In certain applications of linear models, the results can best be presented and interpreted graphically, rather than in terms of parameter estimates and measures of precision. This approach can be especially effective when consulting with clients having limited statistical background or experience. Plots or graphs of both primary and secondary parameters can also be of great value to the analyst in exploratory data analysis, comparison of different models, analysis of residuals, and plotting of different groups or treatments over a range of interest of one of the independent variables.

This paper will present a PROCLIMATE macro for plotting expected mean response for a given set of values of the independent variables. The confidence bounds for the mean response estimates can also be plotted by specifying the appropriate limits. Examples of plotting both primary and secondary parameters will be shown, and instructions for the use of the macro and a listing will be included.

The macro was designed for use with the GLM-D system, but can be used with any linear models program, assuming the parameter estimates, error covariance and $X'X$-inverse matrices are available and can be FETCHed into PROC MATRIX. This system creates a SAS dataset that can be used with PROC PLOT or with digital plotter software. The user required code necessary to generate the plots is minimal and can be almost routinely included in most linear models runs with GLM-D.

**THEORY**

Consider a linear model of the form:

$$y = X\beta + \epsilon$$

with estimated mean and variance $\hat{\beta}$ and $\hat{s}^2$. Define $\hat{\epsilon}$ as a vector of $X$ values corresponding to one observational unit, that is, one row of the $X$ matrix:

$$\hat{\epsilon} = \epsilon_1, \epsilon_2, \ldots, \epsilon_p$$

The expected mean response of the vector $\hat{\epsilon}$ is then given as:

$$\gamma = \hat{\epsilon} \hat{\beta}$$

with variance

$$s^2(\gamma) = s^2 \hat{\beta} (X'X)^{-1} \hat{\beta}$$

The confidence interval for $\gamma$ is then:

$$\hat{\gamma} - Ts(\gamma) < \gamma < \hat{\gamma} + Ts(\gamma)$$

where $T$ is the Student's $t$ statistic with $n-p$ degrees of freedom.

The confidence region for the entire regression surface can be obtained using the Working-Botelling confidence bounds:

$$\hat{\gamma} - Ws(\gamma) < \gamma < \hat{\gamma} + Ws(\gamma)$$

where $W$ is $p$ times the $F$ statistic with $p$ and $n-p$ degrees of freedom. The use of the Working-Botelling bounds allows any number of mean responses to be estimated from all possible $\hat{\epsilon}$ vectors.

**PLOTTING RESPONSE CURVES AND CONFIDENCE BANDS**

In order to plot the expected mean response vs. a particular $X$ variable, a matrix composed of $k$ observational vectors is formed:

$$D_1 \quad D_2 \quad \ldots \quad X_1 \quad \ldots \quad D_p$$

$$D = \begin{bmatrix} \vdots \end{bmatrix}$$

$$D_1 \quad D_2 \quad \ldots \quad X_k \quad \ldots \quad D_p$$

The $D_j$ elements correspond to reference values of the columns of $X$ that remain constant from observation to observation. The $X_j$ elements represent the $k$ values of the $X$ variable of interest. The mean response vector is then given by:

$$\hat{\gamma} = DB$$

The vector of standard errors associated with each mean response is then calculated as:

$$\hat{s} = s \text{ VECDIAG}(D(X'X)^{-1}D')$$

where VECDIAG is a column vector composed of the diagonal elements of the matrix following it. In order to plot the response line and its upper and lower confidence bands, the following matrix of plotting points is constructed:

$$\begin{bmatrix} \hat{\gamma} \\ s \end{bmatrix}$$

$$\text{POINTS} = \frac{1}{2}(\hat{\gamma} - Ws(\gamma))$$

$$\begin{bmatrix} \hat{\gamma} \\ s \end{bmatrix}$$

This matrix can then be OUTPUTed to a SAS dataset and plotted with PROC PLOT.
EXTENSION TO THE MULTIVARIABLE CASE

In the case where there are \( q \) dependent variables and thus \( q \) columns in \( Y \), the model becomes:

\[
Y = XB + E
\]

with parameter estimates \( \hat{\theta} \) and \( \Sigma \). A row vector \( y \) must be defined in order to select the linear combination of the \( Y \) variables to be plotted. In this case the estimated mean and standard error vectors are given by:

\[
\hat{\mu} = \hat{\theta} y
\]

where \( y \) is a \((q \times 1)\) vector and

\[
\hat{\Sigma} = \text{SORT}(\hat{\theta} \Sigma \hat{\theta}^T) \cdot \text{VECDIAG}(Q(\hat{\theta}'X)^{-1}Q')
\]

In most cases \( y \) will be of the form:

\[
y = (0, 0, \ldots, 1, \ldots, 0)
\]

and will simply select the single dependent variable to be plotted on the \( Y \) axis. In some repeated measures and growth curve applications, however, combinations of the dependent variables such as averages and contrasts may be appropriate (2, 3).

APPLICATIONS

The macro GENPLOTS creates the matrix of plotting points and outputs a SAS dataset named PLOT DATA containing the variables LINE, \( X \), and \( Y \). LINE is a single character representing the plotting symbols to be used. This allows different symbols to be used for the response line and for the confidence bounds. It also permits multiple lines to be printed on the same plot. The code for GENPLOTS is shown in Appendix I. The macro assumes that the matrices \( _{\text{BETA}}_\_\Sigma \) and \( _{\text{XPATH}}_\_\Sigma \) exist and are available, as will always be the case with GLIM-5. In order to invoke the macro, the user must perform the following steps:

Step 1 - Indicate the desired linear combination of dependent variables by defining:

\[
y = a \ (q \times 1) \ \text{vector}
\]

Step 2 - Specify the region of interest for the \( X \) axis by defining the scalars:

- \( \text{START} \) = beginning value
- \( \text{END} \) = ending value
- \( \text{INCR} \) = increment between points

where \((\text{END} - \text{START})/\text{INCR}\) is a positive integer.

Step 3 - Specify the columns of the \( D \) matrix that remain constant for a given plotted line by defining a matrix:

\[
D_{11} \ldots 0 \ldots D_{1p} \\
D_{21} \ldots 0 \ldots D_{2p} \\
\vdots \\
D_{q1} \ldots 0 \ldots D_{qp}
\]

where each row represents a different line to be plotted. The \( D_{ij} \) elements represent the elements that remain constant, and the zeros indicate the location of the \( X \) axis variable.

Step 4 - Specify the columns of the \( D \) matrix that represent the variable to be plotted on the \( X \) axis:

\[
D_{\text{VAR}} = 0 \ldots 1 \ldots 0 \\
0 \ldots 1 \ldots 0
\]

\( D_{\text{VAR}} \) must conform to \( D_{\text{CONST}} \) and complement it in terms of zero and non-zero elements.

Step 5 - Set the line plotting symbols for each line defined above by defining:

\[
\text{PLINE} = a \ 1 \ \text{by} \ m \ \text{vector of single characters}
\]

Step 6 - If confidence bands are not needed then set \( \text{CBAND} = 0 \) and go to Step 7. Otherwise set:

\[
\text{CBAND} = 1 \\
\text{UBLIN} = \text{symbols for upper bound} \\
\text{BLBLIN} = \text{symbols for lower bound} \\
\text{STAT} = \text{confidence coefficient}
\]

Step 7 - If multiple plots are being generated or if the dataset is to be saved, then redefine the output dataset name by:

\[
\text{MACRO PLOTDATA} \ \text{dataset name} \%
\]

Step 8 - Invoke the macro to generate the plotting points:

\[
\text{GENPLOTS};
\]

EXAMPLES

Consider an experiment where four blood measurements (cholesterol, triglyceride, diastolic and systolic blood pressure) are made on a number of subjects (2). The sex, age, height and weight of the subjects are also recorded. Assume we fit a model with the four blood measurements as dependent variables:

\[
Y = \text{CHOL TRIG DIAS SIST}
\]

The independent variables are separate intercepts and age slopes for each sex.
and common height and weight slope increments:

\[ Y = M\text{\_INT} F\text{\_INT} M\text{\_AGE} F\text{\_AGE} HT WT \]

After fitting the model with GLM-5 as discussed in (2), the following plots are generated.

In the first example, a plot of male(M) and female(F) triglyceride levels vs. age from 15 to 70 years with height and weight held constant at their sample mean values is desired. The 95% confidence band for each line is also to be printed using the same plotting symbols defined for the lines. Using the previously outlined steps, the following SAS statements are used to generate the plots:

Step 1  \( Y = 0 / 1 / 0 / 0; \)
Step 2  \( \text{START} = 15; \)
\( \text{END} = 70; \)
\( \text{INCD} = 5; \)
Step 3  \( \text{D\_CONST} = 1 0 0 0 66.5 155/0 1 0 0 66.5 155; \)
Step 4  \( \text{D\_VAR} = 0 0 1 0 0 0 0 /0 0 0 1 0 0 0; \)
Step 5  \( \text{PLINE} = 'M' 'F'; \)
Step 6  \( \text{C(balance) = 1}; \)
\( \text{UBLINE} = \text{PLINE}; \)
\( \text{LLINE} = \text{PLINE}; \)
\( \text{STAT} = 3.68; \)
Step 7  \( \text{MACRO PLOTPDATA PLOT1 %} \)
Step 8  \( \text{GENPLOTS;} \)

The resulting PROC PLOT output is shown in Figure 1.

Two other examples are also displayed and the actual code used to generate the plots is listed in Appendix II. In Figure 2, the dependent variable is changed to diastolic blood pressure and the confidence bands are not plotted. Note that the user need change only those matrices that differ from the values given in the first example. Figure 3 shows a more complicated plot. The average triglyceride of males and females is plotted vs. weight at ages 20, 40 and 60 years, using plotting symbols 2, 4 and 6 respectively.

**SUMMARY**

The power and flexibility of the PROC MATRIX operators, combined with the accessibility of the internal matrices and ease of use of the GLM-5 linear models system, provide a readily usable method of graphically displaying both primary and secondary model parameters. The resulting plots can be used for data analysis as well as for presentation of results. Since the data points for plotting are in SAS datasets, they can be saved for later use and further processing by other SAS procedures.

The examples presented in this paper represent only a small subset of the applications of graphics to linear models procedures in particular, and to statistical analysis in general. For example, the plotting of different models on the same graph may lend insights that are not apparent from examination of the matrices of their parameter estimates. This approach can be especially helpful in stepwise and polynomial applications(3). By replacing \( \hat{y} \) with the original \( y \) used to fit the model, the user can generate the expected responses, calculate the residuals, and plot them using the same general method presented here. Thus the user of GLM-5 with moderate PROC MATRIX experience has a useful and flexible analysis tool at his or her disposal.

**ACKNOWLEDGEMENT**

This research was supported in part by the Bureau of Health Manpower under a Project Grant for Graduate Training in Public Health, Grant No. 1 D04 AR01460 entitled "Public Health Data Management."

**REFERENCES**


252
Figure 1.
Figure 2.
Figure 3.
APPENDIX I - Macro Listing

MACRO GENPLOTS * - A MACRO TO GENERATE PLOTS FROM GLM 5. ;
* WRITTEN BY DAVID CHRISTIANSEN, UNC CHAPEL HILL JAN 79 ;
* CALCULATE VECTOR OF POINTS FOR THE X_AXIS;
X_POINTS = *((START*/INCP);(END*/INCP))*INCP;
K = NROW(X_POINTS);
* CALCULATE D MATRIX;
A = NROW(D_VAR);
D = D_CONST @ J(K,1,1) + D_VAR @ X_POINTS;
PRINT D COLNAME = _XNAME_;
PRINT V ROWNAME = _YNAME_;
* CALCULATE MEAN RESPONSE VECTOR;
y = D* _BETA_*';
* EXPAND PLOTTING SYMBOLS TO CONFORM TO POINTS;
LINE = PLINE @ J(1,K,1);
* CREATE MATRIX OF PLOT POINTS;
POINTER = III *(J(A,1,1) @ X_POINTS);
* IF CONFIDENCE BANDS ARE DESIRED, ADD THEM TO POINTS;
IF CBAND THEN DC;
* CALCULATE STD. DEV VECTOR S;
S = (ROWSUM(D*XPXINV_)*D)*((V*SIGMA_*V))**.5;
* CALCULATE UPPER AND LOWER BOUNDARY POINTS;
UPPER = (Y + 5*STAT) @ POINTS(*,2);
LOWER = (Y - 5*STAT) @ POINTS(*,2);
* ADD BOUNDARY POINTS;
POINTS = Y | (J(A,1,1) @ X_POINTS);
* IF CONFIDENCE BANDS ARE DESIRED, ADD THEM TO POINTS;
IF CBAND THEN DC;
* OUTPUT POINTS TO SAS DS, USING ROWNAME TO PASS PLOTTING SYMBOL;
PNAME = 'Y' 'X';
OUTPUT POINTS DATA=PLOTDATA(RENAME=(RCW=LINE))
ROWNAME=LINE
COLNAME=PNAME;
% * END MACRO GENPLOTS;

APPENDIX II - Examples

* EXAMPLE 2 ;
* CHANGE DEPENDENT VAR TO DIASTOLIC BP AND SUPPRESS CONFIDENCE BANDS:
V = 0 / 0 / 0 / 1;
CBAND = 0;
MACRO PLOTDATA PLOT2 % GENPLOTS

* EXAMPLE 3 ;
* AVERAGE TRIGLYCERIDE OF MALES AND FEMALES VS WEIGHT ;
* AT AGES 20, 40, 60, ;
V = 0 / 1 / 0 / 0 ;
D_CONST = .5 .5 10 10 66.5 0 /
-5 .5 20 20 66.5 0 /
D_VAR = 0 0 0 0 1 /
0 0 0 0 1 /
0 0 0 0 1 ;
START = 100;
END = 280;
INCR = 10;
PLINE = '2' '4' '6';
CBAND = 0;
MACRO PLOTDATA PLOT3 % GENPLOTS