SAS Institute is funding a software research and development project at the University of North Carolina at Chapel Hill whose purpose is to develop a package of Psychometric and Market Research procedures. To date, this project has developed three SAS procedures, one stand-alone program, and one SAS macro. Four additional procedures are under development. In this paper we present a brief overview of the entire project, and then discuss PROC CONJOINT and PROC TRANSREG, new procedures for transformation regression which perform conjoint analysis, preference mapping, external unfolding, nonmetric regression, and a wide variety of additional analyses.

1. Project Overview

The Psychometric and Market Research software that has already been developed at the University of North Carolina includes PROC IDPLOT (Kuhfeld, 1986), a currently released version 5 procedure; PRINQUAL, a currently released version 5 macro for principal component analysis of qualitative data (Kuhfeld & Sarle, 1987); PROCs PRINQUAL and CORRESP (correspondence analysis), experimental version 6 procedures currently available for testing (Kuhfeld, Edds, Kent and Young, 1986; Young & Kuhfeld, 1986a; 1986b); and VISUALS (dynamic hyperdimensional graphics), a stand-alone program which is currently available for testing (Young, Edds, Kent & Kuhfeld, 1986; Kent, Edds, Kuhfeld & Young, 1986; Young, Kent & Kuhfeld, 1987). The design and function of these programs has been reported elsewhere and will not be presented here.

Two additional experimental version 6 procedures, PROC CONJOINT (conjoint analysis) and PROC TRANSREG (transformation regression), will be available for testing soon. This paper focuses on these procedures.

Finally, PROC's PROXSCAL (multidimensional scaling) and COSAN (covariance structure analysis) are in the initial stages of development. These version 6 procedures are not discussed here.

2. Introduction

PROC TRANSREG (Young and Kuhfeld, 1987) obtains linear and nonlinear transformations of variables to optimize the least-squares fit of the data to a variety of linear models including: Conjoint Analysis; vector and ideal point preference mapping (metric or nonmetric); external unfolding; simple, multiple, and multivariate regression with variable transformation; ANOVA models including generalized conjoint analysis; redundancy analysis; canonical correlation analysis; and response surface regression.

TRANSREG is a transformation procedure. It transforms variables to optimize regression models. The independent and dependent variables may be categorical, ordinal or quantitative. Any mix is allowed. Categorical variables may be transformed by scoring the categories to minimize least-squares error, or may be expanded into dummy variables. Ordinal variables may be transformed monotonically by scoring the ordered categories so that order is preserved and least-squares error is minimized. Ties may be optimally untied or left tied. Ordinal variables may instead be transformed to ranks. Quantitative variables may be smoothly transformed using spline or monotone spline transformations, or may be linearly transformed. Missing data may estimated without constraint, or with category or order constraints.

TRANSREG is a scoring procedure. It generates very little printed output. It produces an output data set which contains the transformed (optimally scored) variables. These variables may be input to PROCs REG, CANCORR, ANOVA or GLM to obtain final analyses; or input to a plotting procedure for obtaining preference biplots, transformations plots, and so on.

PROC TRANSREG is an alternating least-squares optimal scaling procedure. This means its algorithms are convergent. At every step, the algorithms minimize least-squares error, and hence maximize (depending on which iterative algorithms is specified) the squared multiple correlation (UNIVARIATE and MORALS), the average of squared multiple correlations (REDUNDANCY), or the average of squared canonical correlations (CANALS).

For more background on alternating least-squares optimal scaling methods and transformation regression methods see Young, de Leeuw, and Takane (1976), Gifi (1981), van der Burg and de Leeuw (1983), Young (1981), Izraels (1984), de Leeuw (1986), de Leeuw and Kuhfeld (in press), Breiman and Friedman (1985), Hastie and Tibshirani (1986), and Schiffman, Reynolds, and Young (1981), to name a few of the many relevant works. Also see the PROC TRANSREG documentation (Young and Kuhfeld, 1987).

3. Expanding and Transforming Variables

The PROC TRANSREG expands or transforms variables so as to optimize regression models. In this section we describe the variable expansion methods and the least-squares transformation families.

3.1. Variable Expansions. The following are ways in which variables may be preprocessed prior to
the start of the iterative algorithms. The variable expansions either replace the original variable with one (RANK) or more (CLASS) new columns, or add new variables (POINT, EPOINT, and QPOINT) to the same set (independent or dependent variables) as the originals. The POINT, EPOINT and QPOINT functions are used in preference (PREFMAP or external unfolding, Carroll, 1972) analyses, for ideal point regression analyses, and for response surface regressions. POINT, EPOINT, and QPOINT create circular, elliptical, and quadratic response or preference surfaces. Those variables are not transformed by the iterative algorithms after the initial preprocessing (except for a linear scaling). Observations with missing values for these types of variables are excluded from the analysis. The names of the expansion functions, and their effects, are as follows:

POINT names continuous numeric variables used for a circular response surface regression or circular ideal point regression. POINT creates a new variable whose value for each observation is the sum of squares of all the POINT variables. This new variable is added to the set of variables and is used in the regression analysis.

EPOINT names continuous numeric variables used for elliptical response surface regression or elliptical ideal point regression. EPOINT creates a new variable for each original EPOINT variable. The value of each new variable is the square of each observed value for the corresponding original variable. Both sets of variables (original and squared) are then used in the regression analysis.

QPOINT names continuous numeric variables used for a quadratic response surface regression or quadratic ideal point regression. QPOINT creates a set of new variables by crossing the QPOINT variables. If there are m QPOINT variables, the m(m + 1)/2 unique pairs of variable i times (element-wise product) variable j (where i may equal j) are created. Both sets (original and crossed) are combined to be used in the regression analysis.

CLASS names character or discrete numeric variables used for classification. A CLASS variable is expanded to dummy variables. Up to the first eight characters of the formatted variable’s value are used to determine class membership.

RANK names numeric variables that are transformed to ranks before the analysis is begun.

3.2. Variable Transformation Families

The following are ways in which variables may be iteratively transformed. All transformations can be used with either independent or dependent variables. Missing values for these types of variables may be optimally estimated.

OPSCORE names character or discrete numeric variables that are to be optimally scored. OPSCORE assigns scores to each class (level) of the variable using Fisher’s (1938) optimal scoring method.

MONOTONE names numeric, usually discrete, variables that are to be transformed monotonically with the restriction that ties are preserved. The Kruskal and Shepard (1974) secondary least-squares monotonic transformation is used. This transformation weakly preserves order and category membership (ties).

UNTIE names numeric, usually discrete variables that are to be transformed monotonically without the restriction that ties be preserved. The Kruskal and Shepard (1974) primary least-squares monotonic transformation method is used. This transformation weakly preserves order but not category membership (it unties tied values).

LINEAR names continuous numeric variables that are subject to a linear transformation - change of origin and scale only.

SPLINE names continuous numeric variables that are subject to a piece-wise polynomial B-spline (de Boor, 1978) transformation. By default a cubic polynomial transformation is used. Knots and other degrees may be specified.

ISPLINE names continuous numeric variables that are subject to an increasing (monotonic) B-spline transformation. By default a quadratic polynomial is used. Knots and other degrees may be specified.

3.3. Expansion and Transformation Name Usage

Expansion and transformation names are used on the MODEL statement. Any mix of the expansions and transformations can be used to fit a mixed measurement level model. Here are some examples:

MODEL LINEAR(Y1-Y5) = LINEAR(X1-X4); specifies ordinary multivariate multiple regression analysis if METHOD=REDUNDANCY, or canonical analysis if METHOD=CAKALS.

MODEL MONOTONE(Y1) = CLASS(X1-X2); specifies simple conjoint analysis.

MODEL MONOTONE(PREF1-PREF10) = POINT(DIM1-DIM3); specifies nonmetric preference analysis using the ideal point preference model.

MODEL MONOTONE(Y1-Y2) OPSCORE(Y3-Y5) = LINEAR(X1) UNTIE(X2) MONOTONE(X3-X4); specifies a mixed measurement level multiple regression analysis.
4. Regression Methods

PROC TRANSREG provides for a wide variety of regression models. These are based on four classes of algorithms, which are reviewed in this section.

4.1. METHOD=UNIVARIATE This method is based on the Young, de Leeuw, and Takane (1976) algorithm. It specifies that each dependent variable be transformed to maximize the squared multiple correlation while the independent variables are not transformed. The METHOD=UNIVARIATE algorithm is a generalization of the ordinary univariate general linear model to allow for dependent variable transformations. This algorithm is used for conjoint analysis, metric and nonmetric preference analysis, external unfolding analyses, and multiple regression analyses with dependent variable transformations.

4.2. METHOD=MORALS This method (Multiple Optimal Regression by Alternating Least Square) is based on the Young, de Leeuw, and Takane (1976) algorithm. It specifies that each dependent variable be transformed along with the set of independent variables to maximize the squared multiple correlation. The METHOD=MORALS algorithm is a univariate generalization of the METHOD=UNIVARIATE algorithm to allow for both dependent and independent variable transformations. While more than one (say m) dependent variables may be specified on the MODEL statement, METHOD=MORALS is a univariate generalization because m linear models are fit, with each linear model containing only one dependent variable at a time.

4.3. METHOD=REDUNDANCY This method is an extension of Young, de Leeuw, and Takane (1976). It specifies that all dependent and independent variables be jointly transformed to maximize the average of the squared multiple correlations. The METHOD=REDUNDANCY algorithm is a multivariate generalization of the METHOD=UNIVARIATE algorithm to allow for both dependent and independent variable transformations. One linear model is fit which transforms each variable only once, maximizing the average squared multiple correlation.

4.4. METHOD=CANALS This method (CAnonical correlation analysis with Alternating Least Squares) is based on the van der Burg and de Leeuw (1983) algorithm. It specifies that all dependent variables and independent variables be jointly transformed to maximize the average of the first r canonical correlations. METHOD=CANALS is a multivariate method which transforms all of the variables together, finding only one transformation of each variable.

5. PROC TRANSREG Options

There are a number of options available in PROC TRANSREG. A very brief description of the options will be provided here. The data set options are: DATA= data set name and OUT= data set name.

The options that control the iterative algorithms are: METHOD=name specifies the iterative algorithms, MAXITER=n specifies the maximum number of iterations, CONVERGE=n specifies the convergence criterion, NCAN=n specifies the number of canonical variables, and SINGULAR=n specifies the singularity criterion.

The output data set score partition options control what is placed in the output data set. The options are: REPLACE specifies independent and dependent variables are replaced by their transformed values, TREPLACE specifies independent variables are replaced by their transformed values, DREPLACE specifies dependent variables are replaced by their transformed values, APPROXIMATIONS specifies that both dependent and independent variable approximations are output, MRC specifies that multiple regression coefficients be output, MPC specifies that point model ideal point coordinates be output, MEC specifies that elliptical point model ideal point coordinates be output, NPC specifies that quadratic point model ideal point coordinates be output, CCC specifies that canonical coefficients be output, CPC specifies that canonical point model ideal point coordinates be output, CEC specifies that canonical elliptical point model ideal point coordinates be output, and CQC specifies that canonical quadratic point model ideal point coordinates be output.

The output data set coefficient partition options are: COEFFICIENTS specifies that the linear model coefficients be included in the output data, UTILITIES specifies that conjoint utilities be output, MEANS specifies that marginal means for CLASS variable columns be output, UTILITIES specifies that conjoint utilities be output, NOSCORES specifies that no score partition information be output.

The output data set variable name prefixes provide prefixes for naming the transformed and approximation variables in the output data set: TDPREFIX=name specifies the transformed dependent variables prefix, TIPREFIX=name specifies the transformed dependent variables prefix, ADPREFIX=name specifies the approximations to the dependent variables prefix, AIPREFIX=name specifies the approximations to the independent variables prefix, and CPREFIX=n specifies the number of first characters of a CLASS variable's name that are to be used in constructing binary variable names for the output data set.

The design matrix options are used to specify the details of how the intercept and binary CLASS variables are created. The options are: NOINT specifies that no intercept term be added.
to the independent variables, DINT specifies that an intercept term be added to the dependent variables, NODINT specifies that an intercept term not be added to the dependent variables, and ALLCATS says there is one binary variable created per category of each CLASS variable.

The remaining options are: TSTANDARD+ specifies that how the means and variances of the transformed variables should be set, NOMISS specifies that all observations with missing values to be excluded from the analysis, NOPRINT specifies that there be no printed output, ORDCAT=names specifies special missing values estimated with within variable order constraints.

A POLYNOMIAL statement may be used to specify knots and polynomial degrees for variables that may be subjected to a SPLINE or ISPLINE transformation: DEGREE=n specifies the degree of the polynomial spline transformation, KNOTS=number-list specifies the interior knots or break points, NNKNOTS=n specifies that PROC TRANSREG should create n knots at evenly spaced percentiles. Knots may be repeated to indicate breaks in lower order derivatives.

Finally, an ID statement can be used to list additional character and/or numeric variables that are to be included in the output data set, and a BY statement can be used to obtain separate analyses on observation groups.

6. Missing Value Estimation

PROC TRANSREG has very flexible missing value estimation capabilities. Missing values within OPSCORE, MONOTONE, UNTIE, LINEAR, SPLINE, and ISPLINE transformation variables may be estimated so that the variance accounted for by the linear model is maximized. A variety of estimation restrictions are provided. No category or order restrictions are placed on the estimates of ordinary missing values (.). The 27 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 may be ordered. The ORDCAT= 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 that they appear.

The OPSCORE, MONOTONE, UNTIE, LINEAR, SPLINE, and ISPLINE functions can be combined with nonmissing values, ordinary missing values, special missing values, and ordered categorical missing values, in any way, as long as there are nonmissing values within each variable that have some variance. The missing value estimation facilities allow for partitioned variables. For example, a LINEAR variable with special missing and ordered categorical missing can be part interval, part ordinal, and part nominal. A MONOTONE variable may have two independent ordinal parts and nominal classes. An UNTIE variable may have an ordered categorical part and an ordered part without category restrictions. There are many more possible examples.

7. Examples

7.1. Conjoint Analysis These statements

```
PROC TRANSREG COEFFICIENTS;
MODEL MONOTONE(Y) = CLASS(X1 X2);
```

perform a simple conjoint analysis, monotonically transforming the variable Y to fit a main effects analysis of variance model with factors X1 and X2. PROC TRANSREG's output data set includes:

```
Y  TY _INTER_  X1A  X1B  X2A  X2B  X1  X2
6 7.39  1.00  . 00  0.0  a  a
7 7.39  1.00  . 00  .0  a  a
4 4.25  1.00  . 00  1.0  a  b
3 4.25  1.00  . 00  1.0  a  b
5 4.25  1.00  1.00  8.0  b  a
4 4.25  1.00  1.00  8.0  b  a
2 1.11  1.00  1.00  1.0  b  b
1 1.11  1.00  1.00  1.0  b  b
. . 7.39  -3.1  -3.1
. . . . 5.62  2.68  5.62  2.68
```

Y is the original dependent variable and TY contains the monotone transformation. The transformed variable has the same total sum of squares as the original, but the within cell and interaction sum of squares have been completely eliminated. X1 and X2 are the independent classification variables, and _INTER_ -- X2B contain the main effects design matrix. The coefficients of the final linear model are displayed in the second last observation and the marginal means in the last observation.

7.2. Nonlinear Redundancy Analysis The following example illustrates using PROC TRANSREG for maximizing the average squared multiple correlation of a set of mixed measurement level variables, with (far too many) missing values. The input data are:

```
Y1  Y2  Y3  X1  X2  X3  X4  X5  X6  X7
A  A  9  1  1  1  A  1  b  b
A  A  9  1  1  1  A  1  a  a
B  B  . 2  1  1  B  G  . 1
B  B  . 2  1  1  B  G  . 1
C  C  . 2  1  1  C  B  . 1
C  C  . 2  1  1  C  B  . 1
7 7  . 2  1  1  7  1  1  1
8 8  . 2  1  1  8  1  1  1
9 9  . 2  1  1  9  1  1  1
```

Y1 and Y2 are between-subjects classification variables, X1 and X2 are the independent variables, X3 through X7 are the dependent variables, and the statements:

```
PROC TRANSREG MAXITER=50 CONVERGE=-1 REPLACE ORDCAT=QI
METHOD=REDUNDANCY;
MODEL MONOTONE(Y1) OPSCORE(Y2-Y3) = OPSCORE(X1-X2) LINEAR(X3) UNTIE(X4);
```

were invoked with the statements:
original variables with the transformed variables in the output data set. ORDCAT=GI specifies that missing values .G, .H, and .I are to be estimated with both order and category constraints, and METHOD=REDUNDANCY specifies the iterative algorithm.

The MODEL statement indicates dependent variables before the equal sign and independent variables after, with transformation names and applicable variables specified. PROC TRANSREG increased the average squared multiple correlation from 0.82893 (computed using the initial scoring of nonmissing data and missing data initialized to the column mean) to a value greater than 0.99999.

It can be seen that the final scoring of nonmissing values conform to the transformation family restrictions, ordinary missing values are unconstrained, special missing values .A, .B, and .C conform to the category constraints, and special missing values .G, .H, and .I are weakly ordered. The transformed data are:

<table>
<thead>
<tr>
<th>Y1</th>
<th>Y2</th>
<th>Y3</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.43</td>
<td>3.10</td>
<td>-1.98</td>
<td>8.82</td>
<td>3.74</td>
<td>0.56</td>
<td>0.85</td>
</tr>
<tr>
<td>7.43</td>
<td>3.10</td>
<td>13.26</td>
<td>0.82</td>
<td>3.39</td>
<td>2.86</td>
<td>-6.68</td>
</tr>
<tr>
<td>8.22</td>
<td>5.58</td>
<td>2.00</td>
<td>0.82</td>
<td>3.39</td>
<td>1.76</td>
<td>2.81</td>
</tr>
<tr>
<td>8.22</td>
<td>5.92</td>
<td>3.14</td>
<td>1.84</td>
<td>3.39</td>
<td>1.85</td>
<td>2.98</td>
</tr>
<tr>
<td>7.91</td>
<td>4.68</td>
<td>6.10</td>
<td>1.84</td>
<td>3.39</td>
<td>1.85</td>
<td>2.89</td>
</tr>
<tr>
<td>7.91</td>
<td>4.75</td>
<td>5.10</td>
<td>1.84</td>
<td>3.39</td>
<td>1.76</td>
<td>2.17</td>
</tr>
<tr>
<td>6.28</td>
<td>2.37</td>
<td>5.10</td>
<td>3.34</td>
<td>2.22</td>
<td>0.66</td>
<td>2.38</td>
</tr>
<tr>
<td>8.63</td>
<td>0.86</td>
<td>6.42</td>
<td>3.34</td>
<td>5.00</td>
<td>2.08</td>
<td>2.28</td>
</tr>
<tr>
<td>9.96</td>
<td>10.30</td>
<td>5.85</td>
<td>3.34</td>
<td>3.57</td>
<td>3.49</td>
<td>7.85</td>
</tr>
</tbody>
</table>

7.3. Splines  PROC TRANSREG can be used to detect nonlinear relationships among quantitative variables. For example, 400 observations from the function \( Y = x / 4 \cdot \sin x \) were generated, then both variables were normalized to mean zero and variance one. The following:

```
PROC TRANSREG MAXITER=20;
   MODEL SPLINE(Y) = SPLINE(X);
   POLYNOMIAL X Y / NAKNOTS=4;
```

request cubic polynomial spline transformations of both variables with four knots. PROC TRANSREG increased the squared correlation from the original value of 0.53053 to 0.99988. The plot of \((Y \times X)\) shows the original function and the plot of \((TY \times TX)\) shows the relationship between the transformed variables. The plot was linearized by finding a nonlinear transformation of X (shown below) and an almost linear transformation of Y. This is a particularly easy problem for PROC TRANSREG since the data perfectly fit a curved line, and portions of sine curves can be closely approximated by cubic polynomials.

6. Availability

If you are interested in testing PROC TRANSREG or any of the other Psychometric and Market Research procedures mentioned in this paper, contact Forrest Young at The L. L. Thurstone Psychometric Laboratory, Davie Hall - Q13-A, The University of North Carolina, Chapel Hill, NC, 27514, USA. Two additional version six procedures, PROXSCAL (multidimensional scaling) and COSAN (covariance structure analysis) are in an early stage of development as of this writing. We do not know at this time when they will be available for testing. For more information about any of these procedures contact either Warren Kuhfeld or Forrest Young at the above address. In addition, comments and design suggestions are welcome.
References


1. SAS is a registered trademark of SAS Institute Inc., Cary, NC, USA. This paper describes experimental version 6 SAS software. There is no commitment on the part of UNC or the Institute to support or distribute this software, nor if it is distributed, that it will be in accordance with the descriptions given in this paper.

2. Note that PROC CONJOINT is an alias of PROC TRANSREG. Thus, everything described in this paper about TRANSREG applies to CONJOINT. The authors wish to thank the staff of the Department of Data Theory, Leiden University, The Netherlands, where the first author developed much of PROC TRANSREG. Many of the pieces that comprise our procedures can be traced to work done by the staff of the Department of Data Theory.