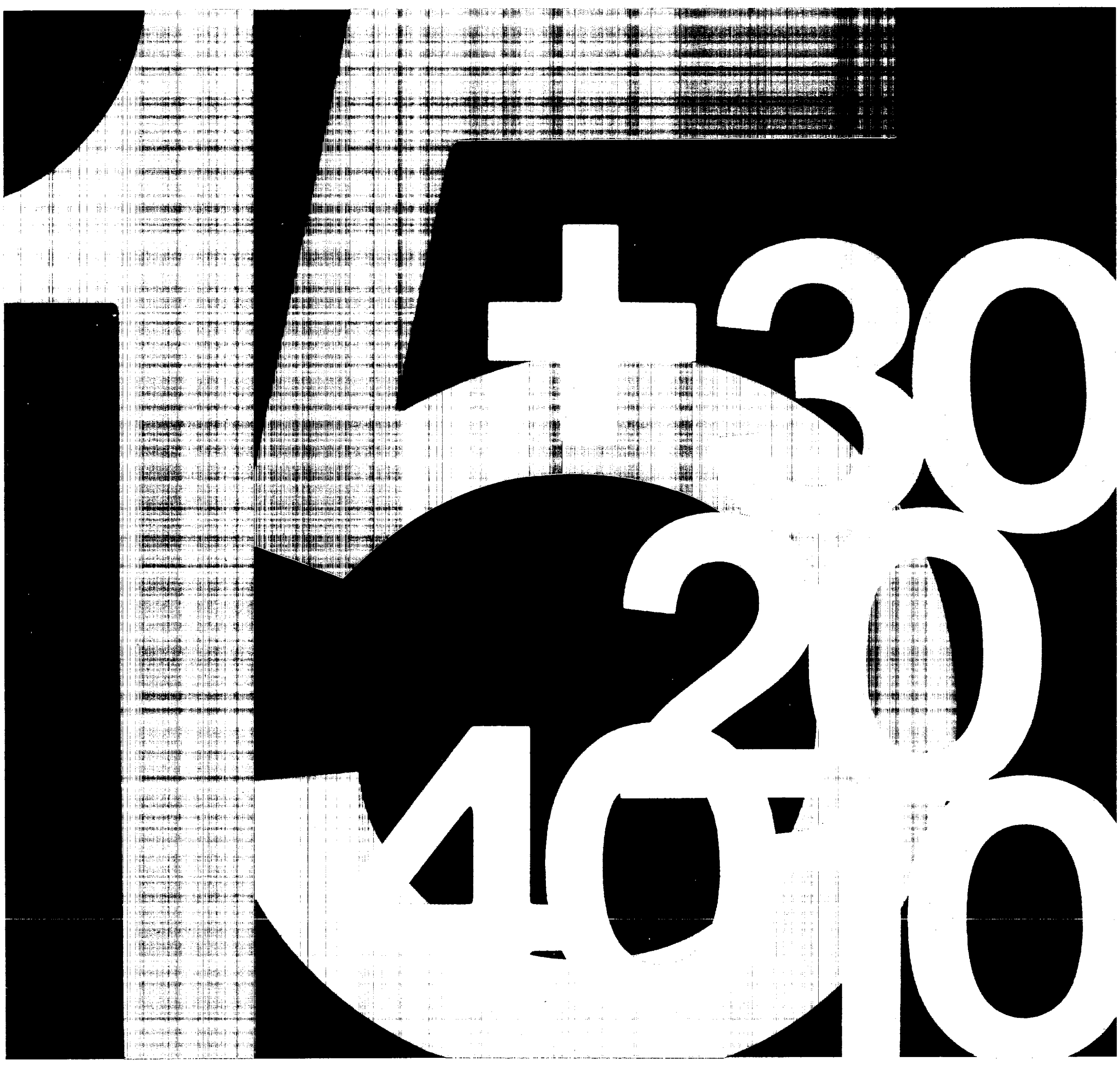
Digital Equipment Corporation
Maynard, Massachusetts

digital

## STATPAC User's Guide

# PDP-15 Systems



DEC-15-UFZA-D

# STATPAC USER'S GUIDE

## PDP-15/20
## PDP-15/30
## PDP-15/40
## PDP-9 ADVANCED SOFTWARE SYSTEMS

For additional copies order No. DEC-15-UFZA-D from Program Library,

Digital Equipment Corporation, Maynard, Mass. Price $3.00

DIGITAL EQUIPMENT CORPORATION □ MAYNARD, MASSACHUSETTS

1st Printing June 1969

# CONTENTS


Page

CHAPTER 1
INTRODUCTION

| | | |
|---|---|---|
| 1.1 | Control Module | 1-1 |
| 1.2 | Input Module | 1-3 |
| 1.3 | SMMRY Module | 1-4 |
| 1.3.1 | SMMRY Statistics | 1-4 |
| 1.3.2 | SMMRY Options | 1-6 |
| 1.4 | STPRG and MLTRG Modules | 1-10 |
| 1.4.1 | Regression Analysis | 1-11 |
| 1.4.2 | Regression Options | 1-13 |
| 1.4.3 | Regression Plots | 1-14 |

CHAPTER 2
MODULE OPERATING PROCEDURES

| | | |
|---|---|---|
| 2.1 | Control Module | 2-1 |
| 2.1.1 | Command Dialogue | 2-1 |
| 2.1.2 | Error Messages | 2-1 |
| 2.2 | Input Module | 2-2 |
| 2.2.1 | Command Dialogue | 2-2 |
| 2.2.2 | Error Messages | 2-3 |
| 2.3 | Descriptive Statistics Module | 2-3 |
| 2.3.1 | Command Dialogue | 2-3 |
| 2.3.2 | Error Messages | 2-5 |
| 2.4 | Regression Analysis Modules | 2-5 |
| 2.4.1 | Command Dialogue | 2-5 |
| 2.4.2 | Error Messages | 2-7 |

CHAPTER 3
IMPLEMENTING AND AUGMENTING STATPAC

| | | |
|---|---|---|
| 3.1 | Building an Executable File | 3-2 |
| 3.2 | Adding Processing Modules to STATPAC | 3-7 |

CHAPTER 4
SAMPLE OPERATION

| | | |
|---|---|---|
| 4.1 | Input Example | 4-2 |
| 4.2 | SMMRY Example | 4-4 |
| 4.3 | STPRG Example | 4-9 |
| 4.4 | MLTRG Example | 4-18 |

APPENDIX A
DESCRIPTIVE STATISTICS ALGORITHMS

CONTENTS (Cont)

Page

APPENDIX B
REGRESSION ANALYSIS ALGORITHMS

ILLUSTRATIONS

| | | |
|---|---|---|
| 1-1 | STATPAC Logic Modules, Flow Diagram | 1-2 |

TABLES

| | | |
|---|---|---|
| 1-1 | STATPAC Symbol Definitions | 1-6 |

## PREFACE

The PDP-15/9 statistics package (STATPAC) is a FORTRAN-coded program used to perform statistical analysis on user-supplied data. STATPAC runs under control of the PDP-15 Advanced and Background/Foreground Monitor Systems and the PDP-9 Keyboard Monitor System, and requires some form of auxiliary bulk storage, such as DECtape or disk. This guide is intended to set forth operating procedures for the user, and does not contain detailed descriptions of the internal operations of the package. The Guide is organized as follows:

| | |
|---|---|
| Chapter 1 | Introduction to STATPAC |
| Chapter 2 | Module Operating Procedures |
| Chapter 3 | Implementing and Augmenting STATPAC |
| Chapter 4 | Sample Operation |

Chapter 1 provides general descriptions of each of the modules in the package. Chapter 2 details the command dialogue and possible error messages. Chapter 3 contains information related to building an executable file and augmenting basic systems either through addition of user software modules or through expanding the hardware configuration. Chapter 4 contains sample dialogue and output for all STATPAC modules. The Appendix contains detailed algorithms for computations performed within the package which will be of interest to the more demanding reader. Finally, a bibliography of statistical texts and applicable manuals is included for convenient reference.

No attempt is made within the Guide to educate the novice statistician. It is assumed that the user has a good background in statistics and can use the package as a tool to achieve the desired results.

## CHAPTER 1
## INTRODUCTION

STATPAC is a FORTRAN-coded program used to perform statistical analysis on user-supplied data. The package is designed to run under control of PDP-15/9 monitor systems in a hardware configuration that includes 8K of core memory, a console Teletype, a high-speed paper tape reader and punch, and two bulk storage units. Due to the limitations of 8K core memory, the package is divided into logical modules, each of which consists of one or more core loads (i.e., chains or overlays). The modules (Figure 1-1) reside on a bulk storage device (logical –4) and include CONTROL, INPUT, SMMRY, STPRG, and MLTRG. Basic operation of the package requires that the user supply data to the INPUT module which prepares standardized binary data files. The user then can, depending upon his next task, select any one of the modules for operation. Briefly, the SMMRY (Summary) module provides the user with a set of descriptive statistics based upon his input files. The descriptive statistics include mean, variance, standard deviation, standard error of the mean, skewness, kurtosis, maximum, minimum, range, and a correlation matrix. The other two modules (STPRG and MLTRG) can be selected to perform stepwise linear regression and multiple linear regression, respectively.

The following paragraphs provide a general description of each of the modules. User dialogues are presented in Chapter 2 and detailed algorithms for the internal computations are given in Appendix A.

### 1.1 CONTROL MODULE

The CONTROL module acts as an executive routine, performing miscellaneous control functions while providing a means for communications between modules. Initially, the CONTROL module is loaded into core (see Chapter 3). Once loaded, it types the message

*PROG

The user must respond by typing one of the following names:

INPUT
SMMRY
STPRG
MLTRG
EXIT

By responding with EXIT, the user terminates all processing by STATPAC and control is returned to the monitor. Responding with one of the module names causes the corresponding module (or a portion of it) to be loaded from the STATPAC tape (logical –4), overlaying the CONTROL module. Control is transferred to the module that has been loaded, and it requests and obtains the remaining control parameters required to perform an analysis by conducting a dialogue

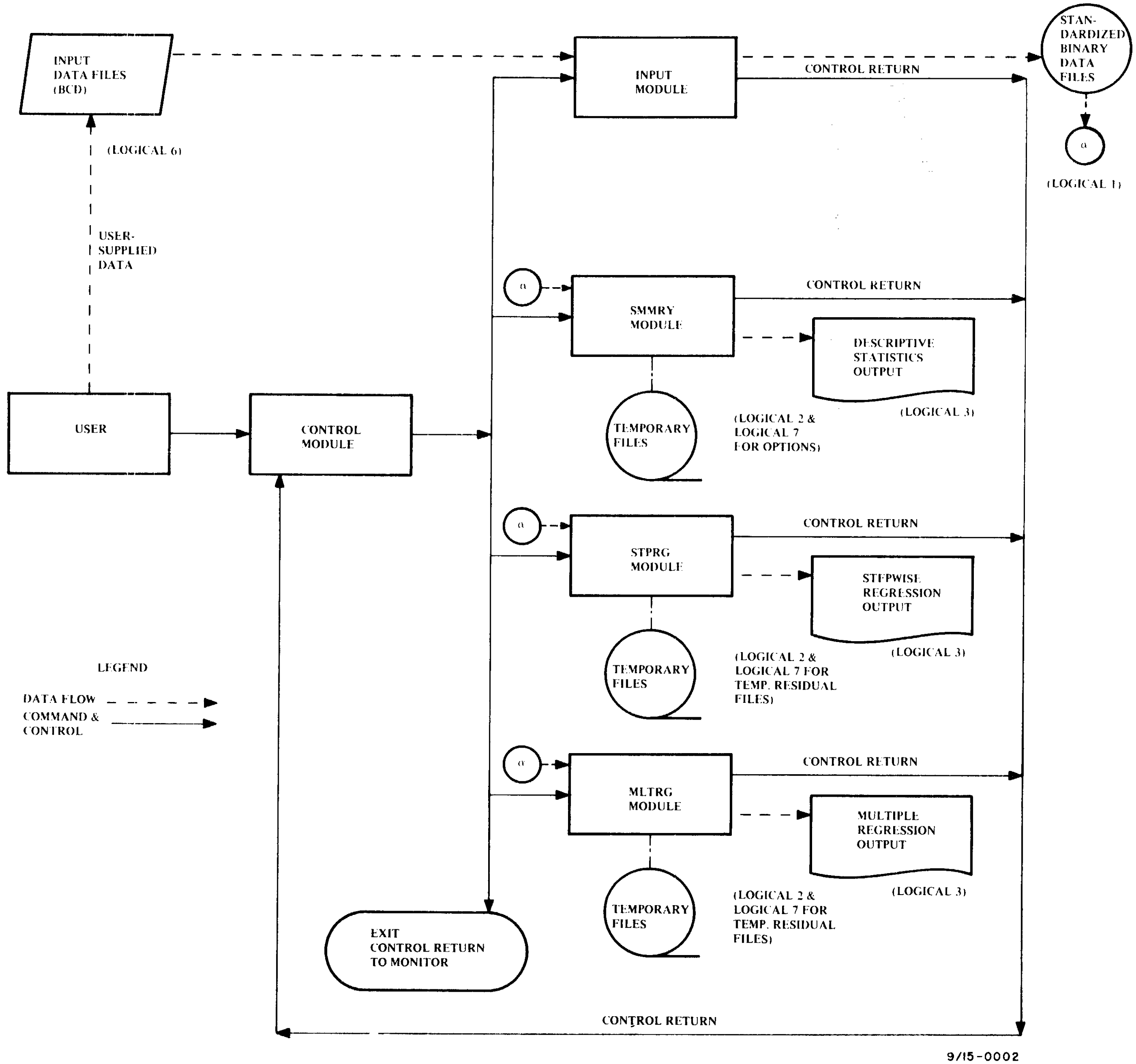



Figure 1-1 STATPAC Logic Modules, Flow Diagram

with the user (see Chapter 2). When the selected module has completed its task, it requests that the user supply the name of the next module to be loaded. If the user requests the module that is already in core, the module again requests the required control parameters. This continues until the user requests a different module, at which time the CONTROL module is loaded into core and, in turn, loads the selected module.

## 1.2 INPUT MODULE

The INPUT module performs two basic tasks:

a. Conversion of user-supplied BCD data to binary.

b. Preparation and storage of the standardized binary data files on a file structured bulk storage device.

The user's input data consists of observations, with each observation consisting of a number of variables. For example, each person living in a town could be considered an observation consisting of the variables age, weight, height, etc. One can think of a data file, then, as a rectangular array or matrix of the form:

| | Variable (1) | Variable (2) | | Variable (L) |
|---|---|---|---|---|
| observation 1 | $X_{1,1}$ | $X_{1,2}$ | . . . | $X_{1,L}$ |
| observation 2 | $X_{2,1}$ | $X_{2,2}$ | . . . | $X_{2,L}$ |
| . | . | . | | . |
| . | . | . | | . |
| . | . | . | | . |
| observation N | $X_{N,1}$ | $X_{N,2}$ | . . . | $X_{N,L}$ |

Note: $X_{ij}$ $i = i^{th}$ observation
$j = i^{th}$ variable

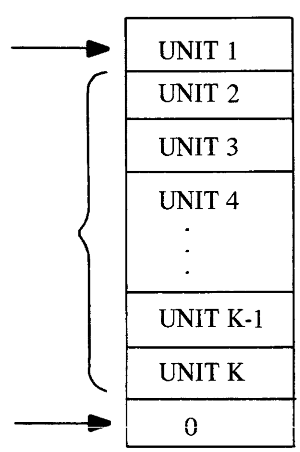
This data file consists of N observations, with each observation consisting of L variables. One can think of the observations as rows and the variables as columns. In using the statistics program, the user will frequently be asked: "What are the variables?", to which he must respond by enumerating the column ordinals of the variables he wants analyzed. In brief, given the subscripts 1, 2, ..., L, where each subscript is associated with one variable, the program is interested in how many and which variables were chosen.

The standardized binary data files are organized on the tape written by the INPUT module as follows:

Unit 1 has one record which contains L, the number of variables in each observation and the names of each variable.

Unit 2 through K have one record which specifies the number of observations within the unit, N; and N records which contain the values of the variables for each observation. All units, except possibly the $K^{th}$, have the same number of observations. Unit K may have less than N observations.

The last unit (K+1) has one record which contains 0 to signal the end of data.

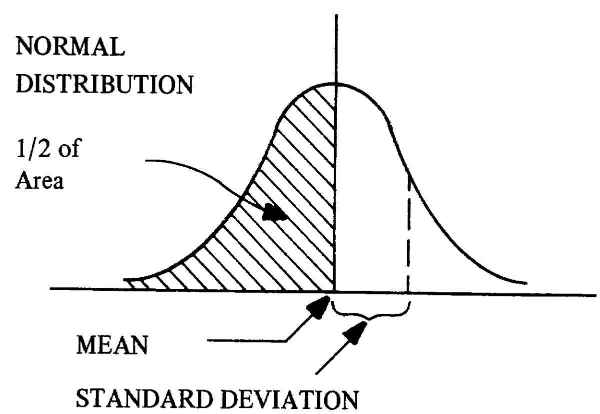


### 1.3 SMMRY MODULE

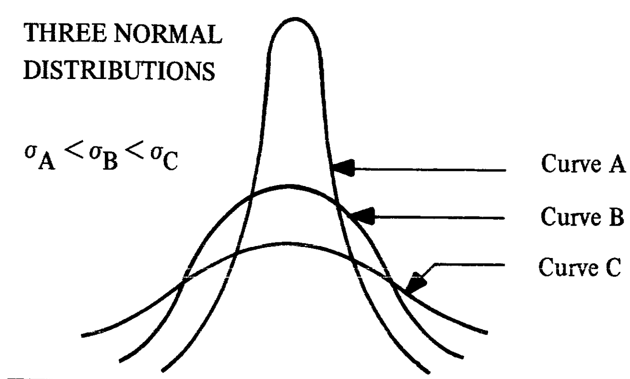
The SMMRY module reads data files designated by the user, analyzes the data, and outputs the following statistics for each variable which was selected by the user for analysis:

- Mean
- Variance
- Standard Deviation
- Standard Error of the Mean
- Skewness
- Kurtosis
- Maximum
- Minimum
- Range
- Correlation Matrix

#### 1.3.1 SMMRY Statistics

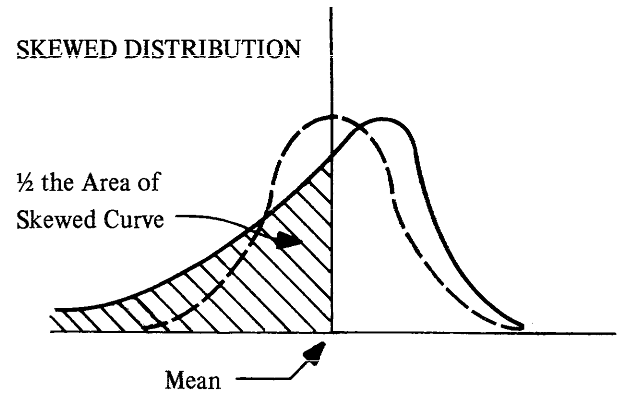



The previous statistics are estimates of the corresponding parameters of the populations from which the samples were drawn. The *mean* serves to specify the "center" of the data, while the *standard deviation* is a measure of the scatter, or dispersion, of the data from the center. The *variance* is the square of the standard deviation.

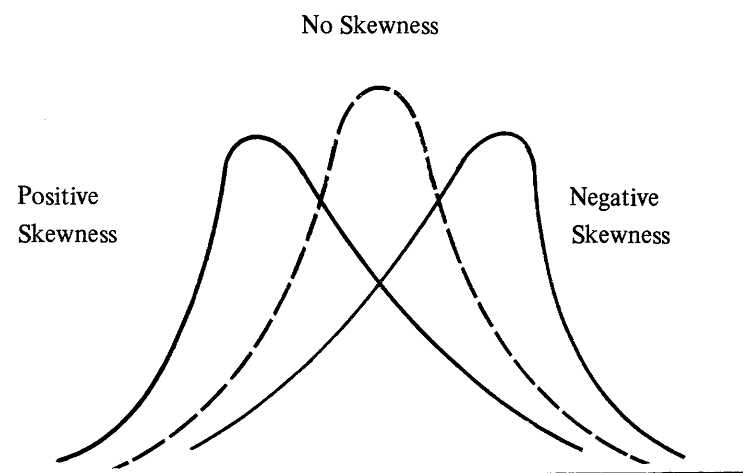



The figure at the left shows the changes in the shape of a curve effected by varying the standard deviation, $\sigma$.

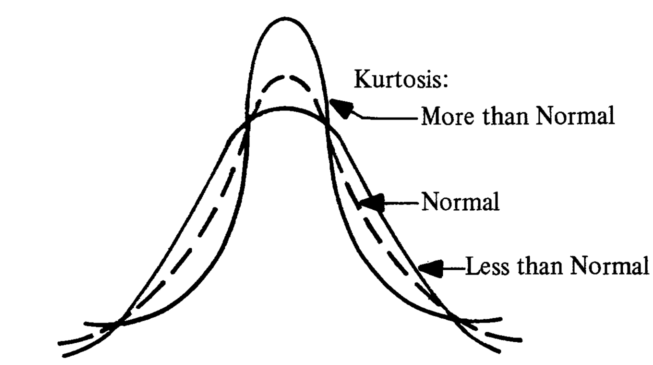
The *skewness* is used to measure the symmetry of a distribution about the mean. Since the normal distribution is symmetric, skewness is used to test whether a distribution is not normal.




The sign of the skewness statistic indicates the direction of the skew as seen at the right.




*Kurtosis* measures the relative concentration of values of a sample; i.e., about the "center", the "tails", and the "shoulders" of the distribution. The illustration at the right compares curves with different degrees of kurtosis.




The *maximum* is the highest observed value and the *minimum* is the lowest observed value. Their difference is the *range.*

The *correlation matrix* indicates whether any pairs of variables in a file are highly correlated. Independent variables which are too highly correlated should not be used in the same regression analysis problem.

The descriptive statistics module (SMMRY) will also enable the user to perform hypotheses testing.

Table 1-1
STATPAC Symbol Definitions

| Symbol | Definition |
|---|---|
| $L$ | the number of variables in a user data file |
| $j$ | the ordinal of a particular variable in a file $(1 \leqslant j \leqslant L)$ |
| $K$ | the number of files in an analysis |
| $i$ | the ordinal of a particular file $(1 \leqslant i \leqslant K)$ |
| $N_i$ | the number of observations in the $i^{th}$ file |
| $m$ | the observation ordinal $(1 \leqslant m \leqslant N_i)$ |
| $x_{jim}$ | the $j^{th}$ variable in the $m^{th}$ observation of the $i^{th}$ file |
| $\sigma_{ji}$ | the standard deviation of the $j^{th}$ variable in the $i^{th}$ file |
| $\sigma_{ji}^2$ | the variance of the $j^{th}$ variable in the $i^{th}$ file |
| $\overline{x}_{ji}$ | *the calculated mean of the $j^{th}$ variable in the $i^{th}$ file |
| $\mu_{ji}$ | *the actual mean of the $j^{th}$ variable in the $i^{th}$ file |
| $\dot{\mu}_j$ | the user supplied test mean for the $j^{th}$ variable |
| $\sigma_j^2$ | the user supplied test variance for the $j^{th}$ variable |
| $y_i$ | the observed value of the dependent variable |
| $\hat{y}_i$ | the predicted value of the dependent variable determined using the regression model |
| $\overline{y}$ | the calculated mean of the observed dependent variable |
| $b_i$ | the coefficient of the $i^{th}$ variable in a regression model |
| $b_o$ | the constant term of a regression model. |

*These symbols are used interchangeably in descriptions of hypotheses.

### 1.3.2 SMMRY Options

The SMMRY module of STATPAC includes six hypothesis test options. Each option permits the user to test one or more actual hypotheses. The user requests a specific option in response to the initial dialogue as described in Chapter 2.

*SMMRY Option 1* allows the user to test hypotheses which relate the calculated means for variables to user-supplied test means. These hypothesis tests may be performed upon one or more data files (up to 10 files). The statistic calculated for each file, however, is independent of that calculated for any other data file.

STATPAC calculates the following t-statistic when option 1 is requested:

$$t_{ji} = \left\{(\bar{x}_{ji} - \mu_j)\sqrt{N_i}\right\} / \sigma_{ji}$$

Under the assumption that the sample came from a normal population, the user can use the statistic $t_{ji}$ to test hypotheses which relate the calculated mean of a variable ($\mu_{ji}$) to the user-supplied test mean for that variable ($\mu_j$), as summarized below.

| Hypothesis | Acceptance Criteria | Alternative Hypothesis |
|---|---|---|
| $\mu_{ji} = \mu_j$ | $-t_{(1-\alpha/2)(N_i-1)} < t_{ji} < t_{(1-\alpha/2)(N_i-1)}$ | $\mu_{ji} \neq \mu_j$ |
| $\mu_{ji} \leqslant \mu_j$ | $t_{ji} < t_{(1-\alpha)(N_i-1)}$ | $\mu_{ji} > \mu_j$ |
| $\mu_{ji} \geqslant \mu_j$ | $t_{ji} > -t_{(1-\alpha)(N_i-1)}$ | $\mu_{ji} < \mu_j$ |

When the acceptance criteria is not satisfied at the user-specified significance level ($\alpha$), the alternate hypothesis is accepted. The t-values are obtained from a statistical table using the values of $N_i - 1$ (the degrees of freedom) and the expression in $\alpha$ as the parameters for selecting the t-value from the table.

*SMMRY Option 2* allows the user to test hypotheses which relate the variance to a user-supplied test variance. These hypothesis tests may be performed upon one or more data files (up to 10 files). The statistic calculated for each file, however, is independent of that calculated for any other data file.

STATPAC calculates the following chi-square statistic when option 2 is requested:

$$\chi^2_{ji} = \left\{ \sum_{m=1}^{N_i} (X_{jim} - \bar{X}_{ji})^2 \right\} / \sigma^2_j$$

Assuming a normal population, the $\chi^2_{ji}$ statistic may then be used by the statistician to test hypotheses which relate the calculated variance ($\sigma^2_{ji}$) with the user-supplied variance ($\sigma^2_j$), as summarized below.

| Hypothesis | Acceptance Criteria | Alternative Hypothesis |
|---|---|---|
| $\sigma_{ji}^2 = \sigma_j^2$ | $\chi^2_{(\alpha/2)(N_i-1)} < \chi^2_{ji} < \chi^2_{(1-\alpha/2)(N_i-1)}$ | $\sigma_{ji}^2 \neq \sigma_j^2$ |
| $\sigma_{ji}^2 \leqslant \sigma_j^2$ | $\chi^2_{ji} < \chi^2_{(1-\alpha)(N_i-1)}$ | $\sigma_{ji}^2 > \sigma_j^2$ |
| $\sigma_{ji}^2 \geqslant \sigma_j^2$ | $\chi^2_{ji} > \chi^2_{(\alpha)(N_i-1)}$ | $\sigma_{ji}^2 < \sigma_j^2$ |

When the acceptance criteria is not satisfied at the user-specified significance level ($\alpha$), the alternate hypothesis is accepted. The chi-square values are obtained from a statistical table using the values of $N_i - 1$ (degrees of freedom) and the expression in $\alpha$ as the parameters for selecting the chi-square value from the table.

*SMMRY Option 3* allows the user to test hypotheses which relate the mean of a variable in one file to the mean of the corresponding variable (i.e., same ordinal) in another file. Thus, at least 2 files must be included in the analysis, but not more than 10 files may be analyzed.

STATPAC calculates the following t-statistics when option 3 is requested by the user:

$$t_{jrs} = (\overline{X}_{jr} - \overline{X}_{js}) / (S^2/N_r + S^2/N_s)^{1/2}$$

where

$$S^2 = \left\{ (N_r - 1)\,\sigma_{jr}^2 + (N_s - 1)\,\sigma_{js}^2 \right\} / \left\{ N_r + N_s - 2 \right\}$$

In the calculation, r and s vary from 1,2,...,K for each value of j. Results are provided by STATPAC for each value of j (i.e., for each variable being analyzed). Thus, for every value of j, there is a K x K matrix generated (where K is the number of files in the analysis).

Under the assumption that the samples came from normal populations with $\sigma_{jr}^2 = \sigma_{js}^2$, the user can perform the following hypothesis tests for the variables of each possible pair of files in the analysis using the statistic $t_{jrs}$. Each hypothesis test relates variables with the same ordinal, but contained in different files.

| Hypothesis | Acceptance Criteria | Alternate Hypothesis |
|---|---|---|
| $\mu_{jr} = \mu_{js}$ | $-t_{(1-\alpha/2)(N_r + N_s - 2)} < t_{jrs} < t_{(1-\alpha/2)(N_r + N_s - 2)}$ | $\mu_{jr} \neq \mu_{js}$ |
| $\mu_{jr} \leqslant \mu_{js}$ | $t_{jrs} < t_{(1-\alpha)(N_r + N_s - 2)}$ | $\mu_{jr} > \mu_{js}$ |
| $\mu_{jr} \geqslant \mu_{js}$ | $t_{jrs} > -t_{(1-\alpha)(N_r + N_s - 2)}$ | $\mu_{jr} < \mu_{js}$ |

When the acceptance condition is not satisfied at the user specified significance level ($\alpha$), the alternate hypothesis is accepted. The t-values are obtained from statistical tables using the values of $N_r + N_s - 2$ (the sum of the separate degrees of freedom $N_r - 1$ and $N_s - 1$), and the expression in $\alpha$ as the parameters for selecting the value from the table.

*SMMRY Option 4* allows the user to test hypotheses which relate the variance of a variable in one file with the variance of the corresponding variable (i.e., same ordinal) in a second file. Analysis of at least two files must be performed for this option to be executed, but no more than 10 files may be included in the analysis.

When option 4 is requested, STATPAC computes the following F-statistic:

$$F_{jrs} = \sigma_{jr}^2 / \sigma_{js}^2$$

where r and s vary from 1,2,...,K for each value of j. These F-values are output by STATPAC for each variable in the analysis (j, where $1 \leqslant j \leqslant L$) and for all combinations of values for r and s (r, s = 1,2,...,K). Thus, for every value of j, there is a K x K matrix generated (where K = the number of files in the analysis).

Under the assumption that the samples were drawn from normal populations, the user can perform the following hypothesis tests for a fixed variable and for each pair of files in the analysis, using the computed statistic $F_{jrs}$. Each hypothesis test relates the variance of a variable with the variance of a variable having the same ordinal, but contained in a different file.

| Hypothesis | Acceptance Criteria | Alternate Hypothesis |
| --- | --- | --- |
| $\sigma_{jr}^2 = \sigma_{js}^2$ | $F_{(\alpha/2)(N_r - 1, N_s - 1)} < F_{jrs} < F_{(1-\alpha/2)(N_r - 1, N_s - 1)}$ | $\sigma_{jr}^2 \neq \sigma_{js}^2$ |
| $\sigma_{jr}^2 \leqslant \sigma_{js}^2$ | $F_{jrs} < F_{(1-\alpha)(N_r - 1, N_s - 1)}$ | $\sigma_{jr}^2 > \sigma_{js}^2$ |
| $\sigma_{jr}^2 \geqslant \sigma_{js}^2$ | $F_{jrs} > F_{(\alpha)(N_r - 1, N_s - 1)}$ | $\sigma_{jr}^2 < \sigma_{js}^2$ |

When the acceptance condition is not satisfied at the user-specified significance level ($\alpha$), the alternate hypothesis is accepted. The F-values are obtained from statistical tables using the values of the degrees of freedom for each file, $N_r - 1$ and $N_s - 1$, and the expression in $\alpha$ as the parameters for selecting the F-value from the table.

*SMMRY Option 5* allows the user to test the hypothesis that, for a particular variable, the means of that variable in all files of the analysis are equal at the user-specified significance level. Analysis of at least 2 files must be performed for this option to be executed, but no more than 10 files may be included. Option 5 is a generalization of option 3.

When option 5 is requested, STATPAC computes the following F-statistic for each variable j analyzed:

$$F_j = \frac{\left( \sum_{i=1}^{K} N_i (\overline{X}_{ji} - \overline{\overline{X}}_j)^2 \right) / (K - 1)}{\left( \sum_{i=1}^{K} \sum_{m=1}^{N_i} (X_{jim} - \overline{X}_{ji})^2 \right) / \sum_{i=1}^{K} (N_i - 1)}$$

where

$$\overline{\overline{X}}_j = \left( \sum_{i=1}^{K} \overline{X}_{ji} \right) / K$$

Under the assumption that all samples were drawn from normal populations with equal variance (i.e., $\sigma_{jr}^2 = \sigma_{js}^2$ for all r, s = 1,2,...,K) the user may test the following hypothesis:

| Hypothesis | Acceptance Criteria | Alternate Hypothesis |
| --- | --- | --- |
| $\mu_{j1} = \mu_{j2} = ... = \mu_{jK}$ | $F_j < F_{(1-\alpha)(V1, V2)}$<br>where<br>$V1 = K - 1$<br>$V2 = \sum_{i=1}^{K} (N_i - 1)$ | $\mu_{jr} \neq \mu_{js}$<br>for some r and s |

When the acceptance condition is not satisfied at the user-specified significance level ($\alpha$), the alternate hypothesis is accepted. The F-values are obtained from statistical tables using the values of $1 - \alpha$, V1, and V2 as the parameters for selecting the F-value from the table.

*SMMRY Option 6* allows the user to perform Bartlett's test for equal variances for a particular variable in all data files in the analysis (normal populations are assumed). Analysis of at least 2 files must be performed for this option to be executed, but no more than 10 files may be included. Option 6 is a generalization of option 4.

When option 6 is requested by the user, STATPAC computes the following chi-square statistic and correction:

$$
{}_j\chi^2_{(K-1)} = (\log_e 10)\left\{\left[\log_{10}\left\{\left(\sum_{i=1}^{K_i}\sum_{m=1}^{N_i}(X_{jim}-\overline{X}_{ji})^2\right) / \sum_{i=1}^{K}(N_i-1)\right\}\right]\sum_{i=1}^{K}(N_i-1) - \sum_{i=1}^{K}(N_i-1)\log_{10}\left(\sum_{m=1}^{N_i}\frac{(X_{jim}-\overline{X}_{ji})^2}{N_i-1}\right)\right\}
$$

$$
C = 1 + \left[(K-1)/3\right]\left\{\sum_{i=1}^{K}\frac{1}{(N_i-1)} - \frac{1}{\sum_{i=1}^{K}(N_i-1)}\right\}
$$

$$
\text{corrected } {}_j\chi^2_{(K-1)} = {}_j\chi^2_{(K-1)} / C
$$

where C is the correction factor. Under the assumption that all samples are drawn from normal populations, the user may test the following hypothesis:

| Hypothesis | Acceptance Criteria | Alternate Hypothesis |
|---|---|---|
| $\sigma_{j1}^2 = \sigma_{j2}^2 = \dots = \sigma_{jK}^2$ | ${}_j\chi^2_{(K-1)} < \chi^2_{(1-\alpha)(K-1)}$ | $\sigma_r^2 \neq \sigma_s^2$ for some r and s |

where ${}_j\chi^2_{(K-1)}$ may be the corrected or the uncorrected value computed by STATPAC.

When the acceptance condition is not satisfied at the user specified significance level ($\alpha$), the alternate hypothesis is accepted. The chi-square values are obtained from statistical tables using the values of $1-\alpha$ and $K-1$ as the parameters for selecting the chi-square value from the table.

For a more complete description of the options, the reader is referred to Chapter 7 of *Statistics in Research* by Bernard Ostle.

### 1.4 STPRG AND MLTRG MODULES

STPRG denotes the Stepwise Linear Regression Module and MLTRG denotes the Multiple Linear Regression Module. These modules are logically separate, but still have much in common (including a similar algorithm, input/output format, and internal organization). Because of their similarities, these modules are described together, with differences clearly noted where they exist.

### 1.4.1 Regression Analysis

Assuming a set of N observations, where each observation consists of L + 1 variables, consider the first L-variables to be independent and denoted by $X_i$, i = 1,2,...,L. Consider the last variable to be a dependent variable, denoted y. To summarize, the data will appear as follows:

| | | | | | |
|---|---|---|---|---|---|
| $X_{1,1}$ | $X_{1,2}$ | $X_{1,3}$ | ... | $X_{1,L}$ | $Y_1$ |
| $X_{2,1}$ | $X_{2,2}$ | $X_{2,3}$ | ... | $X_{2,L}$ | $Y_2$ |
| . | . | . | | . | . |
| . | . | . | | . | . |
| . | . | . | | . | . |
| $X_{N,1}$ | $X_{N,2}$ | $X_{N,3}$ | ... | $X_{N,L}$ | $Y_N$ |

Now, let us assume that there exists a model (i.e., a rule, relationship, formula, or equation) which defines y as a function of the X's. This model can be expressed in the following manner:

$$Y = B_0 + B_1X_1 + B_2X_2 + ... + B_LX_L + E$$

where $B_0,B_1,...,B_L$ are the parameters of the model, and E is a true error to compensate for any discrepancies in the model. The task of regression analysis is to estimate or approximate this model, as follows:

$$Y = b_0 + b_1x_1 + b_2x_2 + ... + b_Lx_L + e$$

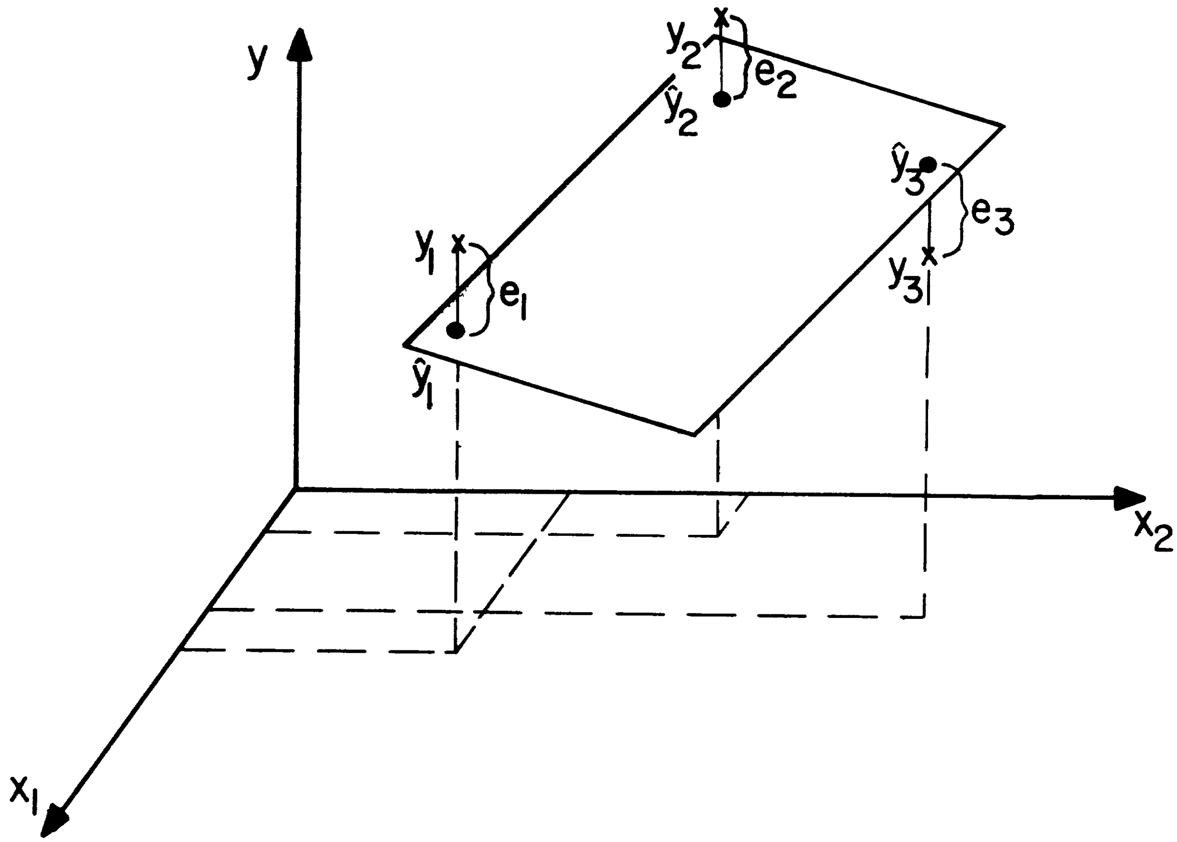
where the b's are estimates of the parameters and e is the residual of the estimating model. The estimating model is applied to all observations in the set of data as follows:

$$\hat{y}_i = b_0 + \left( \sum_{j=1}^{L} b_j x_{ij} \right) + e_i \qquad i = 1,2,..., N$$

where $y_i$ is the estimate of $y_i$ in the data, $e_i = y_i - \hat{y}_i$ is the residual for the $i^{th}$ estimate, and N is the number of observations. The "goodness" criteria of the estimating model is that the sum of the squares of the residuals must be a minimum (i.e., least squares). The criteria may be expressed as follows:

$$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} e_i^2 = \text{minimum}$$

A regression model with two independent variables is illustrated below. This hyperplane is determined as the best fit for the equation $y = b_0 + b_1x_1 + b_2x_2$ by STATPAC regression analysis. The predicted values ($\hat{y}_i$), the observed values ($y_i$), and the residuals ($e_i$) are shown for three observations.




Because each $b_i$ in the model is raised to the first power and because there are several x's this analysis is referred to as multiple linear regression. At times, one may suspect that some of the independent variables do not significantly contribute to the prediction quality of the model. In such a case, one can examine the contribution of each variable to the model and, using the following criteria, include or exclude variables. To delete a variable, which at this point in time is in the model, the *increase* in the residual variance caused by the elimination of the variable from the model is calculated. If the increase is significant according to a user pre-selected level, the variable is deleted from the regression model; otherwise it remains in the model. To enter a variable, which at this point in time is not in the model, the *decrease* in the residual variance caused by the inclusion of this variable is calculated. If this decrease is significant, according to a user pre-selected level, the variable is added to the regression equation; otherwise it is not added. The technique of examining variables individually is denoted stepwise linear regression, (note that stepwise implies multiple variables, hence multiple linear regression).

The output of regression analysis consists of:

a. Correlation of each independent variable with the dependent variable
b. For stepwise regression, output is for each step or iteration and includes:
   (1) Variable entering (or leaving) the model
   (2) Sequential F-test which is compared with the user supplied value of F-IN (or F-OUT) to determine inclusion (exclusion) of the variable in the model
   (3) The degrees of freedom for this iteration
   (4) R-squared, the multiple correlation coefficient
   (5) The change in multiple correlation from the previous iteration

(6) The standard error of the dependent variable

(7) An analysis of variance table which includes an overall F-value which is used to test the hypothesis that all coefficients (except the constant term) are zero

(8) A table of variables in the regression consisiting of the respective coefficients, standard errors, and F-values needed to remove these terms from the model

(9) The constant term of the model ($b_0$)

(10) A table of variables not in the regression model and including the tolerance level, partial correlation with the dependent variable, and the F-value needed to enter these terms in the model

c. For multiple linear regression, the output includes:

(1) The value of R-squared, the multiple correlation coefficient

(2) The standard error of y

(3) An analysis of variance table, as in stepwise regression analysis

(4) The regression coefficients with their standard errors

(5) The constant term of the model ($b_0$)

### 1.4.2 Regression Options

The regression analysis modules have four options and four plots which may be requested by the user. The first three options allow the user to perform hypothesis tests. The last option is the output of residuals. (Analyses which involve many observations require considerable amounts of time to output residuals.)

*Option 1* of STPRG or MLTRG allows the user to test hypotheses $H_0 : b_i = 0$ or $H_1 : b_i \neq 0$. STATPAC computes the value of $t_i = b_i/\text{s.e.}(b_i)$ for $i = 1,2,...,p - 1$. If $|t_i| \leqslant t(N - p, 1 - ½\, \alpha)$, $H_0$ is accepted; otherwise, $H_1$ is accepted. The hypothesis is accepted or rejected at the $100\,(1 - \alpha)\%$ significance (confidence) level. In the expression above,

N is the number of observations in the data file

p is defined by the model such that N-p = degrees of freedom (p is the number of coefficients in the model, $y = b_0 + b_1 y_1 + ... + b_{p-1} y_{p-1}$)

$\alpha$ is the user selected significance level

Option 1 output consists of the t-value ($t_i$) for each term in the regression model, and the degrees of freedom.

*Option 2* is a generalization of option 1 which tests the hypotheses $H_0 : b_i = b_i'$ and $H_1 : b_i \neq b_i'$ where the value of $b_i'$ is supplied by the user. (Thus option 1 is option 2 with $b_i' = 0$.) STATPAC computes $t_i = (b_i - b_i')/\text{s.e.}(b_i)$ for $i = 1,2,...,p - 1$. If $|t_i| \leqslant t(N - p, 1 - ½\, \alpha)$, the hypothesis $H_0$ is accepted at the $100(1 - \alpha)\%$ significance level. Otherwise the hypothesis $H_1$ is accepted. The user is asked to supply the values for $b_i'$ in the command dialogue whenever this option is selected. Option 2 output consists of the t-values ($t_i$) for each term in the regression model, the user supplied coefficients from question 8 of the dialogue, and the degrees of freedom.

*Option 3* is a further generalization of options 1 and 2 and tests the hypothesis $H_0 : L \leqslant b_i \leqslant H$, where H and L are limits of the confidence interval for each coefficient in the regression model. The confidence interval is computed by STATPAC at the $100(1 - \alpha)\ \%$ confidence level as represented by the following equation:

$$b_i \pm t(N - p, 1 - ½\, \alpha)\ (\text{s.e.}(b_i))$$

The value of $t(N - p, 1 - ½\, \alpha)$ is supplied by the user from a t-table, for a statistical value of $\alpha$, in response to STATPAC dialogue. Option 3 output consists of the upper and lower bounds of the confidence intervals, and the degrees of freedom.

*Option 4* is requested by the user when he desires output of residual values (output is on logical 3). Regression analysis with many variables will require considerable time to output the residuals. The residual output for each observation in the original file includes the observation ordinal, the observed dependent variable ($y_i$), the predicted value of y ($\hat{y}_i$) by the model, and the residual for each observation ($e_i = y_i - \hat{y}_i$).

### 1.4.3 Regression Plots

*Plot 1* is predicted $\hat{y}$ versus residual for each observation. The plot should appear as a broad horizontal band with no trend. A plot which shows an increasing broadness indicates that the variance is not constant and the regression is suspect. Each element of the grid is a counter with maximum value equal to 9.

*Plot 2* is predicted $\hat{y}$ versus observed y. The plot indicates the ability of the model to predict the observed y's. The ideal situation would be the straight line, $\hat{y} = y$. As in plot 1, each element of the grid is a counter with maximum value equal to 9.

*Plot 3* is residuals versus ordinal. The residuals are plotted versus the ordinal of the observation to check for time factors in the model. A non-horizontal plot is an indication that the model may be time dependent. That is, either variance increases with time or terms in time should have been included in the model.

*Plot 4* is an overall plot of residuals. This histogram of residuals indicates normal distribution of the residuals.

For further description of the plots and their implications, refer to Chapter 3 of *Applied Regression Analysis,* by Draper and Smith (see Bibliography). Options 1, 2 and 3 are also described in Section 1.4 of the same reference.

# CHAPTER 2
# MODULE OPERATING PROCEDURES

STATPAC is loaded as a compiled and chained FORTRAN program (see Chapter 3 of this manual and the applicable Users' Guide). Once STATPAC is loaded, the CONTROL module assumes control. The command dialogue and possible error messages for the five modules are described in this chapter.

NOTES

1. If the user has assigned the Teletype handler (TTA) to logical 5, he may use the RUBOUT key to erase one character to the left for each striking of the key. Control U may be typed to delete the whole line. These keys can only be used to erase up to the last carriage return.
2. Characters typed by STATPAC are underlined in the following text to distinguish them from those typed by the user.
3. Sample input and dialogue for each module is contained in Chapter 4.

## 2.1 CONTROL MODULE

### 2.1.1 Command Dialogue

CONTROL contains one dialogue message which it types on logical 4.

*PROG

On logical 5, CONTROL expects one of the following five responses, left justified:

INPUT
SMMRY
STPRG
MLTRG
EXIT

### 2.1.2 Error Messages

If the user does not respond with one of the above legal responses, the CONTROL module will type:

*ERROR
*PROG

CONTROL then awaits the input of another module name.

## 2.2 INPUT MODULE

### 2.2.1 Command Dialogue

The execution of the INPUT module is directed by the following dialogue. The ASCII data file must be on logical 6. and the new file will be written on logical 1.

Question 1:
*FILE (OLD)
AAAAAAAAA

INPUT requests the user to identify the name of his data file with its nine character name (i.e., six character file name and three-character extension). If the device assigned to logical 6 is not a bulk storage device, this name is of no consequence and a carriage return or blank card will suffice. If the device is bulk storage, the nine character name may be obtained from the directory listing and must be given exactly as given on the listing.

Question 2:
*FILE (NEW)
AAAAA

Question 2 of INPUT requests the name to be associated with the standardized binary file which it will create. The user must respond with five characters in A5 format as described in the FORTRAN Manual. The file will be identified with this name for all further analysis by STATPAC. If a non-bulk storage device is assigned to logical 1, the name is of no consequence, and a carriage return or blank card will suffice.

Question 3:
*FORMAT
$(S_1, S_2, \ldots, S_n)$

The user must respond by typing the format needed to read *one* observation of his BCD (ASCII) data. The number of fields specified in the format must correspond to the number of variables per observation (see Question 5). For a discussion of format statements, see the FORTRAN Manual.

Question 4:
*NO. OBS.
IIIII

INPUT requests the number of observations in the user's data file. The user responds with an integer value of five characters right justified in the five character field. (Preceding spaces or 0s must be supplied by the user for numbers with less than five characters.)

Question 5:
*VARS
VARII=AAAAA
VARII= AAAAA
.
.
.
VARII=AAAAA
⤸

Question 5 requests the user to specify the names of the variables in his data file. He need not specify the names of all of the variables, but he must include the highest variable subscript in the list which may not be greater than 15. *The highest subscript given by the user defines the number of variables per observation.* The response must be in the form given below:

| Character Position | Content |
|---|---|
| 1 through 3 | VAR |
| 4 and 5 | subscript II ($01 \leq II \leq 15$) |
| 6 | = |
| 7 through 11 | variable name in A5 |

The list of variables is terminated by a blank record (e.g., a simple carriage return or blank card).

Question 6:
*ZERO OBS.
AAA

The user is asked if he wants those observations which contain at least one 0 variable printed. If the user wants these observations, he responds by typing "YES" in A3 format. Any other response (e.g., "NO", or a simple carriage return) will suppress the typing of such observations. (See Section 2.2.2 for a more complete description.)

Comment: *O.K.

INPUT outputs this message to terminate the dialogue, and to indicate that processing will now begin. No response is necessary.

NOTE

Once a user BCD data file has been written in standardized binary format on logical 1 by INPUT, the user need not generate the binary file again.

### 2.2.2 Error Messages

If the user types an illegal subscript (greater than 15) to Question 5 (*VARS), INPUT will type the following message and will not accept the line which was typed.

*IGNORED

The user may continue with legal responses to the question.

If a subscript less than 1 is typed, the input list is terminated and the following question will be asked (ZERO OBS.).

If a bulk storage device is assigned to logical 6 and the user does not respond with the exact name of a file in response to Question 1 (*FILE (OLD)), IOPS 13 will result.[1] No recovery is possible except by restarting execution of STATPAC. Restarting is accomplished by typing CNTRL/C after IOPS 13. The monitor from which execution may be restarted will be loaded.

If a variable of a particular observation does not conform to the user-specified format (answer to the question *FORMAT), the variable in question may be recorded as 0 in the standardized file created by INPUT. The question "ZERO OBS." allows the user to monitor 0 values for such losses. If the user requests such output, INPUT will type the entire observation on logical 4 together with the observation ordinal. The option cannot, however, distinguish between valid and assigned 0 values.

NOTE

1. Only five characters (A5) for a file name are supplied by the user in Question 2. STATPAC supplies ␣ STP as the remaining four characters of the file name in the bulk storage directory.

2. All variables in BCD data on logical 6 must be *real*; i.e., E-, F- or G-type conversion.

## 2.3 DESCRIPTIVE STATISTICS MODULE

### 2.3.1 Command Dialogue

Question 0: *PROG SMMRY

The module currently in command requests the user to specify the next module to be used. Assume the user answers with the name of the descriptive statistics module, SMMRY.

Question 1: *FILE AAAAA AAAAA . . . AAAAA ⤸

The user is asked to specify the files to be analyzed by the SMMRY module. The response consists of the names assigned to the files during execution of the INPUT module (INPUT Question 2). Response must be in A5 format, left justified. As many as ten files may be analyzed at one time. The user terminates the list of file names by supplying a blank record (e.g., a simple carriage return or a blank card).

[1]Consult the Users' Guide of your computer system for a description of the IOPS 13 error.

Question 2:

```
*VARS
II
II
 .
 .
 .
II
↲
```

The user is asked to list the subscripts of the variables to be analyzed. The response is in I2 format, right justified. At most, 15 subscripts may be listed, and the values of II must be $01 \leqslant II \leqslant 15$. The list of subscripts is terminated by a blank record.

Question 3:

```
*OPTS
IIIIII
```

The user is requested to indicate which, if any, hypothesis tests are desired. The user responds with a 0 if he does not want a specific option and with a 1 (or any positive integer) if he does want an option. The first position represents option 1, the second position represents option 2, etc. If no options are desired, the user may respond with a blank record. Examples:

| | |
|---|---|
| 001101 | User requests options 3, 4, and 6. |
| 100000 | User requests option 1 only. |

Question 4:

```
*MEAN
MENII=±XXX.XXXX
MENII=±XXX.XXXX
      .
      .
      .
MENII=±XXX.XXXX
↲
```

If option 1 is requested in response to Question 3, the user is requested to provide a test mean for each variable in the analysis. The user must respond in the following form.

| Character Position | Content |
|---|---|
| 1 through 3 | MEN |
| 4 through 5 | subscript II ($01 \leqslant II \leqslant 15$) |
| 6 | = |
| 7 through 15 | test mean in F9.6 |

The list is terminated by a blank record. An entry for a subscript may be retyped and only the last appearance will be used. If a variable which is in the analysis by virtue of being listed in Question 2 is not assigned a test mean, a default test mean of 0 is assumed.

Question 5:

```
*VRNC
VARII=±XXX.XXXX
VARII=±XXX.XXXX
      .
      .
      .
VARII=±XXX.XXXX
↲
```

The user who has requested option 2 in Question 3 is requested to provide a test variance for each variable in the analysis. The user must respond in the following form.

| Character Position | Content |
|---|---|
| 1 through 3 | VAR |
| 4 through 5 | subscript II ($01 \leqslant II \leqslant 15$) |
| 6 | = |
| 7 through 15 | test variance in F9.5 |

The list is terminated by a blank record. An entry for a subscript may be retyped after it has already been entered and only the last appearance will be used. If a variable in the analysis is not assigned a test mean, because it was listed in Question 2 a default test mean of 0 is assumed by STATPAC.

Comment:
*O.K.

STATPAC indicates the termination of dialogue and the start of the requested analysis by typing the comment O.K.

### 2.3.2 Error Messages

If the user responds with a subscript greater than 15 in response to Questions 2, 4, or 5, that particular line is completely ignored but no message is typed. If the user responds with a non-positive subscript to one of these same questions, this is treated as a list terminator and the dialogue proceeds to the next question.

If the user responds to Question 2 (*VARS) with a subscript which is greater than the number of variables/observations in some data file which he has listed in Question 1 (*FILE), the message

*ERR1

will be output on logical 4. Note that the number of variables per observation is defined by the user's response to Question 5 of INPUT. When this error condition exists, the file in question is excluded from further analysis and processing continues. If all the user listed files from Question 1 are eliminated from analysis, the following message will be output on logical 4:

*ERR2

If the user requests option 3, 4, 5, or 6 and has listed only one file name for analysis in response to Question 1, these options will not be processed and no output will result, since they are meaningless for only one file.

NOTES

1. When the elements of the correlation matrix are being calculated, the terms

$$\sum_{m=1}^{N} (x_{jm} - \bar{x}_j)^2, \quad \sum_{m=1}^{N} (x_{im} - \bar{x}_i)^2$$

are used in the denominator of the expression for calculating $c_{ij}$. If both of these terms are not larger than TOL = .1E-9, then $c_{ij}$ is given the default value 2.0.

2. If, during the processing of option 2, a user supplied variance is found to be less than or equal to TOL = .1E-9, the corresponding statistic is given the default value of 1.E76. Similarly, if during the processing of option 1, a standard deviation is calculated from the data file which is less than .TOL, the corresponding statistic is assigned the default value .1E76. Option 4 operates similarly.

## 2.4 REGRESSION ANALYSIS MODULES

### 2.4.1 Command Dialogue

Question 0:
*PROG
STPRG
or MLTRG

The module currently in command types this question. The user is assumed to have typed either STPRG or MLTRG.

Question 1:
*FILE
AAAAA

STPRG or MLTRG requests the name of the data file to be analyzed. The user responds in A5 with the exact name of the file which was given in answer to INPUT Question 2.

Question 2:
*VARS
II
II
II
.
.
.
II
↵

STPRG or MLTRG requests the subscripts of the variables to be analyzed. The user must respond in I2, followed by a carriage return, and right justified in the two character field. Values must be $01 \leqslant II \leqslant 15$. *The last subscript of the list will be considered the dependent variable.* Each subscript is terminated by a carriage return, and the list is terminated by a blank record (e.g., an extra carriage return or blank card). If all variables of the file are to be analyzed, the user need only type the subscript of the dependent variable.

Question 3:
*FIN
XXXX.XXXX

STPRG requests the F-value which will be used to determine if a variable not in the model makes a significant contribution to the model and, therefore, should be added. The user responds in F9.5 followed by a carriage return.

NOTE

This question is asked only when the user response to Question 0 is STPRG.

Question 4:
*FOUT
XXXX.XXXX

Question 4 of STPRG requests the F-value to determine if the contribution of a variable, which is in the estimating model, is insignificant and should therefore be excluded from the model. Response is in F9.5. See NOTE for Question 3.

Question 5:
*LIM
II

Question 5 of STPRG requests the user to specify the number of iterations or cycles to be allowed in the calculation of the estimating model. The limit prevents STPRG from getting into a nonproductive loop of successively including and excluding variables in the model. If the user responds with other than a positive integer, STPRG will use a default limit equal to twice the number of independent variables being analyzed. See NOTE for Question 3.

Question 6:
*TOL
XXXX.XXXX

Question 6 requests the tolerance factor used by STPRG and MLTRG. The tolerance is used to check for constant observations and to check the diagonal elements of the correlation matrix to avoid trying to invert a badly behaved matrix. Values for TOL are usually between 0.001 and 0.0001. If the user responds with a blank record, STATPAC uses a default value of TOL = .001.

Question 7:
*OPTS
IIII

STPRG or MLTRG requests the user to specify which options are desired. The four options are described in detail in Section 1.4.2. The user responds with a 0 if he does not want an option or with a 1 if he does want the option. Examples:

| | |
|---|---|
| 1000 | Option 1 only |
| 1010 | Options 1 and 3 only |
| 0011 | Options 3 and 4 only |

Question 8:

*COEF
COFII=XXXX.XXXX
COFII=XXXX.XXXX
.
.
.
COFII=XXXX.XXXX
⤸

This question is asked only if the user requests option 2 in Question 7 (i.e., he types 1 in the second position of the response to Question 7). The user must respond as outlined below.

| Character Position | Content |
|---|---|
| 1 through 3 | COF |
| 4 and 5 | subscript II ($01 \leqslant II \leqslant 15$) |
| 6 | = |
| 7 through 15 | test coefficient in F9.5 |

The list is terminated by a blank record.

Question 9:

*FCTR
XXXX.XXXX

This question is asked only if the user requests option 3 in Question 7 (i.e., he types a 1 in the third position of the response to Question 7). The user must respond with the value of $t(N - p, 1 - \frac{1}{2}\alpha)$ in F9.5. The t-value is obtained from a table by estimating the degrees of freedom (N-p) and specifying a confidence level. The actual degrees of freedom are output by the regression module and may be checked against the estimated value used to obtain the response.

Question 10:

*PLTS
IIII

The user is requested to specify the output plots which he would like. The plots are described in Section 1.4.3, and are listed below. The user types a 1 in the position corresponding to those plots he wants and a 0 in the positions corresponding to plots he does not want. Examples:

| | |
|---|---|
| 1000 | Plot 1 only |
| 0010 | Plot 3 only |
| 1011 | Plot 1, 3, and 4 |

Comment:

*O.K.

The regression analysis modules type this message to indicate termination of the dialogue and start of the processing.

NOTE

An "A" appears in plotted output when the value to be plotted is a counter which exceeds 9.

### 2.4.2 Error Messages

If the user responds to Question 2 (*VARS) or Question 8 (*COEF) with a subscript value greater than 15, the line in question will be ignored. If a non-positive subscript is typed, the list in question is terminated and STATPAC types the next question in the dialogue.

If the following expression is less than the user-supplied tolerance for a specific variable with subscript j:

$$\left( \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2 \right)^{\frac{1}{2}} \leqslant TOL$$

where N is the number of observations, the $j^{th}$ variable is considered constant by STATPAC and the following error message is output on logical 4:

*ERR 1 j

(The value j is the subscript of the variable which caused the error.) ERR 1 will terminate processing and a new dialogue will begin.

If the MLTRG module is being used and not all the independent variables can be entered (due to the choice of TOL in part) the following error message will be output on logical 4:

*ERR 2

ERR 2 will terminate processing and STATPAC will begin a new dialogue.

If the user has responded with a subscript greater than the number of variables/observation, STATPAC outputs the following error message:

*ERR 3

ERR 3 will terminate processing and cause STATPAC to begin a new dialogue. Note that the number of variables per observation is defined by the user's response to Question 5 of INPUT.

When the STPRG module is being used, a limit factor (LIM) is used to limit the number of passes in the stepping algorithm. Exceeding this limit causes the following error message on logical 4:

*ERR 4

ERR 4 terminates processing and STATPAC begins a new dialogue.

If no variables are entered into the regression model when using either the MLTRG or the STPRG modules, the following message is output on logical 4:

*ERR 5

ERR 5 terminates processing and STATPAC begins a new dialogue.

NOTE

When any of the above error conditions occur, the user should not indiscriminately adjust the values of TOL, FIN, LIM, etc., to force a complete analysis. The user should closely examine the variables and limits involved before making any such adjustments.

# CHAPTER 3
# IMPLEMENTING AND AUGMENTING STATPAC

This chapter describes the procedure to be followed when building a STATPAC executable file using the PDP-15 and PDP-9 Monitors. Refer to CHAIN and EXECUTE of the Users' Guide for the general description for chaining the STATPAC program. The procedure for building an executable file with specific hardware and handler assignments is described later in this chapter.

STATPAC modules make use of the following logical units. Specific device handlers must be assigned to these units when the executable file is built.

| Logical Unit | Function within STATPAC |
|---|---|
| -4 | Contains STATPAC in executable form in IOPS binary. |
| 1 | Contains standardized binary data files written by the INPUT module in IOPS binary. |
| 2 | Stores temporary files during processing by a STATPAC module in IOPS binary. |
| 3 | Hard copy statistical output in IOPS ASCII. |
| 4 | Program queries and error messages in IOPS ASCII. |
| 5 | User responses to queries in IOPS ASCII. |
| 6 | User supplied BCD data files as input to the INPUT module in IOPS ASCII. |
| 7 | Temporary storage of residuals (regression) and temporary storage of option output for SMMRY in IOPS binary. |

If the user assigns a bulk storage device to logical unit 3, the output of STATPAC will be recorded in files with the following file names:

Descriptive Statistics

| | |
|---|---|
| SMMRY STP | Contains the standard SMMRY module output |
| OPTON STP | Contains the output of the options for SMMRY |

Regression Analysis

| | |
|---|---|
| REGRS STP | Contains the standard regression output (for each step in the case of STPRG) and the residual output, if requested |
| OPTPL STP | Contains the output for regression options 1, 2, and 3 and the plotted output, if requested |

### 3.1 BUILDING AN EXECUTABLE FILE

STATPAC is supplied to the user in three forms:

a. Source files for each STATPAC chain.
b. Binary files of the FORTRAN compiled STATPAC source for each chain.
c. An executable file with fixed handler assignments, chained as described in this section.

The files are organized according to the following chart.

| Module | Source File | Binary File |
|---|---|---|
| CONTROL | CH01 SRC | CH01 BIN |
| INPUT | CH03 SRC | CH03 BIN |
| SMMRY | CH06 SRC | CH06 BIN |
| SMMRY | CH07 SRC | CH07 BIN |
| SMMRY | CH08 SRC | CH08 BIN |
| SMMRY | CH09 SRC | CH09 BIN |
| STPRG & MLTRG | CH10 SRC | CH10 BIN |
| STPRG & MLTRG | CH11 SRC | CH11 BIN |
| STPRG & MLTRG | CH12 SRC | CH12 BIN |
| STPRG & MLTRG | CH13 SRC | CH13 BIN |

NOTE

16K of core memory is required to build the STATPAC executable file, although it will operate in an 8K memory.

An executable file is produced from the above chains by following the steps outlined below. The description assumes that the user has two DECtapes as bulk storage devices for the Monitor System. Assuming also that the compiled binary files are on DECtape unit 1; the user should make the following handler assignments:

```
$A DTA0 -1
$A DTA1 -4,-6
$A DTB1 1,2,6,7
$A TTA 3,4,5
$CHAIN
```

Once CHAIN has been loaded from the system tape, the message "CHAIN V2A" is typed as shown below. The user then types the command "BUILD STATPC", thereby identifying the executable file to be built. The user then types all responses which are preceded by a > in the following listing. (Lines preceded by a > indicate user response is required. Lines not preceded by a > are typed by CHAIN.)

```
CHAIN V2A
>BUILD STATPC
>C 1
>CH01
>END
 CH01    36655
 BCDIO   33662
 .SS     33603
 GOTO    33555
 STOP    33542
 SPMSG   33447
 FIOPS   32713
 OTSER   32617
 INTEGE  32467
 REAL    31466

 CHAIN#  1
 LOWEST  31466
 COMSZE  00010

>C 3
>CH03
>END
 CH03    34672
 DTB.    32647
 FILE    31325
 .DA     31256
 BCDIO   26263
 BINIO   26012
 .SS     25733
 FIOPS   25177
 OTSER   25103
 INTEGE  24753
 REAL    23752

 CHAIN#  3
 LOWEST  23752
 COMSZE  00010

>C 6
>CH06
>END
 CH06    35077
 DTB.    32054
 FILE    31532
 .DA     31463
 BCDIO   26470
 BINIO   26217
 .SS     26140
 GOTO    26112
 FIOPS   25356
 OTSER   25262
 INTEGE  25132
 REAL    24131
```

```
 CHAIN# 6
 LOWEST 24131
 COMSZE 00167

>C 7
>CH07
>END
 CH06   33571
 DTB.   30546
 FILE   30224
 FLOAT  30213
 SQRT   30125
 .DA    30056
 BINIO  27605
 .SS    27526
 FIOPS  26772
 OTSER  26676
 INTEGE 26546
 REAL   25545

 CHAIN# 7
 LOWEST 22545
 COMSZE 00167

>C 8
>CH08
>END
 CH08   33620
 DTB.   30575
 FILE   30253
 ABS    30235
 FLOAT  30224
 SQRT   30136
 ALOG10 30116
 .EE    30025
 .EC    27761
 .DA    27712
 BINIO  27441
 .SS    27362
 GOTO   27334
 FIOPS  26600
 OTSER  26504
 INTEGE 26354
 REAL   25353

 CHAIN# 10
 LOWEST 25353
 COMSZE 00167

>C 9
>CH09
>END
 CH09   35511
 DTB.   32466
 FILE   32144
 .DA    32075
 BCDIO  27102
 BINIO  26631
 .SS    26552
 GOTO   26524
 FIOPS  25770
 OTSER  25674
 INTEGE 25544
 REAL   24543
```

```
 CHAIN# 11
 LOWEST 24543
 COMSZE 00167

>C 10
>CH10
>END
 CH10    34420
 DTB.    31375
 FILE    31053
 .DA     31004
 BCDIO   26011
 BINIO   25540
 .SS     25461
 GOTO    25433
 FIOPS   24677
 OTSER   24603
 INTEGE  24453
 REAL    23452

 CHAIN# 12
 LOWEST 23452
 COMSZE 00125

>C 11
>CH11
>END
 CH11    31553
 DTB.    26530
 FILE    26206
 ABS     26170
 IABS    26154
 FLOAT   26143
 SQRT    26055
 .DA     26006
 BINIO   25535
 .SS     25456
 FIOPS   24722
 OTSER   24626
 INTEGE  24476
 REAL    23475

 CHAIN# 13
 LOWEST 23475
 COMSZE 00273

>C 12
>CH12
>END
 CH12    33003
 DTB.    27760
 FILE    27436
 ABS     27420
 .DA     27351
 BINIO   27100
 .SS     27012
 GOTO    26773
 FIOPS   26237
 OTSER   26143
 INTEGE  26013
 REAL    25012
```

```
 CHAIN# 14
 LOWEST 25Ø12
 COMSZE ØØ273

>C 13
>CH13
>END
 CH13    35217
 DTB.    32174
 FILE    31652
 .DA     316Ø3
 BCDIO   2661Ø
 BINIO   26337
 .SS     2626Ø
 GOTO    26232
 FIOPS   25476
 OTSER   254Ø2
 INTEGE  25252
 REAL    24251

 CHAIN# 15
 LOWEST 24251
 COMSZE ØØØ25

>CLOSE

CHAIN V2A
>EXIT

MONITOR V4B
```

NOTE

Chain numbers are typed by the user in decimal, but CHAIN prints the chain numbers in octal.

When the message "CHAIN V2A" is typed at the end of the listing, STATPC is on DECtape 1 in executable format. After calling the MONITOR, STATPC is executed by typing:

A DTC1 -4
E STATPC

NOTE

STATPC XCT is the name of the executable file which is stored on logical unit 1 by CHAIN.

The assignments made to build the executable file described in this chapter result in all files (ASCII, binary, or temporary) being stored on DECtape 1, and all hard copy output being on the Teletype unit, which is also used for the dialogue of the modules.

The user can increase the processing and output speed by optimally assigning peripheral handlers to the STATPAC logical units.

## 3.2 ADDING PROCESSING MODULES TO STATPAC

The user has the ability to add FORTRAN coded modules to STATPAC by following a few simple conventions. For example, assume the user wishes to add a module which is written in one chain to STATPAC. The name of the new module is ABCDE. The chain must include the following statements:

```
           .
           .
           .
      COMMON ICNTRL, CPBLTY, IFLAG, ...
           .
           .
           .
      DATA ABCDE/5HABCDE/
           .

C     PROCESSING BEGINS HERE
100   CONTINUE
        :
C     PROCESSING IS FINISHED
200   WRITE (4,201)
201   FORMAT (6H*PROG)
      READ (5,202) CPBLTY
202   FORMAT (A5)
      IF (CPBLTY.EQ.ABCDE) GO TO 100
C     DIFFERENT MODULE REQUESTED, CALL CONTROL MODULE
      CALL CHAIN (1)
      END
```

If the newly added module must read a data file which was written by INPUT, the file (on logical-1) could be read with the following coding:

```
        .
        .
        .
C     READ AAAAA, WHICH WAS GIVEN BY USER IN QUEST. 2 OF INPUT.
      READ (5,100) FILE (1)
100   FORMAT (A5)
C     FILE (2)=4H STP
      CALL SEEK (1,FILE)
      READ (1) L, (NAME(I),I=1,L)
C     INITIALIZE OBSERVATION COUNTER
      N=0
103   READ (1) NO
      IF (NO.EQ.0) GO TO 102
      N=N+NO
      DO 101  NO1=1,NO
      READ (1) (X(I),I=1,L)
        .
        .
        .
```

```
101   CONTINUE
      GO TO 103
C     ALL OBSERVATIONS READ
C     N TOTAL NUMBER OF OBSERVATIONS
102   CALL CLOSE (1)
  .
  .
  .
```

The format of the files formed by INPUT which must be read by a user written module for analysis is given in Section 1.2. This format is summarized below. The file is stored on a bulk storage device with directory entry AAAAA STP where the name AAAAA is supplied by the user in response to Question 2 of INPUT and the STP is automatically supplied by STATPAC. In the following description, L is the number of variables (the highest acceptable subscript supplied in Question 5 of INPUT).

| | |
|---|---|
| [ L, NAME(1),...,NAME(L) ] | Contains the number of variables per observation and their respective names |
| [ NO ] | NO is the number of observations which follow |
| [ $X_1,X_2,...X_L$ ] | Contains one observation |
| . . . | |
| [ $X_1,X_2,...,X_L$ ] | Contains one observation |
| [ NO ] | NO is the number of observations which follow |
| [ $X_1,X_2,...X_L$ ] | Contains one observation |
| . . . | |
| [ $X_1,X_2,...,X_L$ ] | Contains one observation |
| . . . | |
| [ 0 ] | Contains 0 to indicate that zero observations follow, i.e., the end of the data file |

The entries in COMMON which must be made in the user supplied module are the following:

| | |
|---|---|
| ICNTRL | Used by the CONTROL module |
| CPBLTY | Used if the user requests a different module once processing is completed by the module presently in core (used to read the answer to "*PROG") |
| IFLAG | May be used in the user supplied program to define multiple entry points into a chain if the module occupies more than one chain (set to 1 by CONTROL) |

When the user adds a module to STATPAC, the CONTROL module must be modified to allow the new module to be called. The following changes must be made to CONTROL, assuming that the new module is named ABCDE and that chain number 20 is assigned to it when the executable file is built. (The CONTROL module is chain 1.)

a. Increase the dimension of TABLE by 1 (i.e., for the first addition, TABLE (6) is the correct dimension).

b. Add the following DATA statement to the CONTROL module:

   DATA ABCDE/5HABCDE/

c. Increase the value of MODULS by 1 (i.e., for the first addition, MODULS/5/ is changed to MODULS/6/).

d. The module selector statement "GOTO (101,102,103,104,105),I" should be changed to add the statement number of a CALL CHAIN command. For the first addition, the GOTO statement is changed to:

```
GOTO (101,102,103,104,105,106),I
```

and the following statement is added after statement 105:

```
106 CALL CHAIN (20)
```

## CHAPTER 4
## SAMPLE OPERATION

The user dialogue with STATPAC and the possible data output are illustrated in this chapter. Each of the options and plots which may be requested in the modules is included in the output. The responses to the initial dialogue are for illustrative purposes only, and are not intended as examples of statistically meaningful responses.

The user is referred to the following books for more complete descriptions of the statistical applications of the various options and plots. The selection of tolerances and test means, variances, etc., is discussed in these references.

| | |
|---|---|
| Descriptive Statistics: | *Statistics in Research*<br>by Bernard Ostle<br>Chapter 7 |
| | *Quality Control and Industrial Statistics*<br>by Acheson J. Duncan, PH. D.<br>Chapter 4 |
| Regression Analysis: | *Applied Regression Analysis*<br>by N. R. Draper and H. Smith<br>Chapters 1,2,3,4, and 6 |
| | *BMD Biomedical Computer Programs*<br>edited by W. J. Dixon<br>Pages 233-257 |
| | *Mathematical Methods for Digital Computers*<br>edited by Anthony Ralston, PH. D.<br>and Herbert S. Wilf, Ph. D.<br>Chapter 17 |

The operation of the CONTROL module is not illustrated explicitly since the only question used specifies the analysis module desired by the user.

The data used to illustrate the STATPAC modules was obtained from *BMD Biomedical Computer Programs,* published by the University of California Press, used with permission of the editor, Mr. W. J. Dixon.

## 4.1 INPUT EXAMPLE

The following Teletype listing is an example of the dialogue for the INPUT module. Any printed line which is started with an asterisk (*) is typed by STATPAC; all other lines are typed by the user. Comments are added to the listing to aid the reader.

The message

***WARNING - COMMON SIZE DIFFERS***

is often typed during the execution of STATPAC. This message is typed by CHAIN and should be ignored by the STATPAC user.

Following the INPUT dialogue, a complete listing of the BMD data is provided to illustrate the format used. This data (its source was credited at the beginning of this chapter) is analyzed by the STATPAC modules and is used throughout this chapter to illustrate the operation of the various modules.

```
$E STATPC

*PROG

INPUT

*FILE (OLD)
                                      ← 3 spaces
BMD   SRC
*FILE (NEW)
BMD
*FORMAT                               ← note use of RUBOUT key
(F7.2,7\F7.0,2F7.2,2F7.0)
*NO.   OBS.
00068
*VARS
VAR01=ONEBM
VAR02=TWOBM
VAR03=TREBM
VAR04=QRTBM
VAR06=SIXBM
VAR05=FIVBM
                                      ← blank record (i.e., simple carriage return)
*ZERO OBS.

*0.K.
```

| | | | | | |
|---|---|---|---|---|---|
| 00250 | 00025 | 02500 | 00150 | 00034 | 00064 |
| 01300 | 00021 | 02100 | 00087 | 00036 | 00065 |
| 00350 | 00022 | 02200 | 00043 | 00041 | 00082 |
| 00175 | 00009 | 00130 | 00180 | 00015 | 00023 |
| 00300 | 00023 | 02300 | 00200 | 00033 | 00064 |
| 00200 | 00010 | 00060 | 00330 | 00013 | 00016 |
| 00550 | 00007 | 00140 | 00340 | 00016 | 00012 |
| 00600 | 00006 | 00080 | 00500 | 00011 | 00027 |
| 00130 | 00008 | 00270 | 00150 | 00019 | 00048 |
| 00500 | 00018 | 00360 | 00180 | 00027 | 00050 |
| 00500 | 00003 | 00100 | 00140 | 00014 | 00012 |
| 00300 | 00008 | 00270 | 00100 | 00025 | 00013 |
| 00200 | 00006 | 00300 | 00150 | 00021 | 00020 |
| 00200 | 00008 | 00100 | 00250 | 00018 | 00023 |
| 00100 | 00022 | 02200 | 00110 | 00046 | 00118 |
| 00400 | 00013 | 01300 | 00280 | 00017 | 00050 |
| 00050 | 00026 | 00120 | 00073 | 00048 | 00063 |
| 00025 | 00023 | 02300 | 00010 | 00036 | 00150 |
| 01400 | 00003 | 00100 | 00350 | 00005 | 00072 |
| 00250 | 00015 | 00250 | 00028 | 00033 | 00054 |
| 00350 | 00028 | 01400 | 00001 | 00046 | 00109 |
| 00350 | 00006 | 00060 | 00500 | 00010 | 00010 |
| 00250 | 00035 | 03500 | 00570 | 00038 | 00125 |
| 00050 | 00011 | 00200 | 00340 | 00016 | 00044 |
| 00200 | 00011 | 01100 | 00050 | 00020 | 00048 |
| 00700 | 00032 | 03200 | 00660 | 00038 | 00105 |
| 00400 | 00008 | 00100 | 00450 | 00012 | 00009 |
| 01500 | 00023 | 02300 | 00015 | 00049 | 00130 |
| 00100 | 00038 | 03800 | 00220 | 00043 | 00160 |
| 00350 | 00015 | 00500 | 00150 | 00033 | 00048 |
| 01300 | 00006 | 00120 | 00370 | 00009 | 00036 |
| 00200 | 00025 | 02500 | 00100 | 00035 | 00150 |
| 01200 | 00005 | 00170 | 00030 | 00021 | 00078 |
| 00400 | 00009 | 00075 | 00190 | 00017 | 00023 |
| 00300 | 00007 | 00350 | 00260 | 00012 | 00042 |
| 00800 | 00020 | 02000 | 00220 | 00030 | 00072 |
| 00900 | 00006 | 00086 | 00250 | 00015 | 00020 |
| 00600 | 00012 | 00400 | 00120 | 00020 | 00036 |
| 00800 | 00026 | 00160 | 00110 | 00035 | 00056 |
| 00150 | 00015 | 00300 | 00160 | 00029 | 00036 |
| 00700 | 00010 | 00090 | 01000 | 00012 | 00026 |
| 00800 | 00028 | 02800 | 00420 | 00040 | 00108 |
| 00200 | 00034 | 03400 | 00090 | 00042 | 00106 |
| 00600 | 00004 | 00080 | 00360 | 00011 | 00016 |
| 01500 | 00032 | 03200 | 00180 | 00044 | 00104 |
| 01700 | 00011 | 01100 | 00230 | 00014 | 00047 |
| 01600 | 00002 | 00050 | 00180 | 00011 | 00027 |
| 00300 | 00018 | 00160 | 00110 | 00032 | 00012 |
| 00600 | 00003 | 00040 | 00130 | 00015 | 00007 |
| 01400 | 00008 | 00110 | 00200 | 00017 | 00018 |
| 00600 | 00014 | 00090 | 00070 | 00029 | 00028 |
| 00180 | 00012 | 00240 | 00150 | 00021 | 00025 |
| 01500 | 00003 | 00150 | 00080 | 00013 | 00011 |
| 01800 | 00006 | 00550 | 00570 | 00009 | 00020 |
| 00500 | 00012 | 00200 | 00410 | 00016 | 00014 |
| 03000 | 00011 | 01100 | 00200 | 00022 | 00038 |
| 02900 | 00008 | 00800 | 00100 | 00022 | 00103 |
| 00180 | 00024 | 02400 | 00110 | 00038 | 00106 |
| 01300 | 00026 | 02600 | 00170 | 00038 | 00063 |
| 01900 | 00029 | 02900 | 04800 | 00029 | 00208 |
| 01100 | 00017 | 01700 | 00160 | 00025 | 00032 |
| 01000 | 00015 | 00500 | 00350 | 00019 | 00028 |
| 00600 | 00010 | 00500 | 00100 | 00026 | 00032 |
| 00500 | 00022 | 02200 | 00120 | 00039 | 00100 |
| 00100 | 00015 | 00500 | 00080 | 00029 | 00050 |
| 01700 | 00009 | 00300 | 01300 | 00010 | 00080 |
| 00500 | 00030 | 03500 | 00090 | 00058 | 00065 |
| 00130 | 00010 | 00130 | 00900 | 00010 | 00025 |

## 4.2 SMMRY EXAMPLE

The following statistical output is the result of analysis by SMMRY on the previously presented BMD data and two arbitrary files, DATA1 and DATA2. The arbitrary data files are included to allow the user to select all six options of SMMRY for demonstration purposes. Thus the complete printed output is presented with options.

```
*PROG

SMMRY
***WARNING - COMMON SIZE DIFFERS***
***WARNING - COMMON SIZE DIFFERS***
```

This message should be ignored.

```
*FILE
```

User requests analysis of three data files - BMD, DATA1, and DATA2.

```
BMD
DATA1
DATA2

*VARS
01
04
05
02
03
```

User types the variables to be analyzed - no specific order necessary.

```
*OPTS
221355
```

User requests all six options - any nonzero digit may be used to request an option.

```
*MEAN
MEN01=1.6
MEN02=1.7
MEN03=1.8
MEN04=2.3
MEN05=2.5
```

User types the values for the test mean of each variable - Option 1 was requested.

```
*VRNC
VAR01=4.1
VAR02=4.2
VAR03=4.3
VAR04=4.4
VAR05=3.5
```

User types the values of the test variances of each variable - option 2 was requested.

```
*O.K.
```

Processing will now begin.

Carriage returns typed to terminate lists.

DESCRIPTIVE STATISTICS

BMD ← Name of file being analyzed.

NO. OBS. = 68

| VARIABLE NO. | NAME | MEAN | VARIANCE | STANDARD DEVIATION | STANDARD ERROR |
|---|---|---|---|---|---|
| 1 | ONEBM | 0.69956E+01 | 0.41909E+02 | 0.64738E+01 | 0.78506E+00 |
| 2 | TWOBM | 0.15250E+02 | 0.87563E+02 | 0.93575E+01 | 0.11348E+01 |
| 3 | TREBM | 0.10425E+02 | 0.13519E+03 | 0.11627E+02 | 0.14100E+01 |
| 4 | ORTBM | 0.30996E+01 | 0.35966E+02 | 0.59971E+01 | 0.72726E+00 |
| 5 | FIVBM | 0.25397E+02 | 0.15574E+03 | 0.12479E+02 | 0.15134E+01 |

| | SKEWNESS | KURTOSIS | MAX | MIN | RANGE |
|---|---|---|---|---|---|
| 1 | 0.15008E+01 | 0.21928E+01 | 0.30000E+02 | 0.25000E+00 | 0.29750E+02 |
| 2 | 0.57906E+00 | -0.80053E+00 | 0.38000E+02 | 0.20000E+01 | 0.36000E+02 |
| 3 | 0.89081E+00 | -0.72163E+00 | 0.38000E+02 | 0.40000E+00 | 0.37600E+02 |
| 4 | 0.62546E+01 | 0.43371E+02 | 0.48000E+02 | 0.10000E-01 | 0.47990E+02 |
| 5 | 0.43154E+00 | -0.85983E+00 | 0.58000E+02 | 0.50000E+01 | 0.53000E+02 |

------------------------CORRELATION MATRIX--------------------------

| | ONEBM 1 | TWOBM 2 | TREBM 3 | QRTBM 4 | FIVBM 5 |
|---|---|---|---|---|---|
| 1 ONEBM | 1.000000 | | | | |
| 2 TWOBM | -0.176465 | 1.000000 | | | |
| 3 TREBM | 0.005134 | 0.867992 | 1.000000 | | |
| 4 QRTBM | 0.255482 | 0.100719 | 0.125866 | 1.000000 | |
| 5 FIVBM | -0.195689 | 0.879120 | 0.751937 | -0.140437 | 1.000000 |

**The name and subscript of the third variable in the analysis.**

DATA1

NO. OBS. = 25

| VARIABLE NO. NAME | MEAN | VARIANCE | STANDARD DEVIATION | STANDARD ERROR |
|---|---|---|---|---|
| 1 ONE01 | 0.45600E+02 | 0.12417E+02 | 0.35237E+01 | 0.70475E+00 |
| 2 TWO01 | 0.31300E+03 | 0.14840E+05 | 0.12182E+03 | 0.24364E+02 |
| 3 TRE01 | 0.24136E+03 | 0.15286E+04 | 0.39097E+02 | 0.78194E+01 |
| 4 QRT01 | 0.10868E+03 | 0.32631E+03 | 0.18064E+02 | 0.36128E+01 |
| 5 FIV01 | 0.37440E+02 | 0.23059E+03 | 0.15185E+02 | 0.30370E+01 |

| | SKEWNESS | KURTOSIS | MAX | MIN | RANGE |
|---|---|---|---|---|---|
| 1 | -0.46517E+00 | 0.12996E+01 | 0.51000E+02 | 0.35000E+02 | 0.16000E+02 |
| 2 | 0.68285E+00 | 0.10441E+01 | 0.67700E+03 | 0.13900E+03 | 0.53800E+03 |
| 3 | 0.14744E+01 | 0.18508E+01 | 0.35700E+03 | 0.19800E+03 | 0.15900E+03 |
| 4 | 0.46925E+00 | -0.84454E+00 | 0.14700E+03 | 0.79000E+02 | 0.68000E+02 |
| 5 | 0.33407E+00 | -0.13492E+01 | 0.64000E+02 | 0.15000E+02 | 0.49000E+02 |

------------------------CORRELATION MATRIX--------------------------

| | ONE01 1 | TWO01 2 | TRE01 3 | QRT01 4 | FIV01 5 |
|---|---|---|---|---|---|
| 1 ONE01 | 1.000000 | | | | |
| 2 TWO01 | 0.011551 | 1.000000 | | | |
| 3 TRE01 | -0.291373 | -0.179528 | 1.000000 | | |
| 4 QRT01 | 0.352039 | -0.029065 | -0.460303 | 1.000000 | |
| 5 FIV01 | 0.310231 | -0.164700 | -0.333221 | 0.938052 | 1.000000 |

DATA2

NO. OBS. = 10

| VARIABLE NO. NAME | MEAN | VARIANCE | STANDARD DEVIATION | STANDARD ERROR |
|---|---|---|---|---|
| 1 ONE02 | 0.78000E+01 | 0.39956E+02 | 0.63210E+01 | 0.19989E+01 |
| 2 TWO02 | 0.51500E+02 | 0.20828E+03 | 0.14432E+02 | 0.45637E+01 |
| 3 TRE02 | 0.12400E+02 | 0.47822E+02 | 0.69154E+01 | 0.21868E+01 |
| 4 QRT02 | 0.25800E+02 | 0.18907E+03 | 0.13750E+02 | 0.43482E+01 |
| 5 FIV02 | 0.98340E+02 | 0.20214E+03 | 0.14218E+02 | 0.44961E+01 |

| | SKEWNESS | KURTOSIS | MAX | MIN | RANGE |
|---|---|---|---|---|---|
| 1 | 0.58356E+00 | -0.70511E+00 | 0.21000E+02 | 0.10000E+01 | 0.20000E+02 |
| 2 | -0.12735E+00 | -0.15380E+01 | 0.71000E+02 | 0.31000E+02 | 0.40000E+02 |
| 3 | 0.37643E+00 | -0.16785E+01 | 0.23000E+02 | 0.40000E+01 | 0.19000E+02 |
| 4 | 0.12898E+00 | -0.14738E+01 | 0.47000E+02 | 0.60000E+01 | 0.41000E+02 |
| 5 | -0.35647E+00 | -0.13386E+01 | 0.11590E+03 | 0.72500E+02 | 0.43400E+02 |

------------------------CORRELATION MATRIX------------------------

| | ONE02 1 | TWO02 2 | TRE02 3 | QRT02 4 | FIV02 5 |
|---|---|---|---|---|---|
| 1 ONE02 | 1.000000 | | | | |
| 2 TWO02 | 0.095004 | 1.000000 | | | |
| 3 TRE02 | -0.872371 | -0.249384 | 1.000000 | | |
| 4 QRT02 | -0.107896 | -0.972027 | 0.150505 | 1.000000 | |
| 5 FIV02 | 0.729910 | 0.728978 | -0.714846 | -0.737734 | 1.000000 |

$N_i$ - 1 (degrees of freedom for each data file)

OPTION 1

| VAR | USER MEAN | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| 1 | 0.1600E+01 | 0.6873E+01 | 0.6243E+02 | 0.3102E+01 |
| 2 | 0.1700E+01 | 0.1194E+02 | 0.1278E+02 | 0.1091E+02 |
| 3 | 0.1800E+01 | 0.6117E+01 | 0.3064E+02 | 0.4847E+01 |
| 4 | 0.2300E+01 | 0.1099E+01 | 0.2945E+02 | 0.5405E+01 |
| 5 | 0.2500E+01 | 0.1513E+02 | 0.1150E+02 | 0.2132E+02 |

Answers to Question 4

t-values to be compared with values obtained from t-table.

OPTION 2

| VAR | USER VARNC | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| 1 | 0.4100E+01 | 0.6849E+03 | 0.7268E+02 | 0.8771E+02 |
| 2 | 0.4200E+01 | 0.1397E+04 | 0.8480E+05 | 0.4463E+03 |
| 3 | 0.4300E+01 | 0.2106E+04 | 0.8532E+04 | 0.1001E+03 |
| 4 | 0.4400E+01 | 0.5477E+03 | 0.1780E+04 | 0.3867E+03 |
| 5 | 0.3500E+01 | 0.2981E+04 | 0.1581E+04 | 0.5198E+03 |

Answers to Question 5

chi-square values to be compared with values obtained from chi-square table.

OPTION 3

VAR = 1

| | | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| BMD | 67 | 0.0000E+00 | -0.2825E+02 | -0.3679E+00 |
| DATA1 | 24 | 0.2825E+02 | 0.0000E+00 | 0.2263E+02 |
| DATA2 | 9 | 0.3679E+00 | -0.2263E+02 | 0.0000E+00 |

t-statistic comparing variable 1 in BMD and DATA1.

VAR = 2

| | | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| BMD | 67 | 0.0000E+00 | -0.2018E+02 | -0.1061E+02 |
| DATA1 | 24 | 0.2018E+02 | 0.0000E+00 | 0.6710E+01 |
| DATA2 | 9 | 0.1061E+02 | -0.6710E+01 | 0.0000E+00 |

VAR = 3

| | | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| BMD | 67 | 0.0000E+00 | -0.4404E+02 | -0.5219E+00 |
| DATA1 | 24 | 0.4404E+02 | 0.0000E+00 | 0.1825E+02 |
| DATA2 | 9 | 0.5219E+00 | -0.1825E+02 | 0.0000E+00 |

VAR = 4

| | | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| BMD | 67 | 0.0000E+00 | -0.4255E+02 | -0.9113E+01 |
| DATA1 | 24 | 0.4255E+02 | 0.0000E+00 | 0.1303E+02 |
| DATA2 | 9 | 0.9113E+01 | -0.1303E+02 | 0.0000E+00 |

VAR = 5

| | | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| BMD | 67 | 0.0000E+00 | -0.3887E+01 | -0.1696E+02 |
| DATA1 | 24 | 0.3887E+01 | 0.0000E+00 | -0.1090E+02 |
| DATA2 | 9 | 0.1696E+02 | 0.1090E+02 | 0.0000E+00 |

OPTION 4

VAR = 1

| | | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| BMD | 67 | 0.1000E+01 | 0.3375E+01 | 0.1049E+01 |
| DATA1 | 24 | 0.2963E+00 | 0.1000E+01 | 0.3108E+00 |
| DATA2 | 9 | 0.9534E+00 | 0.3218E+01 | 0.1000E+01 |

VAR = 2

| | | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| BMD | 67 | 0.1000E+01 | 0.5901E-02 | 0.4204E+00 |
| DATA1 | 24 | 0.1695E+03 | 0.1000E+01 | 0.7125E+02 |
| DATA2 | 9 | 0.2379E+01 | 0.1404E-01 | 0.1000E+01 |

F-statistic comparing variable 2 in DATA2 and DATA1.

VAR = 3

| | | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| BMD | 67 | 0.1000E+01 | 0.8844E-01 | 0.2827E+01 |
| DATA1 | 24 | 0.1131E+02 | 0.1000E+01 | 0.3196E+02 |
| DATA2 | 9 | 0.3537E+00 | 0.3129E-01 | 0.1000E+01 |

VAR = 4

| | | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| BMD | 67 | 0.1000E+01 | 0.1102E+00 | 0.1902E+00 |
| DATA1 | 24 | 0.9073E+01 | 0.1000E+01 | 0.1726E+01 |
| DATA2 | 9 | 0.5257E+01 | 0.5794E+00 | 0.1000E+01 |

VAR = 5

| | | BMD 67 | DATA1 24 | DATA2 9 |
|---|---|---|---|---|
| BMD | 67 | 0.1000E+01 | 0.6754E+00 | 0.7704E+00 |
| DATA1 | 24 | 0.1481E+01 | 0.1000E+01 | 0.1141E+01 |
| DATA2 | 9 | 0.1298E+01 | 0.8766E+00 | 0.1000E+01 |

```
OPTION 5

 VAR        F-VALUE

   1     0.4252E+03
   2     0.2429E+03
   3     0.1142E+04   <---- F-statistic comparing variable 3 in all files of the analysis.
   4     0.9504E+03
   5     0.2280E+03

   V1  =     2  }
   V2  =   100  }  <---- Degrees of freedom parameters.
```

```
OPTION 6

 VAR    UNCORRECTED    CORRECTED

   1    0.1062E+02    0.9609E+01
   2    0.2417E+03    0.2187E+03
   3    0.7398E+02    0.6694E+02
   4    0.5215E+02    0.4719E+02
   5    0.1527E+01    0.1381E+01

   K-1  = 2
   C    = 0.1105E+01  <---- Correction factor
```

### 4.3 STPRG EXAMPLE

The following example illustrates the STPRG module of STATPAC including all options and plots. The previously presented BMD data is analyzed with variable 6 as the dependent variable. STPRG performs four steps and stops after including independent variables 1,3,4, and 5 in the regression model. Variable 2 was not entered into the model because the value of F (ENTER) never exceeded the value of F-IN (=0.5).

```
*PROG
STPRG
***WARNING - COMMON SIZE DIFFERS***     This message should be ignored.

*FILE                      User types the name of the file to be analyzed.

BMD
*VARS                      User types the variables to be considered for the model. The last
01                         variable listed is used as the dependent variable.
04
05
02
03
06

*FIN                       User types the F-value to determine entry.
.5
*FOUT                      User types the F-value to determine exit.
.3
*LIM                       User types the maximum number of iterations.
06
*TOL                       User requests the default tolerance (0.001).

*OPTS                      User requests all options.
1111
*COEF                      User supplies the test coefficients for option 2.
COF01=1.1
COF02=1.2
```

```
COF03=1.3
COF05=1.4
COF04=1.5
COF02=1.2678
                         User types simple carriage return.
*FCTR                    User supplies the value for t(N-p,1-½α).
2.0
*PLTS                    User requests all plots.
1111
*O.K.                    Processing will now begin.
```

REGRESSION OUTPUT (STPRG)   Stepwise regression

```
DATA FILE BMD <------------ Name of file (Question 1).
NO. OBS.        68
RESP.   6   SIXBM <-------- Subscript and name of dependent variable.
TOL.      0.00100 <-------- Response to Question 6.
F-IN      0.50000 }
F-OUT     0.30000 } <------ Responses to Questions 3 and 4.

 VARIABLE      CORR. X.VS.Y
 ---------------------------- Response to Question 2.
 1 ONEBM        0.082056
 2 TWOBM        0.749617
 3 TREBM        0.786058
 4 ORTBM        0.342568
 5 FIVBM        0.645167
 ---------------------------- Responses to Question 5 of INPUT.
```

```
STEP NO.   1

VAR. ENTERING      3   TREBM <---- Subscript and name of first variable in model
SEQ. F-TEST          106.724 <---- F-test to determine entry (106.724 > 0.5)
DEGREES OF FREEDOM        66
CHANGE IN R-SQ       0.617888 <--- Change in multiple correlation coefficient
R-SQ                 0.617888 <--- New multiple correlation coefficient
STD. ERR. Y          27.1238 <---- Standard error of dependent variable.

ANOVA <------------------------- Analysis of Variance
```

| SOURCE | D.F. | SUM OF SQUARES | MEAN SQUARE | OVERALL F |
|---|---|---|---|---|
| TOTAL | 67 | 0.127073E+06 | | |
| REGRS. | 1 | 0.785169E+05 | 0.785169E+05 | 0.106724E+03 |
| RESID. | 66 | 0.485562E+05 | 0.735700E+03 | |

VAR. IN REG.

| VARIABLE | COEFFICIENT | STD. ERROR | F (REMOVE) |
|---|---|---|---|
| 3 TREBM | 0.294426E+01 | 0.285000E+00 | 0.106724E+03 |

F-value needed to remove this variable from the present regression model

```
B0  =        26.0998 <--- Constant term of the model
```

VAR. NOT IN REG.

| VARIABLE | TOLERANCE | PARTIAL CORR. | F (ENTER) |
|---|---|---|---|
| 1 ONEBM | 0.999974 | 0.126216 | 0.105225E+01 |
| 2 TWOBM | 0.246590 | 0.219325 | 0.328473E+01 |
| 4 ORTBM | 0.984158 | 0.397285 | 0.121821E+02 |
| 5 FIVBM | 0.434591 | 0.132760 | 0.116621E+01 |
| | ↑ $a_{ii}$ | ↑ Partial correlation with y | ↑ F-values needed to enter these variables in the present regression model |

STEP NO. 2

```
VAR. ENTERING       4  ORTRBM
SEQ. F-TEST           12.1821
DEGREES OF FREEDOM       65
CHANGE IN R-SQ    0.060311
R-SQ              0.678199
STD. ERR. Y         25.0821
```

← Variable 4 enters regression model

ANOVA

| SOURCE | D.F. | SUM OF SQUARES | MEAN SQUARE | OVERALL F |
|---|---|---|---|---|
| TOTAL | 67 | 0.127073E+06 | | |
| REGRS. | 2 | 0.861808E+05 | 0.430904E+05 | 0.684940E+02 |
| RESID. | 65 | 0.408923E+05 | 0.629112E+03 | |

VAR. IN REG.

| VARIABLE | COEFFICIENT | STD. ERROR | F (REMOVE) |
|---|---|---|---|
| 3 TREBM | 0.282755E+01 | 0.265660E+00 | 0.113284E+03 |
| 4 ORTBM | 0.179768E+01 | 0.515052E+00 | 0.121821E+02 |

B0 = 21.7445

VAR. NOT IN REG.

| VARIABLE | TOLERANCE | PARTIAL CORR. | F (ENTER) |
|---|---|---|---|
| 1 ONEBM | 0.933987 | 0.027242 | 0.475306E-01 |
| 2 TWOBM | 0.246516 | 0.246529 | 0.414141E+01 |
| 5 FIVBM | 0.378439 | 0.321789 | 0.739256E+01 |

STEP NO. 3

```
VAR. ENTERING       5  FIVBM
SEQ. F-TEST           7.39256
DEGREES OF FREEDOM       64
CHANGE IN R-SQ    0.033322
R-SQ              0.711521
STD. ERR. Y         23.9328
```

← Variable 5 enters the regression model

ANOVA

| SOURCE | D.F. | SUM OF SQUARES | MEAN SQUARE | OVERALL F |
|---|---|---|---|---|
| TOTAL | 67 | 0.127073E+06 | | |
| REGRS. | 3 | 0.904151E+05 | 0.301384E+05 | 0.526176E+02 |
| RESID. | 64 | 0.366580E+05 | 0.572781E+03 | |

VAR. IN REG.

| VARIABLE | COEFFICIENT | STD. ERROR | F (REMOVE) |
|---|---|---|---|
| 3 TREBM | 0.195840E+01 | 0.407974E+00 | 0.230429E+02 |
| 4 QRTBM | 0.231239E+01 | 0.526652E+00 | 0.192786E+02 |
| 5 FIVBM | 0.103553E+01 | 0.380860E+00 | 0.739256E+01 |

B0 = 2.91072

VAR. NOT IN REG.

| VARIABLE | TOLERANCE | PARTIAL CORR. | F (ENTER) |
|---|---|---|---|
| 1 ONEBM | 0.883169 | 0.111114 | 0.787547E+00 |
| 2 TWOBM | 0.113447 | 0.015737 | 0.156066E-01 |

STEP NO. 4 (← **Last step - no more variables to be entered or removed**)

VAR. ENTERING 1 ONEBM (← **Variable 1 enters regression model**)
SEQ. F-TEST 0.787547
DEGREES OF FREEDOM 63
CHANGE IN R-SQ 0.003562
R-SQ 0.715082
STD. ERR. Y 23.9727

ANOVA

| SOURCE | D.F. | SUM OF SQUARES | MEAN SQUARE | OVERALL F |
|---|---|---|---|---|
| TOTAL | 67 | 0.127073E+06 | | |
| REGRS. | 4 | 0.908677E+05 | 0.227169E+05 | 0.395291E+02 |
| RESID. | 63 | 0.362054E+05 | 0.574689E+03 | |

VAR. IN REG.

**All values are greater than F-OUT (0.3)**

| VARIABLE | COEFFICIENT | STD. ERROR | F (REMOVE) |
|---|---|---|---|
| 1 ONEBM | 0.427208E+00 | 0.481394E+00 | 0.787547E+00 |
| 3 TREBM | 0.189677E+01 | 0.414512E+00 | 0.209389E+02 |
| 4 QRTBM | 0.223335E+01 | 0.534995E+00 | 0.174266E+02 |
| 5 FIVBM | 0.111674E+01 | 0.392316E+00 | 0.810276E+01 |

B0 = -1.25284

**Less than F-IN (0.5)**

VAR. NOT IN REG.

| VARIABLE | TOLERANCE | PARTIAL CORR. | F (ENTER) |
|---|---|---|---|
| 2 TWOBM | 0.102908 | 0.052408 | 0.170755E+00 |

RESIDUAL OUTPUT

| NO. | OBS | PRED | RESID |
|---|---|---|---|
| 1 | 0.640000E+02 | 0.885535E+02 | -0.245535E+02 |
| 2 | 0.650000E+02 | 0.862786E+02 | -0.212786E+02 |
| 3 | 0.820000E+02 | 0.887180E+02 | -0.671796E+01 |
| 4 | 0.230000E+02 | 0.227317E+02 | 0.268293E+00 |
| 5 | 0.640000E+02 | 0.849735E+02 | -0.209735E+02 |
| 6 | 0.160000E+02 | 0.226273E+02 | -0.662731E+01 |
| 7 | 0.120000E+02 | 0.292135E+02 | -0.172135E+02 |
| 8 | 0.270000E+02 | 0.262787E+02 | 0.721302E+00 |
| 9 | 0.480000E+02 | 0.289919E+02 | 0.190081E+02 |
| 10 | 0.500000E+02 | 0.418836E+02 | 0.811641E+01 |
| 11 | 0.120000E+02 | 0.215410E+02 | -0.954102E+01 |
| 12 | 0.130000E+02 | 0.353019E+02 | -0.223019E+02 |
| 13 | 0.200000E+02 | 0.320935E+02 | -0.120935E+02 |
| 14 | 0.230000E+02 | 0.271830E+02 | -0.418304E+01 |
| 15 | 0.118000E+03 | 0.947300E+02 | 0.232700E+02 |
| 16 | 0.500000E+02 | 0.503519E+02 | -0.351914E+00 |
| 17 | 0.630000E+02 | 0.564708E+02 | 0.652920E+01 |
| 18 | 0.150000E+03 | 0.829056E+02 | 0.670944E+02 |
| 19 | 0.720000E+02 | 0.200252E+02 | 0.519748E+02 |
| 20 | 0.540000E+02 | 0.420349E+02 | 0.119651E+02 |
| 21 | 0.109000E+03 | 0.781895E+02 | 0.308105E+02 |
| 22 | 0.100000E+02 | 0.237146E+02 | -0.137146E+02 |
| 23 | 0.125000E+03 | 0.121368E+03 | 0.363178E+01 |
| 24 | 0.440000E+02 | 0.282155E+02 | 0.157845E+02 |
| 25 | 0.480000E+02 | 0.439175E+02 | 0.408250E+01 |
| 26 | 0.105000E+03 | 0.119610E+03 | -0.146104E+02 |
| 27 | 0.900000E+01 | 0.258037E+02 | -0.168037E+02 |
| 28 | 0.130000E+03 | 0.103836E+03 | 0.261638E+02 |
| 29 | 0.160000E+03 | 0.124185E+03 | 0.358153E+02 |
| 30 | 0.480000E+02 | 0.499287E+02 | -0.192869E+01 |
| 31 | 0.360000E+02 | 0.248910E+02 | 0.111090E+02 |
| 32 | 0.150000E+03 | 0.883400E+02 | 0.616600E+02 |
| 33 | 0.780000E+02 | 0.312197E+02 | 0.467803E+02 |
| 34 | 0.230000E+02 | 0.251065E+02 | -0.210652E+01 |
| 35 | 0.420000E+02 | 0.258751E+02 | 0.161249E+02 |
| 36 | 0.720000E+02 | 0.785157E+02 | -0.651573E+01 |
| 37 | 0.200000E+02 | 0.265577E+02 | -0.655773E+01 |
| 38 | 0.360000E+02 | 0.339123E+02 | 0.208769E+01 |
| 39 | 0.560000E+02 | 0.467423E+02 | 0.925773E+01 |
| 40 | 0.360000E+02 | 0.410371E+02 | -0.503711E+01 |
| 41 | 0.260000E+02 | 0.391791E+02 | -0.131791E+02 |
| 42 | 0.108000E+03 | 0.109324E+03 | -0.132396E+01 |
| 43 | 0.106000E+03 | 0.113005E+03 | -0.700476E+01 |
| 44 | 0.160000E+02 | 0.231520E+02 | -0.715201E+01 |
| 45 | 0.104000E+03 | 0.119008E+03 | -0.150084E+02 |
| 46 | 0.470000E+02 | 0.476452E+02 | -0.645188E+00 |
| 47 | 0.270000E+02 | 0.228350E+02 | 0.416496E+01 |
| 48 | 0.120000E+02 | 0.412560E+02 | -0.292560E+02 |
| 49 | 0.700000E+01 | 0.217236E+02 | -0.147236E+02 |
| 50 | 0.180000E+02 | 0.302658E+02 | -0.122658E+02 |
| 51 | 0.280000E+02 | 0.369663E+02 | -0.896633E+01 |
| 52 | 0.250000E+02 | 0.308700E+02 | -0.586995E+01 |
| 53 | 0.110000E+02 | 0.243047E+02 | -0.133047E+02 |
| 54 | 0.200000E+02 | 0.396499E+02 | -0.196499E+02 |
| 55 | 0.140000E+02 | 0.317013E+02 | -0.177013E+02 |
| 56 | 0.380000E+02 | 0.614628E+02 | -0.234628E+02 |

| | | | |
|---|---|---|---|
| 57 | 0.103000E+03 | 0.531120E+02 | 0.498880E+02 |
| 58 | 0.106000E+03 | 0.899314E+02 | 0.160686E+02 |
| 59 | 0.630000E+02 | 0.998496E+02 | -0.368496E+02 |
| 60 | 0.208000E+03 | 0.201456E+03 | 0.654361E+01 |
| 61 | 0.320000E+02 | 0.671833E+02 | -0.351833E+02 |
| 62 | 0.280000E+02 | 0.415379E+02 | -0.135379E+02 |
| 63 | 0.320000E+02 | 0.420629E+02 | -0.100629E+02 |
| 64 | 0.100000E+03 | 0.888450E+02 | 0.111550E+02 |
| 65 | 0.500000E+02 | 0.428304E+02 | 0.716963E+01 |
| 66 | 0.800000E+02 | 0.519009E+02 | 0.280991E+02 |
| 67 | 0.650000E+02 | 0.134051E+03 | -0.690510E+02 |
| 68 | 0.250000E+02 | 0.330358E+02 | -0.803584E+01 |

OPTION 1

| VARIABLE | T-VALUE |
|---|---|
| 1 ONEBM | 0.887438E+00 |
| 3 TREBM | 0.457590E+01 |
| 4 QRTBM | 0.417451E+01 |
| 5 FIVBM | 0.284654E+01 |

**Note that t-statistics are computed only for variables included in the model**

OPTION 2

**Response to Question 8** (→ USER COEFF.)

| VARIABLE | USER COEFF. | T-VALUE |
|---|---|---|
| 1 ONEBM | 0.110000E+01 | -0.139759E+01 |
| 3 TREBM | 0.130000E+01 | 0.143968E+01 |
| 4 QRTBM | 0.150000E+01 | 0.137075E+01 |
| 5 FIVBM | 0.140000E+01 | -0.722017E+00 |

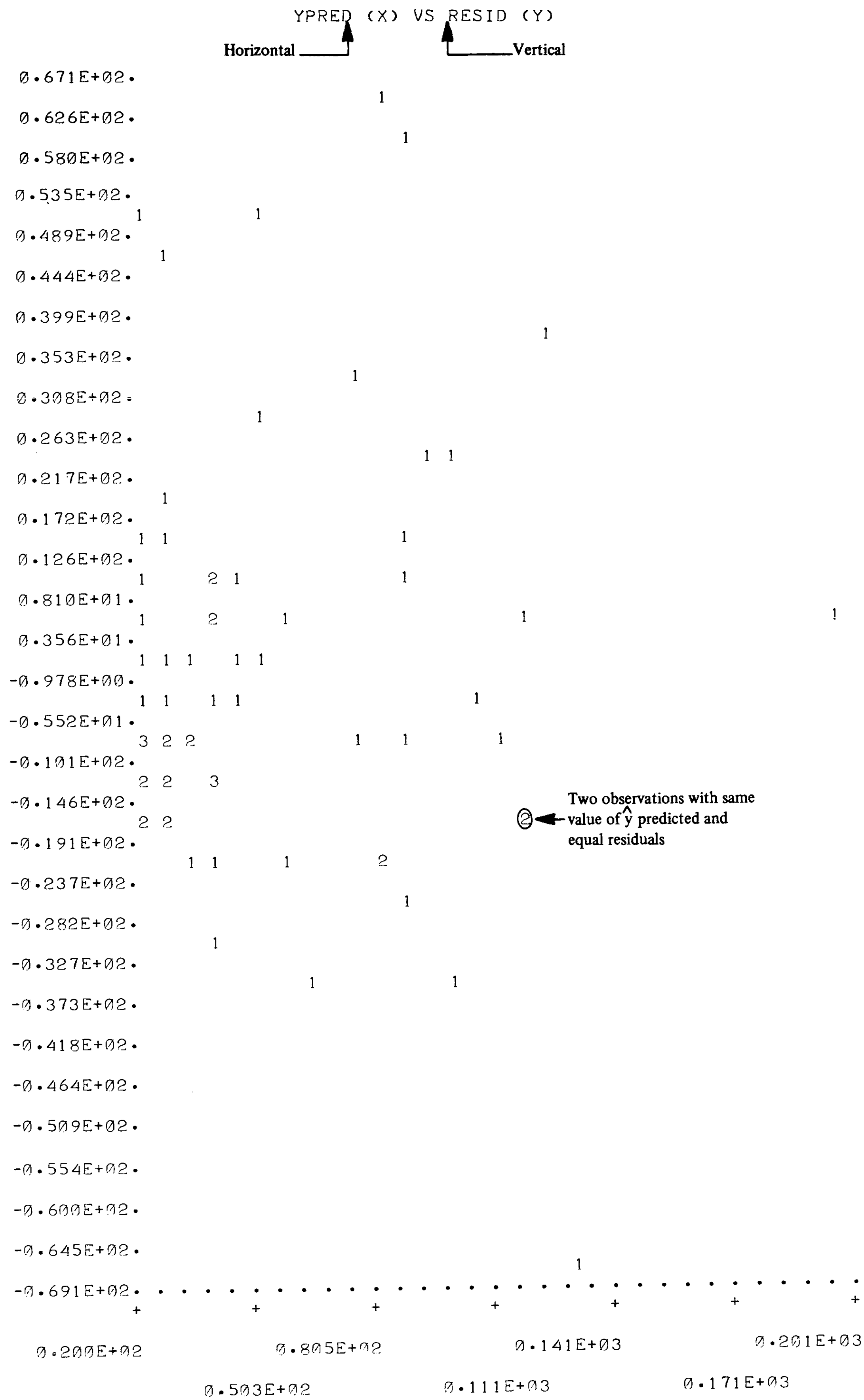
OPTION 3

T( N-P , 1-ALPHA/2 ) = 2.00000 ← **Response to Question 9 (FCTR)**

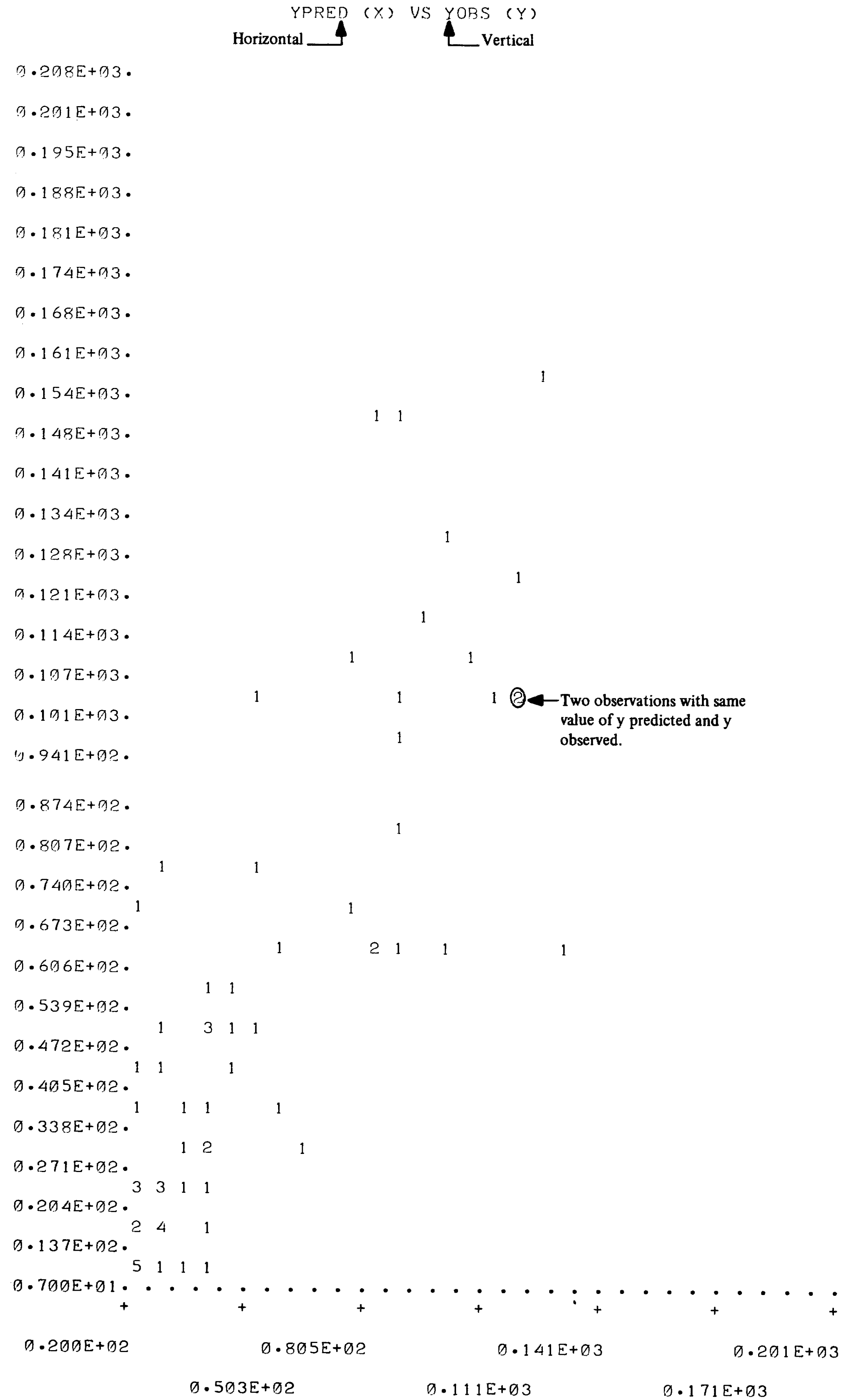
| VARIABLE | LOWER BOUND | UPPER BOUND |
|---|---|---|
| 1 ONEBM | -0.535581E+00 | 0.139000E+01 |
| 3 TREBM | 0.106774E+01 | 0.272579E+01 |
| 4 QRTBM | 0.116335E+01 | 0.330334E+01 |
| 5 FIVBM | 0.332109E+00 | 0.190137E+01 |

N-P = 63 ← **Degrees of freedom**

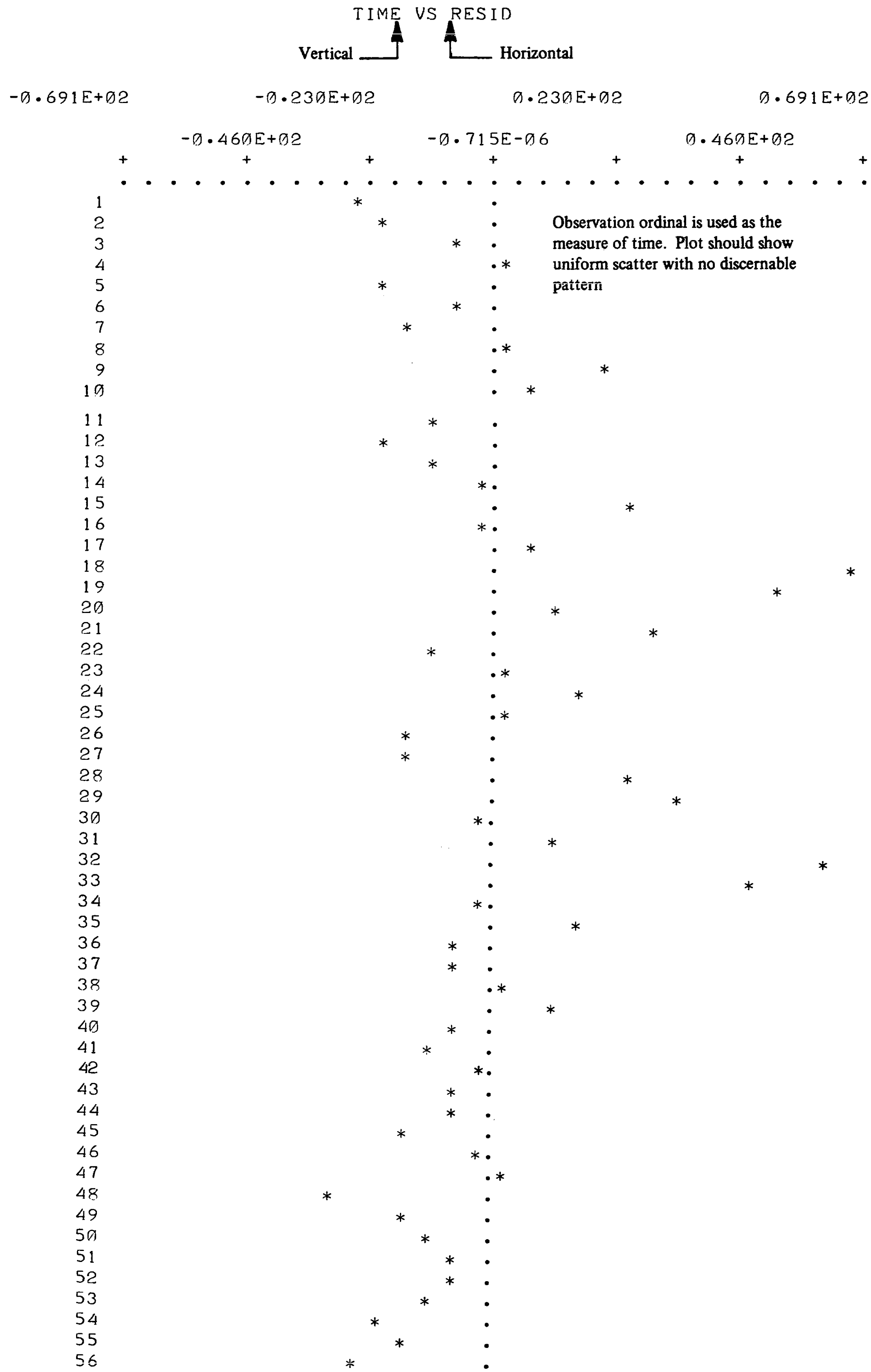
PLOT 1

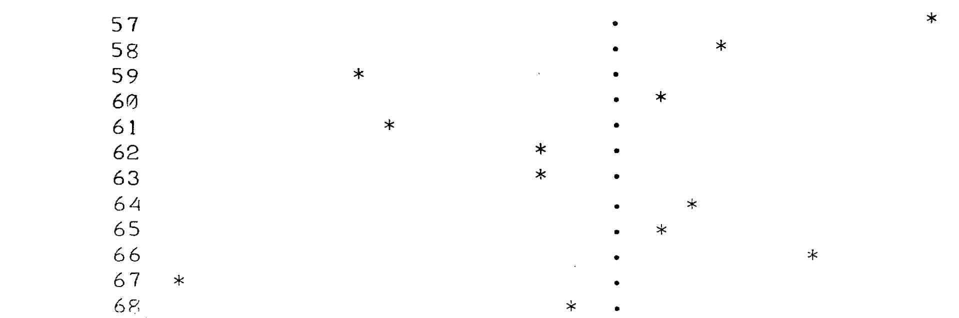

YPRED (X) VS RESID (Y)
Horizontal
Vertical
0.671E+02
0.626E+02
0.580E+02
0.535E+02
0.489E+02
0.444E+02
0.399E+02
0.353E+02
0.308E+02
0.263E+02
0.217E+02
0.172E+02
0.126E+02
0.810E+01
0.356E+01
-0.978E+00
-0.552E+01
-0.101E+02
-0.146E+02
-0.191E+02
-0.237E+02
-0.282E+02
-0.327E+02
-0.373E+02
-0.418E+02
-0.464E+02
-0.509E+02
-0.554E+02
-0.600E+02
-0.645E+02
-0.691E+02
0.200E+02
0.503E+02
0.805E+02
0.111E+03
0.141E+03
0.171E+03
0.201E+03
Two observations with same value of ŷ predicted and equal residuals

PLOT 2


YPRED (X) VS YOBS (Y)
Horizontal
Vertical
0.208E+03
0.201E+03
0.195E+03
0.188E+03
0.181E+03
0.174E+03
0.168E+03
0.161E+03
0.154E+03
0.148E+03
0.141E+03
0.134E+03
0.128E+03
0.121E+03
0.114E+03
0.107E+03
0.101E+03
0.941E+02
0.874E+02
0.807E+02
0.740E+02
0.673E+02
0.606E+02
0.539E+02
0.472E+02
0.405E+02
0.338E+02
0.271E+02
0.204E+02
0.137E+02
0.700E+01
Two observations with same value of y predicted and y observed.
0.200E+02
0.805E+02
0.141E+03
0.201E+03
0.503E+02
0.111E+03
0.171E+03

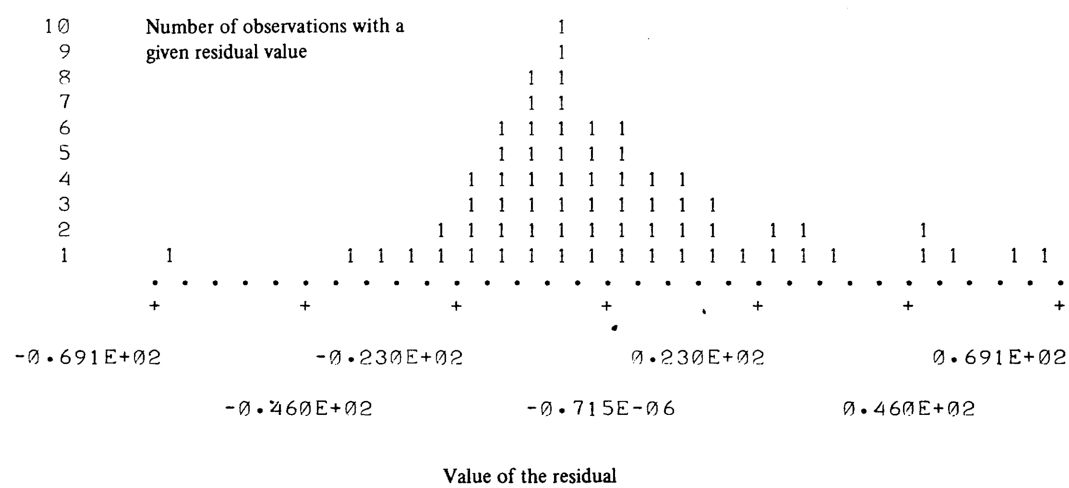
PLOT 3


TIME VS RESID
Vertical
Horizontal
-0.691E+02
-0.230E+02
0.230E+02
0.691E+02
-0.460E+02
-0.715E-06
0.460E+02
Observation ordinal is used as the measure of time. Plot should show uniform scatter with no discernable pattern




PLOT 4

RESID HISTOGRAM




## 4.4 MLTRG EXAMPLE

The following sample output is the result of analysis of the DATA1 file with the MLTRG module of STATPAC. The output does not include options and plots, as the MLTRG options and plots are the same as those of STPRG previously presented.

```
*PROG

MLTRG
***WARNING - COMMON SIZE DIFFERS***    This message should be ignored

*FILE

DATA1
*VARS
06                        Last variable typed is the dependent variable (all variables in the file
                          will be entered in the regression model)
*TOL
.00103
*OPTS
1001                      Only options 1 and 4 are requested
*PLTS
0111                      Only plots 2, 3 and 4 are requested
*O.K.
```

REGRESSION OUTPUT (MLTRG)

```
DATA FILE DATA1
NO. OBS.       25
RESP.   6   SIX01  <------ Dependent variables
TOL.      0.00103

 VARIABLE      CORR. X.VS.Y

  1 ONE01        0.068706 )
  2 TWO01        0.468219 |
  3 TRE01        0.111343 >  All independent variables are included
  4 QRT01        0.308764 |  in the model
  5 FIV01        0.290697 )


R-SQ               0.494342
STD. ERR. Y         1.15539

ANOVA

  SOURCE     D.F.  SUM OF SQUARES      MEAN SQUARE        OVERALL F

  TOTAL       24     0.501600E+02
  REGRS.       5     0.247962E+02    0.495924E+01      0.371496E+01
  RESID.      19     0.253638E+02    0.133494E+01

VAR. IN REG.

    VARIABLE       COEFFICIENT        STD. ERROR

     1 ONE01      0.430884E-02     0.724055E-01
     2 TWO01      0.722737E-02     0.212183E-02
     3 TRE01      0.158192E-01     0.728874E-02
     4 QRT01      0.148276E-01     0.444280E-01
     5 FIV01      0.339402E-01     0.505993E-01

B0 =      -6.71895
```

## APPENDIX A
## DESCRIPTIVE STATISTICS ALGORITHMS

$\overline{X}_{ji}$: *mean* of the $j^{th}$ variable in the $i^{th}$ file.

$$\overline{X}_{ji} = \left( \sum_{m=1}^{N_i} X_{jim} \right) / N_i$$

$\sigma_{ji}^2$: *variance* of the $j^{th}$ variable in the $i^{th}$ file.

$$\sigma_{ji}^2 = \left( \sum_{m=1}^{N_i} (X_{jim} - \overline{X}_{ji})^2 \right) / (N_i - 1)$$

$\sigma_{ji}$: *standard deviation* of the $j^{th}$ variable in the $i^{th}$ file.

$$\sigma_{ji} = \sqrt{\sigma_{ji}^2}$$

$S.E._{ji}$: *standard error of the mean* of the $j^{th}$ variable in the $i^{th}$ file.

$$S.E._{ji} = \sigma_{ji} / \sqrt{N_i}$$

$SKEWNESS_{ji}$: *coefficient of skewness* of the $j^{th}$ variable in the $i^{th}$ file.

$$SKEWNESS_{ji} = \frac{\left( \sum_{m=1}^{N_i} (X_{jim} - \overline{X}_{ji})^3 \right) / N_i}{\sigma_{ji}^3}$$

$KURTOSIS_{ji}$: *coefficient of kurtosis* of the $j^{th}$ variable in the $i^{th}$ file.

$$KURTOSIS_{ji} = \frac{\left( \sum_{m=1}^{N_i} (X_{jim} - \overline{X}_{ji})^4 \right) / N_i}{\sigma_{ji}^4} - 3$$

$C_{rs}$: *simple correlation coefficient* between the $r^{th}$ and $s^{th}$ variable in the $i^{th}$ file.

$$C_{rs} = \frac{\sum_{m=1}^{N_i} (X_{rim} - \overline{X}_{ri})(X_{sim} - \overline{X}_{si})}{\left[ \left\{ \sum_{m=1}^{N_i} (X_{rim} - \overline{X}_{ri})^2 \right\} \left\{ \sum_{m=1}^{N_i} (X_{sim} - \overline{X}_{si})^2 \right\} \right]^{1/2}}$$

## APPENDIX B
## REGRESSION ANALYSIS ALGORITHMS

corr $(x_i,y)$: correlation of the $i^{th}$ dependent variable with the independent variable.

$$\text{corr}(x_i,y) = \frac{\sum_{m=1}^{N} (X_{im} - \overline{X}_i)(y_m - \overline{y})}{\sqrt{\sum_{m=1}^{N} (X_{im} - \overline{X}_i)^2}\sqrt{\sum_{m=1}^{N} (y_m - \overline{y})^2}}$$

$r^2$: multiple correlation.

$$r^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2} = 1 - a_{nn}$$

$s_y$: standard error of y.

$$s_y = \sqrt{\sum_{m=1}^{N} (y_m - \overline{y})^2} \quad \sqrt{a_{nn} / (N-1-p)}$$

Sequential F-test (Entering).

$$F = \left\{V_{max}\,(\phi - 1)\right\} / (a_{nn} - V_{max})$$

where $V_{max}$ = maximum $V_i$, $V_i = a_{in}\,a_{ni}/a_{ii}$

$a_{ij}$ = elements of the correlation matrix

n = total number of variables being analyzed

$\phi$ = degrees of freedom (N-1-p)

Sequential F-test (Leaving)

$$F = \left\{(|V_{min}|) \cdot \phi\right\} / a_{nn}$$

where $V_{min}$ = minimum $V_i$, $V_i = a_{in}\,a_{ni}/a_{ii}$

and other symbols as above.

**ANOVA Analysis of Variance Table**

| | | |
|---|---|---|
| D.F. (Degrees of Freedom) | Total | N - 1 |
| | Regression | p |
| | Residual | N - 1 - p |
| Sum of Squares | Total | $SS_y = \sum_{m=1}^{N} (y_m - \bar{y})^2$ |
| | Regression | $SS_{reg} = SS_y\,(1 - a_{nn})$ |
| | Residual | $SS_{resid} = SS_y - SS_{reg}$ |
| Mean Square | Regression | $ms_{reg} = SS_{reg}/p$ |
| | Residual | $ms_{resid} = SS_{resid}/(N\text{-}1\text{-}p)$ |
| Overall F | Regression | $ms_{reg}/ms_{resid}$ |

**Table of Variables in Regression**

$b_i$: coefficient of the $i^{th}$ variable.

$$b_i = b_{in}\frac{\sigma_n}{\sigma_i}$$

where $b_{in}$ = $i^{th}$ element of last ($n^{th}$) column of inverted correlation matrix.

$$\sigma_n = \sqrt{\sum_{m=1}^{N} (y_m - \bar{y})^2} \qquad \sigma_i = \sqrt{\sum_{m=1}^{N} (X_{im} - \bar{X}_i)^2}$$

$s_i$: standard error of $b_i$.

$$s_i = \frac{s_y}{\sigma_i} \sqrt{b_{ii}}$$

where $s_y$ is standard error of y;

$\sigma_i$ is defined as above;

$b_{ii}$ is diagonal element of the correlation matrix.

$F_i$: F-test to remove the $i^{th}$ variable.

$$F_i = \left[ \frac{b_i}{s_i} \right]^2$$

$b_0$: constant term of model.

$$b_0 = \bar{y} - \sum_{i=1}^{T} b_i \bar{X}_i$$

where T = number of variables in regression

**Table of Variables not in Regression**

$tol_i$ : tolerance of $i^{th}$ variable.

$tol_i = a_{ii}$ (the $i^{th}$ diagonal element of the inverted correlation matrix)

Partial correlation of the $i^{th}$ variable.

$$\text{part. corr.}_i = a_{in} / \sqrt{a_{ii}\, a_{nn}}$$

$F_i$: F-test to enter the $i^{th}$ variable.

$$F_i = \left\{ a_{in}^2 (\phi - 1) \right\} / (a_{ii}\, a_{nn} - a_{in})^2$$

where $\phi = N - 1 - p$

**Residual Output**

$\hat{y}_i$: predicted value of the dependent variable for the $i^{th}$ observation.

$$\hat{y}_i = b_0 + \sum_{j=i}^{T} b_j X_{ji}$$

where T = number of variables in regression.

$e_i$: residual for the $i^{th}$ observation.

$$e_i = y_i - \hat{y}_i$$

## BIBLIOGRAPHY

The following PDP-15 and PDP-9 manuals describe the software and operating procedures used by STATPAC:


PDP-15/20 System

*PDP-15/20 Advanced Monitor Software System*
Digital Equipment Corporation, Maynard, Mass.;
Order No. DEC-15-MR2A-D.

*PDP-15/20 User's Guide*
Digital Equipment Corporation, Maynard, Mass.;
Order No. DEC-15-MG2A-D.

PDP-15/30 System

*PDP-15/30 Background/Foreground Monitor Software System*
Digital Equipment Corporation, Maynard, Mass.;
Order No. DEC-15-MR3A-D.

*PDP-15/30 User's Guide*
Digital Equipment Corporation, Maynard, Mass.;
Order No. DEC-15-MG3A-D.

PDP-15/40 System

*PDP-15/40 Disk Oriented Background/Foreground Monitor Software System*
Digital Equipment Corporation, Maynard, Mass.;
Order No. DEC-15-MR4A-D.

*PDP-15/40 User's Guide*
Digital Equipment Corporation, Maynard, Mass.;
Order No. DEC-15-MG4A-D.

PDP-9 Advanced Software System

*Keyboard Monitor Guide*
PDP-9 ADVANCED Software System;
Digital Equipment Corporation, Maynard, Mass.;
Order No. DEC-9A-NGBA-D.

*PDP-9 ADVANCED Software System Monitors*
Digital Equipment Corporation, Maynard, Mass.;
Order No. DEC-9A-MADO-D.

FORTRAN Manuals

*PDP-15 FORTRAN IV Manual*
Digital Equipment Corporation, Maynard, Mass.;
Order No. DEC-15-KFZA-D.

*FORTRAN IV Programmer's Reference Manual*
PDP-9 ADVANCED Software System;
Digital Equipment Corporation, Maynard, Mass.;
Order No. DEC-9A-KFZA-D.

The following statistics texts contain useful information relating to the descriptive statistics module:

*Quality Control and Industrial Statistics* (Revised edition)
by Acheson J. Duncan, Ph.D.; Richard D. Erwin, Inc., 1959; Chapter 4.

*Statistics in Research* (Second edition),
by Bernard Ostle; The Iowa State University Press, Ames, Iowa; 1963; Chapter 7.

The following statistics texts contain useful information relating to the regression analysis modules:

*Applied Regression Analysis*
by N. K. Draper and H. Smith; John Wiley & Sons, Inc.; 1968; Chapters 1, 2, 3, 4, and 6.

*BMD Biomedical Computer Programs* (Second edition)
edited by W. J. Dixon; University of California Press, Berkeley, California; 1967; Pages 233-257.

*"Multiple Regression Analysis"*
by M. A. Efroymson, contained in *Mathematical methods for Digital Computers* (Volume 1); edited by Anthony Ralston, PH.D., and Herbert S. Wilf, PH.D.; John Wiley & Sons, Inc.; 1967; Chapter 17.

# INDEX


A

Adding Processing Modules To Statpac, 3-7
  File format, 3-8

Algorithms
  Descriptive statistics, A-1
  Regression analysis, B-1

B

Bartlett's test, 1-10

Building an executable file, 3-2

C

CHAIN, 3-3

COEF, 2-7

Command dialogue, 2-1, 2-2, 2-3, 2-5
  COEF, 2-7
  FCTR, 2-7
  FILE, 2-2, 2-3, 2-6
  FILE (NEW), 2-2
  FILE (OLD), 2-2, 2-3
  FIN, 2-6
  FORMAT, 2-2
  FOUT, 2-6
  LIM, 2-6, 2-8
  NO. OBS, 2-2
  O.K, 2-3, 2-5, 2-7
  OPTS, 2-4, 2-6
  PLTS, 2-7
  TOL, 2-6
  VARS, 2-2, 2-4, 2-6
  VRNC, 2-4

Control Module, 1-1, 2-1
  EXIT, 1-1
  PROG, 1-1, 2-1

Correlation matrix, 1-5, 2-5

D

Descriptive statistics, 1-4
  Algorithms, A-1
  Dialogue, 2-3

Device handlers, 3-1

E

Error messages, 2-1, 2-3, 2-5, 2-7
  ERR1, 2-5, 2-7
  ERR2, 2-5, 2-7
  ERR3, 2-8
  ERR4, 2-8
  ERR5, 2-8
  ERROR, 2-1
  IGNORED, 2-3

Example,
  INPUT, 4-2
  MLTRG, 4-18
  SMMRY, 4-4
  STPRG, 4-9

EXIT, 1-1

F

FCTR, 2-7

FILE, 2-2, 2-3, 2-6

FILE (NEW), 2-2

FILE (OLD), 2-2, 2-3

File format, 1-3, 3-2

File name, 2-3, 3-2

FIN, 2-6

FORMAT, 2-2

FOUT, 2-6

H

Hyperplane, 1-11

Hypothesis test, 1-6, 1-13

I

IGNORED, 2-3

Implementing and Augmenting STATPAC, 3-1
- Device handlers, 3-1
- Executable file, 3-1
- Logical units, 3-1

INPUT module, 1-1, 1-3,
- Dialogue, 2-2
- Example, 4-2
- Observation, 1-3
- Standardized binary data file, 1-3
- Variables, 1-3

K

Kurtosis, 1-5

L

LIM, 2-6, 2-8

Logical units, 3-1

M

Maximum, 1-5

Mean, 1-4

MEAN, 2-7

Minimum, 1-5

MLTRG module, 1-1, 1-10
- Dialogue, 2-5
- Example, 4-18
- Output, 1-12

Model, 1-11

Module operating procedures, 2-1
- CONTROL command dialogue, 2-1
- INPUT command dialogue, 2-2
- RUBOUT, 2-1
- SMMRY command dialogue, 2-3
- STPRG & MLTRG command dialogue, 2-5

Modules, 1-1
- CONTROL, 1-1, 2-1
- INPUT, 1-1, 1-3, 2-2, 4-2
- SMMRY, 1-1, 1-4, 2-3, 4-4
- STPRG, 1-1, 1-10, 2-5, 4-9
- MLTRG, 1-1, 1-10, 2-5, 4-18

Multiple Linear Regression Module,
(see MLTRG module)

N

NO. OBS, 2-2

O

Observation, 1-3

O.K, 2-3, 2-5, 2-7

Output, Regression Analysis, 1-12

P

Predicted $\hat{Y}$ Versus Observed, 1-14

Predicted $\hat{Y}$ Versus Residual, 1-14

PLTS, 2-7

PROG, 1-1, 2-1

R


Range, 1-5

Regression Analysis, 1-11
  Algorithms, B-1
  Dialogue, 2-5
  Hyperplane, 1-11
  Model, 1-11
  Output, 1-12
  Residuals, 1-11

Regression Module Operation, 2-5

Regression Options, 1-13

Regression plots, 1-14
  Overall plot of Residuals, 1-14, 4-18
  Predicted $\hat{Y}$ Versus Observed Y, 1-14, 4-16
  Predicted $\hat{Y}$ Versus Residual, 1-14, 4-15
  Residuals, Versus Ordinal, 1-14, 4-17

Residuals, 1-11, 1-14


S


Sample Operation, 4-1

Skewness, 1-5

SMMRY module, 1-1, 1-4
  Dialogue, 2-3
  Example, 4-4

SMMRY Options, 1-6
  Hypothesis test, 1-6

SMMRY statistics, 1-4
  Correlation matrix, 1-5
  Kurtosis, 1-5
  Maximum, 1-5
  Mean, 1-4
  Minimum, 1-5
  Range, 1-5
  Skewness, 1-5
  Standard Deviation, 1-4
  Variance, 1-4

Standard deviation, 1-4

Standardized binary data file, 1-3

Stepwise Linear Regression Module,
  (see STPRG module)

STP, 2-3

STPRG module, 1-1, 1-10
  Dialogue, 2-5
  Example, 4-9
  Output, 1-12
  Symbol definitions, 1-6


T


TOL, for SMMRY, 2-5

TOL, for regression analysis, 2-6


V


Variables, 1-3

Variance, 1-4

VARS, 2-2, 2-4, 2-6


W


WARNING, 4-2


Z


ZERO OBS, 2-2, 2-3

## HOW TO OBTAIN SOFTWARE INFORMATION

Announcements for new and revised software, as well as programming notes, software problems, and documentation corrections are published monthly by Software Information Service in the following newsletters.

Digital Software News for the PDP-8 Family
Digital Software News for the PDP-9 Family
Digital Software News for the PDP-15 Family

These newsletters contain information applicable to software available from Digital's Program Library (see title page for address). Software products and documents are usually shipped only after the Program Library receives a specific request from a user.

Digital Equipment Computer Users Society (DECUS) maintains a user library and publishes a catalog of programs as well as the DECUSCOPE magazine for its members and non-members who request it.

Please complete the card below to receive information on DECUS membership or to place your name on the newsletter mailing list.

Please send

☐ DECUS membership information,

or add my name to the

☐ DECUSCOPE non-membership list.

And, send me the Digital Software News for the

☐ PDP-8 ☐ PDP-9 ☐ PDP-15

NAME ______________________________

COMPANY ______________________________

ADDRESS ______________________________

CITY ____________ STATE __________ ZIP ______

Fold Here

Do Not Tear - Fold Here and Staple

FIRST CLASS
PERMIT NO. 33
MAYNARD, MASS.

BUSINESS REPLY MAIL
NO POSTAGE STAMP NECESSARY IF MAILED IN THE UNITED STATES

Postage will be paid by:

digital

**DECUS**
**Digital Equipment Corporation**
**146 Main Street**
**Maynard, Mass. 01754**

## READER'S COMMENTS

Digital Equipment Corporation maintains a continuous effort to improve the quality and usefulness of its publications. To do this effectively we need user feedback – your critical evaluation of this manual.

Please comment on this manual's completeness, accuracy, organization, usability, and readability.

Did you find errors in this manual?

How can this manual be improved?

DEC also strives to keep its customers informed of current DEC software and publications. Thus, the following periodically distributed publications are available upon request. Please check the appropriate box(s) for a current issue of the publication(s) desired.

- ☐ PDP-15 Software Manual Update, a quarterly collection of revisions to current software manuals.
- ☐ PDP-15 User's Bookshelf, a bibliography of current software manuals.
- ☐ Program Library Price List, a list of currently available PDP-15 software programs and manuals.

Please describe your position.

Name Organization

Street Department

City State Zip or Country

Fold Here

Do Not Tear - Fold Here and Staple

FIRST CLASS
PERMIT NO. 33
MAYNARD, MASS.

BUSINESS REPLY MAIL
NO POSTAGE STAMP NECESSARY IF MAILED IN THE UNITED STATES

Postage will be paid by:

digital

Digital Equipment Corporation
Software Information Services
146 Main Street, Bldg. 3-5
Maynard, Massachusetts 01754

**Digital Equipment Corporation**
**Maynard, Massachusetts**

printed in U.S.A.