MULTREG USERS MANUAL

Version 3.0

by

Sanford Weisberg

Revision of TR # 275
Also printed as TR # 298

UNIVERSITY OF MINNESOTA

SCHOOL OF
STATISTICS

October 1977
Reprinted July 1979

July, 1980

Multreg Users' Manual Update

The following pages in the Version 3.0 Multreg User's Manual of October 1977 should be replaced with the attached pages:

1-9, 21-26, 29-30, 33-36, 53-56.

CONTENTS

-continued-

weighted least squares, Σ is a diagonal matrix such that the i-th diagonal element is the reciprocal of the case weight used in weighted least squares. The least squares estimate of β is $\hat{\beta} = (X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1}Y$. The fitted values $\hat{Y}$ are given by $\hat{Y} = X\hat{\beta}$ and the residuals by $\hat{e} = Y-\hat{Y}$. We define the matrix V to be the nxn matrix $\Sigma^{-\frac{1}{2}}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-\frac{1}{2}}$. $\sigma^2 V$ is the variance matrix for $\hat{Y}$. The matrix V is called a projection matrix. The diagonal elements of V, called v_{ii}, are used in many of the computations.

4. Commands

The following gives a description of all MULTREG commands. Any command may be abbreviated by its first four letters.

Continuations. Any command in MULTREG may be continued on several lines. To do this, simply end the line to be continued with an ampersand (&) or plus sign (+) or the letter C preceded by at least one blank.

4.1 READ.

This command is used to read both formatted and unformatted data files, including many files that have been written or can be read by the program MATTER. (See pages 4-6 of this manual.) The syntax of the command is as follows:

For unformatted or MATTER type files (with or without format statements):

READ [FROM] fname [list of column numbers]

READ fname [list of column numbers]

READ [list of column numbers] FROM fname

For MATTER type files with more than one matrix on the file:

READ [matname][list of column numbers] FROM fname

In the above,

fname (required) is the name of the local file containing the data

matname (optional) is the name of the data matrix on the file that is to be read. The matrix name is needed only if more than one matrix is given on the file, and the matrix to be used is not the first one on the file.

The list of column numbers (optional) specifies the columns on the file to be used as Multreg variables. If the list is omitted or else consists of a single zero (0), then all columns (or the first 33 if more than 33) are used. This is a change from Version 3.2.

The separator FROM must be used if the filename on the command is not given immediately after the word READ.

MATTER files in "COLUMNS" format cannot be read by Multreg. However, files with more than one matrix will be read.

To read the file given on page 7 of this manual, any of the following commands may be used:

```
NEXT? READ HALD
NEXT? READ FROM HALD
NEXT? READ HALD 1 2 3 4 5
NEXT? READ MULTREG1 0 FROM HALD
NEXT? READ 1 2 3 4 5 FROM HALD
```

WARNING --- An unformatted file will not be properly read if (1) the file has 3 or 4 columns and (2) the first column contains non-numeric information. To read such a file, simply add a header statement. For example, if file FNAME has 75 cases and 3 columns, and Column 1 = alphabetic label, insert as the first line of the file the following: MAT 75 3. The file can then be read via READ FNAME 2 3.

4.1a. FREAD. This command is now equivalent to READ, section 4.1.

4.2 SET. Form: SET [keywords and parameters]

This command is used to change the values of several parameters as described below. Each keyword except BATCH requires that the user give a parameter value after the keyword. Several keywords may appear on one set command. All parameters have default values, so SET is never required.

SET WIDTH [value] is used to tell the program the number of columns on the terminal being used. Some commands will use the full width of the terminal. The default value of WIDTH is 72, which is appropriate for a standard teletype. The maximum value of 132 is appropriate for wide printers, such as DECwriters. WIDTH must be set to at least 40, and no more than 132.

SET TOL [value between 0 and 1]. This sets the pivot tolerance, as explained on page 17 . The default value is .0001; in version 2.1 it was .0000001.

SET BATCH. This command should be used as the <u>first</u> command in a Kultreg run when the job is run from a batch environment (e.g. cards as input; see section 5.3). This command sets the WIDTH to 132, generates a page feed, and will echo all input lines <u>exactly</u> as they are read, and slightly modifies the handling of input errors.

SET PLOTSIZE [NCOLS, NROWS] sets the size of scatter plots in command SCPLOT. The default size is 51, 31, producing a plot that is approximately 13 cm x 13 cm. The command SET PLOTSIZE 61 52 will give a plot that fills a standard sheet of paper, while SET PLOTSIZE 112, 52, the maximum allowed, will fill a sheet of standard computer paper.

SET WEIGHT [VARIABLE label or number]. This command instructs the computer to use weighted least squares, using the values in the specified column as weights. All values in the specified column must be either zero or positive. Cases given zero weight will be ignored in computing estimates. Whenever a SET WEIGHT command is entered, any case that was previously deleted (Section 4.4) will be restored.

SET PEEPSILON [Value] sets the bin-width for groupings cases on the "pureerror" command (section 4.34). The default values is 1.E-09.

To change from weighted least squares to unweighted least squares, the command SET WEIGHTS 0 or SET WEIGHTS VO should be used.

TEKPLOT parameters. Several parameters are available for use with the TEKPLOT command (relevant only to users with Tektronix terminals). All of these begin with a T

SET TSHAPE 0 (default): rectangular plotting area
1 : square plotting area

SET TSYM 2 (default): plot a cross (x)
0 : plot no symbol
See Tektronix PLOT10 Advanced Graphics II manual, p. 65, for other options.

SET TLINE -1 (default): no line joining points
0 : solid line
1 : dotted line

SET TLABEL 0 (default): use curser
1 : turn curser off

SET TXAXIS xmin xmax : sets minimum and maximum plotting values for X-axis. If xmin > xmax, the program computes optimum values.

SET TYAXIS ymin ymax : sets minimum and maximum values for Y-axis. If ymin > ymax, the program computes optimum values.

A typical command might be

```
      NEXT? SET WIDTH 132 TOL 1.E-6 WEIGHT 2 PLOTSIZE 51 31 +
? TLABEL 1 TSYM 0 TXAXIS 0 10.5
 WIDTH FOR PRINTING CHANGED TO 132
 PIVOT TOLERANCE CHANGED TO  .100000E-05
 VARIABLE X2   USED AS WEIGHTS.
 PLOT SIZE:   51 COLUMNS BY  31 ROWS.
 TEKTRONIX PLOT POINT LABELING ON.
 TEKTRONIX SYMBOL CODE IS   0
 TEKTRONIX PLOT XAXIS MIN, MAX ARE      0      10.50
```

4.3 **LABEL. Form: LABEL [list 1] AS [list 2]**

This command is used to assign up to 4 character labels to variable numbers. Any group of 4 or fewer characters not recognizable as a number is a valid label, except for a few reserved words like VS, FOR, AS.

The elements in list 1 may be either variable numbers or old labels. Each variable has a null label of V1 for variable 1, V2 for variable 2, etc. After the READ command is executed all new labels are lost.

For our example, the appropriate label command for the example is:

```
NEXT? LABEL 1 2 3 4 5 AS X1 X2 X3 X4 Y
```

Any number of labels may be specified in a single command. Any labeled variable may be refered to by its number or by its label.

4.4 DELETE. Form: DELETE [list of case numbers].

This command is used to delete the specified cases from the current data set. One would wish to delete a case if it were suspected of being an outlier, or an influential case. A careful analyst would wish to compare the fit of a model both with and without the questionable case. The result of this command is to treat the specified cases as if they had zero case weight. Deleted cases are not used in the computation of estimates, R^2, degrees of freedom, etc, but predicted values, and standard errors of prediction produced by the RESID command are given for deleted cases.

The computational method used by the DELETE command is outlined in the discussion of the updating techniques in Sec. 3.

Deleted observations may be restored in two ways. The usual method is to use the RESTORE command (Sec. 4.5). However, all deleted cases are restored when the SET WEIGHTS is used.

```
NEXT? DELETE 1 3 3
DELETED CASES ARE   1   3   9
```

If more than 25 cases are deleted, only the case numbers of the first 25 can be printed. The user should be aware that severe round off error may result from deleting many cases.

4.5 RESTORE. Form: RESTORE [case numbers]

This command will restore the specified cases (that have previously been DELETED) to the data set. If the list of case numbers is left off, all deleted cases are restored.

If a list of cases is specified, then the computations done by this command correspond to single steps of the updating methods given in section 3. If the list is left off, then the cross product matrix is recomputed from scratch. This is done to avoid the accumulation of roundoff errors.

```
      NEXT? RESTORE 1
 REMAINING DELETED CASES ARE   3   8

      NEXT? RESTORE
 ALL CASES RESTORED TO THE DATA SET.
```

Weighted least squares. If the SET WEIGHTS command had been used, the command RESTORE will also automatically reset to unweighted least squares.

4.6 LIST. Form: List [Var. list]

All the data for the variables specified are printed; if the list is off, all the data is printed.

Example:

```
NEXT? LIST
     X1          X2          X3          X4          Y
   7.000       26.00       6.000       60.00       78.50
   1.000       29.00       15.00       52.00       74.30
   11.00       56.00       8.000       20.00       104.3
   11.00       31.00       8.000       47.00       87.60
   7.000       52.00       6.000       33.00       95.90
   11.00       55.00       9.000       22.00       109.2
   3.000       71.00       17.00       6.000       102.7
   1.000       31.00       22.00       44.00       72.50
   2.000       54.00       18.00       22.00       93.10
   21.00       47.00       4.000       26.00       115.9
   1.000       40.00       23.00       34.00       83.80
   11.00       66.00       9.000       12.00       113.3
   10.00       68.00       8.000       12.00       109.4
**AVERAGE**
   7.462       48.15       11.77       30.00       95.42
**STD. DEV.**
   5.882       15.56       6.405       16.74       15.04
```

Saving data. In previous versions of Multreg, data was saved by a command LIST SAVE. This command has been replaced by a new command SAVE (see 4.6a, following). The command LIST SAVE will be recognized, and treated as exactly equivalent to SAVE.

4.6a. SAVE. Form: SAVE [list][AS fname]

This command is used to write the variables in the list as a formatted file on file fname. For example, the command SAVE 1 V3 V5 AS MYDATA would create a new formatted file called MYDATA with V1 = column 1, V3 = column 2, V5 = column 3. Options: If the list is left off, all variables are written on the file. If the phrase AS fname is left off the file written is called SAVER. The resulting file can be read by MATTER or by Multreg. More than one data matrix can be written on a single file.

4.12 REGS. Form: REGS [model]

Result: This command uses sweep as described in section 2 to obtain least squares estimates of the parameters in the model, their standard errors and corresponding t statistics. Additionally, R^2, the residual mean square and its square root and its degrees of freedom are printed. A typical command (without labels) is: REGS 4 ON 1 2 3. For regression through the origin, see section 1.4. If the set of regressions is linearly dependent (or colinear) (e.g., the tolerance test, Sec. 3, fails), a linearly dependent subset of maximal dimension will be used and appropriated error messages will be printed.

The regression of Y on X1 X2 X3 is given by

```
    NEXT? REGS Y ON X1 X2 X3
REGS     Y     ON    X1    X2    X3
VARIABLE      COEF'T          ST. ERROR          T VALUE
 BO         48.19363          3.913305            12.32
 X1         1.695890          .2045820             8.29
 X2         .6569149          .4423423E-01        14.85
 X3         .2500176          .1847109             1.35
         DEGREES OF FREEDOM  =       9
         RESIDUAL MEAN SQUARE=   5.345624
         ROOT MEAN SQUARE    =   2.312061
         R-SQUARED           =      .9823
```

4.13 ANOVA. Form: ANOVA [model] [BRIEF]

Result: An analysis of variance table is produced, with the independent variables entered <u>in the order given in the model specification</u>, from left to right. Cumulative R^2 are also given unless regression is through the origin. Note that ANOVA 1 ON 2 3 4 is different from ANOVA 1 ON 4 2 3.

If the keyword BRIEF is included, then only the regression and residual sums of squares are printed.

See section 1.4 for a discussion of models and section 4.12 for a discussion of linear dependence.

Below we give two anova tables for two different models.

```
      NEXT? ANOVA Y ON X1 X2 X3 X4
  ANOVA    Y     ON    X1    X2    X3    X4
                 INDIVIDUAL                      CUMULATIVE
SOURCE     DF       SS          MS        DF       SS          MS          R**2
 MEAN       1  .1184E+06   .1184E+06
 X1         1    1450.       1450.         1     1450.       1450.        .5339
 X2         1    1208.       1208.         2     2658.       1329.        .9787
 X3         1    9.794       9.794         3     2668.       889.2        .9823
 X4         1    .2470       .2470         4     2668.       667.0        .9824
RESIDUAL    8    47.86       5.983        12     2716.       226.3

      NEXT? ANOVA Y ON X1 X3 X2
  ANOVA    Y     ON    X1    X3    X2
                 INDIVIDUAL                      CUMULATIVE
SOURCE     DF       SS          MS        DF       SS          MS          R**2
 MEAN       1  .1184E+06   .1184E+06
 X1         1    1450.       1450.         1     1450.       1450.        .5339
 X3         1    38.61       38.61         2     1489.       744.3        .5482
 X2         1    1179.       1179.         3     2668.       889.2        .9823
RESIDUAL    9    48.11       5.346        12     2716.       226.3
```

In both of the above tables, X1 is fit first, giving a sum of squares of 1450. From the first table, the sum of squares for X3, adjusting for X1 and X2 is 9.794, while from the second table, the sum of squares for X3 adjusting for X1 (but ignoring X2) is 38.61. Fitting variables in differing orders will result in differing results.

The columns marked "cumulative" in the ANOVA table give the cumulative sums of squares usually referred to as the "regression" sums of squares. For example, in the first table, the 4 d.f. sum of squares 2668. is the sum of squares for regression on X1, X2, X3, and X4; this number will be the

estimated variance of a predicted value is varpred = $\hat{\sigma}^2$ (1 + v), the "1" due to the additional variability that is inherent in the new observation.

Besides the prediction, fitted value, and their standard errors, the value of v is also printed out. This quantity is of special interest in prediction, as it gives a rough measure of the reliability of the prediction. If v is large, then the new values of the independent variables are not like the values used in estimation, and the predicted or fitted values from the model might be very poor. "Large" corresponds to being bigger than any of the v_{ii} in the RESID output; certainly, if $v \geq 1$, predictions will be unreliable. We refer to the large v case as being extrapolation.

On the other hand, if v is small, say about equal in size to the smallest value of the v_{ii} from the residual output, then we should expect predictions to be quite good, as we are attempting to predict in the same region as the original data. This is called interpolation.

An example of the command is given below.

```
     NEXT? PREDICT Y ON X1 X3 FOR 8 12
PREDICTION OR FITTED VALUE =   96.7824
STD. ERROR(FITTED VALUE)   =   3.14972
STD. ERROR(PREDICTION)     =   11.5164
V = V(FIT. VAL)/RES. M.S.  =   .808490E-01

     NEXT? PREDICT Y ON X1 X3 FOR 8 80
PREDICTION OR FITTED VALUE =   130.406
STD. ERROR(FITTED VALUE)   =   60.6464
STD. ERROR(PREDICTION)     =   61.6498
V = V(FIT. VAL)/RES. M.S.  =   29.9737
```

4.17 PARCOR. Form: PARCOR [model]

This command prints the partial correlations of the dependent variable with the variables not in the model controlling for those in the model. It is useful in so-called stepwise "step-up" regressions.

```
      NEXT? PARCOR 5 ON 4
PARTIAL CORRELATIONS WITH Y
   X1          X2          X3
 .9568       .1302      -.8951

      NEXT? PARCOR 5 ON X4 X1
PARTIAL CORRELATIONS WITH Y
   X2          X3
 .5986      -.5657
```

The first command gives the partial correlation of Y with X1, and X2 and X3 adjusted for X4; the second gives the partial correlations adjusted for X4 and X1.

4.18 TRAN. Form: TRAN [Var1] = [type][variables][parameters]

The transformation command is used to create new variables from the old ones already in the data set. A typical command might look like

TRAN V3 = POWER V1 2 .5

which is read as instructing the program to take old variable V1, add .5 to each value of the variable, square the result, and store these values in variable V3. Each of the specifications that are used is explained below.

Var1. This is the variable number or label of the variable that is to be replaced or created by the transformed values. Var1 may correspond to any variable currently in the data set except for the column of weights if weighted least squares is being used. Alternatively, Var1 may be the number of the first unused variable in the data. For example, if the data currently has 10 columns or variables, then the command TRAN V11 = LOG10 V10 would create a new variable, V11, whose values are the logarithms (to the

base 10) of the data in V10.

If Var1 is left off, then the equals sign must also be left off. The destination is then assumed to be the first variable in the variable list following the "type" specification. For example, TRAN SQRT V3 would replace V3 by its square root.

Type. The type is the name of the transformation. A complete list of types is given in the table on the next page. (Note that some of the types listed in the table are not yet available, but should be added in the near future. Type TRAN HELP for a complete list of available types.) The names of the types are meant to convey their function, e.g., SQRT, POWER, PRODUCT, etc. The type specification is required.

Variables. This is the list of variables to be transformed. Most of the transformations only require one variable in this list, but some (SUM, DIFF, PROD, etc.) require two. The values of the variables in the list will not be damaged except as noted under "Var1" above.

Parameters. From zero to two parameters (e.g., numerical values) are required to perform the transformations. In the table, any parameter in square brackets, e.g., [C], may be left off, and it will assume the default value shown in the table. All two parameter transformations require one parameter, and the second is optional; all one parameter transformations have the parameter optional. Thus, the following two commands are identical: TRAN 2 = POWER 2 .33, and TRAN 2 = POWER 2 .33 0.

Table 4.18

Keyword	Number of Variables	Parameters	Results	Comments
POWER	1	P [C]	$(V+C)^P$	If P=0, use $\log_e$
LN	1	[C]	$\log_e(V+C)$	V+C>0
LOG10	1	[C]	$\log_{10}(V+C)$	V+C>0
EXP	1	[C]	$e^{(V+C)}$	$\|V+C\|<710$
EXP10	1	[C]	$10^{(V+C)}$	$\|V+C\|<310$
SQRT	1	[C]	$\sqrt{V+C}$	V+C>0
FT	1	[C]	$\sqrt{V+C} + \sqrt{V+C+1}$	V+C>0
LINEAR	1	A, [C]	A*V+C	V+C 0
SCALE	1	A, [C]	(V-C)/A	A≠0
SUM	2	none	V1+V2	
DIFF	2	none	V1-V2	
PROD	2	none	V1*V2	
RATIO	2	none	V1/V2	V2≠0
ANGLIT	1	[B]	$\sin^{-1}(\sqrt{V/B})$	0<V/B<∞, result in radians
ABS	1	[C]	$\|V+C\|$	
XLOGX	1	none	$(V)\log_e(V)$	V>0
LAG	1	[B]	$Y_i=V_{i-B}$	-N/2<B<N/2, circu lag. (see discuss
BCPOWER	1	[B]	$\frac{V^B-1}{B\,G^{B-1}}$	G=geometric mean V must be positiv (see discussion)
LSCORE	1	none	"Score" vector	See discussion
LOGIT, WLOGIT, ALOGIT, RLOGIT	2	[C]	Logit transform of p = (V1+C)/(V2+2 See discussion	
PROBIT, WPROBIT APROBIT, RPROBIT	2	[C]	Probit transform of p = (V1+C)/(V2+ See discussion	
DUMMY	none	List of cases	Cases listed set to 1, all others se	

[C] has default value C=0, is optional

[B] has default value B=1, is optional

See discussion

Lagged variables - If cases in a data set are ordered, usually in time, it is sometimes desirable to create new variables that are the same as old variables except that they are shifted up or down (lagged) one or more places. In Multreg, this is done via the command TRAN LAG. For example, in the table below, V2 = LAG V1 -2, V3 = LAG V1 -1, V4 = LAG V1 1, V5 = LAG V1 2

	LAG -2	LAG -1	LAG 1	LAG 2
V1	V2	V3	V4	V5
1	3	2	6	5
2	4	3	1	6
3	5	4	2	1
4	6	5	3	2
5	1	6	4	3
6	2	1	5	4

The LAG transformation is defined circularly: if the parameter B, the number of lagged periods, is positive, the number for the last B periods of V1 "wrap around" and are placed in the first B periods of the new variable and if B<0, the last |B| periods of the new variable have wrapped around values. In most applications the cases with the wrapped around values should not be used in computations. For example, if the LAG +2 variable is to be used in a regression, type

```
NEXT? TRAN V5 = LAG V1 2
NEXT? DELETE 1 2
```

If LAG -1 is to be used, and there are n=100 cases, type

```
NEXT? TRAN V3 = LAG V1 -1
NEXT? DELETE 100
```

BCPOWER. The form of this transformation is

$$VAR1 = \begin{cases} \dfrac{(V1)^B - 1}{B\ G^{B-1}} & B \neq 0 \\ G\ \ln(V1) & B = 0 \end{cases} \qquad (1)$$

where G = geometric mean of V1 = $\exp(\Sigma \ln(V1)/n)$. This family of the transform tion was defined by G.E.P. Box and D.R. Cox, in "An analysis of Transformatio J. Roy. Statist. Soc., Sec B (1964), 26, 211-46, see also Weisberg (1980), Applied Linear Regression, p. 136.

Generally, the Box-Cox power family transformation is applied to the depend variable (Y) in a regression equation. For each of several values of B, the regress is computed, and the residual sum of squares is found. The value of B that g the minimum residual sum of squares is the maximum likelihood estimate of B fo model

$$Y^B = X\beta + e\ ,\ e \sim N(0,\sigma^2 I) \qquad (1)$$

The geometric mean is included in the transformation to make the residual sum of squares be in the same scale for all B.

LSCORE. This transformation is related to the Box and Cox transformatio It is defined to be minus the score function, or

$$VAR1 = -\frac{\partial}{\partial B}\left[\frac{V1^B - 1}{B\ G^{B-1}}\right]_{B=1} = V1\ (1 - \ln(V1/C))$$

where G = geometric mean of V1.

This transformation was suggested by A.C. Atkinson, "Testing transformations to normality," J. Roy. Statist. Soc., Ser. B, (1973), 35, 473-479. Its use is briefly summarized below.

1. Suppose that fitting a model of the form (1) is of interest. As a first try, set B=1, and fit the model $Y = X\beta+e$. For example, suppose Y = V4 , and the independent variables are V1 V2 V3

2. To test the adequacy of this model (e.g. of B = 1), define a new variable V5 by

```
TRAN V5 = LSCORE Y
```

Then, fit the model Y ON V1 V2 V3 V5

3. The t-test of the coefficient for V5 for this model is an approximate t-test for B = 1. The number of d.f., $n - p' - 1$, is given in the REGS output.

4. If the t-statistic is large, then an estimate of B is given by $1+\hat{\gamma}$ where $\hat{\gamma}$ is the estimated coefficient for V5.

5. Let $\hat{B}$ be the maximum likelihood estimate of B obtained by the Box and Cox method. Then continuing the above example, compute

```
TRAN V6 = BCPOWER Y B̂
TRAN V7 = LSCORE V6
REGS V6 ON V1 V2 V3 V7
```

Then the standard error of the coefficient estimate for V7 can be used with the t distribution with $n - p' - 1$ degrees of freedom to get a confidence interval for B.

XLOGX. This transformation is helpful in finding trasnformation of independent variables via the method of G.E.P. Box and P.W. Tidwell, "Transformations of the independent variables," Technometrics (1962) 4, 531-50; see also Weisberg (1980) op.cit., p.141.

Counted data, proportions and percentages. Transformations that should be used for counted data, proportions or percentage data are now available in Multreg. The first of these, the ANGLIT or angular transformation, is used as a variance stabilizing transformation when the counts are binomially distributed. The second transformation, the PROBIT or integrated normal, is an empirical transformation used for counted data. The third, the LOGIT transformation, is used to fit logistic regression models. If either the probit or the logit transforms are to be applied to a dependent variable in a regression problem, it is usually appropriate to use empirical weights in weighted least squares. Additional transformations to obtain these weights and to perform iteratively reweighted least squares are described below.

Anglit. The form of this command is

TRAN VAR1 = ANGLIT V1 (B)

where: VAR1 is the destination variable, V1 is the variable to be transformed, B is a scale factor that has a default value of B=1 and V1/B must fall in the range [0,1]. The transformation computed is

$$VAR1 = \sin^{-1}(V1/B)^{1/2},$$

and the resulting angle is in radians. If V1 is a percentage rather than a proportion, it is appropriate to set B=100 . If V1 contains the number of successes, and V2 contains the number of trials, then the command TRAN V1 = RATIO V1 V2 should be used before calling TRAN ANGLIT.

Use of the anglit transformation is discussed in Snedecor and Cochran, Statistical Methods (1967), 6th edition, pp 327.

Logit and probit. These commands are invoked the same way:

TRAN VAR1 = PROBIT V1 V2 (C)

TRAN VAR1 = LOGIT V1 V2 (C)

where VAR1 is the result, V1 and V2 are existing variables such that

if V1 is the number of successes, V2 is the number of trials

if V1 is the proportion of successes, V2 is 0, indicating the column of 1's.

If V1 is a percentage, it must first be transformed to a proportion by use of the TRAN SCALE command. The optional parameter C has default value 0, and is used to handle cases with observed values of 0% and 100%. When these extreme values occur in the data, the value of C = .5 is recommended.

For both of these transformations, first a value p is computed where, for each case, $p = (V1 + C)/(V2 + 2C)$; each p must be between 0 and 1. The transformations are defined by:

LOGIT: $VAR1 = \ln(p/(1-p))$

PROBIT: $VAR1 = \Phi^{-1}(p) + 5.0$

where $\Phi^{-1}(p)$ is the inverse of the standard normal distribution evaluated at p.

If data transformed via the logit or probit transformation are to be used as dependent variables in a regression problem, it is usually appropriate to use weighted least squares with empirically determined weights. These weights are determined by the following commands:

```
TRAN VAR2 = WLOGIT  V1 V2 [C]
TRAN VAR2 = WPROBIT V1 V2 [C]
```

where V1, V2, and C should be identical to the values given in the corresponding probit or logit commands. Defining p as above, the transformations are:

WLOGIT: $VAR2 = V2(p(1-p))$

WPROBIT: $VAR2 = V2(\exp(-y^2)/p(1-p)(2\pi)$, with $y = \Phi^{-1}(p)$

For example, to do a weighted regression using the logit transform, with

V1 = number of successes, V2 = number of trials, V3 and V4 = independent variables, the following commands are appropriate:

```
NEXT? TRAN V5 = LOGIT V1 V2 .5
NEXT? TRAN V6 = WLOGIT V1 V2 .5
NEXT? SET WEIGHTS V6
NEXT? REGS V5 ON V3 V4
```

Two warnings are in order here: First, since the weights are empirically determined, usual interpretation of tests and the like do not apply exactly to the results obtained. Thus, tests, standard errors, residual plots, and similar Multreg results are only approximately correct. Second, methods other than least squares are usually used to fit these models; for logistic regression, maximum liklihood is a preferred estimation method , and the weighted least squares estimates obtained by one iteration in Multreg are not equivalent.

Inverse logit and probit. Suppose that in the above example, we add the following command

```
NEXT? KEEP 5 ON 3 4 PRED AS 7
```

The predicted values (V7) can be transformed back to the original scale via the inverse or antilogit on antiprobit command:

```
NEXT? TRAN V8 = ALOGIT V7 V2 [C]
NEXT? TRAN V8 = APROBIT V7 V2 [C]
```

where V7 = column of predicted values, V2 = column of sample sizes for fitted counts and, V2 = 0 for fitted proportions, C = optional constant, probably C=0. The computed transformation is for ALOGIT

$$V8 = \frac{1}{1+e^{-V7}}(V2 + 2C) - C$$

For the analogous APROBIT command, the transformation is

$$V8 = \Phi(V7 - 5)(V2 + 2C) - C$$

where Φ is the cumulative normal distribution function.

Iteratively reweighted least squares. A second weighted regression can be computed using the fitted values from the first regression to compute weights. This is facilitated with the commands RLOGIT, PROBIT:

```
TRAN V9 = RLOGIT V7 V2
TRAN V9 = RPROBIT V7 V2
```

where V7 = column of fitted values in the probit or logit scale (from KEEP command), V2 = column of sample sizes (or V2 = 0 if sample sizes are all equal). There are no parameters. The transformation are:

$$\text{RLOGIT} \quad V9 = V2(e^{-V7}/(1 + e^{-V7})^2)$$

$$\text{RPROBIT} \quad V9 = \frac{V2}{\Phi(V7 - 5)(1 - \Phi(V7 - 5))2\pi} e^{-(V7 - 5)^2}$$

The following commands are then appropriate (continuing the example)

```
NEXT? SET WEIGHTS V9
NEXT? REGS V5 ON V3 V4
```

Another iteration would be done by repeating KEEP, TRAN RLOGIT, SET and REGS.

For information on logit transform, see D.R. Cox, The Analysis of Bindry Data (1970), London: Methuen. The probit transform is discussed by D. J. Finney, Probit Analysis (1971), 3nd Edition, Cambridge U. Press.

Dummy variables. A dummy variable, with values 0 or 1 , is created in analogy to the following:

NEXT? TRAN V3 = DUMMY 1 3 5 TO 9 15 TO N

The new variable (V3) has all values equal to zero, except in the listed cases, where the value is one. Note the use of "TO", indicating cases 5,6,7,8, and cases 15 to case N, the last case in the data set. Case numbers must be ordered, smallest to largest. The created variable can replace any existing variable, or be the first unused variable. (Remember that Multreg can only read in 99 items on any command. Thus, a DUMMY transformation with 96 or more singly listed case numbers will not be properly read.)

Limitations. At most two statistics may be saved on one KEEP command; thus KEEP RESID AS 3 PRED AS 4 is permitted, but KEEP RESID AS 3 PRED AS 4 V AS 5 is not.

```
      NEXT? KEEP Y ON X1 X2 X3 X4 RESID AS 6 PRED AS 7
KEEP      Y    ON   X1   X2   X3   X4
PREDICTED KEPT AS V7
RESIDUALS KEPT AS V6

      NEXT? STAT
VARIABLE    N   AVERAGE      VARIANCE     ST. DEV.      MIN          MAX
  X1       13    7.462         34.60        5.882      1.000        21.00
  X2       13   48.15         242.1        15.56      26.00        71.00
  X3       13   11.77          41.03        6.405      4.000        23.00
  X4       13   30.00         280.2        16.74       6.000        60.00
  Y        13   95.42         226.3        15.04      72.50        115.9
  V6       13   .1366E-11       3.989       1.997     -3.175        3.925
  V7       13   95.42         222.3        14.91      72.79        115.6
```

4.26 EIGEN. Form: EIGEN [variable list] [VECTORS] [CORR]

This command is used to print the eigenvalues and, if requested, the eigenvectors of the corrected cross product matrix (or, if requested, the correlation matrix) of the variables specified in the list. If the list is left off, the eigenvalues of the whole data set are computed; for regression applications, one would usually want the eigenvalues for the set of independent variables only.

For the Hald data, this command is given by

```
NEXT? EIGEN X1 X2 X3 X4 VECTORS
EIGENVALUES
          6214.        810.0        148.9        2.846
PCT.     86.60        11.29         2.07          .04

EIGENVECTORS
X1     -.6780E-01      .6460       -.5673        .5062
X2      -.6785       .1999E-01      .5440        .4933
X3      .2902E-01      -.7553      -.4036        .5156
X4       .7309         .1085        .4684        .4844

NEXT? EIGEN X1 X2 X3 X4 CORR VECTORS
EIGENVALUES
          2.236        1.576        .1866      .1624E-02
PCT.     55.89        39.40         4.67          .04

EIGENVECTORS
X1       .4760         .5090        .6755        .2411
X2       .5639        -.4139       -.3144        .6418
X3      -.3941        -.6050        .6377        .2685
X4      -.5479         .4512       -.1954        .6767
```

If the word "VECTORS" did not appear on the command, only the eigenvalues would have been computed. If CORR appears on the command, the correlation matrix is used to find eigenvalues and eigenvectors.

The line of output immediately under the eigenvalues gives each eigenvalue as a percent of the sum of the eigenvalues.

Computations. The computations in this command are done by the triangular method using subroutine RS, part of the EISPACK library, written at the Argonne National Laboratories.

The order of printing the eigenvalues and eigenvectors has changed from version 3.2.

4.27 PRINCOMP (list 1) AS (list 2) [CORR] [VECTORS]

This command prints eigenvalues (and eigenvectors if keyword VECTORS is used) and saves principal components based on list 1, putting the results in the variable numbers in list 2, in decreasing order of the eigenvalues. For the Hald data, a typical command might be

```
NEXT? PRINCOMP X1 X2 X3 X4 AS 6 7 8 9
EIGENVALUES
      6214.      810.0      148.9      2.846
PCT.  86.60      11.29      2.07       .04

THE FIRST  4 PRINCIPAL COMPONENTS SAVED AS
V6   V7   V8   V9
```

List 1 must include column numbers currently in the data set. List 2 may include the first few previously unused columns, but if so, the new column numbers must be ordered smallest to largest with no unused columns. If list 2 has fewer elements than does list 1, then only the principal components corresponding to the largest eigenvalues are saved.

Computations. The $n \times p$ matrix Z of principal component scores is computed from the formula $Z = XU$ where X is determined from list 1 and U is a matrix of eigenvalues, see S. Weisberg, Applied Linear Regression (1980), Section 7.6. X, U are computed as follows:

Without keyword CORR, X = matrix of data given by list 1, U = eigenvectors of corrected cross product matrix for X .

With keyword CORR, X = as above except each entry is standarized by (value - column average)/column standard deviation; U = eigenvectors of correlation matrix.

With * in list 1: X = as above, U computed from uncorrected cross product matrix.

If weighted least squares is used, then U = eigenvalues of appropriate

weighted matrix; however, the column of weights may not appear in list 1 or in list 2.

4.27a SENSIT. Form: SENSIT [model][USING value list]

This command is used to estimate the sensitivities of estimated coefficients to small perturbations in the independent variables. The coefficient used is minus the common logarithm of the relative sensitivity coefficient suggested by G. W. Stewart, and discussed in S. Weisberg, Applied Linear Regression, (1980), pages 67-72. In the command, the value list contains estimated standard deviations for the perturbations in the independent variables in the same order as the independent variables appear in the model. If the list is not given, these standard deviations are all set to one. If fewer values are given than there are independent variables, ones are added to the end of the list.

If an $n \times p'$ independent variable matrix X is observed without error, with dependent variable Y, the least squares estimator for the model $Y = X\beta + e$ is $\hat{\beta} = (X^T X)^{-1} X^T Y$. Now, suppose that rather than observe X, we observe $\tilde{X} = X + E = \tilde{X}$, where E is a matrix of small observation errors or perturbations. We would say that perturbations are negligible if $\tilde{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T Y$ is nearly the same as $\hat{\beta}$.

The relative sensitivity coefficients g_{jk} are defined by

$$g_{jk} = \frac{\sigma_k (\hat{\beta}_k^2 c_{jj} + RSS\ c_{jk}^2)^{1/2}}{|\hat{\beta}_j|} \qquad \begin{matrix} j = 0, 1, \ldots, p \\ k = 1, \ldots, p \end{matrix} .$$

where $\hat{\beta}_k$, $\hat{\beta}_j$ are the observed least squares estimators, RSS is the residual sum of squares, and c_{jj}, c_{jk} are elements of $(X^T X)^{-1}$, and σ_k is the presumed standard deviation of the perturbations in column k. One can show that, if only column k of X is subject to relatively small perturbations (of magnitude σ_k), then,

$$\frac{|\tilde{\beta}_j - \hat{\beta}_j|}{|\hat{\beta}_j|} \leq s_{jk}$$

with high probability. Thus, if all the s_{jk} are small, the regression problem is insensitive to perturbations. The values printed by the command are G_{jk}, where

$$G_{jk} = \begin{cases} -\log(s_{jk}) & \text{if } |\hat{\beta}_j| > 1. \times 10^{-9} \\ -99 & \text{if } |\hat{\beta}_j| < 1. \times 10^{-9} \end{cases}$$

If G_{jk} is positive, then we expect that if X_k were perturbed, $\tilde{\beta}_j$ would agree with $\hat{\beta}_j$ to about G_{jk} digits; if $G_{jk} < 0$, then no agreement is expected. If any of the G_{jk} are small or negative, than the regression problem is sensitive to perturbations. In the computer output, rows correspond to coefficients estimated, and columns correspond to perturbed columns.

We now compute the sensitivities for the Hald data. Each of X1 to X4 is a percentage, and we can probably assume that each of these is accurately measured, except possibly for round-off error on the least significant digit. If we assume that the round-off error is uniformly distributed on the interval (-.5, .5), then variance of the round-off error is well known to be 1/12, so the standard deviation is $(1/12)^{1/2} \doteq 0.29$. Using this as the standard deviation for each of the X's, we find

```
     NEXT? SENS Y ON X1 X2 X3 X4 USING .29 .29 .29 .29
  SENSIT  Y    ON   X1   X2   X3   X4

              PERTURBED COLUMN
EST.      X1       X2       X3       X4
V0         .5       .6       .6       .6
X1         .8       .9       .9      1.0
X2         .4       .4       .5       .5
X3        -.4      -.3      -.3      -.2
X4        -.2      -.1      -.1      -.1
```

For example, if X4 were perturbed by adding in round-off error, the new estimate for the coefficient for X1 should agree with the current estimate to about 1.0 digits (e.g., $G_{14} = 1.0$). All in all, these values are very small, and many are negative, so the calculations are very sensitive to round-off error. Regression calculations for these data and the full model are not reliable.

4.28 END [FNAME]

End is the usual termination for the program. If a filename is given, then a binary file is written that includes all current information, including labels, set values of parameters, etc. This file can be read by Multreg only via the RELOAD command.

4.28a RELOAD [filename]

This command reads a file, created by an END command, restoring all values as they were before when the END was encountered, and losing any current values

```
      NEXT? VARS
NUMBER     LABEL
   1        X1
   2        X2
   3        X3
   4        X4
   5        Y
UNWEIGHTED (ORDINARY) LEAST SQUARES.
PIVOT TOLERANCE =    .10OE-05
NO. OF CASES WITH POSITIVE WEIGHT =    13

      NEXT? END FNAME
FILE FNAME      CAN BE RELOADED .
TOTAL CP TIME USED IS   1.276 SECONDS.
:X,MULTREG
M U L T R E G -- VERSION 3.4 80/06/20.    14.01.02.
80/06/16 --- NEW VERSION: USERS TYPE MESSAGE
FIRST STEP (OR TYPE 'READ HELP')
? RELOAD FNAME
FILE FNAME      RELOADED.

      NEXT? VARS
NUMBER     LABEL
   1        X1
   2        X2
   3        X3
   4        X4
   5        Y
UNWEIGHTED (ORDINARY) LEAST SQUARES.
PIVOT TOLERANCE =    .100E-05
NO. OF CASES WITH POSITIVE WEIGHT =    13
```

4.29 HELP.

This command results in the printing of a brief description of all the available commands. In addition, HELP may be used as a modifier for any command. For example, REGS HELP would get information concerning the regression command.

4.30 MESSAGE

This command allows communication between the user and the writers of the program. Recent information concerning changes in the program will be available here. MESSAGE will be updated periodically.

4.31 OUTPUT. Form: OUTPUT SAVE [ON or OFF or fname] PRINT [ON or OFF]

This command is used to control printing. If SAVE is ON, the results of certain commands are output on local file SAVER as they appear at the teletype (including headings); the default for SAVE is OFF. If a local file name is given in place of ON or OFF, the output is written on that file.

If PRINT is OFF, output from certain commands, except for error messages, is not printed at the user's terminal. The default of PRINT is ON.

For example, suppose a user wanted to save the values of C_p, but, because there are many variables the user does not want the output printed. The following commands are then appropriate

NEXT? OUTPUT SAVE ON PRINT OFF
OUTPUT FROM ALL AND RESIDUAL PRINTED ON FILE SAVER

Mixing output from SAVE and OUTPUT on one file, while permitted, is not recommended. The results of OUTPUT connot be read by Multreg or by MATTER.

4.32 CPTIME

This command will print the amount of central processor time used since the last "READ" command.

4.33 $COMMENT

Any input line beginning with a "$" will be ignored by the program. This command is used primarily to allow annotation of output for future reference.

```
NEXT? $ANY INPUT BEGINNING WITH A "$" IS IGNORED.
NEXT?
```

4.35 RPCPLOT [MODEL] vs [VARIABLE LIST]

The result of this command is to produce scatter plots. On the X axis, the values of the specified independent variable are plotted. The Y-axis has values of the "residuals plus components" plotted. These values are defined to be the sum of the residual for the model that is fit plus a component for the variable plotted on the X-axis (namely, that variable times its estimated slope from the regression in the model specified. This plot is useful in studying the relationship between the variable on the X axis and the independent variable after adjusting for all other variables in the model. These plots are also called "partial residual plots". For further discussion, see W. A. Larsen, and S. A. McCleary (1972), "The use of partial residual plots in regression analysis", Technometrics 14 781-90 or F. Wood (1973), "The use of individual effects and residuals in fitting equations to data", Technometrics 15, 677-95.

4.36 TEKPLOT YVAR VS [XVAR or option] [keyword],

where YVAR and XVAR are existing variables, and the optional keyword is described below. This command draws plots on Tektronix 4000 - series terminals. In addition, it provides the necessary information to turn on and off a Tektronix model 4662 pen plotter (provided that the plotter is referred to as unit "D"). This command should not be used by non-Tektronix users.

Options for X-axis

RANKIT - normal probability plot of YVAR
INDEX - plot YVAR vs index number
HNORM - half normal probability plot of YVAR

Keywords

ALL - Plot all cases

INCLUDED - (default) plot cases with positive case weight

DELETED - Plot cases with negative or zero case weight

Parameters. Many aspects of the plot can be altered using the SET command (parameters for TEKPLOT are TSHAPE, TLINE, TSYM, TLABL, TXAXIS, TYAXIS). See Sec. 4.3 for details.

Curser. On terminals equipped with a curser, if TLABL = 0 (the default), cross hairs will appear on the screen after plotting. Position cross hairs appropriately and type:

C to print the case number of the nearest case

V to print the vertical axis value for the nearest case

H to print the horizontal axis value for the nearest case

T to type a title starting at the location chosen and ending with a carriage return

A to replot the same data, but using user set minima and maxima for the axes (the screen will be automatically erased, and the user will be asked to supply minima and maxima) (Again)

S to terminate the command and turn off the curser. (Stop)

Residual plots. Because only existing columns can be used in TEKPLOT, a residual plot is produced by commands analogous to the following

```
NEXT? KEEP 5 CN 1 2 3 4 STUDRES AS 6 PRED AS 7

NEXT? LABEL 6 7 AS EHAT FIT   [OPTIONAL]

NEXT? TEKPLOT EHAT VS FIT
```

MULTREG User's Manual

by

Sanford Weisberg

Technical Report #298

Version 3.0

Reprinted with minor corrections July 1979

Corrected 1 March 1979

15 October 1977

School of Statistics
University of Minnesota
St. Paul, Minnesota 55108

The work is supported by grants from the Educational Development Program, University of Minnesota.

Version 3.0
Fall 1977

MULTREG USERS MANUAL

by
Sanford Weisberg
School of Statistics
University of Minnesota
St. Paul, Minnesota 55108

CONTENTS

-continued-

MULTREG User's Manual

0.

Introduction

MULTREG is a versatile interactive computer program for the analysis of data using the techniques of multiple regression. In addition to computing estimates, standard errors, and tests, the program can be used for a wide variety of statistical calculations, such as predictions, residual analysis, plotting, transformations, variable selection, weighted least squares, and cross validation.

This version of the program represents a significant change from earlier versions available at the University of Minnesota (from 1974-77), as both input and output features have been improved, and new operations have been added. The program was primarily designed by S. Weisberg, with important contributions by Christopher Bingham. Several others have written code for the program: Sharon Yang, Peter Stenberg, Lewis Paper, Ronald R. Regal, Paul Smith and Keith Howar-Lowe. The program is written in FORTRAN, and runs with the KRONOS operating system on CDC Cyber computers.

Support for this program has been provided by grants from the Educational Development Program of the University of Minnesota for 1975-76 and 1976-77.

Disclaimer. This computer program has been extensively tested and checked for accuracy and, to the best of our knowledge, contains no errors. However, neither the University of Minnesota, the Department of Applied Statistics, nor any of the authors of this program claim any responsibility for any errors that do arise.

Any comments, questions or problems with this program should be directed to Sanford Weisberg, University of Minnesota, Department of Applied Statistics, 1994 Buford Avenue, St. Paul, Minnesota 55108, (612) 373-1068, preferably in writing.

Using this manual. The relevance of this manual to the current version of MULTREG can be determined by the version number of the program printed as a header by the computer when the program is accessed. Changes in the digit to the right of the decimal (e.g., from 3.2 to 3.3) indicate that a minor change has occurred in the program, probably an addition of a new command or implementation of a new option. In this case, the current manual will still be in use. Information about the changes in the program can be obtained by entering the command MESSAGE (see section 4.30 page 55; MESSAGE may not be available on some systems). If the version number to the left of the decimal point changes, a new manual will be available.

This manual was written for Version 3.2.

1. General Information

1.1 Unformatted files. Data for this program must be available on a local KRONOS coded file. A case consists of one number or measurement for each of several columns. The use of the word "column" does not imply that the numbers for a column for different cases will always line up one below the other. Columns must be in the same order for each observation and all observations must have the same number of columns, separated by blanks or commas: thus 10,402 is read as two numbers, 10 and 402. A typical unformatted file is given on page 13.

Although a case is usually equivalent to a line in the data file, each observation may take any number of lines with up to 150 spaces per line (including blanks) provided that each line to be continued ends with the characters CONT. In particular, output from the ISIS transformation routine TRDAT2 at the University of Minnesota is valid input to MULTREG. Non-numeric columns may appear on the file but may not be specified as data. Missing values are not permitted.

The numbers on a file may or may not have decimal points. Also, scientific notation (E or D format) is allowed. However, all numbers must have ten or fewer characters; otherwise, leftmost characters will be lost. [This is subject to change to 20 characters in the future.]

Missing data. As stated above, Multreg will not accept files with missing data. Users with large data sets with many missing values are urged to use the program BMDPAM, in the UCLA Health Sciences Computer Facility BMDP series. This program gives the user good editing capabilities for processing data with missing values.

Limitations: No more than 1000 observations on no more than 99 columns are permitted; no more than 33 variables are allowed.

Choosing the data set. The user may specify up to 33 columns to be copied from the unformatted file and thus become MULTREG variables. Variables are renumbered internally by MULTREG in the order they are specified. (Note: Line numbers on the data file, if any, count as a column.) Thus, specifying columns 4, 2, 5, in this order would cause the data in column 4 to become variable 1 (V_1), column 2 to become V_2 and column 5 to become V_3. Formatted files always use all columns (or the first 33) as variables.

Options: If Ø (zero) is entered, all the columns (the first 33 if more than 33) become MULTREG variables in the order that they occur. Any other column numbers entered with Ø are ignored.

Multiple specifications, e.g. "5 2 2" are no longer needed by the TRAN command.

Unformatted files are read via the READ command (sec. 4.1)

1.2 Formatted files. As a new feature of Version 3, Multreg will now read formatted files, as long as they follow a very strict pattern. In particular, Multreg can now read files created by the program "Matter," and can also read files from the LIST SAVE command (sec. 4.6). Formatted files must have the following form:

1. The first line of the file consists of a one word alphabetical identifier (optional) followed by the number of cases on the data file and then the number of variables. Finally, the word "FORMAT" may optionally be at the end of the line (this is required for Matter files). EXAMPLE:
DATANAME 20 3 FORMAT

2. The second line consists of the format statement, e.g. (3F4.0, 2F6.1), or whatever the appropriate format is. Format statements are surrounded by parenthesis. Only "F", "G", "E" and "X" formats may be used.

3. The rest of the file contains the data, one case at a time (of course a case may take as many lines in the file as specified by the format statement).

Unlike unformatted files, a read of a formatted file adds all of the variables specified by the format statement (or the first 33 if more than 33) to the data set.

Files from Matter. Matter files may have more than one data set on them. Currently, only the first can be read by Multreg. Similarly, the LIST SAVE command can be used to save more than one data set; all of these can be read by Matter; however, only the first can be read by Multreg.

Formatted files are read using the FREAD command (sec. 4.1a)

An example of a formatted file is given below.

```
LNH,F=FHALD
 MULTREG1          13          5 FORMAT
(7G15.8)
  7.0000000       26.000000       6.0000000       60.000000       78.500000
  1.0000000       29.000000       15.000000       52.000000       74.300000
  11.000000       56.000000       8.0000000       20.000000       104.30000
  11.000000       31.000000       8.0000000       47.000000       87.600000
  7.0000000       52.000000       6.0000000       33.000000       95.900000
  11.000000       55.000000       9.0000000       22.000000       109.20000
  3.0000000       71.000000       17.000000       6.0000000       102.70000
  1.0000000       31.000000       22.000000       44.000000       72.500000
  2.0000000       54.000000       18.000000       22.000000       93.100000
  21.000000       47.000000       4.0000000       26.000000       115.90000
  1.0000000       40.000000       23.000000       34.000000       83.800000
  11.000000       66.000000       9.0000000       12.000000       113.30000
  10.000000       68.000000       8.0000000       12.000000       109.40000
```

1.3 Commands. A command is an instruction to the computer to perform a specified task, such as compute a regression, plot residuals, perform a transformation, etc.

Any command may be abbreviated by its first four characters (e.g., PREDICT becomes PRED, etc.); any alphabetic character after the fourth is ignored.

THE USER IS WARNED THAT "Ø" (ZERO) IS NOT THE SAME CHARACTER AS THE LETTER "O" (OH).

1.4 Models. Many commands require specification of a model. The choice of a linear model means the specification of a dependent variable and at least one independent variable, usually given by an equation such as

$$V_6 = \beta_{\emptyset} + \beta_4 V_4 + \beta_2 V_2 + \beta_1 V_1 + e.$$

In the above equation, V_6 is the dependent variable and V_4, V_2 and V_1 are the independent variables. (The "e" in the equation above is the usual error term). For this program, the above equation (model) is specified by "6 ON 4 2 1". More generally, a model is specified by typing the variable number or label of the dependent variable followed by "ON" followed by a string of variable numbers or labels specifying the independent variables. The variable numbers or labels must be separated by blanks or commas. A typical command would be "REGS 6 ON 4 2 1", which should be read as "compute the estimates of parameters for the regression in the model 6 on 4 2 1." For commands in which the order of the independent variables matters (e.g., ANOVA), independent variables are entered in the order given. See Section 4.13 for further details.

Options: Models requiring regression through the origin ($\beta_\emptyset = \emptyset$) are specified by including a star ("*") in the variable number list; thus, "6 ON * 4 2 1" fits the model

$$V_6 = \beta_4 V_4 + \beta_2 V_2 + \beta_1 V_1 + e.$$

In the ANOVA command it may be of interest to fit the mean after the variables in the model, the reverse of the usual fitting order. This can be accomplished by using a "*" and a $\emptyset$ (zero) in the variable list: "ANOVA 7 ON * 1 5 3 $\emptyset$". Variable $V_\emptyset$ is always considered to be a constant, ($V_\emptyset = 1$).

The star, if used, should appear to the right of "ON" in the model specification.

Restrictions: No variable number should appear more than once in a model, otherwise an error message will be printed.

Defaults: If a model is not specified where required, the last specified model is assumed. Thus, the command ANOVA preceded by REGS 7 ON 1 4 5 will give the analysis of variance table for the model 7 ON 1 4 5.

1.5 Weighted least squares. When computing begins, Multreg assumes that all computations are to be done using unweighted (ordinary) least squares. This means that every case in a study is given equal weight. In some applications, it is desirable to use unequal weights for the observations. Weighted least squares is now an option in Multreg. The weights, if not all equal, must be included in the data set as one of the columns of data (the column of weights can even be used as a variable in the model, if desired). The variable to be used as weights is set by the command SET WEIGHTS as described in section 4.2. Cases will be weighted in proportion to their case weight, cases with higher weight being more important (the exact formulas for weighted least squares are given in section 3 of this manual). If one wishes to weight "inversely proportional to variance", then a column containing the variances of the observations would have to be transformed via the TRAN command to reciprocals before weighting is done.

Zero case weights. One use for the weighted least squares is the option of having zero case weights. Cases with zero weight are ignored in all the computations, except in plotting, and residual analysis. This allows a simple way to do cross validation.

A more complete discussion of the use of weights is given in section 6.3. Use of the SET WEIGHTS command is given in section 4.2.

1.6 New features in Version 3.2. Version 3.2 represents almost a complete rewriting of the program. For this version, a field length of only 45000B is needed, as compared to 54000B for version 2.1. In addition, the program should be considerably faster, and therefore less expensive to run. The major changes in the program are briefly described below.

1. The program no longer asks the user to "name a data file". Rather, the user requests that a file be read via a READ or FREAD command (section 4.1 and 4.1a). This change was made in part for batch use of the program, where the fi step of a job would be SET BATCH, rather than READ (see sections 4.2 and 5.3).

2. Weighted least squares is now an option (sec. 4.2 and 6.3).

3. Cases may be deleted from the data set, and then restored (sections 4.4 and 4.5)

4. A new command, SET, has been added for the setting of parameters for the computations. The parameters include width for printing, the value of the tolerance for adding variables to an equation, the column for weights in weighted least squares, the size of scatter plots in the SCPLOT command, and setting a flag if input is from cards (BATCH input) rather than from a teletype. See sec. 4.2 for the details.

5. The transformation routine has been completely rewritten, has a new form, and is greatly expanded. See sec. 4.18.

6. Scatter plots are now available by command SCPLOT (sec. 4.20).

7. The output from the RESIDUAL command has been increased, and a new command, YHAT, has been added. If the printing width is greater than 120, the output from both YHAT and RESID is printed in the RESID command. New statistics computed include the so-called "predicted" residuals, the predicted

residual sum of squares (PRESS), and, if no cases are deleted, the Durbin-Watson statistic which assumes that the observations are ordered and equally spaced). See secs. 4.21 and 4.22 for details.

8. The output from ALL has been slightly modified to agree with that for SCREEN (sec. 4.23).

9. A new command, SCREEN, is an implementation of the "leaps and bounds algorithm" of Furnival and Wilson (1974) for finding the "best" few subset regressions. Multreg uses a subroutine that was written by Furnival and Wilson. (see sec. 4.24).

10. The residuals and fitted values from a regression can now be stored as variables using the KEEP command. This may be useful either in plotting, or in iteratively reweighted least squares. (sec. 4.25).

11. Eigenvalues and eigenvectors (e.g., principal components) can now be computed using the EIGEN command (sec. 4.26). In addition, the command PRINCOMP will allow the storing of the so-called "principal component scores" for doing regression on principal components. (sec. 4.27)

1.7 HELP: an instant manual. Help is available to the user at a terminal. The command HELP will cause the printing of a list of commands. Furthermore, any command followed by the work "HELP", for example REGS HELP, will result in the printing of information about that command. HELP is available for every command.

An "instant manual," actually a listing of all the HELP output, can be obtained by listing a file. On MERITSS, the file REGHELP (type FETCH, REGHELP) all the help output; on the other systems, enquire locally.

The instant manual should not be considered to be a substitute for this manual.

2. A Running Example

To illustrate the uses of the program and the discussion of the computations, an example will be carried throughout most of this manual. The example used, first given in the statistical literature by A. Hald, in Statistical Theory with Engineering Applications (1952), p. 647, has been extensively analyzed elsewhere, especially by Draper and Smith (1966), pp. 164-67, 365-402, and by Daniel and Wood (1971), Chapters 6 and 9. A complete analysis is given in Chapter 9 of the last reference; no attempt at a complete analysis will be given here.

The problem is to relate the cumulative heat of hardening after 180 days of thirteen different cements to the percent content of the cements of 4 different components. The variables are

Y = Cumulative heat of hardening after 180 days

X1 = percent tricalcium aluminate

X2 = percent tricalcium silicate

X3 = percent calcium aluminum ferate

X4 = percent dicalcium silicate

We treat HEAT as a dependent variable, and the others as independent variables.

The data, in a form suitable for MULTREG, is available on Meritss file HALD/UN=2051999, with Column 1 = X1, Column 2 = X2; Column 3 = X3; Column 4 = X4; Column 5 - Y. The file is listed below.

```
LNH,F=HALD
7 26 6 60 78.5
1 29 15 52 74.3
11 56 8 20 104.3
11 31 8 47 87.6
7 52 6 33 95.9
11 55 9 22 109.2
3 71 17 6 102.7
1 31 22 44 72.5
2 54 18 22 93.1
21 47 4 26 115.9
1 40 23 34 83.8
11 66 9 12 113.3
10 68 8 12 109.4
:
```

3. Computations

This section describes the method of computation used in the program. The user is encouraged to read this section, although the program can still be run without it, except, perhaps the commands SWEEP and CSWEEP.

Assume that we have k variables X_1, X_2, . . . X_k. The program implicitly adds an additional variable $V_\emptyset$ with constant value of 1, $V_\emptyset = 1$ for all cases. The data is read in casewise, e.g. one case at a time. A matrix of corrected cross-products, say C is computed by the method of updating, which is described below. The method of updating permits reading in the data in a single pass in a way that is numerically stable. It is a continuous technique so that after n_o cases, we will have computed the means and cross product matrix for those n_o cases. If we add an additional case, we will update the means and cross product matrix to be correct for $n_o + 1$ data points.

Let $x_{i\ell}$ be the value of the i-th variable for the ℓ-th case, and let w_ℓ be the case weight for the ℓ-th case (if unweighted least squares, that $w_\ell = 1$ for all ℓ.) Let $\bar{x}_i(n_o)$ be the means and $c_{ij}(n_o)$ be the elements of cross products matrix after n_o cases. These quantities are defined by

$$(1) \qquad \bar{x}_i(n_o) = \frac{\sum_{\ell=1}^{n_o} w_\ell x_{i\ell}}{\sum_{\ell=1}^{n_o} w_\ell}$$

$$(2) \qquad c_{ij}(n_o) = \frac{\sum_{\ell=1}^{n_o} w_\ell (x_{i\ell} - \bar{x}_i(n_o))(x_{j\ell} - \bar{x}_j(n_o))}{\sum_{\ell=1}^{n_o} w_\ell}$$

Now, suppose that the $(n_o + 1)$-st case is read in, and has values x_{i,n_o+1}, $i=0,\ldots k$ and case weight w_{n_o+1}. We first update the means to get $\bar{x}_i(n_o + 1)$ by the formula

$$(3) \qquad \bar{x}_i(n_o + 1) = \frac{\sum_{\ell=1}^{n_o} w_\ell}{\sum_{\ell=1}^{n_o} w_\ell + w_{n_o+1}} \bar{x}_i(n_o) + \frac{1}{\sum_{\ell=1}^{n_o} w_\ell + w_{n_o+1}} x_{i,n_o+1}$$

and we update the elements of c_{ij} by

$$(4) \qquad c_{ij}(n_o + 1) = c_{ij}(n_o) + \frac{(w_{n_o+1})\left(\sum_{\ell=1}^{n_o} w_\ell\right)}{\sum_{\ell=1}^{n_o} w_\ell + w_{n_o+1}} (x_{j,n_o+1} - \bar{x}_i(n_o))(x_{j,n_o+1} - \bar{x}_j(n_o))$$

These relationships are repeatedly applied until all n cases have been read.

A similar updating formula can be found for <u>deleting</u> cases (as in the DELETE command) by solving (3) and (4) for $\bar{x}_i(n_o), c_{ij}(n_o)$ in terms of $\bar{x}_i(n_o + 1)$ and $c_{ij}(n_o + 1)$.

In Multreg, computations, are done using an <u>augmented</u> corrected cross product matrix S, with elements s_{ij} defined by

$$s_{ij} = c_{ij}(n)$$

$$s_{i\emptyset} = \bar{x}_i(n)$$

$$s_{\emptyset j} = -\bar{x}_j(n)$$

$$s_{\emptyset\emptyset} = 1/\sum_{\ell=1}^{n} w_\ell$$

For the running example described in section 2, we have k = 5 variables, X_1, X_2, X_3, X_4 and Y. In MULTREG, the correct cross product matrix is printed

by the command SWEEP. Only the lower triangular part of the matrix is printed as the matrices printed are symmetric (except, perhaps, for a sign change). (In the example, all $w_\ell = 1$, e.g., unweighted least squares.)

```
      NEXT? SWEEP
SWEEP
V0    .7692E-01
X1    7.462      415.2
X2    48.15      251.1      2906.
X3    11.77      -372.6     -166.5     492.3
X4    30.00      -290.0     -3041.     38.00      3362.
Y     95.42      776.0      2293.      -618.2     -2482.     2716.
        V0         X1         X2         X3         X4         Y
```

The first column in SWEEP gives the variable means; e.g., the mean of X1 is 7.462. The remainder of the matrix is (n-1) times the sample (weighted) covariance matrix, where n is the number of observations. Thus $(n-1)\hat{\text{Var}}\ X1 = 415.2$ and $(n-1)\hat{\text{Cov}}(X1,X2) = 251.1$. The appropriate divisions needed to obtain the covariance and correlation matrices are carried out by the COVARIANCE and CORRELATION commands respectively.

Sweep. The fundamental computational tool used by the program is the sweep operator (Beaton (1964), Dempster (1969)). Following Beaton, a square matrix $S = (s_{ij})$ is said to have been swept on the r-th pivot when it has been transformed into a matrix $T = (t_{ij})$ such that:

$$t_{rr} = 1/s_{rr}$$

$$t_{ir} = -s_{ir}/s_{rr} \qquad i \neq r$$

$$t_{rj} = s_{rj}/s_{rr} \qquad j \neq r$$

$$t_{ij} = s_{ij} - s_{ir}\, s_{rj}/s_{rr} \qquad i,j \neq r$$

The sweep operator possesses the following properties:

1. Sweep is reversible; that is, sweeping a matrix twice on the same pivot is equivalent (except for rounding error) to not having swept at all.

2. Sweep is commutative; that is, sweeping a matrix first on pivot r and then pivot q is equivalent (except for rounding error) to sweeping first on the q-th pivot and then the r-th.

3. If a non-singular matrix S is swept once on each pivot, the resulting matrix is, except for rounding error, equal to S^{-1}.

4. Tolerance. In the above calculation, if s_{rr} is near zero, then t_{rr} will be near infinity. $s_{rr} = 0$ implies that the matrix S is singular, or that the r-th variable can be expressed as a linear combination of the variables that have already been swept into the model. Since computers are subject to round off error, we check to see if s_{rr} is too small; if so, we do not sweep on the r-th pivot (and an appropriate message is printed). The value for failing to sweep can be set by the user by choosing a minimum value for the "tolerance" in the command SET (section 4.2). The tolerance can be thought of in much the same way as a coefficient of determination R^2. The tolerance is equal to one minus the square of the multiple correlation coefficient for the regression of the r-th variable on the variables that have already been swept. The default value of the minimum tolerance in this program is .0001, so variables will be as swept only if .01% or more of their variability is not explained by other variables that have been swept.

(In version 2.1 of Multreg, the minimum tolerance was set by the program at .00000001, so that, using the default tolerance, it is possible that a data set that had no evidence of singularity in version 2.1 would now show one. However, the default minimum tolerance can be changed by the user.)

5. Suppose we sweep the matrix S on pivots 1, 2, 3, . . . , k-1. Then, the resulting matrix can be interpreted as:

$$\text{SWEEP } [1, 2, \ldots, k-1]\, S = \left[\begin{array}{c|c} (X'X)^{-1} & \begin{matrix} -\hat{\beta}_{\emptyset} \\ \vdots \\ -\hat{\beta}_{k-1} \end{matrix} \\ \hline \hat{\beta}_{\emptyset}\hat{\beta}_1 \cdots \hat{\beta}_{k-1} & (n-k)\hat{\sigma}^2 \end{array}\right]$$

where $\hat{\beta} = (\hat{\beta}_{\emptyset}, \ldots, \hat{\beta}_{k-1})$ are the least squares estimates of the parameters in the model

$$X_k = \beta_{\emptyset} + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-1} X_{k-1} + e ,$$

$\hat{\sigma}^2$ is the residual mean square estimate of $\text{Var}(e) = \sigma^2$, and $(X'X)^{-1}$ is the matrix such that $\text{Var}(\hat{\beta}) = (X'X)^{-1}\sigma^2$. More generally, the regression of any variable X_p with unswept pivot on a set $X_{i_1}, \ldots, X_{i_r}$ is obtained and can be analyzed by sweeping on pivotals $i_1, \ldots, i_r$ (but no others).

In the example, suppose we sweep on pivot 1 (or variable X1):

```
      NEXT? SWEEP 1
SWEEP      X1
V0      .2110
X1    -.1797E-01    .2408E-02
X2      43.64         .6047     |  2754.
X3      18.46        -.8974     |  58.77      157.9
X4      35.21        -.6984     | -2866.     -222.2      3159.
Y       81.48         1.869     |  1824.      78.09      -1940.      1266.
          V0           X1       |   X2         X3          X4          Y
```

This output contains a wealth of information. If we consider Y to be dependent variable, then the diagonal element in the column labeled Y, 1266., is the residual sum of squares for the model Y on X1. (Note that we also get the residual SS for the models X4 on X1, X3 on X1 and X2 on X1 if these are of interest.)

The lower hand corner gives the estimates of the β coefficients; for Y on X1, the estimates are $\hat{\beta}_0 = 81.48$, $\hat{\beta}_{X1} = 1.869$. The upper left hand corner corresponds to $(X'X)^{-1}$, and is needed in estimating the standard errors of the $\hat{\beta}$'s, and of estimation, prediction, etc. Thus, for example, all of the information for the REGS and PREDICT command can be obtained from a single sweep command.

Continuing, we give below the result of sweeping on 2 (and only on 2). Note that the matrix has been reordered so that swept rows and columns come first, and unswept ones last. This command gives information for Y ON X2. Note the residual SS for this model, 906.3.

```
      NEXT? SWEEP 2
SWEEP      X2
V0      .8749
X2   -.1657E-01   .3442E-03
X1      3.301      .8641E-01 | 393.5
X3      14.53     -.5731E-01 |-358.2      482.8
X4      80.40      -1.047    |-27.23      -136.3      179.4
Y       57.42       .7891    | 577.8      -486.8     -81.97      906.3
         V0          X2      |  X1         X3          X4          Y
```

Finally, we give the output for SWEEP 1 2 :

```
      NEXT? SWEEP 1 2
SWEEP      X1     X2
V0      .9026
X1   -.8387E-02   .2541E-02
X2   -.1585E-01 -.2196E-03   .3631E-03
X3      17.53      -.9103      .2134E-01 | 156.7
X4      80.62   -.6920E-01     -1.041    |-161.1      177.5
Y       52.58       1.468       .6623    | 39.17      -41.99      57.90
         V0          X1          X2      |  X3         X4          Y
```

This fits the model HEAT ON X1 X2. The residual SS, 57.90, is the residual for fitting X1 and X2. If we subtract this from 906.3 from the last output, we get 906.3 - 57.90 = 848.4, which is the sum of squares for X1 after X2. Similarly, the sum of squares for X2 after X1 is given by 1266 - 57.9 = 1208.1. This suggests a technique for finding the analysis of variance table: sweep S, column by column in the order specified by the model. The sum of squares for each variable in order is then found by subtracting residual sums of squares.

Regression through the origin. Multreg fits an intercept as its default; however, regression through the origin is obtained by specifying a "*" in a model. Computationally, the technique used is to find an "uncorrected" cross product matrix, by sweeping once on column 0. For the model $Y = \hat{\beta}_1 X_1$,

SWEEP * X1

X1	.8780E-03				
X2	4.321	.1178E+05			
X3	.6752	3878.	1774.		
X4	2.300	4417.	2859.	9035.	
Y	8.808	.1868E+05	7208.	.1166E+05	.3273E+05
	X1	X2	X3	X4	Y

Mathematical notation. In this section, we give the basic regression model in matrix notation; this notation will be used in the description of commands where the computations are not straightforward.

The model fit by Multreg is given by

(5) $\quad Y = X\beta + e$

where Y is an n x 1 vector, X is n x k′, where k′ = k, the number of independent variables if regression is through the origin, and k′ = k + 1 if regression is not through the origin. Rows of X correspond to cases; columns of X correspond to variables. e is an n x 1 random vector such that $\mathrm{Var}(e) = \sigma^2 \Sigma$, where, for unweighted least squares $\Sigma = I$, and for

weighted least squares, Σ is a diagonal matrix such that the i-th diagonal element is the reciprocal of the case weight used in weighted least squares. The least squares estimate of β is $\hat{\beta} = (X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1}Y$. The fitted values $\hat{Y}$ are given by $\hat{Y} = X\hat{\beta}$ and the residuals by $\hat{e} = Y-\hat{Y}$. We define the matrix V to be the nxn matrix $\Sigma^{-\frac{1}{2}}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-\frac{1}{2}}$. $\sigma^2 V$ is the variance matrix for $\hat{Y}$. The matrix V is called a projection matrix. The diagonal elements of V, called v_{ii}, are used in many of the computations.

4. Commands

The following gives a description of all MULTREG commands. Any command may be abbreviated by its first four letters.

Continuations. Any command in MULTREG may be continued on several lines. To do this, simply end the line to be continued with an ampersand (&) or plus sign (+) or the letter C preceded by at least one blank.

4.1 READ. Form: READ [FILE NAME] [COLUMN NUMBERS].

This command instructs the program to read data from the local file named. The list of column numbers gives the columns from the file to be used as variables, see section 1.2. If the number Ø (zero) is entered in place of the column numbers, then all the columns (or the first 33 if more than 33) will be the variables. A typical command is given by

```
:X,DO,MULTREG
M U L T R E G - VERSION 3.2
*****NEW VERSION OF SUMMER, 1977.  TYPE 'MESSAGE'.
FIRST STEP (OR TYPE 'READ HELP')
? READ HALD 0
NO. OF COLUMNS ON FILE =   5
NO. OF CASES ON FILE   =  13
NO. OF COLUMNS CHOSEN  =   5
```

If column numbers or the file name are not given, prompts will be provided.

Note that READ may be used repeatedly, each time changing to a different data file.

4.1a FREAD. Form: FREAD filename

This command is used to read formatted files, as described in section 1.2. Files written by LIST SAVE or by Matter can be read by this command. An example of a formatted file is given on page 7.

4.2 SET. Form: SET [keywords and parameters]

This command is used to change the values of several parameters as described below. Each keyword except BATCH requires that the user give a parameter value after the keyword. Several keywords may appear on one set command. All parameters have default values, so SET is never required.

SET WIDTH [value] is used to tell the program the number of columns on the terminal being used. Some commands will use the full width of the terminal. The default value of WIDTH is 72, which is appropriate for a standard teletype. The maximum value of 132 is appropriate for wide printers, such as DECwriters. WIDTH must be set to at least 40, and no more than 132.

SET TOL [value between 0 and 1]. This sets the pivot tolerance, as explained on page 17 . The default value is .0001; in version 2.1 it was .0000001.

SET BATCH. This command should be used as the first command in a Multreg run when the job is run from a batch environment (e.g. cards as input; see section 5.3). This command sets the WIDTH to 132, generates a page feed, and will echo all input lines exactly as they are read, and slightly modifies the handling of input errors.

SET PLOTSIZE [NCOLS, NROWS] sets the size of scatter plots in command SCPLOT. The default size is 51, 31, producing a plot that is approximately 13 cm x 13 cm. The command SET PLOTSIZE 61 52 will give a plot that fills a standard sheet of paper, while SET PLOTSIZE 112, 52, the maximum allowed, will fill a sheet of standard computer paper.

SET WEIGHT [VARIABLE label or number]. This command instructs the computer to use weighted least squares, using the values in the specified column as weights. All values in the specified column must be either zero or positive. Cases given zero weight will be ignored in computing estimates. Whenever a SET WEIGHT command is entered, any case that was previously deleted (Section 4.4) will be restored.

SET PEEPSILON [Value] sets the bin-width for groupings cases on the "pureerror" command (section 4.34). The default values is 1.E-09.

To change from weighted least squares to unweighted least squares, the command SET WEIGHTS 0 or SET WEIGHTS V0 should be used.

A typical command might be

```
      NEXT? SET WIDTH 132 TOL 1.E-6 WEIGHT 2 PLOTSIZE 51 31
WIDTH FOR PRINTING CHANGED TO 132
PIVOT TOLERANCE CHANGED TO    .100000E-05
VARIABLE V2   USED AS WEIGHTS.
PLOT SIZE:   51 COLUMNS BY  31 ROWS.
```

4.3 LABEL. Form: LABEL [list 1] AS [list 2]

This command is used to assign up to 4 character labels to variable numbers. Any group of 4 or fewer characters not recognizable as a number is a valid label, except for a few reserved words like VS, FOR, AS.

The elements in list 1 may be either variable numbers or old labels. Each variable has a null label of V1 for variable 1, V2 for variable 2, etc. After the READ command is executed all new labels are lost.

For our example, the appropriate label command for the example is:

```
NEXT? LABEL 1 2 3 4 5 AS X1 X2 X3 X4 Y
```

Any number of labels may be specified in a single command. Any labeled variable may be refered to by its number or by its label.

4.4 DELETE. Form: DELETE [list of case numbers].

This command is used to delete the specified cases from the current data set. One would wish to delete a case if it were suspected of being an outlier, or an influential case. A careful analyst would wish to compare the fit of a model both with and without the questionable case. The result of this command is to treat the specified cases as if they had zero case weight. Deleted cases are not used in the computation of estimates, R^2, degrees of freedom, etc, but predicted values, and standard errors of prediction produced by the RESID command are given for deleted cases.

The computational method used by the DELETE command is outlined in the discussion of the updating technique in Sec. 3.

Deleted observations may be restored in two ways. The usual method is to use the RESTORE command (Sec. 4.5). However, all deleted cases are restored when the SET WEIGHTS is used.

```
NEXT? DELETE 1 8 3
DELETED CASES ARE   1   3   8
```

If more than 25 cases are deleted, only the case numbers of the first 25 can be printed. The user should be aware that severe round off error may result from deleting many cases.

4.5 RESTORE. Form: RESTORE [case numbers]

This command will restore the specified cases (that have previously been DELETED) to the data set. If the list of case numbers is left off, all deleted cases are restored.

If a list of cases is specified, then the computations done by this command correspond to single steps of the updating methods given in section 3. If the list is left off, then the cross product matrix is recomputed from scratch. This is done to avoid the accumulation of roundoff errors.

```
      NEXT? RESTORE 1
REMAINING DELETED CASES ARE    3    8

      NEXT? RESTORE
ALL CASES RESTORED TO THE DATA SET.
```

Weighted least squares. If the SET WEIGHTS command had been used, the command RESTORE will also automatically reset to unweighted least squares.

4.6 LIST. Form: List [Var. list] or SAVE

This command has 2 functions. If the word "SAVE" does not appear on the command, all the data for the variables specified are printed; if the list is off, all the data is printed.

Example:

```
NEXT? LIST
   X1          X2          X3          X4          Y
 7.000       26.00       6.000       60.00       78.50
 1.000       29.00       15.00       52.00       74.30
 11.00       56.00       8.000       20.00       104.3
 11.00       31.00       8.000       47.00       87.60
 7.000       52.00       6.000       33.00       95.90
 11.00       55.00       9.000       22.00       109.2
 3.000       71.00       17.00       6.000       102.7
 1.000       31.00       22.00       44.00       72.50
 2.000       54.00       18.00       22.00       93.10
 21.00       47.00       4.000       26.00       115.9
 1.000       40.00       23.00       34.00       83.80
 11.00       66.00       9.000       12.00       113.3
 10.00       68.00       8.000       12.00       109.4
**MEAN**
 7.462       48.15       11.77       30.00       95.42
**STDEV**
 5.882       15.56       6.405       16.74       15.04

NEXT?
```

LIST SAVE. This command will result in the entire data set (all variables and cases) being printed as a formatted file on file SAVER. The data can then be read in at a later time by the FREAD (sec. 4.1) command.

Please note the following for use with LIST SAVE: (1) end your session with END, not STOP. If STOP is used, SAVER will be lost. (2) Rename SAVER (outside of MULTREG) via a "system" command like "RENAME, Newname = SAVER". (3) If LIST SAVE is entered several times, Multreg will only be able to retrieve the first set saved; however, MATTER will be able to retrieve any set saved. (4) The OUTPUT command (sec. 4.31) should not be used before LIST SAVE, or else SAVER will not be readable.

4.7 STAT. Form: STAT [variable list]

The basic statistics for the variables in the list are printed. If the list is missing, statistics are printed for all variables.

```
      NEXT? STAT
VARIABLE     N      MEAN        VARIANCE      ST.DEV.        MIN          MAX
  X1         13     7.462         34.60         5.882        1.000        21.00
  X2         13     48.15         242.1         15.56        26.00        71.00
  X3         13     11.77         41.03         6.405        4.000        23.00
  X4         13     30.00         280.2         16.74        6.000        60.00
  Y          13     95.42         226.3         15.04        72.50        115.9
```

4.8 CORRELATION. Form: CORR [variable list]

This command prints the correlations between all variables in the list. If the list is not given, the correlations between all the variables in the data set are printed. The rows are in the same order as the columns.

```
      NEXT? CORR X1 X2 X3 X4
 CORRELATION MATRIX
 X1     1.000
 X2      .2286        1.000
 X3    -.8241        -.1392         1.000
 X4    -.2454        -.9730         .2954E-01    1.000
         X1            X2             X3           X4
```

4.9 COVARIANCE. Form: COVA [variable list]

This command prints the sample covariances and variances of the variables in the list. If the list is not present, the default is to print the covariances between all the variables.

```
      NEXT? COVA
 COVARIANCE MATRIX
 X1     34.60
 X2     20.92        242.1
 X3    -31.05        -13.88         41.03
 X4    -24.17        -253.4         3.167        280.2
 Y      64.66         191.1        -51.52       -206.8        226.3
         X1            X2             X3           X4            Y
```

4.10 VARS. Form: VARS

This command is used to print the labels corresponding to variable numbers, and give the values of the parameters that can be set by the SET command (section 4.1), and a list of deleted cases. A typical example, showing all the default values for the parameters is

```
      NEXT? VARS
NUMBER      LABEL
   1        X1
   2        X2
   3        X3
   4        X4
   5        Y
UNWEIGHTED (ORDINARY) LEAST SQUARES.
PIVOT TOLERANCE =    .100E-03
```

4.11 MODEL. Form: MODEL [model specification]

Result: This command does not initiate any computations. It permits the user to input a model without specifying any computations or print out an existing model if no specification is made. It would be primarily of use when the model to be used is long, so that the model can be set on any command that requires a model (e.g., REGS, ANOVA).

```
      NEXT? MODEL 5 ON 1 2 3 4
DEP. VAR.    Y
INDEP. VARS.    X1    X2    X3    X4
```

4.12 REGS. Form: REGS [model]

Result: This command uses sweep as described in section 2 to obtain least squares estimates of the parameters in the model, their standard errors and corresponding t statistics. Additionally, R^2, the residual mean square and its square root and its degrees of freedom are printed. A typical command (without labels) is: REGS 4 ON 1 2 3. For regression through the origin, see section 1.4. If the set of regressions is linearly dependent (or colinear) (e.g., the tolerance test, Sec. 3, fails), a linearly dependent subset of maximal dimension will be used and appropriated error messages will be printed.

The regression of Y on X1 X2 X3 is given by

```
    NEXT? REGS Y ON X1 X2 X3
 REGS     Y     ON    X1    X2    X3
VARIABLE      COEF'T          ST. ERROR          T VALUE
  BO        48.19363          3.913305           12.32
  X1        1.695890          .2045820            8.29
  X2        .6569149          .4423423E-01       14.85
  X3        .2500176          .1847109            1.35
        DEGREES OF FREEDOM  =       9
        RESIDUAL MEAN SQUARE=   5.345624
        ROOT MEAN SQUARE    =   2.312061
        R-SQUARED           =      .9823
```

4.13 ANOVA. Form: ANOVA [model]

Result: An analysis of variance table is produced, with the independent variables entered in the order given in the model specification, from left to right. Cumulative R^2 are also given unless regression is through the origin. Note that ANOVA 1 ON 2 3 4 is different from ANOVA 1 ON 4 2 3.

See section 1.4 for a discussion of models and section 4.12 for a discussion of linear dependence.

Below we give two anova tables for two different models.

```
      NEXT? ANOVA Y ON X1 X2 X3 X4
  ANOVA    Y     ON    X1    X2    X3    X4
                 INDIVIDUAL                    CUMULATIVE
SOURCE     DF       SS          MS        DF       SS          MS          R**2
 MEAN       1   .1184E+06   .1184E+06
 X1         1   1450.       1450.          1   1450.       1450.       .5339
 X2         1   1208.       1208.          2   2658.       1329.       .9787
 X3         1   9.794       9.794          3   2668.       889.2       .9823
 X4         1   .2470       .2470          4   2668.       667.0       .9824
RESIDUAL    8   47.86       5.983         12   2716.       226.3

      NEXT? ANOVA Y ON X1 X3 X2
  ANOVA    Y     ON    X1    X3    X2
                 INDIVIDUAL                    CUMULATIVE
SOURCE     DF       SS          MS        DF       SS          MS          R**2
 MEAN       1   .1184E+06   .1184E+06
 X1         1   1450.       1450.          1   1450.       1450.       .5339
 X3         1   38.61       38.61          2   1489.       744.3       .5482
 X2         1   1179.       1179.          3   2668.       889.2       .9823
RESIDUAL    9   48.11       5.346         12   2716.       226.3
```

In both of the above tables, X1 is fit first, giving a sum of squares of 1450. From the first table, the sum of squares for X3, adjusting for X1 and X2 is 9.794, while from the second table, the sum of squares for X3 adjusting for X1 (but ignoring X2) is 38.61. Fitting variables in differing orders will result in differing results.

The columns marked "cumulative" in the ANOVA table give the cumulative sums of squares usually referred to as the "regression" sums of squares. For example, in the first table, the 4 d.f. sum of squares 2668. is the sum of squares for regression on X1, X2, X3, and X4; this number will be the

same regardless of order of fitting. The usual F-test for regression is, in this case, 667.0/5.983 = 111.48, with (4,8) degrees of freedom. (Multreg generally does not provide for automatic computation of F-tests, since the variety of possible F tests of interest is so great.) The sum of squares on the "residual" line under the cumulative heading is the total sum of squares (corrected for the mean).

4.14 SWEEP. Form: SWEEP [variable list]

Result: Sweep is the fundamental operation on MULTREG. It is used by all of the computation commands to obtain the results needed for least squares regression. This command allows the user to directly invoke the Sweep algorithm, as described in section 3 of this manual. The variable list gives the columns that are to be swept. If the variable list is excluded, e.g., the command is simply SWEEP, the corrected cross product matrix defined in section 2 is printed. SWEEP * will result in printing the uncorrected cross product matrix; SWEEP * (variable list) would be used as the command for regression through the origin.

Examples of SWEEP are given in section 3.

4.15 CSWEEP. Form: CSWEEP [variable list]

CSWEEP is a relatively complicated command, and is best illustrated by the output from the command, given in one example as follows:

```
      NEXT? CSWEEP X2 X3 4
CSWE       X2      X3      X4
X2      1.000
X3      .4786          1.000
X4      .9788          .4631          1.000
X1     -.3494         -.9177         -.3201  |      .2598E-01
Y      -.8927         -3.401         -1.399  |      .5929          .2718E-01
         X2             X3             X4    |       X1             Y
```

The output is really just a complicated rescaling of the output from SWEEP 2 3 4. The upper left corner of the output is the correlations between the estimated β-coefficients; in sweep, this area is a multiple of the covariances. The area in the box gives the standardized β's, that is the values of the β-coefficients with the variables all rescaled to have zero sample mean and unit sample variance; SWEEP gives the orginary β's in this area.

The area at the right gives, on the diagonal, the proportion of unexplained variability ($1 - R^2$). The off-diagonal terms are the partial correlations between the variables that are not dependent variables.

With a little work, most of the information in CSWEEP can be obtained by using other commands.

4.16 PREDICT. Form: PREDICT [model] FOR [values]

This command is used to obtain fitted values and/or predicted values that correspond to specified values of the independent variables in the model. Also, the standard errors of the fitted value and of the prediction are printed.

Referring to the mathematical notation for the linear model given briefly in Section 3, we consider the linear model $Y = X\beta + e$ (in matrix terms), with the covariance of e, and the estimator of β as given in section 3. Let $\underset{\sim}{x}$ correspond to the vector of values specified for the independent variables. The predicted value = the fitted value = $\underset{\sim}{x}'\hat{\beta}$. The estimated variance of the fitted value is given by $\hat{\sigma}^2\underset{\sim}{x}'(X'X)^{-1}\underset{\sim}{x}$, which we write as $\hat{\sigma}^2 v$. It is useful to think of v as a "distance" measure, since it measures, in a sense, the distance from the vector x to the center of the n cases used in the estimation. See section 6.1 for more on v. The

estimated variance of a predicted value is varpred = $\hat{\sigma}^2$ (1 + v), the "1" due to the additional variability that is inherent in the new observation.

Besides the prediction, fitted value, and their standard errors, the value of v is also printed out. This quantity is of special interest in prediction, as it gives a rough measure of the reliability of the prediction. If v is large, then the new values of the independent variables are not like the values used in estimation, and the predicted or fitted values from the model might be very poor. "Large" corresponds to being bigger than any of the v_{ii} in the RESID output; certainly, if $v \geq 1$, predictions will be unreliable. We refer to the large v case as being extrapolation.

On the other hand, if v is small, say about equal in size to the smallest value of the v_{ii} from the residual output, then we should expect predictions to be quite good, as we are attempting to predict in the same region as the original data. This is called interpolation.

An example of the command is given below.

```
      NEXT? PREDICT Y ON X1 X3 FOR 8 12
PREDICTION OR FITTED VALUE =    96.7824
STD. ERROR(FITTED VALUE)   =    3.14972
STD. ERROR(PREDICTION)     =    11.5164
V = V(FIT. VAL)/RES. M.S.  =    .808490E-01
V/(1.-V)                   =    .879605E-01

      NEXT? PREDICT Y ON X1 X3 FOR 8 80
PREDICTION OR FITTED VALUE =    130.406
STD. ERROR(FITTED VALUE)   =    60.6464
STD. ERROR(PREDICTION)     =    61.6498
V = V(FIT. VAL)/RES. M.S.  =    29.9737
```

4.17 PARCOR. Form: PARCOR [model]

This command prints the partial correlations of the dependent variable with the variables not in the model controlling for those in the model. It is useful in so-called stepwise "step-up" regressions.

```
      NEXT? PARCOR 5 ON 4
PARTIAL CORRELATIONS WITH Y
   X1            X2            X3
 .9568         .1302        -.8951

      NEXT? PARCOR 5 ON X4 X1
PARTIAL CORRELATIONS WITH Y
   X2            X3
 .5986        -.5657
```

The first command gives the partial correlation of Y with X1, and X2 and X3 adjusted for X4; the second gives the partial correlations adjusted for X4 and X1.

4.18 TRAN. Form: TRAN [Var1] = [type][variables][parameters]

The transformation command is used to create new variables from the old ones already in the data set. A typical command might look like

TRAN V3 = POWER V1 2 .5

which is read as instructing the program to take old variable V1, add .5 to each value of the variable, square the result, and store these values in variable V3. Each of the specifications that are used is explained below.

Var1. This is the variable number or label of the variable that is to be replaced or created by the transformed values. Var1 may correspond to any variable currently in the data set except for the column of weights if weighted least squares is being used. Alternatively, Var1 may be the number of the first unused variable in the data. For example, if the data currently has 10 columns or variables, then the command TRAN V11 = LOG10 V10 would create a new variable, V11, whose values are the logarithms (to the

base 10) of the data in V10.

If Var1 is left off, then the equals sign must also be left off. The destination is then assumed to be the first variable in the variable list following the "type" specification. For example, TRAN SQRT V3 would replace V3 by its square root.

Type. The type is the name of the transformation. A complete list of types is given in the table on the next page. (Note that some of the types listed in the table are not yet available, but should be added in the near future. Type TRAN HELP for a complete list of available types.) The names of the types are meant to convey their function, e.g., SQRT, POWER, PRODUCT, etc. The type specification is required.

Variables. This is the list of variables to be transformed. Most of the transformations only require one variable in this list, but some (SUM, DIFF, PROD, etc.) require two. The values of the variables in the list will not be damaged except as noted under "Var1" above.

Parameters. From zero to two parameters (e.g., numerical values) are required to perform the transformations. In the table, any parameter in square brackets, e.g., [C], may be left off, and it will assume the default value shown in the table. All two parameter transformations require one parameter, and the second is optional; all one parameter transformations have the parameter optional. Thus, the following two commands are identical: TRAN 2 = POWER 2 .33, and TRAN 2 = POWER 2 .33 0.

TABLE 4.18

Keyword	Number of Variables (required)	Parameters	Results	Comments
POWER	1	P [C]	$(V + C)^P$	If P=0 then the transformati is $\log_e (V+C)$. If P is not an integer, then V+C>0 for all cases.
SQRT	1	[C]	$\sqrt{V+C}$	V+C>0
LN	1	[C]	$\log_e(V+C)$	V+C>0
LOG10	1	[C]	$\log(V+C)$	V+C>0
EXP	1	[C]	$\exp(V+C)$	$\lvert V+C \rvert < 385$
EXP10*	1	[C]	$10^{(V+C)}$	
LINEAR	1	A [C]	A*V+C	
SCALE	1	A [C]	(V-C)/A	$A \neq 0$
SUM	2	None	V1+V2	
DIFF	2	None	V1-V2	
PROD	2	None	V1*V2	
RATIO*	2	None	V1/V2	
ANGLIT*	1	[B]	$\text{SIN}^{-1}\sqrt{V1/B}$	$0 \leq V1/B \leq 1$
LOGIT*	1	[B]	$\log_e((1-V/B)/(V/B))$	
FT*	1	None	$\sqrt{V} + \sqrt{V+1}$	

*Not currently available (9/1/77)

[C] has default value of C = 0

[B] has default value of B = 1

Complex transformations. Suppose that we wanted to create a variable that has its value equal to $V2^2/(V3 + \exp(V2))$. We might use the following commands

```
NEXT? TRAN V4 = EXP V2          (so that V4 = exp (V2))
NEXT? TRAN V4 = SUM V3 V4       (now, V4 = V2 + exp(V4))
NEXT? TRAN V4 = POWER V4 -1     (now, V4 = 1/(V2 + exp (V4))
NEXT? TRAN V5 = POWER V2  2     (V5 = V2^2)
NEXT? TRAN V4 = PRCD V4 V5      (the desired result)
```

4.19 PLOT. Form: PLOT [model] VS [list for x-axis variables] [keyword]

This routine produces a "6-line plot" of the studentized residuals from the specified model (on the y-axis), plotted against (on the x-axis) any variables in the data set (specified by its number or label), or plotted against observation number (specify Ø, zero), or predicted values from the regression (specify PRED), or the ordered residuals are plotted in a normal or rankit plot (specify RANKIT). Several plots of the same residuals may be specified by giving a list of variables for the x-axis; e.g.

PLOT 4 ON 2 3 VS Ø PREDICT RANKIT 2

would be a typical command, giving a total of four plots.
If the model specified is a single variable (e.g. PLOT 4 VS 2), the numbers plotted on the y-axis are found by the transformation $(V_4 - \overline{V}_4)/\sqrt{\text{Var}(V_4)}$, that is, they are the standardized values. In this case, plot against PRED is not applicable. The keyword is used only if there are deleted or zero weight cases on the file.

Each plot produced is computed by an algorithm by Andrews and Tukey [1]. In addition, RANKIT plots also result in printing of the Wilk-Shapiro W statistic, which is a test statistic for non-normality. (See section 6.1 for details.)

At first glance, the 6-line plots appear to be very difficult to read and to interpret; however, with some practice they can become a powerful tool in exploratory analysis. They are used in this program because of their speed in printing: with them, it is possible to produce several plots in the time needed to produce a single scatter diagram. Recall that in most instances we will be plotting residuals (actually, studentized residuals) on the y-axis. Most of the information contained in residuals come from looking at large (positive or negative) residuals and by looking at patterns in the residuals. This information is completely retained in 6-line plots.

The six lines in the plot are labeled as +TWO +ONE +ZERO and -ZERO, -ONE -TWO. The idea is to put large residuals on the top and bottom lines of the plot (e.g. the +TWO and -TWO lines respectively), while the less large residuals will go on the intermediate lines on the graph. Here, the use of standardized residuals plays an important part: no matter what the scale of the data is, the meaning of "large" for standardized residuals is always the same: the neighborhood of +2 or bigger or -2 or smaller is a minimal criteria for "large."

A simple further refinement that could be used in the plots is to make a clever choice of what symbol to actually plot. For example, suppose we wanted to plot the value 1.3 on the y-axis. One way of indicating this would be to plot the symbol "3" on the line +ONE; we could then have the rule that

would tell us to combine this to reconstruct the number 1.3. The problem with this is that we are unable to handle numbers bigger than 3 or smaller than -3 on a six line plot (since 3.1 would want to be plotted on the +THREE line, if we had one). This is a vexing problem since we would usually be especially interested in values of the standardized residuals that are outside the range (-3, 3). Andrews and Tukey (1973) came up with a solution to this problem: for the plotting symbol, instead of plotting the number of tenths in the number on the line name, plot the number of quarters on the line name. Thus, for example, 1.3 would be thought of as 1 and 1/4 (plus a little extra); we could plot this as a 1 on the line +ONE. Admittedly, this technique would lose a little information, at least on the residuals that are not large. However, on the large residuals, a great deal of information is saved: for example, a residual of -3.8 can be thought of as being approximately -2 and 7/4, so that we could plot a character 7 on the line -TWO. In this way, our 6-line plot can handle any residuals in the range -4.25 to +4.25; numbers outside this range would be coded as 9 on the top or bottom line, depending on sign. In almost all examples, this range is adequate for plotting the residuals.

The following table may be useful in interpreting the plotted values:

LINE NAME	Plotted Character 0	1	2	3	4	5	6	7	8	9
+TWO	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00	4.25	+ ∞
+ONE	1.00	1.25	1.50	1.75	2.00						
+	.00	.25	.50	.75	1.00						
-	-.00	-.25	-.50	-.75	-1.00						
-ONE	-1.00	-1.25	-1.50	-1.75	-2.00						
-TWO	-2.00	-2.25	-2.50	-2.75	-3.00	-3.25	-3.50	-3.75	-4.00	-4.25	- ∞

Break points for values corresponding to plotted characters in a 6-line plot. (e.g., a "2" on the line "-ONE" represents a value less than or equal to -1.50 but greater than -1.75.)

Note: On plots with a large number of observations, it is quite likely that several of the observations should be plotted at the same location on the plot. From the above, it is clear that no provision is available for multiple observations. The algorithm will currently print only the last character to fall on a specific print position; hence, for example, if the values +TWO 9 and +TWO 1 were to be printed at the same location, it is possible that either of the characters might end up being printed, depending on the order of the data in the file. Clearly, in a case such as this, a scatter plot would be desirable. See section 4.11.

Deleted or zero weight cases. Cases that are not used in computations are not ordinarily plotted. This default can be modified by putting a keyword on the control card. The keywords are ALL, DELETE, INCLUDED, where ALL means plot all cases, DELETE means plot only deleted and zero weight cases, and INCLUDED means plot only included cases (the default). A typical command might be PLOT 4 ON 3 VS PRED ALL. The keyword must be the last item on the command.

Two typical plots are given below.

```
      NEXT? PLOT Y ON X1 X4 VS PRED RANKIT

PLOT:    X-AXIS = PREDICTED VALUE
         Y-AXIS = STANDARDIZED RESIDUALS OF Y    ON    X1    X4
+TWO
+ONE                                  0                1
   +  2    3          0             1                          2
   -                                                1    3    1          3
-ONE                          0
-TWO          0
         ^.........^.........^.........^.........^.........^.........^.....
LOBOUNDS=    72.61     /UPBOUNDS=    117.4     /INCRE.=    8.000

PLOT:    X-AXIS = RANKITS
         Y-AXIS = STANDARDIZED RESIDUALS OF Y    ON    X1    X4
+TWO
+ONE                                                 0      1
   +                             0  1  2   2   3
   -                3   3   1  1
-ONE           0
-TWO   0
         ^.........^.........0.........^.........^.
LOBOUNDS=   -1.673     /UPBOUNDS=    1.673     /INCRE.=    .8000
          APPROX. WILK-SHAPIRO W = .9714
```

4.20 SCPLOT. Form: SCPLOT [model] VS [list for the x-axis] [keyword].

The control for this command is identical to that for the 6-line plots, sec. 4.19. This command produces scatter plots of about 13 cm (5 inches) on a side. The size of the plot can be altered by use of the SET PLOTSIZE command (sec. 4.2). The default plotsize of 51 (columns) by 31 (rows) is a good size for interactive computing using timesharing. A plot the size of a standard sheet of typewriter paper can be obtained by the command SET PLOTSIZE 61 52. A plot the size of a sheet of computer paper is 112, or the width of paper set by the SET WIDTH command, whichever is smaller.

Unlike the six line plots, the plotted character refers to the number of points plotted rather than their magnitude. A plotted * means that one value is plotted at the point, 2 through 9 mean, respectively 2 to 9 points. Letters are then used, with A, B, C, ... , Y corresponding to 10, 11 12, ..., 35 points. A plotted Z means 36 or more points.

Deleted or zero weight cases. Cases that are not used in computations are not ordinarily plotted. This default can be modified by putting a keyword at the end of a SCPLOT command. The keywords are ALL (plot all cases), DELETE

(plot only deleted or zero weight cases), and INCLUDE (plot only cases with positive weight that have not been deleted, the default). A typical command, to plot all the cases, might be SCPLOT 4 ON 2 3 VS PRED ALL.

Plots of variables against variables. In a scatter plot of two variables, neither will be standardized as in the 6-line plot, but rather they will be plotted in their natural scale.

The typical plot below is equivalent to the first 6-line plot given in the last section.

```
      NEXT? SCPLOT Y ON X1 X4 VS PRED

SCPLOT: X-AXIS = PREDICTED VALUE
        Y-AXIS = STANDARDIZED RESIDUALS OF Y    ON   X1   X4
+.......^.........^.........^.........^.........^.........^...
.
.
.
.
.
.
.                                 *
.
.                     *
.    *
.
.*                                          *
.
.
.                     *
.
0-----------*-----------------------------------------------
.
.                                       *
.                                  *
.
.                                             *
.                                    *
.                   *
.
.
.
.
.
.      *
+.......^.........^.........^.........^.........^.........^...

X AXIS: 1-ST TICK =     80.00     /INCREMENT =     10.00
  DATA: MINIMUM   =     72.61     /MAXIMUM   =     117.4
Y AXIS: 1-ST TICK =    -1.500     /INCREMENT =     1.500
  DATA: MINIMUM   =    -2.062     /MAXIMUM   =     1.467
```

The legend at the bottom of the plot should be interpreted as follows. The value at the first tick is the value along the axis at the leftmost (or bottom) pointer; the increment is the amount of increase between tick marks. The maximum and minimum of the plotted variables are also given.

4.21 RESIDUALS. Form: RESID [model]

This command is used to compute the "case analysis" statistics for the specified regression linear model. These statistics are used to judge the adequacy of fit of the model, and the influence of individual cases on the estimates, etc. The output from this command will depend upon the value of the width for printing chosen by the SET WIDTH command. If the width is greater than 120 (but less than 132), then the output from both the RESID and the YHAT commands are produced by the RESID command. If the width is less than 120, the output is necessarily split into two parts. In this section we describe the RESIDUAL output, and discuss YHAT in the next section. Typical (narrow) output from the residual command is:

```
      NEXT? RESIDUALS Y ON 1 2 3 4
  RESID     Y      ON    X1     X2     X3     X4
CASE        Y            RESIDUAL   STUD. RES        V          DISTANCE       T
   1     78.50          .4760E-02     .0029      .5503          .0000        .00
   2     74.30           1.511         .7566      .3332          .0572        .73
   3     104.3          -1.671       -1.0503      .5769          .3009      -1.06
   4     87.60          -1.727        -.8411      .2952          .0593       -.82
   5     95.90           .2508         .1279      .3576          .0018        .12
   6     109.2           3.925        1.7148      .1242          .0834       2.02
   7     102.7          -1.449        -.7445      .3671          .0643       -.72
   8     72.50          -3.175       -1.6878      .4085          .3935      -1.97
   9     93.10           1.378         .6708      .2943          .0375        .65
  10     115.9           .2815         .2103      .7004          .0207        .20
  11     83.80           1.991        1.0739      .4255          .1708       1.09
  12     113.3           .9730         .4634      .2630          .0153        .44
  13     109.4          -2.294       -1.1241      .3037          .1102      -1.15
DURBIN-WATSON=     2.0526
RESIDUAL SS   =  47.86363935
PRESS         =  110.3465569
```

Referring to the mathematical notation in section 3, all of the values in this printout can be defined. The first column is the case number. Following this, a * will be printed for any case that is deleted, or has zero weight. The next column gives the value of the dependent variable, y_i. Next the residual = $\hat{e}_i$ = (observed) - (fit) = $y_i - x_i'\hat{\beta}$ is printed. The next column contains the value of the Studentized residual, τ_i, which is defined by

$$\tau_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-v_{ii}}}$$

where $\hat{\sigma}^2$ is the residual mean square, and $v_{ii} = x_i'(X'X)^{-1}x_i$ is a measure of the similarity of x_i to the other vectors in the data set. Also, $1 - v_{ii}$ is the known function of the x's so that $\text{Var}(\hat{e}_i) = \sigma^2(1 - v_{ii})$. The Studentized residuals have zero mean and unit variance (however, they are not distributed as Student's t, although a transformation of τ_i is, as is given below).

The next column gives the value of v_{ii}, which, as pointed out measures how unusual the vector x_i' is relative to the other rows of the matrix X. v_{ii} is related to the Mahalanobis distance MD_i by the equation $MD_i = (n-1)(v_{ii}$-1

The next column gives the value of D_i, Cook's (1977) distance measure, which is defined from the data. D_i is computed from the equation

$$D_i = \frac{\tau_i^2}{k'} \frac{v_{ii}}{1-v_{ii}}$$

where k' is the number of parameters in the model. D_i is discussed in section 6.1.

The final column in this output gives the transformation of τ_i that is distributed as Student's t. This t statistic is equivalent to the test that would be computed if the i-th case were deleted from the model, and then were tested to see if the observed value y_i was too far away from the value $\hat{y}_i$ predicted from the remaining n - 1 cases. It has n-k'-1 degrees of freedom, where k' is the number of estimated parameters. See section 6.1 for more information.

Summary statistics. At the foot of the residual output some summary statistics are computed and printed. The residual sum of squares is computed by actually squaring and adding up the residuals. This is done to provide a check on the numerical accuracy of the computations. The computation is done separately for cases that are deleted or have zero case weights. The value of PRESS, the predicted residual sum of squares, is explained in the next section. The last value computed is the Durbin-Watson statistic for serial correlation, and is defined by the equation

$$\text{Durbin-Watson statistic} = \frac{\Sigma\ (e_{i+1} - e_i)^2}{\Sigma\ e_i^2}$$

(if weighted least squares is used, then in the above equation replace all $\hat{e}_i$ by $\sqrt{w_i}\ (\hat{e}_i)$ (and similarly for e_{i+1}).) The Durbin-Watson statistic is a measure of serial correlation in the residuals, and is a useful measure if the order of the cases in the file is meaningful; for example, if the cases are ordered in time. If this is not the case, then the Durbin-Watson statistic does not contain useful information.

Small values of the Durbin-Watson statistic are indicative of autocorrelation among the residuals; values of the statistic of about 2 or greater are usually considered to be large. The use of this statistic is a relatively complex problem, and is not discussed in this manual. Both a discussion of the statistic, and tables for it, are given by H. Theil, (1971), page 199.

Deleted or zero weight cases. If a case is deleted or has zero weight, then some of the statistics computed by RESID do not apply. For these cases, the following are printed: for the residual, we get the value of y_i minus the predicted value based on the cases with positive weight. If only one case is deleted, this is equivalent to the "predicted residual" described in the next section. The value in the "V" column is the distance from the deleted case to the center of the remaining cases; if only one case is deleted, then the value in this column will be the same as the value for this case in the column $v_{ii}/(1-v_{ii})$ in the YHAT command before deletion. The value in the T column is the t-test to see if this case is adequately described by the model fit to the rest of the data. If only one case is deleted, the t value for the deleted case should equal the t value for this case before deletion.

If any cases are deleted or have zero weight, the Durbin-Watson statistic is not computed.

Most of the output produced by this command is not generally available elsewhere, and few general texts cover their use. A modest treatment on the use of them is given in Section 6.1.

4.22 YHAT. Form: YHAT [model]

This command is a continuation of the RESIDUAL command when the width for printing is less than 120. Typical output is shown below

```
     NEXT? YHAT 5 ON 1 2 3 4
  YHAT     Y     ON    X1    X2    X3    X4
CASE     Y            YHAT     PRED RES     SE(YHAT)     SE(PRED)     V/(1-V)
   1    78.50        78.50       .1059E-01    1.814        3.046        1.224
   2    74.30        72.79        2.266       1.412        2.824        .4998
   3    104.3        106.0       -3.950       1.858        3.072        1.364
   4    87.60        89.33       -2.451       1.329        2.784        .4189
   5    95.90        95.65        .3903       1.463        2.850        .5567
   6    109.2        105.3        4.482       .8619        2.593        .1418
   7    102.7        104.1       -2.289       1.482        2.860        .5800
   8    72.50        75.67       -5.368       1.563        2.903        .6907
   9    93.10        91.72        1.953       1.327        2.783        .4170
  10    115.9        115.6        .9398       2.047        3.190        2.338
  11    83.80        81.81        3.466       1.596        2.920        .7407
  12    113.3        112.3        1.320       1.254        2.749        .3568
  13    109.4        111.7       -3.295       1.348        2.793        .4362
DURBIN-WATSON=      2.0526
RESIDUAL SS   =   47.86363935
PRESS         =   110.3465569
```

For cases with positive case weights, the values printed are defined as follows. First, the case number and value of the dependent variable are printed. A * following the case number indicates that the case has zero weight or has been deleted from computations by the DELETE command. The fitted values $\hat{y}_i$ are then printed (YHAT). The next column is called the predicted residual, and is defined by

$$\text{PRED. RES}_i = \hat{e}_i/(1-v_{ii})$$

The predicted residual is the difference between y_i and the predicted value for y_i obtained <u>without</u> the inclusion of the i-th case in the data for estimation. The next two columns given the standard errors of fitted and predicted values with independent variable vectors equal to the values for the i-th cases. The value $v_{ii}/(1-v_{ii})$ gives the distance from the i-th case to the center of the remaining cases in the data set.

For deleted or zero weight cases, the values printed are the same as they are for the included cases.

The predicted residual sum of squares (PRESS) is defined by (Allen (1971))

$$\text{PRESS} = \Sigma\ (\text{PRED. RES}_i)^2$$

In a sense, PRESS measures how successful the regression equation is at predicting future values, since it is the sum of squared differences between the observed values and the fitted values based on all the other cases.

The use of YHAT is described in section 6.1.

4.23 ALL. Form: ALL [model] FORCING [list1] OMIT [list2] CPMAX [val1] CPMIN [val2]

This command will compute all possible regressions of the dependent variable on subsets of the independent variables in the specified model. For each regression, 4 summary statistics will be computed: C_p, R^2, Adjusted R^2, and the residual sum of squares (these statistics are defined in section 6.2). The parameters in the control language are used to limit the amount of output printed. HOWEVER, IT IS RECOMMENDED THAT THE NEW COMMAND SCREEN BE USED IN PLACE OF ALL IF THERE ARE MORE THAN 6 OR 7 INDEPENDENT VARIABLES. SCREEN is faster, somewhat easier to use, and produces managable output; see the next section.

Keyword FORCING. The list following this keyword gives the labels or numbers of variables in the model that are to be forced into every equation. Each variable forced reduces the computations done by one half.

Keyword OMIT. The list following this keyword gives the labels or numbers of all variables in the model that are to be excluded from all regression equations. However, these variables will be used in the computation of C_p as described in section 6.2.

Keywords CPMAX and CPMIN. Models will be printed only if the computed value of C_p is less than CPMAX but greater than CPMIN. If not specified, CPMAX = 500, CPMIN = -100. If the number of true independent variables exceeds 6, CPMAX must be specified. A suggested value for CPMAX = k, where k is the number of independent variables in the full model. The parameter CPMIN was made available to encourage users to set CPMAX as small as feasible. If the chosen value is then found to be too small, then the user can ask for models between the old value of CPMAX and some new value. For example, if the user enters ALL CPMAX 5, and then decides to ask for models with C_p less than 10, one could ask for ALL CPMIN 5 CPMAX 10, to print out all remaining values.

Saving output. Output from ALL is potentially quite long. The user should consult section 4.31 for techniques of having output sent to a file.

A complete discussion of this command, is given in section 6.2.

```
      NEXT? ALL 5 ON 1 2 3 4
ALL REGRESSIONS WITH  Y    DEPENDENT
MODELS WITH CP LESS THAN  500.0
 P      C(P)     R2ADJ     R**2      RSS          VARIABLES
 1    442.917         0        0  2716.      0
 2    202.549    .4916    .5339  1266.      1
 3      2.678    .9744    .9787  57.90      1  2
 2    142.486    .6359    .6663  906.3      2
 3     62.438    .8164    .8470  415.4      2  3
 4      3.041    .9764    .9823  48.11      2  3  1
 3    198.095    .4578    .5482  1227.      3  1
 2    315.154    .2210    .2859  1939.      3
 3     22.373    .9223    .9353  175.7      3  4
 4      3.497    .9750    .9813  50.84      3  4  1
 5      5.000    .9736    .9824  47.86      3  4  1  2
 4      7.337    .9638    .9728  73.81      3  4  2
 3    138.226    .6161    .6801  868.9      4  2
 4      3.018    .9764    .9823  47.97      4  2  1
 3      5.496    .9670    .9725  74.76      4  1
 2    138.731    .6450    .6745  883.9      4
TIME USED IS     .087 SECONDS.
```

4.24 SCREEN. Form: SCREEN [MODEL] FORCING [LIST1] OMIT [LIST2]
MBEST [number] METHOD [type] [PRINT]

This command gives an alternative to the ALL command when the number of variables in a model that are neither FORCEd or OMITted is large. A very efficient algorithm is used that will compute only a fraction of all possible regressions, and then print either the MBEST regression equations, or the MBEST of each subset size, depending on the METHOD used.

The algorithm was written by G. M. Furnival and R. W. Wilson of the School of Forestry, Yale University. The identical code is also used in several generally available places, notably in the program BMDP9R (Dixon(1975)), and in the IMSL library of subroutines (subroutine RLEAP). See Furnival and Wilson (1974) for details on the working of the algorithm.

Keywords FORCing and OMITting. The use of these keywords is identical to their use in the ALL command (section 4.23).

Keyword MBEST. The program will compute and print the MBEST regression equations. If MBEST is not specified, then the default value is 5. The

maximum value is 10. If the METHod R2 is chosen, then the MBEST regressions of each subset size are printed.

Keyword METHOD. This keyword determines the criterion function to be used for defining the "best" equations. The methods available are CP, ADJR2, and R2 where CP means the C_p statistic, ADJR2 is the adjusted R^2, and R2 is the ordinary R^2. The use of these statistics is discussed in section 6.2. METHod R2 will result in much more printing and more computing, and hence will be more expensive than the other methods.

Keyword PRINT. If this keyword is included in the command, then, in addition to the MBEST models, the summary statistics for all models examined by the algorithm will be printed. This additional output is rarely of interest. The default of PRINT is to only print the MBEST models. In the output, a * indicates one of the MBEST models.

Restrictions. The only restriction in the use of this command is that the number of variables that are in the model after omitting and forcing must be less than 21 but at least 3.

Output. The output from SCREEN is similar to the output from ALL. For each model, one line of information is given, with the values of the summary statistics, and the numbers of the independent variables in the model. An example of the output, with all options, is on the next page.

Saving Output. Output from SCREEN can be saved on file SAVER. See section 4.31 for details.

```
      NEXT? SCREEN Y ON 1 2 3 4 MBEST 5 METHOD CP PRINT
BEST   5 REGRESSION WITH Y    DEPENDENT, USING CP
 P        C(P)    R2(ADJ)    R**2     RSS        VARIABLES
 2      138.731    .6450    .6745   883.9        4
 2      142.486    .6359    .6663   906.3        2
 2      202.549    .4916    .5339   1266.        1
 2      315.154    .2210    .2859   1939.        3
 3   *    2.678    .9744    .9787   57.90        1  2
 3        5.496    .9670    .9725   74.76        1  4
 3       22.373    .9223    .9353   175.7        3  4
 3      138.226    .6161    .6801   868.9        2  4
 3      198.095    .4578    .5482   1227.        1  3
 4   *    3.018    .9764    .9823   47.97        1  2  4
 4   *    3.041    .9764    .9823   48.11        1  2  3
 4   *    3.497    .9750    .9813   50.84        1  3  4
 4        7.337    .9638    .9728   73.81        2  3  4
 5   *    5.000    .9736    .9824   47.86        1  2  3  4
CP TIME USED IS    .097 SECONDS.
```

4.25 KEEP. Form: KEEP [model] [statistic] AS [variable name]

This command is used to save some of the output from the RESID or YHAT commands as variables in the data set. For example, the command KEEP 4 ON 2 1 RESID AS V3 would save the residuals from the regression of V4 on V2 V1 and replace the values of V3 by the residuals. The variable to receive the values that are kept may either be any existing variable, or the first unused column as in the TRAN command (section 4.18). The list of statistics that can be stored via the KEEP command is:

Keyword	Statistic
PRED	Predicted Values
RESID	Residuals (regular, not studentized)
STUDRES	Studentized residuals
PRERES	Predicted residuals
DISTANCE	Cook's distance measure
V	Diagonal elements of $X(X'X)^{-1}X'$

Limitations. At most two statistics may be saved on one KEEP command; thus KEEP RESID AS 3 PRED AS 4 is permitted, but KEEP RESID AS 3 PRED AS 4 V AS 5 is not.

```
     NEXT? KEEP Y ON X1 X4 RESID AS 6 PRED AS 7
RESIDUALS WRITTEN AS V6
PREDICTED WRITTEN AS V7

     NEXT? STAT
```

VARIABLE	N	MEAN	VARIANCE	ST.DEV.	MIN	MAX
X1	13	7.462	34.60	5.882	1.000	21.00
X2	13	48.15	242.1	15.56	26.00	71.00
X3	13	11.77	41.03	6.405	4.000	23.00
X4	13	30.00	280.2	16.74	6.000	60.00
Y	13	95.42	226.3	15.04	72.50	115.9
V6	13	.1644E-11	6.230	2.496	-5.023	3.770
V7	13	95.42	220.1	14.84	72.61	117.4

4.26 EIGEN. Form: EIGEN [variable list] [VECTORS] [CORR]

This command is used to print the Eigenvalues and, if requested, the Eigenvectors of the corrected cross product matrix (or, if requested, the correlation matrix) of the variables specified in the list. If the list is left off, the eigenvalues of the whole data set are computed; for regression applications, one would usually want the eigenvalues for the set of independent variables only.

For the Hald data, this command is given by

```
      NEXT? EIGEN X1 X2 X3 X4 VECTORS
    EIGENVALUES
       2.846        148.9        810.0        6214.
PCT.     .04         2.07        11.29        86.60

    EIGENVECTORS
X1      .5062       -.5673        .6460      -.6780E-01
X2      .4933        .5440        .1999E-01  -.6785
X3      .5156       -.4036       -.7553       .2902E-01
X4      .4844        .4684        .1085       .7309

      NEXT? EIGEN X1 X2 X3 X4 O_CORR VECTORS
    EIGENVALUES
       .1624E-02    .1866        1.576        2.236
PCT.     .04         4.67        39.40        55.89

    EIGENVECTORS
X1      .2411        .6755        .5090        .4760
X2      .6418       -.3144       -.4139        .5639
X3      .2685        .6377       -.6050       -.3941
X4      .6767       -.1954        .4512       -.5479
```

If the word "VECTORS" did not appear on the command, only the eigenvalues would have been computed. If CORR appears on the command, the correlation matrix is used in the computations.

The line of output immediately under the eigenvalues gives each eigenvalue as a percent of the sum of the eigenvalues.

Computations. The computations in this command are done by the subroutine RS, part of the EISPACK library, written at the Argonne National Laboratories.

4.27 PRINCOMP. Form: PRINCOMP [list] SAVE [numbers] AS [variable numbers]

This command is not available.

4.28 END

This is the usual command for termination of the program.

4.29 HELP.

This command results in the printing of a brief description of all the available commands. In addition, HELP may be used as a modifier for any command. For example, REGS HELP would get information concerning the regression command. HELP may not be available on all systems.

4.30 MESSAGE

This command allows communication between the user and the writers of the program. Recent information concerning changes in the program will be available here. MESSAGE will be updated periodically.

4.31 OUTPUT. Form: OUTPUT SAVE [ON OR OFF]

This command is used to control printing. If SAVE is ON, the results of certain commands are output on local file SAVER exactly as they appear at the teletype (including headings); the default for SAVE is OFF.

If PRINT is OFF, output from certain commands, except for error messages, is not printed at the user's terminal. The default of PRINT is ON.

For example, suppose a user wanted to save residuals from a regression, but, because there are many observations, the user does not want the output printed. The following commands are then appropriate.

```
NEXT? OUTPUT SAVE ON PRINT OFF
OUTPUT FROM ALL AND RESID PRINTED ON FILE SAVER.
```

4.32 CPTIME

This command will print the amount of central processor time used since the last "READ" command.

4.33 $COMMENT

Any input line beginning with a "$" will be ignored by the program. This command is used primarily to allow annotation of output for future reference.

```
NEXT? $ANY INPUT BEGINNING WITH A "$" IS IGNORED.
NEXT?
```

4.34 PUREERROR [model] [USING Keyword] [GROUPS]

This command used to compute the sum of squares due to pure error (replication of X values) for the regression model specified. If the word "GROUPS" appears at the end of the command, then the sum of squares within each group of replicated cases is printed; if groups is left off, then only the sum of squares for pure error is printed.

The keyword "USING" determines the method of comupting this sum of squares. As a default, cases are grouped together if a random linear combination of the X's in the model agrees within a limit, set by the parameter "PEEPSILON", set by the "SET PEEPS" command (the default value is peepsilon = 1.E.09). If the phrase "USING YHAT" appears on the command, then cases are groups by values of YHAT. This will usually give the same results as the default method for simple linear regression, but for more complicated models, this may result in a pseudo-pure error being computed, since different sets of X-values can lead to the same value of YHAT.

Finally, cases can be grouped using any variable, or subset of variables in the data set by typing, for example, "USING V3 V6 V1", which would group cases on the basis of values of V1 V3 V6.

Examples:

PUREERROR 5 ON 1 2 GROUPS (Compute the sum of squares for pure error for 5 on 1 2, printing the within groups sum of squares)

PUREERROR 5 ON 1 2 USING YHAT GROUPS (as above, except cases are grouped by values of the predicted values yhat, not be values of the variables V1 V2).

SET PEEPSILON 1.5
PUREERROR 5 ON 1 2 USING X2 (Groups cases together if the difference between the X2 values for any two cases is less than or equal to 1.5)

4.35 RPCPLOT [MODEL] vs [VARIABLE LIST]

The result of this command is to produce scatter plots. On the X axis, the values of the specified independent variable are plotted. The Y-axis has values of the "residuals plus components" plotted. These values are defined to be the sum of the residual for the model that is fit plus a component for the variable plotted on the X-axis (namely, that variable times its estimated slope from the regression in the model specified. This plot is useful in studying the relationship between the variable on the X axis and the independent variable after adjusting for all other variables in the model. These plots are also called "partial residual plots". For further discussion, see W. A. Larsen, and S. A. McCleary (1972), "The use of partial residual plots in regression analysis", Technometrics 14 781-90 or F. Wood (1973), "The use of individual effects and residuals in fitting equations to data", Technometrics 15, 677-95.

5. Accessing the program

In this section we discuss the method of accessing the program on the CDC computers at the University of Minnesota. On all systems, before the porgram is run, a local file containing the data to be used must be available, usually by a GET, OLD, or NEW command. We describe Meritss, Mirje, and Batch input on the CYBER 74 separately. Instructions for use of the program on the MECC CYBER 73 are not available at this time.

5.1 Meritss. Access on Meritss is the simplest of all the systems. The user needs to enter the command

```
X,DO,MULTREG
```

and the program will be automatically loaded and run.

5.2 Mirje. Mirje is the timesharing network on the CYBER 74 and 172 computers. To run Multreg, use the following control cards:

```
BATCH,45000
/FETCH,MULTREG
/MULTREG
```

5.3 Batch input on the CYBER 74. Multreg is primarily an interactive program, therefore, batch users (that is, users with input from cards rather than from a terminal) may encounter some difficulties with its use. The primary problem is that, when Multreg detects an error, a message is sent to output (that is, to a terminal), whereas on batch computing, it would be

better to terminate the program when a error is encountered. It is therefore important that the user carefully edit the job input to Multreg. Most important of all is to be sure that the READ or FREAD command is correct, as no computing can be done if this command is not properly read.

We will distinguish two cases in batch input: the data is on cards, and the data is on a (disk) file, called, for example, DATAFIL.

Data on disk. If the data is on disk, the following job is recommended:

```
JOB,T5,CM45000.
ACCOUNT,ACCTNO,PASSWORD.
GET,DATAFIL.
FETCH,MULTREG.
MULTREG.
COST.
  --EOR--
SET BATCH
FREAD DATAFIL  [IF FILE NOT FORMATTED, USE READ DATAFIL 0]
STAT
LIST
LABEL 1 2 3 AS A B C
...ETC.
END
```

USE THE FOLLOWING FOR DATA ON CARDS:

```
JOB,T5,CM45000.
ACCOUNT,ACCTNO,PASSWORD.
 CCR,INPUT,DATAFIL.
FETCH,MULTREG.
MULTREG.
COST.
  --EOR--
DATAFILE   NO. OF CASES(ROWS)   NO. OF VARIABLES(COLUMNS)   FORMA
(FORMAT STATEMENT)
...DATA CARDS GO HERE...
  --EOR--
SET BATCH
FREAD DATAFIL
STAT
LIST
LABEL 1 2 3 AS A B C
...ETC.
END
```

In the above, --EOR-- means a card with the 7, 8, and 9 holes all punched in column number 1. In the Multreg control cards, the first card should always be SET BATCH for a batch input job. This will result in some optimization for batch jobs, including some page feeds, error checking and so on. The second card is then a READ or FREAD, to read the data file. From this point on, all the input should look exactly as it would for timesharing (terminal) input. Commands are free format, as long as numbers and keywords are separated by spaces or commas. Continuations are indicated as noted at the beginning of section 4.

Advice on Batch computing. A field length of 45000B is required. We suggest that the maximum time be kept low until one becomes experienced with the program; if the number of cases is under a 100 or so, most of the computations, except for ALL can be done in a fraction of a second; for larger data sets, the time should be increased somewhat. The command SCREEN is very cheap, and an excellent substitute for ALL.

In large data sets, the following commands may be more expensive than other commands: ALL, TRAN, RESTORE, PLOT, SCPLOT. If several plots with the same Y-axis are to be used, the command SCPLOT VS V1 V2 V3 will run in about one third the time of the three separate commands SCPLOT VS V1, SCPLOT VS V2, SCPLOT VS V3.

6. Appendices

6.1 Residual analysis with Multreg.

The usual model in (multiple) linear regression can be expressed as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

where Y_i, $i = 1,\ldots,n$ are the values of the dependent variable Y, and $X_{i1}, X_{i2}, \ldots, X_{ik}$ are the observed values of the k independent variables $X_1, \ldots, X_k$ associated with Y_i, and $\epsilon_1, \ldots, \epsilon_n$ are $N(0,\sigma^2)$. It is not unusual in applications for there to be one or more data points that do not seem to conform to this model. In the simple linear regression case (k=1) this may be obvious from a plot of Y against $X = X_i$. Most of the points may lie near a line but one (or more) do not fit the pattern. If a point really does not fit in with the remaining ones, there are several possible explanations:

1. An improbable but perfectly conforming observation was made. That is, Y_i satisfied the linear regression model but the ϵ_i associated with Y_i happened to be "large" as normal deviates will occasionally be. In such a case, theory suggests the point should be included in the estimation process.
2. After checking records or otherwise investigating, the observation is found to be associated with some exceptional circumstance such as a power failure, the first day of a newly hired technician, data known to come from a poorly drained plot in a field, etc. In such a case, one is probably justified in eliminating the point from the set of data and estimating the regression model without it.
3. An exceptional event in fact occurred such as in (2) but no record or evidence exists. Again, we would like to be able to leave out the point.
4. The point is perfectly legitimate, nothing exceptional or even improbable occurred. However, the expected value of Y for that set of independent variable values does <u>not</u> lie on a line or plane with (most of) the other points. If this is the case, the data point may be

the most important of all. It may represent new and unexpected information. Often it does not contain information about the relationship governing the remaining points and should be omitted from the estimation equations. However, the information it does contain should be reported.

In (2), (3), and (4) the conclusion is the same for estimation of regression coefficients: Compute the regression excluding the point that doesn't belong. Even when it is possible to identify "outlying" points, however, one can rarely distinguish between (3) and (4). However, correct estimation in (2), (3), and (4) does require identification of outliers. Since (1) is always a possibility there is the chance that if we exclude any point we are discarding real information about the regression coefficient. Moreover, because we discard only data points with large residuals, there is a tendency to under-estimate the true error variance which is based on the sum of squares of the residuals. For these reasons we need solutions to two problems: (A) How to identify possible outliers; and (B) How to protect ourselves from discarding good data. In practice, B is a requirement for a test of the null hypothesis that the suspected point does conform to the same model as the remaining points.

There is an additional problem associated with B. Since generally when tests for outliers are made the data points to be tested are determined after seeing the data, there is substantial danger of selection bias. Because only "large" residuals will be tested, naive application of rejection t-tests will almost certainly lead to a much higher rate of Type I error when in fact all the data is "good". For this reason, even though the test statistics proposed below take the form of t-tests, modified critical values allowing for selection are needed except in rare cases. The modified values are given in an attached table.

The approaches that are used for problem A are manifold and no pretence is made here of exhaustive presentation. First and foremost is an examination of observed residuals from a regression using all the data, as is done by the RESIDUAL command in MULTREG. It is straightforward to show, under the assumptions stated, that the variance of the i-th residual, $\hat{e}_i$, is given by $\mathrm{Var}(\hat{e}_i) = \sigma^2_{Y\cdot X}\,(1-v_{ii})$, where v_{ii} is a known constant that depends on the values of $X_1, X_2, \ldots, X_k$ for the i-th data point and on the sample means $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ of the independent variables. For

example, in simple regression (k=1), $v_{ii} = \frac{1}{n} + (X_{i1}-\bar{X}_1)^2 / \sum_{j=1}^{n}(X_{j1}-\bar{X}_1)^2$. In general, v_{ii} will be a quadratic function in $(X_{ij}-\bar{X}_j)$, and, hence may vary greatly from data point to data point. We can, however, give the following rules:

1) the variance of residuals near the center of the data (e.g., with x-values near their mean) will have v_{ii} <u>small</u>, and hence $\text{Var}(\hat{e}_i)$ will be large.

2) the variance of residuals with values of X's far from the center of the data will have $\text{Var}(\hat{e}_i)$ small.

(For further discussion, see D.W. Behnken and Draper, N.R., (1972), "Residuals and their variance patterns," <u>Technometrics</u> <u>14</u>, 101-114.) As consequences of these rules, it will be more difficult to find large residuals at unusual values of the X's then at usual ones. This suggests that a convenient rescaling of the residuals would be to divide each $\hat{e}_i$ by an estimate of its standard error, $\sqrt{s^2(1-v_{ii})}$, where s^2 is the residual mean square from the regression. These numbers are given in the column "STUD. RES" in the RESID output in Multreg. These numbers are called <u>studentized</u> or <u>standardized</u> residuals. (Note: Few, if any, other computer programs offer automatic studentization of residuals. At best, other programs will use the ratio $\hat{e}_i/s$ as their "standardized" residuals. While some of the techniques of analysis discussed here are appropriate for $\hat{e}_i/s$, the analyst must be careful in interpretation of results, since the problem of unequal variances has <u>not</u> been solved.)

To a first approximation, if there are no outliers, the studentized residuals should behave approximately like a sample of size n from N(0,1). In particular, values greater than, say, $\pm$ 2 may be worthy of attention. A rankit plot of the residuals (in Multreg, PLOT VS RANKIT) may also be useful. Apparent curvature in the plot gives evidence that the normality assumption is not valid. One or more residuals that "stick out" at either end of the plot provide candidates for outliers. As a summary statistic for normal plots, MULTREG computes an approximation to the Wilk-Shapiro W statistic. W as computed is simply defined to be the square of the correlation between the ordered studentized residuals and the rankits (i.e., expected normal order statistics). The hypothesis of normality is rejected if W is too <u>small</u>. Table 3 at the end of the manual gives critical values for W for selected sample sizes less than 100.

Plotting residuals against fitted or predicted values $\hat{Y}$ can be helpful in the multiple regression case. Large residuals associated with one of the most extreme predicted values may indicate that values of the independent variables at that point are outside the region of validity of the model. Such a plot can be obtained easily in MULTREG by the command, PLOT[model] VS PRED. Plots against predicted values can be used to locate a number of failures in a model, especially in finding failures that depend upon magnitude of response, most notably failures of the linear model (curvature), or heterogeneity of variance (usually with the magnitude of the dispersion increasing with level of response). Plots of residuals versus the dependent variable should not be used as the residuals and the observed values are positively correlated and hence interpretation of such a plot is nearly impossible. The slope of the regression of the residuals on the predicted values, on the other hand, is zero.

Once one has determined a point or points that need closer attention, some objective method of assessing the significance of the size of the residual is needed. The most widely used method is as follows: Delete the suspect point for the set of data and recompute the regression. We then need to have a technique to assess the effect of the suspected data point on the regression. We shall do this using three statistics, suggested by R.D. Cook (1977), "Detection of influential observations on linear regression," Technometrics 19, 15-18.

Cook's distance measure. Suppose we write the linear model in matrix form $Y = X\beta + \varepsilon$. Then, a simultaneous $(1 - \alpha) \times 100\%$ confidence interval for $(\beta_0, \beta_1, \ldots, \beta_k)$ is given, by the Scheffé method, to be the set of all β which satisfy

$$\frac{(\beta-\hat{\beta})' X'X (\beta-\hat{\beta})}{(k+1)s^2} \leq F(k+1, n-k-1; 1-\alpha)$$

where $F(k+1, n-k-1; 1-\alpha)$ is the $(1-\alpha) \times 100\%$ point of the F distribution, with $k+1$, $n-k-1$ df and $\hat{\beta} = (X'X)^{-1} X'Y$ is the least squares estimate of β. Now, to determine the influence of the i-th data point on regression, one could

delete that point, and re-estimate the β-vector. Following Cook, let $\hat{\beta}_{(-i)}$ be the estimate of β with the i-th data point removed. If the i-th point has little influence, one might expect $\hat{\beta}_{(-i)}$ to nearly equal $\hat{\beta}$; however, if the i-th point is influential $\hat{\beta}_{(-i)}$ should be quite different from $\hat{\beta}$. This suggests looking at the distance measure

$$D_i = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(-i)} - \hat{\beta})}{(k+1)s^2} \qquad i = 1,\ldots,n \ .$$

We can interpret D_i as the distance that $\hat{\beta}$ moves when the i-th data point is removed. We can compare D_i to the F distribution with k+1 and n-k-1 degrees of freedom. For example, if $D_i \approx F(k+1, n-k-1, .5)$, then the i-th data point moves the least squares estimate to the edge of a 50% confidence region for β based on $\hat{\beta}$. Such a situation may cause concern. For an uncomplicated analysis, one would like each D_i to give a 10%, say, confidence region.

The D_i are given in the column marked "DISTANCE" in the RESIDUAL output from Multreg.

(Notes on F distributions: Usually the D_i will be much smaller than 1, while most F-tables give only values for the upper tail (e.g., 95% and 99% points). While better tables do exist (c.f. Dixon and Massey (1969)), it is useful to know the following two rules about F distributions: (a) the median (50% point) of an F-distribution is approximately equal to 1; (b) suppose $a = F(n_1, n_2; 1-\alpha)$; that is, a is the $(1-\alpha)$ x 100% point of F with (n_1, n_2) degrees of freedom. Then $\frac{1}{a} = F(n_2, n_1; \alpha)$; that is, 1/a is the α x 100% point of F with (n_2, n_1) degrees of freedom. For example, $2.81 = F(3, 9, .90)$, so that $1/2.81 = .36 = F(9, 3; .10)$. With this rule, percentage points in the lower tail can be found from tabled values in the upper tail.)

From the derivation of D_i given, it appears that a separate regression must be computed with each data point deleted. However, this is not the case, and D_i can be computed as

$$D_i = \frac{(\text{i-th studentized residual})^2}{k+1} \ \frac{v_{ii}}{1 - v_{ii}}$$

and hence D_i depends only on the studentized residuals, the number of parameters k+1, and the ratio $v_{ii}/(1 - v_{ii})$. As noted earlier, $\sigma^2(1 - v_{ii})$

is the variance of the i-th residual; also, $\sigma^2 v_{ii}$ is the variance of the i-th fitted value $\hat{Y}_i$. The ratio $v_{ii}/(1 - v_{ii})$ measures the relative sensitivity of the estimate $\hat{\beta}$, to potential outlying values at the i-th data point. Large values mean that the point is important in the determination of $\hat{\beta}$. The values $v_{ii}/(1 - v_{ii})$ are given in the column V(YHAT)/V(RES) in the RESID output.

A test for a single outlier. Once an outlier is suspected, it is reasonable to test to see if it is an outlier. The classical procedure is to eliminate the suspected point, fit the regression line, and test to see if this point can be adequately described by the line fit from the remaining points. Again, using the results of Cook, the necessary computations can be done from computations on the complete data, since the test for a single outlier in multiple regression is a monotonic transformation of the largest studentized residual. The appropriate t-test is given by

$$t_i = r_i \sqrt{\frac{n-k-2}{n-k-1-r_i^2}}$$

where r_i is the i-th studentized residual. t_i is distributed as a t-statistic with n-k-2 degrees of freedom. We declare an observation to be an outlier if t_i is too big. (The t_i are in the column headed T on the RESID output.)

When, as is usually the case, the points tested for being outliers are chosen after seeing the data, the usual Student's critical values should not be used in this test. Since often we will only test the largest of n residuals, our true probability of Type I error will be on the order of $n\alpha$, if α is the level of significance sought. The answer is the application of what is known as Bonferroni's inequality which in this case states that the probability of a Type I error is in fact not greater than $n\alpha$. Thus if we use $t_{\alpha/n}(n-k-2)$ as the cut-off value in the test, our (Type I) error rate is no greater than $n(\alpha/n) = \alpha$, as desired. To ease the application of this result, attached are tables of $t_{.05/n}(n-k-2)$ and $t_{.01/n}(n-k-2)$. In the table, n is the number of cases in the data and p' is the total number of parameters in the model (p' = k + 1 if there is a constant in the model and p' = k if no constant). For example, for n = 39, α = .05, k = 5 (so p' = 5 + 1 = 6) we compare t_i to 3.53.

6.2 Variable Selection in Linear Regression

Suppose we have a linear model

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_i X_{ij} + e_i \qquad i = 1, \ldots, n ,$$

with $Var(e_i) = \sigma^2$, and each e_i uncorrelated. A common problem is that of selecting a subset of the k X-variables that can be used to provide an adequate model for Y. This problem can arise in many different contexts, for example:

1. In a screening experiment, many X's may be measured, and a selection procedure is needed to obtain a subset of the X's for further study.

2. In some studies, variable selection is used to find a parsimonious model for description of a process. Finding relatively small models is often useful.

3. Selection of subsets may be desirable if the X's are very closely related (e.g. are nearly colinear, or, in matrix terms, if X'X is nearly singular)

4. If the goal of regression is prediction, predictions based on subsets are generally more precise than predictions from a model with extraneous variables.

5. The coefficients for the X's in the model are generally estimated more precisely from a subset than from a model with extraneous variables included.

Because of their wide applicability, the problems of variable selection have attracted wide interest in the statistical literature. A recent survey of the literature is given by R.R. Hocking (1976), who gives a good overview of the problem and provides extensive references. Here, we discuss the use of Multreg in using two principal selection techniques: stepwise regression and

all possible regressions. No attempt will be made at a complete description of the techniques.

In using these techniques, the user should be aware of the additional implicit assumptions of selection procedures: (1) all X's that are relevant are included in the set of k under study, along with a few extraneous variables, and (2) that all the X's are measured in proper scale. The latter point should be critically tested through residual analysis. The former point is more difficult and usually cannot be tested directly.

Techniques

Stepwise procedures. The most common selection procedures in use are the so-called stepwise procedures (reference: Draper and Smith (1966), Chapter 6). These techniques were formulated when computer time was very expensive, and a systematic way of dealing with the $2^k - 1$ possible models had to be developed. The simplest technique is a step up procedure, which works as follows (we use the Hald data, given in section 2, as an example; here, $k = 4$):

1. Start with the model $Y_i = \beta_0 + e_i$; i.e. the model with none of the X's included. Look at the correlations between Y and all the X's. Add to the equation that X with the highest correlation with Y. For the Hald data, X_4 is the best single predictor as shown below:

```
        NEXT? CORR
CORRELATION MATRIX
X1     1.000
X2      .2286       1.000
X3    -.8241      -.1392        1.000
X4    -.2454      -.9730        .2954E-01   1.000
Y      .7307       .8163       -.5347      -.8213       1.000
    .   X1          X2           X3          X4           Y
```

Add the best predictor, here X_4, to the model.

2. Now, consider the partial correlations between Y and all the X's not in the model, controlling for X's in the model. In Multreg, this is done with the PARCOR command:

```
NEXT? PARCOR Y ON X4
PARTIAL CORRELATIONS WITH Y
   X1          X2          X3
 .9568       .1302      -.8951
```

Add to the equation that X with the highest partial correlation; in Hald's data X_1 gets added after X_4 to give the model $Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + e$.

3. Repeat step 2 until some stopping rule is satisfied. The usual rules are: stop with subset of some predetermined size p; stop if the t-statistic or partial F-test for the variable last added is less than some predetermined value, t = 2 or F = 4 are common values (recent research suggests that choosing the upper 15% point of t or F is a good choice); stop if the change in R^2 is sufficiently small.

Using the t = 2 stopping rule on the Hald data, leads to the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + e$, since in the model with all 4 predictors the t-test for X_3 last is t = .1350, much less than 2.

```
NEXT? PARCOR Y ON X4 X1
PARTIAL CORRELATIONS WITH Y
   X2          X3
 .5986      -.5657

NEXT? REGS Y ON X1 X2 X4
 REGS     Y     ON    X1    X2    X4
VARIABLE      COEF'T          ST. ERROR          T VALUE
  BO        71.64831          14.14239            5.07
  X1         1.451938          .1169976           12.41
  X2         .4161098          .1856105            2.24
  X4        -.2365402          .1732878           -1.37
        DEGREES OF FREEDOM  =       9
        RESIDUAL MEAN SQUARE=  5.330303
        ROOT MEAN SQUARE    =  2.308745
        R-SQUARED           =      .9823
```

The step down procedure starts with the full model, and deletes variables one at a time. At each step, that variable with the smallest partial F or t test is deleted; one might stop when the t or F to delete is too big, again t = 2 or F = 4 is a common value. For the Hald data, the following computations lead to the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$. This is not the same model as obtained by step up.

```
     NEXT? REGS Y ON X1 X2 X3 X4
  REGS     Y     ON    X1    X2    X3    X4
VARIABLE       COEF'T          ST. ERROR          T VALUE
  BO         62.40537           70.07096            .89
  X1         1.551103           .7447699           2.08
  X2         .5101676           .7237880            .70
  X3         .1019094           .7547090            .14
  X4        -.1440610           .7090521           -.20
          DEGREES OF FREEDOM  =       8
          RESIDUAL MEAN SQUARE=   5.982955
          ROOT MEAN SQUARE    =   2.446008
          R-SQUARED           =      .9824

     NEXT? REGS Y ON X1 X2 X4
  REGS     Y     ON    X1    X2    X4
VARIABLE       COEF'T          ST. ERROR          T VALUE
  BO         71.64831           14.14239            5.07
  X1         1.451938           .1169976           12.41
  X2         .4161098           .1856105            2.24
  X4        -.2365402           .1732878           -1.37
          DEGREES OF FREEDOM  =       9
          RESIDUAL MEAN SQUARE=   5.330303
          ROOT MEAN SQUARE    =   2.308745
          R-SQUARED           =      .9823

     NEXT? REGS Y ON X1 X2
  REGS     Y     ON    X1    X2
VARIABLE       COEF'T          ST. ERROR          T VALUE
  BO         52.57735           2.286174           23.00
  X1         1.468306           .1213009           12.10
  X2         .6622505           .4585472E-01       14.44
          DEGREES OF FREEDOM  =      10
          RESIDUAL MEAN SQUARE=   5.790448
          ROOT MEAN SQUARE    =   2.406335
          R-SQUARED           =      .9787
```

There are many variants of stepwise procedures. The most important variation is a combination of step-up and step-down, where, at each step, a variable is added to the model or deleted from the model, according to some rule. The program BMDP2R (Dixon (1975)), which is probably the most versitile stepwise program, allows 4 different rules for stepping (this is in addition to the user being able to specify values for the t-to-enter or t-to-delete). This procedure, due to Efroymson, is so popular that the term, "stepwise" is often used to refer specifically to it.

Using stepwise methods. Stepwise methods are versitile, relatively easy to use, and, unfortunately, quite often are very misleading. None of the stepwise procedures will always (or even usually) find a "best" model (judged by any objective criterion). Step up and step down, for example, often lead to different models, as in the Hald data. Stepwise methods can give the unexperienced analyst a false sense of security, since it produces an artificial (and often meaningless) ordering to the variables, and appears to give the answer and find the model.

If the number of variables k is very large ($k > 27$) or if the number of variables is greater than the number of cases ($k > n$), only stepwise selection methods are available, either because of cost considerations in computations, or because of insufficient degrees of freedom for estimation. In these cases, the analyst should carefully consider his or her problem before doing any computations.

All possible regressions. A good strategy in subset selection would be to compare all possible subsets in terms of some criterion functions, and then carefully analyze the best few regression models in terms of their values on the criterion function. In Multreg, there are two commands that can be used to find the best subsets on the basis of a criterion, ALL and SCREEN. Additionally there are 4 possible criterion functions, R^2, the adjusted R^2, Mallows' C_p and the residual sum of squares (which is actually equivalent to R^2). We first discuss the computational methods used in ALL and SCREEN, and then discuss the criterion functions.

ALL. This command uses an efficient algorithm to literally compute all possible regressions of a dependent variable on subsets of k predictor variables. The algorithm, suggested by Schatzoff, Tsao and Fienberg (1968),

allows the computations of all 2^k subset regressions in exactly 2^k "sweeps" (sec. 3). For each regression, the number of parameters in the model (= p = number of variables in the subset + 1), and the values of the 4 summary statistics (defined below) are printed.

This method is computationally efficient in that very little storage is required in the computer; however for k moderate (say 10 or more), the amount of computation, and the amount of output, can be large. Consequently, several modifications to the ALL command are possible that can either limit computations or printing or both. They are as follows:

OMIT. [variable list]. OMITted variables are used in computing C_p, but are not included in any model. Each OMITted variable reduces computations by 50%.

FORCE [variable list]. FORCEd variables are included in every model. Each FORCEd variable reduces computations by 50%.

CPMIN, CPMAX. Models with values of C_p between CPMIN and CPMAX are printed (defaults are CPMIN = -10, CPMAX = 500). This reduces printing but not computations. The use of CPMAX = k + 1 is recommended if, in the regression in the full model, most variables have $|t|$ statistics $\geq \sqrt{2}$. If this is not the case, then a smaller value of CPMAX is recommended.

SCREEN. The ALL command computes all possible regressions in a fixed order, and may therefore be considered a "dumb" command. SCREEN, on the other hand, is an "intelligent" command. Rather than compute in a fixed order, the alogrithm decides which regression to compute next on the basis of the values of the criterion function as specified by the user. This algorithm is called "leaps and bounds" by its authors, George Furnival and Robert Wilson (1974), the bounds refering to bounds determined by previous computations, and the leaps refer to jumping from place to place to do computing.

Consequently, the leaps and bounds algorithm is very efficient, and is much less expensive than ALL. Its only drawback is that it requires a very large amount of storage, so that the number of variables is limited to 20.

The algorithm in the SCREEN command was obtained from Furnival and Wilson (1974). Nearly identical code is also available in the IMSL subroutine RLEAP, and in the program BMDP9R. This latter program is very similar to Multreg, except it is a batch (card input) program, not timesharing.

The OMIT and FORCE keywords can also be used in SCREEN.

Criteria functions. Several criteria functions have been proposed for choosing the "best" for all possible regressions. The simplest of these is R^2, the proportion of variability explain. All subsets of the same size can be compared by their value of R^2. This will lead to one or two potentially best subsets of each size to be studied further. The user must remember that, as the size of the subset increases, so does R^2.

Since R^2 is nondecreasing when variables are added, an adjusted version of R^2 has been proposed, defined by

$$\text{Adjusted } R^2 = 1 - \frac{n-1}{n-p} (1-R^2)$$

where p is the number of parameters in the model. In large samples, adjusted R^2 will differ little from R^2; in small samples, the adjusted version may be quite different from R^2. If R^2 is small, adjusted R^2 can be negative.

The adjusted R^2 tends to choose larger subsets over smaller ones. It is a popular criterion function in econometrics.

The C_p statistic is the fundamental criteria for model selection in Multreg. It was first suggested by Colin Mallows, and is discussed in detail by Daniel and Wood (1971).

Mallows considered problems in which the goal is the estimation of fitted values corresponding to the observed data. Suppose that the "true" regression function would yield fitted values $\nu_1, \nu_2, \ldots, \nu_n$ (these, of course, can never be observed). A model under study with p-1 predictors (and p parameters) will yield fitted values $\eta_1, \eta_2, \ldots, \eta_n$. The term $\sum_{j=1}^{n} (\nu_j - \eta_j)^2$ is called the squared bias in the equation under consideration; in principle this bias should be small. Mallows suggests that a good criteria for estimation would be the total error of estimation,

$$\Sigma(\nu_j - \eta_j)^2 + \Sigma \mathrm{Var}(\hat{Y}_j) \quad .$$

This equation immediately simplifies since $\Sigma \mathrm{Var}(\hat{Y}_j) = p\sigma^2$ ($= \sigma^2 \mathrm{trace}\ (X(X'X)^{-1}X')$), thus

$$\Gamma_p = \frac{\Sigma(\nu_j - \eta_j)^2}{\sigma^2} + p \ .$$

We now need to replace $\Sigma(\nu_j - \eta_j)^2$ and σ^2 by estimates to get a useful measure. From the p-1 variate model (with p parameters), the residual sum of squares will have expectation

$$E(RSS) = \Sigma(\nu_j - \eta_j)^2 + (n - p)\sigma^2 \quad .$$

Solving for $\Sigma(\nu_j - \eta_j)^2$, and substituting into the equation for Γ_p gives

$$\Gamma_p = \frac{E(RSS)}{\sigma^2} - (n - 2p) \quad .$$

Estimating E(RSS) by its observed value, and estimating σ^2 from the model with all predictors included, we estimate Γ_p by

$$C_p = \frac{RSS}{\hat{\sigma}^2} - (n - 2p) \quad .$$

Properties of C_p. The first important property of C_p is that, if the bias is zero, $E(C_p) = p$, so one should seek models with $C_p \simeq p$; usually models with $C_p \leq p$ are considered to be candidates for good models. The model including all variables will have $c_{k+1} = k + 1$ by inspection of the formula for C_p. Also, since C_p is simply a function of n, p, RSS, and $\hat{\sigma}^2$, C_p can be expressed in terms of other possible criterion functions, such as R^2 or partial F tests or overall F tests. The minimum possible value of C_p is 2p - k, which may be negative.

The output from the SCREEN command for the Hald data is given below. Similar output for ALL is given in Section 4.23.

```
      NEXT? SCREEN 5 ON 1 2 3 4
BEST  5 REGRESSIONS WITH Y    DEPENDENT, USING CP
 P       C(P)    R2(ADJ)    R**2     RSS        VARIABLES
 3   *   2.678    .9744     .9787   57.90       1  2
 4   *   3.018    .9764     .9823   47.97       1  2  4
 4   *   3.041    .9764     .9823   48.11       1  2  3
 4   *   3.497    .9750     .9813   50.84       1  3  4
 5   *   5.000    .9736     .9824   47.86       1  2  3  4
CP TIME USED IS    .070 SECONDS.
```

Mallows suggests plotting C_p vs p. This plot, for the Hald data, is given by Daniel and Wood (1971) p. 88; it is not produced by Multreg. From either the plot or the list of C_p's, four sets of predictors merit further study: 1,2; 1,2,3; 1,2,3; and 1,3,4. These four models should then be studied in detail.

In some circumstances the user may wish to obtain the estimate $\hat{\sigma}^2$ using some variables that are not to be included in variable selection. This may be accomplished in Multreg by use of the OMIT keyword. For example, ALL 4 ON 1 2 3 5 OMIT 5 would use V5 is estimating $\hat{\sigma}^2$, but would not use V5 in any further computations.

6.3 Using weighted least squares.

The column of weights can be used in many ways in Multreg. The primary uses are weighted least squares, comparing regression lines, and cross validation. We consider these three uses separately.

Weighted least squares. Suppose we have a linear model $Y = X\beta + e$ where the variance of e_i is $\text{Var}(e_i) = \sigma^2\eta_i$, such that the η_i are known numbers. This can occur if, for example, σ^2 is a known function of some data (e.g. "the variance is proportional to square of the amount of impurity of material"), or in using Multreg for unbalanced analysis of variance. In this case, the column of weights, w_i should consist of the reciprocals of the η_i, since we want to weight inversely proportional to variance, cases with large variance having smaller weight. All computations that are done in Multreg will be the correct weighted least squares computations.

In general, the use of a column of weights (via the SET WEIGHTS command) will make cases with large weights more important, while cases with small weights will be less important. The exact computing formulas are given in sec. 3.

Comparing regression lines. Any case with weight 0 will be ignored in computing estimates, test statistics, and so on. This can be exploited to give a relatively easy way of comparing regression lines for different groups.

The data in the table below gives the IQ scores of identical twins, with one twin raised by natural parents and the other raised in a foster home. The twins are classified by social class of their natural parents (high or low), as given in the data. The data were originally given by Sir Cyril Burt, and may well have been fabricated (see Science, 26 November 1976, p.916, and 21 January 1977, p.246). We shall see if the regression of Foster IQ (FOST) on Home IQ (HOME) is the same in each class. To do this we perform the following calculations, given below.

HOME	FOST	HIGH	LOW
82.00	82.00	1.000	0
90.00	80.00	1.000	0
91.00	88.00	1.000	0
115.0	108.0	1.000	0
115.0	116.0	1.000	0
129.0	117.0	1.000	0
131.0	132.0	1.000	0
68.00	63.00	0	1.000
73.00	77.00	0	1.000
81.00	86.00	0	1.000
85.00	83.00	0	1.000
87.00	93.00	0	1.000
87.00	97.00	0	1.000
93.00	87.00	0	1.000
94.00	94.00	0	1.000
95.00	96.00	0	1.000
97.00	112.0	0	1.000
97.00	113.0	0	1.000
103.0	106.0	0	1.000
106.0	107.0	0	1.000
111.0	98.00	0	1.000
MEAN			
96.67	96.90	.3333	.6667
STDEV			
16.58	16.38	.4830	.4830

```
      NEXT? $FIRST, COMPUTE ANOVA FOR THE FULL DATA
      NEXT? ANOVA FOSTER ON HOME
  ANOVA   FOST ON   HOME
                INDIVIDUAL                    CUMULATIVE
SOURCE     DF      SS          MS        DF      SS          MS          R**2
 MEAN       1   .1972E+06   .1972E+06
 HOME       1   4239.       4239.          1   4239.       4239.       .7904
RESIDUAL   19   1125.       59.19         20   5364.       268.2

      NEXT? $NOW, SELECT THE 'HIGH' GROUP
      NEXT? SET WEIGHTS HIGH
VARIABLE HIGH USED AS WEIGHTS.

      NEXT? $GET THE ANOVA FOR THE HIGH GROUP
      NEXT? ANOVA
  ANOVA   FOST ON   HOME
WEIGHTED BY HIGH
                INDIVIDUAL                    CUMULATIVE
SOURCE     DF      SS          MS        DF      SS          MS          R**2
 MEAN       1   .7468E+05   .7468E+05
 HOME       1   2251.       2251.          1   2251.       2251.       .9282
RESIDUAL    5   174.2       34.85          6   2425.       404.2

      NEXT? $SET WEIGHTS TO SELECT THE 'LOW' GROUP
      NEXT? SET WEIGHTS LOW
ALL CASES RESTORED TO THE DATA SET.
VARIABLE LOW  USED AS WEIGHTS.

      NEXT? $GET ANOVA FOR THE LOW GROUP
      NEXT? ANOVA
  ANOVA   FOST ON   HOME
WEIGHTED BY LOW
                INDIVIDUAL                    CUMULATIVE
SOURCE     DF      SS          MS        DF      SS          MS          R**2
 MEAN       1   .1230E+06   .1230E+06
 HOME       1   1700.       1700.          1   1700.       1700.       .6772
RESIDUAL   12   810.5       67.54         13   2511.       193.1

      NEXT? $THE F-TEST FOR EQUALITY OF LINES MUST BE
      NEXT? $COMPUTED BY HAND.  THE FORMULA IS GIVEN ON THE
      NEXT? $NEXT PAGE.
      NEXT? $NOW, RESTORE ALL DATA AND COMPUTE THE REGRESSION.
      NEXT? SET WEIGHTS 0
ALL CASES RESTORED TO THE DATA SET.
VARIABLE V0   USED AS WEIGHTS.

      NEXT? REGS
  REGS    FOST ON   HOME
VARIABLE      COEF'T          ST. ERROR          T VALUE
  B0        11.99596          10.17207            1.18
  HOME       .8783669          .1037853           8.46
         DEGREES OF FREEDOM  =     19
         RESIDUAL MEAN SQUARE=  59.18516
         ROOT MEAN SQUARE    =  7.693189
         R-SQUARED           =     .7904
```

1) Compute the ANOVA for the full data. The residual sum of squares, SSE_{ALL} = 1125. with 19 d.f.

2) Set weights to the high group indicator to eliminate the low group from estimation. Compute the Anova, get RSS_{HIGH} = 174.2, 5 d.f.

3) Set weights for the low group indicator. The Anova gives RSS_{LOW} = 810.5, 12 d.f.

4) The F test to compare regression lines is found, from the general F - test, to be

$$F = \frac{(SSE_{ALL} - SSE_{HIGH} - SSE_{LOW})/(19-5-12)}{(SSE_{HIGH} + SSE_{LOW})/5+12)} = 1.18$$

The F is computed by hand, and has (2, 19) d.f. clearly, there is no difference between regression lines, no "class effect". The regression line fit to the whole data set, given in the computations, is adequate.

Validation. Validation means testing a model fitted to one set of data on additional cases. This is done in Multreg using SET WEIGHTS and the RESID command (additionally, YHAT and PREDICT may be helpful). For example, suppose we fit a model only to the high group in the IQ data as is shown below. Suppose we want to see if the model fit to the high group is an adequate predictor of the model fit to the low group; this suggests looking at the RESIDUAL command. All "starred" cases were not used in computing the regression line. We see that the predictor based on the high group tends to underpredict, on the average, for the low group, since the T-values are mostly positive. However, none of the predictions are very far off since

the largest t-value is only 3.11.

As a summary of the fit, we might compare the PRESS for the included and excluded cases (or perhaps, PRESS/(no. of cases in the group).

```
      NEXT? SET WEIGHTS HIGH
VARIABLE HIGH USED AS WEIGHTS.

      NEXT? REGS FOSTER ON HOME
  REGS    FOST ON   HOME
WEIGHTED BY HIGH
VARIABLE      COEF'T          ST. ERROR          T VALUE
  BO       -1.872044          13.27250            -.14
  HOME      .9775622          .1216272            8.04
         DEGREES OF FREEDOM  =       5
         RESIDUAL MEAN SQUARE=   34.84851
         ROOT MEAN SQUARE    =   5.903263
         R-SQUARED           =      .9282

      NEXT? $TO SEE IF THE MODEL FIT TO THE HIGH GROUP ADEQUATELY
      NEXT? $DESCRIBES THE LOW GROUP, USE THE RESIDUAL COMMAND.
      NEXT? $
      NEXT? RESIDUALS
  RESID   FOST ON   HOME
WEIGHTED BY HIGH
CASE       FOST      RESIDUAL  STUD. RES       V          DISTANCE      T
   1      82.00        3.712      .8260      .4204         .2475       .79
   2      80.00       -6.109     -1.2144     .2739         .2782     -1.29
   3      88.00        .9139      .1799      .2594         .0057       .16
   4      108.0       -2.548      -.4726     .1663         .0223      -.43
   5      116.0        5.452      1.0115     .1663         .1020      1.01
   6      117.0       -7.233     -1.5058     .3378         .5782     -1.82
   7      132.0        5.811      1.2461     .3759         .4675      1.34
   8 *    63.00       -1.602                 .8076                    -.20
   9 *    77.00        7.510                 .6502                     .99
  10 *    86.00        8.690                 .4426                    1.23
  11 *    83.00        1.779                 .3591                     .26
  12 *    93.00        9.824                 .3225                    1.45
  13 *    97.00        13.82                 .3225                    2.04
  14 *    87.00       -2.041                 .2330                    -.31
  15 *    94.00        3.981                 .2210                     .61
  16 *    96.00        5.004                 .2099                     .77
  17 *    112.0        19.05                 .1903                    2.96
  18 *    113.0        20.05                 .1903                    3.11
  19 *    106.0        7.183                 .1517                    1.13
  20 *    107.0        5.250                 .1439                     .83
  21 *    98.00       -8.637                 .1478                   -1.37
                  INCLUDED CASES     EXCLUDED CASES(*)
RESIDUAL SS       174.2425713        1388.871427
PRESS             371.4412429        1388.871427
```

The t-values for the omitted cases (under usual assumptions) are distributed as students t with degrees of freedom equal to the degrees of freedom of the estimate of σ^2, in the example, the d.f. is 5. (Recall that for the included cases, the number of degrees of freedom is one less than the number of d.f. for $\hat{\sigma}^2$, in this case 4.) The Bonferroni inequality should be used to compute p-values for these t-statistics.

References

D.M. Allen, The sum of prediction errors as a criterion for selecting prediction variables. Tech. Report #23, Department of Statistics, University of Kentucky (1971).

D.F. Andrews and J.W. Tukey, "Teletypewriter plots for data analysis can be fast: 6-line plots including probability plots, "Applied Statistics 22 (1973), 192-203.

A.E. Beaton, The Use of Special Matrix Operators in Statistical Calculus. Research Bulletin RB-64-51 (1964), Educational Testing Service, Princeton, N.J.

R.D. Cook, "Detection of Influential observations in linear regression," Technometrics (1977), 19, 15-18.

C. Daniel and F.S. Wood, Fitting Equations to Data. (1971). John Wiley.

A.P. Dempster, Elements of Continuous Multivariate Analysis. (1969) Addison-Wesley.

W. Dixon and M. Brown, BMDP 77. University of California Press (1977).

W.J. Dixon and F. Massey, Introduction to Statistical Analysis, Third Edition (1969), McGraw-Hill.

N. Draper and H. Smith, Applied Regression Analysis. (1966). John Wiley.

G. Furnival and R. Wilson, Regression by leaps and bounds. Technometrics 16 (1974), 499-511.

R.R. Hocking, "The Analysis and selection of variables in linear regression," Biometrics 32 (March 1976), 1-49.

M. Schatzoff, R. Tsao and S.E. Fienberg, "Efficient calculation of all possible regressions," Technometrics 10 (1968), 769-779.

H. Theil, Principles of Econometrics, (1971) John Wiley.

Critical values for outlier test, $\alpha = .01$

n / p'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
6	7.53	10.87	24.46	382.0														
7	6.35	7.84	11.45	26.43	445.6													
8	5.71	6.54	8.12	11.98	28.26	509.3												
9	5.31	5.84	6.71	8.38	12.47	29.97	573.0											
10	5.04	5.41	5.96	6.87	8.61	12.92	31.60	636.6										
11	4.85	5.12	5.50	6.07	7.01	8.83	13.35	33.14	700.3									
12	4.71	4.91	5.19	5.58	6.17	7.15	9.03	13.75	34.62	763.9								
13	4.60	4.76	4.97	5.25	5.66	6.26	7.27	9.22	14.12	36.03	827.6							
14	4.51	4.64	4.81	5.02	5.32	5.73	6.35	7.39	9.40	14.48	37.40	891.3						
15	4.44	4.55	4.68	4.85	5.08	5.37	5.80	6.43	7.50	9.57	14.82	38.71	954.9					
16	4.38	4.48	4.59	4.72	4.90	5.12	5.43	5.86	6.51	7.60	9.73	15.15	39.98					
17	4.34	4.41	4.51	4.62	4.76	4.94	5.17	5.48	5.92	6.59	7.70	9.88	15.46	41.21				
18	4.30	4.36	4.44	4.54	4.66	4.80	4.98	5.21	5.53	5.98	6.66	7.80	10.03	15.76	42.41			
19	4.26	4.32	4.39	4.47	4.57	4.69	4.83	5.01	5.25	5.57	6.03	6.72	7.89	10.17	16.05			
20	4.23	4.29	4.35	4.42	4.50	4.60	4.72	4.86	5.05	5.29	5.62	6.08	6.79	7.98	10.31			
21	4.21	4.26	4.31	4.37	4.44	4.52	4.62	4.74	4.89	5.08	5.33	5.66	6.13	6.85	8.06			
22	4.19	4.23	4.28	4.33	4.39	4.46	4.55	4.65	4.77	4.92	5.11	5.36	5.70	6.18	6.91			
23	4.17	4.21	4.25	4.30	4.35	4.41	4.49	4.57	4.67	4.80	4.95	5.14	5.40	5.74	6.22	47.94		
24	4.15	4.19	4.22	4.27	4.32	4.37	4.43	4.51	4.59	4.70	4.82	4.98	5.17	5.43	5.78	17.36		
25	4.14	4.17	4.20	4.24	4.28	4.33	4.39	4.45	4.53	4.62	4.72	4.85	5.00	5.20	5.46	10.92		
26	4.12	4.15	4.18	4.22	4.26	4.30	4.35	4.41	4.47	4.55	4.64	4.74	4.87	5.03	5.23	8.43		
27	4.11	4.14	4.17	4.20	4.24	4.27	4.32	4.37	4.43	4.49	4.57	4.66	4.76	4.89	5.05	7.17		
28	4.10	4.13	4.15	4.18	4.21	4.25	4.29	4.33	4.38	4.44	4.51	4.59	4.68	4.78	4.91	6.43	52.90	
29	4.09	4.12	4.14	4.17	4.20	4.23	4.26	4.30	4.35	4.40	4.46	4.53	4.60	4.69	4.80	5.94	18.50	
30	4.09	4.11	4.13	4.15	4.18	4.21	4.24	4.28	4.32	4.36	4.42	4.47	4.54	4.62	4.71	5.60	11.44	
31	4.08	4.10	4.12	4.14	4.17	4.19	4.22	4.26	4.29	4.33	4.38	4.43	4.49	4.56	4.64	5.35	8.75	
32	4.07	4.09	4.11	4.13	4.15	4.18	4.21	4.24	4.27	4.31	4.35	4.39	4.45	4.50	4.57	5.16	7.40	
33	4.07	4.08	4.10	4.12	4.14	4.17	4.19	4.22	4.25	4.28	4.32	4.36	4.41	4.46	4.52	5.01	6.60	57.43
34	4.06	4.08	4.09	4.11	4.13	4.15	4.18	4.20	4.23	4.26	4.29	4.33	4.37	4.42	4.47	4.89	6.09	19.51
35	4.06	4.07	4.09	4.11	4.12	4.14	4.16	4.19	4.21	4.24	4.27	4.31	4.34	4.39	4.43	4.79	5.72	11.90
36	4.05	4.07	4.08	4.10	4.12	4.13	4.15	4.18	4.20	4.22	4.25	4.28	4.32	4.36	4.40	4.71	5.46	9.03
37	4.05	4.06	4.08	4.09	4.11	4.13	4.14	4.16	4.19	4.21	4.24	4.26	4.29	4.33	4.37	4.64	5.26	7.60
38	4.05	4.06	4.07	4.09	4.10	4.12	4.13	4.15	4.17	4.20	4.22	4.25	4.27	4.31	4.34	4.59	5.10	6.76
39	4.04	4.06	4.07	4.08	4.10	4.11	4.13	4.14	4.16	4.18	4.21	4.23	4.26	4.28	4.32	4.54	4.97	6.21
40	4.04	4.05	4.06	4.08	4.09	4.10	4.12	4.14	4.15	4.17	4.19	4.22	4.24	4.27	4.29	4.49	4.87	5.83
50	4.03	4.03	4.04	4.05	4.06	4.07	4.07	4.08	4.09	4.10	4.12	4.13	4.14	4.15	4.17	4.25	4.38	4.59
60	4.03	4.03	4.04	4.04	4.05	4.05	4.06	4.06	4.07	4.08	4.08	4.09	4.10	4.11	4.12	4.17	4.23	4.32
70	4.03	4.03	4.04	4.04	4.05	4.05	4.05	4.06	4.06	4.07	4.07	4.08	4.08	4.09	4.09	4.13	4.17	4.22
80	4.04	4.04	4.04	4.05	4.05	4.05	4.06	4.06	4.06	4.07	4.07	4.07	4.08	4.08	4.09	4.11	4.13	4.17
90	4.05	4.05	4.05	4.05	4.06	4.06	4.06	4.06	4.07	4.07	4.07	4.07	4.08	4.08	4.08	4.10	4.12	4.14
100	4.06	4.06	4.06	4.06	4.06	4.07	4.07	4.07	4.07	4.07	4.08	4.08	4.08	4.08	4.09	4.10	4.11	4.13
200	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.16	4.16	4.16
300	4.21	4.21	4.21	4.21	4.21	4.21	4.22	4.22	4.22	4.22	4.22	4.22	4.22	4.22	4.22	4.22	4.22	4.22
400	4.26	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27

n / p'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
6	4.85	6.23	10.89	76.39														
7	4.38	5.07	6.58	11.77	89.12													
8	4.12	4.53	5.26	6.90	12.59	101.9												
9	3.95	4.22	4.66	5.44	7.18	13.36	114.6											
10	3.83	4.03	4.32	4.77	5.60	7.45	14.09	127.3										
11	3.75	3.90	4.10	4.40	4.88	5.75	7.70	14.78	140.1									
12	3.69	3.81	3.96	4.17	4.49	4.98	5.89	7.94	15.44	152.8								
13	3.65	3.74	3.86	4.02	4.24	4.56	5.08	6.02	8.16	16.08	165.5							
14	3.61	3.69	3.79	3.91	4.07	4.30	4.63	5.16	6.14	8.37	16.69	178.2						
15	3.58	3.65	3.73	3.83	3.95	4.12	4.36	4.70	5.25	6.25	8.58	17.28	191.0					
16	3.56	3.62	3.68	3.77	3.87	4.00	4.17	4.41	4.76	5.33	6.36	8.77	17.85	203.7				
17	3.54	3.59	3.65	3.72	3.80	3.90	4.04	4.21	4.46	4.82	5.40	6.47	8.95	18.40	216.4			
18	3.53	3.57	3.62	3.68	3.75	3.83	3.94	4.08	4.26	4.51	4.88	5.47	6.57	9.13	18.93			
19	3.52	3.56	3.60	3.65	3.71	3.78	3.86	3.97	4.11	4.30	4.55	4.93	5.54	6.67	9.30			
20	3.51	3.54	3.58	3.62	3.67	3.73	3.81	3.89	4.00	4.15	4.33	4.59	4.98	5.60	6.76			
21	3.50	3.53	3.57	3.60	3.65	3.70	3.76	3.83	3.92	4.03	4.18	4.37	4.64	5.03	5.67			
22	3.50	3.52	3.55	3.59	3.63	3.67	3.72	3.78	3.86	3.95	4.06	4.21	4.40	4.68	5.08	280.1		
23	3.49	3.52	3.54	3.57	3.61	3.65	3.69	3.75	3.81	3.88	3.98	4.09	4.24	4.44	4.71	21.41		
24	3.49	3.51	3.53	3.56	3.59	3.63	3.67	3.71	3.77	3.83	3.91	4.00	4.12	4.27	4.47	10.07		
25	3.48	3.50	3.53	3.55	3.58	3.61	3.65	3.69	3.73	3.79	3.85	3.93	4.02	4.14	4.30	7.17		
26	3.48	3.50	3.52	3.54	3.57	3.60	3.63	3.66	3.70	3.75	3.81	3.87	3.95	4.05	4.17	5.95		
27	3.48	3.50	3.52	3.54	3.56	3.58	3.61	3.65	3.68	3.72	3.77	3.83	3.89	3.97	4.07	5.29	343.8	
28	3.48	3.50	3.51	3.53	3.55	3.58	3.60	3.63	3.66	3.70	3.74	3.79	3.84	3.91	3.99	4.88	23.63	
29	3.48	3.49	3.51	3.53	3.55	3.57	3.59	3.62	3.64	3.68	3.71	3.76	3.81	3.86	3.93	4.61	10.74	
30	3.48	3.49	3.51	3.52	3.54	3.56	3.58	3.60	3.63	3.66	3.69	3.73	3.77	3.82	3.88	4.42	7.53	
31	3.48	3.49	3.50	3.52	3.54	3.55	3.57	3.59	3.62	3.64	3.67	3.71	3.74	3.79	3.84	4.28	6.18	
32	3.48	3.49	3.50	3.52	3.53	3.55	3.57	3.59	3.61	3.63	3.66	3.69	3.72	3.76	3.80	4.17	5.47	407.4
33	3.48	3.49	3.50	3.52	3.53	3.54	3.56	3.58	3.60	3.62	3.64	3.67	3.70	3.74	3.77	4.08	5.03	25.66
34	3.48	3.49	3.50	3.51	3.53	3.54	3.56	3.57	3.59	3.61	3.63	3.66	3.68	3.71	3.75	4.01	4.74	11.34
35	3.48	3.49	3.50	3.51	3.52	3.54	3.55	3.57	3.58	3.60	3.62	3.64	3.67	3.70	3.73	3.96	4.53	7.84
36	3.48	3.49	3.50	3.51	3.52	3.54	3.55	3.56	3.58	3.60	3.61	3.63	3.66	3.68	3.71	3.91	4.37	6.39
37	3.48	3.49	3.50	3.51	3.52	3.53	3.55	3.56	3.57	3.59	3.61	3.62	3.65	3.67	3.69	3.87	4.26	5.62
38	3.48	3.49	3.50	3.51	3.52	3.53	3.54	3.56	3.57	3.58	3.60	3.62	3.64	3.66	3.68	3.84	4.16	5.16
39	3.49	3.49	3.50	3.51	3.52	3.53	3.54	3.55	3.57	3.58	3.59	3.61	3.63	3.65	3.67	3.81	4.09	4.84
40	3.49	3.49	3.50	3.51	3.52	3.53	3.54	3.55	3.56	3.58	3.59	3.60	3.62	3.64	3.66	3.79	4.03	4.62
50	3.51	3.51	3.51	3.52	3.53	3.53	3.54	3.54	3.55	3.56	3.57	3.57	3.58	3.59	3.60	3.66	3.75	3.88
60	3.53	3.53	3.53	3.54	3.54	3.54	3.55	3.55	3.56	3.56	3.57	3.57	3.58	3.58	3.59	3.62	3.67	3.73
70	3.55	3.55	3.55	3.55	3.56	3.56	3.56	3.56	3.57	3.57	3.57	3.58	3.58	3.59	3.59	3.61	3.64	3.67
80	3.57	3.57	3.57	3.57	3.57	3.58	3.58	3.58	3.58	3.58	3.59	3.59	3.59	3.60	3.60	3.61	3.63	3.66
90	3.58	3.59	3.59	3.59	3.59	3.59	3.59	3.60	3.60	3.60	3.60	3.60	3.60	3.61	3.61	3.62	3.63	3.65
100	3.60	3.60	3.60	3.60	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.62	3.62	3.62	3.62	3.63	3.64	3.65
200	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.74	3.74	3.74	3.74
300	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.82	3.82	3.82	3.82	3.82	3.82	3.82	3.82
400	3.87	3.87	3.87	3.87	3.87	3.87	3.87	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88
500	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92

Table 3. Selected percentage points of the W statistic*

n/α	.01	.05	.10	.50
5	.675	.777	.817	.922
10	.776	.842	.869	.940
15	.815	.878	.903	.954
20	.858	.902	.921	.962
35	.919	.943	.952	.976
50	.935	.953	.963	.981
75	.956	.969	.973	.986
99	.967	.976	.980	.989

(Note: Reject hypothesis of normality for small values of W)

*From Weisberg, S., (1974), "An empirical comparison of W and W′, Biometrika 61, 645-646, and Shapiro, S.S. and Francis, R.S. (1972), "An approximate analysis of variance test for normality," J. Amer. Statist. Assoc., 67, 215-216.