New Developments in Psychometric and Market Research Procedures
Warren F. Kuhfeld, SAS Institute Inc., Cary, NC
Forrest W. Young, University of North Carolina at Chapel Hill

ABSTRACT
Three new SAS® procedures, PROC TRANSREG (regression analysis with nonlinear variable transformations), PROC PRINQUAL (principal components analysis with nonlinear variable transformations), and PROC CORRESP (correspondence analysis) will be available in a later version of Release 6.03 of SAS/STAT software. These procedures were developed initially at the University of North Carolina at Chapel Hill under a grant from SAS Institute. This paper discusses the current design of these procedures.

INTRODUCTION
For three years, from spring of 1984 to spring of 1987, SAS Institute funded a software research and development project at the University of North Carolina at Chapel Hill. During the course of this project, we presented eight SUGI papers, which are reviewed here. In these papers we describe our plans, and later our progress, in making new graphical, psychometric, and market research data analysis methods available to SAS software users. The project ended almost a year ago, at the end of the work that started at UNC continues at SAS Institute. This paper describes the current state of the software.

At SUGI 10, Kuhfeld, Sarle, and Young (1985) described the algorithms available in the PRINQUAL macro. Later, the PRINQUAL macro was enhanced and made available in the SAS Sample Library (see Young and Kuhfeld, 1985). PROC PRINQUAL was released for testing in Release 6.02, and an enhanced version of the procedure will be a late addition to SAS/STAT software in Release 6.03. More is said about this procedure later in this paper.

At SUGI 10, Young (1985) announced plans for the TRANSREG procedure (transformation regression: fitting a linear model to optimally transformed data) and the CONJOINT procedure (conjoint analysis: fitting an ANOVA model to an ordinal dependent variable). At SUGI 11, Kuhfeld, et. al. described a Release 6.02 CONJOINT procedure; however, after many enhancements the procedure evolved into PROC TRANSREG. The conjoint analysis model is a special case TRANSREG model, so a separate CONJOINT procedure is not planned. Conjoint analysis is also available in Version 5 using the CONJOINT macro (see Young, 1985). The current design of this procedure is discussed later in this paper.

Young (1985) announced plans for PROC MCORESP, a multiple correspondence analysis procedure. The CORRESP procedure for simple and multiple correspondence analysis was made available for testing in Release 6.02, and an enhanced version of the procedure will be a late addition to Release 6.03 of SAS/STAT software. The early design was discussed by Kuhfeld, et. al. (1986). The current design of this procedure is discussed later in this paper.


Also at SUGI 10, Young (1985) announced plans for a PATHANAL procedure (path analysis or covariance structure analysis). This procedure, under the new name CALIS procedure (covariance analysis of linear structures), is currently being developed at SAS Institute by Wolfgang Hartmann. The release date is unknown.

The HYPERSPACE procedure, an interactive hyperdimensional scatterplot program, was first mentioned at SUGI 10 (Young, 1985). Initially, the program was based on a metaphor of the data analyst being a pilot, touring a multidimensional data space by flying through three-dimensional subspaces. The metaphor was changed so that the analyst sat outside and rotated and manipulated the data space, and the name was changed to VISUALS (Kent, Edds, Kuhfeld, and Young, 1986; Young, Kent, Edds, and Kuhfeld, 1986; Young, Kent, and Kuhfeld, 1987). VISUALS is an acronym for Viewing Internal Structure Using Alternating Least Squares. Alternating least squares methodology, however, was never incorporated into the program.

The PROXSCAL procedure was planned for future releases. At SUGI 10, Young (1985) announced several other procedures. The PROXSCAL procedure was planned to be a multidimensional scaling procedure that optimizes the fit of distances to transformed data using the algorithm of de Leeuw and Heiser (1980). A stand-alone FORTRAN program is available from them, but as of this writing it has not been converted to a SAS procedure. Multidimensional scaling is still available under OS and CMS with the ALSCAL procedure (Young, Lewyckyj, and Takane, 1986, in SUGI Supplemental Library User's Guide, Version 5 Edition). No work has been done on PRIN3WAY, a three-mode principal components analysis program. Also, the SPREADSPACE procedure, which would produce a matrix of three-dimensional scatterplots, was never developed.

THE CORRESP PROCEDURE
The CORRESP procedure performs simple and multiple correspondence analysis (Lebart, Morineau, and Warwick, 1984; Greenacre, 1984). Input can be raw category responses on two or more classification variables, category frequency data on two or more classification variables, a binary category response indicator matrix, or a two-way contingency table. Burt tables can be input directly or created from categorical variables. Supplemen-
tary variables and observations can also be input. The CORRESP procedure produces an output data set that can be used by plotting procedures such as the IDPLOT, PLOT, GPLOT, and G3D procedures.

The CORRESP procedure is controlled by the following statements:

```
PROC CORRESP options;
  TABLES variables[,variables];
  VAR variables;
  WEIGHT variables;
  SUPVAR variables;
  ID variable;
  BY variables;
```

Correspondence analysis can be described as a weighted principal component analysis of a contingency table. Let \( N \) be the contingency table formed from those observations and variables that are not supplementary and those observations that have no missing values and a positive weight. This table is an \((n_x n_y)\) rank \( r \) matrix of nonnegative numbers with nonzero row and column sums. Define \( P = (1/f) N \), where \( f \) is the total frequency (grand sum) of the elements of \( N \). Let \( B \) be a vector of ones of the appropriate order, and let \( \text{diag}(b) \) be a matrix-valued function that creates a diagonal matrix from a vector. Define \( r = P1, c = P'1, D, = \text{diag}(r), \) and \( D_c = \text{diag}(c) \).

The correspondence analysis is defined in terms of the generalized singular value decomposition of \( P \) or \( P' \). Subtracting \( re' \) centers the matrix of cell proportions, giving differences between the observed proportions and the expected proportions under the hypothesis of row and column independence. \( P = AD'B' \), where \( A'D'r' = I - 1 \) and \( I \) is an identity matrix. \( A \) is the matrix of left singular vectors, having \( n_x \) rows and \( r \) columns; \( D, \) is a diagonal matrix of singular values, having \( r \) rows and columns; and \( B \) is a rectangular matrix of right singular vectors having \( n_y \) rows and \( r \) columns. The first singular value, one, the first left vector, \( r \), and the first right vector, \( c \), are discarded before any correspondence analysis results are displayed. There are six choices of standardized row coordinates: \( D,^{1/2}A, \) \( D,^{1/2}A'\), \( D,^{1/2}A, \) \( D,^{1/2}A'\), \( D,^{1/2}A\) \( A'\), and \( D,^{1/2}A\) \( A'\). Similarly, there are six choices of standardized column coordinates: \( B, \) \( B'\), \( B, \) \( B'\), \( B\) \( B'\), and \( B, \) \( B'\). The usual sets of coordinates are the defaults: \( D,^{1/2}A, \) and \( D,^{1/2}B, \).

There are two new options for row and column points. Row coordinates \( D,^{1/2}A, \) and column coordinates \( D,^{1/2}B, \) form a symmetric decomposition of \( D,^{1/2}P, \) (Gill, 1981; van der Heijden and de Leeuw, 1985). In all of the pairs above, distances between row and column points are not meaningful. When the row coordinates are \( D,^{1/2}A\) \( A'\), distances both between and within sets are comparable (Carroll, Green, and Schaffer, 1986).

There are several other new options available in the Release 6.03 CORRESP procedure. The most notable change is that all correspondence analysis results are available in the output coordinate (OUTC=) data set. The output coordinate data set contains inertias, coordinates, partial contributions to inertia, squared cosines, quality of representation, mass, and indexes of which points contribute most to each dimension's inertia. The index statistics are new. Another new feature is the ability to designate categorical TABLES statement input variables as supplementary with the SUPVAR statement. This provides a simple mechanism for designating Burt table partitions as supplementary.

### THE TRANSREG PROCEDURE

The TRANSREG procedure obtains linear and nonlinear transformations of variables to optimize the least-squares fit of the data to a variety of linear models. These include simple, multiple, and multivariate regression; main effects, reference cell, and cell mean ANOVA models including joint analysis; vector and ideal point preference regression or external unfolding (Carroll, 1972); redundancy analysis; canonical correlation analysis; and response surface regression.

The TRANSREG procedure is controlled by the following statements:

```
PROC TRANSREG options;
  MODEL t1{independents[options]} [t2{dependents[options]}
    ...]
    [t3{independents[options]}
    ...]
    [options];
  OUTPUT options;
  ID variables;
  BY variables;
```

The MODEL statement is used to name the dependent and independent variables, to name the transformations of these variables, to supply transformation-specific options, and to specify analysis parameters. \( t1, t2, \) and so on are transformation (or expansion) names that specify transformation families for the enclosed variables. For some transformation families, transformation options may be specified after a slash, within the paren-

---

**References:**
- Gill, 1981.
- Carroll, Green, and Schaffer, 1986.
theses. Any transformation can be used with either dependent or independent variables. There are three classes of names:

- Variable expansions preprocess the specified variables, replacing them with more variables.
- Nonoptimal transformations preprocess the specified variables, replacing each one with a single new nonoptimal, nonlinear transformation.
- Optimal transformations are derived iteratively. Each specified variable is replaced by a single new optimal transformation, which fits the specified model better than the original variable (except for contrived cases where the transformation fits the model exactly as well as the original variable). The optimal transformation capabilities should never be used as a preliminary step in hypothesis testing. The usual assumptions that must be made to test hypotheses in general linear model analyses cannot be made after variables are optimally transformed. The usual test statistics reported by the REG and GLM procedures are not valid for optimally transformed variables.

Variable Expansions

Variable expansions replace the original variables with a larger set of new variables. The resulting variables are not transformed by the iterative algorithms after the initial preprocessing. The POINT, EPOINT, and QPOINT expansions are used in preference analyses (PREFMAP, external unfolding, ideal-point regression; Carroll, 1972) and for response-surface regressions.

- CLASS names discrete classification variables that are expanded to dummy variables.
- EPOINT names variables for elliptical response-surface regression or elliptical ideal-point regression. EPOINT adds a set of new variables, the squares of the original variables.
- POINT names variables for a circular response-surface regression or circular ideal-point regression. POINT adds a new variable whose value for each observation is the sum of squares of all the POINT variables.
- QPOINT names variables for a quadratic response-surface regression or quadratic ideal-point regression. QPOINT adds a set of new variables equal to the squares and crossproducts of the original variables.

Nonoptimal Variable Transformations

The nonoptimal transformations replace each original variable by a single transformed variable:

- ARSIN names variables for an inverse trigonometric sine transformation.
- EXP names variables to be exponentiated ($a^x$ for data value $x$ and parameter $a$).
- LOG names variables to be transformed to logarithms ($\log_x$ for data value $x$ and parameter $a$).
- LOGIT names variables for a logit transformation ($\log(x / (1 - x))$).
- POWER names variables that are to be raised to a power ($a^x$ for data value $x$ and parameter $a$).

RANK names variables that are to be transformed to ranks, which are averaged within ties.

Optimal Variable Transformations

The following name families from which iteratively derived optimal transformations are computed. Missing values for these types of variables can be optimally estimated. The optimal transformation capabilities should never be used as a preliminary step in hypothesis testing. The usual assumptions that must be made to test hypotheses in general linear model analyses cannot be made after variables are optimally transformed. The usual test statistics reported by the REG and GLM procedures are not valid for optimally transformed variables.

- LINEAR names variables for a linear transformation (change of origin and scale only).
- MONOTONE names variables for a monotonic transformation with the restriction that ties are preserved (Kruskal, 1964).
- MSPLINE names variables for a monotonic B-spline transformation (de Boor, 1978; de Leeuw, 1986).
- OPSCORE names variables whose categories are to be optimally scored (Fisher, 1938).
- SPLINE names variables for a B-spline transformation (de Boor, 1978).
- UNTIE names variables for a monotonic transformation without the restriction that ties be preserved (Kruskal, 1964).

TRANSREG Procedure Algorithms

The TRANSREG procedure has four iterative algorithms. The UNIVARIATE algorithm (based on Young, de Leeuw, and Takane, 1976) transforms each dependent variable to maximize the squared multiple correlation. Independent variables are not transformed. This algorithm is used for conjoint analysis, metric and nonmetric PREFMAP and external unfolding analyses (Carroll, 1972), and multiple regression analyses with dependent variable transformations.

The MORALS algorithm (multiple optimal regression using alternating least squares based on Young, et al., 1976) specifies that each dependent variable be transformed along with the set of independent variables to maximize the squared multiple correlation. If there are $m$ dependent variables, $m$ single dependent variable linear models are fit.

The REDUNDANCY algorithm (an extension of Young, et al., 1976) specifies that all dependent variables and independent variables be jointly transformed to maximize the average of the squared multiple correlations.

The CANALS algorithm (canonical correlation analysis with alternating least squares based on the CANALS method of van der Burg and de Leeuw, 1983) specifies that all dependent variables and independent variables be jointly transformed to maximize the average of the first $r$ squared canonical correlations.

Missing Value Estimation

The TRANSREG procedure has very flexible missing value estimation capabilities. Missing values within OPSCORE, MONOTONE, UNTIE, LINEAR, SPLINE, and MSPLINE transformation variables can be estimated so that the variance accounted
for by the linear model is maximized. No category or order restrictions are placed on the estimates of ordinary missing values (._). The twenty-seven special missing values (._ and .A through .Z) can be used to indicate categorical missing values whose estimates within class and variable must be identical. In addition, some missing value categories can be ordered. The MONOTONE= option can be used to indicate a range of special missing values from the list .A to .Z whose estimates must be weakly ordered within each variable in which they appear. The UNtie= option can be used to indicate a range of special missing values whose estimates must be weakly ordered but can be untied.

THE PRINQUAL PROCEDURE

The PRINQUAL procedure has three algorithms for transforming a set of variables. One algorithm follows from a principal-components model, the second from a regression model, and the third from a regression-like model. The PRINQUAL procedure can be viewed as a generalization of ordinary principal-components analysis to a method capable of analyzing data that are not quantitative. It can be used to perform metric and nonmetric MDPref analyses (Carroll 1972). It can be viewed as a data preprocessor, used to transform data prior to their use in any other data analysis. In addition, the PRINQUAL procedure has the same missing data estimation facilities discussed in the previous section on the TRANSREG procedure.

The PRINQUAL procedure is controlled by the following statements:

PROC PRINQUAL options;
VAR variables [/ options];
ID variables;
BY variables;

The PRINQUAL procedure has been discussed in detail in previous SUGI papers, so only its new features are mentioned here. The PRINQUAL procedure in Release 6.03 has all of the optimal and nonoptimal transformations discussed with the TRANSREG procedure (but not the expansions). The SPLINE and MSPLINE transformations and the nonoptimal transformations are new with Release 6.03.

REFERENCES


SAS is a registered trademark of SAS Institute Inc., Cary, NC.
SAS/STAT is a trademark of SAS Institute Inc., Cary, NC.

IBM is a registered trademark of International Business Machines Corporation, Armonk, NY.