System 35 Software

Regression Analysis Methods

HEWLETT-PACKARD

Regression Analysis Methods

HP 9835A Desktop Computer

Hewlett-Packard Desktop Computer Division
3404 East Harmony Road, Fort Collins, Colorado 80525

HP Computer Museum
www.hpmuseum.net

For research and education purposes only.

Table of Contents

Commentary

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 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.

We are confident that you will find this collection of regression programs to be useful in your work. We will continue to develop other regression procedures in the future.

Thomas J. Boardman, Ph.D.
Professor in Charge
Statistical Laboratory
Colorado State University
Fort Collins, CO. 80523

Basic Statistics and Data Manipulation

I. DESCRIPTION:

This set of programs allows the user to enter a data matrix into memory and to then perform various operations on the data. The data entry may be made via keyboard, mass storage device, card reader, or other devices. The operations on the data set include editing, transforming, recoding, sorting, creating subfiles, naming, storing, and listing.

Features include a provision for missing data values, a provision for incorporating a subfile structure, the ability to store the data matrix and related information, the ability to join two data sets, error detection, the ability to correct many possible errors, and YES/NO keys to speed program use.

More specific objectives and features are listed with the instructions for the individual programs.

II. SYSTEM CONFIGURATON:

9835A Desktop Computer

Optional:

External printer
9872A Plotter (requires 98337A Plotter ROM)
9885M Flexible disk (requires 98331A Mass Storage ROM)
9869A Hopper card reader (requires 98332A I/O ROM)

III. SPECIAL CONSIDERATIONS:

A. Data Matrix Configuration: The data matrix incorporated in this program should be thought of as a p x n array whose columns correspond to observations and whose rows correspond to variables as shown below.

		OBSERVATIONS				
		O_1	O_2	O_3	. . .	O_n
VARIABLES	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.

	Subfile 1	Subfile 2		Subfile S	
	$O_1\ O_2 \ldots O_{n_1}$	$O_{n_1+1}\ O_{n_1+2} \ldots O_{n_1+n_2}$	$\ldots$	$O_{n_1+\ldots+n_{s-1}+1}$	$\ldots\quad O_{n_1+\ldots+n_s}$
V_1					
V_2					
.					
Variables .					
.					
V_P					

B. Limitations: The programs have been designed to operate in the basic machine. The maximum number of elements is 1500. Hence, for two variables, a maximum of 750 observations may be input; for three variables, the maximum is 500 observations. This may be changed if more memory is available. However, the scratch file "DATA" on the "REGRESSION ANALYSIS METHODS" tape cartridge may not be able to contain the increased amount of data. A new "DATA" file could be created to contain the increased data, or with a few program changes the data could be stored automatically on a data file on another mass storage medium (e.g., the 9885 floppy disk) and used as the "DATA" file on the cartridge presently is used.

If more than 1500 elements are desired, a number of other changes must also be made. In file "START", line 90, change Mno = 1500 to Mno = N, where N is the total number of elements desired. Also, in all COM statements, the array D (*,*) must be dimensioned as D(1,N). The following table gives the location of each COM statement:

File Name	Lines
"START"	40,1460
"BASIC"	30
"HCOPY"	40
"LIST"	40
"EDIT"	50
"STORE"	40
"TRANSF"	40
"RECODE"	50
"SUBFIL"	40
"NAME"	40
"JOIN"	30,140
"SORT"	50,150
"CARDS"	20
"MLR"	30
"STEP"	40
"POLY"	40

The maximum number of variables is 20. To increase this, change Mnv in file "START", line 80, from 20 to the number desired, N. In all COM statements, the vector Vn$(*)[*] must be dimensioned as Vn$(N)[10]. If more than 40 variables are desired, two further changes must be made: in file "STORE", line 200, the CREATE statement should be changed to CREATE F$,2+Nv*No*8 DIV M, M (where M=484+N*100); and the scratch file "DATA" must be made large enough to accommodate the larger data array.

The "REGRESSION ANALYSIS METHODS" tape cartridge contains the data points used in the examples for the BASIC STATISTICS AND DATA MANIPULATION routines on the file "DATA". The user may wish to page through the manual and try each of the programs available in the pack, 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 editted.

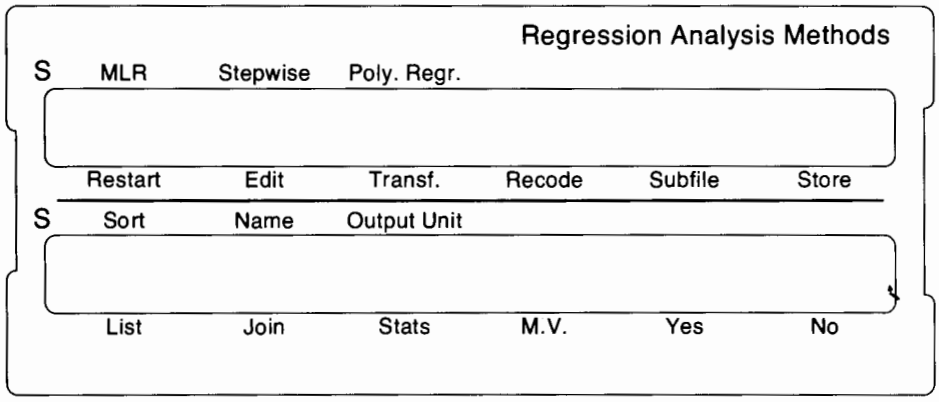
C. Data File Configuration: The scratch file on the program medium, i.e., "DATA", and any files created to hold stored data and related information are configured as follows. The data file is broken into logical records of 700 bytes each. The first logical record is a "header file", which contains information pertinent to the data set stored in the remaining logical records. The header file contains the following information (variables): data set title (T$), number of observations (No), number of variables (Nv), variable names (Vn$(*)), number of subfiles (Ns), subfile names (Sn$(*)), and subfile characteristics (Sc(*)). The remaining logical records contain D(*,*) -- the data matrix.

D. Missing Values: If a data value is missing, it can be so designated by pressing the key labeled M.V. when entering the data from keyboard. The number used to designate a missing value is −9999999.99999. The justification for this number is (besides seeming unlikely to occur as a legitimate data point) that it is easily picked out in a listing of the data. If the input is to be via cards, it may be more desirable to designate a missing value by a more easily typed number, e.g., by 0 if zero is not a legitimate data point. The zeros could then be converted to the missing value recognized by the programs via a user-defined transformation.

E. 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 STOP, then RUN.

F. Special Function Keys: Most of the operations are selected by pressing the appropriate special function keys shown below. Two of the keys, those labeled YES and NO, may be used as responses to any question requiring a yes or no answer.



G. Component Parts: The component parts of the "REGRESSION ANALYSIS METHODS" package (09835-15010) may be ordered separately:

User Instructions	09835-15011
Cartridge	09835-15014

Start

I. OBJECT OF PROGRAM:

This program allows you to place a data matrix into memory. The data may be entered from the keyboard, from cards, or from some other input device. Conversely, the data may have been entered previously and stored in the program medium's scratch file ("DATA") or in a user created file on a tape cartridge or flexible disk. In this case, the function of this program is to retrieve the previously stored data and place it into memory so that further operations can be performed. After the data is in memory, a listing option is provided.

II. SPECIAL CONSIDERATIONS:

A. The displayed prompts concerning the scratch file ("DATA"), whether the data was stored by this program, and whether the data is in the proper configuration are explained in the Special Considerations section of BASIC STATISTICS AND DATA MANIPULATION (page 3).

B. The prompts concerning the data medium and program medium may cause confusion. The word "medium" is used since the set of programs making up the "REGRESSION ANALYSIS METHODS" package may be on tape or floppy disk. Thus, the "program medium" refers to either the cartridge or the disk on which the programs making up this package are stored. Conversely, the "data medium" refers either to the cartridge or to the disk on which the file containing the data matrix resides. In some cases, the program medium and the data medium may be the same. However, this cannot be determined by the program and hence, the prompts are still displayed to make sure the correct medium is in the correct device.

C. When entering data from the keyboard, an option to enter data one case at a time is offered. The following example will serve to explain this feature. Suppose an investigator has collected four observations on each of three variables. He has the following data matrix:

		VARIABLE		
		1	2	3
OBSERVATION	1	10	2	5
	2	11	2	6
	3	9	3	7
	4	9	2	6

He elects to enter the data one case at a time. Then, when the prompt "Observation #1, all variables (separated by commas)=?" is displayed he enters 10, 2, 5 and presses CONTINUE, etc. This allows for quick entry of the data.

D. When entering data from cards, a few specifications should be noted. Information for the header file (see Data File Configuration, page 5) is entered via the keyboard as responses to program queries. Therefore, the only information on the cards is the observed data.

   Each data card should contain all data values for one observation, including missing values. (NOTE: A blank is **not** a legitimate missing value.) In addition, each value must be separated by a comma. Therefore, in order to enter the data for the matrix given, four cards would be punched with three data values in each.

E. If the user elects to use a method of data input other than the keyboard, mass storage device, or card reader, he needs to edit file "START" beginning at line 2770. Here, he should insert the code necessary to input the data in the configuration explained in the Special Considerations section (page 3). File "START" should then be RE-STORE'd. Program execution is begun by following the User Instructions for START.

III. USER INSTRUCTIONS:

Getting Started

1. With the machine turned on, insert the "REGRESSION ANALYSIS METHODS" cartridge into the tape drive.
2. Load the program into memory:
   a. Type: LOAD "AUTOST"
   b. Press: EXECUTE
3. Press: RUN
4. When "Is REGRESSION ANALYSIS METHODS overlay placed on keys?" is displayed:
   a. Press: YES when the overlay is in place.

Hard-Copy Options

5. When "HARD-COPY OPTIONS" is printed and "Option number=?" is displayed:
   a. Enter: 1, if no hard copy is desired.
   b. Press: CONT.
   c. Go to step 9.

   or

   a. Enter: 2, if hard copy is desired from an external printer.
   b. Press: CONT.

6. When "Are you using an HP-IB Printer?" is displayed:

Yes

   a. Press: YES if an HP-IB printer is being used.
   b. Go to step 7.

   or

   a. Press: NO if the printer is not HP-IB.
   b. Go to step 8.

7. When "Printer select code, bus address=? (for example, 7,1)" is displayed:

7, 8

   a. Enter the select code, bus address for the HP-IB printer.
   b. Press: CONT.
   c. Go to step 9.

8. When "Printer select code=?" is displayed:

   a. Enter the select code of the printer desired for hard copy.
   b. Press: CONT.

Data Entry Modes

9. When "DATA ENTRY MODES" is printed and "Mode number=?" is displayed:

   a. Enter: 1, if data is to be entered from the keyboard.
   b. Press: CONT.
   c. Go to step 10.

   or

   a. Enter: 2, if data is to be entered from a cartridge or a disk.
   b. Press: CONT.
   c. Go to step 17.

   or

   a. Enter: 3, if data is to be entered via cards.
   b. Press: CONT.
   c. Go to step 10.

   or

   a. Enter: 4, if data is to be entered from some other device.
   b. Press: CONT.

Creating the Data Set

10. When "Project title for this data set (<= 80 characters)=?" is displayed:

    a. Enter up to 80 characters to be used as the name of the data set.

    b. Press: CONT.

11. When "Number of variables=?" is displayed:

    a. Enter the number of variables you want in the data set.

    b. Press: CONT.

12. When "Number of observations/variable=?" is displayed:

    a. Enter the number of observations for each variable in the data set.

    b. Press: CONT.

13. When "VARIABLE NAMES" is printed and "Variable #I name (<= 10 characters)=?" is displayed:

    a. Enter the name of the Ith variable - it may be up to 10 characters long.

    b. Press: CONT.

    c. Repeat steps a and b until all variables have been named, then go to step 14.

14. When "Is above information correct?" is displayed:

    a. Press: YES if the variable names are as desired.

    b. Go to step 23.

    or

    a. Press: NO if changes to the variable names are desired.

15. When "Which variable do you wish to change?" is displayed:

    a. Enter the number beside the variable name that needs changing.

    b. Press: CONT.

16. When "Variable #I name (<= 10 characters)=?" is displayed:

    a. Type in new variable name.

    b. Press: CONT.

    c. Go to step 14.

Using an Existing Data Set

17. When "Is data stored on the program medium's scratch file (DATA)?" is displayed:

    a. Press: YES if the data is stored in "DATA", the scratch file on the program tape.

    b. The data and related information are loaded into memory at this point.

    c. Go to step 23.

    or

    a. Press: NO if the data is stored in a user-created file.

18. When file name specifications are printed and "Data file name=?" is displayed:

    a. Enter the data file name (less than or equal to six characters) followed by a colon and the mass storage unit specifier.

    b. Press: CONT.

19. When "Was data stored by this program?" is displayed:

    a. Press: YES if the data file was created by the BASIC STATISTICS AND DATA MANIPULATION routines or by any routine which stores a header file with the data (see Data File Configuration, page 5).

    b. Go to step 20.

    or

    a. Press: NO if the file was created by a program which does not store header information along with the data.

    b. Go to step 10.

20. When "Is data medium placed in device [msus]?" is displayed:

    a. Press: YES when the data medium is in place.

21. The data and related information are loaded into memory at this point.

22. When "Is program medium replaced in device?" is displayed:

    a. Press: YES when the program medium is in place (or if it was never removed).

Summary of the Data Set

23. At this point a summary of the data set is printed.

24. If the input mode is via keyboard, card reader or "OTHER", go to step 28.

25. If data was not entered from a file created by the BASIC STATISTICS AND DATA MANIPULATION routines (i.e., NO was the response given in step 19), go to step 28.

26. When "List data?" is displayed:

    a. Press: YES if a hard-copy listing of the data is desired.

    b. Go to step 1 of the User Instructions for the LIST program.

    or

    a. Press: NO if a hard-copy listing is not desired.

Selecting a Program

27. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

    b. Go to the User Instructions for the selected key.

Data Input

28. When "DATA INPUT" is printed:

    a. If the data is being entered from the keyboard, go to step 29.

    or

    a. If the data is being entered from cartridge or disk, go to step 34.

    or

    a. If the data is being entered via cards, go to step 42.

    or

    a. If the data is being entered from some other device, go to step 50.

29. When the program notes are printed:

    a. If there is only one variable in the data set, go to step 30.

    or

    a. If there are two or more variables in the data set, go to step 31.

30. When the heading is printed and "Observation #I" is displayed:

    a. Enter the value of the Ith observation.

    b. Press: CONT.

    c. Repeat steps a and b until all observations have been entered.

    d. Go to step 51.

31. When "Enter data one case at a time, that is, by observation?" is displayed:

    a. Press: YES if you desire to key in the values of all variables for a given observation at once.

    b. Go to step 33.

    or

    a. Press: NO if you desire to key in the value of one variable at a time for a given observation.

32. When "Observation #I, Variable #J" is displayed:

    a. Enter the value of the Ith observation, Jth variable.

    b. Press: CONT.

    c. Repeat steps a and b until all variables for all observations are entered.

    d. Go to step 51.

33. When "Observation #I, all variables (separated by commas)=?" is displayed:

    a. Enter variable 1, comma, variable 2, comma, . . ., variable P for Observation #I.

    b. Press: CONT.

    c. Repeat steps a and b until all variables for all observations are entered.

    d. Go to step 51.

34. When "Are MISSING VALUES denoted by −9999999.99999?" is displayed:

    a. YES if missing values were specified by pressing MISSING VALUE (key labeled M.V.) when the data was originally entered, or if there are no missing values.

    b. Go to step 36.

    or

    a. Press: NO if missing values are specified by some other value.

35. When "Missing value=?" is displayed:

    a. Enter the value that has been used to specify a missing data point.

    b. Press: CONT.

36. When "Is data in proper configuration, that is, variables = rows, observations = columns?" is displayed:

    a. Press: YES if the data matrix contains a variable in each row and an observation in each column.

b. Go to step 38.

or

a. Press: NO if the data matrix is in a different configuration.

37. When "Data stored as contiguous array with observations = rows, variables = columns?" is displayed:

    a. Press: YES if the data matrix contains an observation in each row and a variable in each column.

    b. Go to step 38.

    or

    a. Press: NO if the data is configured in a different manner.

    b. Execution stops - the data set is unacceptable for use by this program and therefore must be keyed in.

38. When "Is data medium placed in device?" is displayed:

    a. Press: YES when data medium is in place.

39. At this point the data is loaded into memory.

40. When "Is program medium replaced in device?" is displayed:

    a. Press: YES when program medium is in place (or if it was never removed).

41. At this point the data is stored on the scratch file of the program medium ("DATA"). Go to step 51.

42. When "Is subprogram 'CARDS' already linked on?" is displayed:

    a. Press: YES if subprogram CARDS currently resides in memory.

    b. Go to step 43.

    or

    a. Press: NO if file "CARDS" has not yet been linked on.

43. When "What is the select code of the card reader (default is 3)?" is displayed:

    a. Enter the card reader select code.

    b. Press: CONT.

44. When "Enter value used to indicate last card." is displayed:

    a. Enter the numeric value used to signal the end of the data stream.

    NOTE: The value used should be one that will not be confused with a valid data value. Example: 10E99.

b. Press: CONT.

45. When the program notes are printed, make sure the data cards are properly loaded in the card reader and that the card reader is ready.

    a. Press: CONT. when you are ready to read in the data.

46. When "Are MISSING VALUES, if any, denoted by −9999999.99999?" is displayed:

    a. Press: YES if −9999999.99999 was used for any missing values, or if there are no missing values.

    b. Go to 48.

    or

    a. Press: NO if missing values are specified by some other value.

47. When "Missing value = ?" is displayed:

    a. Enter the value that was used to specify a missing point.

    b. Press: CONT.

48. When "Is program medium replaced in device?" is displayed:

    a. Press: YES when program medium is in place (or if it was never removed).

49. At this point the data is stored on the scratch file of the program medium ("DATA"). Go to 51.

50. At this point the code necessary to receive the desired input from some other input device will be executed. An explanation for supplying that code is given on page 10.

51. If the hard-copy option chosen in step 5 was 1 (no hard copy), go to step 53.

52. When "List data?" is displayed:

    a. Press: YES if a hard-copy listing of the data is desired.

    b. Go to step 1 of the User Instructions for LIST.

    or

    a. Press: NO if no listing is desired.

53. When "Are corrections to the data necessary?" is displayed:

    a. Press: YES if corrections are needed.

    b. Go to step 1 of the User Instructions for EDIT.

    or

    a. Press: NO if the data is correct.

Selecting a Program

54. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys defined on the "REGRESSION ANALYSIS METHODS" overlay.

    b. Go to step 1 of the User Instructions for the selected key.

IV. EXAMPLE:

The data listed below will be used as a data sample for most of the operations performed by this package. There are five variables:

1. average monthly temperature (°C)
2. monthly production (kg.)
3. number of working days
4. number of people on the payroll
5. monthly water use (litrés)

There are 17 observations, one per month for 17 months.

```
*****************************************************************************
*                              DATA MANIPULATION                            *
*****************************************************************************
                                    SAMPLE

Data file name: DATA
Number of observations: 17
Number of variables: 5

Variables names:
   1. Temp(C)
   2. Production
   3. Days
   4. Payroll
   5. Water Use

Subfiles:  NONE

                                    SAMPLE
```

OBS#	Variable # 1	Variable # 2	Variable # 3	Variable # 4	Variable # 5
1	14.90000	6396.00000	21.00000	134.00000	3373.00000
2	18.40000	5736.00000	22.00000	146.00000	3110.00000
3	21.60000	6116.00000	22.00000	158.00000	3180.00000
4	25.20000	8287.00000	20.00000	171.00000	3293.00000
5	26.30000	13313.00000	25.00000	198.00000	3390.00000
6	27.20000	13108.00000	23.00000	194.00000	4287.00000
7	22.20000	10768.00000	20.00000	180.00000	3852.00000
8	17.70000	12173.00000	23.00000	191.00000	3366.00000
9	12.50000	11390.00000	20.00000	195.00000	3532.00000
10	6.90000	12707.00000	20.00000	192.00000	3614.00000
11	6.40000	15022.00000	22.00000	200.00000	3896.00000
12	13.30000	13114.00000	19.00000	211.00000	3437.00000
13	18.20000	12257.00000	22.00000	203.00000	3324.00000
14	22.80000	13118.00000	22.00000	197.00000	3214.00000
15	26.10000	13100.00000	21.00000	196.00000	4345.00000
16	26.30000	16716.00000	21.00000	205.00000	4936.00000
17	4.20000	14056.00000	22.00000	205.00000	3624.00000

Edit

I. OBJECT OF PROGRAM:

This program is designed to allow you to perform a variety of editing procedures on the data matrix. The editing capabilities include: the changing of an incorrect data value, the deletion of a variable, the deletion of an observation, the addition of an observation, the insertion of an observation (if the data is ordered), and the addition of a variable. All of the above operations can be performed repeatedly; for example, three variables could be deleted in succession. After the data matrix has been edited, you are given the option of listing the edited data.

II. SPECIAL CONSIDERATIONS:

A. Order of Corrections: As stated in the program note printed on the screen, the data is renumbered after deletions or insertions are performed. For this reason, if more than one deletion (insertion) is to be performed, it is recommended that the highest-numbered observation (or variable) be deleted, then the next highest-numbered, etc. For example, if observations three and eight are to be deleted, then it is recommended to delete observation eight first, then observation three. Notice that if observation three were deleted first, the subsequent renumbering would move observation eight to position seven. The recommendation is meant to alleviate confusion which may occur due to the renumbering.

B. Subfiles: Insertions or deletions of observations will affect the content of subfiles which are in existence at the time of editing; for example, if subfile one consists of the first 10 observations while subfile two consists of the last 20 and if observation five is deleted, then observation ten (formerly numbered 11) will have jumped from subfile two to subfile one. Thus, it may be necessary to change the subfile structure after editing. It is recommended that subfiles be created only after all editing has been performed.

III. METHODS AND FORMULAE:

A. The data matrix is redimensioned into a row vector to facilitate the shuffling of elements necessitated by the editing operations. The vector contains all the observations of variable one, followed by the observations of variable two, etc. When an observation is inserted, for example, the elements of the data vector are shuffled one at a time to make room for the incoming observation. Similarly, when an observation is deleted, the remaining observations are "packed" together so that the resultant data vector has no "holes" between observations.

IV. USER INSTRUCTIONS:

Press: The special function key labeled Edit.

Correcting a Data Value

1. When "Correct a data value?" is displayed:

   a. Press: NO if there are no data values present that need to be changed.

b. Go to step 6.

or

a. Press: YES if there is a data point in memory that needs to be changed.

2. When "Observation number=?" is displayed:

   a. Enter the observation number of the data value that is to be changed.
   b. Press: CONT.
   c. If there is only one variable in the data set, go to step 4.

3. When "Variable number=?" is displayed:

   a. Enter the variable number of the data value that is to be changed.
   b. Press: CONT.

4. When "Old value='n' - Correct value=?" is displayed:
   (The old value is displayed so one can be sure that the correction is being made as anticipated.)

   a. Enter the correct data value.
   b. Press: CONT.

5. When "Correct another value?" is displayed:

   a. Press: YES if there is another data point in memory that needs to be changed.
   b. Go to step 2.

   or

   a. Press: NO if there are no more changes to be made to existing data points.

6. If there is only one variable in the data set, go to step 11.

Deleting a Variable

7. When "Delete a variable?" is displayed:

   a. Press: NO if no variables are to be removed from the data set.
   b. Go to step 11.

   or

   a. Press: YES if you desire to remove one or more variables from the data set.

8. When the program note is printed, and "Number of the variable to be deleted=?" is displayed:

a. Enter the highest-numbered variable to be deleted; for example, if variables 2 and 7 are to deleted, enter 7 first (since the variables are renumbered after a deletion, and if 2 were deleted first, then variable 7 would be referred to as variable 6).

b. Press: CONT.

9. If there is only one variable remaining in the data set, go to step 11.

10. When "Delete another variable?" is displayed:

    a. Press: YES if you desire to remove another variable from the data set.

    b. Go to step 8.

    or

    a. Press: NO if no more variables are to be deleted from the data set.

Deleting an Observation

11. When "Delete an observation?" is displayed:

    a. Press: NO if no observations are to be removed from the data set.

    b. Go to step 14.

    or

    a. Press: YES if you desire to remove one or more observations from the data set.

12. When the program note is printed and "Number of the observation to be deleted=?" is displayed:

    a. Enter the highest-numbered observation to be deleted; for example, if observations 2 and 5 are to be deleted, enter 5 first (since the observations are renumbered after each deletion, if 2 were deleted first, then observation 5 would become observation 4).

    b. Press: CONT.

13. When "Delete another observation?" is displayed:

    a. Press: YES if it is desired to remove another observation from the data set.

    b. Go to step 12.

    or

    a. Press: NO if no more observations are to be removed from the data set.

Adding an Observation

14. If the addition of one more observation will exceed available memory, go to step 23.

15. When "Add an observation?" is displayed:

a. Press: NO if no observations are to be added to the end of the data set or inserted into the data set.

b. Go to step 23.

or

a. Press: YES if you desire to add more observations to the end of the data set or to insert observations into the data set.

16. When "Are observations ordered, that is, should additions be inserted?" is displayed:

a. Press: YES if the data follows some type of order and if you desire to insert an observation, for example, between observations 4 and 5.

b. Go to step 19.

or

a. Press: NO if the observations are to be added to the end of the data set.

17. When "How many observations are to be added?" is displayed:

a. Enter the number of observations to be added to the data set.

b. Press: CONT.

18. When "Observation #'I', Variable #'J'=?" is displayed:

a. Enter the data point corresponding to the Jth variable of the Ith observation.

b. Press: CONT.

c. Repeat steps a-b until data has been entered for all variables of each of the observations added.

d. Go to step 23.

19. When the program note is printed and "Insertion to precede observation #?" is displayed:

a. Enter the number of the observation which the insertion will precede; for example, if an observation is to be inserted between observations 8 and 9, enter 9. Since the observations are renumbered after each insertion, insertions should be made between the highest-numbered observations first; for example, if insertions are to be made between observations 4 and 5 as well as between observation 8 and 9, the latter should be performed first, that is, 9 should be entered first.

b. Press: CONT.

20. When "Observation #'I', Variable #'J'=?" is displayed:

a. Enter the data point corresponding to variable J of the inserted observation.

b. Press: CONT.

c. Repeat steps a-b until data has been entered for all variables.

21. If the addition of one more observation will exceed program limitations, go to step 23.
22. When "Insert another observation?" is displayed:

    a. Press: YES if you desire to insert another observation.
    b. Go to step 19.

    or

    a. Press: NO if no more observations are to be inserted.

Adding a Variable

23. If the addition of a variable will exceed program limitations, go to step 29.
24. When "Add a variable?" is displayed:

    a. Press: NO if no variables are to be added from the keyboard.
    b. Go to step 29.

    or

    a. Press: YES if you desire to add one or more variables from the keyboard.

25. When "Variable name (<= 10 characters)" is displayed:

    a. Enter the name of the variable to be added - it may be up to 10 characters long.
    b. Press: CONT.

26. When "Variable #'I', Observation #'J' =?" is displayed:

    a. Enter the Jth observation of the variable being added.
    b. Press: CONT.
    c. Repeat steps a-b until all observations have been entered for the added variable.

27. If the addition of one more variable will exceed available memory, go to step 29.
28. When "Add another variable?" is displayed:

    a. Press: YES if you desire to add another variable from the keyboard.
    b. Go to step 25.

    or

    a. Press: NO if no more variables are to be added from the keyboard.

29. When "More corrections?" is displayed:

    a. Press: YES if more corrections to the data matrix are required.

    b. Go to step 1.

    or

    a. Press: NO if no more corrections to the data set are necessary.

30. At this point the corrected data matrix and related information are recorded in file "DATA" of the program medium.

31. When "List data?" is displayed:

    a. Press: YES if a listing of the corrected data set is desired.

    b. Go to step 1 of the User Instructions for LIST.

    or

    a. Press: NO if a listing of the data is not required.

32. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

    b. Go to step 1 of the User Instructions for the selected key.

V. EXAMPLE:

The program EDIT was run in the print all mode. The printout shows a correction, deletion of an observation, and addition of an observation. The edited data was then listed.


```
Correct a data value?
"YES"
Observation number = ?
11
Variable number = ?
2
Old value =  15022 -- Correct value =
?
15024
Observation # 11 Variable # 2 -- correct value =  15024
Correct another value?
"NO"

Delete a variable?
"NO"
Delete an observation?
"YES"
Number of the observation to be deleted = ?
10
Observation # 10 has been deleted, 16 observations remain.
Delete another observation?
"NO"

Add an observation?
"YES"
Are observations ordered, that is, should additions be inserted?
"NO"
How many observations are to be added?
1
Observation # 17 Variable # 1  =
?
4.2
Observation # 17 Variable # 1  =  4.2
Observation # 17 Variable # 2  =
?
12707
Observation # 17 Variable # 2  =  12707
Observation # 17 Variable # 3  =
?
20
Observation # 17 Variable # 3  =  20
Observation # 17 Variable # 4  =
?
192
Observation # 17 Variable # 4  =  192
Observation # 17 Variable # 5  =
?
3614
Observation # 17 Variable # 5  =  3614

  Total number of observations now =  17

Add a variable?
"NO"
More corrections?
"NO"
List data?
"YES"
```

SAMPLE

OBS#	Variable # 1	Variable # 2	Variable # 3	Variable # 4	Variable # 5
1	14.90000	6396.00000	21.00000	134.00000	3373.00000
2	18.40000	5736.00000	22.00000	146.00000	3110.00000
3	21.60000	6116.00000	22.00000	158.00000	3180.00000
4	25.20000	8287.00000	20.00000	171.00000	3293.00000
5	26.30000	13313.00000	25.00000	198.00000	3390.00000
6	27.20000	13108.00000	23.00000	194.00000	4287.00000
7	22.20000	10768.00000	20.00000	180.00000	3852.00000
8	17.70000	12173.00000	23.00000	191.00000	3366.00000
9	12.50000	11390.00000	20.00000	195.00000	3532.00000
10	6.40000	15024.00000	22.00000	200.00000	3896.00000
11	13.30000	13114.00000	19.00000	211.00000	3437.00000
12	18.20000	12257.00000	22.00000	203.00000	3324.00000
13	22.80000	13118.00000	22.00000	197.00000	3214.00000
14	26.10000	13100.00000	21.00000	196.00000	4345.00000
15	26.30000	16716.00000	21.00000	205.00000	4936.00000
16	4.20000	14056.00000	22.00000	205.00000	3624.00000
17	4.20000	12707.00000	20.00000	192.00000	3614.00000

Transform

I. OBJECT OF PROGRAM:

This program allows you to transform one or two variables in the data matrix via 16 pre-specified functions or through a function which you specify. The transformed data may then be treated as a new variable, or it may replace the elements of an existing variable. Hence, transformations on more than two variables may be made iteratively or via a user-defined transformation.

The transformations available are:

1. $aX^b + c$
2. $a \log(bX) + c$
3. $a \ln(bX) + c$
4. $a \exp(bX) + c$
5. $a(b^{cX})$
6. $a \cos(bX) + c$
7. $a \sin(bX) + c$
8. $a\sqrt{\arcsin(bX) + c}$
9. $aX + bY + c$
10. aX^bY^c
11. $a \log(bX + cY)$
12. $a \ln(bX + cY)$
13. $a \cos(bX + cY)$
14. $a \sin(bX + cY)$
15. PROUND (X,a) [round to specified power of 10]
16. DROUND (X,a) [round to specified no. of digits]
17. User Defined

II. SPECIAL CONSIDERATIONS:

A. Missing Values: None of the 16 pre-specified transformations are applied to missing values. Thus, missing values are unaffected by these transformations. However, this is not necessarily the case with the user-defined transformation. If you define a transformation and there are missing values, you must make provisions to ensure that the transformation is not applied to the missing values (unless, of course, this is desired). This may be accomplished as explained below.

B. User-Defined Transformation: At step three of the User Instructions, up to 10 lines of code may be used to define a transformation, namely lines 1310 through 1319. These lines may be typed in and stored successively prior to continuing on to the next instruction step. The following example shows the form of a typical user-defined transformation. Suppose the data set consists of four variables with 20 observations each. There are missing values, which are not to be used. You desire to form variable five as the sum of the exponentials of variables one and three. The following sequence should be carried out at step three of the TRANSFORM User Instructions:

a. Type: 1310 D(Z,I)=EXP(D(1,I))+EXP(D(3,I))

b. Press: STORE

c. Type: 1311 IF D(1,I)=−9999999.99999 OR
D(3,I)=−9999999.99999 THEN D(Z,I)=1

d. Press: STORE

The variable Z is used to identify the variable in which the result of the transformation is to be stored. Notice that the elements of D(*,*) are the data - the first subscript refers to variable number while the second subscript refers to observation number (and should always be 'I' here). Line 1311 sets D(Z,I)=1 if a missing value is present in either of variables one or three.

III. USER INSTRUCTIONS:

Press: The special function key labeled Transf.

1. When the program notes and transformations are printed and "Transformation number=?" is displayed:

   a. Enter the number of the transformation which is desired (1-17, inclusive).

   b. Press: CONT.

2. If the selected transformation was not 'User defined', that is, not transformation number 17, go to step 4.
3. When the program notes are printed and "Ready to continue?" is displayed:

   a. Type: "1310 D(Z,I)='defined transformation' "; the form of the defined transformation is discussed in Special Considerations.

   b. Press: STORE

   c. Press: YES when ready to continue.

   d. Go to step 11.

4. When the selected transformation is printed and "Number of the variable corresponding to X=?" is displayed:

   a. Enter the variable number corresponding to X, where X refers to the variable in the selected transformation.

   b. Press: CONT.

5. If the selected transformation number is less than 9 or greater than 14, go to step 7.
6. When "Number of the variable corresponding to Y=?" is displayed:

   a. Enter the variable number corresponding to Y, where Y refers to the variable in the selected transformation.

b. Press: CONT.

7. When "Parameter a=?" is displayed:

   a. Enter the value of parameter a, where a refers to the parameter in the selected transformation.

   b. Press: CONT.

8. If the selected transformation number is greater than 14, go to step 11.

9. When "Parameter b=?" is displayed:

   a. Enter the value of parameter b, where b refers to the parameter in the selected transformation.

   b. Press: CONT.

10. When "Parameter c=?" is displayed:

    a. Enter the value of parameter c, where c refers to the parameter in the selected transformation.

    b. Press: CONT.

11. When "Store transformed data in variable # (<='I')" is displayed:

    a. Enter the number of the variable in which the transformed data is to be stored - it may be less than or equal to I.

    b. Press: CONT.

12. If the variable specified in step 11 existed previously, that is, if the transformed data is being stored in place of previously existing data, go to step 14.

13. When "Variable name (<= 10 characters)=?" is displayed:

    a. Enter the name of the variable to contain the transformed data - it may be up to 10 characters long.

    b. Press: CONT.

14. If the transformation was 'User defined', go to step 16.

15. When "Is above information correct?" is displayed:

    a. Press: NO if a mistake has been made during entry of the information printed.

    b. Go to step 1.

    or

    a. Press: YES if no corrections are to be made.

16. At this point the transformation is carried out.
17. When "More transformations?" is displayed:

    a. Press: YES if more transformations are desired.

    b. Go to step 1.

    or

    a. Press: NO if no more transformations are desired.

18. The altered data matrix and related information are now stored on file "DATA" of the program medium.
19. When "List data?" is displayed:

    a. Press: YES if a listing of the data is desired.

    b. Go to step 1 of the User Instructions for LIST.

    or

    a. Press: NO if a listing of the data is not required.

20. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

    b. Go to step 1 of the User Instructions for the selected key.

IV. EXAMPLE:

Variable five, Water Use, was transformed, converting litres to U.S. gallons. Transformation 1 was used, with a=0.2642, b=1,c=0.

```
Transformation number = ?
1
Number of the Variable number corresponding to X = ?
5
Parameter a = ?
.2642
Parameter b = ?
1
Parameter c = ?
0
Store transformed data in Variable # ( <= 6 )
?
5
Is above information correct?
"YES"
The following transformation was performed: a*(X^b)+c
  where a =  .2642
        b =  1
        c =  0
        X is Variable # 5
        Transformed data is stored in Variable # 5 (Water Use).


More transformations?
"NO"
List data?
"YES"
```

SAMPLE

OBS#	Variable # 1	Variable # 2	Variable # 3	Variable # 4	Variable # 5
1	14.90000	6396.00000	21.00000	134.00000	891.14660
2	18.40000	5736.00000	22.00000	146.00000	821.66200
3	21.60000	6116.00000	22.00000	158.00000	840.15600
4	25.20000	8287.00000	20.00000	171.00000	870.01060
5	26.30000	13313.00000	25.00000	198.00000	895.63800
6	27.20000	13108.00000	23.00000	194.00000	1132.62540
7	22.20000	10768.00000	20.00000	180.00000	1017.69840
8	17.70000	12173.00000	23.00000	191.00000	889.29720
9	12.50000	11390.00000	20.00000	195.00000	933.15440
10	6.90000	12707.00000	20.00000	192.00000	954.81880
11	6.40000	15022.00000	22.00000	200.00000	1029.32320
12	13.30000	13114.00000	19.00000	211.00000	908.05540
13	18.20000	12257.00000	22.00000	203.00000	878.20080
14	22.80000	13118.00000	22.00000	197.00000	849.13880
15	26.10000	13100.00000	21.00000	196.00000	1147.94900
16	26.30000	16716.00000	21.00000	205.00000	1304.09120
17	4.20000	14056.00000	22.00000	205.00000	957.46080

Recode

I. OBJECT OF PROGRAM:

This program allows you to assign codes to various categories or classes of data. The categories are intervals along the real number line, 20 of which may be specified. The recoding is done on one variable at a time. The same coding scheme may be used iteratively on successive variables. A summary of the coding intervals, codes, and number of observations assigned to each code may be printed as hard copy.

II. SPECIAL CONSIDERATIONS:

A. Coding Schemes: Four coding schemes are available for the sole purpose of eliminating unnecessary entries from the keyboard. If the coding intervals are all of the same length and are contiguous, that is, together they form a connected interval, then the interval construction can be accomplished internally knowing only the interval length and lower limit for the first interval. Similarly, if the intervals are of equal length but noncontiguous, for example, [10,20), [25,35), [35,45), [50,60), then the lower limit of each interval needs to be specified but the upper limit may be computed internally. Hence, the coding schemes are meant only to minimize the amount of information which needs to be entered from the keyboard. Clearly, the coding intervals could all be constructed by requiring you to enter the lower and upper limits for each and every interval (which is necessary, and what is done if the intervals are unequal and non-contiguous).

B. Same Coding Scheme: The coding is carried out on one variable at a time. However, if you desire to code both variables one and two according to the same coding intervals, these intervals need to be constructed only once. A positive response to the option offered by "Use same coding scheme?" allows variable two to be coded according to the same scheme without constructing the intervals a second time. If, however, you desire to code variable two according to a different scheme, it is possible to construct a second set of coding intervals by giving a negative response to the above prompt.

C. The brackets used to denote the coding intervals are meant to follow their usual mathematical interpretation, that is, the intervals are closed on the left and open on the right.

D. If an observation does not fall into any of the coding intervals, its value is not changed during the coding process.

III. USER INSTRUCTIONS:

Press: The special function key labeled Recode.

1. When the title is printed and "Store recoded data in variable # (<='I')?" is displayed:

   a. Enter the number of the variable in which the recoded data is to be stored - it may be less than or equal to I.

   b. Press: CONT.

2. If the variable specified in step 1 existed previously, that is, if the recorded data is being stored in place of previously existing data, go to step 4.
3. When "Variable name (<= 10 characters)=?" is displayed:

   a. Enter the name of the variable which will contain the recoded data - it may be up to 10 characters long.

   b. Press: CONT.

4. When "Number of the variable to be recoded=?" is displayed:

   a. Enter the number of the variable which is to be recoded.

   b. Press: CONT.

5. When the data coding schemes are printed and "Option number=?" is displayed:

   a. Enter the option number associated with the desired recoding scheme.

   b. Press: CONT.

6. When "Number of recoding intervals to be specified (<=20)=?" is displayed:

   a. Enter the number of categories into which the data will be divided, or equivalently, the number of codes which will be assigned.

   b. Press: CONT.

7. If the specified coding scheme does not have equal intervals, go to step 9.
8. When "Length of each interval=?" is displayed:

   a. Enter the common increment to be used to generate the intervals.

   b. Press: CONT.

9. If the data recoding scheme is non-contiguous, go to step 11.
10. When "Lower limit of first interval=?" is displayed:

    a. Enter the smallest number which will be assigned code #1.

    b. Press: CONT.

11. NOTE: The following steps, 12 through 14, are repeated the number of times specified in step 6 until all information is input for each of the recoding intervals. Some of the steps are skipped for certain schemes, since the necessary information is computed internally. For scheme #1, only step 14 needs to be repeated. For scheme #2, only steps 13 and 14 need to be repeated. For scheme #3, only steps 12 and 14 need to be repeated. For scheme #4, steps 12-14 need to be repeated.

12. When "Lower limit of interval #'I'=?" is displayed:

    a. Enter the smallest number which will be assigned the code associated with interval #I.

    b. Press: CONT.

13. When "Upper limit of interval #'I'=?" is displayed:

    a. Enter the upper bound of numbers which will be assigned the code associated with interval #I.

    b. Press: CONT.

14. When "For data falling in interval #'I', code=?" is displayed:

    a. Enter the code which will be assigned to those data values falling in the Ith coding interval.

    b. Press: CONT.

    c. Repeat as noted in instruction 11.

15. When "Is above information correct?" is displayed:

    a. Press: NO if a mistake has been made in entering the information.

    b. Go to step 5.

    or

    a. Press: YES if the recoding scheme is as desired.

16. At this point the recoding is carried out and a note of the coding is printed.

17. When "Recode more data?" is displayed:

    a. Press: YES if you desire to recode more data.

    b. Go to step 18.

    or

    a. Press: NO if no more recoding is desired.

    b. Go to step 19.

18. When "Use same coding scheme?" is displayed:

    a. Press: YES if the same intervals and codes are desired, that is, if the identical recoding scheme is to be applied to another variable.

    b. Go to steps 1-4, then directly to 16.

    or

    a. Press: NO if a different recoding scheme is desired.

    b. Go to step 1.

19. At this point the altered data matrix and related information are stored on file "DATA" of the program medium.
20. When "List data?" is displayed:

    a. Press: YES if a listing of the data is desired.

    b. Go to step 1 of the User Instructions for LIST.

    or

    a. Press: NO if no listing is desired.

21. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

    b. Go to step 1 of the User Instructions for the selected key.

IV. EXAMPLE:

Using the original data, variable 5, Water Use, was recoded, and the resulting data was stored as a new variable, 6. The purpose of the recoding operation was to simplify the data, replacing entire intervals by single numbers.

```
Store recoded data in Variable # (<= 6 )
?
6
Variable name (<= 10 characters) = ?
Coded Use
Number of the variable to be recoded = ?
5
Option number = ?
1
Number of recoding intervals to be specified (<=20) = ?
5
Length of each interval = ?
400
Lower limit of first interval = ?
3000
For data falling in interval 1 , code =
?
30
For data falling in interval 2 , code =
?
34
For data falling in interval 3 , code =
?
38
For data falling in interval 4 , code =
?
42
For data falling in interval 5 , code =
?
46
Is above information correct?
"YES"

Variable # 5 is recoded into 5 categories, and the recoded
values are stored in Variable # 6 , where:
```

CATEGORY BOUNDS		# OBS	
LOWER	UPPER	CODED	CODE
3000.000	3400.000	8	30.000
3400.000	3800.000	4	34.000
3800.000	4200.000	2	38.000
4200.000	4600.000	2	42.000
4600.000	5000.000	1	46.000

```
Recode more data?
"NO"
List data?
"YES"
```

SAMPLE

OBS#	Variable # 1	Variable # 2	Variable # 3	Variable # 4	Variable # 5	Variable # 6
1	14.90000	6396.00000	21.00000	134.00000	3373.00000	30.00000
2	18.40000	5736.00000	22.00000	146.00000	3110.00000	30.00000
3	21.60000	6116.00000	22.00000	158.00000	3180.00000	30.00000
4	25.20000	8287.00000	20.00000	171.00000	3293.00000	30.00000
5	26.30000	13313.00000	25.00000	198.00000	3390.00000	30.00000
6	27.20000	13108.00000	23.00000	194.00000	4287.00000	42.00000
7	22.20000	10768.00000	20.00000	180.00000	3852.00000	38.00000
8	17.70000	12173.00000	23.00000	191.00000	3366.00000	30.00000
9	12.50000	11390.00000	20.00000	195.00000	3532.00000	34.00000
10	6.90000	12707.00000	20.00000	192.00000	3614.00000	34.00000
11	6.40000	15022.00000	22.00000	200.00000	3896.00000	38.00000
12	13.30000	13114.00000	19.00000	211.00000	3437.00000	34.00000
13	18.20000	12257.00000	22.00000	203.00000	3324.00000	30.00000
14	22.80000	13118.00000	22.00000	197.00000	3214.00000	30.00000
15	26.10000	13100.00000	21.00000	196.00000	4345.00000	42.00000
16	26.30000	16716.00000	21.00000	205.00000	4936.00000	46.00000
17	4.20000	14056.00000	22.00000	205.00000	3624.00000	34.00000

Sort

I. OBJECT OF PROGRAM:

This program allows the data matrix, or subfiles thereof, to be sorted according to the values of one variable. For example, suppose an investigator has five observations of three variables, say height, weight and age and wanted to arrange the observations in ascending order according to age. This is accomplished by sorting the data matrix according to variable three.

II. SPECIAL CONSIDERATIONS:

A. Subfile Structure Options: If subfiles are ignored, the entire data set will be sorted and, in the process, the composition of the subfiles is subject to change. The option of sorting certain subfiles may be used to sort a single subfile or a set of successive subfiles according to one variable. The option of sorting all subfiles may be used to sort every subfile. The options of sorting certain subfiles and sorting all subfiles treat each subfile as if it were a separate data set. Thus, the sort is done with respect to one subfile at a time.

B. It is important to note that entire observations are moved when the sort is carried out. Thus, referring to the example given in Object of Program, a person's height and weight remain with the person's age as shown below.

ORIGINAL MATRIX

		VARIABLE		
		Height	Weight	Age
OBSERVATION	1	72	170	21
	2	70	165	25
	3	69	150	20
	4	71	155	22
	5	73	160	19

MATRIX SORTED BY AGE

		VARIABLE		
		Height	Weight	Age
OBSERVATION	1	73	160	19
	2	69	150	20
	3	72	170	21
	4	71	155	22
	5	70	165	25

III. USER INSTRUCTIONS:

Press: The special function key labeled Sort.

1. At this point the subprogram "Sort" is called. A subprogram is used here for the sole purpose of variable dimensioning.
2. When "Number of the Variable on which to sort=?" is displayed:

   a. Enter the number of the variable on which the matrix will be sorted, that is, the variable whose observations will be arranged in ascending order.

   b. Press: CONT.

3. If the data set contains no subfiles, go to step 7.
4. When the subfile structure options are printed and "Option number=?" is displayed:

   a. Enter 1 if you desire to ignore the subfile structure, that is, to group all the observations together for the sort. Note that this may shift observations from subfile to subfile.

   b. Press: CONT.

   c. Go to step 7.

   or

   a. Enter 2 if you desire to sort a set of one or more successive subfiles; for example, to sort subfiles 3 through 5. The sort will be done by subfile: subfile 3 will be sorted, then 4, and finally 5.

   b. Press: CONT.

   c. Go to step 5.

   or

   a. Enter 3 if you desire to sort all of the subfiles. Each subfile will be sorted as if it were a separate data set.

   b. Press: CONT.

   c. Go to step 7.

5. When "Number of first subfile=?" is displayed:

   a. Enter the number of the first subfile to be sorted.

   b. Press: CONT.

6. When "Number of last subfile=?" is displayed:

   a. Enter the number of the last subfile to be sorted. All subfiles between and including the first and last specified will be sorted as though they were separate data sets.

   b Press: CONT.

7. At this point the desired sorting is performed and the sorted data is stored on file "DATA" of the program medium.
8. When "List data?" is displayed:

   a. Press: YES if a listing of the sorted data is desired.

   b. Go to step 1 of the User Instructions for LIST.

   or

   a. Press: NO if no listing is desired.

9. When "SELECT ANY KEY" is displayed:

   a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

   b. Go to step 1 of the User Instructions for the selected key.

IV. EXAMPLE:

The original data set was sorted by variable 2. That is, the observations were arranged so that monthly production was in ascending order. Notice that the other variables are carried by the sort operation, so that although the order of the observations has changed, each observation remains intact.

```
Number of the Variable on which to sort = ?
2
Data set:
                                              SAMPLE
has been arranged in ascending order according to Variable # 2
(Subfiles ignored)


List data?
"YES"
```

SAMPLE

OBS#	Variable # 1	Variable # 2	Variable # 3	Variable # 4	Variable # 5
1	18.40000	5736.00000	22.00000	146.00000	3110.00000
2	21.60000	6116.00000	22.00000	158.00000	3180.00000
3	14.90000	6396.00000	21.00000	134.00000	3373.00000
4	25.20000	8287.00000	20.00000	171.00000	3293.00000
5	22.20000	10768.00000	20.00000	180.00000	3852.00000
6	12.50000	11390.00000	20.00000	195.00000	3532.00000
7	17.70000	12173.00000	23.00000	191.00000	3366.00000
8	18.20000	12257.00000	22.00000	203.00000	3324.00000
9	6.90000	12707.00000	20.00000	192.00000	3614.00000
10	26.10000	13100.00000	21.00000	196.00000	4345.00000
11	27.20000	13108.00000	23.00000	194.00000	4287.00000
12	13.30000	13114.00000	19.00000	211.00000	3437.00000
13	22.80000	13118.00000	22.00000	197.00000	3214.00000
14	26.30000	13313.00000	25.00000	198.00000	3390.00000
15	4.20000	14056.00000	22.00000	205.00000	3624.00000
16	6.40000	15022.00000	22.00000	200.00000	3896.00000
17	26.30000	16716.00000	21.00000	205.00000	4936.00000

Subfiles

I. OBJECT OF PROGRAM:

This program allows you to specify subfiles or logical groupings of the observations. This may be accomplished by entering the number of observations in each subfile or by entering the observation number of the first observation in each subfile. Names for the subfiles are entered in both cases. A third option allows you to destroy the existing subfile structure.

II. SPECIAL CONSIDERATIONS:

A. Use of Subfiles: Subfiles may be created in order to specify logical groupings of observations. A subfile structure allows you to consider each subfile as a separate data set or to group all the subfiles together and analyze the overall data set. For example, suppose an investigator wished to measure several variables on 50 trout. He would like to analyze the data separately for each of the three varieties of the trout. He could form three separate data sets and do the individual analyses, then later join the three sets together for the overall analysis. However, since the same variables were measured on each variety of fish, this situation is well-handled by specifying a subfile for each variety. The subfile structure options make it possible to do the analysis by subfile as well as for the overall data set.

B. Certain operations in the editing and sorting programs may cause observations to move from one subfile to another. To avoid undesired results such as this, it is recommended that subfiles be specified after any editing or sorting has been carried out.

III. USER INSTRUCTIONS:

Press: The special function key labeled Subfile.

1. When the subfile characterization options are printed and "Option number=?" is displayed:

   a. Enter 1 if you desire to specify subfiles by entering the number of observations in each subfile.

   b. Press: CONT.

   c. Go to step 2.

   or

   a. Enter 2 if you desire to specify subfiles by entering the number of the first observation in each subfile.

   b. Press: CONT.

   c. Go to step 2.

   or

   a. Enter 3 if you desire to destroy the existing subfiles structure, that is, group all the data together.

   b. Press: CONT.

c. Go to step 8.

2. When "Number of subfiles (<= 10)=?" is displayed:

   a. Enter the number of subfiles which will be specified.
   b. Press: CONT.

3. When "Name of subfile #'I' (<= 10 characters)=?" is displayed:

   a. Enter the name of the Ith subfile - it may be up to 10 characters long.
   b. Press: CONT.

4. If the option number you chose was 2, go to step 6.
5. When "Subfile #'I', number of observations=?" is displayed:

   a. Enter the number of observations which will be in the Ith subfile.
   b. Press: CONT.
   c. Repeat steps 3 and 5 until all subfiles have been specified.
   d. Go to step 7.

6. When "Subfile 'I', number of first observation=?" is displayed:

   a. Enter the number of the first observation in the Ith subfile.
   b. Press: CONT.
   c. Repeat steps 3 and 6 until all subfiles have been specified.

7. When "Is above information correct?" is displayed:

   a. Press: NO if a mistake has been made while entering the required information.
   b. Go to step 2.

   or

   a. Press: YES if the subfile characterizations are as desired.

8. At this point a record of the subfiles will be printed on file "DATA" of the program medium.
9. When "SELECT ANY KEY" is displayed:

   a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.
   b. Go to step 1 of the User Instructions for the selected key.

IV. EXAMPLE:

The original data (before sorting and recoding) was grouped into two subfiles. The first 12 observations make up the first subfile (Fiscal year '76) and the remaining observations make up the second subfile (Fiscal year '77).


```
Option number = ?
2
Number of subfiles ( <=10 ) = ?
2
Name of Subfile # 1 ( <= 10 characters ) =
?
FY'76
Name of Subfile # 2 ( <= 10 characters ) =
?
FY'77
Subfile # 2 ; number of first observation =
?
13
Is above information correct?
"YES"
Subfile name:     beginning observation--number of observations
 1. FY'76                              1                        12
 2. FY'77                             13                         5

SELECT ANY KEY
```

Rename

I. OBJECT OF PROGRAM:

This program allows you to rename a data set, variables and/or subfiles. These names are then stored, along with the data, on the program medium's scratch file ("DATA").

II. USER INSTRUCTIONS:

Press: The special function key labeled Name.

1. When the title is printed and "Rename data set?" is displayed:

   a. Press: NO to leave the data set name unchanged.

   b. Go to step 3.

   or

   a. Press: YES to change the data set name.

2. When "Name of data set (<= 80 characters)=?" is displayed:

   a. Enter the project title for the data set.

   b. Press: CONT.

3. When "Rename variables?" is displayed:

   a. Press: NO if no variables are to be renamed.

   b. Go to step 9.

   or

   a. Press: YES if you desire to rename some variables.

4. When "Rename beginning with variable #?" is displayed:

   a. Enter the number of the first variable to be renamed.

   b. Press: CONT.

5. When "Rename ending with variable #?" is displayed:

   a. Enter the number of the last variable to be renamed.

   b. Press: CONT.

6. When "Name of variable #'I' (<= 10 characters)=?" is displayed:

   a. Enter the name of the Ith variable - it may be up to 10 characters long.

   b. Press: CONT.

   c. Repeat parts a and b until all specified variables have been renamed.

7. When "Is above information correct?" is displayed:

    a. Press: NO if a mistake has been made while entering variable names.
    b. Go to step 3.

    or

    a. Press: YES if the variable names are as desired.

8. At this point a list of all the current variable names will be printed.
9. If there are no subfiles in the data set, go to 16.
10. When "Rename subfiles?" is displayed:

    a. Press: NO if no subfiles are to be renamed.
    b. Go to step 16.

    or

    a. Press: YES if you desire to rename some subfiles.

11. When "Rename beginning with subfile #?" is displayed:

    a. Enter the number of the first subfile to be renamed.
    b. Press: CONT.

12. When "Rename ending with subfile #?" is displayed:

    a. Enter the number of the last subfile to be renamed.
    b. Press: CONT.

13. When "Name of subfile #'I' (< = 10 characters)=?" is displayed:

    a. Enter the name of the Ith subfile - it may be up to 10 characters long.
    b. Press: CONT.
    c. Repeat parts a and b until all specified subfiles have been renamed.

14. When "Is above information correct?" is displayed:

    a. Press: NO if a mistake has been made while entering subfile names.
    b. Go to step 10.

15. At this point a list of the current subfile names is printed.
16. The revised information is stored on file "DATA" of the program medium.

17. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

    b. Go to step 1 of the User Instructions for the selected key.

III. EXAMPLE:

The name of the data set and the name of variable four were changed.

```
Rename data set?
"YES"
Name of data set (<= 80 characters) = ?
SAMPLE DATA
Rename variables?
"YES"
Rename beginning with Variable # ?
4
Rename ending with Variable # ?
4
Name of Variable # 4 ( <= 10 characters ) =
?
Payroll(#)
Is above information correct?
"YES"

CURRENT VARIABLE NAMES
 1. Temp(C)
 2. Production
 3. Days
 4. Payroll(#)
 5. Water Use

Rename subfiles?
"NO"
SELECT ANY KEY
```

Store

I. OBJECT OF PROGRAM:

This program allows you to store the data matrix and related information in a file so that it may be retrieved at a later date for further analysis. The program also allows you to specify the file name.

II. SPECIAL CONSIDERATIONS:

A. Use of Program: The store feature will be useful in two different situations. First, if an investigator has a data set which he may want to analyze further at a later date, he may store it now using this routine and retrieve it then via the START routine. Secondly, if several people have access to the data input programs, it becomes mandatory that each be able to store his data set in a unique place. Note that if only one person uses the routine on one data set it is unnecessary to use the store feature since the data and related information are kept in "DATA" - the scratch file on the program medium.

B. The existence of a file is checked in the program in an attempt to avoid the accidental loss of existing data. Thus, when a file is specified to receive the data, an attempt is made to ensure that you are not accidentally storing the new data in a file which you did not know existed.

III. USER INSTRUCTIONS:

Press: The special function key labeled Store.

1. When the program notes are printed and "Name of data file=?" is displayed:

   a. Enter the name of the file in which the data matrix and related information are to be stored - refer to the program note for the proper form of the file name.

   b. Press: CONT.

2. When "Is data medium placed in device [msus]?" is displayed:

   a. Press: YES when the medium on which the data and related information are to be stored is in place.

3. When "Does this file already exist?" is displayed:

   a. Press: YES if the file has been previously created (and is of the form specified in the Special Considerations section of BASIC STATISTICS AND DATA MANIPULATION) and you desire to "print over" the information that is currently in the file.

   b. Go to step 4.

   or

a. Press: NO if the file has not been previously created.

NOTE: This step is an attempt to prevent you from accidentally writing over an existing file. If the response to the question in step 3 is negative and the program finds a file by this name, a note will be displayed and control will pass to step 1.

4. At this point, the file will be created if necessary and the data matrix along with the related information will be stored in the specified file. A note of the storage will be printed.
5. When "Is program medium replaced in device?" is displayed:

   a. Press: YES when the program medium is in place (or if it was never removed).

6. When "SELECT ANY KEY" is displayed:

   a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

   b. Go to step 1 of the User Instructions for the selected key.

List

I. OBJECT OF PROGRAM:

This program allows you to obtain a listing of the data matrix. The listing will appear on the device that has been specified in the START routine or in the OUTPUT UNIT routine.

II. USER INSTRUCTIONS:

Press: The special function key labeled List.

1. When the data listing options are printed and "Option number=?" is displayed:

   a. Enter 1 if you desire to obtain a listing of all data - the listing will be by observation.

   b. Press: CONT.

   c. Go to step 6.

   or

   a. Enter 2 if you desire to obtain a listing of data by variable - it may be all the data or just a set of successive variables.

   b. Press: CONT.

   c. Go to step 2.

   or

   a. Enter 3 if you desire to obtain a listing of data by observation - the observations are to be successive but need not be the entire data set.

   b. Press: CONT.

   c. Go to step 4.

2. When "List beginning with Variable #?" is displayed:

   a. Enter the number of the first variable to be included in the list.

   b. Press: CONT.

3. When "Ending with Variable #?" is displayed:

   a. Enter the number of the last variable to be included in the list (all variables between and including the first and last variables specified will be listed).

   b. Press: CONT.

   c. Go to step 6.

4. When "List beginning with Observation #?" is displayed:

    a. Enter the number of the first observation to be included in the list.

    b. Press: CONT.

5. When "Ending with Observation #?" is displayed:

    a. Enter the number of the last observation to be included in the list.

    b. Press: CONT

6. At this point the listing of the data will be printed.
7. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

    b. Go to step 1 of the User Instructions for the selected key.

Join

I. OBJECT OF PROGRAM:

This program allows you to join or combine two data sets into a single unit. One data set must be in memory and the other data set must have been previously stored by the STORE routine. Two options are available. First, observations may be added together (if both sets have the same number of variables). Second, variables may be added together (if both sets have the same number of observations).

II. SPECIAL CONSIDERATIONS:

A. Adding Observations: Suppose data on six variables was gathered in each of the 52 weeks in 1975, analyzed, and stored on an auxiliary data tape. Suppose the same variables were measured in 1976, analyzed, and stored. If the investigator is interested in combining the two sets of data together for an overall analysis, he may use the Add Observations option of the joining routine. One set of data must be retrieved via the START routine. Then, after pressing the Join key, the second set may be retrieved and the joining carried out. Notice that the variables must be in the same order in the two data sets.

B. Adding Variables: Suppose an investigator measured five variables on each of 50 mice in an experiment. These were analyzed and stored on disk. Later, he realized that three more variables were of interest. He measured these variables on the mice in the same order as before and analyzed them. All eight variables measured on each mouse could be combined into a single data set via the joining routine.

C. Subfiles: If variables are added, the subfile structure assigned to the resultant data set is the subfile structure of data set #1, that is, the data set that is in the machine prior to the joining operation. If observations are added, the following procedures are employed: 1) If no subfiles exist in either data set, the resultant set has no subfiles. 2) If data set #1 has no subfiles, but data set #2 does, then a subfile named "SET #1" is created which consists of data set #1 and the subfiles of data set #2 remain unchanged. 3) If data set #1 contains subfiles, but data set #2 does not, then a subfile named "SET #2" is created which consists of data set #2 and the subfiles of data set #1 remain unchanged. 4) If both data sets contain subfiles, all of the subfiles of data set #1 are retained and as many subfiles of data set #2 are retained as possible - the upper limit of total subfiles for the resultant set being 10.

III. USER INSTRUCTIONS:

Press: The special function key labeled Join.

1. At this point the subprogram "Join" is called. A subprogram is used here for the sole purpose of variable dimensioning.
2. When the joining options are printed and "Option number=?" is displayed:

   a. Enter 1 if you desire to add variables from a data medium to those which are currently in memory.

   b. Press: CONT.

   c. Go to step 3.

   or

   a. Enter 2 if you desire to add observations from a data medium to those which are currently in memory.

   b. Press: CONT.

3. When the requirements are printed and "OK to continue?" is displayed:

   a. Press: NO if the requirements have not been met.

   b. Go to step 16.

   or

   a. Press: YES if you feel that the requirements have been satisfied.

4. When "Project title for the combined data set (<= 80 characters)=?" is displayed:

   a. Enter up to 80 characters to be used as the name of the joined data set.

   b. Press: CONT.

5. When the program note is printed and "File name of data set #2=?" is displayed:

   a. Enter the name of the data file in which the data to be joined to the set already in memory resides. Refer to the program note for the proper form of the file name.

   b. Press: CONT.

6. When "Is data set #2 medium placed in device [msus]?" is displayed:

   a. Press: YES when the medium on which data set #2 resides is in place.

7. At this point, the number of variables, number of observations, and subfile structures are printed for data sets 1 and 2 as well as for the resultant set. The requirements are checked internally and if they have not been met, control is passed to step 16.

8. When "OK to continue?" is displayed:

    a. Press: NO if, after reviewing the summary, you realize that the resultant data set will not turn out as expected (remember, the subfile structure may be changed at a later time).

    b. Go to step 12.

    or

    a. Press: YES if the resultant data set will be as expected.

9. At this point a summary of the variable names for data sets 1 and 2 as well as for the resulting joined data set are printed.

10. When "OK to continue?" is displayed:

    a. Press: NO if, after reviewing the summary, you desire to abort the joining operation (remember, the variable names may be changed at a later time).

    b. Go to step 12.

    or

    a. Press: YES if the joining operation is to be carried out.

11. At this point the joining operation is performed and a summary of the resultant data set is printed.

12. When "Is program medium replaced in device?" is displayed:

    a. Press: YES when the program medium is in place (or if it was never removed).

13. If the joining operation was aborted, go to step 16.

14. At this point the resultant data set and related information are printed on file "DATA" of the program medium.

15. When "List data?" is displayed:

    a. Press: YES if a listing of the resultant data set is desired.

    b. Go to step 1 of the User Instructions for LIST.

    or

    a. Press: NO if no listing is desired.

16. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

    b. Go to step 1 of the User Instructions for the selected key.

IV. EXAMPLE:

Three observations were entered by keyboard and stored in file "TDATA5". Then the JOIN program added these observations to the end of the original data set to form a new set.

```
Option number = ?
2
OK to continue?
"YES"
Project title for combined data set (<= 80 characters) = ?
SAMPLE
File name of data set #2 = ?
TDATA5:T15
Is data set #2 medium placed in device T15
?
"YES"
OK to continue?
"YES"
OK to continue?
"YES"


                                        SAMPLE

Number of variables: 5
Number of observations: 20

Variable names:
 1. Temp(C)
 2. Production
 3. Days
 4. Payroll
 5. Water Use

Subfiles: NONE
Is program medium replaced in device?
"YES"
List data?
"YES"
```

SAMPLE

OBS#	Variable # 1	Variable # 2	Variable # 3	Variable # 4	Variable # 5
1	14.90000	6396.00000	21.00000	134.00000	3373.00000
2	18.40000	5736.00000	22.00000	146.00000	3110.00000
3	21.60000	6116.00000	22.00000	158.00000	3180.00000
4	25.20000	8287.00000	20.00000	171.00000	3293.00000
5	26.30000	13313.00000	25.00000	198.00000	3390.00000
6	27.20000	13108.00000	23.00000	194.00000	4287.00000
7	22.20000	10768.00000	20.00000	180.00000	3852.00000
8	17.70000	12173.00000	23.00000	191.00000	3366.00000
9	12.50000	11390.00000	20.00000	195.00000	3532.00000
10	6.90000	12707.00000	20.00000	192.00000	3614.00000
11	6.40000	15022.00000	22.00000	200.00000	3896.00000
12	13.30000	13114.00000	19.00000	211.00000	3437.00000
13	18.20000	12257.00000	22.00000	203.00000	3324.00000
14	22.80000	13118.00000	22.00000	197.00000	3214.00000
15	26.10000	13100.00000	21.00000	196.00000	4345.00000
16	26.30000	16716.00000	21.00000	205.00000	4936.00000
17	4.20000	14056.00000	22.00000	205.00000	3624.00000
18	25.30000	9315.00000	20.00000	183.00000	3356.00000
19	12.40000	11298.00000	19.00000	203.00000	4205.00000
20	18.60000	14653.00000	21.00000	189.00000	4256.00000

Output Unit

I. OBJECT OF PROGRAM:

This program allows you to change the device on which the hard-copy output will be printed, or conversely, to specify that no hard copy is desired.

In lieu of this routine, the hard-copy device may be changed at any time by the following procedure:

Non– HP-IB Printer

1. Type: Hc=
2. Key in the select code of the printer desired.
3. Press: EXECUTE
4. Type: Hcbus=999
5. Press: EXECUTE

HP-IB Printer

1. Type: Hc=
2. Key in select code of the printer desired.
3. Press: EXECUTE
4. Type: Hcbus=
5. Key in the bus address of the HP-IB device.
6. Press: EXECUTE

"HP-IB" stands for Hewlett-Packard Interface Bus and is our version of IEEE Standard 488, 1975 which is a universal standard interface. The select code may be found on the interface card and the bus address (if the device is an HP-IB device) is found on the back of the device itself in binary form. For further information, consult the manual for the output device.

II. USER INSTRUCTIONS:

Press: The special function key labeled Output Unit.

1. When the hard-copy options are printed and "Option number=?" is displayed:

   a. Enter 1 if no hard copy is desired.

   b. Press: CONT.

   c. Go to step 5.

   or

   a. Enter 2 if a hard copy is desired from an external printer.

   b. Press: CONT.

2. When "Are you using an HP-IB Printer?" is displayed:

   a. Press: YES if an HP-IB printer is being used.

   b. Go to step 3.

   or

   a. Press: NO if the printer is not HP-IB.

   b. Go to step 4.

3. When "Printer select code, bus address=? (for example, 7,1)" is displayed:

   a. Enter the select code, bus address for the HP-IB printer.

   b. Press: CONT.

   c. Go to step 5.

4. When "Printer select code=?" is displayed:

   a. Enter the select code of the printer desired for hard copy.

   b. Press: CONT.

5. When "SELECT ANY KEY" is displayed:

   a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

   b. Go to step 1 of the User Instructions for the selected key.

Basic Statistics

I. OBJECT OF PROGRAM:

This program computes a variety of summary statistics for data which was entered via the BASIC STATISTICS AND DATA MANIPULATION program set. The statistics may be computed by subfile or for the entire data set (ignoring subfiles). Basic statistics which are computed include: number of observations, number of missing values, sum, mean, variance, standard deviation, coefficient of skewness, coefficient of kurtosis, coefficient of variation, standard error of the mean, and confidence interval of the mean. An option is available to obtain the correlation matrix. Order statistics computed include: the maximum, the minimum, range, and midrange. Additional order statistics which may be obtained include: the median, 25th percentile, 75th percentile, Tukey's middlemeans, and user-specified percentiles.

II. SPECIAL CONSIDERATIONS:

A. If a hard copy of the statistics is not being made, the program halts occasionally so that you may study the results. In this case, it is necessary only to press CONTINUE to continue program execution.

B. If the option to obtain additional order statistics (Tukey's middlemeans and percentiles) is exercised, the data matrix is sorted and the observations of each variable are arranged in ascending order. At the end of the program the original data matrix is reloaded. Thus, if the program is aborted, that is, if another key is pressed before the reloading can occur, the data matrix will be in the sorted state. Hence, if the portion of the program used to calculate additional order statistics is accessed, abortion of the program should be discouraged.

III. METHODS AND FORMULAE:

A. Let N(J) be the number of observations of the Jth variable in the data set or subfile, whichever is applicable. Let D(I,J) be the Jth observations of the Ith variable. The following formulas are computed for the Ith variable.

1. Sum: $S(I) = \sum_{J=1}^{N(I)} D(I,J)$

2. Mean: $M(I) = \frac{S(I)}{N(I)}$

3. Variance: $V(I) = \frac{\sum_{J=1}^{N(I)} (D(I,J))^2 - N(I)(M(I))^2}{N(I) - 1}$

4. Standard deviation: $Sd(I) = (V(I))^{1/2}$

5. Second moment: $M_0(I) = \frac{(N(I) - 1)\, V(I)}{N(I)}$

6. Skewness: $Sk(I) = \frac{\sum_{J=1}^{N(I)} (D(I,J))^3 - 3M(I) \sum_{J=1}^{N(I)} (D(I,J))^2 + (2M(I))^3}{(M_0(I))^{3/2}\, N(I)}$

7. Kurtosis:

$$K(I) = \frac{\sum_{J=1}^{N(I)} (D(I,J))^4 - 4M(I) \sum_{J=1}^{N(I)} (D(I,J))^3 + 6(M(I))^2 \sum_{J=1}^{N(I)} (D(I,J))^2 - 3(M(I))^4 N(I)}{(M_0(I))^2 N(I)} - 3$$

B. Let C be the confidence coefficient for a confidence interval on the mean. The following operations are used to obtain the desired t-value.

$$P = \frac{1 - \frac{c}{100}}{2}$$

$$V = \left(\ln\left(\frac{1}{p^2}\right)\right)^{1/2}$$

$$X = 2.5155174 + .802853V + .010328V^2$$

$$Y = 1 + 1.432788V + .189269V^2 + .001308V^3$$

$$Z = V - \frac{X}{Y}$$

$$M = N(J) - 1$$

1. Then the desired t-value is:

$$T = Z + \frac{Z^3 + Z}{4M} + \frac{5Z^5 + 16Z^3 + 3Z}{96M^2} + \frac{3Z^7 + 19Z^5 + 17Z^3 - 15Z}{384M^3} + \frac{79Z^9 + 776Z^7 + 1482Z^5 - 1920Z^3 - 945Z}{92160M^4}$$

2. Standard error: $Se(I) = \frac{(V(I))^{1/2}}{(N(I))^{1/2}}$

3. Confidence interval on mean: $M(I) \pm T(Se(I))$

4. Coefficient of variation: $Cv(I) = \left|\frac{(V(I))^{1/2}}{(M(J)}\right| (100)$

5. Correlations: Suppose we have the following data matrix:

		OBSERVATION				
		1	2	3	4	5
VARIABLE	1	5	M	3	4	5
	2	6	7	M	6	4
	3	1	3	2	1	1

An M denotes a missing value. When computing the correlation between variables 1 and 2, we discard observations 2 and 3 since variable 1 is missing a data value for

observation 2 and variable 2 is missing the data value for observation 3. However, when computing the correlation between variables 1 and 3, we need only discard observation 2. Similarly, the correlation between 2 and 3 is computed by discarding only observation 3. Hence, the correlations may be based on different numbers of observations. An observation is thrown out if and only if a data value from that observation is missing from one of the two variables for which the correlation is being computed. With this in mind, let N(I,J) be the number of observations used to compute the correlation between variables I and J. Then, the correlation is:

$$C(I,J) = \frac{\sum_{K=1}^{N(I,J)} D(I,K)D(J,K) - \frac{\sum_{K=1}^{N(I,J)} D(I,K) \sum_{K=1}^{N(I,J)} D(J,K)}{N(I,J)}}{\left[\sum_{K=1}^{N(I,J)} (D(I,K))^2 - \frac{\sum_{K=1}^{N(I,J)} D(I,K)^2}{N(I,J)}\right]^{\frac{1}{2}} \left[\sum_{K=1}^{N(I,J)} (D(J,K))^2 - \frac{\sum_{K=1}^{N(I,J)} D(I,K)^2}{N(I,J)}\right]^{\frac{1}{2}}}$$

C. Let M(I) be the largest data value of the Ith variable, m(I) be the smallest data value of the Ith variable.

1. Range: $R(I) = M(I) - m(I)$
2. Midrange: $Mr(I) = \frac{M(I) + m(I)}{2}$
3. The percentiles are computed as follows: Let P be the percentile in question. If P*N(I) is an integer, then $P(I) = \frac{D(I,P*N(I)) + D(1,Q)}{2}$, where Q is the next integer value between P*N(I) and the observation index of the median. If P*N(I) is not an integer, then P(I)=D(I,N(I)*P+Q) where

$$Q = \begin{cases} 1 \text{ if } P \leq 50 \\ -1 \text{ if } P > 50 \end{cases}$$

The median refers to the 50th percentile.

D. Tukey's Middlemeans:

1. Midmean: $Mm(I) = \frac{1}{N} \Sigma$ all observations between (and including, if applicable) 25th and 75th percentile.
2. Trimean: Tm(I) = ¼ (25th percentile + 2(median) + 75th percentile).
3. Midspread: Ms(I) = 75th percentile − 25th percentile.

IV. USER INSTRUCTIONS:

Press: The special function key labeled Stats.

1. If there is only one variable in the data set, go to step 5.
2. When "Summary statistics on all variables?" is displayed:

a. Press: YES if summary statistics are desired for all of the variables in the data set.

b. Go to step 5.

or

a. Press: NO if summary statistics are desired for only one variable or a set of successive variables in the data set.

3. When "Number of first variable=?" is displayed:

a. Enter the number of the first variable for which summary statistics are desired.

b. Press: CONT.

4. When "Number of last variable=?" is displayed:

a. Enter the number of the last variable for which summary statistics are desired. Summary statistics will be calculated for all variables between and including the first and last specified.

b. Press: CONT.

5. If the data set contains no subfiles, go to step 9.

6. When the subfile structure options are printed and "Option number=?" is displayed:

a. Enter 1 if you desire to ignore the subfile structure, that is, to group all the observations together and calculate summary statistics on the data set as a whole.

b. Press: CONT.

c. Go to step 9.

or

a. Enter 2 if you desire to calculate summary statistics for only one subfile or for each of a set of adjacent subfiles; for example, for subfiles 3 through 5. The statistics will be calculated by subfile, that is, calculated as though each subfile were a separate data set.

b. Press: CONT.

c. Go to step 7.

or

a. Enter 3 if you desire to calculate summary statistics for all of the subfiles. The statistics will be calculated as though each subfile were a separate data set.

b. Press: CONT.

c. Go to step 9.

7. When "Number of first subfile=?" is displayed:

a. Enter the number of the first subfile for which summary statistics will be calculated.

b. Press: CONT.

8. When "Number of last subfile=?" is displayed:

   a. Enter the number of the last subfile for which summary statistics will be calculated. Statistics will be calculated for all subfiles between and including the first and last subfiles specified as though they were separate data sets.

   b. Press: CONT.

9. At this point a heading and basic statistics will be output. If no hard-copy printer has been specified, the program will pause occasionally to allow examination of the output and "Press 'CONT' when ready." will be displayed. To resume execution, press CONTINUE.

10. When "Confidence coefficient for confidence interval on the mean=?" is displayed:

    a. Enter the confidence coefficient or confidence level to be used in constructing a confidence interval on the mean; for example, enter 95 for a 95% confidence on the mean.

    b. Press: CONT.

11. At this point several more statistics will be output. If the output is to the screen, it will be necessary to press CONTINUE to resume program execution after examining the output.

12. If summary statistics are being computed for a single variable, go to step 15.

13. When "Correlation matrix?" is displayed:

    a. Press: NO if correlations are not desired.

    b. Go to step 15.

    or

    a. Press: YES to obtain a matrix of correlations among the variables.

14. At this point the correlation matrix will be computed and output. If output is to the screen, it will be necessary to press CONTINUE to resume program execution after examining the correlations.

15. Several order statistics are now computed and output.

16. When "More order statistics (Tukey's middlemeans & percentiles)?" is displayed:

    a. Press: NO if no further order statistics are desired.

    b. Go to step 21.

    or

    a. Press: YES if Tukey's middlemeans (see Methods and Formulae), the 25th, 50th, and 75th percentiles, are desired.

17. At this point the data matrix is sorted by variable. Then, several order statistics are computed and output.

18. When "Other percentiles?" is displayed:

    a. Press: NO if no further percentiles are desired.

    b. Go to step 21.

    or

    a. Press: YES if additional percentiles are desired.

19. When "Press 'NO' when finished - Percentile=?" is displayed:

    a. Enter the percentile which is desired; for example, enter 90 to obtain the 90th percentile. The number entered must be between 1 and 99, inclusive.

    b. Press: CONT.

    or

    a. Press: NO if no more percentiles are desired.

    b. Go to step 21.

20. At this point, the desired percentile is computed for each variable under consideration and the results are output. Go to step 19.

21. If summary statistics are being computed for more than one subfile, steps 9-20 are repeated for each subfile.

22. If the additional order statistics were computed for the data set (or any subfile), the original data matrix is reloaded at this point.

23. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys identified on the "REGRESSION ANALYSIS METHODS" overlay.

    b. Go to step 1 of the User Instructions for the selected key.

V. EXAMPLE:

The following sample outputs were obtained using the original data set which was divided into two subfiles. For the first sample, subfile structures were ignored. The second sample shows statistics for each subfile.

```
*****************************************************************************
*                              SUMMARY STATISTICS                           *
*                                 ON DATA SET:                              *
*                                    SAMPLE                                 *
*****************************************************************************

-----------------------------------------------------------------------------
(SUBFILES IGNORED)
-----------------------------------------------------------------------------

                                BASIC STATISTICS

VARIABLE      # OBSERVATIONS    # MISS. VALUES            SUM             MEAN
Temp(C)                  17                 0       310.20000         18.24706
Production               17                 0    197377.00000      11610.41176
Days                     17                 0       365.00000         21.47059
Payroll                  17                 0      3176.00000        186.82353
Water Use                17                 0     61773.00000       3633.70588
-----------------------------------------------------------------------------

                                                      COEF OF          COEF OF
VARIABLE            VARIANCE    STANDARD DEV.       SKEWNESS         KURTOSIS
Temp(C)             56.43265         7.51217         -.52698          -.93974
Production   10075542.13250      3174.19945         -.68503          -.48477
Days                 2.13971         1.46277          .49186           .23008
Payroll            483.77941        21.99499        -1.25111           .41009
Water Use       241372.97063       491.29723         1.32313           .99803
-----------------------------------------------------------------------------
Confidence coefficient for confidence interval on the mean = ?
95
                              STANDARD ERROR    95 % CONFIDENCE INTERVAL ON MEAN
VARIABLE      COEF VARIATION     OF THE MEAN       LOWER LIMIT      UPPER LIMIT
Temp(C)              41.16919         1.82197          14.38370         22.11042
Production           27.33925       769.85645        9977.98475      13242.83878
Days                  6.81292          .35477          20.71831         22.22286
Payroll              11.77313         5.33457         175.51195        198.13511
Water Use            13.52056       119.15708        3381.04158       3886.37018
-----------------------------------------------------------------------------
Correlation matrix?
"YES"
                              CORRELATION MATRIX

             Production         Days      Payroll   Water Use
Temp(C)       -.0923502    .2685939    -.1073726    .2503744
Production                 .1057436     .9184717    .6308869
Days                                    .0318812   -.0887635
Payroll                                             .4134486

-----------------------------------------------------------------------------
```

ORDER STATISTICS

VARIABLE	MAXIMUM	MINIMUM	RANGE	MIDRANGE
Temp(C)	27.20000	4.20000	23.00000	15.70000
Production	16716.00000	5736.00000	10980.00000	11226.00000
Days	25.00000	19.00000	6.00000	22.00000
Payroll	211.00000	134.00000	77.00000	172.50000
Water Use	4936.00000	3110.00000	1826.00000	4023.00000

More order statistics (Tukey's middlemeans & percentiles)?
"YES"

		TUKEY'S HINGES	
VARIABLE	MEDIAN	25-th %-ile	75-th %-ile
Temp(C)	18.40000	13.30000	22.80000
Production	12707.00000	10768.00000	13114.00000
Days	22.00000	20.00000	22.00000
Payroll	195.00000	180.00000	198.00000
Water Use	3437.00000	3324.00000	3624.00000

	TUKEY'S MIDDLEMEANS		
VARIABLE	MIDMEAN	TRIMEAN	MIDSPREAD
Temp(C)	19.36667	18.22500	9.50000
Production	12415.00000	12324.00000	2346.00000
Days	21.44444	21.50000	2.00000
Payroll	193.66667	192.00000	18.00000
Water Use	3501.33333	3455.50000	300.00000

Other percentiles?
"YES"
Press 'NO' when finished--Percentile = ?
10

VARIABLE	10-th percentile
Temp(C)	6.40000
Production	6116.00000
Days	20.00000
Payroll	146.00000
Water Use	3180.00000

Press 'NO' when finished--Percentile = ?
90

VARIABLE	90-th percentile
Temp(C)	26.30000
Production	14056.00000
Days	23.00000
Payroll	205.00000
Water Use	4287.00000

Press 'NO' when finished--Percentile = ?
"NO"
SELECT ANY KEY

```
*******************************************************************************
*                              SUMMARY STATISTICS                             *
*                                 ON DATA SET:                                *
*                                    SAMPLE                                   *
*******************************************************************************


-------------------------------------------------------------------------------
Subfile: FY'76
-------------------------------------------------------------------------------

                                BASIC STATISTICS

VARIABLE      # OBSERVATIONS    # MISS. VALUES              SUM             MEAN
Temp(C)                  12                 0        212.60000         17.71667
Production               12                 0     128130.00000      10677.50000
Days                     12                 0        257.00000         21.41667
Payroll                  12                 0       2170.00000        180.83333
Water Use                12                 0      42330.00000       3527.50000
-------------------------------------------------------------------------------
                                                        COEF OF         COEF OF
VARIABLE            VARIANCE     STANDARD DEV.        SKEWNESS        KURTOSIS
Temp(C)              50.45242          7.10299          -.24164        -1.10813
Production     10346097.90910       3216.53508          -.47514        -1.24452
Days                  2.99242          1.72986           .53169         -.47442
Payroll             563.60606         23.74039          -.78573         -.59264
Water Use        113867.00000        337.44185           .94917          .15129
-------------------------------------------------------------------------------
Confidence coefficient for confidence interval on the mean = ?
95
                                STANDARD ERROR    95 % CONFIDENCE INTERVAL ON MEAN
VARIABLE      COEF VARIATION      OF THE MEAN       LOWER LIMIT       UPPER LIMIT
Temp(C)              40.09212          2.05046          13.20247          22.23086
Production           30.12442        928.53370        8633.28011       12721.71989
Days                  8.07718           .49937          20.31728          22.51605
Payroll              13.12833          6.85326         165.74549         195.92118
Water Use             9.56603         97.41107        3313.04397        3741.95603
-------------------------------------------------------------------------------
Correlation matrix?
"YES"
                              CORRELATION MATRIX

             Production          Days       Payroll    Water Use
Temp(C)       -.2644635     .4396054     -.1782662     -.0077412
Production                  .1294077      .9312213      .6450021
Days                                      .0040583      .0388568
Payroll                                                 .4892023
Payroll          211.00000          134.00000            77.00000          172.50000
Water Use       4287.00000         3110.00000          1177.00000         3698.50000
-------------------------------------------------------------------------------
```

ORDER STATISTICS

VARIABLE	MAXIMUM	MINIMUM	RANGE	MIDRANGE
Temp(C)	27.20000	6.40000	20.80000	16.80000
Production	15022.00000	5736.00000	9286.00000	10379.00000
Days	25.00000	19.00000	6.00000	22.00000

More order statistics (Tukey's middlemeans & percentiles)?
"YES"

	TUKEY'S HINGES		
VARIABLE	MEDIAN	25-th %-ile	75-th %-ile
Temp(C)	18.05000	12.90000	23.70000
Production	11781.50000	7341.50000	13111.00000
Days	21.50000	20.00000	22.50000
Payroll	191.50000	164.50000	196.50000
Water Use	3413.50000	3329.50000	3733.00000

	TUKEY'S MIDDLEMEANS		
VARIABLE	MIDMEAN	TRIMEAN	MIDSPREAD
Temp(C)	18.01667	18.17500	10.80000
Production	11405.50000	11003.87500	5769.50000
Days	21.16667	21.37500	2.50000
Payroll	187.16667	186.00000	32.00000
Water Use	3452.00000	3472.37500	403.50000

Other percentiles?
"YES"
Press 'NO' when finished--Percentile = ?
10

VARIABLE	10-th percentile
Temp(C)	6.90000
Production	6116.00000
Days	20.00000
Payroll	146.00000
Water Use	3180.00000

Press 'NO' when finished--Percentile = ?
90

VARIABLE	90-th percentile
Temp(C)	25.20000
Production	13114.00000
Days	23.00000
Payroll	198.00000
Water Use	3852.00000

Press 'NO' when finished--Percentile = ?
"NO"

Subfile: FY'77

BASIC STATISTICS

VARIABLE	# OBSERVATIONS	# MISS. VALUES	SUM	MEAN
Temp(C)	5	0	97.60000	19.52000
Production	5	0	69247.00000	13849.40000
Days	5	0	108.00000	21.60000
Payroll	5	0	1006.00000	201.20000
Water Use	5	0	19443.00000	3888.60000

VARIABLE	VARIANCE	STANDARD DEV.	COEF OF SKEWNESS	COEF OF KURTOSIS
Temp(C)	84.11700	9.17153	-1.07462	-.38577
Production	2973090.80000	1724.26529	1.02039	-.36407
Days	.30000	.54772	-.40825	-1.83333
Payroll	19.20000	4.38178	-.32931	-1.75395
Water Use	537304.80000	733.01078	.52418	-1.31081

Confidence coefficient for confidence interval on the mean = ?
95

VARIABLE	COEF VARIATION	STANDARD ERROR OF THE MEAN	95 % CONFIDENCE INTERVAL ON MEAN LOWER LIMIT	UPPER LIMIT
Temp(C)	46.98531	4.10163	8.13155	30.90845
Production	12.45011	771.11488	11708.34913	15990.45087
Days	2.53575	.24495	20.91988	22.28012
Payroll	2.17782	1.95959	195.75907	206.64093
Water Use	18.85025	327.81239	2978.40748	4798.79252

Correlation matrix?
"YES"

CORRELATION MATRIX

	Production	Days	Payroll	Water Use
Temp(C)	.1745645	-.6648810	-.5152082	.4719269
Production		-.5604504	.5221663	.7908537
Days			.1458333	-.9363950
Payroll				.1882379

ORDER STATISTICS

VARIABLE	MAXIMUM	MINIMUM	RANGE	MIDRANGE
Temp(C)	26.30000	4.20000	22.10000	15.25000
Production	16716.00000	12257.00000	4459.00000	14486.50000
Days	22.00000	21.00000	1.00000	21.50000
Payroll	205.00000	196.00000	9.00000	200.50000
Water Use	4936.00000	3214.00000	1722.00000	4075.00000

More order statistics (Tukey's middlemeans & percentiles)?
"YES"

		TUKEY'S HINGES	
VARIABLE	MEDIAN	25-th %-ile	75-th %-ile
Temp(C)	22.80000	18.20000	22.80000
Production	13118.00000	13100.00000	13118.00000
Days	22.00000	21.00000	22.00000
Payroll	203.00000	197.00000	203.00000
Water Use	3624.00000	3324.00000	3624.00000

	TUKEY'S MIDDLEMEANS		
VARIABLE	MIDMEAN	TRIMEAN	MIDSPREAD
Temp(C)	22.36667	21.65000	4.60000
Production	13424.66667	13113.50000	18.00000
Days	21.66667	21.75000	1.00000
Payroll	201.66667	201.50000	6.00000
Water Use	3764.33333	3549.00000	300.00000

Other percentiles?
"YES"
Press 'NO' when finished--Percentile = ?
10

VARIABLE	10-th percentile
Temp(C)	4.20000
Production	12257.00000
Days	21.00000
Payroll	196.00000
Water Use	3214.00000

Press 'NO' when finished--Percentile = ?
90

VARIABLE	90-th percentile
Temp(C)	26.10000
Production	14056.00000
Days	22.00000
Payroll	205.00000
Water Use	4345.00000

Press 'NO' when finished--Percentile = ?
"NO"
SELECT ANY KEY

Regression Analysis Methods

I. OBJECT OF PROGRAMS:

The regression package is made up of three regression routines — 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.

II. SPECIAL CONSIDERATIONS:

A. Limitations: The program is capable of handling 20 variables with a total of 1500 data points, since these are the maximum numbers allowed by the BASIC STATISTICS AND DATA MANIPULATION routine. If one of the larger memory options is purchased and larger problems arise, it is possible to alter the programs so that more than 20 variables may be used in the regression. First, the general data input routine must be altered as shown in the Special Considerations section of the BASIC STATISTICS AND DATA MANIPULATION routine. The alterations necessary to each regression program are detailed in the Special Considerations section of that program.

B. Graphics: If the 9835A is not equipped with a 98337A Plotter ROM, two of the programs in this section will behave somewhat abnormally: POLYNOMIAL REGRESSION and RESIDUAL ANALYSIS. Please refer to the User Instructions for those sections for details.

6. When "F-value for inclusion = ?" is displayed:

   a. Enter the F value for inclusion, F1. A variable must have an F-to-enter greater than F1 for entry into the regression model via the stepwise or forward selection procedures. A typical value is 4.

   b. Press: CONT.

7. If the forward selection procedure was chosen, go to 9.

8. When "F-value for deletion = ?" is displayed:

   a. Enter the F value for deletion, F2. A variable may be deleted from the regression via the stepwise or backward selection procedures only if its F-to-delete is less than F2. A typical value is 4. If using the stepwise procedure, we must have F1 $\geq$ F2.

   b. Press: CONT.

9. When the information input in steps 3-8 is printed and "Is above information correct?" is displayed:

   a. Press: NO if changes are necessary.

   b. Go to 3.

   or

   a. Press: YES if no changes are required.

10. When "Number of the dependent variable = ?" is displayed:

    a. Enter the number corresponding to the dependent variable, i.e., if the model is to be $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$, enter Y.

    b. Press: CONT.

11. When "Of the remaining 'I' variables, number to be considered in the regression = ?" is displayed:

    a. Enter the number of variables to be considered as **candidates for independent variables**. Clearly, the response must be $\leq$ I.

    b. Press: CONT.

12. When "Variable numbers (separated by commas) = ?" is displayed:

    a. Enter the numbers corresponding to candidates for independent variables separated by commas. There should be as many entries here as are specified in step 11.

    b. Press: CONT.

13. When the variable numbers are printed and "Is above information correct?" is displayed:

    a. Press: NO if changes are necessary.

b. Go to 10.

or

a. Press: YES if no changes are required.

14. A summary of the regression problem to be performed will be printed.
15. When the methods of computing correlations are printed and "Method number = ?" is displayed:

    a. Enter 1 if there are missing values and it is desired to compute correlations by using an observation only if data values are present for all variables. (See Special Considerations for an example.)

    b. Press: CONT.

    or

    a. Enter 2 if there are no missing values, or, if it is desired to use every data value possible to compute each correlation. (See Special Considerations for an example.)

16. At this point the correlation matrix is computed and output. Then the regression is carried out automatically as far as possible. The manual variable selection mode is entered.
17. When "Input 'K', delete '−K', or, enter Ø to end regression...." is displayed:

    a. Enter Ø to end the regression.

    b. Press: CONT.

    c. Go to 18.

    or

    a. Enter the number of the variable to be included in the model.

    b. Press: CONT.

    c. The variable will be added to the model and associated output will be printed.

    d. Go to 17.

    or

    a. Enter the number of the variable to be deleted from the regression **preceded by a negative sign**.

    b. Press: CONT.

    c. The variable will be deleted from the model and the associated output will be printed.

    d. Go to 17.

18. When "Residual analysis?" is printed:

    a. Press: NO if no residual analysis is desired.

b. Go to 19.

or

a. Press: YES if a residual analysis is desired.

b. Go to step 1 of the RESIDUAL ANALYSIS User Instructions.

19. When "SELECT ANY KEY" is displayed:

a. Press: Any of the keys defined on the "REGRESSION ANALYSIS METHODS" overlay.

b. Go to the User Instructions for the selected key.

V. EXAMPLE:

Following are the results from the stepwise and backward selection procedures. The data used is the same as used in MULTIPLE LINEAR REGRESSION.

```
*******************************************************************************
                              STEPWISE REGRESSION ON
                                SAMPLE PROBLEM
*******************************************************************************

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        X2^2       X1*X2           Y
X1               1.0000000   0.0000000   .9747877   0.0000000   .8120711  -.4209438
X2                           1.0000000  0.0000000   .9897433   .4802402   .5916875
X1^2                                    1.0000000  0.0000000   .7915969  -.3905355
X2^2                                               1.0000000   .4753145   .6250961
X1*X2                                                          1.0000000  -.2314209
Y                                                                         1.0000000

*******************************************************************************

                 F TO  PART         F TO        REGRESSION COEFFICIENTS        STD
#--VARIABLE     ENTER  CORR    TOL DELETE  STD.FORMAT             E-FORMAT     ERROR
 1.X1             1.51  .421 1.000
 2.X2             3.77  .592 1.000
 4.X1^2           1.26  .391 1.000
 5.X2^2           4.49  .625 1.000
 6.X1*X2           .40  .231 1.000

*******************************************************************************
STEP NUMBER 1
VARIABLE'X2^2' ADDED
R-SQUARED =  .3907451639

                                      AOV

SOURCE              DF      SUM OF SQUARES         MEAN SQUARE         F-VALUE
TOTAL                8              .20052
REGRESSION           1              .07835              .07835            4.49
RESIDUAL             7              .12217              .01745

STANDARD ERROR =  .132107402858

                 F TO  PART         F TO        REGRESSION COEFFICIENTS        STD
#--VARIABLE     ENTER  CORR    TOL DELETE  STD.FORMAT             E-FORMAT     ERROR
 1.X1             2.46  .539 1.000
 2.X2              .37  .242  .020
 4.X1^2           2.00  .500 1.000
 5.X2^2                              4.49       .00177  .176721938776E-02     .0008
 6.X1*X2          8.72  .770  .774
Constant = -.04761904762
```

D. The program is currently capable of plotting on the HP 9872A plotter. If some other plotter is to be used, say "INCPLOT", then change line 480 to PLOTTER IS Psc, Bn, "INCPLOT", where Psc is the plotter select code and Bn is the bus number or device address.

E. If your machine is not equipped with the 98337A Plotter ROM when this program is loaded, a series of error messages will appear in the display, accompanied by an audible tone. The program will then operate normally (after pressing RUN), provided no plotting is requested.

III. 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 inversion method produces less round-off error than other procedures. This method, as well as all other methods and formulae may be found in F.A. Graybill's Theory and Application of the Linear Modes.

IV. USER INSTRUCTIONS:

Press: The special function key labeled Poly. Regr.

1. If the data set currently in memory contains no subfiles, go to 3.
2. When "Subfile # (enter Ø to ignore subfiles) = ?" is displayed:

   a. Enter the number of the subfile for which the regression is desired.

   b. Press: CONT.

   c. Go to 3.

   or

   a. Enter Ø to ignore the subfile structure, i.e., to carry out the regression on the data set as a whole.

   b. Press: CONT.

3. When "Number of the dependent variable = ?" is displayed:

   a. Enter the number corresponding to the dependent variable, i.e., if the model is $Y = \beta_0 + \beta_1 X + \ldots + \beta_p X^p$, enter the number corresponding to variable Y.

   b. Press: CONT.

4. When "Number of the independent variable = ?" is displayed:

   a. Enter the number corresponding to the independent variable, i.e., for the model $Y = \beta_0 + \beta_1 X + \ldots + \beta_p X^p$, enter the number corresponding to variable X.

   b. Press: CONT.

5. A copy of the above input will be printed.
6. When "Is a plot of the regression desired?" is displayed:

a. Press: NO if no plotting hardware is available or if no plot of the data and regression curve is desired.

b. Go to 20.

or

a. Press: YES if plotting capabilities are available and an X-Y scatter plot as well as the regression curve is desired.

7. When "Plotter select code, Bus # = ?" is displayed:

a. Enter the select code of the plotting device followed by its bus number (separated by a comma).

7,7

b. Press: CONT.

8. When the maximums and minimums are printed and "X-min = ?" is displayed:

a. Refer to the printed information and enter the X value at which the X-axis will begin.

b. Press: CONT.

9. When "X-max = ?" is displayed:

a. Enter the X value at which the X-axis will end.

b. Press: CONT.

10. When "Y-min = ?" is displayed:

a. Enter the Y value at which the Y-axis will begin.

b. Press: CONT.

11. When "Y-max = ?" is displayed:

a. Enter the Y value at which the Y-axis will end.

b. Press: CONT.

12. When "Y-axis crosses X-axis at X = ?" is displayed:

a. Enter the X value at which the Y-axis will cross the X-axis, e.g., this could be X-min.

b. Press: CONT.

13. When "X-axis crosses Y-axis at Y = ?" is displayed:

a. Enter the Y value at which the X-axis will cross the Y-axis, e.g., this could be Y-min.

b. Press: CONT.

14. When "Distance between X-ticks = ?" is displayed:

29. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys defined on the "REGRESSION ANALYSIS METHODS" overlay.

    b. Go to step 1 of the User Instructions for the selected key.

V. EXAMPLE:

Bus Passenger Service Time

The time required to service boarding passengers at a bus stop was measured together with the actual number of passengers boarding. The service time was recorded from the moment that the bus stopped and the door opened until the last passenger boarded the bus. The objective is to determine a model for predicting passenger service time, given knowledge of the number boarding at a particular stop. Let Variable 1 = number boarding and Variable 2 = passenger service time. The following data was gathered during the month of May, 1968, at twelve downtown locations in Louisville, Kentucky.

This data is found in file "DATA3" of the program medium.

```
*******************************************************************************
*                                DATA MANIPULATION                            *
*******************************************************************************
                            Bus Passenger Service Time

Data file name: DATA3:T15
Number of observations: 31
Number of variables: 2

Variables names:
   1. NUMBER
   2. TIME

Subfiles:  NONE
```

Bus Passenger Service Time

OBS#	Variable # 1	Variable # 2
1	1.00000	1.40000
2	1.00000	2.80000
3	1.00000	3.00000
4	1.00000	1.80000
5	1.00000	2.00000
6	2.00000	4.70000
7	2.00000	8.00000
8	2.00000	3.00000
9	2.00000	2.50000
10	3.00000	5.20000
11	3.00000	6.20000
12	3.00000	9.40000
13	4.00000	11.70000
14	5.00000	7.50000
15	5.00000	11.90000
16	6.00000	13.60000
17	6.00000	12.40000
18	6.00000	11.60000
19	7.00000	14.70000
20	7.00000	13.50000
21	8.00000	12.00000
22	8.00000	14.10000
23	8.00000	26.00000
24	9.00000	19.00000
25	10.00000	21.20000

```
26
        11.00000        22.90000
27
        11.00000        22.60000
28
        13.00000        25.20000
29
        17.00000        33.50000
30
        19.00000        33.70000
31
        25.00000        54.20000
```

****
* SUMMARY STATISTICS *
* ON DATA SET: *
* Bus Passenger Service Time *
****

BASIC STATISTICS

VARIABLE	# OBSERVATIONS	# MISS. VALUES	SUM	MEAN
NUMBER	31	0	207.00000	6.67742
TIME	31	0	431.30000	13.91290

VARIABLE	VARIANCE	STANDARD DEV.	COEF OF SKEWNESS	COEF OF KURTOSIS
NUMBER	33.22581	5.76418	1.43125	1.90790
TIME	139.39983	11.80677	1.48977	2.55645

VARIABLE	COEF VARIATION	STANDARD ERROR OF THE MEAN	95 % CONFIDENCE INTERVAL ON MEAN LOWER LIMIT	UPPER LIMIT
NUMBER	86.32351	1.03528	4.56260	8.79223
TIME	84.86202	2.12056	9.58113	18.24468

CORRELATION MATRIX

	TIME
NUMBER	.9743533

ORDER STATISTICS

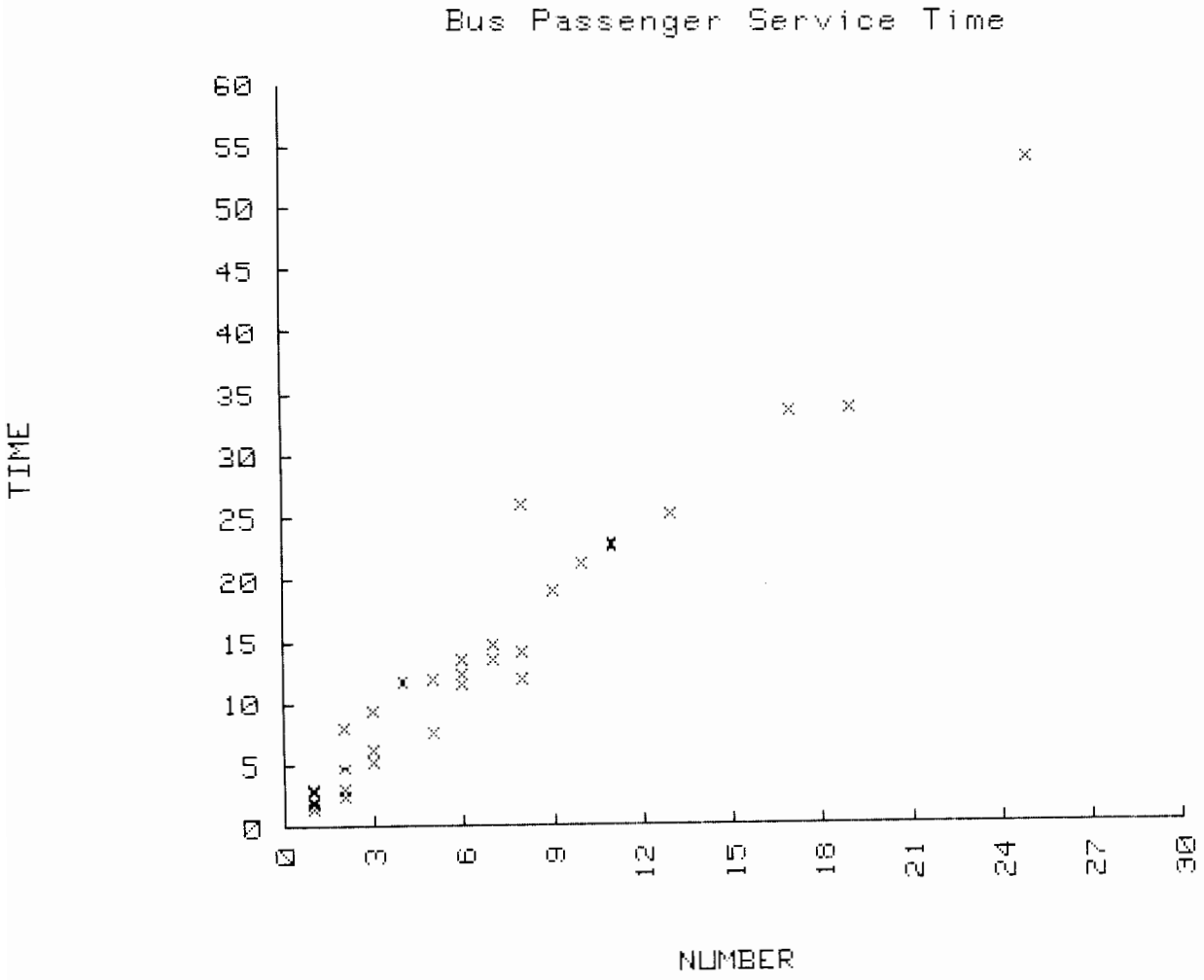
VARIABLE	MAXIMUM	MINIMUM	RANGE	MIDRANGE
NUMBER	25.00000	1.00000	24.00000	13.00000
TIME	54.20000	1.40000	52.80000	27.80000

	TUKEY'S HINGES		
VARIABLE	MEDIAN	25-th %-ile	75-th %-ile
NUMBER	6.00000	2.00000	8.00000
TIME	11.90000	4.70000	19.00000

	TUKEY'S MIDDLEMEANS		
VARIABLE	MIDMEAN	TRIMEAN	MIDSPREAD
NUMBER	5.41176	5.50000	6.00000
TIME	11.57059	11.87500	14.30000

```
*******************************************************************************
POLYNOMIAL REGRESSION ON DATA SET:
                              Bus Passenger Service Time
*******************************************************************************

--where:  Dependent variable = TIME
          Independent variable = NUMBER
```



VARIABLE	N	MEAN	VARIANCE	STANDARD DEVIATION	COEFFICIENT OF VARIATION
NUMBER	31	6.67742	33.22581	5.76418	86.32351
TIME	31	13.91290	139.39983	11.80677	84.86202

CORRELATION = .974353347879

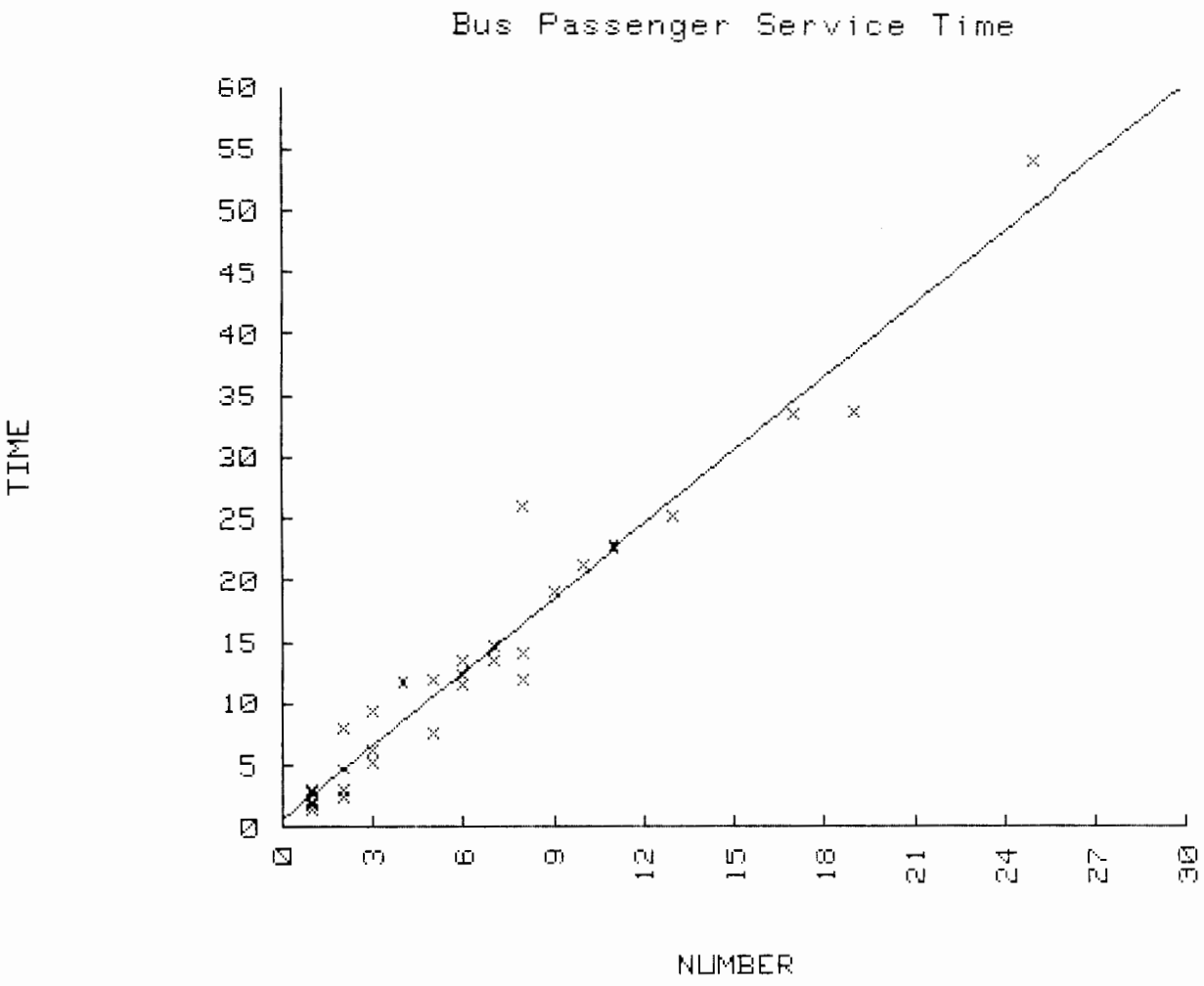
Selected degree of regression = 1
R-SQUARED = .949364446536
STANDARD ERROR OF ESTIMATE = 2.70221890456

AOV

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F-VALUE
TOTAL	30	4181.99484		
REGRESSION	1	3970.23722	3970.23722	543.72
X^1	1	3970.23722	3970.23722	543.72
RESIDUAL	29	211.75762	7.30199	

VARIABLE	REGRESSION COEFFICIENTS STD. FORMAT	E-FORMAT	STANDARD ERROR REG. COEFFICIENT	T-VALUE
'CONSTANT'	.58633	.586330096900E+00	.74979	.78
X^1	1.99577	.199576699031E+01	.08559	23.32

	COEFFICIENT	95 % CONFIDENCE INTERVAL LOWER LIMIT	UPPER LIMIT
'CONSTANT'	.58633	-.94752	2.12018
X^1	1.99577	1.82068	2.17086


Bus Passenger Service Time
TIME
NUMBER
0
5
10
15
20
25
30
35
40
45
50
55
60
0
3
6
9
12
15
18
21
24
27
30

Residual Analysis

I. OBJECT PROGRAM:

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. A final column shows which residuals are significantly large. In this column, two asterisks are printed if the standardized residual is between two and three standard deviations away from zero. Similarly, three asterisks are printed for standardized residuals between three and four standard deviations away from zero, and four asterisks are printed for standardized residuals four or more standard deviations away from zero.

The residual plot allows the user to plot the standardized residuals versus time or versus any of variables in the model. The plot may be carried out on the HP 9872A plotter.

II. SPECIAL CONSIDERATIONS:

A. 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.

B. The program is currently capable of plotting residuals on the HP 9872A plotter. If some other plotter is to be used, say "INCPLOT", then change line 340 of file "RESID" to PLOTTER IS Psc, Bn, "INCPLOT", where Psc is the plotter select code and Bn is the bus number or device address.

C. If your machine is not equipped with the 98337A Plotter ROM when this program is loaded, a series of error messages will appear on the screen and the program will halt. To resume, type CONT 1Ø and press EXECUTE. The program will then run normally, provided no plotting is requested.

III. METHODS AND FORMULAE:

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

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

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

1. Predicted Y: $\hat{Y}(n) = \hat{\beta}_0 + \hat{\beta}_1 D(X_1,n) + \ldots + \hat{\beta}_p D(X_p,n)$
2. Residual: $R(n) = D(Y,I) - \hat{Y}(n)$
3. Standard error of residuals: Ser = residual mean square

The residual mean square is calculated in the regression routine.

4. Standardized residual: $Sr(n) = \frac{R(n)}{Ser}$

IV. USER INSTRUCTIONS:

1. When "Print out residuals?" is displayed:

   a. Press: NO if a printout of the residuals is not desired.

   b. Go to 3.

   or

   a. Press: YES if a printout of the residuals is desired.

2. At this point the residuals, standardized residuals, etc., are printed.
3. When "Residual plots?" is displayed:

   a. Press: NO if no plots are desired, or if a plotting device is not available.

   b. Go to 13.

   or

   a. Press: YES if one or more residual plots are desired.

4. When "Plotter select code, Bus # = ?" is displayed:

   a. Enter the select code of the plotting device followed by the bus number of the device (separated by a comma).

   b. Press: CONT.

5. When "Independent variable (enter 0 for plot versus sequence #) = ?" is displayed:

   a. Enter the number of the variable that will serve as the independent (X) variable in the residual plot.

   b. Press: CONT.

   c. Go to 6.

   or

   a. Enter 0 to plot the residuals versus sequence or observation number, e.g., versus time.

   b. Press: CONT.

6. When the maximum and minimum values of the independent variable are printed and "X-min = ?" is displayed:

   a. Refer to the printed information and enter the minimum value of the independent variable to be included in the plot area. Note: The X-axis will begin at this point.

b. Press: CONT.

7. When "X-max = ?" is displayed:

   a. Enter the maximum value of the independent variable to be included in the plot area. Note: The X-axis will end at this point. Clearly, Xmin < Xmax is a requirement.

   b. Press: CONT.

8. When "Distance between X-ticks = ?" is displayed:

   a. Enter the spacing desired between tick marks on X-axis. Note: The first tick mark will coincide with X-min. Every tick mark will be labelled. Hence, an uncluttered plot should contain no more than 10 tick marks.

   b. Press: CONT.

9. When "# of decimals for labelling X-axis (<=7) = ?" is displayed:

   a. Enter the number of decimal places to appear in the labelling of tick marks along the X-axis. The number must be between 0 and 7 inclusive.

   b. Press: CONT.

10. When the input information is printed and "Is above information correct?" is displayed:

    a. Press: NO if changes are desired.

    b. Go to 3.

    or

    a. Press: YES if no changes are necessary.

11. The residual plot is now drawn.

12. When "Another residual plot?" is displayed:

    a. Press: YES if another plot is desired.

    b. Go to 5.

    or

    a. Press: NO if no more plots are desired.

13. When "SELECT ANY KEY" is displayed:

    a. Press: Any of the keys defined on the "REGRESSION ANALYSIS METHODS" overlay.

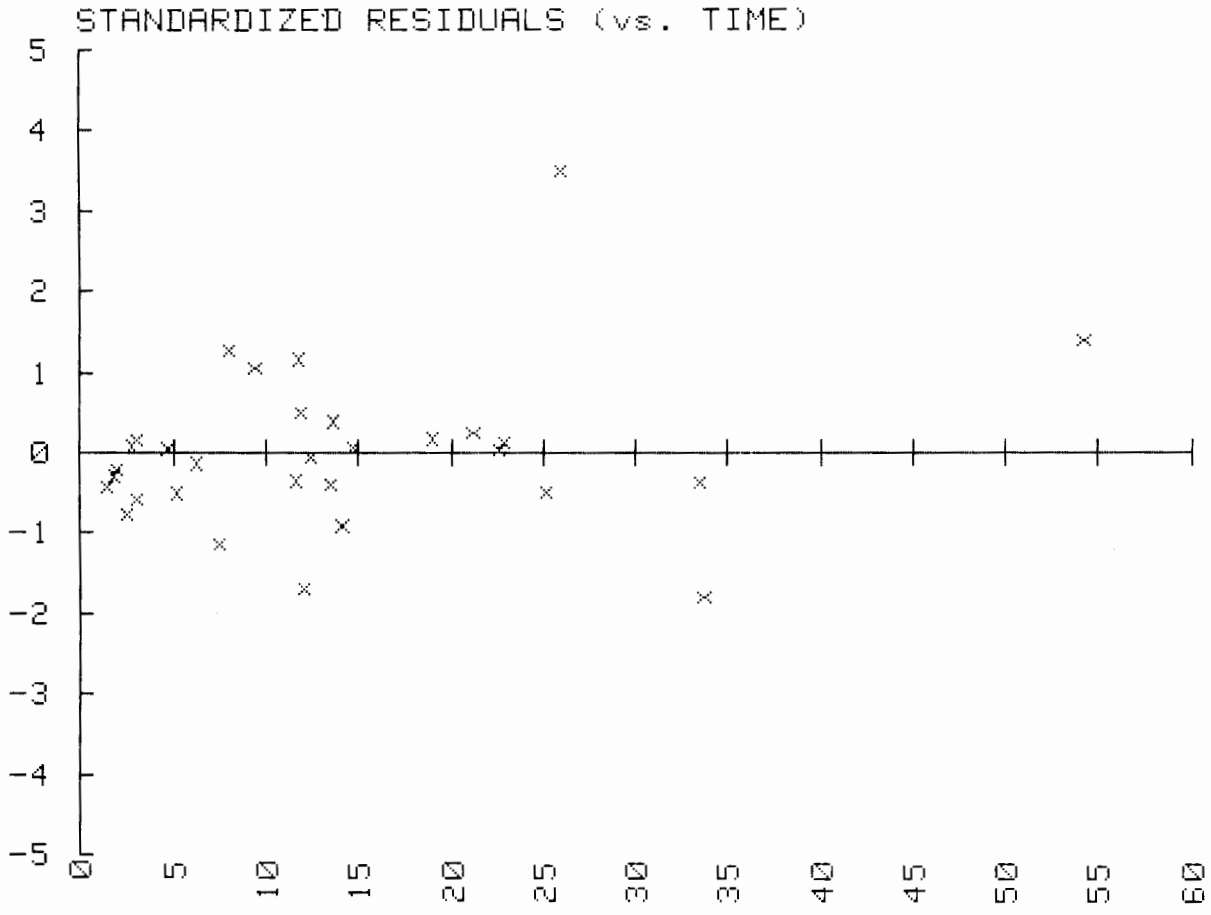
    b. Go to step 1 of the User Instructions for the selected key.

V. EXAMPLE:

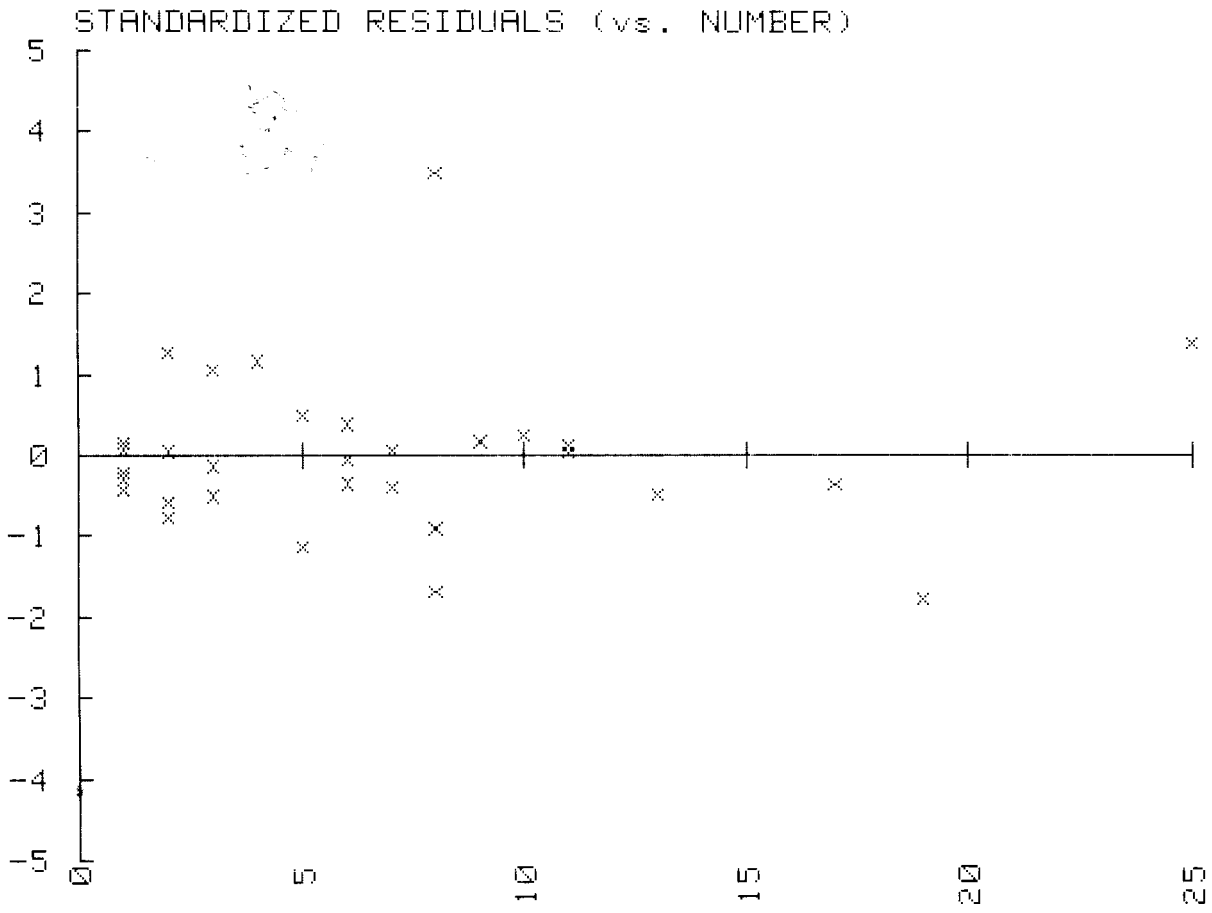
Residual analysis was performed on the data used in POLYNOMIAL REGRESSION. First is a table of the residuals, followed by residual plots. Notice that the independent variable for residual plots can be the observation number or either of the variables in the regression problem.

OBS#	OBSERVED Y	PREDICTED Y	RESIDUAL	STAND.RES.	SIGNIF.
1	1.40000	2.58210	-1.18210	-.43745	
2	2.80000	2.58210	.21790	.08064	
3	3.00000	2.58210	.41790	.15465	
4	1.80000	2.58210	-.78210	-.28943	
5	2.00000	2.58210	-.58210	-.21541	
6	4.70000	4.57786	.12214	.04520	
7	8.00000	4.57786	3.42214	1.26642	
8	3.00000	4.57786	-1.57786	-.58391	
9	2.50000	4.57786	-2.07786	-.76895	
10	5.20000	6.57363	-1.37363	-.50833	
11	6.20000	6.57363	-.37363	-.13827	
12	9.40000	6.57363	2.82637	1.04594	
13	11.70000	8.56940	3.13060	1.15853	
14	7.50000	10.56517	-3.06517	-1.13431	
15	11.90000	10.56517	1.33483	.49398	
16	13.60000	12.56093	1.03907	.38452	
17	12.40000	12.56093	-.16093	-.05956	
18	11.60000	12.56093	-.96093	-.35561	
19	14.70000	14.55670	.14330	.05303	
20	13.50000	14.55670	-1.05670	-.39105	
21	12.00000	16.55247	-4.55247	-1.68471	
22	14.10000	16.55247	-2.45247	-.90757	
23	26.00000	16.55247	9.44753	3.49621	***
24	19.00000	18.54823	.45177	.16718	
25	21.20000	20.54400	.65600	.24276	
26	22.90000	22.53977	.36023	.13331	
27	22.60000	22.53977	.06023	.02229	
28	25.20000	26.53130	-1.33130	-.49267	
29	33.50000	34.51437	-1.01437	-.37538	
30	33.70000	38.50590	-4.80590	-1.77850	
31	54.20000	50.48050	3.71950	1.37646	

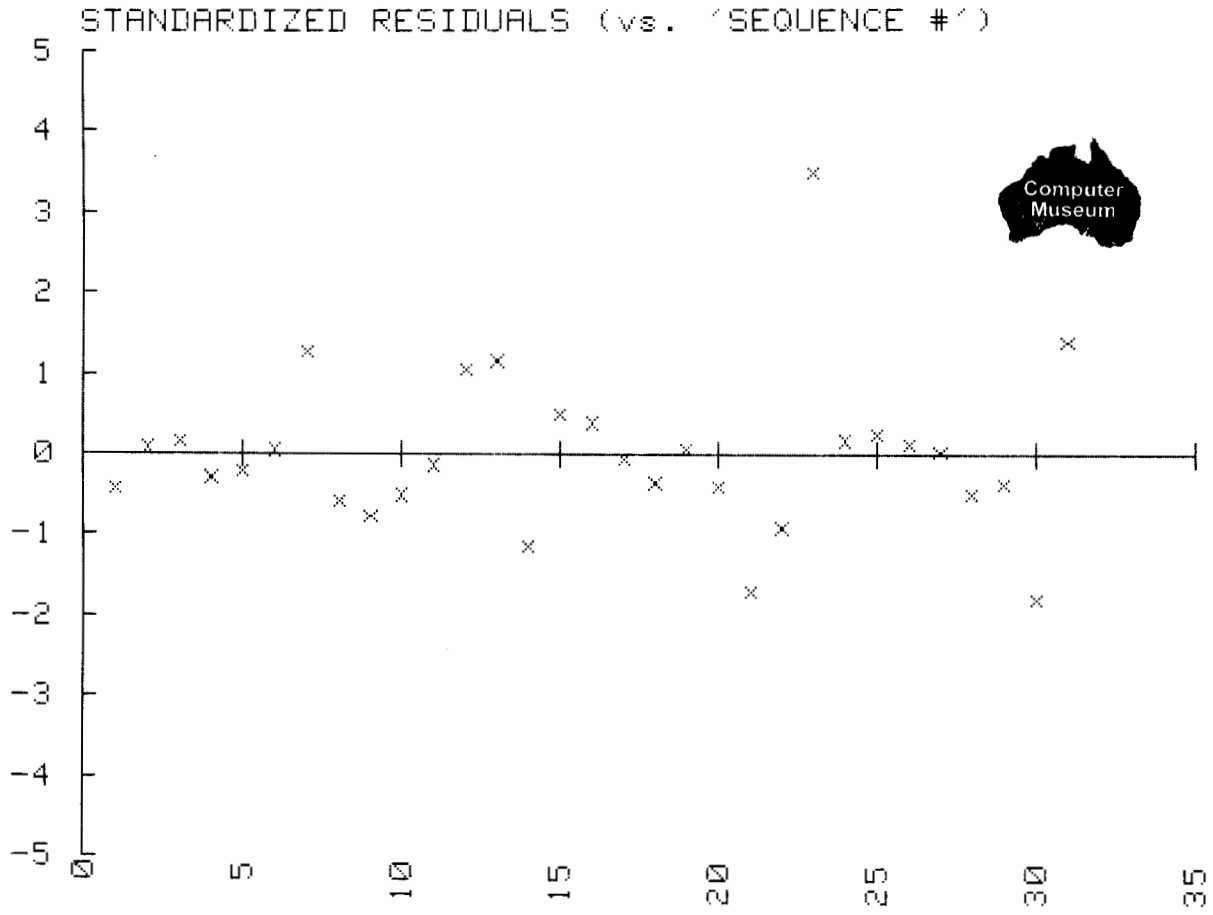
Bus Passenger Service Time



Bus Passenger Service Time



Bus Passenger Service Time



Notes