.2973 *A*-efficiency
94.8683 *G*-efficiency

Variance Matrix

	Int	X1	X2	X3	-	X4	-	X5	-
Int	63	-21	0	0	0	0	0	0	0
X1	-21	63	0	0	0	0	0	0	0
X2	0	0	56	0	0	0	0	0	0
X3	0	0	0	56	0	0	0	0	0
-	0	0	0	0	56	0	0	0	0
X4	0	0	0	0	0	56	0	0	0
-	0	0	0	0	0	0	56	0	0
X5	0	0	0	0	0	0	0	56	0
-	0	0	0	0	0	0	0	0	56

Note: multiply variance matrix values by 0.001.

Table 3
Green & Wind Orthogonal Design
Information Matrix

	Int	X1	X2	X3	-	X4	-	X5	-
Int	18	-6	-6	0	0	0	0	0	0
X1	-6	18	2	0	0	0	0	0	0
X2	-6	2	18	0	0	0	0	0	0
X3	0	0	0	18	0	0	0	0	0
-	0	0	0	0	18	0	0	0	0
X4	0	0	0	0	0	18	0	0	0
-	0	0	0	0	0	0	18	0	0
X5	0	0	0	0	0	0	0	18	0
-	0	0	0	0	0	0	0	0	18

97.4166 *D*-efficiency
94.7368 *A*-efficiency
90.4534 *G*-efficiency

Variance Matrix

	Int	X1	X2	X3	-	X4	-	X5	-
Int	69	21	21	0	0	0	0	0	0
X1	21	63	0	0	0	0	0	0	0
X2	21	0	63	0	0	0	0	0	0
X3	0	0	0	56	0	0	0	0	0
-	0	0	0	0	56	0	0	0	0
X4	0	0	0	0	0	56	0	0	0
-	0	0	0	0	0	0	56	0	0
X5	0	0	0	0	0	0	0	56	0
-	0	0	0	0	0	0	0	0	56

Notes: multiply variance matrix values by 0.001.

balanced in all the factors, but X1 and X2 are slightly correlated, shown by the 2's off the diagonal. There is no *completely* orthogonal (that is, both balanced and orthogonal) $2^2 3^3$ design in 18 runs.[‡] The nonorthogonality in Table 4 has a much smaller effect on the variances of X1 and X2 (1.2%) than the lack of balance in the orthogonal design in Table 2 has on the variance of X2 (12.5%). In optimizing efficiency, the search algorithms effectively optimize both balance and orthogonality. In contrast, in some orthogonal designs, balance and efficiency may be sacrificed to preserve orthogonality.

This example shows that a nonorthogonal design may be more efficient than an unbalanced orthogonal design. We have seen this phenomenon with other orthogonal designs and in other situations as well. *Preserving orthogonality at all costs can lead to decreased efficiency.* Orthogonality was extremely important in the days before general linear model software became widely available. Today, it is more important to consider efficiency when choosing a design. These comparisons are interesting because they illustrate in a simple example how lack of orthogonality and imbalance affect efficiency. Nonorthogonal designs will never be more efficient than balanced orthogonal designs, when they exist. However, nonorthogonal designs may well be more efficient than unbalanced orthogonal designs. Although this point is interesting and important, what is most important is that good nonorthogonal designs exist

[‡]In order for the design to be both balanced and orthogonal, the number of runs must be divisible by 2, 3, 2×2, 3×3, and 2×3. Since 18 is not divisible by 2×2, orthogonality and balance are not both simultaneously possible for this design.

Table 4
Information-Efficient Orthogonal Design
Information Matrix

	Int	X1	X2	X3	-	X4	-	X5	-
Int	18	0	0	0	0	0	0	0	0
X1	0	18	2	0	0	0	0	0	0
X2	0	2	18	0	0	0	0	0	0
X3	0	0	0	18	0	0	0	0	0
-	0	0	0	0	18	0	0	0	0
X4	0	0	0	0	0	18	0	0	0
-	0	0	0	0	0	0	18	0	0
X5	0	0	0	0	0	0	0	18	0
-	0	0	0	0	0	0	0	0	18

99.8621 *D*-efficiency
99.7230 *A*-efficiency
98.6394 *G*-efficiency

Variance Matrix

	Int	X1	X2	X3	-	X4	-	X5	-
Int	56	0	0	0	0	0	0	0	0
X1	0	56	-6	0	0	0	0	0	0
X2	0	-6	56	0	0	0	0	0	0
X3	0	0	0	56	0	0	0	0	0
-	0	0	0	0	56	0	0	0	0
X4	0	0	0	0	0	56	0	0	0
-	0	0	0	0	0	0	56	0	0
X5	0	0	0	0	0	0	0	56	0
-	0	0	0	0	0	0	0	0	56

Notes: multiply variance matrix values by 0.001.
The diagonal entries for X1 and X2 are slightly larger at 0.0563 than the other diagonal entries of 0.0556.

Table 5
Unrealistic Combinations Excluded
Information Matrix

	Int	X1	X2	X3	-	X4	-	X5	-
Int	18	0	0	0	0	0	0	0	0
X1	0	18	2	0	0	0	0	0	0
X2	0	2	18	0	0	0	0	0	0
X3	0	0	0	18	0	0	0	0	0
-	0	0	0	0	18	0	0	0	0
X4	0	0	0	0	0	18	0	-6	5
-	0	0	0	0	0	0	18	5	0
X5	0	0	0	0	0	-6	5	18	0
-	0	0	0	0	0	5	0	0	18

96.4182 *D*-efficiency
92.3190 *A*-efficiency
91.0765 *G*-efficiency

Variance Matrix

	Int	X1	X2	X3	-	X4	-	X5	-
Int	56	0	0	0	0	0	0	0	0
X1	0	56	-6	0	0	0	0	0	0
X2	0	-6	56	0	0	0	0	0	0
X3	0	0	0	56	0	0	0	0	0
-	0	0	0	0	56	0	0	0	0
X4	0	0	0	0	0	69	-7	25	-20
-	0	0	0	0	0	-7	61	-20	2
X5	0	0	0	0	0	25	-20	69	-7
-	0	0	0	0	0	-20	2	-7	61

Notes: multiply variance matrix values by 0.001.

in many situations in which no orthogonal designs exist.

Design Considerations

Codings and Efficiency. The specific design matrix coding does not affect the relative *D*-efficiency of competing designs. Rank-preserving linear transformations are immaterial, whether they are from full-rank indicator variables to effects coding or to an orthogonal coding such as the one shown in Table A2. Any full-rank coding is equivalent to any other. The absolute *D*-efficiency values will change, but the ratio of two *D*-efficiencies for competing designs is constant. Similarly, scale for quantitative factors does not affect relative efficiency. The proof is simple. If design $\mathbf{X}_1$ is recoded to $\mathbf{X}_1\mathbf{A}$, then $|(\mathbf{X}_1\mathbf{A})'(\mathbf{X}_1\mathbf{A})| = |\mathbf{A}'\mathbf{X}_1'\mathbf{X}_1\mathbf{A}| = |\mathbf{A}\mathbf{A}'||\mathbf{X}_1'\mathbf{X}_1|$. The relative efficiency of design $\mathbf{X}_1$ compared to $\mathbf{X}_2$ is the same as $\mathbf{X}_1\mathbf{A}$ compared to $\mathbf{X}_2\mathbf{A}$, since the $|\mathbf{A}\mathbf{A}'|$'s terms in efficiency ratios will cancel. We prefer the orthogonal coding because it yields "nicer" information matrices with the number of runs on the

diagonal and efficiency values scaled so that 100 means perfect efficiency.

Quantitative Factors. The factors in an experimental design are usually qualitative (nominal), but quantitative factors such as price are also important. With quantitative factors, the choice of levels depends on the function of the original variable that is modeled. To illustrate, consider a pricing study in which price ranges from \$0.99 to \$1.99. If a linear function of price is modeled, only two levels of price should be used − the end points (\$0.99 and \$1.99). Using prices that are closer together is inefficient; the variances of the estimated coefficients will be larger. The efficiency of a given design is affected by the coding of quantitative factors, even though the relative efficiency of competing designs is unaffected by coding. Consider treating the second factor of the Chakravarti L_{18}, $2^2 3^3$ as linear. It is nearly three times more *D*-efficient to use \$0.99 and \$1.99 as levels instead of \$1.49 and \$1.50 ($58.6652/21.0832 = 2.7826$). To visualize this, imagine supporting a yard stick (line) on your two index fingers (with two points). The effect on the slope of the yard stick of small vertical changes in finger locations is much greater when your fingers are closer together than when they are near the ends.

Of course there are other considerations besides the numerical measure of efficiency. It would not make sense to use prices of \$0.01 and \$1,000,000 just because that is more efficient than using \$0.99 and \$1.99. The model is almost certainly not linear over this range. To maximize efficiency, the range of experimentation for quantitative factors should be as large as possible, given that the model is plausible.

The number of levels also affects efficiency. Because two points define a line, it is inefficient to use more than two points to model a linear function. When a quadratic function is used (x and x^2 are included in the model), three points are needed − the two extremes and the midpoint. Similarly, four points are needed for a cubic function. More levels are needed when the functional form is unknown. Extra levels allow for the examination of complicated nonlinear functions, with a cost of decreased efficiency for the simpler functions. When the function is assumed to be linear, experimental points should not be spread throughout the range of experimentation. See Kuhfeld and Garratt (1992) for a discussion of nonlinear functions of quantitative factors in conjoint analysis. See Kuhfeld (2003) page 643.

Most of the discussion outside this section has concerned qualitative (nominal) factors, even if that was not always explicitly stated. Quantitative factors complicate general design characterizations. For example, we previously stated that "if a balanced and orthogonal design exists, then it has optimum efficiency." This statement must be qualified to be absolutely correct. The design would not be optimal if, for example, a three-level factor were treated as quantitative and linear.

Nonstandard Algorithms and Criteria. Other researchers have proposed other algorithms and criteria. Steckel, DeSarbo, and Mahajan (SDM) (1991) proposed using computer-generated experimental designs for conjoint analysis to exclude unacceptable combinations from the design. They considered a nonstandard measure of design goodness based on the determinant of the (m-factor × m-factor) correlation matrix ($|\mathbf{R}|$) instead of the customary determinant of the (p-parameter × p-parameter) variance matrix ($|(\mathbf{X}'\mathbf{X})^{-1}|$). The SDM approach represents each factor by a single column rather than as a set of coded indicator variables. Designs generated using nonstandard criteria will not generally be efficient in terms of standard criteria like *A*-efficiency and *D*-efficiency, so the parameter estimates will have larger variances. To illustrate graphically, refer to Figure 1. The criterion $|\mathbf{R}|$ cannot distinguish between any of the nine different four-point designs, constructed from this candidate set, that form a square or a rectangle. All are orthogonal; only one is optimal.

We generated a *D*-efficient design for SDM's example, treating the variables as all quantitative (as they did). The $|\mathbf{R}|$ for the SDM design is 0.9932, whereas the $|\mathbf{R}|$ for the information-efficient design is 0.9498. The SDM approach works quite well in maximizing $|\mathbf{R}|$; hence the SDM design is close to orthogonal. However, efficiency is not always maximized when orthogonality is maximized. The SDM design is approximately 75% as *D*-efficient as a design generated with standard criteria and algorithms ($70.1182/93.3361 = 0.7512$).

Choosing a Design. Computerized search algorithms generate many designs, from which the researcher must choose one. Often, several designs are tied or nearly tied for the best *D*, *A*, and *G* information efficiencies. A design should be chosen after examining the design matrix, its information matrix, its variance matrix, factor correlations, and levels frequencies. *It is important to look at the results and not just routinely choose the design from the top of the list.*

For studies involving human subjects, achieving at least nearly-balanced designs is an important consideration. Consider for example a two-level factor in an 18-run design in which one level occurs 12 times and the other level occurs 6 times versus a design in which each level occurs 9 times. Subjects who see one level more often than the other may try to read something into the study and adjust their responses in some way. Alternatively, subjects who see one level most often may respond differently than those who see the second level most often. These are not concerns with nearly balanced designs. One design selection strategy is to choose the most balanced design from the top few.

Many other strategies can be used. Perhaps correlation and imprecision are tolerable in some variables but not in others. Perhaps imbalance is tolerable, but the correlations between the factors should be minimal. Goals will no doubt change from experiment to experiment. Choosing a suitable design can be part art and part science. Efficiency should always be considered when choosing between alternative designs, even manually created designs, but it is not the only consideration.*

Adding Observations or Variables. These techniques can be extended to augment an existing design. A design with r runs can be created by augmenting m specified combinations (established brands or existing combinations) with $r - m$ combinations chosen by the algorithm. Alternatively, combinations that must be used for certain variables can be specified, and then the algorithm picks the levels for the other variables (Cook and Nachtsheim 1989). This can be used to ensure that some factors are balanced or uncorrelated; another application is blocking factors. Using design algorithms, we are able to establish numbers of runs and blocking patterns that fit into practical fielding schedules.

Designs with Interactions. There is a growing interest in using both main effects and interactions in discrete-choice models, because interaction and cross-effect terms may improve aggregate models (Elrod, Louviere, and Davey 1992). The current standard for choice models is to have all main-effects estimable both within and between alternatives. It is often necessary to estimate interactions within alternatives, such as in modeling separate price elasticities for product forms, sizes or packages. For certain classes of designs, in which a brand appears in only a subset of runs, it is often necessary to have estimable main-effects, own-brand interactions, and cross-effects in the submatrix of the design in which that brand is present. One way to ensure estimability is to include in the model interactions between the alternative-specific variables of interest and the indicator variables that control for presence or absence of the brand in the choice set. Orthogonal designs that allow for estimation of interactions are usually very large, whereas efficient nonorthogonal designs can be generated for any linear model, including models with interactions, and for any (reasonable) number of runs.

*See the `%MktEval` macro, page 542, for a tool that helps evaluate designs.

Unrealistic Combinations. It is sometimes useful to exclude certain combinations from the candidate set. SDM (1991) have also considered this problem. Consider a discrete-choice model for several brands and their line extensions. It may not make sense to have a choice set in which the line extension is present and the "flagship" brand absent. Of course, as we eliminate combinations, we may introduce unavoidable correlation between the parameter estimates. In Tables 5 and A5, the twenty combinations where (X1 = 1 and X2 = 1 and X3 = 1) or (X4 = 1 and X5 = 1) were excluded and an 18-run design was generated with the modified Fedorov algorithm. With these restrictions, all three efficiency criteria dropped, for example 96.4182/99.8621 = 0.9655. This shows that the design with excluded combinations is almost 97% as *D*-efficient as the best (unrestricted) design. The information matrix shows that X1 and X2 are correlated, as are X4 and X5. This is the price paid for obtaining a design with only realistic combinations.

In the "Quantitative Factors" section, we stated "Because two points define a line, it is inefficient to use more than two points to model a linear function." When unrealistic combinations are excluded, this statement may no longer be true. For example, if minimum price with maximum size is excluded, an efficient design may involve the median price and size.

Choosing the Number of Runs. Deciding on a number of runs for the design is a complicated process; it requires balancing statistical concerns of estimability and precision with practical concerns like time and subject fatigue. Optimal design algorithms can generate designs for any number of runs greater than or equal to the number of parameters. The variances of the least-squares estimates of the part-worth utilities will be roughly inversely proportional to both the *D*-efficiency and the number of runs. In particular, for a given number of runs, a *D*-efficient design will give more accurate estimates than would be obtained with a less efficient design. A more precise value for the number of choices depends on the ratio of the inherent variability in subject ratings to the absolute size of utility that is considered important. Subject concerns probably outweigh the statistical concerns, and the best course is to provide as many products as are practical for the subjects to evaluate. In any case, the use of information-efficient designs provides more flexibility than manual methods.

Asymmetry in the Number of Levels of Variables. In many practical applications of discrete-choice modeling, there is asymmetry in the number of factor levels, and interaction and polynomial parameters must be estimated. One common method for generating choice model designs is to create a resolution III orthogonal array and modify it. The starting point is a $q^{\Sigma M_j}$ design, where q represents a fixed number of levels across all attributes and M_j represents the number of attributes for brand j. For example, in the "Consumer Food Product" example in a subsequent section, with five brands with 1, 3, 1, 2, and 1 attributes and with each attribute having at most four levels, the starting point is a 4^8 orthogonal array. Availability cross-effect designs are created by letting one of the M_j variables function as an indicator for presence/absence of each brand or by allowing one level of a common variable (price) to operate as the indicator. These methods are fairly straightforward to implement in designs in which the factor levels are all the same, but they become quite difficult to set up when there are different numbers of levels for some factors or in which specific interactions must be estimable.

Asymmetry in the number of levels of factors may be handled either by using the "coding down" approach (Addelman 1962b) or by expansion. In the coding down approach, designs are created using factors that have numbers of levels equal to the largest number required in the design. Factors that have fewer levels are created by recoding. For example, a five-level factor {1, 2, 3, 4, 5} can be recoded into a three-level factor by duplicating levels {1, 1, 2, 2, 3}. The variables will still be orthogonal because the indicator variables for the recoding are in a subspace of the original space. However, recoding introduces imbalance and inefficiency. The second method is to expand a factor at k-levels

into several variables at some fraction of k-levels. For example, a four-level variable can be expanded into three orthogonal two-level variables. In many cases, both methods must be used to achieve the required design.

These approaches are difficult for a simple main-effect design of resolution III and extremely difficult when interactions between asymmetric factors must be considered. In practical applications, asymmetry is the norm. Consider for example the form of an analgesic product. One brand may have caplet and tablet varieties, another may have tablet, liquid, and chewable forms. In a discrete-choice model, these two brand/forms must be modeled as asymmetric alternative-specific factors. If we furthermore anticipated that the direct price elasticity might vary, depending on the form, we would need to estimate the interaction of a quantitative price variable with the nominal-level form variable.

Computerized search methods are simpler to use by an order of magnitude. They provide asymmetric designs that are usually nearly balanced, as well as providing easy specification for interactions, polynomials and continuous by class effects.

Strategies for Many Variables. Consider generating a 3^{15} design in 36 runs. There are 14,348,907 combinations in the full-factorial design, which is too many to use even for a candidate set. For problems like this, the coordinate exchange algorithm (Meyer and Nachtsheim 1995)] works well. The `%MktEx` macro which uses coordinate exchange with a partial orthogonal array initialization easily finds design over 98.9% D-efficient. Even designs with over 100 variables can be created this way.

Examples

Choice of Consumer Food Products. Consider the problem of using a discrete choice model to study the effect of introducing a retail food product. This may be useful, for example, to refine a marketing plan or to optimize a product prior to test market. A typical brand team will have several concerns such as knowing the potential market share for the product, examining the source of volume, and providing guidance for pricing and promotions. The brand team may also want to know what brand attributes have competitive clout and want to identify competitive attributes to which they are vulnerable.

To develop this further, assume our client wishes to introduce a line extension in the category of frozen entrees. The client has one nationally branded competitor, a regional competitor in each of three regions, and a profusion of private label products at the grocery chain level. The product comes in two different forms: stove-top or microwaveable. The client believes that the private labels are very likely to mimic this line extension and to sell it at a lower price. The client suspects that this strategy on the part of private labels may work for the stove-top version but not for the microwaveable, in which they have the edge on perceived quality. They also want to test the effect of a shelf-talker that will draw attention to their product.

This problem may be set up as a discrete choice model in which a respondent's choice among brands, given choice set C_a of available brands, will correspond to the brand with the highest utility. For each brand i, the utility U_i is the sum of a systematic component V_i and a random component e_i. The probability of choosing brand i from choice set C_a is therefore:

$$P(i|C_a) = P(U_i > \max(U_j)) = P(V_i + e_i > \max(V_j + e_j)) \quad \forall \quad (j \neq i) \in C_a$$

Table 6
Factors and Levels

Alternative	Factor	Levels	Brand	Description
1	X1	4	Client	3 prices + absent
2	X2	4	Client Line Extension	3 prices + absent
	X3	2		microwave/stove-top
	X4	2		shelf-talker yes/no
3	X5	3	Regional	2 prices + absent
4	X6	3	Private Label	2 prices + absent
	X7	2		microwave/stove-top
5	X8	3	Competitor	2 prices + absent

Assuming that the e_i follow an extreme value type I distribution, the conditional probabilities $P(i|C_a)$ can be found using the MNL formulation of McFadden (1974)

$$P(i|C_a) = \exp(V_i)/\sum_{j\in C_a}\exp(V_j)$$

One of the consequences of the MNL formulation is the property of independence of irrelevant alternatives (IIA). Under the assumption of IIA, all cross-effects are assumed to be equal, so that if a brand gains in utility, it draws share from all other brands in proportion to their current shares. Departures from IIA exist when certain subsets of brands are in more direct competition and tend to draw a disproportionate amount of share from each other than from other members in the category. One way to capture departures from IIA is to use the mother logit formulation of McFadden (1974). In these models, the utility for brand i is a function of both the attributes of brand i and the attributes of other brands. The effect of one brand's attributes on another is termed a *cross-effect*. In the case of designs in which only subsets C_a of the full shelf set C appear, the effect of the presence or absence of one brand on the utility of another is termed an *availability cross-effect*.

In the frozen entree example, there are five alternatives: the client, the client's line extension, a national branded competitor, a regional brand and a private label brand. Several regional and private labels can be tested in each market, then aggregated for the final model. Note that the line extension is treated as a separate alternative rather than as a "level" of the client brand. This enables us to model the source of volume for the new entry and to quantify any cannibalization that occurs. Each brand is shown at either two or three price points. Additional price points are included so that quadratic models of price elasticity can be tested. The indicator for the presence or absence of any brand in the shelf set is coded using one level of the price variable. The layout of factors and levels is given in Table 6.

In addition to intercepts and main effects, we also require that all two-way interactions within alternatives be estimable: X2*X3, X2*X4, X3*X4 for the line extension and X6*X7 for private labels. This will enable us to test for different price elasticities by form (stove-top versus microwaveable) and to

see if the promotion works better combined with a low price or with different forms. Using a linear model for X1-X8, the total number of parameters including the intercept, all main effects, and two-way interactions with brand is 25. This assumes that price is treated as qualitative. The actual number of parameters in the choice model is larger than this because of the inclusion of cross-effects. Using indicator variables to code availability, the systematic component of utility for brand i can be expressed as:

$$V_i = a_i + \sum_k (b_{ik} \times x_{ik}) + \sum_{j \neq i} z_j (d_{ij} + \sum_l (g_{ijl} \times x_{jl}))$$

where

a_i = intercept for brand i
b_{ik} = effect of attribute k for brand i, where $k = 1, .., K_i$
x_{ik} = level of attribute k for brand i
d_{ij} = availability cross-effect of brand j on brand i
z_j = availability code $= \begin{cases} 1 & \text{if } j \in C_a, \\ 0 & \text{otherwise} \end{cases}$
g_{ijl} = cross-effect of attribute l for brand j on brand i, where $l = 1, .., L_j$
x_{jl} = level of attribute l for brand j.

The x_{ik} and x_{jl} might be expanded to include interaction and polynomial terms. In an availability cross-effects design, each brand is present in only a fraction of choice sets. The size of this fraction or subdesign is a function of the number of levels of the alternative-specific variable that is used to code availability (usually price). For example, if price has three valid levels and a fourth "zero" level to indicate absence, then the brand will appear in only three out of four runs. Following Lazari and Anderson (1994), the size of each subdesign determines how many model equations can be written for each brand in the discrete choice model. If X_i is the subdesign matrix corresponding to V_i, then each X_i must be full rank to ensure that the choice set design provides estimates for all parameters.

To create the design, a full candidate set is generated consisting of 3456 runs. It is then reduced to 2776 runs that contain between two and four brands so that the respondent is never required to compare more than four brands at a time. In the algorithm model specification, we designate all variables as classification variables and require that all main effects and two-way interactions within brands be estimable. The number of runs to use follows from a calculation of the number of parameters that we wish to estimate in the various submatrices $\mathbf{X}_i$ of $\mathbf{X}$. Assuming that there is a category "None" used as a reference cell, the numbers of parameters required for various alternatives are shown in the Table 7 along with the size of submatrices (rounded down) for various numbers of runs. Parameters for quadratic price models are given in parentheses. Note that the effect of private label being in a microwaveable or stove-top form (stove/micro cross-effect) is an explicit parameter under the client line extension.

The number of runs chosen was N=26. This number provides adequate degrees of freedom for the linear price model and will also allow estimation of direct quadratic price effects. To estimate quadratic cross-effects for price would require 32 runs at the very least. Although the technique of using two-way interactions between nominal level variables will usually guarantee that all direct and cross-effects are estimable, it is sometimes necessary and a good practice to check the ranks of the submatrices for more complex models (Lazari and Anderson 1994). Creating designs for cross effects can be difficult, even with the aid of a computer.

Table 7
Parameters

Effect	Client	Client Line Extension	Regional	Private Label	Competitor
intercept	1	1	1	1	1
availability cross-effects	4	4	4	4	4
direct price effect	1 (2)	1 (2)	1	1	1
price cross-effects	4 (8)	4 (8)	4	4	4
stove versus microwave	-	1	-	1	-
stove/micro cross-effects	-	1	-	-	-
shelf-talker	-	1	-	-	-
price*stove/microwave	-	1 (2)	-	1	-
price*shelf-talker	-	1 (2)	-	-	-
stove/micro*shelf-talker	-	1	-	-	-
Total	10 (15)	16 (23)	10	12	10
Subdesign size					
22 runs	16	16	14	14	14
26 runs	19	19	17	17	17
32 runs	24	24	21	21	21

It took approximately 4.5 minutes to generate a design. The final (unrandomized) design in 26 runs is in table A6.† The coded choice sets are presented in Table A7 and the level frequencies are presented in Table A8. Note that the runs have been ordered by the presence/absence of the shelf-talker. This ordering is done because it is unrealistic to think that once the respondent's attention has been drawn in by the promotion, it can just be "undrawn." The two blocks that result may be shown to two groups of people or to the same people sequentially. It would be extremely difficult and time consuming to generate a design for this problem without a computerized algorithm.

Conjoint Analysis with Aggregate Interactions. This example illustrates creating a design for a conjoint analysis study. The goal is to create a 3^6 design in 90 runs. The design consists of five blocks of 18 runs each, so each subject will only have to rate 18 products. Within each block, main-effects must be estimable. In the aggregate, all main-effects and two-way interactions must be estimable. (The utilities from the main-effects models will be used to cluster subjects, then in the aggregate analysis, clusters of subjects will be pooled across blocks and the blocking factor ignored.) Our goal is to create a design that is simultaneously efficient in six ways. Each of the five blocks should be an efficient design for a first-order (main-effects) model, and the aggregate design should be efficient for the second-order (main-effects and two-way interactions) model. The main-effects models for the five blocks have $5(1 + 6(3 - 1)) = 65$ parameters. In addition, there are $(6 \times 5/2)(3 - 1)(3 - 1) = 60$ parameters for interactions in the aggregate model. There are more parameters than runs, but not all

†This is the design that was presented in the original 1994 paper, which due to differences in the random number seeds, is not reproduced by today's tools.

parameters will be simultaneously estimated.

One approach to this problem is the Bayesian regression method of DuMouchell and Jones (1994). Instead of optimizing $|\mathbf{X}'\mathbf{X}|$, we optimized $|\mathbf{X}'\mathbf{X}+\mathbf{P}|$, where $\mathbf{P}$ is a diagonal matrix of prior precisions. This is analogous to ridge regression, in which a diagonal matrix is added to a rank-deficient $\mathbf{X}'\mathbf{X}$ to create a full-rank problem. We specified a model with a blocking variable, main effects for the six factors, block-effect interactions for the six factors, and all two-way interactions. We constructed $\mathbf{P}$ to contain zeros for the blocking variable, main effects, and block-effect interactions, and 45s (the number of runs divided by 2) for the two-way interactions. Then we used the modified Fedorov algorithm to search for good designs.

With an appropriate coding for $\mathbf{X}$, the value of the prior precision for a parameter roughly reflects the number of runs worth of prior information available for that parameter. The larger the prior precision for a parameter, the less information about that parameter is in the final design. Specifying a nonzero prior precision for a parameter reduces the contribution of that parameter to the overall efficiency. For this problem, we wanted maximal efficiency for the within-subject main-effects models, so we gave a nonzero prior precision to the aggregated two-way interactions.

Our best design had a *D*-efficiency for the second-order model of 63.9281 (with a *D*-efficiency for the aggregate main-effects model of 99.4338) and *D*-efficiencies for the main-effects models within each block of 100.0000, 100.0000, 100.0000, 99.0981, and 98.0854. The design is completely balanced within all blocks. We could have specified other values in $\mathbf{P}$ and gotten better efficiency for the aggregate design but less efficiency for the blocks. Choice of $\mathbf{P}$ depends in part on the primary goals of the experiment. It may require some simulation work to determine a good choice of $\mathbf{P}$.

All of the examples in this article so far have been straight-forward applications of computerized design methodology. A set of factors, levels, and estimable effects was specified, and the computer looked for an efficient design for that specification. Simple problems, such as those discussed previously, require only a few minutes of computer time. This problem was much more difficult, so we let a work station generate designs for about 72 hours. (We could have found less efficient but still acceptable designs in much less time.) We were asking the computer to find a good design out of over 9.6×10^{116} possibilities. This is like looking for a needle in a haystack, when the haystack is the size of the entire known universe. With such problems, we may do better if we use our intuition to give the computer "hints," forcing certain structure into the design. To illustrate, we tried this problem again, this time using a different approach.

We used the modified Fedorov algorithm to generate main-effects only 3^6 designs in 18 runs. We stopped when we had ten designs all with 100% efficiency. We then wrote an ad hoc program that randomly selected five of the ten designs, randomly permuted columns within each block, and randomly permuted levels within each block. These operations do not affect the first-order efficiencies but do affect the overall efficiency for the aggregate design. When an operation increased efficiency, the new design was kept. We iterated over the entire design 20 times. We let the program run for about 16 hours, which generated 98 designs, and we found our best design in three hours. Our best design had a *D*-efficiency for the second-order model of 68.0565 (versus 63.9281 previously), and all first-order efficiencies of 100.

Many other variations on this approach could be tried. For example, columns and blocks could be chosen at random, instead of systematically. We performed excursions of up to eight permutations before we reverted to the previous design. This number could be varied. It seemed that permuting the levels helped more than permuting the columns, though this was not thoroughly investigated. Whatever is done, it is important to consider efficiency. For example, just randomly permuting levels can create very inefficient designs.

For this particular problem, the ad hoc algorithm generated better designs than the Bayesian method, and it required less computer time. In fact, 91 out of the 98 ad hoc designs were better than the best Bayesian design. However, the ad hoc method required much more programmer time. It is possible to manually create a design for this situation, but it would be extremely difficult and time consuming to find an efficient design without a computerized algorithm for all but the most sophisticated of human designers. The best designs were found when used both our human design skills and a computerized search. We have frequently found this to be the case.

Conclusions

Computer-generated experimental designs can provide both better and more general designs for discrete-choice and preference-based conjoint studies. Classical designs, obtained from books or computerized tables, can be good options when they exist, but they are not the only option. The time-consuming and potentially error-prone process of finding and manually modifying an existing design can be avoided. When the design is nonstandard and there are restrictions, a computer can generate a design, and it can be done quickly. In most situations, a good design can be generated in a few minutes or hours, though for certain difficult problems more time may be necessary. Furthermore, when the circumstances of the project change, a new design can again be generated quickly.

We do not argue that computerized searches for *D*-efficient designs are *uniformly* superior to manually generated designs. The human designer, using intuition, experience, and heuristics, can recognize structure that an optimization algorithm cannot. On the other hand, the computerized search usually does a good job, it is easy to use, and it can create a design faster than manual methods, especially for the nonexpert. Computerized search methods and the use of efficiency criteria can benefit expert designers as well. For example, the expert can manually generate a design and then use the computer to evaluate and perhaps improve its efficiency.

In nonstandard situations, simultaneous balance and orthogonality may be unobtainable. Often, the best that can be hoped for is optimal efficiency. Computerized algorithms help by searching for the most efficient designs from a potentially very large set of possible designs. Computerized search algorithms for *D*-efficient designs do not supplant traditional design-creation skills. Rather, they provide helpful tools for finding good, efficient experimental designs.

Table A1
Chakravarti's L_{18}, Factor Levels

X1	X2	X3	X4	X5
-1	-1	-1	-1	-1
-1	-1	0	0	1
-1	-1	1	1	0
-1	1	-1	1	0
-1	1	0	-1	-1
-1	1	1	0	1
1	-1	-1	0	0
1	-1	-1	1	1
1	-1	0	-1	0
1	-1	0	1	-1
1	-1	1	-1	1
1	-1	1	0	-1
1	1	-1	-1	1
1	1	-1	0	-1
1	1	0	0	0
1	1	0	1	1
1	1	1	-1	0
1	1	1	1	-1

Table A2
Chakravarti's L_{18}, Orthogonal Coding

X1	X2	X3	-	X4	-	X5	-
-1	-1	-1.225	-0.707	-1.225	-0.707	-1.225	-0.707
-1	-1	0.000	1.414	0.000	1.414	1.225	-0.707
-1	-1	1.225	-0.707	1.225	-0.707	0.000	1.414
-1	1	-1.225	-0.707	1.225	-0.707	0.000	1.414
-1	1	0.000	1.414	-1.225	-0.707	-1.225	-0.707
-1	1	1.225	-0.707	0.000	1.414	1.225	-0.707
1	-1	-1.225	-0.707	0.000	1.414	0.000	1.414
1	-1	-1.225	-0.707	1.225	-0.707	1.225	-0.707
1	-1	0.000	1.414	-1.225	-0.707	0.000	1.414
1	-1	0.000	1.414	1.225	-0.707	-1.225	-0.707
1	-1	1.225	-0.707	-1.225	-0.707	1.225	-0.707
1	-1	1.225	-0.707	0.000	1.414	-1.225	-0.707
1	1	-1.225	-0.707	-1.225	-0.707	1.225	-0.707
1	1	-1.225	-0.707	0.000	1.414	-1.225	-0.707
1	1	0.000	1.414	0.000	1.414	0.000	1.414
1	1	0.000	1.414	1.225	-0.707	1.225	-0.707
1	1	1.225	-0.707	-1.225	-0.707	0.000	1.414
1	1	1.225	-0.707	1.225	-0.707	-1.225	-0.707

Table A3
Green & Wind
Orthogonal Design
Example

X1	X2	X3	X4	X5
-1	-1	-1	-1	-1
-1	-1	-1	1	0
-1	-1	0	-1	-1
-1	-1	0	0	1
-1	-1	0	1	-1
-1	-1	1	-1	0
-1	-1	1	0	1
-1	-1	1	1	0
-1	1	-1	1	1
-1	1	-1	-1	1
-1	1	0	0	0
-1	1	1	0	-1
1	-1	-1	0	-1
1	-1	-1	0	0
1	-1	0	1	1
1	-1	1	-1	1
1	1	1	1	-1
1	1	0	-1	0

Table A4
Information-Efficient
Design,
Factor Levels

X1	X2	X3	X4	X5
-1	-1	-1	0	-1
-1	-1	0	-1	0
-1	-1	0	1	-1
-1	-1	1	0	1
-1	-1	1	1	1
-1	1	-1	-1	0
-1	1	-1	0	-1
-1	1	0	-1	1
-1	1	1	1	0
1	-1	-1	-1	1
1	-1	-1	1	0
1	-1	0	0	0
1	-1	1	-1	-1
1	1	-1	1	1
1	1	0	0	1
1	1	0	1	-1
1	1	1	-1	-1
1	1	1	0	0

Table A5
Information-Efficient
Design, Unrealistic
Combinations Excluded

X1	X2	X3	X4	X5
-1	-1	-1	1	0
-1	-1	-1	-1	1
-1	-1	-1	0	-1
-1	-1	0	-1	1
-1	-1	0	0	0
-1	1	1	1	0
-1	1	1	-1	-1
-1	1	1	0	1
-1	1	0	1	-1
1	-1	1	1	-1
1	-1	1	-1	0
1	-1	1	0	1
1	-1	0	1	-1
1	1	-1	1	0
1	1	-1	-1	-1
1	1	-1	0	1
1	1	0	-1	1
1	1	0	0	0

Table A6
Consumer Food Product (Raw) Design

X1	X2	X3	X4	X5	X6	X7	X8
1	1	2	1	1	2	1	3
1	2	2	1	2	3	1	2
1	4	1	1	1	3	1	3
2	2	1	1	3	2	1	1
2	3	2	1	2	2	2	3
2	4	2	1	3	3	2	2
3	1	1	1	3	2	2	2
3	3	2	1	3	1	2	1
3	4	2	1	2	1	1	1
4	1	1	1	2	3	2	1
4	1	2	1	3	3	1	1
4	2	2	1	1	2	2	3
4	3	1	1	1	1	1	2
1	3	1	2	3	2	2	1
1	3	2	2	3	1	1	3
1	4	2	2	1	1	2	1
2	1	1	2	1	3	1	1
2	2	2	2	3	2	1	1
2	3	1	2	2	1	2	3
2	4	1	2	3	1	1	2
3	1	2	2	2	3	2	2
3	2	1	2	1	3	2	3
3	4	2	2	2	3	1	3
4	1	1	2	3	2	1	3
4	2	1	2	2	1	2	2
4	3	2	2	1	2	1	2

Table A7
Consumer Food Product Choice Set

Block 1: Shelf-Talker Absent For Client Line Extension

Choice Set	Client Brand	Client Line Extension	Regional Brand	Private Label	National Competitor
1	$1.29	$1.39/stove	$1.99	$2.29/micro	N/A
2	$1.29	$1.89/stove	$2.49	N/A	$2.39
3	$1.29	N/A	$1.99	N/A	N/A
4	$1.69	$1.89/micro	N/A	$2.29/micro	$1.99
5	$1.69	$2.39/stove	$2.49	$2.29/stove	N/A
6	$1.69	N/A	N/A	N/A	$2.39
7	$2.09	$1.39/micro	N/A	$2.29/stove	$2.39
8	$2.09	$2.39/stove	N/A	$1.49/stove	$1.99
9	$2.09	N/A	$2.49	$1.49/micro	$1.99
10	N/A	$1.39/micro	$2.49	N/A	$1.99
11	N/A	$1.39/stove	N/A	N/A	$1.99
12	N/A	$1.89/stove	$1.99	$2.29/stove	N/A
13	N/A	$2.39/micro	$1.99	$1.49/micro	$2.39

Block 2: Shelf-Talker Present For Client Line Extension

Choice Set	Client Brand	Client Line Extension	Regional Brand	Private Label	National Competitor
14	$1.29	$2.39/micro	N/A	$2.29/stove	$1.99
15	$1.29	$2.39/stove	N/A	$1.49/micro	N/A
16	$1.29	N/A	$1.99	$1.49/stove	$1.99
17	$1.69	$1.39/micro	$1.99	N/A	$1.99
18	$1.69	$1.89/stove	N/A	$2.29/micro	$1.99
19	$1.69	$2.39/micro	$2.49	$1.49/stove	N/A
20	$1.69	N/A	N/A	$1.49/micro	$2.39
21	$2.09	$1.39/stove	$2.49	N/A	$2.39
22	$2.09	$1.89/micro	$1.99	N/A	N/A
23	$2.09	N/A	$2.49	N/A	N/A
24	N/A	$1.39/micro	N/A	$2.29/micro	N/A
25	N/A	$1.89/micro	$2.49	$1.49/stove	$2.39
26	N/A	$2.39/stove	$1.99	$2.29/micro	$2.39

Table A8
Consumer Food Product Design Level Frequencies

Level	X1	X2	X3	X4	X5	X6	X7	X8
1	6	7	12	13	8	8	14	9
2	7	6	14	13	8	9	12	8
3	6	7			10	9		9
4	7	6						

Table A9
Consumer Food Product Design Creation Code

```
*-------------------------------------------------------*
|                Construct the Design.                  |
*-------------------------------------------------------;

%macro bad;
  bad = (x1 < 4) + (x2 < 4) + (x5 < 3) + (x6 < 3) + (x8 < 3);
  bad = abs(bad - 3) * ((bad < 2) | (bad > 4));
  %mend;

%mktex( 4 4 2 2 3 3 2 3, n=26, interact=x2*x3 x2*x4 x3*x4 x6*x7,
        restrictions=bad, outr=sasuser.choicdes )

*-------------------------------------------------------*
|                  Print the Design.                    |
*-------------------------------------------------------;

proc format;
   value yn    1 = 'No'    2 = 'Talker';
   value micro 1 = 'Micro' 2 = 'Stove';
   run;

data key;
   missing N;
   input x1-x8;
   format x1 x2 x5 x6 x8 dollar5.2
          x4 yn. x3 x7 micro.;
   label x1 = 'Client Brand'
         x2 = 'Client Line Extension'
         x3 = 'Client Micro/Stove'
         x4 = 'Shelf Talker'
         x5 = 'Regional Brand'
         x6 = 'Private Label'
         x7 = 'Private Micro/Stove'
         x8 = 'National Competitor';
   datalines;
1.29 1.39 1 1 1.99 1.49 1 1.99
1.69 1.89 2 2 2.49 2.29 2 2.39
2.09 2.39 . . N    N    .    N
N    N    . . .    .    .    .
;

%mktlab(data=sasuser.choicdes, key=key)

proc sort out=sasuser.finchdes; by x4; run;

proc print label; id x4; by x4; run;
```

A General Method for Constructing Efficient Choice Designs

Klaus Zwerina

Joel Huber

Warren F. Kuhfeld

Abstract

Researchers have traditionally built choice designs using extensions of concepts from the general linear design literature. We show that a computerized search strategy can generate efficient choice designs with standard personal computers. This approach holds three important advantages over previous design strategies. First, it allows the incorporation of anticipated model parameters, thereby increasing design efficiency and considerably reducing the number of required choices. Second, complex choice designs can be easily generated, allowing researchers to conduct choice experiments that more closely mirror actual market conditions. Finally, researchers can explore model and design modifications and examine trade-offs between a design's statistical benefits and its operational and behavioral costs.*

Introduction

Discrete choice experiments are becoming increasingly popular in marketing, economics, and transportation. These experiments enable researchers to model choice in an explicit competitive context, thus realistically emulating market decisions. A choice design consists of choice sets composed of several alternatives, each defined as combinations of different attribute levels. A good choice design is efficient, meaning that the parameters of the choice model are estimated with maximum precision.

A number of methods have been suggested for building choice designs (Anderson and Wiley 1992, Bunch, Louviere, and Anderson 1996, Krieger and Green 1991, Kuhfeld 2003 (page 81), Lazari and Anderson 1994, Louviere and Woodworth 1983). Most of the methods use extensions of standard linear experimental designs (Addelman 1962b, Green 1974). However, the use of linear designs in choice experiments may be nonoptimal due to two well-known differences between linear and choice

*Klaus Zwerina is a consultant at BASF AG, Ludwigshafen, Germany. Joel Huber is Professor of Marketing, Fuqua School of Business, Duke University. Warren F. Kuhfeld is Manager, Multivariate Models R&D, Statistical Research and Development, SAS Institute Inc. We would like to thank Jim Bettman and Richard Johnson for their helpful comments on an earlier version of this chapter. Copies of this chapter (TS-689D) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market .

models. First, probabilistic choice models are nonlinear in the parameters, implying that the statistical efficiency of a choice design depends on an (unknown) parameter vector. This property implies the need to bring anticipated parameter values in choice designs. Second, choice design efficiency depends both on the creation of appropriate profiles and properly placing them into several choice sets. For example, in a linear design, the order of the 16 profiles in a conjoint exercise does not affect its formal efficiency, whereas the efficiency of the same 16 profiles broken into four choice sets depends critically on the grouping. Despite its limitations, linear design theory has been used to produce satisfactory choice designs for many years, drawing on readily available tables and processes. Such carefully selected linear designs are reasonable, general-purpose choice designs, but are generally not optimal in a statistical sense.

We present a general strategy for the computerized construction of efficient choice designs. This contribution can be viewed as an extension of the work of Kuhfeld, Tobias, and Garratt (1994) and of Huber and Zwerina (1996). Kuhfeld et al. recommended using a search algorithm to find efficient *linear* designs. Huber and Zwerina show how to modify *choice* designs using anticipated model parameters in order to improve design efficiency. We adapt the optimization procedure outlined in Kuhfeld et al. to the principles of choice design efficiency described by Huber and Zwerina. Our approach holds several important advantages over previous choice design strategies. It (1) optimizes the "correct" criterion of minimizing estimation error rather than following linear design principles, (2) it can generate choice designs that accommodate any anticipated parameter vector, (3) it can accommodate virtually any level of model complexity, and finally (4) it can be built using widely available software. To illustrate, we include a SAS/IML program that generates relatively simple choice designs. This program can be easily generalized to handle far more complex problems.

The chapter begins with a review of criteria for efficient choice designs and illustrates how they can be built with a computer. Then, beginning with simple designs, we illustrate how the algorithm works and how our linear design intuition must be changed when coping with choice designs. Next, we generate more complex choice designs and show how to evaluate the impact on efficiency of different design and model modifications. We conclude with a discussion of the proposed choice design approach and directions for future research.

Criteria For Choice Design Efficiency

Measure Of Choice Design Efficiency. First, we derive a measure of efficiency in choice designs from the well-known multinomial logit model (McFadden 1974). This model assumes that consumers make choices among alternatives that maximize their perceived utility, u, given by

$$u = \mathbf{x}_i\boldsymbol{\beta} + e \tag{1}$$

where $\mathbf{x}_i$ is a row vector of attributes characterizing alternative i, $\boldsymbol{\beta}$ is a column vector of K weights associated with these attributes, and e is an error term that captures unobserved variations in utility. Suppose that there are N choice sets, C_n, indexed by $n = 1, 2, \ldots, N$, where each choice set is characterized by a set of alternatives $C_n = \{x_{1n}, K, x_{J_n n}\}$. If the errors, e, are independently and identically Gumbel distributed, then it can be shown that the probability of choosing an alternative i from a choice set C_n is

$$P_{in}(\mathbf{X_n}, \boldsymbol{\beta}) = \frac{e^{\mathbf{x}_{in}\boldsymbol{\beta}}}{\sum_{j=1}^{J_n} e^{\mathbf{x}_{jn}\boldsymbol{\beta}}} \tag{2}$$

where $\mathbf{X_n}$ is a matrix that consists of J_n row vectors, each describing the characteristics of the alternatives, $\mathbf{x}_{jn}$. The vertical concatenation of the $\mathbf{X_n}$ matrices is called a choice design matrix $\mathbf{X}$.

The task of the analyst is to find a parameter estimate for $\boldsymbol{\beta}$ in Equation (2) that maximizes the likelihood given the data. Under very general conditions, the maximum likelihood estimator is consistent and asymptotically normal with covariance matrix

$$\boldsymbol{\Sigma} = (\mathbf{Z}'\mathbf{P}\mathbf{Z})^{-1} = \left[\sum_{n=1}^{N}\sum_{j=1}^{J_N} \mathbf{z}'_{jn} P_{jn} \mathbf{z}_{jn}\right]^{-1} \tag{3}$$

$$\text{where} \quad \mathbf{z}_{jn} = \mathbf{x}_{jn} - \sum_{i=1}^{J_n} \mathbf{x}_{in} P_{in} \; .$$

Equation (3) reveals some important properties of (nonlinear) choice models. In linear models, centering occurs across all profiles whereas in choice models, centering occurs within choice sets. This shows that in choice designs both the profile selection and the assignment of profiles to choice sets affects the covariance matrix. Moreover, in linear models, the covariance matrix does not depend on the true parameter vector, whereas in choice models the probabilities, P_{jn}, are functions of $\boldsymbol{\beta}$ and hence the covariance matrix. Assuming $\boldsymbol{\beta} = \mathbf{0}$ simplifies the design problem, however Huber and Zwerina (1996) recently demonstrated that this assumption may be costly. They showed that incorrectly assuming that $\boldsymbol{\beta} = \mathbf{0}$ may require from 10% to 50% more respondents than those built from reasonably anticipated parameters.

The goal in choice designs is to define a group of choice sets, given the anticipated $\boldsymbol{\beta}$, that minimizes the "size" of the covariance matrix, $\boldsymbol{\Sigma}$, defined in Equation (3). There are various summary measures of error size that can be derived from the covariance matrix (see, e.g., Raktoe, Hedayat, and Federer 1981). Perhaps the most intuitive summary measure is the average variance around the estimated parameters of a model. This measure is referred to in the literature as *A*-efficiency or its inversely related counterpart,

$$A - \mathit{error} = \text{trace}\,(\boldsymbol{\Sigma})/K \tag{4}$$

where K is the number of parameters. Two problems with this measure limit its suitability as an overall measure of design efficiency. First, relative *A*-error is not invariant under (nonsingular) recodings of the design matrix, i.e., design efficiency depends on the type of coding. Second, it is computationally expensive to update. A related measure,

$$D - \mathit{error} = |\boldsymbol{\Sigma}|^{1/K} \tag{5}$$

is based on the determinant as opposed to the trace of the covariance matrix. *D*-error is computationally efficient to update, and the ratios of *D*-errors are invariant under different codings of the design matrix. Since *A*-error is the arithmetic mean and *D*-error is the geometric mean of the eigenvalues of $\boldsymbol{\Sigma}$, they

are generally highly correlated. *D*-error thereby provides a reasonable way to find designs that are "good" on alternative criteria. For example, if *A*-error is the ultimate criterion, we can first minimize *D*-error and then select the design with minimum *A*-error rather than minimizing *A*-error directly. For these reasons, *D*-error (or its inverse, *D*-efficiency or *D*-optimality) is the most common criterion for evaluating linear designs and we advocate it as a criterion for choice designs.

Next, we discuss four principles of choice design efficiency defined by Huber and Zwerina (1996). Choice designs that satisfy these principles are optimal, however, these principles are only satisfied for a few special cases and under quite restrictive assumptions. The principles of orthogonality and balance that figured so prominently in linear designs remain important in understanding what makes a good choice design, but they will be less useful in generating one.

Four Principles Of Efficient Choice Designs. Huber and Zwerina (1996) identify four principles which when jointly satisfied indicate that a design has minimal *D*-error. These principles are orthogonality, level balance, minimal overlap, and utility balance. *Orthogonality* is satisfied when the levels of each attribute vary independently of one another. *Level balance* is satisfied when the levels of each attribute appear with equal frequency. *Minimal overlap* is satisfied when the alternatives within each choice set have nonoverlapping attribute levels. *Utility balance* is satisfied when the utilities of alternatives within choice sets are the same, i.e., the design will be more efficient as the expected probabilities within a choice set C_n among J_n alternatives approach $1/J_n$.

These principles are useful in understanding what makes a choice design efficient, and improving any principle, holding the others constant, improves efficiency. However, for most combinations of attributes, levels, alternatives, and assumed parameter vectors, it is impossible to create a design that satisfies these principles. The proposed approach does not build choice designs from these formal principles, but instead uses a computer to directly minimize *D*-error. As a result, these principles may only be approximately satisfied in our designs, but they will generally be more efficient than those built directly from the principles.

A General Method For Efficient Choice Designs

Figure 1 provides a flowchart of the proposed design approach. The critical aspect of this approach involves an adaptation of Cook and Nachtsheim's (1980) modification of the Fedorov (1972) algorithm that has successfully been used to generate efficient *linear* designs (e.g., Cook and Nachtsheim 1980, Kuhfeld et al. 1994). We will first describe the proposed choice design approach conceptually and then define the details in a context of a particular search.

The process begins by building a *candidate set*, which is a list of potential alternatives. A random selection of these alternatives is the *starting design*. The algorithm alters the starting design by exchanging its alternatives with the candidate alternatives. The algorithm finds the best exchange (if one exists) for the first alternative in the starting design. The first iteration is completed once the algorithm has sequentially found the best exchanges for all of the alternatives in the starting design. After that, the process moves back to the first alternative and continues until no substantial efficiency improvement is possible. To avoid poor local optima, the whole process can be restarted with different random starting designs and the most efficient design is selected. For example, if there are 300 alternatives in the candidate set and 50 alternatives in the choice design, then each iteration requires testing 15,000 possible exchanges, which is a reasonable problem on today's desktop computers and workstations. While there is no guarantee that it will converge to an optimal design, our experience

with relatively small problems suggests that the algorithm works very well.

To illustrate the process we first generate choice designs for simple models that reveal the characteristics of efficient choice designs. In examining these simple designs, our focus is on the benefits and the insights that derive from using this approach. Then, we apply the approach to more complex design problems, such as alternative-specific designs and designs with constant alternatives. As we illustrate more complex designs, we will focus on the use of the approach, *per se*. We provide illustrative computer code in the appendix.

Choice Design Applications

Simple Designs

Generic Models. The simplest choice models involve alternatives described by generic attributes. The utility functions for these models consist of attribute parameters that are the same for all alternatives, for example, a common price slope across all alternatives. Generic designs are appealing because they are simple and analogous to main-effects conjoint experiments. Bunch et al. (1996) evaluate ways to generate generic choice designs and show that shifted or cyclic designs generally have superior efficiency compared with other strategies for generating main effects designs. These shifted designs use an orthogonal fractional factorial to provide the "seed" alternatives for each choice set. Subsequent alternatives within a choice set are cyclically generated. The attribute levels of the new alternatives add one to the level of the previous alternative until it is at its highest level, at which point the assignment re-cycles to the lowest level.

For certain families of plans and assuming that all coefficients are zero, these shifted designs satisfy all four principles, and thus are optimal.[†] For example, consider a choice experiment with three attributes, each at three levels, defining three alternatives in each of nine choice sets. The left-hand panel of Table 1 shows a plan using the Bunch et al. (1996) method.

In this special case, all four efficiency principles are perfectly satisfied. Level balance is satisfied since each level occurs in precisely 1/3 of the cases, and orthogonality can be confirmed by noting the all pairs of attribute levels occur in precisely 1/9 of the attributes (Addelman 1962b). There is perfect minimal overlap since each level occurs exactly once in each choice set, and finally, utility balance is trivially satisfied with the assumption that $\boldsymbol{\beta} = \mathbf{0}$. More formally, it is useful to examine the covariance matrix of the (effects-coded) parameters, reported in the first panel of Table 2. The equal variances across attributes and the zero covariances across attributes both indicate optimality.

A simple design such as this could have been built from our algorithm, although using a standard orthogonal array and cyclic permutations ensured optimality. Our next example, encompassing a model with just one interaction term, illustrates the case when a computerized search is very useful in finding a statistically efficient design.

[†]We were not able to analytically prove this, but after examining scores of designs, we have never found more efficient designs than those that satisfy all four principles.

Figure 1
Flowchart of Algorithm for Constructing Efficient Choice Designs

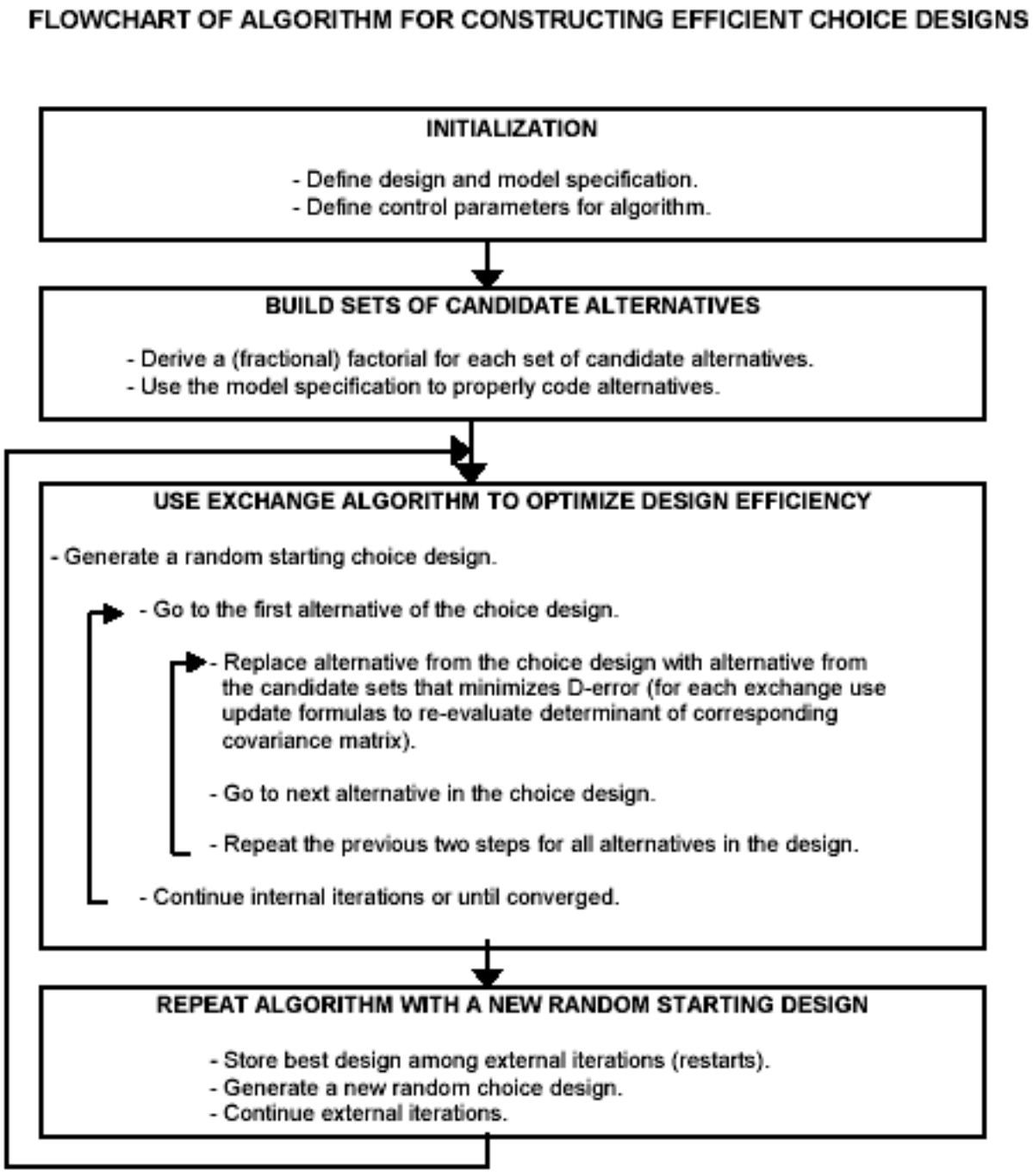


Table 1
Main Effects and A×B-Interaction Effects Choice Design

		$\boldsymbol{\beta}_0$-Efficient Main-Effects Design ($\boldsymbol{\beta}_0$=0 0 0 0 0 0)			$\boldsymbol{\beta}_0$-Efficient Interaction-Effects ($\boldsymbol{\beta}_0$=0 0 0 0 0 0 0 0 0)				$\boldsymbol{\beta}_1$-Efficient Interaction-Effects ($\boldsymbol{\beta}_1$=-1 0 -1 0 -1 0 0 0 0)			
Set	Alt	A	B	C	A	B	C	p($\boldsymbol{\beta}_1$)	A	B	C	p($\boldsymbol{\beta}_1$)
1	I	1	1	1	2	3	2	.495	2	1	3	.422
	II	2	2	2	2	2	3	.495	1	3	2	.422
	III	3	3	3	1	1	1	.009	3	1	1	.155
2	I	1	2	2	3	1	1	.155	3	2	2	.422
	II	2	3	3	2	2	2	.422	2	1	3	.155
	III	3	1	1	1	2	3	.422	3	3	1	.422
3	I	1	3	3	1	1	2	.042	2	2	3	.155
	II	2	1	1	1	3	1	.114	3	3	2	.422
	III	3	2	2	3	1	3	.844	2	3	3	.422
4	I	2	1	3	2	1	2	.018	2	1	2	.422
	II	3	2	1	3	3	3	.965	1	1	3	.422
	III	1	3	2	2	2	1	.018	1	2	1	.155
5	I	2	2	1	1	3	3	.245	3	1	2	.422
	II	3	3	2	3	3	2	.665	1	1	3	.155
	III	1	1	3	2	3	1	.090	3	2	1	.422
6	I	2	3	2	2	1	3	.468	1	3	3	.422
	II	3	1	3	1	2	1	.063	2	3	2	.422
	III	1	2	1	1	3	2	.468	3	2	1	.155
7	I	3	1	2	3	2	3	.665	1	2	1	.212
	II	1	2	3	3	3	1	.245	2	2	1	.576
	III	2	3	1	3	1	2	.090	1	1	2	.212
8	I	3	2	3	1	2	2	.042	1	2	3	.576
	II	1	3	1	2	3	3	.844	1	3	1	.212
	III	2	1	2	3	2	1	.114	3	1	1	.212
9	I	3	3	1	1	1	3	.114	2	2	2	.212
	II	1	1	2	2	1	1	.042	3	1	3	.576
	III	2	2	3	3	2	2	.844	2	3	1	.212
								avemaxp = .690				avemaxp = .474
		D-error($\boldsymbol{\beta}_0$) = .192						D-error($\boldsymbol{\beta}_0$) = .306				D-error($\boldsymbol{\beta}_0$) = .365
								D-error($\boldsymbol{\beta}_1$) = .630				D-error($\boldsymbol{\beta}_1$) = .399

Estimating an A×B Interaction. For the previous example with nine choice sets, let us assume that the researcher is confident that there are no A×C or B×C interactions, but the A×B interaction must be estimated. The middle panel of Table 1 shows the best design we were able to find which includes this one interaction. Note that in this design, the principle of minimal overlap on attributes A and B is violated, in that attribute levels are frequently repeated within a set. In general, interactions *require* overlap of attribute levels to produce the contrasts necessary to estimate the effects.

The covariance matrix of this design, depicted in the lower half of Table 2, highlights the effects of incorporating the A×B interaction. Violating minimal overlap permits the estimation of the A×B

Table 2
Covariance Matrix of Main Effects and A×B-Interaction
Effects Choice Design

$\boldsymbol{\beta}_0$-Efficient Main Effects Design

	a1	a2	b1	b2	c1	c2
a1	.222	-.111	.000	.000	.000	.000
a2	-.111	.222	.000	.000	.000	.000
b1	.000	.000	.222	-.111	.000	.000
b2	.000	.000	-.111	.222	.000	.000
c1	.000	.000	.000	.000	.222	-.111
c1	.000	.000	.000	.000	-.111	.222

D-error($\boldsymbol{\beta}_0$)=.192

$\boldsymbol{\beta}_0$-Efficient Interaction Effects Design

	a1	a2	b1	b2	c1	c2	ab11	ab12	ab21	ab22
a1	.296	-.130	.019	-.019	.000	.000	-.037	.000	.000	-.019
a2	-.130	.296	-.019	.019	.000	.000	.037	.000	.000	.019
b1	.019	-.019	.296	-.130	.000	.000	.019	-.056	.000	.037
b2	-.019	.019	-.130	.296	.000	.000	-.019	.056	.000	-.037
c1	.000	.000	.000	.000	.222	-.111	.000	.000	.000	.000
c2	.000	.000	.000	.000	-.111	.222	.000	.000	.000	.000
ab11	-.037	.037	.019	-.019	.000	.000	.630	-.333	-.333	.148
ab12	.000	.000	-.056	.056	.000	.000	-.333	.556	.167	-.278
ab21	.000	.000	.000	.000	.000	.000	-.333	.167	.667	-.333
ab22	-.019	.019	.037	-.037	.000	.000	.148	-.278	-.333	.630

D-error($\boldsymbol{\beta}_0$) of main effects = .239
D-error($\boldsymbol{\beta}_0$) of all effects = .306

interaction by sacrificing efficiency on the main effects of attribute A and B, reflected in higher variances of the main effects estimates (a1, a2, b1, and b2). The *D*-error of the main effect estimates increases by 24%, from .192 to .239, and the covariances across attributes A and B are no longer zero. Note also that the errors around attribute C are unchanged − they are unaffected by the A×B interaction, indicating that the algorithm was able to find a design that allowed the A×B interaction to be uncorrelated with C.

There are several important lessons from this simple example. First, it illustrates that a design that is "perfect" for one model may be far from optimal for a slightly different model. Adding one interaction strongly altered the covariance matrix, so efficient designs generally violate the formal principles. Second, the example shows that estimating new interactions is not without cost; being able to estimate one interaction increased by 24% the error on the main effects. Finally, the trade-off of efficiency with estimability demonstrates one of the primary benefits of this approach−it allows the analyst to understand the efficiency implications of changes in the design structure and/or model specification. This use of the approach will be illustrated again in the context of more complex choice designs.

The Impact Of Non-Zero Betas. The preceding discussion has assumed that the true parameters are zero. This assumption is justified when there is very little information about the model parameters; however, typically the analyst has some information on the relative importance of attributes or the relative value of their levels (Huber and Zwerina 1996). To show the potential gain that can come from nonzero parameters, assume that the anticipated partworths of the main effects for the three level attributes discussed previously are not 0, 0, 0, but -1, 0, 1, while the A×B-interaction effect continues to have zero parameters.[‡] Calling the new parameter vector $\boldsymbol{\beta}_1$ to distinguish it from the zero parameter vector, $\boldsymbol{\beta}_0$, the third panel of Table 1 displays the efficient design using these parameters. This new design has a D-error($\boldsymbol{\beta}_1$) of 0.399. However, if instead we had used the design in the center panel, its error given $\boldsymbol{\beta}_1$ is true would have been .630, implying that 37% (1 - .399/.630) fewer respondents are needed for the "utility balanced" over the "utility neutral" design.

Comparing the last two panels in Table 1 reveals how the algorithm used the anticipated nonzero parameters to produce a more efficient design. As an index of utility balance, we calculated the average of the maximum within-choice-set choice probabilities (avemaxp). The smaller this index the harder is the average choice task and the greater is "utility balance." We can see, by using $\boldsymbol{\beta}_1$, the new design is more utility balanced than the previous design, which results in an average maximum probability of .474 compared with one of .690. We also see that the increase in utility balance sacrifices somewhat the three formal principles, reflected in an increase of D-error($\boldsymbol{\beta}_0$) from .306 to .365. The new design does not have perfect orthogonality, level balance, utility balance, or minimal overlap, but it is more efficient than any design that is perfect on any of those criteria.

More Complex Choice Designs. The proposed algorithm is very general and can be applied to virtually any level of design complexity. We will use it next to generate an alternative-specific choice design, which has a separate set of parameters for each alternative. Suppose, the researcher is interested in simulating the market behavior of three brands, Coke, Pepsi, and RC Cola, with the attribute combinations shown in Table 3.

This kind of choice experiment, which we call a market emulation study, is quite different from the generic choice design presented previously. In a market emulation study, emphasis is on predicting the impact of brand, flavor, and container decisions in the context of a realistic market place offering. What this kind of study gains in realism, it loses in the interpretability of its results. For example, since each brand only occurs at specific prices, it is much harder to disentangle the independent effects of brand and price. These designs are, however, useful in assessing the managerially critical question of the impact of, say, a 60 cent drop in the price of Coke's 16 ounce case in a realistic competitive configuration.

Since we assume that the impact of price depends on the brand to which it is attached, it is important that the impact of price be estimable within each brand.[§] Further, let us assume that the reaction to price additionally depends on the number of ounces, so that it is necessary to estimate the brand×price×container interaction. Using standard ANOVA-coding, these assumptions require four main effects (brand, flavor, container, and price for 8 *df*), four two-way interactions (brand×price, brand×flavor, brand×container, and price×container for 16 *df*), and one three-way interaction (brand×price×container for 8 *df*), resulting in a total of 32 parameters.[¶]

[‡]We assume for simplicity that the interaction has parameter values of zero. Note, this also produces minimal variance of estimates around zero, implying greatest power of a test in the region of that null hypothesis.

[§]The assumption that price has a different impact depending on the brand is testable. The ability to make that test is just one of the advantages of these choice designs.

[¶]We need the fourth two-way interaction, price×container, to be able to estimate the three-way interaction brand×price×container. Of course, there are many other ways of coding a design.

Table 3
Attributes/Levels for an Alternative-Specific Choice Experiment

	Alternative-Specific Levels		
Attributes	Coke	Pepsi	RC Cola
Price per case	$5.69	$5.39	$4.49
	$6.89	$5.99	$5.39
	$7.49	$6.59	$5.99
Container	12 oz cans	12 oz cans	12 oz cans
	10 oz bottle	10 oz bottle	16 oz bottle
	16 oz bottle	18 oz bottle	22 oz bottle
Flavor	Regular	Regular	Regular
	Cherry Coke	Pepsi Lite	Cherry
	Diet Coke	Diet Pepsi	Diet

Suppose we want to precisely estimate these effects with a choice design consisting of 27 choice sets each composed of three alternatives.* The candidate set of alternatives comprises the $3^4 = 81$ possible alternatives, and the initial design is a random selection from these. The algorithm exchanges alternatives between the candidate set and the starting design until the efficiency gain becomes negligible. In the example with 27 choice sets and 32 parameters, *D*-error is .167. This statistic provides a baseline for evaluating other related designs, which we will generate in the following section.

Evaluating Design Modifications. The proposed approach can be used to evaluate design modifications. Typically, efficiency is meaningful within a relatively narrow family of designs, limited to a particular attribute structure, model specification, and number of alternatives per choice set. For many applications, optimizing a design within such a narrow design family is too restrictive. Most analysts are not tightly bound to a particular number of alternatives per choice set or even particular attributes, but are interested in exploring the impact of changes in these specifications on the precision of the parameter estimates. We will demonstrate how comparing designs across design families allows a reasoned trade-off of design structure against estimation precision.

Consider the following questions an analyst might ask concerning the alternative-specific choice design just presented.

1. How much does efficiency increase if 54 choice sets are used instead of two replications of 27 choice sets?
2. What is the efficiency loss if each of the brands (Coke, Pepsi, RC) must be present in a choice set?
3. What is the gain in efficiency if a fourth alternative is added to each choice set?
4. What happens to efficiency if this fourth alternative is constant (e.g., "keep on shopping")?

*The appendix contains a SAS/IML program that performs the search for this design. Focusing on the principles of the algorithm, the program was deliberately kept simple, specific, and small. A general macro for searching for choice designs, `%ChoicEff`, is documented in Kuhfeld (2003) starting on pages 479 and 481. See page 303 for an example.

Table 4
Impact of Design Modifications on *D*-Error

Design Modification	*D*-error	Efficiency per Choice Set	Comments
27 sets, 3 alternatives per set.	.167	100%	Original design.
Double the number of sets.	.079	106%	Limited benefit from doubling the number of sets.
Require each alternative to contain one of each brand.	.175	95%	Shows minor cost of constraining a design.
Add a fourth alternative.	.144	116%	Diminishing returns from adding additional alternatives.
Fourth alternative is constant.	.195	86%	Design is less efficient because constant alternative is chosen 25% of the time.

The first question assesses the benefit of building a design with 54 choice sets rather than using the original 27 choice sets twice. As Table 4 shows, specifying twice as many choice sets produces a *D*-error of .079 compared with .084 (=.167/2) for two independent runs of the 27 choice set design. This relatively small 6% benefit in efficiency indicates that the original 27 choice set design, while highly fractionated, appears to have suffered little due to this fact.

The second question evaluates the impact of constraints on the choice sets that respondents face. The original design often paired the same brand against itself within a choice set. For example, a choice set with Coke in a 12 oz bottle for $5.69 per case might include Coke in a 16 oz bottle for $7.49 per case. For managerial reasons it might be desirable to have each brand (Coke, Pepsi, RC) represented in every set of three alternatives. To examine the cost of this constraint, Coke is assigned to the first alternative, Pepsi to the second alternative, and RC to the third alternative within each of the 27 choice sets. With this constraint, the *D*-error is .175. This relatively moderate decrease in efficiency of 5% should be acceptable if there are managerially-based reasons to constrain the choice sets.

The third question investigates the benefits of adding a fourth alternative to each choice set. This change increases by 25% the number of alternatives, although the marginal effect of an additional alternative should not be as great. With this modification, *D*-error becomes .144, producing a 16% efficiency gain over three alternatives per choice set. The decision whether to include a fourth alternative now pits the analyst's appraisal of the trade-off between the value of this 16% efficiency gain and the cost in respondent time and reliability.

What happens if this fourth alternative is common and constrained to be constant in all choice sets? With a constant alternative, respondents are not forced to make a choice among undesirable alternatives. Moreover, a constant alternative permits an estimate of demand volume rather than just market shares (Carson et al. 1994). A constant alternative can take many forms, ranging from the respondent's "current brand," to an indication that "none are acceptable," or simply "keep on shopping." While constant alternatives are often added to choice sets, little is known about the efficiency implications

of this practice. To create designs with a constant alternative, this alternative must be added to the candidate set. Also, a model with a constant alternative has one more parameter. Comparing a design with a constant alternative to one without, it is necessary to calculate *D*-error with respect to the original 32 parameters using the corresponding submatrix of $\mathbf{\Sigma}$.

Adding a constant alternative to the original design increases the *D*-error of the original 32 parameters by 17% and is nearly 35% worse than allowing the fourth alternative to be variable. Some part of this loss in efficiency is due to the one additional degree of freedom from the constant alternative. A larger part is due to the efficiency lost when respondents are assumed to select the constant alternative. Every time it is chosen, one obtains less information about the values of the other parameters. In this case, the assumption that $\boldsymbol{\beta} = \mathbf{0}$ is not benign, as it assumes the constant alternative, along with all others in the four-option choice sets, will be chosen 25% of the time. We can reduce the efficiency cost to the other parameters by using a smaller $\boldsymbol{\beta}$ for the constant alternative, reflecting the assumption that it will be chosen less often.

In summary, the analysis suggests that adding a constant alternative to a three-alternative choice set can degrade the precision of estimates around the original parameters. Two caveats are important. First, this result will not always occur. We have found some highly fractionated designs where a constant alternative adds to the resolution of the original design. Second, there are studies where a major goal is the estimation of the constant alternative; in that case "oversampling" the constant ensures that its coefficient will be known with greater precision.

An important lesson across these four examples is that one cannot rely on heuristics to guide design strategies, ignoring statistical efficiency. It is generally necessary to test specific design strategies, given anticipated model parameters, to find a good choice design.

Evaluating Model Modifications. The proposed approach can be used to assess modifications of the model specification. This allows one, for example, to estimate the cost of "assumption insurance," i.e., building a design that is robust to false assumptions. Often we assume that factors are independent; for example, that the utility of price does not depend on brand or container. In many instances this assumption would be better termed a "presumption" in that if it is wrong, the estimates are biased, but there is no way to know given the design. Assessing the cost of assumption insurance involves four steps:

1. Find the best design for the unrestricted model (possibly including interactions).
2. Find the best design for the restricted model.
3. Evaluate *D*-error for that unrestricted design under the restricted model.
4. Evaluate *D*-error for the best design for the restricted model.

The cost of assumption insurance is the percent difference between steps 3 and 4, reflecting the loss of efficiency of the core parameters for the two designs. We illustrate how to assess this cost for a design that permits the price term to interact with brand and container versus one that assumes they are independent. To simplify the example, we take the same case as before, but assume that price is a linear rather than a categorical variable.[†]

[†]Substituting a linear price term for a three-level categorical one has two immediate implications. First, any change in coding results in quite different absolute values of *D*-error. Second, in optimizing a linear coding for price, the search routine will try to eliminate alternatives with the middle level of price within brand. This focus on extremes is appropriate given the linear assumption, but, may preclude estimation of quadratic effects.

The first step involves finding an efficient design with all price interactions with brand and brand×container estimable. This unrestricted model has 7 *df* for main effects (two for brand, two for container, two for flavor, and one for price), 12 *df* for two-way interactions (brand×price, brand×flavor, brand×container, container×price), and 4 *df* for the three-way interaction (brand×container×price). An efficient design for this unrestricted model has a *D*-error of .148. If this design is used for a restricted model in which price does not interact (7 *df* for main effects and 8 *df* for two-way interactions) then *D*-error drops to .118. The critical question is how much better still can one do by searching for the best design in the 15-parameter restricted model. The best design we find has a *D*-error of .110. Thus, assumption insurance in this case imposes a 6% (1 - .110/.118) efficiency loss, a reasonable cost given that prices will often interact with brands and containers.

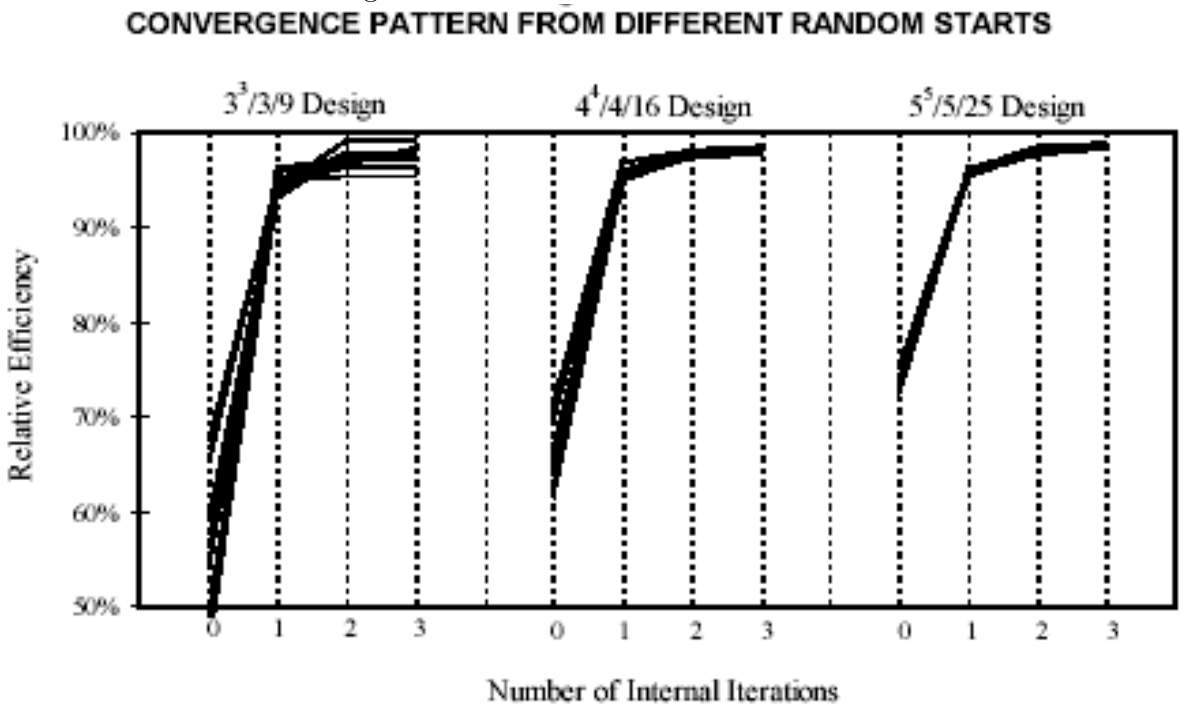
To summarize, the search routine allows estimates of the cost in efficiency of various design modifications and even changes in the model specification. Again, the important lesson is not the generalizations from the results of these particular examples, but rather an understanding of how these and similar questions can be answered in the context of any research study.

How Good Are These Designs? The preceding discussion has shown that our adaptation of the modified Fedorov algorithm can find estimable choice designs and answer a variety of useful questions. We still need to discuss the question, how close to optimal are these designs? The search is nonexhaustive, and there is no guarantee that the solutions are optimal or even nearly so. For some designs, such as the alternative-specific one shown previously, we can never be completely certain that the search process does not consistently find poor local optima. However, one can achieve some confidence from the pattern of results based on different random restarts; similar efficiencies emerging from different random starts indicate robustness of the resultant designs. An even stronger test is to assess efficiencies of the search process in cases where an optimal solution is known. While this cannot be done generally, we can test the absolute efficiency of certain symmetric designs, where the optimal design can be built using formal methods. We illustrated this kind of design in the three attribute, three level, three alternative, nine choice set ($3^3/3/9$) design discussed earlier, and found that the search routine was not able to find a better design. Now, we ask how good are our generated designs relative to three optimal designs: the design mentioned previously and two corresponding, but bigger designs − $4^4/4/16$ and a $5^5/5/25$ generic design.

For these types of designs we apply the proposed algorithm and compare our designs with the analytically generated ones. For each design, we used ten different random starts and three internal iterations. Figure 2 displays the impact of efficiency on different starting points and different numbers of internal iterations.

Figure 2 reveals important properties of the proposed algorithm. After the first iteration, the algorithm finds a choice design with about 90% relative efficiency, after a few more iterations, relative efficiencies approach 95%-99%. Further, this property appears to be independent of any initial starting design − the process converges just as quickly from a random start as from a rational one. These encouraging properties suggest important advantages for the practical use of the approach. First, in contrast to Huber and Zwerina (1996), the process does not require a rational starting design (which may be difficult to build). Second, since the process yields very efficient designs after only one or two iterations, most practical problems involving even large choice designs can be accommodated.

Figure 2
Convergence Pattern From Different Random Starts



Conclusions

We propose an adaptation of the modified Fedorov algorithm to construct statistically efficient choice designs with standard personal computers. The algorithm generates efficient designs quickly and is appropriate for all but the largest choice designs. The approach is illustrated with a SAS/IML program. SAS has the advantage of a general model statement that facilitates the building of choice designs with different model specifications. The cost of using SAS/IML software, however, is that the algorithm generally runs slower than a program developed in, for example, PASCAL or C.

There are three major advantages of using a computer to construct choice designs rather than deriving them from formal principles. First, computers are the only way we know to build designs that allow one to incorporate anticipated model parameters. Since the incorporation of this information can increase efficiency by 10% to 50% (see Huber and Zwerina 1996), this benefit alone justifies the use of computer search routines to find efficient choice designs.

The second advantage is that one is less restricted in design selection. Symmetric designs may not reflect the typically asymmetric characteristics of the real market. The adaptability of computerized searches is particularly important in choice studies that simulate consumer choice in a real market (Carson et al. 1994). Moreover, the process we propose allows the analyst to generate choice designs that account for any set of interactions, or alternative-specific effects of interest and critical tests of these assumptions. We illustrated a market emulation design that permits brand to interact with price, container, and flavor and can test the three-way interaction of brand by container by price. This pattern of alternative-specific effects would be very hard to build with standard designs, but it is easy to do with the computerized search routine by simply setting the model statement. The process can handle even more complex models, such as availability and attribute cross effects models (Lazari and Anderson 1994).

Finally, the ability to assess expected errors in parameters permits the researcher to examine the impact of different modifications in a given design, such as adding more choice sets or dropping a level from a factor. Most valuable, perhaps, is the ability to easily test designs for robustness. We provide one example of assumption insurance, but others are straightforward to generate. What happens to the efficiency of a design if there are interactions, but they are not included in the model statement? What kind of model will do a good job given a linear representation of price, but will also permit a test of curvature? What happens to the efficiency of the design if one's estimate of $\boldsymbol{\beta}$ is wrong?

There are several areas in which future research is needed. The first of these involves studies of the search process per se. We chose the modified Fedorov algorithm because it is robust and runs fast enough on today's desktop computers. As computing power increases, more exhaustive searches should be evaluated. For extremely large problems, faster and less reliable algorithms may be appropriate. Furthermore, while the approach builds efficient choice designs for multinomial logit models, efficiency issues with respect to other models, for example, nested logit and probit models, have yet to be explored.

A second area in which research would be fruitful involves the behavioral impact of different choice designs. The evaluations of our designs all implicitly assumed that the error level is constant regardless of the design. Many choice experiments use relatively small set sizes and few attributes reflecting an implicit recognition that "better" information comes from making the choice less complex. However, from a statistical perspective it is easy to show that smaller set sizes reduce statistical efficiency. In one example, we demonstrated that increasing the number of alternatives per choice set from three to four can increase efficiency by 16%. This gain depends on the assumption that respondent's error levels do not change. If they do increase, then that 16% percent gain might be lessened or even reversed. Thus, there is a need for a series of studies measuring respondents' error levels to tasks at different levels of complexity. Also, it is important to measure the degree of correspondence between the experimental tasks and the actual market behavior, choice experiments are intended to simulate. Such information is critical for correct trade-offs between design efficiency, measured here, and survey effectiveness, measured in the marketplace.

The purpose of this article is to demonstrate the important advantages of a flexible computerized search in generating efficient choice designs. The proposed adaptation of the modified Fedorov algorithm solves many of the practical problems involved in building choice designs, thus enabling more researchers to conduct choice experiments. Nevertheless, we want to emphasize that it does not preclude traditional design skills; they remain critical in determining the model specification and in assessing the choice designs produced by the computerized search.

Appendix

SAS/IML Code for the Proposed Choice Design Algorithm

The SAS code shows a simple implementation of the algorithm. In this example, the program finds a design with 27 choice sets and three alternatives per set. There are four attributes (brand, price, container, and flavor) each with three levels. A design is requested in which all main effects, the two-way interactions between brand and the other attributes, the two-way interaction between container and price, and the brand by price by container three-way interaction are estimable. Here, the parameters are assumed to be zero, but could be easily changed by setting other values.

A computer that evaluated all possible ($81^{81}/3!27! = 5 \times 9 \times 10^{125}$) designs would take numerous billion years. Instead, we use the modified Fedorov algorithm, which uses the following heuristic: find the best exchange for each design point given all of the other candidate points. With 81 candidate alternatives, 27 choice sets, 3 alternatives per set, (say) 3 internal iterations, and 2 random starts, $81 \times 27 \times 3 \times 3 \times 2 = 39,366$ exchanges must be evaluated. The algorithm tries to maximize $|\mathbf{X}'\mathbf{X}|$ rather than minimizing $|(\mathbf{X}'\mathbf{X})^{-1}|$ (note that $|(\mathbf{X}'\mathbf{X})^{-1}| = |\mathbf{X}'\mathbf{X}|^{-1}$). Each exchange requires then the evaluation of a matrix determinant, $|\mathbf{X}'\mathbf{X}|$. Fortunately, we do not have to evaluate this determinant from scratch for each exchange since $|\mathbf{X}'\mathbf{X} + \mathbf{x}'\mathbf{x}| = |\mathbf{X}'\mathbf{X}||I + \mathbf{x}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'|$ (Mardia et al. 1979). Each exchange evaluates a quadratic form, and in this example with three alternatives per choice set, the determinant of a 3×3 matrix. It should also be noted that this algorithm can handle a rank-deficient covariance matrix by operating on $|\mathbf{X}'\mathbf{X} + I\epsilon|$, where ϵ is a small number. This eliminates zero determinants so that less-than-full-rank codings and singular starting designs are not a problem. With these short cuts, one iteration required about 30 seconds on an ordinary 486 PC, implying that the algorithm is reasonable for many marketing contexts.

This appendix is provided simply to show the algorithm for those who might wish to implement or better understand it. If you want to use the algorithm, use the `%ChoicEff` autocall SAS macro documented in Kuhfeld (2003) starting on pages 479 and 481. See page 303 for an example. The `%ChoicEff` is much larger and more full-featured than the code shown in this appendix.

```
/*------------------------------Initial Set Up-------------------------------*/
%let beta  = 0 0 0 0 0 0 0 0              /* 8 main effects                    */
             0 0 0 0 0 0 0 0              /* brand x price, brand x container,*/
             0 0 0 0                      /* brand x flavor,                   */
             0 0 0 0                      /* price x container interactions    */
             0 0 0 0 0 0 0 0;             /* brand x price x container         */
%let nalts = 3;                           /* Number of alternatives            */
%let nsets = 27;                          /* Number of choice sets             */

proc plan ordered;                        /* Create candidate alternatives     */
   factors brand=3 price=3 contain=3 flavor=3 / noprint;
   output out=candidat;
   run;

proc transreg design data=candidat;       /* Code the candidate alternatives   */
model class(brand price contain flavor brand*price brand*contain
            brand*flavor contain*price brand*contain*price / effects);
   output out=tmp_cand;
   run;

proc contents p data=tmp_cand(keep=&_trgind); run;

/*-----------------------Begin Efficient Design Search----------------------*/
proc iml; file log;
   use tmp_cand(keep=&_trgind);           /* Identify candidate set for input */
   read all into cand;                    /* Read candidate set into IML      */
   utils  = exp(cand * {&beta}`);         /* exp(alternative utilities)       */
   np     = 1 / ncol(cand);               /* Exponent applied to determinant  */
   imat   = i(&nalts);                    /* Identity matrix                  */
   nobs   = &nsets # &nalts;              /* Total n of alts in choice design */
   ncands = nrow(cand);                   /* Number of candidates             */
   fuzz   = i(ncol(cand)) # 1e-8;         /* X`X ridge factor, avoid singular */

   start center(x, exputil);              /* Probability centering subroutine */
      do i = 1 to nrow(x) / &nalts;       /* Do for each choice set           */
         k = (i-1)#&nalts+1 : i#&nalts;   /* Choice set index vector          */
         p = exputil[k,]; p = p / sum(p); /* Probability of choice            */
         z = x[k,];                       /* Get choice set                   */
         x[k,] = (z - j(&nalts,1,1) *     /* Center choice set, absorb p's    */
                  p` * z) # sqrt(p);
         end;
      finish;

   /*---------------Create Designs With Different Random Starts--------------*/
   do desnum = 1 to 2;                    /* Number of designs to create      */

      indvec  = ceil(ncands *             /* Random index vector (indvec)     */
                uniform(j(1, nobs, 0)));  /* into candidates                  */
      des     = cand[indvec,];            /* Initial random design            */
```

```
run center(des, utils[indvec,]);    /* Probability center              */
currdet = det(des‘ * des);          /* Initial determinants, eff’s     */
maxdet  = currdet; oldeff = currdet ## np; fineff = oldeff;
if fineff <= 0 then err = .; else err = 1 / fineff;
put /// +8 ’Design   Iteration  D-Efficiency       D-Error’ /
        +8 ’-----------------------------------------’;
put +6 desnum 6. 0 10. +6 fineff best12. +2 err best12.;

/*-------------------------Internal Iterations-------------------------*/
do iter = 1 to 8 until(converge);   /* Iterate until convergence       */

   /*---------Consider Replacing Each Alternative in the Design---------*/
   do desi = 1 to nobs;             /* Process each alt in design      */
      ind = ceil(desi / &nalts);    /* Choice set number               */
      ind = (ind - 1) # &nalts + 1  /* Choice set index vector         */
            : ind # &nalts;
      besttry = des[ind,];          /* Store current choice set        */
      des[ind,] = 0;                /* Remove current choice set       */
      do i = 0 to 100 until(d ## np > 1e-8);
         xpx = des‘*des + i#i*fuzz; /* X‘X, ridged if necessary        */
         d   = det(xpx);            /* Determinant, if 0 then X‘X will */
         end;                       /* be ridged to make it nonsingular*/
      xpxinv = inv(xpx);            /* Inverse (all but current set)   */
      indcan = indvec[,ind];        /* Indvec for this choice set      */
      alt = mod(desi-1, &nalts) + 1;/* Alternative number              */

      /*-----------------Loop Over All of the Candidates----------------*/
      do candi = 1 to ncands;       /* Consider each candidate         */
         indcan[,alt] = candi;      /* Update indvec for this candidate*/
         tryit = cand[indcan,];     /* Candidate choice set            */
         run center(tryit,          /* Probability center              */
                    utils[indcan,]);
         currdet = d *              /* Update determinant              */
                   det(imat + tryit * xpxinv * tryit‘);

         /*------------Store Results When Efficiency Improves-----------*/
         if currdet > maxdet then do;
            maxdet        = currdet;/* Best determinant so far         */
            indvec[,desi] = candi;  /* Indvec of best design so far    */
            besttry       = tryit;  /* Best choice set so far          */
            end;
         end;

      des[ind,] = besttry;          /* Update design with new choice set*/
      end;

   /*----------Evaluate Efficiency/Convergence, Report Results----------*/
   neweff = maxdet ## np;           /* Newest efficiency               */
   converge = ((neweff - oldeff) /  /* Less than 1/2 percent           */
```

```
            max(oldeff,1e-8) < 0.005); /* improvement means convergence    */
        oldeff = neweff;               /* Store for use in next iteration  */
        fineff = det(des‘ * des) ## np;  /* Efficiency at end of iteration   */
        if fineff <= 0 then err = .; else err = 1 / fineff;
        put +12 iter 10. +6 fineff best12. +2 err best12.;
        end;

     /*----Store Efficiency, Index of Efficient Design, Covariance Matrix----*/
     final = final // (shape(desnum || fineff, nobs, 2) || indvec‘);
     cov = cov // (shape(desnum || fineff, ncol(des), 2) ||
                 sweep(des‘ * des, 1 : ncol(des)));
     end;

  /*----------------------Write Results to SAS Data Sets---------------------*/
  create cov     var({design effic &_trgind});   append from cov;
  create results var({design effic index   });   append from final;
  quit;

/*-------------Store Actual Design Points, Using Indices from IML-------------*/
data results; set results; i=index; n=_n_; set candidat point=i; run;
proc sort; by descending effic n; run;    /* Put most eff design first        */
proc print; run;                          /* Print designs, best to worst     */
```

Discrete Choice

Warren F. Kuhfeld

Abstract

Discrete choice modeling is a popular technique in marketing research, transportation, and other areas, used for understanding people's stated choice among alternatives. We will discuss designing a choice experiment, preparing the questionnaire, inputting and processing the data, performing the analysis, and interpreting the results.*

Introduction

This chapter shows you how to use the multinomial logit model (McFadden, 1974; Manski and McFadden, 1981; Louviere and Woodworth, 1983) to investigate consumer's stated choices. The multinomial logit model is an alternative to full-profile conjoint analysis and is extremely popular in marketing research (Louviere, 1991; Carson et. al., 1994). Discrete choice, using the multinomial logit model, is sometimes referred to as "choice-based conjoint." However, discrete choice uses a different model from full-profile conjoint analysis. Discrete choice applies a nonlinear model to aggregate choice data, whereas full-profile conjoint analysis applies a linear model to individual-level rating or ranking data.

Several examples are discussed.†

- The candy example (page 96) is a first, very simple example that discusses the multinomial logit model, the input data, analysis, results, and computing probability of choice.
- The fabric softener example (page 108) is a small, somewhat more realistic example that discusses designing the choice experiment, randomization, generating the questionnaire, entering and processing the data, analysis, results, probability of choice, and custom questionnaires.
- The first vacation example (page 134) is a larger, symmetric example that discusses designing the choice experiment, blocks, randomization, generating the questionnaire, entering and processing the data, coding, and alternative-specific effects.
- The second vacation example (page 178) is a larger, asymmetric example that discusses designing the choice experiment, blocks, blocking an existing design, interactions, generating the questionnaire, generating artificial data, reading, processing, and analyzing the data, aggregating the data to save time and memory.

*Copies of this chapter (TS-689E) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market . This document would not be possible without the help of Randy Tobias who contributed to the discussion of experimental design and Ying So who contributed to the discussion of analysis. Randy Tobias wrote PROC FACTEX and PROC OPTEX. Ying So wrote PROC PHREG. Warren F. Kuhfeld wrote PROC TRANSREG and the macros.

†All of the example data sets are artificially generated.

- The brand choice example (page 205) is a small example that discusses the processing of aggregate data, the mother logit model, and the likelihood function.
- The food product example (page 228) is a medium sized example that discusses asymmetry, coding, checking the design to ensure that all effects are estimable, availability cross effects, interactions, overnight design searches, modeling subject attributes, and designs when balance is of primary importance.
- The drug allocation example (page 284) is a small example that discusses data processing for studies where respondents potentially make multiple choices.
- The chair example (page 303) is a purely generic-attributes study, and it uses the `%ChoicEff` macro to create experimental designs.
- The last example sections (page 323) contain miscellaneous examples including improving an existing design, augmenting a design with some choice sets are fixed in advance, and partial profiles. Also see page 566 for an example of a choice design with a complicated set of restrictions.

This chapter relies heavily on a number of macros and procedures.

- We use the `%MktRuns` autocall macro to suggest design sizes. See page 600 for documentation.
- We use the `%MktEx` autocall macro to generate most of our experimental designs. See page 546 for documentation.
- We use the `%MktEval` autocall macro to evaluate our designs. See page 542 for documentation.
- We use the `%ChoicEff` autocall macro to generate certain specialized choice designs. See page 481 for documentation.
- We use the autocall macros `%MktRoll`, `%MktMerge`, and `%MktAllo` to prepare the data and design for analysis. See pages 595, 588, and 512 for documentation.
- We use PROC TRANSREG to do all of our design coding.
- We use the `%PhChoice` autocall macro to customize our printed output. This macro uses PROC TEMPLATE and ODS (Output Delivery System) to customize the output from PROC PHREG, which fits the multinomial logit model. See page 606 for documentation.
- The `%MktBal` macro can be used to make perfectly balanced designs. See page 515 for documentation.
- The `%MktBlock` macro can be used to block a linear or choice design. See page 518 for documentation.
- The `%MktDups` macro can be used to search for duplicate runs or choice sets. See page 534 for documentation.
- The `%MktLab` macro can be used to assign different variable names, labels and levels to experimental designs and to add an intercept. See page 577 for documentation.
- The `%MktOrth` macro can be used to list orthogonal experimental designs that the `%MktEx` macro can produce. See page 590 for documentation.

All of these macros are distributed with SAS 9.1 as autocall macros (see page 479 for more information on autocall macros). If you are running SAS 9 or any earlier version of SAS, get the latest macros from the web or by writing Warren.Kuhfeld@sas.com. This chapter and the macros are available from the Technical Support web site at http://support.sas.com/techsup/tnote/tnote_stat.html#market .

Preliminaries

This section defines some design terms that we will use later and shows how to customize the multinomial logit output listing.

Experimental Design Terminology

An *experimental design* is a plan for running an experiment. The *factors* of an experimental design are the columns or variables that have two or more fixed values, or *levels*. The rows of a design are called *runs* and correspond to product profiles in a full-profile conjoint study or choice sets in a discrete choice study. Experiments are performed to study the effects of the factor levels on the dependent or response variable. In a discrete-choice study, the factors are the attributes of the hypothetical products or services, and the response is choice. For example, the following table contains an experimental design in 8 runs with three factors, Brand 1 price, Brand 2 price, and Brand 3 price. Each factor has two levels, \$1.99 and \$2.99.

Linear Design
For a Choice Model

Brand 1 Price	Brand 2 Price	Brand3 Price
1.99	1.99	1.99
1.99	1.99	2.99
1.99	2.99	1.99
1.99	2.99	2.99
2.99	1.99	1.99
2.99	1.99	2.99
2.99	2.99	1.99
2.99	2.99	2.99

This is an example of a *full-factorial design*. It consists of all possible combinations of the levels of the factors ($2^3 = 2 \times 2 \times 2 = 8$ runs). Full-factorial designs allow you to estimate both main effects and interactions. A *main effect* is a simple effect, such as a price or brand effect. In a main-effects model, for example, the brand effect is the same at the different prices and the price effect is the same for the different brands. *Interactions* involve two or more factors, such as a brand by price interaction. In a model with interactions, for example, brand preference is different at the different prices and the price effect is different for the different brands. In Figure 1, there is a main effect for price, and utility increases by one when price goes from \$2.99 to \$1.99 for all brands. Similarly, the change in utility from Brand 1 to Brand 2 to Brand 3 does not depend on price. In contrast, in there are interactions in Figure 2, so the price effect is different depending on brand, and the brand effect is different depending on price.

In a full-factorial design, all main effects, all two-way interactions, and all higher-order interactions are estimable and uncorrelated. The problem with a full-factorial design is that, for most practical situations, it is too cost-prohibitive and tedious to have subjects consider all possible combinations. For example, with five factors, two at four levels and three at five levels (denoted $4^2 5^3$), there are $4 \times 4 \times 5 \times 5 \times 5 = 2000$ combinations in the full-factorial design. For this reason, researchers often use *fractional-factorial designs*, which have fewer runs than full-factorial designs. The price of having fewer runs is that some effects become confounded. Two effects are *confounded* or *aliased* when they

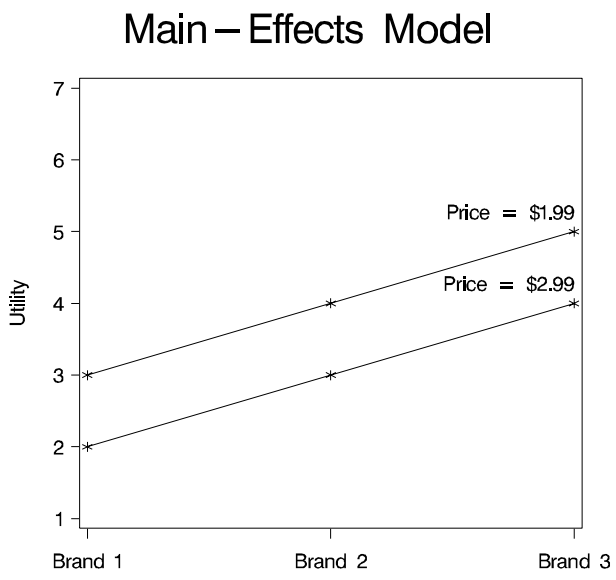


Figure 1

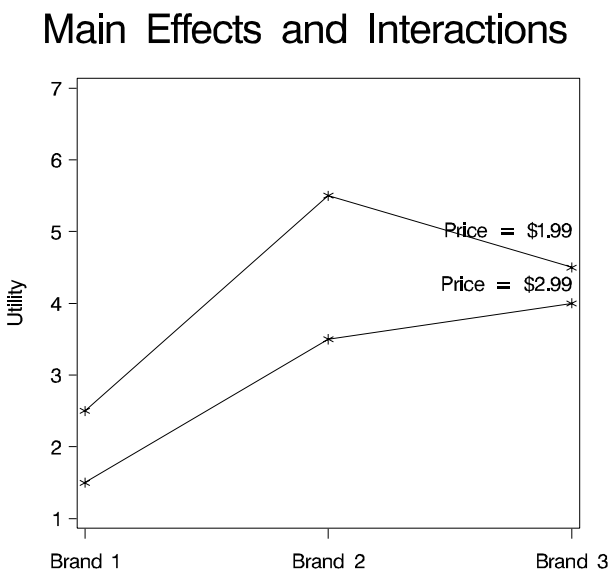


Figure 2

are not distinguishable from each other. This means that lower-order effects such as main effects or two-way interactions may be aliased with higher order interactions in most of our designs. We estimate lower-order effects by assuming that higher-order effects are zero or negligible. See page 246 for an example of aliasing.

An *orthogonal array* is an experimental design in which all estimable effects are uncorrelated. Orthogonal arrays are categorized by their *resolution*. The resolution identifies which effects, possibly including interactions, are estimable. For example, for resolution III designs, all main effects are estimable free of each other, but some of them are confounded with two-factor interactions. For resolution V designs, all main effects and two-factor interactions are estimable free of each other. More generally, if resolution (r) is odd, then effects of order $e = (r-1)/2$ or less are estimable free of each other. However, at least some of the effects of order e are confounded with interactions of order $e+1$. If r is even, then effects of order $e = (r-2)/2$ are estimable free of each other and are also free of interactions of order $e+1$. Higher resolutions require larger designs. Orthogonal arrays come in specific numbers of runs (such as 16, 18, 20, 24, 27, 28, ...) for specific numbers of factors with specific numbers of levels. Resolution III orthogonal arrays are frequently used in marketing research.

When each level occurs equally often within each factor the design is said to be *balanced*. When every *pair* of levels occurs equally often across all pairs of factor, the design is said to be *orthogonal*. Another way in which a design can be orthogonal is when the frequencies for level pairs are proportional instead of equal. For example, with 2 two-level factors, an orthogonal design could have pair frequencies proportional to 2, 4, 4, 8. Such a design will not be balanced − one level will occur twice as often as the other.

An orthogonal array is both balanced and orthogonal, and hence 100% efficient and optimal. Efficiency, which is explained in the next section, is a measure of the goodness of the experimental design. The term "orthogonal array," as it is sometimes used in practice, is imprecise. It is correctly used to refer to designs that are both orthogonal and balanced, and hence optimal. The term is sometimes also used to refer to designs that are orthogonal but not balanced, and hence not 100% efficient and sometimes not even optimal. Imbalance is a generalized form of nonorthogonality. In a design that is not balanced,

the intercept is not orthogonal to the unbalanced factors. Imbalance increases the variances of the parameter estimates and decreases efficiency.

Experimental Design Efficiency

The goodness or *efficiency* of an experimental design can be quantified. Common measures of the efficiency of an $(N_D \times p)$ design matrix $\mathbf{X}$ are based on the *information matrix* $\mathbf{X}'\mathbf{X}$. The variance-covariance matrix of the vector of parameter estimates $\hat{\boldsymbol{\beta}}$ in a least-squares analysis is proportional to $(\mathbf{X}'\mathbf{X})^{-1}$. An efficient design will have a "small" variance matrix, and the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}$ provide measures of its "size." The two most prominent efficiency measures are based on quantifying the idea of matrix size by averaging (in various ways) the eigenvalues or variances. *A-efficiency* is a function of the arithmetic mean of the eigenvalues, which is also the arithmetic mean of the variances and is given by trace $((\mathbf{X}'\mathbf{X})^{-1})/p$. (The trace is the sum of the diagonal elements of a matrix, which is the sum of the eigenvalues.) *D-efficiency* is a function of the geometric mean of the eigenvalues, which is given by $|(\mathbf{X}'\mathbf{X})^{-1}|^{1/p}$. (The determinant, $|(\mathbf{X}'\mathbf{X})^{-1}|$, is the product of the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}$.) A third common efficiency measure, *G-efficiency* is based on σ_M, the maximum standard error for prediction over the candidate set. All three of these criteria are convex functions of the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}$ and hence are usually highly correlated.

For all three criteria, if a balanced and orthogonal design exists, then it has optimum *efficiency*; conversely, the more efficient a design is, the more it tends toward balance and orthogonality. A design is balanced and orthogonal when $(\mathbf{X}'\mathbf{X})^{-1}$ is diagonal, $\frac{1}{N_D}\mathbf{I}$, for a suitably coded $\mathbf{X}$. A design is orthogonal when the submatrix of $(\mathbf{X}'\mathbf{X})^{-1}$, excluding the row and column for the intercept, is diagonal; there may be off-diagonal nonzeros for the intercept. A design is balanced when all off-diagonal elements in the intercept row and column are zero.

These measures of efficiency can be scaled to range from 0 to 100 (see page 91 for the orthogonal coding of $\mathbf{X}$ that must be used with these formulas):

$$A\text{-efficiency} = 100 \times \frac{1}{N_D \text{ trace}\left((\mathbf{X}'\mathbf{X})^{-1}\right)/p}$$

$$D\text{-efficiency} = 100 \times \frac{1}{N_D \left|(\mathbf{X}'\mathbf{X})^{-1}\right|^{1/p}}$$

$$G\text{-efficiency} = 100 \times \frac{\sqrt{p/N_D}}{\sigma_M}$$

These efficiencies measure the goodness of the design relative to hypothetical orthogonal designs that may not exist, so they are not useful as absolute measures of design efficiency. Instead, they should be used relatively, to compare one design to another for the same situation. Designs with efficiencies that are not near 100 may be perfectly satisfactory. Throughout this book, we will use the `%MktEx` macro to find good, efficient experimental designs.

Conjoint, Linear, and Choice Designs

Consider a simple example of three brands each at two prices. We could use linear model theory to create a design for a full-profile conjoint study. The full-profile conjoint design has two factors, one for brand and one for price. For the same brands and prices, we could instead use linear model theory to create a *linear design* from which we could construct a *choice design* to use in a discrete choice study. The conjoint and linear designs are shown next.

Full-Profile Conjoint Design

Brand	Price
1	1.99
1	2.99
2	1.99
2	2.99
3	1.99
3	2.99

Linear Design Used to Make a Choice Design

Brand 1 Price	Brand 2 Price	Brand3 Price
1.99	1.99	1.99
1.99	2.99	2.99
2.99	1.99	2.99
2.99	2.99	1.99

The linear design for a pricing study with three brands has three factors (Brand 1 Price, Brand 2 Price, and Brand 3 Price) and one row for each choice set. More generally, the linear design has one factor for each attribute of each alternative (or brand), and brand is not a factor in the linear design. Each brand is a "bin" into which its factors are collected. Before we fit the choice model, we will construct a choice design from the linear design and code the choice design. See the three following tables.

Linear Design

1	2	3
1.99	1.99	1.99
1.99	2.99	2.99
2.99	1.99	2.99
2.99	2.99	1.99

Choice Design

Set	Brand	Price
1	1	1.99
	2	1.99
	3	1.99
2	1	1.99
	2	2.99
	3	2.99
3	1	2.99
	2	1.99
	3	2.99
4	1	2.99
	2	2.99
	3	1.99

Choice Design Coding

	Brand Effects			Brand by Price		
Set	1	2	3	1	2	3
1	1	0	0	1.99	0	0
	0	1	0	0	1.99	0
	0	0	1	0	0	1.99
2	1	0	0	1.99	0	0
	0	1	0	0	2.99	0
	0	0	1	0	0	2.99
3	1	0	0	2.99	0	0
	0	1	0	0	1.99	0
	0	0	1	0	0	2.99
4	1	0	0	2.99	0	0
	0	1	0	0	2.99	0
	0	0	1	0	0	1.99

The linear design has one row per choice set. The choice design has three rows for each choice set. The linear design and the choice design contain different arrangements of the exact same information. In the linear design, brand is a bin into which its factors are collected (in this case one factor per brand). In the choice design, brand and price are both factors, because the design has been rearranged from one row per choice set to one row per alternative per choice set. For this problem, with only one attribute per brand, the first row of the choice design matrix corresponds to the first value in the linear design matrix, Brand 1 at \$1.99. The second row of the choice design matrix corresponds to the second value in the linear design matrix, Brand 2 at \$1.99. The third row of the choice design matrix corresponds to the third value in the linear design matrix, Brand 3 at \$1.99, and so on.

A design is coded by replacing each factor with one more columns of *indicator variables* (which are often referred to as "dummy variables") or other codings. In this example, a brand factor is replaced by the three binary variables. We will go through how to construct and code linear and choice designs many times in the examples using a number of different codings. For now, just notice that the conjoint design is different from the linear design, which is different from the choice design. They aren't even the same size! Also note that we *cannot* use linear efficiency criteria to directly construct the choice design bypassing the linear design step. Usually, we will use the `%MktEx` macro to make a linear design, the `%MktRoll` macro to convert it into a choice design, and the TRANSREG procedure to code the choice design.

Efficiency of a Choice Design

All of the efficiency theory discussed so far concerned linear models. In linear models, the parameter estimates $\hat{\boldsymbol{\beta}}$ have variances proportional to $(\mathbf{X}'\mathbf{X})^{-1}$. In contrast, the variances of the parameter estimates in the discrete choice multinomial logit model are given by

$$V(\hat{\beta}) = -\left[\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\right]^{-1} = \left[\Sigma_{k=1}^{n} N \left[\frac{\Sigma_{j=1}^{m} \exp(x_j'\beta) x_j x_j'}{\Sigma_{j=1}^{m} \exp(x_j'\beta)} - \frac{(\Sigma_{j=1}^{m} \exp(x_j'\beta) x_j)(\Sigma_{j=1}^{m} \exp(x_j'\beta) x_j)'}{(\Sigma_{j=1}^{m} \exp(x_j'\beta))^2}\right]\right]^{-1}$$

where

$$\ell(\beta) = \Pi_{k=1}^{n} \frac{\exp((\Sigma_{j=1}^{m} f_j x_j')\beta)}{(\Sigma_{j=1}^{m} \exp(x_j'\beta))^N}$$

m − brands
n − choice sets
N − people

In the choice model, ideally we would like to pick $\mathbf{x}$'s that make this variance matrix "small." Unfortunately, we cannot do this unless we know $\boldsymbol{\beta}$, and if we knew $\boldsymbol{\beta}$, we would not need to do the experiment. However, in the chair example on pages 303−322, we will see how to make an efficient choice design when we are willing to make assumptions about $\boldsymbol{\beta}$.

Because we do not know $\boldsymbol{\beta}$, we will often create experimental designs for choice models using efficiency criteria for linear models. We make a good design for a linear model by picking $\mathbf{x}$'s that minimize functions of $(\mathbf{X}'\mathbf{X})^{-1}$ then convert our linear design into a choice design. Certain assumptions must be made before applying ordinary general-linear-model theory to problems in marketing research. The usual goal in linear modeling is to estimate parameters and test hypotheses about those parameters. Typically, independence and normality are assumed. In full-profile conjoint analysis, each subject rates all products and separate ordinary-least-squares analyses are run for each subject. This is not a standard general linear model; in particular, observations are not independent and normality cannot be assumed. Discrete choice models, which are nonlinear, are even more removed from the general linear model.

Marketing researchers have always made the critical assumption that designs that are good for general linear models are also good designs for conjoint analysis and discrete choice models. We also make this assumption. We will assume that an efficient design for a linear model is a good design for the multinomial logit model used in discrete choice studies. We assume that if we create the linear design (one row per choice set and all of the attributes of all of the alternatives comprise that row), and if we strive for linear-model efficiency (near balance and orthogonality), then we will have a good design for measuring the utility of each alternative and the contributions of the factors to that utility. When we construct choice designs in this way, our designs will have two nice properties. 1) Each attribute level will occur equally often (or at least nearly equally often) for each attribute of each alternative across all choice sets. 2) Each attribute will be independent of every other attribute (or at least nearly independent), both those in the current alternative and those in all of the other alternatives. The design techniques discussed in this book, based on the assumption that linear design efficiency is a good surrogate for choice design goodness, have been used quite successfully in the field for many years.

In most of the examples, we will use the `%MktEx` macro to create a good linear design, from which we will construct our choice design. This seems to be a good safe strategy. It is a good strategy because it gives designs where all attributes, both within and between alternatives, are orthogonal or at least nearly so. It is safe in the sense that you have enough choice sets and collect the right information so that very complex models, including models with alternative-specific effects, availability effects, and cross effects, can be fit. However, it is good to remember that when you run the `%MktEx` macro and you get an efficiency value, it corresponds to the linear design, not the choice design. It is a surrogate for the criterion of interest, the efficiency of the choice design, which is unknowable unless you know the parameters.

Canonical Correlations

We will use canonical correlations to evaluate nonorthogonal designs and the extent to which factors are correlated or not independent. To illustrate, consider a design with four three-level factors in 9 runs shown next along with its coding.

Linear Design

x1	x2	x3	x4
1	1	1	1
1	2	3	3
1	3	2	2
2	1	2	3
2	2	1	2
2	3	3	1
3	1	3	2
3	2	2	1
3	3	1	3

Coded Linear Design

x1			x2			x3			x4		
1	2	3	1	2	3	1	2	3	1	2	3
1	0	0	1	0	0	1	0	0	1	0	0
1	0	0	0	1	0	0	0	1	0	0	1
1	0	0	0	0	1	0	1	0	0	1	0
0	1	0	1	0	0	0	1	0	0	0	1
0	1	0	0	1	0	1	0	0	0	1	0
0	1	0	0	0	1	0	0	1	1	0	0
0	0	1	1	0	0	0	0	1	0	1	0
0	0	1	0	1	0	0	1	0	1	0	0
0	0	1	0	0	1	1	0	0	0	0	1

Each three-level factor can be coded with three columns that contains a one in the first column when the factor level is 1 and zeros in columns 2 and 3, or a one in the second column when the factor level is 2 and zeros in columns 1 and 3, or a one in the third column when the factor level is 3 and zeros in columns 1 and 2. Furthermore, a factor can be recoded by applying a coefficient vector $\boldsymbol{\alpha}' = (\alpha_1\ \alpha_2\ \alpha_3)$ or $\boldsymbol{\beta}' = (\beta_1\ \beta_2\ \beta_3)$ to a coded factor to create a single column. In other words, the original coding of (1 2 3) can be replaced with arbitrary $(\alpha_1\ \alpha_2\ \alpha_3)$ or $(\beta_1\ \beta_2\ \beta_3)$ If two factors are orthogonal, then for all choices of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the simple correlation between recoded columns is zero. A *canonical correlation* shows the maximum correlation between two recoded factors that can be obtained with the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This design, 3^4 in 9 runs is orthogonal so for all pairs of factors and all choices of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the simple correlations between recoded factors will be zero. The canonical correlation between a factor and itself is 1.0.

For nonorthogonal designs and designs with interactions, the canonical-correlation matrix is not a substitute for looking at the variance matrix discussed on pages 146, 191, and 556. It just provides a quick and more-compact picture of the correlations between the factors. The variance matrix is sensitive to the actual model specified and the actual coding. The canonical-correlation matrix just tells you if there is some correlation between the main effects. A matrix of canonical correlations provides a useful picture of the orthogonality or lack of orthogonality in a design. For example, this canonical-correlation matrix from the vacation example, page 144, shown next shows that a design with 16 factors that is mostly orthogonal, but `x13-x15` are not orthogonal to each other. However, with $r^2 = 0.25^2 = 0.0625$, these factors are nearly independent.

Canonical Correlation Matrix

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16
x1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
x4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
x5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
x6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
x7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
x8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
x9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
x10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
x11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
x12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
x13	0	0	0	0	0	0	0	0	0	0	0	0	1	0.25	0.25	0
x14	0	0	0	0	0	0	0	0	0	0	0	0	0.25	1	0.25	0
x15	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	1	0
x16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Coding, Efficiency, Balance, and Orthogonality

We mentioned on page 86 that we use a special orthogonal coding of **X** when computing design efficiency. This section shows that coding and other codings. Even if you gloss over the mathematical details, this section is informative, because it provides insights into coding and the meaning of 100% efficiency and less than 100% efficiency designs.

Here are nonorthogonal less than-full-rank binary or indicator codings for two-level through five-level factors. We will use these codings in many places throughout the examples.

Two-Level

a	1	0
b	0	1

Three-Level

a	1	0	0
b	0	1	0
c	0	0	1

Four-Level

a	1	0	0	0
b	0	1	0	0
c	0	0	1	0
d	0	0	0	1

Five-Level

a	1	0	0	0	0
b	0	1	0	0	0
c	0	0	1	0	0
d	0	0	0	1	0
e	0	0	0	0	1

Here are nonorthogonal full-rank binary or indicator codings for two-level through five-level factors. We will use these codings in many places throughout the examples.

Two-Level

a	1
b	0

Three-Level

a	1	0
b	0	1
c	0	0

Four-Level

a	1	0	0
b	0	1	0
c	0	0	1
d	0	0	0

Five-Level

a	1	0	0	0
b	0	1	0	0
c	0	0	1	0
d	0	0	0	1
e	0	0	0	0

Here are nonorthogonal effects codings for two-level[‡] through five-level factors. We will use these codings in many places throughout the examples.

Two-Level

a	1
b	-1

Three-Level

a	1	0
b	0	1
c	-1	-1

Four-Level

a	1	0	0
b	0	1	0
c	0	0	1
d	-1	-1	-1

Five-Level

a	1	0	0	0
b	0	1	0	0
c	0	0	1	0
d	0	0	0	1
e	-1	-1	-1	-1

Here are the orthogonal codings for two-level through five-level factors that the software uses internally for design evaluation.

Two-Level

a	1.00
b	-1.00

Three-Level

a	1.22	-0.71
b	0	1.41
c	-1.22	-0.71

Four-Level

a	1.41	-0.82	-0.58
b	0	1.63	-0.58
c	0	0	1.73
d	-1.41	-0.82	-0.58

Five-Level

a	1.58	-0.91	-0.65	-0.50
b	0	1.83	-0.65	-0.50
c	0	0	1.94	-0.50
d	0	0	0	2.00
e	-1.58	-0.91	-0.65	-0.50

Notice that the sum of squares for the orthogonal coding of the two-level factor is 2; for all of the columns of the three-level factor, the sums of squares are 3; for the four-level factor, the sums of squares are all 4; and for the five-level factor, the sums of squares are all 5. Also notice that each column within a factor is orthogonal to all of the other columns − the sum of cross products is zero. For example, in the last two columns of the five-level factor, $-0.65\times-0.5+-0.65\times-0.5+1.94\times-.05+0\times2+-0.65\times-0.5=0$. Finally notice that the codings for each level form a contrast − the *ith* level versus all of the preceding levels and the last level.

Recall that our measures of design efficiency are scaled to range from 0 to 100.

$$A\text{-efficiency} = 100\times\frac{1}{N_D\ \text{trace}\left((\mathbf{X}'\mathbf{X})^{-1}\right)/p}$$

$$D\text{-efficiency} = 100\times\frac{1}{N_D\ |(\mathbf{X}'\mathbf{X})^{-1}|^{1/p}}$$

When computing *D*-efficiency or *A*-efficiency, we code $\mathbf{X}$ so that when the design is orthogonal and balanced, $\mathbf{X}'\mathbf{X}=N_D\mathbf{I}$ where $\mathbf{I}$ is a $p\times p$ identity matrix. When our design is orthogonal and balanced, $(\mathbf{X}'\mathbf{X})^{-1}=\frac{1}{N_D}\mathbf{I}$, and $\text{trace}\left((\mathbf{X}'\mathbf{X})^{-1}\right)/p=|(\mathbf{X}'\mathbf{X})^{-1}|^{1/p}=1/N_D$. In this case, the two denominator terms cancel and efficiency is 100%. As the average variance increases, efficiency decreases.

This example shows the coding of a 2×6 full-factorial design in 12 runs using a coding function that requires that the factors levels are consecutive positive integers beginning with one and ending with m for an m-level factor. Note that the IML operator `#` performs ordinary (scalar) multiplication, and `##` performs exponentiation.

[‡]The two-level effects coding is orthogonal, but the three-level and beyond codings are not.

```
proc iml; /* orthogonal coding, levels must be 1, 2, ..., m */
   reset fuzz;

   start orthogcode(x);
      levels = max(x);
      xstar  = shape(x, levels - 1, nrow(x))‘;
      j = shape(1 : (levels - 1), nrow(x), levels - 1);
      r = sqrt(levels # (x / (x + 1))) # (j = xstar) -
          sqrt(levels / (j # (j + 1))) # (j > xstar | xstar = levels);
      return(r);
      finish;
   design = (1:2)‘ @ j(6, 1, 1) || {1, 1} @ (1:6)‘;
   x = j(12, 1, 1) || orthogcode(design[,1]) || orthogcode(design[,2]);
   print design[format=1.]  ’  ’ x[format=5.2 colname={’Int’ ’Two’ ’Six’}];

   xpx = x‘ * x;     print xpx[format=best5.];
   inv = inv(xpx);   print inv[format=best5.];
   d_eff = 100 / (nrow(x) #   det(inv) ## (1 / ncol(inv)));
   a_eff = 100 / (nrow(x) # trace(inv)       / ncol(inv));
   print    ’D-efficiency =’ d_eff[format=6.2]
         ’   A-efficiency =’ a_eff[format=6.2];
```

	X						
DESIGN	Int	Two	Six				
1 1	1.00	1.00	1.73	-1.00	-0.71	-0.55	-0.45
1 2	1.00	1.00	0.00	2.00	-0.71	-0.55	-0.45
1 3	1.00	1.00	0.00	0.00	2.12	-0.55	-0.45
1 4	1.00	1.00	0.00	0.00	0.00	2.19	-0.45
1 5	1.00	1.00	0.00	0.00	0.00	0.00	2.24
1 6	1.00	1.00	-1.73	-1.00	-0.71	-0.55	-0.45
2 1	1.00	-1.00	1.73	-1.00	-0.71	-0.55	-0.45
2 2	1.00	-1.00	0.00	2.00	-0.71	-0.55	-0.45
2 3	1.00	-1.00	0.00	0.00	2.12	-0.55	-0.45
2 4	1.00	-1.00	0.00	0.00	0.00	2.19	-0.45
2 5	1.00	-1.00	0.00	0.00	0.00	0.00	2.24
2 6	1.00	-1.00	-1.73	-1.00	-0.71	-0.55	-0.45

XPX

12	0	0	0	0	0	0
0	12	0	0	0	0	0
0	0	12	0	0	0	0
0	0	0	12	0	0	0
0	0	0	0	12	0	0
0	0	0	0	0	12	0
0	0	0	0	0	0	12

```
                                INV

      0.083     0     0     0     0     0     0
          0 0.083     0     0     0     0     0
          0     0 0.083     0     0     0     0
          0     0     0 0.083     0     0     0
          0     0     0     0 0.083     0     0
          0     0     0     0     0 0.083     0
          0     0     0     0     0     0 0.083

                     D_EFF                    A_EFF

     D-efficiency = 100.00    A-efficiency = 100.00
```

With this orthogonal and balanced design, $\mathbf{X}'\mathbf{X} = N_D\mathbf{I} = 12\mathbf{I}$, which means $(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{N_D}\mathbf{I} = \frac{1}{12}\mathbf{I}$, and D-efficiency = 100%.

With a nonorthogonal design, for example with the first 10 rows of the 2×6 full-factorial design, D-efficiency and A-efficiency are less than 100%.

```
design = design[1:10,];
x = j(10, 1, 1) || orthogcode(design[,1]) || orthogcode(design[,2]);
inv = inv(x‘ * x);
d_eff = 100 / (nrow(x) #   det(inv) ## (1 / ncol(inv)));
a_eff = 100 / (nrow(x) # trace(inv)       / ncol(inv));
print    ’D-efficiency =’ d_eff[format=6.2]
      ’   A-efficiency =’ a_eff[format=6.2];
quit;
```

```
                          D_EFF                    A_EFF

     D-efficiency =  92.90    A-efficiency =  84.00
```

In this case, $|(\mathbf{X}'\mathbf{X})^{-1}|^{1/p}$ and $\text{trace}\,((\mathbf{X}'\mathbf{X})^{-1})/p$ are multiplied in the denominator of the efficiency formulas by $\frac{1}{N_D} = \frac{1}{10}$. If an orthogonal and balanced design were available for this problem, then $(\mathbf{X}'\mathbf{X})^{-1}$ would equal $\frac{1}{N_D}\mathbf{I} = \frac{1}{10}\mathbf{I}$. Since an orthogonal and balanced design is not possible (6 does not divide 10), both D-efficiency and A-efficiency will be less than 100%, even with the optimal design. A main-effects, orthogonal and balanced design, with a variance matrix equal to $\frac{1}{N_D}\mathbf{I}$, is the standard by which 100% efficiency is gauged, even when we know such a design cannot exist. The standard is the average variance for the maximally efficient *potentially hypothetical* design, which is knowable, not the average variance for the optimal design, which for many practical problems we have no way of knowing.

For our purposes in this book, we will only consider experimental design with at least as many runs as parameters. A *saturated* or *tight* design has as many runs as there are parameters. The number of parameters in a main-effects model is 1 (for the intercept) plus the sum of the numbers of levels of all of the factors, minus the number of factors. Equivalently, since there are $m - 1$ parameters in an m-level factor, the number of parameters is $1 + \sum_{j=1}^{k}(m_j - 1)$ for k factors, each with m_j levels.

If a main-effects design is orthogonal and balanced, then the design must be at least as large as the saturated design and the number of runs must be divisible by the number of levels of all the factors and by the products of the number of levels of all pairs of factors. For example, a $2 \times 2 \times 3 \times 3 \times 3$ design cannot be orthogonal and balanced unless the number of runs is divisible by 2 (twice because there are two 2's), 3 (three times because there are three 3's), $2 \times 2 = 4$ (once, because there is one pair of 2's), $2 \times 3 = 6$ (six times, two 2's times three 3's), and $3 \times 3 = 9$ (three times, three pairs of 3's). If the design is orthogonal and balanced, then all of the divisions will work without a remainder. However, all of the divisions working is a necessary but not sufficient condition for the existence of an orthogonal and balanced design. For example, 45 is divisible by 3 and $3 \times 3 = 9$, but an orthogonal and balanced saturated design 3^{22} (22 three-level factors) in 45 runs does not exist.

Customizing the Multinomial Logit Output

The multinomial logit model for discrete choice experiments is fit using the SAS/STAT procedure PHREG (proportional hazards regression), with the `ties=breslow` option. The likelihood function of the multinomial logit model has the same form as a survival analysis model fit by PROC PHREG. The output from PROC PHREG is primarily designed for survival analysis studies. Before we fit the multinomial logit model with PROC PHREG, we can customize the output to make it more appropriate for choice experiments. We will use the autocall macro `%PhChoice` macro. See page 479 for information on autocall macros. You can run the following macro to customize PROC PHREG output.

```
%phchoice(on)
```

The macro uses PROC TEMPLATE and ODS (Output Delivery System) to customize the output from PROC PHREG. Running this code edits the templates and stores copies in `sasuser`. These changes will remain in effect until you delete them, so typically, you only have to run this macro once. Note that these changes assume that each effect in the choice model has a variable label associated with it, so there is no need to print variable names. If you are coding with PROC TRANSREG, this will usually be the case. To return to the default output from PROC PHREG, run the following macro.

```
%phchoice(off)
```

See page 606 for more information on the `%PhChoice` macro.

Candy Example

We begin with a very simple example. In this example, we will discuss the multinomial logit model, data input and processing, analysis, results, interpretation, and probability of choice. In this example, each of ten subjects was presented with eight different chocolate candies and asked to choose one. The eight candies consist of the 2^3 combinations of dark or milk chocolate, soft or chewy center, and nuts or no nuts. Each subject saw all eight candies and made one choice. Experimental choice data such as these are typically analyzed with a multinomial logit model.

The Multinomial Logit Model

The multinomial logit model assumes that the probability that an individual will choose one of the m alternatives, c_i, from choice set C is

$$p(c_i|C) = \frac{\exp(U(c_i))}{\sum_{j=1}^{m} \exp(U(c_j))} = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{\sum_{j=1}^{m} \exp(\mathbf{x}_j\boldsymbol{\beta})}$$

where $\mathbf{x}_i$ is a vector of alternative attributes and $\boldsymbol{\beta}$ is a vector of unknown parameters. $U(c_i) = \mathbf{x}_i\boldsymbol{\beta}$ is the utility for alternative c_i, which is a linear function of the attributes. The probability that an individual will choose one of the m alternatives, c_i, from choice set C is the exponential of the utility of the alternative divided by the sum of all of the exponentiated utilities.

There are $m = 8$ attribute vectors in this example, one for each alternative. Let $\mathbf{x}$ = (Dark/Milk, Soft/Chewy, Nuts/No Nuts) where Dark/Milk = (1 = Dark, 0 = Milk), Soft/Chewy = (1 = Soft, 0 = Chewy), Nuts/No Nuts = (1 = Nuts, 0 = No Nuts). The eight attribute vectors are

$\mathbf{x}_1 = (0\ 0\ 0)$ (Milk, Chewy, No Nuts)
$\mathbf{x}_2 = (0\ 0\ 1)$ (Milk, Chewy, Nuts)
$\mathbf{x}_3 = (0\ 1\ 0)$ (Milk, Soft, No Nuts)
$\mathbf{x}_4 = (0\ 1\ 1)$ (Milk, Soft, Nuts)
$\mathbf{x}_5 = (1\ 0\ 0)$ (Dark, Chewy, No Nuts)
$\mathbf{x}_6 = (1\ 0\ 1)$ (Dark, Chewy, Nuts)
$\mathbf{x}_7 = (1\ 1\ 0)$ (Dark, Soft, No Nuts)
$\mathbf{x}_8 = (1\ 1\ 1)$ (Dark, Soft, Nuts)

Say, hypothetically that $\boldsymbol{\beta}' = (4 \quad -2 \quad 1)$. That is, the part-worth utility for dark chocolate is 4, the part-worth utility for soft center is -2, and the part-worth utility for nuts is 1. The utility for each of the combinations, $\mathbf{x}_i\boldsymbol{\beta}$, would be as follows.

U(Milk, Chewy, No Nuts)	=	0×4	+	0×-2	+	0×1	=	0
U(Milk, Chewy, Nuts)	=	0×4	+	0×-2	+	1×1	=	1
U(Milk, Soft, No Nuts)	=	0×4	+	1×-2	+	0×1	=	-2
U(Milk, Soft, Nuts)	=	0×4	+	1×-2	+	1×1	=	-1
U(Dark, Chewy, No Nuts)	=	1×4	+	0×-2	+	0×1	=	4
U(Dark, Chewy, Nuts)	=	1×4	+	0×-2	+	1×1	=	5
U(Dark, Soft, No Nuts)	=	1×4	+	1×-2	+	0×1	=	2
U(Dark, Soft, Nuts)	=	1×4	+	1×-2	+	1×1	=	3

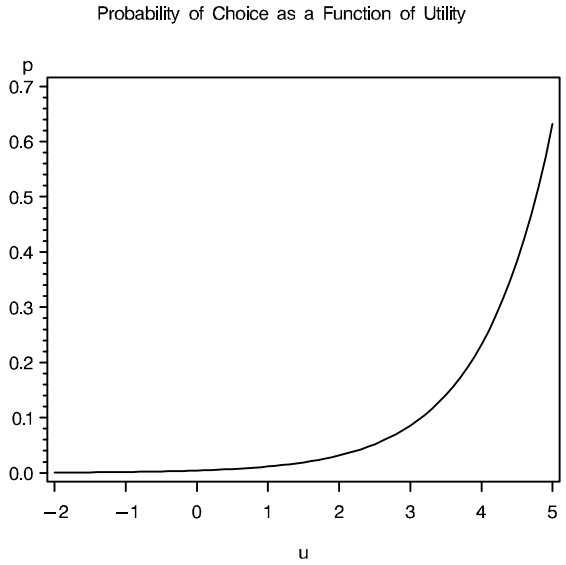
The denominator of the probability formula, $\sum_{j=1}^{m} \exp(\mathbf{x}_j\boldsymbol{\beta})$, is $\exp(0)+\exp(1)+\exp(-2)+\exp(-1)+\exp(4)+\exp(5)+\exp(2)+\exp(3) = 234.707$. The probability that each alternative is chosen, $\exp(\mathbf{x}_i\boldsymbol{\beta})/\sum_{j=1}^{m} \exp(\mathbf{x}_j\boldsymbol{\beta})$, is

p(Milk, Chewy, No Nuts)	=	exp(0)	/ 234.707	=	0.004
p(Milk, Chewy, Nuts)	=	exp(1)	/ 234.707	=	0.012
p(Milk, Soft, No Nuts)	=	exp(-2)	/ 234.707	=	0.001
p(Milk, Soft, Nuts)	=	exp(-1)	/ 234.707	=	0.002
p(Dark, Chewy, No Nuts)	=	exp(4)	/ 234.707	=	0.233
p(Dark, Chewy, Nuts)	=	exp(5)	/ 234.707	=	0.632
p(Dark, Soft, No Nuts)	=	exp(2)	/ 234.707	=	0.031
p(Dark, Soft, Nuts)	=	exp(3)	/ 234.707	=	0.086

Note that even combinations with a negative or zero utility have a nonzero probability of choice. Also note that adding a constant to the utilities will not change the probability of choice, however multiplying by a constant will.

Probability of choice is a nonlinear and increasing function of utility. The following plot shows the relationship between utility and probability of choice for this hypothetical situation.

```
data x;
   do u = -2 to 5 by 0.1;
      p = exp(u) / 234.707;
      output;
      end;
   run;
proc gplot;
   title h=1 'Probability of Choice as a Function of Utility';
   plot p * u;
   symbol1 i=join;
   run; quit;
```



This plot shows the function exp(−2) to exp(5), scaled into the range zero to one, the range of probability values. For the small negative utilities, the probability of choice is essentially zero. As utility increases beyond two, the function starts rapidly increasing.

In this example, the chosen alternatives are $\mathbf{x}_5$, $\mathbf{x}_6$, $\mathbf{x}_7$, $\mathbf{x}_5$, $\mathbf{x}_2$, $\mathbf{x}_6$, $\mathbf{x}_2$, $\mathbf{x}_6$, $\mathbf{x}_6$, $\mathbf{x}_6$. Alternative $\mathbf{x}_2$ was chosen 2 times, $\mathbf{x}_5$ was chosen 2 times, $\mathbf{x}_6$ was chosen 5 times, and $\mathbf{x}_7$ was chosen 1 time. The choice model likelihood for these data is the product of ten terms, one for each choice set for each subject. Each term consists of the probability that the chosen alternative is chosen. For each choice set, the utilities for all of the alternatives enter into the denominator, and the utility for the chosen alternative enters into the numerator. The choice model likelihood for these data is

$$\begin{aligned}
\mathcal{L}_C &= \frac{\exp(\mathbf{x}_5\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]} \times \frac{\exp(\mathbf{x}_6\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]} \times \frac{\exp(\mathbf{x}_7\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]} \times \frac{\exp(\mathbf{x}_5\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]} \times \\
&\quad \frac{\exp(\mathbf{x}_2\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]} \times \frac{\exp(\mathbf{x}_6\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]} \times \frac{\exp(\mathbf{x}_2\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]} \times \frac{\exp(\mathbf{x}_6\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]} \times \\
&\quad \frac{\exp(\mathbf{x}_6\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]} \times \frac{\exp(\mathbf{x}_6\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]} \\
&= \frac{\exp((2\mathbf{x}_2 + 2\mathbf{x}_5 + 5\mathbf{x}_6 + \mathbf{x}_7)\boldsymbol{\beta})}{\left[\sum_{j=1}^{8}\exp(\mathbf{x}_j\boldsymbol{\beta})\right]^{10}}
\end{aligned}$$

The Input Data

The data set consists of one observation for each alternative of each choice set for each subject. (A typical choice study has more than one choice set per person. This first example only has one choice set to help keep it simple.) All of the chosen and unchosen alternatives must appear in the data set. The data set must contain variables that identify the subject, the choice set, which alternative was chosen, and the set of alternatives from which it was chosen. In this example, the data set contains $10 \times 1 \times 8 = 80$ observations: 10 subjects each saw 1 choice set with 8 alternatives.

Typically, two variables are used to identify the choice sets, subject ID and choice set within subject. In this simple case where each subject only made one choice, the choice set variable is not necessary. However, we use it here to illustrate the general case. The variable `Subj` is the subject number, and `Set` identifies the choice set within subject. The chosen alternative is indicated by `c=1`, which means first choice. All second and subsequent choices are unobserved, so the unchosen alternatives are indicated by `c=2`, which means that all we know is that they would have been chosen after the first choice. Both the chosen and unchosen alternatives must appear in the input data set since both are needed to construct the likelihood function. The `c=2` observations enter into the denominator of the likelihood function, and the `c=1` observations enter into both the numerator and the denominator of the likelihood function. In this input DATA step, the data for four alternatives appear on one line, and all of the data for a choice set of eight alternatives appear on two lines. The DATA step shows data entry in the way that requires the fewest programming statements. Each execution of the `input` statement reads information about one alternative. The `@@` in the `input` statement specifies that SAS should not automatically go to a new input data set line when it reads the next row of data. This specification is needed here because each line in the input data set contains the data for four output data set rows. The data from the first two subjects is printed.

```
title 'Choice of Chocolate Candies';

data chocs;
   input Subj c Dark Soft Nuts @@;
   Set = 1;
   datalines;
 1 2 0 0 0    1 2 0 0 1    1 2 0 1 0    1 2 0 1 1
 1 1 1 0 0    1 2 1 0 1    1 2 1 1 0    1 2 1 1 1
 2 2 0 0 0    2 2 0 0 1    2 2 0 1 0    2 2 0 1 1
 2 2 1 0 0    2 1 1 0 1    2 2 1 1 0    2 2 1 1 1
 3 2 0 0 0    3 2 0 0 1    3 2 0 1 0    3 2 0 1 1
 3 2 1 0 0    3 2 1 0 1    3 1 1 1 0    3 2 1 1 1
 4 2 0 0 0    4 2 0 0 1    4 2 0 1 0    4 2 0 1 1
 4 1 1 0 0    4 2 1 0 1    4 2 1 1 0    4 2 1 1 1
 5 2 0 0 0    5 1 0 0 1    5 2 0 1 0    5 2 0 1 1
 5 2 1 0 0    5 2 1 0 1    5 2 1 1 0    5 2 1 1 1
 6 2 0 0 0    6 2 0 0 1    6 2 0 1 0    6 2 0 1 1
 6 2 1 0 0    6 1 1 0 1    6 2 1 1 0    6 2 1 1 1
 7 2 0 0 0    7 1 0 0 1    7 2 0 1 0    7 2 0 1 1
 7 2 1 0 0    7 2 1 0 1    7 2 1 1 0    7 2 1 1 1
 8 2 0 0 0    8 2 0 0 1    8 2 0 1 0    8 2 0 1 1
 8 2 1 0 0    8 1 1 0 1    8 2 1 1 0    8 2 1 1 1
 9 2 0 0 0    9 2 0 0 1    9 2 0 1 0    9 2 0 1 1
 9 2 1 0 0    9 1 1 0 1    9 2 1 1 0    9 2 1 1 1
10 2 0 0 0   10 2 0 0 1   10 2 0 1 0   10 2 0 1 1
10 2 1 0 0   10 1 1 0 1   10 2 1 1 0   10 2 1 1 1
;
proc print data=chocs noobs;
   where subj <= 2;
   var subj set c dark soft nuts;
   run;
```

Choice of Chocolate Candies

Subj	Set	c	Dark	Soft	Nuts
1	1	2	0	0	0
1	1	2	0	0	1
1	1	2	0	1	0
1	1	2	0	1	1
1	1	1	1	0	0
1	1	2	1	0	1
1	1	2	1	1	0
1	1	2	1	1	1

```
2     1     2     0     0     0
2     1     2     0     0     1
2     1     2     0     1     0
2     1     2     0     1     1
2     1     2     1     0     0
2     1     1     1     0     1
2     1     2     1     1     0
2     1     2     1     1     1
```

These next steps illustrate a more typical form of data entry. The experimental design is stored in a separate data set from the choices and is merged with the choices as the data are read, which produces the same results as the preceding steps.

```
title 'Choice of Chocolate Candies';

* Alternative Form of Data Entry;

data combos;                      /* Read the design matrix.     */
   input Dark Soft Nuts;
   datalines;
0 0 0
0 0 1
0 1 0
0 1 1
1 0 0
1 0 1
1 1 0
1 1 1
;
data chocs;                       /* Create the data set.        */
   input Choice @@; drop choice; /* Read the chosen combo num.  */
   Subj = _n_; Set = 1;           /* Store subj, choice set num. */
   do i = 1 to 8;                 /* Loop over alternatives.     */
      c = 2 - (i eq choice);      /* Designate chosen alt.       */
      set combos point=i;         /* Read design matrix.         */
      output;                     /* Output the results.         */
      end;
   datalines;
5 6 7 5 2 6 2 6 6 6
;
```

The variable `Choice` is the number of the chosen alternative. For each choice set, each of the eight observations in the experimental design is read. The `point=` option on the `set` statement is used to read the *ith* observation of the data set COMBOS. When `i` (the alternative index) equals `Choice` (the number of the chosen alternative), the logical expression `(i eq choice)` equals 1; otherwise it is 0. The statement `c = 2 - (i eq choice)` sets `c` to 1 (two minus one) when the alternative is chosen and 2 (two minus zero) otherwise. All eight observations in the COMBOS data set are read 10 times, once per subject. The resulting data set is the same as the one we created previously. In all of the remaining examples, we will simplify this process by using the `%MktMerge` macro to merge the design and data. The basic logic underlying this macro is shown in the preceding step. The number of a

chosen alternative is read, then each alternative of the choice set is read, the chosen alternative is flagged `(c = 1)`, and the unchosen alternatives are flagged `(c = 2)`. One observation per choice set per subject is read from the input data stream, and one observation per alternative per choice set per subject is written.

Choice and Survival Models

In SAS, the multinomial logit model is fit with the SAS/STAT procedure PHREG (proportional hazards regression), with the `ties=breslow` option. The likelihood function of the multinomial logit model has the same form as a survival analysis model fit by PROC PHREG.

In a discrete choice study, subjects are presented with sets of alternatives and asked to choose the most preferred alternative. The data for one choice set consist of one alternative that was chosen and $m-1$ alternatives that were not chosen. First choice was observed. Second and subsequent choices were not observed; it is only known that the other alternatives would have been chosen after the first choice. In survival analysis, subjects (rats, people, light bulbs, machines, and so on) are followed until a specific event occurs (such as failure or death) or until the experiment ends. The data are event times. The data for subjects who have not experienced the event (such as those who survive past the end of a medical experiment) are *censored*. The exact event time is not known, but it is known to have occurred after the censored time. In a discrete choice study, first choice occurs at time one, and all subsequent choices (second choice, third choice, and so on) are unobserved or censored. The survival and choice models are the same.

Fitting the Multinomial Logit Model

The data are now in the right form for analysis. To fit the multinomial logit model, use PROC PHREG as follows.

```
proc phreg data=chocs outest=betas;
   strata subj set;
   model c*c(2) = dark soft nuts / ties=breslow;
   label dark = 'Dark Chocolate' soft = 'Soft Center'
         nuts = 'With Nuts';
   run;
```

The `data=` option specifies the input data set. The `outest=` option requests an output data set called BETAS with the parameter estimates. The `strata` statement specifies that each combination of the variables `Set` and `Subj` forms a set from which a choice was made. Each term in the likelihood function is a *stratum*. There is one term or stratum per choice set per subject, and each is composed of information about the chosen and all the unchosen alternatives.

In the left side of the `model` statement, you specify the variables that indicate which alternatives were chosen and not chosen. While this could be two different variables, we will use one variable `c` to provide both pieces of information. The response variable `c` has values 1 (chosen or first choice) and 2 (unchosen or subsequent choices). The first `c` of the `c*c(2)` in the `model` statement specifies that `c` indicates which alternative was chosen. The second `c` specifies that `c` indicates which alternatives were not chosen, and `(2)` means that observations with values of 2 were not chosen. When `c` is set up such that 1 indicates the chosen alternative and 2 indicates the unchosen alternatives, always specify

`c*c(2)` on the left of the equal sign in the `model` statement. The attribute variables are specified after the equal sign. Specify `ties=breslow` after a slash to explicitly specify the likelihood function for the multinomial logit model. (Do not specify any other `ties=` options; `ties=breslow` specifies the most efficient and always appropriate way to fit the multinomial logit model.) The `label` statement is added since we are using a template that assumes each variable has a label.

Note that the `c*c(`n`)` syntax allows second choice `(c=2)` and subsequent choices (`c=3`, `c=4`, ...) to be entered. Just enter in parentheses one plus the number of choices actually made. For example, with first and second choice data specify `c*c(3)`. Note however that some experts believe that second and subsequent choice data are much less reliable than first choice data.

Multinomial Logit Model Results

The output is shown next. Recall that we used `%phchoice(on)` on page 95 to customize the output from PROC PHREG.

Choice of Chocolate Candies

The PHREG Procedure

Model Information

Data Set	WORK.CHOCS
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	80
Number of Observations Used	80

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Stratum	Subj	Set	Number of Alternatives	Chosen Alternatives	Not Chosen
1	1	1	8	1	7
2	2	1	8	1	7
3	3	1	8	1	7
4	4	1	8	1	7
5	5	1	8	1	7
6	6	1	8	1	7
7	7	1	8	1	7
8	8	1	8	1	7
9	9	1	8	1	7
10	10	1	8	1	7
Total			80	10	70

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	41.589	28.727
AIC	41.589	34.727
SBC	41.589	35.635

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12.8618	3	0.0049
Score	11.6000	3	0.0089
Wald	8.9275	3	0.0303

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Dark Chocolate	1	1.38629	0.79057	3.0749	0.0795
Soft Center	1	-2.19722	1.05409	4.3450	0.0371
With Nuts	1	0.84730	0.69007	1.5076	0.2195

The first table, `Model Information`, contains the input data set name, dependent variable name, censoring information, and tie handling option.

The `Summary of Subjects, Sets, and Chosen and Unchosen Alternatives` table is printed by default and should be used to check the data entry. In general, there are as many strata as there are combinations of the `Subj` and `Set` variables. In this case, there are ten strata. Each stratum must be composed of m alternatives. In this case, there are eight alternatives. The number of chosen alternatives should be 1, and the number of unchosen alternatives is $m - 1$ (in this case 7). **Always check the summary table to ensure that the data are arrayed correctly.**

The `Convergence Status` table shows that the iterative algorithm successfully converged. The next tables, `Model Fit Statistics` and `Testing Global Null Hypothesis: BETA=0` contain the overall fit of the model. The -2 LOG L statistic under `With Covariates` is 28.727 and the Chi-Square statistic is 12.8618 with 3 *df* (p=0.0049), which is used to test the null hypothesis that the attributes do not influence choice. At common alpha levels such as 0.05 and 0.01, we would reject the null hypothesis of no relationship between choice and the attributes. Note that 41.589 (-2 LOG L Without Covariates, which is -2 LOG L for a model with no explanatory variables) minus 28.727 (-2 LOG L With Covariates, which is -2 LOG L for a model with all explanatory variables) equals 12.8618 (Model Chi-Square, which is used to test the effects of the explanatory variables).

Next is the 'Multinomial Logit Parameter Estimates' table. For each effect, it contains the maximum likelihood parameter estimate, its estimated standard error (the square root of the corresponding diagonal element of the estimated variance matrix), the Wald Chi-Square statistic (the square of the parameter estimate divided by its standard error), the *df* of the Wald Chi-Square statistic (1 unless the corresponding parameter is redundant or infinite, in which case the value is 0), and the p-value of the Chi-Squared statistic with respect to a chi-squared distribution with one *df*. The parameter estimate with the smallest p-value is for soft center. Since the parameter estimate is negative, chewy is the more preferred level. Dark is preferred over milk, and nuts over no nuts, however only the p-value for Soft is less than 0.05.

Fitting the Multinomial Logit Model, All Levels

It is instructive to perform some manipulations on the data set and analyze it again. These steps will perform the same analysis as before, only now, coefficients for both levels of the three attributes are printed. Binary variables for the missing levels are created by subtracting the existing binary variables from 1.

```
data chocs2;
   set chocs;
   Milk = 1 - dark; Chewy = 1 - Soft; NoNuts = 1 - nuts;
   label dark = 'Dark Chocolate' milk   = 'Milk Chocolate'
         soft = 'Soft Center'    chewy  = 'Chewy Center'
         nuts = 'With Nuts'      nonuts = 'No Nuts';
   run;

proc phreg data=chocs2;
   strata subj set;
   model c*c(2) = dark milk soft chewy nuts nonuts / ties=breslow;
   run;
```

Choice of Chocolate Candies

The PHREG Procedure

Model Information

Data Set	WORK.CHOCS2
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	80
Number of Observations Used	80

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Stratum	Subj	Set	Number of Alternatives	Chosen Alternatives	Not Chosen
1	1	1	8	1	7
2	2	1	8	1	7
3	3	1	8	1	7
4	4	1	8	1	7
5	5	1	8	1	7
6	6	1	8	1	7
7	7	1	8	1	7
8	8	1	8	1	7
9	9	1	8	1	7
10	10	1	8	1	7
Total			80	10	70

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	41.589	28.727
AIC	41.589	34.727
SBC	41.589	35.635

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12.8618	3	0.0049
Score	11.6000	3	0.0089
Wald	8.9275	3	0.0303

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Dark Chocolate	1	1.38629	0.79057	3.0749	0.0795
Milk Chocolate	0	0	.	.	.
Soft Center	1	-2.19722	1.05409	4.3450	0.0371
Chewy Center	0	0	.	.	.
With Nuts	1	0.84730	0.69007	1.5076	0.2195
No Nuts	0	0	.	.	.

Now the zero coefficients for the reference levels, milk, chewy, and no nuts are printed. The part-worth utility for Milk Chocolate is a structural zero, and the part-worth utility for Dark Chocolate is larger at 1.38629. Similarly, the part-worth utility for Chewy Center is a structural zero, and the part-worth utility for Soft Center is smaller at -2.19722. Finally, the part-worth utility for No Nuts is a structural zero, and the part-worth utility for Nuts is larger at 0.84730.

Probability of Choice

The parameter estimates are used next to construct the estimated probability that each alternative will be chosen. The DATA step program uses the following formula to create the choice probabilities.

$$p(c_i|C) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{\sum_{j=1}^{m}\exp(\mathbf{x}_j\boldsymbol{\beta})}$$

```
* Estimate the probability that each alternative will be chosen;

data p;
   retain sum 0;
   set combos end=eof;

   * On the first pass through the DATA step (_n_ is the pass
     number), get the regression coefficients in B1-B3.
     Note that they are automatically retained so that they
     can be used in all passes through the DATA step.;

   if _n_ = 1 then
      set betas(rename=(dark=b1 soft=b2 nuts=b3));
   keep dark soft nuts p;
   array x[3] dark soft nuts;
   array b[3] b1-b3;
   * For each combination, create x * b;
   p = 0;
   do j = 1 to 3;
      p = p + x[j] * b[j];
      end;
```

```
   * Exponentiate x * b and sum them up;
   p   = exp(p);
   sum = sum + p;

   * Output sum exp(x * b) in the macro variable '&sum';
   if eof then call symput('sum',put(sum,best12.));
   run;

proc format;
   value df 1 = 'Dark' 0 = 'Milk';
   value sf 1 = 'Soft' 0 = 'Chewy';
   value nf 1 = 'Nuts' 0 = 'No Nuts';
   run;
* Divide each exp(x * b) by sum exp(x * b);
data p;
   set p;
   p = p / (&sum);
   format dark df. soft sf. nuts nf.;
   run;

proc sort;
   by descending p;
   run;
proc print;
  run;
```

Choice of Chocolate Candies

Obs	Dark	Soft	Nuts	p
1	Dark	Chewy	Nuts	0.50400
2	Dark	Chewy	No Nuts	0.21600
3	Milk	Chewy	Nuts	0.12600
4	Dark	Soft	Nuts	0.05600
5	Milk	Chewy	No Nuts	0.05400
6	Dark	Soft	No Nuts	0.02400
7	Milk	Soft	Nuts	0.01400
8	Milk	Soft	No Nuts	0.00600

The three most preferred alternatives are Dark/Chewy/Nuts, Dark/Chewy/No Nuts, and Milk/Chewy/Nuts.

Fabric Softener Example

In this example, subjects are asked to choose among fabric softeners. This example shows all of the steps in a discrete choice study, including experimental design creation and evaluation, creating the questionnaire, inputting the raw data, creating the data set for analysis, coding, fitting the discrete choice model, interpretation, and probability of choice. In addition, custom questionnaires are discussed. We assume the reader is familiar with the experimental design issues discussed in Kuhfeld, Tobias, and Garratt (1994), which starts on page 39. Some of these concepts are reviewed starting on page 84.

Set Up

The study involves four fictitious fabric softeners *Sploosh*, *Plumbbob*, *Platter*, and *Moosey*.[§] Each choice set consists of each of these four brands and a constant alternative *Another*. Each of the brands is available at three prices, \$1.49, \$1.99, and \$2.49. *Another* is only offered at \$1.99. There are 50 subjects, each of which will see the same choice sets. We can use the `%MktRuns` autocall macro to help us choose the number of choice sets. All of the autocall macros used in this book are documented starting on page 479. To use this macro, you specify the number of levels for each of the factors. With four brands each with three prices, you specify four 3's (or `3 ** 4`).

```
title 'Choice of Fabric Softener';

%mktruns( 3 3 3 3 )
```

The output first tells us that we specified a design with four factors, each with three levels. The next table reports the size of the saturated design, which is the number of parameters in the linear design, and suggests design sizes.

```
                    Choice of Fabric Softener

                        Design Summary

                     Number of
                     Levels       Frequency

                          3               4

                    Choice of Fabric Softener

              Saturated      = 9
              Full Factorial = 81
```

[§]Of course real studies would use real brands. Since we have not collected real data, we cannot use real brand names. We picked these silly names so no one would confuse our artificial data with real data.

```
Some Reasonable                    Cannot Be
   Design Sizes        Violations  Divided By

              9 *              0
             18 *              0
             12                6  9
             15                6  9
             10               10  3 9
             11               10  3 9
             13               10  3 9
             14               10  3 9
             16               10  3 9
             17               10  3 9

* - 100% Efficient Design can be made with the MktEx Macro.

                Choice of Fabric Softener

  n    Design                          Reference

  9           3 **  4                  Fractional-Factorial
 18    2 **  1  3 **  7                Orthogonal Array
 18           3 **  6  6 **  1         Orthogonal Array
```

The output from this macro tells us that the saturated design has nine runs and the full-factorial design has 81 runs. It also tells us that 9 and 18 are optimal design sizes with zero violations. The macro tells us that in nine runs, an orthogonal design with 4 three-level factors is available, and in 18 runs, two orthogonal and balanced designs are available: one with a two-level factor and 7 three-level factors, and one with 6 three-level factors and a six-level factor. There are zero violations with these designs because these sizes can be divided by 3 and $3 \times 3 = 9$. Twelve and 15 are also reported as potential design sizes, but each has 6 violations. Six times (the $4(4-1)/2 = 6$ pairs of the four threes) 12 and 15 cannot be divided by $3 \times 3 = 9$. Ideally, we would like to have a manageable number of choice sets for people to evaluate and a design that is both orthogonal and balanced. When violations are reported, orthogonal and balanced designs are not possible. While orthogonality and balance are not required, they are nice properties to have. With 4 three-level factors, the number of choice sets in all orthogonal and balanced designs must be divisible by $3 \times 3 = 9$.

Nine choice sets is a bit small. Furthermore, there are no error *df*. We set the number of choice sets to 18 since it is small enough for each person to see all choice sets, large enough to have reasonable error *df*, and an orthogonal and balanced design is available. It is important to remember however that the concept of number of parameters and error *df* discussed here applies to the linear design and not to the choice design.[¶] We could use the nine-run design for a discrete choice model and have error *df* in the choice model. If we were to instead use this design for a full-profile conjoint (not recommended), there would be no error *df*.

To make the code easier to modify for future use, the number of choice sets and alternatives are stored in macro variables and the prices are put into a format. Our design, in raw form, will have values for price of 1, 2, and 3. We will use a format to assign the actual prices: \$1.49, \$1.99, and \$2.49. The

[¶]See page 87 for an illustration of linear versus choice designs.

format also creates a price of $1.99 for missing, which will be used for the constant alternative.

```
%let n = 18;                    /* n choice sets                    */
%let m = 5;                     /* m alternative including constant */
%let mm1 = %eval(&m - 1);       /* m - 1                            */

proc format;                    /* create a format for the price    */
   value price 1 = '$1.49' 2 = '$1.99' 3 = '$2.49' . = '$1.99';
   run;
```

Designing the Choice Experiment

In the next steps, an efficient experimental design is created. We will use an autocall macro `%MktEx` to create most of our designs. (All of the autocall macros used in this book are documented starting on page 479.) When you invoke the `%MktEx` macro for a simple problem, you only need to specify the numbers of levels, and number of runs. The macro does the rest. Here is the `%MktEx` macro usage for this example:

```
%mktex(3 ** 4, n=&n)
```

The syntax '*n* ** *m*' means *m* factors each at *n* levels. This example has four factors, `x1` through `x4`, all with three levels. A design with 18 runs is requested. The `n=` option specifies the number of runs. These are all the options that are needed for a simple problem such as this one. However, throughout this book, random number seeds are explicitly specified with the `seed=` option so that the results will be reproducible.[‖] Here is the macro call with the random number seed specified:

```
%mktex(3 ** 4, n=&n, seed=17)

proc print; run;
```

Here are the results.

Choice of Fabric Softener

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	100.0000	100.0000	Tab
1	End	100.0000		

[‖] By specifying a random number seed, results should be reproducible within a SAS release for a particular operating system. However, due to machine differences, some results may not be exactly reproducible on other machines. For most orthogonal and balanced designs, the results should be reproducible. When computerized searches are done, it is likely that you will not get the same design as the one in the book, although you would expect the efficiency differences to be slight.

Choice of Fabric Softener

The OPTEX Procedure

Class Level Information

Class	Levels	-Values-
x1	3	1 2 3
x2	3	1 2 3
x3	3	1 2 3
x4	3	1 2 3

Choice of Fabric Softener

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	100.0000	100.0000	100.0000	0.7071

Choice of Fabric Softener

Obs	x1	x2	x3	x4
1	1	1	1	1
2	1	1	2	3
3	1	2	1	3
4	1	2	3	2
5	1	3	2	2
6	1	3	3	1
7	2	1	1	2
8	2	1	3	3
9	2	2	2	2
10	2	2	3	1
11	2	3	1	3
12	2	3	2	1
13	3	1	2	1
14	3	1	3	2
15	3	2	1	1
16	3	2	2	3
17	3	3	1	2
18	3	3	3	3

The macro found a perfect, orthogonal and balanced, 100% efficient design consisting of 4 three-level factors, `x1-x4`. The levels are the integers 1 to 3. For this problem, the macro generated the design directly. For other problems, the macro may have to use a computerized search. See page 142 for more information on how the `%MktEx` macro works.

Examining the Design

You should run basic checks on all designs, even orthogonal designs such as this one. You can use the `%MktEval` macro to display information about the design. The macro first prints a matrix of canonical correlations between the factors. We hope to see an identity matrix (a matrix of ones on the diagonal and zeros everywhere else) which means the design is orthogonal. Next, the macro prints all one-way frequencies for all attributes, all two-way frequencies, and all n-way frequencies (in this case four-way) frequencies. We hope to see equal or at least nearly equal one-way and two-way frequencies, and we want to see that each combination occurs only once.

```
%mkteval;
```

Choice of Fabric Softener
Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

	x1	x2	x3	x4
x1	1	0	0	0
x2	0	1	0	0
x3	0	0	1	0
x4	0	0	0	1

Choice of Fabric Softener
Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316

Frequencies

x1	6 6 6
x2	6 6 6
x3	6 6 6
x4	6 6 6
x1 x2	2 2 2 2 2 2 2 2 2
x1 x3	2 2 2 2 2 2 2 2 2
x1 x4	2 2 2 2 2 2 2 2 2
x2 x3	2 2 2 2 2 2 2 2 2
x2 x4	2 2 2 2 2 2 2 2 2
x3 x4	2 2 2 2 2 2 2 2 2
N-Way	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

A *canonical correlation* is the maximum correlation between linear combinations of the coded factors (see page 90). All zeros off the diagonal show that this design is orthogonal for main effects. If any off-diagonal canonical correlations had been greater than 0.316 ($r^2 > 0.1$), the macro would have listed them in a separate table. The last title line tells you that none of them was this large. For nonorthogonal designs and designs with interactions, the canonical-correlation matrix is not a substitute for looking at the variance matrix (with `examine=v`, discussed on pages 146, 191, and 556). It just provides a quick and more-compact picture of the correlations between the factors. The variance matrix is sensitive to the actual model specified and the actual coding. The canonical-correlation matrix just tells you if there is some correlation between the main effects. In this case, there are no correlations.

The equal one-way frequencies show you that this design is balanced. The equal two-way frequencies show you that this design is orthogonal. The n-way frequencies, all equal to one, show there are no duplicate profiles. This is a perfect design for a main-effects model.

You should always check the n-way frequencies to ensure that you do not have duplicates. For this situation for example, a 100% efficient design exists where each run appears twice. It consists of two copies of the fractional-factorial design 3^4 in 9 runs. When you get duplicates, specify `options=nodups` in the `%MktEx` macro, or sometimes you can just change the random number seed. Most designs will not have duplicates, so it is better to specify `options=nodups` only after you have found a design with duplicates. The no-duplicates constraint greatly slows down the algorithm.

The `%MktEval` macro produces a very compact summary of the design, hence some information, for example the levels to which the frequencies correspond, is not shown. You can use the `print=freqs` option to get a less compact and more detailed display.

```
%mkteval(data=design, print=freqs);
```

Here are some of the results.

Choice of Fabric Softener
Frequencies
There are 0 Canonical Correlations Greater Than 0.316

Effects	Frequency	x1	x2	x3	x4
x1	6	1	.	.	.
	6	2	.	.	.
	6	3	.	.	.
x2	6	.	1	.	.
	6	.	2	.	.
	6	.	3	.	.
.					
.					
.					

```
x1 x2          2          1     1     .     .
               2          1     2     .     .
               2          1     3     .     .
               2          2     1     .     .
               2          2     2     .     .
               2          2     3     .     .
               2          3     1     .     .
               2          3     2     .     .
               2          3     3     .     .
.
.
.
x3 x4          2          .     .     1     1
               2          .     .     1     2
               2          .     .     1     3
               2          .     .     2     1
               2          .     .     2     2
               2          .     .     2     3
               2          .     .     3     1
               2          .     .     3     2
               2          .     .     3     3
N-Way          1          1     1     1     1
               1          1     1     2     3
               1          1     2     1     3
               1          1     2     3     2
               1          1     3     2     2
               1          1     3     3     1
               1          2     1     1     2
               1          2     1     3     3
               1          2     2     2     2
               1          2     2     3     1
               1          2     3     1     3
               1          2     3     2     1
               1          3     1     2     1
               1          3     1     3     2
               1          3     2     1     1
               1          3     2     2     3
               1          3     3     1     2
               1          3     3     3     3
```

Randomizing the Design, Postprocessing

The design we just looked at and examined was in the default output data set DESIGN. The DESIGN data set is sorted and often has a first row consisting entirely of ones. For these reasons, you should actually use the *randomized* design. In the randomized design, the choice sets are presented in a

random order and the levels have been randomly reassigned. Neither of these operations affects the design efficiency, balance, or orthogonality. The macro automatically randomizes the design and stores the results in a data set called RANDOMIZED. The next steps assign formats and labels, and store the results in a SAS data set SASUSER.DES. It is important to store the design in a permanent SAS data set or in some other permanent form so that it will be available for analysis after the data are collected.

```
data sasuser.des;
   set randomized;
   format x1-x&mm1 price.;
   label x1 = 'Sploosh' x2 = 'Plumbbob' x3 = 'Platter' x4 = 'Moosey';
   run;
```

This is the final design.

```
proc print data=sasuser.des label; /* print final design */
   title2 'Efficient Design';
   run;

title2;
```

Choice of Fabric Softener
Efficient Design

Obs	Sploosh	Plumbbob	Platter	Moosey
1	$1.99	$1.49	$1.99	$2.49
2	$2.49	$2.49	$1.49	$1.99
3	$1.49	$1.99	$2.49	$1.49
4	$2.49	$1.99	$2.49	$1.99
5	$1.49	$2.49	$2.49	$2.49
6	$1.49	$1.99	$1.49	$2.49
7	$2.49	$1.49	$1.49	$1.49
8	$2.49	$1.49	$2.49	$2.49
9	$1.99	$2.49	$2.49	$1.49
10	$1.49	$1.49	$1.99	$1.49
11	$1.99	$2.49	$1.49	$2.49
12	$1.49	$1.49	$1.49	$1.99
13	$1.99	$1.49	$2.49	$1.99
14	$1.49	$2.49	$1.99	$1.99
15	$2.49	$2.49	$1.99	$1.49
16	$1.99	$1.99	$1.99	$1.99
17	$1.99	$1.99	$1.49	$1.49
18	$2.49	$1.99	$1.99	$2.49

Generating the Questionnaire

A questionnaire based on the design is printed using the DATA step. The statement `array brands[&m] $ _temporary_ ('Sploosh' 'Plumbbob' 'Platter' 'Moosey' 'Another')` creates a constant array so that `brands[1]` accesses the string `'Sploosh'`, `brands[2]` accesses the string `'Plumbbob'`, and so on. The `_temporary_` specification means that no output data set variables are created for this array. The `linesleft=` specification in the `file` statement creates the variable `ll`, which contains the number of lines left on a page. This ensures that each choice set is not split over two pages.

```
options ls=80 ps=60 nonumber nodate;
title;

data _null_;                        /* print questionnaire */
   array brands[&m] $ _temporary_ ('Sploosh' 'Plumbbob' 'Platter'
                                   'Moosey' 'Another');
   array x[&m] x1-x&m;
   file print linesleft=ll;
   set sasuser.des;

   x&m = 2;                        /* constant alternative */
   format x&m price.;

   if _n_ = 1 or ll < 12 then do;
      put _page_;
      put @60 'Subject: _________' //;
      end;
   put _n_ 2. ') Circle your choice of '
       'one of the following fabric softeners:' /;
   do brnds = 1 to &m;
      put '    ' brnds 1. ') ' brands[brnds] 'brand at '
          x[brnds] +(-1) '.' /;
      end;
   run;
```

In the interest of space, only the first two choice sets are printed. The questionnaire is printed, copied, and the data are collected.

```
                                                    Subject: _________

1) Circle your choice of one of the following fabric softeners:

   1) Sploosh brand at $1.99.

   2) Plumbbob brand at $1.49.

   3) Platter brand at $1.99.

   4) Moosey brand at $2.49.

   5) Another brand at $1.99.

2) Circle your choice of one of the following fabric softeners:

   1) Sploosh brand at $2.49.

   2) Plumbbob brand at $2.49.

   3) Platter brand at $1.49.

   4) Moosey brand at $1.99.

   5) Another brand at $1.99.
```

In practice, data collection may be much more elaborate than this. It may involve art work or photographs, and the choice sets may be presented and the data may be collected through personal interview or over the web. However the choice sets are presented and the data are collected, the essential ingredients remain the same. Subjects are shown sets of alternatives and are asked to make a choice, then they go on to the next set.

Entering the Data

The data consist of a subject number followed by 18 integers in the range 1 to 5. These are the alternatives that were chosen for each choice set. For example, the first subject chose alternative 2 (*Plumbbob* brand at $1.49) in the first choice set, alternative 3 (*Platter* brand at $1.49) in the second choice set, and so on. In the interest of space, data from three subjects appear on one line.

```
title 'Choice of Fabric Softener';

data results;                     /* read choice data set */
   input Subj (choose1-choose&n) (1.) @@;
   datalines;
  1 234513324233214433  2 334213324433233335  3 333313323333333333
  4 334413324434414453  5 235413324233514333  6 234433424433214233
  7 234413324433234333  8 234413424433444332  9 334413324353233333
 10 235233325233234333 11 334213324433233332 12 234313324433214353
 13 534313334333234333 14 234411424433214443 15 334313325433434335
 16 334433325333335333 17 534313424453214433 18 334415524433444543
 19 334313325433234433 20 231413324233215533 21 234353534433514333
 22 234313354233214533 23 334333333333234333 24 535212324232214243
 25 254313324433234333 26 234415525233214333 27 234353524233254333
 28 234333334333434333 29 334413524335253333 30 334453324533334333
 31 354313324333233433 32 254313324233234333 33 234413424353413335
 34 334413324433214333 35 234553324453214345 36 234233524433244333
 37 234413324433514343 38 254333324433254533 39 334353324333233533
 40 334413324433434333 41 334433424433444443 42 234413324433214333
 43 334413454433434343 44 234413324233214333 45 234453524432214444
 46 234413425433514433 47 544413524433244333 48 335453324433233353
 49 234513324133234433 50 234413324333234333
;
```

Processing the Data

Next, we prepare the experimental design for analysis. Our design, which we stored in a permanent SAS data set SASUSER.DES, is arranged with one row per choice set. We call this the *linear design* (see page 87). The linear design, which came directly from the `%MktEx` macro, is conveniently arrayed for generating the questionnaire, however it is not in the right form for analysis. For analysis, we need a *choice design* with one row for each alternative of each choice set. We will use the macro `%MktRoll` to "roll out" the linear design into the choice design, which is in proper form for analysis. First, we must create a data set that describes how the design will be processed. We call this data set the *design key*.

In this example, we want a choice design with two factors, `Brand` and `Price`. `Brand` has levels `'Sploosh'`, `'Plumbbob'`, `'Platter'`, `'Moosey'`, and `'Another'`. `Price` has levels $1.49, $1.99, and $2.49. `Brand` and `Price` are created by different processes. The `Price` factor will be constructed from the factors of the linear design matrix. In contrast, there is no `Brand` factor in the linear design. Each brand is a bin into which its factors are collected. The variable `Brand` will be named on the `alt=` option of the `%MktRoll` macro as the alternative variable, so its values will be read directly out of the

KEY data set. `Price` will not be named on the `alt=` macro option, so its values in the KEY data set are variable names from the linear design data set. The values of `Price` in the final choice design will be read from the named variables in the linear design data set. The `Price` factor in the choice design is created from the four linear design factors (`x1` for *Sploosh*, `x2` for *Plumbbob*, `x3` for *Platter*, `x4` for *Moosey*, and no attribute for *Another*, the constant alternative). Here is how the KEY data set is created. The `Brand` factor levels and the `Price` linear design factors are stored in the KEY data set.

```
title2 'Key Data Set';

data key;
   input Brand $ Price $;
   datalines;
Sploosh   x1
Plumbbob  x2
Platter   x3
Moosey    x4
Another   .
;

proc print; run;

title2;
```

Choice of Fabric Softener
Key Data Set

Obs	Brand	Price
1	Sploosh	x1
2	Plumbbob	x2
3	Platter	x3
4	Moosey	x4
5	Another	

Note that the value of `Price` for alternative *Another* in the KEY data set is blank (character missing). The period in the in-stream data set is simply a placeholder, used with list input to read both character and numeric missing data. A period is not stored with the data. Next, we use the `%MktRoll` macro to process the design.

```
%mktroll(design=sasuser.des, key=key, alt=brand, out=rolled)
```

The `%MktRoll` step processes the `design=sasuser.des` linear design data set using the rules specified in the `key=key` data set, naming the `alt=brand` variable as the alternative name variable, and creating an output SAS data set called ROLLED, which contains the choice design. The input `design=sasuser.des` data set has 18 observations, one per choice set, and the output `out=rolled` data set has $5 \times 18 = 90$ observations, one for each alternative of each choice set. Here are the first three observations of the linear design data set.

```
title2 'Linear Design (First 3 Sets)';

proc print data=sasuser.des(obs=3); run;

title2;
```

Choice of Fabric Softener
Linear Design (First 3 Sets)

Obs	x1	x2	x3	x4
1	$1.99	$1.49	$1.99	$2.49
2	$2.49	$2.49	$1.49	$1.99
3	$1.49	$1.99	$2.49	$1.49

These observations define the first three choice sets. Here are those same observations, arrayed for analysis in the choice design data set.

```
title2 'Choice Design (First 3 Sets)';

proc print data=rolled(obs=15); format price price.; run;

title2;
```

Choice of Fabric Softener
Choice Design (First 3 Sets)

Obs	Set	Brand	Price
1	1	Sploosh	$1.99
2	1	Plumbbob	$1.49
3	1	Platter	$1.99
4	1	Moosey	$2.49
5	1	Another	$1.99
6	2	Sploosh	$2.49
7	2	Plumbbob	$2.49
8	2	Platter	$1.49
9	2	Moosey	$1.99
10	2	Another	$1.99
11	3	Sploosh	$1.49
12	3	Plumbbob	$1.99
13	3	Platter	$2.49
14	3	Moosey	$1.49
15	3	Another	$1.99

The choice design data set has a choice set variable `Set`, an alternative name variable `Brand`, and a price variable `Price`. The prices come from the linear design, and the price for *Another* is a constant $1.99. Recall that the prices are assigned by the following format.

```
proc format;                    /* create a format for the price */
   value price 1 = '$1.49' 2 = '$1.99' 3 = '$2.49' . = '$1.99';
   run;
```

The next step merges the choice data with the choice design using the `%MktMerge` macro.

```
%mktmerge(design=rolled, data=results, out=res2,
          nsets=&n, nalts=&m, setvars=choose1-choose&n)
```

This step reads the `design=rolled` choice design and the `data=results` data set and creates the `out=res2` output data set. The data are from an experiment with `nsets=&n` choice sets, `nalts=&m` alternatives, with variables `setvars=choose1-choose&n` containing the numbers of the chosen alternatives. Here are the first 15 observations.

```
title2 'Choice Design and Data (First 3 Sets)';

proc print data=res2(obs=15); run;

title2;
```

Choice of Fabric Softener
Choice Design and Data (First 3 Sets)

Obs	Subj	Set	Brand	Price	c
1	1	1	Sploosh	2	2
2	1	1	Plumbbob	1	1
3	1	1	Platter	2	2
4	1	1	Moosey	3	2
5	1	1	Another	.	2
6	1	2	Sploosh	3	2
7	1	2	Plumbbob	3	2
8	1	2	Platter	1	1
9	1	2	Moosey	2	2
10	1	2	Another	.	2
11	1	3	Sploosh	1	2
12	1	3	Plumbbob	2	2
13	1	3	Platter	3	2
14	1	3	Moosey	1	1
15	1	3	Another	.	2

The data set contains the subject ID variable `Subj` from the `data=results` data set, the `Set`, `Brand`, and `Price` variables from the `design=rolled` data set, and the variable `c`, which indicates which alternative was chosen. The variable `c` indicates the chosen alternatives: 1 for first choice and 2 for second or subsequent choice. This subject chose the second alternative, *Plumbbob*, in the first choice set, *Platter* in the second, and *Moosey* in the third. This data set has 4500 observations: 50 subjects times 18 choice sets times 5 alternatives.

Since we did not specify a format, we see in the design the raw design values for `Price`: 1, 2, 3 and missing for the constant alternative. If we were going to treat `Price` as a categorical variable for analysis, this would be fine. We would simply assign our price format to `Price` and designate it as a `class` variable. However, in this analysis we are going to treat price as quantitative and use the actual prices in the analysis. Hence, we must convert our design values of 1, 2, 3, and . to 1.49, 1.99, 2.49, and 1.99. We cannot do this by simply assigning a format. Formats create character strings that are printed in place of the original value. We need to convert a numeric variable from one set of numbers to another. We could use `if` and assignment statements. We could also use the `%MktLab` macro, which is used in later examples. However, instead we will use the `put` function to write the formatted value into a character string, then we read it back using a dollar format and the `input` function. For example, the expression `put(price, price.)` converts a number, say 2, into a string (in this case '$1.99'), then the `input` function reads the string and converts it to a numeric 1.99. This step also assigns a label to the variable `Price`.

```
data res3;   /* Create a numeric actual price */
   set res2;
   price = input(put(price, price.), dollar5.);
   label price = 'Price';
   run;
```

Binary Coding

One more thing must be done to these data before they can be analyzed. The factors must be coded. In this example, we use a *binary* or zero-one coding for the brand effect. This can be done with PROC TRANSREG.

```
proc transreg design=5000 data=res3 nozeroconstant norestoremissing;
   model class(brand / zero=none order=data)
         identity(price) / lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   id subj set c;
   run;
```

The `design` option specifies that no model is fit; the procedure is just being used to code a design. When `design` is specified, dependent variables are not required. Optionally, `design` can be followed by "`=`n" where n is the number of observations to process at one time. By default, PROC TRANSREG codes all observations in one big group. For very large data sets, this can consume large amounts of memory and time. Processing blocks of smaller numbers of observations is more efficient. The option `design=5000` processes observations in blocks of 5000. For smaller computers, try something like `design=1000`.

The `nozeroconstant` and `norestoremissing` options are not necessary for this example but are included here because sometimes they are very helpful in coding choice models. The `nozeroconstant` option specifies that if the coding creates a constant variable, it should not be zeroed. The `nozeroconstant` option should always be specified when you specify `design=`n because the last group of observations may be small and may contain constant variables. The `nozeroconstant` option is also important if you do something like coding `by subj set` because sometimes an attribute is constant within a choice set. The `norestoremissing` option specifies that missing values should not be restored when the `out=` data set is created. By default, the coded `class` variable contains a row of missing values for observations in which the `class` variable is missing. When you specify the `norestoremissing` option, these observations contain a row of zeros instead. This option is useful when there is a constant alternative

indicated by missing values. Both of these options, like almost all options in PROC TRANSREG, can be abbreviated to three characters (`noz` and `nor`).

The `model` statement names the variables to code and provides information about how they should be coded. The specification `class(brand / ...)` specifies that the variable `Brand` is a classification variable and requests a binary coding. The `zero=none` option creates binary variables for all categories. In contrast, by default, a binary variable is not created for the last category − the parameter for the last category is a structural zero. The `zero=none` option is used when there are no structural zeros or when you want to see the structural zeros in the multinomial logit parameter estimates table. The `order=data` option sorts the levels into the order they were first encountered in the data set. The specification `identity(price)` specifies that `Price` is a quantitative factor that should be analyzed as is (not expanded into indicator variables).

The `lprefix=0` option specifies that when labels are created for the binary variables, zero characters of the original variable name should be used as a prefix. This means that the labels are created only from the level values. For example, `'Sploosh'` and `'Plumbbob'` are created as labels not `'Brand Sploosh'` and `'Brand Plumbbob'`.

An `output` statement names the output data set and drops variables that are not needed. These variables do not have to be dropped. However, since they are variable names that are often found in special data set types, PROC PHREG prints warnings when it finds them. Dropping the variables prevents the warnings. Finally, the `id` statement names the additional variables that we want copied from the input to the output data set. The next steps print the first three coded choice sets.

```
proc print data=coded(obs=15) label;
   title2 'First 15 Observations of Analysis Data Set';
   id subj set c;
   run;

title2;
```

Choice of Fabric Softener
First 15 Observations of Analysis Data Set

Subj	Set	c	Sploosh	Plumbbob	Platter	Moosey	Another	Price	Brand
1	1	2	1	0	0	0	0	1.99	Sploosh
1	1	1	0	1	0	0	0	1.49	Plumbbob
1	1	2	0	0	1	0	0	1.99	Platter
1	1	2	0	0	0	1	0	2.49	Moosey
1	1	2	0	0	0	0	1	1.99	Another
1	2	2	1	0	0	0	0	2.49	Sploosh
1	2	2	0	1	0	0	0	2.49	Plumbbob
1	2	1	0	0	1	0	0	1.49	Platter
1	2	2	0	0	0	1	0	1.99	Moosey
1	2	2	0	0	0	0	1	1.99	Another

```
1    3    2    1         0         0         0         0       1.49  Sploosh
1    3    2    0         1         0         0         0       1.99  Plumbbob
1    3    2    0         0         1         0         0       2.49  Platter
1    3    1    0         0         0         1         0       1.49  Moosey
1    3    2    0         0         0         0         1       1.99  Another
```

Fitting the Multinomial Logit Model

The next step fits the discrete choice, multinomial logit model.

```
proc phreg data=coded outest=betas brief;
   title2 'Discrete Choice Model';
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;

title2;
```

The `brief` option requests a brief summary for the strata. As with the candy example, `c*c(2)` designates the chosen and unchosen alternatives in the `model` statement. We specify the `&_trgind` macro variable for the `model` statement independent variable list. PROC TRANSREG automatically creates this macro variable. It contains the list of coded independent variables generated by the procedure. This is so you do not have to figure out what names TRANSREG created and specify them. In this case, PROC TRANSREG sets `&_trgind` to contain the following list.

```
BrandSploosh BrandPlumbbob BrandPlatter BrandMoosey BrandAnother Price
```

The `ties=breslow` option specifies a PROC PHREG model that has the same likelihood as the multinomial logit model for discrete choice. The `strata` statement specifies that the combinations of `Set` and `Subj` indicate the choice sets. This data set has 4500 observations consisting of $18 \times 50 = 900$ strata and five observations per stratum.

Each subject rated 18 choice sets, but the multinomial logit model assumes each stratum is independent. That is, the multinomial logit model assumes each person makes only one choice. The option of collecting only one datum from each subject is too expensive to consider for many problems, so multiple choices are collected from each subject, and the repeated measures aspect of the problem is ignored. This practice is typical, and it usually works well.

Multinomial Logit Model Results

The output is shown next. (Recall that we used `%phchoice(on)` on page 95 to customize the output from PROC PHREG.)

Choice of Fabric Softener
Discrete Choice Model

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	4500
Number of Observations Used	4500

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	900	5	1	4

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2896.988	1391.313
AIC	2896.988	1401.313
SBC	2896.988	1425.325

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1505.6751	5	<.0001
Score	1345.8861	5	<.0001
Wald	658.0940	5	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Sploosh	1	-1.45430	0.21487	45.8095	<.0001
Plumbbob	1	-0.56238	0.18579	9.1622	0.0025
Platter	1	1.83425	0.14640	156.9854	<.0001
Moosey	1	0.51937	0.16029	10.4990	0.0012
Another	0	0	.	.	.
Price	1	-4.79503	0.21735	486.7239	<.0001

The procedure output begins with information about the data set, variables, options, and number of observations read. This is followed by information about the 900 strata. Since the `brief` option was specified, this table contains one row for each stratum pattern. In contrast, the default table would have 900 rows, one for each choice set and subject combination. Each subject and choice set combination consists of a total of five observations, one that was chosen and four that were not chosen. This pattern was observed 900 times. This table provides a check on data entry. Unless we have an availability or allocation study (page 271) or a nonconstant number of alternatives in different choice sets, we would expect to see one pattern of results where one of the m alternatives was chosen for each choice set. If you do not observe this for a study like this, there was probably a mistake in the data entry or processing.

The most to least preferred brands are: *Platter*, *Moosey*, *Another*, *Plumbbob*, and *Sploosh*. Increases in price have a negative utility. For example, the predicted utility of *Platter* brand at \$1.99 is $\mathbf{x}_i\boldsymbol{\beta}$ which is $(0 \quad 0 \quad 1 \quad 0 \quad 0 \quad \$1.99) \quad (-1.45 \quad -0.56 \quad 1.83 \quad 0.52 \quad 0 \quad -4.80)' = 1.83 + 1.99 \times -4.80 = -7.72$ Since `Price` was analyzed as a quantitative factor, we can see for example that the utility of *Platter* at \$1.89, which was not in any choice set, is $1.83 + 1.89 \times -4.80 = -7.24$, which is a $\$0.10 \times 4.80 = 0.48$ increase in utility.

Probability of Choice

These next steps compute the expected probability that each alternative is chosen within each choice set. This code could easily be modified to compute expected market share for hypothetical marketplaces that do not directly correspond to the choice sets. Note however, that a term like "expected market share," while widely used, is a misnomer. Without purchase volume data, it is unlikely that these numbers would mirror true market share.

First, PROC SCORE is used to compute the predicted utility for each alternative.

```
proc score data=coded(where=(subj=1) drop=c)
           score=betas type=parms out=p;
   var &_trgind;
   run;
```

The data set to be scored is named with the `data=` option, and the coefficients are specified in the option `score=beta`. Note that we only need to read all of the choice sets once, since the parameter estimates were computed in an aggregate analysis. This is why we specified `where=(subj=1)`. We do not need $\mathbf{x}_j\hat{\boldsymbol{\beta}}$ for each of the different subjects. We dropped the variable `c` from the CODED data set since this

name will be used by PROC SCORE for the results ($\mathbf{x}_j\hat{\boldsymbol{\beta}}$). The option `type=parms` specifies that the `score=` data set contains the parameters in `_TYPE_ = 'PARMS'` observations. The output data set with the predicted utilities is named P. Scoring is based on the coded variables from PROC TRANSREG, whose names are contained in the macro variable `&_trgind`. The next step exponentiates $\mathbf{x}_j\hat{\boldsymbol{\beta}}$.

```
data p2;
   set p;
   p = exp(c);
   run;
```

Next, $\exp(\mathbf{x}_j\hat{\boldsymbol{\beta}})$ is summed for each choice set.

```
proc means data=p2 noprint;
   output out=s sum(p) = sp;
   by set;
   run;
```

Finally, each $\mathbf{x}_j\hat{\boldsymbol{\beta}}$ is divided by $\sum_{j=1}^{m}\mathbf{x}_j\hat{\boldsymbol{\beta}}$.

```
data p;
   merge p2 s(keep=set sp);
   by set;
   p = p / sp;
   keep brand set price p;
   run;
```

Here are the results for the first three choice sets.

```
proc print data=p(obs=15);
   title2 'Choice Probabilities for the First 3 Choice Sets';
   run;

title2;
```

Choice of Fabric Softener
Choice Probabilities for the First 3 Choice Sets

Obs	Price	Brand	Set	p
1	1.99	Sploosh	1	0.01679
2	1.49	Plumbbob	1	0.45037
3	1.99	Platter	1	0.44998
4	2.49	Moosey	1	0.01099
5	1.99	Another	1	0.07188
6	2.49	Sploosh	2	0.00030
7	2.49	Plumbbob	2	0.00072
8	1.49	Platter	2	0.96153
9	1.99	Moosey	2	0.02348
10	1.99	Another	2	0.01397

11	1.49	Sploosh	3	0.11074
12	1.99	Plumbbob	3	0.02457
13	2.49	Platter	3	0.02455
14	1.49	Moosey	3	0.79702
15	1.99	Another	3	0.04312

Custom Questionnaires

In this part of the example, a custom questionnaire is printed for each person. Previously, each subject saw the same questionnaire, with the same choice sets, each containing the same alternatives, with everything in the same order. In this example, the order of the choice sets and all alternatives within choice sets are randomized for each subject. Randomizing avoids any systematic effects due to the order of the alternatives and choice sets. The constant alternative is always printed last. If you have no interest in custom questionnaires, you can skip ahead to page 134.

First, the macro variable `&forms` is created. It contains the number of separate questionnaires (or forms or subjects, in this case 50). We can use the `%MktEx` macro to create a data set with one observation for each alternative of each choice set for each person. The specification `%mktex(&forms &n &mm1, n=&forms * &n * &mm1)` is `%mktex(50 18 4, n=50 * 18 * 4)` and creates a $50 \times 18 \times 4$ full-factorial design. Note that the `n=` specification allows expressions. The macro `%MktLab` is then used to assign the variable names `Form`, `Set`, and `Alt` instead of the default `x1 - x3`. The data set is sorted by `Form`. Within `Form`, the choice sets are sorted into a random order, and within choice set, the alternatives are sorted into a random order. The 72 observations for each choice set contain 18 blocks of 4 observations − one block per choice set in a random order and the 4 alternatives within each choice set, again in a random order. Note that we store these in a permanent SAS data set so they will be available after the data are collected.

```
%let forms = 50;
title2 'Create 50 Custom Questionnaires';

*---Make the design---;
%mktex(&forms &n &mm1, n=&forms * &n * &mm1)

*---Assign Factor Names---;
%mktlab(data=design, vars=Form Set Alt)

*---Set up for Random Ordering---;
data sasuser.orders;
   set final;
   by form set;
   retain r1;
   if first.set then r1 = uniform(17);
   r2 = uniform(17);
   run;

*---Random Sort---;
proc sort out=sasuser.orders(drop=r:); by form r1 r2; run;

proc print data=sasuser.orders(obs=16); run;
```

The first 16 observations in this data set are shown next.

Choice of Fabric Softener
Create 50 Custom Questionnaires

Obs	Form	Set	Alt
1	1	4	3
2	1	4	1
3	1	4	2
4	1	4	4
5	1	8	2
6	1	8	3
7	1	8	1
8	1	8	4
9	1	16	1
10	1	16	2
11	1	16	3
12	1	16	4
13	1	1	3
14	1	1	1
15	1	1	4
16	1	1	2

The data set is transposed, so the resulting data set contains $50 \times 18 = 900$ observations, one per subject per choice set. The alternatives are in the variables `Col1-Col4`. The first 18 observations, which contain the ordering of the choice sets for the first subject, are shown next.

```
proc transpose data=sasuser.orders out=sasuser.orders(drop=_name_);
   by form notsorted set;
   run;
proc print data=sasuser.orders(obs=18);
   run;
```

Choice of Fabric Softener
Create 50 Custom Questionnaires

Obs	Form	Set	COL1	COL2	COL3	COL4
1	1	4	3	1	2	4
2	1	8	2	3	1	4
3	1	16	1	2	3	4
4	1	1	3	1	4	2
5	1	6	2	4	1	3
6	1	7	4	1	3	2
7	1	12	3	2	1	4
8	1	2	2	4	1	3
9	1	17	3	4	1	2

10	1	15	4	2	3	1
11	1	14	1	2	3	4
12	1	10	2	4	3	1
13	1	5	1	4	2	3
14	1	9	2	4	1	3
15	1	13	3	2	1	4
16	1	3	3	4	2	1
17	1	18	4	2	1	3
18	1	11	3	1	4	2

The following DATA step prints the 50 custom questionnaires.

```
options ls=80 ps=60 nodate nonumber;
title;

data _null_;
   array brands[&mm1] $ _temporary_
                  ('Sploosh' 'Plumbbob' 'Platter' 'Moosey');
   array x[&mm1] x1-x&mm1;
   array c[&mm1] col1-col&mm1;
   format x1-x&mm1 price.;
   file print linesleft=ll;
   do frms = 1 to &forms;
      do choice = 1 to &n;
         if choice = 1 or ll < 12 then do;
            put _page_;
            put @60 'Subject: ' frms //;
            end;
         put choice 2. ') Circle your choice of '
             'one of the following fabric softeners:' /;
         set sasuser.orders;
         set sasuser.des point=set;
         do brnds = 1 to &mm1;
            put '    ' brnds 1. ') ' brands[c[brnds]] 'brand at '
                x[c[brnds]] +(-1) '.' /;
            end;
         put '    5) Another brand at $1.99.' /;
         end;
      end;
   stop;
   run;
```

The loop `do frms = 1 to &forms` creates the 50 questionnaires. The loop `do choice = 1 to &n` creates the alternatives within each choice set. On the first choice set and when there is not enough room for the next choice set, we skip to a new page (`put _page_`) and print the subject (forms) number. The data set SASUSER.ORDERS is read and the `Set` variable is used to read the relevant observation from SASUSER.DES using the `point=` option in the `set` statement. The order of the alternatives is in the `c` array and variables `col1-col&mm1` from the SASUSER.ORDERS data set. In the first observation of SASUSER.ORDERS, `Set=4`, `Col1=3`, `Col2=1`, `Col3=2`, and `Col4=4`. The first brand, is `c[brnds] = c[1] = col1 = 3`, so `brands[c[brnds]] = brands[c[1]] = brands[3]`

= 'Platter', and the price, from observation Set=4 of SASUSER.DES, is x[c[brnds]] = x[3] = $2.49. The second brand, is c[brnds] = c[2] = col2 = 1, so brands[c[brnds]] = brands[c[2]] = brands[1] = 'Sploosh', and the price, from observation Set=4 of SASUSER.DES, is x[c[brnds]] = x[1] = $2.49.

In the interest of space, only the first two choice sets are printed. Note that the subject number is printed on the form. This information is needed to restore all data to the original order.

```
                                                    Subject: 1


   1) Circle your choice of one of the following fabric softeners:

      1) Platter brand at $2.49.

      2) Sploosh brand at $2.49.

      3) Plumbbob brand at $1.99.

      4) Moosey brand at $1.99.

      5) Another brand at $1.99.

   2) Circle your choice of one of the following fabric softeners:

      1) Plumbbob brand at $1.49.

      2) Platter brand at $2.49.

      3) Sploosh brand at $2.49.

      4) Moosey brand at $2.49.

      5) Another brand at $1.99.
```

Processing the Data for Custom Questionnaires

Here are the data. (Actually, these are the data that would have been collected if the same people as in the previous situation made the same choices, without error and uninfluenced by order effects.) Before these data are analyzed, the original order must be restored.

```
title 'Choice of Fabric Softener';

data results;                    /* read choice data set */
   input Subj (choose1-choose&n) (1.) @@;
   datalines;
  1 514443141111122241  2 532231422321124311  3 224113221414144231
  4 421413132322544334  5 123244311522341532  6 233214431443133321
  7 313214224433422312  8 244132344421114412  9 115422242443114224
 10 521432445311432321 11 331243123313423222 12 431234434313123245
 13 313313411243435334 14 443443434342333114 15 344423243531141345
 16 425444321454433414 17 234241535433442432 18 325222352241521311
 19 134113342433542213 20 315321253334442412 21 513453254212232224
 22 312314223544113125 23 433443444431143432 24 353234433451334321
 25 322411331352444431 26 243135451131141445 27 553253331223333111
 28 233443212333231424 29 442454334541231533 30 223133332135132542
 31 422222434323513242 32 434144312354323423 33 414212243433154445
 34 133114112312232331 35 425311432255122522 36 142334254232324432
 37 511321124112341323 38 542353113412342543 39 221542432333512212
 40 124232222244211211 41 243411341423133213 42 214122324311222114
 43 323251321313324342 44 114344444144214422 45 131252434312452121
 46 314215325411422113 47 133254134453111432 48 231335341342551314
 49 125124122444221224 50 444112131412134341
;
```

The data set is transposed, and the original order is restored.

```
proc transpose data=results   /* create one obs per choice set */
               out=res2(rename=(col1=choose) drop=_name_);
   by subj;
   run;

data res3(keep=subj set choose);
   array c[&mm1] col1-col&mm1;
   merge sasuser.orders res2;
   if choose < 5 then choose = c[choose];
   run;


proc sort; by subj set; run;
```

The actual choice number, stored in `Choose`, indexes the alternative numbers from SASUSER.ORDERS to restore the original alternative orders. For example, for the first subject, the first choice was 5, which is the *Another* constant alternative. Since the first subject saw the fourth choice set first, the fourth data value for the first subject in the processed data set will have a value of 5. The choice in the second choice set for the first subject was 1, and the first alternative the subject saw was *Plumbbob*. The data set SASUSER.ORDERS shows in the second observation that this choice of 1 corresponds to the second (original) alternative (in the first column variable, `Col1 = 2`) of choice set `Set= 8`. In

the original ordering, *Plumbbob* is the second alternative. Hence the eighth data value in the processed data set will have a value of 2. This DATA step writes out the data after the original order has been restored. It matches the data on page 118.

```
data _null_;
   set res3;
   by subj;
   if first.subj then do;
      if mod(subj, 3) eq 1 then put;
      put subj 4. +1 @@;
      end;
   put choose 1. @@;
   run;
```

```
 1 234513324233214433   2 334213324433233335   3 333313323333333333
 4 334413324434414453   5 235413324233514333   6 234433424433214233
 7 234413324433234333   8 234413424433444332   9 334413324353233333
10 235233325233234333  11 334213324433233332  12 234313324433214353
13 534313334333234333  14 234411424433214443  15 334313325433434335
16 334433325333335333  17 534313424453214433  18 334415524433444543
19 334313325433234433  20 231413324233215533  21 234353534433514333
22 234313354233214533  23 334333333333234333  24 535212324232214243
25 254313324433234333  26 234415525233214333  27 234353524233254333
28 234333334333434333  29 334413524335253333  30 334453324533334333
31 354313324333233433  32 254313324233234333  33 234413424353413335
34 334413324433214333  35 234553324453214345  36 234233524433244333
37 234413324433514343  38 254333324433254533  39 334353324333233533
40 334413324433434333  41 334433424433444443  42 234413324433214333
43 334413454433434343  44 234413324233214333  45 234453524432214444
46 234413425433514433  47 544413524433244333  48 335453324433233353
49 234513324133234433  50 234413324333234333
```

The data can be combined with the design and analyzed as in the previous example.

Vacation Example

This example illustrates the design and analysis for a larger choice experiment. We will discuss designing a choice experiment, evaluating the design, generating the questionnaire, processing the data, binary coding, generic attributes, quantitative price effects, quadratic price effects, effects coding, alternative-specific effects, analysis, and interpretation of the results. In this example, a researcher is interested in studying choice of vacation destinations. There are five destinations (alternatives) of interest: Hawaii, Alaska, Mexico, California, and Maine. Here are two summaries of the design, one with factors first grouped by attribute and one grouped by destination.

Factor	Destination	Attribute	Levels
X1	Hawaii	Accommodations	Cabin, Bed & Breakfast, Hotel
X2	Alaska	Accommodations	Cabin, Bed & Breakfast, Hotel
X3	Mexico	Accommodations	Cabin, Bed & Breakfast, Hotel
X4	California	Accommodations	Cabin, Bed & Breakfast, Hotel
X5	Maine	Accommodations	Cabin, Bed & Breakfast, Hotel
X6	Hawaii	Scenery	Mountains, Lake, Beach
X7	Alaska	Scenery	Mountains, Lake, Beach
X8	Mexico	Scenery	Mountains, Lake, Beach
X9	California	Scenery	Mountains, Lake, Beach
X10	Maine	Scenery	Mountains, Lake, Beach
X11	Hawaii	Price	$999, $1249, $1499
X12	Alaska	Price	$999, $1249, $1499
X13	Mexico	Price	$999, $1249, $1499
X14	California	Price	$999, $1249, $1499
X15	Maine	Price	$999, $1249, $1499

Factor	Destination	Attribute	Levels
X1	Hawaii	Accommodations	Cabin, Bed & Breakfast, Hotel
X6		Scenery	Mountains, Lake, Beach
X11		Price	$999, $1249, $1499
X2	Alaska	Accommodations	Cabin, Bed & Breakfast, Hotel
X7		Scenery	Mountains, Lake, Beach
X12		Price	$999, $1249, $1499
X3	Mexico	Accommodations	Cabin, Bed & Breakfast, Hotel
X8		Scenery	Mountains, Lake, Beach
X13		Price	$999, $1249, $1499
X4	California	Accommodations	Cabin, Bed & Breakfast, Hotel
X9		Scenery	Mountains, Lake, Beach
X14		Price	$999, $1249, $1499
X5	Maine	Accommodations	Cabin, Bed & Breakfast, Hotel
X10		Scenery	Mountains, Lake, Beach
X15		Price	$999, $1249, $1499

Each alternative is composed of three factors: package cost ($999, $1,249, $1,499), scenery (mountains, lake, beach), and accommodations (cabin, bed & breakfast, and hotel). There are five destinations, each with three attributes, for a total of 15 factors. This problem requires a design with 15 three-level factors, denoted 3^{15}. Each row of the design matrix contains the description of the five alternatives in one choice set. Note that the levels do not have to be the same for all destinations. For example, the cost for Hawaii and Alaska could be different from the other destinations. However, for this example, each destination will have the same attributes.

Set Up

We can use the `%MktRuns` autocall macro to suggest design sizes. (All of the autocall macros used in this book are documented starting on page 479.) To use this macro, you specify the number of levels for each of the factors. With 15 attributes each with three prices, you specify fifteen 3's (`3 3 3 3 3 3 3 3 3 3 3 3 3 3 3`) or you can use the more compact syntax of `3 ** 15`.

```
title 'Vacation Example';

%mktruns( 3 ** 15 )
```

The output tells us the size of the saturated design, which is the number of parameters in the linear design, and suggests design sizes.

Vacation Example

Design Summary

Number of Levels	Frequency
3	15

Vacation Example

Saturated = 31
Full Factorial = 14,348,907

Some Reasonable Design Sizes		Violations	Cannot Be Divided By
36		0	
45		0	
54	*	0	
63		0	
72	*	0	
33		105	9
39		105	9
42		105	9
48		105	9
51		105	9

* - 100% Efficient Design can be made with the MktEx Macro.

Vacation Example

n	Design	Reference
54	2 ** 1 3 ** 25	Orthogonal Array
54	2 ** 1 3 ** 21 9 ** 1	Orthogonal Array
54	3 ** 24 6 ** 1	Orthogonal Array
54	3 ** 20 6 ** 1 9 ** 1	Orthogonal Array
54	3 ** 18 18 ** 1	Orthogonal Array
72	2 ** 23 3 ** 24	Orthogonal Array
72	2 ** 22 3 ** 20 6 ** 1	Orthogonal Array
72	2 ** 21 3 ** 16 6 ** 2	Orthogonal Array
72	2 ** 20 3 ** 24 4 ** 1	Orthogonal Array
72	2 ** 19 3 ** 20 4 ** 1 6 ** 1	Orthogonal Array
72	2 ** 18 3 ** 16 4 ** 1 6 ** 2	Orthogonal Array
72	2 ** 16 3 ** 25	Orthogonal Array
72	2 ** 15 3 ** 21 6 ** 1	Orthogonal Array
72	2 ** 14 3 ** 24 6 ** 1	Orthogonal Array
72	2 ** 14 3 ** 17 6 ** 2	Orthogonal Array
72	2 ** 13 3 ** 25 4 ** 1	Orthogonal Array
72	2 ** 13 3 ** 20 6 ** 2	Orthogonal Array
72	2 ** 12 3 ** 24 12 ** 1	Orthogonal Array
72	2 ** 12 3 ** 21 4 ** 1 6 ** 1	Orthogonal Array
72	2 ** 12 3 ** 16 6 ** 3	Orthogonal Array
72	2 ** 11 3 ** 24 4 ** 1 6 ** 1	Orthogonal Array
72	2 ** 11 3 ** 20 6 ** 1 12 ** 1	Orthogonal Array
72	2 ** 11 3 ** 17 4 ** 1 6 ** 2	Orthogonal Array
72	2 ** 10 3 ** 20 4 ** 1 6 ** 2	Orthogonal Array
72	2 ** 10 3 ** 16 6 ** 2 12 ** 1	Orthogonal Array
72	2 ** 9 3 ** 16 4 ** 1 6 ** 3	Orthogonal Array
72	3 ** 25 8 ** 1	Orthogonal Array
72	3 ** 24 24 ** 1	Orthogonal Array

In this design, there are $15 \times (3-1)+1 = 31$ parameters, so at least 31 choice sets must be created. With all three-level factors, the number of choice sets in all orthogonal and balanced designs must be divisible by $3 \times 3 = 9$. Hence, optimal designs for this problem have at least 36 choice sets (the smallest number ≥ 31 and divisible by 9) and the number of choice sets must be a multiple of 9. Note however, that zero violations does not guarantee that a 100% efficient design exists. It just means that 100% efficiency is not precluded by unequal cell frequencies. In fact, the `%MktEx` orthogonal design catalog does not include orthogonal designs for this problem in 36, 45, and 63 runs (because they do not exist).

Thirty-six would be a good design size (2 blocks of size 18) as would 54 (3 blocks of size 18). Fifty-four would probably be the best choice, and that is what we would recommend for this study. However, we will instead create an efficient experimental design with 36 choice sets using the `%MktEx` macro. In practice, with more difficult designs, an orthogonal design is not available, and using 36 choice sets will allow us to see an example of using the `%Mkt` family of macros to get a nonorthogonal design.

We can see what orthogonal designs with three-level factors are available in 36 runs as follows. The `%MktOrth` macro creates a data set with information about the orthogonal designs that the `%MktEx` macro knows how to make. This macro produces a data set called MKTDESLEV that contains variables `n`, the number of runs; `Design`, a description of the design; and `Reference`, which contains the type of the design. In addition, there are variables: `x1`, the number of 1-level factors (which is always zero); `x2`, the number of 2-level factors; `x3`, the number of 3-level factors; and so on. We specify that `%MktOrth` only output `n=36` run designs and sort this list so that designs with the most three-level factors are printed first.

```
%mktorth(range=n=36)

proc sort data=mktdeslev out=list(drop=x:);
   by descending x3;
   run;

proc print; run;
```

Vacation Example

Obs	n	Design	Reference
1	36	2 ** 4 3 ** 13	Orthogonal Array
2	36	3 ** 13 4 ** 1	Orthogonal Array
3	36	2 ** 11 3 ** 12	Orthogonal Array
4	36	2 ** 2 3 ** 12 6 ** 1	Orthogonal Array
5	36	3 ** 12 12 ** 1	Orthogonal Array
6	36	2 ** 3 3 ** 9 6 ** 1	Orthogonal Array
7	36	2 ** 10 3 ** 8 6 ** 1	Orthogonal Array
8	36	2 ** 1 3 ** 8 6 ** 2	Orthogonal Array
9	36	3 ** 7 6 ** 3	Orthogonal Array
10	36	2 ** 2 3 ** 5 6 ** 2	Orthogonal Array
11	36	2 ** 13 3 ** 4	Orthogonal Array
12	36	2 ** 9 3 ** 4 6 ** 2	Orthogonal Array
13	36	2 ** 1 3 ** 3 6 ** 3	Orthogonal Array
14	36	2 ** 20 3 ** 2	Orthogonal Array
15	36	2 ** 11 3 ** 2 6 ** 1	Orthogonal Array

```
16    36    2 **  3  3 **  2  6 **  3    Orthogonal Array
17    36    2 ** 27  3 **  1             Orthogonal Array
18    36    2 ** 18  3 **  1  6 **  1    Orthogonal Array
19    36    2 ** 10  3 **  1  6 **  2    Orthogonal Array
20    36    2 **  4  3 **  1  6 **  3    Orthogonal Array
21    36    2 ** 35                       Hadamard
22    36    2 ** 13           9 **  1    Orthogonal Array
23    36    2 ** 13           6 **  2    Orthogonal Array
24    36    2 **  8           6 **  3    Orthogonal Array
25    36    2 **  2          18 **  1    Orthogonal Array
```

There are 13 two-level factors available in 36 runs, and we need 15, only two more, so we would expect to make a pretty good nonorthogonal design.

Designing the Choice Experiment

The following code creates a design.

```
%let m   = 6;                    /* m alts including constant        */
%let mm1 = %eval(&m - 1);        /* m - 1                            */
%let n   = 18;                   /* number of choice sets per person */
%let blocks = 2;                 /* number of blocks                 */

%mktex(3 ** 15 2, n=&n * &blocks, seed=151)
```

The specification `3 ** 15` requests a design with 15 factors, `x1`−`x15`, each with three levels. This specification also requests a two-level factor (the `2` following the `3 ** 15`). This is because 36 choice sets may be too many for one person to rate, so we may want to block the design into two blocks, and we can use a two-level factor to do this. A design with $18 \times 2 = 36$ runs is requested, which will mean 36 choice sets. A random number seed is explicitly specified so we will be able to reproduce these exact results.*

Here are some of the log messages.

```
NOTE: Generating the candidate set.
NOTE: Performing 20 searches of 81 candidates, full-factorial=28,697,814.
NOTE: Generating the orthogonal array design, n=36.
```

The macro searches a fractional-factorial candidate set of 81 runs, and it also generates a tabled design in 36 runs to try as part of the design. This will be explained in more detail on page 142.

*By specifying a random number seed, results should be reproducible within a SAS release for a particular operating system. However, due to machine differences, some results may not be exactly reproducible on other machines. For most orthogonal and balanced designs, the results should be reproducible. When computerized searches are done, it is likely that you will not get the same design as the one in the book, although you would expect the efficiency differences to be slight.

Here are some of the results from the `%MktEx` macro.

Vacation Example

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	82.2324	82.2324	Can
1	End	82.2324		
2	Start	78.4867		Tab,Ran
2	6 15	82.6447	82.6447	
2	7 15	82.7537	82.7537	
2	7 15	83.2331	83.2331	
2	8 15	84.6947	84.6947	
.				
.				
.				
2	End	98.7672		
3	Start	81.7827		Tab,Ran
3	11 15	98.9438	98.9438	
3	End	98.9438		
.				
.				
.				
11	Start	82.7854		Tab,Ran
11	End	97.8091		
12	Start	53.6699		Ran,Mut,Ann
12	End	92.8450		
.				
.				
.				
21	Start	48.8411		Ran,Mut,Ann
21	End	93.1010		

Design Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	98.9438	98.9438	Ini
1	Start	76.7520		Tab,Ran
1	End	98.6368		
2	Start	74.2345		Tab,Ran
2	End	98.8567		
.				
.				
.				
51	Start	80.4775		Tab,Ran
51	4 15	98.9438	98.9438	
51	End	98.9438		
.				
.				
.				
90	Start	78.6219		Tab,Ran
90	14 15	98.9438	98.9438	
90	End	98.9438		
.				
.				
.				
138	Start	81.4007		Tab,Ran
138	32 15	98.9438	98.9438	
138	End	98.9438		

NOTE: Stopping since it appears that no improvement is possible.

Vacation Example

Design Refinement History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	98.9438	98.9438	Ini

```
1      Start       97.5641                Pre,Mut,Ann
1     14  13       98.9438       98.9438
1     15   1       98.9438       98.9438
1     18  13       98.9438       98.9438
1     21   4       98.9438       98.9438
1        End       97.9925
.
.
.

9      Start       96.3118                Pre,Mut,Ann
9     14  10       98.9438       98.9438
9     15   6       98.9438       98.9438
9        End       98.5372
```

NOTE: Stopping since it appears that no improvement is possible.

Vacation Example

The OPTEX Procedure

Class Level Information

Class	Levels	-Values-
x1	3	1 2 3
x2	3	1 2 3
x3	3	1 2 3
x4	3	1 2 3
x5	3	1 2 3
x6	3	1 2 3
x7	3	1 2 3
x8	3	1 2 3
x9	3	1 2 3
x10	3	1 2 3
x11	3	1 2 3
x12	3	1 2 3
x13	3	1 2 3
x14	3	1 2 3
x15	3	1 2 3
x16	2	1 2

Vacation Example

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	98.9437	97.9592	98.9743	0.9428

The %MktEx macro used 32 seconds and found a design that is almost 99% efficient. (Differences in the fourth decimal place between the iteration history and the final table, in this case 98.9438 versus 98.9437, are due to rounding error and differences in ridging strategies between the macro and PROC OPTEX and are nothing to worry about.)

The %MktEx Macro Algorithm

The %MktEx macro creates efficient linear experimental designs using several approaches. The macro will try to create a tabled design, it will search a set of candidate runs (rows of the design), and it will use a coordinate-exchange algorithm using both random initial designs and also a partial tabled design initialization. The macro stops if at any time it finds a perfect, 100% efficient, orthogonal and balanced design. This first phase is the algorithm search phase. In it, the macro determines which approach is working best for this problem. At the end of this phase, the macro chooses the method that has produced the best design and performs another set of iterations using exclusively the chosen approach. Finally, the macro performs a third set of iterations where it takes the best design it found so far and tries to improve it.

The %MktEx macro can directly generate, without iterations, thousands of different 100% *D*-efficient, orthogonal and balanced, tabled designs. It does this using its design catalog and many different general and ad hoc algorithms. The closest design that the macro knows how to make for this problem is $2^1 3^{13}$ in 36 runs.

The candidate-set search has two parts. First, either PROC PLAN is run to create a full-factorial design for small problems, or PROC FACTEX is run to create a fractional-factorial design for large problems. Either way, this larger design is a *candidate set* that in the second part is searched by PROC OPTEX using the modified Fedorov algorithm. A design is built from a selection of the rows of the candidate set (Fedorov, 1972; Cook and Nachtsheim, 1980). The modified Fedorov algorithm considers each run in the design and each candidate run. Candidate runs are swapped in and design runs are swapped out if the swap improves *D*-efficiency. In this case, since the full-factorial design is large (over 28 million runs), the candidate-set search step calls PROC FACTEX to make the candidate set and then PROC OPTEX to do the search. The Can line of the iteration history shows that this step found a design that was 82.2324% efficient.

Next, the %MktEx macro uses the *coordinate-exchange algorithm*, based on Meyer and Nachtsheim (1995). The coordinate-exchange algorithm considers each level of each factor, and considers the effect on *D*-efficiency of changing a level ($1 \rightarrow 2$, or $1 \rightarrow 3$, or $2 \rightarrow 1$, or $2 \rightarrow 3$, or $3 \rightarrow 1$, or $3 \rightarrow 2$, and so on). Exchanges that increase efficiency are performed. In this step, the macro first tries to initialize the

design with a tabled design (`Tab`) and a random design (`Ran`) both. In this case, 14 of the 16 columns can be initialized with 14 columns of $2^4 3^{13}$, and the other two columns are randomly initialized. Levels that are not orthogonally initialized may be exchanged for other levels if the exchange increases efficiency. For example, the iteration history for this example shows that the macro exchanged levels in row 6 column 15, row 7 column 15, ..., row 11 column 15, and so on.

The initialization may be more complicated in other problems. Say you asked for the design $4^1 5^1 3^4$ in 18 runs. The macro would use the tabled design $3^6 6^1$ in 18 runs to initialize the three-level factors orthogonally, and the five-level factor with the six-level factor coded down to five levels (and hence unbalanced). The four-level factor would be randomly initialized. The macro would also try the same initialization but with a random rather than unbalanced initialization of the five-level factor, as a minor variation on the first initialization. In the next initialization variation, the macro would use a fully random initialization. If the number of runs requested were smaller than the number or runs in the initial tabled design, the macro would initialize the design with just the first n rows of the tabled design. Similarly, if the number of runs requested were larger than the number or runs in the initial tabled design, the macro would initialize part of the design with the orthogonal tabled design and the remaining rows and columns randomly. The coordinate-exchange algorithm considers each level of each factor that is not orthogonally initialized, and it exchanges a level if the exchange improves D-efficiency. When the number or runs in the tabled design does not match the number of runs desired, none of the design is initialized orthogonally.

The coordinate-exchange algorithm is not restricted by having a candidate set and hence can *potentially* consider any possible design. In practice, however, both the candidate-set-based and coordinate-exchange algorithms consider only a tiny fraction of the possible designs. When the number of runs in the full-factorial design is very small (say 100 or 200 runs), the modified Fedorov algorithm and coordinate exchange algorithms usually work equally well. When the number of runs in the full-factorial design is small (up to several thousand), the modified Fedorov algorithm is often superior to coordinate exchange. When the full-factorial design is larger, coordinate exchange is usually the superior approach. However, heuristics like these are sometimes wrong, which is why the macro tries both methods to see which one is really best for each problem.

In the first attempt at coordinate exchange (Design 2), the macro found a design that is 98.7672% efficient (Design 2, `End`). In design 3 and subsequent designs, the macro uses this same approach, but different random initializations of the remaining two columns. In design 3, the `%MktEx` macro finds a design that is 98.9438% efficient. Designs 12 through 21 use a purely random initialization and simulated annealing and are not as good as previous designs. During these iterations, the macro is considering exchanging every level of every factor with all of the other levels, one level of one factor at a time.

At this point, the `%MktEx` macro determines that the combination of tabled and random initialization is working best and tries more iterations using that approach. It starts by printing the initial (`Ini`) best efficiency of 98.9438. In designs 51, 90, 138, and others not shown, the macro finds a design that is 98.9438% efficient. After iteration 138, the macro stops since it keeps finding the same design over and over. This does not necessarily mean the macro found *the* optimal design; it means it found a very attractive (perhaps local) optimum, and it is unlikely it will do better using this approach.

Next, the `%MktEx` macro tries to improve the best design it found previously. Using the previous best design as an initialization (`Pre`), and random mutations of the initialization (`Mut`) and simulated annealing (`Ann`), the macro uses the coordinate-exchange algorithm to try to find a better design. This step is important because the best design that the macro found may be an intermediate design and not be the final design at the end of an iteration. Sometimes the iterations deliberately make the designs

less efficient, and sometimes, the macro never finds a design as efficient or more efficient again. Hence it is worthwhile to see if the best design found so far can be improved. In this case the macro fails to improve the design. At the end, PROC OPTEX is called to print the levels of each factor and the final *D*-efficiency.

Random mutations add random noise to the initial design before iterations start (levels are randomly changed). This may eliminate the perfect balance that will often be in the initial design. By default, random mutations are used with designs with fully random initializations and in the design refinement step; orthogonal initial designs are not mutated.

Simulated annealing allows the design to get worse occasionally but with decreasing probability as the number of exchanges increases. For design 1, for the first level of the first factor, by default, the macro may execute an exchange (say change a 2 to a 1), that makes the design worse, with probability 0.05. As more and more exchanges occur, this probability decreases so at the end of the processing of design 1, exchanges that decrease efficiency are hardly ever done. For design 2, this same process is repeated, again starting by default with an annealing probability of 0.05. This often helps the algorithm overcome local efficiency maxima. To envision this, imagine that you are standing on a molehill next to a mountain. The only way you can start going up the mountain is to first step down off the molehill. Once you are on the mountain, you may occasionally hit a dead end, where all you can do is step down and look for a better place to continue going up. Simulated annealing, by occasionally stepping down the efficiency function, often allows the macro to go farther up it than it would otherwise. The simulated annealing is why you will sometimes see designs getting worse in the iteration history. However, the macro keeps track of the best design, not the final design in each step. By default, annealing is used with designs with fully random initializations and in the design refinement step; simulated annealing is not used with orthogonal initial designs.

For this example, the `%MktEx` macro ran in around 30 seconds. If an orthogonal design had been available, run time would have been a few seconds. If the fully random initialization method had been the best method, run time might have been on the order of 10 to 45 minutes. Since the tabled initialization worked best, run time was much shorter. While it is possible to construct huge problems that will take much longer, for any design that most marketing researchers are likely to encounter, run time should be less than one hour. One of the macro options, `maxtime=`, ensures this.

Examining the Design

Before you use a design, you should always look at its characteristics. We will use the `%MktEval` macro.

```
%mkteval;
```

Here are some of the results.

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16
x1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
x4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
x5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
x6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
x7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
x8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
x9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
x10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
x11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
x12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
x13	0	0	0	0	0	0	0	0	0	0	0	0	1	0.25	0.25	0
x14	0	0	0	0	0	0	0	0	0	0	0	0	0.25	1	0.25	0
x15	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	1	0
x16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Vacation Example
Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316
* - Indicates Unequal Frequencies

Frequencies

x1	12 12 12
x2	12 12 12
x3	12 12 12
x4	12 12 12
x5	12 12 12
x6	12 12 12
x7	12 12 12
x8	12 12 12
x9	12 12 12
x10	12 12 12
x11	12 12 12
x12	12 12 12
x13	12 12 12
x14	12 12 12
x15	12 12 12
x16	18 18

```
      x1 x2      4 4 4 4 4 4 4 4 4
      x1 x3      4 4 4 4 4 4 4 4 4
      x1 x4      4 4 4 4 4 4 4 4 4
      x1 x5      4 4 4 4 4 4 4 4 4
      x1 x6      4 4 4 4 4 4 4 4 4
      x1 x7      4 4 4 4 4 4 4 4 4
      x1 x8      4 4 4 4 4 4 4 4 4
      x1 x9      4 4 4 4 4 4 4 4 4
      x1 x10     4 4 4 4 4 4 4 4 4
      x1 x11     4 4 4 4 4 4 4 4 4
      x1 x12     4 4 4 4 4 4 4 4 4
      x1 x13     4 4 4 4 4 4 4 4 4
      x1 x14     4 4 4 4 4 4 4 4 4
      x1 x14     4 4 4 4 4 4 4 4 4
      x1 x15     4 4 4 4 4 4 4 4 4
      x1 x16     6 6 6 6 6 6
      .
      .
      .
*     x13 x14    6 3 3 3 3 6 3 6 3
*     x13 x15    6 3 3 3 3 6 3 6 3
      x13 x16    6 6 6 6 6 6
*     x14 x15    6 3 3 3 6 3 3 3 6
      x14 x16    6 6 6 6 6 6
      x15 x16    6 6 6 6 6 6
      N-Way      1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

This design looks great! The factors `x1-x13` form an orthogonal design, `x14` and `x15` are slightly correlated with each other and with `x13`. The blocking factor `x16` is orthogonal to all the other factors. All of the factors are perfectly balanced. The N-Way frequencies show that each choice set appears once.

What if there had been some larger canonical correlations? Would this be a problem? That depends. You have to decide this for yourself based on your particular study. You do not want large correlations between your most important factors. If you have high correlations between the wrong factors, you can swap them with other factors with the same number of levels, or try to make a new design with a different seed, or change the number of choice sets, and so on. While this design looks great, we should make one minor adjustment based on these results. Since our correlations are in the factors we originally planned to make price factors, we should change our plans slightly and use those factors for less important attributes like scenery.

You can run the `%MktEx` macro to provide additional information about a design, for example asking to examine the information matrix (`I`) and its inverse (`V`), which is the variance matrix of the parameter estimates. You hope to see that all of the off-diagonal elements of the variance matrix, the covariances, are small relative to the variances on the diagonal. When `options=check` is specified, the macro evaluates an initial design instead of generating a design. The option `init=randomized` names the design to evaluate, and the `examine=` option displays the information and variance matrices. The blocking variable was dropped.

```
%mktex(3 ** 15,  n=&n * &blocks, init=randomized(drop=x16),
       options=check, examine=i v)
```

Here is a small part of the output.

Vacation Example

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	98.9099	97.8947	98.9418	0.9280

Information Matrix

	Intercept	x11	x12	x21	x22	x31	x32	x41	x42	x51	x52
Intercept	36	0	0	0	-0	0	-0	0	-0	0	-0
x11	0	36	0	-0	0	0	0	0	-0	-0	0
x12	0	0	36	0	-0	0	-0	-0	-0	-0	0
x21	0	-0	0	36	0	-0	-0	0	-0	-0	0
x22	-0	0	-0	0	36	-0	0	-0	0	-0	-0
x31	0	0	0	-0	-0	36	0	-0	-0	-0	0
x32	-0	0	-0	-0	0	0	36	-0	-0	0	0
x41	0	0	-0	0	-0	-0	-0	36	0	0	-0
x42	-0	-0	-0	-0	0	-0	-0	0	36	0	-0
x51	0	-0	-0	-0	-0	-0	0	0	0	36	0

.
.
.

Information Matrix

	x112	x121	x122	x131	x132	x141	x142	x151	x152
x112	36	-0	0	0	-0	-0	0	-0	-0
x121	-0	36	0	0	-0	0	0	0	-0
x122	0	0	36	0	-0	0	0	-0	-0
x131	0	0	0	36	0	9	0	9	0
x132	-0	-0	-0	0	36	0	-9	0	-9
x141	-0	0	0	9	0	36	0	9	0
x142	0	0	0	0	-9	0	36	0	9
x151	-0	0	-0	9	0	9	0	36	0
x152	-0	-0	-0	0	-9	0	9	0	36

Variance Matrix

	Intercept	x11	x12	x21	x22	x31
Intercept	0.0278	-0.0000	-0.0000	0.0000	0.0000	-0.0000
x11	-0.0000	0.0278	-0.0000	0.0000	-0.0000	-0.0000
x12	-0.0000	-0.0000	0.0278	-0.0000	0.0000	-0.0000
x21	0.0000	0.0000	-0.0000	0.0278	-0.0000	0.0000
x22	0.0000	-0.0000	0.0000	-0.0000	0.0278	0.0000
x31	-0.0000	-0.0000	-0.0000	0.0000	0.0000	0.0278

.
.
.

Variance Matrix

	x131	x132	x141	x142	x151	x152
x131	0.0309	-0.0000	-0.0062	-0.0000	-0.0062	-0.0000
x132	-0.0000	0.0309	-0.0000	0.0062	-0.0000	0.0062
x141	-0.0062	-0.0000	0.0309	-0.0000	-0.0062	0.0000
x142	-0.0000	0.0062	-0.0000	0.0309	0.0000	-0.0062
x151	-0.0062	-0.0000	-0.0062	0.0000	0.0309	-0.0000
x152	-0.0000	0.0062	0.0000	-0.0062	-0.0000	0.0309

This design still looks good. The *D*-efficiency for the design excluding the blocking factor is 98.9099%. We can see that the nonorthogonality between `x13-x15` make their variances larger than the other factors (0.0309 versus 0.0278).

This variance matrix is a little hard to look at. All of the 0.0000 and -0.0000's tend to obscure the nonzeros. We can use ODS along with PROC FORMAT and PROC PRINT to make a better display. The variance matrix is excluded from the printed output and instead is output to a SAS data set. The `persist` option is used since the ODS statements need to persist through the macro steps until the macro reaches the PROC OPTEX step. In SAS 9 and previous SAS versions, the `match_all` option must be specified with `persist` on the `ods output` statement. PROC FORMAT is used to construct a format so that the values within rounding error of zero print as '0'. PROC PRINT is called to print the results. The label statement gives the row ID variable, `rowname` a null header.

```
ods exclude 'variance matrix'(persist);
ods output 'variance matrix'(persist match_all)=v;
%mktex(3 ** 15,  n=&n * &blocks, init=randomized(drop=x16),
       options=check, examine=v)

proc format; value zer -1e-8 - 1e-8 = ' 0     '; run;

proc print label data=v(drop=_:);
   format _numeric_ zer7.4;
   label rowname = '00'x;
   id rowname;
   run;
```

Vacation Example

	Intercept	x11	x12	x21	x22	x31	x32	x41
Intercept	0.0278	0	0	0	0	0	0	0
x11	0	0.0278	0	0	0	0	0	0
x12	0	0	0.0278	0	0	0	0	0
x21	0	0	0	0.0278	0	0	0	0
x22	0	0	0	0	0.0278	0	0	0
x31	0	0	0	0	0	0.0278	0	0
x32	0	0	0	0	0	0	0.0278	0
x41	0	0	0	0	0	0	0	0.0278

.
.
.

	x122	x131	x132	x141	x142	x151	x152
x112	0	0	0	0	0	0	0
x121	0	0	0	0	0	0	0
x122	0.0278	0	0	0	0	0	0
x131	0	0.0309	0	-0.0062	0	-0.0062	0
x132	0	0	0.0309	0	0.0062	0	0.0062
x141	0	-0.0062	0	0.0309	0	-0.0062	0
x142	0	0	0.0062	0	0.0309	0	-0.0062
x151	0	-0.0062	0	-0.0062	0	0.0309	0
x152	0	0	0.0062	0	-0.0062	0	0.0309

These next steps use the `%MktLab` macro to reassign the variable names, store the design in a permanent SAS data set, SASUSER.BLOCKDES, and then use the `%MktEx` macro to check the results. The `vars=` option provides the new variable names: the first variable (originally `x1`) becomes `x1` (still), ..., the fifth variable (originally `x5`) becomes `x5` (still), the sixth variable (originally `x6`) becomes `x11`, ... the tenth variable (originally `x10`) becomes `x15`, the eleventh through fifteenth original variables become `x6`, `x9`, `x7`, `x8`, `x10`, and finally the last variable becomes `Block`. We made the correlated variables correspond to the least important attributes in different alternatives (in this case the scenery factors

for Alaska, Mexico, and Maine).

```
%mktlab(data=randomized, vars=x1-x5 x11-x15 x6 x9 x7 x8 x10 Block,
        out=sasuser.blockdes)

%mkteval(blocks=block)
```

Here is the output from the %MktLab macro, which shows the correspondence between the original and new variable names.

```
Variable Mapping:
   x1  : x1
   x2  : x2
   x3  : x3
   x4  : x4
   x5  : x5
   x6  : x11
   x7  : x12
   x8  : x13
   x9  : x14
   x10 : x15
   x11 : x6
   x12 : x9
   x13 : x7
   x14 : x8
   x15 : x10
   x16 : Block
```

Here is some of the output from the %MktEval macro.

Vacation Example
Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

	Block	x1	x2	x3	x4	x5	x11	x12	x13	x14	x15	x6	x9	x7	x8	x10
Block	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
x3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
x4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
x5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
x11	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
x12	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
x13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
x14	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
x15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

x6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
x9	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
x7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.25	0.25
x8	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	1	0.25
x10	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	1

Vacation Example
Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316
* - Indicates Unequal Frequencies

	Frequencies
Block	18 18
x1	12 12 12
x2	12 12 12
x3	12 12 12
x4	12 12 12
x5	12 12 12
x11	12 12 12
x12	12 12 12
x13	12 12 12
x14	12 12 12
x15	12 12 12
x6	12 12 12
x9	12 12 12
x7	12 12 12
x8	12 12 12
x10	12 12 12
Block x1	6 6 6 6 6 6
Block x2	6 6 6 6 6 6
Block x3	6 6 6 6 6 6
Block x4	6 6 6 6 6 6
Block x5	6 6 6 6 6 6
Block x11	6 6 6 6 6 6
Block x12	6 6 6 6 6 6
Block x13	6 6 6 6 6 6
Block x14	6 6 6 6 6 6
Block x15	6 6 6 6 6 6
Block x6	6 6 6 6 6 6
Block x9	6 6 6 6 6 6
Block x7	6 6 6 6 6 6
Block x8	6 6 6 6 6 6
Block x10	6 6 6 6 6 6

```
      x1 x2        4 4 4 4 4 4 4 4 4
      x1 x3        4 4 4 4 4 4 4 4 4
      x1 x4        4 4 4 4 4 4 4 4 4
      .
      .
      .
      x9 x7        4 4 4 4 4 4 4 4 4
      x9 x8        4 4 4 4 4 4 4 4 4
      x9 x10       4 4 4 4 4 4 4 4 4
*     x7 x8        6 3 3 3 3 6 3 6 3
*     x7 x10       6 3 3 3 3 6 3 6 3
*     x8 x10       6 3 3 3 6 3 3 3 6
       N-Way       1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                   1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Generating the Questionnaire

This next DATA step prints the questionnaires. They are then copied and the data are collected.

```
title;
proc sort data=sasuser.blockdes; by block; run;

options ls=80 ps=60 nodate nonumber;

data _null_;
   array dests[&mm1] $ 10 _temporary_ ('Hawaii' 'Alaska' 'Mexico'
                                       'California' 'Maine');
   array prices[3]   $  5 _temporary_ ('$999' '$1249' '$1499');
   array scenes[3]   $ 13 _temporary_
                     ('the Mountains' 'a Lake' 'the Beach');
   array lodging[3]  $ 15 _temporary_
                     ('Cabin' 'Bed & Breakfast' 'Hotel');
   array x[15];
   file print linesleft=ll;
   set sasuser.blockdes;
   by block;

   if first.block then do;
      choice = 0;
      put _page_;
      put @50 'Form: ' block  ' Subject: ________' //;
      end;
   choice + 1;
```

```
if ll < 19 then put _page_;
put choice 2. ') Circle your choice of '
    'vacation destinations:' /;
do dest = 1 to &mm1;
   put '    ' dest 1. ') ' dests[dest]
       +(-1) ', staying in a ' lodging[x[dest]]
       'near ' scenes[x[&mm1 + dest]] +(-1) ',' /
       '       with a package cost of '
      prices[x[2 * &mm1 + dest]] +(-1) '.' /;
   end;
put "    &m) Stay at home this year." /;
run;
```

In this design, there are five destinations, and each destination has three attributes. Each destination name is accessed from the array `dests`. Note that destination is not a factor in the design; it is a bin into which the attributes are grouped. The factors in the design are named in the statement `array x[15]`, which is a short-hand notation for `array x[15] x1-x15`. The first five factors are used for the lodging attribute of the five destinations. The actual descriptions of lodging are accessed by `lodging[x[dest]]`. The variable `Dest` varies from 1 to 5 destinations, so `x[dest]` extracts the levels for the `Dest` destination. Similarly for scenery, `scenes[x[&mm1 + dest]]` extracts the descriptions of the scenery. The index `&mm1 + dest` accesses factors 6 through 10, and `x[&mm1 + dest]` indexes the `scenes` array. For prices, `prices[x[2 * &mm1 + dest]]`, the index `2 * &mm1 + dest` accesses the factors 11 through 15. Here are the first two choice sets.

```
                                          Form: 1  Subject: ________


  1) Circle your choice of vacation destinations:

     1) Hawaii, staying in a Cabin near a Lake,
        with a package cost of $999.

     2) Alaska, staying in a Cabin near a Lake,
        with a package cost of $1249.

     3) Mexico, staying in a Hotel near the Mountains,
        with a package cost of $1499.

     4) California, staying in a Bed & Breakfast near a Lake,
        with a package cost of $1499.

     5) Maine, staying in a Cabin near the Mountains,
        with a package cost of $1499.

     6) Stay at home this year.
```

```
2) Circle your choice of vacation destinations:

   1) Hawaii, staying in a Hotel near the Mountains,
      with a package cost of $1499.

   2) Alaska, staying in a Hotel near a Lake,
      with a package cost of $1499.

   3) Mexico, staying in a Hotel near the Beach,
      with a package cost of $1499.

   4) California, staying in a Hotel near the Mountains,
      with a package cost of $1499.

   5) Maine, staying in a Bed & Breakfast near the Beach,
      with a package cost of $1499.

   6) Stay at home this year.
```

In practice, data collection may be much more elaborate than this. It may involve art work or photographs, and the choice sets may be presented and the data may be collected through personal interview or over the web. However the choice sets are presented and the data collected, the essential ingredients remain the same. Subjects are shown sets of alternatives and asked to make a choice, then they go on to the next set.

Entering and Processing the Data

Here are some of the input data. Data from a total of 200 subjects were collected, 100 per form.

```
data results;
   input Subj Form (choose1-choose&n) (1.) @@;
   datalines;
  1  1 132513243441314151   2  2 455113115112113413   3  1 132153331144534151
  4  2 431133511214334413   5  1 133113141141321443   6  2 151114113242134413
  7  1 143553241511354153   8  2 451113513244113511   9  1 111124231141311151
 10  2 454113133144133513  11  1 153513231543321153  12  2 431153113414334513
 13  1 153514531414314151  14  2 431114113142133113  15  1 143513111511314151
 16  2 515113143414334313  17  1 133125333424111151  18  2 525113113111134111
 19  1 143113131141344211  20  2 435114333112433413  21  1 133113531543215143
.
.
.
;
```

These next steps prepare the design for analysis. We need to covert our linear design into a choice design.† We need to create a data set KEY that describes how the factors in our linear design will be used to make the choice design for analysis. The KEY data set will contain all of the factor names,

†See page 87 for an illustration of linear versus choice designs.

x1, x2, x3, ... x15. We can run the %MktKey macro to get these names in the SAS log, for cutting and pasting into the program without typing them.

```
%mktkey(x1-x15)
```

The %MktKey macro produced the following line.

```
x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15
```

This code makes the KEY data set and processes the design.

```
title 'Vacation Example';

data key;
   input Place $ 1-10 (Lodge Scene Price) ($);
   datalines;
Hawaii      x1  x6   x11
Alaska      x2  x7   x12
Mexico      x3  x8   x13
California  x4  x9   x14
Maine       x5  x10  x15
Home        .   .    .
;

%mktroll(design=sasuser.blockdes, key=key, alt=place, out=rolled)
```

For analysis, the design will have four factors as shown by the variables in the data set KEY. Place is the alternative name; its values are directly read from the KEY in-stream data. Lodge is an attribute whose values will be constructed from the SASUSER.BLOCKDES data set. Lodge is created from x1 for Hawaii, x2 for Alaska, ..., x5 for Maine, and no attribute for Home. Similarly, Scene is created from x6-x10, and Price is created from x11-x15. The macro %MktRoll is used to create the data set ROLLED from SASUSER.BLOCKDES using the mapping in KEY and using the variable Place as the alternative ID variable.

The macro warns us:

```
WARNING: The variable block is in the DESIGN= data set but not
         the KEY= data set.
```

While this message could indicate a problem, in this case it does not. The variable Block in the design= sasuser.blockdes data set will not appear in the final design. The purpose of the variable Block (sorting the design into blocks) has already been achieved. You can specify options=nowarn if you want to suppress this warning.

These next steps show the results for the first two choice sets. The data set is converted from a design matrix with one row per choice set to a design matrix with one row per alternative per choice set.

```
proc print data=sasuser.blockdes(obs=2);
   id Block;
   var x1-x15;
   run;

proc print data=rolled(obs=12); run;
```

Vacation Example

Block	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15
1	1	1	3	2	1	2	2	1	2	1	1	2	3	3	3
1	3	3	3	3	2	1	2	3	1	3	3	3	3	3	3

Vacation Example

Obs	Set	Place	Lodge	Scene	Price
1	1	Hawaii	1	2	1
2	1	Alaska	1	2	2
3	1	Mexico	3	1	3
4	1	California	2	2	3
5	1	Maine	1	1	3
6	1	Home	.	.	.
7	2	Hawaii	3	1	3
8	2	Alaska	3	2	3
9	2	Mexico	3	3	3
10	2	California	3	1	3
11	2	Maine	2	3	3
12	2	Home	.	.	.

The next steps assign formats, convert the variable `Price` to contain actual prices, and recode the constant alternative.

```
proc format;
   value price 1 = ' 999'      2 = '1249'
               3 = '1499'      0 = '   0';
   value scene 1 = 'Mountains' 2 = 'Lake'
               3 = 'Beach'     0 = 'Home';
   value lodge 1 = 'Cabin'     2 = 'Bed & Breakfast'
               3 = 'Hotel'     0 = 'Home';
   run;
data rolled2;
   set rolled;
   if place = 'Home' then do; lodge = 0; scene = 0; price = 0; end;
   price = input(put(price, price.), 5.);
   format scene scene. lodge lodge.;
   run;
proc print data=rolled2(obs=12); run;
```

Vacation Example

Obs	Set	Place	Lodge	Scene	Price
1	1	Hawaii	Cabin	Lake	999
2	1	Alaska	Cabin	Lake	1249
3	1	Mexico	Hotel	Mountains	1499
4	1	California	Bed & Breakfast	Lake	1499
5	1	Maine	Cabin	Mountains	1499
6	1	Home	Home	Home	0
7	2	Hawaii	Hotel	Mountains	1499
8	2	Alaska	Hotel	Lake	1499
9	2	Mexico	Hotel	Beach	1499
10	2	California	Hotel	Mountains	1499
11	2	Maine	Bed & Breakfast	Beach	1499
12	2	Home	Home	Home	0

It is not necessary to recode the missing values for the constant alternative. In practice, we usually will not do this step. However, for this first analysis, we will want all nonmissing values of the attributes so we can see all levels in the final printed output. We also recode `Price` so that for a later analysis, we can analyze `Price` as a quantitative effect. For example, the expression `put(price, price.)` converts a number, say 2, into a string (in this case '1249'), then the `input` function reads the string and converts it to a numeric 1249. Next, we use the macro `%MktMerge` to combine the data and design and create the variable `c`, indicating whether each alternative was a first choice or a subsequent choice.

```
%mktmerge(design=rolled2, data=results, out=res2, blocks=form,
          nsets=&n, nalts=&m, setvars=choose1-choose&n)

proc print data=res2(obs=12); run;
```

This macro takes the `design=rolled2` experimental design, merges it with the `data=result` data set, creating the `out=res2` output data set. The RESULTS data set contains the variable `Form` that contains the block number. Since there are two blocks, this variable must have values of 1 and 2. This variable must be specified in the `blocks=` option. The experiment has `nsets=&n` choice sets, `nalts=6` alternatives, and the variables `setvars=choose1-choose&n` contain the numbers of the chosen alternatives. The output data set RES2 has 21600 observations (200 subjects who each saw 18 choice sets with 6 alternatives). Here are the first two choice sets.

Obs	Subj	Form	Set	Place	Lodge	Scene	Price	c
1	1	1	1	Hawaii	Cabin	Lake	999	1
2	1	1	1	Alaska	Cabin	Lake	1249	2
3	1	1	1	Mexico	Hotel	Mountains	1499	2
4	1	1	1	California	Bed & Breakfast	Lake	1499	2
5	1	1	1	Maine	Cabin	Mountains	1499	2
6	1	1	1	Home	Home	Home	0	2

7	1	1	2	Hawaii	Hotel	Mountains	1499	2
8	1	1	2	Alaska	Hotel	Lake	1499	2
9	1	1	2	Mexico	Hotel	Beach	1499	1
10	1	1	2	California	Hotel	Mountains	1499	2
11	1	1	2	Maine	Bed & Breakfast	Beach	1499	2
12	1	1	2	Home	Home	Home	0	2

Binary Coding

One more thing must be done to these data before they can be analyzed. The binary design matrix is coded for each effect. This can be done with PROC TRANSREG.

```
proc transreg design=5000 data=res2 nozeroconstant norestoremissing;
   model class(place / zero=none order=data)
         class(price scene lodge / zero=none order=formatted) /
         lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   id subj set form c;
   run;
```

The `design` option specifies that no model is fit; the procedure is just being used to code a design. When `design` is specified, dependent variables are not required. Optionally, `design` can be followed by "=n" where n is the number of observations to process at one time. By default, PROC TRANSREG codes all observations in one big group. For very large data sets, this can consume large amounts of memory and time. Processing blocks of smaller numbers of observations is more efficient. The option `design=5000` processes observations in blocks of 5000. For smaller computers, try something like `design=1000`.

The `nozeroconstant` and `norestoremissing` options are not necessary for this example but are included here because sometimes they are very helpful in coding choice models. The `nozeroconstant` option specifies that if the coding creates a constant variable, it should not be zeroed. The `nozeroconstant` option should always be specified when you specify `design=`n because the last group of observations may be small and may contain constant variables. The `nozeroconstant` option is also important if you do something like coding `by subj set` because sometimes an attribute is constant within a choice set. The `norestoremissing` option specifies that missing values should not be restored when the `out=` data set is created. By default, the coded `class` variable contains a row of missing values for observations in which the `class` variable is missing. When you specify the `norestoremissing` option, these observations contain a row of zeros instead. This option is useful when there is a constant alternative indicated by missing values. Both of these options, like almost all options in PROC TRANSREG, can be abbreviated to three characters (`noz` and `nor`).

The `model` statement names the variables to code and provides information about how they should be coded. The specification `class(place / ...)` specifies that the variable `Place` is a classification variable and requests a binary coding. The `zero=none` option creates binary variables for all categories. The `order=data` option sorts the levels into the order they were first encountered in the data set. It is specified so `'Home'` will be the last destination in the analysis, as it is in the data set. The `class(price scene lodge / ...)` specification names the variables `Price`, `Scene`, and `Lodge` as categorical variables and creates binary variables for all of the levels of all of the variables. The levels are sorted into order based on their formatted values. The `lprefix=0` option specifies that when

labels are created for the binary variables, zero characters of the original variable name should be used as a prefix. This means that the labels are created only from the level values. For example, `'Mountains'` and `'Bed & Breakfast'` are created as labels not `'Scene Mountains'` and `'Lodge Bed & Breakfast'`.

An `output` statement names the output data set and drops variables that are not needed. These variables do not have to be dropped. However, since they are variable names that are often found in special data set types, PROC PHREG prints warnings when it finds them. Dropping the variables prevents the warnings. Finally, the `id` statement names the additional variables that we want copied from the input to the output data set. The next steps print the first coded choice set.

```
proc print data=coded(obs=6);
   id place;
   var subj set form c price scene lodge;
   run;
proc print data=coded(obs=6) label;
   var pl:;
   run;
proc print data=coded(obs=6) label;
   id place;
   var sc:;
   run;
proc print data=coded(obs=6) label;
   id place;
   var lo: pr:;
   run;
```

Vacation Example

Place	Subj	Set	Form	c	Price	Scene	Lodge
Hawaii	1	1	1	1	999	Lake	Cabin
Alaska	1	1	1	2	1249	Lake	Cabin
Mexico	1	1	1	2	1499	Mountains	Hotel
California	1	1	1	2	1499	Lake	Bed & Breakfast
Maine	1	1	1	2	1499	Mountains	Cabin
Home	1	1	1	2	0	Home	Home

Vacation Example

Obs	Hawaii	Alaska	Mexico	California	Maine	Home	Place
1	1	0	0	0	0	0	Hawaii
2	0	1	0	0	0	0	Alaska
3	0	0	1	0	0	0	Mexico
4	0	0	0	1	0	0	California
5	0	0	0	0	1	0	Maine
6	0	0	0	0	0	1	Home

Vacation Example

Place	Beach	Home	Lake	Mountains	Scene
Hawaii	0	0	1	0	Lake
Alaska	0	0	1	0	Lake
Mexico	0	0	0	1	Mountains
California	0	0	1	0	Lake
Maine	0	0	0	1	Mountains
Home	0	1	0	0	Home

Vacation Example

Place	Bed & Breakfast	Cabin	Home	Hotel	Lodge	0	999	1249	1499	Price
Hawaii	0	1	0	0	Cabin	0	1	0	0	999
Alaska	0	1	0	0	Cabin	0	0	1	0	1249
Mexico	0	0	0	1	Hotel	0	0	0	1	1499
California	1	0	0	0	Bed & Breakfast	0	0	0	1	1499
Maine	0	1	0	0	Cabin	0	0	0	1	1499
Home	0	0	1	0	Home	1	0	0	0	0

The coded design consists of binary variables for destinations Hawaii − Home, scenery Beach − Mountains, lodging Bed & Breakfast − Hotel, and price 0 − 1499. For example, in the last printed panel of the first choice set, the Bed & Breakfast column has a 1 for Hawaii since Hawaii has B & B lodging in this choice set. The Bed & Breakfast column has a 0 for Alaska since Alaska does not have B & B lodging in this choice set. These binary variables will form the independent variables in the analysis.

Note that we are fitting a model with *generic attributes*. Generic attributes are assumed to be the same for all alternatives. For example, our model is structured so that the part-worth utility for being on a lake will be the same for Hawaii, Alaska, and all of the other destinations. Similarly, the part-worth utilities for the different prices will not depend on the destinations. In contrast, on page 171, using the same data, we will code alternative-specific effects where the part-worth utilities are allowed by the model to be different for each of the destinations.

PROC PHREG is run in the usual way to fit the choice model.

```
proc phreg data=coded brief;
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;
```

We specify the `&_trgind` macro variable for the `model` statement independent variable list. PROC TRANSREG automatically creates this macro variable. It contains the list of coded independent variables generated by the procedure. This is so you do not have to figure out what names TRANSREG created and specify them. In this case, PROC TRANSREG sets `&_trgind` to contain the following list.

```
PlaceHawaii PlaceAlaska PlaceMexico PlaceCalifornia PlaceMaine PlaceHome
Price0 Price999 Price1249 Price1499 SceneBeach SceneHome SceneLake
SceneMountains LodgeBed___Breakfast LodgeCabin LodgeHome LodgeHotel
```

The analysis is stratified by subject and choice set. Each stratum consists of a set of alternatives from which a subject made one choice. In this example, each stratum consists of six alternatives, one of which was chosen and five of which were not chosen. (Recall that we used `%phchoice(on)` on page 95 to customize the output from PROC PHREG.) The full table of the strata would be quite large with one line for each of the 3600 strata, so the `brief` option was specified on the PROC PHREG statement. This option produces a brief summary of the strata. In this case, we see there were 3600 choice sets that all fit one response pattern. Each consisted of 6 alternatives, 1 of which was chosen and 5 of which were not chosen. There should be one pattern for all choice sets in an example like this one − the number of alternatives, number of chosen alternatives, and the number not chosen should be constant.

Vacation Example

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	21600
Number of Observations Used	21600

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	3600	6	1	5

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	12900.668	6344.976
AIC	12900.668	6366.976
SBC	12900.668	6435.051

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6555.6923	11	<.0001
Score	5631.6233	11	<.0001
Wald	2472.1970	11	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Hawaii	1	3.54935	0.45628	60.5117	<.0001
Alaska	1	0.78414	0.46602	2.8312	0.0924
Mexico	1	2.77734	0.45831	36.7232	<.0001
California	1	2.22583	0.45901	23.5147	<.0001
Maine	1	1.39482	0.46227	9.1043	0.0026
Home	0	0	.	.	.
0	0	0	.	.	.
999	1	3.64557	0.09529	1463.7685	<.0001
1249	1	1.36008	0.08161	277.7719	<.0001
1499	0	0	.	.	.
Beach	1	1.44385	0.07158	406.9211	<.0001
Home	0	0	.	.	.
Lake	1	0.78337	0.08023	95.3493	<.0001
Mountains	0	0	.	.	.
Bed & Breakfast	1	0.61542	0.05554	122.7790	<.0001
Cabin	1	-1.45370	0.06818	454.6628	<.0001
Home	0	0	.	.	.
Hotel	0	0	.	.	.

The destinations, from most preferred to least preferred, are Hawaii, Mexico, California, Maine, Alaska, and then stay at home. The utility for lower price is greater than the utility for higher price. The beach is preferred over a lake, which is preferred over the mountains. A bed & breakfast is preferred over a hotel, which is preferred over a cabin. Notice that the coefficients for the constant alternative (home and zero price) are all zero. Also notice that for each factor, destination, price, scenery and accommodations, the coefficient for the last level is always zero. This will always occur when we code with `zero=none`. The last level of each factor is a reference level, and the other coefficients will have values relative to this zero. For example, all of the coefficients for the destination are positive relative to the zero for staying at home. For scenery, all of the coefficients are positive relative to the zero for the mountains. For accommodations, the coefficient for cabin is less than the zero for hotel, which is less than the coefficient for bed & breakfast. In some sense, each `class` variable in a choice model with a constant alternative has two reference levels or two levels that will always have a zero coefficient: the level corresponding to the constant alternative and the level corresponding to the last level. At first, it is reassuring to run the model with all levels represented to see that all the right levels get zeroed. Later, we will see ways to eliminate these levels from the output.

Quantitative Price Effect

These data can also be analyzed in a different way. The `Price` variable can be specified directly as a quantitative variable, instead of with indicator variables for a qualitative price effect. You could print the independent variable list and copy and edit it, removing the `Price` indicator variables and adding `Price`.

```
%put &_trgind;
```

Alternatively, you could run PROC TRANSREG again with the new coding. We use this latter approach, because it is easier, and it will allow us to illustrate other options. In the previous analysis, there were a number of structural-zeros in the parameter estimate results due to the usage of the `zero=none` option in the PROC TRANSREG coding. This is a good thing, particularly for a first attempt at the analysis. It is good to specify `zero=none` and check the results and make sure you have the right pattern of zeros and nonzeros. Later, you can run again excluding some of the structural zeros. This time, we will explicitly specify the `'Home'` level in the `zero=` option as the reference level so it will be omitted from the `&_trgind` variable list. The first `class` specification specifies `zero='Home'` since there is one variable. The second `class` specification specifies `zero='Home' 'Home'` specifying the reference level for each of the two variables. The variable `Price` is designated as an `identity` variable. The `identity` transformation is the no-transformation option, which is used for variables that need to enter the model with no further manipulations. The `identity` variables are simply copied into the output data set and added to the `&_trgind` variable list. The statement `label price = 'Price'` is specified to explicitly set a label for the `identity` variable price. This is because we explicitly modified PROC PHREG output using `%phchoice(on)` so that the rows of the parameter estimate table would be labeled only with variable labels not variable names. A label for `Price` must be explicitly specified in order for the output to contain a label for the price effect.

```
proc transreg design data=res2 nozeroconstant norestoremissing;
   model class(place / zero='Home' order=data) identity(price)
         class(scene lodge / zero='Home' 'Home' order=formatted) /
         lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   label price = 'Price';
   id subj set form c;
   run;

proc phreg data=coded brief;
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;
```

Here are the results.

Vacation Example

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	21600
Number of Observations Used	21600

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	3600	6	1	5

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	12900.668	6407.005
AIC	12900.668	6427.005
SBC	12900.668	6488.892

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6493.6633	10	<.0001
Score	5432.8215	10	<.0001
Wald	2485.3726	10	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Hawaii	1	14.72561	0.51256	825.3759	<.0001
Alaska	1	12.03600	0.50269	573.2784	<.0001
Mexico	1	13.95733	0.50734	756.8369	<.0001
California	1	13.41320	0.50709	699.6586	<.0001
Maine	1	12.65675	0.50650	624.4204	<.0001
Price	1	-0.00766	0.0001935	1567.3937	<.0001
Beach	1	1.46904	0.06967	444.5790	<.0001
Lake	1	0.74802	0.07753	93.0983	<.0001
Mountains	0	0	.	.	.
Bed & Breakfast	1	0.62936	0.05532	129.4274	<.0001
Cabin	1	-1.36891	0.06567	434.5382	<.0001
Hotel	0	0	.	.	.

The results of the two different analyses are similar. The coefficients for the destinations all increase by a nonconstant amount (approximately 11.21) but the pattern is the same. There is still a negative effect for price. Also, the fit of this model is slightly worse, Chi-Square = 6493.6633, compared to the previous value of 6555.6923 (bigger values mean better fit), because price has one fewer parameter.

Quadratic Price Effect

Previously, we saw price treated as a qualitative factor with two parameters and two *df*, then we saw price treated as a quantitative factor with one parameter and one *df*. Alternatively, we could treat price as quantitative and add a *quadratic* price effect (price squared). Like treating price as a qualitative factor, there are two parameters and two *df* for price. First, we create `PriceL`, the linear price term by centering the original price and dividing by the price increment (250). This maps (999, 1249, 1499) to (-1, 0, 1). Next, we run PROC TRANSREG and PROC PHREG with the new price variables.

```
data res3;
   set res2;
   PriceL = price;
   if price then pricel = (price - 1249) / 250;
   run;

proc transreg design=5000 data=res3 nozeroconstant norestoremissing;
   model class(place / zero='Home' order=data)
         pspline(pricel / degree=2)
         class(scene lodge / zero='Home' 'Home' order=formatted) /
         lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   label pricel = 'Price';
   id subj set form c;
   run;
```

The pspline or polynomial spline expansion with the degree=2 option replaces the variable PriceL with two coded variables, PriceL_1 (which is the same as the original PriceL) and PriceL_2 (which is PriceL squared). A degree=2 spline with no knots (neither knots= nor nknots= were specified) simply expands the variable into a quadratic polynomial.

```
proc phreg data=coded brief;
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;
```

This step produced the following results.

Vacation Example

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	21600
Number of Observations Used	21600

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	3600	6	1	5

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	12900.668	6344.976
AIC	12900.668	6366.976
SBC	12900.668	6435.051

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6555.6923	11	<.0001
Score	5631.6233	11	<.0001
Wald	2472.1970	11	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Hawaii	1	4.90943	0.45336	117.2666	<.0001
Alaska	1	2.14422	0.46144	21.5926	<.0001
Mexico	1	4.13742	0.45449	82.8741	<.0001
California	1	3.58591	0.45503	62.1037	<.0001
Maine	1	2.75491	0.45911	36.0068	<.0001
Price 1	1	-1.82278	0.04764	1463.7685	<.0001
Price 2	1	0.46270	0.05879	61.9512	<.0001
Beach	1	1.44385	0.07158	406.9211	<.0001
Lake	1	0.78337	0.08023	95.3493	<.0001
Mountains	0	0	.	.	.
Bed & Breakfast	1	0.61542	0.05554	122.7790	<.0001
Cabin	1	-1.45370	0.06818	454.6628	<.0001
Hotel	0	0	.	.	.

The fit is exactly the same as when price was treated as qualitative, Chi-Square = 6555.6923. This is because both models are the same except for the different but equivalent 2 *df* codings of price. The coefficients for the destinations in the two models differ by a constant 1.36008. The coefficients for the factors after price are unchanged. The part-worth utility for \$999 is $-1.82278 \times (999 - 1249)/250 + 0.46270 \times ((999 - 1249)/250)^2 = 2.28548$, the part-worth utility for \$1249 is $-1.82278 \times (1249 - 1249)/250 + 0.46270 \times ((1249 - 1249)/250)^2 = 0$, and the part-worth utility for \$1499 is $-1.82278 \times (1499-1249)/250+0.46270 \times ((1499-1249)/250)^2 = -1.36008$, which differ from the coefficients when price was treated as qualitative, by a constant -1.36008.

Effects Coding

In the previous analyses, *binary* (1, 0) codings were used for the variables. The next analysis illustrates *effects* (1, 0, -1) coding. The two codings differ in how the final reference level is coded. In binary coding, the reference level is coded with zeros. In effects coding, the reference level is coded with minus ones.

	Binary Coding		Effects Coding	
Levels	One	Two	One	Two
1	1	0	1	0
2	0	1	0	1
3	0	0	-1	-1

In this example, we will use a binary coding for the destinations and effects codings for the attributes.

PROC TRANSREG can be used for effects coding. The `effects` option used inside the parentheses after `class` asks for a (0, 1, -1) coding. The `zero=` option specifies the levels that receive the -1's. PROC PHREG is run with almost the same variable list as before, except now the variables for the reference levels, those whose parameters are structural zeros are omitted. Refer back to the parameter estimates table on page 162, a few select lines of which are reproduced next:

(Some Lines in the)
Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Home	0	0	.	.	.
0	0	0	.	.	.
1499	0	0	.	.	.
Home	0	0	.	.	.
Mountains	0	0	.	.	.
Home	0	0	.	.	.
Hotel	0	0	.	.	.

Notice that the coefficients for the constant alternative (home and zero price) are all zero. Also notice that for each factor, destination, price, scenery and accommodations, the coefficient for the last level is always zero. In some sense, each `class` variable in a choice model with a constant alternative has two reference levels or two levels that will always have a zero coefficient: the level corresponding to the constant alternative and the level corresponding to the last level. In some of the preceding examples, we eliminated the `'Home'` levels by specifying `zero=Home`. Now we will see how to eliminate all of the structural zeros from the parameter estimate table.

First, for each classification variable, we change the level for the constant alternative to missing. (Recall that they were originally missing and we only made them nonmissing to deliberately produce the zero coefficients.) This will cause PROC TRANSREG to ignore those levels when constructing indicator variables. When you use this strategy, you must specify the `norestoremissing` option in the PROC

TRANSREG statement. During the first stage of design matrix creation, PROC TRANSREG puts zeros in the indicator variables for observations with missing `class` levels. At the end, it replaces the zeros with missings, "restoring the missing values." When the `norestoremissing` option is specified, missing values are not restored and we get zeros in the indicator variables for missing `class` levels. The DATA step `if` statements recode the constant levels to missing. Next, in PROC TRANSREG, the reference levels `'Mountains'` and `'Hotel'` are listed in the `zero=` option in the `class` specification.

```
data res4;
   set res3;
   if scene = 0 then scene = .;
   if lodge = 0 then lodge = .;
   run;

proc transreg design=5000 data=res4 nozeroconstant norestoremissing;
   model class(place / zero='Home' order=data)
         pspline(pricel / degree=2)
         class(scene lodge /
               effects zero='Mountains' 'Hotel' order=formatted) /
         lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   label pricel = 'Price';
   id subj set form c;
   run;
```

The coded data and design matrix are printed for the first choice set. The coded design matrix begins with five binary columns for the destinations, `'Hawaii'` through `'Maine'`. There is not a column for the stay-at-home destination and the row for stay at home has all zeros in the coded variables. Next is the linear price effect, `'Price 1'`, consisting of 0, 1, and -1. It is followed by the quadratic price effect, `'Price 2'`, which is `'Price 1'` squared. Next are the scenery terms, effects coded. `'Beach'` and `'Lake'` have values of 0 and 1; -1's in the fourth row for the reference level, `'Mountains'`; and zeros in the last row for the stay-at-home alternative. Next are the lodging terms, effects coded. `'Bed & Breakfast'` and `'Cabin'` have values of 0 and 1; -1's in the first, third and fourth row for the reference level, `'Hotel'`; and zeros in the last row for the stay-at-home alternative.

```
proc print data=coded(obs=6) label; run;
```

Vacation Example

Obs	Hawaii	Alaska	Mexico	California	Maine	Price 1	Price 2	Beach	Lake	Bed & Breakfast
1	1	0	0	0	0	-1	1	0	1	0
2	0	1	0	0	0	0	0	0	1	0
3	0	0	1	0	0	1	1	-1	-1	-1
4	0	0	0	1	0	1	1	0	1	1
5	0	0	0	0	1	1	1	-1	-1	0
6	0	0	0	0	0	0	0	0	0	0

Obs	Cabin	Place	Price	Scene	Lodge	Subj	Set	Form	c
1	1	Hawaii	-1	Lake	Cabin	1	1	1	1
2	1	Alaska	0	Lake	Cabin	1	1	1	2
3	-1	Mexico	1	Mountains	Hotel	1	1	1	2
4	0	California	1	Lake	Bed & Breakfast	1	1	1	2
5	1	Maine	1	Mountains	Cabin	1	1	1	2
6	0	Home	0	.	.	1	1	1	2

```
proc phreg data=coded brief;
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;
```

Vacation Example

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	21600
Number of Observations Used	21600

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	3600	6	1	5

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	12900.668	6344.976
AIC	12900.668	6366.976
SBC	12900.668	6435.051

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6555.6923	11	<.0001
Score	5631.6233	11	<.0001
Wald	2472.1970	11	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Hawaii	1	5.37241	0.45045	142.2507	<.0001
Alaska	1	2.60720	0.45694	32.5563	<.0001
Mexico	1	4.60040	0.45097	104.0650	<.0001
California	1	4.04889	0.45189	80.2794	<.0001
Maine	1	3.21788	0.45464	50.0964	<.0001
Price 1	1	-1.82278	0.04764	1463.7685	<.0001
Price 2	1	0.46270	0.05879	61.9512	<.0001
Beach	1	0.70144	0.03720	355.5596	<.0001
Lake	1	0.04097	0.04268	0.9214	0.3371
Bed & Breakfast	1	0.89485	0.03559	632.2530	<.0001
Cabin	1	-1.17428	0.04228	771.3878	<.0001

It is instructive to compare the results of this analysis to the previous analysis on page 166. First, the model fit and chi-square statistics are the same indicating the models are equivalent. The coefficients for the destinations differ by a constant 0.46298, the price coefficients are the same, the scenery coefficients differ by a constant -0.7424, and the lodging coefficients differ by a constant 0.27943. Notice that $0.46298 + 0 + -0.74241 + 0.27943 = 0$, so the utility for each alternative is unchanged by the different but equivalent codings.

Alternative-Specific Effects

In all of the analyses presented so far in this example, we have assumed that the effects for price, scenery, and accommodations are generic or constant across the different destinations. Equivalently, we assumed that destination does not interact with the attributes. Next, we show a model with *alternative-specific effects* that does not make this assumption. Our new model allows for different price, scenery and lodging effects for each destination. The coding can be done with PROC TRANSREG and its syntax for interactions. Before we do the coding, let's go back to the design preparation stage and redo it in a more normal fashion so reference levels will be omitted from the analysis.

We start by creating the data set KEY. This step differs from the one we saw on page 155 only in that now we have a missing value for `Place` for the constant alternative.

```
data key;
   input Place $ 1-10 (Lodge Scene Price) ($);
   datalines;
Hawaii      x1  x6   x11
Alaska      x2  x7   x12
Mexico      x3  x8   x13
California  x4  x9   x14
Maine       x5  x10  x15
.           .   .    .
;
```

Next, we use the `%MktRoll` macro to process the design and the `%MktMerge` macro to merge the design and data.

```
%mktroll(design=sasuser.blockdes, key=key, alt=place, out=rolled)

%mktmerge(design=rolled, data=results, out=res2, blocks=form,
          nsets=&n, nalts=&m, setvars=choose1-choose&n,
          stmts=%str(price = input(put(price, price.), 5.);
                     format scene scene. lodge lodge.;))

proc print data=res2(obs=12); run;
```

The usage of the `%MktRoll` macro is exactly the same as we saw on page 155. The `%MktMerge` macro usage differs from page 157 in that instead of assigning labels and recoding price in a separate DATA step, we now do it directly in the macro. The `stmts=` option is used to add a `price =` assignment statement and `format` statement to the DATA step that merges the two data sets. The statements were included in a `%str( )` macro since they contain semicolons. Here are the first two choice sets.

Vacation Example

Obs	Subj	Form	Set	Place	Lodge	Scene	Price	c
1	1	1	1	Hawaii	Cabin	Lake	999	1
2	1	1	1	Alaska	Cabin	Lake	1249	2
3	1	1	1	Mexico	Hotel	Mountains	1499	2
4	1	1	1	California	Bed & Breakfast	Lake	1499	2
5	1	1	1	Maine	Cabin	Mountains	1499	2
6	1	1	1		.	.	.	2
7	1	1	2	Hawaii	Hotel	Mountains	1499	2
8	1	1	2	Alaska	Hotel	Lake	1499	2
9	1	1	2	Mexico	Hotel	Beach	1499	1
10	1	1	2	California	Hotel	Mountains	1499	2
11	1	1	2	Maine	Bed & Breakfast	Beach	1499	2
12	1	1	2		.	.	.	2

Notice that the attributes for the constant alternative are all missing. Next, we code with PROC TRANSREG. Since we are using missing values for the constant alternative, we must specify the `norestoremissing` option in the PROC TRANSREG statement. When you specify the `norestoremissing` option, the indicator variables created for missing `class` variable values contain all zeros instead of all missings. First, we specify the variable `Place` as a `class` variable. Next, we interact `Place` with all of the attributes, `Price`, `Scene`, and `Lodge`, to create the alternative-specific effects.

```
proc transreg design=5000 data=res2 nozeroconstant norestoremissing;
   model class(place / zero=none order=data)
         class(place * price place * scene place * lodge /
               zero=none order=formatted) / lprefix=0 sep=' ' ', ';
   output out=coded(drop=_type_ _name_ intercept);
   id subj set form c;
   run;
proc print data=coded(obs=6) label noobs; run;
```

The coded design matrix consists of:

- five binary columns, `'Hawaii'` through `'Maine'`, for the five destinations,
- fifteen binary columns (5 destinations times 3 prices), `'Alaska, 999'` through `'Mexico, 1499'`, for the alternative-specific price effects,
- fifteen binary columns (5 destinations times 3 sceneries), `'Alaska, Beach'` through `'Mexico, Mountains'`, for the alternative-specific scenery effects,
- fifteen binary columns (5 destinations times 3 lodgings), `'Alaska, Bed & Breakfast'` through `'Mexico, Hotel'`, for the alternative-specific lodging effects.

The entire sixth row of the coded design matrix, the stay-at-home alternative, consists of zeros.

Vacation Example

Hawaii	Alaska	Mexico	California	Maine	Alaska, 999	Alaska, 1249	Alaska, 1499
1	0	0	0	0	0	0	0
0	1	0	0	0	0	1	0
0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0

California, 999	California, 1249	California, 1499	Hawaii, 999	Hawaii, 1249	Hawaii, 1499	Maine, 999	Maine, 1249	Maine, 1499
0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0

Mexico, 999	Mexico, 1249	Mexico, 1499	Alaska, Beach	Alaska, Lake	Alaska, Mountains	California, Beach	California, Lake	California, Mountains
0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Hawaii, Beach	Hawaii, Lake	Hawaii, Mountains	Maine, Beach	Maine, Lake	Maine, Mountains	Mexico, Beach	Mexico, Lake	Mexico, Mountains
0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0

Alaska, Bed & Breakfast	Alaska, Cabin	Alaska, Hotel	California, Bed & Breakfast	California, Cabin	California, Hotel
0	0	0	0	0	0
0	1	0	0	0	0
0	0	0	0	0	0
0	0	0	1	0	0
0	0	0	0	0	0
0	0	0	0	0	0

Hawaii, Bed & Breakfast	Hawaii, Cabin	Hawaii, Hotel	Maine, Bed & Breakfast	Maine, Cabin	Maine, Hotel	Mexico, Bed & Breakfast	Mexico, Cabin	Mexico, Hotel
0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0

Place	Price	Scene	Lodge	Subj	Set	Form	c
Hawaii	999	Lake	Cabin	1	1	1	1
Alaska	1249	Lake	Cabin	1	1	1	2
Mexico	1499	Mountains	Hotel	1	1	1	2
California	1499	Lake	Bed & Breakfast	1	1	1	2
Maine	1499	Mountains	Cabin	1	1	1	2
	.	.	.	1	1	1	2

Analysis proceeds by running PROC PHREG as before.

```
proc phreg data=coded brief;
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;
```

Vacation Example

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	21600
Number of Observations Used	21600

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	3600	6	1	5

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	12900.668	6311.908
AIC	12900.668	6381.908
SBC	12900.668	6598.512

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6588.7600	35	<.0001
Score	6160.3313	35	<.0001
Wald	2346.1138	35	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Hawaii	1	3.74693	0.46426	65.1372	<.0001
Alaska	1	0.46811	0.66017	0.5028	0.4783
Mexico	1	2.58269	0.48906	27.8883	<.0001
California	1	2.17449	0.49548	19.2603	<.0001
Maine	1	0.78218	0.52903	2.1860	0.1393
Alaska, 999	1	4.27617	0.42448	101.4861	<.0001
Alaska, 1249	1	1.59027	0.46383	11.7551	0.0006
Alaska, 1499	0	0	.	.	.
California, 999	1	3.76692	0.17969	439.4715	<.0001
California, 1249	1	1.29591	0.18948	46.7767	<.0001
California, 1499	0	0	.	.	.
Hawaii, 999	1	3.44903	0.13117	691.3525	<.0001
Hawaii, 1249	1	1.16508	0.12842	82.3111	<.0001
Hawaii, 1499	0	0	.	.	.
Maine, 999	1	4.17560	0.25328	271.7974	<.0001
Maine, 1249	1	1.87079	0.27427	46.5261	<.0001
Maine, 1499	0	0	.	.	.
Mexico, 999	1	3.70719	0.16564	500.9042	<.0001
Mexico, 1249	1	1.56587	0.16249	92.8695	<.0001
Mexico, 1499	0	0	.	.	.
Alaska, Beach	1	1.25655	0.24081	27.2281	<.0001
Alaska, Lake	1	0.68725	0.24822	7.6656	0.0056
Alaska, Mountains	0	0	.	.	.
California, Beach	1	1.49280	0.14678	103.4326	<.0001
California, Lake	1	0.68987	0.16263	17.9940	<.0001
California, Mountains	0	0	.	.	.
Hawaii, Beach	1	1.45943	0.11718	155.1081	<.0001
Hawaii, Lake	1	0.81501	0.12575	42.0042	<.0001
Hawaii, Mountains	0	0	.	.	.
Maine, Beach	1	1.60677	0.20249	62.9642	<.0001
Maine, Lake	1	0.79942	0.20513	15.1884	<.0001
Maine, Mountains	0	0	.	.	.
Mexico, Beach	1	1.66083	0.15218	119.1038	<.0001
Mexico, Lake	1	0.89927	0.15165	35.1617	<.0001
Mexico, Mountains	0	0	.	.	.

Alaska, Bed & Breakfast	1	0.50806	0.18663	7.4104	0.0065
Alaska, Cabin	1	-1.95733	0.33673	33.7882	<.0001
Alaska, Hotel	0	0	.	.	.
California, Bed & Breakfast	1	0.65346	0.13179	24.5850	<.0001
California, Cabin	1	-1.43215	0.17478	67.1409	<.0001
California, Hotel	0	0	.	.	.
Hawaii, Bed & Breakfast	1	0.48132	0.11148	18.6408	<.0001
Hawaii, Cabin	1	-1.43085	0.12552	129.9497	<.0001
Hawaii, Hotel	0	0	.	.	.
Maine, Bed & Breakfast	1	0.87057	0.17456	24.8723	<.0001
Maine, Cabin	1	-1.61015	0.22579	50.8554	<.0001
Maine, Hotel	0	0	.	.	.
Mexico, Bed & Breakfast	1	0.59476	0.12173	23.8715	<.0001
Mexico, Cabin	1	-1.53314	0.15499	97.8448	<.0001
Mexico, Hotel	0	0	.	.	.

There are zero coefficients for the reference level. Do we need this more complicated model instead of the simpler model? To answer this, first look at the coefficients. Are they similar across different destinations? In this case, they seem to be. This suggests that the simpler model may be sufficient.

More formally, the two models can be statistically compared. You can test the null hypothesis that the two models are not significantly different by comparing their likelihoods. The difference between two $-2\log(\mathcal{L}_C)$'s (the number reported under 'With Covariates' in the output) has a chi-square distribution. We can get the *df* for the test by subtracting the two *df* for the two likelihoods. The difference $6588.76 - 6555.6923 = 33.0677$ is distributed χ^2 with $35 - 11 = 24$ *df* ($p < 0.10265$). This more complicated model does not account for significantly more variance than the simpler model.

Vacation Example, with Alternative-Specific Attributes

This example discusses choosing the number of choice sets, designing the choice experiment, ensuring that certain key interactions are estimable, examining the design, blocking an existing design, generating the questionnaire, generating artificial data, reading, processing, and analyzing the data, binary coding, generic attributes, alternative-specific effects, aggregating the data, analysis, and interpretation of the results. In this example, a researcher is interested in studying choice of vacation destinations. Here and on the next page are two summaries of the design, one with factors grouped by attribute and one grouped by destination.

Factor	Destination	Attribute	Levels
X1	Hawaii	Accommodations	Cabin, Bed & Breakfast, Hotel
X2	Alaska	Accommodations	Cabin, Bed & Breakfast, Hotel
X3	Mexico	Accommodations	Cabin, Bed & Breakfast, Hotel
X4	California	Accommodations	Cabin, Bed & Breakfast, Hotel
X5	Maine	Accommodations	Cabin, Bed & Breakfast, Hotel
X6	Hawaii	Scenery	Mountains, Lake, Beach
X7	Alaska	Scenery	Mountains, Lake, Beach
X8	Mexico	Scenery	Mountains, Lake, Beach
X9	California	Scenery	Mountains, Lake, Beach
X10	Maine	Scenery	Mountains, Lake, Beach
X11	Hawaii	Price	$1249, $1499, $1749
X12	Alaska	Price	$1249, $1499, $1749
X13	Mexico	Price	$999, $1249, $1499
X14	California	Price	$999, $1249, $1499, $1749
X15	Maine	Price	$999, $1249, $1499
X16	Hawaii	Side Trip	Yes, No
X17	Mexico	Side Trip	Yes, No

This example is a modification of the previous example. Now, all alternatives do not have the same factors, and all factors do not have the same numbers of levels. There are still five destinations of interest: Hawaii, Alaska, Mexico, California, and Maine. Each alternative is composed of three factors like before: package cost, scenery, and accommodations, only now they do not all have the same levels, and the Hawaii and Mexico alternatives are composed of one additional attribute. For Hawaii and Alaska, the costs are $1,249, $1,499, and $1,749; for California, the prices are $999, $1,249, $1,499, and $1,749; and for Mexico and Maine, the prices are $999, $1,249, and $1,499. Scenery (mountains, lake, beach) and accommodations (cabin, bed & breakfast, and hotel) are the same as before. The Mexico trip now has the option of a side trip to sites of archaeological significance, via bus, for an additional cost of $100. The Hawaii trip has the option of a side trip to an active volcano, via helicopter, for an additional cost of $200. This is typical of the problems that marketing researchers face. We have lots of factors and *asymmetry* – each alternative is not composed of the same factors, and the factors do not all have the same numbers of levels.

Factor	Destination	Attribute	Levels
X1	Hawaii	Accommodations	Cabin, Bed & Breakfast, Hotel
X6		Scenery	Mountains, Lake, Beach
X11		Price	$1249, $1499, $1749
X16		Side Trip	Yes, No
X2	Alaska	Accommodations	Cabin, Bed & Breakfast, Hotel
X7		Scenery	Mountains, Lake, Beach
X12		Price	$1249, $1499, $1749
X3	Mexico	Accommodations	Cabin, Bed & Breakfast, Hotel
X8		Scenery	Mountains, Lake, Beach
X13		Price	$999, $1249, $1499
X17		Side Trip	Yes, No
X4	California	Accommodations	Cabin, Bed & Breakfast, Hotel
X9		Scenery	Mountains, Lake, Beach
X14		Price	$999, $1249, $1499, $1749
X5	Maine	Accommodations	Cabin, Bed & Breakfast, Hotel
X10		Scenery	Mountains, Lake, Beach
X15		Price	$999, $1249, $1499

Choosing the Number of Choice Sets

We can use the `%MktRuns` autocall macro to suggest experimental design sizes. (All of the autocall macros used in this book are documented starting on page 479.) As before, we specify a list containing the number of levels of each factor.

```
title 'Vacation Example with Asymmetry';

%mktruns( 3 ** 14 4 2 2 )
```

The output tells us the size of the saturated design, which is the number of parameters in the linear design, and suggests design sizes.

Vacation Example with Asymmetry

Design Summary

Number of Levels	Frequency
2	2
3	14
4	1

```
                Vacation Example with Asymmetry

     Saturated      = 34
     Full Factorial = 76,527,504

     Some Reasonable                        Cannot Be
        Design Sizes       Violations       Divided By

                  72 *              0
                 144 *              0
                  36                2      8
                 108                2      8
                  54               18      4  8 12
                  90               18      4  8 12
                 126               18      4  8 12
                  45               48      2  4  6  8 12
                  63               48      2  4  6  8 12
                  81               48      2  4  6  8 12

  * - 100% Efficient Design can be made with the MktEx Macro.

                Vacation Example with Asymmetry

  n                    Design                        Reference

  72    2 ** 20  3 ** 24  4 **  1                 Orthogonal Array
  72    2 ** 19  3 ** 20  4 **  1  6 **  1        Orthogonal Array
  72    2 ** 18  3 ** 16  4 **  1  6 **  2        Orthogonal Array
  72    2 ** 13  3 ** 25  4 **  1                 Orthogonal Array
  72    2 ** 12  3 ** 21  4 **  1  6 **  1        Orthogonal Array
  72    2 ** 11  3 ** 24  4 **  1  6 **  1        Orthogonal Array
  72    2 ** 11  3 ** 17  4 **  1  6 **  2        Orthogonal Array
  72    2 ** 10  3 ** 20  4 **  1  6 **  2        Orthogonal Array
  72    2 **  9  3 ** 16  4 **  1  6 **  3        Orthogonal Array
 144    2 ** 92  3 ** 24  4 **  1                 Orthogonal Array
 144    2 ** 91  3 ** 20  4 **  1  6 **  1        Orthogonal Array
 144    2 ** 90  3 ** 16  4 **  1  6 **  2        Orthogonal Array
 144    2 ** 85  3 ** 25  4 **  1                 Orthogonal Array
 144    2 ** 84  3 ** 21  4 **  1  6 **  1        Orthogonal Array
 144    2 ** 83  3 ** 24  4 **  1  6 **  1        Orthogonal Array
 144    2 ** 83  3 ** 17  4 **  1  6 **  2        Orthogonal Array
 144    2 ** 82  3 ** 20  4 **  1  6 **  2        Orthogonal Array
 144    2 ** 81  3 ** 16  4 **  1  6 **  3        Orthogonal Array
```

We need at least 34 choice sets, as shown by `(Saturated=34)` in the listing. Any size that is a multiple of 72 would be optimal. We would recommend 72 choice sets, four blocks of size 18. However, like the previous vacation example, we will use fewer choice sets so that we can illustrate getting an efficient but nonorthogonal design. A design with 36 choice sets is pretty good. Thirty-six is not divisible by $8 = 2 \times 4$, so we cannot have equal frequencies in the California price and Mexico and Hawaii side trip

combinations. This should not pose any problem. This leaves only 2 error *df* for the linear model, but in the choice model, we will have adequate error *df*.

Designing the Choice Experiment

This problem requires a design with 1 four-level factor for price and 4 three-level factors for price. There are 10 three-level factors for scenery and accommodations as before. Also, we need 2 two-level factors for the two side trips. Note that we do not need a factor for the price or mode of transportation of the side trips since they are constant within each trip. With the `%MktEx` macro, making an asymmetric design is no more difficult than making a symmetric design.

```
%mktex(3 ** 13 4 3 2 2, n=36, seed=205)
%mkteval;
```

Here is the last part of the results.

Vacation Example with Asymmetry

The OPTEX Procedure

Class Level Information

Class	Levels	-Values-
x1	3	1 2 3
x2	3	1 2 3
x3	3	1 2 3
x4	3	1 2 3
x5	3	1 2 3
x6	3	1 2 3
x7	3	1 2 3
x8	3	1 2 3
x9	3	1 2 3
x10	3	1 2 3
x11	3	1 2 3
x12	3	1 2 3
x13	3	1 2 3
x14	4	1 2 3 4
x15	3	1 2 3
x16	2	1 2
x17	2	1 2

Vacation Example with Asymmetry

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	98.8874	97.5943	97.4925	0.9718

Vacation Example with Asymmetry
Canonical Correlations Between the Factors
There are 2 Canonical Correlations Greater Than 0.316

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17
x1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
x5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
x6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
x7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
x8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
x9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
x10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
x11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
x12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
x13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.25	0	0
x14	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.33	0.33
x15	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0	1	0	0
x16	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33	0	1	0
x17	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33	0	0	1

The macro found a very nice, almost orthogonal and almost 99% efficient design in 40 seconds. However, we will not use this design. Instead, we will make a larger design with interactions.

Ensuring that Certain Key Interactions are Estimable

Next, we will ensure that certain key interactions are estimable. Specifically, it would be good if in the aggregate, the interactions between price and accommodations were estimable for each destination. We would like the following interactions to be estimable: `x1*x11 x2*x12 x3*x13 x4*x15 x5*x15`. We will again use the `%MktEx` macro.

```
%mktex(3 ** 13 4 3 2 2, n=36,
       interact=x1*x11 x2*x12 x3*x13 x4*x14 x5*x15,
       seed=205)
```

We immediately get this message.

```
ERROR: More parameters than runs.
       If you really want to do this, specify RIDGE=.
       There are 36 runs with 56 parameters.
ERROR: The MKTEX macro ended abnormally.
```

If we want interactions to be estimable, we will need more choice sets. The number of parameters is 1 for the intercept, $14 \times (3-1) + (4-1) + 2 \times (2-1) = 33$ for main effects, and $4 \times (3-1) \times (3-1) + (4-1) \times (3-1) = 22$ for interactions for a total of $1 + 33 + 22 = 56$ parameters. This means we need at least 56 choice sets, and ideally for this design with 2, 3, and 4 level factors, we would like the number of sets to be divisible by 2×2, 2×3, 2×4, 3×3, and 3×4. Sixty is divisible by 2, 3, 4, 6, and 12 so is a reasonable design size. Sixty choice sets could be divided into three blocks of size 20, four blocks of size 15, or five blocks of size 12. Seventy-two choice sets would be better, since unlike 60, 72 can be divided by 9. Unfortunately, 72 would require one more block.

We can also run the `%MktRuns` macro to help us choose the number of choice sets. However, the `%MktRuns` does not have a special syntax for interactions, you have to specify the main effects and interactions of two factors as if it were a single factor. For example, for the interaction of 2 three-level factors, you specify 9 in the list. For the interaction of a three-level factor and a four-level factor, you specify 12 in the list. Do not specify '3 3 9' or '3 4 12'; just specify '3' and '12'. In this example, we specify four 9's for the four accommodation/price interactions involving only three-level factors, one 12 for the California accommodation/price interaction, five 3's for scenery, and two 2's for the side trips. We also specified a keyword option `max=` to consider only the 45 design sizes from the minimum of 56 up to 100.

```
title 'Vacation Example with Asymmetry';

%mktruns(9 9 9 9 12 3 3 3 3 3 2 2, max=45)
```

Vacation Example with Asymmetry

Design Summary

Number of Levels	Frequency
2	2
3	5
9	4
12	1

Vacation Example with Asymmetry

```
Saturated      = 56
Full Factorial = 76,527,504
```

Some Reasonable Design Sizes	Violations	Cannot Be Divided By								
72	30	27	81	108						
81	33	2	4	6	12	18	24	36	108	
90	39	4	12	24	27	36	81	108		
96	57	9	18	27	36	81	108			
60	59	9	18	24	27	36	81	108		
63	59	2	4	6	12	18	24	27	36	81 108
84	59	9	18	24	27	36	81	108		
99	59	2	4	6	12	18	24	27	36	81 108
66	61	4	9	12	18	24	27	36	81	108
78	61	4	9	12	18	24	27	36	81	108

We see that 72 cannot be divided by $27 = 9 \times 3$ so for example the Maine accommodation/price combinations cannot occur with equal frequency with each of the three-level factors. We see that 72 cannot be divided by $81 = 9 \times 9$ so for example the Mexico accommodation/price combinations cannot occur with equal frequency with each of the Hawaii accommodation/price combinations. We see that 72 cannot be divided by $108 = 9 \times 12$ so for example the California accommodation/price combinations cannot occur with equal frequency with each of the Maine accommodation/price combinations. With interactions, there are many higher-order opportunities for nonorthogonality. However, usually we will not be overly concerned about potential unequal cell frequencies on combinations of attributes in different alternatives.

The smallest number of runs in the table is 60. While 72 is better in that it can be divided by more numbers, either 72 or 60 should work fine. We will pick the larger number and run the `%MktEx` macro again with `n=72` specified.

```
%mktex(3 ** 13 4 3 2 2, n=72, seed=368,
       interact=x1*x11 x2*x12 x3*x13 x4*x14 x5*x15)
```

The macro printed these notes to the log.

```
NOTE: Generating the candidate set.
NOTE: Performing 20 searches of 243 candidates, full-factorial=76,527,504.
NOTE: Generating the orthogonal array design, n=72.
```

The candidate-set search is using a fractional-factorial candidate set with $3^5 = 243$ candidates. The two-level factors in the candidate set are made from three-level factors by coding down. *Coding down* replaces an m-level factor with a factor with fewer than m levels, for example a two-level factor could be created from a three-level factor: $((1\ 2\ 3) \Rightarrow (1\ 2\ 1))$. The four-level factor in the candidate set is made from 2 three-level factors and coding down. $((1\ 2\ 3) \times (1\ 2\ 3) \Rightarrow (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9) \Rightarrow (1\ 2\ 3\ 4\ 1\ 2\ 3\ 4\ 1))$. The tabled design used for the partial initialization in the coordinate-exchange steps has 72 runs. Here are some of the results.

Vacation Example with Asymmetry

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	84.8774	84.8774	Can
1	End	84.8774		
2	Start	59.2427		Tab
2	35 1	84.8786	84.8786	
2	35 13	84.8989	84.8989	
2	36 15	84.9311	84.9311	
2	39 13	84.9506	84.9506	
2	45 5	84.9835	84.9835	
2	47 1	85.0125	85.0125	
2	47 5	85.0377	85.0377	
2	48 15	85.0422	85.0422	
2	49 5	85.0514	85.0514	
2	58 13	85.0854	85.0854	
2	61 3	85.1230	85.1230	
2	64 2	85.1699	85.1699	
2	15 4	85.1913	85.1913	
2	17 12	85.2029	85.2029	
2	39 1	85.2635	85.2635	
2	39 13	85.2714	85.2714	
2	43 12	85.2730	85.2730	
2	55 2	85.2758	85.2758	
2	55 12	85.3610	85.3610	
2	58 12	85.3972	85.3972	
2	43 2	85.4009	85.4009	
2	55 2	85.4463	85.4463	
2	69 2	85.4590	85.4590	
2	3 3	85.4733	85.4733	
2	17 12	85.4950	85.4950	
2	29 5	85.4956	85.4956	
2	49 5	85.5033	85.5033	
2	59 15	85.5195	85.5195	
2	65 5	85.5611	85.5611	
2	65 15	85.6228	85.6228	
2	67 2	85.6288	85.6288	
2	15 4	85.6392	85.6392	
2	17 2	85.6499	85.6499	
2	29 5	85.6701	85.6701	
2	33 1	85.6813	85.6813	
2	29 1	85.7072	85.7072	
2	49 5	85.7076	85.7076	
2	End	85.7076		

```
 3      Start       59.2427                Tab
 3     49   5       85.7076       85.7076
 3        End       85.7076

 4      Start       59.2427                Tab
 4     49   5       85.7076       85.7076
 4        End       85.7076
 .
 .
 .
11      Start       59.2427                Tab
11     49   5       85.7076       85.7076
11        End       85.7076

12      Start       55.7719                Ran,Mut,Ann
12     51   1       85.7134       85.7134
12     51   7       85.7200       85.7200
 .
 .
 .
12     41   9       89.6452       89.6452
12        End       89.6452
 .
 .
 .
21      Start       60.4950                Ran,Mut,Ann
21        End       89.5099
```

Vacation Example with Asymmetry

Design Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	89.6452	89.6452	Ini
1	Start	58.1158		Ran,Mut,Ann
1	End	89.2388		
2	Start	58.9937		Ran,Mut,Ann
2	40 9	89.6514	89.6514	
2	47 6	89.6517	89.6517	
.				
.				
.				

20	Start	58.2187		Ran,Mut,Ann
20	60 10	89.7016	89.7016	
20	68 10	89.7039	89.7039	
.				
.				
.				
20	64 14	90.0208	90.0208	
20	End	90.0208		
.				
.				
.				
25	Start	54.6536		Ran,Mut,Ann
25	End	89.1925		

Vacation Example with Asymmetry

Design Refinement History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	90.0208	90.0208	Ini
1	Start	87.8506		Pre,Mut,Ann
1	End	89.5491		
.				
.				
.				
10	Start	88.7072		Pre,Mut,Ann
10	End	89.4804		

Vacation Example with Asymmetry

The OPTEX Procedure

Vacation Example with Asymmetry

The OPTEX Procedure

Class Level Information

Class	Levels	-Values-
x1	3	1 2 3
x2	3	1 2 3
x3	3	1 2 3
x4	3	1 2 3
x5	3	1 2 3
x6	3	1 2 3
x7	3	1 2 3
x8	3	1 2 3
x9	3	1 2 3
x10	3	1 2 3
x11	3	1 2 3
x12	3	1 2 3
x13	3	1 2 3
x14	4	1 2 3 4
x15	3	1 2 3
x16	2	1 2
x17	2	1 2

Vacation Example with Asymmetry

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	90.0208	79.9719	95.1716	0.8819

The macro ran in just under 8 minutes. The algorithm search history shows that the candidate-set approach (`Can`) used in design 1 found a design that was 84.8774% efficient. The macro makes no attempt to improve on this design, unless there are restriction on the design, until the end in the design refinement step, and only if it is the best design found.

Designs 2 through 11 used the coordinate-exchange algorithm with a tabled design initialization (`Tab`). This process found a design that was 85.7076% efficient. For this problem, the tabled design initialization initializes all 72 rows; For other problems, when the number of runs in the design is greater than the number of runs in the nearest tabled design, the remaining rows would be randomly initialized. The tabled design initialization usually works very well when all but at most a very few rows and columns are left uninitialized and there are no interactions or restrictions. That is not the case in this problem, and when the algorithm switches to a fully random initialization in design 12, it immediately does better.

The algorithm search phase picked the coordinate-exchange algorithm with a random initialization, random mutations, and simulated annealing as the algorithm to use in the next step, the design search step. The design search history is initialized with the best design (D-efficiency = 89.6452) found so far. The design search phase starts out with the initial design (`Ini`) found in the algorithm search phase. In iteration 20, the macro finds a design with D-efficiency = 90.0208.

The final set of iterations tries to improve the best design found so far. Random mutations (`Ran`), simulated annealing (`Ann`), and level exchanges are used on the previous best (`Pre`) design. The random mutations are responsible for making the efficiency of the starting design worse than the previous best efficiency. In this case, the design refinement step was unable to improve the best design found by the design search step.

Examining the Design

We can use the `%MktEval` macro to evaluate the goodness of this design.

```
%mkteval(data=design);
```

Here are some of the results.

Vacation Example with Asymmetry
Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

	x1	x2	x3	x4	x5	x6	x7	x8	x9
x1	1	0.07	0.13	0.17	0.07	0.03	0.10	0.06	0.13
x2	0.07	1	0.15	0.17	0.10	0.07	0.13	0.14	0.10
x3	0.13	0.15	1	0.15	0.08	0.09	0.10	0.14	0.09
x4	0.17	0.17	0.15	1	0.09	0.10	0.09	0.03	0.11
x5	0.07	0.10	0.08	0.09	1	0.09	0.10	0.07	0.11
x6	0.03	0.07	0.09	0.10	0.09	1	0.11	0.07	0.09
x7	0.10	0.13	0.10	0.09	0.10	0.11	1	0.10	0.06
x8	0.06	0.14	0.14	0.03	0.07	0.07	0.10	1	0.03
x9	0.13	0.10	0.09	0.11	0.11	0.09	0.06	0.03	1

.
.
.

```
            Vacation Example with Asymmetry
                 Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316
          * - Indicates Unequal Frequencies

                 Frequencies

*    x1          25 24 23
*    x2          26 24 22
*    x3          23 26 23
*    x4          24 23 25
*    x5          23 25 24
*    x6          25 24 23
*    x7          25 24 23
*    x8          26 22 24
*    x9          24 25 23
*    x10         26 24 22
*    x11         26 23 23
*    x12         23 23 26
*    x13         25 23 24
*    x14         18 17 18 19
*    x15         24 25 23
*    x16         42 30
     x17         36 36
*    x1 x2       9 9 7 9 8 7 8 7 8
*    x1 x3       7 11 7 8 7 9 8 8 7
*    x1 x4       8 8 9 6 8 10 10 7 6
*    x1 x5       7 9 9 8 8 8 8 8 7
*    x1 x6       9 8 8 8 8 8 8 8 7
*    x1 x7       10 7 8 7 9 8 8 8 7
*    x1 x8       10 7 8 8 8 8 8 7 8
*    x1 x9       9 8 8 6 9 9 9 8 6
*    x1 x10      9 8 8 9 7 8 8 9 6
*    x1 x11      9 8 8 9 8 7 8 7 8
*    x1 x12      7 9 9 8 8 8 8 6 9
*    x1 x13      9 8 8 7 8 9 9 7 7
*    x1 x14      6 7 8 4 6 5 5 8 6 5 5 7
*    x1 x15      8 9 8 6 9 9 10 7 6
*    x1 x16      15 10 13 11 14 9
*    x1 x17      14 11 13 11 9 14
.
.
.
*    x12 x13     7 7 9 8 7 8 10 9 7
*    x12 x14     7 5 4 7 6 4 6 7 5 8 8 5
*    x12 x15     7 9 7 7 8 8 10 8 8
*    x12 x16     13 10 14 9 15 11
*    x12 x17     11 12 11 12 14 12
```

```
.
.
.
     N-Way      1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

We can use the `%MktEx` macro to check the design and print the information matrix and variance matrix.

```
%mktex(3 ** 13 4 3 2 2, n=72, examine=i v, options=check, init=randomized,
       interact=x1*x11 x2*x12 x3*x13 x4*x14 x5*x15)
```

In the interest of space, the results from this step are not shown.

Blocking an Existing Design

An existing design is blocked using the `%MktBlock` macro. The macro takes the observations in an existing design and optimally sorts them into blocks. Here we are seeing how to block the linear version of the choice design, but the macro can also be used directly on the choice design.

```
%mktblock(data=randomized, nblocks=4, out=sasuser.blockdes, seed=114)
```

This step took 2 seconds. Here are some of the results including the one-way frequencies within blocks. They should be examined to ensure that each level is well represented in each block. The design is nearly balanced in most of the factors and blocks. Perfect balance is impossible for the three level factors.

Vacation Example with Asymmetry
Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

	Block	x1	x2	x3	x4	x5	x6	x7	x8
Block	1	0.06	0.10	0.10	0.15	0.10	0.10	0.10	0.10
x1	0.06	1	0.07	0.13	0.17	0.07	0.03	0.10	0.06
x2	0.10	0.07	1	0.15	0.17	0.10	0.07	0.13	0.14
x3	0.10	0.13	0.15	1	0.15	0.08	0.09	0.10	0.14
x4	0.15	0.17	0.17	0.15	1	0.09	0.10	0.09	0.03
x5	0.10	0.07	0.10	0.08	0.09	1	0.09	0.10	0.07
x6	0.10	0.03	0.07	0.09	0.10	0.09	1	0.11	0.07
x7	0.10	0.10	0.13	0.10	0.09	0.10	0.11	1	0.10
x8	0.10	0.06	0.14	0.14	0.03	0.07	0.07	0.10	1

.
.
.

```
                Vacation Example with Asymmetry
                     Summary of Frequencies
 There are 0 Canonical Correlations Greater Than 0.316
           * - Indicates Unequal Frequencies

                     Frequencies

      Block          18 18 18 18
*     x1             25 23 24
*     x2             24 26 22
*     x3             23 26 23
*     x4             24 23 25
*     x5             23 24 25
*     x6             24 25 23
*     x7             23 24 25
*     x8             22 24 26
*     x9             24 23 25
*     x10            26 24 22
*     x11            26 23 23
*     x12            26 23 23
*     x13            24 23 25
*     x14            17 18 18 19
*     x15            23 25 24
*     x16            42 30
      x17            36 36

*     Block x1       7 5 6 6 6 6 6 6 6 6 6 6
*     Block x2       7 5 6 6 7 5 6 7 5 5 7 6
*     Block x3       6 7 5 6 6 6 6 7 5 5 6 7
*     Block x4       7 6 5 7 5 6 6 6 6 4 6 8
*     Block x5       6 7 5 6 6 6 5 6 7 6 5 7
*     Block x6       7 6 5 5 6 7 6 7 5 6 6 6
*     Block x7       6 6 6 5 6 7 6 7 5 6 5 7
*     Block x8       6 5 7 6 7 5 5 6 7 5 6 7
*     Block x9       6 7 5 6 5 7 6 5 7 6 6 6
*     Block x10      7 6 5 6 6 6 6 6 6 7 6 5
*     Block x11      5 7 6 7 6 5 8 5 5 6 5 7
*     Block x12      6 6 6 6 6 6 7 6 5 7 5 6
*     Block x13      7 6 5 6 6 6 6 5 7 5 6 7
*     Block x14      4 5 5 4 4 4 5 5 4 6 4 4 5 3 4 6
*     Block x15      5 7 6 6 6 6 6 6 6 6 6 6
*     Block x16      10 8 11 7 10 8 11 7
*     Block x17      8 10 10 8 9 9 9 9
.
.
.
```

```
                    N-Way        1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                                 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                                 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                                 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Generating the Questionnaire

These next steps print the questionnaire.

```
%let m   = 6;                    /* m alts including constant  */
%let mm1 = %eval(&m - 1);        /* m - 1                      */
%let n   = 18;                   /* number of choice sets      */
%let blocks = 4;                 /* number of blocks           */

title;
options ls=80 ps=60 nonumber nodate;

data _null_;
   array dests[&mm1] $ 10 _temporary_ ('Hawaii' 'Alaska' 'Mexico'
                                       'California' 'Maine');
   array scenes[3]   $ 13 _temporary_
                     ('the Mountains' 'a Lake' 'the Beach');
   array lodging[3]  $ 15 _temporary_
                     ('Cabin' 'Bed & Breakfast' 'Hotel');
   array x[15];
   array p[&mm1];
   length price $ 6;
   file print linesleft=ll;
   set sasuser.blockdes;
   by block;

   p1 = 1499 + (x[11] - 2) * 250;
   p2 = 1499 + (x[12] - 2) * 250;
   p3 = 1249 + (x[13] - 2) * 250;
   p4 = 1374 + (x[14] - 2.5) * 250;
   p5 = 1249 + (x[15] - 2) * 250;
   if first.block then do;
      choice = 0;
      put _page_;
      put @50 'Form: ' block  ' Subject: ________' //;
      end;
   choice + 1;
   if ll < (19 + (x16 = 1) + (x17 = 1)) then put _page_;
   put choice 2. ') Circle your choice of '
       'vacation destinations:' /;
```

```
do dest = 1 to &mm1;
   price = left(put(p[dest], dollar6.));
   put '    ' dest 1. ') ' dests[dest]
       +(-1) ', staying in a ' lodging[x[dest]]
       'near ' scenes[x[&mm1 + dest]] +(-1) ',' /
       +7 'with a package cost of ' price +(-1) @@;
   if dest = 3 and x16 = 1 then
      put ', and an optional visit' / +7
          'to archaeological sites for an additional $100' @@;
   else if dest = 1 and x17 = 1 then
      put ', and an optional helicopter' / +7
          'flight to an active volcano for an additional $200' @@;
   put '.' /;
   end;
put "    &m) Stay at home this year." /;
run;
```

Here are the first two choice sets for the first subject.

```
                                          Form: 1  Subject: ________


1) Circle your choice of vacation destinations:

   1) Hawaii, staying in a Cabin near the Mountains,
      with a package cost of $1,249, and an optional helicopter
      flight to an active volcano for an additional $200.

   2) Alaska, staying in a Hotel near a Lake,
      with a package cost of $1,749.

   3) Mexico, staying in a Cabin near a Lake,
      with a package cost of $1,499.

   4) California, staying in a Bed & Breakfast near the Mountains,
      with a package cost of $1,499.

   5) Maine, staying in a Bed & Breakfast near a Lake,
      with a package cost of $999.

   6) Stay at home this year.
```

```
2) Circle your choice of vacation destinations:

   1) Hawaii, staying in a Hotel near a Lake,
      with a package cost of $1,499, and an optional helicopter
      flight to an active volcano for an additional $200.

   2) Alaska, staying in a Cabin near a Lake,
      with a package cost of $1,499.

   3) Mexico, staying in a Bed & Breakfast near a Lake,
      with a package cost of $1,249.

   4) California, staying in a Cabin near a Lake,
      with a package cost of $1,499.

   5) Maine, staying in a Hotel near a Lake,
      with a package cost of $1,499.

   6) Stay at home this year.
```

In practice, data collection may be much more elaborate than this. It may involve art work or photographs, and the choice sets may be presented and the data may be collected through personal interview or over the web. However the choice sets are presented and the data collected, the essential ingredients remain the same. Subjects are shown sets of alternatives and asked to make a choice, and then they go on to the next set.

Generating Artificial Data

This next step generates an artificial set of data. Collecting data is time consuming and expensive. Generating some artificial data before the data are collected to test your code and make sure the analysis will run is a good idea. It helps avoid the "How am I going to analyze this?" question from occurring after the data have already been collected. See page 256 for an alternative method of testing your design. This step generates data for 300 subjects, 100 per block.

```
data _null_;
   array dests[&mm1] _temporary_ (5 -1 4 3 2);
   array scenes[3]   _temporary_ (-1 0 1);
   array lodging[3]  _temporary_ (0 3 2);
   array u[&m];
   array x[15];
```

```
do rep = 1 to 100;
   n = 0;
   do i = 1 to &blocks;
      k + 1;
      if mod(k,3) = 1 then put;
      put k 3. +1 i 1. +2 @@;
      do j = 1 to &n; n + 1;
         set sasuser.blockdes point=n;
         do dest = 1 to &mm1;
            u[dest] = dests[dest] + lodging[x[dest]] +
                      scenes[x[&mm1 + dest]] -
                      x[2 * &mm1 + dest] +
                      2 * normal(17);
         end;
         u[1] = u[1] + (x16 = 1);
         u[3] = u[3] + (x17 = 1);
         u&m  = -3 + 3 * normal(17);
         m = max(of u1-u&m);
         if      abs(u1 - m) < 1e-4 then c = 1;
         else if abs(u2 - m) < 1e-4 then c = 2;
         else if abs(u3 - m) < 1e-4 then c = 3;
         else if abs(u4 - m) < 1e-4 then c = 4;
         else if abs(u5 - m) < 1e-4 then c = 5;
         else                            c = 6;
         put +(-1) c @@;
         end;
      end;
   end;
stop;
run;
```

The `dests`, `scenes`, and `lodging` arrays are initialized with part-worth utilities for each level. The utilities for each of the destinations are computed and stored in the array `u` in the statement `u[dest] = ...`, which includes an error term `2 * normal(17)`. The utilities for the side trips are added in separately with `u[1] = u[1] + (x16 = 1)` and `u[3] = u[3] + (x17 = 1)`. The utility for the stay-at-home alternative is `-3 + 3 * normal(17)`. The maximum utility is computed, `m = max(of u1-u&m)` and the alternative with the maximum utility is chosen. The `put` statement writes out the results to the log.

Reading, Processing, and Analyzing the Data

The results from the previous step are pasted into a DATA step and run to mimic reading real input data.

```
title 'Vacation Example with Asymmetry';

data results;
   input Subj Form (choose1-choose&n) (1.) @@;
   datalines;
  1 1 413131153351111535   2 2 111151141153511152   3 3 331151344433111341
  4 4 313311134131311114   5 1 533133551541441321   6 2 311151111113311511
  7 3 313113111311351331   8 4 413341341134141131   9 1 533335315133141661
  .
  .
  .
;
```

The analysis proceeds in a fashion similar to before in the simpler vacation example on page 154. We start by creating some formats for the factor levels and the key to converting the linear design into a choice design.‡

```
proc format;
   value price 1 = ' 999'      2 = '1249' 3 = '1499' 4 = '1749';
   value scene 1 = 'Mountains' 2 = 'Lake'            3 = 'Beach';
   value lodge 1 = 'Cabin'     2 = 'Bed & Breakfast' 3 = 'Hotel';
   value side  1 = 'Side Trip' 2 = 'No';
   run;

data key;
   input Place $ 1-10 (Lodge Scene Price Side) ($);
   datalines;
Hawaii      x1  x6   x11  x16
Alaska      x2  x7   x12  .
Mexico      x3  x8   x13  x17
California  x4  x9   x14  .
Maine       x5  x10  x15  .
.           .   .    .    .
;
```

For analysis, the design will have five attributes. `Place` is the alternative name. `Lodge`, `Scene`, `Price` and `Side` are created from the design using the indicated factors. See page 155 for more information on creating the design key. Notice that `Side` only applies to some of the alternatives and hence has missing values for the others. Processing the design and merging it with the data are similar to what was done on pages 155 and 157. One difference is now there are asymmetries in `Price`. For Hawaii's price, `x11`, we need to change 1, 2, and 3 to \$1249, \$1499, and \$1749. For Alaska's price, `x12`, we need to change 1, 2, and 3 to \$1249, \$1499, and \$1749. For Mexico's price, `x13`, we need to change 1, 2, and 3 to \$999, \$1249, and \$1499. For California's price, `x14`, we need to change 1, 2, 3, and 4 to \$999, \$1249, \$1499, and \$1749. For Maine's price, `x11`, we need to change 1, 2, and 3 to \$999, \$1249, and \$1499. We can simplify the problem by adding 1 to `x11` and `x12`, which are the factors that start at

‡See page 87 for an illustration of linear versus choice designs.

$1249 instead of $999. This will allow us to use a common format to set the price. See pages 251 and 502 for examples of handling more complicated asymmetries.

```
data temp;
   set sasuser.blockdes;
   x11 + 1;
   x12 + 1;
   run;

%mktroll(design=temp, key=key, alt=place, out=rolled, options=nowarn)

%mktmerge(design=rolled, data=results, out=res2, blocks=form,
          nsets=&n, nalts=&m, setvars=choose1-choose&n,
          stmts=%str(price = input(put(price, price.), 5.);
                     format scene scene. lodge lodge. side side.;))

proc print data=res2(obs=18); run;
```

Here are the first three choice sets.

Vacation Example with Asymmetry

Obs	Subj	Form	Set	Place	Lodge	Scene	Price	Side	c
1	1	1	1	Hawaii	Cabin	Mountains	1249	No	2
2	1	1	1	Alaska	Hotel	Lake	1749	.	2
3	1	1	1	Mexico	Cabin	Lake	1499	Side Trip	2
4	1	1	1	California	Bed & Breakfast	Mountains	1499	.	1
5	1	1	1	Maine	Bed & Breakfast	Lake	999	.	2
6	1	1	1		.	.	.	.	2
7	1	1	2	Hawaii	Hotel	Lake	1499	No	1
8	1	1	2	Alaska	Cabin	Lake	1499	.	2
9	1	1	2	Mexico	Bed & Breakfast	Lake	1249	Side Trip	2
10	1	1	2	California	Cabin	Lake	1499	.	2
11	1	1	2	Maine	Hotel	Lake	1499	.	2
12	1	1	2		.	.	.	.	2
13	1	1	3	Hawaii	Cabin	Beach	1499	Side Trip	2
14	1	1	3	Alaska	Cabin	Mountains	1499	.	2
15	1	1	3	Mexico	Hotel	Beach	999	Side Trip	1
16	1	1	3	California	Hotel	Lake	1249	.	2
17	1	1	3	Maine	Bed & Breakfast	Beach	999	.	2
18	1	1	3		.	.	.	.	2

Indicator variables and labels are created using PROC TRANSREG like before.

```
proc transreg design=5000 data=res2 nozeroconstant norestoremissing;
   model class(place / zero=none order=data)
         class(price scene lodge / zero=none order=formatted)
         class(place * side / zero=' ' 'No' separators='' ' ') /
         lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   id subj set form c;
   run;
proc print data=coded(obs=6) label;
   run;
```

The `design=5000` option specifies that no model is fit; the procedure is just being used to code a design in blocks of 5000 observations at a time. The `nozeroconstant` option specifies that if the coding creates a constant variable, it should not be zeroed. The `norestoremissing` option specifies that missing values should not be restored when the `out=` data set is created. The `model` statement names the variables to code and provides information about how they should be coded. The specification `class(place / ...)` specifies that the variable `Place` is a classification variable and requests a binary coding. The `zero=none` option creates binary variables for all categories. The `order=data` option sorts the levels into the order they were first encountered in the data set. Similarly, the variables `Price`, `Scene`, and `Lodge` are classification variables. The specification `class(place * side / ...)` creates alternative-specific side trip effects. The option `zero=' ' 'No'` specifies that indicator variables should be created for all levels of `Place` except blank, and all levels of `Side` except `'No'`. The specification `zero=' '` is almost the same as `zero=none`. The `zero=' '` specification names a missing level as the reference level creating indicator variables for all nonmissing levels of the `class` variables, just like `zero=none`. The difference is `zero=none` applies to all of the variables named in the `class` specification. When you want `zero=none` to apply to only some variables, then you must use `zero=' '`, as in `zero=' ' 'No'` instead. In this case, `zero=none` applies to the first variable and `zero='No'` applies to the second. With `zero=' '`, TRANSREG prints the following warning, which can be safely ignored.

```
WARNING: Reference level ZERO='' was not found for variable Place.
```

The `separators='' ' '` option (`separators=` quote quote space quote space quote) allows you to specify two label component separators for the main effect and interaction terms, respectively. By specifying a blank for the second value, we request labels for the side trip effects like `'Mexico Side Trip'` instead of the default `'Mexico * Side Trip'`. This option is explained in more detail on page 209.

The `lprefix=0` option specifies that when labels are created for the binary variables, zero characters of the original variable name should be used as a prefix. This means that the labels are created only from the level values. An `output` statement names the output data set and drops variables that are not needed. Finally, the `id` statement names the additional variables that we want copied from the input to the output data set.

Vacation Example with Asymmetry

Obs	Hawaii	Alaska	Mexico	California	Maine	999	1249	1499	1749	Beach	Lake
1	1	0	0	0	0	0	1	0	0	0	0
2	0	1	0	0	0	0	0	0	1	0	1
3	0	0	1	0	0	0	0	1	0	0	1
4	0	0	0	1	0	0	0	1	0	0	0
5	0	0	0	0	1	1	0	0	0	0	1
6	0	0	0	0	0	0	0	0	0	0	0

Obs	Mountains	Bed & Breakfast	Cabin	Hotel	Alaska Side Trip	California Side Trip	Hawaii Side Trip	Maine Side Trip	Mexico Side Trip
1	1	0	1	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0	0	1
4	1	1	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0

Obs	Place	Price	Scene	Lodge	Side	Subj	Set	Form	c
1	Hawaii	1249	Mountains	Cabin	No	1	1	1	2
2	Alaska	1749	Lake	Hotel	.	1	1	1	2
3	Mexico	1499	Lake	Cabin	Side Trip	1	1	1	2
4	California	1499	Mountains	Bed & Breakfast	.	1	1	1	1
5	Maine	999	Lake	Bed & Breakfast	.	1	1	1	2
6		.	.	.	.	1	1	1	2

The PROC PHREG specification is the same as we have used before. (Recall that we used `%phchoice(on)` on page 95 to customize the output from PROC PHREG.)

```
proc phreg data=coded brief;
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;
```

Here are the results.

Vacation Example with Asymmetry

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	43200
Number of Observations Used	43200

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	7200	6	1	5

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	25801.336	12547.295
AIC	25801.336	12575.295
SBC	25801.336	12671.641

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	13254.0409	14	<.0001
Score	12457.0987	14	<.0001
Wald	5078.7045	14	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Hawaii	1	3.63561	0.23933	230.7503	<.0001
Alaska	1	-0.96812	0.28146	11.8309	0.0006
Mexico	1	2.30666	0.24388	89.4600	<.0001
California	1	1.50750	0.24418	38.1165	<.0001
Maine	1	0.80856	0.24780	10.6467	0.0011
999	1	2.12422	0.07017	916.4316	<.0001
1249	1	1.44097	0.06097	558.5902	<.0001
1499	1	0.79010	0.06105	167.4903	<.0001
1749	0	0	.	.	.
Beach	1	1.43665	0.04754	913.2623	<.0001
Lake	1	0.72145	0.04540	252.5637	<.0001
Mountains	0	0	.	.	.
Bed & Breakfast	1	0.72398	0.04091	313.1374	<.0001
Cabin	1	-1.39630	0.04876	820.0340	<.0001
Hotel	0	0	.	.	.
Alaska Side Trip	0	0	.	.	.
California Side Trip	0	0	.	.	.
Hawaii Side Trip	1	0.69058	0.05802	141.6519	<.0001
Maine Side Trip	0	0	.	.	.
Mexico Side Trip	1	0.60142	0.06145	95.8045	<.0001

You would not expect the part-worth utilities to match those that were used to generate the data, but you would expect a similar ordering within each factor, and in fact that does occur. These data can also be analyzed with quantitative price effects and destination by attribute interactions, as in the previous vacation example.

Aggregating the Data

This data set is rather large with 43,200 observations. You can make the analysis run faster and with less memory by aggregating. Instead of stratifying on each choice set and subject combination, you can stratify just on choice set and specify the number of times each alternative was chosen or unchosen. First, use PROC SUMMARY to count the number of times each observation occurs. Specify all the analysis variables, and in this example, also specify `Form`. The variable `Form` was added to the list because `Set` designates choice set within form. It is the `Form` and `Set` combinations that identify the choice sets. (In the previous PROC PHREG step, since the `Subj * Set` combinations uniquely identified each stratum, `Form` was not needed.) PROC SUMMARY stores the number of times each unique observation appears in the variable `_freq_`. PROC PHREG is then run with a `freq` statement. Now, instead of analyzing a data set with 43,200 observations and 7200 strata, we analyze a data set with at most $2 \times 6 \times 72 = 864$ observations and 72 strata. For each of the 6 alternatives and 72 choice sets, there are typically 2 observations in the aggregate data set: one that contains the number of times it was chosen and one that contains the number of times it was not chosen. When one of those counts is zero, there will be one observation. In this case, the aggregate data set has 726 observations.

```
proc summary data=coded nway;
   class form set c &_trgind;
   output out=agg(drop=_type_);
   run;

proc phreg data=agg;
   model c*c(2) = &_trgind / ties=breslow;
   freq _freq_;
   strata form set;
   run;
```

PROC SUMMARY ran in three seconds, and PROC PHREG ran in less than one second. The parameter estimates and Chi-Square statistics (not shown) are the same as before. The summary table shows the results of the aggregation, 100 out of 600 alternatives were chosen in each stratum. The log likelihood statistics are different, but that does not matter since the Chi-Square statistics are the same. Page 226 provides more information about this.

Vacation Example with Asymmetry

The PHREG Procedure

Model Information

Data Set	WORK.AGG
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Frequency Variable	_FREQ_
Ties Handling	BRESLOW

Number of Observations Read	729
Number of Observations Used	729
Sum of Frequencies Read	43200
Sum of Frequencies Used	43200

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Stratum	Form	Set	Number of Alternatives	Chosen Alternatives	Not Chosen
1	1	1	600	100	500
2	1	2	600	100	500
3	1	3	600	100	500
4	1	4	600	100	500
.					
.					
.					

71	4	71	600	100	500
72	4	72	600	100	500
Total			43200	7200	36000

Brand Choice Example with Aggregate Data

In this next example, subjects were presented with brands of a product at different prices. There were four brands and a constant alternative, eight choice sets, and 100 subjects. This example shows how to handle data that come to you already aggregated. It also illustrates comparing the fits of two competing models, the mother logit model, cross effects, IIA, and techniques for handling large data sets. The choice sets, with the price of each alternative and the number of times it was chosen, are shown next.

Set	Brand 1		Brand 2		Brand 3		Brand 4		Other	
1	$3.99	4	$5.99	29	$3.99	16	$5.99	42	$4.99	9
2	$5.99	12	$5.99	19	$5.99	22	$5.99	33	$4.99	14
3	$5.99	34	$5.99	26	$3.99	8	$3.99	27	$4.99	5
4	$5.99	13	$3.99	37	$5.99	15	$3.99	27	$4.99	8
5	$5.99	49	$3.99	1	$3.99	9	$5.99	37	$4.99	4
6	$3.99	31	$5.99	12	$5.99	6	$3.99	18	$4.99	33
7	$3.99	37	$3.99	10	$5.99	5	$5.99	35	$4.99	13
8	$3.99	16	$3.99	14	$3.99	5	$3.99	51	$4.99	14

The first choice set consists of Brand 1 at $3.99, Brand 2 at $5.99, Brand 3 at $3.99, Brand 4 at $5.99, and Other at $4.99. From this choice set, Brand 1 was chosen 4 times, Brand 2 was chosen 29 times, Brand 3 was chosen 16 times, Brand 4 was chosen 42 times, and Other was chosen 9 times.

Processing the Data

As in the previous examples, we will process the data to create a data set with one stratum for each choice set within each subject and m alternatives per stratum. This example will have 100 people times 5 alternatives times 8 choice sets equals 4000 observations. The first five observations are for the first subject and the first choice set, the next five observations are for the second subject and the first choice set, ..., the next five observations are for the one-hundredth subject and the first choice set, the next five observations are for the first subject and the second choice set, and so on. Subject 1 in the first choice set is almost certainly not the same as subject 1 in subsequent choice sets since we were given aggregate data. However, that is not important. What is important is that we have a subject and choice set variable whose unique combinations identify each choice set within each subject. In previous examples, we specified `strata Subj Set` with PROC PHREG, and our data were sorted by choice set within subject. We can still use the same specification even though our data are now sorted by subject within choice set. This next step reads and prepares the data.

```
%let m = 5;  /* Number of Brands in Each Choice Set */
             /* (including Other)                   */

title 'Brand Choice Example, Multinomial Logit Model';

proc format;
   value brand 1 = 'Brand 1' 2 = 'Brand 2' 3 = 'Brand 3'
               4 = 'Brand 4' 5 = 'Other';
   run;
```

```
data price;
   array p[&m] p1-p&m; /* Prices for the Brands */
   array f[&m] f1-f&m; /* Frequency of Choice   */

   input p1-p&m f1-f&m;
   keep subj set brand price c p1-p&m;
   * Store choice set and subject number to stratify;
   Set = _n_; Subj = 0;
   do i = 1 to &m;           /* Loop over the &m frequencies    */
      do ci = 1 to f[i];     /* Loop frequency of choice times  */
         subj + 1;           /* Subject within choice set       */
         do Brand = 1 to &m; /* Alternatives within choice set  */

            Price = p[brand];

            * Output first choice: c=1, unchosen: c=2;
            c = 2 - (i eq brand); output;
            end;
         end;
      end;

format brand brand.;
datalines;
3.99 5.99 3.99 5.99 4.99   4 29 16 42  9
5.99 5.99 5.99 5.99 4.99  12 19 22 33 14
5.99 5.99 3.99 3.99 4.99  34 26  8 27  5
5.99 3.99 5.99 3.99 4.99  13 37 15 27  8
5.99 3.99 3.99 5.99 4.99  49  1  9 37  4
3.99 5.99 5.99 3.99 4.99  31 12  6 18 33
3.99 3.99 5.99 5.99 4.99  37 10  5 35 13
3.99 3.99 3.99 3.99 4.99  16 14  5 51 14
;

proc print data=price(obs=15);
   var subj set c price brand;
   run;
```

The inner loop `do Brand = 1 to &m` creates all of the observations for the m alternatives within a person/choice set combination. Within a choice set (row of input data), the outer two loops, `do i = 1 to &m` and `do ci = 1 to f[i]` execute the code inside 100 times, the variable `Subj` goes from 1 to 100. In the first choice set, they first create the data for the four subjects that chose Brand 1, then the data for the 29 subjects that chose Brand 2, and so on. Here are the first 15 observations of the data set.

Brand Choice Example, Multinomial Logit Model

Obs	Subj	Set	c	Price	Brand
1	1	1	1	3.99	Brand 1
2	1	1	2	5.99	Brand 2
3	1	1	2	3.99	Brand 3
4	1	1	2	5.99	Brand 4
5	1	1	2	4.99	Other
6	2	1	1	3.99	Brand 1
7	2	1	2	5.99	Brand 2
8	2	1	2	3.99	Brand 3
9	2	1	2	5.99	Brand 4
10	2	1	2	4.99	Other
11	3	1	1	3.99	Brand 1
12	3	1	2	5.99	Brand 2
13	3	1	2	3.99	Brand 3
14	3	1	2	5.99	Brand 4
15	3	1	2	4.99	Other

Note that the data set also contains the variables `p1-p5` which contain the prices of each of the alternatives. These variables, which are used in constructing the cross effects, will be discussed in more detail on page 212.

```
proc print data=price(obs=5); run;
```

Brand Choice Example, Multinomial Logit Model

Obs	p1	p2	p3	p4	p5	Set	Subj	Brand	Price	c
1	3.99	5.99	3.99	5.99	4.99	1	1	Brand 1	3.99	1
2	3.99	5.99	3.99	5.99	4.99	1	1	Brand 2	5.99	2
3	3.99	5.99	3.99	5.99	4.99	1	1	Brand 3	3.99	2
4	3.99	5.99	3.99	5.99	4.99	1	1	Brand 4	5.99	2
5	3.99	5.99	3.99	5.99	4.99	1	1	Other	4.99	2

Simple Price Effects

The data are coded using PROC TRANSREG.

```
proc transreg design data=price nozeroconstant norestoremissing;
   model class(brand / zero=none) identity(price) / lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   label price = 'Price';
   id subj set c;
   run;
```

The `design` option specifies that no model is fit; the procedure is just being used to code a design. The `nozeroconstant` option specifies that if the coding creates a constant variable, it should not be zeroed. The `norestoremissing` option specifies that missing values should not be restored when the `out=` data set is created. The `model` statement names the variables to code and provides information about how they should be coded. The specification `class(brand / zero=none)` specifies that the variable `Brand` is a classification variable and requests a binary coding. The `zero=none` option creates binary variables for all categories. The specification `identity(price)` specifies that the variable `Price` is quantitative and hence should directly enter the model without coding. The `lprefix=0` option specifies that when labels are created for the binary variables, zero characters of the original variable name should be used as a prefix. This means that the labels are created only from the level values. An `output` statement names the output data set and drops variables that are not needed. Finally, the `id` statement names the additional variables that we want copied from the input to the output data set.

```
proc phreg data=coded brief;
   title2 'Discrete Choice with Common Price Effect';
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;

title2;
```

Here are the results. (Recall that we used `%phchoice(on)` on page 95 to customize the output from PROC PHREG.)

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Common Price Effect

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	4000
Number of Observations Used	4000

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	800	5	1	4

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2575.101	2425.214
AIC	2575.101	2435.214
SBC	2575.101	2458.637

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	149.8868	5	<.0001
Score	153.2328	5	<.0001
Wald	142.9002	5	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Brand 1	1	0.66727	0.12305	29.4065	<.0001
Brand 2	1	0.38503	0.12962	8.8235	0.0030
Brand 3	1	-0.15955	0.14725	1.1740	0.2786
Brand 4	1	0.98964	0.11720	71.2993	<.0001
Other	0	0	.	.	.
Price	1	0.14966	0.04406	11.5379	0.0007

Alternative-Specific Price Effects

In the next step, the data are coded for fitting a multinomial logit model with brand by price effects.

```
proc transreg design data=price nozeroconstant norestoremissing;
   model class(brand / zero=none separators='' ' ') |
         identity(price) / lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   label price = 'Price';
   id subj set c;
   run;
```

The PROC TRANSREG `model` statement has a vertical bar, "`|`", between the `class` specification and the `identity` specification. Since the `zero=none` option is specified with `class`, the vertical bar creates two sets of variables: five indicator variables for the brand effects and five more variables for the brand by price interactions. The `separators=` option allows you to specify two label component separators as quoted strings. The specification `separators='' ' '` (`separators=` quote quote space quote space quote) specifies a null string (quote quote) and a blank (quote space quote). The `separators='' ' '` option in the `class` specification specifies the separators that are used to construct the labels for the main effect and interaction terms, respectively. By default, the alternative-specific price effects –

the brand by price interactions − would have labels like `'Brand 1 * Price'` since the default second value for `separators=` is `' * '` (a quoted space asterisk space). Specifying `' '` (a quoted space) as the second value creates labels of the form `'Brand 1 Price'`. Since `lprefix=0`, the main-effects separator, which is the first `separators=` value, `''` (quote quote), is ignored. Zero name or input variable label characters are used to construct the label. The label is simply the formatted value of the `class` variable. The next steps print the first two coded choice sets and perform the analysis.

```
proc print data=coded(obs=10) label;
   title2 'Discrete Choice with Brand by Price Effects';
   var subj set c brand price &_trgind;
   run;
proc phreg data=coded brief;
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;

title2;
```

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Brand by Price Effects

Obs	Subj	Set	c	Brand	Price	Brand 1	Brand 2	Brand 3	Brand 4
1	1	1	1	Brand 1	3.99	1	0	0	0
2	1	1	2	Brand 2	5.99	0	1	0	0
3	1	1	2	Brand 3	3.99	0	0	1	0
4	1	1	2	Brand 4	5.99	0	0	0	1
5	1	1	2	Other	4.99	0	0	0	0
6	2	1	1	Brand 1	3.99	1	0	0	0
7	2	1	2	Brand 2	5.99	0	1	0	0
8	2	1	2	Brand 3	3.99	0	0	1	0
9	2	1	2	Brand 4	5.99	0	0	0	1
10	2	1	2	Other	4.99	0	0	0	0

Obs	Other	Brand 1 Price	Brand 2 Price	Brand 3 Price	Brand 4 Price	Other Price
1	0	3.99	0.00	0.00	0.00	0.00
2	0	0.00	5.99	0.00	0.00	0.00
3	0	0.00	0.00	3.99	0.00	0.00
4	0	0.00	0.00	0.00	5.99	0.00
5	1	0.00	0.00	0.00	0.00	4.99
6	0	3.99	0.00	0.00	0.00	0.00
7	0	0.00	5.99	0.00	0.00	0.00
8	0	0.00	0.00	3.99	0.00	0.00
9	0	0.00	0.00	0.00	5.99	0.00
10	1	0.00	0.00	0.00	0.00	4.99

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Brand by Price Effects

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	4000
Number of Observations Used	4000

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	800	5	1	4

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2575.101	2424.812
AIC	2575.101	2440.812
SBC	2575.101	2478.288

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	150.2891	8	<.0001
Score	154.2563	8	<.0001
Wald	143.1425	8	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Brand 1	1	-0.00972	0.43555	0.0005	0.9822
Brand 2	1	-0.62230	0.48866	1.6217	0.2028
Brand 3	1	-0.81250	0.60318	1.8145	0.1780
Brand 4	1	0.31778	0.39549	0.6456	0.4217
Other	0	0	.	.	.
Brand 1 Price	1	0.13587	0.08259	2.7063	0.1000
Brand 2 Price	1	0.20074	0.09210	4.7512	0.0293
Brand 3 Price	1	0.13126	0.11487	1.3057	0.2532
Brand 4 Price	1	0.13478	0.07504	3.2255	0.0725
Other Price	0	0	.	.	.

The likelihood for this model is essentially the same as for the simpler, common-price-slope model fit previously, $-2\log(\mathcal{L}_C) = 2425.214$ compared to 2424.812. You can test the null hypothesis that the two models are not significantly different by comparing their likelihoods. The difference between two $-2\log(\mathcal{L}_C)$'s (the number reported under 'With Covariates' in the output) has a chi-square distribution. We can get the *df* for the test by subtracting the two *df* for the two likelihoods. The difference $2425.214 - 2424.812 = 0.402$ is distributed χ^2 with $8 - 5 = 3$ *df* and is not statistically significant.

Mother Logit Model

This next step fits the so-called "mother logit" model. This step creates the full design matrix, including the brand, price, and cross effects. A cross effect represents the effect of one alternative on the utility of another alternative. First, let's look at the input data set for the first choice set.

```
proc print data=price(obs=5) label;
   run;
```

Brand Choice Example, Multinomial Logit Model

Obs	p1	p2	p3	p4	p5	Set	Subj	Brand	Price	c
1	3.99	5.99	3.99	5.99	4.99	1	1	Brand 1	3.99	1
2	3.99	5.99	3.99	5.99	4.99	1	1	Brand 2	5.99	2
3	3.99	5.99	3.99	5.99	4.99	1	1	Brand 3	3.99	2
4	3.99	5.99	3.99	5.99	4.99	1	1	Brand 4	5.99	2
5	3.99	5.99	3.99	5.99	4.99	1	1	Other	4.99	2

The input consists of `Set`, `Subj`, `Brand`, `Price`, and a choice time variable `c`. In addition, it contains five variables `p1` through `p5`. The first observation of the `Price` variable shows us that the first alternative costs \$3.99; `p1` contains the cost of alternative 1, \$3.99, which is the same for all alternatives. It does not matter which alternative you are looking at, `p1` shows that alternative 1 costs \$3.99. Similarly, the

second observation of the `Price` variable shows us that the second alternative costs \$5.99; `p2` contains the cost of alternative 2, \$5.99, which is the same for all alternatives. There is one price variable, `p1` through `p5`, for each of the five alternatives.

In all of the previous examples, we have used models that were coded so that the utility of an alternative only depended on the attributes of that alternative. For example, the utility of Brand 1 would only depend on the Brand 1 name and its price. In contrast, `p1-p5` contain information about each of the *other* alternatives' attributes. We will construct cross effects using the interaction of `p1-p5` and the `Brand` variable. In a model with cross effects, the utility for an alternative depends on both that alternative's attributes *and* the other alternatives' attributes. The IIA (independence from irrelevant alternatives) property states that utility only depends on an alternative's own attributes. Cross effects add other alternative's attributes to the model, so they can be used to test for violations of IIA. (See pages 219, 228, 354, and 358 for other discussions of IIA.) Here is the PROC TRANSREG code for the cross-effects model.

```
proc transreg design data=price nozeroconstant norestoremissing;
   model class(brand / zero=none separators='' ' ') | identity(price)
         identity(p1-p&m) *
            class(brand / zero=none lprefix=0 separators='' ' on ') /
         lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   label price = 'Price'
         p1 = 'Brand 1' p2 = 'Brand 2' p3 = 'Brand 3'
         p4 = 'Brand 4' p5 = 'Other';
   id subj set c;
   run;
```

The `class(brand / ...) | identity(price)` specification in the `model` statement is the same as the previous analysis. The additional terms, `identity(p1-p&m) * class(brand / ...)` create the cross effects. The second value of the `separators=` option, `' on'` is used to create labels like `'Brand 1 on Brand 2'` instead of the default `'Brand 1 * Brand 2'`. It is important to note that you must specify the cross effect by specifying `identity` with the price factors, followed by the asterisk, followed by `class` and the brand effect, *in that order*. The order of the specification determines the order in which brand names are added to the labels. Do not specify the brand variable first; doing so will create incorrect labels.

With m alternatives, there are $m \times m$ cross effects, but as we will see, many of them are zero. The first coded choice set is printed with the following PROC PRINT steps. Multiple steps are used to facilitate explaining the coding.

```
title2 'Discrete Choice with Cross Effects, Mother Logit';
proc format; value zer 0 = '  0' 1 = '  1'; run;
proc print data=coded(obs=5) label; var subj set c brand price; run;
proc print data=coded(obs=5) label; var Brand:;
   format brand: zer5.2 brand brand.; run;
proc print data=coded(obs=5) label; var p1B:; format p: zer5.2; id brand; run;
proc print data=coded(obs=5) label; var p2B:; format p: zer5.2; id brand; run;
proc print data=coded(obs=5) label; var p3B:; format p: zer5.2; id brand; run;
proc print data=coded(obs=5) label; var p4B:; format p: zer5.2; id brand; run;
proc print data=coded(obs=5) label; var p5B:; format p: zer5.2; id brand; run;
```

The coded data set contains the strata variable `Subj` and `Set`, choice time variable `c`, and `Brand` and `Price`. `Brand` and `Price` were used to create the coded independent variables but they are not used

in the analysis with PROC PHREG.

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Cross Effects, Mother Logit

Obs	Subj	Set	c	Brand	Price
1	1	1	1	Brand 1	3.99
2	1	1	2	Brand 2	5.99
3	1	1	2	Brand 3	3.99
4	1	1	2	Brand 4	5.99
5	1	1	2	Other	4.99

The effects 'Brand 1' through 'Other' in the next output are the binary brand effect variables. They indicate the brand for each alternative. The effects 'Brand 1 Price' through 'Other Price' are alternative-specific price effects. They indicate the price for each alternative. All ten of these variables are independent variables in the analysis, and their names are part of the &_trgind macro variable list, as are all of the cross effects that are described next.

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Cross Effects, Mother Logit

Obs	Brand 1	Brand 2	Brand 3	Brand 4	Other	Brand 1 Price	Brand 2 Price	Brand 3 Price	Brand 4 Price	Other Price	Brand
1	1	0	0	0	0	3.99	0	0	0	0	Brand 1
2	0	1	0	0	0	0	5.99	0	0	0	Brand 2
3	0	0	1	0	0	0	0	3.99	0	0	Brand 3
4	0	0	0	1	0	0	0	0	5.99	0	Brand 4
5	0	0	0	0	1	0	0	0	0	4.99	Other

The effects 'Brand 1 on Brand 1' through 'Brand 1 on Other' in the next output are the first five cross effects.

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Cross Effects, Mother Logit

Brand	Brand 1 on Brand 1	Brand 1 on Brand 2	Brand 1 on Brand 3	Brand 1 on Brand 4	Brand 1 on Other
Brand 1	3.99	0	0	0	0
Brand 2	0	3.99	0	0	0
Brand 3	0	0	3.99	0	0
Brand 4	0	0	0	3.99	0
Other	0	0	0	0	3.99

They represent the effect of Brand 1 at its price on the utility of each alternative. The label `'Brand` n `on Brand` m`'` is read as 'the effect of Brand n at its price on the utility of Brand m.' For the first choice set, these first five cross effects consist entirely of zeros and \$3.99's, where \$3.99 is the price of Brand 1 in this choice set. The nonzero value is constant across all of the alternatives in each choice set since Brand 1 has only one price in each choice set. Notice the `'Brand 1 on Brand 1'` term, which is the effect of Brand 1 at its price on the utility of Brand 1. Also notice the `'Brand 1 Price'` effect, which is shown in the previous output. The description "the effect of Brand 1 at its price on the utility of Brand 1" is just a convoluted way of describing the Brand 1 price effect. The `'Brand 1 on Brand 1'` cross effect is the same as the Brand 1 price effect, hence when we do the analysis, we will see that the coefficient for the `'Brand 1 on Brand 1'` cross effect is zero.

The effects `'Brand 2 on Brand 1'` through `'Brand 2 on Other'` in the next output are the next five cross effects.

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Cross Effects, Mother Logit

Brand	Brand 2 on Brand 1	Brand 2 on Brand 2	Brand 2 on Brand 3	Brand 2 on Brand 4	Brand 2 on Other
Brand 1	5.99	0	0	0	0
Brand 2	0	5.99	0	0	0
Brand 3	0	0	5.99	0	0
Brand 4	0	0	0	5.99	0
Other	0	0	0	0	5.99

They represent the effect of Brand 2 at its price on the utility of each alternative. For the first choice set, these five cross effects consist entirely of zeros and \$5.99's, where \$5.99 is the price of Brand 2 in this choice set. The nonzero value is constant across all of the alternatives in each choice set since Brand 2 has only one price in each choice set. Notice the `'Brand 2 on Brand 2'` term, which is the effect of Brand 2 at its price on the utility of Brand 2. The description "the effect of Brand 2 at its price on the utility of Brand 2" is just a convoluted way of describing the Brand 2 price effect. The `'Brand 2 on Brand 2'` cross effect is the same as the Brand 2 price effect, hence when we do the analysis, we will see that the coefficient for the `'Brand 2 on Brand 2'` cross effect is zero.

The effects `'Brand 3 on Brand 1'` through `'Brand 3 on Other'` in the next output are the next five cross effects.

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Cross Effects, Mother Logit

Brand	Brand 3 on Brand 1	Brand 3 on Brand 2	Brand 3 on Brand 3	Brand 3 on Brand 4	Brand 3 on Other
Brand 1	3.99	0	0	0	0
Brand 2	0	3.99	0	0	0
Brand 3	0	0	3.99	0	0
Brand 4	0	0	0	3.99	0
Other	0	0	0	0	3.99

They represent the effect of Brand 3 at its price on the utility of each alternative. For the first choice set, these five cross effects consist entirely of zeros and \$3.99's, where \$3.99 is the price of Brand 3 in this choice set. Notice that the `'Brand 3 on Brand 3'` term is the same as the Brand 3 price effect, hence when we do the analysis, we will see that the coefficient for the `'Brand 3 on Brand 3'` cross effect is zero.

Here are the remaining cross effects. They follow the same pattern that was described for the previous cross effects.

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Cross Effects, Mother Logit

Brand	Brand 4 on Brand 1	Brand 4 on Brand 2	Brand 4 on Brand 3	Brand 4 on Brand 4	Brand 4 on Other
Brand 1	5.99	0	0	0	0
Brand 2	0	5.99	0	0	0
Brand 3	0	0	5.99	0	0
Brand 4	0	0	0	5.99	0
Other	0	0	0	0	5.99

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Cross Effects, Mother Logit

Brand	Other on Brand 1	Other on Brand 2	Other on Brand 3	Other on Brand 4	Other on Other
Brand 1	4.99	0	0	0	0
Brand 2	0	4.99	0	0	0
Brand 3	0	0	4.99	0	0
Brand 4	0	0	0	4.99	0
Other	0	0	0	0	4.99

We have been describing variables by their labels. While it is not necessary to look at it, the `&_trgind` macro variable name list that PROC TRANSREG creates for this problem is as follows:

```
%put &_trgind;
```

```
BrandBrand_1 BrandBrand_2 BrandBrand_3 BrandBrand_4 BrandOther
BrandBrand_1Price BrandBrand_2Price BrandBrand_3Price BrandBrand_4Price
BrandOtherPrice p1BrandBrand_1 p1BrandBrand_2 p1BrandBrand_3 p1BrandBrand_4
p1BrandOther p2BrandBrand_1 p2BrandBrand_2 p2BrandBrand_3 p2BrandBrand_4
p2BrandOther p3BrandBrand_1 p3BrandBrand_2 p3BrandBrand_3 p3BrandBrand_4
p3BrandOther p4BrandBrand_1 p4BrandBrand_2 p4BrandBrand_3 p4BrandBrand_4
p4BrandOther p5BrandBrand_1 p5BrandBrand_2 p5BrandBrand_3 p5BrandBrand_4
p5BrandOther
```

The analysis proceeds in exactly the same manner as before.

```
proc phreg data=coded brief;
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;
```

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Cross Effects, Mother Logit

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	4000
Number of Observations Used	4000

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	800	5	1	4

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2575.101	2349.325
AIC	2575.101	2389.325
SBC	2575.101	2483.018

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	225.7752	20	<.0001
Score	218.4500	20	<.0001
Wald	190.0257	20	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Brand 1	1	1.24963	1.31259	0.9064	0.3411
Brand 2	1	-0.16269	1.38579	0.0138	0.9065
Brand 3	1	-3.90179	1.56511	6.2150	0.0127
Brand 4	1	2.49435	1.25537	3.9480	0.0469
Other	0	0	.	.	.
Brand 1 Price	1	0.51056	0.13178	15.0096	0.0001
Brand 2 Price	1	-0.04920	0.13411	0.1346	0.7137
Brand 3 Price	1	-0.27594	0.15517	3.1623	0.0754
Brand 4 Price	1	0.28951	0.12192	5.6389	0.0176
Other Price	0	0	.	.	.
Brand 1 on Brand 1	0	0	.	.	.
Brand 1 on Brand 2	1	0.51651	0.13675	14.2653	0.0002
Brand 1 on Brand 3	1	0.66122	0.15655	17.8397	<.0001
Brand 1 on Brand 4	1	0.32806	0.12664	6.7105	0.0096
Brand 1 on Other	0	0	.	.	.
Brand 2 on Brand 1	1	-0.39876	0.12832	9.6561	0.0019
Brand 2 on Brand 2	0	0	.	.	.
Brand 2 on Brand 3	1	-0.01755	0.15349	0.0131	0.9090
Brand 2 on Brand 4	1	-0.33802	0.12220	7.6512	0.0057
Brand 2 on Other	0	0	.	.	.
Brand 3 on Brand 1	1	-0.43868	0.13119	11.1823	0.0008
Brand 3 on Brand 2	1	-0.31541	0.13655	5.3356	0.0209
Brand 3 on Brand 3	0	0	.	.	.
Brand 3 on Brand 4	1	-0.54854	0.12528	19.1723	<.0001
Brand 3 on Other	0	0	.	.	.

Brand 4 on Brand 1	1	0.24398	0.12781	3.6443	0.0563
Brand 4 on Brand 2	1	-0.01214	0.13416	0.0082	0.9279
Brand 4 on Brand 3	1	0.40500	0.15285	7.0211	0.0081
Brand 4 on Brand 4	0	0	.	.	.
Brand 4 on Other	0	0	.	.	.
Other on Brand 1	0	0	.	.	.
Other on Brand 2	0	0	.	.	.
Other on Brand 3	0	0	.	.	.
Other on Brand 4	0	0	.	.	.
Other on Other	0	0	.	.	.

The results consist of:

- four nonzero brand effects and a zero for the constant alternative
- four nonzero alternative-specific price effects and a zero for the constant alternative
- $5 \times 5 = 25$ cross effects, the number of alternatives squared, but only $(5-1) \times (5-2) = 12$ of them are nonzero (four brands not counting Other affecting each of the remaining three brands).
  - There are three cross effects for the effect of Brand 1 on Brands 2, 3, and 4.
  - There are three cross effects for the effect of Brand 2 on Brands 1, 3, and 4.
  - There are three cross effects for the effect of Brand 3 on Brands 1, 2, and 4.
  - There are three cross effects for the effect of Brand 4 on Brands 1, 2, and 3.

All coefficients for the constant (other) alternative are zero as are the cross effects of a brand on itself.

The mother logit model is used to test for violations of IIA (independence from irrelevant alternatives). IIA means the odds of choosing alternative c_i over c_j do not depend on the other alternatives in the choice set. Ideally, this more general model will not significantly explain more variation in choice than the restricted models. Also, if IIA is satisfied, few if any of the cross-effect terms should be significantly different from zero. (See pages 213, 228, 354, and 358 for other discussions of IIA.) In this case, it appears that IIA is *not* satisfied (the data are artificial), so the more general mother logit model is needed. The chi-square statistic is $2424.812 - 2349.325 = 75.487$ with $20 - 8 = 12$ *df* ($p < 0.0001$).

You could eliminate some of the zero parameters by changing `zero=none` to `zero='Other'` and eliminating `p5` (`p&m`) from the model.

```
proc transreg design data=price nozeroconstant norestoremissing;
   model class(brand / zero='Other' separators='' ' ') | identity(price)
         identity(p1-p4) * class(brand / zero='Other' separators='' ' on ') /
         lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   label price = 'Price'
         p1 = 'Brand 1' p2 = 'Brand 2' p3 = 'Brand 3'
         p4 = 'Brand 4';
   id subj set c;
   run;
```

You could also eliminate the brand by price effects and instead capture brand by price effects as the cross effect of a variable on itself.

```
proc transreg design data=price nozeroconstant norestoremissing;
   model class(brand / zero='Other' separators='' ' ')
         identity(p1-p4) * class(brand / zero='Other' separators='' ' on ') /
         lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   label price = 'Price'
         p1 = 'Brand 1' p2 = 'Brand 2' p3 = 'Brand 3'
         p4 = 'Brand 4';
   id subj set c;
   run;
```

In both cases, the analysis (not shown) would be run in the usual manner. Except for the elimination of zero terms, and in the second case, the change to capture the price effects in the cross effects, the results are identical.

Aggregating the Data

In all examples so far (except the last part of the last vacation example), the data set has been created for analysis with one stratum for each choice set and subject combination. Such data sets can be large. The data can also be arrayed with a frequency variable and each choice set forming a separate stratum. This example illustrates how.

```
title 'Brand Choice Example, Multinomial Logit Model';
title2 'Aggregate Data';

%let m = 5;  /* Number of Brands in Each Choice Set */
             /* (including Other)                   */

proc format;
   value brand 1 = 'Brand 1' 2 = 'Brand 2' 3 = 'Brand 3'
               4 = 'Brand 4' 5 = 'Other';
   run;

data price2;
   array p[&m] p1-p&m; /* Prices for the Brands */
   array f[&m] f1-f&m; /* Frequency of Choice   */
   input p1-p&m f1-f&m;
   keep set price brand freq c p1-p&m;

   * Store choice set number to stratify;
   Set = _n_;
```

```
      do Brand = 1 to &m;

         Price = p[brand];

         * Output first choice: c=1, unchosen: c=2;
         Freq = f[brand]; c = 1; output;

         * Output number of times brand was not chosen.;
         freq = sum(of f1-f&m) - freq; c = 2; output;

         end;

   format brand brand.;
   datalines;
   3.99 5.99 3.99 5.99 4.99   4 29 16 42  9
   5.99 5.99 5.99 5.99 4.99  12 19 22 33 14
   5.99 5.99 3.99 3.99 4.99  34 26  8 27  5
   5.99 3.99 5.99 3.99 4.99  13 37 15 27  8
   5.99 3.99 3.99 5.99 4.99  49  1  9 37  4
   3.99 5.99 5.99 3.99 4.99  31 12  6 18 33
   3.99 3.99 5.99 5.99 4.99  37 10  5 35 13
   3.99 3.99 3.99 3.99 4.99  16 14  5 51 14
   ;
   proc print data=price2(obs=10);
      var set c freq price brand;
      run;
```

Brand Choice Example, Multinomial Logit Model
Aggregate Data

Obs	Set	c	Freq	Price	Brand
1	1	1	4	3.99	Brand 1
2	1	2	96	3.99	Brand 1
3	1	1	29	5.99	Brand 2
4	1	2	71	5.99	Brand 2
5	1	1	16	3.99	Brand 3
6	1	2	84	3.99	Brand 3
7	1	1	42	5.99	Brand 4
8	1	2	58	5.99	Brand 4
9	1	1	9	4.99	Other
10	1	2	91	4.99	Other

This data set has 5 brands times 2 observations times 8 choice sets for a total of 80 observations, compared to $100 \times 5 \times 8 = 4000$ using the standard method. Two observations are created for each alternative within each choice set. The first contains the number of people who chose the alternative, and the second contains the number of people who did not choose the alternative.

To analyze the data, specify `strata Set` and `freq Freq`.

```
proc transreg design data=price2 nozeroconstant norestoremissing;
   model class(brand / zero=none) identity(price) / lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   label price = 'Price';
   id freq set c;
   run;

proc phreg data=coded;
   title2 'Discrete Choice with Common Price Effect, Aggregate Data';
   model c*c(2) = &_trgind / ties=breslow;
   strata set;
   freq freq;
   run;

title2;
```

These steps produced the following results.

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Common Price Effect, Aggregate Data

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Frequency Variable	Freq
Ties Handling	BRESLOW

Number of Observations Read	80
Number of Observations Used	80
Sum of Frequencies Read	4000
Sum of Frequencies Used	4000

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Stratum	Set	Number of Alternatives	Chosen Alternatives	Not Chosen
1	1	500	100	400
2	2	500	100	400
3	3	500	100	400
4	4	500	100	400
5	5	500	100	400
6	6	500	100	400
7	7	500	100	400
8	8	500	100	400
Total		4000	800	3200

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	9943.373	9793.486
AIC	9943.373	9803.486
SBC	9943.373	9826.909

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	149.8868	5	<.0001
Score	153.2328	5	<.0001
Wald	142.9002	5	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Brand 1	1	0.66727	0.12305	29.4065	<.0001
Brand 2	1	0.38503	0.12962	8.8235	0.0030
Brand 3	1	-0.15955	0.14725	1.1740	0.2786
Brand 4	1	0.98964	0.11720	71.2993	<.0001
Other	0	0	.	.	.
Price	1	0.14966	0.04406	11.5379	0.0007

The summary table is small with eight rows, one row per choice set. Each row represents 100 chosen alternatives and 400 unchosen. The 'Analysis of Maximum Likelihood Estimates' table exactly matches the one produced by the standard analysis. The -2 LOG L statistics are different than before: 9793.486 now compared to 2425.214 previously. This is because the data are arrayed in this example so that the partial likelihood of the proportional hazards model fit by PROC PHREG with the `ties=breslow` option is now proportional to − not identical to − the likelihood for the choice model. However, the Model Chi-Square statistics, *df*, and *p*-values are the same as before. The two corresponding pairs of -2 LOG L's differ by a constant $9943.373-2575.101 = 9793.486-2425.214 = 7368.272 = 2\times800\times\log(100)$. Since the χ^2 is the -2 LOG L without covariates minus -2 LOG L with covariates, the constants cancel and the χ^2 test is correct for both methods.

The technique of aggregating the data and using a frequency variable can be used for other models as well, for example with brand by price effects.

```
proc transreg design data=price2 nozeroconstant norestoremissing;
   model class(brand / zero=none separators='' ' ') |
         identity(price) / lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   label price = 'Price';
   id freq set c;
   run;

proc phreg data=coded;
   title2 'Discrete Choice with Brand by Price Effects, Aggregate Data';
   model c*c(2) = &_trgind / ties=breslow;
   strata set;
   freq freq;
   run;
```

This step produced the following results. The only thing that changes from the analysis with one stratum for each subject and choice set combination is the likelihood.

Brand Choice Example, Multinomial Logit Model
Discrete Choice with Brand by Price Effects, Aggregate Data

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Frequency Variable	Freq
Ties Handling	BRESLOW

Number of Observations Read	80
Number of Observations Used	80
Sum of Frequencies Read	4000
Sum of Frequencies Used	4000

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Stratum	Set	Number of Alternatives	Chosen Alternatives	Not Chosen
1	1	500	100	400
2	2	500	100	400
3	3	500	100	400
4	4	500	100	400
5	5	500	100	400
6	6	500	100	400
7	7	500	100	400
8	8	500	100	400
Total		4000	800	3200

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	9943.373	9793.084
AIC	9943.373	9809.084
SBC	9943.373	9846.561

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	150.2891	8	<.0001
Score	154.2562	8	<.0001
Wald	143.1425	8	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Brand 1	1	-0.00972	0.43555	0.0005	0.9822
Brand 2	1	-0.62230	0.48866	1.6217	0.2028
Brand 3	1	-0.81250	0.60318	1.8145	0.1780
Brand 4	1	0.31778	0.39549	0.6456	0.4217
Other	0	0	.	.	.

Brand 1 Price	1	0.13587	0.08259	2.7063	0.1000
Brand 2 Price	1	0.20074	0.09210	4.7512	0.0293
Brand 3 Price	1	0.13126	0.11487	1.3057	0.2532
Brand 4 Price	1	0.13478	0.07504	3.2255	0.0725
Other Price	0	0	.	.	.

Previously, with one stratum per choice set within subject, we compared these models as follows: "The difference $2425.214 - 2424.812 = 0.402$ is distributed χ^2 with $8 - 5 = 3$ *df* and is not statistically significant." The difference between two $-2\log(\mathcal{L}_C)$'s equals the difference between two $-2\log(\mathcal{L}_B)$'s, since the constant terms $(800 \times \log(100))$ cancel, $9793.486 - 9793.084 = 2425.214 - 2424.812 = 0.402$.

Choice and Breslow Likelihood Comparison

This section explains why the -2 LOG L values differ by a constant with aggregate data versus individual data. It may be skipped by all but the most dedicated readers.

Consider the choice model with a common price slope. Let x_0 represent the price of the brand. Let x_1, x_2, x_3, and x_4 be indicator variables representing the choice of brands. Let $\mathbf{x} = (x_0\ x_1\ x_2\ x_3\ x_4)$ be the vector of alternative attributes. (A sixth element for 'Other' is omitted, since its parameter is always zero given the other brands.)

Consider the first choice set. There are five distinct vectors of alternative attributes
$\mathbf{x}_1 = (3.99\ 1\ 0\ 0\ 0)$ $\mathbf{x}_2 = (5.99\ 0\ 1\ 0\ 0)$ $\mathbf{x}_3 = (3.99\ 0\ 0\ 1\ 0)$ $\mathbf{x}_4 = (5.99\ 0\ 0\ 0\ 1)$
$\mathbf{x}_5 = (4.99\ 0\ 0\ 0\ 0)$

The vector $\mathbf{x}_2$, for example, represents choice of Brand 2, and $\mathbf{x}_5$ represents the choice of Other. One hundred individuals were asked to choose one of the $m = 5$ brands from each of the eight sets. Let f_1, f_2, f_3, f_4, and f_5 be the number of times each brand was chosen. For the first choice set, $f_1 = 4$, $f_2 = 29$, $f_3 = 16$, $f_4 = 42$, and $f_5 = 9$. Let N be the total frequency for each choice set, $N = \sum_{j=1}^{5} f_j = 100$. The likelihood L_1^C for the first choice set data is

$$L_1^C = \frac{\exp\left(\left(\sum_{j=1}^{5} f_j \mathbf{x}_j\right)\boldsymbol{\beta}\right)}{\left[\sum_{j=1}^{5} \exp(\mathbf{x}_j\boldsymbol{\beta})\right]^N}$$

The joint likelihood for all eight choice sets is the product of the likelihoods

$$\mathcal{L}_C = \prod_{k=1}^{8} L_k^C$$

The Breslow likelihood for this example, L_k^B, for the *kth* choice set, is the same as the likelihood for the choice model, except for a multiplicative constant.

$$L_k^C = N^N L_k^B = 100^{100} L_k^B$$

Therefore, the Breslow likelihood for all eight choice sets is

$$\mathcal{L}_B = \prod_{k=1}^{8} L_k^B = N^{-8N}\mathcal{L}_C = 100^{-800}\mathcal{L}_C$$

The two likelihoods are not exactly the same, because each choice set is designated as a separate stratum, instead of each choice set within each subject.

The log likelihood for the choice model is

$$\begin{aligned}
\log(\mathcal{L}_C) &= 800 \times \log(100) + \log(\mathcal{L}_B), \\
\log(\mathcal{L}_C) &= 800 \times \log(100) + (-0.5) \times 9793.486, \\
\log(\mathcal{L}_C) &= -1212.607
\end{aligned}$$

and $-2\log(\mathcal{L}_C) = 2425.214$, which matches the earlier output. However, it is usually not necessary to obtain this value.

Food Product Example with Asymmetry and Availability Cross Effects

This is the choice example from Kuhfeld, Tobias, and Garratt (1994), which starts on page 39. This example discusses the multinomial logit model, number of parameters, choosing the number of choice sets, designing the choice experiment, long design searches, examining the design, examining the sub-designs, examining the aliasing structure, blocking the design, testing the design before data collection, generating artificial data, processing the data, coding, cross effects, availability, multinomial logit model results, modeling subject attributes, results, and interpretation.

Consider the problem of using a discrete choice model to study the effect of introducing a retail food product. This may be useful, for instance, to refine a marketing plan or to optimize a product prior to test market. A typical brand team will have several concerns such as knowing the potential market share for the product, examining the source of volume, and providing guidance for pricing and promotions. The brand team may also want to know what brand attributes have competitive clout and want to identify competitive attributes to which they are vulnerable.

To develop this further, assume our client wishes to introduce a line extension in the category of frozen entrées. The client has one nationally branded competitor, a regional competitor in each of three regions, and a profusion of private label products at the grocery chain level. The product may come in two different forms: stove-top or microwaveable. The client believes that the private labels are very likely to mimic this line extension and to sell it at a lower price. The client suspects that this strategy on the part of private labels may work for the stove-top version but not for the microwaveable, where they have the edge on perceived quality. They also want to test the effect of a shelf-talker that will draw attention to their product.

The Multinomial Logit Model

This problem can be set up as a discrete choice model in which a respondent's choice among brands, given choice set C_a of available brands, will correspond to the brand with the highest utility. For each brand i, the utility U_i is the sum of a systematic component V_i and a random component e_i. The probability of choosing brand i from choice set C_a is therefore:

$$P(i|C_a) = P(U_i > \max(U_j)) = P(V_i + e_i > \max(V_j + e_j)) \quad \forall \quad (j \neq i) \in C_a$$

Assuming that the e_i follow an extreme value type I distribution, the conditional probabilities $P(i|C_a)$ can be found using the multinomial logit (MNL) formulation of McFadden (1974).

$$P(i|C_a) = \exp(V_i) / \sum_{j \in C_a} \exp(V_j)$$

One of the consequences of the MNL formulation is the property of independence from irrelevant alternatives (IIA). Under the assumption of IIA, all cross effects are assumed to be equal, so that if a brand gains in utility, it draws share from all other brands in proportion to their current shares. Departures from IIA exist when certain subsets of brands are in more direct competition and tend to draw a disproportionate amount of share from each other than from other members in the category.

IIA is frequently described using a transportation example. Say you have three alternatives for getting to work: bicycle, car, or a blue bus. If a fourth alternative became available, a red bus, then according to IIA the red bus should draw riders from the other alternatives in proportion to their current usage. However, in this case, IIA would be violated, and instead the red bus would draw more riders from the blue bus than from car drivers and bicycle riders.

The mother logit formulation of McFadden (1974) can be used to capture departures from IIA. In a mother logit model, the utility for brand i is a function of both the attributes of brand i and the attributes of other brands. The effect of one brand's attributes on another is termed a cross effect. In the case of designs in which only subsets C_a of the full shelf set C appear, the effect of the presence/absence of one brand on the utility of another is termed an *availability cross effect*. (See pages 213, 219, 354, and 358 for other discussions of IIA.)

Set Up

In the frozen entrée example, there are five alternatives: the client's brand, the client's line extension, a national branded competitor, a regional brand and a private label brand. Several regional and private labels can be tested in each market, then aggregated for the final model. Note that the line extension is treated as a separate alternative rather than as a level of the client brand. This enables us to model the source of volume for the new entry and to quantify any cannibalization that occurs. Each brand is shown at either two or three price points. Additional price points are included so that quadratic models of price elasticity can be tested. The indicator for the presence or absence of a brand in the shelf set is coded using one level of the `Price` variable. The layout of factors and levels is given in the following table.

Factors and Levels

Alternative	Factor	Levels	Brand	Description
1	X1	4	Client	1.29, 1.69, 2.09 + absent
2	X2	4	Client Line Extension	1.39, 1.89, 2.39, + absent
	X3	2		microwave/stove-top
	X4	2		shelf-talker yes/no
3	X5	3	Regional	1.99, 2.49 + absent
4	X6	3	Private Label	1.49, 2.29 absent
	X7	2		microwave/stove-top
5	X8	3	National	1.99 + 2.39 + absent

In addition to intercepts and main effects, we also require that all two-way interactions within alternatives be estimable: `x2*x3, x2*x4, x3*x4` for the line extension and `x6*x7` for private labels. This will enable us to test for different price elasticities by form (stove-top versus microwaveable) and to see if the promotion works better combined with a low price or with different forms. Using a linear model for `x1-x8`, the total number of parameters including the intercept, all main effects, and two-way interactions with brand is 25. This assumes that price is treated as qualitative. The actual number of parameters in the choice model is larger than this because of the inclusion of cross effects. Using indicator variables to code availability, the systematic component of utility for brand i can be expressed as:

$$V_i = a_i + \sum_k (b_{ik} \times x_{ik}) + \sum_{j \neq i} z_j (d_{ij} + \sum_l (g_{ijl} \times x_{jl}))$$

where

a_i = intercept for brand i
b_{ik} = effect of attribute k for brand i, where $k = 1, .., K_i$
x_{ik} = level of attribute k for brand i
d_{ij} = availability cross effect of brand j on brand i
z_j = availability code $= \begin{cases} 1 & \text{if } j \in C_a, \\ 0 & \text{otherwise} \end{cases}$
g_{ijl} = cross effect of attribute l for brand j on brand i, where $l = 1, .., L_j$
x_{jl} = level of attribute l for brand j.

The x_{ik} and x_{jl} could be expanded to include interaction and polynomial terms. In an availability-cross-effects design, each brand is present in only a fraction of the choice sets. The size of this fraction or subdesign is a function of the number of levels of the alternative-specific variable that is used to code availability (usually price). For instance, if price has three valid levels and a fourth zero level to indicate absence, then the brand will appear in only three out of four runs. Following Lazari and Anderson (1994), the size of each subdesign determines how many model equations can be written for each brand in the discrete choice model. If X_i is the subdesign matrix corresponding to V_i, then each X_i must be full rank to ensure that the choice set design provides estimates for all parameters.

To create the design, a full-factorial candidate set is generated consisting of 3456 runs. It is then reduced to 2776 runs that contain between two and four brands so that the respondent is never required to compare more than four brands at a time. In the model specification, we designate all variables as classification variables and require that all main effects and two-way interactions within brands be estimable. The number of runs calculations are based on the number of parameters that we wish to estimate in the various subdesigns $\mathbf{X}_i$ of $\mathbf{X}$. Assuming that there is a None alternative used as a reference level, the numbers of parameters required for various alternatives are shown in the next table along with the sizes of the subdesigns (rounded down) for various numbers of runs. Parameters for quadratic price models are given in parentheses. Note that the effect of private label being in a microwaveable or stove-top form (stove/micro cross effect) is an explicit parameter under the client line extension.

The subdesign sizes are computed by taking the floor of the number of runs from the marginal times the expected proportion of runs in which the alternative will appear. For example, for the client brand which has three prices and not available and 22 runs, floor($22 \times 3/4$) = 16; for the competitor and 32 runs, floor($32 \times 2/3$) = 21. The number of runs chosen was `n=26`. This number provides adequate degrees of freedom for the linear price model and will also allow estimation of direct quadratic price effects. To estimate quadratic cross effects for price would require 32 runs at the very least. Although the technique of using two-way interactions between nominal level variables will usually guarantee that all direct and cross effects are estimable, it is sometimes necessary and good practice to check the ranks of the subdesigns for more complex models (Lazari and Anderson 1994).

Parameters

Effect	Client	Client Line Extension	Regional	Private Label	Competitor
intercept	1	1	1	1	1
availability cross effects	4	4	4	4	4
direct price effect	1 (2)	1 (2)	1	1	1
price cross effects	4 (8)	4 (8)	4	4	4
stove versus microwave	-	1	-	1	-
stove/micro cross effects	-	1	-	-	-
shelf-talker	-	1	-	-	-
price*stove/microwave	-	1 (2)	-	1	-
price*shelf-talker	-	1 (2)	-	-	-
stove/micro*shelf-talker	-	1	-	-	-
Total	10 (15)	16 (23)	10	12	10
Subdesign size					
22 runs	16	16	14	14	14
26 runs	19	19	17	17	17
32 runs	24	24	21	21	21

Designing the Choice Experiment

This example originated with Kuhfeld, Tobias, and Garratt (1994), long before the `%MktRuns` macro was programmed. At least for now, we will skip the customary step of running the `%MktRuns` macro to suggest a design size and instead use the original size of 26 choice sets.

We will use the `%MktEx` autocall macro to create the design. (All of the autocall macros used in this book are documented starting on page 479.) To recap, we want to make the design $2^3 3^3 4^2$ in 26 runs, and we want the following interactions to be estimable: `x2*x3 x2*x4 x3*x4 x6*x7`. Furthermore, there are restrictions on the design. Each of the price variables, `x1`, `x2`, `x5`, `x6`, and `x8`, has one level − the maximum level − that indicates the alternative is not available in the choice set. We use this to create choice sets with 2, 3, or 4 alternatives available. If `(x1 < 4)` then the first alternative is available, if `(x2 < 4)` then the second alternative is available, if `(x5 < 3)` then the third alternative is available, and so on. A Boolean term such as `(x1 < 4)` is one when true and zero otherwise. Hence,

```
((x1 < 4) + (x2 < 4) + (x5 < 3) + (x6 < 3) + (x8 < 3))
```

is the number of available alternatives. This is simply the sum of some 1's if available and 0's if not available.

We impose restrictions with the `%MktEx` macro by writing a macro, with IML statements, that quantifies the badness of each run (or in this case, each choice set). We do this so `bad` = 0 is good and values larger than zero are increasingly worse. We write our restrictions using an IML row vector `x` that contains the levels (integers beginning with 1) of each of the factors in the *ith* choice set, the one the macro is currently seeking to improve. The *jth* factor is `x[j]`. or we may also use the factor names (for example, `x1`, `x2`). (See pages 337 and 566 for other examples of restrictions.)

We must use IML logical operators, which are not as rich as DATA step operators:

=	equals	not: EQ
∧ = or ¬ =	not equals	not: NE
<	less than	not: LT
<=	less than or equal to	not: LE
>	greater than	not: GT
>=	greater than or equal to	not: GE
&	and	not: AND
|	or	not: OR
∧ or ¬	not	not: NOT

To restrict the design, we must specify `restrictions=`*macro-name*, in this case `restrictions=bad`, that names the macro that quantifies badness. The first statement counts up the number of available alternatives. The second sets the actual badness values. If bad (the number available) is less than two or greater than 4, then the Boolean expression `((bad < 2) | (bad > 4))` is true or 1. When the expression is true, then bad gets set to the absolute difference between the number available and 3. Hence, zero available corresponds to `bad = 3`, one available corresponds to `bad = 2`, two through four available corresponds to `bad = 0`, and five available corresponds to `bad = 2`. Do not just set `bad` to zero when everything is fine and one otherwise, but the macro needs to know that when it switches from zero available to one available, it is going in the right direction. For simple restrictions like this, it does not matter very much. However, for complicated sets of restrictions, it is critical that the *bad* variable is set to a count of the number of current restriction violations. Here is the code.

```
title 'Consumer Food Product Example';

%macro bad;
  bad = (x1 < 4) + (x2 < 4) + (x5 < 3) + (x6 < 3) + (x8 < 3);
  bad = abs(bad - 3) * ((bad < 2) | (bad > 4));
  %mend;

%mktex( 4 4 2 2 3 3 2 3, n=26, interact=x2*x3 x2*x4 x3*x4 x6*x7,
        restrictions=bad, seed=377, outr=sasuser.choicdes )
```

Here are the initial messages the macro prints.

```
NOTE: Generating the fractional-factorial design, n=27.
NOTE: Generating the candidate set.
NOTE: Performing 60 searches of 2,776 candidates, full-factorial=3,456.
```

The tabled design initialization part of the coordinate-exchange algorithm iterations will be initialized with the first 26 rows of a 27 run fractional-factorial design. This design has 13 three-level factors, ten of which are used to make $2^3 3^3 4^2$. The initial design will be unbalanced and one row short of orthogonal, so we would expect that other methods would be better for this problem. The macro also tells us that it is performing 60 PROC OPTEX searches of 2776 candidates, and that the full-factorial design has 3456 runs. The macro is searching the full-factorial design minus the excluded choice sets. Since the full-factorial design is not too large (less than 5000), and since there is no tabled design that is very good for this problem, this is the kind of problem where we would expect the PROC OPTEX algorithm to work best. The macro chose 60 OPTEX iterations. In the fabric softener example, the macro did not try any OPTEX iterations, because it knew it could directly make a 100% efficient design. In the vacation examples, it ran the default minimum of 20 OPTEX iterations because the macro's heuristics

concluded that OPTEX would probably not be the best approach for those problems. In this example, the macro's heuristics tried more iterations since this is the kind of example where OPTEX works best.

Here is some of the output.

Consumer Food Product Example

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	84.3176		Can
1	2 1	84.3176	84.3176	Conforms
1	End	84.3176		
2	Start	27.8626		Tab,Unb,Ran
2	1 1	76.5332		Conforms
2	End	80.4628		
.				
.				
.				
11	Start	24.5507		Tab,Ran
11	26 1	78.6100		Conforms
11	End	81.8604		
12	Start	26.3898		Ran,Mut,Ann
12	1 1	67.0450		Conforms
12	End	83.0114		
.				
.				
.				
21	Start	45.9310		Ran,Mut,Ann
21	15 1	67.1046		Conforms
21	End	82.1657		

NOTE: Performing 600 searches of 2,776 candidates.

Consumer Food Product Example

Design Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	84.3176	84.3176	Ini
1	Start	84.7548		Can
1	2 1	84.7548	84.7548	Conforms
1	End	84.7548		

Consumer Food Product Example

Design Refinement History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	84.7548	84.7548	Ini
1	Start	84.7548		Pre,Mut,Ann
1	2 1	82.6737		Conforms
1	14 1	84.7548	84.7548	
1	End	82.6386		
.				
.				
.				
8	Start	84.7548		Pre,Mut,Ann
8	2 1	84.7548	84.7548	Conforms
8	14 1	84.7548	84.7548	
8	21 2	84.7548	84.7548	
8	12 3	84.7548	84.7548	
8	12 6	84.7548	84.7548	
8	18 1	84.7548	84.7548	
8	2 2	84.7548	84.7548	
8	End	84.7548		

NOTE: Stopping since it appears that no improvement is possible.

Consumer Food Product Example

The OPTEX Procedure

Class Level Information

Class	Levels	-Values-
x1	4	1 2 3 4
x2	4	1 2 3 4
x3	2	1 2
x4	2	1 2
x5	3	1 2 3
x6	3	1 2 3
x7	2	1 2

Consumer Food Product Example

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	84.7548	71.1686	98.0583	0.9806

Design 1 (`Can`), which was created by the candidate-set search (using PROC OPTEX), had *D*-efficiency or 84.3176, and the macro confirms that the design conforms to our restrictions. The tabled, unbalanced, and random initializations do not work as well. For each design, the macro iteration history states the *D*-efficiency for the initial design (27.8626 in design 2), the *D*-efficiency when the restrictions are met (76.5332, `Conforms`), and the *D*-efficiency for the final design (80.4628). The fully random initialization tends to work a little better than the tabled initialization for this problem, but not as well as PROC OPTEX. At the end of the algorithm search phase, the macro decides to use PROC OPTEX and performs 600 more searches, and it finds a design with 84.7548% *D*-efficiency. The design refinement step fails to improve on the best design. This step took 3.5 minutes.

When You Have a Long Time to Search for an Efficient Design

With a moderate sized candidate set such as this one (2000 to 6000 runs), we might be able to do better with more iterations. To test this, PROC OPTEX was run 10,000 times over the winter holiday vacation, from December 22 through January 2, creating a total of 200,000 designs, 20 designs on each try. Here is a summary of the results.

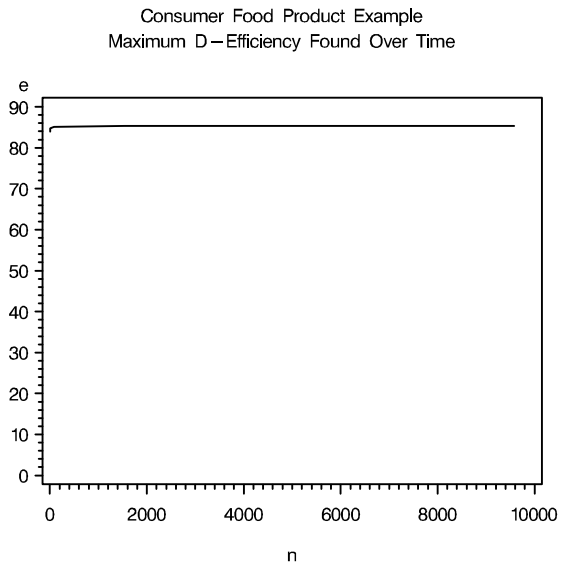
PROC OPTEX Run	*D*-Efficiency	Percent Improvement
1	83.8959	
2	83.9890	0.11%
3	84.3763	0.46%
6	84.7548	0.45%
84	85.1561	0.47%
1535	85.3298	0.20%
9576	85.3985	0.08%

This example is interesting, because it shows the diminishing value of increasing the number of iterations. Six minutes into the search, in the first six passes through PROC OPTEX ($6 \times 20 = 120$ total iterations), we found a design with reasonably good *D*-efficiency=84.7548. Over an hour into the search, with $(84-6) \times 20 = 1560$ more iterations, we get a small 0.47% increase in efficiency to 85.1561. About one day into the search, with $(1535 - 84) \times 20 = 29,020$ more iterations, we get another small 0.20% increase in efficiency, 85.3298. Finally, almost a week into the search, with $(9576-1535) \times 20 = 160,820$ more iterations, we get another small 0.08% increase in efficiency to 85.3985. Our overall improvement over the best design found in 120 iterations was 0.75952%, about three-quarters of a percent. These numbers will change with other problems and other seeds. However, as these results show, usually the first few iterations will give you a good, efficient design, and usually, subsequent iterations will give you slight improvements but with a cost of much greater run times. Next, we will construct a plot of this table.

```
data; input n e; datalines;
    1  83.8959
    2  83.9890
    3  84.3763
    6  84.7548
   84  85.1561
 1535  85.3298
 9576  85.3985
;
proc gplot;
   title  h=1 'Consumer Food Product Example';
   title2 h=1 'Maximum D-Efficiency Found Over Time';
   plot e * n / vaxis=axis1;
   symbol i=join;
   axis1 order=(0 to 90 by 10);
   run; quit;

title2;
```

The plot of maximum *D*-efficiency as a function of PROC OPTEX run number clearly shows that the gain in efficiency that comes from a large number of iterations is very slight.



If you have a lot of time to search for a good design, you can specify some of the time and maximum number of iteration parameters. Sometimes you will get lucky and find a better design. In this next example, `maxtime=300 300 60` was specified. This gives the macro up to 300 minutes for the algorithm search step, 300 minutes for the design search step, and 60 minutes for the refinement step. The option `maxiter=` increases the number iterations to 10000 for each of the three steps (or the maximum time). With this specification, you would expect the macro to run overnight. See the macro documentation (starting on page 546) for more iteration options. Note that you must increase the number of iterations and the maximum amount of time if you want the macro to run longer. With this specification, the macro performs 1800 OPTEX iterations initially (compared to 60 by default).

```
title 'Consumer Food Product Example';

%macro bad;
  bad = (x1 < 4) + (x2 < 4) + (x5 < 3) + (x6 < 3) + (x8 < 3);
  bad = abs(bad - 3) * ((bad < 2) | (bad > 4));
  %mend;

%mktex( 4 4 2 2 3 3 2 3, n=26, interact=x2*x3 x2*x4 x3*x4 x6*x7,
        restrictions=bad, seed=151,
        maxtime=300 300 60, maxiter=10000 )
```

The results from this step are not shown.

Examining the Design

We can use the `%MktEval` macro to start to evaluate the design.

```
%mkteval(data=sasuser.choicdes);
```

Here are the results.

Consumer Food Product Example
Canonical Correlations Between the Factors
There are 4 Canonical Correlations Greater Than 0.316

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1	0.30	0.20	0.11	0.42	0.26	0.09	0.33
x2	0.30	1	0.10	0.10	0.13	0.17	0.51	0.18
x3	0.20	0.10	1	0.08	0.09	0.30	0	0.10
x4	0.11	0.10	0.08	1	0.09	0.10	0	0.10
x5	0.42	0.13	0.09	0.09	1	0.24	0.05	0.43
x6	0.26	0.17	0.30	0.10	0.24	1	0.14	0.13
x7	0.09	0.51	0	0	0.05	0.14	1	0.14
x8	0.33	0.18	0.10	0.10	0.43	0.13	0.14	1

Consumer Food Product Example
Canonical Correlations > 0.316 Between the Factors
There are 4 Canonical Correlations Greater Than 0.316

		r	r Square
x2	x7	0.51	0.26
x5	x8	0.43	0.18
x1	x5	0.42	0.18
x1	x8	0.33	0.11

Consumer Food Product Example
Summary of Frequencies
There are 4 Canonical Correlations Greater Than 0.316
* - Indicates Unequal Frequencies

		Frequencies
*	x1	7 8 6 5
*	x2	6 7 7 6
	x3	13 13
	x4	13 13
*	x5	9 8 9
*	x6	7 10 9
*	x7	12 14
*	x8	7 9 10

```
*    x1 x2    2 2 1 2 2 2 2 2 1 1 2 2 1 2 2 0
*    x1 x3    3 4 4 4 4 2 2 3
*    x1 x4    4 3 4 4 3 3 2 3
*    x1 x5    4 2 1 2 1 5 2 2 2 1 3 1
*    x1 x6    2 3 2 2 4 2 2 1 3 1 2 2
*    x1 x7    3 4 4 4 3 3 2 3
*    x1 x8    1 2 4 2 4 2 2 1 3 2 2 1
*    x2 x3    3 3 3 4 4 3 3 3
*    x2 x4    3 3 3 4 4 3 3 3
*    x2 x5    2 2 2 3 2 2 2 2 3 2 2 2
*    x2 x6    2 2 2 2 3 2 2 2 3 1 3 2
*    x2 x7    1 5 4 3 2 5 5 1
*    x2 x8    2 2 2 1 3 3 2 2 3 2 2 2
*    x3 x4    7 6 6 7
*    x3 x5    5 4 4 4 4 5
*    x3 x6    2 5 6 5 5 3
*    x3 x7    6 7 6 7
*    x3 x8    4 4 5 3 5 5
*    x4 x5    4 4 5 5 4 4
*    x4 x6    4 5 4 3 5 5
*    x4 x7    6 7 6 7
*    x4 x8    4 4 5 3 5 5
*    x5 x6    2 4 3 2 2 4 3 4 2
*    x5 x7    4 5 4 4 4 5
*    x5 x8    1 2 6 4 2 2 2 5 2
*    x6 x7    3 4 4 6 5 4
*    x6 x8    2 2 3 2 4 4 3 3 3
*    x7 x8    4 4 4 3 5 6
     N-Way    1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
              1 1 1 1 1 1 1
```

Some of the canonical correlations are bigger than we would like. They all involve attributes in different alternatives, so they should not pose huge problems. Still, they are large enough to make some researchers uncomfortable. The frequencies are pretty close to balanced. Perfect balance is not possible with 26 choice sets and this design. If we were willing to consider blocking the design, we might do better with more choice sets.

Designing the Choice Experiment, More Choice Sets

Let's run the `%MktRuns` macro to see what looks good. For now, we will ignore the interactions.

```
%mktruns( 4 4 2 2 3 3 2 3 )
```

Consumer Food Product Example

Design Summary

Number of Levels	Frequency
2	3
3	3
4	2

Consumer Food Product Example

Saturated = 16
Full Factorial = 3,456

Some Reasonable Design Sizes	Violations	Cannot Be Divided By
144 *	0	
72	1	16
48	3	9
96	3	9
192	3	9
24	4	9 16
120	4	9 16
168	4	9 16
36	7	8 16
108	7	8 16

* - 100% Efficient Design can be made with the MktEx Macro.

Consumer Food Product Example

n	Design	Reference
144	2 ** 48 3 ** 3 4 ** 2	Orthogonal Array
144	2 ** 44 3 ** 3 4 ** 3	Orthogonal Array
144	2 ** 41 3 ** 4 4 ** 2	Orthogonal Array
144	2 ** 39 3 ** 3 4 ** 2 6 ** 1	Orthogonal Array
144	2 ** 37 3 ** 4 4 ** 3	Orthogonal Array
144	2 ** 37 3 ** 3 4 ** 2 12 ** 1	Orthogonal Array
144	2 ** 35 3 ** 3 4 ** 3 6 ** 1	Orthogonal Array
.		
.		
.		

The smallest suggestion larger than 26 is 36. With this mix of factor levels, we would have to have 144 runs to get an orthogonal design, so we will definitely want to stick with a nonorthogonal design. Balance will be possible in 36 runs, but 36 cannot be divided by $2 \times 4 = 8$ and $4 \times 4 = 16$. With 36 runs, a blocking factor will be required (2 blocks of 18 or 3 blocks of 12). We would like the shelf-talker to appear in half of the choice sets within block, so with two blocks, we will want the number of choice sets to be divisible by $2 \times 2 = 4$, and 36 can be divided by 4. The `%MktRuns` macro cannot provide us with much guidance with the interactions. We "tricked" it in the past by substituting products of levels, like $9 = 3 \times 3$, but in this case, factors like `x2`, `x3`, and `x4` interact multiple times, so it would not be that simple. We will try making a design in 36 runs, and see how it looks.

```
title 'Consumer Food Product Example';

%macro bad;
  bad = (x1 < 4) + (x2 < 4) + (x5 < 3) + (x6 < 3) + (x8 < 3);
  bad = abs(bad - 3) * ((bad < 2) | (bad > 4));
  %mend;

%mktex( 4 4 2 2 3 3 2 3, n=36, interact=x2*x3 x2*x4 x3*x4 x6*x7,
        restrictions=bad, seed=377, outr=sasuser.choicdes )

%mkteval;
```

Here is the last part of the output from the `%MktEx` macro.

Consumer Food Product Example

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	94.3544	88.6147	92.8475	0.8333

D-efficiency at 94.68% looks good. Here is part of the `%MktEval` results.

Consumer Food Product Example
Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1	0.16	0.05	0.05	0.17	0.22	0.20	0.13
x2	0.16	1	0.06	0.06	0.12	0.22	0.13	0.16
x3	0.05	0.06	1	0.00	0.04	0.04	0.06	0.11
x4	0.05	0.06	0.00	1	0.11	0.10	0.06	0.04
x5	0.17	0.12	0.04	0.11	1	0.09	0.12	0.11
x6	0.22	0.22	0.04	0.10	0.09	1	0.07	0.16
x7	0.20	0.13	0.06	0.06	0.12	0.07	1	0.07
x8	0.13	0.16	0.11	0.04	0.11	0.16	0.07	1

Consumer Food Product Example
Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316
* - Indicates Unequal Frequencies

		Frequencies
*	x1	8 8 11 9
*	x2	9 10 8 9
*	x3	17 19
*	x4	17 19
*	x5	9 12 15
*	x6	11 11 14
	x7	18 18
*	x8	10 11 15

The correlations are better, however the balance is worse than we might like. We can run the macro again, this time specifying `balance=2`, which forces better balance. The specification of 2 allows the maximum frequency for a level in a factor to be no more than two greater than the minimum frequency.

```
%mktex( 4 4 2 2 3 3 2 3, n=36, interact=x2*x3 x2*x4 x3*x4 x6*x7,
        restrictions=bad, seed=377, outr=sasuser.choicdes, balance=2 )

%mkteval;
```

Here is the last part of the output from the `%MktEx` macro.

Consumer Food Product Example

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	93.8799	87.7982	93.4280	0.8333

The *D*-efficiency looks good. It is a little lower than before, but not much. Here is the first part of the output from the `%MktEval` macro.

Consumer Food Product Example
Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1	0.13	0.06	0.13	0.16	0.22	0.18	0.07
x2	0.13	1	0.10	0.11	0.17	0.29	0.11	0.21
x3	0.06	0.10	1	0.06	0.15	0.08	0.06	0.04
x4	0.13	0.11	0.06	1	0.12	0.14	0	0.12
x5	0.16	0.17	0.15	0.12	1	0.17	0.07	0.18
x6	0.22	0.29	0.08	0.14	0.17	1	0	0.10
x7	0.18	0.11	0.06	0	0.07	0	1	0.07
x8	0.07	0.21	0.04	0.12	0.18	0.10	0.07	1

The canonical correlations look good. Here is the last part of the output from the `%MktEval` macro.

Consumer Food Product Example
Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316
* - Indicates Unequal Frequencies

Frequencies

*	x1	8 9 9 10
	x2	9 9 9 9
*	x3	17 19
	x4	18 18
*	x5	11 12 13
	x6	12 12 12
	x7	18 18
*	x8	11 13 12

```
*    x1 x2    2 2 2 2 2 2 3 2 2 2 2 3 3 3 2 2
*    x1 x3    4 4 4 5 4 5 5 5
*    x1 x4    4 4 4 5 4 5 6 4
*    x1 x5    2 2 4 3 3 3 2 4 3 4 3 3
*    x1 x6    3 2 3 4 2 3 2 4 3 3 4 3
*    x1 x7    5 3 5 4 4 5 4 6
*    x1 x8    2 3 3 3 3 3 3 3 3 3 4 3
*    x2 x3    5 4 4 5 4 5 4 5
*    x2 x4    5 4 5 4 4 5 4 5
*    x2 x5    3 3 3 3 2 4 3 3 3 2 4 3
*    x2 x6    2 3 4 3 3 3 3 2 4 4 4 1
*    x2 x7    4 5 5 4 5 4 4 5
*    x2 x8    2 4 3 4 3 2 2 3 4 3 3 3
*    x3 x4    9 8 9 10
*    x3 x5    4 6 7 7 6 6
*    x3 x6    5 6 6 7 6 6
*    x3 x7    8 9 10 9
*    x3 x8    5 6 6 6 7 6
*    x4 x5    5 7 6 6 5 7
*    x4 x6    6 7 5 6 5 7
     x4 x7    9 9 9 9
*    x4 x8    5 6 7 6 7 5
*    x5 x6    4 3 4 3 4 5 5 5 3
*    x5 x7    5 6 6 6 7 6
*    x5 x8    4 3 4 3 4 5 4 6 3
     x6 x7    6 6 6 6 6 6
*    x6 x8    4 4 4 3 5 4 4 4 4
*    x7 x8    5 7 6 6 6 6
     N-Way    1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
              1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

This design looks much better. It is possible to get designs with better balance by specifying `balance=1`, but for this problem, the price in efficiency is too high. We do want to ensure that `x4`, the shelf talker factor is balanced, since we will be dividing the design into two parts, depending on whether the shelf talker is there or not. It is balanced in this design. If it had not been, we could have switched it with another two-level factor or tried again with a different seed. If nothing else worked, we could have added it after the fact by blocking (running the `%MktBlock` macro as if we were adding a blocking factor).

The `balance=` option in the `%MktEx` macro works by adding restrictions to the design. The approach it uses often works quite well, but sometimes it does not. Forcing balance gives the macro much less freedom in its search, and makes it easy for the macro to get stuck in suboptimal designs. Most of our restrictions are imposed within each row. Those kinds of restrictions do not pose a problem for the macro. Balance restrictions are imposed across rows within a column. We know of better ways to impose balance, but they tend to be very slow. This is an area where more research is needed, and the way this option works will quite likely be different in future releases. If perfect balance is critical, try the `%MktBal` macro.

Examining the Subdesigns

As we mentioned previously, "it is sometimes necessary and good practice to check the ranks of the subdesigns for more complex models (Lazari and Anderson 1994)." Here is a way to do that with PROC OPTEX. This is the only usage of PROC OPTEX in this book that is too specialized to be run from one of the `%Mkt` macros (because not all variables are designated as `class` variables). For convenience, we call PROC OPTEX from an ad hoc macro, since it must be run five times, once per alternative, with only a change in the `where` statement. We need to evaluate the design when the client's alternative is available `(x1 ne 4)`, when the client line extension alternative is available `(x2 ne 4)`, when the regional competitor is available `(x5 ne 3)`, when the private label competitor is available `(x6 ne 3)`, and when the national competitor is available `(x8 ne 3)`. We need to use a `model` statement that lists all of the main effects and interactions. We do not designate all of the variables on the `class` statement because we only have enough runs to consider linear price effects within each availability group. The statement `generate method=sequential initdesign=desv` specifies that we will be evaluating the initial design `desv`, using the sequential algorithm, which ensures no swaps between the candidate set and the initial design. The other option of note here appears on the `class` statement, and that is `param=orthref`. This specifies an orthogonal parameterization of the effects that gives us a nice 0 to 100 scale for the *D*-efficiencies.

```
%macro evaleff(where);
data desv / view=desv; set sasuser.choicdes(where=(&where)); run;

proc optex data=desv;
   class x3 x4 x7 / param=orthref;
   model x1-x8 x2*x3 x2*x4 x3*x4 x6*x7;
   generate method=sequential initdesign=desv;
   run; quit;

%mkteval(data=desv)
%mend;

%evaleff(x1 ne 4)
%evaleff(x2 ne 4)
%evaleff(x5 ne 3)
%evaleff(x6 ne 3)
%evaleff(x8 ne 3)
```

Each step took just over two seconds. We hope to not see any efficiencies of zero, and we hope to not get the message `WARNING: Can't estimate model parameters in the final design.` Here are some of the results.

Consumer Food Product Example

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	67.3274	57.4665	83.6185	0.7071

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	73.8239	67.9782	87.2117	0.6939

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	68.8274	57.9474	82.7738	0.7518

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	72.0566	60.2545	87.0855	0.7360

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	68.8448	56.4859	86.7440	0.7360

Examining the Aliasing Structure

It is also good to look at the aliasing structure of the design. We use PROC GLM to do this, so we must create a dependent variable. We will use a constant `y=1`. The first PROC GLM step just checks the model to make sure none of the specified effects are aliased with each other. This step is not necessary since our *D*-efficiency value greater than zero already guarantees this.

```
data temp;
   set sasuser.choicdes;
   y = 1;
   run;
proc glm data=temp;
   model y = x1-x8 x2*x3 x2*x4 x3*x4 x6*x7 / e aliasing;
   run; quit;
```

Here are the results, ignoring the ANOVA and regression tables, which are not of interest. Each of these lines is a linear combination that is estimable. It is simply a list of the effects.

```
Intercept
x1
x2
x3
x4
x5
x6
x7
x8
x2*x3
x2*x4
x3*x4
x6*x7
```

Contrast this with a specification that includes all simple effects and two-way and three-way interactions. We specify the model of interest first, `x1-x8 x2*x3 x2*x4 x3*x4 x6*x7`, so all of those terms will be listed first, then we specify all main effects and two-way and three-way interactions using the notation `x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8@3`. It is not a problem that some of the terms were both explicitly specified and also generated by the `x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8@3` list since PROC GLM automatically eliminates duplicate terms.

```
proc glm data=temp;
   model y = x1-x8 x2*x3 x2*x4 x3*x4 x6*x7
             x1|x2|x3|x4|x5|x6|x7|x8@3 / e aliasing;
   run; quit;
```

```
Intercept + 19.841*x4*x6 + 92.073*x1*x4*x6 + 77.513*x2*x4*x6 + 33.148*x3*x4*x6 -
56.771*x5*x6 - 99.617*x1*x5*x6 - 202.92*x2*x5*x6 - 86.937*x3*x5*x6 -
58.893*x4*x5*x6 - 54.98*x1*x7 + 34.901*x2*x7 - 75.033*x1*x2*x7 - 27.184*x3*x7 -
195.12*x1*x3*x7 - 29.802*x2*x3*x7 - 4.7757*x4*x7 - 114.1*x1*x4*x7 +
59.836*x2*x4*x7 - 25.33*x3*x4*x7 + 26.719*x5*x7 - 44.435*x1*x5*x7 +
153.6*x2*x5*x7 + 11.575*x3*x5*x7 + 43.714*x4*x5*x7 - 61.007*x1*x6*x7 +
55.783*x2*x6*x7 - 65.069*x3*x6*x7 + 31.255*x4*x6*x7 - 3.375*x5*x6*x7 +
44.572*x1*x8 + 97.812*x2*x8 + 371.54*x1*x2*x8 + 8.5783*x3*x8 + 88.885*x1*x3*x8 +
205.6*x2*x3*x8 - 62.399*x4*x8 - 110.69*x1*x4*x8 - 7.1257*x2*x4*x8 -
107.69*x3*x4*x8 - 30.287*x5*x8 + 6.4007*x1*x5*x8 + 132.7*x2*x5*x8 -
20.893*x3*x5*x8 - 143.95*x4*x5*x8 - 98.515*x6*x8 - 160.95*x1*x6*x8 +
26.253*x2*x6*x8 - 73.35*x3*x6*x8 - 170.62*x4*x6*x8 - 366.07*x5*x6*x8 +
8.8425*x7*x8 - 42.501*x1*x7*x8 + 269.69*x2*x7*x8 - 51.733*x3*x7*x8 -
92.401*x4*x7*x8 + 21.357*x5*x7*x8 - 127.07*x6*x7*x8
```

```
x1 - 29.027*x4*x6 - 83.628*x1*x4*x6 - 72.701*x2*x4*x6 - 44.896*x3*x4*x6 +
5.2367*x5*x6 + 22.835*x1*x5*x6 + 28.119*x2*x5*x6 + 10.102*x3*x5*x6 -
56.162*x4*x5*x6 + 20.701*x1*x7 - 8.8641*x2*x7 + 20.33*x1*x2*x7 - 2.5242*x3*x7 +
32.551*x1*x3*x7 - 14.5*x2*x3*x7 + 3.7689*x4*x7 + 33.213*x1*x4*x7 -
13.188*x2*x4*x7 + 3.7564*x3*x4*x7 - 13.69*x5*x7 + 2.5647*x1*x5*x7 -
55.758*x2*x5*x7 - 31.65*x3*x5*x7 - 17.662*x4*x5*x7 + 37.267*x1*x6*x7 -
17.642*x2*x6*x7 - 2.1979*x3*x6*x7 - 38.7*x4*x6*x7 - 20.918*x5*x6*x7 +
32.967*x1*x8 - 42.314*x2*x8 - 62.836*x1*x2*x8 + 9.0633*x3*x8 + 61.14*x1*x3*x8 -
63.984*x2*x3*x8 + 8.2713*x4*x8 + 78.112*x1*x4*x8 - 41.884*x2*x4*x8 +
33.735*x3*x4*x8 + 26.387*x5*x8 + 124.95*x1*x5*x8 - 27.697*x2*x5*x8 +
63.499*x3*x5*x8 + 49.955*x4*x5*x8 + 4.8752*x6*x8 + 105.91*x1*x6*x8 -
76.88*x2*x6*x8 + 24.902*x3*x6*x8 - 37.29*x4*x6*x8 + 80.345*x5*x6*x8 +
7.1641*x7*x8 + 104.12*x1*x7*x8 - 82.917*x2*x7*x8 + 22.174*x3*x7*x8 +
39.304*x4*x7*x8 + 38.632*x5*x7*x8 + 20.425*x6*x7*x8
.
.
.
```

Again, we have a list of linear combinations that are estimable. This shows that the `Intercept` cannot be estimated independently of the `x4*x6` interaction and a bunch of others including four-way though eight-way interactions which were not specified and hence not shown. Similarly, `x1` is confounded with a bunch of interactions, and so on. This is why we want to be estimable the two-way interactions between factors that are combined to create an alternative. We did not want something like `x2*x3`, the client-line extension's price and microwave/stove top interaction to be confounded with say another brand's price.

Blocking the Design

At 36 choice sets, this design is a bit large, so we will block it into two blocks of 18 choice sets. Within each block we will want the shelf talker to be on half the time.

```
%mktblock(data=sasuser.choicdes, out=sasuser.blockdes, nblocks=2, seed=289)
```

The first attempt (not shown) produced a design where `x4`, the shelf talker did not occur equally often within each block. Changing the seed took care of the problem. Here are the canonical correlations.

Consumer Food Product Example
Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

	Block	x1	x2	x3	x4	x5	x6	x7	x8
Block	1	0.08	0.11	0.06	0	0.07	0	0	0.07
x1	0.08	1	0.13	0.06	0.13	0.16	0.22	0.18	0.07
x2	0.11	0.13	1	0.10	0.11	0.17	0.29	0.11	0.21
x3	0.06	0.06	0.10	1	0.06	0.15	0.08	0.06	0.04
x4	0	0.13	0.11	0.06	1	0.12	0.14	0	0.12
x5	0.07	0.16	0.17	0.15	0.12	1	0.17	0.07	0.18
x6	0	0.22	0.29	0.08	0.14	0.17	1	0	0.10
x7	0	0.18	0.11	0.06	0	0.07	0	1	0.07
x8	0.07	0.07	0.21	0.04	0.12	0.18	0.10	0.07	1

The blocking variable is not highly correlated with any of the factors. Here are some of the frequencies.

Consumer Food Product Example
Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316
* - Indicates Unequal Frequencies

		Frequencies
	Block	18 18
*	x1	8 9 9 10
	x2	9 9 9 9
*	x3	17 19
	x4	18 18
*	x5	11 12 13
	x6	12 12 12
	x7	18 18
*	x8	11 13 12
*	Block x1	4 5 4 5 4 4 5 5
*	Block x2	4 5 5 4 5 4 4 5
*	Block x3	8 10 9 9
	Block x4	9 9 9 9
*	Block x5	6 6 6 5 6 7
	Block x6	6 6 6 6 6 6
	Block x7	9 9 9 9
*	Block x8	6 6 6 5 7 6

The blocking variable is perfectly balanced, as it is guaranteed to be if the number of blocks divides the number of runs. Balance within blocks, that is the cross-tabulations of the factors with the blocking variable, looks good. The macro also prints canonical correlations within blocking variables. These can sometimes be quite high, even 1.0, but that is *not* a problem.[§] Here is the design, as it is printed by the `%MktBlock` macro.

Consumer Food Product Example

Block	Run	x1	x2	x3	x4	x5	x6	x7	x8
1	1	3	2	2	2	2	2	2	3
	2	1	1	1	2	2	3	1	3
	3	4	2	2	1	1	2	1	1
	4	1	4	1	1	3	1	1	3
	5	2	2	2	1	3	1	2	1
	6	1	1	2	1	1	2	2	3
	7	3	3	2	1	2	2	1	3
	8	2	2	1	2	3	1	1	1
	9	4	2	2	2	1	3	1	2
	10	2	1	2	1	2	3	1	2
	11	1	4	2	2	3	2	1	2
	12	2	3	1	1	1	1	2	3
	13	2	3	2	2	1	3	1	1
	14	3	4	1	2	1	1	2	2
	15	4	1	1	1	2	1	2	1
	16	4	4	1	2	3	2	2	1
	17	4	3	2	1	3	3	2	2
	18	3	3	1	2	2	3	2	2

[§]Ideally, each subject would only make one choice, since the choice model is based on this assumption (which is almost always ignored). As the number of blocks increases, the correlations will mostly go to one, and ultimately be undefined when there is only one choice set per block.

Consumer Food Product Example

Block	Run	x1	x2	x3	x4	x5	x6	x7	x8
2	1	1	2	1	2	1	3	2	1
	2	3	1	2	2	3	3	2	1
	3	1	3	2	2	3	1	2	2
	4	2	1	1	2	3	2	2	2
	5	4	1	1	1	1	3	2	3
	6	2	4	2	2	2	3	2	3
	7	4	2	1	1	2	2	2	2
	8	4	1	2	2	1	1	1	2
	9	2	4	1	1	2	2	1	2
	10	3	4	2	1	1	2	2	1
	11	1	3	1	1	2	3	1	1
	12	3	4	2	2	2	1	1	1
	13	4	4	2	1	2	1	2	3
	14	4	3	1	2	3	2	1	3
	15	3	2	1	1	3	3	1	3
	16	3	1	1	1	3	2	1	2
	17	1	2	2	1	3	1	1	2
	18	2	3	2	2	1	1	1	3

The Final Design

The following code creates the final choice design, stored in SASUSER.FINCHDES, sorted by the blocking and shelf-talker variable. We will use the `%MktLab` macro to assign values, formats, and labels to the design. Previously, we have used the `%MktLab` macro to reassign factor names when we wanted something more descriptive than the default, `x1`, `x2`, and so on, and when we wanted to reassign the names of two m-level factors to minimize the problems associated with correlated factors. This time, we will use the `%MktLab` macro primarily to deal with the asymmetry in the price factors. Recall our factor levels.

Factors and Levels

Alternative	Factor	Levels	Brand	Description
1	X1	4	Client	1.29, 1.69, 2.09 + absent
2	X2	4	Client Line Extension	1.39, 1.89, 2.39, + absent
	X3	2		microwave/stove-top
	X4	2		shelf-talker yes/no
3	X5	3	Regional	1.99, 2.49 + absent
4	X6	3	Private Label	1.49, 2.29 absent
	X7	2		microwave/stove-top
5	X8	3	National	1.99 + 2.39 + absent

The choice design will need a quantitative price factor, make from all five of the linear price factors, that contains the prices of each of the alternatives. At this point, our factor `x1` contains 1, 2, 3, 4, and not 1.29, 1.69, 2.09, and absent, which is different from `x2` and from all of the other factors. A 1 in `x1` will need to become a price of 1.29 in the choice design, a 1 in `x2` will need to become a price of 1.39 in the choice design, a 1 in `x3` will need to become a price of 1.99 in the choice design, and so on. Before we use the `%MktRoll` macro to turn the linear design into a choice design, we need to use the `%MktLab` macro to assign the actual prices to the price factors.

The `%MktLab` macro is like the `%MktRoll` macro in the sense that it can use as input a `key=` data set that contains the rules for customizing a design for our particular usage. In the `%MktRoll` macro, the `key=` data set provides the rules for turning a linear design into a choice design. In contrast, in the `%MktLab` macro, the `key=` data set contains the rules for turning a linear design into another linear design, changing one or more of the following: factor names, factor levels, factor types (numeric to character), level formats, and factor labels.

We could use the `%MktLab` macro to change the names of the variables and their types, but we will not do that for this example. Ultimately, we will use the `%MktRoll` macro to assign all of the price factors to a variable called `Price` and similarly provide meaningful names for all of the factors in the choice design, just as we have in previous examples. We could also change a variable like `x3` with values of 1 and 2 to something like `Stove` with values `'Stove'` and `'Micro'`. We will not do that because we want to make a design with a simple list of numeric factors, with simple names like `x1-x8` that we can run through the `%MktRoll` macro to get the final choice design. We will assign formats and labels, so we can print the design in a meaningful way, but ultimately, our only goal at this step is to handle the price asymmetries by assigning the actual price values to the factors.

The `key=` data set contains the rules for customizing our design. The data set has as many rows as the maximum number of levels, in this case four. Each variable is one of the factors in the design, and the values are the factor levels that we want in the final design. The first factor, `x1`, is the price factor for the client brand. Its levels are 1.29, 1.69, and 2.09. In addition, one level is 'not available', which is flagged by the SAS special missing value `.N`. In order to read special missing values in an input data set, you must use the `missing` statement and name the expected missing values. The factor `x2` has the same structure as `x1`, but with different levels. The factor `x3` has two levels, hence the `key=` data set has missing values in the third and fourth row. Since the design has only 1's and 2's for `x3`, this missing values will never be used. Notice that we are keeping `x3` as a numeric variable with values 1 and 2 using a format to supply the character levels 'micro' and 'stove'. The other factors are created in a similar fashion. By default, ordinary missing values `'.'` are not permitted as levels. By default, you may only use ordinary missing values as place holders for factors that have fewer levels than the maximum. If you want missing values in the levels, you must use one of the special missing values `.A`, `.B`, ..., `.Z`, and `._`[¶] or the `cfill=` or `nfill=` options.

The `%MktLab` macro specification names the input SAS data set with the design and the key data set. By default, it creates an output SAS data set called FINAL. The data set is sorted by block and shelf talker and printed.

```
proc format;
   value yn    1 = 'No'    2 = 'Talker';
   value micro 1 = 'Micro' 2 = 'Stove';
   run;
```

[¶]Note that the `'.'` in `'.N'` is not typed in the data, nor is it typed in the `missing` statement. Furthermore, it does not appear in the printed output. However, you need to type it if you ever refer to a special missing value in code: `if x1 eq .N then ...`.

```
data key;
   missing N;
   input x1-x8;
   format x1 x2 x5 x6 x8 dollar5.2
          x4 yn. x3 x7 micro.;
   label x1 = 'Client Brand'
         x2 = 'Client Line Extension'
         x3 = 'Client Micro/Stove'
         x4 = 'Shelf Talker'
         x5 = 'Regional Brand'
         x6 = 'Private Label'
         x7 = 'Private Micro/Stove'
         x8 = 'National Competitor';
   datalines;
1.29 1.39 1 1 1.99 1.49 1 1.99
1.69 1.89 2 2 2.49 2.29 2 2.39
2.09 2.39 . . N    N    .    N
N    N    . . .    .    .    .
;

%mktlab(data=sasuser.blockdes, key=key)

proc sort out=sasuser.finchdes(drop=run); by block x4; run;

proc print label; id block x4; by block x4; run;
```

The %MktLab macro prints the variable mapping that it uses, old names followed by new names. In this case, none of the names change, but it is good to make sure that you have the expected correspondence.

```
Variable Mapping:
   x1 : x1
   x2 : x2
   x3 : x3
   x4 : x4
   x5 : x5
   x6 : x6
   x7 : x7
   x8 : x8
```

Here is the design.

Consumer Food Product Example

Block	Shelf Talker	Client Brand	Client Line Extension	Client Micro/ Stove	Regional Brand	Private Label	Private Micro/ Stove	National Competitor
1	No	N	$1.89	Stove	$1.99	$2.29	Micro	$1.99
		$1.29	N	Micro	N	$1.49	Micro	N
		$1.69	$1.89	Stove	N	$1.49	Stove	$1.99
		$1.29	$1.39	Stove	$1.99	$2.29	Stove	N
		$2.09	$2.39	Stove	$2.49	$2.29	Micro	N
		$1.69	$1.39	Stove	$2.49	N	Micro	$2.39
		$1.69	$2.39	Micro	$1.99	$1.49	Stove	N
		N	$1.39	Micro	$2.49	$1.49	Stove	$1.99
		N	$2.39	Stove	N	N	Stove	$2.39
1	Talker	$2.09	$1.89	Stove	$2.49	$2.29	Stove	N
		$1.29	$1.39	Micro	$2.49	N	Micro	N
		$1.69	$1.89	Micro	N	$1.49	Micro	$1.99
		N	$1.89	Stove	$1.99	N	Micro	$2.39
		$1.29	N	Stove	N	$2.29	Micro	$2.39
		$1.69	$2.39	Stove	$1.99	N	Micro	$1.99
		$2.09	N	Micro	$1.99	$1.49	Stove	$2.39
		N	N	Micro	N	$2.29	Stove	$1.99
		$2.09	$2.39	Micro	$2.49	N	Stove	$2.39
2	No	N	$1.39	Micro	$1.99	N	Stove	N
		N	$1.89	Micro	$2.49	$2.29	Stove	$2.39
		$1.69	N	Micro	$2.49	$2.29	Micro	$2.39
		$2.09	N	Stove	$1.99	$2.29	Stove	$1.99
		$1.29	$2.39	Micro	$2.49	N	Micro	$1.99
		N	N	Stove	$2.49	$1.49	Stove	N
		$2.09	$1.89	Micro	N	N	Micro	N
		$2.09	$1.39	Micro	N	$2.29	Micro	$2.39
		$1.29	$1.89	Stove	N	$1.49	Micro	$2.39
2	Talker	$1.29	$1.89	Micro	$1.99	N	Stove	$1.99
		$2.09	$1.39	Stove	N	N	Stove	$1.99
		$1.29	$2.39	Stove	N	$1.49	Stove	$2.39
		$1.69	$1.39	Micro	N	$2.29	Stove	$2.39
		$1.69	N	Stove	$2.49	N	Stove	N
		N	$1.39	Stove	$1.99	$1.49	Micro	$2.39
		$2.09	N	Stove	$2.49	$1.49	Micro	$1.99
		N	$2.39	Micro	N	$2.29	Micro	N
		$1.69	$2.39	Stove	$1.99	$1.49	Micro	N

In contrast, here are the actual values without formats and labels.

```
proc print data=sasuser.finchdes; format _numeric_; run;
```

Consumer Food Product Example

Obs	x1	x2	x3	x4	x5	x6	x7	x8	Block
1	N	1.89	2	1	1.99	2.29	1	1.99	1
2	1.29	N	1	1	N	1.49	1	N	1
3	1.69	1.89	2	1	N	1.49	2	1.99	1
4	1.29	1.39	2	1	1.99	2.29	2	N	1
5	2.09	2.39	2	1	2.49	2.29	1	N	1
6	1.69	1.39	2	1	2.49	N	1	2.39	1
7	1.69	2.39	1	1	1.99	1.49	2	N	1
8	N	1.39	1	1	2.49	1.49	2	1.99	1
9	N	2.39	2	1	N	N	2	2.39	1
10	2.09	1.89	2	2	2.49	2.29	2	N	1
11	1.29	1.39	1	2	2.49	N	1	N	1
12	1.69	1.89	1	2	N	1.49	1	1.99	1
13	N	1.89	2	2	1.99	N	1	2.39	1
14	1.29	N	2	2	N	2.29	1	2.39	1
15	1.69	2.39	2	2	1.99	N	1	1.99	1
16	2.09	N	1	2	1.99	1.49	2	2.39	1
17	N	N	1	2	N	2.29	2	1.99	1
18	2.09	2.39	1	2	2.49	N	2	2.39	1
19	N	1.39	1	1	1.99	N	2	N	2
20	N	1.89	1	1	2.49	2.29	2	2.39	2
21	1.69	N	1	1	2.49	2.29	1	2.39	2
22	2.09	N	2	1	1.99	2.29	2	1.99	2
23	1.29	2.39	1	1	2.49	N	1	1.99	2
24	N	N	2	1	2.49	1.49	2	N	2
25	2.09	1.89	1	1	N	N	1	N	2
26	2.09	1.39	1	1	N	2.29	1	2.39	2
27	1.29	1.89	2	1	N	1.49	1	2.39	2
28	1.29	1.89	1	2	1.99	N	2	1.99	2
29	2.09	1.39	2	2	N	N	2	1.99	2
30	1.29	2.39	2	2	N	1.49	2	2.39	2
31	1.69	1.39	1	2	N	2.29	2	2.39	2
32	1.69	N	2	2	2.49	N	2	N	2
33	N	1.39	2	2	1.99	1.49	1	2.39	2
34	2.09	N	2	2	2.49	1.49	1	1.99	2
35	N	2.39	1	2	N	2.29	1	N	2
36	1.69	2.39	2	2	1.99	1.49	1	N	2

One issue remains to be resolved regarding this design and that concerns the role of the shelf-talker when the client line extension is not available. The second part of each block of the design consists of choice sets in which the shelf-talker is present and calls attention to the client line extension. However, in five of those choice sets, the client line extension is unavailable. This problem can be handled in several ways. Here are a few:

- Rerun the design creation and evaluation programs excluding all choice sets with shelf-talker present and client line extension unavailable. However, this requires changing the model because the excluded cell will make inestimable the interaction between client-line-extension price and shelf-talker. Furthermore, the shelf-talker variable will almost certainly no longer be balanced.
- Move the choice sets with client line extension unavailable to the no-shelf-talker block and rerandomize. The shelf-talker is then on for all of the last nine choice sets.
- Let the shelf-talker go on and off as needed.
- Let the shelf-talker call attention to a brand that happens to be out of stock. It is easy to imagine this happening in a real store.

Other options are available as well. No one approach is obviously superior to the alternatives. For this example, we will take the latter approach and allow the shelf-talker to be on even when the client line extension is not available. Note that if the shelf-talker is turned off when the client line extension is not available then the design must be manually modified to reflect this fact.

Testing the Design Before Data Collection

This is a complicated design that will be used to fit a complicated model with alternative specific and availability cross effects. Collecting data is time consuming and expensive. It is good practice, particularly when there are cross effects, to make sure that the design will work with the most complicated model that we anticipate fitting. We saw, starting on page 195, an example of generating artificial data to test the design before collecting real data. Here, we will explore an alternative approach. Before we collect any data, we will convert the linear design to a choice design* and use the `%ChoicEff` macro to evaluate its efficiency for a multinomial logit model with availability cross effects.

For analysis, the design will have four factors, `Brand`, `Price`, `Micro`, `Shelf`. We will use the `%MktRoll` macro and a `key=` data set (although not the same one as before) to make the choice design. `Brand` is the alternative name; its values are directly read from the `key=KEY` in-stream data. `Price` is an attribute whose values will be constructed from the factors `x1`, `x2`, `x5`, `x6`, and `x8` in SASUSER.FINCHDES data set. `Micro`, the microwave factor, is constructed from `x3` for the client line extension and `x7` for the private label. `Shelf`, the shelf talker factor, is created from `x4` for the extension. The `keep=` option on the `%MktRoll` macro is used to keep the original price factors in the design, since we will need them for the price effects. Normally, they would be dropped.

*See page 87 for an illustration of linear versus choice designs.

```
data key;
   input Brand $ 1-10 (Price Micro Shelf) ($);
   datalines;
Client     x1 .  .
Extension  x2 x3 x4
Regional   x5 .  .
Private    x6 x7 .
National   x8 .  .
None       .  .  .
;

%mktroll(design=sasuser.finchdes, key=key, alt=brand, out=rolled,
         keep=x1 x2 x5 x6 x8)

proc print data=sasuser.finchdes(obs=2); run;

proc print data=rolled(obs=12);
   format price dollar5.2 shelf yn. micro micro.;
   id set; by set;
   run;
```

Consider the first two choice sets in the linear design.

Consumer Food Product Example

Obs	x1	x2	x3	x4	x5	x6	x7	x8	Block
1	N	$1.89	Stove	No	$1.99	$2.29	Micro	$1.99	1
2	$1.29	N	Micro	No	N	$1.49	Micro	N	1

Here they are in the rolled out choice design.

Consumer Food Product Example

Set	Brand	Price	Micro	Shelf	x1	x2	x5	x6	x8
1	Client	N	.	.	N	$1.89	$1.99	$2.29	$1.99
	Extension	$1.89	Stove	No	N	$1.89	$1.99	$2.29	$1.99
	Regional	$1.99	.	.	N	$1.89	$1.99	$2.29	$1.99
	Private	$2.29	Micro	.	N	$1.89	$1.99	$2.29	$1.99
	National	$1.99	.	.	N	$1.89	$1.99	$2.29	$1.99
	None	.	.	.	N	$1.89	$1.99	$2.29	$1.99

2	Client	$1.29	.	.	$1.29	N	N	$1.49	N
	Extension	N	Micro	No	$1.29	N	N	$1.49	N
	Regional	N	.	.	$1.29	N	N	$1.49	N
	Private	$1.49	Micro	.	$1.29	N	N	$1.49	N
	National	N	.	.	$1.29	N	N	$1.49	N
	None	.	.	.	$1.29	N	N	$1.49	N

Set 1, Alternative 1

`Brand`	=	'Client'	the brand for this alternative
`Price`	=	`x1` = N	the price of this alternative
`Micro`	=	.	does not apply to this brand
`Shelf`	=	.	does not apply to this brand
`x1`	=	N	client brand unavailable in this choice set
`x2`	=	$1.89	the price of the extension in this choice set
`x5`	=	$1.99	the price of the regional competitor in this choice set
`x6`	=	$2.29	the price of the private label in this choice set
`x8`	=	$1.99	national competitor unavailable in this choice set

Set 1, Alternative 2

`Brand`	=	'Extension'	the brand for this alternative
`Price`	=	`x2` = $1.89	the price of this alternative
`Micro`	=	Stove	Stove top version
`Shelf`	=	No	Shelf Talker, No
`x1`	=	N	client brand unavailable in this choice set
`x2`	=	$1.89	the price of the extension in this choice set
`x5`	=	$1.99	the price of the regional competitor in this choice set
`x6`	=	$2.29	the price of the private label in this choice set
`x8`	=	$1.99	national competitor unavailable in this choice set

Notice that `x1` through `x8` are constant within each choice set. The variable `x1` is the price of alternative one, which is the same no matter which alternative it is stored with.

We need to do a few more things to this design before we are ready to use it. Since we will be treating all of the price factors as a quantitative (not as `class` variables), we need to convert the missing prices to zero. We also need to convert the missings for when `Micro` and `Shelf` do not apply to 2 for 'Stove' and 1 for 'No'. We also need to assign formats. Eventually, we will also need to output just the alternatives that are available (those with a nonzero price and also the none alternative). For now, we will just make a variable `w` that flags the available alternatives (`w` = 1).

```
data sasuser.choicedes(drop=i);
   set rolled;
   array x[6] price x1 -- x8;
   do i = 1 to 6; if nmiss(x[i]) then x[i] = 0; end;
   if nmiss(micro) then micro = 2;
   if nmiss(shelf) then shelf = 1;
   w = brand eq 'None' or price ne 0;
   format price dollar5.2 shelf yn. micro micro.;
   run;

proc print data=sasuser.choicedes(obs=12); by set; id set; run;
```

Here are the first two choice sets.

Consumer Food Product Example

Set	Brand	Price	Micro	Shelf	x1	x2	x5	x6	x8	w
1	Client	$0.00	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	0
	Extension	$1.89	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1
	Regional	$1.99	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1
	Private	$2.29	Micro	No	$0.00	$1.89	$1.99	$2.29	$1.99	1
	National	$1.99	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1
	None	$0.00	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1
2	Client	$1.29	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	1
	Extension	$0.00	Micro	No	$1.29	$0.00	$0.00	$1.49	$0.00	0
	Regional	$0.00	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	0
	Private	$1.49	Micro	No	$1.29	$0.00	$0.00	$1.49	$0.00	1
	National	$0.00	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	0
	None	$0.00	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	1

Now our choice design is done except for the final coding for the analysis. We can now use the `%ChoicEff` macro to evaluate our choice design. Normally, you would use this macro to search a candidate set for an efficient choice design. You can also use it to evaluate a design created by other means. Here is some sample code, omitting for now the details of the model (indicated by `model= ...`). The complicated part of this is the model due to the alternative-specific price effects and cross effects. For now, let's concentrate on everything else.

```
%choiceff(data=sasuser.choicedes,
          model= ..., /* model specification skipped for now */
          nsets=36, nalts=6, weight=w,
          beta=zero, init=sasuser.choicedes(keep=set),
          intiter=0);
```

The way you check the efficiency of a design like this is to first name it on the `data=` option. This will be the candidate set that contains all of the choice sets that we will consider. In addition, the same design is named on the `init=` option. The full specification is `init=sasuser.choicedes(keep=set)`. Just the variable `Set` is kept. It will be used to bring in just the indicated choice sets from the `data=` design, which in this case is all of them. The option `nsets=36` specifies the number of choice sets, and

`nalts=6` specifies the number of alternatives. This macro requires a constant number of alternatives in each choice set for ease of data management. However, not all of the alternatives have to be used. In this case, we have an availability study. We need to keep the unavailable alternatives in the design for this step, but we do not want them to contribute to the analysis, so we specify a weight variable with `weight=w` and flag the available alternatives with `w=1` and the unavailable alternatives with `w=0`. The option `beta=zero` specifies that we are assuming for design evaluation purpose all zero betas. We can specify other values and get other results for the variances and standard errors. Finally, we specify `intiter=0` which specifies zero internal iterations. We use zero internal iterations when we want to evaluate an initial design, but not attempt to improve it. Here is the actual specification we will use, complete with the model specification.

```
%choiceff(data=sasuser.choicedes,
          model=class(brand / zero='None')
                class(brand / zero='None' separators='' ' ') *
                   identity(price)
                class(shelf micro / lprefix=5 0 zero='No' 'Stove')
                identity(x1 x2 x5 x6 x8) *
                   class(brand / zero='None' separators=' ' ' on ') /
                lprefix=0 order=data,
          nsets=36, nalts=6, weight=w,
          beta=zero, init=sasuser.choicedes(keep=set),
          intiter=0);
```

The specification `class(brand / zero='None')` specifies the brand effects. This specification will create indicator variables for brand with the constant alternative being the reference brand. The option `zero='None'` ensures that the reference level will be `'None'` instead of the default last sorted level (`'Regional'`). Indicator variables will be created for the brands Client, Extension, Regional, Private, and National, but not None. The `zero='None'` option, like `zero='Home'` and other `zero=`*'literal-string'* options we have used in previous examples, names the actual formatted value of the `class` variable that should be excluded from the coded variables because the coefficient will be zero. Do not confuse `zero=none` and `zero='None'`. The `zero=none` option specifies that you want all indicator variables to be created, even including one for the last level. In contrast, the option `zero='None'` (or `zero=` any quoted string) names a specific formatted value, in this case `'None'`, for which indicator variables are not to be created.

The specification `class(brand / ...) * identity(price)` creates the alternative-specific price effects. They are specified as an interaction between a categorical variable `Brand` and a quantitative factor `Price`. The `separators='' ' '` option in the `class` specification specifies the separators that are used to construct the labels for the main effect and interaction terms. The main-effects separator, which is the first `separators=` value, `''`, is ignored since `lprefix=0`. Specifying `' '` as the second value creates labels of the form *brand-blank-price* instead of the default *brand-blank-asterisk-blank-price*.

The specification `class(shelf micro / ...)` names the shelf talker and microwave variables as categorical variables and creates indicator variables for the `'Talker'` category, not the `'No'` category and the `'Micro'` category not the `'Stove'` category. In `zero='No' 'Stove'`, the `'No'` applies to the first variable, `Shelf` and the second value, `'Stove'`, applies to second variable, `Micro`.

The specification `identity(x1 x2 x5 x6 x8) * class(brand / ...)` creates the cross effects. The `separators=` option is specified with a second value of `' on '` to create cross effect labels like `'Client on Extension'`. More will be said on the cross effects when we look at the actual coded values in the next few pages.

Note that PROC TRANSREG produces the following warning twice.

```
WARNING: This usage of * sets one group's slope to zero.  Specify |
         to allow all slopes and intercepts to vary.  Alternatively,
         specify CLASS(vars) * identity(vars) identity(vars) for
         separate within group functions and a common intercept.
         This is a change from Version 6.
```

This is because on two occasions `class` was interacted with `identity` using the asterisk instead of the vertical bar. In a linear model, this may be a sign of a coding error, so the procedure prints a warning. If you get this warning while coding a choice model specifying `zero=`*'constant-alternative-level'*, you can safely ignore it. Still, it is always good to print out one or more coded choice sets to check the coding as we will do later. Here is the last part of the output from the `%ChoicEff` macro.

Consumer Food Product Example

n	Variable Name	Label	Variance	DF	Standard Error
1	BrandClient	Client	9.0879	1	3.01462
2	BrandExtension	Extension	9.0029	1	3.00049
3	BrandRegional	Regional	27.1773	1	5.21319
4	BrandPrivate	Private	8.9994	1	2.99991
5	BrandNational	National	36.4079	1	6.03389
6	BrandClientPrice	Client Price	2.3795	1	1.54256
7	BrandExtensionPrice	Extension Price	1.4509	1	1.20452
8	BrandRegionalPrice	Regional Price	4.5531	1	2.13381
9	BrandPrivatePrice	Private Price	1.6487	1	1.28403
10	BrandNationalPrice	National Price	6.4858	1	2.54672
11	ShelfTalker	Shelf Talker	0.9072	1	0.95249
12	MicroMicro	Micro	0.4957	1	0.70408
13	x1BrandClient	Client Brand on Client	.	0	.
14	x1BrandExtension	Client Brand on Extension	0.4377	1	0.66156
15	x1BrandRegional	Client Brand on Regional	0.4908	1	0.70054
16	x1BrandPrivate	Client Brand on Private	0.4724	1	0.68730
17	x1BrandNational	Client Brand on National	0.5053	1	0.71082
18	x2BrandClient	Client Line Extension on Client	0.3856	1	0.62097
19	x2BrandExtension	Client Line Extension on Extension	.	0	.
20	x2BrandRegional	Client Line Extension on Regional	0.4397	1	0.66310
21	x2BrandPrivate	Client Line Extension on Private	0.3720	1	0.60990
22	x2BrandNational	Client Line Extension on National	0.6656	1	0.81587
23	x5BrandClient	Regional Brand on Client	0.2677	1	0.51741
24	x5BrandExtension	Regional Brand on Extension	0.2617	1	0.51158
25	x5BrandRegional	Regional Brand on Regional	.	0	.
26	x5BrandPrivate	Regional Brand on Private	0.2742	1	0.52365
27	x5BrandNational	Regional Brand on National	0.2877	1	0.53639

```
28 x6BrandClient       Private Label on Client             0.3785  1  0.61521
29 x6BrandExtension    Private Label on Extension          0.3242  1  0.56938
30 x6BrandRegional     Private Label on Regional           0.3811  1  0.61729
31 x6BrandPrivate      Private Label on Private              .      0   .
32 x6BrandNational     Private Label on National           0.5760  1  0.75893
33 x8BrandClient       National Competitor on Client       0.2777  1  0.52695
34 x8BrandExtension    National Competitor on Extension    0.2806  1  0.52976
35 x8BrandRegional     National Competitor on Regional     0.2904  1  0.53887
36 x8BrandPrivate      National Competitor on Private      0.3066  1  0.55372
37 x8BrandNational     National Competitor on National       .     0   .
                                                                   ==
                                                                   32
```

First we see estimable brand effects for each of the five brands, excluding the constant alternative 'None'. Next we see quantitative alternative-specific price effects for each of the brands. The next two effects that are single *df* effects for the shelf talker and the microwave option. Then we see five sets of cross effects, each consisting of four effects of a brand on another brand, plus one more zero *df* cross effect of a brand on itself. The zero *df* and missing variances and standard errors are correct since the cross effect of an alternative on itself is perfectly aliased with its alternative-specific price effect. These results look fine. Everything that should be estimable is estimable, and everything that should not be estimable is not.

Next, we will run some further checks by looking at the coded design. Before we look at the coded design, recall that the design for the first five choice sets is as follows.

Consumer Food Product Example

Block	Shelf Talker	Client Brand	Client Line Extension	Client Micro/ Stove	Regional Brand	Private Label	Private Micro/ Stove	National Competitor
1	No	N	$1.89	Stove	$1.99	$2.29	Micro	$1.99
		$1.29	N	Micro	N	$1.49	Micro	N
		$1.69	$1.89	Stove	N	$1.49	Stove	$1.99
		$1.29	$1.39	Stove	$1.99	$2.29	Stove	N
		$2.09	$2.39	Stove	$2.49	$2.29	Micro	N

The coded design that the %ChoiceEff macro creates is called TMP_CAND. We will look at the coded data set in several ways. First, here are the Brand, Price, microwave and shelf talker factors, for just the available alternatives for the first five choice sets.

```
proc print data=tmp_cand(obs=23) label;
   var Brand Price Shelf Micro;
   where w;
   run;
```

Consumer Food Product Example

Obs	Brand	Price	Shelf	Micro
2	Extension	$1.89	No	Stove
3	Regional	$1.99	No	Stove
4	Private	$2.29	No	Micro
5	National	$1.99	No	Stove
6	None	$0.00	No	Stove
7	Client	$1.29	No	Stove
10	Private	$1.49	No	Micro
12	None	$0.00	No	Stove
13	Client	$1.69	No	Stove
14	Extension	$1.89	No	Stove
16	Private	$1.49	No	Stove
17	National	$1.99	No	Stove
18	None	$0.00	No	Stove
19	Client	$1.29	No	Stove
20	Extension	$1.39	No	Stove
21	Regional	$1.99	No	Stove
22	Private	$2.29	No	Stove
24	None	$0.00	No	Stove
25	Client	$2.09	No	Stove
26	Extension	$2.39	No	Stove
27	Regional	$2.49	No	Stove
28	Private	$2.29	No	Micro
30	None	$0.00	No	Stove

Unlike all previous examples, the number of alternatives is not the same in all of the choice sets. The first choice set consists of five alternatives including `'None'`. The client brand is not available in this choice set. The second choice set consists of three alternatives including `'None'`. The extension, regional, and national competitors are not available in this choice set. The third choice set consists of five alternatives including `'None'`, and so on.

Here are the coded factors for the brand effects and alternative-specific price effects for the first choice set.

```
proc print data=tmp_cand(obs=5) label;
   id Brand;
   var BrandClient -- BrandNational;
   where w;
   run;
```

```
proc format; value zer 0 = '  0'; run;

proc print data=tmp_cand(obs=5) label;
   id Brand Price;
   var BrandClientPrice -- BrandNationalPrice;
   format BrandClientPrice -- BrandNationalPrice zer5.2;
   where w;
   run;
```

Consumer Food Product Example

Brand	Client	Extension	Regional	Private	National
Extension	0	1	0	0	0
Regional	0	0	1	0	0
Private	0	0	0	1	0
National	0	0	0	0	1
None	0	0	0	0	0

Consumer Food Product Example

Brand	Price	Client Price	Extension Price	Regional Price	Private Price	National Price
Extension	$1.89	0	1.89	0	0	0
Regional	$1.99	0	0	1.99	0	0
Private	$2.29	0	0	0	2.29	0
National	$1.99	0	0	0	0	1.99
None	$0.00	0	0	0	0	0

The brand effects and alternative-specific price effect codings are similar to those we have used previously. The difference is the presence of all zero columns for unavailable alternatives, in this case the client brand. Note that `Brand Price` are just ID variables and do not enter into the analysis.

Here are the shelf talker and microwave coded factors (along with the `Brand`, `Price`, `Shelf`, and `Micro` factors).

```
proc print data=tmp_cand(obs=5) label;
   id Brand Price Shelf Micro;
   var shelftalker micromicro;
   where w;
   run;
```

Consumer Food Product Example

Brand	Price	Shelf	Micro	Shelf Talker	Micro
Extension	$1.89	No	Stove	0	0
Regional	$1.99	No	Stove	0	0
Private	$2.29	No	Micro	0	1
National	$1.99	No	Stove	0	0
None	$0.00	No	Stove	0	0

The following code prints the cross effects along with `Brand` and `Price` for the first choice set.

```
proc print data=tmp_cand(obs=5) label;
   id Brand Price;
   var x1Brand:; format x1Brand: zer5.2;
   where w;
   run;
proc print data=tmp_cand(obs=5) label;
   id Brand Price;
   var x2Brand:; format x2Brand: zer5.2;
   where w;
   run;
proc print data=tmp_cand(obs=5) label;
   id Brand Price;
   var x5Brand:; format x5Brand: zer5.2;
   where w;
   run;
proc print data=tmp_cand(obs=5) label;
   id Brand Price;
   var x6Brand:; format x6Brand: zer5.2;
   where w;
   run;
proc print data=tmp_cand(obs=5) label;
   id Brand Price;
   var x8Brand:; format x8Brand: zer5.2;
   where w;
   run;
```

The cross effects are printed in panels. This first panel shows the terms that capture the effect of the client brand, which is unavailable, on the utility of the other brands. An unavailable brand has no effect on any other brand's utility in that choice set. The second panel shows the terms that capture the effect of the line extension being available at $1.89 on the other available alternatives. Subsequent panels shows the terms that capture the effect of each available alternative on the other available alternatives.

Consumer Food Product Example

Brand	Price	Client Brand on Client	Client Brand on Extension	Client Brand on Regional	Client Brand on Private	Client Brand on National
Extension	$1.89	0	0	0	0	0
Regional	$1.99	0	0	0	0	0
Private	$2.29	0	0	0	0	0
National	$1.99	0	0	0	0	0
None	$0.00	0	0	0	0	0

Consumer Food Product Example

Brand	Price	Client Line Extension on Client	Client Line Extension on Extension	Client Line Extension on Regional	Client Line Extension on Private	Client Line Extension on National
Extension	$1.89	0	1.89	0	0	0
Regional	$1.99	0	0	1.89	0	0
Private	$2.29	0	0	0	1.89	0
National	$1.99	0	0	0	0	1.89
None	$0.00	0	0	0	0	0

Consumer Food Product Example

Brand	Price	Regional Brand on Client	Regional Brand on Extension	Regional Brand on Regional	Regional Brand on Private	Regional Brand on National
Extension	$1.89	0	1.99	0	0	0
Regional	$1.99	0	0	1.99	0	0
Private	$2.29	0	0	0	1.99	0
National	$1.99	0	0	0	0	1.99
None	$0.00	0	0	0	0	0

Consumer Food Product Example

Brand	Price	Private Label on Client	Private Label on Extension	Private Label on Regional	Private Label on Private	Private Label on National
Extension	$1.89	0	2.29	0	0	0
Regional	$1.99	0	0	2.29	0	0
Private	$2.29	0	0	0	2.29	0
National	$1.99	0	0	0	0	2.29
None	$0.00	0	0	0	0	0

Consumer Food Product Example

Brand	Price	National Competitor on Client	National Competitor on Extension	National Competitor on Regional	National Competitor on Private	National Competitor on National
Extension	$1.89	0	1.99	0	0	0
Regional	$1.99	0	0	1.99	0	0
Private	$2.29	0	0	0	1.99	0
National	$1.99	0	0	0	0	1.99
None	$0.00	0	0	0	0	0

A column like 'Private Label on Extension' in the second last panel, for example captures the effect of the private label brand being available at $2.29 on the utility of the extension. In the previous panel, 'Regional Brand on Extension' captures the effect of the regional brand being available at $1.99 on the utility of the extension.

The design looks good, it has reasonably good balance and correlations, it can be used to estimate all of the effects of interest, and we have shown we know how to specify the model to get all the right codings. We are ready to collect data.

Generating Artificial Data

We will not illustrate questionnaire generation for this example since we have done it several times before in previous examples. Instead we will go straight to data processing and analysis. This DATA step generates some artificial data. Creating artificial data and trying the analysis before collecting data is another way to test the design before going to the expense of data collection.

```
%let m   = 6;
%let mm1 = %eval(&m - 1);
%let n   = 36;
```

```
proc format;
   value yn    1 = 'No'    2 = 'Talker';
   value micro 1 = 'Micro' 2 = 'Stove';
   run;

data _null_;
   array brands[&m] _temporary_ (5 7 1 2 3 -2);
   array u[&m];
   array x[&mm1] x1 x2 x5 x6 x8;

   do rep = 1 to 300;
      if mod(rep, 2) then put;
      put rep 3. +2 @@;
      do j = 1 to &n;
         set sasuser.finchdes point=j;
         do brand = 1 to &m; u[brand] = brands[brand] + 2 * normal(17); end;
         do brand = 1 to &mm1;
            if n(x[brand]) then u[brand] + -x[brand]; else u[brand] = .;
            end;
         if n(u2) and x4 = 2 then u2 + 1; /* shelf-talker */
         if n(u2) and x3 = 1 then u2 + 1; /* microwave    */
         if n(u4) and x7 = 1 then u4 + 1; /* microwave    */
         * Choose the most preferred alternative.;
         m = max(of u1-u&m);

         do brand = 1 to &m;
            if n(u[brand]) then if abs(u[brand] - m) < 1e-4 then c = brand;
            end;
         put +(-1) c @@;
         end;
      end;
   stop;
   run;
```

This DATA step reads the data.

```
data results;
   input Subj (choose1-choose&n) (1.) @@;
   datalines;
  1 41222225222221255222512422222221212 2   2 21222222222221214222111322222121242 4
  3 21122222222124115222112622112211252 2   4 21211222222221215222512322122221212 1
  5 21222522222224216222141422222221242 2   6 24212222212222121522241242222112124 22
  .
  .
  .
297 21224222222221245225412422222223212 2 298 21222222222255245222152422222121452 2
299 21222122222221244122512621221421262 2 300 21224212222221114222141422212221232 2
;
```

Processing the Data

The analysis proceeds in a fashion similar to before. We have already made the choice design, so we just have to merge it with the data. The data and design are merged in the usual way using the `%MktMerge` macro. Notice at this point that the unavailable alternatives are still in the design. The `%MktMerge` macro has an `nalts=` alternative and expects a constant number of alternatives in each choice set.

```
%mktmerge(design=sasuser.choicedes, data=results, out=res2,
          nsets=&n, nalts=&m, setvars=choose1-choose&n)

proc print data=res2(obs=12); id subj set; by subj set; run;
```

Here are the data and design for the first two choice sets for the first subject, including the unavailable alternatives.

Consumer Food Product Example

Subj	Set	Brand	Price	Micro	Shelf	x1	x2	x5	x6	x8	w	c
1	1	Client	$0.00	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	0	2
		Extension	$1.89	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1	2
		Regional	$1.99	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1	2
		Private	$2.29	Micro	No	$0.00	$1.89	$1.99	$2.29	$1.99	1	1
		National	$1.99	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1	2
		None	$0.00	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1	2
1	2	Client	$1.29	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	1	1
		Extension	$0.00	Micro	No	$1.29	$0.00	$0.00	$1.49	$0.00	0	2
		Regional	$0.00	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	0	2
		Private	$1.49	Micro	No	$1.29	$0.00	$0.00	$1.49	$0.00	1	2
		National	$0.00	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	0	2
		None	$0.00	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	1	2

These next steps aggregate the data. The data set is fairly large at 64,800 observations, and aggregating greatly reduces its size, which makes both the TRANSREG and the PHREG steps run in just a few seconds. This step also excludes the unavailable alternatives. When `w` is 1 (true) the alternative is available and counted, otherwise when `w` is 0 (false) the alternative is unavailable and excluded by the `where` clause and not counted. There is nothing in subsequent steps that assumes a fixed number of alternatives.

```
proc summary data=res2 nway;
   class set brand price shelf micro x1 x2 x5 x6 x8 c;
   output out=agg(drop=_type_);
   where w; /* exclude unavailable, w = 0 */
   run;

proc print; where set = 1; run;
```

All of the variables used in the analysis are named as `class` variables in PROC SUMMARY, which reduces the data set from 64,800 observations to 286. Here are the aggregated data for the first choice set.

Consumer Food Product Example

Obs	Set	Brand	Price	Shelf	Micro	x1	x2	x5	x6	x8	c	_FREQ_
1	1	Extension	$1.89	No	Stove	$0.00	$1.89	$1.99	$2.29	$1.99	1	257
2	1	Extension	$1.89	No	Stove	$0.00	$1.89	$1.99	$2.29	$1.99	2	43
3	1	National	$1.99	No	Stove	$0.00	$1.89	$1.99	$2.29	$1.99	1	20
4	1	National	$1.99	No	Stove	$0.00	$1.89	$1.99	$2.29	$1.99	2	280
5	1	None	$0.00	No	Stove	$0.00	$1.89	$1.99	$2.29	$1.99	1	1
6	1	None	$0.00	No	Stove	$0.00	$1.89	$1.99	$2.29	$1.99	2	299
7	1	Private	$2.29	No	Micro	$0.00	$1.89	$1.99	$2.29	$1.99	1	22
8	1	Private	$2.29	No	Micro	$0.00	$1.89	$1.99	$2.29	$1.99	2	278
9	1	Regional	$1.99	No	Stove	$0.00	$1.89	$1.99	$2.29	$1.99	2	300

In the first choice set, the extension was chosen (c = 1) a total of _freq_ = 257 times and not chosen (c = 2) a total of _freq_ = 43 times. Each alternative was chosen and not chosen a total of 300 times, which is the number of subjects. These next steps code and run the analysis.

Cross Effects

This next step codes the design for analysis. This coding was discussed on page 260. PROC TRANSREG is run like before, except now the data set AGG is specified and the ID variable includes _freq_ (the frequency variable) but not Subj (the subject number variable).

```
proc transreg data=agg design=5000 nozeroconstant norestoremissing;
   model class(brand / zero='None')
         class(brand / zero='None' separators='' ' ') * identity(price)
         class(shelf micro / lprefix=5 0 zero='No' 'Stove')
         identity(x1 x2 x5 x6 x8) *
            class(brand / zero='None' separators=' ' ' on ') /
         lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   id set c _freq_;
   label x1 = 'CE, Client'
         x2 = 'CE, Extension'
         x5 = 'CE, Regional'
         x6 = 'CE, Private'
         x8 = 'CE, National'
         shelf = 'Shelf Talker'
         micro = 'Microwave';
   run;
```

Note that like we saw in the `%ChoicEff` macro, PROC TRANSREG produces the following warning twice.

```
WARNING: This usage of * sets one group's slope to zero.  Specify |
         to allow all slopes and intercepts to vary.  Alternatively,
         specify CLASS(vars) * identity(vars) identity(vars) for
         separate within group functions and a common intercept.
         This is a change from Version 6.
```

This is because on two occasions `class` was interacted with `identity` using the asterisk instead of the vertical bar. In a linear model, this may be a sign of a coding error, so the procedure prints a warning. If you get this warning while coding a choice model specifying `zero=`*'constant-alternative-level'*, you can safely ignore it.

The analysis is the same as we have done previously with aggregate data. PROC PHREG is run to fit the mother logit model, complete with availability cross effects.

```
proc phreg data=coded;
   strata set;
   model c*c(2) = &_trgind / ties=breslow;
   freq _freq_;
   run;
```

Multinomial Logit Model Results

These steps produced the following results. (Recall that we used `%phchoice(on)` on page 95 to customize the output from PROC PHREG.)

Consumer Food Product Example

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Frequency Variable	_FREQ_
Ties Handling	BRESLOW

Number of Observations Read	286
Number of Observations Used	286
Sum of Frequencies Read	48000
Sum of Frequencies Used	48000

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Stratum	Set	Number of Alternatives	Chosen Alternatives	Not Chosen
1	1	1500	300	1200
2	2	900	300	600
3	3	1500	300	1200
4	4	1500	300	1200
5	5	1500	300	1200
6	6	1500	300	1200
7	7	1500	300	1200
8	8	1500	300	1200
9	9	900	300	600
10	10	1500	300	1200
11	11	1200	300	900
12	12	1500	300	1200
13	13	1200	300	900
14	14	1200	300	900
15	15	1500	300	1200
16	16	1500	300	1200
17	17	900	300	600
18	18	1500	300	1200
19	19	900	300	600
20	20	1500	300	1200
21	21	1500	300	1200
22	22	1500	300	1200
23	23	1500	300	1200
24	24	900	300	600
25	25	900	300	600
26	26	1500	300	1200
27	27	1500	300	1200
28	28	1500	300	1200
29	29	1200	300	900
30	30	1500	300	1200
31	31	1500	300	1200
32	32	900	300	600
33	33	1500	300	1200
34	34	1500	300	1200
35	35	900	300	600
36	36	1500	300	1200
Total		48000	10800	37200

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	154978.05	134843.82
AIC	154978.05	134907.82
SBC	154978.05	135141.01

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	20134.2372	32	<.0001
Score	22265.5761	32	<.0001
Wald	6639.0479	32	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Client	1	3.79393	0.49306	59.2081	<.0001
Extension	1	5.12295	0.52452	95.3934	<.0001
National	1	3.23179	0.81642	15.6697	<.0001
Private	1	1.31050	0.54937	5.6905	0.0171
Regional	1	-0.10024	1.58232	0.0040	0.9495
Client Price	1	-0.30395	0.20501	2.1982	0.1382
Extension Price	1	0.23579	0.28797	0.6705	0.4129
National Price	1	-1.37790	0.34576	15.8812	<.0001
Private Price	1	0.02406	0.28896	0.0069	0.9336
Regional Price	1	-0.17397	0.59454	0.0856	0.7698
Shelf Talker	1	0.71515	0.06909	107.1472	<.0001
Micro	1	0.76468	0.06150	154.5960	<.0001
CE, Client on Client	0	0	.	.	.
CE, Client on Extension	1	0.46604	0.18222	6.5414	0.0105
CE, Client on National	1	0.36788	0.17994	4.1800	0.0409
CE, Client on Private	1	0.38018	0.16846	5.0931	0.0240
CE, Client on Regional	1	0.46363	0.22151	4.3808	0.0363
CE, Extension on Client	1	1.19048	0.27575	18.6385	<.0001
CE, Extension on Extension	0	0	.	.	.
CE, Extension on National	1	1.06140	0.28152	14.2152	0.0002
CE, Extension on Private	1	0.95278	0.27909	11.6544	0.0006
CE, Extension on Regional	1	0.68052	0.32046	4.5097	0.0337
CE, Regional on Client	1	0.08894	0.12876	0.4771	0.4897
CE, Regional on Extension	1	0.15695	0.13012	1.4548	0.2278
CE, Regional on National	1	0.10450	0.13716	0.5805	0.4461
CE, Regional on Private	1	0.06719	0.12210	0.3028	0.5821
CE, Regional on Regional	0	0	.	.	.

CE, Private on Client	1	0.45727	0.24562	3.4660	0.0626
CE, Private on Extension	1	0.42913	0.24719	3.0138	0.0826
CE, Private on National	1	0.55401	0.26339	4.4242	0.0354
CE, Private on Private	0	0	.	.	.
CE, Private on Regional	1	0.45098	0.29187	2.3874	0.1223
CE, National on Client	1	-0.20225	0.17594	1.3216	0.2503
CE, National on Extension	1	-0.14792	0.18001	0.6753	0.4112
CE, National on National	0	0	.	.	.
CE, National on Private	1	-0.15185	0.17687	0.7371	0.3906
CE, National on Regional	1	-0.27439	0.22642	1.4687	0.2256

Since the number of alternatives is not constant within each choice set, the summary table has non-constant numbers of alternatives and numbers of alternatives not chosen. The number chosen, 300 (or one per subject per choice set), is constant, since each subject always chooses one alternative from each choice set regardless of the number of alternatives. The first choice set has 1500 alternatives, 5 available times 300 subjects; whereas the second choice set has 900 alternatives, 3 available times 300 subjects.

The most to least preferred brands are: client line extension, client brand, national brand, private label, the none alternative (with an implicit part-worth utility of zero), and regional competitor. The price effects are mostly negative, and the positive effects are nonsignificant. Both the shelf-talker and the microwaveable option have positive utility. The cross effects are mostly nonsignificant. The most significant cross effect is the effect of the extension on the original client brand.

Modeling Subject Attributes

This example uses the same design and data as we just saw, but this time we have some demographic information about our respondents that we wish to model. The following DATA step reads a subject number, the choices, and respondent age and income (in thousands of dollars).

```
data results;
   input Subj (choose1-choose&n) (1.) age income;
   datalines;
  1 41222225222221255222512422222221212 2  33  44
  2 21222222222221214222111322222121242 4  52  82
  3 21122222222124115222112622112211252 2  51 136
  4 21211222222221215222512322122221212 1  60 108
  .
  .
  .
299 21222122222221244122512621221421262 2  48  49
300 21224212222221114222141422212221232 2  38  51
;
```

Merging the data and design is no different from what we saw previously.

```
%mktmerge(design=sasuser.choicedes, data=results, out=res2,
          nsets=&n, nalts=&m, setvars=choose1-choose&n)

proc print data=res2;
   by subj set; id subj set;
   where (subj = 1 and set = 1) or
         (subj = 2 and set = 2) or
         (subj = 3 and set = 3) or
         (subj = 300 and set = 36);
   run;
```

Here is a small sample of the data. Note that like before, the unavailable alternatives are required for the merge step.

Consumer Food Product Example

Subj	Set	Age	Income	Brand	Price	Micro	Shelf	x1	x2	x5	x6	x8	w	c
1	1	33	44	Client	$0.00	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	0	2
		33	44	Extension	$1.89	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1	2
		33	44	Regional	$1.99	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1	2
		33	44	Private	$2.29	Micro	No	$0.00	$1.89	$1.99	$2.29	$1.99	1	1
		33	44	National	$1.99	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1	2
		33	44	None	$0.00	Stove	No	$0.00	$1.89	$1.99	$2.29	$1.99	1	2
2	2	52	82	Client	$1.29	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	1	1
		52	82	Extension	$0.00	Micro	No	$1.29	$0.00	$0.00	$1.49	$0.00	0	2
		52	82	Regional	$0.00	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	0	2
		52	82	Private	$1.49	Micro	No	$1.29	$0.00	$0.00	$1.49	$0.00	1	2
		52	82	National	$0.00	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	0	2
		52	82	None	$0.00	Stove	No	$1.29	$0.00	$0.00	$1.49	$0.00	1	2
3	3	51	136	Client	$1.69	Stove	No	$1.69	$1.89	$0.00	$1.49	$1.99	1	1
		51	136	Extension	$1.89	Stove	No	$1.69	$1.89	$0.00	$1.49	$1.99	1	2
		51	136	Regional	$0.00	Stove	No	$1.69	$1.89	$0.00	$1.49	$1.99	0	2
		51	136	Private	$1.49	Stove	No	$1.69	$1.89	$0.00	$1.49	$1.99	1	2
		51	136	National	$1.99	Stove	No	$1.69	$1.89	$0.00	$1.49	$1.99	1	2
		51	136	None	$0.00	Stove	No	$1.69	$1.89	$0.00	$1.49	$1.99	1	2
300	36	38	51	Client	$1.69	Stove	No	$1.69	$2.39	$1.99	$1.49	$0.00	1	2
		38	51	Extension	$2.39	Stove	Talker	$1.69	$2.39	$1.99	$1.49	$0.00	1	1
		38	51	Regional	$1.99	Stove	No	$1.69	$2.39	$1.99	$1.49	$0.00	1	2
		38	51	Private	$1.49	Micro	No	$1.69	$2.39	$1.99	$1.49	$0.00	1	2
		38	51	National	$0.00	Stove	No	$1.69	$2.39	$1.99	$1.49	$0.00	0	2
		38	51	None	$0.00	Stove	No	$1.69	$2.39	$1.99	$1.49	$0.00	1	2

You can see that the demographic information matches the raw data and is constant within each subject. The rest of the data processing is virtually the same as well. Since we have demographic information, we will not aggregate. There would have to be ties in both the demographics and choice for aggregation to have any effect.

We use PROC TRANSREG to code, adding `Age` and `Income` to the analysis.

```
proc transreg data=res2 design=5000 nozeroconstant norestoremissing;
   model class(brand / zero='None')
         identity(age income) * class(brand / zero='None' separators='' ', ')
         class(brand / zero='None' separators='' ' ') * identity(price)
         class(shelf micro / lprefix=5 0 zero='No' 'Stove')
         identity(x1 x2 x5 x6 x8) *
            class(brand / zero='None' separators=' ' ' on ') /
         lprefix=0 order=data;
   output out=code(drop=_type_ _name_ intercept);
   id subj set c w;
   label x1 = 'CE, Client'
         x2 = 'CE, Extension'
         x5 = 'CE, Regional'
         x6 = 'CE, Private'
         x8 = 'CE, National'
         shelf = 'Shelf Talker'
         micro = 'Microwave';
   run;

data coded(drop=w); set code; where w; run; /* exclude unavailable */
```

The `Age` and `Income` variables are incorporated into the analysis by interacting them with `Brand`. Demographic variables must be interacted with attributes to have any effect. If `identity(age income)` had been specified instead of `identity(age income) * class(brand / ...)` the coefficients for age and income would be zero. This is because age and income are constant within each choice set and subject combination, which means they are constant within each stratum. The second separator `', '` is used to create names for the brand/demographic interaction terms like `'Age, Client'`.

These next steps print the first coded choice set.

```
proc print data=coded(obs=5) label;
   id brand price;
   var BrandClient -- BrandPrivate Shelf Micro c;
   run;
proc print data=coded(obs=5 drop=Age) label;
   id brand price;
   var Age:;
   run;
proc print data=coded(obs=5 drop=Income) label;
   id brand price;
   var Income:;
   run;
```

```
proc print data=coded(obs=5) label;
   id brand price;
   var BrandClientPrice -- BrandPrivatePrice;
   format BrandClientPrice -- BrandPrivatePrice best4.;
   run;
proc print data=coded(obs=5 drop=x1) label;
   id brand price; var x1:; format x1: best4.;
   run;
proc print data=coded(obs=5 drop=x2) label;
   id brand price; var x2:; format x2: best4.;
   run;
proc print data=coded(obs=5 drop=x5) label;
   id brand price; var x5:; format x5: best4.;
   run;
proc print data=coded(obs=5 drop=x6) label;
   id brand price; var x6:; format x6: best4.;
   run;
proc print data=coded(obs=5 drop=x8) label;
   id brand price; var x8:; format x8: best4.;
   run;
```

Here is the coded data set for the first subject and choice set. The part that is new is the second and third panel, which will be used to capture the brand by age and brand by income effects.

Here are the attributes and the brand effects.

Consumer Food Product Example

Brand	Price	Client	Extension	Regional	Private	Shelf Talker	Microwave	c
Extension	$1.89	0	1	0	0	No	Stove	2
Regional	$1.99	0	0	1	0	No	Stove	2
Private	$2.29	0	0	0	1	No	Micro	1
National	$1.99	0	0	0	0	No	Stove	2
None	$0.00	0	0	0	0	No	Stove	2

Here are the age by brand effects.

Consumer Food Product Example

Brand	Price	Age, Client	Age, Extension	Age, Regional	Age, Private	Age, National
Extension	$1.89	0	33	0	0	0
Regional	$1.99	0	0	33	0	0
Private	$2.29	0	0	0	33	0
National	$1.99	0	0	0	0	33
None	$0.00	0	0	0	0	0

Here are the income by brand effects.

Consumer Food Product Example

Brand	Price	Income, Client	Income, Extension	Income, Regional	Income, Private	Income, National
Extension	$1.89	0	44	0	0	0
Regional	$1.99	0	0	44	0	0
Private	$2.29	0	0	0	44	0
National	$1.99	0	0	0	0	44
None	$0.00	0	0	0	0	0

Here are the alternative-specific price effects.

Consumer Food Product Example

Brand	Price	Client Price	Extension Price	Regional Price	Private Price
Extension	$1.89	0	1.89	0	0
Regional	$1.99	0	0	1.99	0
Private	$2.29	0	0	0	2.29
National	$1.99	0	0	0	0
None	$0.00	0	0	0	0

Here are the client cross effects.

Consumer Food Product Example

Brand	Price	CE, Client on Client	CE, Client on Extension	CE, Client on Regional	CE, Client on Private	CE, Client on National
Extension	$1.89	0	0	0	0	0
Regional	$1.99	0	0	0	0	0
Private	$2.29	0	0	0	0	0
National	$1.99	0	0	0	0	0
None	$0.00	0	0	0	0	0

Here are the extension cross effects.

Consumer Food Product Example

Brand	Price	CE, Extension on Client	CE, Extension on Extension	CE, Extension on Regional	CE, Extension on Private	CE, Extension on National
Extension	$1.89	0	1.89	0	0	0
Regional	$1.99	0	0	1.89	0	0
Private	$2.29	0	0	0	1.89	0
National	$1.99	0	0	0	0	1.89
None	$0.00	0	0	0	0	0

Here are the regional competitor cross effects.

Consumer Food Product Example

Brand	Price	CE, Regional on Client	CE, Regional on Extension	CE, Regional on Regional	CE, Regional on Private	CE, Regional on National
Extension	$1.89	0	1.99	0	0	0
Regional	$1.99	0	0	1.99	0	0
Private	$2.29	0	0	0	1.99	0
National	$1.99	0	0	0	0	1.99
None	$0.00	0	0	0	0	0

Here are the private label cross effects.

Consumer Food Product Example

Brand	Price	CE, Private on Client	CE, Private on Extension	CE, Private on Regional	CE, Private on Private	CE, Private on National
Extension	$1.89	0	2.29	0	0	0
Regional	$1.99	0	0	2.29	0	0
Private	$2.29	0	0	0	2.29	0
National	$1.99	0	0	0	0	2.29
None	$0.00	0	0	0	0	0

Here are the national competitor cross effects.

Consumer Food Product Example

Brand	Price	CE, National on Client	CE, National on Extension	CE, National on Regional	CE, National on Private	CE, National on National
Extension	$1.89	0	1.99	0	0	0
Regional	$1.99	0	0	1.99	0	0
Private	$2.29	0	0	0	1.99	0
National	$1.99	0	0	0	0	1.99
None	$0.00	0	0	0	0	0

The PROC PHREG specification is the same as we have used before with nonaggregated data.

```
proc phreg data=coded brief;
   model c*c(2) = &_trgind / ties=breslow;
   strata subj set;
   run;
```

This step took just about one minute and produced the following results.

Consumer Food Product Example

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW

Number of Observations Read	48000
Number of Observations Used	48000

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	2400	3	1	2
2	1200	4	1	3
3	7200	5	1	4

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	31776.351	11624.646
AIC	31776.351	11708.646
SBC	31776.351	12014.713

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	20151.7051	42	<.0001
Score	22272.9627	42	<.0001
Wald	6625.8828	42	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Client	1	3.47166	0.59909	33.5805	<.0001
Extension	1	4.98005	0.63031	62.4251	<.0001
Regional	1	-0.68856	1.64851	0.1745	0.6762
Private	1	1.25265	0.64553	3.7655	0.0523
National	1	2.62920	0.89165	8.6947	0.0032
Age, Client	1	0.00833	0.01304	0.4086	0.5227
Age, Extension	1	0.00128	0.01333	0.0092	0.9237
Age, Regional	1	0.01362	0.01690	0.6495	0.4203
Age, Private	1	0.00318	0.01298	0.0601	0.8063
Age, National	1	0.02564	0.01366	3.5224	0.0605
Income, Client	1	-0.0003997	0.00518	0.0060	0.9385
Income, Extension	1	0.00150	0.00530	0.0797	0.7777
Income, Regional	1	0.0000957	0.00672	0.0002	0.9886
Income, Private	1	-0.00112	0.00515	0.0472	0.8281
Income, National	1	-0.00723	0.00545	1.7623	0.1843
Client Price	1	-0.30497	0.20503	2.2124	0.1369
Extension Price	1	0.23198	0.28780	0.6497	0.4202
Regional Price	1	-0.17593	0.59465	0.0875	0.7673
Private Price	1	0.02219	0.28887	0.0059	0.9388
National Price	1	-1.38361	0.34604	15.9877	<.0001
Shelf Talker	1	0.71563	0.06911	107.2177	<.0001
Micro	1	0.76531	0.06152	154.7423	<.0001
CE, Client on Client	0	0	.	.	.
CE, Client on Extension	1	0.46507	0.18224	6.5123	0.0107
CE, Client on Regional	1	0.46621	0.22160	4.4260	0.0354
CE, Client on Private	1	0.37948	0.16846	5.0744	0.0243
CE, Client on National	1	0.36584	0.18002	4.1300	0.0421
CE, Extension on Client	1	1.18750	0.27558	18.5687	<.0001
CE, Extension on Extension	0	0	.	.	.
CE, Extension on Regional	1	0.67797	0.32030	4.4803	0.0343
CE, Extension on Private	1	0.95072	0.27892	11.6182	0.0007
CE, Extension on National	1	1.05294	0.28142	13.9988	0.0002
CE, Regional on Client	1	0.08818	0.12882	0.4685	0.4937
CE, Regional on Extension	1	0.15624	0.13018	1.4404	0.2301
CE, Regional on Regional	0	0	.	.	.
CE, Regional on Private	1	0.06658	0.12215	0.2971	0.5857
CE, Regional on National	1	0.10271	0.13723	0.5602	0.4542
CE, Private on Client	1	0.45631	0.24549	3.4550	0.0631
CE, Private on Extension	1	0.42821	0.24707	3.0039	0.0831
CE, Private on Regional	1	0.44903	0.29180	2.3680	0.1238
CE, Private on Private	0	0	.	.	.
CE, Private on National	1	0.55255	0.26330	4.4039	0.0359

CE, National on Client	1	-0.20108	0.17594	1.3062	0.2531
CE, National on Extension	1	-0.14686	0.18001	0.6656	0.4146
CE, National on Regional	1	-0.27254	0.22645	1.4485	0.2288
CE, National on Private	1	-0.15075	0.17688	0.7264	0.3941
CE, National on National	0	0	.	.	.

In previous examples, when we used the `brief` option to produce a brief summary of the strata, the table had only one line. In this case, since our choice sets have 3, 4, or 5 alternatives, we have three rows, one for each choice set size. The coefficients for the age and income variables are generally not very significant in this analysis.

Allocation of Prescription Drugs

This example discusses an allocation study, which is a technique often used in the area of prescription drug marketing research. This example discusses designing the allocation experiment, processing the data, analyzing frequencies, analyzing proportions, coding, analysis, and results. The principles of designing an allocation study are the same as for designing a first-choice experiment, as is the coding and final analysis. However, processing the data before analysis is different.

The previous examples have all modeled simple choice. However, sometimes the response of interest is not simple first choice. For example, in prescription drug marketing, researchers often use allocation studies where multiple, not single choices are made. Physicians are asked questions like "For the next ten prescriptions you write for a particular condition, how many would you write for each of these drugs?" The response, for example, could be "5 for drug 1, none for drug 2, 3 for drug 3, and 2 for drug 4."

Designing the Allocation Experiment

In this study, physicians were asked to specify which of ten drugs they would prescribe to their next ten patients. In this study, ten drugs, Drug 1 − Drug 10, were available each at three different prices, $50, $75, and $100. In real studies, real brand names would be used and there would probably be more attributes. Since experimental design has been covered in some detail in other examples, we chose a simple design for this experiment so that we could concentrate on data processing. First, we use the `%MktRuns` autocall macro to suggest a design size. (All of the autocall macros used in this book are documented starting on page 479.) We specify `3 ** 10` for the 10 three-level factors.

```
title 'Allocation of Prescription Drugs';

%mktruns( 3 ** 10 )
```

```
                Allocation of Prescription Drugs

                      Design Summary

                Number of
                Levels       Frequency

                   3             10
```

```
                Allocation of Prescription Drugs

          Saturated      = 21
          Full Factorial = 59,049

          Some Reasonable                        Cannot Be
             Design Sizes       Violations       Divided By

                      27 *                0
                      36 *                0
                      45 *                0
                      54 *                0
                      21                 45       9
                      24                 45       9
                      30                 45       9
                      33                 45       9
                      39                 45       9
                      42                 45       9

        * - 100% Efficient Design can be made with the MktEx Macro.

                Allocation of Prescription Drugs

 n    Design                                          Reference

27             3 ** 13                              Fractional-Factorial
36    2 ** 11  3 ** 12                              Orthogonal Array
36    2 **  4  3 ** 13                              Orthogonal Array
36    2 **  2  3 ** 12  6 **  1                     Orthogonal Array
36             3 ** 13  4 **  1                     Orthogonal Array
36             3 ** 12 12 **  1                     Orthogonal Array
45             3 ** 10  5 **  1                     Orthogonal Array
54    2 **  1  3 ** 25                              Orthogonal Array
54    2 **  1  3 ** 21  9 **  1                     Orthogonal Array
54             3 ** 24  6 **  1                     Orthogonal Array
54             3 ** 20  6 **  1  9 **  1            Orthogonal Array
54             3 ** 18 18 **  1                     Orthogonal Array
```

We need at least 21 choice sets and we see the optimal sizes are all divisible by nine. We will use 27 choice sets, which can give us up to 13 three-level factors.

Next, we use the `%MktEx` macro to create the design. In addition, one more factor is added to the design. This factor will be used to block the design into three blocks of size 9.

```
%let nalts = 10;

%mktex(3 ** &nalts 3, n=27, seed=396)
```

The macro finds a 100% *D*-efficient design.

Allocation of Prescription Drugs

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	100.0000	100.0000	Tab
1	End	100.0000		

Allocation of Prescription Drugs

The OPTEX Procedure

Class Level Information

Class	Levels	-Values-
x1	3	1 2 3
x2	3	1 2 3
x3	3	1 2 3
x4	3	1 2 3
x5	3	1 2 3
x6	3	1 2 3
x7	3	1 2 3
x8	3	1 2 3
x9	3	1 2 3
x10	3	1 2 3
x11	3	1 2 3

Allocation of Prescription Drugs

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	100.0000	100.0000	100.0000	0.9230

The `%MktEx` macro always creates factor names of `x1`, `x2`, and so on with values of 1, 2, ... You can create a data set with the names and values you want and use it to rename the factors and reset the levels. This first step creates a data set with 11 variables, `Block` and `Brand1 - Brand10`. `Block` has values 1, 2, and 3, and the brand variables have values of 50, 75, and 100 with a `dollar` format. The `%MktLab` macro takes the `data=Randomized` design data set and uses the names, values, and formats

in the `key=Key` data set to make the `out=Final` data set. This data set is sorted by block and printed. The `%MktEval` macro is called to check the results.

```
data key(drop=i);
   input Block Brand1;
   array Brand[10];
   do i = 2 to 10; brand[i] = brand1; end;
   format brand: dollar4.;
   datalines;
1  50
2  75
3 100
;

proc print; run;

%mktlab(key=key);

proc sort out=sasuser.allocdes; by block; run;

proc print; id block; by block; run;

%mkteval(blocks=block)
```

Here is the `key=` data set.

Allocation of Prescription Drugs

Obs	Block	Brand1	Brand2	Brand3	Brand4	Brand5	Brand6	Brand7	Brand8	Brand9	Brand10
1	1	$50	$50	$50	$50	$50	$50	$50	$50	$50	$50
2	2	$75	$75	$75	$75	$75	$75	$75	$75	$75	$75
3	3	$100	$100	$100	$100	$100	$100	$100	$100	$100	$100

The `%MktLab` macro prints the following mapping information.

```
Variable Mapping:
   x1  : Block
   x2  : Brand1
   x3  : Brand2
   x4  : Brand3
   x5  : Brand4
   x6  : Brand5
   x7  : Brand6
   x8  : Brand7
   x9  : Brand8
   x10 : Brand9
   x11 : Brand10
```

Here is the design.

Allocation of Prescription Drugs

Block	Brand1	Brand2	Brand3	Brand4	Brand5	Brand6	Brand7	Brand8	Brand9	Brand10
1	$50	$75	$50	$75	$100	$100	$100	$100	$50	$100
	$100	$50	$100	$75	$75	$75	$100	$50	$100	$75
	$50	$50	$75	$100	$50	$75	$50	$75	$50	$75
	$75	$50	$50	$50	$100	$75	$75	$100	$75	$75
	$75	$75	$100	$100	$75	$100	$50	$50	$75	$100
	$50	$100	$100	$50	$75	$50	$75	$50	$50	$50
	$100	$75	$75	$50	$50	$100	$75	$75	$100	$100
	$100	$100	$50	$100	$100	$50	$50	$100	$100	$50
	$75	$100	$75	$75	$50	$50	$100	$75	$75	$50
2	$100	$75	$50	$100	$75	$50	$100	$75	$75	$75
	$100	$100	$100	$75	$50	$75	$75	$100	$75	$100
	$50	$75	$100	$50	$50	$50	$50	$100	$100	$75
	$75	$50	$100	$100	$50	$100	$100	$100	$50	$50
	$50	$100	$75	$100	$100	$75	$100	$50	$100	$100
	$100	$50	$75	$50	$100	$100	$50	$50	$75	$50
	$50	$50	$50	$75	$75	$100	$75	$75	$100	$50
	$75	$75	$75	$75	$100	$50	$75	$50	$50	$75
	$75	$100	$50	$50	$75	$75	$50	$75	$50	$100
3	$100	$75	$100	$75	$100	$75	$50	$75	$50	$50
	$75	$75	$50	$50	$50	$75	$100	$50	$100	$50
	$50	$75	$75	$100	$75	$75	$75	$100	$75	$50
	$50	$100	$50	$75	$50	$100	$50	$50	$75	$75
	$50	$50	$100	$50	$100	$50	$100	$75	$75	$100
	$75	$50	$75	$75	$75	$50	$50	$100	$100	$100
	$75	$100	$100	$100	$100	$100	$75	$75	$100	$75
	$100	$50	$50	$100	$50	$50	$75	$50	$50	$100
	$100	$100	$75	$50	$75	$100	$100	$100	$50	$75

Here are some of the evaluation results.

Allocation of Prescription Drugs
Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

	Block	Brand1	Brand2	Brand3	Brand4	Brand5	Brand6	Brand7	Brand8	Brand9	Brand10
Block	1	0	0	0	0	0	0	0	0	0	0
Brand1	0	1	0	0	0	0	0	0	0	0	0
Brand2	0	0	1	0	0	0	0	0	0	0	0
Brand3	0	0	0	1	0	0	0	0	0	0	0
Brand4	0	0	0	0	1	0	0	0	0	0	0
Brand5	0	0	0	0	0	1	0	0	0	0	0
Brand6	0	0	0	0	0	0	1	0	0	0	0
Brand7	0	0	0	0	0	0	0	1	0	0	0
Brand8	0	0	0	0	0	0	0	0	1	0	0
Brand9	0	0	0	0	0	0	0	0	0	1	0
Brand10	0	0	0	0	0	0	0	0	0	0	1

Allocation of Prescription Drugs
Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316

	Frequencies
Block	9 9 9
Brand1	9 9 9
Brand2	9 9 9
Brand3	9 9 9
Brand4	9 9 9
Brand5	9 9 9
Brand6	9 9 9
Brand7	9 9 9
Brand8	9 9 9
Brand9	9 9 9
Brand10	9 9 9

```
Block Brand1      3 3 3 3 3 3 3 3 3
Block Brand2      3 3 3 3 3 3 3 3 3
Block Brand3      3 3 3 3 3 3 3 3 3
Block Brand4      3 3 3 3 3 3 3 3 3
Block Brand5      3 3 3 3 3 3 3 3 3
Block Brand6      3 3 3 3 3 3 3 3 3
Block Brand7      3 3 3 3 3 3 3 3 3
Block Brand8      3 3 3 3 3 3 3 3 3
Block Brand9      3 3 3 3 3 3 3 3 3
Block Brand10     3 3 3 3 3 3 3 3 3
.
.
.
N-Way             1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                  1 1 1 1 1 1 1 1
```

Processing the Data

Questionnaires are generated and data collected using a minor modification of the methods discussed in earlier examples. The difference is instead of asking for first choice data, allocation data are collected instead. Each row of the input data set contains a block, subject, and set number, followed by the number of times each of the ten alternatives was chosen. If all of the choice frequencies are zero, then the constant alternative was chosen. The `if` statement is used to check data entry. For convenience, choice set number is recoded to run from 1 to 27 instead of consisting of three blocks of nine sets. This gives us one fewer variable on which to stratify.

```
data results;
   input Block Subject Set @9 (freq1-freq&nalts) (2.);
   if not (sum(of freq:) in (0, &nalts)) then put _all_;
   set = (block - 1) * 9 + set;
   datalines;
1   1 1  0 0 8 0 2 0 0 0 0 0
1   1 2  0 0 8 0 0 0 2 0 0 0
1   1 3  0 0 0 0 0 0 0 010 0
1   1 4  1 0 0 1 3 3 0 0 2 0
1   1 5  2 0 8 0 0 0 0 0 0 0
1   1 6  0 1 3 1 0 0 0 0 1 4
1   1 7  0 1 3 1 1 2 0 0 2 0
1   1 8  0 0 3 0 0 2 1 0 0 4
1   1 9  0 2 5 0 0 0 0 0 3 0
2   2 1  1 1 0 2 0 3 0 1 1 1
2   2 2  1 0 3 1 0 1 1 0 2 1
.
.
.
;
```

In the first step, in creating an analysis data set for an allocation study, we reformat the data from one row per choice set per block per subject ($9 \times 3 \times 100 = 2700$ observations) to one per alternative (including the constant) per choice set per block per subject ($(10+1)\times 9\times 3\times 100 = 29700$ observations). For each choice set, 11 observations are written storing the choice frequency in the variable `Count` and the brand in the variable `Brand`. If no alternative is chosen, then the constant alternative is chosen ten times, otherwise it is chosen zero times.

```
data allocs(keep=block set brand count);
   set results;

   array freq[&nalts];

   * Handle the &nalts alternatives;
   do b = 1 to &nalts;
      Brand = 'Brand ' || put(b, 2.);
      Count = freq[b];
      output;
      end;
   * Constant alt choice is implied if nothing else is chosen.
     brand = ' ' is used to flag the constant alternative.;

   brand = ' ';
   count = 10 * (sum(of freq:) = 0);
   output;
   run;
proc print data=results(obs=3) label noobs; run;
proc print data=allocs(obs=33); id block set; by block set; run;
```

The PROC PRINT steps show how the first three observations of the RESULTS data set are transposed into the first 33 observations of the ALLOCS data set.

Allocation of Prescription Drugs

Block	Subject	Set	Freq1	Freq2	Freq3	Freq4	Freq5	Freq6	Freq7	Freq8	Freq9	Freq10
1	1	1	0	0	8	0	2	0	0	0	0	0
1	1	2	0	0	8	0	0	0	2	0	0	0
1	1	3	0	0	0	0	0	0	0	0	10	0

Allocation of Prescription Drugs

Block	Set	Brand	Count
1	1	Brand 1	0
		Brand 2	0
		Brand 3	8
		Brand 4	0
		Brand 5	2
		Brand 6	0
		Brand 7	0
		Brand 8	0
		Brand 9	0
		Brand 10	0
			0
1	2	Brand 1	0
		Brand 2	0
		Brand 3	8
		Brand 4	0
		Brand 5	0
		Brand 6	0
		Brand 7	2
		Brand 8	0
		Brand 9	0
		Brand 10	0
			0
1	3	Brand 1	0
		Brand 2	0
		Brand 3	0
		Brand 4	0
		Brand 5	0
		Brand 6	0
		Brand 7	0
		Brand 8	0
		Brand 9	10
		Brand 10	0
			0

The next step aggregates the data. It stores in the variable `Count` the number of times each alternative of each choice set was chosen. This creates a data set with 297 observations (3 blocks × 9 sets × 11 alternatives = 297).

```
* Aggregate, store the results back in count.;

proc summary data=allocs nway missing;
   class set brand;
   output sum(count)=Count out=allocs(drop=_type_ _freq_);
   run;
```

These next steps prepare the design for analysis. We need to create a data set KEY that describes how the factors in our design will be used for analysis. It will contain all of the factor names, `Brand1`, `Brand2`, ..., `Brand10`. We can run the `%MktKey` macro to get these names in the SAS log for cutting and pasting into the program without typing them.

```
%mktkey(Brand1-Brand10)
```

The `%MktKey` macro produced the following line.

```
Brand1 Brand2 Brand3 Brand4 Brand5 Brand6 Brand7 Brand8 Brand9 Brand10
```

The next step rolls out the experimental design data set to match the choice allocations data set. The data set is transposed from one row per choice set to one row per alternative per choice set. This data set also has 297 observations. As we saw in many previous examples, the `%MktRoll` macro can be used to process the design.

```
data key(keep=Brand Price);
   input Brand $ 1-8 Price $;
   datalines;
Brand  1    Brand1
Brand  2    Brand2
Brand  3    Brand3
Brand  4    Brand4
Brand  5    Brand5
Brand  6    Brand6
Brand  7    Brand7
Brand  8    Brand8
Brand  9    Brand9
Brand 10    Brand10
.           .
;
%mktroll(design=sasuser.allocdes, key=key, alt=brand, out=rolled)

proc print data=rolled(obs=11); format price dollar4.; run;
```

Allocation of Prescription Drugs

Obs	Set	Brand	Price
1	1	Brand 1	$50
2	1	Brand 2	$75
3	1	Brand 3	$50
4	1	Brand 4	$75
5	1	Brand 5	$100
6	1	Brand 6	$100
7	1	Brand 7	$100
8	1	Brand 8	$100
9	1	Brand 9	$50
10	1	Brand 10	$100
11	1		.

Both data sets must be sorted the same way before they can be merged. The constant alternative, indicated by a missing brand, is last in the design choice set and hence is out of order. Missing must come before nonmissing for the merge. The order is correct in the ALLOCS data set since it was created by PROC SUMMARY with `Brand` as a `class` variable.

```
proc sort data=rolled; by set brand; run;
```

The data are merged along with error checking to ensure that the merge proceeded properly. Both data sets should have the same observations and `Set` and `Brand` variables, so the merge should be one to one.

```
data allocs2;
   merge allocs(in=flag1) rolled(in=flag2);
   by set brand;
   if flag1 ne flag2 then put 'ERROR: Merge is not 1 to 1.';
   format price dollar4.;
   run;

proc print data=allocs2(obs=22);
   var brand price count;
   sum count;
   by notsorted set;
   id set;
   run;
```

In the aggregate and combined data set, we see how often each alternative was chosen for each choice set. For example, in the first choice set, the constant alternative was chosen zero times, Brand 1 at $100 was chosen 103 times, and so on. The 11 alternatives were chosen a total of 1000 times, 100 subjects times 10 choices each.

Allocation of Prescription Drugs

Set	Brand	Price	Count
1		.	0
	Brand 1	$50	103
	Brand 2	$75	58
	Brand 3	$50	318
	Brand 4	$75	99
	Brand 5	$100	54
	Brand 6	$100	83
	Brand 7	$100	71
	Brand 8	$100	58
	Brand 9	$50	100
	Brand 10	$100	56
---			-----
1			1000
2		.	10
	Brand 1	$100	73
	Brand 2	$50	76
	Brand 3	$100	342
	Brand 4	$75	55
	Brand 5	$75	50
	Brand 6	$75	77
	Brand 7	$100	95
	Brand 8	$50	71
	Brand 9	$100	72
	Brand 10	$75	79
---			-----
2			1000

At this point, the data set contains 297 observations (27 choice sets times 11 alternatives) showing the number of times each alternative was chosen. This data set must be augmented to also include the number of times each alternative was not chosen. For example, in the first choice set, brand 1 was chosen 103 times, which means it was not chosen $0+58+318+99+54+83+71+58+100+56 = 897$ times. We use a macro, `%MktAllo` for "marketing allocation study" to process the data. We specify the input `data=allocs2` data set, the output `out=allocs3` data set, the number of alternatives including the constant (`nalts=%eval(&nalts + 1)`), the variables in the data set except the frequency variable (`vars=set brand price`), and the frequency variable (`freq=Count`). The macro counts how many times each alternative was chosen and not chosen and writes the results to the `out=` data set along with the usual `c` $= 1$ for chosen and `c` $= 2$ for unchosen.

```
%mktallo(data=allocs2, out=allocs3, nalts=%eval(&nalts + 1),
         vars=set brand price, freq=Count)

proc print data=allocs3(obs=22);
   var set brand price count c;
   run;
```

The first 22 records of the allocation data set are shown next.

Allocation of Prescription Drugs

Obs	Set	Brand	Price	Count	c
1	1		.	0	1
2	1		.	1000	2
3	1	Brand 1	$50	103	1
4	1	Brand 1	$50	897	2
5	1	Brand 2	$75	58	1
6	1	Brand 2	$75	942	2
7	1	Brand 3	$50	318	1
8	1	Brand 3	$50	682	2
9	1	Brand 4	$75	99	1
10	1	Brand 4	$75	901	2
11	1	Brand 5	$100	54	1
12	1	Brand 5	$100	946	2
13	1	Brand 6	$100	83	1
14	1	Brand 6	$100	917	2
15	1	Brand 7	$100	71	1
16	1	Brand 7	$100	929	2
17	1	Brand 8	$100	58	1
18	1	Brand 8	$100	942	2
19	1	Brand 9	$50	100	1
20	1	Brand 9	$50	900	2
21	1	Brand 10	$100	56	1
22	1	Brand 10	$100	944	2

In the first choice set, the constant alternative is chosen zero times and not chosen 1000 times, Brand 1 is chosen 103 times and not chosen $1000 - 103 = 897$ times, Brand 2 is chosen 58 times and not chosen $1000 - 58 = 942$ times, and so on. Note that allocation studies do not always have fixed sums, so it is important to use the `%MktAllo` macro or some other approach that actually counts the number of times each alternative was not chosen. It is not always sufficient to simply subtract from a fixed constant (in this case 1000).

Coding and Analysis

The next step codes the design for analysis. Indicator variables are created for `Brand` and `Price`. All of the PROC TRANSREG options have been discussed in other examples.

```
proc transreg design data=allocs3 nozeroconstant norestoremissing;
   model class(brand price / zero=none) / lprefix=0;
   output out=coded(drop=_type_ _name_ intercept);
   id set c count;
   run;
```

Analysis proceeds like it has in all other examples. We stratify by choice set number. We do not need to stratify by `Block` since choice set number does not repeat within block.

```
proc phreg data=coded;
   where count > 0;
   model c*c(2) = &_trgind / ties=breslow;
   freq count;
   strata set;
   run;
```

We used the `where` statement to exclude observations with zero frequency; otherwise PROC PHREG complains about them.

Multinomial Logit Model Results

Here are the results. Recall that we used `%phchoice(on)` on page 95 to customize the output from PROC PHREG.

Allocation of Prescription Drugs

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Frequency Variable	Count
Ties Handling	BRESLOW

Number of Observations Read	583
Number of Observations Used	583
Sum of Frequencies Read	297000
Sum of Frequencies Used	297000

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Stratum	Set	Number of Alternatives	Chosen Alternatives	Not Chosen
1	1	11000	1000	10000
2	2	11000	1000	10000
3	3	11000	1000	10000
4	4	11000	1000	10000
5	5	11000	1000	10000
6	6	11000	1000	10000
7	7	11000	1000	10000
8	8	11000	1000	10000
9	9	11000	1000	10000
10	10	11000	1000	10000
11	11	11000	1000	10000
12	12	11000	1000	10000
13	13	11000	1000	10000
14	14	11000	1000	10000
15	15	11000	1000	10000
16	16	11000	1000	10000
17	17	11000	1000	10000
18	18	11000	1000	10000
19	19	11000	1000	10000
20	20	11000	1000	10000
21	21	11000	1000	10000
22	22	11000	1000	10000
23	23	11000	1000	10000
24	24	11000	1000	10000
25	25	11000	1000	10000
26	26	11000	1000	10000
27	27	11000	1000	10000
Total		297000	27000	270000

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	502505.13	489062.66
AIC	502505.13	489086.66
SBC	502505.13	489185.11

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	13442.4676	12	<.0001
Score	18340.8415	12	<.0001
Wald	14087.6778	12	<.0001

Multinomial Logit Parameter Estimates

		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Brand	1	1	2.09906	0.06766	962.5297	<.0001
Brand	2	1	2.09118	0.06769	954.5113	<.0001
Brand	3	1	3.54204	0.06484	2984.4698	<.0001
Brand	4	1	2.09710	0.06766	960.5277	<.0001
Brand	5	1	2.08523	0.06771	948.4791	<.0001
Brand	6	1	2.03530	0.06790	898.6218	<.0001
Brand	7	1	2.06920	0.06777	932.3154	<.0001
Brand	8	1	2.08573	0.06771	948.9824	<.0001
Brand	9	1	2.11705	0.06759	980.9640	<.0001
Brand	10	1	2.06363	0.06779	926.7331	<.0001
$50		1	0.00529	0.01628	0.1058	0.7450
$75		1	0.0005304	0.01629	0.0011	0.9740
$100		0	0	.	.	.

The output shows that there are 27 strata, one per choice set, each consisting of 1000 chosen alternatives (10 choices by 100 subjects) and 10,000 unchosen alternatives. All of the brand coefficients are "significant," with the Brand 3 effect being by far the strongest. (We will soon see that statistical significance should be ignored with allocation studies.) There is no price effect.

Analyzing Proportions

Recall that we collected data by asking physicians to report which brands they would prescribe the next ten times they write prescriptions. Alternatively, we could ask them to report the *proportion* of time they would prescribe each brand. We can simulate having proportion data by dividing our count data by 10. This means our frequency variable will no longer contain integers, so we need to specify the `notruncate` option on PROC PHREG `freq` statement to allow noninteger "frequencies."

```
data coded2;
   set coded;
   count = count / 10;
   run;
```

```
proc phreg data=coded2;
   where count > 0;
   model c*c(2) = &_trgind / ties=breslow;
   freq count / notruncate;
   strata set;
   run;
```

When we do this, we see the number of alternatives and the number chosen and not chosen decrease by a factor of 10 as do all of the Chi-Square tests. The coefficients are unchanged. This implies that market share calculations are invariant to the different scalings of the frequencies. However, the p-values are not invariant. The sample size is artificially inflated when counts are used so p-values are not interpretable in an allocation study. When proportions are used, each subject is contributing 1 to the number chosen instead of 10, just like a normal choice study, so p-values have meaning.

Allocation of Prescription Drugs

The PHREG Procedure

Model Information

Data Set	WORK.CODED2
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Frequency Variable	Count
Ties Handling	BRESLOW

Number of Observations Read	583
Number of Observations Used	583
Sum of Frequencies Read	29700
Sum of Frequencies Used	29700

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Stratum	Set	Number of Alternatives	Chosen Alternatives	Not Chosen
1	1	1100.0	100.0	1000.0
2	2	1100.0	100.0	1000.0
3	3	1100.0	100.0	1000.0
4	4	1100.0	100.0	1000.0
5	5	1100.0	100.0	1000.0
6	6	1100.0	100.0	1000.0
7	7	1100.0	100.0	1000.0
8	8	1100.0	100.0	1000.0
9	9	1100.0	100.0	1000.0
10	10	1100.0	100.0	1000.0

11	11	1100.0	100.0	1000.0
12	12	1100.0	100.0	1000.0
13	13	1100.0	100.0	1000.0
14	14	1100.0	100.0	1000.0
15	15	1100.0	100.0	1000.0
16	16	1100.0	100.0	1000.0
17	17	1100.0	100.0	1000.0
18	18	1100.0	100.0	1000.0
19	19	1100.0	100.0	1000.0
20	20	1100.0	100.0	1000.0
21	21	1100.0	100.0	1000.0
22	22	1100.0	100.0	1000.0
23	23	1100.0	100.0	1000.0
24	24	1100.0	100.0	1000.0
25	25	1100.0	100.0	1000.0
26	26	1100.0	100.0	1000.0
27	27	1100.0	100.0	1000.0
Total		29700.0	2700.0	27000.0

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	37816.553	36472.307
AIC	37816.553	36496.307
SBC	37816.553	36567.119

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1344.2468	12	<.0001
Score	1834.0841	12	<.0001
Wald	1408.7678	12	<.0001

Multinomial Logit Parameter Estimates

	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Brand 1	1	2.09906	0.21395	96.2530	<.0001
Brand 2	1	2.09118	0.21404	95.4511	<.0001
Brand 3	1	3.54204	0.20503	298.4470	<.0001
Brand 4	1	2.09710	0.21398	96.0528	<.0001
Brand 5	1	2.08523	0.21411	94.8479	<.0001
Brand 6	1	2.03530	0.21470	89.8622	<.0001
Brand 7	1	2.06920	0.21430	93.2315	<.0001
Brand 8	1	2.08573	0.21411	94.8982	<.0001
Brand 9	1	2.11705	0.21375	98.0964	<.0001
Brand 10	1	2.06363	0.21436	92.6733	<.0001
$50	1	0.00529	0.05148	0.0106	0.9181
$75	1	0.0005304	0.05152	0.0001	0.9918
$100	0	0	.	.	.

Chair Design with Generic Attributes

This study illustrates creating an experimental design for a purely generic choice model. This example discusses generic attributes, alternative swapping, choice set swapping, and constant alternatives. In a purely generic study, there are no brands, just bundles of attributes. Say a manufacturer is interested in designing one or more new chairs. The manufacturer can vary the attributes of the chairs, present subjects with competing chair designs, and model the effects of the attributes on choice. Here are the attributes of interest.

Factor	Attribute	Levels
X1	Color	3 Colors
X2	Back	3 Styles
X3	Seat	3 Styles
X4	Arm Rest	3 Styles
X5	Material	3 Materials

Since seeing descriptions of chairs is not the same as seeing and sitting in the actual chairs, the manufacturer is going to actually make sample chairs for people to try and choose from. Subjects will be shown groups of three chairs at a time. If we were to make our design using the approach discussed in previous examples, we would use the `%MktEx` autocall macro to create a design with 15 factors, five for the first chair, five for the second chair, and five for the third chair. This design would have to have at least $15 \times (3-1)+1 = 31$ runs and 93 sample chairs. Here is how we could have made the design.

```
title 'Generic Chair Attributes';

* This design will not be used;
%mktex(3 ** 15, n=36, seed=238)

data key;
   input (x1-x5) ($) @@;
   datalines;
 x1  x2  x3  x4  x5
 x6  x7  x8  x9 x10
x11 x12 x13 x14 x15
;

%mktroll(design=randomized, key=key, out=cand);
```

The `%MktEx` approach to designing an experiment like this allows you to fit very general models including models with alternative-specific effects and even mother logit models. However, at analysis time for this purely generic model, we will fit a model with 10 parameters, two for each of the five factors, `class(x1-x5)`. Creating a design with over $31 \times 3 = 93$ chairs is way too expensive. In ordinary linear designs, we need at least as many runs as parameters. In choice designs, we need to count the total number of alternatives across all choice sets, subtract the number the number of choice sets, and this number must be at least as large as the number of parameters. Equivalently, each choice set allows us to estimate $m-1$ parameters, where m is the number of alternatives in that choice set. In this case, we could fit our purely generic model with as few as $10/(3-1) = 5$ choice sets.

Since we only need a simple generic model for this example, and since our chair manufacturing for our research will be expensive, we will not use the `%MktEx` approach for designing our choice experiment. Instead, we will use a different approach that will allow us to get a smaller design that is adequate for our model and budget. Recall the discussion of linear design efficiency, choice model design efficiency,

and using linear design efficiency as a surrogate for choice design goodness starting on page 86. Instead of using linear design efficiency as a surrogate for choice design goodness, we can directly optimize choice design efficiency given an assumed model and parameter vector $\boldsymbol{\beta}$. This approach uses the `%ChoicEff` macro.

Generic Attributes, Alternative Swapping, Large Candidate Set

This part of the example illustrates using the `%ChoicEff` macro for efficient choice designs, using its algorithm that builds a design from candidate alternatives (as opposed to candidates consisting of entire choice sets). First, we will use the `%MktRuns` macro to suggest a candidate-set size.

```
%mktruns(3 ** 5)
```

Here are some of the results.

Generic Chair Attributes

Design Summary

Number of Levels	Frequency
3	5

Saturated = 11
Full Factorial = 243

Some Reasonable Design Sizes	Violations	Cannot Be Divided By
18 *	0	
27 *	0	
36 *	0	
12	10	9
15	10	9
21	10	9
24	10	9
30	10	9
33	10	9
11	15	3 9

* - 100% Efficient Design can be made with the MktEx Macro.

Generic Chair Attributes

n	Design								Reference
18	2 **	1	3 **	7					Orthogonal Array
18			3 **	6	6 **	1			Orthogonal Array
27			3 **	13					Fractional-Factorial
27			3 **	9	9 **	1			Fractional-Factorial
36	2 **	11	3 **	12					Orthogonal Array
36	2 **	10	3 **	8	6 **	1			Orthogonal Array
36	2 **	4	3 **	13					Orthogonal Array
36	2 **	3	3 **	9	6 **	1			Orthogonal Array
36	2 **	2	3 **	12	6 **	1			Orthogonal Array
36	2 **	2	3 **	5	6 **	2			Orthogonal Array
36	2 **	1	3 **	8	6 **	2			Orthogonal Array
36			3 **	13	4 **	1			Orthogonal Array
36			3 **	12	12 **	1			Orthogonal Array
36			3 **	7	6 **	3			Orthogonal Array

We could use candidate sets of size: 18, 27 or 36. Additionally, since this problem is small, we could try an 81-run fractional-factorial design or the 243-run full-factorial design. We will choose the 243-run full-factorial design, since it is reasonably small and it will give the macro the most freedom to find a good design.[†]

We will use the `%MktEx` macro to create a candidate set. The candidate set will consist of 5 three-level factors, one for each of the five generic attributes. We will add three flag variables to the candidate set, `f1-f3`, one for each alternative. Since there are three alternatives, the candidate set must contain those observations that may be used for alternative 1, those observations that may be used for alternative 2, and those observations that may be used for alternative 3. The flag variable for each alternative consists of ones for those candidates that may be included for that alternative and zeros or missings for those candidates that may not be included for that alternative. The candidates for the different alternatives may be all different, all the same, or something in between depending on the problem. For example, the candidate set may contain one observation that is only used for the last, constant alternative. In this purely generic case, each flag variable consists entirely of ones indicating that any candidate can appear in any alternative. The `%MktEx` macro will not allow you to create constant or one-level factors. We can instead use the `%MktLab` macro to add the flag variables, essentially by specifying that we have multiple intercepts. The option `int=f1-f3` creates three variables with values all one. The default output data set is called FINAL. The following code creates the candidates.

```
%mktex(3 ** 5, n=243)
%mktlab(data=design, int=f1-f3)

proc print data=final(obs=27); run;
```

[†]Later, we will see we could have chosen 18.

The columns `f1-f3` are the flags, and `x1-x5` are the generic attributes. Here is part of the candidate set.

Generic Chair Attributes

Obs	f1	f2	f3	x1	x2	x3	x4	x5
1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	2
3	1	1	1	1	1	1	1	3
4	1	1	1	1	1	1	2	1
5	1	1	1	1	1	1	2	2
6	1	1	1	1	1	1	2	3
7	1	1	1	1	1	1	3	1
8	1	1	1	1	1	1	3	2
9	1	1	1	1	1	1	3	3
10	1	1	1	1	1	2	1	1
11	1	1	1	1	1	2	1	2
12	1	1	1	1	1	2	1	3
13	1	1	1	1	1	2	2	1
14	1	1	1	1	1	2	2	2
15	1	1	1	1	1	2	2	3
16	1	1	1	1	1	2	3	1
17	1	1	1	1	1	2	3	2
18	1	1	1	1	1	2	3	3
19	1	1	1	1	1	3	1	1
20	1	1	1	1	1	3	1	2
21	1	1	1	1	1	3	1	3
22	1	1	1	1	1	3	2	1
23	1	1	1	1	1	3	2	2
24	1	1	1	1	1	3	2	3
25	1	1	1	1	1	3	3	1
26	1	1	1	1	1	3	3	2
27	1	1	1	1	1	3	3	3

Next, we will search that candidate set for an efficient design for the model specification `class(x1-x5)` and the assumption $\boldsymbol{\beta} = \mathbf{0}$. We will use the `%ChoicEff` autocall macro to do this. (All of the autocall macros used in this book are documented starting on page 479.) This approach is based on the work of Huber and Zwerina (1996) who proposed constructing efficient experimental designs for choice experiments under an assumed model and $\boldsymbol{\beta}$. The `%ChoicEff` macro uses a modified Fedorov algorithm (Fedorov, 1972; Cook and Nachtsheim, 1980) to optimize the choice model variance matrix. We will be using the largest possible candidate set for this problem, the full-factorial design, and we will ask for more than the default number of iterations, so run time will be slower than it could be. However, we will be requesting a very small number of choice sets. Building the chairs will be expensive, so we want to get a really good but small design. This specification requests a generic design with six choice sets each consisting of three alternatives.

```
%choiceff(data=final, model=class(x1-x5), nsets=6, maxiter=100,
          seed=121, flags=f1-f3, beta=zero);
```

The `data=final` option names the input data set of candidates. The `model=class(x1-x5)` option specifies the most general model that will be considered at analysis time. The `nsets=6` option specifies the number of choice sets. Note that this is considerably smaller than the minimum of 31 that would be required if we were just using the `%MktEx` linear-design approach ($6 \times 3 = 18$ chairs instead of $31 \times 3 = 93$ chairs). The `maxiter=100` option requests 100 designs based on 100 random initial designs (by default, `maxiter=2`). The `seed=121` option specifies the random number seed. The `flags=f1-f3` specifies the flag variables for alternatives 1 to 3. Implicitly, this option also specifies the fact that there are three alternatives since three flag variables were specified. The `beta=zero` option specifies the assumption $\boldsymbol{\beta} = \mathbf{0}$. A vector of numbers like `beta=-1 0 -1 0 -1 0 -1 0 -1 0 -1 0` could be specified. (See page 489 for an example of this.) When you wish to assume all parameters are zero, you can specify `beta=zero` instead of typing a vector of the zeros. You can also omit the `beta=` option if you just want the macro to list the parameters. You can use this list to ensure that you specify the parameters in the right order.

The first part of the output from the macro is a list of all of the effects generated and the assumed values of $\boldsymbol{\beta}$. It is very important to check this list and make sure it is correct. In particular, when you are explicitly specifying the $\boldsymbol{\beta}$ vector, you need to make sure you specified all of the values in the right order.

Generic Chair Attributes

n	Name	Beta	Label
1	x11	0	x1 1
2	x12	0	x1 2
3	x21	0	x2 1
4	x22	0	x2 2
5	x31	0	x3 1
6	x32	0	x3 2
7	x41	0	x4 1
8	x42	0	x4 2
9	x51	0	x5 1
10	x52	0	x5 2

Next, the macro produces the iteration history, which is different from the iteration histories we are used to seeing in the `%MktEx` macro. The `%ChoicEff` macro uses PROC IML and a modified Fedorov algorithm to iteratively improve the efficiency of the choice design given the specified candidates, model, and $\boldsymbol{\beta}$. Note that these efficiencies are *not* on a 0 to 100 scale. This step took about 12 minutes. Here are some of the results.

Generic Chair Attributes

Design	Iteration	D-Efficiency	D-Error
1	0	0.352304	2.838455
	1	0.946001	1.057081
	2	1.001164	0.998838
	3	1.041130	0.960494
	4	1.044343	0.957540

.
.
.

Design	Iteration	D-Efficiency	D-Error
34	0	0.469771	2.128698
	1	0.919074	1.088051
	2	1.058235	0.944970
	3	1.154701	0.866025
	4	1.154701	0.866025

.
.
.

Design	Iteration	D-Efficiency	D-Error
100	0	0.456308	2.191501
	1	1.006320	0.993719
	2	1.042702	0.959046
	3	1.042702	0.959046

Next, the macro shows which design it chose and the final efficiency and *D*-Error (*D*-Efficiency = 1 / *D*-Error).

Final Results

Design	34
Choice Sets	6
Alternatives	3
D-Efficiency	1.154701
D-Error	0.866025

Next, it shows the variance, standard error, and *df* for each effect. It is important to ensure that each effect is estimable: ($df = 1$). Usually, when all of the variances are constant, like we see in this table, it means that the macro has found the optimal design.

Generic Chair Attributes

n	Variable Name	Label	Variance	DF	Standard Error
1	x11	x1 1	1	1	1
2	x12	x1 2	1	1	1
3	x21	x2 1	1	1	1
4	x22	x2 2	1	1	1
5	x31	x3 1	1	1	1
6	x32	x3 2	1	1	1
7	x41	x4 1	1	1	1
8	x42	x4 2	1	1	1
9	x51	x5 1	1	1	1
10	x52	x5 2	1	1	1
				==	
				10	

The data set BEST contains the final, best design found.

```
proc print; by set; id set; run;
```

The data set contains: `Design` - the number of the design with the maximum efficiency, `Efficiency` - the efficiency of this design, `Index` - the candidate set observation number, `Set` - the choice set number, `Prob` - the probability that this alternative will be chosen given $\boldsymbol{\beta}$, `n` - the observation number, `x1-x5` - the design, and `f1-f3` - the flags.

Generic Chair Attributes

Set	Design	Efficiency	Index	Prob	n	f1	f2	f3	x1	x2	x3	x4	x5
1	34	1.15470	183	0.33333	595	1	1	1	3	1	3	1	3
	34	1.15470	62	0.33333	596	1	1	1	1	3	1	3	2
	34	1.15470	121	0.33333	597	1	1	1	2	2	2	2	1
2	34	1.15470	217	0.33333	598	1	1	1	3	3	1	1	1
	34	1.15470	45	0.33333	599	1	1	1	1	2	2	3	3
	34	1.15470	104	0.33333	600	1	1	1	2	1	3	2	2
3	34	1.15470	215	0.33333	601	1	1	1	3	2	3	3	2
	34	1.15470	147	0.33333	602	1	1	1	2	3	2	1	3
	34	1.15470	4	0.33333	603	1	1	1	1	1	1	2	1
4	34	1.15470	78	0.33333	604	1	1	1	1	3	3	2	3
	34	1.15470	178	0.33333	605	1	1	1	3	1	2	3	1
	34	1.15470	110	0.33333	606	1	1	1	2	2	1	1	2
5	34	1.15470	90	0.33333	607	1	1	1	2	1	1	3	3
	34	1.15470	46	0.33333	608	1	1	1	1	2	3	1	1
	34	1.15470	230	0.33333	609	1	1	1	3	3	2	2	2

6	34	1.15470	195	0.33333	610	1	1	1	3	2	1	2	3
	34	1.15470	11	0.33333	611	1	1	1	1	1	2	1	2
	34	1.15470	160	0.33333	612	1	1	1	2	3	3	3	1

This design has 18 runs (6 choice sets × 3 alternatives). Notice that in this design, each level occurs exactly once in each factor and each choice set. To use this design for analysis, you would only need the variables `Set` and `x1-x5`. Since it is already in choice design format, it would not need to be processed using the `%MktRoll` macro. Since data collection, processing, and analysis have already been covered in detail in other examples, this example will concentrate solely on experimental design.

Generic Attributes, Alternative Swapping, Small Candidate Set

In this part of this example, we will try to make an equivalent design to the one we just made, only this time using a smaller candidate set. Here is the code.

```
%mktex(3 ** 5, n=18)

%mktlab(data=design, int=f1-f3)

%choiceff(data=final, model=class(x1-x5), nsets=6, maxiter=20,
          seed=121, flags=f1-f3, beta=zero);

proc print; run;
```

This time, instead of creating a full-factorial candidate set, we asked for 5 three-level factors from the L_{18}, an orthogonal tabled design in 18 runs. We also asked for fewer iterations in the `%ChoicEff` macro. Since the candidate set is much smaller, the macro should be able to find the best design available in this candidate set fairly easily. Here are some of the results.

Generic Chair Attributes

n	Name	Beta	Label
1	x11	0	x1 1
2	x12	0	x1 2
3	x21	0	x2 1
4	x22	0	x2 2
5	x31	0	x3 1
6	x32	0	x3 2
7	x41	0	x4 1
8	x42	0	x4 2
9	x51	0	x5 1
10	x52	0	x5 2

Generic Chair Attributes

Design	Iteration	D-Efficiency	D-Error
1	0	0	.
	1	0.913290	1.094943
	2	1.008888	0.991191
	3	1.042878	0.958885
	4	1.154701	0.866025
	5	1.154701	0.866025

.
.
.

Design	Iteration	D-Efficiency	D-Error
20	0	0.364703	2.741954
	1	0.851038	1.175036
	2	1.008888	0.991191
	3	1.042878	0.958885
	4	1.154701	0.866025
	5	1.154701	0.866025

Final Results

Design	1
Choice Sets	6
Alternatives	3
D-Efficiency	1.154701
D-Error	0.866025

Generic Chair Attributes

n	Variable Name	Label	Variance	DF	Standard Error
1	x11	x1 1	1	1	1
2	x12	x1 2	1	1	1
3	x21	x2 1	1	1	1
4	x22	x2 2	1	1	1
5	x31	x3 1	1	1	1
6	x32	x3 2	1	1	1
7	x41	x4 1	1	1	1
8	x42	x4 2	1	1	1
9	x51	x5 1	1	1	1
10	x52	x5 2	1	1	1
				==	
				10	

Generic Chair Attributes

Obs	Design	Efficiency	Index	Set	Prob	n	f1	f2	f3	x1	x2	x3	x4	x5
1	1	1.15470	11	1	0.33333	1	1	1	1	2	3	1	3	1
2	1	1.15470	13	1	0.33333	2	1	1	1	3	1	2	1	2
3	1	1.15470	4	1	0.33333	3	1	1	1	1	2	3	2	3
4	1	1.15470	3	2	0.33333	4	1	1	1	1	2	1	3	2
5	1	1.15470	12	2	0.33333	5	1	1	1	2	3	2	1	3
6	1	1.15470	14	2	0.33333	6	1	1	1	3	1	3	2	1
7	1	1.15470	5	3	0.33333	7	1	1	1	1	3	2	2	1
8	1	1.15470	8	3	0.33333	8	1	1	1	2	1	3	3	2
9	1	1.15470	15	3	0.33333	9	1	1	1	3	2	1	1	3
10	1	1.15470	9	4	0.33333	10	1	1	1	2	2	2	2	2
11	1	1.15470	1	4	0.33333	11	1	1	1	1	1	1	1	1
12	1	1.15470	18	4	0.33333	12	1	1	1	3	3	3	3	3
13	1	1.15470	10	5	0.33333	13	1	1	1	2	2	3	1	1
14	1	1.15470	17	5	0.33333	14	1	1	1	3	3	1	2	2
15	1	1.15470	2	5	0.33333	15	1	1	1	1	1	2	3	3
16	1	1.15470	6	6	0.33333	16	1	1	1	1	3	3	1	2
17	1	1.15470	7	6	0.33333	17	1	1	1	2	1	1	2	3
18	1	1.15470	16	6	0.33333	18	1	1	1	3	2	2	3	1

Notice we got the same D-efficiency and variances as before (D-efficiency = 1.1547005384 and all variances 1). Also notice the `Index` variable in the design (which is the candidate set row number). Each candidate appears in the design exactly once. We have frequently found for problems like this (all generic attributes, no brands, no constant alternative, total number of alternatives equal to the number of runs in an orthogonal design, all factors available in that orthogonal design, and an assumed $\boldsymbol{\beta}$ vector of zero) that the optimal design can be created by optimally sorting the rows of an orthogonal design into choice sets, and the `%ChoicEff` macro can do this quite well.

Six choice sets is a bit small. If you can afford a larger number, it would be good to try a larger design. In this case, nine choice sets are requested using a fractional-factorial candidate set in 27 runs. Notice that like before, the number of runs in the candidate set was chosen to be the product of the number of choice sets and the number of alternatives in each choice set.

```
%mktex(3 ** 5, n=27, seed=382)

%mktlab(data=design, int=f1-f3)

%choiceff(data=final, model=class(x1-x5), nsets=9, maxiter=20,
          seed=121, flags=f1-f3, beta=zero);

proc print; id set; by set; var index prob x:; run;
```

Here are the variances and the design.

Generic Chair Attributes

n	Variable Name	Label	Variance	DF	Standard Error
1	x11	x1 1	0.66667	1	0.81650
2	x12	x1 2	0.66667	1	0.81650
3	x21	x2 1	0.66667	1	0.81650
4	x22	x2 2	0.66667	1	0.81650
5	x31	x3 1	0.66667	1	0.81650
6	x32	x3 2	0.66667	1	0.81650
7	x41	x4 1	0.66667	1	0.81650
8	x42	x4 2	0.66667	1	0.81650
9	x51	x5 1	0.66667	1	0.81650
10	x52	x5 2	0.66667	1	0.81650
				==	
				10	

Generic Chair Attributes

Set	Index	Prob	x1	x2	x3	x4	x5
1	9	0.33333	1	3	3	1	2
	13	0.33333	2	2	1	3	3
	20	0.33333	3	1	2	2	1
2	25	0.33333	3	3	1	2	2
	5	0.33333	1	2	2	1	3
	12	0.33333	2	1	3	3	1
3	6	0.33333	1	2	3	3	1
	26	0.33333	3	3	2	1	3
	10	0.33333	2	1	1	2	2
4	22	0.33333	3	2	1	1	1
	2	0.33333	1	1	2	3	2
	18	0.33333	2	3	3	2	3
5	11	0.33333	2	1	2	1	3
	4	0.33333	1	2	1	2	2
	27	0.33333	3	3	3	3	1
6	8	0.33333	1	3	2	2	1
	19	0.33333	3	1	1	3	3
	15	0.33333	2	2	3	1	2
7	3	0.33333	1	1	3	2	3
	23	0.33333	3	2	2	3	2
	16	0.33333	2	3	1	1	1

8	17	0.33333	2	3	2	3	2
	24	0.33333	3	2	3	2	3
	1	0.33333	1	1	1	1	1
9	7	0.33333	1	3	1	3	3
	14	0.33333	2	2	2	2	1
	21	0.33333	3	1	3	1	2

Notice that like before, the variances are constant, but in this case smaller at 2/3, and each candidate appears once. This appears to be an optimal design in 9 choice sets.

Generic Attributes, a Constant Alternative, and Alternative Swapping

Now let's make a design for the same problem but this time with a constant alternative. We will first use the `%MktEx` macro just like before to make a design for the nonconstant alternatives. We will then use a DATA step to add the flags and a constant alternative.

```
title 'Generic Chair Attributes';

%mktex(3 ** 5, n=243, seed=306)

data final(drop=i);
   set design end=eof;
   retain f1-f3 1 f4 0;
   output;
   if eof then do;
      array x[9] x1-x5 f1-f4;
      do i = 1 to 9; x[i] = i le 5 or i eq 9; end;
      output;
      end;
   run;

proc print data=final(where=(x1 eq x3 and x2 eq x4 and x3 eq x5 or f4)); run;
```

Here is a sample of the observations in the candidate set.

Generic Chair Attributes

Obs	x1	x2	x3	x4	x5	f1	f2	f3	f4
1	1	1	1	1	1	1	1	1	0
31	1	2	1	2	1	1	1	1	0
61	1	3	1	3	1	1	1	1	0
92	2	1	2	1	2	1	1	1	0
122	2	2	2	2	2	1	1	1	0
152	2	3	2	3	2	1	1	1	0
183	3	1	3	1	3	1	1	1	0
213	3	2	3	2	3	1	1	1	0
243	3	3	3	3	3	1	1	1	0
244	1	1	1	1	1	0	0	0	1

The first 243 observations may be used for any of the first three alternatives and the *244th* observation may only be used for fourth or constant alternative. In this example, the constant alternative is composed solely from the first level of each factor. Of course this could be changed depending on the situation. The `%ChoicEff` macro invocation is the same as before, except now we have four flags.

```
%choiceff(data=final, model=class(x1-x5), nsets=6, maxiter=100,
          seed=121, flags=f1-f4, beta=zero);

proc print; by set; id set; run;
```

You can see in the final design that there are now four alternatives and the last alternative in each choice set is constant and is always flagged by `f4=1`. In the interest of space, most of the iteration histories are omitted.

Generic Chair Attributes

n	Name	Beta	Label
1	x11	0	x1 1
2	x12	0	x1 2
3	x21	0	x2 1
4	x22	0	x2 2
5	x31	0	x3 1
6	x32	0	x3 2
7	x41	0	x4 1
8	x42	0	x4 2
9	x51	0	x5 1
10	x52	0	x5 2

Generic Chair Attributes

Design	Iteration	D-Efficiency	D-Error
1	0	0.424723	2.354476
	1	0.900662	1.110294
	2	0.939090	1.064861
	3	0.943548	1.059830

.
.
.

Design	Iteration	D-Efficiency	D-Error
13	0	0.494007	2.024263
	1	0.873818	1.144404
	2	0.915135	1.092735
	3	0.960392	1.041241
	4	0.999769	1.000231
	5	1.003398	0.996614

.
.
.

Design	Iteration	D-Efficiency	D-Error
100	0	0.528399	1.892509
	1	0.883854	1.131408
	2	0.924346	1.081846
	3	0.939811	1.064044
	4	0.942047	1.061518

Generic Chair Attributes

Final Results

Design	13
Choice Sets	6
Alternatives	4
D-Efficiency	1.003398
D-Error	0.996614

Generic Chair Attributes

n	Variable Name	Label	Variance	DF	Standard Error
1	x11	x1 1	1.14695	1	1.07096
2	x12	x1 2	1.33333	1	1.15470
3	x21	x2 1	1.14695	1	1.07096
4	x22	x2 2	1.33333	1	1.15470
5	x31	x3 1	1.19793	1	1.09450
6	x32	x3 2	1.27439	1	1.12889
7	x41	x4 1	1.19793	1	1.09450
8	x42	x4 2	1.27439	1	1.12889
9	x51	x5 1	1.13102	1	1.06350
10	x52	x5 2	1.27439	1	1.12889
				==	
				10	

Generic Chair Attributes

Set	Design	Efficiency	Index	Prob	n	x1	x2	x3	x4	x5	f1	f2	f3	f4
1	13	1.00340	152	0.25	289	2	3	2	3	2	1	1	1	0
	13	1.00340	213	0.25	290	3	2	3	2	3	1	1	1	0
	13	1.00340	15	0.25	291	1	1	2	2	3	1	1	1	0
	13	1.00340	244	0.25	292	1	1	1	1	1	0	0	0	1
2	13	1.00340	154	0.25	293	2	3	3	1	1	1	1	1	0
	13	1.00340	15	0.25	294	1	1	2	2	3	1	1	1	0
	13	1.00340	197	0.25	295	3	2	1	3	2	1	1	1	0
	13	1.00340	244	0.25	296	1	1	1	1	1	0	0	0	1
3	13	1.00340	108	0.25	297	2	1	3	3	3	1	1	1	0
	13	1.00340	220	0.25	298	3	3	1	2	1	1	1	1	0
	13	1.00340	38	0.25	299	1	2	2	1	2	1	1	1	0
	13	1.00340	244	0.25	300	1	1	1	1	1	0	0	0	1
4	13	1.00340	121	0.25	301	2	2	2	2	1	1	1	1	0
	13	1.00340	182	0.25	302	3	1	3	1	2	1	1	1	0
	13	1.00340	63	0.25	303	1	3	1	3	3	1	1	1	0
	13	1.00340	244	0.25	304	1	1	1	1	1	0	0	0	1
5	13	1.00340	111	0.25	305	2	2	1	1	3	1	1	1	0
	13	1.00340	77	0.25	306	1	3	3	2	2	1	1	1	0
	13	1.00340	178	0.25	307	3	1	2	3	1	1	1	1	0
	13	1.00340	244	0.25	308	1	1	1	1	1	0	0	0	1

6	13	1.00340	228	0.25	309	3	3	2	1	3	1	1	1	0
	13	1.00340	52	0.25	310	1	2	3	3	1	1	1	1	0
	13	1.00340	86	0.25	311	2	1	1	2	2	1	1	1	0
	13	1.00340	244	0.25	312	1	1	1	1	1	0	0	0	1

When there were three alternatives, each alternative had a probability of choice of 1/3, and now with four alternatives, the probability is 1/4. They are all equal because of the assumption $\boldsymbol{\beta} = \mathbf{0}$. With other assumptions about $\boldsymbol{\beta}$, typically the probabilities will not all be equal. To use this design for analysis, you would only need the variables `Set` and `x1-x5`. Since it is already in choice design format (one row per alternative), it would not need to be processed using the `%MktRoll` macro. Note that when you make designs with the `%ChoicEff` macro, the `model` statement in PROC TRANSREG should match or be no more complicated than the `model` specification that generated the design:

```
model class(x1-x5);
```

A model with fewer degrees of freedom is safe, although the design will be suboptimal. For example, if `x1-x5` are quantitative attributes, this would be safe:

```
model identity(x1-x5);
```

However, specifying interactions, or using this design in a branded study and specifying alternative-specific effects like this could lead to quite a few inestimable parameters.

```
* Bad idea for this design!!;
model class(x1-x5 x1*x2 x4*x5);

* Another bad idea for this design!!;
model class(brand)
      class(brand * x1 brand * x2 brand * x3 brand * x4 brand * x5);
```

Generic Attributes, a Constant Alternative, and Choice Set Swapping

The `%ChoicEff` macro can be used in a very different way. Instead of providing a candidate set of alternatives to swap in and out of the design, you can provide a candidate set of entire choice sets. For this particular example, swapping alternatives will almost certainly be better (see page 322). However, sometimes, if you need to impose restrictions on which alternative can appear with which other alternative, then you must use the set-swapping options. We will start by using the `%MktEx` macro to make a candidate design, with one run per choice set and one factor for each attribute of each alternative (just like we did in the vacation, fabric softener, and food examples). We will then process the candidates from one row per choice set to one row per alternative per choice set using the `%MktRoll` macro.

```
%mktex(3 ** 15, n=81 * 81, seed=522)

%mktkey(x1-x15)

data key;
   input (x1-x5) ($);
   datalines;
 x1  x2  x3  x4  x5
 x6  x7  x8  x9 x10
x11 x12 x13 x14 x15
  .   .   .   .   .
;

%mktroll(design=randomized, key=key, out=rolled)

* Code the constant alternative;
data final;
   set rolled;
   if _alt_ = '4' then do; x1 = 1; x2 = 1; x3 = 1; x4 = 1; x5 = 1; end;
   run;

proc print; by set; id set; where set in (1, 100, 1000, 5000, 6561); run;
```

The %MktKey macro produced the following line, which we copied, pasted, and edited to make the KEY data set.

```
x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15
```

Here are a few of the candidate choice sets.

Generic Chair Attributes

Set	_Alt_	x1	x2	x3	x4	x5
1	1	2	1	1	2	3
	2	3	3	1	1	1
	3	2	2	1	1	3
	4	1	1	1	1	1
100	1	1	2	1	2	1
	2	2	3	3	2	1
	3	2	2	3	3	2
	4	1	1	1	1	1
1000	1	2	1	2	1	3
	2	2	1	2	1	2
	3	1	3	2	2	2
	4	1	1	1	1	1

```
5000      1       3     1     3     2     3
          2       3     3     3     3     2
          3       1     2     1     2     3
          4       1     1     1     1     1

6561      1       1     3     1     2     2
          2       3     2     2     2     2
          3       1     3     3     1     3
          4       1     1     1     1     1
```

Next, we will then run the `%ChoicEff` macro, only this time we will specify `nalts=4` instead of `flags=f1-f4`. Since there are no alternative flag variables to count, we have to tell the macro how many alternatives are in each choice set. We will also ask for fewer iterations since the candidate set is large.

```
%choiceff(data=final, model=class(x1-x5), nsets=6, nalts=4, maxiter=10,
          beta=zero, seed=109);
```

Generic Chair Attributes

n	Name	Beta	Label
1	x11	0	x1 1
2	x12	0	x1 2
3	x21	0	x2 1
4	x22	0	x2 2
5	x31	0	x3 1
6	x32	0	x3 2
7	x41	0	x4 1
8	x42	0	x4 2
9	x51	0	x5 1
10	x52	0	x5 2

Generic Chair Attributes

Design	Iteration	D-Efficiency	D-Error
1	0	0.536166	1.865093
	1	0.848201	1.178966
	2	0.872298	1.146398
	3	0.872298	1.146398

.
.
.

Design	Iteration	D-Efficiency	D-Error
5	0	0.529592	1.888245
	1	0.836422	1.195568
	2	0.861051	1.161372
	3	0.898936	1.112426
	4	0.904411	1.105692
	5	0.904411	1.105692

.
.
.

Design	Iteration	D-Efficiency	D-Error
10	0	0.539774	1.852627
	1	0.820582	1.218648
	2	0.846874	1.180814
	3	0.869219	1.150458
	4	0.869219	1.150458

Generic Chair Attributes

Final Results

Design	5
Choice Sets	6
Alternatives	4
D-Efficiency	0.904411
D-Error	1.105692

Generic Chair Attributes

n	Variable Name	Label	Variance	DF	Standard Error
1	x11	x1 1	1.14609	1	1.07056
2	x12	x1 2	2.32530	1	1.52489
3	x21	x2 1	1.48741	1	1.21959
4	x22	x2 2	1.95354	1	1.39769
5	x31	x3 1	1.16334	1	1.07858
6	x32	x3 2	1.50116	1	1.22522
7	x41	x4 1	1.34713	1	1.16066
8	x42	x4 2	1.35845	1	1.16552
9	x51	x5 1	1.27405	1	1.12874
10	x52	x5 2	1.54939	1	1.24475
				==	
				10	

This design is less efficient than we found using the alternative-swapping algorithm, so we will not use it.

Design Algorithm Comparisons

It is instructive to compare the three approaches outlined in this chapter in the context of this problem. There are $3^{3\times5} = 14,348,907$ choice sets for this problem (three-level factors and 3 alternatives times 5 factors per alternative). If we were to use the pure linear design approach using the `%MktEx` macro, we could never begin to consider all possible candidate choice sets. Similarly, with the choice-set-swapping algorithm of the `%ChoicEff` macro, we could never begin to consider all possible candidate choice sets. Furthermore, with the linear design approach, we could not create a design with six choice sets since the minimum size is $2 \times 15 + 1 = 31$. Now consider the alternative-swapping algorithm. It uses at most a candidate set with only 244 observations ($3^5 + 1$). From it, every possible choice set can potentially be constructed, although the macro will only consider a tiny fraction of the possibilities. Hence, the alternative swapping will usually find a better design, because the candidate set does not limit it.

Both uses of the `%ChoicEff` macro have the advantage that they are explicitly minimizing the variances of the parameter estimates given a model and a $\boldsymbol{\beta}$ vector. They can be used to produce smaller, more specialized, and better designs. However, if the $\boldsymbol{\beta}$ vector or model is badly misspecified, the designs could be horrible. How badly do things have to be misspecified before you will have problems? Who knows. More research is needed. In contrast, the linear model `%MktEx` approach is very conservative and safe in that it should let you specify a very general model and still produce estimable parameters. The cost is you may be using many more choice sets than you need, particularly for nonbranded generic attributes. If you really have some information about your parameters, you should use them to produce a smaller and better design. However, if you have little or no information about parameters and if you anticipate specifying very general models like mother logit, then you probably want to use the linear design approach.

Initial Designs

This section illustrates some design strategies that involve improving on or augmenting initial designs. We will not actually use any designs from this section.

Improving an Existing Design

Sometimes, it is useful to try to improve an existing design. In this example, we use the `%MktEx` macro to create a design in 80 runs for 25 four-level factors. In the next step, we specify `init=`, and the macro goes straight into the design refinement history seeking to refine the input design. You might want to do this for example whenever you have a good, but not 100% efficient design, and you are willing to wait a few minutes to see if the macro can make it any better.

```
title 'Try to Improve an Existing Design';

%mktex(4 ** 25, n=80, seed=368)
%mktex(4 ** 25, n=80, seed=306, init=design, maxtime=20)
```

Here is the *D*-efficiency of the final design from the first step.

Try to Improve an Existing Design

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	91.4106	83.9583	97.6073	0.9747

This is a large problem. One in which the `maxtime=` option may cause the macro to stop before it reaches the maximum number of iterations. Running a second refinement step might help improve the design by adding a few more iterations. Here are the results from the second step.

Design Refinement History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	91.4106	91.4106	Ini
1	Start	90.0771		Pre,Mut,Ann
1	End	91.3476		

```
2      Start       88.8927                Pre,Mut,Ann
2     36  12       91.4181       91.4181
2     56   6       91.4285       91.4285
2      7  17       91.4372       91.4372
2     13  10       91.4373       91.4373
2     23  18       91.4404       91.4404
2     17  16       91.4445       91.4445
2     34   6       91.4572       91.4572
2     56  19       91.4673       91.4673
2     56  21       91.4768       91.4768
2     77   1       91.4821       91.4821
2     23  18       91.4827       91.4827
2     48   3       91.4848       91.4848
2     48   9       91.4863       91.4863
2     40  18       91.4863       91.4863
2        End       91.4863
.
.
.

6      Start       90.2194                Pre,Mut,Ann
6     63  19       91.5811       91.5811
6     68  18       91.5835       91.5835
6        End       91.5751


7      Start       89.4607                Pre,Mut,Ann
7     25   4       91.5851       91.5851
7     34   7       91.5902       91.5902
7     47   2       91.5913       91.5913
7     48  14       91.5930       91.5930
7     56   4       91.5955       91.5955
7     56  15       91.5999       91.5999
7     60   6       91.6142       91.6142
7     68   7       91.6172       91.6172
7     78   5       91.6172       91.6172
7     13  21       91.6249       91.6249
7     18  19       91.6249       91.6249
7     43  10       91.6249       91.6249
7     48  14       91.6282       91.6282
7     50  22       91.6408       91.6408
7     61   4       91.6417       91.6417
7     80  15       91.6430       91.6430
7     46  12       91.6430       91.6430
7     48   6       91.6430       91.6430
7        End       91.6430
.
.
.
```

```
10      Start       89.8707                Pre,Mut,Ann
10        End       91.6629
 .
 .
 .
```

```
        Try to Improve an Existing Design

              The OPTEX Procedure
```

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	91.7082	83.9853	97.5951	0.9747

The macro skips the normal first steps, algorithm search and design search, and goes straight into the design refinement search. In this example a small improvement was found, although often, no improvement is found.

When Some Choice Sets are Fixed in Advance

Sometimes certain runs or choice sets are fixed in advance and must be included in the design. The `%MktEx` macro can be used to efficiently augment a starting design with other choice sets. Suppose that you can make a choice design from the L_{36} ($2^{11}3^{12}$). In addition, you want to optimally add four more choice sets to use as holdouts. First we will look at how to do this using the `fixed=` option. This option can be used for fairly general design augmentation and refinement problems. On page 330, we will see an easier way to handle this particular problem using the `holdouts=` option.

You can create the design in 36 runs as before. Next, a DATA step is used to add a flag variable `f` that has values of 1 for the original 36 runs. In addition, four more runs are added (just copies of the last run) but with a flag value of missing. When this variable is specified on the `fixed=f` option, it indicates that the first 36 runs of the `init=init` design are *fixed* − they may not change. The remaining 4 runs are to be randomly initialized and optimally refined to maximize the *D*-efficiency of the overall 40-run design. We specified `options=nosort` so that the additional runs would stay at the end of the design.

```
title 'Augment a Design';

%mktex(n=36, seed=292)
```

```
data init;
   set randomized end = eof;
   f = 1;
   output;
   if eof then do;
      f = .;
      do i = 1 to 4; output; end;
      drop i;
      end;
   run;

proc print; run;

%mktex(2 ** 11 3 ** 12, n=40, init=init, fixed=f, seed=513, options=nosort)

proc print; run;
```

Here is the initial design.

Augment a Design

O										x	x	x	x	x	x	x	x	x	x	x	x	x	
b	x	x	x	x	x	x	x	x	x	1	1	1	1	1	1	1	1	1	1	2	2	2	2	
s	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	f
1	2	1	2	2	2	2	1	1	1	2	1	2	3	1	1	1	2	3	2	3	1	3	2	1
2	2	1	1	2	2	1	2	1	2	1	2	3	1	2	1	3	2	3	1	2	2	3	1	1
3	2	2	1	1	1	1	1	1	1	2	2	1	2	3	1	1	2	2	3	2	3	3	3	1
4	1	1	1	1	2	1	1	2	2	2	1	1	2	3	2	3	2	3	2	1	2	1	2	1
5	1	2	1	2	2	2	1	2	1	1	2	1	1	2	2	2	1	3	3	3	3	3	2	1
6	2	1	1	2	2	1	2	1	2	1	2	1	2	1	3	2	1	2	2	3	1	1	3	1
7	1	2	2	1	2	1	2	1	1	1	1	3	3	2	2	1	1	2	2	1	2	3	3	1
8	2	2	1	1	1	1	1	1	1	2	2	2	3	2	3	3	1	1	1	3	2	1	2	1
9	2	2	2	2	1	1	1	2	2	1	1	2	2	2	3	3	3	3	3	1	1	3	3	1
10	2	2	2	2	1	1	1	2	2	1	1	1	1	3	1	1	1	1	2	3	2	2	1	1
11	1	1	2	1	1	2	1	1	2	1	2	3	2	2	2	1	2	1	3	3	1	1	1	1
12	2	2	2	1	2	2	2	2	2	2	2	1	3	3	2	3	3	2	1	3	1	3	1	1
13	1	2	1	2	2	2	1	2	1	1	2	3	3	3	3	3	2	1	2	2	1	2	3	1
14	1	1	2	1	1	2	1	1	2	1	2	1	3	1	1	3	1	3	1	1	3	2	3	1
15	2	1	1	1	1	2	2	2	1	1	1	1	3	2	3	1	3	3	2	2	3	1	1	1
16	1	2	2	1	2	1	2	1	1	1	1	2	2	3	3	2	2	3	1	3	3	2	1	1
17	2	1	1	1	1	2	2	2	1	1	1	3	2	3	1	2	1	1	1	1	1	3	2	1
18	1	1	1	1	2	1	1	2	2	2	1	3	1	1	3	1	3	1	1	3	3	3	3	1
19	2	1	2	2	2	2	1	1	1	2	1	1	2	2	2	2	3	1	1	2	2	2	3	1
20	2	1	1	1	1	2	2	2	1	1	1	2	1	1	2	3	2	2	3	3	2	2	3	1

21	1	2	1	2	1	2	2	1	2	2	1	1	3	1	3	2	2	1	3	1	2	3	1	1
22	1	1	2	1	1	2	1	1	2	1	2	2	1	3	3	2	3	2	2	2	2	3	2	1
23	2	2	2	1	2	2	2	2	2	2	2	3	2	1	3	1	1	3	3	2	2	2	2	1
24	1	2	2	1	2	1	2	1	1	1	1	1	1	1	1	3	3	1	3	2	1	1	2	1
25	1	1	2	2	1	1	2	2	1	2	2	1	1	2	3	1	2	2	1	1	1	2	2	1
26	1	1	2	2	1	1	2	2	1	2	2	2	2	1	2	3	1	1	2	2	3	3	1	1
27	1	1	2	2	1	1	2	2	1	2	2	3	3	3	1	2	3	3	3	3	2	1	3	1
28	1	2	1	2	2	2	1	2	1	1	2	2	2	1	1	1	3	2	1	1	2	1	1	1
29	2	1	2	2	2	2	1	1	1	2	1	3	1	3	3	3	1	2	3	1	3	1	1	1
30	2	2	1	1	1	1	1	1	1	2	2	3	1	1	2	2	3	3	2	1	1	2	1	1
31	1	2	1	2	1	2	2	1	2	2	1	2	1	3	2	1	1	3	1	2	1	1	3	1
32	2	1	1	2	2	1	2	1	2	1	2	2	3	3	2	1	3	1	3	1	3	2	2	1
33	2	2	2	2	1	1	1	2	2	1	1	3	3	1	2	2	2	2	1	2	3	1	2	1
34	1	2	1	2	1	2	2	1	2	2	1	3	2	2	1	3	3	2	2	3	3	2	2	1
35	2	2	2	1	2	2	2	2	2	2	2	2	1	2	1	2	2	1	2	1	3	1	3	1
36	1	1	1	1	2	1	1	2	2	2	1	2	3	2	1	2	1	2	3	2	1	2	1	1
37	1	1	1	1	2	1	1	2	2	2	1	2	3	2	1	2	1	2	3	2	1	2	1	.
38	1	1	1	1	2	1	1	2	2	2	1	2	3	2	1	2	1	2	3	2	1	2	1	.
39	1	1	1	1	2	1	1	2	2	2	1	2	3	2	1	2	1	2	3	2	1	2	1	.
40	1	1	1	1	2	1	1	2	2	2	1	2	3	2	1	2	1	2	3	2	1	2	1	.

Here is the iteration history for the augmentation.

Augment a Design

Design Refinement History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	97.0559	97.0559	Ini
1	Start	97.0788	97.0788	Pre,Mut,Ann
1	37 5	97.0805	97.0805	
1	37 8	97.0816	97.0816	
1	37 9	97.1049	97.1049	
1	37 11	97.1133	97.1133	
1	37 13	97.1177	97.1177	
1	38 1	97.1410	97.1410	
1	38 2	97.1605	97.1605	
1	38 3	97.1729	97.1729	
1	38 4	97.1822	97.1822	
1	38 6	97.1905	97.1905	
1	38 8	97.1944	97.1944	
1	39 2	97.1991	97.1991	
1	39 13	97.2007	97.2007	
1	39 19	97.2007	97.2007	

```
1     40   9       97.2007       97.2007
1     40  10       97.2007       97.2007
1     37  18       97.2023       97.2023
1     37   3       97.2028       97.2028
1     37   4       97.2028       97.2028
1     37  23       97.2043       97.2043
1        End       97.2043

2      Start       97.2043       97.2043  Pre,Mut,Ann
2     40  21       97.2043       97.2043
2     38   2       97.2043       97.2043
2     40   9       97.2043       97.2043
2     40  21       97.2043       97.2043
2        End       97.2043

3      Start       97.2043       97.2043  Pre,Mut,Ann
3        End       97.2043

4      Start       97.2002                Pre,Mut,Ann
4     39  12       97.2043       97.2043
4     39  23       97.2043       97.2043
4     39  16       97.2043       97.2043
4        End       97.2043

5      Start       97.2043       97.2043  Pre,Mut,Ann
5     37   3       97.2043       97.2043
5     37  15       97.2043       97.2043
5        End       97.2043

6      Start       97.2043       97.2043  Pre,Mut,Ann
6     40   1       97.2043       97.2043
6     39  16       97.2043       97.2043
6     38  16       97.2043       97.2043
6        End       97.2043
NOTE: Stopping since it appears that no improvement is possible.
```

Notice that the macro goes straight into the design refinement stage. Also notice that in the iteration history, only rows 37 through 40 are changed. Here is the design. The last four rows are the holdouts.

Augment a Design

0										x	x	x	x	x	x	x	x	x	x	x	x	x	x	
b	x	x	x	x	x	x	x	x	x	1	1	1	1	1	1	1	1	1	1	2	2	2	2	
s	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	f
1	2	1	2	2	2	2	1	1	1	2	1	2	3	1	1	1	2	3	2	3	1	3	2	1
2	2	1	1	2	2	1	2	1	2	1	2	3	1	2	1	3	2	3	1	2	2	3	1	1
3	2	2	1	1	1	1	1	1	1	2	2	1	2	3	1	1	2	2	3	2	3	3	3	1
4	1	1	1	1	2	1	1	2	2	2	1	1	2	3	2	3	2	3	2	1	2	1	2	1
5	1	2	1	2	2	2	1	2	1	1	2	1	1	2	2	2	1	3	3	3	3	3	2	1
6	2	1	1	2	2	1	2	1	2	1	2	1	2	1	3	2	1	2	2	3	1	1	3	1
7	1	2	2	1	2	1	2	1	1	1	1	3	3	2	2	1	1	2	2	1	2	3	3	1
8	2	2	1	1	1	1	1	1	1	2	2	2	3	2	3	3	1	1	1	3	2	1	2	1
9	2	2	2	2	1	1	1	2	2	1	1	2	2	2	3	3	3	3	3	1	1	3	3	1
10	2	2	2	2	1	1	1	2	2	1	1	1	1	3	1	1	1	1	2	3	2	2	1	1
11	1	1	2	1	1	2	1	1	2	1	2	3	2	2	2	1	2	1	3	3	1	1	1	1
12	2	2	2	1	2	2	2	2	2	2	2	1	3	3	2	3	3	2	1	3	1	3	1	1
13	1	2	1	2	2	2	1	2	1	1	2	3	3	3	3	3	2	1	2	2	1	2	3	1
14	1	1	2	1	1	2	1	1	2	1	2	1	3	1	1	3	1	3	1	1	3	2	3	1
15	2	1	1	1	1	2	2	2	1	1	1	1	3	2	3	1	3	3	2	2	3	1	1	1
16	1	2	2	1	2	1	2	1	1	1	1	2	2	3	3	2	2	3	1	3	3	2	1	1
17	2	1	1	1	1	2	2	2	1	1	1	3	2	3	1	2	1	1	1	1	1	3	2	1
18	1	1	1	1	2	1	1	2	2	2	1	3	1	1	3	1	3	1	1	3	3	3	3	1
19	2	1	2	2	2	2	1	1	1	2	1	1	2	2	2	2	3	1	1	2	2	2	3	1
20	2	1	1	1	1	2	2	2	1	1	1	2	1	1	2	3	2	2	3	3	2	2	3	1
21	1	2	1	2	1	2	2	1	2	2	1	1	3	1	3	2	2	1	3	1	2	3	1	1
22	1	1	2	1	1	2	1	1	2	1	2	2	1	3	3	2	3	2	2	2	2	3	2	1
23	2	2	2	1	2	2	2	2	2	2	2	3	2	1	3	1	1	3	3	2	2	2	2	1
24	1	2	2	1	2	1	2	1	1	1	1	1	1	1	1	3	3	1	3	2	1	1	2	1
25	1	1	2	2	1	1	2	2	1	2	2	1	1	2	3	1	2	2	1	1	1	2	2	1
26	1	1	2	2	1	1	2	2	1	2	2	2	2	1	2	3	1	1	2	2	3	3	1	1
27	1	1	2	2	1	1	2	2	1	2	2	3	3	3	1	2	3	3	3	3	2	1	3	1
28	1	2	1	2	2	2	1	2	1	1	2	2	2	1	1	1	3	2	1	1	2	1	1	1
29	2	1	2	2	2	2	1	1	1	2	1	3	1	3	3	3	1	2	3	1	3	1	1	1
30	2	2	1	1	1	1	1	1	1	2	2	3	1	1	2	2	3	3	2	1	1	2	1	1
31	1	2	1	2	1	2	2	1	2	2	1	2	1	3	2	1	1	3	1	2	1	1	3	1
32	2	1	1	2	2	1	2	1	2	1	2	2	3	3	2	1	3	1	3	1	3	2	2	1
33	2	2	2	2	1	1	1	2	2	1	1	3	3	1	2	2	2	2	1	2	3	1	2	1
34	1	2	1	2	1	2	2	1	2	2	1	3	2	2	1	3	3	2	2	3	3	2	2	1
35	2	2	2	1	2	2	2	2	2	2	2	2	1	2	1	2	2	1	2	1	3	1	3	1
36	1	1	1	1	2	1	1	2	2	2	1	2	3	2	1	2	1	2	3	2	1	2	1	1
37	1	2	2	2	2	1	1	2	1	1	1	3	2	3	3	2	1	2	1	3	1	3	1	.
38	2	2	2	1	2	1	2	2	1	1	1	2	2	1	2	3	3	3	1	1	3	2	3	.
39	1	2	2	2	2	2	2	1	2	2	2	3	3	3	1	2	1	3	2	1	3	2	3	.
40	2	1	1	1	1	1	1	2	2	2	1	2	1	2	3	2	1	3	2	1	1	2	1	.

This code does the same thing only using the `holdouts=4` option instead.

```
title 'Augment a Design';

%mktex(n=36, seed=292)
%mktex(2 ** 11 3 ** 12, n=40, init=randomized,
       holdouts=4, seed=513, options=nosort)

proc print data=design(firstobs=37); run;
```

Here are the holdout observations, which are the same as we saw previously.

Augment a Design

0										x	x	x	x	x	x	x	x	x	x	x	x	
b	x	x	x	x	x	x	x	x	x	1	1	1	1	1	1	1	1	1	1	2	2	2	2	
s	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	w
37	1	2	2	2	2	1	1	2	1	1	1	3	2	3	3	2	1	2	1	3	1	3	1	.
38	2	2	2	1	2	1	2	2	1	1	1	2	2	1	2	3	3	3	1	1	3	2	3	.
39	1	2	2	2	2	2	2	1	2	2	2	3	3	3	1	2	1	3	2	1	3	2	3	.
40	2	1	1	1	1	1	1	2	2	2	1	2	1	2	3	2	1	3	2	1	1	2	1	.

Partial Profiles and Restrictions

Partial-profile designs (Chrzan and Elrod, 1995) are used when there are many attributes but no more than a few of them are allowed to vary at a time. Chrzan and Elrod show an example where respondents must choose between vacuum cleaners that vary along 20 different attributes: Brand, Price, Warranty, Horsepower, and so on. It is difficult for respondents to simultaneously evaluate that many attributes, so it is better if they are only exposed to a few at a time.

Pair-wise Partial Profile Choice Design

Here for example is a partial profile design for 20 two-level factors, with 5 varying at a time, with the factors that are not shown printed with an ordinary missing value.

```
. . . . . . . . . 2 . 2 . 2 . 1 . . 2 .
. . . 2 1 . . . . . . . . 2 . . 2 2 . .
. . 1 . . . . . . . . 1 . . 1 . . . 2 1
. . 2 . 2 . . . . . . 1 . . . . 1 1 . .
1 . . . 2 . . . . . . . . 1 . 2 . . . 1
. . . . . . . 2 . . 1 . . . . 1 . 2 1 .
. . . 1 . . . 1 . . 2 1 . 1 . . . . . .
. . . . . . . . . 1 2 . . . 1 2 1 . . .
. . . . . 2 . . . . . . . . 2 . . 1 2 2
. . . 2 . . . 2 . 2 . . 1 . 2 . . . . .
. . 2 . . 1 . 2 . . . . 2 2 . . . . . .
. 1 2 . . . . . 2 . . . . 1 1 . . . . .
. . 2 . . . 1 1 . . . . . . . 2 . . 2 .
1 1 . . . . 1 . . . 2 1 . . . . . . . .
. 1 1 1 . . . . . . . . . . . 2 . 1 . .
2 . . . 1 2 . . 1 . . . . . . 1 . . . .
. . . . 1 . . . . 2 . . 2 . . . . . 1 2
1 . 1 . . . . 1 2 . . . . . . . 1 . . .
. 2 1 . 2 . . . . . 2 . . . . . . . 2 .
. 1 . . 2 . . 1 . 1 . . . . . . . . . 1
. . . . . 1 . . . . . . . 1 . . 1 2 . 2 .
. . . . 1 1 . . . . 1 2 . . 2 . . . . .
. . . 2 . 1 . . 1 . . 2 . . . . . . 1 .
. . 2 . . . . . 2 . 2 . . . . . . 1 . 1
2 . . 2 . . . . . . 1 . 2 . . . . . 2 .
. . . . . 2 2 . . . . 2 1 . . . . 2 . .
. 2 . 1 . . 1 . . . . . . . . . . 1 . . 1
2 . . . . . . 2 . . . 2 . . . . 2 . . 2
1 2 . . . . . 1 . . . . . . . 1 . . 1 . .
. . 2 2 . . 2 . . . . . . . . . 1 . . . 2
```

```
. . . . . 1 . 1 . 1 2 . . . . . 2
. . . 1 2 . 1 . 2 . . . . . 1 . . . . .
1 . 2 1 . . . . . 1 . . . . . . . . 2 .
2 . . . . 1 . . . 2 . . . . . . . 2 . 2
. 2 . . . . . . 2 1 . 1 . . . 2 . . . .
2 . . . . . 2 . . . . . . 2 2 . . . 1 .
. . . . 1 . 2 2 1 . . . . . . . . . 2 .
. . 1 . . . 1 . . 1 . . . 1 . . . 1 . .
. . . . . . . . . 1 . . . 2 . 2 . 2 2 . .
. . . . . 2 2 . . 2 1 . . . . . 2 . . .
```

A design like this could be used to make a binary choice experiment. For example, the last run has factors 6, 7, 10, 11, and 17 varying. Assume they are all yes-no factors (1 yes, 2 no). Subjects could be offered a choice between these two profiles:

```
x6 = no,   x7 = no,   x10 = no,   x11 = yes,  x17 = no
x6 = yes,  x7 = yes,  x10 = yes,  x11 = no,   x17 = yes
```

The first profile came directly from the design and the second came from shifting the design: yes → no, and no → yes. Partial-profile designs have become very popular among some researchers.

Here is the code that generated and printed the partial profile design.

```
title 'Partial Profiles';

%mktex(3 ** 20, n=41, partial=5, seed=292, maxdesigns=1)

%mktlab(values=. 1 2, nfill=99)

data _null_; set final(firstobs=2); put (x1-x20) (2.); run;
```

A 3^{20} design is requested in 41 runs. The three levels are yes, no, and not shown. Forty-one runs will give us 40 partial profiles and one more run with just all attributes not shown (all ones in the original design before reassigning levels). When we ask for partial profiles, in this case `partial=5`, we are imposing a constraint that the number of 2's and 3's in each run equal 5 and the number of 1's equal 15. This makes the sum of the coded variables constant in each run and hence introduces a linear dependency (the sum of the coded variables is proportional to the intercept). The way we avoid having the linear dependency is by adding this additional row where all attributes are set to the not shown level. The sum of the coded variables for this row will be different than the constant sum for the other rows and hence will eliminate the linear dependency we would otherwise have.

The `%MktLab` macro reassigns the levels 1, 2, 3 to ., 1, 2 where . will mean not shown. Normally the `%MktLab` macro complains about using missing values for levels, because missing values are used in the `key=` data set as fillers when some factors have more levels than others. Encountering missing levels normally indicates an error. We can allow for missing levels by specifying `nfill=99`. Then the macro considers levels of 99 to be invalid, not missing. A DATA step prints the design excluding the constant (all not shown) first row.

This next section of code takes this design and turns it into a partial-profile choice design. It reads each profile in the design, and outputs it. If the level is not missing, the code changes 1 to 2 and 2 to 1 and outputs the new profile. The next step uses the `%ChoicEff` macro to evaluate the design.

We specified `zero=none` for now to see exactly which parameters we can estimate and which ones we cannot. This usage of the `%ChoicEff` macro is similar to what we saw in the food product example on page 260. Our choice design is specified on `data=` and the same data set, with just the `Set` variable kept, is specified on the `init=` option. The number of choice sets, 40 (we drop the constant choice set), number of alternatives, 2, and assumed betas, a vector of zeros, are also specified. Zero internal iterations are requested since we want a design evaluation, not an attempt to improve the design.

```
data des(drop=i);
   Set = _n_;
   set final(firstobs=2);
   array x[20];
   output;
   do i = 1 to 20;
      if n(x[i]) then do; if x[i] = 1 then x[i] = 2; else x[i] = 1; end;
      end;
   output;
   run;

%choiceff(data=des,
          model=class(x1-x20 / zero=none),
          nsets=40, nalts=2,
          beta=zero, init=des(keep=set),
          intiter=0)
```

Here is the last part of the output.

Partial Profiles 39

n	Variable Name	Label	Variance	DF	Standard Error
1	x11	x1 1	0.51210	1	0.71561
2	x12	x1 2	.	0	.
3	x21	x2 1	0.53230	1	0.72959
4	x22	x2 2	.	0	.
5	x31	x3 1	0.37021	1	0.60845
6	x32	x3 2	.	0	.
7	x41	x4 1	0.49818	1	0.70582
8	x42	x4 2	.	0	.
9	x51	x5 1	0.51016	1	0.71425
10	x52	x5 2	.	0	.

11	x61	x6 1	0.52252	1	0.72285
12	x62	x6 2	.	0	.
13	x71	x7 1	0.54267	1	0.73666
14	x72	x7 2	.	0	.
15	x81	x8 1	0.49567	1	0.70404
16	x82	x8 2	.	0	.
17	x91	x9 1	0.51694	1	0.71898
18	x92	x9 2	.	0	.
19	x101	x10 1	0.50651	1	0.71170
20	x102	x10 2	.	0	.
21	x111	x11 1	0.48852	1	0.69894
22	x112	x11 2	.	0	.
23	x121	x12 1	0.50283	1	0.70910
24	x122	x12 2	.	0	.
25	x131	x13 1	0.56443	1	0.75128
26	x132	x13 2	.	0	.
27	x141	x14 1	0.55872	1	0.74747
28	x142	x14 2	.	0	.
29	x151	x15 1	0.49168	1	0.70120
30	x152	x15 2	.	0	.
31	x161	x16 1	0.50164	1	0.70827
32	x162	x16 2	.	0	.
33	x171	x17 1	0.59184	1	0.76931
34	x172	x17 2	.	0	.
35	x181	x18 1	0.45818	1	0.67689
36	x182	x18 2	.	0	.
37	x191	x19 1	0.35190	1	0.59321
38	x192	x19 2	.	0	.
39	x201	x20 1	0.46177	1	0.67954
40	x202	x20 2	.	0	.
				==	
				20	

We see that one parameter is estimable for each factor and that is the parameter for the 1 or yes level. The `%ChoicEff` macro prints a list of all redundant variables.

```
Redundant Variables:

x12 x22 x32 x42 x52 x62 x72 x82 x92 x102 x112 x122 x132 x142 x152 x162 x172 x182
x192 x202
```

We can cut this list into our program and drop those terms.

```
%choiceff(data=des,
          model=class(x1-x20 / zero=none),
          nsets=40, nalts=2,
          beta=zero, init=des(keep=set),
          intiter=0, drop=x12 x22 x32 x42 x52 x62 x72 x82 x92 x102
          x112 x122 x132 x142 x152 x162 x172 x182 x192 x202);
```

Here is the last part of the output.

Partial Profiles

n	Variable Name	Label	Variance	DF	Standard Error
1	x11	x1 1	0.51210	1	0.71561
2	x21	x2 1	0.53230	1	0.72959
3	x31	x3 1	0.37021	1	0.60845
4	x41	x4 1	0.49818	1	0.70582
5	x51	x5 1	0.51016	1	0.71425
6	x61	x6 1	0.52252	1	0.72285
7	x71	x7 1	0.54267	1	0.73666
8	x81	x8 1	0.49567	1	0.70404
9	x91	x9 1	0.51694	1	0.71898
10	x101	x10 1	0.50651	1	0.71170
11	x111	x11 1	0.48852	1	0.69894
12	x121	x12 1	0.50283	1	0.70910
13	x131	x13 1	0.56443	1	0.75128
14	x141	x14 1	0.55872	1	0.74747
15	x151	x15 1	0.49168	1	0.70120
16	x161	x16 1	0.50164	1	0.70827
17	x171	x17 1	0.59184	1	0.76931
18	x181	x18 1	0.45818	1	0.67689
19	x191	x19 1	0.35190	1	0.59321
20	x201	x20 1	0.46177	1	0.67954
				==	
				20	

Linear Partial Profile Design

Here is another example. Say you would like to make a design in 36 runs with 12 three-level factors, but you want only four of them to be considered at a time. You would need to create four-level factors with one of the levels meaning not shown. You also need to ask for a design in 37 runs, because with partial profiles, one run must be all-constant. Here is a partial profile request with the `%MktEx` macro using the `partial=` option.

```
title 'Partial Profiles';

%mktex(4 ** 12, n=37, partial=4, seed=462, maxdesigns=1)
%mktlab(values=. 1 2 3, nfill=99)

proc print; run;
```

The iteration history will proceed like before, so we will not discuss it. Here is the final *D*-efficiency.

Partial Profiles

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	49.4048	22.4346	100.0000	1.0000

With partial-profile designs, *D*-efficiency will typically be much less than we are accustomed to seeing with other types of designs. Here is the design.

Partial Profiles

Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
1	.	.	.	.	.	.	.	.	.	.	.	.
2	.	.	.	.	.	1	.	1	1	.	1	.
3	3	1	.	2	.	.	.	2	.	.	.	.
4	.	.	.	3	.	2	.	.	3	.	3	.
5	2	3	.	.	.	.	.	.	.	1	3	.
6	1	2	.	3	3	.	.	.	.	.	.	.
7	.	3	.	.	.	.	2	3	3	.	.	.
8	.	2	.	2	.	.	3	.	.	3	.	.
9	.	.	1	.	.	3	.	1	.	2	.	.
10	.	.	.	2	2	1	.	.	.	.	2	.
11	1	.	2	.	.	.	3	.	.	.	2	.
12	.	.	3	.	1	.	.	.	3	1	.	.
13	3	2	.	.	.	.	.	.	.	.	2	3
14	.	.	2	2	3	.	.	.	.	.	.	3
15	.	.	1	.	2	.	.	.	.	2	1	.
16	2	.	3	.	.	2	.	3	.	.	.	.
17	.	2	2	.	.	1	1	.	.	.	.	.
18	1	.	.	.	.	1	.	.	.	3	.	3
19	.	.	.	.	2	.	1	1	.	.	.	1
20	.	3	.	.	1	2	.	.	.	.	.	2
21	3	.	.	.	3	1	3	.	.	.	.	.
22	.	3	3	1	.	.	.	.	2	.	.	.
23	3	.	2	3	.	.	.	.	.	3	.	.
24	.	.	.	.	.	3	1	.	.	.	1	1
25	.	.	1	.	.	.	1	2	.	.	2	.
26	2	.	.	1	.	.	.	.	3	.	.	2
27	.	.	.	1	.	2	2	.	.	1	.	.

28	.	.	.	1	1	.	.	3	.	.	3	.
29	.	1	.	.	3	.	.	.	.	3	2	.
30	.	.	.	.	.	1	.	2	.	2	.	1
31	.	.	.	2	.	.	1	.	1	2	.	.
32	.	1	1	.	.	.	.	.	1	.	.	1
33	.	.	3	.	.	.	2	.	.	.	3	2
34	2	.	.	.	1	.	2	.	2	.	.	.
35	.	1	.	3	.	.	3	.	.	.	.	3
36	.	.	.	.	.	.	.	3	2	1	.	2
37	.	.	.	.	2	3	.	2	1	.	.	.

Notice that the first run is constant. For all other runs, exactly four factors vary and have levels not missing.

Choice from Triples; Partial Profiles Constructed Using Restrictions

The approach we just saw, constructing partial profiles using the `partial=` option, would be fine for a full-profile conjoint study or a pair-wise choice study with level shifts. However, it would not be good for a more general choice experiment with more alternatives. For a choice experiment, you would have to have full-profile restrictions on each alternative, and you must have the same attributes varying in each choice set. There is currently no automatic way to request this in the `%MktEx` macro, so you have to program the restrictions yourself. To specify restrictions for choice designs, you need to take into consideration the number of attributes that may vary within each alternative, which ones, and which attributes go with which alternatives. Fortunately, that is not too difficult. See page 231 for another example of restrictions.

In this section, we will construct a partial-profile design for a purely generic study (unbranded), with ten attributes and three alternatives. Each attribute will have three levels, and each alternative will be a bundle of attributes. Partial-profile designs have the advantage that subjects do not have to consider all attributes at once. However, this is also a bit of a disadvantage as well in the sense that the subjects must constantly shift from considering one set of attributes to considering a different set. For this reason, it can be helpful to get more information out of each choice, and having more than two alternatives per choice set accomplishes this.

This example will have several parts. As we mentioned in the chair study, we will usually not directly use the `%MktEx` macro to generate designs for generic studies. Instead, we will use the `%MktEx` macro to generate a candidate set of partial profile choice sets. Next, the design will be checked and turned into a candidate set of generic choice sets. Next, the `%MktDups` macro will be called to ensure there are no duplicate choice sets. Finally, the `%ChoiceEff` macro will be used to create an efficient generic partial-profile choice design.

Before we go into any more detail on making this design, let's skip ahead and look at a couple of potential choice sets so it will be clear what we are trying to accomplish and why. Here are two potential choice sets still in linear design format.

2 2 1 3 1 2 1 2 2 2	2 1 3 2 1 1 2 1 2 2	2 3 2 1 1 3 3 3 2 2
2 2 1 3 2 2 1 3 2 3	3 2 2 3 1 2 1 1 1 1	1 2 3 3 3 2 1 2 3 2

Here are the same two potential choice sets, but now arrayed in choice design format.

Partial Profiles

Set	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	2	2	1	3	1	2	1	2	2	2
	2	1	3	2	1	1	2	1	2	2
	2	3	2	1	1	3	3	3	2	2
2	2	2	1	3	2	2	1	3	2	3
	3	2	2	3	1	2	1	1	1	1
	1	2	3	3	3	2	1	2	3	2

Each choice set has 10 three-level factors and three alternatives. Four attributes are constant in each choice set: `x1`, `x5`, `x9`, and `x10` in the first choice set, and `x2`, `x4`, `x6`, and `x7` in the second choice set. We do not need an all-constant choice set like we saw in our earlier partial-profile designs, nor do we need an extra level for not varying. In this approach, we will simply construct choice sets for four constant attributes (they may be constant at 1, 2, or 3) and six varying attributes (with levels: 1, 2, and 3). Respondents will be given a choice task along the lines of "Given a set of products that differ on these attributes but are identical in all other respects, which one would you choose?". They would then be shown a list of differences.

Here is the code for making the candidate set.

```
title 'Partial Profiles';

%macro partprof;
   sum = 0;
   do k = 1 to 10;
      sum = sum + (x[k] = x[k+10]   &   x[k] = x[k+20]);
      end;
   bad = abs(sum - 4);
   %mend;

%mktex(3 ** 30, n=198, optiter=0, tabiter=0, maxtime=0, order=random,
       out=sasuser.cand, restrictions=partprof, seed=382)
```

We requested a design in 198 runs with 30 three-level factors. The 198 was chosen arbitrarily as a number divisible by $3 \times 3 = 9$ that would give us approximately 200 candidate sets. The first ten factors, `x1-x10`, will make the first alternative, the next ten, `x11-x20`, will make the second alternative, and the last ten, `x21-x30`, will make the third alternative. We will want six attributes to be nonconstant at a time. The `PartProf` macro will count the number of constant attributes: `x1 = x11 = x21`, `x2 = x12 = x22`, ..., and `x10 = x20 = x30`. If the number of constant attributes is four, our choice set conforms. If it is more or less than four, our choice set is in violation of the restrictions. The badness is the absolute difference between four and the number of constant attributes.

We specified `order=random`, which specifies that the columns are to be looped over in a random order in the coordinate exchange algorithm. When `partial=` is specified, as it was in the previous partial-profile examples, `order=random` is the default. Whenever you are imposing partial-profile restrictions without using the `partial=` option, you should specify `order=random`. Without `order=random`, `%MktEx` will tend to put the nonconstant levels close together in each row.

Our goal in this step is to make a candidate set of potential partial-profile choice sets, not to make a final experimental design. Ideally, it would be nice if we had more than random candidates – it would be nice if our candidate generation code at least made some attempt to ensure that our attributes are approximately orthogonal and balanced across attributes both between and within alternatives. This is a big problem (30 factors and 198 runs) with restrictions, so `%MktEx` macro will run slowly by default. It is not critical that we allow the macro to spend a great deal of time optimizing linear model *D*-efficiency. For this reason, we use some of the more esoteric number-of-iterations options. We specify `optiter=0`, which specifies no OPTEX iterations, since with large partial profile studies, we will never have a good candidate set for PROC OPTEX to search. We also specify `tabiter=0` since a tabled initial design will be horrible for this problem. We specified the `maxtime=0` option so that the macro will just create two candidate designs using the coordinate-exchange algorithm with a random initialization and make one attempt to refine the best one.

Partial Profiles

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	85.1531		Ran,Mut,Ann
1	192 1	93.4101	93.4101	Conforms
1	192 14	93.4135	93.4135	
1	192 25	93.4181	93.4181	
1	192 11	93.4181	93.4181	
.				
.				
.				
1	105 12	96.5419	96.5419	
1	110 17	96.5421	96.5421	
1	182 22	96.5422	96.5422	
1	4 17	96.5438	96.5438	
1	End	96.5433		

NOTE: Quitting the algorithm search step after 2.46 minutes and 1 designs.

.
.
.

Partial Profiles

Design Refinement History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	96.5438	96.5438	Ini
1	Start	96.0296		Pre,Mut,Ann
1	163 1	95.8964		Conforms
1	77 25	96.5438	96.5438	
1	156 9	96.5442	96.5442	
1	2 17	96.5442	96.5442	
1	7 23	96.5451	96.5451	
1	24 2	96.5453	96.5453	
1	54 7	96.5473	96.5473	
1	131 19	96.5475	96.5475	
1	172 7	96.5483	96.5483	
1	180 5	96.5484	96.5484	
1	24 2	96.5493	96.5493	
1	180 5	96.5501	96.5501	
1	119 29	96.5501	96.5501	
1	End	96.5501		

NOTE: Quitting the refinement step after 1.90 minutes and 1 designs.

Partial Profiles

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	96.5501	93.4564	95.5808	0.5551

The macro finds a design that conforms to the restrictions (shown by the `Conforms` note) that is 93.41% *D*-efficient, and the final design is 96.55% *D*-efficient. This step took 7 minutes.

Here is the rest of the code for making the partial-profile choice design.

```
%mktkey(x1-x30)

data key;
   input (x1-x10) ($);
   datalines;
 x1  x2  x3  x4  x5  x6  x7  x8  x9 x10
x11 x12 x13 x14 x15 x16 x17 x18 x19 x20
x21 x22 x23 x24 x25 x26 x27 x28 x29 x30
;

%mktroll(design=sasuser.cand, key=key, out=rolled)

%mktdups(generic, data=rolled, out=nodups, factors=x1-x10, nalts=3);

proc print data=nodups(obs=9); id set; by set; run;

%choiceff(data=nodups, model=class(x1-x10), seed=495,
          iter=10, nsets=27, nalts=3, options=nodups, beta=zero)

proc print data=best; id set; by notsorted set; var x1-x10; run;
```

The %MktKey macro is run to generate the full list of names in the range `x1 - x30` for pasting into the next step. A KEY data set is created and the %MktRoll macro is run to create a generic choice design from the linear candidate design.

The next step runs the %MktDups macro, which we have not used in previous examples. The %MktDups macro can check a design to see if there are any duplicate runs and output just the unique sets. For a generic study like this, it can also check to make sure there are no duplicate choice sets taking into account the fact that two choice sets can be duplicates even if the alternatives are not in the same order. The %MktDups step names in a positional parameter the type of design as a `generic` choice design. It names the input data set and the output data set that will contain the design with any duplicates removed. It names the factors in the choice design `x1-x10` and the number of alternatives. The result is a data set called NODUPS. Here are the first 3 candidate choice sets.

Partial Profiles

Set	_Alt_	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	1	1	1	1	1	1	1	2	1	3	2
	2	3	3	2	2	1	3	3	1	3	2
	3	2	2	3	3	1	2	1	1	3	2
2	1	1	1	1	3	1	3	1	3	3	3
	2	1	2	2	1	1	1	2	3	3	1
	3	1	3	3	2	1	2	3	3	3	2
3	1	1	1	1	3	2	3	1	2	3	1
	2	1	1	1	2	2	1	3	3	1	3
	3	1	1	1	1	2	2	2	1	2	2

The `%ChoicEff` macro is called to search for an efficient choice design. The model specification `class(x1-x10)` specifies a generic model with 10 attributes. The option `iter=10` specifies more than the default number of iterations (the default is 2 designs). We ask for a design with 27 sets and 3 alternatives. Furthermore, we ask for no duplicate choice sets and specify an assumed beta vector of zero. Here are some of the results from the `%ChoicEff` macro.

Partial Profiles

Design	Iteration	D-Efficiency	D-Error
1	0	2.368953	0.422127
	1	2.904996	0.344235
	2	2.912352	0.343365

.
.
.

Design	Iteration	D-Efficiency	D-Error
10	0	2.489370	0.401708
	1	2.873127	0.348053
	2	2.918003	0.342700
	3	2.936918	0.340493
	4	2.940485	0.340080

Partial Profiles

Final Results

Design	10
Choice Sets	27
Alternatives	3
D-Efficiency	2.940485
D-Error	0.340080

Partial Profiles

n	Variable Name	Label	Variance	DF	Standard Error
1	x11	x1 1	0.39481	1	0.62834
2	x12	x1 2	0.37025	1	0.60848
3	x21	x2 1	0.38889	1	0.62361
4	x22	x2 2	0.39423	1	0.62788
5	x31	x3 1	0.42249	1	0.64999
6	x32	x3 2	0.44601	1	0.66784
7	x41	x4 1	0.40120	1	0.63340
8	x42	x4 2	0.44805	1	0.66937
9	x51	x5 1	0.36874	1	0.60724
10	x52	x5 2	0.37101	1	0.60910

11	x61	x6 1	0.37513	1	0.61248
12	x62	x6 2	0.36982	1	0.60813
13	x71	x7 1	0.48395	1	0.69566
14	x72	x7 2	0.46304	1	0.68047
15	x81	x8 1	0.41310	1	0.64273
16	x82	x8 2	0.41780	1	0.64637
17	x91	x9 1	0.42968	1	0.65550
18	x92	x9 2	0.43238	1	0.65755
19	x101	x10 1	0.48179	1	0.69411
20	x102	x10 2	0.42340	1	0.65069
				==	
				20	

Here is part of the design.

Partial Profiles

Set	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
30	1	2	2	2	1	1	1	1	3	3
	1	2	3	1	2	1	2	2	3	3
	1	3	1	3	3	1	3	3	3	3
70	2	1	1	1	1	2	1	3	1	2
	3	2	1	3	2	3	1	1	1	2
	1	3	1	2	3	1	1	2	1	2
.										
.										
.										
104	2	2	2	3	2	2	1	2	3	1
	2	3	2	1	1	2	3	3	2	1
	2	1	2	2	3	2	2	1	1	1

The design has 27 choice sets. The choice set numbers shown in this output correspond to the *original* set numbers in the candidate design not the choice set numbers in the final design.

Multinomial Logit Models

Ying So

Warren F. Kuhfeld

Abstract

Multinomial logit models are used to model relationships between a polytomous response variable and a set of regressor variables. The term "multinomial logit model" includes, in a broad sense, a variety of models. The cumulative logit model is used when the response of an individual unit is restricted to one of a finite number of ordinal values. Generalized logit and conditional logit models are used to model consumer choices. This article focuses on the statistical techniques for analyzing discrete choice data and discusses fitting these models using SAS/STAT software.*

Introduction

Multinomial logit models are used to model relationships between a polytomous response variable and a set of regressor variables. These polytomous response models can be classified into two distinct types, depending on whether the response variable has an ordered or unordered structure.

In an ordered model, the response Y of an individual unit is restricted to one of m ordered values. For example, the severity of a medical condition may be: none, mild, and severe. The cumulative logit model assumes that the ordinal nature of the observed response is due to methodological limitations in collecting the data that results in lumping together values of an otherwise continuous response variable (McKelvey and Zavoina 1975). Suppose Y takes values $y_1, y_2, \ldots, y_m$ on some scale, where $y_1 < y_2 < \ldots < y_m$. It is assumed that the observable variable is a categorized version of a continuous latent variable U such that

$$Y = y_i \Leftrightarrow \alpha_{i-1} < U \leq \alpha_i, i = 1, \ldots, m$$

where $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_m = \infty$. It is further assumed that the latent variable U is determined by the explanatory variable vector $\mathbf{x}$ in the linear form $U = -\boldsymbol{\beta}'\mathbf{x} + \epsilon$, where $\boldsymbol{\beta}$ is a vector of regression coefficients and ϵ is a random variable with a distribution function F. It follows that

$$\Pr\{Y \leq y_i | \mathbf{x}\} = F(\alpha_i + \boldsymbol{\beta}'\mathbf{x})$$

If F is the logistic distribution function, the cumulative model is also known as the proportional odds model. You can use PROC LOGISTIC or PROC PROBIT directly to fit the cumulative logit models. Although the cumulative model is the most widely used model for ordinal response data, other useful models include the adjacent-categories logit model and the continuation-ratio model (Agresti 1990).

*This chapter was presented at SUGI 20 by Ying So and can also be found in the SUGI 20 proceedings. Copies of this article (TS-689F) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market .

In an unordered model, the polytomous response variable does not have an ordered structure. Two classes of models, the generalized logit models and the conditional logit models, can be used with nominal response data. The generalized logit model consists of a combination of several binary logits estimated simultaneously. For example, the response variable of interest is the occurrence or nonoccurrence of infection after a Caesarean section with two types of (I,II) infection. Two binary logits are considered: one for type I infection versus no infection and the other for type II infection versus no infection. The conditional logit model has been used in biomedical research to estimate relative risks in matched case-control studies. The nuisance parameters that correspond to the matched sets in an unconditional analysis are eliminated by using a conditional likelihood that contains only the relative risk parameters (Breslow and Day 1980). The conditional logit model was also introduced by McFadden (1974) in the context of econometrics.

In studying consumer behavior, an individual is presented with a set of alternatives and asked to choose the most preferred alternative. Both the generalized logit and conditional logit models are used in the analysis of discrete choice data. In a conditional logit model, a choice among alternatives is treated as a function of the characteristics of the alternatives, whereas in a generalized logit model, the choice is a function of the characteristics of the individual making the choice. In many situations, a mixed model that includes both the characteristics of the alternatives and the individual is needed for investigating consumer choice.

Consider an example of travel demand. People are asked to choose between travel by auto, plane or public transit (bus or train). The following SAS statements create the data set TRAVEL. The variables `AutoTime`, `PlanTime`, and `TranTime` represent the total travel time required to get to a destination by using auto, plane, or transit, respectively. The variable `Age` represents the age of the individual being surveyed, and the variable `Chosen` contains the individual's choice of travel mode.

```
data travel;
   input AutoTime PlanTime TranTime Age Chosen $;
   datalines;
10.0      4.5     10.5    32  Plane
 5.5      4.0      7.5    13  Auto
 4.5      6.0      5.5    41  Transit
 3.5      2.0      5.0    41  Transit
 1.5      4.5      4.0    47  Auto
10.5      3.0     10.5    24  Plane
 7.0      3.0      9.0    27  Auto
 9.0      3.5      9.0    21  Plane
 4.0      5.0      5.5    23  Auto
22.0      4.5     22.5    30  Plane
 7.5      5.5     10.0    58  Plane
11.5      3.5     11.5    36  Transit
 3.5      4.5      4.5    43  Auto
12.0      3.0     11.0    33  Plane
18.0      5.5     20.0    30  Plane
23.0      5.5     21.5    28  Plane
 4.0      3.0      4.5    44  Plane
 5.0      2.5      7.0    37  Transit
 3.5      2.0      7.0    45  Auto
12.5      3.5     15.5    35  Plane
 1.5      4.0      2.0    22  Auto
;
```

In this example, `AutoTime`, `PlanTime`, and `TranTime` are alternative-specific variables, whereas `Age` is a characteristic of the individual. You use a generalized logit model to investigate the relationship between the choice of transportation and `Age`, and you use a conditional logit model to investigate how travel time affects the choice. To study how the choice depends on both the travel time and age of the individual, you need to use a mixed model that incorporates both types of variables.

A survey of the literature reveals a confusion in the terminology for the nominal response models. The term "multinomial logit model" is often used to describe the generalized logit model. The mixed logit is sometimes referred to as the multinomial logit model in which the generalized logit and the conditional logit models are special cases.

The following sections describe discrete choice models, illustrate how to use SAS/STAT software to fit these models, and discuss cross-alternative effects.

Modeling Discrete Choice Data

Consider an individual choosing among m alternatives in a choice set. Let Π_{jk} denote the probability that individual j chooses alternative k, let $\mathbf{X}_j$ represent the characteristics of individual j, and let $\mathbf{Z}_{jk}$ be the characteristics of the kth alternative for individual j. For example, $\mathbf{X}_j$ may be an age and each $\mathbf{Z}_{jk}$ a travel time.

The generalized logit model focuses on the individual as the unit of analysis and uses individual characteristics as explanatory variables. The explanatory variables, being characteristics of an individual, are constant over the alternatives. For example, for each of the m travel modes, $\mathbf{X}_j = (1\ age)'$, and for the first subject, $\mathbf{X}_1 = (1\ 32)'$. The probability that individual j chooses alternative k is

$$\Pi_{jk} = \frac{\exp(\boldsymbol{\beta}_k'\mathbf{X}_j)}{\sum_{l=1}^{m}\exp(\boldsymbol{\beta}_l'\mathbf{X}_j)} = \frac{1}{\sum_{l=1}^{m}\exp[(\boldsymbol{\beta}_l - \boldsymbol{\beta}_k)'\mathbf{X}_j)]}$$

$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ are m vectors of unknown regression parameters (each of which is different, even though $\mathbf{X}_j$ is constant across alternatives). Since $\sum_{k=1}^{m}\Pi_{jk} = 1$, the m sets of parameters are not unique. By setting the last set of coefficients to null (that is, $\boldsymbol{\beta}_m = \mathbf{0}$), the coefficients $\boldsymbol{\beta}_k$ represent the effects of the $\mathbf{X}$ variables on the probability of choosing the kth alternative over the last alternative. In fitting such a model, you estimate $m - 1$ sets of regression coefficients.

In the conditional logit model, the explanatory variables $\mathbf{Z}$ assume different values for each alternative and the impact of a unit of $\mathbf{Z}$ is assumed to be constant across alternatives. For example, for each of the m travel modes, $\mathbf{Z}_{jk} = (time)'$, and for the first subject, $\mathbf{Z}_{11} = (10)'$, $\mathbf{Z}_{12} = (4.5)'$, and $\mathbf{Z}_{13} = (10.5)'$. The probability that the individual j chooses alternative k is

$$\Pi_{jk} = \frac{\exp(\boldsymbol{\theta}'\mathbf{Z}_{jk})}{\sum_{l=1}^{m}\exp(\boldsymbol{\theta}'\mathbf{Z}_{jl})} = \frac{1}{\sum_{l=1}^{m}\exp[\boldsymbol{\theta}'(\mathbf{Z}_{jl} - \mathbf{Z}_{jk})]}$$

$\boldsymbol{\theta}$ is a single vector of regression coefficients. The impact of a variable on the choice probabilities derives from the difference of its values across the alternatives.

For the mixed logit model that includes both characteristics of the individual and the alternatives, the choice probabilities are

$$\Pi_{jk} = \frac{\exp(\boldsymbol{\beta}_k'\mathbf{X}_j + \boldsymbol{\theta}'\mathbf{Z}_{jk})}{\sum_{l=1}^{m}\exp(\boldsymbol{\beta}_l'\mathbf{X}_j + \boldsymbol{\theta}'\mathbf{Z}_{jl})}$$

$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{m-1}$ and $\boldsymbol{\beta}_m \equiv \mathbf{0}$ are the alternative-specific coefficients, and $\boldsymbol{\theta}$ is the set of global coefficients.

Fitting Discrete Choice Models

The CATMOD procedure in SAS/STAT software directly fits the generalized logit model. SAS/STAT software does not yet have a procedure that is specially designed to fit the conditional or mixed logit models. However, with some preliminary data processing, you can use the PHREG procedure to fit these models.

The PHREG procedure fits the Cox proportional hazards model to survival data (refer to SAS/STAT documentation). The partial likelihood of Breslow has the same form as the likelihood in a conditional logit model.

Let z_l denote the vector of explanatory variables for individual l. Let $t_1 < t_2 < \ldots < t_k$ denote k distinct ordered event times. Let d_i denote the number of failures at t_i. Let s_i be the sum of the vectors z_l for those individuals that fail at t_i, and let $\mathcal{R}_i$ denote the set of indices for those who are at risk just before t_i.

The Breslow (partial) likelihood is

$$L_B(\boldsymbol{\theta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\theta}' s_i)}{[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\theta}' z_l)]^{d_i}}$$

In a stratified analysis, the partial likelihood is the product of the partial likelihood for each individual stratum. For example, in a study of the time to first infection from a surgery, the variables of a patient consist of `Time` (time from surgery to the first infection), `Status` (an indicator of whether the observation time is censored, with value 2 identifying a censored time), `Z1` and `Z2` (explanatory variables thought to be related to the time to infection), and `Grp` (a variable identifying the stratum to which the observation belongs). The specification in PROC PHREG for fitting the Cox model using the Breslow likelihood is as follows:

```
proc phreg;
   model time*status(2) = z1 z2 / ties=breslow;
   strata grp;
   run;
```

To cast the likelihood of the conditional logit model in the form of the Breslow likelihood, consider m artificial observed times for each individual who chooses one of m alternatives. The kth alternative is chosen at time 1; the choices of all other alternatives (second choice, third choice, ...) are not observed and would have been chosen at some later time. So a choice variable is coded with an observed time value of 1 for the chosen alternative and a larger value, 2, for all unchosen (unobserved or censored alternatives). For each individual, there is exactly one event time (1) and $m - 1$ nonevent times, and the risk set just prior to this event time consists of all the m alternatives. For individual j with alternative-specific characteristics $\mathbf{Z}_{jl}$, the Breslow likelihood is then

$$L_B(\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}' \mathbf{Z}_{jk})}{\sum_{l=1}^{m} \exp(\boldsymbol{\theta}' \mathbf{Z}_{jl})}$$

This is precisely the probability that individual j chooses alternative k in a conditional logit model. By stratifying on individuals, you get the likelihood of the conditional logit model. Note that the observed time values of 1 and 2 are chosen for convenience; however, the censored times have to be larger than the event time to form the correct risk set.

Before you invoke PROC PHREG to fit the conditional logit, you must arrange your data in such a way that there is a survival time for each individual-alternative. In the example of travel demand, let `Subject` identify the individuals, let `TravTime` represent the travel time for each mode of transportation, and let `Choice` have a value 1 if the alternative is chosen and 2 otherwise. The `Choice` variable is used as the artificial time variable as well as a censoring variable in PROC PHREG. The following SAS statements reshape the data set TRAVEL into data set CHOICE and display the first nine observations:

```
data choice(keep=subject mode travtime choice);
   array times[3] autotime plantime trantime;
   array allmodes[3] $ _temporary_ ('Auto' 'Plane' 'Transit');
   set travel;
   Subject = _n_;
   do i = 1 to 3;
      Mode = allmodes[i];
      TravTime = times[i];
      Choice = 2 - (chosen eq mode);
      output;
   end;
   run;

proc print data=choice(obs=9);
   run;
```

Obs	Subject	Mode	Trav Time	Choice
1	1	Auto	10.0	2
2	1	Plane	4.5	1
3	1	Transit	10.5	2
4	2	Auto	5.5	1
5	2	Plane	4.0	2
6	2	Transit	7.5	2
7	3	Auto	4.5	2
8	3	Plane	6.0	2
9	3	Transit	5.5	1

Notice that each observation in TRAVEL corresponds to a block of three observations in CHOICE, exactly one of which is chosen.

The following SAS statements invoke PROC PHREG to fit the conditional logit model. The Breslow likelihood is requested by specifying `ties=breslow`. `Choice` is the artificial time variable, and a value of 2 identifies censored times. `Subject` is used as a stratification variable.

```
proc phreg data=choice;
   model choice*choice(2) = travtime / ties=breslow;
   strata subject;
   title 'Conditional Logit Model Using PHREG';
   run;
```

Conditional Logit Model Using PHREG

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
TravTime	1	-0.26549	0.10215	6.7551	0.0093	0.767

To study the relationship between the choice of transportation and the age of people making the choice, the analysis is based on the generalized logit model. You can use PROC CATMOD directly to fit the generalized logit model (refer to *SAS/STAT User's Guide, Vol. 1*). In the following invocation of PROC CATMOD, `Chosen` is the response variable and `Age` is the explanatory variable:

```
proc catmod data=travel;
   direct age;
   model chosen=age;
   title 'Multinomial Logit Model Using Catmod';
   run;
```

Response Profiles

Response	Chosen
1	Auto
2	Plane
3	Transit

Analysis of Maximum Likelihood Estimates

Parameter	Function Number	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	3.0449	2.4268	1.57	0.2096
	2	2.7212	2.2929	1.41	0.2353
Age	1	-0.0710	0.0652	1.19	0.2762
	2	-0.0500	0.0596	0.70	0.4013

Note that there are two intercept coefficients and two slope coefficients for `Age`. The first `Intercept` and the first `Age` coefficients correspond to the effect on the probability of choosing auto over transit, and the second intercept and second age coefficients correspond to the effect of choosing plane over transit.

Let $\mathbf{X}_j$ be a $(p+1)$-vector representing the characteristics of individual j. The generalized logit model can be cast in the framework of a conditional model by defining the global parameter vector $\boldsymbol{\theta}$ and the alternative-specific regressor variables $\mathbf{Z}_{jk}$ as follows:

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_{m-1} \end{bmatrix} \quad \mathbf{Z}_{j1} = \begin{bmatrix} \mathbf{X}_j \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \quad \mathbf{Z}_{j2} = \begin{bmatrix} \mathbf{0} \\ \mathbf{X}_j \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \quad \dots \quad \mathbf{Z}_{j,m-1} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{X}_j \end{bmatrix} \quad \mathbf{Z}_{jm} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

where the $\mathbf{0}$ is a $(p+1)$-vector of zeros. The probability that individual j chooses alternative k for the generalized logit model is put in the form that corresponds to a conditional logit model as follows:

$$\begin{aligned} \Pi_{jk} &= \frac{\exp(\boldsymbol{\beta}_k' \mathbf{X}_j)}{\sum_{l=1}^{m} \exp(\boldsymbol{\beta}_l' \mathbf{X}_j)} \\ &= \frac{\exp(\boldsymbol{\theta}' \mathbf{Z}_{jk})}{\sum_{l=1}^{m} \exp(\boldsymbol{\theta}' \mathbf{Z}_{jl})} \end{aligned}$$

Here, the vector $\mathbf{X}_j$ representing the characteristics of individual j includes the element 1 for the intercept parameter (provided that the intercept parameters are to be included in the model).

By casting the generalized logit model into a conditional logit model, you can then use PROC PHREG to analyze the generalized logit model. In the example of travel demand, the alternative-specific variables `Auto`, `Plane`, `AgeAuto`, and `AgePlane` are created from the individual characteristic variable `Age`. The following SAS statements reshape the data set TRAVEL into data set CHOICE2 and display the first nine observations:

```
data choice2;
   array times[3] autotime plantime trantime;
   array allmodes[3] $ _temporary_ ('Auto' 'Plane' 'Transit');
   set travel;
   Subject = _n_;
   do i = 1 to 3;
      Mode = allmodes[i];
      TravTime = times[i];
      Choice = 2 - (chosen eq mode);
      Auto = (i eq 1);
      Plane = (i eq 2);
      AgeAuto = auto * age;
      AgePlane = plane * age;
      output;
   end;
   keep subject mode travtime choice auto plane ageauto ageplane;
   run;
proc print data=choice2(obs=9);
   run;
```

Obs	Subject	Mode	Trav Time	Choice	Auto	Plane	Age Auto	Age Plane
1	1	Auto	10.0	2	1	0	32	0
2	1	Plane	4.5	1	0	1	0	32
3	1	Transit	10.5	2	0	0	0	0
4	2	Auto	5.5	1	1	0	13	0
5	2	Plane	4.0	2	0	1	0	13
6	2	Transit	7.5	2	0	0	0	0
7	3	Auto	4.5	2	1	0	41	0
8	3	Plane	6.0	2	0	1	0	41
9	3	Transit	5.5	1	0	0	0	0

The following SAS statements invoke PROC PHREG to fit the generalized logit model:

```
proc phreg data=choice2;
   model choice*choice(2) = auto plane ageauto ageplane /
         ties=breslow;
   strata subject;
   title 'Generalized Logit Model Using PHREG';
   run;
```

Generalized Logit Model Using PHREG

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Auto	1	3.04494	2.42682	1.5743	0.2096	21.009
Plane	1	2.72121	2.29289	1.4085	0.2353	15.199
AgeAuto	1	-0.07097	0.06517	1.1859	0.2762	0.931
AgePlane	1	-0.05000	0.05958	0.7045	0.4013	0.951

By transforming individual characteristics into alternative-specific variables, the mixed logit model can be analyzed as a conditional logit model.

Analyzing the travel demand data for the effects of both travel time and age of individual requires the same data set as the generalized logit model, only now the `TravTime` variable will be used as well. The following SAS statements use PROC PHREG to fit the mixed logit model:

```
proc phreg data=choice2;
   model choice*choice(2) = auto plane ageauto ageplane travtime /
         ties=breslow;
   strata subject;
   title 'Mixed Logit Model Using PHREG';
   run;
```

Mixed Logit Model Using PHREG

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Auto	1	2.50069	2.39585	1.0894	0.2966	12.191
Plane	1	-2.77912	3.52929	0.6201	0.4310	0.062
AgeAuto	1	-0.07826	0.06332	1.5274	0.2165	0.925
AgePlane	1	0.01695	0.07439	0.0519	0.8198	1.017
TravTime	1	-0.60845	0.27126	5.0315	0.0249	0.544

A special case of the mixed logit model is the conditional logit model with alternative-specific constants. Each alternative in the model can be represented by its own intercept, which captures the unmeasured desirability of the alternative.

```
proc phreg data=choice2;
   model choice*choice(2) = auto plane travtime / ties=breslow;
   strata subject;
   title 'Conditional Logit Model with Alternative Specific Constants';
   run;
```

Conditional Logit Model with Alternative Specific Constants

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Auto	1	-0.11966	0.70820	0.0285	0.8658	0.887
Plane	1	-1.63145	1.24251	1.7241	0.1892	0.196
TravTime	1	-0.48665	0.20725	5.5139	0.0189	0.615

With transit as the reference mode, the intercept for auto, which is negative, may reflect the inconvenience of having to drive over traveling by bus/train, and the intercept for plane may reflect the high expense of traveling by plane over bus/train.

Cross-Alternative Effects

Discrete choice models are often derived from the principle of maximum random utility. It is assumed that an unobserved utility V_k is associated with the kth alternative, and the response function Y is

determined by

$$Y = k \Leftrightarrow V_k = \max\{V_l, 1 \le l \le m\}$$

Both the generalized logit and the conditional logit models are based on the assumption that $V_1, \ldots, V_m$ are independently distributed and each follows an extreme maxima value distribution (Hoffman and Duncan, 1988). An important property of such models is Independence from Irrelevant Alternatives (IIA); that is, the ratio of the choice probabilities for any two alternatives for a particular observation is not influenced systematically by any other alternatives. IIA can be tested by fitting a model that contains all the cross-alternative effects and examining the significance of these effects. The cross-alternative effects pick up a variety of IIA violations and other sources of error in the model. (See pages 213, 219, 228, and 358 for other discussions of IIA.)

In the example of travel demand, there may be separate effects for the three travel modes and travel times. In addition, there may be cross-alternative effects for travel times. Not all the effects are estimable, only two of the three intercepts and three of the six cross-alternative effects can be estimated. The following SAS statements create the design variables for all the cross-alternative effects and display the first nine observations:

```
* Number of alternatives in each choice set;
%let m = 3;
data choice3;
   drop i j k autotime plantime trantime;

   * Values of the variable CHOSEN;
   array allmodes[&m] $
      _temporary_ ('Auto' 'Plane' 'Transit');
   * Travel times for the alternatives;
   array times[&m] autotime plantime trantime;
   * New variables that will contain the design:;
   array inters[&m]
      Auto      /*intercept for auto              */
      Plane     /*intercept for plane             */
      Transit;  /*intercept for transit           */
   array cross[%eval(&m * &m)]
      TimeAuto  /*time of auto alternative        */
      PlanAuto  /*cross effect of plane on auto   */
      TranAuto  /*cross effect of transit on auto */
      AutoPlan  /*cross effect of auto on plane   */
      TimePlan  /*time of plane alternative       */
      TranPlan  /*cross effect of transit on plane*/
      AutoTran  /*cross effect of auto on transit */
      PlanTran  /*cross effect of plane on transit*/
      TimeTran; /*time of transit alternative     */
   set travel;
   subject = _n_;
```

```
   * Create &m observations for each choice set;
   do i = 1 to &m;
      Mode = allmodes[i];  /* this alternative    */
      Travtime = times[i]; /* travel time         */
      Choice = 2 - (chosen eq mode);/* 1 - chosen */
      do j = 1 to &m;
         inters[j] = (i eq j);  /* mode indicator */
         do k = 1 to &m;
            * (j=k) - time, otherwise, cross effect;
            cross[&m*(j-1)+k]=times[k]*inters[j];
            end;
         end;
      output;
      end;
   run;
proc print data=choice3(obs=9) label noobs;
   var subject mode travtime choice auto plane transit
       timeauto timeplan timetran autoplan autotran planauto
       plantran tranauto tranplan;
   run;
```

subject	Mode	Travtime	Choice	Auto	Plane	Transit
1	Auto	10.0	2	1	0	0
1	Plane	4.5	1	0	1	0
1	Transit	10.5	2	0	0	1
2	Auto	5.5	1	1	0	0
2	Plane	4.0	2	0	1	0
2	Transit	7.5	2	0	0	1
3	Auto	4.5	2	1	0	0
3	Plane	6.0	2	0	1	0
3	Transit	5.5	1	0	0	1

Time Auto	Time Plan	Time Tran	Auto Plan	Auto Tran	Plan Auto	Plan Tran	Tran Auto	Tran Plan
10.0	0.0	0.0	0.0	0.0	4.5	0.0	10.5	0.0
0.0	4.5	0.0	10.0	0.0	0.0	0.0	0.0	10.5
0.0	0.0	10.5	0.0	10.0	0.0	4.5	0.0	0.0
5.5	0.0	0.0	0.0	0.0	4.0	0.0	7.5	0.0
0.0	4.0	0.0	5.5	0.0	0.0	0.0	0.0	7.5
0.0	0.0	7.5	0.0	5.5	0.0	4.0	0.0	0.0
4.5	0.0	0.0	0.0	0.0	6.0	0.0	5.5	0.0
0.0	6.0	0.0	4.5	0.0	0.0	0.0	0.0	5.5
0.0	0.0	5.5	0.0	4.5	0.0	6.0	0.0	0.0

PROC PHREG allows you to specify `test` statements for testing linear hypotheses of the parameters. The test is a Wald test, which is based on the asymptotic normality of the parameter estimators. The following SAS statements invoke PROC PHREG to fit the so called "Mother Logit" model that includes all the cross-alternative effects. The TEST statement, with label IIA, specifies the null hypothesis that cross-alternative effects `AutoPlan`, `PlanTran`, and `TranAuto` are 0. Since only three cross-alternative effects are estimable and these are the first cross-alternative effects specified in the model, they account for all the cross-alternative effects in the model.

```
proc phreg data=choice3;
   model choice*choice(2) = auto plane transit timeauto timeplan
         timetran autoplan plantran tranauto planauto tranplan
         autotran / ties=breslow;
   IIA: test autoplan,plantran,tranauto;
   strata subject;
   title 'Mother Logit Model';
   run;
```

Mother Logit Model

The PHREG Procedure

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	46.142	24.781
AIC	46.142	40.781
SBC	46.142	49.137

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.3607	8	0.0062
Score	15.4059	8	0.0517
Wald	6.2404	8	0.6203

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Auto	1	-0.73812	3.05933	0.0582	0.8093	0.478
Plane	1	-3.62435	3.48049	1.0844	0.2977	0.027
Transit	0	0	.	.	.	.
TimeAuto	1	-2.23433	1.89921	1.3840	0.2394	0.107
TimePlan	1	-0.10112	0.68621	0.0217	0.8829	0.904
TimeTran	1	0.09785	0.70096	0.0195	0.8890	1.103
AutoPlan	1	0.44495	0.68616	0.4205	0.5167	1.560
PlanTran	1	-0.53234	0.63481	0.7032	0.4017	0.587
TranAuto	1	1.66295	1.51193	1.2097	0.2714	5.275
PlanAuto	0	0	.	.	.	.
TranPlan	0	0	.	.	.	.
AutoTran	0	0	.	.	.	.

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
IIA	1.6526	3	0.6475

The χ^2 statistic for the Wald test is 1.6526 with 3 degrees of freedom, indicating that the cross-alternative effects are not statistically significant ($p = .6475$). A generally more preferable way of testing the significance of the cross-alternative effects is to compare the likelihood of the "Mother logit" model with the likelihood of the reduced model with the cross- alternative effects removed. The following SAS statements invoke PROC PHREG to fit the reduced model:

```
proc phreg data=choice3;
   model choice*choice(2) = auto plane transit timeauto
         timeplan timetran / ties=breslow;
   strata subject;
   title 'Reduced Model without Cross-Alternative Effects';
   run;
```

Reduced Model without Cross-Alternative Effects

The PHREG Procedure

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	46.142	27.153
AIC	46.142	37.153
SBC	46.142	42.376

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.9886	5	0.0019
Score	14.4603	5	0.0129
Wald	7.3422	5	0.1964

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Auto	1	1.71578	1.80467	0.9039	0.3417	5.561
Plane	1	-3.60073	3.30555	1.1866	0.2760	0.027
Transit	0	0	.	.	.	.
TimeAuto	1	-0.79543	0.36327	4.7946	0.0285	0.451
TimePlan	1	0.12162	0.58954	0.0426	0.8366	1.129
TimeTran	1	-0.42184	0.25733	2.6873	0.1012	0.656

The chi-squared statistic for the likelihood ratio test of IIA is $(27.153 - 24.781) = 2.372$, which is not statistically significant ($p = .4989$) when compared to a χ^2 distribution with 3 degrees of freedom. This is consistent with the previous result of the Wald test. (See pages 213, 219, 228, and 354 for other discussions of IIA.)

Final Comments

For some discrete choice problems, the number of available alternatives is not the same for each individual. For example, in a study of consumer brand choices of laundry detergents as prices change, data are pooled from different locations, not all of which offer a brand that contains potash. The varying choice sets across individuals can easily be accommodated in PROC PHREG. For individual j who chooses from a set of m_j alternatives, consider m_j artificial times in which the chosen alternative has an event time 1 and the unchosen alternatives have a censored time of 2. The analysis is carried out in the same fashion as illustrated in the previous section.

Unlike the example of travel demand in which data for each individual are provided, choice data are often given in aggregate form, with choice frequencies indicating the repetition of each choice. One way of dealing with aggregate data is to expand the data to the individual level and carry out the analysis as if you have nonaggregate data. This approach is generally not recommended, because it

defeats the purpose of having a smaller aggregate data set. PROC PHREG provides a FREQ statement that allows you to specify a variable that identifies the frequency of occurrence of each observation. However, with the specification of a FREQ variable, the artificial event time is no longer the only event time in a given stratum, but has ties of the given frequency. With proper stratification, the Breslow likelihood is proportional to the likelihood of the conditional logit model. Thus PROC PHREG can be used to obtain parameter estimates and hypothesis testing results for the choice models.

The `ties=discrete` option should not be used instead of the `ties=breslow` option. This is especially detrimental with aggregate choice data because the likelihood that PROC PHREG is maximizing may no longer be the same as the likelihood of the conditional logit model. `ties=discrete` corresponds to the discrete logistic model for genuinely discrete time scale, which is also suitable for the analysis of case-control studies when there is more than one case in a matched set (Gail, Lubin, and Rubinstein, 1981). For nonaggregate choice data, all `ties=` options give the same results; however, the resources required for the computation are not the same, with `ties=breslow` being the most efficient.

Once you have a basic understanding of how PROC PHREG works, you can use it to fit a variety of models for the discrete choice data. The major involvement in such a task lies in reorganizing the data to create the observations necessary to form the correct risk sets and the appropriate design variables. There are many options in PROC PHREG that can also be useful in the analysis of discrete choice data. For example, the `offset=` option allows you to restrict the coefficient of an explanatory variable to the value of 1; the `selection=` option allows you to specify one of four methods for selecting variables into the model; the `outest=` option allows you to specify the name of the SAS data set that contains the parameter estimates, based on which you can easily compute the predicted probabilities of the alternatives.

This article deals with estimating parameters of discrete choice models. There is active research in the field of marketing research to use design of experiments to study consumer choice behavior. If you are interested in this area, refer to Carson et al. (1994), Kuhfeld et al. (1994), and Lazari et al. (1994).

Conjoint Analysis

Warren F. Kuhfeld

Abstract

Conjoint analysis is used to study consumers' product preferences and simulate consumer choice. This chapter describes conjoint analysis and provides examples using SAS. Topics include metric and nonmetric conjoint analysis, efficient experimental design, data collection and manipulation, holdouts, brand by price interactions, maximum utility and logit simulators, and change in market share.[†]

Introduction

Conjoint analysis is used to study the factors that influence consumers' purchasing decisions. Products possess attributes such as price, color, ingredients, guarantee, environmental impact, predicted reliability, and so on. Consumers typically do not have the option of buying the product that is best in every attribute, particularly when one of those attributes is price. Consumers are forced to make *trade-offs* as they decide which products to purchase. Consider the decision to purchase a car. Increased size generally means increased safety and comfort. The trade off is an increase in cost and environmental impact and a decrease in gas mileage and maneuverability. Conjoint analysis is used to study these trade-offs.

Conjoint analysis is a popular marketing research technique. It is used in designing new products, changing or repositioning existing products, evaluating the effects of price on purchase intent, and simulating market share. Refer to Green and Rao (1971) and Green and Wind (1975) for early introductions to conjoint analysis, Louviere (1988) for a more recent introduction, and Green and Srinivasan (1990) for a review article.

Conjoint Measurement

Conjoint analysis grew out of the area of *conjoint measurement* in mathematical psychology. Conjoint measurement is used to investigate the joint effect of a set of independent variables on an ordinal-scale-of-measurement dependent variable. The independent variables are typically nominal and sometimes interval-scaled variables. Conjoint measurement simultaneously finds a monotonic scoring of the dependent variable and numerical values for each level of each independent variable. The goal is to monotonically transform the ordinal values to equal the sum of their attribute level values. Hence, conjoint measurement is used to derive an interval variable from ordinal data. The conjoint measurement model is a mathematical model, not a statistical model, since it has no statistical error term.

[†]Copies of this chapter (TS-689G) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market .

Conjoint Analysis

Conjoint analysis is based on a main effects analysis-of-variance model. Subjects provide data about their preferences for hypothetical products defined by attribute combinations. Conjoint analysis decomposes the judgment data into components, based on qualitative attributes of the products. A numerical *part-worth utility* value is computed for each level of each attribute. Large part-worth utilities are assigned to the most preferred levels, and small part-worth utilities are assigned to the least preferred levels. The attributes with the largest part-worth utility range are considered the most important in predicting preference. Conjoint analysis is a statistical model with an error term and a loss function.

Metric conjoint analysis models the judgments directly. When all of the attributes are nominal, the metric conjoint analysis is a simple main-effects ANOVA with some specialized output. The attributes are the independent variables, the judgments comprise the dependent variable, and the part-worth utilities are the β's, the parameter estimates from the ANOVA model. The following is a metric conjoint analysis model for three factors:

$$y_{ijk} = \mu + \beta_{1i} + \beta_{2j} + \beta_{3k} + \epsilon_{ijk}$$

where

$$\sum \beta_{1i} = \sum \beta_{2j} = \sum \beta_{3k} = 0$$

This model could be used, for example, to investigate preferences for cars that differ on three attributes: mileage, expected reliability, and price. The y_{ijk} term is one subject's stated preference for a car with the *ith* level of mileage, the *jth* level of expected reliability, and the *kth* level of price. The grand mean is μ, and the error is ϵ_{ijk}. The predicted utility for the *ijk* product is:

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\beta}_{1i} + \hat{\beta}_{2j} + \hat{\beta}_{3k}$$

Nonmetric conjoint analysis finds a monotonic transformation of the preference judgments. The model, which follows directly from conjoint measurement, iteratively fits the ANOVA model until the transformation stabilizes. The R^2 increases during every iteration until convergence, when the change in R^2 is essentially zero. The following is a nonmetric conjoint analysis model for three factors:

$$\Phi(y_{ijk}) = \mu + \beta_{1i} + \beta_{2j} + \beta_{3k} + \epsilon_{ijk}$$

where $\Phi(y_{ijk})$ designates a monotonic transformation of the variable `y`.

The R^2 for a nonmetric conjoint analysis model will always be greater than or equal to the R^2 from a metric analysis of the same data. The smaller R^2 in metric conjoint analysis is not necessarily a disadvantage, since results should be more stable and reproducible with the metric model. Metric conjoint analysis was derived from nonmetric conjoint analysis as a special case. Today, metric conjoint analysis is probably used more often than nonmetric conjoint analysis.

In the SAS System, conjoint analysis is performed with the SAS/STAT procedure TRANSREG (transformation regression). Metric conjoint analysis models are fit using ordinary least squares, and nonmetric conjoint analysis models are fit using an alternating least squares algorithm (Young, 1981; Gifi,

1990). Conjoint analysis is explained more fully in the examples. The "PROC TRANSREG Specifications" section of this chapter starting on page 467 documents the PROC TRANSREG statements and options that are most relevant to conjoint analysis. The "Samples of PROC TRANSREG Usage" section starting on page 476 shows some typical conjoint analysis specifications. This chapter shows some of the SAS programming that is used for conjoint analysis. Alternatively, there is a marketing research GUI that performs conjoint analysis available from the main display manager PMENU by selecting: `Solutions` → `Analysis` → `Market Research`.

Choice-Based Conjoint

The meaning of the word "conjoint" has broadened over the years from conjoint measurement to conjoint analysis (which at first always meant what we now call nonmetric conjoint analysis) and later to metric conjoint analysis. Metric and nonmetric conjoint analysis are based on a linear ANOVA model. In contrast, a different technique, discrete choice, is based on the nonlinear multinomial logit model. Discrete choice is sometimes referred to as "choice-based conjoint." This technique is not discussed in this chapter, however it is discussed in detail starting on page 81.

Preliminaries

Design of Experiments

The design of experiments is a fundamental part of conjoint analysis.[‡] During conjoint analysis data collection, subjects are asked to judge their preferences for hypothetical products defined by attribute combinations. Experimental designs are used to select the attribute combinations. The *factors* of an experimental design are variables that have two or more fixed values, or *levels*. Experiments are performed to study the effects of the factor levels on the *response*, or dependent variable. In a conjoint study, the factors are the attributes of the hypothetical products or services, and the response is a rating or ranking of product preference. The rows of a design are called *runs* and correspond to product profiles in a full-profile conjoint study. For example, the following table contains an experimental design with two factors, brand and price. Brand has three levels, Acme, Ajax, and Widget, and price has two levels, $1.99 and $2.99.

Full-Profile Conjoint Design	
Brand	Price
Acme	1.99
Acme	2.99
Ajax	1.99
Ajax	2.99
Widget	1.99
Widget	2.99

This is an example of a *full-factorial design*. It consists of all possible combinations of the levels of the factors. With five factors, two at four levels and three at five levels (denoted $4^2 5^3$), there are $4 \times 4 \times 5 \times 5 \times 5 = 2000$ combinations in the full-factorial design. Factorial designs allow you to estimate main effects and interactions. A *main effect* is a simple effect, such as a price or brand effect. In a main effects model, for example, the brand effect is the same at the different prices and the price effect is the same for the different brands. *Interactions* involve two or more factors, such as a brand by price interaction. In a model with interactions, for example, brand preference is different at the different prices and the price effect is different for the different brands.

In a full-factorial design, all main effects, all two-way interactions, and all higher-order interactions are estimable and uncorrelated. The problem with a full-factorial design is that, for most practical situations, it is too cost-prohibitive and tedious to have subjects consider all possible combinations. For this reason, researchers often use *fractional-factorial designs*, which have fewer runs than full-factorial designs. The price of having fewer runs is that some effects become confounded. Two effects are *confounded* or *aliased* when they are not distinguishable from each other.

A special type of fractional-factorial design is the *orthogonal array*. An orthogonal array or orthogonal design is one in which all estimable effects are uncorrelated. The term "orthogonal array," as it is sometimes used in practice, is imprecise. It is correctly used to refer to designs that are both orthogonal and balanced, and hence optimal. The term is sometimes also used to refer to designs that are orthogonal but not balanced, and hence not 100% efficient and sometimes not even optimal.

[‡]See pages 39 and 84 for more detailed discussions of experimental design.

A design is *balanced* when each level occurs equally often within each factor, which that means the intercept is orthogonal to each effect. Imbalance is a generalized form of nonorthogonality, hence it increases the variances of the parameter estimates and decreases efficiency. Orthogonal designs are often practical for main-effects models when the number of factors is small (say six or fewer) and the number of levels of each factor is small (say four or fewer). However, there are some situations in which orthogonal designs are not practical, such as when

- not all combinations of factor levels are feasible or make sense
- the desired number of runs is not available in an orthogonal design
- a nonstandard model is being used, such as a model with interactions, polynomials, or splines.

When an orthogonal and balanced design is not practical, you must make a choice. One choice is to change the factors and levels to fit some known orthogonal design. This choice is undesirable for obvious reasons. When a suitable orthogonal and balanced design does not exist, efficient nonorthogonal designs can be used instead. Nonorthogonal designs, where some coefficients may be slightly correlated, can be used in all of the situations listed previously. You do not have to adapt every experiment to fit some known orthogonal array. First you choose the number of runs. You are not restricted by the sizes of orthogonal arrays, which come in specific numbers of runs (such as 16, 18, 27, 32, 36, and so forth) for specific numbers of factors with specific numbers of levels. Then you specify the levels of each of the factors and the number of runs. Algorithms for generating efficient designs select a set of *design points* that optimize an efficiency criterion. Throughout this book, we will use the `%MktEx` macro to find good, efficient experimental designs. The `%MktEx` macro is a part of the SAS autocall library. See page 479 for information on installing and using SAS autocall macros. Refer to Kuhfeld, Tobias, and Garratt (1994), page 39, for more information on efficient experimental designs.

The rest of the *Design of Experiments* section discusses precisely what is meant by an efficient design. It is not critical that you understand the rest of this section before doing a conjoint analysis, and if you wish, you may now skip to *The Output Delivery System* section on page 366.

The goodness or *efficiency* of an experimental design can be quantified. Common measures of the efficiency of an $(N_D \times p)$ design matrix $\mathbf{X}$ are based on the *information matrix* $\mathbf{X}'\mathbf{X}$. The variance-covariance matrix of the vector of parameter estimates $\hat{\boldsymbol{\beta}}$ in a least-squares analysis is proportional to $(\mathbf{X}'\mathbf{X})^{-1}$. An efficient design will have a "small" variance matrix, and the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}$ provide measures of its "size." The two most prominent efficiency measures are based on quantifying the idea of matrix size by averaging (in some sense) the eigenvalues or variances. *A-efficiency* is a function of the arithmetic mean of the eigenvalues, which is also the arithmetic mean of the variances and is given by trace $((\mathbf{X}'\mathbf{X})^{-1})/p$. (The trace is the sum of the diagonal elements of a matrix, which is the sum of the eigenvalues.) *D-efficiency* is a function of the geometric mean of the eigenvalues, which is given by $|(\mathbf{X}'\mathbf{X})^{-1}|^{1/p}$. (The determinant, $|(\mathbf{X}'\mathbf{X})^{-1}|$, is the product of the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}$.) A third common efficiency measure, *G-efficiency*, is based on σ_M, the maximum standard error for prediction over the candidate set. All three of these criteria are convex functions of the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}$ and hence are usually highly correlated.

For all three criteria, if a balanced and orthogonal design exists, then it has optimum efficiency; conversely, the more efficient a design is, the more it tends toward balance and orthogonality. A design is balanced and orthogonal when $(\mathbf{X}'\mathbf{X})^{-1}$ is diagonal, $\frac{1}{N_D}\mathbf{I}$, for a suitably coded $\mathbf{X}$. A design is orthogonal when the submatrix of $(\mathbf{X}'\mathbf{X})^{-1}$, excluding the row and column for the intercept, is diagonal; there may be off-diagonal nonzeros for the intercept. A design is balanced when all off-diagonal elements in the intercept row and column are zero.

These measures of efficiency can be scaled to range from 0 to 100 (see page 91 for the orthogonal coding of $\mathbf{X}$ that must be used with these formulas):

$$A\text{-efficiency} = 100 \times \frac{1}{N_D \text{ trace}\left((\mathbf{X}'\mathbf{X})^{-1}\right)/p}$$

$$D\text{-efficiency} = 100 \times \frac{1}{N_D \left|(\mathbf{X}'\mathbf{X})^{-1}\right|^{1/p}}$$

$$G\text{-efficiency} = 100 \times \frac{\sqrt{p/N_D}}{\sigma_M}$$

These efficiencies measure the goodness of the design relative to hypothetical orthogonal designs that may not exist, so they are not useful as absolute measures of design efficiency. Instead, they should be used relatively, to compare one design to another for the same situation. Efficiencies that are not near 100 may be perfectly satisfactory.

The Output Delivery System

The Output Delivery System (ODS) can be used to customize the output of SAS procedures including PROC TRANSREG, the procedure we use for conjoint analysis. PROC TRANSREG can produce a great deal of information for conjoint analysis, more than we often wish to see. We will use ODS primarily to exclude certain portions of the default conjoint output in which we are usually not interested. This will create a better, more parsimonious display for typical analyses. However, when we need it, we can revert back to getting the full array of information. See page 95 for other examples of customizing output using ODS. You can run the following step once to customize PROC TRANSREG conjoint analysis output.

```
proc template;
   edit Stat.Transreg.ParentUtilities;
      column Label Utility StdErr tValue Probt Importance Variable;
      header title;
      define title; text 'Part-Worth Utilities'; space=1; end;
      define Variable; print=off; end;
      end;
   run;
```

Running this step edits the templates for the main conjoint analysis results table and stores a copy in SASUSER. These changes will remain in effect until you delete them. These changes move the variable label to the first column, turn off printing the variable names, and set the table header to "Part-Worth Utilities". These changes assume that each effect in the model has a variable label associated with it, so there is no need to print variable names. This will usually be the case. To return to the default output, run the following.

```
* Delete edited template, restore original template;
proc template;
   delete Stat.Transreg.ParentUtilities;
   run;
```

By default, PROC TRANSREG prints an ANOVA table for metric conjoint analysis and both univariate and multivariate ANOVA tables for nonmetric conjoint analysis. With nonmetric conjoint analysis, PROC TRANSREG sometimes prints liberal and conservative ANOVA tables. All of the possible

ANOVA tables, along with some header notes, can be suppressed by specifying the following statement before running PROC TRANSREG.

```
ods exclude notes anova liberalanova conservanova
            mvanova liberalmvanova conservmvanova;
```

For metric conjoint analysis, this statement can be abbreviated as follows.

```
ods exclude notes mvanova anova;
```

The rest of this section gives more details about what the PROC TEMPLATE step does and why. The rest of this section can be helpful if you wish to further customize the output from TRANSREG or some other procedure. Impatient readers may skip ahead to the candy example on page 369.

We are most interested in the part-worth utilities table in conjoint analysis, which contains the part-worth utilities, their standard errors, and the importance of each attribute. We can first use PROC TEMPLATE to identify the template for the utilities table and then edit the template. First, let's have PROC TEMPLATE display the templates for PROC TRANSREG. The `source stat.transreg` statement specifies that we want to see PROC TEMPLATE source code for the STAT product and the TRANSREG procedure.

```
proc template;
   source stat.transreg;
   run;
```

If we search the results for "Utilities", we find the template for the table `Stat.Transreg.ParentUtilities`. Here is the template for the part-worth utilities table.

```
define table Stat.Transreg.ParentUtilities;
   notes "Parent Utilities Table for Proc Transreg";
   dynamic FootMessages TitleText;
   column Label Utility StdErr tValue Probt Importance Variable;
   header Title;
   footer Foot;
   define Title;
      text TitleText;
      space = 1;
      spill_margin;
      first_panel;
   end;
   define Label;
      parent = Stat.Transreg.Label;
      style = RowHeader;
   end;
   define Utility;
      header = "Utility";
      format_width = 7;
      parent = Stat.Transreg.Coefficient;
   end;
   define StdErr;
      parent = Stat.Transreg.StdErr;
   end;
```

```
      define tValue;
         parent = Stat.Transreg.tValue;
         print = OFF;
      end;
      define Probt;
         parent = Stat.Transreg.Probt;
         print = OFF;
      end;
      define Importance;
         header = %nrstr(";Importance;%(%% Utility;Range%)");
         translate _val_=._ into " ";
         format = 7.3;
      end;
      define Variable;
         parent = Stat.Transreg.Variable;
      end;
      define Foot;
         text FootMessages;
         just = l;
         maximize;
      end;
      control = control;
      required_space = 20;
   end;
```

Recall that we ran the following step to customize the output.

```
   proc template;
      edit Stat.Transreg.ParentUtilities;
         column Label Utility StdErr tValue Probt Importance Variable;
         header title;
         define title; text 'Part-Worth Utilities'; space=1; end;
         define Variable; print=off; end;
         end;
      run;
```

We specified the `edit Stat.Transreg.ParentUtilities` statement to name the table that we wish to change. The `column` statement was copied from the PROC TEMPLATE source listing, and it names all of the columns in the table. Some, like `tValue` and `Probt` do not print by default. We will change the `Variable` column to also not print. We redefine `Variable` with the `print=off` option specified. We also redefine the table header to read "Part-Worth Utilities". The names in the `column` and `header` statements must match the names in the original template.

Chocolate Candy Example

This example illustrates conjoint analysis with rating scale data and a single subject. The subject was asked to rate his preference for eight chocolate candies. The covering was either dark or milk chocolate, the center was either chewy or soft, and the candy did or did not contain nuts. The candies were rated on a 1 to 9 scale where 1 means low preference and 9 means high preference. Conjoint analysis is used to determine the importance of each attribute and the part-worth utility for each level of each attribute.

Metric Conjoint Analysis

After data collection, the attributes and the rating data are entered into a SAS data set.

```
title 'Preference for Chocolate Candies';

data choc;
   input Chocolate $ Center $ Nuts $& Rating;
   datalines;
Dark  Chewy  Nuts       7
Dark  Chewy  No Nuts    6
Dark  Soft   Nuts       6
Dark  Soft   No Nuts    4
Milk  Chewy  Nuts       9
Milk  Chewy  No Nuts    8
Milk  Soft   Nuts       9
Milk  Soft   No Nuts    7
;
```

Note that the "`&`" specification in the `input` statement is used to read character data with embedded blanks.

PROC TRANSREG is used to perform a metric conjoint analysis.

```
ods exclude notes mvanova anova;
proc transreg utilities separators=', ' short;
   title2 'Metric Conjoint Analysis';
   model identity(rating) = class(chocolate center nuts / zero=sum);
   run;
```

Printed output from the metric conjoint analysis is requested by specifying the `utilities` option in the `proc` statement. The value specified in the `separators=` option, in this case a comma followed by a blank, is used in constructing the labels for the part-worth utilities in the printed output. With these options, the labels will consist of the `class` variable name, a comma, a blank and the values of the `class` variables. We specify the `short` option to suppress the iteration history. PROC TRANSREG will still print a convergence summary table so we will know if there are any convergence problems. Since this is a metric conjoint analysis, there should be only one iteration and there should not be any problems. We specified `ods exclude notes mvanova anova` to exclude ANOVA information (which we usually want to ignore) and provide more parsimonious output. The analysis variables, the transformation of each variable, and transformation specific options are specified in the `model` statement.

The `model` statement provides for general transformation regression models, so it has a markedly different syntax from other SAS/STAT procedure `model` statements. Variable lists are specified in parentheses after a transformation name. The specification `identity(rating)` requests an `identity` transformation of the dependent variable `Rating`. A transformation name must be specified for all variable lists, even for the dependent variable in metric conjoint analysis, when no transformation is desired. The `identity` transformation of `Rating` will not change the original scoring. An equal sign follows the dependent variable specification, then the attribute variables are specified along with their transformation. The specification

```
class(chocolate center nuts / zero=sum)
```

designates the attributes as `class` variables with the restriction that the part-worth utilities sum to zero within each attribute. A slash must be specified to separate the variables from the transformation option `zero=sum`. The `class` specification creates a main-effects design matrix from the specified variables. This example produces only printed output; later examples will show how to store results in output SAS data sets.

The output is shown next. Recall that we used an `ods exclude` statement and we used PROC TEMPLATE on page 366 to customize the output from PROC TRANSREG.

Preference for Chocolate Candies
Metric Conjoint Analysis

The TRANSREG Procedure

Dependent Variable Identity(Rating)

Class Level Information

Class	Levels	Values
Chocolate	2	Dark Milk
Center	2	Chewy Soft
Nuts	2	No Nuts Nuts

Number of Observations Read	8
Number of Observations Used	8

Identity(Rating)
Algorithm converged.

Root MSE	0.50000	R-Square	0.9500
Dependent Mean	7.00000	Adj R-Sq	0.9125
Coeff Var	7.14286		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	7.0000	0.17678	
Chocolate, Dark	-1.2500	0.17678	50.000
Chocolate, Milk	1.2500	0.17678	
Center, Chewy	0.5000	0.17678	20.000
Center, Soft	-0.5000	0.17678	
Nuts, No Nuts	-0.7500	0.17678	30.000
Nuts, Nuts	0.7500	0.17678	

We see `Algorithm converged` in the output indicating no problems with the iterations. We also see $R^2 = 0.95$. The last table displays the part-worth utilities. The part-worth utilities show the most and least preferred levels of the attributes. Levels with positive utility are preferred over those with negative utility. Milk chocolate (part-worth utility = 1.25) was preferred over dark (-1.25), chewy center (0.5) over soft (-0.5), and nuts (0.75) over no nuts (-0.75).

Conjoint analysis provides an approximate decomposition of the original ratings. The predicted utility for a candy is the sum of the intercept and the part-worth utilities. The conjoint analysis model for the preference for chocolate type i, center j, and nut content k is

$$y_{ijk} = \mu + \beta_{1i} + \beta_{2j} + \beta_{3k} + \epsilon_{ijk}$$

for $i = 1, 2$; $j = 1, 2$; $k = 1, 2$; where

$$\beta_{11} + \beta_{12} = \beta_{21} + \beta_{22} = \beta_{31} + \beta_{32} = 0$$

The part-worth utilities for the attribute levels are the parameter estimates $\hat{\beta}_{11}$, $\hat{\beta}_{12}$, $\hat{\beta}_{21}$, $\hat{\beta}_{22}$, $\hat{\beta}_{31}$, and $\hat{\beta}_{32}$ from this main-effects ANOVA model. The estimate of the intercept is $\hat{\mu}$, and the error term is ϵ_{ijk}.

The predicted utility for the ijk combination is

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\beta}_{1i} + \hat{\beta}_{2j} + \hat{\beta}_{3k}$$

For the most preferred milk/chewy/nuts combination, the predicted utility and actual preference values are

$$7.0 + 1.25 + 0.5 + 0.75 = 9.5 = \hat{y} \approx y = 9.0$$

For the least preferred dark/soft/no nuts combination, the predicted utility and actual preference values are

$$7.0 + -1.25 + -0.5 + -0.75 = 4.5 = \hat{y} \approx y = 4.0$$

The predicted utilities are regression predicted values; the squared correlation between the predicted utilities for each combination and the actual preference ratings is the R^2.

The *importance* value is computed from the part-worth utility range for each factor (attribute). Each range is divided by the sum of all ranges and multiplied by 100. The factors with the largest part-worth utility ranges are the most important in determining preference. Note that when the attributes have a varying number of levels, attributes with the most levels sometimes have inflated importances (Wittink, Krishnamurthi, and Reibstein ; 1989).

The importance values show that type of chocolate, with an importance of 50%, was the most important attribute in determining preference.

$$\frac{100 \times (1.25 - -1.25)}{(1.25 - -1.25) + (0.50 - -0.50) + (0.75 - -0.75)} = 50\%$$

The second most important attribute was whether the candy contained nuts, with an importance of 30%.

$$\frac{100 \times (0.75 - -0.75)}{(1.25 - -1.25) + (0.50 - -0.50) + (0.75 - -0.75)} = 30\%$$

Type of center was least important at 20%.

$$\frac{100 \times (0.50 - -0.50)}{(1.25 - -1.25) + (0.50 - -0.50) + (0.75 - -0.75)} = 20\%$$

Nonmetric Conjoint Analysis

In the next part of this example, PROC TRANSREG is used to perform a nonmetric conjoint analysis of the candy data set. The difference between requesting a nonmetric and metric conjoint analysis is the dependent variable transformation; a `monotone` transformation of `Rating` variable is requested instead of an `identity` transformation. Also, we did not specify the `short` option this time so that we could see the iteration history table. The `output` statement is used to put the transformed rating into the `out=` output data set.

```
ods exclude notes anova liberalanova conservanova
            mvanova liberalmvanova conservmvanova;
proc transreg utilities separators=', ';
   title2 'Nonmetric Conjoint Analysis';
   model monotone(rating) = class(chocolate center nuts / zero=sum);
   output;
   run;
```

Nonmetric conjoint analysis iteratively derives the monotonic transformation of the ratings. Recall that we used an `ods exclude` statement and we used PROC TEMPLATE on page 366 to customize the output from PROC TRANSREG.

Preference for Chocolate Candies
Nonmetric Conjoint Analysis

The TRANSREG Procedure

Dependent Variable Monotone(Rating)

Class Level Information

Class	Levels	Values
Chocolate	2	Dark Milk
Center	2	Chewy Soft
Nuts	2	No Nuts Nuts

Number of Observations Read	8
Number of Observations Used	8

TRANSREG Univariate Algorithm Iteration History for Monotone(Rating)

Iteration Number	Average Change	Maximum Change	R-Square	Criterion Change	Note
1	0.08995	0.23179	0.95000		
2	0.01263	0.03113	0.96939	0.01939	
3	0.00345	0.00955	0.96981	0.00042	
4	0.00123	0.00423	0.96984	0.00003	
5	0.00050	0.00182	0.96985	0.00000	
6	0.00021	0.00078	0.96985	0.00000	
7	0.00009	0.00033	0.96985	0.00000	
8	0.00004	0.00014	0.96985	0.00000	
9	0.00002	0.00006	0.96985	0.00000	
10	0.00001	0.00003	0.96985	0.00000	Converged

Algorithm converged.

Root MSE	0.38829	R-Square	0.9698
Dependent Mean	7.00000	Adj R-Sq	0.9472
Coeff Var	5.54699		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	7.0000	0.13728	
Chocolate, Dark	-1.3143	0.13728	53.209
Chocolate, Milk	1.3143	0.13728	
Center, Chewy	0.4564	0.13728	18.479
Center, Soft	-0.4564	0.13728	
Nuts, No Nuts	-0.6993	0.13728	28.312
Nuts, Nuts	0.6993	0.13728	

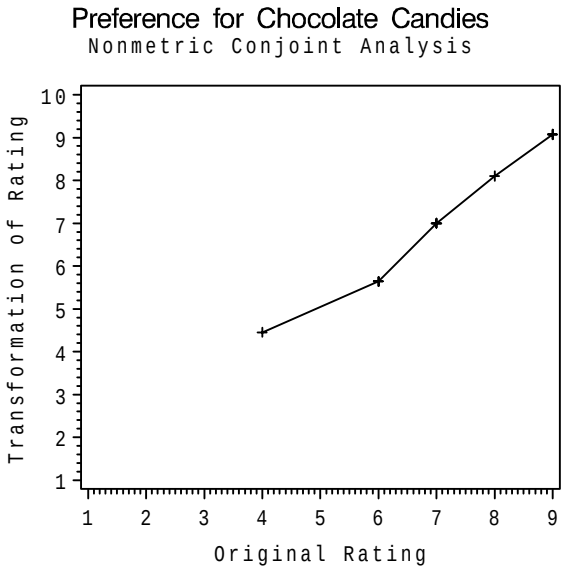
The standard errors are not adjusted for the fact that the dependent variable was transformed and so are generally liberal (too small).

The R^2 increases from 0.95 for the metric case to 0.96985 for the nonmetric case. The importances and part-worth utilities are slightly different from the metric analysis, but the overall pattern of results is the same.

The GPLOT procedure is used to plot the transformation of the ratings.

```
proc sort; by rating; run;

proc gplot;
   title  h=1.5 'Preference for Chocolate Candies';
   title2 h=1 'Nonmetric Conjoint Analysis';
   plot trating * rating = 1 / frame haxis=axis2 vaxis=axis1;
   symbol1 v=plus i=join;
   axis1 order=(1 to 10)
         label=(angle=90 'Transformation of Rating');
   axis2 order=(1 to  9) label=('Original Rating');
   run; quit;
```



In this case, the transformation is nearly linear. In practice, the R^2 may increase much more than it did in this example, and the transformation may be markedly nonlinear.

Frozen Diet Entrées Example (Basic)

This example uses PROC TRANSREG to perform a conjoint analysis to study preferences for frozen diet entrées. The entrées have four attributes: three with three levels and one with two levels. The attributes are shown in the table.

Factor	Levels		
Main Ingredient	Chicken	Beef	Turkey
Fat Claim Per Serving	8 Grams	5 Grams	2 Grams
Price	$2.59	$2.29	$1.99
Calories	350	250	

Choosing the Number of Stimuli

Ideally, for this design, we would like the number of *runs* in the experimental design to be divisible by 2 (because of the two-level factor), 3 (because of the three-level factors), $2 \times 3 = 6$ (to have equal numbers of products in each two-level and three-level factor combinations), and $3 \times 3 = 9$ (to have equal numbers of products in each pair of three-level factor combinations). If we fit a main-effects model, we need at least $1 + 3 \times (3 - 1) + (2 - 1) = 8$ runs. We can avoid doing this math ourselves and instead use the `%MktRuns` autocall macro to help us choose the number of products. See page 479 for macro documentation and information on installing and using SAS autocall macros. To use this macro, you specify the number of levels for each of the factors. For this example, specify three 3's and one 2.

```
title 'Frozen Diet Entrees';

%mktruns( 3 3 3 2 )
```

```
                    Frozen Diet Entrees

                      Design Summary

                   Number of
                   Levels       Frequency

                       2            1
                       3            3

                    Frozen Diet Entrees

              Saturated      = 8
              Full Factorial = 54
```

```
 Some Reasonable                      Cannot Be
    Design Sizes        Violations    Divided By

              18 *              0
              36 *              0
              12                3     9
              24                3     9
              30                3     9
               9                4     2 6
              27                4     2 6
              15                7     2 6 9
              21                7     2 6 9
              33                7     2 6 9

* - 100% Efficient Design can be made with the MktEx Macro.

   n    Design                          Reference

  18    2 **  1  3 **  7                Orthogonal Array
  36    2 ** 13  3 **  4                Orthogonal Array
  36    2 ** 11  3 ** 12                Orthogonal Array
  36    2 ** 10  3 **  8  6 **  1       Orthogonal Array
  36    2 **  9  3 **  4  6 **  2       Orthogonal Array
  36    2 **  4  3 ** 13                Orthogonal Array
  36    2 **  3  3 **  9  6 **  1       Orthogonal Array
  36    2 **  2  3 ** 12  6 **  1       Orthogonal Array
  36    2 **  2  3 **  5  6 **  2       Orthogonal Array
  36    2 **  1  3 **  8  6 **  2       Orthogonal Array
  36    2 **  1  3 **  3  6 **  3       Orthogonal Array
```

The output tells us that we need at least eight products, shown by the "Saturated = 8". The sizes 18 and 36 would be optimal. Twelve is a good size but three times it cannot be divided by $9 = 3 \times 3$. The "three times" comes from the $3(3-1)/2 = 3$ pairs of three-level factors. Similarly, the size 9 has four violations because it cannot be divided once by 2 and three times by $6 = 2 \times 3$ (once for each three-level factor and two-level factor pair). We could use a size smaller than 18 and not have equal frequencies everywhere, but 18 is a manageable number so we will use 18.

When an orthogonal and balanced design is available from the `%MktEx` macro, the `%MktRuns` macro tells us about it. For example, the macro tells us that our design, which can be designated $2^1 3^3$, is available in 18 runs, and can be constructed from a design with 1 two-level factor (`2 ** 1` or 2^1) and 7 three-level factors (`3 ** 7` or 3^7). Both the `%MktRuns` and `%MktEx` macros accept this '$n * *m$' exponential syntax as input, which means m factors each at n levels. Hence, `2 3 ** 7` or `2 ** 1 3 ** 7` or `2 3 3 3 3 3 3 3` are all equivalent level-list specifications for the experimental design $2^1 3^7$, which has 1 two-level factor and 7 three-level factors.

Generating the Design

We can use the %MktEx autocall macro to find a design. When you invoke the %MktEx macro for a simple problem, you only need to specify the numbers of levels and number of runs. The macro does the rest. The %MktEx macro can create designs in a number of ways. For this problem, it simply looks up an orthogonal design. Here is the %MktEx macro call for this example:

```
%mktex(3 3 3 2, n=18)
```

The first argument to the %MktEx macro is a list of factor levels, and the second is the number of runs (n=18). These are all the options that are needed for a simple problem such as this one. However, throughout this book, random number seeds are explicitly specified with the seed= option so that you can reproduce these results.[§] Here is the code for creating our design with the random number seed and the actual factor names specified:

```
%mktex(3 3 3 2, n=18, seed=151)
%mktlab(vars=Ingredient Fat Price Calories)
```

The %MktEx macro always creates factors named x1, x2, and so on. The %MktLab autocall macro is used to change the names when you want to provide actual factor names. This example has four factors, Ingredient, Fat, and Price, each with three levels and Calories with two levels.

Here is the output:

Frozen Diet Entrees

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	100.0000	100.0000	Tab
1	End	100.0000		

Frozen Diet Entrees

The OPTEX Procedure

Class Level Information

Class	Levels	-Values-
x1	3	1 2 3
x2	3	1 2 3
x3	3	1 2 3
x4	2	1 2

[§]By specifying a random number seed, results should be reproducible within a SAS release for a particular operating system. However, due to machine differences, some results may not be exactly reproducible on other machines. For most orthogonal and balanced designs, the results should be reproducible. When computerized searches are done, it is likely that you will not get the same design as the one in the book, although you would expect the efficiency differences to be slight.

Frozen Diet Entrees

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	100.0000	100.0000	100.0000	0.6667

We see that the macro had no trouble finding an optimal, 100% efficient experimental design through table look up. The value `Tab` in the `Notes` column of the algorithm search history tells us the macro was able to use table look up. See pages 479, 546, and the discrete choice examples starting on page 81 for more information on how the `%MktEx` macro works.

The `%MktEx` macro creates two output data sets with the experimental design, DESIGN and RANDOMIZED. The DESIGN data set is sorted and for a number of the tabled designs, often has a first row consisting entirely of ones. For these reasons, you should typically use the *randomized* design. In the randomized design, the profiles are presented in a random order and the levels have been randomly reassigned. Neither of these operations affects the design efficiency, balance, or orthogonality. When there are restrictions on the design (see for example page 433), the profiles are sorted into a random order, but the levels are *not* randomly reassigned. The randomized design is the default input to the `%MktLab` macro.

Evaluating and Preparing the Design

We will use the FORMAT procedure to create descriptive labels for the levels of the attributes. By default, the values of the factors are positive integers. For example for `ingredient`, we create a format `if` (for Ingredient Format) that assigns the descriptive value label "Chicken" for level 1, "Beef" for level 2, and "Turkey" for level 3. A permanent SAS data set is created with the formats assigned (although, as we will see in the next example, we could have done this previously in the `%MktLab` step). Finally, the design is printed.

```
proc format;
   value if  1='Chicken'  2='Beef'     3='Turkey';
   value ff  1='8 Grams'  2='5 Grams'  3='2 Grams';
   value pf  1='$2.59'    2='$2.29'    3='$1.99';
   value cf  1='350'      2='250';
   run;
data sasuser.dietdes;
   set final;
   format ingredient if. fat ff. price pf. calories cf.;
   run;
proc print; run;
```

Here is the design.

Frozen Diet Entrees

Obs	Ingredient	Fat	Price	Calories
1	Turkey	5 Grams	$1.99	350
2	Turkey	8 Grams	$2.29	350
3	Chicken	8 Grams	$1.99	350
4	Turkey	2 Grams	$2.59	250
5	Beef	8 Grams	$2.59	350
6	Beef	2 Grams	$1.99	350
7	Beef	5 Grams	$2.29	350
8	Beef	5 Grams	$2.29	250
9	Chicken	2 Grams	$2.29	350
10	Beef	8 Grams	$2.59	250
11	Turkey	8 Grams	$2.29	250
12	Chicken	5 Grams	$2.59	350
13	Chicken	5 Grams	$2.59	250
14	Chicken	2 Grams	$2.29	250
15	Turkey	5 Grams	$1.99	250
16	Turkey	2 Grams	$2.59	350
17	Beef	2 Grams	$1.99	250
18	Chicken	8 Grams	$1.99	250

Even when you know the design is 100% *D*-efficient, orthogonal, and balanced, it is good to run basic checks on your designs. You can use the `%MktEval` autocall macro to display information about the design.

```
%mkteval;
```

The macro first prints a matrix of canonical correlations between the factors. We hope to see an identity matrix (a matrix of ones on the diagonal and zeros everywhere else), which would mean that all of the factors are uncorrelated. Next, the macro prints all one-way frequencies for all attributes, all two-way frequencies, and all *n*-way frequencies (in this case four-way frequencies). We hope to see equal or at least nearly equal one-way and two-way frequencies, and we want to see that each combination occurs only once.

Frozen Diet Entrees
Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

	Ingredient	Fat	Price	Calories
Ingredient	1	0	0	0
Fat	0	1	0	0
Price	0	0	1	0
Calories	0	0	0	1

```
                    Frozen Diet Entrees
                  Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316

                          Frequencies

Ingredient                6 6 6
Fat                       6 6 6
Price                     6 6 6
Calories                  9 9
Ingredient Fat            2 2 2 2 2 2 2 2 2
Ingredient Price          2 2 2 2 2 2 2 2 2
Ingredient Calories       3 3 3 3 3 3
Fat Price                 2 2 2 2 2 2 2 2 2
Fat Calories              3 3 3 3 3 3
Price Calories            3 3 3 3 3 3
N-Way                     1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

A *canonical correlation* is the maximum correlation between linear combinations of the coded factors (see page 90). All zeros off the diagonal show that this design is orthogonal for main effects. If any off-diagonal canonical correlations had been greater than 0.316 ($r^2 > 0.1$), the macro would have listed them in a separate table. The last title line tells you that none of them were this large. For nonorthogonal designs and designs with interactions, the canonical-correlation matrix is not a substitute for looking at the variance matrix (with `examine=v`) in the `%MktEx` macro. The `%MktEx` macro just provides a quick and more-compact picture of the correlations between the factors. The variance matrix is sensitive to the actual model specified and the coding. The canonical-correlation matrix just tells you if there is some correlation between the main effects. In this case, there are no correlations.

The equal one-way frequencies show you that this design is balanced. The equal two-way frequencies show you that this design is orthogonal. Equal one-way and two-way frequencies together show you that this design is 100% D-efficient. The n-way frequencies, all equal to one, show you that there are no duplicate profiles. This is a perfect design for a main effects model. However, there are other 100% efficient designs for this problem with duplicate observations. In the last part of the output, the N-Way frequencies may contain some 2's for those designs. You can specify `options=nodups` in the `%MktEx` macro to ensure that there are no duplicate profiles.

The `%MktEval` macro produces a very compact summary of the design, hence some information, for example the levels to which the frequencies correspond, is not shown. You can use the `print=freqs` option in the `%MktEval` macro to get a less compact and more detailed display.

Printing the Stimuli and Data Collection

Next, we generate the stimuli. The `data _null_` step uses the `file` statement to set the print destination to the printed output destination. The design data set is read with the `set` statement. A `put` statement prints the attributes along with some constant text and the combination number. The `put` statement option `+3` skips 3 spaces, `@50` starts printing in column 50, `+(-1)` skips one space *backwards* getting rid of the blank that would by default appear after the stimulus number, and `/` skips to a new line. Text enclosed in quotes is literally copied to the output. For our attribute variables, the formatted

values are printed. The variable _n_ is the number of the current pass through the DATA step, which in this case is the stimulus number. The if statement causes six descriptions to be printed on a page.

```
title;
data _null_;
   file print;
   set sasuser.dietdes;
   put ///
       +3 ingredient 'Entree' @50 '(' _n_ +(-1) ')' /
       +3 'With ' fat 'of Fat and ' calories  'Calories' /
       +3 'Now for Only ' Price +(-1) '.'///;
   if mod(_n_, 6) = 0 then put _page_;
   run;
```

```
Turkey Entree                                (1)
With 5 Grams of Fat and 350 Calories
Now for Only $1.99.

Turkey Entree                                (2)
With 8 Grams of Fat and 350 Calories
Now for Only $2.29.

Chicken Entree                               (3)
With 8 Grams of Fat and 350 Calories
Now for Only $1.99.

Turkey Entree                                (4)
With 2 Grams of Fat and 250 Calories
Now for Only $2.59.

Beef Entree                                  (5)
With 8 Grams of Fat and 350 Calories
Now for Only $2.59.

Beef Entree                                  (6)
With 2 Grams of Fat and 350 Calories
Now for Only $1.99.

Beef Entree                                  (7)
With 5 Grams of Fat and 350 Calories
Now for Only $2.29.

Beef Entree                                  (8)
With 5 Grams of Fat and 250 Calories
Now for Only $2.29.

Chicken Entree                               (9)
With 2 Grams of Fat and 350 Calories
Now for Only $2.29.
```

```
Beef Entree                                (10)
With 8 Grams of Fat and 250 Calories
Now for Only $2.59.

Turkey Entree                              (11)
With 8 Grams of Fat and 250 Calories
Now for Only $2.29.

Chicken Entree                             (12)
With 5 Grams of Fat and 350 Calories
Now for Only $2.59.

Chicken Entree                             (13)
With 5 Grams of Fat and 250 Calories
Now for Only $2.59.

Chicken Entree                             (14)
With 2 Grams of Fat and 250 Calories
Now for Only $2.29.

Turkey Entree                              (15)
With 5 Grams of Fat and 250 Calories
Now for Only $1.99.

Turkey Entree                              (16)
With 2 Grams of Fat and 350 Calories
Now for Only $2.59.

Beef Entree                                (17)
With 2 Grams of Fat and 250 Calories
Now for Only $1.99.

Chicken Entree                             (18)
With 8 Grams of Fat and 250 Calories
Now for Only $1.99.
```

Next, we print the stimuli, produce the cards, and ask a subject to sort the cards from most preferred to least preferred. The combination numbers (most preferred to least preferred) are entered as data. For example, this subject's most preferred combination is 17, which is the "Beef Entree, With 2 Grams of Fat and 250 Calories, Now for Only $1.99", and her least preferred combination is 18, "Chicken Entree, With 8 Grams of Fat and 250 Calories, Now for Only $1.99".

Data Processing

The data are transposed, going from one observation and 18 variables to 18 observations and one variable named `Combo`. The next DATA step creates the variable `Rank`: 1 for the first and most preferred combination, ..., and 18 for the last and least preferred combination. The data are sorted by combination number and merged with the design.

```
title 'Frozen Diet Entrees';

data results;
   input combo1-combo18;
   datalines;
17 6 8 7 10 5 4 16 15 1 11 2 9 14 12 13 3 18
;
proc transpose out=results(rename=(col1=combo)); run;

data results; set results; Rank = _n_; drop _name_; run;

proc sort; by combo; run;

data results(drop=combo);
   merge sasuser.dietdes results;
   run;
proc print; run;
```

Frozen Diet Entrees

Obs	Ingredient	Fat	Price	Calories	Rank
1	Turkey	5 Grams	$1.99	350	10
2	Turkey	8 Grams	$2.29	350	12
3	Chicken	8 Grams	$1.99	350	17
4	Turkey	2 Grams	$2.59	250	7
5	Beef	8 Grams	$2.59	350	6
6	Beef	2 Grams	$1.99	350	2
7	Beef	5 Grams	$2.29	350	4
8	Beef	5 Grams	$2.29	250	3
9	Chicken	2 Grams	$2.29	350	13
10	Beef	8 Grams	$2.59	250	5
11	Turkey	8 Grams	$2.29	250	11
12	Chicken	5 Grams	$2.59	350	15
13	Chicken	5 Grams	$2.59	250	16
14	Chicken	2 Grams	$2.29	250	14
15	Turkey	5 Grams	$1.99	250	9
16	Turkey	2 Grams	$2.59	350	8
17	Beef	2 Grams	$1.99	250	1
18	Chicken	8 Grams	$1.99	250	18

Recall that the seventeenth combination was most preferred, and it has a rank of 1. The eighteenth combination was least preferred and it has a rank of 18.

Nonmetric Conjoint Analysis

PROC TRANSREG is used to perform the nonmetric conjoint analysis of the ranks.

```
ods exclude notes anova liberalanova conservanova
            mvanova liberalmvanova conservmvanova;
proc transreg utilities order=formatted separators=', ';
   model monotone(rank / reflect) =
         class(Ingredient Fat Price Calories / zero=sum);
   output out=utils p ireplace;
   run;
```

The `utilities` option prints the part-worth utilities and importance table. The `order=formatted` option sorts the levels of the attributes by the formatted values. By default, levels are sorted by their internal unformatted values (in this case the integers 1, 2, 3). The `model` statement names the variable `Rank` as the dependent variable and specifies a `monotone` transformation for the nonmetric conjoint analysis. The `reflect` transformation option is specified with rank data. With rank data, small values mean high preference and large values mean low preference. The `reflect` transformation option reflects the ranks around their mean (-(rank - mean rank) + mean rank) so that in the results, large part-worth utilities will mean high preference. With ranks ranging from 1 to 18, `reflect` transforms 1 to 18, 2 to 17, ..., r to $(19 - r)$, ..., and 18 to 1. (Note that the mean rank is the midpoint, in this case $(18+1)/2 = 9.5$, and $-(r-\bar{r})+\bar{r} = 2\bar{r}-r = 2(\max(r)+\min(r))/2-r = 19-r$.) The `class` specification names the attributes and scales the part-worth utilities to sum to zero within each attribute.

The `output` statement creates the `out=` data set, which contains the original variables, transformed variables, and indicator variables. The predicted utilities for all combinations are written to this data set by the `p` option (for predicted values). The `ireplace` option specifies that the transformed independent variables replace the original independent variables, since both are the same.

Here are the results of the conjoint analysis. Recall that we used an `ods exclude` statement and we used PROC TEMPLATE on page 366 to customize the output from PROC TRANSREG.

Frozen Diet Entrees

The TRANSREG Procedure

Dependent Variable Monotone(Rank)

Class Level Information

Class	Levels	Values
Ingredient	3	Beef Chicken Turkey
Fat	3	2 Grams 5 Grams 8 Grams
Price	3	$1.99 $2.29 $2.59
Calories	2	250 350

Number of Observations Read	18
Number of Observations Used	18

TRANSREG Univariate Algorithm Iteration History for Monotone(Rank)

Iteration Number	Average Change	Maximum Change	R-Square	Criterion Change	Note
1	0.07276	0.10014	0.99174		
2	0.00704	0.01074	0.99977	0.00802	
3	0.00468	0.00710	0.99990	0.00013	
4	0.00311	0.00470	0.99995	0.00006	
5	0.00207	0.00312	0.99998	0.00003	
6	0.00138	0.00208	0.99999	0.00001	
7	0.00092	0.00138	1.00000	0.00001	
8	0.00061	0.00092	1.00000	0.00000	
9	0.00041	0.00061	1.00000	0.00000	
10	0.00027	0.00041	1.00000	0.00000	
11	0.00018	0.00027	1.00000	0.00000	
12	0.00012	0.00018	1.00000	0.00000	
13	0.00008	0.00012	1.00000	0.00000	
14	0.00005	0.00008	1.00000	0.00000	
15	0.00004	0.00005	1.00000	0.00000	
16	0.00002	0.00004	1.00000	0.00000	
17	0.00002	0.00002	1.00000	0.00000	
18	0.00001	0.00002	1.00000	0.00000	
19	0.00001	0.00001	1.00000	0.00000	Converged

Algorithm converged.

Frozen Diet Entrees

The TRANSREG Procedure

The TRANSREG Procedure Hypothesis Tests for Monotone(Rank)

Root MSE	0.00007166	R-Square	1.0000
Dependent Mean	9.50000	Adj R-Sq	1.0000
Coeff Var	0.00075429		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	9.5000	0.00002	
Ingredient, Beef	6.0281	0.00002	74.999
Ingredient, Chicken	-6.0281	0.00002	
Ingredient, Turkey	-0.0000	0.00002	
Fat, 2 Grams	2.0094	0.00002	25.000
Fat, 5 Grams	0.0000	0.00002	
Fat, 8 Grams	-2.0094	0.00002	
Price, $1.99	0.0000	0.00002	0.000
Price, $2.29	0.0000	0.00002	
Price, $2.59	-0.0000	0.00002	
Calories, 250	0.0001	0.00002	0.001
Calories, 350	-0.0001	0.00002	

The standard errors are not adjusted for the fact that the dependent variable was transformed and so are generally liberal (too small).

We see in the conjoint output that main ingredient was the most important attribute at almost 75% and that beef was preferred over turkey, which was preferred over chicken. We also see that fat content was the second most important attribute at 25% and lower fat is preferred over higher fat. Price and calories only account for essentially none of the preference.

Next, the products in the `out=` data set are sorted by their predicted utility and the combinations are printed along with their rank, transformed and reflected rank, and predicted values (predicted utility). The variable `Rank` is the original rank variable; `TRank` contains the transformation of rank, in this case the reflection and monotonic transformation; and `PRank` contains the predicted utilities or predicted values. The first letter of the variable name comes from the first letter of "Transformation" and "Predicted".

```
proc sort; by descending prank; run;

proc print label;
   var ingredient fat price calories rank trank prank;
   label trank = 'Reflected Rank'
         prank = 'Utilities';
   run;
```

Frozen Diet Entrees

Obs	Ingredient	Fat	Price	Calories	Rank	Reflected Rank	Utilities
1	Beef	2 Grams	$1.99	250	1	17.5375	17.5375
2	Beef	2 Grams	$1.99	350	2	17.5373	17.5373
3	Beef	5 Grams	$2.29	250	3	15.5282	15.5281
4	Beef	5 Grams	$2.29	350	4	15.5279	15.5280
5	Beef	8 Grams	$2.59	250	5	13.5188	13.5188
6	Beef	8 Grams	$2.59	350	6	13.5186	13.5186
7	Turkey	2 Grams	$2.59	250	7	11.5095	11.5094
8	Turkey	2 Grams	$2.59	350	8	11.5092	11.5093
9	Turkey	5 Grams	$1.99	250	9	9.5001	9.5001
10	Turkey	5 Grams	$1.99	350	10	9.4999	9.4999
11	Turkey	8 Grams	$2.29	250	11	7.4908	7.4907
12	Turkey	8 Grams	$2.29	350	12	7.4905	7.4906
13	Chicken	2 Grams	$2.29	250	14	5.4813	5.4814
14	Chicken	2 Grams	$2.29	350	13	5.4813	5.4812
15	Chicken	5 Grams	$2.59	250	16	3.4719	3.4720
16	Chicken	5 Grams	$2.59	350	15	3.4719	3.4719
17	Chicken	8 Grams	$1.99	250	18	1.4626	1.4627
18	Chicken	8 Grams	$1.99	350	17	1.4626	1.4625

It is interesting to see that the sorted combinations support the information in the utilities table. The combinations are perfectly sorted on beef, turkey, and chicken. Furthermore, within ties in the main ingredient, the products are sorted by fat content.

Frozen Diet Entrées Example (Advanced)

This example is an advanced version of the previous example. It illustrates conjoint analysis with more than one subject. It has six parts.

- The `%MktEx` macro is used to generate an experimental design.
- Holdout observations are generated.
- The descriptions of the products are printed for data collection.
- The data are collected, entered, and processed.
- The metric conjoint analysis is performed.
- Results are summarized across subjects.

Creating a Design with the %MktEx Macro

The first thing you need to do in a conjoint study is decide on the product attributes and levels. Then you create the experimental design. We will use the same experimental design as we used in the previous example. The attributes and levels are shown in the table.

Factor	Levels		
Main Ingredient	Chicken	Beef	Turkey
Fat Claim Per Serving	8 Grams	5 Grams	2 Grams
Price	$2.59	$2.29	$1.99
Calories	350	250	

We will create our designs in the same way as we did in the previous example, starting on page 378. Only the random number seed has changed. Like before, we use the `%MktEval` macro to check the one-way and two-way frequencies and to ensure that each combination only appears once. See page 479 for macro documentation and information on installing and using SAS autocall macros.

```
title 'Frozen Diet Entrees';

proc format;
   value if  1='Chicken'  2='Beef'     3='Turkey';
   value ff  1='8 Grams'  2='5 Grams'  3='2 Grams';
   value pf  1='$2.59'    2='$2.29'    3='$1.99';
   value cf  1='350'      2='250';
   run;
%mktex(3 3 3 2, n=18, seed=205)
%mktlab(vars=Ingredient Fat Price Calories)
%mkteval;
```

Frozen Diet Entrees

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	100.0000	100.0000	Tab
1	End	100.0000		

Frozen Diet Entrees

The OPTEX Procedure

Class Level Information

Class	Levels	-Values-
x1	3	1 2 3
x2	3	1 2 3
x3	3	1 2 3
x4	2	1 2

Frozen Diet Entrees

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	100.0000	100.0000	100.0000	0.6667

Frozen Diet Entrees

Canonical Correlations Between the Factors

There are 0 Canonical Correlations Greater Than 0.316

	Ingredient	Fat	Price	Calories
Ingredient	1	0	0	0
Fat	0	1	0	0
Price	0	0	1	0
Calories	0	0	0	1

Frozen Diet Entrees
Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316

	Frequencies
Ingredient	6 6 6
Fat	6 6 6
Price	6 6 6
Calories	9 9
Ingredient Fat	2 2 2 2 2 2 2 2 2
Ingredient Price	2 2 2 2 2 2 2 2 2
Ingredient Calories	3 3 3 3 3 3
Fat Price	2 2 2 2 2 2 2 2 2
Fat Calories	3 3 3 3 3 3
Price Calories	3 3 3 3 3 3
N-Way	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

This design is 100% efficient, perfectly balanced and orthogonal, and each product occurs exactly once.

Designing Holdouts

The next steps add *holdout* observations to the design and display the results. Holdouts are ranked by the subjects but are analyzed with zero weight to exclude them from contributing to the utility computations. The correlation between the ranks for holdouts and their predicted utilities provide an indication of the validity of the results of the study.

The first `%MktEx` step recreates the formats and the design (just so you can see all of the code for a design with holdouts in one step). The next `%MktEx` step adds four holdouts to the randomized design created from the previous step. The specification `options=nodups` (no duplicates) ensures that the holdouts do not match products already in the design. The first `%MktEval` step evaluates just the original design, excluding the holdouts. The second `%MktEval` step evaluates the entire design. Both `%MktEval` steps ensure that the variable `w`, which flags the active and holdout observations, is excluded and not treated as a factor. The `%MktLab` step gives the factors informative names and assigns formats. Unlike the previous examples, this time we directly assign the formats in the `%MktLab` macro using the `statements=` option, specifying a complete `format` statement.

```
title 'Frozen Diet Entrees';

proc format;
   value if  1='Chicken'  2='Beef'     3='Turkey';
   value ff  1='8 Grams'  2='5 Grams'  3='2 Grams';
   value pf  1='$2.59'    2='$2.29'    3='$1.99';
   value cf  1='350'      2='250';
   run;

%mktex(3 3 3 2, n=18, seed=205)

%mktex(3 3 3 2, n=22, init=randomized, holdouts=4, options=nodups, seed=368)
proc print data=randomized; run;

%mkteval(data=randomized(where=(w=1)), factors=x:);
%mkteval(data=randomized(drop=w));

%mktlab(data=randomized, out=sasuser.dietdes,
        vars=Ingredient Fat Price Calories,
        statements=format Ingredient if. fat ff. price pf. calories cf.)

proc print; run;
```

Here is the last part of the output from the first %MktEx step, which shows that the macro found a 100% efficient design.

Frozen Diet Entrees

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	100.0000	100.0000	100.0000	0.6667

Next, is some of the output from the %MktEx step that finds the holdouts. Notice that the macro immediately enters the design refinement step.

Frozen Diet Entrees

Design Refinement History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	98.0764		Ini
1	Start	98.0764		Pre,Mut,Ann
1	22 1	98.2421	98.2421	Conforms
1	End	98.2421		
2	Start	98.2421		Pre,Mut,Ann
2	2 1	98.2421	98.2421	Conforms
2	End	98.2421		
3	Start	98.2421		Pre,Mut,Ann
3	2 1	98.2421	98.2421	Conforms
3	End	98.2421		
4	Start	98.2421		Pre,Mut,Ann
4	2 1	98.2421	98.2421	Conforms
4	End	98.2421		
5	Start	98.2421		Pre,Mut,Ann
5	2 1	98.2421	98.2421	Conforms
5	End	98.2421		

NOTE: Stopping since it appears that no improvement is possible.

Next, the raw design is printed. Observations with `w` equal to 1 comprise the original design. The observations with a missing `w` are the holdouts.

Frozen Diet Entrees

Obs	x1	x2	x3	x4	w
1	2	3	1	2	.
2	2	2	1	2	1
3	3	3	3	1	1
4	3	3	3	2	1
5	3	1	1	1	1
6	1	3	1	1	1

7	1	3	1	2	1
8	1	1	2	1	1
9	2	2	1	1	1
10	3	2	2	2	1
11	2	2	3	1	.
12	2	3	2	2	1
13	3	1	1	2	1
14	2	1	3	2	1
15	3	2	1	1	.
16	1	2	3	1	1
17	1	1	2	2	1
18	1	2	2	2	.
19	2	3	2	1	1
20	3	2	2	1	1
21	1	2	3	2	1
22	2	1	3	1	1

Here is the evaluation of the original design.

Frozen Diet Entrees
Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

	x1	x2	x3	x4
x1	1	0	0	0
x2	0	1	0	0
x3	0	0	1	0
x4	0	0	0	1

Frozen Diet Entrees
Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316

Frequencies

x1	6 6 6
x2	6 6 6
x3	6 6 6
x4	9 9
x1 x2	2 2 2 2 2 2 2 2 2
x1 x3	2 2 2 2 2 2 2 2 2
x1 x4	3 3 3 3 3 3
x2 x3	2 2 2 2 2 2 2 2 2
x2 x4	3 3 3 3 3 3
x3 x4	3 3 3 3 3 3
N-Way	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Here is an evaluation of the design with the holdouts.

```
                    Frozen Diet Entrees
          Canonical Correlations Between the Factors
There are 0 Canonical Correlations Greater Than 0.316

                  x1      x2      x3      x4

          x1      1       0.09    0.17    0.11
          x2      0.09    1       0.09    0.11
          x3      0.17    0.09    1       0.11
          x4      0.11    0.11    0.11    1

                    Frozen Diet Entrees
                   Summary of Frequencies
There are 0 Canonical Correlations Greater Than 0.316
          * - Indicates Unequal Frequencies

                Frequencies

 *      x1       7 8 7
 *      x2       6 9 7
 *      x3       8 7 7
        x4       11 11
 *      x1 x2    2 3 2 2 3 3 2 3 2
 *      x1 x3    2 3 2 3 2 3 3 2 2
 *      x1 x4    3 4 4 4 4 3
 *      x2 x3    2 2 2 3 3 3 3 2 2
 *      x2 x4    3 3 5 4 3 4
 *      x3 x4    4 4 3 4 4 3
        N-Way    1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                 1 1 1
```

Here is the design, printed with descriptive factor names and formats.

Frozen Diet Entrees

Obs	Ingredient	Fat	Price	Calories	w
1	Beef	2 Grams	$2.59	250	.
2	Beef	5 Grams	$2.59	250	1
3	Turkey	2 Grams	$1.99	350	1
4	Turkey	2 Grams	$1.99	250	1
5	Turkey	8 Grams	$2.59	350	1
6	Chicken	2 Grams	$2.59	350	1
7	Chicken	2 Grams	$2.59	250	1
8	Chicken	8 Grams	$2.29	350	1
9	Beef	5 Grams	$2.59	350	1
10	Turkey	5 Grams	$2.29	250	1
11	Beef	5 Grams	$1.99	350	.
12	Beef	2 Grams	$2.29	250	1
13	Turkey	8 Grams	$2.59	250	1
14	Beef	8 Grams	$1.99	250	1
15	Turkey	5 Grams	$2.59	350	.
16	Chicken	5 Grams	$1.99	350	1
17	Chicken	8 Grams	$2.29	250	1
18	Chicken	5 Grams	$2.29	250	.
19	Beef	2 Grams	$2.29	350	1
20	Turkey	5 Grams	$2.29	350	1
21	Chicken	5 Grams	$1.99	250	1
22	Beef	8 Grams	$1.99	350	1

Print the Stimuli

Once the design is generated, the *stimuli* (descriptions of the combinations) must be generated for data collection. They are printed using the exact same step as we used on page 382.

```
title;
data _null_;
   file print;
   set sasuser.dietdes;
   put ///
       +3 ingredient 'Entree' @50 '(' _n_ +(-1) ')' /
       +3 'With ' fat 'of Fat and ' calories  'Calories' /
       +3 'Now for Only ' Price +(-1) '.'///;
   if mod(_n_, 6) = 0 then put _page_;
   run;
```

In the interest of space, only the first three are shown.

```
Beef Entree                                    (1)
With 2 Grams of Fat and 250 Calories
Now for Only $2.59.

Beef Entree                                    (2)
With 5 Grams of Fat and 250 Calories
Now for Only $2.59.

Turkey Entree                                  (3)
With 2 Grams of Fat and 350 Calories
Now for Only $1.99.
```

Data Collection, Entry, and Preprocessing

The next step in the conjoint analysis study is data collection and entry. Each subject was asked to take the 22 cards and rank them from the most preferred combination to the least preferred combination. The combination numbers are entered as data. The data follow the `datalines` statement in the next DATA step. For the first subject, 4 was most preferred, 3 was second most preferred, ..., and 5 was the least preferred combination. The DATA step validates the data entry and converts the input to ranks.

```
title 'Frozen Diet Entrees';

%let m = 22; /* number of combinations */

* Read the input data and convert to ranks;
data ranks(drop=i k c1-c&m);
   input c1-c&m;
   array c[&m];
   array r[&m];
   do i = 1 to &m;
      k = c[i];
      if 1 le k le &m then do;
         if r[k] ne . then
            put 'ERROR: For subject ' _n_ +(-1) ', combination ' k
                'is given more than once.';
         r[k] = i; /* Convert to ranks. */
         end;
      else put 'ERROR: For subject ' _n_ +(-1) ', combination ' k
                'is invalid.';
      end;
```

```
   do i = 1 to &m;
      if r[i] = . then
         put 'ERROR: For subject ' _n_ +(-1) ', combination ' i
             'is not given.';
      end;
   name = 'Subj' || put(_n_, z2.);
   datalines;
4 3 7 21 12 10 6 19 1 16 18 11 20 14 17 15 2 22 9 8 13 5
4 12 3 1 19 7 10 6 11 21 16 2 18 20 15 9 14 22 13 17 5 8
4 3 7 12 19 21 1 6 10 18 16 11 20 15 2 14 9 17 22 8 13 5
4 12 1 10 21 14 18 3 7 2 17 13 19 11 22 20 16 15 6 9 5 8
4 21 14 11 16 3 12 22 19 18 10 17 8 20 7 1 6 2 9 13 15 5
4 21 16 12 3 14 11 22 18 19 7 10 1 17 8 6 2 20 9 13 15 5
12 4 19 1 3 7 6 21 18 11 16 2 10 20 9 15 14 17 22 8 13 5
4 21 3 16 14 11 12 22 18 10 19 20 17 8 7 6 1 2 13 15 9 5
4 21 3 16 11 14 22 12 18 10 20 19 17 8 7 6 1 13 15 2 9 5
4 3 14 11 21 12 16 22 19 10 18 20 17 1 7 8 2 13 9 6 15 5
15 22 17 21 6 11 13 19 4 12 3 18 9 7 1 10 8 20 14 16 5 2
12 4 3 7 21 19 1 18 11 6 16 2 14 10 17 22 20 9 15 8 13 5
;
```

The macro variable `&m` is set to 22, the number of combinations. This is done to make it easier to modify the code for future use with different sized studies. For each subject, the numbers of the 22 products are read into the variables `c1` through `c22`. The do loop, `do i = 1 to &m`, loops over each of the products. Consider the first product: `k` is set to `c[i]`, which is `c[1]`, which is 4 since the fourth product was ranked first by the first subject. The first data integrity check, `if 1 le k le &m then do` ensures that the number is in the valid range, 1 to 22. Otherwise an error is printed. Since the number is valid, `r[k]` is checked to see if it is missing. If it is not missing, another error is printed. The array `r` consists of 22 variables `r1` through `22`. These variables start out each pass through the DATA step as missing and end up as the ranks. If `r[k] eq .`, then the *kth* combination has not had a rank assigned yet so everything is fine. If `r[k] ne .`, the same number appears twice in a subject's data so there is something wrong with the data entry. The statement `r[k] = i` assigns the ranks. For subject 1 and the first product, `k = c[i] = c[1] = 4` so the rank of the fourth product is set to 1 (`r[k] = r[4] = i = 1`). For subject 1 and the second product, `k = c[i] = c[2] = 3` so the rank of the third product is set to 2 (`r[k] = r[3] = i = 2`). For subject 1 and the last product, `k = c[i] = c[22] = 5` so the rank of the fifth product is set to 22 (`r[k] = r[5] = i = 22`). At the end of the `do i = 1 to &m` loop, each of the 22 variables in `r1-r22` should have been set to exactly one rank. If any of these variables are missing, then one or more product numbers did not appear in the data, so this is flagged as an error. The statement `name = 'Subj' || put(_n_, z2.)` creates a subject ID of the form `Subj01`, `Subj02`, ..., `Subj12`.

Say there was a mistake in data entry for the first subject − say product 17 had been entered as 7 instead of 17. We would get the following error messages.

```
ERROR: For subject 1, combination 7 is given more than once.
ERROR: For subject 1, combination 17 is not given.
```

If for the first subject, the 17 had been entered as 117 instead of 17, we would get the following error messages.

```
ERROR: For subject 1, combination 117 is invalid.
ERROR: For subject 1, combination 17 is not given.
```

The next step transposes the data set from one row per subject to one row per product. The `id name` statement on PROC TRANSPOSE names the rank variables `Subj01` through `Subj12`. Later, we will need to sort by these names. That is why we used leading zeros and names like `Subj01` instead of `Subj1`. Next, the input data set is merged with the design.

```
proc transpose data=ranks out=ranks2;
   id name;
   run;
data both;
   merge sasuser.dietdes ranks2;
   drop _name_;
   run;
proc print label;
   title2 'Data and Design Together';
   run;
```

Frozen Diet Entrees
Data and Design Together

Obs	Ingredient	Fat	Price	Calories	w	Subj01	Subj02	Subj03	Subj04
1	Beef	2 Grams	$2.59	250	.	9	4	7	3
2	Beef	5 Grams	$2.59	250	1	17	12	15	10
3	Turkey	2 Grams	$1.99	350	1	2	3	2	8
4	Turkey	2 Grams	$1.99	250	1	1	1	1	1
5	Turkey	8 Grams	$2.59	350	1	22	21	22	21
6	Chicken	2 Grams	$2.59	350	1	7	8	8	19
7	Chicken	2 Grams	$2.59	250	1	3	6	3	9
8	Chicken	8 Grams	$2.29	350	1	20	22	20	22
9	Beef	5 Grams	$2.59	350	1	19	16	17	20
10	Turkey	5 Grams	$2.29	250	1	6	7	9	4
11	Beef	5 Grams	$1.99	350	.	12	9	12	14
12	Beef	2 Grams	$2.29	250	1	5	2	4	2
13	Turkey	8 Grams	$2.59	250	1	21	19	21	12
14	Beef	8 Grams	$1.99	250	1	14	17	16	6
15	Turkey	5 Grams	$2.59	350	.	16	15	14	18
16	Chicken	5 Grams	$1.99	350	1	10	11	11	17
17	Chicken	8 Grams	$2.29	250	1	15	20	18	11
18	Chicken	5 Grams	$2.29	250	.	11	13	10	7
19	Beef	2 Grams	$2.29	350	1	8	5	5	13
20	Turkey	5 Grams	$2.29	350	1	13	14	13	16
21	Chicken	5 Grams	$1.99	250	1	4	10	6	5
22	Beef	8 Grams	$1.99	350	1	18	18	19	15

Obs	Subj05	Subj06	Subj07	Subj08	Subj09	Subj10	Subj11	Subj12
1	16	13	4	17	17	14	15	7
2	18	17	12	18	20	17	22	12
3	6	5	5	3	3	2	11	3
4	1	1	2	1	1	1	9	2
5	22	22	22	22	22	22	21	22
6	17	16	7	16	16	20	5	10
7	15	11	6	15	15	15	14	4
8	13	15	20	14	14	16	17	20
9	19	19	15	21	21	19	13	18
10	11	12	13	10	10	10	16	14
11	4	7	10	6	5	4	6	9
12	7	4	1	7	8	6	10	1
13	20	20	21	19	18	18	7	21
14	3	6	17	5	6	3	19	13
15	21	21	16	20	19	21	1	19
16	5	3	11	4	4	7	20	11
17	12	14	18	13	13	13	3	15
18	10	9	9	9	9	11	12	8
19	9	10	3	11	12	9	8	6
20	14	18	14	12	11	12	18	17
21	2	2	8	2	2	5	4	5
22	8	8	19	8	7	8	2	16

One more data set manipulation is sometimes necessary − the addition of *simulation* observations. Simulation observations are not rated by the subjects and do not contribute to the analysis. They are scored as passive observations. Simulations are *what-if* combinations. They are combinations that are entered to get a prediction of what their utility would have been if they had been rated. In this example, all combinations are added as simulations. The `%MktEx` macro is called to make a full-factorial design. The `n=` specification accepts expressions, so `n=3*3*3*2` and `n=54` are equivalent. The `data all` step reads in the design and data followed by the simulation observations. The flag variable `f` indicates when the simulation observations are being processed. Simulation observations are given a weight of 0 to exclude them from the analysis and to distinguish them from the holdouts. Notice that the dependent variable has missing values for the simulations and nonmissing values for the holdouts and active observations.

```
proc format;
   value wf 1 = 'Active'
            . = 'Holdout'
            0 = 'Simulation';
   run;

%mktex(3 3 3 2, n=3*3*3*2)
%mktlab(data=design, vars=Ingredient Fat Price Calories)

data all;
   set both final(in=f);
   if f then w = 0;
   format w wf.;
   run;
```

```
proc print data=all(Obs=25 drop=subj04-subj12) label;
   title2 'Some of the Final Data Set';
   run;
```

Here the data for the first three subjects and the first 25 rows of the data set.

Frozen Diet Entrees
Some of the Final Data Set

Obs	Ingredient	Fat	Price	Calories	w	Subj01	Subj02	Subj03
1	Beef	2 Grams	$2.59	250	Holdout	9	4	7
2	Beef	5 Grams	$2.59	250	Active	17	12	15
3	Turkey	2 Grams	$1.99	350	Active	2	3	2
4	Turkey	2 Grams	$1.99	250	Active	1	1	1
5	Turkey	8 Grams	$2.59	350	Active	22	21	22
6	Chicken	2 Grams	$2.59	350	Active	7	8	8
7	Chicken	2 Grams	$2.59	250	Active	3	6	3
8	Chicken	8 Grams	$2.29	350	Active	20	22	20
9	Beef	5 Grams	$2.59	350	Active	19	16	17
10	Turkey	5 Grams	$2.29	250	Active	6	7	9
11	Beef	5 Grams	$1.99	350	Holdout	12	9	12
12	Beef	2 Grams	$2.29	250	Active	5	2	4
13	Turkey	8 Grams	$2.59	250	Active	21	19	21
14	Beef	8 Grams	$1.99	250	Active	14	17	16
15	Turkey	5 Grams	$2.59	350	Holdout	16	15	14
16	Chicken	5 Grams	$1.99	350	Active	10	11	11
17	Chicken	8 Grams	$2.29	250	Active	15	20	18
18	Chicken	5 Grams	$2.29	250	Holdout	11	13	10
19	Beef	2 Grams	$2.29	350	Active	8	5	5
20	Turkey	5 Grams	$2.29	350	Active	13	14	13
21	Chicken	5 Grams	$1.99	250	Active	4	10	6
22	Beef	8 Grams	$1.99	350	Active	18	18	19
23	Chicken	8 Grams	$2.59	350	Simulation	.	.	.
24	Chicken	8 Grams	$2.59	250	Simulation	.	.	.
25	Chicken	8 Grams	$2.29	350	Simulation	.	.	.

Metric Conjoint Analysis

In this part of this example, the conjoint analysis is performed with PROC TRANSREG.

```
ods exclude notes mvanova anova;
proc transreg data=all utilities short separators=', '
   method=morals outtest=utils;
   title2 'Conjoint Analysis';
   model identity(subj: / reflect) =
         class(Ingredient Fat Price Calories / zero=sum);
   weight w;
   output p ireplace out=results coefficients;
   run;
```

The `proc`, `model`, and `output` statements are typical for a conjoint analysis of rank-order data with more than one subject. (In this analysis, we perform a metric conjoint analysis. It is more typical to perform nonmetric conjoint analysis of rank-order data. However, it is not absolutely required.) The `proc` statement specifies `method=morals`, which fits the conjoint analysis model separately for each subject. The `proc` statement also requests an `outtest=` data set, which contains the ANOVA and part-worth utilities tables from the printed output. In the `model` statement, the dependent variable list `subj:` specifies all variables in the DATA= data set that begin with the prefix `subj` (in this case `subj01-subj12`). The `weight` variable designates the active (`weight` = 1), holdout (`weight` = .), and simulation (`weight` = 0) observations. Only the active observations are used to compute the part-worth utilities. However, predicted utilities are computed for all observations, including active, holdouts, and simulations, using those part-worths. The `output` statement creates an `out=` data set beginning with all results for the first subject, followed by all subject two results, and so on.

Here are the results. Recall that we used an `ods exclude` statement and we used PROC TEMPLATE on page 366 to customize the output from PROC TRANSREG. There is one set of output for each subject. Conjoint analysis fits individual-level models.

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Class Level Information

Class	Levels	Values
Ingredient	3	Chicken Beef Turkey
Fat	3	8 Grams 5 Grams 2 Grams
Price	3	$2.59 $2.29 $1.99
Calories	2	350 250

Number of Observations Read	76
Number of Observations Used	18
Sum of Weights Read	18
Sum of Weights Used	18

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj01)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj01)

Root MSE	1.81046	R-Square	0.9618
Dependent Mean	11.38889	Adj R-Sq	0.9351
Coeff Var	15.89675		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.3889	0.42673	
Ingredient, Chicken	1.5556	0.60349	13.095
Ingredient, Beef	-2.1111	0.60349	
Ingredient, Turkey	0.5556	0.60349	
Fat, 8 Grams	-6.9444	0.60349	50.000
Fat, 5 Grams	-0.1111	0.60349	
Fat, 2 Grams	7.0556	0.60349	
Price, $2.59	-3.4444	0.60349	23.810
Price, $2.29	0.2222	0.60349	
Price, $1.99	3.2222	0.60349	
Calories, 350	-1.8333	0.42673	13.095
Calories, 250	1.8333	0.42673	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj02)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj02)

Root MSE	1.30809	R-Square	0.9788
Dependent Mean	11.77778	Adj R-Sq	0.9640
Coeff Var	11.10646		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.7778	0.30832	
Ingredient, Chicken	-1.0556	0.43603	8.451
Ingredient, Beef	0.1111	0.43603	
Ingredient, Turkey	0.9444	0.43603	
Fat, 8 Grams	-7.7222	0.43603	64.789
Fat, 5 Grams	0.1111	0.43603	
Fat, 2 Grams	7.6111	0.43603	
Price, $2.59	-1.8889	0.43603	15.493
Price, $2.29	0.1111	0.43603	
Price, $1.99	1.7778	0.43603	
Calories, 350	-1.3333	0.30832	11.268
Calories, 250	1.3333	0.30832	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj03)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj03)

Root MSE	1.15470	R-Square	0.9844
Dependent Mean	11.66667	Adj R-Sq	0.9735
Coeff Var	9.89743		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.6667	0.27217	
Ingredient, Chicken	0.6667	0.38490	6.667
Ingredient, Beef	-1.0000	0.38490	
Ingredient, Turkey	0.3333	0.38490	
Fat, 8 Grams	-7.6667	0.38490	62.000
Fat, 5 Grams	-0.1667	0.38490	
Fat, 2 Grams	7.8333	0.38490	
Price, $2.59	-2.6667	0.38490	20.667
Price, $2.29	0.1667	0.38490	
Price, $1.99	2.5000	0.38490	
Calories, 350	-1.3333	0.27217	10.667
Calories, 250	1.3333	0.27217	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj04)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj04)

Root MSE	1.05935	R-Square	0.9849
Dependent Mean	11.72222	Adj R-Sq	0.9743
Coeff Var	9.03711		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.7222	0.24969	
Ingredient, Chicken	-2.1111	0.35312	13.490
Ingredient, Beef	0.7222	0.35312	
Ingredient, Turkey	1.3889	0.35312	
Fat, 8 Grams	-2.7778	0.35312	22.484
Fat, 5 Grams	-0.2778	0.35312	
Fat, 2 Grams	3.0556	0.35312	
Price, $2.59	-3.4444	0.35312	25.054
Price, $2.29	0.3889	0.35312	
Price, $1.99	3.0556	0.35312	
Calories, 350	-5.0556	0.24969	38.972
Calories, 250	5.0556	0.24969	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj05)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj05)

Root MSE	1.02198	R-Square	0.9854
Dependent Mean	11.22222	Adj R-Sq	0.9752
Coeff Var	9.10676		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.2222	0.24088	
Ingredient, Chicken	0.5556	0.34066	7.407
Ingredient, Beef	0.5556	0.34066	
Ingredient, Turkey	-1.1111	0.34066	
Fat, 8 Grams	-1.7778	0.34066	17.037
Fat, 5 Grams	-0.2778	0.34066	
Fat, 2 Grams	2.0556	0.34066	
Price, $2.59	-7.2778	0.34066	63.704
Price, $2.29	0.2222	0.34066	
Price, $1.99	7.0556	0.34066	
Calories, 350	-1.3333	0.24088	11.852
Calories, 250	1.3333	0.24088	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj06)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj06)

Root MSE	1.67000	R-Square	0.9636
Dependent Mean	11.27778	Adj R-Sq	0.9381
Coeff Var	14.80785		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.2778	0.39362	
Ingredient, Chicken	1.1111	0.55667	11.015
Ingredient, Beef	0.6111	0.55667	
Ingredient, Turkey	-1.7222	0.55667	
Fat, 8 Grams	-2.8889	0.55667	24.622
Fat, 5 Grams	-0.5556	0.55667	
Fat, 2 Grams	3.4444	0.55667	
Price, $2.59	-6.2222	0.55667	51.836
Price, $2.29	-0.8889	0.55667	
Price, $1.99	7.1111	0.55667	
Calories, 350	-1.6111	0.39362	12.527
Calories, 250	1.6111	0.39362	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj07)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj07)

Root MSE	1.06979	R-Square	0.9857
Dependent Mean	11.88889	Adj R-Sq	0.9756
Coeff Var	8.99821		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.8889	0.25215	
Ingredient, Chicken	0.2222	0.35660	7.353
Ingredient, Beef	0.7222	0.35660	
Ingredient, Turkey	-0.9444	0.35660	
Fat, 8 Grams	-7.6111	0.35660	68.382
Fat, 5 Grams	-0.2778	0.35660	
Fat, 2 Grams	7.8889	0.35660	
Price, $2.59	-1.9444	0.35660	15.441
Price, $2.29	0.3889	0.35660	
Price, $1.99	1.5556	0.35660	
Calories, 350	-1.0000	0.25215	8.824
Calories, 250	1.0000	0.25215	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj08)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj08)

Root MSE	0.79582	R-Square	0.9915
Dependent Mean	11.16667	Adj R-Sq	0.9855
Coeff Var	7.12677		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.1667	0.18758	
Ingredient, Chicken	0.5000	0.26527	4.412
Ingredient, Beef	-0.5000	0.26527	
Ingredient, Turkey	0.0000	0.26527	
Fat, 8 Grams	-2.3333	0.26527	20.588
Fat, 5 Grams	0.0000	0.26527	
Fat, 2 Grams	2.3333	0.26527	
Price, $2.59	-7.3333	0.26527	64.706
Price, $2.29	-0.0000	0.26527	
Price, $1.99	7.3333	0.26527	
Calories, 350	-1.1667	0.18758	10.294
Calories, 250	1.1667	0.18758	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj09)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj09)

Root MSE	1.05935	R-Square	0.9850
Dependent Mean	11.27778	Adj R-Sq	0.9745
Coeff Var	9.39325		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.2778	0.24969	
Ingredient, Chicken	0.6111	0.35312	7.389
Ingredient, Beef	-1.0556	0.35312	
Ingredient, Turkey	0.4444	0.35312	
Fat, 8 Grams	-2.0556	0.35312	18.473
Fat, 5 Grams	-0.0556	0.35312	
Fat, 2 Grams	2.1111	0.35312	
Price, $2.59	-7.3889	0.35312	65.764
Price, $2.29	-0.0556	0.35312	
Price, $1.99	7.4444	0.35312	
Calories, 350	-0.9444	0.24969	8.374
Calories, 250	0.9444	0.24969	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj10)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj10)

Root MSE	0.90062	R-Square	0.9889
Dependent Mean	11.27778	Adj R-Sq	0.9812
Coeff Var	7.98577		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.2778	0.21228	
Ingredient, Chicken	-1.3889	0.30021	9.722
Ingredient, Beef	0.9444	0.30021	
Ingredient, Turkey	0.4444	0.30021	
Fat, 8 Grams	-2.0556	0.30021	18.750
Fat, 5 Grams	-0.3889	0.30021	
Fat, 2 Grams	2.4444	0.30021	
Price, $2.59	-7.2222	0.30021	59.028
Price, $2.29	0.2778	0.30021	
Price, $1.99	6.9444	0.30021	
Calories, 350	-1.5000	0.21228	12.500
Calories, 250	1.5000	0.21228	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj11)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj11)

Root MSE	7.42369	R-Square	0.2393
Dependent Mean	12.16667	Adj R-Sq	-0.2932
Coeff Var	61.01660		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	12.1667	1.74978	
Ingredient, Chicken	1.6667	2.47456	23.950
Ingredient, Beef	-0.1667	2.47456	
Ingredient, Turkey	-1.5000	2.47456	
Fat, 8 Grams	0.6667	2.47456	45.378
Fat, 5 Grams	-3.3333	2.47456	
Fat, 2 Grams	2.6667	2.47456	
Price, $2.59	-1.5000	2.47456	21.429
Price, $2.29	0.1667	2.47456	
Price, $1.99	1.3333	2.47456	
Calories, 350	-0.6111	1.74978	9.244
Calories, 250	0.6111	1.74978	

Frozen Diet Entrees
Conjoint Analysis

The TRANSREG Procedure

Identity(Subj12)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Subj12)

Root MSE	1.49443	R-Square	0.9717
Dependent Mean	11.66667	Adj R-Sq	0.9519
Coeff Var	12.80944		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	11.6667	0.35224	
Ingredient, Chicken	0.8333	0.49814	8.974
Ingredient, Beef	0.6667	0.49814	
Ingredient, Turkey	-1.5000	0.49814	
Fat, 8 Grams	-6.1667	0.49814	51.923
Fat, 5 Grams	-1.1667	0.49814	
Fat, 2 Grams	7.3333	0.49814	
Price, $2.59	-2.8333	0.49814	23.718
Price, $2.29	-0.5000	0.49814	
Price, $1.99	3.3333	0.49814	
Calories, 350	-2.0000	0.35224	15.385
Calories, 250	2.0000	0.35224	

Next, we will print some of the output data set to see the predicted utilities for the first two subjects.

```
proc print data=results(drop=_depend_ t_depend_ intercept &_trgind) label;
   title2 'Predicted Utility';
   where w ne 0 and _depvar_ le 'Identity(Subj02)' and not (_type_ =: 'M');
   by _depvar_;
   label p_depend_ = 'Predicted Utility';
   run;
```

We print _TYPE_, _NAME_, and the weight variable, w; drop the original and transformed dependent

variable, `_depend_` and `t_depend_`; print the predicted values (predicted utilities), `p_depend_`; drop the intercept and coded independent variables; and print the original `class` variables. Note that the macro variable `&_trgind` is automatically created by PROC TRANSREG and its value is a list of the names of the coded variables. The `where` statement is used to exclude the simulation observations and just show results for the first two subjects. Here are the predicted utilities for each of the rated products for the first two subjects.

Frozen Diet Entrees
Predicted Utility

---------- Dependent Variable Transformation(Name)=Identity(Subj01) ----------

Obs	_TYPE_	_NAME_	w	Predicted Utility	Ingredient	Fat	Price	Calories
1		ROW1	Holdout	14.7222	Beef	2 Grams	$2.59	250
2	SCORE	ROW2	Active	7.5556	Beef	5 Grams	$2.59	250
3	SCORE	ROW3	Active	20.3889	Turkey	2 Grams	$1.99	350
4	SCORE	ROW4	Active	24.0556	Turkey	2 Grams	$1.99	250
5	SCORE	ROW5	Active	-0.2778	Turkey	8 Grams	$2.59	350
6	SCORE	ROW6	Active	14.7222	Chicken	2 Grams	$2.59	350
7	SCORE	ROW7	Active	18.3889	Chicken	2 Grams	$2.59	250
8	SCORE	ROW8	Active	4.3889	Chicken	8 Grams	$2.29	350
9	SCORE	ROW9	Active	3.8889	Beef	5 Grams	$2.59	350
10	SCORE	ROW10	Active	13.8889	Turkey	5 Grams	$2.29	250
11		ROW11	Holdout	10.5556	Beef	5 Grams	$1.99	350
12	SCORE	ROW12	Active	18.3889	Beef	2 Grams	$2.29	250
13	SCORE	ROW13	Active	3.3889	Turkey	8 Grams	$2.59	250
14	SCORE	ROW14	Active	7.3889	Beef	8 Grams	$1.99	250
15		ROW15	Holdout	6.5556	Turkey	5 Grams	$2.59	350
16	SCORE	ROW16	Active	14.2222	Chicken	5 Grams	$1.99	350
17	SCORE	ROW17	Active	8.0556	Chicken	8 Grams	$2.29	250
18		ROW18	Holdout	14.8889	Chicken	5 Grams	$2.29	250
19	SCORE	ROW19	Active	14.7222	Beef	2 Grams	$2.29	350
20	SCORE	ROW20	Active	10.2222	Turkey	5 Grams	$2.29	350
21	SCORE	ROW21	Active	17.8889	Chicken	5 Grams	$1.99	250
22	SCORE	ROW22	Active	3.7222	Beef	8 Grams	$1.99	350

---------- Dependent Variable Transformation(Name)=Identity(Subj02) ----------

Obs	_TYPE_	_NAME_	w	Predicted Utility	Ingredient	Fat	Price	Calories
79		ROW1	Holdout	18.9444	Beef	2 Grams	$2.59	250
80	SCORE	ROW2	Active	11.4444	Beef	5 Grams	$2.59	250
81	SCORE	ROW3	Active	20.7778	Turkey	2 Grams	$1.99	350
82	SCORE	ROW4	Active	23.4444	Turkey	2 Grams	$1.99	250
83	SCORE	ROW5	Active	1.7778	Turkey	8 Grams	$2.59	350
84	SCORE	ROW6	Active	15.1111	Chicken	2 Grams	$2.59	350

85	SCORE	ROW7	Active	17.7778	Chicken	2 Grams	$2.59	250
86	SCORE	ROW8	Active	1.7778	Chicken	8 Grams	$2.29	350
87	SCORE	ROW9	Active	8.7778	Beef	5 Grams	$2.59	350
88	SCORE	ROW10	Active	14.2778	Turkey	5 Grams	$2.29	250
89		ROW11	Holdout	12.4444	Beef	5 Grams	$1.99	350
90	SCORE	ROW12	Active	20.9444	Beef	2 Grams	$2.29	250
91	SCORE	ROW13	Active	4.4444	Turkey	8 Grams	$2.59	250
92	SCORE	ROW14	Active	7.2778	Beef	8 Grams	$1.99	250
93		ROW15	Holdout	9.6111	Turkey	5 Grams	$2.59	350
94	SCORE	ROW16	Active	11.2778	Chicken	5 Grams	$1.99	350
95	SCORE	ROW17	Active	4.4444	Chicken	8 Grams	$2.29	250
96		ROW18	Holdout	12.2778	Chicken	5 Grams	$2.29	250
97	SCORE	ROW19	Active	18.2778	Beef	2 Grams	$2.29	350
98	SCORE	ROW20	Active	11.6111	Turkey	5 Grams	$2.29	350
99	SCORE	ROW21	Active	13.9444	Chicken	5 Grams	$1.99	250
100	SCORE	ROW22	Active	4.6111	Beef	8 Grams	$1.99	350

Analyzing Holdouts

The next steps display the correlations between the predicted utility for holdout observations and their actual ratings. These correlations provide a measure of the validity of the results, since the holdout observations have zero weight and do not contribute to any of the calculations. The Pearson correlations are the ordinary correlation coefficients, and the Kendall Tau's are rank-based measures of correlation. These correlations should always be large. Subjects whose correlations are small may be unreliable.

PROC CORR is used to produce the correlations. Since the output is not very compact, ODS is used to suppress the normal printed output (`ods listing close`), output the Pearson correlations to an output data set P (`PearsonCorr=p`), and output the Kendall correlations to an output data set K (`KendallCorr=k`). The listing is reopened for normal output (`ods listing`), the two tables are merged renaming the variables to identify the correlation type, the subject number is pulled out of the subject variable names, and the results are printed.

```
ods output KendallCorr=k PearsonCorr=p;
ods listing close;
proc corr nosimple noprob kendall pearson
   data=results(where=(w=.));
   title2 'Holdout Validation Results';
   var p_depend_;
   with t_depend_;
   by notsorted _depvar_;
   run;
ods listing;
```

```
data both(keep=subject pearson kendall);
   length Subject 8;
   merge p(rename=(p_depend_=Pearson))
         k(rename=(p_depend_=Kendall));
   subject = input(substr(_depvar_, 14, 2), best2.);
   run;

proc print; run;
```

Here are the results.

Frozen Diet Entrees
Holdout Validation Results

Obs	Subject	Pearson	Kendall
1	1	0.93848	0.66667
2	2	0.94340	1.00000
3	3	0.99038	1.00000
4	4	0.97980	1.00000
5	5	0.98930	1.00000
6	6	0.98649	1.00000
7	7	0.99029	1.00000
8	8	0.99296	1.00000
9	9	0.99873	1.00000
10	10	0.99973	1.00000
11	11	-0.98184	-1.00000
12	12	0.92920	1.00000

Most of the correlations look great! However, the results from subject 11 look suspect. Subject 11's holdout correlations are negative. We can return to page 414 and look at the conjoint results. Subject 11 has an R^2 of 0.2393. In contrast, all of the other subjects have an R^2 over 0.95. Subject 11 almost certainly did not take the task seriously, so his or her results will be discarded.

```
data results2;
   set results;
   if not (index(_depvar_, '11'));
   run;
data utils2;
   set utils;
   if not (index(_depvar_, '11'));
   run;
```

Simulations

The next steps display simulation observations. The most preferred combinations are printed for each subject.

```
proc sort data=results2(where=(w=0)) out=sims(drop=&_trgind);
   by _depvar_ descending p_depend_;
   run;
data sims; /* Pull out first 10 for each subject. */
   set sims;
   by _depvar_;
   retain n 0;
   if first._depvar_ then n = 0;
   n = n + 1;
   if n le 10;
   drop w _depend_ t_depend_ n _name_ _type_ intercept;
   run;
proc print data=sims label;
   by _depvar_ ;
   title2 'Simulations Sorted by Decreasing Predicted Utility';
   title3 'Just the Ten Most Preferred Combinations are Printed';
   label p_depend_ = 'Predicted Utility';
   run;
```

Frozen Diet Entrees
Simulations Sorted by Decreasing Predicted Utility
Just the Ten Most Preferred Combinations are Printed

---------- Dependent Variable Transformation(Name)=Identity(Subj01) ----------

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
1	25.0556	Chicken	2 Grams	$1.99	250
2	24.0556	Turkey	2 Grams	$1.99	250
3	22.0556	Chicken	2 Grams	$2.29	250
4	21.3889	Chicken	2 Grams	$1.99	350
5	21.3889	Beef	2 Grams	$1.99	250
6	21.0556	Turkey	2 Grams	$2.29	250
7	20.3889	Turkey	2 Grams	$1.99	350
8	18.3889	Chicken	2 Grams	$2.59	250
9	18.3889	Chicken	2 Grams	$2.29	350
10	18.3889	Beef	2 Grams	$2.29	250

---------- Dependent Variable Transformation(Name)=Identity(Subj02) ----------

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
11	23.4444	Turkey	2 Grams	$1.99	250
12	22.6111	Beef	2 Grams	$1.99	250
13	21.7778	Turkey	2 Grams	$2.29	250
14	21.4444	Chicken	2 Grams	$1.99	250
15	20.9444	Beef	2 Grams	$2.29	250
16	20.7778	Turkey	2 Grams	$1.99	350
17	19.9444	Beef	2 Grams	$1.99	350
18	19.7778	Chicken	2 Grams	$2.29	250
19	19.7778	Turkey	2 Grams	$2.59	250
20	19.1111	Turkey	2 Grams	$2.29	350

---------- Dependent Variable Transformation(Name)=Identity(Subj03) ----------

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
21	24.0000	Chicken	2 Grams	$1.99	250
22	23.6667	Turkey	2 Grams	$1.99	250
23	22.3333	Beef	2 Grams	$1.99	250
24	21.6667	Chicken	2 Grams	$2.29	250
25	21.3333	Chicken	2 Grams	$1.99	350
26	21.3333	Turkey	2 Grams	$2.29	250
27	21.0000	Turkey	2 Grams	$1.99	350
28	20.0000	Beef	2 Grams	$2.29	250
29	19.6667	Beef	2 Grams	$1.99	350
30	19.0000	Chicken	2 Grams	$2.29	350

---------- Dependent Variable Transformation(Name)=Identity(Subj04) ----------

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
31	24.2778	Turkey	2 Grams	$1.99	250
32	23.6111	Beef	2 Grams	$1.99	250
33	21.6111	Turkey	2 Grams	$2.29	250
34	20.9444	Beef	2 Grams	$2.29	250
35	20.9444	Turkey	5 Grams	$1.99	250
36	20.7778	Chicken	2 Grams	$1.99	250
37	20.2778	Beef	5 Grams	$1.99	250
38	18.4444	Turkey	8 Grams	$1.99	250
39	18.2778	Turkey	5 Grams	$2.29	250
40	18.1111	Chicken	2 Grams	$2.29	250

---------- Dependent Variable Transformation(Name)=Identity(Subj05) ----------

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
41	22.2222	Chicken	2 Grams	$1.99	250
42	22.2222	Beef	2 Grams	$1.99	250
43	20.5556	Turkey	2 Grams	$1.99	250
44	19.8889	Chicken	5 Grams	$1.99	250
45	19.8889	Beef	5 Grams	$1.99	250
46	19.5556	Chicken	2 Grams	$1.99	350
47	19.5556	Beef	2 Grams	$1.99	350
48	18.3889	Beef	8 Grams	$1.99	250
49	18.3889	Chicken	8 Grams	$1.99	250
50	18.2222	Turkey	5 Grams	$1.99	250

---------- Dependent Variable Transformation(Name)=Identity(Subj06) ----------

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
51	24.5556	Chicken	2 Grams	$1.99	250
52	24.0556	Beef	2 Grams	$1.99	250
53	21.7222	Turkey	2 Grams	$1.99	250
54	21.3333	Chicken	2 Grams	$1.99	350
55	20.8333	Beef	2 Grams	$1.99	350
56	20.5556	Chicken	5 Grams	$1.99	250
57	20.0556	Beef	5 Grams	$1.99	250
58	18.5000	Turkey	2 Grams	$1.99	350
59	18.2222	Chicken	8 Grams	$1.99	250
60	17.7222	Beef	8 Grams	$1.99	250

---------- Dependent Variable Transformation(Name)=Identity(Subj07) ----------

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
61	23.0556	Beef	2 Grams	$1.99	250
62	22.5556	Chicken	2 Grams	$1.99	250
63	21.8889	Beef	2 Grams	$2.29	250
64	21.3889	Chicken	2 Grams	$2.29	250
65	21.3889	Turkey	2 Grams	$1.99	250
66	21.0556	Beef	2 Grams	$1.99	350
67	20.5556	Chicken	2 Grams	$1.99	350
68	20.2222	Turkey	2 Grams	$2.29	250
69	19.8889	Beef	2 Grams	$2.29	350
70	19.5556	Beef	2 Grams	$2.59	250

---------- Dependent Variable Transformation(Name)=Identity(Subj08) ----------

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
71	22.5000	Chicken	2 Grams	$1.99	250
72	22.0000	Turkey	2 Grams	$1.99	250
73	21.5000	Beef	2 Grams	$1.99	250
74	20.1667	Chicken	5 Grams	$1.99	250
75	20.1667	Chicken	2 Grams	$1.99	350
76	19.6667	Turkey	5 Grams	$1.99	250
77	19.6667	Turkey	2 Grams	$1.99	350
78	19.1667	Beef	5 Grams	$1.99	250
79	19.1667	Beef	2 Grams	$1.99	350
80	17.8333	Chicken	8 Grams	$1.99	250

---------- Dependent Variable Transformation(Name)=Identity(Subj09) ----------

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
81	22.3889	Chicken	2 Grams	$1.99	250
82	22.2222	Turkey	2 Grams	$1.99	250
83	20.7222	Beef	2 Grams	$1.99	250
84	20.5000	Chicken	2 Grams	$1.99	350
85	20.3333	Turkey	2 Grams	$1.99	350
86	20.2222	Chicken	5 Grams	$1.99	250
87	20.0556	Turkey	5 Grams	$1.99	250
88	18.8333	Beef	2 Grams	$1.99	350
89	18.5556	Beef	5 Grams	$1.99	250
90	18.3333	Chicken	5 Grams	$1.99	350

---------- Dependent Variable Transformation(Name)=Identity(Subj10) ----------

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
91	23.1111	Beef	2 Grams	$1.99	250
92	22.6111	Turkey	2 Grams	$1.99	250
93	20.7778	Chicken	2 Grams	$1.99	250
94	20.2778	Beef	5 Grams	$1.99	250
95	20.1111	Beef	2 Grams	$1.99	350
96	19.7778	Turkey	5 Grams	$1.99	250
97	19.6111	Turkey	2 Grams	$1.99	350
98	18.6111	Beef	8 Grams	$1.99	250
99	18.1111	Turkey	8 Grams	$1.99	250
100	17.9444	Chicken	5 Grams	$1.99	250

```
---------- Dependent Variable Transformation(Name)=Identity(Subj12) ----------
```

Obs	Predicted Utility	Ingredient	Fat	Price	Calories
101	25.1667	Chicken	2 Grams	$1.99	250
102	25.0000	Beef	2 Grams	$1.99	250
103	22.8333	Turkey	2 Grams	$1.99	250
104	21.3333	Chicken	2 Grams	$2.29	250
105	21.1667	Chicken	2 Grams	$1.99	350
106	21.1667	Beef	2 Grams	$2.29	250
107	21.0000	Beef	2 Grams	$1.99	350
108	19.0000	Chicken	2 Grams	$2.59	250
109	19.0000	Turkey	2 Grams	$2.29	250
110	18.8333	Beef	2 Grams	$2.59	250

Summarizing Results Across Subjects

Conjoint analyses are performed on an individual basis, but usually the goal is to summarize the results across subjects. The `outtest=` data set contains all of the information in the printed output and can be manipulated to create additional reports including a list of the individual R^2s and the average of the importance values across subjects. Here is a listing of the variables in the `outtest=` data set.

```
proc contents data=utils2 position;
   ods select position;
   title2 'Variables in the OUTTEST= Data Set';
   run;
```

Frozen Diet Entrees
Variables in the OUTTEST= Data Set

The CONTENTS Procedure

Variables in Creation Order

#	Variable	Type	Len	Label
1	_DEPVAR_	Char	42	Dependent Variable Transformation(Name)
2	_TYPE_	Char	8	
3	Title	Char	80	Title
4	Variable	Char	42	Variable
5	Coefficient	Num	8	Coefficient
6	Statistic	Char	24	Statistic

7	Value	Num	8	Value
8	NumDF	Num	8	Num DF
9	DenDF	Num	8	Den DF
10	SSq	Num	8	Sum of Squares
11	MeanSquare	Num	8	Mean Square
12	F	Num	8	F Value
13	NumericP	Num	8	Numeric (Approximate) p Value
14	P	Char	9	Formatted p Value
15	LowerLimit	Num	8	95% Lower Confidence Limit
16	UpperLimit	Num	8	95% Upper Confidence Limit
17	StdError	Num	8	Standard Error
18	Importance	Num	8	Importance (% Utility Range)
19	Label	Char	256	Label

The individual R^2s are displayed by printing the `Value` variable for observations whose `Statistic` value is "R-Square".

```
proc print data=utils2 label;
   title2 'R-Squares';
   id _depvar_;
   var value;
   format value 4.2;
   where statistic = 'R-Square';
   label value = 'R-Square' _depvar_ = 'Subject';
   run;
```

Frozen Diet Entrees
R-Squares

Subject	R-Square
Identity(Subj01)	0.96
Identity(Subj02)	0.98
Identity(Subj03)	0.98
Identity(Subj04)	0.98
Identity(Subj05)	0.99
Identity(Subj06)	0.96
Identity(Subj07)	0.99
Identity(Subj08)	0.99
Identity(Subj09)	0.99
Identity(Subj10)	0.99
Identity(Subj12)	0.97

The next steps extract the importance values and create a table. The DATA step extracts the importance values and creates row and column labels. The PROC TRANSPOSE step creates a subjects by attributes matrix from a vector (of the number of subjects times the number of attribute values). PROC PRINT displays the importance values, and PROC MEANS displays the average importances.

```
data im;
   set utils2;
   if n(importance);   /* Exclude all missing, including specials.*/
   _depvar_ = scan(_depvar_, 2);   /* Discard transformation.     */
   label    = scan(label, 1, ','); /* Use up to comma for label.  */
   keep importance _depvar_ label;
   run;
proc transpose data=im out=im(drop=_name_ _label_);
   id label;
   by notsorted _depvar_;
   var importance;
   label _depvar_ = 'Subject';
   run;
proc print label;
   title2 'Importances';
   format _numeric_ 2.;
   id _depvar_;
   run;
proc means mean;
   title2 'Average Importances';
   run;
```

Frozen Diet Entrees
Importances

Subject	Ingredient	Fat	Price	Calories
Subj01	13	50	24	13
Subj02	8	65	15	11
Subj03	7	62	21	11
Subj04	13	22	25	39
Subj05	7	17	64	12
Subj06	11	25	52	13
Subj07	7	68	15	9
Subj08	4	21	65	10
Subj09	7	18	66	8
Subj10	10	19	59	13
Subj12	9	52	24	15

```
Frozen Diet Entrees
Average Importances

The MEANS Procedure

Variable              Mean
--------------------------
Ingredient       8.9069044
Fat             38.0953010
Price           39.0198700
Calories        13.9779245
--------------------------
```

On the average, price is the most important attribute followed very closely by fat content. These two attributes on the average account for 77% of preference. Calories and main ingredient account for the remaining 23%. Note that everyone does not have the same pattern of importance values. However, it is a little hard to compare subjects just by looking at the numbers.

We can make a nicer display of importances with stars flagging the most important attributes for each product as follows. These steps replace each importance variable with its formatted value followed by zero stars for 0 - 30, one star for 30 - 45, two stars for 45 - 60, three stars for 60 - 75, and so on. The value returned by the `ceil` function is the number of characters that are extracted from the string `' ******'`.

```
data im2;
   set im;
   label c1 = 'Ingredient' c2 = 'Fat' c3 = 'Price' c4 = 'Calories';
   c1 = put(ingredient, 2.) || substr('  ******', 1, ceil(ingredient / 15));
   c2 = put(fat       , 2.) || substr('  ******', 1, ceil(fat        / 15));
   c3 = put(price     , 2.) || substr('  ******', 1, ceil(price      / 15));
   c4 = put(calories  , 2.) || substr('  ******', 1, ceil(calories   / 15));
   run;
proc print label;
   title2 'Importances';
   var c1-c4;
   id _depvar_;
   run;
```

Frozen Diet Entrees
Importances

Subject	Ingredient	Fat		Price		Calories	
Subj01	13	50	**	24		13	
Subj02	8	65	***	15		11	
Subj03	7	62	***	21		11	
Subj04	13	22		25		39	*
Subj05	7	17		64	***	12	
Subj06	11	25		52	**	13	
Subj07	7	68	***	15		9	
Subj08	4	21		65	***	10	
Subj09	7	18		66	***	8	
Subj10	10	19		59	**	13	
Subj12	9	52	**	24		15	

Subject 4 is more concerned about calories. However, most individuals seem to fall into one of two groups, either primarily price conscious then fat conscious, or primarily fat conscious then price conscious.

Both the `out=` data set and the `outtest=` data set contain the part-worth utilities. In the `out=` data set, they are contained in the observations whose `_type_` value is `'M COEFFI'`. The part-worth utilities are the multiple regression coefficients. The names of the variables that contain the part-worth utilities are stored in the macro variable `&_trgind`, which is automatically created by PROC TRANSREG.

```
proc print data=results2 label;
   title2 'Part-Worth Utilities';
   where _type_ = 'M COEFFI';
   id _name_;
   var &_trgind;
   run;
```

Frozen Diet Entrees
Part-Worth Utilities

NAME	Ingredient, Chicken	Ingredient, Beef	Ingredient, Turkey	Fat, 8 Grams	Fat, 5 Grams
Subj01	1.55556	-2.11111	0.55556	-6.94444	-0.11111
Subj02	-1.05556	0.11111	0.94444	-7.72222	0.11111
Subj03	0.66667	-1.00000	0.33333	-7.66667	-0.16667
Subj04	-2.11111	0.72222	1.38889	-2.77778	-0.27778
Subj05	0.55556	0.55556	-1.11111	-1.77778	-0.27778
Subj06	1.11111	0.61111	-1.72222	-2.88889	-0.55556
Subj07	0.22222	0.72222	-0.94444	-7.61111	-0.27778
Subj08	0.50000	-0.50000	0.00000	-2.33333	0.00000
Subj09	0.61111	-1.05556	0.44444	-2.05556	-0.05556
Subj10	-1.38889	0.94444	0.44444	-2.05556	-0.38889
Subj12	0.83333	0.66667	-1.50000	-6.16667	-1.16667

NAME	Fat, 2 Grams	Price, $2.59	Price, $2.29	Price, $1.99	Calories, 350	Calories, 250
Subj01	7.05556	-3.44444	0.22222	3.22222	-1.83333	1.83333
Subj02	7.61111	-1.88889	0.11111	1.77778	-1.33333	1.33333
Subj03	7.83333	-2.66667	0.16667	2.50000	-1.33333	1.33333
Subj04	3.05556	-3.44444	0.38889	3.05556	-5.05556	5.05556
Subj05	2.05556	-7.27778	0.22222	7.05556	-1.33333	1.33333
Subj06	3.44444	-6.22222	-0.88889	7.11111	-1.61111	1.61111
Subj07	7.88889	-1.94444	0.38889	1.55556	-1.00000	1.00000
Subj08	2.33333	-7.33333	-0.00000	7.33333	-1.16667	1.16667
Subj09	2.11111	-7.38889	-0.05556	7.44444	-0.94444	0.94444
Subj10	2.44444	-7.22222	0.27778	6.94444	-1.50000	1.50000
Subj12	7.33333	-2.83333	-0.50000	3.33333	-2.00000	2.00000

These part-worth utilities can be clustered, for example using PROC FASTCLUS.

```
proc fastclus data=results2 maxclusters=3 out=clusts;
   where _type_ = 'M COEFFI';
   id _name_;
   var &_trgind;
   run;
proc sort; by cluster; run;
proc print label;
   title2 'Part-Worth Utilities, Clustered';
   by cluster;
   id _name_;
   var &_trgind;
   run;
```

Frozen Diet Entrees
Part-Worth Utilities, Clustered

---------- Cluster=1 ----------

NAME	Ingredient, Chicken	Ingredient, Beef	Ingredient, Turkey	Fat, 8 Grams	Fat, 5 Grams
Subj05	0.55556	0.55556	-1.11111	-1.77778	-0.27778
Subj06	1.11111	0.61111	-1.72222	-2.88889	-0.55556
Subj08	0.50000	-0.50000	0.00000	-2.33333	0.00000
Subj09	0.61111	-1.05556	0.44444	-2.05556	-0.05556
Subj10	-1.38889	0.94444	0.44444	-2.05556	-0.38889

NAME	Fat, 2 Grams	Price, $2.59	Price, $2.29	Price, $1.99	Calories, 350	Calories, 250
Subj05	2.05556	-7.27778	0.22222	7.05556	-1.33333	1.33333
Subj06	3.44444	-6.22222	-0.88889	7.11111	-1.61111	1.61111
Subj08	2.33333	-7.33333	-0.00000	7.33333	-1.16667	1.16667
Subj09	2.11111	-7.38889	-0.05556	7.44444	-0.94444	0.94444
Subj10	2.44444	-7.22222	0.27778	6.94444	-1.50000	1.50000

---------- Cluster=2 ----------

NAME	Ingredient, Chicken	Ingredient, Beef	Ingredient, Turkey	Fat, 8 Grams	Fat, 5 Grams
Subj01	1.55556	-2.11111	0.55556	-6.94444	-0.11111
Subj02	-1.05556	0.11111	0.94444	-7.72222	0.11111
Subj03	0.66667	-1.00000	0.33333	-7.66667	-0.16667
Subj07	0.22222	0.72222	-0.94444	-7.61111	-0.27778
Subj12	0.83333	0.66667	-1.50000	-6.16667	-1.16667

NAME	Fat, 2 Grams	Price, $2.59	Price, $2.29	Price, $1.99	Calories, 350	Calories, 250
Subj01	7.05556	-3.44444	0.22222	3.22222	-1.83333	1.83333
Subj02	7.61111	-1.88889	0.11111	1.77778	-1.33333	1.33333
Subj03	7.83333	-2.66667	0.16667	2.50000	-1.33333	1.33333
Subj07	7.88889	-1.94444	0.38889	1.55556	-1.00000	1.00000
Subj12	7.33333	-2.83333	-0.50000	3.33333	-2.00000	2.00000

---------- Cluster=3 ----------

NAME	Ingredient, Chicken	Ingredient, Beef	Ingredient, Turkey	Fat, 8 Grams	Fat, 5 Grams
Subj04	-2.11111	0.72222	1.38889	-2.77778	-0.27778

NAME	Fat, 2 Grams	Price, $2.59	Price, $2.29	Price, $1.99	Calories, 350	Calories, 250
Subj04	3.05556	-3.44444	0.38889	3.05556	-5.05556	5.05556

The clusters reflect what we saw looking at the importance information. Subject 4, who is the only subject that is primarily calorie conscious, is in a separate cluster from everyone else. Cluster 1 subjects 5, 6, 8, 9, and 10 are primarily price conscious. Cluster 2 subjects 1, 2, 3, 7, and 12 are primarily fat conscious.

Spaghetti Sauce

This example uses conjoint analysis in a study of spaghetti sauce preferences. The goal is to investigate the main effects for all of the attributes and the interaction of brand and price, and to simulate market share. Rating scale data are gathered from a group of subjects. The example has eight parts.

- An efficient experimental design is generated with the `%MktEx` macro.
- Descriptions of the spaghetti sauces are generated.
- Data are collected, entered, and processed.
- The metric conjoint analysis is performed with PROC TRANSREG.
- Market share is simulated with the maximum utility model.
- Market share is simulated with the Bradley-Terry-Luce and logit models.
- The simulators are compared.
- Change in market share is investigated.

Create an Efficient Experimental Design with the %MktEx Macro

In this example, subjects were asked to rate their interest in purchasing hypothetical spaghetti sauces. The table shows the attributes, the attribute levels, and the number of *df* associated with each effect.

Experimental Design		
Effects	**Levels**	*df*
Intercept		1
Brand	Pregu, Sundance, Tomato Garden	2
Meat Content	Vegetarian, Meat, Italian Sausage	2
Mushroom Content	Mushrooms, No Mention	1
Natural Ingredients	All Natural Ingredients, No Mention	1
Price	\$1.99, \$2.29, \$2.49, \$2.79, \$2.99	4
Brand × Price		8

The brand names "Pregu", "Sundance", and "Tomato Garden" are artificial. Usually, real brand names would be used − your client's or company's brand and the competitors' brands. The absence of a feature (for example, no mushrooms) is not mentioned in the product description, hence the "No Mention" in the table.

In this design there are 19 model *df*. A design with more than 19 runs must be generated if there are to be error *df*. A popular heuristic is to limit the design size to at most 30 runs. In this example, 30 runs allow us to have two observations in each of the 15 brand by price cells. Note however that when subjects are required to make that many judgments, there is the risk that the quality of the data will be poor. Caution should be used when generating designs with this many runs. We can use the `%MktRuns` macro to evaluate this and other design sizes. See page 479 for macro documentation

and information on installing and using SAS autocall macros. We specify the number of levels of each factor as the argument.

```
title 'Spaghetti Sauces';

%mktruns( 3 3 2 2 5 )
```

Spaghetti Sauces

Design Summary

Number of Levels	Frequency
2	2
3	2
5	1

Saturated = 11
Full Factorial = 180

Some Reasonable Design Sizes	Violations	Cannot Be Divided By
180 *	0	
60	1	9
90	1	4
120	1	9
30	2	4 9
150	2	4 9
210	2	4 9
36	5	5 10 15
72	5	5 10 15
108	5	5 10 15

* - 100% Efficient Design can be made with the MktEx Macro.

We see that 30 is a reasonable size, although it cannot be divided by $9 = 3 \times 3$ and $4 = 2 \times 2$, so perfect orthogonality will not be possible. We would need a much larger size like 60 or 180 to do better. Note that this output states "Saturated=11" referring to a main-effects model. In this example, we are also interested in the brand by price interaction. We can run the `%MktRuns` macro again, this time specifying the three-level factor, the five-level factor and the 3×5 interaction as one 15-level factor.

```
%mktruns( 3 2 2 15 )
```

Spaghetti Sauces

Design Summary

Number of Levels	Frequency
2	2
3	1
15	1

Saturated = 19
Full Factorial = 180

Some Reasonable Design Sizes	Violations	Cannot Be Divided By
180 *	0	
60	1	45
90	1	4
120	1	45
30	2	4 45
150	2	4 45
210	2	4 45
24	4	15 30 45
36	4	15 30 45
48	4	15 30 45

* - 100% Efficient Design can be made with the MktEx Macro.

Now the output states "Saturated=19", which includes the 8 *df* for the interaction. We see as before that 30 cannot be divided by $4 = 2 \times 2$. We also see that 30 cannot be divide by $45 = 3 \times 15$ so each level of meat content will not appear equally often in each brand/price cell. Since we would need a much larger size to do better, we will use 30 runs.

The next steps create and evaluate the design. First, formats for each of the factors are created using PROC FORMAT. The `%MktEx` macro is called to create the design. The factors `x1 = Brand` and `x2 = Meat` are designated as three-level factors, `x3 = Mushroom` and `x4 = Ingredients` as two-level factors, and `x5 = Price` as a five-level factor. The `interact=1*5` option specifies that the interaction between the first and fifth factors must be estimable (`x1` × `x5` which is brand by price), `n=30` specifies the number of runs, and `seed=289` specifies the random number seed. The `where` macro provides restrictions that eliminate unrealistic combinations. Specifically, products at the cheapest price, \$1.99, with meat, and products with Italian Sausage with All Natural Ingredients are eliminated from consideration.

We impose restrictions with the `%MktEx` macro by writing a macro, with IML statements, that quantifies the badness of each run of the design. The variable `bad` is set to zero when everything is fine, and values larger than zero when the row of the design does not conform to the restrictions. Ideally, when there are multiple restrictions, as there are here, the variable `bad` should be set to the number of violations, so the macro can know when it is moving in the right direction as it changes the design. Our first

restriction (contribution to the badness value) is `(x2 = 3 & x4 = 1)` and our second is `(x5 = 1 & (x2 = 2 | x2 = 3))`, where `&` means `and` and `|` means `or`.[¶] The restrictions correspond to `(Meat = 'Italian Sausage' & Ingredients = 'All Natural')` and `(Price = 1.99 & (Meat = 'Meat' | Meat = 'Italian Sausage'))`. Each of these Boolean or logical expressions evaluates to 1 when the expression is true and 0 when it is false. The sum of the two restrictions is: 0 - no problem, 1 - one restriction violation, or 2 - two restriction violations.

The `%MktLab` macro assigns actual descriptive factor names instead of the default `x1-x5` and formats for the levels. The default input to the `%MktLab` macro is the data set RANDOMIZED, which is the randomized design created by the `%MktEx` macro.

The default output from the `%MktLab` macro is a data set called FINAL. We instead use the `out=` option to store the results in a permanent SAS data set. The `%MktEx` macro is used to display the frequencies for each level, the two-way frequencies, and the number of times each product occurs in the design (five-way frequencies).

```
title 'Spaghetti Sauces';

proc format;
   value br 1='Pregu'       2='Sundance'  3='Tomato Garden';
   value me 1='Vegetarian'  2='Meat'      3='Italian Sausage';
   value mu 1='Mushrooms'   2='No Mention';
   value in 1='All Natural' 2='No Mention';
   value pr 1='1.99' 2='2.29' 3='2.49' 4='2.79' 5='2.99';
   run;

%macro where;
   bad = (x2 = 3 & x4 = 1) + (x5 = 1 & (x2 = 2 | x2 = 3));
   %mend;

%mktex(3 3 2 2 5, interact=1*5, n=30, seed=289, restrictions=where)
%mktlab(vars=Brand Meat Mushroom Ingredients Price,
        statements=format brand br. meat me. mushroom mu.
                   ingredients in. price pr.,
        out=sasuser.spag);
%mkteval;

proc print data=sasuser.spag; run;
```

Here is some of the output from the `%MktEx` macro.

[¶]In the restrictions macro, you must use the logical symbols `| & ∧ ¬ > < >= <= = ∧= ¬=` and *not* the logical words
`OR AND NOT GT LT GE LE EQ NE`.

Spaghetti Sauces

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	92.6280		Can
1	2 1	92.6280	92.6280	Conforms
1	End	92.6280		
2	Start	78.9640		Tab,Unb
2	28 1	91.5726		Conforms
2	End	91.6084		
3	Start	78.9640		Tab,Unb
3	1 1	91.5434		Conforms
3	End	91.6084		
4	Start	77.5906		Tab,Ran
4	28 1	91.9486		Conforms
4	5 4	92.6280	92.6280	
4	End	92.6280		
.				
.				
.				
21	Start	74.7430		Ran,Mut,Ann
21	24 1	89.9706		Conforms
21	End	91.6084		

Spaghetti Sauces

Design Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	92.6280	92.6280	Ini
1	Start	92.6280		Can
1	2 1	92.6280	92.6280	Conforms
1	End	92.6280		

Spaghetti Sauces

Design Refinement History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	92.6280	92.6280	Ini
1	Start	90.4842		Pre,Mut,Ann
1	2 1	91.2145		Conforms
1	End	91.6084		
.				
.				
.				
6	Start	91.1998		Pre,Mut,Ann
6	2 1	91.6084		Conforms
6	End	91.6084		

NOTE: Stopping since it appears that no improvement is possible.

Spaghetti Sauces

The OPTEX Procedure

Class Level Information

Class	Levels	-Values--
x1	3	1 2 3
x2	3	1 2 3
x3	2	1 2
x4	2	1 2
x5	5	1 2 3 4 5

Spaghetti Sauces

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	92.6280	82.6056	97.6092	0.7958

The *D*-Efficiency looks reasonable at 92.63. For this problem, the full-factorial design is small (180 runs), and the macro found the same *D*-efficiency several times. This suggests that we have probably

indeed found the optimal design for this situation. Here is the output from the `%MktEval` macro.

```
                          Spaghetti Sauces
             Canonical Correlations Between the Factors
        There are 2 Canonical Correlations Greater Than 0.316

                 Brand    Meat    Mushroom    Ingredients    Price

    Brand         1       0.21      0            0.17         0
    Meat          0.21    1         0.08         0.42         0.52
    Mushroom      0       0.08      1            0            0
    Ingredients   0.17    0.42      0            1            0.17
    Price         0       0.52      0            0.17         1

                          Spaghetti Sauces
          Canonical Correlations > 0.316 Between the Factors
        There are 2 Canonical Correlations Greater Than 0.316

                                          r    r Square

                Meat    Price           0.52      0.27
                Meat    Ingredients     0.42      0.17

                          Spaghetti Sauces
                       Summary of Frequencies
        There are 2 Canonical Correlations Greater Than 0.316
                  * - Indicates Unequal Frequencies

                                  Frequencies

        Brand                     10 10 10
   *    Meat                      15 9 6
        Mushroom                  15 15
   *    Ingredients               12 18
        Price                     6 6 6 6 6
   *    Brand Meat                4 3 3 5 4 1 6 2 2
        Brand Mushroom            5 5 5 5 5 5
   *    Brand Ingredients         3 7 5 5 4 6
        Brand Price               2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
   *    Meat Mushroom             7 8 5 4 3 3
   *    Meat Ingredients          7 8 5 4 0 6
   *    Meat Price                6 3 2 2 2 0 2 2 3 2 0 1 2 1 2
   *    Mushroom Ingredients      6 9 6 9
        Mushroom Price            3 3 3 3 3 3 3 3 3 3
   *    Ingredients Price         3 3 2 2 2 3 3 4 4 4
        N-Way                     1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                                  1 1 1 1 1 1 1 1 1 1 1
```

The meat and price factors are correlated, as are the meat and ingredients factors. This is not surprising since we excluded cells for these factor combinations and hence forced some correlations. The rest of the correlations are small.

The frequencies look good. The n-way frequencies at the end of this listing show that each product occurs only once. Each brand, price, and brand/price combination occurs equally often, as does each mushroom level. There are more vegetarian sauces (the first formatted level) than either of the meat sauces because of the restrictions that meat cannot occur at the lowest price and Italian sausage cannot be paired with all-natural ingredients. The design is shown next.

Spaghetti Sauces

Obs	Brand	Meat	Mushroom	Ingredients	Price
1	Pregu	Meat	No Mention	No Mention	2.79
2	Tomato Garden	Vegetarian	No Mention	No Mention	2.79
3	Pregu	Meat	Mushrooms	All Natural	2.29
4	Tomato Garden	Vegetarian	Mushrooms	All Natural	2.49
5	Sundance	Vegetarian	Mushrooms	No Mention	1.99
6	Pregu	Italian Sausage	No Mention	No Mention	2.49
7	Tomato Garden	Vegetarian	No Mention	No Mention	2.99
8	Tomato Garden	Italian Sausage	Mushrooms	No Mention	2.29
9	Pregu	Vegetarian	Mushrooms	No Mention	2.49
10	Pregu	Vegetarian	No Mention	No Mention	2.29
11	Sundance	Vegetarian	Mushrooms	No Mention	2.79
12	Tomato Garden	Vegetarian	Mushrooms	No Mention	1.99
13	Sundance	Meat	No Mention	No Mention	2.29
14	Sundance	Meat	Mushrooms	No Mention	2.99
15	Pregu	Italian Sausage	Mushrooms	No Mention	2.79
16	Tomato Garden	Italian Sausage	Mushrooms	No Mention	2.99
17	Sundance	Vegetarian	Mushrooms	All Natural	2.29
18	Pregu	Meat	Mushrooms	All Natural	2.99
19	Tomato Garden	Meat	No Mention	No Mention	2.49
20	Sundance	Meat	Mushrooms	All Natural	2.49
21	Pregu	Vegetarian	No Mention	All Natural	1.99
22	Sundance	Meat	No Mention	All Natural	2.79
23	Tomato Garden	Vegetarian	No Mention	All Natural	1.99
24	Sundance	Italian Sausage	No Mention	No Mention	2.49
25	Sundance	Vegetarian	No Mention	All Natural	1.99
26	Sundance	Vegetarian	No Mention	All Natural	2.99
27	Pregu	Italian Sausage	No Mention	No Mention	2.99
28	Tomato Garden	Vegetarian	No Mention	All Natural	2.29
29	Pregu	Vegetarian	Mushrooms	No Mention	1.99
30	Tomato Garden	Meat	Mushrooms	All Natural	2.79

Generating the Questionnaire

Next, preparations are made for data collection. A DATA step is used to print descriptions of each product combination. Here is an example:

```
Try Pregu brand vegetarian spaghetti sauce, now with
mushrooms.  A 26 ounce jar serves four adults for only
$1.99.
```

Remember that "No Mention" is not mentioned. The following step prints the questionnaires including a cover sheet.

```
options ls=80 ps=74 nonumber nodate;
title;

data _null_;
   set sasuser.spag;
   length lines $ 500 aline $ 60;
   file print linesleft=ll;

   * Format meat level, preserve 'Italian' capitalization;
   aline = lowcase(put(meat, me.));
   if aline =: 'ita' then substr(aline, 1, 1) = 'I';

   * Format meat differently for 'vegetarian';
   if meat > 1
      then lines = 'Try ' || trim(put(brand, br.)) ||
                   ' brand spaghetti sauce with ' || aline;
      else lines = 'Try ' || trim(put(brand, br.)) ||
                   ' brand ' || trim(aline) || ' spaghetti sauce ';
   * Add mushrooms, natural ingredients to text line;
   n = (put(ingredients, in.) =: 'All');
   m = (put(mushroom,    mu.) =: 'Mus');

   if n or m then do;
      lines = trim(lines) || ', now with';

      if m then do;
         lines = trim(lines) || ' ' || lowcase(put(mushroom, mu.));
         if n then lines = trim(lines) || ' and';
         end;
      if n then lines = trim(lines) || ' ' ||
                        lowcase(put(ingredients, in.)) || ' ingredients';
      end;
   * Add price;
   lines = trim(lines) ||
          '.  A 26 ounce jar serves four adults for only $' ||
          put(price, pr.) || '.';
```

```
* Print cover page, with subject number, instructions, and rating scale;
if _n_ = 1 then do;
   put ///// +41 'Subject: ________' ////
       +5 'Please rate your willingness to purchase the following' /
       +5 'products on a nine point scale.' ///
       +9 '1   Definitely Would Not Purchase This Product' ///
       +9 '2'  ///
       +9 '3   Probably Would Not Purchase This Product' ///
       +9 '4'  ///
       +9 '5   May or May Not Purchase This Product' ///
       +9 '6'  ///
       +9 '7   Probably Would Purchase This Product' ///
       +9 '8'  ///
       +9 '9   Definitely Would Purchase This Product' /////
       +5 'Please rate every product and be sure to rate' /
       +5 'each product only once.' //////
       +5 'Thank you for your participation!';
   put _page_;
   end;
if ll < 8 then put _page_;
* Break up description, print on several lines;

start = 1;
do l = 1 to 10 until(aline = ' ');

   * Find a good place to split, blank or punctuation;
   stop = start + 60;
   do i = stop to start by -1 while(substr(lines, i, 1) ne ' '); end;
   do j = i to max(start, i - 8) by -1;
      if substr(lines, j, 1) in ('.' ',') then do; i = j; j = 0; end;
      end;

   stop = i; len = stop + 1 - start;
   aline = substr(lines, start, len);
   start = stop + 1;
   if l = 1 then put +5 _n_ 2. ') ' aline;
   else          put +9 aline;
   end;
* Print rating scale;
put +9 'Definitely                                        Definitely ' /
    +9 'Would Not   1   2   3   4   5   6   7   8   9   Would      ' /
    +9 'Purchase                                        Purchase   ' //;
run;

options ls=80 ps=60 nonumber nodate;
```

In the interest of space, not all questions are printed.

Subject: _____

Please rate your willingness to purchase the following products on a nine point scale.

1 Definitely Would Not Purchase This Product

2

3 Probably Would Not Purchase This Product

4

5 May or May Not Purchase This Product

6

7 Probably Would Purchase This Product

8

9 Definitely Would Purchase This Product

Please rate every product and be sure to rate each product only once.

Thank you for your participation!

1) Try Pregu brand spaghetti sauce with meat. A 26 ounce jar serves four adults for only $2.79.

Definitely Would Not Purchase	1	2	3	4	5	6	7	8	9	Definitely Would Purchase

2) Try Tomato Garden brand vegetarian spaghetti sauce. A 26 ounce jar serves four adults for only $2.79.

Definitely Would Not Purchase	1	2	3	4	5	6	7	8	9	Definitely Would Purchase

3) Try Pregu brand spaghetti sauce with meat, now with mushrooms and all natural ingredients. A 26 ounce jar serves four adults for only $2.29.

Definitely Would Not Purchase	1	2	3	4	5	6	7	8	9	Definitely Would Purchase

4) Try Tomato Garden brand vegetarian spaghetti sauce, now with mushrooms and all natural ingredients. A 26 ounce jar serves four adults for only $2.49.

Definitely Would Not Purchase	1	2	3	4	5	6	7	8	9	Definitely Would Purchase

5) Try Sundance brand vegetarian spaghetti sauce, now with mushrooms. A 26 ounce jar serves four adults for only $1.99.

Definitely Would Not Purchase	1	2	3	4	5	6	7	8	9	Definitely Would Purchase

6) Try Pregu brand spaghetti sauce with Italian sausage. A 26 ounce jar serves four adults for only $2.49.

Definitely Would Not Purchase	1	2	3	4	5	6	7	8	9	Definitely Would Purchase

7) Try Tomato Garden brand vegetarian spaghetti sauce. A 26 ounce jar serves four adults for only $2.99.

Definitely Would Not Purchase	1	2	3	4	5	6	7	8	9	Definitely Would Purchase

```
        .
        .
        .

      30) Try Tomato Garden brand spaghetti sauce with meat, now with
          mushrooms and all natural ingredients.  A 26 ounce jar
          serves four adults for only $2.79.

          Definitely                                        Definitely
          Would Not   1   2   3   4   5   6   7   8   9   Would
          Purchase                                          Purchase
```

Data Processing

The data are entered next. Some cases have ordinary '.' missing values. This code was used at data entry for no response. When there were multiple responses or the response was not clear, the special underscore missing value was used. The statement `missing _` specifies that underscore missing values are to be expected in the data. The `input` statement reads the subject number and the 30 ratings. A name like `Subj001`, `Subj002`, ..., `Subj030` is created from the subject number. If there are any missing data, all data for that subject are excluded by the `if nmiss(of rate:) = 0` statement.

```
title 'Spaghetti Sauces';

data rawdata;
   missing _;
   input subj @5 (rate1-rate30) (1.);
   name = compress('Sub' || put(subj, z3.));
   if nmiss(of rate:) = 0;
   datalines;
  1 319591129691132168146121171191
  2 749173216928911175549891841791
  3 449491116819413186158171961791
  .
  .
  .
 14 1139812951994_9466149198915699
  .
  .
  .
 19 2214922399981121.1116161941991
  .
  .
  .
;
```

Next, the data are transposed from one row per subject and 30 columns to one column per subject and 30 rows, one for each product rated. Then the data are merged with the experimental design.

```
proc transpose data=rawdata(drop=subj) out=temp(drop=_name_);
    id name;
    run;

data inputdata; merge sasuser.spag temp; run;
```

Metric Conjoint Analysis

Next, we use PROC TRANSREG to perform the conjoint analysis.

```
ods exclude notes mvanova anova;
proc transreg data=inputdata utilities short separators='' ', '
   lprefix=0 outtest=utils method=morals;
   title2 'Conjoint Analysis';
   model identity(sub:) =
         class(brand | price meat mushroom ingredients / zero=sum);
   output p ireplace out=results1 coefficients;
   run;
```

The `utilities` option requests conjoint analysis output, and the `short` option suppresses the iteration histories. The `lprefix=0` option specifies that zero variable name characters are to be used to construct the labels for the part-worths; the labels will simply consist of formatted values. The `outtest=` option creates an output SAS data set, UTILS, that contains all of the statistical results. The `method=morals`, algorithm fits the conjoint analysis model separately for each subject. We specify `ods exclude notes mvanova anova` to exclude ANOVA information (which we usually want to ignore) and provide more parsimonious output.

The `model` statement names the ratings for each subject as dependent variables and the factors as independent variables. Since this is a metric conjoint analysis, `identity` is specified for the ratings. The `identity` transformation is the no-transformation option, which is used for variables that need to enter the model with no further manipulations. The factors are specified as `class` variables, and the `zero=sum` option is specified to constrain the parameter estimates to sum to zero within each effect. The `brand | price` specification asks for a simple `brand` effect, a simple `price` effect, and the `brand * price` interaction.

The `p` option in the `output` statement requests predicted values, the `ireplace` option suppresses the output of transformed independent variables, and the `coefficients` option requests that the part-worth utilities be output. These options control the contents of the `out=results` data set, which contains the ratings, predicted utilities for each product, indicator variables, and the part-worth utilities.

In the interest of space, only the results for the first subject are printed here. Recall that we used an `ods exclude` statement and we used PROC TEMPLATE on page 366 to customize the output from PROC TRANSREG.

Conjoint Analysis

The TRANSREG Procedure

Class Level Information

Class	Levels	Values
Brand	3	Pregu Sundance Tomato Garden
Price	5	1.99 2.29 2.49 2.79 2.99
Meat	3	Vegetarian Meat Italian Sausage
Mushroom	2	Mushrooms No Mention
Ingredients	2	All Natural No Mention

Number of Observations Read	30
Number of Observations Used	30

Conjoint Analysis

The TRANSREG Procedure

Identity(Sub001)
Algorithm converged.

The TRANSREG Procedure Hypothesis Tests for Identity(Sub001)

Root MSE	2.09608	R-Square	0.8344
Dependent Mean	3.73333	Adj R-Sq	0.5635
Coeff Var	56.14499		

Part-Worth Utilities

Label	Utility	Standard Error	Importance (% Utility Range)
Intercept	3.0675	0.45364	
Pregu	2.0903	0.55937	28.924
Sundance	0.2973	0.55886	
Tomato Garden	-2.3876	0.55205	

1.99	-0.6836	0.91331	7.134
2.29	0.3815	0.77035	
2.49	0.4209	0.78975	
2.79	-0.5397	0.79677	
2.99	0.4209	0.78975	
Pregu, 1.99	0.7430	1.09161	15.639
Pregu, 2.29	0.9491	1.13055	
Pregu, 2.49	-0.7433	1.14528	
Pregu, 2.79	-1.0115	1.13157	
Pregu, 2.99	0.0626	1.13769	
Sundance, 1.99	0.0361	1.09135	
Sundance, 2.29	-1.2578	1.09310	
Sundance, 2.49	-0.1443	1.16287	
Sundance, 2.79	1.1633	1.09574	
Sundance, 2.99	0.2027	1.12077	
Tomato Garden, 1.99	-0.7791	1.08788	
Tomato Garden, 2.29	0.3087	1.16798	
Tomato Garden, 2.49	0.8876	1.16026	
Tomato Garden, 2.79	-0.1518	1.10376	
Tomato Garden, 2.99	-0.2654	1.13455	
Vegetarian	2.2828	0.68783	27.813
Meat	-0.2596	0.70138	
Italian Sausage	-2.0231	0.86266	
Mushrooms	1.5514	0.38441	20.042
No Mention	-1.5514	0.38441	
All Natural	-0.0347	0.45814	0.448
No Mention	0.0347	0.45814	

The next steps process the `outtest=` data set, saving the R^2, adjusted R^2, and *df*. Subjects whose adjusted R^2 is less than 0.3 (R^2 approximately 0.73) are flagged for exclusion. We want the final analysis to be based on subjects who seemed to be taking the task seriously. The next steps flag the subjects whose fit seems bad and create a macro variable `&droplist` that contains a list of variables to be dropped from the final analysis.

```
data model;
   set utils;
   if statistic in ('R-Square', 'Adj R-Sq', 'Model');
   Subj = scan(_depvar_, 2);
   if statistic = 'Model' then do;
      value = numdf;
      statistic = 'Num DF';
      output;
      value = dendf;
      statistic = 'Den DF';
      output;
      value = dendf + numdf + 1;
      statistic = 'N';
      end;
   output;
   keep statistic value subj;
   run;
proc transpose data=model out=summ;
   by subj;
   idlabel statistic;
   id statistic;
   run;
data summ2(drop=list);
   length list $ 1000;
   retain list;
   set summ end=eof;
   if adj_r_sq < 0.3 then do;
      Small = '*';
      list = trim(list) || ' ' || subj;
      end;
   if eof then call symput('droplist', trim(list));
   run;
%put &droplist;
proc print label data=summ2(drop=_name_ _label_); run;
```

The `outtest=` data set contains for each subject the ANOVA, R^2, and part-worth utility tables. The numerator *df* is found in the variable `NumDF`, the denominator *df* is found in the variable `DenDF`, and the R^2 and adjusted R^2 are found in the variable `Value`. The first DATA step processes the `outtest=` data set, stores all of the statistics of interest in the variable `Value`, and discards the extra observations and variables. The PROC TRANSPOSE step creates a data set with one observation per subject. Here is the `&droplist` macro variable.

```
Sub011 Sub021 Sub031 Sub051 Sub071 Sub081 Sub092 Sub093 Sub094 Sub096
```

Here is some of the R^2 and *df* summary. We see the *df* are right, and most of the R^2's look good.

Conjoint Analysis

Obs	Subj	Num DF	Den DF	N	R-Square	Adj R-Sq	Small
1	Sub001	18	11	30	0.83441	0.56345	
2	Sub002	18	11	30	0.91844	0.78497	
3	Sub003	18	11	30	0.92908	0.81302	
.							
.							
.							
10	Sub010	18	11	30	0.97643	0.93786	
.							
.							
.							
84	Sub091	18	11	30	0.85048	0.60581	
85	Sub092	18	11	30	0.64600	0.06673	*
86	Sub093	18	11	30	0.45024	-0.44936	*
87	Sub094	18	11	30	0.62250	0.00477	*
88	Sub095	18	11	30	0.85996	0.63081	
89	Sub096	18	11	30	0.73321	0.29664	*
90	Sub097	18	11	30	0.94155	0.84589	
91	Sub099	18	11	30	0.88920	0.70789	
92	Sub100	18	11	30	0.90330	0.74507	

We can run the conjoint again, this time using the `drop=&droplist` data set option to drop the subjects with poor fit. In the interest of space, the `noprint` option was specified on this step. The printed output will be the same as in the previous step, except for the fact that a few subject's tables are deleted.

```
proc transreg data=inputdata(drop=&droplist) utilities short noprint
   separators='' ', ' lprefix=0 outtest=utils method=morals;
   title2 'Conjoint Analysis';
   model identity(sub:) =
         class(brand | price meat mushroom ingredients / zero=sum);
   output p ireplace out=results2 coefficients;
   run;
```

Simulating Market Share

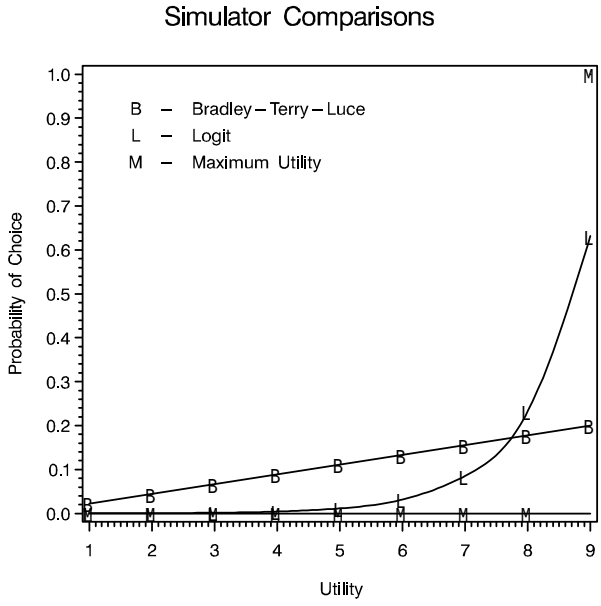
In many conjoint analysis studies, the conjoint analysis is not the primary goal. The conjoint analysis is used to generate part-worth utilities, which are then used as input to consumer choice and market share simulators. The end result for a product is its expected "preference share," which when properly weighted can be used to predict the proportion of times that the product will be purchased. The effects on market share of introducing new products can also be simulated.

One of the most popular ways to simulate market share is with the maximum utility model, which assumes each subject will buy with probability one the product for which he or she has the highest utility. The probabilities for each product are averaged across subjects to get predicted market share.

Other simulation methods include the Bradley-Terry-Luce (BTL) model and the logit model. Unlike the maximum utility model, the BTL and the logit models do not assign all of the probability of choice to the most preferred alternative. Probability is a continuous function of predicted utility. In the maximum utility model, probability of choice is a binary step function of utility. In the BTL model, probability of choice is a linear function of predicted utility. In the logit model, probability of choice is an increasing nonlinear logit function of predicted utility. The BTL model computes the probabilities by dividing each utility by the sum of the predicted utilities within each subject. The logit model divides the exponentiated predicted utilities by the sum of exponentiated utilities, again within subject.

$$\begin{aligned}
\text{Maximum Utility:} \quad p_{ijk} &= 1.0 \quad \text{if } y_{ijk} = \text{MAX}(y_{ijk}), \\
&\quad\ 0.0 \quad \text{otherwise} \\
\text{BTL:} \quad p_{ijk} &= y_{ijk} / \sum\sum\sum y_{ijk} \\
\text{Logit:} \quad p_{ijk} &= \exp(y_{ijk}) / \sum\sum\sum \exp(y_{ijk})
\end{aligned}$$

The following plot shows the different assumptions made by the three choice simulators. This plot shows expected market share for a subject with utilities ranging from one to nine.



The maximum utility line is flat at zero until it reaches the maximum utility, where it jumps to 1.0. The BTL line increases from 0.02 to 0.20 as utility ranges from 1 to 9. The logit function increases exponentially, with small utilities mapping to near-zero probabilities and the largest utility mapping to a proportion of 0.63.

The maximum utility, BTL, and logit models are based on different assumptions and produce different results. The maximum utility model has the advantage of being scale-free. Any strictly monotonic transformation of each subject's predicted utilities will produce the same market share. However, this model is unstable because it assigns a zero probability of choice to all alternatives that do not have the maximum predicted utility, including those that have predicted utilities near the maximum. The disadvantage of the BTL and logit models is that results are not invariant under linear transformations of the predicted utilities. These methods are considered inappropriate by some researchers for this reason. With negative predicted utilities, the BTL method produces negative probabilities, which are

invalid. The BTL results change when a constant is added to the predicted utilities but do not change when a constant is multiplied by the predicted utilities. Conversely, the logit results change when a constant is multiplied by the predicted utilities but do not change when a constant is added to the predicted utilities. The BTL method is not often used in practice, the logit model is sometimes used, and the maximum utility model is most often used. Refer to Finkbeiner (1988) for a discussion of conjoint analysis choice simulators. Do not confuse a logit model choice simulator and the multinomial logit model; they are quite different.

The three simulation methods will produce different results. This is because all three methods make different assumptions about how consumers translate utility into choice. To see why the models differ, imagine a product that is everyone's second choice. Further imagine that there is wide-spread disagreement on first choice. Every other product is someone's first choice, and all other products are preferred about equally often. In the maximum utility model, this second choice product will have zero probability of choice because no one would choose it first. In the other models, it should be the most preferred, because for every individual it will have a high, near-maximum probability of choice. Of course, preference patterns are not usually as weird as the one just described. If consumers are perfectly rational and always choose the alternative with the highest utility, then the maximum utility model is correct. However, you need to be aware that your results will depend on the choice of simulator model and in BTL and logit, the scaling of the utilities. One reason why the discrete choice model is popular in marketing research is discrete choice models choices directly, whereas conjoint simulates choices indirectly.

Here is the code that made the plot. You can try this program with different minima and maxima to see the effects of linear transformations of the predicted utilities.

```
%let min = 1;
%let max = 9;
%let by  = 1;
%let list = &min to &max by &by;
data a;
   sumb = 0;
   suml = 0;
   do u = &list;
      logit = exp(u);
      btl  = u;
      sumb = sumb + btl;
      suml = suml + logit;
      end;
   do u = &list;
      logit = exp(u);
      btl = u;
      max = abs(u - (&max)) < (0.5 * (&by));
      btl = btl / sumb;
      logit = logit / suml;
      output;
      end;
   run;
```

```
goptions ftext=swiss colors=(black) hsize=4.5in vsize=4.5in;
proc gplot;
   title h=1.5 'Simulator Comparisons';
   plot max * u = 1 btl * u = 2 logit * u = 3 /
      vaxis=axis2 haxis=axis1 overlay frame;
   symbol1 v=M i=step;
   symbol2 v=B i=join;
   symbol3 v=L i=spline;
   axis1 order=(&list) label=('Utility');
   axis2 order=(0 to 1 by 0.1)
         label=(angle=90 "Probability of Choice");
   note move=(2.5cm, 9.2cm)
        font=swissu 'B - ' font=swiss 'Bradley-Terry-Luce';
   note move=(2.5cm, 8.7cm)
        font=swissu 'L - ' font=swiss 'Logit';
   note move=(2.5cm, 8.2cm)
        font=swissu 'M - ' font=swiss 'Maximum Utility';
   run; quit;
```

Simulating Market Share, Maximum Utility Model

This section shows how to use the predicted utilities from a conjoint analysis to simulate choice and predict market share. The end result for a hypothetical product is its expected market share, which is a prediction of the proportion of times that the product will be purchased. Note however, that a term like "expected market share," while widely used, is a misnomer. Without purchase volume data, it is unlikely that these numbers would mirror true market share. Nevertheless, conjoint analysis is a useful and popular marketing research technique.

A SAS macro is used to simulate market share. It takes a `method=morals` output data set from PROC TRANSREG and creates a data set with expected market share for each combination. First, market share is computed with the maximum utility model. The macro finds the most preferred combination(s) for each subject, which are those combinations with the largest predicted utility, and assigns the probability that each combination will be purchased. Typically for each subject, one product will have a probability of purchase of 1.0, and all other produces will have zero probability of purchase. However, when two predicted utilities are tied for the maximum, that subject will have two probabilities of 0.5 and the rest will be zero. The probabilities are averaged across subjects for each product to get market share. Subjects can be differentially weighted.

```
                        /*--------------------------------------*/
                        /* Simulate Market Share                */
                        /*--------------------------------------*/
%macro sim(data=_last_, /* SAS data set with utilities.         */
           idvars=,     /* Additional variables to display with  */
                        /* market share results.                */
           weights=,    /* By default, each subject contributes  */
                        /* equally to the market share          */
                        /* computations.  To differentially     */
                        /* weight the subjects, specify a vector */
                        /* of weights, one per subject.         */
                        /* Separate the weights by blanks.      */
           out=shares,  /* Output data set name.                */
           method=max   /* max   - maximum utility model.       */
                        /* btl   - Bradley-Terry-Luce model.    */
                        /* logit - logit model.                 */
                        /* WARNING: The Bradley-Terry-Luce model */
                        /* and the logit model results are not   */
                        /* invariant under linear               */
                        /* transformations of the utilities.    */
          );            /*--------------------------------------*/

options nonotes;

%if &method = btl or &method = logit %then
   %put WARNING: The Bradley-Terry-Luce model and the logit model
results are not invariant under linear transformations of the
utilities.;
%else %if &method ne max %then %do;
   %put WARNING: Invalid method &method..  Assuming method=max.;
   %let method = max;
   %end;
* Eliminate coefficient observations, if any;
data temp1;
   set &data(where=(_type_ = 'SCORE' or _type_ = ' '));
   run;

* Determine number of runs and subjects.;
proc sql;
   create table temp2 as select nruns,
      count(nruns) as nsubs, count(distinct nruns) as chk
      from (select count( _depvar_) as nruns
      from temp1 where _type_ in ('SCORE', ' ') group by _depvar_);
   quit;
```

```
data _null_;
   set temp2;
   call symput('nruns', compress(put(nruns, 5.0)));
   call symput('nsubs', compress(put(nsubs, 5.0)));
   if chk > 1 then do;
      put 'ERROR: Corrupt input data set.';
      call symput('okay', 'no');
      end;
   else call symput('okay', 'yes');
   run;
%if &okay ne yes %then %do;
   proc print;
     title2 'Number of runs should be constant across subjects';
     run;
   %goto endit;
   %end;

%else %put NOTE: &nruns runs and &nsubs subjects.;
%let w = %scan(&weights, %eval(&nsubs + 1), %str( ));
%if %length(&w) > 0 %then %do;
   %put ERROR: Too many weights.;
   %goto endit;
   %end;

* Form nruns by nsubs data set of utilities;
data temp2;
   keep _u1 - _u&nsubs &idvars;
   array u[&nsubs] _u1 - _u&nsubs;
   do j = 1 to &nruns;

      * Read ID variables;
      set temp1(keep=&idvars) point = j;

      * Read utilities;
      k = j;
      do i = 1 to &nsubs;
         set temp1(keep=p_depend_) point = k;
         u[i] = p_depend_;
         %if &method = logit %then u[i] = exp(u[i]);;
         k = k + &nruns;
         end;

      output;
      end;

   stop;
   run;
```

```
* Set up for maximum utility model;
%if &method = max %then %do;

   * Compute maximum utility for each subject;
   proc means data=temp2 noprint;
      var _u1-_u&nsubs;
      output out=temp1 max=_sum1 - _sum&nsubs;
      run;
   * Flag maximum utility;
   data temp2(keep=_u1 - _u&nsubs &idvars);
      if _n_ = 1 then set temp1(drop=_type_ _freq_);
      array u[&nsubs] _u1 - _u&nsubs;
      array m[&nsubs] _sum1 - _sum&nsubs;
      set temp2;
      do i = 1 to &nsubs;
         u[i] = ((u[i] - m[i]) > -1e-8); /* < 1e-8 is considered 0 */
         end;
      run;

   %end;
* Compute sum for each subject;
proc means data=temp2 noprint;
   var _u1-_u&nsubs;
   output out=temp1 sum=_sum1 - _sum&nsubs;
   run;
* Compute expected market share;
data &out(keep=share &idvars);
   if _n_ = 1 then set temp1(drop=_type_ _freq_);
   array u[&nsubs] _u1 - _u&nsubs;
   array m[&nsubs] _sum1 - _sum&nsubs;
   set temp2;
   * Compute final probabilities;

   do i = 1 to &nsubs;
      u[i] = u[i] / m[i];
      end;
   * Compute expected market share;

   %if %length(&weights) = 0 %then %do;
      Share = mean(of _u1 - _u&nsubs);
      %end;
```

```
      %else %do;
         Share = 0;
         wsum  = 0;
         %do i = 1 %to &nsubs;
            %let w = %scan(&weights, &i, %str( ));
            %if %length(&w) = 0 %then %let w = .;
            if &w < 0 then do;
               if _n_ > 1 then stop;
               put "ERROR: Invalid weight &w..";
               call symput('okay', 'no');
               end;
            share = share + &w * _u&i;
            wsum  = wsum  + &w;
            %end;
         share = share / wsum;
         %end;
      run;
   options notes;

   %if &okay ne yes %then %goto endit;
   proc sort;
      by descending share &idvars;
      run;
   proc print label noobs;
      title2 'Expected Market Share';
      title3 %if       &method = max %then "Maximum Utility Model";
             %else %if &method = btl %then "Bradley-Terry-Luce Model";
             %else                         "Logit Model";;
      run;

   %endit:

   %mend;
   title 'Spaghetti Sauces';

   %sim(data=results2, out=maxutils, method=max,
        idvars=price brand meat mushroom ingredients);
```

Spaghetti Sauces
Expected Market Share
Maximum Utility Model

Brand	Price	Meat	Mushroom	Ingredients	Share
Sundance	1.99	Vegetarian	Mushrooms	No Mention	0.18293
Pregu	1.99	Vegetarian	No Mention	All Natural	0.14228
Tomato Garden	2.29	Italian Sausage	Mushrooms	No Mention	0.12195
Pregu	2.29	Vegetarian	No Mention	No Mention	0.10976
Pregu	1.99	Vegetarian	Mushrooms	No Mention	0.10366
Tomato Garden	1.99	Vegetarian	Mushrooms	No Mention	0.09146
Tomato Garden	1.99	Vegetarian	No Mention	All Natural	0.07520
Sundance	2.29	Vegetarian	Mushrooms	All Natural	0.07317
Sundance	1.99	Vegetarian	No Mention	All Natural	0.05081
Pregu	2.29	Meat	Mushrooms	All Natural	0.02439
Sundance	2.29	Meat	No Mention	No Mention	0.01220
Sundance	2.49	Italian Sausage	No Mention	No Mention	0.01220
Tomato Garden	2.29	Vegetarian	No Mention	All Natural	0.00000
Pregu	2.49	Vegetarian	Mushrooms	No Mention	0.00000
Pregu	2.49	Italian Sausage	No Mention	No Mention	0.00000
Sundance	2.49	Meat	Mushrooms	All Natural	0.00000
Tomato Garden	2.49	Vegetarian	Mushrooms	All Natural	0.00000
Tomato Garden	2.49	Meat	No Mention	No Mention	0.00000
Pregu	2.79	Meat	No Mention	No Mention	0.00000
Pregu	2.79	Italian Sausage	Mushrooms	No Mention	0.00000
Sundance	2.79	Vegetarian	Mushrooms	No Mention	0.00000
Sundance	2.79	Meat	No Mention	All Natural	0.00000
Tomato Garden	2.79	Vegetarian	No Mention	No Mention	0.00000
Tomato Garden	2.79	Meat	Mushrooms	All Natural	0.00000
Pregu	2.99	Meat	Mushrooms	All Natural	0.00000
Pregu	2.99	Italian Sausage	No Mention	No Mention	0.00000
Sundance	2.99	Vegetarian	No Mention	All Natural	0.00000
Sundance	2.99	Meat	Mushrooms	No Mention	0.00000
Tomato Garden	2.99	Vegetarian	No Mention	No Mention	0.00000
Tomato Garden	2.99	Italian Sausage	Mushrooms	No Mention	0.00000

The largest market share (18.29%) is for Sundance brand vegetarian sauce with mushrooms costing $1.99. The next largest share (14.23%) is Pregu brand vegetarian sauce with all natural ingredients costing $1.99. Five of the seven most preferred sauces all cost $1.99 – the minimum. It is not clear from this simulation if any brand is the leader.

Simulating Market Share, Bradley-Terry-Luce and Logit Models

The Bradley-Terry-Luce model and the logit model are also available in the %SIM macro.

```
title 'Spaghetti Sauces';

%sim(data=results2, out=btl, method=btl,
     idvars=price brand meat mushroom ingredients);

%sim(data=results2, out=logit, method=logit,
     idvars=price brand meat mushroom ingredients);
```

Spaghetti Sauces
Expected Market Share
Bradley-Terry-Luce Model

Brand	Price	Meat	Mushroom	Ingredients	Share
Pregu	1.99	Vegetarian	Mushrooms	No Mention	0.053479
Sundance	1.99	Vegetarian	Mushrooms	No Mention	0.052990
Tomato Garden	1.99	Vegetarian	Mushrooms	No Mention	0.051751
Pregu	1.99	Vegetarian	No Mention	All Natural	0.050683
Sundance	1.99	Vegetarian	No Mention	All Natural	0.050193
Tomato Garden	1.99	Vegetarian	No Mention	All Natural	0.048955
Sundance	2.29	Vegetarian	Mushrooms	All Natural	0.048236
Pregu	2.29	Vegetarian	No Mention	No Mention	0.043972
Tomato Garden	2.29	Vegetarian	No Mention	All Natural	0.042035
Pregu	2.49	Vegetarian	Mushrooms	No Mention	0.041532
Pregu	2.29	Meat	Mushrooms	All Natural	0.041063
Sundance	2.29	Meat	No Mention	No Mention	0.036321
Tomato Garden	2.29	Italian Sausage	Mushrooms	No Mention	0.032995
Sundance	2.79	Vegetarian	Mushrooms	No Mention	0.032067
Sundance	2.49	Meat	Mushrooms	All Natural	0.031310
Tomato Garden	2.49	Vegetarian	Mushrooms	All Natural	0.031057
Sundance	2.99	Vegetarian	No Mention	All Natural	0.026879
Pregu	2.49	Italian Sausage	No Mention	No Mention	0.026046
Pregu	2.99	Meat	Mushrooms	All Natural	0.025318
Pregu	2.79	Meat	No Mention	No Mention	0.025038
Tomato Garden	2.79	Vegetarian	No Mention	No Mention	0.024325
Pregu	2.79	Italian Sausage	Mushrooms	No Mention	0.024263
Sundance	2.49	Italian Sausage	No Mention	No Mention	0.022383
Sundance	2.99	Meat	Mushrooms	No Mention	0.022264
Tomato Garden	2.99	Vegetarian	No Mention	No Mention	0.022113
Sundance	2.79	Meat	No Mention	All Natural	0.021858
Tomato Garden	2.79	Meat	Mushrooms	All Natural	0.021415
Tomato Garden	2.49	Meat	No Mention	No Mention	0.019142
Pregu	2.99	Italian Sausage	No Mention	No Mention	0.016391
Tomato Garden	2.99	Italian Sausage	Mushrooms	No Mention	0.013926

Spaghetti Sauces
Expected Market Share
Logit Model

Brand	Price	Meat	Mushroom	Ingredients	Share
Sundance	1.99	Vegetarian	Mushrooms	No Mention	0.10463
Pregu	1.99	Vegetarian	No Mention	All Natural	0.09621
Tomato Garden	1.99	Vegetarian	Mushrooms	No Mention	0.09001
Pregu	1.99	Vegetarian	Mushrooms	No Mention	0.08358
Pregu	2.29	Vegetarian	No Mention	No Mention	0.07755
Sundance	2.29	Vegetarian	Mushrooms	All Natural	0.07102
Tomato Garden	1.99	Vegetarian	No Mention	All Natural	0.06872
Tomato Garden	2.29	Italian Sausage	Mushrooms	No Mention	0.06735
Sundance	1.99	Vegetarian	No Mention	All Natural	0.06419
Pregu	2.29	Meat	Mushrooms	All Natural	0.04137
Pregu	2.49	Vegetarian	Mushrooms	No Mention	0.03578
Sundance	2.29	Meat	No Mention	No Mention	0.03273
Sundance	2.49	Italian Sausage	No Mention	No Mention	0.02081
Tomato Garden	2.99	Italian Sausage	Mushrooms	No Mention	0.02055
Sundance	2.79	Vegetarian	Mushrooms	No Mention	0.02022
Tomato Garden	2.29	Vegetarian	No Mention	All Natural	0.01996
Pregu	2.79	Italian Sausage	Mushrooms	No Mention	0.01233
Pregu	2.49	Italian Sausage	No Mention	No Mention	0.01199
Sundance	2.49	Meat	Mushrooms	All Natural	0.01010
Sundance	2.99	Meat	Mushrooms	No Mention	0.00964
Pregu	2.79	Meat	No Mention	No Mention	0.00763
Pregu	2.99	Italian Sausage	No Mention	No Mention	0.00637
Pregu	2.99	Meat	Mushrooms	All Natural	0.00547
Tomato Garden	2.49	Vegetarian	Mushrooms	All Natural	0.00538
Tomato Garden	2.79	Meat	Mushrooms	All Natural	0.00516
Sundance	2.99	Vegetarian	No Mention	All Natural	0.00399
Sundance	2.79	Meat	No Mention	All Natural	0.00266
Tomato Garden	2.79	Vegetarian	No Mention	No Mention	0.00209
Tomato Garden	2.99	Vegetarian	No Mention	No Mention	0.00162
Tomato Garden	2.49	Meat	No Mention	No Mention	0.00088

The three methods produce different results.

Change in Market Share

The following steps simulate what would happen to the market if new products were introduced. Simulation observations are added to the data set and given zero weight. The conjoint analyses are rerun to compute the predicted utilities for the active observations and the simulations. The maximum utility model is used.

Recall that the design has numeric variables with values like 1, 2, and 3. Formats are used to print the descriptions of the levels of the attributes. The first thing we want to do is read in products

to simulate. We could read in values like 1, 2, and 3 or we could read in more descriptive values and convert them to numerics using informats. We chose the latter approach. First we use PROC FORMAT to create the informats. Previously, we created formats with PROC FORMAT by specifying a `value` statement followed by pairs of the form *numeric-value=descriptive-character-string*. We create an informat with PROC FORMAT by specifying an `invalue` statement followed by pairs of the form *descriptive-character-string=numeric-value*.

```
title 'Spaghetti Sauces';

proc format;
   invalue inbrand 'Preg'=1 'Sun' =2 'Tom' =3;
   invalue inmeat  'Veg' =1 'Meat'=2 'Ital'=3;
   invalue inmush  'Mush'=1 'No'  =2;
   invalue iningre 'Nat' =1 'No'  =2;
   invalue inprice '1.99'=1 '2.29'=2 '2.49'=3 '2.79'=4 '2.99'=5;
   run;
```

Next, we read the observations we want to consider for a sample market using the informats we just created. An `input` statement specification of the form "*variable* : *informat*" reads values starting with the first nonblank character.

```
data simulat;
   input brand       : inbrand.
         meat        : inmeat.
         mushroom    : inmush.
         ingredients : iningre.
         price       : inprice.;
   datalines;
Preg  Veg   Mush  Nat  1.99
Sun   Veg   Mush  Nat  1.99
Tom   Veg   Mush  Nat  1.99
Preg  Meat  Mush  Nat  2.49
Sun   Meat  Mush  Nat  2.49
Tom   Meat  Mush  Nat  2.49
Preg  Ital  Mush  Nat  2.79
Sun   Ital  Mush  Nat  2.79
Tom   Ital  Mush  Nat  2.79
;
```

Next, the original input data set is combined with the simulation observations. The subjects with poor fit are dropped and a `weight` variable is created to flag the simulation observations. The `weight` variable is not strictly necessary since all of the simulation observations will have missing values on the ratings so will be excluded from the analysis that way. Still, it is good practice to explicitly use weights to exclude observations.

```
data inputdata2(drop=&droplist);
   set inputdata(in=w) simulat;
   Weight = w;
   run;
```

```
proc print;
   title2 'Simulation Observations Have a Weight of Zero';
   id weight;
   var brand -- price;
   run;
```

Spaghetti Sauces
Simulation Observations Have a Weight of Zero

Weight	Brand	Meat	Mushroom	Ingredients	Price
1	Pregu	Meat	No Mention	No Mention	2.79
1	Tomato Garden	Vegetarian	No Mention	No Mention	2.79
1	Pregu	Meat	Mushrooms	All Natural	2.29
1	Tomato Garden	Vegetarian	Mushrooms	All Natural	2.49
1	Sundance	Vegetarian	Mushrooms	No Mention	1.99
1	Pregu	Italian Sausage	No Mention	No Mention	2.49
1	Tomato Garden	Vegetarian	No Mention	No Mention	2.99
1	Tomato Garden	Italian Sausage	Mushrooms	No Mention	2.29
1	Pregu	Vegetarian	Mushrooms	No Mention	2.49
1	Pregu	Vegetarian	No Mention	No Mention	2.29
1	Sundance	Vegetarian	Mushrooms	No Mention	2.79
1	Tomato Garden	Vegetarian	Mushrooms	No Mention	1.99
1	Sundance	Meat	No Mention	No Mention	2.29
1	Sundance	Meat	Mushrooms	No Mention	2.99
1	Pregu	Italian Sausage	Mushrooms	No Mention	2.79
1	Tomato Garden	Italian Sausage	Mushrooms	No Mention	2.99
1	Sundance	Vegetarian	Mushrooms	All Natural	2.29
1	Pregu	Meat	Mushrooms	All Natural	2.99
1	Tomato Garden	Meat	No Mention	No Mention	2.49
1	Sundance	Meat	Mushrooms	All Natural	2.49
1	Pregu	Vegetarian	No Mention	All Natural	1.99
1	Sundance	Meat	No Mention	All Natural	2.79
1	Tomato Garden	Vegetarian	No Mention	All Natural	1.99
1	Sundance	Italian Sausage	No Mention	No Mention	2.49
1	Sundance	Vegetarian	No Mention	All Natural	1.99
1	Sundance	Vegetarian	No Mention	All Natural	2.99
1	Pregu	Italian Sausage	No Mention	No Mention	2.99
1	Tomato Garden	Vegetarian	No Mention	All Natural	2.29
1	Pregu	Vegetarian	Mushrooms	No Mention	1.99
1	Tomato Garden	Meat	Mushrooms	All Natural	2.79
0	Pregu	Vegetarian	Mushrooms	All Natural	1.99
0	Sundance	Vegetarian	Mushrooms	All Natural	1.99
0	Tomato Garden	Vegetarian	Mushrooms	All Natural	1.99
0	Pregu	Meat	Mushrooms	All Natural	2.49
0	Sundance	Meat	Mushrooms	All Natural	2.49

```
0      Tomato Garden    Meat              Mushrooms     All Natural    2.49
0      Pregu            Italian Sausage   Mushrooms     All Natural    2.79
0      Sundance         Italian Sausage   Mushrooms     All Natural    2.79
0      Tomato Garden    Italian Sausage   Mushrooms     All Natural    2.79
```

The next steps run the conjoint analyses suppressing the printed output using the `noprint` option. The statement `weight weight` is specified since we want the simulation observations (which have zero weight) excluded from contributing to the analysis. However, the procedure will still compute an expected utility for every observation including observations with zero, missing, and negative weights. The `outtest=` data set is created like before so we can check to make sure the *df* and R^2 look reasonable.

```
ods exclude notes mvanova anova;
proc transreg data=inputdata2 utilities short noprint
   separators=', ' lprefix=0 method=morals outtest=utils;
   title2 'Conjoint Analysis';
   model identity(sub:) =
         class(brand | price meat mushroom ingredients / zero=sum);
   output p ireplace out=results3 coefficients;
   weight weight;
   run;
data model;
   set utils;
   if statistic in ('R-Square', 'Adj R-Sq', 'Model');
   Subj = scan(_depvar_, 2);
   if statistic = 'Model' then do;
      value = numdf;
      statistic = 'Num DF';
      output;
      value = dendf;
      statistic = 'Den DF';
      output;
      value = dendf + numdf + 1;
      statistic = 'N';
      end;
   output;
   keep statistic value subj;
   run;
proc transpose data=model out=summ;
   by subj;
   idlabel statistic;
   id statistic;
   run;
proc print label data=summ(drop=_name_ _label_); run;
```

The SAS log tells us that the nine simulation observations were deleted both because of zero weight and because of missing values in the dependent variables.

```
NOTE: 9 observations were deleted from the analysis but not from the
      output data set due to missing values.
NOTE: 9 observations were deleted from the analysis but not from the
      output data set due to nonpositive weights.
NOTE: A total of 9 observations were deleted.
```

The *df* and R^2 results, some of which are shown next, look fine.

Spaghetti Sauces
Conjoint Analysis

Obs	Subj	Num DF	Den DF	N	R-Square	Adj R-Sq
1	Sub001	18	11	30	0.83441	0.56345
2	Sub002	18	11	30	0.91844	0.78497
3	Sub003	18	11	30	0.92908	0.81302
.						
.						
.						
81	Sub099	18	11	30	0.88920	0.70789
82	Sub100	18	11	30	0.90330	0.74507

The simulation observations are pulled out of the `out=` data set, and the %SIM macro is run to simulate market share.

```
data results4;
   set results3;
   where weight = 0;
   run;
%sim(data=results4, out=shares2, method=max,
     idvars=price brand meat mushroom ingredients);
```

Spaghetti Sauces
Expected Market Share
Maximum Utility Model

Brand	Price	Meat	Mushroom	Ingredients	Share
Pregu	1.99	Vegetarian	Mushrooms	All Natural	0.35976
Sundance	1.99	Vegetarian	Mushrooms	All Natural	0.29878
Tomato Garden	1.99	Vegetarian	Mushrooms	All Natural	0.19512
Tomato Garden	2.79	Italian Sausage	Mushrooms	All Natural	0.08537
Sundance	2.79	Italian Sausage	Mushrooms	All Natural	0.02439
Pregu	2.49	Meat	Mushrooms	All Natural	0.01220
Sundance	2.49	Meat	Mushrooms	All Natural	0.01220
Pregu	2.79	Italian Sausage	Mushrooms	All Natural	0.01220
Tomato Garden	2.49	Meat	Mushrooms	All Natural	0.00000

For this set of products, the inexpensive vegetarian sauces have the greatest market share with Pregu brand preferred over Sundance and Tomato Garden. Now we'll consider adding six more products to the market, the six meat sauces we just saw, but at a lower price.

```
data simulat2;
   input brand       : inbrand.
         meat        : inmeat.
         mushroom    : inmush.
         ingredients : iningre.
         price       : inprice.;
   datalines;
Preg Meat Mush Nat 2.29
Sun  Meat Mush Nat 2.29
Tom  Meat Mush Nat 2.29
Preg Ital Mush Nat 2.49
Sun  Ital Mush Nat 2.49
Tom  Ital Mush Nat 2.49
;
data inputdata3(drop=&droplist);
   set inputdata(in=w) simulat simulat2;
   weight = w;
   run;
ods exclude notes mvanova anova;
proc transreg data=inputdata3 utilities short noprint
   separators=', ' lprefix=0 method=morals outtest=utils;
   title2 'Conjoint Analysis';
   model identity(sub:) =
         class(brand | price meat mushroom ingredients / zero=sum);
   output p ireplace out=results5 coefficients;
   weight weight;
   run;
```

Now we see that 15 simulation observations were excluded.

```
NOTE: 15 observations were deleted from the analysis but not from the
      output data set due to missing values.
NOTE: 15 observations were deleted from the analysis but not from the
      output data set due to nonpositive weights.
NOTE: A total of 15 observations were deleted.
```

These steps extract the *df* and R^2.

```
data model;
   set utils;
   if statistic in ('R-Square', 'Adj R-Sq', 'Model');
   Subj = scan(_depvar_, 2);
   if statistic = 'Model' then do;
      value = numdf;
      statistic = 'Num DF';
      output;
      value = dendf;
      statistic = 'Den DF';
      output;
      value = dendf + numdf + 1;
      statistic = 'N';
      end;
   output;
   keep statistic value subj;
   run;
proc transpose data=model out=summ;
   by subj;
   idlabel statistic;
   id statistic;
   run;
proc print label data=summ(drop=_name_ _label_); run;
```

The *df* and R^2 still look fine.

Spaghetti Sauces
Conjoint Analysis

Obs	Subj	Num DF	Den DF	N	R-Square	Adj R-Sq
1	Sub001	18	11	30	0.83441	0.56345
2	Sub002	18	11	30	0.91844	0.78497
3	Sub003	18	11	30	0.92908	0.81302
.						
.						
.						
81	Sub099	18	11	30	0.88920	0.70789
82	Sub100	18	11	30	0.90330	0.74507

Now we'll run the simulation with all 15 simulation observations.

```
data results6;
   set results5;
   where weight = 0;
   run;

%sim(data=results6, out=shares3, method=max,
     idvars=price brand meat mushroom ingredients);
```

Spaghetti Sauces
Expected Market Share
Maximum Utility Model

Brand	Price	Meat	Mushroom	Ingredients	Share
Sundance	1.99	Vegetarian	Mushrooms	All Natural	0.25813
Pregu	1.99	Vegetarian	Mushrooms	All Natural	0.20935
Pregu	2.29	Meat	Mushrooms	All Natural	0.19512
Tomato Garden	1.99	Vegetarian	Mushrooms	All Natural	0.15447
Sundance	2.49	Italian Sausage	Mushrooms	All Natural	0.08537
Sundance	2.29	Meat	Mushrooms	All Natural	0.03659
Tomato Garden	2.49	Italian Sausage	Mushrooms	All Natural	0.01829
Tomato Garden	2.29	Meat	Mushrooms	All Natural	0.01220
Pregu	2.49	Italian Sausage	Mushrooms	All Natural	0.01220
Tomato Garden	2.79	Italian Sausage	Mushrooms	All Natural	0.01220
Sundance	2.79	Italian Sausage	Mushrooms	All Natural	0.00610
Pregu	2.49	Meat	Mushrooms	All Natural	0.00000
Sundance	2.49	Meat	Mushrooms	All Natural	0.00000
Tomato Garden	2.49	Meat	Mushrooms	All Natural	0.00000
Pregu	2.79	Italian Sausage	Mushrooms	All Natural	0.00000

These steps merge the data set containing the old market shares with the data set containing the new market shares to show the effect of adding the new products.

```
title 'Spaghetti Sauces';

proc sort data=shares2;
   by price brand meat mushroom ingredients;
   run;
proc sort data=shares3;
   by price brand meat mushroom ingredients;
   run;
data both;
   merge shares2(rename=(share=OldShare)) shares3;
   by price brand meat mushroom ingredients;
   if oldshare = . then Change = 0;
   else change = oldshare;
   change = share - change;
   run;
```

```
proc sort;
   by descending share price brand meat mushroom ingredients;
   run;
options missing=' ';
proc print noobs;
   title2 'Expected Market Share and Change';
   var price brand meat mushroom ingredients
       oldshare share change;
   format oldshare -- change 6.3;
   run;
options missing=.;
```

Spaghetti Sauces
Expected Market Share and Change

Price	Brand	Meat	Mushroom	Ingredients	Old Share	Share	Change
1.99	Sundance	Vegetarian	Mushrooms	All Natural	0.299	0.258	-0.041
1.99	Pregu	Vegetarian	Mushrooms	All Natural	0.360	0.209	-0.150
2.29	Pregu	Meat	Mushrooms	All Natural		0.195	0.195
1.99	Tomato Garden	Vegetarian	Mushrooms	All Natural	0.195	0.154	-0.041
2.49	Sundance	Italian Sausage	Mushrooms	All Natural		0.085	0.085
2.29	Sundance	Meat	Mushrooms	All Natural		0.037	0.037
2.49	Tomato Garden	Italian Sausage	Mushrooms	All Natural		0.018	0.018
2.29	Tomato Garden	Meat	Mushrooms	All Natural		0.012	0.012
2.49	Pregu	Italian Sausage	Mushrooms	All Natural		0.012	0.012
2.79	Tomato Garden	Italian Sausage	Mushrooms	All Natural	0.085	0.012	-0.073
2.79	Sundance	Italian Sausage	Mushrooms	All Natural	0.024	0.006	-0.018
2.49	Pregu	Meat	Mushrooms	All Natural	0.012	0.000	-0.012
2.49	Sundance	Meat	Mushrooms	All Natural	0.012	0.000	-0.012
2.49	Tomato Garden	Meat	Mushrooms	All Natural	0.000	0.000	0.000
2.79	Pregu	Italian Sausage	Mushrooms	All Natural	0.012	0.000	-0.012

We see that the vegetarian sauces are most preferred, but we predict they would lose share if the new meat sauces were entered in the market. In particular, the Sundance and Pregu meat sauces would gain significant market share under this model.

PROC TRANSREG Specifications

PROC TRANSREG (transformation regression) is used to perform conjoint analysis and many other types of analyses, including simple regression, multiple regression, redundancy analysis, canonical correlation, analysis of variance, and external unfolding, all with nonlinear transformations of the variables. This section documents the statements and options available in PROC TRANSREG that are commonly used in conjoint analyses. Refer to "The TRANSREG Procedure" in the *SAS/STAT User's Guide* for more information on PROC TRANSREG. This section documents only a small subset of the capabilities of PROC TRANSREG.

The following statements are used in the TRANSREG procedure for conjoint analysis:

```
PROC TRANSREG <DATA=SAS-data-set> <OUTTEST=SAS-data-set>
              <a-options> <o-options>;
   MODEL transform(dependents   </ t-options>) =
         transform(independents </ t-options>)
        <transform(independents </ t-options>) ...> </ a-options>;
   OUTPUT <OUT=SAS-data-set> <o-options>;
   WEIGHT variable;
   ID variables;
   BY variables;
```

Specify the `proc` and `model` statements to use PROC TRANSREG. The `output` statement is required to produce an `out=` output data set, which contains the transformations, indicator variables, and predicted utility for each product. The `outtest=` data set, which contains the ANOVA, regression, and part-worth utility tables, is requested in the `proc` statement. All options can be abbreviated to their first three letters.

PROC TRANSREG *Statement*

```
PROC TRANSREG <DATA=SAS-data-set> <OUTTEST=SAS-data-set>
              <a-options> <o-options>;
```

The `data=` and `outtest=` options can appear only in the PROC TRANSREG statement. The algorithm options *(a-options)* appear in the `proc` or `model` statement. The output options *(o-options)* can appear in the `proc` or `output` statement.

`DATA=`*SAS-data-set*
specifies the input SAS data. If the `data=` option is not specified, PROC TRANSREG uses the most recently created SAS data set.

`OUTTEST=`*SAS-data-set*
specifies an output data set that will contain the ANOVA table, R^2, and the conjoint analysis part-worth utilities, and the attribute importances.

Algorithm Options

```
PROC TRANSREG <DATA=SAS-data-set> <OUTTEST=SAS-data-set>
              <a-options> <o-options>;
   MODEL transform(dependents   </ t-options>) =
         transform(independents </ t-options>)
        <transform(independents </ t-options>) ...> </ a-options>;
```

Algorithm options can appear in the `proc` or `model` statement as *a-options*.

`CONVERGE=`*n*
specifies the minimum average absolute change in standardized variable scores that is required to continue iterating. By default, `converge=0.00001`.

`DUMMY`
requests a canonical initialization. When `spline` transformations are requested, specify `dummy` to solve for the optimal transformations without iteration. Iteration is only necessary when there are monotonicity constraints.

`LPREFIX=`*n*
specifies the number of first characters of a `class` variable's label (or name if no label is specified) to use in constructing labels for part-worth utilities. For example, the default label for `Brand=Duff` is "Brand Duff". If you specify `lprefix=0` then the label is simply "Duff".

`MAXITER=`*n*
specifies the maximum number of iterations. By default, `maxiter=30`.

`NOPRINT`
suppresses the display of all output.

`ORDER=FORMATTED`
`ORDER=INTERNAL`
specifies the order in which the CLASS variable levels are reported. The default, `order=internal`, sorts by unformatted value. Specify `order=formatted` when you want the levels sorted by formatted value. Sort order is machine dependent. Note that in Version 6 and Version 7 of the SAS System, the default sort order was `order=formatted`. The default was changed to `order=internal` in Version 8 to be consistent with Base SAS procedures.

`METHOD=MORALS`
`METHOD=UNIVARIATE`
specifies the iterative algorithm. Both `method=morals` and `method=univariate` fit univariate multiple regression models with the possibility of nonlinear transformations of the variables. They differ in the way they structure the output data set when there is more than one dependent variable. When it can be used, `method=univariate` is more efficient than `method=morals`.

You can use `method=univariate` when no transformations of the independent variables are requested, for example when the independent variables are all designated `class`, `identity`, or `pspline`. In this case, the final set of independent variables will be the same for all subjects. If transformations such as

`monotone`, `identity`, `spline` or `mspline` are specified for the independent variables, the transformed independent variables may be different for each dependent variable and so must be output separately for each dependent variable. In conjoint analysis, there will typically be one dependent variable for each subject. This is illustrated in the examples.

With `method=univariate` and more than one dependent variable, PROC TRANSREG creates a data set with the same number of score observations as the original but with more variables. The untransformed dependent variable names are unchanged. The default transformed dependent variable names consist of the prefix "T" and the original variable names. The default predicted value names consist of the prefix "P" and the original variable names. The full set of independent variables appears once.

When more than one dependent variable is specified, `method=morals` creates a *rolled-out* data set with the dependent variable in `_depend_`, its transformation in `t_depend_`, and its predicted values in `p_depend_`. The full set of independents is repeated for each (original) dependent variable.

The procedure chooses a default method based on what is specified in the `model` statement. When transformations of the independent variables are requested, the default method is `morals`. Otherwise the default method is `univariate`.

`SEPARATORS=`*string-1* ¡*string-2*¿
specifies separators for creating labels for the part-worth utilities. By default, `separators=' ' ' * '` ("blank" and "blank asterisk blank"). The first value is used to separate variable names and values in interactions. The second value is used to separate interaction components. For example, the default label for `Brand=Duff` is "Brand Duff". If you specify `separators=', '` then the label is "Brand, Duff". Furthermore, the default label for the interaction of `Brand=Duff` and `Price=3.99` is "Brand Duff * Price 3.99". You could specify `lprefix=0` and `separators='' ' @ '` to instead create labels like "Duff @ 3.99". You use the `lprefix=0` option when you want to construct labels using zero characters of the variable name, that is when you want to construct labels from just the formatted level. The option `separators='' ' @ '` specifies in the second string a separator of the form "blank at blank". In this case, the first string is ignored because with `lprefix=0` there is no name to separate from the level.

`SHORT`
suppresses the iteration histories. For most standard metric conjoint analyses, no iterations are necessary, so specifying `short` eliminates unnecessary output. PROC TRANSREG will print a message if it ever fails to converge, so it is usually safe to specify the `short` option.

`UTILITIES`
prints the part-worth utilities and importances table and an ANOVA table. Note that you can use an `ods exclude` statement to exclude ANOVA tables and unnecessary notes from the conjoint output (see page 367).

Output Options

```
PROC TRANSREG <DATA=SAS-data-set> <OUTTEST=SAS-data-set>
              <a-options> <o-options>;
   OUTPUT <OUT=SAS-data-set> <o-options>;
```

The `out=` option can only appear in the `output` statement. The other output options can appear in the `proc` or `output` statement as *o-options*.

`COEFFICIENTS`
outputs the part-worth utilities to the `out=` data set.

`P`
includes the predicted values in the `out=` output data set, which are the predicted utilities for each product. By default, the predicted values variable name is the original dependent variable name prefixed with a "P".

`IREPLACE`
replaces the original independent variables with the transformed independent variables in the output data set. The names of the transformed variables in the output data set correspond to the names of the original independent variables in the input data set.

`OUT=`*SAS-data-set*
names the output data set. When an `output` statement is specified without the `out=` option, PROC TRANSREG creates a data set and uses the DATA*n* convention. To create a permanent SAS data set, specify a two-level name. The data set will contain the original input variables, the coded indicator variables, the transformation of the dependent variable, and the optionally predicted utilities for each product.

`RESIDUALS`
outputs to the `out=` data set the differences between the observed and predicted utilities. By default, the residual variable name is the original dependent variable name prefixed with an "R".

Transformations and Expansions

```
MODEL transform(dependents   </ t-options>) =
      transform(independents </ t-options>)
     <transform(independents </ t-options>) ...> </ a-options>;
```

The operators "`*`", "`|`", and "`@`" from the GLM procedure are available for interactions with `class` variables.

```
class(a * b ...
      c | d ...
      e | f ... @ n)
```

For example, this statement fits 100 individual main-effects models:

```
model identity(rating1-rating100) = class(x1-x5 / zero=sum);
```

This fits models with main effects and all two-way interactions:

```
model identity(rating1-rating100) = class(x1|x2|x3|x4|x5@2 / zero=sum);
```

This fits models with main effects and some two-way interactions:

```
model identity(rating1-rating100) = class(x1-x5 x1*x2 x3*x4 / zero=sum);
```

You can also fit separate price functions within each brand by specifying:

```
model identity(rating1-rating100) =
      class(brand / zero=none) | spline(price);
```

The list `x1-x5` is equivalent to `x1 x2 x3 x4 x5`. The vertical bar specifies all main effects and interactions, and the at sign limits the interactions. For example, `@2` limits the model to main effects and two-way interactions. The list `x1|x2|x3|x4|x5@2` is equivalent to `x1 x2 x1 * x2 x3 x1 * x3 x2 * x3 x4 x1 * x4 x2 * x4 x3 * x4 x5 x1 * x5 x2 * x5 x3 * x5 x4 * x5`. The specification `x1 * x2` indicates the two-way interaction between `x1` and `x2`, and `x1 * x2 * x3` indicates the three-way interaction between `x1`, `x2`, and `x3`.

Each of the following can be specified in the `model` statement as a *transform*. The `pspline` and `class` expansions create more than one output variable for each input variable. The rest are transformations that create one output variable for each input variable.

`CLASS`
designates variables for analysis as nominal-scale-of-measurement variables. For conjoint analysis, the `zero=sum` *t-option* is typically specified: `class(`*variables* `/ zero=sum)`. Variables designated as `class` variables are expanded to a set of indicator variables. Usually the number output variables for each `class` variable is the number of different values in the input variables. Dependent variables should not be designated as `class` variables.

`IDENTITY`
variables are not changed by the iterations. The `identity(`*variables*`)` specification designates interval-scale-of-measurement variables when no transformation is permitted. When small data values mean high preference, you will need to use the `reflect` transformation option.

`MONOTONE`
monotonically transforms variables; ties are preserved. When `monotone(`*variables*`)` is used with dependent variables, a nonmetric conjoint analysis is performed. When small data values mean high preference, you will need to use the `reflect` transformation option. The `monotone` specification can also be used with independent variables to impose monotonicity on the part-worth utilities. When it is known that monotonicity should exist in an attribute variable, using `monotone` instead of `class` for that attribute may improve prediction. An option exists in PROC TRANSREG for optimally untying tied values, but this option should not be used because it almost always produces a degenerate result.

`MSPLINE`
monotonically and smoothly transforms variables. By default, `mspline(`*variables)* fits a monotonic quadratic spline with no knots. Knots are specified as *t-options*, for example `mspline(`*variables* `/ nknots=3)` or `mspline(`*variables* `/ knots=5 to 15 by 5)`. Like `monotone`, `mspline`, finds a monotonic transformation. Unlike `monotone`, `mspline` places a bound on the *df* (number of knots + degree) used by the transformation. With `mspline`, it is possible to allow for nonlinearity in the responses and still have error *df*. This is not always possible with `monotone`. When small data values mean high preference, you will need to use the `reflect` transformation option. You can also use `mspline` with attribute variables to impose monotonicity on the part-worth utilities.

PSPLINE
expands each variable to a piece-wise polynomial spline basis. By default, `pspline(`*variables*`)` uses a cubic spline with no knots. Knots are specified as *t-options*. Specify `pspline(`*variable* `/ degree=2)` for an attribute variable to fit a quadratic model. For each `pspline` variable, $d + k$ output variables are created, where d is the degree of the polynomial and k is the number of knots. You should not specify `pspline` with the dependent variables.

RANK
performs a rank transformation, with ranks averaged within ties. Rating-scale data can be transformed to ranks by specifying `rank(`*variables*`)`. When small data values mean high preference, you will need to use the `reflect` transformation option. Typically, `rank` is only used for dependent variables. For example, if a rating-scale variable has sorted values 1, 1, 1, 2, 3, 3, 4, 5, 5, 5, then the rank transformation is 2, 2, 2, 4, 5.5, 5.5, 7, 9, 9, 9. A conjoint analysis of the original rating-scale variable will not usually be the same as a conjoint analysis of a rank transformation of the ratings. With ordinal-scale-of-measurement data, it is often good to analyze rank transformations instead of the original data. An alternative is to specify `monotone`, which performs a nonmetric conjoint analysis. For real data, `monotone` will always find a better fit than `rank`, but `rank` may lead to better prediction.

SPLINE
smoothly transforms variables. By default, `spline(`*variables*`)` fits a cubic spline with no knots. Knots are specified as *t-options*. Like `pspline`, `spline` models nonlinearities in the attributes.

Transformation Options

```
MODEL transform(dependents   </ t-options>) =
      transform(independents </ t-options>)
     <transform(independents </ t-options>) ...> </ a-options>;
```

The following are specified in the `model` statement as *t-options*'s.

DEGREE=*n*
specifies the degree of the spline. The defaults are `degree=3` (cubic spline) for `spline` and `pspline`, and `degree=2` (quadratic spline) for `mspline`. For example, to request a quadratic spline, specify `spline(`*variables* `/ degree=2)`.

EVENLY
is used with the `nknots=` option to evenly space the knots for splines. For example, if `spline(x / nknots=2 evenly)` is specified and `x` has a minimum of 4 and a maximum of 10, then the two interior knots are 6 and 8. Without `evenly`, the `nknots=` option places knots at percentiles, so the knots are not evenly spaced.

KNOTS=*numberlist*
specifies the interior knots or break points for splines. By default, there are no knots. For example, to request knots at 1, 2, 3, 4, 5, specify `spline(`*variable* `/ knots=1 to 5)`.

`NKNOTS=`*k*
creates *k* knots for splines: the first at the $100/(k+1)$ percentile, the second at the $200/(k+1)$ percentile, and so on. Unless `evenly` is specified, knots are placed at data values; there is no interpolation. For example, with `spline(`*variable* `/ NKNOTS=3)`, knots are placed at the twenty-fifth percentile, the median, and the seventy-fifth percentile. By default, `nknots=0`.

`REFLECT`
reflects the transformation around its mean, $\mathrm{Y} = -(\mathrm{Y} - \mathrm{Y}) + \mathrm{Y}$, after the iterations are completed and before the final standardization and results calculations. This option is particularly useful with the dependent variable. When the dependent variable consists of ranks with the most preferred combination assigned 1.0, `identity(`*variable* `/ reflect)` will reflect the transformation so that positive utilities mean high preference.

`ZERO=SUM`
constrains the part-worth utilities to sum to zero within each attribute. The specification `class(`*variables* `/ zero=sum)` creates a less than full rank model, but the coefficients are uniquely determined due to the sum-to-zero constraint.

BY *Statement*

```
BY variables;
```

A `by` statement can be used with PROC TRANSREG to obtain separate analyses on observations in groups defined by the `by` variables. When a `by` statement appears, the procedure expects the input data set to be sorted in order of the `by` variables.

If the input data set is not sorted in ascending order, use one of the following alternatives:

- Use the SORT procedure with a similar `by` statement to sort the data.
- Use the `by` statement options `notsorted` or `descending` in the `by` statement for the TRANSREG procedure. As a cautionary note, the `notsorted` option does not mean that the data are unsorted. It means that the data are arranged in groups (according to values of the `by` variables), and these groups are not necessarily in alphabetical or increasing numeric order.
- Use the DATASETS procedure (in base SAS software) to create an index on the `by` variables.

For more information on the `by` statement, refer to the discussion in *SAS Language: Reference*. For more information on the DATASETS procedure, refer to the discussion in *SAS Procedures Guide*.

ID *Statement*

```
ID variables;
```

The `id` statement includes additional character or numeric variables from the input data set in the `out=` data set.

WEIGHT *Statement*

```
WEIGHT variable;
```

A `weight` statement can be used in conjoint analysis to distinguish ordinary active observations, holdouts, and simulation observations. When a `weight` statement is used, a weighted residual sum of squares is minimized. The observation is used in the analysis only if the value of the `weight` statement variable is greater than zero. For observations with positive weight, the `weight` statement has no effect on *df* or number of observations, but the weights affect most other calculations.

Assign each active observation a weight of 1. Assign each holdout observation a weight that excludes it from the analysis, such as missing. Assign each simulation observation a different weight that excludes it from the analysis, such as zero. Holdouts are rated by the subjects and so have nonmissing values in the dependent variables. Simulation observations are not rated and so have missing values in the dependent variable. It is useful to create a format for the `weight` variable that distinguishes the three types of observations in the input and output data sets.

```
proc format;
   value wf 1  = 'Active'
            .  = 'Holdout'
            0  = 'Simulation';
   run;
```

PROC TRANSREG does not distinguish between weights that are zero, missing, or negative. All nonpositive weights exclude the observations from the analysis. The holdout and simulation observations are given different nonpositive values and a format to make them easy to distinguish in subsequent analyses and listings. The part-worth utilities for each attribute are computed using only those observations with positive weight. The predicted utility is computed for all products, even those with nonpositive weights.

Monotone, Spline, and Monotone Spline Comparisons

When you choose the transformation of the ratings or rankings, you choose among

`identity` - model the data directly

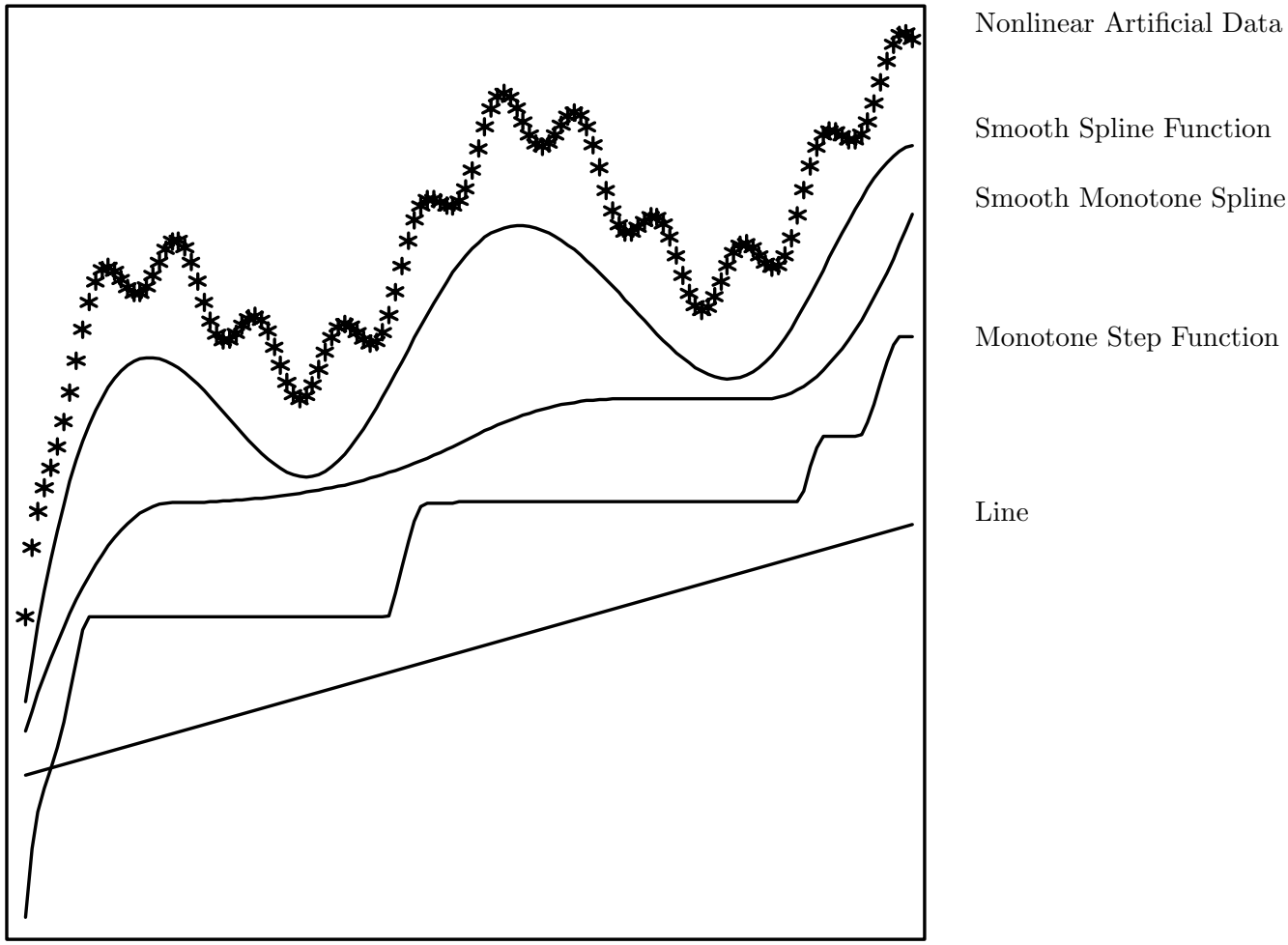
`monotone` - model an increasing step function of the data

`mspline` - model a nonlinear but smooth and increasing function of the data

`spline` - model a smooth function of the data

The following plot shows examples of the different types of functions you can fit in PROC TRANSREG. At the top of the plot are some artificial nonlinear data. Below that is a spline function, created by `spline`. It is smooth and nonlinear. It follows the overall shape of the data, but smoothes out the smaller bumps. Below that is a monotone spline function, created by `mspline`. Like the spline function, it is smooth and nonlinear. Unlike the spline function, it is monotonic. The function never decreases; it always rises or stays flat. The monotone spline function follows the overall upward trend in the data, and it shows the changes in upward trend, but it smoothes out all the dips and bumps in the function. Below the monotone spline function is a monotone step function, created by `monotone`. It is not smooth, but it is monotonic. Like the monotone spline, the monotone step function follows the

Functions Available in PROC TRANSREG



overall upward trend in the data, and it smoothes out all the dips and bumps in the function. However, the function is not smooth, and it typically requires many more parameters be fit than with monotone splines. Below the monotone step function is a line, created by `identity`. It is smooth and linear. It follows the overall upward trend in the data, but it smoothes over all the dips, bumps, and changes in upward trend.

Typical conjoint analyses are metric (using `identity`) or nonmetric (using `monotone`). While not often used in practice, monotone splines have a lot to recommend them. They allow for nonlinearities in the transformation of preference, but unlike `monotone`, they are smooth and do not use up all of your error *df*. One would typically never use `spline` on the ratings or rankings in a conjoint analysis, but if for some reason, you had a lot of price points,[‖] you could fit a spline function of the price attribute. This would allow for nonlinearities in preferences for different prices while constraining the part-worth utility function to be smooth.

[‖] For design efficiency reasons, you typically should not.

Samples of PROC TRANSREG Usage

Conjoint analysis can be performed in many ways with PROC TRANSREG. This section provides sample specifications for some typical and some more esoteric conjoint analyses. The dependent variables typically contain ratings or rankings of products by a number of subjects. The independent variables, `x1-x5`, are the attributes. For metric conjoint analysis, the dependent variable is designated `identity`. For nonmetric conjoint analysis, `monotone` is used. Attributes are usually designated as `class` variables with the restriction that the part-worth utilities within each attribute sum to zero.

The `utilities` option requests an overall ANOVA table, a table of part-worth utilities, their standard errors, and the importance of each attribute. The `p` (predicted values) option outputs to a data set the predicted utility for each product. The `ireplace` option suppresses the separate output of transformed independent variables since the independent variable transformations are the same as the raw independent variables. The `weight` variable is used to distinguish active observations from holdouts and simulation observations. The `reflect` transformation option reflects the transformation of the ranking so that large transformed values, positive utility, and positive evaluation will all correspond.

Today, metric conjoint analysis is used more often than nonmetric conjoint analysis, and rating-scale data are collected more often than rankings.

Metric Conjoint Analysis with Rating-Scale Data

This is a metric conjoint analysis with rating-scale data.

```
ods exclude notes mvanova anova;
proc transreg data=a utilities short method=morals;
   model identity(rating1-rating100) = class(x1-x5 / zero=sum);
   output p ireplace;
   weight w;
   run;
```

Nonmetric Conjoint Analysis

This is a nonmetric conjoint analysis specification, which has *many* parameters for the transformations.

```
ods exclude notes anova liberalanova conservanova
            mvanova liberalmvanova conservmvanova;
proc transreg data=a utilities short maxiter=500 method=morals;
   model monotone(ranking1-ranking100 / reflect) = class(x1-x5 / zero=sum);
   output p ireplace;
   weight w;
   run;
```

Monotone Splines

This is a conjoint analysis that is more restrictive than a nonmetric analysis but less restrictive than a metric conjoint analysis. By default, the monotone spline transformation has two parameters (degree two with no knots).

```
ods exclude notes anova liberalanova conservanova
            mvanova liberalmvanova conservmvanova;
proc transreg data=a utilities short maxiter=500 method=morals;
   model mspline(ranking1-ranking100 / reflect) =
         class(x1-x5 / zero=sum);
   output p ireplace;
   weight w;
   run;
```

If less smoothness is desired, specify knots. For example:

```
ods exclude notes anova liberalanova conservanova
            mvanova liberalmvanova conservmvanova;
proc transreg data=a utilities short maxiter=500 method=morals;
   model mspline(ranking1-ranking100 / reflect nknots=3) =
         class(x1-x5 / zero=sum);
   output p ireplace;
   weight w;
   run;
```

Each knot requires estimation of an additional parameter.

Constraints on the Utilities

Here is a metric conjoint analysis specification with linearity constraints imposed on `x4` and monotonicity constraints imposed on `x5`.

```
ods exclude notes anova liberalanova conservanova
            mvanova liberalmvanova conservmvanova
            liberalutilities liberalfitstatistics;
proc transreg data=a utilities short maxiter=500 method=morals;
   model identity(rating1-rating100) = class(x1-x3 / zero=sum)
         identity(x4) monotone(x5);
   output p ireplace;
   weight w;
   run;
```

With the monotonic constraints on the part-worth utilities, PROC TRANSREG prints some extra information, liberal and conservative part-worth utility and fit statistics tables. These tables report the same part-worth utilities, but are based on different methods of counting the number of parameters estimated. The liberal test tables can be suppressed by adding `liberalutilities liberalfitstatistics` to the `ods exclude` statement.

Here is another example, specifying monotonic step-function constraints on `x1-x5` and a smooth, monotonic transformation of price:

```
ods exclude notes anova liberalanova conservanova
            mvanova liberalmvanova conservmvanova
            liberalutilities liberalfitstatistics;
proc transreg data=a utilities short maxiter=500 method=morals;
   model identity(rating1-rating100) = monotone(x1-x5) mspline(price);
   output p ireplace;
   weight w;
   run;
```

A Discontinuous Price Function

The utility of price may not be a continuous function of price. It has been frequently found that utility is discontinuous at *round* numbers such as $1.00, $2.00, $100, $1000, and so on. If `price` has many values in the data set, say over the range $1.05 to $3.95, then a monotone function of price with discontinuities at $2.00 and $3.00 can be requested as follows.

```
ods exclude notes anova liberalanova conservanova
            mvanova liberalmvanova conservmvanova
            liberalutilities liberalfitstatistics;
proc transreg data=a utilities short maxiter=500 method=morals;
   model identity(rating1-rating100) =
         class(x1-x5 / zero=sum)
         mspline(price / knots=2 2 2 3 3 3);
   output p ireplace;
   weight w;
   run;
```

The monotone spline is degree two. The *order* of the spline is one greater than the degree; in this case the order is three. When the same knot value is specified *order* times, the transformation is discontinuous at the knot. Refer to Kuhfeld and Garratt (1992), page 643, for some applications of splines to conjoint analysis.

Experimental Design and Choice Modeling Macros

Warren F. Kuhfeld

Abstract

SAS provides a set of macros for designing experiments and analyzing choice data. Syntax and usage of these macros is discussed in this chapter. An additional macro for scatter plots of labeled points is documented here as well.*

Introduction

The following SAS autocall macros are used for generating efficient factorial designs, efficient choice designs, processing and evaluating designs, and processing data from choice experiments.

Macro	Page	Purpose
`%ChoicEff`	481	efficient choice design
`%MktAllo`	512	processes allocation data
`%MktBal`	515	balanced main-effects designs
`%MktBlock`	518	block a linear or choice design
`%MktDes`	527	efficient linear experimental design via candidate set search
`%MktDups`	534	identify duplicate choice sets or runs
`%MktEval`	542	evaluate an experimental design
`%MktEx`	546	efficient experimental designs, our flagship factorial-design macro
`%MktKey`	576	aid creation of the `key=` data set
`%MktLab`	577	relabel, rename and assign levels to design factors
`%MktMerge`	588	merges a choice design with choice data
`%MktOrth`	590	lists 12,356 of the orthogonal designs that `%MktEx` can make
`%MktRoll`	595	rolls a linear design into a choice design
`%MktRuns`	600	experimental design size, aids in selecting the number or runs
`%PhChoice`	606	customizes the printed output from a choice model
`%PlotIt`	611	graphical scatter plots of labeled points

*Copies of this chapter (TS-689H) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market .

Once you have these macros installed, you can call them and use them just as you would use a SAS procedure. Only the syntax for macros is a little different from SAS procedures. Here is an example of making an experimental design with a two-level and 7 three-level factors with 18 runs or profiles.

```
%mktex( 2 3 ** 7, n=18 )
```

Installation

If your site has installed the autocall libraries supplied by SAS and uses the standard configuration of SAS supplied software, you need to ensure that the SAS system option `mautosource` is in effect to begin using the autocall macros. Note however, that if you are using any SAS Version released prior to SAS 9.1, there are differences between the macros used in this documentation and those that were shipped with your Version of SAS. Be sure to get the latest macros from Warren.Kuhfeld@sas.com or the Web. These macros will not work with Version 6.12, however they should work with Version 8.2 and later versions of SAS. You should install *all* of these macros, not just one or some. Some of the macros call other macros and will not work if the other macros are not there or if only older versions of the other macros are there. For example, the `%MktEx` macro calls: the `%MktRuns` macro to parse the factors list, the `%MktDes` macro for candidate set generation and search, and the `%MktOrth` macro to get the orthogonal array catalog.

The macros do not have to be included (for example, with a `%include` statement). They can be called directly once they are properly installed. For more information about autocall libraries, refer to *SAS Macro Language: Reference*. On a PC for example, the autocall library may be installed in the `stat\sasmacro` directory off of your SAS root directory. The name of your SAS root directory could be anything, but it is probably something like SAS or SAS\V8. One way to find the right directory is to use `Start` → `Find` to find one of the existing autocall macros such as `mktdes.sas` or `plotit.sas`. Unix should have a similar directory structure to the PC. The autocall library in Unix may be installed in the `stat/sasmacro` directory off of your SAS root directory. On MVS, each macro will be a different member of a PDS. For details on installing autocall macros, consult your host documentation.

Usually, an autocall library is a directory containing individual files, each of which contains one macro definition. An autocall library can also be a SAS catalog. To use a directory as a SAS autocall library, store the source code for each macro in a separate file in the directory. The name of the file must be the same as the macro name, typically followed by `.sas`. For example, the macro `%MktEx` must typically be stored in a file named `mktex.sas`. On most hosts, the reserved `fileref sasautos` is assigned at invocation time to the autocall library supplied by SAS or another one designated by your site. If you are specifying your own autocall libraries, remember to concatenate the autocall library supplied by SAS with your autocall libraries so that these macros will also be available. For details, refer to your host documentation and SAS macro language documentation.

%ChoicEff Macro

The %ChoicEff autocall macro is used to find efficient experimental designs for choice experiments. See pages 260, 303, 333, and 340 for examples. You supply sets of candidate alternatives. The macro searches the candidates for an efficient experimental design − a design in which the variances of the parameter estimates are minimized, given an assumed parameter vector $\boldsymbol{\beta}$.

There are two ways you can use the macro:

- You can create a candidate set of alternatives, and the macro will create a design consisting of choice sets built from the alternatives you supplied. You must designate for each candidate alternative the design alternative(s) for which it is a candidate. For a branded study with m brands, you must create m lists of candidate alternatives, one for each brand.
- You can create a candidate set of choice sets, and the macro will build a design from the choice sets that you supplied. Typically, you would only use this approach when there are restrictions across alternatives (certain alternatives may not appear with certain other alternatives).

The %ChoicEff macro uses a modified Fedorov candidate-set-search algorithm, just like PROC OPTEX and the %MktEx macro. Typically, you use as a candidate set a full-factorial, fractional-factorial, or a tabled design created with the %MktEx macro. First, the %ChoicEff macro either constructs a random initial design from the candidates or it uses an initial design that you specified. The macro considers swapping out every design alternative/set and replacing it with each candidate alternative/set. Swaps that increase efficiency are performed. The process of evaluating and swapping continues until efficiency stabilizes. This process is repeated with different initial designs, and the best design is output for use. The key differences between the %ChoicEff macro and the %MktEx macro are as follows. The %ChoicEff macro requires you to specify the true (or assumed true) parameters and it optimizes the variance matrix for a multinomial logit model, whereas the %MktEx macro optimizes the variance matrix for a linear model, which does not depend on the parameters.

Here is an example. This example creates a design for a generic model with 3 three-level factors. First, the %MktEx macro is used to create a set of candidate alternatives, where x1-x3 are the factors. Note that the n= specification allows expressions. Our candidate set must also contain flag variables, one for each alternative, that flag which candidates can be used for which alternative(s). Since this is a generic model, each candidate can appear in any alternative, so we need to add flags that are constant: f1=1 f2=1 f3=1. The %MktEx macro does not allow you to create constant factors. Instead, we can use the %MktLab macro to add the flag variables, essentially by specifying that we have multiple intercepts. The option int=f1-f3 creates three variables with values all one. The default output data set is called FINAL. Next, the %ChoicEff macro is run to find an efficient design for the unbranded, purely generic model assuming $\boldsymbol{\beta} = \mathbf{0}$. Here is the code.

```
%mktex(3 ** 3, n=3**3, seed=238)
%mktlab(int=f1-f3)

%choiceff(data=final, model=class(x1-x3), nsets=9,
          flags=f1-f3, beta=zero, seed=145)

proc print; var x1-x3; id set; by set; run;
```

The option `data=final` names the input data set, `model=class(x1-x3)` specifies the PROC TRANSREG `model` statement for coding the design, `nsets=9` specifies nine choice sets, `flags=f1-f3` specifies the three alternative flag variables, `beta=zero` specifies all zero parameters, and `seed=145` specifies the random number seed. Here is the output.

n	Name	Beta	Label
1	x11	0	x1 1
2	x12	0	x1 2
3	x21	0	x2 1
4	x22	0	x2 2
5	x31	0	x3 1
6	x32	0	x3 2

Design	Iteration	D-Efficiency	D-Error
1	0	0.720845	1.387261
	1	1.636787	0.610953
	2	1.657858	0.603188
	3	1.678027	0.595938
	4	1.678027	0.595938

Design	Iteration	D-Efficiency	D-Error
2	0	1.078369	0.927326
	1	1.702866	0.587245
	2	1.702866	0.587245

Final Results

Design	2
Choice Sets	9
Alternatives	3
D-Efficiency	1.702866
D-Error	0.587245

n	Variable Name	Label	Variance	DF	Standard Error
1	x11	x1 1	0.71089	1	0.84314
2	x12	x1 2	0.68354	1	0.82677
3	x21	x2 1	0.67544	1	0.82185
4	x22	x2 2	0.67544	1	0.82185
5	x31	x3 1	0.67544	1	0.82185
6	x32	x3 2	0.67544	1	0.82185
				==	
				6	

Set	x1	x2	x3
1	3	3	1
	1	2	2
	2	1	3
2	3	1	2
	2	3	1
	1	2	3
3	3	1	3
	2	3	2
	1	2	1
4	1	1	2
	3	3	3
	2	2	1
5	3	1	2
	2	2	3
	1	3	1
6	3	2	1
	1	1	3
	2	3	2
7	1	2	2
	3	3	3
	2	1	1
8	3	1	1
	1	3	2
	2	2	3
9	2	1	1
	1	3	3
	3	2	2

The output from the `%ChoicEff` macro consists of a list of the parameter names, values and labels, followed by two iteration histories (each based on a different random initial design), then a brief report on the most efficient design found, and finally a table with the parameter names, variances, *df*, and standard errors. The design is printed using PROC PRINT.

Here is another example. These next steps directly create an optimal design for this generic model and evaluate its efficiency using the `%ChoicEff` macro and the initial design options. The DATA step creates a cyclic design. In a cyclic design, the factor levels increase cyclically from one alternative to the next. The levels for a factor for the three alternatives will always be one of the following: (1, 2, 3) or (2, 3, 1) or (3, 1, 2).

```
* Cyclic (Optimal) Design;
data x(keep=f1-f3 x1-x3);
   retain f1-f3 1;
   d1 = ceil(_n_ / 3); d2 = mod(_n_ - 1, 3) + 1; input d3 @@;
   do i = -1 to 1;
      x1 = mod(d1 + i, 3) + 1;
      x2 = mod(d2 + i, 3) + 1;
      x3 = mod(d3 + i, 3) + 1;
      output;
      end;
   datalines;
1 2 3 3 1 2 2 3 1
;

proc print data=x; var x: f:; run;
```

Here is part of the cyclic design. Notice the cyclical pattern. Each level in the second or third alternative is one greater than the level in the previous alternative, where 3+1 is defined to be 1. The flag variables `f1-f3` contain all ones showing that each candidate can be used in any alternative.

Obs	x1	x2	x3	f1	f2	f3
1	1	1	1	1	1	1
2	2	2	2	1	1	1
3	3	3	3	1	1	1
4	1	2	2	1	1	1
5	2	3	3	1	1	1
6	3	1	1	1	1	1
.						
.						
.						
25	3	3	1	1	1	1
26	1	1	2	1	1	1
27	2	2	3	1	1	1

This is the code that evaluates the design.

```
%choiceff(data=x, model=class(x1-x3), nsets=9, flags=f1-f3,
          beta=zero, init=x, initvars=x1-x3, intiter=0);
```

The option `init=x` specifies the initial design, `initvars=x1-x3` specifies the factors in the initial design, and `intiter=0` specifies the number of internal iterations. Specify `intiter=0` when you just want to evaluate the efficiency of a given design. Here is the output from the `%ChoicEff` macro.

```
                        n    Name    Beta    Label

                        1    x11        0     x1 1
                        2    x12        0     x1 2
                        3    x21        0     x2 1
                        4    x22        0     x2 2
                        5    x31        0     x3 1
                        6    x32        0     x3 2

Design   Iteration  D-Efficiency       D-Error
-----------------------------------------------
   1         0          1.732051      0.577350

                          Final Results

                        Design                  1
                        Choice Sets             9
                        Alternatives            3
                        D-Efficiency     1.732051
                        D-Error          0.577350

             Variable                                    Standard
          n    Name      Label    Variance     DF        Error

          1    x11       x1 1      0.66667      1      0.81650
          2    x12       x1 2      0.66667      1      0.81650
          3    x21       x2 1      0.66667      1      0.81650
          4    x22       x2 2      0.66667      1      0.81650
          5    x31       x3 1      0.66667      1      0.81650
          6    x32       x3 2      0.66667      1      0.81650
                                                ==
                                                 6
```

These next steps use the `%MktEx` and `%MktRoll` macros to create a candidate set of choice sets and the `%ChoicEff` macro to search for an efficient design using the candidate-set-swapping algorithm.

```
%mktex(3 ** 9, n=2187)

data key;
   input (x1-x3) ($);
   datalines;
x1 x2 x3
x4 x5 x6
x7 x8 x9
;
%mktroll(design=design, key=key, out=rolled)

%choiceff(data=rolled, model=class(x1-x3), nsets=9, nalts=3,
          beta=zero, seed=17);
```

The first steps create a candidate set of choice sets. The %MktEx macro creates a design with nine factors, three for each of the three alternatives. The KEY data set specifies that the first alternative is made from the linear design factors x1-x3, the second alternative is made from x4-x6, and the third alternative is made from x7-x9. The %MktRoll macro turns a linear design into a choice design using the rules specified in the KEY data set.

In the %ChoicEff macro, the nalts=3 option specifies that there are three alternatives. There must always be a constant number of alternatives in each choice set, even if all of the alternatives will not be used. When a nonconstant number of alternatives is desired, you must use a weight variable to flag those alternatives that the subject will not see. When you swap choice sets, you need to specify nalts=. The output from these steps is not appreciably different from what we saw previously, so it is not shown.

This next example has brand effects and uses the alternative-swapping algorithm.

```
%mktex(3 ** 4, n = 3**4)
%mktlab(data=design, vars=x1-x3 Brand)

data full(drop=i);
   set final;
   array f[3];
   do i = 1 to 3; f[i] = (brand eq i); end;
   run;

proc print data=full(obs=9); run;
```

The %MktEx macro makes the linear candidate design. The %MktLab macro changes the name of the variable x4 to Brand while retaining the original names for x1-x3 and original values for all factors of 1, 2, and 3. The DATA step creates the flags. The flag f1 flags brand 1 candidates as available for the first alternative, f2 flags brand 2 candidates as available for the second alternative, and so on. The Boolean expression (brand eq i) evaluates to 1 if true and 0 if false. Here is the first part of the candidate set.

Obs	x1	x2	x3	Brand	f1	f2	f3
1	1	1	1	1	1	0	0
2	1	1	1	2	0	1	0
3	1	1	1	3	0	0	1
4	1	1	2	1	1	0	0
5	1	1	2	2	0	1	0
6	1	1	2	3	0	0	1
7	1	1	3	1	1	0	0
8	1	1	3	2	0	1	0
9	1	1	3	3	0	0	1

Here is the %ChoicEff macro call for making the design.

```
%choiceff(data=full, seed=151,
          model=class(brand brand*x1 brand*x2 brand*x3 / zero=' '),
          nsets=15, flags=f1-f3, beta=zero, converge=1e-12);
```

The `model=` specification states that `Brand` and `x1-x3` are classification or categorical variables and brand effects and brand by attribute interactions (alternative-specific effects) (see page 171) are desired. The `zero=' '` specification is like `zero=none` except `zero=none` applies to all factors in the specification whereas `zero=' '` applies to just the first. This `zero=' '` specification specifies that there is no reference level for the first factor (`Brand`), and last level will by default be the reference category for the other factors (`x1-x3`). Hence, binary variables are created for all three brands, but only two binary variables are created for the 3 three-level factors. We need to do this because we need the alternative-specific effects for all brands, including Brand 3. Notice that the candidate set consists of branded alternatives with flags such that only brand n is considered for the nth alternative of each choice set. In the interest of space, not all of the output is shown. Here is some of the output.

```
Design   Iteration  D-Efficiency       D-Error
-----------------------------------------------
   1         0                 0             .
             1                 0             .
               0.300256 (Ridged)
             2                 0             .
               0.302184 (Ridged)
             3                 0             .
               0.303659 (Ridged)
             4                 0             .
               0.305192 (Ridged)
             5                 0             .
               0.305192 (Ridged)

Design   Iteration  D-Efficiency       D-Error
-----------------------------------------------
   2         0                 0             .
             1                 0             .
               0.295570 (Ridged)
             2                 0             .
               0.303106 (Ridged)
             .
             .
             .
             7                 0             .
               0.304929 (Ridged)

                Final Results

          Design                 1
          Choice Sets           15
          Alternatives           3
          D-Efficiency           0
          D-Error                .
```

n	Variable Name	Label	Variance	DF	Standard Error
1	Brand1	Brand 1	5.70845	1	2.38924
2	Brand2	Brand 2	3.89246	1	1.97293
3	Brand3	Brand 3	.	0	.
4	Brand1x11	Brand 1 * x1 1	1.99395	1	1.41207
5	Brand1x12	Brand 1 * x1 2	2.09289	1	1.44668
6	Brand2x11	Brand 2 * x1 1	1.80635	1	1.34400
7	Brand2x12	Brand 2 * x1 2	2.09299	1	1.44672
8	Brand3x11	Brand 3 * x1 1	2.12281	1	1.45699
9	Brand3x12	Brand 3 * x1 2	2.16706	1	1.47209
.					
.					
.					
21	Brand3x32	Brand 3 * x3 2	2.18010	1	1.47652
				==	
				20	

The following is printed to the log.

```
Redundant Variables:

Brand3
```

Notice that at each step, the efficiency is zero, but a nonzero ridged value is printed. This model contains a structural-zero coefficient in `Brand3`. While we need alternative-specific effects for Brand 3 (like `Brand3x11` and `Brand3x12`), we do not need the Brand 3 effect (`Brand3`) This can be seen from both the `Redundant Variables` list and from looking at the variance and *df* table. The inclusion of the `Brand3` term in the model makes the efficiency of the design zero. However, the `%ChoicEff` macro can still optimize the goodness of the design by optimizing a ridged efficiency criterion. That is what is shown in the iteration history. The option `converge=1e-12` was specified because for this example, iteration stops prematurely with the default convergence criterion. This next step switches to a full-rank coding, dropping the redundant variable `Brand3`, and using the output from the last step as the initial design.

```
%choiceff(data=full, init=best(keep=index), drop=brand3, seed=522,
          model=class(brand brand*x1 brand*x2 brand*x3 / zero=' '),
          nsets=15, flags=f1-f3, beta=zero, converge=1e-12);
```

The option `drop=brand3` is used to drop the parameter with the zero coefficient. We could have moved the brand specification into its own `class` specification (separate from the alternative-specific effects) and not specified `zero=' '` with it (see for example page 489). However, sometimes it is easier to specify a model with more terms than you really need, and then list the terms to drop, so that is what we illustrate here.

In this usage of `init=` with alternative swapping, the only part of the initial design that is required is the `Index` variable. It contains indices into the candidate set of the alternatives that are used to make the initial design. This usage is for the situation where the initial design was output from the macro. (In contrast, in the example usage on page 483, the option `initvars=x1-x3` was specified because the initial design was not created by the `%ChoicEff` macro.) Here is some of the output. Notice that now there are no zero parameters so *D*-efficiency can be directly computed.

Design	Iteration	D-Efficiency	D-Error
1	0	0.683825	1.462362
	1	0.683825	1.462362

Final Results

Design	1
Choice Sets	15
Alternatives	3
D-Efficiency	0.683825
D-Error	1.462362

n	Variable Name	Label	Variance	DF	Standard Error
1	Brand1	Brand 1	5.70845	1	2.38924
2	Brand2	Brand 2	3.89246	1	1.97293
3	Brand1x11	Brand 1 * x1 1	1.99395	1	1.41207
4	Brand1x12	Brand 1 * x1 2	2.09289	1	1.44668
5	Brand2x11	Brand 2 * x1 1	1.80635	1	1.34400
6	Brand2x12	Brand 2 * x1 2	2.09299	1	1.44672
7	Brand3x11	Brand 3 * x1 1	2.12281	1	1.45699
8	Brand3x12	Brand 3 * x1 2	2.16706	1	1.47209
9	Brand1x21	Brand 1 * x2 1	2.38373	1	1.54393
10	Brand1x22	Brand 1 * x2 2	2.17427	1	1.47454
11	Brand2x21	Brand 2 * x2 1	2.28422	1	1.51136
12	Brand2x22	Brand 2 * x2 2	2.25113	1	1.50038
13	Brand3x21	Brand 3 * x2 1	2.21430	1	1.48805
14	Brand3x22	Brand 3 * x2 2	2.09383	1	1.44701
15	Brand1x31	Brand 1 * x3 1	2.39416	1	1.54731
16	Brand1x32	Brand 1 * x3 2	2.10564	1	1.45108
17	Brand2x31	Brand 2 * x3 1	2.55697	1	1.59905
18	Brand2x32	Brand 2 * x3 2	2.25251	1	1.50084
19	Brand3x31	Brand 3 * x3 1	2.29769	1	1.51581
20	Brand3x32	Brand 3 * x3 2	2.18010	1	1.47652
				==	
				20	

These next steps handle the same problem, only this time, we use the set-swapping algorithm, and we will specify a parameter vector that is not zero. At first, we will omit the `beta=` option, just to see the coding. We specified the `effects` option in the PROC TRANSREG `class` specification to get -1, 0, 1 coding.

```
%mktex(3 ** 9, n=2187, seed=121)

data key;
   input (Brand x1-x3) ($);
   datalines;
1 x1 x2 x3
2 x4 x5 x6
3 x7 x8 x9
;

%mktroll(design=design, key=key, alt=brand, out=rolled)

%choiceff(data=rolled, nsets=15, nalts=3,
          model=class(brand)
                class(brand*x1 brand*x2 brand*x3 / effects zero=' '))
```

The output tells us the parameter names and the order in which we need to specify parameters.

n	Name	Beta	Label
1	Brand1	.	Brand 1
2	Brand2	.	Brand 2
3	Brand1x11	.	Brand 1 * x1 1
4	Brand1x12	.	Brand 1 * x1 2
5	Brand2x11	.	Brand 2 * x1 1
6	Brand2x12	.	Brand 2 * x1 2
7	Brand3x11	.	Brand 3 * x1 1
8	Brand3x12	.	Brand 3 * x1 2
9	Brand1x21	.	Brand 1 * x2 1
10	Brand1x22	.	Brand 1 * x2 2
11	Brand2x21	.	Brand 2 * x2 1
12	Brand2x22	.	Brand 2 * x2 2
13	Brand3x21	.	Brand 3 * x2 1
14	Brand3x22	.	Brand 3 * x2 2
15	Brand1x31	.	Brand 1 * x3 1
16	Brand1x32	.	Brand 1 * x3 2
17	Brand2x31	.	Brand 2 * x3 1
18	Brand2x32	.	Brand 2 * x3 2
19	Brand3x31	.	Brand 3 * x3 1
20	Brand3x32	.	Brand 3 * x3 2

Now that we are sure we know the order of the parameters, we can specify the assumed betas on the `beta=` option. These numbers are based on prior research or our expectations of approximately what we expect the parameter estimates will be. We also specified `n=100` on this run, which is a sample size we are considering.

```
%choiceff(data=rolled, nsets=15, nalts=3, n=100, seed=462,
          beta=1 2 -0.5 0.5 -0.75 0.75 -1 1
               -0.5 0.5 -0.75 0.75 -1 1 -0.5 0.5 -0.75 0.75 -1 1,
          model=class(brand)
                class(brand*x1 brand*x2 brand*x3 / effects zero=' '))
```

Here is some of the output. Notice that parameters and test statistics are incorporated into the output. The `n=` value is incorporated into the variance matrix and hence the efficiency statistics, variances and tests.

n	Variable Name	Label	Variance	Assumed Beta	DF	Standard Error	Wald	Prob > Squared Wald
1	Brand1	Brand 1	0.011889	1.00	1	0.10903	9.1714	0.0001
2	Brand2	Brand 2	0.020697	2.00	1	0.14386	13.9021	0.0001
3	Brand1x11	Brand 1 * x1 1	0.008617	-0.50	1	0.09283	-5.3865	0.0001
4	Brand1x12	Brand 1 * x1 2	0.008527	0.50	1	0.09234	5.4147	0.0001
5	Brand2x11	Brand 2 * x1 1	0.009283	-0.75	1	0.09635	-7.7842	0.0001
6	Brand2x12	Brand 2 * x1 2	0.012453	0.75	1	0.11159	6.7208	0.0001
7	Brand3x11	Brand 3 * x1 1	0.021764	-1.00	1	0.14753	-6.7784	0.0001
8	Brand3x12	Brand 3 * x1 2	0.015657	1.00	1	0.12513	7.9917	0.0001
9	Brand1x21	Brand 1 * x2 1	0.012520	-0.50	1	0.11189	-4.4685	0.0001
10	Brand1x22	Brand 1 * x2 2	0.010685	0.50	1	0.10337	4.8370	0.0001
11	Brand2x21	Brand 2 * x2 1	0.010545	-0.75	1	0.10269	-7.3035	0.0001
12	Brand2x22	Brand 2 * x2 2	0.012654	0.75	1	0.11249	6.6672	0.0001
13	Brand3x21	Brand 3 * x2 1	0.018279	-1.00	1	0.13520	-7.3964	0.0001
14	Brand3x22	Brand 3 * x2 2	0.012117	1.00	1	0.11008	9.0845	0.0001
15	Brand1x31	Brand 1 * x3 1	0.009697	-0.50	1	0.09848	-5.0774	0.0001
16	Brand1x32	Brand 1 * x3 2	0.010787	0.50	1	0.10386	4.8141	0.0001
17	Brand2x31	Brand 2 * x3 1	0.009203	-0.75	1	0.09593	-7.8181	0.0001
18	Brand2x32	Brand 2 * x3 2	0.013923	0.75	1	0.11800	6.3562	0.0001
19	Brand3x31	Brand 3 * x3 1	0.016546	-1.00	1	0.12863	-7.7742	0.0001
20	Brand3x32	Brand 3 * x3 2	0.014235	1.00	1	0.11931	8.3815	0.0001
					==			
					20			

These next steps create a design for a cross-effects model with five brands at three prices and a constant alternative. (See the examples beginning on pages 205 and 228 for more information on cross effects.) Note the choice-set-swapping algorithm can handle cross effects but not the alternative-swapping algorithm.

```
%mktex(3 ** 5, n=3**5)
```

```
data key;
   input (Brand Price) ($);
   datalines;
1 x1
2 x2
3 x3
4 x4
5 x5
. .
;
%mktroll(design=design, key=key, alt=brand, out=rolled, keep=x1-x5)

proc print; by set; id set; where set in (1, 48, 101, 243); run;
```

The `keep=` option on the `%MktRoll` macro is used to keep the price variables that are needed to make the cross effects. Here are a few of the candidate choice sets.

Set	Brand	Price	x1	x2	x3	x4	x5
1	1	1	1	1	1	1	1
	2	1	1	1	1	1	1
	3	1	1	1	1	1	1
	4	1	1	1	1	1	1
	5	1	1	1	1	1	1
		.	1	1	1	1	1
48	1	1	1	2	3	1	3
	2	2	1	2	3	1	3
	3	3	1	2	3	1	3
	4	1	1	2	3	1	3
	5	3	1	2	3	1	3
		.	1	2	3	1	3
101	1	2	2	1	3	1	2
	2	1	2	1	3	1	2
	3	3	2	1	3	1	2
	4	1	2	1	3	1	2
	5	2	2	1	3	1	2
		.	2	1	3	1	2
243	1	3	3	3	3	3	3
	2	3	3	3	3	3	3
	3	3	3	3	3	3	3
	4	3	3	3	3	3	3
	5	3	3	3	3	3	3
		.	3	3	3	3	3

Notice that `x1` contains the price for Brand 1, `x2` contains the price for Brand 2, and so on, and the price of brand i in a choice set is the same, no matter which alternative it is stored with.

Here is the `%ChoicEff` macro call for creating the choice design with cross effects.

```
%choiceff(data=rolled, seed=17,
          model=class(brand brand*price / zero=none)
                identity(x1-x5) * class(brand / zero=none),
          nsets=20, nalts=6, beta=zero);
```

Cross effects are created by interacting the price factors with brand. See pages 212 and 260 for more information about cross effects.

Here is the redundant variable list from the log.

```
Redundant Variables:

Brand1Price3 Brand2Price3 Brand3Price3 Brand4Price3 Brand5Price3 x1Brand1
x2Brand2 x3Brand3 x4Brand4 x5Brand5
```

Next, we will run the macro again, this time requesting a full-rank model. The list of dropped names was created by copying from the redundant variable list. Also, `zero=none` was changed to `zero=' '` so no level would be zeroed for `Brand` but the last level of `Price` would be zeroed.

```
%choiceff(data=rolled, seed=17,
          model=class(brand brand*price / zero=' ')
                identity(x1-x5) * class(brand / zero=none),
          drop=x1Brand1 x2Brand2 x3Brand3 x4Brand4 x5Brand5,
          nsets=20, nalts=6, beta=zero);
```

Here is the last part of the output. Notice that we have five brand parameters, two price parameters for each of the five brands, and four cross effect parameters for each of the five brands.

n	Variable Name	Label	Variance	DF	Standard Error
1	Brand1	Brand 1	13.8149	1	3.71683
2	Brand2	Brand 2	13.5263	1	3.67782
3	Brand3	Brand 3	13.2895	1	3.64547
4	Brand4	Brand 4	13.5224	1	3.67728
5	Brand5	Brand 5	16.3216	1	4.04000
6	Brand1Price1	Brand 1 * Price 1	2.8825	1	1.69779
7	Brand1Price2	Brand 1 * Price 2	3.5118	1	1.87399
8	Brand2Price1	Brand 2 * Price 1	2.8710	1	1.69441
9	Brand2Price2	Brand 2 * Price 2	3.5999	1	1.89733
10	Brand3Price1	Brand 3 * Price 1	2.8713	1	1.69448
11	Brand3Price2	Brand 3 * Price 2	3.5972	1	1.89662
12	Brand4Price1	Brand 4 * Price 1	2.8710	1	1.69441
13	Brand4Price2	Brand 4 * Price 2	3.5560	1	1.88574
14	Brand5Price1	Brand 5 * Price 1	2.8443	1	1.68649
15	Brand5Price2	Brand 5 * Price 2	3.8397	1	1.95953
16	x1Brand2	x1 * Brand 2	0.7204	1	0.84878
17	x1Brand3	x1 * Brand 3	0.7209	1	0.84908
18	x1Brand4	x1 * Brand 4	0.7204	1	0.84878
19	x1Brand5	x1 * Brand 5	0.7204	1	0.84877

```
20    x2Brand1        x2 * Brand 1          0.7178     1     0.84722
21    x2Brand3        x2 * Brand 3          0.7178     1     0.84724
22    x2Brand4        x2 * Brand 4          0.7178     1     0.84720
23    x2Brand5        x2 * Brand 5          0.7248     1     0.85133
24    x3Brand1        x3 * Brand 1          0.7178     1     0.84722
25    x3Brand2        x3 * Brand 2          0.7178     1     0.84721
26    x3Brand4        x3 * Brand 4          0.7178     1     0.84720
27    x3Brand5        x3 * Brand 5          0.7248     1     0.85133
28    x4Brand1        x4 * Brand 1          0.7178     1     0.84722
29    x4Brand2        x4 * Brand 2          0.7178     1     0.84721
30    x4Brand3        x4 * Brand 3          0.7178     1     0.84724
31    x4Brand5        x4 * Brand 5          0.7293     1     0.85402
32    x5Brand1        x5 * Brand 1          0.7111     1     0.84325
33    x5Brand2        x5 * Brand 2          0.7180     1     0.84737
34    x5Brand3        x5 * Brand 3          0.7248     1     0.85135
35    x5Brand4        x5 * Brand 4          0.7179     1     0.84731
                                                      ==
                                                      35
```

In this final `%ChoicEff` macro example, the goal is to create a design for a pricing study with ten brands plus a constant alternative. Each brand has a single attribute, price. However, the prices are potentially different for each brand and they do not even have the same numbers of levels. A model is desired with brand and alternative-specific price effects. Here are the design specifications.

Brand	Levels	Prices
Brand 1	8	0.89 0.94 0.99 1.04 1.09 1.14 1.19 1.24
Brand 2	8	0.94 0.99 1.04 1.09 1.14 1.19 1.24 1.29
Brand 3	6	0.99 1.04 1.09 1.14 1.19 1.24
Brand 4	6	0.89 0.94 0.99 1.04 1.09 1.14
Brand 5	6	1.04 1.09 1.14 1.19 1.24 1.29
Brand 6	4	0.89 0.99 1.09 1.19
Brand 7	4	0.99 1.09 1.19 1.29
Brand 8	4	0.94 0.99 1.14 1.19
Brand 9	4	1.09 1.14 1.19 1.24
Brand 10	4	1.14 1.19 1.24 1.29

There are two challenging aspects of this problem: creating the candidate set and coping with the price asymmetries. The candidate set must contain 8 rows for the eight Brand 1 prices, 8 rows for the eight Brand 2 prices, 6 rows for the six Brand 3 prices, ..., and 4 rows for the four Brand 10 prices. It also must contain a constant alternative. Furthermore, if we are to use the alternative-swapping algorithm, the candidate set must contain 11 flag variables, each of which will designate the appropriate group of alternatives. We could run the `%MktEx` macro ten times to make a candidate set for each of the brands, but since we have only one factor per brand, it would be much easier to generate the candidate set with a DATA step. Before we discuss the code, here is the candidate set that we need.

Brand	Price	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11
1	1	1	0	0	0	0	0	0	0	0	0	0
1	2	1	0	0	0	0	0	0	0	0	0	0
1	3	1	0	0	0	0	0	0	0	0	0	0
1	4	1	0	0	0	0	0	0	0	0	0	0
1	5	1	0	0	0	0	0	0	0	0	0	0
1	6	1	0	0	0	0	0	0	0	0	0	0
1	7	1	0	0	0	0	0	0	0	0	0	0
1	8	1	0	0	0	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0	0	0	0	0
2	2	0	1	0	0	0	0	0	0	0	0	0
2	3	0	1	0	0	0	0	0	0	0	0	0
2	4	0	1	0	0	0	0	0	0	0	0	0
2	5	0	1	0	0	0	0	0	0	0	0	0
2	6	0	1	0	0	0	0	0	0	0	0	0
2	7	0	1	0	0	0	0	0	0	0	0	0
2	8	0	1	0	0	0	0	0	0	0	0	0
3	1	0	0	1	0	0	0	0	0	0	0	0
3	2	0	0	1	0	0	0	0	0	0	0	0
3	3	0	0	1	0	0	0	0	0	0	0	0
3	4	0	0	1	0	0	0	0	0	0	0	0
3	5	0	0	1	0	0	0	0	0	0	0	0
3	6	0	0	1	0	0	0	0	0	0	0	0
4	1	0	0	0	1	0	0	0	0	0	0	0
4	2	0	0	0	1	0	0	0	0	0	0	0
4	3	0	0	0	1	0	0	0	0	0	0	0
4	4	0	0	0	1	0	0	0	0	0	0	0
4	5	0	0	0	1	0	0	0	0	0	0	0
4	6	0	0	0	1	0	0	0	0	0	0	0
5	1	0	0	0	0	1	0	0	0	0	0	0
5	2	0	0	0	0	1	0	0	0	0	0	0
5	3	0	0	0	0	1	0	0	0	0	0	0
5	4	0	0	0	0	1	0	0	0	0	0	0
5	5	0	0	0	0	1	0	0	0	0	0	0
5	6	0	0	0	0	1	0	0	0	0	0	0
6	1	0	0	0	0	0	1	0	0	0	0	0
6	2	0	0	0	0	0	1	0	0	0	0	0
6	3	0	0	0	0	0	1	0	0	0	0	0
6	4	0	0	0	0	0	1	0	0	0	0	0
7	1	0	0	0	0	0	0	1	0	0	0	0
7	2	0	0	0	0	0	0	1	0	0	0	0
7	3	0	0	0	0	0	0	1	0	0	0	0
7	4	0	0	0	0	0	0	1	0	0	0	0
8	1	0	0	0	0	0	0	0	1	0	0	0
8	2	0	0	0	0	0	0	0	1	0	0	0
8	3	0	0	0	0	0	0	0	1	0	0	0
8	4	0	0	0	0	0	0	0	1	0	0	0

9	1	0	0	0	0	0	0	0	0	1	0	0
9	2	0	0	0	0	0	0	0	0	1	0	0
9	3	0	0	0	0	0	0	0	0	1	0	0
9	4	0	0	0	0	0	0	0	0	1	0	0
10	1	0	0	0	0	0	0	0	0	0	1	0
10	2	0	0	0	0	0	0	0	0	0	1	0
10	3	0	0	0	0	0	0	0	0	0	1	0
10	4	0	0	0	0	0	0	0	0	0	1	0
.	.	0	0	0	0	0	0	0	0	0	0	1

It begins with eight alternatives for the eight prices for the first brand (`Brand` = 1 `f1` = 1, `f2-f11` = 0). It is followed by eight alternatives for the eight prices for the second brand (`Brand` = 2 `f2` = 1, `f1` = 0, `f3` through `f11` = 0). At the end is the constant alternative. For now, we do not need to worry about the actual price levels, since price will be treated as a qualitative factor. Here is the code that generated the candidate design.

```
data cand;
   array n[10] _temporary_ (8 8 6 6 6 4 4 4 4 4);
   retain f1-f11 0;
   array f[11];
   do Brand = 1 to 10;
      f[brand] = 1;
      do Price = 1 to n[brand]; output; end;
      f[brand] = 0;
      end;
   brand = .; price = .; f11 = 1; output;
   run;

proc print; id brand price; run;
```

It has the statement `do Brand = 1 to 10` that creates the ten brands, plus an `output` statement at the end to generate the constant alternative. Inside the `do Brand` loop is another `do` loop that creates the `n[brand]` prices. The `n` array is a temporary array, which means it will not create any variables to go into the output data set. It just gives us a convenient way to access the number of levels for each of the ten brands.

This call to the `%ChoicEff` macro creates the design naming `Brand` and `Price` as classification variables. Dummy variables will be created for all nonmissing levels of brand.

```
%choiceff(data=cand, seed=462,
          model=class(brand / zero=none) class(brand*price / zero=' '),
          nsets=24, flags=f1-f11, beta=zero)
```

Here is some of the output.

Design	Iteration	D-Efficiency	D-Error
1	0	0	.
	1	0	.
		0.001445 (Ridged)	
	2	0	.
		0.001445 (Ridged)	

Design	Iteration	D-Efficiency	D-Error
2	0	0	.
	1	0	.
		0.001445 (Ridged)	
	2	0	.
		0.001445 (Ridged)	

Final Results

Design	1
Choice Sets	24
Alternatives	11
D-Efficiency	0
D-Error	.

n	Variable Name	Label	Variance	DF	Standard Error
1	Brand1	Brand 1	4.51944	1	2.12590
2	Brand2	Brand 2	4.50242	1	2.12189
3	Brand3	Brand 3	3.50008	1	1.87085
4	Brand4	Brand 4	3.49251	1	1.86882
5	Brand5	Brand 5	3.48040	1	1.86558
6	Brand6	Brand 6	2.46342	1	1.56953
7	Brand7	Brand 7	2.46617	1	1.57041
8	Brand8	Brand 8	2.47129	1	1.57203
9	Brand9	Brand 9	2.47975	1	1.57472
10	Brand10	Brand 10	2.46659	1	1.57054
11	Brand1Price1	Brand 1 * Price 1	8.18241	1	2.86049
12	Brand1Price2	Brand 1 * Price 2	8.22704	1	2.86828
13	Brand1Price3	Brand 1 * Price 3	8.20656	1	2.86471
14	Brand1Price4	Brand 1 * Price 4	8.28067	1	2.87762
15	Brand1Price5	Brand 1 * Price 5	8.20668	1	2.86473
16	Brand1Price6	Brand 1 * Price 6	8.21663	1	2.86647
17	Brand1Price7	Brand 1 * Price 7	8.25795	1	2.87366

18	Brand2Price1	Brand	2 * Price 1	8.16766	1	2.85791
19	Brand2Price2	Brand	2 * Price 2	8.25283	1	2.87277
20	Brand2Price3	Brand	2 * Price 3	8.18178	1	2.86038
21	Brand2Price4	Brand	2 * Price 4	8.22588	1	2.86808
22	Brand2Price5	Brand	2 * Price 5	8.24887	1	2.87208
23	Brand2Price6	Brand	2 * Price 6	8.19937	1	2.86345
24	Brand2Price7	Brand	2 * Price 7	8.21050	1	2.86540
25	Brand3Price1	Brand	3 * Price 1	6.18856	1	2.48768
26	Brand3Price2	Brand	3 * Price 2	6.16883	1	2.48371
27	Brand3Price3	Brand	3 * Price 3	6.22013	1	2.49402
28	Brand3Price4	Brand	3 * Price 4	6.17914	1	2.48579
29	Brand3Price5	Brand	3 * Price 5	6.17185	1	2.48432
30	Brand3Price6	Brand	3 * Price 6	.	0	.
31	Brand3Price7	Brand	3 * Price 7	.	0	.
32	Brand4Price1	Brand	4 * Price 1	6.16116	1	2.48217
33	Brand4Price2	Brand	4 * Price 2	6.18716	1	2.48740
34	Brand4Price3	Brand	4 * Price 3	6.16633	1	2.48321
35	Brand4Price4	Brand	4 * Price 4	6.25094	1	2.50019
36	Brand4Price5	Brand	4 * Price 5	6.13517	1	2.47693
37	Brand4Price6	Brand	4 * Price 6	.	0	.
38	Brand4Price7	Brand	4 * Price 7	.	0	.
39	Brand5Price1	Brand	5 * Price 1	6.15820	1	2.48157
40	Brand5Price2	Brand	5 * Price 2	6.17572	1	2.48510
41	Brand5Price3	Brand	5 * Price 3	6.14151	1	2.47821
42	Brand5Price4	Brand	5 * Price 4	6.17153	1	2.48426
43	Brand5Price5	Brand	5 * Price 5	6.14552	1	2.47902
44	Brand5Price6	Brand	5 * Price 6	.	0	.
45	Brand5Price7	Brand	5 * Price 7	.	0	.
46	Brand6Price1	Brand	6 * Price 1	4.14170	1	2.03512
47	Brand6Price2	Brand	6 * Price 2	4.15481	1	2.03834
48	Brand6Price3	Brand	6 * Price 3	4.09000	1	2.02238
49	Brand6Price4	Brand	6 * Price 4	.	0	.
50	Brand6Price5	Brand	6 * Price 5	.	0	.
51	Brand6Price6	Brand	6 * Price 6	.	0	.
52	Brand6Price7	Brand	6 * Price 7	.	0	.
53	Brand7Price1	Brand	7 * Price 1	4.12837	1	2.03184
54	Brand7Price2	Brand	7 * Price 2	4.09248	1	2.02299
55	Brand7Price3	Brand	7 * Price 3	4.16503	1	2.04084
56	Brand7Price4	Brand	7 * Price 4	.	0	.
57	Brand7Price5	Brand	7 * Price 5	.	0	.
58	Brand7Price6	Brand	7 * Price 6	.	0	.
59	Brand7Price7	Brand	7 * Price 7	.	0	.
60	Brand8Price1	Brand	8 * Price 1	4.16889	1	2.04179
61	Brand8Price2	Brand	8 * Price 2	4.16045	1	2.03972
62	Brand8Price3	Brand	8 * Price 3	4.09355	1	2.02325
63	Brand8Price4	Brand	8 * Price 4	.	0	.
64	Brand8Price5	Brand	8 * Price 5	.	0	.
65	Brand8Price6	Brand	8 * Price 6	.	0	.
66	Brand8Price7	Brand	8 * Price 7	.	0	.

67	Brand9Price1	Brand	9 * Price 1	4.11932	1	2.02961
68	Brand9Price2	Brand	9 * Price 2	4.15745	1	2.03898
69	Brand9Price3	Brand	9 * Price 3	4.17574	1	2.04346
70	Brand9Price4	Brand	9 * Price 4	.	0	.
71	Brand9Price5	Brand	9 * Price 5	.	0	.
72	Brand9Price6	Brand	9 * Price 6	.	0	.
73	Brand9Price7	Brand	9 * Price 7	.	0	.
74	Brand10Price1	Brand	10 * Price 1	4.12770	1	2.03167
75	Brand10Price2	Brand	10 * Price 2	4.12731	1	2.03158
76	Brand10Price3	Brand	10 * Price 3	4.11729	1	2.02911
77	Brand10Price4	Brand	10 * Price 4	.	0	.
78	Brand10Price5	Brand	10 * Price 5	.	0	.
79	Brand10Price6	Brand	10 * Price 6	.	0	.
80	Brand10Price7	Brand	10 * Price 7	.	0	.
					==	
					54	

There are singularities in our model, and for the moment, that is fine. We see 10 parameters for `Brand`. The constant alternative (not shown) is the reference alternative. We see 7 parameters for Brand 1's price (8 prices - 1 = 7), 7 parameters for Brand 2's price, 5 parameters for Brand 3's price (6 prices - 1 = 5), ..., and 3 parameters for Brand 10's price (4 prices - 1 = 3). This all looks correct.

The log file contains the lines:

```
Redundant Variables:

Brand3Price6 Brand3Price7 Brand4Price6 Brand4Price7 Brand5Price6 Brand5Price7
Brand6Price4 Brand6Price5 Brand6Price6 Brand6Price7 Brand7Price4 Brand7Price5
Brand7Price6 Brand7Price7 Brand8Price4 Brand8Price5 Brand8Price6 Brand8Price7
Brand9Price4 Brand9Price5 Brand9Price6 Brand9Price7 Brand10Price4 Brand10Price5
Brand10Price6 Brand10Price7
```

Here they are again, manually reformatted to one brand per line:

```
Redundant Variables:

Brand3Price6 Brand3Price7
Brand4Price6 Brand4Price7
Brand5Price6 Brand5Price7
Brand6Price4 Brand6Price5 Brand6Price6 Brand6Price7
Brand7Price4 Brand7Price5 Brand7Price6 Brand7Price7
Brand8Price4 Brand8Price5 Brand8Price6 Brand8Price7
Brand9Price4 Brand9Price5 Brand9Price6 Brand9Price7
Brand10Price4 Brand10Price5 Brand10Price6 Brand10Price7
```

For Brands 1 and 2 we have 7 parameters and the last level for price 8 is the reference level and does not appear in the model. The specification `class(brand*price / zero=' ')` sets the reference level for brand to blank so it will use all 10 brands and uses the ordinary default last level for the reference level for price. This `zero=` specification names a list of reference levels, blank for the first variable and nothing specified, and hence the default, for the second.

For Brands 3 through 5, level 8, which does not appear, is the reference level as it is for all the brands. In addition, since these brands have only six levels, two more terms are not estimable, the terms for price levels 6 and 7. Hence the factors `Brand3Price6, Brand3Price7, Brand4Price6, Brand4Price7, Brand5Price6,` and `Brand5Price7` are not needed. Similarly for Brands 6 through 10, we can drop the terms for the fourth through seventh price levels. We can run the macro again, this time deleting all these terms.

```
%choiceff(data=cand, seed=462,
          model=class(brand / zero=none) class(brand*price / zero=' '),
          nsets=24, flags=f1-f11, beta=zero,
          drop=
          brand3price6 brand3price7
          brand4price6 brand4price7
          brand5price6 brand5price7
          brand6price4 brand6price5 brand6price6 brand6price7
          brand7price4 brand7price5 brand7price6 brand7price7
          brand8price4 brand8price5 brand8price6 brand8price7
          brand9price4 brand9price5 brand9price6 brand9price7
          brand10price4 brand10price5 brand10price6 brand10price7
          );
```

Here is some of the output.

Design	Iteration	D-Efficiency	D-Error
1	0	0.313841	3.186325
	1	0.340743	2.934765
	2	0.340743	2.934765

Design	Iteration	D-Efficiency	D-Error
2	0	0	.
	1	0.341011	2.932458
	2	0.341011	2.932458

Final Results

Design	2
Choice Sets	24
Alternatives	11
D-Efficiency	0.341011
D-Error	2.932458

n	Variable Name	Label	Variance	DF	Standard Error
1	Brand1	Brand 1	4.50417	1	2.12230
2	Brand2	Brand 2	4.52567	1	2.12736
3	Brand3	Brand 3	3.47776	1	1.86487
4	Brand4	Brand 4	3.48724	1	1.86742
5	Brand5	Brand 5	3.49982	1	1.87078
6	Brand6	Brand 6	2.47337	1	1.57270
7	Brand7	Brand 7	2.45738	1	1.56760
8	Brand8	Brand 8	2.45142	1	1.56570
9	Brand9	Brand 9	2.45282	1	1.56615
10	Brand10	Brand 10	2.44575	1	1.56389
11	Brand1Price1	Brand 1 * Price 1	8.19264	1	2.86228
12	Brand1Price2	Brand 1 * Price 2	8.19269	1	2.86229
13	Brand1Price3	Brand 1 * Price 3	8.18940	1	2.86171
14	Brand1Price4	Brand 1 * Price 4	8.23067	1	2.86891
15	Brand1Price5	Brand 1 * Price 5	8.21587	1	2.86633
16	Brand1Price6	Brand 1 * Price 6	8.19365	1	2.86246
17	Brand1Price7	Brand 1 * Price 7	8.23031	1	2.86885
18	Brand2Price1	Brand 2 * Price 1	8.16830	1	2.85802
19	Brand2Price2	Brand 2 * Price 2	8.23185	1	2.86912
20	Brand2Price3	Brand 2 * Price 3	8.21687	1	2.86651
21	Brand2Price4	Brand 2 * Price 4	8.23295	1	2.86931
22	Brand2Price5	Brand 2 * Price 5	8.27059	1	2.87586
23	Brand2Price6	Brand 2 * Price 6	8.22612	1	2.86812
24	Brand2Price7	Brand 2 * Price 7	8.28203	1	2.87785
25	Brand3Price1	Brand 3 * Price 1	6.15558	1	2.48104
26	Brand3Price2	Brand 3 * Price 2	6.17640	1	2.48524
27	Brand3Price3	Brand 3 * Price 3	6.13255	1	2.47640
28	Brand3Price4	Brand 3 * Price 4	6.14840	1	2.47960
29	Brand3Price5	Brand 3 * Price 5	6.11249	1	2.47234
30	Brand4Price1	Brand 4 * Price 1	6.17231	1	2.48441
31	Brand4Price2	Brand 4 * Price 2	6.22760	1	2.49552
32	Brand4Price3	Brand 4 * Price 3	6.12111	1	2.47409
33	Brand4Price4	Brand 4 * Price 4	6.19792	1	2.48956
34	Brand4Price5	Brand 4 * Price 5	6.12131	1	2.47413
35	Brand5Price1	Brand 5 * Price 1	6.21514	1	2.49302
36	Brand5Price2	Brand 5 * Price 2	6.15748	1	2.48143
37	Brand5Price3	Brand 5 * Price 3	6.17697	1	2.48535
38	Brand5Price4	Brand 5 * Price 4	6.16121	1	2.48218
39	Brand5Price5	Brand 5 * Price 5	6.20067	1	2.49011
40	Brand6Price1	Brand 6 * Price 1	4.16170	1	2.04002
41	Brand6Price2	Brand 6 * Price 2	4.11324	1	2.02811
42	Brand6Price3	Brand 6 * Price 3	4.13298	1	2.03297
43	Brand7Price1	Brand 7 * Price 1	4.10703	1	2.02658
44	Brand7Price2	Brand 7 * Price 2	4.11083	1	2.02752
45	Brand7Price3	Brand 7 * Price 3	4.10632	1	2.02641

46	Brand8Price1	Brand	8 * Price 1	4.12107	1	2.03004
47	Brand8Price2	Brand	8 * Price 2	4.10075	1	2.02503
48	Brand8Price3	Brand	8 * Price 3	4.08366	1	2.02081
49	Brand9Price1	Brand	9 * Price 1	4.11157	1	2.02770
50	Brand9Price2	Brand	9 * Price 2	4.10049	1	2.02497
51	Brand9Price3	Brand	9 * Price 3	4.10522	1	2.02614
52	Brand10Price1	Brand	10 * Price 1	4.07896	1	2.01964
53	Brand10Price2	Brand	10 * Price 2	4.09065	1	2.02253
54	Brand10Price3	Brand	10 * Price 3	4.11148	1	2.02768
					==	
					54	

We can see that we now have all the terms for the final model.

```
proc print data=best(obs=22); id set; by set; var brand price; run;
```

Here are the first two choice sets.

Set	Brand	Price
1	1	2
	2	8
	3	6
	4	6
	5	4
	6	4
	7	2
	8	4
	9	4
	10	3
	.	.
2	1	4
	2	8
	3	4
	4	3
	5	2
	6	4
	7	4
	8	3
	9	4
	10	3
	.	.

Because of the asymmetry, assigning the actual prices is not as simple as using a format. You could write a lot of code of the form `if brand = 1 and price = 1 then price = 0.89; else ...`, however that would be difficult. Instead, we will start by transposing our choice design.

```
proc transpose data=best out=lin(keep=x1-x10) prefix=x;
   var price; by set;
   run;

proc print; run;
```

The transposed data set has one row per choice set and each of the 10 prices for the ten nonconstant alternatives in each choice set. This is the linear version of our choice design. The factors `x1` and `x2` have 8 prices, 1 to 8, the factors `x3` through `x5` have 6 prices, 1 to 6. and the factors `x6` through `x10` have 4 prices, 1 to 4.

Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	2	8	6	6	4	4	2	4	4	3
2	4	8	4	3	2	4	4	3	4	3
3	8	4	3	5	5	4	1	4	4	4
4	2	1	4	3	4	1	1	3	3	2
5	6	8	6	6	5	2	3	3	3	2
6	6	1	5	5	2	4	3	4	4	3
7	8	6	5	6	4	3	3	3	2	4
8	4	5	2	3	3	3	3	1	3	4
9	5	3	5	1	6	2	4	1	3	4
10	3	4	2	2	4	2	4	2	1	1
11	3	5	5	1	3	3	2	4	2	2
12	2	3	1	5	1	4	2	2	1	3
13	6	2	3	3	6	2	3	1	2	3
14	5	1	1	4	2	3	1	2	3	2
15	8	6	4	4	5	2	2	2	1	2
16	5	2	6	5	6	3	4	3	2	1
17	1	7	1	4	5	1	1	3	1	1
18	3	3	2	6	6	4	1	1	3	3
19	4	5	1	2	1	1	3	4	2	1
20	1	6	3	1	1	3	4	2	4	1
21	7	7	3	2	1	1	2	2	1	4
22	7	2	2	4	3	1	1	1	4	1
23	7	4	4	2	3	1	2	1	2	2
24	1	7	6	1	2	2	4	4	1	4

Now we can use formats or other means such as the `%MktLab` macro to assign prices within the brand/price factors and then restore the choice design format. Here is the `key=` data set for the `%MktLab` macro. It contains all of the different prices. The `%MktLab` macro assigns the prices and the `%MktRoll` macro is used to convert our linear design back into a choice design.[†] The `%MktLab` macro uses a KEY data set to specify the prices (see page 577). The `%MktRoll` macro uses a different KEY data set to specify which linear design factor applies to which brand (see page 595).

[†]See page 87 for an illustration of linear versus choice designs.

```
data key;
   input x1-x10;
   datalines;
0.89 0.94 0.99 0.89 1.04 0.89 0.99 0.94 1.09 1.14
0.94 0.99 1.04 0.94 1.09 0.99 1.09 0.99 1.14 1.19
0.99 1.04 1.09 0.99 1.14 1.09 1.19 1.14 1.19 1.24
1.04 1.09 1.14 1.04 1.19 1.19 1.29 1.19 1.24 1.29
1.09 1.14 1.19 1.09 1.24 .    .    .    .    .
1.14 1.19 1.24 1.14 1.29 .    .    .    .    .
1.19 1.24 .    .    .    .    .    .    .    .
1.24 1.29 .    .    .    .    .    .    .    .
;

proc print; run;

%mktlab(data=lin, key=key, out=final)

%mktkey(x1-x10)

data key;
   if _n_ le 10 then Brand = put(_n_, 2.);
   input Price $ @@;
   datalines;
x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 .
;

proc print; run;

%mktroll(design=final,
         alt=brand, key=key, out=sasuser.prices)

proc print data=sasuser.prices(obs=22); id set; by set; var brand price; run;
```

Here is the %MktLab KEY data set.

Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	0.89	0.94	0.99	0.89	1.04	0.89	0.99	0.94	1.09	1.14
2	0.94	0.99	1.04	0.94	1.09	0.99	1.09	0.99	1.14	1.19
3	0.99	1.04	1.09	0.99	1.14	1.09	1.19	1.14	1.19	1.24
4	1.04	1.09	1.14	1.04	1.19	1.19	1.29	1.19	1.24	1.29
5	1.09	1.14	1.19	1.09	1.24	.	.	.	.	.
6	1.14	1.19	1.24	1.14	1.29	.	.	.	.	.
7	1.19	1.24	.	.	.	.	.	.	.	.
8	1.24	1.29	.	.	.	.	.	.	.	.

Here is the `%MktRoll` KEY data set.

Obs	Brand	Price
1	1	x1
2	2	x2
3	3	x3
4	4	x4
5	5	x5
6	6	x6
7	7	x7
8	8	x8
9	9	x9
10	10	x10
11		

Here are the first two choice sets with the actual prices assigned.

Set	Brand	Price
1	1	0.94
	2	1.29
	3	1.24
	4	1.14
	5	1.19
	6	1.19
	7	1.09
	8	1.19
	9	1.24
	10	1.24
		.
2	1	1.04
	2	1.29
	3	1.14
	4	0.99
	5	1.09
	6	1.19
	7	1.29
	8	1.14
	9	1.24
	10	1.24
		.

%ChoicEff Macro Options

The following options can be used with the `%ChoicEff` macro.

Option	Description
`bestcov=`*SAS-data-set*	covariance matrix for the best design
`bestout=`*SAS-data-set*	best design
`beta=`*list*	true parameters
`converge=`*n*	convergence criterion
`cov=`*SAS-data-set*	all of the covariance matrices
`data=`*SAS-data-set*	input choice candidate set
`drop=`*variable-list*	variables to drop from the model
`fixed=`*variable-list*	variable that flags fixed alternatives
`flags=`*variable-list*	variables that flag the alternative(s)
`init=`*SAS-data-set*	input initial design data set
`initvars=`*variable-list*	initial variables
`intiter=`*n*	maximum number of internal iterations
`iter=`*n*	maximum iterations (designs to create)
`maxiter=`*n*	maximum iterations (designs to create)
`model=`*model-specification*	`model` statement list of effects
`morevars=`*variable-list*	more variables to add to the model
`n=`*n*	number of observations
`nalts=`*n*	number of alternatives
`nsets=`*n*	number of choice sets desired
`options=`*options-list*	binary options
`out=`*SAS-data-set*	all designs data set
`seed=`*n*	random number seed
`submat=`*number-list*	submatrix for efficiency calculations
`types=`*integer-list*	number of sets of each type
`typevar=`*variable*	choice set types variable
`weight=`*weight-variable*	optional weight variable

Required Options

You must specify both the `model=` and `nsets=` options and either the `flags=` or `nalts=` options. You can omit `beta=` if you just want a listing of effects, however you must specify `beta=` to create a design. The rest of the options are optional.

model= *model-specification*

specifies a PROC TRANSREG `model` statement list of effects. There are many potential forms for the model specification and a number of options. See the SAS/STAT PROC TRANSREG documentation.

Generic effects example:

```
model=class(x1-x3),
```

Brand and alternative-specific effects example:

```
model=class(b)
      class(b*x1 b*x2 b*x3 / effects zero=' '),
```

Brand, alternative-specific, and cross effects:

```
model=class(b b*p / zero=' ')
      identity(x1-x5) * class(b / zero=none),
```

See pages 481 through 505 for other examples of `model` syntax. Furthermore, all of the PROC TRANSREG and `%ChoicEff` macro examples from pages 122 through 343 show examples of model syntax for choice models.

nsets= *n*

specifies the number of choice sets desired.

Other Required Options

You must specify exactly one of these next two options. When the candidate set consists of individual alternatives to be swapped, specify the alternative flags with `flags=`. When the candidate set consists of entire sets of alternatives to be swapped, specify the number of alternatives in each set with `nalts=`.

flags= *variable-list*

specifies variables that flag the alternative(s) for which each candidate may be used. There must be one flag variable per alternative. If every candidate can be used in all alternatives, then the flags are constant. For example, with three alternatives, create these constant flags: `f1=1 f2=1 f3=1`. Otherwise, with three alternatives, specify `flags=f1-f3` and create a candidate set where: alternative 1 candidates are indicated by `f1=1 f2=0 f3=0`, alternative 2 candidates are indicated by `f1=0 f2=1 f3=0`, and alternative 3 candidates are indicated by `f1=0 f2=0 f3=1`.

nalts= *n*

specifies the number of alternatives in each choice set for the set-swapping algorithm.

Other Options

The rest of the parameters are optional. You may specify zero or more of them.

bestcov= *SAS-data-set*

specifies a name for the data set containing the covariance matrix for the best design. By default, this data set is called BESTCOV.

bestout= *SAS-data-set*

specifies a name for the data set containing the best design. By default, this data set is called BEST. Often, you will want to specify a two-level name to create a permanent SAS data set so the design will be available later for analysis.

beta= *list*

specifies the true parameters. By default, when `beta=` is not specified, the macro just reports on coding. You can specify `beta=zero` to assume all zeros. Otherwise specify a number list: `beta=1 -1 2 -2 1 -1`.

converge= *n*

specifies the *D*-efficiency convergence criterion. By default, `converge=0.005`.

cov= *SAS-data-set*

specifies a name for the data set containing all of the covariance matrices for all of the designs. By default, this data set is called COV.

data= *SAS-data-set*

specifies the input choice candidate set. By default, the macro uses the last data set created.

drop= *variable-list*

specifies a list of variables to drop from the model. If you specified a less-than-full-rank `model=` specification, you can use `drop=` to produce a full rank coding. When there are redundant variables, the macro prints a list that you can use in the `drop=` option on a subsequent run.

fixed= *variable-list*

specifies the variable that flags the fixed alternatives. When `fixed=variable` is specified, the `init=` data set must contain the named variable, which indicates which alternatives are fixed (cannot be swapped out) and which ones may be changed. Example: `fixed=fixed, init=init, initvars=x1-x3`. Values of the `fixed=` variable include:

1 - means this alternative can never be swapped out.

0 - means this alternative is used in the initial design, but it may be swapped out.

. - means this alternative should be randomly initialized, and it may be swapped out.

The `fixed=` option may be specified only when both `init=` and `initvars=` are specified.

init= *SAS-data-set*

specifies an input initial design data set. Null means a random start. One usage is to specify the `bestout=` data set for an initial start. When `flags=` is specified, `init=` must contain the index variable. Example: `init=best(keep=index)`. When `nalts=` is specified, `init=` must contain the choice set variable. Example: `init=best(keep=set)`.

Alternatively, the `init=` data set can contain an arbitrary design, potentially created outside this macro. In that case, you must also specify `initvars=factors`, where factors are the factors in the design, for example `initvars=x1-x3`. When alternatives are swapped, this data set must also contain the `flags=` variables. When `init=` is specified with `initvars=`, the data set may also contain a variable specified on the `fixed=` option, which indicates which alternatives are fixed, and which ones can be swapped in and out.

intiter= *n*
specifies the maximum number of internal iterations. Specify `intiter=0` to just evaluate efficiency of an existing design. By default, `intiter=10`.

initvars= *variable-list*
specifies the factor variables in the `init=` data set that must match up with the variables in the `data=` data set. See `init=`. All of these variables must be of the same type.

maxiter= *n*

iter= *n*
specifies the maximum iterations (designs to create). By default, `maxiter=10`.

morevars= *variable-list*
specifies more variables to add to the model. This option gives you the ability to specify a list of variables to copy along as is, through the TRANSREG coding, then add them to the model.

n= *n*
specifies the number of observations to use in the variance matrix formula. By default, `n=1`.

options= *options-list*
specifies binary options. By default, none of these options are specified. Specify one or more of the following values after `options=`.

`coded`
prints the coded candidate set.

`detail`
prints the details of the swaps.

`nocode`
skips the PROC TRANSREG coding stage, assuming that WORK.TMP_CAND was created by a previous step. This is most useful with set swapping when the candidate set can be big. It is important with `options=nocode` to note that the effect of `morevars=` and `drop=` in previous runs has already been taken care of, so do not specify them (unless for instance you want to drop still more variables).

`nodups`
prevents the same choice set from coming out more than once. This option does not affect the initialization, so the random initial design may have duplicates. This option forces duplicates out during the iterations, so do not set `intiter=` to a small value. It may take several iterations to eliminate all duplicates. It is possible that efficiency will decrease as duplicates are forced out. With set swapping, this macro checks the candidate choice set numbers to avoid duplicates. With alternative swapping, this macro checks the candidate alternative index to avoid duplicates. The macro does not look at the actual factors. This makes the checks faster, but if the candidate set contains duplicate choice sets or alternatives, the macro may not succeed in eliminating all duplicates. Run the `%MktDups` macro (which looks at the actual factors) on the design to check and make sure all duplicates are eliminated. If you are using set swapping to make a generic design make sure you run the `%MktDups` macro on the candidate set to eliminate duplicate choice sets in advance.

`notests`
suppresses printing the diagonal of the covariance matrix, and hypothesis tests for this n and $\boldsymbol{\beta}$. When $\boldsymbol{\beta}$ is not zero, the results include a Wald test statistic (β divided by the standard error), which is normally distributed, and the probability of a larger squared Wald statistic.

`orthcan`
orthogonalizes the candidate set.

out= *SAS-data-set*
specifies a name for the output SAS data set with all of the final designs. The default is `out=results`.

seed= *n*
specifies the random number seed. By default, `seed=0`, and clock time is used as the random number seed. By specifying a random number seed, results should be reproducible within a SAS release for a particular operating system. However, due to machine differences, some results may not be exactly reproducible on other machines, although you would expect the efficiency differences to be slight.

submat= *number-list*
specifies a submatrix for which efficiency calculations are desired. Specify an index vector. For example, with 3 three-level factors, `a`, `b`, and `c`, and the model `class(a b c a*b)`, specify `submat=1:6`, to see the efficiency of just the 6×6 matrix of main effects. Specify `submat=3:6`, to see the efficiency of just the 4×4 matrix of `b` and `c` main effects.

types= *integer-list*
specifies the number of sets of each type to put into the design. This option is used when you have multiple types of choice sets and you want the design to consist of only certain numbers of each type. This option can be specified with the set-swapping algorithm. The argument is an integer list. When you specify `types=`, you must also specify `typevar=`. Say you are creating a design with 30 choice sets, and you want the first 10 sets to consist of sets whose `typevar=` variable in the candidate set is type 1, and you want the rest to be type 2. You would specify `types=10 20`.

typevar= *variable*
specifies a variable in the candidate data set that contains choice set types. The types must be integers starting with 1. This option can only be specified with the set-swapping algorithm. When you specify `typevar=`, you must also specify `types=`.

weight= *weight-variable*
specifies an optional weight variable. Typical usage is with an availability design. Give unavailable alternatives a weight of zero and available alternatives a weight of one. The number of alternatives must always be constant, so varying numbers of alternatives are handled by giving unavailable or unseen alternatives a weight of zero.

%ChoicEff Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktAllo Macro

The %MktAllo autocall macro is used for manipulating data for an allocation choice experiment. See page 284 for an example. The %MktAllo macro takes as input a data set with one row for each alternative of each choice set. For example, in a study with 10 brands plus a constant alternative and 27 choice sets, there are $27 \times 11 = 297$ observations in the input data set. Here is an example of an input data set. It contains a choice set variable, product attributes (Brand and Price) and a frequency variable (Count) that contains the total number of times that each alternative was chosen.

Obs	Set	Brand	Price	Count
1	1			0
2	1	Brand 1	$50	103
3	1	Brand 2	$75	58
4	1	Brand 3	$50	318
5	1	Brand 4	$100	99
6	1	Brand 5	$100	54
7	1	Brand 6	$100	83
8	1	Brand 7	$75	71
9	1	Brand 8	$75	58
10	1	Brand 9	$75	100
11	1	Brand 10	$50	56
.				
.				
.				
296	27	Brand 9	$100	94
297	27	Brand 10	$50	65

The end result is a data set with twice as many observations that contains the number of times each alternative was chosen and the number of times it was not chosen. This data set also contains a variable c with values 1 for first choice and 2 for second or subsequent choice.

Obs	Set	Brand	Price	Count	c
1	1			0	1
2	1			1000	2
3	1	Brand 1	$50	103	1
4	1	Brand 1	$50	897	2
5	1	Brand 2	$75	58	1
6	1	Brand 2	$75	942	2
7	1	Brand 3	$50	318	1

```
  8     1     Brand  3     $50      682    2
  .
  .
  .
593    27     Brand 10     $50       65    1
594    27     Brand 10     $50      935    2
```

Here is an example of usage:

```
%mktallo(data=allocs2, out=allocs3, nalts=11,
         vars=set brand price, freq=Count)
```

The option `data=` names the input data set, `out=` names the output data set, `nalts=` specifies the number of alternatives, `vars=` names the variables in the data set that will be used in the analysis excluding the `freq=` variable, and `freq=` names the frequency variable.

%MktAllo Macro Options

The following options can be used with the `%MktAllo` macro.

Option	Description
`data=`*SAS-data-set*	input SAS data set
`freq=`*variable*	frequency variable
`nalts=`*n*	number of alternatives
`out=`*SAS-data-set*	output SAS data set
`vars=`*variable-list*	input variables

You must specify the `nalts=`, `freq=`, and `vars=` options.

data= *SAS-data-set*
specifies the input SAS data set. By default, the macro uses the last data set created.

freq= *variable*
specifies the frequency variable, which contains the number of times this alternative was chosen. This option must be specified.

nalts= *n*
specifies the number of alternatives (including if appropriate the constant alternative). This option must be specified.

out= *SAS-data-set*
specifies the output SAS data set. The default is `out=allocs`.

vars= *variable-list*
specifies the variables in the data set that will be used in the analysis but not the `freq=` variable. This option must be specified.

%MktAllo Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktBal Macro

The %MktBal macro creates factorial designs using an algorithm that ensures that the design is perfectly balanced, or when the number of levels of a factor does not divide the number of runs, as close to perfectly balanced as possible. Do not use the %MktBal macro until you have tried the %MktEx macro and determined that it does not make a design that is balanced enough for your needs. The %MktEx macro can directly create thousands of orthogonal and balanced designs that the %MktBal algorithm will never be able to find. Even when the %MktEx macro cannot create an orthogonal and balanced design, it will usually find a nearly balanced design. Designs created with the %MktBal macro, while perfectly balanced, may be less efficient than designs found with the %MktEx macro, and for large problems, the %MktBal macro can be slow. It is likely that the current algorithm used by the %MktBal macro will be changed in the future to use some now unknown algorithm that is both faster and better.

The %MktBal macro is *not* a full-featured experimental design generator. For example, you cannot specify interactions that you want to estimate or specify restrictions such as which levels may or may not appear together. You must use the %MktEx macro for that. The %MktBal macro builds a design by creating a balanced first factor, optimally blocking it to create the second factor, then optimally blocking the first two factors to create the third, and so on. Once it creates all factors, it refines each factor. Each factor is in turn removed from the design, and the rest of the design is reblocked, replacing the initial factor if the new design is more D-efficient.

Here is a simple example of creating a design with 2 two-level factors and 3 three-level factors in 18 runs. The %MktEval macro evaluates the results. This design is in fact optimal.

```
%mktbal(2 2 3 3 3, n=18, seed=151)
%mkteval;
```

In all cases, the factors are named x1, x2, x3, and so on.

This next example, at 120 runs and with factor levels greater than 5, is starting to get big, and by default, it will run slowly. You can use the maxstarts=, maxtries=, and maxiter= options to make the macro run more quickly. For example, the second example shown next runs much faster than the first.

```
%mktbal(2 3 4 5 6 7 8 9 10, n=120, options=progress, seed=17)

%mktbal(2 3 4 5 6 7 8 9 10, n=120, options=progress, seed=17,
        maxstarts=1, maxiter=1, maxtries=1)
```

%MktBal Macro Options

The following options can be used with the `%MktBal` macro.

Option	Description
`iter=`*n*	maximum iterations (designs to create)
`list`	list of the numbers of levels
`maxiter=`*n*	maximum iterations (designs to create)
`maxstarts=`*n*	maximum number of random starts
`maxtries=`*n*	times to try refining each factor
`n=`*n*	number of runs in the design
`options=`*options-list*	binary options
`out=`*SAS-data set*	output experimental design
`seed=`*n*	random number seed

list

specifies a list of the numbers of levels of all the factors. For example, for 3 two-level factors specify either `2 2 2` or `2 ** 3`. Lists of numbers, like `2 2 3 3 4 4` or a *levels**number of factors* syntax like: `2**2 3**2 4**2` can be used, or both can be combined: `2 2 3**4 5 6`. The specification `3**4` means 4 three-level factors. You must specify a list. Note that the factor list is a positional parameter. This means it must come first, and unlike all other parameters, it is not specified after a name and an equal sign.

n= *n*

specifies the number of runs in the design. You must specify `n=`. You can use the `%MktRuns` macro to get suggestions for values of `n=`.

out= *SAS-data set*

specifies the output experimental design. The default is `out=design`.

These next options control some of the details of the `%MktBal` macro.

maxiter= *n*

iter= *n*

specifies the maximum iterations (designs to create). By default, `maxiter=5`.

maxstarts= *n*

specifies the maximum number of random starts for each factor. With larger values, the macro tends to find slightly better designs at a cost of slower run times. The default is `maxstarts=10`.

maxtries= *n*

specifies the maximum number of times to try refining each factor after the initialization stage. The default is `maxtries=10`.

options= *options-list*

specifies binary options. By default, none of these options are specified. Specify one or more of the following values after `options=`.

`noprint`
specifies that the final *D*-efficiency should not be printed.

`progress`
reports on the macro's progress. For large numbers of factors, a large number or runs, or when the number of levels is large, this macro is slow. The `options=progress` specification gives you information about which step is being executed.

seed= *n*

specifies the random number seed. By default, `seed=0`, and clock time is used to make the random number seed. By specifying a random number seed, results should be reproducible within a SAS release for a particular operating system. However, due to machine differences, some results may not be exactly reproducible on other machines, although you would expect the efficiency differences to be slight.

%MktBal Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktBlock Macro

The %MktBlock autocall macro is used to block a choice design or an ordinary linear experimental design. See pages 191 and 248 for examples. When a choice design is too large to show all choice sets to each subject, the design is blocked and a block of choice sets is shown to each subject. For example, if there are 36 choice sets, instead of showing each subject 36 sets, you could instead create 2 blocks and show 2 groups of subjects 18 sets each. You could also create 3 blocks of 12 choice sets or 4 blocks of 9 choice sets. You can also request just one block if you want to see the correlations and frequencies among all of the attributes of all of the alternatives of a choice design.

The design can be in one of two formats. Typically, a choice design has one row for each alternative of each choice set and one column for each of the attributes. Typically, this kind of design is produced by either the %ChoiceEff or %MktRoll macro. Alternatively, a "linear" design is an intermediate step in preparing some choice designs.[‡] The linear design has one row for each choice set and one column for each attribute of each alternative. Typically, the linear design is produced by the %MktEx macro. The output from the %MktBlock macro is a data set containing the design, with the blocking variable added and hence not in the original order, with runs or choice sets nested within blocks.

The macro tries to create a blocking factor that is uncorrelated with every attribute of every alternative. In other words, the macro is trying to optimally add one additional factor, a blocking factor, to the linear design. It is trying to make a factor that is orthogonal to all of the attributes of all of the alternatives. For linear designs, you can usually ask for a blocking factor directly as just another factor in the design, and then use the %MktLab macro to provide a name like Block, or you can use the %MktBlock macro.

Here is an example of creating the blocking variable directly.

```
%mktex(3 ** 7, n=27, seed=350)

%mktlab(vars=x1-x6 Block)
```

Here is an example of creating a design and then blocking it.

```
%mktex(3 ** 6, n=27, seed=350)

%mktblock(data=randomized, nblocks=3, seed=377, maxiter=50)
```

The output shows that the blocking factor is uncorrelated with all of the factors in the design. This output comes from the %MktEval macro, which is called by the %MktBlock macro.

[‡] See page 87 for an illustration of linear versus choice designs.

```
                Canonical Correlations Between the Factors
            There are 0 Canonical Correlations Greater Than 0.316

                 Block    x1       x2       x3       x4       x5       x6

        Block    1        0        0        0        0        0        0
        x1       0        1        0        0        0        0        0
        x2       0        0        1        0        0        0        0
        x3       0        0        0        1        0        0        0
        x4       0        0        0        0        1        0        0
        x5       0        0        0        0        0        1        0
        x6       0        0        0        0        0        0        1

                          Summary of Frequencies
            There are 0 Canonical Correlations Greater Than 0.316

                             Frequencies

                 Block       9 9 9
                 x1          9 9 9
                 x2          9 9 9
                 x3          9 9 9
                 x4          9 9 9
                 x5          9 9 9
                 x6          9 9 9
                 Block x1    3 3 3 3 3 3 3 3 3
                 Block x2    3 3 3 3 3 3 3 3 3
                 Block x3    3 3 3 3 3 3 3 3 3
                 Block x4    3 3 3 3 3 3 3 3 3
                 Block x5    3 3 3 3 3 3 3 3 3
                 Block x6    3 3 3 3 3 3 3 3 3
                 x1 x2       3 3 3 3 3 3 3 3 3
                 x1 x3       3 3 3 3 3 3 3 3 3
                 x1 x4       3 3 3 3 3 3 3 3 3
                 x1 x5       3 3 3 3 3 3 3 3 3
                 x1 x6       3 3 3 3 3 3 3 3 3
                 x2 x3       3 3 3 3 3 3 3 3 3
                 x2 x4       3 3 3 3 3 3 3 3 3
                 x2 x5       3 3 3 3 3 3 3 3 3
                 x2 x6       3 3 3 3 3 3 3 3 3
                 x3 x4       3 3 3 3 3 3 3 3 3
                 x3 x5       3 3 3 3 3 3 3 3 3
                 x3 x6       3 3 3 3 3 3 3 3 3
                 x4 x5       3 3 3 3 3 3 3 3 3
                 x4 x6       3 3 3 3 3 3 3 3 3
                 x5 x6       3 3 3 3 3 3 3 3 3
                 N-Way       1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                             1 1 1 1 1 1 1 1
```

Canonical Correlations Between the Factors by Block

Block		x1	x2	x3	x4	x5	x6
1	x1	1	0.58	0	0.58	0.58	0
	x2	0.58	1	0	0	0.58	0.58
	x3	0	0	1	0	0	0
	x4	0.58	0	0	1	0.58	0.58
	x5	0.58	0.58	0	0.58	1	0.58
	x6	0	0.58	0	0.58	0.58	1
2	x1	1	0.58	0	0.58	0.58	0
	x2	0.58	1	0	0	0.58	0.58
	x3	0	0	1	0	0	0
	x4	0.58	0	0	1	0.58	0.58
	x5	0.58	0.58	0	0.58	1	0.58
	x6	0	0.58	0	0.58	0.58	1
3	x1	1	0.58	0	0.58	0.58	0
	x2	0.58	1	0	0	0.58	0.58
	x3	0	0	1	0	0	0
	x4	0.58	0	0	1	0.58	0.58
	x5	0.58	0.58	0	0.58	1	0.58
	x6	0	0.58	0	0.58	0.58	1

Notice that even with a perfect blocking variable like we have in this example, canonical correlations within each block will not be all zero.

Here is the blocked linear design (3 blocks of nine choice sets). Note that in the linear version of the design, there is one row for each choice set and all of the attributes of all of the alternatives are in the same row.

Block	Run	x1	x2	x3	x4	x5	x6
1	1	1	1	3	1	3	1
	2	1	1	1	3	1	3
	3	1	2	2	1	3	2
	4	3	3	2	3	2	1
	5	2	1	2	2	1	3
	6	3	2	1	2	3	2
	7	2	3	1	1	2	1
	8	2	3	3	2	1	2
	9	3	2	3	3	2	3

Block	Run	x1	x2	x3	x4	x5	x6
2	1	2	1	3	1	2	2
	2	3	1	2	2	3	1
	3	2	2	2	1	2	3
	4	2	1	1	3	3	1
	5	3	3	3	2	3	3
	6	3	3	1	1	1	2
	7	1	3	2	3	1	2
	8	1	2	3	3	1	1
	9	1	2	1	2	2	3

Block	Run	x1	x2	x3	x4	x5	x6
3	1	3	1	3	1	1	3
	2	2	3	2	3	3	3
	3	3	2	2	1	1	1
	4	1	1	2	2	2	2
	5	2	2	1	2	1	1
	6	2	2	3	3	3	2
	7	3	1	1	3	2	2
	8	1	3	1	1	3	3
	9	1	3	3	2	2	1

Next, we will create and block a choice design with two blocks of nine sets instead of blocking the linear version of a choice design.

```
%mktex(3 ** 6, n=3**6)

* Create an efficient choice design;
data key;
   input (x1-x3) ($);
   datalines;
x1 x2 x3
x4 x5 x6
;

%mktroll(design=design, key=key, out=out)

%choiceff(data=out, model=class(x1-x3), nsets=18, nalts=2,
          beta=zero, options=nodups, seed=151)

* Block the choice design.  Ask for 2 blocks;
%mktblock(data=best, nalts=2, nblocks=2, factors=x1-x3, seed=472)
```

(Note that if this had been a branded example, and if you were running SAS version 8.2 or an earlier release, specify `id=brand`; do not add your brand variable to the factor list. For SAS 9.0 and later SAS releases, it is fine to add your brand variable to the factor list.)

Both the design and the blocking are not as good this time. The variable names in the output are composed of `Alt`, the alternative number, and the factor name. Since there are two alternatives each composed of three factors plus one blocking variable ($2 \times 3 + 1 = 7$), a 7×7 correlation matrix is reported. Here is some of the output.

Canonical Correlations Between the Factors
There are 11 Canonical Correlations Greater Than 0.316

	Block	Alt1_x1	Alt1_x2	Alt1_x3	Alt2_x1	Alt2_x2	Alt2_x3
Block	1	0.13	0	0	0.15	0.13	0
Alt1_x1	0.13	1	0.36	0.33	0.63	0.29	0.26
Alt1_x2	0	0.36	1	0.47	0.34	0.59	0.47
Alt1_x3	0	0.33	0.47	1	0.37	0.30	0.60
Alt2_x1	0.15	0.63	0.34	0.37	1	0.23	0.36
Alt2_x2	0.13	0.29	0.59	0.30	0.23	1	0.35
Alt2_x3	0	0.26	0.47	0.60	0.36	0.35	1

Summary of Frequencies
There are 11 Canonical Correlations Greater Than 0.316
* - Indicates Unequal Frequencies

		Frequencies
	Block	9 9
*	Alt1_x1	7 7 4
*	Alt1_x2	8 2 8
*	Alt1_x3	8 4 6
*	Alt2_x1	5 5 8
*	Alt2_x2	5 9 4
*	Alt2_x3	4 8 6
*	Block Alt1_x1	4 3 2 3 4 2
*	Block Alt1_x2	4 1 4 4 1 4
*	Block Alt1_x3	4 2 3 4 2 3
*	Block Alt2_x1	2 3 4 3 2 4
*	Block Alt2_x2	3 4 2 2 5 2
*	Block Alt2_x3	2 4 3 2 4 3
*	Alt1_x1 Alt1_x2	3 1 3 4 1 2 1 0 3
*	Alt1_x1 Alt1_x3	4 2 1 3 1 3 1 1 2
*	Alt1_x1 Alt2_x1	0 4 3 2 0 5 3 1 0
*	Alt1_x1 Alt2_x2	3 3 1 1 4 2 1 2 1
*	Alt1_x1 Alt2_x3	1 4 2 2 2 3 1 2 1
*	Alt1_x2 Alt1_x3	4 1 3 2 0 0 2 3 3
*	Alt1_x2 Alt2_x1	1 3 4 1 0 1 3 2 3

```
*    Alt1_x2 Alt2_x2    0 5 3 1 0 1 4 4 0
*    Alt1_x2 Alt2_x3    2 2 4 0 2 0 2 4 2
*    Alt1_x3 Alt2_x1    3 3 2 1 1 2 1 1 4
*    Alt1_x3 Alt2_x2    2 5 1 2 1 1 1 3 2
*    Alt1_x3 Alt2_x3    0 5 3 1 0 3 3 3 0
*    Alt2_x1 Alt2_x2    1 3 1 1 3 1 3 3 2
*    Alt2_x1 Alt2_x3    0 3 2 2 2 1 2 3 3
*    Alt2_x2 Alt2_x3    0 3 2 3 3 3 1 2 1
     N-Way              1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Note that in this example, the input is a choice design (as opposed to the linear version of a choice design) so the results are in choice design format. There is one row for each alternative of each choice set.

Block	Set	Alt	x1	x2	x3
1	1	1	3	1	3
		2	2	3	1
1	2	1	2	1	1
		2	1	2	3
.					
.					
.					
Block	Set	Alt	x1	x2	x3
2	1	1	2	1	1
		2	3	2	3
2	2	1	2	2	1
		2	1	3	2
.					
.					
.					

%MktBlock Macro Options

The following options can be used with the `%MktBlock` macro.

Option	Description
`alt=`*variable*	alternative number variable
`block=`*variable*	block number variable
`data=`*SAS-data-set*	either the choice or linear design
`factors=`*variable-list*	factors in the design
`id=`*variable-list*	variables to copy to output data set
`initblock=`*variable*	initial blocking variable
`iter=`*n*	times to try to block the design
`maxiter=`*n*	times to try to block the design
`nalts=`*n*	number of alternatives in choice set
`nblocks=`*n*	number of blocks to create
`next=`*n*	where to look for the next exchange
`out=`*SAS-data-set*	output data set with the block numbers
`outr=`*SAS-data-set*	randomized output data set
`print=`*print-options*	printing options
`ridge=`*n*	ridging factor
`seed=`*n*	random number seed
`set=`*variable*	choice set number variable
`vars=`*variable-list*	factors in the design

alt= *variable*
specifies the alternative number variable. If this variable is in the input data set, it is excluded from the factor list. The default is `alt=Alt`.

block= *variable*
specifies the block number variable. If this variable is in the input data set, it is excluded from the factor list. The default is `block=Block`.

data= *SAS-data-set*
specifies either the choice or linear design. The choice design has one row for each alternative of each choice set and one column for each of the attributes. Typically this design is produced by either the `%ChoicEff` or `%MktRoll` macro. For choice designs, you must also specify the `nalts=` option. By default, the macro uses the last data set created. The linear design has one row for each choice set and one column for each attribute of each alternative. Typically this design is produced by the `%MktEx` macro. This is the design that is input into the `%MktRoll` macro.

factors= *variable-list*

vars= *variable-list*
specifies the factors in the design. By default, all numeric variables are used, except variables with names matching those in the `block=`, `set=`, and `alt=` options. (By default, the variables `Block`, `Set`, `Run`, and `Alt` are excluded from the factor list.) If you are using version 8.2 or an earlier SAS release

with a branded choice design (assuming the brand factor is called `Brand`), specify `id=Brand`. Do not add the brand factor to the factor list unless you are using SAS 9.0 or a later SAS release.

id= *variable-list*
specifies the `data=` data set variables to copy to the output data set. If you are using version 8.2 or an earlier SAS release with a branded choice design (assuming the brand factor is called `Brand`), specify `id=Brand`. Do not add the brand factor to the factor list unless you are using SAS 9.0 or a later SAS release.

initblock= *variable*
specifies the name of the variable in the data set that is to be used as the initial blocking variable for the first iteration.

maxiter= *n*

iter= *n*
specifies the number of times to try to block the design starting with a different random blocking. By default, the macro tries five random starts, and iteratively refines each until *D*-efficiency quits improving, then in the end selects the blocking with the best *D*-efficiency.

nalts= *n*
specifies the number of alternatives in each choice set. If you are inputting a choice design, you must specify `nalts=`, otherwise the macro assumes you are inputting a linear design.

nblocks= *n*
specifies the number of blocks to create. The option `nblocks=1` just reports information about the design. The `nblocks=` option must be specified.

next= *n*
specifies how far into the design to go to look for the next exchange. The specification `next=1` specifies that the macro should try exchanging the level for each run with the level for the next run and all other runs. The specification `next=2` considers exchanges with half of the other runs, which makes the algorithm run more quickly. The macro considers exchanging the level for run i with run $i+1$ then uses the `next=` value to find the next potential exchanges. Other values, including nonintegers can be specified as well. For example `next=1.5` considers exchanging observation 1 with observations 2, 4, 5, 7, 8, 10, 11, and so on. With smaller values, the macro tends to find a slightly better blocking variable at a cost of much slower run time.

out= *SAS-data-set*
specifies the output data set with the block numbers. The default is `out=blocked`. Often, you will want to specify a two-level name to create a permanent SAS data set so the design will be available later for analysis.

outr= *SAS-data-set*
specifies the randomized output data set if you would like the design randomly sorted within blocks. Often, you will want to specify a two-level name to create a permanent SAS data set so the design will be available later for analysis.

print= *print-options*
specifies both the `%MktBlock` and the `%MktEval` macro printing options, which control the printing of the results. The default is `print=normal`. Values include:

`corr`	canonical correlations
`list`	list of big canonical correlations
`freqs`	long frequencies list
`summ`	frequency summaries
`block`	canonical correlations within blocks
`design`	blocked design
`note`	blocking note
`all`	all of the above
`noprint`	no printed output
`normal`	`corr list summ block design note`
`short`	`corr summ note`

ridge= *n*
specifies the value to add to the diagonal of $(\mathbf{X}'\mathbf{X})^{-1}$ to make it nonsingular. Usually, you will not need to change this value. If you do, you probably will not notice any effect. Specify `ridge=0` to use a generalized inverse instead of ridging. The default is `ridge=0.01`.

seed= *n*
specifies the random number seed. By default, `seed=0`, and clock time is used to make the random number seed. By specifying a random number seed, results should be reproducible within a SAS release for a particular operating system. However, due to machine differences, some results may not be exactly reproducible on other machines, although you would expect the efficiency differences to be slight.

set= *variable*
specifies the choice set number variable. When `nalts=` is specified, the default is `Set`, otherwise the default is `Run`. If this variable is in the input data set, it is excluded from the factor list.

%MktBlock Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktDes Macro

The %MktDes autocall macro creates efficient experimental designs. Usually, we will not need to call the %MktDes macro directly. Instead, we will usually use the %MktEx autocall macro, which calls the %MktDes macro as one of its many tools. At the heart of the %MktDes macro are PROC PLAN, PROC FACTEX, and PROC OPTEX. We use a macro instead of calling these procedures directly because the macro has a simpler syntax.

The %MktDes macro creates efficient experimental designs. You specify the names of the factors and the number of levels for each factor. You also specify the number of runs you want in your final design. Here for example is how you can create a design in 18 runs with 2 two-level factors (x1 and x2) and 3 three-level factors (x3, x4, and x5).

```
%mktdes(factors=x1-x2=2 x3-x5=3, n=18)
```

You can also optionally specify interactions that you want to be estimable. The macro creates a candidate design in which every effect you want to be estimable is estimable, but the candidate design is bigger than you want. By default, the candidate set is stored in a SAS data set called CAND1. The macro then uses PROC OPTEX to search the candidate design for an efficient final design. By default, the final experimental design is stored in a SAS data set called DESIGN.

When the full-factorial design is small (by default less than 2189 runs, although sizes up to 5000 or 6000 runs are reasonably small), the experimental design problem is straightforward. First, the macro uses PROC PLAN to create a full-factorial candidate set. Next, PROC OPTEX searches the full-factorial candidate set. For very small problems (a few hundred candidates) PROC OPTEX will often find the optimal design, and for larger problems, it may not find *the* optimal design, but given sufficient iteration (for example, specify iter=100 or more) it will find very good designs. Run time will typically be a few seconds or a few minutes, but it could be longer. Here is a typical example of using the %MktDes macro to find an optimal nonorthogonal design when the full-factorial design is small (108 runs):

```
*---2 two-level factors and 3 three-level factors in 18 runs---;
%mktdes(factors=x1-x2=2 x3-x5=3, n=18, maxiter=500)
```

When the full-factorial design is larger, the macro uses PROC FACTEX to create a fractional-factorial candidate set. In those cases, the methods found in the %MktEx macro usually make better designs than those found with the %MktDes macro.

%MktDes Macro Options

The following options can be used with the `%MktDes` macro.

Option	Description
`big=`*n*	size of big candidate set
`cand=`*SAS-data-set*	candidate design
`classopts=`*options*	`class` statement options
`coding=`*name*	`coding=` option
`examine=`I | V	matrices that you want to examine
`facopts=`*options*	PROC FACTEX statement options
`factors=`*factor-list*	factors and levels for each factor
`generate=`*options*	`generate` statement options
`interact=`*interaction-list*	interactions that must be estimable
`iter=`*n*	number of designs
`keep=`*n*	number of designs to keep
`maxiter=`*n*	number of designs
`method=`*name*	search method
`n=`*n* | SATURATED	number of runs
`nlev=`*n*	number of levels for pseudo-factors
`options=`*options-list*	binary options
`otherfac=`*variable-list*	other factors
`otherint=`*terms*	multi-step interaction terms
`out=`*SAS-data-set*	output experimental design
`procopts=`*options*	PROC OPTEX statement options
`run=`*procedure-list*	list of procedures that may be run
`seed=`*n*	random number seed
`size=`*n* | MIN	candidate-set size
`step=`*n*	step number
`where=`*where-clause*	`where` clause

big= *n*

specifies the size at which the candidate set is considered to be big. By default, `big=2188`. If the size of the full-factorial design is less than or equal to this size, and if PROC PLAN is in the `run=` list, the macro uses PROC PLAN instead of PROC FACTEX to create the candidate set. The default of 2188 is $\max(2^{11}, 3^7) + 1$. Specifying values as large as `big=6000` or even slightly more is often reasonable. However, run time is slower as the size of the candidate set increases. The `%MktEx` macro coordinate-exchange algorithm will usually work better than a candidate-set search when the full-factorial design has more than several thousand runs.

cand= *SAS-data-set*

specifies the output data set with the candidate design (from PROC FACTEX or PROC PLAN). The default name is `Cand` followed by the step number, for example: `Cand1` for step 1, `Cand2` for step 2, and so on. You should only use this option when you are reading an external candidate set. When you specify `step=` values greater than 1, the macro assumes the default candidate set names, CAND1, CAND2, and so on, were used in previous steps. Specify just a data set name, no data set options.

classopts= *options*
specifies PROC OPTEX `class` statement options. The default, is `classopts=param=orthref`. You probably never want to change this option.

coding= *name*
specifies the PROC OPTEX `coding=` option. This option is usually not needed.

examine= I | V
specifies the matrices that you want to examine. The option `examine=I` prints the information matrix, $\mathbf{X}'\mathbf{X}$; `examine=V` prints the variance matrix, $(\mathbf{X}'\mathbf{X})^{-1}$; and `examine=I V` prints both. By default, these matrices are not printed.

facopts= *options*
specifies PROC FACTEX statement options.

factors= *factor-list*
specifies the factors and the number of levels for each factor. The `factors=` option must be specified. All other options are not required. Here is a simple example of creating a design with 10 two-level factors.

```
%mktdes(factors=x1-x10=2)
```

First, a factor list, which is a valid SAS variable list, is specified. The factor list must be followed by an equal sign and an integer, which gives the number of levels. Multiple lists can be specified. For example, to create 5 two-level factors, 5 three-level factors, and 5 five-level factors, specify:

```
%mktdes(factors=x1-x5=2 x6-x10=3 x11-x15=5)
```

By default, this macro creates each factor in a fractional-factorial candidate set from a minimum number of pseudo-factors. Pseudo-factors are not output; they are used to create the factors of interest and then discarded. For example, with `nlev=2`, a three-level factor `x1` is created from 2 two-level pseudo-factors (`_1` and `_2`) and their interaction by coding down:

```
(_1=1, _2=1) -> x1=1
(_1=1, _2=2) -> x1=2
(_1=2, _2=1) -> x1=3
(_1=2, _2=2) -> x1=1
```

This creates imbalance − the 1 level appears twice as often as 2 and 3. Somewhat better balance can be obtained by instead using three pseudo-factors. The number of pseudo-factors may be specified in parentheses after the number of levels. Example:

```
%mktdes(factors=x1-x5=2 x6-x10=3(3))
```

The levels 1 to 8 are coded down to 1 2 3 1 2 3 1 3, which is better balanced. The cost is candidate-set size may increase and efficiency may actually decrease. Some researchers are willing to sacrifice a little bit of efficiency in order to achieve better balance.

generate= *options*
specifies the PROC OPTEX `generate` statement options. By default, additional options are not added to the `generate` statement.

interact= *interaction-list*
specifies interactions that must be estimable. By default, no interactions are guaranteed to be estimable. Examples:

```
interact=x1*x2
interact=x1*x2 x3*x4*x5
interact=x1|x2|x3|x4|x5@2
```

The interaction syntax is like PROC GLM's and many of the other modeling procedures. It uses "`*`" for simple interactions (`x1*x2` is the interaction between `x1` and `x2`), "`|`" for main effects and interactions (`x1|x2|x3` is the same as `x1 x2 x1*x2 x3 x1*x3 x2*x3 x1*x2*x3`) and "`@`" to eliminate higher-order interactions (`x1|x2|x3@2` eliminates `x1*x2*x3` and is the same as `x1 x2 x1*x2 x3 x1*x3 x2*x3`). The specification "`@2`" allows only main effects and two-way interactions. Only "`@`" values of `2` or `3` are allowed.

iter= *n*

maxiter= *n*
specifies the PROC OPTEX `iter=` option which creates *n* designs. By default, `iter=10`.

keep= *n*
specifies the PROC OPTEX `keep=` option which keeps the *n* best designs. By default, `keep=5`.

nlev= *n*
specifies the number of levels from which factors are constructed through pseudo-factors and coding down. The value must be a prime or a power of a prime: 2, 3, 4, 5, 7, 8, 9, 11 .... This option is used with PROC FACTEX:

```
factors factors / nlev=&nlev;
```

By default, the macro uses the minimum prime or power of a prime from the `factors=` list or 2 if no suitable value is found.

method= *name*
specifies the PROC OPTEX `method=` search method option. The default is `method=m_Fedorov` (modified Fedorov).

n= *n* | **SATURATED**
specifies the PROC OPTEX `n=` option, which is the number of runs in the final design. The default is the PROC OPTEX default and depends on the problem. Typically, you will not want to use the default. Instead, you should pick a value using the information produced by the `%MktRuns` macro as guidance (see page 600). The `n=saturated` option creates a design with the minimum number of runs.

options= *options-list*
specifies binary options. By default, none of these options are specified. Specify one or more of the following values after `options=`.

`check`
checks the efficiency of a given design, specified in `cand=`.

`nocode`
suppresses printing the PROC PLAN, PROC FACTEX, and PROC OPTEX code.

`allcode`
shows all code, even code that will not be run.

otherfac= *variable-list*
specifies other terms to mention in the `factors` statement of PROC FACTEX. These terms are not guaranteed to be estimable. By default, there are no other factors.

otherint= *terms*
specifies interaction terms that will only be specified with PROC OPTEX for multi-step macro invocations. By default, no interactions are guaranteed to be estimable. Normally, interactions that are specified via the `interact=` option affect both the PROC FACTEX and the PROC OPTEX `model` statements. In multi-step problems, part of an interaction may not be in a particular PROC FACTEX step. In that case, the interaction term must only appear in the PROC OPTEX step. For example, if `x1` is created in one step and `x4` is created in another, and if the `x1*x4` interaction must be estimable, specify `otherint=x1*x4` on the final step, the one that runs PROC OPTEX.

```
%mktdes(step=1, factors=x1-x3=2, n=30, run=factex)

%mktdes(step=2, factors=x4-x6=3, n=30, run=factex)

%mktdes(step=3, factors=x7-x9=5, n=30, run=factex optex,
        otherint=x1*x4)
```

out= *SAS-data-set*
specifies the output experimental design (from PROC OPTEX). By default, `out=design`. Often, you will want to specify a two-level name to create a permanent SAS data set so the design will be available later for analysis.

procopts= *options*
specifies PROC OPTEX statement options. By default, no options are added to the PROC OPTEX statement.

run= *procedure-list*
specifies the list of procedures that the macro may run. Normally, the macro runs either PROC FACTEX or PROC PLAN and then PROC OPTEX. By default, `run=plan factex optex`. You can skip steps by omitting procedure names from this list. When both PLAN and FACTEX are in the list, the macro chooses between them based on the size of the full-factorial design and the value of `big=`.

When PLAN is not in the list, the macro generates code for PROC FACTEX.

seed= n

specifies the random number seed. By default, `seed=0`, and clock time is used to make the random number seed. By specifying a random number seed, results should be reproducible within a SAS release for a particular operating system. However, due to machine differences, some results may not be exactly reproducible on other machines, although you would expect the efficiency differences to be slight.

size= n | MIN

specifies the candidate-set size. Start with the default `size=min` and see how big that design is. If you want, subsequently you can specify larger values that are `nlev=`n multiples of the minimum size. This option is used with PROC FACTEX:

```
size design=&size;
```

When `nlev=`n, increase the `size=` value by a factor of n each time. For example, when `nlev=2`, increase the `size=` value by a factor of two each time. If `size=min` implies `size=128`, then 256, 512, 1024, and 2048 are reasonable sizes to try. Integer expressions like `size=128*4` are allowed.

step= n

specifies the step number. By default, there is only one step. However, sometimes, a better design can be found using a multi-step approach. Do not specify the `cand=` option on any step of a multi-step run. Consider the problem of making a design with 3 two-level factors, 3 three-level factors, and 3 five-level factors. The simplest approach is to do something like this − create a design from two-level factors using pseudo-factors and coding down.

```
%mktdes(factors=x1-x3=2 x4-x6=3 x7-x9=5, n=30)
```

However, for small problems like this, the following three-step approach will usually be better.

```
%mktdes(step=1, factors=x1-x3=2, n=30, run=factex)
%mktdes(step=2, factors=x4-x6=3, n=30, run=factex)
%mktdes(step=3, factors=x7-x9=5, n=30, run=factex optex)
```

Note however, that the following `%MktEx` macro call will usually be better still.

```
%mktex(2 2 2 3 3 3 5 5 5, n=30)
```

The first `%MktDes` macro step uses PROC FACTEX to create a fractional-factorial design for the two-level factors. The second step uses PROC FACTEX to create a fractional-factorial design for the three-level factors and cross it with the two-level factors. The third step uses PROC FACTEX to create a fractional-factorial design for the five-level factors and cross it with the design for the two and three-level factors and then run PROC OPTEX.

Each step globally stores two macro variables (`&class1` and `&inter1` for the first step, `&class2` and `&inter2` for the second step, ...) that are used to construct the PROC OPTEX `class` and `model` statements. When `step` > 1, variables from the previous steps are used in the `class` and `model` statements. In this example, the following PROC OPTEX code is created by step 3:

```
proc optex data=Cand3;
   class
      x1-x3
      x4-x6
      x7-x9
      / param=orthref;
   model
      x1-x3
      x4-x6
      x7-x9
      ;
   generate n=30 iter=10 keep=5 method=m_fedorov;
   output out=Design;
   run; quit;
```

This step uses the previously stored macro variables `&class1=x1-x3` and `&class2=x4-x6`.

where= *where-clause*
specifies a SAS `where` clause for the candidate design, which is used to restrict the candidates. By default, the candidate design is not restricted.

%MktDes Macro Notes

This macro specifies `options nonotes` throughout much of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktDups Macro

The `%MktDups` autocall macro detects duplicate choice sets and duplicate alternatives within generic choice sets. See page 340 for an example. To illustrate, consider a simple experiment with these two choice sets. These choice sets are completely different and are not duplicates.

```
a b c       a b c
1 2 1       1 1 1
2 1 2       2 2 2
1 1 2       2 2 1
2 1 1       1 2 2
```

Now consider these two choice sets:

```
a b c       a b c
1 2 1       2 1 2
2 1 2       1 1 2
1 1 2       2 1 1
2 1 1       1 2 1
```

They are the same for a generic study because all of the same alternatives are there, they are just in a different order. However, for a branded study they are different. For a branded study, there would be a different brand for each alternative, so the choice sets would be the same only if all the same alternatives appeared in the same order. For both a branded and generic study, these choice sets are duplicates:

```
a b c       a b c
1 2 1       1 2 1
2 1 2       2 1 2
1 1 2       1 1 2
2 1 1       2 1 1
```

Now consider these choice sets for a generic study.

```
a b c       a b c
1 2 1       1 2 1
2 1 1       1 2 1
1 1 2       1 1 2
2 1 1       2 1 1
```

First, each of these choice sets has duplicate alternatives (2 1 1 in the first and 1 2 1 in the second). Second, these two choice sets are flagged as duplicates, even though they are not exactly the same. They are flagged as duplicates because every alternative in choice set one is also in choice set two, and every alternative in choice set two is also in choice set one. In generic studies, two choice sets are considered duplicates unless one has one or more alternatives that are not in the other choice set.

Here is an example. A design is created with the `%ChoicEff` macro choice-set-swapping algorithm for a branded study, then the `%MktDups` macro is run to check for and eliminate duplicate choice sets.

```
%mktex(3 ** 9, n=27, seed=424)

data key;
   input (Brand x1-x3) ($);
   datalines;
Acme   x1 x2 x3
Ajax   x4 x5 x6
Widgit x7 x8 x9
;

%mktroll(design=randomized, key=key, alt=brand, out=cand);

%choiceff(data=cand, model=class(brand x1-x3), seed=420,
          nsets=18, nalts=3, beta=zero);

proc freq; tables set; run;

%mktdups(branded, data=best, factors=brand x1-x3, nalts=3, out=out)

proc freq; tables set; run;
```

The first PROC FREQ output shows us that several candidate choice sets occur more than once in the design.

The FREQ Procedure

Set	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	11.11	6	11.11
3	3	5.56	9	16.67
4	6	11.11	15	27.78
13	6	11.11	21	38.89
15	6	11.11	27	50.00
16	3	5.56	30	55.56
21	3	5.56	33	61.11
22	3	5.56	36	66.67
23	6	11.11	42	77.78
25	9	16.67	51	94.44
27	3	5.56	54	100.00

The `%MktDups` macro prints the following information to the log:

```
Design:          Branded
Factors:         brand x1-x3
                 Brand
                 x1 x2 x3
Duplicate Sets:  7
```

The output from the %MktDups macro contains the following table:

Choice Set	Duplicate Choice Sets To Delete
1	5
2	16
3	9
	17
7	13
8	15
11	14

The first line of the first table tells us that this is a branded design as opposed to generic. The second line tells us the factors as specified on the factors= option. These are followed by the actual variable names for the factors. The last line reports the number of duplicates. The second table tells us that choice set 1 is the same as choice set 5. Similarly, 2 and 16 are the same as are 3, 9, and 17, and so on. The out= data set will contain the design with the duplicate choice set eliminated.

Now consider an example with purely generic alternatives.

```
%mktex(2 ** 5, n=2**5, seed=109)
%mktlab(int=f1-f4)

%choiceff(data=final, model=class(x1-x5), seed=93,
          nsets=42, flags=f1-f4, beta=zero);

%mktdups(generic, data=best, factors=x1-x5, nalts=4, out=out)
```

The macro produces the following tables:

```
Design:          Generic
Factors:         x1-x5
                 x1 x2 x3 x4 x5
Sets w Dup Alts: 1
Duplicate Sets:  1
```

Choice Set	Duplicate Choice Sets To Delete
2	25
39	Alternatives

For each choice set listed in the choice set column, either the other choice sets it duplicates are listed or the word `Alternatives` is printed if the problem is with duplicate alternatives.

Here are just the choice sets with duplication problems:

```
proc print data=best;
   var x1-x5;
   id set; by set;
   where set in (2, 25, 39);
   run;
```

Set	x1	x2	x3	x4	x5
2	1	2	1	1	1
	2	2	1	1	1
	1	1	2	2	2
	2	1	2	2	2
25	1	1	2	2	2
	2	1	2	2	2
	2	2	1	1	1
	1	2	1	1	1
39	1	1	2	1	1
	1	1	2	1	1
	2	2	1	2	2
	2	2	1	2	2

You can see that the macro detects duplicates even though the alternatives do not always appear in the same order in the different choice sets.

Now consider another example.

```
%mktex(2 ** 6, n=2**6)

data key;
   input (x1-x2) ($) @@;
   datalines;
x1 x2 x3 x4 x5 x6
;

%mktroll(design=design, key=key, out=cand);

%mktdups(generic, data=cand, factors=x1-x2, nalts=3, out=out)

proc print; by set; id set; run;
```

Here is some of the output. The output lists, for each set of duplicates, the choice set that will be kept (in the first column) and all the matching choice sets that will be deleted (in the second column).

```
Design:          Generic
Factors:         x1-x2
                 x1 x2
Sets w Dup Alts: 40
Duplicate Sets:  50
```

Choice Set	Duplicate Choice Sets To Delete	
1		Alternatives
2		Alternatives
	5	
	6	
	17	
	18	
	21	
.		
.		
.		

Here are the unique choice sets.

Set	_Alt_	x1	x2
7	1	1	1
	2	1	2
	3	2	1

```
 8      1       1      1
        2       1      2
        3       2      2

12      1       1      1
        2       2      1
        3       2      2

28      1       1      2
        2       2      1
        3       2      2
```

This next example creates a conjoint design[§] and tests it for duplicates.

```
%mktex( 3 ** 3 2 ** 2, n=19, seed=121)

%mktdups(linear, factors=x1-x5);
```

```
Design:          Linear
Factors:         x1-x5
                 x1 x2 x3 x4 x5
Duplicate Runs:  2
```

Run	Duplicate Runs To Delete
3	4
10	11

%MktDups Macro Options

The following options can be used with the `%MktDups` macro.

Option	Description
`data=`*SAS-data-set*	input choice design
`factors=`*variable-list*	factors in the design
`nalts=`*n*	number of alternatives
`options`	binary options
`out=`*SAS-data-set*	output data set
`outlist=`*SAS-data-set*	output data set with duplicates
`vars=`*variable-list*	factors in the design

[§]Normally, we would use 18 runs and not a prime number like 19 that is not divisible by any of the numbers of levels, 2 and 3. We picked a silly number like 19 to ensure duplicates for this example.

Positional Parameter

The options list is a positional parameter. This means it must come first, and unlike all other parameters, it is not specified after a name and an equal sign.

options

For the first option, specify one or more of the following. You may specify `noprint` and one of the following: `generic`, `branded`, or `linear`.

`branded`
specifies that since one of the factors is brand, the macro only needs to compare corresponding alternatives in each choice set.

`generic`
specifies a generic design and is the default. This means that there are no brands, so options are interchangeable, so the macro needs to compare each alternative with every other alternative in every choice set.

`linear`
specifies a linear not a choice design. Specify `linear` for a full-profile conjoint design, for an ANOVA design, or for the linear version of a branded choice design.

`noprint`
specifies no printed output. This option will be used when you are only interested in the output data set or macro variable.

Example:

```
%mktdups(branded noprint, nalts=3)
```

Required Options

This next option is mandatory with choice designs.

nalts= *n*

specifies the number of alternatives. This option must be specified with generic or branded designs. It is ignored with linear designs. For generic or branded designs, the `data=` data set must contain `nalts=` observations for the first choice set, `nalts=` observations for the second choice set, and so on.

Other Options

Here are the other options.

data= *SAS-data-set*

specifies the input choice design. By default, the macro uses the last data set created.

out= *SAS-data-set*
specifies an output data set that contains the design with duplicate choice sets excluded. By default, no data set is created, and the macro just reports on duplicates. Often, you will want to specify a two-level name to create a permanent SAS data set so the design will be available later for analysis.

outlist= *SAS-data-set*
specifies the output data set with the list of duplicates. By default, `outlist=outdups`.

vars= *variable-list*

factors= *variable-list*
specifies the factors in the design. By default, all numeric variables are used.

%MktDups Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktEval Macro

The %MktEval autocall macro helps you evaluate an experimental design. There are numerous examples of its usage from pages 113 through 287. The %MktEval macro reports on balance and orthogonality. Typically, you will call it immediately after running the %MktEx macro. The output from this macro contains two default tables. The first table shows the canonical correlations between pairs of coded factors. A canonical correlation is the maximum correlation between linear combinations of the coded factors. See page 90 for more information about canonical correlations. All zeros off the diagonal show that the design is orthogonal for main effects. Off-diagonal canonical correlations greater than 0.316 ($r^2 > 0.1$) are listed in a separate table.

For nonorthogonal designs and designs with interactions, the canonical-correlation matrix is not a substitute for looking at the variance matrix with the %MktEx macro. It just provides a quick and more-compact picture of the correlations between the factors. The variance matrix is sensitive to the actual model specified and the coding. The canonical-correlation matrix just tells you if there is some correlation between the main effects. When is a canonical correlation too big? You will have to decide that for yourself. In part, the answer depends on the factors and how the design will be used. A high correlation between the client's and the main competitor's price factor is a serious problem meaning you will need to use a different design. In contrast, a moderate correlation in a choice design between one brand's minor attribute and another brand's minor attribute may be perfectly fine.

The macro also prints one-way, two-way and n-way frequencies. Equal one-way frequencies occur when the design is balanced. Equal two-way frequencies occur when the design is orthogonal. Equal n-way frequencies, all equal to one, occur when there are no duplicate runs or choice sets.

Here is a typical usage:

```
%mktex(2 2 3 ** 6, n=18, unbalanced=0, seed=289)
%mkteval;
```

Canonical Correlations Between the Factors
There is 1 Canonical Correlation Greater Than 0.316

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1	0.33	0	0	0	0	0	0
x2	0.33	1	0	0	0	0	0	0
x3	0	0	1	0	0	0	0	0
x4	0	0	0	1	0	0	0	0
x5	0	0	0	0	1	0	0	0
x6	0	0	0	0	0	1	0	0
x7	0	0	0	0	0	0	1	0
x8	0	0	0	0	0	0	0	1

Canonical Correlations > 0.316 Between the Factors
There is 1 Canonical Correlation Greater Than 0.316

		r	r Square
x1	x2	0.33	0.11

```
                Summary of Frequencies
There is 1 Canonical Correlation Greater Than 0.316
        * - Indicates Unequal Frequencies

                Frequencies

      x1        9 9
      x2        9 9
      x3        6 6 6
      x4        6 6 6
      x5        6 6 6
      x6        6 6 6
      x7        6 6 6
      x8        6 6 6
 *    x1 x2     3 6 6 3
      x1 x3     3 3 3 3 3 3
      x1 x4     3 3 3 3 3 3
      x1 x5     3 3 3 3 3 3
      x1 x6     3 3 3 3 3 3
      x1 x7     3 3 3 3 3 3
      x1 x8     3 3 3 3 3 3
      x2 x3     3 3 3 3 3 3
      x2 x4     3 3 3 3 3 3
      x2 x5     3 3 3 3 3 3
      x2 x6     3 3 3 3 3 3
      x2 x7     3 3 3 3 3 3
      x2 x8     3 3 3 3 3 3
      x3 x4     2 2 2 2 2 2 2 2 2
      x3 x5     2 2 2 2 2 2 2 2 2
      x3 x6     2 2 2 2 2 2 2 2 2
      x3 x7     2 2 2 2 2 2 2 2 2
      x3 x8     2 2 2 2 2 2 2 2 2
      x4 x5     2 2 2 2 2 2 2 2 2
      x4 x6     2 2 2 2 2 2 2 2 2
      x4 x7     2 2 2 2 2 2 2 2 2
      x4 x8     2 2 2 2 2 2 2 2 2
      x5 x6     2 2 2 2 2 2 2 2 2
      x5 x7     2 2 2 2 2 2 2 2 2
      x5 x8     2 2 2 2 2 2 2 2 2
      x6 x7     2 2 2 2 2 2 2 2 2
      x6 x8     2 2 2 2 2 2 2 2 2
      x7 x8     2 2 2 2 2 2 2 2 2
      N-Way     1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

All factors in this design are perfectly balanced, and almost all are orthogonal, but `x1` and `x2` are correlated with each other.

%MktEval Macro Options

The following options can be used with the `%MktEval` macro.

Option	Description
`blocks=`*variable*	blocking variable
`data=`*SAS-data-set*	input data set with design
`factors=`*variable-list*	factors in the design
`format=`*format*	format for canonical correlations
`freqs=`*frequency-list*	frequencies to print
`list=`*n*	minimum canonical correlation to list
`outcb=`*SAS-data-set*	within-block canonical correlations
`outcorr=`*SAS-data-set*	canonical correlation matrix
`outfreq=`*SAS-data-set*	frequencies
`outfsum=`*SAS-data-set*	frequency summaries
`outlist=`*SAS-data-set*	list of largest canonical correlations
`print=list`	controls the printing of the results
`vars=`*variable-list*	list of the factors

blocks= *variable*
specifies a blocking variable. This option prints separate canonical correlations within each block. By default, there is one block.

data= *SAS-data-set*
specifies the input SAS data set with the experimental design. By default, the macro uses the last data set created.

factors= *variable-list*

vars= *variable-list*
specifies a list of the factors in the experimental design. The default is all of the numeric variables in the data set.

freqs= *frequency-list*
specifies the frequencies to print. By default, `freqs=1 2 n`, and 1-way, 2-way, and n-way frequencies are printed. Do not specify the exact number of ways instead of `n`. For ways other than `n`, the macro checks for and prints zero cell frequencies. For n-ways, the macro does not output or print zero frequencies. Only the full-factorial design will have nonzero cells, so specifying something like `freqs=1 2 20` will make the macro take a *long* time, and it will try to create *huge* data sets and will probably run out of memory or disk space before it is done. However, `freqs=1 2 n` runs very reasonably.

format= *format*
specifies the format for printing canonical correlations. The default format is `4.2`.

list= *n*
specifies the minimum canonical correlation to list. The default is 0.316, the square root of $r^2 = 0.1$.

outcorr= *SAS-data-set*
specifies the output SAS data set for the canonical correlation matrix. The default data set name is CORR.

outcb= *SAS-data-set*
specifies the output SAS data set for the within-block canonical correlation matrices. The default data set name is CB.

outlist= *SAS-data-set*
specifies the output data set for the list of largest canonical correlations. The default data set name is LIST.

outfreq= *SAS-data-set*
specifies the output data set for the frequencies. The default data set name is FREQ.

outfsum= *SAS-data-set*
specifies the output data set for the frequency summaries. The default data set name is FSUM.

print= `short|corr|list|freqs|summ|all`
controls the printing of the results. Specify one or more values from the following list.

`corr`	prints the canonical correlations matrix.
`block`	prints the canonical correlations within block.
`list`	prints the list of canonical correlations greater than the `list=` value.
`freqs`	prints the frequencies, specified by the `freqs=` option.
`summ`	prints the frequency summaries.
`all`	prints all of the above.
`short`	is the default and is equivalent to: `corr list summ block`.
`noprint`	specifies no printed output.

By default, the frequency list, which contains the factor names, levels, and frequencies is not printed, but the more compact frequency summary list, which contains the factors and frequencies but not the levels is printed.

%MktEval Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktEx Macro

The %MktEx autocall macro is designed for marketing researchers and any one else who wants to make good, efficient experimental designs. There are numerous examples of its usage from pages 110 through 338. The %MktEx macro is designed to be very simple to use and to run in seconds for trivial problems, minutes for small problems, and in less than an hour for larger and difficult problems. This macro is a full-featured linear designer that can handle simple problems like main-effects designs and more complicated problems including designs with interactions and restrictions on which levels can appear together. The macro is particularly designed to easily create the kinds of linear designs that marketing researches need for conjoint and choice experiments. For any linear design problem, you can simply run the macro once, specifying only the number of runs and the numbers of levels of all the factors. You will no longer have to try different algorithms and different approaches to see which one works best. The macro does all of that for you. Kuhfeld, Tobias, and Garratt (2003) state on page 40 "The best approach to design creation is to use the computer as a tool along with traditional design skills, not as a substitute for thinking about the problem." With the %MktEx macro, we try to automate some of the thought processes of the expert designer.

Here is an example of using the %MktEx macro to create a design with 5 two-level factors, 4 three-level factors, 3 five-level factors, 2 six-level factors, all in 60 runs (rows or conjoint profiles or choice sets).

```
%mktex( 2 ** 5  3 ** 4  5 5 5  6 6,  n=60 )
```

The notation `m ** n` means m^n or n m-level factors. For example `2 ** 5` means $2 \times 2 \times 2 \times 2 \times 2$ or 5 two-level factors.

The %MktEx macro creates efficient linear experimental designs using several approaches. The macro will try to directly create an orthogonal design, it will search a set of candidate runs (rows of the design), and it will use a coordinate-exchange algorithm using both random initial designs and also a partial orthogonal design initialization. The macro stops if at any time it finds a perfect, 100% efficient, orthogonal and balanced design. This first phase is the algorithm search phase. In it, the macro determines which approach is working best for this problem. At the end of this phase, the macro chooses the method that has produced the best design and performs another set of iterations using exclusively the chosen approach. Finally, the macro performs a third set of iterations where it takes the best design it found so far and tries to improve it.

In all phases, the macro attempts to optimize D-efficiency (sometimes known as D-optimality), which is a standard measure of the goodness of the experimental design. As D-efficiency increases, the standard errors of the parameter estimates in the linear model decrease. A perfect design is orthogonal and balanced and has 100% D-efficiency. A design is orthogonal when all of the parameter estimates are uncorrelated. A design is balanced when all of the levels within each of the factors occur equally often. A design is orthogonal and balanced when the variance matrix, which is proportional to $(\mathbf{X}'\mathbf{X})^{-1}$ is diagonal, where $\mathbf{X}$ is a suitable orthogonal coding (see page 91) of the design matrix. See page 84 and also Kuhfeld, Tobias, and Garratt (1994), which starts on page 39, for more information on efficient experimental designs.

For most problems, you only need to specify the levels of all the factors and the number of runs. For more complicated problems, you may need to also specify the interactions that you want to be estimable or restrictions on which levels may not appear together. Other than that, you should not need any other options for most problems. This macro is not like other design tools that you have to tell what to do. With this macro, you just tell it what you want, and it figures out a good way to do it. For some problems, the sophisticated user, with a lot of work, may be able to adjust the options

to come up with a better design. However, this macro should always produce a very good design with minimal effort for even the most unsophisticated users.

The %MktEx macro has the world's largest catalog of orthogonal arrays. The orthogonal arrays are constructed using methods and arrays from a variety of sources, for example: Dey (1985), Wang and Wu (1991), Wang (1996a, 1996b), Hedayat, Sloane, and Stufken (1999), De Cock and Stufken (2000), Zhang, Pang, and Wang (2001), Xu (2002), Sloane (2002), and the SAS FACTEX procedure. Many of the designs were created using difference schemes and the methods of Wang and Wu (1991) and Wang (1996a, 1996b). Other relevant references include Hadamard (1893), Rao (1947), Addelman (1962a, 1962b), Bose (1947), Taguchi (1987), Suen (1989a, 1989b), and Wang and Wu (1989).

When n=n is a multiple of 4, the %MktEx macro can construct orthogonal designs with up to $n-1$ two-level factors. The two-level designs are constructed from Hadamard matrices (Hadamard, 1893; Paley, 1933; Williamson, 1944; Hedayat, Sloane, and Stufken, 1999). The next table shows the available sizes up through n=1000:

Hadamard Matrix Sizes Up to n=1000															
4	8	12	16	20	24	28	32	36	40	44	48	52	56	60	64
68	72	76	80	84	88	92	96	100	104	108	112	116	120	124	128
132	136	140	144	148	152	156	160	164	168	172	176	180	184	188	192
196	200	204	208	212	216	220	224	228	232	236	240	244	248	252	256
264	272	276	280	284	288	296	300	304	308	312	316	320	328	332	336
344	348	352	360	364	368	376	380	384	388	392	396	400	408	416	420
424	432	440	444	448	456	460	464	468	472	480	484	488	492	496	500
504	512	516	524	528	540	544	548	552	556	560	564	568	572	576	588
592	600	608	616	620	624	628	632	636	640	644	656	660	664	672	676
684	688	692	696	700	704	708	720	728	736	740	748	752	760	768	776
780	784	788	792	796	800	804	812	816	820	828	832	840	844	848	860
864	868	880	884	888	896	900	908	912	916	920	924	928	936	944	948
960	968	972	976	984	992	1000									

Larger sizes are available as well. The %MktEx macro can construct these designs when n is a multiple of 4 and one or more of the following hold:

- $n \leq 256$
- $n-1$ is prime
- $n/2-1$ is prime and $\mathrm{mod}(n/2, 4) = 2$
- n is a power of 2 (2, 4, 8, 16, ...) times the size of a smaller Hadamard matrix that is available.

For some of these sizes, the macro can create orthogonal designs with a small number (say m) four-level factors in place of $3 \times m$ of the two-level factors (for example, $2^{70}\ 4^3$ in 80 runs).

You can see more sizes of Hadamard matrices that the macro knows how to make by running the next program. Note however that the fact that a number appears in this program's listing, does not guarantee that your computer will have enough memory and other resources to create it.

```
data x;
   length Method $ 12;
   do n = 4 to 10000 by 4;
      HadSize = n; method = ' ';
      do while(mod(hadsize, 8) eq 0); hadsize = hadsize / 2; end;
      ispm1 = 1; ispm2 = mod(hadsize / 2, 4) eq 2;
      do i = 2 to sqrt(hadsize) while(ispm1);
         ispm1 = mod(hadsize - 1, i);
         end;
      do i = 2 to sqrt(hadsize / 2) while(ispm2);
         ispm2 = mod(hadsize / 2 - 1, i);
         end;
      if      ispm1          then method = 'Paley 1';
      else if ispm2          then method = 'Paley 2';
      else if hadsize le 256 then method = 'Williamson';
      if method ne ' ' then do; Change = n - lag(n); output; end;
      end;
   run;

proc print label noobs;
   label hadsize = 'Reduced Hadamard Matrix Size';
   var n hadsize method change;
   run;
```

Here is a simple example of using the `%MktEx` macro to request the L_{36} design, $2^{11}3^{12}$, which has 11 two-level factors and 12 three-level factors.

```
%mktex( n=36 )
```

No iterations are needed, and the macro immediately creates the L_{36}, which is 100% efficient. This example runs in a few seconds. The factors are always named `x1`, `x2`, ... and the levels are always consecutive integers starting with 1. You can use the `%MktLab` macro to assign different names and levels (see page 577).

By default, the macro creates two output data sets with the design.

- `out=Design` - the experimental design, sorted by the factor levels.
- `outr=Randomized` - the randomized experimental design.

The two designs are equivalent and have the same D-efficiency. The `out=Design` data set is sorted and hence is usually easier to look at, however the `outr=Randomized` design is the better one to use. The randomized design has the rows sorted into a random order, and all of the factor levels are randomly reassigned. For example with two-level factors, approximately half of the original (1, 2) mappings will be reassigned (2, 1). Similarly, with three level factors, the mapping (1, 2, 3) will be changed to one of the following: (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), or (3, 2, 1). The reassignment of levels is usually not critical for the iteratively derived designs, but it can be very important the orthogonal designs, many of which have all ones in the first row.

The candidate-set search has two parts. First, either PROC PLAN is run to create a full-factorial design for small problems, or PROC FACTEX is run to create a fractional-factorial design for large problems. Either way, this design is a candidate set that in the second part is searched by PROC

OPTEX using the modified Fedorov algorithm. A design is built from a selection of the rows of the candidate set (Fedorov, 1972; Cook and Nachtsheim, 1980). The modified Fedorov algorithm considers each run in the design and each candidate run. Candidate runs are swapped in and design runs are swapped out if the swap improves *D*-efficiency.

Next, the `%MktEx` macro uses the coordinate-exchange algorithm, based on Meyer and Nachtsheim (1995). The coordinate-exchange algorithm considers each level of each factor, and considers the effect on *D*-efficiency of changing a level ($1 \rightarrow 2$, or $1 \rightarrow 3$, or $2 \rightarrow 1$, or $2 \rightarrow 3$, or $3 \rightarrow 1$, or $3 \rightarrow 2$, and so on). Exchanges that increase efficiency are performed. Typically, the macro first tries to initialize the design with an orthogonal (or tabled) design (`Tab`) and a random design (`Ran`) both. Levels that are not orthogonally initialized may be exchanged for other levels if the exchange increases efficiency.

The initialization may be more complicated. Say you asked for the design $4^1 5^1 3^5$ in 18 runs. The macro would use the orthogonal design $3^6 6^1$ in 18 runs to initialize the three-level factors orthogonally, and the five-level factor with the six-level factor coded down to five levels (and hence unbalanced). The four-level factor would be randomly initialized. The macro would also try the same initialization but with a random rather than unbalanced initialization of the five-level factor, as a minor variation on the first initialization. In the next initialization variation, the macro would use a fully random initialization. If the number of runs requested were smaller than the number or runs in the initial orthogonal design, the macro would initialize the design with just the first n rows of the orthogonal design. Similarly, if the number of runs requested were larger than the number or runs in the initial orthogonal design, the macro would initialize part of the design with the orthogonal design and the remaining rows and columns randomly. The coordinate-exchange algorithm considers each level of each factor that is not orthogonally initialized, and it exchanges a level if the exchange improves *D*-efficiency. When the number or runs in the orthogonal design does not match the number of runs desired, none of the design is initialized orthogonally.

The coordinate-exchange algorithm is not restricted by having a candidate set and hence can *potentially* consider every possible design. That is, no design is precluded from consideration due to the limitations of a candidate set. In practice, however, both the candidate-set-based and coordinate-exchange algorithms consider only a tiny fraction of the possible designs. When the number of runs in the full-factorial design is very small (say 100 or 200 runs), the modified Fedorov algorithm and coordinate exchange algorithms usually work equally well. When the number of runs in the full-factorial design is small (up to several thousand), the modified Fedorov algorithm is usually superior to coordinate exchange. When the full-factorial design is larger, coordinate exchange is usually the superior approach. However, heuristics like these are sometimes wrong, which is why the macro tries both methods to see which one is really best for each problem.

Next, the `%MktEx` macro determines which algorithm (candidate set search, coordinate exchange with partial orthogonal initialization, or coordinate exchange with random initialization) is working best and tries more iterations using that approach. It starts by printing the initial (`Ini`) best efficiency.

Next, the `%MktEx` macro tries to improve the best design it found previously. Using the previous best design as an initialization (`Pre`), and random mutations of the initialization (`Mut`) and simulated annealing (`Ann`), the macro uses the coordinate-exchange algorithm to try to find a better design. This step is important because the best design that the macro found may be an intermediate design and may not be the final design at the end of an iteration. Sometimes the iterations deliberately make the designs less efficient, and sometimes, the macro never finds a design as efficient or more efficient again. Hence it is worthwhile to see if the best design found so far can be improved. At the end, PROC OPTEX is called to print the levels of each factor and the final *D*-efficiency.

Random mutations involve adding random noise to the initial design before iterations start (levels are randomly changed). This may eliminate the perfect balance that will often be in the initial design. By default, random mutations are used with designs with fully random initializations and in the design refinement step; orthogonal initial designs are not mutated.

Coordinate exchange can be combined with the simulated annealing optimization technique (Kirkpatrick, Gellat, and Vecchi 1983). Annealing refers to the cooling of a liquid in a heat bath. The structure of the solid depends on the rate of cooling. Coordinate exchange without simulated annealing seeks to maximize *D*-efficiency at every step. Coordinate exchange with simulated annealing allows *D*-efficiency to occasionally decrease with a probability that decreases with each iteration. This is analogous to slower cooling, and it helps overcome local optima.

For design 1, for the first level of the first factor, by default, the macro may execute an exchange (say change a 2 to a 1) that makes the design worse with probability 0.05. As more and more exchanges occur, this probability decreases so at the end of the processing of design 1, exchanges that decrease efficiency are hardly ever done. For design 2, this same process is repeated, again starting by default with an annealing probability of 0.05. This often helps the algorithm overcome local efficiency maxima. To envision this, imagine that you are standing on a molehill next to a mountain. The only way you can start going up the mountain is to first step down off the molehill. Once you are on the mountain, you may occasionally hit a dead end, where all you can do is step down and look for a better place to continue going up. Simulated annealing, by occasionally stepping down the efficiency function, often allows the macro to go farther up it than it would otherwise. The simulated annealing is why you will sometimes see designs getting worse in the iteration history. The macro keeps track of the best design, not the final design in each step. By default, annealing is used with designs with fully random initializations and in the design refinement step. Simulated annealing is not used with orthogonally initialized designs.

%MktEx Macro Notes

The `%MktEx` macro prints notes to the SAS log to show you what it is doing while it is running. Most of the notes that would normally come out of the macro's procedure and DATA steps are suppressed by default by an `options nonotes` statement. This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro. This section describes the notes that are normally not suppressed.

The macro will usually start by printing one of the following notes (filling in a value after `n=`).

```
NOTE: Generating the Hadamard design, n=.
NOTE: Generating the full-factorial design, n=.
NOTE: Generating the fractional-factorial design, n=.
NOTE: Generating the orthogonal array design, n=.
```

These messages tell you which type of orthogonal design the macro is constructing. The design may be the final design, or it may provide an initialization for the coordinate exchange algorithm. In some cases, it may not have the same number of runs, `n`, as the final design. Usually this step is fast, but constructing some fractional-factorial designs may be time consuming.

If the macro is going to use PROC OPTEX to search a candidate set, it will print this note.

```
NOTE: Generating the candidate set.
```

This step will usually be fast. Next, when a candidate set is searched, the macro will print this next note, substituting in values for the ellipses.

```
NOTE: Performing ... searches of ... candidates.
```

This step may be take a while depending on the size of the candidate set and the size of the design. When there are a lot of restrictions and a fractional-factorial candidate set is being used, the candidate set may be so restricted that it does not contain enough information to make the design. In that case, you will get this message.

```
NOTE: The candidate-set initialization failed,
      but the MKTEX macro is continuing.
```

Even though part of the macro's algorithm failed, it is *not* a problem. The macro just goes on to the coordinate-exchange algorithm, which will almost certainly work better than searching any severely-restricted candidate set.

Sometimes you will get this note.

```
NOTE: Stopping since it appears that no improvement is possible.
```

When the macro keeps finding the same maximum D-efficiency over and over again in different designs, it may stop early. This may mean that the macro has found the optimal design, or it may mean that the macro keeps finding a very attractive local optimum. Either way, it is unlikely that the macro will do any better. You can control this using the `stopearly=` option.

The macro has options that control the amount of time it spends trying different techniques. When time expires, the macro may switch to other techniques before it completes the usual maximum number of iterations. When this happens, the macro tells you.

```
NOTE: Switching to a random initialization after ... minutes and
      ... designs.
NOTE: Quitting the algorithm search after ... minutes and ... designs.
NOTE: Quitting the design search after ... minutes and ... designs.
NOTE: Quitting the refinement step after ... minutes and ... designs.
```

When there are restrictions, or when you specify that you do not want duplicate runs, you may also specify `options=accept`. This means you are willing to accept designs that violate the restrictions. With `options=accept`, the macro will tell you if the restrictions are not met.

```
NOTE: The restrictions were not met.
NOTE: The design has duplicate runs.
```

The macro ends with one of the following two messages.

```
NOTE: The MKTEX macro used ... seconds.
NOTE: The MKTEX macro used ... minutes.
```

%MktEx Macro Iteration History

This section provides information on interpreting the iteration history table produced by the `%MktEx` macro. Here is part of a table.

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	82.2172	82.2172	Can
1	End	82.2172		
2	Start	78.5039		Tab,Ran
2	5 14	83.2098	83.2098	
2	6 14	83.3917	83.3917	
2	6 15	83.5655	83.5655	
2	7 14	83.7278	83.7278	
2	7 15	84.0318	84.0318	
2	7 15	84.3370	84.3370	
2	8 14	85.1449	85.1449	
.				
.				
.				
2	End	98.0624		
.				
.				
.				
12	Start	51.8915		Ran,Mut,Ann
12	End	93.0214		
.				
.				
.				

Design Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	98.8933	98.8933	Ini
1	Start	80.4296		Tab,Ran
1	End	98.8567		
.				
.				
.				

Design Refinement History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
0	Initial	98.9438	98.9438	Ini
1	Start	94.7490		Pre,Mut,Ann
1	End	92.1336		
.				
.				
.				

The first column, `Design`, is a design number. Each design corresponds to a complete iteration using a different initialization. Initial designs are numbered zero. The second column is `Row,Col`, which shows the design row and column that is changing in the coordinate-exchange algorithm. This column also contains `Start` for displaying the initial efficiency, `End` for displaying the final efficiency, and `Initial` for displaying the efficiency of a previously created initial design (perhaps created externally or perhaps created in a previous step). The `Current D-Efficiency` column contains the *D*-efficiency for the design including starting, intermediate and final values. The next column is `Best D-Efficiency`. Values are put in this column for initial designs and when a design is found that is as good as or better than the previous best design. The last column, `Notes`, contains assorted algorithm and explanatory details. Values are added to the table at the beginning of an iteration, at the end of an iteration, when a better design is found, and when a design first conforms to restrictions. Details of the candidate search iterations are not shown. Only the *D*-efficiency for the best design found through candidate search is shown.

Here are the notes.

`Can`	- the results of a candidate-set search
`Tab`	- tabled (orthogonal array, full, or fractional factorial) initialization (full or in part)
`Ran`	- random initialization (full or in part)
`Unb`	- unbalanced initialization (usually in part)
`Ini`	- initial design
`Mut`	- random mutations of the initial design were performed
`Ann`	- simulated annealing was used in this iteration
`Pre`	- using previous best design as a starting point
`Conforms`	- design conforms to restrictions
`Violations`	- number of restriction violations

Often, more than one note appears. For example, the triples `Ran,Mut,Ann` and `Pre,Mut,Ann` frequently appear together.

The iteration history consists of three tables.

`Algorithm Search History`	- searches for a design and the best algorithm for this problem
`Design Search History`	- uses the best algorithm to search further
`Design Refinement History`	- tries to refine the best design

%MktEx Macro Options

The following options can be used with the `%MktEx` macro.

Option	Description
`anneal=`$n1 < n2 < n3 >>$	starting probability for annealing
`annealfun=`*function*	annealing probability function
`anniter=`$n1 < n2 < n3 >>$	first annealing iteration
`balance=`n	maximum allowed level-frequency range
`big=`$n <$ *choose* $>$	size of big full-factorial design
`canditer=`$n1 < n2 >$	iterations for OPTEX designs
`detfuzz=`n	determinants change increment
`examine=`I | V	matrices that you want to examine
`exchange=`n	number of factors to exchange
`fixed=`*variable*	indicates runs that are fixed
`holdouts=`n	adds holdout observations
`imlopts=`*options*	IML PROC statement options
`init=`*SAS-data-set*	initial (input) experimental design
`interact=`*interaction-list*	interactions that must be estimable
`iter=`$n1 < n2 < n3 >>$	maximum number of iterations
`list`	list of the numbers of factor levels
`maxdesigns=`n	maximum number of designs to make
`maxiter=`$n1 < n2 < n3 >>$	maximum number of iterations
`maxstages=`n	maximum number of algorithm stages
`maxtime=`$n1 < n2 < n3 >>$	approximate maximum run time
`mutate=`$n1 < n2 < n3 >>$	mutation probability
`mutiter=`$n1 < n2 < n3 >>$	first iteration to consider mutating
`n=`n	number of runs in the design
`options=`*options-list*	binary options
`optiter=`$n1 < n2 >$	OPTEX iterations
`order=`*value*	coordinate exchange column order
`out=`*SAS-data-set*	output experimental design
`outall=`*SAS-data-set*	output data set with all designs found
`outr=`*SAS-data-set*	randomized output experimental design
`partial=`n	partial profile design
`restrictions=`*macro-name*	restrictions macro
`ridge=`n	ridging factor
`seed=`n	random number seed
`stopearly=`n	that the macro may stop early
`tabiter=`$n1 < n2 >$	tabled design iterations
`tabsize=`n	orthogonal array size
`target=`n	target efficiency criterion
`unbalanced=`$n1 < n2 >$	unbalanced factors iterations

Required Options

These options are almost always required.

list

specifies a list of the numbers of levels of all the factors. For example, for 3 two-level factors specify either `2 2 2` or `2 ** 3`. Lists of numbers, like `2 2 3 3 4 4` or a *levels**number of factors* syntax like: `2**2 3**2 4**2` can be used, or both can be combined: `2 2 3**4 5 6`. The specification `3**4` means 4 three-level factors. Note that the factor list is a positional parameter. This means that if it is specified, it must come first, and unlike all other parameters, it is not specified after a name and an equal sign. Usually, you have to specify a list. However, in some cases, you can just specify `n=` and omit the list and a default list is implied. For example, `n=18` implies a list of `2 3 ** 7`. When the list is omitted, and if there are no interactions, restrictions, or duplicate exclusions, then by default there are no OPTEX iterations (`optiter=0`).

n= *n*

specifies the number of runs in the design. You must specify `n=`. Here is an example of using the `%MktRuns` macro to get suggestions for values of `n=`:

```
%mktruns( 4 2 ** 5 3 ** 5 )
```

In this case, this macro suggests several sizes including orthogonal designs with `n=72` and `n=144` runs and some smaller nonorthogonal designs including `n=36, 24, 48, 60`.

Basic Options

This next group of options contains some of the more commonly used options.

balance= *n*

specifies the maximum allowed level-frequency range. The `balance=` option allows you to tell the macro that it should make an extra effort to ensure that the design is nearly balanced. By default, the macro does not try to ensure balance beyond the fact that lack of balance decreases *D*-efficiency. Specify a positive integer, usually 1 or 2, that specifies the degree of imbalance that is acceptable. You may need to also specify `options=accept` with `balance=`. The macro usually does a good job of producing nearly balanced design, but if balance is critically important, and your designs are not balanced enough, you can sometimes achieve better balance by specifying `balance=`, but usually at the price of worse efficiency, sometimes much worse. The `balance=` option specifies additional restrictions (see `restrictions=`) that help achieve better balance. By default, no additional restrictions are added. The `balance=`*n* option specifies that for each factor, the difference between the frequencies, for the most and least frequently occurring levels, should be no larger than *n*. You may specify `balance=0`, however this is usually not a good idea. The macro needs the flexibility to have imbalance as it refines the design. Another option is to instead use the `%MktBal` macro, which produces perfectly balanced main effects plans. It is likely that the algorithms used by both the `balance=` option and the `%MktBal` macro will be changed in the future to use some now unknown algorithms that are both faster and better.

examine= I | V

specifies the matrices that you want to examine. The option `examine=I` prints the information matrix, $\mathbf{X}'\mathbf{X}$; `examine=V` prints the variance matrix, $(\mathbf{X}'\mathbf{X})^{-1}$; and `examine=I V` prints both. By default, these matrices are not printed.

interact= *interaction-list*

specifies interactions that must be estimable. By default, no interactions are guaranteed to be estimable. Examples:

```
interact=x1*x2
interact=x1*x2 x3*x4*x5
interact=x1|x2|x3|x4|x5@2
```

The interaction syntax is like PROC GLM's and many of the other modeling procedures. It uses "`*`" for simple interactions (`x1*x2` is the interaction between `x1` and `x2`), "`|`" for main effects and interactions (`x1|x2|x3` is the same as `x1 x2 x1*x2 x3 x1*x3 x2*x3 x1*x2*x3`) and "`@`" to eliminate higher-order interactions (`x1|x2|x3@2` eliminates `x1*x2*x3` and is the same as `x1 x2 x1*x2 x3 x1*x3 x2*x3`). The specification "`@2`" allows only main effects and two-way interactions. Only "`@`" values of `2` or `3` are allowed. For the factor names, you must specify either the actual variable names (for example, `x1 x2 ...`) or you can just specify the number without the "`x`" (for example, `x1*x2` is equivalent to `1*2`). You can also specify `interact=@2` for all main effects and two-way interactions. For example, these two specifications are equivalent:

```
%mktex(2 ** 5, interact=@2, n=32)
%mktex(2 ** 5, interact=1|2|3|4|5@2, n=32)
```

options= *options-list*

specifies binary options. By default, none of these options are specified. Specify one or more of the following values after `options=`.

`accept`
allows the macro to output designs that violate restrictions imposed by `restrictions=`, `balance=`, or `partial=`, or have duplicates with `options=nodups`. Normally the macro will not output such designs. With `options=accept`, a design becomes eligible for output when the macro can no longer improve on the restrictions or eliminate duplicates. Without `options=accept`, a design is only eligible when all restrictions are met and all duplicates are eliminated.

`check`
checks the efficiency of a given design, specified in `init=`, and disables the `out=`, `outr=`, and `outall=` options. If `init=` is not specified, `options=check` is ignored.

`justinit`
specifies that the macro should stop processing as soon as it is done making the initial design, even if that design would not normally be the final design. Sometimes, this design will be an orthogonal array, but there are no guarantees. The specification `options=justinit` implies `optiter=0` and `outr=`. Specifying `options=justinit nofinal` both stops processing and prevents the final design from being evaluated. Particularly when you specify `options=nofinal`, you must ensure that this design has a suitable efficiency.

`largedesign`
allows the macro to stop after `maxtime=` minutes have elapsed in the coordinate exchange algorithm. Typically you would use this with `maxstages=1` and other options that make the algorithm run faster. By default, the macro checks time after it finishes with a design. With this option, the macro checks the time at the end of each row, after it has completed the first full pass through the design, and after any restrictions have been met, so the macro may stop before *D*-efficiency has converged. For really large problems and problems with restrictions, this option may make the macro run much faster but at a price of lower *D*-efficiency. For example, for large problems with restrictions, you might just want to try one run through the coordinate exchange algorithm with no candidate set search, orthogonal arrays, or mutations.

`nodups`
eliminates duplicate runs.

`nofinal`
skips calling PROC OPTEX to print the efficiency of the final experimental design.

`nohistory`
does not print the iteration history.

`resrep`
reports on the progress of the restrictions. You may want to specify this option with large problems with lots of restrictions or if you try to create a design and find that `%MktEx` is unable to make a design that conforms to the restrictions. By default, the iteration history is not printed for the stage where `%MktEx` is trying to make the design conform to the restrictions. Specify `options=resrep` when you want to see the progress in making the design conform.

`nosort`
does not sort the design. One use of this option is with Hadamard matrices. Hadamard matrices are generated with a banded structure that is lost when the design is sorted. If you want to see the original Hadamard matrix, and not just a design constructed from the Hadamard matrix, specify `options=nosort`.

partial= *n*

specifies a partial profile design (Chrzan and Elrod, 1995). The default is an ordinary linear design. Specify for example `partial=4` if you only want 4 attributes to vary in each row of the design (except the first run, in which none vary). This option works by adding restrictions to the design (see `restrictions=`) and specifying `order=random` and `exchange=2`. Because of the default `exchange=2` with partial profile designs, the construction is slow, so you may want to specify `maxdesigns=1` or other options to make `%MktEx` run faster. For large problems, you may get faster but less good results by specifying `order=seqran`. Specifying `options=accept` or `balance=` with `partial=` is *not* a good idea. Here is the first part of a partial profile design with twelve factors, each of which has three levels that vary and one level that means the attribute is not shown.

```
%mktex(4 ** 12, n=48, partial=4, seed=205, maxdesigns=1)
%mktlab(values=. 1 2 3, nfill=99)
options missing=' ';
proc print data=final(obs=10); run;
options missing='.';
```

Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
1												
2	1			2	1		2					
3				3	2		3					1
4			1					2	1			1
5		2	2		1							2
6				1			2				1	2
7			3		2				1			3
8				1	1		3		3			
9			2			2	2		2			
10		3				3	3				1	

restrictions= *macro-name*

specifies the name of a macro that places restrictions on the design. By default, there are no restrictions. If you have restrictions on the design, what combinations can appear with what other combinations, then you must create a macro that creates a variable called `bad` that contains a numerical summary of how bad the row of the design is. When everything is fine, set `bad` to zero. Otherwise set `bad` to a larger value that is a function of the number of restriction violations. The `bad` variable must not be binary (0 - ok, 1 - bad) unless there is only one simple restriction. You must set `bad` so that the macro knows if changes to factor levels are moving in the right direction. More complicated cases may have more complicated badness functions (see page 566). The macro must consist of PROC IML statements and possibly some macro statements.

Be sure to check the log when you specify `restrictions=`. The macro cannot always ensure that your statements are syntax-error free and stop if they are not.

Your macro can look at several things in quantifying badness, and must store its results in `bad`.

`i` - is a scalar that contains the number of the row currently being changed. If you are writing restrictions that use the variable `i`, you almost certainly should specify `options=nosort`.

`x` - is a row vector of factor levels, always containing integer values beginning with 1 and continuing on to the number of levels for each factor.

`x1` is the same as `x[1]`, `x2` is the same as `x[2]`, and so on.

`j1` - is a scalar that contains the number of the column currently being changed.

`j2` - is a scalar that contains the number of the other column currently being changed (along with `j1`) with `exchange=2` and larger `exchange=` values.

`j3` - is a scalar that contains the number of the third column currently being changed (along with `j1` and `j2`) with `exchange=3` and larger `exchange=` values.

`xmat` - is the entire `x` matrix. Note that the *ith* row of `xmat` may not be `x` since `x` will contain information on the exchanges being considered.

`bad` - results: 0 - fine, or the number of violations of restrictions.

Do not use these names (other than `bad`*) for intermediate values!*

Other than that, you can create intermediate variables without worrying about conflicts with the names in the macro. The levels of the factors for one row of the experimental design are stored in a vector `x`, and the first level is always 1, the second always 2, and so on. All restrictions must be defined in terms of `x[j]` (or alternatively, `x1`, `x2`, ..., and perhaps the other matrices). For example, if there are 5 three-level factors and if it is bad if the level of a factor equals the level for the following factor, create a macro `restrict` as follows and specify `restrictions=restrict`.

```
%macro restrict;
   bad = (x1 = x2) +
         (x2 = x3) +
         (x3 = x4) +
         (x4 = x5);
   %mend;
```

Note that you specify just the macro name and no percents on the `restrictions=` option. Also note that IML does not have the full set of Boolean operators that the DATA step and other parts of SAS have. For example, these are *not* available: `OR AND NOT GT LT GE LE EQ NE`. Here are the operators you can use along with their meaning.

=	equals	not: EQ
∧ = or ¬ =	not equals	not: NE
<	less than	not: LT
<=	less than or equal to	not: LE
>	greater than	not: GT
>=	greater than or equal to	not: GE
&	and	not: AND
|	or	not: OR
∧ or ¬	not	not: NOT

Restrictions seriously slow down the algorithm.

With restrictions, the `Current D-Efficiency` column of the iteration history table may contain values larger than the `Best D-Efficiency` column. This is because the design corresponding to the current *D*-efficiency may have restriction violations. Values are only reported in the best *D*-efficiency column after all of the restriction violations have been removed. You can specify `options=accept` with `restrictions=` when it is okay if the restrictions are not met.

See page 566 for more information on restrictions. See pages 231 and 337 for examples of restrictions.

seed= *n*

specifies the random number seed. By default, `seed=0`, and clock time is used to make the random number seed. By specifying a random number seed, results should be reproducible within a SAS release for a particular operating system. However, due to machine differences, some results may not be exactly reproducible on other machines. For most orthogonal and balanced designs, the results should be reproducible. When computerized searches are done, it is likely that you will not get the

same design across different computers, operating systems and different SAS releases, although you would expect the efficiency differences to be slight.

Data Set Options

These next options specify the names of the input and output data sets.

init= *SAS-data-set*
specifies the initial (input) experimental design. By default, there is no initial design. Use `init=` when you want to evaluate the efficiency of a design (along with `options=check`) or when you want to try to improve a design.

out= *SAS-data-set*
specifies the output experimental design. The default is `out=Design`. By default, this design is sorted unless you specify `options=nosort`. This is the output data set to look at in evaluating the design. See the `outr=` option for a randomized version of the same design, which is generally more suitable for actual use. Specify a null value for `out=` if you do not want this data set created. Often, you will want to specify a two-level name to create a permanent SAS data set so the design will be available later for analysis.

outall= *SAS-data-set*
specifies the output data set containing all designs found. By default, this data set is not created.

outr= *SAS-data-set*
specifies the randomized output experimental design. The default is `outr=Randomized`. Levels are randomly reassigned within factors, and the runs are sorted into a random order. Neither of these operations affects efficiency. When `restrictions=` or `partial=` is specified, only the random sort is performed. Specify a null value for `outr=` if you do not want a randomized design created. Often, you will want to specify a two-level name to create a permanent SAS data set so the design will be available later for analysis.

Iteration Options

These next options control some of the details of the iterations. Some of these options can take three arguments, one for each set of iterations. The macro can perform three sets of iterations. The `Algorithm Search` set of iterations looks for efficient designs using three different approaches. It then determines which approach appears to be working best and uses that approach exclusively in the second set of `Design Search` iterations. The third set or `Design Refinement` iterations tries to refine the best design found so far by using level exchanges combined with random mutations and simulated annealing.

The first set of iterations can have up to three parts. The first part uses either PROC PLAN or PROC FACTEX followed by PROC OPTEX, called through the `%MktDes` macro, to create and search a candidate set for an optimal initial design. The second part may use an orthogonal array or fractional-factorial design as an initial design. The next part consists of level exchanges starting with random initial designs.

In the first part, if the full-factorial design is manageable (arbitrarily defined as < 5185 runs), it is used as a candidate set, otherwise a fractional-factorial candidate set is used. The macro tries `optiter=` iterations to make an optimal design using the `%MktDes` macro and PROC OPTEX.

In the second part, the macro will try to generate and improve a standard orthogonal array or fractional-factorial design. Sometimes, this can lead immediately to an optimal design, for example with $2^{11}3^{12}$ and $n = 36$. In other cases, when only part of the desired design matches some standard design, only part of the design is initialized with the standard design and multiple iterations are run using the standard design as a partial initialization with the rest of the design randomly initialized.

In the third part, the macro uses the coordinate-exchange algorithm with random initial designs.

anneal= $n1 < n2 < n3 >>$

specifies the starting probability for simulated annealing in the coordinate-exchange algorithm. The default is `anneal=.05 .05 .01`. Specify a zero or null value for no annealing. You can specify more than one value if you would like to use a different value for the algorithm search, design search, and design refinement iterations. Specifying a value (greater than zero and less than one, for example 0.1) allows the design to get worse with decreasing probability as the number of iterations increases. This often helps the algorithm overcome local efficiency maxima. Allowing efficiency to decrease can help get past the bumps on the efficiency function.

Examples: `anneal=` or `anneal=0` specifies no annealing, `anneal=0.1` specifies an annealing probability of 0.1 during all three sets of iterations, `mutate=0 0.1 0.05` specifies no annealing during the initial iterations, an annealing probability of 0.1 during the search iterations, and an annealing probability of 0.05 during the refinement iterations.

anniter= $n1 < n2 < n3 >>$

specifies the first iteration to consider using annealing on the design. The default is `anniter=. . .`, which means that the macro chooses values to use. The default is the first iteration that uses a fully random initial design in each of the three sets of iterations. Hence by default, there is no random annealing in any part of the initial design when part of the initial design comes from an orthogonal design.

canditer= $n1 < n2 >$

specifies the number of coordinate-exchange iterations that will be used to try to improve a candidate-set based, OPTEX-generated initial design. The default is `canditer=1 1`. Note that `optiter=` controls the number of OPTEX iterations. Unless you are using annealing or mutation in the `canditer=` iterations (by default you are not) or unless you are using `options=nodups`, do not change theses values. The default value of `canditer=1 1`, along with the default `mutiter=` and `anniter=` values of missing, mean that the results of the OPTEX iterations are presented once in the algorithm iteration history, and if appropriate, once in the design search iteration history. Furthermore, by default, OPTEX generated designs are not improved with level exchanges except in the design refinement phase.

maxdesigns= n

specifies that the macro should stop after `maxdesigns=` designs have been created. This option may be useful for big, slow problems with restrictions. You could specify for example `maxdesigns=3` and `maxtime=0` and the macro would perform one candidate-set-based iteration, one orthogonal design initialization iteration, and one random initialization iteration and then stop. By default, this option

is ignored and stopping is based on the other iteration options. For large designs with restrictions, a typical specification is `optiter=0, tabiter=0, maxdesigns=1, options=largedesign`.

maxiter= *n1 < n2 < n3 >>*

iter= *n1 < n2 < n3 >>*

specifies the maximum number of iterations or designs to generate. The default is `maxiter=21 25 10`. With larger values, the macro tends to find better designs at a cost of slower run times. You can specify more than one value if you would like to use a different value for the algorithm search, design search, and design refinement iterations. The second value is only used if the second set of iterations consists of coordinate-exchange iterations. Otherwise, the number of iterations for the second set is specified with the `tabiter=`, or `canditer=` and `optiter=` options. If you want more iterations, be sure to set the `maxtime=` option as well, because iteration stops when the maximum number of iterations is reached or the maximum amount of time, whichever comes first. Examples: `maxiter=10` specifies 10 iterations for the initial, search, and refinement iterations, and `maxiter=10 10 5` specifies 10 initial iterations, followed by 10 search iterations, followed by 5 refinement iterations.

maxstages= *n*

specifies that the macro should stop after `maxstages=` algorithm stages have been completed. This option may be useful for big, slow problems with restrictions. You could specify `maxstages=1` and the macro will stop after the algorithm search stage, or `maxstages=2` and the macro will stop after the design search stage. The default is `maxstages=3`, which means the macro will stop after the design refinement stage.

maxtime= *n1 < n2 < n3 >>*

specifies the approximate maximum amount of time in minutes to run each phase. The default is `maxtime=10 20 5`. When an iteration completes (a design is completed), if more than the specified amount of time has elapsed, the macro quits iterating in that phase. Usually, run time will be no more than 10% or 20% larger than the specified values. However, for large problems, with restrictions, and with `exchange=` values other than 1, run time may be quite a bit larger than the specified value, since the macro only checks time after a design finishes. You can specify more than one value if you would like to use a different value for the algorithm search, design search, and design refinement iterations. By default, the macro spends up to 10 minutes on the algorithm search iterations, 20 minutes on the design search iterations, and 5 minutes in the refinement stage. Most problems run in much less time than this. Note that the second value is ignored for OPTEX iterations since OPTEX does not have any timing options. This option also affects, in the algorithm search iterations, when the macro switches between using an orthogonal initial design to using a random initial design. If the macro is not done using orthogonal initializations, and one half of the first time value has passed, it switches. Examples: `maxtime=60` specifies up to one hour for each phase. `maxtime=20 30 10` specifies 20 minutes for the first phase and 30 minutes for the second, and 10 for the third. The option `maxtime=0` provides a way to get a quick run, with no more than one iteration in each phase. However, even with `maxtime=0`, run time can be several minutes or more for large problems. See the `maxdesigns=` and `maxstages=` options for other ways to drastically cut run time for large problems.

If you specify really large time values (anything more than hours), you probably need to also specify `optiter=` since the default values depend on `maxtime=`.

mutate= $n1 < n2 < n3 >>$
specifies the probability at which each value in an initial design may mutate or be assigned a different random value before the coordinate-exchange iterations begin. The default is `mutate=.05 .05 .01`. Specify a zero or null value for no mutation. You can specify more than one value if you would like to use a different value for the algorithm search, design search, and design refinement iterations. Examples: `mutate=` or `mutate=0` specifies no random mutations. The `mutate=0.1` option specifies a mutation probability of 0.1 during all three sets of iterations. The `mutate=0 0.1 0.05` option specifies no mutations during the first iterations, a mutation probability of 0.1 during the search iterations, and a mutation probability of 0.05 during the refinement iterations.

mutiter= $n1 < n2 < n3 >>$
specifies the first iteration to consider mutating the design. The default is `mutiter=. . .`, which means that the macro chooses values to use. The default is the first iteration that uses a fully random initial design in each of the three sets of iterations. Hence by default, there are no random mutations of any part of the initial design when part of the initial design comes from an orthogonal design.

optiter= $n1 < n2 >$
specifies the number of iterations to use in the OPTEX candidate-set based searches in the algorithm and design search iterations. The default is `optiter=. .`, which means that the macro chooses values to use. When the first value is "`.`" (missing), the macro will choose a value usually no smaller than 20 for larger problems and usually no larger than 200 for smaller problems. However, `maxtime=` values other than the defaults can make the macro choose values outside this range. When the second value is missing, the macro will choose a value based on how long the first OPTEX run took and the value of `maxtime=`, but no larger than 5000. When a missing value is specified for the first `optiter=` value, the default, the macro may choose to not perform any OPTEX iterations to save time if it thinks it can find a perfect design without them.

tabiter= $n1 < n2 >$
specifies the number of times to try to improve an orthogonal or fractional-factorial initial design. The default is `tabiter=10 200`, which means 10 iterations in the algorithm search and 200 iterations in the design search.

unbalanced= $n1 < n2 >$
specifies the proportion of the `tabiter=` iterations to consider using unbalanced factors in the initial design. The default is `unbalanced=.2 .1`. One way that unbalanced factors occur is through coding down. Coding down for example creates a three-level factor from a four-level factor: (1 2 3 4) $\Rightarrow$ (1 2 3 3) or a two-level factor from a three-level factor: (1 2 3) $\Rightarrow$ (1 2 2). For any particular problem, this strategy is probably either going to work really well or not well at all, without much variability in the results, so it is not tried very often by default. This option will try to create two-level through five-level factors from three-level through six-level factors. It will not attempt for example to code down a twenty-level factor into a nineteen-level factor (although the macro is often capable of in effect doing just that through level exchanges).

Miscellaneous Options

This section contains some miscellaneous options that some users may occasionally find useful.

big= $n <$ `choose` $>$

specifies the full-factorial-design size that is considered to be big. The default is `big=5185 choose`. The default value was chosen because 5185 is approximately 5000 and greater than $2^6 3^4 = 5184$, $2^{12} = 4096$, and $2 \times 3^7 = 4374$. When the full-factorial design is smaller than the `big=` value, the `%MktEx` macro searches a full-factorial candidate set. Otherwise, it searches a fractional-factorial candidate set. When `choose` is specified as well (the default), the macro is allowed to choose to use a fractional-factorial even if the full-factorial design is not too big if it appears that the final design can be created from the fractional-factorial design. This may be useful for example when you are requesting a fractional-factorial design with interactions. Using FACTEX to create the fractional-factorial design may be a better strategy than searching a full-factorial design with PROC OPTEX.

exchange= n

specifies the number of factors to consider at a time when exchanging levels. You can specify `exchange=2` to do pair-wise exchanges. Pair-wise exchanges are *much* slower, but may produce better designs. For this reason, you may want to specify `maxtime=0` or `maxdesigns=1` or other iteration options to make fewer designs and make the macro run faster. The `exchange=` option interacts with the `order=` option. The `order=seqran` option is faster with `exchange=2` than `order=sequential` or `order=random`. The default is `exchange=2` when `partial=` is specified, otherwise, the default is `exchange=1`.

With partial-profile designs and certain other highly restricted designs, it is important to do pair-wise exchanges. Consider for example, the following design row with `partial=4`

```
1 1 2 3 1 1 1 2 1 1 1 3
```

The `%MktEx` macro cannot consider changing a 1 to a 2 or 3 unless it can also consider changing one of the current 2's or 3's to 1 to maintain the partial-profile restriction of exactly four values not equal to 1. Specifying the `exchange=2` option gives `%MktEx` that flexibility.

fixed= *variable*

specifies an `init=` data set variable that indicates which runs are fixed (cannot be changed) and which ones may be changed. By default, no runs are fixed.

1 - (or any nonmissing) means this run may never change.
0 - means this run is used in the initial design, but it may be swapped out.
. - means this run should be randomly initialized, and it may be swapped out.

This option can be used to add holdout runs to a conjoint design, but see `holdouts=` for an easier way.

holdouts= n

adds holdout observations to the `init=` data set. This option augments an initial design. Specifying `holdouts=`n optimally adds n runs to the `init=` design. The option `holdouts=`n works by adding a `fixed=` variable and extra runs to the `init=` data set. Do not specify both `fixed=` and `holdouts=`. The number of rows in the `init=` design, plus the value specified in `holdouts=` must equal the `n=`

value.

order= col=n | random | ranseq | sequential
specifies the order in which the columns are worked on in the coordinate exchange algorithm. Valid values include:

`col=`n - process n random columns in each row
`random` - random order
`ranseq` - sequential from a random first column
`seqran` - alias for `ranseq` `sequential` - 1, 2, 3, ...
null, the default - `random` when there are partial-profile restrictions, `ranseq` when there are other restrictions, and `sequential` otherwise.

For `order=col=`n, specify an integer for n, for example `order=col=2`. This option should only be used for huge problems where you do not care if you hit every column. Typically, this option will be used in conjunction with `options=largedesign, maxdesigns=1, optiter=0, tabiter=0`. You would use it when you have a large problem and you do not have enough time for one complete pass through the design. You just want to iterate for approximately the `maxtime=` amount of time then stop. You should not use `order=col=` with restrictions.

The `order=` option interacts with the `exchange=` option. With a random order and `exchange=2`, the variable `j1` loops over the columns of the design in a random order and for each `j1`, `j2` loops over the columns greater than `j1` in a random order. With a sequential order and `exchange=2`, the variable `j1` loops over the columns in 1, 2, 3 order and for each `j1`, `j2` loops over the columns greater than `j1` in a `j1`+1, `j1`+2, `j1`+3 order. The `order=ranseq` option is a bit different. With `exchange=2`, the variable `j1` loops over the columns in an order r, r+1, r+2, ..., m, 1, 2, ..., r-1 (for random r), and for each `j1` there is a single random `j2`. Hence, `order=ranseq` is the fastest option since it does not consider all pairs, just one pair. The `order=ranseq` option provides the only situation where you might try `exchange=3` or larger values.

stopearly= n
specifies that the macro may stop early when it keeps finding the same maximum D-efficiency over and over again in different designs. The default is `stopearly=5`. By default, during the design search iterations and refinement iterations, the macro will stop early if 5 times, the macro finds a D-efficiency essentially equal to the maximum but not greater than the maximum. This may mean that the macro has found the optimal design, or it may mean that the macro keeps finding a very attractive local optimum. Either way, it is unlikely it will do any better. When the macro stops for this reason, the macro will print

```
NOTE: Stopping since it appears that no improvement is possible.
```

Specify either 0 or a very large value to turn off the stop-early checking.

tabsize= n
specifies which orthogonal array (or FACTEX or Hadamard) design is used for the partial initialization when an exact match to an orthogonal design is not found. Specify the number of runs in the orthogonal design. By default, the macro chooses an orthogonal design that bests matches the specified design.

target= *n*
specifies the target efficiency criterion. The default is `target=100`. The macro stops when it finds an efficiency value greater than or equal to this number. If you know what the maximum efficiency criterion is, or you know how big is big enough, you can sometimes make the macro run faster by allowing it to stop when it reaches the specified efficiency. You can also use this option if you just want to see the initial design that `%MktEx` is using: `target=1, optiter=0`. By specifying `target=1`, the macro will stop after the initialization as long as the initial efficiency is ≥ 1.

Esoteric Options

This last set of options contains all of the other miscellaneous options. Most of the time, most users should not specify options from this list.

annealfun= *function*
specifies the function that controls how the simulated annealing probability changes with each pass through the design. The default is `annealfun=anneal # 0.85`. Note that the IML operator `#` performs ordinary (scalar) multiplication. Most users will never need this option.

detfuzz= *n*
specifies the value used to determine if determinants are changing. The default is `detfuzz=1e-8`. If `newdeter > olddeter * (1 + detfuzz)` then the new determinant is larger. Otherwise if `newdeter > olddeter * (1 - detfuzz)` then the new determinant is the same. Otherwise the new determinant is smaller. Most users will never need this option.

imlopts= *options*
specifies IML PROC statement options. For example, for very large problems, you can use this option to specify the IML `symsize=` or `worksize=` options: `imlopts=symsize=`*n* `worksize=`*m*, substituting numeric values for *n* and *m*. The defaults for these options are host dependent. Most users will never need this option.

ridge= *n*
specifies the value to add to the diagonal of $\mathbf{X}'\mathbf{X}$ to make it nonsingular. The default is `ridge=1e-7`. Usually, for normal problems, you will not need to change this value. If you want the macro to create designs with more parameters than runs, you must specify some other value, usually something like 0.01. By default, the macro will quit when there are more parameters than runs. Specifying a `ridge=` value other than the default (even if you just change the 'e' in 1e-7 to 'E') allows the macro to create a design with more parameters than runs. Most users will never need this option.

Advanced Restrictions

It is extremely important with restrictions to appropriately quantify the badness of the run. The macro has to know when it considers an exchange if it is considering

- eliminating restriction violations making the design better,
- causing more restriction violations making the design worse,

- a change that neither increases nor decreases the number of violations.

The macro must tell %MktEx when it is making progress in the right direction. If it does not, %MktEx will probably not find an acceptable design.

Complicated Restrictions

Consider designing a choice experiment with two alternatives each composed of 25 attributes, the first 22 of which will have restrictions on them. Attribute one in the choice design will be made from x1 and x23, attribute two in the choice design will be made from x2 and x24, ..., and attribute 22 in the choice design will be made from x22 and x44. The remaining attributes will be made from x45 - x50. The restrictions are as follows: each choice attribute must contain two 1's between 5 and 9 times, each choice attribute must contain exactly one 1 between 5 and 9 times, and each choice attribute must contain two 2's between 5 and 9 times. Here is an example of how *NOT* to accomplish this.

```
%macro sumres;
   allone = 0; oneone = 0; alltwo = 0;
   do k = 1 to 22;
      if      (x[k] = 1 & x[k+22] = 1) then allone = allone + 1;
      else if (x[k] = 1 | x[k+22] = 1) then oneone = oneone + 1;
      else if (x[k] = 2 & x[k+22] = 2) then alltwo = alltwo + 1;
      end;

   * Bad example.  Need to quantify badness.;
   bad = (^((5 <= allone & allone <= 9) &
            (5 <= oneone & oneone <= 9) &
            (5 <= alltwo & alltwo <= 9)));
   %mend;

%mktex(3 ** 50, n=135, optiter=0, tabiter=0, maxdesigns=1,
       restrictions=sumres, seed=289, options=resrep);
```

The problem with the preceding code is there are complicated restrictions but badness is binary. If all the counts are in the right range, badness is 0, otherwise it is 1. You need to let the macro know when it is going in the right direction or it will probably never find a suitable design. One thing that is correct about the preceding code is the compound Boolean range expressions like (5 <= allone & allone <= 9). Abbreviated expressions like (5 <= allone <= 9) that work correctly in the DATA step work incorrectly and without warning in IML. Here is a slightly better but still bad example of the macro.

```
%macro sumres;
   allone = 0; oneone = 0; alltwo = 0;
   do k = 1 to 22;
      if      (x[k] = 1 & x[k+22] = 1) then allone = allone + 1;
      else if (x[k] = 1 | x[k+22] = 1) then oneone = oneone + 1;
      else if (x[k] = 2 & x[k+22] = 2) then alltwo = alltwo + 1;
      end;
   * Better, badness is quantified, and almost correctly too!;
   bad = (^((5 <= allone & allone <= 9) &
            (5 <= oneone & oneone <= 9) &
            (5 <= alltwo & alltwo <= 9))) #
         (abs(allone - 7) + abs(oneone - 7) + abs(alltwo - 7));
   %mend;

%mktex(3 ** 50, n=135, optiter=0, tabiter=0, maxdesigns=1,
       restrictions=sumres, seed=289, options=resrep);
```

This restrictions macro seems at first glance to do everything right − it quantifies badness. We need to examine this macro more closely. It counts in `allone`, `oneone`, and `alltwo` the number of times choice attributes are all one, have exactly one 1, or are all two. Everything is fine when the all one count is in the range 5 to 9 (`5 <= allone & allone <= 9`), and the exactly one 1 count is in the range 5 to 9 (`5 <= oneone & oneone <= 9`), and the all two count is in the range 5 to 9 (`5 <= alltwo & alltwo <= 9`). It is bad when this is not true (`^((5 <= allone & allone <= 9) & (5 <= oneone & oneone <= 9) & (5 <= alltwo & alltwo <= 9))`), the Boolean not operator '^' performs the logical negation. This Boolean expression is 1 for bad and 0 for OK. It is multiplied times a quantitative sum of how far these counts are outside the right range (`abs(allone - 7) + abs(oneone - 7) + abs(alltwo - 7)`). When the run meets all restrictions, this sum of absolute differences will be multiplied by zero. Otherwise badness gets larger as the counts get farther away from the middle of the 5 to 9 interval.

In the `%MktEx` macro, we specified `options=resrep` which produces a report in the iteration history on the process of meeting the restrictions. When you run `%MktEx` and it is having trouble making a design that conforms to restrictions, this report can be extremely helpful. Next, we will examine some of the output from running the preceding macros.

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	59.7632		Ran,Mut,Ann
1	1	60.0363		0 Violations
1	2	60.3715		0 Violations
1	3	60.9507		0 Violations
1	4	61.2319		5 Violations
1	5	61.6829		0 Violations
1	6	62.1529		0 Violations
1	7	62.4004		0 Violations
1	8	62.9747		3 Violations

```
.
.
.
1    132          70.4482          6 Violations
1    133          70.3394          4 Violations
1    134          70.4054          0 Violations
1    135          70.4598          0 Violations
```

So far we have seen the results from the first pass through the design. With `options=resrep` the macro prints one line per row with the number of violations when it is done with the row. Notice that the macro is succeeding in eliminating violations in some but not all rows. This is the first thing you should look for. If it is not succeeding in any rows, you may have written a set of restrictions that is impossible to satisfy. Let's look next at some of the output from the second pass through the design.

```
1      1          70.5586          0 Violations
1      2          70.7439          0 Violations
1      3          70.7383          0 Violations
1      4          70.7429          5 Violations
1      4          70.6392          4 Violations
1      4          70.7081          4 Violations
1      4          70.7717          4 Violations
1      4          70.7717          4 Violations
1      4          70.7717          4 Violations
1      4          70.7717          4 Violations
1      4          70.7717          4 Violations
1      4          70.7202          4 Violations
1      4          70.7717          4 Violations
1      4          70.7717          4 Violations
1      4          70.7717          4 Violations
1      4          70.7264          4 Violations
1      4          70.7717          4 Violations
1      4          70.7717          4 Violations
1      4          70.7717          4 Violations
1      4          70.7717          4 Violations
1      4          70.7717          4 Violations
1      4          70.7274          4 Violations
1      4          70.7717          4 Violations
1      4          70.7515          4 Violations
1      4          70.7636          4 Violations
1      4          70.7717          4 Violations
1      4          70.7591          4 Violations
1      4          70.7717          4 Violations
1      5          70.7913          0 Violations
1      6          70.9467          0 Violations
1      7          71.0102          0 Violations
1      8          71.0660          0 Violations
```

In the second pass, in situations where the macro had some reasonable success in the first pass, %MktEx tries extra hard to impose restrictions. We see it trying over and over again without success to impose the restrictions in the fourth row. All it manages to do is lower the number of violations from 5 to 4. We also see it has no trouble removing all violations in the eighth row that were still there after the first pass. The macro produces volumes of output like this. For several iterations it will devote extra attention to rows with some violations but in this case without complete success. When you see this pattern, some success but also some stubborn rows that the macro cannot fix, there may be something wrong with your restrictions macro. Are you *really* telling %MktEx when it is doing a better job? These preceding steps illustrate some of the things that can go wrong with restrictions macros. It is important to carefully evaluate the results − look at the design, look at the iteration history, specify options=resrep, and so on to ensure your restrictions are doing what you want. The problem in this case is in the quantification of badness, which is shown again next.

```
bad = (^((5 <= allone & allone <= 9) &
         (5 <= oneone & oneone <= 9) &
         (5 <= alltwo & alltwo <= 9))) #
       (abs(allone - 7) + abs(oneone - 7) + abs(alltwo - 7));
```

For one thing, notice that we have three nonindependent contributors to the badness function, the three counts. As a level gets changed, it could increase one count and decrease another. There is a larger problem too. Say that allone and oneone are in the right range but alltwo is not. Then the function fragments abs(allone - 7) and abs(oneone - 7) incorrectly contribute to the badness function. The fix is to clearly differentiate the three sources of badness *and* weight the pieces so that one part never trades off against the other. Here is an example.

```
%macro sumres;
   allone = 0; oneone = 0; alltwo = 0;
   do k = 1 to 22;
      if      (x[k] = 1 & x[k+22] = 1) then allone = allone + 1;
      else if (x[k] = 1 | x[k+22] = 1) then oneone = oneone + 1;
      else if (x[k] = 2 & x[k+22] = 2) then alltwo = alltwo + 1;
      end;
   bad = 100 # (^(5 <= allone & allone <= 9)) # abs(allone - 7) +
          10 # (^(5 <= oneone & oneone <= 9)) # abs(oneone - 7) +
               (^(5 <= alltwo & alltwo <= 9)) # abs(alltwo - 7);
   %mend;

%mktex(3 ** 50, n=135, optiter=0, tabiter=0, maxdesigns=1,
       restrictions=sumres, seed=289, options=resrep);
```

Now a component of the badness only contributes to the function when it is really part of the problem. We gave the first part weight 100 and the second part weight 10. Now the macro will never change oneone or alltwo if that causes a problem for allone, and it will never change alltwo if that causes a problem for oneone. Previously the macro was getting stuck in some rows because it could never figure out how to fix one component of badness without making another component worse. Here is some of the output from the first pass through the design.

The SAS System

Algorithm Search History

Design	Row,Col	Current D-Efficiency	Best D-Efficiency	Notes
1	Start	59.7632		Ran,Mut,Ann
1	1	60.1415		0 Violations
1	2	60.5303		0 Violations
1	3	61.0148		0 Violations
1	4	61.4507		0 Violations
1	5	61.7717		0 Violations
1	6	62.2353		0 Violations
1	7	62.5967		0 Violations
1	8	63.1628		3 Violations
.				
.				
.				
1	126	72.3566		4 Violations
1	127	72.2597		0 Violations
1	128	72.3067		0 Violations
1	129	72.3092		0 Violations
1	130	72.0980		0 Violations
1	131	71.8163		0 Violations
1	132	71.3795		0 Violations
1	133	71.4446		0 Violations
1	134	71.2805		0 Violations
1	135	71.3253		0 Violations

We can see that in the first pass, the macro is imposing all restrictions for most but not all of the rows. Here is some of the output from the second pass.

1	1	71.3968		0 Violations
1	2	71.5017		0 Violations
1	3	71.7295		0 Violations
1	4	71.7839		0 Violations
1	5	71.8671		0 Violations
1	6	71.9544		0 Violations
1	7	72.0444		0 Violations
1	8	72.0472		0 Violations
.				
.				
.				

```
1    126          77.1597              0 Violations
1    127          77.1604              0 Violations
1    128          77.1323              0 Violations
1    129          77.1584              0 Violations
1    130          77.0708              0 Violations
1    131          77.1013              0 Violations
1    132          77.1721              0 Violations
1    133          77.1651              0 Violations
1    134          77.1651              0 Violations
1    135          77.2061              0 Violations
```

In the second pass, %MktEx has imposed all the restrictions in rows 8 and 126, the rows that still had violations after the first pass (and all of the other not shown rows too). The third pass ends like this.

```
1    126          78.7813              0 Violations
1    127   1      78.7813     78.7813  Conforms
1    127  18      78.7899     78.7899
1    127  19      78.7923     78.7923
1    127  32      78.7933     78.7933
1    127  40      78.7971     78.7971
1    127  44      78.8042     78.8042
1    127  47      78.8250     78.8250
1    127  50      78.8259     78.8259
1    127   1      78.8296     78.8296
1    127   5      78.8296     78.8296
1    127   8      78.8449     78.8449
1    127  10      78.8456     78.8456
1    128  48      78.8585     78.8585
1    128  49      78.8591     78.8591
1    128   7      78.8591     78.8591
```

The %MktEx macro completes a full pass through row 126, the place of the last violation, without finding any new violations so the macro states in row 127 that the design conforms to the restrictions and the iteration history proceeds in the normal fashion from then on (not shown). Here is the final efficiency.

The OPTEX Procedure

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	85.0645	72.2858	95.6858	0.8650

This next code creates the choice design. Notice the slightly unusual arrangement of the KEY data set

due to the fact that the first 22 attributes get made from the first 44 factors of the linear design.

```
%mktkey(x1-x50)

data key;
   input (x1-x25) ($);
   datalines;
x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13
x14 x15 x16 x17 x18 x19 x20 x21 x22                  x45 x46 x47
x23 x24 x25 x26 x27 x28 x29 x30 x31 x32 x33 x34
x35 x36 x37 x38 x39 x40 x41 x42 x43 x44              x48 x49 x50
;
%mktroll(design=design, key=key, out=chdes);

proc print; by set; id set; where set le 2 or set ge 134; run;
```

Here are a few of the choice sets.

```
                _
               A
             S l                   x x x x x x x x x x x x x x x x x
             e t x x x x x x x x x 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2
             t _ 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5

             1 1 1 1 1 1 1 2 1 3 1 1 2 3 2 3 2 3 2 2 2 1 2 2 1 1 1
               2 2 2 1 1 3 2 1 3 3 1 3 2 2 3 2 1 2 3 3 1 1 2 3 2 3

             2 1 1 1 1 1 1 3 1 1 2 2 3 2 2 2 1 2 3 1 1 3 1 2 1 1 2
               2 3 3 1 1 1 1 3 3 2 2 2 1 2 2 2 3 1 1 3 3 1 2 1 2 2

           134 1 3 3 2 3 3 2 1 1 1 1 1 1 2 2 1 3 2 2 1 3 3 1 2 1 2
               2 1 1 2 2 1 2 1 1 1 1 1 3 2 2 3 1 1 2 1 1 3 3 1 3 3

           135 1 3 3 3 1 3 1 1 1 1 2 2 3 1 2 3 3 1 3 2 1 2 1 2 3 1
               2 2 1 1 1 1 1 1 3 1 2 2 1 3 2 1 3 3 1 2 1 2 1 2 2 2
```

Where the Restrictions Macro Gets Called

There is one more aspect to restrictions that must be understood for the most sophisticated usages of restrictions. The macro that imposes the restrictions is defined and called in four distinct places in the `%MktEx` macro. First, the restrictions macro is called in a separate, preliminary IML step, just to catch some syntax errors you might have made. Next, it is called in between calling PROC PLAN or PROC FACTEX and calling PROC OPTEX. Here, the restrictions macro is used to impose restrictions on the candidate set. Next, it is used in the obvious way during design creation and the coordinate-exchange algorithm. Finally, when `options=accept` is specified, which means that restriction violations are acceptable, the macro is called after all of the iterations have completed to report on restriction violations in the final design. For some advanced restrictions, we may not want exactly the same code running in all four places. When the restrictions are purely written in terms of restrictions on `x`, which

is the *ith* row of the design matrix, there is no problem. The same macro will work fine for all uses. However, when `xmat` (the full `x` matrix) or `i` or `j1` (the row or column number) are used, the same code typically cannot be used for all applications, although sometimes it does not matter. Next are some notes on each of the four phases.

Syntax Check. In this phase, the macro is defined and called just to check for syntax errors. This step allows the macro to end more gracefully than it would otherwise if there are errors. Your restrictions macro can recognize when it is in this phase because the macro variable `&main` is set to 0 and the macro variable `&pass` is set to null. The pass variable is null before the iterations begin, 1 for the algorithm search phase, 2 for the design search phase, 3 for the design refinement stage, and 4 after the iterations end. You can conditionally execute code in this step or not using the following macro statements.

```
%if      &main eq 0 and &pass eq  %then %do;  /* execute in syntax check     */
%if not (&main eq 0 and &pass eq) %then %do;  /* not execute in syntax check */
```

You will usually not need to worry about this step. It just calls the macro once and ignores the results to check for syntax errors. For this step, `xmat` (and hence `x`) is a vector of ones (since the design does not exist yet) and `j1 = j2 = j3 = i = 1`. If you have complicated restrictions involving the row or column exchange indices (`i`, `j1`, `j2`, `j3`) you may need to worry about this step. You may need to either not execute your restrictions in this step or *conditionally* execute some assignment statements (just for this step) that set up `j1`, `j2`, and `j3` more appropriately. If you have syntax errors in your restrictions macro and you cannot figure out what they are, sometimes the best thing to do is directly submit the statements in your restrictions macro to IML so you can see the normal IML syntax errors. First submit the following statements.

```
%let n = 27; /* substitute number of runs    */
%let m = 10; /* substitute number of factors */
proc iml;
   xmat = j(&n, &m, 1);
   i = 1; j1 = 1; j2 = 1; j3 = 1; bad = 0; x = xmat[i,];
```

Candidate Check. In this phase, the macro is used to impose restrictions on the candidate set created by PROC PLAN or PROC FACTEX before it is searched by PROC OPTEX. For some problems, such as most partial profile problems, the restrictions are so severe that virtually none of the candidates will conform. Also, restrictions that are based on row number and column number do not make sense in the context of a candidate design. Your restrictions macro can recognize when it is in this phase because the macro variable `&main` is set to 0 and the macro variable `&pass` is set to 1 or 2. You can conditionally execute code in this step or not using the following macro statements.

```
%if      &main eq 0 and &pass ge 1 and &pass le 2
         %then %do;                          /* execute on candidates     */
%if not (&main eq 0 and &pass ge 1 and &pass le 2)
         %then %do;                          /* not execute on candidates */
```

For simple restrictions, not involving the column exchange indices (`j1`, `j2`, `j3`), you probably do not need to worry about this step. If you use `j1`, `j2`, or `j3`, you will need to either not execute your restrictions in this step or conditionally execute some assignment statements that set up `j1`, `j2`, and `j3` appropriately. Ordinarily for this step, `xmat` contains the candidate design, `x` contains the *ith* row, `j1 = 0; j2 = 0; j3 = 0;` and `i` is set to the candidate row number.

Main Coordinate-Exchange Algorithm. In this phase, the macro is used to impose restrictions on the design as it is being built in the coordinate-exchange algorithm. Your restrictions macro can recognize when it is in this phase because the macro variable `&main` is set to 1 and the macro variable `&pass` is set to 1, 2, or 3. You can conditionally execute code in this step or not using the following macro statements.

```
%if      &main eq 1 and &pass ge 1 and &pass le 3
         %then %do;                    /* execute on coordinate exchange     */
%if not (&main eq 1 and &pass ge 1 and &pass le 3)
         %then %do;                    /* not execute on coordinate exchange */
```

For this step, `xmat` contains the candidate design, `x` contains the *ith* row, `j1` contains the column index, `j2` and `j3` are zero (unless you are using `exchange=`, in which case `j1` and `j2` are indexes of other columns being exchanged), and `i` is the row number.

Restrictions Violations Check. In this phase, the macro is used to check the design when there are restrictions and `options=accept`. Your restrictions macro can recognize when it is in this phase because the macro variable `&main` is set to 1 and the macro variable `&pass` is greater than 3. You can conditionally execute code in this step or not using the following macro statements.

```
%if      &main eq 1 and &pass gt 3  %then %do; /* execute on final check     */
%if not (&main eq 1 and &pass gt 3) %then %do; /* not execute on final check */
```

For this step, `xmat` contains the candidate design, `x` contains the *ith* row, `j1 = 0; j2 = 0; j3 = 0;` and `i` is the row number.

Here is an example of a partial profile macro that does what the `partial=4` option does.

```
%macro partprof;
   nvary = sum(x ^= 1);
   %if &main %then %do;
      if i = 1 then bad = nvary;
      else          bad = abs(nvary - 4);
      %end;
   %else %do;
      bad = ^ (nvary = 0 | nvary = 4);
      %end;
   %mend;
```

In the main algorithm, when imposing restrictions on the design, we restrict the first run to be constant and all other runs to have four attributes varying. For the candidate-set restrictions, when MAIN is zero, any observation with zero or four varying factors is acceptable. For the candidate-set restrictions, there is no reason to count the number of violations. A candidate run is either acceptable or not. We do not worry about the syntax error or final check steps; both versions will work fine in either.

%MktKey Macro

The %MktKey macro creates expanded lists of variable names.

```
%mktkey(x1-x15)
```

The %MktKey macro produced the following line.

```
x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15
```

You can cut and paste this list to make it easier to construct the key= data set for the %MktRoll macro.

```
data key;
   input (x1-x5) ($);
   datalines;
 x1  x2  x3  x4  x5
 x6  x7  x8  x9 x10
x11 x12 x13 x14 x15
  .   .   .   .   .
;
```

%MktKey Macro Options

The only argument to the %MktKey macro is a variable list.

list

specifies a variable list. Note that the variable list is a positional parameter and it is not specified after a name and an equal sign.

%MktLab Macro

The macro %MktLab is used to process an experimental design, usually created by the %MktEx macro, and assign the final variable names and levels. There are numerous examples of its usage from pages 128 through 335. For example, say you used the %MktEx macro to create a design with 11 two-level factors (with default levels of 1 and 2).

```
%mktex(n=12, options=nosort)

proc print noobs; run;
```

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
1	1	1	1	1	1	1	1	1	1	1
1	2	2	2	1	1	2	1	1	2	2
2	2	1	1	2	1	2	2	1	2	1
1	1	1	2	1	2	2	2	2	2	1
1	2	2	2	2	1	1	2	2	1	1
1	2	1	1	2	2	2	1	2	1	2
2	2	1	2	1	2	1	2	1	1	2
2	1	1	2	2	1	1	1	2	2	2
2	1	2	2	2	2	2	1	1	1	1
1	1	2	1	2	2	1	2	1	2	2
2	2	2	1	1	2	1	1	2	2	1
2	1	2	1	1	1	2	2	2	1	2

The %MktLab macro can be used to assign levels of -1 and 1, add an intercept, and change the variable name prefixes from x to Had. This creates a Hadamard matrix (although, of course, the Hadamard matrix can have any set of variable names).

```
%mktlab(data=design, values=1 -1, int=Had0, prefix=Had);

proc print noobs; run;
```

Here is the resulting Hadamard matrix:

Had0	Had1	Had2	Had3	Had4	Had5	Had6	Had7	Had8	Had9	Had10	Had11
1	1	1	1	1	1	1	1	1	1	1	1
1	1	-1	-1	-1	1	1	-1	1	1	-1	-1
1	-1	-1	1	1	-1	1	-1	-1	1	-1	1
1	1	1	1	-1	1	-1	-1	-1	-1	-1	1
1	1	-1	-1	-1	-1	1	1	-1	-1	1	1
1	1	-1	1	1	-1	-1	-1	1	-1	1	-1
1	-1	-1	1	-1	1	-1	1	-1	1	1	-1
1	-1	1	1	-1	-1	1	1	1	-1	-1	-1
1	-1	1	-1	-1	-1	-1	-1	1	1	1	1
1	1	1	-1	1	-1	-1	1	-1	1	-1	-1
1	-1	-1	-1	1	1	-1	1	1	-1	-1	1
1	-1	1	-1	1	1	1	-1	-1	-1	1	-1

Here is an alternative way of doing the same thing using a `key=` data set.

```
data key;
  array Had[11];
  input Had1 @@;
  do i = 2 to 11; Had[i] = Had1; end;
  drop i;
  datalines;
1 -1
;
proc print data=key; run;
```

Here is the `key=` data set.

Obs	Had1	Had2	Had3	Had4	Had5	Had6	Had7	Had8	Had9	Had10	Had11
1	1	1	1	1	1	1	1	1	1	1	1
2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

```
%mktlab(data=design, key=key, int=Had0);
```

The Hadamard matrix from this step (not shown) is exactly the same as above.

The `key=` data set contains all of the variables that you want in the design and all of their levels. This information will be applied to the design, by default the one stored in a data set called RANDOMIZED, which is the default `outr=` data set name from the `%MktEx` macro. The results are stored in a new data set, FINAL, with the desired factor names and levels.

Consider the consumer food product example from Kuhfeld, Tobias, and Garratt (1994). Here is one possible design.

```
data randomized;
  input x1-x8 @@;
  datalines;
4 2 1 1 1 2 2 2 2 1 1 2 1 3 1 3 3 4 2 2 1 3 2 3 4 3 2 1 3 2 2 3 4 1 2 1
1 1 1 1 2 4 1 2 1 2 1 1 1 2 1 2 3 3 2 1 2 2 2 2 2 2 2 3 1 4 2 1 1 2 2 2
3 2 2 1 3 1 2 1 1 4 1 2 2 3 1 2 1 3 2 2 1 3 1 1 3 2 1 2 2 1 2 3 3 4 1 1
3 1 1 3 4 1 2 2 2 1 2 1 2 3 2 1 2 3 2 2 2 1 2 1 3 3 1 3 4 2 2 2 1 3 1 2
2 4 2 2 3 1 1 2 3 1 2 2 3 2 1 2 3 3 1 1 2 3 1 1 4 4 2 1 2 2 1 3 1 1 1 1
3 2 1 2 4 3 1 2 3 3 2 2 1 2 2 1 2 1 1 3 1 3 1 1 1 1 2 3
;
```

Designs created by the `%MktEx` macro always have factor names `x1`, `x2`, ..., and so on, and the levels are consecutive integers beginning with 1 (1, 2 for two-level factors; 1, 2, 3 for three-level factors; and so on). The `%MktLab` macro provides you with a convenient way to change the names and levels to more meaningful values. The data set KEY contains the variable names and levels that you ultimately want.

```
data key;
  missing N;
  input Client ClientLineExtension ClientMicro $ ShelfTalker $
        Regional Private PrivateMicro $ NationalLabel;
  format _numeric_ dollar5.2;
  datalines;
1.29 1.39 micro Yes 1.99 1.49 micro 1.99
1.69 1.89 stove No  2.49 2.29 stove 2.39
2.09 2.39 .     .   N    N    .     N
N    N    .     .   .    .    .     .
;
%mktlab(key=key);

proc sort; by shelftalker; run;

proc print; by shelftalker; run;
```

The variable `Client` with 4 levels will be made from `x1`, `ClientLineExtension` with 4 levels will be made from `x2`, `ClientMicro` with 2 levels will be made from `x3`. The `N` (for not available) is treated as a special missing value. The KEY data set has four rows because the maximum number of levels is four. Factors with fewer than four levels are filled in with ordinary missing values. The `%MktLab` macro takes the default `data=randomized` data set from `%MktEx` and uses the rules in the `key=key` data set, to create the information in the `out=final` data set, which is shown next, sorted by the shelf talker variable.

Here is some of the design:

```
------------------------------ ShelfTalker=No ------------------------------

                Client
                 Line       Client                            Private   National
Obs   Client   Extension   Micro    Regional   Private        Micro     Label

  1   $1.69      $1.39     micro     $1.99        N           micro       N
  2   $2.09        N       stove     $1.99        N           stove       N
  3   $1.69        N       micro     $1.99      $2.29         micro     $1.99
  .
  .
  .

------------------------------ ShelfTalker=Yes -----------------------------

                Client
                 Line       Client                            Private   National
Obs   Client   Extension   Micro    Regional   Private        Micro     Label

 14     N        $1.89     micro     $1.99      $2.29         stove     $2.39
 15     N        $2.39     stove       N        $2.29         stove       N
 16     N        $1.39     stove     $1.99      $1.49         micro     $1.99
  .
  .
  .
```

This macro creates the `out=` data set by repeatedly reading and rereading the `key=` data set, one datum at a time, using the information in the `data=` data set to determine which levels to read from the `key=` data set. In this example, for the first observation, `x1`=4 so the fourth value of the first `key=` variable is read, then `x2`=2 so the second value of the second `key=` variable is read, then `x3`=1 so the first value of the third `key=` variable is read, ..., then `x8`=2 so the second value of the eighth `key=` variable is read, then the first observation is output. This continues for all observations. This is why the `data=` data set must have integer values beginning with 1.

This example creates the L_{36}, renames the two-level factors `two1-two11` and assigns them values -1, 1, and renames the three-level factors `thr1-thr12` and assigns them values -1, 0, 1.

```
%mktex(n=36, seed=420)

data key;
  array x[23] two1-two11 thr1-thr12;
  input two1 thr1;
  do i =  2 to 11; x[i] = two1; end;
  do i = 13 to 23; x[i] = thr1; end;
  drop i;
  datalines;
-1 -1
 1  0
 .  1
;
%mktlab(key=key);

proc print data=key noobs; var two:; run;
proc print data=key noobs; var thr:; run;

proc print data=final(obs=5) noobs; var two:; run;
proc print data=final(obs=5) noobs; var thr:; run;
```

Here is the KEY data set.

two1	two2	two3	two4	two5	two6	two7	two8	two9	two10	two11
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	1	1	1	1	1	1	1	1	1
.	.	.	.	.	.	.	.	.	.	.

thr1	thr2	thr3	thr4	thr5	thr6	thr7	thr8	thr9	thr10	thr11	thr12
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1

Here are the first five rows of the design.

two1	two2	two3	two4	two5	two6	two7	two8	two9	two10	two11
-1	-1	1	1	1	1	1	1	1	1	-1
1	1	1	1	-1	1	1	-1	-1	1	1
-1	-1	-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	-1	-1	1	1	1	-1	-1	-1	1
-1	1	-1	-1	-1	-1	1	1	1	1	1

thr1	thr2	thr3	thr4	thr5	thr6	thr7	thr8	thr9	thr10	thr11	thr12
0	0	0	1	1	1	-1	1	1	1	-1	-1
1	-1	-1	1	0	0	-1	1	1	0	1	0
0	-1	0	0	-1	-1	-1	0	-1	0	-1	0
0	1	0	1	1	0	1	-1	-1	-1	1	0
1	-1	1	1	1	1	0	-1	-1	0	-1	1

This next step creates a design and blocks it. This example shows that it is okay if not all of the variables in the input design are used. The variables `Block`, `Run`, and `x4` are just copied from the input to the output.

```
%mktex(n=18, seed=396)

%mktblock(nblocks=2, factors=x1-x4, seed=292)

data key;
   input Brand $ Price Size;
   format price dollar5.2;
   datalines;
Acme 1.49   6
Apex 1.79   8
.    1.99  12
;

%mktlab(data=blocked, key=key)

proc print; id block run; by block; run;
```

Here are the results:

Block	Run	Brand	Price	Size	x4
1	1	Acme	$1.49	12	2
	2	Acme	$1.79	6	3
	3	Acme	$1.79	8	2
	4	Acme	$1.99	8	1
	5	Acme	$1.99	12	1
	6	Apex	$1.49	8	3
	7	Apex	$1.49	12	3
	8	Apex	$1.79	6	1
	9	Apex	$1.99	6	2

```
                    2       1     Acme     $1.49      6      1
                            2     Acme     $1.49      6      2
                            3     Acme     $1.79      8      3
                            4     Acme     $1.99     12      3
                            5     Apex     $1.49      8      1
                            6     Apex     $1.79     12      1
                            7     Apex     $1.79     12      2
                            8     Apex     $1.99      6      3
                            9     Apex     $1.99      8      2
```

This next example illustrates using the `labels=` option. This option is more typically used with `values=` input, rather than when you construct the `key=` data set yourself, but it can be used either way. This example is from a vacation choice example.

```
%mktex(3 ** 15, n=36, seed=17, maxtime=0)

%mktblock(data=randomized, nblocks=2, factors=x1-x15, seed=448)

%macro lab;
   label X1  = 'Hawaii, Accommodations'
         X2  = 'Alaska, Accommodations'
         X3  = 'Mexico, Accommodations'
         X4  = 'California, Accommodations'
         X5  = 'Maine, Accommodations'
         X6  = 'Hawaii, Scenery'
         X7  = 'Alaska, Scenery'
         X8  = 'Mexico, Scenery'
         X9  = 'California, Scenery'
         X10 = 'Maine, Scenery'
         X11 = 'Hawaii, Price'
         X12 = 'Alaska, Price'
         X13 = 'Mexico, Price'
         X14 = 'California, Price'
         X15 = 'Maine, Price';
   format x11-x15 dollar5.;
%mend;
```

```
data key;
   length x1-x5 $ 16 x6-x10 $ 8 x11-x15 8;
   input x1 & $ x6 $ x11;
   x2  = x1;    x3 = x1;    x4 = x1;    x5 = x1;
   x7  = x6;    x8 = x6;    x9 = x6;   x10 = x6;
   x12 = x11;  x13 = x11;  x14 = x11;  x15 = x11;
   datalines;
Cabin            Mountains   999
Bed & Breakfast  Lake       1249
Hotel            Beach      1499
;

%mktlab(data=blocked, key=key, labels=lab)

proc contents p; ods select position; run;
```

Here is the variable name, label, and format information.

The CONTENTS Procedure

Variables in Creation Order

#	Variable	Type	Len	Format	Label
1	x1	Char	16		Hawaii, Accommodations
2	x2	Char	16		Alaska, Accommodations
3	x3	Char	16		Mexico, Accommodations
4	x4	Char	16		California, Accommodations
5	x5	Char	16		Maine, Accommodations
6	x6	Char	8		Hawaii, Scenery
7	x7	Char	8		Alaska, Scenery
8	x8	Char	8		Mexico, Scenery
9	x9	Char	8		California, Scenery
10	x10	Char	8		Maine, Scenery
11	x11	Num	8	DOLLAR5.	Hawaii, Price
12	x12	Num	8	DOLLAR5.	Alaska, Price
13	x13	Num	8	DOLLAR5.	Mexico, Price
14	x14	Num	8	DOLLAR5.	California, Price
15	x15	Num	8	DOLLAR5.	Maine, Price
16	Block	Num	8		
17	Run	Num	8		

%MktLab Macro Options

The following options can be used with the `%MktLab` macro.

Option	Description
`cfill=`*character-string*	character fill value
`data=`*SAS-data-set*	input design data set
`dolist=`*do-list*	new values using a do-list syntax
`int=`*variable-list*	name of an intercept variable
`key=`*SAS-data-set*	key data set
`labels=`*macro-name*	macro that provides labels and formats
`nfill=`*number*	numeric fill value
`out=`*SAS-data-set*	output data set with recoded design
`prefix=`*variable-prefix*	prefix for naming variables
`statements=`*SAS-code*	add extra statements
`values=`*value-list*	the new values for all of the variables
`vars=`*variable-list*	list of variable names

cfill= *character-string*
specifies the fill value in the `key=` data set for character variables. See the `nfill=` option for more information on fill values. The default is `cfill=' '`.

data= *SAS-data-set*
specifies the input data set with the experimental design, usually created by the `%MktEx` macro. The default is `data=Randomized`. The factor levels in the `data=` data set must be consecutive integers beginning with 1.

dolist= *do-list*
specifies the new values, using a do-list syntax (`n TO m <BY p>`), for example: `dolist=1 to 10` or `dolist=0 to 9`. With asymmetric designs (not all factors have the same levels), specify the levels for the largest number of levels. For example, with two-level and three-level factors and `dolist= 0 to 2`, the two-level factors will be assigned levels 0 and 1, and the three-level factors will be assigned levels 0, 1, and 2. Do not specify both `values=` and `dolist=`. By default, when `key=`, `values=`, and `dolist=` are all not specified, the default value list comes from `dolist=1 to 100`.

int= *variable-list*
specifies the name of an intercept variable (column of ones), if you want an intercept added to the `out=` data set. You can also specify a variable list instead of a variable name if you would like to make a list of variables with values all one. This can be useful, for example, for creating flag variables for generic choice models when the design is going to be used as a candidate set for the `%ChoicEff` macro.

key= *SAS-data-set*
specifies the input data set with the key to recoding the design. When `values=` or `dolist=` is specified, this data set is made for you. By default, when `key=`, `values=`, and `dolist=` are all not specified, the default value list comes from `dolist=1 to 100`.

labels= *macro-name*
specifies the name of a macro that provides labels, formats, or other additional information to the `key=` data set. For a simple format specification, it is easier to use `statements=`. For more involved specifications, use `labels=`. Note that you specify just the macro name, no percents on the `labels=` option. Example:

```
%mktex(3 ** 4, n=18, seed=205)
%macro labs;
   label x1 = 'Sploosh' x2 = 'Plumbob'
         x3 = 'Platter' x4 = 'Moosey';
   format x1-x4 dollar5.2;
   %mend;
%mktlab(values=1.49 1.99 2.49, labels=labs)
proc print label; run;
```

Obs	Sploosh	Plumbob	Platter	Moosey
1	$2.49	$2.49	$2.49	$1.49
2	$2.49	$2.49	$1.99	$1.99
3	$1.49	$1.49	$1.49	$1.49
4	$1.99	$1.99	$2.49	$2.49
.				
.				
.				

nfill= *number*
specifies the fill value in the `key=` data set for numeric variables. For example when the maximum number of levels is three, the last value in the `key=` data set for numeric two-level factors should have a value of `nfill=`, which by default is ordinary missing. If the macro tries to access one of these values, it prints a warning. If you would like ordinary missing (.) to be a legitimate level, specify a different `nfill=` value and use it for the extra places in the `key=` data set.

out= *SAS-data-set*
specifies the output data set with the final, recoded design. The default is `out=final`. Often, you will want to specify a two-level name to create a permanent SAS data set so the design will be available later for analysis.

prefix= *variable-prefix*
specifies a prefix for naming variables when `values=` is specified. For example `prefix=Var` creates variables `Var1`, `Var2`, and so on. By default, the variables are `x1`, `x2`, ... This option is ignored when `vars=` is specified.

statements= *SAS-code*
is an alternative to `labels=` that you can use to add extra statements to the `key=` data set. For a simple format specification, it is easier to use `statements=`. For more involved specifications, use `labels=`. Example:

```
%mktex(3 ** 4, n=18, seed=205)

%mktlab(values=1.49 1.99 2.49,
        vars=Sploosh Plumbob Platter Moosey,
        statements=format Sploosh Plumbob Platter Moosey dollar5.2)

proc print; run;
```

Obs	Sploosh	Plumbob	Platter	Moosey
1	$2.49	$2.49	$2.49	$1.49
2	$2.49	$2.49	$1.99	$1.99
3	$1.49	$1.49	$1.49	$1.49
4	$1.99	$1.99	$2.49	$2.49
.				
.				
.				

values= *value-list*
specifies the new values for all of the variables. If all variables will have the same value, it is easier to specify `values=` or `dolist=` than `key=`. When you specify `values=`, the `key=` data set is created for you. Specify a list of levels separated by blanks. If your levels contain blanks, separate them with two blanks. With asymmetric designs (not all factors have the same levels) specify the levels for the largest number of levels. For example, with two-level and three-level factors and `values=a b c`, the two-level factors will be assigned levels `'a'` and `'b'`, and the three-level factors will be assigned levels `'a'`, `'b'`, and `'c'`. Do not specify both `values=` and `dolist=`. By default, when `key=`, `values=`, and `dolist=` are all not specified, the default value list comes from `dolist=1 to 100`.

vars= *variable-list*
specifies a list of variable names when `values=` or `dolist=` is specified. If `vars=` is not specified with `values=`, then `prefix=` is used.

%MktLab Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktMerge Macro

The %MktMerge autocall macro merges a data set containing an experimental design for a choice model with the data for the choice model. There are numerous examples of its usage from pages 121 through 275. Here is a typical usage of the macro.

```
%mktmerge(design=rolled, data=results, out=res2,
          nsets=18, nalts=5, setvars=choose1-choose18)
```

The design= data set comes from the %MktRoll macro. The data= data set contains the data, and the setvars= variables in the data= data set contain the numbers of the chosen alternatives for each of the 18 choice sets. The nsets= option specifies the number of choice sets, and the nalts= option specifies the number of alternatives. The out= option names the output SAS data set that contains the experimental design and a variable c that contains 1 for the chosen alternatives (first choice) and 2 for unchosen alternatives (second or subsequent choice).

When the data= data set contains a blocking variable, name it on the blocks= option. When there is blocking, it is assumed that the design= data set contains blocks of *nalts* × *nsets* observations. The blocks= variable must contain values 1, 2, ..., n for n blocks. Here is an example of using the %MktMerge macro with blocking.

```
%mktmerge(design=rolled, data=results, out=res2, blocks=form,
          nsets=18, nalts=5, setvars=choose1-choose18)
```

%MktMerge Macro Options

The following options can be used with the %MktMerge macro.

Option	Description
blocks=1 | *variable*	blocking variable
data=*SAS-data-set*	input SAS data set
design=*SAS-data-set*	input SAS choice design data set
nalts=*n*	number of alternatives
nsets=*n*	number of choice sets
out=*SAS-data-set*	output SAS data set
setvars=*variable-list*	variables with the data
statements=SAS-statements	additional statements

You must specify the design=, nalts=, nsets=, and setvars= options.

blocks= 1 | *variable*
specifies either a 1 (the default) if there is no blocking or the name of a variable in the data= data set that contains the block number. When there is blocking, it is assumed that the design= data set contains blocks of *nalts* × *nsets* observations, one set per block. The blocks= variable must contain values 1, 2, ..., n for n blocks.

data= *SAS-data-set*
specifies an input SAS data set with data for the choice model. By default, the `data=` data set is the last data set created.

design= *SAS-data-set*
specifies an input SAS data set with the choice design. This data set could have been created for example with the `%MktRoll` macro. This option must be specified.

nalts= *n*
specifies the number of alternatives. This option must be specified.

nsets= *n*
specifies the number of choice sets. This option must be specified.

out= *SAS-data-set*
specifies the output SAS data set. If `out=` is not specified, the DATAn convention is used. This data set contains the experimental design and a variable `c` that contains 1 for the chosen alternatives (first choice) and 2 for unchosen alternatives (second or subsequent choice).

setvars= *variable-list*
specifies a list of variables, one per choice set, in the `data=` data set that contains the numbers of the chosen alternatives. It is assumed that the values of these variables range from 1 to *nalts*. This option must be specified.

statements= SAS-statements
specifies additional statements like `format` and `label` statements. Example:

```
%mktmerge(design=rolled, data=results, out=res2, blocks=form,
          nsets=&n, nalts=&m, setvars=choose1-choose&n,
          statements=%str(price = input(put(price, price.), 5.);
                          format scene scene. lodge lodge.;))
```

%MktMerge Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktOrth Macro

The %MktOrth macro lists some of the 100% orthogonal main-effects plans that the %MktEx macro can generate. See page 137 for an example. The %MktOrth macro can help you see what orthogonal designs are available and decide which ones to use. Here is a typical usage.

```
%mktorth;
```

The macro creates data sets and no printed output except log notes.

```
NOTE: The data set WORK.MKTDESLEV has 12356 observations and 54 variables.
NOTE: The data set WORK.MKTDESCAT has 12356 observations and 3 variables.
```

This next step would print the entire catalog. The results of this step are not shown.

```
proc print data=mktdescat;
   id n; by n;
   run;
```

Here are the first few and last few designs in the catalog.

```
proc print data=mktdeslev(where=(n le 12 or n eq 512 and x2 le 4));
   var design reference;
   id n; by n;
   run;
```

n	Design	Reference
4	2 ** 3	Hadamard
6	2 ** 1 3 ** 1	Full-Factorial
8	2 ** 7	Hadamard
	2 ** 4 4 ** 1	Hadamard
9	3 ** 4	Fractional-Factorial
10	2 ** 1 5 ** 1	Full-Factorial
12	2 ** 11	Hadamard
	2 ** 4 3 ** 1	Orthogonal Array
	2 ** 2 6 ** 1	Orthogonal Array
	3 ** 1 4 ** 1	Full-Factorial
512	2 ** 4 4 **169	Fractional-Factorial
	2 ** 3 4 **167 8 ** 1	Fractional-Factorial
	2 ** 3 4 **159 32 ** 1	Fractional-Factorial
	4 **168 8 ** 1	Fractional-Factorial
	4 **160 32 ** 1	Fractional-Factorial

Here are the first few designs and variables in the MKTDESLEV data set.

```
proc print data=mktdeslev(where=(n le 12));
   var design reference x1-x6;
   id n; by n;
   run;
```

n	Design	Reference	x1	x2	x3	x4	x5	x6
4	2 ** 3	Hadamard	0	3	0	0	0	0
6	2 ** 1 3 ** 1	Full-Factorial	0	1	1	0	0	0
8	2 ** 7	Hadamard	0	7	0	0	0	0
	2 ** 4 4 ** 1	Hadamard	0	4	0	1	0	0
9	3 ** 4	Fractional-Factorial	0	0	4	0	0	0
10	2 ** 1 5 ** 1	Full-Factorial	0	1	0	0	1	0
12	2 ** 11	Hadamard	0	11	0	0	0	0
	2 ** 4 3 ** 1	Orthogonal Array	0	4	1	0	0	0
	2 ** 2 6 ** 1	Orthogonal Array	0	2	0	0	0	1
	3 ** 1 4 ** 1	Full-Factorial	0	0	1	1	0	0

If you just want to display a list of designs, possibly selecting on n, the number of runs, you can use the MKTDESCAT data set. However, if you would like to do more advanced processing, based on the numbers of levels of some of the factors, you can use the `outlev=mktdeslev` data set to select potential designs. You can look at the level information in MKTDESLEV and see the number of two-level factors in `x2`, the number of three-level factors in `x3`, ..., and the number of fifty-level factors is in `x50`. The number of one level factors, `x1`, is always zero, but `x1` is available so you can make arrays (for example, `array x[50]`) and have `x[2]` refer to `x2`, the number of two-level factors.

Say you are interested in the design $2^5 3^5 4^1$. Here are some of the ways in which it is available.

```
proc print data=mktdeslev;
   where x2 ge 5 and x3 ge 5 and x4 ge 1 and n le 100;
   var n design reference;
   run;
```

Obs	n	Design	Reference
347	72	2 ** 44 3 ** 12 4 ** 1	Orthogonal Array
356	72	2 ** 37 3 ** 13 4 ** 1	Orthogonal Array
362	72	2 ** 35 3 ** 12 4 ** 1 6 ** 1	Orthogonal Array
366	72	2 ** 20 3 ** 24 4 ** 1	Orthogonal Array
.			
.			
.			

Here is one way you can see all the designs in a certain range of sizes.

```
proc print; where 12 le n le 20; run;
```

Obs	n	Design	Reference
7	12	2 ** 11	Hadamard
8	12	2 ** 4 3 ** 1	Orthogonal Array
9	12	2 ** 2 6 ** 1	Orthogonal Array
10	12	3 ** 1 4 ** 1	Full-Factorial
11	14	2 ** 1 7 ** 1	Full-Factorial
12	15	3 ** 1 5 ** 1	Full-Factorial
13	16	2 ** 15	Hadamard
14	16	2 ** 12 4 ** 1	Hadamard
15	16	2 ** 9 4 ** 2	Hadamard
16	16	2 ** 8 8 ** 1	Fractional-Factorial
17	16	2 ** 6 4 ** 3	Hadamard
18	16	2 ** 3 4 ** 4	Fractional-Factorial
19	16	4 ** 5	Fractional-Factorial
20	18	2 ** 1 3 ** 7	Orthogonal Array
21	18	2 ** 1 9 ** 1	Full-Factorial
22	18	3 ** 6 6 ** 1	Orthogonal Array
23	20	2 ** 19	Hadamard
24	20	2 ** 8 5 ** 1	Orthogonal Array
25	20	2 ** 2 10 ** 1	Orthogonal Array
26	20	4 ** 1 5 ** 1	Full-Factorial

The %MktOrth macro can output the lineage of each design, which is the set of steps that the %MktEx macro uses to create it. Here is an example.

```
%mktorth(range=n=36, options=lineage)

proc print; where index(design, '2 ** 11') and index(design, '3 ** 12'); run;
```

Obs	n	Design	Reference
8	36	2 ** 11 3 ** 12	Orthogonal Array

Obs	Lineage
8	36 ** 1 : 36 ** 1 > 3 ** 12 12 ** 1 : 12 ** 1 > 2 ** 11

The design $2^{11}3^{12}$ in 36 runs starts out as a single 36-level factor, 36^1. Then 36^1 is replaced by $3^{12}12^1$. Finally, 12^1 is replaced by 2^{11} resulting in $2^{11}3^{12}$.

%MktOrth Macro Options

The following options can be used with the `%MktOrth` macro.

Option	Description
`maxn=`*n*	maximum number of runs of interest
`options=`*options-list*	binary options
`outall=`*SAS-data-set*	output data set with all designs
`outcat=`*SAS-data-set*	design catalog data set
`outlev=`*SAS-data-set*	output data set with the list of
`range=`*range-specification*	number of runs of interest

options= *options-list*

specifies binary options. By default, none of these options are specified. Specify one or more of the following values after `options=`.

`mktex`
specifies that the macro is being called from the `%MktEx` macro and just the `outlev=` data set is needed. The macro takes short cuts to make it run faster doing only what the `%MktEx` macro needs.

`mktruns`
specifies that the macro is being called from the `%MktRuns` macro and just the `outlev=` data set is needed. The macro takes short cuts to make it run faster doing only what `%MktRuns` needs.

`lineage`
outputs the design lineage. The `%MktEx` macro uses this information to construct the design.

outall= *SAS-data-set*

specifies the output data set with all designs. This data set is not created by default. This is like the `outlev=` data set, except larger. The `outall=` data set includes *all* of the `%MktEx` design catalog, including all of the smaller designs that can be trivially made from larger designs by dropping factors. For example, when the `outlev=` data set has `x2`=2 `x3`=2, then the `outall=` data set has that design and also `x1=2 x3=1`, `x1=1 x3=2`, and `x1=1 x2=1`. When you specify `outall=` you must also specify a reasonably small `range=` or `maxn=` value. Otherwise, the `outall=` specification will take a *long* time and create a *huge* data set, which will very likely be too large to store on your computer.

outcat= *SAS-data-set*

specifies the output data set with the catalog of designs that the `%MktEx` macro can create. The default is `outcat=MktDesCat`.

outlev= *SAS-data-set*

specifies the output data set with the list of designs and 50 more variables, `x1-x50`, which includes: `x2` - the number of two-level factors, `x3` - the number of three-level factors and so on. The default is `outlev=MktDesLev`.

maxn= *n*
specifies the maximum number of runs of interest. For small numbers (e.g. $n \leq 36$) this can make the macro run faster.

range= *range-specification*
specifies the number of runs of interest. Specify a range involving `n`, where `n` is the number of runs. Your range specification must be a logical expression involving `n`. Examples:

```
range=n=36
range=18 le n le 36
range=n eq 18 or n eq 36
```

%MktOrth Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktRoll Macro

The %MktRoll autocall macro is used for manipulating the experimental design for choice experiments. There are numerous examples of its usage from pages 119 through 341. The %MktRoll macro takes as input a SAS data set containing an experimental design with one row per choice set, the *linear design*, for example a design created by the %MktEx macro. This data set is specified in the design= option. This data set has one variable for each attribute of each alternative in the choice experiment. The output from this macro is an out= SAS data set is the *choice design* containing the experimental design with one row per alternative per choice set. There is one column for each different attribute. For example, in a simple branded study, design= could contain the variables x1-x5 which contain the prices of each of five alternative brands. The output data set would have one factor, Price, that contains the price of each of the five alternatives. In addition, it would have the number (or optionally the name) of each alternative.¶

The rules for determining the mapping between factors in the design= data set and the out= data set are contained in the key= data set. For example, assume that the design= data set contains the variables x1-x5 which contain the prices of each of five alternative brands: Brand A, B, C, D, and E. Here is how you would create the key= data set. The choice design has two factors, Brand and Price. Brand A price is made from x1, Brand B price is made from x2, ..., and Brand E price is made from x5.

A convenient way to get all the names in a variable list like x1-x5 is with the %MktKey macro.

```
%mktkey(x1-x5)
```

The %MktKey macro produced the following line.

```
x1 x2 x3 x4 x5
```

Here is the KEY data set.

```
data key;
   input (Brand Price) ($);
   datalines;
A x1
B x2
C x3
D x4
E x5
;
```

This data set has two variables. Brand contains the brand names, and Price contains the names of the factors that are used to make the price effects for each of the alternatives. The out= data set will contain the variables with the same names as the variables in the key= data set.

Here is how you can create the linear design with one row per choice set:

```
%mktex(3 ** 5, n=12)
```

Here is how you can create the choice design with one row per alternative per choice set:

```
%mktroll(design=randomized, key=key, out=sasuser.design, alt=brand)
```

For example, if the data set RANDOMIZED contains the row:

¶See page 87 for an illustration of linear versus choice designs.

Obs	x1	x2	x3	x4	x5
9	3	1	1	2	1

then the data set SASUSER.DESIGN contains the rows:

Obs	Set	Brand	Price
41	9	A	3
42	9	B	1
43	9	C	1
44	9	D	2
45	9	E	1

The price for Brand A is made from `x1=3`, ..., and the price for Brand E is made from `x5=1`.

Now assume that there are three alternatives, each a different brand, and each composed of four factors: `Price`, `Size`, `Color`, and `Shape`. In addition, there is a constant alternative. First, the `%MktEx` macro is used to create a design with 12 factors, one for each attribute of each alternative.

```
%mktex(2 ** 12, n=16, seed=109)
```

Next, the `key=` data set is created. It shows that there are three brands, A, B, and C, and also None.

```
data key;
   input (Brand Price Size Color Shape) ($); datalines;
         A     x1     x2   x3    x4
         B     x5     x6   x7    x8
         C     x9    x10  x11   x12
         None   .      .    .     .
;
```

Brand A is created from `Brand` = 'A', `Price` = `x1`, `Size` = `x2`, `Color` = `x3`, `Shape` = `x4`.

Brand B is created from `Brand` = 'B', `Price` = `x5`, `Size` = `x6`, `Color` = `x7`, `Shape` = `x8`.

Brand C is created from `Brand` = 'C', `Price` = `x9`, `Size` = `x10`, `Color` = `x11`, `Shape` = `x12`.

The constant alternative is created from `Brand` = 'None' and none of the attributes. The "`.`" notation is used to indicate missing values in input data sets. The actual values in the KEY data set will be blank (character missing).

Here is how you create the design with one row per alternative per choice set:

```
%mktroll(key=key, design=randomized, out=sasuser.design, alt=brand)
```

For example, if the data set RANDOMIZED contains the row:

Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
8	2	2	2	2	2	1	1	2	2	2	1	2

then the data set SASUSER.DESIGN contains the rows:

29	8	A	2	2	2	2
30	8	B	2	1	1	2
31	8	C	2	2	1	2
32	8	None	.	.	.	.

Now assume like before that there are three branded alternatives, each composed of four factors: `Price`, `Size`, `Color`, and `Shape`. In addition, there is a constant alternative. Also, there is an alternative-specific factor, `Pattern`, that only applies to Brand A and Brand C. First, the `%MktEx` macro is used to create a design with 14 factors, one for each attribute of each alternative.

```
%mktex(2 ** 14, n=16, seed=114)
```

Next, the `key=` data set is created. It shows that there are three brands, A, B, and C, plus None.

```
data key;
   input (Brand Price Size Color Shape Pattern) ($);
   datalines;
A     x1     x2   x3    x4   x13
B     x5     x6   x7    x8     .
C     x9    x10  x11   x12   x14
None   .      .    .     .     .
;
```

Brand A is created from `Brand` = 'A', `Price` = `x1`, `Size` = `x2`, `Color` = `x3`, `Shape` = `x4`, `Pattern` = `x13`.

Brand B is created from `Brand` = 'B', `Price` = `x5`, `Size` = `x6`, `Color` = `x7`, `Shape` = `x8`.

Brand C is created from `Brand` = 'C', `Price` = `x9`, `Size` = `x10`, `Color` = `x11`, `Shape` = `x12`, `Pattern` = `x14`.

The constant alternative is `Brand` = 'None' and none of the attributes.

Here is how you can create the design with one row per alternative per choice set:

```
%mktroll(key=key, design=randomized, out=sasuser.design, alt=brand)
```

For example, if the data set RANDOMIZED contains the row:

Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
8	2	1	1	2	1	2	1	2	1	1	1	2	1	2

then the data set SASUSER.DESIGN contains the rows:

Obs	Set	Brand	Price	Size	Color	Shape	Pattern
29	8	A	2	1	1	2	1
30	8	B	1	2	1	2	.
31	8	C	1	1	1	2	2
32	8	None	.	.	.	.	.

Now assume we are going to fit a model with price cross effects so we need `x1`, `x5`, and `x9` (the three price effects) available in the `out=` data set. See pages 205 and 228 for other examples of cross effects.

```
%mktroll(key=key, design=randomized, out=sasuser.design, alt=brand,
         keep=x1 x5 x9)
```

Now the data set also contains the three original price variables.

Obs	Set	Brand	Price	Size	Color	Shape	Pattern	x1	x5	x9
29	8	A	2	1	1	2	1	2	1	1
30	8	B	1	2	1	2	.	2	1	1
31	8	C	1	1	1	2	2	2	1	1
32	8	None	.	.	.	.	.	2	1	1

Every value in the `key=` data set must appear as a variable in the `design=` data set. The macro prints a warning if it encounters a variable name in the `design=` data set that does not appear as a value in the `key=` data set.

%MktRoll Macro Options

The following options can be used with the `%MktRoll` macro.

Option	Description
`alt=`*variable*	variable with name of each alternative
`design=`*SAS-data-set*	input SAS data set
`keep=`*variable-list*	factors to keep
`key=`*SAS-data-set*	key data set
`options=`*options-list*	binary options
`out=`*SAS-data-set*	output SAS data set
`set=`*variable*	choice set number variable

You must specify the `design=` and `key=` options.

alt= *variable*

specifies the variable in the `key=` data set that contains the name of each alternative. Often this will be something like `alt=Brand`. When `alt=` is not specified, the macro creates a variable `_Alt_` that contains the alternative number.

design= *SAS-data-set*
specifies an input SAS data set with one row per choice set. The `design=` option must be specified.

keep= *variable-list*
specifies factors from the `design=` data set that should also be kept in the `out=` data set. This option is useful to keep terms that will be used to create cross effects.

key= *SAS-data-set*
specifies an input SAS data set containing the rules for mapping the `design=` data set to the `out=` data set. The `key=` option must be specified.

options= *options-list*
specifies binary options. By default, none of these options are specified. Specify one or more of the following values after `options=`.

`nowarn`
do not print a warning when the `design=` data set contains variables not mentioned in the `key=` data set. Sometimes this is perfectly fine.

out= *SAS-data-set*
specifies the output SAS data set. If `out=` is not specified, the DATAn convention is used.

set= *variable*
specifies the variable in the `out=` data set that will contain the choice set number. By default, this variable is named `Set`.

%MktRoll Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%MktRuns Macro

The %MktRuns autocall macro suggests reasonable sizes for main-effects experimental designs. There are numerous examples of its usage from pages 108 through 304. The %MktRuns macro tries to find sizes in which perfect balance and orthogonality can occur, or at least sizes in which violations of orthogonality and balance are minimized. Typically, the macro takes one argument, a list of the number of levels of each factor.

For example, with 3 two-level and 4 three-level factors, specify either of the following.

```
%mktruns( 2 2 2 3 3 3 3 )

%mktruns( 2 ** 3 3 ** 4 )
```

The output from the macro in this example is:

Design Summary

Number of Levels	Frequency
2	3
3	4

Saturated = 12
Full Factorial = 648

Some Reasonable Design Sizes	Violations	Cannot Be Divided By
36 *	0	
72 *	0	
18	3	4
54	3	4
12	6	9
24	6	9
48	6	9
60	6	9
30	9	4 9
42	9	4 9

* - 100% Efficient Design can be made with the MktEx Macro.

```
 n    Design                                    Reference

36    2 ** 13  3 **  4                        Orthogonal Array
36    2 ** 11  3 ** 12                        Orthogonal Array
36    2 ** 10  3 **  8  6 **  1               Orthogonal Array
36    2 **  9  3 **  4  6 **  2               Orthogonal Array
36    2 **  4  3 ** 13                        Orthogonal Array
36    2 **  3  3 **  9  6 **  1               Orthogonal Array
72    2 ** 49  3 **  4                        Orthogonal Array
72    2 ** 47  3 ** 12                        Orthogonal Array
72    2 ** 46  3 **  8  6 **  1               Orthogonal Array
72    2 ** 45  3 **  4  6 **  2               Orthogonal Array
72    2 ** 44  3 ** 12  4 **  1               Orthogonal Array
 .
 .
 .
```

The macro reports that the saturated design has 12 runs and that 36 and 72 are optimal design sizes. The macro picks 36, because it is the smallest integer >= 12 that can be divided by 2, 3, 2×2, 2×3, and 3×3. The macro also reports 18 as a reasonable size. There are three violations with 18, because 18 cannot be divided by each of the three pairs of 2×2, so perfect orthogonality in the two-level factors will not be possible with 18 runs. Larger sizes are reported as well. The macro prints orthogonal designs that are available from the `%MktEx` macro that match your specification.

To see every size the macro considered, simply run PROC PRINT after the macro finishes. The output from this step is not shown.

```
proc print label data=nums split='-';
   id n;
   run;
```

For 2 two-level factors, 2 three-level factors, 2 four-level factors, and 2 five-level factors specify:

```
%mktruns( 2 2 3 3 4 4 5 5 )
```

Here are the results:

```
                    Design Summary

                  Number of
                  Levels       Frequency

                       2               2
                       3               2
                       4               2
                       5               2

Saturated      = 21
Full Factorial = 14,400
```

Some Reasonable Design Sizes	Violations	Cannot Be Divided By
120	3	9 16 25
180	6	8 16 25
60	7	8 9 16 25
144	15	5 10 15 20 25
48	16	5 9 10 15 20 25
72	16	5 10 15 16 20 25
80	16	3 6 9 12 15 25
96	16	5 9 10 15 20 25
160	16	3 6 9 12 15 25
192	16	5 9 10 15 20 25

Among the smaller design sizes, 60 or 48 look like good possibilities.

The macro has an optional keyword parameter: `max=`. It specifies the maximum number of sizes to try. Usually you will not need to specify the `max=` option. The smallest design that is considered is the saturated design. This next specification tries 5000 sizes (21 to 5020) and reports that a perfect design can be found with 3600 runs The `%MktEx` macro does not explicitly know how to make this design, however, it can usually find it or come extremely close with the coordinate exchange algorithm.

```
%mktruns(2 2 3 3 4 4 5 5, max=5000)
```

Design Summary

Number of Levels	Frequency
2	2
3	2
4	2
5	2

Saturated = 21
Full Factorial = 14,400

Some Reasonable Design Sizes	Violations	Cannot Be Divided By
3600	0	
720	1	25
1200	1	9
1440	1	25
1800	1	16

2160	1	25
2400	1	9
2880	1	25
4320	1	25
4800	1	9

Now consider again the problem with 3 two-level and 4 three-level factors, but this time we want to be estimable the interaction of two of the two-level factors. Now, instead of specifying `%mktruns( 2 2 2 3 3 3 3 )`, we replace two of the 2's with a 4.

```
%mktruns( 2 4 3 3 3 3 )
```

Design Summary

Number of Levels	Frequency
2	1
3	4
4	1

Saturated = 13
Full Factorial = 648

Some Reasonable Design Sizes	Violations	Cannot Be Divided By
72 *	0	
144 *	0	
36	1	8
108	1	8
18	6	4 8 12
24	6	9
48	6	9
54	6	4 8 12
90	6	4 8 12
96	6	9

* - 100% Efficient Design can be made with the MktEx Macro.

```
n     Design                                                  Reference

72    2 ** 44  3 ** 12  4 **  1                          Orthogonal Array
72    2 ** 37  3 ** 13  4 **  1                          Orthogonal Array
72    2 ** 35  3 ** 12  4 **  1  6 **  1                 Orthogonal Array
72    2 ** 20  3 ** 24  4 **  1                          Orthogonal Array
 .
 .
 .
```

Now we need 72 runs for perfect balance and orthogonality and there are six violations in 18 runs (4, 4×2, 4×3, 4×3, 4×3, and 4×3).

%MktRuns Macro Options

The following options can be used with the `%MktRuns` macro.

Option	Description
`list`	numbers of levels of all the factors
`max=`n $< m >$	maximum number of design sizes to try
`n=`n	design size to evaluate
`options=`*options-list*	binary options
`out=`*SAS-data-set*	data set with the suggested sizes

The `%MktRuns` macro has one positional parameter, `list`, and several keyword parameters.

list

specifies a list of the numbers of levels of all the factors. For example, for 3 two-level factors specify either `2 2 2` or `2 ** 3`. Lists of numbers, like `2 2 3 3 4 4` or a *levels**number of factors* syntax like: `2**2 3**2 4**2` can be used, or both can be combined: `2 2 3**4 5 6`. The specification `3**4` means 4 three-level factors. You must specify a list. Note that the factor list is a positional parameter. This means it must come first, and unlike all other parameters, it is not specified after a name and an equal sign.

max= n $< m >$

specifies the maximum number of design sizes to try. By default, `max=200 2`. The macro tries up to n sizes starting with the saturated design. The macro stops trying larger sizes when it finds a design size with zero violations that is m times as big as a previously found size with zero violations. The macro reports the best 10 sizes. For example, if the saturated design has 10 runs, and there are zero violations in 16 runs, then by default, the largest size that the macro will consider is $32 = 2 \times 16$ runs.

n= n

specifies the design size to evaluate. By default, this option is not specified, and the `max=` option specification provides a range of design sizes to evaluate.

options= *options-list*
specifies binary options. By default, none of these options are specified. Specify one the following values after `options=`.

`justparse`
is used by other `Mkt` macros to have this macro just parse the list argument and return it as a simple list of integers.

out= *SAS-data-set*
specifies the name of a SAS data set with the suggested sizes. The default is `out=nums`.

%MktRuns Macro Notes

This macro specifies `options nonotes` throughout most of its execution. If you want to see all of the notes, submit the statement `%let mktopts = notes;` before running the macro.

%PhChoice Macro

The %PhChoice autocall macro is used to customize the discrete choice output from PROC PHREG. Typically, you run the following macro once to customize the PROC PHREG output.

```
%phchoice(on)
```

The macro uses PROC TEMPLATE and ODS (Output Delivery System) to customize the output from PROC PHREG. Running this code edits the templates and stores copies in SASUSER. These changes will remain in effect until you delete them. Note that these changes assume that each effect in the choice model has a variable label associated with it so there is no need to print variable names. If you are coding with PROC TRANSREG, this will usually be the case. To return to the default output from PROC PHREG, run the following macro.

```
%phchoice(off)
```

If you ever have errors running this macro, like invalid page errors, see "Macro Errors" on page 641. The rest of this section discusses the details of what the %PhChoice macro does and why. Unless you are interested in further customization of the output, you should skip to "%PhChoice Macro Options" on page 609.

We are most interested in the Analysis of Maximum Likelihood Estimates table, which contains the parameter estimates. We can first use PROC TEMPLATE to identify the template for the parameter estimates table and then edit the template. First, let's have PROC TEMPLATE display the templates for PROC PHREG. The source stat.phreg statement specifies that we want to see PROC TEMPLATE source code for the STAT product and the PHREG procedure.

```
proc template;
   source stat.phreg;
   run;
```

If we search the results for the Analysis of Maximum Likelihood Estimates table we find the following code, which defines the Stat.Phreg.ParameterEstimates table.

```
define table Stat.Phreg.ParameterEstimates;
   notes "Parameter Estimates Table";
   dynamic Confidence NRows;
   column Variable DF Estimate StdErr StdErrRatio ChiSq ProbChiSq HazardRatio
      HRLowerCL HRUpperCL Label;
   header h1 h2;
   define h1;
      text "Analysis of Maximum Likelihood Estimates";
      space = 1;
      spill_margin;
   end;
   define h2;
      text Confidence BEST8. %nrstr("%% Hazard Ratio Confidence Limits");
      space = 0;
      end = HRUpperCL;
      start = HRLowerCL;
      spill_margin = OFF;
   end;
```

```
define Variable;
   header = "Variable";
   style = RowHeader;
   id;
end;
define DF;
   parent = Common.ParameterEstimates.DF;
end;
define Estimate;
   header = ";Parameter;Estimate;";
   format = D10.;
   parent = Common.ParameterEstimates.Estimate;
end;
define StdErr;
   header = ";Standard;Error;";
   format = D10.;
   parent = Common.ParameterEstimates.StdErr;
end;
define StdErrRatio;
   header = ";StdErr;Ratio;";
   format = 6.3;
end;
define ChiSq;
   parent = Stat.Phreg.ChiSq;
end;
define ProbChiSq;
   parent = Stat.Phreg.ProbChiSq;
end;
define HazardRatio;
   header = ";Hazard;Ratio;";
   glue = 2;
   format = 8.3;
end;
define HRLowerCL;
   glue = 2;
   format = 8.3;
   print_headers = OFF;
end;
define HRUpperCL;
   format = 8.3;
   print_headers = OFF;
end;
define Label;
   header = "Variable Label";
end;
```

```
      col_space_max = 4;
      col_space_min = 1;
      required_space = NRows;
   end;
```

It contains header, format, spacing and other information for each column in the table. Most of this need not concern us now. The template contains this `column` statement, which lists the columns of the table.

```
      column Variable DF Estimate StdErr StdErrRatio ChiSq ProbChiSq HazardRatio
         HRLowerCL HRUpperCL Label;
```

Since we will usually have a label that adequately names each parameter, we do not need the variable column. We also do not need the hazard information. If we move the label to the front of the list and drop the variable column and the hazard columns, we get this.

```
   column Label DF Estimate StdErr ChiSq ProbChiSq;
```

We use the `edit` statement to edit the template. We can also modify some headers. We specify the new `column` statement and the new headers. We can also modify the Summary table, which is `Stat.Phreg.CensoredSummary`, to use the vocabulary of choice models instead of survival analysis models. The code is grabbed from the PROC TEMPLATE step with the `source` statement. The overall header 'Summary of the Number of Event and Censored Values' is changed to 'Summary of Subjects, Sets, and Chosen and Unchosen Alternatives', 'Total' is changed to 'Number of Alternatives', 'Event' is changed to 'Chosen Alternatives', 'Censored' is changed to 'Not Chosen', and 'Percent Censored' is dropped. Finally `Style=RowHeader` was specified on the label column. This sets the color, font, and general style for HTML output. The `RowHeader` style is typically used on first columns that provide names or labels for the rows. Here is the code that the `%phchoice(on)` macro runs.

```
   proc template;
      edit stat.phreg.ParameterEstimates;
         column Label DF Estimate StdErr ChiSq ProbChiSq;
         header h1;
         define h1;
            text "Multinomial Logit Parameter Estimates";
            space = 1;
            spill_margin;
            end;
         define Label;
            header = " " style = RowHeader;
            end;
         end;
         edit Stat.Phreg.CensoredSummary;
            column Stratum Pattern Freq GenericStrVar Total
                   Event Censored;
            header h1;
            define h1;
               text "Summary of Subjects, Sets, "
                    "and Chosen and Unchosen Alternatives";
               space = 1;
               spill_margin;
               first_panel;
            end;
```

```
      define Freq;
        header=";Number of;Choices" format=6.0;
      end;
   define Total;
      header = ";Number of;Alternatives";
      format_ndec = ndec;
      format_width = 8;
   end;
   define Event;
      header = ";Chosen;Alternatives";
      format_ndec = ndec;
      format_width = 8;
   end;
   define Censored;
      header = "Not Chosen";
      format_ndec = ndec;
      format_width = 8;
   end;
   end;

run;
```

Here is the code that `%phchoice(off)` runs.

```
* Delete edited templates, restore original templates;
proc template;
   delete Stat.Phreg.ParameterEstimates;
   delete Stat.Phreg.CensoredSummary;
   run;
```

Our editing of the multinomial logit parameter estimates table assumes that each independent variable has a label. If you are coding with PROC TRANSREG, this will be true of all variables created by `class` expansions. You may have to provide labels for `identity` and other variables. Alternatively, if you want variable names to appear in the table, you can do that as follows. This may be useful when you are not coding with PROC TRANSREG.

```
%phchoice(on, Variable DF Estimate StdErr ChiSq ProbChiSq Label)
```

The optional second argument provides a list of the column names to print. The available columns are: `Variable DF Estimate StdErr StdErrRatio ChiSq ProbChiSq HazardRatio HRLowerCL HRUpperCL Label`. (`HRLowerCL` and `HRUpperCL` are confidence limits on the hazard ratio.) For very detailed customizations, you may have to run PROC TEMPLATE directly.

%PhChoice Macro Options

The `%PhChoice` macro has two positional parameters, `onoff` and `column`. Positional parameters must come first, and unlike all other parameters, are not specified after a name and an equal sign.

onoff

`ON` specifies choice model customization.

`OFF` turns off the choice model customization and returns to the default PROC PHREG templates.

`EXPB` turns on choice model customization and adds the hazard ratio to the output.

Upper/lower case does not matter.

column

specifies an optional column list for more extensive customizations.

%PlotIt Macro

The %PlotIt macro is used to make graphical scatter plots of labeled points. It is particularly designed to display raw data and results from analyses such as regression, correspondence analysis, MDPREF, PREFMAP, and MDS. However, it can make many other types of graphical displays as well. It can plot points, labeled points, vectors, circles and density. See pages 15–26 and 661–687 for example plots and more on these methods.

By default, the %PlotIt macro creates a graphical scatter plot on your screen. The macro will by default use the last data set created, so you must specify data= if you run %PlotIt a second time. The macro creates an output Annotate data set that cannot be used as input to the macro. If no graphics device has been previously specified (either directly or indirectly), you will be prompted for a device as follows:

```
No device name has been given--please enter device name:
```

Enter your graphics device. This name will be remembered for the duration of your SAS session or until you change the device. You can modify the gopprint= and gopplot= options to set default devices so that you will not be prompted. Note that all graphics options specified on a goptions statement (except device=) are ignored by default. Use the macro options gopprint=, gopplot=, gopts2=, and gopts= to set goptions.

To display a plot on your screen using the default goptions, specify:

```
%plotit(data=coor, datatype=corresp)
```

To create a color postscript file named myplot.ps, suitable for printing on a cljps device, specify:

```
%plotit(data=coor, datatype=corresp,
        method=print, post=myplot.ps, gopts=device=cljps)
```

Alternatively, change the default for gopprint= below to name your typical device, for example from

```
gopprint=gsfmode=replace gaccess=gsasfile,
```

to

```
gopprint=gsfmode=replace gaccess=gsasfile device=cljps,
```

Then to create a postscript file, specify:

```
%plotit(data=coor, datatype=corresp, method=print, post=myplot.ps)
```

Then the file may be previewed and printed. Another alternative is to send the plot directly to the printer. In this example, chpljr51 is a printer that prints on ordinary 8.5 × 11 paper.

```
%plotit(data=coor, datatype=corresp, gopts=device=chpljr51 cback=white)
```

To just see the printer plot, specify method=plot. Use gout= to write the plot to a catalogue.

If you do not like the default background color, specify gopplot=cback=*some-color* (substituting your favorite color for *some-color*) or gopplot= to use the default background. Similarly, you can change the color defaults color=cyan, and colors=blue red green cyan magenta orange gold lilac olive purple brown gray rose violet salmon yellow, as you see fit. The default color, color=cyan, was chosen because it would show up on both a dark background and a white background, not because it is ever the color of choice. You may want to change the default color to black for printed plots and white otherwise or to some other more-suitable colors. You can permanently change the defaults by

modifying a local copy of the macro.

Sample Usage

This example performs a simple correspondence analysis. For many plots, you only the need to specify the `data=` and `datatype=` options.

```
*------Simple Correspondence Analysis------;
proc corresp all data=cars outc=coor;
   tables marital, origin;
   title 'Simple Correspondence Analysis';
   run;

%plotit(data=coor, datatype=corresp)
```

This next example performs multiple correspondence analysis.

```
*------Multiple Correspondence Analysis------;
proc corresp mca observed data=cars outc=coor;
   tables origin size type income home marital sex;
   title 'Multiple Correspondence Analysis';
   run;

%plotit(data=coor, datatype=mca)
```

This next example performs multidimensional preference analysis. The vector lengths are increased by a factor of 2.5 to make a better graphical display.

```
*------MDPREF------;
proc prinqual data=carpref out=results n=2
              replace standard scores correlations;
   id model mpg reliable ride;
   transform ide(judge1-judge25);
   title 'Multidimensional Preference (MDPREF) Analysis';
   run;

%plotit(data=results, datatype=mdpref 2.5)
```

This next example performs a preference mapping, vector model. Again, the vector lengths are increased by a factor of 2.5 to make a better graphical display.

```
*------PREFMAP, Vector Model------;
proc transreg data=results(where=(_type_ = 'SCORE'));
   model ide(mpg reliable ride)=identity(prin1 prin2);
   output tstandard=center coefficients replace out=tresult1;
   id model;
   title 'Preference Mapping (PREFMAP) Analysis - Vector';
   run;

%plotit(data=tresult1,datatype=vector 2.5)
```

This next example performs a preference mapping, ideal point model. The `antiidea=1` option is specified to handle anti-ideal points when large data values are positive or ideal.

```
*------PREFMAP, Ideal Point------;
proc transreg data=results(where=(_type_ = 'SCORE'));
   model identity(mpg reliable ride)=point(prin1 prin2);
   output tstandard=center coordinates replace out=tresult1;
   id model;
   title 'Preference Mapping (PREFMAP) Analysis - Ideal';
   run;

%plotit(data=tresult1,datatype=ideal,antiidea=1)
```

This next example performs multidimensional preference analysis. The `mdpref2` specification means MDPREF and label the vectors *too*. The vector lengths are increased by a factor of 3 to make a better graphical display. The `symlen=2` option specifies two-character symbols. The specification `vechead=`, (a null value) means no vector heads since there are labels. The `adjust1=` option is used to add full SAS DATA step statements to the preprocessed data set. This example processes `_type_ = 'CORR'` observations (those that contain vector the coordinates) the original variable names (`sub1`, `sub2`, `sub3`, ..., from the `activity` variable) and creates symbol values (1, 2, 3, ...) of size 0.7. The result is a plot with each vector labeled with a subject number.

```
*------MDPREF, labeled vector end points------;
proc prinqual cor data=recreate out=rec score std rep;
   transform identity(sub:);
   id activity active relaxing spectato;
   title 'MDPREF of Recreational Activities';
   run;

%plotit(data=rec,datatype=mdpref2 3,
        symlen=2,vechead=,adjust1=%str(
        if _type_ = 'CORR' then do;
           __symbol = substr(activity,4);
           __ssize  = 0.7;
           activity = ' ';
           end;))
```

This next example creates a contour plot, displaying density with color. The `paint=z black blue magenta red` option specifies that color interpolation is based on the variable `z`, going from black (zero density) through blue, magenta, and to red (maximum density). Observations with a computed color of black (`CX000000`) are excluded for efficiency so PROC GANNO has fewer observations to process. This color list is designed for `cback=black`.

```
*------Bivariate Normal Density Function------;
proc iml;
   title 'Bivariate Normal Density Function';
   s = inv({1 0.25 , 0.25 1});
   m = -2.5; n = 2.5; inc = 0.05; k = 0;
   x = j((1 + (n - m) / inc) ** 2, 3, 0);
   c = sqrt(det(s)) / (2 * 3.1415);
   do i = m to n by inc;
      do j = m to n by inc;
         v = i || j; z = c * exp(-0.5 * v * s * v`);
         k = k + 1; x[k,] = v || z;
         end;
      end;
   create x from x[colname={'x' 'y' 'z'}]; append from x;
   quit;

%plotit(datatype=contour, data=x, excolors=CX000000,
         paint=z black blue magenta red, gopts=cback=black)
```

The goal of this next example is to create a plot of the IRIS data set with each observation identified by its species. Species name is centered at each point's location, and each species name is plotted in a different color. This scatter plot is overlaid on the densities used by PROC DISCRIM to classify the observations. There are three densities, one for each species. Density is portrayed by a color contour plot with black (the assumed background color) indicating a density of essentially zero. Yellow, orange, and red indicate successively increasing density.

The `data=` option names the input SAS data set. The `plotvars=` option names the y-axis and x-axis variables. The `labelvar=_blank_` option specifies that all labels are blank. This example does not use any of PROC PLOT's label collision avoidance code. It simply uses PROC PLOT to figure out how big to make the plot, and then the macro puts everything inside the plot independently of PROC PLOT, so the printer plot is blank. The `symlen=10` option specifies that the maximum length of a symbol value is 10 characters. This is because the full species names are used as symbols. The `exttypes=symbol contour` option explicitly specifies that PROC PLOT will know nothing about the symbols or the contours. They are external types that will be added to the graphical plot by the macro after PROC PLOT has finished. The `ls=100` option specifies a constant line size. Since no label avoidance is done, there can be no collisions, and the macro will not iteratively determine the plot size. The default line size of 65 is too small for this example, whereas `ls=100` makes a better display. The `paint=` option specifies that based on values the variable `density`, colors should be interpolated ranging from black (minimum `density`) to yellow to orange to red (maximum `density`). The `rgbtypes=contour` option specifies that the `paint=` option should apply to contour type observations.

The grid (created with the loops: `do sepallen = 30 to 90 by 0.6;` and `do petallen = 0 to 80 by 0.6;`) is not square, so for optimal results the macro must be told the number of horizontal and vertical positions. The PLOTDATA DATA step creates these values and stores them in macro variables `&hnobs` and `&vnobs`, so the specification `hnobs=&hnobs, vnobs=&vnobs`, specifies the grid size. Of course these values could have been specified directly instead of through symbolic variables. The `excolors=CX000000` option is included for efficiency. The input data consist of a large grid for the contour plot. Most of the densities are essentially zero, so many of the colors will be `CX000000`, which is black, computed by `paint=`, which is the same color as the background. Excluding them from processing makes the macro run faster and creates smaller datasets.

This example shows how to manually do the kinds of things that the datatype= option does for you with standard types of data sets. The macro expects the data set to contain observations of one or more types. Each type is designated by a different value in a variable, usually named _type_. In this example, there are four types of observations, designated by the _type_ variable's four values, 1, 2, 3, 4, which are specified in the types= option. The symtype= option specifies the symbol types for these four observation types. The first three types of observations are symbol and the last type, _type_ = 4, designates the contour observations. The first three symbols are the species names (symbols= values) printed in symfont=swiss font. The last symbol is null because contours do not use symbols. The first three symbols, since they are words as opposed to a single character, are given a small size (symsize=0.7). A value of 1 is specified for the symbol size for contour type observations. The macro determines the optimal size for each color rectangle of the contour plot. Constant colors are only specified for the noncontour observations since a variable color is computed for contour observations.

```
*------Discriminant Analysis------;
data plotdata;  * Create a grid over which DISCRIM outputs densities.;
   do sepallen = 30 to 90 by 0.6;
      h + 1; * Number of horizontal cells;
      do petallen = 0 to 80 by 0.6;
         n + 1; * Total number of cells;
         output;
         end;
      end;
   call symput('hnobs', compress(put(h    , best12.))); * H grid size;
   call symput('vnobs', compress(put(n / h, best12.))); * V grid size;
   drop n h;
   run;
proc discrim data=iris testdata=plotdata testoutd=plotd
             method=normal pool=no short noclassify;
   class species;
   var petallen sepallen;
   title 'Discriminant Analysis of Fisher (1936) Iris Data';
   title2 'Using Normal Density Estimates with POOL=NO';
   run;
data all;
   * Set the density observations first so the scatter plot points
     will be on top of the contour plot.  Otherwise the contour plot
     points will hide the scatter plot points.;
   set plotd iris(in=iris);
   if iris then do;
      _type_ = s; * unformatted species number 1, 2, 3;
      output;
      end;
   else do;
      _type_ = 4; * density observations;
      density = max(setosa,versicol,virginic);
      output;
      end;
   run;
```

```
%plotit(data=all,plotvars=petallen sepallen,labelvar=_blank_,
        symlen=10,exttypes=symbol contour,ls=100,
        paint=density black yellow orange red,rgbtypes=contour,
        hnobs=&hnobs,vnobs=&vnobs,excolors=CX000000,
        types  =1      2          3          4,
        symtype=symbol symbol     symbol     contour,
        symbols=Setosa Versicolor Virginica  '',
        symsize=0.7    0.7        0.7        1,
        symfont=swiss  swiss      swiss      solid,
        colors =blue   magenta    cyan
        )
```

How %PlotIt Works

You create a data set either with a DATA step or with a procedure. Then you run the macro to create a graphical scatter plot. This macro is not a SAS/GRAPH procedure and does not behave like a typical SAS/GRAPH procedure. The `%PlotIt` macro performs the following steps.

1. The `%PlotIt` macro reads an input data set and preprocesses it. The preprocessed data set contains information such as the axis variables, the point-symbol and point-label variables, and symbol and label types, sizes, fonts, and colors. The nature of the preprocessing depends on the type of data analysis that generated the input data set. For example, if the option `datatype=mdpref` was specified with an input data set created by PROC PRINQUAL for a multidimensional preference analysis, then the `%PlotIt` macro creates blue points for `_type_ = 'SCORE'` observations and red vectors for `_type_ = 'CORR'` observations.

2. A DATA step, using the DATA Step Graphics Interface, determines how big to make the graphical plot.

3. PROC PLOT determines where to position the point labels. By default, if some of the point label characters are hidden, the `%PlotIt` macro recreates the printer plot with a larger line and page size, and hence creates more cells and more room for the labels. Note that when there are no point labels, the printer plot may be empty. All of the information that is in the graphical scatter plot may be stored in the `extraobs=` data set. All results from PROC PLOT are written to data sets with ODS. The macro will clear existing `ods select` and `ods exclude` statements.

4. The printer plot is read and information from it, the preprocessed data set, and the extra observations data set are combined to create an Annotate data set. The label position information is read from the PROC PLOT output, and all of the symbol, size, font, and color information is extracted from the preprocessed (or extra observations) data set. The Annotate data set contains all of the instructions for drawing the axes, ticks, tick marks, titles, point symbols, point labels, axis labels, and so on. Circles can be drawn around certain locations, and vectors can be drawn from the origin to other locations.

5. The Annotate data set is displayed with the GANNO procedure. The `%PlotIt` macro does not use PROC GPLOT.

Debugging

When you have problems, try `debug=vars` to see what the macro thinks you specified. It is also helpful to specify: `debug=mprint notes`. You can also print the final Annotate data set and the preprocessing data set:

```
options ls=180;
proc print data=anno uniform;
   format text $20. comment $40.;
   run;

proc print data=preproc uniform;
   run;
```

Advanced Processing

You can post-process the Annotate DATA step to change colors, fonts, undesirable placements, and so on. Sometimes, this can be done with the `adjust4=` option. Alternatively, when you specify `method=none`, you create an Annotate data set without displaying it. The data set name is by default WORK.ANNO. You can then manipulate it further with a DATA step or PROC FSEDIT to change colors, fonts, or sizes for some labels; move some labels; and so on. If the final result is a new data set called ANNO2, display it by running:

```
proc ganno annotate=anno2;
   run;
```

Notes

With `method=print`, the macro creates a file. See the `filepref=` and `post=` options and make sure that the file name does not conflict with existing names.

This macro creates variable names that begin with two underscores and assumes that these names will not conflict with any input data set variable names.

It is not feasible with a macro to provide the full range of error checking that is provided with a procedure. Extensive error checking is provided, but not all errors will be diagnosed.

Not all options will work with all other options. Some combinations of options may produce macro errors or Annotate errors.

This macro may not be fully portable. When you switch operating systems or graphics devices, some changes may be necessary to get the macro to run successfully again.

Graphics device differences may also be a problem. We do not know of any portability problems, but the macro has not been tested on all supported devices.

This macro tries to create a plot with equated axes, where a centimeter on one axis represents the same data range as a centimeter on the other axis. The only way to accomplish this is by explicitly and jointly controlling the `hsize=`, `vsize=`, `hpos=`, and `vpos= goptions`. By default, the macro tries to ensure that all of the values work for the specific device. See `makefit=`, `xmax=`, and `ymax=`. By default the macro uses GASK to determine `xmax` and `ymax`. If you change any of these options, your axes may not be equated. Axes are equated when `vsize` × `hpos` / `hsize` × `hpos` = `vtoh`.

When you are plotting variables that have very different scales, you may need to specify appropriate tick increments for both axes to get a reasonable plot. Here is an example: `plotopts=haxis=by 20 vaxis=by 5000`. Alternatively, just specifying the smaller increment is often sufficient: `plotopts=haxis=by 20`. Alternatively, specify `vtoh=`, (null value) to get a plot like PROC GPLOT's, with the window filled.

By default, the macro iteratively creates and recreates the plot, increasing the line size and the flexibility in the `placement=` list until there are no penalties.

The SAS system option `ovp` (overprint) is not supported by this macro.

%PlotIt Macro Options

The following options can be used with the `%PlotIt` macro.

Option	Description
`adjust1=`*SAS-statements*	adjust the preprocessing data set
`adjust2=`*SAS-statements*	includes statements with PROC PLOT
`adjust3=`*SAS-statements*	extra statements for the final DATA step
`adjust4=`*SAS-statements*	extra statements for the final DATA step
`adjust5=`*SAS-statements*	extra statements for the final DATA step
`antiidea=`*n*	eliminates PREFMAP anti-ideal points
`blue=`*expression*	blue part of RGB colors
`bright=`*n*	generates random label colors
`britypes=`*type*	types to which `bright=` applies
`cframe=`*color*	color of background within the frame
`cirsegs=`*n*	circle smoothness parameter
`color=`*color*	default color
`colors=`*colors-list*	default label and symbol color list
`cursegs=`*n*	number of segments in a curve
`curvecol=`*color*	color of curve
`data=`*SAS-data-set*	input data set
`datatype=`*data-type*	data analysis that generated data set
`debug=`*values*	debugging output
`excolors=`*color-list*	excludes from the Annotate data set
`extend=`*axis-extensions*	extend the x and y axes
`extraobs=`*SAS-data-set*	extra observations data set
`exttypes=`*type*	types for `extraobs=` data set
`filepref=`*prefix*	file name prefix
`font=`*font*	default font
`framecol=`*color*	color of frame
`gdesc=`*description*	catalog description
`gname=`*name*	catalog entry
`gopplot=`*goptions*	`goptions` for plotting to screen
`gopprint=`*goptions*	`goptions` for printing
`gopts2=`*goptions*	`goptions` that are always used
`gopts=`*goptions*	additional `goptions`
`gout=`*catalog*	`proc anno gout=` catalog

`green=`*expression*	green part of RGB colors
`hminor=`*n* | *do-list*	horizontal axis minor tick marks
`hnobs=`*n*	horizontal observations for contour plots
`hpos=`*n*	horizontal positions in graphics area
`href=`*do-list*	horizontal reference lines
`hsize=`*n*	horizontal graphics area size
`inc=`*n*	`haxis=by inc`, `vaxis=by inc`
`interpol=`*method*	axis interpolation method
`labcol=`*label-colors*	colors for the point labels
`label=`*label-statement*	`label` statement
`labelcol=`*color*	color of variable labels
`labelvar=`*label-variable*	point label variable
`labfont=`*label-fonts*	fonts for the point labels
`labsize=`*label-sizes*	sizes for the point labels
`ls=`*n*	how line sizes are generated
`lsinc=`*n*	increment to line size
`lsizes=`*number-list*	line sizes (thicknesses)
`makefit=`*n*	proportion of graphics window to use
`maxiter=`*n*	maximum number of iterations
`maxokpen=`*n*	maximum acceptable penalty sum
`method=`*value*	where to send the plot
`monochro=`*color*	overrides all other colors
`nknots=`*n*	number of knots option
`offset=`*n*	move symbols for coincident points
`options=`*options-list*	binary options
`out=`*SAS-data-set*	output Annotate data set
`outward=`none | *'c'*	PLOT statement `outward=`
`paint=`*interpolation*	color interpolation
`place=`*placement*	generates a `placement=` option
`plotopts=`*options*	PLOT statement options
`plotvars=`*variable-list*	*y*-axis and *x*-axis variables
`post=`*filename*	graphics stream file name
`preproc=`*SAS-data-set*	preprocessed `data=` data set
`procopts=`*options*	PROC PLOT statement options
`ps=`*n*	page size
`radii=`*do-list*	radii of circles
`red=`*expression*	red part of RGB colors
`regdat=`*SAS-data-set*	intermediate regression results data set
`regopts=`*options*	regression curve fitting options
`regprint=`noprint	regression options
`rgbround=`*RGB-rounding*	`paint=` rounding factors
`rgbtypes=`*type*	types to which RGB options apply
`symbols=`*symbol-list*	plotting symbols
`symcol=`*symbol-colors*	colors of the symbols
`symfont=`*symbol fonts*	symbol fonts
`symlen=`*n*	length of the symbols
`symsize=`*symbol-sizes*	sizes of symbols
`symtype=`*symbol-types*	types of symbols
`symvar=`*symbol-variable*	plotting symbol variable

`tempdat1=`*SAS-data-set*	intermediate results data set
`tempdat2=`*SAS-data-set*	intermediate results data set
`tempdat3=`*SAS-data-set*	intermediate results data set
`tempdat4=`*SAS-data-set*	intermediate results data set
`tempdat5=`*SAS-data-set*	intermediate results data set
`tempdat6=`*SAS-data-set*	intermediate results data set
`tickaxes=`*axis-string*	axes to draw tick marks
`tickcol=`*color*	color of ticks
`tickfor=`*format*	tick format used by `interpol=tick`
`ticklen=`*n*	length of tick mark in horizontal cells
`titlecol=`*color*	color of title
`tsize=`*n*	default text size
`types=`*observation-types*	observations types
`typevar=`*variable*	observation types variable
`unit=`in | cm	`hsize=` and `vsize=` unit
`vechead=`*vector-head-size*	how to draw vector heads
`vminor=`*n* | *do-list*	vertical axis minor tick marks
`vnobs=`*n*	vertical observations for contour plots
`vpos=`*n*	vertical positions in graphics area
`vref=`*do-list*	vertical reference lines
`vsize=`*n*	vertical graphics area size
`vtoh=`*n*	PROC PLOT `vtoh=` option
`xmax=`*n*	maximum horizontal graphics area size
`ymax=`*n*	maximum vertical graphics area size

Note that for many analyses, the only options you need to specify are `data=`, `datatype=`, and sometimes `method=`. To specify variables to plot, specify `plotvars=`, `labelvar=`, and `symvar=`.

Overriding Options

This macro looks for a special global macro variable named `plotitop`. If it exists, its values are used to override the macro options. Say you have a series of calls to the plotting macro and you want to route them all to a postscript file, you can specify this once:

```
%let plotitop = gopts=gsfmode=append gaccess=gsasfile device=qmscolor;
```

and then run the macro repeatedly without change. The value of the macro variable must begin with a name, followed by an equal sign, followed by a value. Optionally, it may continue with a comma, followed by a `name=`*value*, and so on. Option values must not contain commas. Here is another example:

```
%let plotitop = color=black, gopts=cback=cyan;
```

Destination and GOPTIONS

The options in this section specify the plot destination and SAS `goptions`. Note that with the `%PlotIt` macro, you do not specify a `goptions` statement. If you do, it will be overridden. All `goptions` (except `device=`) are specified with macro options. If you would prefer to specify your own `goptions` statement and have the macro use it, just specify or change the default for these four options to null: `gopplot=, gopprint=, gopts2=, gopts=`. If you use a locally installed copy of the macro, you can modify the `gopprint=` and `gopplot=` options defaults to include the devices that you typically use. Otherwise,

the macro checks the `goptions` to get a device.

gopplot= *goptions*
specifies the `goptions` for directly plotting on the screen. There are no default goptions for `gopplot=`.

gopprint= *goptions*
specifies the `goptions` for printing (creating a graphics stream file). The default is `gopprint=gsfmode=replace gaccess=gsasfile`.

Here is an example of how you might modify the defaults for `gopprint=` and `gopplot=` option defaults to set default devices.

```
gopprint=gsfmode=replace gaccess=gsasfile device=qmscolor,
gopplot=cback=black device=win,
```

gopts= *goptions*
provides a way to specify additional `goptions` that are always used. There are no default goptions for `gopts=`. For example, to rotate to a landscape orientation with a black background color, specify `gopts=rotate cback=black`.

gopts2= *goptions*
specifies the `goptions` that are always used, no matter which `method=` is specified. The default is `gopts2=reset=goptions erase`.

method= gplot | plot | print | none
specifies where to send the plot. The default is `method=gplot`.

`gplot` - displays a graphical scatter plot on your screen using the `goptions` from `gopplot=`. The `gopplot=` option should contain the `goptions` that only apply to plots displayed on the screen.

`plot` - creates a printer plot only.

`print` - routes the plot to a graphics stream file, such as a postscript file, using the `goptions` from `gopprint=`. The `gopprint=` option should contain the `goptions` that only apply to hard-copy plots. Specify the file name with `post=`.

`none` - just creates the Annotate data set and sets up `goptions` using `gopplot=`.

Data Set and Catalog Options

These options specify the input SAS data set, output Annotate data set, and options for writing plots to files and catalogs.

data= *SAS-data-set*
specifies the input data set. The default input data set is the last data set created. You should always specify the `data=` option since the macro creates data sets that are not suitable for use as input.

filepref= *file-name-prefix*
specifies the file name prefix. The default is `filepref=sasplot`.

gdesc= *description*
specifies the name of a catalog description. This option can optionally be used with `proc anno gout=` to provide the `description=`.

gname= *name*
specifies the name of a catalog entry. This option can optionally be used with `proc anno gout=` to provide the `name=`.

gout= *catalog*
specifies the `proc anno gout=` catalog. With `gout=gc.slides`, first specify: `libname gc '.'`; Then to replay, run: `proc greplay igout=gc.slides; run;` Note that replayed plots will not in general have the correct aspect ratio.

out= *SAS-data-set*
specifies the output Annotate data set. This data set contains all of the instructions for drawing the graph. The default is `out=anno`.

post= *filename*
specifies the graphics stream file name. The default name is constructed from the `filepref=` value and '`ps`' in a host-specific way.

Typical Options

These are some of the most frequently used options.

datatype= *data-type*
specifies the type of data analysis that generated the data set. This option is used to set defaults for other options and to do some preprocessing of the data.

When the data type is `corresp`, `mds`, `mca`, `row`, `column`, `mdpref`, `mdpref2`, `vector`, or `ideal`, the `label=typical` option is implied when `label=` is not otherwise specified. The default point label variable is the last character variable in the data set.

Some data types (`mdpref`, `vector`, `ideal`, `corresp`, `row`, `mca`, `column`, `mds`) expect certain observation types and set the `types=` list accordingly. For example, `mdpref` expects `_type_ = 'SCORE'` and `_type_ = 'CORR'` observations. The remaining data types do not expect any specific value of the `typevar=` variable. So if you do not get the right data types associated with the right observation types, specify `types=`, and specify the `types=` values in an order that corresponds to the order of the symbol types in the `Types Legend` table. Unlike `symtype=`, the order in which you specify `datatype=` values is irrelevant.

A null value (`datatype=`, the default) specifies no special processing, and the default plotting variables are the first two numeric variables in the data set. Specifying `corresp`, `mds`, `mca`, `row`, or `column` will set the default `plotvars` to `dim2` and `dim1`. Otherwise, when a nonnull value is specified, the default

`plotvars` are `prin2` and `prin1`.

Here are the various data types.

`datatype=column`
specifies a `proc corresp profile=column` analysis. Row points are plotted as vectors.

`datatype=contour`
draws solid color contour plots. When the number of row points is not the same as the number of column points in the grid, use `hnobs=` and `vnobs=` to specify the number of points. This method creates an `hnobs=` by `vnobs=` grid of colored rectangles. Each of the rectangles should touch all adjacent rectangles. This method works well with a regular grid of points. The `method=square` option is a good alternative when the data do not fall in a regular grid.

`datatype=corresp`
specifies an ordinary correspondence analysis.

`datatype=curve`
fits a curve through the scatter plot.

`datatype=curve2`
fits a curve through the scatter plot and tries to make the labels avoid overprinting the curve.

`datatype=function`
draws functions. Typically, no labels or symbols are drawn. This option has a similar effect to the PROC GPLOT `symbol` statement options `i=join v=none`.

`datatype=ideal`
specifies a PREFMAP ideal point model. See the `antiidea=`, `radii=`, and `cirsegs=` options.

`datatype=mca`
specifies a multiple correspondence analysis.

`datatype=mdpref`
specifies multidimensional preference analysis with vectors with blank labels. Note that `datatype=mdpref` can also be used for ordinary principal component analysis.

`datatype=mdpref2`
specifies MDPREF with vector labels (MDPREF and labels *too*).

`datatype=mds`
specifies multidimensional scaling.

`datatype=mds ideal`
specifies PREFMAP ideal point after the MDS.

`datatype=mds vector`
specifies PREFMAP after MDS.

`datatype=row`
specifies a `proc corresp profile=row` analysis. Column points are plotted as vectors.

`datatype=square`
plots each point as a solid square. The `datatype=square` option is useful as a form of contour plotting when the data do not form a regular grid. The `datatype=square` option, unlike `datatype=contour`, does not try to scale the size of the square so that each square will touch another square.

`datatype=symbol`
specifies an ordinary scatter plot.

`datatype=vector`
specifies a PREFMAP vector model.

`datatype=vector ideal`
specifies both PREFMAP vectors and ideal points.

For some `datatype=` values, a number may be specified after the name. This is primarily useful for biplot data sets produced by PROC PRINQUAL and PREFMAP data sets produced by PROC TRANSREG. This number specifies that the lengths of vectors should be changed by this amount. The number must be specified last. Examples: `datatype=mdpref 2`, `datatype=mds vector 1.5`.

The primary purpose of the `datatype=` option is to provide an easy mechanism for specifying defaults for the options in the next section (`typevar=` through `outward=`).

labelvar= *label-variable* | `_blank_`

specifies the variable that contains the point labels. The default is the last character variable in the data set. If `labelvar=_blank_` is specified, the macro will create a blank label variable.

options= *options-list*

specifies binary options. Specify zero, one, or more in any order. For example: `options=nocenter nolegend`.

`border`
draws a border box around the outside of the graphics area, like the border `goption`.

`close`
if a border is being drawn, perform the same adjustments on the border that are performed on the axes. This option is most useful with contour plots.

`diag`
draws a diagonal reference line.

`expand`
specifies Annotate data set post processing, typically for use with `extend=close` and contour plots. This option makes the plot bigger to fill up more of the window.

`nocenter`
do not center. By default, when `nocenter` is not specified, `vsize=` and `hsize=` are set to their maximum values, and the `vpos=` and `hpos=` values are increased accordingly. The x and y coordinates are increased to position the plot in the center of the full graphics area.

`noclip`
do not clip. By default, when `noclip` is not specified, labels that extend past the edges of the plot are clipped. This option will not absolutely prevent labels from extending beyond the plot, particularly when sizes are greater than 1.

`nocode`
suppresses the printing of the PROC PLOT and `goptions` statements.

`nodelete`
do not delete intermediate data sets.

`nohistory`
suppresses the printing of the iteration history table.

`nolegend`
suppresses the printing of the legends.

`noprint`
equivalent to `nolegend`, `nocode`, and `nohistory`.

`square`
uses the same ticks for both axes and tries to make the plot square by tinkering with the `extend=` option. Otherwise, ticks may be different.

`textline`
put text in the data set, followed by lines, so lines overwrite text. Otherwise text overwrites lines.

plotvars= *two-variable-names*

specifies the *y*-axis variable then the *x*-axis variable. To plot `dim2` and `dim3`, specify `plotvars=dim2 dim3`. The `datatype=` option controls the default variable list.

symlen= *n*

specifies the length of the symbols. By default, symbols are single characters, but the macro can center longer strings at the symbol location.

symvar= *symbol-variable* | `_symbol_`

specifies the variable that contains the plotting symbol for input to PROC PLOT. When `_symbol_` is specified, which is the default, the symbol variable is created, typically from the `symbols=` list, which may be constructed inside the macro. (Note that the variable `__symbol` is created to contain the symbol for the graphical scatter plot. The variables `_symbol_` and `__symbol` may or may not contain the same values.) Variables can be specified, and the first `symlen=` characters are used for the symbol. When a null value (`symvar=`) or a constant value is specified, the symbol from the printer plot will be used (which is always length one, no matter what is specified for `symlen=`). To get PROC PLOT pointer symbols, specify `symvar='00'x`, (hex null constant). To center labels with no symbols, specify: `symvar=, place=0`.

Observation-Type List Options

Data sets for plotting can have different types of observations that are plotted differently. These options allow you to specify the types of observations, the variable that contains the observation types, and the different ways the different types should be plotted. For many types of analyses, these can all be handled easily with the `datatype=` option, which sets analysis-specific defaults for the list options. When you can, you should use `datatype=` instead of the list options. If you do use the list options, specify a variable, in `typevar=`, whose values distinguish the observation types. Specify the list of valid types in `types=`. Then specify colors, fonts, sizes, and so on for the various observation types. Alternatively, you can use these options with `datatype=`. Specify lists for just those label or symbol characteristics you want to change, for example colors, fonts or sizes.

The lists do not all have to have the same number of elements. The number of elements in `types=` determines how many of the other list elements are used. When an observation type does not match one of the `type=` values, the results are the same as if the first type were matched. If one of the other lists is shorter than the `types=` list, the shorter list is extended by adding copies of the last element to the end. Individual list values may be quoted, but quotes are not required. *Embedded blanks are not permitted. If you embed blanks, you will not get the right results.* Values of the `typevar=` variable are compressed before they are used, so for example, an _type_ value of `'M COEFFI'` must be specified as `'MCOEFFI'`.

britypes= *type*

specifies the types to which `bright=` applies. The default is `britypes=symbol`.

colors= *colors-list*

specifies the default color list for the `symcol=` and `labcol=` options. The default is `colors=blue red green cyan magenta orange gold lilac olive purple brown gray rose violet salmon yellow`.

exttypes= *type*

specifies the types to always put in the `extraobs=` data set when they have blank labels. The default is `exttypes=vector`.

labcol= *label-colors*

specifies the colors for the point labels. The default list is constructed from the `colors=` option. Examples:

```
labcol='red'
labcol='red' 'white' 'blue'
```

labfont= *label-fonts*

specifies the fonts for the point labels. Examples:

```
labfont='swiss'
labfont='swiss' 'swissi'
```

labsize= *label-sizes*
specifies the sizes for the point labels. Examples:

```
labsize=1
labsize=1 1.5
labsize=1 0
```

rgbtypes= *type*
specifies the types to which `paint=`, `red=`, `green=`, and `blue=` apply. The default is `rgbtypes=symbol`.

symbols= *symbol-list*
specifies the plotting symbols. Symbols may be more than a single character. You must specify `symlen=n` for longer lengths. Blank symbols must be specified as `''` with no embedded blanks. Examples:

```
symbols='*'
symbols='**'
symbols='*' '+' '*' ''
symbols='NC' 'OH' 'NJ' 'NY'
```

symcol= *symbol-colors*
specifies the colors of the symbols. The default list is constructed from the `colors=` option. Examples:

```
symcol='red'
symcol='red' 'white' 'blue'
```

symtype= *symbol-types*
specifies the types of symbols. Valid values are `symbol`, `vector`, `circle`, `contour`, and `square`. Examples:

```
symtype='symbol'
symtype='symbol' 'vector'
symtype='symbol' 'circle'
```

symfont= *symbol fonts*
specifies the symbol fonts. The font is ignored for vectors with no symbols. Examples:

```
symfont='swiss'
symfont='swiss' 'swissi'
```

symsize= *symbol-sizes*
specifies the sizes of symbols. Examples:

```
symsize=1
symsize=1 1.5
```

types= *observation-types*
specifies the observations types. Observation types are usually values of a variable like _type_. Embedded blanks are not permitted. Examples:

```
types='SCORE'
types='OBS' 'SUPOBS' 'VAR' 'SUPVAR'
types='SCORE'
types='SCORE' 'MCOEFFI'
```

The order in which values are specified for the other options depends on the order of the types. The default types for various datatype= values are given next:

```
corresp:  'VAR' 'OBS' 'SUPVAR' 'SUPOBS'
row:      'VAR' 'OBS' 'SUPVAR' 'SUPOBS'
mca:      'VAR' 'OBS' 'SUPVAR' 'SUPOBS'
column:   'VAR' 'OBS' 'SUPVAR' 'SUPOBS'
mdpref:   'SCORE' 'CORR'
vector:   'SCORE' 'MCOEFFI'
ideal:    'SCORE' 'MPOINT'
mds:      'SCORE' 'CONFIG'
```

For combinations of options, these lists are combined in order, but without repeating 'SCORE', for example with datatype=mdpref vector ideal, the default types= list is: 'SCORE' 'CORR' 'MCOEFFI' 'MPOINT'.

typevar= *variable*
specifies a variable that is looked at for the observation types. By default, this will be _type_ if it is in the input data set.

Internal Data Set Options

The macro creates one or more of these data sets internally to store intermediate results.

extraobs= *SAS-data-set*
specifies a data set used to contain the extra observations that do not go through PROC PLOT. The default is extraobs=extraobs.

preproc= *SAS-data-set*
specifies a data set used to contain the preprocessed data= data set. The default is preproc=preproc.

regdat= *SAS-data-set*
specifies a data set used to contain intermediate regression results for curve fitting. The default is regdat=regdat.

tempdat1= *SAS-data-set*
specifies a data set used to hold intermediate results. The default is `tempdat1=tempdat1`.

tempdat2= *SAS-data-set*
specifies a data set used to hold intermediate results. The default is `tempdat2=tempdat2`.

tempdat3= *SAS-data-set*
specifies a data set used to hold intermediate results. The default is `tempdat3=tempdat3`.

tempdat4= *SAS-data-set*
specifies a data set used to hold intermediate results. The default is `tempdat4=tempdat4`.

tempdat5= *SAS-data-set*
specifies a data set used to hold intermediate results. The default is `tempdat5=tempdat5`.

tempdat6= *SAS-data-set*
specifies a data set used to hold intermediate results. The default is `tempdat6=tempdat6`.

Miscellaneous Options

Here are some options that are sometimes needed for certain situations to control the details of the plots.

antiidea= *n*
eliminates PREFMAP anti-ideal points. The TRANSREG ideal-point model assumes that small attribute ratings mean that the object is similar to the attribute and large ratings imply dissimilarity to the attribute. For example, if the objects are food and the attribute is "sweetness," then the analysis assumes that 1 means sweet and 9 is much less sweet. The resulting coordinates are usually ideal points, representing an ideal amount of the attribute, but sometimes they are anti-ideal points and need to be converted to ideal points. This option is used to specify the nature of the data (small ratings mean similar or dissimilar) and to request automatic conversion of anti-ideal points.

null value - (`antiidea=`, the default) - do nothing.

1 - reverses in observations whose `_TYPE_` contains 'POINT' when `_issq_` > 0. Specify `antiidea=1` with `datatype=ideal` for the unusual case when large data values are positive or ideal.

-1 - reverses in observations whose `_type_` contains 'POINT' when `_issq_` < 0. Specify `antiidea=-1` with `datatype=ideal` for the typical case when small data values are positive or ideal.

extend= *axis-extensions*
is used to extend the x and y axes beyond their default lengths. Specify four values, for the left, right, top, and bottom axes. If the word `close` is specified somewhere in the string, then macro moves the axes in close to the extreme data values, and the computed values are added to the specified values (if any). Sample specifications: `extend=2 2`, or `extend=3 3 -0.5 0.5`. Specifying a positive value n

extends the axis n positions in the indicated direction. Specifying a negative value shrinks the axis. The defaults are in the range -2 to 2, and are chosen in an attempt to add a little extra horizontal space and make equal the extra space next to each of the four extreme ticks. When there is enough space, the horizontal axis is slightly extended by default to decrease the chance of a label slightly extending outside the plot. PROC PLOT usually puts one or two more lines on the top of the plot than in the bottom. The macro tries to eliminate this discrepancy. This option does not add any tick marks; it just extends or shrinks the ends of the axis lines. So typically, only small values should be specified. Be careful with this option and a positive `makefit=` value.

font= *font*
specifies the default font. The default is `font=swiss`.

hminor= *n* | *do-list*
specifies the number of horizontal axis minor tick marks between major tick marks. A typical value is 9. The number cannot be specified when `haxis=` is specified with `plotopts=`. Alternatively, specify a DATA step `do` list. Note that with log scaling, specify $\log_{10}$'s of the data values. For example, specify `hminor=0.25 to 5 by 0.25`, with data ranging up to 10**5.

href= *do-list*
specifies horizontal-axis reference lines (which are drawn vertically). Specify a DATA step `do` list. By default, there are no reference lines.

inc= *n*
specifies `haxis=by inc` and `vaxis=by inc` values. The specified increments apply to both axes. To individually control the increments, you must specify the PLOT statement `haxis=` and `vaxis=` options on the `plotopts=` option. When you are plotting variables that have very different scales, you may need to independently specify appropriate tick increments for both axes to get a reasonable plot. Here is an example: `plotopts=haxis=by 20 vaxis=by 5000`.

interpol= ls | tick | no | hlog | vlog | yes
specifies the axis interpolation method.

`ls` - uses the least-squares method only. This method computes the mapping between data and positions using ordinary least-squares linear regression. Usually, you should not specify `interpol=ls` because slight inaccuracies may result, producing aesthetically unappealing plots.

`hlog` - specifies that the x-axis is on a log scale.

no - does not interpolate.

`tick` - uses the tick mark method. This method computes the slope and intercept using tick marks and their values. Tick marks are read using the `tickfor=` format.

`vlog` - specifies that the y-axis is on a log scale.

`yes` - the default, interpolates symbol locations, starting with least squares but replacing them with tick-based estimates when they are available.

This option makes the symbols, vectors, and circles map to the location they would in a true graphical scatter plot, not the cell locations from PROC PLOT. This option has no effect on labels, the frame, reference lines, titles, or ticks. With `interpol=no`, plots tend to look nicer whereas `interpol=yes` plots are slightly more accurate. Note that the strategy used to interpolate can be defeated in certain cases. If the horizontal axis tick values print vertically, specify `interpol=ls`. The `hlog` and `vlog` values are specified in addition to the method. For example, `interpol=yes vlog hlog`.

label= *label-statement*
specifies a `label` statement. Note that specifying the keyword `label` to begin the statement is optional. You can specify `label=typical` to request a label statement constructed with `'Dimension'` and the numeric suffix of the variable name, for example, `label dim1 = 'Dimension 1' dim2 = 'Dimension 2';` when `plotvars=dim2 dim1`. The `label=typical` option can only be used with variable names that consist of a prefix and a numeric suffix.

ls= *n* | *iterative-specification*
specifies how line sizes are generated. The default is `ls=compute search`. When the second word is `search`, the macro searches for an optimal line size. See the `place=` option for more information on searches. When the first word is `compute`, the line size is computed from the iteration number so that the line sizes are: 65 80 100 125 150 175 200. Otherwise the first word is the first linesize and with each iteration the linesize is incremented by the `lsinc=` amount. Example: `ls=65 search`.

lsinc= *n*
specifies the increment to line size in iterations when line size is not computed. The default is `lsinc=15`.

lsizes= *number-list*
specifies the line sizes (thicknesses) for frame, ticks, vectors, circles, curves, respectively. The default is `lsizes=1 1 1 1 1`.

maxiter= *n*
specifies the maximum number of iterations. The default is `maxiter=15`.

maxokpen= *n*
specifies the maximum acceptable penalty sum. The default is `maxokpen=0`. Penalties accrue when label characters collide, labels get moved too far from their symbols, or words get split.

offset= *n*
move symbols for coincident points `offset=` spaces up/down and left/right. This helps to better display coincident symbols. Specify a null value (`offset=,`) to turn off offsetting. The default is `offset=0.25`.

place= *placement-specification*
generates a `placement=` option for the plot request. The default is `place=2 search`. Specify a non-negative integer. Values greater than 13 are set to 13. As the value gets larger, the procedure is given more freedom to move the labels farther from the symbols. The generated placement list will be printed in the log. You can still specify `placement=` directly on the `plotopts=` option. This option just gives

you a shorthand notation. For example:

```
place=0 - placement=((s=center))

place=1 - placement=((h=2 -2 : s=right left)
                     (v=1 * h=0 -1 to -2 by alt))

place=2 - placement=((h=2 -2 : s=right left)
                     (v=1 -1 * h=0 -1 to -5 by alt))

place=3 - placement=((h=2 -2 : s=right left)
                     (v=1 to 2 by alt * h=0 -1 to -10 by alt))

place=4 - placement=((h=2 -2 : s=right left)
                     (v=1 to 2 by alt * h=0 -1 to -10 by alt)
                     (s=center right left * v=0 1 to 2 by alt *
                      h=0 -1 to -6 by alt * l= 1 to 2))
```

and so on.

The `place=` option, along with the `ls=` option can be used to search for an optimal placement list and an optimal line size. By default, the macro will create and recreate the plot until it avoids all collisions. The search is turned off when a `placement=` option is detected in the plot request or plot options.

If search is not specified with `place=` or `ls=`, the specified value is fixed. If search is specified with the other option, only that option's value is incremented in the search.

plotopts= *options*

specifies PLOT statement options. The `box` option will be specified, even if you do not specify it. Reference lines should not be specified using the PROC PLOT `href=` and `vref=` options. Instead, they should be specified directly using the `href=` and `vref=` macro options. By default, no PLOT statement options are specified except `box`.

procopts= *options*

specifies PROC PLOT statement options. The default is `procopts=nolegend`.

tickaxes= *axis-string*

specifies the axes to draw tick marks. The default, `tickaxes=`LRTBFlb, means major ticks on left (L), right (R), top (T), and bottom (B), and the full frame (F) is to be drawn, and potentially minor tick marks on the left (l) and bottom (b). Minor ticks on the right (r) and top (t) can also be requested. To just have major tick marks on the left and bottom axes, and no full frame, specify `tickaxes=LB`. Order and spacing do not matter. `hminor=` and `vminor=` must also be specified to get minor ticks.

tickfor= *format*

specifies the tick format used by `interpol=tick`. You should change this if the tick values in the PROC PLOT output cannot be read with the default `tickfor=32.` format. For example, specify `tickfor=date7.` with dates.

ticklen= *n*
specifies the length of tick marks in horizontal cells. A negative value can be specified to indicate that only half ticks should be used, which means the ticks run to but not across the axes. The default is `ticklen=1.5`.

tsize= *n*
specifies the default text size. The default is `tsize=1`.

vminor= *n* | *do-list*
specifies the number of vertical axis minor tick marks between major tick marks. A typical value is 9. The number cannot be specified when `vaxis=` is specified with `plotopts=`. Alternatively, specify a DATA step `do` list. Note that with log scaling, specify $\log_{10}$'s of the data values. For example, specify `vminor=0.25 to 5 by 0.25`, with data ranging up to 10**5.

vref= *do-list*
specifies vertical reference lines (which are drawn horizontally). Specify a DATA step `do` list. By default, there are no reference lines.

Color Options

The symbol and point label colors are set by the `labcol=` and `symcol=` options. Here are the other color options.

bright= *n*
generates random label colors for `britypes=` values. In congested plots, it may be easier to see which labels and symbols go together if each label/symbol pair has a different random color. Colors are computed so that the mean RGB (red, green, blue) components equal the specified `bright=` value. The valid range is $5 \leq bright \leq 250$. 128 is a good value. Small values will produce essentially black labels and large values will produce essentially white labels, and so should be avoided. The default is a null value, `bright=`, and there are no random label colors. If you get a color table full error message, you need to specify larger values for the `rgbround=` option.

cframe= *color*
specifies the color of the background within the frame. This is analogous to the `cframe=` SAS/GRAPH option. By default, when `cframe=` is null, this option has no effect.

color= *color*
specifies the default color that is used when no other color is set. The default color, `color=cyan`, was chosen because it will show up on both a dark background and a white background, not because it is ever the color of choice. You may want to change the default color to black for printed plots and white otherwise or to some other more-suitable colors.

curvecol= *color*
specifies the color of curves in a regression plot. The default comes from `color=`.

excolors= *color-list*
excludes observations from the Annotate data set with colors in this list. For example, with a black background, to exclude all observations that have a color set to black as well as those with a computed black color, for example from `bright=` or `paint=`, specify `excolors=black CX000000`. This is done for efficiency, to make the Annotate data set smaller.

framecol= *color*
specifies the color of the frame. The default comes from `color=`.

labelcol= *color*
specifies the color of the variable labels. The default comes from `color=`.

monochro= *color*
overrides all other specified colors. This option is useful when you have specified colors and you want to temporarily override them to send the plot to a monochrome device. By default, when `monochro=` is null, this option has no effect. Typical usage: `monochro=black`.

tickcol= *color*
specifies the color of ticks. The default comes from `color=`.

titlecol= *color*
specifies the color of the title. The default comes from `color=`.

Color Interpolation and Painting

These next options are used to create label and symbol colors using some function of the input data set variables. For example, you can plot the first two principal components on the x and y axes and show the third principal component in the same plot by using it to control the label colors. The `paint=` option gives you a simple and fairly general way to interpolate colors. The `red=`, `green=`, and `blue=` options are used together for many other types of interpolations, but these options are much harder to use. These options apply to `rgbtypes=` observations. If `red=`, `green=`, and `blue=` are not flexible enough, for example if you need full statements, specify `red=128` (so later code will know you are computing colors) then insert the full statements you need to generate the colors using `adjust1=`.

paint= *color-interpolation-specification*
is used to interpolate between colors based on the values of a variable. The simplest specification is `paint=variable`. More generally, specify:

```
paint=variable optional-color-list optional-data-value-list
```

The following color names are recognized: red, green, blue, yellow, magenta, cyan, black, white, orange, brown, gray, olive, pink, purple, violet. For other colors, specify the RGB color name. Colors can be

represented as `CX`*rrggbb* where *rr* is the red value, *gg* is the green, and *bb* is blue, all three specified in hex. The base ten numbers 0 to 255 map to 00 to FF hex. For example, white is `CXFFFFFF`, black is `CX000000`, red is `CXFF0000`, blue is `CX0000FF` and magenta is `CXFF00FF`. When a variable named `z` is specified with no other arguments, the default is `paint=z blue magenta red`. The option `paint=z red green 1 10` interpolates between red and green, based on the values of the variable `z`, where values of 1 or less map to red, values of 10 or more map to green, and values in between map to colors in between. The specification `paint=z red yellow green 1 5 10`, interpolates between red at `z=1`, yellow at `Z=5`, and green at `Z=10`. If the data value list is omitted, it is computed from the data.

red= *expression*

green= *expression*

blue= *expression*

specify for arithmetic expressions that produce integers in the range 0 to 255. Colors will be created as follows:

```
__color = 'CX' ||
          put(%if &red   ne %then round(&red,  __roured); %else 128; ,hex2.) ||
          put(%if &green ne %then round(&green,__rougre); %else 128;, hex2.) ||
          put(%if &blue  ne %then round(&blue, __roublu); %else 128; ,hex2.);}
```

The `__rou` variables are extracted from the second through fourth values of the `rgbround=` option. Example: `red = min(100 + (z - 10) * 3, 255)`, `blue=50`, `green=50`. Then all labels are various shades of red, depending on the value of `z`. Be aware that light colors (small red-green-blue values) do not show up well on white backgrounds and dark colors do not show up well on dark backgrounds. Typically, you will not want to use the full range of possible red-green-blue values. Computed values greater than 255 will be set to 255.

rgbround= *RGB-rounding-specification*

specifies rounding factors used for the `paint=` variable and RGB values. The default is `rgbround=-240 1 1 1`. The first value is used to round the `paint=` variable. Specify a positive value to have the variable rounded to multiples of that value. Specify a negative value $-n$ to have a maximum of n colors. For the other three values, specify positive values. The last three are rounding factors used to round the values for the red, green, and blue component of the color (see `red=`). If more than 256 colors are generated, you will get the error that a color was not added because the color table is full. By default, when a value is missing, there is no rounding. Rounding the `paint=` variable is useful with contour plots.

Contour Options

Use these options with contour plots. For example if the grid for a contour plot was generated as follows.

```
do x = -4 to 4 by 0.1;
   do y = -2 to 2 by 0.1;
      ... statements ...
      end;
   end;
```

then specify `hnobs=81`, `vnobs=41`. By default, the square root of the number of contour type observations is used for both `hnobs=` and `vnobs=` (which assumes a square grid).

hnobs= *n*

specifies the number of horizontal observations in the grid for contour plots.

vnobs= *n*

specifies the number of vertical observations in the grid for contour plots.

Advanced Plot Control Options

You can use the these next options to add full SAS DATA step statements to strategic places in the macro, such as the PROC PLOT step, the end of the preprocessing, and last full data steps. These options do minor adjustments before the final plot is produced. These options allow very powerful customization of your results to an extent not typically found in procedures. However, they may require a fair amount of work and some trial and error to understand and get right.

adjust1= *SAS-statements*

The following variables are created in the preprocessing data set:

`__lsize` - label size
`__lfont` - label font
`__lcolor` - label color
`__ssize` - symbol size
`__sfont` - symbol font
`__scolor` - symbol color
`__stype` - symbol type
`__symbol` - symbol value
`__otype` - observation type

Use `adjust1=` to adjust these variables in the preprocessing data set. You must specify complete statements with semicolons. Examples:

```
adjust1=%str(__lsize = 1.2; __lcolor = green;)}

adjust1=%str(if z > 20 then do;
__scolor = 'green'; __lcolor = 'green'; end;)}
```

adjust2= *SAS-statements*
includes statements with PROC PLOT such as `format` statements. Just specify the full statement.

adjust3= *SAS-statements*

adjust4= *SAS-statements*
specify options to adjust the final Annotate data set. For example, in Swiss fonts, asterisks are not vertically centered when printed, so `adjust3=` converts to use the SYMBOL function, so by default, `adjust3=%str(if text = '*' and function = 'LABEL' then do; style = ' '; text = 'star'; function = 'SYMBOL'; end;)`. The default for `adjust4=` is null, so you can use it to add new statements. If you add new variables to the data set, you must also include a `keep` statement. Here is an example of using `adjust4=` to vertically print the y-axis label, like it would be in PROC PLOT.

```
adjust4=%str(if angle = 90 then do; angle = 270; rotate = 90; keep rotate; end;)
```

This example changes the size of title lines.

```
adjust4=%str(if index(comment, 'title') then size = 2;)
```

adjust5= *SAS-statements*
adds extra statements to the final DATA step that is used only for `datatype=function`. For example, to periodically mark the function with pluses, specify:

```
adjust5=%str( if mod(_n_,30) = 0 then do;
                size=0.25; function = 'LABEL'; text = '+'; output; end;)
```

Other Options

Here are the remaining options for the `%PlotIt` macro.

cirsegs= *n*
specifies a circle smoothness parameter used in determining the number of line segments in each circle. Smaller values create smoother circles. The `cirsegs=` value is approximately related to the length of the line segments that compose the circle. The default is `cirsegs=.1`.

cursegs= *n*
specifies the number of segments in a regression function curve. The default is `cursegs=200`.

debug= vars | dprint | notes | time | mprint
specifies values that control debugging output.

`dprint` - print intermediate data sets.

`mprint` - run with `options mprint`.

`notes` - do not specify `options nonotes` during most of the macro.

`time` - prints total macro run time, ignored with `options nostimer;`

`vars` - print macro options and macro variables for debugging.

You should provide a list of names for more than one type of debugging. Example: `debug=vars dprint notes time mprint`. The default is `debug=time`.

hpos= *n*
specifies the number of horizontal positions in the graphics area.

hsize= *n*
specifies the horizontal graphics area size in `unit=` units. The default is the maximum size for the device. By default, when `options=nocenter` is not specified, `hsize=` affects the size of the plot but not the `hsize= goption`. When `options=nocenter` is specified, `hsize=` affects both the plot size and the `hsize= goption`. If you specify just the `hsize=` but not `vsize=`, the vertical size will be scaled accordingly.

makefit= *n*
specifies the proportion of the graphics window to use. When the `makefit=` value is negative, the absolute value will be used, and the final value may be changed if the macro thinks that part of the plot may extend over the edge. When a positive value is specified, it will not be changed by the macro. When nonnull, the macro uses GASK to determine the minimum and maximum graphics window sizes and makes sure the plot can fit in them. The macro uses `gopprint=` or `gopplot=` to determine the device. The default is `makefit=-0.95`.

nknots= *n*
specifies the PROC TRANSREG number of knots option for regression functions.

outward= none | *'c'*
specifies a string for the PLOT statement `outward=` option. Normally, this option's value is constructed from the symbol that holds the place for vectors. Specify `outward=none` if you want to not have `outward=` specified for vectors. The `outward=` option is used to greatly increase the likelihood that labels from vectors will be printed outward − away from the origin.

ps= *n*
specifies the page size.

radii= *do-list*
specifies the radii of circles (in a DATA step do list). The unit corresponds to the horizontal axis variable. The `radii=` option can also specify a variable in the input data set when radii vary for each point. By default, no circles are drawn.

regopts= *options*
specifies the PROC TRANSREG options for curve fitting. Example: `regopts=nknots=10 evenly`.

regfun= *regression-function*
specifies the function for curve fitting. Possible values include:

`linear` - line

`spline` - nonlinear spline function, perhaps with knots

`mspline` - nonlinear but monotone spline function, perhaps with knots

`monotone` - nonlinear, monotone step function

See PROC TRANSREG documentation for more information

regprint= noprint
specifies the PROC TRANSREG PROC statement options, typically printing options such as:

`noprint` - no regression printed output

`short` - suppress iteration histories

`ss2` - regression results

To see the regression table, specify: `regprint=ss2 short`. The default is `regprint=noprint`.

unit= in | cm
specifies the `hsize=` and `vsize=` unit (in or cm). The default is `unit=in`.

vechead= *vector-head-size* specifies how to draw vector heads. For example, the default specification `vechead=0.2 0.05`, specifies a head consisting of two hypotenuses from triangles with sides 0.2 units long along the vector and 0.05 units on the side perpendicular to the vector.

vpos= *n*
specifies the number of vertical positions in the graphics area.

vsize= *n*
specifies the vertical graphics area size in `unit=` units. The default is the maximum size for the device. By default when `options=nocenter` is not specified, `vsize=` affects the size of the plot but not the `vsize= goption`. When `options=nocenter` is specified, `vsize=` affects both the plot size and the `vsize= goption`. If you specify just the `vsize=` but not `hsize=`, the horizontal size will be scaled accordingly.

vtoh= *n*
specifies the PROC PLOT `vtoh=` option. The `vtoh=` option specifies the ratio of the vertical height of a typical character to the horizontal width. The default is `vtoh=2`. Do not specify values much different than 2, especially by default when you are using proportional fonts. There is no one-to-one correspondence between characters and cells and character widths vary, but characters tend to be approximately twice as high as they are wide. When you specify `vtoh=` values larger than 2, near-by labels may overlap, even when they do not collide in the printer plot. The macro uses this option to equate the axes so that a centimeter on one axis represents the same data range as a centimeter on the

other axis. A null value can be specified, `vtoh=`, when you want the macro to just fill the window, like a typical GPLOT.

Smaller values give you more lines and smaller labels. The specification `vtoh=1.75` is a good alternative to `vtoh=2` when you need more lines to avoid collisions. The specification `vtoh=1.75` means 7 columns for each 4 rows between ticks (7 / 4 = 1.75). The `vtoh=2` specification means the plot will have 8 columns for each 4 rows between ticks. Note that PROC PLOT sometimes takes this value as a hint, not as a rigid specification so the actual value may be slightly different, particularly when a value other than 2.0 is specified. This is generally not a problem; the macro adjusts accordingly.

xmax= *n*
specifies the maximum horizontal size of the graphics area.

ymax= *n*
specifies the maximum vertical size of the graphics area.

Macro Errors

Usually, if you make a mistake in specifying macro options, the macro will print an informative message and quit. These macros go to great lengths to check their input and issue informative errors. However, *complete* error checking like we have with procedures is impossible in macros, and sometimes you will get a cascade of less than helpful error messages.* In that case, you will have to check the input and hunt for errors. One of the more common errors is a missing comma between options. Sometimes for harder errors, specifying `options mprint;` will help you locate the problem. You may get a listing with a *lot* of code, almost all of which you can ignore. Search for the error and look at the code that comes before the error for ideas about what went wrong. Once you think you know which option is involved, be sure to also check the option before and after in your macro invocation, because that might be where the problem really is.

The `%PhChoice` macro use PROC TEMPLATE and ODS to create customized output tables. Typically, the instructions for this customization, created by PROC TEMPLATE, are stored in a file under the `sasuser` directory with a host dependent name. On some hosts, this name is `templat.sas7bitm`. On other hosts, the name is some variation of the name `templat`. Sometimes this file can be corrupted. When this happens, these macros will not run correctly, and you will see error messages including errors about invalid pages. The solution is to find the corrupt file under `sasuser` and delete it (using your ordinary operating system file deletion method). After that, this macros should run fine again. If you have run any other PROC TEMPLATE customizations, you will need to rerun them after deleting the file. For more information, see "Template Store" or "Item Store" in the SAS ODS documentation.

Sometimes, you will run the `%MktEx` macro, and everything will seem to run fine in the entire job, but at the end of your SAS log, you will see the message:

```
ERROR: Errors printed on page ....
```

Typically, this is caused by one or more PROC FACTEX steps failing to find the requested design. When this happens, the macro recovers and continues searching. The macro does not always know in advance if PROC FACTEX will succeed. The only way for it to find out is for it to try. The macro suppresses the PROC FACTEX error messages along with most other notes and warnings that would ordinarily come out. However, SAS still knows that a procedure tried to print an error message, and prints an error at the end of the log. This error can be ignored.

*If this happens, please write Warren.Kuhfeld@sas.com, and I will see if I can make the macros better handle that problem in the next release. Send all the code necessary to reproduce what you have done.

Linear Models and Conjoint Analysis with Nonlinear Spline Transformations

Warren F. Kuhfeld

Mark Garratt

Abstract

Many common data analysis models are based on the general linear univariate model, including linear regression, analysis of variance, and conjoint analysis. This chapter discusses the general linear model in a framework that allows nonlinear transformations of the variables. We show how to evaluate the effect of each transformation. Applications to marketing research are presented.*

Why Use Nonlinear Transformations?

In marketing research, as in other areas of data analysis, relationships among variables are not always linear. Consider the problem of modeling product purchasing as a function of product price. Purchasing may decrease as price increases. For consumers who consider price to be an indication of quality, purchasing may increase as price increases but then start decreasing as the price gets too high. The number of purchases may be a discontinuous function of price with jumps at "round numbers" such as even dollar amounts. In any case, it is likely that purchasing behavior is not a linear function of price. Marketing researchers who model purchasing as a linear function of price may miss valuable nonlinear information in their data. A transformation regression model can be used to investigate the nonlinearities. The data analyst is not required to specify the form of the nonlinear function; the data suggest the function.

The primary purpose of this chapter is to suggest the use of linear regression models with nonlinear transformations of the variables—*transformation regression* models. It is common in marketing research to model nonlinearities by fitting a quadratic polynomial model. Polynomial models often have collinearity problems, but that can be overcome with orthogonal polynomials. The problem that polynomials cannot overcome is the fact that polynomial curves are rigid; they do not do a good job of locally fitting the data. Piecewise polynomials or *splines* are generalizations of polynomials that provide more flexibility than ordinary polynomials.

*This chapter is a revision of a paper that was presented to the American Marketing Association, Advanced Research Techniques Forum, June 14–17, 1992, Lake Tahoe, Nevada. The authors are: Warren F. Kuhfeld, Manager, Multivariate Models R&D, SAS Institute Inc., Cary NC 27513-2414. Mark Garratt was with Conway | Milliken & Associates, when this paper was presented and is now with Miller Brewing Company. Copies of this chapter (TS-689I) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market .

Background and History

The foundation for our work can be found mostly in the psychometric literature. Some relevant references include: Kruskal & Shepard (1974); Young, de Leeuw, & Takane (1976); de Leeuw, Young, & Takane (1976); Perreault & Young (1980); Winsberg & Ramsay (1980); Young (1981); Gifi (1981, 1990); Coolen, van Rijckevorsel, & de Leeuw (1982); van Rijckevorsel (1982); van der Burg & de Leeuw (1983); de Leeuw (1986), and many others. The transformation regression problem has also received attention in the statistical literature (Breiman & Friedman 1985, Hastie & Tibshirani, 1986) under the names *ACE* and *generalized additive models*.

Our work is characterized by the following statements:

- Transformation regression is an inferential statistical technique, not a purely descriptive technique.
- We prefer smooth nonlinear spline transformations over step-function transformations.
- Transformations are found that minimize a squared-error loss function.

Many of the models discussed in this chapter can be directly fit with some data manipulations and any multiple regression or canonical correlation software; some models require specialized software. Algorithms are given by Kuhfeld (1990), de Boor (1978), and in SAS/STAT documentation.

Next, we present notation and review some fundamentals of the general linear univariate model.

The General Linear Univariate Model

A general linear univariate model has the scalar form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \epsilon$$

and the matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The dependent variable $\mathbf{y}$ is an $(n \times 1)$ vector of observations; $\mathbf{y}$ has expected value $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and expected variance $V(\mathbf{y}) = \sigma^2\mathbf{I}_n$. The vector $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ contains the unobservable deviations from the expected values. The assumptions on $\mathbf{y}$ imply $E(\epsilon) = \mathbf{0}$ and $V(\epsilon) = \sigma^2\mathbf{I}_n$. The columns of $\mathbf{X}$ are the independent variables; $\mathbf{X}$ is an $(n \times m)$ matrix of constants that are assumed to be known without appreciable error. The elements of the column vector $\boldsymbol{\beta}$ are the parameters. The objectives of a linear models analysis are to estimate the parameter vector $\boldsymbol{\beta}$, estimate interesting linear combinations of the elements of $\boldsymbol{\beta}$, and test hypotheses about the parameters $\boldsymbol{\beta}$ or linear combinations of $\boldsymbol{\beta}$.

We discuss fitting linear models with nonlinear spline transformations of the variables. Note that we do *not* discuss models that are nonlinear in the parameters such as

$$y = e^{x\beta} + \epsilon$$

$$y = \beta_0 x^{\beta_1} + \epsilon$$

$$y = \frac{\beta_1 x_1 + \beta_2 x_1^2}{\beta_3 x_2 + \beta_4 x_2^2} + \epsilon$$

Our nonlinear transformations are found within the framework of the general linear model.

There are numerous special cases of the general linear model that are of interest. When all of the columns of **y** and **X** are interval variables, the model is a multiple regression model. When all of the columns of **X** are indicator variables created from nominal variables, the model is a main-effects analysis of variance model, or a metric conjoint analysis model. The model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

is of special interest. It is a *linear* model because it is linear in the parameters, and it models y as a *nonlinear* function of x. It is a *cubic polynomial* regression model, which is a special case of a spline.

Polynomial Splines

Splines are curves that are typically required to be continuous and smooth. Splines are usually defined as piecewise polynomials of degree d whose function values and first $d-1$ derivatives agree at the points where they join. The abscissa values of the join points are called *knots*. The term spline is also used for polynomials (splines with no knots), and piecewise polynomials with more than one discontinuous derivative. Splines with more knots or more discontinuities fit the data better and use more degrees of freedom (*df*). Fewer knots and fewer discontinuities provide smoother splines that user fewer *df*. A spline of degree three is a cubic spline, degree two splines are quadratic splines, degree one splines are piecewise linear, and degree zero splines are step functions. Higher degrees are rarely used.

A simple special case of a spline is the line,

$$\beta_0 + \beta_1 x$$

from the simple regression model

$$y = \beta_0 + \beta_1 x + \epsilon$$

A line is continuous and completely smooth. However, there is little to be gained by thinking of a line as a spline. A special case of greater interest was mentioned previously. The polynomial

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

from the linear model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

is a cubic spline with no knots. This equation models y as a *nonlinear* function of x, but does so with a *linear* regression model; y is a linear function of x, x^2, and x^3. Table 1 shows the **X** matrix, ($\mathbf{1}$ $\mathbf{x}$ $\mathbf{x}^2$ $\mathbf{x}^3$), for a cubic polynomial, where $x = -5, -4, ..., 5$. Figure 1 plots the polynomial terms (except the intercept). See Smith (1979) for an excellent introduction to splines.

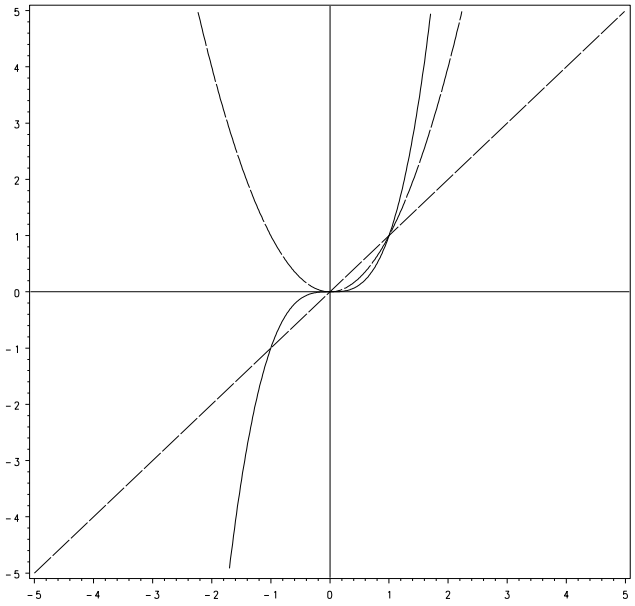


Figure 1. Linear, Quadratic, and Cubic Curves

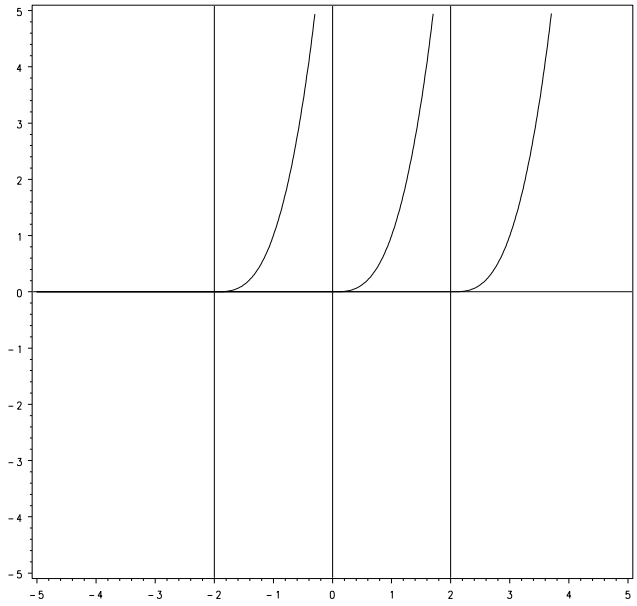


Figure 2. Curves For Knots at −2, 0, 2

Splines with Knots

Here is an example of a polynomial spline model with three knots at t_1, t_2, and t_3.

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \\ &\quad \beta_4 (x > t_1)(x - t_1)^3 + \\ &\quad \beta_5 (x > t_2)(x - t_2)^3 + \\ &\quad \beta_6 (x > t_3)(x - t_3)^3 + \epsilon \end{aligned}$$

The Boolean expression $(x > t_1)$ is 1 if $x > t_1$, and 0 otherwise. The term

$$\beta_4 (x > t_1)(x - t_1)^3$$

is zero when $x \leq t_1$ and becomes nonzero, letting the curve change, as x becomes greater than knot t_1. This spline uses more *df* and is less smooth than the polynomial model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

Assume knots at -2, 0, and 2; the spline model is:

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \\ &\quad \beta_4 (x > -2)(x - -2)^3 + \\ &\quad \beta_5 (x > \ \ 0)(x - \ \ 0)^3 + \\ &\quad \beta_6 (x > \ \ 2)(x - \ \ 2)^3 + \epsilon \end{aligned}$$

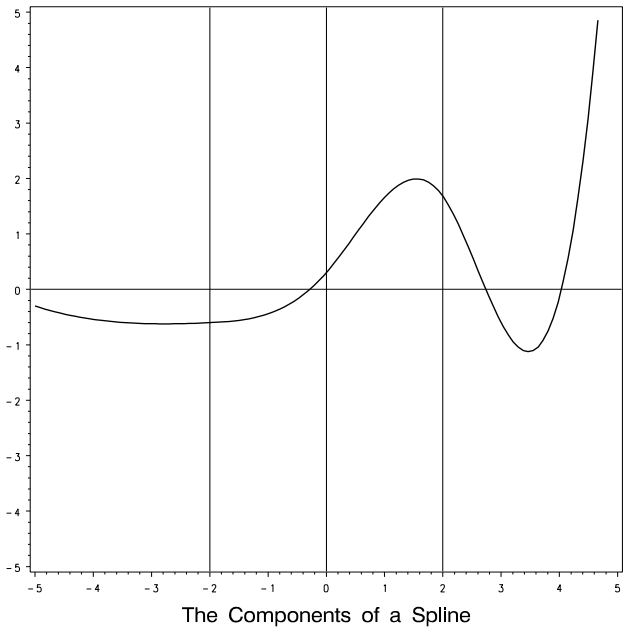


Figure 3. A Spline Curve With Knots at −2, 0, 2

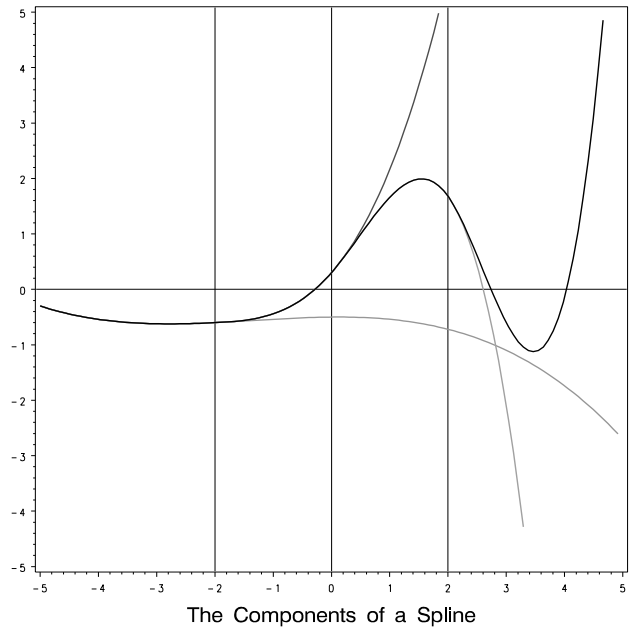


Figure 4. The Components of the Spline

Table 2 shows an **X** matrix for this model, Figure 1 plots the polynomial terms, and Figure 2 plots the knot terms.

The β_0, $\beta_1 x$, $\beta_2 x^2$, and $\beta_3 x^3$ terms contribute to the overall shape of the curve. The

$$\beta_4(x > -2)(x - -2)^3$$

term has no effect on the curve before $x = -2$, and allows the curve to change at $x = -2$. The $\beta_4(x > -2)(x - -2)^3$ term is exactly zero at $x = -2$ and increases as x becomes greater than -2. The

Table 1
Cubic Polynomial
Spline Basis

1	-5	25	-125
1	-4	16	-64
1	-3	9	-27
1	-2	4	-8
1	-1	1	-1
1	0	0	0
1	1	1	1
1	2	4	8
1	3	9	27
1	4	16	64
1	5	25	125

Table 2
Cubic Polynomial
With Knots at −2, 0, 2

1	-5	25	-125	0	0	0
1	-4	16	-64	0	0	0
1	-3	9	-27	0	0	0
1	-2	4	-8	0	0	0
1	-1	1	-1	1	0	0
1	0	0	0	8	0	0
1	1	1	1	27	1	0
1	2	4	8	64	8	0
1	3	9	27	125	27	1
1	4	16	64	216	64	8
1	5	25	125	343	125	27

Table 3
Basis for a Discontinuous (at 0) Spline

1	-5	25	-125	0	0	0	0
1	-4	16	-64	0	0	0	0
1	-3	9	-27	0	0	0	0
1	-2	4	-8	0	0	0	0
1	-1	1	-1	0	0	0	0
1	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
1	2	4	8	1	2	4	8
1	3	9	27	1	3	9	27
1	4	16	64	1	4	16	64
1	5	25	125	1	5	25	125

$\beta_4(x > -2)(x - -2)^3$ term contributes to the shape of the curve even beyond the next knot at $x = 0$, but at $x = 0$,

$$\beta_5(x > 0)(x - 0)^3$$

allows the curve to change again. Finally, the last term

$$\beta_6(x > 2)(x - 2)^3$$

allows one more change. For example, consider the curve in Figure 3. It is the spline

$$\begin{aligned} y &= -0.5 + 0.01x + -0.04x^2 + -0.01x^3 + \\ &\quad 0.1(x > -2)(x - -2)^3 + \\ &\quad -0.5(x > 0)(x - 0)^3 + \\ &\quad 1.5(x > 2)(x - 2)^3 \end{aligned}$$

It is constructed from the curves in Figure 4. At $x = -2.0$ there is a branch;

$$y = -0.5 + 0.01x + -0.04x^2 + -0.01x^3$$

continues over and down while the curve of interest,

$$\begin{aligned} y &= -0.5 + 0.01x + -0.04x^2 + -0.01x^3 + \\ &\quad 0.1(x > -2)(x - -2)^3 \end{aligned}$$

starts heading upwards. At $x = 0$, the addition of

$$-0.5(x > 0)(x - 0)^3$$

slows the ascent until the curve starts decreasing again. Finally, the addition of

$$1.5(x > 2)(x - 2)^3$$

produces the final change. Notice that the curves do not immediately diverge at the knots. The function and its first two derivatives are continuous, so the function is smooth everywhere.

Derivatives of a Polynomial Spline

The next equations show a cubic spline model with a knot at t_1 and its first three derivatives with respect to x.

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \\ &\quad \beta_4(x > t_1)(x - t_1)^3 + \epsilon \end{aligned}$$

$$\begin{aligned} \frac{dy}{dx} &= \beta_1 + 2\beta_2 x + 3\beta_3 x^2 + \\ &\quad 3\beta_4(x > t_1)(x - t_1)^2 \end{aligned}$$

$$\begin{aligned} \frac{d^2y}{dx^2} &= 2\beta_2 + 6\beta_3 x + \\ &\quad 6\beta_4(x > t_1)(x - t_1) \end{aligned}$$

$$\frac{d^3y}{dx^3} = 6\beta_3 + 6\beta_4(x > t_1)$$

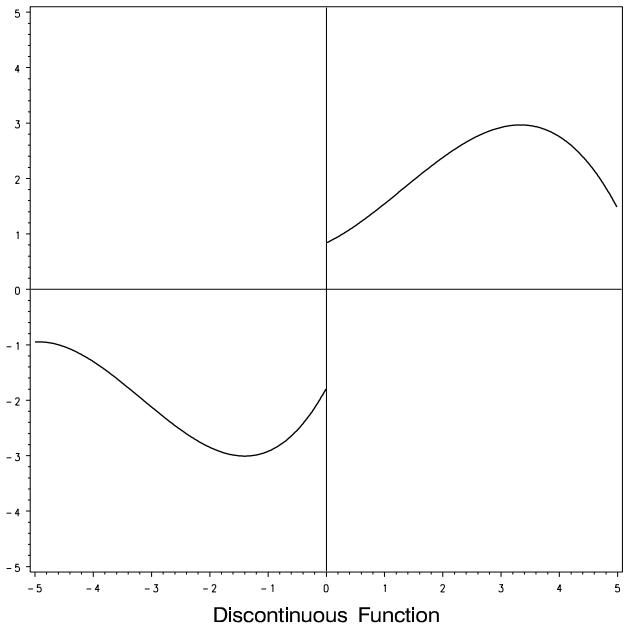


Figure 5. A Discontinuous Spline Function

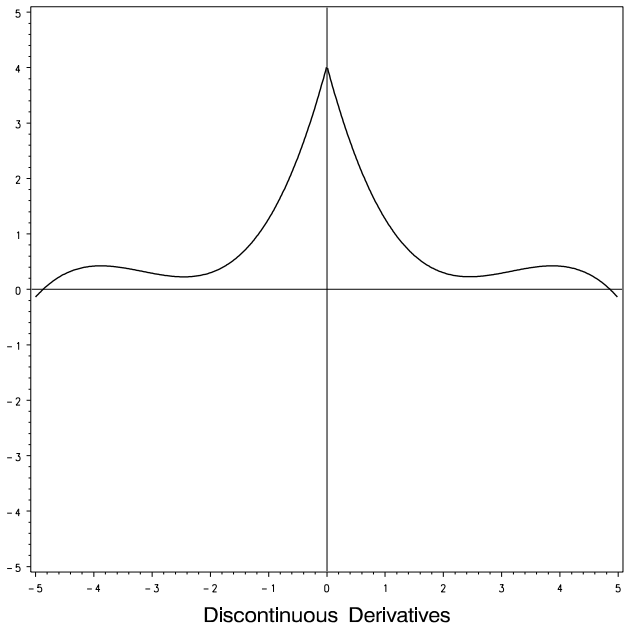


Figure 6. A Spline With a Discontinuous Slope

The first two derivatives are continuous functions of x at the knots. This is what gives the spline function its smoothness at the knots. In the vicinity of the knots, the curve is continuous, the slope of the curve is a continuous function, and the rate of change of the slope function is a continuous function. The third derivative is discontinuous at the knots. It is the horizontal line $6\beta_3$ when $x \leq t_1$ and jumps to the horizontal line $6\beta_3 + 6\beta_4$ when $x > t_1$. In other words, the cubic part of the curve changes at the knots, but the linear and quadratic parts do not change.

Discontinuous Spline Functions

Here is an example of a spline model that is discontinuous at $x = t_1$.

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \\ &\quad \beta_4 (x > t_1) + \\ &\quad \beta_5 (x > t_1)(x - t_1) + \\ &\quad \beta_6 (x > t_1)(x - t_1)^2 + \\ &\quad \beta_7 (x > t_1)(x - t_1)^3 + \epsilon \end{aligned}$$

Figure 5 shows an example, and Table 3 shows a design matrix for this model with $t_1 = 0$. The fifth column is a binary (zero/one) vector that allows the discontinuity. It is a change in the intercept parameter. Note that the sixth through eighth columns are necessary if the spine is to consist of two independent polynomials. Without them, there is a jump at $t_1 = 0$, but both curves are based on the

same polynomial. For $x \leq t_1$, the spline is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

and for $x > t_1$, the spline is

$$\begin{aligned} y &= \beta_0 + \beta_4 + \\ &\quad \beta_1 x + \beta_5 (x - t_1) + \\ &\quad \beta_2 x^2 + \beta_6 (x - t_1)^2 + \\ &\quad \beta_3 x^3 + \beta_7 (x - t_1)^3 + \epsilon \end{aligned}$$

The discontinuities are as follows:

$$\beta_7 (x > t_1)(x - t_1)^3$$

specifies a discontinuity in the third derivative of the spline function at t_1,

$$\beta_6 (x > t_1)(x - t_1)^2$$

specifies a discontinuity in the second derivative at t_1,

$$\beta_5 (x > t_1)(x - t_1)$$

specifies a discontinuity in the derivative at t_1, and

$$\beta_4 (x > t_1)$$

specifies a discontinuity in the function at t_1. The function consists of two separate polynomial curves, one for $-\infty < x \leq t_1$ and the other for $t_1 < x < \infty$. This kind of spline can be used to model a discontinuity in price.

Here is an example of a spline model that is continuous at $x = t_1$ but does not have $d - 1$ continuous derivatives.

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \\ &\quad \beta_4 (x > t_1)(x - t_1) + \\ &\quad \beta_5 (x > t_1)(x - t_1)^2 + \\ &\quad \beta_6 (x > t_1)(x - t_1)^3 + \epsilon \end{aligned}$$

$$\begin{aligned} \frac{dy}{dx} &= \beta_1 + 2\beta_2 x + 3\beta_3 x^2 + \\ &\quad \beta_4 (x > t_1) + \\ &\quad 2\beta_5 (x > t_1)(x - t_1) + \\ &\quad 3\beta_6 (x > t_1)(x - t_1)^2 \end{aligned}$$

Since the first derivative is not continuous at $t_1 = x$, the slope of the spline is not continuous at $t_1 = x$. Figure 6 contains an example with $t_1 = 0$. Notice that the slope of the curve is indeterminate at zero.

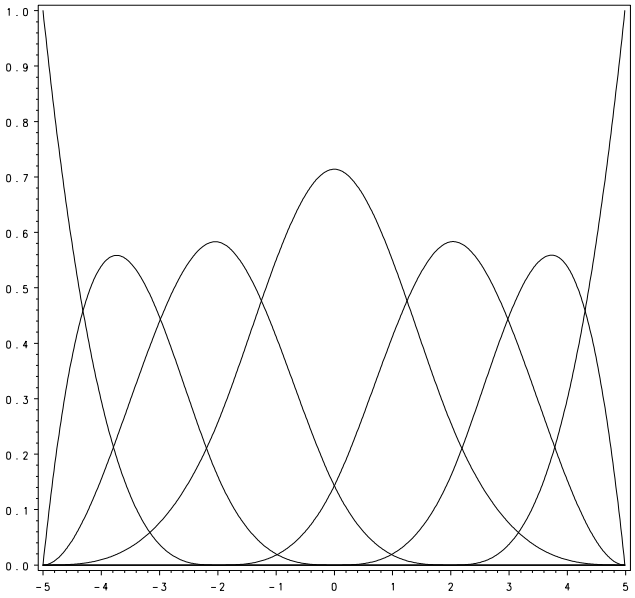


Table 4
Cubic B-Spline With Knots at −2, 0, 2

1.00	0.00	0.00	0.00	0.00	0.00	0.00
0.30	0.54	0.15	0.01	0.00	0.00	0.00
0.04	0.45	0.44	0.08	0.00	0.00	0.00
0.00	0.16	0.58	0.26	0.00	0.00	0.00
0.00	0.02	0.41	0.55	0.02	0.00	0.00
0.00	0.00	0.14	0.71	0.14	0.00	0.00
0.00	0.00	0.02	0.55	0.41	0.02	0.00
0.00	0.00	0.00	0.26	0.58	0.16	0.00
0.00	0.00	0.00	0.08	0.44	0.45	0.04
0.00	0.00	0.00	0.01	0.15	0.54	0.30
0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 7. B-Splines With Knots at −2, 0, 2

Monotone Splines and B-Splines

An increasing *monotone spline* never decreases; its slope is always positive or zero. Decreasing monotone splines, with slopes that are always negative or zero, are also possible. Monotone splines cannot be conveniently created from polynomial splines. A different basis, the *B-spline* basis, is used instead. Polynomial splines provide a convenient way to describe splines, but B-splines provide a better way to fit spline models.

The columns of the B-spline basis are easily constructed with a recursive algorithm (de Boor, 1978, pages 134−135). A basis for a vector space is a linearly independent set of vectors; every vector in the space has a unique representation as a linear combination of a given basis. Table 4 shows the B-spline $\mathbf{X}$ matrix for a model with knots at −2, 0, and 2. Figure 7 shows the B-spline curves. The columns of the matrix in Table 4 can all be constructed by taking linear combinations of the columns of the polynomial spline in Table 2. Both matrices form a basis for the same vector space.

The numbers in the B-spline basis are all between zero and one, and the marginal sums across columns are all ones. The matrix has a diagonally banded structure, such that the band moves one position to the right at each knot. The matrix has many more zeros than the matrix of polynomials and much smaller numbers. The columns of the matrix are not orthogonal like a matrix of orthogonal polynomials, but collinearity is not a problem with the B-spline basis like it is with a polynomial spline. The B-spline basis is very stable numerically.

To illustrate, 1000 evenly spaced observations were generated over the range -5 to 5. Then a B-spline basis and polynomial spline basis were constructed with knots at −2, 0, and 2. The eigenvalues for the centered $\mathbf{X}'\mathbf{X}$ matrices, excluding the last structural zero eigenvalue, are given in Table 5. In the

Table 5

Polynomial and B-Spline Eigenvalues

B-Spline Basis		
Eigenvalue	Proportion	Cumulative
0.107872	0.358718	0.35872
0.096710	0.321599	0.68032
0.046290	0.153933	0.83425
0.030391	0.101062	0.93531
0.012894	0.042878	0.97819
0.006559	0.021810	1.00000

Polynomial Spline Basis		
Eigenvalue	Proportion	Cumulative
10749.8	0.941206	0.94121
631.8	0.055317	0.99652
37.7	0.003300	0.99982
1.7	0.000148	0.99997
0.3	0.000029	1.00000
0.0	0.000000	1.00000

polynomial splines, the first two components already account for more than 99% of the variation of the points. In the B-splines, the cumulative proportion does not pass 99% until the last term. The eigenvalues show that the B-spline basis is better conditioned than the piecewise polynomial basis. B-splines are used instead of orthogonal polynomials or Box-Cox transformations because B-splines allow knots and more general curves. B-splines also allow monotonicity constraints.

A transformation of x is monotonically increasing if the coefficients that are used to combine the columns of the B-spline basis are monotonically increasing. Models with splines can be fit directly using ordinary least squares (OLS). OLS does not work for monotone splines because OLS has no method of ensuring monotonicity in the coefficients. When there are monotonicity constraints, an alternating least square (ALS) algorithm is used. Both OLS and ALS attempt to minimize a squared error loss function. See Kuhfeld (1990) for a description of the iterative algorithm that fits monotone splines. See Ramsay (1988) for some applications and a different approach to ensuring monotonicity.

Transformation Regression

If the dependent variable is not transformed and if there are no monotonicity constraints on the independent variable transformations, the transformation regression model is the same as the OLS regression model. When only the independent variables are transformed, the transformation regression model is nothing more than a different rendition of an OLS regression. All of the spline models presented up to this point can be reformulated as

$$y = \beta_0 + \Phi(x) + \epsilon$$

The nonlinear transformation of x is $\Phi(x)$; it is solved for by fitting a spline model such as

$$\begin{aligned} y \;=\; & \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \\ & \beta_4 (x > t_1)(x - t_1)^3 + \\ & \beta_5 (x > t_2)(x - t_2)^3 + \\ & \beta_6 (x > t_3)(x - t_3)^3 + \epsilon \end{aligned}$$

where

$$\begin{aligned}\Phi(x) &= \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \\ &\quad \beta_4 (x > t_1)(x - t_1)^3 + \\ &\quad \beta_5 (x > t_2)(x - t_2)^3 + \\ &\quad \beta_6 (x > t_3)(x - t_3)^3\end{aligned}$$

Consider a model with two polynomials:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_2 + \beta_5 x_2^2 + \beta_6 x_2^3 + \epsilon$$

It is the same as a transformation regression model

$$y = \beta_0 + \Phi_1(x_1) + \Phi_2(x_2) + \epsilon$$

where $\Phi_\bullet(\bullet)$ designates cubic spline transformations with no knots. The actual transformations in this case are

$$\widehat{\Phi}_1(x_1) = \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2 + \widehat{\beta}_3 x_1^3$$

and

$$\widehat{\Phi}_2(x_2) = \widehat{\beta}_4 x_2 + \widehat{\beta}_5 x_2^2 + \widehat{\beta}_6 x_2^3$$

There are six model *df*. The test for the effect of the transformation $\Phi_1(x_1)$ is the test of the linear hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$, and the $\Phi_2(x_2)$ transformation test is a test that $\beta_4 = \beta_5 = \beta_6 = 0$. Both tests are *F*-tests with three numerator *df*. When there are monotone transformations, constrained least-squares estimates of the parameters are obtained.

Degrees of Freedom

In an ordinary general linear model, there is one parameter for each independent variable. In the transformation regression model, many of the variables are used internally in the bases for the transformations. Each linearly independent basis column has one parameter and one model *df*. If a variable is not transformed, it has one parameter. Nominal classification variables with c categories have $c - 1$ parameters. For degree d splines with k knots and $d - 1$ continuous derivatives, there are $d + k$ parameters.

When there are monotonicity constraints, counting the number of scoring parameters is less precise. One way of handling a monotone spline transformation is to treat it as if it were simply a spline transformation with $d + k$ parameters. However, there are typically fewer than $d + k$ *unique* parameter estimates since some of those $d + k$ scoring parameter estimates may be tied to impose the order constraints. Imposing ties is equivalent to fitting a model with fewer parameters. So, there are two available scoring parameter counts: $d + k$ and a potentially smaller number that is determined during the analysis. Using $d + k$ as the model *df* is *conservative* since the scoring parameter estimates are restricted. Using the smaller count is too *liberal* since the data and the model together are being used to determine the number of parameters. Our solution is to report tests using both liberal and conservative *df* to provide lower and upper bounds on the "true" *p*-values.

Dependent Variable Transformations

When a dependent variable is transformed, the problem becomes multivariate:

$$\Phi_0(y) = \beta_0 + \Phi_1(x_1) + \Phi_2(x_2) + \epsilon$$

Hypothesis tests are performed in the context of a multivariate linear model, with the number of dependent variables equal to the number of scoring parameters for the dependent variable transformation. Multivariate normality is assumed even though it is known that the assumption is *always* violated. This is one reason that all hypothesis tests in the presence of a dependent variable transformation should be considered approximate at best.

For the transformation regression model, we redefine three of the usual multivariate test statistics: Pillai's Trace, Wilks' Lambda, and the Hotelling-Lawley Trace. These statistics are normally computed using all of the squared canonical correlations, which are the eigenvalues of the matrix $\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}$. Here, there is only one linear combination (the transformation) and hence only one squared canonical correlation of interest, which is equal to the R^2. We use R^2 for the first eigenvalue; all other eigenvalues are set to zero since only one linear combination is used. Degrees of freedom are computed assuming that all linear combinations contribute to the Lambda and Trace statistics, so the F-tests for those statistics are conservative. In practice, the adjusted Pillai's Trace is very conservative—perhaps too conservative to be useful. Wilks' Lambda is less conservative, and the Hotelling-Lawley Trace seems to be the least conservative.

It may seem that the Roy's Greatest Root statistic, which always uses only the largest squared canonical correlation, is the only statistic of interest. Unfortunately, Roy's Greatest Root is very liberal and only provides a lower bound on the p-value. The p-values for the liberal and conservative statistics are used together to provide approximate lower and upper bounds on p.

Scales of Measurement

Early work in scaling, such as Young, de Leeuw, & Takane (1976), and Perreault & Young (1980) was concerned with fitting models with mixed nominal, ordinal, and interval scale of measurement variables. Nominal variables were optimally scored using Fisher's (1938) optimal scoring algorithm. Ordinal variables were optimally scored using the Kruskal and Shepard (1974) monotone regression algorithm. Interval and ratio scale of measurement variables were left alone nonlinearly transformed with a polynomial transformation.

In the transformation regression setting, the Fisher optimal scoring approach is equivalent to using an indicator variable representation, as long as the correct *df* are used. The optimal scores are category means. Introducing optimal scaling for nominal variables does not lead to any increased capability in the regression model.

For ordinal variables, we believe the Kruskal and Shepard monotone regression algorithm should be reserved for the situation when a variable has only a few categories, say five or fewer. When there are more levels, a monotone spline is preferred because it uses fewer model *df* and because it is less likely to capitalize on chance.

Interval and ratio scale of measurement variables can be left alone or nonlinearly transformed with splines or monotone splines. When the true model has a nonlinear function, say

$$y = \beta_0 + \beta_1 log(x) + \epsilon$$

or

$$y = \beta_0 + \beta_1/x + \epsilon$$

the transformation regression model

$$y = \beta_0 + \Phi(x) + \epsilon$$

can be used to hunt for parametric transformations. Plots of $\widehat{\Phi}(x)$ may suggest log or reciprocal transformations.

Conjoint Analysis

Green & Srinivasan (1990) discuss some of the problems that can be handled with a transformation regression model, particularly the problem of degrees of freedom. Consider a conjoint analysis design where a factor with $c > 3$ levels has an inherent ordering. By finding a quadratic monotone spline transformation with no knots, that variable will use only two *df* instead of the larger $c - 1$. The model *df* in a spline transformation model are determined by the data analyst, not by the number of categories in the variables. Furthermore, a "*quasi-metric*" conjoint analysis can be performed by finding a monotone spline transformation of the dependent variable. This model has fewer restrictions than a metric analysis, but will still typically have error *df*, unlike the nonmetric analysis.

Curve Fitting Applications

With a simple regression model, you can fit a line through a $y \times x$ scatter plot. With a transformation regression model, you can fit a curve through the scatter plot. The y-axis coordinates of the curve are

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\Phi}(x)$$

from the model

$$y = \beta_0 + \Phi(x) + \epsilon$$

With more than one group of observations and a multiple regression model, you can fit multiple lines, lines with the same slope but different intercepts, and lines with common intercepts but different slopes. With the transformation regression model, you can fit multiple curves through a scatter plot. The curves can be monotone or not, constrained to be parallel, or constrained to have the same intercept. Consider the problem of modeling the number of product purchases as a function of price. Separate curves can be simultaneously fit for two groups who may behave differently, for example those who are making a planned purchase and those who are buying impulsively. Later in this chapter, there is an example of plotting brand by price interactions.

Figure 8 contains an artificial example of two separate spline functions; the shapes of the two curves are independent of each other, and $R^2 = 0.87$. In Figure 9, the splines are constrained to be parallel, and $R^2 = 0.72$. The parallel curve model is more restrictive and fits the data less well than the unconstrained model. In Figure 8, each curve follows its swarm of data. In Figure 9, the curves find paths through the data that are best on the average considering both swarms together. In the vicinity of $x = -2$, the top curve is high and the bottom curve is low. In the vicinity of $x = 1$, the top curve is low and the bottom curve is high.

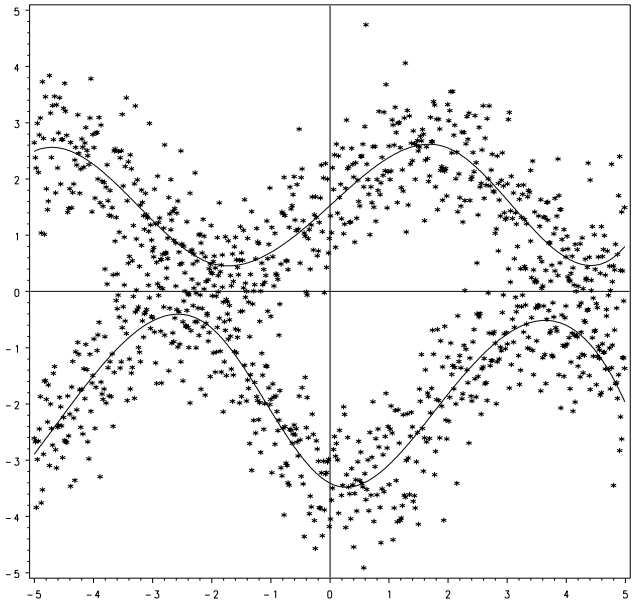


Figure 8. Separate Spline Functions, Two Groups

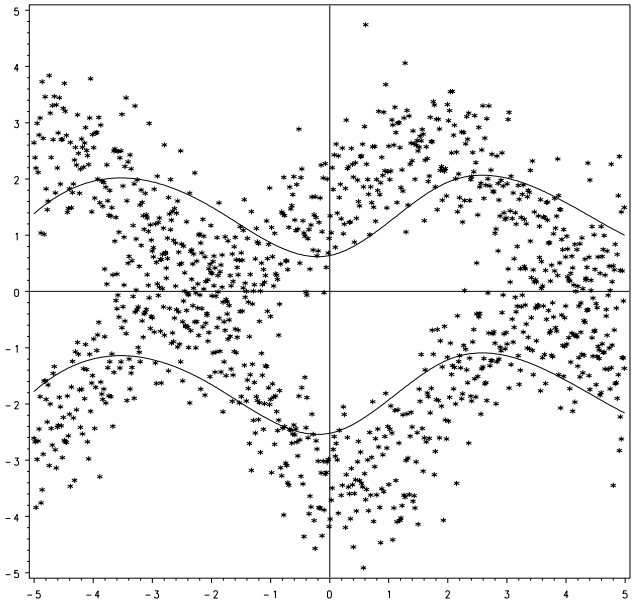


Figure 9. Parallel Spline Functions, Two Groups

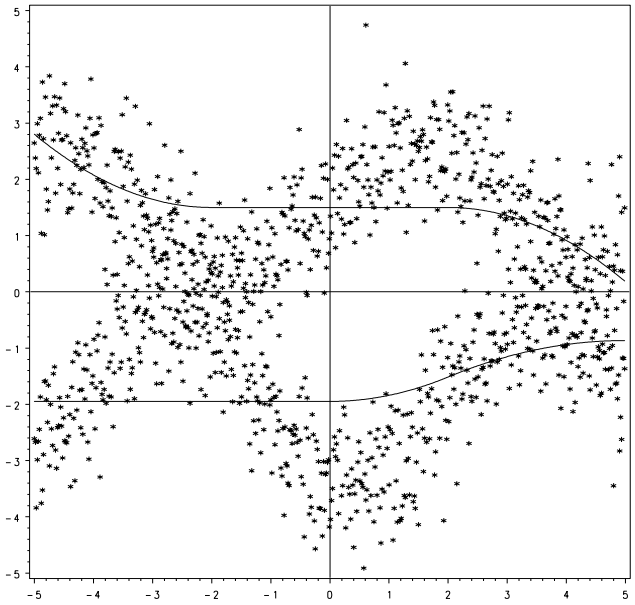


Figure 10. Monotone Spline Functions, Two Groups

Figure 10 contains the same data and two monotonic spline functions; the shapes of the two curves are independent of each other, and $R^2 = 0.71$. The top curve is monotonically decreasing, whereas the bottom curve is monotonically increasing. The curves in Figure 10 flatten where there is nonmonotonicity in Figure 8.

Parallel curves are very easy to model. If there are two groups and the variable g is a binary variable indicating group membership, fit the model

$$y = \beta_0 + \beta_1 g + \Phi_1(x) + \epsilon$$

where $\Phi_1(x)$ is a linear, spline, or monotone spline transformation. Plot $\widehat{y}$ as a function of x to see the two curves. Separate curves are almost as easy; the model is

$$y = \beta_0 + \beta_1 g + \Phi_1(x \times (1 - g)) + \Phi_2(x \times g) + \epsilon$$

When $x \times (1 - g)$ is zero, $x \times g$ is x, and vice versa.

Spline Functions of Price

This section illustrates splines with an artificial data set. Imagine that subjects were asked to rate their interest in purchasing various types of spaghetti sauces on a one to nine scale, where nine indicated definitely will buy and one indicated definitely will *not* buy. Prices were chosen from typical retail trade prices, such as \$1.49, \$1.99, \$2.49, and \$2.99; and one penny more than a typical price, \$1.00, \$1.50, \$2.00, and \$2.50. Between each "round" number price, such as \$1.00, and each typical price, such as \$1.49, three additional prices were chosen, such as \$1.15, \$1.25, and \$1.35. The goal is to allow a model with a separate spline for each of the four ranges: \$1.00 — \$1.49, \$1.50 — \$1.99, \$2.00 — \$2.49, and \$2.50 — \$2.99. For each range, a spline with zero or one knot can be fit.

One rating for each price was constructed and various models were fit to the data. Figures 11 through 18 contain results. For each figure, the number of model *df* are printed. One additional *df* for the intercept is also used. The SAS/STAT procedure TRANSREG was used to fit all of the models in this chapter.

Figure 11 shows the linear fit, Figure 12 uses a quadratic polynomial, and Figure 13 uses a cubic polynomial. The curve in Figure 13 has a slight nonmonotonicity in the tail, and since it is a polynomial, it is rigid and cannot locally fit the data values.

Figure 14 shows a monotone spline. It closely follows the data and never increases. A range for the model *df* is specified; the larger value is a conservative count and the smaller value is a liberal count.

The curves in Figures 12 through 14 are all continuous and smooth. These curves do a good job of following the data, but inspection of the data suggests that a different model may be more appropriate. There is a large drop in purchase interest when price increases from \$1.49 to \$1.50, a smaller drop between \$1.99 and \$2.00, and a still smaller drop between \$2.49 and \$2.50.

In Figure 15, a separate quadratic polynomial is fit for each of the four price ranges: \$1.00 — \$1.49, \$1.50 — \$1.99, \$2.00 — \$2.49, and \$2.50 — \$2.99. The functions are connected. The function over the range \$1.00 — \$1.49 is nearly flat; there is high purchase interest for all of these prices. Over the range \$1.50 — \$1.99, purchase interest drops more rapidly with a slight leveling in the low end; the slope decreases as the function increases. Over the range \$2.00 — \$2.49, purchase interest drops less

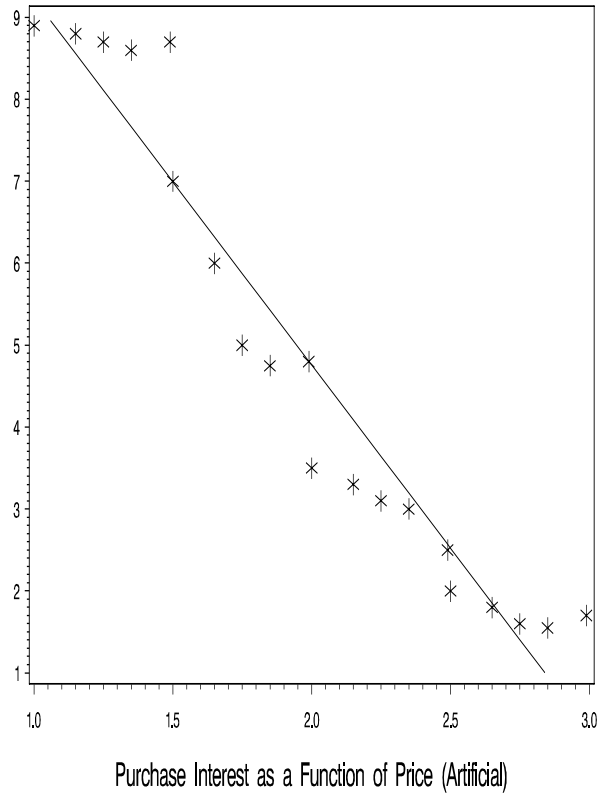


Figure 11. Linear Function, 1 df

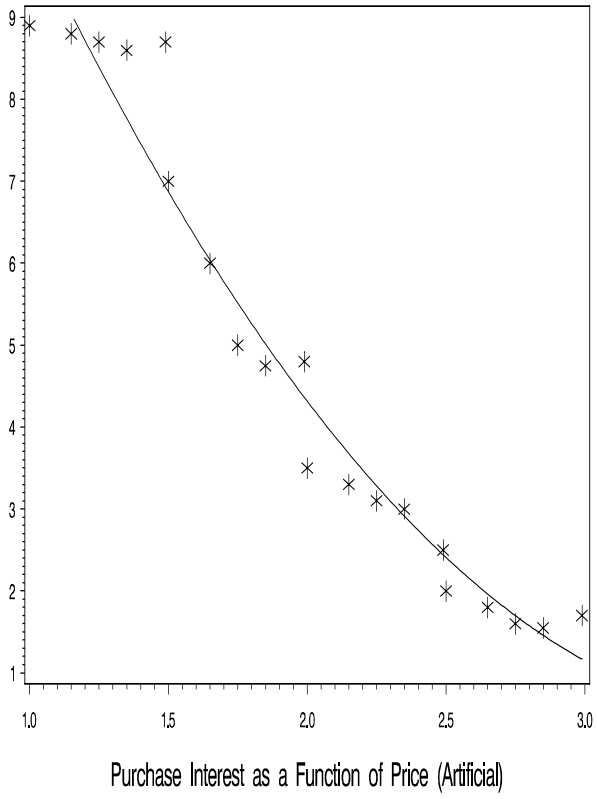


Figure 12. Quadratic Function, 2 df

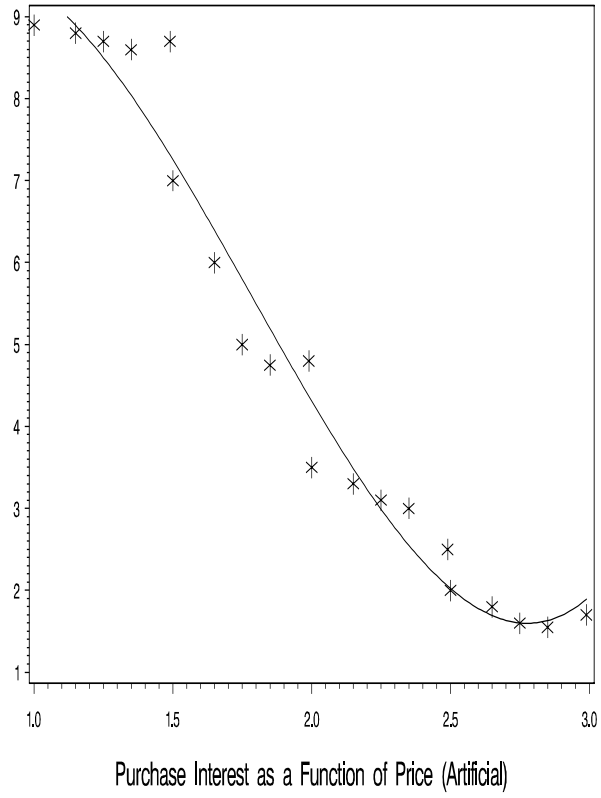


Figure 13. Cubic Function, 3 df

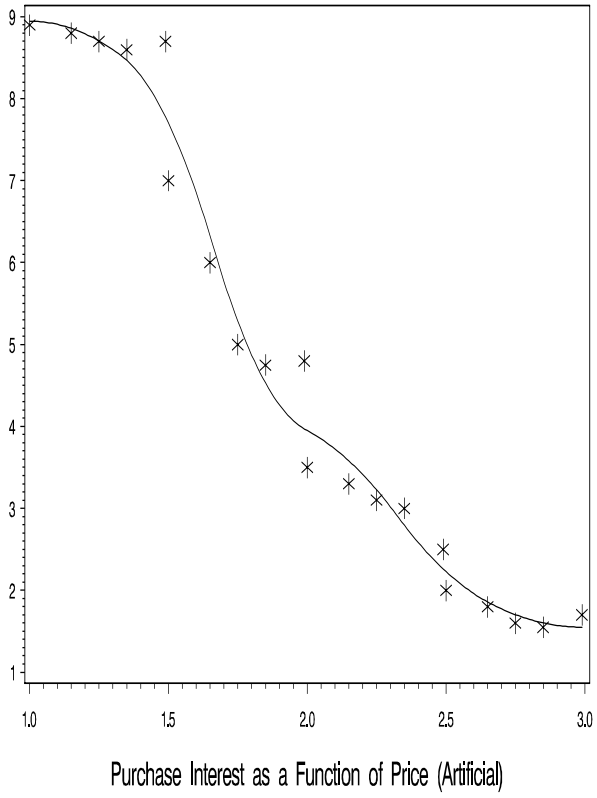


Figure 14. Monotone Function, 5–7 df

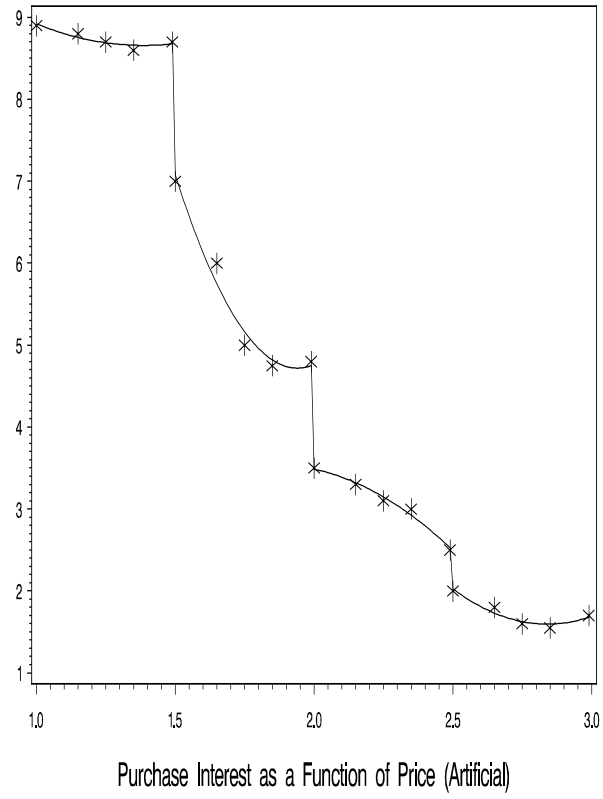


Figure 15. Discontinuous Polynomial, 11 df

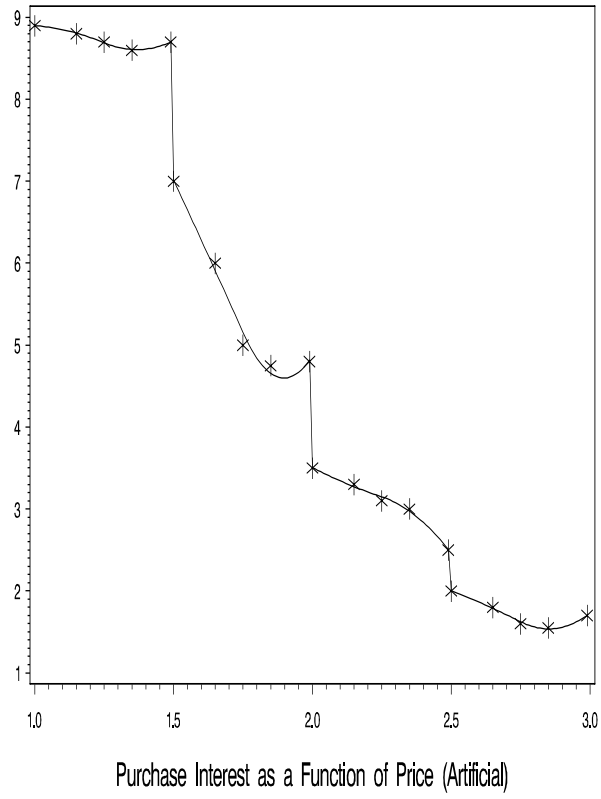


Figure 16. Discontinuous Spline Function, 15 df

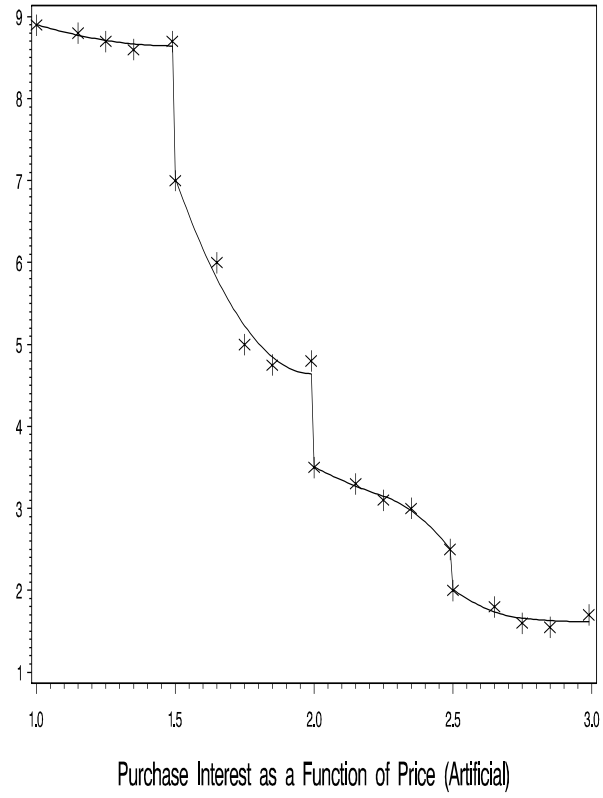


Figure 17. Discontinuous Monotone Spline, 12–15 df

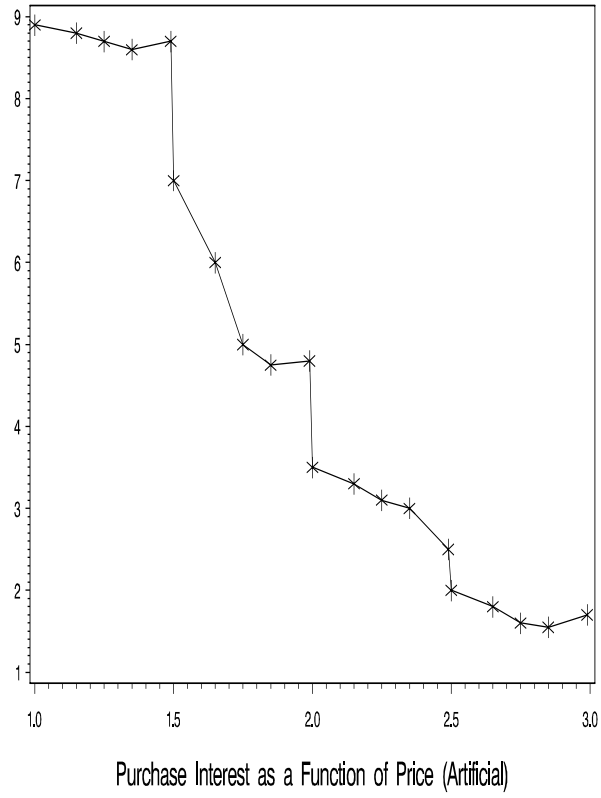


Figure 18. Cell Means, 19 df

rapidly; the slope increases as the function increases. Over the range \$2.50 — \$2.99, the function is nearly flat. At \$1.99, \$2.49, and \$2.99 there is a slight increase in purchase interest.

In Figure 16, there is a knot in the middle of each range. This gives the spline more freedom to follow the data. Figure 17 uses the same model as Figure 16, but monotonicity is imposed. When monotonicity is imposed the curves touch fewer of the data values, passing in between the nonmonotonic points. In Figure 18, the means for each price are plotted and connected. This analysis uses the most model *df* and is the least smooth of the plots.

Benefits of Splines

In marketing research and conjoint analysis, the use of spline models can have several benefits. Whenever a factor has three or more levels and an inherent ordering of the levels, that factor can be modeled as a quadratic monotone spline. The *df* used by the variable is controlled at data analysis time; it is not simply the number of categories minus one. When the alternative is a model in which a factor is designated as nominal, splines can be used to fit a more restrictive model with fewer model *df*. Since the spline model has fewer *df*, it should yield more reproducible results.

The opposite alternative is also important. Consider a variable with many values, like price in some examples. Instead of using a restrictive single *df* linear model, splines can be used to fit a more general model with more *df*. The more general model may show information in the data that is not apparent from the ordinary linear model. This can be a benefit in conjoint analyses that focus on price, in the analysis of scanner data, and in survey research. Splines give you the ability to examine the nonlinearities that may be very important in predicting consumer behavior.

Fitting quadratic and cubic polynomial models is common in marketing research. Splines extend that capability by adding the possibility of knots and hence more general curves. Spline curves can also be restricted to be monotone. Monotone splines are less restrictive than a line and more restrictive than splines that can have both positive and negative slopes. You are no longer restricted to fitting just a line, polynomial, or a step function. Splines give you possibilities in between.

Conclusions

Splines allow you to fit curves to your data. Splines may not revolutionize the way you analyze data, but they will provide you with some new tools for your data analysis toolbox. These new tools allow you to try new methods for solving old problems and tackle new problems that could not be adequately solved with your old tools. We hope you will find these tools useful, and we hope that they will help you to better understand your marketing data.

Graphical Scatter Plots of Labeled Points

Warren F. Kuhfeld

Abstract

The %PlotIt (PLOT ITeratively) macro creates graphical scatter plots of labeled points. It is designed to make it easy to display raw data, regressions, and the results of many other data analyses. You can draw curves, vectors, and circles, and you can control the colors, sizes, fonts, and general appearance of the plots. The %PlotIt macro is a part of the SAS autocall library.*

Introduction

SAS has provided software for producing scatter plots for many years (for example, the PLOT and GPLOT procedures). For many types of data analyses, it is useful to have each point in the plot labeled with the value of a variable. However, before the creation of the %PlotIt macro, there was not a satisfactory way to do this. PROC GPLOT produces graphical scatter plots. Combined with the Annotate facility, it allows long point labels, but it does not provide any way to optimally position them. The PLOT procedure can optimally position long point labels in the scatter plot, however PROC PLOT cannot create a graphical scatter plot. The PROC PLOT label-placement algorithm was developed by Kuhfeld (1991), and the PROC PLOT options are documented in Base SAS documentation.

The macro, %PlotIt (PLOT ITeratively), creates graphical scatter plots of labeled points. It can fit curves, draw vectors, and draw circles. It has many options, but only a small number are needed for many types of plots. The %PlotIt macro uses DATA steps and multiple procedures, including PLOT and GANNO. The %PlotIt macro is over 5700 lines long, so it is not printed here. It is fully documented in the header comments. This article illustrates through examples some of the main features of the %PlotIt macro.

*This chapter originally appeared in the SAS Journal **Observations**, Fourth Quarter, 1994, pages 23−37. This version of the chapter has been updated for SAS Version 8.2. Copies of this chapter (TS-689J) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market . Color plots are available on the web in a separate file.

An Overview of the %PlotIt Macro

The `%PlotIt` macro performs the following steps.

1. It reads an input data set and preprocesses it. The preprocessed data set contains information such as the axis variables, the point-symbol and point-label variables, and symbol and label types, sizes, fonts, and colors. The nature of the preprocessing depends on the type of data analysis that generated the input data set. For example, if the option DATATYPE=MDPREF was specified with an input data set created by PROC PRINQUAL for a multidimensional preference analysis, then the `%PlotIt` macro creates blue points for `_TYPE_ = 'SCORE'` observations and red vectors for `_TYPE_ = 'CORR'` observations.
2. A DATA step, using the DATA Step Graphics Interface, determines how big to make the graphical plot.
3. PROC PLOT determines where to position the point labels. The results are sent to output SAS data sets using ODS. By default, if some of the point label characters are hidden, the `%PlotIt` macro recreates the printer plot with a larger line and page size, and hence creates more cells and more room for the labels.
4. The PROC PLOT output data sets are read, and information from them and the preprocessed data set are combined to create an Annotate data set. The label position information is read from the PROC PLOT output, and all of the symbol, size, font, and color information is extracted from the preprocessed data set. The Annotate data set contains all of the instructions for drawing the axes, ticks, tick marks, titles, point symbols, point labels, axis labels, and so on.
5. The Annotate data set is displayed with the GANNO procedure. The `%PlotIt` macro does not use PROC GPLOT.

With the `%PlotIt` macro, you can:

- display plots and create graphics stream files and `gout=` entries
- easily display the results of correspondence analysis, multidimensional preference analysis, preference mapping, multidimensional scaling, regression analysis, and density estimation
- use single or multi-character symbols and control their color, font, and size
- use multi-character point labels and control their color, font, and size
- use fixed, variable, and random colors, and use colors to display changes in a third dimension
- automatically determine a good line size, page size, and list of point label placements
- automatically equate the axes for all devices
- control the colors, sizes, fonts, and general appearance of all aspects of the plot
- pre- and post-process the data
- specify many `goptions`

Since %PlotIt is a macro, you can modify it, change the defaults, add new options, and so on. The %PlotIt macro is heavily commented to make it easier for you to modify it to suit your needs. There is substantial error checking and options to print intermediate results for debugging when you do not get the results you expect. Furthermore, you have complete access to all of the data sets it creates, including the preprocessed version of the input and the Annotate data set. You can modify the results by editing the Annotate and preprocessed data sets.

Examples

This section contains examples of some of the capabilities of the %PlotIt macro. Rather than interpreting the plots or discussing the details of the statistical analyses, this section concentrates on showing what the %PlotIt macro can do. Most of the examples are based on SAS/STAT example data sets. Data for all of the examples can be found in the SAS/STAT sample program plotitex.sas.

Example 1: Principal Components of Mammal's Teeth
Principal component analysis computes a low-dimensional approximation to a set of data. Principal components are frequently displayed graphically. This example is based on the Mammal's Teeth data set. To perform a principal component analysis, specify:

```
proc princomp data=teeth out=scores(keep=prin1 prin2 mammal);
   title "Principal Components of Mammals' Teeth";
   run;

%plotit()
```

The plot is shown in Figure 1. No options were specified in the %PlotIt macro, so by default a plot is constructed from the first two numeric variables and the last character variable in the last data set created. The %PlotIt macro printed the following information to the log:

```
                Iterative Scatter Plot of Labeled Points Macro

            Iteration    Place    Line Size    Page Size    Penalty
            ------------------------------------------------------
                1          2         65           45           34
                2          3         80           50            0

   The following code will create the printer plot on which the graphical
   plot is based:

   options nonumber ls=80 ps=50;
   proc plot nolegend formchar='|----|+|---' data=preproc vtoh=2;
      plot Prin2 * Prin1 $ mammal = _symbol_ /
           haxis=by 1 vaxis=by 1 box list=1
           placement=((h=2 -2 : s=right left) (v=1 to 2 by alt * h=0 -1 to -10
           by alt));
      label Prin2 = '#' Prin1 = '#';
      run; quit;
```

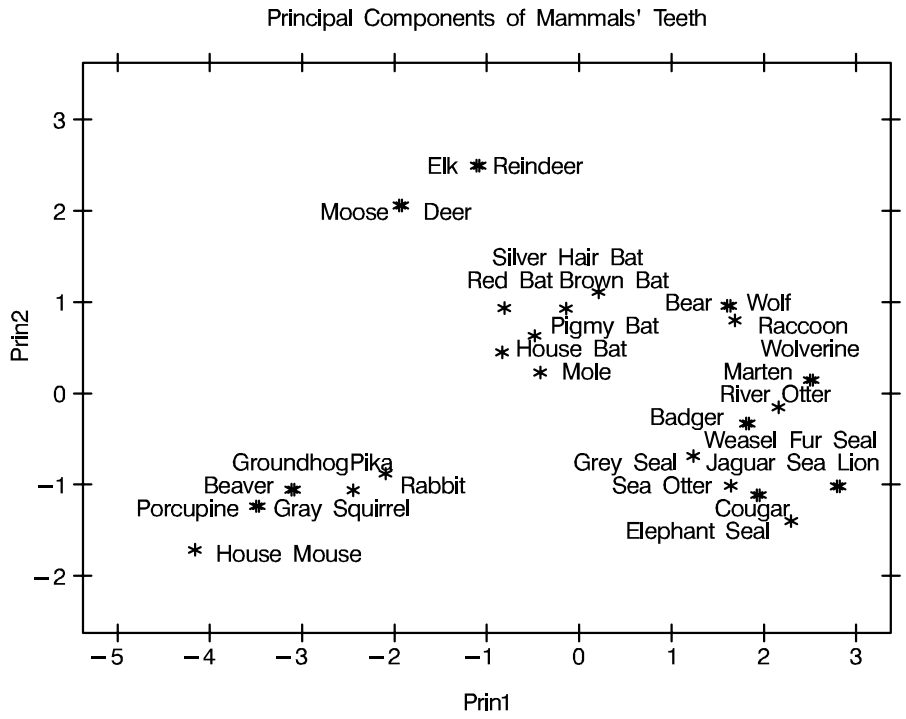


Figure 1

```
The plot was created with the following goptions:

goptions reset=goptions erase hpos=99 vpos=34 hsize=11.71in vsize=7.98in
         device=XCOLOR;

The OUT=anno Annotate data set has 148 observations.
The PLOTIT macro used 8.4 seconds to create OUT=anno.
```

The iteration table shows that the `%PlotIt` macro tried twice to create the plot, with line sizes of 65 and 80. It stopped when all point label characters were plotted with zero penalty.[†] The `%PlotIt` macro displays PROC PLOT code for the printer plot, on which the graphical plot is based. It also displays the `goptions` statement that was used with PROC GANNO.[‡]

[†]Penalties accrue when point labels are nonoptimally placed, such as when two label characters map to the same location. PROC PLOT tries to minimize the penalties for all point labels. See PROC PLOT documentation for more information.

[‡]This code could be different depending on your device. By default, the macro uses your default device.

There are several notable features of the plot in Figure 1.

1. Symbols for several pairs of points, such as Elk and Reindeer, are coincident. By default, the `%PlotIt` macro slightly offsets coincident symbols so that it is clear that more than one point maps to the same location.
2. Point labels map into discrete rows, just as they would in a printer plot produced by PROC PLOT. However, unlike printer plots, the `%PlotIt` macro uses proportional fonts.
3. Symbols are not restricted to fixed cells. Their mapping is essentially continuous, more like PROC GPLOT's than PROC PLOT's.
4. A fixed distance represents the same data range along both axes, which means the axes are equated so that distances and angles will have meaning. In contrast, procedures such as PLOT and GPLOT fill the available space by default, so the axes are not equated.

Example 2: Principal Components of Crime Rates
A typical plot has for each point a single-character symbol and a multi-character label; however, this is not required. This example is based on the Crime data set. The point labels are state names, and the symbol for each label is a two-character postal code.

```
proc princomp data=crime out=crime2;
   title 'Crime Rates Per 100,000 Population by State';
   run;

%plotit(data=crime2,plotvars=prin2 prin1,
        symvar=postcode,symlen=2,symsize=0.6,paint=larceny,
        labelvar=state,label=typical)
```

This plot request specifies:
- the input data set: `crime2`
- the y-axis and x-axis variables: `prin2` and `prin1`
- the symbol variable: `postcode`
- the number of symbol characters: 2
- the size of the symbol font in the plot: 0.6
- the colors are based on the variable: `larceny`
- the point label variable: `state`
- the typical method of generating variable labels for the plot axes

A symbol size of 0.6 instead of the normal 1.0 is specified to make the symbol small, because two characters are mapping to a location where there is usually just one. The option `paint=larceny` creates interpolated label and symbol colors, by default between blue, magenta, and red, so that states that have a low larceny rate are blue and high-rate states are red.[§] `Label=typical` for variables `prin2` and `prin1` generates the following label statement:

```
label prin2 = 'Dimension 2' prin1 = 'Dimension 1';
```

This plot request is much more complicated than most. Often, you need to specify only the type of analysis that generated the data set.

The plot is shown in Figure 2.

[§]If you are viewing a black and white version of this chapter, all colors for all plots will be shown in black or shades of gray.

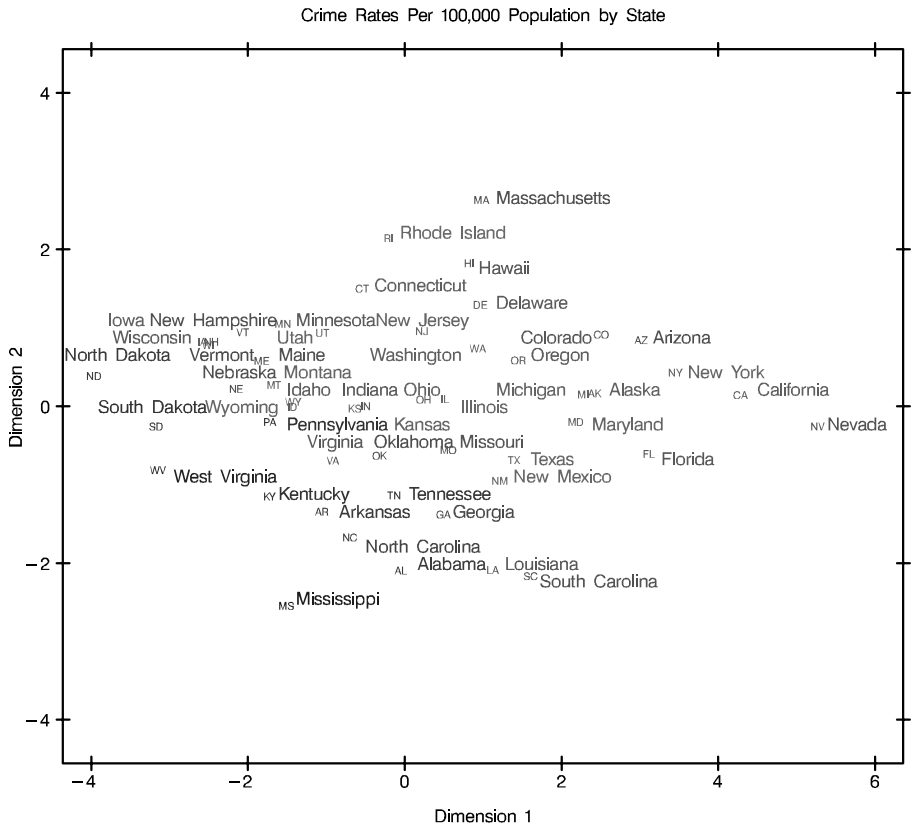


Figure 2

Examples 3 & 4: Correspondence Analysis of the Car Ownership Survey

Correspondence analysis graphically displays crosstabulations. These examples use the Car Survey data. To perform a correspondence analysis, specify:

```
proc corresp data=cars outc=coors;
   title 'Car Owners and Car Origin';
   tables marital, origin;
   run;

%plotit(data=coors,datatype=corresp)
```

The plot is shown in Figure 3. With `datatype=corresp`, the `%PlotIt` macro automatically incorporates the proportion of inertia[¶] into the axis labels and plots the row points in red and the column points in blue.

[¶]Inertia in correspondence analysis is analogous to variance in principal component analysis.

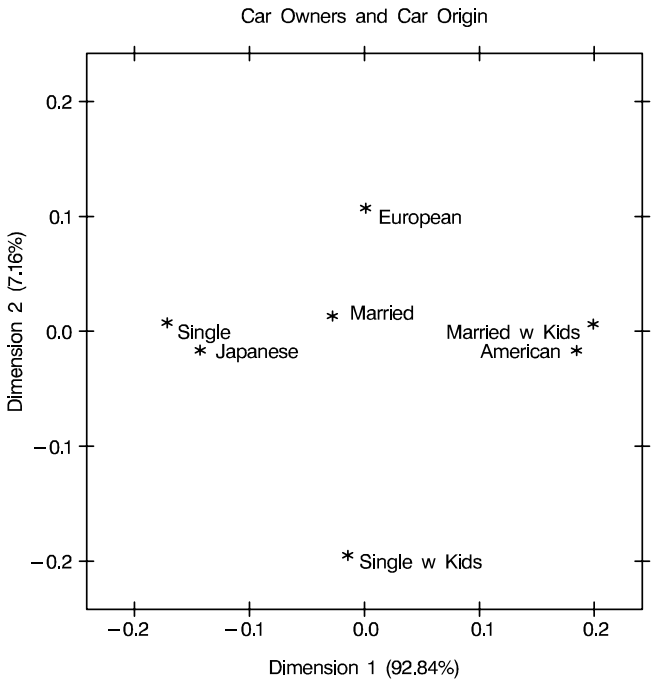


Figure 3

For a multiple correspondence analysis, specify:

```
proc corresp mca observed data=cars outc=coors;
   title 'MCA of Car Owners and Car Origin';
   tables origin size type income home marital sex;
   run;

%plotit(data=coors,datatype=mca)
```

The plot is shown in Figure 4.

Examples 5 & 6: Multidimensional Preference Analysis of Recreational Activities
Multidimensional preference analysis is a variant on principal component analysis that simultaneously displays people and their preferences for objects. Each person is a variable in the input data set, and each object is a row. Each person is represented in the plot as a vector that points in approximately the direction of his or her most preferred objects. These examples use the Preferences for Recreational Activities data set. For multidimensional preference analysis, specify:

```
proc prinqual cor data=recreate out=rec score std rep;
   title1 'Multidimensional Preference Analysis of Recreational Activities';
   transform identity(sub1-sub56);
   id activity;
   run;

%plotit(data=rec,datatype=mdpref 3)
```

The plot is shown in Figure 5. With `datatype=mdpref`, the `%PlotIt` macro automatically displays the people as vectors and the activities as points (based on the `_TYPE_` variable). The `3` after the MDPREF

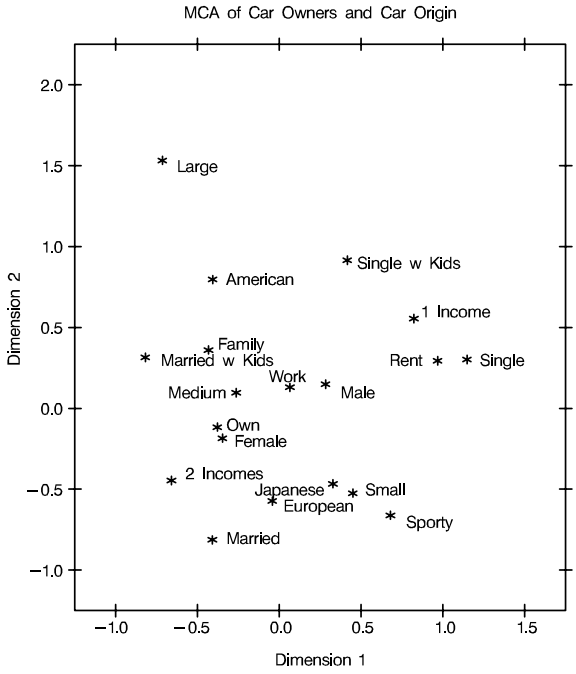


Figure 4

is a scaling factor for the vectors. The lengths of all the vectors are increased by a factor of three to create a better graphical display. You can also label the vectors by specifying:

```
%plotit(data=rec,datatype=mdpref2 3)
```

MDPREF2 specifies a MDPREF analysis with labeled vectors (the `2` means labels *too*). This plot is not shown because in this input data set, each subject is identified by a variable name of the form `sub1`, `sub2`, ..., and the graphical display looks cluttered with all those `sub`'s. The default point label variable is the ID statement variable `activity`, because it is the last character variable in the data set. PROC PRINQUAL fills in this variable for the `_TYPE_ = 'CORR'` observations (the people that plot as vectors) with the variable names: `sub1-sub56`. You can preprocess the input data set directly in the `%PlotIt` macro to remove the `sub`'s as follows:

```
%plotit(data=rec,datatype=mdpref2 3,
        adjust1=%str(if _type_ = 'CORR' then
                       activity = substr(activity,4);))
```

The plot is shown in Figure 6. The `adjust1` option adds DATA step statements to the end of the preprocessing step. By default, the `%PlotIt` macro tries to position the vector labels outward, not between the vector head and the origin.

Examples 7 & 8: Preference Mapping of Cars

Preference mapping simultaneously displays objects and attributes of those objects. These examples use the Car Preference data set to illustrate preference mapping. The following code fits a preference mapping vector model:

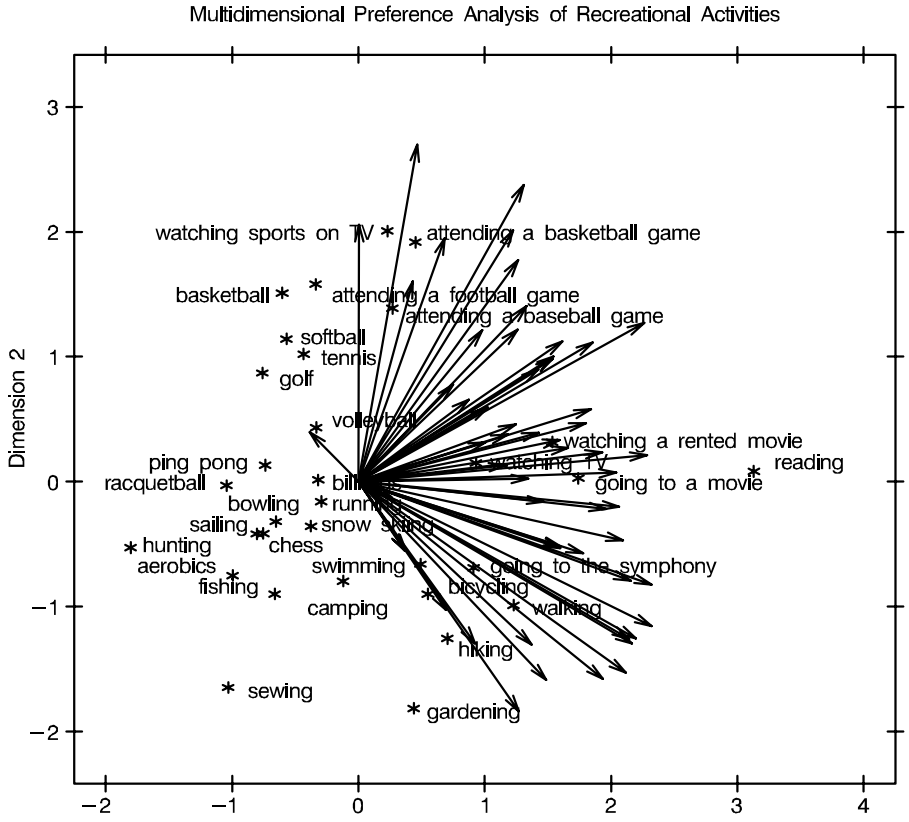


Figure 5

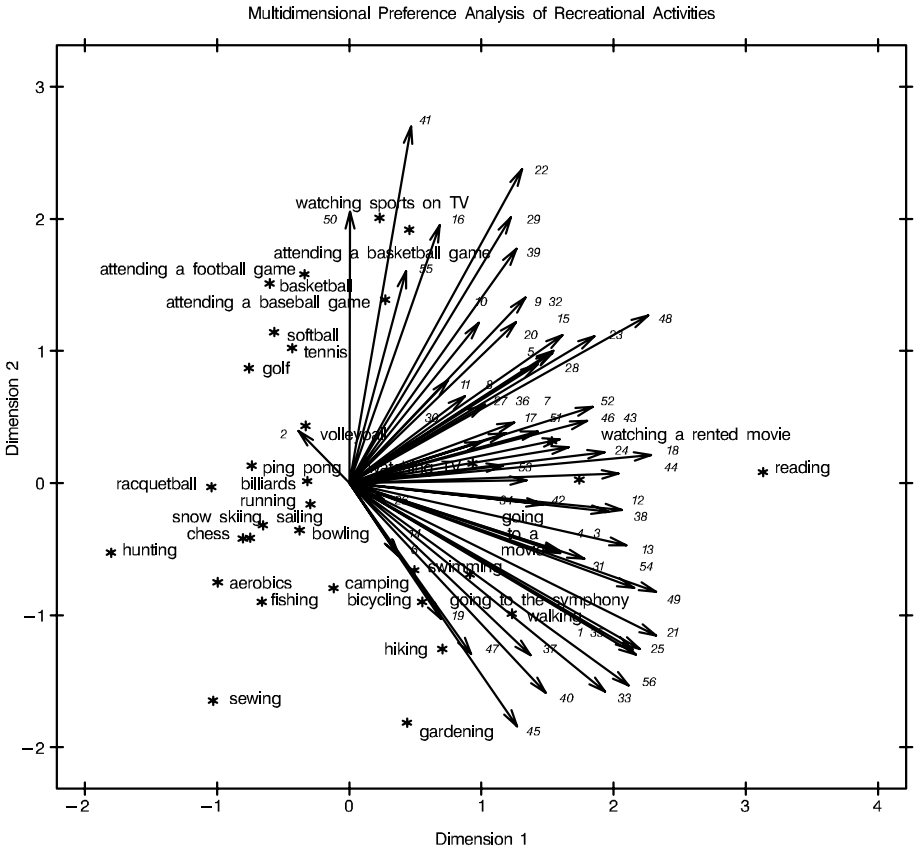


Figure 6

```
*---Compute Coordinates for a 2-Dimensional Scatter plot of Cars---;
proc prinqual data=carpref out=presults(drop=judge1-judge25) n=2
              replace standard scores;
   title 'Preference Ratings for Automobiles Manufactured in 1980';
   id model mpg reliable ride;
   transform ide(judge1-judge25);
   run;
*---Compute Endpoints for Vectors---;
proc transreg data=presults;
   title2 'Preference Mapping, Vector Model';
   model ide(mpg reliable ride)=identity(prin1 prin2);
   output tstandard=center coefficients replace out=vector;
   id model;
   run;
%plotit(data=vector,datatype=vector 2.5)
```

The plot is shown in Figure 7. Each attribute is represented as a vector that points in approximately the direction of the objects with larger values of the attribute. The `datatype=vector 2.5` option requests the vector model, with the vectors stretched by a factor of 2.5. Alternatively, you can represent attributes as points by specifying:

```
*---Compute Ideal Point Coordinates---;
proc transreg data=presults;
   title2 'Preference Mapping, Ideal Point Model';
   model identity(mpg reliable ride)=point(prin1 prin2);
   output tstandard=center coordinates replace out=ideal;
   id model;
   run;

%plotit(data=ideal,datatype=ideal,antiidea=1)
```

The plot is shown in Figure 8. The option `datatype=ideal` requests a preference mapping, with each attribute represented as an ideal point. Circles are drawn to show distances between the cars and the ideal points, which are locations of hypothetical cars that have the ideal amount of the attribute. The `antiidea=1` option specifies how anti-ideal points are recognized.[‖] By default, the labels for the attributes are larger than the other point labels and hence sometimes extend slightly beyond the plot. This happened with "Miles per gallon" in Figure 8. You can move the label up one character unit and to the left 12 character units by adding the following option:

```
adjust4=%str(if text =: 'Miles' then do; y = y + 1; x = x - 12; end;)
```

The `adjust4` option adds DATA step statements to the end of the final Annotate DATA step. The `%PlotIt` macro has no sense of esthetics; sometimes a little human intervention is needed for the final production plots.

Examples 9 & 10: Curve Fitting of Birth and Death Rates

It is often useful to display a set of points along with a regression line or nonlinear function. The `%PlotIt` macro can fit and display lines and curves (and optionally print the regression and ANOVA table). These examples use the Vital Statistics data set. The following requests a cubic-polynomial regression function:

[‖] Anti-ideal points have their signs wrong, so the macro must reverse them before plotting. When small ratings are good, specify `antiidea=-1`, and when small ratings are not good, specify `antiidea=1`.

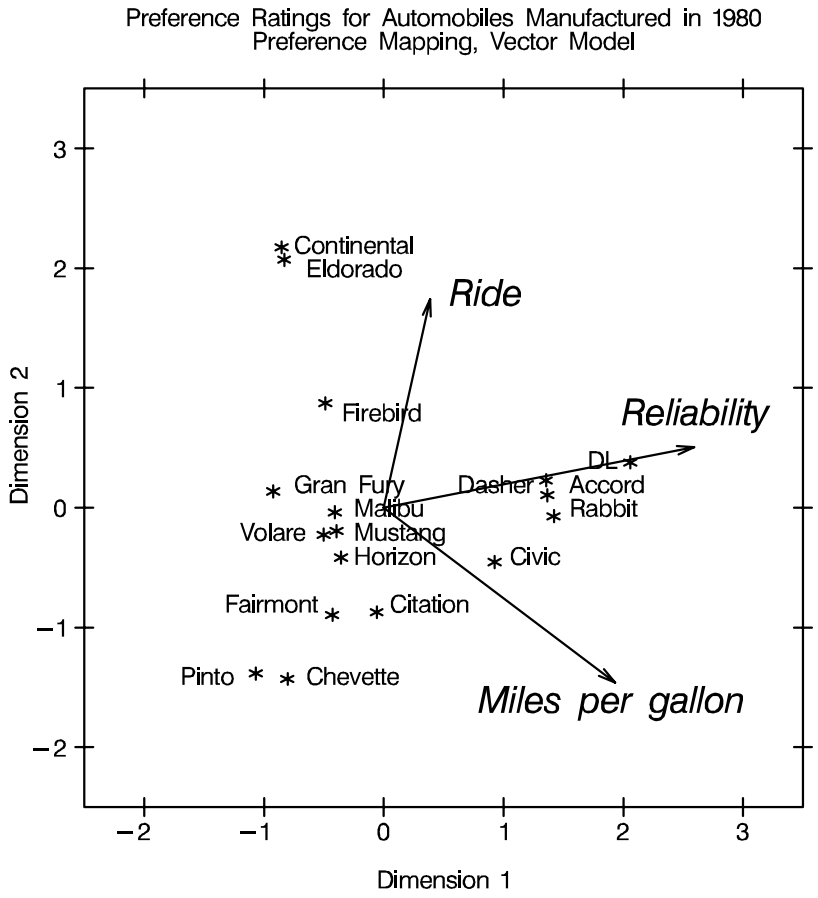


Figure 7

```
title 'Crude Death Rate as a Function of Crude Birth Rate';

%plotit(data=vital,vtoh=1.75,datatype=curve2)
```

The plot is shown in Figure 9. The option `vtoh=1.75` specifies the PROC PLOT aspect ratio (vertical to horizontal). The default is 2.0. Smaller values create plots with more cells for label characters, which is helpful when the point cloud is relatively dense. The option `datatype=curve2` instructs the `%PlotIt` macro to fit a curve and have the point labels avoid the curve (the `2` means label avoidance *too*).

You can control the type of curve. The `%PlotIt` macro uses PROC TRANSREG to fit the curve, and you can specify PROC TRANSREG options. For example, to request a monotone spline regression function with two knots, specify:

```
%plotit(data=vital,datatype=curve,bright=128,maxiter=4,
        symvar=country,regfun=mspline,nknots=2)
```

The plot is shown in Figure 10. There are several differences between Figures 9 and 10, in addition to the difference in the regression function. The option `datatype=curve` was specified, not `datatype=curve2`, so there is more overlap between the point labels and the curve. For each point in the plot, the plotting symbol is the first letter of the country and the point label is the country. Each label/symbol pair is

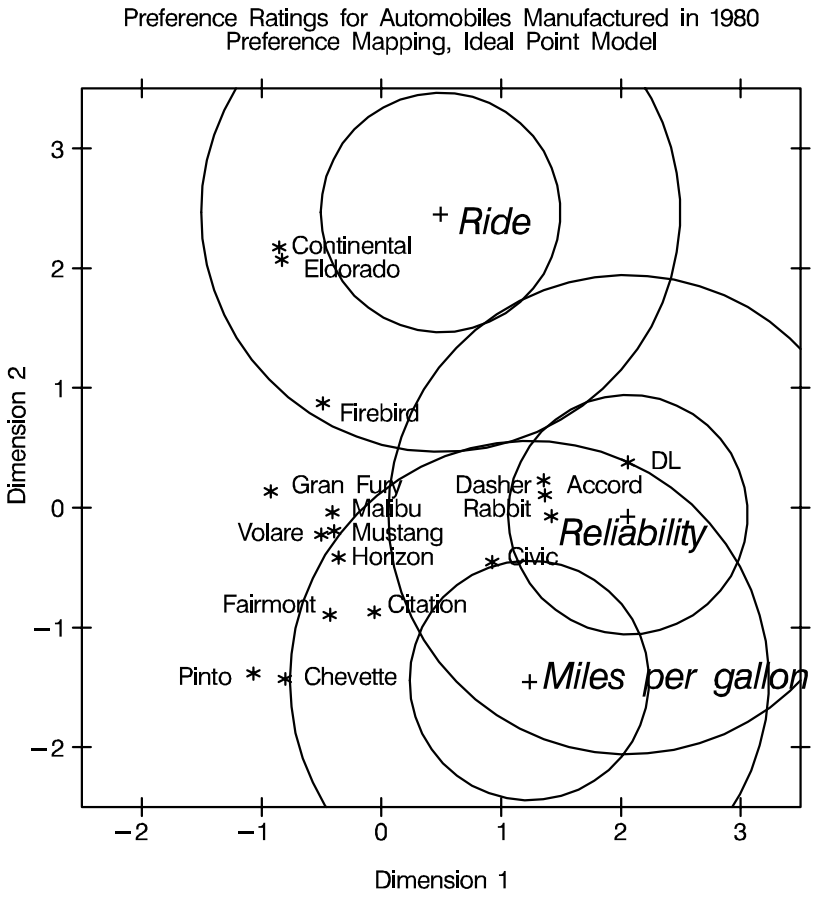


Figure 8

a different random color with brightness (average RGB or red-green-blue value) of 128. These options make it much easier to find the symbol that corresponds to each label. Also, the default `vtoh=2` was used to decrease the number of cells and make the labels larger. With these data, the penalty sum at iteration four is eight. Specifying `maxiter=4` prevents the algorithm from reaching iteration 5, which prevents the line size from increasing from 125 to 150. This also makes the labels larger. The price is that some label characters collide (for example, "Germany" and "S") and the plot looks more cluttered because there are fewer cells with white space.

Availability

If your site has installed the autocall libraries supplied by SAS and uses the standard configuration of SAS supplied software, you need only to ensure that the SAS system option `mautosource` is in effect to begin using autocall macros. That is, the macros do *not* have to be included (for example with a `%include` statement). They can be called directly. For more information about autocall libraries, refer to *SAS Macro Language: Reference* and page 479. On a PC for example, the autocall library may be installed in the `stat\sasmacro` directories. On MVS, each macro will be a different member of a PDS. For details on installing autocall macros, consult your host documentation.

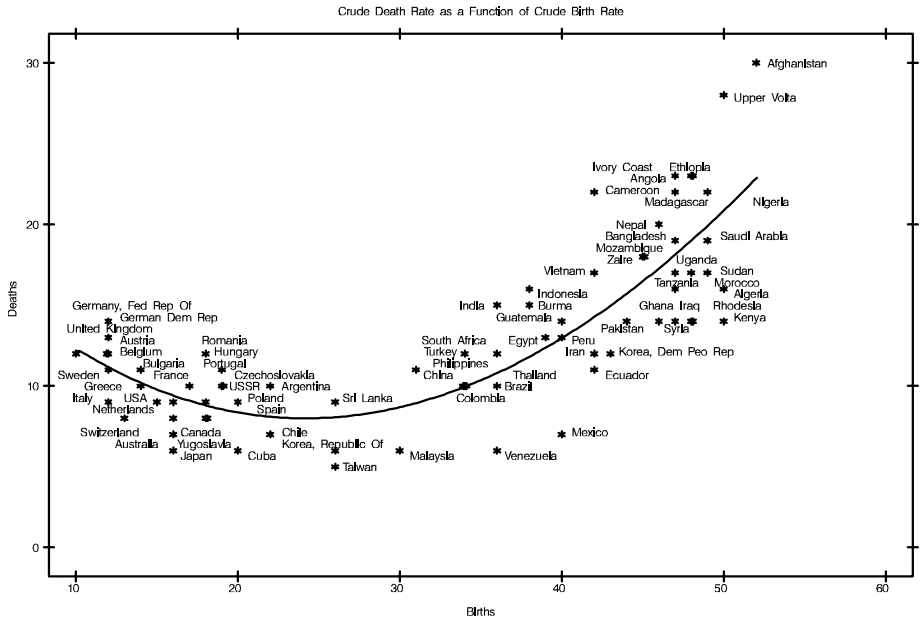


Figure 9

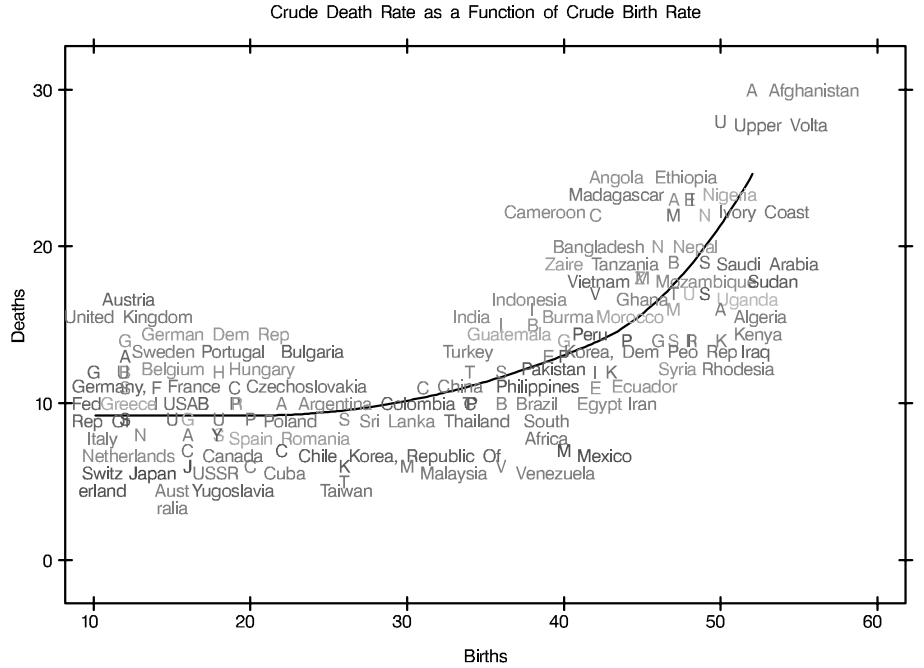


Figure 10

Base SAS and SAS/GRAPH software are required to run the `%PlotIt` macro. The `datatype=curve` and `datatype=curve2` options use PROC TRANSREG, which is in SAS/STAT. All of the other `datatype=` values assume an input data set in a form created by a SAS/STAT procedure.

Conclusions

The `%PlotIt` macro provides a convenient way to display the results from many types of data analyses. Usually, only a small number of options are needed; the macro does the rest. The `%PlotIt` macro does not replace procedures like GPLOT and PLOT. Instead, it makes it easy to generate many types of plots that are extremely difficult to produce with standard procedures.

Graphical Methods for Marketing Research

Warren F. Kuhfeld

Abstract

Correspondence analysis, multiple correspondence analysis, preference mapping, and multidimensional preference analysis are descriptive statistical methods that generate graphical displays from data matrices. These methods are used by marketing researchers to investigate relationships among products and individual differences in preferences for those products. The end result is a two- or three-dimensional scatter plot that shows the most salient information in the data matrix. This chapter describes these methods, shows examples of the graphical displays, and discusses marketing research applications.*

Introduction

Correspondence analysis (CA), multiple correspondence analysis (MCA), preference mapping (PREFMAP), and multidimensional preference analysis (MDPREF) are descriptive statistical methods that generate graphical displays from data matrices. These methods are sometimes referred to as perceptual mapping methods. They simultaneously locate two or more sets of points in a single plot, and all emphasize presenting the geometry of the data. CA simultaneously displays in a scatter plot the row and column labels from a two-way contingency table or crosstabulation constructed from two categorical variables. MCA simultaneously displays in a scatter plot the category labels from more than two categorical variables. MDPREF displays both the row labels (products) and column labels (people) from a data matrix of continuous variables. PREFMAP shows rating scale data projected into a plot of row labels−for example, from an MDPREF analysis. These methods are used by marketing researchers to investigate relationships among products and individual differences in preferences for those products.

This chapter will only discuss these techniques as methods of generating two-dimensional scatter plots. However, three-dimensional and higher-dimensional results can also be generated and displayed with modern interactive graphics software and with scatter plot matrices.

*This chapter is a revision of a paper that was published in the 1992 National Computer Graphics Association Conference Proceedings. Copies of this chapter (TS-689K) are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market .

Methods

This section presents the algebra and example plots for MDPREF, PREFMAP, CA, and MCA. These methods are all similar in spirit to the biplot, which is discussed first to provide a foundation for the other methods.

The Biplot. A *biplot* (Gabriel 1981) simultaneously displays the rows and columns of a data matrix in a low-dimensional (typically two-dimensional) plot. The "bi" in "biplot" refers to the *joint* display of rows and columns, not to the dimensionality of the plot. Consider an $(n \times m)$ data matrix $\mathbf{Y}$, an $(n \times q)$ matrix $\mathbf{A}$ with row vectors $\mathbf{a}_1'$, $\mathbf{a}_2'$, ..., $\mathbf{a}_n'$, and an $(m \times q)$ matrix $\mathbf{B}$ with row vectors $\mathbf{b}_1'$, $\mathbf{b}_2'$, ..., $\mathbf{b}_m'$. The n rows of $\mathbf{A}$ correspond to the rows of $\mathbf{Y}$, and the m columns of $\mathbf{B}'$ correspond to the columns of $\mathbf{Y}$. The rank of $\mathbf{Y}$ is $q \leq \text{MIN}(n, m)$. $\mathbf{A}$ and $\mathbf{B}$ are chosen such that $y_{ij} = \mathbf{a}_i'\mathbf{b}_j$. If $q = 2$ and the rows of $\mathbf{A}$ and $\mathbf{B}$ are plotted in a two-dimensional scatter plot, the scalar product of the coordinates $\mathbf{a}_i'$ and $\mathbf{b}_j'$ *exactly* equals the data value y_{ij}. This kind of scatter plot is a biplot; it geometrically shows the algebraic relationship $\mathbf{AB}' = \mathbf{Y}$. Typically, the row coordinates are plotted as points, and the column coordinates are plotted as vectors.

When $q > 2$ and two dimensions are plotted, then $\mathbf{a}_i'\mathbf{b}_j$ is *approximately* equal to y_{ij}, and the display is an *approximate biplot.*[*] The approximate biplot geometrically shows the algebraic relationship $\mathbf{AB}' \approx \mathbf{Y}$. The best values for $\mathbf{A}$ and $\mathbf{B}$, in terms of minimum squared error in approximating $\mathbf{Y}$, are found using a singular value decomposition (SVD),[†] $\mathbf{Y} = \mathbf{AB}' = \mathbf{UDV}'$, where $\mathbf{D}$ is a $(q \times q)$ diagonal matrix and $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_q$, a $(q \times q)$ identity matrix. Solutions for $\mathbf{A}$ and $\mathbf{B}$ include $\mathbf{A} = \mathbf{U}$ and $\mathbf{B} = \mathbf{VD}$, or $\mathbf{A} = \mathbf{UD}$ and $\mathbf{B} = \mathbf{V}$, or more generally $\mathbf{A} = \mathbf{UD}^r$ and $\mathbf{B} = \mathbf{VD}^{(1-r)}$, for $0 \leq r \leq 1$. See Gabriel (1981) for more information on the biplot.

Multidimensional Preference Analysis. Multidimensional Preference Analysis (Carroll 1972) or MDPREF is a biplot analysis for preference data. Data are collected by asking respondents to rate their preference for a set of objects. Typically in marketing research, the objects are products−the client's products and the competitors'. Questions that can be addressed with MDPREF analyses include: Who are my customers? Who else should be my customers? Who are my competitors' customers? Where is my product positioned relative to my competitors' products? What new products should I create? What audience should I target for my new products?

For example, consumers can be asked to rate their preference for a group of automobiles on a 0 to 9 scale, where 0 means no preference and 9 means high preference. $\mathbf{Y}$ is an $(n \times m)$ matrix that contains ratings of the n products by the m consumers. The data are stored as the transpose of the typical data matrix, since the columns are the people. The goal is to produce a plot with the cars represented as points and the consumers represented as vectors. Each person's vector points in *approximately* the direction of the cars that the person most preferred and away from the cars that are least preferred.

Figure 1 contains an example in which 25 consumers rated their preference for 17 new (at the time) 1980 automobiles. This plot is based on a principal component model. It differs from a proper biplot of $\mathbf{Y}$ due to scaling factors. In principal components, the columns in data matrix $\mathbf{Y}$ are standardized to mean zero and variance one. The SVD is $\mathbf{Y} = \mathbf{UDV}'$, and the principal component model is $\mathbf{Y} = ((n-1)^{1/2}\mathbf{U})\,((n-1)^{-1/2}\mathbf{D})\,(\mathbf{V}')$. The standardized principal component scores matrix, $\mathbf{A} = (n-1)^{1/2}\mathbf{U}$, and the component structure matrix, $(n-1)^{-1/2}\mathbf{DV}'$, are plotted. The advantage of creating a biplot based on $(n-1)^{1/2}\mathbf{U}$ and $(n-1)^{-1/2}\mathbf{DV}'$ instead of $\mathbf{U}$ and $\mathbf{DV}'$ is that the coordinates

[*]In practice, the term biplot is sometimes used without qualification to refer to an approximate biplot.

[†]SVD is sometimes referred to in the psychometric literature as an Eckart-Young (1936) decomposition.

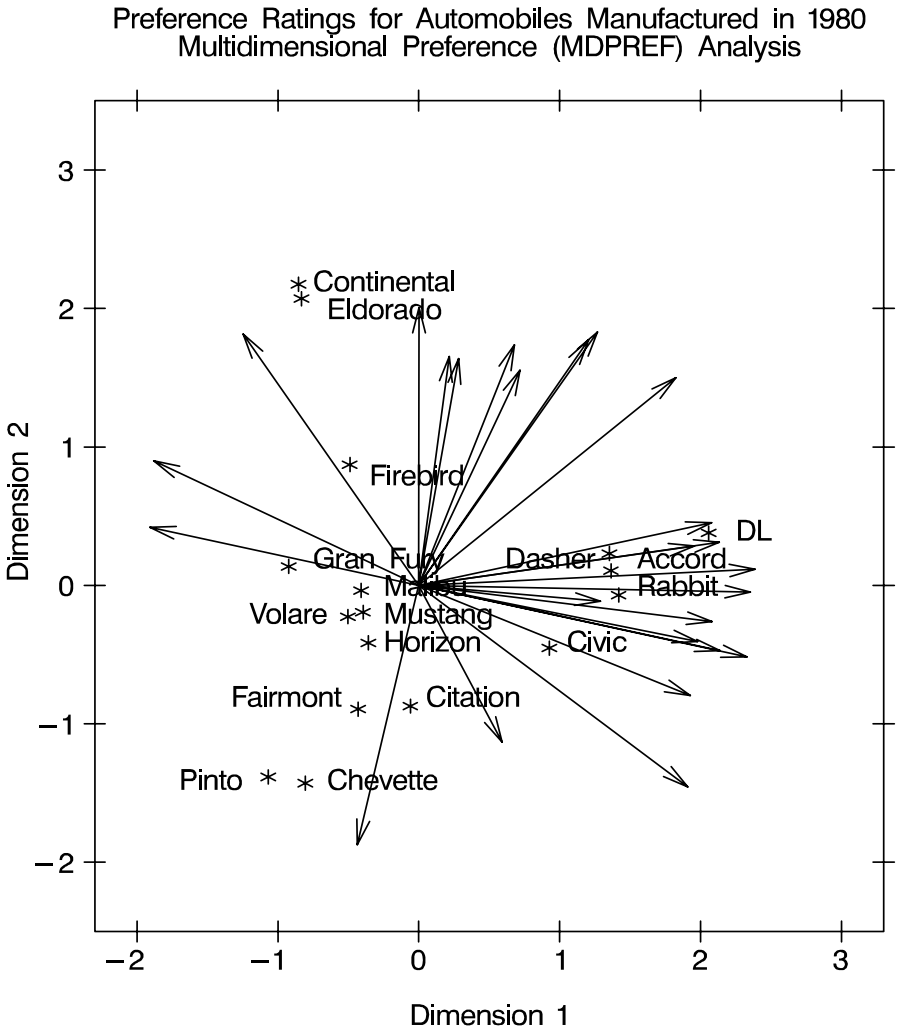


Figure 1. Multidimensional Preference Analysis

do not get smaller as sample size increases. The fit, or proportion of the variance in the data accounted for by the first two dimensions, is the sum of squares of the first two elements of $(n-1)^{-1/2}\mathbf{D}$ divided by the sum of squares of all of the elements of $(n-1)^{-1/2}\mathbf{D}$.

The dimensions of the MDPREF biplot are the first two principal components. The first principal component represents the information that is most salient to the preference judgments. At one end of the plot of the first principal component are the most preferred automobiles; the least preferred automobiles are at the other end of the plot. The second principal component represents the direction that is most salient to the preference judgments that is orthogonal to the first principal component. The automobile point coordinates are the scores of the automobile on the first two principal components. The judge vectors point in *approximately* the direction of judges most preferred cars, with preference

increasing as the vector moves from the origin.

Let $\mathbf{a}_i'$ be row i of $\mathbf{A} = (n-1)^{1/2}\mathbf{U}$, $\mathbf{b}_j'$ be row j of $\mathbf{B} = (n-1)^{-1/2}\mathbf{VD}$, $\|\mathbf{a}_i\|$ be the length of $\mathbf{a}_i$, $\|\mathbf{b}_j\|$ be the length of $\mathbf{b}_j$, and θ be the angle between the vectors $\mathbf{a}_i$ and $\mathbf{b}_j$. The predicted degree of (scaled) preference that an individual judge has for an automobile is $\mathbf{a}_i'\mathbf{b}_j = \|\mathbf{a}_i\|\,\|\mathbf{b}_j\| cos\theta$. Each car point can be orthogonally projected onto each judge's vector. The projection of the ith car on the jth judge vector is $\mathbf{b}_j((\mathbf{a}_i'\mathbf{b}_j)/(\mathbf{b}_j'\mathbf{b}_j))$, and the length of this projection is $\|\mathbf{a}_i\| cos\theta$. The automobile that projects farthest along a judge vector has the highest predicted preference. The length of this projection, $\|\mathbf{a}_i\| cos\theta$, differs from the predicted preference, $\|\mathbf{a}_i\|\,\|\mathbf{b}_j\| cos\theta$, only by $\|\mathbf{b}_j\|$, which is constant within each judge. Since the goal is to look at projections of points onto the vectors, the absolute length of a judge's vector is unimportant. The relative lengths of the vectors indicate fit, with longer vectors indicating better fit. The coordinates for the endpoints of the vectors were multiplied by 2.5 to extend the vectors and create a better graphical display. The direction of the preference scale is important. The vectors point in the direction of increasing values of the data values. If the data had been ranks, with 1 the most preferred and n the least preferred, then the vectors would point in the direction of the least preferred automobiles.

The people in the top left portion of the plot most prefer the large American cars. Other people, with vectors pointing up and nearly vertical, also show this pattern of preference. There is a large cluster of people who prefer the Japanese and European cars. A few people, most notably the person whose vector passes through the "e" in "Chevette", prefer the small and inexpensive American Cars. There are no vectors pointing through the bottom left portion of the of the plot, which suggests that the smaller American cars are generally not preferred by anyone within this group.

The first dimension, which is a measure of overall evaluation, discriminates between the American cars on the left and the Japanese and European cars on the right. The second dimension seems to reflect the sizes of the automobiles. Some cars have a similar pattern of preference, most notably Continental and Eldorado, which share a symbol in the plot. Marketers of Continental or Eldorado may want to try to distinguish their car from the competition. Dasher, Accord, and Rabbit were rated similarly, as were Malibu, Mustang, Volare, and Horizon.

This 1980 example is quite prophetic even though it is based on a small nonrandom sample. Very few vectors point toward the smaller American cars, and Mustang is the only one of them that is still being made. Many vectors are pointing toward the European and Japanese cars, and they are still doing quite well in the market place. Many vectors are pointing in the one to two o'clock range where there are no cars in the plot. One can speculate that these people would prefer Japanese and European luxury cars such as Accura, Lexus, Infinity, BMW, and Mercedes.

Preference Mapping. Preference mapping[‡] (Carroll 1972) or PREFMAP plots resemble biplots, but are based on a different model. The goal in PREFMAP is to take a set of coordinates for a set of objects, such as the MDPREF car coordinates in example in Figure 1, and project in external information that can aid in interpreting the configuration of points. Questions that can be addressed with PREFMAP analyses include: Where is my product positioned relative to my competitors' products? Why is my product positioned there? How can I reposition my existing products? What new products should I create?

[‡]Preference mapping is sometimes referred to as external unfolding.

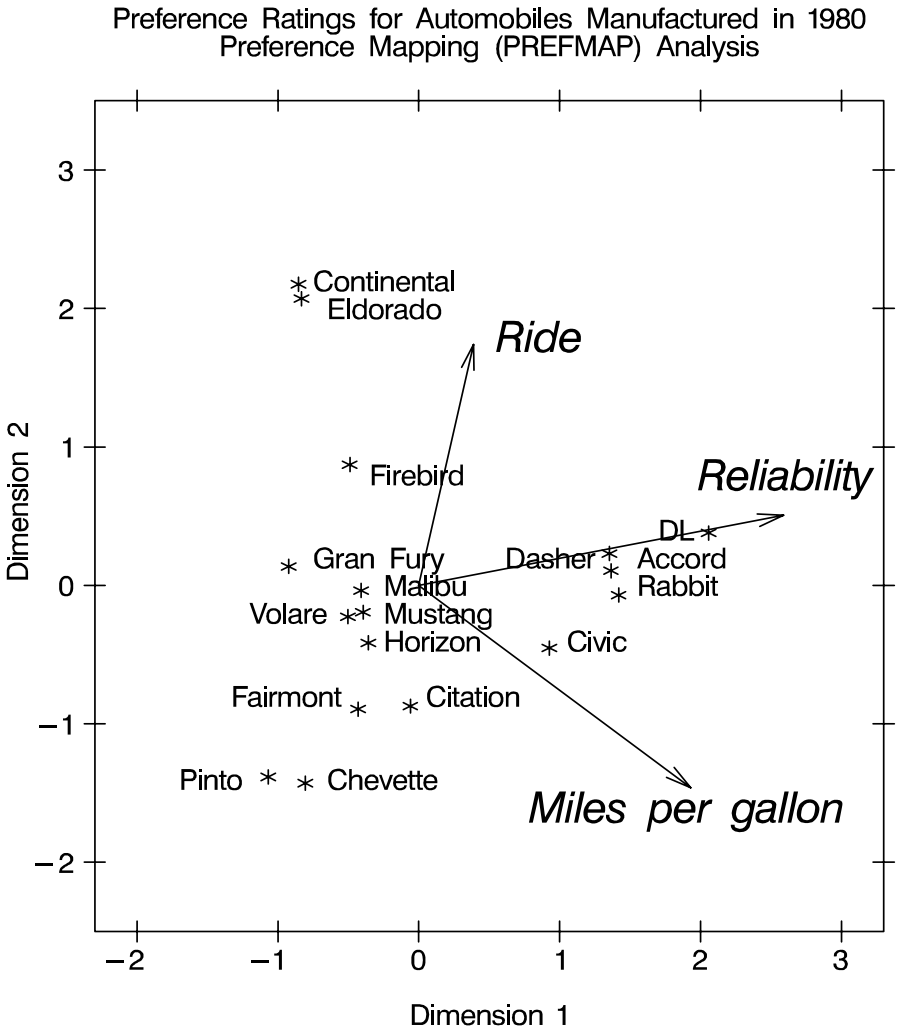


Figure 2. Preference Mapping, Vector Model

The Preference Mapping Vector Model. Figure 2 contains an example in which three attribute variables (ride, reliability, and miles per gallon) are displayed in the plot of the first two principal components of the car preference data. Each of the automobiles was rated on these three dimensions on a 1 to 5 scale, where 1 is poor and 5 is good. Figure 2 is based on the simplest version of PREFMAP−the *vector model*. The vector model assumes that some is good and more is *always* better. This model is appropriate for miles per gallon and reliability−the more miles a motorist can travel without refueling or breaking down, the better. The end points for the attribute vectors are obtained by projecting the attribute variables into the car space. If the attribute ratings are stored in matrix $\mathbf{R}$, then the coordinates for the end points are in the matrix $\boldsymbol{\beta}$ from the multivariate linear regression model $\mathbf{R} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. $\mathbf{A}$ is the matrix of standardized principal component scores, or $\mathbf{A}$ could be the coordinates from a multidimensional scaling analysis. The relative lengths of the vectors indicate fit, which is given

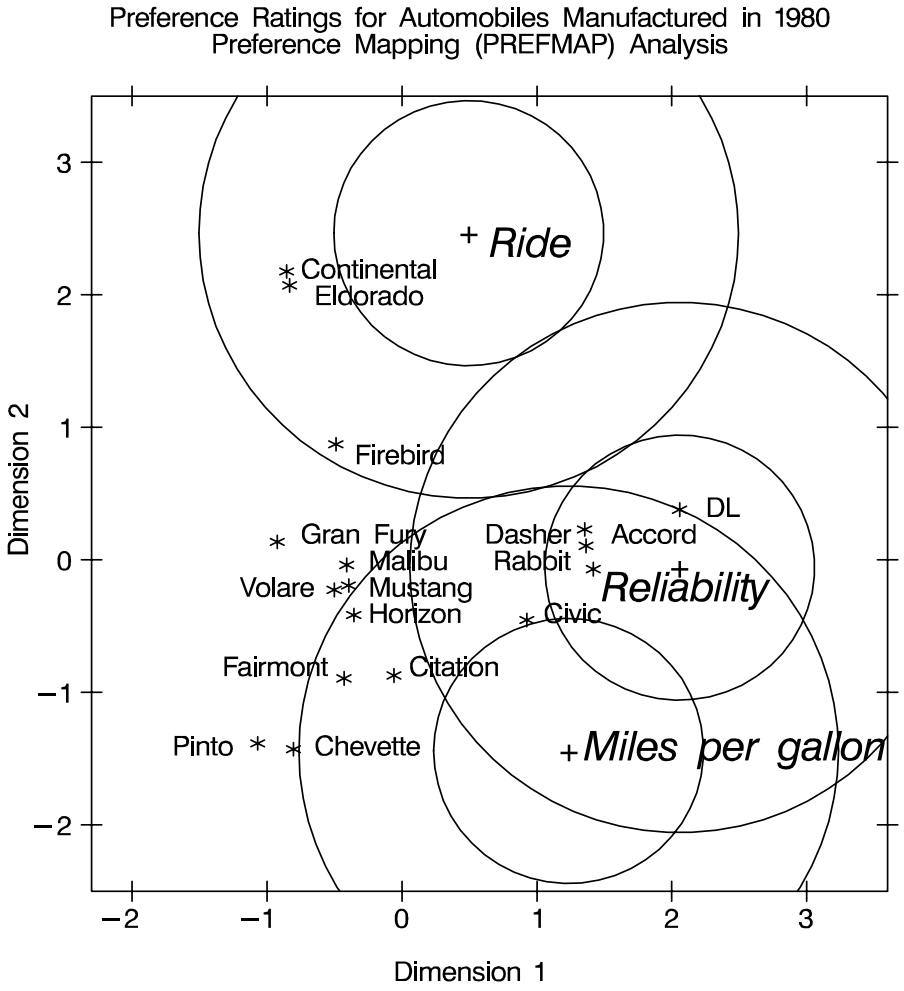


Figure 3. Preference Mapping, Ideal Point Model

by the R^2. As with MDPREF, the lengths of all vectors can be scaled by the same constant to make a better graphical display.

PREFMAP analyses can help in the interpretation of principal component, multidimensional scaling, and MDPREF analyses by projecting in external information that helps explain the configuration. Orthogonal projections of the product points on an attribute vector give an *approximate* ordering of the products on the attribute. The ride vector points almost straight up showing that the larger cars, such as the Eldorado and Continental, have the best ride. In Figure 1, it was shown that most people preferred the DL, Japanese cars, and larger American cars. Figure 2 shows that the DL and Japanese cars were rated as the most reliable and have the best fuel economy. The small American cars are not rated highly on any of the three dimensions, although some are on the positive end of miles per gallon.

The Preference Mapping Ideal Point Model. The *ideal point* model differs from the vector model in that the ideal point model does not assume that more is better, *ad infinitum*. Consider the sugar content of cake. There is an ideal amount of sugar that cake should contain−not enough sugar is not good, and too much sugar is also not good. In the current example, the ideal number of miles per gallon and the ideal reliability are unachievable. It makes sense to consider a vector model, because the ideal point is infinitely far away. This argument is less compelling for ride; the point for a car with smooth, quiet ride may not be infinitely far away. Figure 3 shows the results of fitting an ideal point model for the three attributes. In the vector model, results are interpreted by orthogonally projecting the car points on the attribute vectors. In the ideal point model, Euclidean distances between car points and ideal points are compared. Eldorado and Continental have the best predicted ride, because they are closest to the ride ideal point. The concentric circles drawn around the ideal points help to show distances between the cars and the ideal points. The numbers of circles and their radii are arbitrary. The overall interpretations of Figures 2 and 3 are the same. All three ideal points are at the edge of the car points, which suggests the simpler vector model is sufficient for these data.

The ideal point model is fit with a multiple regression model and some pre- and post-processing. First the $\mathbf{A}$ matrix is augmented by a variable that is the sum of squares of columns of $\mathbf{A}$ creating $\mathbf{A}^*$. Then solve for $\boldsymbol{\beta}$ from $\mathbf{R} = \mathbf{A}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}$. For a two-dimensional scatter plot, the ideal point coordinates are given by dividing each coefficient for the two axes by the coefficient for the sum-of-squares variable, then multiplying the resulting values by -0.5. The coordinates are $-0.5\,\boldsymbol{\beta}\,\mathrm{diag}(\boldsymbol{\beta}_3)^{-1}$, where $\mathrm{diag}(\boldsymbol{\beta}_3)$ is a diagonal matrix constructed from the third row of $\boldsymbol{\beta}$. This is a constrained response-surface model. The fit is given by the R^2. See Carroll (1972) for the justification for the formula.

The results in Figure 3 were modified from the raw results to eliminate *anti-ideal points*. The ideal point model is a distance model. The rating data are interpreted as distances between attribute ideal points and the products. In this example, each of the automobiles was rated on these three dimensions, on a 1 to 5 scale, where 1 is poor and 5 is good. The data are the reverse of what they should be−a ride rating of 1 should mean this car is similar to a car with a good ride, and a rating of 5 should mean this car is different from a car with a good ride. So the raw coordinates must be multiplied by -1 to get ideal points. Even if the scoring had been reversed, anti-ideal points can occur. If the coefficient for the sum-of-squares variable is negative, the point is an anti-ideal point. In this example, there is the possibility of *anti-anti-ideal points*. When the coefficient for the sum-of-squares variable is negative, the two multiplications by -1 cancel, and the coordinates are ideal points. When the coefficient for the sum-of-squares variable is positive, the coordinates are multiplied by -1 to get an ideal point.

Other PREFMAP Models. The ideal point model presented here is based on an ordinary Euclidean distance model. All points falling on a circle centered around an ideal point are an equal distance from the ideal point. Two more PREFMAP models are sometimes used. The more general models allow for differential weighting of the axes and rotations, so ellipses, not circles, show equal weighted distances. All three ideal point models are response surface models. See Carroll (1972) for more information.

Correspondence Analysis. Correspondence analysis (CA) is a weighted SVD of a contingency table. It is used to find a low-dimensional graphical representation of the association between rows and columns of a table. Each row and column is represented by a point in a Euclidean space determined from cell frequencies. Like MDPREF, CA is based on a singular value decomposition, but ordinary SVD of a contingency table does not portray a desirable geometry.

Questions that can be addressed with CA and MCA include: Who are my customers? Who else should be my customers? Who are my competitors' customers? Where is my product positioned relative to my competitors' products? Why is my product positioned there? How can I reposition my existing products? What new products should I create? What audience should I target for my new products?

CA is a popular data analysis method in France and Japan. In France, CA was developed under the strong influence of Jean-Paul Benzécri; in Japan, under Chikio Hayashi. CA is described in Lebart, Morineau, and Warwick (1984); Greenacre (1984); Nishisato (1980); Tenenhaus and Young (1985); Gifi (1990); Greenacre and Hastie (1987); and many other sources. Hoffman and Franke (1986) provide a good introductory treatment using examples from marketing research.

Simple CA. This section is primarily based on the theory of CA found in Greenacre (1984). Let $\mathbf{N}$ be an $(n_r \times n_c)$ contingency table of rank $q \leq MIN(n_r, n_c)$. Let $\mathbf{1}$ be a vector of ones of the appropriate order, $\mathbf{I}$ be an identity matrix, and diag() be a matrix-valued function that creates a diagonal matrix from a vector. Let $f = \mathbf{1'N1}$, $\mathbf{P} = (1/f)\mathbf{N}$, $\mathbf{r} = \mathbf{P1}$, $\mathbf{c} = \mathbf{P'1}$, $\mathbf{D_r} = \text{diag}(\mathbf{r})$, and $\mathbf{D_c} = \text{diag}(\mathbf{c})$. The scalar f is the sum of all elements in $\mathbf{N}$. $\mathbf{P}$ is a matrix of relative frequencies. The vector $\mathbf{r}$ contains row marginal proportions or row *masses*. The vector $\mathbf{c}$ contains column marginal proportions or column masses. $\mathbf{D_r}$ and $\mathbf{D_c}$ are diagonal matrices of marginals. The coordinates of the CA are based on the generalized singular value decomposition of $\mathbf{P}$, $\mathbf{P} = \mathbf{AD_uB'}$, where $\mathbf{A'D_r}^{-1}\mathbf{A} = \mathbf{B'D_c}^{-1}\mathbf{B} = \mathbf{I}$. $\mathbf{A}$ is an $(n_r \times q)$ matrix of left generalized singular vectors, $\mathbf{D_u}$ is a $(q \times q)$ diagonal matrix of singular values, and $\mathbf{B}$ is an $(n_c \times q)$ matrix of right generalized singular vectors. The first (trivial) column of $\mathbf{A}$ and $\mathbf{B}$ and the first singular value in $\mathbf{D_u}$ are discarded before any results are displayed. This step centers the table and is analogous to centering the data in ordinary principal component analysis. In practice, this centering is done by subtracting ordinary chi-square expected values from $\mathbf{P}$ before the SVD. The columns of $\mathbf{A}$ and $\mathbf{B}$ define the principal axes of the column and row point clouds, respectively. The fit, or proportion of the *inertia* (analogous to variance) in the data accounted for by the first two dimensions, is the sum of squares of the first two singular values, divided by the sum of squares of all of the singular values. Three sets of coordinates are typically available from CA, one based on rows, one based on columns, and the usual set is based on both rows and columns.

The *row profile* (conditional probability) matrix is defined as $\mathbf{R} = \mathbf{D_r}^{-1}\mathbf{P} = \mathbf{D_r}^{-1}\mathbf{AD_uB'}$. Each (i, j) element of $\mathbf{R}$ contains the observed probability of being in column j given membership in row i. The values in each row of $\mathbf{R}$ sum to one. The row coordinates, $\mathbf{D_r}^{-1}\mathbf{AD_u}$, and column coordinates, $\mathbf{D_c}^{-1}\mathbf{B}$, provide a CA based on the row profile matrix. The *principal* row coordinates, $\mathbf{D_r}^{-1}\mathbf{AD_u}$, and *standard* column coordinates, $\mathbf{D_c}^{-1}\mathbf{B}$, provide a decomposition of $\mathbf{D_r}^{-1}\mathbf{AD_uB'D_c}^{-1} = \mathbf{D_r}^{-1}\mathbf{PD_c}^{-1} = \mathbf{RD_c}^{-1}$. Since $\mathbf{D_r}^{-1}\mathbf{AD_u} = \mathbf{RD_c}^{-1}\mathbf{B}$, the row coordinates are weighted centroids of the column coordinates. Each column point, with coordinates scaled to standard coordinates, defines a vertex in $(n_c - 1)$-dimensional space. All of the principal row coordinates are located in the space defined by the standard column coordinates. Distances among row points have meaning, but distances among column points and distances between row and column points are not interpretable.

The formulas for the analysis of the *column profile* matrix can easily be derived by applying the row profile formulas to the transpose of $\mathbf{P}$. The principal column coordinates $\mathbf{D_c}^{-1}\mathbf{BD_u}$ are weighted centroids of the standard row coordinates $\mathbf{D_r}^{-1}\mathbf{A}$. Each row point, with coordinates scaled to standard

coordinates, defines a vertex in $(n_r - 1)$-dimensional space. All of the principal column coordinates are located in the space defined by the standard row coordinates. Distances among column points have meaning, but distances among row points and distances between row and column points are not interpretable.

The usual sets of coordinates[§] are given by $\mathbf{D_r}^{-1}\mathbf{AD_u}$ and $\mathbf{D_c}^{-1}\mathbf{BD_u}$. One advantage of using these coordinates is that both sets are postmultiplied by the diagonal matrix $\mathbf{D_u}$, whose diagonal values are all less than or equal to one. When $\mathbf{D_u}$ is a part of the definition of only one set of coordinates, that set forms a tight cluster near the centroid, whereas the other set of points is more widely dispersed. Including $\mathbf{D_u}$ in both sets makes a better graphical display. However, care must be taken in interpreting such a plot. No correct interpretation of distances between row points and column points can be made. Less specific statements, such as "two points are on the same side of the plot" have meaning.

Another property of this choice of coordinates concerns the geometry of distances between points within each set. Distances between row (or column) profiles are computed using a *chi-square metric*. The rationale for computing distances between row profiles using the non-Euclidean chi-square metric is as follows. Each row of the contingency table may be viewed as a realization of a multinomial distribution conditional on its row marginal frequency. The null hypothesis of row and column independence is equivalent to the hypothesis of homogeneity of the row profiles. A significant chi-square statistic is geometrically interpreted as a significant deviation of the row profiles from their centroid, $\mathbf{c}'$. The chi-square metric is the Mahalanobis metric between row profiles based on their estimated covariance matrix under the homogeneity assumption (Greenacre and Hastie 1987). A parallel argument can be made for the column profiles.

The row coordinates are $\mathbf{D_r}^{-1}\mathbf{AD_u} = \mathbf{D_r}^{-1}\mathbf{AD_uB'D_c}^{-1}\mathbf{B} = (\mathbf{D_r}^{-1}\mathbf{P})(\mathbf{D_c}^{-1/2})(\mathbf{D_c}^{-1/2}\mathbf{B})$. They are row profiles $\mathbf{D_r}^{-1}\mathbf{P}$ rescaled by $\mathbf{D_c}^{-1/2}$ (rescaled so that distances between profiles are transformed from a chi-square metric to a Euclidean metric), then orthogonally rotated with $\mathbf{D_c}^{-1/2}\mathbf{B}$ to a principal axes orientation. Similarly, the column coordinates are column profiles rescaled to a Euclidean metric and orthogonally rotated to a principal axes orientation.

CA Example. Figure 4 contains a plot of the results of a simple CA of a survey of car owners. The questions included origin of the car (American, Japanese, European), and marital/family status (single, married, single and living with children, and married living with children). Both variables are categorical. Table 1 contains the crosstabulation and the observed minus expected frequencies. It can be seen from the observed minus expected frequencies that four cells have values appreciably different from zero (Married w Kids/American, Single/American, Married w Kids/Japanese, Single/Japanese). More people who are married with children drive American cars than would be expected if the rows and columns are independent, and more people who are single with no children drive Japanese cars than would be expected if the rows and columns are independent.

CA graphically shows the information in the observed minus expected frequencies. The right side of Figure 4 shows the association between being married with children and owning an American Car. The left side of the plot shows the association between being single and owning a Japanese Car. This interpretation is based on points being located in approximately the same direction from the origin and in approximately the same region of the space. Distances between row points and column points are not defined.

[§]This set is often referred to as the French standardization due to its popularity in France.

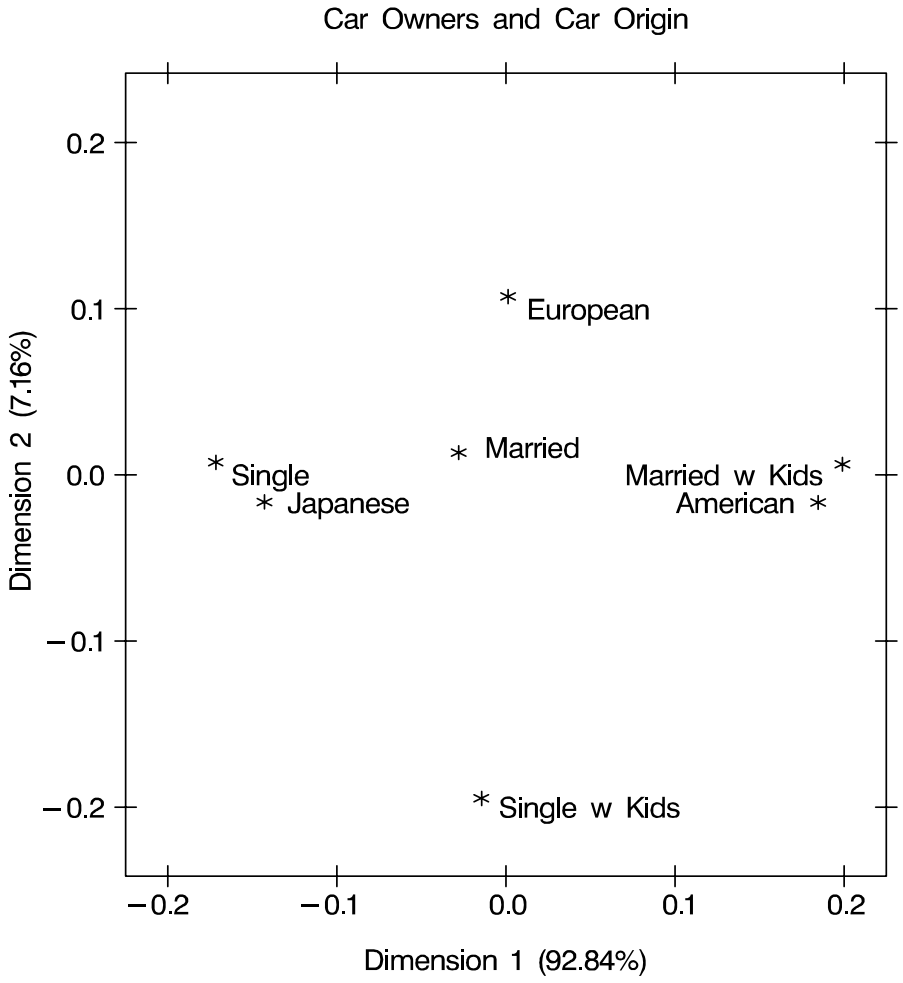


Figure 4. Simple Correspondence Analysis

Table 1
Simple Correspondence Example Input

	Contingency Table			Observed Minus Expected Values		
	American	European	Japanese	American	European	Japanese
Married	37	14	51	-1.5133	0.4602	1.0531
Married w Kids	52	15	44	10.0885	0.2655	-10.3540
Single	33	15	63	-8.9115	0.2655	8.6460
Single w Kids	6	1	8	0.3363	-0.9912	0.6549

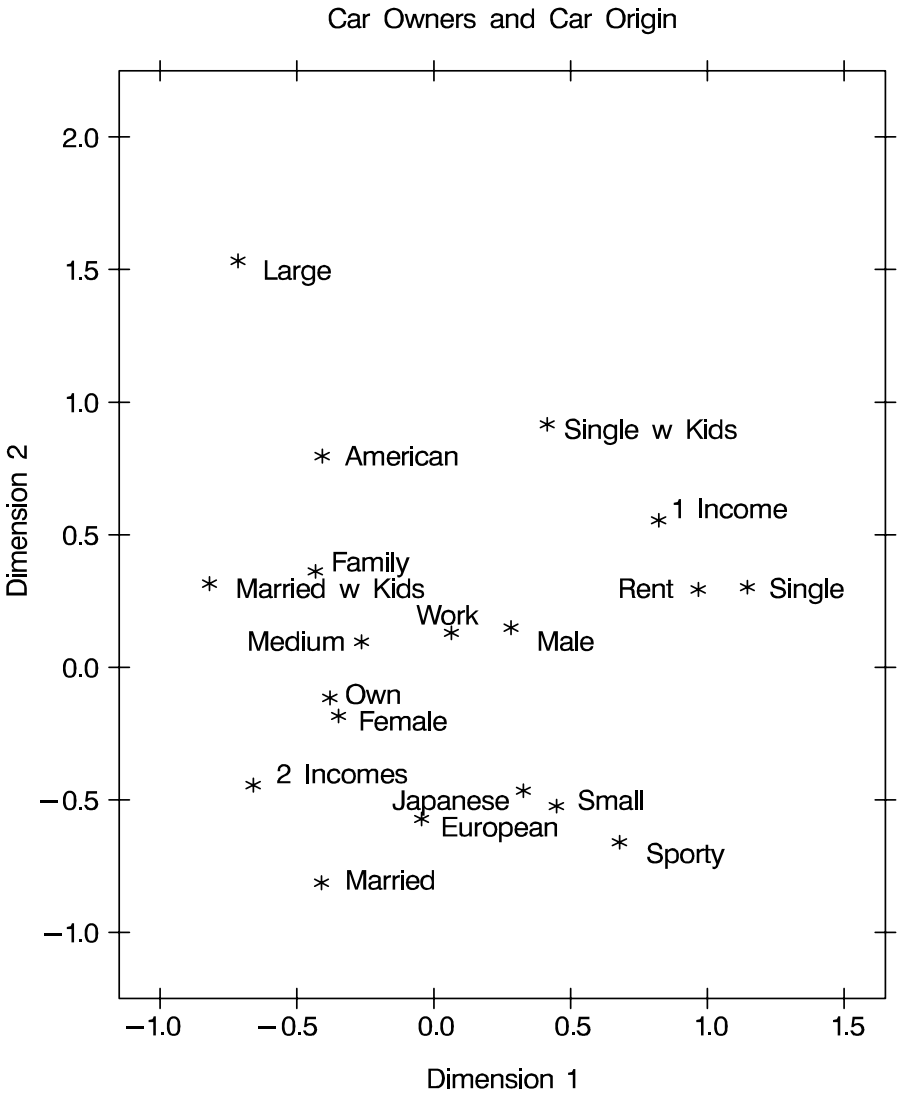


Figure 5. Multiple Correspondence Analysis

Multiple Correspondence Analysis. Multiple correspondence analysis (MCA) is a generalization of simple CA for more than two variables. The input is a *Burt table*, which is a partitioned symmetric matrix containing all pairs of crosstabulations among a set of categorical variables. Each diagonal partition is a diagonal matrix containing marginal frequencies (a crosstabulation of a variable with itself). Each off-diagonal partition is an ordinary contingency table. Each contingency table above the diagonal has a transposed counterpart below the diagonal. A Burt table is the inner product of a partitioned design matrix. There is one partition per categorical variable, and each partition is a binary design matrix. Each design matrix has one column per category, and a single 1 in each row. The partitioned design matrix has exactly m ones in each row, where m is the number of categorical variables. The results of a MCA of a Burt table, $\mathbf{N}$, are the same as the column results from a simple CA of the design matrix whose inner product is the Burt table. MCA is not a simple CA of the

Burt table. The coordinates for MCA are $\mathbf{D_c}^{-1}\mathbf{B}\mathbf{D_u}$, from $(1/f)\mathbf{N} = \mathbf{P} = \mathbf{P}' = \mathbf{B}\mathbf{D_u}^2\mathbf{B}'$, where $\mathbf{B}'\mathbf{D_c}^{-1}\mathbf{B} = \mathbf{I}$.

MCA Example. Figure 5 contains a plot of the results of an MCA of a survey of car owners. The questions included origin of the car (American, Japanese, European), size of car (small, medium, large), type of car (family, sporty, work vehicle), home ownership (owns, rents), marital/family status (single, married, single and living with children, and married living with children), and sex (male, female). The variables are all categorical.

The top-right quadrant of the plot shows that the categories single, single with kids, 1 income, and renting a home are associated. Proceeding clockwise, the categories sporty, small, and Japanese are associated. In the bottom-left quadrant we see the association between being married, owning your own home, and having two incomes. Having children is associated with owning a large American family car. Such information could be used in marketing research to identify target audiences for advertisements. This interpretation is based on points being located in approximately the same direction from the origin and in approximately the same region of the space. Distances between points are not interpretable in MCA.

Other CA Standardizations. Other standardizations have been proposed for CA by several authors. The usual goal is to provide a standardization that avoids the problem of row and column distances being undefined. Unfortunately, this problem remains unsolved. Carroll, Green, and Schaffer (1986) proposed that simple CA coordinates should be transformed to MCA coordinates before they are plotted. They argued that all distances are then comparable, but Greenacre (1989) showed that their assertion was incorrect. Others have also claimed to have discovered a method of defining the between row and column differences, but so far no method has been demonstrated to be correct.

Notes

The Geometry of the Scatter Plots. All of the scatterplots in this chapter were created with the axes equated so that a centimeter on the y-axis represents the same data range as a centimeter on the x-axis. *This is important.* Distances, angles between vectors, and projections are evaluated to interpret the plots. When the axes are equated, distances and angles are correctly represented in the plot. When axes are scaled independently, for example to fill the page, then the correct geometry is not presented. The important step of equating the axes is often overlooked in practice.

In a true biplot, $\mathbf{A} = \mathbf{U}\mathbf{D}^r$ and $\mathbf{B} = \mathbf{V}\mathbf{D}^{(1-r)}$ are plotted, and the elements of $\mathbf{Y}$ can be approximated from $y_{ij} \approx \mathbf{a}_i'\mathbf{b}_j$. For MDPREF and PREFMAP, the absolute lengths of the vectors are not important since the goal is to project points on vectors, not look at scalar products of row points and column vectors. It is often necessary to change the lengths of *all* of the vectors to improve the graphical display. If all of the vectors are relatively short with end points clustered near the origin, the display will not look good. To avoid this problem in Figure 1, *both* the x-axis and y-axis coordinates were multiplied by the constant 2.5, to lengthen all vectors by the same relative amount. The coordinates must not be scaled independently.

Software. All data analyses were performed with Release 8.00 of the SAS System. MDPREF is performed with PROC PRINQUAL, simple and multiple correspondence analysis are performed with PROC CORRESP, and PREFMAP is performed with PROC TRANSREG. The plots are prepared with the SAS `%PlotIt` autocall macro. If your site has installed the autocall libraries supplied by SAS Institute and uses the standard configuration of SAS software supplied by the Institute, you need only to ensure that the SAS system option `mautosource` is in effect to begin using the autocall macros. For more information about autocall libraries, refer to *SAS Macro Language: Reference*. The macro is documented in the macro comments.

Conclusions

Marketing research helps marketing decision makers understand their customers and their competition. Correspondence analysis compactly displays survey data to aid in determining what kinds of consumers are buying certain products. Multidimensional preference analysis shows product positioning, group preferences, and individual preferences. The plots may suggest how to reposition products to appeal to a broader audience. They may also suggest new groups of customers to target. Preference mapping is used as an aid in understanding MDPREF and multidimensional scaling results. PREFMAP displays product attributes in the same plot as the products. The insight gained from perceptual mapping methods can be a valuable asset in marketing decision making. These techniques can help marketers gain insight into their products, their customers, and their competition.

Concluding Remarks

I hope you like this book and the new macros. In particular, I hope you find the `%MktEx` macro to be very powerful and useful. My goal in writing this book and tool set is to help you do better research and do it more quickly and more easily. I would like to hear what you think. Many of my examples and enhancements to the software are based on feedback from people like you. This book has already been revised many times, and future revisions are likely. If you have comments or suggestions for future revisions, write Warren F. Kuhfeld, (Warren.Kuhfeld@sas.com) at SAS Institute Inc. Almost all of our examples are artificial. We would welcome any real data sets that we could use in future examples. Please direct questions to the technical support division. My goal to provide you with enough examples so that you can easily adapt aspects of one or more examples to fit your particular needs. When I do not succeed, tell me about it and I will try to add a new example to the next revision. Please email me. I would like to hear from you!

I would like to put in a plug for the American Marketing Association's Advanced Research Techniques Forum (ART Forum), which is a conference held each year in June. I have been to every one since 1991, and that is a streak that I hope to keep going for a long time. It is a great place to meet academic researchers and top practitioners in the areas of conjoint, choice, and other branches of marketing research. It always draws a diverse and international crowd. There are a number of great tutorials including (most years including 2003) one by Don Anderson and myself on choice designs.

I leave you with this old Irish blessing.

May the road rise up
to meet you
may the wind be
always at your back
May the sun shine warm
on your face
And the rain fall soft
upon your fields

... along with this additional thought ...

May your designs always be efficient
and your standard errors small

The Kuhfeld Conundrum

What do all of the random number seeds used in the "Conjoint Analysis Examples," "Multinomial Logit, Discrete Choice Modeling," and "Experimental Design and Choice Modeling Macros" have in common? Send answers to Warren.Kuhfeld@sas.com. I will send a small prize to the first person to send me the answer that I have in mind. Hint: The relationship is not mathematical. An answer like "they are all less than 619," while true, is not what I have in mind.* I first put this challenge out there quite a few years ago now. I keep hoping that one of these days, one of you will send me the answer. Heed this admonition: stick to the middle of the road.

*If you can't get that one, here is an easy one—no prize for this one though. Find the joke embedded in the index.

References

Addelman, S. (1962a), "Orthogonal Main-Effects Plans for Asymmetrical Factorial Experiments," *Technometrics*, 4, 21–46.

Addelman, S. (1962b), "Symmetrical and Asymmetrical Fractional Factorial Plans," *Technometrics*, 4, 47–58.

Agresti, A. (1990), *Categorical Data Analysis*. New York: John Wiley and Sons.

Anderson, D.A. and Wiley, J.B. (1992), "Efficient Choice Set Designs for Estimating Cross-Effects Models," *Marketing Letters*, 3, 357–370.

de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer Verlag.

Bose, R.C. (1947), "Mathematical Theory of the Symmetrical Factorial Design," *Sankhya*, 8, 107–166.

Booth, K.H.V. and Cox, D.R. (1962), "Some Systematic Super-Saturated Designs," *Technometrics*, 4, 489–495.

Breiman, L. and Friedman, J.H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," (with discussion), *Journal of the American Statistical Association*, 77, 580–619.

Breslow, N. and Day, N.E. (1980), *Statistical Methods in Cancer Research, Vol. II: The Design and Analysis of Cohort Studies,* Lyon: IARC.

Bunch, D.S., Louviere, J.J., and Anderson, D.A. (1996), "A Comparison of Experimental Design Strategies for Choice-Based Conjoint Analysis with Generic-Attribute Multinomial Logit Models," Working Paper, Graduate School of Management, University of California at Davis.

van der Burg, E. and de Leeuw, J. (1983), "Non-linear Canonical Correlation," *British Journal of Mathematical and Statistical Psychology*, 36, 54–80.

Carmone, F.J. and Green, P.E. (1981), "Model Misspecification in Multiattribute Parameter Estimation," *Journal of Marketing Research*, 18 (February), 87–93.

Carroll, J.D. (1972), "Individual Differences and Multidimensional Scaling," in *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences (Volume 1)*, in Shepard, R.N., Romney, A.K., and Nerlove, S.B. (ed.), New York: Seminar Press.

Carroll, J.D, Green, P.E., and Schaffer, C.M. (1986), "Interpoint Distance Comparisons in Correspondence Analysis," *Journal of Marketing Research*, 23, 271–280.

Carson, R.T., Louviere, J.J, Anderson, D.A., Arabie, P., Bunch, D., Hensher, D.A., Johnson, R.M., Kuhfeld, W.F., Steinberg, D., Swait, J., Timmermans, H., and Wiley, J.B. (1994), "Experimental Analysis of Choice," *Marketing Letters,* 5(4), 351–368.

Chakravarti, I.M. (1956), "Fractional Replication in Asymmetrical Factorial Designs and Partially Balanced Arrays," *Sankhya*, 17, 143–164.

Chrzan, K. and Elrod, T. (1995), "Partial Profile Choice Experiments: A Choice-Based Approach for Handling Large Numbers of Attributes," paper presented at the AMA's 1995 Advanced Research Techniques Forum, Monterey, CA.

Cook, R.D. and Nachtsheim, C.J. (1980), "A Comparison of Algorithms for Constructing Exact *D*-optimal Designs," *Technometrics*, 22, 315–324.

Cook, R.D. and Nachtsheim, C.J. (1989), "Computer-Aided Blocking of Factorial and Response-Surface Designs," *Technometrics* 31 (August), 339–346.

Coolen, H., van Rijckevorsel, J., and de Leeuw, J. (1982), "An Algorithm for Nonlinear Principal Components with B-splines by Means of Alternating Least Squares," in H. Caussinus, P. Ettinger, and R. Tomassone (ed.), *COMPUSTAT 1982*, Part 2, Vienna: Physica Verlag.

De Cock, D. and Stufken, J. (2000), "On Finding Mixed Orthogonal Arrays of Strength 2 With Many 2-Level Factors," *Statistics and Probability Letters*, 50, 383–388.

Dey, A. (1985), *Orthogonal Fractional Factorial Designs*, New York: Wiley.

DuMouchell, W. and Jones, B. (1994), "A Simple Bayesian Modification of *D*-Optimal Designs to Reduce Dependence on an Assumed Model," *Technometrics* 36 (February), 37–47.

Dykstra, O. (1971), "The Augmentation of Experimental Data to Maximize $|(\mathbf{X}'\mathbf{X})^{-1}|$," *Technometrics*, 13 (August), 682–688.

Eckart, C. and Young, G. (1936), "The Approximation of One Matrix by Another of Lower Rank," *Psychometrika*, 1, 211–218.

Elrod, T., Louviere, J.J, and Davey, K.S. (1992), "An Empirical Comparison of Ratings-Based and Choice-Based Conjoint Models," *Journal of Marketing Research*, 29 (August), 368–377.

Fedorov, V.V. (1972), *Theory of Optimal Experiments*, translated and edited by W.J. Studden and E.M. Klimko, New York: Academic Press.

Finkbeiner, C.T. (1988), "Comparison of Conjoint Choice Simulators," Sawtooth Software Conference Proceedings.

Fisher, R. (1938), *Statistical Methods for Research Workers*, 10th Edition, Edinburgh: Oliver and Boyd Press.

Gabriel, K.R. (1981), "Biplot Display of Multivariate Matrices for Inspection of Data and Diagnosis," *Interpreting Multivariate Data*, V. Barnett (ed.), London: John Wiley and Sons, Inc.

Gail, M.H., Lubin, J.H., and Rubinstein, L.V. (1981), "Likelihood Calculations for Matched Case-control Studies and Survival Studies with Tied Death Times," *Biometrika*, 68, 703–707.

Gifi, A. (1981), *Nonlinear Multivariate Analysis*, Department of Data Theory, The University of Leiden, The Netherlands.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, New York: Wiley.

Green, P.E. (1974), "On the Design of Choice Experiments involving Multifactor Alternatives," *Journal of Consumer Research*, 1, 61–68.

Green, P.E. and Rao, V.R. (1971), "Conjoint Measurement for Quantifying Judgmental Data," *Journal of Marketing Research*, 8, 355–363.

Green, P.E. and Srinivasan, V. (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing*, 54, 3–19.

Green, P.E. and Wind, Y. (1975), "New Way to Measure Consumers' Judgments," *Harvard Business Review*, July–August, 107–117.

Green, P.E. and Rao, V.R. (1971), "Conjoint Measurement for Quantifying Judgmental Data," *Journal of Marketing Research*, 8, 355–363.

Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*, London: Academic Press.

Greenacre, M.J. (1989), "The Carroll-Green-Schaffer Scaling in Correspondence Analysis: A Theoretical and Empirical Appraisal," *Journal of Market Research*, 26, 358–365.

Greenacre, M.J. and Hastie, T. (1987), "The Geometric Interpretation of Correspondence Analysis," *Journal of the American Statistical Association*, 82, 437–447.

Hadamard, J. (1893), "Resolution d'une Question Relative Aux Determinants," *Bull. des Sciences Math*, (2), 17, 240–246.

Hastie, T. and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 3, 297–318.

Hedayat, A.S., Sloane, N.J.A., and Stufken, J. (1999), *Orthogonal Arrays*, New York: Springer.

Hoffman, D.L., and Franke, G.R. (1986), "Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research," *Journal of Marketing Research*, 23, 213–227.

Hoffman, S.D. and Duncan, G.J. (1988), "Multinomial and Conditional Logit Discrete-choice Models in Demography," *Demography,* 25 (3), 415–427.

Huber, J. and Zwerina, K. (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33, 307–317.

Huber, J., Wittink, D.R., Fiedler, J.A., and Miller, R. (1993), "The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice," *Journal of Marketing Research*, 30 (February), 105–114.

Kirkpatrick, S., Gellat, C.D., and Vecchi, M.P. (1983), "Optimization by Simulated Annealing," *Science*, 220, 671–680.

Krieger, A.B. and Green, P.E. (1991), "Designing Pareto Optimal Stimuli for Multiattribute Choice Experiments," *Marketing Letters*, 2, 337–348.

Kruskal, J.B. and Wish, M. (1978), *Multidimensional Scaling*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07–011, Beverly Hills and London: Sage Publications.

Kruskal, J.B. and Shepard, R.N. (1974), "A Nonmetric Variety of Linear Factor Analysis," *Psychometrika*, 38, 123–157.

Kuhfeld, W.F. (1991), "A Heuristic Procedure for Label Placement in Scatter Plots," Presented to the joint meeting of the Psychometric Society and Classification Society of North America, Rutgers University, New Brunswick NJ, June 13–16, 1991.

Kuhfeld, W.F. (2003), "Marketing Research Methods in SAS,"
[http://support.sas.com/techsup/tnote/tnote_stat.html#market].

Kuhfeld, W.F. (1990), *SAS Technical Report R-108: Algorithms for the PRINQUAL and TRANSREG Procedures*, Cary NC: SAS Institute Inc.

Kuhfeld, W.F., Tobias, R.D., and Garratt, M. (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, 31, 545–557.

Kuhfeld, W.F. and Garratt, M. (1992), "Linear Models and Conjoint Analysis with Nonlinear Spline Transformations," Paper presented to the American Marketing Association Advanced Research Techniques Forum, Lake Tahoe, Nevada.

Lazari, A.G. and Anderson, D.A. (1994), "Designs of Discrete Choice Set Experiments for Estimating Both Attribute and Availability Cross Effects," *Journal of Marketing Research*, 31, 375–383.

Lebart, L., Morineau, A., and Warwick, K.M. (1984), *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*, New York: Wiley.

de Leeuw, J. (1986), "Regression with Optimal Scaling of the Dependent Variable," Department of Data Theory, The University of Leiden, The Netherlands.

de Leeuw, J., Young, F.W., and Takane, Y. (1976), "Additive Structure in Qualitative Data: an Alternating Least Squares Method with Optimal Scaling features," *Psychometrika*, 41, 471–503.

Louviere, J.J. (1988), *Analyzing Decision Making, Metric Conjoint Analysis*, Sage University Papers, Beverly Hills: Sage.

Louviere, J.J. (1991), "Consumer Choice Models and the Design and Analysis of Choice Experiments," Tutorial presented to the American Marketing Association Advanced Research Techniques Forum, Beaver Creek, Colorado.

Louviere, J.J. and Woodworth, G. (1983), "Design and Analysis of Simulated Consumer Choice of Allocation Experiments: A Method Based on Aggregate Data," *Journal of Marketing Research*, 20 (November), 350–367.

Louviere, J.J. (1991), "Consumer Choice Models and the Design and Analysis of Choice Experiments," Tutorial presented to the American Marketing Association Advanced Research Techniques Forum, Beaver Creek, Colorado.

Manski, C.F. and McFadden, D. (1981), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge: MIT Press.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), Multivariate Analysis, New York: Academic Press.

McFadden, D. (1974), "Conditional logit Analysis Of Qualitative Choice Behavior," in P. Zarembka (ed.) *Frontiers in Econometrics,* New York: Academic Press, 105–142.

McKelvey, R.D. and Zavoina, W. (1975), "A Statistical Model for the Analysis Of Ordinal Level Dependent Variables," *Journal of Mathematical Sociology,* 4, 103–120.

Meyer, R.K. and Nachtsheim, C.J. (1995), "The Coordinate-Exchange Algorithm for Constructing Exact Optimal Experimental Designs," *Technometrics*, 37, 60–69.

Mitchell, T.J. and Miller, F.L. Jr. (1970), "Use of Design Repair to Construct Designs for Special Linear Models," *Math. Div. Ann. Progr. Rept. (ORNL-4661)*, 130–131, Oak Ridge, TN: Oak Ridge National Laboratory.

Mitchell, T.J. (1974), "An Algorithm for the Construction of *D*-optimal Experimental Designs," *Technometrics*, 16 (May), 203–210.

Nishisato, S. (1980), *Analysis of Categorical Data: Dual Scaling and Its Applications*, Toronto: University of Toronto Press.

Paley, R.E.A.C (1933), "On Orthogonal Matrices," *J. Math. Phys*, 12, 311–320.

Perreault, W.D. and Young, F.W. (1980), "Alternating Least Squares Optimal Scaling: Analysis of Nonmetric Data in Marketing Research," *Journal of Marketing Research*, 17, 1–13.

Raktoe, B.L., Hedayat, A.S., and Federer, W.T. (1981), *Factorial Designs*, New York: John Wiley and Sons.

Ramsay, J.O. (1988), "Monotone Regression Splines in Action," *Statistical Science*, 3, 425–461.

Rao, C.R. (1947), "Factorial Experiments Derivable from Combinatorial Arrangements of Arrays," *Journal of the Royal Statistical Society*, Suppl., 9, 128–139.

van Rijckevorsel, J. (1982), "Canonical Analysis with B-splines," in H. Caussinus, P. Ettinger, and R. Tomassone (ed.), *COMPUSTAT 1982*, Part I, Vienna: Physica Verlag.

Schiffman, S.S., Reynolds, M.L., and Young, F.W. (1981), *Introduction to Multidimensional Scaling*, New York: Academic Press.

Sloane, N.J.A. (2002), "A Library of Orthogonal Arrays," [http://www.research.att.com/∼njas/oadir].

Smith, P.L. (1979), "Splines as a Useful and Convenient Statistical Tool," *The American Statistician*, 33, 57–62.

Steckel, J.H., DeSarbo, W.S., and Mahajan, V. (1991), "On the Creation of Acceptable Conjoint Analysis Experimental Designs," *Decision Sciences*, 22, 435–442.

Suen, C.Y. (1989a), "A Class of Orthogonal Main Effects Plans," *Journal of Statistical Planning and Inference*, 21, 391–394.

Suen, C.Y. (1989b), "Some Resolvable Orthogonal Arrays with Two Symbols," *Communications in Statistics, Theory and Methods*, 18, 3875–3881.

Taguchi, G. (1987), *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Costs. White Plains*, NY: UNIPIB, and Dearborn, MI: American Supplier Institute.

Tenenhaus, M. and Young, F.W. (1985), "An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis, and Other Methods of Quantifying Categorical Multivariate Data," *Psychometrika*, 50, 91–119.

Wang, J.C. (1996a), "Mixed Difference Matrices and the Construction of Orthogonal Arrays," *Statist. Probab. Lett.*, 28, 121–126.

Wang, J.C. (1996b), *A Recursive Construction of Orthogonal Arrays*, Preprint.

Wang, J.C. and Wu, C.F.J. (1989), "An Approach to the Construction of Asymmetrical Orthogonal Arrays," *IIQP Research Report RR-89-01*, University of Waterloo.

Wang, J.C. and Wu, C.F.J. (1991), "An Approach to the Construction of Asymmetrical Orthogonal Arrays," *Journal of the American Statistical Association*, 86, 450–456.

Williamson, J. (1944), "Hadamard's Determinant Theorem and the Sum of Four Squares," Duke Math. J., 11, 65–81.

Winsberg, S. and Ramsay, J.O. (1980), "Monotonic Transformations to Additivity Using Splines," *Biometrika*, 67, 669–674.

Wittink, D.R. and Cattin, P. (1989), "Commercial Use of Conjoint Analysis: An Update," *Journal of Marketing*, 53 (July), 91–96.

Wittink, D.R., Krishnamurthi, L., and Reibstein , D.J. (1989), "The Effect of Differences in the Number of Attribute Levels in Conjoint Results," *Marketing Letters*, *1:2*, 113–123.

Xu, H. (2002), "An Algorithm for Constructing Orthogonal and Nearly Orthogonal Arrays with Mixed Levels and Small Runs," *Technometrics*, 44, 356–368.

Young, F.W. (1981), "Quantitative Analysis of Qualitative Data," *Psychometrika*, 46, 357–388.

Young, F.W. (1987), *Multidimensional Scaling: History, Theory, and Applications*, R.M. Hamer (ed.), Hillsdale, NJ: Lawrence Erlbaum Associates.

Young, F.W., de Leeuw, J., and Takane, Y. (1976), "Regression with Qualitative and Quantitative Variables: an Alternating Least Squares Approach with Optimal Scaling Features," *Psychometrika*, 41, 505–529.

Zhang, Y.S., Lu, Y., and Pang, S. (1999), "Orthogonal Arrays Obtained by Orthogonal Decompositions of Projection Matrices," *Statistica Sinica*, 9, 595–604.

Zhang, Y.S., Pang, S., and Wang, Y. (2001), "Orthogonal Arrays Obtained by Generalized Hadamard Product," *Discrete Mathematics*, 238, 151–170.

Index