HEWLETT-PACKARD

HP-85

REGRESSION ANALYSIS PAC

HEWLETT PACKARD

HP-83/85
Regression Analysis Pac

September 1980

00085-90472

Printed in U.S.A.

Introduction

The regression procedures that have been included in this collection of programs should be an important tool for you in determining whether an appropriate multiple linear model exists between a set of independent variables and a dependent variable. We have included three distinct programs: Stepwise Selection Procedure, Multiple Regression, and Polynomial Regression. All three programs assume that the operator has previously stored the data using the Basic Statistics and Data Manipulation routines.

The programs included in the stepwise procedure actually include four model building algorithms. The most popular is the stepwise selection algorithm. However, we have included the backward and forward algorithm as well. Actually, the procedure we use most frequently is the manual selection procedure, which allows the user to decide the variables to include or delete at each step. With a little experience, you will find that these procedures are useful in selecting appropriate variables for your regression model.

The multiple regression procedure allows you to obtain the regression coefficients, the analysis of variance, etc., for a model that you specify. This algorithm uses the Cholesky square-root procedure, which is the most accurate and efficient procedure available for use on desktop computers.

The polynomial regression program allows you to develop a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p.$$

Even though the algorithm used here is the Cholesky procedure, we caution the operator to use realistic values for p, or the computational accuracy may be such that the program will inform the operator to select a lower degree. Keep in mind that the X values must be raised to the 2p power (X^{2p}) in the computation of the estimates for β_i. Hence, if the original X has several significant digits, raising X to the 2p power may be computationally impossible. Conclusion: Use only realistic values for p depending on your data set and plot the data first to see what values of p make sense for your data.

All three of the programs discussed above use a residual analysis routine which can also plot the standardized residuals. We strongly suggest that you study the residuals from any regression model you develop in order to "see" the adequacy of this model.

Hewlett-Packard would like to acknowledge the work of Thomas J. Boardman, Ph.D., Statistical Laboratory, Colorado State University, in the development of this pac.

Contents

Program Usage 6
- General 6
- Program Flow 8
- Multiple Linear Regression 8
- Stepwise Regression 9
- Polynomial Regression 11
- Residual Analysis 11
- Examples 14

Appendix A: Limitations 38
Appendix B: Data File Configuration 40
Appendix C: Program Documentation 41
Appendix D: Using the 7225A Plotter 42
Appendix E: Using the Disc Version 43

Program Operation Hints

These programs have been designed to execute with a minimum of difficulty, but problems may occur which you can easily solve during program operation. There are four different types of errors or warnings that can occur while executing a program; input errors, math errors, tape errors and image format string errors.

The input errors include errors 43, 44, and 45. These errors will cause a message to be output followed by a new question mark as a prompt for the input. You should verify your mistake and then enter the correct input. The program will not proceed until the input is acceptable. There is a complete discussion of INPUT in your Owner's Manual if you need more detail.

The second type of error which might occur is a math error (1 thru 13). With DEFAULT ON, the first eight errors listed in Appendix E of your Owner's Manual cause a warning message to be output, but program execution will not be halted. The cause of these errors can usually be attributed to specific characteristics of your data and the type of calculations being performed. In most cases, there is no cause for alarm, but you should direct your attention to a possible problem. An example of such a case is found in the Standard Pac when the curve fitting program computes a curve fit to your data which has a value of 1 for the coefficient of determination, r^2. The computation of the F ratio results in a divide by zero, Warning 8.

The third type of error, tape errors (60 thru 75) may be due to several different problems. Some of the most likely causes are the tape being write-protected, the wrong cartridge (or no cartridge) being inserted, a bad tape cartridge, or wrong data file name specification during program execution. Appendix E of your Owner's Manual should be consulted for a complete listing.

The fourth type of error is due to generalizing the output to anticipated data ranges. In many cases, the output has assumed ranges which may or may not be appropriate with your data. Adjusting the image format string for your data will solve this type of problem. You may also want to change the image string if you require more digits to the right of the decimal point.

These are the more common problems which may occur during program operation. Your Owner's Manual should be consulted if you need more assistance.

Two versions of the program have been designed to run specifically on either a tape or a disc. The operation of the disc version is explained in Appendix E of this manual.

Program Usage

General

The regression package is made up of three regression routines–a multiple linear regression, a regression routine incorporating various variable selection procedures, and a polynomial regression routine. A residual analysis routine may be accessed upon completion of any of the three regression programs.

The multiple linear regression routine performs a least-squares regression on a set of predetermined variables. The variable selection program performs regressions iteratively on a set of variables determined by one of four selection procedures–stepwise, forward, backward, or manual. The polynomial regression routine builds a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p$$

where the degree of the regression is chosen by the user with the aid of a preliminary analysis of variance table and, if desired, an X-Y scatter plot. All of the programs provide an analysis of variance table, correlations, and the regression coefficients, as well as their standard errors.

The residual analysis routine provides a list of the residuals as well as the plot of the standardized residuals if desired.

Special Considerations

Data Matrix Configuration

The data matrix incorporated in this program should be thought of as a p-by-n array whose columns correspond to observations and whose rows correspond to variables as shown below.

	OBSERVATIONS
VARIABLES	O_1 O_2 O_3 ... O_n
V_1	...
V_2	...
V_3	...
.	
.	
.	
V_p	...

Subfiles may be created, in which case the structure becomes only slightly more complex as shown below.

VARIABLES	SUBFILE 1 $O_1O_2...O_{n_1}$	SUBFILE 2 $O_{n_1+1}O_{n_1+2}...O_{n_1+n_2}$	...	SUBFILE S $O_{n_1+\cdots+n_{s-1}+1}...O_{n_1+\cdots+n_s}$
V_1				
V_2				
.				
.				
.				
V_p				

Missing Values

The number used to designate a missing value is −999999.9999. The justification for this number is that (besides seeming unlikely to occur as a legitimate data point) it is easily picked out in a listing of the data. It may be more desirable to designate a missing value by a more easily typed number, for example, by 0 if zero is not a legitimate data point. The zeros could then be converted to the missing value recognized by the programs. This may be accomplished in two ways. First, during the input procedure, answer "NO" to the prompt, "Are missing values denoted by −999999.9999?" Then input the value used to denote missing values. Or, in the second case, the transformation program contains a missing value option. A value denoting a missing value can be specified for any or all variables. Any observation having the denoted value is transformed to the value −999999.9999.

Incorrect Responses

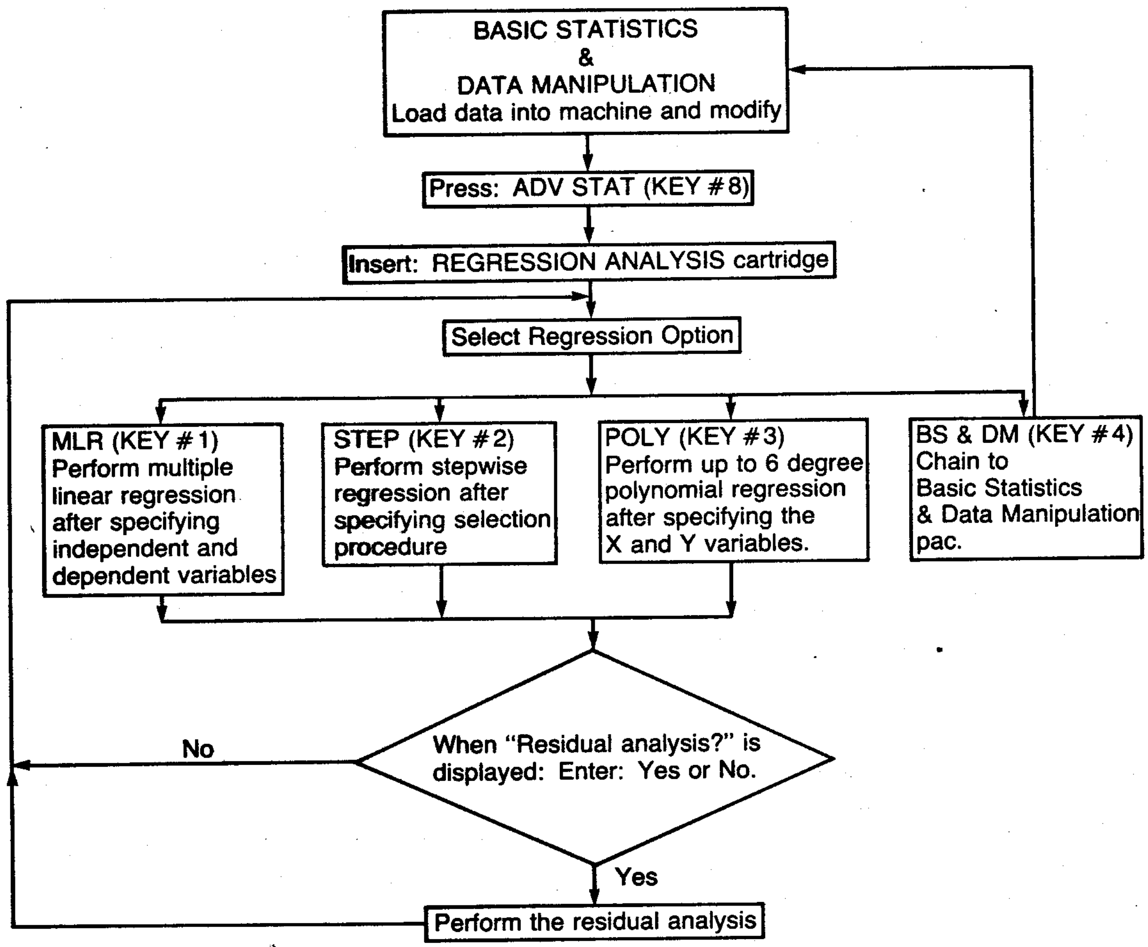
If a response outside the range of plausible responses is input from the keyboard, a message so stating will be displayed for about three seconds. Program execution is resumed by asking the question or a previous question again.

If a plausible response is given, but yet one which is not correct from the user's standpoint, one of three possibilities exist. First, if an incorrect value has been entered for a data point, it may be corrected in the EDIT program. Second, in many cases, responses to several questions are printed on the CRT and then a question such as "Is the above information correct?" is asked. This allows any of the printed information to be changed. Lastly, if a YES/NO question is incorrectly answered or if the above options are not offered, the program can be restarted by pressing KEY LABEL and the soft key for the procedure you want.

Memory Size

Many of the programs in the Regression Analysis Pac take the entire 16K memory. Thus, if you have a 16K machine, all ROMs must be removed from the HP-85 before attempting to run any regression problems. This means that unless you have a 32K machine, regression graphics are only available on the CRT since the Printer/Plotter ROM and other ROMs take a small portion of the user's memory.

Program Flow



Multiple Linear Regression

This program is designed to perform a least-squares multiple linear regression on a predetermined set of variables.

Several basic statistics, as well as the correlation matrix, are output. An analysis of variance table is printed. The regression coefficients and their standard errors are output and confidence intervals are constructed about them. In addition, a residual analysis may be performed.

Special Considerations

Method of Computing the Sums of Squares and Cross-Products Matrix

If a data value is missing for one or more variables, the entire observation is not used in computing the sums of squares and cross-products matrix (and correlations). Hence, in the following matrix where missing values are denoted by M,

OBSERVATION	VARIABLE		
	1	2	3
1	M	3	2
2	1	3	4
3	2	2	3
4	M	4	M
5	1	3	3

Observation 1 is omitted since the data value is missing for variable 1 and observation 4 is omitted since the data value is missing for variables 1 and 3. Hence, only observations 2, 3, and 5 will be used to compute the sums of squares and cross-products matrix as well as the correlations.

Methods and Formulae

The Cholesky square-root method is used to factor the sum of squares and cross-products matrix. It is felt that this method produces less round-off error than other inversion techniques. This method, as well as all other methods and formulae used may be found in F.A. Graybill's *Theory and Application of the Linear Model*. Duxbury Press, 1976.

Stepwise Regression

This program allows a regression model to be built iteratively using one of four variable selection procedures. The procedures are stepwise, forward, backward, and manual. A correlation matrix is calculated and output. An analysis of variance table, as well as partial correlations, F values for deletion and inclusion, and the regression coefficients are output at each step of the regression. In addition, a residual analysis may be performed.

The four selection procedures operate as follows:

Stepwise– The user inputs an F-to-enter and an F-to-delete, and the program begins with no variables in the model. If any of the variables has an F value larger than the F-to-enter, then that variable with the largest F value enters the model. This process is repeated with the remaining variables. At this point, the F values of the variables in the model are compared with the F-to-delete. If a variable has a smaller F value than the F-to-delete, it is removed from the model. This process of adding and deleting variables continues until the F values of all the variables in the model have F values larger than the F-to-delete and all the variables not in the model have F values smaller than the F-to-enter, or until the tolerance value becomes too small (i.e., the matrix becomes unstable).

Forward - The user inputs an F-to-enter. The program operates in the same manner as the stepwise selection procedure, except that variables are not deleted. The process continues until all variables not in the model have F values smaller than the F-to-enter, or until the tolerance value becomes too small.

Backward - The user inputs an F-to-delete and the program begins with all the variables in the model. If any variable has an F value smaller than the F-to-delete, then that with the smallest F value is deleted from the model. This process continues until all the variables in the model have F values larger than the F-to-delete or until the tolerance value becomes too small.

Manual - As the name implies, variables are added or deleted manually until the user is satisfied with the model.

Special Considerations

If one of the stepwise, forward, or backward procedures are used in the selection of variables, the program will proceed automatically by entering and/or removing variables from the model until the F values are insufficient for further computation or until the tolerance value is not met. At this point the program reverts to the manual mode. For example, this allows the user to enter a variable whose F value is just slightly less than the specified F-to-enter.

Methods of Computing Correlations

Two methods of computing correlations are available. The first method will use an observation only if data values are present for each variable. The second method uses all possible data values to compute each correlation. If no missing values are present, method two should be used to speed computation.

A simple example will show the difference between the two methods. Suppose we have the following data set:

OBSERVATION	VARIABLE		
	1	2	3
1	2	3	M
2	3	2	4
3	1	3	5
4	M	1	4

If method one is used to compute the correlations, only observations 2 and 3 will be used. Observation 1 will be omitted since the data value is missing for variable 3. Similarly, observation 4 will be omitted since the data value is missing for variable 2.

Conversely, suppose method two is chosen. The correlation between variables 1 and 2 will be computed using the data values of observations 1, 2, and 3. The correlation between variables 1 and 3 will use data values associated with observations 2 and 3. Similarly, the correlation between variables 2 and 3 will use data values associated with observations 2, 3, and 4. Hence, data values from a given observation are used if the data points are present for the two variables under consideration.

Methods and Formulae

All methods and formulae used in this program may be found in *Statistical Methods for Digital Computers* by K. Enslein, et.al., John Wiley and Sons, 1977.

Polynomial Regression

This program is designed to build a polynomial regression model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p$$

where $p <= 6$ and the β's are computed via the Cholesky method.

The degree of the regression, p, is chosen by the user with the aid of a "preliminary" analysis of variance table and, if desired, an X-Y scatter plot. The preliminary analysis of variance table shows the additional sum of squares explained by models of successive degrees as well as the associated F values and R-squared values.

After the degree of the regression is selected, an analysis of variance table for the model is printed, the regression coefficients and their standard errors are printed and confidence intervals are constructed about the coefficients. In addition, a residual analysis may be performed.

Special Considerations

Degree of Model

The maximum degree of the model has been set (somewhat arbitrarily) at 6. Models of degree six involve arithmetic operations using ΣX^{12} where X is the independent variable. Hence substantial round-off errors may occur with models of high degree. In general, a model of degree p will involve numbers of magnitude ΣX^{2p}. It is, therefore, suggested to use extreme caution in choosing the degree of the model.

Method of Computing Sums of Squares and Cross-Products Matrix

If a data value is missing for one or more variables, the entire observation is deleted, i.e., not used in the computation of sums of squares and cross products. See Special Considerations for the MULTIPLE LINEAR REGRESSION routine for an example.

Residual Analysis

This program allows the user to analyze the residuals from a regression problem in order to check the adequacy of the regression model. The residuals may be printed and/or plotted.

The residual printout includes the observed value, predicted value, residual, and standardized residual. If the standardized residual is between two and three standard deviations away from zero, an asterisk will be printed beside the standardized residual. If the standardized residual is more than three standard deviations away from zero two asterisks will be printed. The Durbin-Watson statistic is output after the above is printed. The statistic is a measure of correlation among the residuals.

The residual plot allows the user to plot the standardized residuals versus time or versus any of variables in the model.

Special Considerations

The standardized residuals are plotted in a range from -5 to 5. If any standardized residuals are outside this range they will not be plotted, but a note showing the number off scale will be added to the graph.

Methods and Formulae

Suppose the model has been determined by one of the regression routines and is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p.$$

We will refer to the nth predicted Y as Y(n), the nth residual as R(n), etc. Let D(I,J) be the Jth observation of the Ith variable in the data matrix.

- Predicted Y: $\hat{Y}(n) = \hat{\beta}_0 + \hat{\beta}_1 D(X_1,n) + \ldots + \hat{\beta}_p D(X_p,n)$
- Residual: $R(n) = D(Y,I) - \hat{Y}(n)$
- Standard error of residuals:
  S8 = residual mean square
- Standardized residual: $S9(n) = R(n)/\sqrt{S8}$

The residual mean square is calculated in the regression routine.

Examples

```
MLR

******************************
*                            *
*     DATA MANIPULATION      *
*                            *
******************************

          MLR EXAMPLE
Data file name: EX-MLR
Number of obs: 9
Number of variables:  3
Variable names:
   1.  X1
   2.  X2
   3.  Y

Subfiles:  NONE

******************************
*                            *
*    DATA TRANSFORMATIONS    *
*                            *
******************************

The following transformation was
performed: a*(X^b)+c
where X is Variable # 1
      a =  1
      b =  2
      c =  0
Transformed data is stored in
Variable # 4 (X1^2  ).

The following transformation was
performed: a*(X^b)+c
where X is Variable # 2
      a =  1
      b =  2
      c =  0
Transformed data is stored in
Variable # 5 (X2^2  ).

The following transformation was
performed: a*(X^b)*(Y^c)
where X is Variable # 1
      Y is Variable # 2
      a =  1
      b =  1
      c =  1
```

This example will show most of the features of the MLR program.

We are using the transformation to create:

$X_4 = X1 \wedge 2$	quadratic X_1 term
$X_5 = X2 \wedge 2$	quadratic X2 term
$X_6 = X1*X2$	linear by linear interaction term

Data listing of the six variables

```
Transformed data is stored in
Variable # 6 (X1*X2 ).

*******************************
*        DATA LISTING         *
*        ON DATA SET:         *
*         MLR EXAMPLE         *
*******************************
```

First four variables

```
          X1              X2
             Y               X1^2
OBS#
 1        7.8000          4.0000
             0.0000          60.8400

 2        7.8000          8.0000
              .0310          60.8400

 3        7.8000         12.0000
              .4750          60.8400

 4       39.0000          4.0000
              .0160        1521.0000

 5       39.0000          8.0000
              .0080        1521.0000

 6       39.0000         12.0000
              .1900        1521.0000

 7       78.0000          4.0000
             0.0000        6084.0000

 8       78.0000          8.0000
              .0390        6084.0000

 9       78.0000         12.0000
             0.0000        6084.0000
```

Last two variables

```
          X2^2            X1*X2
OBS#
 1       16.0000         31.2000

 2       64.0000         62.4000

 3      144.0000         93.6000

 4       16.0000        156.0000

 5       64.0000        312.0000

 6      144.0000        468.0000
```

7	16.0000	312.0000
8	64.0000	624.0000
9	144.0000	936.0000

```
********************************
*      SUMMARY STATISTICS      *
*         ON DATA SET:         *
*          MLR EXAMPLE         *
********************************
```

BASIC STATISTICS

Var. Names	# of Obs.	# of Missing
X1	9	0
X2	9	0
Y	9	0
X1^2	9	0
X2^2	9	0
X1*X2	9	0

Basic statistics on all six variables.

Var. Names	Mean	Std. Dev.
X1	41.6000	30.4600
X2	8.0000	3.4641
Y	.0843	.1583
X1^2	2555.2800	2721.0176
X2^2	74.6667	56.0000
X1*X2	332.8000	300.0720

Var. Names	Std.Error	Coef of Variation
X1	10.1533	73.2211
X2	1.1547	43.3013
Y	.0528	187.7295
X1^2	907.0059	106.4861
X2^2	18.6667	75.0000
X1*X2	100.0240	90.1659

Var. Names	Coef of Skewness	Coef of Kurtosis
X1	.1351	-1.5000
X2	0.0000	-1.5000
Y	1.9377	2.2910
X1^2	.5392	-1.5000
X2^2	.2948	-1.5000
X1*X2	.8842	-.2633

95% CONFIDENCE INTERVAL ON MEAN

Var. Names	Lower Limit	Upper Limit
X1	18.1801	65.0199
X2	5.3365	10.6635
Y	-.0374	.2061
X1^2	463.1579	4647.4021
X2^2	31.6097	117.7237
X1*X2	102.0822	563.5178

CORRELATION MATRIX

	X2	Y	X1^2
X1	0.0000	-.4209	.9748
X2		.5917	0.0000
Y			-.3905

	X2^2	X1*X2
X1	0.0000	.8121
X2	.9897	.4802
Y	.6251	-.2314
X1^2	0.0000	.7916
X2^2		.4753

We would expect that X1 and X1 ^ 2 should be highly correlated (.9748).

ORDER STATISTICS

Var. Names	Maximum	Minimum
X1	78.0000	7.8000
X2	12.0000	4.0000
Y	.4750	0.0000
X1^2	6084.0000	60.8400
X2^2	144.0000	16.0000
X1*X2	936.0000	31.2000

Var. Names	Range	Midrange
X1	70.2000	42.9000
X2	8.0000	8.0000
Y	.4750	.2375
X1^2	6023.1600	3072.4200
X2^2	128.0000	80.0000
X1*X2	904.8000	483.6000

```
Var.
Names          Median
X1            39.0000
X2             8.0000
Y               .0160
X1^2        1521.0000
X2^2          64.0000
X1*X2        312.0000

Var.
Names         25-th %        75-th %
X1             7.8000        39.0000
X2             4.0000         8.0000
Y              0.0000          .0310
X1^2          60.8400      1521.0000
X2^2          16.0000        64.0000
X1*X2         93.6000       312.0000
```

```
********************************
*  MULTIPLE LINEAR REGRESSION  *
*          ON DATA SET         *
*          MLR EXAMPLE         *
********************************

Dependent variable : Y
Independent variable(s) : X1
                          X2
                          X1^2
                          X2^2
                          X1*X2
```

MLR with dependent variable of Y = X_3. Certain basic statistics are output.

VARIABLE	N	MEAN
X1	9	41.60000
X2	9	8.00000
X1^2	9	2555.28000
X2^2	9	74.66667
X1*X2	9	332.80000
Y	9	.08433

VARIABLE	STANDARD DEVIATION	COEF. OF VARIATION
X1	30.45997	73.2211
X2	3.46410	43.3013
X1^2	2721.01756	106.4861
X2^2	56.00000	75.0000
X1*X2	300.07199	90.1659
Y	.15832	187.7295

CORRELATION MATRIX

	X2	X1^2	X2^2
X1	0.0000	.9748	0.0000
X2		0.0000	.9897
X1^2			0.0000

	X1*X2	Y
X1	.8121	-.4209
X2	.4802	.5917
X1^2	.7916	-.3905
X2^2	.4753	.6251
X1*X2		-.2314

AOV TABLE

SOURCE	DF	MEAN SQUARE	F-VALUE
TOTAL	8		
REGR.	5	.03554	4.67
X1	1	.03553	4.67
X2	1	.07020	9.23
X1^2	1	.00158	.21
X2^2	1	.01531	2.01
X1*X2	1	.05507	7.24
RESID	3	.00761	

R-SQUARED = .886151704515
STD. ERROR OF EST. = .08723

The AOV table with all five independent variables. The "partial" F statistics show the additional contribution of each variable (X1, X2, X1 ^ 2, etc.) given the previous variables. The five variables account for 88.6% of the variation in Y. Not bad for a simple example with n = 9.

REGRESSION COEFFICIENTS

VAR.	STD. FORMAT	STD. ERROR
CONST.	-.00218	.25209
X1	.00247	.00517
X2	-.02576	.06364
X1^2	.00002	.00005
X2^2	.00547	.00386
X1*X2	-.00083	.00031

VAR.	E-FORMAT	T-VALUE
CONST.	-2.181542194E-003	-.01
X1	2.469641773E-003	.48
X2	-2.576434426E-002	-.40
X1^2	2.313292911E-005	.46
X2^2	5.468750000E-003	1.42
X1*X2	-8.339901219E-004	-2.69

The coefficients of the regression equation are shown in two formats:

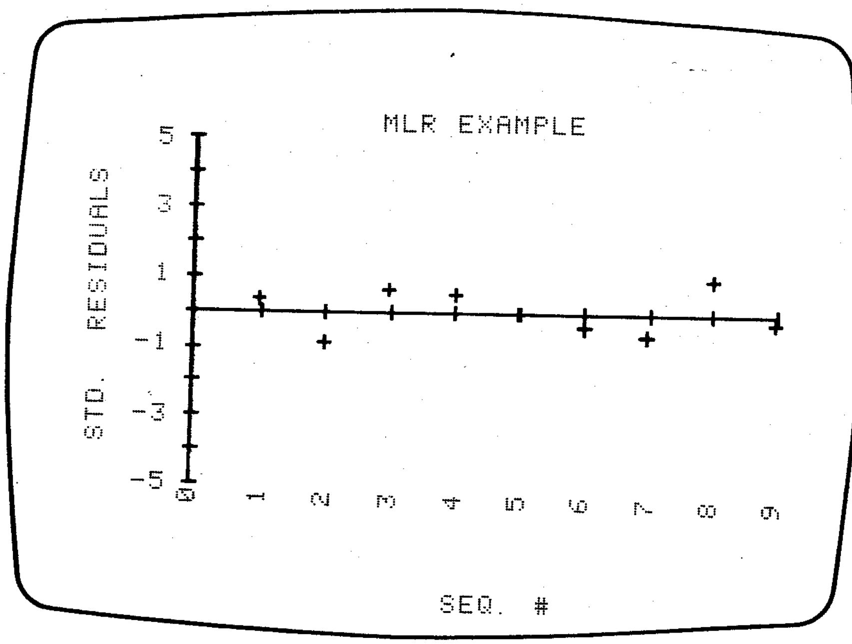
$$\hat{y} = -\ .00218 + .00247X1 - .02576X2$$

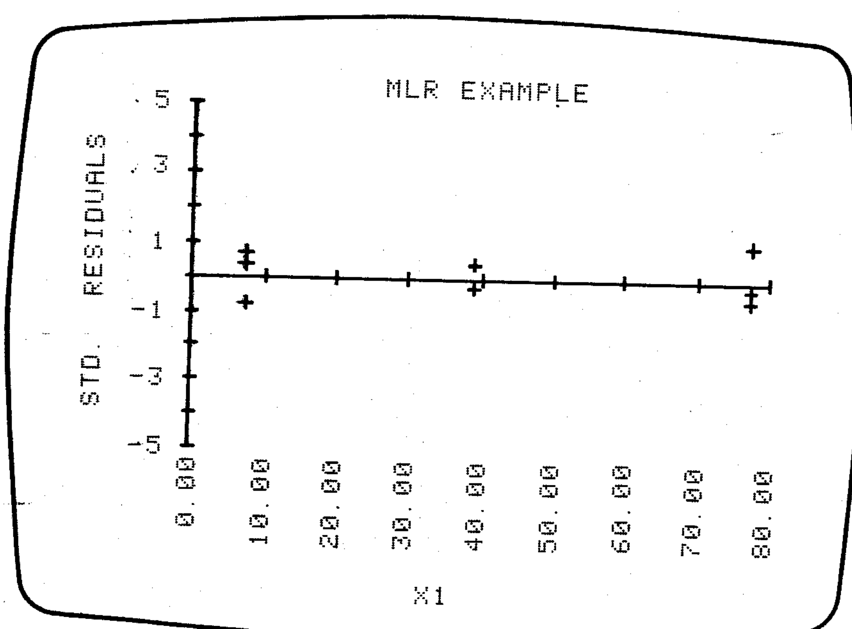
$$+ -\ .00083X1*X2.$$

```
        95 % CONFIDENCE INTERVAL.
VAR.    LOWER LIMIT  UPPER LIMIT
CONST.      -.80382       .79946
X1          -.01397       .01891
X2          -.22814       .17661
X1^2        -.00014       .00018
X2^2        -.00679       .01773
X1*X2       -.00182       .00015


********************************
*                              *
*      RESIDUAL ANALYSIS       *
*                              *
********************************


OBS#     Observed Y   Predicted Y
   1       0.00000       -.02309
   2        .03100        .11033
   3        .47500        .41876
   4        .01600       -.01634
   5        .00800        .01300
   6        .19000        .21734
   7       0.00000        .05543
   8        .03900       -.04533
   9       0.00000        .02890
```

Residual analysis is a useful diagnostic tool. The standardized residuals exhibit no large values.

```
OBS#        Residual    Std.Res.
   1         .02309      .26468
   2        -.07933     -.90944
   3         .05624      .64476
   4         .03234      .37073
   5        -.00500     -.05732
   6        -.02734     -.31342
   7        -.05543     -.63541
   8         .08433      .96676
   9        -.02890     -.33135


Durbin-Watson stat. =  2.8246
```


MLR EXAMPLE
STD. RESIDUALS
5
3
1
-1
-3
-5
0
1
2
3
4
5
6
7
8
9
SEQ. #


MLR EXAMPLE
STD. RESIDUALS
5
3
1
-1
-3
-5
0.00
10.00
20.00
30.00
40.00
50.00
60.00
70.00
80.00
X1

Stepwise Regression Example

```
STEP

*******************************
*                             *
*      DATA MANIPULATION      *
*                             *
*******************************

           MLR EXAMPLE
Data file name: DATA
Number of obs: 9
Number of variables:  6
Variable names:
   1.  X1
   2.  X2
   3.  Y
   4.  X1^2
   5.  X2^2
   6.  X1*X2

Subfiles:  NONE
```

This is the same data set as we used for the MLR example.

```
*******************************
*     STEPWISE  REGRESSION    *
*         ON DATA SET         *
*          MLR EXAMPLE        *
*******************************

Dependent variable : Y
Independent variable(s) : X1
                          X2
                          X1^2
                          X2^2
                          X1*X2

Tolerance = .01
F-value for inclusion = 4
F-value for deletion = 4

       CORRELATION MATRIX

          X1       X2       X1^2

X1       1.0000   0.0000    .9748
X2                1.0000   0.0000
X1^2                       1.0000
```

```
            X2^2      X1*X2     Y

X1        0.0000     .8121   -.4209
X2         .9897     .4802    .5917
X1^2      0.0000     .7916   -.3905
X2^2      1.0000     .4753    .6251
X1*X2               1.0000   -.2314
Y                            1.0000
```

```
*** VARIABLES IN REGRESSION ****
           REG. COEF.
VAR.      STD. FORMAT   STD. ERROR
CONST          .08433
                              F TO
VAR.    REG COEF E-FORMAT   DELETE
CONST   8.433333333E-002

CONST = MEAN OF DEP. VAR.
```

At step 0, before any variables are in the regression, this coefficient is the overall mean for Y.

```
* VARIABLES NOT IN REGRESSION **
               F TO      PART
VAR.           ENTER     CORR      TOL
X1              1.51    .4209   1.000
X2              3.77    .5917   1.000
X1^2            1.26    .3905   1.000
X2^2            4.49    .6251   1.000
X1*X2            .40    .2314   1.000
```

Note that variable X_5 = X2 ∧ 2, the quadratic effect of X2 will be the first to enter the regression.

```
*********************************
STEP NUMBER 1
VARIABLE ADDED : X2^2

            AOV TABLE
SOURCE   DF   MEAN SQUARE   F-VALUE
TOTAL     8
REGR.     1        .07835      4.49
RESID.    7        .01745

R-SQUARED =  .390745163902
STD. ERROR OF EST. =       .13211
```

This is confirmed at step 1.

```
*** VARIABLES IN REGRESSION ****
           REG. COEF.
VAR.      STD. FORMAT   STD. ERROR
X2^2           .00177       .00083
CONST        -.04762
                              F TO
VAR.    REG COEF E-FORMAT   DELETE
X2^2     1.767219388E-003     4.49
CONST   -4.761904762E-002
```

```
* VARIABLES NOT IN REGRESSION **
              F TO      PART
VAR.          ENTER     CORR      TOL
X1             2.46     .5393   1.000
X2              .37     .2421    .020
X1^2           2.00     .5003   1.000
X1*X2          8.72     .7696    .774


*******************************
STEP NUMBER 2
VARIABLE ADDED : X1*X2

           AOV TABLE
SOURCE   DF   MEAN SQUARE   F-VALUE
TOTAL     8
REGR.     2        .07536      9.08
RESID.    6        .00830

R-SQUARED =   .75163029247
STD. ERROR OF EST. =       .09111


*** VARIABLES IN REGRESSION ****
          REG. COEF.
VAR.      STD. FORMAT   STD. ERROR
X2^2          .00268        .00065
X1*X2       -.00036         .00012
CONST         .00376
                              F TO
VAR.    REG COEF E-FORMAT   DELETE
X2^2     2.684743302E-003    16.86
X1*X2   -3.602457676E-004     8.72
CONST    3.762291577E-003

* VARIABLES NOT IN REGRESSION **
              F TO      PART
VAR.          ENTER     CORR      TOL
X1             4.71     .6963    .148
X2              .45     .2861    .020
X1^2           4.53     .6895    .190


********************************
STEP NUMBER 3
VARIABLE ADDED : X1

           AOV TABLE
SOURCE   DF   MEAN SQUARE   F-VALUE
TOTAL     8
REGR.     3        .05829     11.36
RESID.    5        .00513

R-SQUARED =   .872061969697
STD. ERROR OF EST. =       .07163
```

Variable X_6 = X1 * X2, the linear by linear interaction, is the next variable to enter, since its F is >=4, our specified value, and it has an F larger than the rest.

```
*** VARIABLES IN REGRESSION ****
          REG. COEF.
VAR.     STD. FORMAT   STD. ERROR
X1             .00469      .00216
X2^2           .00396      .00078
X1*X2        -.00086       .00025
CONST        -.12004
                            F TO
VAR.   REG COEF E-FORMAT   DELETE
X1      4.687491529E-003    4.71
X2^2    3.956117661E-003   25.76
X1*X2  -8.594232003E-004   11.89
CONST  -1.200403919E-001
```

```
* VARIABLES NOT IN REGRESSION **
              F TO      PART
VAR.          ENTER     CORR     TOL
X2              .20    .2205    .020
X1^2            .26    .2480    .050
```

After 3 steps, the model involves X1, X2 ^ 2, and X1*X2, plus, of course, the intercept = const. The R^2 = .87 for these 3 terms.

```
Tolerance value too small and/or
F-values insufficient to proceed
```

```
********************************
*     BACKWARD  REGRESSION     *
*         ON DATA SET          *
*          MLR EXAMPLE         *
********************************
```

In order to confirm the stepwise model selection, many data analyses suggest using the backward elimation procedure.

```
Dependent variable : Y
Independent variable(s) : X1
                          X2
                          X1^2
                          X2^2
                          X1*X2
```

```
Tolerance = .01
F-value for deletion = 4
```

```
       CORRELATION MATRIX

          X1        X2        X1^2

X1       1.0000    0.0000     .9748
X2                 1.0000    0.0000
X1^2                         1.0000
```

	X2^2	X1*X2	Y
X1	0.0000	.8121	-.4209
X2	.9897	.4802	.5917
X1^2	0.0000	.7916	-.3905
X2^2	1.0000	.4753	.6251
X1*X2		1.0000	-.2314
Y			1.0000

AOV TABLE

SOURCE	DF	MEAN SQUARE	F-VALUE
TOTAL	8		
REGR.	5	.03554	4.67
RESID.	3	.00761	

R-SQUARED = .886151704527
STD. ERROR OF EST. = .08723

*** VARIABLES IN REGRESSION ****

VAR.	REG. COEF. STD. FORMAT	STD. ERROR
X1	.00247	.00517
X2	-.02576	.06364
X1^2	.00002	.00005
X2^2	.00547	.00386
X1*X2	-.00083	.00031
CONST	-.00218	

VAR.	REG COEF E-FORMAT	F TO DELETE
X1	2.469641773E-003	.23
X2	-2.576434427E-002	.16
X1^2	2.313292912E-005	.21
X2^2	5.468750001E-003	2.01
X1*X2	-8.339901219E-004	7.24
CONST	-2.181542172E-003	

****
STEP NUMBER 1
VARIABLE DELETED : X2

AOV TABLE

SOURCE	DF	MEAN SQUARE	F-VALUE
TOTAL	8		
REGR.	4	.04411	7.33
RESID.	4	.00602	

R-SQUARED = .87993229594
STD. ERROR OF EST. = .07758

```
*** VARIABLES IN REGRESSION ****
           REG. COEF.
VAR.       STD. FORMAT   STD. ERROR
X1              .00267       .00458
X1^2            .00002       .00005
X2^2            .00396       .00084
X1*X2          -.00086       .00027
CONST          -.09535
                               F TO
VAR.     REG COEF E-FORMAT   DELETE
X1        2.673106400E-003      .34
X1^2      2.313292912E-005      .26
X2^2      3.956117661E-003    21.96
X1*X2    -8.594232003E-004    10.13
CONST    -9.535308167E-002


* VARIABLES NOT IN REGRESSION **
               F TO     PART
VAR.          ENTER     CORR     TOL
X2              .16    .2276    .020
```

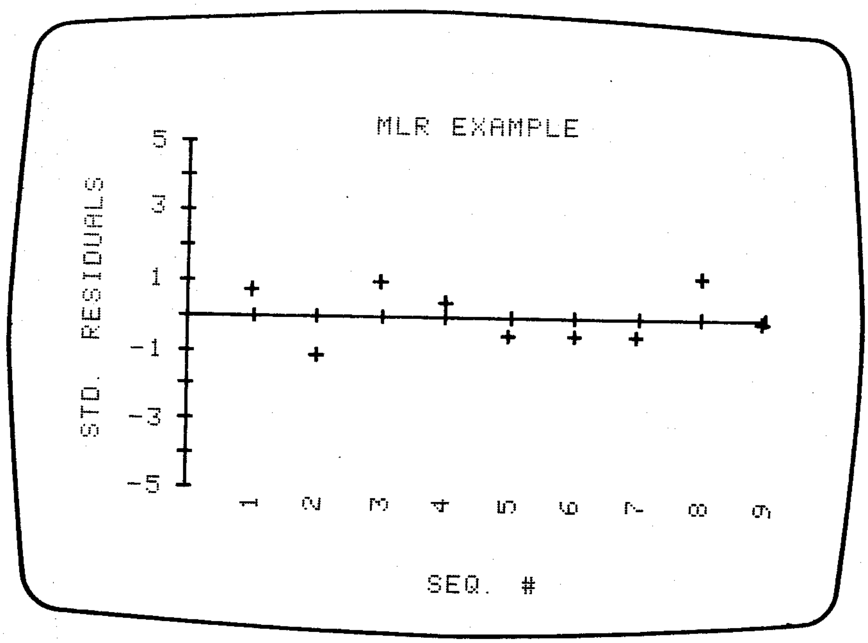
```
********************************
STEP NUMBER 2
VARIABLE DELETED : X1^2

           AOV TABLE
SOURCE   DF   MEAN SQUARE   F-VALUE
TOTAL     8
REGR.     3        .05829     11.36
RESID.    5        .00513

R-SQUARED =   .872061969702
STD. ERROR OF EST. =        .07163
```

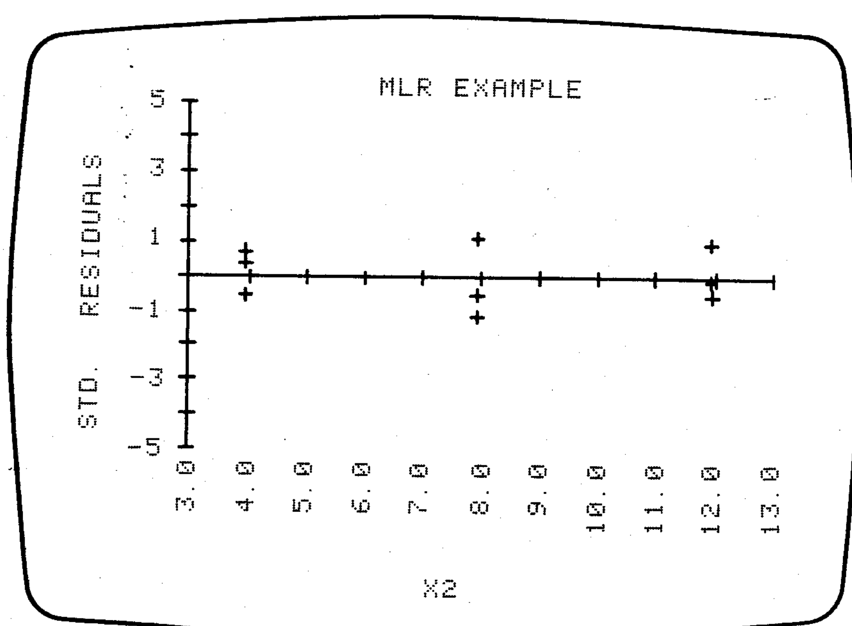
```
*** VARIABLES IN REGRESSION ****
           REG. COEF.
VAR.       STD. FORMAT   STD. ERROR
X1              .00469       .00216
X2^2            .00396       .00078
X1*X2          -.00086       .00025
CONST          -.12004
                               F TO
VAR.     REG COEF E-FORMAT   DELETE
X1        4.687491530E-003     4.71
X2^2      3.956117661E-003    25.76
X1*X2    -8.594232003E-004    11.89
CONST    -1.200403919E-001
```

After several steps, the backward elimination procedure ends up with the same model as the stepwise algorithm. Other data sets may not result in the same confirmation.

```
* VARIABLES NOT IN REGRESSION **
               F TO     PART
VAR.           ENTER    CORR      TOL
X2               .20    .2205     .020
X1^2             .26    .2480     .050


Tolerance value too small and/or
F-values insufficient to proceed


*******************************
*                             *
*     RESIDUAL ANALYSIS       *
*                             *
*******************************


OBS#    Observed Y     Predicted Y
   1       0.00000          -.04699
   2        .03100           .11609
   3        .47500           .40576
   4        .01600          -.00800
   5        .00800           .04782
   6        .19000           .23024
   7       0.00000           .04074
   8        .03900          -.03750
   9       0.00000           .01084


OBS#        Residual    Std.Res.
   1          .04699      .65607
   2         -.08509    -1.18786
   3          .06924      .96663
   4          .02400      .33506
   5         -.03982     -.55596
   6         -.04024     -.56182
   7         -.04074     -.56879
   8          .07650     1.06806
   9         -.01084     -.15140


Durbin-Watson stat. =  2.6802
```

The plots do not show any patterns suggesting that the regression equation on this small data set is adequate.





```
POLY

********************************
*                              *
*      DATA MANIPULATION       *
*                              *
********************************

       POLYNOMIAL EXAMPLE
Data file name: EX-POL
Number of obs: 31
Number of variables:  2
Variable names:
   1.  NUMBER
   2.  TIME

Subfiles:  NONE


********************************
*         DATA LISTING         *
*         ON DATA SET:         *
*      POLYNOMIAL EXAMPLE      *
********************************
```

Polynomial Regression Example

OBS#	NUMBER	TIME
1	1.0000	1.4000
2	1.0000	2.8000
3	1.0000	3.0000
4	1.0000	1.8000
5	1.0000	2.0000
6	2.0000	4.7000
7	2.0000	8.0000
8	2.0000	3.0000
9	2.0000	2.5000
10	3.0000	5.2000
11	3.0000	6.2000
12	3.0000	9.4000
13	4.0000	11.7000
14	5.0000	7.5000

Data listing of X = number of passengers boarding a bus and Y = the number of seconds required to have these people get on the bus (passenger service time).

15	5.0000	11.9000
16	6.0000	13.6000
17	6.0000	12.4000
18	6.0000	11.6000
19	7.0000	14.7000
20	7.0000	13.5000
21	8.0000	12.0000
22	8.0000	14.1000
23	8.0000	26.0000
24	9.0000	19.0000
25	10.0000	21.2000
26	11.0000	22.9000
27	11.0000	22.6000
28	13.0000	25.2000
29	17.0000	33.5000
30	19.0000	33.7000
31	25.0000	54.2000

```
*******************************
*      SUMMARY STATISTICS     *
*        ON DATA SET:         *
*      POLYNOMIAL EXAMPLE     *
*******************************
```

Basic statistics on the data set.

```
        BASIC STATISTICS
        ----------------
Var.               # of        # of
Names              Obs.       Missing
NUMBER              31           0
TIME                31           0

Var.
Names               Mean     Std. Dev.
NUMBER            6.6774       5.7642
TIME             13.9129      11.8068
```

Var. Names	Std.Error	Coef of. Variation
NUMBER	1.0353	86.3235
TIME	2.1206	84.8620

Var. Names	Coef of Skewness	Coef of Kurtosis
NUMBER	1.4313	1.9079
TIME	1.4898	2.5565

95% CONFIDENCE INTERVAL ON MEAN

Var. Names	Lower Limit	Upper Limit
NUMBER	4.5626	8.7922
TIME	9.5811	18.2447

CORRELATION MATRIX

	TIME
NUMBER	.9744

ORDER STATISTICS

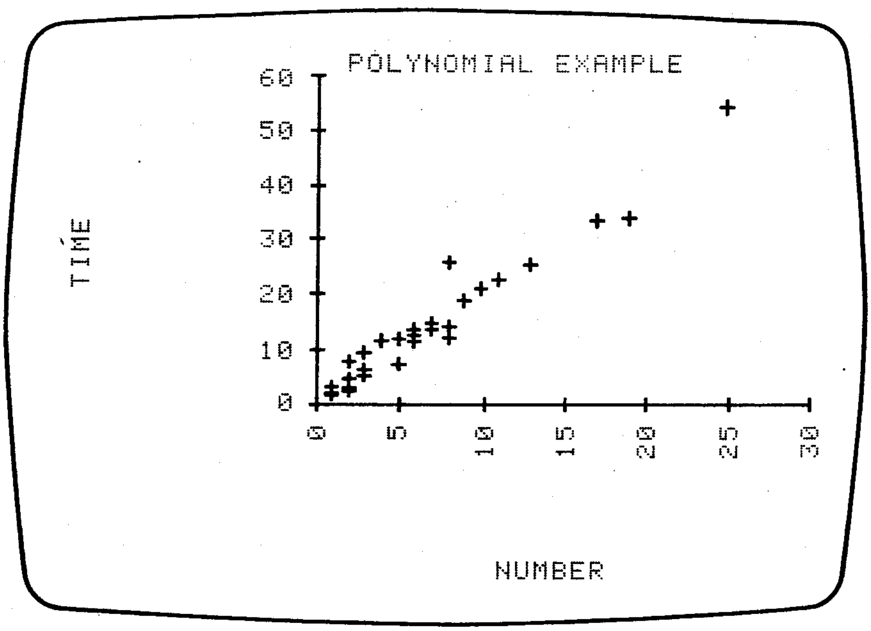
Var. Names	Maximum	Minimum
NUMBER	25.0000	1.0000
TIME	54.2000	1.4000

Var. Names	Range	Midrange
NUMBER	24.0000	13.0000
TIME	52.8000	27.8000

Var. Names	Median
NUMBER	6.0000
TIME	11.9000

Var. Names	25-th %	75-th %
NUMBER	2.0000	8.0000
TIME	4.7000	19.0000

```
********************************
*    POLYNOMIAL REGRESSION     *
*         ON DATA SET          *
*     POLYNOMIAL EXAMPLE       *
********************************

Dependent var. = TIME
Independent var. = NUMBER
```



Scatter plot of X vs. Y.

```
Variable       N            Mean
NUMBER        31         6.67742
TIME          31        13.91290

            Standard      Coef. of
Variable   Deviation     Variation
NUMBER       5.76418       86.3235
TIME        11.80677       84.8620

Correlation =   .974353347879

Selected deg. of regr. =   1
R-squared = .94936444652
Std. error of est. =      2.70222
```

Simple (straight line) linear correlation between X and Y.

Good fit accounting for almost 95% of the variation in the passenger service term, Y.

AOV TABLE

SOURCE	DF	MEAN SQUARE	F-VALUE
TOTAL	30		
REGR.	1	3970.23722	543.72
X^1	1	3970.23722	543.72
RESID.	29	7.30199	

REGRESSION COEFFICIENTS

Var.	Std.Format	E-Format
CONS	.5863	5.86330097E-001
X^1	1.9958	1.99576699E+000

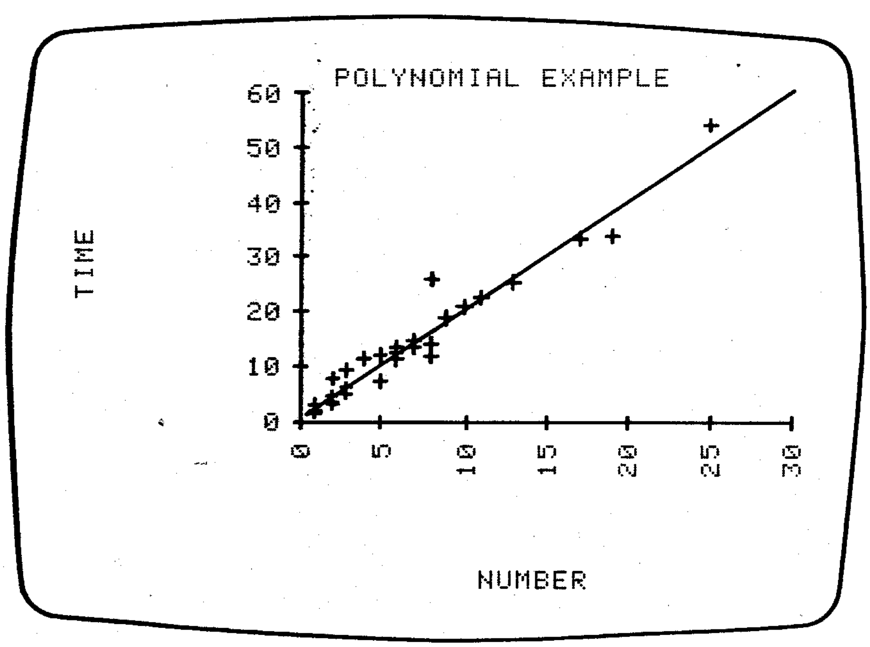
$\hat{y} = .5863 + 1.9958\ X$

Var.	Std.Error of Coef.	T-Value
CONS	.74979	.78
X^1	.08559	23.32

Approximately .6 second start up time (open the doors), plus 2 seconds per passenger.

95 % CONFIDENCE INTERVAL

Var.	Lower Limit	Upper Limit
CONS	-.94752	2.12018
X^1	1.82068	2.17086



Regression line placed on graph.

```
*******************************
*                             *
*      RESIDUAL ANALYSIS      *
*                             *
*******************************
```

Residual analysis.

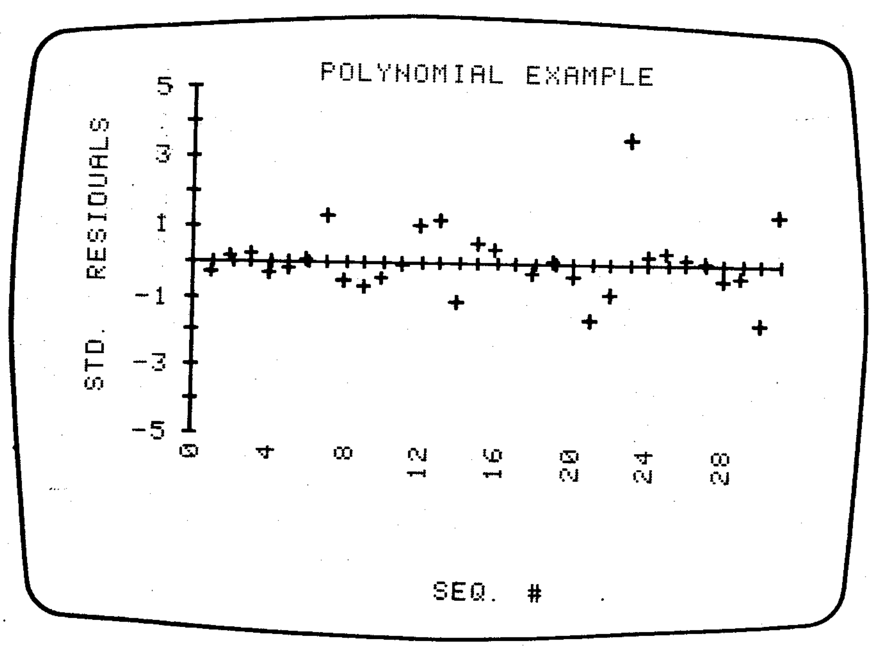
OBS#	Observed Y	Predicted Y
1	1.40000	2.58210
2	2.80000	2.58210
3	3.00000	2.58210
4	1.80000	2.58210
5	2.00000	2.58210
6	4.70000	4.57786
7	8.00000	4.57786
8	3.00000	4.57786
9	2.50000	4.57786
10	5.20000	6.57363
11	6.20000	6.57363
12	9.40000	6.57363
13	11.70000	8.56940
14	7.50000	10.56517
15	11.90000	10.56517
16	13.60000	12.56093
17	12.40000	12.56093
18	11.60000	12.56093
19	14.70000	14.55670
20	13.50000	14.55670
21	12.00000	16.55247
22	14.10000	16.55247
23	26.00000	16.55247
24	19.00000	18.54823
25	21.20000	20.54400
26	22.90000	22.53977
27	22.60000	22.53977
28	25.20000	26.53130
29	33.50000	34.51437
30	33.70000	38.50590
31	54.20000	50.48050

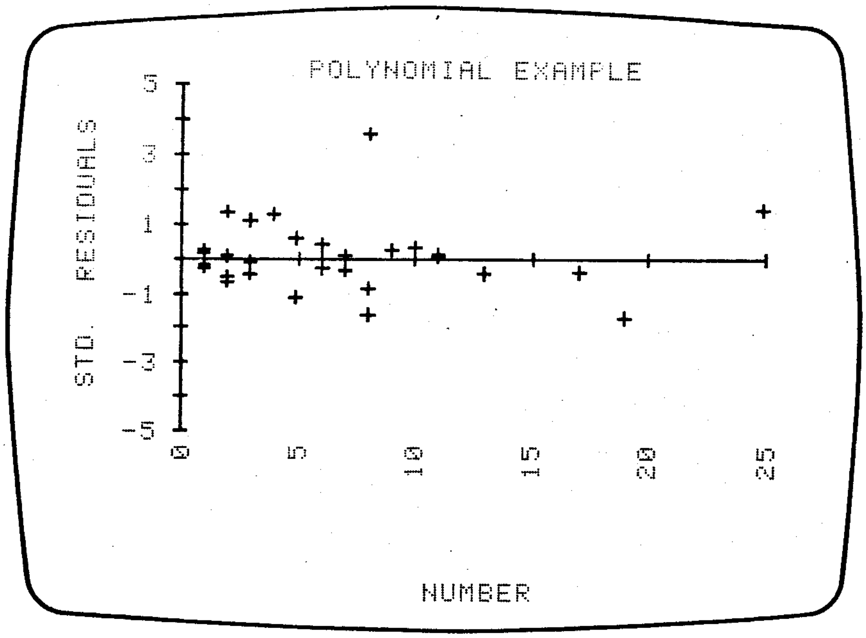
OBS#	Residual	Std.Res.
1	-1.18210	-.43745
2	.21790	.08064
3	.41790	.15465
4	-.78210	-.28943
5	-.58210	-.21541
6	.12214	.04520
7	3.42214	1.26642
8	-1.57786	-.58391
9	-2.07786	-.76895
10	-1.37363	-.50833
11	-.37363	-.13827
12	2.82637	1.04594

13	3.13060	1.15853
14	-3.06517	-1.13431
15	1.33483	.49398
16	1.03907	.38452
17	-.16093	-.05956
18	-.96093	-.35561
19	.14330	.05303
20	-1.05670	-.39105
21	-4.55247	-1.68471
22	-2.45247	-.90757
23	9.44753	3.49621**
24	.45177	.16718
25	.65600	.24276
26	.36023	.13331
27	.06023	.02229
28	-1.33130	-.49267
29	-1.01437	-.37538
30	-4.80590	-1.77850
31	3.71950	1.37646

One point seems out of control, although our original data sheets offer no explanation.

Durbin-Watson stat. = 2.0920




POLYNOMIAL EXAMPLE
STD. RESIDUALS
5
3
1
-1
-3
-5
0
5
10
15
20
25
NUMBER

Appendix A
Limitations

The programs have been designed to operate in the basic machine with a maximum of 500 elements. Hence, for two variables a maximum of 250 observations may be input. This may be changed if more memory is available.

If more than 500 elements are desired a number of changes must be made. All the COM statements containing the array D(?,?) must be changed so that D is dimensioned to D(1,N), where N=maximum number of observations (maximum variables * sample size) desired. The following table gives the location of these COM statements.

File Name	Line
"ADVST"	40
"REENT"	30
"MLR1"	40
"MLR2"	40
"STEP1"	40
"STEP2"	42
"POLY1"	30
"POLY2"	30
"RESID"	30

To increase the maximum number of variables, from 12, V1$[?] must be redimensioned to V1$[M] where M=6*V and V is the number of variables desired. V1$ is located in the COM statements listed above.

In addition, the following lines should also be changed if you want to increase the number of variables:

In "MLR1"
```
60 DIM X(V+1,V+1),V2(V),B(V),C(V),D3$[6],C$[8]
```
In "MLR2"
```
60 DIM X(V+1,V+1),V2(V),B(V),C(V),D3$[6],C$[8]
```
In "STEP1"
```
50 DIM P$[40],V2(V),M(V),V(V),B(V),C(V,V),C$[8],V4(V),V5(V),D2(V)
```
In "STEP2"
```
42 DIM V2(V),M(V),V(V),B(V),C(V,V),V4(V),V5(V),D2(V)
```
In "RESID"
```
50 DIM B(V+1),V2(V+1),R$[36]
```
where V = number of variables.

Also, all COM lines containing E(?) must be changed to E(M) where M= V*(V+1)/2+ V+15 and V=number of variables. If M≤125 this change does not have to be made. These COM statements immediately follow the other

COM statements mentioned above. Remember the E array must also be changed on all files in the "BASIC STATISTICS AND DATA MANIPULATION" cartridge too.

With any change in the limitations, a new "DATA" file must be created. First, purge or rename the old "DATA" file. Then create a new one with the following statement:
CREATE"DATA",2+N*8 DIV M,M
where N=maximum number of observations, M=288+V*6 and V=number of variables.

For instructions on how to modify the "BASIC STATISTICS AND DATA MANIPULATION" cartridge, see the "BASIC STATISTICS AND DATA MANIPULATION" manual.

The REGRESSION ANALYSIS tape cartridge contains two example data sets. "EX-MLR" contains the data used in the multiple linear regression and stepwise regression examples, and "EX-POL" contains the data for the polynomial regression example. The user may wish to page through the manual and try each of the programs available in the pac, then compare the results with those in the examples. It should be noted, however, that each example was run using the original data and not data which had been transformed or edited.

Appendix B
Data File Configuration

The scratch file on the program medium, i.e., ''DATA'', and any files created to hold stored data and related information are configured as follows. The data file is broken into logical records of 300 bytes each. The first logical record is a ''header record'', which contains information pertinent to the data set stored in the remaining logical records. The header record contains the following information (variables): data set title (T$), number of observations (O1), number of variables (N1), variable names (V1$), number of subfiles (S1), subfile names (S1$), and subfile characteristics (S2(*)). The remaining logical records contain D(* *) -- the data matrix.

Appendix C

Program Documentation

The documentation for the Regression Analysis Pac is contained in the DOCRG1 and DOCRG2 programs. The major variables are defined in addition to comments for major sections of code. To obtain the documentation, load and run the program.

Appendix D

Using the 7225A Plotter

As noted in Program Usage, regression graphics on the 7225A requires a 32K machine. The programs have been designed to do all graphs on the CRT, but by changing the programs as noted, this pac can be set up to plot the various graphs on the 7225A. Each program to be changed must first be loaded and converted by executing the TRANSLATE command. After performing the TRANSLATE command, make the noted changes and then store the revised program. Three programs need to be changed to take advantage of the 7225A Plotter.

The new lines are shown for each program as well as the lines which must be deleted.

Program: POLY1

```
1050 DISP "Prepare plotter & press 'CONT' when ready." @ PAUSE

1055 PLOTTER IS 705 @ CSIZE 7 @ DEG @ LORG 1

1480 LDIR 0 @ LORG 5 @ CSIZE 3

1500 MOVE D(X,I),D(Y,I)

1530 BEEP @ DISP "PLOT COMPLETE" @ PAUSE

1550 CLEAR @ DISP "Proceed with regression(Type 'E' to Exit)";@ INPUT N$

1560 ON FNA(N$) GOTO 1550,1640,1640,1665
```

Delete lines 1570 to 1630.

Program: POLY2

```
1800 BEEP @ DISP "PLOT COMPLETE" @ PAUSE
```

Delete lines 1801 to 1850.

Program: RESID

```
1170 DISP "Prepare plotter, press 'CONT' when ready." @ PAUSE

1180 PLOTTER IS 705 @ CSIZE 7 @ LORG 1 @ DEG

1450 N9=0 @ LORG 5 @ CSIZE 3

1570 IF X=0 THEN MOVE I-B1+1,S8

1580 IF X<>0 THEN MOVE D(X,I),S8

1610 CSIZE 7 @ LORG 1

1740 BEEP @ DISP "PLOT COMPLETE" @ PAUSE
```

Delete lines 1760 to 1830.

By making these modifications, the programs will produce reasonable plots. If the labeling is still not as you like it, you may easily change it in the programs POLY1 and RESID.

Appendix E
Using the Disc Version

The following information will increase your understanding of the disc version of this pac, and hopefully facilitate operation of the programs.

Printer Prompt

You have the ability to choose the output device by selecting the proper output code. After loading the program and pressing (RUN), the printer prompt will ask you to specify the output device with the following codes:

Enter: 1 (END LINE) will direct system output to the CRT

Enter: 2 (END LINE) will direct system output to the internal printer

other numbers of specific printers will direct system output to an external printer

A system output test is included with the above entry which will advance the desired printer one line if the system is operating properly.

Output via the CRT

When the CRT is chosen as the output device, the program will pause when displaying more than one full screen to allow full retention of output data. Simply press (CONT) to continue viewing until output is complete.

Operating Limits

The maximum operating limits of some of the programs have been slightly modified to accommodate the disc version of this pac. This need only be of concern as you approach these maximum operating limits.

References to Tape

All references to tape in this manual will be understood as references to the current mass storage medium, and therefore will apply to the disc version of this pac.

HEWLETT
PACKARD

For additional information please contact the nearest authorized HP-85 dealer
or your local Hewlett-Packard sales office.

00085-90158

Printed in U.S.A.