FIT

PURPOSE

Estimate the parameters for a linear, polynomial, or nonlinear least squares fit.

DESCRIPTION

This is one of DATAPLOT's most powerful and heavily used commands. Among the more common nonlinear models that can be analyzed with DATAPLOT are exponential models, models involving powers, square root models, exponential over polynomial models, Lorentzian models, Gaussian models, Bessel function models, Chebychev models, and rational function models.

After a fit, DATAPLOT prints the following analytic information.

1. The parameter estimates, the parameter standard deviations, and the parameter t-values are printed to the terminal. The t-value is used to determine if a given parameter is significant in the FIT. The parameters and their standard deviations are written to the file DPST1F.DAT (the name may vary slightly on some implementations). To store these values in the arrays COEF and COEFSD, enter:

   READ DPST1F.DAT COEF COEFSD

2. The residual standard deviation and its corresponding degrees of freedom are written to the terminal. These are stored in the parameters RESSD and RESDF respectively. RESDF is the number of observations minus the number of independent variables in the fit (including the constant term). The formula for RESSD is:

$$RESSD = \sqrt{\frac{(y-\hat{y})^2}{RESDF}} \qquad \textbf{(EQ 3-42)}$$

3. If there is replication in the independent variables, the replication standard deviation and corresponding degrees of freedom are printed to the terminal. In addition, a lack of fit F test is performed. These are stored in the parameters REPSD, REPDF, and LOFCDF respectively. The formulas for REPDF and REPSD are (s is the number of replication sets):

$$REPDF = \sum_{j=1}^{s} (N_s - 1) \qquad \textbf{(EQ 3-43)}$$

$$REPSD = \sqrt{\frac{\sum (y_{ij} - \bar{y}_j)^2}{REPDF}} \qquad \textbf{(EQ 3-44)}$$

4. The predicted values are saved in the variable PRED and the residual values are saved in the variable RES. These variables can be used in subsequent LET and PLOT commands to generate diagnostic plots of residuals and predicted values.
5. For linear fits, the standard deviation of the predicted values and the lower and upper limits for 95% and 99% confidence intervals for the predicted values are written to the file DPST2F.DAT (the name may vary on some operating systems). The standard deviation can be used to generate additional intervals and tests (the PROGRAM 3 example demonstrates this).
6. For linear fits, the following diagnostic information is printed to the file DPST3F.DAT (the name may vary on some implementations):

   Variable 1 - the diagonals of the hat matrix (the hat matrix is $X(X^TX)^{-1}X^T$);
   Variable 2 - the variance of the residuals;
   Variable 3 - the standardized residuals;
   Variable 4 - the internally studentized residuals;
   Variable 5 - the deleted residuals;
   Variable 6 - the externally studentized residuals;
   Variable 7 - Cook's distance;
   Variable 8 - the DFFITS statistic.

   These statistics are discussed in more detail in the NOTE - REGRESSION DIAGNOSTICS section below.

7. For linear fits, the variance-covariance matrix of the parameters and the inverse of the (X'X) matrix are written to the file DPST4F.DAT (the name may vary on some implementations). These values can be used in deriving additional statistics, intervals, and tests. The PROGRAM 3 example shows how to read in these matrices and use them to generate additional intervals.

Linear multiple regression is one of the most used statistical techniques and there is a large number of intervals and diagnostic tools used in regression analysis. In addition, there is a wide discrepancy among analysts regarding which tools and diagnostics should be used. The various NOTE sections below summarize most of the more commonly used ones (most modern statistics textbooks that cover regression discuss many of them in detail, specifically see chapters 7, 8, and 11 of the Neter, Wasserman, and Kuntner book listed in the REFERENCE section). The PROGRAM 3 example below demonstrates how to generate many of these procedures in DATAPLOT. This program example is intended only to show the mechanics of generating commonly used statistics, plots, and intervals with DATAPLOT and is not intended to be a case study. Consult one of the books listed in the REFERENCE section for guidance on when these various statistics and intervals are appropriate and on how to interpret them.

Since DATAPLOT is intended to be used interactively, it writes only a limited amount of diagnostic information to the screen when performing a fit. Additional information is saved in internal variables or written to files, which the analyst can either use or ignore. Since most analysts typically have a certain subset of these techniques that they routinely use, it is recommended that you write a few general purpose macros to handle these routine analyses (use PROGRAM 3 as a guide).

SYNTAX 1

FIT <y1> = <f> <SUBSET/EXCEPT/FOR qualification>

where <y1> is the response (= dependent) variable;

<f> is:

1. any general Fortran-like expression; or

2. any function name that the user has already created via the LET FUNCTION command;

and where the <SUBSET/EXCEPT/FOR qualification> is optional.

This first syntax is appropriate for all models--linear, polynomial, multi-linear (up to 15 independent variables), and nonlinear (up to 15 independent variables). It uses an iterative modified Levenberg-Marquardt algorithm. Linear fits are handled as a special case (the fits are still done iteratively). Only the DPST1F.DAT file is created.

SYNTAX 2

<d> FIT <y> <x1> ... <xn> <SUBSET/EXCEPT/FOR qualification>

where <d> is the optional specification of the desired degree:

LINEAR	or	FIRST-DEGREE (the default)
QUADRATIC	or	SECOND-DEGREE
CUBIC	or	THIRD-DEGREE
QUARTIC	or	FOURTH-DEGREE
QUINTIC	or	FIFTH-DEGREE
SEXTIC	or	SIXTH-DEGREE
SEPTIC	or	SEVENTH-DEGREE
OCTIC	or	EIGHT-DEGREE
NONIC	or	NINTH-DEGREE
DEXIC	or	TENTH-DEGREE;

<y> is the response (= dependent) variable;

<x1> ... <xn> is a list of 1 to 35 independent variables;

and where the <SUBSET/EXCEPT/FOR qualification> is optional.

The estimated parameters are stored in A0, A1, ... , AN.

This second syntax is appropriate only for linear and polynomial models. In practice, the linear and quadratic fits receive heavy use while the other degrees are rarely used. Up to 35 independent variables can be specified in the FIT. It uses a modified Gramm-Schmidt algorithm (based on the QR decomposition) with iterative refinement. Since this generates an exact fit rather than an iterative fit and calculates much more diagnostic information, this syntax is recommended for linear fits. The code was adapted from the OMNITAB statistical package.

EXAMPLES

```
FIT Y = A+B*EXP(-C*X)
FIT Y = (A+B*X+C*X**D)/(SIN(EXP(-ALPHA*X2+BETA*X3)))
FIT Y = F1
LINEAR FIT Y X
Y X1 X2 X5
Y X1 X2 X5 SUBSET TAG > 1
QUADRATIC FIT PRESSURE TEMP
```

```
CUBIC FIT V R
```

NOTE 1

The paper “Techniques for Fitting and Verification of Linear/Non-Linear Models using DATAPLOT” contains a large number of fitting examples. It is particularly useful for guidance in fitting nonlinear and rational function models in DATAPLOT. This paper is distributed as part of the standard DATAPLOT documentation.

NOTE 2

Starting values are not required. However, they can sometimes speed up nonlinear fits or prevent nonlinear fits from getting stuck in a local minimum or maximum if they are specified. To specify them, simply assign values to the coefficients before doing the fit. For example:

```
LET ALPHA = 0.15
LET A = 0.004
LET B = 0.01
FIT Y = EXP(-ALPHA*X)/(A+B*X)
```

The PRE-FIT command can be used to determine better starting values.

NOTE 3

The nonlinear algorithm is iterative with two commands for controlling the iterations. By default, a maximum of 50 iterations are allowed before DATAPLOT assumes the fit is not converging. You can change this maximum with the FIT ITERATIONS command. DATAPLOT tests the value of the residual standard deviation and the ratio of successive values of the residual standard deviation to check for convergence. The cutoff value for the residual standard deviation can be specified with the FIT STANDARD DEVIATION command. See the documentation for FIT ITERATIONS and FIT STANDARD DEVIATION for details.

NOTE 4

In addition to standard least squares fits, DATAPLOT also supports:

1. Weighted fits via the WEIGHT command.
2. Robust fits via the BIWEIGHT and TRICUBE LET subcommands. The documentation for the WEIGHTS command gives a macro for iteratively reweighted least squares.
3. Locally weighted least squares via the LOWESS SMOOTH command.
4. Smoothing via the SMOOTH command.
5. Cubic spline interpolation via the INTERPOLATION LET subcommand and spline fitting via the SPLINE FIT command.
6. Exact rational function fitting via the EXACT RATIONAL FIT command.
7. Principle components regression via the PRINCIPLE COMPONENTS LET subcommand.

NOTE 5

The following plots are often desired after a fit:

1. The predicted values with the dependent variable values. This gives an overall impression of how good the fit is.
2. The dependent variable against each of the independent variables. This gives an indication of the individual relationships and can identify outliers and gaps in the data.
3. Each of the independent variables against the other independent variables (e.g., PLOT X1 VS X2). In addition, interaction terms are sometimes desired (e.g., LET X12 = X1*X2, PLOT X1 VS X12). These plots can be used to identify highly correlated independent variables and to identify joint outliers and gaps in the data.
4. Various types of residual plots. Typically, a sequence plot and a normal probability plot are generated. These are used to verify the regression assumptions (that the errors are independent, normally distributed, and have constant variance). In addition, the residuals can be plotted against the predicted values to test the constant variance assumption. Some analysts also like to plot the residuals against each of the independent variables to check for remaining structure.

In any event, all of these plots are straightforward to generate since RES and PRED can be plotted like any other variable. The MULTIPLOT and LOOP commands can be used to generate some of them systematically while avoiding too many pages to look at. These plots are useful for both linear and nonlinear fits. The command 6-PLOT generates 6 commonly used plots after a fit on one page.

NOTE 6

For linear regression, some analysts like an ANOVA table. The needed values are easy to derive. This is demonstrated in the PROGRAM 3 example.

NOTE 7

Data transformations can be generated easily if needed via the LET command. Some types of nonlinear fits can be restated as linear fits with an appropriate transformation. The BOX-COX LINEARITY PLOT can be a useful command for determining an appropriate transformation.

Some analysts prefer to “standardize” the independent variables and the dependent variable by subtracting the mean and dividing by the standard deviation. This is done to provide numerical stability (DATAPLOT scales the data internally before doing the regression calculations) and also so that the data and regression coefficients are on a common scale. The original regression and standardized model are related as follows:

$$x' = \frac{x_i - \bar{x}}{s_x} \qquad \textbf{(EQ 3-45)}$$

$$y' = \frac{y_i - \bar{y}}{s_y} \qquad \textbf{(EQ 3-46)}$$

$$\beta_k = \left(\frac{s_y}{s_k}\right)\beta'_k \qquad \textbf{(EQ 3-47)}$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}_1 - \ldots - \beta_p\bar{x}_p \qquad \textbf{(EQ 3-48)}$$

A variation on this is the correlation transformation (also called the standardized regression model). It does:

$$y'_i = \frac{1}{\sqrt{N-1}} \times \left(\frac{y_i - \bar{y}}{s_y}\right) \qquad \textbf{(EQ 3-49)}$$

$$x'_{ik} = \frac{1}{\sqrt{N-1}} \times \left(\frac{x_{ik} - \bar{x}_k}{s_k}\right) \qquad \textbf{(EQ 3-50)}$$

With this transformation, the X^TX matrix reduces to correlation matrix of the independent variables. Either one of the above transformations is easy to generate in DATAPLOT. For example, if there are P independent variables:

```
LET FACT = 1/SQRT(N - 1)
LOOP FOR K = 1 1 P
    LET XMEAN = MEAN X^K
    LET XSD = STANDARD DEVIATION X^K
    LET Z^K = FACT*((X^K - XMEAN)/XSD)
END OF LOOP
LET YMEAN = MEAN Y
LET YSD = STANDARD DEVIATION Y
LET YT = FACT*((Y - YMEAN)/YSD)
FIT YT Z1 TO Z^P
```

NOTE - INTERVALS AND TESTS

For linear regression, the theory for generating confidence and prediction intervals for the data points and confidence intervals for the parameters are well developed. As noted above, DATAPLOT writes the coefficients and the coefficient standard deviations to the file DPST1F.DAT. It also writes the standard deviations of the predicted values and the 95% and 99% confidence intervals for the predicted values to the file DPST2F.DAT. The following statistics and intervals can be computed from these values.

1. A function to provide a point estimate for new data points.
2. If you want different significance levels for the confidence interval for predicted values, do something like the following (assume NP is the number of observations minus the number of variables in the fit and the predicted standard deviation was read into PREDSD):

```
LET ALPHA = <value>
LET ALPHA2 = ALPHA/2
LET T = TPPF(ALPHA2,NP)
LET UPPER = PRED + T*PREDSD
```

LET LOWER = PRED - T*PREDSD

3. Bonferroni joint confidence limits for the parameters (the t-values listed in the standard output can be used to test single parameters for significance).
4. Bonferroni or Hotelling joint confidence limits for the response function. This is for observations that were used to generate the regression.

Additional statistics can be generated from the inverse of the X^TX matrix and the variance-covariance matrix of the parameters (these vary only by a scalar multiplication). As noted above, DATAPLOT writes these to the file DPST4F.DAT. The following can be derived from these matrices:

1. The variance for a new point.
2. A confidence interval for the new point.
3. A prediction interval for the new point.
4. A Bonferroni or Hotelling joint confidence interval for more than one new point.
5. A Bonferroni or Scheffe joint prediction interval for more than one new point.

The numerical details for reading these matrices and generating the above types of intervals are demonstrated in the PROGRAM 3 example. Be aware that earlier versions of DATAPLOT may not write out these values to the files DPST1F.DAT and DPST2F.DAT or may write different statistics to these files.

NOTE - REGRESSION DIAGNOSTICS

Regression diagnostics are used to identify outliers in the dependent variable, identify influential points, to identify high leverage points, or in general to uncover features of the data that could cause difficulties for the fitted regression. High leverage points are those that are outliers with respect to the independent variables. Influential points are those that cause large changes in the fitted function when they are deleted. Although an influential point typically has high leverage, a high leverage point is not necessarily influential. The books by Belsley, Kuh, and Welsch and by Cook and Weisberg listed in the REFERENCE discuss regression diagnostics in detail. Chapter 11 of the Neter, Wasserman, and Kuntner book listed in the REFERENCE section discusses them in a less theoretical way. Consult one of these books or some other statistics text that covers regression diagnostics for guidance on when these various diagnostic statistics are appropriate and on how to interpret them.

At a minimum, diagnostic analysis typically includes various plots of the residuals and predicted values as outlined in the NOTE section above. For more complete diagnostics, the variables written to the file DPSTS3F.DAT can be analyzed (see the DESCRIPTION section above). This file contains the diagonals of the hat matrix, 4 alternate forms of the residuals, Cook's distance, the DFFITS values, and the variance of the residuals.

The standardized residuals are the residuals divided by the square root of the mean square error. The internally studentized residuals are the residuals divided by their standard deviations. Many analysts prefer to use one of these forms in the standard residual analysis. The deleted residuals are the residuals obtained from deleting one case at a time from the regression (i.e., the ith deleted residual is the difference between the original Y data value and the predicted value obtained when the ith case is deleted from the fit). The externally studentized residuals (also called the studentized deleted residuals) are the deleted residuals divided by their standard deviations. Deleted residuals and externally deleted residuals are used to identify outlying Y observations that the normal residuals do not identify (cases with high leverage tend to generate small residuals even if they are outliers).

Many recently developed regression diagnostics depend on the Hat matrix. This matrix is $X(X^TX)^{-1}X^T$. The limit of 100 columns for matrices limits the Hat matrix to cases with 100 observations or less. Fortunately, most of the relevant statistics can be derived from the diagonal elements of this matrix (which can be read from the DPST3F.DAT file). These are also referred to as the leverage points. The minimum leverage is (1/N), the maximum leverage is 1.0, and the average leverage is (P/N) where P is the number of variables in the fit. As a rule of thumb, leverage points greater than twice the average leverage can be considered high leverage. High leverage points are outliers in terms of the X matrix and have an unduly large influence on the predicted values. High leverage points also tend to have small residuals, so they are not typically detected by standard residual analysis.

The DFFITS values are a measure of influence that observation *i* has on the *i*th predicted value. As a rule of thumb, for small or moderate size data sets, values greater than 1 indicate an influential case. For large data sets, values greater than 2*SQRT(P/N) indicate influential cases.

Cook's distance is a measure of the combined impact of the ith observation on all of the regression coefficients. It is typically compared to the F distribution. Values near or above the 50th percentile imply that observation has substantial influence on the regression parameters.

The DFFITS values and the Cook's distance are used to identify high leverage points that are also influential (in the sense that they have a large effect on the fitted model).

These variables can be read in as follows (you can modify the SET READ FORMAT line to skip over variables you don't want):

```
FIT ....
SKIP 1
SET READ FORMAT 8(E15.7,1X)
READ DPST3F.DAT PREDSD HII VARRES STDRES ISTUDRES DELRES ESTUDRES COOK DFFITS
SKIP 0
SET READ FORMAT
```

Once these variables have been read in, they can be printed, plotted, and used in further calculations like any other variable. This is demonstrated in the PROGRAM 3 example. They can also be used to derive additional diagnostic statistics. The PROGRAM 3 example shows how to compute the Mahalanobis distance and Cook's V statistic. The Mahalanobis distance is a measure of the distance of an observation from the "center" of the observations and is essentially an alternate measure of leverage. The Cook's V statistic is the ratio of the variances of the predicted values and the variances of the residuals.

Another popular diagnostic is the DFBETA statistic. This is similar to Cook's distance in that it tries to identify points that have a large influence on the estimated regression coefficients. The distinction is that DFBETA assesses the influence on each individual parameter rather than the parameters as a whole. The DFBETA statistics require the catcher matrix, which is described in the NOTE - MULTI-COLLINEARITY section below, for easy computation. The usual recommendation for DFBETA is that absolute values larger than 1 for small and medium size data sets and larger than (2/SQRT(N)) for large data sets should be considered influential.

The variables written to file DPST3F.DAT are calculated without computing any additional regressions. The statistics based on a case being deleted are computed from mathematically equivalent formulas that do not require additional regressions to be performed. The Neter, Wasserman, and Kunter text gives the formulas that are actually used.

Robust regression is often recommended when there are significant outliers in the data. One common robust technique is called iteratively reweighted least squares (or M-estimation). Note that these techniques protect against outlying Y values, but they do not protect against outliers in the X matrix. Also, they test for single outliers and are not as sensitive for a group of similar outliers. See the documentation for WEIGHTS, BI-WEIGHT, and TRI-CUBE for more information on performing iteratively reweighted least squares regression in DATAPLOT. Techniques for protecting against outliers in the X matrix use alternatives to least squares. Two such methods are least median squares regression (also called LSQ regression) and least trimmed squares regression (also called LTS regression). DATAPLOT does not support these techniques at this time. The documentation for the WEIGHTS command in the Support chapter discusses one approach for dealing with outliers in the X matrix in the context of iteratively reweighted least squares.

NOTE - MULTI-COLLINEARITY

Multi-collinearity results when the columns of the X matrix have significant interdependence (that is, one column is close to a linear combination of some collection of other columns). Multi-collinearity typically results in numerically unstable estimates in the sense that small changes in the X matrix can result in significant changes in the estimated regression coefficients. It can also cause other significant problems. Pairwise collinearity can be determined from correlation coefficients between independent variables (or from pairwise plots). However, this does not detect higher order multi-collinearity. One measure of this is the multiple correlation coefficient between the *j*th variable and the rest of the independent variables. The Variance Inflation Factor (VIF) is a scaled version of this with the following formula:

$$VIF_j = \frac{1}{(1 - R_j^2)} \qquad \textbf{(EQ 3-51)}$$

The VIF values are often given as their reciprocals (this is called the tolerance) . Fortunately, these values can be computed without performing any additional regressions. The computing formulas are based on the catcher matrix, which is $X(X^TX)^{-1}$. The equation is:

$$VIF_j = \sum_{i=1}^{N} c_{ij}^2 \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2 \qquad \textbf{(EQ 3-52)}$$

where *c* is the catcher matrix. This is a straightforward calculation in DATAPLOT if N is less than the maximum number of rows that DATAPLOT matrix commands can handle (typically 500 if the maximum number of rows for a variable is 10,000 and 1,000 if the maximum number of rows for a variable is 20,000). If N is greater than this maximum, the catcher matrix cannot be computed easily. The PROGRAM 3 example shows how to compute the catcher matrix (for less than 500 rows), the R_j^2, and the VIF's.

Another measure of multi-collinearity are the condition indices. The condition indices are calculated as follows:

1. Scale the columns of the X matrix to have unit sums of squares.
2. Calculate the singular values of the scaled X matrix and square them.

Condition indices between 30 and 100 indicate moderate to strong collinearity. The PROGRAM 3 example shows how to compute the condition indices.

If significant multi-collinearity is present, a model with fewer (or at least different) independent variables is usually called for. However, several alternative techniques are available which use all the independent variables. Principal components regression performs a principal component analysis on the X matrix to get a reduced set of variables that contain most of the information in the original X matrix. See the documentation for the PRINCIPAL COMPONENTS command for an example of this with DATAPLOT. Ridge regression can generate estimates with smaller variance than ordinary least squares (at the expense of some bias in the estimators) and that are stable even with significant multi-collinearity. Although DATAPLOT does not support ridge regression directly, the sample macro RIDGE.DP in the DATAPLOT reference directory shows an example of ridge regression with DATAPLOT. Principal components regression is discussed in the Draper and Smith text (see REFERENCE) and ridge regression is discussed in both the Draper and Smith and the Neter, Wasserman, and Kunter texts (see REFERENCE).

NOTE 8

Two related plots for determining if a variable should be deleted (or added) to a regression are the partial residual plot and the partial regression plot. Plotting an independent variable against the dependent variable does not take into account the effect of the other independent variables in the fit. Both of these plots are essentially attempts to show the dependence of the dependent variable on the variable given the effect of the other dependent variables already in the model.

In the partial residual plot, the values b_kX_{ik} are added to the ordinary residuals and then plotted against the X_{ik} (b_k is the kth regression coefficient and X_{ik} is the ith row of the kth independent variable)

The partial regression plot (also referred to as added variable plots) is a plot of the residuals when X_j is removed from the regression (referred to as $Y_{.[j]}$) against the residuals of a fit of X_j against the remaining independent variables (referred to as $X_{j.[j]}$). Once the catcher matrix (see the NOTE - MULTI-COLLENARITY above) is computed, it is straightforward to compute partial regression plots. This plot reduces to $b_jX_{j.[j]}$+e against $X_{j.[j]}$ where b_j is the jth parameter estimate. The $X_{j.[j]}$ values are calculated from the catcher matrix by:

$$(X_{j.[\mathrm{j}]})_i = \frac{c_{ij}}{\sum_{i=1}^{N} c_{ij}^2} \qquad \textbf{(EQ 3-53)}$$

The PROGRAM 3 example demonstrates how to generate both partial residual and partial regression plots. In addition, it generates partial leverage plots. Partial leverage plots show how the leverage changes as a variable is added (or deleted).

Although partial residual plots are easier to generate (they do not require the computation of the catcher matrix), the partial regression plot is generally considered to have better statistical properties.

NOTE 9

DATAPLOT's primary limitation for linear regression is that it does not do any type of automatic subset selection. Statistics such as Mallow's C_p are not conveniently generated by DATAPLOT.

DEFAULT

None

SYNONYMS

None

RELATED COMMANDS

FIT ITERATIONS	=	Sets the maximum number of iterations for the fit command.
FIT STANDARD DEVIATION	=	Sets the minimum standard deviation for the fit command.
WEIGHTS	=	Sets the weights for the fit command.
PRED	=	A variable where predicted values are stored.
RES	=	A variable where residuals are stored.
RESSD	=	A parameter where the residual standard deviation is stored.
RESDF	=	A parameter where the residual degrees of freedom is stored.
REPSD	=	A parameter where the replication standard deviation is stored.
REPDF	=	A parameter where the replication degrees of freedom is stored.
LOFCDF	=	A parameter where the lack of fit cdf is stored.
EXACT RATIONAL FIT	=	Carries out an exact rational fit.

PRE-FIT	=	Carries out a least squares pre-fit.
SPLINE FIT	=	Carries out a spline fit.
SMOOTH	=	Carries out a smoothing.
ANOVA	=	Carries out an ANOVA.
MEDIAN POLISH	=	Carries out a median polish.
LOWESS SMOOTH	=	Generate a locally weighted least squares smoothing.
PLOT	=	Generates a data or function plot.

REFERENCES

“Applied Linear Statistical Models,” 3rd ed., Neter, Wasserman, and Kunter, Irwin, 1990.

“Applied Regression Analysis,” 2nd ed., Draper and Smith, John Wiley, 1981.

“Data Analysis and Regression,” Mosteller and Tukey, Addison-Wesley, 1977.

“Residuals and Influence in Regression,” Cook and Weisberg, Chapman and Hall, 1982.

“Regression Diagnostics,” Belsley, Kuh, and Welsch, John Wiley, 1980.

“Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis,” Atkinson, Oxford University Press, 1985.

“Efficient Computing of Regression Diagnostics,” Velleman and Welsch, The American Statistician, November, 1981.

“Techniques for Fitting and Verification of Linear/Non-Linear Models using DATAPLOT,” J. J. Filliben, unpublished manuscript,

APPLICATIONS

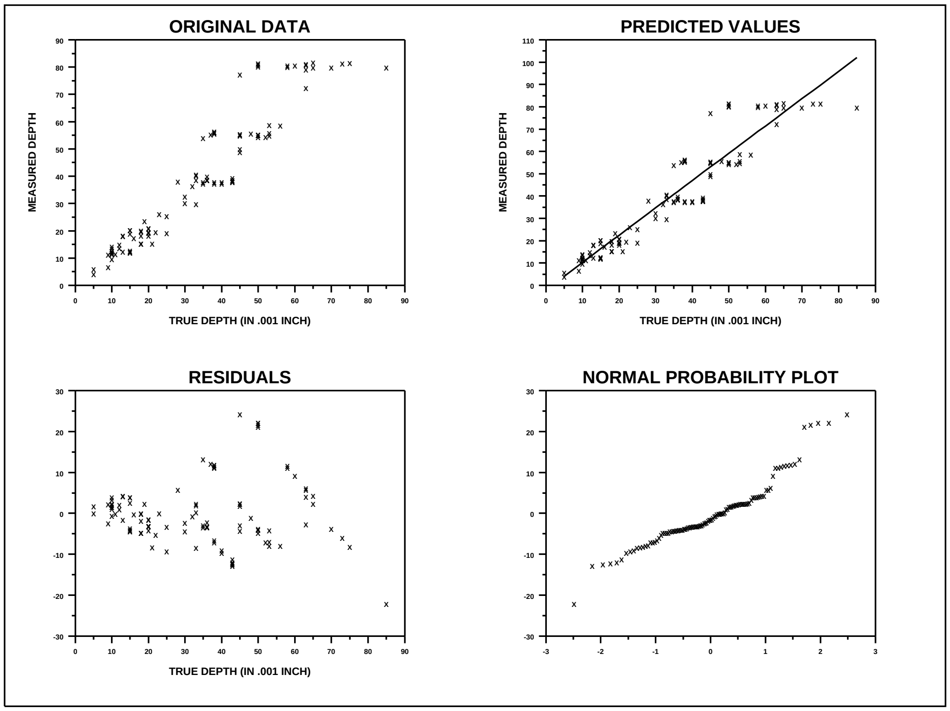
Non-linear and linear fitting

IMPLEMENTATION DATE

Pre-1987

PROGRAM 1 (LINEAR EXAMPLE)

```
. ALASKA PIPELINE RADIOGRAPHIC DEFECT BIAS CURVE
. PERFORM A LINEAR REGRESSION
SKIP 25
READ BERGER1.DAT MEAS TRUE
FIT MEAS TRUE
.
MULTIPLOT 2 2
MULTIPLOT CORNER COORDINATES 0 0 100 100
TITLE ORIGINAL DATA
X1LABEL TRUE DEPTH (IN .001 INCH)
Y1LABEL MEASURED DEPTH
CHARACTERS X
LINES BLANK
PLOT MEAS TRUE
TITLE PREDICTED VALUES
PLOT MEAS PRED VS TRUE
TITLE RESIDUALS
Y1LABEL
PLOT RES VS TRUE
X1LABEL
TITLE NORMAL PROBABILITY PLOT
NORMAL PROBABILITY PLOT RES
END OF MULTIPLOT
```

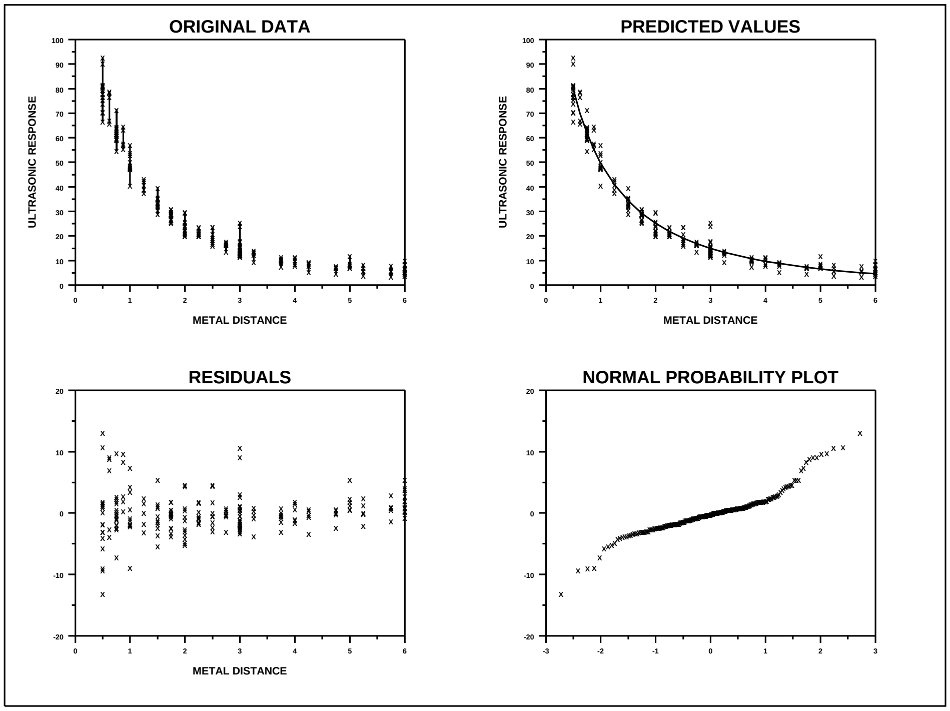


This command generates the following output.

```
 LEAST SQUARES MULTILINEAR FIT
       SAMPLE SIZE N       =      107
       NUMBER OF VARIABLES =        1
       REPLICATION CASE
       REPLICATION STANDARD DEVIATION =     0.6112687111D+01
       REPLICATION DEGREES OF FREEDOM =          29
       NUMBER OF DISTINCT SUBSETS     =          78


             PARAMETER ESTIMATES           (APPROX. ST. DEV.)    T VALUE
       1  A0                   4.99368       ( 1.126    )          4.4
       2  A1       TRUE       0.731111       (0.2455E-01)          30.

       RESIDUAL    STANDARD DEVIATION =         6.0809240341
       RESIDUAL    DEGREES OF FREEDOM =         105
       REPLICATION STANDARD DEVIATION =         6.1126871109
       REPLICATION DEGREES OF FREEDOM =          29
       LACK OF FIT F RATIO =       0.9857 = THE  46.3056% POINT OF THE
       F DISTRIBUTION WITH     76 AND     29 DEGREES OF FREEDOM
       COEF AND SD(COEF) WRITTEN TO FILE DPST1F.DAT
```

PROGRAM 2 (NON-LINEAR EXAMPLE)

```
. DAN CHWIRUT ULTRASONIC REFERENCE BLOCK ANALYSIS
. PERFORM A NON-LINEAR REGRESSION
SKIP 25; READ CHWIRUT1.DAT Y X
LET ALPHA = 0.15
LET A = 0.004
LET B = 0.01
FIT Y = EXP(-ALPHA*X)/(A+B*X)
.
MULTIPLOT 2 2 ; MULTIPLOT CORNER COORDINATES 0 0 100 100
TITLE ORIGINAL DATA
CHARACTERS X ALL
X1LABEL METAL DISTANCE
Y1LABEL ULTRASONIC RESPONSE
PLOT Y X X
TITLE PREDICTED VALUES
LINE BLANK SOLID
CHARACTER X BLANK
PLOT Y PRED VS X
TITLE RESIDUALS
Y1LABEL
PLOT RES VS X
X1LABEL
TITLE NORMAL PROBABILITY PLOT
NORMAL PROBABILITY PLOT RES
END OF MULTIPLOT
```



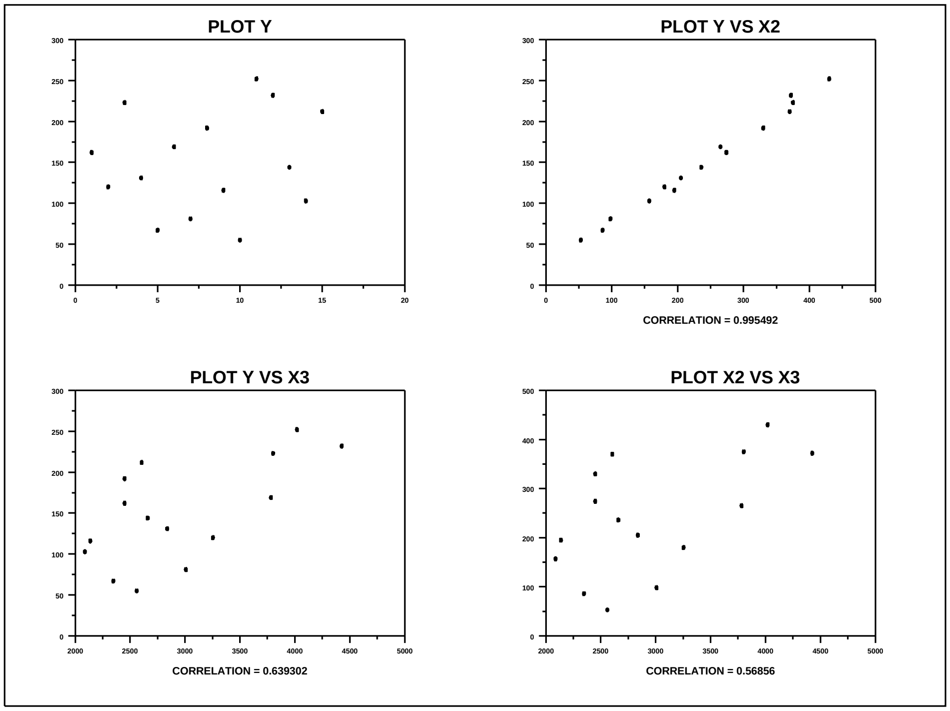
This program generates the following output.

```
THE COMPUTED VALUE OF THE CONSTANT ALPHA    =   0.1500000E+00


THE COMPUTED VALUE OF THE CONSTANT A        =   0.4000000E-02


THE COMPUTED VALUE OF THE CONSTANT B        =   0.1000000E-01


LEAST SQUARES NON-LINEAR FIT
      SAMPLE SIZE N =      214
      MODEL--Y =EXP(-ALPHA*X)/(A+B*X)
      REPLICATION CASE
      REPLICATION STANDARD DEVIATION =     0.3281762600D+01
      REPLICATION DEGREES OF FREEDOM =         192
      NUMBER OF DISTINCT SUBSETS     =          22

ITERATION  CONVERGENCE  RESIDUAL  *  PARAMETER
 NUMBER      MEASURE    STANDARD  *  ESTIMATES
                        DEVIATION *
----------------------------------*-----------
    1--  0.10000E-01  0.10779E+02 * 0.15000E+00 0.40000E-02 0.10000E-01
    2--  0.50000E-02  0.37219E+01 * 0.18074E+00 0.55543E-02 0.10717E-01
    3--  0.25000E-02  0.33620E+01 * 0.19055E+00 0.61191E-02 0.10520E-01
    4--  0.12500E-02  0.33617E+01 * 0.19045E+00 0.61338E-02 0.10525E-01

      FINAL PARAMETER ESTIMATES           (APPROX. ST. DEV.)    T VALUE
      1  ALPHA               0.190408       (0.2207E-01)          8.6
      2  A                   0.613298E-02   (0.3493E-03)          18.
      3  B                   0.105266E-01   (0.8027E-03)          13.

      RESIDUAL    STANDARD DEVIATION =         3.3616721630
      RESIDUAL    DEGREES OF FREEDOM =         211
      REPLICATION STANDARD DEVIATION =         3.2817625999
      REPLICATION DEGREES OF FREEDOM =         192
      LACK OF FIT F RATIO =       1.5474 = THE  92.6461% POINT OF THE
      F DISTRIBUTION WITH     19 AND    192 DEGREES OF FREEDOM
      COEF AND SD(COEF) WRITTEN TO FILE DPST1F.DAT
```

PROGRAM 3 (MULTI-LINEAR REGRESSION)

```
. ZARTHAN COMPAY EXAMPLE FROM
. “APPLIED LINEAR STATISTICAL MODELS” BY NETER, WASSERMAN, KUTNER
. Y = SALES
. X1 = CONSTANT TERM
. X2 = TARGET POPULATION (IN THOUSANDS)
. X3 = PER CAPITA DISCRETIONARY INCOME (DOLLARS)
.
DIMENSION 200 COLUMNS
LET NVAR = 2
READ DISTRICT Y POP INCOME
 1  162  274 2450
 2  120  180 3254
 3  223  375 3802
 4  131  205 2838
 5   67   86 2347
 6  169  265 3782
 7   81   98 3008
 8  192  330 2450
 9  116  195 2137
10   55   53 2560
11  252  430 4020
12  232  372 4427
13  144  236 2660
14  103  157 2088
15  212  370 2605
END OF DATA
.
LET N = SIZE Y
LET P = NVAR + 1
LET X1 = 1 FOR I = 1 1 N
LET X2 = POP
LET X3 = INCOME
```

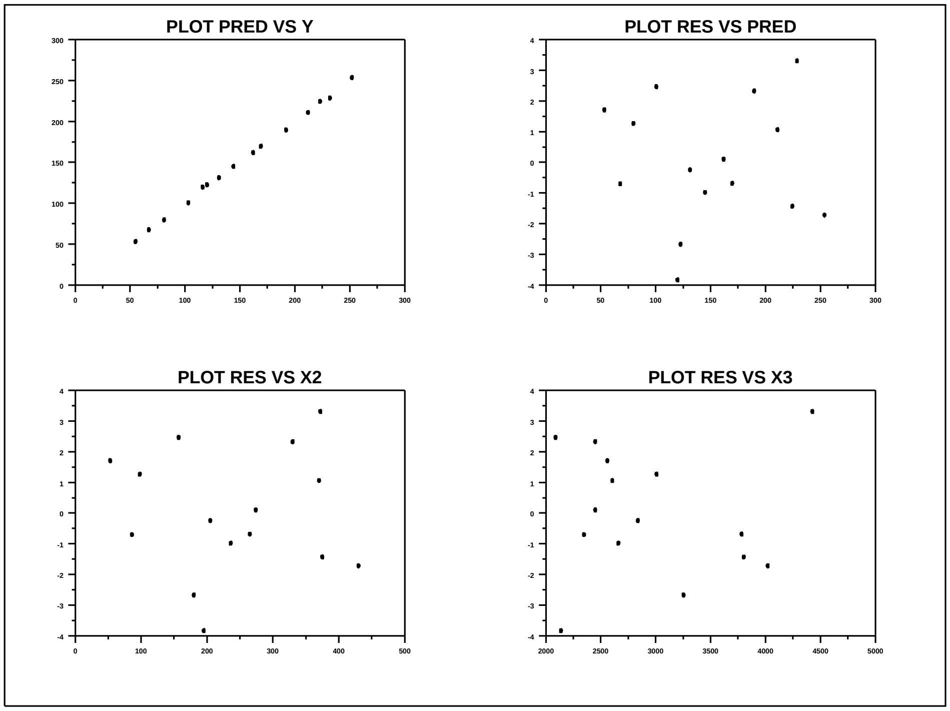
```
. DO PRELIMANARY PLOTS
.  1) INDEPENDENT AGAINST DEPENDENT
.  2) INDEPENDENT AGAINST INDEPENDENT
TITLE AUTOMATIC
LINE BLANK
CHARACTER FONT SIMPLEX ALL
CHARACTERS CIRCLE BLANK
CHARACTER SIZE 1 ALL
CHARACTER FILL ON ALL
MULTIPLOT CORNER COORDINATES 0 0 100 100; MULTIPLOT 2 2
PLOT Y
LOOP FOR K = 2 1 P
    LET RIJ = CORRELATION Y X^K
    X1LABEL CORRELATION = ^RIJ
    PLOT Y VS X^K
END OF LOOP
LOOP FOR K = 2 1 P
    LET IK1 = K + 1
    LOOP FOR J = IK1 1 P
        LET RIJ = CORRELATION X^K X^J
        X1LABEL CORRELATION = ^RIJ
        PLOT X^K VS X^J
    END OF LOOP
END OF LOOP
END OF MULTIPLOT
X1LABEL
```



```
. DO THE LINEAR FIT WITH ALL VARIABLES
.  1) THE RESIDUAL AND PREDICTED PLOTS
.  2) THE PARTIAL RESIDUAL PLOTS
FIT Y X2 TO X^P
READ DPST1F.DAT COEF COEFSD
MULTIPLOT 2 2
PLOT PRED VS Y
PLOT RES VS PRED
LOOP FOR K = 2 1 P
    PLOT RES VS X^K
END OF LOOP
END OF MULTIPLOT
MULTIPLOT 2 2
PLOT RES
NORMAL PROBABILITY PLOT RES
LOOP FOR J = 2 1 P
    LET AJUNK = COEF(J)
    LET PARTRES = RES + AJUNK*X^J
    LEGEND 1 COEF = ^AJUNK; X1LABEL X^J; TITLE PARTIAL RESIDUAL PLOT
    PLOT PARTRES X^J
END OF LOOP
LEGEND 1
TITLE
X1LABEL
DELETE PARTRES
END OF MULTIPLOT
```

The FIT command generates the following output.

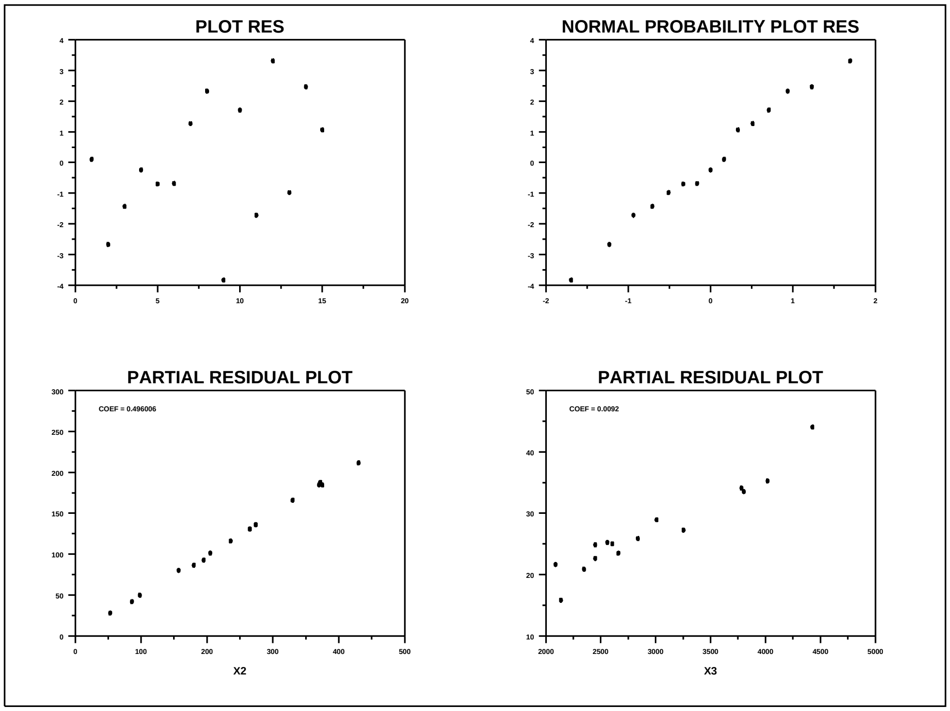
```
LEAST SQUARES MULTILINEAR FIT
    SAMPLE SIZE N       =       15
    NUMBER OF VARIABLES =        2
    NO REPLICATION CASE


            PARAMETER ESTIMATES           (APPROX. ST. DEV.)    T VALUE
     1  A0                   3.45261       ( 2.431    )          1.4
     2  A1       X2         0.496005       (0.6054E-02)          82.
     3  A2       X3         0.919908E-02   (0.9681E-03)          9.5

    RESIDUAL    STANDARD DEVIATION =         2.1772222519
    RESIDUAL    DEGREES OF FREEDOM =          12
    COEF AND SD(COEF) WRITTEN OUT TO FILE DPST1F.DAT
    SD(PRED),95LOWER,95UPPER,99LOWER,99UPPER
                      WRITTEN OUT TO FILE DPST2F.DAT
    REGRESSION DIAGNOSTICS WRITTEN OUT TO FILE DPST3F.DAT
    PARAMETER VARIANCE-COVARIANCE MATRIX AND
    INVERSE OF X-TRANSPOSE X MATRIX
    WRITTEN OUT TO FILE DPST4F.DAT
```


PLOT PRED VS Y

PLOT RES VS PRED

PLOT RES VS X2

PLOT RES VS X3


PLOT RES

NORMAL PROBABILITY PLOT RES

PARTIAL RESIDUAL PLOT

PARTIAL RESIDUAL PLOT

```
. CALCULATE:
. 1) ANOVA TERMS (SSE, SSR, SSTO, MSR, MSE, R2, R2ADJ)
.
LET NP = N - P
LET YBAR = MEAN Y
LET MSE = RESSD**2
LET SSE = MSE*RESDF
LET TEMP = (PRED - YBAR)**2
LET SSR = SUM TEMP
LET MSR = SSR/(P-1)
LET TEMP = (Y - YBAR)**2
LET SSTO = SUM TEMP
LET R2 = 1 - SSE/SSTO
LET R2ADJ = 1- ((N-1)/(N-P))*(SSE/SSTO)
PRINT “SOURCE: REGRESSION”
PRINT “ SSR = ^SSR”
LET DF = P - 1
PRINT “ DEGREES OF FREEDOM = ^DF”
PRINT “ MSR = ^MSR”
PRINT “SOURCE: ERROR”
PRINT “ SSE = ^SSE”
PRINT “ DEGREES OF FREEDOM = ^NP”
PRINT “ MEAN SQUARE ERROR = ^MSE”
PRINT “SOURCE: TOTAL”
PRINT “ SSTO = ^SSTO”
LET FSTAT = MSR/MSE
PRINT “ F STATSISTIC = ^FSTAT”
LET FCRIT = FPPF(.95,DF,NP)
PRINT “ F CRITICAL VALUE (ALPHA = 0.05) = ^FCRIT”
PRINT “ “
PRINT “ “
PRINT “R2 = ^R2”
PRINT “ADJUSTED R2 = ^R2ADJ”
```

The following output is generated.

```
SOURCE: REGRESSION
  SSR  = 53844.72
  DEGREES OF FREEDOM = 2
  MSR  = 26922.36
SOURCE: ERROR
  SSE  = 56.88357
  DEGREES OF FREEDOM = 12
  MEAN SQUARE ERROR = 4.740297
SOURCE: TOTAL
  SSTO = 53901.6
  F STATSISTIC = 5679.467
  F CRITICAL VALUE (ALPHA = 0.05) = 3.885294


R2 = 0.998945
ADJUSTED R2 = 0.998769
```

```
.  READ
.    1) PREDSD = PREDICTED VALUE STANDARD DEVIATIONS
.    2) 95% AND 99% CONFIDENCE LIMITS FOR PREDICTED VALUES
.  CALCULATE
.    1) JOINT CONFIDENCE INTERVAL FOR THE GIVEN DATA POINTS
.       BONFERRONI  = JOINTBL, JOINTBU
.       HOTELLING   = JOINTWL, JOINTWU
.
READ DPST2F.DAT PREDSD  LOWER  UPPER LOWER99 UPPER99
LET ALPHA = .10
LET ALPHA2 = 1 - (ALPHA/(2*N))
LET B = TPPF(ALPHA2,NP)
LET JOINTBU = PRED + B*PREDSD
LET JOINTBL = PRED - B*PREDSD
LET W2 = P*FPPF(.90,P,NP)
LET W = SQRT(W2)
LET JOINTWU = PRED + W*PREDSD
LET JOINTWL = PRED - W*PREDSD
PRINT “PREDICTED VALUE, 95% LOWER AND UPPER CONFIDENCE LIMITS, ..”
PRINT “    90% JOINT BONFERRONI LOWER AND UPPER CONFIDENCE LIMITS, ..”
PRINT “    90% JOINT HOTELLING LOWER AND UPPER CONFIDENCE LIMITS, ..”
PRINT “ PREDICTED LOWER 95% UPPER 95% LOWER BON UPPER BON LOWER HOT UPPER HOT”
SET WRITE FORMAT 7F10.3
PRINT PRED LOWER UPPER JOINTBL JOINTBU JOINTWL JOINTWU
SET WRITE FORMAT
```

The following output is generated.

```
PREDICTED VALUE, 95% LOWER AND UPPER CONFIDENCE LIMITS, ..
    90% JOINT BONFERRONI LOWER AND UPPER CONFIDENCE LIMITS, ..
    90% JOINT HOTELLING LOWER AND UPPER CONFIDENCE LIMITS, ..
 PREDICTED LOWER 95% UPPER 95% LOWER BON UPPER BON LOWER HOT UPPER HOT
   161.896   160.060   163.731   159.138   164.653   159.540   164.251
   122.667   120.903   124.432   120.017   125.318   120.403   124.932
   224.429   222.383   226.476   221.355   227.504   221.803   227.056
   131.241   129.952   132.529   129.306   133.176   129.588   132.894
    67.699    65.608    69.790    64.558    70.840    65.016    70.383
   169.685   167.689   171.681   166.687   172.683   167.124   172.246
    79.732    77.427    82.037    76.269    83.195    76.774    82.690
   189.672   187.337   192.007   186.165   193.179   186.676   192.668
   119.832   117.912   121.752   116.947   122.717   117.368   122.296
    53.291    50.837    55.744    49.606    56.975    50.143    56.438
   253.715   251.196   256.234   249.931   257.499   250.482   256.948
   228.691   225.869   231.513   224.451   232.930   225.069   232.312
   144.979   143.617   146.342   142.933   147.026   143.231   146.728
   100.533    98.583   102.484    97.603   103.463    98.030   103.036
   210.938   208.413   213.464   207.144   214.732   207.697   214.179
```

```
.  CALCULATE
.    1) BONFERONI JOINT CONFIDENCE LIMITS FOR PARAMETERS (BONU, BONL)
.    2) FUNCTION TO CALCULATE REGRESSION ESTIMATE FOR NEW DATA (F)
LET A0 = COEF(1)
LET FUNCTION F = A0
LET DUMMY = PREDSD(1)
.  USE (2*NVAR) RATHER THAN (2*P) IF NO CONSTANT TERM IN JOINT INTERVAL
LET ALPHA2 = 1 - (ALPHA/(2*P))
LET B = TPPF(ALPHA2,NP)
LET BONU(1) = A0 + B*DUMMY
LET BONL(1) = A0 - B*DUMMY
LOOP FOR K = 1 1 NVAR
   LET INDX = K + 1
   LET A^K = COEF(INDX)
   LET FUNCTION F = F + (A^K)*(Z^K)
   LET DUMMY = COEFSD(INDX)
   LET BONU(INDX) = A^K + B*DUMMY
   LET BONL(INDX) = A^K - B*DUMMY
END OF LOOP
.
PRINT “90% BONFERRONI JOINT CONFIDENCE INTERVALS FOR PARAMETERS”
SET WRITE FORMAT 3F10.3
PRINT BONL COEF BONU
SET WRITE FORMAT
.
LET Z1 = DATA 220 375
LET Z2 = DATA 2500 3500
LET YNEW = F
.
PRINT “ “
PRINT “ “
PRINT “NEW X VALUES, ESTIMATED NEW VALUE”
PRINT Z1 Z2 YNEW
```

The following output is generated.

```
90% BONFERRONI JOINT CONFIDENCE INTERVALS FOR PARAMETERS
     1.428     3.453     5.477
     0.481     0.496     0.511
     0.007     0.009     0.012


NEW X VALUES, ESTIMATED NEW VALUE

VARIABLES--Z1             Z2             YNEW

  0.2200000E+03  0.2500000E+04  0.1355714E+03
  0.3750000E+03  0.3500000E+04  0.2216513E+03
```

```
.  READ:
.    1) INVERSE OF (X'X)
.    2) PARAMETER VARIANCE-COVARIANCE MATRIX
.
.  THEY ARE READ IN AS COLUMNS AND THEN COVERTED TO MATRICES.
.  CREATE MATRICES IN SEPARATE LOOPS SINCE DATAPLOT EXPECTS COLUMNS FOR
.  MATRIX DEFINITION TO BE CONTIGUOUS.
READ DPST4F.DAT TEMP1 TEMP2
LET TAG = SEQUENCE 1 P 1 P
LOOP FOR K = 1 1 P
   LET S2B^K = TEMP1
   RETAIN S2B^K SUBSET TAG = K
END OF LOOP
LET S2B = MATRIX DEFINITION S2B1 P P
LOOP FOR K = 1 1 P
   LET XTXINV^K = TEMP2
   RETAIN XTXINV^K SUBSET TAG = K
END OF LOOP
LET XTXINV = MATRIX DEFINITION XTXINV1 P P
PRINT " "; PRINT " "; PRINT "THE X'X INVERSE MATRIX"; PRINT XTXINV
PRINT " "; PRINT " "
PRINT "THE PARAMETER VARIANCE-COVARIANCE MATRIX"; PRINT S2B
DELETE TEMP1 TEMP2
```

The following output is generated.

```
THE X'X INVERSE MATRIX

VARIABLES--XTXINV1        XTXINV2        XTXINV3

  0.1246348E+01  0.2129666E-03 -0.4156712E-03
  0.2129666E-03  0.7732902E-05 -0.7030254E-06
 -0.4156712E-03 -0.7030254E-06  0.1977185E-06


THE PARAMETER VARIANCE-COVARIANCE MATRIX

VARIABLES--S2B1           S2B2           S2B3

  0.5908062E+01  0.1009525E-02 -0.1970405E-02
  0.1009525E-02  0.3665626E-04 -0.3332548E-05
 -0.1970405E-02 -0.3332548E-05  0.9372444E-06
```

```
.  CALCULATE:
.     1) THE VARIANCE OF A NEW POINT (S2YHAT)
.     2) A CONFIDENCE INTERVAL FOR A NEW POINT
.     3) A JOINT CONFIDENCE INTERVAL FOR MORE THAN ONE POINT
.     4) A PREDICTION INTERVAL FOR A NEW POINT
.     5) A SCHEFFE JOINT PREDICTION INTERVAL FOR MORE THAN ONE POINT
.
LET NPT = SIZE YNEW
LOOP FOR DUMMY = 1 1 NPT
   LET XNEW(1) = 1
   LET XNEW(2) = Z1(DUMMY)
   LET XNEW(3) = Z2(DUMMY)
   LOOP FOR K = 1 1 P
      LET DUMMY2 = VECTOR DOT PRODUCT XNEW S2B^K
      LET SUM(K) = DUMMY2
   END OF LOOP
   LET S2YHAT = VECTOR DOT PRODUCT SUM XNEW
   LET S2YPRED = MSE + S2YHAT
   LET YHATS2(DUMMY) = S2YHAT
   LET YPREDS2(DUMMY) = S2YPRED
   LET SYHAT = SQRT(S2YHAT)
   LET YHATS(DUMMY) = SYHAT
   LET SYPRED = SQRT(S2YPRED)
   LET YPREDS(DUMMY) = SYPRED
   LET YHAT = YNEW(DUMMY)
   PRINT “ “; PRINT “ “
   PRINT “THE PREDICTED VALUE FOR THE NEW POINT = ^YHAT”
   PRINT “THE VARIANCE OF THE NEW VALUE = ^S2YHAT”
   PRINT “THE VARIANCE FOR PREDICTION INTERVALS = ^S2YPRED”
   LET T = TPPF(.975,NP)
   LET YHATU = YHAT + T*SYHAT
   LET YHATL = YHAT - T*SYHAT
   LET YPREDU = YHAT + T*SYPRED
   LET YPREDL = YHAT - T*SYPRED
   PRINT “ “
   PRINT “95% CONFIDENCE INTERVAL FOR YHAT: ^YHATL <= YHAT <= ^YHATU”
   PRINT “95% PREDICTION INTERVAL FOR YHAT: ^YPREDL <= YHAT <= ^YPREDU”
END OF LOOP
```

The following output is generated.

```
THE PREDICTED VALUE FOR THE NEW POINT = 135.5715
THE VARIANCE OF THE NEW VALUE = 0.466366
THE VARIANCE FOR PREDICTION INTERVALS = 5.206663

95% CONFIDENCE INTERVAL FOR YHAT: 134.0835 <= YHAT <= 137.0594
95% PREDICTION INTERVAL FOR YHAT: 130.5998 <= YHAT <= 140.543


THE PREDICTED VALUE FOR THE NEW POINT = 221.6513
THE VARIANCE OF THE NEW VALUE = 0.760463
THE VARIANCE FOR PREDICTION INTERVALS = 5.50076

95% CONFIDENCE INTERVAL FOR YHAT: 219.7513 <= YHAT <= 223.5514
95% PREDICTION INTERVAL FOR YHAT: 216.5412 <= YHAT <= 226.7614
```

```
LET ALPHA = 0.10
LET DUMMY = 1 - ALPHA/(2*NPT)
LET B = TPPF(DUMMY,NP)
LET JOINTBU = YNEW + B*YHATS
LET JOINTBL = YNEW - B*YHATS
PRINT “ “
PRINT “90% BONFERRONI JOINT CONFIDENCE INTERVALS FOR NEW VALUES”
PRINT JOINTBL YNEW JOINTBU
LET W = P*FPPF(.90,P,NP); LET W = SQRT(W)
LET JOINTWU = YNEW + W*YHATS
LET JOINTWL = YNEW - W*YHATS
PRINT “ “
PRINT “90% HOTELLING JOINT CONFIDENCE INTERVALS FOR NEW VALUES”
PRINT JOINTWL YNEW JOINTWU
LET JOINTBU = YNEW + B*YPREDS
LET JOINTBL = YNEW - B*YPREDS
PRINT “ “
PRINT “90% BONFERRONI JOINT PREDICTION INTERVALS FOR NEW VALUES”
PRINT JOINTBL YNEW JOINTBU
LET S = NPT*FPPF(.90,NPT,NP); LET S = SQRT(S)
LET JOINTSU = YNEW + S*YPREDS
LET JOINTSL = YNEW - S*YPREDS
PRINT “ “
PRINT “90% SCHEFFE JOINT PREDICTION INTERVALS FOR NEW VALUES”
PRINT JOINTSL YNEW JOINTSU
```

The following output is generated.

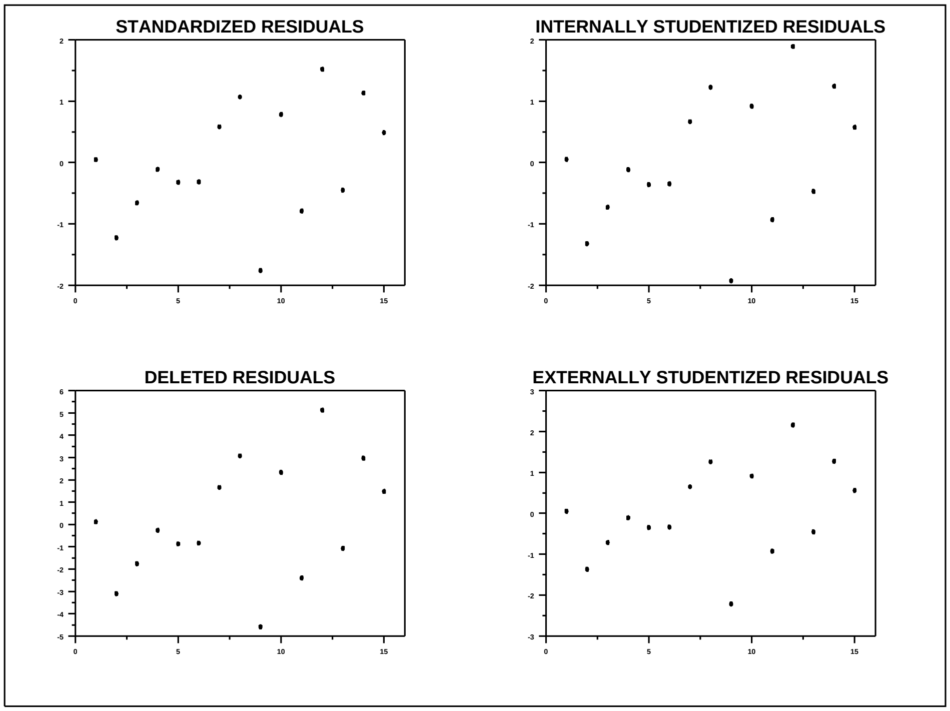
```
90% BONFERRONI JOINT CONFIDENCE INTERVALS FOR NEW VALUES

VARIABLES--JOINTBL        YNEW           JOINTBU

  0.1340835E+03  0.1355714E+03  0.1370594E+03
  0.2197513E+03  0.2216513E+03  0.2235513E+03

90% HOTELLING JOINT CONFIDENCE INTERVALS FOR NEW VALUES

VARIABLES--JOINTWL        YNEW           JOINTWU

  0.1336621E+03  0.1355714E+03  0.1374807E+03
  0.2192132E+03  0.2216513E+03  0.2240894E+03

90% BONFERRONI JOINT PREDICTION INTERVALS FOR NEW VALUES

VARIABLES--JOINTBL        YNEW           JOINTBU

  0.1305998E+03  0.1355714E+03  0.1405431E+03
  0.2165412E+03  0.2216513E+03  0.2267614E+03

90% SCHEFFE JOINT PREDICTION INTERVALS FOR NEW VALUES

VARIABLES--JOINTSL        YNEW           JOINTSU

  0.1301651E+03  0.1355714E+03  0.1409777E+03
  0.2160944E+03  0.2216513E+03  0.2272082E+03
```

```
.  READ IN VARIOUS DIAGNOSTIC STATISTICS:
.      1) DIAGONAL OF HAT MATRIX (HII)
.      2) VARIANCE OF RESIDUALS (RESVAR)
.      3) STANDARIZED RESIDUALS (STDRES)
.      4) INTERNALLY STUDENTIZED RESIDUALS (STUDRES)
.      5) DELETED RESIDUALS (DELRES)
.      6) EXTERNALLY STUDENTIZED RESIDUALS (ESTUDRES)
.      7) COOK' DISTANCE (COOK)
.      8) DFFITS STATISTIC (DFFITS)
.      9) DERIVE PRESS STATISTIC, COOK'S V STATISTIC, AND MAHALANOBIS DISTANCE
.
SKIP 1
READ DPST3F.DAT HII RESVAR STDRES STUDRES DELRES ESTUDRES COOK DFFITS
SKIP 0
LET V = PREDSD**2/RESVAR
LET TEMP = DELRES*DELRES
LET PRESSP = SUM TEMP
LET MAHAL = ((HII-1/N)/(1-HII))*(N*(N-2)/(N-1))
.
LET HBAR = P/N
LET DUMMY = SUM HII
SET WRITE FORMAT 5F10.5
PRINT " "
PRINT "       HII      COOK    DFFITS         V     MAHAL"
PRINT HII COOK DFFITS V MAHALSET WRITE FORMAT
```

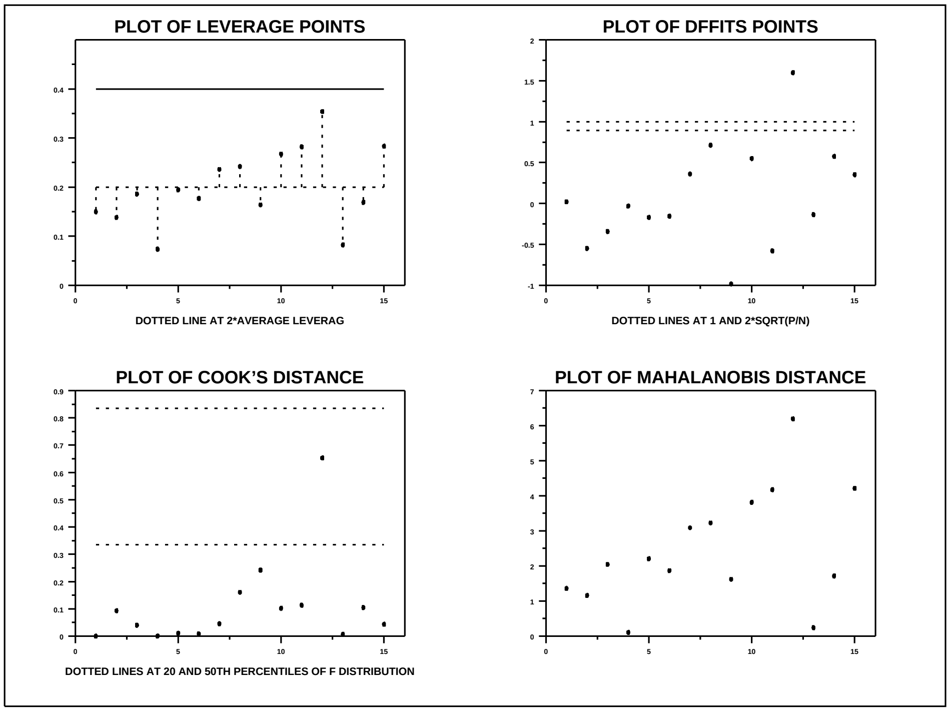
The following output is generated.

```
       HII      COOK    DFFITS         V     MAHAL
   0.14974   0.00016   0.02087   0.17612   1.36094
   0.13837   0.09324  -0.54768   0.16059   1.15907
   0.18613   0.04037  -0.34081   0.22870   2.04457
   0.07374   0.00035  -0.03104   0.07961   0.10632
   0.19432   0.01029  -0.16916   0.24119   2.20696
   0.17701   0.00862  -0.15474   0.21508   1.86753
   0.23617   0.04577   0.36153   0.30919   3.09086
   0.24224   0.16078   0.71115   0.31969   3.22734
   0.16388   0.24206  -0.98132   0.19600   1.61948
   0.26740   0.10238   0.55026   0.36501   3.81652
   0.28203   0.11316  -0.57910   0.39281   4.17798
   0.35396   0.65306   1.59945   0.54789   6.19399
   0.08250   0.00661  -0.13608   0.08992   0.24043
   0.16906   0.10478   0.57510   0.20346   1.71639
   0.28343   0.04377   0.35185   0.39554   4.21346
```

```
. PLOT VARIOUS RESIDUALS
X1LABEL
XLIMITS 0 15
MAJOR XTIC MARK NUMBER 4
XTIC OFFSET 0 1
MULTIPLOT 2 2
TITLE STANDARDIZED RESIDUALS
PLOT STDRES
TITLE INTERNALLY STUDENTIZED RESIDUALS
PLOT STUDRES
TITLE DELETED RESIDUALS
X1LABEL PRESS STATISTIC = ^PRESSP
PLOT DELRES
X1LABEL
TITLE EXTERNALLY STUDENTIZED RESIDUALS
PLOT ESTUDRES
END OF MULTIPLOT
```



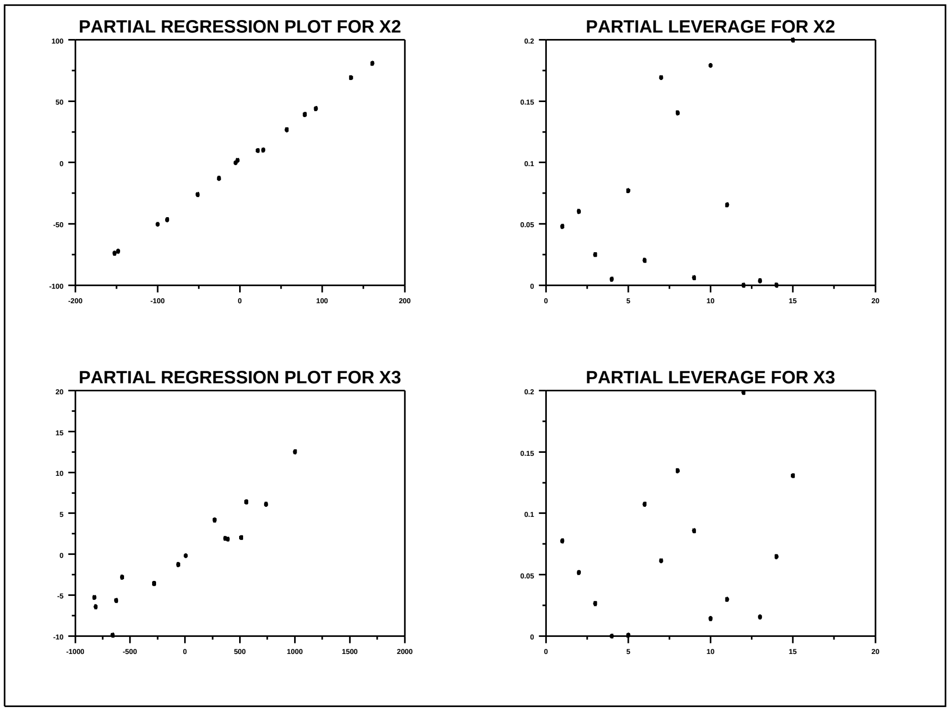
```
.  PLOT SEVERAL DIAGNOSTIC STATISTICS
MULTIPLOT 2 2
CHARACTER FILL ON OFF; CHARACTER CIRCLE BLANK; LINE BLANK SOLID DOTTED
TITLE PLOT OF LEVERAGE POINTS; Y1LABEL; X1LABEL DOTTED LINE AT 2*AVERAGE LEVERAGE
YTIC OFFSET 0 0.1
LET TEMP6 = DATA 1 N; LET DUMMY = 2*HBAR
LET TEMP4 = DATA DUMMY DUMMY; LET TEMP5 = DATA HBAR HBAR
SPIKE ON; SPIKE BASE HBAR; SPIKE LINE DOTTED
PLOT HII AND
PLOT TEMP4 TEMP5 VS TEMP6
SPIKE OFF; YTIC OFFSET 0 0
.
CHARACTER CIRCLE BLANK BLANK; LINE BLANK DOTTED DOTTED; Y1LABEL
TITLE PLOT OF DFFITS POINTS; X1LABEL DOTTED LINES AT 1 AND 2*SQRT(P/N)
LET TEMP4 = DATA 1 1; LET DUMMY = 2*SQRT(P/N); LET TEMP5 = DATA DUMMY DUMMY
PLOT DFFITS AND
PLOT TEMP4 TEMP5 VS TEMP6
.
TITLE PLOT OF COOK’S DISTANCE
X1LABEL DOTTED LINES AT 20 AND 50TH PERCENTILES OF F DISTRIBUTION
LET DUMMY = FPPF(.20,P,NP); LET TEMP4 = DATA DUMMY DUMMY
LET DUMMY = FPPF(.50,P,NP); LET TEMP5 = DATA DUMMY DUMMY
PLOT COOK AND
PLOT TEMP4 TEMP5 VS TEMP6
.
TITLE PLOT OF MAHALANOBIS DISTANCE; X1LABEL
PLOT MAHAL
END OF MULTIPLOT
```



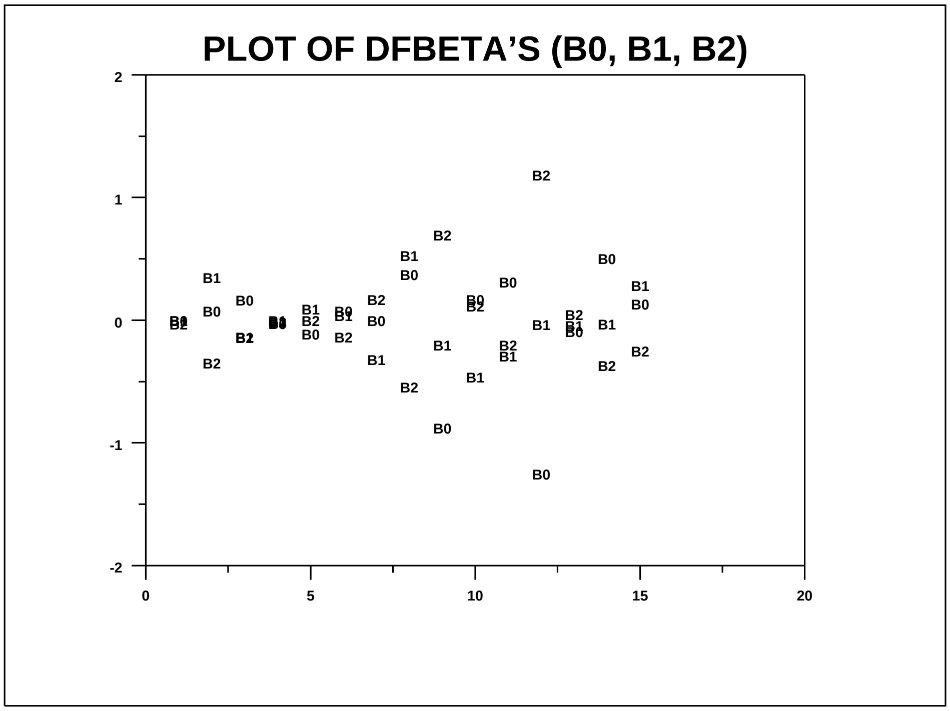
```
.  CALCULATE:
. 1) CATCHER MATRIX
. 2) VARIANCE INFLATION FACTORS
. 3) CONDITION NUMBERS OF X’X (BASED ON SINGULAR VALUES OF SCALED X)
. 4) PARTIAL REGRESSION PLOTS (ALSO CALLED ADDED VARIABLE PLOTS)
. 5) PARTIAL LEVERAGE PLOTS
. 6) DFBETA’S
IF N > 500
    PRINT “MORE THAN 500 POINTS, CAN’T CALCLUATE C MATRIX”
    QUIT
END OF IF
LET XMAT = MATRIX DEFINITION X1 N P
LET C = MATRIX MULTIPLY XMAT XTXINV
MULTIPLOT 2 2
CHARACTER CIRCLE; LINE BLANK
SPIKE BLANK; LIMITS DEFAULT
TIC OFFSET 0 0; MAJOR TIC MARK NUMBER DEFAULT
LET ICOUNT = 0
LOOP FOR K = 2 1 P
    LET ICOUNT = ICOUNT + 1
    LET TEMP1 = C^K*C^K
    LET XMEAN = MEAN X^K
    LET DENOM = SUM TEMP1
    LET TEMP2 = DENOM*(X^K - XMEAN)**2
    LET DUMMY = SUM TEMP2
    LET VIF(ICOUNT) = DUMMY
    LET XJDOTJ = C^K/DENOM
    LET DUMMY = COEF(K)
    LET TEMP1 = DUMMY*XJDOTJ + RES
    TITLE PARTIAL REGRESSION PLOT FOR X^K
    PLOT TEMP1 VS XJDOTJ
    LET XJDOTJ2 = XJDOTJ*XJDOTJ
    LET DUMMY = SUM XJDOTJ2
    LET PARTLEV = XJDOTJ2/DUMMY
    TITLE PARTIAL LEVERAGE FOR X^K
    PLOT PARTLEV
    LET DUMMY = XTXINV^K(K)
    LET DFBETA^K = (C^K*ESTUDRES)/SQRT(DUMMY*(1-HII))
END OF LOOP
END OF MULTIPLOT
LET RJ = 1 - 1/VIF
LET TOL = 1/VIF
. CALCULATE CONDITION INDICES
. (SCALE EACH COLUMN TO UNIT SUM OF SQUARES, SQUARE SINGULAR VALUES)
LOOP FOR K = 1 1 P
    LET JUNK = XMAT^K*XMAT^K
    LET ATEMP = SUM JUNK
    LET ATEMP = SQRT(ATEMP)
    LET XMAT^K = XMAT^K/ATEMP
END OF LOOP
LET SVALUES = SINGULAR VALUES XMAT; LET SVALUES = SVALUES*SVALUES
LET DUMMY = MAXIMUM SVALUES; LET DCOND = DUMMY/SVALUES
.
PRINT “ “; PRINT “ “
PRINT “ Rj-SQUARE      VIF TOLERANCE CONDITION INDICES”
SET WRITE FORMAT F5.3,5X,F10.5,F10.5,F10.5
PRINT RJ VIF TOL DCOND
```

The following output is generated.

```
Rj-SQUARE        VIF TOLERANCE CONDITION INDICES
0.323        1.47767   0.67674   1.00000
0.323        1.47767   0.67674  28.86588
0.000        0.00000   0.00000 126.81018
```



```
.
LET DUMMY = XTXINV1(1)
TITLE PLOT OF DFBETA'S (B0, B1, B2)
LINE BLANK ALL
CHARACTER B0 B1 B2
CHARACTER SIZE 2 ALL
LET TEMP4 = SEQUENCE 1 1 N
LET DFBETA1 = (C1*ESTUDRES)/SQRT(DUMMY*(1-HII))
PLOT DFBETA1 DFBETA2 DFBETA3 VS TEMP4
```


PLOT OF DFBETA'S (B0, B1, B2)
B0
B1
B2