# THE SCA STATISTICAL SYSTEM

## REFERENCE MANUAL FOR GENERAL STATISTICAL ANALYSIS


**by**

**Gregory B. Hudak**

**Lon-Mu Liu**

**in collaboration with**

**George E. P. Box**

**Mervin E. Muller**

**George C. Tiao**



**This manual is published by**

**Scientific Computing Associates® Corp.**

**913 West Van Buren Street, Suite 3H**

**Chicago, Illinois 60607-3528**

**U.S.A.**

**April 2002**

# PREFACE

This edition of the SCA manual for general statistical analysis encompasses the current capabilities of the SCA-GSA product. It replaces all previous versions related to the GSA product both in scope and style of presentation.

The manual is written for novices to statistics and to the SCA System, as well as more experienced statisticians or SCA users. Tools for data analysis, from data plotting through specific statistical methods can be found in Chapter 3 through 12. A listing of the organization of various capabilities is presented in Chapter 1. Chapter 2 provides basic information on the use of the SCA System. Chapters 3 through 12 may be used in any order, as each concentrates on those capabilities most directly related to it. Cross-references are provided whenever necessary, although the use of capabilities that are explained in other chapters is usually self-evident.

Material in this manual is presented in a "data analysis" form. That is, SCA System capabilities, commands, and output are usually presented within the context of a data analysis. Examples have been chosen from a variety of sources in order to both demonstrate the use of the SCA System, and to provide some guidelines for a statistical analysis. For this reason, this manual could also be used as a supplementary text in an introductory statistics course.

Within a chapter, information regarding specific capabilities and features of the SCA System are presented from those most frequently used to those that are less commonly employed. All detailed information regarding the command structure of the SCA System is presented at the end of the chapter.

As features are added to the SCA-GSA product, this manual will be supplemented with other documents. These documents will be provided to users of this manual.

## Acknowledgments

The development of the capabilities comprising the SCA-GSA product began in 1981 as part of the SCA System for forecasting and time series analysis. The product was enhanced and separated as a self-contained package in 1984. Current and future developments are keyed to the applications of our user community.

The SCA-GSA product was designed and developed by Lon-Mu Liu with the assistance of the SCA programming staff. In particular, we wish to thank Philip Burns for the development of the cross tabulation and nonparametric test capabilities, and Houston Stokes of The University of Illinois at Chicago for his programming work related to dispersion plots. We are grateful to Alan Montgomery and Ki-Kan Chan for their programming and testing efforts.

We are indebted to Professor George Tiao of The University of Chicago for his suggestions to improve the regression, time series and ANOVA capabilities. We thank Professors George Box, Søren Bisgaard, and Conrad Fung of The Center for Quality and Productivity Improvement at the University of Wisconsin-Madison, and Professor Mervin Muller of The Ohio State University for their valuable comments and suggestions.

This manual was prepared by Gregory Hudak, Scientific Computing Associates, and Lon-Mu Liu, The University of Illinois at Chicago. Chapters were entered and edited by Ching-Te Liu and Diana Hass. Without their tireless editing skills, we could not have completed this manual.

**Scientific Computing Associates**
May 1991

# TABLE OF CONTENTS

# CHAPTER 1

# INTRODUCTION

This manual describes the capabilities that comprise the SCA-GSA product of the SCA Statistical System. The SCA-GSA product is a self-contained module of the SCA System. This module contains a broad range of statistical methods for data analysis. Capabilities described in this manual (and chapters containing them) include:

| | |
|---|---|
| Plotting data:<br>(Chapter 3) | Scatter plots of two or more variables, and plots of one or more variables over time. |
| Descriptive statistics:<br>(Chapter 4) | Frequently used sample statistics of a data set, exploratory data analyses, a simple tabular display, and measures of correlation. |
| Plots of location, dispersion and distribution:<br>(Chapter 5) | Histograms of single and pooled variables, dispersion plots of samples and sub-samples (with the possible inclusion of a reference distribution), confidence intervals, and probability plots. |
| Cross tabulation:<br>(Chapter 6) | One-way through n-way cross classification tables and statistics related to these tables. |
| Comparing two samples:<br>(Chapter 7) | Testing the difference between the means of two paired or independent samples. |
| Analysis of variance:<br>(Chapter 8) | One-way through n-way analysis of variance, including the incorporation of a Box-Cox power transformation, and the analysis of covariance. |
| Linear regression analysis:<br>(Chapter 9) | Multiple linear regression, weighted least squares, ridge regression, and regression with serially correlated errors. |
| Box-Jenkins Modeling:<br>(Chapter 10) | Univariate time series analysis and forecasting using Box-Jenkins ARIMA and transfer function models. The capabilities of this chapter are a subset of those contained in the SCA-UTS product of the SCA System (see *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*). |
| Nonparametric statistics:<br>(Chapter 11) | Calculation of a wide variety of statistics for nonparametric tests. |

| | |
|---|---|
| Distribution and model simulation: (Chapter 12) | User specified distribution or time series model simulation. |
| Analytic functions and matrix operations: (Appendix A) | Analytic functions and matrix operations that supplement the SCA System's statistical capabilities. |
| Data generation: (Appendix B) | User specified data generation, editing and other data manipulation. |
| Macro procedures: (Appendix C) | Creation and use of sequences of SCA statements to either perform SCA data analyses or augment SCA capabilities. |
| Utility information: (Appendix D) | Output saving and review, management of files, internal workspace (memory), and other utility related tasks of an SCA session. |

The information contained in the Appendices are condensed from those described in *The SCA Statistical System: Reference Manual for Fundamental Capabilities*. Selected information regarding the basic use of the SCA System and data entry are found in Chapter 2. This Chapter and the Appendices of this manual are designed to provide self-contained documentation for the SCA-GSA product.

## 1.1 Statistical Analysis Using the SCA Statistical System

A statistical analysis is most effective when it is used in an inductive-deductive fashion. Observation and basic knowledge first lead to the postulation of an initial theory or model. This theory or model is then tried and the results are reviewed to provide insight for modification or correction where necessary. The process is repeated until a satisfactory result is obtained. This inductive-deductive type of analysis is greatly facilitated by the flexibility in the SCA Statistical System, which allows both analytic and English-like statements to be blended by the user in any logical order when using the System.

## 1.2 The SCA System

Scientific Computing Associates (SCA) provides several self-contained modules in its statistical software system. At present the SCA Statistical System includes the SCA-GSA module for general statistical analysis, the SCA-QPI module for industrial quality and process improvement, the SCA-UTS module for univariate time series analysis and forecasting, the SCA-MTS module for multivariate time series analysis and forecasting, and the SCA-ECON/M module for econometric modeling and forecasting. The capabilities of other modules are discussed in other documents. In addition to its own unique capabilities, each module of the SCA System also contains a complete set of SCA fundamental capabilities,

including data input and output, analytic functions and matrix operations, data manipulation and editing, histograms and plots, macro procedures and other utility capabilities. Details regarding these capabilities are also described in *The SCA Statistical System: Reference Manual for Fundamental Capabilities*.

The modules described above are available as components in three statistical packages offered by SCA. These packages and their component modules are:

**General Application Package**: GSA

**Quality Improvement Package**: QPI and GSA

**Forecasting and Modeling Package:** UTS, MTS, ECON/M and GSA

In addition to the statistical modules described above, SCA provides software for employing windows and graphics, the **SCA Windows/Graphics Package**. This package provides an innovative means to integrate the computing power of mainframe computers and workstations with the user-friendly features and high-resolution graphics capabilities available on personal computers. The **SCA Windows/Graphics Package provides** for:

- High resolution graphics for scatter plots, contour plots and plots over time,
- A window environment for the SCA System,
- Menus to access all SCA capabilities,
- Convenient on-line help for SCA capabilities, and
- Two-way data transfer between mainframe computers and a PC.

# CHAPTER 2

# SYSTEM BASICS

Every software system has its own vocabulary and language to put user's "words" into action. This chapter provides the basics of the SCA command language and the use of the SCA System. In addition, information concerning the entry of data to the SCA System is also provided. More complete information can be found in *The SCA Statistical System: Reference Manual for Fundamental Capabilities*.

## 2.1 Getting Started

The SCA System is a command driven system. That is, the System responds to user instructions (commands) rather than to user chosen options from a menu. When the SCA System is used through the **SCA Windows/Graphics Package**, a Command Builder creates necessary commands from menu selections. In this manner, the SCA System has the same command language at all computing levels. All command lines must be followed by a carriage return. For easier reading in the remainder of this manual, we shall not explicitly display '<cr>' (carriage return) when presenting command lines. However, all command lines of the SCA System are preceded with the symbols '-->' as a means to indicate a line entered by the user. The symbols '-->' themselves should not be entered.

**Mainframe and workstation computers**

To access the SCA System on a mainframe computer, we enter

**SCA** ( or **sca** )

If this command does not invoke the System, a local computer consultant should be contacted regarding the appropriate command. It is possible a computing center may have installed the SCA System under a different command name.

**Personal computers**

The SCA System is also available for use on personal computers having a DOS, OS/2 or Macintosh operating system. Within the DOS or OS/2 environment, we first enter the subdirectory in which the SCA System was installed. The PC SCA System installation guide advises that the subdirectory be named SCA for DOS operating systems and OS2-SCA for OS/2 operating system. Thus enter

**CD \SCA** (or **CD \OS2-SCA**).

To invoke the SCA System in this directory, enter

**SCA**

To invoke the SCA System on Macintosh, we can simply double click the SCA icon from the folder in which it is stored. The icon should be created when the SCA System is installed.

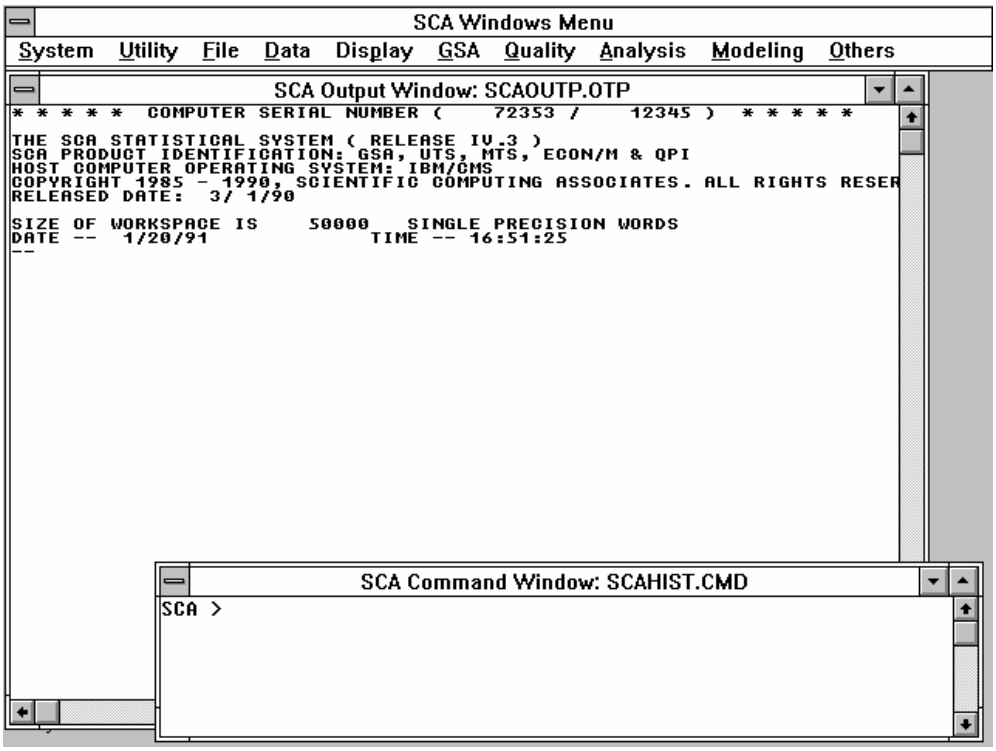
## System heading and prompt

When the SCA System is appropriately invoked, a set of short descriptive information appears. For example, the heading at an IBM/CMS mainframe site will be something like

```
* * * * *  COMPUTER SERIAL NUMBER (   172353 /     12345 )  * * * * *

THE SCA STATISTICAL SYSTEM ( RELEASE IV.3 )
SCA PRODUCT IDENTIFICATION: GSA, UTS, MTS, ECON/M & QPI
HOST COMPUTER OPERATING SYSTEM: IBM/CMS
COPYRIGHT 1985 - 1990, SCIENTIFIC COMPUTING ASSOCIATES. ALL RIGHTS RESERVED
RELEASED DATE:  3/ 1/90

SIZE OF WORKSPACE IS   50000   SINGLE PRECISION WORDS
DATE --  11/30/90             TIME -- 10:10:43
--
```

This set of information includes SCA release version, product names, host computer and operating system, and workspace (memory) size. The heading information is followed by a double dash, '--'. The double dash is a prompt issued by the SCA System. This indicates we can now enter an SCA command.

When the SCA System on a mainframe or workstation computer is invoked through the SCA Windows/Graphics Package (see the related document *SCA Windows/Graphics Package User's Guide* for more information), the following windows appear on the PC screen.




The heading information and subsequent SCA output are contained in the output window SCAOUTP.OTP. SCA commands are entered in the SCA command window, SCAHIST.CMD, or are generated from menu selections through the SCA Command Builder. The command history (i.e., the set of all SCA commands entered) of the SCA session are maintained in this window.

## Creating a larger workspace environment

We can designate a larger workspace (memory) size for an SCA session when we invoke the SCA System. This is a useful feature when we are dealing with larger data sets or complex computations. The amount of workspace that can be designated may be restricted due to local computer installation constraints or an SCA System constraint, depending on the subscription level. The maximum workspace size for the SCA System on personal computers varies between 30K and 35K words (1K words = 1000 words), while the maximum workspace for the SCA System on mainframe and workstation computers usually does not have a specific limit.

The designation of a larger workspace varies somewhat between computers and operating systems. For most operating systems, invoking the SCA System with

**SCA n**

where n is an integer, will allocate nK words of memory for the session. The instruction is different for IBM TSO and CMS operating systems where we must use either

**SCA SIZE(n)** (for an IBM TSO operating system)

or

**SCA SIZE n** (for an IBM CMS operating system)

If none of the above instructions affect the workspace size, it is necessary to check with a local computer consultant to determine what to do.

## 2.2 General Syntax of System Commands

Once we are in the SCA System, we have begun an **SCA session**. All SCA commands within a session are the same across all computer types. These commands are also called "statements".

Each statement is entered after the '--' prompt. We can use blanks freely in a statement to space words, but blanks cannot be used within names or numbers. Usually command lines are limited to 72 spaces and most commands can be written in one line. If we need to continue to another line, the current line must be ended with the character '@'. We refer to the symbol '@' as the **continuation character**. It must be the last non-blank character of any line being continued. It cannot be used as a hyphenation character. That is, words and numbers cannot be divided with '@'. The SCA System processes a command whenever a line is entered that does not end with '@'.

**Analytic statements**

There are two types of statements that we can use during an SCA session, analytic or "English-like". **Analytic statements** are used for most vector and matrix operations or manipulations. These statements have the general form

v = e

where "e" is an expression involving a combination of operators and variable names (the labels used to retain data in the SCA workspace); and "v" is a variable name (label) that will be used to hold results. For example,

LNY = LN(Y)

will take the natural logarithm of the data currently being held in the variable Y and store the result into the variable LNY. The statement

TEMP = INV(A) # B

will multiply the matrix B by the inverse of the matrix A (i.e., A-1B), then store the results into the variable TEMP.

A complete list of SCA analytic functions and matrix operators can be found in Appendix A. Some examples are also provided. A more detailed discussion regarding analytic statements can be found in *The SCA System: Reference Manual for Fundamental Capabilities*.

## English-like statements

**English-like statements** (or **paragraphs**) are used to accomplish most operations in an SCA session. These statements consist of a paragraph name that can be followed by one or more modifying sentences. For example,

```
PRINT  VARIABLE  IS  GROWTH
```

is an English-like statement. The paragraph name is PRINT and the modifying sentence is VARIABLE IS GROWTH. Here the function of the statement is implicit in the paragraph name. Information contained in the single modifying sentence is sufficient for the execution of the command.

The first word of a paragraph must be a valid paragraph name. This name is then followed by any number of modifying sentences. Sentences have no specific order of entry. **A sentence must be ended with a period if another sentence is to follow**. Each line within the paragraph, except for the last line, must have the continuation character ('@') as its last character.

Modifying sentences fall into two categories: **required** and **optional**. A sentence is optional if there is a default condition (or value) that can be used during the execution of the paragraph. An optional sentence is used only if we wish to change a default condition. A sentence is required if no default condition (or value) exists. If we omit any required sentence, the System will issue prompts requesting the information omitted.

For example, suppose there are two variables in the SCA workspace, TAX and INCOME, each containing 200 values. If we enter

```
PLOT  VARIABLES  ARE  TAX,  INCOME
```

then the System will produce a scatter plot based on all 200 data pairs (see Appendix C for more information on scatter plots). If we enter

```
PLOT  VARIABLES  ARE  TAX, INCOME.  SPAN IS 1,150
```

then the System will produce a scatter plot based on the first 150 pairs of data. The sentences VARIABLES and SPAN must be separated by a period. If only the statement

```
PLOT  SPAN  IS  1, 150
```

is entered, the System will prompt us for the variables to be used in the plot, since VARIABLES is a required sentence.

**Most frequently used required sentence**

For our convenience, the subject and verb of the "most frequently used sentence" of a paragraph can be omitted provided the sentence is the first sentence used after the paragraph name. For example, the VARIABLE sentence is the most frequently used sentence of both the PRINT and PLOT paragraphs. If we desire, we can omit the words VARIABLES ARE in these paragraphs. That is, the statement

```
PRINT   GROWTH
```

is equivalent to the statement

```
PRINT   VARIABLE  IS  GROWTH
```

The statement

```
PLOT   TAX,  INCOME.   SPAN  IS  1,150
```

is processed by the SCA System in the same fashion as the statement

```
PLOT   VARIABLES  ARE  TAX, INCOME.   SPAN  IS 1, 150
```

Note that if the statement

```
PLOT    SPAN  IS  1, 150.   TAX,  INCOME
```

is entered, then an error occurs. The System would interpret TAX as the first three letters of a sentence name and not as variable information. Very often, the "most frequently used sentence" is the only sentence specified in a paragraph. The portion of the "most frequently used sentence" that can be omitted is highlighted in the syntax description for every paragraph of the SCA System.

## 2.3 An Example

To illustrate the types of commands and using the SCA System, we will examine some data taken from the text Statistics for Experimenters by Box, Hunter and Hunter (1978). The data, shown below, are the growth rate (in coded units) of experimental rats and the amount (in grams) of a dietary substance fed to the rats.

| *Growth rate* | *Dietary supplement* |
|---|---|
| 73 | 10 |
| 78 | 10 |
| 85 | 15 |
| 90 | 20 |
| 91 | 20 |
| 87 | 25 |
| 86 | 25 |
| 91 | 25 |
| 75 | 30 |
| 65 | 35 |

We first want to transmit (or enter) data into the System's workspace (memory). There are many ways in which data can be entered. Complete information on the entry of data into the SCA workspace is provided in Chapter 3 of *The SCA Statistical System: Reference Manual for Fundamental Capabilities*. A summary of some frequently used methods for data entry is given in Section 10 of this Chapter. In this example we will enter both columns of data directly from the terminal. To enter the growth rate data we can enter

```
-->INPUT   GROWTH
```

Note that '-->' denotes a line we are entering (and should not be typed). We also must press the carriage return key to end our entry. We have informed the System that we will be transmitting data to it and want it retained in the System's workspace (memory) under the label GROWTH. Any valid name (see Section 2.4) can be used as a label for a variable. GROWTH has been chosen since this label is well suited to designate the data. The System responds with

```
READY  FOR  DATA  INPUT
```

The '--' prompt is not displayed because the System is not expecting any sort of instruction, just data. We can enter the data on one line by entering:

```
-->73  78  85  90  91  87  86  91  75  65
```

In order to tell the System that we are finished entering data for GROWTH, we now type

```
-->END  OF  DATA
```

The System responds with

```
GROWTH  , A    10   BY    1 VARIABLE, IS STORED IN THE WORKSPACE
```

Now we enter the dietary supplement data and retain it in the workspace under the label DIET.

```
-->INPUT   DIET
   READY FOR DATA INPUT

-->10  10  15  20  20  25  25  25  30  35
-->END  OF  DATA
   DIET   , A   10   BY   1 VARIABLE, IS STORED IN THE WORKSPACE
```

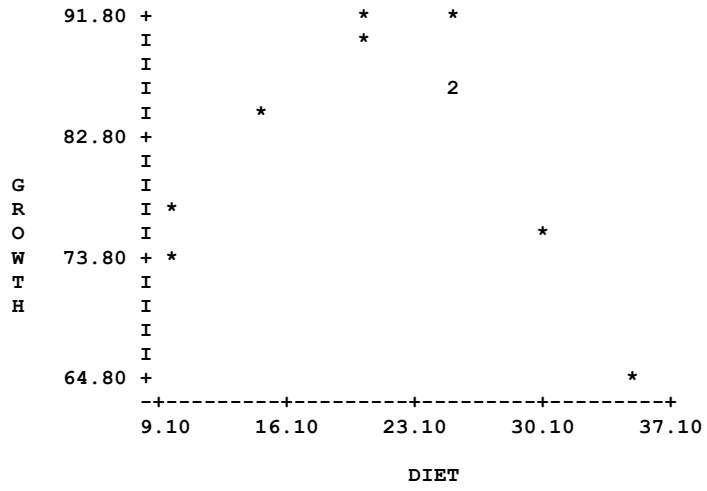
Before we continue, we can display the data that has been transmitted. We do this by entering the following paragraph

```
-->PRINT   GROWTH,  DIET
```

```
GROWTH    IS  A    10  BY     1  VARIABLE
DIET      IS  A    10  BY     1  VARIABLE

VARIABLE    GROWTH      DIET
 COLUMN-->       1         1
  ROW
    1       73.000    10.000
    2       78.000    10.000
    3       85.000    15.000
    4       90.000    20.000
    5       91.000    20.000
    6       87.000    25.000
    7       86.000    25.000
    8       91.000    25.000
    9       75.000    30.000
   10       65.000    35.000
```

To get an idea of how growth rate and dietary supplement are related, we display a scatter plot (see Appendix C) by entering

```
-->PLOT   GROWTH,  DIET
```



We observe that the effect of the dietary supplement on the growth rate increases to a peak level, then falls off. As a result we may wish to use regression analysis (see Chapter 9) to estimate the model.

$$Y = b_0 + b_1X + b_2X^2 + \text{error}$$

where Y is the growth rate and X is the amount of dietary supplement. We do not have the quadratic term, $X^2$ at present, but we can create it by using an analytic statement (see Appendix A). One means to create $X^2$ is to enter

```
-->DIET2 = DIET**2
```

The data generated by this command are retained in the workspace under the label DIET2. We are now ready for a regression analysis. We can fit the model above by entering

```
-->REGRESS  GROWTH, DIET, DIET2
```

The output generated from this command is suppressed at this time. Other options are available to us within the REGRESS paragraph, for example diagnostic checking, retaining calculated values and methods of fitting (see Chapter 9 for more information).

## 2.4 Names and Abbreviations

All data and models are stored in the SCA workspace (memory). We are required to provide names for all data and models that we place in the workspace. Other names used in an SCA session (i.e., paragraph and sentence names) are a part of the System's command language.

The names we specify for data or models can be of any length, although **only the first eight characters are interpreted by the System**. The first character of a name (label) must be a letter. The other characters may be letters, numbers or the underscore symbol, '_'. **Blanks cannot be used as part of a name**. Examples of valid names that we may specify are:

X, XDATA, X_DATA, X1, SERIES1, SERIES_1, DATASET1,
XDCDDEA, S33E45, F55XX_2, INFORMATION_FOR_SERIES_1

Examples of some invalid names are:

| | |
|---|---|
| 1X | (**the first character is not a letter**) |
| X DATA | (**blanks are not permitted**) |
| X-DATA | (**the special character ' - ', hyphen, is not permitted**) |

**Abbreviation rules**

All names used in an SCA session can be abbreviated. Names and labels that we specify are identified by the SCA System by their first eight (8) characters only. Hence the name

INFORMATION_FOR_SERIES_1

is interpreted by the SCA System as INFORMAT. The remaining characters are not maintained in memory, but may be used for readability. Thus, the name

INFORMATION_FOR_SERIES_2

is also interpreted by the System as INFORMAT. As a result, if we transmit data sequentially using these two names then all data first stored in the workspace under the label INFORMAT would be overwritten by the latter.

All sentence names are uniquely defined by their first three characters. Paragraph names are likewise defined, with a few exceptions due to name multiplicity (e.g., DESCRIBE and DESIGN). These names may be reduced to the first four characters. For example, the System internally interprets the statement

```
-->PLOT  VARIABLES  ARE  WITHHOLDING, INCOME.  @
-->      SPAN  IS  1, 30.
```

as

```
-->PLO   VAR  ARE  WITHHOLD, INCOME.   SPA  IS  1, 30.
```

## 2.5 Reserved Words and Symbols

Certain words and symbols have special meaning to the SCA System. They are summarized below and should only be used in their special context. More details can be found in *The SCA Statistical System: Reference Manual for Fundamental Capabilities*.

(1) FOR, TO, BY and $ are used to specify an implied list of arguments.

(2) The apostrophe ( ‘ ) is used in the identification of character strings.

(3) @ is a continuation symbol. It can also be used within macro procedures.

(4) -- is interpreted as an in-line comment when it is specified by the user.

(5) . specifies either a decimal point or a period.

(6) IS, ARE, IN, and ON are used as verbs within SCA sentences provided they immediately follow a sentence name. Otherwise, they are interpreted as variable names.

(7) The exclamation mark ! is used to cancel a statement when it appears as the last character of a statement.

## 2.6 Obtaining On-Line Help

The SCA System provides interactive on-line help on the capabilities and syntax of statements of the System. To obtain help information, enter the statement

```
-->HELP
```

More complete information is then provided. To obtain information on a specific SCA paragraph, enter

```
-->HELP   PARAGRAPH-NAME
```

To terminate a help session on mainframe computers, enter QUIT. To terminate the help session on a PC, press the ESC key. The System will then display the prompt '--' and the user will be at that position in an SCA session where help was requested. (If the DOS or OS/2 prompt 'C>' appears in the PC environment, enter the command QUIT.)

## 2.7 Responding to Prompts

Whenever a required sentence of a paragraph is either omitted or incomplete, the System will prompt for information it requires. When the System issues prompts, it only wants a direct response to its inquiries. For example, if we enter the statement

```
-->PLOT
```

rather than the statement

```
-->PLOT   TAX, INCOME
```

then the System will issue a prompt for the variable names omitted. Although the sentence that has been omitted is VARIABLES ARE TAX, INCOME, the System does not want the entry of a full sentence. It only wants the information omitted, i.e., TAX and INCOME. The response we need to provide is simply

```
-->TAX, INCOME
```

Prompts will continue until the System has all the necessary information it requires to proceed with the specified paragraph. If we wish to terminate the prompting session, we can

do so by entering the instruction QUIT. In addition to terminating the prompting session, the QUIT command will also abort the execution of the specified paragraph.

## 2.8 Panic Buttons

Occasionally, we may want to stop what is currently happening and get back to the basic command level (‘ -- ‘). The following are useful “panic” buttons:

(1) CTRL-C The execution of any paragraph can be terminated by simultaneously holding down the CTRL and C keys (or Break key for IBM MVS and IBM CMS operating systems). Output may not stop immediately as some output may already have been sent to a print buffer. In the IBM MVS and IBM CMS environments, be careful not to enter the Break key continuously as three successive entries of the Break key will terminate the SCA session.

(2) QUIT The instruction QUIT will terminate any prompting session. This will also terminate the execution of the specified command.

(3) ! The exclamation mark will cancel any statement, provided it is the last character of the statement. For example, suppose we enter the lines

```
-->PLOT   TEX, INCOME.   @
-->       SPAN  IS  1, 30
```

If we realize we have misspelled TAX as TEX before we transmit the second line, we can cancel the entire command by ending the second line with ‘ ! ‘.

## 2.9 Ending an SCA Session

To exit from an SCA session enter the command

```
-->STOP
```

## 2.10 Entering Data

There are many ways in which data can be transmitted to the SCA System. This section presents examples of the most common ways to enter data. The SCA paragraph INPUT may be used to transmit any data to the SCA System. Other paragraphs, BINPUT and FINPUT, are also available for special types of data.

### 2.10.1 Entering data from the terminal

We will first demonstrate how to enter data directly from a terminal during an SCA session. We will use the two data sets presented in Section 2.3 of this Chapter, growth rate and dietary supplement. The data sets are small enough that we may consider entering the

data directly from the keyboard. Previously, all the data of one variable were entered, then all the data of the other were entered. This is called **variable by variable** data entry. Alternatively, we could choose to enter both variables at the same time by entering the first pair of data, then the second, and so on. This is called **case by case** data entry.

## Entering data of a single variable

To enter the data for growth rate in a **variable by variable** fashion and store the data in the SCA workspace under the label GROWTH, enter

```
-->INPUT   GROWTH
```

This is equivalent to the statement

```
-->INPUT   VARIABLE  IS  GROWTH
```

in which the complete VARIABLE sentence is specified. The System responds with

```
READY FOR DATA INPUT
```

We now can enter data using free format (that is, data are separated by one or more blanks). We can enter all data on the same line, for example

```
-->73  78  85  90  91  87  86  91  75  65
```

or

```
-->73   78  85    90  91  87  86   91      75     65
```

We can also enter one data value per line, for example

```
-->73
-->     78
-->  85
--> 90
-->       91
-->87
-->  86
-->91
-->     75
--> 65
```

or we could enter the data on multiple lines

```
-->73  78  85
-->      90  91
-->87 86   91   75  65
```

As soon as we are through entering data, we enter

-->END OF DATA (OR -->END)

This completes the data entry for the variable GROWTH. The System will then respond with the message

```
GROWTH ,  A 10  BY  1 VARIABLE,  IS STORED IN THE WORKSPACE
```

**Entering data for more than one variable**

Instead of entering the two data sets in a variable by variable fashion, we could transmit both data sets simultaneously (i.e., in a **case by case** fashion) by entering

-->INPUT GROWTH, DIET

After the System prompt for data, we enter the ten cases of data using free format. Each case must be on a new line (record). This is, we enter

```
-->73  10
-->78  10
-->85  15
-->90  20
-->91  20
-->87  25
-->86  25
-->91  25
-->75  30
-->65  35
-->END OF DATA
```

The System will then respond with the message

```
GROWTH ,  A 10  BY  1 VARIABLE,  IS STORED IN THE WORKSPACE
DIET   ,  A 10  BY  1 VARIABLE,  IS STORED IN THE WORKSPACE
```

Each case (or record, or row) is transmitted in free format, so that the alignment shown above is arbitrary. Each line of data can be written in any convenient form.

## 2.10.2 Options related to the INPUT paragraph

When we enter data from the terminal, the only required sentence associated with the INPUT paragraph is the VARIABLES sentence. When data of only one variable are entered, the System will assume the data to be in free format, be a single column vector, be of single precision, and have no missing values. If we need to change any of these default conditions then an appropriate modifying sentence must be added.

**Entering a matrix of data**

When we transmit a matrix of data to the SCA System, we need to indicate the number of columns (NCOL) in the matrix. The number of rows is determined from the number of rows of data entered. For example, suppose the growth rate data was actually a matrix consisting of two columns of data. The value in the first column is the growth rate in week 1 and the value in the second column is the growth rate in week 2. To enter the data as a 10 x 2 matrix, GROWTH we may enter

```
-->INPUT   GROWTH.   NCOL  ARE  2.
```

and now enter data in a case by case fashion after the System prompt, for example

```
-->73    70
-->78    81
-->85    86
-->90    87
-->91    92
-->87    86
-->86    87
-->91    89
-->75    79
-->65    62
-->END OF DATA
```

The default value of NCOL for each variable is 1. If NCOL is changed from 1 for any variable, then data must be transmitted in a case by case fashion as above. For example, if we enter

```
-->INPUT XVECTOR, YMATRIX.  NCOL ARE 1, 3.
```

and enter the following data

```
-->1 2 3 4 5 6 7 8
-->8 7 6 5 4 3 2 1
-->0 1 1 2 2 3 3 3
-->END OF DATA
```

Then XVECTOR will be a 3 x 1 vector consisting of the values 1, 8, and 0; and YMATRIX will be the 3 x 3 matrix

```
2 3 4
7 6 5
1 1 2
```

All values after the 1 + 3 = 4th column of any row are ignored by the System.

**Entering non-numeric data, the PRECISION sentence**

The SCA System assumes that all data transmitted are single precision numeric data. To alter this default, we need to employ the PRECISION sentence. For example, suppose dietary data to be transmitted consist of the type of diet the rat was fed, A, B or C (i.e., character data) as well as the above two weeks worth of growth data. We can enter the statement

```
-->INPUT   GROWTH, DIET.  NCOLS  ARE  2, 1.    @
-->        PRECISIONS  ARE  SINGLE,  CHARACTER
```

Here two modifying sentences, NCOL and PRECISIONS, are used. NCOL specifies there are two columns in the data that for the variable GROWTH and one column of data for DIET. The PRECISION sentence is used to specify that DIET consists of character information. Since the default condition of the PRECISION sentence was changed for one variable (DIET), we need to specify the appropriate modifier for all variables of the sentence. Also note that since we were unable to write the INPUT statement entirely on one line, we used the continuation symbol, '@'.

## 2.10.3 Entering data from a file

In practice, we do not always enter data directly from a terminal. Often data exists on an external "flat file". A "flat file" is one that can be created or edited by a text editor. Flat files generally contain only one data set, or one set of case by case data records. When we enter data from an external file, we need to include the modifying sentence FILE in the INPUT paragraph to inform the SCA System that the data exists on a file and the file's name. If the FILE sentence is omitted, the System will assume that the data will be entered directly from the keyboard. Specification of the FILE sentence does not affect other default conditions of the INPUT paragraph (e.g., free format, single precision, no missing data). The line END OF DATA is not necessary in the external file, as the System will understand when it encounters the physical end of the file. For example, to enter the single variable GROWTH from file, we enter

```
-->INPUT   GROWTH.   FILE  IS  'FILE-NAME'
```

where "file-name" represents the appropriate name of the file containing the data. The actual name will be dependent on the conventions of the computer environment we are in.

Other modifying sentences, such as FORMAT, NCOL, and PRECISION can be included as in the case that data are transmitted from a keyboard. The FORMAT sentence is one that could be used if the data has been written onto the external file according to a specific format.

**File name conventions**

The convention used to name files varies according to the type of the computer and operating system. For example GROWTH.DAT is a valid file name on VAX VMS computers, GROWTH DATA A1 is a valid file name on IBM CMS computers, and U01234.GROWTH.DAT is a valid file name on IBM MVS computers. The file name GROWTH.DAT is also valid on IBM PC's and compatibles operating under DOS. On PC DOS computers, a drive may be added to a file name (e.g., A:GROWTH.DAT). If we are on a VAX with a VMS operating system and our data is stored in the file GROWTH.DAT, we would enter

```
-->INPUT  GROWTH.  FILE IS  'GROWTH.DAT'
```

If we are on a PC with GROWTH.DAT in drive A, we would enter

```
-->INPUT  GROWTH.  FILE IS  `A:GROWTH.DAT'
```

Note that the file name must be enclosed within the pair of single apostrophes ( ‘ ). In the remainder of this document, we will employ data set names appropriate in a VAX VMS or PC DOS setting, unless otherwise noted.

## 2.10.4 More examples of data entry

This section provides more examples on data entry using the INPUT paragraph. In addition to the INPUT paragraph, the FINPUT and BINPUT paragraphs can be used to access data that are stored on external files containing internal documentation specific for SCA usage. Information on SCA files and related paragraphs can be found in *The SCA Statistical System: Reference Manual for Fundamental Capabilities*.

In the following examples, we do not provide specific data. Instead data are only described and illustrated when necessary.

**(1) Entry of character and numeric data from a terminal**

Three variables will be entered from the terminal in a case by case fashion. The first variable is a list of names (last name and first name). The second and third are mathematics and English scores. We need to alter both the defaults for PRECISION and COL as the first variable is character data and has two columns of data. An appropriate statement is

```
-->INPUT  NAMES, MATH, ENGLISH.  NCOLS  ARE  2, 1, 1.  @
-->            PRECISION  IS  CHARACTER, SINGLE, SINGLE
```

**(2) Entry of character and numeric data from a file**

Same data as in (1), but the data is on an IBM CMS file TESTDATA DATA A1. An appropriate statement is

```
-->INPUT   NAMES, MATH, ENGLISH.   NCOL  ARE  2, 1, 1.   @
-->        FILE  IS  'TESTDATA DATA A1'.              @
-->        PRECISION  IS  CHARACTER, SINGLE, SINGLE
```

**(3) Specifying a format for data**

Some sales data has been downloaded from a mainframe computer to a PC. The name of the file on PC is SALES.DAT. The data are of one variable. There are 15 years of data, with each record having the sales totals (in thousands of dollars) for each month of the year. The data have been compressed so that a typical record on the file looks like

```
95.3 88.2 87.1 90.2 88.1 91.4101.3 87.2 88.6 91.6 95.8100.4
```

That is, the sales for January were $95,300, the sales for February $88,200, and so on. We need to include a FORMAT statement indicating that every record has 12 sets of numbers, each number is in a field of 5 characters of the form “xxx.x”. An appropriate statement for this data is

```
-->INPUT   SALES.   FILE  IS  'SALES.DAT'.   @
-->        FORMAT  IS  '12F5.1'
```

**(4) Data having missing data code as values**

We will transmit the same data as in (3), but some months had missing sales figures. In those cases the missing data code ***** appears in the five character string for the month. For example, suppose the third value of the “typical record” is missing. Then this record is

```
95.3 88.2***** 90.2 88.1 91.4101.3 87.2 88.6 91.6 95.8100.4
```

In this case the statement given in (3) is still appropriate for data entry.

**(5) Data having a numeric substitute for missing values**

Same data as in (4), except those missing entries are recorded as -99.9. We can either use the INPUT statement of (3) above and work with the value -99.9, or we can redefine -99.9 to an internal missing data code. In the latter case, we can employ the statement

```
-->INPUT   SALES.   FILE IS 'SALES.DAT'.   @
-->        FORMAT IS '12F5.1'.    REDEFINE -99.9
```

## REFERENCE

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). *Statistics for Experimenters*. New York: Wiley.

# CHAPTER 3

# PLOTTING DATA

Data displays in various forms are essential tools in the analyses of a data set. Often the best way to comprehend data comes from visual depictions, rather than from extensive statistical analyses. We can immediately realize the need to account for trend or the seasonal behavior of time series data through a time plot, a plot of the data over time. We can check if a process is in a state of statistical control through the use of Shewhart charts. Relationships that may exist between variables can be discerned through scatter plots, plots of one variable against another. Moreover, we may be able to determine the basic functional form of relationships (e.g., linear, quadratic) with these plots. We may discover that it may be more appropriate statistically to analyze the data in a metric other than the one in which the data are recorded. For example, a logarithmic, square root, or other type of transformation, may be appropriate. Spurious observations, or typographical errors in data entry, may be quickly spotted in a data plot. For such reasons, it is important that we should always view data first and we should not rely on statistical summaries alone.

The SCA System provides a number of paragraphs useful in the display of data. Scatter plots and time plots are discussed in this Chapter. Histograms and probability plots are covered in Chapter 5. Shewhart charts are discussed in Section 3. Other plots specific to experimental design and analysis are found in the SCA reference manual *Quality and Productivity Improvement Using the SCA Statistical System*.

## 3.1 Scatter Plots

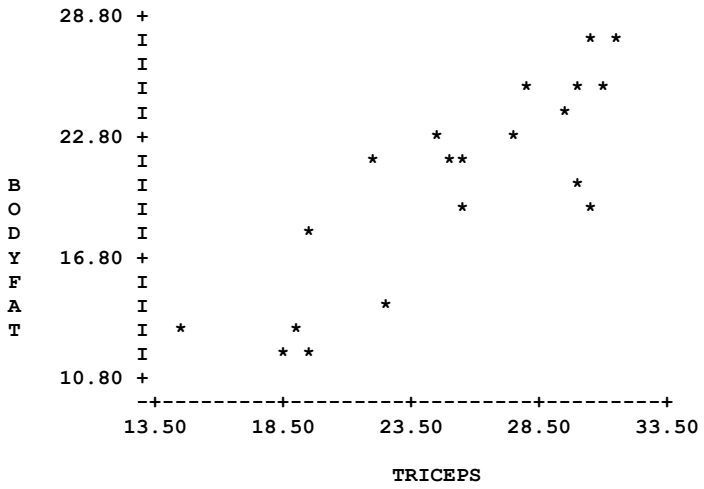
To illustrate plots of one or more variables against another, we will consider a data set analyzed in Neter, Wasserman, and Kutner (1983, Chapters 8 and 11). The data came from a study of the relation of bodyfat to triceps skinfold thickness and thigh circumference of 20 subjects. The data are shown in Table 1 and are stored in the SCA workspace under the labels, BODYFAT, TRICEPTS, and THIGH, respectively. A more complete analysis of these data may be found in Chapter 9.

**Table 1 Bodyfat study data**

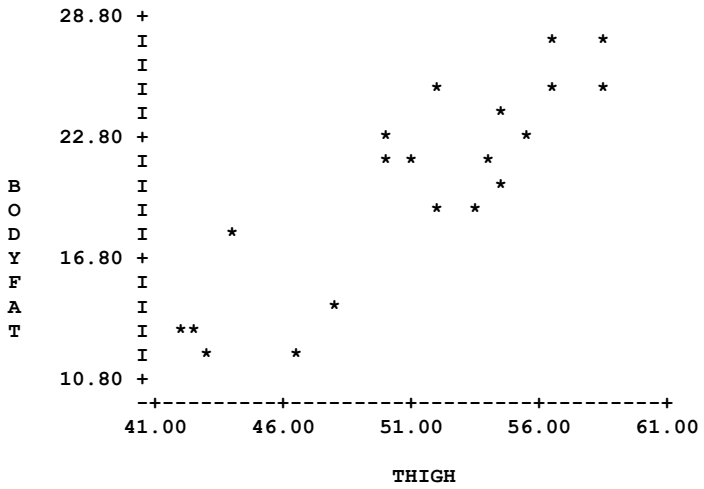
| *Subject* | *Triceps Skinfold Thickness TRICEPS* | *Thigh Circumference THIGH* | *Body Fat BODYFAT* |
|---|---|---|---|
| 1 | 19.5 | 43.1 | 11.9 |
| 2 | 24.7 | 49.8 | 22.8 |
| 3 | 30.7 | 51.9 | 18.7 |
| 4 | 29.8 | 54.3 | 20.1 |
| 5 | 19.1 | 42.2 | 12.9 |
| 6 | 25.6 | 53.9 | 21.7 |
| 7 | 31.4 | 58.5 | 27.1 |
| 8 | 27.9 | 52.1 | 25.4 |
| 9 | 22.1 | 49.9 | 21.3 |
| 10 | 25.5 | 53.5 | 19.3 |
| 11 | 31.1 | 56.6 | 25.4 |
| 12 | 30.4 | 56.7 | 27.2 |
| 13 | 18.7 | 46.5 | 11.7 |
| 14 | 19.7 | 44.2 | 17.8 |
| 15 | 14.6 | 42.7 | 12.8 |
| 16 | 29.5 | 54.4 | 23.9 |
| 17 | 27.7 | 55.3 | 22.6 |
| 18 | 30.2 | 58.6 | 25.4 |
| 19 | 22.7 | 48.2 | 14.8 |
| 20 | 25.2 | 51.0 | 21.1 |

We wish to discover the relationships, if any, that exist between BODYFAT and the variables TRICEPS and THIGH. One set of visual representations are the individual plots of the values of the BODYFAT variable with the associated values of both the TRICEPS and THIGH variables. These scatter plots may be obtained using the PLOT paragraph as follows.

-->PLOT BODYFAT, TRICEPS




-->PLOT BODYFAT, THIGH



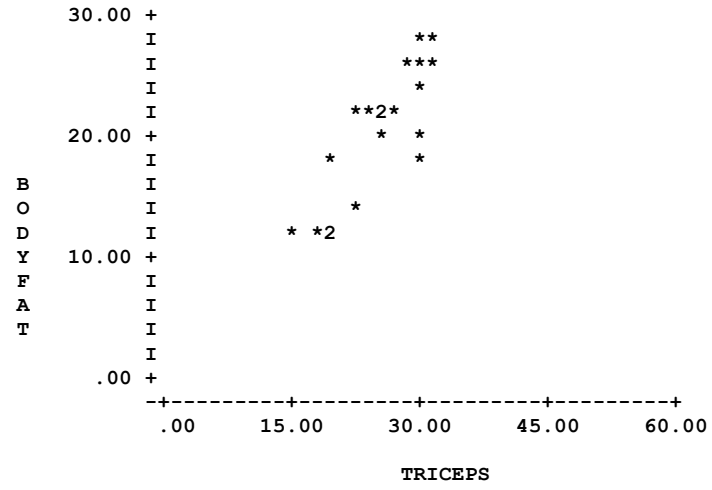

The PLOT paragraph provides us with a display of symbols on an L-shaped frame. The frame is composed of a vertical Y-axis for the first variable specified, BODYFAT, and a horizontal X-axis for the second variable specified, TRICEPS or THIGH. The symbol '*' is used to indicate a data point; that is, one of the (x,y) pairs displayed.

The SCA System automatically chooses suitable intervals for the values of the axes based on the range of values assumed by the 'X' and 'Y' variables and the amount of space available for the display. In the plots above, the range for the Y-axis is the same for both plots, since the same variable is used; but the ranges for the X-axes are different. The values of TRICEPS range between 13.50 and 33.50, and those of THIGH range between 41.00 and 60.00.

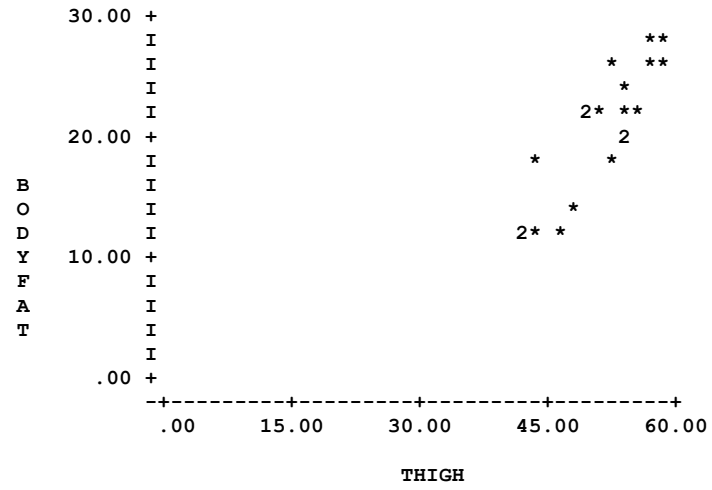
We observe what appears to be a linear relationship between BODYFAT and TRICEPS as well as between BODYFAT and THIGH. For illustrative purposes, we can re-scale the plots so that the ranges for the axes are the same in both plots. We can see from the plots, and

from Table 1, that the largest value of BODYFAT, the Y variable, is under 30, and the largest value of either TRICEPS or THIGH, the X variables, is less than 60. We can construct plots in which 0.0 is used as the lower end-point of both axes and 30.0 or 60.0 is used as the upper end-point of the Y or X axis, respectively. We can accomplish this by including the RANGE sentence as follows:

-->PLOT BODYFAT, TRICEPS. RANGE IS Y(0.0,30.0), X(0.0,60.0)




-->PLOT BODYFAT, THIGH. RANGE IS Y(0.0,30.0), X(0.0,60.0)

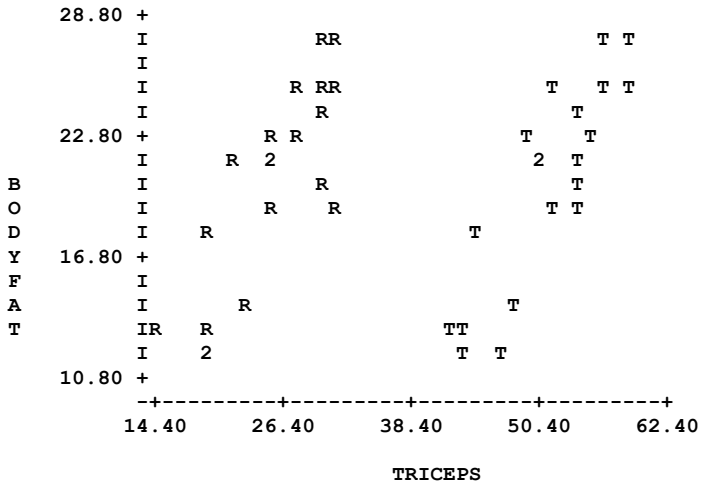



Now we can observe the data on the same scales for all variable involved. In the above two plots the symbol '2' appears several times. The symbol '2' indicates there are two data points so close together that they cannot be shown uniquely. The reason for this is immediate. Since we have imposed an arbitrary scale for the X-axis, the resultant data points are "bunched" together a little more than before. As a result, all data pairs cannot be displayed distinctly. The same inference can be made for the symbols '3', '4', . . . , '9' should any

appear. 'A' through 'Z' represent 10 through 35 data points, and '#' is used for 36 or more. Other "tagging" of points is possible (see Section 3.1).

In the plots above, we have plotted exactly one Y variable against one X variable in the same frame. If we wished to display other scatter plots, we must use separate frames. However, we can display multiple plots on the same frame through the MPLOT paragraph. To display the scatter plots of BODYFAT against TRICEPS and BODYFAT against THIGH on the same frame, we can enter the following.

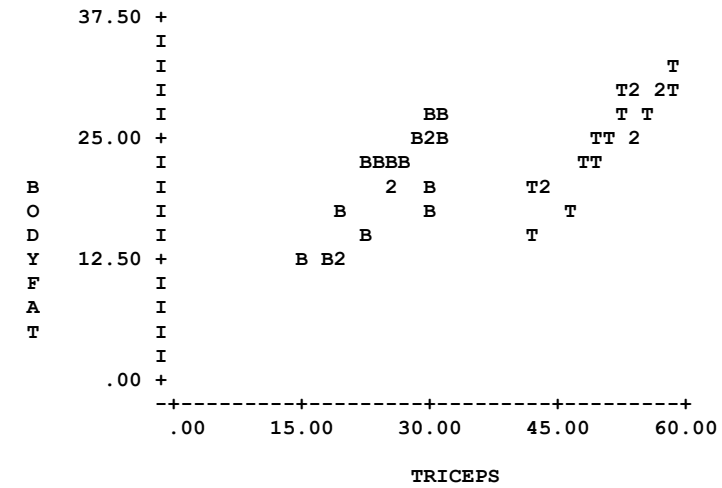
```
-->MPLOT   Y-VARIABLES ARE BODYFAT, BODYFAT.  @
-->        X-VARIABLES ARE THIGH, TRICEPS.    @
-->        SYMBOLS ARE  'T', 'R'.
```




Note the values of the axes have been determined automatically by the SCA System. In addition, we have "distinguished" the two scatter plots by using the symbol 'T' for the data points of the first plot (X-variable is THIGH and Y-variable is BODYFAT), and 'R' for the second plot (TRICEPS and BODYFAT).

It may appear redundant that we specified the Y-VARIABLES above as BODYFAT and BODYFAT; but it was necessary. The MPLOT paragraph does not place any limitation on the X or Y variables that can appear on the same frame. For example, we can display the scatter plots of two distinct Y variables against two distinct X variables on the same frame. For the purpose of illustration, we will display the scatter plots of BODYFAT against TRICEPS and TRICEPS against THIGH on the same frame. Here TRICEPS is used as both an X and a Y variable. the symbols 'B' and 'T' will be used to distinguish the Y variable. We will also force the ranges for the X-axis and Y-axis to be 0.0 to 60.0 and 0.0 to 40.0, respectively.

```
-->MPLOT   Y-VARIABLES  ARE  TRICEPS, BODYFAT.    @
-->        X-VARIABLES  ARE  THIGH, TRICEPS.      @
-->        SYMBOLS ARE 'T', 'B'.   RANGES ARE Y(0.0, 40.0), X(0.0, 60.0)
```




The SCA System will use the names of the last X and Y variables specified for axes labels.

## 3.2 Plotting Data Over Time

Data collected over time usually embodies some time dependent characteristics. The exact nature of these characteristics are not always obvious. Some may be suspected or assumed, such as a trend or seasonal behavior, as occur often in business data. Others may be hidden. For example, an experiment may be conducted in which the cutting precision of a tool on metals of various alloy compositions is measured. It may be the case that the tool is subject to wear regardless of the metal being cut, hence it may be necessary to include time as a factor in the analysis. In general, if data are gathered or recorded in any sort of time dependent order, it is a good practice to plot the data against time.

### 3.2.1 Plots of a single variable over time

A set of data from the *Commodity Year Book* (1986) will be used to illustrate plots over time. The data, listed in Table 2, are comprised of monthly observations, From January 1980 through December 1986 of the following prices:

(1) The average wholesale price of gasoline (regular grade, leaded)
(2) The average price of crude petroleum at wells

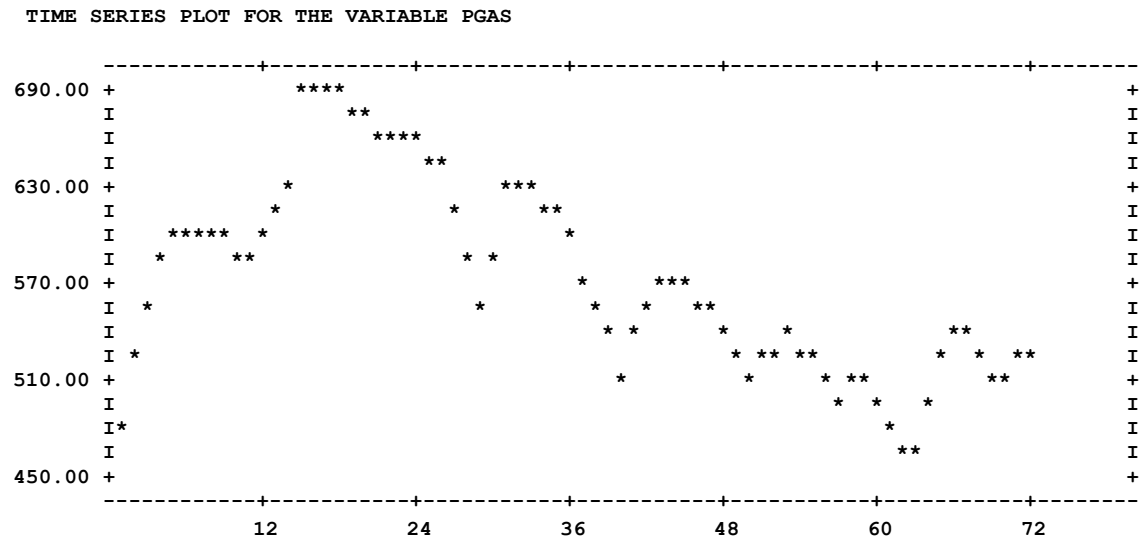
The data are stored in the SCA workspace under the names PGAS and PCRUDE, respectively. A more complete description and analysis of these data can be found in Chapters 9 and 10.

**Table 2 Gasoline data**

| *Obs.* | *Month* | *Gasoline Price PGAS* | *Crude oil Price PCRUDE* | *Obs.* | *Month* | *Gasoline Price PGAS* | *Crude oil Price PCRUDE* |
|---|---|---|---|---|---|---|---|
| 1 | 1/80 | 481.1 | 447.8 | 37 | 1/83 | 576.7 | 627.5 |
| 2 | 2/80 | 517.5 | 449.1 | 38 | 2/83 | 551.4 | 604.1 |
| 3 | 3/80 | 560.4 | 455.8 | 39 | 3/83 | 533.5 | 591.1 |
| 4 | 4/80 | 585.4 | 465.5 | 40 | 4/83 | 515.3 | 591.1 |
| 5 | 5/80 | 595.5 | 470.9 | 41 | 5/83 | 537.2 | 591.1 |
| 6 | 6/80 | 598.6 | 478.6 | 42 | 6/83 | 559.5 | 591.0 |
| 7 | 7/80 | 601.1 | 480.7 | 43 | 7/83 | 566.6 | 589.1 |
| 8 | 8/80 | 602.9 | 494.2 | 44 | 8/83 | 571.2 | 588.6 |
| 9 | 9/80 | 599.6 | 498.1 | 45 | 9/83 | 566.3 | 589.1 |
| 10 | 10/80 | 591.5 | 505.3 | 46 | 10/83 | 559.2 | 589.1 |
| 11 | 11/80 | 590.8 | 523.6 | 47 | 11/83 | 548.2 | 589.0 |
| 12 | 12/80 | 596.1 | 551.7 | 48 | 12/83 | 535.8 | 588.0 |
| 13 | 1/81 | 607.5 | 614.1 | 49 | 1/84 | 518.3 | 589.0 |
| 14 | 2/81 | 632.9 | 734.7 | 50 | 2/84 | 512.4 | 589.0 |
| 15 | 3/81 | 683.2 | 734.8 | 51 | 3/84 | 517.9 | 589.0 |
| 16 | 4/81 | 694.7 | 734.5 | 52 | 4/84 | 520.5 | 587.5 |
| 17 | 5/81 | 690.4 | 732.3 | 53 | 5/84 | 532.6 | 587.5 |
| 18 | 6/81 | 685.6 | 711.3 | 54 | 6/84 | 531.0 | 587.0 |
| 19 | 7/81 | 677.4 | 696.5 | 55 | 7/84 | 520.9 | 586.4 |
| 20 | 8/81 | 668.4 | 694.7 | 56 | 8/84 | 504.6 | 585.1 |
| 21 | 9/81 | 666.4 | 694.7 | 57 | 9/84 | 500.3 | 584.7 |
| 22 | 10/81 | 666.1 | 687.2 | 58 | 10/84 | 509.8 | 584.0 |
| 23 | 11/81 | 661.7 | 685.2 | 59 | 11/84 | 511.3 | 571.8 |
| 24 | 12/81 | 657.7 | 686.3 | 60 | 12/84 | 502.0 | 566.2 |
| 25 | 1/82 | 651.7 | 686.3 | 61 | 1/85 | 480.5 | 550.3 |
| 26 | 2/82 | 642.3 | 671.6 | 62 | 2/85 | 458.4 | 536.3 |
| 27 | 3/82 | 621.1 | 649.3 | 63 | 3/85 | 467.2 | 536.6 |
| 28 | 4/82 | 578.6 | 625.9 | 64 | 4/85 | 493.9 | 538.4 |
| 29 | 5/82 | 555.7 | 625.8 | 65 | 5/85 | 522.5 | 541.3 |
| 30 | 6/82 | 582.7 | 626.2 | 66 | 6/85 | 535.7 | 540.6 |
| 31 | 7/82 | 628.8 | 626.3 | 67 | 7/85 | 539.3 | 539.6 |
| 32 | 8/82 | 636.3 | 626.3 | 68 | 8/85 | 526.7 | 535.4 |
| 33 | 9/82 | 628.4 | 626.7 | 69 | 9/85 | 513.6 | 536.6 |
| 34 | 10/82 | 617.2 | 641.1 | 70 | 10/85 | 506.1 | 539.2 |
| 35 | 11/82 | 611.0 | 640.0 | 71 | 11/85 | 520.1 | 541.8 |
| 36 | 12/82 | 600.7 | 628.1 | 72 | 12/85 | 523.0 | 544.3 |

Since these data are collected on a monthly basis, we would like to indicate the end of each year of data. We will plot the PGAS data using the TSPLOT (Time Series PLOT) paragraph.

-->TSPLOT PGAS. SEASONALITY IS 12. SYMBOL IS '*'.




We see the data are plotted against a horizontal time axis. Marks along the axis are at multiples of 12, that specified in the SEASONALITY sentence. The use of the SYMBOLS sentence is explained in detail in Section 3.3, but its purpose is evident.

**Remark**: The SEASONALITY sentence is a replacement of the sentence, TIC-MARK. In the event your version of the SCA System does not recognize the SEASONALITY sentence, it is likely you have an older version of the System. In such a case, please substitute TIC-MARK for SEASONALITY.

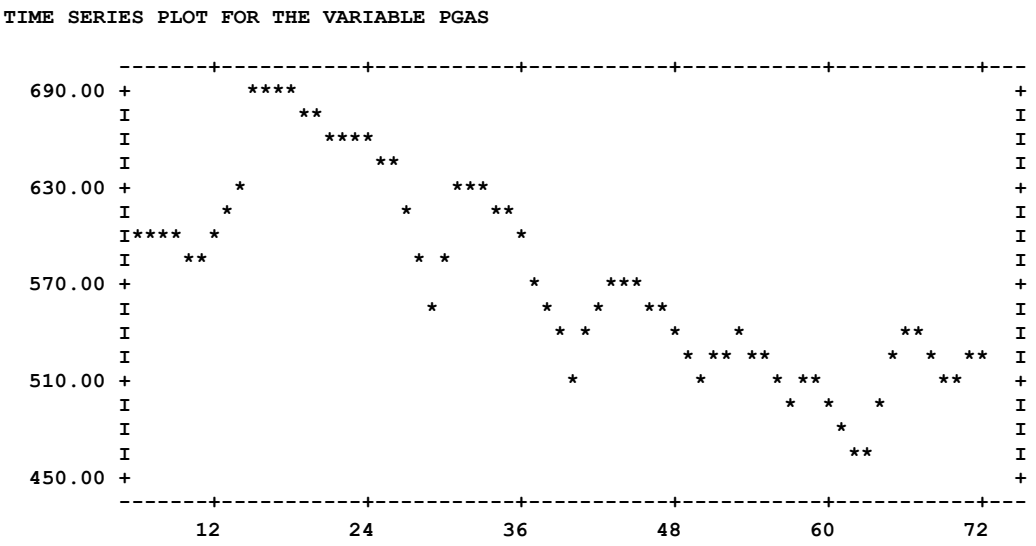
The display provided by the TSPLOT paragraph is dependent on the output width available to the SCA System. As in the PLOT paragraph, the SCA System automatically scales a plot to fit within the space available for display. Unlike the PLOT paragraph, the TSPLOT paragraph will uniquely represent any data point displayed. Consequently, if the SCA System does not have "enough space" available to present the complete time plot, it will truncate the data displayed. Since the last data points are often the most influential in forecasting a time series, the SCA System plots all data it can from the end of the series forward. Any truncation of data occurs at the beginning of the series.

The display of the above plot was generated on a "wide screen". The default output width assumed by the SCA System is 80 characters. This value is appropriate for virtually all output devices (terminals, printers, files). This output width can be altered by the PROFILE paragraph (see *The SCA Statistical System: Reference Manual For Fundamental Capabilities*). We can increase the output width to 132 characters (i.e., that of "large" computer paper) by entering

PROFILE OWIDTH IS 132

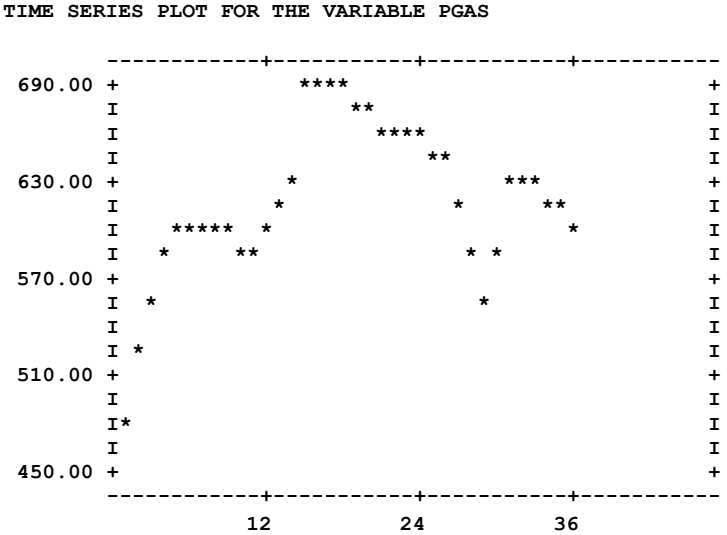
If we are limited to 80 characters of output width, the following display occurs

```
-->TSPLOT  PGAS.  SEASONALITY IS 12.  SYMBOL IS  '*'.
```




If we are confined to a limited output space yet desire a plot of the complete series, there are two things we may do. One is to plot the series vertically rather than horizontally. This may be done using the TPLOT paragraph (shown later). The second option is to split the plot into pieces using the SPAN sentence. We will do this here, by displaying the first 36 observations then the last 36 observations. Since the range of values may be different in the two plots, we will impose a range of 450 to 700. This appears reasonable given the values of the above plot.

```
-->TSPLOT   PGAS.   SPAN IS 1, 36.  SEASONALITY  IS  12.  @
-->         SYMBOL IS `*'.    RANGE IS 450, 700.
```



```
-->TSPLOT   PGAS.   SPAN IS 37,72.    SYMBOL IS  '*'.     @
-->         SEASONALITY IS 12, 37.   RANGE IS 450, 700.
```

```
     TIME SERIES PLOT FOR THE VARIABLE PGAS

            ------------+-----------+-----------+-----------
     690.00 +                                               +
            I                                               I
            I                                               I
            I                                               I
     630.00 +                                               +
            I                                               I
            I                                               I
            I                                               I
     570.00 +*     ***                                      +
            I *   *   **                                    I
            I  * *      *    *           **                 I
            I            * ** **         *  *  **          I
     510.00 +   *         *     * **        **              +
            I                  *  *   *                     I
            I                     *                         I
            I                      **                       I
     450.00 +                                               +
            ------------+-----------+-----------+-----------
                        48          60          72
```

## 3.2.2 Plots of more than one variable over time

We have several options available if we wish to display the plots of more than one variable over time. One option is to use the TSPLOT separately for each variable. We can also specify more than one variable in the TSPLOT paragraph. for example, suppose both PGAS and PCRUDE are specified in TSPLOT. We have

```
-->TSPLOT   PGAS, PCRUDE.   SEASONALITY IS 12.   SYMBOL IS  '*'.
```

```
     TIME SERIES PLOT FOR THE VARIABLE PGAS

            ------------+-----------+-----------+-----------+-----------+-----------+--------
     720.00 +                                                                         +
            I              *                                                         I
            I             * ***                                                      I
            I                  *****                                                 I
     648.00 +                      **                                                +
            I            *          *   ***                                        I
            I           *                  **                                      I
            I   *********                    *                                     I
     576.00 +                          * *      *      *                           +
            I  *                        *        *   ** **                         I
            I                                  * *     **    **          **         I
            I *                                *        * **  *         *  ** **    I
     504.00 +                                             *     *****         *     +
            I*                                                     *  *             I
            I                                                       *               I
            I                                                      *                I
     432.00 +                                                                         +
            ------------+-----------+-----------+-----------+-----------+-----------+--------
                        12          24          36          48          60          72
```

```
TIME SERIES PLOT FOR THE VARIABLE PCRUDE

        ------------+-----------+-----------+-----------+-----------+-----------+--------
 720.00 +             *****                                                      +
        I                  ***                                                  I
        I                     ****                                              I
        I                         *                                             I
 648.00 +                         *      **                                    +
        I                          ******  **                                   I
        I             *                      *                                  I
        I                                    ******************                 I
 576.00 +                                                    ***                +
        I            *                                         **               I
        I                                                        ***********     I
        I          *                                                           I
 504.00 +        **                                                           +
        I     ***                                                             I
        I   **                                                                I
        I***                                                                  I
 432.00 +                                                                     +
        ------------+-----------+-----------+-----------+-----------+-----------+--------
                    12          24          36          48          60          72
```

We obtain two separate time series plots, but the same range of values is used as the Y axis of both plots. The SCA System automatically determines a range appropriate for all variables involved.

We may wish to view the variables in the same display frame. This can be useful in determining if the values assumed by one variable may be influenced by the values of another. Perhaps one series "leads" another in some way. For example, a low value for one series may indicate a low (or high) value of another series in a future time period. Similarly, a turn in one series (e.g., a decreasing set of values that change to increasing) may indicate a subsequent turn in another series.

The MTSPLOT (Multiple Time Series PLOT) paragraph may be used to display the plots of two or more series, or variables, over time on the same frame. Data are distinguished by letters. Unless we specify our own set of symbols, the symbol 'A' is used to represent the first variable specified, 'B' for the second, and so on. The symbol '*' is used if any displayed values are concurrent. We can specify our own symbols by including the SYMBOLS sentence in the paragraph.

We will display the time plots of PGAS and PCRUDE in the same frame to illustrate the use of the MTSPLOT paragraph. We will use the symbol 'X' to represent PGAS data and '+' for PCRUDE data. As before, we will also include the SEASONALITY sentence. We have increased the display width to assure plots of the complete data sets.

```
-->MTSPLOT   PGAS, PCRUDE.   SEASONALITY IS 12.   SYMBOLS ARE 'X', '+'.

 TIME SERIES PLOT FOR VARIABLES     PGAS   AND   PCRUDE

       ------------+-----------+-----------+-----------+-----------+-----------+--------
720.00 +             +++++                                                          +
       I               X  +++                                                       I
       I              X XXX  ++++                                                   I
       I                   XXXXX +                                                  I
648.00 +                        XX+      ++                                        +
       I             X           X+++***  ++                                       I
       I            *                    XX  +                                     I
       I   XXXXXXXXX                      X  ++++++++++++++++++                    I
576.00 +                               X X      X      X            +++             +
       I  X        +                   X        X   XX XX             ++           I
       I                                        X X     XX    XX       +++++*++++   I
       I X        +                             X        X XX  X         X  XX XX   I
504.00 +        ++                                       X     XXXXX         X     +
       IX    +++                                                 X  X              I
       I   ++                                                      X               I
       I+++                                                        X               I
432.00 +                                                                          +
       ------------+-----------+-----------+-----------+-----------+-----------+--------
                  12          24          36          48          60          72
```

The MTSPLOT paragraph can be a useful visual tool if two variables are slightly "out of synch", or if we wish to display the actual values of a series together with forecasted values (and standard errors). For more information on the latter, see Chapter 10. However, it is possible that the overlap of the two or more plots presents a more confusing pattern than we may like. Even less useful information may be obtained when either the range of values of one variable dwarf those of another, or if the combined ranges of all variables are extreme.

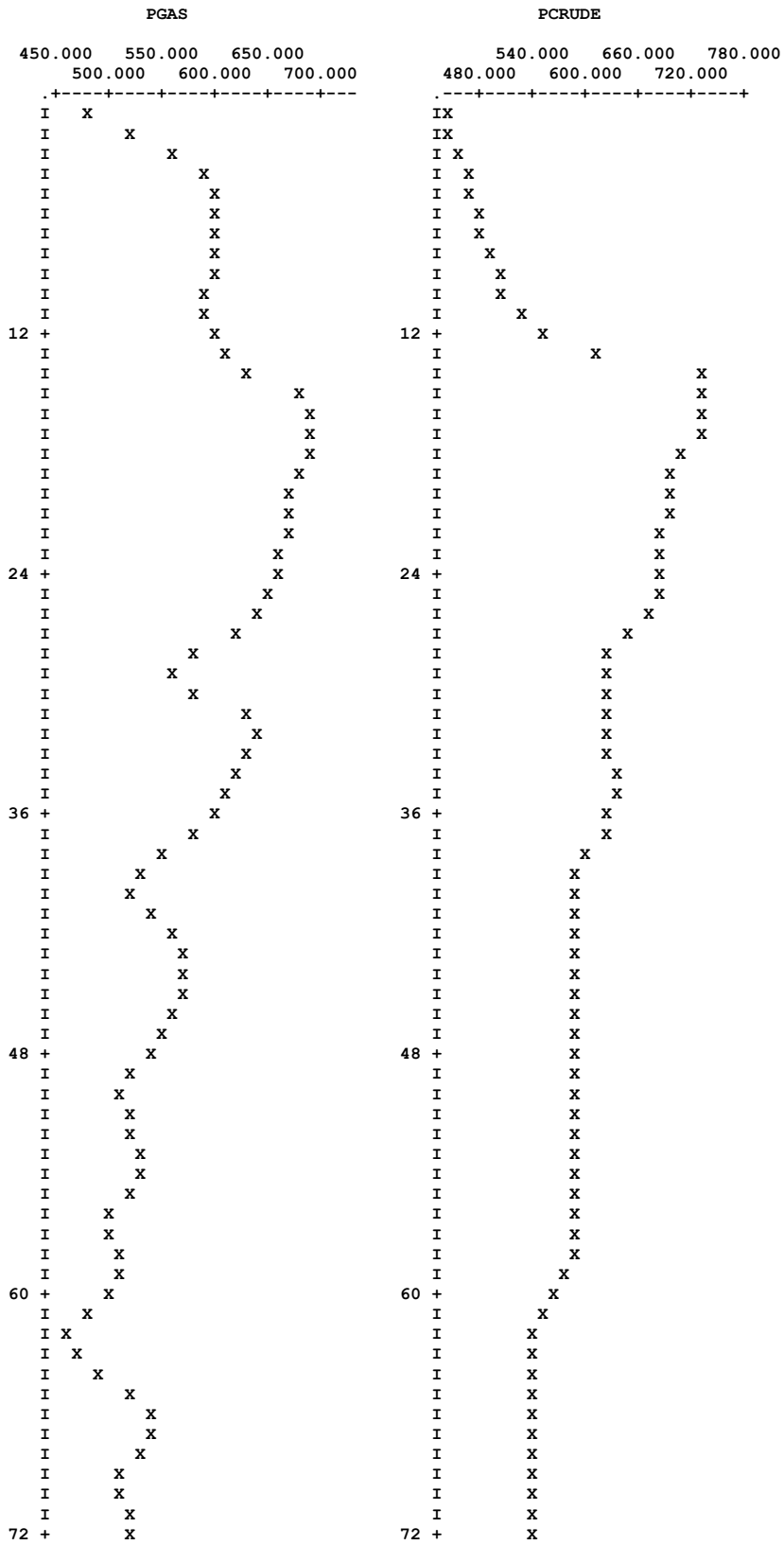
### 3.2.3 Vertical time plots

The time axis for all plots above has been horizontal. This can be convenient for the visual display of a relatively short series of data, but it can be limiting if a data set is lengthy. As an alternative, we can choose to have a vertical time axis. This will permit the time plot of a data set of any length, but the display will usually run over several pages, or screens. It is advised that when a vertical time axis is used, the plot should be routed to a printer or to a file.

Two paragraphs are provided for plotting data over a vertical time axis, TPLOT and MTPLOT. We can plot one or more data sets using TPLOT, and we can display multiple plots on the same time frame using MTPLOT. MTPLOT offers more clarity than MTSPLOT in its display of multiple plots since more "space" is available to it. Options for these paragraphs are the same as for TSPLOT and MTSPLOT.

TPLOT provides us with an additional means to display more than one series. If more than one variable is specified, then all variables will be shown in parallel to one another on the display device. For example, consider a time plot of PGAS and PCRUDE in the same TPLOT paragraph (the display has been edited for presentation purposes).

-->TPLOT PGAS, PCRUDE. SEASONALITY IS 12. SYMBOL IS 'X'.



The advantage in this sort of display is that concurrent observations are aligned for variables that may be related, but the individual pattern of each series is still separate from all others. A disadvantage is that the width of the display device will diminish the resolution for each series as more series are plotted in parallel. As with TSPLOT, we can increase the display width through the PROFILE paragraph. Alternatively, we can limit the number of series that are displayed. It is recommended that no more than three or four variables be displayed at one time, depending on the width of the display device. There is a caution that accompanies this recommendation. Since the width of any plot is a function of the number of plots being displayed, the width and resolution of the display of the time plot of the same series will be different if it is plotted alone, with one other series, or with more series. This problem can be resolved easily.

Suppose we find that the resolution associated with the parallel display of three series is what we want, but we need to plot five different series. The easiest "solution" to this problem is to use TPLOT with any three of the series, then use TPLOT again with the remaining two series and one of the first three plotted. By artificially "padding" the total number of series, we have achieved the desired resolution for all plots that are displayed.

## 3.3 Altering Basic Displays

The plotting paragraphs of the SCA System are designed so that we only need to specify the names of the variables involved in order to generate a plot. While the default options taken by a paragraph are sufficient in most situations, other features are available for specific needs. This section explains and illustrates many of these features.

### 3.3.1 Symbols for scatter plots

The SCA System displays a symbol to represent a data point; that is, a specific realization of a coordinate pair of values. Symbols are not connected to others in any way. Specific symbols used are dependent upon the paragraph or those defined by the user.

**PLOT paragraph**

When a single pair of variables is plotted in a frame, the default symbol displayed at any coordinate is '*'. If two or more data points are required to be displayed at the coordinate, the following symbol is used:

2, 3, . . . , 9 occurrences : '2', '3', . . . , '9', respectively;
10, . . . , 35 occurrences : 'A', . . . , 'Z', respectively;
36 or more occurrences : '#'

In lieu of the symbol '*', we can define a variable of of symbolic "tags" that are to be used in the display for each data pair. This "tagging" information can be useful to keep track of occurrences that share some common trait. For example, in our plots of BODYFAT

against TRICEPS and THIGHS we may wish to distinguish individuals based on age (under 20, over 20) or race. We may also wish to "tag" data recorded according to, or otherwise follow, a periodic pattern.

The number of symbols contained in the "tagging" variable must be the same as the number of data points displayed. The coordinate pair is represented by the first symbol of the tagging variable, the second pair by the second symbol, and so on. The distinct "tags" that are available are the symbol '*', the values 2-9, and the letters A-Z. The SCA System makes the following association between the value in the tagging variable and the symbol that is displayed:

| If the value of tagging variable is | the symbol displayed is |
|---|---|
| 1 | * |
| 2, 3, ..., 9 | 2, 3, ..., 9 |
| 10, 11, ..., 35 | A, B, ..., Z |

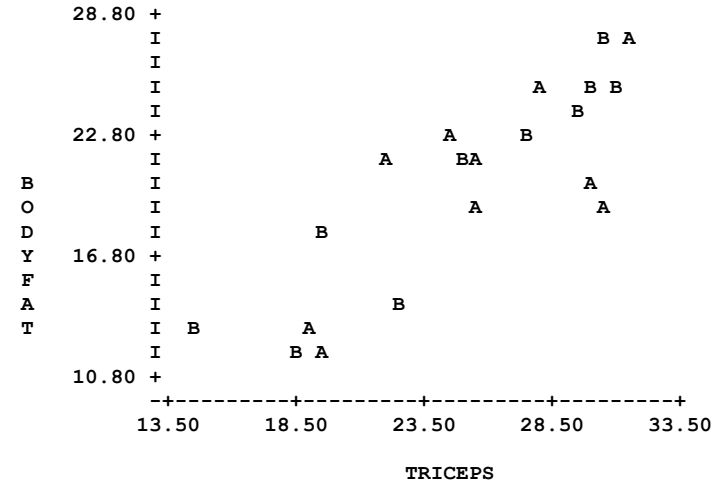
Values may be repeated within the tagging variable. This variable must be created outside of the PLOT paragraph, either by using the INPUT paragraph (see Chapter 2) or by the GENERATE or other data editing paragraphs (see Appendix B).

To illustrate the creation and use of tags, the scatter plot of BODYFAT against TRICEPS of Section 1 will be displayed. The symbol 'A' will be used to represent the first 10 cases, and the symbol 'B' will be used to represent the last 10 cases. First, we will generate a variable of tags, TAGS, using the GENERATE paragraph. The number 10 (associated with 'A') is assigned to the first 10 values and 11 (associated with 'B') is assigned to the next 10 values.

```
-->GENERATE  TAGS.  NROWS ARE 20.  VALUES ARE 10 FOR 10, 11 FOR 10.

 THE SINGLE PRECISION VARIABLE  TAGS  IS GENERATED
```

We now use the TAGSET sentence within the PLOT paragraph.

```
-->PLOT  BODYFAT, TRICEPS.  TAGSET IS TAGS.
```



The tags show that the levels of bodyfat and triceps do not seem to be affected by the order in which measurements were taken (or recorded).

## MPLOT paragraph

When multiple pairs of variables are displayed on the same frame, the symbol 'A' represents the coordinate of a value from the first pair of variables, 'B' represents the coordinate of a value from the second pair of variables, and so on. The symbol '*' is used to represent any overlapped data points. No distinctions are made regarding which data points overlap. For example, the '*' symbol will be displayed if two coordinates of values from the first pair of variables are the same, if two coordinates of values from the second pair of variables are the same, or if the coordinate of a value from the first pair of variables is the same as the coordinate of a value from the second pair of variables. Hence we may need to employ some caution in interpreting the '*' symbol should it appear.

We can designate a specific symbol for each pair of variables, as we did in the MPLOT examples of Section 3.1. The SYMBOLS sentence is used for this purpose.

### 3.3.2 Scatter plot displays

Scatter plots are displayed with a horizontal X-axis and vertical Y-axis. The name of the variable of each axis is also displayed. In the case of multiple plots on the same frame, the name of the last X and Y variables are displayed.

#### Display layouts

Three types of display layouts are available. The type of layout may be changed by using the LAYOUT sentence. Available layouts (and associated keywords) are:

| | | |
|---|---|---|
| L-shape (L) | -- | Axes form to resemble the letter 'L'. This is the default. |
| Box-type (BOX) | -- | 'L' above is "completed" to resemble a rectangle. |
| Grid-type (GRID) | -- | Cross hatch markings are included in a box-type layout. Markings occur at tic-marks. |

#### Titles for plots

A title can be included with any plot. The TITLE sentence is included in the paragraph with the desired title. The title may be 72 characters or less and must be enclosed in a pair of apostrophes (‘),

To illustrate a box-type and grid-type layout, and the use of titles, the scatter plot BODYFAT against TRICEPS will be shown in both forms.

-->PLOT BODYFAT, TRICEPS. LAYOUT IS BOX. TITLE IS @
--> ' SCATTER PLOT OF BODYFAT VS TRICEPS WITH A BOX-TYPE LAYOUT '.

```
          SCATTER PLOT OF BODYFAT VS TRICEPS WITH A BOX-TYPE LAYOUT

           13.50     18.50     23.50     28.50     33.50
            -+---------+---------+---------+---------+-
       28.80 +                                       +
             I                                 * *     I
             I                                       I
             I                               *   * *   I
             I                                 *     I
       22.80 +                       *     *         +
             I                   *     **            I
   B         I                                 *     I
   O         I                         *         *   I
   D         I             *                         I
   Y   16.80 +                                       +
   F         I                                       I
   A         I                   *                   I
   T         I  *         *                          I
             I           * *                         I
       10.80 +                                       +
             -+---------+---------+---------+---------+-
           13.50     18.50     23.50     28.50     33.50

                             TRICEPS
```

-->PLOT BODYFAT, TRICEPS. LAYOUT IS GRID. TITLE IS @
--> ' SCATTER PLOT OF BODYFAT VS TRICEPS WITH A GRID-TYPE LAYOUT '.

```
          SCATTER PLOT OF BODYFAT VS TRICEPS WITH A GRID-TYPE LAYOUT

           13.50     18.50     23.50     28.50     33.50
            -+---------+---------+---------+---------+-
       28.80 +          I         I         I         +
             I          I         I         I   * *   I
             I          I         I         I         I
             I          I         I        *I  * *    I
             I          I         I         I *       I
       22.80 +----------I---------I-*-----*-I---------+
             I          I      *  I  **     I         I
   B         I          I         I         I  *      I
   O         I          I         I   *     I   *     I
   D         I          I *       I         I         I
   Y   16.80 +----------I---------I---------I---------+
   F         I          I         I         I         I
   A         I          I       * I         I         I
   T         I  *       I*        I         I         I
             I          * *       I         I         I
       10.80 +          I         I         I         +
             -+---------+---------+---------+---------+-
           13.50     18.50     23.50     28.50     33.50

                             TRICEPS
```

### 3.3.3 Symbols for plots over time

As in the case of scatter plots, the SCA System displays a symbol to represent a data point, and symbols are not connected to others in any way. Specific symbols used are dependent upon the paragraph or those defined by the user.

**TSPLOT and TPLOT paragraphs**

The default set of symbols used for data in the TSPLOT paragraph is '1', '2', . . . , '9', '0'. This set is repeated as needed. The default symbol to designate a data point in the TPLOT paragraph is 'X'. As in the case of scatter plots, we can provide an alternative set of symbols. However, the purpose of providing an alternative set of symbols here is slightly different than for scatter plots. Symbols we provide here are used specifically to display the periodic occurrences of data. As a result, we do not have to specify a tagging set for all points, only for the number of points that comprise a period. The symbol set is then repeated over and over until the data set to be plotted is exhausted. For example, if the data in a series represents daily observations recorded on a weekly basis, then we may specify seven distinct symbols. As a consequence, when the plots are displayed all "Mondays" will have the same symbol, all "Tuesdays" will have the same symbol, and so on. Symbols are limited to 0 to 9 and A to Z, hence a maximum period of 36.

For our convenience a default set of symbols is generated automatically in the TSPLOT paragraph that corresponds to the value specified in the SEASONALITY sentence. The default symbol set generated is the first i symbols from

'1', '2', . . . , '9', '0', 'A', 'B', . . . , 'Z'

where i is the value in "SEASONALITY IS i". Hence the default set generated for the examples of TSPLOT presented in Section 2 should be

'1', '2', . . . , '9', '0', 'A', 'B'.

This sequence of symbols would be repeated in the display. However, this default symbol set as overridden by our inclusion of the sentence

SYMBOL IS '*'

The plot of PGAS over time is now shown with the System generated default set of symbols.

-->TSPLOT PGAS. SEASONALITY IS 12.

```
 TIME SERIES PLOT FOR THE VARIABLE PGAS

       ------------+-----------+-----------+-----------+-----------+-----------+--------
690.00 +              3456                                                              +
       I                  78                                                            I
       I                    90AB                                                        I
       I                        12                                                      I
630.00 +             2               789                                                +
       I            1             3      0A                                            I
       I    56789  B                      B                                           I
       I   4     0A                4 6                                                 I
570.00 +                                    1     789                                 +
       I  3                          5        2   6   0A                               I
       I                                       3 5      B    5           67            I
       I 2                                              1 34 67         5  8  AB       I
510.00 +                                      4         2     8 0A         90         +
       I                                                    9  B   4                   I
       I1                                                      1                     I
       I                                                        23                   I
450.00 +                                                                              +
       ------------+-----------+-----------+-----------+-----------+-----------+--------
                  12          24          36          48          60          72
```

## MTSPLOT and MTPLOT paragraphs

As in the case of multiple scatter plots on the same frame, the symbol ‘A’ is used for all data points from the first series, `B' for the second series, and so on unless the user specifies otherwise. When a symbol set is specified, the symbols replace ‘A’, ‘B’, and so on; but cannot be used to indicate observations of the same period (e.g., day or month) as in the TSPLOT and TPLOT paragraphs.

### 3.3.4 Tic marks, seasonality

Tic marks appear along the time axis at specific multiples. The default multiple for the TSPLOT and MTSPLOT paragraphs is 10; that is at 10, 20, 30, . . . . The default multiple for the TPLOT and MTPLOT paragraphs is 5.

It is also assumed that the index for the first observation of a series is 1. However, we may wish to specify a different multiple for the tic mark, as well as a beginning index value. The former is useful when plotting periodic data such as hourly (24), weekly (7), or monthly (12) observations (as we did in Section 3.2 and above). The SEASONALITY sentence provides a new multiple for the tic marks.

The latter specification is useful in those cases when the data set being plotted does not begin at the start of a period. For example, if a series is of monthly observations, we may want tic marks every December. If the data actually begins in March, then we want to associate the first observation with the number 3. In such a case the initial index for the data to be plotted may be specified as a second value in the SEASONALITY sentence. For example,

SEASONALITY IS 12, 3.

indicates a periodicity of 12, but the first data point is the 3rd observation in a period (e.g., March). If the SPAN sentence is used in conjunction with the SEASONALITY sentence, the System will determine tic-marks and symbols as if the entire data set is to be plotted, but only display the plot of the specified span. This was evident in the TSPLOT of PGAS on page 3.11. For example, if we had entered

```
-->TSPLOT PGAS. SEASONALITY IS 12. SPAN IS 39, 65.
```

then the plot displayed would have tic-marks at 48 and 60, and the symbol for the first observation plotted would be '3'.

**Remark**: The SEASONALITY sentence is a replacement of the older sentence, TIC-MARK. In the event your version of the SCA System does not recognize the SEASONALITY sentence, it is likely you have an older version of the System. In such a case, please substitute TIC-MARK for SEASONALITY.

## 3.4 Shewhart Control Charts

A control chart is a statistical technique developed by W.A. Shewhart (1931, 1939) for studying and monitoring a process. The basic idea behind a control chart is that in any process there will always exist inherent or natural variation, which is the cumulative effect of many small causes of variation about which little can be done besides changing the process. Besides these random causes, there are relatively large variations attributable to special causes (such as differences in materials, machines, operators, etc.). A process operating in the presence of special (or assignable) causes is said to be out of statistical control.

We would like to quickly detect any change in the process output which is due to a special cause so that the situation can be investigated and corrective action can be taken. It can be expensive and inefficient to adjust a process because of a suspected special cause when the variation is in fact due to ordinary random variation. It is thus important to be able to distinguish between the random variation in the process and the variation that occurs due to special causes, and a control chart can be used to assist in such situations.

### 3.4.1 Example, forged piston rings data

To illustrate the use of Shewhart control charts, we consider a data set from Montgomery (1985, p.176) regarding measurements of piston rings in a forging process. The process is believed to be in a state of control. Twenty-five samples, each of size five, are taken from this process. The data are displayed in Table 3. The first through fifth observation of each sample are stored in the SCA workspace under the labels PISTON1 through PISTON5, respectively. Also listed in Table 3 is the mean and range of each sample.

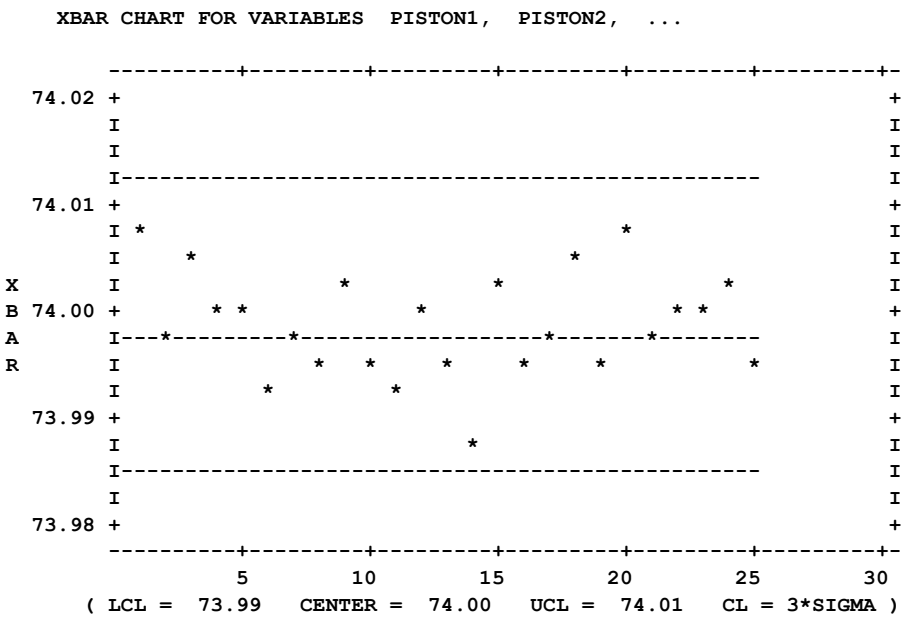
**Table 3 Forged piston ring data**

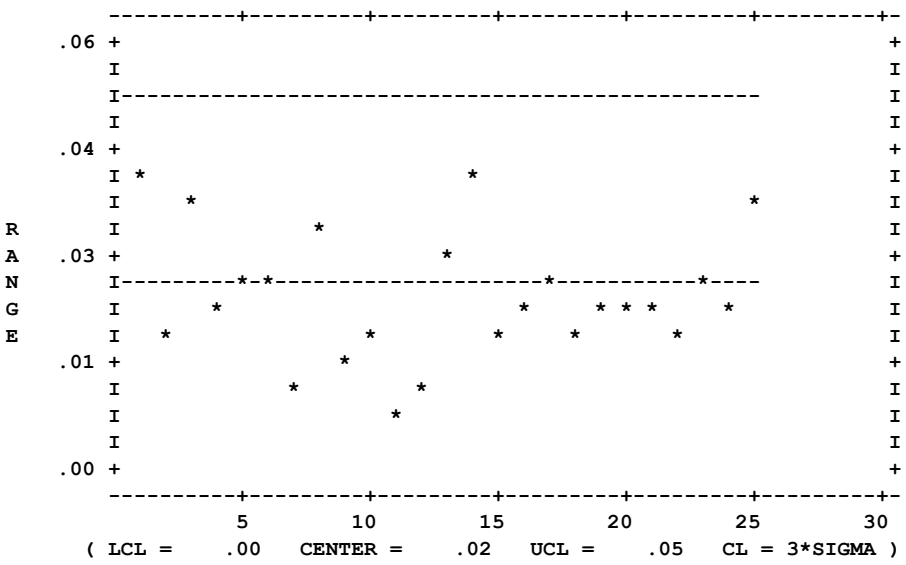
| *Sample Number* | *Observations* PISTON1 | PISTON2 | PISTON3 | PISTON4 | PISTON5 | *Sample mean* | *Sample ranges* |
|---|---|---|---|---|---|---|---|
| 1 | 74.030 | 74.002 | 74.019 | 73.992 | 74.008 | 74.010 | 0.038 |
| 2 | 73.995 | 73.992 | 74.001 | 74.011 | 74.004 | 74.001 | 0.019 |
| 3 | 73.988 | 74.024 | 74.021 | 74.005 | 74.002 | 74.008 | 0.036 |
| 4 | 74.002 | 73.996 | 73.993 | 74.015 | 74.009 | 74.003 | 0.022 |
| 5 | 73.992 | 74.007 | 74.015 | 73.989 | 74.014 | 74.003 | 0.026 |
| 6 | 74.009 | 73.994 | 73.997 | 73.985 | 73.993 | 73.996 | 0.024 |
| 7 | 73.995 | 74.006 | 73.994 | 74.000 | 74.005 | 74.000 | 0.012 |
| 8 | 73.985 | 74.003 | 73.993 | 74.015 | 73.988 | 73.997 | 0.030 |
| 9 | 74.008 | 73.995 | 74.009 | 74.005 | 74.004 | 74.004 | 0.014 |
| 10 | 73.998 | 74.000 | 73.990 | 74.007 | 73.995 | 73.998 | 0.017 |
| 11 | 73.994 | 73.998 | 73.994 | 73.995 | 73.990 | 73.994 | 0.008 |
| 12 | 74.004 | 74.000 | 74.007 | 74.000 | 73.996 | 74.001 | 0.011 |
| 13 | 73.983 | 74.002 | 73.998 | 73.997 | 74.012 | 73.998 | 0.029 |
| 14 | 74.006 | 73.967 | 73.994 | 74.000 | 73.984 | 73.990 | 0.039 |
| 15 | 74.012 | 74.014 | 73.998 | 73.999 | 74.007 | 74.006 | 0.016 |
| 16 | 74.000 | 73.984 | 74.005 | 73.998 | 73.996 | 73.997 | 0.021 |
| 17 | 73.994 | 74.012 | 73.986 | 74.005 | 74.007 | 74.001 | 0.026 |
| 18 | 74.006 | 74.010 | 74.018 | 74.003 | 74.000 | 74.007 | 0.018 |
| 19 | 73.984 | 74.002 | 74.003 | 74.005 | 73.997 | 73.998 | 0.021 |
| 20 | 74.000 | 74.010 | 74.013 | 74.020 | 74.003 | 74.009 | 0.020 |
| 21 | 73.988 | 74.001 | 74.009 | 74.005 | 73.996 | 73.996 | 0.033 |
| 22 | 74.004 | 73.999 | 73.990 | 74.006 | 74.009 | 74.002 | 0.019 |
| 23 | 74.010 | 73.989 | 73.990 | 74.009 | 74.014 | 74.002 | 0.025 |
| 24 | 74.015 | 74.008 | 73.993 | 74.000 | 74.010 | 74.005 | 0.022 |
| 25 | 73.982 | 73.984 | 73.995 | 74.017 | 74.013 | 73.998 | 0.035 |

We can obtain Shewhart control charts for the piston data be entering

```
--> SHEWHART   PISTON1, PISTON2, PISTON3, PISTON4, PISTON5
```

We obtain the two following charts


XBAR CHART FOR VARIABLES PISTON1, PISTON2, ...
74.02
74.01
74.00
73.99
73.98
XBAR
5
10
15
20
25
30
( LCL = 73.99 CENTER = 74.00 UCL = 74.01 CL = 3*SIGMA )


RANGE CHART FOR VARIABLES PISTON1, PISTON2, ...
.06
.04
.03
.01
.00
RANGE
5
10
15
20
25
30
( LCL = .00 CENTER = .02 UCL = .05 CL = 3*SIGMA )

### 3.4.2 Types of charts

We observe two separate charts; one is called XBAR ($\bar{x}$), while the other is of the RANGE (R). Each chart shows a series of points plotted in time order together with three horizontal lines. The center line of each chart represents the average level when the process is in statistical control (i.e., only random variations are present). The upper and lower control limits are chosen so that nearly all of the points will fall between the limits if the process is in statistical control. A point that falls outside of these limits is viewed as an indication that there is a special cause of variation present and that the process is out of control. Furthermore, if the points vary in a systematic or non-random way within the limits, then that would be considered as evidence that there is a special cause of variation. Examples of systematic variation include a cyclic pattern and a trend.

The general control chart can be considered as follows. Let x be a quality characteristic of interest, $\mu$ its mean value, and $\sigma$ its standard deviation, then we have

| | |
|---|---|
| Upper control limit (UCL) | $\mu + k\sigma$ |
| Center line | $\mu$ |
| Lower control limit (LCL) | $\mu - k\sigma$ |

where the multiplier k is the number of standard deviations from the center line to the control limits. The values used as the LCL, Center line and UCL are printed with each chart. The integer displayed for CL is the value of k, here 3. The most frequently used value for k is 3, and is the default value for the SHEWHART paragraph. If the computed LCL is negative, and the variable is non-negative, then 0.0 is used as the LCL.

We note that no value is outside the control limits of either of our charts for the piston ring data, nor are there any discernable systematic patterns. Hence, based on these charts, we are led to conclude the process is indeed in control. We will now discuss why these charts were used.

#### $\bar{x}$- and R-charts

For each of our 25 samples we can use the sample mean, $\bar{x}$, as a measure of the mean quality level of the sample; and we can use the sample range,

R = maximum - minimum value

as a measure of the sample's variation. The Shewhart charts calculated above, known as **$\bar{x}$-R control charts**, display the computed values of $\bar{x}$ and R for all the samples. The center line of the$\bar{x}$ chart is the average of all computed sample means; that is $\bar{x}$ . The center line of the R chart is the average of all sample ranges; that is $\bar{R}$ . The control limits for the charts are

| | $\bar{x}$-*chart* | *R-chart* |
|---|---|---|
| UCL | $\bar{x} + k\left(\frac{\bar{R}}{d_2\sqrt{n}}\right)$ | $\bar{R} + k\left(\bar{R}\frac{d_3}{d_2}\right)$ |
| LCL | $\bar{x} - k\left(\frac{\bar{R}}{d_2\sqrt{n}}\right)$ | $\bar{R} - k\left(\bar{R}\frac{d_3}{d_2}\right)$ |

where typically k is 3. The term $\bar{R}/(d_2\sqrt{n})$ is an estimate of the standard deviation of $\bar{x}$, and $\bar{R}(d_3/d_2)$ is an estimate of the standard deviation of $\bar{R}$. The coefficients $d_2$ and $d_3$ are functions of the sample size n and are computed by the SCA System automatically. These values can also be found in a number of books, for example, Montgomery (1985) and Duncan (1974). In the R-chart, if the LCL is negative, it is set to 0.

When examining an $\bar{x}$-R chart, we should first inspect the R-chart since the control limits on the $\bar{x}$-chart depend on the average process variability ($\bar{R}$). Consequently the $\bar{x}$-chart limits will not have much meaning unless the process variability is in control. We should not attempt to interpret the $\bar{x}$-chart if the R-chart indicates that the variability is out of control.

In addition, the validity of the $\bar{x}$- and $\bar{R}$-charts is dependent on the assumption of normality. The R-chart is more sensitive to departures from normality than the $\bar{x}$-chart. If normality does not hold, then a process may appear out of control even though in fact that might not be the case.

## $\bar{x}$- and S-charts

When the sample size is greater than 10 or so, the range is not statistically efficient for estimating the standard deviation. Instead we should use the sample standard deviation, s, in place of R and replace the $\bar{x}$-R chart by an $\bar{x}$-S chart. The center line and control limits under such a situation are listed below:

| | $\bar{x}$-*chart* | *S-chart* |
|---|---|---|
| UCL | $\bar{x} + k\left(\frac{\bar{s}}{c_4\sqrt{n}}\right)$ | $\bar{s} + k\left(\frac{\bar{s}\sqrt{1-c_4^2}}{c_4}\right)$ |
| Center line | $\bar{x}$ | $\bar{s}$ |
| LCLx | $\bar{x} - k\left(\frac{\bar{R}}{d_2\sqrt{n}}\right)$ | $\bar{s} - k\left(\frac{\bar{s}\sqrt{1-c_4^2}}{c_4}\right)$ |

Similar to $d_2$ and $d_3$ of the $\bar{x}$-R chart, the constant $c_4$ is also a function of n and is computed by the SCA System automatically. The values of $c_4$ can also be found in Montgomery (1985) and Duncan (1974). In the S-chart, if the LCL is negative, it is also set to 0.

We can request the creation of $\bar{x}$-S control charts instead of the default by including the TYPE sentence in the SHEWHART paragraph. To obtain the $\bar{x}$-S charts for the piston ring data (not displayed here) we can enter

```
SHEWHART    PISTON1, PISTON2, PISTON3, PISTON4, PISTON5.    TYPE IS XS.
```

### 3.4.3 Control guidelines

We have noted that the presence of points outside control limits of systematic patterns in the points within the control limits are indications of an assignable cause. The Western Electric Handbook (1956) gives guidelines for detecting such patterns. Generally it can be concluded that the process is out of control if

(1) there are points outside 3σ limits;

(2) 2 out of 3 consecutive points appear beyond 2σ limits;

(3) 4 out of 5 consecutive points appear beyond 1σ limit;

(4) there is a run of 7 points with the run consistently either up or down, or above or below the centerline; or

(5) any cyclic pattern, or hugging of the center line, is seen.

# SUMMARY OF THE SCA PARAGRAPHS IN CHAPTER 3

This section provides a summary of those SCA paragraphs employed in this chapter. The syntax for each paragraph is presented in both a brief and full form. The brief display of the syntax contains the most frequently used sentences of a paragraph, while the full display presents all possible modifying sentences of a paragraph. In addition, special remarks related to a paragraph may also be presented with the description.

Each SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

The paragraphs to be explained in this summary are PLOT, MPLOT, TSPLOT, MTSPLOT, TPLOT, MTPLOT, and SHEWHART

Legend (see Chapter 2 for further explanation)

v : variable name
i : integer
r : real value
w : keyword
'c' : character data (must be enclosed within single apostrophes)

## PLOT Paragraph

The PLOT paragraph is used to construct and display the scatter plot of a single pair of variables or the plots of multiple pairs of variables on separate frames, each frame having the same X and Y scaling.

## Syntax for the PLOT Paragraph

### Brief syntax

```
PLOT  VARIABLES ARE  v1, v2
```

### Full syntax

```
PLOT  VARIABLES ARE  v1, v2.                    @
      X-VARIABLES ARE  v1, v2, --- .            @
      Y-VARIABLES ARE  v1, v2, --- .            @
      TITLE IS 'c'.                             @
      SPAN IS i1, i2.                           @
      TAGSETS ARE  v1, v2, --- .                @
      RANGES ARE X(r1,r2), Y(r3,r4)             @
      LAYOUT IS w.                              @
      SIZE IS X(i1), Y(i2).                     @
      TIC-MARK IS X(i1), Y(i2).                 @
      GRID IS X(i1), Y(I2).

Required sentences:  VARIABLES, or X-VARIABLES and Y-VARIABLES
```

## Sentences Used in the PLOT Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the names (labels) of the Y (vertical) variable, v1, and X (horizontal) variable, v2. Note that when this sentence is used, the X-VARIABLE and Y-VARIABLE sentences are ignored. It is invalid to specify more than one pair of variable names in this sentence.

**X-VARIABLE sentence**

The X-VARIABLE sentence is used to specify the names of the variables to be plotted along the horizontal axis. The number of variables specified in this sentence must be the same as that in the Y-VARIABLE sentence.

**Y-VARIABLE sentence**

The Y-VARIABLE sentence is used to specify the names of the variables to be plotted along the vertical axis. The number of variables specified in this sentence must be the same as that in the X-VARIABLE sentence.

**TITLE sentence**

The TITLE sentence is used to specify the title for the plot(s). The specified title must be enclosed in a pair of apostrophes and have no more than 72 characters. The default is that no title will be displayed.

**SPAN sentence**

The SPAN sentence is used to specify the span of indices, from i1 to i2, for which the values of the co-ordinates will be plotted. The default is to plot all cases.

**TAGSETS sentence**

The TAGSETS sentence is used to specify the name(s) of variable(s) containing the "tags" to be used in plotting data. The default is none. See Section 3.3.1 for the way the values of the TAGSET variable(s) are converted to symbols. If the TAGSET sentence is used, one variable must be specified for each Y-VARIABLE specified.

**RANGES sentence**

The RANGES sentence is used to specify the upper and lower limits for the X and Y variable values to be plotted. The default are limits determined automatically by the SCA System.

**LAYOUT sentence**

The LAYOUT sentence is used to specify the layout type for the axes of the plot. The valid keywords are L for L-shape layout, BOX for box-type layout, and GRID for grid-type layout. The default layout is L-shape.

**SIZE sentence**

The SIZE sentence is used to specify the number of character units for the width of the X-axis and Y-axis. The default is 50 characters for the X-axis and 30 characters for the Y-axis.

**TIC-MARK sentence**

The TIC-MARK sentence is used to specify the intervals (in number of character units) for the printing of tic-marks on the X and Y axes. The default is 10 units for the X-axis and 5 units for the Y-axis.

**GRID sentence**

The GRID sentence is used to specify the number of tic-marks on each axis within a grid for hatch markings. This sentence can be specified only if the plot layout is GRID. The default is 1 for both X and Y.

## MPLOT Paragraph

The MPLOT paragraph is used to display the scatter plot(s) as one or more pair(s) of variables on the same frame.

## Syntax for the MPLOT Paragraph

### Brief syntax

```
MPLOT     X-VARIABLES ARE v1, v2, --- .     @
          Y-VARIABLES ARE v1, v2, --- .
```

### Full syntax

```
MPLOT     X-VARIABLES ARE v1, v2, --- .     @
          Y-VARIABLES ARE v1, v2, --- .     @
          TITLE IS 'c'.                     @
          SPAN IS i1, i2.                   @
          RANGES ARE X(r1,r2), Y(r3,r4).    @
          SYMBOLS ARE 'c1', 'c2', --- .     @
          LAYOUT IS w.                      @
          SIZE IS X(i1), Y(i2).             @
          TIC-MARK IS X(i1), Y(i2).         @
          GRID IS X(i1), Y(i2).
```

Required sentences: **X-VARIABLES** and **Y-VARIABLES**

## Sentences Used in the MPLOT Paragraph

**X-VARIABLE sentence**

The X-VARIABLE sentence is used to specify the names of the variables to be plotted along the horizontal axis. The number of variables specified in this sentence must be the same as that in the Y-VARIABLE sentence.

**Y-VARIABLE sentence**

The Y-VARIABLE sentence is used to specify the names of the variables to be plotted along the vertical axis. The number of variables specified in this sentence must be the same as that in the X-VARIABLE sentence.

**TITLE sentence**

The TITLE sentence is used to specify the title for the plot(s). The specified title must be enclosed in a pair of apostrophes and have no more than 72 characters. The default is that no title will be displayed.

**SPAN sentence**

The SPAN sentence is used to specify the span of indices, from i1 to i2, for which the values of the co-ordinates will be plotted. The default is all cases.

**RANGES sentence**

The RANGES sentence is used to specify the upper and lower limits for the X and Y variable values to be plotted. Default is all the values.

**SYMBOLS sentence**

The SYMBOLS sentence is used to specify the SYMBOLS that will represent co-ordinates of different pairs of variables. If no set of symbols is specified, 'A' represents co-ordinates of the first pair of variables, and `B' represents co-ordinates of the second pair, etc.

**LAYOUT sentence**

The LAYOUT sentence is used to specify the layout type for the axes of the plot. The valid keywords are L for L-shape layout, BOX for box-type layout, and GRID for grid-type layout. The default layout is L-shape.

**SIZE sentence**

The SIZE sentence is used to specify the number of character units for the width of the X-axis and Y-axis. The default is 50 characters for the X-axis and 30 characters for the Y-axis.

**TIC-MARK sentence**

The TIC-MARK sentence is used to specify the intervals (in number of character units) for the printing of tic-marks on the X and Y axes. The default is 10 units for the X-axis and 5 units for the Y-axis.

**GRID sentence**

The GRID sentence is used to specify the number of tic-marks on each axis within a grid for hatch markings. This sentence can be specified only if the plot layout is GRID. The default is 1 for both X and Y.

## TSPLOT, TPLOT Paragraphs

The TSPLOT paragraph is used to specify the horizontal time plot of one or more series in separate frames. The TPLOT paragraph is used to display the vertical time plot of one or more series in separate, parallel frames on the display device.

## Syntax of the TSPLOT or TPLOT Paragraph

### Brief syntax

```
TSPLOT      VARIABLES ARE v1, v2, --- .

      or

TPLOT       VARIABLES ARE v1, v2, --- .
```

### Full syntax

```
TSPLOT          VARIABLES ARE v1, v2, --- .          @
(or TPLOT)      SEASONALITY IS i1, i2.               @
                SPAN IS i1, i2.                      @
                TITLE IS 'c'.                        @
                SYMBOLS ARE 'c1', 'c2', --- .        @
                RANGE IS r1, r2.

Required sentence:  VARIABLE(S)
```

## Sentences Used in the TSPLOT or TPLOT Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the names of the series to be plotted.

**SEASONALITY sentence**

The SEASONALITY sentence is used to specify the multiple (i1) at which a tic-mark is printed along the time axis and the value of the index (i2) of the first observation. The default value of i1 is 10 and of i2 is 1 (or the lower limit of the SPAN sentence if this sentence is specified). Specification of a seasonality will also generate a default set of symbols (unless overwritten by the SYMBOLS sentence). See Section 3.3 for a further explanation. Note SEASONALITY replaces the sentence TIC-MARK of older versions of the SCA System.

**SPAN sentence**

The SPAN sentence is used to specify the span of time indices, from i1 to i2, for which values will be plotted. Default is that all observations in the series will be used.

**TITLE sentence**

The TITLE sentence is used to specify the title for the plot(s). The specified title must be enclosed in a pair of apostrophes and have no more than 72 characters. The default is that no title will be displayed.

**SYMBOLS sentence**

The SYMBOLS sentence is used to specify a sequence of symbols repeated in the plot. The default symbols used are the first i characters of the set '1', '2', ... '9', '0', 'A', 'B', ..., 'Z' where i is the distance between axis tic-marks. The value of i corresponds to the SEASONALITY specified (default is i=10). Specification of the SYMBOLS sentence overrides this default set of symbols.

**RANGES sentence**

The RANGES sentence is used to specify the upper and lower limits for the series to be plotted. The default are limits determined automatically be the SCA System.

## MTSPLOT, MTPLOT Paragraphs

The MTSPLOT paragraph is used to display the time plot of more than one series on the same horizontal frame. The MTPLOT paragraph is used to display the time plot of more than one series on the same vertical time frame.

## Syntax for the MTSPLOT or MTPLOT Paragraph

### Brief syntax

```
MTSPLOT     VARIABLES ARE v1, v2, --- .

            or

MTPLOT      VARIABLES ARE v1, v2, --- .
```

### Full syntax

```
MTSPLOT      VARIABLES ARE v1, v2, --- .        @
(or MTPLOT)  SEASONALITY IS i1, i2.             @
             SPAN IS i1, i2.                    @
             TITLE IS 'c'.                      @
             SYMBOLS ARE 'c1', 'c2', --- .      @
             SPAN IS i1, i2.                    @
             RANGE IS r1, r2.

        Required sentence: VARIABLES
```

## Sentences Used in the MTSPLOT or MTPLOT Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the names of the series to be plotted.

**SEASONALITY sentence**

The SEASONALITY sentence is used to specify the multiple (i1) at which a tic-mark is printed along the time axis and the value of the index (i2) of the first observation. The default value of i1 is 10 and of i2 is 1 (or the lower limit of the SPAN sentence if this sentence is specified). See Section 3.3 for a further explanation. Note SEASONALITY replaces the sentence TIC-MARK of older versions of the SCA System.

**SPAN sentence**

The SPAN sentence is used to specify the span of time indices, from i1 to i2, for which values will be plotted. Default is that all observations in the series will be used.

**TITLE sentence**

The TITLE sentence is used to specify the title for the plot(s). The specified title must be enclosed in a pair of apostrophes and have no more than 72 characters. The default is that no title will be displayed.

**SYMBOLS sentence**

The SYMBOLS sentence is used to specify the SYMBOLS for distinguishing that different series. If this sentence is omitted, 'A' represents the first series, `B' the second, etc.

**RANGES sentence**

The RANGES sentence is used to specify the upper and lower limits for the series to be plotted. The default are limits determined automatically by the SCA System.

## SHEWHART Paragraph

The SHEWHART paragraph is used to graphically display the level (mean) and the dispersion (range or standard deviation) for samples taken from a process. Either the $\bar{x}$-R or $\bar{x}$-S chart may be displayed. These charts can be used to monitor a process and indicate whether it is in a state of statistical control.

## Syntax for the SHEWHART Paragraph

### Brief syntax

```
SHEWHART  VARIABLES ARE  v1, v2, ---. @
          TYPE IS w.     @
          TITLE IS 'c'.

Required sentence:  VARIABLES
```

### Full syntax

```
SHEWHART  VARIABLES ARE  v1, v2, ---.      @
          TYPE IS w.                       @
          TITLE IS 'c'.                    @
          CRITERIA IS r.                   @
          LIMITS ARE r1, r2, ---.          @
          SPAN IS i1, i2.                  @
          RANGE IS r1, r2.                 @
          TIC-MARK IS i1, i2.

Required sentence:  VARIABLES
```

## Sentences Used in the SHEWHART paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the names of the variables to be used in the construction of the charts. A single variable usually represents the observations at a particular time or order of sequence within each sample. This is a required sentence.

**TYPE sentence**

The TYPE sentence is used to specify the type of Shewhart chart to be displayed. Valid keywords are:

| | |
|---|---|
| X | : all observations are displayed |
| XBAR | : only the $\bar{x}$-chart is displayed |
| R | : only the R-chart is displayed |
| S | : only the S-chart is displayed |
| XR | : both the $\bar{x}$- and R-charts are displayed |
| XS | : both the $\bar{x}$- and S-charts are displayed |

The default is XR.

**TITLE sentence**

The TITLE sentence is used to specify the title for the chart. The specified title must be enclosed in a pair of apostrophes and have no more than 72 characters. The default is that no title will be displayed.

**CRITERIA sentence**

The CRITERIA sentence is used to specify the multiplier for the computation of control limits. The control limits are computed based on

mean level ± (control criteria value) x (estimate of standard deviation)

The default control criteria value is 3.0.

**LIMITS sentence**

The LIMITS sentence allows the user to specify fixed upper and lower control limits for the Shewhart charts. The first two values are control limits for the X, and the third and fourth values, if specified, are control limits for the range or the standard deviation. The default control limits are computed by the System internally.

**SPAN sentence**

The SPAN sentence is used to specify the span of indices, from i1 to i2, for which the chart is to be displayed. Default is all observations.

**RANGES sentence**

The RANGES sentence is used to specify the upper and lower values for the display of the charts. The default values are determined by the System internally.

**TIC-MARK sentence**

The TIC-MARK sentence is used to specify the multiple (i1) at which a tic-mark is printed along the time (vertical) axis and the value of the index (i2) of the first observation. The defaults are 5 and 1.

## REFERENCES

*Commodity Year Book* (1986). New York: Commodity Research Bureau.

Duncan, A.J. (1974). *Quality Control and Industrial Statistics*, 4th edition. Homewood, IL: Richard D. Irwin, Inc.

Montgomery, D.C. (1985). *Introduction to Statistical Quality Control*, New York: Wiley.

Neter, J., Wasserman, W., and Kutner, M.H. (1983). *Applied Linear Regression Models*, Homewood, IL: Richard D. Irwin, Inc.

Shewhart, W.A. (1931). *Economic Control of Quality Manufactured Product*, Milwaukee, WI: American Society for Quality Control.

Shewhart, W.A. (1939). *Statistical Methods from the Viewpoint of Quality Control*, W.E. Deming, ed., Washington, DC: Graduate School, Department of Agriculture.

Western Electric Company, Inc. (1956). *Statistical Quality Control Handbook*, Indianapolis, IN: I.D.C. Commercial Sales, Western Electric Co.

# CHAPTER 4

# DESCRIPTIVE STATISTICS AND CORRELATION

In the previous chapter we illustrated some visual descriptions of data in the form of plots. We may also use various statistical measures to describe the data at hand. In this manner, a large set of data can be reduced to a more manageable, and possibly more meaningful, set of values. The resultant descriptive measures can go hand in hand with plots to provide a better understanding of the data we wish to analyze or explore.

This chapter provides information on the capabilities of the SCA System that may be used to generate such descriptive information. Included in these capabilities are basic descriptive statistics of one or more data sets and subgroups, measures of correlation, and the visual display of descriptive information.

## 4.1 Summary Statistics of Data

The DESCRIBE paragraph may be used to calculate various sample statistics for one or more sets of data. To illustrate the use of the DESCRIBE paragraph, we shall first consider a data set consisting of the 1970 population (in millions) of the 50 states. The data are listed in Table 1, and are assumed to be in the SCA workspace under the variables name USPOP.

**Table 1 1970 population (in millions) of the 50 United States.**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3.44 | .30 | 1.77 | 1.92 | 19.95 | 2.21 | 3.03 | .55 | 6.79 | 4.59 |
| .77 | .71 | 11.01 | 5.19 | 2.83 | 2.25 | 3.22 | 3.64 | .99 | 3.92 |
| 5.69 | 8.88 | 3.81 | 2.22 | 4.68 | .69 | 1.48 | .49 | .74 | 7.17 |
| 1.02 | 18.24 | 5.08 | .62 | 10.65 | 2.56 | 2.09 | 11.79 | .95 | 2.59 |
| .67 | 3.92 | 11.20 | 1.06 | .44 | 4.65 | 3.41 | 1.74 | 4.42 | .33 |

To obtain descriptive statistics for this variable, we may simply enter

```
-->DESCRIBE  USPOP

VARIABLE     NAME   IS    USPOP
NUMBER OF OBSERVATIONS       50
NUMBER OF MISSING VALUE       0

                  STATISTIC      STD. ERROR     STATISTIC/S.E.

MEAN                4.0472           .6123          6.6103
VARIANCE           18.7430
STD DEVIATION       4.3293
C.  V.              1.0697
SKEWNESS            1.9335           .3366
KURTOSIS            3.7093           .6619
```

```
                    QUARTILE

MINIMUM               .3000
1ST QUARTILE          .9300
MEDIAN               2.5900
3RD QUARTILE         4.6350
MAXIMUM             19.9500

                      RANGE

MAX - MIN           19.6500
Q3  - Q1             3.7050
```

The sample statistics listed in Table 2 are calculated and displayed. Most of the computed statistics can be retained in the SCA workspace under variable names that we may specify. The HOLD sentence must also be specified in the DESCRIBE paragraph for this purpose. The HOLD sentence is described with the syntax information for the DESCRIBE paragraph at the end of this Chapter. Also listed in Table 2 are the keywords that are used in the HOLD sentence to denote the computed statistic.

**Table 2 Statistics computed in the DESCRIBE paragraph**
(Keywords related to the HOLD sentence are also provided)

| Statistic (Keyword) | Mathematical representation or comment |
|---|---|
| Number of observations (NCASES) | n |
| Number missing (NMISS) | Number of missing observations |
| Sample mean (MEAN) | $\sum x_i / n$ |
| Sample median (MEDIAN) | the value equaled or exceeded by exactly one half the values of the data set. In the case of a central pair of data, the average of these values is displayed. |
| Sample variance (VARIANCE) | $\sum (x_i - \bar{x})^2 / n$ |
| Sample standard deviation (STD) | $\left\{\sum (x_i - \bar{x})^2 / (n-1)\right\}^{1/2}$ |
| Coefficient of variation (CV) | sample standard deviation divided by sample mean |
| Minimum (MINIMUM) | smallest value of the data set |
| Maximum (MAXIMUM) | largest value of the data set |
| Sample range (RANGE) | difference between maximum and minimum |
| First quartile (Q1) | sample 25th percentile value |
| Third quartile (Q3) | sample 75th percentile value |
| Sample skewness (SKEWNESS) | $\left(\sum (x_i - \bar{x})^3 / n\right) / STD^3$ |
| Sample kurtosis (KURTOSIS) | $\left(\sum (x_i - \bar{x})^4 / n\right) / STD^4$ |

### 4.1.1 Exploratory data analysis plots

In addition to the plots described in the previous chapter, the DESCRIBE paragraph can provide us with two other visual displays of one or more data sets. The DESCRIBE paragraph can generate a stem-and-leaf display and a box (and whisker) plot, two displays for exploratory data analysis (see Tukey, 1977). A histogram is another useful tool for exploratory data analysis. Its use is discussed in Chapter 5.

#### Stem-and-leaf display

A stem-and-leaf display is a presentation of the distribution of the values of a data set. This presentation is in columnar form, and provides a quick overview of both the distribution of the data and actual data values. The stem-and-leaf display has two parts: the stem, a column of equally spaced values ranging from the minimum value to the maximum value of the data; and the leaves, row bar extensions from the column that provide information on both data values and the frequency of their occurrence.

Data values represented in a stem-and-leaf display are reduced to pairs of integers. The first integer is used for the columnar stem. The second single digit is used in the row leaf. For example, consider the data of USPOP. We see from the output of the DESCRIBE paragraph that the data ranges from a minimum of 0.3 to a maximum of 19.95. We can "reduce" the fifty values to fifty pairs of integers. The first is the integer portion of the value ( 0, 1, 2, . . . , 19 ). The second single digit is the "tenths" unit after rounding. For example, the first four values of Table 1, (3.44, 0.30, 1.77, and 1.92) can be presented as the integer pairs

( 3, 4), ( 0, 3), ( 1, 8), and ( 1, 9).

The first integer of these pairs is given on the columnar stem. The leaves of pairs having the same "stem" (first value) are ordered and displayed along the row "leaf". As a result we are provided with both a picture of the distribution of the data, and we can also roughly "see" the data values (to the approximation provided).

The logical sentence STEMLEAF must be included in the DESCRIBE paragraph to obtain a stem-and-leaf display. The display is provided in as concise, yet readable, fashion as possible. In some cases, to reduce the amount of display to a meaningful amount, not all stem values are listed along the column. However, the actual stem and leaf values can be inferred from the display. In the case of USPOP, the actual output that would be provided for the stem-and-leaf display is

```
STEM-AND-LEAF DISPLAY FOR THE VARIABLE USPOP


  0  .3345667777890015789
  2  .122366802446899
  4  .4677127
  6  .82
  8  .9
 10  .6028
 12  .
 14  .
 16  .
 18  .2
 20  .0
```

We see that not all stem values are part of the column, only the even values from 0 to 20, inclusive, are listed. However, we can approximate all values by looking at the leaf rows. In the row with stem value 6, the leaf values are 8 and 2. We know that leaf values are presented in order. Then, the 8 must correspond to (6, 8) and the 2 must correspond to (7, 2). We can see from Table 1 that the actual data values are 6.79 and 7.17. However, the interpretation of the value of stem row "8" is unclear. The stem-leaf pair could either be (8, 9) or (9, 9). The actual value is 8.88. Similar questions arise for stem rows "18" and "20". However, since the purpose of a stem-and-leaf display is to provide summary information, potential different interpretations of stem-leaf values are relatively unimportant.

Clearly, the stem and leaf values displayed will not always be the integer portion and "tenths" value of a set of data. The integer pairs are dependent on the magnitude of the values of the data set and the range encompassed by them. For example, if a data set ranges from 3180 to 7455 then the stem-leaf pair (50, 2) can represent a value between 5015 and 5025. Similarly, the same pair can represent a value between .5015 and .5025 if the data ranges from 0 to 1.

## Box plots

A box plot (or box-and-whisker plot, see Tukey 1977) is a useful display of the relative location and spread of the values of a data set. To illustrate a box plot, the following plot is displayed if we include the logical sentence BOXPLOT in the DESCRIBE paragraph for USPOP.

```
BOX-PLOT FOR THE SPECIFIED VARIABLE(S)

     19.95  .    E
            .
            .    E
            .
            .
            .
            .
            .
            .    E2
            .    O2
            .
            .    X
            .    |
            .    |
            .    |
            .  --+--
            .  |   |
            .  | * |
            .  --+--
      .30   .    X
```

Part of this plot consists of a rectangular box that contains the symbol '*'. This symbol depicts the sample median, here 2.59. The upper and lower "sides" of the box indicate the upper and lower hinges of the data. Hinge values are very close to the 75th and 25th percentile values of the data, .93 and 4.635, respectively. A more precise description of a hinge can be found in Tukey (1977). The 'X' symbols indicate the location of the actual data values closest to, but within, the upper and lower inner fences. An inner fence occurs at a point one step beyond a hinge, where a step is defined as 1.5 of the length between hinges. If we use the percentile approximation for the hinge values, we see the upper and lower fences for the USPOP data are around 10.2 and -4.7, respectively. The closest data values to these fences, but contained by them, are 8.88 and 0.30, as indicated by the 'X' symbols.

An outer fence occurs one step beyond an inner fence. Data points that fall outside of fences may be outliers or extreme points of the data set. Data values that occur between the inner and outer fences may be spurious or outliers. These values are displayed with the symbol 'O'. Data values that are beyond the outer fence are extreme, and merit our attention. These values are displayed with the symbol 'E'. If two or more data points have nearly the same value, the number of values "in the area" is indicated next to the 'O' or 'E'.

In the USPOP data set we have two values between the upper inner fence and the upper outer fence. This is indicated by the display of 'O2', and represent the values 10.65 and 11.01. There are four values that fall outside the upper outer fence: 11.20, 11.79, 18.24, and 19.95 . No value occurs below the lower inner fence.

### 4.1.2 Descriptive statistics for more than one data set

We can compute and display descriptive statistics for more than one data set in two ways. We may use the DESCRIBE sentence for each data set separately, or we may specify more than one variable in the DESCRIBE sentence. If more than one variable are specified, then we obtain a separate descriptive summary for each variable. Separate stem-and-leaf displays are also presented. However, box plots are displayed on a single frame in the same order as used in the specification of variables.

To illustrate the display for more than one variable, we will use data of student scores on authoritarianism and social status strivings. These data are found in Siegel (1956) and are listed in Table 3. The data are stored in the SCA workspace under the names AUTHORIT and STATUS, respectively. The data are also used in Chapter 11. Basic descriptive statistics and box plots will be computed and displayed for these data sets.

**Table 3 Student score data**

| *Student* | *Authoritarianism score AUTHORIT* | *Social Status Striving STATUS* |
|---|---|---|
| 1 | 82.000 | 42.000 |
| 2 | 98.000 | 46.000 |
| 3 | 87.000 | 39.000 |
| 4 | 40.000 | 37.000 |
| 5 | 116.000 | 65.000 |
| 6 | 113.000 | 88.000 |
| 7 | 111.000 | 86.000 |
| 8 | 83.000 | 56.000 |
| 9 | 85.000 | 62.000 |
| 10 | 126.000 | 92.000 |
| 11 | 106.000 | 54.000 |
| 12 | 117.000 | 81.000 |

-->DESCRIBE AUTHORIT, STATUS. BOXPLOT

```
VARIABLE     NAME   IS AUTHORIT
NUMBER OF OBSERVATIONS       12
NUMBER OF MISSING VALUE       0

                  STATISTIC      STD. ERROR     STATISTIC/S.E.

MEAN               97.0000          6.7700         14.3278
VARIANCE          550.0000
STD DEVIATION      23.4521
C.  V.               .2418
SKEWNESS            -.9511           .6373
KURTOSIS             .2463          1.2322
```

```
                  QUARTILE

MINIMUM             40.0000
1ST QUARTILE        83.0000
MEDIAN              98.0000
3RD QUARTILE       113.0000
MAXIMUM            126.0000

                    RANGE

MAX - MIN           86.0000
Q3  - Q1            30.0000


VARIABLE     NAME   IS   STATUS
NUMBER OF OBSERVATIONS       12
NUMBER OF MISSING VALUE       0

                  STATISTIC      STD. ERROR     STATISTIC/S.E.

MEAN               62.3333          5.7936         10.7590
VARIANCE          402.7879
STD DEVIATION      20.0696
C.  V.               .3220
SKEWNESS             .2086           .6373
KURTOSIS           -1.6642          1.2322

                  QUARTILE

MINIMUM             37.0000
1ST QUARTILE        42.0000
MEDIAN              56.0000
3RD QUARTILE        81.0000
MAXIMUM             92.0000

                    RANGE

MAX - MIN           55.0000
Q3  - Q1            39.0000


BOX-PLOT FOR THE SPECIFIED VARIABLE(S)

   126.00  .    X
           .    |
           .  --+--
           .  |   |
           .  |   |
           .  | * |
           .  |   |
           .  |   |     X
           .  |   |     |
           .  --+--   --+--
           .          |   |
           .          |   |
           .          |   |
           .          |   |
           .          | * |
           .          |   |
           .          |   |
           .          |   |
           .    o     --+--
    37.00  .            X
```

The first box plot shown above is of the variable AUTHORIT, the second is of STATUS. As a second illustration, we will generate complete descriptive information for the variables PGAS and PCRUDE used in Chapter 3.

```
-->DESCRIBE  PGAS, PCRUDE.  STEMLEAF.  BOXPLOT

 VARIABLE     NAME   IS     PGAS
 NUMBER OF OBSERVATIONS       72
 NUMBER OF MISSING VALUE       0

                   STATISTIC      STD. ERROR     STATISTIC/S.E.

MEAN               571.6181          7.2042         79.3456
VARIANCE          3736.7913
STD DEVIATION       61.1293
C.  V.                .1069
SKEWNESS              .3252           .2829
KURTOSIS             -.9471           .5588

                   QUARTILE

MINIMUM            458.4000
1ST QUARTILE       520.1000
MEDIAN             560.4000
3RD QUARTILE       611.0000
MAXIMUM            694.7000

                      RANGE

MAX - MIN          236.3000
Q3  - Q1            90.9000


STEM-AND-LEAF DISPLAY FOR THE VARIABLE PGAS

   44  .8
   46  .7
   48  .014
   50  .025601245888
   52  .0113371346679
   54  .8169
   56  .1067179
   58  .3512669
   60  .0113817
   62  .18936
   64  .228
   66  .26687
   68  .3605


VARIABLE     NAME   IS   PCRUDE
NUMBER OF OBSERVATIONS       72
NUMBER OF MISSING VALUE       0

                   STATISTIC      STD. ERROR     STATISTIC/S.E.

MEAN               590.5212          8.4940         69.5220
VARIANCE          5194.6756
STD DEVIATION       72.0741
C.  V.                .1221
SKEWNESS              .1152           .2829
KURTOSIS             -.4766           .5588
```

```
                  QUARTILE

MINIMUM           447.7768
1ST QUARTILE      539.5815
MEDIAN            589.0148
3RD QUARTILE      626.6783
MAXIMUM           734.7864

                     RANGE

MAX - MIN         287.0096
Q3  - Q1           87.0967


STEM-AND-LEAF DISPLAY FOR THE VARIABLE PCRUDE

    4  .
    4  .55677889
    5  .01244444444444
    5  .557788999999999999999999
    6  .013333333344
    6  .57999999
    7  .013333


 BOX-PLOT FOR THE SPECIFIED VARIABLE(S)

   734.79  .            O4
           .
           .            X
           .    X       |
           .    |       |
           .    |       |
           .    |       |
           .    |     --+--
           .  --+--   |   |
           .  |   |   |   |
           .  |   |   | * |
           .  | * |   |   |
           .  |   |   |   |
           .  |   |   --+--
           .  --+--     |
           .    |       |
           .    |       |
           .    |       |
           .    X       |
   447.78  .            X
```

The first box plot displayed is that of the variable PGAS, the second is of PCRUDE.

## 4.2 Tables for subgroups or subsamples

Often recorded responses can be divided into groups according to a common factor or treatment of the group. It is informative to observe how responses differ from group to group. The TABLE paragraph computes and displays the sample mean and standard deviation of groups within a response variable. Groups are defined according to the levels assumed by one or two external factors.

To illustrate the use of the TABLE paragraph, we consider the survival times of 48 animals exposed to three different poisons and treated with four different antidotes. The data are used in Box, Hunter and Hunter (1978, Sections 7.7 through 7.9), and are listed in Table 4. Data are stored in the SCA workspace in the variables SURVIVAL, TREATMNT and POISON.

**Table 4 Toxic agents data**

Survival times (in tens of hours)

| | *Treatment* | | | |
|---|---|---|---|---|
| *Poison* | A | B | C | D |
| 1 | 0.31 | 0.82 | 0.43 | 0.45 |
| | 0.45 | 1.10 | 0.45 | 0.71 |
| | 0.46 | 0.88 | 0.63 | 0.66 |
| | 0.43 | 0.72 | 0.76 | 0.62 |
| 2 | 0.36 | 0.92 | 0.44 | 0.56 |
| | 0.29 | 0.61 | 0.35 | 1.02 |
| | 0.40 | 0.49 | 0.31 | 0.71 |
| | 0.23 | 1.24 | 0.40 | 0.38 |
| 3 | 0.22 | 0.30 | 0.23 | 0.30 |
| | 0.21 | 0.37 | 0.25 | 0.36 |
| | 0.18 | 0.38 | 0.24 | 0.31 |
| | 0.23 | 0.29 | 0.22 | 0.33 |

We can create a table for average survival times based on the type of poison used by entering

```
-->TABLE   SURVIVAL, POISON
```

we obtain the following

```
DEPENDENT   VARIABLE : SURVIVAL
CATEGORICAL VARIABLES: POISON


  ROW      MEAN     STD. DEV.   COUNTS
-----------------------------------------
   1        .617        .209     16
   2        .544        .289     16
   3        .276        .062     16
```

We are provided with the sample mean, standard deviation and frequency counts of survival times for each of the 3 poisons. We can obtain a similar tabular breakdown of survival times based on antidote used (i.e., TREATMNT) by entering

```
-->TABLE   SURVIVAL, TREATMNT.    STORE   TREATMN, TREATSTD, TREATCNT.
```

We obtain

```
DEPENDENT   VARIABLE : SURVIVAL
CATEGORICAL VARIABLES: TREATMNT


  ROW      MEAN     STD. DEV.   COUNTS
---------------------------------------
   1        .314        .102     12
   2        .677        .321     12
   3        .392        .167     12
   4        .534        .219     12
```

The tabular information is presented in the same form as before, though the treatments are displayed as columns in Table 4. The "rows" of the categorical variable TREATMNT are presented in "ascending" order according to the value (or level) of the variable. Here the numeric order corresponds to A, B, C, and D, respectively.

We also included the STORE sentence within the TABLE paragraph. In this way we retain the sample means, standard deviations, and number of observations per group in the variables TREATMN, TREATSTD and TREATCNT, respectively. Retaining this information can be especially useful if we wish to display the sample means of the groups jointly with a reference distribution (see Chapter 5).

We can obtain a (3x4) table of the sample mean and standard deviation for each poison/treatment combination by entering

```
-->TABLE   SURVIVAL, POISON, TREATMNT.    STORE SM,SSTD,SCNT.
```

```
DEPENDENT   VARIABLE : SURVIVAL
CATEGORICAL VARIABLES: POISON, TREATMNT

LINE 1: MEANS       LINE 2: STANDARD DEVIATIONS

          1        2        3        4
   1     .413     .880     .567     .610
         .069     .161     .157     .113

   2     .320     .815     .375     .668
         .075     .336     .057     .271

   3     .210     .335     .235     .325
         .022     .047     .013     .026
```

As before, we use the STORE sentence to retain the sample mean, standard deviation, and number of observations in each category in the variables SM, SSTD, and SCNT, respectively. Information in these variables are used in later chapters of this manual.

If we desire additional categorical information for the response variable (e.g., minimum or maximum), we can employ the CROSSTAB paragraph (see Chapter 6).

## 4.3 Covariance and Correlation

When we analyze more than one data set, or variable, we often want to obtain a measure of how variables are associated with each other, in addition to the "usual" descriptive information on each variable separately. Scatter plots and multiple plots over time are useful visual tools for this purpose (see Chapter 3). Two statistical measures that may be of use are the covariance and correlation between variables.

### 4.3.1 Pearson correlation coefficient

The Pearson correlation coefficient for the variables X and Y is the statistic.

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{S_X X_Y}$$

where $\bar{X}$ , $S_X$ , $\bar{Y}$ and $S_Y$ are the sample mean and standard deviation of the variables X and Y, respectively. The numerator of the above expression is called the sample covariance.

The computed correlation coefficient is a measure of the degree of linear association that may exist between X and Y. However, we must remember that correlation alone does not imply any causal relationships. This statistic is a measure of an assumed linear relationship between the variables, and not how well we may predict one using the other. In fact, correlation between two variables may be attributable to their common relationship to other variables. Furthermore, a small value of the correlation coefficient does not preclude the existence of a non-linear relationship between the variables. We should take these cautions into account in reviewing any correlation results.

We can calculate and display the correlation and/or covariance between two or more variables through the CORRELATE paragraph. As an illustration, the correlation and covariance of the authoritarianism and social status data used in Section 4.1.2 will be calculated and displayed.

```
-->CORRELATE   AUTHORIT, STATUS.    TYPES ARE COVAR, CORR.

NUMBER OF CASES TO BE ANALYZED . . .        12
NUMBER OF COMPLETE CASES . . . . . .        12


VARIABLE          MEAN    STD. DEVIATION   COEFF. OF VARIATION

AUTHORIT        97.00000        23.45208          .24177
STATUS          62.33333        20.06958          .32197
```

```
CORRELATION MATRIX

AUTHORIT     1.0000
STATUS        .7745   1.0000
           AUTHORIT   STATUS


COVARIANCE MATRIX

AUTHORIT   550.0000
STATUS     364.5455    402.7879
           AUTHORIT      STATUS
```

The display provides basic summary statistics for the data sets including individual sample means, standard deviations and coefficients of variation. Correlation between variables are obtained from the CORRELATION MATRIX portion of the display. The correlation between AUTHORIT and STATUS is .7745. The correlation between any variable and itself is 1.0. From the information listed under COVARIANCE MATRIX, we see the correlation coefficient was calculated as 364.5455/((23.45208)(20.06958)).

The TYPES sentence was included in the above paragraph in order to obtain COVAR, the covariance matrix, in addition to CORR, the correlation matrix. The default calculation is of the correlation matrix only; but when the TYPES sentence is employed, all matrices desired must be specified. To illustrate the default display, the correlation matrix of the variables PGAS and PCRUDE (used in Section 4.1) is computed and displayed.

```
-->CORRELATE   PGAS, PCRUDE.

NUMBER OF CASES TO BE ANALYZED . . .        72
NUMBER OF COMPLETE CASES . . . . . .        72

VARIABLE        MEAN    STD. DEVIATION   COEFF. OF VARIATION

PGAS         571.61806        61.12930          .10694
PCRUDE       590.52117        72.07410          .12205

CORRELATION MATRIX

PGAS    1.0000
PCRUDE   .6553   1.0000
          PGAS   PCRUDE
```

**Remark**: It is important to note that in the definition of correlation, it is assumed that the X variable and the Y variable have a mean level that is relatively constant throughout their data spans. For the above gasoline and crude oil price data, neither PGAS nor PCRUDE has a fixed mean level since there is a noticeable trend present. Therefore the above correlation coefficient in fact is not meaningful.

## Missing data

The following action is taken in the computation of the correlation matrix for variables containing missing data. Whenever a missing value is encountered, the corresponding

observation of all other specified variables will be excluded from the calculation. In this way, each set of bivariate computations will have the same number of observations.

### 4.3.2 Autocorrelation and cross correlation

Sometimes we may observe that a variable maintains some degree of "memory" of its past values. This phenomenon is often exhibited by variables whose values are recorded over time. There may be a strong correlation of where a process "is" in relation to where it last "was". For example, the temperature we may record now is highly correlated with what was recorded an hour ago, two hours ago, and even longer. Analysis of time dependent data, time series analysis, is discussed in Chapters 9 and 10.

For time dependent data, it is important that we consider the correlation that exists between current and past values in time. This is called the autocorrelation of a series. We may be understand the computation of autocorrelation in the following manner. Suppose we list the values a variable assumes, in the order they are recorded, in a column. Now we can "create" a new variable by writing the previously recorded value of the variable next to each value of the column. In doing so, we have created a lagged variable, lagged in time by one time period. We can create other variables lagged in time by two periods, three periods, and so on. We can now correlate our original series with the one lagged by one period of time, then with the one lagged by two periods of time, and so on. In doing so we are computing the values of the sample autocorrelation function (ACF) for various lags. We can calculate and display this function using the ACF paragraph. As an illustration, we will compute and display the ACF for the PGAS data, used previously, for 15 lags. The display shown has been edited for presentation purposes.

```
-->ACF  PGAS.  MAXLAG IS 15.
```

```
AUTOCORRELATIONS

          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+
                                     I
  1    .94                    +     IXXXXX+XXXXXXXXXXXXXXXXX
  2    .84                +         IXXXXXXXXX+XXXXXXXXXX
  3    .75               +          IXXXXXXXXXX+XXXXXXX
  4    .68              +           IXXXXXXXXXX+XXXX
  5    .65             +            IXXXXXXXXXXXX+XX
  6    .62            +             IXXXXXXXXXXXXX+X
  7    .61           +              IXXXXXXXXXXXXXX+
  8    .59          +               IXXXXXXXXXXXXXX +
  9    .57         +                IXXXXXXXXXXXXX   +
 10    .53         +                IXXXXXXXXXXXX    +
 11    .48        +                 IXXXXXXXXXXX      +
 12    .40        +                 IXXXXXXXXX        +
 13    .31        +                 IXXXXXXX          +
 14    .24       +                  IXXXXX             +
 15    .19       +                  IXXXX              +
```

From the above graph, we can infer that there is a strong memory present in the data. A more complete discussion related to this series can be found in Chapters 9 and 10.

When there are two or more series, or variables, that are recorded in time we should consider the lagged cross correlations that may exist between pairs of series. The concept of cross correlation, and the cross correlation function (CCF), is similar to that of autocorrelation. However, here we also need to keep track of how one series is lagged in relation to the other. That is, we need to remember which series may be "leading" the other. To illustrate the computation and display of the CCF, we will again consider the variables PGAS and PCRUDE of the previous chapter. The display is again edited for presentation purposes.

-->CCF PCRUDE, PGAS. MAXLAG IS 15.

```
CROSS CORRELATION BETWEEN   PCRUDE(T)  AND     PGAS(T-L)

          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+
                                      I
-15   -.37                XXX+XXXXXI     +
-14   -.37                XXX+XXXXXI     +
-13   -.32                 XX+XXXXXI     +
-12   -.25                   XXXXXXI     +
-11   -.16                   + XXXXI     +
-10   -.08                   +   XXI     +
 -9    .00                   +     I     +
 -8    .07                   +     IXX   +
 -7    .13                   +     IXXX  +
 -6    .18                   +     IXXXXX+
 -5    .24                   +     IXXXXXX
 -4    .31                   +     IXXXXX+XX
 -3    .40                   +     IXXXXX+XXXX
 -2    .49                   +     IXXXXX+XXXXXX
 -1    .58                   +     IXXXXX+XXXXXXXXX
  0    .66                   +     IXXXXX+XXXXXXXXXX
  1    .65                   +     IXXXXX+XXXXXXXXXX
  2    .63                   +     IXXXXX+XXXXXXXXXX
  3    .61                   +     IXXXXX+XXXXXXXXX
  4    .59                   +     IXXXXX+XXXXXXXXX
  5    .58                   +     IXXXXX+XXXXXXXX
  6    .56                   +     IXXXXX+XXXXXXXX
  7    .55                   +     IXXXXX+XXXXXXXX
  8    .54                   +     IXXXXX+XXXXXXX
  9    .51                   +     IXXXXX+XXXXXXX
 10    .47                   +     IXXXXX+XXXXXX
 11    .41                   +     IXXXXX+XXXX
 12    .33                   +     IXXXXX+XX
 13    .26                   +     IXXXXX+X
 14    .23                   +     IXXXXXX
 15    .21                   +     IXXXXX+
```

Thirty-one cross correlations are computed and displayed. The legend indicates the relative lagging schemes. Positive lags indicate the correlation between "currently" observed values of PCRUDE with that of those values of PGAS "L" periods before. Negative lags indicate the correlation between "currently" observed values of PGAS with those of PCRUDE "L" periods before. The value for L=0 is the cross correlation of contemporaneous periods. This value, .66, is the Pearson correlation coefficient computed previously. It is important to note that cross correlations may not reveal actual "leading" or "lagging" relationships for

series that exhibit autocorrelation. A more detailed discussion can be found in Box and Jenkins (1970).

# SUMMARY OF THE SCA PARAGRAPHS IN CHAPTER 4

This section provides a summary of those SCA paragraphs employed in this chapter. The syntax for each paragraph is presented in both a brief and full form. The brief display of the syntax contains the most frequently used sentences of a paragraph, while the full display presents all possible modifying sentences of a paragraph. In addition, special remarks related to a paragraph may also be presented with the description.

Each SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

The paragraphs to be explained in this summary are DESCRIBE, TABLE, CORRELATE, ACF, and CCF.

Legend (see Chapter 2 for further explanation)

v : variable name
i : integer
w : keyword

## DESCRIBE Paragraph

The DESCRIBE paragraph is used to calculate display descriptive statistics of a single variable. If multiple variables are specified, then statistics for each variable will be displayed. Descriptive statistics include measures of location, spread, skewness and kurtosis. In addition, a stem-and-leaf display or box plot may be produced for each variable.

## Syntax for the DESCRIBE Paragraph

### Brief syntax

```
DESCRIBE    VARIABLES ARE v1, v2, ---.        @
            BOXPLOT./NO BOXPLOT.              @
            STEMLEAF./NO STEMLEAF.

Required sentence:  VARIABLES
```

### Full syntax

```
DESCRIBE    VARIABLES ARE v1, v2, ---.        @
            BOXPLOT./NO BOXPLOT.              @
            STEMLEAF./NO STEMLEAF.            @
            WEIGHT IS v.                      @
            SUMMARY./NO SUMMARY.              @
            HOLD w1(v1), w2(v2), ---.

Required sentence:  VARIABLES
```

## Sentences used in the DESCRIBE Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the names of the variables for which descriptive statistics will be produced. Only variables with numeric values should be specified.

**BOXPLOT sentence**

The BOXPLOT sentence is used to specify that a box plot be produced for each variable. The default is NO BOXPLOT.

**STEMLEAF sentence**

The STEMLEAF sentence is used to specify that a stem-and-leaf display be produced for each variable. The default is NO STEMLEAF.

**WEIGHT sentence**

The WEIGHT sentence is used to specify a variable containing the frequency count (i.e., weights) to be used for each observation of each variable. The default frequency is a count of 1 for each observation.

**SUMMARY sentence**

The SUMMARY sentence is used to specify the calculation and display of summary output. The default is SUMMARY.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. Default is that none of the values of the above statistics will be retained after the paragraph is used. The values that may be retained are:

| | |
|---|---|
| MEAN | : the sample mean |
| MEDIAN | : the sample median |
| VARIANCE | : the sample variance |
| STD | : the sample standard deviation |
| CV | : the coefficient of variation |
| MINIMUM | : the minimum value of the data set |
| MAXIMUM | : the maximum value of the data set |
| RANGE | : sample range |
| Q1 | : the sample 25th percentile |
| Q3 | : the sample 75th percentile |
| SKEWNESS | : the sample value of skewness |
| KURTOSIS | : the sample value of kurtosis |
| NCASE | : the number of cases |
| NMISS | : the number of missing values |

## TABLE Paragraph

The TABLE paragraph computes and displays the sample mean and standard deviation of groups within a response (dependent) variable. Groups are defined according to the levels assumed by one or two external factors. Values in the may be numeric or character. The mean, standard deviation and count of each group can be maintained in variables in the SCA workspace.

## Syntax for the TABLE Paragraph

### Brief syntax

```
TABLE       VARIABLES ARE v1, v2, v3.  @
            STORE IN v1, v2, v3.
```

Required sentence: **VARIABLES**

### Full syntax

```
TABLE       VARIABLES ARE v1, v2, v3.  @
            WEIGHT IS v.               @
            MISSING./NO MISSING.       @
            SPAN IS i1, i2.            @
            STORE IN v1, v2, v3.
```

Required sentence: **VARIABLES**

## Sentences Used in the TABLE Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the response variable, v1, and one or two variables (v2, v3) whose levels will be used in the construction of a categorization table. Within each cell of the table, the mean and standard deviation of all associated observations of the response variable will be displayed. The levels assumed by the factor variables (v2, v3) are used as levels in the construction of the categorization table.

**WEIGHT sentence**

The WEIGHT sentence is used to specify the name of a case weight variable. Each observation of the factor variable(s) is entered into group counts as many times as the corresponding value of the weight variable. Any case with a negative weight is ignored in the computations. The default is to assign each case a weight of one.

**MISSING sentence**

The MISSING sentence is used to specify whether missing values are to be counted in the computation of table statistics. Specify MISSING if missing data are to be included in computations and NO MISSING if missing data are to be excluded from computations. NO MISSING is the default.

**SPAN sentence**

The SPAN sentence is used to specify the span of cases, from i1 to i2, of each variable from which the table(s) will be constructed. The default is all observations.

**STORE sentence**

The STORE sentence is used to specify one, two, or three variables in which the sample means, standard deviations, and number of observations per cell in the table are stored. The sample means are stored in v1; the standard deviations are stored in v2 (if specified); and the number of observations per cell are stored in v3 (if specified). Data are stored in a column by column manner. The default is that no information is retained.

## CORRELATION Paragraph

The CORRELATION paragraph is used to calculate and display the correlation between variables. Certain summary statistics are also displayed and many statistics may be retained by the user in the SCA workspace.

## Syntax for the CORRELATION Paragraph

### Brief syntax

```
CORRELATION  VARIABLES ARE v1. v2. ---.
```

### Full syntax

```
CORRELATION     VARIABLES ARE v1, v2, ---.          @
                TYPES ARE w1, w2, ---.              @
                WEIGHT IS v.                        @
                SPAN IS i1, i2.                     @
                SUMMARY./NO SUMMARY.                @
                HOLD CORR(v), NOBS(v), SSCP(v), XPX(v), @
                    COVAR(v), MEAN(v), STD(v).

Required sentence: VARIABLES
```

## Sentences Used in the CORRELATION Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the labels of the variables for which calculations will be made. At least two variables must appear.

**TYPES sentence**

The TYPES sentence is used to specify the kind of cross products matrix to produce. CORR specifies a correlation matrix, this is the default. COVAR specifies a variance-covariance matrix, and SSCP specifies a sum of squares and cross products matrix. More than one type of matrix can be specified at a time. The default type is CORR.

**WEIGHTS sentence**

The WEIGHTS sentence is used to specify a variable containing frequency count (i.,e., weights) for each observation. The default is a frequency count of 1.0 for each observation.

**SPAN sentence**

The SPAN sentence is used to specify the span of cases, from i1 to i2, of each variable from which statistics will be calculated. Default is all observations.

**SUMMARY sentence**

The SUMMARY sentence is used to specify that basic statistics be displayed in addition to the cross-products matrix. The summary statistics produced for each variable are sample mean, variance, standard deviation, total number of observations and number of missing observations. NO SUMMARY suppresses the production of these statistics. The default is SUMMARY.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. Default is that none of the values of the above statistics will be retained after the paragraph is executed. The values that may be retained are:

| | |
|---|---|
| CORR | : Pearson's correlation coefficient |
| NOBS | : number of cases comprising each pair of correlations |
| SSCP | : adjusted sum of squares and cross products matrix |
| XPX | : unadjusted sum of squares and cross products matrix |
| COVAR | : variance-covariance matrix |
| MEAN | : mean of each variable |
| STD | : standard deviation of each variable |

## ACF Paragraph

The ACF paragraph is used to compute the sample autocorrelation function of a time series. The paragraph also displays some descriptive statistics including the sample mean, standard deviation and a t-statistic on the significance of a constant term. The ACF paragraph is discussed in more detail in Chapter 10.

## Syntax for the ACF paragraph

### Brief syntax

```
ACF   VARIABLE IS  v.
```

### Full syntax

```
ACF   VARIABLE IS v.      @
      MAXLAG IS i.        @
      SPAN IS i1, i2.     @
      HOLD ACF(v), SDACF(v).

Required sentence:  VARIABLE
```

## Sentences Used in the ACF paragraph

**VARIABLE sentence**

The VARIABLE sentence is used to specify the name of the series to be analyzed.

**MAXLAG sentence**

The MAXLAG sentence is used to specify the maximum order (lag) of sample ACF to compute. Default is 36.

**SPAN sentence**

The SPAN sentence is used to specify the span of time indices, from i1 to i2, for which the data will be analyzed. Default is the maximum span available for the series.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. Default is that none of the values of the above statistics will be retained after the paragraph is executed. The values that may be retained are:

ACF : the sample ACF of the series
SDACF : the standard deviations of the sample ACF for the series.

## CCF Paragraph

The CCF paragraph is used to compute the cross correlation function between two specified time series. The paragraph also displays for each series some descriptive statistics including the sample mean, standard deviation and a t-statistic on the significance of a constant term. The CCF paragraph is discussed in more detail in Chapter 10.

## Syntax for the CCF Paragraph

### Brief syntax

```
CCF   VARIABLES ARE  v1, v2.
```

### Full syntax

```
CCF   VARIABLES ARE  v1, v2.     @
      MAXLAG IS  i.              @
      SPAN IS  i1, i2.           @
      HOLD CCF(v), SDCCF(v).

Required sentence:  VARIABLES
```

## Sentences Used in the CCF Paragraph

**VARIABLES sentence**
The VARIABLES sentence is used to specify the names of the series to be analyzed.

**MAXLAG sentence**
The MAXLAG sentence is used to specify the maximum order (lag) of CCF to compute. Default is 36. Note there will be the same number of negative and positive lags calculated.

**SPAN sentence**
The SPAN sentence is used to specify the span of time indices, i1 to i2, for which the data will be analyzed. Default is the maximum span available for the series.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. Default is that none of the values of the above statistics will be retained after the paragraph is used. The values that may be retained are:

CCF : the sample CCF of the series
SDCCF : the standard deviations of the sample CCF of the series

## REFERENCES

Box, G.E.P., and Jenkins, G.H. (1970). *Time-Series Analysis: Forecasting and Control*, San Francisco: Holden Day.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Tukey, J.W. (1977). *Exploratory Data Analyses*. Reading, MA: Addison-Wesley.

# CHAPTER 5

# DISPLAYS OF LOCATION, DISPERSION AND DISTRIBUTION

The characteristics of a data set can often be represented by its center and the variability (dispersion) about it. We often learn much through the examination of the "center" of data (e.g., mean, median, or mode) and how the data distributes itself about this value. In this manual, we use the mean as the principal statistic to represent the center of a data set. As a result, we may be concerned both in the distribution of data around a sample mean, and the distribution of the sample mean itself. It is useful to visually examine data whenever we have any specific distributional assumptions of it. This can be of particular importance in model building and in the examination of residuals from a fitted model.

Questions also arise when data can be divided into groups. It is natural to wonder if all subgroups are "alike". If not, important information may be gained by learning how and why the groups differ. It is useful to compute and retain information on groups as well as to obtain visual representations of groups and of group means. When only two groups are present, we may wish to determine whether the measured responses of one group (e.g., yield, survival rate, bad parts) are somehow superior to the other. If so, then we may be able to isolate a key factor that is responsible for such behavior.

In this chapter we present SCA capabilities useful for simple data description and distributional references. Capabilities are shown for both an entire sample, and for subgroups within the sample. Some methods for visualization are described. In addition, we provide a capability for calculating a confidence interval for a data set. In the last section, we present a method, probability plot, to examine if the data from a sample follow a normal distribution.

## 5.1 Histograms

Histograms are useful for the display of the frequency of occurrence of observations of one of more variables. To illustrate a histogram, we shall consider the data set USPOP consisting of the 1970 population (in millions) of the 50 states. The data were used in Chapter 4 and are listed in Table 1.

**Table 1 1970 Population (in millions) for the 50 United States.**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3.44 | .30 | 1.77 | 1.92 | 19.95 | 2.21 | 3.03 | .55 | 6.79 | 4.59 |
| .77 | .71 | 11.01 | 5.19 | 2.83 | 2.25 | 3.22 | 3.64 | .99 | 3.92 |
| 5.69 | 8.88 | 3.81 | 2.22 | 4.68 | .69 | 1.48 | .49 | .74 | 7.17 |
| 1.02 | 18.24 | 5.08 | .62 | 10.65 | 2.56 | 2.09 | 11.79 | .95 | 2.59 |
| .67 | 3.92 | 11.20 | 1.06 | .44 | 4.65 | 3.41 | 1.74 | 4.42 | .33 |

To obtain the histogram of this variable, we may simply enter

```
-->HISTOGRAM  USPOP
```

```
VARIABLE     NAME   IS    USPOP
NUMBER OF OBSERVATIONS       50
NUMBER OF MISSING VALUE       0

    LOWER     UPPER  FREQ- 0    5   10   15   20   25   30   35   40
    BOUND     BOUND  UENCY +----+----+----+----+----+----+----+----+----
                            I
     .000 -    4.000    34 IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                           IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                            I
    4.000 -    8.000     9 IXXXXXXXXX
                           IXXXXXXXXX
                            I
    8.000 -   12.000     5 IXXXXX
                           IXXXXX
                            I
   12.000 -   16.000     0 I
                            I
                            I
   16.000 -   20.000     2 IXX
                           IXX
                            I
                           +----+----+----+----+----+----+----+----+----
                           0    5   10   15   20   25   30   35   40
```

We obtain a horizontal layout of information. There is a column of sequentially increasing ranges of values assumed by the data. For each interval, the number of occurrences within the designated interval and a row "bar" representing this number are displayed.

## Types of histograms

The histogram displayed above is the default generated by the HISTOGRAM paragraph; that is, the frequency (number) of values observed in an interval. Row bars can also reflect any of the following:

| | | |
|---|---|---|
| cumulative frequency | -- | the accumulated number of values observed up to and including the present interval |
| percentage | -- | the fraction of values observed in the interval |
| cumulative percentage | -- | the accumulated fraction of values observed up to and including the present interval |

We will restrict our attention to the default, frequency, throughout this section.

## 5.2 Histogram of More Than One Variable

The HISTOGRAM paragraph will provide us with exactly one histogram regardless of the number of variables specified. When more than one variable are specified, the data of all variables will be utilized for the construction of a pooled histogram; that is, a histogram of all data.

To illustrate the histogram construction of more than one variable, we will consider data from a study of the length of a time (in months) to spoilage for various types of apples (see Barker, 1985, p.170). The types of apples studied were Ida Red, McIntosh, and Delicious. The data are shown in Table 2. The time to spoilage for each apple are stored in the SCA workspace under the labels IDA, MAC and DEL, respectively.

**Table 2 Months to spoilage for various types of apples**

| Ida Red<br>IDA | McIntosh<br>MAC | Delicious<br>DEL |
|---|---|---|
| 6.5 | 7.5 | 5.0 |
| 7.5 | 8.5 | 6.0 |
| 8.0 | 9.5 | 7.0 |
| 7.5 | 9.5 | 6.0 |
| 5.0 | 7.5 | 5.0 |
| 6.0 | 8.0 | 4.5 |
| 8.0 | 9.0 | 6.5 |
| 8.5 | 9.0 | 7.5 |
| 7.0 | 9.0 | 5.5 |
| 4.5 | 7.0 | 4.5 |

To obtain the histogram of the time to spoilage for Ida Red apples, we enter

-->HISTOGRAM IDA

```
VARIABLE     NAME   IS      IDA
NUMBER OF OBSERVATIONS       10
NUMBER OF MISSING VALUE       0

    LOWER     UPPER  FREQ- 0    5   10   15   20   25   30   35   40
    BOUND     BOUND  UENCY +----+----+----+----+----+----+----+----+----
                            I
    4.000 -    5.000     2 IXX
                           IXX
                            I
    5.000 -    6.000     1 IX
                           IX
                            I
    6.000 -    7.000     2 IXX
                           IXX
                            I
    7.000 -    8.000     4 IXXXX
                           IXXXX
                            I
    8.000 -    9.000     1 IX
                           IX
                            I
                           +----+----+----+----+----+----+----+----+----
                           0    5   10   15   20   25   30   35   40
```

To obtain the histogram of the time to spoilage for all types of apples, we may simply enter

-->HISTOGRAM IDA, MAC, DEL

```
VARIABLE     NAME   IS      IDA
NUMBER OF OBSERVATIONS       10
NUMBER OF MISSING VALUE       0

VARIABLE     NAME   IS      MAC
NUMBER OF OBSERVATIONS       10
NUMBER OF MISSING VALUE       0

VARIABLE     NAME   IS      DEL
NUMBER OF OBSERVATIONS       10
NUMBER OF MISSING VALUE       0

LOWER     UPPER  FREQ- 0    5   10   15   20   25   30   35   40
BOUND     BOUND  UENCY +----+----+----+----+----+----+----+----+----
                       I
3.600 -    4.800     3 IXXX
                       IXXX
                       I
4.800 -    6.000     7 IXXXXXXX
                       IXXXXXXX
                       I
6.000 -    7.200     5 IXXXXX
                       IXXXXX
                       I
7.200 -    8.400     8 IXXXXXXXX
                       IXXXXXXXX
                       I
8.400 -    9.600     7 IXXXXXXX
                       IXXXXXXX
                       I
                       +----+----+----+----+----+----+----+----+----
                       0    5   10   15   20   25   30   35   40
```

We see that for the histogram displayed, the time to spoilage for the three types of apples has been combined. We obtain an initial summary of the number of observations (cases) of each type of apple that is used, but no distinction for type is provided on the row bars of the histogram.

The range of values (minimum to maximum) of the values of IDA, MAX and DEL was used in the construction of the histogram above. We can alter the range, and as a result the interval sizes, by including the RANGE sentence in the HISTOGRAM paragraph. For example, if we enter

```
-->HISTOGRAM IDA, MAC, DEL.  RANGE IS 3.0, 10.5.
```

we obtain the following histogram

```
LOWER     UPPER  FREQ- 0    5   10   15   20   25   30   35   40
BOUND     BOUND  UENCY +----+----+----+----+----+----+----+----+----
                       I
3.000 -    4.500     3 IXXX
                       IXXX
                       I
4.500 -    6.000     7 IXXXXXXX
                       IXXXXXXX
                       I
6.000 -    7.500    10 IXXXXXXXXXX
                       IXXXXXXXXXX
                       I
7.500 -    9.000     8 IXXXXXXXX
                       IXXXXXXXX
                       I
9.000 -   10.500     2 IXX
                       IXX
                       I
                       +----+----+----+----+----+----+----+----+----
                       0    5   10   15   20   25   30   35   40
```

We can obtain information regarding the dispersion of each type of apple by doing any of the following

(1) Sequentially using the HISTOGRAM paragraph for each variable separately;

(2) Employing the SYMBOLS sentence (see the syntax section at end of this chapter) to define a separate symbol for the representation of individual variables along the row bars; or

(3) Use the display capability of the DPLOT paragraph, as will now be discussed.

## 5.3 Dispersion Plots

A HISTOGRAM provides us with a display of the distribution of a possibly pooled set of data. We can also obtain a concise display of the distribution if we use the DPLOT paragraph. The DPLOT paragraph will not pool together data, but will create "sub-plots" from the entire population. In addition to the dispersion plot constructed by the DPLOT paragraph, we can also super-impose a plot of a reference t-distribution using the DTPLOT

paragraph (see Section 5.4). The reference distribution permits us to more easily spot any departures from normality. More complete information on both the DPLOT and DTPLOT paragraphs can be found in Chapter 3 of Quality and Productivity Improvement Using the SCA Statistical System.

To illustrate the use of the DPLOT paragraph, we will again consider the apple data used above. However, we will consider the data in a slightly different form. The data are shown in Table 3. Apple types are denoted by I for Ida Red, M for McIntosh, and D for Delicious. We see that in addition to data regarding the type of apple, we have temperature levels at which apples were kept. Each apple and temperature combination was replicated in obtaining a time to spoilage. Data from the table are stored in the SCA workspace under the labels REP1, REP2, TYPE and TEMP, as noted in the table. The data are then edited to form three "columns" of data. One column contains the number of months to spoilage. A second contains the corresponding type of apple. The third is the corresponding temperature. The names of these new "columns", or variables are SPOILAGE, APPLE and STORAGE, respectively. Please see Section 8.2.2 for the commands used to edit the data.

**Table 3 Apple data**

| *Months to spoilage* | | *Type of* | *Storage* |
|---|---|---|---|
| *Replication 1* | *Replication 2* | *Apple* | *Temperature* |
| *REP1* | *REP2* | *TYPE* | *TEMP* |
| 6.5 | 6.0 | I | 36 |
| 7.5 | 8.0 | I | 38 |
| 8.0 | 8.5 | I | 40 |
| 7.5 | 7.0 | I | 42 |
| 5.0 | 4.5 | I | 44 |
| 7.5 | 8.0 | M | 36 |
| 8.5 | 9.0 | M | 38 |
| 9.5 | 9.0 | M | 40 |
| 9.5 | 9.0 | M | 42 |
| 7.5 | 7.0 | M | 44 |
| 5.0 | 4.5 | D | 36 |
| 6.0 | 6.5 | D | 38 |
| 7.0 | 7.5 | D | 40 |
| 6.0 | 5.5 | D | 42 |
| 5.0 | 4.5 | D | 44 |

A HISTOGRAM of the variable SPOILAGE will yield the same histogram as that of the pooled histogram of IDA, MAC and DEL presented before (except no individual component information will be given). If we would like this display more condensed, and on a horizontal axis, we may enter

-->DPLOT SPOILAGE

```
            X
            X
X X   X   X X X   X
X X   X X X X X X X X
X X X X X X X X X X X
+---------*---------+
4.5       7.0       9.5
MEAN=7.0      SD=1.6
```

This display provides us with some basic visual and statistical dispersion information for the months to spoilage. The sample mean is always used as the center for the plot. It may be interesting to know how spoilage is affected by either the type of apple or the storage temperature. We can obtain a dispersion breakdown by the type of apple by entering

-->DPLOT SPOILAGE, APPLE

```
X X   X
X X X X X X X
+----*----+---------+
4.5       7.0       9.5
MEAN=5.8      SD=1.0

 APPLE  =     D


            X X
X X   X X X X X X
+--------*+---------+
4.5       7.0       9.5
MEAN=6.9      SD=1.3

 APPLE  =     I


                  X
            X     X X
          X X X X X X
+---------+-----*---+
4.5       7.0       9.5
MEAN=8.5      SD=.9

 APPLE  =     M
```

This breakdown by type provides us with an interesting set of visual displays. We are presented with a display of the months to spoilage for each of the three types of apples. Each display uses the same axis as that of the full display, permitting easy visual comparisons. The sample mean and standard deviation for each subgroup is also provided. We observe that type M apples (McIntosh) appear to have the greatest "life" to spoilage, while Delicious tends to spoil faster. Except for two observations, Ida Red have a slightly better than (pooled) average time to spoilage. From Table 3 we see the two low times to spoilage for Ida Red are associated with the highest storage temperature.

We can obtain five separate dispersion displays (one for each storage temperature) if we enter

```
-->DPLOT  SPOILAGE, STORAGE
```

These displays are not shown here. We should note that we will obtain the same basic information if we request a dispersion plot for each apple type separately. However, if the plots are requested separately, the horizontal axes used in the displays will be slightly different. To see this, we now request a DPLOT of each apple type separately.

```
-->DPLOT  IDA

                 X  X
X X    X  X  X X  X X
+---------+-*-------+
4.5      6.5      8.5
MEAN=6.9     SD=1.3
```

```
-->DPLOT  MAC

                 X
    X            X   X
X   X   X   X    X   X
+---------+-*-------+
7.0      8.3      9.5
MEAN=8.5     SD=.9
```

```
-->DPLOT  DEL

X  X      X
X  X  X   X  X   X  X
+-------*-+---------+
4.5      6.0      7.5
MEAN=5.8     SD=1.0
```

The plots are similar to the DPLOT of SPOILAGE on APPLE type. Each separate plot provides us with information regarding the "internal" variability of each type of apple. We do not have as striking a visual representation of how overall spoilage is affected by type as we did before, since the SCA System does not "know" these plots are related to one another and we can not restrict each plot to have the same axis for display. In this example, the combined specification of SPOILAGE and APPLE in the DPLOT paragraph provides a more meaningful graphical display.

## 5.4 Reference Distributions

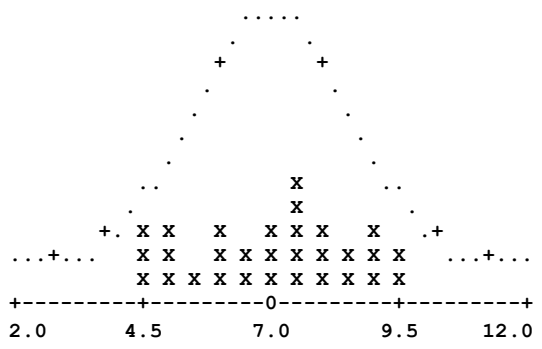
It may be of use to view the dispersion of a data set jointly with an external reference distribution. In this manner we can better understand how well the data "conform" to what may be an underlying distribution. The DTPLOT paragraph displays, in the same frame, the dispersion plot of a variable with one (or more) t-distributions. The DTPLOT paragraph can be used to see how well a data set adheres to an assumption of normality.

More importantly, the DTPLOT paragraph can be used to provide a reference t-distribution for the sample means for groups of data (see Section 5.4.2). If these means represent the average of responses recorded for various levels (or types) of a factor, we then have a visual indication of whether the average responses are alike, or not. If not, we may be able to determine a factor level (or type) that produces a significantly better response than the rest. The DTPLOT also permits the reference distribution to be centered at user designated positions to aid in the visual "grouping" of data. More complete information on the DTPLOT paragraph can be found in Chapter 3 of *Quality and Productivity Improvement Using the SCA Statistical System*.

### 5.4.1 Dispersion plots with a reference distribution

To illustrate the use of the DTPLOT paragraph for a single sample (variable), we will continue with the apple data used in the prior sections. With the data in columns, we can obtain a dispersion plot of the variable SPOILAGE, with a reference t-distribution by entering

```
-->DTPLOT  SPOILAGE
```



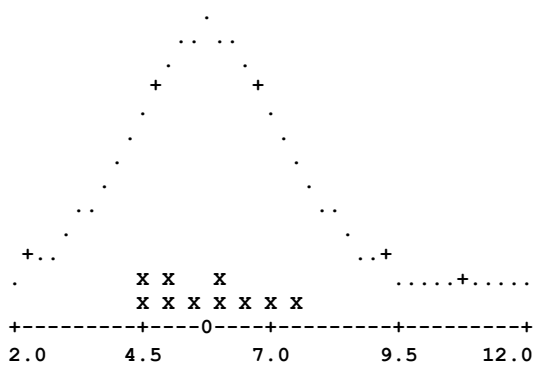

```
MEAN  =7.017          SD= 1.551
CENTER=7.017          SD= 1.551
```

We are provided with a dispersion plot similar to the one obtained before. The summary information lists the sample mean and standard deviation of the sample, and the center and standard deviation used for the reference distribution that appears above the dispersion plot. The default is that the center and standard deviation for the reference distribution are the same as the sample mean and standard deviation. The symbol '+' appears 6 times on the plot of the reference distribution. These occur in 3 separate pairs. The inner-most pair indicate the end points of a 50% confidence intervals for a sample having the reference distribution. The middle pair indicate the end points of a 95% confidence interval, and the outer-most pair indicate a 99% confidence interval. Some pairs may not be displayed if there is not sufficient space in the plot. In the above plot, we observe that all data of SPOILAGE are with the 95% limits.

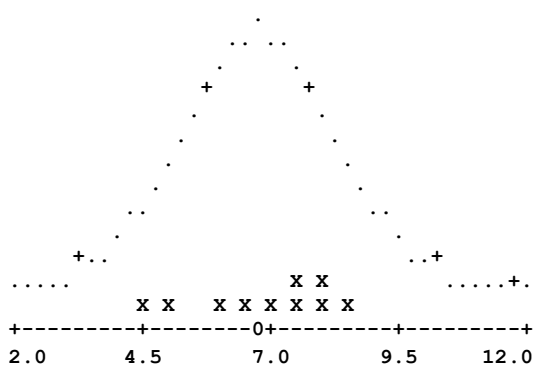
We can also use the DTPLOT paragraph to obtain a reference t-distribution with dispersion plots of subsamples. For example, if we enter

```
-->DTPLOT   SPOILAGE, APPLE
```

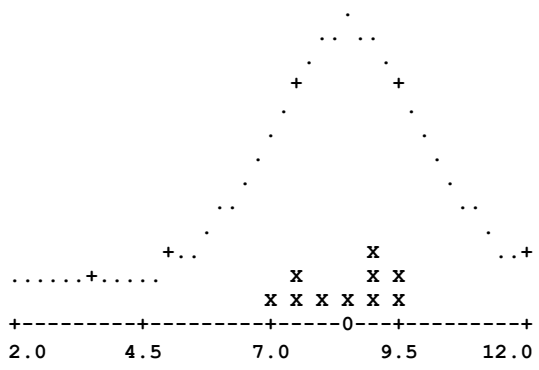
we obtain the following set of displays



```
MEAN  =5.750         SD= 1.034
CENTER=5.750         SD= 1.551
APPLE   =   D
```



```
MEAN  =6.850         SD= 1.334
CENTER=6.850         SD= 1.551
APPLE   =   I
```



```
MEAN  =8.450         SD= .896
CENTER=8.450         SD= 1.551
APPLE   =   M
```

The plots are similar to the ones shown in Section 5.3, but a reference distribution, with center at each group's sample mean and having the standard deviation of the entire sample. In

this manner we can observe how the distribution "shifts" its location (center) from group to group, and how the dispersion of a group compares with the dispersion of the complete sample.

The DTPLOT paragraph provides us with a useful option for the visualization of subsample data. We can designate a center or standard deviation for the reference distribution. For example, in the dispersion plots of SPOILAGE, grouped according to the type of apple stored, we note that apples D and I (Delicious and Ida Red) appear to be similar to one another while the spoilage time for apple M (McIntosh) appears to have higher mean level. To better observe if this is true or not, we may wish to designate one or more centers for the reference distribution. Two centers we may consider in this example are 8.5, the approximate sample mean for time to spoilage associated with apple M, and 6.3, the average of the other times to spoilage. The CENTER sentence is used for this purpose. More information regarding the CENTER sentence can be found in Section 3.2.3 of *Quality and Productivity Improvement Using the SCA Statistical System*.

### 5.4.2 Reference distributions for subsample means

By using the DPLOT paragraph, we obtain a sense of whether different factors produce the same effect, or not. Another way to discern if differences are produced by various factors is to plot the sample mean for each factor level. Since each sample mean is an estimate of its population mean, we may also wish to super-impose an external reference distribution to observe if any sample mean is significantly distinct from the rest.

The TABLE and DTPLOT paragraphs can be conveniently used for this purpose. The TABLE paragraph may be used to obtain the sample means of various groups (see Section 2 of Chapter 4). The DTPLOT paragraph can be used to plot these values jointly with an external reference distribution. However, we also need to calculate an appropriate standard deviation for the reference distribution. This value is computed from the **pooled standard deviation** derived from the standard deviations of the individual groups. The pooled value is the square root of

$$S_p^2 = \frac{\left[ s_1^2(n_1 - 1) + s_2^2(n_2 - 1) + \cdots + s_k^2(n_k - 1) \right]}{\left[ (n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1) \right]}$$

where $s_i$ and $n_i$ are the sample standard deviation and number of observations for the $i^{th}$ group, respectively. The degrees of freedom for the reference t-distribution is the denominator in the above expression and the standard error of the mean is the pooled standard error divided by the number of sample means. These values can be computed easily using analytic operators (see Appendix A) and information retained from the TABLE paragraph.

To illustrate the above, we will consider the toxic agents data used to illustrate the TABLE paragraph in Chapter 4. Now we will construct a reference distribution for the survival times associated with specific poisons with an appropriate reference distribution. The pooled standard error and degrees of freedom for the t-distribution that are appropriate

for both treatment means and poison means is derived from the matrix table presented in Chapter 4. At that time we retained standard deviation and frequency information of the table in the variables SSTD and SCNT, respectively. The pooled standard error and the degrees of freedom may be computed by entering

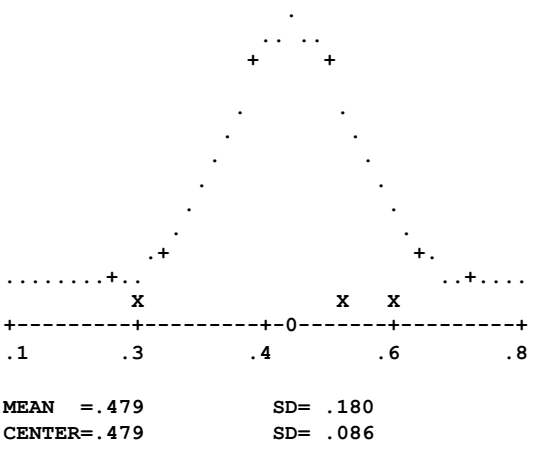
```
-->SPOOL = SQRT (SUM (SSTD*SSTD*(SCNT-1) ) / SUM( SCNT-1 ) )

-->SDF = SUM(SCNT-1)
```

These values are .149 and 36, respectively. The sample means for poisons can be computed using the TABLE paragraph by entering (output from the TABLE paragraph is suppressed)

```
-->TABLE   SURVIVAL, POISON.   STORE  MPOISON.
```

Since we have 3 sample averages, the standard error of the mean is $.149/\sqrt{3} = .086$. We can plot these values against their reference distribution by entering

```
-->DTPLOT   MPOISON.   STDEV IS .086 .   DF ARE 36.
```




We see that one sample mean (corresponding to poison 3) is clearly apart from the other two and lies outside the 95% confidence interval of the reference distribution. We have strong visual evidence that the responses to poisons are different.

## 5.5 Confidence Intervals

The CINTERVAL paragraph may be used to calculate and display the sample mean and the sample standard deviation of a data set, and confidence intervals of the mean. If we are interested in other sample statistics (such as quartiles, skewness, kurtosis) or the display of a stem-and-leaf or box-and-whisker plot, the DESCRIBE paragraph can be used (see Chapter 4).

A **confidence interval for the mean** is a range of values, calculated from a sample, that is likely to include the mean of the population, μ, from which the sample was drawn. The

range computed may not include the population mean, μ. The percentage associated with the interval, say 95%, indicates the percentage of intervals that will include μ if we construct intervals for repeated samples drawn from the population.

The CINTERVAL paragraph provides an estimate of a lower end-point, say LOWER, and upper end-point, say UPPER, for the interval. The computations for these values are dependent on whether the standard deviation of the population, σ, is known or not. If σ is known, then

$$\text{LOWER} = \quad \bar{y} - z(\sigma/\sqrt{n})$$

$$\text{UPPER} = \quad \bar{y} + z(\sigma/\sqrt{n}),$$

where $\bar{y}$ is the sample mean, n is the number of observations in the sample, and z is the value from the standard normal distribution that corresponds to our desired percentage of confidence. However, it is usually the case that σ is not known. In such a case the sample standard deviation, s, is used in its place and

$$\text{LOWER} = \quad \bar{y} - t(s/\sqrt{n})$$

$$\text{UPPER} = \quad \bar{y} + t(s/\sqrt{n}),$$

where t is the value from the t distribution (with n-1 degrees of freedom) that corresponds to the percentage of confidence. The default assumption of the CINTERVAL paragraph is that a 95% confidence interval is to be computed using the sample standard deviation, s. We can designate an alternative percentage value and a value to use as the population standard deviation.

To illustrate the CINTERVAL paragraph, we consider the chemical yields recorded in a manufacturing process (Box, Hunter and Hunter, 1978, Section 2.3). The data are listed in Table 4 and are stored in the SCA workspace in the variable YIELD.

**Table 4 Successive yields from a manufacturing process (Read data across)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 85.5 | 81.7 | 80.6 | 84.7 | 88.2 | 84.9 | 81.8 | 84.9 | 85.2 | 81.9 |
| 89.4 | 79.0 | 81.4 | 84.8 | 85.9 | 88.0 | 80.3 | 82.6 | 83.5 | 80.2 |
| 85.2 | 87.2 | 83.5 | 84.3 | 82.9 | 84.7 | 82.9 | 81.5 | 83.4 | 87.7 |
| 81.8 | 79.6 | 85.8 | 77.9 | 89.7 | 85.4 | 86.3 | 80.7 | 83.8 | 90.5 |
| 84.5 | 82.4 | 86.7 | 83.0 | 81.8 | 89.3 | 79.3 | 82.7 | 88.0 | 79.6 |
| 87.8 | 83.6 | 79.5 | 83.3 | 88.4 | 86.6 | 84.6 | 79.7 | 86.0 | 84.2 |
| 83.0 | 84.8 | 83.6 | 81.8 | 85.9 | 88.2 | 83.5 | 87.2 | 83.7 | 87.3 |
| 83.0 | 90.5 | 80.7 | 83.1 | 86.5 | 90.0 | 77.5 | 84.7 | 84.6 | 87.2 |
| 80.5 | 86.1 | 82.6 | 85.4 | 84.7 | 82.8 | 81.9 | 83.6 | 86.8 | 84.0 |
| 84.2 | 82.8 | 83.0 | 82.0 | 84.7 | 84.4 | 88.9 | 82.4 | 83.0 | 85.0 |
| 82.2 | 81.6 | 86.2 | 85.4 | 82.1 | 81.4 | 85.0 | 85.8 | 84.2 | 83.5 |
| 86.5 | 85.0 | 80.4 | 85.7 | 86.7 | 86.7 | 82.3 | 86.4 | 82.5 | 82.0 |
| 79.5 | 86.7 | 80.5 | 91.7 | 81.6 | 83.9 | 85.6 | 84.8 | 78.4 | 89.9 |
| 85.0 | 86.2 | 83.0 | 85.4 | 84.4 | 84.5 | 86.2 | 85.6 | 83.2 | 85.7 |
| 83.5 | 80.1 | 82.2 | 88.6 | 82.0 | 85.0 | 85.2 | 85.3 | 84.3 | 82.3 |
| 89.7 | 84.8 | 83.1 | 80.6 | 87.4 | 86.8 | 83.5 | 86.2 | 84.1 | 82.3 |
| 84.8 | 86.6 | 83.5 | 78.1 | 88.8 | 81.9 | 83.3 | 80.0 | 87.2 | 83.3 |
| 86.6 | 79.5 | 84.1 | 82.2 | 90.8 | 86.5 | 79.7 | 81.0 | 87.2 | 81.6 |
| 84.4 | 84.4 | 82.2 | 88.9 | 80.9 | 85.1 | 87.1 | 84.0 | 76.5 | 82.7 |
| 85.1 | 83.3 | 90.4 | 81.0 | 80.3 | 79.8 | 89.0 | 83.7 | 80.9 | 87.3 |
| 81.1 | 85.6 | 86.6 | 80.0 | 86.6 | 83.3 | 83.1 | 82.3 | 86.7 | 80.2 |

We will also consider only the first 20 data points of YIELD, stored in the SCA workspace in the variable Y20. To compute a 95% confidence interval for the mean using the data of YIELD, we can simply enter

```
-->CINTERVAL  YIELD

SUMMARY INFORMATION FOR VARIABLE:  YIELD

                                 STD ERR
NOBS      MEAN       STD DEV     OF MEAN      DF     T-VALUE
210     84.1214       2.8809      0.1988     209      1.971

95.0% CONFIDENCE INTERVAL OF THE MEAN:  ( 83.730    ,  84.513   )
```

In addition to the lower and upper limits of the 95% confidence interval, the SCA System displays a one-line summary. This summary includes the number of observations, the sample mean, the sample standard deviation, the standard error of the mean (i.e., $s/\sqrt{n}$), and the t-value appropriate for computing the confidence interval. We can obtain a 90% confidence interval for the population mean by entering

```
-->CINTERVAL   YIELD.   PERCENT IS 90.

SUMMARY INFORMATION FOR VARIABLE:  YIELD

                                 STD ERR
NOBS      MEAN       STD DEV     OF MEAN      DF     T-VALUE
```

```
210    84.1214      2.8809       0.1988    209     1.652

90.0% CONFIDENCE INTERVAL OF THE MEAN:  ( 83.793    , 84.450    )
```

The summary information is essentially the same as before. The only change we observe is the t-value for a 90% confidence interval and the limits of the resultant interval. The PERCENT sentence was used to specify a 90% level.

Confidence intervals are often described as (1-α)100% intervals. Hence we may wonder what would occur if we specify 1-α (e.g., .90 instead of 90). The SCA System interprets a value under 1.00 as 1-α. As a result, if we enter

```
-->CINTERVAL   YIELD.    PERCENT IS .90 .
```

we will create the same confidence interval for the mean as given above.

We can obtain a 95% confidence interval for the first 20 observations of YIELD (i.e., Y20) if we enter

```
-->CINTERVAL  Y20
```

```
SUMMARY INFORMATION FOR VARIABLE:  Y20

                                 STD ERR
NOBS      MEAN       STD DEV     OF MEAN      DF     T-VALUE
 20     83.7250       2.8977       0.6479      19      2.093

95.0% CONFIDENCE INTERVAL OF THE MEAN:  ( 82.369    ,  85.081   )
```

The sample mean and standard deviation of Y20 are almost the same as YIELD. The confidence interval constructed is wider since the number of observations is 20. As a result, the standard error of the mean ( $s/\sqrt{n}$ ) is larger, as is the t-value.

If we assume σ, the standard deviation of the population, is known (say 2.88), we can construct a "z-interval" for the population mean by entering

```
-->CINTERVAL  Y20.   STDEV IS 2.88.   HOLD LOWER(ZLOW), UPPER(ZUPP)
```

```
SUMMARY INFORMATION FOR VARIABLE:  Y20

                                 STD ERR      NORMAL
NOBS      MEAN       STD DEV     OF MEAN      Z-VALUE
 20     83.7250       2.8977       0.6440       1.960

95.0% CONFIDENCE INTERVAL OF THE MEAN:  ( 82.463    ,  84.987   )
```

The STDEV sentence is used to specify a value to use in place of the sample standard deviation for the computation of the standard error of the mean. The standard normal distribution is then used as the reference distribution. Hence a z-value is displayed above and is used in the construction of the confidence interval. If we also provide a value for the "degrees of freedom" (as the second argument in the PERCENT sentence), a t-value will be used instead of a z-value.

The HOLD sentence is used to retain the lower and upper limits of the confidence interval that is constructed. In this case the values are retained in the SCA workspace under the variable names ZLOW and ZUPP, respectively.

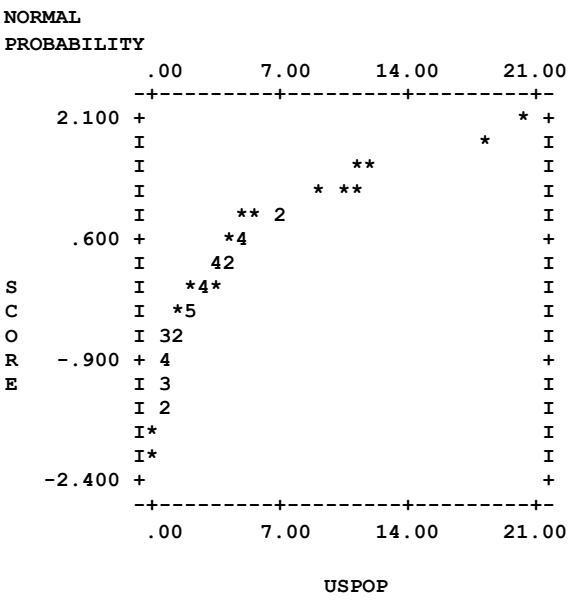
## 5.6 Probability Plots

In this section, we discuss a statistical technique that is commonly employed to verify if a data set is normally distributed. If we assume that a data set should roughly approximate that of a sample from a normal distribution, then we may employ normal probability plots to visually inspect the data. We must remember that any type of normal probability plot is only a graphical representation of the data and will not provide us with an automatic decision regarding the underlying distribution. We should exercise our own judgment regarding the information displayed. In addition, we can employ specific "goodness of fit" tests of how well the data "fit" an assumption of normality (see Chapter 11).

The idea behind a normal probability plot is simple. Suppose we standardize a data set (i.e., subtract the sample mean from all values then divide the resultant values by the sample standard error) and then order these standardized values. We could also, and separately, create a set of values we may expect to observe from a data set of equal size that is drawn randomly from a standard normal distribution. We can also order these values. If we plot our ordered standardized set of data against this ``typical'' ordered standardized set of normal data, then we should observe points aligning close to a straight line. This will be true provided our data follows a normal distribution. In broad terms, this is what is done in the construction of a normal probability plot.

To illustrate probability plots, we consider the data set USPOP we have employed previously. The histogram of this data was skewed appreciably. In effect, it did not come close to an approximation of the "bell-shape" of the normal distribution. Hence, we anticipate that a probability plot will not produce an approximation to a straight line. To obtain a probability plot, we enter

-->PPLOT USPOP



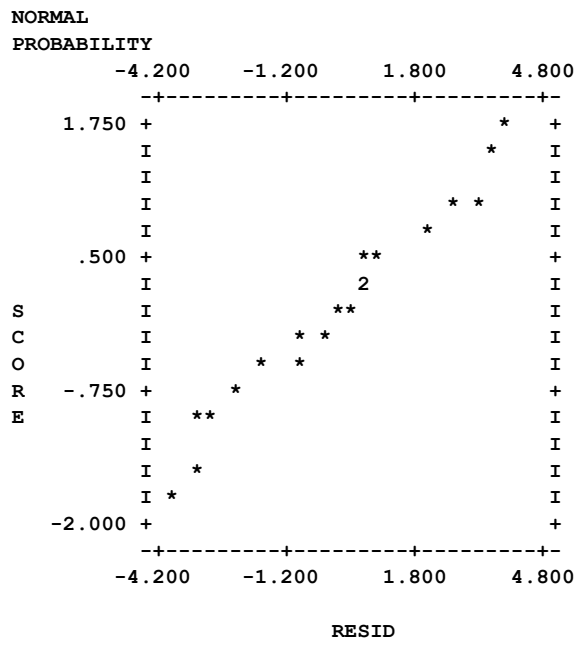

Indeed, the plot above looks more curved than straight. This is a clear indication that the data are not normally distributed.

Often normality is an assumption in the statistical analysis of a data set, especially in a regression analysis. In such cases, we will want to examine the residuals from a fit (i.e., observed values - fitted values). These residuals should be consonant with the normal distribution. For example, consider the data related to bodyfat. Scatter plots of the data are shown in Chapters 3 and 9, and a regression analysis of the data is provided in Chapter 9. The residuals are listed in Table 5 and are stored in the SCA workspace under the label RESID. We can request a probability plot of the residuals to visually inspect how well they conform to a normal distribution.

**Table 5 Residuals from a regression analysis of the bodyfat data**

| | | | | |
|---|---|---|---|---|
| -1.683 | 3.643 | -3.176 | -3.158 | .000 |
| -.361 | .716 | 4.015 | 2.655 | -2.475 |
| .336 | 2.226 | -3.947 | 3.447 | .571 |
| .642 | -.851 | -.783 | -2.857 | 1.040 |

```
-->PPLOT   RESID
```




Here the residuals appear to follow a relatively straight line. Hence one diagnostic check appears to have been passed. We should not restrict our diagnostic checks to that of PPLOT only; other checks should be performed (see Sections 2 and 5 of Chapter 9).

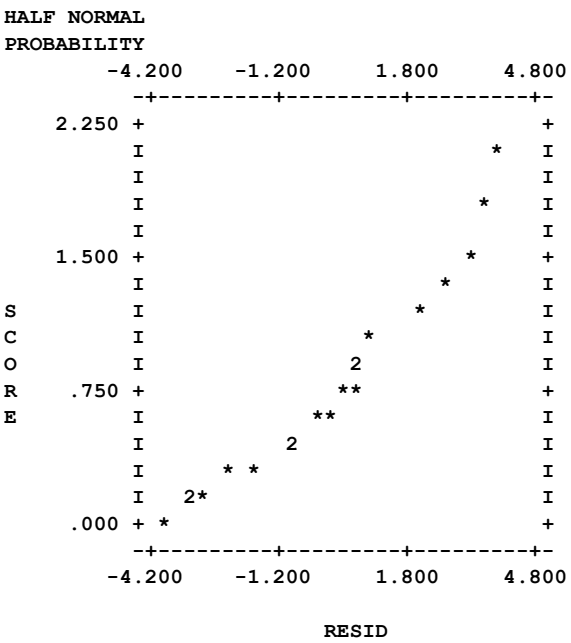
## Types of plots

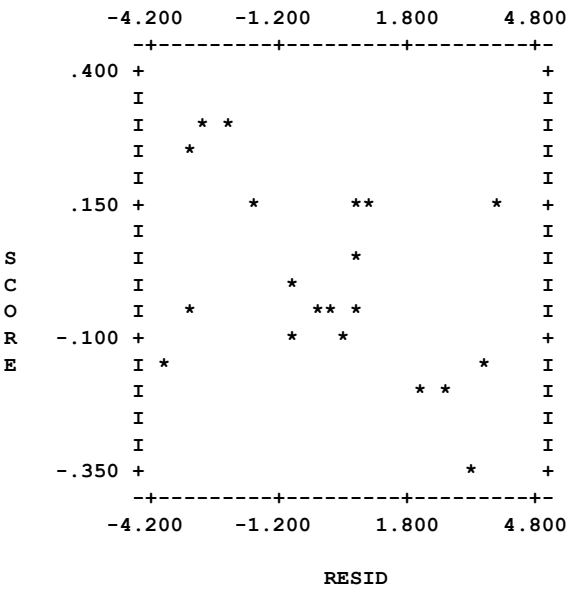
Although the basic idea behind probability plots is simple, we have a number of choices available to us regarding the type of plot we may use and the method with which we derive "expected" values for the set of data used as a reference distribution.

We can request a normal, detrended normal, or half-normal probability plot. The default plot displayed is the normal probability plot. The detrended normal probability plot is similar to the normal probability plot, except the slope of the line is removed. Data that is distributed according to a normal distribution should then randomly oscillate about a horizontal zero line. A half-normal plot is an alternative to a normal probability plot in those cases when a zero mean is known (or assumed) for a data set. Here, the "full" line produced in a normal plot is effectively folded over so that extreme values "coincide". The half-normal plot is then a means to observe all extreme values together.

The normal probability plot of the residuals from the regression did not indicate any significant deviation from a straight line. We can observe the detrended normal and half-normal plots by entering

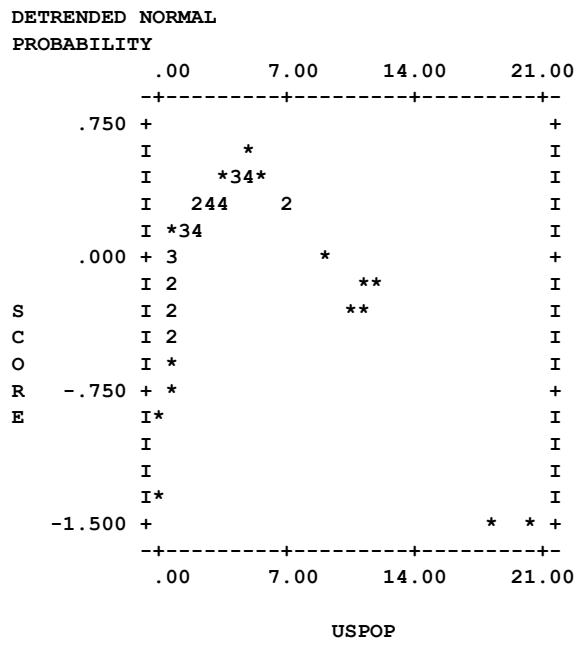
-->PPLOT RESID. TYPES ARE DNORMAL, HNORMAL.




Neither plot appears to significantly differ from a straight line (horizontal in the case of the detrended plot). In contrast we can look at the detrended plot of the U.S. population data by entering

-->PPLOT USPOP. TYPE IS DNORMAL.




Here we have both reverse "bunching" of points and what appears to be a curve. As before, we conclude these data are not normally distributed.

All of the above plots require an approximation to the values we may "expect" to see for a set of standard normal observations that are randomly drawn, then ordered. That is, we need to approximate the value the smallest of these ranked standard normal numbers is likely to be, the value of the next largest, and so on to the largest value of the set. It is these values that our standardized data are plotted against.

All plots are based on the inverse cumulative distribution function, $\Phi^{-1}(x)$. The cumulative distribution function of a standard normal, denoted by $\Phi(z)$, provides the probability of observing a value less than z. The inverse of this function $\Phi^{-1}(x)$ denotes the value, say z, so that $\Phi(z) = x$. The values making up the vertical axis of our plot are the values of the inverse probability function for a function of the integers 1, 2, . . . , n, where n is the number of observations in our data set. Mathematically, we plot our standardized data against $\Phi^{-1}(u(i))$, for some function u. The function u is called the score (or scoring) function of ranks (where 1 is the first rank, 2 is the second rank, and so on).

Three methods for the creation of "scores" are available for either the normal or detrended normal probability plots. These are:

Tukey : $u(i) = (i - 1/3)/(n + 1/3)$

Blom : $u(i) = (i - 3/8)/(n + 1/4)$

van der Waerden : $u(i) = i/(n+1)$

The Tukey scoring function is the default method employed. For a normal probability plot $\Phi^{-1}(u(i))$ is plotted against z(i), where z(i) are the ordered values of our data set, after they are standardized. For a detrended normal probability plot, the values for the Y-axis are

$$\Phi^{-1}(u(i)) - z(i)$$

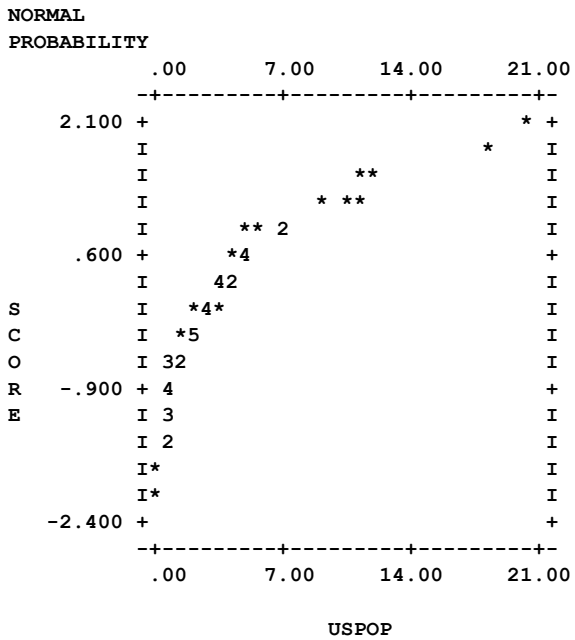
There is no choice in the scoring function used in a half-normal probability plot. Here $\Phi^{-1}(u(i))$ is plotted against z(i), where

$$u(i) = (3n - 3i - 1)/(6n + 1)$$

We can select the type(s) of plots to be produced by including the TYPES sentence in the PPLOT paragraph. We can alter the scoring function for the normal and detrended normal probability plots by including the METHOD sentence in the PPLOT paragraph.

To illustrate the choice of a different scoring function, we will generate a probability plot of the USPOP data using the Blom score function. We simply enter

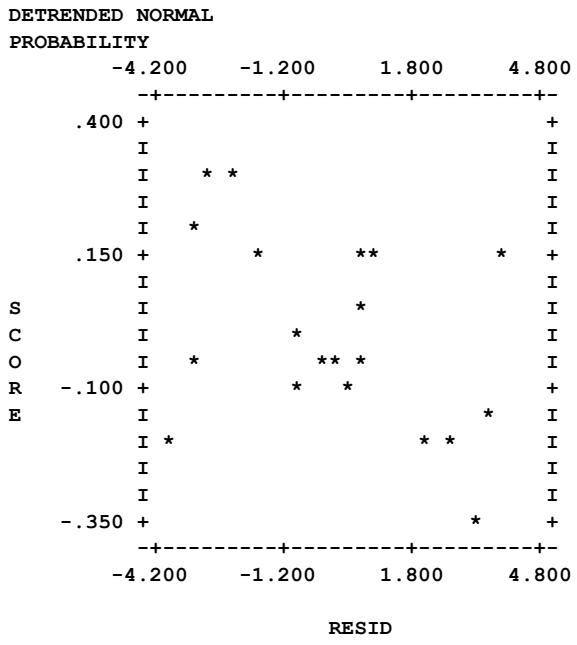
-->PPLOT USPOP. METHOD IS BLOM.



Again the plot is not well represented by a straight line. The choice of scoring function has no effect.

To illustrate how to request both a different type of scoring function and plot, we will create a detrended normal plot of the variable RESID using the Blom function. We enter

-->PPLOT RESID. METHOD IS BLOM. TYPE IS DNORMAL.



# SUMMARY OF THE SCA PARAGRAPHS IN CHAPTER 5

This section provides a summary of those SCA paragraphs employed in this chapter. The syntax for each paragraph is presented in both a brief and full form. The brief display of the syntax contains the most frequently used sentences of a paragraph, while the full display presents all possible modifying sentences of a paragraph. In addition, special remarks related to a paragraph may also be presented with the description.

Each SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

The paragraphs to be explained in this summary are HISTOGRAM, DPLOT, DTPLOT, CINTERVAL and PPLOT.

Legend (see Chapter 2 for further explanation)

v : variable name
i : integer
r : real value
w : keyword
'c' : character data (must be enclosed within single apostrophes)

## HISTOGRAM Paragraph

The HISTOGRAM paragraph is used to create and display the histogram and associated descriptive statistics of a single variable or the pooled histogram of several variables. Descriptive statistics provided include sample mean, standard deviation, coefficient of variation, skewness, kurtosis and quartiles of the distribution. The paragraph will not display multiple histograms in a single use of the paragraph, even if several variables are designated.

## Syntax for the HISTOGRAM Paragraph

### Brief syntax

```
HISTOGRAM      VARIABLES ARE v1, v2, --- .      @
               SPAN IS i1, i2.

Required sentence: VARIABLE(S)
```

### Full syntax

```
HISTOGRAM      VARIABLES ARE v1, v2, --- .      @
               TYPE IS w.                       @
               TITLE IS 'c'.                    @
               INTERVAL IS i.                   @
               BOUNDARIES ARE r1, r2, ---.      @
               RANGE IS r1, r2.                 @
               SPAN IS i1, i2.                  @
               SYMBOLS ARE 'c1', 'c2', ---.     @
               SCALE IS i.                      @
               WIDTH IS i1, i2.                 @
               HEIGHT IS i.                     @
               ZERO./NO ZERO.

Required sentence: VARIABLE(S)
```

## Sentences used in the HISTOGRAM Paragraph

### VARIABLE sentence

The VARIABLES sentence is used to specify the variable(s) used in the construction of the histogram. If several variables are specified, a histogram of the pooled variables is displayed.

**TYPE sentence**

The TYPE sentence is used to specify the criterion to be used in the construction of the histogram. The keywords are

| | | |
|---|---|---|
| FREQ | -- | frequency |
| CFREQ | -- | cumulative frequency |
| PERCENT | -- | percentage |
| CPERCENT | -- | cumulative percentage |

The default type is FREQ.

**TITLE sentence**

The TITLE sentence is used to specify the title for a histogram. The specified title must be enclosed in a pair of apostrophes and have no more than 72 characters. The default is that no title will be displayed.

**INTERVAL sentence**

The INTERVAL sentence is used to specify the maximum number of equally spaced intervals for the histogram. The default is 5 intervals.

**BOUNDARIES sentence**

The BOUNDARIES sentence is used to specify the interval boundaries for the histogram. The default is an internally computed boundary appropriate to the range and the number of intervals requested.

**RANGE sentence**

The RANGE sentence is used to specify the upper and lower limits in which the histogram will be constructed. The default is the range of the variable.

**SPAN sentence**

The SPAN sentence is used to specify the span of indices, i1 to i2, of each variable from which the histogram will be constructed. The default is all observations for each variable.

**SYMBOLS sentence**

The SYMBOLS sentence is used to specify the symbols representing different variables in displaying the histogram bars. The default symbol is `X' for all variables.

**SCALE sentence**

The SCALE sentence is used to specify the number of data points a symbol represents in the histogram. The default number is calculated by the system, which usually is 1.

**WIDTH sentence**

The WIDTH sentence is used to specify the width (number of character symbols) for each histogram bar and the space between histogram bars. The default is 2 characters for the width (i1) and 1 character for the space (i2).

**HEIGHT sentence**

The HEIGHT sentence is used to specify the (row) height (number of character symbols) for the histogram. The default is 40 characters.

**ZERO sentence**

The ZERO sentence is used to specify whether an interval of zero frequency will be plotted. The default is ZERO, i.e., intervals of zero frequency will be displayed. If the intervals of zero frequency are to be suppressed, then NO ZERO must be specified.

## DPLOT Paragraph

The DPLOT paragraph is used to create and display the dispersion of a variable, or a sequence of the dispersion of a variable conditional on the values assumed by one or two associated variables. Also displayed are the sample mean and standard deviation of the variable.

## Syntax for the DPLOT Paragraph

### Brief syntax

```
DPLOT   VARIABLES ARE v1, v2, v3.
```

### Full syntax

```
DPLOT       VARIABLES ARE v1, v2, v3. @
            SIZE IS i.                @
            NPLOTS IS i.              @
            SPAN IS i1, i2.

Required sentence: VARIABLES
```

## Sentences Used in the DPLOT Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the name of the variable to be plotted and the names of the variables, if any, whose levels will be used for the creation of dispersion sub-plots. The values of the first variable specified are used in all plots. The second and the third variables, if specified, are used to define a sequence of mutually exclusive occurrences. A dispersion plot is generated for all possible combinations of values (or levels) of the second and third levels. The plot is composed of data in variable one that is present for a specific combination. The axis used for all generated plots will be the same and is that appropriate for all values of the first variable. No more than three variables may be specified. This sentence is required.

**SIZE sentence**

The SIZE sentence is used to specify the size of the plot in characters. The default size is 20.

**NPLOTS sentence**

The NPLOTS sentence is used to specify the number of dispersion plots to be displayed across a page. The default is 1 if v3 is not present and is the number of levels in v3 if v3 is present.

**SPAN sentence**

The SPAN sentence is used to specify the span of indices from i1 to i2, for which the dispersion plot is based. The default is all indices.

## DTPLOT Paragraph

The DTPLOT paragraph is used to create and display the dispersion of a variable, and one or more t-distributions for reference purposes.

## Syntax for the DTPLOT Paragraph

### Brief syntax

```
DTPLOT      VARIABLE IS v1, v2, v3.      @
            CENTERS ARE r1, r2, ---.     @
            STDEV ARE r1, r2, ---.

Required sentence: VARIABLES
```

### Full syntax

```
DTPLOT      VARIABLE IS v1, v2, v3.      @
            CENTERS ARE r1, r2, ---.     @
            STDEV ARE r1, r2, ---.       @
            SIZE IS i.                   @
            HEIGHTS IS i.                @
            NPLOTS IS i.                 @
            SPAN IS i1, i2.

Required sentence: VARIABLES
```

## Sentences Used in the DTPLOT Paragraph

**VARIABLE sentence**

The VARIABLES sentence is used to specify the name of the variable to be plotted and the names of the variables, if any, whose levels will be used for the creation of dispersion sub-plots. The values of the first variable specified are used in all plots. The second and the third variables, if specified, are used to define a sequence of mutually exclusive occurrences. A dispersion plot is generated for all possible combinations of values (or levels) of the second and third levels. The plot is composed of data in variable one that is present for a specific combination. The axis used for all generated plots will be the same and is that appropriate for all values of the first variable. No more than three variables may be specified. This sentence is required.

**CENTERS sentence**

The CENTERS sentence is used to specify the values of the center points to be used for the reference t-distributions that will be displayed. The default is that one center point will be used, the sample mean of the data set.

**STDEV sentence**

The STDEV sentence is used to specify the values of the standard deviations to be used in the construction of the reference t-distribution. The default is that the sample standard deviation will be used for all reference distributions that are constructed. If the sentence is specified, the number of standard deviations specified must be the same as center points specified in the CENTERS sentence. The first standard deviation is used with the first center point value, and so on.

**SIZE sentence**

The SIZE sentence is used to specify the size of the plot in characters. The default size is 20.

**NPLOTS sentence**

The NPLOTS sentence is used to specify the number of dispersion plots to be displayed across a page. The default is 1.

**SPAN sentence**

The SPAN sentence is used to specify the span of indices from i1 to i2, for which the dispersion plot is based. The default is all indices.

## CINTERVAL Paragraph

The CINTERVAL paragraph computes and displays a confidence interval for the mean of one or more variables. If more than one variable is specified, confidence intervals are computed for each variable separately. The default is the construction of a 95% t-interval for each variable specified. If a standard deviation is specified, then a 95% z-interval is computed. Alternative percentage value or degrees of freedom used for a t-value can also be specified.

## Syntax for the CINTERVAL Paragraph

```
CINTERVAL       VARIABLE(S) ARE v1, v2, --- .      @
                STDEV IS r.                        @
                PERCENT IS r, i.                   @
                HOLD LOWER(v), UPPER(v).

Required sentence: VARIABLE
```

## Sentences Used in the CINTERVAL Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the variable(s) for whom confidence intervals will be constructed. If more than one variable is specified, a confidence interval will be computed for each variable separately.

**STDEV sentence**

The STDEV sentence is used to specify a standard deviation to use in the construction of a z-interval for the mean. The default is to use the sample standard deviation of each variable specified, and the t-distribution, for a confidence interval for each variable.

**PERCENT sentence**

The PERCENT sentence is used to specify the percentage value to use in the computation of a confidence interval. The default percent is 95. If a value under 1.0 is specified, the percent value used is r*100.

In addition to the percent value, an integer value can be specified to override, when necessary, the degrees of freedom used in determining a t-value. The default is to use one less than the sample size of each variable for the degrees of freedom in the t-distribution.

**HOLD sentence**

The HOLD sentence is used to specify those values to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. The default is that none of the values of the statistics below will be retained after the paragraph is used. The values that may be retained are:

LOWER: the lower endpoint(s) of the confidence interval(s) (the variable specified will have the same number of rows as the number of confidence intervals created)

UPPER: the upper endpoint(s) of the confidence interval(s) (the variable specified will have the same number of rows as the number of confidence intervals created)

## PPLOT Paragraph

The PPLOT paragraph is used to create and display one or more univariate probability plots, including normal, half-normal, and detrended normal probability plots. These plots are useful in checking the distributional properties of variables.

## Syntax for the PPLOT Paragraph

### Brief syntax

```
PPLOT       VARIABLE IS v1.     @
            METHOD IS w.        @
            TYPES ARE w1, w2, ---.
```

Required sentence: **VARIABLE**

### Full syntax

```
PPLOT       VARIABLE IS v1.             @
            METHOD IS w.                @
            TYPES ARE w1, w2, ---.      @
            TITLE IS 'c'.               @
            SPAN IS i1, i2.
```

Required sentence: **VARIABLE**

## Sentences Used in the PPLOT Paragraph

**VARIABLE sentence**

The VARIABLE sentence is used to specify the name of the variable for which probability plots will be produced. For the best plots it is advised that only variables having numeric and non-discrete values be used.

**METHOD sentence**

The METHOD sentence is used to specify the kind of normal scoring to be used in the plot (see the previous discussion). The available kinds are:

TUKEY -- Tukey scores (the default)
BLOM -- Blom scores
VANDW -- van der Waerden scores

**TYPES sentence**

The TYPES sentence is used to specify the kinds of plots to be produced. NORMAL requests a normal plot, HNORMAL requests a half-normal plot, and DNORMAL requests a detrended normal plot. The default is NORMAL. Any number of these types may be specified.

**TITLE sentence**

The TITLE sentence is used to specify the title for the plot(s). The specified title must be enclosed in a pair of apostrophes and have no more than 72 characters. The default is that no title will be displayed.

**SPAN sentence**

The SPAN sentence is used to specify the span of indices, from i1 to i2, for which the values will be plotted. The default is all cases.

# REFERENCE

Barker, T.B. (1985). *Quality by Experimental Design*. New York: Marcel Dekker.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). *Statistics for Experimenters*. New York: Wiley.

# CHAPTER 6

# CROSS TABULATION

In order to investigate relationships between two or more variables, we may wish to calculate and display joint frequency distributions and other relevant information. We have discussed previously, basic tabular information that can be computed and retained for a response variable and one or two explanatory variables using the TABLE paragraph (see Section 2 of Chapter 4). Cross classification tables, or cross tabulations, can also be used for this purpose. We can use information from these tables in various analyses of the level of association between the variables. In addition, we may observe how values of possibly related variables are affected by different combinations of our principal variables of interest. In this chapter we will examine the cross tabulation capabilities of the SCA System.

## 6.1 Two-way Classification

To illustrate the cross tabulation of two variables, we will consider a data set of Glass (1954) concerning occupations and possible upward mobility in England. Occupations for 3500 sets of fathers and sons were recorded. Each occupation was placed in one of five status categories. The number of father and son pairs that fell in each of the 25 possible combinations was recorded. This summary data are shown in Table 1. The status categories for fathers and sons, and the frequency counts assumed, are stored in the SCA workspace under the labels FATHER, SON, and COUNT, respectively. Also listed in Table 1 is an artificially generated variable representing the average of the income of the SONS for each status combination. This information is stored in the SCA workspace under the label INCOME.

We will use the CROSSTAB paragraph to generate and display a cross classification table between the occupation status of fathers with that of their sons. This is accomplished by entering

```
-->CROSSTAB   FATHER, SON.    WEIGHT IS COUNT.
```

**Table 1 Father-Son Data**

| *Status* | | *Frequency* | *Associated variable* |
|---|---|---|---|
| *FATHER* | *SON* | *COUNT* | *INCOME* |
| 1 | 1 | 50 | 361 |
| 1 | 2 | 45 | 512 |
| 1 | 3 | 8 | 634 |
| 1 | 4 | 18 | 684 |
| 1 | 5 | 8 | 718 |
| 2 | 1 | 28 | 431 |
| 2 | 2 | 174 | 631 |
| 2 | 3 | 84 | 652 |
| 2 | 4 | 154 | 793 |
| 2 | 5 | 55 | 814 |
| 3 | 1 | 11 | 492 |
| 3 | 2 | 78 | 670 |
| 3 | 3 | 110 | 732 |
| 3 | 4 | 223 | 831 |
| 3 | 5 | 96 | 842 |
| 4 | 1 | 14 | 496 |
| 4 | 2 | 150 | 683 |
| 4 | 3 | 185 | 763 |
| 4 | 4 | 714 | 852 |
| 4 | 5 | 447 | 852 |
| 5 | 1 | 3 | 582 |
| 5 | 2 | 42 | 696 |
| 5 | 3 | 72 | 762 |
| 5 | 4 | 320 | 864 |
| 5 | 5 | 411 | 875 |

The WEIGHT sentence is included because the FATHER and SON variables do not consist of the actual 3500 records of information. These variables comprise only the values of the 25 possible occupation status combinations. The number of observations recorded for each combination, or cell, is contained in the variable COUNT. Hence COUNT contains the relative "weight" of each cell. The following is displayed.

```
CROSS-TABULATION OF FATHER   (ROWS) BY SON      (COLUMNS)

   TABLE ENTRIES ARE CELL COUNTS

             SON-1    SON-2    SON-3    SON-4    SON-5    TOTAL
          +--------+--------+--------+--------+--------+
FATHER-1 I     50       45        8       18        8 I    129
          +                                            +
FATHER-2 I     28      174       84      154       55 I    495
          +                                            +
FATHER-3 I     11       78      110      223       96 I    518
          +                                            +
FATHER-4 I     14      150      185      714      447 I   1510
          +                                            +
FATHER-5 I      3       42       72      320      411 I    848
          +--------+--------+--------+--------+--------+
TOTAL         106      489      459     1429     1017     3500
STATISTICS FOR TABLE FREQUENCIES

CHI-SQUARE IS                                  1176.5278
  DEGREES OF FREEDOM ARE                              16
  SIGNIFICANCE LEVEL IS                        LT  .0001
```

The matrix display includes the cell count (i.e., the observed total) for each combination of values in FATHER and SON, as well as row and column totals. A summary statistic of the table, its distribution, and its statistical significance are also displayed.

The summary $\chi^2$ statistic may be used in testing whether the occupation status of the sons is independent of that of their fathers. If these two variables are independent, then the probability of an occurrence in any cell is the product of the probability of the individual occurrence of each variable separately. That is, for any choice of i, j = 1, 2, 3, 4, 5, we have

$$\text{Prob(FATHER=i and SON=j)} = \text{Prob(FATHER=i) Prob(SON=j)}$$

Based on this probability of joint occurrence, the number of counts we would "expect" to see in a particular cell is the product of these of probabilities and 3500, the total number of observations in the study. For example, in the cell where FATHER = 3 and SON = 2 we would expect a value about

$$3500\ (518/3500)\ (489/3500) = 72.37$$

as 518/3500 and 489/3500 are our best estimates of Prob(FATHER=3) and Prob(SON=2), respectively. We can compute such values for every cell. The computed $\chi^2$ statistic is a measure of how likely it is for the occurrence of the observed cell counts given the row and column totals, and an assumption of independence. Since the significance level is less than .01%, it is a highly unlikely occurrence. Hence it appears the status of a son's occupation is strongly related to his father's.

### 6.1.1 Types of variables used in classification displays

In Table 1, and in the SCA workspace, we have 4 variables, FATHER, SON, COUNT and INCOME. FATHER and SON data were cross tabulated with one another. COUNT provided information regarding the frequency of occurrence for each data pair. INCOME has not, as yet, been used. This (artificially generated) variable represents the mean level of SON's income for each FATHER and SON combination.

In the construction of classification tables in the CROSSTAB paragraph a variable in the SCA workspace can be used in one of three contexts.

A variable that will be cross tabulated with other variables is referred to as a categorization variable. The CROSSTAB paragraph requires the specification of at least one categorization variable. The categorization variables of our present example are FATHER and SON.

Each observation of a categorization variable is assumed to have a frequency (weight) of one in the construction of a cross classification table. That is, we would have one "record" for each occurrence in a study. Other weights can be specified through the use of a case weight variable. If a case weight variable is specified, each observation of every categorization variable is recorded in group counts as many times as the corresponding value of the case weight variable. For example, the cross tabulation of the above father-son data can be constructed in one of two fashions. In one case, all 3500 pairs of data could be transmitted and used. Alternatively, a case weight variable could be used specifying each of the twenty five cell counts that appear as the table entries. The latter is the way data was transmitted to the SCA System. In this example COUNT was our case weight variable. Specification of such a variable is optional, and only one case weight variable may be specified in a CROSSTAB paragraph. Whenever a case weight variable is specified, its weights are used for all other variables. In the event a case weight value is negative, then corresponding observations are removed from any display or calculation.

An associated variable is one for which statistics are computed and displayed in the same tabular form as that used for the categorization variables of the paragraph. An associated variable is not cross tabulated with any other variable. Various statistics computed for an associated variable can be useful in observing the influence categorization variables may have on it. The specification of associated variables is optional.

As an illustration the father-son data are cross tabulated once more; but now with the associated variable INCOME. Recall the variable INCOME contains artificial values that represent the mean level of the income of all sons of a particular cell.

```
-->CROSSTAB   VARIABLES ARE FATHER,  SON.    WEIGHT IS COUNT.      @
-->           ASSOCIATED IS INCOME.

   CROSS-TABULATION OF FATHER    (ROWS) BY SON       (COLUMNS)

      TABLE ENTRIES ARE CELL COUNTS

                  SON-1     SON-2     SON-3     SON-4     SON-5     TOTAL
            +---------+---------+---------+---------+---------+
   FATHER-1 I      50        45         8        18         8 I     129
            +                                                  +
   FATHER-2 I      28       174        84       154        55 I     495
            +                                                  +
   FATHER-3 I      11        78       110       223        96 I     518
            +                                                  +
   FATHER-4 I      14       150       185       714       447 I    1510
            +                                                  +
   FATHER-5 I       3        42        72       320       411 I     848
            +---------+---------+---------+---------+---------+
   TOTAL          106       489       459      1429      1017      3500

   STATISTICS FOR TABLE FREQUENCIES

   CHI-SQUARE IS                                    1176.5278
     DEGREES OF FREEDOM ARE                               16
     SIGNIFICANCE LEVEL IS                          LT  .0001

   STATISTICS FOR INCOME
      BROKEN DOWN BY FATHER    (ROWS) AND SON       (COLUMNS)

      TABLE ENTRIES ARE MEANS

                  SON-1     SON-2     SON-3     SON-4     SON-5
            +---------+---------+---------+---------+---------+
   FATHER-1 I     361       512       634       684       718 I
            +                                                  +
   FATHER-2 I     431       631       652       793       814 I
            +                                                  +
   FATHER-3 I     492       670       732       831       842 I
            +                                                  +
   FATHER-4 I     496       683       763       852       852 I
            +                                                  +
   FATHER-5 I     582       696       762       864       875 I
            +---------+---------+---------+---------+---------+
```

We obtain the same classification information as before for FATHER and SON. We are also provided with an additional table, listing the mean value of the values of INCOME for each FATHER-SON cell. The data of this variable are completely artificial. Since there is only one value per cell, all data of INCOME are displayed. If the INCOME data were meaningful, we may note the income level of sons is related to their status and is uniformly higher as a function of their father's status.

### 6.1.2 Classification table entries

In the previous examples, the only information provided in the father-son classification tables was the number of occurrences per cell. The only entry provided in the table related to

the associated variable INCOME was an average value per cell. Other table entries are available for both types of variables.

### 6.1.3 Table entries for categorization variables

Seven different statistics may be computed and displayed in classification tables for categorization variables. These statistics to be computed are specified in the ENTRIES sentence of the CROSSTAB paragraph. Keywords are used to indicate the desired statistic(s). Statistics we can compute and display, the keyword associated with each statistic, and a brief description of the statistic are listed below.

| | |
|---|---|
| Cell counts (COUNT) | Frequency tabulation for cell and row and column entries |
| Total percentage (TPCT) | Frequency tabulation for cell and row and column entries given as a percent of table total |
| Row percentage (RPCT) | Frequency tabulation for cell entries as a percent of row total |
| Column percentage (CPCT) | Frequency tabulation for cell entries as a percent of column total |
| Expected values (EXPECTED) | Theoretic number of cell entries that should be present assuming independence of row and column frequencies. The theoretic number of entries for the $ij^{th}$ cell is $R_i C_j / T$ where $R_i$ is the row count total, $C_j$ is the column count total and T is the total count for the table. |
| Residuals (RESI) | $X_{ij} - E_{ij}$ , where $X_{ij}$ is the observed frequency count and $E_{ij}$ is the expected frequency for the $ij^{th}$ cell entry |
| Standardized residuals (SRESI) | Residuals for the $ij^{th}$ cell entry scaled by the square root of the expected value for the cell; i.e., $(X_{ij} - E_{ij})/(E_{ij})^{1/2}$ |

The default display is that of COUNT only. We may use the keyword ALL if we want the computation and display of all statistics listed above.

To illustrate the calculation and display of more than one table entry, we will create a cross classification table for the father-son data that also includes expected values (under the assumption of independence).

```
-->CROSSTAB   VARIABLES ARE FATHER,  SON.    WEIGHT IS COUNT.      @
-->           ENTRIES ARE COUNTS,  EXPECTED.
```

```
CROSS-TABULATION OF FATHER   (ROWS) BY SON      (COLUMNS)

   TABLE ENTRIES ARE ...
    LINE  1 IS CELL COUNTS
    LINE  2 IS EXPECTED VALUES

            SON-1    SON-2    SON-3    SON-4    SON-5    TOTAL
          +--------+--------+--------+--------+--------+
FATHER-1 I     50       45        8       18        8 I    129
         I   3.91    18.02    16.92    52.67    37.48 I
          +--------+--------+--------+--------+--------+
FATHER-2 I     28      174       84      154       55 I    495
         I  14.99    69.16    64.92   202.10   143.83 I
          +--------+--------+--------+--------+--------+
FATHER-3 I     11       78      110      223       96 I    518
         I  15.69    72.37    67.93   211.49   150.52 I
          +--------+--------+--------+--------+--------+
FATHER-4 I     14      150      185      714      447 I   1510
         I  45.73   210.97   198.03   616.51   438.76 I
          +--------+--------+--------+--------+--------+
FATHER-5 I      3       42       72      320      411 I    848
         I  25.68   118.48   111.21   346.23   246.40 I
          +--------+--------+--------+--------+--------+
TOTAL         106      489      459     1429     1017     3500


STATISTICS FOR TABLE FREQUENCIES

CHI-SQUARE IS                                  1176.5278
  DEGREES OF FREEDOM ARE                              16
  SIGNIFICANCE LEVEL IS                        LT  .0001
```

### 6.1.4 Table entries for associated variables

We can request the calculation and display of seven statistics of associated variables. Statistics are based on the values of the associated variable corresponding to the row-column entries of the categorization variables. No statistics are computed for row or column totals. We specify those statistics to be computed and displayed for the associated variables in the AENTRIES sentence of the CROSSTAB paragraph. As in the case of categorization variables, keywords are used in this sentence. Available statistics, associated keywords, and a brief description of each statistic are given below. The default display statistic is the sample mean.

| | |
|---|---|
| Mean values (MEAN) | Sample mean of corresponding entries of the associated variable |
| Variance (VARIANCE) | Sample variance of corresponding entries of the associated variable |
| Standard deviation (SD) | Square root of the above variance |
| Standard error of the mean (SEMEAN) | Standard deviation divided by the square root of the frequency count of corresponding entries of the associated variable |

| | |
|---|---|
| Maximum value (MAXIMUM) | Maximum of corresponding entries of the associated variable |
| Minimum value (MINIMUM) | Minimum of corresponding entries of the associated variable |
| Sample range (RANGE) | Range for corresponding entries of the associated variable |

## 6.2 Statistical Measures Derived From Two-Way Classification Tables

In addition to the $\chi^2$ statistic described previously, several other summary statistics of classification tables are available. These statistics may be useful in determining the level of association that may exist between categorization variables. The STATISTICS sentence is used to specify those statistics we want to compute and display. Keywords are employed to indicate the particular statistics. Listed below are those statistics we may request, the keyword associated with the statistic, and a brief description of the statistic. The default is that only the $\chi^2$ statistic is calculated. If we want all available statistics, we may use the keyword ALL.

**Chi-square (CHISQ)**

A sum of squares statistic based on the observed and expected frequency counts in each cell of the table

$$\chi^2 = \sum (f_{i0} - f_{iE})^2 / f_{iE} \text{ where}$$

$f_{i0}$ = frequency count of cell i, $f_{iE}$ = expected frequency for cell i.

**Cramer's V-statistic (V)**

An adjustment to the $\chi 2$ statistic above

$$V = \left[ \chi^2 / N(\min(r-1, c-1)) \right]^{1/2} \text{ where}$$

N = total number of counts of table; r = number of rows; c = number of columns

Note: When we have a 2x2 table (i.e., r=c=2) V is also known as the phi-statistic of the table.

**Contingency coefficient (C)**

A measure of association based on the $\chi 2$ statistic

$$C = (\chi^2 / (\chi^2 + N))^{1/2} \text{ where } N = \text{total number of counts of table}$$

**Lambda (LAMBDA)**

A measure of the information in a row (column) that may be used as a predictor of information of the other index (Goodman and Kruskal, 1954). The value is based on the observed cell frequencies and it is most applicable to categorization variables of nominal levels (that is, when values of a variable are used to separate categories only and have no mathematic consequence). There are three measures that are provided: for a column's dependency on a row; for a row's dependency on a column; and symmetric dependency. Calculations are given below

$$\lambda(\text{column dependent on row}) = (\sum_i \max\ f_{R_i} - \max\ f_C)/(N - \max\ f_C)$$

$$\lambda(\text{row dependent on column}) = (\sum_i \max\ f_{C_j} - \max\ f_R)/(N - \max\ f_R)$$

$$\lambda(\text{symmetric}) = (\sum_i \max\ f_{R_i} + \sum_i \max\ f_{C_j} - \max\ f_C - \max\ f_R)/(2N - \max\ f_C - \max\ f_R)$$

where $\max\ f_{R_i}$ = maximum cell frequency count in row i;

$\max\ f_{C_j}$ = maximum cell frequency count in column j;

$\max\ f_R$ = maximum row frequency total;

$\max\ f_C$ = maximum column frequency total; and

$N$ = total number of counts of table

**Uncertainty coefficient (U)**

A measure of association for categorization variables of nominal levels (see above). The statistic is based on proportions of "uncertainty". The uncertainty coefficient is an indication of the reduction of the "uncertainty" of the level of one variable given that information from another variable is known. As with the lambda statistic above, values can be calculated for row (column) depending on column (row), or symmetric dependency. Calculations are given below.

$$U(\text{column given row}) = (U_j - U_i - U_{ij})/U_j$$

$$U(\text{row given column}) = (U_i + U_j - U_{ij})/U_i$$

$$U(\text{symmetric}) = 2(U_i + U_j - U_{ij})/(U_i + U_j)$$

where

$$U_{ij} = -(1/N)\sum_i \sum_j f_{ij}(\ln(\sum f_{ij}/N))$$

$$U_i = -(1/N)\sum_i [(\sum_j f_{ij})(\ln(\sum_j f_{ij}/N))]$$

$$U_j = -(1/N)\sum_j[(\sum_i f_{ij})(\ln(\sum_i f_{ij}/N))], \text{ and}$$

$f_{ij}$ = cell frequency count, N = total number of counts in table

**Tau B (BTAU)**

A measure of correlation in square tables (i.e., the number of rows equals the number of columns) between ordinal variables. An ordinal variable is one in which observations can be ordered from smallest value to largest value. The statistic is based on the ordering of all pairs of information from the classification table

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + X)(P + Q + Y)}}$$

where

P = number of concordant pairs in table, i.e., the number of pairs of observations $(x_{ij}, x_{km})$ for which $x_{ij} \le x_{km}$ and either i < k and j < m or i > k and j > m;

Q = number of discordant pairs in table, i.e., the number of pairs of observations $(x_{ij}, x_{km})$ for which $x_{ij} \le x_{km}$ and either i > k and j < m or i < k and j > m ;

X = number of pairs of observations $(x_{ij}, x_{km})$ for which $x_{ij} \le x_{km}$, i = k and $j \ne m$ ; and

Y = number of pairs of observations $(x_{ij}, x_{km})$ for which $x_{ij} \le x_{km}$, $i \ne k$ and j = m

**Tau C (CTAU)**

An adjustment to τb described above when the table is not square (i.e., the number of rows and the number of column differ)

$$\tau_c = [2(P - Q)\min(r,c)]/[N^2(\min(r,c) - 1)]$$

where P and Q are defined as in $\tau_b$ above and N = total number of counts

**Conditional Gamma (GAMMA)**

A measure of monotonicity. The statistic is the difference in the number of concordant and discordant pairs divided by the number of all discordant or concordant pairs

$\gamma = (P - Q)/(P + Q)$, where P and Q are defined as in $\tau_b$ above

**Somer's D (D)**

A version of the statistic $\tau_b$ above for ordinal variables where $\tau_b$ may be adjusted for row or column ties in pairs

$$d_{R|C} = (P - Q)/(P + Q + X),$$

$$d_{C|R} = (P - Q)/(P + Q + Y),$$

$$d_{RC} = (P - Q)/(P + Q + \tfrac{1}{2}(X + Y))$$

where P, Q, X and Y are defined as in $\tau_b$ above.

To illustrate the statistics that may be generated, we will cross tabulate the father-son data again. We will now request that all possible statistics from the table be computed and displayed.

```
-->CROSSTAB   VARIABLES ARE FATHER,  SON.    WEIGHT IS COUNT.  @
-->           ENTRIES ARE COUNTS,  EXPECTED.   NO COMBINED.  @
-->           STATISTICS ARE ALL.
```

```
CROSS-TABULATION OF FATHER   (ROWS) BY SON      (COLUMNS)

   TABLE ENTRIES ARE CELL COUNTS

             SON-1    SON-2    SON-3    SON-4    SON-5    TOTAL
          +--------+--------+--------+--------+--------+
FATHER-1  I     50       45        8       18        8 I    129
          +                                             +
FATHER-2  I     28      174       84      154       55 I    495
          +                                             +
FATHER-3  I     11       78      110      223       96 I    518
          +                                             +
FATHER-4  I     14      150      185      714      447 I   1510
          +                                             +
FATHER-5  I      3       42       72      320      411 I    848
          +--------+--------+--------+--------+--------+
TOTAL         106      489      459     1429     1017     3500


   TABLE ENTRIES ARE EXPECTED VALUES

             SON-1    SON-2    SON-3    SON-4    SON-5
          +--------+--------+--------+--------+--------+
FATHER-1  I   3.91    18.02    16.92    52.67    37.48 I
          +                                             +
FATHER-2  I  14.99    69.16    64.92   202.10   143.83 I
          +                                             +
FATHER-3  I  15.69    72.37    67.93   211.49   150.52 I
          +                                             +
FATHER-4  I  45.73   210.97   198.03   616.51   438.76 I
          +                                             +
FATHER-5  I  25.68   118.48   111.21   346.23   246.40 I
          +--------+--------+--------+--------+--------+
```

```
STATISTICS FOR TABLE FREQUENCIES

CHI-SQUARE                                      1176.5278
  DEGREES OF FREEDOM                                   16
  SIGNIFICANCE LEVEL                            LT  .0001
CRAMER'S V                                         0.2899
CONTINGENCY COEFFICIENT                             0.5016
LAMBDA ASYMMETRIC, SON      DEPENDENT               0.0302
LAMBDA ASYMMETRIC, FATHER   DEPENDENT               0.0299
LAMBDA SYMMETRIC                                    0.0500
UNCERTAINTY ASYMMETRIC, SON      DEPENDENT          0.0825
UNCERTAINTY ASYMMETRIC, FATHER   DEPENDENT          0.0816
UNCERTAINTY SYMMETRIC                               0.0820
KENDALL'S TAU B                                     0.3457
KENDALL'S TAU C                                     0.3075
CONDITIONAL GAMMA                                   0.2733
SOMER'S D ASYMMETRIC, SON      DEPENDENT            0.3459
SOMER'S D ASYMMETRIC, FATHER   DEPENDENT            0.3456
SOMER'S D SYMMETRIC                                 0.3457
```

## 6.3 Table Displays for One or Two Variables

In both of the last two examples, we requested the display of the frequency count for each cell and the expected count if the variables were independent. The actual displays differed slightly as in one case all information appeared in one table, while in the other two separate tables appeared. We will now explain the differences in the table displays for a two-way table, and how we can access various display forms.

### 6.3.1 Displays for a two-way table

When exactly two categorization variables are cross tabulated we can display computed cell statistics either jointly in one two-way table or have a separate two-way table displayed for each statistic. In the preceding two examples, the cell frequency counts and expected number of counts per cell are calculated. In the default display, information was provided in the following single table.

```
CROSS-TABULATION OF FATHER   (ROWS) BY SON      (COLUMNS)

   TABLE ENTRIES ARE ...
    LINE  1 IS CELL COUNTS
    LINE  2 IS EXPECTED VALUES

            SON-1    SON-2    SON-3    SON-4    SON-5    TOTAL
         +--------+--------+--------+--------+--------+
FATHER-1 I     50       45        8       18        8 I    129
         I   3.91    18.02    16.92    52.67    37.48 I
         +--------+--------+--------+--------+--------+
FATHER-2 I     28      174       84      154       55 I    495
         I  14.99    69.16    64.92   202.10   143.83 I
         +--------+--------+--------+--------+--------+
FATHER-3 I     11       78      110      223       96 I    518
         I  15.69    72.37    67.93   211.49   150.52 I
         +--------+--------+--------+--------+--------+
FATHER-4 I     14      150      185      714      447 I   1510
         I  45.73   210.97   198.03   616.51   438.76 I
         +--------+--------+--------+--------+--------+
FATHER-5 I      3       42       72      320      411 I    848
         I  25.68   118.48   111.21   346.23   246.40 I
         +--------+--------+--------+--------+--------+
TOTAL         106      489      459     1429     1017     3500
```

Such a joint display of cell statistics is termed a <u>combined display</u> of the tabulated information.

Alternatively, we can choose not to have a combined display. We did this in the latter example by including the logical sentence NO COMBINED in the paragraph. As a result, the following two tables appeared

```
CROSS-TABULATION OF FATHER   (ROWS) BY SON      (COLUMNS)

   TABLE ENTRIES ARE CELL COUNTS

            SON-1    SON-2    SON-3    SON-4    SON-5    TOTAL
         +--------+--------+--------+--------+--------+
FATHER-1 I     50       45        8       18        8 I    129
         +                                             +
FATHER-2 I     28      174       84      154       55 I    495
         +                                             +
FATHER-3 I     11       78      110      223       96 I    518
         +                                             +
FATHER-4 I     14      150      185      714      447 I   1510
         +                                             +
FATHER-5 I      3       42       72      320      411 I    848
         +--------+--------+--------+--------+--------+
TOTAL         106      489      459     1429     1017     3500
```

```
   TABLE ENTRIES ARE EXPECTED VALUES

             SON-1    SON-2    SON-3    SON-4    SON-5
          +--------+--------+--------+--------+--------+
FATHER-1  I   3.91    18.02    16.92    52.67    37.48 I
          +                                            +
FATHER-2  I  14.99    69.16    64.92   202.10   143.83 I
          +                                            +
FATHER-3  I  15.69    72.37    67.93   211.49   150.52 I
          +                                            +
FATHER-4  I  45.73   210.97   198.03   616.51   438.76 I
          +                                            +
FATHER-5  I  25.68   118.48   111.21   346.23   246.40 I
          +--------+--------+--------+--------+--------+
```

### 6.3.2 One-way display

Although not frequently used, we can request the tabulation of a single variable. Such one-way tables may be useful for the display of frequency information of grouped data or character data of a single variable. In addition, one-way tables may be useful for the display of statistics of associated variables.

To obtain a one-way table for the variable FATHER alone, we simply enter

```
-->CROSSTAB   FATHER.   WEIGHT IS COUNT.
```

```
FREQUENCY DISTRIBUTION OF FATHER

   TABLE ENTRIES ARE CELL COUNTS

   FATHER-1 FATHER-2 FATHER-3 FATHER-4 FATHER-5      TOTAL
   +--------+--------+--------+--------+--------+
   |    129      495      518     1510      848 |     3500
   +--------+--------+--------+--------+--------+
```

Frequency information for a variable can also be generated using either the HISTOGRAM or DPLOT paragraph (see Chapter 5). The categorization variable must contain only numeric values in order to use either of these paragraphs.

## 6.4 Grouping and Labeling Observations of a Categorization Variable

Categorization variables of the CROSSTAB paragraph may consist of either character or numeric values. Associated variables may assume numeric values only. If a categorization variable is composed of character values, the character strings will be used as labels on all tables.

If we have a categorization variable whose observations are measurements of a continuous process, or are coded values of traits (e.g., sex, hair color, eye color), it may be advantageous to group values into sets (or categories) and label these sets for display purposes. The CATEGORIES sentence is used to group individual values or ranges of values of a categorization variable into specified categories and associate a label with each category.

In addition to grouping values into levels, we can also "translate" numeric values to labels for display purposes.

To illustrate grouping and labeling, consider the values of the FATHER and SON variables. The values represent status levels for occupations. We could group and label values in the following manner

1 or 2 : BELOW average
3 : AVERAGE
4 or 5 : ABOVE average

We can designate these groups for the values of FATHER only by entering the following

```
-->CROSSTAB   FATHER, SON.    WEIGHT IS COUNT.                @
-->         CATEGORIES ARE FATHER ( 1,2, 'BELOW' / 3, 'AVERAGE' / 4,5, 'ABOVE').

CASES PROCESSED:   1 THROUGH    25  (   25 CASES ARE USED)

CROSS-TABULATION OF FATHER   (ROWS) BY SON      (COLUMNS)

   TABLE ENTRIES ARE CELL COUNTS

            SON-1    SON-2    SON-3    SON-4    SON-5    TOTAL
          +--------+--------+--------+--------+--------+
   BELOW I     78       219        92       172        63 I    624
          +                                              +
 AVERAGE I     11        78       110       223        96 I    518
          +                                              +
   ABOVE I     17       192       257      1034       858 I   2358
          +--------+--------+--------+--------+--------+
TOTAL         106       489       459      1429      1017     3500

STATISTICS FOR TABLE FREQUENCIES

CHI-SQUARE IS                                    694.2610
  DEGREES OF FREEDOM ARE                            8
  SIGNIFICANCE LEVEL IS                          LT  .0001
```

The column designations (SON) have remained unchanged but the first two rows, and last two rows, of our original table are now grouped (added) together. We specified this grouping (and labeling) with the sentence

CATEGORIES ARE FATHER (1, 2, 'BELOW' / 3, 'AVERAGE' / 4, 5, 'ABOVE')

This sentence can be interpreted as "Group the following information together for the variable named FATHER:

(a) 1 or 2 (and label as 'BELOW'),

(b) 3 (and label as 'AVERAGE'), and

(c) 4 or 5 (and label as 'ABOVE')" .

Labels are not required, but they are handy. Labels are distinguished by being between apostrophes. The symbol '/' is used to separate group "definitions".

We can group the data of SON in the same manner by "extending" the CATEGORIES sentence

```
-->CROSSTAB   FATHER, SON.    WEIGHT IS COUNT.                    @
-->   CATEGORIES ARE FATHER ( 1,2, 'BELOW' / 3, 'AVERAGE' / 4,5, 'ABOVE'),   @
-->   SON ( 1,2, 'BELOW' / 3, 'AVERAGE' / 4,5, 'ABOVE').
```

```
CASES PROCESSED:   1 THROUGH    25  (   25 CASES ARE USED)

CROSS-TABULATION OF FATHER   (ROWS) BY SON      (COLUMNS)

   TABLE ENTRIES ARE CELL COUNTS

              BELOW  AVERAGE    ABOVE    TOTAL
          +--------+--------+--------+
   BELOW  I    297       92      235 I    624
          +                         +
 AVERAGE  I     89      110      319 I    518
          +                         +
   ABOVE  I    209      257     1892 I   2358
          +--------+--------+--------+
TOTAL         595      459     2446     3500


STATISTICS FOR TABLE FREQUENCIES

CHI-SQUARE IS                                  605.5792
  DEGREES OF FREEDOM ARE                              4
  SIGNIFICANCE LEVEL IS                        LT  .0001
```

If the values of a categorization variable spanned a wider range of values, we can define intervals for various categories. For example, suppose the status level for FATHER is a value between 0 and 100 (instead of the current 1, 2, 3, 4, 5). Moreover, suppose that due to a coding problem the following ranges of values make up our three intervals

'BELOW' average : 11 to 45

'AVERAGE' : 46 to 75

'ABOVE' average : 0 to 10 or 76 to 100

We can group and label these values by the inclusion of the following sentence in the CROSSTAB paragraph

```
CATEGORIES ARE FATHER (11 thru 45,  'BELOW' / 46 thru 75,  'AVERAGE' / @
     0 thru 10,  76 thru 100,  'ABOVE').
```

The word THRU can be used in the specification of ranges. In addition, more than one range can be specified for a grouping.

The CATEGORIES sentence, like all modifying sentences, does not have to begin on a new line and may be continued over a number of lines. We may need to extend information over a few lines, depending on the number of categorization variables involved and the extent of our grouping instructions.

To illustrate the "breaking" of the CATEGORIES sentence, we will re-specify the last example. We will also include the ENTRIES sentence to obtain more information per cell.

```
-->CROSSTAB   FATHER, SON.   WEIGHT IS COUNT.        @
-->   ENTRIES ARE COUNTS, EXPECTED.   @
-->   CATEGORIES ARE FATHER ( 1,2, 'BELOW' / 3, 'AVERAGE' /   @
-->   4,5, 'ABOVE'), SON ( 1,2, 'BELOW' / 3, 'AVERAGE' / 4,5, 'ABOVE').

CASES PROCESSED:   1 THROUGH    25  (   25 CASES ARE USED)

CROSS-TABULATION OF FATHER   (ROWS) BY SON      (COLUMNS)

   TABLE ENTRIES ARE ...
    LINE  1 IS CELL COUNTS
    LINE  2 IS EXPECTED VALUES

             BELOW  AVERAGE    ABOVE    TOTAL
          +--------+--------+--------+
   BELOW I    297       92      235 I    624
          I 106.08    81.83   436.09 I
          +--------+--------+--------+
 AVERAGE I     89      110      319 I    518
          I  88.06    67.93   362.01 I
          +--------+--------+--------+
   ABOVE I    209      257     1892 I   2358
          I 400.86   309.23  1647.91 I
          +--------+--------+--------+
TOTAL         595      459     2446     3500


STATISTICS FOR TABLE FREQUENCIES

CHI-SQUARE IS                                  605.5792
  DEGREES OF FREEDOM ARE                              4
  SIGNIFICANCE LEVEL IS                        LT  .0001
```

If a categorization variable contains only integer values, then the SCA System automatically creates default categories and labels. One category is provided for each integer, beginning with the smallest value and ending with largest value. Hence a large number of categories could be generated if the integer values do not represent the desirable grouping and, as a result, a table may be created that is not meaningful. The CATEGORIES sentence can be used to reduce the number of groups that will be used or displayed as well as to provide more meaningful labeling information of each group. Note that all categorization variables that have any non-integer values must be grouped through the CATEGORIES sentence. Otherwise, a large number of meaningless categories may be generated.

Additional information regarding the CATEGORIES sentence is presented in Section 6.1.

## 6.5 Tables For Three or More Variables

The CROSSTAB paragraph is not limited to one or two categorization variables. To illustrate cross classification of more than two variables, we will consider data of Morrison et al (1973) on a study of the survival of breast cancer patients who received radiation therapy. The data are shown in Table 2. The classification categories (together with the label used to store the data in the SCA workspace) for the 767 recorded observations are:

(1) CENTER -- the cancer center where the patient was diagnosed: Boston, Glamorgan, or Tokyo.

(2) AGE -- the age at which the patient was diagnosed: 1 (0-49), 2 (50-69), or 3 (70 and older)

(3) SURVIVED -- whether or not the patient survived for three years: yes or no

(4) INFLAMM -- a measure of inflammation: 1 (minimal) or 2 (greater)

(5) APPEAR -- a measure of appearance: 1 (malignant) or 2 (benign)

The number of observations in each of the 72 categories is listed and stored in the SCA workspace under the label COUNTS.

## TABLE 2 Survival Data

| Cancer center<br>CENTER | Age of diagnosis<br>AGE | Survival information<br>SURVIVED | Measure of inflammation<br>INFLAMM | Measure of appearance<br>APPEAR | Number of observations in the category<br>COUNTS |
|---|---|---|---|---|---|
| TOKYO | 1 | NO | 1 | 1 | 9 |
| TOKYO | 1 | NO | 1 | 2 | 7 |
| TOKYO | 1 | NO | 2 | 1 | 4 |
| TOKYO | 1 | NO | 2 | 2 | 3 |
| TOKYO | 1 | YES | 1 | 1 | 26 |
| TOKYO | 1 | YES | 1 | 2 | 68 |
| TOKYO | 1 | YES | 2 | 1 | 25 |
| TOKYO | 1 | YES | 2 | 2 | 9 |
| TOKYO | 2 | NO | 1 | 1 | 9 |
| TOKYO | 2 | NO | 1 | 2 | 9 |
| TOKYO | 2 | NO | 2 | 1 | 11 |
| TOKYO | 2 | NO | 2 | 2 | 2 |
| TOKYO | 2 | YES | 1 | 1 | 20 |
| TOKYO | 2 | YES | 1 | 2 | 46 |
| TOKYO | 2 | YES | 2 | 1 | 18 |
| TOKYO | 2 | YES | 2 | 2 | 5 |
| TOKYO | 3 | NO | 1 | 1 | 2 |
| TOKYO | 3 | NO | 1 | 2 | 3 |
| TOKYO | 3 | NO | 2 | 1 | 1 |
| TOKYO | 3 | NO | 2 | 2 | 0 |
| TOKYO | 3 | YES | 1 | 1 | 1 |
| TOKYO | 3 | YES | 1 | 2 | 6 |
| TOKYO | 3 | YES | 2 | 1 | 5 |
| TOKYO | 3 | YES | 2 | 2 | 1 |
| BOSTON | 1 | NO | 1 | 1 | 6 |
| BOSTON | 1 | NO | 1 | 2 | 7 |
| BOSTON | 1 | NO | 2 | 1 | 6 |
| BOSTON | 1 | NO | 2 | 2 | 0 |
| BOSTON | 1 | YES | 1 | 1 | 11 |
| BOSTON | 1 | YES | 1 | 2 | 24 |
| BOSTON | 1 | YES | 2 | 1 | 4 |
| BOSTON | 1 | YES | 2 | 2 | 0 |
| BOSTON | 2 | NO | 1 | 1 | 8 |
| BOSTON | 2 | NO | 1 | 2 | 20 |
| BOSTON | 2 | NO | 2 | 1 | 3 |
| BOSTON | 2 | NO | 2 | 2 | 2 |
| BOSTON | 2 | YES | 1 | 1 | 18 |
| BOSTON | 2 | YES | 1 | 2 | 58 |
| BOSTON | 2 | YES | 2 | 1 | 10 |
| BOSTON | 2 | YES | 2 | 2 | 3 |
| BOSTON | 3 | NO | 1 | 1 | 9 |
| BOSTON | 3 | NO | 1 | 2 | 18 |
| BOSTON | 3 | NO | 2 | 1 | 3 |
| BOSTON | 3 | NO | 2 | 2 | 0 |
| BOSTON | 3 | YES | 1 | 2 | 26 |
| BOSTON | 3 | YES | 1 | 1 | 15 |
| BOSTON | 3 | YES | 2 | 1 | 1 |
| BOSTON | 3 | YES | 2 | 2 | 1 |
| GLAMORGN | 1 | NO | 1 | 1 | 16 |
| GLAMORGN | 1 | NO | 1 | 2 | 7 |
| GLAMORGN | 1 | NO | 2 | 1 | 3 |
| GLAMORGN | 1 | NO | 2 | 2 | 0 |
| GLAMORGN | 1 | YES | 1 | 1 | 16 |
| GLAMORGN | 1 | YES | 1 | 2 | 20 |
| GLAMORGN | 1 | YES | 2 | 1 | 8 |
| GLAMORGN | 1 | YES | 2 | 2 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| GLAMORGN | 2 | NO | 1 | 1 | 14 |
| GLAMORGN | 2 | NO | 1 | 2 | 12 |
| GLAMORGN | 2 | NO | 2 | 1 | 3 |
| GLAMORGN | 2 | NO | 2 | 2 | 0 |
| GLAMORGN | 2 | YES | 1 | 1 | 27 |
| GLAMORGN | 2 | YES | 1 | 2 | 39 |
| GLAMORGN | 2 | YES | 2 | 1 | 10 |
| GLAMORGN | 2 | YES | 2 | 2 | 4 |
| GLAMORGN | 3 | NO | 1 | 1 | 3 |
| GLAMORGN | 3 | NO | 1 | 2 | 7 |
| GLAMORGN | 3 | NO | 2 | 1 | 3 |
| GLAMORGN | 3 | NO | 2 | 2 | 0 |
| GLAMORGN | 3 | YES | 1 | 1 | 12 |
| GLAMORGN | 3 | YES | 1 | 2 | 11 |
| GLAMORGN | 3 | YES | 2 | 1 | 4 |
| GLAMORGN | 3 | YES | 2 | 2 | 1 |

In the example that follows we will use 4 categorization variables: CENTER, AGE, SURVIVED, and a variable to be constructed from INFLAMM and APPEAR.

### 6.5.1 Available displays

We can choose to have our classification tables displayed as either a series of two-way tables or as a multi-dimensional display. When a series of two-way tables is displayed, the last two variables are used as the row and column variable for all two-way tables. The remaining categorization variables are used as control variables. That is, a two-way table is displayed for each combination of categories for the control variables. Each table contains a label denoting the particular combination of values of control variables being used. Any statistic derived from this ``two-dimensional slice'' is displayed before another two-way table is given. Such a sequence of two-way tables can result in a great amount of output, depending on the number of variables and levels assumed by the variables.

If we include the logical sentence MTABLE in the CROSSTAB paragraph, we obtain a multi-dimensional display. In a multi-dimensional display, the two-way format is extended so that more than one variable generates rows for the tables. Variables are “nested” within another so that any single “row” represents a specific combination of variable values (or categories). Hence we have a table in which a column variable is cross tabulated jointly with more than one variable.

We will now cross tabulate the cancer study survival data using a multi-dimensional table. First we will use an analytic statement (see Appendix A) to create a new variable from INFLAMM and APPEAR.

```
-->INFLAPP = (INFLAMM - 1)*2 + APPEAR
```

We now have a variable, INFLAPP, that assumes 4 values:

1 : if minimal inflammation and malignant appearance (MIN-MAL);
2 : if minimal inflammation and benign appearance (MIN-BEN);
3 : if greater inflammation and malignant appearance (GRT-MAL); or
4 : if greater inflammation and benign appearance (GRT-BEN).

We will use the CATEGORIES sentence to re-label the values of INFLAPP to those noted in parenthesis above and also provide labels for the three categories of AGE.

```
-->CROSSTAB  VARIABLES ARE CENTER, AGE, SURVIVED, INFLAPP.      @
-->   CATEGORIES ARE AGE(1, 'UNDER50'/ 2, '50-60'/ 3, 'ABOVE69'),      @
-->   INFLAPP(1, 'MIN-MAL'/ 2, 'MIN-BEN'/ 3, 'GRT-MAL'/ 4, 'GRT-BEN').  @
-->   WEIGHT IS COUNTS.  MTABLE.
```

**CROSS-TABULATION OF SURVIVED (ROWS) BY INFLAPP (COLUMNS)**

**TABLE ENTRIES ARE CELL COUNTS**

| CENTER | AGE | SURVIVED | INFLAPP | | | | |
|---|---|---|---|---|---|---|---|
| | | | MIN-MAL | MIN-BEN | GRT-MAL | GRT-BEN | TOTAL |
| BOSTON | UNDER 50 | NO | 6 | 7 | 6 | 0 I | 19 |
| | | YES | 11 | 24 | 4 | 0 I | 39 |
| | | TOTAL | 17 | 31 | 10 | 0 I | 58 |
| | 50-69 | NO | 8 | 20 | 3 | 2 I | 33 |
| | | YES | 18 | 58 | 10 | 3 I | 89 |
| | | TOTAL | 26 | 78 | 13 | 5 I | 122 |
| | ABOVE 69 | NO | 9 | 18 | 3 | 0 I | 30 |
| | | YES | 15 | 26 | 1 | 1 I | 43 |
| | | TOTAL | 24 | 44 | 4 | 1 I | 73 |
| GLAMORGN | UNDER 50 | NO | 16 | 7 | 3 | 0 I | 26 |
| | | YES | 16 | 20 | 8 | 1 I | 45 |
| | | TOTAL | 32 | 27 | 11 | 1 I | 71 |
| | 50-69 | NO | 14 | 12 | 3 | 0 I | 29 |
| | | YES | 27 | 39 | 10 | 4 I | 80 |
| | | TOTAL | 41 | 51 | 13 | 4 I | 109 |
| | ABOVE 69 | NO | 3 | 7 | 3 | 0 I | 13 |
| | | YES | 12 | 11 | 4 | 1 I | 28 |
| | | TOTAL | 15 | 18 | 7 | 1 I | 41 |
| TOKYO | UNDER 50 | NO | 9 | 7 | 4 | 3 I | 23 |
| | | YES | 26 | 68 | 25 | 9 I | 128 |
| | | TOTAL | 35 | 75 | 29 | 12 I | 151 |
| | 50-69 | NO | 9 | 9 | 11 | 2 I | 31 |
| | | YES | 20 | 46 | 18 | 5 I | 89 |
| | | TOTAL | 29 | 55 | 29 | 7 I | 120 |
| | ABOVE 69 | NO | 2 | 3 | 1 | 0 I | 6 |
| | | YES | 1 | 6 | 5 | 1 I | 13 |
| | | TOTAL | 3 | 9 | 6 | 1 I | 19 |

The last categorical variable that we specified, INFLAPP, has been used as the column variable for the entire table. Other variables are sequentially nested "backwards" to the first variable we specified, CENTER. That is, the levels of the variable SURVIVED are varied within those of the variable AGE, that in turn are varied within those of CENTER.

We should be aware that the display of a row of a multi-dimensional table along a single unbroken line may not be possible because there may be too many categorical variables or too many categories assumed by the "column" variable (i.e., the last variable). If we want an unbroken row in the table we may need to either appropriately "choose" our "column" variable or adjust the number of categories displayed for our "column" variable.

### 6.5.2 Display of cell statistics

Similar to that of a two-way display, we have two choices regarding the display of computed statistics of a cell in the classification table(s) for three or more categorization variables. We may have either a combined display or non-combined display for these cells. The default, as before, is a combined display.

If our cross classification is displayed as a sequence of two-way tables, then cell statistics are displayed exactly as that described previously for two-way tables (see Section 6.2). That is, all cell information is presented either as a single combined display or as an additional sub-sequence of two-way tables, one for each cell statistic.

For a multi-dimensional table, a combined display is similar to that of a non-combined one. The difference between displays occurs only in the display of cell information for the "inner-most" categorization variable of a row. This variable (the second to last variable specified) is the one whose labeling information is changed most frequently along the left most side of the table.

As an illustration, consider the breast cancer example above. The "column" variable for the table is INFLAPP, while the "rows" of the table are CENTER, AGE and SURVIVED, with SURVIVED being the inner-most variable. If we had included the sentence

```
ENTRIES ARE COUNT, TPCT, RPCT.
```

in the above CROSSTAB paragraph, we are provided with frequency counts, total percents and row percents for each cell.

In a combined display we would obtain the following (output is edited for presentation purposes)

```
CASES PROCESSED:   1 THROUGH    72  (   72 CASES ARE USED)

CROSS-TABULATION OF SURVIVED (ROWS) BY INFLAPP  (COLUMNS)

   TABLE ENTRIES ARE ...
    LINE  1 IS CELL COUNTS
    LINE  2 IS TOTAL PERCENTS
    LINE  3 IS ROW PERCENTS

 CENTER   AGE      SURVIVED                   INFLAPP
 -------- -------- --------                   --------
                             MIN-MAL  MIN-BEN  GRT-MAL  GRT-BEN   TOTAL
+--------+--------+--------+--------+--------+--------+--------+--------

   BOSTON  UNDER 50      NO       6        7        6        0 I     19
                               10.34    12.07    10.34      .00 I  32.76
                               31.58    36.84    31.58      .00 I
                          YES      11       24        4        0 I     39
                               18.97    41.38     6.90      .00 I  67.24
                               28.21    61.54    10.26      .00 I
                      +--------+--------+--------+--------+--------+--------
                       TOTAL         17       31       10        0 I     58
                               29.31    53.45    17.24      .00 I 100.00

             50-69       NO       8       20        3        2 I     33
                                6.56    16.39     2.46     1.64 I  27.05
                               24.24    60.61     9.09     6.06 I
                          YES      18       58       10        3 I     89
                               14.75    47.54     8.20     2.46 I  72.95
                               20.22    65.17    11.24     3.37 I
                      +--------+--------+--------+--------+--------+--------
                       TOTAL         26       78       13        5 I    122
                               21.31    63.93    10.66     4.10 I 100.00
                      .                         .                   .
                      .                         .                   .
                      .                         .                   .
```

Here all cell information is displayed within one cell entry, as in a two-way table.

In a non-combined display, the above portion of the display is presented as

```
   TABLE ENTRIES ARE ...
   SUBTABLE 1 CONTAINS CELL COUNTS
   SUBTABLE 2 CONTAINS TOTAL PERCENTS
   SUBTABLE 3 CONTAINS ROW PERCENTS

 CENTER   AGE      SURVIVED                  INFLAPP
 -------- -------- --------                  --------
                             MIN-MAL  MIN-BEN  GRT-MAL  GRT-BEN   TOTAL
+--------+--------+--------+--------+--------+--------+--------+--------

   BOSTON  UNDER 50      NO       6        7        6        0 I     19
                         YES      11       24       4        0 I     39
                  +--------+--------+--------+--------+--------+--------
                   TOTAL          17       31       10       0 I     58

                         NO    10.34    12.07    10.34     .00 I  32.76
                         YES   18.97    41.38     6.90     .00 I  67.24
                  +--------+--------+--------+--------+--------+--------
                   TOTAL       29.31    53.45    17.24     .00 I 100.00

                         NO    31.58    36.84    31.58     .00 I
                         YES   28.21    61.54    10.26     .00 I
                  +--------+--------+--------+--------+--------+--------

              50-69       NO       8       20        3        2 I     33
                         YES      18       58       10        3 I     89
                  +--------+--------+--------+--------+--------+--------
                   TOTAL          26       78       13        5 I    122

                         NO     6.56    16.39     2.46     1.64 I  27.05
                         YES   14.75    47.54     8.20     2.46 I  72.95
                  +--------+--------+--------+--------+--------+--------
                   TOTAL       21.31    63.93    10.66     4.10 I 100.00

                         NO    24.24    60.61     9.09     6.06 I
                         YES   20.22    65.17    11.24     3.37 I
                  +--------+--------+--------+--------+--------+--------
```

Here the cell information has been separated into sub-tables within the larger display, but separate multi-dimensional tables for each of the cell statistics are not created.

We see the only change between the two displays is how cell information for the last two variables is presented (with other variables held "fixed"). If we want a separate display of tables for each cell statistic, then we need to invoke the CROSSTAB paragraph repeatedly, each time specifying a new cell statistic.

### 6.5.3 Statistical measures derived from classification tables of three or more categorization variables

Section 2 describes statistical measures that can be derived from two-way tables. These statistics can also be calculated and displayed for every pair of variables when three or more categorization variables are specified.

However, since these table statistics are computed from two-way displays, they may only be requested when the classification tables are displayed in a series of two-way tables. In this case statistics for each two-way table are provided before another two-way table is displayed. If we choose a multi-dimensional display of the variables (i.e., MTABLE is specified), then no statistical measure of the table may be calculated.

## 6.6 Miscellaneous Display Information

This section provides information on infrequently used features of the CROSSTAB paragraph. It may be omitted at first reading and used for later reference.

### 6.6.1 Secondary categories of the CATEGORIES sentence

The CATEGORIES sentence (see Section 6.4) is useful for grouping or labeling information contained in a variable. An example used to illustrate this sentence for the father-son data was

```
CATEGORIES ARE FATHER(1, 2,  'BELOW' / 3,  'AVERAGE' / 4, 5,  'ABOVE').
```

It is possible that we may not designate groups for all data of a variable. However, we may wish to somehow keep track of this data. Secondary categories are those categories that will be displayed in tables but whose values are not used in the generation of statistics pertaining to the table. Secondary categories must be the last categories specified in the CATEGORIES sentence. There are two secondary categories that may be specified:

```
OTHER (or OTHERS)  : for those values not categorized into any group, and
MISSING            : for missing data values.
```

Only data defined to be in a category will be used for computational purposes in a cross tabulation. In this manner cross tabulations can be made for subsets of categorization variables.

For example, if the variable FATHER contained missing data, we could include this as a category by augmenting the above CATEGORIES sentence to

```
CATEGORIES ARE FATHER (1, 2, 'BELOW'/ 3, 'AVERAGE'/ 4, 5, 'ABOVE'/ MISSING)
```

### 6.6.2 Missing data

Missing data are usually treated as a secondary category and are not used in a cross tabulation. Information regarding where missing data occur is not displayed on tables unless the secondary category MISSING is specified in the CATEGORIES sentence (see above). On occasion we may want missing data included in the calculation of statistics from a cross classification table. The logical sentence MISSING is used to specify whether missing data

are to be counted in the computation of table statistics. We may use this logical sentence in conjunction with the specification of a MISSING secondary category to control both the display and the use of missing data in the calculations of statistical measures of a table. The MISSING sentence interacts with the secondary category specification of MISSING of the CATEGORIES sentence in the following manner:

| Specification within CATEGORIES sentence | Specification given MISSING sentence | Result |
|---|---|---|
| no MISSING category specified | NO MISSING | Missing values do not appear in table and are not used in statistical measures of the table |
| no MISSING category specified | MISSING | Missing values do not appear in table but are used in statistical measures of the table |
| MISSING category specified | NO MISSING | Missing values appear in table but are excluded from statistical measures of the table |
| MISSING category specified | MISSING | Missing values appear in table and are included in statistical measures of the table |

### 6.6.3 Displays for associated variables

We may have any number of associated variables in the CROSSTAB paragraph. Statistics of associated variables are displayed in tables that follow the cross classification tables of the categorical variables.

All tables related to one associated variable are displayed before those of another associated variable. The row and column labels of these tables are the same as those used in the display of the categorization variables, but the cells contain those statistics that are to be calculated for the associated variable (see Section 2.2 of this Chapter). Cell statistics are presented in the same combined or non-combined fashion as the categorization variables.

# SUMMARY OF THE SCA PARAGRAPH IN CHAPTER 6

This section provides a summary of the SCA paragraph employed in this chapter. The syntax is presented in both a brief and full form. The brief display of the syntax contains the most frequently used sentences of the paragraph, while the full display presents all possible modifying sentences of the paragraph. In addition, special remarks related to a paragraph may also be presented with the description.

Each SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

In this section we provide a summary of the CROSSTAB paragraph.

Legend (see Chapter 2 for further explanation)

v : variable name
i : integer
r : real value
w : keyword (label)

## CROSSTAB Paragraph

The CROSSTAB paragraph is used to perform one-way to n-way cross tabulation. The TABLE paragraph (see Chapter 4) can also be used to provide simple tabular statistics for a response variable with one or two explanatory variables. Values of categorization variables may be set into groups. Group frequencies, row, column and total percentages, expected group counts, residuals, and standardized residuals are available. Chi-square statistics and a variety of measures of association can be produced. In addition, statistics may be requested on an associated set of variables. This allows, for example, the computation of marginal means or group variances.

## Syntax of the CROSSTAB paragraph

### Brief syntax

```
CROSSTAB  VARIABLES ARE v1, v2, ---.                          @
          ASSOCIATED-VARIABLES ARE v1, v2, ---.               @
          CATEGORIES ARE v1(group 1 values, 'label1'/         @
          group 2 values, 'label2'/ . . .                     @
          /OTHER/MISSING),                                    @
          v2(group 1 values, 'label1'/                        @
          group 2 values, 'label2'/ . . .                     @
          /OTHER/MISSING), ---.                               @
          ENTRIES ARE w1, w2, ---.                            @
          AENTRIES ARE w1, w2, ---.                           @
          MTABLE. / NO MTABLE.
```

Required sentence: **VARIABLES**

### Full syntax

```
CROSSTAB  VARIABLES ARE v1, v2, ---.                          @
          ASSOCIATED-VARIABLES ARE v1, v2, ---.               @
          WEIGHT IS v.                                        @
          CATEGORIES ARE v1(group 1 values, 'label1'/         @
          group 2 values, 'label2'/ . . .                     @
          /OTHER/MISSING),                                    @
          v2(group 1 values, 'label1'/                        @
          group 2 values, 'label2'/ . . .                     @
          /OTHER/MISSING), ---.                               @
          COMBINED./NO COMBINED.                              @
          MISSING./NO MISSING.                                @
          SPAN IS i1, i2.                                     @
          EMPTY IS 'c'.                                       @
          TITLE IS 'c'.                                       @
          OWIDTH IS i.                                        @
          ENTRIES ARE w1, w2, ---.                            @
          AENTRIES ARE w1, w2, ---.                           @
          STATISTICS ARE w1, w2, ---.                         @
          MTABLE./NO MTABLE.
```

Required sentence: **VARIABLES**

## Sentences Used in the CROSSTAB Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the categorization variables for the table(s). It is the only required sentence for the CROSSTAB paragraph. The table has as many dimensions as the number of variables specified here. See Section 6.1.1 for a further explanation.

**ASSOCIATED-VARIABLES sentence**

The ASSOCIATED-VARIABLES sentence is used to specify the names of associated variables (if any) for which statistics are to be computed. See Section 6.1.1 for a further explanation.

**WEIGHT sentence**

The WEIGHT sentence is used to specify the name of a case weight variable. Each observation of the categorization variables is entered into group counts as many times as the corresponding value of the weight variable. Any case with a negative weight is ignored in the computations. The default is to assign each case a weight of one.

**CATEGORIES sentence**

The CATEGORIES sentence is used to specify groupings to be used for categorization variable, values and labels for those groupings. For each categorization variable, a category grouping is defined by a list of values (r1, r2, ...) and an associated label (`label'). The list of values may contain ranges in the form r1 THRU r2 as well as individual values. If the variable takes on integer values from 1 to p, then no values need to be specified (if default labels are acceptable). If a variable does not appear in the CATEGORIES sentence, each different observed value of that variable will be used as a separate category. Secondary categories (OTHER or MISSING) must appear at the end of the category specification for each variable. Information in the secondary categories are not used in computing statistics. See Sections 6.4 and 6.1 for a further explanation.

**COMBINED sentence**

The COMBINED sentence is used to specify whether the requested table entries (see ENTRIES and AENTRIES sentences) are to be displayed in combined tables or not. COMBINED requests joint tables for the entries specified in ENTRIES and AENTRIES. However the entries specified in the ENTRIES and AENTRIES sentences are never displayed within the same cell. The default option is COMBINED. See Section 6.3.1 for a further explanation.

**MISSING sentence**

The MISSING sentence is used to specify whether missing values are to be counted in the computation of table statistics or not. Specify MISSING if missing data are to be included in computations and NO MISSING if missing data are to be excluded from computations. NO MISSING is the default. See Section 6.6 for a further explanation.

**SPAN sentence**

The SPAN sentence is used to specify the span of cases, from i1 to i2, of each variable from which the table(s) will be constructed. The default is all observations.

**EMPTY sentence**

The EMPTY sentence is used to specify a character string to be displayed in place of the cell with no information. The maximum width of the string is 8 characters. The default is to display NONE.

**TITLE sentence**

The TITLE sentence is used to specify a title for the table produced. The length of the title is limited to 72 characters. The default is no title.

**OWIDTH sentence**

The OWIDTH sentence is used to specify the maximum print width allowed for displaying the tables and statistics. The default is to use the current system-wide output width.

**ENTRIES sentence**

The ENTRIES sentence is used to specify statistics that will be computed and displayed for each cell of the classification tables. Available keywords are:

| | | |
|---|---|---|
| COUNT | -- | cell counts |
| TPCT | -- | total percentage |
| RPCT | -- | row percentage |
| CPCT | -- | column percentage |
| EXPECTED | -- | expected values |
| RESI | -- | residuals |
| SRESI | -- | standardized residuals |
| ALL | -- | all of the above |

The default is COUNT, report group counts only. See Section 6.1.3 for a further explanation.

**AENTRIES sentence**

The AENTRIES sentence is used to specify those descriptive statistics that are to be displayed for in each cell of the classification table for any associated variables. Available keywords are:

| | | |
|---|---|---|
| MEAN | -- | variable mean |
| VARIANCE | -- | variable variance |
| SD | -- | variable standard deviation |
| SEMEAN | -- | standard error of the mean |
| MAXIMUM | -- | maximum value in each group |
| MINIMUM | -- | minimum value in each group |
| RANGE | -- | range of values in each group |
| ALL | -- | all of the above |

The default is MEAN to display variable means. See Section 6.1.4 for a further explanation.

**STATISTICS sentence**

The STATISTICS sentence is used to specify the statistical measures to be reported for each table that is constructed from categorical variables. See Section 6.2 for a further discussion. Available keywords are:

| | | |
|---|---|---|
| CHISQ | -- | chi-square |
| V | -- | Cramer's V (Phi for 2 x 2) |
| C | -- | contingency coefficient |
| LAMBDA | -- | lambda |
| U | -- | uncertainty coefficients |
| BTAU | -- | tau B |
| CTAU | -- | tau C |
| GAMMA | -- | conditional gamma |
| D | -- | Somer's D statistics |
| ALL | -- | all of the above |

Note: The specification of the MTABLE sentence precludes the computation of summary statistical measures of the table.

**MTABLE sentence**

The MTABLE sentence is used to specify that the resultant tables be displayed in a multi-dimensional form (see Section 6.5.1). The display is adjusted as required so that the table will fit properly in the given page width.

If NO MTABLE is specified, then the complete n-dimensional table is displayed in two-dimensional slices, each a two-way contingency table. The first (n-2) variables are used as control values for the last two variables. Note that specifying MTABLE precludes the computation of summary statistical measures of the table. NO MTABLE is the default.

## ACKNOWLEDGEMENT

Scientific Computing Associates gratefully appreciates the programming assistance of Philip Burns in the development of the CROSSTAB paragraph.

## REFERENCES

Bailey, B.J.R. (1980). "Large Sample Simultaneous Confidence Intervals for the Multinomial Probabilities Based on Transformations of the Cell Frequencies". *Technometrics 22*: 583-589.

Fisher, R.A. (1935). "The Logic of Inductive Inference (with discussion)". *Journal of the Royal Statistical Society 98*: 39-54.

Glass, D.V. (ed.) (1954). *Social Mobility in Britain*, Glencoe, IL: The Free Press.

Goodman, L.A., and Kruskal, W.H. (1954). "Measures of Association for Cross-Classification". *Journal of the American Statistical Association 49*: 732-764.

Morrison, A.S., Black, M.M., Lowe, C.R., MacMahon, B., and Yuasa, S. (1973). "Some International Differences in Historical and Survival in Breast Cancer". *International Journal of Cancer II*: 261-267.

Yule, G.U. (1912). "On the Methods of Measuring Association Between Two Attributes". *Journal of the Royal Statistical Society 75*: 579-642.

# CHAPTER 7

# COMPARING TWO SAMPLES

Often it is informative to measure the differences between two data sets. A study of the magnitude of a difference between related samples can highlight causes for any disparities. For example, one production process may produce better results than another, or one medical treatment may prove substantially better than another. One way we can study the differences between two data sets is by examining the difference of the means of the samples. To assess the magnitude of a difference, we need to take the variability of each set of data into consideration. The TTEST paragraph of the SCA System permits the study of the difference of means under a variety of situations.

The TTEST paragraph provides:

(1) Summary information for each data set, including sample mean and a 95% confidence interval for the mean. This provides us with a quick visual representation of the two data sets.

(2) The test statistic $\overline{x}_1 - \overline{x}_2$ and its t-value. The t-value is

$$\frac{\overline{x}_1 - \overline{x}_2}{S},$$

where S is an estimate of the standard error of the difference. The estimate of standard error can vary, depending upon the assumptions we make for the data sets. These assumptions will also affect the degrees of freedom associated with the t-statistic. This information is also provided.

(3) Confidence intervals for the difference in means. In addition to implications concerning whether means are equal, we can examine other hypotheses.

## 7.1 Paired Comparisons

A direct method to study the difference between effects is through a paired comparison. In such a study we derive measurements from pairs of "like" entities. These could be observations taken from the same group of people at two distinct time points or from any two sets of things that have been paired according to a similarity of traits. One member of each pair is provided "Treatment A" and the other member is provided "Treatment B". "Treatments" can be different drugs, environments, production processes, and the like. "No treatment" (i.e., not exposing a subject to a factor) can itself be a "treatment" in the case of a controlled experiment. Whenever possible the assignment of "treatments" should be random.

**Example: Shoe wear data**

As an example of a paired comparison, we will analyze the measurements of the amount of wear of the soles of shoes worn by 10 boys (Box, Hunter, and Hunter, 1978, page 98). Two different synthetic fibers are used, A and B. The fibers are allocated to the right or left shoe of each boy at random. The data are displayed in Table 1. The amounts of wear for fibers A and B are stored in the SCA workspace under the labels WEARA and WEARB, respectively.

**Table 1 Shoe wear data**

| *Boy* | *Material A WEARA* | *Material B WEARB* | *Difference in wear* |
|---|---|---|---|
| 1 | 13.2 | 14.0 | -.8 |
| 2 | 8.2 | 8.8 | -.6 |
| 3 | 10.9 | 11.2 | -.3 |
| 4 | 14.3 | 14.2 | .1 |
| 5 | 10.7 | 11.8 | -1.1 |
| 6 | 6.6 | 6.4 | .2 |
| 7 | 9.5 | 9.8 | -.3 |
| 8 | 10.8 | 11.3 | -.5 |
| 9 | 8.8 | 9.3 | -.5 |
| 10 | 13.3 | 13.6 | -.3 |

We observe the wear associated with fiber B is greater than that of fiber A in all but two cases. In order to determine if this is significant we will derive a paired t-test for the difference in sample means.

```
-->TTEST   WEARA, WEARB.  PAIRED.

SUMMARY AND CONFIDENCE INTERVAL PLOT FOR EACH GROUP

                                  8.100     9.600    11.100    12.600    14.100
GROUP  N    MEAN   STD DEV   SE MEAN +---------+---------+---------+---------+
  1   10   10.630    2.451     .775      (-----------*----------)
  2   10   11.040    2.518     .796         (-----------*----------)
                                    +---------+---------+---------+---------+
DIFFERENCE OF MEANS =     -.410     STANDARD ERROR OF THE DIFFERENCE =     .122
T-VALUE             =    -3.349     DEGREES OF FREEDOM               =       9

95% CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS: (    -0.687,    -0.133 )
99% CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS: (    -0.808,    -0.012 )
```

The initial summary data displayed of each sample does not indicate a discernable difference. The means are about the same and both samples have about the same variance. However, this is a paired comparison. We need to derive conclusions from the sample mean and standard deviation of the differenced data, and not just the difference in sample means.

The sample mean of the differenced data (column 3 of Table 1) is -0.410, and its sample standard error is 0.122. The t-value, mean divided by its standard error, measures whether the mean is statistically different from zero or not. If the mean is not different from zero, then the wear due to material A would not be distinguishable from that of material B. The computed t-value is -3.3449 which is significant at both the 5% and 1% levels for a t-value with 9 degrees of freedom. Hence there is a significant difference between materials. This can also be seen in the information provided for the confidence interval of the difference of means. Both the 95% and 99% confidence intervals do not contain zero.

## 7.2 Independent Samples

If we do not have a paired study, then we will examine the difference of sample means directly. That is, we obtain $\bar{x}_1 - \bar{x}_2$ and its t-value. The t-value is $(\bar{x}_1 - \bar{x}_2)/s$, where s is the standard error of the difference.

### 7.2.1 Pooled and unpooled estimate of standard error

The standard error of the difference of two independent samples is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} \tag{7.1}$$

If we are testing the hypothesis that there is no difference between treatments, then the notion of this hypothesis is sometimes extended to an assumption that both samples have approximately the same variability. With such an assumption the information of both samples can be pooled to calculate the common variance.

The pooled estimate of the common variance, denoted by $S_p^2$, is

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{7.2}$$

where $s_1$ and $s_2$ are the estimated standard errors of the first and second samples, respectively. The size of the samples are $n_1$ and $n_2$, respectively. The estimate of S is then

$$S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{7.3}$$

The t-statistic computed has $n_1 + n_2 - 2$ degrees of freedom.

In some situations, the assumption that both samples have approximately the same variance can be seriously flawed. If there is reason to doubt the validity of such an assumption, we should not pool information to obtain a common variance. In such cases, the variance of each sample is estimated separately and used in the calculation of S. Here

$$S_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and the computed t-statistic has approximately the following degrees of freedom

$$\text{d.f.} = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} \tag{7.4}$$

All calculations are automatically computed within the SCA System.

### 7.2.2 An experiment of equal sample sizes, hormone data

An example from Snedecor and Cochran (1989, page 91) is used to demonstrate a comparison of data from groups of equal size. The 15-day average comb weights of two lots of male chicks, one receiving hormone A (testosterone) and the other hormone C (dehydroandrosterone), are compared. The data are presented in Table 2 and are stored in the SCA workspace under the labels WEIGHTA and WEIGHTC, respectively.

**Table 2 Hormone Data**

| Chick | Weight using Hormone A WEIGHTA | Weight using Hormone C WEIGHTC |
|---|---|---|
| 1 | 57.0 | 89.0 |
| 2 | 120.0 | 30.0 |
| 3 | 101.0 | 82.0 |
| 4 | 137.0 | 50.0 |
| 5 | 119.0 | 39.0 |
| 6 | 117.0 | 22.0 |
| 7 | 104.0 | 57.0 |
| 8 | 73.0 | 32.0 |
| 9 | 53.0 | 96.0 |
| 10 | 68.0 | 31.0 |
| 11 | 118.0 | 88.0 |

We will employ the TTEST paragraph twice. First we will calculate the standard error of the difference using a pooled estimate of a common variance. This is the default of the TTEST paragraph.

-->TTEST WEIGHTA, WEIGHTC

```
SUMMARY AND CONFIDENCE INTERVAL PLOT FOR EACH GROUP

                                  30.000    55.000    80.000   105.000   130.000
GROUP  N    MEAN   STD DEV   SE MEAN +---------+---------+---------+---------+
  1   11   97.000   28.478    8.586                     (------*------)
  2   11   56.000   28.478    8.586    (------*-------)
                                     +---------+---------+---------+---------+

POOLED VARIANCE =      811.0000     POOLED STANDARD ERROR =     28.4781

DIFFERENCE OF MEANS =    41.000     STANDARD ERROR OF THE DIFFERENCE =   12.143
T-VALUE             =     3.376     DEGREES OF FREEDOM               =       20

95% CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS: (    15.670,    66.330 )
99% CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS: (     6.449     75.551 )
```

The computed t-value is highly significant, indicating a difference in weights due to the hormones. Hormone A produces significantly higher weights than Hormone C.

As a check to see if a pooled estimate of variance is appropriate in this case, we will re-calculate the t-statistic using (7.1) to calculate the standard error, and (7.4) to determine the appropriate degrees of freedom. Since our sample sizes are equal, (7.4) reduces to

$$\text{d.f.} = \frac{(n-1)(s_1^2 + s_2^2)^2}{s_1^4 + s_2^4} \tag{7.5}$$

-->TTEST WEIGHTA, WEIGHTC. NO POOLED

```
SUMMARY AND CONFIDENCE INTERVAL PLOT FOR EACH GROUP

                                  30.000    55.000    80.000   105.000   130.000
GROUP  N    MEAN   STD DEV   SE MEAN +---------+---------+---------+---------+
  1   11   97.000   29.107    8.776                     (------*------)
  2   11   56.000   27.835    8.393     (-----*------)
                                     +---------+---------+---------+---------+

DIFFERENCE OF MEANS =    41.000     STANDARD ERROR OF THE DIFFERENCE =   12.143
T-VALUE             =     3.376     DEGREES OF FREEDOM               =       20

95% CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS: (    15.670,    66.330 )
99% CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS: (     6.449,    75.551 )
```

We obtain the same results as before. The reason for the same results is directly attributable to the fact we have equal sample sizes and the standard deviations of each series are approximately the same.

### 7.2.3 An example with unequal sample sizes, tomato data

To illustrate the analysis involving unequal sample sizes, we will use the data from a randomized design with two factors. The data are the yield, in pounds, of tomatoes that have been treated with two different fertilizers, A and B (Box, Hunter, and Hunter, 1978, page 94). Eleven plants were used, each randomly assigned one of the fertilizers. The data are given in Table 3 and are stored in the SCA workspace under the labels YIELDA and YIELDB.

**Table 3 Tomato data**

| *Yield from fertilizer A YIELDA* | *Yield from fertilizer B YIELDB* |
|---|---|
| 29.9 | 26.6 |
| 11.4 | 23.7 |
| 25.3 | 28.5 |
| 16.5 | 14.2 |
| 21.1 | 17.9 |
| | 24.3 |

As in the previous example, the TTEST paragraph will be used twice. In the first case a pooled estimate of standard error will be used. In the second case, (7.1) is used directly.

```
-->TTEST   YIELDA, YIELDB

SUMMARY AND CONFIDENCE INTERVAL PLOT FOR EACH GROUP

                                  12.000    17.000    22.000    27.000    32.000
GROUP  N    MEAN   STD DEV   SE MEAN +---------+---------+---------+---------+
  1    5   20.840    6.303    2.819       (------------*-----------)
  2    6   22.533    6.303    2.819            (-----------*-----------)
                                     +---------+---------+---------+---------+

POOLED VARIANCE =       39.7250     POOLED STANDARD ERROR =      6.3028

DIFFERENCE OF MEANS =    -1.693     STANDARD ERROR OF THE DIFFERENCE =    3.817
T-VALUE             =     -.444     DEGREES OF FREEDOM               =        9

95% CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS: (   -10.327,     6.940 )
99% CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS: (   -14.096,    10.710 )
```

-->TTEST YIELDA, YIELDB. NO POOLED.

```
SUMMARY AND CONFIDENCE INTERVAL PLOT FOR EACH GROUP

                                    11.000    16.000    21.000    26.000    31.000
GROUP  N    MEAN    STD DEV   SE MEAN +---------+---------+---------+---------+
  1    5   20.840    7.246    3.240     (---------------*--------------)
  2    6   22.533    5.432    2.218               (---------*----------)
                                      +---------+---------+---------+---------+

DIFFERENCE OF MEANS =    -1.693     STANDARD ERROR OF THE DIFFERENCE =    3.926
T-VALUE             =     -.431     DEGREES OF FREEDOM               =        7

95% CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS: (   -10.978,     7.591 )
99% CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS: (   -15.434,    12.047 )
```

Since the sample sizes of each group are nearly equal and the standard errors are not too different, the standard error of the difference, and resultant t-values, are about the same in each case. There does not appear to be any difference in yield due to fertilizers. This is confirmed in the confidence intervals that are displayed, as they all contain 0.0. We note the confidence intervals are different in the two cases due to differing standard errors and degrees of freedom for the t-value employed.

# SUMMARY OF THE SCA PARAGRAPH IN CHAPTER 7

This section provides a summary of the SCA paragraph employed in this chapter. The paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

Legend for variables (see Chapter 2 for further explanation):

v: variable name

## TTEST Paragraph

The TTEST paragraph is used to calculate and display a two-sample t-test.

## Syntax of the TTEST Paragraph

```
TTEST     VARIABLES ARE v1, v2.     @
          PAIRED. / NO PAIRED.      @
          POOLED. / NO POOLED.

Required: VARIABLES
```

## Sentences Used in the TTEST Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to list the names of the two variables to be analyzed.

**PAIRED sentence**

The PAIRED sentence is used to indicate the data comes from a paired study and the paired t-test will be used. The default is NO PAIRED.

**POOLED sentence**

The POOLED sentence is used to specify that the pooled variance will be used in the calculation of the standard error of the difference. This is the default condition.

# REFERENCES

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). *Statistics for Experimenters*. New York: Wiley.

Snedecor, G.W., and Cochran, W.G. (1989). *Statistical Methods, 8th Edition*. Ames, IA: The Iowa State Press.

# CHAPTER 8

# ANALYSIS OF VARIANCE

The analysis of variance is an extension of the comparison of means as presented in the previous chapter. By analyzing variances in each group jointly, we will be able to construct a test of the hypothesis whether the means of several groups of a factor are the same or not. The method can be extended so that we may also examine interactions between factors. The SCA System provides three paragraphs for the analysis of variance:

OWAY: for one-way analysis of variance

TWAY: for one-way or two-way analysis of variance

NWAY: for multi-way analysis of variance

The TWAY and NWAY paragraphs also permit incorporation of a Box-Cox transformation analysis (see Section 8.3) and the inclusion of covariate variables (see Section 8.4).

## 8.1 One-Way Analysis of Variance

A simple extension of the comparison of the means of two samples can be made for more than two samples. In Chapter 7, we examined the difference of sample means as a check of whether the effects of two "treatments" are the same, or not. When the "treatments" cause a significant difference in the sample mean of a response variable, we are able to determine which treatment is more effective.

In a one-way analysis of variance, we examine the hypothesis that two or more "treatments" have the same effect. We do so by representing a "typical" treatment response as

$$\hat{Y}_j = (\text{grand mean}) + (\text{effect of treatment}_j), \tag{8.1}$$

where $\hat{Y}_j$ is the fitted value of the $j^{th}$ "treatment". We can interpret (8.1) as follows. In the absence of "treatments", the best representation we may provide for a typical response is the (grand) mean of our sample. If "treatments" are present, we can make adjustments from this mean level according to the effect of each treatment. Clearly, if all treatments have the same effect, we do not need to make any adjustment from the grand mean.

We will study whether "treatments" are the same, or not, by examining the variation in the data using the model (8.1). If all "treatments" are the same, then the common effect of all treatments can be incorporated into the grand mean. As a result, the total variation we will observe is that corresponding to random error about the grand mean. That is,

total variation = variation due to error (i.e., from the mean) .

If there is some appreciable difference in treatments, then we should be able to "partition" the total variation observed as

(total variation) = (variation due to treatments )
+ (variation due to error) (8.2)

The OWAY paragraph is used to perform a one-way analysis of variance. There are two distinct ways to specify response information in the OWAY paragraph, as discussed in Sections 8.1.1 and 8.1.4.

### 8.1.1 Example: Blood coagulation times

To illustrate a one-way analysis of variance, we consider the coagulation time (in seconds) for samples of blood drawn from 24 animals, each subjected to one of four diets, A, B, C, and D. The data, and the related analysis, may be found in Box, Hunter and Hunter (1978, Chapter 6). Diets were assigned in a random order. The data are presented in Table 1. The coagulation times and assignment orderings for each diet are stored in the SCA workspace in the variables COAGA, COAGB, COAGC, COAGD, ORDERA, ORDERB, ORDERC and ORDERD, respectively.

**Table 1 Coagulation data from Box, Hunter and Hunter**

| Diet (randomization order) | | | |
|---|---|---|---|
| A<br>COAGA | B<br>COAGB | C<br>COAGC | D<br>COAGD |
| 62 (20) | 63 (12) | 68 (16) | 56 (23) |
| 60 ( 2) | 67 ( 9) | 66 ( 7) | 62 ( 3) |
| 63 (11) | 71 (15) | 71 ( 1) | 60 ( 6) |
| 59 (10) | 64 (14) | 67 (17) | 61 (18) |
| | 65 ( 4) | 68 (13) | 63 (22) |
| | 66 ( 8) | 68 (21) | 64 (19) |
| | | | 63 ( 5) |
| | | | 59 (24) |

It is good practice to use some exploratory data techniques before conducting a more extensive statistical analysis, such as an analysis of variance. However, we will defer using such techniques at the moment. We can use the OWAY paragraph to perform an analysis of variance by entering

-->OWAY COAGA, COAGB, COAGC, COAGD

we obtain the following

```
ANALYSIS OF VARIANCE FOR THE FOLLOWING VARIABLES:
COAGA    COAGB    COAGC    COAGD

S =        2.3664      R**2 =  67.1%      R**2(ADJ) =  62.1%

--------------------------
ANALYSIS OF VARIANCE TABLE
--------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

TREATMNT (A)        228.000     3        76.000      13.571
RESIDUAL            112.000    20         5.600
ADJ. TOTAL          340.000    23


SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =     64.000 )

FACTOR:  TREATMNT                58.500    61.500    64.500    67.500    70.500
LEVEL  N    MEAN    STD DEV  STD ERR +---------+---------+---------+---------+
  1    4   61.000     1.826    1.183 (------*--------)
  2    6   66.000     2.828     .966                (------*------)
  3    6   68.000     1.673     .966                       (------*-----)
  4    8   61.000     2.619     .837   (----*-----)
                                     +---------+---------+---------+---------+
```

## 8.1.2 Sums of squares of the ANOVA table

The entries within the ANALYSIS OF VARIANCE TABLE correspond to the components of (8.2). The sum of squares corresponding to "total variation" is given in the ADJ. TOTAL row. This sum of squares, 340.0, is further divided into a TREATMNT and RESIDUAL component.

The sum of squares associated with RESIDUAL corresponds to the sum of squares of "variation due to error". This value is also known as the **within-treatment sum of squares**. The sum of square corresponding to "variation due to treatments" is given in the TREATMNT row. This corresponds to the coagulation times associated with various diets used by the animals. This sum of squares is also known as the **between-treatment sum of squares**.

Each of the between-treatment sum of squares and the within-treatment sum of squares may be divided by its associated degrees of freedom (DF) to yield an estimate for the variance of an observation. Each estimate is displayed in the MEAN SQUARE column. We assume the variance of any individual observation is the same.

If the true means of the treatments actually vary, then the TREATMNT mean square will be an inflated estimate of the variance while the RESIDUAL mean square will be unaffected. The quotient of these terms, 76.00/5.6 = 13.571, can then be used as a test

statistic. It should be compared to values of the F(k-1, n-k) distribution, where k is the number of treatments. We observe an inflated value of the mean square for TREATMNT compared to that of RESIDUAL. The computed F-value is significant at the .001 level. Hence, the means are not the same and we may conclude the diets produce different effects.

### 8.1.3 Other output from the OWAY paragraph

In addition to the ANOVA table, the OWAY paragraph produces a set of summary information for the various treatment levels (here diets) used in the analysis.

A one-line summary for each treatment (diet) is provided after the analysis of variance table. Information is provided for the factors according to the order they are specified in the OWAY paragraph. Level 1 corresponds to COAGA, level 2 corresponds to COAGB, and so on. We are provided with the number of observations having the factor, and the sample mean and standard deviation from the responses of this group. In addition to the standard deviation of the group (STD DEV), we are also provided with the standard error of the mean for the factor (STD ERR). The value for the standard error is $s/\sqrt{n_i}$, where the value s is the pooled estimate of standard error (here 2.3664) and $n_i$ is the number of observations for the level (4 for COAGA, 6 for COAGB and COAGC, and 8 for COAGD). Also displayed is a 95% confidence interval for the mean. This interval is constructed using the STD ERR value.

### 8.1.4 Using OWAY with treatment information contained in a single variable

It is often convenient to maintain all treatment information in a single variable and their corresponding responses in another variable. We can use the JOIN paragraph (see Appendix B) to append all the coagulation time information into the single variable COAGTIME. Similarly, the randomization orders are appended into the variable ORDER.

```
-->JOIN  COAGA, COAGB, COAGC, COAGD.  NEW IS COAGTIME.
-->JOIN  ORDERA, ORDERB, ORDERC, ORDERD.  NEW IS ORDER.
```

We can create a variable of diets that corresponds to the above variables by using the INPUT paragraph (see Chapter 2). We may enter (SCA System prompts and messages are not shown)

```
-->INPUT   DIET.  PRECISION  IS  CHARACTER.
-->A A A A   B B B B B B   C C C C C C   D D D D D D D D
-->END OF DATA
```

The content of the resultant variables is shown in Table 2.

We now have our responses in a single variable, COAGTIME, and their corresponding treatment is in another variable, DIET. We can still employ the OWAY paragraph when our data in such columnar form. We can obtain a one-way analysis of variance for this data if we enter

```
-->OWAY   COAGTIME, DIET.
```

Unlike our previous OWAY specification, the names of the variables specified here are those of the responses, COAGTIME (the first variable specified), and of the factor levels, DIET. This is the default assumed by the OWAY paragraph when only two variables are specified.

The output from the above paragraph is suppressed as it is virtually the same as presented previously. The only significant differences that may be observed using the above two-variable specification (i.e., the specification of a single response variable and a "treatment" variable) are labeling information and the order in which summary information is displayed after the ANOVA table. Here, the factor summary is presented relative to the values assumed by the levels of the factor. That is, levels assumed by the factors are ordered (either numerically in ascending order or alphabetically), and summaries are provided for that ordering. In this case, the alphabetic ordering (A, B, C, D) results in the same display as the previous factor specification (i.e., COAGA, COAGB, COAGC, COAGD).

**Table 2 Blood coagulation data in vector form**

| Coagulation time COAGTIME | Diet DIET | Test order ORDER |
|---|---|---|
| 62.00 | A | 20 |
| 60.00 | A | 2 |
| 63.00 | A | 11 |
| 59.00 | A | 10 |
| 63.00 | B | 12 |
| 67.00 | B | 9 |
| 71.00 | B | 15 |
| 64.00 | B | 14 |
| 65.00 | B | 4 |
| 66.00 | B | 8 |
| 68.00 | C | 16 |
| 66.00 | C | 7 |
| 71.00 | C | 1 |
| 67.00 | C | 17 |
| 68.00 | C | 13 |
| 68.00 | C | 21 |
| 56.00 | D | 23 |
| 62.00 | D | 3 |
| 60.00 | D | 6 |
| 61.00 | D | 18 |
| 63.00 | D | 22 |
| 64.00 | D | 19 |
| 63.00 | D | 5 |
| 59.00 | D | 24 |

### 8.1.5 Exploratory analysis of the blood coagulation data

Having the blood coagulation data in a single variable permits us to make better use of the exploratory data analysis features in the SCA System. For example, we can produce a summary table of the coagulation times for each diet by using the TABLE paragraph (see Chapter 4).

```
-->TABLE  COAGTIME, DIET.

DEPENDENT   VARIABLE : COAGTIME
CATEGORICAL VARIABLES: DIET

  ROW      MEAN     STD. DEV.   COUNTS
-----------------------------------
   1     61.000       1.826      4
   2     66.000       2.828      6
   3     68.000       1.673      6
   4     61.000       2.619      8
```

The tabular information suggests that diets produce different effects on blood coagulation time. We can visually inspect for differences by plotting the data. We use the dispersion plot capability of SCA, DPLOT (see Chapter 5), to construct a histogram for the combined data set and then for each of the diet groupings

```
-->DPLOT  COAGTIME

          X
         X      X
     XX  XX X  XXX   X
X    XXX XX XX XXX   X
+---------+*--------+
56.0    63.5     71.0
MEAN=61.0    SD=3.8


-->DPLOT  COAGTIME, DIET

    XX  XX
+-----*---+---------+
56.0    63.5     71.0
MEAN=61.0    SD=1.8

  DIET  =   A


         X XX XX    X
+---------+---*-----+
56.0    63.5     71.0
MEAN=66.0    SD=2.8

  DIET  =   B
```

```
                X
                X
               XXX   X
+---------+-----*---+
56.0    63.5     71.0
MEAN=68.0    SD=1.7

  DIET  =   C


         X
X   XXX XX X
+-----*---+---------+
56.0    63.5     71.0
MEAN=61.0    SD=2.6

  DIET  =   D
```

Our plots our revealing. Although the combined sample has a mean of 64.0, the four diet sub-groups vary extensively. Moreover, the standard errors of each of the sub-groups are approximately the same (between 1.7 and 2.8); but these values are considerable less than the standard error of the complete sample.

Even if no further analyses are conducted, we would suspect that coagulation times are affected by diet. The previous OWAY results confirm this belief.

### 8.1.6 Diagnostic checks of a fitted model

As noted in previous chapters, it is important that an analysis include diagnostic checks of the fitted model. It is useful that the residuals and fitted values of the model be maintained for diagnostic checking. We can retain the residuals and fitted values in the previous example (in the variables RES and FIT, respectively) if we enter (depending on the form our data are in) either

```
-->OWAY   COAGA,COAGB,COAGC,COAGD.   HOLD  RESIDUALS(RES), FITTED(FIT).
```

or

```
-->OWAY   COAGTIME, DIET.   HOLD  RESIDUALS(RES), FITTED(FIT).
```

For this example, the values in RES and FIT will be in the identical order regardless which form of OWAY is used.

**Dispersion plots of the residuals**

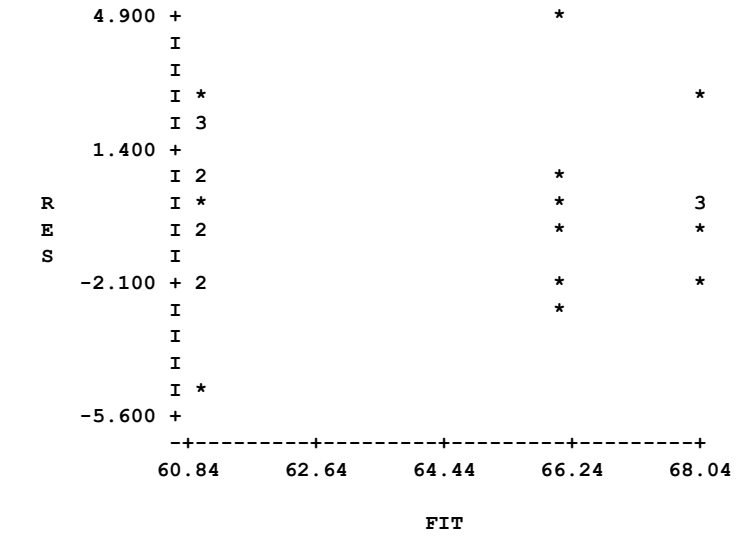
A one-way analysis of variance is similar to fitting a regression model to the data. The model fit consists of simply fitting a different mean to each group. We can then check the adequacy of our fitted model as we would a regression model (see Chapter 9.2). We can visually check the distribution of the residuals by constructing dispersion plots for all residuals, and for the residuals corresponding to each treatment. We do not observe any gross abnormalities.

```
-->DPLOT   RES

            X
         X X X
         X X X X X
         X X X X X X
   X   X X X X X X X   X
   +---------*---------+
   -5.0     -.0      5.0
   MEAN=.0      SD=2.2


-->DPLOT   RES, DIET

         X X   X X
   +---------*---------+
   -5.0     -.0      5.0
   MEAN=.0      SD=1.8

     DIET  =   A


       X X X X X       X
   +---------*---------+
   -5.0     -.0      5.0
   MEAN=.0      SD=2.8

     DIET  =   B


            X
            X
         X X X     X
   +---------*---------+
   -5.0     -.0      5.0
   MEAN=.0      SD=1.7

     DIET  =   C


               X
   X     X X X X X X
   +---------*---------+
   -5.0     -.0      5.0
   MEAN=.0      SD=2.6

     DIET  =   D
```

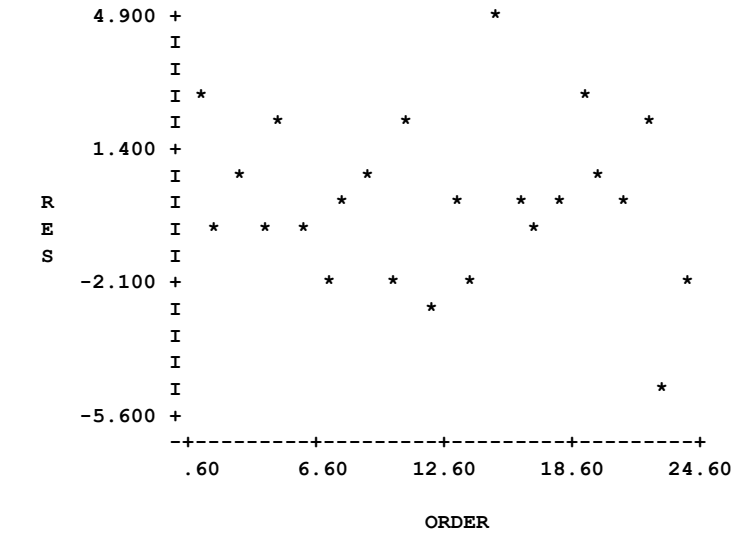
## Plots of residuals

We also consider two plots of the residuals. A plot of residuals against fitted values can reveal anomalies from the model. In addition, residuals may be plotted in time (or assignment) order to check for any time (or assignment) effect. Again, no model inadequacies are apparent.

```
-->PLOT   RES, FIT
```




```
-->PLOT   RES, ORDER
```




## 8.1.7 Example: Doughnut data

As a second example of the use of the OWAY paragraph for a one-way analysis of variance, we consider the amount of fat absorbed by doughnuts during cooking (Snedecor and Cochran, 1989, page 217). Four types of fat are examined. The amount of grams of fat absorbed per batch for each type of fat is shown in Table 3. The data are stored in the SCA workspace under the labels FAT1 through FAT4, respectively.

**Table 3 Doughnut data**

Grams of fat absorbed per batch by type of fat

| *Fat* | | | |
|---|---|---|---|
| *1* | *2* | *3* | *4* |
| *FAT1* | *FAT2* | *FAT3* | *FAT4* |
| 64 | 78 | 75 | 55 |
| 72 | 91 | 93 | 66 |
| 68 | 97 | 78 | 49 |
| 77 | 82 | 71 | 64 |
| 56 | 85 | 63 | 70 |
| 95 | 77 | 76 | 68 |

We can now use the OWAY paragraph perform a one-way analysis of variance as before. We will not perform any diagnostic checks of this model.

```
-->OWAY   FAT1, FAT2, FAT3, FAT4

ANALYSIS OF VARIANCE FOR THE FOLLOWING VARIABLES:
FAT1     FAT2     FAT3     FAT4

S =       10.0449      R**2 =  44.8%      R**2(ADJ) =  36.5%

--------------------------
ANALYSIS OF VARIANCE TABLE
--------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

TREATMNT (A)       1636.500     3       545.500       5.406
RESIDUAL           2018.000    20       100.900
ADJ. TOTAL         3654.500    23


SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =     73.750 )

FACTOR:  TREATMNT                53.000    63.000    73.000    83.000    93.000
LEVEL  N    MEAN    STD DEV  STD ERR +---------+---------+---------+---------+
  1    6   72.000    13.342    4.101          (--------*--------)
  2    6   85.000     7.772    4.101                    (--------*--------)
  3    6   76.000     9.879    4.101              (--------*--------)
  4    6   62.000     8.222    4.101 (-------*--------)
                                    +---------+---------+---------+---------+
```

The F-statistic for this data is 5.40, significant at the 1% level. Our conclusion is that the fats are absorbed differently by the doughnuts. As we see in the summary information, fat 2 appears to have the highest absorption rate, and fat 4 the lowest.

### 8.1.8 Using the TWAY paragraph for a one-way analysis of variance

If the treatment responses are contained within a single variable, and the corresponding treatment levels are in another variable, we can also employ the TWAY paragraph for a one-way analysis of variance. To illustrate this, we consider the coagulation data used previously. We can obtain a one-way analysis of variance by entering

```
-->TWAY COAGTIME, DIET
```

The variable of responses (here COAGTIME) is the first variable specified followed by the name of treatment variable (here DIET).

The ANOVA table produced by the TWAY paragraph is identical to that produced by the OWAY paragraph (with DIET replacing the generic TREATMNT label in the display). However, the factor summary information is not the same. The factor summary information of the TWAY paragraph is

```
SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =     64.000 )

FACTOR:  DIET                    58.800    61.800    64.800    67.800    70.800
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    4   61.000    61.000     .974 (-----*------)
  2    6   66.000    66.000     .845                (-----*-----)
  3    6   68.000    68.000     .845                       (-----*-----)
  4    8   61.000    61.000     .773  (----*-----)
                                    +---------+---------+---------+---------+
```

In place of the STD DEV column of the OWAY paragraph, we are provided with an ESTIMATE column. Values in this column are the least squares estimates for the means of the responses in each factor level. The data in the STD ERR column are the standard errors of these estimates (derived from a least squares fit of a corresponding linear model) and are not the standard errors of the mean. The 95% confidence interval displayed corresponds to the estimate of the mean. Hence these intervals also differ from those produced by the OWAY paragraph.

## 8.2 Two-Way Analysis of Variance

To this point we have examined the effects of two or more "treatments". However, the "treatments" may be viewed as changes made to a similar component or of the same basic factor. In the previous examples, we examined the effect of different dietary supplements or the type of fat. Each observation was subjected to only one of the available "treatments". A natural extension to this type of analysis is to concurrently apply "treatments" of two different factors. This is what is done in a two-way analysis of variance.

Terminology is changed slightly in two-way analysis of variance in order to distinguish factors and levels within factors. The components of a two-way analysis are distinguished by the terms "treatments" and "blocks". We are now able to display results in a matrix form, usually with various "blocks" as rows and "treatments" as columns. For the purpose of our discussion, we assume there are r different "blocks" and c different "treatments".

We will extend the rationale of the one-way model to incorporate two distinct factors. In a two-way model, an observation could be affected by both the "treatment" as well as the "block" it is in. The model in (8.1) can be extended to provide a fitted value for each block-treatment combination according to

$$\hat{Y}_{ij} = (\text{grand mean}) + (\text{effect of block}_i) + (\text{effect of treatment}_j) \tag{8.3}$$

The analysis of variation described in (8.2) can be extended as

$$\begin{aligned}(\text{total variation}) = &(\text{variation due to block}) \\ &+ (\text{variation due to treatment}) \\ &+ (\text{variation due to random error})\end{aligned} \tag{8.4}$$

It is also possible that, in addition to these "main" effects (i.e., caused by either block or treatment), there can also be an effect due to the **interaction** between "block" and "treatment". In such a case our fitted value of (8.3) is changed to

$$\begin{aligned}\hat{Y}_{ij} = &(\text{grand mean}) + (\text{effect of block}_i) + (\text{effect of treatment}_j) \\ &+ (\text{effect of interaction}_{ij})\end{aligned} \tag{8.5}$$

This is studied through the following decomposition of variance:

$$\begin{aligned}(\text{total variation}) = &(\text{variation due to block}) \\ &+ (\text{variation due to treatment}) \\ &+ (\text{variation due to block-treatment interaction}) \\ &+ (\text{variation due to random error})\end{aligned} \tag{8.6}$$

In many cases, we are restricted to the case of the simple **additive model** (i.e., the formulation of (8.3) and (8.4)). This restriction may be result of our inability to provide a design to estimate an interaction effect (as two or more responses for each block and treatment combination are required).

### 8.2.1 Example: Penicillin production process data

To illustrate a two-way analysis of variance, we consider an example of a randomized block experiment. The example is found in Box, Hunter and Hunter (1978, Sections 7.1 through 7.6). A process for the manufacture of penicillin was being investigated, with yield as the primary interest. Four variants of the basic process were studied, and are denoted as treatments A, B, C, and D.

One important raw material, corn steep liquor, was quite variable. As a result, it was important to block the experiment on this factor. For any blend of the liquor there was a sufficient amount for use in four process runs. Thus all four treatments could be used in any blend. Five blends were used, twenty runs in all. Data are listed in Table 4. The data are stored in a file under the names YIELD, BLEND, and TREATMNT. Each variable in the SCA workspace has 20 elements.

**Table 4 Penicillin production process data**

| *Blend number (of corn steep liquor)* | *Treatment* | | | |
|---|---|---|---|---|
| | *A* | *B* | *C* | *D* |
| 1 | 89 | 88 | 97 | 94 |
| 2 | 84 | 77 | 92 | 79 |
| 3 | 81 | 87 | 87 | 85 |
| 4 | 87 | 92 | 89 | 84 |
| 5 | 79 | 81 | 80 | 88 |

**Exploratory data analysis**

We can use the TABLE and DPLOT paragraphs (see Chapters 4 and 5, respectively) to get an overview of the data and how responses are affected by the factors. From what is known about the process, we expect to see variability in blends.

First we will construct tables to observe the sample mean and standard deviation for the various treatment and blend levels of YIELD.

```
-->TABLE   YIELD, TREATMNT

DEPENDENT   VARIABLE : YIELD
CATEGORICAL VARIABLES: TREATMNT

  ROW      MEAN     STD. DEV.   COUNTS
-----------------------------------------
   1     84.000       4.123      5
   2     85.000       5.958      5
   3     89.000       6.285      5
   4     86.000       5.523      5
```

-->TABLE YIELD, BLEND.

```
DEPENDENT   VARIABLE : YIELD
CATEGORICAL VARIABLES: BLEND

  ROW     MEAN    STD. DEV.  COUNTS
-----------------------------------
   1     92.000      4.243     4
   2     83.000      6.683     4
   3     85.000      2.828     4
   4     88.000      3.367     4
   5     82.000      4.082     4
```

There do not appear to be great differences in the four treatment sample means, although one (89.0) is apart from the rest. This could be accounted for by the overall variability present. The five blend sample means are not as "bunched" together. Clearly, the sample mean of blend 1 (92.0) could be statistically different from the rest.

We can observe the variation in both the treatment and blend sub-groups through the DPLOT paragraph.

-->DPLOT YIELD, TREATMNT

```
  X X  X  X X
+------*--+---------+
77.0    87.0     97.0
MEAN=84.0    SD=4.1

TREATMNT=   A



X   X     XX   X
+-------*-+---------+
77.0    87.0     97.0
MEAN=85.0    SD=6.0

TREATMNT=   B



   X      X X  X    X
+---------+-*-------+
77.0    87.0     97.0
MEAN=89.0    SD=6.3

TREATMNT=   C



  X    XX  X     X
+--------*+---------+
77.0    87.0     97.0
MEAN=86.0    SD=5.5

TREATMNT=   D
```

-->DPLOT YIELD, BLEND

```
           XX    X  X
+---------+----*----+
77.0    87.0     97.0
MEAN=92.0    SD=4.2

 BLEND  =     1.0


X X    X       X
+-----*---+---------+
77.0    87.0     97.0
MEAN=83.0    SD=6.7

 BLEND  =     2.0


          X
    X   X X
+-------*-+---------+
77.0    87.0     97.0
MEAN=85.0    SD=2.8

 BLEND  =     3.0


       X  X X  X
+---------+*--------+
77.0    87.0     97.0
MEAN=88.0    SD=3.4

 BLEND  =     4.0


  XXX      X
+----*----+---------+
77.0    87.0     97.0
MEAN=82.0    SD=4.1

  BLEND  =     5.0
```

The dispersion plots provide more information than the simple summary information of TABLE. The plots broken by treatment level indicate that the high sample mean for treatment C is more the result of a single observation than a noticeable difference in treatments. However, there seems to be a discernible difference between the various blends used.

## Analysis of variance for an additive model

Since blend was used as a blocking variable in this experiment, it was assumed that an interaction between treatment and blends did not exist. As a result, an additive model (8.3) was postulated as an appropriate model for the data. Moreover, since there is only one

observation per cell (i.e., block-treatment combination), an interaction effect cannot be estimated here.

We will employ the TWAY paragraph for the two-way analysis of variance. The TWAY paragraph requires that two or three variables be specified. The first variable is a column of responses. The remaining variable(s) provide the categories or levels at which the treatments (and blocks) are set for a corresponding response. We obtain an analysis of variance for this model (and retain residuals and fitted values for diagnostic checks of the model) by entering

```
-->TWAY   YIELD, BLEND, TREATMNT.   HOLD  RESIDUALS(RES), FITTED(FIT).
```

```
ANALYSIS OF VARIANCE FOR THE VARIABLE :      YIELD
FACTOR(S) IN THE MODEL:    BLEND TREATMNT

S =        4.3397      R**2 =  59.6%      R**2(ADJ) =  36.1%

-----------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
-----------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

BLEND    (A)         264.000     4        66.000       3.504
TREATMNT (B)          70.000     3        23.333       1.239
RESIDUAL             226.000    12        18.833
ADJ. TOTAL           560.000    19


SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =     86.000 )

FACTOR:  BLEND                   77.500    82.500    87.500    92.500    97.500
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    4   92.000    92.000    1.941                     (-------*-------)
  2    4   83.000    83.000    1.941   (-------*-------)
  3    4   85.000    85.000    1.941       (-------*-------)
  4    4   88.000    88.000    1.941             (-------*-------)
  5    4   82.000    82.000    1.941 (-------*-------)
                                     +---------+---------+---------+---------+


FACTOR:  TREATMNT                77.500    82.500    87.500    92.500    97.500
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    5   84.000    84.000    1.681      (------*------)
  2    5   85.000    85.000    1.681        (------*------)
  3    5   89.000    89.000    1.681                (------*------)
  4    5   86.000    86.000    1.681          (------*------)
                                     +---------+---------+---------+---------+
```

## Interpreting the output

The output for this two-way analysis of variance is similar to that of the one-way display. In the one-way display, sums of squares were broken into between-treatment and within-treatment categories. These categories correspond to the "treatments" and "error" components of (8.2).

In the above table the adjusted total sum of squares is broken into three categories: BLEND, TREATMNT, and RESIDUAL. These correspond to the "block", "treatment" and "error" components of (8.6), respectively. Each category provides an estimate of the variance of an observation, as given under the column MEAN SQUARE. The ratio derived from the mean square for "block" or "treatment" divided by the residual mean square provides us with an F-ratio that can be used for hypotheses testing.

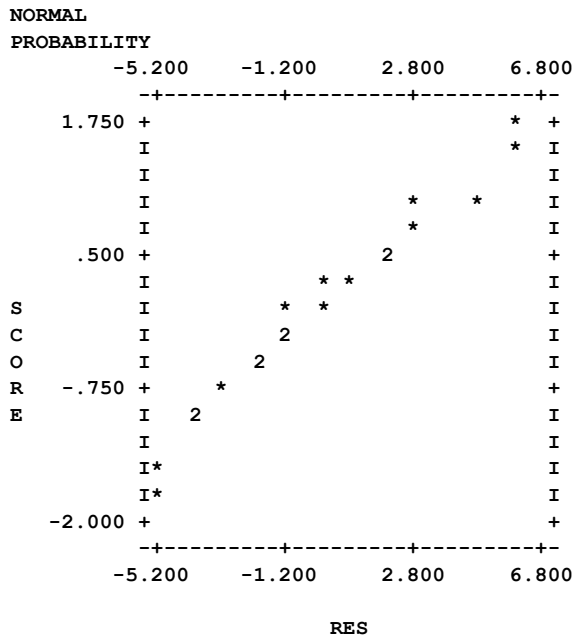
The F-value for BLEND (3.504) is significant at the 5% level, indicating blend is an important factor in the manufacturing process. However, the F-value for TREATMNT (1.239) is not significant at even the 25% level, indicating that the type of treatment is not an important factor in the process.

The summary information for blends and treatments provides us with information similar to that we obtained from TABLE and DPLOT. Note the STD ERR values are of the estimate of sub-group means and the confidence intervals are 95% intervals for these estimates.

## Diagnostic checks of the model

The above conclusions are valid only if the model is valid. Hence, it is important that we diagnostically check the fitted model. If the model is adequate, the residuals are expected to be approximately normally distributed with zero mean and constant variance. One important check of the model is a normal probability plot of the residuals. If the model is adequate, then the residuals should fall roughly on a straight line in this plot. We can create a probability plot using the PPLOT paragraph (see Chapter 5).

-->PPLOT RES



We observe no model inadequacy from this plot. We can also use the DPLOT paragraph (see Chapter 5) to observe if there is any abnormal distribution within the sub-groups of either blend or treatment. Again, no inadequacies are observed as the means for each block and treatment are about the same with the same relative variability in each group.

-->DPLOT RES

```
          X
XX   X X X   X X    X
XX X X X X X X X   XX
+--------*+---------+
-5.0     .5       6.0
MEAN=.0      SD=3.4
```

-->DPLOT RES, BLEND

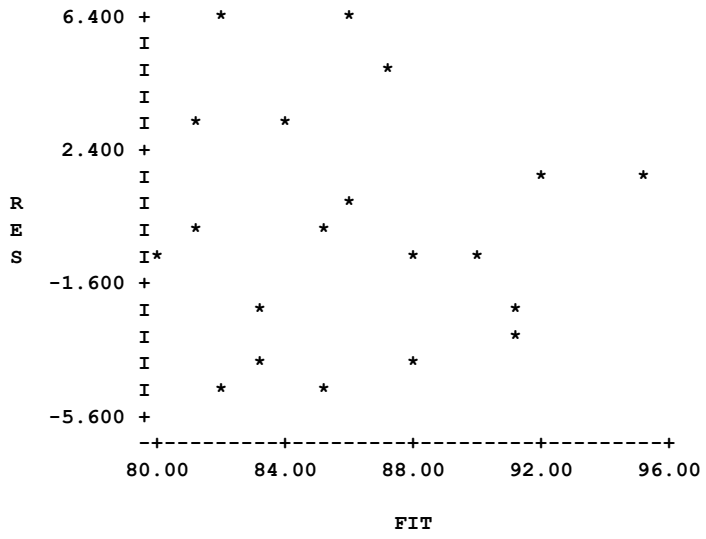
```
            X
   X   X    X
+--------*+---------+
-5.0     .5       6.0
MEAN=.0      SD=2.4

 BLEND  =     1.0


XX            X    X
+--------*+---------+
-5.0     .5       6.0
MEAN=.0      SD=5.4

 BLEND  =     2.0


     X X X    X
+--------*+---------+
-5.0     .5       6.0
MEAN=.0      SD=2.2

 BLEND  =     3.0


 X   X     X       X
+--------*+---------+
-5.0     .5       6.0
MEAN=.0      SD=3.9

 BLEND  =     4.0


X      X X          X
+--------*+---------+
-5.0     .5       6.0
MEAN=.0      SD=4.5

 BLEND  =     5.0
```

-->DPLOT RES, TREATMNT

```
       X
     X X   X   X
+--------*+---------+
-5.0     .5       6.0
MEAN=.0      SD=2.0

TREATMNT=   A


X  X     X     X   X
+--------*+---------+
-5.0     .5       6.0
MEAN=.0      SD=4.1

TREATMNT=   B


X    X X     X      X
+--------*+---------+
-5.0     .5       6.0
MEAN=.0      SD=4.2

TREATMNT=   C


 X
 X       X   X      X
+--------*+---------+
-5.0     .5       6.0
MEAN=.0      SD=4.2

TREATMNT=   D
```

Another useful diagnostic check is a plot of residuals against fitted values. This plot does not reveal any model abnormalities.

-->PLOT RES, FIT




## A formal test for nonadditivity

We have employed an additive model in our analysis. An additive model was employed since no interaction was believed to exist between TREATMNT and the blocking variable BLEND, nor could an interaction estimate be computed. Tukey (1949) provides a formal test for the assumption of an additive model. Although this test is not provided directly by the SCA System, we can calculate the test statistic in a straight forward manner using the TWAY paragraph and selected analytic statements (see Appendix A). To compute the statistic, we must conduct an analysis of variance on the square of the fitted values from the above model. We can create the squared values by entering

-->FITSQRD = FIT*FIT

In our second use of TWAY, we employ FITSQRD in place of our original response variable (YIELD). The purpose of this analysis of variance is to obtain a sum of squares and the residuals from the fit. Hence we can ignore most of the information provided (and much of the output from the paragraph is suppressed below)

```
-->TWAY  FITSQRD, BLEND, TREATMNT.  HOLD  RESIDUALS(RESF).

 ANALYSIS OF VARIANCE FOR THE VARIABLE :    FITSQRD
 FACTOR(S) IN THE MODEL:    BLEND TREATMNT
 S =       17.5499      R**2 = 100.0%      R**2(ADJ) =  99.9%
 ------------------------------------------------------------
 ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
 ------------------------------------------------------------
 SOURCE        SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO
 BLEND    (A)    7994923.000     4   1998730.750    6489.386
 TREATMNT (B)    2102085.000     3    700695.000    2274.984
 RESIDUAL           3696.000    12       308.000
 ADJ. TOTAL     10100704.200    19
```

The F-statistic to compute is $\dfrac{S_{na}}{(S_R - S_{na})/df_D}$

where

$S_R$ = the residual sum of squares of our additive ANOVA table (We see in the ANOVA table on page 8.17 the value is 226.0.)

$S_{na}$ = $P^2/Q$ with P the sum of the cross-products of the two sets of residuals (here RES and RESF) and Q the residual sum of squares of the second ANOVA (here 3696.0), and

$df_D$ = 1 less than the df of the residual sum of squares of the additive ANOVA table (in this example, 12 - 1 = 11)

We can compute the statistic using the following sequence of analytic statements

```
-->P = SUM (RES*RESF)
-->SNA = P*P/3696.0
-->FSTAT = SNA/((226.0-SNA)/11)
```

We can compute the significance level for the statistic using the inverse function of the F distribution (see Appendix A).

```
-->FLEVEL = 1 - IDFF(FSTAT,1,11)
```

Printing FSTAT and FLEVEL reveals the value of the F-statistic (.098) is insignificant.

**Use of reference distributions for interpretations**

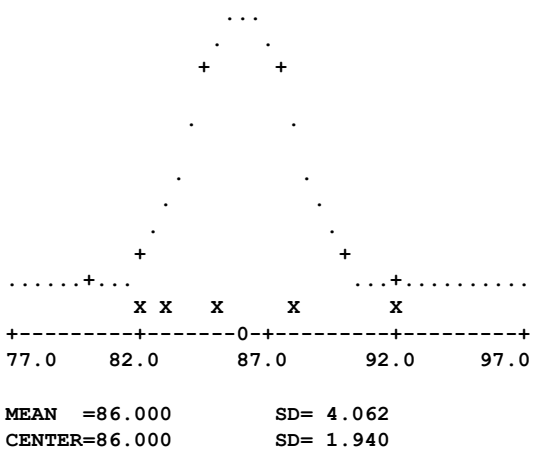
Box, Hunter and Hunter (1978, Section 7.6) discuss the use of reference distributions for the comparison of individual means, and provide a warning for the reliance on the F-test alone. The reference distribution for treatment and blend group averages is the t-distribution with $df_{res}$ degrees of freedom, where $df_{res}$ are the degrees of freedom for the RESIDUAL in our ANOVA table (in this case, 12). The standard deviation for a group is $s/\sqrt{n}$, where s is the standard error displayed in the TWAY output (here 4.3397) and n is the number of treatments (or blends). Hence the standard error for the treatment means is 2.17 ($4.3397/\sqrt{4}$) and the standard error for the blend means is 1.94 ($4.3397/\sqrt{5}$).

We can display these means, together with their reference distribution, directly in the SCA System using the TABLE and DTPLOT paragraphs (see Chapters 4 and 5, respectively). The TABLE paragraph is used to compute and hold the sample means. We can display the means of each factor, and retain the means, by entering

```
-->TABLE   YIELD, BLEND.  STORE  MBLEND.
-->TABLE   YIELD, TREATMNT.  STORE  MTREAT.
```

The STORE sentence is used to retain the sample means in the variable that is specified. Output from these paragraphs is suppressed as it is the same as that shown earlier. We can now display the blend means and their reference distribution by entering

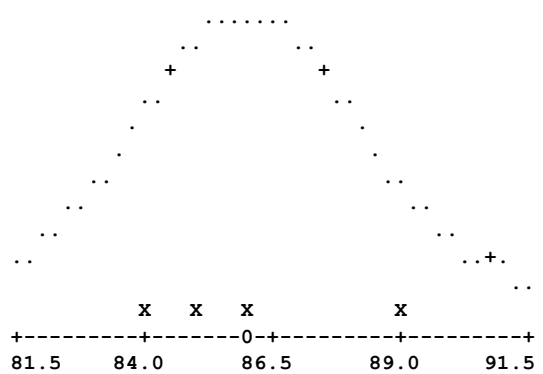
```
-->DTPLOT   MBLEND.  STDEV IS 1.94.  DF ARE 12.
```




We see that the largest group mean value (92.0) appears under the third '+' symbol from the center; that is, at the boundary of a 99% confidence interval. This is a clear indication that the value is different from the rest.

In like manner, we can display the treatment means with their reference distribution by entering

```
-->DTPLOT  MTREAT.   STDEV IS 2.17.    DF ARE 12.
```



```
          X   X   X           X
+---------+-------0-+---------+---------+
81.5    84.0      86.5      89.0     91.5

MEAN  =86.000        SD= 2.160
CENTER=86.000        SD= 2.170
```

All treatment means are well within a 95% confidence interval of the reference distribution. We have good visual evidence that there is no significant difference in yields based on treatments.

## 8.2.2 Example: Apple storage data

As a second example of a two-way analysis of variance, we consider a study of the length of time (in months) to spoilage for various types of apples under different storage conditions (see Barker, 1985, page 170). The data are shown in Table 5. The types of apples studied were Ida Red, McIntosh, and Delicious (denoted by I, M and D, respectively). Storage conditions indicate the temperature level at which the apples were kept. Each apple and temperature combination was replicated.

**Table 5 Apple data from Barker**

| *Months to spoilage* | | *Type of* | *Storage* |
|---|---|---|---|
| *Replication 1* | *Replication 2* | *Apple* | *Temperature* |
| *REP1* | *REP2* | *TYPE* | *TEMP* |
| 6.5 | 6.0 | I | 36 |
| 7.5 | 8.0 | I | 38 |
| 8.0 | 8.5 | I | 40 |
| 7.5 | 7.0 | I | 42 |
| 5.0 | 4.5 | I | 44 |
| 7.5 | 8.0 | M | 36 |
| 8.5 | 9.0 | M | 38 |
| 9.5 | 9.0 | M | 40 |
| 9.5 | 9.0 | M | 42 |
| 7.5 | 7.0 | M | 44 |
| 5.0 | 4.5 | D | 36 |
| 6.0 | 6.5 | D | 38 |
| 7.0 | 7.5 | D | 40 |
| 6.0 | 5.5 | D | 42 |
| 5.0 | 4.5 | D | 44 |

The data are stored in the SCA workspace under the labels REP1, REP2, TYPE, and TEMP. Although it is appropriate to first plot the average time to spoilage for the various combinations of type and temperature, we will postpone this until after the analysis of variance.

We will again employ the TWAY paragraph. We noted in the previous example, the TWAY paragraph requires columns of data, one for the responses and the rest for the treatment and blocks. As a result, we will first create three columns (variables) for the apple data. One variable consists of the months to spoilage (we will label this as SPOILAGE). Another variable consists of apple types (APPLE), and the third of storage temperatures (STORAGE). The JOIN paragraph is used for this purpose (see Appendix B).

```
-->JOIN      REP1, REP2.   NEW IS SPOILAGE.
-->JOIN      TYPE, TYPE.   NEW IS APPLE.
-->JOIN      TEMP, TEMP.   NEW IS STORAGE.
```

To compute and display the two-way analysis of variance for this data we may enter

```
-->TWAY   SPOILAGE, APPLE, STORAGE

ANALYSIS OF VARIANCE FOR THE VARIABLE :     SPOILAGE
FACTOR(S) IN THE MODEL:    APPLE  STORAGE

S =         .3536      R**2 =  97.3%      R**2(ADJ) =  94.8%

-----------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
-----------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

APPLE    (A)         36.867     2        18.433     147.467
STORAGE  (B)         27.867     4         6.967      55.733
AB                    3.133     8          .392       3.133
RESIDUAL 24           1.875    15          .125
ADJ. TOTAL           69.742    29


SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =      7.017 )

FACTOR:  APPLE                    5.100     6.100     7.100     8.100     9.100
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   10    6.850     6.850     .091               (--)
  2   10    8.450     8.450     .091                                  (--)
  3   10    5.750     5.750     .091     (--)
                                     +---------+---------+---------+---------+


FACTOR:  STORAGE                  5.100     6.100     7.100     8.100     9.100
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    6    6.250     6.250     .129          (--*-)
  2    6    7.583     7.583     .129                       (--*--)
  3    6    8.250     8.250     .129                              (--*-)
  4    6    7.417     7.417     .129                     (--*--)
  5    6    5.583     5.583     .129  (--*--)
                                     +---------+---------+---------+---------+
```

The above analysis of variance table has four categories: APPLE, STORAGE, AB and RESIDUAL. The AB category corresponds to the interaction between APPLE and STORAGE. An interaction category is automatically provided by the TWAY paragraph whenever one can be determined, unless we specify otherwise. A more complete discussion follows later in this section.

The F-ratios for the **main effects** of APPLE and STORAGE are 147.5 and 55.7, respectively. These values are highly significant. The F-ratio for interaction is also significant at the 5% level. The summary information shows that apple 2 (McIntosh) has a significantly higher storage life than the other two, while storage condition 3 (40o) produces the highest average storage life.

As another visual display of the results, we will plot the average time to spoilage of each variety against temperature using the PLOT paragraph (see Chapter 3). We will first calculate the average time to spoilage using an analytic statement (see Appendix A). Next we generate "tags" for each variety of apple (see Appendix B); then we PLOT the averages using the "tags".

```
-->AVERAGE = (REP1 + REP2)/2

-->GENERATE  TAGS.   NROWS ARE 15.                @
-->          VALUES ARE 18 FOR 5, 22 FOR 5, 13 FOR 5.

   THE SINGLE PRECISION VARIABLE  TAGS   IS GENERATED


-->PLOT  AVERAGE, TEMP.   TAGSET IS TAGS.

      9.000 +                    M         M
            I          M
            I                     I
            I
            IM         I
      7.500 +
            I                     D         I         M
  A         I
  V         I
  E         II         D
  R   6.000 +
  A         I                               D
  G         I
  E         I
            ID                                        2
      4.500 +
            -+---------+---------+---------+---------+
           36.00     38.00     40.00     42.00     44.00

                              TEMP
```

We see that McIntosh has the longest time to spoilage for each temperature. Ida Red has a higher time to spoilage than Delicious except for the highest temperature (this somewhat explains our interaction effect). The best temperature at which to store all apples is 40o. Barker (1985) notes that once Ida Red and Delicious are removed from storage, the temperature can be set higher, since the McIntosh survives as well at 42o as it does at 40o.

## Analysis without an interaction term

We can control whether or not to include an interaction term in our two-way analysis of variance. The default of the SCA System is to calculate and display the interaction effect if it can be calculated. This is possible whenever there are at least two observations for each "block" and "treatment" combination. We noted that this was not possible in the previous example. In those cases where an interaction term can be calculated, but we only wish to consider an additive model, we need to inform the TWAY paragraph that we do not want to include an interaction term. In such cases, the RESIDUAL sum of squares will be the total of the sum of squares for error and interaction. To see this, we will re-compute the analysis of variance table without an interaction term.

```
-->TWAY   SPOILAGE, APPLE, STORAGE.   NO INTERACTION.

ANALYSIS OF VARIANCE FOR THE VARIABLE :     SPOILAGE
FACTOR(S) IN THE MODEL:    APPLE  STORAGE

S =       26 .4666      R**2 =  92.8%      R**2(ADJ) =  90.9%

-----------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
-----------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

APPLE    (A)         36.867     2        18.433      84.652
STORAGE  (B)         27.867     4         6.967      31.993
RESIDUAL              5.008    23          .218
ADJ. TOTAL           69.742    29


SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =      7.017 )

FACTOR:  APPLE                    5.100     6.100     7.100     8.100     9.100
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   10    6.850     6.850     .120               (-*--)
  2   10    8.450     8.450     .120                                  (-*--)
  3   10    5.750     5.750     .120    (--*-)
                                     +---------+---------+---------+---------+

FACTOR:  STORAGE                  5.100     6.100     7.100     8.100     9.100
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    6    6.250     6.250     .170        (---*--)
  2    6    7.583     7.583     .170                     (---*--)
  3    6    8.250     8.250     .170                            (---*--)
  4    6    7.417     7.417     .170                    (--*---)
  5    6    5.583     5.583     .170 (---*--)
                                     +---------+---------+---------+---------+
```

If we compare the above results with the one displayed previously, we see the RESIDUAL sum of squares, here (5.008), is the sum of squares for AB and RESIDUAL, previously (3.133 and 1.875). The degrees of freedom is also the sum of the previous sources. A new mean square for RESIDUAL is obtained. This then changes the F-ratio's for APPLE and STORAGE.

## 8.3 Example: Toxic agents data

As a third example of two-way analysis of variance, we consider data of a two-way factorial experiment used by Box and Cox (1964), and Box, Hunter and Hunter (1978, page 228). The survival times (in tens of hours) of groups of animals subjected to three poisons and four treatments are used. The data are given in Table 6. The data are stored in the SCA workspace in three variables, SURVIVAL, POISON, and TREATMNT. The variable SURVIVAL consists of the 48 survival times presented in Table 6. The poison and treatment associated with the survival time are maintained in POISON and TREATMNT, respectively.

**Table 6 Toxic agents data**

Survival times (in tens of hours)

| *Poison* | *Treatment* *A* | *B* | *C* | *D* |
|---|---|---|---|---|
| 1 | 0.31 | 0.82 | 0.43 | 0.45 |
| | 0.45 | 1.10 | 0.45 | 0.71 |
| | 0.46 | 0.88 | 0.63 | 0.66 |
| | 0.43 | 0.72 | 0.76 | 0.62 |
| 2 | 0.36 | 0.92 | 0.44 | 0.56 |
| | 0.29 | 0.61 | 0.35 | 1.02 |
| | 0.40 | 0.49 | 0.31 | 0.71 |
| | 0.23 | 1.24 | 0.40 | 0.38 |
| 3 | 0.22 | 0.30 | 0.23 | 0.30 |
| | 0.21 | 0.37 | 0.25 | 0.36 |
| | 0.18 | 0.38 | 0.24 | 0.31 |
| | 0.23 | 0.29 | 0.22 | 0.33 |

A two-way analysis of variance yields the following:

```
-->TWAY   SURVIVAL, TREATMNT, POISON.     @
        HOLD  RESIDUALS (RESID), FITTED (FITTED).

ANALYSIS OF VARIANCE FOR THE VARIABLE :   SURVIVAL
FACTOR(S) IN THE MODEL: TREATMNT   POISON

S =         .1491      R**2 =  73.4%      R**2(ADJ) =  65.2%

-----------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
-----------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

TREATMNT (A)           .921     3          .307      13.806
POISON   (B)          1.033     2          .517      23.222
AB                     .250     6          .042       1.874
RESIDUAL               .801    36          .022
ADJ. TOTAL            3.005    47
```

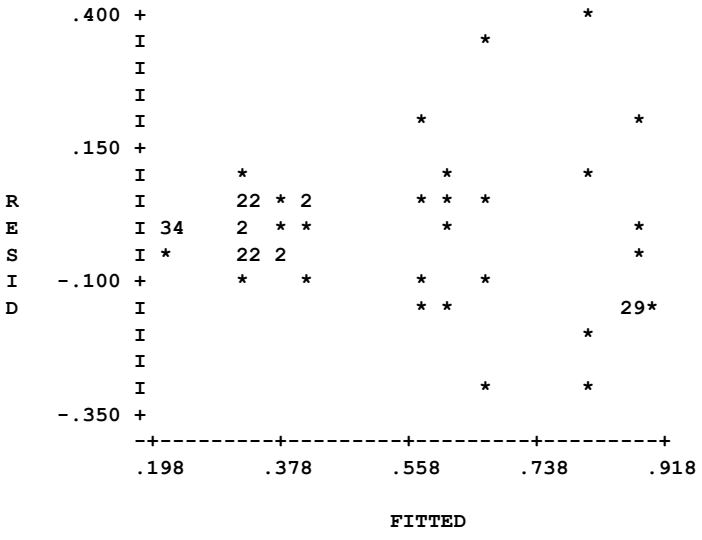
```
SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =       .479 )

FACTOR:  TREATMNT                  .195      .345      .495      .645      .795
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   12     .314      .314     .037   (----*----)
  2   12     .677      .677     .037                          (----*----)
  3   12     .392      .392     .037        (----*----)
  4   12     .534      .534     .037                  (----*---)
                                     +---------+---------+---------+---------+

FACTOR:  POISON                    .195      .345      .495      .645      .795
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   16     .617      .618     .030                        (---*---)
  2   16     .544      .544     .030                   (---*---)
  3   16     .276      .276     .030 (---*---)
                                     +---------+---------+---------+---------+
```

The F-ratios associated with both main effects (treatment and poison) are significant at the 1% level. This indicates that effects of poisons and treatments differ within themselves. The summary information indicates that the second treatment (B) is the most effective overall and poison 3 is the quickest acting poison.

### 8.3.1 Reducing the complexity of an analysis

The above analysis is useful <u>provided</u> the model employed in the analysis is adequate for the data. The residuals and fitted values from the model have been retained in the SCA workspace under the labels RESID and FITTED, respectively. As one diagnostic check, residuals are plotted against the fitted values.

-->PLOT RESID, FITTED




We observe a funnel-like appearance in the plot. This suggests the standard deviation increases with mean level. We need to stabilize the variance before we can properly analyze the data.

An alternative method to access whether a transformation may be in order is to plot the sample standard deviation of each cell of the data (i.e., each poison/treatment combination) against the sample mean of the cell. We can do this conveniently using the TABLE and PLOT paragraphs. We can compute and retain sample means and standard deviations of cells using the STORE sentence of the TABLE paragraph (see Section 2 of Chapter 4). In Chapter 4 we retained the means and standard deviations for this data set in SM and SSTD, respectively. If we now plot SSTD against SM, we obtain the following

```
-->PLOT  SSTD, SM

      .375 +
           I
           I                                  *
           I
           I                          *
      .250 +
           I
 S         I
 S         I
 T         I                    *                 *
 D    .125 +                       *
           I
           I       *    *
           I        * *
           I **    *
      .000 +
           -+---------+---------+---------+---------+
           .198      .378      .558      .738      .918

                              SM
```

We observe that the value of the standard deviation is proportional to the value of the sample mean. This is an indication that the survival data should be transformed before an analysis (usually with a logarithmic transformation).

## 8.3.2 Transformations and a transformation analysis of the toxic data

Typical assumptions made in linear models (of which the models used for analysis of variance are a special case) include the homogeneity of variance and normality of the random errors associated with the model. Box and Cox (1964) showed how a nonlinear transformation of data can improve matters in those cases where one or more of these assumptions are not satisfied. A procedure for estimating the proper transformation was also discussed. This procedure has been widely used with great success and has been incorporated within the TWAY (and NWAY) paragraph.

Box and Cox worked with a family of power transformations expressed as

$$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \ell n(y) & \text{if } \lambda = 0 \end{cases}$$

In this parameterization, the family of transformations is continuous in the transformation parameter λ. This family of data transformations has come to be known in the literature as the "Box-Cox transformation".

Box and Cox showed how λ can be estimated jointly with the other model parameters by maximum likelihood. The procedure is based on an analysis of the following scaled response

$$z^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/(\lambda \dot{y}^{\lambda-1}) & \text{if } \lambda \neq 0 \\ \dot{y}\ \ell n(y) & \text{if } \lambda = 0 \end{cases}$$

where $\dot{y}$ denotes the geometric mean of the observations. A sequence of standard analyses are conducted for these scaled responses by varying the choice of λ.

A grid of λ is chosen to start. The maximum likelihood estimate of λ is the value of λ which yields the smallest mean square error, MSE(λ), in all analyses conducted. In practice, the exact value of λ may not be used. Instead a value close to it may be chosen on the basis of a better physical interpretation. We will now incorporate a transformation analysis in the toxic data example.

From previous experience with the data, the grid chosen for the transformation parameter is between -3.0 and 1.0 at 0.1 spacings. We will retain residuals and fitted values as before. In addition, the values of λ, the associated MSE, and associated F-values for the treatment, poison and interaction effects will be retained in the variables labeled LAMBDA, MSERROR, FT, FP and FTP, respectively.

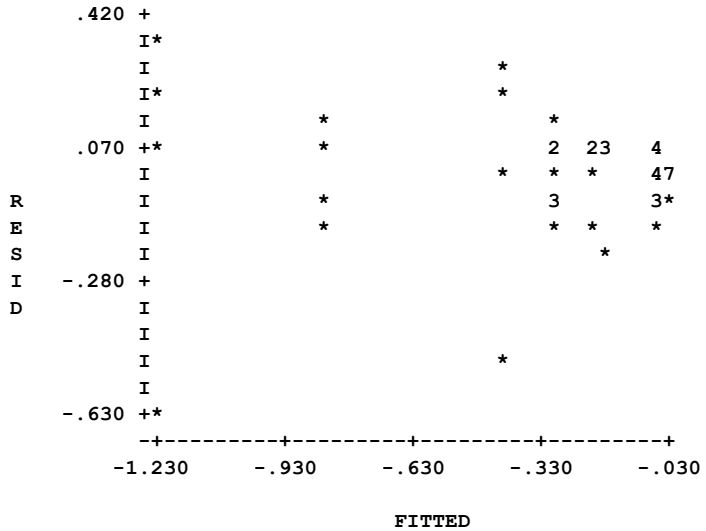
```
-->TWAY   SURVIVAL, TREATMNT, POISON.   POWERS ARE 0.1, -3.0, 1.0 .        @
      HOLD  RESIDUALS (RESID), FITTED (FITTED), POWERS (LAMBDA),          @
      MSE (MSERROR), FVALUES (FT, FP, FTP).
```

```
ANALYSIS OF VARIANCE FOR THE VARIABLE :   SURVIVAL
FACTOR(S) IN THE MODEL: TREATMNT   POISON

POWER VALUE OF TRANSFORMATION IS   -.800

S =         .0880      R**2 =  86.3%      R**2(ADJ) =  82.2%

-----------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
-----------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

TREATMNT (A)           .641     3          .214      27.624
POISON   (B)          1.074     2          .537      69.430
AB                     .044     6          .007        .949
RESIDUAL               .278    36          .008
ADJ. TOTAL            2.038    47
```

```
SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =      -.302 )

FACTOR:  TREATMNT                -.564     -.444     -.324     -.204     -.084
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   12    -.459     -.459     .022     (---*--)
  2   12    -.164     -.164     .022                          (--*---)
  3   12    -.361     -.361     .022              (---*--)
  4   12    -.223     -.223     .022                        (--*---)
                                     +---------+---------+---------+---------+


FACTOR:  POISON                  -.564     -.444     -.324     -.204     -.084
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   16    -.157     -.157     .018                              (--*--)
  2   16    -.241     -.241     .018                        (--*--)
  3   16    -.508     -.508     .018  (--*--)
                                     +---------+---------+---------+---------+
```

The values in the analysis of variance table have changed significantly, but the interpretations are the same as before. The λ value for the power transformation is -.8, since this value produces the smallest MSE. To assess the effect of this transformation, we again plot the residuals from the fit against the fitted values.

-->PLOT RESID, FITTED




The funnel-like behavior of before has been eliminated, indicating a more adequate fit of the data for λ = -.8. In order to examine the effect of different transformations, we can now plot MSE(λ) against λ, and the F-ratios for main effects and interactions obtained for each λ against λ.

```
-->PLOT   MSERROR, LAMBDA

       .030 +*
            I *
            I
            I  *
            I
  M    .023 +   *                                  *
  S         I    *
  E         I                                     *
  R         I     *                              *
  R         I      *                            *
  O    .015 +       **                          *
  R         I         *                       **
            I          **                     *
            I            **                ***
            I              ****          ***
       .008 +                  *********
            -+---------+---------+---------+---------+
          -3.000    -2.000    -1.000     -.000     1.000

                              LAMBDA

-->MPLOT   Y-VARIABLES  ARE  FT, FP, FTP.                    @
           X-VARIABLES  ARE  LAMBDA, LAMBDA, LAMBDA.        @
           SYMBOLS  ARE 'T', 'P', 'I'.

      75.00 +            PPPPPPPPP
            I         PPP         PP
            I      PPP              PPP
            I   PPP                    P
            I PP                        PP
      50.00 +P                            PP
            I                               PP
  F         I                                 P
  T         I                                  PP
  P         I             TTTTTTTTTT             PP
      25.00 +    TTTTTTTT           TTTTTTT        PPP
            ITTTT                             TTTTTT
            I                                       TTTTT
            I
            IIIIIIIIIIIII
        .00 +             IIIIIIIIIIIIIIIIIIIIIIIIIIII
            -+---------+---------+---------+---------+
          -3.000    -2.000    -1.000     -.000     1.000

                              LAMBDA
```

The plot of MSE(λ) against λ shows that MSE(λ) is relatively flat in the interval (-1.2, -0.5). Any λ value in this range will produce results similar to that of -0.80 . In particular, we could use λ = -1.00, corresponding to the reciprocal transformation. This has the advantage of both scientific plausibility and ease of interpretation.

F-values were displayed using the MPLOT paragraph (see Chapter 3). The F-ratios corresponding to poison, treatment and interaction are displayed with the symbols P, T and I, respectively. This plot clearly shows that the reciprocal transformation produces a better separation of main effects than that of λ = -0.80, implying a more efficient analysis. In

addition, interaction effects are reduced to insignificance if a transformation with either $\lambda = -.8$ or $\lambda = -1.0$ is used. Hence we could now re-analyze the reciprocal of survival times using a simpler model.

## 8.4 The Analysis of Covariance

The analysis of covariance is a statistical technique that combines the analysis of variance and regression. The analysis of covariance arises as follows. When an experiment is conducted, the observations of one or more variables are recorded in addition to the response variable under study. The values of these other variables, called **concomitant variables** or **covariates**, are not controlled by the experimenter; they are only observed and recorded. If the covariates are related linearly to the response variable, then an adjustment can be made to the response variable before the analysis of variance is conducted. Without such an adjustment, true differences between treatments (factors) may be hard to detect.

As a result, the model fitted in the analysis of covariance is a hybrid of a linear regression of the response variable on its covariates and the analysis of variance models. A more complete discussion of the analysis of covariance can be found in Chapter 16 of Montgomery (1984) or Chapter 18 of Snedecor and Cochran (1989). In this section we will illustrate its use in a one-way analysis of variance and a two-way analysis of variance.

### 8.4.1 One-way ANOVA with a single covariate: A breaking strength example

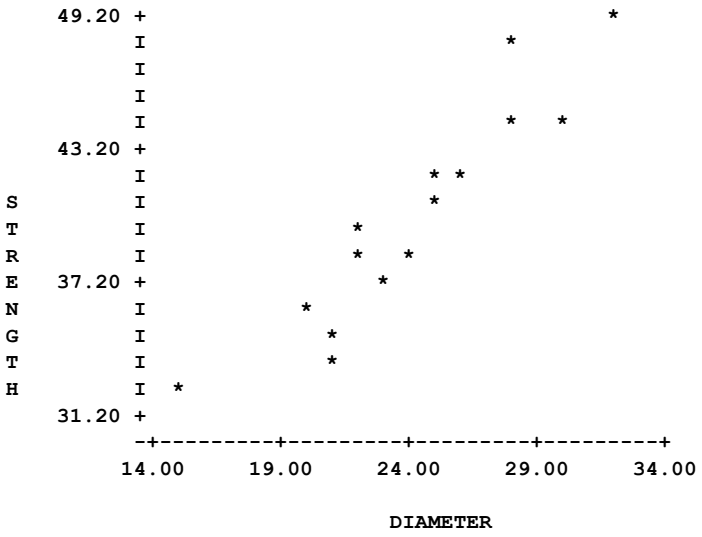
To illustrate the analysis of covariance for a single covariate and a one-way analysis of variance, we consider an example from Section 16.2 of Montgomery (1984). In this experiment, the breaking strength (in pounds) of a fiber produced by three different machines is recorded. An analysis of variance is used to determine if there are any differences in the breaking strengths that is due to the machines. In addition to the breaking strength, the diameter (in 10-3 inches) of each fiber is also recorded since the strength of a fiber is related to its diameter. The data are shown in Table 7 and are stored in the SCA workspace under the labels STRENGTH, DIAMETER and MACHINE.

**Table 7 Breaking strength data**

| *Machine 1* | | *Machine 2* | | *Machine 3* | |
|---|---|---|---|---|---|
| *Breaking strength* | *Diameter of fiber* | *Breaking strength* | *Diameter of fiber* | *Breaking strength* | *Diameter of fiber* |
| 36 | 20 | 40 | 22 | 35 | 21 |
| 41 | 25 | 48 | 28 | 37 | 23 |
| 39 | 24 | 39 | 22 | 42 | 26 |
| 42 | 25 | 45 | 30 | 34 | 21 |
| 49 | 32 | 44 | 28 | 32 | 15 |

We can observe that a linear relationship exists between breaking strength and diameter through a simple scatter plot (see Chapter 3) of BREAKING against DIAMETER

-->PLOT STRENGTH, DIAMETER




It would be reasonable to include DIAMETER as a covariate in an analysis of variance. However, to better illustrate the need for an analysis of covariance in this example, we will first perform a simple one-way ANOVA without a covariate.

We can perform a one-way analysis of variance by using either the OWAY or TWAY paragraph. However, the OWAY paragraph cannot be used for an analysis of covariance. Hence the TWAY paragraph is used. We can obtain a one-way ANOVA by entering

-->TWAY STRENGTH, MACHINE. HOLD RESIDUALS(RES),FITTED(FIT).

Residuals and fitted values are maintained in the variables RES and FIT, respectively for diagnostic checking purposes. We obtain the following

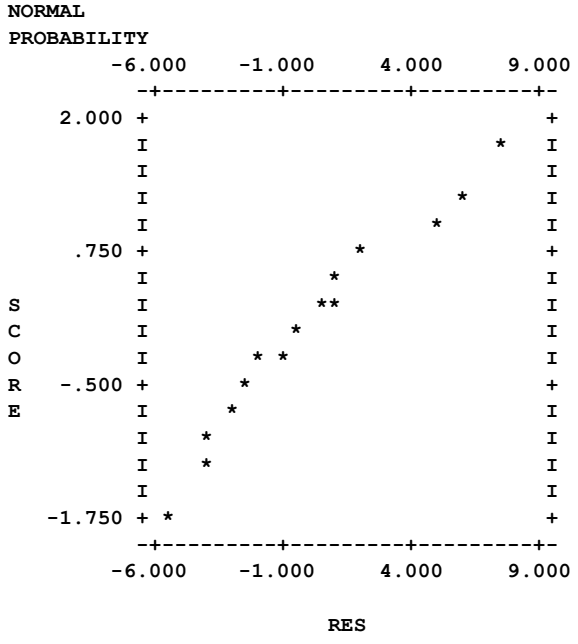
```
ANALYSIS OF VARIANCE FOR THE VARIABLE :   STRENGTH
FACTOR(S) IN THE MODEL:  MACHINE

S =        4.1433      R**2 =  40.5%      R**2(ADJ) =  30.6%

------------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
------------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

MACHINE  (A)        140.400     2        70.200       4.089
RESIDUAL            206.000    12        17.167
ADJ. TOTAL          346.400    14
```

```
SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =     40.200 )

FACTOR:  MACHINE                32.550    36.050    39.550    43.050    46.550
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    5   41.400    41.400    1.513               (--------*---------)
  2    5   43.200    43.200    1.513                    (--------*---------)
  3    5   36.000    36.000    1.513 (--------*--------)
                                     +---------+---------+---------+---------+
```

The F-statistic corresponding to MACHINE is significant at the 5% level. The above analysis appears to indicate that there are significant differences in breaking strength due to machines.

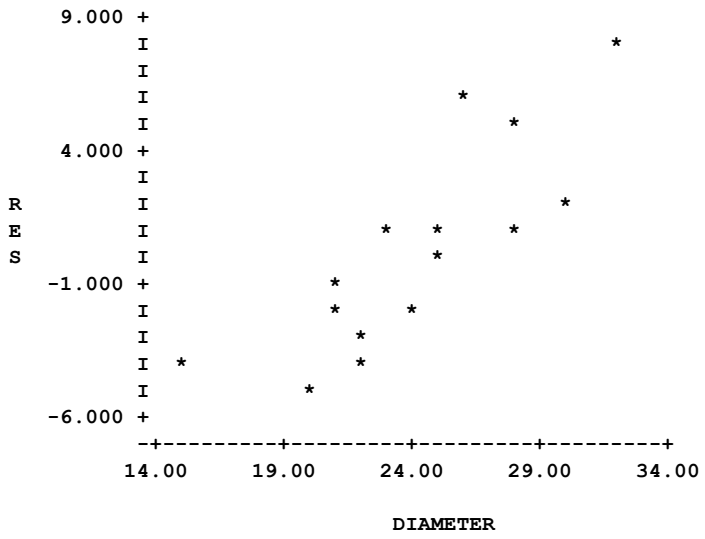
Diagnostic checking is necessary to validate the adequacy of the above model. Plots to consider for this example are a probability plot of the residuals and scatter plots of residuals against fitted values, the input variable (MACHINE), and DIAMETER. The probability plot of residuals does not indicate any gross inadequacies in the model.

-->PPLOT RES




Similarly, plots of residuals against fitted values and against the input variable (MACHINE) do not reveal any gross inadequacies. However, the plot of residuals against diameter shows a strong linear relationship. It is clear that DIAMETER needs to be incorporated into the model.

-->PLOT RES, DIAMETER




We will now include DIAMETER as a covariate in the analysis. The TWAY paragraph will be used again, and the variable DIAMETER is specified after a slash ('/'), indicating it is a covariate.

-->TWAY STRENGTH,MACHINE /DIAMETER. HOLD RESIDUALS(RES),FITTED(FIT)

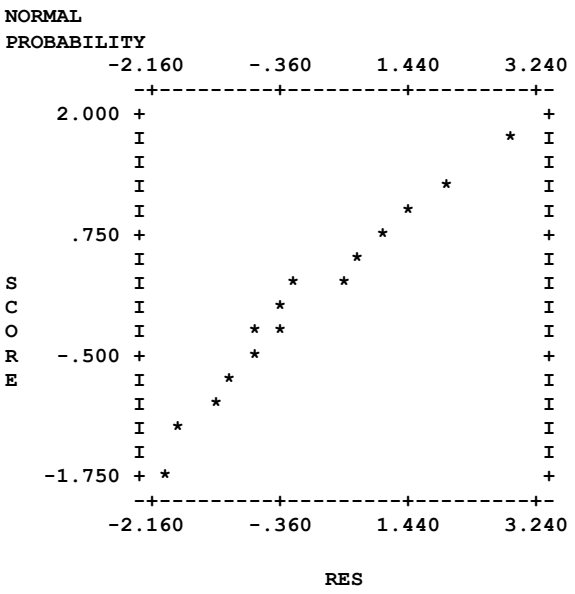
As before, we will maintain the residuals and fitted values from the model in the variables RES and FIT, respectively for use in diagnostic checks of the model. We obtain the following

```
ANALYSIS OF VARIANCE FOR THE VARIABLE :   STRENGTH
FACTOR(S) IN THE MODEL: DIAMETER  MACHINE

S =        1.5950      R**2 =  91.9%      R**2(ADJ) =  89.7%

-----------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
-----------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

DIAMETER            305.130     1       305.130     119.933
MACHINE  (A)         13.284     2         6.642       2.611
RESIDUAL             27.986    11         2.544
ADJ. TOTAL          346.400    14


SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =     40.200 )

FACTOR:  MACHINE                37.200    38.700    40.200    41.700    43.200
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    5   41.400    40.382     .595            (--------*--------)
  2    5   43.200    41.419     .620                     (--------*--------)
  3    5   36.000    38.798     .672 (---------*---------)
                                    +---------+---------+---------+---------+
```

The sum of squares for MACHINES and error (i.e., RESIDUAL) are now adjusted for the presence of the covariate DIAMETER. Now the F-statistic corresponding to MACHINE is not significant at the 5% (nor 10%) level, and the hypothesis that there are no differences in breaking strength due to machines cannot be rejected.

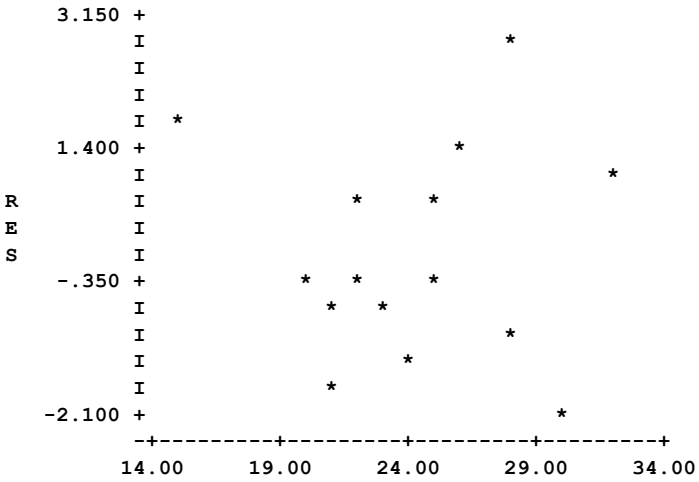
As before, the probability plot of the residuals does not reveal any anomalies with the model.

-->PPLOT RES




Similarly, the scatter plots of residuals against fitted values and residuals against the input variable reveal no gross model inadequacies. Now the plot of residuals against diameter supports the model, since we have included diameter within the model.

-->PLOT RES, DIAMETER



### 8.4.2 Two-way ANOVA with a single covariate: A corn yield example

To illustrate the use of two-way ANOVA with a single covariate, we consider a corn yield example from Section 18.4 of Snedecor and Cochran (1989). Here an experiment was conducted whose response was the yields of six varieties of corn. The experiment was blocked into different plots. Another potential source of variation from plot to plot was the number of plants (called stands) in a plot. Stands could be a covariate.

The data from this experiment are listed in Table 8 and data are stored in the SCA workspace under the labels YIELD, VARIETY, BLOCK and STAND.

**Table 8 Data of the corn yield experiment**

| | *Block 1* | | *Block 2* | | *Block 3* | | *Block 4* | |
|---|---|---|---|---|---|---|---|---|
| *Variety* | *Stand* | *Yield* | *Stand* | *Yield* | *Stand* | *Yield* | *Stand* | *Yield* |
| A | 28 | 202 | 22 | 165 | 27 | 191 | 19 | 134 |
| B | 23 | 145 | 26 | 201 | 28 | 203 | 24 | 180 |
| C | 27 | 188 | 24 | 185 | 27 | 185 | 28 | 220 |
| D | 24 | 201 | 28 | 231 | 30 | 238 | 30 | 261 |
| E | 30 | 202 | 26 | 178 | 26 | 198 | 29 | 226 |
| F | 30 | 228 | 25 | 221 | 27 | 207 | 24 | 204 |

We can conduct a two-way ANOVA of YIELD on VARIETY and BLOCK by entering

-->TWAY YIELD, VARIETY, BLOCK. HOLD RESIDUALS(RES), FITTED(FIT).

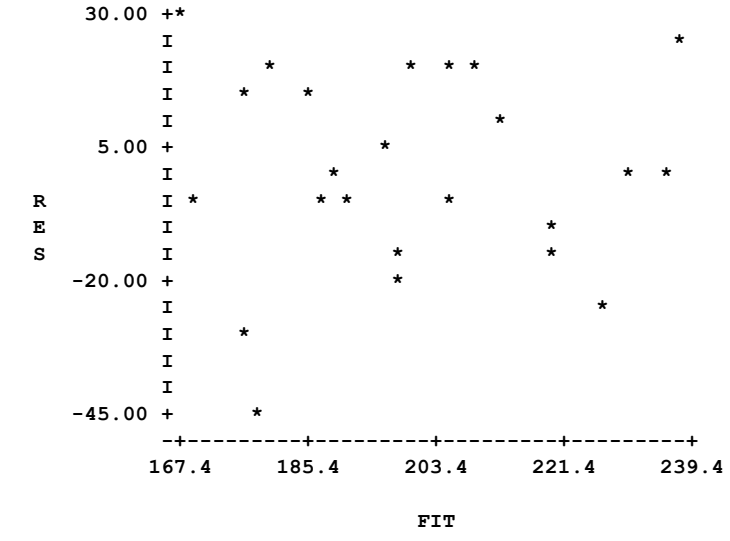
We obtain the following

```
ANALYSIS OF VARIANCE FOR THE VARIABLE :      YIELD
FACTOR(S) IN THE MODEL:  VARIETY    BLOCK

S =       24.1555      R**2 =  53.1%      R**2(ADJ) =  28.2%

-------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
-------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

VARIETY  (A)       9490.000     5      1898.000       3.253
BLOCK    (B)        436.167     3       145.389        .249
RESIDUAL           8752.333    15       583.489
ADJ. TOTAL        18678.500    23
```

```
SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =    199.750 )

FACTOR:  VARIETY               147.000   177.000   207.000   237.000   267.000
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    4  173.000   173.000   11.025 (-------*-------)
  2    4  182.250   182.250   11.025    (-------*-------)
  3    4  194.500   194.500   11.025        (-------*-------)
  4    4  232.750   232.750   11.025                    (-------*------)
  5    4  201.000   201.000   11.025          (-------*-------)
  6    4  215.000   215.000   11.025              (-------*-------)
                               +---------+---------+---------+---------+

FACTOR:  BLOCK                 147.000   177.000   207.000   237.000   267.000
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    6  194.333   194.333    8.540          (-----*-----)
  2    6  196.833   196.833    8.540           (-----*-----)
  3    6  203.667   203.667    8.540             (-----*-----)
  4    6  204.167   204.167    8.540             (-----*-----)
                               +---------+---------+---------+---------+
```

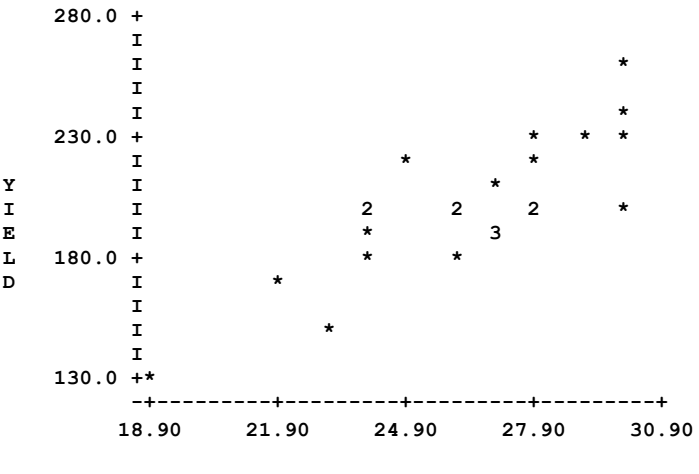
The results above appear reasonable, although STAND is not used as a covariate. The F-statistic corresponding to VARIETY, 3.25, is significant at the 5% level. In addition the summary information for VARIETY clearly indicates a difference of corn yield based on variety. A diagnostic check, the plot of residuals against residuals does not indicate any gross abnormality in the model.

-->PLOT RES, FIT




A scatter plot of corn yield against the number of plants (i.e., stands) reveals there is a strong linear relationship between the two variables. This supports an analysis of covariance with STAND as a covariate.

-->PLOT YIELD, STAND




We obtain an analysis of covariance for this example by entering

-->TWAY YIELD,VARIETY,BLOCK / STAND. HOLD RESIDUALS(RES),FITTED(FIT).

```
ANALYSIS OF VARIANCE FOR THE VARIABLE :      YIELD
FACTOR(S) IN THE MODEL:    STAND  VARIETY    BLOCK

S =        9.8608      R**2 =  92.7%      R**2(ADJ) =  88.0%

------------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
------------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

STAND             12161.167     1     12161.167     125.070
VARIETY  (A)       3653.652     5       730.730       7.515
BLOCK    (B)       1502.394     3       500.798       5.150
RESIDUAL           1361.286    14        97.235
ADJ. TOTAL        18678.500    23

SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =    199.750 )

FACTOR:  VARIETY               177.000   192.000   207.000   222.000   237.000
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    4  173.000   191.802    4.991   (------*------)
  2    4  182.250   190.979    4.611   (-----*------)
  3    4  194.500   193.157    4.503    (------*-----)
  4    4  232.750   219.320    4.757                  (------*------)
  5    4  201.000   189.585    4.687  (-----*------)
  6    4  215.000   213.657    4.503                (-----*------)
                                     +---------+---------+---------+---------+


FACTOR:  BLOCK                 177.000   192.000   207.000   222.000   237.000
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    6  194.333   188.961    3.540   (----*----)
  2    6  196.833   206.234    3.649             (----*-----)
  3    6  203.667   194.266    3.649      (-----*----)
  4    6  204.167   209.539    3.540                (----*----)
                                     +---------+---------+---------+---------+
```

The introduction of a covariate into the analysis has a remarkable effect. We note the residual standard error has been reduced from 24.16 to 9.86. The differences in yield due to variety have become more apparent both visually (in the summary confidence intervals) and statistically (as the F-statistic is almost significant at the 0.1% level). Moreover, we see it was important to block the experiment as differences between blocks are revealed. As before, there is no apparent model inadequacy.

## 8.5 Multi-Way Analysis of Variance

The extension of the analysis of variance from two factors to three or more factors is direct. The only problem we encounter in the analysis of more than two factors may be an inability to study the effect of all possible interactions. In order to analyze all possible interactions, we need to be able to replicate factor combination at all levels. This may not be possible, depending upon the process involved and the factors under study.

An experimenter is almost always interested in the "main effect" of a factor. That is, we will want to test to see if any given factor has a significant effect on the mean level observed. This is an immediate extension of (8.3), and we will extend the analysis of variation of (8.4) to include more than two factors. Interactions are often incorporated into the model according to the level of knowledge we have of the process under study. For example, we may know that a chemical is likely to cause an effect when combined with a specific catalyzing agent, but the effect should be the same within a temperature range. Hence, we may wish to examine the effect of the interaction of chemical and catalyst, but not that of chemical and temperature, nor that of catalyst and temperature.

The NWAY paragraph is used to conduct a multi-way analysis of variance in the SCA System. The paragraph will always provide results for the main effects of each factor, but it will only calculate effects for designated interactions, even if all interactions can be computed. In this way, the NWAY paragraph and the TWAY paragraph differ slightly. We will use two examples to illustrate the NWAY paragraph.

### 8.5.1 Spinning synthetic yarn example

Our first example will be the analysis of variance for data from combinations of Latin square designs. In a Latin square design all observations are subjected to a combination of factor levels exactly once. As a result, only main effects can be studied. The data to be used are the breaking strength of a synthetic yarn (Box, Hunter, and Hunter, 1978, Chapter 8.1). The yarn is subjected to one of three draw ratios ("usual", denoted by A; a 5% increase, B; and a 10% increase, C) as it is spun. Three spinnerets are employed, with each spinneret supplying yarn to three individual bobbins. New bobbins are obtained each time a machine is "doffed" (when a wound bobbin is replaced completely with an empty bobbin). There are 12 doffs, consisting of four sets of Latin squares in the factors of draw ratio and spinneret head. Data are given in Table 9. Information on doff, spinneret head, draw ratio, and breaking strength are stored in the SCA workspace under the labels DOFF, HEAD, DRAW, and STRENGTH, respectively.

**Table 9 Synthetic Yarn Data**

Draw ratio (A, B, or C) and Yarn Breaking Strength for Doff and Spinneret Combination

| *Doff* | *Spinneret Head* 1 | | *Spinneret Head* 2 | | *Spinneret Head* 3 | |
|---|---|---|---|---|---|---|
| 1 | A | 19.56 | B | 23.16 | C | 29.72 |
| 2 | B | 22.94 | C | 27.51 | A | 23.71 |
| 3 | C | 25.06 | A | 17.70 | B | 22.32 |
| 4 | B | 23.24 | C | 23.54 | A | 18.75 |
| 5 | A | 16.28 | B | 22.29 | C | 28.09 |
| 6 | C | 18.53 | A | 19.89 | B | 20.42 |
| 7 | C | 23.98 | A | 20.46 | B | 19.28 |
| 8 | A | 15.33 | B | 23.02 | C | 24.97 |
| 9 | B | 24.41 | C | 22.44 | A | 19.23 |
| 10 | A | 16.65 | B | 22.69 | C | 24.94 |
| 11 | B | 18.96 | C | 24.19 | A | 21.95 |
| 12 | C | 21.49 | A | 15.78 | B | 24.65 |

We now conduct an analysis of variance for the simple additive model.

-->NWAY STRENGTH, DOFF, HEAD, DRAW

```
ANALYSIS OF VARIANCE FOR THE VARIABLE :   STRENGTH
FACTOR(S) IN THE MODEL:     DOFF     HEAD     DRAW

S =        2.2715      R**2 =  74.9%      R**2(ADJ) =  56.1%

-------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
-------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

DOFF     (A)         63.884    11         5.808       1.126
HEAD     (B)         41.617     2        20.809       4.033
DRAW     (C)        202.483     2       101.241      19.621
RESIDUAL            103.198    20         5.160
ADJ. TOTAL          411.183    35
```

```
SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =     21.865 )

FACTOR:  DOFF                    14.000    18.000    22.000    26.000    30.000
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1    3   24.147    24.147    1.256            (-----------*-----------)
  2    3   24.720    24.720    1.256             (-----------*-----------)
  3    3   21.693    21.693    1.256      (-----------*-----------)
  4    3   21.843    21.843    1.256      (-----------*-----------)
  5    3   22.220    22.220    1.256       (-----------*-----------)
  6    3   19.613    19.613    1.256 (-----------*-----------)
  7    3   21.240    21.240    1.256     (-----------*-----------)
  8    3   21.107    21.107    1.256    (-----------*-----------)
  9    3   22.027    22.027    1.256       (-----------*-----------)
 10    3   21.427    21.427    1.256     (-----------*-----------)
 11    3   21.700    21.700    1.256      (-----------*-----------)
 12    3   20.640    20.640    1.256   (-----------*-----------)
                               +---------+---------+---------+---------+


FACTOR:  HEAD                    14.000    18.000    22.000    26.000    30.000
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   12   20.536    20.536     .535           (----*-----)
  2   12   21.889    21.889     .535              (-----*----)
  3   12   23.169    23.169     .535                 (-----*-----)
                               +---------+---------+---------+---------+


FACTOR:  DRAW                    14.000    18.000    22.000    26.000    30.000
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   12   18.774    18.774     .535      (-----*-----)
  2   12   22.282    22.282     .535               (-----*----)
  3   12   24.538    24.538     .535                    (----*-----)
                               +---------+---------+---------+---------+
```

We observe little effect attributed to doffs, and some effect due to the spinneret head (the F-ratio of 4.033 is significant at the 5% level for an F(2, 20) distribution). The effect of draw ratio is appreciable, as is seen both in the F-ratio and the summary information for DRAW. The greatest breaking strength is achieved at draw ratio C (corresponding to a 10% increase in the draw ratio).

We should be aware that the above results are valid only if the model employed is an adequate characterization for the process. A careful analysis should also include an analysis of the residuals of the fit. In this case the fit is adequate, but diagnostic checks are not displayed.

### 8.5.2 Multi-way ANOVA with interactions: Weight gain example

Our second example illustrates the specification and estimation of interaction effects in a three-factor experiment. The example consists of the weight gains observed in male pigs (Snedecor and Cochran, 1989, page 317. The data may be found in Snedecor and Cochran, 1974, page 361.). The factors of the study are food supplements to corn in the pig feeding, these supplements being:

Percentage of Lysine (0, .05, .10, or .15),

Percentage of Methionine (0, .025, or .05), and

Percentage of Protein (12 or 14).

The example is of a 2 x 3 x 4 factorial arrangement in a randomized block design. The blocking factor reflects the replication of the factorial arrangement. The data are shown in Table 10. Data are stored in the SCA workspace under the labels WGTGAINS, BLOCK, METHION, PROTEIN, and LYSINE. The percentage values for both Methionine and Lysine are scaled by multiplication by 1000 and 100, respectively.

**Table 10 Weight Gain Data for Pigs**

| *Percentage of supplement to feed* | | | *Average daily weight gain* | |
|---|---|---|---|---|
| *Lysine (x 100)* | *Methionine (x 1000)* | *Protein* | *Replication 1* | *Replication 2* |
| 0 | 0 | 12 | 1.11 | 0.97 |
| 0 | 0 | 14 | 1.52 | 1.45 |
| 0 | 25 | 12 | 1.09 | 0.99 |
| 0 | 25 | 14 | 1.27 | 1.22 |
| 0 | 50 | 12 | 0.85 | 1.21 |
| 0 | 50 | 14 | 1.67 | 1.24 |
| 5 | 0 | 12 | 1.30 | 1.00 |
| 5 | 0 | 14 | 1.55 | 1.53 |
| 5 | 25 | 12 | 1.03 | 1.21 |
| 5 | 25 | 14 | 1.24 | 1.34 |
| 5 | 50 | 12 | 1.12 | 0.96 |
| 5 | 50 | 14 | 1.76 | 1.27 |
| 10 | 0 | 12 | 1.22 | 1.13 |
| 10 | 0 | 14 | 1.38 | 1.08 |
| 10 | 25 | 12 | 1.34 | 1.41 |
| 10 | 25 | 14 | 1.40 | 1.21 |
| 10 | 50 | 12 | 1.34 | 1.19 |
| 10 | 50 | 14 | 1.46 | 1.39 |
| 15 | 0 | 12 | 1.19 | 1.03 |
| 15 | 0 | 14 | 0.80 | 1.29 |
| 15 | 25 | 12 | 1.36 | 1.16 |
| 15 | 25 | 14 | 1.42 | 1.39 |
| 15 | 50 | 12 | 1.46 | 1.03 |
| 15 | 50 | 14 | 1.62 | 1.27 |

We will illustrate the use of the NWAY paragraph for three cases. As the purpose of the examples is illustrative only, a complete analysis is not presented here. Although the data permits the calculation of all possible interactions, we will first calculate the ANOVA table for main effects. We will also ignore the fact the data are blocked into two groups.

-->NWAY WGTGAINS, METHION, PROTEIN, LYSINE

```
ANALYSIS OF VARIANCE FOR THE VARIABLE :   WGTGAINS
FACTOR(S) IN THE MODEL:  METHION  PROTEIN   LYSINE

S =         .1855      R**2 =  30.9%      R**2(ADJ) =  20.8%


-----------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
-----------------------------------------------------------------
SOURCE        SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

METHION  (A)            .053     2          .026        .764
PROTEIN  (B)            .536     1          .536      15.570
LYSINE   (C)            .043     3          .014        .413
RESIDUAL               1.410    41          .034
ADJ. TOTAL             2.041    47


SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =      1.260 )

FACTOR:  METHION                  1.096     1.176     1.256     1.336     1.416
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   16    1.222     1.222     .038      (---------*--------)
  2   16    1.255     1.255     .038          (---------*---------)
  3   16    1.303     1.303     .038                (---------*--------)
                                    +---------+---------+---------+---------+


FACTOR:  PROTEIN                  1.096     1.176     1.256     1.336     1.416
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   24    1.154     1.154     .027 (-----*------)
  2   24    1.365     1.365     .027                           (------*-----)
                                    +---------+---------+---------+---------+


FACTOR:  LYSINE                   1.096     1.176     1.256     1.336     1.416
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   12    1.216     1.216     .046   (-----------*-----------)
  2   12    1.276     1.276     .046           (-----------*-----------)
  3   12    1.296     1.296     .046             (-----------*-----------)
  4   12    1.252     1.252     .046        (-----------*-----------)
                                    +---------+---------+---------+---------+
```

We observe the only significant main effect is that for protein. We will now incorporate the blocking factor into the analysis. To do this, we will include the BLOCKING sentence in our specification. This sentence specifies the variable containing blocking information (here BLOCK).

-->NWAY WGTGAINS, METHION, PROTEIN, LYSINE. BLOCKING IS BLOCK.

```
ANALYSIS OF VARIANCE FOR THE VARIABLE :   WGTGAINS
FACTOR(S) IN THE MODEL:    BLOCK  METHION  PROTEIN   LYSINE

S =         .1787      R**2 =  37.4%      R**2(ADJ) =  26.5%


-----------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
-----------------------------------------------------------------
SOURCE       SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

BLOCK                  .133     1          .133       4.178
METHION  (A)           .053     2          .026        .823
PROTEIN  (B)           .536     1          .536      16.777
LYSINE   (C)           .043     3          .014        .445
RESIDUAL              1.277    40          .032
ADJ. TOTAL            2.041    47


SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =      1.260 )

FACTOR:  BLOCK                    1.096     1.176     1.256     1.336     1.416
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   24    1.313     1.313     .026                    (------*------)
  2   24    1.207     1.207     .026       (------*-----)
                                  +---------+---------+---------+---------+


FACTOR:  METHION                  1.096     1.176     1.256     1.336     1.416
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   16    1.222     1.222     .036      (---------*--------)
  2   16    1.255     1.255     .036           (--------*--------)
  3   16    1.303     1.303     .036                (--------*--------)
                                  +---------+---------+---------+---------+


FACTOR:  PROTEIN                  1.096     1.176     1.256     1.336     1.416
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   24    1.154     1.154     .026 (-----*------)
  2   24    1.365     1.365     .026                          (------*-----)
                                  +---------+---------+---------+---------+


FACTOR:  LYSINE                   1.096     1.176     1.256     1.336     1.416
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   12    1.216     1.216     .045    (----------*----------)
  2   12    1.276     1.276     .045          (----------*----------)
  3   12    1.296     1.296     .045             (----------*----------)
  4   12    1.252     1.252     .045        (----------*----------)
                                  +---------+---------+---------+---------+
```

Due to orthogonality in the design (see Box, Hunter, and Hunter, 1978, Appendices 6B and 14D), the only change from the previous results is that the residual sum of squares has been "partitioned" into that portion due to blocking and remaining error. This decreases the value of s2 from .034 to .032, and increases the F-values of main effects accordingly. We note that in addition to protein, the effect due to blocking is also significant.

We will now incorporate interactions into the model. Note that each "source of variation" in the ANOVA table is associated with a letter. Methionine (METHION) is associated with the letter A, protein with B, and lysine with C. We will use the INTERACTIONS sentence to specify those interactions we wish to include in the model. Specifying AB results in the inclusion of the methionine-protein interaction. Similar associations are made for AC, BC, or ABC. We will include all interactions.

```
-->NWAY   WGTGAINS, METHION, PROTEIN, LYSINE.      @
      INTERACTIONS ARE AB, AC, BC, ABC.            @
      BLOCKING IS BLOCK.
```

```
ANALYSIS OF VARIANCE FOR THE VARIABLE :   WGTGAINS
FACTOR(S) IN THE MODEL:    BLOCK  METHION  PROTEIN   LYSINE

S =         .1658      R**2 =  69.0%      R**2(ADJ) =  36.7%

------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE (BASED ON SEQUENTIAL SUM OF SQUARES)
------------------------------------------------------------
SOURCE        SUM OF SQUARES    DF   MEAN SQUARE     F-RATIO

BLOCK                  .133     1          .133       4.854
METHION  (A)           .053     2          .026        .956
PROTEIN  (B)           .536     1          .536      19.492
LYSINE   (C)           .043     3          .014        .518
AB                     .082     2          .041       1.495
AC                     .254     6          .042       1.543
BC                     .240     3          .080       2.911
ABC                    .068     6          .011        .415
RESIDUAL               .632    23          .027
ADJ. TOTAL            2.041    47

SUMMARY AND CONFIDENCE INTERVALS FOR EACH FACTOR   ( GRAND MEAN =      1.260 )

FACTOR:  BLOCK                    1.104     1.184     1.264     1.344     1.424
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   24    1.313     1.313     .024                    (-----*-----)
  2   24    1.207     1.207     .024         (-----*-----)
                               +---------+---------+---------+---------+

FACTOR:  METHION                  1.104     1.184     1.264     1.344     1.424
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   16    1.222     1.222     .034       (--------*-------)
  2   16    1.255     1.255     .034            (--------*-------)
  3   16    1.303     1.303     .034                  (--------*-------)
                               +---------+---------+---------+---------+

FACTOR:  PROTEIN                  1.104     1.184     1.264     1.344     1.424
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   24    1.154     1.154     .024 (----*-----)
  2   24    1.365     1.365     .024                           (-----*-----)
                               +---------+---------+---------+---------+

FACTOR:  LYSINE                   1.104     1.184     1.264     1.344     1.424
LEVEL  N    MEAN   ESTIMATE  STD ERR +---------+---------+---------+---------+
  1   12    1.216     1.216     .041     (---------*---------)
  2   12    1.276     1.276     .041            (---------*----------)
  3   12    1.296     1.296     .041               (---------*---------)
  4   12    1.252     1.252     .041         (---------*----------)
                               +---------+---------+---------+---------+
```

Again, due to orthogonality, the sum of squares for main effects (and now blocking) have not changed. The residual sum of square is further partitioned into that portion due to each indicated interaction. In addition to the effects of blocking and protein, we see there is some indication of an interaction between protein and lysine (BC). Hence the analysis could continue concentrating on these factors alone to determine the best factor levels and combinations for greatest average daily weight gain.

# SUMMARY OF THE SCA PARAGRAPHS IN CHAPTER 8

This section provides a summary of those SCA paragraphs employed in this chapter. The syntax for each paragraph is presented in both a brief and full form. The brief display of the syntax contains the most frequently used sentences of a paragraph, while the full display presents all possible modifying sentences of a paragraph. In addition, special remarks related to a paragraph may also be presented with the description.

Each SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

The paragraphs to be explained in this summary are OWAY, TWAY and NWAY.

Legend (see Chapter 2 for further explanation)

v : variable name
i : integer
r : real value
w : character data (must be enclosed within single apostrophes)

## OWAY Paragraph

The OWAY paragraph is used to perform one-way analysis of variance (or covariance). The paragraph can be used for both balanced and unbalanced designs. Estimation is based on a least squares fit of the linear model corresponding to the appropriate analysis of variance.

## Syntax of the OWAY Paragraph

### Brief syntax

```
OWAY        VARIABLES ARE v1, v2, ---.

Required sentence: VARIABLES
```

### Full syntax

```
OWAY        VARIABLES ARE v1, v2, --- .                    @
            CONSTANT. / NO CONSTANT.                       @
            SPAN IS i1 i2.                                 @
            BLOCKING IS v.                                 @
            WEIGHT IS v.                                   @
            TABLE. / NO TABLE.                             @
            CIPLOT. / NO CIPLOT.                           @
            ANOVA IS w.                                    @
            OUTPUT IS LEVEL(w), PRINT(w1, w2, ---),        @
                  NOPRINT(w1, w2, ---).

Required sentence: VARIABLES
```

## Sentences Used in the OWAY Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the variables containing the responses for various factor levels. When only two variables are specified, it is assumed the first is the variable of responses and the second is factor levels.

**CONSTANT sentence**

The CONSTANT sentence is used to specify whether a constant term is included in the analysis. The default is to include the constant term in the model.

**SPAN sentence**

The SPAN sentence is used to specify the span of cases, from i1 to i2, of the response variable and corresponding factors to be used in the analysis. The default is to use all observations.

**BLOCKING sentence**

The BLOCKING sentence is used to specify the variable containing blocking information. The default is no blocking variable.

*The following are infrequently used sentences of the paragraph*

**WEIGHT sentence (see Chapter 9.6.5)**

The WEIGHT sentence is used to specify a variable containing a weight for each response. The default is 1.0 for each observation.

**TABLE sentence**

The TABLE sentence is used to specify the display of a table of factor means and standard deviations. The default is TABLE.

**CIPLOT sentence**

The CIPLOT sentence is used to specify the display of the confidence interval plots of estimated main effects. The default is CIPLOT.

**ANOVA sentence (see Chapter 9.6.1)**

The ANOVA sentence is used to obtain different analysis of variance tables. The keyword may be PARTIAL (for partial sum of squares), SEQUENTIAL (for the sequential sum of squares), BOTH, or NONE. The default is SEQUENTIAL. The partial sum of squares table shows how each explanatory variable of a regression contributes to the total sum of squares if all other factors in the model are included. The sequential sum of squares table shows the contribution to the total sum of squares of each factor in the regression model, assuming each factor is fitted in the sequential order specified in the VARIABLES sentence.

**OUTPUT sentence**

The OUTPUT sentence is used to control the amount of output printed for computed statistics. Control is achieved in a two stage procedure. First a basic LEVEL of output is specified. Output may then be increased from this level by use of PRINT, or decreased from this level by use of NOPRINT.

The keywords for LEVEL and output printed are:

BRIEF : ESTIMATES
NORMAL : ESTIMATES
DETAILED : ESTIMATES, RCORR, and CORR

where the reserved words on the right denote:

CORR : the correlation matrix for all variables in regression analysis
RCORR : the correlation matrix for the estimates of the regression coefficients
ESTIMATES : the estimates of the regression coefficients

These reserved words are also keywords for PRINT and NOPRINT. The default for LEVEL is NORMAL, or the level specified in the PROFILE paragraph.

## TWAY Paragraph

The TWAY paragraph may be used to perform one-way or two-way analysis of variance (or covariance). A Box-Cox transformation analysis can be incorporated within the analysis. The paragraph can be used for both balanced and unbalanced designs. Estimation is based on a least squares fit of the linear model corresponding to the appropriate analysis of variance (or covariance).

## Syntax of the TWAY Paragraph

### Brief syntax

```
TWAY        VARIABLES ARE v1, v2, v3.

        Required sentence: VARIABLES
```

### Full syntax

```
TWAY        VARIABLES ARE v1, v2, v3 / --- .                    @
            INTERACTION. / NO INTERACTION.                      @
            CONSTANT. / NO CONSTANT.                            @
            POWERS ARE r1, r2, r3.                              @
            SPAN IS i1 i2.                                      @
            BLOCKING IS v.                                      @
            HOLD POWER(v), ESTIMATES(v1, v2, ---).              @
                TVALUES(v1, v2, ---),                           @
                FVALUES(v1, v2, ---),                           @
                RESIDUALS(v1, v2, ---),                         @
                FITTED(v1, v2, ---),                            @
                MSE(v), MEAN(v), STDV(v).                       @
            WEIGHT IS v.                                        @
            TABLE. / NO TABLE.                                  @
            CIPLOT. / NO CIPLOT.                                @
            ANOVA IS w.                                         @
            OUTPUT IS LEVEL(w), PRINT(w1, w2, ---),             @
                    NOPRINT(w1, w2, ---).

Required sentence: VARIABLES
```

## Sentences Used in the TWAY Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the names of the responses and factors of the ANOVA model. The first variable is the response variable, and the others are all treated as factors. All variables listed before the symbol ' / ' must contain integer or character information. If a variable is specified after the symbol ' / ', it is assumed to be a concomitant variable (covariate). The concomitant variables can be either a vector or a matrix. See Section 8.4 for more information on covariates.

**INTERACTION sentence**

The INTERACTION sentence is used to specify that an interaction term be included in ANOVA model. The default is to include an interaction term if the interaction effects can be estimated.

**CONSTANT sentence**

The CONSTANT sentence is used to specify whether a constant term is included in the analysis. The default is to include the constant term in the model.

**POWER sentence**

The POWER sentence is used to specify the values to be employed in a Box-Cox power transformation analysis. The value r1 is the increment for the power (lambda) value, within the interval beginning at r2 and ending at r3. The default values are 0.1, -1.0 and 1.0 respectively. The number of arguments in this sentence may be 1, 2 or 3. If only a particular transformation is desired, the arguments can be specified as 0.0, r2, where 0.0 indicates no increment, and r2 indicates the specific power value to be used in the analysis.

**SPAN sentence**

The SPAN sentence is used to specify the span of cases, from i1 to i2, of the response variable and corresponding factors to be used in the analysis. The default is to use all observations.

**BLOCKING sentence**

The BLOCKING sentence is used to specify the variable containing blocking information. The default is no blocking variable.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. The default is that no values are retained after the paragraph is executed. The values that may be retained are:

| | | |
|---|---|---|
| POWER | : | the powers (lambda) employed in the transformation analysis |
| ESTIMATES | : | the parameter estimates for each lambda. Each parameter is represented by a label (variable). If the constant term is present, it is represented by v1 |
| TVALUES | : | the corresponding t-values of the above estimates |
| FVALUES | : | the corresponding F-values for each group of estimates |
| RESIDUALS | : | the residuals of the fitted model. The number of variable names specified must be the same as the number of lambda in the transformation analysis |
| FITTED | : | the value for each dependent variable based on the estimated model. The number of variables specified must be the same as the number of lambdas in the transformation analysis |
| MSE | : | the mean square error of the model for each lambda |
| MEAN | : | the cell means for each lambda |
| STDV | : | the cell standard deviations for each lambda |

*The following are infrequently used sentences of the paragraph*

**WEIGHT sentence (see Chapter 9.6.5)**

The WEIGHT sentence is used to specify a variable containing a weight for each response. The default is 1.0 for each observation.

**TABLE sentence**

The TABLE sentence is used to specify the display of a table of factor means and standard deviations. The default is TABLE.

**CIPLOT sentence**

The CIPLOT sentence is used to specify the display of the confidence interval plots of estimated main effects. The default is CIPLOT.

**ANOVA sentence (see Chapter 9.6.1)**

The ANOVA sentence is used to obtain different analysis of variance tables. The keyword may be PARTIAL (for partial sum of squares), SEQUENTIAL (for the sequential sum of squares), BOTH, or NONE. The default is SEQUENTIAL. The partial sum of squares table shows how each explanatory variable of a regression contributes to the total sum of squares if all other factors in the model are included. The sequential sum of squares table shows the contribution to the total sum of squares of each factor in the regression model, assuming each factor is fitted in the sequential order specified in the VARIABLES sentence.

**OUTPUT sentence**

The OUTPUT sentence is used to control the amount of output printed for computed statistics. Control is achieved in a two stage procedure. First a basic LEVEL of output is specified. Output may then be increased from this level by use of PRINT, or decreased from this level by use of NOPRINT.

The keywords for LEVEL and output printed are:

BRIEF : ESTIMATES
NORMAL : ESTIMATES
DETAILED : ESTIMATES, RCORR, and CORR

where the reserved words on the right denote:

CORR : the correlation matrix for all variables in regression analysis
RCORR : the correlation matrix for the estimates of the regression coefficients
ESTIMATES : the estimates of the regression coefficients

These reserved words are also keywords for PRINT and NOPRINT. The default for LEVEL is NORMAL, or the level specified in the PROFILE paragraph.

## NWAY Paragraph

The NWAY paragraph is used to perform multi-way analysis of variance (or covariance). A Box-Cox transformation analysis can be incorporated within the analysis. The paragraph can be used for both balanced and unbalanced designs. Estimation is based on a least squares set of the linear model corresponding to the appropriate analysis of variance (or covariance).

## Syntax of the NWAY Paragraph

### Brief syntax

```
NWAY        VARIABLES ARE v1, v2, ---. @
            INTERACTIONS ARE e1, e2, ---.
```

Required sentence: **VARIABLES**

### Full syntax

```
NWAY        VARIABLES ARE v1, v2, --- / ---.                  @
            INTERACTIONS ARE e1, e2. --- .                    @
            BLOCKING IS v.                                    @
            CONSTANT. / NO CONSTANT.                          @
            POWERS ARE r1, r2, r3.                            @
            SPAN IS il, i2.                                   @
            HOLD POWER(v), ESTIMATES(v1, v2, ---).            @
               TVALUES(v1, v2, ---),                          @
               FVALUES(v1, v2, ---),                          @
               RESIDUALS(v1, v2, ---),                        @
               FITTED(v1, v2, ---),                           @
               MSE(v), MEAN(v), STDV(v).                      @
            WEIGHT IS v.                                      @
            ANOVA IS w.                                       @
            OUTPUT IS LEVEL(w), PRINT(w1, w2, ---),           @
               NOPRINT(w1, w2, ---).
```

Required sentence: **VARIABLES**

## Sentences Used in the NWAY Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the names of the response and factors of the ANOVA model. The first variable is the response variable, and the others are all treated as factors. All variables listed before the symbol " / " must contain integer or character information. If a variable is specified after the symbol " / ", it is assumed to be a concomitant variable (covariate). The concomitant variables can be either a vector or a matrix.

**INTERACTIONS sentence (see Section 8.5.2)**

The INTERACTIONS sentence is used to specify that the interaction terms be included in ANOVA model. The term AB denotes the interaction between the first and the second factors; BC denotes the interaction between the second and third factors; ABC denotes the interaction between the first, the second, and the third factors, etc. The default is no interaction terms.

**BLOCKING sentence**

The BLOCKING sentence is used to specify the variable containing blocking information. The default is no blocking variable.

**CONSTANT sentence**

The CONSTANT sentence is used to specify whether a constant term is included in the analysis. The default is to include the constant term in the model.

**POWER sentence**

The POWER sentence is used to specify the values to be employed in a Box-Cox power transformation analysis. The value r1 is the increment for the power (lambda) value, within the interval beginning at r2 and ending at r3. The default values are 0.1, -1.0 and 1.0 respectively. The number of arguments in this sentence may be 1, 2 or 3. If only a particular transformation is desired, the arguments can be specified as 0.0, r2, where 0.0 indicates no increment, and r2 indicates the specific power value to be used in the analysis.

**SPAN sentence**

The SPAN sentence is used to specify the span of cases, from i1 to i2, of the response variable and corresponding factors to be used in the analysis. Default is to use all observations.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. The default is that no values are retained after the paragraph is executed. The values that may be retained are:

| | | |
|---|---|---|
| POWER | : | the powers (lambda) employed in the transformation analysis |
| ESTIMATES | : | the parameter estimates for each lambda. Each parameter is represented a label (variable). If the constant term is present, it is represented by v1 |
| TVALUES | : | the corresponding t-values of the above estimates |
| FVALUES | : | the corresponding F-values for each group of estimates |
| RESIDUALS | : | the residuals of the fitted model. The number of variable names specified must be the same as the number of lambdas in the transformation analysis |
| FITTED | : | the value for each dependent variable based on the estimated model. The number of variables specified must be the same as the number of lambdas in the transformation analysis |
| MSE | : | the mean square error of the model for each lambda |
| MEAN | : | the cell means for each lambda |
| STDV | : | the cell standard deviations for each lambda |

*The following are infrequently used sentences of the paragraph*

**WEIGHT sentence (see Chapter 9.6.5)**

The WEIGHT sentence is used to specify a variable containing a weight for each response. The default is 1.0 for each observation.

**ANOVA sentence (see Chapter 9.6.1)**

The ANOVA sentence is used to obtain different analysis of variance tables. The keyword may be PARTIAL (for partial sum of squares), SEQUENTIAL (for sequential sum of squares), BOTH or NONE. The default is SEQUENTIAL. The partial sum of squares table shows how each explanatory variable of a regression contributes to the total sum of squares if all other factors in the model are included. The sequential sum of squares table shows the contribution to the total sum of squares of each factor in the regression model, assuming each factor is fitted in the sequential order specified in the VARIABLES sentence.

**OUTPUT sentence**

The OUTPUT sentence is used to control the amount of output printed for computed statistics. Control is achieved in a two stage procedure. First a basic LEVEL of output is specified. Output may then be increased from this level by use of PRINT, or decreased from this level by use of NOPRINT.

The keywords for LEVEL and output printed are:

BRIEF : ESTIMATES
NORMAL : ESTIMATES
DETAILED : ESTIMATES, RCORR, and CORR

where the reserved words on the right denote:

CORR : the correlation matrix for all variables in regression analysis
RCORR : the correlation matrix for the estimates of the regression coefficients
ESTIMATES : the estimates of the regression coefficients

These reserved words are also keywords for PRINT and NOPRINT. The default for LEVEL is NORMAL, or the level specified in the PROFILE paragraph.

## REFERENCES

Barker, T.B. (1985). *Quality by Experimental Design*. New York: Marcel Dekker.

Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society*, B26: 211-243.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). *Statistics for Experimenters*. New York: Wiley.

Montgomery, D.C. (1984). *Design and Analysis of Experiments, 2nd edition*. New York: Wiley.

Snedecor, G.W. and Cochran, W.G. (1974). *Statistical Methods, 6th edition*. Ames, IA: Iowa State Press.

Snedecor, G.W. and Cochran, W.G. (1989). *Statistical Methods, 8th edition*. Ames, IA: Iowa State Press.

Tukey, J.W. (1949). One Degree of Freedom for Non-Additivity. *Biometrics*, 5: 232-242

# CHAPTER 9

# LINEAR REGRESSION ANALYSIS

Regression analysis is a statistical method used in modeling the relationships that may exist between variables. In a regression analysis we relate the response of a dependent variable to the values of potential explanatory variables. Once a model is established, it may be used to make inferences about the formulated relationships, or to make predictions for future responses when the explanatory variables are at designated levels.

The simplest of all such relationships occurs when the responses for the dependent variable appear to nearly follow a straight line when plotted against the values of a single explanatory variable. In such a relationship, the predicted value of the dependent variable, $\hat{Y}$, can be obtained from the linear equation

$$\hat{Y} = a + b\,X$$

where X is an explanatory variable and a and b are estimated values. We can extend this linear relation to include more than one explanatory variables with the equation

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

Model fitting of the above linear regression model can be easily accomplished using the REGRESS paragraph. To demonstrate this and other regression capabilities we will consider several examples.

## 9.1 Multiple Regression Analysis

To illustrate the use of regression analysis in an industrial application, we will analyze a set of data pertaining to beer distribution (Montgomery, 1984, page 424). In an effort to analyze the delivery system of a beer distributor, in particular, the time required to service a retail outlet, the following data and factors are studied:

(1) The delivery time (in minutes) to service an outlet,

(2) The number of cases of beer delivered to the outlet, and

(3) The maximum distance the delivery man must travel.

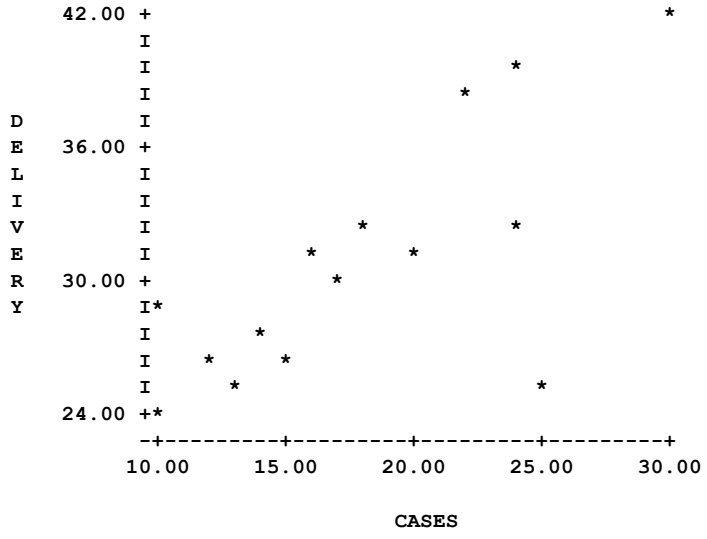
The data are shown in Table 1 and are stored in the SCA workspace under the labels DELIVERY, CASES, and DISTANCE, respectively.

**Table 1 Beer delivery time data**

| *Observation Number* | *Number of Cases* *CASES* | *Distance* *DISTANCE* | *Delivery Time (minutes)* *DELIVERY* |
|---|---|---|---|
| 1 | 10 | 30 | 24 |
| 2 | 15 | 25 | 27 |
| 3 | 10 | 40 | 29 |
| 4 | 20 | 18 | 31 |
| 5 | 25 | 22 | 25 |
| 6 | 18 | 31 | 33 |
| 7 | 12 | 26 | 26 |
| 8 | 14 | 34 | 28 |
| 9 | 16 | 29 | 31 |
| 10 | 22 | 37 | 39 |
| 11 | 24 | 20 | 33 |
| 12 | 17 | 25 | 30 |
| 13 | 13 | 27 | 25 |
| 14 | 30 | 23 | 42 |
| 15 | 24 | 33 | 40 |

We first plot DELIVERY against both CASES and DISTANCE to check if there are any obvious relationships or unusual occurrences in the data.

-->PLOT DELIVERY, CASES



```
-->PLOT   DELIVERY, DISTANCE

      42.00 +        *
            I
            I                          *
            I                                 *
            I
  D         I
  E   36.00 +
  L         I
  I         I
  V         I   *                 *
  E         I*                *
  R   30.00 +            *
  Y         I                                    *
            I                           *
            I           **
            I       *       *
      24.00 +                    *
            -+---------+---------+---------+---------+
           18.00     24.00     30.00     36.00     42.00

                           DISTANCE
```

In the scatter plot between DELIVERY and CASES, we observe a strong linear relationship between the number of cases delivered and delivery time. However, there appears to be an aberration from linearity for the delivery time when 25 cases are delivered. This corresponds to observation number 5. No clear patterns are seen in the scatter plot between DELIVERY and DISTANCE.

We now will regress DELIVERY on CASES and DISTANCE. That is, we will use the REGRESS paragraph to obtain the fitted equation (omitting the "hat")

$$\text{DELIVERY} = b_0 + b_1 \text{ CASES} + b_2 \text{ DISTANCE}.$$

To obtain this fit, we specify the dependent and explanatory variables as

```
REGRESS   DELIVERY, CASES, DISTANCE
```

The actual REGRESS command is shown below together with other modifying (or optional) sentences that will be explained later. The continuation character (@) is used to continue our commands to a second line.

```
-->REGRESS   DELIVERY, CASES, DISTANCE.  DIAGNOSTICS ARE FULL.    @
-->          HOLD  RESIDUALS(RESID), FITTED(FIT)
```

We obtain the following:

```
REGRESSION ANALYSIS FOR THE VARIABLE     DELIVERY

PREDICTOR       COEFFICIENT    STD. ERROR     T-VALUE
INTERCEPT          2.31120       5.85730         .39
   CASES            .87720        .15303        5.73
DISTANCE            .45592        .14676        3.11

CORRELATION MATRIX OF REGRESSION COEFFICIENTS

   CASES      1.00
DISTANCE       .41     1.00
            CASES DISTANCE

S =        3.1408      R**2 =  73.7%      R**2(ADJ) =  69.3%


---------------------------
ANALYSIS OF VARIANCE TABLE
---------------------------

  SOURCE     SUM OF SQUARES    DF    MEAN SQUARE     F-RATIO
REGRESSION          331.359     2        165.679      16.795
 RESIDUAL           118.375    12          9.865
ADJ. TOTAL          449.733    14

  SOURCE      SEQUENTIAL SS    DF    MEAN SQUARE     F-RATIO
  CASES             236.161     1        236.161      23.940
  DISTANCE           95.198     1         95.198       9.650


DIAGNOSTIC STATISTICS:
                                        STUDENTIZED
 CASE   OBSERVED            STANDARDIZED  DELETED    COOK'S
  NO.     VALUE    RESIDUAL  RESIDUAL   RESIDUAL   DISTANCE   LEVERAGE
   1    24.0000     -.7609      -.27       -.26       .006       .198
   2    27.0000      .1327       .05        .04       .000       .124
   3    29.0000     -.3201      -.13       -.12       .003       .356
   4    31.0000     2.9381      1.09       1.09       .136       .258
   5    25.0000    -9.2716     -3.27 *    -9.44 *     .803       .184
   6    33.0000      .7656       .26        .24       .002       .086
   7    26.0000     1.3084       .46        .45       .016       .183
   8    28.0000    -2.0934      -.72       -.70       .028       .139
   9    31.0000     1.4318       .47        .46       .006       .075
  10    39.0000      .5212       .21        .20       .008       .348
  11    33.0000      .5175       .18        .18       .003       .203
  12    30.0000     1.3783       .46        .45       .007       .094
  13    25.0000    -1.0247      -.35       -.34       .007       .137
  14    42.0000     2.8865      1.14       1.16       .237       .352
  15    40.0000     1.5905       .59        .57       .041       .262

"*" DENOTES AN OBSERVATION WITH A LARGE RESIDUAL
```

The fitted equation from the above regression can be obtained from the first few lines of output as

DELIVERY = 2.31 + .88 CASES + .46 DISTANCE.

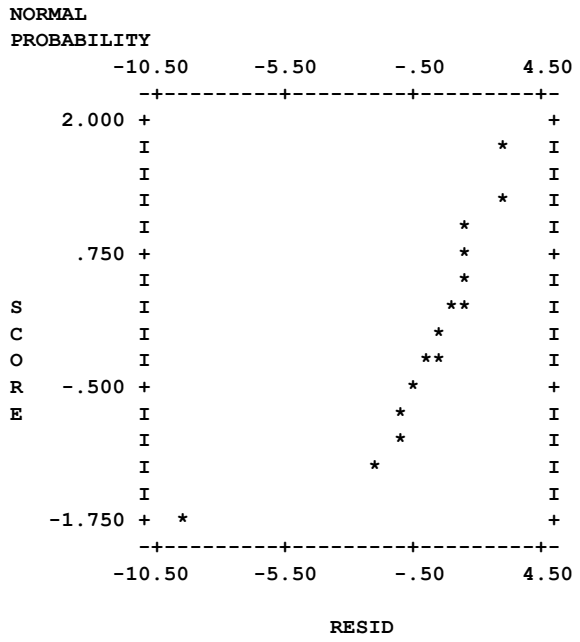
The estimates associated with CASES and DISTANCE are statistically significant as the absolute value of their t-values are greater than 2.15 (the approximate 5% critical level for the sample size). The small t-value associated with the intercept term, 0.39, implies that this estimate cannot be distinguished statistically from zero. Hence we may wish to exclude this

term from our model. However, before we employ this equation, we need to check the models's validity.

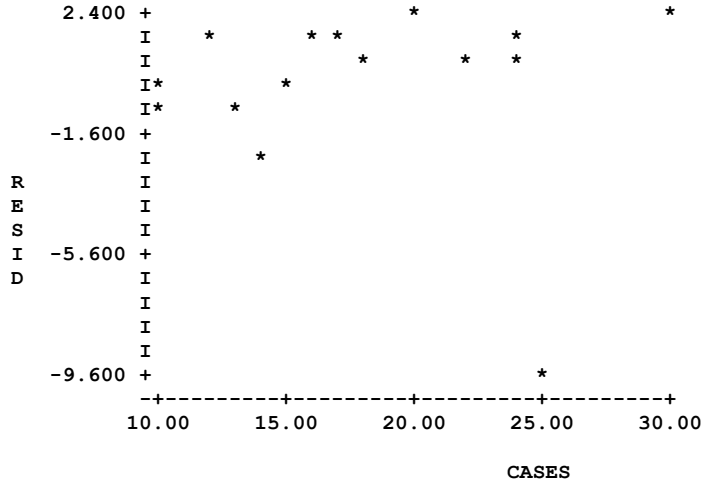
In an effort to assess the model's validity we requested a FULL display of a set of diagnostic statistics with the inclusion of the DIAGNOSTICS sentence in the paragraph. The value of the standardized residual, studentized deleted residual and Cook's distance (see Section 9.2.1) for case number 5 mark it as a potential outlier.

The values obtained using the fitted equation have been retained under the label FIT. The residuals of the fit (i.e., DELIVERY - FIT) are stored in the variable RESID. The residuals should approximate a set of values that are randomly drawn from a standard normal distribution. We can observe the spurious nature of this observation (case number 5) in the probability plot of the residuals and in the plots of the residual series RESID against the explanatory variables CASES and DISTANCE (see Section 9.4.3). In each case there is only one observation that leads us to question the adequacy of the fitted model, observation 5.

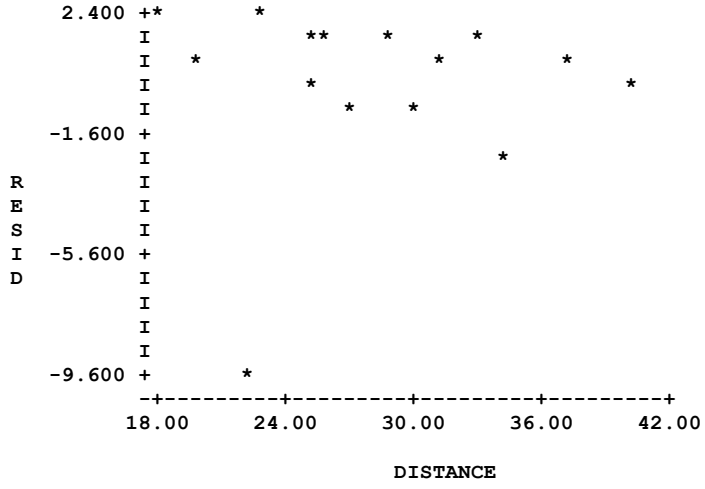
-->PPLOT RESID



--->PLOT RESID, CASES




-->PLOT RESID, DISTANCE




Montgomery (1984, page 426) suggests that a data recording error could have been made at observation 5 (DELIVERY entered as 25 instead of 35). However, there was no way to verify this. To observe the effect of a possible recording error, we will recode the value to 35 and re-run the regression analysis. We can recode the value directly using an analytic assignment statement (see Appendix A).

```
-->DELIVERY(5) = 35
   --

-->REGRESS  DELIVERY, CASES, DISTANCE.  DIAGNOSTICS ARE FULL.  @
-->       HOLD RESIDUALS (RESID),  FITTED (FIT)

   REGRESSION ANALYSIS FOR THE VARIABLE     DELIVERY

   PREDICTOR       COEFFICIENT    STD. ERROR     T-VALUE
   INTERCEPT          2.84270       2.05241        1.39
      CASES            .98803        .05362       18.43
   DISTANCE            .38951        .05143        7.57

   CORRELATION MATRIX OF REGRESSION COEFFICIENTS

      CASES      1.00
   DISTANCE       .41     1.00
                CASES DISTANCE

   S =        1.1005      R**2 =  96.6%      R**2(ADJ) =  96.0%
```

```
---------------------------
ANALYSIS OF VARIANCE TABLE
---------------------------

  SOURCE     SUM OF SQUARES    DF    MEAN SQUARE     F-RATIO
REGRESSION          411.199     2        205.600     169.750
 RESIDUAL            14.534    12          1.211
ADJ. TOTAL          425.733    14

  SOURCE      SEQUENTIAL SS    DF    MEAN SQUARE     F-RATIO
  CASES             341.716     1        341.716     282.133
  DISTANCE           69.483     1         69.483      57.368


DIAGNOSTIC STATISTICS:
                                          STUDENTIZED
 CASE    OBSERVED           STANDARDIZED   DELETED     COOK'S
  NO.      VALUE    RESIDUAL  RESIDUAL    RESIDUAL   DISTANCE   LEVERAGE
   1     24.0000     -.4081      -.41       -.40       .014       .198
   2     27.0000     -.4007      -.39       -.37       .007       .124
   3     29.0000      .6968       .79        .78       .115       .356
   4     31.0000     1.3857      1.46       1.54       .247       .258
   5     35.0000    -1.1125     -1.12      -1.13       .094       .184
   6     33.0000      .2981       .28        .27       .003       .086
   7     26.0000     1.1738      1.18       1.20       .104       .183
   8     28.0000    -1.9183     -1.88      -2.14 *     .190       .139
   9     31.0000     1.0532       .99        .99       .027       .075
  10     39.0000      .0090       .01        .01       .000       .348
  11     33.0000    -1.3454     -1.37      -1.43       .160       .203
  12     30.0000      .6232       .60        .58       .012       .094
  13     25.0000    -1.2037     -1.18      -1.20       .074       .137
  14     42.0000      .5579       .63        .61       .072       .352
  15     40.0000      .5910       .63        .61       .046       .262

"*" DENOTES AN OBSERVATION WITH A LARGE RESIDUAL
```

We observe that the fitted equation is only slightly changed from

TIME = 2.31 + .88 CASES + .46 DISTANCE

to

TIME = 2.84 + .99 CASES + .39 DISTANCE

However, recoding the single point has an appreciable effect on variance. We see:

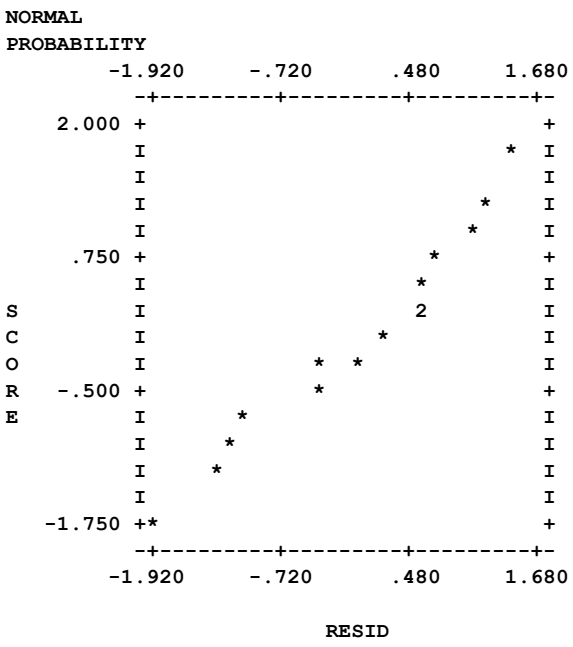
(1) Standard errors of coefficients for CASES and DISTANCE are 1/3 of what they were previously (resulting in a dramatic change in the t-values of the coefficients);

(2) A substantial change in the amount of the REGRESSION sum of squares in the ANOVA table (from 331.359 to 411.199); and hence a

(3) Change in R2 from 73.7% to 96.6%. (Please see Section 9.4.2 for a more complete discussion on the interpretation of R2.)

The probability plot of the residuals reveals no apparent model inadequacy.

-->PPLOT RESID




Similarly, as would be expected, the plots of RESID against the explanatory variables CASES and DISTANCE now show no evidence of model inadequacy. Hence it is possible a simple recording error has affected the results of the analysis dramatically. This indicates the need for a careful diagnostic check of a model (see Section 9.4.3). We will now concentrate on one aspect of diagnostic checking, the identification of influential observations.

## 9.2 Statistical Measures for Spurious and Influential Observations in a Regression Analysis

The previous example demonstrates the need to diagnostically check a model, both for model adequacy and to discover spurious and influential observations. The example presented in this section and the prior beer data example are used to illustrate the diagnostic statistics computed in the SCA System to help highlight such observations. The data used in this section are taken from Neter, Wasserman, and Kutner (1983, Chapters 8 and 11). Discussions related to the identification of spurious and influential observations, and remedial measures, can be found in Neter, Wasserman, and Kutner (1983, Sections 11.5 and 11.6).

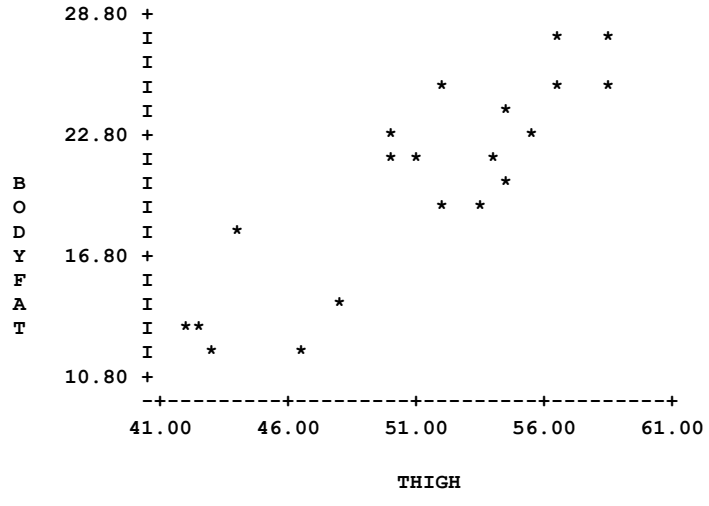
The data came from a study of the relation of bodyfat to thigh circumference and triceps skinfold thickness of 20 subjects. The data are shown in Table 2 and are stored in the SCA workspace under the labels BODYFAT, THIGH, and TRICEPS, respectively.

**Table 2 Bodyfat study data**

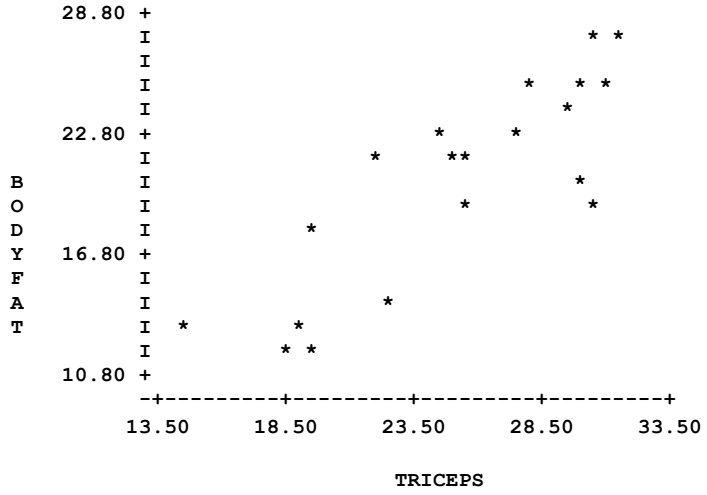
| *Subject* | *Thigh Circumference THIGH* | *Triceps Skinfold Thickness TRICEPS* | *Body Fat BODYFAT* |
|---|---|---|---|
| 1 | 43.1 | 19.5 | 11.9 |
| 2 | 49.8 | 24.7 | 22.8 |
| 3 | 51.9 | 30.7 | 18.7 |
| 4 | 54.3 | 29.8 | 20.1 |
| 5 | 42.2 | 19.1 | 12.9 |
| 6 | 53.9 | 25.6 | 21.7 |
| 7 | 58.5 | 31.4 | 27.1 |
| 8 | 52.1 | 27.9 | 25.4 |
| 9 | 49.9 | 22.1 | 21.3 |
| 10 | 53.5 | 25.5 | 19.3 |
| 11 | 56.6 | 31.1 | 25.4 |
| 12 | 56.7 | 30.4 | 27.2 |
| 13 | 46.5 | 18.7 | 11.7 |
| 14 | 44.2 | 19.7 | 17.8 |
| 15 | 42.7 | 14.6 | 12.8 |
| 16 | 54.4 | 29.5 | 23.9 |
| 17 | 55.3 | 27.7 | 22.6 |
| 18 | 58.6 | 30.2 | 25.4 |
| 19 | 48.2 | 22.7 | 14.8 |
| 20 | 51.0 | 25.2 | 21.1 |

Scatter plots of BODYFAT against both explanatory variables THIGH and TRICEPS and a scatter plot of the explanatory variables are constructed to obtain an overview of the data.

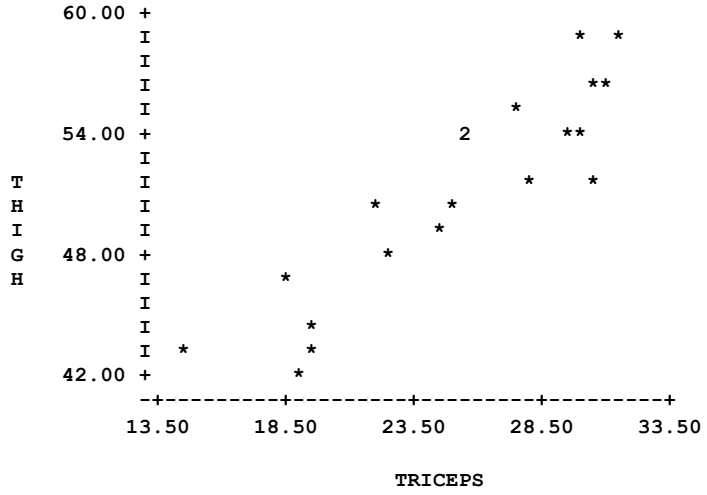
-->PLOT BODYFAT, THIGH



-->PLOT BODYFAT, TRICEPS




-->PLOT THIGH, TRICEPS




The plot of THIGH versus TRICEPS shows that a discernible linear relationship is visible between the explanatory variables. Hence a strong linear relationship between one of these explanatory variables and BODYFAT will also be present with the other variable and BODYFAT. In fact, such is indeed the case.

If we carefully examine the scatter plot of THIGH versus TRICEPS, we may observe two possible outlying, or influential, observations. The point nearest the two axes with TRICEPS = 14.6 and THIGH = 42.7 (subject 15) is clearly apart from the other points. This may be an indication that a linear relationship may not exist at low levels. Less obvious, the point with TRICEPS = 30.7 and THIGH = 51.9 (subject 3, near the top of the points) seems to "stick out" from the remaining subjects. We will regress BODYFAT on THIGH and TRICEPS and observe closely the diagnostic statistics associated with these observations.

-->REGRESS BODYFAT, THIGH, TRICEPS. DIAGNOSTICS ARE FULL. @
--> HOLD RESIDUAL(RESID), FITTED (FATFIT)

**REGRESSION ANALYSIS FOR THE VARIABLE BODYFAT**

| PREDICTOR | COEFFICIENT | STD. ERROR | T-VALUE |
|---|---|---|---|
| INTERCEPT | -19.17425 | 8.36064 | -2.29 |
| THIGH | .65942 | .29119 | 2.26 |
| TRICEPS | .22235 | .30344 | .73 |

**CORRELATION MATRIX OF REGRESSION COEFFICIENTS**

| | THIGH | TRICEPS |
|---|---|---|
| THIGH | 1.00 | |
| TRICEPS | -.92 | 1.00 |

**S = 2.5432 R**2 = 77.8% R**2(ADJ) = 75.2%**

**ANALYSIS OF VARIANCE TABLE**

| SOURCE | SUM OF SQUARES | DF | MEAN SQUARE | F-RATIO |
|---|---|---|---|---|
| REGRESSION | 385.439 | 2 | 192.719 | 29.797 |
| RESIDUAL | 109.951 | 17 | 6.468 | |
| ADJ. TOTAL | 495.390 | 19 | | |

| SOURCE | SEQUENTIAL SS | DF | MEAN SQUARE | F-RATIO |
|---|---|---|---|---|
| THIGH | 381.966 | 1 | 381.966 | 59.057 |
| TRICEPS | 3.473 | 1 | 3.473 | .537 |

**DURBIN-WATSON STATISTIC = 2.36**

**DIAGNOSTIC STATISTICS:**

| CASE NO. | OBSERVED VALUE | RESIDUAL | STANDARDIZED RESIDUAL | STUDENTIZED DELETED RESIDUAL | COOK'S DISTANCE | LEVERAGE | |
|---|---|---|---|---|---|---|---|
| 1 | 11.9000 | -1.6827 | -.74 | -.73 | .046 | .201 | |
| 2 | 22.8000 | 3.6429 | 1.48 | 1.53 | .045 | .059 | |
| 3 | 18.7000 | -3.1760 | -1.58 | -1.65 | .490 | .372 | X |
| 4 | 20.1000 | -3.1585 | -1.32 | -1.35 | .072 | .111 | |
| 5 | 12.9000 | -.0003 | .00 | .00 | .000 | .248 | |
| 6 | 21.7000 | -.3608 | -.15 | -.15 | .001 | .129 | |
| 7 | 27.1000 | .7162 | .31 | .30 | .006 | .156 | |
| 8 | 25.4000 | 4.0147 | 1.66 | 1.76 | .098 | .096 | |
| 9 | 21.3000 | 2.6551 | 1.11 | 1.12 | .053 | .115 | |
| 10 | 19.3000 | -2.4748 | -1.03 | -1.03 | .044 | .110 | |
| 11 | 25.4000 | .3358 | .14 | .14 | .001 | .120 | |
| 12 | 27.2000 | 2.2255 | .93 | .92 | .035 | .109 | |
| 13 | 11.7000 | -3.9469 | -1.71 | -1.83 | .212 | .178 | |
| 14 | 17.8000 | 3.4475 | 1.47 | 1.52 | .125 | .148 | |
| 15 | 12.8000 | .5706 | .27 | .27 | .013 | .333 | X |
| 16 | 23.9000 | .6423 | .27 | .26 | .002 | .095 | |
| 17 | 22.6000 | -.8509 | -.35 | -.34 | .005 | .106 | |
| 18 | 25.4000 | -.7829 | -.34 | -.33 | .010 | .197 | |
| 19 | 14.8000 | -2.8573 | -1.16 | -1.18 | .032 | .067 | |
| 20 | 21.1000 | 1.0405 | .42 | .41 | .003 | .050 | |

**"X" DENOTES AN OBSERVATION WITH AN INFLUENTIAL INPUT VECTOR**

The inclusion of the DIAGNOSTICS sentence in the REGRESS paragraph provides us with a number of useful statistical measures for the identification of both spurious and influential observations.

## Leverage

An outlying or spurious observation may have little influence on the fitted regression equation. However, any point can be very influential based on its relative position to the other observations used in the fit. These observations should be studied to see if, in addition, they are outliers. One measure of the "importance" of a single observation is the leverage it has on a fit. A large leverage indicates the observation is distant from the center of the remaining observations. As a result the mass of other observation set as a fulcrum for the leverage applied by the single point.

The calculation of the leverage associated with a point is given in Section 9.6.6. In order to establish the "significance" of the leverage value, we may check to see if it is greater than 2p/n where n is the number of observations in the regression and p is the total number of parameters calculated. This rule of thumb is useful in spotting influential points (Neter, Wasserman and Kutner, 1983, page 403).

In the example above, 2p/n = 2(3/20) = .30. Both observations 3 and 15 have leverage values greater than this "cut off" value and have been "flagged" in the display of diagnostic statistics. Although an observation may have great leverage, it does not imply that it is outlying, or spurious. The converse is also true, an outlier does not need to have great leverage. For example, the outlier found in the prior beer data example did not have statistically significant leverage, but it affected fitted results greatly. For this reason we now need to determine how observations 3 and 15 affect the regression equation.

## Cook's distance

An overall measure of the impact of a single observation on the fit of a regression equation is given by Cook's distance. If an observation has a substantial effect on a fit and is determined to be spurious or an outlier, then a decision regarding possible remedial measures is required (see page 409 of Neter, Wasserman and Kutner, 1983, for a discussion).

The SCA System calculates Cook's distance with other diagnostic statistics (see Section 9.6.6 for the method of calculation). This value should be compared with percentage points of the F(p, n-p) distribution (n and p are the same as defined above) to determine its significance.

In the bodyfat example, we will refer to the F(3, 17) distribution. The Cook's distance of observations 3 and 15 are .490 and .013, respectively. Neither are significant at the 5% level. Hence even if these points are found to be outliers, remedial measures are not necessarily required. The Cook's distance associated with observation 5 of the beer data is also not significant at the 5% level of the F(3,15) distribution. We could conclude that no

remedial measures are required, but we have seen the consequence of one such measure (that is, recoding the value of the response from 25 to 35).

## Standardized residual

The residuals of the fitted equation, $Y_i - \hat{Y}_i$ , are usually assumed to approximate a normal or t distribution with a zero mean. If these values are divided by their standard error, they should then be consonant with the standard normal or t distribution.

For each observation, the REGRESS paragraph can display the observed value, Yi, the residual, and the standardized value of the residual (see Section 9.6.6 for the method of calculation). In the REGRESS paragraph, each residual is standardized using an estimate of the standard error based on its leverage and the value s2. Residuals standardized in this manner are also known as studentized residuals. These values can be compared with percentage points of the standard normal or t distribution.

The value of the standardized residual of observation 5 of the beer data, -3.27, is clearly significant. This indicates the observation merits further study, or some remedial measure. Conversely, the standardized residuals of both observations 3 and 15 of the bodyfat data do not imply inconsistency with the standard normal distribution.

## Studentized deleted residual

As a refinement to the standardized (studentized) residual, we can also calculate the residual at the $j^{th}$ observation when the fitted regression is based on all observations except the $j^{th}$ observation. In this manner the individual observation cannot influence the regression. Residual values obtained are appropriately standardized. It should be noted if two or more outliers are almost coincident, this measure may fail to be useful. Hence it is always important to plot the residuals.

The method of calculation for this deleted studentized residual is given in Section 9.6.6. Values are compared with the t(n-p-1) distribution, with n and p as before.

In the beer data example, the t(11) distribution is used. The $5^{th}$ observation is a clear aberration (the value -9.44 is significant at almost all levels) and warrants study. In the bodyfat example, the t(16) distribution is used for reference. No point is significant enough to merit further study.

## 9.3 Specifying a Regression Model

The SCA System affords a number of ways to specify information regarding a regression or a fit of a linear model. This section describes the most frequently used information used.

### 9.3.1 Specifying dependent and independent variables

The basic information required for a regression analysis are the names of the dependent and independent variables. We had three variables in each of the prior examples. In the first example, DELIVERY was regressed on CASE and DISTANCE. In the second example, the variable BODYFAT was the dependent variable and variables THIGH and TRICEPS were regressors, or independent variables.

These variables are easily specified by listing their names immediately after the REGRESS command. The first variable specified is used as the dependent variable. All other variables are used as regressors in the model. Hence

```
REGRESS  VARIABLES ARE DELIVERY, CASES, DISTANCE.
```

or, as we used in abbreviated form,

```
REGRESS  DELIVERY, CASES, DISTANCE.
```

is interpreted as a regression specification of DELIVERY on CASES and DISTANCE. Similarly,

```
REGRESS  VARIABLES ARE BODYFAT, THIGH, TRICEPS.
```

or more simply,

```
REGRESS  BODYFAT, THIGH, TRICEPS.
```

is interpreted as a regression specification of BODYFAT on THIGH and TRICEPS.

### 9.3.2 Including a constant term

Whenever we list the variables involved in a regression, a constant term is also included. This is the default condition used by the SCA System. The constant term is usually important in a regression analysis as we try to determine if more information than mean level alone can be obtained from the dependent variable. If we do not want a constant term in the regression, we need to add the logical sentence NO CONSTANT after the variable specification. That is, if we did not want to include a constant in the beer example, we need to state

```
REGRESS  DELIVERY, CASES, DISTANCE.  NO CONSTANT.
```

## 9.4 A Brief Overview of Linear Regression Analysis

The linear regression model is part of a more general class of linear models. Properties of linear models and regression analysis have been considered by many authors including Draper and Smith (1981), Seber (1977), Neter and Wasserman (1974), Neter, Wasserman, and Kutner (1983), Searle (1971), Daniel and Wood (1980), Graybill (1961), Rao (1973) and

references contained therein. This section will briefly review the linear regression model, the interpretation of some output, and diagnostic checks for a fitted regression model.

### 9.4.1 Linear regression model

The general form of the linear regression model can be written as

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_m X_{mj} + \varepsilon_j, \qquad j = 1, 2, ..., n;$$

where

$Y_j$ is the $j^{th}$ observation (trial, case) of a response, or dependent, variable;

$X_{ij}$ is the $j^{th}$ observation of the $i^{th}$ explanatory, or independent, variable (i.e., a variable whose values are known);

$\beta_0$ , $\beta_1$ , $\beta_2$ , . . . , $\beta_m$ are parameters to be estimated, and

$\varepsilon_j$ is an error term.

The error terms are assumed to be uncorrelated random variables with mean zero and unknown variance, $\sigma^2$ . The estimates for parameters in the above equation, $\hat{\beta}_0, \hat{\beta}_1, ...., \hat{\beta}_m$ , are chosen to minimize the sum of the squared errors, i.e.,

$$SSE = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2$$

where $\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_{1j} + \hat{\beta}_2 X_{2j} + \cdots + \hat{\beta}_m X_{mj}$

A usual assumption is that the error terms follow a normal distribution (i.e., N(0, $\sigma^2$ )). In such a case, the estimates for the parameters are also maximum likelihood estimates. Note that in this chapter we use p to indicate the number of parameters to be estimated. We observe that p=m+1 if a constant is included in the model (i.e., $\beta_0$ ) and p=m otherwise.

### 9.4.2 Interpreting SCA output

The SCA System generates and displays important information regarding a regression. This information can be used in several contexts, including inference and prediction. It is important to note that the validity of the estimates of the regression equation, and any inference or prediction made from a regression, is based on the data at hand and the validity of the model being fit. Hence it is important to carefully check any model for outliers or spurious observations and for deviations from the assumptions of the model. This is discussed briefly in Section 9.4.3 and in Section 9.5.

To illustrate the use of SCA output for inference and prediction, we will consider the output of our initial example. The fitted equation derived from the beer data is

DELIVERY = 2.311 + .877 CASES + .456 DISTANCE

As a result, if we want to predict the value of DELIVERY for observation number 1 (CASES = 10, DISTANCE = 30) we would use the above equation and obtain, approximately,

DELIVERY = 2.311 + .877(10) + .456(30) = 24.76 .

We may also wish to predict the value of DELIVERY at other plausible combinations of values for CASES and DISTANCE that are not part of our sample. For example, if we wish to predict a value of DELIVERY for CASES = 20 and DISTANCE = 30, we would use the fitted equation and obtain

DELIVERY = 2.311 + .877(20) + .456(30) = 33.53

We should realize that if we were to obtain another sample of 15 observations, it is likely the fitted equation will change. We could still predict DELIVERY for the two combinations of CASES and DISTANCE illustrated above and obtain different predictions in each instance. Thus, it may be important to know, based on our sample:

(1) The amount of variation present in our prediction equation
(2) The significance of the estimated parameters of the model
(3) How much "explanatory power" our fitted equation provides

## Estimate of the variation of the error terms

A prediction, and any other inference, is based on our sample, or the "information at hand". Hence it is important to have some measure of uncertainty (or variation). In examining the linear regression model, we see a key uncertainty is the variability of what is not observed, the error term. The smaller $\sigma^2$ , in relation to the unit of measurement of Y, the more precise our prediction of Y for values of $X_1$ , $X_2$ , . . ., $X_m$.

An estimate of $\sigma$, the standard deviation of the error terms, is calculated from the data. This value, denoted by s, is computed according to

$$s = \sqrt{\frac{SSE}{n - p}}$$

where SSE is the sum of squared errors, n is the number of observations, and p is the number of parameters estimated. SSE and (n - p) are displayed in the analysis of variance table on the line labeled RESIDUAL. We see in the initial fit of the beer data

$s^2$ = mean square error = 118.375/12 = 9.865 ,

so that $s = (9.865)^{1/2} = 3.1408$.

## Deviation of a fitted Y value

If we had included the logical sentence FIT in either of the REGRESS paragraphs of Section 9.1, in addition to the information displayed, we would obtain the following additional output

```
FITTED VALUES AND THEIR STANDARD ERRORS:

 CASE  OBSERVED    FITTED     STD ERR OF
  NO.    VALUE     VALUE   FITTED VALUE  LEVERAGE
   1    24.0000   24.7609     1.3969       .1978
   2    27.0000   26.8673     1.1073       .1243
   3    29.0000   29.3201     1.8736       .3559
   4    31.0000   28.0619     1.5941       .2576
   5    25.0000   34.2716     1.3476       .1841
   6    33.0000   32.2344      .9228       .0863
   7    26.0000   24.6916     1.3436       .1830
   8    28.0000   30.0934     1.1708       .1390
   9    31.0000   29.5682      .8582       .0747
  10    39.0000   38.4788     1.8527       .3480
  11    33.0000   32.4825     1.4161       .2033
  12    30.0000   28.6217      .9641       .0942
  13    25.0000   26.0247     1.1643       .1374
  14    42.0000   39.1135     1.8645       .3524
  15    40.0000   38.4095     1.6079       .2621
```

An estimate of the standard error of fit is provided for each fitted value. This estimate can be used to obtain a confidence interval for the average value of the response for a specific realization of the explanatory variables. The confidence interval is computed from the fitted value, $\hat{Y}$; the estimated standard error of fit; and a value taken from a t-table (or the SCA inverse distribution function, IDFT, shown later). This t-value is based on (n - p) degrees of freedom and the size of the confidence interval we desire. The end points of the interval are

$$\hat{Y} \pm (\text{estimated standard error of fit}) \times (\text{tabled t-value})$$

For the beer data, the tabled t-value for a 95% confidence interval is 2.179. The end points of confidence interval for the average value of TIME for the specific realization CASES = 10, DISTANCE = 30 (observation 1) are

$$24.761 \pm (1.397)(2.179)$$

or

21.717 and 27.805

Hence, given the data, we have a 95% level of confidence that the average time of delivery for all situations in which 10 cases are delivered to a maximum distance of 30 miles is between 21.717 and 27.805 minutes.

## Prediction interval for a single fitted value

The fitted value at a point as calculated above gives us an indication of the average value we could observe for a given realization of values of the explanatory variables. We can also construct a prediction (confidence) interval for the specific values that can occur. The interval is calculated in the same manner as above, except the estimate of standard error is larger. It can be shown this standard error is

$$\sqrt{(\text{estimate of standard error of fitted value})^2 + s^2}$$

Then the end points for a 95% prediction (confidence) interval for the first observation are

$$24.761 \pm 2.179\sqrt{(1.397)^2 + (3.141)^2}$$

or

17.270 and 32.252.

We can compute the end points of this interval using the analytic functions of the SCA System (see Appendix A). The values for the fitted value (24.761), standard error of fitted value (1.397), and s (3.141) may be read directly from SCA output.

The t-value for a 95% (i.e., (1-α) x 100%) confidence interval may be obtained by using the inverse cumulative distribution function of the t-distribution (IDFT). The t-value appropriate for our interval is

IDFT(.975, 12) (i.e., IDFT(1-α/2, n-p))

Hence we can compute the lower and upper end points by sequentially entering

```
-->AMOUNT = IDFT(.975, 12) * SQRT(1.397**2 + 3.141**2)
-->LOWER = 24.761 - AMOUNT
-->UPPER = 24.761 + AMOUNT
```

We can also obtain prediction (confidence) intervals for points not in our sample. This can be done by including additional observations in all explanatory variables of the regression and give the response variable the missing value code. For example suppose we add a 16th observation to the beer sample with CASES = 20 and DISTANCE = 30. If we now use the REGRESS command as before including the FIT sentence, then we will obtain the same results as before with the following additional output.

```
FITTED VALUES AND THEIR STANDARD ERRORS:

 CASE  OBSERVED    FITTED    STD ERR OF
  NO.    VALUE     VALUE   FITTED VALUE  LEVERAGE
   1   24.0000   24.7609     1.3969       .1978
   2   27.0000   26.8673     1.1073       .1243
   3   29.0000   29.3201     1.8736       .3559
   4   31.0000   28.0619     1.5941       .2576
   5   25.0000   34.2716     1.3476       .1841
   6   33.0000   32.2344      .9228       .0863
   7   26.0000   24.6916     1.3436       .1830
   8   28.0000   30.0934     1.1708       .1390
   9   31.0000   29.5682      .8582       .0747
  10   39.0000   38.4788     1.8527       .3480
  11   33.0000   32.4825     1.4161       .2033
  12   30.0000   28.6217      .9641       .0942
  13   25.0000   26.0247     1.1643       .1374
  14   42.0000   39.1135     1.8645       .3524
  15   40.0000   38.4095     1.6079       .2621
  16      *****  33.5329      .9541       .0923
```

We see the fitted value listed for the 16th observation is 33.53, as we calculated before. The end points of 95% prediction interval for this fitted value are

$$33.53 \pm 2.179\sqrt{(.954)^2 + (3.141)^2}$$

or

26.38 and 40.68

Although a prediction and prediction interval can be obtained for any set of values for the explanatory variables, it is important to realize the validity of a prediction is less reliable the further removed we are from the range of values the explanatory variables assume in the regression. That is, although it may be reasonable to predict DELIVERY for CASES = 10 and DISTANCE = 30, it is unreasonable to try to extend a prediction for CASES = 100 or DISTANCE = 75 as these values are far removed from the range of values used to obtain the fitted equation.

## Parameter inference, tests of significance

We can construct tests of significance of the parameters of our model. The test statistic that is used is

$$t = \frac{(\text{estimate}) - (\text{hypothesized value})}{(\text{estimated standard deviation of estimate})}$$

This statistic is then compared with a critical value of the t-distribution with (n-p) degrees of freedom.

The t-value displayed by the SCA System is the value associated with a test of "parameter = 0". In the beer data example, the t-values for both of the estimates associated with CASES and DISTANCE are significant at the 1% level. Hence these estimates are

statistically different from zero. However, the hypothesis that the intercept is zero cannot be rejected at the 5% level.

We can also use displayed information for tests of other specific values. For example, to test the hypothesis that the coefficient of DISTANCE is .5 against the alternative it is not, we compute

$$t = \frac{.45592 - .5}{.14676} = -.3004$$

|t| = .3004 is not significant at the 5% significance level, so the hypothesis cannot be rejected at this level.

We can compute the significance level of a t-value in the SCA System by employing the cumulative distribution function of the t distribution (CDFT). The CDF is an analytic function in the SCA System (see Appendix A). The significance level may be found by using the analytic expression

```
2*(1-CDFT(|t|, df))
```

In this case, we have |t| = .3004 and df = 12. Thus we can use

```
2*(1-CDFT(.3004, 12))
```

and we find the actual significance level is approximately .77.

## Amount of variation explained

A measure of how well a regression model "explains" a response variable is in the amount of the variability of the response variable that can be attributed to the linear model. This value, $R^2$, can be calculated as

$$R^2 = \frac{\text{(Sum of squares due to regression)}}{\text{(Total sum of squares, adjusted for the mean)}}$$

These quantities are all displayed by the SCA System. For the first use of regression in the beer example, we had

$R^2 = 331.359/449.733 = .7368 \approx 73.7\%$

The $R^2$ value is sometimes used as a criterion in choosing the most appropriate regression model from among subsets of possible explanatory variables. Since the $R^2$ value above does not account for the number of parameters present in a model, it is useful to adjust the value for the number of parameters. This value, $R_a^2$, is calculated as

$$R_a^2 = 1 - \left[\frac{n-1}{n-p}\right]\left[\frac{\text{sum of squares due to error}}{\text{total sum of squares, adjusted}}\right]$$

In the beer example we have

$$R_a^2 = 1 - [(15 - 1)/(15 - 3)][\ 118.375/449.733\ ] = .6929 \_ 69.3\%$$

This value is displayed as R**2(ADJ).

### 9.4.3 Diagnostic checks of a fitted model

A careful regression analysis includes more than the specification and estimation of a regression model. A model should be checked carefully to determine if there are any model inadequacies or deviations from the assumptions of the model. The REGRESS paragraph can calculate and display several statistics that are useful in a diagnostic check of a model. In addition, the residuals from a fit, that is, the variable consisting of the values

$$e_j = Y_j - \hat{Y}_j \qquad j = 1, 2, ...., n.$$

can be retained in the SCA workspace for analysis. The analysis of residuals includes, but is not limited to, various plots of residuals and the examination of statistics of the residuals to ascertain if they are consonant with postulated assumptions of the error structure. This section reviews useful diagnostic checks that are readily available within the SCA System. A more complete discussion of these checks can be found in Draper and Smith (1981, Chapter 3) and Neter, Wasserman and Kutner (1983, Chapter 4).

**Residual plots**

The following are useful plots of residuals and should be included, when appropriate, in a regression analysis. Also included are the names of the SCA paragraphs that can be used to create the plot.

(a) Plot against fitted values, $\hat{Y}$ (PLOT): Plots of residuals against $\hat{Y}$ can help reveal non-homogenous variance (a variance that increases with the level of $\hat{Y}$) or model inadequacy. The latter can be due to the need for extra terms in the model (e.g., $X^2$ in addition to X to include quadratic terms) or for a transformation of Y before the analysis (e.g., log or square root).

(b) Plot against explanatory variables (PLOT): The plots here can help to reveal similar anomalies as in (b) above. However, these may be useful in determining specific explanatory variables that could be involved.

(c) Plots against variables not used in model (PLOT): Plotting residuals against variables excluded from a model could reveal the presence of important explanatory variables that should be included in the analysis (see Neter, Wasserman and Kutner, 1983, page 120).

(d) Probability plot of residuals (PPLOT): This is a useful visual check of the residuals. If the assumption of normality is valid, a normal or half-normal plot of residuals should yield an approximate straight line with no point too far apart from the rest.

(e) Time series plot (TSPLOT): Whenever observations are recorded in time order, it is important to plot data over time. This can reveal a variance that is not constant over time, or the presence of linear or quadratic terms in time that should have been included in the model. A plot over time is also useful in observing "runs" of positive or negative residual terms.

(f) Simple plot of residuals (HISTOGRAM or DPLOT): This is useful as a visual check of the normality assumption and to spot potential outlying or spurious observations.

### Statistics of residuals or fit

The SCA System can also calculate and display useful diagnostic statistics of a regression or the residuals of a regression. Listed below is a summary of useful diagnostic statistic and how they may be obtained in the SCA System:

(a) Leverage, Cook's distance, standardized residuals, studentized deleted residuals (DIAGNOSTICS sentence): These are useful in the identification of spurious and influential observations. See Section 9.2 for a discussion and Section 9.6.6 for specifics regarding calculation.

(b) Checks on randomness (DW sentence, ACF and NPAR paragraphs): The Durbin-Watson statistic (DW) can be used to assess the randomness of residuals. DW statistic, the Durbin h statistic, and the alternative autocorrelation function (ACF) are discussed in Section 9.5. The nonparametric RUNS test can also be employed to test the randomness of the residuals.

(c) Tests for normality (NPAR paragraph): The residuals of the fit can be examined by many nonparametric test statistics (see the NPAR paragraph, Chapter 11) to check on "goodness of fit". Possible tests are the Kolmogorov-Smirnov or chi-square test.

## 9.5 A Regression Analysis of Serially Correlated Data

A set of data from the *Commodity Year Book* (1986) will be used to illustrate a business application of regression analysis. The data, listed in Table 3, are comprised of monthly observations, from January 1980 through December 1985, of:

(1) The average wholesale price of gasoline (regular grade, leaded; index = 100 in January 1973)
(2) The average price of crude petroleum at wells (index = 100 in January 1973)

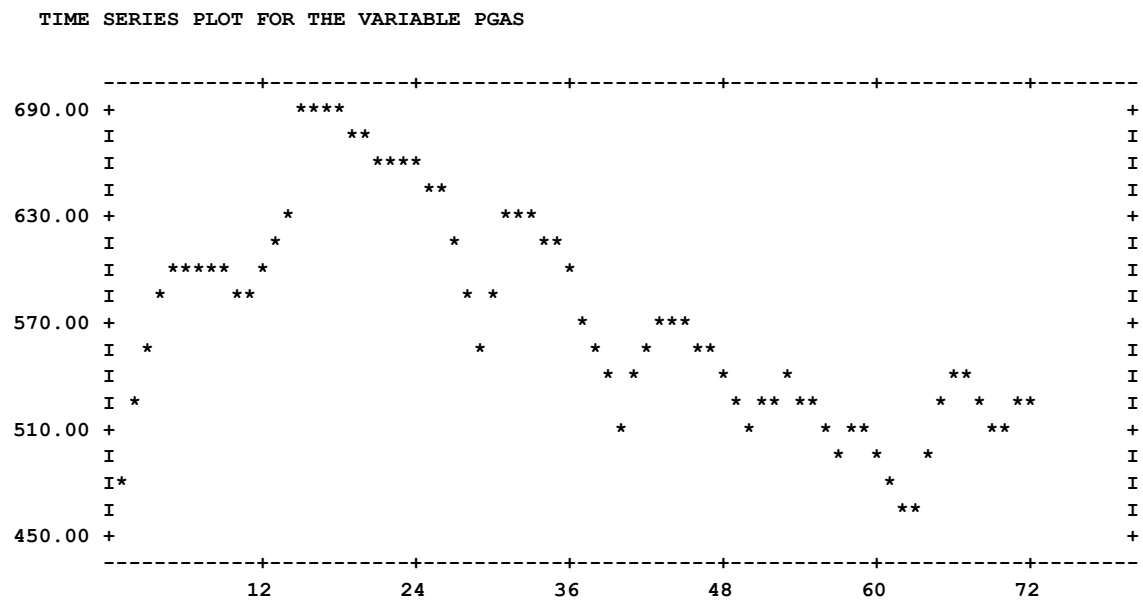
Prices are adjusted for inflation using the base month of January, 1973. The data are stored in the SCA workspace under the names PGAS and PCRUDE respectively.

Table 3 Gasoline data

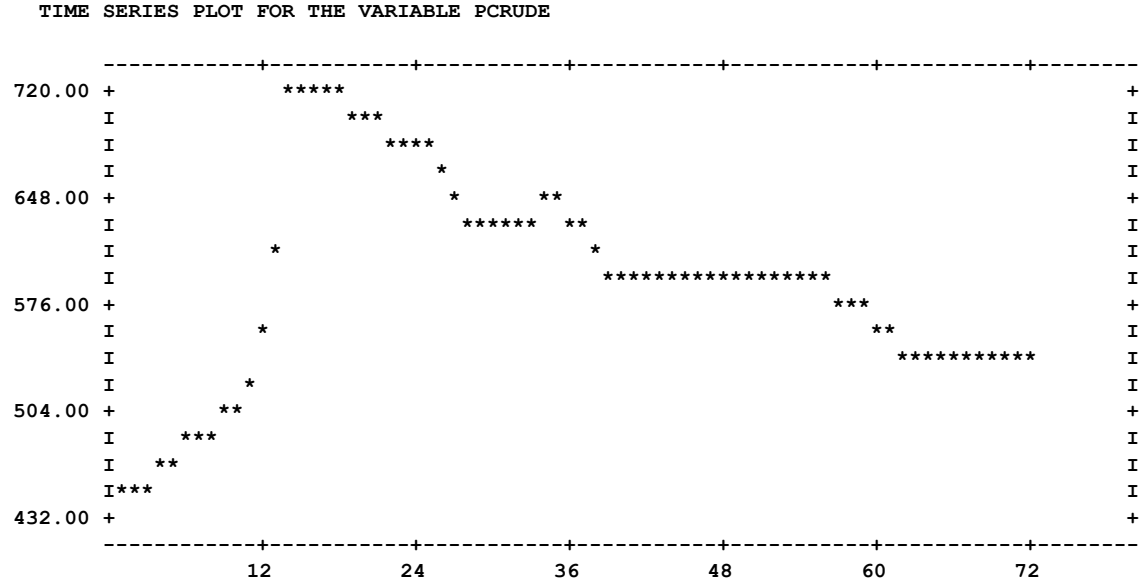
| *Obs.* | *Month* | *Gasoline Price PGAS* | *Crude oil Price PCRUDE* | *Obs.* | *Month* | *Gasoline Price PGAS* | *Crude oil Price PCRUDE* |
|---|---|---|---|---|---|---|---|
| 1 | 1/80 | 481.1 | 447.8 | 37 | 1/83 | 576.7 | 627.5 |
| 2 | 2/80 | 517.5 | 449.1 | 38 | 2/83 | 551.4 | 604.1 |
| 3 | 3/80 | 560.4 | 455.8 | 39 | 3/83 | 533.5 | 591.1 |
| 4 | 4/80 | 585.4 | 465.5 | 40 | 4/83 | 515.3 | 591.1 |
| 5 | 5/80 | 595.5 | 470.9 | 41 | 5/83 | 537.2 | 591.1 |
| 6 | 6/80 | 598.6 | 478.6 | 42 | 6/83 | 559.5 | 591.0 |
| 7 | 7/80 | 601.1 | 480.7 | 43 | 7/83 | 566.6 | 589.1 |
| 8 | 8/80 | 602.9 | 494.2 | 44 | 8/83 | 571.2 | 588.6 |
| 9 | 9/80 | 599.6 | 498.1 | 45 | 9/83 | 566.3 | 589.1 |
| 10 | 10/80 | 591.5 | 505.3 | 46 | 10/83 | 559.2 | 589.1 |
| 11 | 11/80 | 590.8 | 523.6 | 47 | 11/83 | 548.2 | 589.0 |
| 12 | 12/80 | 596.1 | 551.7 | 48 | 12/83 | 535.8 | 588.0 |
| 13 | 1/81 | 607.5 | 614.1 | 49 | 1/84 | 518.3 | 589.0 |
| 14 | 2/81 | 632.9 | 734.7 | 50 | 2/84 | 512.4 | 589.0 |
| 15 | 3/81 | 683.2 | 734.8 | 51 | 3/84 | 517.9 | 589.0 |
| 16 | 4/81 | 694.7 | 734.5 | 52 | 4/84 | 520.5 | 587.5 |
| 17 | 5/81 | 690.4 | 732.3 | 53 | 5/84 | 532.6 | 587.5 |
| 18 | 6/81 | 685.6 | 711.3 | 54 | 6/84 | 531.0 | 587.0 |
| 19 | 7/81 | 677.4 | 696.5 | 55 | 7/84 | 520.9 | 586.4 |
| 20 | 8/81 | 668.4 | 694.7 | 56 | 8/84 | 504.6 | 585.1 |
| 21 | 9/81 | 666.4 | 694.7 | 57 | 9/84 | 500.3 | 584.7 |
| 22 | 10/81 | 666.1 | 687.2 | 58 | 10/84 | 509.8 | 584.0 |
| 23 | 11/81 | 661.7 | 685.2 | 59 | 11/84 | 511.3 | 571.8 |
| 24 | 12/81 | 657.7 | 686.3 | 60 | 12/84 | 502.0 | 566.2 |
| 25 | 1/82 | 651.7 | 686.3 | 61 | 1/85 | 480.5 | 550.3 |
| 26 | 2/82 | 642.3 | 671.6 | 62 | 2/85 | 458.4 | 536.3 |
| 27 | 3/82 | 621.1 | 649.3 | 63 | 3/85 | 467.2 | 536.6 |
| 28 | 4/82 | 578.6 | 625.9 | 64 | 4/85 | 493.9 | 538.4 |
| 29 | 5/82 | 555.7 | 625.8 | 65 | 5/85 | 522.5 | 541.3 |
| 30 | 6/82 | 582.7 | 626.2 | 66 | 6/85 | 535.7 | 540.6 |
| 31 | 7/82 | 628.8 | 626.3 | 67 | 7/85 | 539.3 | 539.6 |
| 32 | 8/82 | 636.3 | 626.3 | 68 | 8/85 | 526.7 | 535.4 |
| 33 | 9/82 | 628.4 | 626.7 | 69 | 9/85 | 513.6 | 536.6 |
| 34 | 10/82 | 617.2 | 641.1 | 70 | 10/85 | 506.1 | 539.2 |
| 35 | 11/82 | 611.0 | 640.0 | 71 | 11/85 | 520.1 | 541.8 |
| 36 | 12/82 | 600.7 | 628.1 | 72 | 12/85 | 523.0 | 544.3 |

An objective of a regression analysis of this set of data could be to forecast gasoline prices based on any relationship that may exist between gasoline and crude oil prices. Since these data are recorded in time, it is useful to plot the data over time. We will use the TSPLOT paragraph (see Chapter 3) for this purpose.

-->TSPLOT PGAS. SEASONALITY IS 12. SYMBOL IS '*'.




-->TSPLOT PCRUDE. SEASONALITY IS 12. SYMBOL IS '*'.



From these plots we observe that PCRUDE appears to lead PGAS. That is, the current value of PCRUDE affects the relative value of PGAS in the next few months. This supports basic beliefs of how these variables may be related. There is still some variability left in PGAS that is not "explained" by PCRUDE. As a result, we may wish to relate the currently observed value of PGAS with some previously observed values of PCRUDE. Before we do this, we may obtain some instructive information by regressing PGAS on (the contemporaneous values of) PCRUDE. That is, we will now use the REGRESS paragraph to obtain the fitted equation

$$PGAS = b_0 + b_1\ PCRUDE \tag{9.1}$$

```
-->REGRESS  PGAS, PCRUDE.  DW.  DIAGNOSTICS ARE BRIEF.  @
-->        HOLD  RESIDUALS(RESID).
```

```
REGRESSION ANALYSIS FOR THE VARIABLE     PGAS

PREDICTOR        COEFFICIENT     STD. ERROR      T-VALUE
INTERCEPT         243.40005       45.54805         5.34
  PCRUDE             .55581         .07657         7.26

S =       46.5025      R**2 =  42.9%      R**2(ADJ) =  42.1%

---------------------------
ANALYSIS OF VARIANCE TABLE
---------------------------

  SOURCE      SUM OF SQUARES     DF     MEAN SQUARE      F-RATIO
REGRESSION        113938.531      1      113938.531       52.689
 RESIDUAL         151373.654     70        2162.481
ADJ. TOTAL        265312.185     71

DURBIN-WATSON STATISTIC =  .13

UNUSUAL OBSERVATIONS:
                                          STUDENTIZED
 CASE    OBSERVED            STANDARDIZED  DELETED     COOK'S
  NO.      VALUE     RESIDUAL  RESIDUAL    RESIDUAL   DISTANCE   LEVERAGE
   1    481.1000   -11.1792      -.25        -.25        .002       .069 X
   2    517.5000    24.4939       .55         .54        .011       .068 X
   3    560.4000    63.6627      1.41        1.42        .067       .063 X
   4    585.4000    83.2839      1.84        1.88        .101       .056 X
   5    595.5000    90.3795      2.00        2.04 *      .111       .053
   6    598.6000    89.1667      1.97        2.01 *      .097       .048
   7    601.1000    90.5038      1.99        2.04 *      .097       .047
  14    632.9000   -18.8538      -.42        -.42        .007       .070 X
  15    683.2000    31.3977       .70         .70        .019       .070 X
  16    694.7000    43.0432       .96         .96        .035       .070 X
  17    690.4000    40.0031       .89         .89        .029       .068 X

"*" DENOTES AN OBSERVATION WITH A LARGE RESIDUAL
"X" DENOTES AN OBSERVATION WITH AN INFLUENTIAL INPUT VECTOR
```

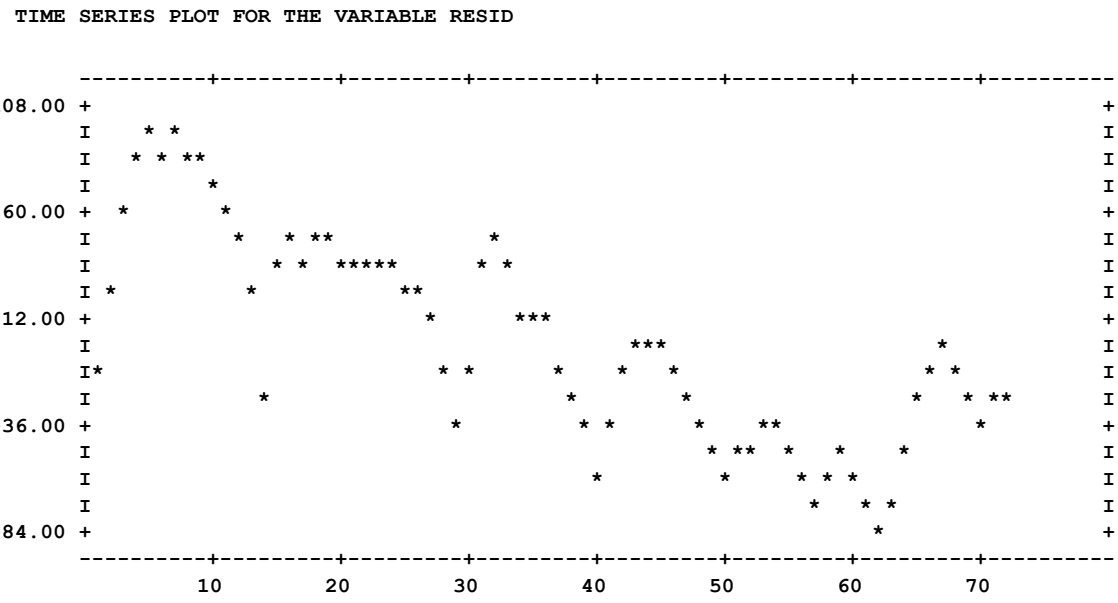
The fitted equation from the above regression is

$$PGAS = 243.40 + .56\ PCRUDE \tag{9.2}$$

All estimates are statistically significant as the absolute value of their t-values are greater than 2 (given the large sample size). Since we have a large data set, we requested a BRIEF set of diagnostic statistics. Here only those cases (observations) with "large" residuals, or leverage are listed. We see that the large residuals are not "that large". We also include the logical sentence DW to obtain the Durbin-Watson statistic (discussed in more detail in Section 9.5.1). Its value, .13, is a warning of first order serial correlation in the residual series. The residual series, representing the actual data values minus the fitted values from the above equation, is a crucial series in diagnostic checks of the model.

As noted previously, the residual series, here maintained in the SCA workspace under the label RESID, should approximate a set of values that are randomly drawn from a normal distribution (also known as a white noise process). However, a distinct pattern is still observed in a plot of the residuals over time.

-->TSPLOT RESID




## 9.5.1 Serial correlation

The error terms of our linear model (see Section 9.4.1) are assumed to be serially uncorrelated. That is, the value of the error associated with one observation should not be related to the value of the error of another observation. If we analyze data that have been recorded over time, it is often the case that this assumption is not true. This is particularly true of business data (as in this example) and of data from industrial experiments that have not been randomized.

If we do not detect the presence of serial correlation, and correct for it, our estimates are inefficient and our analysis can be flawed seriously. For a discussion of the problems that can arise, see Box and Newbold (1971) and Neter, Wasserman, and Kutner (1983, Chapter 13).

We can check for serial correlation in a residual series directly in the SCA System by using the ACF paragraph (see Chapter 10). The ACF paragraph calculates a statistic measuring the correlation present between residual "j" (i.e., $e_j$) and the residual that occurred "$\ell$" periods ahead of it (i.e., $e_{j-\ell}$). The value "$\ell$" is known as the lag. To find if there is lag one autocorrelation in the residual (also called the first order autocorrelation), a statistic is computed from all data one lag apart (i.e., $e_2$ and $e_1$ ; $e_3$ and $e_2$ ; . . . ; $e_n$ and $e_{n-1}$). Similar statistics can be computed from all data two lags apart, three lags apart, and so on. The ACF paragraph can be used to calculate and display the autocorrelation for any number of lags in the residual series. It is useful to observe more than the first order autocorrelation. Autocorrelation of higher lags may also provide us with meaningful information (e.g., a seasonal period). The ACF paragraph will graphically display the calculated values together with a set of 5% significance levels. To obtain the autocorrelation of the above residual series RESID for the first 12 lags, we can simply enter

```
-->ACF   RESID.   MAXLAG IS 12.
```

we obtain

```
TIME PERIOD ANALYZED . . . . . .  1  TO    72
NAME OF THE SERIES . . . . . .        RESID
EFFECTIVE NUMBER OF OBSERVATIONS . . .          72
STANDARD DEVIATION OF THE SERIES . . .     45.8521
MEAN OF THE (DIFFERENCED) SERIES . . .       .0000
STANDARD DEVIATION OF THE MEAN . . . .      5.4037
T-VALUE OF MEAN (AGAINST ZERO) . . . .       .0000


AUTOCORRELATIONS

 1- 12     .93  .81  .71  .63  .57  .54  .51  .49  .48  .48  .48  .46
 ST.E.     .12  .19  .24  .27  .29  .30  .31  .33  .34  .35  .35  .36
  Q       65.2  116  154  185  212  235  257  277  297  316  336  355


        -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
          +----+----+----+----+----+----+----+----+----+----+
                                    I
 1    .93                     +     IXXXXX+XXXXXXXXXXXXXXXXX
 2    .81                  +        IXXXXXXXXX+XXXXXXXXXX
 3    .71                +          IXXXXXXXXXXX+XXXXXX
 4    .63              +            IXXXXXXXXXXXX+XXX
 5    .57             +             IXXXXXXXXXXXXXX
 6    .54            +              IXXXXXXXXXXXXX +
 7    .51            +              IXXXXXXXXXXXXX +
 8    .49           +               IXXXXXXXXXXXX   +
 9    .48           +               IXXXXXXXXXXXX   +
10    .48          +                IXXXXXXXXXXXX    +
11    .48          +                IXXXXXXXXXXXX    +
12    .46         +                 IXXXXXXXXXXX     +
```

A frequently used statistic to assess serial correlation is the Durbin-Watson (DW) statistic. The DW statistic can be used in a test for the presence of a first order autocorrelation in the residual series. Inclusion of the sentence DW will lead to a display of the DW statistic. As noted before. the value of the Durbin-Watson statistic above is .13.

An exact test based on the DW statistic is not always possible. However, tabulated upper and lower bounds for the statistic can be used in one or two tailed tests (see Draper and Smith, 1981, Section 3.11 or Neter, Wasserman and Kutner, 1983, Section 13.3). The DW statistic above is significant at the 1% level. This indicates the presence of serial correlation in the residual series. This conclusion is more apparent by observing the ACF of the residual series. It is worth noting that the DW statistic is approximately equal to $2 - 2r_1$ , where $r_1$ is the lag 1 autocorrelation of the residual series. In the above gasoline example $r_1 = .93$ and $2 - 2(.93) = .14$; the DW value displayed is .13.

The ACF of the residuals adds important information that is missed by the Durbin-Watson statistic. In some situations, the DW statistic may imply there is no correlation present in the residuals. This can be misleading as the DW statistic is based on the first order autocorrelation of the residuals. The ACF provides us with a sequence of autocorrelations. This is particularly important when seasonality are present in the data. Because of the relationship between $r_1$ and the DW statistic, and the fact that more informative statistics can be obtained from the ACF paragraph, it is not recommended the DW statistic be used as the only check for serial correlation.

If autocorrelation is detected, then we may want to modify the regression model by including time dependent variables or correlated error terms in the model. One possible adjustment is the inclusion of lagged terms of the dependent variables as regressors (this will be done in Section 9.5.2). We can also include a serially correlated error term in the model. An example of how to do this is shown in Section 9.7 and Chapter 10. Another possible adjustment is to include a time term as a regressor. This regressor could be as simple as the sequence 1, 2, 3, . . . , n if we are only interested in an inclusion of a simple linear time trend.

### 9.5.2 Use of dynamic regressions

As we anticipated, the model fitted above is inappropriate for the data set. The initial data plots and the checks on the above fit indicate we need to consider incorporating both the leading relationships that may exist between PGAS and PCRUDE as well as the "memory" the data may retain from one month to the next (i.e., serial correlation). Although it is more appropriate to model this data using the time series analysis capabilities of the SCA System (see Section 9.7 and Chapter 10), we will continue in this section by considering a dynamic regression.

For a dynamic regression we will extend the fit of model (9.1) by incorporating variables created by lagging variables of the model. To obtain a better understanding this, we will rewrite (9.1) by explicitly including time subscripts. In this way, (9.1) may also be written as

$$PGAS_t = b_0 + b_1 PCRUDE_t \tag{9.3}$$

Here we see the observed value of PGAS at time " t " is being related (regressed) on the contemporaneous value of PCRUDE (i.e., that occurring during the same month). To allow for the possible leading relationship of PCRUDE, we will want to include the value of PCRUDE at one or more prior months. For example, we can also incorporate the values of PCRUDE of the two preceding periods. That is,

$$PGAS_t = b_0 + b_1 PCRUDE_t + b_2 PCRUDE_{t-1} + b_3 PCRUDE_{t-2} \tag{9.4}$$

To account for the "memory" in the system, we can include past values of PGAS for one or more months. For illustration we will consider one prior period by including

$$b_4 PGAS_{t-1}$$

in the right hand side of the equation (9.4). Hence we will want to obtain the following fitted equation

$$PGAS_t = b_0 + b_1 PCRUDE_t + b_2 pCRUDE_{t-1} + b_3 PCRUDE_{t-2} + b_4 PGAS_{t-1} \tag{9.5}$$

In order to obtain the above fit, we will first create the lagged series for PCRUDE and PGAS. This can be done using the LAG paragraph as shown below.

```
-->LAG  PCRUDE.   NEW ARE PCRUDEL1, PCRUDEL2.   LAGS ARE 1, 2

THE ORIGINAL SERIES IS PCRUDE
THE LAG  1 SERIES IS STORED IN VARIABLE PCRUDEL1,  WHICH HAS   73  ENTRIES
THE LAG  2 SERIES IS STORED IN VARIABLE PCRUDEL2,  WHICH HAS   74  ENTRIES

-->LAG  PGAS.   NEW IS PGASL1.

THE ORIGINAL SERIES IS PGAS
THE LAG  1 SERIES IS STORED IN VARIABLE  PGASL1 ,  WHICH HAS   73  ENTRIES
```

We have now created series with more observations than the original series. PCRUDEL1, representing $PCRUDE_{t-1}$, has 73 observations. The first observation contains the missing value code since we do not know $PCRUDE_0$. The remaining observations are properly "aligned" as in (9.5). The series PCRUDEL2 and PGASL1 (representing $PCRUDE_{t-2}$ and $PGAS_{t-1}$ respectively) have similar properties.

The REGRESS paragraph will automatically handle the case of unequal number of observations per variable and missing data. The SCA System will calculate estimates based on the 69 cases that are completely aligned without missing observations. We will also calculate the Durbin-Watson statistic and retain the residuals of the fit for diagnostic checking.

```
-->REGRESS   PGAS, PCRUDE, PCRUDEL1, PCRUDEL2, PGASL1.     @
-->          DW.  HOLD RESIDUALS(RESID).

REGRESSION ANALYSIS FOR THE VARIABLE     PGAS

PREDICTOR        COEFFICIENT     STD. ERROR      T-VALUE
INTERCEPT          35.91689       17.24426         2.08
  PCRUDE             .07632         .11333          .67
PCRUDEL1             .39770         .18748         2.12
PCRUDEL2            -.49029         .10850        -4.52
  PGASL1             .95301         .03729        25.56

CORRELATION MATRIX OF REGRESSION COEFFICIENTS

  PCRUDE      1.00
PCRUDEL1      -.85     1.00
PCRUDEL2       .52     -.87     1.00
  PGASL1      -.18     -.02      .02     1.00
            PCRUDE PCRUDEL1 PCRUDEL2   PGASL1

S =       14.1175      R**2 =  94.9%      R**2(ADJ) =  94.6%

---------------------------
ANALYSIS OF VARIANCE TABLE
---------------------------

  SOURCE     SUM OF SQUARES    DF    MEAN SQUARE     F-RATIO
REGRESSION       240936.348     4      60234.087     302.224
 RESIDUAL         12954.699    65        199.303
ADJ. TOTAL       253891.047    69

  SOURCE      SEQUENTIAL SS    DF    MEAN SQUARE     F-RATIO
  PCRUDE          103256.147     1     103256.147     518.086
  PCRUDEL1          2559.618     1       2559.618      12.843
  PCRUDEL2          4911.499     1       4911.499      24.643
  PGASL1          130209.082     1     130209.082     653.322

DURBIN-WATSON STATISTIC =  .83
```

The fitted equation from this model is

$$\text{PGAS}_t = 35.92 + .08\ \text{PCRUDE}_t + .40\ \text{PCRUDE}_{t\text{-}1} - .49\ \text{PCRUDE}_{t\text{-}2} + .95\ \text{PGAS}_{t\text{-}1} \qquad (9.6)$$

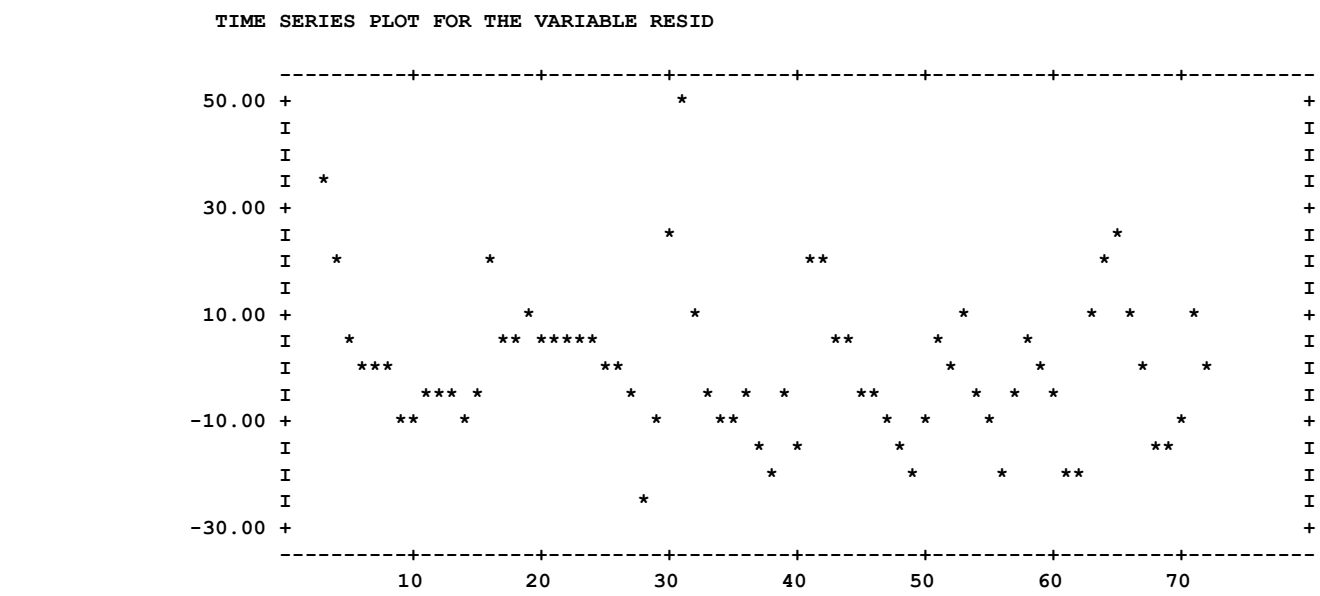
Based on the computed t-values, the estimate of the coefficient of $\text{PCRUDE}_t$ is statistically indistinguishable from 0. Note this differs from the results of the first model considered. The leading relationship of PCRUDE on PGAS is strongly supported. It appears that the inclusion of the lagged observations of PGAS has reduced the Durbin-Watson statistic to insignificance. However, when lagged dependent variables are employed in a regression (here PGASL1), the Durbin-Watson statistic is biased. In these cases we should either examine the autocorrelation structure of the residual series or compute the Durbin h test (Durbin 1970). See Section 9.5.3 for details of this test.

In addition, the estimate of the coefficient of PGASL1 is near 1.0 (and statistically indistinguishable from it). If it was 1.0, then (9.6) could be rewritten as

$$\text{PGAS}_t - \text{PGAS}_{t-1} = 35.92 + .08\ \text{PCRUDE}_t + .40\ \text{PCRUDE}_{t\text{-}1} - .49\ \text{PCRUDE}_{t\text{-}2} \qquad (9.7)$$

As a result, we may wish to model the change in the price of gasoline rather than the observed price. Perhaps because a reformalization of the model is still required, the time plot of the residual series still appears to have a pattern present. The ACF also shows the presence of serial correlation.

-->TSPLOT RESID. SYMBOL IS '*'.




-->ACF RESID. MAXLAG IS 12.

```
          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+

  1    .53                    +     IXXXXX+XXXXXXX
  2    .01                   +      I      +
  3   -.27                   XXXXXXXI      +
  4   -.30                  +XXXXXXXI       +
  5   -.18                  +  XXXXXI       +
  6   -.15                  +   XXXXI       +
  7   -.13                  +    XXXI       +
  8   -.09                  +     XXI       +
  9   -.05                  +      XI       +
 10    .09                  +       IXX     +
 11    .22                  +       IXXXXX  +
 12    .24                 +        IXXXXXX  +
```

### 9.5.3 Durbin h statistic

The Durbin-Watson statistic is biased when the explanatory variables of a regression model include lagged dependent variables. However, an asymptotically valid test of the first-order autocorrelation may be obtained by employing the Durbin h test (Durbin, 1970). When a first-order lagged dependent variable is a regressor, the test statistic to use is

$$DH = r_1 * \sqrt{\frac{n}{1 - n\mathrm{Var}(\hat{\beta}_L)}}$$

where Var ( $\hat{\beta}_L$ ) is the estimated variance of the lagged dependent regressor and $r_1$ is the lag 1 autocorrelation of the residual series. For large n, the standard normal distribution is the reference distribution of this statistic.

A first order lagged dependent variable (PGASL1) was employed as a regressor in our most recent regression. We have n=70 and Var ( $\hat{\beta}_L$ ) = Var (PGASL1) = $(.03729)^2$ , $r_1$ = .53 (see the ACF output following the regression output), hence

$$DH = (.53)\sqrt{\frac{70}{1 - 70(.03729)^2}} = 4.67$$

The value of DH is quite significant with respect to its standard normal reference distribution.

We can compute the Durbin h statistic directly from information retained after the execution of the REGRESS and ACF paragraphs. For example, we can obtain the above DH statistic from the previous analysis if we had done the following

```
-->REGRESS  PGAS, PCRUDE, PCRUDEL1, PCRUDEL2, PGASL1.  @
-->         HOLD RESIDUALS (RESID), INVXPX (XPXINV), MSE (SSQ).

-->ACF  RESID.   MAXLAG IS 12.   HOLD ACF (AUTOC)

-->VBETA = SSQ * XPXINV

-->DH = AUTOC(1) * SQRT(70/(1-70*VBETA(5,5)))
```

The variable XPXINV is designated in the REGRESS paragraph to maintain the matrix $(\mathbf{X'X})^{-1}$, and SSQ=$s^2$. As a result, VBETA is the variance-covariance matrix of the parameter estimates, with VBETA (5,5) the variance of the $5^{th}$ parameter, here PGASL1. The sample ACF are retained in the variable AUTOC, so that r1 = AUTOC(1). The test statistic is then calculated directly.

## 9.6 Other Regression Topics

This section provides a brief overview of topics related to regression analysis or the SCA REGRESS paragraph. Much of the material presented in this section may be considered "advanced". As a consequence this section can be skipped, and selected topics can be referred to as needed. The material presented, and the section containing it are:

| Section | Topics |
|---|---|
| 9.6.1 | ANOVA table |
| 9.6.2 | Fitting Polynomial equations |
| 9.6.3 | Transformations |
| 9.6.4 | Submodel analysis |
| 9.6.5 | Extensions to the linear model<br>• Weighted least squares<br>• Ridge regression<br>• Piecewise setting<br>• Matrix form of regression model |
| 9.6.6 | Computational methods used |

### 9.6.1 ANOVA table

The REGRESS paragraph can provide two separate ANOVA tables, one based on the partial sum of squares, and the other on the sequential sum of squares. Sum of squares entries in the sequential table represent the contribution to the sum of squares as an explanatory variable enters the model in the order listed after the REGRESS command (i.e. in the order specified in the VARIABLE sentence). Thus, the sum of squares entry in the ANOVA table represents the sum of squares for an explanatory variable after accounting for the fact that all variables listed above the entry in the table are in the model. This is the default presentation used for the ANOVA table.

An alternative way to present the ANOVA table is according to partial sum of squares. Sum of squares entries in the partial sum of squares table represent the extra contribution to the sum of squares (see Draper and Smith 1981) when the designated factor is included rather than omitted from the model.

The partial sum of squares table is useful in determining the relative importance of each explanatory variable. The sequential sum of squares table may be useful in tests involving nested hypotheses. For example, suppose a full model contains explanatory variables $X_1, X_2, ...., X_9$ and it is desired to test the importance of subsets of the explanatory variables in the model under certain conditions, such as

(a) including $X_6$ and $X_7$ given $X_1$ through $X_5$ are in the model,
(b) including $X_6$ , $X_7$ and $X_8$ given $X_1$ through $X_5$ are in the model, or
(c) including $X_6$ , $X_7$ , $X_8$ and $X_9$ , given $X_1$ through $X_5$ in the model,

then an appropriate statistic may be computed by adding entries in a sequential ANOVA table in the order $X_1, X_2, ...., X_9$. Obviously the order in which the explanatory variables are specified is important in such analyses.

The sequential table is the default sum of squares presentation used for the ANOVA table. If the alternative presentation is desired, include the sentence.

ANOVA IS PARTIAL.

in the REGRESS paragraph. If we want to see both tables, then we need to include the sentence

ANOVA IS BOTH.

in the REGRESS paragraph.

### 9.6.2 Fitting polynomial equations

The linear model presented in Section 9.4.1 was

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_m X_{mj} + \varepsilon_j .$$

From this representation, as well as the examples presented previously in Sections 9.1, 9.2 and 9.5, it might be inferred that only linear functions can be modeled. However, the "linear" model denotes "linear in the coefficients" of the model. As a result, $X_{1j}$ , $X_{2j}$ and $X_{3j}$ could represent $X_{1j}$ , $(X_{1j})^2$ and $(X_{1j})^3$. This permits us to fit polynomial equations through the REGRESS paragraph.

**Example: Growth rate data**

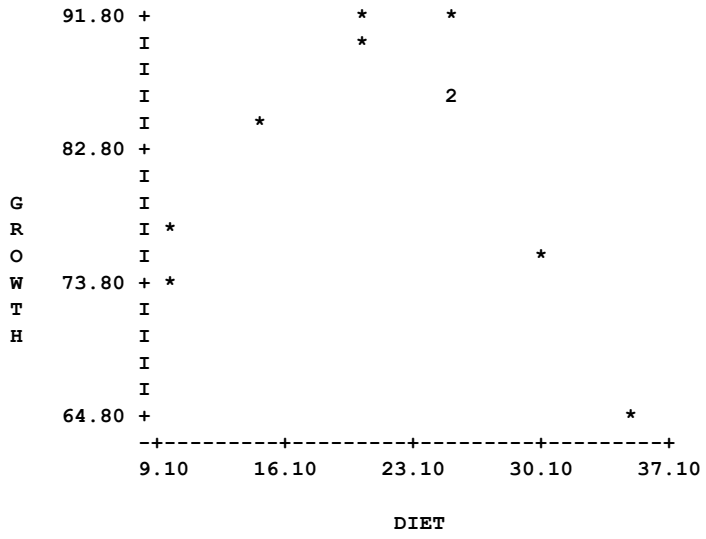
To illustrate fitting a polynomial equation, an example from Box, Hunter and Hunter (1978) is considered. The data to be used are the growth rate (in coded units) for experimental rats and the amount (in grams) of a dietary substance fed to the rats. The data are stored in the SCA workspace under the labels GROWTH and DIET, respectively. The values for the data are given in Table 4.

**Table 4 Growth data**

| *Growth rate GROWTH* | *Dietary Supplement DIET* |
|---|---|
| 73 | 10 |
| 78 | 10 |
| 85 | 15 |
| 90 | 20 |
| 91 | 20 |
| 87 | 25 |
| 86 | 25 |
| 91 | 25 |
| 75 | 30 |
| 65 | 35 |

A scatter plot of GROWTH against DIET reveals a curvature.

**-->PLOT GROWTH, DIET**




The plot indicates a straight line model may not be adequate for the data . We could consider a model of the form

$$\text{GROWTH} = \beta_0 + \beta_1 \text{DIET} + \beta_2 (\text{DIET})^2 + \text{error}$$

We need to define the quadratic term $(\text{DIET})^2$ and then estimate this model. This is done using an analytic capability (See Appendix A) and the REGRESS paragraph.

```
-->DIET2 = DIET**2

-->REGRESS   GROWTH, DIET, DIET2

REGRESSION ANALYSIS FOR THE VARIABLE   GROWTH

PREDICTOR       COEFFICIENT    STD. ERROR     T-VALUE
INTERCEPT         35.65744       5.61793        6.35
    DIET           5.26290        .55802        9.43
   DIET2           -.12767        .01281       -9.97

CORRELATION MATRIX OF REGRESSION COEFFICIENTS

    DIET      1.00
   DIET2      -.98     1.00
              DIET    DIET2

S =        2.5409      R**2 =  93.6%      R**2(ADJ) =  91.8%

---------------------------
ANALYSIS OF VARIANCE TABLE
---------------------------

  SOURCE     SUM OF SQUARES    DF    MEAN SQUARE     F-RATIO
REGRESSION          665.706     2        332.853      51.555
 RESIDUAL            45.194     7          6.456
ADJ. TOTAL          710.900     9

  SOURCE      SEQUENTIAL SS    DF    MEAN SQUARE     F-RATIO
  DIET              24.502     1         24.502       3.795
  DIET2            641.204     1        641.204      99.315
```

## 9.6.3 Transformations

Linear regression analysis is one of the most basic means to build simple relationships between variables. It can be the case that a simple relationship exists between variables, but not in the metric (i.e. the unit of measurement) used to record the data. In such circumstances the assumptions of the linear model may not "quite" fit. Hence the results obtained may not reveal the most informative structure that may exist between variables.

However, if we transform the data, these relationships may be better seen. In addition, a transformation of data may better satisfy the assumptions of the model. Box and Cox (1964) showed how a nonlinear transformation of the response variable can often improve a regression analysis. That is, it may be more informative or appropriate to study a transformation of the response variable (e.g., $Y^{1/2}$, log(Y) or 1/Y) than the variable itself.

Certain rules of thumb are sometimes used in employing a data transformation. For example, if Y has 0 as a natural lower bound, then an assumption of normality of the error term may be more appropriate in an analysis of ln(Y) than of Y. Similarly, ln(Y) is often used as a variance stabilizing transformation or as a means to model the change of a response variable rather than its actual value. Such a transformation may be considered for PGAS in Section 9.5 for the latter reason.

The SCA System permits the inclusion and selection of the transformation parameter in the TWAY and NWAY paragraphs (see Chapter 8) and in the PTRAN and RSM paragraphs (see the document *Quality and Productivity Improvement Using the SCA Statistical System*). These two paragraphs are extensions of the REGRESS paragraph.

### 9.6.4 Modifying a previously specified model and sub-model analyses

Usually a regression analysis consists of a sequence of model fittings for linear models until an "adequate" model is found (see, for example, Chapter 6 of Draper and Smith 1981). The analysis is iterative in nature since either some variables of a previously estimated model are removed or additional variables are added to a model. The System provides capabilities to perform such an iterative analysis efficiently.

One of the most extensive computations in regression analysis is that of computing the **X'X** and **X'Y** matrices of the normal equations. However, the **X'X** and **X'Y** matrices of a sub-model are contained in those of a full model. Thus, the System enables the user to specify the largest model (or a full model) for computing **X'X** and **X'Y** matrices, even though only a portion of the matrices may be used in subsequent analyses.

The System allows the user to perform sub-model analyses from a specified full linear model provided the model is retained in the SCA workspace under a model name. The sentences INCLUDE and EXCLUDE in the REGRESS are used to specify the response and explanatory variables of the sub-model in relation to those variables of the full model. When the INCLUDE sentence is specified, only the variables stated in this sentence that are contained in the full model will be used in the sub-model analysis. On the other hand, when the EXCLUDE sentence is specified, all variables in the full model except those stated in the sentence will be used in the sub-model analysis. Note that the INCLUDE and EXCLUDE sentences must not be used concurrently in a REGRESS paragraph. All variables specified in either of these two sentences must be present in the full model.

The CONSTANT sentence is used to specify, in the present context, whether the constant term should be excluded or included in a sub-model analysis. The specification of NO CONSTANT or CONSTANT in a submodel analysis overrides whatever was assumed about the constant in the full model.

A model name need not be given in the REGRESS paragraph, even if the model specified in the paragraph is a full model that may be used in subsequent analyses. However, if no label (model name) is given to a full model, then we must completely specify the response and explanatory variables of any subsequent analysis as if the sub-model is a completely new model. This complete specification is necessary since, as far as the System is concerned, it is a new model. In such a case possibly redundant calculations will be made.

## Example: Cement data

A set of data from an experiment that studied the effect of the composition of cement on heat evolved during hardening will be used to illustrate submodel analysis. The data used came from Hald (1952). An analyses of this data can be found in Draper and Smith (1981, Chapter 6). Five variables are included:

Y : heat evolved in calories per gram of cement
$X_1$ : amount of tricalcium aluminate
$X_2$ : amount of tricalcium silicate
$X_3$ : amount of tetracalcium alumino ferrite
$X_4$ : amount of dicalcium silicate

$X_1$ , $X_2$ , $X_3$ , and $X_4$ are measured as percent of the weight of the clinkers from which the cement was made. The data are presented in Table 5. The data are stored in the SCA System under the names Y, X1, X2, X3, and X4.

As a first step, Y is regressed on X1, X2, X3, and X4. We will store information associated with the regression in the SCA workspace under the name (label) CEMENT.

**Table 5 Cement data**

| *Experiment* | *Y* | *X1* | *X2* | *X3* | *X4* |
|---|---|---|---|---|---|
| 1 | 78.5 | 7.0 | 26.0 | 6.0 | 60.0 |
| 2 | 74.3 | 1.0 | 29.0 | 16.0 | 52.0 |
| 3 | 104.3 | 11.0 | 56.0 | 8.0 | 20.0 |
| 4 | 87.6 | 11.0 | 31.0 | 8.0 | 47.0 |
| 5 | 96.9 | 7.0 | 52.0 | 5.0 | 33.0 |
| 6 | 109.2 | 11.0 | 55.0 | 9.0 | 22.0 |
| 7 | 102.7 | 3.0 | 71.0 | 17.0 | 6.0 |
| 8 | 72.5 | 1.0 | 31.0 | 22.0 | 44.0 |
| 9 | 93.1 | 2.0 | 54.0 | 18.0 | 22.0 |
| 10 | 115.9 | 21.0 | 47.0 | 4.0 | 26.0 |
| 11 | 83.8 | 1.0 | 40.0 | 23.0 | 34.0 |
| 12 | 113.3 | 11.0 | 66.0 | 9.0 | 12.0 |
| 13 | 109.4 | 10.0 | 68.0 | 8.0 | 12.0 |

-->REGRESS Y, X1, X2, X3, X4. NAME IS CEMENT.

```
REGRESSION ANALYSIS FOR THE VARIABLE        Y

PREDICTOR       COEFFICIENT    STD. ERROR     T-VALUE
INTERCEPT         54.80574      73.05967         .75
      X1           1.59691        .76323        2.09
      X2            .59923        .76096         .79
      X3            .14642        .76407         .19
      X4           -.05999        .74004        -.08


CORRELATION MATRIX OF REGRESSION COEFFICIENTS

      X1      1.00
      X2       .95     1.00
      X3       .99      .97     1.00
      X4       .96     1.00      .97     1.00
                X1       X2       X3       X4

S =        2.4724      R**2 =  98.2%      R**2(ADJ) =  97.3%

---------------------------
ANALYSIS OF VARIANCE TABLE
---------------------------

  SOURCE     SUM OF SQUARES    DF    MEAN SQUARE     F-RATIO
REGRESSION         2668.737     4        667.184     109.143
 RESIDUAL            48.903     8          6.113
ADJ. TOTAL         2717.640    12

  SOURCE      SEQUENTIAL SS    DF    MEAN SQUARE     F-RATIO
  X1               1448.352     1       1448.352     236.933
  X2               1213.252     1       1213.252     198.473
  X3                  7.092     1          7.092       1.160
  X4                   .040     1           .040        .007
```

It appears X1 is a significant explanatory variable while X2, X3 and X4 are not. However, since all coefficients are highly correlated, we may wish to consider submodels. For example, if we wish to regress Y on all variables except X3, we can re-run the regression directly by entering

-->REGRESS NAME IS CEMENT. EXCLUDE X3.

```
REGRESSION ANALYSIS FOR THE VARIABLE        Y

PREDICTOR       COEFFICIENT    STD. ERROR     T-VALUE
INTERCEPT         68.50220      14.31165        4.79
      X1           1.45264        .11840       12.27
      X2            .45847        .18783        2.44
      X4           -.19727        .17536       -1.12

CORRELATION MATRIX OF REGRESSION COEFFICIENTS

      X1      1.00
      X2       .05     1.00
      X4       .10      .97     1.00
                X1       X2       X4

S =        2.3364      R**2 =  98.2%      R**2(ADJ) =  97.6%
```

```
---------------------------
ANALYSIS OF VARIANCE TABLE
---------------------------

  SOURCE     SUM OF SQUARES    DF    MEAN SQUARE     F-RATIO
REGRESSION         2668.512     3        889.504     162.953
 RESIDUAL            49.128     9          5.459
ADJ. TOTAL         2717.640    12

  SOURCE      SEQUENTIAL SS    DF    MEAN SQUARE     F-RATIO
  X1               1448.352     1       1448.352     265.331
  X2               1213.252     1       1213.252     222.262
  X4                  6.908     1          6.908       1.266
```

In a similar fashion, we can regress Y on all variables except X4. We can either exclude X4 or include X1, X2, and X3. For illustrative purposes only, the latter is shown.

-->REGRESS NAME IS CEMENT. INCLUDE X1, X2, X3.

```
REGRESSION ANALYSIS FOR THE VARIABLE        Y

PREDICTOR        COEFFICIENT    STD. ERROR     T-VALUE
INTERCEPT          48.89302       3.92804       12.45
      X1            1.65625        .20369        8.13
      X2             .66080        .04451       14.84
      X3             .20638        .18072        1.14

CORRELATION MATRIX OF REGRESSION COEFFICIENTS

      X1      1.00
      X2      -.18     1.00
      X3       .82     -.06     1.00
                X1       X2       X3

S =        2.3320      R**2 =  98.2%      R**2(ADJ) =  97.6%

---------------------------
ANALYSIS OF VARIANCE TABLE
---------------------------

  SOURCE     SUM OF SQUARES    DF    MEAN SQUARE     F-RATIO
REGRESSION         2668.696     3        889.565     163.578
 RESIDUAL            48.944     9          5.438
ADJ. TOTAL         2717.640    12

  SOURCE      SEQUENTIAL SS    DF    MEAN SQUARE     F-RATIO
  X1               1448.352     1       1448.352     266.330
  X2               1213.252     1       1213.252     223.099
  X3                  7.092     1          7.092       1.304
```

### 9.6.5 Extensions to the linear model

This section briefly discusses the syntax necessary to extend the estimation of a linear model using:

(1) weighted least squares,
(2) ridge regression,
(3) the fitting of a linear model in a piecewise fashion, that is, fits over non-over lapping intervals.

No detailed explanation of the statistical basis for these analyses is given in this manual. The reader is referred to a text such as Draper and Smith (1981) or Neter, Wasserman and Kutner (1983) for an explanation of methods. Fitting a linear model with serially correlated errors is discussed in Section 9.7.

In addition to the above, a discussion of the matrix form of the linear model and specifications in the SCA System will also be given.

#### Weighted least squares

Weighted least squares is useful in those situations where the errors in the linear model do not all have the same variance. In such situations ordinary least squares is inappropriate, but may be used in a modified form if appropriate weights are given to the errors. A "weighted" regression provides those parameter estimates which minimize the weighted sum of squares $\sum w_j(Y_j - \hat{Y}_j)^2$ where $w_j$ , j = 1, . . ., n, is a user supplied set of weights. The sentence WEIGHT is used to specify the name of the variable containing these weights. Note the number of entries for this variable must be the same as that of the dependent variable.

#### Ridge regression

Ridge regression is useful when the correlated cross-product matrix **X'X** is close to singular (i.e., high correlations exist among X's). In such a situation, the parameter estimates are unstable and the constrained solution supplied by ridge regression can be helpful. The SCA System allows the specification of a variable, the vector of the ridge values, consisting of q values, $r_1, r_2, ....., r_q$ , so that **X'X** is adjusted to the matrix (**X'X + R**) where **R** is a diagonal matrix with diagonal elements $r_1, r_2, ....., r_q$ . Note that q = p - 1 if the model has a constant term, and q = p if not. The vector of ridge values is specified in the RIDGE sentence.

## Piecewise fits

Fitting the observations of a linear model in a piecewise fashion (that is, in non-overlapping intervals) may be useful to check if the assumption of uniform variance is valid and if the parameter values of the model remain approximately the same throughout the observational range. The SPAN sentence may be used to specify a range of observational indices over which the linear model is fitted. In order to perform piecewise set, we would repeat our specification of the REGRESS paragraph, but alter the arguments of the SPAN sentence.

## Matrix form of the regression model

The linear model described in Section 9.4.1 can be written in the more general matrix form

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{9.8}$$

where **Y** is an (n x 1) vector of observations,

**X** is an (n x p) matrix of observations or of known form,

**β** is an (p x 1) vector or parameters (regression coefficients), and

**ε** is an (n x 1) vector of errors whose elements are independently and identically distributed with mean zero and variance σ2.

The estimate of **β** that minimizes the sum of the squared error is denoted by $\hat{\beta}$, i.e., the solution of the normal equations (**X'X**) = **X'Y**. If **X** is a full rank matrix of order p, the inverse of **X'X** exists and

$$\hat{\beta} = (X'X)^{-1}X'Y\,.$$

Additionally, if the errors are normally distributed, then $\hat{\beta}$ is the maximum likelihood estimate of **β**. If the rank is less then p, the SCA System provides an error message to draw our attention to the need to reformulate the regression model.

The linear model, equation (9.8) may be written in several forms. If the matrix, **X**, is rewritten in the form of p (n x 1) column vectors, say

$$\mathbf{X} = [\,\mathbf{X}_1\,,\,\mathbf{X}_2\,,\,\ldots\,,\,\mathbf{X}_p\,]\,.$$

then equation (9.8) may be rewritten as

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \times\times\times + \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}\,, \tag{9.9}$$

or

$$\mathbf{Y} = \boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2 + \times\times\times + \boldsymbol{\beta}_p\mathbf{X}_p + \boldsymbol{\varepsilon}. \tag{9.10}$$

The SCA System can accommodate a very general class of linear models; and equations (9.8) through (9.10) also are meaningful, and can be handled by the System, if Y is an (n x m) matrix instead of an (n x 1) vector. In such situations $\beta$ is an (p x m) matrix of parameters instead of a vector. The specification of such a model is based on representation (9.10) and is the same as before.

### 9.6.6 Computational methods in SCA regression

Regression packages differ in the calculations used to estimate the parameters of the linear model and other statistics related to the model. Also the auxiliary statistics and considerations for missing values may differ. In this section we outline the basic calculations used in the REGRESS paragraph.

#### Missing values

The REGRESS paragraph allows missing values and unequal number of cases for dependent and independent variables. However, all incomplete cases in which one or more of the dependent or independent variable values is missing or unavailable are omitted from the analysis.

#### Cross-product matrix and matrix inversion

The cross-products matrix is computed by the provisional mean method stored under double precision. The matrix is pivoted (swept) on all independent variables except those that fail a tolerance test. If the tolerance test is not passed, no further analysis is conducted.

#### Information criterion

In addition to statistics used for an F-test or t-test, the REGRESS paragraph also provides statistics of Akaike's Information Criterion (AIC, Akaike 1973) and Schwarz' Information Criterion (SIC, Schwarz 1978). Each of these statistics has been proposed as a means of comparing models obtained from the same set of data. Models with smaller AIC and SIC are better than those with larger AIC or SIC. The AIC and SIC statistics are computed by the following formulas:

$$\begin{aligned} AIC &= -2 \log_e (\text{maximum of the likelihood function}) + 2m \\ &= n \log_e(2\pi) + n \log_e(SSE/n) + n + 2m \text{ , and} \end{aligned}$$

$$\begin{aligned} SIC &= -2 \log_e (\text{maximum of the likelihood function}) + m \log_e(n) \\ &= n \log_e(2\pi) + n \log_e(SSE/n) + n + m \log_e(n) \end{aligned}$$

where SSE is the sum of squared errors and m is the number of parameters in a regression model (including $\sigma^2$).

## Leverage

The matrix of residuals, **e**, can be expressed as

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ .

The diagonal elements of **H** (i.e., $h_{11}$ , $h_{22}$ , . . . , $h_{nn}$) are useful in determining outliers. The value $h_{jj}$ is called the leverage of the $j^{th}$ observation.

$$0 \le h_{ij} \le 1 \qquad \text{and} \qquad \sum_{j=1}^{n} h_{ij} = p \, .$$

A large leverage value $h_{jj}$ indicates that the j-th observation is distant from the center of the X observations. Hence the mass of other observations act as a fulcrum for the leverage applied by the $j^{th}$ observation.

## Cook's distance

Cook (1977) proposed an overall measure of the impact of an observation on the estimated regression coefficients. This distance measure, $D_j$, can be computed according to

$$D_j = \frac{1}{p}\left\{\frac{e_j^2}{s^2(1-h_{jj})}\right\}\left\{\frac{h_{jj}}{1-h_{jj}}\right\}$$

where $h_{jj}$ is the leverage defined above, and $s^2$ is the estimate of $\sigma^2$ for the full data set.

## Standardized residual

The error components of a linear model, $\varepsilon_j$, is often assumed follow a normal distribution. Thus

$$e_j / s \qquad\qquad j = 1, 2, ....., n$$

are often examined to see if they are consonant with the standard normal distribution. If the variances of the residuals

$$\mathrm{Var}(e_j) = \sigma^2(1-h_{jj})$$

are substantially different, then it is more appropriate to standardize the residuals according to

$$\frac{e_j}{s\sqrt{1-h_{jj}}}$$

with s and $h_{jj}$ defined as above.

The REGRESS paragraph displays the latter standardization as a standardized residuals. These values are also known as studentized residuals.

## Studentized deleted residuals

A refinement to the studentized residual above is to consider the deleted residual

$$d_j = Y_j - \hat{Y}_{(j)}$$

where $\hat{Y}_{(j)}$ is the fitted value for $Y_j$ when parameters are calculated based on all observations except the $j^{th}$ one.

The studentized deleted residual, $d_j^*$ , is a standardization of these values. It is calculated according to

$$d_j^* = e_j\sqrt{\frac{n-p-1}{SSE(1-h_{jj})-e_j^2}}$$

where $h_{jj}$ is the leverage and SSE is the error sum of squares of the full model.

## 9.7 Time Series Regression with Serially Correlated Errors

The REGRESS paragraph permits the estimation of a linear model when serial correlation is present in the error sequence. The effect of serial correlation is discussed in Section 9.5.1. Box and Jenkins (1970) contains a more detailed explanation of considerations for analysis of such data. In order to model in the presence of serially correlated error, the linear model of Section 9.4.1 is rewritten as

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_m X_{mt} + N_t \qquad (9.11)$$

where $N_t$ may be an autoregressive-moving average (ARMA) process (see Chapter 10). That is, instead of the assumption

$$N_t = \varepsilon_t$$

as before, we can hypothesize an error (or disturbance) structure of the form

$$N_t - \phi_1 N_{t-1} - \cdots - \phi_p N_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

in which the $\phi$'s and $\theta$'s may be estimated and $a_t$'s follow a white noise process.

For example, consider the gasoline data modeled in Section 9.5. The regression model

$$PGAS_t = \beta_0 + \beta_1 PCRUDE_t + \varepsilon_t \tag{9.12}$$

was found to have serial correlation. We can allow for a first order autoregressive disturbance if we estimate the model

$$PGAS_t = \beta_0 + \beta_1 PCRUDE_t + N_t \tag{9.13}$$

where the error (disturbance) term follows the relationship

$$N_t - \phi N_{t-1} = a_t . \tag{9.14}$$

(9.14) can be rewritten in terms of backshift operators. The backshift operator is denoted by B and has the property

$$B\ Z_t = Z_{t-1}$$

for a variable, $Z_t$. Hence (9.14) can be rewritten as

$$N_t - \phi B N_t = a_t$$

or

$$(1 - \phi B) N_t = a_t$$

or

$$N_t = \frac{1}{1 - \phi B} a_t \tag{9.15}$$

If we substitute (9.15) into (9.13) we obtain

$$PGAS_t = \beta_0 + \beta_1 PCRUDE_t + \frac{1}{1 - \phi B} a_t \tag{9.16}$$

Equation (9.16) can be specified directly and estimated by the SCA System using the TSMODEL and ESTIM paragraphs, respectively. Details of these paragraphs are found in Chapter 10, but a short discussion related to the above model is given here.

The TSMODEL paragraph requires two sentences. The MODEL sentence is a virtual transcription of the model to be estimated, here (9.16), and the NAME sentence is a label given to maintain information involving the model in the SCA workspace (like the NAME sentence of the REGRESS paragraph). The ESTIM paragraph invokes nonlinear estimation of the model with a label (NAME) given previously. In our example the following could be

used to specify and estimate the model (9.16). Note the name given to the model is MODEL1 and the label NOISE is used to denote at.

```
-->TSMODEL  NAME IS MODEL1.  MODEL IS              @
-->        PGAS = B0 + (B1)PCRUDE + 1/(1-PHI*B)NOISE

   SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- MODEL1

   ------------------------------------------------------------------------
   VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
              VARIABLE   OR CENTERED

     PGAS     RANDOM     ORIGINAL     NONE

    PCRUDE    RANDOM     ORIGINAL     NONE
   ------------------------------------------------------------------------

    PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
      LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

     1    B0              CNST      1       0     NONE      .0000
     2    B1      PCRUDE  NUM.      1       0     NONE      .1000
     3   PHI       PGAS   D-AR      1       1     NONE      .1000


-->ESTIM   MODEL1

   THE FOLLOWING ANALYSIS IS BASED ON TIME SPAN   1  THRU   72

   ITERATION  1, USING STANDARD ERROR =        72.63142241

    ITER.   OBJ.    PARAMETER ESTIMATES
     1   .2034E+05    569.       .980E-01   .931
     2   .1824E+05    375.       .338       .931
     3   .1822E+05    377.       .335       .942
     4   .1822E+05    374.       .339       .942

   ITERATION TERMINATED DUE TO:
   RELATIVE CHANGE IN (OBJECTIVE FUNCTION)**0.5 LESS THAN  .1000D-03

   TOTAL NUMBER OF ITERATIONS . . . . . .          4
   RELATIVE CHANGE IN (OBJECTIVE FUNCTION)**0.5 . . .  .9947D-05
   MAXIMUM RELATIVE CHANGE IN THE ESTIMATES . . . . .  .1278D-01

   REDUCED CORRELATION MATRIX OF PARAMETER ESTIMATES

           1     2     3
     1   1.00
     2   -.90  1.00
     3    .     .    1.00

   THE RECIPROCAL CONDITION VALUE FOR THE CROSS PRODUCT MATRIX OF
   THE PARAMETER PARTIAL DERIVATIVES IS  .531412D-01
```

```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- MODEL1

 ------------------------------------------------------------------------
 VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
           VARIABLE   OR CENTERED

   PGAS     RANDOM     ORIGINAL     NONE

  PCRUDE    RANDOM     ORIGINAL     NONE
 ------------------------------------------------------------------------

  PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
    LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

   1    B0              CNST      1      0     NONE   374.1062   75.1425   4.98
   2    B1      PCRUDE  NUM.      1      0     NONE      .3390     .1094   3.10
   3   PHI       PGAS   D-AR      1      1     NONE      .9417     .0399  23.58


 TOTAL SUM OF SQUARES . . . . . .   .265312E+06
 TOTAL NUMBER OF OBSERVATIONS . . .          72
 RESIDUAL SUM OF SQUARES. . . . . .   .182194E+05
 R-SQUARE . . . . . .          .930
 EFFECTIVE NUMBER OF OBSERVATIONS . .          71
 RESIDUAL VARIANCE ESTIMATE . . . . .   .256611E+03
 RESIDUAL STANDARD ERROR. . . . . .   .160191E+02
```

The fitted equation from this model is

$$\text{PGAS}_t = 374.11 + (.34)\text{PCRUDE}_t + \frac{1}{1-.93\text{B}}\text{a}_t \qquad 9.17)$$

In Section 9.5 we noted that PCRUDE appeared to be a leading indicator of PGAS. We incorporated past, or lagged, values of PCRUDE in the linear model for PGAS in (9.4). As an example of how we can do this in the present discussion, suppose we add a lagged term in (9.16). That is, suppose we want to fit

$$\text{PGAS}_t = \beta_0 + \beta_1\text{PCRUDE}_t + \beta_2\text{PCRUDE}_{t-1} + \frac{1}{1-\phi\text{B}}\text{a}_t \qquad (9.18)$$

We can use the backshift operator to rewrite (9.18) as

$$\text{PGAS}_t = \beta_0 + (\beta_1 + \beta_2\text{B})\text{PCRUDE}_t + \frac{1}{1-\phi\text{B}}\text{a}_t \qquad (9.19)$$

The specification and estimation of this model is similar to that of (9.16). To distinguish this model from MODEL1, we will give it the label MODEL2.

```
-->TSMODEL   NAME IS MODEL2.   MODEL IS          @
-->        PGAS = B0 + (B1 + B2*B)PCRUDE + 1/(1-PHI*B)NOISE

   SUMMARY FOR UNIVARIATE TIME SERIES MODEL --  MODEL2

   ------------------------------------------------------------------------
   VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
              VARIABLE   OR CENTERED

     PGAS     RANDOM     ORIGINAL     NONE

    PCRUDE    RANDOM     ORIGINAL     NONE
   ------------------------------------------------------------------------

    PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
      LABEL       NAME    DENOM.                TRAINT               ERROR  VALUE

     1    B0              CNST      1       0     NONE   374.1062
     2    B1     PCRUDE   NUM.      1       0     NONE      .3390
     3    B2     PCRUDE   NUM.      1       1     NONE      .1000
     4   PHI      PGAS    D-AR      1       1     NONE      .9417

-->ESTIM   MODEL2

   THE FOLLOWING ANALYSIS IS BASED ON TIME SPAN   1  THRU   72


   ITERATION  1, USING STANDARD ERROR =        15.54905444

    ITER.   OBJ.    PARAMETER ESTIMATES
     1   .1362E+05    239.       .134       .422       .946
     2   .1357E+05    226.       .110       .454       .956
     3   .1357E+05    219.       .114       .458       .956

   ITERATION TERMINATED DUE TO:
   RELATIVE CHANGE IN (OBJECTIVE FUNCTION)**0.5 LESS THAN  .1000D-03

   TOTAL NUMBER OF ITERATIONS . . . . . .          3
   RELATIVE CHANGE IN (OBJECTIVE FUNCTION)**0.5 . . .  .5766D-04
   MAXIMUM RELATIVE CHANGE IN THE ESTIMATES . . . . .  .3229D-01


   REDUCED CORRELATION MATRIX OF PARAMETER ESTIMATES

           1     2     3     4
     1   1.00
     2   -.45  1.00
     3   -.44  -.46  1.00
     4    .     .     .    1.00


   THE RECIPROCAL CONDITION VALUE FOR THE CROSS PRODUCT MATRIX OF
   THE PARAMETER PARTIAL DERIVATIVES IS  .531295D-01
```

```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL --  MODEL2

 ------------------------------------------------------------------------
 VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
           VARIABLE   OR CENTERED

   PGAS     RANDOM     ORIGINAL     NONE

  PCRUDE    RANDOM     ORIGINAL     NONE
 ------------------------------------------------------------------------

  PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
    LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

   1    B0              CNST      1      0     NONE   219.0819   77.7405   2.82
   2    B1     PCRUDE   NUM.      1      0     NONE      .1137     .1098   1.04
   3    B2     PCRUDE   NUM.      1      1     NONE      .4584     .1092   4.20
   4   PHI      PGAS    D-AR      1      1     NONE      .9560     .0367  26.03

 TOTAL SUM OF SQUARES . . . . . .   .265312E+06
 TOTAL NUMBER OF OBSERVATIONS . . .          72
 RESIDUAL SUM OF SQUARES. . . . . .   .135684E+05
 R-SQUARE . . . . . .          .947
 EFFECTIVE NUMBER OF OBSERVATIONS . .          70
 RESIDUAL VARIANCE ESTIMATE . . . . .   .193835E+03
 RESIDUAL STANDARD ERROR. . . . . .   .139225E+02
```

The fitted equation from this model is

$$\text{PGAS}_t = 219.08 + (.11 + .46\text{B})\text{PCRUDE}_t + \frac{1}{1 - .96\text{B}}\text{a}_t \qquad (9.20)$$

It may be of interest to compare the above fitted equation with the fitted equation obtained from the dynamic regression that approximates it (that is, equation (9.5) used previously). If we "multiply" all terms of (9.20) by (1-.96B), we obtain

$$(1 - .06\text{B})\text{PGAS}_t = (1 - .96\text{B})219.08 + (1 - .96\text{B})(.11 + .46\text{B})\text{PCRUDE}_t + \text{a}_t \qquad (9.21)$$

Now

$$(1 - .96\text{B})\text{PGAS}_t = \text{PGAS}_t - .96\text{PGAS}_{t-1},$$

$$(1 - .96\text{B})219.08 = 219.08 - 210.32 = 8.76, \text{and}$$

$$(1 - .96\text{B})(.11 + .46\text{B})\text{PCRUDE}_t = (.11 + .46\text{B} - .10\text{B} - .44\text{B}^2)\text{PCRUDE}_t$$

$$= (.11 + .36\text{B} - .44\text{B}^2)\text{PCRUDE}_t$$

$$= .11\ \text{PCRUDE}_t + .36\ \text{PCRUDE}_{t-1} - .44\ \text{PCRUDE}_{t-2}$$

If we substitute the three preceding results into (9.21), and move $.96\ \text{PGAS}_{t-1}$ to the right hand side of the equation, we have the fitted equation

$$\text{PGAS}_t = 8.76 + .11\text{PCRUDE}_t + .36\text{PCRUDE}_{t-1} - .44\text{PCRUDE}_{t-2} + .96\text{PGAS}_t. \qquad (9.22)$$

Equation (9.22) is almost identical to that of (9.6) obtained previously. The only noticeable difference between (9.22) and (9.6) is the constant term of each equation. The reason for this difference is due to the size of the $\phi$ estimate (nearly equal to 1). As a result, the constant term is nearly undefined.

The residuals of the fit of (9.16) or (9.19) can be retained for diagnostic checking purposes if we include the HOLD sentence in the ESTIM paragraph. Note that in the fitted equation for both models, the estimate of the autoregressive parameter, $\phi$, is about 1.0. This is an indication that we should instead model the change in gasoline prices,

$$PGAS_t - PGAS_{t-1} = (1 - B)PGAS_t,$$

instead of the price itself. In addition, the change in crude prices should be used as an explanatory variable. A more complete analysis is presented in Chapter 10.

The above examples were provided to illustrate the estimation of regression models with a first-order autoregressive error structure. Other models can be fit in a similar fashion.

# SUMMARY OF THE SCA PARAGRAPH IN CHAPTER 9

This section provides a summary of the SCA paragraph employed in this chapter. The syntax is presented in both a brief and full form. The brief display of the syntax contains the most frequently used sentences of the paragraph, while the full display presents all possible modifying sentences of the paragraph. In addition, special remarks related to the paragraph may also be presented with the description.

Each SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

In this section, we provide a summary of the REGRESS paragraph.

Legend (see Chapter 2 for further explanation)

v : variable name
i : integer
r : real value
w : keyword

## REGRESS Paragraph

The REGRESS paragraph is used either

(1) to specify and estimate the parameters of a linear model by listing the response (dependent) and explanatory (independent) variables of the model, or

(2) to modify and estimate the parameters of an existing model (see Section 9.6.4).

## Syntax of the REGRESS Paragraph

### Brief syntax

```
REGRESS     VARIABLES ARE v1, v2, ---. @
            DIAGNOSTICS ARE w.      @
            DW. / NO DW.            @
            FIT. / NO FIT.          @
            HOLD RESIDUALS(v1), FITTED(v1).

Required:   List of variables (i.e., VARIABLES sentence)
```

### Full syntax

```
REGRESS     VARIABLES ARE v1, v2, ---.                  @
            NAME IS v.                                  @
            NO CONSTANT. / CONSTANT.                    @
            DIAGNOSTICS ARE w.                          @
            DW. / NO DW.                                @
            FIT. / NO FIT.                              @
            HOLD RESIDUALS (v1,v2,---),                 @
               FITTED(v1,v2,---),                       @
               ESTIMATE(v), INVXPX(v), MSE(v).          @
            SPAN IS i1, i2.                             @
            WEIGHT IS v.                                @
            INCLUDE v1, v2, --- .                       @
            EXCLUDE v1, v2, --- .                       @
            ANOVA IS w.                                 @
            RIDGE IS v.                                 @
            OUTPUT IS LEVEL(w),                         @
                   PRINT(w1, w2, ---), NOPRINT(w).

Required:   List of variables (i.e., VARIABLES sentence)
            or NAME sentence
```

## Sentences Used in the REGRESS Paragraph

**VARIABLES sentence**

A list of variables or the VARIABLES sentence is used to list the dependent and explanatory variables of the regression model. The first variable specified is used as the dependent variable and all other specified variables are used as explanatory variables.

**NAME sentence (see Section 9.6.4)**

The NAME sentence is used to specify a name for the regression model. This is an optional sentence when variables (i.e., the VARIABLES sentence) are specified. If a name is specified, the regression model and related information will be stored under the specified model name and can be used in subsequent analyses. When an existing model is being modified, variable (i.e., the VARIABLES sentence) should not be specified (used).

**NO CONSTANT sentence**

The NO CONSTANT sentence is used to exclude a constant term from an analysis. The default is CONSTANT (that is, include a constant term in the analysis).

**DIAGNOSTICS sentence (see Sections 9.2 and 9.4.3)**

The DIAGNOSTICS sentence is used to specify that diagnostic statistics should be computed and displayed. Valid keywords are FULL and BRIEF. If FULL is specified then the residual, standardized residual, studentized deleted residual, Cook's distance and leverage are computed and displayed for all data points. If BRIEF is specified then the above statistics are displayed for significant values only.

**DW sentence (see Section 9.5.1)**

The DW sentence is used to specify that the Durbin-Watson statistic be computed for the residuals of the model. The default is NO DW, that is, no computation of the statistic.

**FIT sentence (see Section 9.4.2)**

The FIT sentence is used to specify the display of fitted values of the response variable, and associated statistics, for all observations. Also displayed are the standard error of the fitted value and the leverage of the observation. A fitted (predicted) value for points not in the sample can be computed by including additional value(s) in all explanatory variables and the missing value code in the response variable. The default is NO FIT, no display of fitted value information.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. The default is that no values are retained after the paragraph is executed. The values that may be retained are:

| | | |
|---|---|---|
| RESIDUALS | : | The the residuals of the fitted model. The number of variable names specified must be the same as the number of dependent variable columns in the model. |
| FITTED | : | the value for each response variable based on the estimated model. The number of variables specified must be the same as the number of response variable columns in the model. |
| SEFIT | : | the estimated standard error of fit for each fitted value |
| ESTIMATES | : | the complete set of parameter estimates |
| INVXPX | : | the inverse of **X'X** (The product of INVXPX and MSE yields the estimated variance-covariance matrix of the parameter estimates.) |
| MSE | : | the mean square error (matrix) of the model |
| LEVERAGE | : | the leverage of each observation |
| COOK | : | the Cook's distance for each observation |
| SRESID | : | the standardized (studentized) residual value for each observation |
| SDR | : | the studentized deleted residual for each observation |

*The following are infrequently used sentences of the paragraph:*

**SPAN sentence**

The SPAN sentence is used to specify the span of cases, from i1 to i2, of the response variable and corresponding explanatory variables to be used in the analysis. The sentence may be employed for the piecewise fitting of a model (see Section 9.6.5). Default is all observations. The SPAN sentence cannot be used if a model is being re-estimated.

**WEIGHT sentence (see Section 9.6.5)**

The WEIGHT sentence is used to specify a variable containing a weight for each response observation. The default is 1.0 for each observation. The WEIGHT sentence cannot be used if a model is being re-estimated.

**INCLUDE sentence (see Section 9.6.4)**

The INCLUDE sentence is used to modify a previously defined model by specifying those response and explanatory variables to be included in the analysis. Note that the INCLUDE and EXCLUDE sentence are mutually exclusive in the same paragraph.

**EXCLUDE sentence (see Section 9.6.4)**

The EXCLUDE sentence is used to modify a previously defined model by specifying those response or explanatory variables to be excluded from the analysis. Note that the INCLUDE and EXCLUDE sentences are mutually exclusive in the same paragraph.

**ANOVA sentence (see Section 9.6.1)**

The ANOVA sentence is used to obtain different analysis of variance tables. The keyword may be PARTIAL (for partial sum of squares), SEQUENTIAL (for sequential sum of squares), BOTH, or NONE. The default is SEQUENTIAL. The partial sum of squares table shows how each explanatory variable of a regression contributes to the total sum of squares if all other factors in the model are included. the sequential sum of squares table shows the contribution to the total sum of squares of each factor in the regression model, assuming each factor is fitted in the sequential order specified in the VARIABLES sentence.

**RIDGE sentence (see Section 9.6.5)**

The RIDGE sentence is used to specify the name of a vector of q values containing the ridge constants for a ridge regression analysis, where q is the order of the corrected **X'X** matrix (that is the matrix derived using deviations from sample means as entries in the **X** matrix. The corrected **X'X** matrix does not contain elements related to the constant term as each element is subtracted by a mean correction value. Note that q = p-1 if the model has a constant term, and q = p if the model does not have a constant term). The default is 0.0 for all ridge constants, that is, no ridge constraints.

**OUTPUT sentence**

The OUTPUT sentence is used to control the amount of output printed for computed statistics. Control is achieved in a two stage procedure. First a basic LEVEL of output (default NORMAL) is specified. Output may then be increased (decreased) from this level by use of PRINT (NOPRINT).

The keywords for LEVEL and output printed are:

| | | |
|---|---|---|
| BRIEF | : | SUMMARY and ESTIMATES |
| NORMAL | : | SUMMARY, ESTIMATES, and RCORR |
| DETAILED | : | SUMMARY, ESTIMATES, RCORR, CORR, COVAR, and AIC |

where the reserved words (and keywords for PRINT, NOPRINT) on the right denote:

| | | |
|---|---|---|
| SUMMARY | : | the summary of all variables in regression analysis which include sample mean, standard deviation, and coefficient of variation |
| RCORR | : | the correlation matrix for the estimates of the regression coefficients |
| CORR | : | the correlation matrix for all variables in the regression analysis |
| COVAR | : | the covariance matrix for the estimates of the regression coefficients |
| ESTIMATES | : | the estimates of the regression coefficients |
| AIC | : | Akaike's Information Criterion (see Section 9.6.6) and Schwarz' Information Criterion (see Section 9.6.6) |

## REFERENCES


Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." *2nd International Symposium on Information Theory*, ed. B.N. Petrov and F. Csaki, Budapest: Akademiai Kiado 267-281.

Box, G.E.P., and Cox, D.R. (1964). "An Analysis of Transformations." *Journal of the Royal Statistical Society*, B, 26: 211-243.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). *Statistics for Experimenters*, New York: Wiley.

Box, G.E.P., and Jenkins, G.M. (1970). *Time Series Analysis Forecasting and Control*, San Francisco: Holden Day.

Box, G.E.P., and Newbold, P. (1971). "Some Comments on a Paper of Coen, Gomme, and Kendall". *Journal of the Royal Statistical Society*, A, 134: 229-240.

*Commodity Year Book* (1986). New York: Commodity Research Bureau.

Cook, R.D. (1977). "Detection of Influential Observations in Linear Regression." *Technometrics 11*: 15-18.

Daniel, C., and Wood, F.S. (1980). *Fitting Equations to Data*. 2nd edition. New York: Wiley.

Draper, N.R., and Smith, H. (1981). *Applied Regression Analysis*. 2nd edition. New York: Wiley.

Durbin, J. (1970). "Testing for Serial Correlation in Least Squares Regression When Some of the Regressors are Lagged Dependent Variables". *Econometrica 38*: 410-421.

Graybill, F.A. (1961). *An Introduction to Linear Statistical Models*, Vol. 1. New York: McGraw-Hill.

Hald, A. (1952). *Statistical Theory with Engineering Applications*. New York: John Wiley and Sons.

Montgomery, D.C. (1984). *Design and Analysis of Experiments*. 2nd edition. New York: Wiley.

Neter, J., and Wasserman, W. (1974). *Applied Linear Statistical Models*. Homewood, IL: Richard D. Irwin, Inc.

Neter, J., Wasserman, W., and Kutner, M.H. (1983). *Applied Linear Regression Models*. Homewood, IL: Richard D. Irwin, Inc.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. 2nd edition. New York: Wiley.

Schwarz, G. (1978). "Estimating the Dimension of a Model". *Annals of Statistics 5*: 461-464.

Searle, S.R. (1971). *Linear Models*. New York: Wiley.

Seber, G.A.F. (1977). *Linear Regression Analysis*. New York: Wiley.

# CHAPTER 10

# BOX-JENKINS TIME SERIES MODELING AND FORECASTING

In the previous chapter, we observed the inadequacy of regression models in the presence of serial correlation. That is, when a variable maintains a "memory" of its past, any model of the data must incorporate this "memory". This phenomenon is likely to occur whenever data are collected in a time sequence. A set of data generated or obtained sequentially over time is known as a time series.

Time series analysis encompasses both the modeling and forecasting of a time series. There are many different types of models used for time series. One popular class of models has become known as Box-Jenkins ARIMA (autoregressive-integrated moving average) models. These models are popular since they can be interpreted as parsimonious representations of long-lagged autoregressive models, they permit the incorporation of many variables into a model, and a well established procedure for modeling has been developed. Some of the texts and reference sources for these models include Box and Jenkins (1970), Abraham and Ledolter (1983), Pankratz (1983), Vandaele (1983), Granger and Newbold (1987), Cryer (1986), and references contained therein.

## 10.1 Box-Jenkins Modeling

"Box-Jenkins modeling" employs a combination of linear operators for the representation of a time series. This type of representation has a long history, and may be traced to Yule (1921, 1927), Slutsky (1937) and Wold (1938). The landmark contribution of Box and Jenkins (1970) was to both consolidate the models and methodologies that had existed and, more importantly, provide a cohesive framework for model building. Box and Jenkins (1970) proposed an iterative procedure for modeling a time series. This iterative modeling approach encompasses three phases:

(1) Identification, in which we examine characteristics and statistics of a time series and attempt to relate them to those of specific models;

(2) Estimation, in which we estimate the parameters of the tentatively identified model(s) using the data at hand; and

(3) Diagnostic checking, in which we examine the estimated model(s), and residuals of the fit(s), to see if the model(s) make sense and are consonant with our assumptions.

After an appropriate model is determined, we may use it for forecasting or to better understand the structure of the time series. We will first consider two examples to better understand the Box-Jenkins modeling procedure and ARIMA models. Transfer function modeling and forecasting are discussed in the latter part of this Chapter.

### 10.1.1 Example: Series C of Box and Jenkins

As an illustration of time series modeling, we will consider a data set of Box and Jenkins (1970). The data, Series C, consist of 226 temperature readings (one per minute) of an "uncontrolled" chemical process. The data are listed in Table 1, and are stored in the SCA workspace under the name SERIESC.

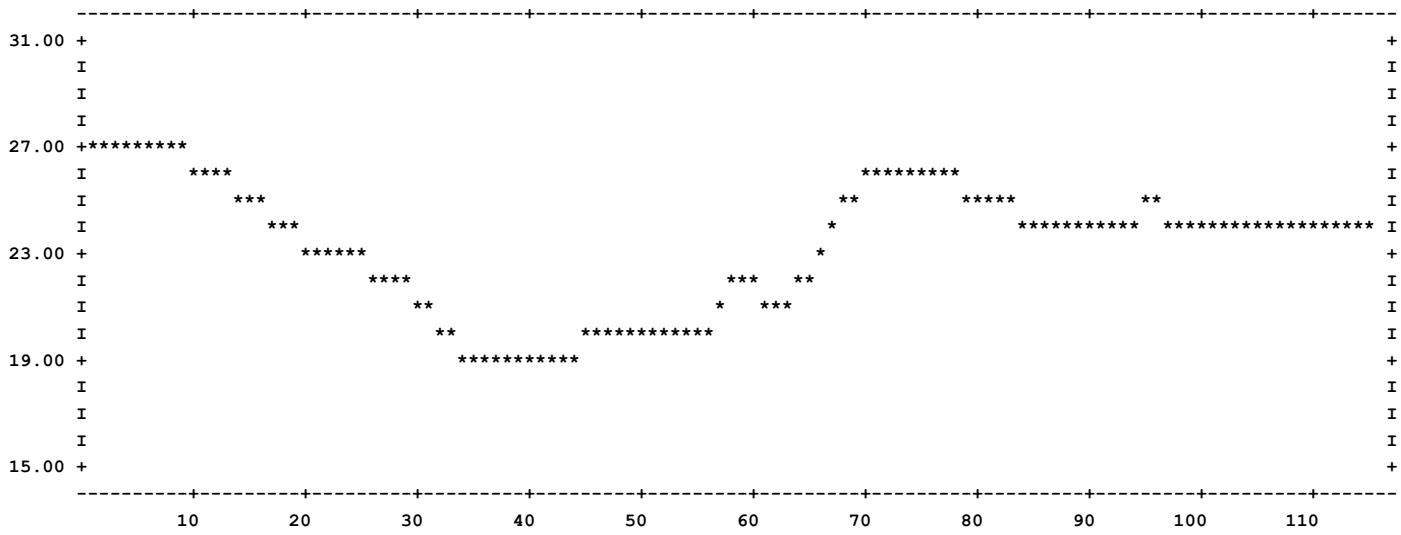
**Table 1 Series C of Box and Jenkins (1970): Temperature readings of a chemical process (Data read across the line)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 26.6 | 27.0 | 27.1 | 27.1 | 27.1 | 27.1 | 26.9 | 26.8 | 26.7 | 26.4 |
| 26.0 | 25.8 | 25.6 | 25.2 | 25.0 | 24.6 | 24.2 | 24.0 | 23.7 | 23.4 |
| 23.1 | 22.9 | 22.8 | 22.7 | 22.6 | 22.4 | 22.2 | 22.0 | 21.8 | 21.4 |
| 20.9 | 20.3 | 19.7 | 19.4 | 19.3 | 19.2 | 19.1 | 19.0 | 18.9 | 18.9 |
| 19.2 | 19.3 | 19.3 | 19.4 | 19.5 | 19.6 | 19.6 | 19.6 | 19.6 | 19.6 |
| 19.7 | 19.9 | 20.0 | 20.1 | 20.2 | 20.3 | 20.6 | 21.6 | 21.9 | 21.7 |
| 21.3 | 21.2 | 21.4 | 21.7 | 22.2 | 23.0 | 23.8 | 24.6 | 25.1 | 25.6 |
| 25.8 | 26.1 | 26.3 | 26.3 | 26.2 | 26.0 | 25.8 | 25.6 | 25.4 | 25.2 |
| 24.9 | 24.7 | 24.5 | 24.4 | 24.4 | 24.4 | 24.4 | 24.4 | 24.3 | 24.4 |
| 24.4 | 24.4 | 24.4 | 24.4 | 24.5 | 24.5 | 24.4 | 24.3 | 24.2 | 24.2 |
| 24.0 | 23.9 | 23.7 | 23.6 | 23.5 | 23.5 | 23.5 | 23.5 | 23.5 | 23.7 |
| 23.8 | 23.8 | 23.9 | 23.9 | 23.8 | 23.7 | 23.6 | 23.4 | 23.2 | 23.0 |
| 22.8 | 22.6 | 22.4 | 22.0 | 21.6 | 21.3 | 21.2 | 21.2 | 21.1 | 21.0 |
| 20.9 | 21.0 | 21.0 | 21.1 | 21.2 | 21.1 | 20.9 | 20.8 | 20.8 | 20.8 |
| 20.8 | 20.9 | 20.8 | 20.8 | 20.7 | 20.7 | 20.8 | 20.9 | 21.2 | 21.4 |
| 21.7 | 21.8 | 21.9 | 22.2 | 22.5 | 22.8 | 23.1 | 23.4 | 23.8 | 24.1 |
| 24.6 | 24.9 | 24.9 | 25.1 | 25.0 | 25.0 | 25.0 | 25.0 | 24.9 | 24.8 |
| 24.7 | 24.6 | 24.5 | 24.5 | 24.5 | 24.5 | 24.5 | 24.5 | 24.5 | 24.4 |
| 24.4 | 24.2 | 24.2 | 24.1 | 24.1 | 24.0 | 24.0 | 24.0 | 23.9 | 23.8 |
| 23.8 | 23.7 | 23.7 | 23.6 | 23.7 | 23.6 | 23.6 | 23.6 | 23.5 | 23.5 |
| 23.4 | 23.3 | 23.3 | 23.3 | 23.4 | 23.4 | 23.3 | 23.2 | 23.3 | 23.3 |
| 23.2 | 23.1 | 22.9 | 22.8 | 22.6 | 22.4 | 22.2 | 21.8 | 21.3 | 20.8 |
| 20.2 | 19.7 | 19.3 | 19.1 | 19.0 | 18.8 | | | | |

The first aspect of a time series analysis, and almost all statistical analyses, is a plot of the data. Here it would be informative if we plot the data as it occurs in time, that is, a time plot. We will use the TSPLOT paragraph (see Chapter 3) for this purpose. Since our data set is relatively "long", we will invoke TSPLOT over two distinct spans.

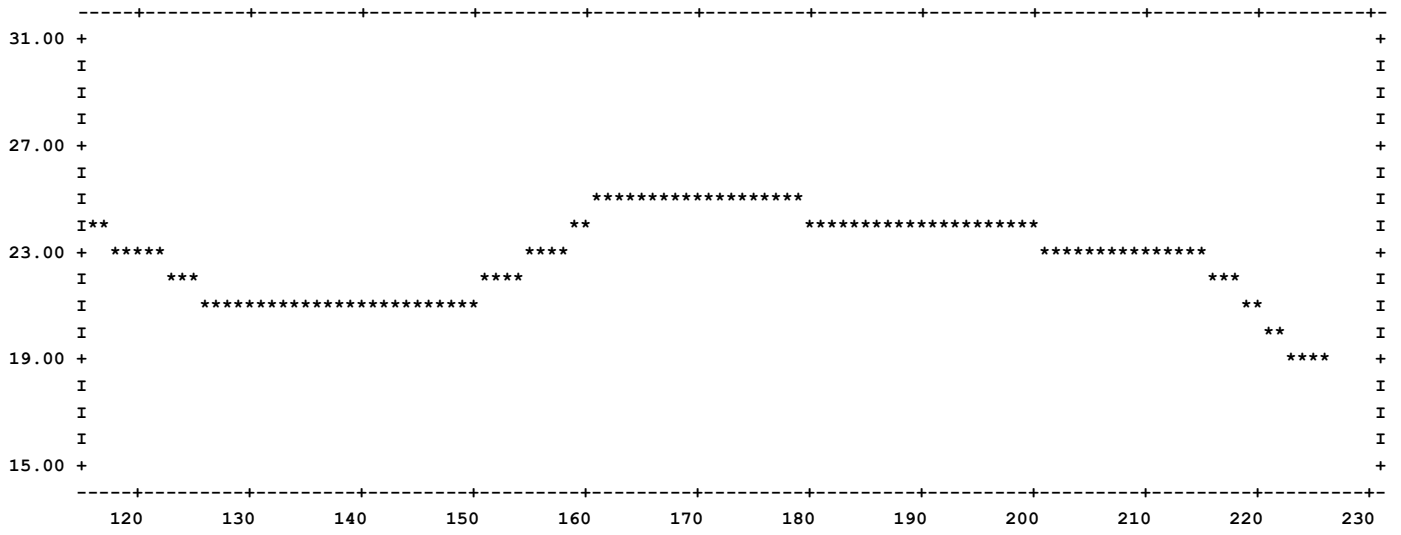
-->TSPLOT SERIESC. SPAN IS 1, 115. SYMBOL IS '*'. RANGE IS 15, 30.

**TIME SERIES PLOT FOR THE VARIABLE SERIESC**




-->TSPLOT SERIESC. SPAN IS 116, 226. SYMBOL IS '*'. RANGE IS 15, 30.

**TIME SERIES PLOT FOR THE VARIABLE SERIESC**




From these plots, we note that the series seems to drift toward a lower level, then it moves up, and then it drifts downwards again. From this plot, it appears that the series does not have a fixed mean level appropriate for all data spans. This is an indication of nonstationary behavior in the time series.

In order to proceed with the identification stage of the analysis, we need to acquire a working knowledge of ARIMA models and notation. If you are familiar with ARIMA models and the backshift operator, you may wish to skip the next section.

### 10.1.2 The univariate ARIMA model

We wish to match the characteristics of our series with those of one or more autoregressive-integrated moving average (ARIMA) models. We have a time series, $Z_t$, t = 1, 2, . . ., n (here n is 226). An autoregressive moving average (ARMA) model has the form

$$Z_t - \phi_1 Z_{t-1} - \phi_2 Z_{t-2} - \cdots - \phi_p Z_{t-p} = C + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \qquad (10.1)$$

where $\{a_t\}$ is a sequence of random variables that are independently and identically distributed with a normal distribution, $N(0, \sigma_a^2)$. If we introduce the backshift operator, B, where

$$BZ_t = Z_{t-1};\ B^2 Z_t = B(BZ_t) = Z_{t-2};\ \text{and so on,}$$

we can rewrite (10.1) as

$$Z_t - \phi_1 BZ_t - \phi_2 B^2 Z_t - \cdots - \phi_p B^p Z_t = C + a_t - \theta_1 Ba_t - \theta_2 B^2 a_t - \cdots - \theta_q B^q a_t \qquad (10.2)$$

or

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) Z_t = C + (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) a_t \qquad (10.3)$$

We can abbreviate (10.3) further by writing it as

$$\phi(B) Z_t = C + \theta(B) a_t . \qquad (10.4)$$

This is known as an ARMA(p,q) model. The value p denotes the order of the auto-regressive operator $\phi(B)$, and q denotes the order of the moving average operator $\theta(B)$. The mathematical properties or requirements of these operators are not discussed here. For a more detailed discussion of these properties see Box and Jenkins (1970).

#### Relationship to a regression model

The ARMA(p,q) model of a series is closely related to a regression model of the series. In Chapter 9 we noted a way to incorporate serially correlation in a model is through a dynamic regression; that is a regression of a series on its own past. We could write such a dynamic regression as (omitting the constant term for notational convenience):

$$Z_t = \pi_1 Z_{t-1} - \pi_2 Z_{t-2} - \pi_3 Z_{t-3} - \cdots + a_t , \qquad (10.5)$$

or, after moving all Z terms to the left-hand side of the equation and employing the backshift operator,

$$\pi(B)Z_t = a_t, \tag{10.6}$$

where

$$\pi(B) = (1 - \pi_1 B - \pi_2 B^2 - \pi_3 B^3 - \cdots). \tag{10.7}$$

There are a large (possibly infinite) number of parameters to estimate here. We may be able greatly reduce the number of parameters if we can approximate $\pi(B)$ as a quotient of polynomials, say $\phi(B)/\theta(B)$ for some choice of p and q. In this manner, we may closely approximate (10.6) as

$$\{\phi(B)/\theta(B)\} Z_t = a_t. \tag{10.8}$$

Multiplication of both sides of (10.8) by $\theta(B)$ yields the ARMA(p,q) model.

If the series is not stationary (i.e., has no fixed mean level), then the autoregressive portion of the ARMA(p,q) model must include a stationary inducing operator. For a series not having seasonality, this is most frequently accomplished through a differencing operator (or product of differencing operators) of the form (1-B). That is, instead of modeling the nonstationary series $Z_t$, we model the series

$$(1 - B)Z_t = Z_t - Z_{t-1}$$

Physically this corresponds to modeling the change in the series rather than the series itself. Usually only a single differencing operator is required. On rare occasions the operator may need to be repeated, a total of d times. The models we then consider is an autoregressive-integrated moving average model of the form

$$\phi(B)(1 - B)^d Z_t = C + \theta(B)a_t. \tag{10.9}$$

This model is also known as an ARIMA(p,d,q) model.

### 10.1.3 Model identification

In the model identification stage, we try to determine "appropriate" orders for p, d, and q in the ARIMA(p,d,q) model. We may not be able to determine a unique set of values, but we may be able to restrict our study to a limited number of models. It may also be the case that not all the autoregressive and moving average parameters of an ARIMA(p,d,q) model are required. For example, if p=3, it may be the case that the lag 2 parameter is zero. We can determine significance during the estimation and diagnostic checking stages.

#### Determining whether or not to difference the data

We have already stated that from its plot, series SERIESC appears to be nonstationary. Hence we expect at least one differencing order. We can confirm this by computing the autocorrelation function (ACF) of the series.

The ACF is a measure of the correlation between a currently observed value and values observed previously. For any positive integer $\ell$, the lag $\ell$ ACF is the correlation between $Z_t$ and $Z_{t-\ell}$. If a series is nonstationary, then its ACF will be high for a number of lags. To compute and display the sample ACF of our series, we may enter

-->ACF SERIESC. MAXLAG IS 24.

```
TIME PERIOD ANALYZED . . . . . .  1  TO   226
NAME OF THE SERIES . . . . . .       SERIESC
EFFECTIVE NUMBER OF OBSERVATIONS . . .          226
STANDARD DEVIATION OF THE SERIES . . .       2.0549
MEAN OF THE (DIFFERENCED) SERIES . . .      22.9739
STANDARD DEVIATION OF THE MEAN . . . .        .1367
T-VALUE OF MEAN (AGAINST ZERO) . . . .     168.0709


AUTOCORRELATIONS

  1- 12     .98  .94  .90  .85  .80  .75  .69  .64  .58  .52  .47  .41
  ST.E.     .07  .11  .14  .17  .19  .20  .21  .22  .23  .24  .24  .25
   Q       219  424  612  781  931 1062 1175 1270 1350 1415 1468 1509

 13- 24     .36  .31  .25  .20  .15  .10  .05 -.00 -.05 -.10 -.15 -.20
  ST.E.     .25  .25  .25  .25  .25  .25  .25  .25  .25  .25  .26  .26
   Q      1540 1563 1578 1588 1594 1596 1597 1597 1597 1600 1606 1616


         -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
           +----+----+----+----+----+----+----+----+----+----+
                                     I
  1    .98                       +  IXX+XXXXXXXXXXXXXXXXXXXX
  2    .94                     +    IXXXXX+XXXXXXXXXXXXXXXXX
  3    .90                    +     IXXXXXX+XXXXXXXXXXXXXXX
  4    .85                   +      IXXXXXXX+XXXXXXXXXXXX
  5    .80                  +       IXXXXXXXX+XXXXXXXXXX
  6    .75                 +        IXXXXXXXXX+XXXXXXXX
  7    .69                 +        IXXXXXXXXX+XXXXXXX
  8    .64                +         IXXXXXXXXXX+XXXXX
  9    .58                +         IXXXXXXXXXX+XXX
 10    .52               +          IXXXXXXXXXXX+X
 11    .47               +          IXXXXXXXXXXXX
 12    .41               +          IXXXXXXXXXX +
 13    .36               +          IXXXXXXXXX  +
 14    .31               +          IXXXXXXXX   +
 15    .25               +          IXXXXXX     +
 16    .20               +          IXXXXX      +
 17    .15               +          IXXXX       +
 18    .10               +          IXX         +
 19    .05               +          IX          +
 20    .00               +          I           +
 21   -.05               +         XI           +
 22   -.10               +        XXXI          +
 23   -.15               +       XXXXI          +
 24   -.20               +       XXXXXI           +
```

We obtain summary information of our data and the display of the ACF for lags 1 through 24. The ACF information is given in two forms. It is listed, together with the standard error of each estimate, and it is plotted. A "Q-value" is also presented in the list of values. We will defer discussion of this statistic until Section 1.5. We note there are several

large values of the ACF before it begins an exponential decay. This behavior and the previous time plot support a decision to difference the series (i.e., to incorporate a d value of at least 1). We will include the differencing operator (1-B) in the remaining modeling of this series.

## Obtaining initial orders for p and q

If our differenced series is stationary we can use its sample ACF and sample partial autocorrelation function (PACF) to determine orders for p and q. We have previously discussed the meaning of the ACF. The PACF is a relative measure of the importance of adding terms in a dynamic regression of a stationary time series. That is, the PACF can be obtained by sequentially fitting

$$Z_t = c + \phi_{11} Z_{t-1} + a_t$$
$$Z_t = c + \phi_{21} Z_{t-1} + \phi_{22} Z_{t-2} + a_t$$
$$Z_t = c + \phi_{31} Z_{t-1} + \phi_{32} Z_{t-2} + \phi_{33} Z_{t-3} + a_t$$
$$\vdots$$

and picking our the estimate of the last term of each fit. The set of parameter estimates of $\phi_{11}, \phi_{22}, \ldots$ is referred to as the sample PACF of the series $Z_t$.

As we may infer from the way that values are computed, the sample PACF provides direct information on the order of autoregressive operator (i.e., p) provided q=0. Alternatively, the ACF provides direct information on the order of the moving average operator (i.e., q) if p=0. More precisely, if we can tentatively identify a pure AR or MA model if we observe the following:

| | **ACF** | **PACF** |
|---|---|---|
| MA(q) | "Cuts off" after lag q | "Dies out" in an exponential or sinusoidal fashion |
| AR(p) | "Dies out" in an exponential or sinusoidal fashion | "Cuts off" after lag p |

By "cut off" we mean that sample estimates of the ACF or PACF are not statistically different from 0 beyond the indicated lag. We judge that an estimate is not significant if it is less (in absolute value) than twice its standard error. We can compute the sample ACF and PACF for the first difference of SERIESC by using the ACF and PACF paragraphs separately, or by simply entering

-->IDEN SERIESC. DFORDER IS 1. MAXLAG IS 12.

The DFORDER sentence specifies the level of differencing we desire (see the note in Section 5.1), and the MAXLAG sentence restricts the number of lags to compute for the sample ACF and PACF to 12 (the default is 36). We obtain the following:

-->IDEN SERIESC. DFORDER IS 1. MAXLAG IS 12.

```
                                          1
DIFFERENCE ORDERS. . . . . . . . . . (1-B  )
TIME PERIOD ANALYZED . . . . . . . .  1  TO   226
NAME OF THE SERIES . . . . . . . . .      SERIESC
EFFECTIVE NUMBER OF OBSERVATIONS . . .        225
STANDARD DEVIATION OF THE SERIES . . .      .2306
MEAN OF THE (DIFFERENCED) SERIES . . .     -.0347
STANDARD DEVIATION OF THE MEAN . . . .      .0154
T-VALUE OF MEAN (AGAINST ZERO) . . . .    -2.2545

AUTOCORRELATIONS

  1- 12     .81  .65  .53  .44  .38  .32  .26  .19  .14  .14  .10  .09
  ST.E.     .07  .10  .12  .13  .13  .14  .14  .14  .15  .15  .15  .15
   Q        148  245  309  354  388  411  427  436  440  445  447  449


         -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
           +----+----+----+----+----+----+----+----+----+----+
                                    I
  1    .81                      +  IXX+XXXXXXXXXXXXXXXXX
  2    .65                    +    IXXXX+XXXXXXXXXXX
  3    .53                   +     IXXXXX+XXXXXXX
  4    .44                   +     IXXXXX+XXXXX
  5    .38                  +      IXXXXXX+XX
  6    .32                  +      IXXXXXX+X
  7    .26                  +      IXXXXXXX
  8    .19                  +      IXXXXX +
  9    .14                  +      IXXX   +
 10    .14                  +      IXXXX  +
 11    .10                  +      IXX    +
 12    .09                  +      IXX    +

PARTIAL AUTOCORRELATIONS

  1- 12     .81  .01 -.01  .05  .03 -.02 -.01 -.08  .02  .12 -.14  .09
  ST.E.     .07  .07  .07  .07  .07  .07  .07  .07  .07  .07  .07  .07

         -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
           +----+----+----+----+----+----+----+----+----+----+
                                    I
  1    .81                      +  IXX+XXXXXXXXXXXXXXXXX
  2    .01                      +  I  +
  3   -.01                      +  I  +
  4    .05                      +  IX +
  5    .03                      +  IX +
  6   -.02                      +  I  +
  7   -.01                      +  I  +
  8   -.08                      +XXI  +
  9    .02                      +  IX +
 10    .12                      +  IXXX
 11   -.14                      XXXI  +
 12    .09                      +  IXX+
```

We see that the PACF cuts off after the first lag and the ACF decays exponentially. These results appear to indicate that an ARMA model with p=1 and q=0 may be appropriate. Hence, we have tentatively identified SERIESC as an ARIMA(1,1,0) model.

## Mixed models

We have relatively simple and effective tools to determine the order of differencing, d, and either p or q, if we have either a pure autoregressive or pure moving average model (after differencing, if necessary). If both p and q are not zero, then the identification of the model can be more difficult if we must rely on the ACF and PACF alone. Box and Jenkins (1970) provide some information on how to determine the orders of p and q from "reading" the sample ACF of a stationary series. However, this approach is usually not very effective in practice.

Tsay and Tiao (1984) have introduced a unified approach to the identification of both the mixed stationary and nonstationary ARMA model. They construct and display a table of values, called the extended autocorrelation function (EACF), to suggest the maximum orders of p and q for an appropriate ARMA(p,q) model. The table of values can be summarized in a condensed form by replacing those values that are within two standard errors of zero by an '0', and by an 'X' otherwise. The order of p and q can then be determined by finding a position $(p_0,q_0)$ in the table so that all values in the table are '0' for the (i,j) coordinates in the triangular region where $i = p_0 + k$, and $j \geq q_0 + k$, $k = 0, 1, 2, \ldots$ .

To illustrate the EACF, we will construct the table for the first difference of SERIESC. To do this, we simply enter

-->EACF SERIESC. DFORDER IS 1.

```
                                          1
DIFFERENCE ORDERS. . . . . . . . . . . (1-B  )
TIME PERIOD ANALYZED . . . . . . . . .  1  TO   226
NAME OF THE SERIES . . . . . . . . . .      SERIESC
EFFECTIVE NUMBER OF OBSERVATIONS . . .          225
STANDARD DEVIATION OF THE SERIES . . .        .2306
MEAN OF THE (DIFFERENCED) SERIES . . .       -.0347
STANDARD DEVIATION OF THE MEAN . . . .        .0154
T-VALUE OF MEAN (AGAINST ZERO) . . . .      -2.2545


THE EXTENDED ACF TABLE

(Q-->)    0    1    2    3    4    5    6    7    8    9   10   11   12
-------------------------------------------------------------------------
(P= 0)  .81  .65  .53  .44  .38  .32  .26  .19  .14  .14  .10  .09  .07
(P= 1)  .01  .01 -.08 -.04  .04  .01  .10 -.00 -.12  .14 -.12  .08 -.08
(P= 2) -.50  .01 -.07 -.07  .06 -.01  .08 -.01 -.07  .04 -.03 -.02 -.05
(P= 3)  .16  .09 -.04 -.05  .03  .03  .05  .01 -.05 -.01 -.04  .01 -.06
(P= 4) -.41  .40 -.26 -.02  .06 -.03  .04  .01 -.06  .02 -.00 -.01 -.00
(P= 5)  .50  .34  .18  .43 -.01  .01  .01  .01 -.08  .04 -.00 -.02  .01
(P= 6) -.16 -.16  .25  .28  .05  .02 -.00  .01 -.06  .02  .02 -.02  .01
```

```
SIMPLIFIED EXTENDED ACF TABLE (5% LEVEL)

(Q-->)  0  1  2  3  4  5  6  7  8  9 10 11 12
-----------------------------------------------
(P= 0)  X  X  X  X  X  X  O  O  O  O  O  O  O
(P= 1)  O  O  O  O  O  O  O  O  O  O  O  O  O
(P= 2)  X  O  O  O  O  O  O  O  O  O  O  O  O
(P= 3)  X  O  O  O  O  O  O  O  O  O  O  O  O
(P= 4)  X  X  X  O  O  O  O  O  O  O  O  O  O
(P= 5)  X  X  X  X  O  O  O  O  O  O  O  O  O
(P= 6)  X  X  X  X  O  O  O  O  O  O  O  O  O
```

We obtain the same summary information as in the previous IDEN output, a table of values of the EACF, and a simplified EACF table. We may observe that a triangular region of '0' values appears to emanate from the vertex where P=1 and Q=0. We have highlighted this region by hand. This confirms our previous conclusion regarding the order of this model.

We noted above that the EACF can be used for nonstationary series as well. To illustrate this, we will compute the EACF for the "undifferenced" SERIESC.

-->EACF SERIESC

```
TIME PERIOD ANALYZED . . . . . .  1  TO   226
NAME OF THE SERIES . . . . . .      SERIESC
EFFECTIVE NUMBER OF OBSERVATIONS . . .    226
STANDARD DEVIATION OF THE SERIES . . .   2.0549
MEAN OF THE (DIFFERENCED) SERIES . . .  22.9739
STANDARD DEVIATION OF THE MEAN . . . .    .1367
T-VALUE OF MEAN (AGAINST ZERO) . . . .  168.0709

THE EXTENDED ACF TABLE

(Q-->)   0    1    2    3    4    5    6    7    8    9   10   11   12
-------------------------------------------------------------------------
(P= 0)  .98  .94  .90  .85  .80  .75  .69  .64  .58  .52  .47  .41  .36
(P= 1)  .81  .66  .55  .48  .43  .38  .34  .28  .25  .25  .22  .22  .20
(P= 2) -.04 -.03 -.12 -.06  .02 -.01  .07 -.04 -.12  .13 -.12  .08 -.08
(P= 3) -.50  .01 -.07 -.11 -.01 -.00  .03 -.03 -.10  .01 -.05 -.03 -.06
(P= 4) -.25 -.27 -.05 -.11 -.01  .03  .00 -.02 -.09 -.01 -.04  .02 -.06
(P= 5) -.48  .28 -.29 -.07  .04 -.05 -.00 -.01 -.08  .07  .00 -.01 -.00
(P= 6) -.08 -.32  .14  .04 -.03 -.04 -.01 -.03 -.08  .07  .00 -.02  .00


SIMPLIFIED EXTENDED ACF TABLE (5% LEVEL)

(Q-->)  0  1  2  3  4  5  6  7  8  9 10 11 12
-----------------------------------------------
(P= 0)  X  X  X  X  X  X  X  X  X  X  X  X  X
(P= 1)  X  X  X  X  X  X  X  O  O  O  O  O  O
(P= 2)  O  O  O  O  O  O  O  O  O  O  O  O  O
(P= 3)  X  O  O  O  O  O  O  O  O  O  O  O  O
(P= 4)  X  X  O  O  O  O  O  O  O  O  O  O  O
(P= 5)  X  X  X  O  O  O  O  O  O  O  O  O  O
(P= 6)  O  X  O  O  O  O  O  O  O  O  O  O  O
```

The initial summary information is the same as that for the ACF of the undifferenced series. Now the triangle of insignificant values appears to emanate from P=2, Q=0. This identification is consistent with our ARIMA(1,1,0) model as the product (1-φB)(1-B) "yields" an AR(2) operator. Hence the EACF, ACF, and PACF can be used to "validate" one another.

Due to sampling fluctuations, the condensed EACF table may not always provide clear cut patterns as shown above. However, it may indicate a few possible candidates for p and q. We should not be concerned by this lack of "uniqueness", since the purpose of the identification stage is to merely suggest a few reasonable models for us to pursue.

### 10.1.4 Model specification and estimation

Now that we have tentatively identified an ARIMA(1,1,0) model as appropriate for our series, we need to estimate the model. This requires two steps. First, we need to specify the model using the TSMODEL paragraph. Once the model is specified, we can estimate the model using the ESTIM paragraph.

We have determined that we will specify a model having a differencing term and one autoregressive parameter. However, should we also include a constant term in the model? Use of a constant term here indicates we believe there may be a trend in the series. Our time plot did not indicate the presence of any definitive trend. We can also examine the summary statistics provided in the IDEN or EACF display of the differenced series. As part of the summary, we are provided with an estimate of the mean of the (differenced) series, its standard error and the associated t-value. This estimate is obtained assuming no serial correlation. We see the t-value here is -2.2545, which may warrant the inclusion of a constant term. However, no constant term is used in our ARIMA (1,1,0) model since the constant term will be insignificant in the final model if it is included. Although we are not including a constant term here, whenever we are in doubt it is often wise to include a constant term. We can then let the data "decide" whether the constant is significant, or not. Omitting a constant term, when one is required, will affect our analysis more than including a constant term when there is no need.

**Model specification**

We want to then specify the following model:

$$(1-\phi B)(1-B)Z_t = a_t . \qquad (10.10)$$

We can specify this model by entering

```
-->TSMODEL   NAME IS CMODEL.   MODEL IS (1-PHI*B)SERIESC((1-B)) = NOISE
```

We need to provide a model in the SCA workspace with a name (label) so that we can refer to it later. Individual names are required since we can maintain more than one model in the workspace in the same SCA session. We have called our model CMODEL. Note that a model name must be distinct from any series name. As a result, we cannot call the model SERIESC, as that is the name of our data. Moreover, if we repeat a model name during an SCA session, the information regarding the newer model completely replaces that of the model that existed previously.

The model specified in the MODEL sentence is a virtual transcription of (10.10), with one exception. The differencing operator (1-B) is specified to the right of our series name, and not on the left as in (10.10). This convention permits the SCA System to distinguish autoregressive operators from "descriptive" operations on the series.

The label PHI used in the specified model is arbitrary. We have chosen it here for convenience. The SCA System permits us to simultaneously maintain and modify many models. Parameter names are used to distinguish and maintain current values of parameters. After we estimate the above model, the estimate of φ will be maintained in the workspace under the label PHI. Since no variable named PHI exists currently, the SCA System will now create one and assign it the initial value 0.10. We see this in the model summary that follows.

```
-->TSMODEL   CMODEL.   MODEL IS (1-PHI*B)SERIESC(1) = NOISE

SUMMARY FOR UNIVARIATE TIME SERIES MODEL --  CMODEL
------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1
SERIESC    RANDOM     ORIGINAL     (1-B  )
------------------------------------------------------------------------
 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT               ERROR  VALUE

  1   PHI    SERIESC    AR       1      1     NONE      .1000
```

We do not always need to be "elaborate" in the specification of a model, as the SCA System only requires information on the order of parameters to be estimated, or differencing operators used. Either of the following can be used to describe the model of (10.10):

```
-->TSMODEL   NAME IS CMODEL.  MODEL IS (1-PHI*B)SERIESC(1) = NOISE.
```
(10.11)

```
-->TSMODEL   NAME IS CMODEL.  MODEL IS (1)SERIESC(1) = NOISE.
```
(10.12)

In (10.11), the differencing operator is reduced to the order of B, that is, 1. If we enter (10.11), the same model summary as given above will occur. In (10.12), we also reduce the autoregressive operator to simply (1). This indicates only a first order term is present in the autoregressive operator. If we enter (10.12) we will obtain the same summary as above, except the parameter estimate will be held internally; that is, no label will be used.

## Model estimation

To estimate the above model we may simply enter

```
-->ESTIM   CMODEL.   HOLD  RESIDUALS(RESIDC).
```

The HOLD sentence is included so that residuals are maintained in the workspace for the purpose of subsequent diagnostic checking. We obtain

```
THE FOLLOWING ANALYSIS IS BASED ON TIME SPAN   1  THRU  226

ITERATION  1, USING STANDARD ERROR =          .21357921
 ITER.   OBJ.    PARAMETER ESTIMATES
  1   .4015E+01    .806
  2   .4014E+01    .813


ITERATION TERMINATED DUE TO:
RELATIVE CHANGE IN (OBJECTIVE FUNCTION)**0.5 LESS THAN  .1000D-03

TOTAL NUMBER OF ITERATIONS . . . . . .          2
RELATIVE CHANGE IN (OBJECTIVE FUNCTION)**0.5 . . .  .7575D-04
MAXIMUM RELATIVE CHANGE IN THE ESTIMATES . . . . .  .8751D-02

THE RECIPROCAL CONDITION VALUE FOR THE CROSS PRODUCT MATRIX OF
THE PARAMETER PARTIAL DERIVATIVES IS  .100000D+01

SUMMARY FOR UNIVARIATE TIME SERIES MODEL --  CMODEL

------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1
SERIESC    RANDOM     ORIGINAL     (1-B  )
------------------------------------------------------------------------

 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT               ERROR  VALUE

  1   PHI    SERIESC    AR       1      1     NONE      .8131     .0383  21.22

TOTAL SUM OF SQUARES . . . . . .   .954336E+03
TOTAL NUMBER OF OBSERVATIONS . . .          226
RESIDUAL SUM OF SQUARES. . . . . .   .401391E+01
R-SQUARE . . . . . .                    .996
EFFECTIVE NUMBER OF OBSERVATIONS . .        224
RESIDUAL VARIANCE ESTIMATE . . . . .   .179192E-01
RESIDUAL STANDARD ERROR. . . . . .   .133863E+00
```

We are provided with a summary of how our parameters change during the estimation process, the reason estimation was terminated, and a summary of the estimated model. We see our estimate of PHI is .8131 with a t-value of 21.22. The t-value indicates that the estimate is clearly significant. The variance of the residuals, that is, the variation in the series that is still not accounted for after our modeling efforts, is .0179. This results in a standard error of about .134. The standard error of our original series (see the ACF summary statistics) is 2.055. Consequently we have accounted for all but $(.134/2.055)^2$, or .4% of the variation of the series. This is reflected in the R-square value of .996.

Note: The high $R^2$ value is misleading since the variation of the modeled series is compared to that of the original series. Since our series is nonstationary, variation is reduced simply by differencing. We can observe that the standard error of the differenced series is .2306 (see the summary statistics on page 10.9 or 10.10). Hence the $R^2$ attributable to differencing is 1 - (.2306/2.055)2 = .987. The $R^2$ related to the differenced data is approximately .66. In ARIMA modeling, $R^2$ is meaningful only if the series is stationary.

### 10.1.5 Diagnostic checks of the model

The final stage of model building is to diagnostically check the model we have estimated. In checking our model(s) we may ask:

(1) Is the model statistically consonant with our assumptions?
(2) Does the model make sense?

The latter is best answered by an individual who "knows" the data. Often when two or more models lead to approximately the same results (e.g., forecasts or explanation of variation), the "best" model may be the one that most closely matches the axioms that apply, if any.

Diagnostic checking model assumptions can be quantified statistically. The most basic assumption made in our model is that the data of the series at are independent and are normally distributed. Such a serially independent series is also referred to as a white noise series. If checks show this assumption is not true, then our model is invalidated and needs to be modified. If the assumption is correct, then the residuals of our model should approximate a serially independent sample and follow a normal distribution with zero mean and constant variance.

We can check our residuals in a number of ways. The most obvious check is a time plot of the residuals.

The plot of the residuals from this fit is not provided here, but no apparent pattern is present. If the residuals approximate white noise, then no autocorrelation should be present. We can check this by computing the ACF of our residual series.

-->ACF RESIDC. MAXLAG IS 12.

```
TIME PERIOD ANALYZED . . . . . .  3  TO   226
NAME OF THE SERIES . . . . . .       RESIDC
EFFECTIVE NUMBER OF OBSERVATIONS . . .          224
STANDARD DEVIATION OF THE SERIES . . .        .1336
MEAN OF THE (DIFFERENCED) SERIES . . .       -.0090
STANDARD DEVIATION OF THE MEAN . . . .        .0089
T-VALUE OF MEAN (AGAINST ZERO) . . . .      -1.0107


AUTOCORRELATIONS

 1- 12     .01  .01 -.05 -.01  .06  .02  .08 -.02 -.09  .13 -.09  .08
 ST.E.     .07  .07  .07  .07  .07  .07  .07  .07  .07  .07  .07  .07
  Q         .0   .0   .7   .7  1.5  1.6  3.0  3.1  4.9  8.9 11.0 12.7


         -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
           +----+----+----+----+----+----+----+----+----+----+
                                     I
 1    .01                        +  I  +
 2    .01                        +  I  +
 3   -.05                        + XI  +
 4   -.01                        +  I  +
 5    .06                        +  IX +
 6    .02                        +  IX +
 7    .08                        +  IXX+
```

```
 8   -.02                      +  I  +
 9   -.09                      +XXI  +
10    .13                      +  IXXX
11   -.09                      +XXI  +
12    .08                      +  IXX+
```

From the summary statistics we see the mean of the residuals is not distinguishable from zero (since the t-value is not significant). In addition, all computed ACF values are within two standard errors of zero. We also are provided with a crude global check on the residuals, a portmanteau test, the Box-Ljung Q statistic (1978). This value, provided in the ACF table in the "Q row", represents a scaled sum of squares of the computed ACF values. It is scaled so that we can use a $\chi^2$ distribution, with $(\ell - p - q)$ degrees of freedom, to determine its significance. The value 12.7 is only marginally significant at the 10% level.

We may also wish to check if we have overfit the series. That is, if some estimates are not statistically different from zero, we may be able to omit them from our model. Here, we have only one parameter in the model, and it is significant, as noted above.

As a final check of the model, we may also wish to test to see if there are any spurious points that may have affected our fit; and if so, how to correct for them. Chang, Tiao and Chen (1988) have proposed a test for outliers in a time series model. This procedure, and a more automatic procedure to iteratively identify and correct for outliers, may be found in *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*.

### 10.1.6 Forecasting an estimated model

Once we have determined that we have an adequate fit, we can forecast the series using the FORECAST paragraph. To forecast SERIESC using our estimated model, we can enter

-->FORECAST CMODEL. NOFS ARE 12.

```
---------------------------------
 12 FORECASTS, BEGINNING AT  226
---------------------------------

 TIME     FORECAST    STD. ERROR    ACTUAL IF KNOWN

  227      18.6374        .1339
  228      18.5051        .2772
  229      18.3976        .4319
  230      18.3102        .5908
  231      18.2391        .7498
  232      18.1813        .9064
  233      18.1343       1.0592
  234      18.0961       1.2073
  235      18.0650       1.3504
  236      18.0398       1.4883
  237      18.0192       1.6212
  238      18.0025       1.7490
```

We are provided with 12 forecasts, together with the standard error of each forecast. The sentence NOFS was included to limit the number of forecasts to 12. If the sentence is omitted, then 24 forecasts are produced.

## 10.2 Modeling the Gasoline Data

As a second example of ARIMA modeling, we will use the gasoline data of Chapter 9. The data, listed in Table 2, are comprised of monthly observations, from January 1980 through December 1985, of:

(1) The average wholesale price of gasoline (regular grade, leaded, index = 100 in January 1973)
(2) The average price of crude petroleum at wells (index = 100 in January 1973)

Prices are adjusted for inflation and are relative to the base month of January, 1973. The data are obtained from the *Commodity Year Book*, (1986). The data are stored in the SCA workspace under the names PGAS and PCRUDE, respectively.

**Table 2 Gasoline data**

| *Obs.* | *Month* | *Gasoline Price PGAS* | *Crude oil Price PCRUDE* | *Obs.* | *Month* | *Gasoline Price PGAS* | *Crude oil Price PCRUDE* |
|---|---|---|---|---|---|---|---|
| 1 | 1/80 | 481.1 | 447.8 | 37 | 1/83 | 576.7 | 627.5 |
| 2 | 2/80 | 517.5 | 449.1 | 38 | 2/83 | 551.4 | 604.1 |
| 3 | 3/80 | 560.4 | 455.8 | 39 | 3/83 | 533.5 | 591.1 |
| 4 | 4/80 | 585.4 | 465.5 | 40 | 4/83 | 515.3 | 591.1 |
| 5 | 5/80 | 595.5 | 470.9 | 41 | 5/83 | 537.2 | 591.1 |
| 6 | 6/80 | 598.6 | 478.6 | 42 | 6/83 | 559.5 | 591.0 |
| 7 | 7/80 | 601.1 | 480.7 | 43 | 7/83 | 566.6 | 589.1 |
| 8 | 8/80 | 602.9 | 494.2 | 44 | 8/83 | 571.2 | 588.6 |
| 9 | 9/80 | 599.6 | 498.1 | 45 | 9/83 | 566.3 | 589.1 |
| 10 | 10/80 | 591.5 | 505.3 | 46 | 10/83 | 559.2 | 589.1 |
| 11 | 11/80 | 590.8 | 523.6 | 47 | 11/83 | 548.2 | 589.0 |
| 12 | 12/80 | 596.1 | 551.7 | 48 | 12/83 | 535.8 | 588.0 |
| 13 | 1/81 | 607.5 | 614.1 | 49 | 1/84 | 518.3 | 589.0 |
| 14 | 2/81 | 632.9 | 734.7 | 50 | 2/84 | 512.4 | 589.0 |
| 15 | 3/81 | 683.2 | 734.8 | 51 | 3/84 | 517.9 | 589.0 |
| 16 | 4/81 | 694.7 | 734.5 | 52 | 4/84 | 520.5 | 587.5 |
| 17 | 5/81 | 690.4 | 732.3 | 53 | 5/84 | 532.6 | 587.5 |
| 18 | 6/81 | 685.6 | 711.3 | 54 | 6/84 | 531.0 | 587.0 |
| 19 | 7/81 | 677.4 | 696.5 | 55 | 7/84 | 520.9 | 586.4 |
| 20 | 8/81 | 668.4 | 694.7 | 56 | 8/84 | 504.6 | 585.1 |
| 21 | 9/81 | 666.4 | 694.7 | 57 | 9/84 | 500.3 | 584.7 |
| 22 | 10/81 | 666.1 | 687.2 | 58 | 10/84 | 509.8 | 584.0 |
| 23 | 11/81 | 661.7 | 685.2 | 59 | 11/84 | 511.3 | 571.8 |
| 24 | 12/81 | 657.7 | 686.3 | 60 | 12/84 | 502.0 | 566.2 |
| 25 | 1/82 | 651.7 | 686.3 | 61 | 1/85 | 480.5 | 550.3 |
| 26 | 2/82 | 642.3 | 671.6 | 62 | 2/85 | 458.4 | 536.3 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 27 | 3/82 | 621.1 | 649.3 | 63 | 3/85 | 467.2 | 536.6 |
| 28 | 4/82 | 578.6 | 625.9 | 64 | 4/85 | 493.9 | 538.4 |
| 29 | 5/82 | 555.7 | 625.8 | 65 | 5/85 | 522.5 | 541.3 |
| 30 | 6/82 | 582.7 | 626.2 | 66 | 6/85 | 535.7 | 540.6 |
| 31 | 7/82 | 628.8 | 626.3 | 67 | 7/85 | 539.3 | 539.6 |
| 32 | 8/82 | 636.3 | 626.3 | 68 | 8/85 | 526.7 | 535.4 |
| 33 | 9/82 | 628.4 | 626.7 | 69 | 9/85 | 513.6 | 536.6 |
| 34 | 10/82 | 617.2 | 641.1 | 70 | 10/85 | 506.1 | 539.2 |
| 35 | 11/82 | 611.0 | 640.0 | 71 | 11/85 | 520.1 | 541.8 |
| 36 | 12/82 | 600.7 | 628.1 | 72 | 12/85 | 523.0 | 544.3 |

## 10.2.1 Model identification of the series PGAS

We will first model the series PGAS alone. In Section 4 we will incorporate the additional information of PCRUDE in order to improve the forecasting accuracy of PGAS. As a first step in the identification of PGAS, we will plot the data using the TSPLOT paragraph (see Chapter 3). Since the data are collected monthly, we will specify a seasonality of 12.

-->TSPLOT PGAS. SEASONALITY IS 12. SYMBOL IS '*'.

```
  TIME SERIES PLOT FOR THE VARIABLE PGAS

      ------------+-----------+-----------+-----------+-----------+-----------+--------
690.00 +              ****                                                          +
      I                  **                                                         I
      I                    ****                                                     I
      I                        **                                                   I
630.00 +             *               ***                                           +
      I             *             *      **                                        I
      I    *****  *                     *                                         I
      I   *     **                 * *                                             I
570.00 +                                  *     ***                               +
      I  *                           *        *   *   **                           I
      I                                        * *      *    *           **        I
      I *                                                * ** **         *  *  **    I
510.00 +                                       *         *     * **         **      +
      I                                                    *  *   *                I
      I*                                                      *                    I
      I                                                        **                  I
450.00 +                                                                           +
      ------------+-----------+-----------+-----------+-----------+-----------+--------
                  12          24          36          48          60          72
```

We observe that the series appears to drift, or fluctuate. Since the series demonstrates no fixed mean level, we can conclude the series is nonstationary. Hence, we may expect that our model will include a differencing term; i.e., (1-B). As a check of nonstationarity, we will compute the ACF of PGAS for 24 lags.

-->ACF PGAS. MAXLAG IS 24.

```
TIME PERIOD ANALYZED . . . . . .  1  TO    72
NAME OF THE SERIES . . . . . .         PGAS
EFFECTIVE NUMBER OF OBSERVATIONS . . .   72
STANDARD DEVIATION OF THE SERIES . . .  60.7033
MEAN OF THE (DIFFERENCED) SERIES . . .  571.6180
STANDARD DEVIATION OF THE MEAN . . . .   7.1540
T-VALUE OF MEAN (AGAINST ZERO) . . . .  79.9024

AUTOCORRELATIONS

 1- 12     .94  .84  .75  .68  .65  .62  .61  .59  .57  .53  .48  .40
 ST.E.     .12  .20  .24  .27  .29  .31  .33  .35  .36  .37  .38  .39
  Q       66.2  120  163  200  233  265  295  324  351  375  395  409

13- 24     .31  .24  .19  .16  .12  .08  .04  .01 -.03 -.07 -.12 -.16
 ST.E.     .40  .40  .40  .40  .40  .40  .40  .40  .40  .40  .40  .40
  Q        418  423  427  429  431  431  431  431  431  432  434  436


         -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
           +----+----+----+----+----+----+----+----+----+----+
                                     I
 1    .94                      +     IXXXXX+XXXXXXXXXXXXXXXXX
 2    .84                   +        IXXXXXXXXX+XXXXXXXXXX
 3    .75                +           IXXXXXXXXXX+XXXXXXX
 4    .68               +            IXXXXXXXXXXX+XXXX
 5    .65              +             IXXXXXXXXXXXX+XX
 6    .62             +              IXXXXXXXXXXXXX+X
 7    .61            +               IXXXXXXXXXXXXXX+
 8    .59           +                IXXXXXXXXXXXXXX +
 9    .57          +                 IXXXXXXXXXXXXX   +
10    .53          +                 IXXXXXXXXXXXX    +
11    .48         +                  IXXXXXXXXXX       +
12    .40         +                  IXXXXXXXXX        +
13    .31         +                  IXXXXXXX          +
14    .24        +                   IXXXXX             +
15    .19        +                   IXXXX              +
16    .16        +                   IXXX               +
17    .12        +                   IXX                +
18    .08        +                   IX                 +
19    .04        +                   I                  +
20    .01        +                   I                  +
21   -.03        +                  XI                  +
22   -.07        +                 XXI                  +
23   -.12        +                XXXI                  +
24   -.16        +               XXXXI                  +
```

The ACF is indicative of a nonstationary series, since the ACF decays relatively slowly as a power of a value between .95 and 1.0. That is, the ACF appears to "die out" according to $\phi^k$, k = 1,2, ... with $\phi$ between .95 and 1.0. This implies we could employ (1-B) in a model.

We will now use the IDEN paragraph to compute 12 lags of the sample ACF and PACF of a first-order differenced series.

-->IDEN PGAS. DFORDER IS 1. MAXLAG IS 12.

```
                                          1
DIFFERENCE ORDERS. . . . . . . . . . (1-B  )
TIME PERIOD ANALYZED . . . . . . . .  1  TO    72
NAME OF THE SERIES . . . . . . . . .         PGAS
EFFECTIVE NUMBER OF OBSERVATIONS . . .         71
STANDARD DEVIATION OF THE SERIES . . .     17.4019
MEAN OF THE (DIFFERENCED) SERIES . . .       .5901
STANDARD DEVIATION OF THE MEAN . . . .      2.0652
T-VALUE OF MEAN (AGAINST ZERO) . . . .       .2858

AUTOCORRELATIONS

  1- 12     .62  .07 -.22 -.25 -.15 -.06 -.03  .00  .09  .22  .31  .23
  ST.E.     .12  .16  .16  .16  .17  .17  .17  .17  .17  .17  .17  .18
   Q       28.3 28.7 32.4 37.2 39.0 39.3 39.4 39.4 40.1 44.3 52.4 57.2


         -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
           +----+----+----+----+----+----+----+----+----+----+
                                     I
  1    .62                    +     IXXXXX+XXXXXXXXX
  2    .07                  +       IXX     +
  3   -.22                  + XXXXXXI       +
  4   -.25                  + XXXXXXI       +
  5   -.15                  +   XXXXI       +
  6   -.06                  +     XXI       +
  7   -.03                  +      XI       +
  8    .00                  +       I       +
  9    .09                  +       IXX     +
 10    .22                  +       IXXXXXX +
 11    .31                 +        IXXXXXXXX+
 12    .23                 +        IXXXXXX  +


PARTIAL AUTOCORRELATIONS

  1- 12     .62 -.50  .04 -.06 -.03 -.04 -.05  .05  .10  .17  .11  .01
  ST.E.     .12  .12  .12  .12  .12  .12  .12  .12  .12  .12  .12  .12


         -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
           +----+----+----+----+----+----+----+----+----+----+
                                     I
  1    .62                    +     IXXXXX+XXXXXXXXX
  2   -.50              XXXXXXX+XXXXXI     +
  3    .04                    +     IX    +
  4   -.06                    +    XI     +
  5   -.03                    +    XI     +
  6   -.04                    +    XI     +
  7   -.05                    +    XI     +
  8    .05                    +     IX    +
  9    .10                    +     IXX   +
 10    .17                    +     IXXXX +
 11    .11                    +     IXXX  +
 12    .01                    +     I     +
```

The PACF of the differenced series appears to cut off after the second lag and the ACF decays, hence an ARIMA(2,1,0) model may be appropriate. We may confirm this by observing the EACF of the differenced series.

-->EACF PGAS. DFORDER IS 1.

```
                                          1
DIFFERENCE ORDERS. . . . . . . . . . (1-B  )
TIME PERIOD ANALYZED . . . . . . . .  1  TO    72
NAME OF THE SERIES . . . . . . . . .         PGAS
EFFECTIVE NUMBER OF OBSERVATIONS . . .         71
STANDARD DEVIATION OF THE SERIES . . .     17.4019
MEAN OF THE (DIFFERENCED) SERIES . . .       .5901
STANDARD DEVIATION OF THE MEAN . . . .      2.0652
T-VALUE OF MEAN (AGAINST ZERO) . . . .       .2858


THE EXTENDED ACF TABLE

(Q-->)   0    1    2    3    4    5    6    7    8    9   10   11   12
-------------------------------------------------------------------------
(P= 0)  .62  .07 -.22 -.25 -.15 -.06 -.03  .00  .09  .22  .31  .23 -.00
(P= 1)  .54  .12 -.34 -.20 -.09  .05 -.02  .00  .03  .10  .21  .20 -.00
(P= 2)  .08 -.12 -.10 -.10 -.12 -.02  .00 -.02  .03  .04  .04  .15 -.04
(P= 3)  .48 -.12  .01  .00 -.13 -.01 -.05 -.01  .03  .03 -.01  .15 -.18
(P= 4) -.46 -.16  .05 -.03 -.03  .04 -.03 -.02 -.02  .07  .01  .13 -.07
(P= 5) -.46 -.00 -.02 -.05 -.07  .01  .01  .01  .05 -.01 -.03  .10 -.04
(P= 6) -.47 -.03 -.09  .27 -.07  .03 -.01 -.03  .05 -.00 -.03  .11 -.04


SIMPLIFIED EXTENDED ACF TABLE (5% LEVEL)

(Q-->)  0  1  2  3  4  5  6  7  8  9 10 11 12
-----------------------------------------------
(P= 0)  X  O  O  O  O  O  O  O  O  O  O  O  O
(P= 1)  X  O  X  O  O  O  O  O  O  O  O  O  O
(P= 2)  O  O  O  O  O  O  O  O  O  O  O  O  O
(P= 3)  X  O  O  O  O  O  O  O  O  O  O  O  O
(P= 4)  X  O  O  O  O  O  O  O  O  O  O  O  O
(P= 5)  X  O  O  O  O  O  O  O  O  O  O  O  O
(P= 6)  X  O  O  O  O  O  O  O  O  O  O  O  O
```

The above EACF also supports an ARIMA(2,1,0) model, as a triangular region of zeroes can be constructed from P=2, Q=0 (as indicated by the solid lines drawn in). However, due to the significant value at P=1 and Q=2, we can also construct a triangular region from P=0, Q=2; i.e., an ARIMA(0,1,2) model. Although this model may merit study, for the purposes of illustration, it will not used in this example.

## 10.2.2 Model specification and estimation for PGAS

We have tentatively identified the model

$$(1-\phi_1 B-\phi_2 B^2)(1-B)Z_t = a_t \tag{10.13}$$

where $Z_t$ represents the series PGAS. The t-value of the mean (against zero) is not significant, as indicated in both the summary statistics of the IDEN and EACF paragraphs. Hence, we have not included a constant term in (10.13). We will now specify the above model and hold it in memory under the label PGASAR.

```
-->TSMODEL   NAME IS PGASAR.   @
       MODEL IS (1 - PHI1*B - PHI2*B**2)PGAS(1) = NOISE.

SUMMARY FOR UNIVARIATE TIME SERIES MODEL --   PGASAR

------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1
  PGAS     RANDOM     ORIGINAL     (1-B  )
------------------------------------------------------------------------

 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

  1   PHI1     PGAS     AR       1      1     NONE      .1000
  2   PHI2     PGAS     AR       1      2     NONE      .1000
```

We now estimate the above model, and will retain the residuals of the fitted model in the variable RGASAR. The output has been condensed for presentation purposes.

```
-->ESTIM   PGASAR.  HOLD  RESIDUALS(RGASAR)

THE FOLLOWING ANALYSIS IS BASED ON TIME SPAN   1  THRU   72

SUMMARY FOR UNIVARIATE TIME SERIES MODEL --   PGASAR

------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1
  PGAS     RANDOM     ORIGINAL     (1-B  )
------------------------------------------------------------------------

 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

  1   PHI1     PGAS     AR       1      1     NONE      .9015     .1017   8.86
  2   PHI2     PGAS     AR       1      2     NONE     -.4884     .0990  -4.93


TOTAL SUM OF SQUARES . . . . . .   .265312E+06
TOTAL NUMBER OF OBSERVATIONS . . .          72
RESIDUAL SUM OF SQUARES. . . . . .   .853368E+04
R-SQUARE . . . . . .          .966
EFFECTIVE NUMBER OF OBSERVATIONS . .          69
RESIDUAL VARIANCE ESTIMATE . . . . .   .123677E+03
RESIDUAL STANDARD ERROR. . . . . .   .111210E+02
```

### 10.2.3 Diagnostic checks of estimated model

The fit of (10.13) yields

$$(1-.90B+.49B^2)(1-B)PGAS = NOISE \qquad (10.14)$$

Both AR parameters have significant t-values, and the residual series has a residual standard error of 11.12. Since our series is nonstationary we must be careful to compare this residual

standard error with that of the standard error of the first-differenced series. The latter value (see the sample EACF summary) is 17.40. Hence we have reduced the variability by 40%.

Our first diagnostic check of the fit is a plot of the residuals over time.

-->TSPLOT RGASAR. SEASONALITY IS 12. SYMBOL IS '*'.

```
  TIME SERIES PLOT FOR THE VARIABLE RGASAR

       ------------+-----------+-----------+-----------+-----------+-----------+--------
 36.00 +                                                                                 +
       I              *                                                                  I
       I                              *          *                                      I
       I                                                                                I
 20.00 +                                                                  *              +
       I             *                                                            *     I
       I                             *                     *                          I
       I    **      *   *              *          *                    **              I
  4.00 +   *  *   *      *  *       *                      **     **        *           +
       I       *   *           *         **          *                  *  **         I
       I        **        ** ** **                * *  *   *  *                        I
       I                                    **  *    *        *  ** *     *           I
-12.00 +                                  *             *    *      *          *       +
       I                         *         *  *                                      I
       I               *              *                                               I
       I                                                                              I
-28.00 +                          *                                                    +
       ------------+-----------+-----------+-----------+-----------+-----------+--------
                  12          24          36          48          60          72
```

No patterns are apparent in the plot. We may note 5 points that appear "apart" from the rest (at 15, 16, 28, 30, and 41). Note the first three residuals are "missing" since they cannot be computed from the difference equation of (10.14). The ACF of the residual series is also "clean".

-->ACF RGASAR. MAXLAG IS 12.

```
TIME PERIOD ANALYZED . . . . . .  4  TO    72
NAME OF THE SERIES . . . . . .       RGASAR
EFFECTIVE NUMBER OF OBSERVATIONS . . .          69
STANDARD DEVIATION OF THE SERIES . . .     11.1138
MEAN OF THE (DIFFERENCED) SERIES . . .      -.3990
STANDARD DEVIATION OF THE MEAN . . . .      1.3379
T-VALUE OF MEAN (AGAINST ZERO) . . . .      -.2983
```

```
AUTOCORRELATIONS

 1- 12     .01 -.08 -.03 -.04 -.07 -.01 -.02  .01  .00  .08  .08  .16
 ST.E.     .12  .12  .12  .12  .12  .12  .12  .12  .12  .12  .12  .12
  Q         .0   .4   .5   .6  1.0  1.0  1.0  1.0  1.0  1.5  2.1  4.5

        -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
          +----+----+----+----+----+----+----+----+----+----+----+
                                    I
 1    .01                   +      I     +
 2   -.08                   +    XXI     +
 3   -.03                   +     XI     +
 4   -.04                   +     XI     +
 5   -.07                   +    XXI     +
 6   -.01                   +      I     +
 7   -.02                   +      I     +
 8    .01                   +      I     +
 9    .00                   +      I     +
10    .08                   +      IXX   +
11    .08                   +      IXX   +
12    .16                   +      IXXXX +
```

## 10.3 Modeling Seasonal Time Series

In the previous sections, we found we could adequately model the given series through the use of nonseasonal ARIMA models. That is, the models we employed did not need to account for any seasonal pattern present in a series. However, we often encounter situations in which a time series exhibits some periodic, or seasonal, pattern. For example, data recorded monthly may exhibit "similar" behavior from year to year; that is, a seasonality of period 12. Data recorded quarterly data may have 4 as its seasonality, and data recorded hourly may have 24 as its periodicity.

To illustrate the modeling of a seasonal time series, we will consider Series G of Box and Jenkins (1970). The data represents the totals of international airline passengers (in thousands) for the period January 1949 through December 1960, inclusive. The data are listed in Table 3, and are stored in the SCA workspace under the label SERIESG.

**Table 3 Series G of Box and Jenkins (1970): Monthly totals (in thousands) of international airline passengers, January 1949 - December 1960**

| Year | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1949 | 112 | 118 | 132 | 129 | 121 | 135 | 148 | 148 | 136 | 119 | 104 | 118 |
| 1950 | 115 | 126 | 141 | 135 | 125 | 149 | 170 | 170 | 158 | 133 | 114 | 140 |
| 1951 | 145 | 150 | 178 | 163 | 172 | 178 | 199 | 199 | 184 | 162 | 146 | 166 |
| 1952 | 171 | 180 | 193 | 181 | 183 | 218 | 230 | 242 | 209 | 191 | 172 | 194 |
| 1953 | 196 | 196 | 236 | 235 | 229 | 243 | 264 | 272 | 237 | 211 | 180 | 201 |
| 1954 | 204 | 188 | 235 | 227 | 234 | 264 | 302 | 293 | 259 | 229 | 203 | 229 |
| 1955 | 242 | 233 | 267 | 269 | 270 | 315 | 364 | 347 | 312 | 274 | 237 | 278 |
| 1956 | 284 | 277 | 317 | 313 | 318 | 374 | 413 | 405 | 355 | 306 | 271 | 306 |
| 1957 | 315 | 301 | 356 | 348 | 355 | 422 | 465 | 467 | 404 | 347 | 305 | 336 |
| 1958 | 340 | 318 | 362 | 348 | 363 | 435 | 491 | 505 | 404 | 359 | 310 | 337 |
| 1959 | 360 | 342 | 406 | 396 | 420 | 472 | 548 | 559 | 463 | 407 | 362 | 405 |
| 1960 | 417 | 391 | 419 | 461 | 472 | 535 | 622 | 606 | 508 | 461 | 390 | 432 |

### 10.3.1 Model identification

We will first plot the data using the TSPLOT paragraph (see Chapter 3). Since the data are monthly totals, we will specify 12 as our seasonal period. We will break the data into two sections and keep the ranges the same on each plot.

```
-->TSPLOT   SERIESG.  SEASONALITY IS 12.   RANGE IS 100, 630.    @
-->         SPAN IS 1,72.
```

```
  TIME SERIES PLOT FOR THE VARIABLE SERIESG

       ------------+-----------+-----------+-----------+-----------+-----------+--------
630.00 +                                                                          +
       I                                                                          I
       I                                                                          I
       I                                                                          I
490.00 +                                                                          +
       I                                                                          I
       I                                                                          I
       I                                                                          I
350.00 +                                                                          +
       I                                                          7               I
       I                                                  78          6 8         I
       I                                          78      3456  9     3 5   90 B  I
210.00 +                             78      3  6  9  B12       0 B1  4      A    +
       I                  789     3456  90 B12 45    0A           A  2             I
       I  34 6789    23456   0 B12        A                                      I
       I12  5    0AB1         A                                                   I
 70.00 +                                                                          +
       ------------+-----------+-----------+-----------+-----------+-----------+--------
                  12          24          36          48          60          72
```

```
-->TSPLOT   SERIESG.  SEASONALITY IS 12.   RANGE IS 100, 630.    @
-->         SPAN IS 73, 144.
```

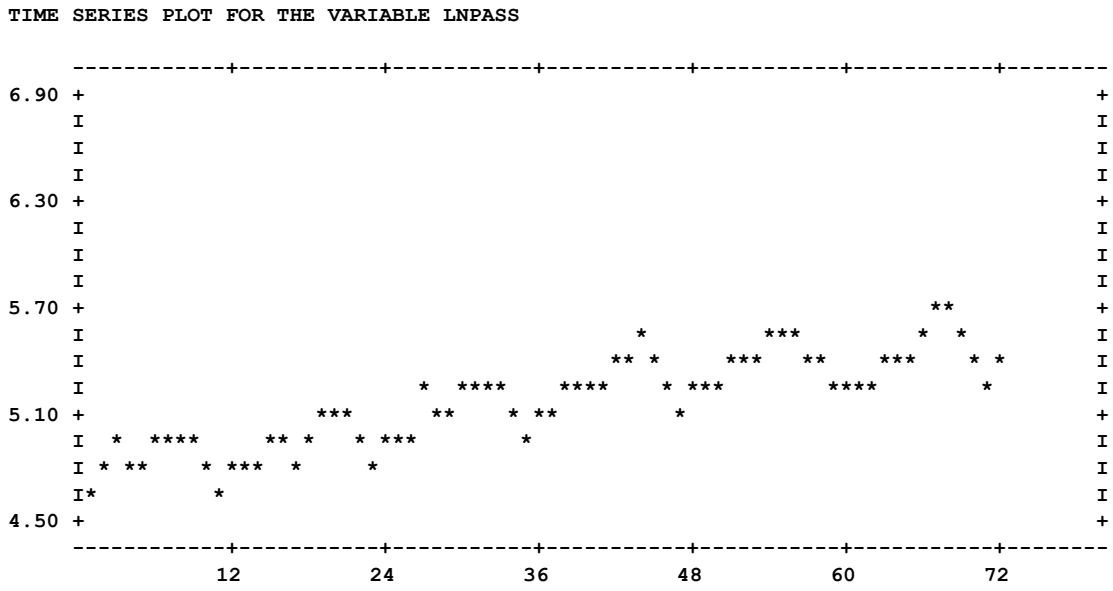
```
  TIME SERIES PLOT FOR THE VARIABLE SERIESG

       ------------+-----------+-----------+-----------+-----------+-----------+--------
630.00 +                                                          7               +
       I                                                           8              I
       I                                                  78                      I
       I                                                          6  9            I
490.00 +                                       78                                 +
       I                             78                     6  9      45    0      I
       I                  78         6  9        6  9     3 5    0 B1 3        B   I
       I                 6                               4         2        A     I
350.00 +      78            9     345    0 B1 345    0 B12        A                +
       I     6  9     345    0 B12        A  2        A                           I
       I  345    0 B12        A                                                   I
       I12        A                                                              I
210.00 +                                                                          +
       I                                                                          I
       I                                                                          I
       I                                                                          I
 70.00 +                                                                          +
       ------------+-----------+-----------+-----------+-----------+-----------+--------
                  84          96         108         120         132         144
```

We observe both a distinct seasonality in the data and the presence of a trend. As a result of the trend, we are certain that the series does not have a stationary mean level. In addition, the variability of the data seems to increase with the mean level. In order to stabilize this variability, it is necessary for us to transform the data. The logarithmic transformation is useful when the variability appears to be a function of the mean. We can use an analytic statement (see Appendix A) to transform the data. We will store the transformed data under the name LNPASS.

```
-->LNPASS = LN(SERIESG)
```

A time plot of the transformed data still exhibits a trend and seasonality, but we seem to have stabilized the variance over the length of the series.

```
-->TSPLOT   LNPASS.   SEASONALITY IS 12.   RANGE IS 4.5, 7.0 .    @
-->         SPAN IS 1, 72.   SYMBOL IS  '*'.
```



```
-->TSPLOT   LNPASS.  SEASONALITY IS 12.  RANGE IS 4.5, 7.0 .   @
-->        SPAN IS 73, 144.  SYMBOL IS  '*'.
```

```
TIME SERIES PLOT FOR THE VARIABLE LNPASS

     ------------+-----------+-----------+-----------+-----------+-----------+-----------+--------
6.90 +                                                                                     +
     I                                                                                     I
     I                                                                                     I
     I                                                                        **           I
6.30 +                                                        **         *  *           +
     I                            **         ***         *  *      **    *           I
     I                  **         *  *          *     ***    * ****       **        I
     I      **         *  *     ***    * ** ***    * ***        *                      I
5.70 +     *  *  ** ***    * ***        *  *        *                                 +
     I* ***    *   *        *                                                        I
     I *        *                                                                    I
     I                                                                               I
5.10 +                                                                               +
     I                                                                               I
     I                                                                               I
     I                                                                               I
4.50 +                                                                               +
     ------------+-----------+-----------+-----------+-----------+-----------+-----------+--------
                 84          96         108         120         132         144
```

We expect that LNPASS is not stationary. This is confirmed when we compute and display the sample ACF of the series.

```
-->ACF   LNPASS
```

```
TIME PERIOD ANALYZED . . . . . .  1  TO   144
NAME OF THE SERIES . . . . . .       LNPASS
EFFECTIVE NUMBER OF OBSERVATIONS . . .          144
STANDARD DEVIATION OF THE SERIES . . .        .4399
MEAN OF THE (DIFFERENCED) SERIES . . .       5.5422
STANDARD DEVIATION OF THE MEAN . . . .        .0367
T-VALUE OF MEAN (AGAINST ZERO) . . . .     151.1774
```

```
AUTOCORRELATIONS

          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+
                                     I
  1    .95                      +   IXXX+XXXXXXXXXXXXXXXXXXX
  2    .90                    +     IXXXXXX+XXXXXXXXXXXXXX
  3    .85                  +       IXXXXXXXX+XXXXXXXXXXX
  4    .81                 +        IXXXXXXXXX+XXXXXXXXX
  5    .78                +         IXXXXXXXXXX+XXXXXXX
  6    .76               +          IXXXXXXXXXXX+XXXXXX
  7    .74              +           IXXXXXXXXXXXX+XXXXX
  8    .73              +           IXXXXXXXXXXXX+XXXXX
  9    .73             +            IXXXXXXXXXXXXX+XXXX
 10    .74            +             IXXXXXXXXXXXXXX+XXXX
 11    .76            +             IXXXXXXXXXXXXXX+XXXX
 12    .76           +              IXXXXXXXXXXXXXXX+XXX
 13    .72           +              IXXXXXXXXXXXXXXX+XX
 14    .66          +               IXXXXXXXXXXXXXXXXX
 15    .62          +               IXXXXXXXXXXXXXXX +
 16    .58         +                IXXXXXXXXXXXXXX   +
 17    .54         +                IXXXXXXXXXXXXXX   +
 18    .52         +                IXXXXXXXXXXXXX    +
 19    .50        +                 IXXXXXXXXXXXXX     +
 20    .49        +                 IXXXXXXXXXXXX      +
 21    .50        +                 IXXXXXXXXXXX       +
 22    .51        +                 IXXXXXXXXXXXX      +
 23    .52        +                 IXXXXXXXXXXXX      +
 24    .52       +                  IXXXXXXXXXXXX       +
 25    .48       +                  IXXXXXXXXXXX        +
 26    .44       +                  IXXXXXXXXXX         +
 27    .40       +                  IXXXXXXXXX          +
 28    .36       +                  IXXXXXXXX           +
 29    .34       +                  IXXXXXXX            +
 30    .31      +                   IXXXXXXX             +
 31    .30      +                   IXXXXXX              +
 32    .29      +                   IXXXXXX              +
 33    .30      +                   IXXXXXX              +
 34    .30      +                   IXXXXXXX             +
 35    .32      +                   IXXXXXXX             +
 36    .32      +                   IXXXXXXX             +
```

The ACF has a "classic" slow die-out pattern that is typical of a nonstationary series. Differencing is required. However, because the data is seasonal, we may wonder if the "proper" differencing factor is (1-B) or $(1-B^{12})$. The former reflects "within year" patterns, while the latter reflects a pattern that is promulgated over years. We can examine the sample ACF for using each of these differencing factors. The output has been edited for presentation purposes.

-->ACF LNPASS. DFORDER IS 1.

```
                                          1
DIFFERENCE ORDERS. . . . . . . . . . (1-B  )

TIME PERIOD ANALYZED . . . . . .  1  TO   144
NAME OF THE SERIES . . . . . .       LNPASS
EFFECTIVE NUMBER OF OBSERVATIONS . . .   143
STANDARD DEVIATION OF THE SERIES . . .  .1062
MEAN OF THE (DIFFERENCED) SERIES . . .  .0094
STANDARD DEVIATION OF THE MEAN . . . .  .0089
T-VALUE OF MEAN (AGAINST ZERO) . . . . 1.0631

AUTOCORRELATIONS

      -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8
       +----+----+----+----+----+----+----+----+
                           I
 1   .20                +   IXXX+X
 2  -.12                +XXXI   +
 3  -.15                XXXXI   +
 4  -.32              XXXX+XXXI   +
 5  -.08                 +  XXI    +
 6   .03                 +    IX   +
 7  -.11                 + XXXI    +
 8  -.34              XXX+XXXXI    +
 9  -.12                 + XXXI    +
10  -.11                 + XXXI    +
11   .21                 +    IXXXXX
12   .84                 +    IXXXX+XXXXXXXXXXXXXXXX
13   .22               +      IXXXXX +
14  -.14               +   XXXI      +
15  -.12               +   XXXI      +
16  -.28               XXXXXXXI      +
17  -.05             +      XI       +
18   .01             +       I       +
19  -.11             +    XXXI       +
20  -.34             XXXXXXXXI       +
21  -.11             +    XXXI       +
22  -.08             +     XXI       +
23   .20             +       IXXXXX  +
24   .74             +       IXXXXXXX+XXXXXXXXXX
25   .20           +         IXXXXX   +
26  -.12           +     XXXI         +
27  -.10           +     XXXI         +
28  -.21           +   XXXXXI         +
29  -.07           +      XXI         +
30   .02           +        I         +
31  -.12           +     XXXI         +
32  -.29           + XXXXXXXI         +
33  -.13           +     XXXI         +
34  -.04         +         XI          +
35   .15         +          IXXXX      +
36   .66         +          IXXXXXXXXX+XXXXXX
```

-->ACF LNPASS. DFORDER IS 12.

```
                                         12
DIFFERENCE ORDERS. . . . . . . . . . (1-B  )

TIME PERIOD ANALYZED . . . . . .  1  TO   144
NAME OF THE SERIES . . . . . .       LNPASS
EFFECTIVE NUMBER OF OBSERVATIONS . . .   132
STANDARD DEVIATION OF THE SERIES . . .  .0614
MEAN OF THE (DIFFERENCED) SERIES . . .  .1198
STANDARD DEVIATION OF THE MEAN . . . .  .0053
T-VALUE OF MEAN (AGAINST ZERO) . . . . 22.4170

AUTOCORRELATIONS

 -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8
  +----+----+----+----+----+----+----+----+
                     I
 1   .71           +   IXXX+XXXXXXXXXXXXXX
 2   .62         +     IXXXXX+XXXXXXXXX
 3   .48        +      IXXXXXX+XXXXX
 4   .44       +       IXXXXXXX+XXX
 5   .39       +       IXXXXXXX+XX
 6   .32       +       IXXXXXXXX
 7   .24      +        IXXXXXX  +
 8   .19      +        IXXXXX   +
 9   .15      +        IXXXX    +
10  -.01      +        I        +
11  -.11      +     XXXI        +
12  -.24      +  XXXXXXI        +
13  -.14      +    XXXXI        +
14  -.14      +    XXXXI        +
15  -.10      +      XXI        +
16  -.15      +    XXXXI        +
17  -.10      +      XXI        +
18  -.11      +     XXXI        +
19  -.14      +    XXXXI        +
20  -.16      +    XXXXI        +
21  -.11      +     XXXI        +
22  -.08      +      XXI        +
23   .00      +        I        +
24  -.05      +       XI        +
25  -.10      +     XXXI        +
26  -.09      +      XXI        +
27  -.13      +     XXXI        +
28  -.15      +    XXXXI        +
29  -.19     +     XXXXXI        +
30  -.20     +     XXXXXI        +
31  -.19     +     XXXXXI        +
32  -.15     +      XXXXI        +
33  -.22     +    XXXXXXI        +
34  -.23     +    XXXXXXI        +
35  -.27     +  XXXXXXXI         +
36  -.22     +    XXXXXXI        +
```

Clearly the use of (1-B) alone does not remove the effects of nonstationarity from the data, since the ACF at lags 12, 24, 36 (and so on) exhibit the same behavior as the ACF of the original series. A seasonal difference is warranted. However, the seasonally differenced series alone is not stationary as indicated by the slow decay of its ACF.

In order to achieve stationarity here, we need to employ both a nonseasonal and a seasonal differencing factor, in the multiplicative form $(1\text{-}B)(1-B^{12})$. We can specify these factors and obtain the sample ACF of the differenced series by entering

-->ACF LNPASS. DFORDERS ARE 1, 12.

```
                                                 1      12
DIFFERENCE ORDERS. . . . . . . . . . . . (1-B  ) (1-B  )
TIME PERIOD ANALYZED . . . . . . . . .  1  TO   144
NAME OF THE SERIES . . . . . . . . . .       LNPASS
EFFECTIVE NUMBER OF OBSERVATIONS . . .          131
STANDARD DEVIATION OF THE SERIES . . .        .0457
MEAN OF THE (DIFFERENCED) SERIES . . .        .0003
STANDARD DEVIATION OF THE MEAN . . . .        .0040
T-VALUE OF MEAN (AGAINST ZERO) . . . .        .0729

AUTOCORRELATIONS

          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+
                                      I
  1   -.34                  XXXXX+XXXI   +
  2    .11                      +    IXXX +
  3   -.20                      XXXXXI    +
  4    .02                      +    IX   +
  5    .06                      +    IX   +
  6    .03                      +    IX   +
  7   -.06                      +   XI    +
  8    .00                      +    I    +
  9    .18                      +    IXXXX+
 10   -.08                      +  XXI    +
 11    .06                      +    IXX  +
 12   -.39                  XXXXX+XXXXI    +
 13    .15                      +     IXXXX +
 14   -.06                      +    XI     +
 15    .15                      +     IXXXX +
 16   -.14                      +  XXXI     +
 17    .07                      +     IXX   +
 18    .02                      +     I     +
 19   -.01                      +     I     +
 20   -.12                      +  XXXI     +
 21    .04                      +     IX    +
 22   -.09                      +   XXI     +
 23    .22                      +     IXXXXXX
 24   -.02                      +     I     +
 25   -.10                      +  XXXI     +
 26    .05                      +     IX    +
 27   -.03                      +    XI     +
 28    .05                      +     IX    +
 29   -.02                      +     I     +
 30   -.05                      +    XI     +
 31   -.05                      +    XI     +
 32    .20                      +     IXXXXX+
 33   -.12                      +  XXXI     +
 34    .08                      +     IXX   +
 35   -.15                      + XXXXI     +
 36   -.01                      +     I     +
```

The sample ACF has significant negative values at lags 1 and 12. Many texts provide guides for the pattern of the ACF for many types of seasonal models. These include Appendix 9.1 of Box and Jenkins (1970), Section 6.2 of Abraham and Ledolter (1983), Section 4.4 of Vandaele (1983), and Section 10.2 of Cryer (1986). The above pattern is indicative of a multiplicative MA(1) and MA(12) model, that is, $(1-\theta_1 B)(1-\theta_{12}B^{12})$.

## Multiplicative seasonal models

Multiplicative seasonal ARIMA models are often described as $(p,d,q)$ x $(P,D,Q)_s$ models, where s is the seasonal period, and P, D, and Q are the orders of the seasonal components. The values of the differencing orders, d and D, of this model are usually either 0 or 1.

$$(1-\phi_1 B-\cdots-\phi_p B^p)(1-\Phi_1 B^s-\cdots-\Phi_p B^{Ps})(1-B)^d(1-B^s)^D Z_t$$
$$= \quad (1-\theta_1 B-\cdots-\theta_q B^q)(1-\theta_1 B^s-\cdots-\theta_P B^{Qs})a_t \qquad (10.15)$$

We have tentatively identified a multiplicative $(0,1,1)$ x $(0,1,1)_{12}$ model for the logged airline data. This particular model

$$(1-B)(1-B^{12})Z_t = (1-\theta B)(1-\theta_{12}B^{12})a_t \qquad (10.16)$$

has become known as the airline model and has been shown to be very useful in modeling many seasonal series. Unfortunately this model is often misused. One common error in ARIMA modeling is to over-difference the original series, which automatically leads to an airline model.

### 10.3.2 Model specification and estimation

The t-value of the mean (against zero) for the multiplicatively differenced series is not significant. Thus, we have tentatively identified the model of the form in (10.16) where $Z_t$ is the natural log of SERIESG (i.e., LNPASS). We can specify this model by entering

```
-->TSMODEL   NAME IS AIRLINE.  MODEL IS        @
-->        LNPASS(1,12) = (1 - THETA1*B)(1 - THETA12*B**12)NOISE.
```

```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- AIRLINE
----------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1      12
 LNPASS    RANDOM     ORIGINAL     (1-B  ) (1-B  )
----------------------------------------------------------------------
 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT               ERROR  VALUE

  1  THETA1   LNPASS    MA       1      1     NONE      .1000
  2 THETA12   LNPASS    MA       2     12     NONE      .1000
```

Note we have specified our differencing operators (1-B)$(1-B^{12})$ as (1,12). This is consistent with the specification of DFORDERS in the ACF, PACF, IDEN and EACF paragraphs. We could also specify these operators as ((1)(12)) or ((1-B)(1-B**12)) if we desire.

## Estimation algorithms for MA parameters

The model is held in the workspace under the label AIRLINE. We can now estimate this model. Whenever we have a model that contains moving average (MA) parameters, we have a choice in the estimation algorithm we can use. The SCA System will estimate parameters using a non-linear method that minimizes an objective function.

The default algorithm used in the minimization of this function employs a conditional estimation method. We can also employ an exact estimation method. This latter method is a more computationally intensive method. However, it is usually good practice to employ the exact algorithm whenever an MA parameter is present (in particular, in a seasonal factor).

The most efficient way to employ exact estimation is to first estimate a model using the default conditional method. Then we can re-estimate the model using the exact method. The advantage in doing this is that the conditional method will provide a good starting point from which the exact method may begin. We can accomplish this easily in the SCA System since each model maintains a "memory" of the last estimate of a parameter.

We will first estimate our airline model using the conditional method by simply entering

-->ESTIM AIRLINE

```
THE FOLLOWING ANALYSIS IS BASED ON TIME SPAN   1  THRU  144


SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- AIRLINE
-----------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1      12
 LNPASS    RANDOM     ORIGINAL     (1-B  ) (1-B  )
-----------------------------------------------------------------------
 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT               ERROR  VALUE

  1  THETA1   LNPASS    MA       1      1     NONE      .3776     .0813   4.64
  2 THETA12   LNPASS    MA       2     12     NONE      .5728     .0776   7.38

TOTAL SUM OF SQUARES . . . . . .   .278684E+02
TOTAL NUMBER OF OBSERVATIONS . . .          144
RESIDUAL SUM OF SQUARES. . . . . .   .181926E+00
R-SQUARE . . . . . .          .993
EFFECTIVE NUMBER OF OBSERVATIONS . .          131
RESIDUAL VARIANCE ESTIMATE . . . . .   .138875E-02
RESIDUAL STANDARD ERROR. . . . . .   .372659E-01
```

We may observe that our MA factor estimates, .3776 and .5728, do not indicate that either of our factors have roots by the unit circle. However, we will still employ the exact estimation method and retain the residuals (in the variable RESIDAIR) after the fit by entering

```
-->ESTIM  AIRLINE.  METHOD IS EXACT.  HOLD  RESIDUALS(RESIDAIR).
```

```
THE FOLLOWING ANALYSIS IS BASED ON TIME SPAN   1  THRU  144


SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- AIRLINE
-----------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1      12
 LNPASS    RANDOM     ORIGINAL     (1-B  ) (1-B  )
-----------------------------------------------------------------------
 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT               ERROR  VALUE

  1  THETA1   LNPASS    MA       1      1     NONE      .4021     .0802   5.01
  2 THETA12   LNPASS    MA       2     12     NONE      .5569     .0728   7.65

TOTAL SUM OF SQUARES . . . . . .   .278684E+02
TOTAL NUMBER OF OBSERVATIONS . . .          144
RESIDUAL SUM OF SQUARES. . . . . .   .176601E+00
R-SQUARE . . . . . .          .993
EFFECTIVE NUMBER OF OBSERVATIONS . .          131
RESIDUAL VARIANCE ESTIMATE . . . . .   .134810E-02
RESIDUAL STANDARD ERROR. . . . . .   .367165E-01
```

Our fitted model is, approximately,

$$(1-B)(1-B^{12})\text{LNPASS} = (1-.40B)(1-.56B^{12})\text{NOISE} \qquad (10.17)$$

The parameters are significant, based on their t-values; and the variation of the residual series (i.e., the variation still unexplained) is 0.00134. The variation after $(1\text{-}B)(1-B^{12})$ differencing is $(.0457)^2$ (see the ACF summary). Hence we have reduced variation by about 36% after differencing.

### 10.3.3 Diagnostic checks of the airline model

A time plot of the residual series (not shown here) does not reveal any abnormalities. We can compute and display 24 lags of the sample ACF of the residuals. We see the ACF is "clean". The output is edited for presentation purposes.

-->ACF RESIDAIR. MAXLAG IS 24.

```
TIME PERIOD ANALYZED . . . . . . . . . . 14  TO   144
NAME OF THE SERIES . . . . . . . . . .     RESIDAIR
EFFECTIVE NUMBER OF OBSERVATIONS . . .          131

AUTOCORRELATIONS


          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+
                                      I
  1    .01                        +   I   +
  2    .02                        +   IX  +
  3   -.13                        +XXXI   +
  4   -.13                        +XXXI   +
  5    .06                        +   IX  +
  6    .06                        +   IXX +
  7   -.06                        + XXI   +
  8   -.02                        +  XI   +
  9    .12                        +   IXXX+
 10   -.08                       +  XXI    +
 11    .01                       +    I    +
 12   -.06                       +   XI    +
 13    .02                       +    I    +
 14    .04                       +    IX   +
 15    .06                       +    IX   +
 16   -.14                       +XXXXI    +
 17    .03                       +    IX   +
 18    .00                       +    I    +
 19   -.10                       + XXXI    +
 20   -.10                       +  XXI    +
 21   -.03                       +   XI    +
 22   -.03                       +   XI    +
 23    .22                       +    IXXXXX
 24    .01                       +    I    +
```

## 10.4 Regression with Serially Correlated Errors: Transfer Function Models

In the previous sections, we derive a model for a series based on information of the series alone. However, it is often the case that a time series is either affected by, or related to, other time series. If we can incorporate the information from these series into our model, then we should be able to improve the ability of the model to explain or forecast the series of interest.

The SCA-GSA product provides for the inclusion of information of another series within a slightly restricted model setting. For the SCA-GSA product, one input series can be employed with a nonseasonal AR(1), AR(2), MA(1) or MA(2) time series model.

Often a forecaster needs to include more than one series in a model. This can arise if a time series is thought to be influenced by two or more explanatory time series, or if calendar variations or interventions are incorporated into a model. The SCA-UTS product has no limit on the number of additional series we may incorporate into a model, nor the form of the time series model related to the "dependent" series. Information related to the SCA-UTS product, and the above topics, can be found in *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*.

To illustrate the incorporation of additional information into a time series model, we will consider the gasoline data of Section 10.2 (and Chapter 9). In Section 2, we constructed a model for monthly gasoline prices. We also have available to us the price of crude petroleum at wells. It is understandable that the price of gasoline is strongly affected by the price of crude oil. In addition, there may be a delay between when the price of crude is set and when the price of gasoline is affected. As a result, we may be better able to model and forecast gasoline prices if we can incorporate this additional information into our model.

The data for these series are given in Table 2, and the prices of gasoline and crude oil are stored in the SCA workspace under the labels PGAS and PCRUDE, respectively. To see if these two series indeed are related, we will plot the data using the TSPLOT paragraph (see Chapter 3). Since the data are collected monthly, we specify a seasonality of 12.

```
-->TSPLOT PGAS.     SEASONALITY IS 12.
```

```
  TIME SERIES PLOT FOR THE VARIABLE PGAS

       ------------+-----------+-----------+-----------+-----------+-----------+--------
690.00 +              ****                                                              +
       I                  **                                                            I
       I                    ****                                                        I
       I                        **                                                      I
630.00 +             *               ***                                               +
       I            *             *      **                                           I
       I    *****  *                      *                                           I
       I   *     **                * *                                                I
570.00 +                                  *     ***                                   +
       I  *                        *        *   *   **                                I
       I                                     * *      *    *           **              I
       I *                                           * ** **         *  *  **          I
510.00 +                                    *         *     * **         **            +
       I                                                  *  *   *                     I
       I*                                                     *                        I
       I                                                       **                      I
450.00 +                                                                               +
       ------------+-----------+-----------+-----------+-----------+-----------+--------
                  12          24          36          48          60          72
```

-->TSPLOT PCRUDE. SEASONALITY IS 12.

```
  TIME SERIES PLOT FOR THE VARIABLE PCRUDE

       ------------+-----------+-----------+-----------+-----------+-----------+--------
720.00 +             *****                                                       +
       I                  ***                                                   I
       I                     ****                                               I
       I                         *                                              I
648.00 +                         *      **                                     +
       I                          ******  **                                    I
       I            *                        *                                  I
       I                                     ******************                 I
576.00 +                                                     ***                +
       I           *                                            **              I
       I                                                          ***********   I
       I          *                                                            I
504.00 +        **                                                            +
       I     ***                                                              I
       I   **                                                                 I
       I***                                                                   I
432.00 +                                                                      +
       ------------+-----------+-----------+-----------+-----------+-----------+--------
                  12          24          36          48          60          72
```

## 10.4.1 The transfer function model

We will extend the ARIMA model, to permit the inclusion of additional explanatory variables, in the following manner. We first distinguish our series as "input", $X_t$, and "output", $Y_t$. Assuming the input and output series are both stationary, the form of our model is

$$Y_t = C + V(B)X_t + N_t, \tag{10.18}$$

or

$$Y_t = C + \{\omega(B)/\delta(B)\}X_t + N_t \tag{10.19}$$

The two basic components of this model are explained below.

### The transfer function

The component V(B), or ω(B)/δ(B), is referred to as the transfer function in the above model. It provides a measure of how information contained in the input series affects the output series. The term

$$V(B) = v_0 + v_1B + v_2B^2 + v_3B^3 + \cdots \tag{10.20}$$

is comprised of values known as impulse response weights, or simply transfer function weights. These weights provide a measure of when an observation of the input series affects the output series, and the weight given to it. That is, $v_0$ is a measure of how the currently observed output is affected by the currently observed input; $v_1$ is a measure of how the

currently observed output is affected by the last value of the input series; $v_3$ is a measure of how the current output is affected by the input series two periods ago; and so on. Since V(B) may consist of many (possibly infinite) terms, in some situations we may approximate it by a fraction of polynomial operators. This rational polynomial, ω(B)/δ(B), is composed of

$$\omega(B) = (\omega_0 + \omega_1 B + \cdots + \omega_{s-1} B^{s-1}) B^b , \tag{10.21}$$

and

$$\delta(B) = (1 - \delta_1 B - \cdots - \delta_r B^r) . \tag{10.22}$$

The value b represents the delay of response in the process. The parameters of ω(B) are similar to the impulse response weights, and the operator δ(B) provides information on how these weights "die out" from the output's "memory".

## The disturbance term

The term $N_t$ represents the time series component of the model involving the error term, at. $N_t$ is called the disturbance term and follows an ARMA model. That is,

$$\phi(B) N_t = \theta(B) a_t , \tag{10.23}$$

or

$$N_t = \{\theta(B) / \phi(B)\} a_t . \tag{10.24}$$

The SCA-GSA product limits the form of $N_t$ to that of a nonseasonal AR(1), AR(2), MA(1) or MA(2) model only. The SCA-UTS product has no such restriction.

Since $N_t$ is that part of the model associated with the noise sequence at, Box and Jenkins (1970, page 362) also refer to the disturbance term as "noise". As a result, and because the letter N is used to represent disturbance, $N_t$ is often confused as actually being the white noise of an ARIMA model, i.e., at. It is important that we realize the distinction between the random errors, $a_t$, and the disturbance term, $N_t$.

### 10.4.2 The identification process for a transfer function model

We can write the complete model in (10.19) as

$$Y_t = C + \{\omega(B) / \delta(B)\} X_t + \{\theta(B) / \phi(B)\} a_t . \tag{10.25}$$

The first step in the implementation of this model is the tentative identification of the orders of the polynomial operators involved. This can prove formidable since the operators can affect each other. Box and Jenkins (1970) outlined a procedure to tentatively identify the orders of these operators. However, this procedure can be tedious and does not extend to more than one input series. A review of this procedure is provided in Section 10.5.

An alternative identification procedure has been proposed by a number of authors, including Box and Jenkins (1970), Liu and Hanssens (1982), Liu and Hudak (1985), and Liu (1987). This procedure can be referred to as the linear transfer function (LTF) method as it employs V(B) directly. In the LTF method, we can simultaneously obtain information related to the transfer function and the disturbance term. We do this by employing representation (10.18) of the transfer function model and utilizing a reasonable approximation for $N_t$. In particular, we will do the following:

(1) Estimate the linear transfer model

$$Y_t = C + (v_0 + v_1 B + v_2 B^2 + \cdots + v_\ell B^\ell) X_t + N_t \tag{10.26}$$

for a sufficiently large $\ell$, and a "reasonable" approximation of $N_t$. For a nonseasonal series (as is the case for the SCA-GSA product), an AR(1) is a useful approximation for $N_t$.

(2) Examine the estimate of the autoregressive parameter in $N_t$. If it is near 1, then we need to appropriately difference the input and output series and repeat step (1) above. In certain situations, it may be appropriate to difference both $Y_t$ and $X_t$ even if the AR parameter is not close to 1.

(3) After any differencing and re-estimation, we should examine the ACF of the residuals of the fitted model. If the ACF does not demonstrate any gross inadequacies, we can tentatively determine orders for the transfer function of model (10.25). If the estimated response weights "cut-off", then we can employ a simple lagged regression model. Otherwise, we may employ a corner table to the estimated response weights (discussed in more detail in *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*).

(4) To obtain a model for the disturbance term we need to obtain

$$\hat{N}_t = Y_t - \hat{C} - (\hat{v}_0 + \hat{v}_1 B + \hat{v}_2 B^2 + \cdots + \hat{v}_\ell B^\ell) X_t \tag{10.27}$$

and employ "standard" identification techniques to determine an ARMA model for $\hat{N}_t$. The SCA System can compute and retain $\hat{N}_t$ of (10.27) after estimation.

We will now illustrate the LTF method in the identification of a transfer function model for our gasoline data.

### 10.4.3 Identification of the gasoline data

Neither the plot of PGAS nor PCRUDE exhibits any seasonal behavior. Hence we will restrict the disturbance to be nonseasonal. We have noted previously that PGAS is nonstationary. Indeed, if we estimate the model

$$\mathrm{PGAS}_t = \mathrm{C} + (\mathrm{v}_0 + \mathrm{v}_1\mathrm{B} + \mathrm{v}_2\mathrm{B}^2 + \cdots + \mathrm{v}_{10}\mathrm{B}^{10})\mathrm{PCRUDE}_t + \{1/(1-\phi\mathrm{B})\}\,\mathrm{N}_t .$$

we will find the estimate of $\phi$ to be close to 1. As a result, we will now specify and estimate the model

$$\nabla\mathrm{PGAS}_t = \mathrm{C} + (\mathrm{v}_0 + \mathrm{v}_1\mathrm{B} + \mathrm{v}_2\mathrm{B}^2 + \cdots + \mathrm{v}_{10}\mathrm{B}^{10})\nabla\mathrm{PCRUDE}_t + \{1/(1-\phi\mathrm{B})\}\,\mathrm{N}_t$$

where $\nabla = (1\text{-}B)$. It will be cumbersome if we must specify all terms of the transfer function in the "usual" scheme, V0+V1*B+V2*B**2+V3*B**3+× × ×+V10*B**10. In addition, the chance for a typographical error in such a specification is great. Fortunately, we can use a shorthand notation for this purpose. We can specify the above transfer function as either

(0 TO 10),

or

(0 TO 10; V0 TO V10).

In the former, we only specify the lag orders in the transfer function. In the latter, we also specify labels in which to store parameter estimates. Here we used the names V0, V1, V2, ..., V10. Since we will eventually want to use these estimated values, the latter specification is used in the following example.

```
-->TSMODEL   GASDATA1.   MODEL IS  PGAS(1) = CONST +     @
-->         (0 TO 10; V0 TO V10)PCRUDE(1) + 1/(1 - P1*B)NOISE.
```

```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- GASDATA2

------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1
  PGAS     RANDOM     ORIGINAL     (1-B  )
                                        1
 PCRUDE    RANDOM     ORIGINAL     (1-B  )
------------------------------------------------------------------------

 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

  1  CONST            CNST      1       0     NONE      .0000
  2    V0     PCRUDE   NUM.      1       0     NONE      .1000
  3    V1     PCRUDE   NUM.      1       1     NONE      .1000
  4    V2     PCRUDE   NUM.      1       2     NONE      .1000
  5    V3     PCRUDE   NUM.      1       3     NONE      .1000
  6    V4     PCRUDE   NUM.      1       4     NONE      .1000
  7    V5     PCRUDE   NUM.      1       5     NONE      .1000
  8    V6     PCRUDE   NUM.      1       6     NONE      .1000
```

```
  9    V7     PCRUDE   NUM.      1      7     NONE      .1000
 10    V8     PCRUDE   NUM.      1      8     NONE      .1000
 11    V9     PCRUDE   NUM.      1      9     NONE      .1000
 12   V10     PCRUDE   NUM.      1     10     NONE      .1000
 13    P1      PGAS    D-AR      1      1     NONE      .1000
```

We now will estimate this model and retain the residuals and estimated disturbance term for future use. The output is edited for presentation purposes.

-->ESTIM GASDATA1. HOLD RESIDUALS(RGAS1), DISTURBANCE(NTGAS1).

```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- GASDATA1

------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1
  PGAS     RANDOM     ORIGINAL     (1-B  )
                                        1
 PCRUDE    RANDOM     ORIGINAL     (1-B  )
------------------------------------------------------------------------

 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT               ERROR  VALUE

  1  CONST             CNST      1      0     NONE    -1.3258    3.2388   -.41
  2    V0     PCRUDE   NUM.      1      0     NONE      .0798     .0844    .95
  3    V1     PCRUDE   NUM.      1      1     NONE      .4453     .0831   5.36
  4    V2     PCRUDE   NUM.      1      2     NONE      .1555     .0828   1.88
  5    V3     PCRUDE   NUM.      1      3     NONE     -.1507     .0828  -1.82
  6    V4     PCRUDE   NUM.      1      4     NONE     -.1173     .0838  -1.40
  7    V5     PCRUDE   NUM.      1      5     NONE     -.0118     .0835   -.14
  8    V6     PCRUDE   NUM.      1      6     NONE      .0264     .0832    .32
  9    V7     PCRUDE   NUM.      1      7     NONE      .0549     .0824    .67
 10    V8     PCRUDE   NUM.      1      8     NONE      .0008     .0822    .01
 11    V9     PCRUDE   NUM.      1      9     NONE      .0159     .0820    .19
 12   V10     PCRUDE   NUM.      1     10     NONE      .0795     .0823    .97
 13    P1      PGAS    D-AR      1      1     NONE      .5719     .1060   5.40


TOTAL SUM OF SQUARES . . . . . .   .265312E+06
TOTAL NUMBER OF OBSERVATIONS . . .          72
RESIDUAL SUM OF SQUARES. . . . . .  .676323E+04
R-SQUARE . . . . . .                      .969
EFFECTIVE NUMBER OF OBSERVATIONS . .        60
RESIDUAL VARIANCE ESTIMATE . . . . .  .112721E+03
RESIDUAL STANDARD ERROR. . . . . .  .106170E+02
```

As a check of our model, we will compute and display the first 12 lags of the ACF of the residual series.

-->ACF RGAS1. MAXLAG IS 12.

```
TIME PERIOD ANALYZED . . . . . . . . . 13  TO    72
NAME OF THE SERIES . . . . . . . . .        RGAS2
EFFECTIVE NUMBER OF OBSERVATIONS . . .         60
STANDARD DEVIATION OF THE SERIES . . .     10.6170
MEAN OF THE (DIFFERENCED) SERIES . . .       .0002
STANDARD DEVIATION OF THE MEAN . . . .      1.3706
T-VALUE OF MEAN (AGAINST ZERO) . . . .       .0001

AUTOCORRELATIONS

  1- 12     .28 -.20 -.40 -.23  .01  .06 -.01 -.07 -.13  .02  .11  .20
  ST.E.     .13  .14  .14  .16  .17  .17  .17  .17  .17  .17  .17  .17
   Q        5.1  7.6 18.3 21.8 21.8 22.1 22.1 22.5 23.7 23.7 24.6 27.6


         -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
           +----+----+----+----+----+----+----+----+----+----+
                                     I
  1    .28                    +     IXXXXX+X
  2   -.20                   + XXXXXI      +
  3   -.40                XXX+XXXXXXI      +
  4   -.23                  + XXXXXXI       +
  5    .01                  +       I       +
  6    .06                  +       IXX     +
  7   -.01                  +       I       +
  8   -.07                  +     XXI       +
  9   -.13                  +    XXXI       +
 10    .02                  +       I       +
 11    .11                  +       IXXX    +
 12    .20                  +       IXXXXX  +
```

We see the lagged correlations are not large. The small spikes at lags 1 and 3 indicate the residuals are not completely serially independent. However, they give us some confidence that the estimates of the model are reasonable. As a result, we can use the estimates of the impulse response weights to determine the transfer function portion of the model. We may use the estimated disturbance term (retained in the workspace in the variable NTGAS1) to determine the orders for $\phi(B)$ and $\theta(B)$.

Since the estimates of $v_1$ through $v_4$ are significant, we will model our series with these four parameters of V(B). We will also include the contemporaneous term vo. In this manner we extend the "regression" model of $Y_t$ on $X_t$, $X_{t-1}$, $X_{t-2}$ and $X_{t-3}$ (see Chapter 9) to one that includes $X_{t-4}$ and serially correlated errors.

In order to model $N_t$ we will compute the first 12 lags of the ACF and PACF of the estimated disturbance term by entering

-->IDEN NTGAS1. MAXLAG IS 12.

```
TIME PERIOD ANALYZED . . . . . . . . . 12  TO    72
NAME OF THE SERIES . . . . . . . . .       NTGAS1
EFFECTIVE NUMBER OF OBSERVATIONS . . .         61
STANDARD DEVIATION OF THE SERIES . . .    12.8418
MEAN OF THE (DIFFERENCED) SERIES . . .      -.1995
STANDARD DEVIATION OF THE MEAN . . . .      1.6442
T-VALUE OF MEAN (AGAINST ZERO) . . . .      -.1213


AUTOCORRELATIONS

          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+
                                     I
  1    .57                    +     IXXXXX+XXXXXXXX
  2   -.01                  +       I       +
  3   -.36                 X+XXXXXXXI       +
  4   -.36                 XXXXXXXXXI        +
  5   -.18                 +   XXXXXI        +
  6   -.08                 +      XXI        +
  7   -.09                 +      XXI        +
  8   -.11                 +     XXXI        +
  9   -.09                 +      XXI        +
 10    .06                +         IX        +
 11    .21                +         IXXXXX    +
 12    .31                +         IXXXXXXXX +


PARTIAL AUTOCORRELATIONS

          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+
                                     I
  1    .57                    +     IXXXXX+XXXXXXXX
  2   -.50              XXXXXXX+XXXXXI     +
  3   -.13                    +  XXXI     +
  4    .00                    +     I     +
  5   -.10                    +   XXI     +
  6   -.18                    + XXXXI     +
  7   -.12                    +  XXXI     +
  8   -.11                    +  XXXI     +
  9   -.11                    +  XXXI     +
 10    .08                    +     IXX   +
 11    .03                    +     IX    +
 12    .13                    +     IXXX  +
```

Since the PACF cuts off after lag 2, we will use an AR(2) model for $N_t$. An MA(1) model could also be considered since the ACF cuts off after lag 1.

### 10.4.4 Specifying and estimating a transfer function model

Since the estimate of the constant term is not significant, we will not include it in our model. We will now specify and estimate the model

$$\text{PGAS}(1) + (v_0 + v_1 B + v_2 B^2 + v_3 B^3 + v_4 B^4)\text{PCRUDE}(1) + 1/(1 - \phi_1 B - \phi_2 B^2)\text{NOISE}.$$

We will use the current estimates of V0 through V4 and P1 as initial estimates of $v_0$ through $v_4$ and $\phi_1$, respectively.

```
-->TSMODEL   NAME IS GASDATA2.   MODEL IS PGAS(1) =       @
-->          (0 TO 4; V0 TO V4)PCRUDE(1) + 1/(1 - P1*B - P2*B**2)NOISE.
```

```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- GASDATA2

------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1
  PGAS     RANDOM     ORIGINAL     (1-B  )
                                        1
 PCRUDE    RANDOM     ORIGINAL     (1-B  )
------------------------------------------------------------------------

 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

  1    V0     PCRUDE   NUM.      1      0     NONE      .0798
  2    V1     PCRUDE   NUM.      1      1     NONE      .4453
  3    V2     PCRUDE   NUM.      1      2     NONE      .1555
  4    V3     PCRUDE   NUM.      1      3     NONE     -.1507
  5    V4     PCRUDE   NUM.      1      4     NONE     -.1173
  6    P1      PGAS    D-AR      1      1     NONE      .5719
  7    P2      PGAS    D-AR      1      2     NONE      .1000
```

We now estimate the model, and will retain the residuals and estimated disturbance term. The residuals are retained for diagnostic purposes and the estimated disturbance term is retained in the event a revision of the model for the disturbance term is required. The output is edited for presentation purposes.

-->ESTIM GASDATA2. HOLD RESIDUALS(RGAS2), DISTURBANCE(NTGAS2).

```
REDUCED CORRELATION MATRIX OF PARAMETER ESTIMATES

        1     2     3     4     5     6     7
  1   1.00
  2    .32  1.00
  3    .     .    1.00
  4   -.36  -.26   .    1.00
  5    .    -.36   .     .32  1.00
  6    .     .     .     .     .    1.00
  7    .     .     .     .     .    -.58  1.00


SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- GASDATA3

-------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1
  PGAS     RANDOM     ORIGINAL     (1-B  )
                                        1
 PCRUDE    RANDOM     ORIGINAL     (1-B  )
-------------------------------------------------------------------------

 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

  1    V0     PCRUDE   NUM.      1      0     NONE      .0828     .0691   1.20
  2    V1     PCRUDE   NUM.      1      1     NONE      .4233     .0725   5.83
  3    V2     PCRUDE   NUM.      1      2     NONE      .1309     .0681   1.92
  4    V3     PCRUDE   NUM.      1      3     NONE     -.1522     .0724  -2.10
  5    V4     PCRUDE   NUM.      1      4     NONE     -.0906     .0690  -1.31
  6    P1      PGAS    D-AR      1      1     NONE      .8681     .1075   8.07
  7    P2      PGAS    D-AR      1      2     NONE     -.5070     .1083  -4.68
TOTAL SUM OF SQUARES . . . . . .265312E+06
TOTAL NUMBER OF OBSERVATIONS . . . .          72
RESIDUAL SUM OF SQUARES. . . . . .530939E+04
R-SQUARE . . . . . .978
EFFECTIVE NUMBER OF OBSERVATIONS . .          65
RESIDUAL VARIANCE ESTIMATE . . . . .816830E+02
RESIDUAL STANDARD ERROR. . . . . .903786E+01
```

The model appears to be relatively good. Based on their t-values, most parameters are significant. In addition, the variance of the residual series is 81.68. In our previous modeling of PGAS, we noted the unexplained variation after an ARIMA model was only 123.68 . As our result, the use of a transfer function model reduces the unexplained variability by an additional 34%.

## Modifying a time series model

Although the above model may be adequate for the data, we note that 3 of the 7 parameters of our model have t-values that are insignificant at the 5% level. Hence, we may be able to achieve a good fit with fewer parameters. We should not simply delete all "insignificant" terms since a low t-value can arise due to correlation between parameters. From the reduced correlation matrix, we see the only correlation of note is between the

estimates of $\phi_1$ and $\phi_2$. We will first reduce the number of parameters by deleting $\omega_0$ and $\omega_4$ from the model.

We can achieve this change by re-specifying the model, or by simply changing only that part of the model that is affected. That is, we can use the TSMODEL paragraph to alter our model by entering

-->TSMODEL GASDATA2. CHANGE (1 TO 3; V1 TO V3)PCRUDE(1).

```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- GASDATA3

------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1
  PGAS     RANDOM     ORIGINAL     (1-B  )
                                        1
 PCRUDE    RANDOM     ORIGINAL     (1-B  )
------------------------------------------------------------------------

 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

  1    V1     PCRUDE   NUM.      1      1     NONE      .4233
  2    V2     PCRUDE   NUM.      1      2     NONE      .1309
  3    V3     PCRUDE   NUM.      1      3     NONE     -.1522
  4    P1      PGAS    D-AR      1      1     NONE      .8681     .1075   8.07
  5    P2      PGAS    D-AR      1      2     NONE     -.5070     .1083  -4.68
```

We see the use of the CHANGE sentence modifies only that part of the model associated with the variable PCRUDE. We now can estimate the model by entering

-->ESTIM GASDATA2. HOLD RESIDUALS(RGAS2).

```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- GASDATA2

------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1
  PGAS     RANDOM     ORIGINAL     (1-B  )
                                        1
 PCRUDE    RANDOM     ORIGINAL     (1-B  )
------------------------------------------------------------------------

 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

  1    V1     PCRUDE   NUM.      1      1     NONE      .3663     .0650   5.63
  2    V2     PCRUDE   NUM.      1      2     NONE      .1297     .0675   1.92
  3    V3     PCRUDE   NUM.      1      3     NONE     -.0944     .0649  -1.45
  4    P1      PGAS    D-AR      1      1     NONE      .8832     .1056   8.36
  5    P2      PGAS    D-AR      1      2     NONE     -.5250     .1061  -4.95
```

```
TOTAL SUM OF SQUARES . . . . . .265312E+06
TOTAL NUMBER OF OBSERVATIONS . . . .          72
RESIDUAL SUM OF SQUARES. . . . . .554205E+04
R-SQUARE . . . . . .977
EFFECTIVE NUMBER OF OBSERVATIONS . .          66
RESIDUAL VARIANCE ESTIMATE . . . . .839705E+02
RESIDUAL STANDARD ERROR. . . . . .916354E+01
```

The results are similar to the previous estimation, but the estimates of $v_2$ and $v_3$ are now insignificant. We should delete these parameters and re-estimate the model, but we will not do so at this time.

If a constant was included in the model, its estimate would be insignificant and we would likely wish to remove it. We can delete a constant term from an existing ARIMA model by including the sentence

DELETE CONSTANT

in the TSMODEL paragraph.

### 10.4.5 Diagnostic checks of a transfer function model

We need to check our transfer function model in much the same manner as an ARIMA model. Our checks may include, but not be limited to

(1) A time plot of the residual series

(2) Examining the ACF of the residual series

(3) Examining the cross correlation function between the residual series and all explanatory (input) series

(4) Outlier analysis of the estimated model

(5) Selected overfitting or underfitting of the "final" model (i.e., refitting the model with additional or fewer parameters)

(6) "Out of sample" check between the forecasts from the model and observed values in the output series. That is, withholding a portion of our data (at the end of the series) from modeling; then examining how well our model forecasts these values.

To obtain a time plot of the residuals we may simply enter

-->TSPLOT RGAS2. SEASONALITY IS 12. SYMBOL IS '*'.

```
  TIME SERIES PLOT FOR THE VARIABLE RGAS2

       ------------+-----------+-----------+-----------+-----------+-----------+--------
 24.00 +                                     *                                           +
       I                           *                                 *                   I
       I                                                                                 I
       I                                                                                 I
 12.00 +                             *                       *           *     *       +
       I                           *   *           *                                     I
       I                 * *                                     **        *             I
       I      *   *                                        **            * *             I
   .00 +       *    ****    ****          *    *   *  *             *        *          +
       I         *          *                        *            *      *       *       I
       I           *             **        *         * **   *                            I
       I        *        *                 *                    *  *        *           I
-12.00 +                          *                       *    *      *          *      +
       I                                  ** *                                         I
       I                               *                                               I
       I                                                                               I
-24.00 +                            *                                                   +
       ------------+-----------+-----------+-----------+-----------+-----------+--------
                  12          24          36          48          60          72
```

No aberrations are apparent in the plot. However, we note that there is less variability in the residuals before t = 24. This corresponds to the period when PGAS was relatively flat.

We can also compute and display 12 lags of the ACF of the residuals. We observe the ACF is clean and the Q statistic does not indicate a bad fit.

-->ACF RGAS2. MAXLAG IS 12.

```
TIME PERIOD ANALYZED . . . . . .  7  TO    72
NAME OF THE SERIES . . . . . .        RGAS2
EFFECTIVE NUMBER OF OBSERVATIONS . . .          66
STANDARD DEVIATION OF THE SERIES . . .      9.1029
MEAN OF THE (DIFFERENCED) SERIES . . .     -1.0528
STANDARD DEVIATION OF THE MEAN . . . .      1.1205
T-VALUE OF MEAN (AGAINST ZERO) . . . .      -.9396
```

```
AUTOCORRELATIONS

 1- 12    -.06  .01 -.10 -.11 -.09 -.11 -.09 -.03 -.08  .15  .11  .20
 ST.E.     .12  .12  .12  .12  .13  .13  .13  .13  .13  .13  .13  .13
  Q         .3   .3  1.0  1.8  2.4  3.4  3.9  4.0  4.5  6.2  7.2 10.5


         -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
           +----+----+----+----+----+----+----+----+----+----+----+
                                      I
 1   -.06                      +   XXI     +
 2    .01                      +     I     +
 3   -.10                      +   XXI     +
 4   -.11                      +  XXXI     +
 5   -.09                      +   XXI     +
 6   -.11                      +  XXXI     +
 7   -.09                      +   XXI     +
 8   -.03                      +    XI     +
 9   -.08                      +   XXI     +
10    .15                      +     IXXXX +
11    .11                     +      IXXX   +
12    .20                     +      IXXXXX +
```

A last diagnostic check that will be performed for this model is a cross correlation between the residual series and the first differenced PCRUDE series. A transfer function model assumes that all variables of the model are statistically independent of one another. In particular our input series, here $(1-B)\text{PCRUDE}_t$, should be independent of the error term $a_t$. The estimates of the error term, the residuals, should thus be independent of $(1-B)\text{PCRUDE}_t$. One measure of independence is to check if there is <u>any</u> type of correlation between these series.

The cross correlation function (CCF) between two series is similar to the ACF of a single series, except we need to distinguish which series is "leading" the other. This results in positive and negative lag terms, one indicating when the first series leads the second and the other when the second series leads the first. To obtain the CCF between $(1-B)\text{PCRUDE}_t$ and $\text{RGAS2}_t$, we will first difference PCRUDE (see Appendix B), then cross correlate the resulting series with RGAS2. We can do this by entering

-->DIFFERENCE PCRUDE. NEW IS LAGCRUDE.

```
                          1
DIFFERENCE ORDERS ARE  (1-B  )
SERIES  PCRUDE  IS DIFFERENCED, THE RESULT IS STORED IN VARIABLE LAGCRUDE
SERIES LAGCRUDE HAS   72  ENTRIES
```

-->CCF LAGCRUDE, RGAS2. MAXLAG IS 12.

```
TIME PERIOD ANALYZED . . . . . .  7  TO    72
NAMES OF THE SERIES  . . . . . .     LAGCRUDE         RGAS2
EFFECTIVE NUMBER OF OBSERVATIONS . . .           66           66
STANDARD DEVIATION OF THE SERIES . . .      18.8668       9.1029
MEAN OF THE (DIFFERENCED) SERIES . . .        .9947      -1.0528
STANDARD DEVIATION OF THE MEAN . . . .       2.3223       1.1205
T-VALUE OF MEAN (AGAINST ZERO) . . . .         .4283       -.9396
CORRELATION       BETWEEN    RGAS2     AND LAGCRUDE IS    .15

CROSS CORRELATION BETWEEN LAGCRUDE(T)  AND    RGAS2(T-L)

  1- 12     .09 -.03  .04 -.03 -.08  .02  .03 -.04 -.05  .06  .02  .04
  ST.E.     .12  .13  .13  .13  .13  .13  .13  .13  .13  .13  .13  .14

CROSS CORRELATION BETWEEN    RGAS2(T)  AND LAGCRUDE(T-L)

  1- 12     .03 -.02 -.09 -.11  .00 -.00  .05  .03  .11  .11 -.03 -.10
  ST.E.     .12  .13  .13  .13  .13  .13  .13  .13  .13  .13  .13  .14


          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+
                                     I
-12   -.10                    +   XXXI      +
-11   -.03                    +     XI      +
-10    .11                    +      IXXX   +
 -9    .11                     +     IXXX  +
 -8    .03                     +     IX    +
 -7    .05                     +     IX    +
 -6    .00                     +     I     +
 -5    .00                     +     I     +
 -4   -.11                     +  XXXI     +
 -3   -.09                     +   XXI     +
 -2   -.02                     +    XI     +
 -1    .03                     +     IX    +
  0    .15                     +     IXXXX +
  1    .09                     +     IXX   +
  2   -.03                     +    XI     +
  3    .04                     +     IX    +
  4   -.03                     +    XI     +
  5   -.08                     +   XXI     +
  6    .02                     +     IX    +
  7    .03                     +     IX    +
  8   -.04                     +    XI     +
  9   -.05                     +    XI     +
 10    .06                    +      IX     +
 11    .02                    +      IX     +
 12    .04                    +      IX     +
```

We do not observe any significant cross correlations in either direction. This is as would be hoped.

### 10.4.6 Forecasting from a transfer function model

We may now wish to forecast from our estimated model. For an ARIMA model involving only one variable, we were able to use the FORECAST paragraph directly for this purpose. However, in the transfer function framework, the forecasts of the output variable are also dependent on forecasts (or known values) of any input variables.

If the values of our input variable(s) are not yet known, as in this case, we must forecast values for both our input and output variables. Values forecasted for our input variables will be incorporated into our forecast of our output variable according to the estimated transfer function equation.

Forecasting the input variable, PCRUDE, requires the construction of an ARIMA model for PCRUDE. Using the techniques described in Sections 1 and 2, we can build a model for PCRUDE. Although the modeling is not shown here, we find that an ARIMA(1,1,0) model is adequate for PCRUDE. The information related to this model is stored in the SCA workspace under the name LEADING.

The forecasts listed are of PCRUDE and PGAS, not the first-differenced series. More information regarding forecasting is given in Section 10.5.2.

First, we will forecast 12 observations for PCRUDE using the model LEADING. We will append the forecasts to the end of PCRUDE by including the JOIN sentence in the FORECAST paragraph.

```
-->FORECAST   LEADING.   NOFS IS 12.   JOIN.

 ---------------------------------
  12 FORECASTS, BEGINNING AT   72
 ---------------------------------

  TIME     FORECAST    STD. ERROR    ACTUAL IF KNOWN

    73     545.5876      15.8189
    74     546.2542      28.6953
    75     546.5965      40.1707
    76     546.7722      50.2886
    77     546.8624      59.2615
    78     546.9088      67.3100
    79     546.9325      74.6186
    80     546.9448      81.3317
    81     546.9510      87.5597
    82     546.9542      93.3869
    83     546.9559      98.8778
    84     546.9568     104.0827
```

Now we will forecast 12 observations for PGAS. Within the FORECAST paragraph we also supply information that this ARIMA model for the input variable PCRUDE is stored under the name LEADING. If we had more than one input variable, we would need to supply this information for each input series.

```
-->FORECAST   GASDATA2.   NOFS IS 12.   IARIMA IS PCRUDE(LEADING).

---------------------------------
 12 FORECASTS, BEGINNING AT   72
---------------------------------

 TIME    FORECAST   STD. ERROR   ACTUAL IF KNOWN

   73    518.9961      9.1635
   74    514.2167     20.3795
   75    512.3013     30.2705
   76    513.3510     36.8937
   77    515.3419     41.4689
   78    516.5790     45.1812
   79    516.6418     48.7041
   80    516.0556     52.2750
   81    515.5090     55.8345
   82    515.3361     59.2272
   83    515.4714     62.3725
   84    515.6822     65.2942
```

## 10.5 Other Time Series Topics

This section provides a brief overview of topics related to time series analysis or the execution of SCA paragraphs related to time series. Much of the material presented in this section can be considered "advanced" or of occasional use. As a consequence, this section can be skipped, and selected topics can be referenced as necessary. The material presented, and the section containing it are:

| Section | Topic |
|---|---|
| 10.5.1 | Use of differencing operators |
| 10.5.2 | Plotting forecasts with confidence limits |
| 10.5.3 | Identification procedures for transfer function models |
| 10.5.4 | Missing data |

### 10.5.1 Use of differencing operators

Sometimes we will find it necessary to use differencing operators to achieve stationarity. Differencing within the usual ARIMA(p,d,q) model is in the form

$$(1-B)^{d} \tag{10.28}$$

In fact, a wider array of stationary inducing operators is available to us. The SCA System extends the representation of (10.28) to that of

$$(1-B^{d1})(1-B^{d2})(1-B^{d3})\cdots(1-B^{dk}) \tag{10.29}$$

where d1, d2, ... , dk are referred to as differencing orders. The representation in (10.29) gives us greater flexibility in the type of differencing we want to use. However, this flexibility can lead to some "quirks" in the specification of "d" when this value is greater than 1.

For example, suppose we wish to analyze a twice-differenced series. Here we want to analyze $(1-B)^2$ of some series. Suppose we specify

DFORDER IS 2

in the ACF, PACF, IDEN, or EACF paragraph; or we include the differencing operator (2) within the MODEL sentence of the TSMODEL paragraph. The SCA System will interpret it as single differencing with order 2 and will base its computations using the differencing operator $(1-B^2)$.

In order to specify the operator (1-B)2, we need to specify

DFORDER IS 1, 1

in an "identification" paragraph, or the operator (1,1) in the MODEL sentence of the TSMODEL paragraph. Although this may seem a bit complicated for the specification of d in a (p,d,q) model, a (p,d,q) model cannot handle the differencing operator

$$(1-B)(1-B^4)(1-B^{12})$$

while it can be handled directly in the SCA System.

We can also difference data outside the SCA paragraphs presented in this chapter. The DIFFERENCE paragraph (see Appendix B) can be used to edit data through differencing. However, use of this paragraph is not advisable in a time series analysis.

### 10.5.2 Plotting forecasts with confidence limits

It is often valuable to plot forecasted values of a time series on the same time frame as that of the original series. In addition, plotting the confidence limits of the forecasts provides us with information on the potential variability of these forecasts.

In order to plot forecasts, we need to create forecasts (using the FORECAST paragraph), possibly modify series using analytic functions or data editing capabilities (see Appendices A and B), and then plot the resultant data (see Chapter 3). As an example, suppose we want to plot 12 forecasted values of PGAS from the model we derived in Section 10.2. In addition, we want to display the 90% confidence limits of the forecasts. The estimated model is in the SCA workspace under the label PGASAR. To forecast the series and retain the forecasts and their standard errors we can enter

```
-->FORECAST   PGASAR.   NOFS ARE 12.      @
-->         HOLD  FORECASTS(PGASFCST), STD_ERR(PGASFERR).

 ---------------------------------
  12 FORECASTS, BEGINNING AT   72
 ---------------------------------

  TIME    FORECAST   STD. ERROR   ACTUAL IF KNOWN

    73    518.7768     11.1210
    74    513.5532     23.8924
    75    510.9069     34.4027
    76    511.0724     41.4428
    77    512.5140     45.9570
    78    513.7328     49.2627
    79    514.1275     52.2500
    80    513.8880     55.3141
    81    513.4793     58.4805
    82    513.2279     61.6039
    83    513.2009     64.5591
    84    513.2992     67.3165
```

Our forecasts are now in the variable PGASFCST and the standard errors are in PGASFERR. The upper and lower confidence limits for a 90% confidence interval can be computed using the following two analytic statements (see Appendix A)

```
UPPER = PGASFCST + 1.645*PGASFERR
LOWER = PGASFCST - 1.645*PGASFERR
```

We can plot the forecasts and limits directly by using the MTSPLOT paragraph (see Chapter 3) and entering

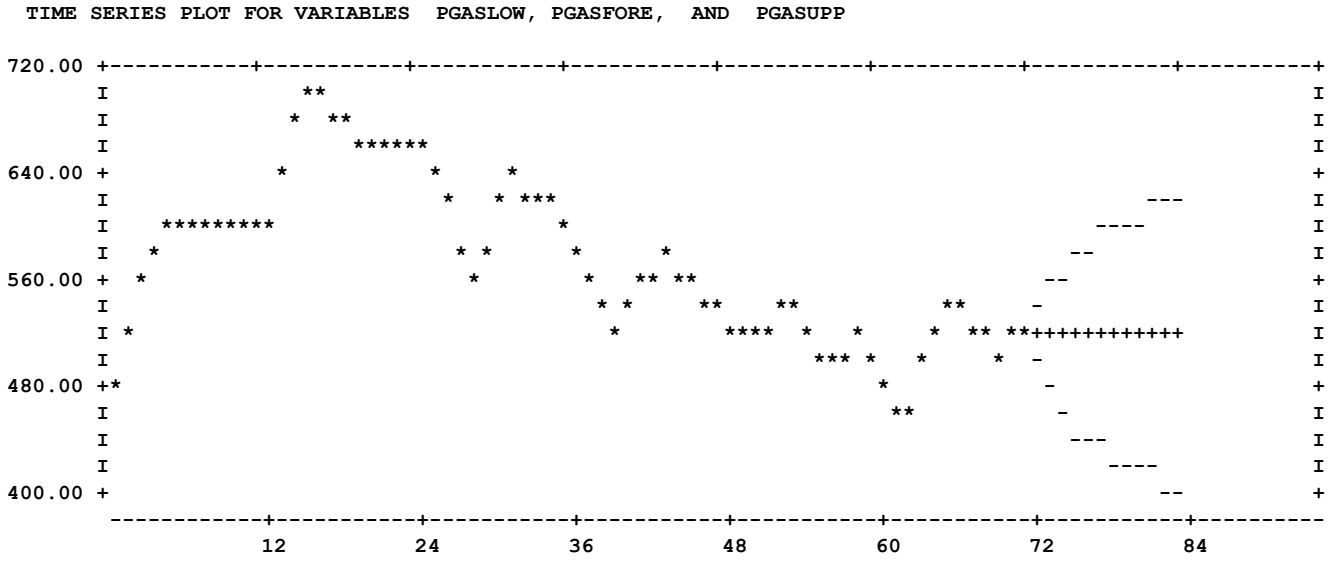
```
-->MTSPLOT   LOWER, PGASFCST, UPPER.   SYMBOLS ARE  '-', '+', '-'.
```

The symbols '-', '+', and '-' are specified here to represent the lower confidence limit, forecasted value, and upper confidence limit, respectively. The plot is not displayed here since we would instead like to plot the forecasts on the same frame as our original series, PGAS. To do this we need to append each of the above three variables to PGAS. We can accomplish this through the JOIN paragraph (see Appendix B).

```
-->JOIN   PGAS, LOWER.   NEW IS PGASLOW.
-->JOIN   PGAS, UPPER.   NEW IS PGASUPP.
-->JOIN   PGAS, PGASFCST.   NEW IS PGASFORE.
```

We may now employ MTSPLOT as before. Here we will specify a seasonal marking on the axis. In addition we will use the symbol '+' to represent our forecasts and '-' to represent the 90% confidence limits. When all variables occupy the same position (i.e., over the original span of PGAS), MTSPLOT will display the value as '*'.

```
-->MTSPLOT   PGASLOW, PGASFORE, PGASUPP.   SYMBOLS ARE '-', '+', '-'.  @
-->          SEASONALITY IS 12.
```




## 10.5.3 Identification procedures for transfer function models

In this section we will review two procedures related to the identification of transfer function models. First, we briefly summarize a "classical" procedure for the modeling of a transfer function. Next, we present a procedure to determine the form of a transfer function, given its impulse response (or transfer function) weights.

### Classical modeling of transfer functions

Box and Jenkins (1970) proposed a procedure for the identification and fitting of the single-input, single-output transfer function model

$$Y_t = C + V(B)X_t + N_t \tag{10.30}$$

or

$$Y_t = C + \{\omega(B)/\delta(B)\}X_t + N_t \tag{10.31}$$

The components of this model are discussed in Section 10.4. The modeling procedure proposed for this model is sequential in nature and employs many steps. Its rationale is to first identify the form of the transfer function, then determine the form of the disturbance term after accounting for the transfer function component. Because of its sequential approach, any error at the beginning stages of modeling affects remaining stages.

As in the approach outlined in Section 10.4, estimates of the impulse response weights (i.e., of V(B)) are the key values required for the identification of the form of the transfer function $\omega(B)/\delta(B)$. Box and Jenkins (1970) noted that if the input series, $X_t$, was white noise then the cross correlation function between the input and output series would be

proportional to the true transfer function weights. Since the input series in not noise, an artificial white noise series is created and (10.30) is modified accordingly to reflect it. This portion of the procedure is known as pre-whitening and has been the cause of substantial confusion in the model building process.

The steps involved in this "classical" modeling procedure can be outlined as follows:

(1) Build an ARIMA model for the input series and retain the residual series. This stage is known as pre-whitening the input series.

(2) Use the above estimated ARIMA model of the input series and apply it to the output series. Retain the resultant "residual series". This stage is known as filtering the output series. The series obtained is a transformation of (10.30), and typically is not consonant with a white noise process.

(3) Compute the cross correlation function (CCF) between the pre-whitened input series and filtered output series. The CCF is proportional to the transfer function weights.

(4) Determine an appropriate $\omega(B)/\delta(B)$ for the transfer function weights.

(5) Use the transfer function determined above in a model, but assume the disturbance term is a white noise process, i.e., let $N_t = a_t$. Estimate this model and retain the residual series from it.

(6) Use the residual series of (5) to identify an ARIMA model for $N_t$.

(7) Re-specify and estimate the resultant model.

As we see, the "classical" procedure outline above is more complicated than the LTF method described and used in Section 10.4.

## Implementation of the "classical" method

The "classical" method described above can be implemented in the SCA System. We will not illustrate this method with a complete example, but we will outline the commands necessary for its implementation.

(1) ARIMA model building. Use of ACF, PACF, IDEN, EACF, TSMODEL and ESTIM as illustrated in Sections 10.1 and 10.2. Residuals should be retained from ESTIM.

(2) Pre-whitening the output series. Use the FILTER paragraph, specifying the ARIMA model name of the input series and the name of the output series. A variable name is specified to retain the transformed series.

(3) Use the CCF paragraph to compute the cross correlation function between the pre-whitened series. If desired, the computed CCF can be retained.

(4) Identifying the form of the transfer function. We can use either the methods described in Box and Jenkins (1970) to determine the form, or a corner table method proposed by Liu and Hanssens (1982) on the weights obtained in step (3) (The corner table method and the SCA-UTS paragraph associated with the method are described in *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*). Note data editing of the retained CCF is necessary (see Appendix B).

(5) Specify and estimate the transfer function with a white noise disturbance term (using TSMODEL). Estimate the model and retain the residual series.

(6) Use standard identification tools on the above residual series to determine an appropriate ARIMA model for the disturbance term.

(7) Use the CHANGE sentence of the TSMODEL paragraph to modify the disturbance term. Re-estimate the resultant model.

Diagnostic checks for the estimated model are as described in Section 10.4.5 .

### 10.5.4 Missing data

We can model series that contain coded missing data, but we must decide how missing data are to be handled. If missing data are present in a time series, and we do not recode the data, then the SCA System will proceed as follows. The SCA System will note the first occurrence of non-missing data and the next occurrence of a missing data point. Only data within this span are used in the calculation of a paragraph.

If we want to use the entire span of data, then we must replace all missing data by some "appropriate" values. We can do this using an SCA data editing paragraph (see Appendix B) or an analytic statement (see Appendix A). "Appropriate" values for missing data might consist of

(1) the average of all observations in a stationary series,

(2) the average of two adjacent observations,

(3) the average of all observations with the same periodicity for nonstationary series that exhibits a distinct seasonal component but no trend, or

(4) the average of two adjacent observations with the same periodicity for a nonstationary series that exhibits a distinct seasonal component and trend.

The PATCH paragraph can be used to accomplish the above (described in Appendix B).

# SUMMARY OF THE SCA PARAGRAPHS IN CHAPTER 10

This section provides a summary of those SCA paragraphs employed in this chapter. The syntax for many paragraphs is presented in both a brief and full form. The brief display of the syntax contains the most frequently used sentences of a paragraph, while the full display presents all possible modifying sentences of a paragraph. In addition, special remarks related to a paragraph may also be presented with the description.

Many paragraphs have more capabilities than those presented in this chapter or that may be inferred by the syntax presented for the paragraphs. More complete information for those paragraphs is presented in *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*. Paragraphs having additional capabilities will be noted.

Each SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

The paragraphs to be explained in this summary are IDEN, ACF, PACF, EACF, CCF, FILTER, CORNER, TSMODEL, ESTIM, FORECAST, and OUTLIER. Paragraphs are categorized by their function, model identification, estimation, forecasting and diagnostic checking. The SCA-GSA product employs only a partial set of the capabilities of several time series related paragraphs. For a complete listing of all capabilities of all paragraphs, please refer to *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*.

Legend (see Chapter 2 for further explanation)

v : variable or model name
i : integer
r : real value
w : keyword

# Paragraphs Related to Model Identification

## A. Univariate Model Identification

## ACF Paragraph

The ACF paragraph is used to compute the sample autocorrelation function of a time series. The paragraph also displays some descriptive statistics including the sample mean, standard deviation and a t-statistic on the significance of a constant term. The sample ACF may also be computed within the IDEN paragraph.

## Syntax for the ACF Paragraph

### Brief syntax

```
ACF   VARIABLE IS v.              @
      DFORDERS ARE i1, i2, --- .  @
      MAXLAG IS i.

Required sentence:  VARIABLE
```

### Full syntax

```
ACF   VARIABLE IS v.              @
      DFORDERS ARE i1, i2, --- .  @
      MAXLAG IS i.                @
      SPAN IS i1, i2.             @
      HOLD ACF(v), SDACF(v).

Required sentence:  VARIABLE
```

## Sentences Used in the ACF Paragraph

**VARIABLE sentence**

The VARIABLE sentence is used to specify the name of the series to be analyzed.

**DFORDERS sentence**

The DFORDERS sentence is used to specify the orders of differencing to be applied on the series when differencing is the stationary inducing transformation being used. For example, the order associated with the differencing operator (1-B) is 1 and that of $(1-B^{12})$ is 12. If a power of an operator is to be used (for example, $(1-B)^2$) then the differencing order must be repeated the appropriate number of times (in this example, 1, 1). Default is none.

**MAXLAG sentence**

The MAXLAG sentence is used to specify the maximum order of sample ACF to be computed. Default is 36.

**SPAN sentence**

The SPAN sentence is used to specify the span of time indices, from i1 to i2, for which the data will be analyzed. Default is the maximum span available for the series.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. Default is that none of the values of the above statistics will be retained after the paragraph is executed. The values that may be retained are:

ACF : the sample ACF of the series
SDACF : the standard deviations of the sample ACF for the series

## PACF Paragraph

The PACF paragraph is used to compute the sample partial autocorrelation function of a time series. The paragraph also displays some descriptive statistics including the sample mean, standard deviation and a t-statistic on the significance of a constant term. The sample PACF may also be computed within the IDEN paragraph.

## Syntax for the PACF Paragraph

### Brief syntax

```
PACF  VARIABLE IS v.              @
      DFORDERS ARE i1, i2, --- .  @
      MAXLAG IS i.

Required sentence:  VARIABLE
```

### Full syntax

```
PACF  VARIABLE IS v.              @
      DFORDERS ARE i1, i2, --- .  @
      MAXLAG IS i.                @
      SPAN IS i1, i2.             @
      HOLD PACF(v), SDPACF(v).

Required sentence:  VARIABLE
```

## Sentences Used in the PACF Paragraph

**VARIABLE sentence**

The VARIABLE sentence is used to specify the name of the series to be analyzed.

**DFORDERS sentence**

The DFORDERS sentence is used to specify the orders of differencing to be applied on the series when differencing is the stationary inducing transformation being used For example, the order associated with the differencing operator (1-B) is 1 and that of $(1-B^{12})$ is 12. If a power of an operator is to be used (for example, $(1-B)^2$) then the differencing order must be repeated the appropriate number of times (in this example, 1, 1). Default is none.

**MAXLAG sentence**

The MAXLAG sentence is used to specify the maximum order of sample PACF to be computed. Default is 36.

**SPAN sentence**

The SPAN sentence is used to specify the span of time indices, from i1 to i2, for which the data will be analyzed. Default is the maximum span available for the series.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. Default is that none of the values of the above statistics will be retained after the paragraph is executed. The values that may be retained are:

PACF : the sample PACF of the series
SDPACF : the standard deviations of the sample PACF for the series

## IDEN Paragraph

The IDEN paragraph can be used when performing the tentative identification of a series or in the diagnostic checking of a residual series. The paragraph is used to co-ordinate the computation of the sample ACF (autocorrelation function) and PACF (partial autocorrelation function) of a univariate time series. If only the sample ACF is desired, it may be computed using the ACF paragraph by itself; similarly for the sample PACF. All three paragraphs also display some descriptive statistics including the sample mean, standard deviation and a t-statistic on the significance of a constant term.

## Syntax for the IDEN Paragraph

### Brief syntax

```
IDEN  VARIABLE IS v.                @
      DFORDERS ARE i1, i2, --- .    @
      MAXLAG IS i.

Required sentence:  VARIABLE
```

### Full syntax

```
IDEN  VARIABLE IS v.                @
      DFORDERS ARE i1, i2, --- .    @
      MAXLAG IS i.                  @
      SPAN IS i1, i2.               @
      HOLD ACF(v), PACF(v), SDACF(v), SDPACF(v).

Required sentence:  VARIABLE
```

## Sentences Used in the IDEN Paragraph

**VARIABLE sentence**

The VARIABLE sentence is used to specify the name of the series to be analyzed.

**DFORDERS sentence**

The DFORDERS sentence is used to specify the orders of differencing to be applied on the series when differencing is the stationary-inducing transformation being used. For example, the order associated with the differencing operator (1-B) is 1 and that of $(1-B^{12})$ is 12. If a power of an operator is to be used (for example, $(1-B)^2$) then the differencing order must be repeated the appropriate number of times (in this example, 1, 1). Default is none.

**MAXLAG sentence**

The MAXLAG sentence is used to specify the maximum order of sample ACF and PACF to be computed. Default is 36.

**SPAN sentence**

The SPAN sentence is used to specify the span of time indices, i1 to i2, for which the data will be analyzed. Default is the maximum span available for the series.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. Default is that none of the values of the above statistics will be retained after the paragraph is executed. The values that may be retained are:

ACF : the sample ACF of the series
PACF : the sample PACF of the series
SDACF : the standard deviations of the sample ACF for the series
SDPACF : the standard deviations of the sample PACF for the series

## EACF Paragraph

The EACF paragraph is used to compute the sample extended autocorrelation function. The paragraph produces a table useful in determining the order of a mixed stationary or nonstationary ARMA process.

## Syntax for the EACF Paragraph

### Brief syntax

```
EACF  VARIABLE IS v.        @
      DFORDERS ARE i1, i2, --- .

Required sentence:  VARIABLE
```

### Full syntax

```
EACF  VARIABLE IS v.                    @
      DFORDERS ARE i1, i2, --- .        @
      MAXLAG IS AR(i1), MA(i2).         @
      SPAN IS i1, i2.

Required sentence:  VARIABLE
```

## Sentences Used in the EACF Paragraph

**VARIABLE sentence**

The VARIABLE sentence is used to specify the name of the series to be analyzed.

**DFORDERS sentence**

The DFORDERS sentence is used to specify the orders of differencing to be applied on the series when differencing is the stationary inducing transformation being used. For example, the order associated with the differencing operator (1-B) is 1 and that of $(1-B^{12})$ is 12. If a power of an operator is to be used (for example, $(1-B)^2$) then the differencing order must be repeated the appropriate number of times (in this example, 1, 1). The default is none.

**MAXLAG sentence**

The MAXLAG sentence is used to specify the maximum autoregressive (AR) and moving average (MA) orders to be computed and displayed. The default maximum AR order is 6 and maximum MA order is 12.

**SPAN sentence**

The SPAN sentence is used to specify the span of time indices, i1 to i2, for which the data will be analyzed. Default is the maximum span available for the series.

## B. Transfer Function Identification Paragraphs

This section gives the syntax for the FILTER, CCF, and CORNER paragraphs. These paragraphs are useful in the identification of the impulse response weights and transfer function in transfer function analysis.

## FILTER Paragraph

The FILTER paragraph is used to filter a time series to a new series according to a specified time series model. A discussion of the use of filtering is found in Section 10.5.4. A special case of this procedure is known as pre-whitening. Common filtering for all input and output series is also useful when the linear transfer function (LTF) method is employed.

### Syntax for the FILTER Paragraph

```
FILTER      MODEL model-name.                    @
            OLD-SERIES  ARE  v1, v2, --- .       @
            NEW-SERIES  ARE  v1, v2, --- .

Required sentence:  MODEL
```

### Sentences Used in the FILTER Paragraph

**MODEL sentence**

The MODEL sentence is used to specify the label (name) of a previously defined univariate time series model that will be used to filter the variable(s) specified in the OLD sentence.

**OLD sentence**

The OLD sentence is used to specify the names of the series to be filtered. If this sentence is omitted, the output variable of the univariate model specified in the MODEL sentence will be filtered.

**NEW sentence**

The NEW sentence is used to specify the variable(s) where the filtered series are stored. The number of variable(s) in this sentence must be the same as that in the OLD sentence if specified. The default are the variable(s) of the OLD sentence.

## CCF Paragraph

The CCF paragraph is used to compute the cross correlation function between two specified time series. The paragraph also displays for each series some descriptive statistics including the sample mean, standard deviation and a t-statistic on the significance of a constant term.

## Syntax for the CCF Paragraph

### Brief syntax

```
CCF   VARIABLES ARE v1, v2.      @
      DFORDERS ARE i1, i2, --- . @
      MAXLAG IS i.

Required sentence: VARIABLE
```

### Full syntax

```
CCF   VARIABLES ARE v1, v2.      @
      DFORDERS ARE i1, i2, --- . @
      MAXLAG IS i.               @
      SPAN IS i1, i2.            @
      HOLD CCF(v), SDCCF(v).

Required sentence: VARIABLE
```

## Sentences Used in the CCF Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the names of the series to be analyzed. Two series names must be specified.

**DFORDERS sentence**

The DFORDERS sentence is used to specify the orders of differencing to be applied on each series when differencing is the stationary-inducing transformation being used. For example, the order associated with the differencing operator (1-B) is 1 and that of $(1-B^{12})$ is 12. If a power of an operator is to be used (for example, $(1-B)^2$) then the differencing order must be repeated the appropriate number of times (in this example, 1, 1). Default is none.

**MAXLAG sentence**

The MAXLAG sentence is used to specify the maximum order of CCF to be computed. Default is 36.

**SPAN sentence**

The SPAN sentence is used to specify the span of time indices, from i1 to i2, for which the data will be analyzed. Default is the maximum span available for the series.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. Default is that none of the values of the above statistics will be retained after the paragraph is used. The values that may be retained are:

CCF : the sample CCF of the series
SDCCF : the standard deviations of the sample CCF of the series

## CORNER Paragraph

The CORNER paragraph is used to compute the corner table for a sequence of impulse response (transfer function) weights.

## Syntax for the CCF Paragraph

```
CORNER      VARIABLE IS v.      @
            SIZE IS NROWS(i1), NCOLS(i2).

Required sentence: VARIABLE
```

## Sentences Used in the CORNER Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the name of the variable that contains the impulse response weights from which the corner table will be computed.

**SIZE sentence**

The SIZE sentence is used to specify the number of rows (NROWS) and columns (NCOLS) for the corner table. Assuming the number of impulse response weights is k, the default value for NROWS is (k+2)/2 and NCOLS is k/2.

# Paragraphs Related to Model Specification and Estimation

## TSMODEL Paragraph
**(ARIMA analysis and single input transfer function modeling)**

The TSMODEL paragraph is used to specify or modify a univariate ARIMA model or a single input transfer function model. The paragraph has more capabilities than the syntax shown below. More complete information can be found in Chapter 3 of *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*.

## Syntax for the TSMODEL Paragraph

For each model specified in a TSMODEL paragraph, a distinguishing label or name must also be given. A number of different models may be specified, each having a unique name, and subsequently employed at a user's discretion. Moreover, the label also enables the information contained under it to be modified.

### Brief syntax

```
TSMODEL   NAME IS model-name.        @
          MODEL IS "model".

Required sentence:  NAME
```

### Full syntax (ARIMA analysis)

```
TSMODEL   NAME IS model-name.        @
          MODEL IS "model".          @
          SHOW./NO SHOW.             @
          CHECK./NO CHECK.           @
          ROOTS./NO ROOTS.           @
          SIMULATION./NO SIMULATION.

Required sentence:  NAME
```

**Full syntax (Single input transfer function modeling)**

```
TSMODEL    NAME IS model-name.          @
           MODEL IS "model".  @
           CHANGE "components of a model".     @
           SHOW./NO SHOW.     @
           CHECK./NO CHECK.   @
           ROOTS./NO ROOTS.   @
           SIMULATION./NO SIMULATION.

Required sentence:  NAME
```

**Sentences Used in the TSMODEL Paragraph**
**(ARIMA analysis and single input transfer function modeling)**

**NAME sentence**

The NAME sentence is used to specify a unique label (name) for the model specified in the paragraph. This label is used to refer to this model in both the ESTIM and FORECAST paragraphs or if the model is to be modified.

**MODEL sentence**

The MODEL sentence is used to specify the univariate Box-Jenkins ARIMA model or the univariate time series model.

**CHANGE sentence**

The CHANGE sentence is used to modify components of an existing model. This sentence is applicable to transfer function models only.

**SHOW sentence**

The SHOW sentence is used to display a summary of the specified model. Default is SHOW. The summary includes series name, differencing (if any), span for data, parameter labels (if any) and current values for parameters.

**CHECK sentence**

The CHECK sentence is used to check whether all roots of the AR, MA, and denominator polynomials lie outside the unit circle. Default is NO CHECK.

**ROOTS sentence**

The ROOTS sentence is used to display all roots of the AR, MA and denominator polynomials. Default is NO ROOTS.

**SIMULATION sentence**

The SIMULATION sentence is used to specify that the model will be used for simulation purposes. Ordinarily this sentence is not specified. See Chapter 12 for more details. The default is NO SIMULATION.

## ESTIM Paragraph

The ESTIM paragraph is used to control the estimation of the parameters of an ARIMA model or transfer function model. The paragraph has more capabilities than the syntax shown below. More complete information can be found in Chapter 3 of *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*.

## Syntax of the ESTIM Paragraph

### Brief syntax

```
ESTIM     MODEL v.      @
          HOLD RESIDUALS(v).

Required sentence:  MODEL
```

### Full syntax

```
ESTIM  MODEL v.                                              @
       METHOD  IS  w.                                        @
       STOP-CRITERIA  ARE  MAXIT(i), LIKELIHOOD(r1),         @
                ESTIMATE(r2).                                @
       SPAN  IS  i1, i2.                                     @
       HOLD RESIDUALS(v), FITTED(v), VARIANCE(v).

 Required sentence:  MODEL
```

## Sentences Used in the ESTIM Paragraph
**(ARIMA analysis and single input transfer function modeling)**

**MODEL sentence**

The MODEL sentence is used to specify the label (name) of the model to be estimated. The label must be one specified in a previous TSMODEL paragraph.

**METHOD sentence**

The METHOD sentence is used to specify the likelihood function used for model estimation. The keyword may be CONDITIONAL for the "conditional" likelihood or EXACT for the "exact" likelihood function. See Section 10.3.2 for a discussion of these two likelihood functions. Default is CONDITIONAL.

**STOP sentence**

The STOP sentence is used to specify the stopping criterion for nonlinear estimation. The argument, i, for the keyword MAXIT specifies the maximum number of iterations (default is i=10); the argument, r1, for the keyword LIKELIHOOD specifies the value of the relative convergence criterion on the likelihood function (default is r1=0.0001); and the argument, r2, for the keyword ESTIMATE specifies the value of the relative convergence criterion on the parameter estimates (default is r2=0.001). Estimation iterations will be terminated when the relative change in the value of the likelihood function or parameter estimates between two successive iterations is less than or equal to the convergence criterion, or if the maximum number of iterations is reached.

**SPAN sentence**

The SPAN sentence is used to specify the span of time indices, from i1 to i2, for which the data will be analyzed. Default is the maximum span available for the series.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. Default is that none of the values of the above statistics will be retained after the paragraph is used. The values that may be retained are:

| | |
|---|---|
| RESIDUAL | : the residual series |
| FITTED | : the one step-ahead forecasts (fitted values) of the series |
| VARIANCE | : variance of the noise |
| DISTURBANCE | : the disturbance series of the model (transfer function model only) |

## FORECAST Paragraph

The FORECAST paragraph is used to compute the forecast of future values of a time series based on a specified ARIMA model or transfer function model. The FORECAST paragraph requires the current estimate of the variance σ2 to compute standard errors of forecasts. The variance for the estimated model is always stored internally during the execution of the ESTIM paragraph, but the internal estimate is overwritten at each subsequent execution of a ESTIM paragraph. Hence if the variance of an estimated model is desired for the computation of standard errors of the forecasts but σ2 has not been stored in a user specified variable, then the FORECAST should be executed before invoking the ESTIM paragraph for the estimation of another model.

The FORECAST paragraph has more capabilities than described below. More complete information can be found in Chapter 3 of *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*.

## Syntax of the FORECAST Paragraph

### Brief syntax

```
FORECAST   MODEL v.                @
           NOFS ARE i1, i2, --- .  @
           ORIGINS ARE i1, i2, ---.

Required sentence:  MODEL
```

### Full syntax (ARIMA analysis)

```
FORECAST   MODEL v.                                  @
           NOFS ARE i1, i2, --- .                    @
           ORIGINS ARE i1, i2, --- .                 @
           JOIN. /NO JOIN.                           @
           HOLD FORECASTS(v1,v2,---), STD_ERRS(v1,v2,---).

Required sentence:  MODEL
```

**Full syntax (Single input transfer function modeling)**

```
FORECAST   MODEL v.                                          @
           NOFS ARE i1, i2, --- .                            @
           ORIGINS ARE i1, i2, --- .                         @
           IARIMA ARE v1(model-name) ,                       @
                  v2(model-name), ---.                       @
           JOIN. /NO JOIN.                                   @
           HOLD FORECASTS(v1,v2,---), STD_ERRS(v1,v2,---).
```

Required sentence: **MODEL**

## Sentences Used in the FORECAST Paragraph

**MODEL sentence**

The MODEL sentence is used to specify the label (name) of the model for the series to be forecasted. The label must be one specified in a previous TSMODEL paragraph.

**NOFS sentence**

The NOFS sentence is used to specify for each time origin the number of time periods ahead for which forecasts will be generated. The number of arguments in this sentence must be the same as that in the ORIGINS sentence. The default is 24 forecasts for each time origin.

**ORIGINS sentence**

The ORIGINS sentence is used to specify the time origins for forecasts. The default is one origin, the last observation.

**IARIMA sentence**

The IARIMA sentence is used to specify the label associated with the ARIMA model of each stochastic input series of a transfer function model. The variable name of each input series must be listed and in parentheses the name (label) for its Box-Jenkins ARIMA model.

**JOIN sentence**

The JOIN sentence is used to specify that the forecasts calculated should be appended to the variable of the model relative to the specified origin. If more than one origin is specified only the last will be used. The default is NO JOIN.

**HOLD sentence**

The HOLD sentence is used to specify those values computed for particular functions to be retained in the workspace. Only those statistics desired to be retained need be named. Values are placed in the variable named in parentheses. Default is that none of the values of the above statistics will be retained after the paragraph is used. The values that may be retained are:

FORECASTS : forecasts for each corresponding time origin
STD_ERRS : standard errors of the forecasts at the last time origin

## Paragraphs Related to Diagnostic Checking an Estimated Model

The ACF (and CCF) paragraphs are useful in the diagnostic checking of a fitted ARIMA (and transfer function) model. In addition time plots of the residuals (see TSPLOT in Chapter 3) are useful in spotting any anomalies in the residual series.

The OUTLIER paragraph (of the SCA-UTS product) is also useful in diagnostic checking a fitted model. This paragraph detects and identifies various outlying observations. More information on this paragraph can be found in *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*.

## REFERENCES

Abraham, B., and Ledolter, J. (1983). *Statistical Methods for Forecasting*, New York: Wiley.

Box, G.E.P., and Jenkins, G.H. (1970). *Time-Series Analysis*: Forecasting and Control, San Francisco: Holden Day.

Chang, I., Tiao, G.C., and Chen, C. (1988). "Estimation of Time Series Parameters in the Presence of Outliers." *Technometrics 30*: 193-204.

*Commodity Year Book* (1986). New York: Commodity Research Bureau.

Cryer, J.D. (1986). *Time Series Analysis*, Boston: Duxbury Press.

Granger, C.W.J., and Newbold, P. (1987). *Forecasting Economic Time Series*, New York: Academic Press.

Hillmer, S.C., and Tiao, G.C. (1979). "Likelihood Function of Stationary Multiple Autoregressive Moving Average Models". *Journal of the American Statistical Association 74*: 652-660.

Liu, L.-M. (1987). "Sales forecasting Using Multi-Equation Transfer Function Models". *Journal of Forecasting 6*: 223-238.

Liu, L.-M., and Hanssens, D.M. (1982). "Identification of Multiple-Input Transfer Function Models". *Communications in Statistics A 11*: 297-314.

Liu, L.-M., and Hudak, G.B. (1985). "Unified Econometric Model Building Using Simultaneous Transfer Function Equations." *Time Series Analysis: Theory and Practice 7* (O.D. Anderson ed.), Amsterdam: North-Holland.

Ljung, G.M., and Box, G.E.P. (1978). "On a Measure of Lack of Fit in time Series Models." *Biometrika 65*: 297-304.

Pankratz, A. (1983). *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*, New York: Wiley.

Slutsky, E. (1937). "The Summation of Random Causes as the Source of Cyclic Processes." *Econometrica 5*: 105 (translation of original 1927 Russian paper).

Tsay, R.S. and Tiao, G.C. (1984). "Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Non-stationary ARMA Models". *Journal of the American Statistical Association 79*: 84-96.

Vandaele, W. (1983). *Applied Time Series Analysis and Box-Jenkins Models*, New York: Academic Press.

Wold, H. (1938). *A Study in the Analysis of Stationary Time Series* (2nd ed. 1954), Uppsala: Almquist and Wicksell.

Yule, G.U. (1921). "On the Time-Correlation Problem with Special Reference to the Variate Difference Correlation Method." *Journal of the Royal Statistical Society 84*: 497-526.

Yule, G.U. (1927). "On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wölfer's Sunspot Numbers." *Philosophical Transactions of the Royal Society of London, Series A, 226*: 267-298.

# CHAPTER 11

# NONPARAMETRIC STATISTICS

Nonparametric statistics refer to a collection of statistical methods for hypothesis testing without making specific assumptions about some aspects of the underlying distribution. Such hypothesis tests encompass location, randomness, goodness of fit and dependency, among others. Properties of nonparametric methods have been considered by many authors including Conover (1980), Gibbons (1985), Siegel (1956), Walsh (1962), Kraft and van Eeden (1968), Noether (1967), Bradley (1968), Hollander and Wolfe (1973), Lehmann (1975), Mosteller and Rourke (1973), and Marascuilo and McSweeney (1977).

Nonparametric methods are sometimes referred to as distribution-free methods. However, we should be aware that these methods are not necessarily free of distributional assumptions. Many tests assume a random sampling hypothesis (see Box, Hunter and Hunter 1978) and can be adversely affected by the presence of dependent errors. In fact, many of these tests are approximations of randomization tests.

The NPAR paragraph is used to access the wide range of nonparametric methods available in the SCA System. A summary of these methods is given in Section 1 below. Only one nonparametric test is performed in any single execution of the paragraph.

## 11.1 Available Nonparametric Tests

Nonparametric tests are usually divided according to the assumptions made on the type of data (or sample) being analyzed, the hypothesis that may be tested, and the way in which observations may be distinguished from one another. Tests are divided in this chapter according to the types of data samples that are assumed. These types are: one sample (that of a single variable), samples of two variables (further divided into the categories of related or independent), and samples of several variables (either related or independent).

Tests may be categorized further according to the measurement scale assumed for the data being analyzed. There are two types of scale used in the NPAR paragraph: nominal, in which observations may be separated according to category, and ordinal, in which observations can be ordered from smallest to largest. It is important to recognize the measurement scale assumed for a test, as this creates a "hierarchy" for the applicability of tests. That is, if only an ordinal scale of measurement is required of a data set, then tests assuming the nominal scale may also be employed. The converse is not true. The minimum scale of measurement for each test is also noted on the chart below. Information regarding the type of hypothesis being tested is listed in specific sections.

| Section | Type of data sample | Name of test (measurement scale required) |
|---|---|---|
| 11.2 | one sample | Binomial (nominal)<br>Runs (nominal)<br>Chi-square (nominal)<br>Kolmogorov-Smirnov (ordinal) |
| 11.3 | two independent samples | Median (ordinal)<br>Mann-Whitney U (ordinal)<br>Kolmogorov-Smirnov (ordinal) |
| 11.4 | several independent samples | Median (ordinal)<br>Kruskal-Wallis H (ordinal) |
| 11.5 | two related samples | Sign (ordinal)<br>Wilcoxon (ordinal)<br>Kendall's rank correlation (ordinal)<br>Spearman's rank correlation (ordinal) |
| 11.6 | several related samples | Cochran Q (nominal)<br>Friedman two-way ANOVA (ordinal)<br>Kendall coefficient of concordance (ordinal) |

## 11.2 Tests Using One Random Sample of a Single Variable

We first consider nonparametric tests that use a single random sample. We will illustrate each test with at least one example. References are provided for examples discussed in texts.

| Section | Test | Hypothesis test for | Minimum measurement scale |
|---|---|---|---|
| 11.2.1 | Binomial test | mean | nominal |
| 11.2.2 | Runs test | randomness | nominal |
| 11.2.3 | Chi-square tests | goodness of fit or independence | nominal |
| 11.2.4 | Kolmogorov-Smirnov test | goodness of fit | ordinal |

### 11.2.1 Binomial test

The binomial test is used in tests of hypothesis involving a probability, p, that an observation will occur in one of two dichotomous classes. The test involves the probability for the population; hence is considered a test of the mean. For some specified constant, p* (where $0 \le p^* \le 1$), the null hypothesis is p = p*. One of the following is the alternative hypothesis:

Two-tailed: $p \neq p^*$
One-tailed: $p > p^*$ or $p < p^*$

**Example: Learning methods**

To illustrate the binomial test we will consider the following example from Siegel (1956, pages 39-40). In an experiment regarding stress, 18 individuals were taught two different methods to tie the same knot. Half were taught one method first, half taught the other first. Later the 18 were required to tie the knot. It was then observed what method was used, the first or second method learned. It is hypothesized that the individuals would revert to using the first method learned. Hence if p is the probability of using the first method, we will test if p* = .5. The data,

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2

are stored in the SCA workspace under the label LEARNING. We will use the value 1.5 as a cut-point for the data (i.e., a value that dichotomizes the data). We could, of course, use any value between 1 and 2 for this purpose. We will test the hypothesis that p* = 0.5.

To request the test we enter

```
-->NPAR  METHOD IS BINOMIAL.  VARIABLE IS LEARNING.  @
-->      CUT-POINT IS 1.5.   PROPORTION IS 0.5.
```

```
BINOMIAL TEST FOR THE VARIABLE   LEARNING
TOTAL NUMBER OF TRIALS. . . . . .          18
NUMBER OF CASES BELOW THE CUT-POINT (1ST GROUP) . . .          16
NUMBER OF CASES ABOVE THE CUT-POINT (2ND GROUP) . . .           2
PROBABILITY OF THE FIRST GROUP. . . . . .    0.50000
CUMULATIVE PROBABILITY (FROM LEFT-HAND SIDE). . . . .    0.99993
CUMULATIVE PROBABILITY (FROM RIGHT-HAND SIDE) . . . .    0.00066
POINT PROBABILITY . . . . . .    0.00058
```

We are provided with the total number of trials (cases, observations) and the number of cases in each group occurring above and below the cut-point. We also are provided with information related to the test itself including the proportion, p*, of the null hypothesis, resultant left-hand and right-hand cumulative probabilities and the resultant point probability.

The resultant cumulative probability from the left-hand side of the distribution denotes the probability of observing less than or equal to the number of cases recorded in the first group, assuming the actual proportion is p*. The resultant cumulative probability from the right-hand side of the distribution denotes the probability of observing greater than or equal to the number of cases recorded in the first group, assuming the proportion p*. In a one-tail test, one of these values (depending upon the direction of the alternative hypothesis) should be compared with the significance level, α. For a two-tailed test, the smaller of these two values should be compared with α/2. We see that, regardless of the alternative, the null hypothesis is rejected at the 1% level. That is, under stress the individuals tend to revert to the first method learned.

In addition to the cumulative probabilities displayed, the probability of exactly the number of cases in the first group assuming the proportion p* is displayed. This information is given since the binomial distribution is discrete and occasionally the probability of the exact occurrence is useful in a significance test.

We need to supply a "cut-point" for the data. However, we are not required to supply a value for p* (specified in the PROPORTION sentence). If a proportion is not specified the value 0.5 is used for p*.

### 11.2.2 Runs test

If a data set consists of a sequence of successive observations, we may employ the runs test as a test for randomness. We will dichotomize the observations into two groups and use the number of runs observed for test purposes. A run is a succession of identical observations (i.e., those of the same group) bounded by observations of a different type (either a different value or the beginning or end of the data).

**Example: Student scores**

To illustrate the runs test, we will consider an example from Siegel (1956, pages 54-56). The data consist of the aggression scores in young children. An experimenter was concerned if any bias was introduced in the manner in which the experiment was conducted. This bias would display itself as a lack of random behavior. The recorded scores

31 23 36 43 51 44 12 26 43 75 2 3 15 18 78 24 13 27 86 61 13 7 6 8

are stored in the SCA workplace under the label SCORE. If we use the value 24.5, the median of the data, to dichotomize the data into two groups (above 24.5, A; below 24.5, B) we can describe the above data as

A B A A A A B A A A B B B B A B B A A A B B B B

We have 12 A and 12 B values with 5 runs of A's and 5 runs of B's, 10 total runs. We can use the RUNS paragraph to ascertain the likelihood of obtaining such a result if the chance of being an A or B is completely random.

-->NPAR METHOD IS RUNS. VARIABLE IS SCORE. CUT-POINT IS 24.5.

```
NONPARAMETRIC  RUNS TEST FOR THE VARIABLE        SCORE

TOTAL NUMBER OF CASES . . . . . .                  24
CUT-POINT VALUE . . . . . .                     24.5000
NUMBER OF CASES IN THE FIRST  GROUP . . .          12
NUMBER OF CASES IN THE SECOND GROUP . . .          12
NUMBER OF RUNS. . . . . .                          10
Z-SCORE OF NORMAL DISTRIBUTION. . . . . .       -1.0436
ONE-TAILED PROBABILITY(NORMAL APPROX.). .        0.1483
TWO-TAILED PROBABILITY(NORMAL APPROX.). .        0.2967

ONE-TAILED  EXACT PROBABILITY . . . . . .        0.1504
TWO-TAILED  EXACT PROBABILITY . . . . . .        0.3009
```

We are provided with basic summary information: the total number of cases or observations, the cut-point used, the number of cases in each group, and the total number of runs. We also obtain the following test information: a computed standard normal value (i.e., Z-score) when a normal approximation is used, its associated (one-tailed and two-tailed) probability values, and exact probability values, when an exact distributional value is computed for small sample sizes (under 25). If the sample size is not small, exact values are not computed. The appropriate probability level (depending on alternative) should be compared to the significance level, $\alpha$, for the test.

We see that the hypotheses of randomness cannot be rejected at the 10% level for either a one-tailed or two-tailed test. We also see that the normal approximation is a relatively good approximation for this size of data (24 observations).

## Example: Box office queue

As a second example of the runs test, an example of the arrangement of men and women in a ticket line is used. This example is also from Siegel (1956, pages 56-58). The data are binary, 0 indicates a woman and 1 a man. The data, stored in the SCA workspace under the label SEX are

```
1 0 1 0 1 1 1 0 0 1 0 1 0 1 0 1 1 1 1 0 1 0 1 0 1 1
0 0 0 1 0 1 0 1 0 1 1 0 1 1 0 1 1 1 1 0 1 0 1 1
```

The data are already in two groups. We will use 0.5 as a cut-off value, although any value between 0 and 1 can be used. Since there are 50 observations, only the normal approximation is employed in the test.

```
-->NPAR  METHOD IS RUNS.  VARIABLE IS SEX.  CUT-POINT IS 0.5.

NONPARAMETRIC  RUNS TEST FOR THE VARIABLE          SEX

TOTAL NUMBER OF CASES . . . . . .                   50
CUT-POINT VALUE . . . . . .                      0.5000
NUMBER OF CASES IN THE FIRST  GROUP . . .           20
NUMBER OF CASES IN THE SECOND GROUP . . .           30
NUMBER OF RUNS. . . . . .                           35
Z-SCORE OF NORMAL DISTRIBUTION. . . . . .        2.9794
ONE-TAILED PROBABILITY(NORMAL APPROX.). .         .0014
TWO-TAILED PROBABILITY(NORMAL APPROX.). .         .0029
```

The hypothesis of randomness is rejected at the 5% level against either a one-tailed or two-tailed alternative. Hence the order of men and women in the line queue does not appear to be random.

### 11.2.3 Chi-square tests

Chi-square tests are usually employed to test how close the number of occurrences of an event come to the number we would "expect" to occur. The test may be used as a goodness-of-fit test of whether grouped data follow some specified distribution function (so that the expected number of occurrences per group can be calculated). A second way a chi-square test may be employed as a test of independence in an r x c contingency table (see Chapter 6).

#### Goodness of fit: GOODNESS

We will present two examples of the goodness-of-fit chi-square test. The first is from Siegel (1956, p.44-46). Here we study whether the post position a horse has at the beginning of a race affects its chance of winning. The data under study are of the number of wins that occurred from each of 8 post positions during a racing season. The number of wins are

29 19 18 25 17 10 15 11

respectively. These are stored in the SCA workspace under the level OBSPOST. If the chance of winning is unaffected by post position, we would expect an equal number of wins, 18, per position. The variable EXPPOST consists of 8 data entries, all equal to 18. We need to specify both variables in the following:

-->NPAR METHOD IS GOODNESS. VARIABLE IS OBSPOST. EXPECTED IS EXPPOST.

```
CHI-SQUARE TEST FOR GOODNESS-OF-FIT BETWEEN     OBSPOST AND  EXPPOST

 CASE   OBSERVED  EXPECTED  RESIDUAL  CHI-SQUARE

    1     29.00     18.00     11.00      6.72
    2     19.00     18.00      1.00      0.06
    3     18.00     18.00       0.0       0.0
    4     25.00     18.00      7.00      2.72
    5     17.00     18.00     -1.00      0.06
    6     10.00     18.00     -8.00      3.56
    7     15.00     18.00     -3.00      0.50
    8     11.00     18.00     -7.00      2.72

TOTAL NUMBER OF CASES . . . . . .          8
CHI-SQUARE STATISTIC. . . . . .    16.3333
DEGREES OF FREEDOM. . . . . .          7
SIGNIFICANCE LEVEL  . . . . . .     0.0222
```

We are provided with a summary of how the $\chi^2$ statistic is computed. For each case (category), we have the observed and expected value, the "residual" (observed - expected), and the component of the $\chi^2$ sum (square of the residual divided by expected). The computed $\chi^2$ value, 16.33, is significant at the 5% level. Hence post position could be a factor in winning.

As a second example we compare the observed and expected totals of the number of 20 randomly generated values that are recorded in four distinct but equally long intervals. The data are used in Conover (1980, pages 191-192). The observed values in the four classes are 8, 4, 3 and 5. If the data has been randomly generated, we expect 5 observations per class. The data are in the SCA workspace under the labels OBSCLASS and EXPCLASS, respectively.

-->NPAR METHOD IS GOODNESS. VAR IS OBSCLASS. EXPECTED IS EXPCLASS.

```
CHI-SQUARE TEST FOR GOODNESS-OF-FIT BETWEEN     OBSCLASS AND   EXPCLASS

 CASE   OBSERVED  EXPECTED  RESIDUAL  CHI-SQUARE

    1      8.00      5.00      3.00      1.80
    2      4.00      5.00     -1.00      0.20
    3      3.00      5.00     -2.00      0.80
    4      5.00      5.00      0.00      0.00

TOTAL NUMBER OF CASES . . . . . .          4
CHI-SQUARE STATISTIC. . . . . .     2.8000
DEGREES OF FREEDOM. . . . . .          3
SIGNIFICANCE LEVEL  . . . . . .     0.4235
```

We have no reason to reject randomness at the 5% level (or any level under 42%)

## Chi-square test for independence: INDEP

To illustrate the use of a chi-square test for independence in an r x c contingency table, we have simulated a 3 x 4 table. The matrix is stored in the SCA workspace under the label VTAB. The data of this matrix are

| | | | |
|---|---|---|---|
| 3.0 | 4.0 | 12.0 | 5.0 |
| 4.0 | 16.0 | 16.0 | 6.0 |
| 1.0 | 7.0 | 7.0 | 3.0 |

We can test for row-column independence by entering

-->NPAR METHOD IS INDEP. VARIABLE IS VTAB.

```
CHI-SQUARE TEST FOR INDEPENDENCE OF THE CONTINGENCY TABLE:     VTAB
-------------------------------------------------
        |       1       2       3       4  TOTALS
-------------------------------------------------
       1|    3.00    4.00   12.00    5.00   24.00
EXPECTED|    2.29    7.71   10.00    4.00
        |
       2|    4.00   16.00   16.00    6.00   42.00
EXPECTED|    4.00   13.50   17.50    7.00
        |
       3|    1.00    7.00    7.00    3.00   18.00
EXPECTED|    1.71    5.79    7.50    3.00
        |
-------------------------------------------------
 TOTALS |    8.00   27.00   35.00   14.00   84.00
-------------------------------------------------

NUMBER OF ROWS. . . . . .          3
NUMBER OF COLUMNS . . . . . .      4
CHI-SQUARE STATISTIC. . . . . .    3.9818
DEGREES OF FREEDOM. . . . . .      6
SIGNIFICANCE LEVEL. . . . . .      0.6791
```

We obtain cell and row-column information for the matrix. The cell information consists of the actual cell value and the expected value if the rows and columns are independent. The $\chi^2$ value for the table, with its significance level, is then presented.

We see that the $\chi^2$ value is not significant unless our $\alpha$ level is at least 68%. Hence we cannot reject the hypothesis of independence.

### 11.2.4 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (also known as the Kolmogorov goodness-of-fit test) is used to test the degree of agreement between a set of observed values and some specified distribution. The test is either one-sided or two-sided. In a two-sided test, the hypothesis that all observed data comes exactly from a specified distribution is tested versus the alternative that at least one observation does not. In a one-sided test, the null hypothesis is that the cumulative distribution function of the observed values is uniformly higher (or lower) than a

specified cumulative distribution function. This hypothesis is tested versus the alternative that for at least one value the inequality is reversed. The test statistic in either case is based upon the greatest absolute distance between the observed and assumed cumulative distribution functions.

We can use the Kolmogorov-Smirnov test in the NPAR paragraph for any of the following distributions:

U(a, b) : uniform distribution between a and b

$N(\mu, \sigma^2)$ : normal distribution with mean $\mu$ and variance $\sigma^2$

POISSON($\lambda$) : Poisson distribution with parameter $\lambda$

We ordinarily supply the distribution, and associated parameters, to use in the test. If no distribution is specified, then a uniform distribution is used with a and b assumed to be the minimum and maximum values of the specified variable.

To illustrate the Kolmogorov-Smirnov tests, we will use a random sample used by Conover (1980, p.348-349). The data

0.621 0.503 0.203 0.477 0.710 0.581 0.329 0.480 0.554 0.382

are stored in the workspace under the label UVAR. It is hypothesized that the underlying distribution is uniform (0,1). We specify the necessary information by entering

```
-->NPAR METHOD IS KS1. VARIABLE IS UVAR. DISTRIBUTION IS U(0,1).

KOLMOGOROV-SMIRNOV TEST FOR THE DISTRIBUTION OF THE VARIABLE       UVAR

THEORETICAL DISTRIBUTION: UNIFORM WITH
    LEFT  ENDPOINT . . . . . .     0.0000
    RIGHT ENDPOINT . . . . . .     1.0000

NUMBER OF CASES. . . . . .            10
KOLMOGOROV-SMIRNOV Z VALUE . . . . .  0.9171
TWO-TAILED P VALUE . . . . . .        0.3696
```

We obtain basic summary information of the data and the reference distribution. Also calculated and displayed is the Kolmogorov-Smirnov (z) value and its (two-tailed) probability value. For this data the uniform (0,1) distribution is supported through the 36% significance level.

## 11.3 Tests Involving Two Independent Samples

We will now consider nonparametric tests involving observations from two independent (random) samples. Each test is illustrated with at least one example. Examples are discussed in cited references.

The tests discussed in this section are

| Section | Test | Hypothesis test for | Minimum measurement scale |
|---|---|---|---|
| 11.3.1 | Median test | location (median) | ordinal |
| 11.3.2 | Mann-Whitney test | location (mean) | ordinal |
| 11.3.3 | Kolmogorov-Smirnov test | identical distributions | ordinal |

### 11.3.1 Median test (two-sample)

The median test is used to determine whether two samples come from populations having the same median. The test is a special application of a 2 x 2 contingency table (see Chapter 6) that has fixed marginal totals. We can extend the test to one of quantiles.

In the two-sample case, a 2 x 2 contingency table is constructed. The values of the first row are the number of observations of each sample that are less than or equal to the grand median of the two samples. The second row of the table consists of the number of observations of each sample that are greater than the grand median. If both samples have the same median then both populations should have approximately the same number of observations above the median. Clearly, we can extend the test to that for quantiles if we specify a value to be used in place of the median. An appropriate 2 x 2 table can be generated using this value to dichotomize samples.

**Example: Illness anxiety scores**

To illustrate the median test for two samples, we consider the following example from Siegel (1956, p.112-115). Child rearing data concerning the explanation (or non-explanation) of illnesses and resultant anxiety are used. The data stored in the workspace in the variable ABSENT represent anxiety scores in societies that do not provide oral explanations of illnesses. The data in PRESENT represent anxiety scores in societies that do provide oral explanation of illnesses. There are 16 observations in ABSENT and 23 in PRESENT. The data are listed in Table 1.

**Table 1 Anxiety scores of illnesses**

| *ABSENT* | | *PRESENT* | | |
|---|---|---|---|---|
| 13 | 8 | 10 | 15 | 13 |
| 12 | 8 | 10 | 15 | 12 |
| 12 | 7 | 10 | 15 | 12 |
| 10 | 7 | 8 | 14 | 12 |
| 10 | 7 | 8 | 14 | 12 |
| 10 | 7 | 6 | 14 | 11 |
| 10 | 7 | 17 | 13 | 11 |
| 9 | 6 | 16 | 13 | |

We obtain the median test for these data by entering

```
-->NPAR  METHOD IS MEDIAN.  VARIABLES ARE ABSENT, PRESENT.
```

```
TWO-SAMPLE MEDIAN TEST

 GROUP --    ABSENT   PRESENT
 LE MEDIAN       13         8
 GT MEDIAN        3        15

 TOTAL NUMBER OF CASES . . . . . .        39
 MEDIAN. . . . . .                   11.0000
 CHI-SQUARE STATISTIC. . . . . .      6.4350
 SIGNIFICANCE LEVEL. . . . . .        0.0112
 FISHER'S EXACT TEST (ONE-TAILED PROBABILITY). . .    0.0049
 FISHER'S EXACT TEST (TWO-TAILED PROBABILITY). . .    0.0082
```

We are provided with the total number of observations (cases); the "median", used to dichotomize the variables, and the resultant 2 x 2 contingency table. Since we did not specify a "median" value (by the inclusion of a MEDIAN sentence), the median of the combined samples is used.

We are provided with the computed chi-square value, and its significance level, for the resultant table. Also displayed are the significance levels associated with Fisher's exact test. We see that we can reject the hypothesis that the medians of ABSENT and PRESENT are the same. The test statistic, 6.435, is significant at the 5% level, and is "almost" significant at the 1% level. We are led to conclude the median anxiety score is higher when oral explanations of illnesses are given than when they are not.

### 11.3.2 Mann-Whitney test

The Mann-Whitney test (also known as the Mann-Whitney U test) is used to determine whether two samples come from populations having the same distribution function. The test is based on the pooled sample of the individual observations. Pooled values are then ranked. If both samples come from populations with the same distribution function, then the observations of one sample should not have ranks uniformly larger (or smaller) than the ranks

of the observations of the other sample. The test is usually two-tailed; that is, the alternative assumption is that there is some difference in the underlying population distribution function. The test can be one-tailed if the alternative assumption is that the location (mean) of one of the populations is significantly greater (or less) than the other.

The test is based on the statistic U, the sum of the number of observations of one group preceding each observation of the other. A normal approximation for the distribution of this sum is used for large samples (i.e., one in which the number of observations in one sample is at least 20). Exact distributions are possible for the sum of small samples.

## Example: Fitness data

To illustrate the Mann-Whitney test, we will consider a set of data examined by Conover (1980, pages 218-220). The fitness scores of 48 boys of a senior high school were obtained. Twelve boys lived on farms, the rest lived in towns. Data are listed in Table 2, and are stored in the SCA workspace under the labels FARM and TOWN, respectively.

**Table 2 Fitness scores**

| Boys who live on farms FARM | | Boys who live in towns TOWN | | | | | |
|---|---|---|---|---|---|---|---|
| 14.8 | 10.6 | 12.7 | 16.9 | 7.6 | 2.4 | 6.2 | 9.9 |
| 7.3 | 12.5 | 14.2 | 7.9 | 11.3 | 6.4 | 6.1 | 10.6 |
| 5.6 | 12.9 | 12.6 | 16.0 | 8.3 | 9.1 | 15.3 | 14.8 |
| 6.3 | 16.1 | 2.1 | 10.6 | 6.7 | 6.7 | 10.6 | 5.0 |
| 9.0 | 11.4 | 17.7 | 5.6 | 3.6 | 18.6 | 1.8 | 2.6 |
| 4.2 | 2.7 | 11.8 | 5.6 | 1.0 | 3.2 | 5.9 | 4.0 |

We can test the hypothesis that the two groups are equal in fitness scores, against the alternative they are not, by entering

```
-->NPAR  METHOD IS MWU.  VARIABLES ARE FARM, TOWN.
```

```
NONPARAMETRIC MANN-WHITNEY U-TEST FOR FARM AND TOWN
NUMBER OF CASES IN THE SMALLER GROUP  . . . . . .          12
NUMBER OF CASES IN THE LARGER  GROUP  . . . . . .          36
VALUE OF THE U-STATISTIC. . . . . .              189.0000

LARGE SAMPLE NORMAL SCORE FOR U-STATISTIC . . . . .      -0.6431
```

We obtain the U statistic value of 189 that, for large samples, has an approximate normal score (i.e., Z-value) of -0.6431. We can now refer to a standard normal table for an approximate significance level. Since the value is well under 2 (in absolute value), we would not reject the null hypothesis at the 5% level (or higher).

**Example: Learning study of rats**

As a second example, we consider data from Siegel (1956, pages 118-121) of a study of whether rats would generalize learned imitation when placed in a new situation. Five of nine rats were trained to follow a leader in a maze in order to obtain food. The mice were then placed in a maze where following a leader would enable them to avoid shocks. Levels of performance were given to each rat.

Scores for rats who learned to follow are stored in the SCA workspace under the label FOLLOW. Scores for those who acted "randomly" are in RANDOM. The data of FOLLOW are 78, 64, 75, 45 and 82. The data of RANDOM are 110, 70, 53 and 51. To calculate the U statistic for these variables, we enter

```
-->NPAR  METHOD IS MWU.  VARIABLES ARE FOLLOW, RANDOM.

NONPARAMETRIC MANN-WHITNEY U-TEST FOR FOLLOW AND RANDOM

NUMBER OF CASES IN THE SMALLER GROUP  . . . . . .          4
NUMBER OF CASES IN THE LARGER  GROUP  . . . . . .          5
VALUE OF THE U-STATISTIC  . . . . . .          9.0000

THE NUMBER OF CASES IN THE LARGER GROUP IS LESS THAN 20.
NORMAL APPROXIMATION IS INAPPROPRIATE. PLEASE CONSULT AN
APPROPRIATE TABLE FOR MEASURE OF SIGNIFICANCE
```

Although we obtain a value of the U statistic, we are not provided with a normal score (nor a significance level). The normal approximation is not valid for small samples (as noted above). In such a case we must consult a table of critical values (e.g., Siegel, 1956). Here we would find the statistic is significant at the 1% level for a one-tailed test. Hence, it appears that previous training was beneficial.

### 11.3.3 Kolmogorov-Smirnov two-sample test

The Kolmogorov-Smirnov two-sample test is used to test whether two independent samples, of possibly different sample sizes, came from populations with the same distribution. The test is similar to the one-sample test (see Section 11.2.4) except no distribution is specified. Here one sample is used as the reference distribution for the other. The test is either one-sided or two-sided. In a two-sided test the hypothesis that all data came from populations with the same distribution (or the same population) is tested versus the alternative that there is at least one difference. In a one-sided test the null hypothesis is that the cumulative distribution function of one sample is uniformly higher (or lower) than that of the other sample. This hypothesis is tested versus the alternative that for at least one value the inequality is reversed. The test statistic in either the one-sided or two-sided case is based upon the greatest absolute distance between the observed cumulative distribution functions.

## Example: Error rate data

To illustrate the Kolmogorov-Smirnov two-sample test, we consider data from Siegel (1956, pages 129-131). Here the percentage of errors made by 10 seventh graders on the first half of a series of material they had learned is compared with the percentage of errors made by 10 eleventh graders on the first half of a series of material they had learned. It is suspected that eleventh graders should make fewer errors as they have developed better ways to learn. The data is listed in Table 3 and stored in the SCA workspace under the labels SEVENTH and ELEVENTH, respectively.

**Table 3 Percentage of errors made on material learned previously**

| 7th graders SEVENTH | | 11th graders ELEVENTH | |
|---|---|---|---|
| 39.1 | 48.7 | 35.2 | 29.1 |
| 41.2 | 55.0 | 39.2 | 41.8 |
| 45.2 | 40.6 | 40.9 | 24.3 |
| 46.2 | 52.1 | 38.1 | 32.4 |
| 48.4 | 47.6 | 34.4 | 32.6 |

To test the hypothesis that these two variables have the same distribution we simply enter

-->NPAR METHOD IS KS2. VARIABLES ARE SEVENTH, ELEVENTH.

```
KOLMOGOROV-SMIRNOV TEST BETWEEN THE DISTRIBUTION OF
THE VARIABLES  SEVENTH AND ELEVENTH

NUMBER OF OBSERVATION IN THE 1ST VARIABLE . . . .          10
NUMBER OF OBSERVATION IN THE 2ND VARIABLE . . . .          10
KOLMOGOROV-SMIRNOV Z VALUE. . . . . .      1.5652
TWO-TAILED P VALUE. . . . . .      0.0149
```

The Kolmogorov-Smirnov Z-value (computed from the greatest absolute distance between the observed cumulative distribution functions) for this data is 1.5652 and has an associated two-tailed significance level of 0.0149. As a result we can reject the hypothesis of the same distribution for both groups at, approximately, the 1% level. That is, we conclude eleventh graders make proportionately fewer errors than seventh graders on material learned previously.

## 11.4 Tests Involving Several Independent Samples

We can now extend our testing to two or more independent (random) samples. Two nonparametric tests are available to us, the median test and the Kruskal-Willis test. At least one example is provided for each test. Examples are discussed in more detail in the cited references.

As a summary of these tests we have

| Section | Test | Hypothesis test for | Minimum measurement scale |
|---|---|---|---|
| 11.4.1 | Median test | location (median) | ordinal |
| 11.4.2 | Kruskal-Wallis test | location (mean) or identical distributions | ordinal |

### 11.4.1 Median test (k-sample)

The k-sample median test is used to determine whether several samples come from populations having the same median. The test is an extension of the two-sample median test (see Section 11.3.1), and may be further extended to be a test of quantiles. The test is a special application of a 2 x k contingency table (see Chapter 6) that has fixed marginal totals. In the k-sample case, a 2 x k contingency table is constructed. The values of the first row are the number of observations of each sample that are less than or equal to the grand median of the k samples. The second row of the table consists of the number of observations of each sample that are greater than the grand median. If all samples have the same median, then all populations should have approximately the same number of observations above the median. Other 2 x k tables can be constructed for a test of quantiles if we specify a value to be used in place of the median.

**Example: Education data**

Data from Siegel (1956, p.180-184) are used to illustrate the k-sample median test. Data representing the interest level in their children's education (as measured by the number of meetings with a teacher) are recorded for 44 mothers. The data are grouped according to the highest grade completed by the mother: 8th grade, 10th grade, 12th grade, some college, college graduate and graduate school. Data are stored in the labels EDUCA1 through EDUCA6, respectively; and are listed in Table 4.

**Table 4 Education & interest level data for mothers based on highest grade completed by mother**

| | | |
|---|---|---|
| 8th Grade | (EDUCA1) | 4 3 0 7 1 2 0 3 5 1 |
| 10th Grade | (EDUCA2) | 2 4 1 6 3 0 2 5 1 2 1 |
| 12th Grade | (EDUCA3) | 2 0 4 3 8 0 5 2 1 7 6 5 1 |
| Some college | (EDUCA4) | 9 4 2 3 |
| College | (EDUCA5) | 2 4 5 2 |
| Graduate school | (EDUCA6) | 2 6 |

We will first attempt a median test for all categories to test if all categories have the same median, or if the interest level (i.e., the number of meetings with teachers) is greater for mothers with more education.

-->NPAR METHOD IS MEDIAN. VARIABLES ARE EDUCA1 TO EDUCA6.

```
K-SAMPLE MEDIAN TEST

GROUP --    EDUCA1   EDUCA2    EDUCA3    EDUCA4    EDUCA5    EDUCA6
LE MEDIAN         5        7         6         1         2         1
GT MEDIAN         5        4         7         3         2         1

*** WARNING: QUESTIONABLE CHI-SQUARE STATISTIC  THE ABOVE TABLE
    HAS    6 CELLS WITH EXPECTED FREQUENCIES LESS THAN 5.
    MINIMUM EXPECTED CELL FREQUENCY IS     1.0

TOTAL NUMBER OF CASES . . . . . .         44
MEDIAN. . . . . .              2.5000
CHI-SQUARE STATISTIC. . . . . .     1.8951
NUMBER OF TIES. . . . . .           5
SIGNIFICANCE LEVEL. . . . . .       0.8635
```

We did not specify a "median" value (by including the MEDIAN sentence) so the median of the combined sample is used to dichotomize groups. For the resultant two-sample test, we are provided with the total number of observations (cases), the value (median) that dichotomizes the data and the computed 2 x 6 contingency table.

Information regarding the derived $\chi^2$ statistic is also given. We note there is a warning that the 2 x 6 table has 6 cells with expected frequencies less than 5. These cells are those that correspond to EDUCA4 through EDUCA6, and can easily make the computed $\chi^2$ value invalid. To correct this we can group these categories into one, mothers with at least some college education. We can use the JOIN paragraph for this purpose (See Appendix A), then use the median test on the 4 categories.

```
-->JOIN    EDUCA4, EDUCA5, EDUCA6.

 THE JOIN OPERATION HAS BEEN COMPLETED, RESULT IS STORED IN VARIABLE  EDUCA4
 VARIABLE  EDUCA4  IS  A   10  BY   1  MATRIX

-->NPAR   METHOD IS MEDIAN.  VARIABLES ARE EDUCA1 TO EDUCA4.

 K-SAMPLE MEDIAN TEST

 GROUP -->    EDUCA1    EDUCA2    EDUCA3    EDUCA4
 LE MEDIAN         5         7         6         4
 GT MEDIAN         5         4         7         6

 TOTAL NUMBER OF CASES . . . . . .          44
 MEDIAN. . . . . . . . . . . . . .      2.5000
 CHI-SQUARE STATISTIC. . . . . . .      1.2951
 NUMBER OF TIES. . . . . . . . . .           3
 SIGNIFICANCE LEVEL. . . . . . . .      0.7303
```

All categories now have cells with "appropriate" expected frequencies (as each group has at least 10 cases). The computed $\chi^2$ value is not significant at the 10% level (to the 70% level). Hence the null hypothesis is supported. It appears that the mothers have relatively the same interest level (i.e., propensity for a teacher visit), regardless of educational background.

### 11.4.2 Kruskal-Wallis test

The Kruskal-Wallis test (also known as the Kruskal-Wallis one-way analysis of variance based on ranks) is used to determine whether K independent samples come from the same distribution or not. The null hypothesis is that all of the K population distribution functions are the same. The alternative hypothesis is that at least one population tends to yield values that are larger than at least one of the other populations. Since the test is sensitive against differences in means in the populations, the test can be used to determine whether all populations have the same mean or not. To compute the test statistic, all observations are combined into a single sample and ranked. Each observation is then replaced by its overall rank and the sum of the ranks for each sample is computed. The test statistic computed determines whether these sums were likely to be drawn from the same population. Since the rankings could yield ties, we can obtain two chi-square statistics, one correcting for possible ties in ranking.

## Examples

The following examples are from Siegel (1956, p.186-192). In the first example authoritarianism scores for three groups of teachers are stored in the workspace under the labels AUTSCOR1, AUTSCOR2, and AUTSCOR3. There are no ties among the overall rankings of these values. In the second example the birth weights of eight litters of pigs are stored under the labels LITTER1 through LITTER8. The second example has ties in its overall rankings.

## Example: Rankings without ties

To illustrate the Kruskal-Wallis test, we consider data on authoritarianism scores of three groups of teachers (Siegel, 1956, pages 186-189). The data are stored in the SCA workspace under the labels AUTSCOR1, AUTSCOR2 and AUTSCOR3, respectively; and are listed below.

AUTSCOR1: 96 128 83 61 101
AUTSCOR2: 82 124 132 135 109
AUTSCOR3: 115 149 166 147

We see that all scores are different so that a ranking of the combined data has no ties. If we wish to test if the distributions of the groups (or means of the groups) are the same we can enter

```
-->NPAR  METHOD IS KWH.  VARIABLES ARE AUTSCOR1, AUTSCOR2, AUTSCOR3.

KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE:

VARIABLE   CASES   MEAN RANK

AUTSCOR1     5        4.40
AUTSCOR2     5        7.40
AUTSCOR3     4       11.50

TOTAL NUMBER OF CASES  . . . . . .          14
DEGREES OF FREEDOM . . . . . .           2
KRUSKAL-WALLIS TEST STATISTIC (CHI-SQUARE) . . . . .       6.4057
SIGNIFICANCE LEVEL . . . . . .       0.0406
```

We obtain summary information on each group as well as the computed $\chi^2$ statistic (and its significance level). The $\chi^2$ value is 6.40 and is significant at the 5% level. Hence we may conclude the distributions (means) are statistically different.

## Example: Rankings with ties

To illustrate the test when ties in overall ranks occur, we consider the birth weights of eight litters of pigs. We wish to examine whether birth weight is affected by the size of the litter. The data, listed in Table 5, were taken from Siegel (1956, pages 189-192) and are stored in the SCA workspace under the labels LITTER1 through LITTER8, respectively

**Table 5 Birth weights (in pounds) for pig litters**

| Litter 1 LITTER1 | | Litter 2 LITTER2 | | Litter 3 LITTER3 | | Litter 4 LITTER4 | | Litter 5 LITTER5 | | Litter 6 LITTER6 | Litter 7 LITTER7 | | Litter 8 LITTER8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2.0 | 3.6 | 3.5 | 2.3 | 3.3 | 3.3 | 3.2 | 3.3 | 2.6 | 2.0 | 3.1 | 2.6 | 2.5 | 2.5 |
| 2.8 | 1.9 | 2.8 | 2.4 | 3.6 | 2.9 | 3.3 | 2.5 | 2.6 | 2.0 | 2.9 | 2.2 | 1.2 | 2.4 |
| 3.3 | 3.3 | 3.2 | 3.0 | 2.6 | 3.4 | 3.2 | 2.6 | 2.9 | 2.1 | 3.1 | 2.2 | 1.2 | 3.0 |
| 3.2 | 2.8 | 3.5 | 1.6 | 3.1 | 3.2 | 2.9 | 2.8 | | | 2.5 | | | 1.5 |
| 4.4 | 1.1 | | | 3.2 | 3.2 | | | | | | | | |

We can compute both $\chi^2$ tests by entering

-->NPAR METHOD IS KWH. VARIABLES ARE LITTER1 TO LITTER8.

```
KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE:

VARIABLE   CASES   MEAN RANK

 LITTER1    10       31.70
 LITTER2     8       27.06
 LITTER3    10       41.40
 LITTER4     8       34.69
 LITTER5     6       17.58
 LITTER6     4       30.50
 LITTER7     6       11.92
 LITTER8     4       18.00

TOTAL NUMBER OF CASES  . . . . . .          56
DEGREES OF FREEDOM . . . . . .           7
KRUSKAL-WALLIS TEST STATISTIC  (CHI-SQUARE). . . . .      18.4639
SIGNIFICANCE LEVEL . . . . . .       0.0100

KRUSKAL-WALLIS TEST STATISTIC  (CORRECTED FOR TIES).      18.5654
SIGNIFICANCE LEVEL . . . . . .       0.0097
```

Since ties in ranks are present, two statistics are computed and displayed. Either statistic is significant at the 1% level, hence the litters' weights are statistically different. Hence litter size may affect litter birth weights.

## 11.5 Tests Involving Two Related Samples

We now consider tests of related samples. First we consider nonparametric tests involving paired observations or two matched samples. At least one example is provided for each test. Examples are discussed in more detail in the cited references.

The tests of this section are

| Section | Test | Hypothesis test for | Minimum measurement scale |
|---|---|---|---|
| 11.5.1 | Sign test | location (median) and independence | ordinal |
| 11.5.2 | Wilcoxon test | location (median) | ordinal |
| 11.5.3 | Kendall's rank correlation | correlation | ordinal |
| 11.5.4 | Spearman's rank correlation | correlation | ordinal |

## 11.5.1 Sign test

The sign test is used to test whether two populations have the same median. Observations consist of matched pair samples from the populations. The test can also be used as a test of trend in a series or as a test for correlation. The sign test is the binomial test (see Section 2.1) with p*=.5, but the test focuses only on the direction of difference between individual pairs of data. That is, only the information of whether the first value is larger than its matched value (denoted by '+') or the converse (denoted by '-') is used. If the populations have the same median then the number of +'s and –'s should be approximately the same.

If we have no basis to pair observations, or if the samples are thought to be independent, the Mann-Whitney test (see Section 11.3.2) is a more powerful test to use.

### Example: Paternal discipline data

We will use an example from Siegel (1956, p.69-71) to illustrate the sign test. Data on ratings given on "insight" into paternal discipline for 17 pairs of mothers and fathers are stored in the workspace under the labels MOTHER and FATHER, respectively. A low value indicates a high "insight". The data are listed below.

```
MOTHER:  4 4 5 5 3 2 5 3 1 5 5 5 4 5 5 5 5
FATHER:  2 3 3 3 3 3 3 3 2 3 2 2 5 2 5 3 1
```

To obtain the sign test for these paired data sets we enter

```
-->NPAR  METHOD IS SIGN.  VARIABLES ARE MOTHER, FATHER.
```

```
NONPARAMETRIC SIGN TEST FOR THE VARIABLES MOTHER AND FATHER

TOTAL NUMBER OF CASES . . . . . .          17
NUMBER OF NEGATIVE DIFFERENCES (1ST LT 2ND) . .          3
NUMBER OF POSITIVE DIFFERENCES (1ST GT 2ND) . .         11
NUMBER OF TIES. . . . . .          3
ONE-TAILED PROBABILITY. . . . . .      0.0287
TWO-TAILED PROBABILITY. . . . . .      0.0574
```

We would reject the hypothesis that the discipline levels are the same, in favor that fathers have a better insight level at the 5% level.

### 11.5.2 The Wilcoxon test

The Wilcoxon matched-pairs signed-rank test is used for testing whether two populations have the same median, when observations are taken in matched pairs. The test is similar to the sign test except the magnitude of a difference is also considered. In a two-tailed test, the null hypothesis that both populations have the same median is tested against the alternative that they do not. In a one-tailed test, the null hypothesis that the median from one population is at least as large as that from the second population is tested versus the converse.

The Wilcoxon test statistic is based on the ranks of the absolute difference of the matched pairs. If the two populations have the same median, then the sum of the ranks of the absolute differences should be approximately the same in the cases that: (1) the first observation is larger than the second, and (2) the second observation is larger than the first.

Two examples from Siegel (1956, p.77-83) are used to illustrate the Wilcoxon test. In the first example the social perceptiveness scores of identical twins, one who attended nursery school and the other who did not, are stored in the workspace under the labels SCHOOL and HOME, respectively. The data are listed below.

```
SCHOOL:  82  69  73  43  58  56  76  85
HOME:    63  42  74  37  51  43  80  82
```

To test whether the scores are the same, we enter

```
-->NPAR  METHOD IS WILCOXON.  VARIABLES ARE SCHOOL, HOME.
```

```
NONPARAMETRIC WILCOXON SIGNED-RANK TEST BETWEEN   SCHOOL AND     HOME

TOTAL NUMBER OF CASES . . . . . .          8
NUMBER OF NEGATIVE DIFFERENCES (1ST LT 2ND) . . . .          2
NUMBER OF POSITIVE DIFFERENCES (1ST GT 2ND) . . . .          6
NUMBER OF TIES. . . . . .          0
VALUE OF THE STANDARDIZED NORMAL SCORE. . . . . -1.9604
ONE-TAILED EXACT PROBABILITY  . . . . . 0.0250
TWO-TAILED EXACT PROBABILITY  . . . . . 0.0499
```

We observe that in 6 of the paired cases, the perceptive score of the twin who attended nursery school is less than the one who did not. This leads to a test statistic that is significant at the 5% level. Hence the hypothesis that the score is not affected by school attendance is rejected at this level.

As a second example of the Wilcoxon test, the utilities (value) of federal inmates related to cigarettes is considered. We can contrast these utilities to the completely ambivalent utility, 0. These utility values are stored in the SCA workspace under the labels UTILITY and ZERO, respectively. The values of these variables are listed below.

```
UTILITY:  -2 0 0 1 0 0 4 4 1 1 5 3 5 3-1 1-1 5 8 2 2 2-3-2 1 4 8 2 3-1

ZERO:     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

To test the hypothesis that the inmates have no feeling regarding cigarettes, we may enter

```
-->NPAR  METHOD IS WILCOXON.  VARIABLES ARE ZERO, UTILITY.
```

```
NONPARAMETRIC WILCOXON SIGNED-RANK TEST BETWEEN     ZERO AND   UTILITY

TOTAL NUMBER OF CASES . . . . . .          30
NUMBER OF NEGATIVE DIFFERENCES (1ST LT 2ND) . . . .          20
NUMBER OF POSITIVE DIFFERENCES (1ST GT 2ND) . . . .           6
NUMBER OF TIES. . . . . .           4
VALUE OF THE STANDARDIZED NORMAL SCORE. . . . . .    -3.1113
ONE-TAILED EXACT PROBABILITY  . . . . . .     0.0009
TWO-TAILED EXACT PROBABILITY  . . . . . .     0.0019
```

We see that the Wilcoxon statistic is significant at the 1% level and the hypothesis is strongly rejected.

### 11.5.3 Kendall's rank correlation

Kendall's rank correlation coefficient, $\tau$, is used as a measure of the correlation between a set of matched pairs of data. Kendall's rank correlation is a measure of the mutual relationship that may exist between two variables. It does not necessarily imply that any causal relationships exist. Kendall's $\tau$ is also useful in the study of partial correlation.

To compute $\tau$, data values for each variable are replaced by their ranks. The ranked values are then re-ordered so that one variable is made to be in ascending order. The coefficient, $\tau$, is calculated from the now re-ordered rankings. For each pair of rankings $(r_{1i}, r_{2i})$, i=1,2, . . . , n-1, the number of concordant pairs above it (that is, $(r_{1j}, r_{2j})$ $j > i$ with $r_{1j} > r_{1i}$ and $r_{2j} > r_{2i}$) is counted. Similarly, the number of discordant pairs (i.e., $(r_{1j}, r_{2j})$ with $r_{1j} > r_{1i}$ and $r_{2j} < r_{2i}$) is counted. Ties are "skipped". If NC is the total for the concordant pairs and ND is the total for the discordant pairs, then

$$\tau = \frac{2(N_C - N_D)}{n(n-1)}$$

A standard deviation for $\tau$ can also be computed so that a normal score can be calculated. This score can then be used for both a one-tailed and two-tailed test of the hypothesis that the two samples are uncorrelated. In the case of ties, the denominator is adjusted to reflect the number, and ranking, of ties (see Section 7.2.2 of Gibbons, 1985).

## Example: Conformity study

We will illustrate Kendall's rank correlation with examples from Siegel (1956, p.216-219). Data are derived from a study on conformity. Scores were given measuring authoritarianism, social status strivings, and the number of times an individual yielded to group pressures. Data are stored in the workspace under the labels AUTH, STATUS and YIELD, respectively. Data are listed in Table 6.

**Table 6 Conformity study data**

| Subject | Authoritarianism Score AUTH | Social Status Measure STATUS | Number of times yielded to group YIELD |
|---|---|---|---|
| 1 | 82 | 42 | 0 |
| 2 | 98 | 46 | 0 |
| 3 | 87 | 39 | 1 |
| 4 | 40 | 37 | 1 |
| 5 | 116 | 65 | 3 |
| 6 | 113 | 88 | 4 |
| 7 | 111 | 86 | 5 |
| 8 | 83 | 56 | 6 |
| 9 | 85 | 62 | 7 |
| 10 | 126 | 92 | 8 |
| 11 | 106 | 54 | 8 |
| 12 | 117 | 81 | 12 |

First we will compute the rank correlation between AUTH and STATUS. Next we calculate $\tau$ for STATUS and YIELD.

```
-->NPAR  METHOD IS KRANK.  VARIABLES ARE AUTH, STATUS.

KENDALL'S RANK CORRELATION FOR     AUTH AND   STATUS

 CASE  1ST VAR      RANK   2ND VAR      RANK

   1     82.00      2.00     42.00      3.00
   2     98.00      6.00     46.00      4.00
   3     87.00      5.00     39.00      2.00
   4     40.00      1.00     37.00      1.00
   5    116.00     10.00     65.00      8.00
   6    113.00      9.00     88.00     11.00
   7    111.00      8.00     86.00     10.00
   8     83.00      3.00     56.00      6.00
   9     85.00      4.00     62.00      7.00
  10    126.00     12.00     92.00     12.00
  11    106.00      7.00     54.00      5.00
  12    117.00     11.00     81.00      9.00

TOTAL NUMBER OF CASES . . . . . .        12
```

```
RANK CORRELATION. . . . . .     0.6667
STANDARD DEVIATION. . . . . .     0.2210
Z STATISTIC (NORMAL APPROXIMATION). . . . .     3.0172
ONE-TAILED PROBABILITY. . . . . .     0.0013
TWO-TAILED PROBABILITY. . . . . .     0.0026
```

-->NPAR METHOD IS KRANK. VARIABLES ARE YIELD, STATUS.

```
KENDALL'S RANK CORRELATION FOR    YIELD AND   STATUS

  CASE  1ST VAR      RANK   2ND VAR      RANK
    1      0.0       1.50     42.00      3.00
    2      0.0       1.50     46.00      4.00
    3      1.00      3.50     39.00      2.00
    4      1.00      3.50     37.00      1.00
    5      3.00      5.00     65.00      8.00
    6      4.00      6.00     88.00     11.00
    7      5.00      7.00     86.00     10.00
    8      6.00      8.00     56.00      6.00
    9      7.00      9.00     62.00      7.00
   10      8.00     10.50     92.00     12.00
   11      8.00     10.50     54.00      5.00
   12     12.00     12.00     81.00      9.00

TOTAL NUMBER OF CASES . . . . . .       12
RANK CORRELATION. . . . . .     0.3722
STANDARD DEVIATION. . . . . .     0.2210
Z STATISTIC (NORMAL APPROXIMATION). . . . .     1.6845
ONE-TAILED PROBABILITY. . . . . .     0.0460
TWO-TAILED PROBABILITY. . . . . .     0.0921
```

In both cases we are provided with relative ranking information for the variables, the value computed for $\tau$, its standard deviation, and one and two tailed probability levels. The probability levels are derived from the normal approximation. This approximation is less appropriate for small sample sizes and is not used when the number of cases is less than 10. In these cases we need to consult other tables (e.g. Table J of Gibbons, 1985).

## 11.5.4 Spearman's rank correlation

Spearman's rank correlation coefficient, $\rho_s$ , is used as a measure of the correlation between a set of matched pairs of data. As we noted in the previous section, correlation is a measure of the mutual relationship that may exist between two variables. It does not necessarily imply that any causal relationships exist.

The Spearman coefficient, $\rho_s$, is an extension of the ordinary Pearson correlation coefficient

$$r_s = \frac{\left[\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)/(n-1)\right]}{s_1 s_2} \text{ where}$$

$\bar{x}_1, s_1, \bar{x}_2$ and $s_2$ are the sample mean and standard deviations of the variables $x_1$ and $x_2$ , respectively. The value $\rho_s$, is obtained by using the above calculation for the ranked values

of two variables. In addition, a t-statistic can be computed based on the value of $\rho_s$. This statistic can be used for both a one-tailed and two-tailed tests of the hypothesis that the two samples are uncorrelated.

We will calculate the Spearman's rank correlation for the conformity study data sets we used in Section 5.3. A discussion can be found in Siegel (1956, pages 211-212). First $\rho_s$ is computed for AUTH and STATUS, then for YIELD and STATUS.

-->NPAR METHOD IS SRANK. VARIABLES ARE AUTH, STATUS.

```
SPEARMAN'S RANK CORRELATION FOR     AUTH AND   STATUS

  CASE  1ST VAR      RANK   2ND VAR      RANK
    1     82.00      2.00     42.00      3.00
    2     98.00      6.00     46.00      4.00
    3     87.00      5.00     39.00      2.00
    4     40.00      1.00     37.00      1.00
    5    116.00     10.00     65.00      8.00
    6    113.00      9.00     88.00     11.00
    7    111.00      8.00     86.00     10.00
    8     83.00      3.00     56.00      6.00
    9     85.00      4.00     62.00      7.00
   10    126.00     12.00     92.00     12.00
   11    106.00      7.00     54.00      5.00
   12    117.00     11.00     81.00      9.00

TOTAL NUMBER OF CASES . . . . . .         12
RANK CORRELATION. . . . . .      0.8182
T-STATISTIC (APPROXIMATE) . . . . . .      4.5000
DEGREES OF FREEDOM. . . . . .         10
ONE-TAILED PROBABILITY. . . . . .      0.0006
TWO-TAILED PROBABILITY. . . . . .      0.0011
```

-->NPAR METHOD IS SRANK. VARIABLES ARE YIELD, STATUS.

```
SPEARMAN'S RANK CORRELATION FOR    YIELD AND   STATUS

  CASE  1ST VAR      RANK   2ND VAR      RANK
    1      0.0       1.50     42.00      3.00
    2      0.0       1.50     46.00      4.00
    3      1.00      3.50     39.00      2.00
    4      1.00      3.50     37.00      1.00
    5      3.00      5.00     65.00      8.00
    6      4.00      6.00     88.00     11.00
    7      5.00      7.00     86.00     10.00
    8      6.00      8.00     56.00      6.00
    9      7.00      9.00     62.00      7.00
   10      8.00     10.50     92.00     12.00
   11      8.00     10.50     54.00      5.00
   12     12.00     12.00     81.00      9.00

TOTAL NUMBER OF CASES . . . . . .         12
RANK CORRELATION. . . . . .      0.6151
T-STATISTIC (APPROXIMATE) . . . . . .      2.4672
DEGREES OF FREEDOM. . . . . .         10
ONE-TAILED PROBABILITY. . . . . .      0.0166
TWO-TAILED PROBABILITY. . . . . .      0.0333
```

We obtain a similar display as for that of Kendall's rank correlation. Information related to the t-statistic of $\rho_s$ replaces the standard normal information related to Kendall's $\tau$.

We note that the $\tau$ values of the two sets are .67 and .37 respectively. The $\rho_s$ values are .82 and .62. This illustrates the underlying different scales used for the two statistics and they are not directly comparable to each other.

## 11.6 Tests Involving Several Related Samples

We now extend nonparametric tests for two related samples to nonparametric tests involving several related samples. As "extensions" to matched, or paired, observations, the number of observations per sample are the same. Examples are discussed in cited references.

The tests of this section are

| Section | Test | Hypothesis test for | Minimum measurement scale |
|---|---|---|---|
| 11.6.1 | Cochran Q test | location (means) | nominal |
| 11.6.2 | Friedman test | location (means) | ordinal |
| 11.6.3 | Kendall's coefficient of concordance | concordance (agreement) | ordinal |

### 11.6.1 Cochran Q test

The Cochran Q test is used to test whether three or more matched sets of frequencies, or proportions, differ significantly among themselves. The data sets used in the test must be binary (i.e., 0 or 1). For example, a "success" is represented by a 1 and a "failure" by a 0. The null hypothesis to be tested is that the probability of a "success" is the same for all samples. The alternative is that the probability of "success" is different in at least one sample. The test statistic is based on the total number of "successes" in a sample (i.e., column or treatment), the total number of "successes" for a matched set (i.e., row or block), and the overall total number of "successes". The test statistic follows an approximate chi-square distribution.

## Example: Interviewing style data

We will use an example from Siegal (1956, pages 163-165) to illustrate the calculation of the Cochran Q statistic. Positive responses from three types of interviewing styles are recorded from 18 sets of housewives. Data are listed in Table 7 and stored in the SCA workspace under the labels INTEREST, FORMAL, and HOSTILE.

**Table 7 Responses from interview styles**

| Subject | Interested *INTEREST* | Formality *FORMAL* | Hostility *HOSTILE* |
|---|---|---|---|
| 1 | 1.00 | 1.00 | 1.00 |
| 2 | 1.00 | 1.00 | 1.00 |
| 3 | 1.00 | 1.00 | .00 |
| 4 | 1.00 | 1.00 | .00 |
| 5 | 1.00 | 1.00 | .00 |
| 6 | 1.00 | 1.00 | 1.00 |
| 7 | 1.00 | 1.00 | .00 |
| 8 | 1.00 | 1.00 | .00 |
| 9 | .00 | .00 | .00 |
| 10 | 1.00 | 1.00 | .00 |
| 11 | .00 | 1.00 | .00 |
| 12 | .00 | .00 | .00 |
| 13 | 1.00 | .00 | .00 |
| 14 | 1.00 | 1.00 | .00 |
| 15 | 1.00 | 1.00 | .00 |
| 16 | .00 | 1.00 | .00 |
| 17 | 1.00 | .00 | .00 |
| 18 | .00 | .00 | .00 |

To calculate the Cochran Q test we enter

```
-->NPAR  METHOD IS COCHRANQ.  VARIABLES ARE INTEREST, FORMAL, HOSTILE.

COCHRAN'S Q-TEST FOR GROUPS OF DICHOTOMOUS DATA
     INTEREST   FORMAL  HOSTILE

VAR. INTEREST   FORMAL  HOSTILE

   1     1.00     1.00     1.00
   2     1.00     1.00     1.00
   3     1.00     1.00      .00
   4     1.00     1.00      .00
   5     1.00     1.00      .00
   6     1.00     1.00     1.00
   7     1.00     1.00      .00
   8     1.00     1.00      .00
   9      .00      .00      .00
  10     1.00     1.00      .00
  11      .00     1.00      .00
  12      .00      .00      .00
  13     1.00      .00      .00
  14     1.00     1.00      .00
  15     1.00     1.00      .00
  16      .00     1.00      .00
  17     1.00      .00      .00
  18      .00      .00      .00

COCHRAN'S Q STATISTIC . . . . . .    16.6667
DEGREES OF FREEDOM  . . . . . .          2
SIGNIFICANCE LEVEL. . . . . .     0.0002
```

## 11.6.2 Friedman test

The Friedman test (also known as the Friedman two-way analysis of variance) is used to test whether K matched samples have been drawn from the same population. Data of each sample can be viewed as responses to a treatment. The null hypothesis tested is that treatments have identical effects. The alternative hypothesis is that at least one treatment tends to yield larger observed values than at least one other treatment. The test statistic is based on the ranks of each matched set of observations (i.e., in each row or block, if "treatments" are viewed as columns). The sum of these ranks over treatments is used to calculate the statistic. If all treatments have identical effects, then each treatment sum of ranks should be approximately the same. The test statistic follows an approximate chi-square distribution.

## Examples

Two examples from Lehmann (1975, p.261-265) are used to illustrate the Friedman test. In the first example, ranked data on three brands of tranquilizers for 4 groups of patients is stored in the workspace under the labels A, B and C. Each patient has assigned a relative ranking to each brand. The rankings are the values of A, B and C and are listed in Table 8.

**Table 8 Rankings of Tranquilizers**

| Subject | Brand | | |
|---|---|---|---|
| | A | B | C |
| 1 | 3 | 2 | 1 |
| 2 | 2 | 3 | 1 |
| 3 | 3 | 1 | 2 |
| 4 | 3 | 1 | 2 |

To calculate the Friedman test statistic for this data we enter

-->NPAR METHOD IS FRIEDMAN. VARIABLES ARE A, B, C.

```
FRIEDMAN TWO-WAY ANALYSIS OF VARIANCE FOR THE VARIABLES:
            A        B        C

VAR.        A        B        C

   1     3.00     2.00     1.00
RANK     3.00     2.00     1.00

   2     2.00     3.00     1.00
RANK     2.00     3.00     1.00

   3     3.00     1.00     2.00
RANK     3.00     1.00     2.00

   4     3.00     1.00     2.00
RANK     3.00     1.00     2.00

FRIEDMAN'S Q STATISTIC . . . . . .     3.5000
DEGREES OF FREEDOM FOR CHI-SQUARE DISTRIBUTION . .          2
SIGNIFICANCE LEVEL . . . . . .     0.1738
```

We obtain a listing of our data (with redundant rankings in this case), the Friedman statistic and its significance. With the computed significant level, we would not reject a hypothesis of identical effects at the 10% level.

As a second example we consider the measurements of skin potential recorded from 8 subjects during hypnosis. The measurements came in response to (randomly ordered) suggestions of fear, happiness, depression and calmness. Data are listed in Table 9 and are stored in the SCA workspace under the labels FEAR, HAPPY, DEPRESS, CALM, respectively.

**Table 9 Skin potential readings**

| Subject | Emotion | | | |
|---|---|---|---|---|
| | Fear *FEAR* | Happiness *HAPPY* | Depression *DEPRESS* | Calmness *CALM* |
| 1 | 23.1 | 22.7 | 22.5 | 22.6 |
| 2 | 57.6 | 53.2 | 53.7 | 53.1 |
| 3 | 10.5 | 9.7 | 10.8 | 8.3 |
| 4 | 23.6 | 19.6 | 21.1 | 21.6 |
| 5 | 11.9 | 13.8 | 13.7 | 13.3 |
| 6 | 54.6 | 47.1 | 39.2 | 37.0 |
| 7 | 21.0 | 13.6 | 13.7 | 14.8 |
| 8 | 20.3 | 23.6 | 16.3 | 14.8 |

-->NPAR METHOD IS FRIEDMAN. VARIABLES ARE FEAR, HAPPY, DEPRESS, CALM.

```
FRIEDMAN TWO-WAY ANALYSIS OF VARIANCE FOR THE VARIABLES:
         FEAR HAPPY DEPRESS CALM

VAR.      FEAR    HAPPY  DEPRESS     CALM

   1     23.10    22.70    22.50    22.60
RANK      4.00     3.00     1.00     2.00

   2     57.60    53.20    53.70    53.10
RANK      4.00     2.00     3.00     1.00

   3     10.50     9.70    10.80     8.30
RANK      3.00     2.00     4.00     1.00

   4     23.60    19.60    21.10    21.60
RANK      4.00     1.00     2.00     3.00

   5     11.90    13.80    13.70    13.30
RANK      1.00     4.00     3.00     2.00

   6     54.60    47.10    39.20    37.00
RANK      4.00     3.00     2.00     1.00

   7     21.00    13.60    13.70    14.80
RANK      4.00     1.00     2.00     3.00

   8     20.30    23.60    16.30    14.80
RANK      3.00     4.00     2.00     1.00

FRIEDMAN'S Q STATISTIC . . . . . .     6.4500
DEGREES OF FREEDOM FOR CHI-SQUARE DISTRIBUTION . .          3
SIGNIFICANCE LEVEL . . . . . .     0.0917
```

We note that the statistic is significant at the 10% level but not at the 5% level. Hence there is some evidence that the skin readings are different for different emotions.

### 11.6.3 Kendall's coefficient of concordance

Kendall's coefficient of concordance, W, is used as a measure of total correlation for three or more matched samples. A chi-square statistic can be computed from W in the construction of a test of whether the samples have been drawn from the same population. However, W was probably intended only as a measure of "agreement in rankings" (Cochran, 1980, page 305). The coefficient is related closely to the test statistic of Friedman (see Section 11.6.2). The test is based on the rankings of the observations within each sample. The sum of these ranks over each matched set (i.e., row or block) is used to calculate the statistic. An associated chi-square statistic can be computed from W.

**Example: Coffee tasting**

We will illustrate the calculation of W with an example from Gibbons (1985, p.307-309). In order to test the objectivity and differences between professional coffee tasters, a coffee manufacturer solicits 4 tasters to rank 10 different blends of coffee. The data are listed in Table 10. Rankings are stored in the SCA System under the labels TASTER1 through TASTER4.

**Table 10 Rankings of Coffee Blends**

| *Blend* | *Taster* | | | |
|---|---|---|---|---|
| | *1* | *2* | *3* | *4* |
| 1 | 10 | 10 | 10 | 9 |
| 2 | 9 | 9 | 9 | 10 |
| 3 | 7 | 8 | 8 | 7 |
| 4 | 8 | 6 | 7 | 8 |
| 5 | 6 | 7 | 6 | 6 |
| 6 | 5 | 3 | 5 | 5 |
| 7 | 4 | 5 | 4 | 4 |
| 8 | 1 | 2 | 1 | 1 |
| 9 | 2 | 1 | 2 | 2 |
| 10 | 3 | 4 | 3 | 3 |

To calculate W and obtain the related $\chi^2$ statistic, we enter

-->NPAR METHOD IS KENDALLW. VARIABLES ARE TASTER1 TO TASTER4.

```
KENDALL COEFFICIENT OF CONCORDANCE
      TASTER1  TASTER2  TASTER3  TASTER4

VAR.  TASTER1  TASTER2  TASTER3  TASTER4

   1    10.00    10.00    10.00     9.00
RANK    10.00    10.00    10.00     9.00

   2     9.00     9.00     9.00    10.00
RANK     9.00     9.00     9.00    10.00

   3     7.00     8.00     8.00     7.00
RANK     7.00     8.00     8.00     7.00

   4     8.00     6.00     7.00     8.00
RANK     8.00     6.00     7.00     8.00

   5     6.00     7.00     6.00     6.00
RANK     6.00     7.00     6.00     6.00

   6     5.00     3.00     5.00     5.00
RANK     5.00     3.00     5.00     5.00

   7     4.00     5.00     4.00     4.00
RANK     4.00     5.00     4.00     4.00

   8     1.00     2.00     1.00     1.00
RANK     1.00     2.00     1.00     1.00

   9     2.00     1.00     2.00     2.00
RANK     2.00     1.00     2.00     2.00

  10     3.00     4.00     3.00     3.00
RANK     3.00     4.00     3.00     3.00


KENDALL'S W STATISTIC . . . . . .      .9636
CHI-SQUARE STATISTIC. . . . . .    34.6909
DEGREES OF FREEDOM. . . . . .           9
SIGNIFICANCE LEVEL. . . . . .      .0001
```

We obtain a listing of our data (with associated ranks, here the same values), W, the $\chi^2$ value and its significance level.

As is apparent in Table 10, where the tasters' rankings are almost identical for each blend, we conclude the rankings are not independent. The actual significance level is .01%, a highly significant level.

# SUMMARY OF THE SCA PARAGRAPH IN CHAPTER 11

The NPAR paragraph is used to access the calculation and display of a nonparametric test statistic. Only one test may be performed in any single execution of the paragraph. The statistics displayed for the paragraph depend upon the method requested. A summary of the statistics displayed is given with the description of each test (see Sections 2 - 6).

## Syntax for the NPAR paragraph

The NPAR paragraph begins with the paragraph name, NPAR, and is followed by modifying sentences. Sentences that may be used as modifiers for this paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portion of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

Note: Not all sentences shown below are appropriate for a specified method. Please consult the sentence description to be sure if the sentence may be used as a modifying sentence in conjunction with the specific method.

Legend (see Chapter 1 for further explanation)

v : variable name
r : real value
w : keyword

```
NPAR METHOD IS w.                                  @
     VARIABLE(S) ARE v1, v2, ---.                  @
     EXPECTED IS v.                                @
     DISTRIBUTION IS distribution (parameters).    @
     CUT-POINT IS r.                               @
     PROPORTION IS r.                              @
     MEDIAN IS r.
```

Required sentences: **METHOD, VARIABLE**

Note: Other sentences may be required depending upon the nonparametric test chosen. Please consult the METHOD sentence description below for other sentences that may be required.

## Sentences Used in the NPAR Paragraph

**METHOD sentence**

The METHOD sentence is used to specify the nonparametric procedure to be performed. Listed below are available keywords, the test associated with the keyword, the section of this chapter in which a brief description of the procedure is given, and other modifying sentences that may be used in conjunction with the given test. Those associated sentences that are underlined are additional required sentences for the paragraph. Note that only one method can be specified for a single use of the NPAR paragraph.

| Keyword | Nonparametric test | Section Reference | Associated sentence(s) |
|---|---|---|---|
| BINOMIAL | Binomial | 2 | CUT-POINT, PROPORTION |
| COCHRANQ | Cochran's Q | 6 | none |
| FRIEDMAN | Friedman two-way ANOVA | 6 | none |
| GOODNESS | Chi-square goodness of fit | 2 | EXPECTED |
| INDEP | Chi-square for independence in an r x c contingency table | 2 | none |
| KENDALLW | Kendall's coefficient of concordance | 6 | none |
| KRANK | Kendall's rank correlation | 5 | none |
| KS1 | Kolmogorov-Smirnov (one sample) | 2 | DISTRIBUTION |
| KS2 | Kolmogorov-Smirnov (two sample) | 3 | none |
| KWH | Kruskal-Wallis H | 4 | none |
| MEDIAN | Median | 3, 4 | MEDIAN |
| MWU | Mann-Whitney U | 3 | none |
| RUNS | Runs | 2 | CUT-POINT |
| SIGN | Sign | 5 | none |
| SRANK | Spearman's rank correlation | 5 | none |
| WILCOXON | Wilcoxon | 5 | none |

**VARIABLE(S) sentence**

The VARIABLE(S) sentence is used to specify name(s) of the variable(s) for which nonparametric statistics are to be produced. The number of variables specified must be consistent with the test chosen in the METHOD sentence. Please refer to the appropriate section of this chapter to determine the number permitted.

**EXPECTED sentence**

The EXPECTED sentence is used to provide the name of the variable containing the expected values of the groups used in the chi-square goodness of fit test. The number of values in the variable specified must be the same as that of the variable specified in the VARIABLE sentence. This is a required sentence when GOODNESS is specified in the METHOD sentence. The sentence is ignored otherwise.

**DISTRIBUTION sentence**

The DISTRIBUTION sentence is used to specify the distribution that will be assumed in the calculation of the Kolmogorov-Smirnov one sample test. The following distributions can be used:

| | | |
|---|---|---|
| U(r1, r2) | : | uniform distribution between r1 and r2 |
| N(r1, r2) | : | normal distribution with mean r1 and variance r2 |
| POISSON(r) | : | Poisson distribution with parameter r |

If KS1 is specified in the METHOD sentence and the DISTRIBUTION sentence is not specified, then a uniform distribution is used with r1 and r2 the minimum and maximum values of the specified variable. The sentence is ignored if KS1 is not specified in the METHOD sentence.

**CUT-POINT sentence**

The CUT-POINT sentence is used to specify a cutting point that will be used to dichotomize a variable when either BINOMIAL or RUNS is specified in the METHOD sentence. The sentence is ignored otherwise.

**PROPORTION sentence**

The PROPORTION sentence is used to specify the proportion, r, that will be used in the significance test when BINOMIAL is specified in the METHOD sentence. The sentence is ignored otherwise. The value specified must be between 0 and 1. The default value used is 0.5. Please see Section 11.2 for a discussion.

**MEDIAN sentence**

The MEDIAN sentence is used to specify a value that will be used to dichotomize samples in the median test. If the sentence is not specified, the combined grand median will be used. The sentence is ignored if the median test is not specified.

**ACKNOWLEDGEMENT**

Scientific Computing Associates gratefully appreciates the programming assistance of Philip Burns in the development of the NPAR paragraph.

## REFERENCES

Box, G.E.P., Hunter W.G., and Hunter, J.S. (1978). *Statistics for Experimenters*, New York: Wiley.

Bradley, J.V. (1968). *Distribution-Free Statistical Tests*, Englewood Cliffs, NJ: Prentice-Hall.

Conover, W.J. (1980). *Practical Nonparametric Statistics*, 2nd edition, New York: Wiley.

Gibbsons, J.D. (1985). *Nonparametric Methods for Quantitative Analysis*, 2nd edition, Columbus, OH: American Sciences Press, Inc.

Hollander, M., and Wolfe, D.A. (1973). *Nonparametric Statistical Methods*, New York: Wiley.

Kraft, C.H. and van Eeden, C. (1968). *A Nonparametric Introduction to Statistics*, New York: MacMillan.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

Marascuilo, L.A. and McSweeney, M. (1977). *Nonparametric and Distribution Free Methods for the Social Sciences*, Belmont, CA: Wadsworth.

Mosteller, F., and Rourke, R.E.K. (1973). *Sturdy Statistics*, Reading, MA: Addison-Wesley.

Noether, G.E. (1976). *Elements of Nonparametric Statistics*, New York: Wiley.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*, New York: McGraw-Hill.

Walsh, J.E. (1962). *Handbook of Nonparametric Statistics*, Princeton, NJ: Van Nostrand.

# CHAPTER 12

# DISTRIBUTION AND MODEL SIMULATION

The simulation of data is often beneficial for both data analyses and scientific research. Simulated data can provide us with a better understanding of various statistical methods, especially when methods are either ad hoc or difficult to understand. In addition, simulated data provide a convenient means to ascertain the sensitivity of an analysis, especially in the study of departures from distributional assumptions.

Simulated data are derived from pseudo random number generators. These data are usually consonant with a specific statistical distribution. We may also find generated data, that is, data completely specified in some manner, are useful in data analyses. Generated data includes indicator variables, i.e., binary data that assume the value 1 under certain circumstances and 0 otherwise.

The SCA System provides us with a great degree of flexibility in the simulation or generation of data. Information regarding the simulation of data according to a distribution or time series model is provided in this chapter. Information regarding the generation of data is found in Appendix B.

The SIMULATE paragraph can be used to simulate data that follow a uniform, normal, or multivariate normal distribution. The paragraph can also be used to simulate data according to a time series model.

## 12.1 Simulating Data According to a Distribution

We will illustrate the simulation of data according to a probability distribution by simulating data from a uniform (-1, 1), standard normal, and a bivariate normal distribution.

### Uniform distribution

To simulate 100 observations of a uniform (□1, 1) distribution we can simply enter

-->SIMULATE XUNIF. NOBS ARE 100. DISTRIBUTION IS U(-1.0, 1.0)

```
THE UNIFORM DISTRIBUTION SPECIFIED IS SIMULATED, THE RESULT IS
STORED IN VARIABLE XUNIF
```

We specify we want to simulate data and store it in the variable with label XUNIF. There should be 100 observations, and the distribution the data should follow is U(-1.0, 1.0). We can use the DESCRIBE paragraph (see Chapter 4) to compare the sample statistics of these simulated data with their theoretic counterparts. A summary of this information is provided below:

| | Sample statistic | Theoretic value |
|---|---|---|
| Mean | -0.0067 | 0.0 |
| Variance | 0.3573 | 1/3 |
| Skewness | 0.0110 | 0.0 |
| Kurtosis | -1.4095 | -1.2 |
| 25th Percentile | -0.5488 | -0.5 |
| 75th Percentile | 0.5495 | 0.5 |

These sample values are in good agreement with the theoretic values. We can observe the actual distribution of the simulated data by creating a histogram (see Chapter 5). We see the data appears rather uniformly distributed between -1 and 1.

```
-->HISTOGRAM   XUNIF.   INTERVAL IS 4.
```

```
  LOWER     UPPER  FREQ- 0    5   10   15   20   25   30   35   40
  BOUND     BOUND  UENCY +----+----+----+----+----+----+----+----+----
                         I
 -1.000 -    -.500    26 IXXXXXXXXXXXXXXXXXXXXXXXXX
                         IXXXXXXXXXXXXXXXXXXXXXXXXX
                         I
  -.500 -     .000    28 IXXXXXXXXXXXXXXXXXXXXXXXXXXX
                         IXXXXXXXXXXXXXXXXXXXXXXXXXXX
                         I
   .000 -     .500    18 IXXXXXXXXXXXXXXXXXX
                         IXXXXXXXXXXXXXXXXXX
                         I
   .500 -    1.000    28 IXXXXXXXXXXXXXXXXXXXXXXXXXXX
                         IXXXXXXXXXXXXXXXXXXXXXXXXXXX
                         I
                         +----+----+----+----+----+----+----+----+----
                         0    5   10   15   20   25   30   35   40
```

## Normal distribution

To simulate 150 values from a standard normal distribution (i.e., normal distribution with zero mean and variance 1), we can enter

```
-->SIMULATE   XNORM.  NOBS ARE 150.  DISTRIBUTION IS N(0,1).  @
-->           SEED IS GSEED.
```

```
THE NORMAL DISTRIBUTION SPECIFIED IS SIMULATED, THE RESULT IS
STORED IN VARIABLE XNORM
```

The specifications here are similar to those used in the simulation of the uniform (-1, 1) data above. We also include the sentence SEED.

Simulated data are derived from a sequence of pseudo random numbers. These pseudo random numbers are created by a random number generator. The generator requires an initial seed value from which to generate its first value. The random number generator creates both a random number and a new seed for the next value. If no initial seed is specified in the SIMULATE paragraph, the default value of 1234567 is used. Since the variable GSEED is

undefined, the default value is used in the simulation of the normal data. After simulation, the value last created as a seed value is stored in GSEED. This seed can be used for subsequent simulations.

It is important to use the SEED sentence when generating more than one data set. If the SEED sentence is not employed, then the same initial seed value (i.e., 1234567) will be used for each data set. If we employ the SEED sentence, in the manner used above, then a new initial seed will be used for each new data set.

We can check the adequacy of our simulation by using the DESCRIBE paragraph to compute summary statistics of XNORM, and the HISTOGRAM paragraph to observe the distribution of the values in XNORM. Statistics are summarized below, together with theoretic values.

| | Sample statistic | Theoretic value |
|---|---|---|
| Mean | 0.1564 | 0.0 |
| Variance | 1.0556 | 1.0 |
| Skewness | -0.1565 | 0.0 |
| Kurtosis | 0.1391 | 0.0 |

-->HISTOGRAM XNORM. INTERVAL IS 6.

```
  LOWER     UPPER  FREQ- 0   10   20   30   40   50   60   70   80
  BOUND     BOUND  UENCY +----+----+----+----+----+----+----+----+----
                         I
 -3.600 -   -2.400     1 IX
                         IX
                         I
 -2.400 -   -1.200    14 IXXXXXXX
                         IXXXXXXX
                         I
 -1.200 -     .000    45 IXXXXXXXXXXXXXXXXXXXXXX
                         IXXXXXXXXXXXXXXXXXXXXXX
                         I
   .000 -    1.200    68 IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                         IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                         I
  1.200 -    2.400    20 IXXXXXXXXXX
                         IXXXXXXXXXX
                         I
  2.400 -    3.600     2 IX
                         IX
                         I
                         +----+----+----+----+----+----+----+----+----
                         0   10   20   30   40   50   60   70   80
```

## Multivariate normal distribution

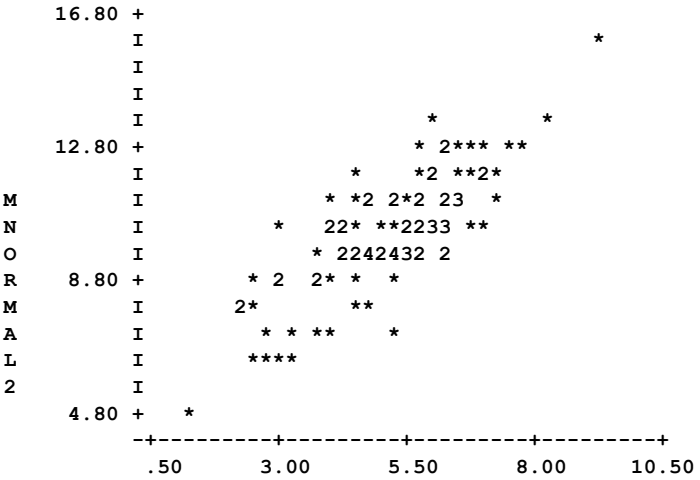
We now illustrate the use of the SIMULATE paragraph to create data from a multivariate normal distribution by simulating a bivariate normal data set. We need to specify the mean vector and covariance matrix when we simulate a multivariate normal distribution. These are specified as variables in the DISTRIBUTION sentence. We will first transmit the necessary vector of mean values and covariance matrix using the INPUT paragraph (see Chapter 2), then simulate two variables of 100 observations each.

```
-->INPUT  XMEAN, XCOV.  NCOL ARE 1,2.
-->        5.0      2.00     1.85
-->        10.0     1.85     3.00
-->END OF DATA

-->SIMULATE   MNORMAL1,MNORMAL2.  NOBS ARE 150.  OMIT 50.  @
-->                DISTRIBUTION IS MN(XMEAN,XCOV).

   THE MULTIVARIATE NORMAL DISTRIBUTION SPECIFIED IS SIMULATED
```

We have specified two variable names, MNORMAL1 and MNORMAL2, to hold the simulated data. In the SIMULATE paragraph we specified that 150 observations be simulated (in the NOBS sentence), but have also specified to OMIT the first 50 observations.

We can check the data we have simulated by observing the scatter plot (see Chapter 3) of MNORMAL1 and MNORMAL2. We can also compute the covariance matrix through the CORRELATE paragraph (see Chapter 4). We will also obtain summary information regarding the mean and standard deviation of each series. We can verify that for the specified mean vector and covariance matrix, MNORMAL1 should have a mean near 5.0 with a standard deviation near 1.414; and the values for MNORMAL2 should be near 10.0 and 1.732, respectively.

```
-->PLOT  MNORMAL2, MNORMAL1
```



-->CORRELATE MNORMAL1, MNORMAL2. TYPE IS COVAR.

```
NUMBER OF CASES TO BE ANALYZED . . .       100
NUMBER OF COMPLETE CASES . . . . . .       100


VARIABLE          MEAN    STD. DEVIATION   COEFF. OF VARIATION

MNORMAL1         5.15538         1.44997          .28125
MNORMAL2        10.20083         1.78777          .17526


CORRELATION MATRIX

MNORMAL1    1.0000
MNORMAL2     .8026   1.0000
          MNORMAL1 MNORMAL2


COVARIANCE MATRIX

MNORMAL1       2.1024
MNORMAL2       2.0806      3.1961
             MNORMAL1    MNORMAL2
```

We employed the OMIT sentence in the simulation of the above multivariate data. The OMIT sentence can be useful because on occasion the data simulated later in a sequence better represent a distribution or model than the beginning portion of the simulated sequence. This is particularly true in the simulation of time dependent data (e.g., a time series) when simulated data values are used in the calculation of subsequent simulated values. In such cases, the recursive relationship being used may be "more valid" later in the simulation sequence. We may then generate more data than the number we actually desire and remove the "excess" from the beginning of the sequence. This is an unobtrusive rule that can be applied in the simulation of data from any distribution or model.

## 12.2 Simulating Time Series Data

We can employ the SIMULATE paragraph and the TSMODEL paragraph (see Chapter 10) to simulate data that follows a univariate time series model. The TSMODEL paragraph is used to specify the time series model the data should follow, and the SIMULATE paragraph generates both the noise series of the model as well as the series itself.

To illustrate this, we will simulate the following AR(1) model

$$(1 - .75B)\,X_t = 5.0 + a_t,$$

where $\sigma_a^2 = 2.5$. We will store the data in the variable XDATA. First, we will specify the AR(1) model using the TSMODEL paragraph (see Chapter 10). We will give the model the name XSIM and use XDATA as a dummy name within the MODEL sentence. We also include the logical sentence SIMULATION to indicate that this model may be used for simulation purposes.

```
-->TSMODEL   NAME IS XSIM.  MODEL IS (1 - .75*B)XDATA =  5.0 + NOISE.  @
-->          SIMULATION.

SUMMARY FOR UNIVARIATE TIME SERIES MODEL --   XSIM

------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
           VARIABLE   OR CENTERED

  XDATA     RANDOM     ORIGINAL     NONE
------------------------------------------------------------------------

 PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-     VALUE      STD     T
   LABEL       NAME    DENOM.                TRAINT              ERROR  VALUE

  1                    CNST      1      0     NONE     5.0000
  2           XDATA     AR       1      1     NONE      .7500
```

We now will use the SIMULATE paragraph to both specify the model being used for simulation and the noise series. The data are stored in the variable XDATA.

```
-->SIMULATE   MODEL IS XSIM.   NOBS ARE 200.   NOISE IS N(0.0, 2.5).

THE UNIVARIATE TIME SERIES  XDATA   IS SIMULATED USING MODEL XSIM
```

We can now check the data simulated. We can verify that the mean and variance of the simulated data should be the following

$$\mu_x = c/(1-\phi) = 5/(1-.75) = 20.0$$

$$\sigma_x^2 = \sigma_a^2/(1-\phi^2) = .25/(1-(.75)^2) \approx 5.71$$

$$\sigma_x \approx 2.39$$

In addition, the ACF of the data should be $(.75)^{\ell}$, $\ell$ = 1, 2, . . . ; and the PACF of the data should be .75 for $\ell$ = 1; and be 0 for $\ell$ = 2, 3, . . . . We can compute and display these statistics using the IDEN paragraph (see Chapter 10).

```
-->IDEN   XDATA.   MAXLAG IS 10.

TIME PERIOD ANALYZED . . . . . .  1  TO   200
NAME OF THE SERIES . . . . . .        XDATA
EFFECTIVE NUMBER OF OBSERVATIONS . . .    200
STANDARD DEVIATION OF THE SERIES . . .   2.1851
MEAN OF THE (DIFFERENCED) SERIES . . .  20.8458
STANDARD DEVIATION OF THE MEAN . . . .    .1545
T-VALUE OF MEAN (AGAINST ZERO) . . . . 134.9183
```

```
AUTOCORRELATIONS

          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+
                                     I
  1    .69                        +  IXX+XXXXXXXXXXXXX
  2    .51                      +    IXXXX+XXXXXXXX
  3    .37                      +    IXXXX+XXXX
  4    .21                     +     IXXXXX+
  5    .07                     +     IXX   +
  6    .02                     +     I     +
  7   -.04                     +    XI     +
  8   -.10                     +   XXI     +
  9   -.15                     + XXXXI     +
 10   -.17                     + XXXXI     +


PARTIAL AUTOCORRELATIONS

          -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
            +----+----+----+----+----+----+----+----+----+----+
                                     I
  1    .69                        +  IXX+XXXXXXXXXXXXX
  2    .07                        +  IXX+
  3    .00                        +  I  +
  4   -.13                        XXXI  +
  5   -.08                        +XXI  +
  6    .03                        +  IX +
  7   -.03                        + XI  +
  8   -.05                        + XI  +
  9   -.09                        +XXI  +
 10   -.04                        + XI  +
```

The sample statistics are in reasonable agreement with the theoretic values. As noted previously, we may also wish to specify a larger number of observations that we desire, and OMIT the excess from the beginning of the series.

We did not use a variable name in the above SIMULATE paragraph as we had embedded the name in the MODEL sentence of the TSMODEL paragraph. If we use a variable name in the SIMULATE paragraph, then the simulated data will be stored under the name specified.

For example, if we had specified

```
-->SIMULATE   YDATA.   MODEL IS XSIM.   NOBS ARE 250.   OMIT 50.  @
-->           NOISE IS N(0.0, 2.5).
```

then the last 200 observations simulated would be stored in the variable YDATA. The variable XDATA (used in the model XSIM) remains unchanged, or may be undefined if it has been defined previously.

# SUMMARY OF THE SCA PARAGRAPH IN CHAPTER 12

This section provides a summary of the SCA paragraph employed in this chapter. An SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

In this section, we provide a summary of the SIMULATE paragraph.

Legend (see Chapter 2 for further explanation)

v : variable name
i : integer
r : real value

## SIMULATE Paragraph

The SIMULATE paragraph is used to generate data according to a user specified distribution or a univariate time series model. The distributions currently available are uniform (U), normal (N), and multivariate normal distribution (MN). A univariate time series model must have been specified previously using the TSMODEL paragraph (see Chapter 10).

## Syntax for the SIMULATE Paragraph

### (I) Distribution Simulation

```
SIMULATE   VARIABLES ARE v1, v2, --- .                    @
           DISTRIBUTION IS distribution(parameters).      @
           NOBS IS i.                                     @
           SEED IS i or v.                                @
           OMIT IS i.

Required sentences: VARIABLE, DISTRIBUTION and NOBS
```

### (II) Model Simulation

```
SIMULATE   VARIABLE IS v.                                 @
           MODEL IS model-name.                           @
           NOISE IS distribution (parameters) or VARIABLE(v). @
           NOBS IS i.                                     @
           SEED IS i                                      @
           OMIT IS i.

Required sentences: MODEL, NOISE and NOBS
```

## Sentences Used in the SIMULATE Paragraph

**VARIABLES sentence**

The VARIABLES sentence is used to specify the name(s) of the variable(s) to store the simulation results. For simulation of univariate data, only one variable may be specified. For simulation of multivariate data, such as that of a vector sequence following a multivariate normal distribution, the number of variables must be the same as the number of variates. The sentence is not required if a univariate time series is generated.

**MODEL sentence**

The MODEL sentence is used to specify the name (label) of the model to be simulated. The model may be an ARIMA model specified in a TSMODEL paragraph (see Chapter 10). The sentence SIMULATION must also appear in the TSMODEL paragraph.

**DISTRIBUTION sentence**

The DISTRIBUTION sentence is used to specify the type of distribution to be simulated. The following distributions can be simulated:

| | | |
|---|---|---|
| U(r1,r2) | : | uniform distribution between r1 and r2 |
| N(r1,r2) | : | normal distribution with mean r1 and variance r2 |
| MN(v1,v2) | : | multivariate normal distribution with mean vector v1 and covariance matrix v2. Note that v1 and v2 must be names of variables defined previously. |

**NOISE sentence**

The NOISE sentence is used to specify the noise sequence for the simulated time series model. Either the distribution for generating the noise sequence or the name of a variable containing values to be used as the sequence is specified. A distribution must be one of those described in the DISTRIBUTION sentence.

**NOBS sentence**

The NOBS sentence is used to specify the number of observations or cases to be simulated.

**SEED sentence**

The SEED sentence is used to specify an integer or the name of a variable for starting the random number generation. When a variable is used, the seven digit value 1234567 is used as a seed if it is not defined yet, or the value of the variable is used if the variable is an existing one. After the simulation, the variable contains the seed last used. The number of digits for the seed must not be more than 8 digits. The default is 1234567.

**OMIT sentence**

The OMIT sentence is used to specify the number of observations to be omitted at the beginning of the simulated data.

# APPENDIX A

# ANALYTIC FUNCTIONS AND MATRIX OPERATIONS

The SCA System provides a wide array of analytic functions and matrix operations to augment its statistical capabilities. This appendix provides basic information regarding these analytic capabilities. More complete information can be found in *The SCA Statistical System: Reference Manual for Fundamental Capabilities*.

## A.1 Basic Operations

The SCA System treats a variable in its workspace as a matrix. For example, a scalar variable is stored as a 1x1 matrix, and a vector variable is stored as a nx1 matrix. By storing data in this manner, analytic operations can be computed more efficiently.

To illustrate the use of some basic mathematical operations in the SCA System, suppose the following vectors are stored in the SCA workspace

$$\text{XDATA} = \begin{bmatrix} 100 \\ 200 \\ 300 \end{bmatrix} \qquad \text{YDATA} = \begin{bmatrix} 20 \\ 50 \\ 30 \end{bmatrix} \qquad \text{ZDATA} = \begin{bmatrix} 5 \\ 8 \\ 4 \end{bmatrix}$$

If we wish to add XDATA and YDATA together, storing the results in NEWDATA, we simply enter

```
-->NEWDATA = XDATA + YDATA
```

NEWDATA now contains the results. The SCA System will not display the result automatically. However, we can print the contents of NEWDATA by entering

```
-->PRINT NEWDATA
```

We also have access to common mathematic functions. For example

```
-->CDATA = LN(YDATA)
-->SDATA = SQRT(ZDATA)
```

stores the natural logarithm of each element of YDATA and the square root of each element of ZDATA in CDATA and SDATA, respectively.

We are not limited to the number of operations used in an assignment statement. For example, suppose we enter

```
-->RESULT = ZDATA * SQRT(YDATA) - (LN(XDATA) + 2 )
```

For corresponding elements in XDATA, YDATA and ZDATA, we will take the natural logarithm of XDATA and add the value 2. This quantity is subtracted from the product of ZDATA and the square root of YDATA.

The SCA System will follow the usual order of mathematical operations for an expression. The following order is observed

| | |
|---|---|
| 1st | Evaluation of a function |
| 2nd | Exponentiation (**) |
| 3rd | Multiplication or division |
| 4th | Addition or subtraction |

The above hierarchy is first applied to all parenthetical expressions. The order is applied again using resultant values, if any, as operations are read in a left to right fashion.

## A.2 Trigonometric and Hyperbolic Functions

We have access to the following trigonometric and hyperbolic functions: sin, cos, tan (and their inverses), sinh, cosh, and tanh. We need to keep in mind that the arguments of sin, cos, tan, sinh, cosh, and tanh are in radians and results of the inverses of sin, cos, and tan will be in radians. For this reason, it is useful to know how to obtain $\pi$ and the conversion factor between radians and degrees within the SCA System.

$\pi$ = 2 * ACOS(0) (i.e., $2\cos^{-1}(0)$)

$1^{\circ} = \frac{\pi}{180}$ radians = [ACOS(0)/90] radians

1 radian = [90/ACOS(0)] degrees

## A.3 Statistical and Probability Distribution Functions

The SCA System provides a wide array of commonly used statistical functions and probability distribution functions. The distribution functions include the cumulative distribution (and inverse distribution) of the standard normal, student's t, $\chi^2$, F and Beta distributions.

### Statistical Functions

To illustrate some statistical functions, suppose the variable X1 consists of the following 17 values

16, 22, 21, 20, 23, 21, 19, 15, 13, 23, 17, 20, 29, 18, 22, 16, 25

We can compute and retain the sample mean, median and the geometric mean of X1 by entering

```
-->X1MEAN = MEAN (X1)
-->X1MEDIAN = MEDN (X1)
-->X1GEOM = GMEN (X1)
```

We can display these values by entering

```
-->PRINT   X1MEAN, X1MEDIAN, X1GEOM

X1MEAN   IS A 1 BY 1 VARIABLE
X1MEDIAN IS A 1 BY 1 VARIABLE
X1GEOM   IS A 1 BY 1 VARIABLE

VARIABLE    X1MEAN   X1MEDIAN   X1GEOM
COLUMN-->      1          1         1
  ROW

   1        20.000    20.000    19.625
```

In similar fashion we can calculate and retain the variance or standard deviation of the data. Descriptive statistics can also be obtained through the DESCRIBE paragraph (see Chapter 4).

## Probability Distribution Functions (CDF)

We can quickly determine the cumulative distribution of a value following a standard normal, t, $\chi^2$, F, or Beta distribution. For example, the CDF for a value of 1.57 of a t-distribution with 16 degrees of freedom can be computed (and stored in the variable CVALUE) by entering

```
-->CVALUE = CDFT (1.57, 16)
```

Similarly, we can obtain values of critical levels from these distributions using the inverse cumulative distribution function. For example, the z-value used for a 90% confidence interval for a standard normal distribution is 1.645. We can confirm this by computing the inverse CDF of the standard normal for the value .95. We can obtain this by entering

```
-->ZSCORE = IDFN(.95)
```

## A.4 Matrix Operations

To illustrate some of the available matrix operations in the SCA System, we will assume the following matrices are in the SCA workspace

$$\text{ADATA} = \begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 0 & 1 \end{bmatrix} \qquad \text{BDATA} = \begin{bmatrix} 1 & 3 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

We can perform matrix multiplication using the symbol '#'. (Note that element by element multiplication occurs if we use the symbol '*'.) For example, if we enter

-->C1DATA = BDATA # ADATA <result>

$$\begin{bmatrix} 10 & 4 \\ 5 & 3 \\ 3 & 0 \end{bmatrix}$$

then C1DATA contains the above matrix product. To display C1DATA we need to employ the PRINT paragraph. We have inserted the values of the resultant matrix above for reference only. We shall continue to do this below.

The matrix product ADATA # BDATA has no sense, since the matrices are not conformable. However, the transpose of ADATA is conformable with BDATA, and we can compute this matrix product by entering

-->C2DATA = T(ADATA)#BDATA <result>

$$\begin{bmatrix} 7 & 6 & 0 \\ 3 & 5 & -1 \end{bmatrix}$$

We may also compute the Kronecker product of ADATA and BDATA, the trace of BDATA and the Cholesky decomposition of BDATA, among other operations. We can compute the determinant, inverse, and adjoint matrix of BDATA by entering

-->DETB = DET(BDATA) <result>

[5]

-->BINVERSE = INV(BDATA) <result>

$$\begin{bmatrix} -.2 & .6 & 0 \\ .4 & -.2 & 0 \\ .4 & -.2 & -1 \end{bmatrix}$$

-->ADJOINTB = DETB * BINVERSE <result>

$$\begin{bmatrix} -1 & 3 & 0 \\ 2 & -1 & 0 \\ 2 & -1 & 5 \end{bmatrix}$$

## Eigenvalues

We can compute the eigenvalues and eigenvectors of any real matrix. For example, suppose we have the following matrix in the SCA workspace

$$\text{EDATA} = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix}$$

We can compute its eigenvalues and eigenvectors by entering

-->EIGEN EDATA. VALUES IN EVAL. VECTORS IN EVEC.

```
EIGENVALUES FOR THE MATRIX EDATA

              1         2         3
  1    4.00000   3.00000   1.00000

EIGENVECTORS FOR THE MATRIX EDATA

              1         2         3
  1     .57735    .70711   -.40825
  2    -.57735 .8412E-16   -.81650
  3     .57735   -.70711   -.40825
```

The VALUES and VECTORS sentences were specified so that the computed eigenvalues and corresponding matrix of eigenvectors would be maintained in the SCA workspace (under the labels EVAL and EVEC, respectively).

## A.5 Summary of Analytic Functions and Syntax for the EIGEN Paragraph

Listed below is a brief list of the analytic capabilities in the SCA System. More complete information is available in Chapter 4 of The SCA Statistical System: Reference Manual for Fundamental Capabilities.

**ABS(A)** -- absolute value of each element in variable A

**AND** -- A AND B; logical operator on binary scalars

**ACOS(A)** -- inverse cosine of each element in variable A

**ASIN(A)** -- inverse sine of each element in variable A

**ATAN(A)** -- inverse tangent of each element in variable A

**CDFB(X,A,B)** -- cumulative distribution function of beta distribution with scale parameters A and B; $0 \le X \le 1$

**CDFC(X,N)** -- cumulative distribution function of chi-square distribution with N degrees of freedom; X positive

**CDFF(X,M,N)** -- cumulative distribution function of F-distribution with M and N d.f.; X positive

**CDFN(X)** -- standard normal cumulative distribution function

**CDFT(X,N)** -- cumulative distribution function of Student's t-distribution with N degrees of freedom

**CDP(A,B)** -- column direct product of matrices A and B

**CHOL(A)** -- Cholesky decomposition of matrix A

**COS(A)** -- cosine of each element in variable A

**COSH(A)** -- hyperbolic cosine of elements in variable A

**DET(A)** -- determinant of matrix A

**EQ** -- A EQ B; logical comparison over all elements

**EIGEN** -- see the EIGEN paragraph

**EXP(A)** -- exponential function applied to elements in A

**FACT(A)** -- factorial value for each element in A

**GAMA(A)** -- gamma function applied to elements in A

**GE** -- A GE B; logical comparison over all elements

**GMEN(A)** -- geometric mean of the elements in variable A

**GT** -- A GT B; logical comparison over all elements

**IDFB(X,A,B)** -- inverse distribution function of beta distribution with scale parameters A and B;

**IDFC(X,N)** -- inverse distribution function of chi-square distribution with N d.f.; $0 \le X \le 1$

**IDFF(X,M,N)** -- inverse distribution function of F-distribution with M and N d.f.; $0 \le X \le 1$

| | | |
|---|---|---|
| **IDFN(X)** | -- | inverse distribution function of standard normal distribution (also known as the PROBIT function); $0 \le X \le 1$ |
| **IDFT(X,N)** | -- | inverse distribution function of t-distribution with N d.f.; $0 \le X \le 1$ |
| **INT(A)** | -- | largest integer value of each element of A |
| **INV(A)** | -- | inverse of matrix A |
| **KP(A,B)** | -- | Kroneker product of matrices A and B |
| **LE** | -- | A LE B; logical comparison over all elements |
| **LN(A)** | -- | natural logarithm of each element in A |
| **LOG(A)** | -- | base 10 logarithm of each element in A |
| **LT** | -- | A LT B; logical comparison over all elements |
| **MAX(A)** | -- | maximum value of the elements in A |
| **MEAN(A)** | -- | arithmetic mean of the elements of A |
| **MEDN(A)** | -- | median value of the elements of A |
| **MIN(A)** | -- | minimum value of the elements in A |
| **MMAX(A,B)** | -- | element by element maximum value in A and B |
| **MMIN(A,B)** | -- | element by element minimum value in A and B |
| **MOD(A,B)** | -- | modular arithmetic; A(i,j)(modula B(i,j)) |
| **NCOL(A)** | -- | number of columns in matrix A |
| **NE** | -- | A NE B; logical comparison over all elements |
| **NMIS(A)** | -- | number of missing values in A |
| **NOT** | -- | NOT A; logical operator on binary scalars |
| **NROW(A)** | -- | number of rows in matrix A |
| **OR** | -- | A OR B; logical operator on binary scalars |
| **RDP(A,B)** | -- | row direct product of matrices A and B |
| **SIGN(A,B)** | -- | transfer of the sign of an element of B to the absolute value of the corresponding element of A |
| **SIN(A)** | -- | sine of each element in A |
| **SINH(A)** | -- | hyperbolic sine of each element in A |
| **SQRT(A)** | -- | square root of each element in A |
| **STD(A)** | -- | sample standard deviation of elements of A |
| **STD1(A)** | -- | unbiased sample st. dev. of elements of A |
| **SUM(A)** | -- | arithmetic sum of all elements in A |
| **T(A)** | -- | transpose of the matrix A |
| **TAN(A)** | -- | tangent of each element in A |
| **TANH(A)** | -- | hyperbolic tangent of each element in A |
| **TR(A)** | -- | trace of the matrix A |
| **VAR(A)** | -- | sample variance of the elements of A |
| **VAR1(A)** | -- | unbiased sample variance of the elements of A |
| + | -- | A + B; element by element addition |

| | | |
|---|---|---|
| - | -- | A - B; element by element subtraction |
| * | -- | A * B; element by element multiplication |
| / | -- | A / B; element by element division |
| ** | -- | A**B ; element by element exponentiation |
| # | -- | A # B; matrix multiplication |

## Syntax for the EIGEN Paragraph

The EIGEN paragraph is used to compute and display the eigenvalues and eigenvectors of any real matrix. The EIGEN paragraph begins with the paragraph name, EIGEN, and may be followed by various modifying sentences. Sentences that may be used as modifiers for this paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not listed as required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line, except the last line, must be the continuation character, '@'.

Legend (see Chapter 2 for further explanation):

v : variable name
w : keyword

| | | |
|---|---|---|
| **EIGEN** | MATRIX IS v. | @ |
| | VALUES IN v. | @ |
| | VECTORS IN v. | @ |
| | ORDER IS w. | |

Required sentence: **MATRIX**

## Sentences Used in the EIGEN Paragraph

**MATRIX sentence**

The MATRIX sentence is used to specify the name of the matrix for which eigenvalues and eigenvectors will be computed.

**VALUES sentence**

The VALUES sentence is used to specify the name of the variable to store the computed eigenvalues of the matrix.

**VECTORS sentence**

The VECTORS sentence is used to specify the name of the variable to store the computed eigenvectors of the matrix. Eigenvectors are stored columnwise; that is, the first column corresponds to the first eigenvalue, and so on.

**ORDER sentence**

The ORDER sentence is used to specify the order that the eigenvalues and their corresponding eigenvectors will be stored. The keyword may be DESCENDING or ASCENDING. The default is DESCENDING.

# APPENDIX B

# DATA GENERATION, EDITING AND MANIPULATION

The SCA System provides several capabilities to generate, edit and manipulate data stored in the SCA workspace. This appendix provides selected information on these capabilities. More complete information can be found in *The SCA Statistical System: Reference Manual for Fundamental Capabilities* and *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*. Features discussed in this appendix, and the section containing them, are:

| Section | Feature(s) |
|---|---|
| B.1 | Generation of a vector or matrix variable |
| B.2 | Modification of the existing values of a variable |
| B.3 | Manipulation of variables |
| B.4 | Editing time series data |

## B.1 Generating Data: the GENERATE Paragraph

We can use the GENERATE paragraph to create data, either by direct value specification or following one of two patterns, and store the data within a vector or matrix. We will illustrate the use of the paragraph with some examples.

### B.1.1 Generating a vector

We will now create four variables, each stored in the SCA workspace as a 10x1 vector (column) of data. Variables created illustrate the various manners that data can be created. First, we will generate and print the data. Afterwards, we will explain what has been created.

```
-->GENERATE  VECTOR1.  NROW ARE 10.  VALUES ARE 0 FOR 5, 1 FOR 5.
   THE SINGLE PRECISION VARIABLE VECTOR IS GENERATED

-->GENERATE  VECTOR2.  NROW ARE 10.  VALUES ARE 0 FOR 5, 1 FOR 2, 0 FOR 3.
   THE SINGLE PRECISION VARIABLE VECTOR2  IS GENERATED

-->GENERATE  VECTOR3.  NROW ARE 10.  PATTERN IS STEP (1.0, 0.5).
   THE SINGLE PRECISION VARIABLE VECTOR3  IS GENERATED

-->GENERATE  VECTOR4.  NROW ARE 10.  PATTERN IS RATE (1.0, 2.0).
   THE SINGLE PRECISION VARIABLE VECTOR4 IS GENERATED
```

-->PRINT VECTOR1, VECTOR2, VECTOR3, VECTOR4

| VARIABLE | VECTOR1 | VECTOR2 | VECTOR3 | VECTOR4 |
|---|---|---|---|---|
| COLUMN--> | 1 | 1 | 1 | 1 |
| ROW | | | | |
| 1 | 0.000 | 0.000 | 1.000 | 1.000 |
| 2 | 0.000 | 0.000 | 1.500 | 2.000 |
| 3 | 0.000 | 0.000 | 2.000 | 4.000 |
| 4 | 0 000 | 0.000 | 2.500 | 8.000 |
| 5 | 0.000 | 0.000 | 3.000 | 16.000 |
| 6 | 1.000 | 1.000 | 3.500 | 32.000 |
| 7 | 1.000 | 1.000 | 4.000 | 64.000 |
| 8 | 1.000 | 0.000 | 4.500 | 128.000 |
| 9 | 1.000 | 0.000 | 5.000 | 256.000 |
| 10 | 1.000 | 0.000 | 5.500 | 512.000 |

In each use of the GENERATE paragraph, we specified the number of rows of data (NROW) to be created as 10. The default number of rows and columns to create is 1. Hence, unless we are creating a scalar, we need to specify the number of rows or/and columns in our variable.

In the above example, we directly entered the values that comprise VECTOR1 and VECTOR2. In VECTOR1, the VALUES of the first 5 points are set to 0 and the next 5 are set to 1. In VECTOR2, the first 5 points are set to 0, the next 2 are set to 1, and the remaining 3 are set to 0. A PATTERN is used to generate the data in both VECTOR3 and VECTOR4. VECTOR3 follows a STEP function. Its first value is 1.0, and each successive value is 0.5 more than the last value. That is, for STEP (a, b) our data are described as

$$X_i = a + (i-1)b, \qquad i = 1, 2, .....$$

The data in VECTOR4 follows a geometric pattern. The initial value is 1.0 and successive values are 2.0 times the previous value. Thus, when we specify the geometric RATE (a,b), our data follow the pattern

$$X_i = a * b^{i-1}, \qquad i = 1, 2, ....$$

## Use of analytic functions

We can use the GENERATE paragraph in conjunction with analytic functions or editing capabilities of the SCA System (see Appendices A and latter sections of this Appendix, and *The SCA Statistical System: Reference Manual for Fundamental Capabilities*) to create variables with more intricate structure.

For example, we could have also created VECTOR2 above by first generating a vector of zeros by entering

-->GENERATE VECTOR2. NROW ARE 10. VALUES ARE 0 FOR 10.
**THE SINGLE PRECISION VARIABLE VECTOR2 IS GENERATED**

Then we could recode the 6th and 7th observations as 1 using the simple assignments

```
-->VECTOR2(6) = 1.0
-->VECTOR2(7) = 1.0
```

As a more intricate illustration, suppose we are to study 15 years of quarterly sales data of a corporation. The end of the fiscal year is June, and some of the sale activity in the second quarter are related to end of year quotas or bonuses. We intend to "isolate" the second quarter by including an indicator variable that is 1 for a second quarter and 0 otherwise. We can use the GENERATE paragraph and row direct product (RDP) analytic function for this purpose.

First we will generate two vectors, one will describe the yearly pattern of the indicator (i.e., 0, 1, 0, 0). The second vector represents the number of times this pattern should be applied. We can enter

```
-->GENERATE  VECTOR5.  NROW ARE 4.  VALUES ARE 0, 1, 0, 0
   THE SINGLE PRECISION VARIABLE VECTOR5  IS GENERATED

-->GENERATE  VECTOR6  NROW ARE 15.  VALUES ARE 1 FOR 15.
   THE SINGLE PRECISION VARIABLE VECTOR6  IS GENERATED
```

We now compute the row direct product (see Appendix A and *The SCA Statistical System: Reference Manual for Fundamental Capabilities*) to create our desired indicator variable. We will call this variable INDC1.

```
-->INDC1 = RDP(VECTOR5, VECTOR6)
```

We have created a variable with 60 values, all are 0 except for the 2nd, 6th, 10th, and so on. These values are all 1. We can see this by printing INDC1.

```
-->PRINT  INDC1.  FORMAT IS  '8F10.2'.

   INDC1    IS  A    60  BY     1  VARIABLE
         .00     1.00       .00      .00       .00      1.00       .00       .00
         .00     1.00       .00      .00       .00      1.00       .00       .00
         .00     1.00       .00      .00       .00      1.00       .00       .00
         .00     1.00       .00      .00       .00      1.00       .00       .00
         .00     1.00       .00      .00       .00      1.00       .00       .00
         .00     1.00       .00      .00       .00      1.00       .00       .00
         .00     1.00       .00      .00       .00      1.00       .00       .00
         .00     1.00       .00      .00
```

## B.1.2 Generating a matrix

We can also use the GENERATE paragraph to create matrices. In such cases, we must include information regarding the number of rows and columns of the matrix (NROW and NCOL, respectively) and the manner in which we want data stored. For example, we can create a 4 x 4 identity matrix by entering

```
-->GENERATE  MATRIX1.  NROW ARE 4.  NCOL ARE 4.     @
-->        VALUES ARE 1 FOR 4.  ORDER IS DIAGONAL.
```

```
THE SINGLE PRECISION VARIABLE MATRIX1  IS GENERATED
```

We have specified that the ORDER to store data is along the DIAGONAL. In this manner all values are entered sequentially on the diagonal of the matrix only. All off diagonal elements are set to zero. If no ORDER is specified, values are stored column by column. That is, data is entered in the first column from "top" to "bottom", then the second column, third column, and so on. Hence if we enter

-->GENERATE MATRIX2. NROW ARE 4. NCOL ARE 4. PATTERN IS STEP(1.0, 2.0)

we create the following matrix

```
1.0       9.0      17.0      25.0
3.0      11.0      19.0      27.0
5.0      13.0      21.0      29.0
7.0      15.0      23.0      31.0
```

We can also choose to have data stored row by row, symmetrically or skew symmetrically. In symmetric storage, data are stored row by row in the lower triangle of the matrix and values of the upper triangle are set equal to their corresponding lower triangular entry. Skew symmetric storage is similar, except the values of the upper triangle are set equal to the negative of their corresponding lower triangular entry. We illustrate this storage in the next section.

## Use of analytic functions

Analytic functions (see Appendix A) can be used in conjunction with the GENERATE paragraph to create matrices of more complicated structure. For example, earlier we created an indicator variable corresponding to the second quarter of each year in a fifteen year period. Now we will construct a matrix whose columns consist of the indicators for the first, second, third and fourth quarters of a year for the same fifteen year period.

To accomplish this we will use the 4 x 4 identity matrix generated earlier and stored as MATRIX1. Each of its columns represents an indicator associated with a quarter of a given year. We also need a matrix equivalent to the number of times this periodic pattern should appear. We can then use the RDP function as before to create the desired matrix.

-->GENERATE MATRIX3. NROW ARE 15. NCOL ARE 4. VALUES ARE 1 FOR 60.

```
THE SINGLE PRECISION VARIABLE MATRIX3  IS GENERATED
```

-->INDC2 = RDP(MATRIX1, MATRIX3)

We will print the first 11 rows of the resultant matrix, INDC2, to observe the pattern we have created.

-->PRINT INDC2. SPAN IS 1, 11.

```
INDC2    IS  A    60  BY     4  VARIABLE

VARIABLE     INDC2     INDC2     INDC2     INDC2
COLUMN-->      1         2         3         4
  ROW
    1        1.000      .000      .000      .000
    2         .000     1.000      .000      .000
    3         .000      .000     1.000      .000
    4         .000      .000      .000     1.000
    5        1.000      .000      .000      .000
    6         .000     1.000      .000      .000
    7         .000      .000     1.000      .000
    8         .000      .000      .000     1.000
    9        1.000      .000      .000      .000
   10         .000     1.000      .000      .000
   11         .000      .000     1.000      .000
```

To illustrate skew symmetric storage and analytic operations, we now create a 4x4 matrix whose lower tridiagonal and diagonal elements are 1 and whose upper tridiagonal elements are 0.

-->GENERATE MATRIX4. NROW ARE 4. NCOL ARE 4. VALUES ARE 1 FOR 16.

```
THE SINGLE PRECISION VARIABLE MATRIX4  IS GENERATED
```

-->GENERATE MATRIX5. NROW ARE 4. NCOL ARE 4. @
--> PATTERN IS STEP (1.0, 0.0). ORDER IS SKEWSYMMETRIC.

```
THE SINGLE PRECISION VARIABLE MATRIX5  IS GENERATED
```

-->MATRIX6 = (MATRIX4 + MATRIX5)/2 + MATRIX1

MATRIX4 is a 4 x 4 matrix of 1's and MATRIX5 is a 4 x 4 matrix whose lower tridiagonal elements are 1's and whose other elements (including the diagonal) are –1's. Adding these matrices together "zeroes out" the upper tridiagonal and the diagonal.

All values in the resultant lower tridiagonal matrix (excluding the diagonal) are 2. If we divide this result by 2 and add the identity matrix (MATRIX1) we obtain our desired matrix. We can observe MATRIX5 and the resultant MATRIX6 by entering

-->PRINT MATRIX5, MATRIX6. FORMAT IS '(4F8.1,2X,4F8.1)'

```
MATRIX5  IS  A     4  BY     4  VARIABLE
MATRIX6  IS  A     4  BY     4  VARIABLE

VARIABLE   MATRIX5 MATRIX5 MATRIX5 MATRIX5   MATRIX6 MATRIX6 MATRIX6 MATRIX6
COLUMN-->      1       2       3       4         1       2       3       4
  ROW
    1        -1.0    -1.0    -1.0    -1.0       1.0      .0      .0      .0
    2         1.0    -1.0    -1.0    -1.0       1.0     1.0      .0      .0
    3         1.0     1.0    -1.0    -1.0       1.0     1.0     1.0      .0
    4         1.0     1.0     1.0    -1.0       1.0     1.0     1.0     1.0
```

## B.2 Modification of Data in a Variable

To illustrate the modification of data in a variable in the SCA workspace, we will suppose the data listed in Table 1 represents the percent concentration of a certain chemical in the yield of some process. The data are stored in the SCA workspace under the label CONC. The value -1.00 is used to denote a missing value.

**Table 1**

**Percent Concentration of Chemical in a Process Yield**
**(Read data across a line)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 24.57 | 24.79 | 22.91 | 25.84 | 25.35 | -1.00 | -1.00 | 29.65 | 226.10 | 23.38 |
| 25.10 | 28.03 | 29.09 | 29.34 | 24.41 | 25.12 | 25.27 | 27.46 | 27.65 | 27.95 |
| 22.87 | 22.95 | 24.36 | 26.32 | 24.05 | 28.27 | 26.57 | -1.00 | 24.35 | 30.04 |
| 25.18 | 27.42 | 24.50 | 23.21 | 25.10 | 23.59 | 26.98 | 22.94 | 25.27 | 25.84 |
| 27.18 | 24.69 | 26.35 | 23.05 | 23.37 | 25.46 | 28.84 | 30.09 | 25.42 | 30.11 |

## Use of analytic statements

The value of the 9th observation, 226.10, stands out. It may be this is a simple entry error that must be corrected. The value should be 26.10. We can quickly change the value by entering

-->CONC(9) = 26.10

We can do the same with data stored in matrix form, all we need to do is to indicate the (i,j) position.

Analytic statements are also convenient for scaling data. For example, suppose the independent variables of a regression are X1DATA and X2DATA, with the values of X1DATA between 1,000,000 and 5,000,000 and the values of X2DATA between 10 and 25. For computational purposes, it is useful to have these two variables around the same scale. We can scale X1DATA by entering

```
-->X1DATA = X1DATA/1000000
```

If we also want the data in our second variable to represent a percentage relative to the first term, we can enter

```
-->X2DATA = X2DATA/X2DATA(1) * 100
```

## Recoding ranges of values

For the data of CONC, suppose we know that the minimum percent of concentration in the yield is 23 and the maximum is 30. Any value outside these limits are due to measurement errors, and within the analysis to be performed it is important that the limits not be exceeded. Further, if regression analysis is employed (see Chapter 9), we know that missing entries are excluded automatically, provided the internal missing value code is used for the value. Hence, we want to do the following:

Recode all values over 30.0 to 30.0,
Recode all values under 23.0 to 23.0, and
Assign the internal missing value code to any value that is presently -1.0 .

We can accomplish this directly using the RECODE paragraph. If we enter

```
-->RECODE  CONC.  NEW IS CONC2.  VALUES ARE (0.0, 23.0, 23.0),  @
          (30.0, 100.0, 30.0), (-1.0, -1.0, MISSING).
```

then all data within the range 0.0 to 23.0 is recoded to 23.0; all data within the range 30.0 to 100.0 is recoded to 30.0; and the value -1.0 is recoded to the internal missing value code. The altered data are stored in the new variable CONC2. If no NEW variable is specified, then the data are stored in the original variable, CONC.

## B.3 Manipulation of Variables

To illustrate some of the capabilities to manipulate data within SCA, we will suppose the following variables are in the SCA workspace:

| A1DATA | | C1 | C2 | C3 |
|---|---|---|---|---|
| 4.5 | 4.7 | 2.9 | 5.8 | 5.3 |
| 1.0 | 0.1 | 9.6 | 1.0 | 3.3 |
| 5.2 | 3.8 | 9.1 | 9.3 | 4.4 |
| 5.1 | 5.2 | 7.4 | 7.6 | 7.9 |
| 2.8 | 2.9 | 4.3 | 6.3 | 4.1 |
| 8.2 | 6.5 | 6.1 | 4.3 | 3.0 |
| 5.1 | 7.4 | 4.5 | 3.2 | 5.1 |
| 3.5 | 6.9 | 2.9 | 5.2 | 5.8 |
| 7.1 | 4.6 | 6.3 | 3.1 | 3.3 |
| 5.4 | 8.8 | 0.9 | 5.4 | 3.1 |

A1DATA is stored as a 10x2 matrix, while C1, C2, and C3 are each vectors of data.

### Selecting and omitting cases

We can select or omit cases of one or more variables according to either its index or its value. For example, suppose we only wish to work with the first 8 cases of C1, C2 and C3. We can enter either

```
-->SELECT  C1,C2,C3.   SPAN IS (1,8).
```

or

```
-->OMIT  C1,C2,C3.  NEW ARE D1,D2,D3.  SPAN IS (9,10).
```

for this purpose. In the SELECT paragraph, data are stored in the original variables since no NEW variables are specified. In the OMIT paragraph, data are stored in the new variables D1, D2, and D3. We can also select or omit cases based on the values assumed by the variable. For example, suppose we only want to use the data in C1 with values under 9.0, and the corresponding entries of C2 and C3. We can accomplish this by entering

```
-->SELECT  C1, C2, C3.  VALUES ARE (0.0, 8.9)
```

Here, all rows, except the 2nd and 3rd, are retained for all variables. We can specify more than one range of indices or values. For example, suppose we wish to omit all values over 7.0 and under 4.0 from C3 (and accompanying cases in C1 and C2). If we enter

```
-->OMIT  C3, C1, C2.  VALUES ARE (7.0, 100.0), (0.0, 4.0).
```

then C1, C2, and C3 will consist of the following

| C1 | C2 | C3 |
|---|---|---|
| 2.9 | 5.8 | 5.3 |
| 9.1 | 9.3 | 4.4 |
| 4.3 | 6.3 | 4.1 |
| 4.5 | 3.2 | 5.1 |
| 2.9 | 5.2 | 5.8 |

The five rows that had values in C3 either over 7.0 or under 4.0 have been removed. We see that C1 and C2 contain values in the "excluded" ranges. These values have not been deleted since the SELECT and OMIT paragraphs apply the selection (or deletion) criteria to the first column of the first variable specified, then selects (or omits) corresponding entries from all other columns and variables. In order to be certain the values of C1 and C2 are within designated ranges, we need to sequentially apply the OMIT or SELECT commands to the variables with C1, then C2, as the first variable.

## Appending data

C1, C2, and C3 are each 10 x 1 vectors. We can create one 30 x 1 vector by appending C2 to the end of C1 and C3 to the end of C2 by entering

```
-->JOIN   C1,C2,C3.  NEW IS D1.
```

The resultant vector is stored in D1. If no NEW variable is specified, then the resultant vector is stored in the first variable specified. We can also append matrices together, provided the number of columns of all matrices are the same. We cannot append vectors to the end of matrices.

If we want to append C1 to the first column of A1DATA and C2 to the second column of A1DATA, we must first create a matrix consisting of columns C1 and C2. We can create this matrix, say CMAT, by entering

```
-->AUGMENT  C1, C2.  NEW IS CMAT.
```

We can now append CMAT to A1DATA by entering

```
-->JOIN  A1DATA, CMAT
```

A1DATA will be changed to a 20 x 2 matrix.

## B.4 Editing Time Series Data

To illustrate editing capabilities for time series data, we will consider the first 40 observations of Series C of Box and Jenkins (1970). These data are assumed to be in the SCA workspace under the label SERIESC. In addition, we will omit a few values, replacing with them with missing values, to illustrate "patching" capabilities. The altered data are stored in the SCA workspace under the label SERIESCP. The data are listed in Table 2.

**Table 2**

Initial forty observations of Series C of Box and Jenkins (1970) (SERIESC) and series with missing data (SERIESCP). (Data are read across a line.)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SERIESC | 26.6 | 27.0 | 27.1 | 27.1 | 27.1 | 27.1 | 26.9 | 26.8 | 26.7 | 26.4 |
| SERIESCP | 26.6 | 27.0 | 27.1 | 27.1 | 27.1 | 27.1 | 26.9 | 26.8 | 26.7 | 26.4 |
| SERIESC | 26.0 | 25.8 | 25.6 | 25.2 | 25.0 | 24.6 | 24.2 | 24.0 | 23.7 | 23.4 |
| SERIESCP | 26.0 | 25.8 | 25.6 | **** | **** | 24.6 | 24.2 | 24.0 | 23.7 | 23.4 |
| SERIESC | 23.1 | 22.9 | 22.8 | 22.7 | 22.6 | 22.4 | 22.2 | 22.0 | 21.8 | 21.4 |
| SERIESCP | 23.1 | 22.9 | 22.8 | 22.7 | 22.6 | 22.4 | 22.2 | 22.0 | 21.8 | 21.4 |
| SERIESC | 20.9 | 20.3 | 19.7 | 19.4 | 19.3 | 19.2 | 19.1 | 19.0 | 18.9 | 18.9 |
| SERIESCP | 20.9 | **** | 19.7 | 19.4 | 19.3 | 19.2 | 19.1 | 19.0 | 18.9 | 18.9 |

### Patching missing data

Series SERIESCP has the missing data code for the value of the 14th, 15th, and 32nd observations. If we wish to analyze this series, it is advisable that we replace the missing data code with some "appropriate" value that will not adversely affect our analysis and may reasonably represent the missing data. We can use an analytic assignment statement (see Section B.1), or we can employ the PATCH paragraph. The PATCH paragraph provides us with some latitude in the recoding of time dependent data. Here, we can observe how well some of these methods perform by comparing the values used in patching the data with the actual values.

One simple scheme is to replace a missing value with the average of the values immediately adjacent to it. We can do that here by entering

```
-->PATCH  SERIESCP.  METHOD IS ADJACENT(1).
```

All missing values are replaced by the average of the values of the observations one time period from it. If two or more missing observations are next to each other, a missing value is replaced by the average of its two nearest, and equidistant, non-missing observations. Here we have

```
THE   14-TH  OBSERVATION IS RECODED TO     25.2000
THE   15-TH  OBSERVATION IS RECODED TO     24.9000
THE   32-TH  OBSERVATION IS RECODED TO     20.3000
```

Here the 32nd observation is recoded to 20.3. Since observation 15 is missing, the 14th observation is recoded to the average of the 12th and 16th values. Similarly the 15th value is recoded to the average of observations 13 and 17. We can average the values of observations two time periods from each missing observation (or span of missing observations) by entering

```
-->PATCH  SERIESCP.  METHOD IS ADJACENT(2).
```

We are informed that

```
THE   14-TH  OBSERVATION IS RECODED TO     25.0000
THE   15-TH  OBSERVATION IS RECODED TO     25.0000
THE   32-TH  OBSERVATION IS RECODED TO     20.4000
```

The recoding for the 14th and 15th observations is as before. We can see that by changing the argument in the METHOD sentence we can average "adjacent" information that is farther and farther away from a missing data point. This may be appropriate if we want to average adjacent, "December" or "1st quarter" data in the case of single missing observations. In such cases the value of the required argument of ADJACENT may be 12 or 4, respectively.

We can also replace missing data by the average of all data, or a periodic average (for seasonal data). We can use the average of all non-missing data as our replacement value by entering

```
-->PATCH  SERIESCP.  METHOD IS MEAN(1).
```

This is a reasonable way to recode missing data of a stationary time series. Since SERIESC is not stationary, and has a downward drift at its beginning, we observe this method of recoding would be inappropriate

```
THE   14-TH  OBSERVATION IS RECODED TO     23.3622
THE   15-TH  OBSERVATION IS RECODED TO     23.3622
THE   32-TH  OBSERVATION IS RECODED TO     23.3622
```

If SERIESCP represented quarterly data, we may wish to use the mean of similar quarters as a "patch". We can specify this by entering

```
-->PATCH  SERIESCP.  METHOD IS MEAN(4).
```

```
THE   14-TH  OBSERVATION IS RECODED TO     23.2889
THE   15-TH  OBSERVATION IS RECODED TO     23.0889
THE   32-TH  OBSERVATION IS RECODED TO     23.3889
```

We may only specify one method in the PATCH paragraph. If different methods are appropriate (e.g., if the structure of the data changes over time), we can combine procedures by invoking the paragraph repeatedly but with different specifications in non-overlapping time spans.

## Lagging and differencing data

The time series capabilities of the SCA System (see Chapter 10) can incorporate differencing in the identification and estimation of time series models. However, it is sometimes useful to be able to lag or to difference data separately. The LAG and DIFFERENCE paragraphs provide these capabilities.

To illustrate the LAG paragraph, suppose we enter

```
-->LAG  SERIESC.  LAGS ARE 1, 2.  NEW ARE LAGC1, LAGC2.
```

```
THE ORIGINAL SERIES IS SERIESC
THE LAG  1 SERIES IS STORED IN VARIABLE  LAGC1  ,  WHICH HAS   41  ENTRIES
THE LAG  2 SERIES IS STORED IN VARIABLE  LAGC2  ,  WHICH HAS   42  ENTRIES
```

We have generated two series, one stored in LAGC1 and the other in LAGC2. LAGC1 contains the first lag of SERIESC (that is, its first lag order). The i-th entry in LAGC1 is the (i-1)st entry of SERIESC. Hence,

$$\text{LAGC}_{15} = \text{SERIESC}_{4},$$
$$\text{LAGC1}_{20} = \text{SERIESC}_{19},$$
$$\text{LAGC1}_{41} = \text{SERIESC}_{40}$$

The value of LAGC1(1) is necessarily undefined. In like manner, LAGC2 contains the second lag order of SERIESC. As a result, the contents of these variables are

| | SERIESC | LAGC1 | LAGC2 |
|---|---|---|---|
| 1 | 26.600 | *** | *** |
| 2 | 27.000 | 26.600 | *** |
| 3 | 27.100 | 27.000 | 26.600 |
| 4 | 27.100 | 27.100 | 27.000 |
| 5 | 27.100 | 27.100 | 27.100 |
| 6 | 27.100 | 27.100 | 27.100 |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| 38 | 19.000 | 19.100 | 19.200 |
| 39 | 18.900 | 19.000 | 19.100 |
| 40 | 18.900 | 18.900 | 19.000 |
| 41 | | 18.900 | 18.900 |
| 42 | | | 18.900 |

A first lag order is assumed if the LAG sentence is not specified. Lagged values are stored as indicated above so that information is properly aligned if we wish to investigate relationships between the currently observed value of one variable and a previous (lagged) observation of another variable.

We difference data in a manner similar to lagging. For example, the first-order differenced series of SERIESC is

$$(1 - B)SERIESC_t = SERIESC_t - B(SERIESC_t) , \quad \text{or}$$
$$= SERIESC_t - SERIESC_{t-1}$$

The subscript t has been included to indicate how values are obtained. We obtain this new series by entering

-->DIFFERENCE SERIESC. NEW IS DIFFC1.

```
                              1
DIFFERENCE ORDERS ARE  (1-B  )
SERIES SERIESC  IS DIFFERENCED, THE RESULT IS STORED IN VARIABLE  DIFFC1
SERIES  DIFFC1  HAS   40  ENTRIES
```

Similarly we can calculate $(1-B)(1-B^4)SERIESC$. This result is related to what we have calculated, since

$$(1-B)(1-B^4)SERIESC_t = (1-B4)DIFFC1_t = DIFFC1_t - DIFFC1_{t-4} .$$

We can obtain this differenced series by entering

-->DIFFERENCE SERIESC. NEW IS DIFFC14. DFORDERS ARE 1, 4.

```
                              1       4
DIFFERENCE ORDERS ARE  (1-B  ) (1-B  )
SERIES SERIESC  IS DIFFERENCED, THE RESULT IS STORED IN VARIABLE DIFFC12
SERIES DIFFC12  HAS   40  ENTRIES
```

A partial listing of the values in these variables is given below

| | SERIESC | DIFFC1 | DIFFC14 |
|---|---|---|---|
| 1 | 26.600 | *** | *** |
| 2 | 27.000 | .400 | *** |
| 3 | 27.100 | .100 | *** |
| 4 | 27.100 | .000 | *** |
| 5 | 27.100 | .000 | *** |
| 6 | 27.100 | .000 | -.400 |
| 7 | 26.900 | -.200 | -.300 |
| 8 | 26.800 | -.100 | -.100 |
| 9 | 26.700 | -.100 | -.100 |
| 10 | 26.400 | -.300 | -.300 |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |

# SUMMARY OF THE SCA PARAGRAPHS IN APPENDIX B

This section provides a summary of those SCA paragraphs employed in this appendix. Each SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

The paragraphs to be explained in this summary are GENERATE, RECODE, OMIT, SELECT, JOIN, AUGMENT, PATCH, LAG, and DIFFERENCE.

Legend (see Chapter 2 for further explanation)

v : variable name
i : integer
r : real value
w(.) : keyword (with argument)

## GENERATE Paragraph

The GENERATE paragraph can be used to create values of a new variable according to user specified conditions. A set of data may be generated in one of two ways. One technique is to specify completely every value of the set. Data may also be created according to a pattern that increases from a specified initial value according to a user specified step size, or rate. The two methods (VALUES and PATTERN) are mutually exclusive and they may not both be specified in the same paragraph. The generated values are then stored into a variable in a user specified order.

## Syntax for the GENERATE Paragraph

```
GENERATE  VARIABLE IS v.          @
          NROW IS i.              @
          NCOL IS i.              @
          ORDER IS w.             @
          VALUES ARE r1, r2, --- .
               or
          PATTERN IS w1(r1,r2), w2(r1,r2).
```

Required sentences: **VARIABLE**, **and** either **VALUES or PATTERN**

## Sentences Used in the GENERATE Paragraph

**VARIABLE sentence**

The VARIABLE sentence is used to specify the name of the vector or matrix to store values that are generated.

**NROW sentence**

The NROW sentence is used to specify the number of rows of values for the variable to be generated. The default is 1.

**NCOL sentence**

The NCOL sentence is used to specify the number of columns of values for the variable to be generated. The default is 1.

**ORDER sentence**

The ORDER sentence is used to specify the order for placing the generated values in a matrix. Keywords available are:

| | | |
|---|---|---|
| COLUMNWISE | -- | values are stored in column 1 first, then column 2, etc. (This is the default) |
| ROWWISE | -- | values are stored in row 1 first, then row 2, etc. |

DIAGONAL -- values are stored in the diagonal elements of the matrix, all off-diagonal elements are set to zero. The matrix must be square. That is, the value specified in the NROW sentence must be the same as that specified in the NCOL sentence.

SYMMETRIC -- values are stored in the lower triangular part of the matrix, row by row. Values in the upper triangular part are set equal to the corresponding lower triangular elements.

SKEWSYMMETRIC -- values are stored in the lower triangular part of the matrix row by row. Values in the upper triangular part are set equal to the negative of the corresponding lower triangular elements.

**VALUES sentence**

The VALUES sentence is used to specify the values to be placed in the variable. The number of values to be specified is NROW*NCOL if the ORDER is COLUMNWISE or ROWWISE, NROW*(NROW+1)/2 if SYMMETRIC or SKEWSYMMETRIC, and NROW if DIAGONAL. Note that the VALUES and the PATTERN (defined below) sentences are mutually exclusive, only one of them can appear in the paragraph.

**PATTERN sentence**

The PATTERN sentence is used to specify the pattern to be used to generate values. The keywords are STEP or RATE. The STEP option will generate an arithmetic sequence with initial value r1 and increment r2 (i.e., the sequence r1, r1+r2, r1+2*r2, ...), and the RATE option will generate a geometric sequence with initial value r1 and rate r2 (i.e., r1, r1*r2, r1*r22, ...). If both STEP and RATE are specified, the result will be the sum of the two sequences. The PATTERN sentence must be specified if the VALUES sentence is not specified.

## RECODE Paragraph

The RECODE paragraph is used to modify or recode the values of an existing variable. Results may be stored in a new or existing variable. The entries of an existing "old variable" falling in a specified range of values are changed to another specified value. Values in a variable may also be modified using analytic statements (see Appendix A).

### Syntax for the RECODE Paragraph

```
RECODE      OLD IS v.                                  @
            NEW IS v.                                  @
            PRECISION IS w.                            @
            VALUES ARE (r1,r2,r3),(r1,r2,r3), ---.
```

Required sentences: **OLD and VALUES**

## Sentences Used in the RECODE Paragraph

**OLD sentence**

The OLD sentence is used to specify the name of the variable to be recoded.

**NEW sentence**

The NEW sentence is used to specify the name of the variable in which the edited results are stored. If a new name is not specified, the recoded variable will be stored under the old name.

**VALUES sentence**

The VALUES sentence is used to specify sets of values consisting of a range (r1,r2) and a recoding value, r3. All data values falling into the range are changed to the recoding value. The reserved word MISSING (that may be abbreviated as MIS) is used to denote the missing value code and can be used in the triplet. To recode missing data to a specific value, the triplet should be specified as (MISSING, MISSING, r) where r is an integer or real number.

**PRECISION sentence**

The PRECISION sentence is used to specify the precision of the storage of the recoded variable. The default is the precision of the old variable.

## OMIT and SELECT Paragraphs

The OMIT and SELECT paragraphs are used to delete or retain elements of a variable according to range (span) or value criteria. Elements are deleted or selected if the element's index falls within the specified range(s). The value criterion is used in a similar manner except that the value is used instead of the range of the values. In addition, the OMIT or SELECT paragraph may operate on more than one variable at a time. If more than one variable is specified in the paragraph, the deletion or selection criteria is only applied to the elements of the first variable while the elements in the corresponding position of all other specified variables are deleted or selected according to the action taken on the entry in the first variable. When more than one variable is specified, the variables need not have the same number of entries but the first variable must have the largest number of rows. Furthermore, values of a variable can be selected even if they have been coded with a missing value code.

## Syntax for the OMIT and SELECT Paragraphs

```
OMIT  OLD ARE  v1, v2, --- .                @
      NEW  ARE  v1, v2, ---.                @
      SPANS  ARE  (i1,i2),(i3,i4), ---.     @
      VALUES  ARE  (r1,r2),(r3,r4), ---.    @
      MISSING.

Required sentence:  OLD
```

```
SELECT      OLD  ARE  v1, v2, --- .                @
            NEW  ARE  v1, v2, --- .                @
            SPANS  ARE  (i1,i2), (i3,i4), --- .    @
            VALUES  ARE  (r1,r2), (r3,r4), --- .   @
            MISSING.

Required sentence:  OLD
```

## Sentences Used in the OMIT and SELECT Paragraphs

**OLD sentence**

The OLD sentence is used to specify the name(s) of the variable(s) for which values will be deleted or selected.

**NEW sentence**

The NEW sentence is used to specify the name(s) of the variable(s) where the results of the deletion or selection operation are stored. The number of variables specified in this sentence must be the same as that in the OLD sentence. The results will be stored in the original variables if the NEW sentence is omitted.

**SPANS sentence**

The SPANS sentence is used to specify the span(s) to be used in the deletion or selection process. Indices falling in i1 to i2, i3 to i4, etc. will be omitted or selected.

**VALUES sentence**

The VALUES sentence is used to specify the range of values to be deleted or selected, values r1 to r2, r3 to r4, etc. This criterion applies to the values of the first variable only, other variables are deleted or selected according to the action taken on the corresponding entry of the first variable.

**MISSING sentence**

The MISSING sentence is used to specify the deletion or the selection of the cases which have been coded with a missing data code. This criterion applies to the first variable. Other variables are deleted or selected according to the action taken on the corresponding entry of the first variable.

## JOIN Paragraph

The JOIN paragraph is used to create a variable by appending the data of one or more variables to the end of a designated variable in the SCA workspace. If all presently defined variables are vectors, the resultant vector is created by appending the entries of the second vector to the last entry of the first, the third to the end of this, and so on. The number of entries in this resultant vector is equal to the sum of the entries of all the present vectors. This procedure is the same if all presently defined variables are matrices. However, each matrix must contain the same number of columns. Vectors may not be joined to matrices. The precision of the resultant variable may also be specified.

## Syntax for the JOIN Paragraph

```
JOIN  OLD ARE v1, v2, --- . @
      NEW IS v.     @
      PRECISION IS w.

Required sentence: OLD
```

## Sentences Used in the JOIN Paragraph

**OLD sentence**

The OLD sentence is used to specify the names of the variables to be joined.

**NEW sentence**

The NEW sentence is used to specify the name of the variable in which the results of the join operation are stored. If the NEW-VARIABLE sentence is not specified, then the results of the join operation will be stored under the name of the first variable listed in the OLD sentence.

**PRECISION sentence**

The PRECISION sentence is used to specify the precision of the storage for the joined results. The keyword, w, may be either SINGLE or DOUBLE. The default is the precision of that of the first variable listed in the OLD sentence.

## AUGMENT Paragraph

The AUGMENT paragraph is used to create a variable by appending the data of one or more variables side by side. All variables (either a vector or a matrix) must have the same number of rows. The number of columns in the resultant matrix is equal to the sum of the columns of all the present variables. The precision of the resultant variable may also be specified.

## Syntax for the AUGMENT Paragraph

```
AUGMENT   OLD ARE v1, v2, ---.      @
          NEW IS v.                 @
          PRECISION IS w.

Required sentence: OLD-VARIABLES
```

## Sentences Used in the AUGMENT Paragraph

**OLD sentence**

The OLD sentence is used to specify the names of the variables to be augmented.

**NEW sentence**

The NEW sentence is used to specify the name of the variable in which the results of the augment operation are stored. If the NEW sentence is not specified, then the results of the join operation will be stored under the name of the first variable listed in the OLD sentence.

**PRECISION sentence**

The PRECISION sentence is used to specify the precision of the storage for the augmented matrix. The keyword, w, may be either SINGLE or DOUBLE. The default is the precision of that of the first variable listed in the OLD sentence.

## PATCH Paragraph

The PATCH paragraph is used to recode missing data of a time series by replacing missing values with one of the following:

(1) the average of the two observations that are i indices adjacent to it,

(2) the mean of all observations or those non-missing observations i indices apart from the missing value, or

(3) a specified value.

In addition, a binary indicator variable can be created to provide a reference variable highlighting those time indices whose values were patched.

## Syntax of the PATCH Paragraph

```
PATCH       OLD IS v.                    @
            NEW IS v.                    @
            METHOD IS w(i).              @
            SPAN IS i1, i2.              @
            INDICATOR IS v.

Required sentence: OLD
```

## Sentences Used in the PATCH Paragraph

**OLD sentence**

The OLD sentence is used to specify the name of the variable containing missing data.

**NEW sentence**

The NEW sentence is used to specify the name of the variable to store the patched series. The default is the name specified in the OLD series.

**METHOD sentence**

The METHOD sentence is used to specify the method used to recode missing data in the OLD variable. Keywords and associated arguments that may be used to specify the method are:

(1) ADJACENT(i): all missing data are recoded to the average of the values of the OLD series with indices (t-i) and (t+i), where t is the index of the missing value.

(2) MEAN(i): all missing data are recoded to the periodic average of the non-missing values of the OLD series. The argument i is used to specify the periodicity of the series. If i=1 then the overall average of all non-missing data will be used to recode the missing observations.

(3) VALUE(r): all missing data are recoded to the value r.

The methods are all mutually exclusive within the execution of a single paragraph. The default is ADJACENT(1).

**SPAN sentence**

The SPAN sentence is used to specify the span of time indices, i1 to i2, in which a patch of missing data will be made. The default span is the whole series.

**INDICATOR sentence**

The INDICATOR sentence is used to specify a name (label) for an indicator variable associated with the patching. The indicator variable contains 1 for missing data that are replaced, and 0 otherwise. The length of the indicator variable is always the same as the old series regardless of the time periods specified in the SPAN sentence. This convention allows employment of the same indicator variable for multiple patches of a series using different methods.

## LAG Paragraph

The LAG paragraph is used to apply the lag (backshift) operator, B, to a variable to create a new lagged variable. For the variable X, the lag operation Y = B(X) is defined as $Y_t = X_{t-1}$ provided it exists (otherwise a missing value code is provided). This definition is for a lag one backshift. Various other lag orders may be specified (e.g., lag k, where $Y_t = X_{t-k}$ for various values of k), hence creating more than one new series.

Lagged values are stored in the following manner. If the variable YDATA stores the k-th order lagged values of the variable XDATA, then

YDATA(t) = the missing value code, j = 1,2,. . .,k
= XDATA(t-k), t = k+1,. . .,k+n

where n is the index of the last observation (value) of XDATA. As a result, YDATA has (n+k) observations, the first k of which containing the missing value code, while XDATA has n observations.

## Syntax for the LAG Paragraph

**LAG** OLD IS v. @
NEW ARE v1, v2, --- .@
LAGS ARE i1, i2, --- .

Required sentence: **OLD**

## Sentences Used in the LAG Paragraph

**OLD sentence**

The OLD sentence is used to specify the name of the series to be lagged.

**NEW sentence**

The NEW sentence is used to specify names of the new series. Results will be stored in the OLD series if the NEW sentence is omitted.

**LAGS sentence**

The LAGS sentence is used to specify the lags to be made on the old series. For example, if there are 3 specified lags, three new series will be generated. The default is 1, creating $Y_t = X_{t-1}$.

## DIFFERENCE Paragraph

The DIFFERENCE paragraph is used to apply the operator $(1-B^j)$ to a variable or a set of variables to create one or more variables. For a variable X, the operation $Y = (1-B^j)X$ is defined as $Y_t = X_t - X_{t-j}$ provided t > j (otherwise a missing value is specified). This definition is given for one differencing order (DFORDER) in the backshift operator B. More than one differencing order may be specified. If differencing orders, i1, i2, . . ., im, are specified, then the operator $(1-B^{i1})(1-B^{i2})\ldots\ldots(1-B^{im})$ will be applied to all designated variables. In such a case the missing value code is stored as the first (i1+i2+×××+im) values of the resulting variable. The missing values may be deleted and the resulting variable compressed to a series containing n-i1+i2+×××+im) values, where n is the number of observations of the original series, if the COMPRESS sentence is specified.

## Syntax for the DIFFERENCE Paragraph

| | | |
|---|---|---|
| **DIFFERENCE** | OLD ARE v1, v2, --- . | @ |
| | NEW ARE v1, v2, --- . | @ |
| | DFORDERS ARE i1, i2, --- . | @ |
| | COMPRESS. /NO COMPRESS. | |

Required sentence: **OLD**

## Sentences Used in the DIFFERENCE Paragraph

**OLD sentence**

The OLD sentence is used to specify the name(s) of the series to be differenced.

**NEW sentence**

The NEW sentence is used to specify the variable name(s) where the differenced series are stored. The default is that the data will be stored in the names specified in the OLD sentence.

**DFORDERS sentence**

The DFORDERS sentence is used to specify the orders in the product of differencing operators to be made on the OLD series. Default is 1, the single operator (1-B). If i1, i2, . . . are specified the operator $(1-B^{i1})(1-B^{i2})$ . . . is applied to the old series.

**COMPRESS sentence**

The COMPRESS sentence is used to indicate whether the missing values caused by differencing will be deleted. When COMPRESS is specified, the resulting NEW variable will have fewer observations than its corresponding OLD variable. The first value of the NEW variable will be the first value for which the differencing operator is valid. Default is NO COMPRESS, i.e., the missing value code will be assigned to all undefined values and the NEW variable will have as many observations as its corresponding OLD variable.

# APPENDIX C

# SCA MACRO PROCEDURES

The SCA System provides us with the capability to create and maintain computations, analyses or procedures specific to our needs. For example, we may find it useful to perform a special sequence of SCA operations with different data during an SCA session. It would simplify our work if such sequences can be written only once and then could be freely referred to subsequently. Many programming languages provide subprograms to help in this situation. SCA offers macro procedures to obtain such flexibility.

The use of an SCA macro procedure enables us to store any set of SCA statements on a file which may be referenced at any point of an SCA session. This enables us to extend the capabilities of the SCA System.

## C.1 SCA Macro Files and Macro Procedures

SCA macro procedures are maintained in files. These files are referred to as SCA macro files. Procedures contained on a macro file may be created by any text editor.

An SCA macro procedure consists of a sequence of SCA statements, including both analytic and English-like statements. A macro procedure is handled as a "subprogram" within an SCA session.

To illustrate SCA macro files and SCA macro procedures, Tables 1 and 2 list the contents of two SCA macro files. These files will be used throughout this Appendix to illustrate SCA macro procedures. The records of Table 1 and Table 2 comprise the files APPENDX.DATA and MACRO.DATA, respectively. The names of the files are for illustration only and may be changed to names appropriate to a local computer.

**Table 1 Contents of the file APPENDX.DATA**

```
==ALL MACRO
  CALL APPENDXA
  CALL APPENDXB
  RETURN
==APPENDXA
--A MACRO PROCEDURE ILLUSTRATING THE MATRIX EXAMPLES
--OF APPENDIX A
INPUT ADATA, BDATA, EDATA.  NCOL ARE 2, 3, 3.
1   1   1   3   0   3  -1   0
3   1   2   1   0  -1   2  -1
0   1   0   1  -1   0  -1   3
END OF DATA
C1DATA = BDATA # ADATA
C2DATA = T(ADATA) # BDATA
DETB = DET(BDATA)
PRINT  C1DATA, C2DATA, DETB
BINVERSE = INV(BDATA)
ADJOINTB = DETB*BINVERSE
PRINT BINVERSE, ADJOINTB
EIGEN EDATA
RETURN
==APPENDXB
--A MACRO PROCEDURE OF THE SCA STATEMENTS IN SECTION B.1.1
--AND B.1.2 OF APPENDXB
GENERATE VECTOR1. NROW ARE 10. VALUES ARE 0 FOR 5, 1 FOR 5.
GENERATE VECTOR2. NROW ARE 10. VALUES ARE 0 FOR 5, 1 FOR 2, 0 FOR 3.
GENERATE VECTOR3. NROW ARE 10. PATTERN IS STEP(1.0, 0.5).
GENERATE VECTOR4. NROW ARE 10. PATTERN IS RATE(1.0, 2.0).
PRINT VECTOR1, VECTOR2, VECTOR3, VECTOR4
RETURN
//
```

## Table 2 Contents of the file MACRO.DATA

```
      ==SCORES
      C                                @
      C   AVERAGE ENGLISH SCORES       @
      C
          INPUT VARIABLE IS ENGLISH.
          82  14  25  67  48  76  23  46  96
          69  66  62  70  88  61  72
          END OF DATA
      C                               @
      C   AVERAGE PHYSICS SCORES      @
      C
          INPUT VARIABLE IS PHYSICS.
          86  72  34  92  68  74  69  35
          75  24  33
          END OF DATA
          RETURN
      ==EXPLORE
      C                                                          @
      C   AS A MEANS TO GET A FEEL FOR A DATA SET, DESCRIPTIVE  @
      C   STATISTICS, A CONFIDENCE INTERVAL AND PLOT OF DATA    @
      C   OVER TIME WILL BE INVOKED.                            @
      C   DATA ARE ASSUMED TO BE STORED IN THE SCA WORKSPACE    @
      C   IN A VARIABLE NAMED X.                                @
      C
          DPLOT  X
          CINTERVAL  X
          TSPLOT  X
          RETURN
==LINREG

      PARAMETER SYMBOLIC-VARIABLES ARE NINDEP, FILE(12).
C                    @
C     READ IN DATA     @
C
      INPUT VARIABLES ARE Y,X.    FILE IS &FILE.  @
            NCOLS ARE 1, &NINDEP.
C                                                                  @
C     COMPUTE REGRESSION COEFFICIENTS, PREDICTED VALUES, RESIDUALS, ETC. @
C
      BETA = INV(T(X)#X)#T(X)#Y  -- COMPUTE REGRESSION COEFFICIENTS
      YHAT = X#BETA              -- COMPUTE PREDICTED VALUES
      RESI = Y-YHAT              -- COMPUTE RESIDUALS
      N    = NROW(X)
      P    = NCOL(X)
      NP   = N-P
      P1   = P-1
      MEAN = SUM(Y)/N
      SST  = SUM((Y-MEAN)**2)
      SSE  = SUM(RESI**2)
      SSB  = SST-SSE
      MSE  = SSE/NP
      MSB  = SSB/P1
      F    = MSB/MSE
C                                  @
C     PRINT REGRESSION COEFFICIENTS   @
C
      DO 100 I=1,P
      I1=I-1
      IF(I1 LE 9) THEN NEXT ELSE GO FORWARD 80
      DISPLAY TEXT IS T5,'BETA',I1('F1.0'),'  = ',BETA('F12.4',I).
      GO FORWARD 100
   80 DISPLAY TEXT IS T5,'BETA',I1('F2.0'), ' = ',BETA('F12.4'',I).
  100 CONTINUE
```

**Table 2 Contents of the file MACRO.DATA (continued)**

```
C                                          @
C     PRINT THE ANALYSIS OF VARIANCE TABLE  @
C
      DISPLAY TEXT IS ///T5,'ANALYSIS OF VARIANCE TABLE'//         @
          T5,'  SOURCE     D.F.   SUM OF SQUARES   MEAN SQUARES   F'/.
      DISPLAY TEXT IS T5,'REGRESSION',P1('F6.0',1),SSB('C17.4',1), @
          MSB('C15.4',1),F('C10.2',1).
      DISPLAY TEXT IS T5,'   ERROR  ',NP('F6.0',1),SSE('C17.4',1), @
          MSE('C15.4',1).
      DISPLAY TEXT IS T5,'   TOTAL  ',N ('F6.0',1),SST('C17.4',1)
      RETURN
//
```

## C.2 Structure of an SCA Macro File

Both files APPENDX.DATA and MACRO.DATA have similar structure. A set of SCA commands, or data, are preceded by a record with double equal signs (i.e., '= =') in columns 1 and 2; and are ended with the statement RETURN. The final entry of each file is '//'.

The alphanumeric characters following '= =' provide the name of the macro procedure. For example, the file APPENDX.DATA consists of the macro procedures named ALLMACRO, APPENDXA and APPENDXB; while MACRO.DATA contains the macro procedures SCORES, EXPLORE and LINREG. The name of a macro procedure may contain from one to eight alphanumeric characters, with a letter as the mandatory first character. If more than eight characters are used as a macro procedure name, only the first eight characters are interpreted.

Any line that begins with a double dash ('--') or the letter C immediately followed by a space ('C ') will be interpreted by the SCA System as a line of comments. Lines beginning with '--' are not printed as they are interpreted, but those beginning with 'C ' will be printed as they are interpreted during an SCA session.

## C.3 Invoking a Macro Procedure

If we enter the command

-->CALL APPENDXB. FILE IS 'APPENDX.DATA'.

then the following set of SCA commands will be interpreted and executed

```
GENERATE VECTOR1. NROW ARE 10. VALUES ARE 0 FOR 5, 1 FOR 5.
GENERATE VECTOR2. NROW ARE 10. VALUES ARE 0 FOR 5, 1 FOR 2, 0 FOR 3.
GENERATE VECTOR3. NROW ARE 10. PATTERN IS STEP(1.0, 0.5).
GENERATE VECTOR4. NROW ARE 10. PATTERN IS RATE(1.0, 2.0).
PRINT VECTOR1, VECTOR2, VECTOR3, VECTOR4
```

These commands will duplicate selected capabilities illustrated in Appendix B. Similarly, if we enter

```
-->CALL   APPENDXA.   FILE IS  'APPENDX.DATA'
```

then selected capabilities illustrated in Appendix A will be computed and results displayed.

The macro procedure SCORES of MACRO.DATA will transmit two variables, stored as ENGLISH and PHYSICS, to the SCA workspace. The procedure named EXPLORE can be used for computing a set of descriptive statistics, a confidence interval and a time series plot of a variable. For example, suppose we have three variables, SERIESA, SERIESB, and SERIESC, in the SCA workspace. We can repeatedly perform these operations by entering the sequence of commands

```
-->X = SERIESA
-->CALL EXPLORE.   FILE IS  'MACRO.DATA'.
-->X = SERIESB
-->CALL EXPLORE
-->X = SERIESC
-->CALL EXPLORE
```

We may note that after the first CALL to the EXPLORE macro procedure the FILE sentence is omitted. Unless it is instructed otherwise, the SCA System assumes a macro procedure being called resides in the last referenced macro file. This default is implicit within the macro procedure ALLMACRO of the APPENDX.DATA file. If we enter

```
CALL   ALLMACRO.   FILE IS  'APPENDX.DATA'.
```

we see that calls to the remaining macro procedures of the file are invoked, hence all macro procedures of the file are executed. Care must be taken if one or more macro procedures is nested within another. That is, an error can occur if one macro procedure calls another. Appropriate allocation and de-allocation of files is required. Please refer to Section 1 of Appendix D for further information.

## C.4 Symbolic Variables in a Macro Procedure

### Symbolic Variables

The term “symbolic variables” refers to any name used in a macro procedure to label a variable or entry that can be given a new value or connotation when a macro procedure is invoked. Symbolic variables add flexibility to macro procedures by labeling actual arguments that may change when a procedure is executed. For example, it is desirable to be able to pass a different series name (or variable name) to the EXPLORE procedure in MACRO.DATA rather than requiring a variable to have X as its name for all uses of the procedure. To facilitate this convenience, the label X may be replaced by the expression &SERIES, in the

procedure. In this manner SERIES is then recognized by the system as a symbolic variable as explained below.

Within the body of a macro procedure a symbolic variable is denoted by preceding a string of alphanumeric characters by an ampersand (&). The first character of the string must be a letter and the last character may be a compound symbol (#). The compound symbol is used as a deliminator if the symbol variable is immediately followed by a number or letter. The name of the symbolic variable is the character string excluding the compound symbol. The compound symbol can be omitted if the symbolic variable name is immediately followed by a special character such as blank, ' . ' or ' , '. If the alphanumeric string denoting the symbolic variable has more than eight characters, only the first eight are interpreted. The special character '&' is used to distinguish symbolic variables from other variables used in a macro procedure. The actual values used for the symbolic variables are supplied when the macro procedure is invoked (by the CALL paragraph, see syntax at the end of this Appendix), or may be those values supplied as default values within the procedure itself. In the SCA interactive mode, if a symbolic variable does not have a default value and is not supplied in the CALL paragraph, the SCA System will issue a prompt for a value. The response to the prompt must be enclosed in a pair of parentheses. A fatal error will occur if such a situation happens in the batch environment.

## Symbolic Substitution

The SCA System scans each line in a macro procedure and replaces symbolic variables with their actual values in an action called symbolic substitution. An actual argument for a symbolic variable is always stored in its exact character form. For example, if a symbolic variable has a value 2.3, it is stored as a string of three characters '2.3', rather than a real number. Hence symbolic substitution will not lose any precision. The rule governing symbolic substitution is simple: the SCA System scans a line in the macro procedure from right to the left and substitutes the first symbolic variable encountered by its associated value (in character form). This scanning is repeated until all symbolic variables are substituted and resolved. This rule allows the user to concatenate symbolic variables to modify existing variable names, or to use multiple ampersands. For example, if &A has the symbolic argument JOHN, and &JOHN has the symbolic argument BOY, then &&A will have the value BOY after the completion of symbolic substitution. The symbolic variables may appear anywhere in a statement in an SCA macro procedure although they usually appear in analytic expressions or argument lists of assignment sentences.

## C.5 A Regression Macro Procedure

To illustrate both the use of symbolic variables and the ability to write our own procedures, we consider the macro procedure LINREG of the MACRO.DATA file (see Table 2). LINREG performs a regression analysis using analytic expressions (see Appendix A). This procedure may be useful in teaching regression analysis, but a more computationally efficient means is available through the SCA REGRESS paragraph (see Chapter 6).

The macro procedure transmits data for the dependent and independent variables from a file. The symbolic argument FILE is used to designate the logical unit number for the file containing the data. If a unit number is not specified in the CALL paragraph, the default unit 12 is used. This default value is specified within the LINREG macro in the PARAMETERS paragraph (its complete syntax is provided at the end of this Appendix).

Within the file FILE the first column of data is transmitted to the dependent variable labeled Y and the remaining p columns contain the independent variables, stored in the matrix X. The value p is represented in the macro by the symbolic argument NINDEP. This argument has no default value, hence we must specify it in our CALL of LINREG.

The data listed in Table 3 is assumed to be on a file that has been associated with the logical unit 12. This assignment may have been accomplished before we invoked the SCA System or through the ASSIGN paragraph (see Appendix D).

**Table 3 Data used in the LINREG example**

| | | | | |
|---|---|---|---|---|
| 101 | 1 | 1 | 1 | 1 |
| 106 | 1 | 1 | 1 | 1 |
| 87 | 1 | 1 | 1 | 1 |
| 131 | 1 | 1 | 1 | 1 |
| 265 | 1 | 1 | 2 | 2 |
| 272 | 1 | 1 | 2 | 2 |
| 279 | 1 | 1 | 2 | 2 |
| 302 | 1 | 1 | 2 | 2 |
| 106 | 1 | 2 | 1 | 2 |
| 89 | 1 | 2 | 1 | 2 |
| 128 | 1 | 2 | 1 | 2 |
| 103 | 1 | 2 | 1 | 2 |
| 291 | 1 | 2 | 2 | 4 |
| 306 | 1 | 2 | 2 | 4 |
| 334 | 1 | 2 | 2 | 4 |
| 272 | 1 | 2 | 2 | 4 |

To invoke the LINREG procedure on this data set we can enter

```
-->CALL   LINREG.   FILE IS  'MACRO.DATA'.   @
        SYMBOLIC IS NINDEP(4).
```

We will obtain output similar to that given below.

```
REGRESSION COEFFICIENTS:

   BETA0  =     -46.2500
   BETA1  =     -20.7500
   BETA2  =     152.2500
   BETA3  =      21.0000

   ANALYSIS OF VARIANCE TABLE

   SOURCE      D.F.   SUM OF SQUARES    MEAN SQUARES        F

   REGRESSION     3     135,959.5000     45,319.8330     117.73
      ERROR      12       4,619.5000        384.9583
      TOTAL      16     140,579.0000
```

## C.6 Global and Local Variables

A variable with '@' as the first character of its name is treated as a local variable within a macro procedure. Others are regarded as global variables. The difference between a local and global variable is that local variables are deleted from the workspace upon completion of a macro procedure, unless otherwise specified. A local variable may be retained in the workspace by using the RETAIN sentence in the RETURN paragraph (see the Syntax section at the end of this Appendix). Global variables may be used anywhere in a session, including in subsequent macro procedures.

# SUMMARY OF THE SCA PARAGRAPHS IN APPENDIX C

This section provides a summary of those SCA paragraphs employed in this chapter. Each SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'.

The paragraphs to be explained in this summary are CALL, PARAMETERS, and RETURN.

Legend (see Chapter 2 for further explanation)

| | | |
|---|---|---|
| v | : | variable name |
| v(a) | : | variable name (with argument) |
| i | : | integer |
| 'c' | : | character data (must be enclosed within single apostrophes) |

## CALL Paragraph

The CALL paragraph is used to invoke an SCA macro procedure. It is also used to specify the actual arguments for the symbolic variables in the macro procedure and repetitions of the execution of the procedure.

## Syntax of the CALL Paragraph

```
CALL PROCEDURE IS procedure-name.      @
     FILE IS 'c' (or i).       @
     SYMBOLIC-VALUES ARE v1(a), v2(a), --- . @
     REPEAT IS i.

Required sentence: PROCEDURE
```

## Sentences used in the CALL paragraph

**PROCEDURE sentence**

The PROCEDURE sentence is used to specify the name of the macro procedure to be executed.

**FILE sentence**

The FILE sentence is used to specify the name of the macro procedure file containing the called macro procedure. A logical unit number may be specified instead. The default unit is 8. More than one macro procedure file may be allocated for an SCA session.

**SYMBOLIC-VALUES sentence**

The SYMBOLIC-VALUES sentence is used to specify the actual values or arguments of of the symbolic variables used in the procedure. The value of a symbolic variable need not be specified if the default value is desirable. If a symbolic variable does not have a default value and is not specified in this sentence, execution of the macro procedure is aborted in batch mode or a prompt message is issued in the interactive mode requesting an appropriate value when the PARAMETER paragraph is executed. The syntax for the arguments in this sentence is the same as that in the SYMBOLIC-VARIABLE sentence of the PARAMETER paragraph.

**REPEAT sentence**

The REPEAT sentence is used to specify the number of times the macro procedure should be executed. This sentence is useful when the macro procedure is used for simulation. The default value is 1.

## PARAMETERS Paragraph

The PARAMETERS paragraph is used to specify the symbolic variables (and their possible default values) of an SCA macro procedure. This paragraph is not required in a macro procedure if the procedure does not have symbolic variables. The PARAMETERS paragraph must be executed before any symbolic variable is used. Usually, it is placed at the beginning of a macro procedure. Note only one PARAMETERS paragraph may be specified in a macro procedure.

## Syntax of the PARAMETERS paragraph

**PARAMETERS** SYMBOLIC-VARIABLES ARE v1(a), v2(a), --- .

Required sentence: **SYMBOLIC**

## Sentence used in the PARAMETERS paragraph

### SYMBOLIC-VARIABLE sentence

The SYMBOLIC-VARIABLE sentence is used to specify those variables that will be used as symbolic variables in a macro procedure. The arguments, v1(a), v2(a), ---, have the following syntax

Symbolic-variable-name(default-symbolic-value)

Specification of a default symbolic value or argument is optional. If a symbolic variable is given no default argument, its argument must be specified in the CALL paragraph. Otherwise a fatal error results in batch mode, or a prompt is issued by the system in the interactive mode. All characters, inside the parentheses, including the leading and trailing blanks, are interpreted as part of the argument. Therefore both NAME(A) and NAME(A ) are acceptable to define the default value of the symbolic variable NAME and are considered to be different. The argument for the former specification has one character, 'A', the latter has two characters, i.e., 'A' and a trailing blank. Due to such differences, the response to a system prompt for the value of a symbolic variable of the paragraph must be enclosed in a pair of parentheses.

**Note**: The names specified in this sentence are the labels of those variables that are symbolic variables in the remainder of the macro procedure. Unlike the designation of symbolic variables in the remainder of the macro, these names must not be preceded by an ampersand (&).

## RETURN Paragraph

The RETURN paragraph is used to signify the end of an execution flow for a set of instructions written as an SCA macro procedure. The paragraph also is used to specify actions to be taken with respect to variables created during the macro procedure.

## Syntax of the RETURN paragraph

**RETURN** RETAIN v1, v2, ---. @
COMPRESSION./NO COMPRESSION.

Required sentences: none

## Sentences used in the RETURN paragraph

**RETAIN sentence**

The RETAIN sentence is used to specify the name(s) of those local variables (i.e., ones that are for temporary use in the macro procedure) that should now be retained (i.e., not deleted) in the workspace after the execution of the macro procedure. Normally, all local variables are deleted from the workspace. All local variables may be retained by specifying

RETAIN ALL@.

**COMPRESSION sentence**

The COMPRESSION sentence is used to specify the compression of the SCA workspace after the execution of the macro procedure. Although all local variables are deleted after an SCA macro procedure is completed, the SCA System does not automatically compress the user workspace. That is, the deleted variables still occupy space in the memory. The keyword COMPRESSION must be specified in the RETURN paragraph if the workspace is to be compressed.

# APPENDIX D

# UTILITY RELATED INFORMATION

The SCA System provides a number of capabilities to manage files, internal workspace (memory), and other utility related tasks effectively within an SCA session. An overview of some of these features are presented in this Appendix. More information may be found in *The SCA Statistical System: Reference Manual for Fundamental Capabilities*. Information for using the SCA System on specific computers is provided with the diskettes or tapes containing the SCA System. In the latter case, this information may be retained by personnel in a computing center and may not be readily available to an SCA user. In such a case, SCA will furnish necessary document(s) upon request.

All information (data) used during an SCA session resides in the main memory of the computer. The SCA System refers to this memory as the workspace of the SCA session. In addition to user defined information, certain control blocks for the SCA System, and temporary work arrays required by some of the operations are also placed in the workspace as variables. The SCA System has a built-in dynamic storage allocator to manage the space available for variables during an SCA session. Usually we do not need to be concerned about the management of external files or of the workspace; but occasionally certain actions may be necessary in order to use the SCA System or the workspace efficiently. We will first examine aspects of file management, then discuss how we can manage the workspace and the presentation of material in it.

## D.1 File Allocation and De-allocation

A file may need to be designated when transmitting data to or from the SCA workspace, when executing a macro procedure (see Appendix C), or managing the SCA workspace (see Section D.2). The FILE sentence is used for this purpose. The syntax of this sentence is

**FILE IS 'file-name'.**

where 'file-name' is a valid file name. Please note that the file name specified must be enclosed within a pair of single quotes. File names with directory path are acceptable.

In some situations, it is necessary to associate (assign) a unit number with a file name. In such cases, the file unit number is an integer and should not be enclosed within single quotes. Some reasons to use unit numbers are provided below. The SCA ASSIGN paragraph can be used for this purpose.

When data are transmitted to or from the SCA workspace, the SCA System dynamically assigns (associates) unit number 7 with the file name specified. Since internal assignment of unit numbers is made in these paragraphs, we need not specify file unit number when using these paragraphs.

The FREE paragraph releases a file from an SCA session and makes the unit number available to other files. However, ASSIGNing the same unit twice implicitly FREEs the first file before ASSIGNing the second one. Thus, it is not necessary to issue a FREE paragraph before re-using a unit number, though it certainly does not hurt.

The ASSIGN paragraph is seldom needed except when (1) recalling the contents of a workspace file with a name different from the default file (or default unit) employed, or (2) a macro procedure calls another macro procedure of a different file. An example is provided to illustrate each situation.

EXAMPLE 1:

As an example of the ASSIGN and WORKSPACE (see Section D.3) paragraphs, the following SCA paragraphs may be used to allocate a file and save the SCA workspace to the file PROJECT1.WRK.

```
ASSIGN      FILE IS 9.                                  @
            EXTERNAL IS  'PROJECT1.WRK'.  NEW.          @
            ATTRIBUTE FILEFORMAT(BINARY),               @
            ACCESS(WRITE).

WORKSPACE       MEMORY IS SAVED(9).
```

The specification ACCESS(WRITE) is not necessary since a NEW file is always writable. However, such specification is necessary if the file to be used is an existing file.

To recall the workspace saved previously, we may enter

```
ASSIGN      FILE IS 9.  EXTERNAL IS  'PROJECT1.WRK'.    @
            ATTRIBUTE FILEFORMAT(BINARY).

WORKSPACE       MEMORY IS RECALLED(9).
```

EXAMPLE 2:

The following example demonstrates the use of the ASSIGN paragraph within an SCA macro procedure (see Appendix C) that has an imbedded CALL to another file. In this example, we assume there are two macro procedure files. One is named MYDATA.DAT, a file consisting of procedures that will transmit data sets to the SCA System. One of the macro procedures of this file is assumed to have the name DATA1.

Suppose there is a second macro procedure file, say MYPROC.DAT, consisting of a number of macro procedures useful for data analysis. In this file, we assume there is a macro procedure named EXAMPLE1 that reads the data contained in the macro procedure DATA1. The portion of this file related to EXAMPLE1 is given below.

```
.
.
.
==EXAMPLE1
  ASSIGN  FILE IS 20.   EXTERNAL IS 'MYDATA.DAT'.
  CALL    DATA1.  FILE IS 20.
      .
      .
      .
  RETURN
  END
==EXAMPLE2
      .
      .
      .
```

The procedure above does the following:

(1) MYDATA.DAT is associated with the file unit 20.

(2) Data are transmitted through the call of the macro procedure DATA1 in the file MYDATA.DAT

(3) Other analyses may follow after the data are transmitted

The above steps are invoked by entering the statement

CALL EXAMPLE1. FILE IS 'MYPROC.DAT'

(See Appendix C regarding the use of the CALL paragraph.) If MYDATA.DAT was not provided with a separate file unit number, then the macro CALL of DATA1 would cause an error. First the macro file MYPROC.DAT would be freed and replaced by MYDATA.DAT as the macro file in use. The SCA System would then be unable to return to EXAMPLE1 as it will have lost track of the file containing it.

## D.2 Control of the SCA Environment: the PROFILE Paragraph

We can "control" our SCA environment through the use of the PROFILE paragraph. The PROFILE paragraph can be used to alter the prompting and display levels of an SCA session, direct output to an external file, or adjust the width of output displayed or assumed for data transmitted to the SCA workspace. More complete information can be found in Chapter 8 of *The SCA Statistical System: Reference Manual for Fundamental Capabilities*.

## D.4 UTILITY RELATED INFORMATION

### D.2.1 Directing output to a file and output review

In some situations, we may wish to simultaneously route to a file all, or portions, of SCA output that are displayed at our terminal screen.

When we enter the SCA System, the System automatically opens a file called SCAOUTP.OTP . This file remains "attached" for the remainder of our SCA session and is assigned an internal unit number of **10**. To simultaneously route the output to this file, we simply enter

**PROFILE REVIEW**

To stop this flow of output to the output file, we enter the SCA statement

**PROFILE NO REVIEW**

Output will then be displayed at our screen only. If we re-specify

**PROFILE REVIEW**

at any point of the session, the output will again be directed to the file SCAOUTP.OTP. In the PC environment, any new output directed to the file is appended so that previous information will not be overwritten. However, previous output will be overwritten in the mainframe or workstation environment.

In the PC environment, we may review the output information on the file at any time by entering

**REVIEW**

The current SCA session will be suspended temporarily and we can review what we have routed to the file. Scrolling instructions at this time are accessed through the movement keys on the numeric keypad (Pageup, Pagedown, Home, End, arrow up, arrow down). To terminate this review of output and continue with our SCA session, we press the ESC key.

In order to review this output information on a mainframe computer or workstation we can temporarily suspend the current SCA session by using the OS paragraph (see Section D.4. The file SCAOUTP.OTP can be viewed using a local editor.

If the SCA System is accessed through the SCA Windows/Graphics Package, output information is automatically stored on the file SCAOUTP.OTP on our PC and appears in the SCA output window. Output information can be reviewed at any time during the SCA session by scrolling the output window. The file SCAOUTP.OTP exists in the PC subdirectory \SCAWIN and is available at the end of an SCA session.

The file SCAOUTP.OTP is automatically opened and rewound when a new SCA session is started. Hence, if we want to keep a permanent copy of this file, we must either rename the file, or copy the file, before we invoke a new SCA session.

### D.2.2 Adjusting input and output width

The default display (output) width for the SCA System is 80 columns. Similarly the default input width is 72 columns. These defaults accommodate all input and output devices. We may find it convenient to "re-adjust" these defaults to better reflect the devices we are employing or the output we will generate. For example, we can extend the input width to 80 columns and display (output) width to 132 columns by entering

**PROFILE IWIDTH IS 80. OWIDTH IS 132.**

To be certain that we have these widths throughout our session, we should make this the first command within our SCA session.

## D.3 Managing the SCA Workspace: the WORKSPACE paragraph

Although the SCA System manages the workspace automatically, on occasion we may need to manage the workspace ourselves. This is especially true if we need to "create" more space in our workspace for large data sets (by deleting current variables from our workspace) or if we wish to copy (or retrieve) our workspace to (from) an external file.

### D.3.1 Saving and retrieving a workspace

We can "suspend" an SCA session, and continue from where we were, by saving the contents of our current workspace to a file, and later retrieving it. The SCA System automatically assigns a workspace file as unit 9 when we start a session. To save workspace to this file, we can enter

**WORKSPACE MEMORY IS SAVED (9)**

or simply

**WORKSPACE SAVED (9)**

To recall this workspace at some later time, we can enter

**WORKSPACE MEMORY IS RECALLED (9)**

or simply

**WORKSPACE RECALLED**

Note that if we use a file name other than the one assigned by the SCA System, we must use the ASSIGN paragraph to associate the file with the appropriate unit number (see Example 1 in Section D.1).

### D.3.2 Deleting variables from the workspace

The WORKSPACE paragraph is used if we need to remove variables from the current workspace. For example, if we need to delete the variables A1DATA, BDATA, and CDATA, we can enter

**WORKSPACE DELETE A1DATA, BDATA, CDATA. COMPRESS.**

The COMPRESS sentence is included to compress the space occupied by remaining variables. If we do not specify this sentence, then the SCA System may not compress the workspace automatically.

### D.3.3 Workspace content

We can display the content of our workspace (i.e., variable and model names) and the amount of space occupied, by entering the command

**WORKSPACE CONTENT**

### D.3.4 Increasing the size of the SCA workspace

On occasions in an SCA session, especially when a large data set is involved or in the estimation of many parameters in a multivariate time series model, the amount of available workspace may not be sufficient. If we find that more workspace is necessary to continue an analysis, the following steps should be taken in an interactive SCA session:

(1) Save the contents of the current SCA workspace to an external file. This is accomplished by the WORKSPACE paragraph (using the SAVED option of the MEMORY sentence).

(2) Exit the SCA System (i.e., STOP).

(3) Re-execute the SCA load module with more workspace allocated. (See Appendix D of *The SCA Statistical System: Reference Manual for Fundamental Capabilities* or a local computer consultant for the instructions appropriate for the host computer environment.)

(4) Once in a new SCA session, we may recall the contents of the old SCA workspace back to the current session by the WORKSPACE paragraph (using the RECALLED option of the MEMORY sentence).

As a result of the above steps, we now have the contents of the previously saved SCA workspace but with a larger size at our disposal. In this way an analysis may continue from the point it was stopped. However, if the SCA System is exited before the current workspace is saved to an external file, the contents of the current memory are lost.

## D.4 Access to the Host Operating System, the OS Paragraph

Frequently it is desirable to be able to access the operating system commands of the host computer while still in an SCA session. The SCA System provides us with such a capability with the use of the OS (Operating System) paragraph. If we enter OS during an SCA session, we temporarily enter the operating system environment. At this time, most of the operating system commands, such as text editing, file allocation, de-allocation (freeing), copying, and listing can be performed. However, some operating system commands may be inaccessible. For more information on what may be accessed, we may need to check with Appendix D of *The SCA Statistical System: Reference Manual for Fundamental Capabilities* or local consultants. We may return to the SCA session by issuing a QUIT or END statement (or **exit** statement on hP/UX.

## D.5 The RESTART Paragraph

In some situations, we may work on several unrelated analyses during the same SCA session. It may be desirable to re-initialize the workspace once a task is completed. This can be achieved by issuing a RESTART statement. This effectively erases the current workspace.

# SUMMARY OF THE SCA PARAGRAPHS IN APPENDIX D

This section provides a summary of those SCA paragraphs employed in this appendix. In most cases, the syntax presented for a paragraph reflects only a portion of the capabilities of the paragraph. More complete information may be found in Chapter 8 of *The SCA Statistical System: Reference Manual for Fundamental Capabilities*.

Each SCA paragraph begins with a paragraph name and is followed by modifying sentences. Sentences that may be used as modifiers for a paragraph are shown below and the types of arguments used in each sentence are also specified. Sentences not designated required may be omitted as default conditions (or values) exist. The most frequently used required sentence is given as the first sentence of the paragraph. The portion of this sentence that may be omitted is underlined. This portion may be omitted only if this sentence appears as the first sentence in a paragraph. Otherwise all portions of the sentence must be used. The last character of each line except the last line must be the continuation character, '@'

The paragraphs to be explained in this summary are ASSIGN, PROFILE, WORKSPACE, OS, and RESTART

Legend (see Chapter 2 for further explanation)

v : variable name
i : integer
w : keyword
'c' : character data (must be enclosed within single apostrophes)

## ASSIGN Paragraph

### Syntax of the ASSIGN Paragraph

(A) Assigning an existing file

```
ASSIGN      FILE IS i.                                          @
            EXTERNAL-NAME IS 'c'.                               @
            ATTRIBUTE IS ACCESS(READ/WRITE/BOTH),               @
                  SHARE(YES/NO).

Required sentences:  FILE and EXTERNAL
```

(B) Assigning a new file

```
ASSIGN      NEW-FILE.                                           @
            FILE IS i.                                          @
            EXTERNAL-NAME IS 'c'.                               @
            ATTRIBUTES ARE                                      @
                  ACCESS(READ/WRITE/BOTH),SHARE(YES/NO),        @
                  FILE_FORMAT(FORMAT/BINARY),                   @
                  TRACKS(i),BLKSIZE(i),RECLENGTH(i),            @
                  DISPOSITION(CATALOG/DELETE).

Required sentences:  FILE, EXTERNAL and NEW-FILE
```

### Sentences Used in the ASSIGN Paragraph

**FILE sentence**

The FILE sentence is used to specify a file unit number for a new or an existing file in an SCA session. On some operating systems, this unit number may only be valid within the same SCA session.

**EXTERNAL-NAME sentence**

The EXTERNAL-NAME sentence specifies the file name used by the host computer's operating system. File name conventions may differ from computer to computer. The user should consult local documentation for external file name conventions.

**NEW-FILE sentence**

The NEW-FILE sentence is used to indicate that the file to be assigned is a new file. The default is NO NEW-FILE, i.e., the file exists.

**ATTRIBUTE sentence**

The ATTRIBUTE sentence is used to specify the characteristics of a file. The keywords in this sentence are:

ACCESS : specifies whether the file is READ only, WRITE only, or both READ and WRITE (BOTH). The specification is only valid within the same SCA session. The default is READ only. Note that a file used for saving data, workspace, or output must be assigned as writable.

SHARE : specifies whether the file will be used in sharing or exclusive mode. Sharing denotes the file may be used by more than one user at the same time. Exclusive denotes that the file may not be shared. The default is YES, i.e., sharing mode.

FILE_FORMAT : specifies whether the file is a FORMATTED or BINARY file. The default is FORMATTED file.

TRACKS : specifies the number of tracks to be initially assigned to the file. The default is 10 tracks.

BLKSIZE : specifies the block size (in characters) of the file. The default is 1600 characters.

RECLENGTH : specifies the logical record length (in characters) of the file. The default is 80 characters.

DISPOSITION : specifies whether the file is to be CATALOGUED or DELETED after file is freed. The default is CATALOG.

## PROFILE Paragraph

The PROFILE paragraph is used to control key features of an SCA session, such as routing information to a file, the width of input/output devices, and the level of output desired.

## Syntax for the PROFILE Paragraph

```
PROFILE     REVIEW/NO REVIEW.      @
            STYLE  IS  w.          @
            ECHO./NO ECHO.         @
            IWIDTH  IS  i.         @
            OWIDTH  IS  i.         @
            OUTPUT-LEVEL  IS  w.

Required sentences:  none
```

## Sentences Used in the PROFILE Paragraph

**REVIEW sentence**

The REVIEW sentence is used to specify that output will be simultaneously displayed on the terminal device and routed to the file SCAOUTP.OTP. This dual routing is continued until the sentence NO REVIEW is specified.

**STYLE sentence**

The STYLE sentence is used to specify the level of prompting provided to the user during an SCA session. The style of an SCA session is either batch or interactive. The keyword BATCH must be specified if the system is used in batch mode. For the interactive mode, the style may be either ALL or PARTIAL. The default style is PARTIAL.

In a PARTIAL session, required sentences and some other important sentences are prompted if they are not provided as basic instructions. All logical sentences and assignment sentences with defaults are not prompted. An ALL style will cause all sentences to be prompted unless the sentence is specified in the basic set of user instructions or the sentence is rarely used.

**ECHO sentence**

The ECHO sentence is used to specify the echo (display) of user's instructions. When ECHO is specified, the SCA System will display user instructions after they are entered. This option is also useful when the input instructions come from cards (e.g., in batch mode) rather than from the terminal or when a macro procedure (see Appendix C) is invoked. When the input instructions come from the terminal, the ECHO option is also useful since the communication line which connects the terminal and the computer may be noisy (defective) on occasions. This option allows the user to know what information the computer actually received. The NO ECHO instruction turns off the display of basic instructions. The default option is ECHO.

**IWIDTH sentence**

The IWIDTH sentence is used to specify the width (in number of characters) for the input device. The width may range from 60 to 80 characters. The IWIDTH also applies to statements from a macro procedure (see Appendix C) or data from a file. The width of records on a data file can also be specified in the INPUT paragraph (see Chapter 2). Since columns 73 to 80 on a record are usually reserved for sequence numbers, the default width is assumed to be 72.

**OWIDTH sentence**

The OWIDTH sentence is used to specify the width (in number of characters) of the output device. Both the analytic and English-like statements automatically adjust the output format according to the specified output device width. The default output width is 80 characters.

**OUTPUT-LEVEL sentence**

The OUTPUT-LEVEL sentence is used to indicate the overall output level desired in an SCA session. The keyword is NONE, BRIEF, NORMAL, or DETAILED. The default output amount is NORMAL. If NONE is specified, the echo of the basic instructions is also turned off. No output is displayed when an analytic statement is used, and the output from an English-like statement is same as in BRIEF level. The user is responsible for most of the output. This option is useful when the SCA System is used strictly as a programming language. If BRIEF, NORMAL, or DETAILED is specified, the SCA System sets a default level of output for each English-like statement according to the specified level. This default option may be modified in a particular paragraph by the OUTPUT sentence of the paragraph.

## WORKSPACE Paragraph

The WORKSPACE paragraph is used to manage the user's workspace, such as displaying current status, deleting unneeded variables, saving or recalling the workspace, or consolidating the unused workspace.

## Syntax for the WORKSPACE Paragraph

| | | |
|---|---|---|
| **WORKSPACE** | MEMORY IS SAVED(i), RECALLED(i). | @ |
| | DELETE v1, v2, --- . | @ |
| | COMPRESSION./NO COMPRESSION. | @ |
| | NOVAR-REQUIRED IS i. | @ |
| | CONTENT./NO CONTENT. | |

Required sentences: none

## Sentences Used in the WORKSPACE Paragraph

**MEMORY sentence**

The MEMORY sentence is used to save the contents of the current SCA workspace to a file or recall a previously saved SCA workspace from a file. The SAVED keyword specifies the logical unit of the file where the workspace will be saved, and RECALLED specifies the logical unit of the file containing the workspace to be recalled. If both SAVED and RECALLED are used, the current workspace is first saved to the designated file and then a previous workspace is recalled from another name. The default logical unit for a workspace file is 9. Therefore if the default file unit is used, the following two statements are both acceptable

WORKSPACE MEMORY IS SAVED. (or simply WORKSPACE SAVED.)

WORKSPACE MEMORY IS RECALLED. (or simply WORKSPACE RECALLED.)

**DELETE sentence**

The DELETE sentence is used to specify the names of the variables and/or models to be deleted. Note that the deletion does not increase the available workspace unless the workspace is compressed.

**COMPRESSION sentence**

The COMPRESSION sentence is used to specify the compression of the SCA workspace. When a variable is deleted, whether implicitly by the processor or explicitly by the user, the SCA System does not compress the workspace immediately. When the user runs out of workspace, unneeded variables and models may be deleted and the workspace compressed in order to release unused workspace. The default option is NO COMPRESSION.

**NOVAR sentence**

The NOVAR sentence is used to specify the number of additional variables desired in an SCA session beyond those already in the workspace. The SCA System initially allows up to 150 variables in the workspace. If the user requires more than 150 variables, the variable list may be expanded to meet the user's requirement.

**CONTENT sentence**

The CONTENT sentence requests the system to display the bookkeeping information of an SCA session. The bookkeeping information includes the names of the variables and models in the workspace, and the amount of workspace used. The default is NO CONTENT.

## OS Paragraph

The OS paragraph is used to access the host computer's operating system commands during an SCA session. Most of the operating system commands, such as file allocation, de-allocation (freeing), copying, listing, and text editing, can then be accessed. However, some operating system commands may be inaccessible. The OS paragraph does not have any modifying sentences. The user may return to the SCA session by issuing a QUIT statement.

## RESTART Paragraph

The RESTART paragraph is used to initialize the SCA workspace and begin another SCA session. The RESTART paragraph has no modifying sentences.