A MACRO LIBRARY FOR CLASSICAL AND ROBUST MULTIVARIATE MODELING

E. James Harner and Kathryn S. Fletcher
West Virginia University

Abstract

This paper describes a macro library for modeling multivariate data. Currently, three statistical macros are available: REG, for regression modeling; PCA, for principal components analysis; and DISC, for discriminant modeling. Each of these macros allows a number of options, including robust analyses and validation techniques.

The macro is geared toward interactive modeling. The commands for invoking the macro are easy to use and are brief. The output displayed at the terminal is carefully selected. Additional output depends on the number of observations, e.g., residuals - output to a SAS data set which can be printed by the PRINT utility macro.

Introduction

The many facets of developing a multivariate model include: determining the functional form of the model, selecting and transforming relevant variables, assessing distributional and other (e.g., linearity) assumptions, assessing the effects of multicollinearity and influential observations, and validating the model. Orchestrating these tasks is complex and can only be done effectively and efficiently in an interactive environment.

An interactive system must be designed carefully. For example, it is important to have a fast (command) response time and a concise output. The response time is system dependent and thus the program designer has limited control in terms of the program structure and efficiency. The output, however, can be skillfully tailored. A reasonable approach is to display only the most important quantitative and graphical summaries at the terminal. Lengthy output, e.g., lists depending on the sample size, should be stored in temporary or permanent data sets for further processing. An additional attractive feature is to have a brief command language which is easy to use.

The SAS® System provides some of these features using the Display Manager interface. However, the data analyst does not have much control over the output, which is often voluminous. Furthermore, many of the current modeling methodologies are either not available or are not easily obtained. The SAS System also provides the tools which allow the development of a modeling environment. Specifically, this macro library is intended to provide a robust regression tool for developing a multivariate model. The available options is now given.

The Macro Library

The macro library currently contains the following four macros: 1) REG -- for regression modeling; 2) PCA -- for principal components analysis; 3) DISC -- for discriminant modeling; 4) PRINT -- for printing lists. These macros are called using either a name-style or command-style invocation. The OPTIONS statement for a name-style call would be:

```
MACRO OPTIONS MAUTOSA (FILEDEF=);
```

where FILEDEF points to the current automatic library. For example, under CMS the FILEDEF statement could be:

```
FILEDEF SASAUTOS DISK STAT MA1.1B;
```

SASAUTOS is the default option for SASAUTOS under the OS and CMS operating systems. If the command-style call is used, it is also necessary to specify the option CMSMAC on the OPTIONS statement.

The structure of the first three macros (REG, PCA, and DISC) is similar. Their contents are illustrated with REG.

```
MACRO REG (keyword parameter list of macro variables);
  macrotect
  $END REG;
```

The parameters specify local macro variables which determine the regression options selected by the data analyst or determined by default. The parameter options for each macro are described in the next three sections. The macrotext is organized sequentially according to a natural flow:

1. macro program statements to check the validity of the parameter options selected by the user;
2. PROC MATRIX FW=SFW/IDU (FZU);
3. MACRO statements to read the input SAS statements;
4. PROC MATRIX statements to check (and delete) observations with missing values;
5. macro program (control) statements to select the appropriate subroutines;
6. "subroutines".

PROC MATRIX is not amenable to structured programming. However, the code was modularized wherever possible.

The PRINT macro is a utility for listing long output. The macro contains a DATA step to format the observation numbers and a PROC step (PRINT) for printing. The original variables as well as computed quantities, such as weights and squared Mahalanobis distances, are printed.

The form of the $MACRO statement for PRINT is given by:

```
$MACRO PRINT(IN=LAST);
```

Since _DATA_ is the default output SAS data set for all three of the statistical macros, it is generally only necessary to invoke the macro by:

```
$PRINT
```

The REG Macro

The REG macro is being developed as a general purpose regression modeling tool. Models are being developed using both classical and robust techniques (Hogg, 1979; Huber, 1981; Barnett et al., 1981). In addition, the model can be validated at any stage by jackknifing or bootstrapping (Silver, 1982). Assessing linearity and distributional assumptions (Kleiberg and others, 1980) is not currently supported. The influence of an observation, although not treated formally, can be partially determined from the weights if a robust regression is requested and from the jackknifed residuals if jackknifing is selected. Variable selection is not available, but the macro can be applied repeatedly, easily and quickly until an appropriate model is found.

The $MACRO statement is given by:

```
$MACRO REG(IN=LAST,OUT=DATA,DEP=,INDEP=,INT=YES,EST=LS);
```

This statement specifies the default parameter options, if any. A full discussion of the available options is now given.

Options for REG

```
IN= gives the input SAS data set name. Data management, transformations, and other tasks involving data manipulation need to be done in DATA step(s) prior to the macro call. The default value is LAST.
```

```
OUT= specifies the output SAS data set name. This data set contains: the original independent and dependent variables; the predicted values and residuals; the weights if a robust regression is requested; the jackknifed predicted values and residuals if jackknifing is specified; the bootstrapped predicted values and residuals if bootstrapping is specified. The default value is _DATA_.
```

```
DEP= gives the name of the dependent variable. This option must be specified; there is no default value.
```

```
INDDEP= gives the names of the independent variables. This option must be specified; there is no default value.
```

```
INT= determines whether or not the intercept is included in the model. A value of NO or any other value other than YES suppresses the intercept. The default is YES.
```

```
EST= specifies the estimation technique. The choices are LS --for least squares regression; HUBER --for Huber's robust regression; BIWGT --for Tukey's biweight regression.
```

Use HUBER for moderate downweighting of outlying residuals and BIWGT for more drastic downweighting. The default choice is LS.
Variable & and are related to AGE in that estimation techniques (Huber, 1981; Maronna, 1976) are supported. The supplemental output could be obtained by invoking the PRINT macro. The PCA macro is a program designed to perform principal component analyses on either the correlation or covariance matrix. Both classical and robust estimation techniques (Huber, 1981; Maronna, 1976) are supported. The user also has the option of obtaining the mean vector and covariance/correlation estimates without performing the eigenanalysis (POC). The output is made as concise as possible. For example, since covariance/correlation type matrices are symmetric, they are printed as one matrix with correlations in the upper triangular part and covariances on the diagonal and lower triangular part. The sMACRO statement for PCA is given by:

```
%MACRO PCAUN=LAST...OUT=DATA...VAR=EIGEN...YES CORR=...;
       PRIN1=0.3992 0.2088 -0.4775 -0.6077;
       PRIN2=-0.2932 -0.8793 0.2867 -0.1998;
       PRIN3=-0.2465 0.4262 0.8298 -0.2825;
       PRIN4=0.7030 -0.0393 0.0044 -0.7101;
       OUTCORR=PCAHP
```

### Options for PCA

**IN=** gives the input SAS data set. The default value is LAST.

**OUT=** specifies the output SAS data set. The data set contains the weights if a robust analysis is requested; the squared Mahalanobis distances; the original variables, the principal variables if the eigenanalysis is selected. The default value is DATA.

**VAR=** gives the name of the input variables. This option must be specified since there is no default value.

**EIGEN=** selects whether or not a principal component analysis is to be done. Choose NO or any other value other than YES to suppress this option. The default value is YES.

**CORR=** determines whether or not the principal component analysis (if selected) is done on the correlation or covariance matrix. Choose NO or any other values other than YES to select the covariance option. The default value is YES.

**STEPS=** specifies the estimation technique. The choices are:

- HUBER — for Huber’s robust estimation. Huber’s method downweights observations based on the Mahalanobis distances. The default choice is CL.

- BPS — gives the minimum change in weights for successful convergence. The default value is 0.95.

- EPS — gives the minimum change in weights and residuals, depending on which option is selected. The default value is 0.005.

**REPS=** in the maximum number of iterations used in the robust estimation techniques. The default is 25.

### Terminal Output

The output is illustrated using data from the National Coal Study. The dependent variable is a measure of lung function, the forced expiratory volume in 1 second (FEV). The independent variables are: height in inches (HGT), age (AGE), average number of cigarettes per day (NIOD), number of years smoking (PACKYR), years underground (UYR). PACKYR and UYR are exposure variables and are related to AGE in that all are time related.

The macro is invoked by typing the following command:

```
%MACRO PCA(IN=COAL,DISP=FEV,JOIND=HOT AGE PACKYR UYR,BPS=HUBER,STEPS=5,REPS=25,PRINT=5);
```

The output for this example would be:

```
The macro is invoked with the following command:

```
%MACRO PCA(IN=COAL,DISP=FEV,JOIND=HOT AGE PACKYR UYR,BPS=HUBER,STEPS=5,REPS=25,PRINT=5);
```

The output from this call would be:

```
```

### Options for PCA

**IN=** gives the input SAS data set. The default value is LAST.

**OUT=** specifies the output SAS data set. The data set contains the weights if a robust analysis is requested; the squared Mahalanobis distances; the original variables, the principal variables if the eigenanalysis is selected. The default value is DATA.

**VAR=** gives the name of the input variables. This option must be specified since there is no default value.

**EIGEN=** selects whether or not a principal component analysis is to be done. Choose NO or any other value other than YES to suppress this option. The default value is YES.

**CORR=** determines whether or not the principal component analysis (if selected) is done on the correlation or covariance matrix. Choose NO or any other values other than YES to select the covariance option. The default value is YES.

**STEPS=** specifies the estimation technique. The choices are:

- HUBER — for Huber’s robust estimation. Huber’s method downweights observations based on the Mahalanobis distances. The default choice is CL.

- BPS — gives the minimum change in weights for successful convergence. The default value is 0.95.

- EPS — gives the minimum change in weights and residuals, depending on which option is selected. The default value is 0.005.

**REPS=** in the maximum number of iterations used in the robust estimation techniques. The default is 25.

### Terminal Output

The output is illustrated using data from the National Coal Study. A robust principal components analysis is performed on the independent variables used in the regression analysis. The variables are: AGE, HGT, PACKYR, and UYR.

The macro is invoked by typing the following command:

```
%MACRO PCA(IN=COAL,VAR=AGE HOT PACKYR UYR,BPS=HUBER,STEPS=5,REPS=25,PRINT=5);
```

The output from this call would be:

```
```

### PCA Macro

```
PCAMACRO
```

### ROBUST MEAN VECTOR ESTIMATE

```
MEAN AGE HOT PACKYR UYR
VALUES 42.87 69.24 15.26 21.27
```

### ROBUST COVARIANCE/CORRELATION MATRIX ESTIMATE

```
COVJR AGE HOT PACKYR UYR
AGE 151.43 -0.0878 0.4738 0.9510
HOT -0.0878 7.2042 -0.3003 -0.0405
PACKYR 91.125 -12.7 211.21 0.04916
UYR 157.8 -5.195 105.82 189.69
```

### PCA BASED ON THE CORRELATION MATRIX

```
VARIANCES OF THE PRINCIPAL VARIABLES (EIGENVALUES)
```

```
EIGENVAL PRIN1 PRIN2 PRIN3 PRIN4 VALUES 2.5225 0.9317 0.5465 0.0074
```

### COEFFICIENTS OF THE PRINCIPAL VARIABLES (EIGENVECTORS)

```
EIGENVEC AGE HOT PACKYR UYR
PRIN1 -0.5992 0.2088 -0.4775 -0.6077
PRIN2 -0.2932 -0.8793 0.2867 -0.1998
PRIN3 -0.2465 0.4262 0.8298 -0.2825
PRIN4 0.7030 -0.0393 0.0044 -0.7101
```

### CORRELATIONS BETWEEN THE PRINCIPAL VARIABLES AND THE ORIGINAL VARIABLES

```
CORR AGE HOT PACKYR UYR
PRIN1 -0.9190 0.3203 -0.7234 -0.9321
PRIN2 -0.2979 -0.8931 0.2932 -0.2436
```

813
The DISC Macro

Discriminant analyses can be done by invoking the DISC macro. Models can be developed using both classical and robust estimation techniques (Harner and Billings, 1983; Huber, 1981; Randies et al., 1978). Currently, assessing methodological and distributional assumptions is not supported. However, the output SAS data set does contain values, e.g., Mahalanobis distances, which can be used as diagnostics. The weights can be used to assess the influence of the observations. Variable selection is not available.

The %MACRO statement for DISC is:

%MACRO DISC(IN=,X=,CLASS=,VAR=,EST=,PV=,WGT=,REP=,WIDTH=);
As before, this statement gives the default values, if any. The options are now discussed.

Options for DISC

- IN= gives the input SAS data set. The default value is .LAST..
- OUT= specifies the output SAS data set. The data set contains the weight if robust estimation is used; the squared Mahalanobis distances, the class variable; the original variables; the canonical or discriminant variables. The default output SAS data set name is DATA..
- CLASS= gives the name of the classification variable used for defining groups. This option must be specified, since there is no default.
- VAR= specifies the name of the discriminant variable. This option must be defined by the user.
- EST= determines the estimation method. The choices are:
  - CL --for classical estimation;
  - HUBER --for Huber's robust estimation.
  These are the same choices as in the PCA macro.
- PV= determines the Huber criterion for downweighting as in the PCA macro. The default value is 0.95.
- EPS= gives the minimum change in weights for successful convergence. The default value is 0.005.
- REP= is the maximum number of iterations used in the robust discriminant analysis. The default is 25.
- VALID= determines the validation method, if any. The options are:
  - NO --no validation;
  - JACK --jackknife validation;
  - BOOT --bootstrap validation.
  Depending on the option either jackknife or bootstrap estimates of error are produced. The default value is NO.
- BREP= is the maximum number of bootstrap replications. The default is 25.

Terminal Output

The output from DISC is also illustrated by data from the National Coal Study. The classification variable is coughing status (GCH), which is 0 for non-cookers and 1 for coughers. The potential discriminators are AGE, PACKT, and UTR. A robust discriminant analysis was run. The macro is invoked by the following command:

`DISC(IN=DATA..,CLASS=AGE,VAR=PACKT,EST=HUBER);`

The output from this call would be:

**DISCRIMINANT MACRO
ROBUST DISCRIMINANT ANALYSIS**