



THE
POWER
TO KNOW.

SAS/INSIGHT[®] 9.1

User's Guide

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2004.
SAS/INSIGHT® 9.1 User's Guide. Cary, NC: SAS Institute Inc.

SAS/INSIGHT® 9.1 User's Guide

Copyright © 2004, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-58025-697-1

All rights reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a Web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

U.S. Government Restricted Rights Notice: Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, March 2004

2nd printing, November 2006

3rd printing, March 2008

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at **support.sas.com/publishing** or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Contents

Part 1. Introduction	1
Chapter 1. Getting Started	3
 Part 2. Techniques	 23
Chapter 2. Entering Data	25
Chapter 3. Examining Data	47
Chapter 4. Exploring Data in One Dimension	69
Chapter 5. Exploring Data in Two Dimensions	85
Chapter 6. Exploring Data in Three Dimensions	107
Chapter 7. Adjusting Axes and Ticks	123
Chapter 8. Labeling Observations	133
Chapter 9. Hiding Observations	143
Chapter 10. Marking Observations	155
Chapter 11. Coloring Observations	167
Chapter 12. Examining Distributions	177
Chapter 13. Fitting Curves	199
Chapter 14. Multiple Regression	217
Chapter 15. Analysis of Variance	241
Chapter 16. Logistic Regression	261
Chapter 17. Poisson Regression	277
Chapter 18. Examining Correlations	293
Chapter 19. Calculating Principal Components	303
Chapter 20. Transforming Variables	317
Chapter 21. Comparing Analyses	337
Chapter 22. Analyzing by Groups	355
Chapter 23. Animating Graphs	367
Chapter 24. Formatting Variables and Values	375
Chapter 25. Editing Windows	391

Chapter 26. Saving and Printing Data	419
Chapter 27. Saving and Printing Graphics	429
Chapter 28. Saving and Printing Tables	443
Chapter 29. Configuring SAS/INSIGHT Software	451
Chapter 30. Working with Other SAS Products	469
Part 3. Reference	483
Chapter 31. Data Windows	485
Chapter 32. Histograms and Bar Charts	497
Chapter 33. Box Plots and Mosaic Plots	505
Chapter 34. Line Plots	519
Chapter 35. Scatter Plots	525
Chapter 36. Contour Plot	533
Chapter 37. Rotating Plot	543
Chapter 38. Distribution Analyses	553
Chapter 39. Fit Analyses	611
Chapter 40. Multivariate Analyses	705
Chapter 41. SAS/INSIGHT Statements	777
Index	791

Part 1

Introduction

Contents

Chapter 1. Getting Started	3
--------------------------------------	---

Introduction

Chapter 1

Getting Started

Chapter Contents

SUMMARY OF FEATURES	6
OF MICE AND MENUS	8
Selecting Objects	8
Choosing from Menus	10
Pop-up Menus	11
Menu State Indicators	13
LEARNING MORE	15
Using This Manual	15
Conventions	15
Getting Help	15
SAMPLE DATA SETS	18
REFERENCES	22

Getting Started

Chapter 1

Getting Started

SAS/INSIGHT software is a tool for data exploration and analysis. With it you can explore data through graphs and analyses linked across multiple windows. You can analyze univariate distributions, investigate multivariate distributions, and fit explanatory models using analysis of variance, regression, and the generalized linear model.

This introduction summarizes important features, describes how to use the product, and explains how to learn more about SAS/INSIGHT software.

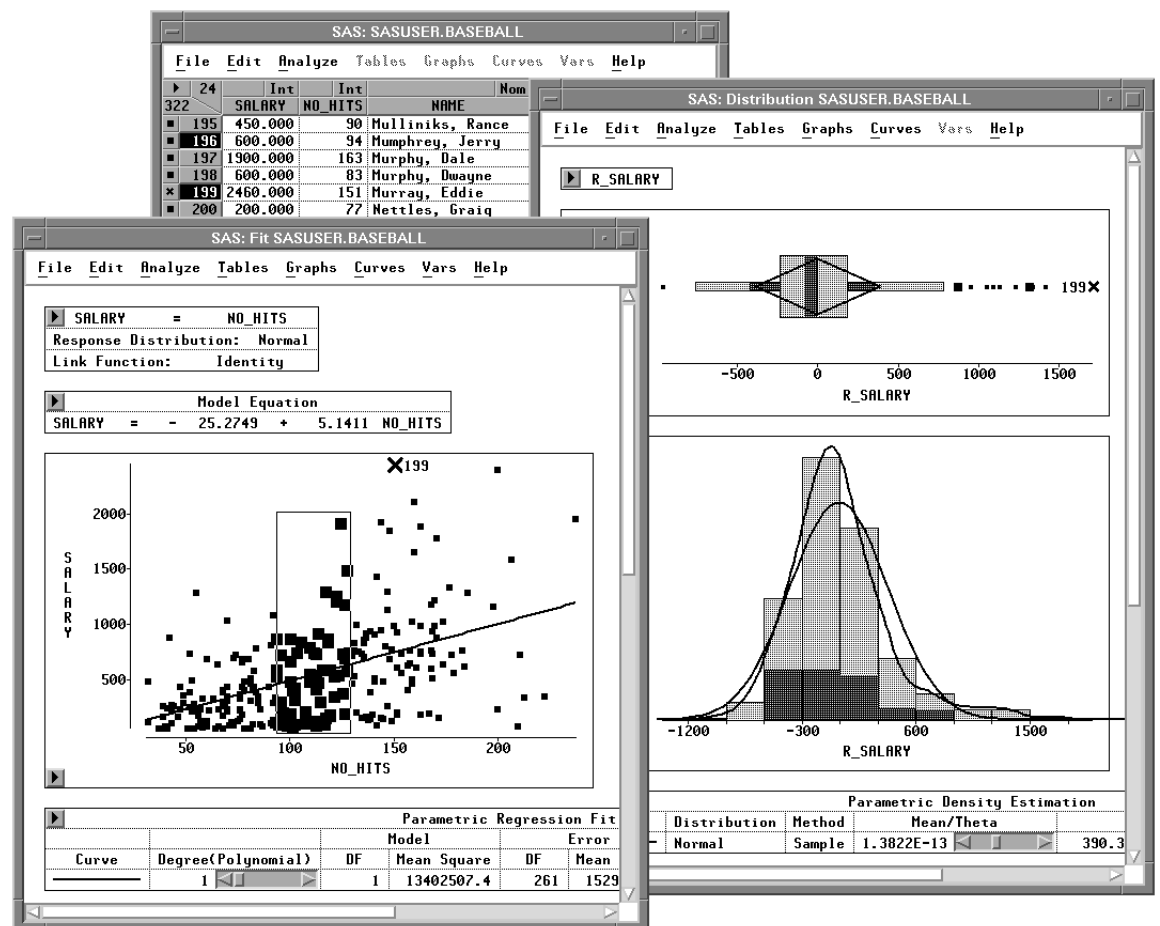


Figure 1.1. Brushing Observations in SAS/INSIGHT Software

Summary of Features

SAS/INSIGHT software provides a comprehensive set of exploratory and analytical tools.

To explore data, you can

- identify observations in plots
- brush observations in linked scatter plots, histograms, box plots, line plots, contour plots, and three-dimensional rotating plots
- exclude observations from graphs and analyses
- search, sort, and edit data
- transform variables
- color observations based on the value of a variable

To analyze distributions, you can

- compute descriptive statistics
- create quantile-quantile plots
- create mosaic plots of cross-classified data
- fit parametric (normal, lognormal, exponential, Weibull) and kernel density estimates for distributions
- fit parametric and empirical cumulative distribution functions
- test hypotheses of completely specified (known parameters) or specific (unknown parameters) parametric distributions based on Kolmogorov's D statistic

To analyze relationships between a response variable and a set of explanatory variables, you can

- fit curves with polynomials, kernels, and smoothing splines
- fit curves with nonparametric local polynomial smoothers using either a fixed bandwidth or loess smoothing
- add confidence bands for mean and predicted values
- fit surfaces with polynomials, kernels, and smoothing splines
- create residual and leverage plots
- fit the general linear model, including classification effects for analysis of variance and analysis of covariance
- fit the generalized linear model, including logistic regression, Poisson regression, and other models

To analyze relationships between variables, you can

- calculate correlation matrices and scatter plot matrices with confidence ellipses for relationships among pairs of variables
- reduce dimensionality of interval variables with principal component analysis
- examine relationships between two sets of interval variables with canonical correlation analysis and maximum redundancy analysis
- examine relationships between a nominal variable and a set of interval variables with canonical discriminant analysis

In addition, you can

- process data by groups
- process multiple data sets
- store option settings to customize SAS/INSIGHT operation
- store results as SAS data sets, SAS/GRAPH catalogs, and text files
- record and submit SAS/INSIGHT statements
- obtain context-sensitive help

Finally, because it is a part of the SAS System, you can use SAS/INSIGHT software to explore results from any SAS procedure. Conversely, you can use any SAS procedure to analyze results from SAS/INSIGHT software.

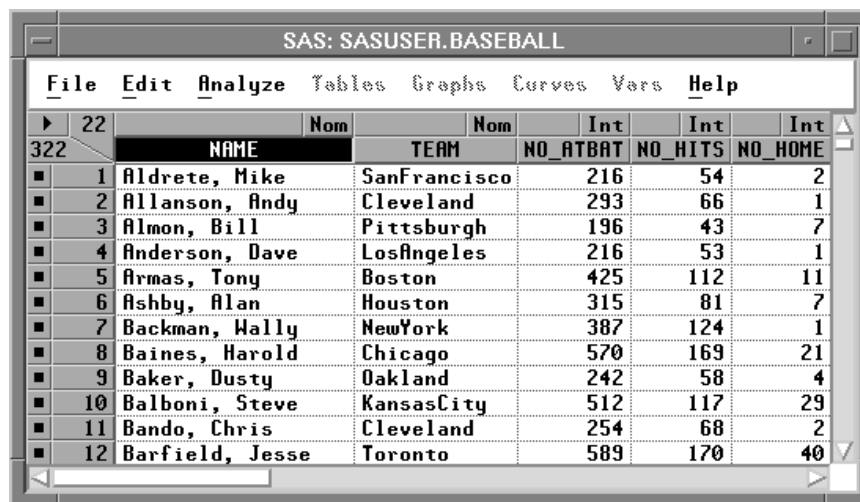
Of Mice and Menus

This section describes how to operate SAS/INSIGHT software and defines terms used in the rest of this book.

Some details depend on your *host*, the specific system of computing hardware and software you use. For example, all hosts present SAS/INSIGHT software in a system of *windows* on the host's *display*, but the appearance of your windows may differ from the figures in this book. You can find more information in the SAS companion for your host and in your host system documentation. On most hosts, you can *point* to objects on the display by using a *mouse*. A mouse is a physical device that controls the location of a *cursor*, a small moveable symbol on the display. The mouse also has *buttons* that work like keys on the computer keyboard. By pointing with the mouse and clicking a button, you can indicate any object on the display. In SAS/INSIGHT software, all operations you may want to perform are listed in *menus*. So to perform any task, you point with the mouse and click the buttons to select objects and choose operations from menus.

Selecting Objects

Objects you can use in SAS/INSIGHT software include variables, observations, values, graphs, curves, and tables. You *select* an object to indicate that it is an object you want to work with. On most hosts, you can select an object by pointing to it and clicking the leftmost button on the mouse. To *click*, press the button down and release it without moving the mouse. [Figure 1.2](#) illustrates the selection of a variable by pointing and clicking.

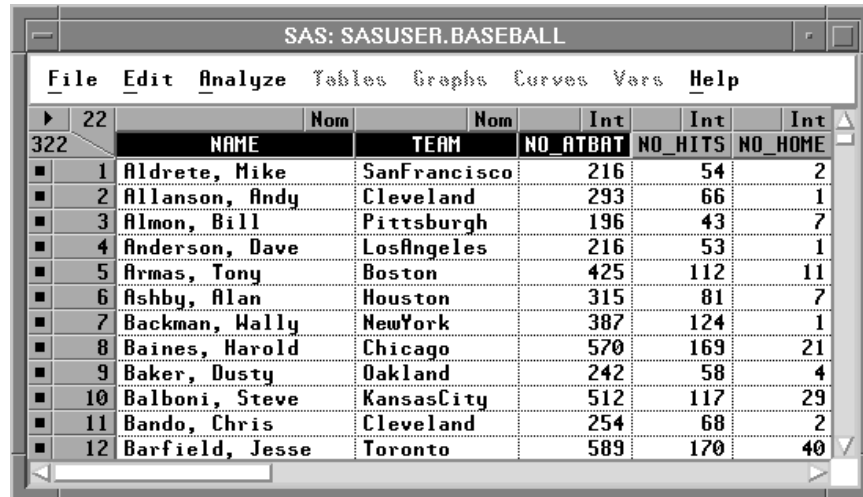


		Nom	Nom	Int	Int	Int
	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME	
1	Aldrete, Mike	SanFrancisco	216	54	2	
2	Allanson, Andy	Cleveland	293	66	1	
3	Almon, Bill	Pittsburgh	196	43	7	
4	Anderson, Dave	LosAngeles	216	53	1	
5	Armas, Tony	Boston	425	112	11	
6	Ashby, Alan	Houston	315	81	7	
7	Backman, Wally	NewYork	387	124	1	
8	Baines, Harold	Chicago	570	169	21	
9	Baker, Dusty	Oakland	242	58	4	
10	Balboni, Steve	KansasCity	512	117	29	
11	Bando, Chris	Cleveland	254	68	2	
12	Barfield, Jesse	Toronto	589	170	40	

Figure 1.2. Selecting by Clicking

You can select multiple objects by *dragging* the mouse. To drag, press the leftmost mouse button down, move the mouse across the objects of interest, then release the mouse button. This selects the object at the cursor location when you pressed the

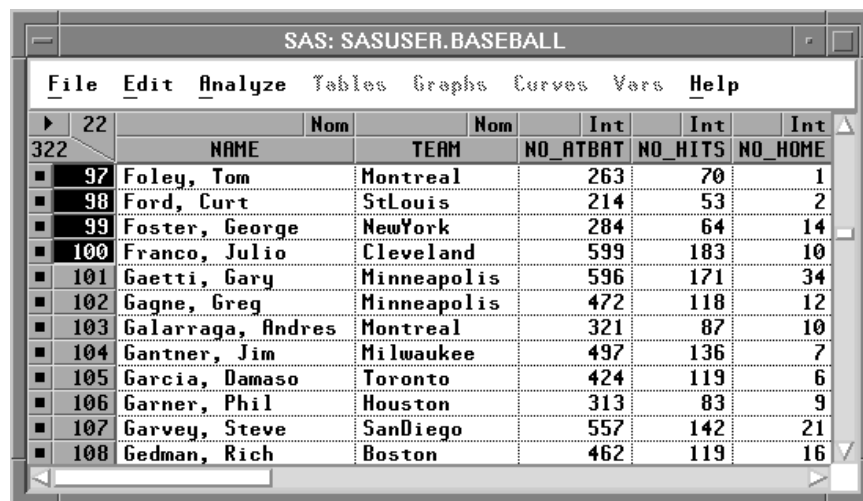
mouse button, the object where you released the button, and all objects in between. Figure 1.3 illustrates the selection of three variables by pointing and dragging.



		Nom	Nom	Int	Int	Int
	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME	
1	Aldrete, Mike	SanFrancisco	216	54	2	
2	Allanson, Andy	Cleveland	293	66	1	
3	Almon, Bill	Pittsburgh	196	43	7	
4	Anderson, Dave	LosAngeles	216	53	1	
5	Armas, Tony	Boston	425	112	11	
6	Ashby, Alan	Houston	315	81	7	
7	Backman, Hally	NewYork	387	124	1	
8	Baines, Harold	Chicago	570	169	21	
9	Baker, Dusty	Oakland	242	58	4	
10	Balboni, Steve	KansasCity	512	117	29	
11	Bando, Chris	Cleveland	254	68	2	
12	Barfield, Jesse	Toronto	589	170	40	

Figure 1.3. Selecting by Dragging

When objects are far apart, it is convenient to use *modifier keys* with the mouse button. On many hosts, you can use the **Shift** key to *extend* a selection. In Figure 1.4, the first observation was clicked on, then the one hundredth observation was clicked on while holding down the **Shift** key. This selects the first observation, the one hundredth observation, and all observations in between.



		Nom	Nom	Int	Int	Int
	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME	
97	Foley, Tom	Montreal	263	70	1	
98	Ford, Curt	StLouis	214	53	2	
99	Foster, George	NewYork	284	64	14	
100	Franco, Julio	Cleveland	599	183	10	
101	Gaetti, Gary	Minneapolis	596	171	34	
102	Gagne, Greg	Minneapolis	472	118	12	
103	Galarrraga, Andres	Montreal	321	87	10	
104	Gantner, Jim	Milwaukee	497	136	7	
105	Garcia, Damaso	Toronto	424	119	6	
106	Garner, Phil	Houston	313	83	9	
107	Garvey, Steve	SanDiego	557	142	21	
108	Gedman, Rich	Boston	462	119	16	

Figure 1.4. Extended Selection

On many hosts, you can use the **Ctrl** key to make a *noncontiguous* selection – that is, a selection of multiple objects not located next to each other. In Figure 1.5, the first observation was clicked on, then the fifth observation was clicked on while holding down the **Ctrl** key. This selects the first observation and the fifth observation without

selecting the observations in between.

	Nom	Nom	Int	Int	Int
	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME
97	Foley, Tom	Montreal	263	70	1
98	Ford, Curt	StLouis	214	53	2
99	Foster, George	NewYork	284	64	14
100	Franco, Julio	Cleveland	599	183	10
101	Gaetti, Gary	Minneapolis	596	171	34
102	Gagne, Greg	Minneapolis	472	118	12
103	Galarraga, Andres	Montreal	321	87	10
104	Gantner, Jim	Milwaukee	497	136	7
105	Garcia, Damaso	Toronto	424	119	6
106	Garner, Phil	Houston	313	83	9
107	Garvey, Steve	SanDiego	557	142	21
108	Gedman, Rich	Boston	462	119	16

Figure 1.5. Noncontiguous Selection

Some hosts use different modifier keys instead of the **Shift** and **Ctrl** keys, so these names do not appear in the remainder of this book. Instead, the terms *extended selection* and *noncontiguous selection* are used. Using single, multiple, extended, and noncontiguous selection, you can precisely indicate the objects you want to work with.

Choosing from Menus

In SAS/INSIGHT software, operations you can perform include creating graphs and analyses, transforming variables, fitting curves, and saving results. On most hosts, you can choose these operations by *pulling down* a menu from a *menu bar*. To pull down a menu, press the left mouse button and hold it down while you drag the cursor across the menu. **Figure 1.6** shows the **Analyze** menu pulled down to create a scatter plot.

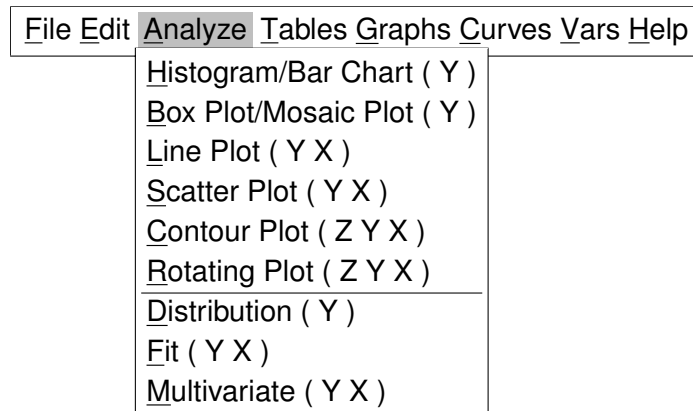


Figure 1.6. Analyze Menu

Depending on your host, each window may display its own menu bar or all windows may share a single menu bar. Workstations with large displays usually provide mul-

multiple menu bars. Personal computers with small displays may allow only one menu bar.

Your host may provide additional choices on the menu bar and within the **File** and **Help** menus. These additional menu choices, if present, are described in the SAS companion for your host.

Pop-up Menus

Pop-up menus enable fast action by providing choices appropriate for the object you point to. Pop-up menus operate on all appropriate selected objects. If no objects are selected, they operate on the object at the cursor location.

Pop-up menus are displayed when you click on *menu buttons* in the data window and in the corners of graphs and tables. On some hosts, you can also display pop-up menus by pressing the right mouse button.

The data window displays a variety of pop-up menus. To display the pop-up menu for data, either click the left mouse button in the upper left corner, as in [Figure 1.7](#), or click and hold the right mouse button anywhere in the data window. See [Chapter 31, “Data Windows,”](#) for a complete description of the pop-up menu choices in the data window.

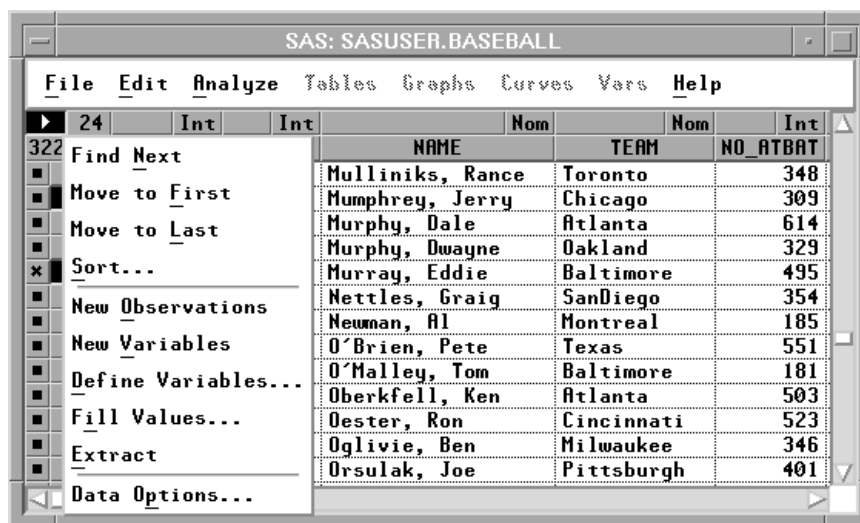


Figure 1.7. Data Pop-up Menu

To display pop-up menus in a graph or table, either click and hold the right mouse button anywhere in the graph or table, or click on the menu button in the corner of the graph or table. [Figure 1.8](#) shows the pop-up menu for a histogram in a distribution analysis.

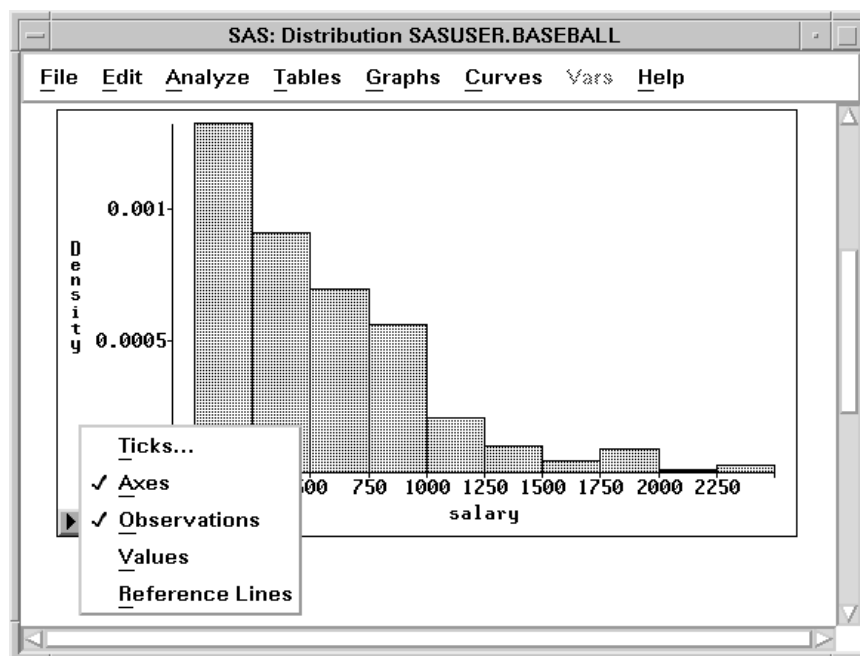


Figure 1.8. Histogram Pop-up Menu

When you are not pointing at a table, graph, or other object, the right mouse button displays the central menu bar, as in [Figure 1.9](#). For more information on pop-up menu choices, see the chapter for the graph or table of interest in the Reference part of this manual.

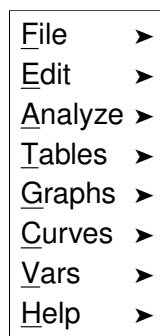


Figure 1.9. Default Pop-up Menu

Menu State Indicators

Menu *state indicators* are either check marks or radio marks. The graphic representation of these marks depends on your host.

Menus with *check marks* always act as *toggles*: they turn a feature on or off. The presence of a check mark indicates the presence of that feature. Toggles are especially useful in graphs, since most graphic features are either on or off.

Menus with *radio marks* do not toggle; they indicate the current state among multiple choices. As with check marks, radio marks help when the current state is not obvious.

For example, the pop-up menu in [Figure 1.10](#) is from a scatter plot. The check marks indicate that axes and observations are displayed and that the marker size is chosen automatically to fit the graph. The radio mark indicates that the current marker size is 4.

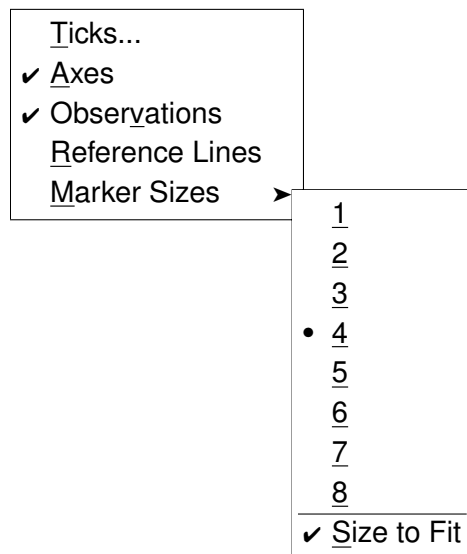


Figure 1.10. Scatter Plot Pop-up Menu

Learning More

Using This Manual

The remainder of this manual is divided into two parts: Techniques and Reference.

Techniques are instructional chapters that explain how to accomplish particular tasks. These chapters use sample data sets shipped with the product, so you can read the techniques and follow the steps on your host at the same time. For more information about sample data sets, see the “[Sample Data Sets](#)” section in this chapter.

Reference chapters provide comprehensive descriptions of data, graphs, and analyses in SAS/INSIGHT software. Use these chapters to answer specific questions about product features.

If you are experienced with SAS/INSIGHT software or experienced using mice and menus, you may learn most quickly by just invoking SAS/INSIGHT software and exploring its capabilities. Use the Table of Contents and the Index to find specific techniques and reference information.

Conventions

This user’s guide employs three special symbols:

⇒ **This symbol and font marks one step in a technique.**

⊕ **Related Reading:** This symbol and label marks a reference to a related chapter.

† **Note:** This symbol and label marks an important note or performance tip.

This user’s guide employs four special typefaces:

- **Bold** is used for steps in techniques.
- *Italic* is used for definitions and for emphasis.
- **Helvetica** is used for words you see on the display.
- `Courier` is used for examples of SAS statements.

Menu items in this user’s guide are separated by colons. For example, the **Bar Chart (Y)** item in the **Analyze** menu is written as **Analyze:Bar Chart (Y)**.

Getting Help

Both beginning and expert users can take advantage of SAS/INSIGHT software’s context-sensitive help system. To receive context-sensitive help, select any graph or table by clicking on its border. Then choose **Help:Help on Selection**, as illustrated in [Figure 1.11](#). [Figure 1.12](#) shows the context-sensitive help when the Quantiles table is selected.

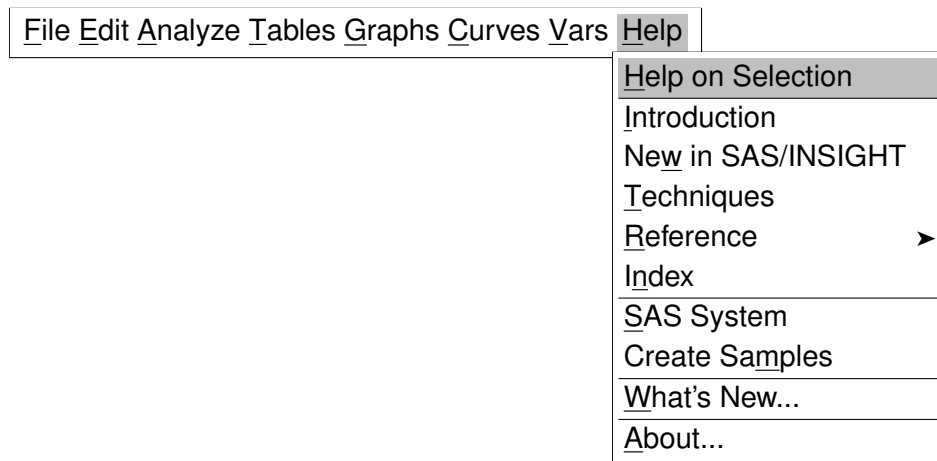


Figure 1.11. Help Menu

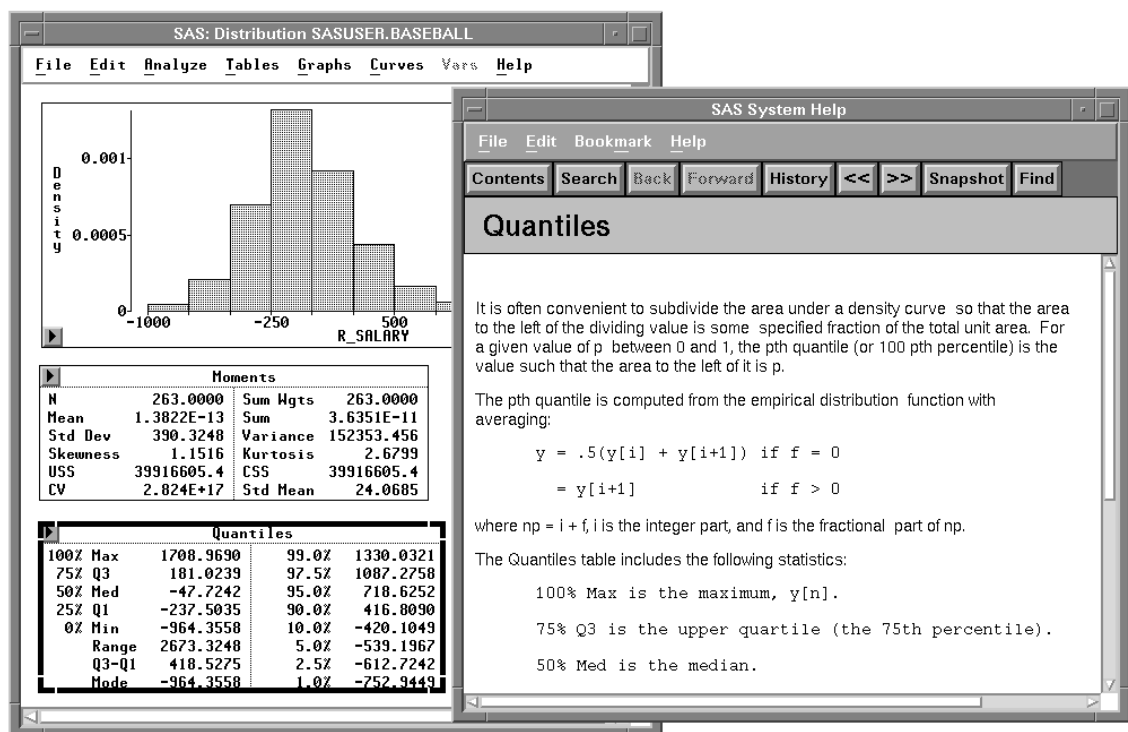


Figure 1.12. Context Sensitive Help

You can also get context-sensitive help with the SAS System *Help* key. This key, usually **F1** on your keyboard, displays help on the object at your present cursor position. You can get context-sensitive help in any SAS/INSIGHT data or analysis window by simply placing the cursor on the item of interest and pressing the **Help** key. Within any help window, you can point and click on individual topics to get further information.

The **Help** menu entries correspond to parts of this manual. Choose **Help:Introduction** to learn about SAS/INSIGHT software; **Help:Techniques** to learn how to perform a particular task; **Help:Reference** to look up detailed information; or **Help:Index** to see an index of all SAS/INSIGHT topics.

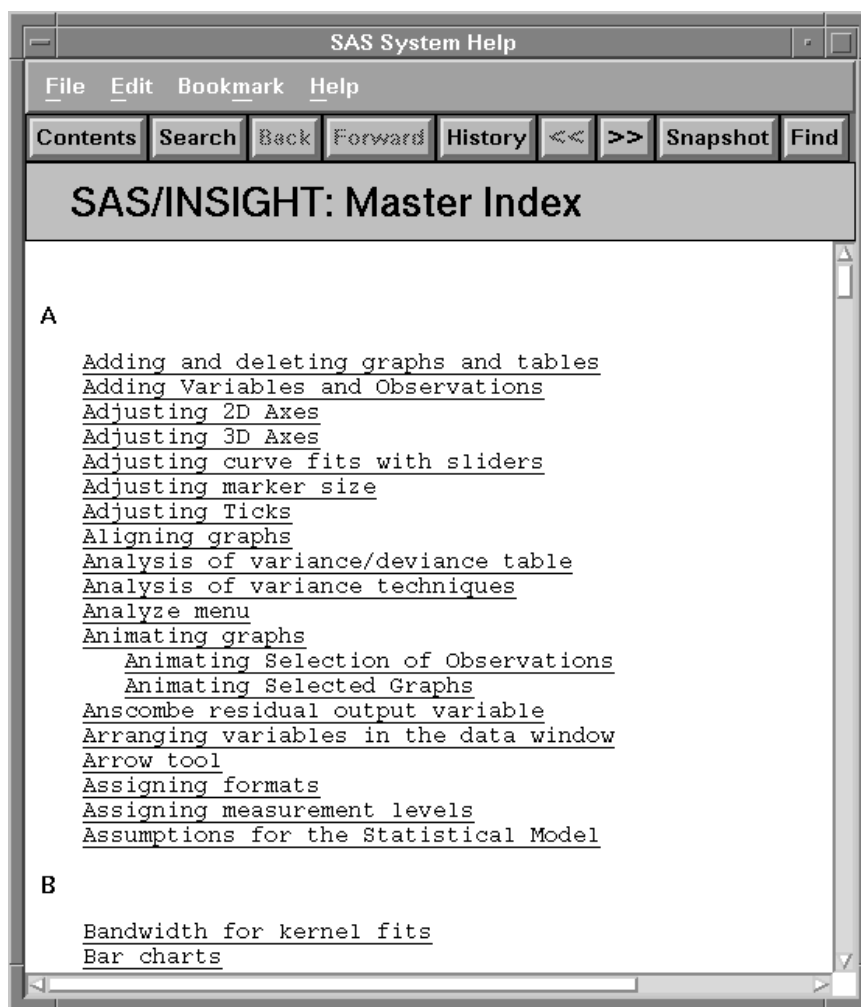


Figure 1.13. Help Index

Choose **Help:SAS System** to see a general index of SAS System topics. Choose **Help:Create Samples** to create sample data sets; examples throughout this manual refer to these data sets. See the following section for more information.

Sample Data Sets

The following sample data sets are included with SAS/INSIGHT software.

The **AIR** data set contains measurements of pollutant concentrations from a city in Germany during a week in November 1989. Variables are

DATETIME	date and hour in SAS DATETIME format
DAY	day of the week
HOURL	hour of the day
CO	carbon monoxide concentration
O3	ozone concentration
SO2	sulfur dioxide concentration
NO	nitrogen oxide concentration
DUST	dust concentration
WIND	wind speed

The **BASEBALL** data set contains performance measures and salary levels for regular hitters and leading substitute hitters in major league baseball for the year 1986 (Collier 1987). There is one observation per hitter. Variables are

NAME	the player's name
NO_ATBAT	number of times at bat in 1986
NO_HITS	number of hits in 1986
NO_HOME	number of home runs in 1986
NO_RUNS	number of runs in 1986
NO_RBI	number of runs batted in in 1986
NO_BB	number of bases on balls in 1986
YR_MAJOR	years in the major leagues
CR_ATBAT	career at bats
CR_HITS	career hits
CR_HOME	career home runs
CR_RUNS	career runs
CR_RBI	career runs batted in
CR_BB	career bases on balls
LEAGUE	player's league at the end of 1986
DIVISION	player's division at the end of 1986

TEAM	player's team at the end of 1986
POSITION	positions played in 1986
NO_OUTS	number of put outs in 1986
NO_ASSTS	number of assists in 1986
NO_ERROR	number of errors in 1986
SALARY	salary in thousands of dollars

The **POSITION** variable in the **BASEBALL** data set is encoded as follows:

13	first base, third base	CS	center field, shortstop
1B	first base	DH	designated hitter
1O	first base, outfield	DO	designated hitter, outfield
23	second base, third base	LF	left field
2B	second base	O1	outfield, first base
2S	second base, shortstop	OD	outfield, designated hitter
32	third base, second base	OF	outfield
3B	third base	OS	outfield, shortstop
3O	third base, outfield	RF	right field
3S	third base, shortstop	S3	shortstop, third base
C	catcher	SS	shortstop
CD	center field, designated hitter	UT	utility
CF	center field		

The **BUSINESS** data set contains information on publicly-held German, Japanese, and U.S. companies in the automotive, chemical, electronics, and oil refining industries. There is one observation for each company. Variables are

NATION	the nationality of the company
INDUSTRY	the company's principal business
EMPLOYS	the number of employees
SALES	sales for 1991 in millions of dollars
PROFITS	profits for 1991 in millions of dollars

The **DRUG** data set contains results of an experiment to evaluate drug effectiveness (Afifi and Azen 1972). Four drugs were tested against three diseases on six subjects; there is one observation for each test. Variables are

DRUG	the drug used in treatment
DISEASE	the disease present
CHANG_BP	the change in systolic blood pressure due to treatment

Introduction ♦ Getting Started

The **GPA** data set contains data collected to determine which applicants at a large midwestern university were likely to succeed in its computer science program (Campbell and McCabe 1984). There is one observation per student. Variables are

GPA	the grade point average of students in the computer science program
HSM	the average high school grade in mathematics
HSE	the average high school grade in English
HSS	the average high school grade in science
SATM	the score on the mathematics portion of the SAT exam
SATV	the score on the verbal portion of the SAT exam
SEX	the student's gender

The **IRIS** data set is Fisher's Iris data (Fisher 1936). Sepal and petal size were measured for fifty specimens from each of three species of iris. There is one observation per specimen. Variables are

SEPALLEN	sepal length in millimeters
SEPALWID	sepal width in millimeters
PETALLEN	petal length in millimeters
PETALWID	petal width in millimeters
SPECIES	the species

The **MINING** data set contains results of an experiment to determine whether drilling time was faster for wet drilling or dry drilling (Penner and Watts 1991). Tests were replicated three times for each method at different test holes. There is one observation per five-foot interval for each replication. Variables are

DRILTIME	the time in minutes to drill the last five feet of the current depth
METHOD	the drilling method, wet or dry
REP	the replicate number
DEPTH	the depth of the hole in feet

The **MININGX** data set is a subset of the **MINING** data set. It contains data from only one of the test holes.

The **PATIENT** data set contains data collected on cancer patients (Lee 1974). There is one observation per patient. Variables are

REMISS	1 if remission occurred and 0 otherwise
CELL	
SMEAR	
INFIL	
LI	
TEMP	
BLAST	measures of patient characteristics

The **SHIP** data set contains data from an investigation of wave damage to cargo ships (McCullagh and Nelder 1989). The purpose of the investigation was to set standards for future hull construction. There is one observation per ship. Variables are

Y	the number of damage incidents
YEAR	year of construction
TYPE	the type of ship
PERIOD	the period of operation
MONTHS	the aggregate months of service

Choose **Help:Create Samples** to create the sample data sets in your **sasuser** directory. When you have created the sample data sets, turn to the Techniques part of this manual to learn how to enter your data and begin exploring it with SAS/INSIGHT software.

† **Note:** If you have an existing data set in your **sasuser** library with the same name as a sample data set, it will be overwritten if you create the sample.

References

- Afi, A.A. and Azen, S.P. (1972), *Statistical Analysis: A Computer-Oriented Approach*, New York: Academic Press, 166.
- Campbell, P.F. and McCabe, G.P. (1984), "Predicting the Success of Freshmen in a Computer Science Major," *Communications of the ACM*, 27, 1108–1113.
- Collier Books (1987), *The 1987 Baseball Encyclopedia Update*, New York: Macmillan Publishing Company.
- Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.
- Lee, E.T. (1974), "A Computer Program for Linear Logistic Regression Analysis," *Computer Programs in Biomedicine*, 80–92.
- McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman and Hall.
- Penner, R. and Watts, D.G. (1991), "Mining Information," *American Statistician*, 45(1), 4–9.

Part 2

Techniques

Contents

Chapter 2. Entering Data	25
Chapter 3. Examining Data	47
Chapter 4. Exploring Data in One Dimension	69
Chapter 5. Exploring Data in Two Dimensions	85
Chapter 6. Exploring Data in Three Dimensions	107
Chapter 7. Adjusting Axes and Ticks	123
Chapter 8. Labeling Observations	133
Chapter 9. Hiding Observations	143
Chapter 10. Marking Observations	155
Chapter 11. Coloring Observations	167
Chapter 12. Examining Distributions	177
Chapter 13. Fitting Curves	199
Chapter 14. Multiple Regression	217
Chapter 15. Analysis of Variance	241
Chapter 16. Logistic Regression	261
Chapter 17. Poisson Regression	277
Chapter 18. Examining Correlations	293

Techniques

Chapter 19. Calculating Principal Components	303
Chapter 20. Transforming Variables	317
Chapter 21. Comparing Analyses	337
Chapter 22. Analyzing by Groups	355
Chapter 23. Animating Graphs	367
Chapter 24. Formatting Variables and Values	375
Chapter 25. Editing Windows	391
Chapter 26. Saving and Printing Data	419
Chapter 27. Saving and Printing Graphics	429
Chapter 28. Saving and Printing Tables	443
Chapter 29. Configuring SAS/INSIGHT Software	451
Chapter 30. Working with Other SAS Products	469

Chapter 2

Entering Data

Chapter Contents

INVOKING SAS/INSIGHT SOFTWARE	28
ENTERING VALUES	32
NAVIGATING THE DATA WINDOW	34
ADDING VARIABLES AND OBSERVATIONS	35
DEFINING VARIABLES	37
FAST DATA ENTRY	40
OTHER OPTIONS	44

Chapter 2

Entering Data

A *SAS data set* consists of variables and observations. *Variables* are quantities or characteristics being measured. *Observations* are sets of variable values for a single entity.

In SAS/INSIGHT software, your data are presented in a window with variables displayed in columns and observations displayed in rows, as in [Figure 2.1](#). You can enter data directly in the data window.

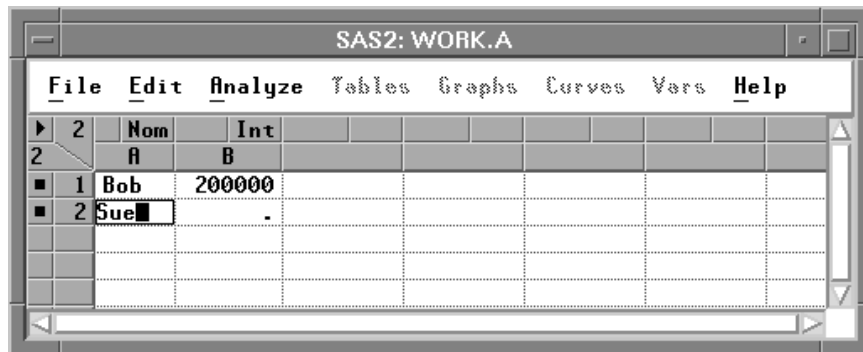


Figure 2.1. Entering Data in the Data Window

Invoking SAS/INSIGHT Software

You can invoke SAS/INSIGHT software in any of three ways.

⇒ You can type **insight** on the command line.

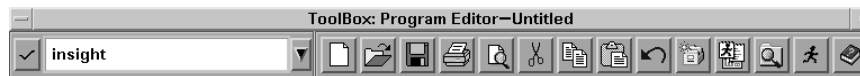


Figure 2.2. Command Line

⇒ If you have menus, you can choose **Solutions:Analyze:Interactive Data Analysis**.

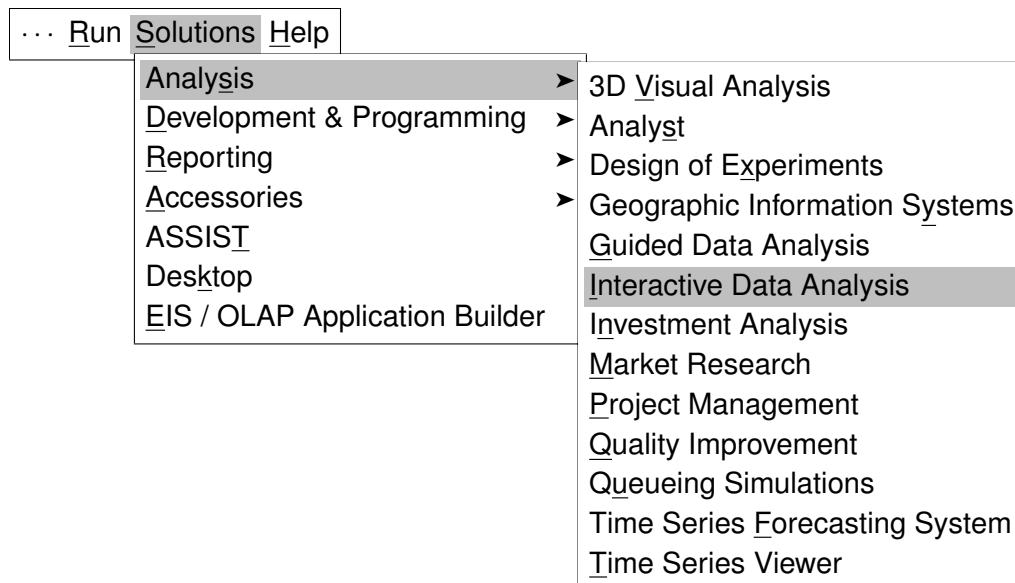


Figure 2.3. SAS Analysis Menu

⇒ You can invoke SAS/INSIGHT software as a SAS procedure.

Choose **Run:Submit** to submit the procedure statement in the Program Editor.

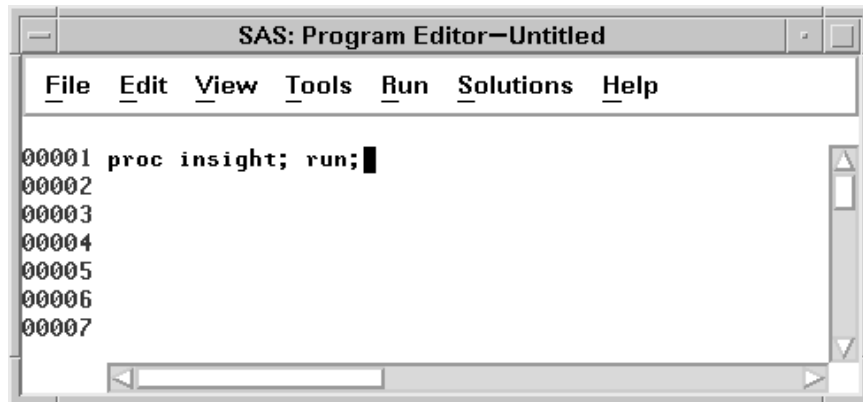


Figure 2.4. Entering a PROC Statement

You may want to access SAS data sets that are located in different libraries than the standard ones. As an example, if you have SAS data sets in a directory named **mypath**, then enter the lines

```
libname mylib 'mypath';  
proc insight;  
run;
```

in the Program Editor window and choose **Run:Submit**. The data set dialog (discussed later) will contain an additional library **mylib** to choose from.

You can invoke SAS/INSIGHT software from the Program Editor window and automatically open a new data window. Enter the lines

```
proc insight data;  
run;
```

in the Program Editor window and choose **Run:Submit**. The data set dialog is skipped and a new data window appears.

You can specify a data set directly. For example, if you have a SAS data set named **mydata** in the **mylib** directory, enter the lines

```
libname mylib 'mypath';  
proc insight data=mylib.mydata;  
run;
```

in the Program Editor window and choose **Run:Submit**. Again the data set dialog is skipped and a data window appears with the specified SAS data set.

Finally, if you have *raw data* that you want to analyze, you most likely need to use the INFILE and INPUT statements in a DATA step. Refer to *SAS Language Reference: Dictionary* for information on how to read in raw data.

† **Note:** It is best to invoke SAS/INSIGHT software from the command line or from the **Solutions** menu. This enables you to use SAS/INSIGHT software simultaneously with other components in the SAS System. If you invoke it as a procedure, you cannot use any other SAS component until you exit SAS/INSIGHT.

Upon invoking SAS/INSIGHT software, you are prompted with a data set dialog.



Figure 2.5. Data Set Dialog

⇒ **Click the New button.**

This opens a new data window in which you can enter data.

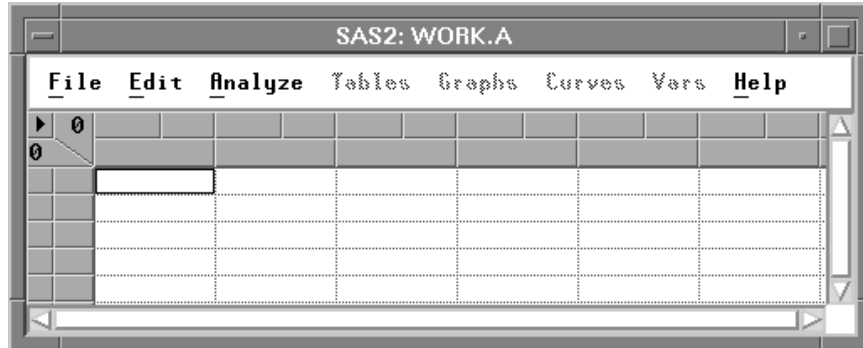


Figure 2.6. New Data Window

Entering Values

By default, the first value in a new data window is selected and is displayed with a frame around it. This *active value* marks your current location in the data window. To enter data, simply begin typing.

⇒ **Enter the name “Bob” in the active value.**

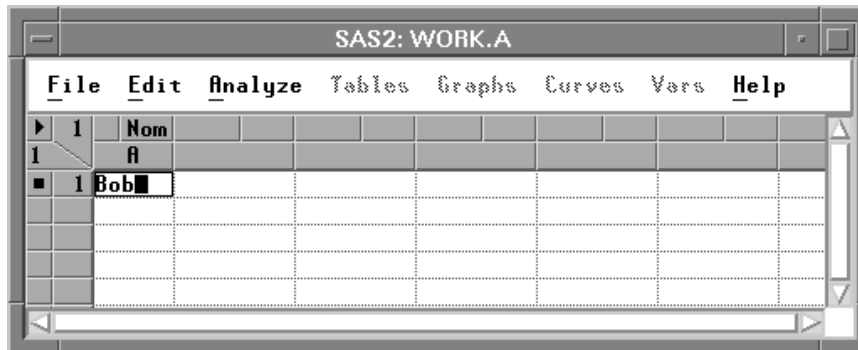


Figure 2.7. Entering a Value

As you type, variables and observations are created for you. The count of variables and observations is shown in the upper left of the data window.

⇒ **Press the Tab key.**

This moves the active value one position to the right.

⇒ **Enter the salary “200000” in the active value.**

Again, a variable is created.

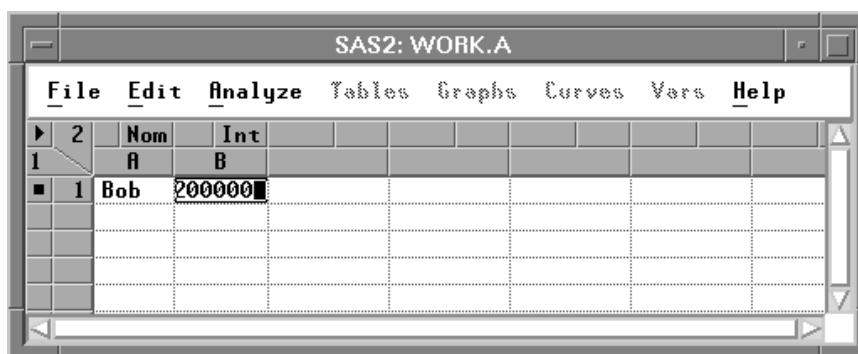
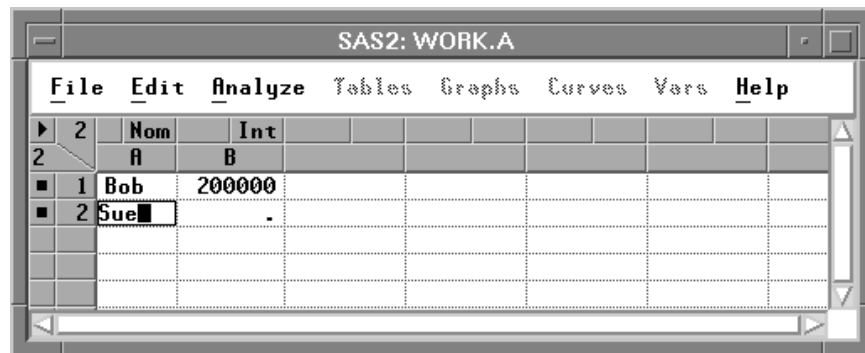


Figure 2.8. A Second Value

⇒ **Press the down arrow key, then press the left arrow key.**

This moves the active value to the first column of the second row.

⇒ **Enter the name “Sue” in the active value.**



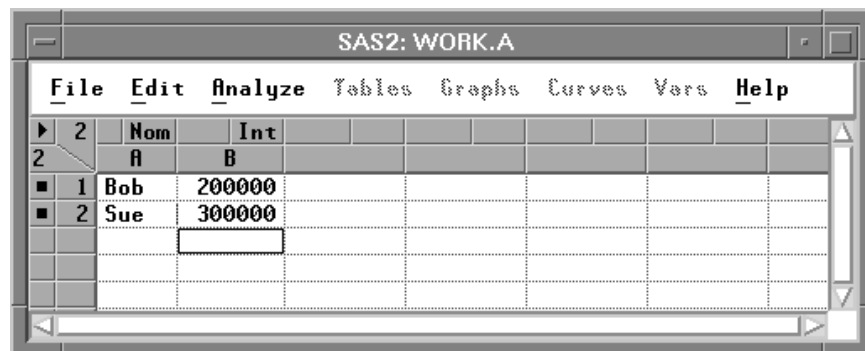
		Nom	Int							
		A	B							
1	Bob	200000								
2	Sue	.								

Figure 2.9. A New Observation

A new observation is created, increasing the observations count to 2. The period (.) in the second value indicates a *missing* value for the numeric variable.

⇒ Press the Tab key to move to the right.

⇒ Enter the salary “300000” to replace the missing value. Then press the down arrow key.



		Nom	Int							
		A	B							
1	Bob	200000								
2	Sue	300000								

Figure 2.10. Replacing the Missing Value

Navigating the Data Window

You can use Tab, BackTab, Enter, Return, and arrow keys to navigate the data window. Tab moves the active value to the right. BackTab, usually defined as Shift-Tab, moves the active value to the left. Enter or Return moves the active value down. Up and down arrow keys move the active value up or down.

When you are not editing any value, left and right arrow keys move the active value left and right. When you are editing a value, left and right arrow keys move the cursor within the active value.

When you have values, variables, or observations selected, the Tab, BackTab, and Return keys navigate within the selected area. This reduces keystrokes when you enter data.

⇒ **Drag a rectangle through several values to select them.**

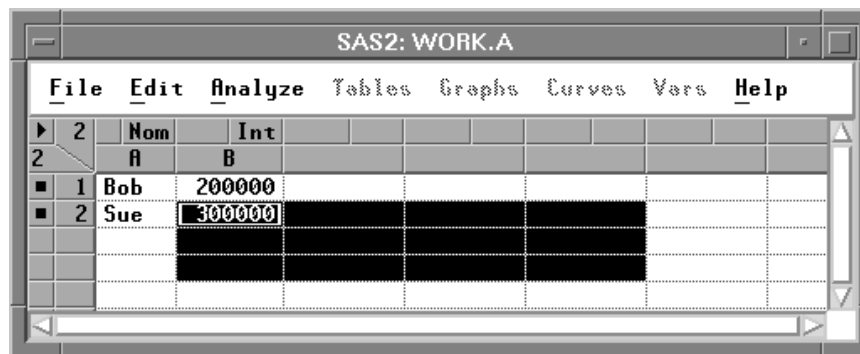


Figure 2.11. Selected Range

⇒ **Press Tab repeatedly.**

⇒ **Press Return repeatedly.**

The active value moves within the range you selected. By default, the Tab key navigates horizontally, and the Return key navigates vertically.

† **Note:** See the section “Data Options” at the end of this chapter for information on defining the direction of Tab and Enter keys.

Adding Variables and Observations

When you have a lot of data to enter, it is more efficient to specify the approximate number of observations rather than to create them one at a time.

⇒ **Click in the upper left corner of the data window.**

This displays the data pop-up menu.

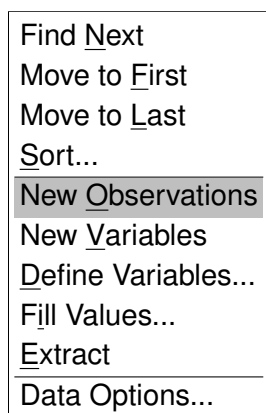


Figure 2.12. Data Pop-up Menu

⇒ **Choose New Observations from the pop-up menu.**

This displays a dialog to prompt you for the number of observations to create.

⇒ **Enter “10” in the observations dialog, then click OK.**

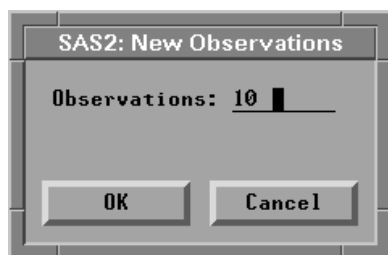
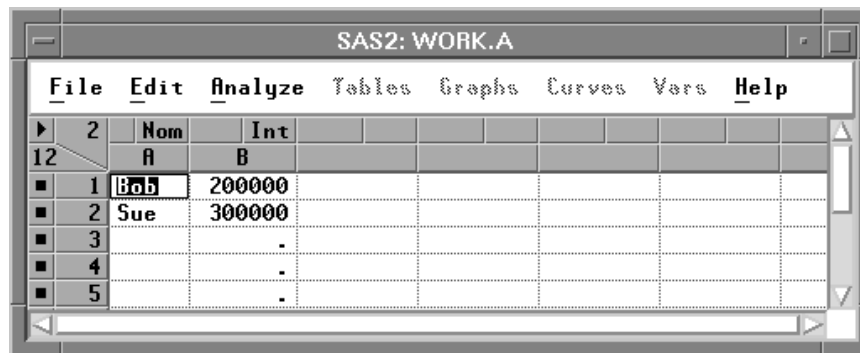


Figure 2.13. Observations Dialog

Observations with missing values are added at the bottom of the data window, increasing the observations count to 12. In the new observations, character values default to blank, while numeric values default to missing.



	2	Nom	Int								
12		A	B								
1	Bob	200000									
2	Sue	300000									
3			.								
4			.								
5			.								

Figure 2.14. New Observations

The **New Variables** menu works like the **New Observations** menu. You can choose **New Variables** to create several variables at once.

Defining Variables

Each variable has a *measurement level* shown in the upper right portion of the column header. By default, numeric values are assigned an *interval* (**Int**) measurement level, indicating values that vary across a continuous range. Character values default to a *nominal* (**Nom**) measurement level, indicating a discrete set of values.

⇒ **Click on the Int measurement level indicator for variable B.**

This displays a pop-up menu.

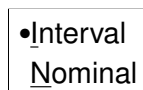


Figure 2.15. Measurement Levels Menu

The radio mark beside **Interval** shows the current measurement level. Because **B** is a numeric variable, it can have either interval or nominal measurement level.

⇒ **Choose Nominal in the pop-up menu to change B's measurement level.**

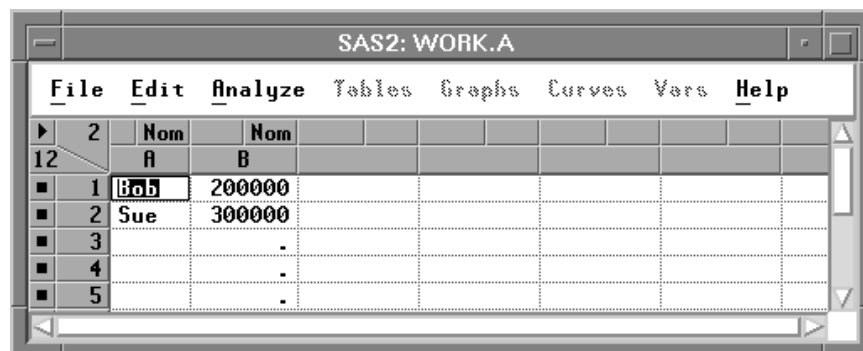


Figure 2.16. Nominal B

You can adjust other variable properties as well. Click in the upper left corner of the data window to display the data pop-up menu.

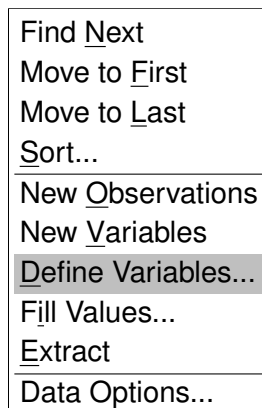


Figure 2.17. Data Pop-up Menu

⇒ Choose **Define Variables** from the pop-up menu.

This displays a dialog. Using this dialog, you can assign variable storage type, measurement level, default roles, name, and label.

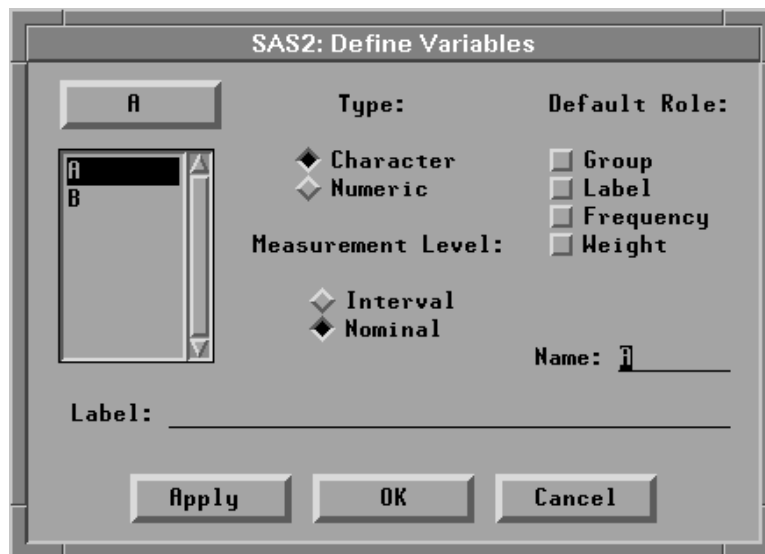


Figure 2.18. Define Variables Dialog

⇒ Enter “NAME” for the name of variable **A**.

⇒ Click the **Apply** button.

In the data window, the variable receives the name you entered.

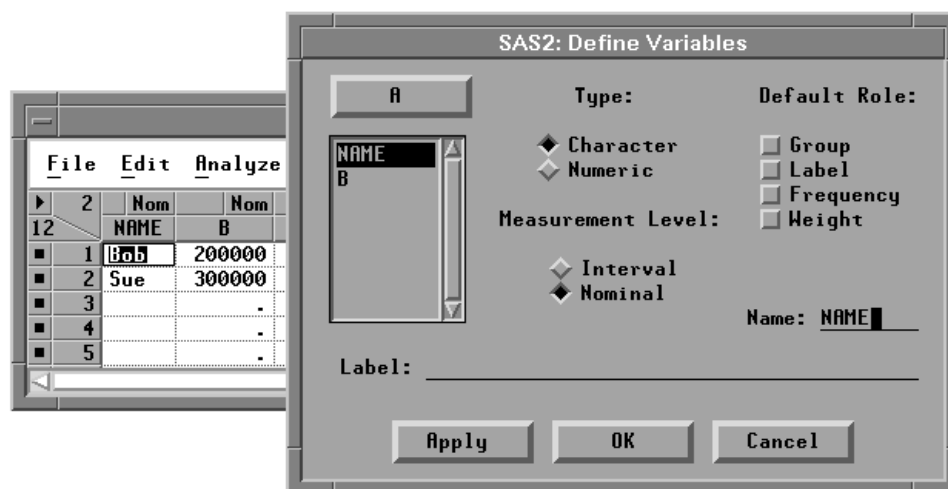


Figure 2.19. Naming a Variable

⇒ Select **B** in the variables list at the left.

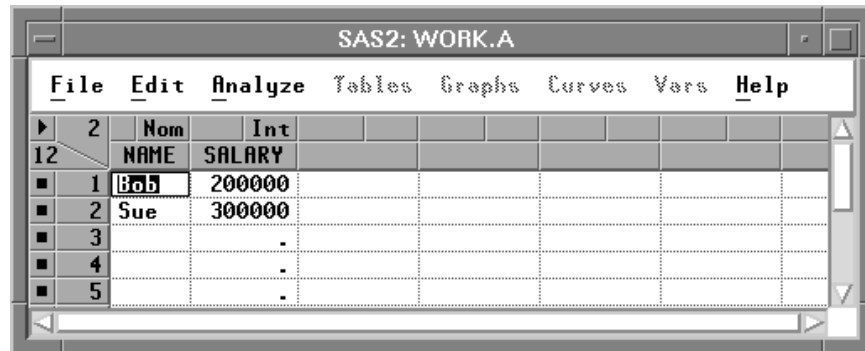
⇒ Enter “SALARY” for the name of variable **B**.

⇒ Click the **Interval** measurement level.

Interval measurement level is appropriate for a variable like salary.

⇒ Click the **OK** button.

This closes the dialog. In the data window, the variable receives the name and measurement level you entered.



	2	Nom	Int										
12		NAME	SALARY										
1		Bob	200000										
2		Sue	300000										
3			.										
4			.										
5			.										

Figure 2.20. Name and Measurement Level Assigned

Fast Data Entry

When you have a lot of data to enter, it is important to be able to do it quickly. Using information from the preceding sections, here is the fastest way to enter data.

⇒ **Open a new data window.**

You can do this when you invoke SAS/INSIGHT software, or you can choose **File:New**.

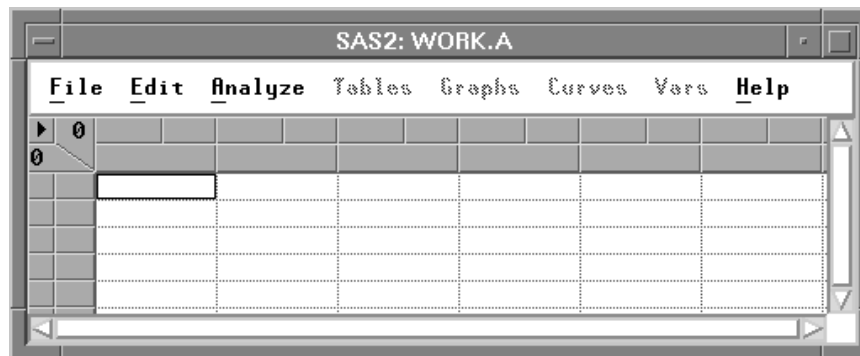


Figure 2.21. New Data Window

⇒ **Create all variables.**

The easiest way to do this is to enter the first observation. Variable types and measurement levels are assigned automatically.

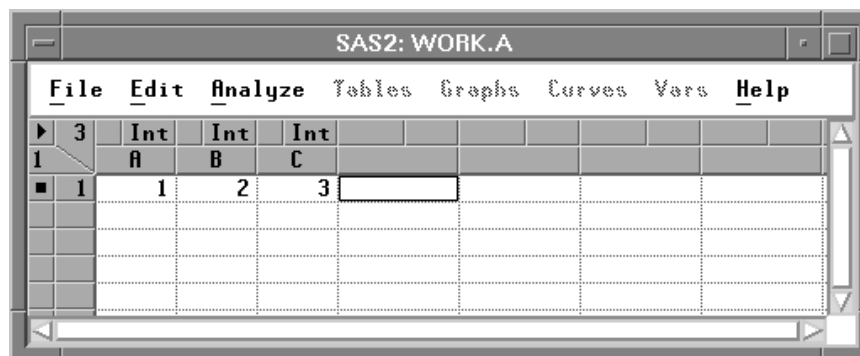


Figure 2.22. Variables Created Automatically

An alternate way to create variables and assign types and measurement levels yourself is by using the data pop-up menu.

⇒ **Click in the upper left corner of the data window.**

This displays the data pop-up menu.

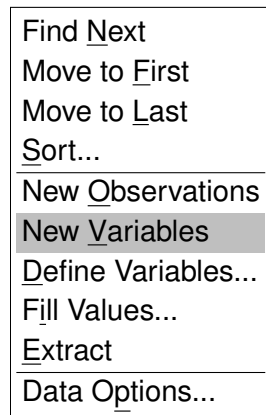


Figure 2.23. Data Pop-up Menu

- ⇒ **Choose New Variables from the pop-up menu.**
This displays a dialog to prompt you for the number of variables to create.
- ⇒ **Enter “3” in the New Variables dialog, then click OK.**

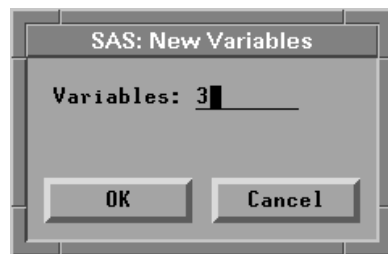


Figure 2.24. New Variables Dialog

The data window should appear as shown in the next figure.

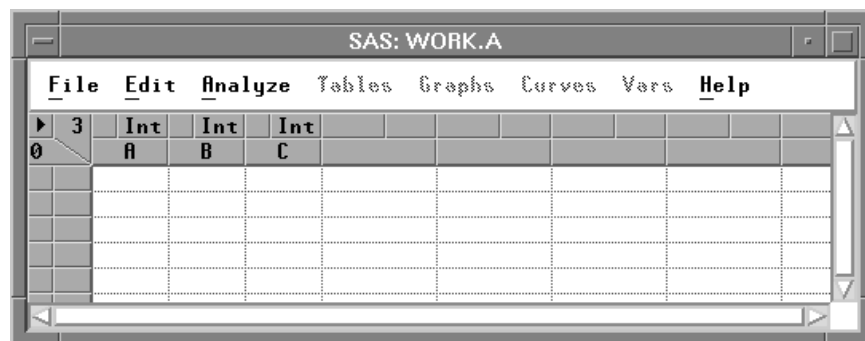


Figure 2.25. Variables Created Manually

The variable names and measurement levels can be selected as shown in the last section.

You can create observations using the following steps.

⇒ **Click in the upper left corner of the data window.**

This displays the data pop-up menu.

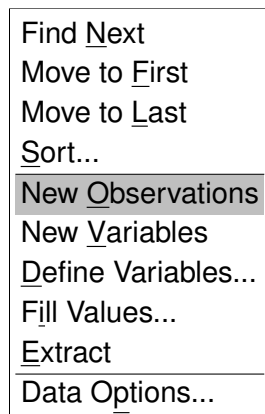


Figure 2.26. Data Pop-up Menu

⇒ **Choose New Observations.**

This displays a dialog prompting you for the number of observations to create.

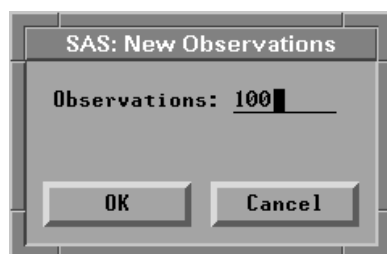


Figure 2.27. Observations Dialog

Enter the number of observations, then click **OK**. If you don't know the number of observations, make it a little larger than you will need. You can delete unused observations later.

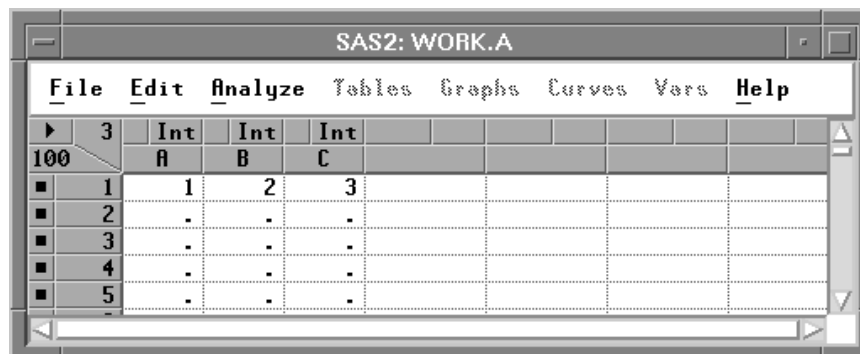


Figure 2.28. Observations Created

⇒ **Select all variables.**

Click the variable count in the upper left corner of the data window.

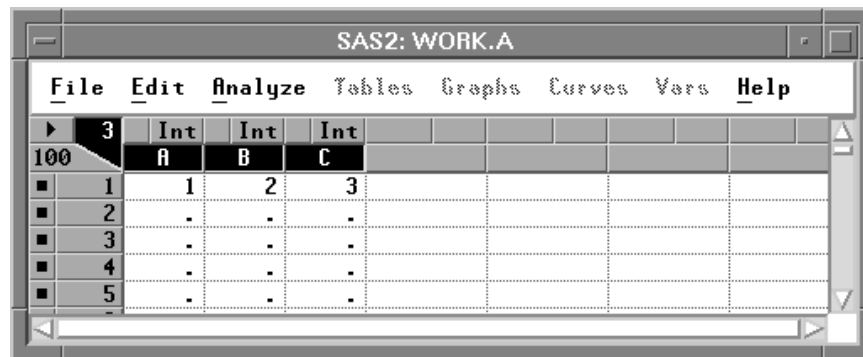


Figure 2.29. Variables Selected

⇒ **Select the active cell.**

Use **Ctrl-click** to avoid deselecting the variables.

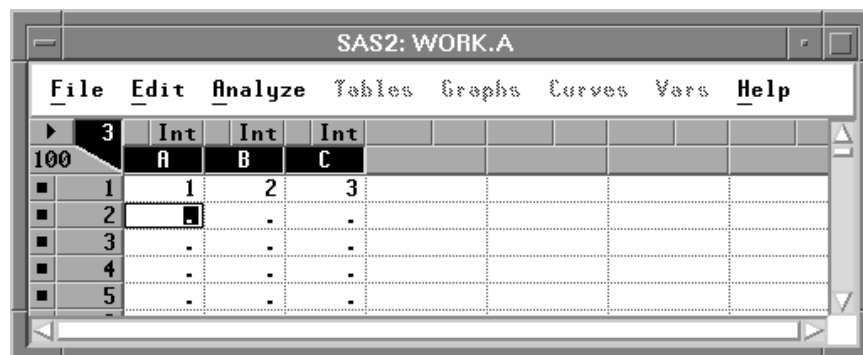


Figure 2.30. Active Value Selected

Now you can enter data, using Tab and BackTab to navigate within the selected variables. You can also fill in blocks of values by using the **Fill Values** option described in the next section. If your keyboard has a numeric keypad, this method enables you to enter numeric data without moving your hand from the keypad.

On some keyboards, the Enter key is easier to hit than the Tab key. So, you may be able to optimize data entry a bit further by defining the direction of the Tab and Enter keys. You can do this by setting the **Data Options** described in the next section. With these options, you can tailor SAS/INSIGHT's data entry to suit your keyboard.

When you have finished entering data, delete any unused observations by selecting them and choosing **Edit:Delete**. If you have not already done so, assign variable names, labels, and other information by choosing **Define Variables**.

Other Options

The pop-up data menu has a couple of useful options for filling in blocks of data and for selecting the actions taken by the Enter and Tab keys.

Click on the button at the upper left corner of the data window to display the data pop-up menu. Choose **Fill Values** to modify selected values in the data window. If you have variables, observations, or values selected, you are prompted to specify a **Value** and an **Increment**. If you have no selections, you are prompted to specify variables and observations.



Figure 2.31. Fill Values Dialog

In the **Fill Values** dialog, the **Value** field can be either character or numeric. If the value is numeric, you can use the **Increment** field to specify an increment or step value. For example, to fill 10 values with ordinals 1 through 10, you can select the values, choose **Fill Values**, and enter 1 for both **Value** and **Increment**.

Choose **Data Options** in the data pop-up menu to set options that control the appearance and operation of the data window. This displays the **Data Options** dialog,

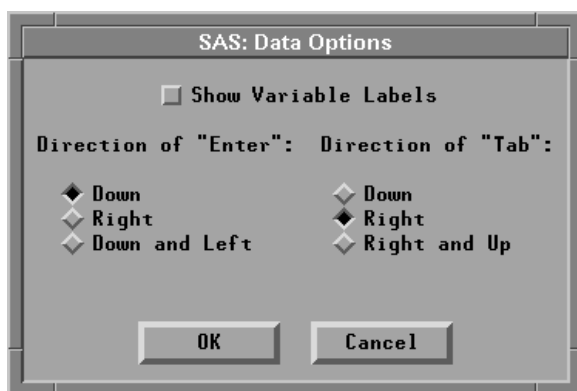


Figure 2.32. Data Options

The dialog contains the following options:

Show Variable Labels

This option controls whether variable labels are displayed. The default is off. If you turn on this option, variable labels are displayed.

Direction of “Enter”

This option controls the interpretation of the Enter and Return keys in the data window. By default, the Enter key moves the active value one position down. If you choose **Right**, the Enter key moves one position to the right. If you choose **Down and Left**, the Enter key moves one position down, and left to the first position.

Direction of “Tab”

This option controls the interpretation of the Tab and BackTab keys in the data window. By default, the Tab key moves the active value one position to the right. If you choose **Down**, the Tab key moves one position down. If you choose **Right and Up**, the Tab key moves one position to the right, and up to the first position.

The options **Down and Left** and **Right and Up** were added in Release 6.11. Not all hosts define a BackTab key, and not all hosts define Enter and Return as the same key. Consult your host documentation for information on key definitions.

You can save data window options by choosing **File:Save:Options**. This enables you to use your preferred option settings as defaults in future SAS/INSIGHT sessions.

Chapter 3

Examining Data

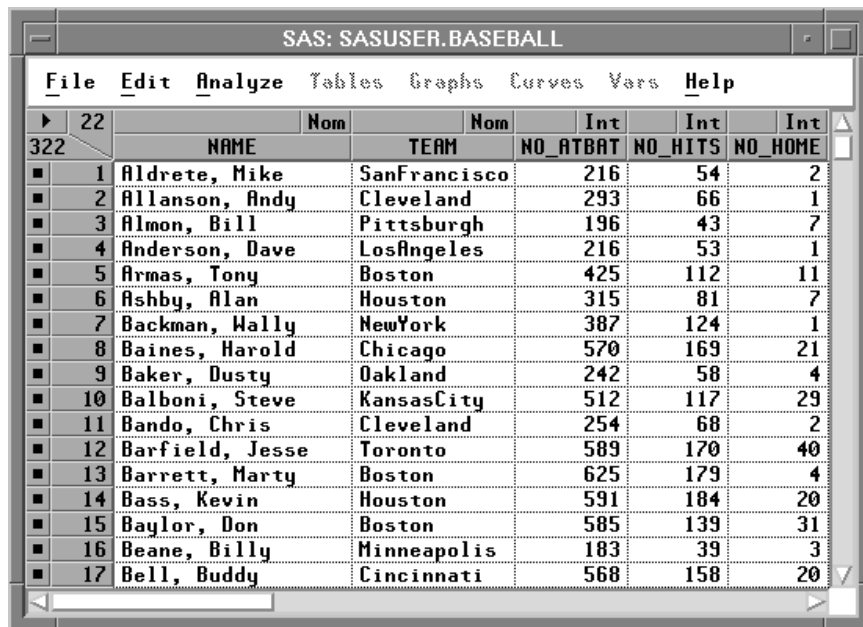
Chapter Contents

INVOKING SAS/INSIGHT SOFTWARE	50
SCROLLING THE DATA WINDOW	51
ARRANGING VARIABLES	52
SORTING OBSERVATIONS	56
FINDING OBSERVATIONS	59
EXAMINING OBSERVATIONS	63
CLOSING THE DATA WINDOW	67

Chapter 3

Examining Data

SAS/INSIGHT software displays your data as a table of rows and columns in which the rows represent observations and the columns represent variables. You can use SAS/INSIGHT software to view your data, arrange variables, sort observations, and find and examine observations of interest.



The screenshot shows the SAS Data Window titled "SAS: SASUSER.BASEBALL". The window contains a menu bar with "File", "Edit", "Analyze", "Tables", "Graphs", "Curves", "Vars", and "Help". Below the menu bar is a table with 6 columns: "NAME", "TEAM", "NO_ATBAT", "NO_HITS", and "NO_HOME". The table lists 17 players, numbered 1 through 17. The first column of the table is labeled "22" and "322".

	Nom	Nom	Int	Int	Int
	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME
1	Aldrete, Mike	SanFrancisco	216	54	2
2	Allanson, Andy	Cleveland	293	66	1
3	Almon, Bill	Pittsburgh	196	43	7
4	Anderson, Dave	LosAngeles	216	53	1
5	Armas, Tony	Boston	425	112	11
6	Ashby, Alan	Houston	315	81	7
7	Backman, Wally	NewYork	387	124	1
8	Baines, Harold	Chicago	570	169	21
9	Baker, Dusty	Oakland	242	58	4
10	Balboni, Steve	KansasCity	512	117	29
11	Bando, Chris	Cleveland	254	68	2
12	Barfield, Jesse	Toronto	589	170	40
13	Barrett, Marty	Boston	625	179	4
14	Bass, Kevin	Houston	591	184	20
15	Baylor, Don	Boston	585	139	31
16	Beane, Billy	Minneapolis	183	39	3
17	Bell, Buddy	Cincinnati	568	158	20

Figure 3.1. Data Window

Invoking SAS/INSIGHT Software

Using one of the methods mentioned in Chapter 2, “Entering Data,” invoke SAS/INSIGHT software to display the data set dialog. ⇒ **In the dialog, point and click to choose a library and data set.**

A *library* is a location where data sets are stored. Point to the list on the left and click on any library to see a list of data sets stored there. Point to the list on the right and click on any data set to select it for opening. Then click on **Open** to open a window on the data.



Figure 3.2. Data Set Dialog

As a shortcut, you can click twice rapidly on the data set (a *double-click*) instead of clicking once on the data set and once on the **Open** button.

	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME
1	Aldrete, Mike	SanFrancisco	216	54	2
2	Allanson, Andy	Cleveland	293	66	1
3	Almon, Bill	Pittsburgh	196	43	7
4	Anderson, Dave	LosAngeles	216	53	1
5	Armas, Tony	Boston	425	112	11
6	Ashby, Alan	Houston	315	81	7
7	Backman, Wally	NewYork	387	124	1
8	Baines, Harold	Chicago	570	169	21
9	Baker, Dusty	Oakland	242	58	4
10	Balboni, Steve	KansasCity	512	117	29
11	Bando, Chris	Cleveland	254	68	2
12	Barfield, Jesse	Toronto	589	170	40

Figure 3.3. Data Window

Each variable in SAS/INSIGHT software has a *measurement level* that determines the way it is treated in graphs and analyses. The measurement level for each variable appears above the variable name. You can assign two measurement levels: *interval* and *nominal*.

Interval variables contain values that vary across a continuous range. For example, **NO_ATBAT** is an interval variable in Figure 3.3.

Nominal variables contain a discrete set of values. For example, **NAME** is a nominal variable in Figure 3.3.

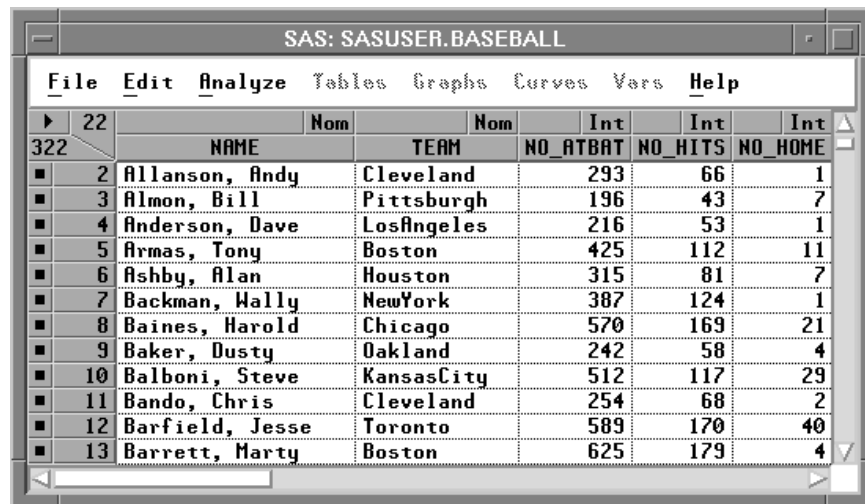
Each observation in SAS/INSIGHT software has a *marker*, a graphic shape that identifies the observation in graphs. The marker for each observation appears to the left of the observation number.

The number of observations and the number of variables in the data set appear in the upper left corner of the data window. The data window in Figure 3.3 shows that **SASUSER.BASEBALL** has 322 observations and 22 variables.

Scrolling the Data Window

Most data sets are too large to fit in a data window, so the window contains *scroll bars* to scroll the data through the window. The appearance of scroll bars varies depending on your host. Most scroll bars have small *arrow buttons* at the ends and a *slider* between the buttons to indicate the current position and relative size of the displayed area.

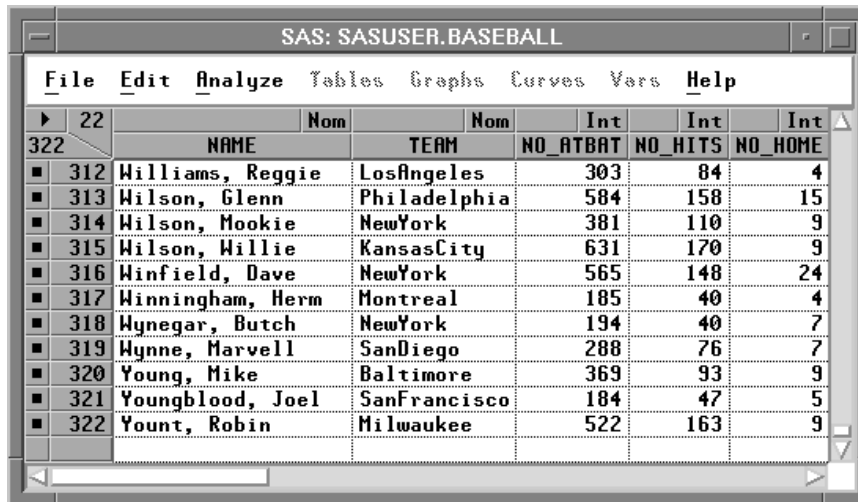
- ⇒ Click the arrow button at the bottom of the vertical scroll bar.
This scrolls down one observation.



		Nom	Nom	Int	Int	Int
	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME	
2	Allanson, Andy	Cleveland	293	66	1	
3	Almon, Bill	Pittsburgh	196	43	7	
4	Anderson, Dave	LosAngeles	216	53	1	
5	Armas, Tony	Boston	425	112	11	
6	Ashby, Alan	Houston	315	81	7	
7	Backman, Wally	NewYork	387	124	1	
8	Baines, Harold	Chicago	570	169	21	
9	Baker, Dusty	Oakland	242	58	4	
10	Balboni, Steve	KansasCity	512	117	29	
11	Bando, Chris	Cleveland	254	68	2	
12	Barfield, Jesse	Toronto	589	170	40	
13	Barrett, Marty	Boston	625	179	4	

Figure 3.4. Scrolling Down One Observation

- ⇒ **Drag the slider on the vertical scroll bar all the way down.**
This scrolls to the last observation.



	Nom	Nom	Int	Int	Int
	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME
312	Williams, Reggie	LosAngeles	303	84	4
313	Wilson, Glenn	Philadelphia	584	158	15
314	Wilson, Mookie	NewYork	381	110	9
315	Wilson, Willie	KansasCity	631	170	9
316	Winfield, Dave	NewYork	565	148	24
317	Winningham, Herm	Montreal	185	40	4
318	Wynegar, Butch	NewYork	194	40	7
319	Wynne, Marvell	SanDiego	288	76	7
320	Young, Mike	Baltimore	369	93	9
321	Youngblood, Joel	SanFrancisco	184	47	5
322	Yount, Robin	Milwaukee	522	163	9

Figure 3.5. Scrolling to the Last Observation

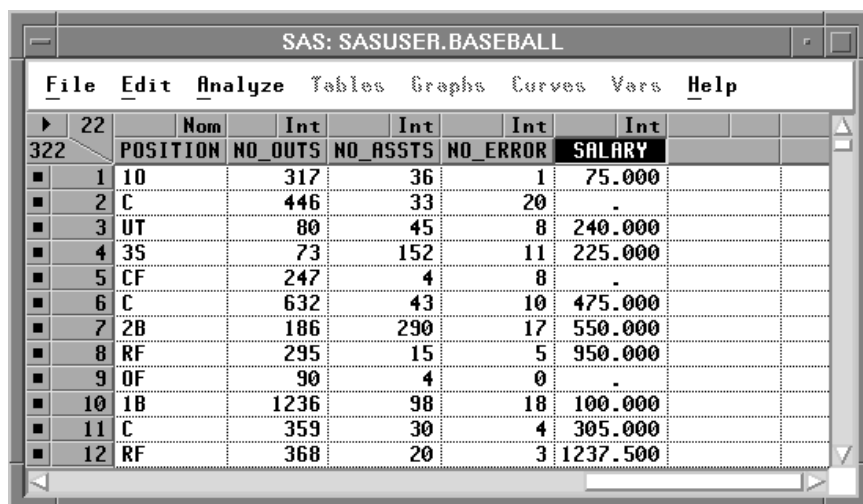
Similarly, clicking the arrow button at the top of the vertical scroll bar scrolls up one observation, and dragging the slider all the way to the top scrolls to the first observation. The horizontal scroll bar works the same way, except that it moves the data by variable instead of by observation.

† **Note:** On many hosts you can click *within* the scroll bar to scroll the width or height of the window. Some hosts offer additional buttons on the scroll bars, and some hosts respond to more than one button on the mouse. Refer to your host documentation for details and experiment by clicking on the scroll bars in the data window.

Arranging Variables

Using scroll bars, you can view all of your data, but the variables and observations may not always be arranged as you would like. For example, suppose you are interested in the salaries of the players in the data set **SASUSER.BASEBALL**. To move the **SALARY** variable to the first position in the data window, follow these steps.

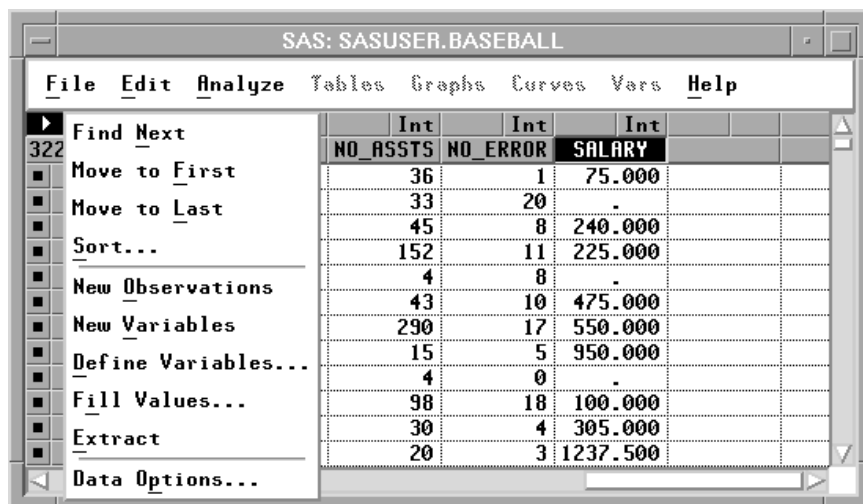
- ⇒ **Scroll the data window to the **SALARY** variable.**
SALARY is the last variable, so drag the slider on the horizontal scroll bar all the way to the right.
- ⇒ **Point to the **SALARY** variable name.**
Then click with the mouse to select the variable **SALARY**. The variable becomes highlighted when you select it.



		Nom	Int	Int	Int	Int
	POSITION	NO_OUTS	NO_ASSTS	NO_ERROR	SALARY	
1	10	317	36	1	75.000	
2	C	446	33	20	.	
3	UT	80	45	8	240.000	
4	3S	73	152	11	225.000	
5	CF	247	4	8	.	
6	C	632	43	10	475.000	
7	2B	186	290	17	550.000	
8	RF	295	15	5	950.000	
9	OF	90	4	0	.	
10	1B	1236	98	18	100.000	
11	C	359	30	4	305.000	
12	RF	368	20	3	1237.500	

Figure 3.6. Selecting the Last Variable

⇒ Click on the menu button in the upper left corner.
This opens the data pop-up menu. Click on **Move to First**.



	Int	Int	Int
	NO_ASSTS	NO_ERROR	SALARY
36	1	75.000	
33	20	.	
45	8	240.000	
152	11	225.000	
4	8	.	
43	10	475.000	
290	17	550.000	
15	5	950.000	
4	0	.	
98	18	100.000	
30	4	305.000	
20	3	1237.500	

Figure 3.7. Data Pop-up Menu

This moves the selected variable to the first position. Note that the **Data** menu also has a **Move to Last** choice, so you can easily move variables to the last position.

SAS: SASUSER.BASEBALL									
File		Edit	Analyze	Tables	Graphs	Curves	Vars	Help	
▶	22	Int	Nom		Nom	Int	Int		
322		SALARY	NAME		TEAM	NO_ATBAT	NO_HITS		
■	1	75.000	Aldrete, Mike		SanFrancisco	216	54		
■	2	.	Allanson, Andy		Cleveland	293	60		
■	3	240.000	Almon, Bill		Pittsburgh	196	43		
■	4	225.000	Anderson, Dave		LosAngeles	216	53		
■	5	.	Armas, Tony		Boston	425	112		
■	6	475.000	Ashby, Alan		Houston	315	8		
■	7	550.000	Backman, Wally		NewYork	387	124		
■	8	950.000	Baines, Harold		Chicago	570	169		
■	9	.	Baker, Dusty		Oakland	242	58		
■	10	100.000	Balboni, Steve		KansasCity	512	112		
■	11	305.000	Bando, Chris		Cleveland	254	60		
■	12	1237.500	Barfield, Jesse		Toronto	589	170		

Figure 3.8. Variable in First Position

You can also move individual variables to different locations by using the *hand tool*.

⇒ Choose **Edit:Windows:Tools**.

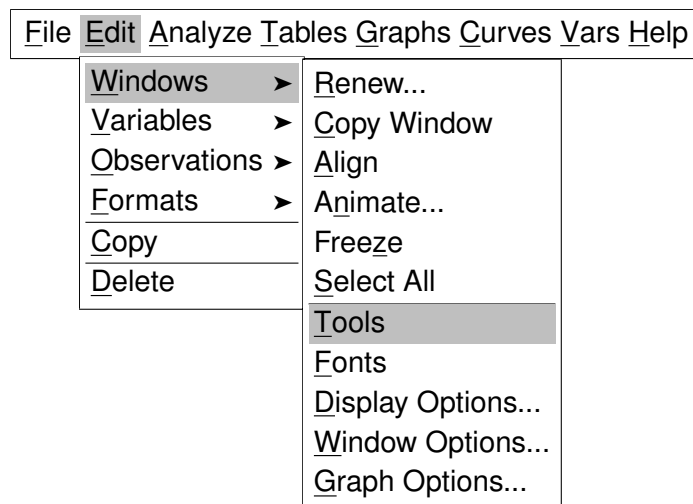


Figure 3.9. Edit:Windows Menu

The tools window is shown in the next figure.

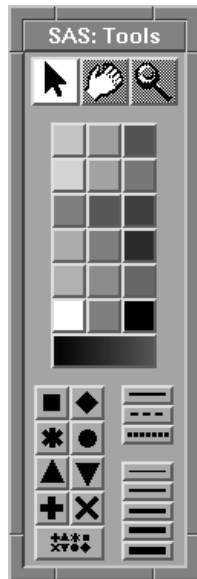


Figure 3.10. Tools Window

- ⇒ **Click the Hand tool at the top of the Tools window.**
The cursor changes to a hand. Move the hand to the variable named **Salary**.
- ⇒ **Press the left mouse button and hold it down.**
A dotted rectangle should appear as the outline of the variable column.
- ⇒ **Drag the rectangle so that its middle is on the border between Name and Team.**
- ⇒ **Release the left mouse button.**
The **Salary** variable has become the second variable in the data window.

SAS: SASUSER.BASEBALL

File Edit Analyze Tables Graphs Curves Vars Help

		Nom	Int	Nom	Int	Int
	NAME	SALARY	TEAM	NO_ATBAT	NO_HI	
1	Aldrete, Mike	75.000	SanFrancisco	216		
2	Allanson, Andy	.	Cleveland	293		
3	Almon, Bill	240.000	Pittsburgh	196		
4	Anderson, Dave	225.000	LosAngeles	216		
5	Armas, Tony	.	Boston	425		
6	Ashby, Alan	475.000	Houston	315		
7	Backman, Wally	550.000	NewYork	387		
8	Baines, Harold	950.000	Chicago	570		
9	Baker, Dusty	.	Oakland	242		
10	Balboni, Steve	100.000	KansasCity	512		
11	Bando, Chris	305.000	Cleveland	254		
12	Barfield, Jesse	1237.500	Toronto	589		
13	Barrett, Marty	575.000	Boston	625		
14	Bass, Kevin	630.000	Houston	591		
15	Baylor, Don	950.000	Boston	585		
16	Beane, Billy	.	Minneapolis	183		
17	Bell, Buddu	725.000	Cincinnati	568		

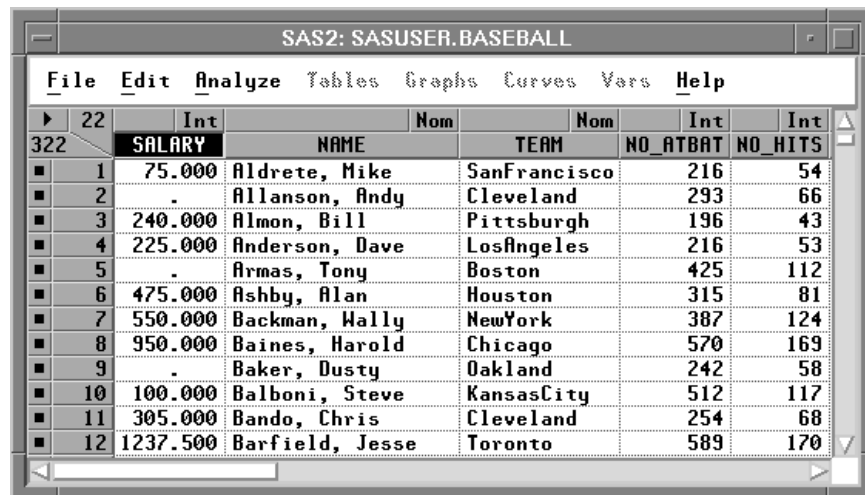
Figure 3.11. Variable in Second Position

- ⇒ Use the Hand tool to move **SALARY** back to the first position.
- ⇒ Click the arrow tool in the Tools window to restore the cursor.

Sorting Observations

It is often useful to examine data ordered by the values of a variable. Suppose you want to sort the baseball data by players' salaries stored in the **SALARY** variable. Follow these steps.

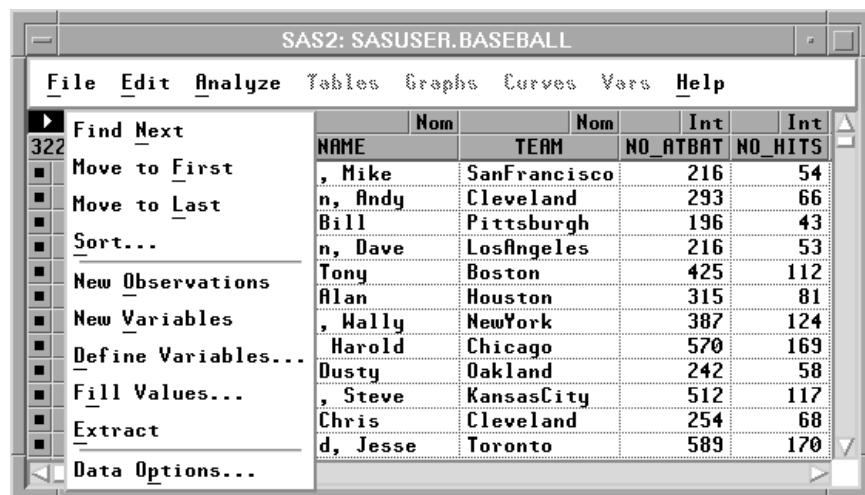
- ⇒ Point and click to select the **SALARY** variable.



	22	Int	Nom	Nom	Int	Int
322	SALARY	NAME	TEAM	NO_ATBAT	NO_HITS	
1	75.000	Aldrete, Mike	SanFrancisco	216	54	
2	.	Allanson, Andy	Cleveland	293	66	
3	240.000	Almon, Bill	Pittsburgh	196	43	
4	225.000	Anderson, Dave	LosAngeles	216	53	
5	.	Armas, Tony	Boston	425	112	
6	475.000	Ashby, Alan	Houston	315	81	
7	550.000	Backman, Wally	NewYork	387	124	
8	950.000	Baines, Harold	Chicago	570	169	
9	.	Baker, Dusty	Oakland	242	58	
10	100.000	Balboni, Steve	KansasCity	512	117	
11	305.000	Bando, Chris	Cleveland	254	68	
12	1237.500	Barfield, Jesse	Toronto	589	170	

Figure 3.12. Selecting a Variable

⇒ Click on the menu button in the upper left corner.
This opens the data pop-up menu. Click on **Sort**.



	Nom	Nom	Int	Int
NAME	TEAM	NO_ATBAT	NO_HITS	
, Mike	SanFrancisco	216	54	
n, Andy	Cleveland	293	66	
Bill	Pittsburgh	196	43	
n, Dave	LosAngeles	216	53	
Tony	Boston	425	112	
Alan	Houston	315	81	
, Wally	NewYork	387	124	
Harold	Chicago	570	169	
Dusty	Oakland	242	58	
, Steve	KansasCity	512	117	
Chris	Cleveland	254	68	
d, Jesse	Toronto	589	170	

Figure 3.13. Sorting Observations

The data are now sorted by **SALARY** in ascending order.

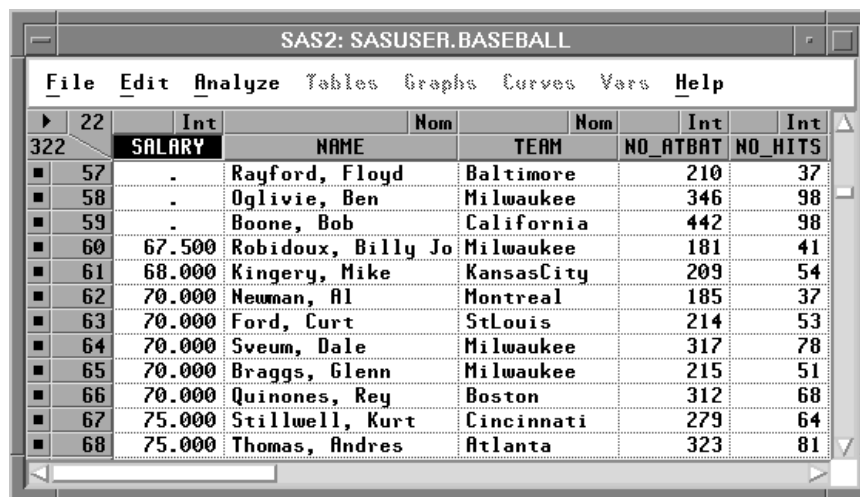


The screenshot shows the SAS2: SASUSER.BASEBALL window with a menu bar (File, Edit, Analyze, Tables, Graphs, Curves, Vars, Help) and a data table. The table is sorted by the SALARY variable. The first 12 rows are visible, showing missing values for salary followed by specific salary values.

	22	Int	Nom	Nom	Int	Int
322	SALARY	NAME	TEAM	NO_ATBAT	NO_HITS	
1	.	Thon, Dickie	Houston	278	69	
2	.	Krenchicki, Hayne	Montreal	221	53	
3	.	Kutcher, Randy	SanFrancisco	186	44	
4	.	Kingman, Dave	Oakland	561	118	
5	.	Cabell, Enos	LosAngeles	277	71	
6	.	Jones, Ruppert	California	393	90	
7	.	Johnson, Cliff	Toronto	336	84	
8	.	Law, Rudy	KansasCity	307	80	
9	.	Lynn, Fred	Baltimore	397	114	
10	.	Brown, Mike	Pittsburgh	243	53	
11	.	Meacham, Bobby	NewYork	161	36	
12	.	Moore, Charlie	Milwaukee	235	61	

Figure 3.14. Sorted Data

The periods (.) displayed in the observations for **SALARY** are *missing values*. Missing values are placeholders that indicate no data are available. Missing values are treated as less than any other value, so when the data are sorted, missing values appear first. If you scroll the data, you can see that the missing values are followed by the smallest salaries.



The screenshot shows the SAS2: SASUSER.BASEBALL window with a menu bar (File, Edit, Analyze, Tables, Graphs, Curves, Vars, Help) and a data table. The table is sorted by the SALARY variable. The first 12 rows are visible, showing missing values for salary followed by specific salary values.

	22	Int	Nom	Nom	Int	Int
322	SALARY	NAME	TEAM	NO_ATBAT	NO_HITS	
57	.	Rayford, Floyd	Baltimore	210	37	
58	.	Oglivie, Ben	Milwaukee	346	98	
59	.	Boone, Bob	California	442	98	
60	67.500	Robidoux, Billy Jo	Milwaukee	181	41	
61	68.000	Kingery, Mike	KansasCity	209	54	
62	70.000	Neuman, Al	Montreal	185	37	
63	70.000	Ford, Curt	StLouis	214	53	
64	70.000	Sveum, Dale	Milwaukee	317	78	
65	70.000	Braggs, Glenn	Milwaukee	215	51	
66	70.000	Quinones, Rey	Boston	312	68	
67	75.000	Stillwell, Kurt	Cincinnati	279	64	
68	75.000	Thomas, Andres	Atlanta	323	81	

Figure 3.15. Sorted Data, Missing and Nonmissing

Finding Observations

Sometimes you want to find observations that share some characteristic. For example, you might want to find all the baseball players who primarily played first base. To do so, follow these steps. The figures in this section are based on the **NAME** variable appearing as the first variable. If you just completed the previous two sections on moving variables and sorting observations, move the **SALARY** variable to the last position and sort the observations on **NAME**. Make sure no variables are selected.

⇒ Choose **Edit:Observations:Find**.

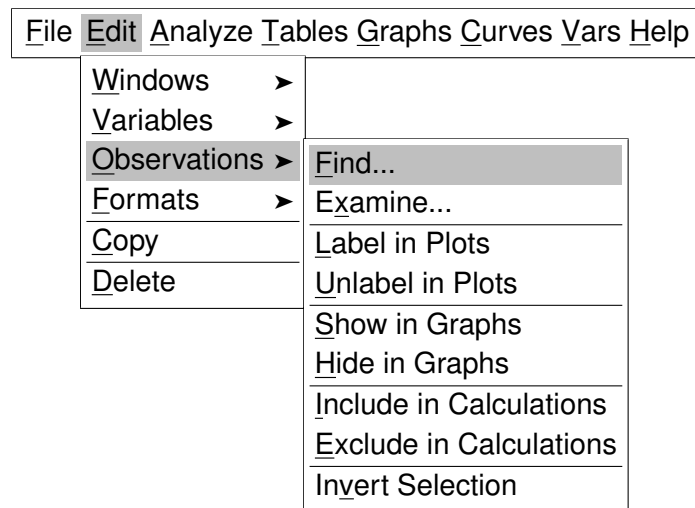


Figure 3.16. Finding Observations

This displays the **Find Observations** dialog.



Figure 3.17. Find Observations Dialog

⇒ Select the **POSITION** variable.

Scroll the list of variables at the left to see the **POSITION** variable. Then point and click to select **POSITION**. Notice that the list of values at the right now contains all the unique values of the **POSITION** variable. By default, the equal (=) test and the first value are selected.



Figure 3.18. Selecting **POSITION**

⇒ Select the values **13**, **1B**, and **10**.

On most hosts, you can either **Shift**-click or **CTRL**-click to select these values. The players selected primarily played first base. Note that players with **POSITION = 01** also played some first base, but they played primarily in the outfield.

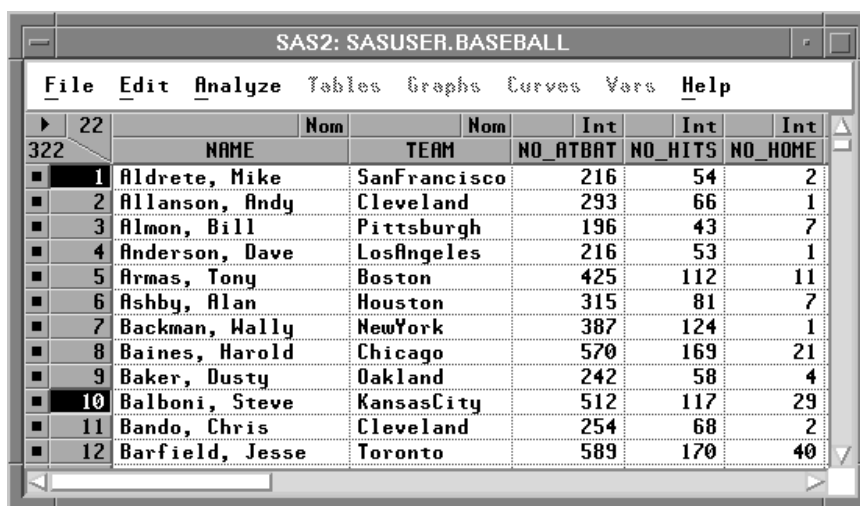
⇒ Click the **Apply** button to find the data.

This selects observations without closing the **Find Observations** dialog. Clicking the **OK** button closes the **Find Observations** dialog after selecting the observations.



Figure 3.19. Selecting First Basemen

Now all observations where **POSITION** is **13**, **1B**, or **1O** are highlighted.

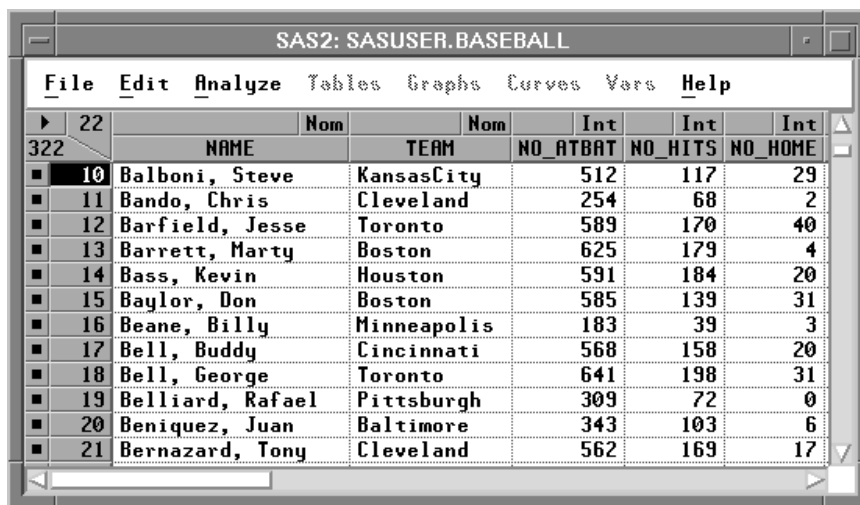


		Nom	Nom	Int	Int	Int
	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME	
1	Aldrete, Mike	SanFrancisco	216	54	2	
2	Allanson, Andy	Cleveland	293	66	1	
3	Almon, Bill	Pittsburgh	196	43	7	
4	Anderson, Dave	LosAngeles	216	53	1	
5	Armas, Tony	Boston	425	112	11	
6	Ashby, Alan	Houston	315	81	7	
7	Backman, Wally	NewYork	387	124	1	
8	Baines, Harold	Chicago	570	169	21	
9	Baker, Dusty	Oakland	242	58	4	
10	Balboni, Steve	KansasCity	512	117	29	
11	Bando, Chris	Cleveland	254	68	2	
12	Barfield, Jesse	Toronto	589	170	40	

Figure 3.20. First Basemen Found

⇒ Choose **Find Next** from the data pop-up menu.

The data window scrolls so the next observation with **POSITION** = **13**, **1B**, or **1O** is at the top.



		Nom	Nom	Int	Int	Int
	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME	
10	Balboni, Steve	KansasCity	512	117	29	
11	Bando, Chris	Cleveland	254	68	2	
12	Barfield, Jesse	Toronto	589	170	40	
13	Barrett, Marty	Boston	625	179	4	
14	Bass, Kevin	Houston	591	184	20	
15	Baylor, Don	Boston	585	139	31	
16	Beane, Billy	Minneapolis	183	39	3	
17	Bell, Buddy	Cincinnati	568	158	20	
18	Bell, George	Toronto	641	198	31	
19	Belliard, Rafael	Pittsburgh	309	72	0	
20	Beniquez, Juan	Baltimore	343	103	6	
21	Bernazard, Tony	Cleveland	562	169	17	

Figure 3.21. Finding the Next Observation

⇒ Choose **Move to First** from the data pop-up menu.

This enables you to see all the selected observations in one place, in this case at the top of the data window.

	22	Nom	Nom	Int	Int	Int
322	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOMI	
<input checked="" type="checkbox"/>	1	Aldrete, Mike	SanFrancisco	216	54	2
<input checked="" type="checkbox"/>	2	Balboni, Steve	KansasCity	512	117	29
<input checked="" type="checkbox"/>	3	Bochte, Bruce	Oakland	407	104	0
<input checked="" type="checkbox"/>	4	Bream, Sid	Pittsburgh	522	140	10
<input checked="" type="checkbox"/>	5	Brock, Greg	LosAngeles	325	76	10
<input checked="" type="checkbox"/>	6	Buckner, Bill	Boston	629	168	18
<input checked="" type="checkbox"/>	7	Cabell, Enos	LosAngeles	277	71	2
<input checked="" type="checkbox"/>	8	Clark, Jack	StLouis	232	55	9
<input checked="" type="checkbox"/>	9	Clark, Will	SanFrancisco	408	117	13
<input checked="" type="checkbox"/>	10	Cooper, Cecil	Milwaukee	542	140	12
<input checked="" type="checkbox"/>	11	Davis, Alan	Seattle	479	130	18
<input checked="" type="checkbox"/>	12	Davis, Glenn	Houston	574	152	31
<input checked="" type="checkbox"/>	13	Durham, Leon	Chicago	484	127	20
<input checked="" type="checkbox"/>	14	Esasky, Nick	Cincinnati	330	76	12
<input checked="" type="checkbox"/>	15	Evans, Darrell	Detroit	507	122	29
<input checked="" type="checkbox"/>	16	Galarraga, Andres	Montreal	321	87	10
<input checked="" type="checkbox"/>	17	Garvey, Steve	SanDiego	557	142	23

Figure 3.22. Collecting the Selected Observations

Examining Observations

You can examine selected observations in detail by following these steps. The figures in this section are based on the data being sorted on the **NAME** variable and the observations selected where **POSITION** is **13**, **1B**, or **1O**. The previous sections on sorting and finding observations provide examples of how to sort and select.

⇒ Choose **Edit:Observations:Examine**.

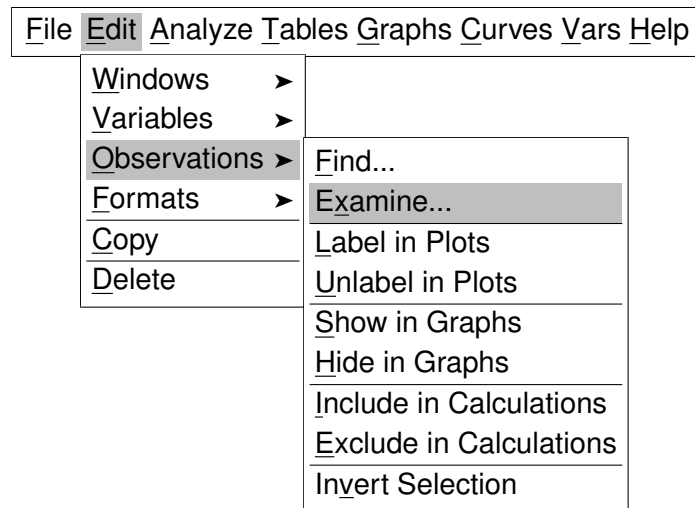


Figure 3.23. Finding Observations

This displays the **Examine Observations** dialog. The list on the left shows the observation number for the selected observations: first basemen. The list on the right displays the variable values for the highlighted observation.

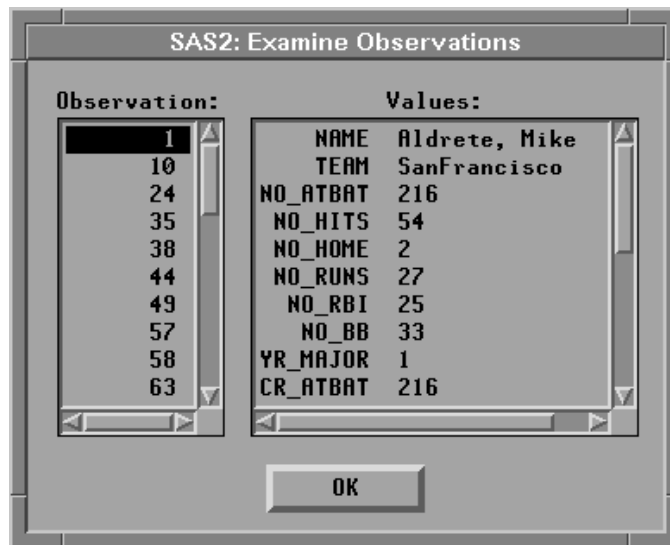


Figure 3.24. Examine Observations Dialog

Scroll down the list on the right to see the rest of Mike Aldrete's statistics. Point and click on observation number **58** to see Will Clark's statistics. Scroll down the list on the left until you can point and click on observation number **246** to see Pete Rose's statistics. Click **OK** to close the dialog.

You can also use the **Examine Observations** dialog directly from a graph or chart. To examine observations from a box plot of player salaries, follow these steps.

⇒ **Choose Analyze:Box Plot/Mosaic Plot (Y).**

This calls up the **Box Plot/Mosaic Plot** dialog.

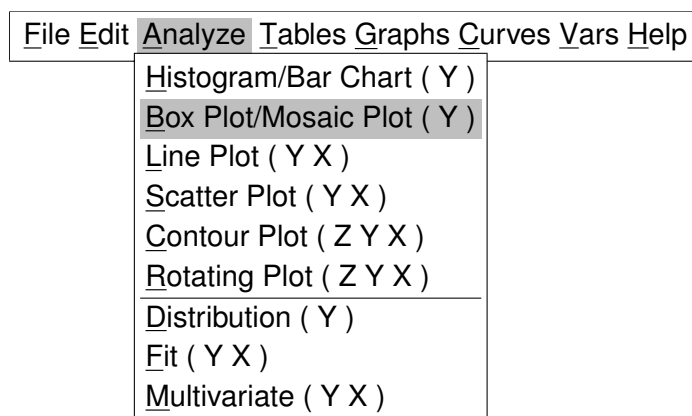


Figure 3.25. Creating a Box Plot

⇒ **Assign SALARY the Y role and LEAGUE the X role.**

Click on **SALARY** in the variable list on the left, then click on **Y** at the top. Similarly, click on **LEAGUE** in the list on the left, then click on **X** at the top.

⇒ **Click OK to create a box plot of SALARY by LEAGUE.**

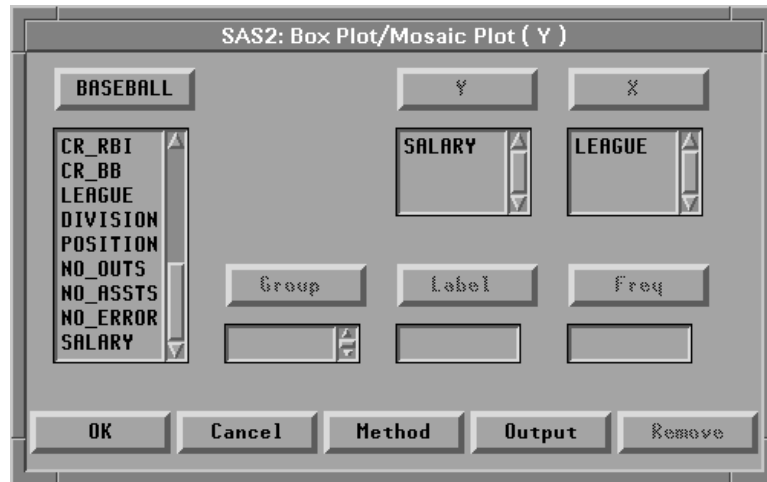


Figure 3.26. Box Plot Variable Roles

⇒ **Double-click on the marker representing the highest salary in the National League.**

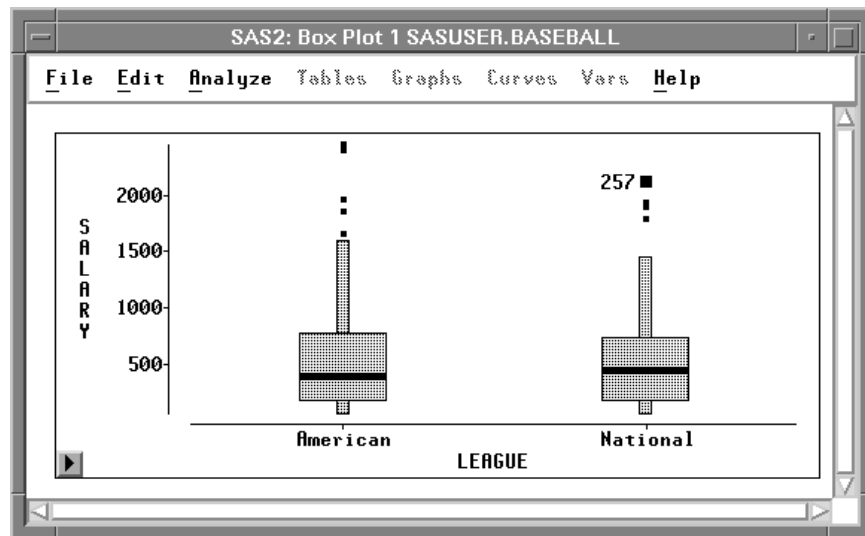


Figure 3.27. Box Plot of **SALARY** by **LEAGUE**

Clicking on the observation identifies the point in the graph with its observation number. Double-clicking displays the **Examine Observations** dialog for the selected observation. In 1986, Mike Schmidt had the highest salary in the National League.

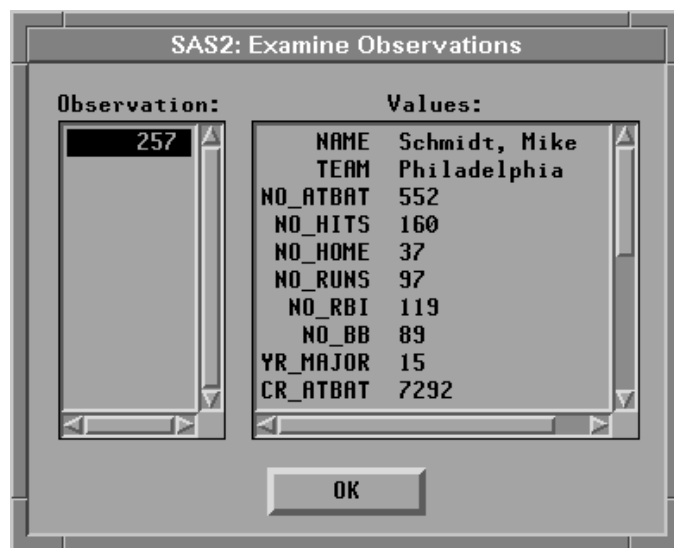


Figure 3.28. Examining Observations

⇒ **Double-click on the upper whisker for the American League.**

This displays the values for all observations within the whisker. Then click in the Observation list to see the values for each observation.

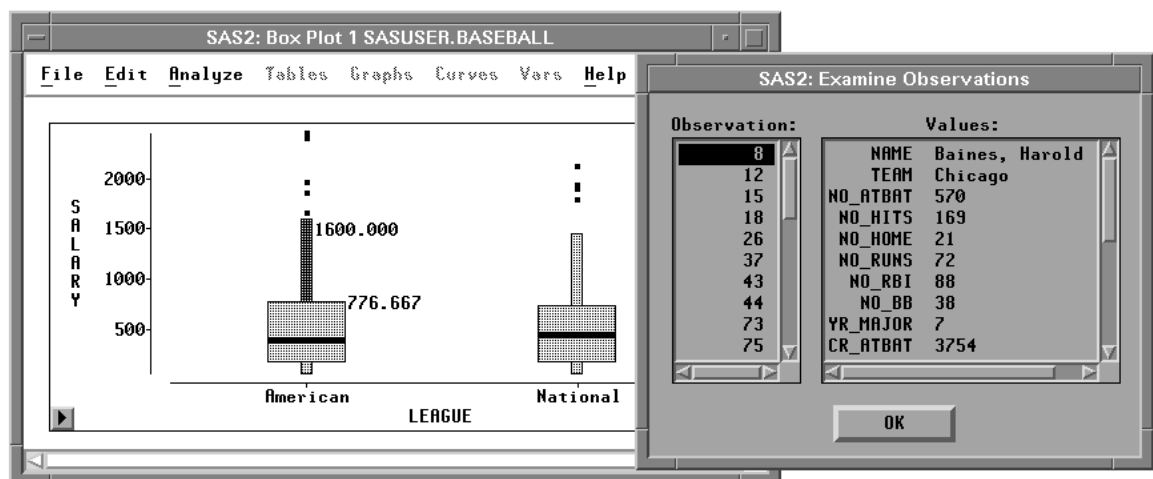


Figure 3.29. Examining Whisker Observations

⇒ **Click OK to close the dialog.**

Closing the Data Window

There are several other features of the data window, and you can find them by exploring the data pop-up menu on your own. For detailed information, see [Chapter 31, “Data Windows,”](#) in the Reference part of this manual. One more feature important enough to describe here concerns what happens when you close a data window.

† **Note:** When you close the data window, you close all windows using that data set. When you close all your data windows, you exit SAS/INSIGHT software.

You can open as many data windows as you like by choosing **File:Open**. You can close any window by choosing **File:End**. Depending on your host, there may be other ways to close windows as well.

You will be prompted with a dialog to confirm that you want to close the data window. In the Confirm dialog, you can click **OK** to close the data window, or you can click **Cancel** to abort the action and leave the data window open. Try it to be sure you know how to exit SAS/INSIGHT software when you are ready, but click **Cancel** in the Confirm dialog to abort the closing.

⇒ **Choose File:End.**

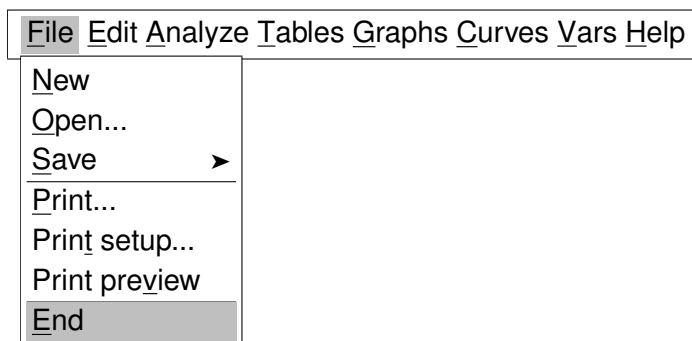


Figure 3.30. File Menu

Choosing **File:End** displays the Confirm dialog.

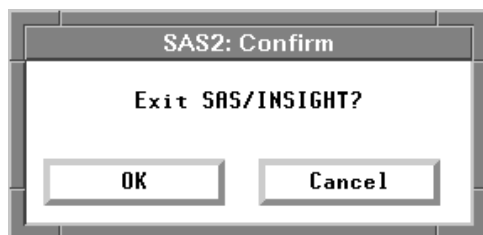


Figure 3.31. Confirm Dialog

⇒ **Click Cancel.**

This aborts the closing and returns you to the data window. If you had clicked **OK**, you would have closed the data window and exited SAS/INSIGHT software.

Techniques ♦ *Examining Data*

Now that you know how to examine data in a data window, turn to the next chapter to learn how to explore data in one dimension.

⊕ **Related Reading:** Data Windows, [Chapter 31](#).

Chapter 4

Exploring Data in One Dimension

Chapter Contents

BAR CHARTS	72
BOX PLOTS	80

Chapter 4

Exploring Data in One Dimension

In SAS/INSIGHT software, you can explore distributions of one variable using bar charts and box plots. *Bar charts* display distributions of interval or nominal variables. *Box plots* display concise summaries of interval variable distributions and show extreme values.

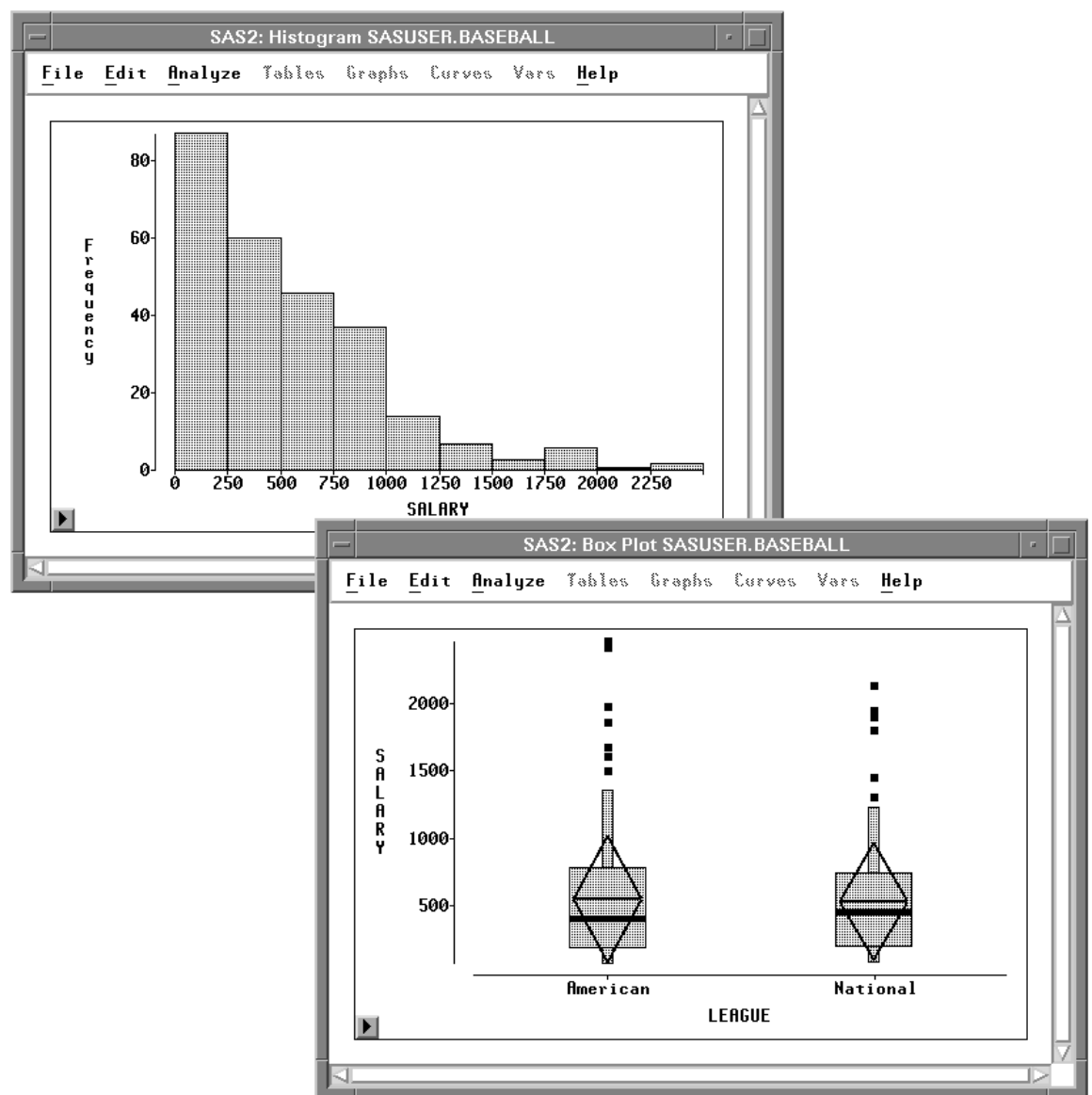


Figure 4.1. A Bar Chart and Box Plot

Bar Charts

Interval variables contain values distributed over a continuous range. For example, in [Figure 4.2](#) baseball players' salaries are stored in **SALARY**, an interval variable. To create a bar chart of players' salaries, follow these steps.

⇒ **Select SALARY in the data window.**

Scroll all the way to the right to find the **SALARY** variable. Point and click on the variable name.

SAS2: SASUSER.BASEBALL															
File		Edit		Analyze		Tables		Graphs		Curves		Vars		Help	
▶	22	Int	Int	Int	Int										
322		NO_OUTS	NO_ASSTS	NO_ERROR	SALARY										
■	1	317	36	1	75.000										
■	2	446	33	20	.										
■	3	80	45	8	240.000										
■	4	73	152	11	225.000										
■	5	247	4	8	.										
■	6	632	43	10	475.000										
■	7	186	290	17	550.000										
■	8	295	15	5	950.000										
■	9	90	4	0	.										
■	10	1236	98	18	100.000										
■	11	359	30	4	305.000										
■	12	368	20	3	1237.500										

Figure 4.2. Selecting the **SALARY** Variable

⇒ **Choose Histogram/Bar Chart (Y) from the Analyze menu.**

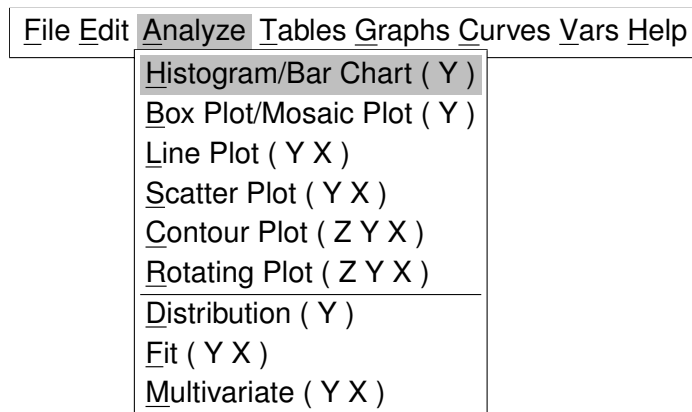


Figure 4.3. Creating a Bar Chart

This creates a bar chart, as shown in [Figure 4.4](#).

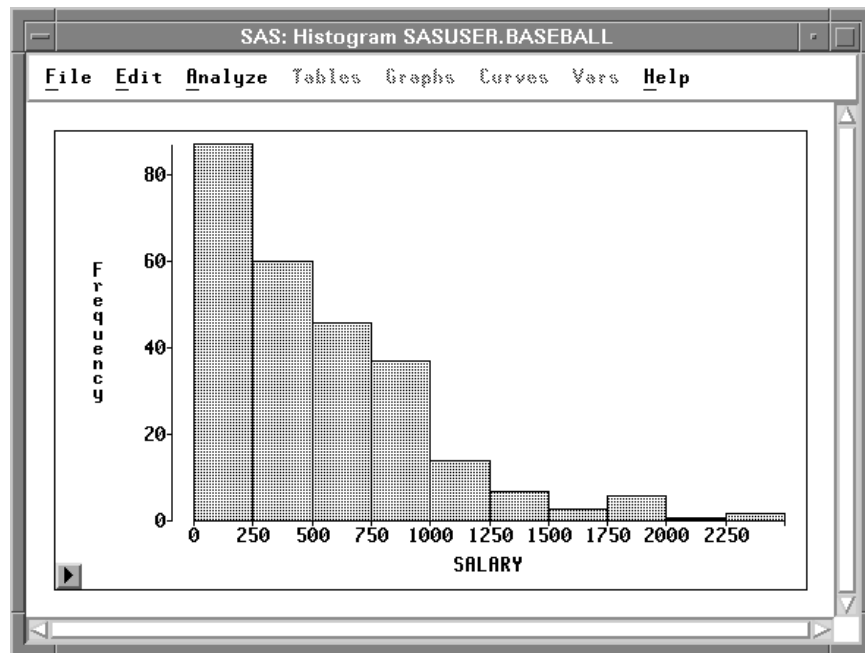


Figure 4.4. Bar Chart

⇒ **Point and click on any bar**

This labels the bar with its frequency and selects all the observations in the bar.

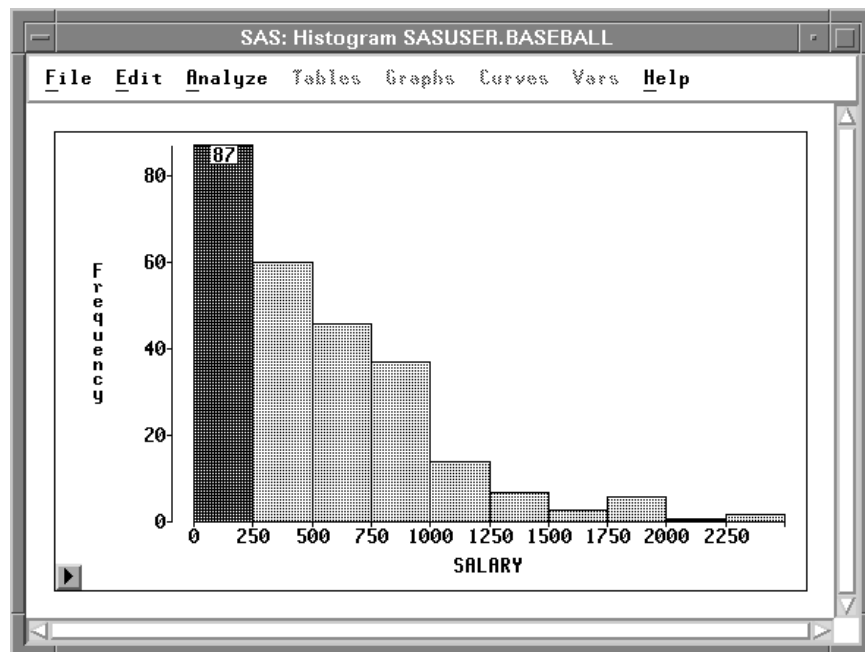


Figure 4.5. Clicking on a Bar

Notice that the observations are selected in the data window as well as in the bar chart window. Windows in SAS/INSIGHT software are just different views of the same data, so observations you select in one window are selected in all other windows.

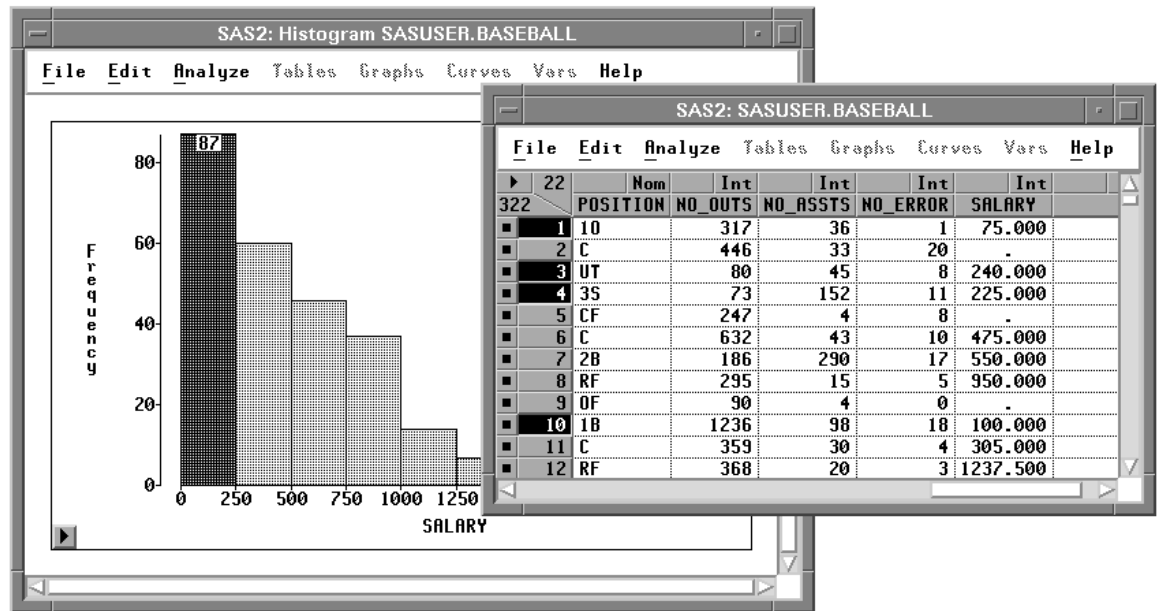


Figure 4.6. Selecting Observations in Multiple Windows

From this bar chart, you can see that the distribution of players' salaries is skewed to the right, with a few players earning high salaries. To find the number of players making the highest salaries, you can label all bars with their heights.

- ⇒ **Click on the menu button in the bottom left corner of the chart.**
 This displays the bar chart pop-up menu in [Figure 4.7](#). Click on **Values**.

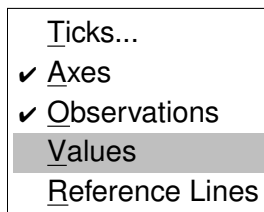


Figure 4.7. Bar Chart Pop-up Menu.

This toggles the display of values for all bar heights. There are three players making salaries above \$2,000,000.

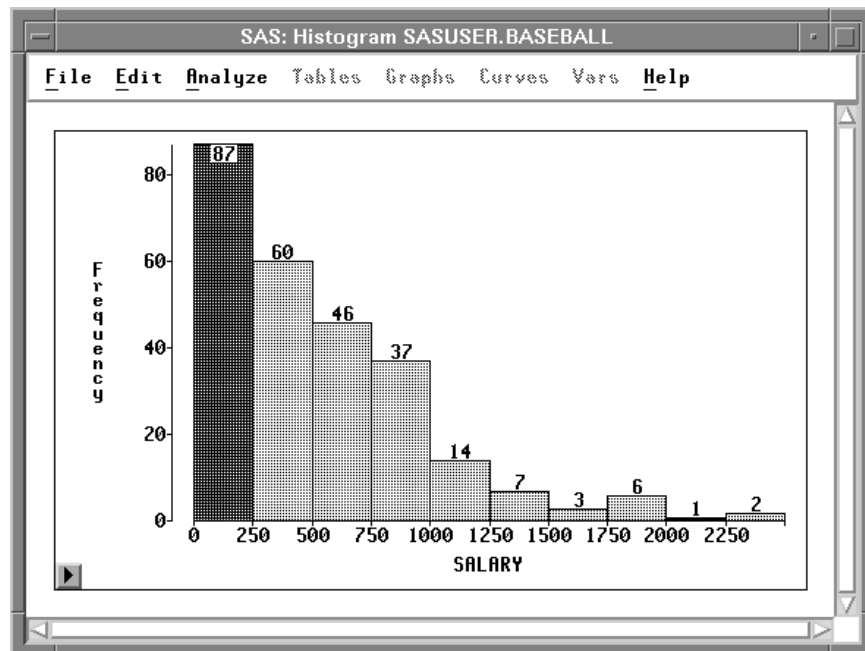


Figure 4.8. Bar Heights

It would be interesting to determine whether salaries differ in the American and National leagues. To compare the distribution of salaries from both leagues, follow these steps.

⇒ Select **LEAGUE** in the data window.

	CR_ATBAT	CR_HITS	CR_HOME	CR_RUNS	CR_RBI	CR_BB	LEAGUE	DIV
1	216	54	2	27	25	33	National	Wes
2	293	66	1	30	29	14	American	Eas
3	3231	825	36	376	290	238	National	Eas
4	926	210	9	118	69	114	National	Wes
5	4513	1134	224	542	727	230	American	Eas
6	3449	835	69	321	414	375	National	Wes
7	1775	506	6	272	125	194	National	Eas
8	3754	1077	140	492	589	263	American	Wes
9	7117	1981	242	964	1013	762	American	Wes
10	1750	412	100	204	276	155	American	Wes
11	999	236	21	108	117	118	American	Eas
12	2325	634	128	371	376	238	American	Eas

Figure 4.9. Selecting **LEAGUE**

Note that **LEAGUE** is a *nominal* variable. Nominal variables contain a discrete set of values. For example, **LEAGUE** contains only two values, **American** and **National**, for the American and National leagues.

⇒ **Choose Histogram/Bar Chart (Y) from the Analyze menu.**

From the bar chart in [Figure 4.10](#) you can see that the **BASEBALL** data set has more observations from the American League.

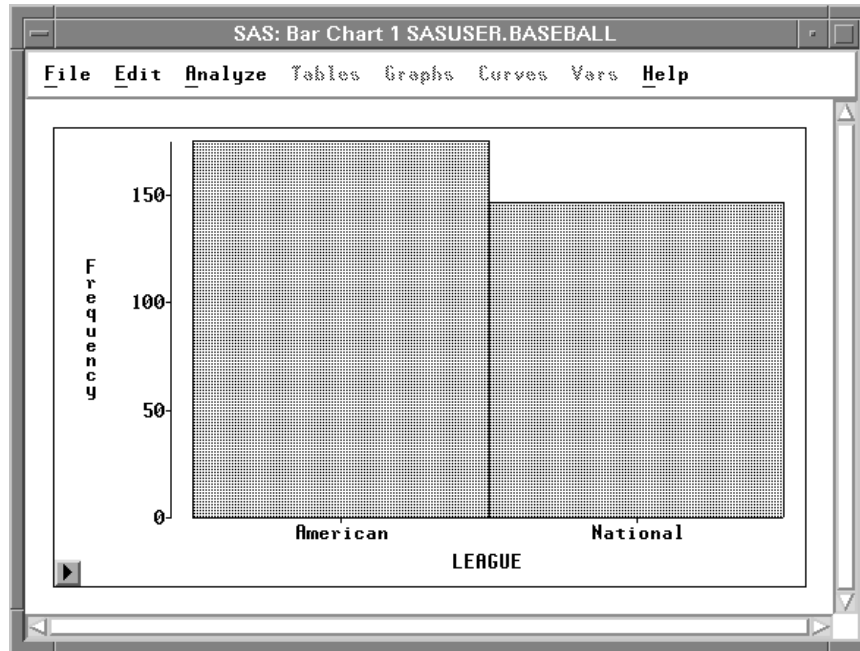


Figure 4.10. Bar Chart of **LEAGUE**

⇒ **Select Values from the bar chart pop-up menu in the new bar chart.**

This displays the frequencies for each of the leagues at the top of the bars on the bar chart.

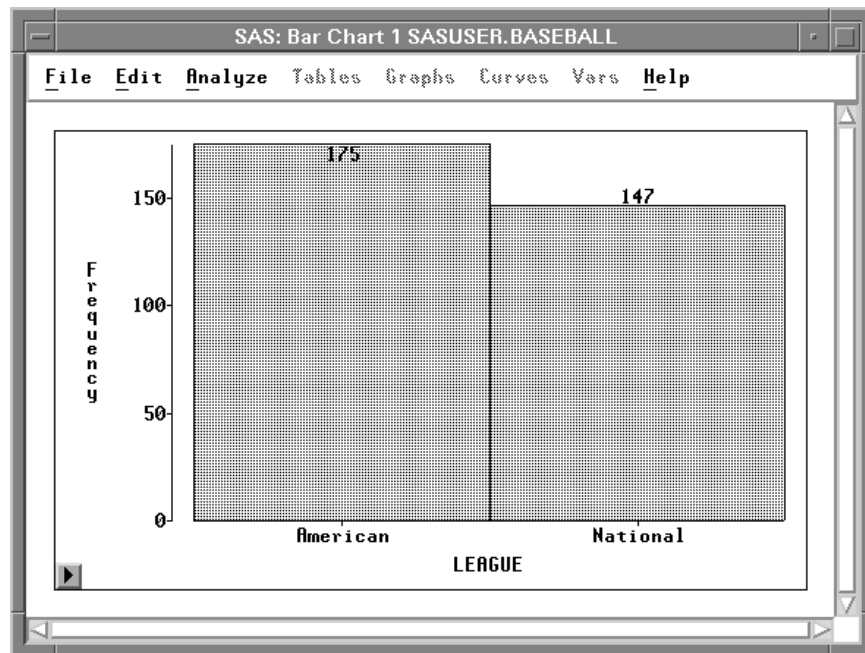


Figure 4.11. Bar Chart with Frequency Values

Techniques ♦ Exploring Data in One Dimension

- ⇒ **Arrange the windows so you can see both bar charts.**
- ⇒ **Click on the bar that represents the American League.**
This selects all observations for players in the American League.

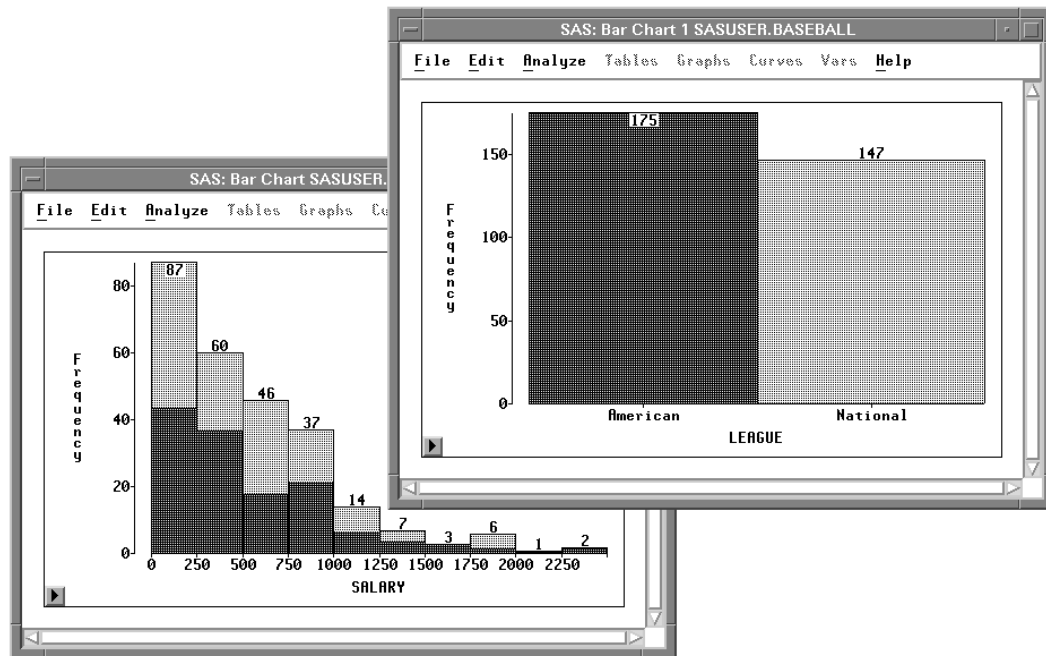


Figure 4.12. Selecting American League Observations

- ⇒ **Click on the bar that represents the National League.**
This selects all observations for players in the National League.

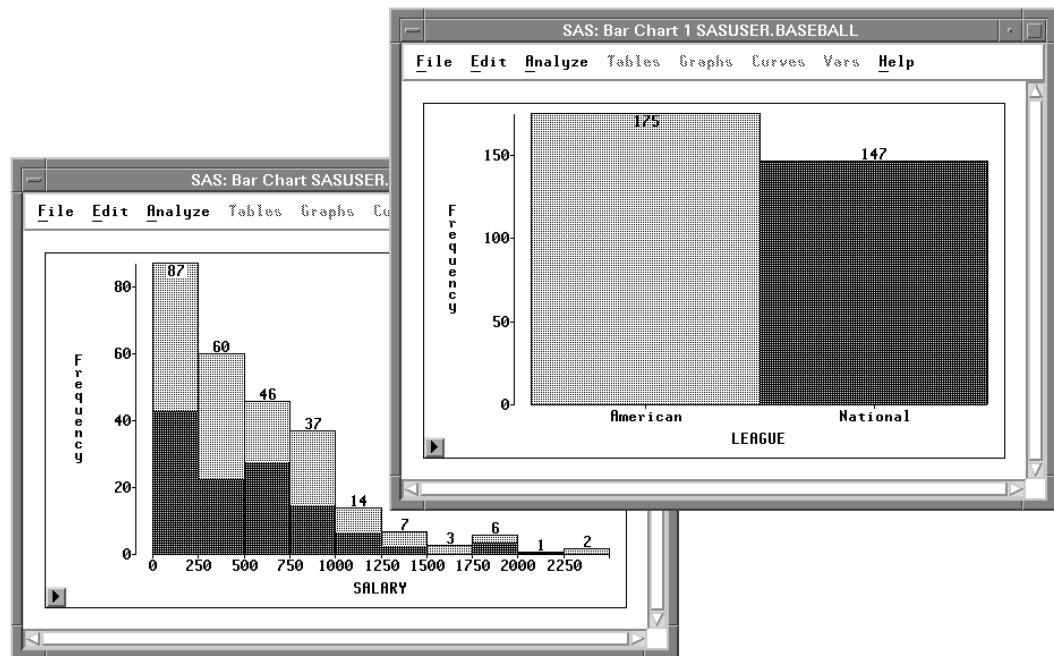


Figure 4.13. Selecting National League Observations

Both leagues have a broad distribution of **SALARY** with most players earning below \$1,000,000 and a few earning much more.

You can examine the distributions in more detail by creating box plots.

⊕ **Related Reading:** Bar Charts, [Chapter 32](#).

Box Plots

Box plots are an effective way to compare distributions of interval data. To create side-by-side box plots comparing the distributions of salaries for the American and National Leagues, follow these steps.

⇒ Choose **Analyze:Box Plot/Mosaic Plot (Y)**.

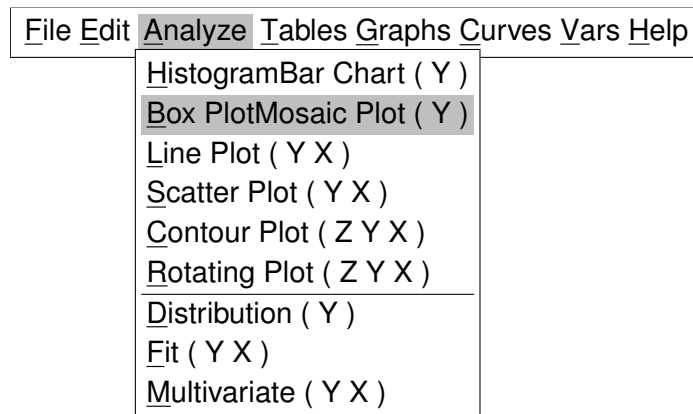


Figure 4.14. Creating a Box Plot

The (Y) in the **Box Plot/Mosaic Plot (Y)** menu indicates that a **Y** variable is *required* to create a box plot. Since you have no variables selected, a variables dialog prompts you to select at least one **Y** variable. Selecting a nominal variable for **Y** creates a mosaic plot; selecting an interval variable for **Y** creates a box plot.

Y is one of several *roles* you can assign to variables in analyses. The variables dialog shows that box plots and mosaic plots can also use **X**, **Group**, **Label**, and **Freq** variables.



Figure 4.15. Box Plot Variables Dialog

† **Note:** You can select variables before choosing from the **Analyze** menu, or you can choose from the **Analyze** menu before selecting variables. Selecting variables first is faster. If you select variables first, they are assigned to the required variable roles listed in the **Analyze** menu. Choosing the analysis first gives you more flexibility. If you choose the analysis first, you can assign optional variable roles such as **Group** and **Label**.

- ⇒ **Select **SALARY** in the list at the left, then click the **Y** button.**
This assigns the **Y** role to **SALARY**. The box plot displays the distribution of the **Y** variable.
- ⇒ **Select **LEAGUE** in the list at the left, then click the **X** button.**
This assigns the **X** role to **LEAGUE**. The box plot displays one schematic distribution plot side-by-side for each unique value of the **X** variable.
- ⇒ **Select **NAME** in the list at the left, then click the **Label** button.**
This assigns the **Label** role to **NAME**. The label variable is used to identify extreme values in the box plot.



Figure 4.16. Assigning Variable Roles

- ⇒ **Click **OK** to create the Box Plot.**

The box plot gives a concise picture of the distributions and places them side-by-side for easy comparison. The horizontal line in the middle of a box marks the *median* or 50th percentile. The top and bottom edges of a box mark the *quartiles*, or the 25th and 75th percentiles. The narrow boxes extending above and below are called *whiskers*. Whiskers extend from the quartiles to the farthest observation not farther than 1.5 times the distance between the quartiles. More extreme data values are plotted with individual markers.

The box plot shows long whiskers above with individual observations beyond the whiskers indicating severe skewness. These are the players making extremely high salaries.

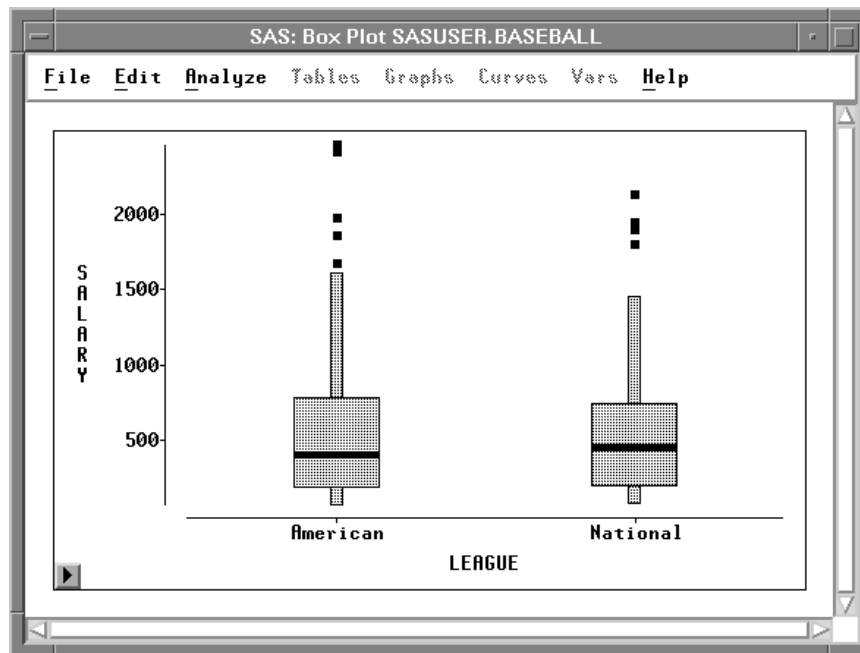


Figure 4.17. Side-By-Side Box Plots

⇒ **Point and click at the extreme values to identify them.**

Eddie Murray and Jim Rice were the highest paid players in the American league, while Mike Schmidt was the highest paid player in the National League.

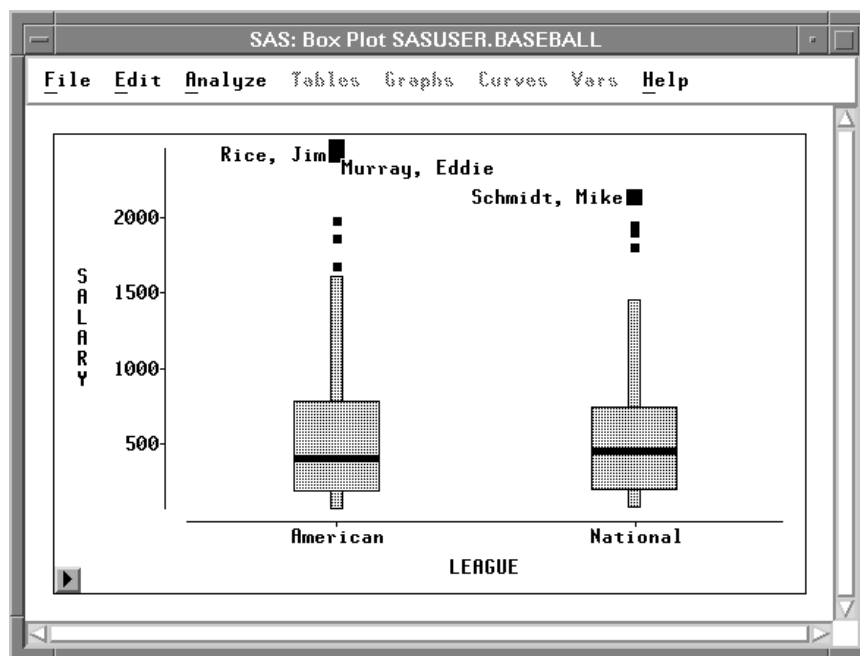


Figure 4.18. Identifying Extreme Values

You can also use a box plot to see the sample mean of a distribution.

- ⇒ **Click on the menu button in the lower left corner of the plot.**
 This displays the box plot pop-up menu. Click on **Means**.

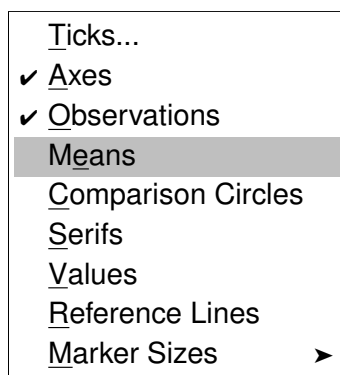


Figure 4.19. Box Plot Pop-up Menu

This toggles the display of mean diamonds on the box plot.

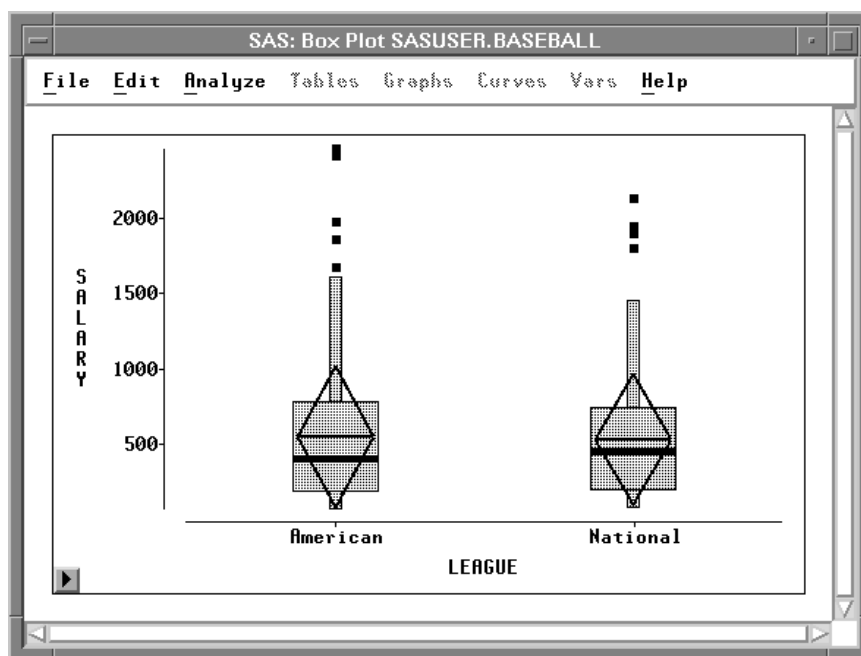


Figure 4.20. Box Plot with Mean Diamonds

The horizontal line in a mean diamond marks the mean salary for each league. The height of a mean diamond is two standard deviations (one on either side of the mean). In this case, the means and standard deviations for each league are almost identical.

Techniques ♦ *Exploring Data in One Dimension*

You can use other choices on the box plot pop-up menu to adjust axis tick marks and marker sizes and to toggle the display of observations, axes, serifs, and values. When there are two or more categories, you can toggle the display of *comparison circles*, which enable you to graphically compare the means of multiple categories.

⊕ **Related Reading:** Box Plots, [Chapter 33](#).

Chapter 5

Exploring Data in Two Dimensions

Chapter Contents

MOSAIC PLOTS	88
SCATTER PLOTS	91
SCATTER PLOT MATRICES	94
Brushing Observations	96
LINE PLOTS	100
REFERENCES	105

Chapter 5

Exploring Data in Two Dimensions

SAS/INSIGHT software provides mosaic plots, scatter plots, and line plots for exploring data in two dimensions. *Mosaic plots* are pictorial representations of frequency counts of nominal variables. *Scatter plots* are graphic representations of the relationship between two interval variables. *Line plots* show the relationships of multiple **Y** variables to a single **X** variable.

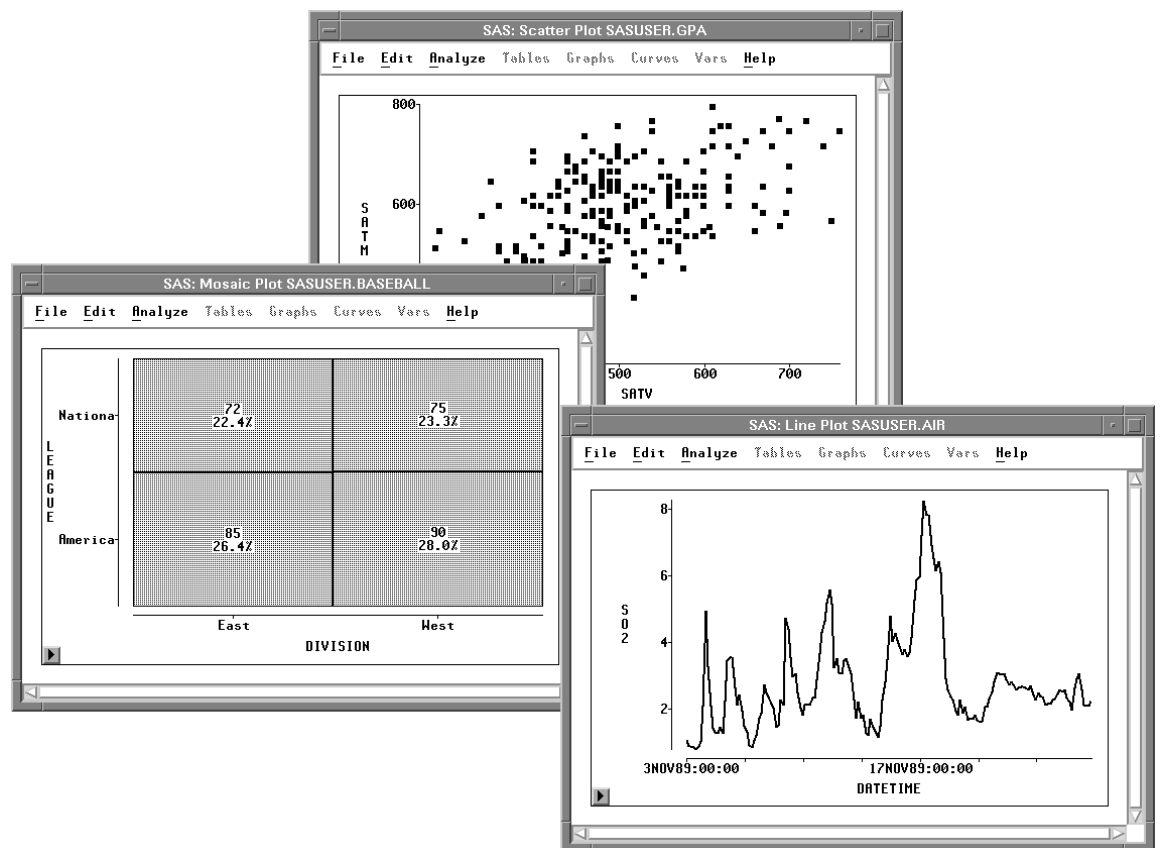


Figure 5.1. A Mosaic Plot, Scatter Plot, and Line Plot

Mosaic Plots

This example illustrates how to create mosaic plots for the **BASEBALL** data cross-classified by **LEAGUE** and **DIVISION**.

- ⇒ **Open the **BASEBALL** data set.**
- ⇒ **Choose **Analyze:Box Plot/Mosaic Plot (Y)**.**
- ⇒ **Assign **LEAGUE** the Y role and **DIVISION** the X role. Then click **OK**.**

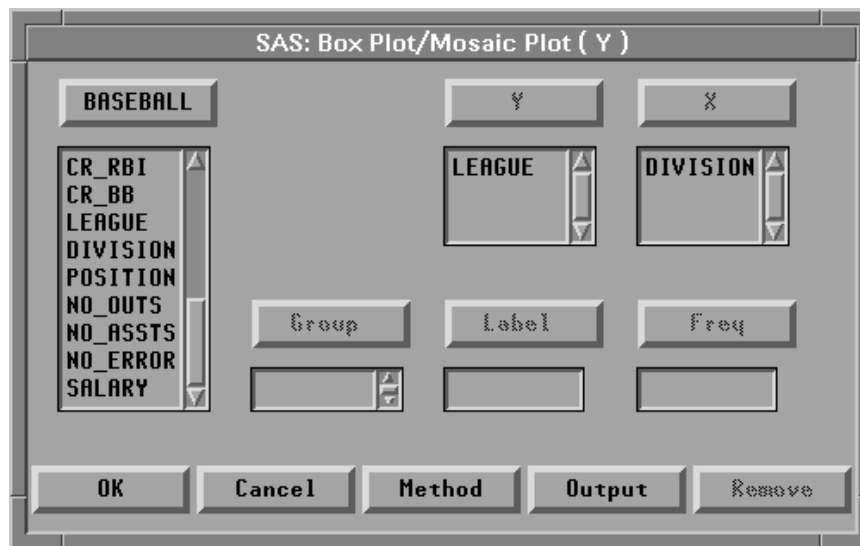


Figure 5.2. Assigning Variables for a Mosaic Plot

This creates a mosaic plot containing four boxes. The areas of the boxes in the mosaic plot are proportional to the number of observations in each category. You can see that, for these data, there are more players in the American League than in the National League and about the same number of players in the East and West Divisions.

You can find out more about specific categories by selecting the boxes.

- ⇒ **Click on the box at the lower left (American League East).**
This selects all the observations in the box and labels the box with its frequency and percentage. For this data, there are 85 players from the East Division of the American League, and these are 26.4% of the total.

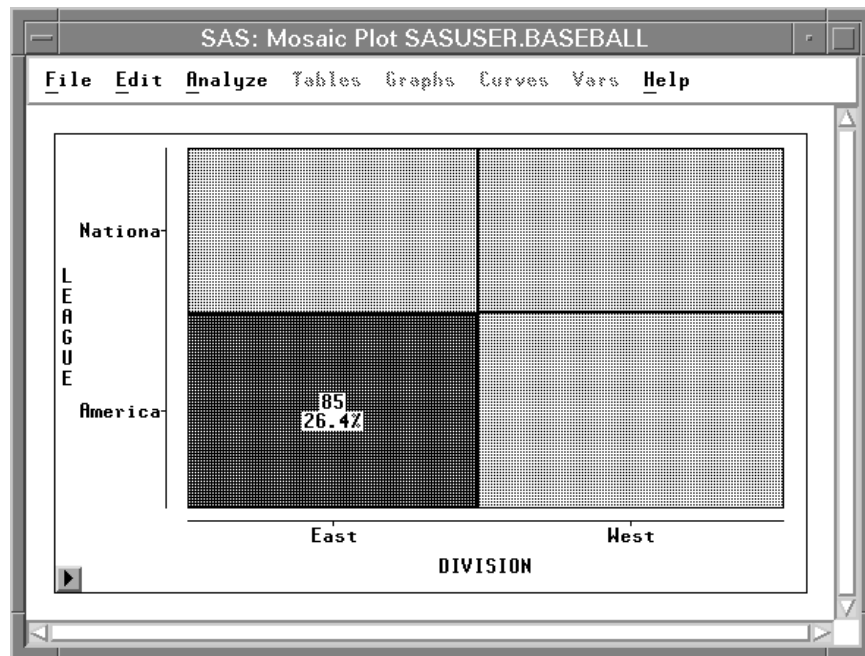


Figure 5.3. Clicking on a Box

⇒ **Double-click on the box to examine the observations.**

This selects all the observations in the box and displays the Examine Observations dialog. By clicking in the Examine Observations dialog, you can get detailed information on all the selected observations.

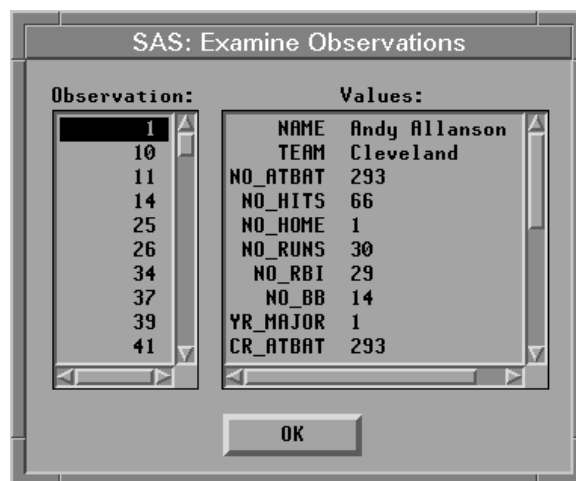


Figure 5.4. Examine Observation Dialog

You can add more information to the mosaic plot by displaying frequency counts and percentages.

⇒ **Choose Values from the pop-up menu.**

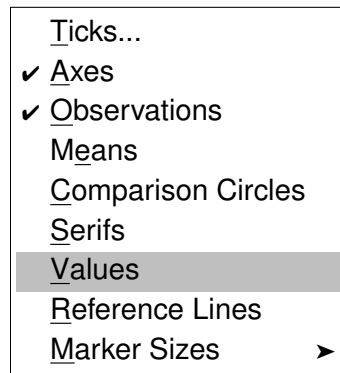


Figure 5.5. Mosaic Plot Pop-up Menu

This toggles the display of frequencies and percentages for all boxes in the mosaic plot.

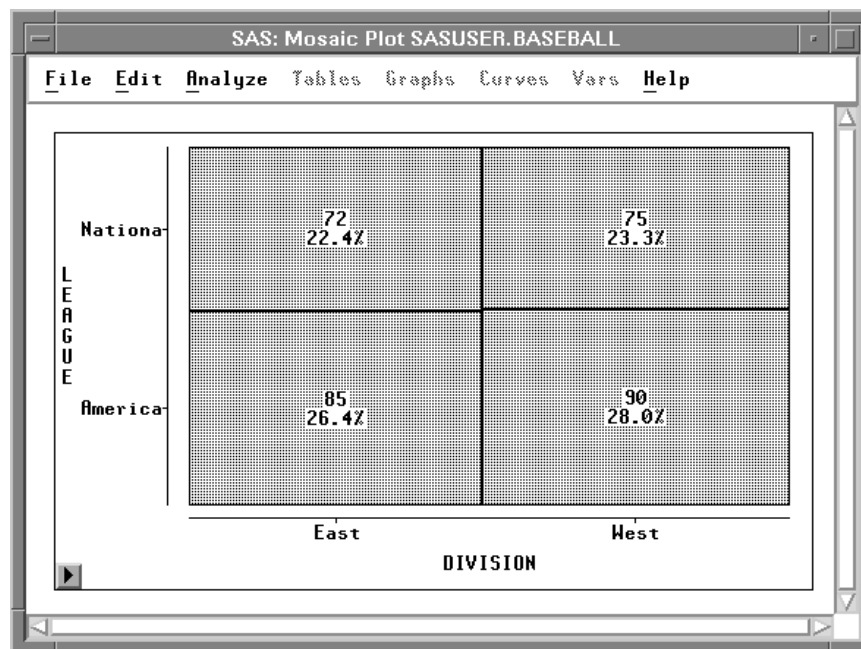


Figure 5.6. Mosaic Plot with Frequencies and Percentages

Scatter Plots

Scatter plots show the relationship between two variables. For example, you can explore the relationship between students' scores on standardized tests of math and verbal ability by following these steps.

⇒ **Open the GPA data set.**

⇒ **Select both the SATM and SATV variables.**

To select both variables, press the mouse button on **SATM**, move the mouse to **SATV**, then release the mouse button.

SAS: SASUSER.GPA

File Edit Analyze Tables Graphs Curves Vars Help

7	Int	Int	Int	Int	Int	Int	Nom		
224	GPA	HSM	HSS	HSE	SATM	SATV	SEX		
1	5.32	10	10	10	670	600	Female		
2	5.14	9	9	10	630	700	Male		
3	3.84	9	6	6	610	390	Female		
4	5.34	10	9	9	570	530	Male		
5	4.26	6	8	5	700	640	Female		
6	4.35	8	6	8	640	530	Female		
7	5.33	9	7	9	630	560	Male		
8	4.85	10	8	8	610	460	Male		
9	4.76	10	10	10	570	570	Male		
10	5.72	7	8	7	550	500	Female		
11	4.08	9	10	7	670	600	Female		
12	5.38	8	9	8	540	580	Female		

Figure 5.7. Selecting Two Variables

⇒ **Choose Analyze:Scatter Plot (Y X).**

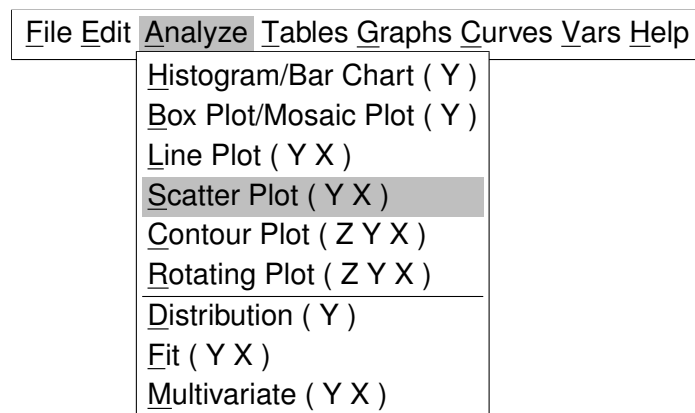


Figure 5.8. Creating a Scatter Plot

Techniques ♦ Exploring Data in Two Dimensions

This creates a scatter plot, as shown in [Figure 5.9](#). Note that the first variable you selected, **SATM**, is plotted on the **Y** axis, while the second variable selected, **SATV**, is plotted on the **X** axis.

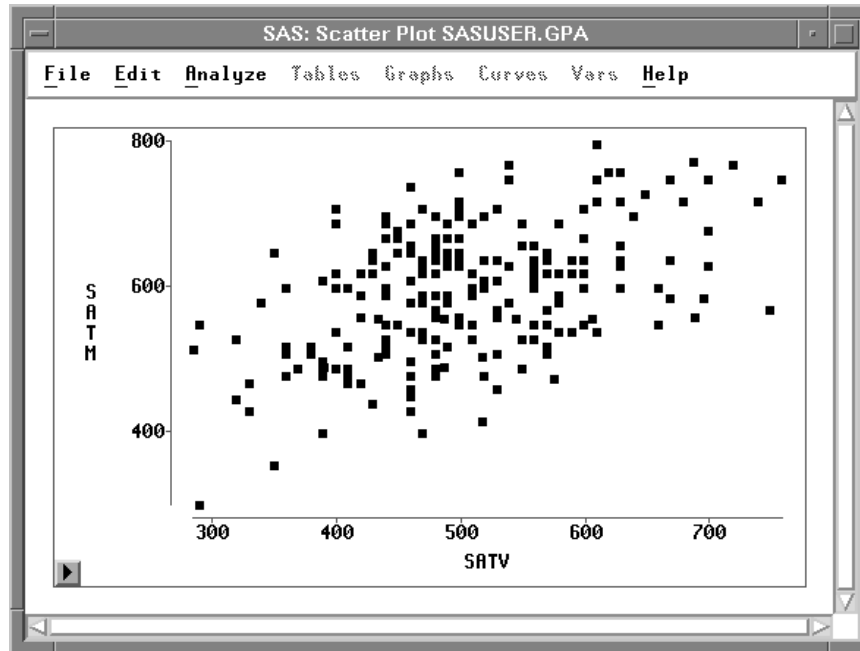


Figure 5.9. Scatter Plot Each *marker* in the scatter plot represents an observation, and its position shows the values of **SATM** and **SATV** for that observation. You can click on any marker to determine which observation it represents.

⇒ **Click on a marker.**

This selects the marker and displays its observation number. For example, observation 20 is selected in [Figure 5.10](#).

Clicking also selects the observation in the data window because windows are linked to their data. Any change to the data is automatically reflected in all windows.

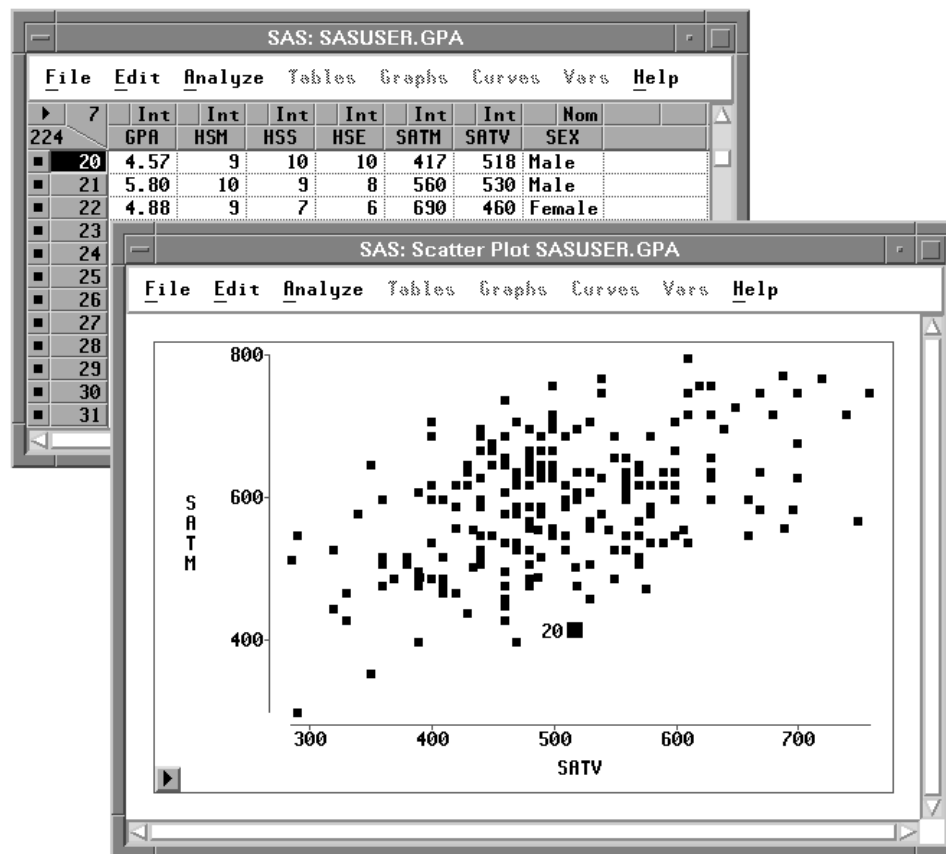


Figure 5.10. Selecting Observations in Multiple Windows

⇒ **Double-click on a marker.**

This selects the marker and displays the Examine Observation dialog. You can examine the values of all variables for the selected observation.

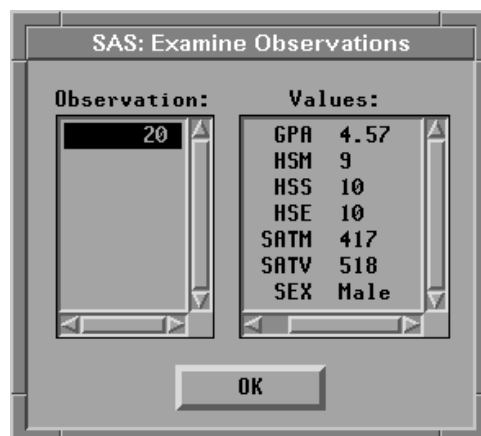


Figure 5.11. Examine Observations Dialog

Scatter Plot Matrices

A scatter plot *matrix* shows relationships among several variables taken two at a time. Scatter plot matrices can reveal a wealth of information, including dependencies, clusters, and outliers.

You can explore the relationships among students' college grade point averages and standardized test scores by following these steps.

⇒ **Select SATM, SATV, and GPA in the data window.**

To select these variables, use noncontiguous selection. On most hosts, you can use the **Ctrl** key to make a noncontiguous selection, as described in [Chapter 1, “Getting Started.”](#)

SAS: SASUSER.GPA															
File		Edit		Analyze		Tables		Graphs		Curves		Vars		Help	
▶	7	Int	Int	Int	Int	Int	Int	Nom							
224		GPA	HSM	HSS	HSE	SATM	SATV	SEX							
■	20	4.57	9	10	10	417	518	Male							
■	21	5.80	10	9	8	560	530	Male							
■	22	4.88	9	7	6	690	460	Female							
■	23	4.28	8	10	10	600	600	Male							
■	24	5.06	8	6	5	540	400	Female							
■	25	5.21	8	8	7	600	400	Female							
■	26	3.60	4	7	7	460	460	Male							
■	27	5.47	10	10	9	720	680	Male							
■	28	4.00	3	7	6	460	530	Female							
■	29	5.18	9	10	8	670	450	Female							
■	30	4.77	6	5	9	590	440	Female							
■	31	4.38	9	9	10	650	570	Male							

Figure 5.12. Selecting Three Variables

⇒ **Choose Analyze:Scatter Plot (Y X).**

This creates the scatter plot matrix shown in [Figure 5.13](#).

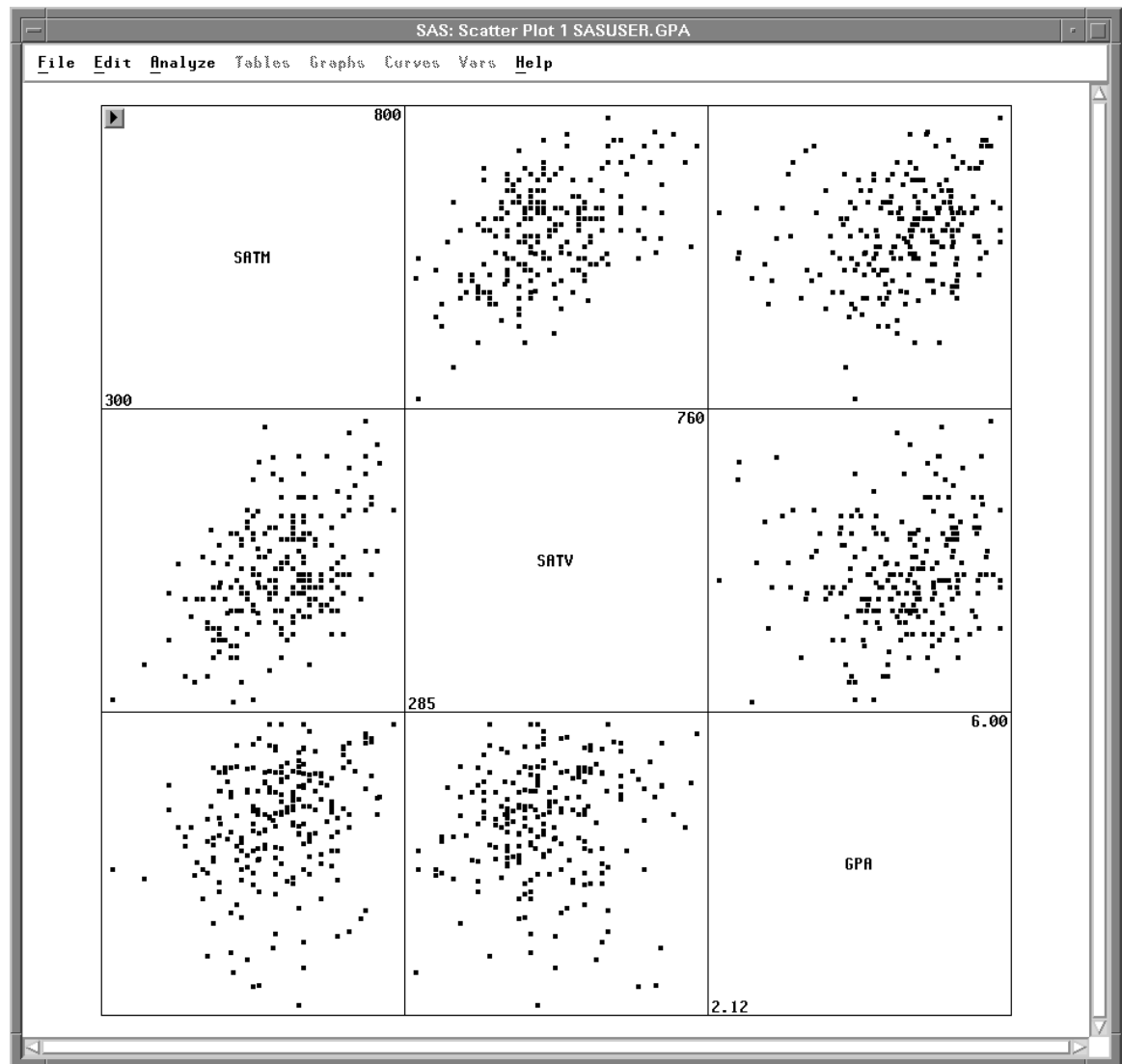


Figure 5.13. Scatter Plot Matrix

The plots are organized in a matrix of all pairwise combinations of the variables **SATM**, **SATV**, and **GPA**. Plots are arranged so that adjacent plots share a common axis. All plots in a row share a common **Y** axis, and all plots in a column share a common **X** axis. The diagonal cells of the matrix contain the names of the variables and their minimum and maximum values.

⇒ **Click on a marker in any scatter plot.**

The observation label is displayed and corresponding markers in all scatter plots are selected, as shown in [Figure 5.14](#). This enables you to explore observations to see, for example, if an outlier in one scatter plot is an outlier in other scatter plots.

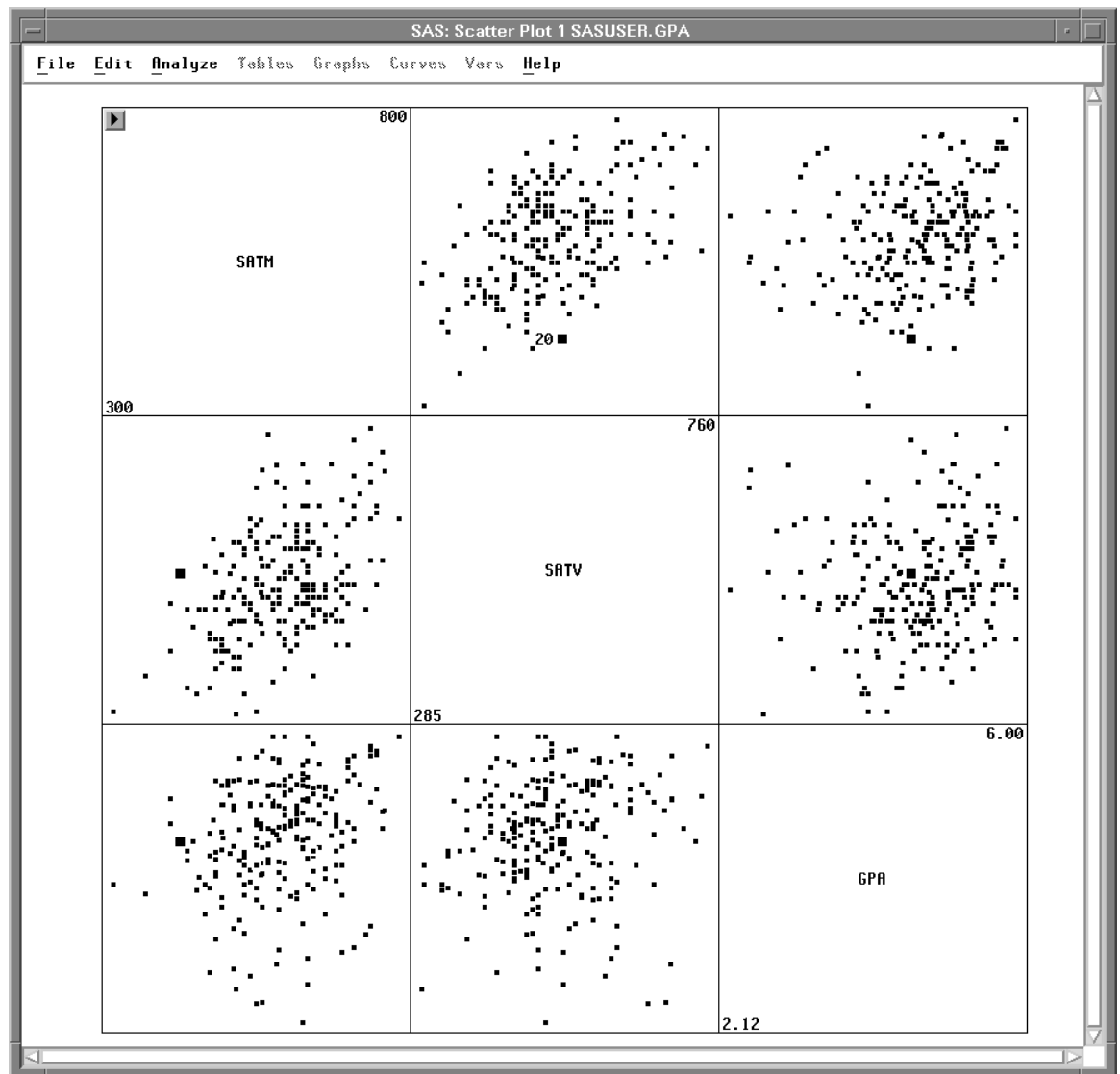


Figure 5.14. Selecting Observations in a Scatter Plot Matrix

Brushing Observations

Brushing is a dynamic method of selecting groups of observations simultaneously in all views of the data. Brushing is an effective technique for investigating multivariate data (Becker, Cleveland, and Wilks, 1987). For example, you can use brushing to find students who performed poorly on their SATs but still had relatively high grade point averages.

⇒ Select observations with low values for **SATM** and **SATV**.

Press the mouse button down, move the mouse, then release the mouse button to create a rectangle in the plot of **SATM** by **SATV**. This rectangle is your *brush*. The observations in the rectangle are selected. Notice that corresponding observations are also highlighted in the other plots.

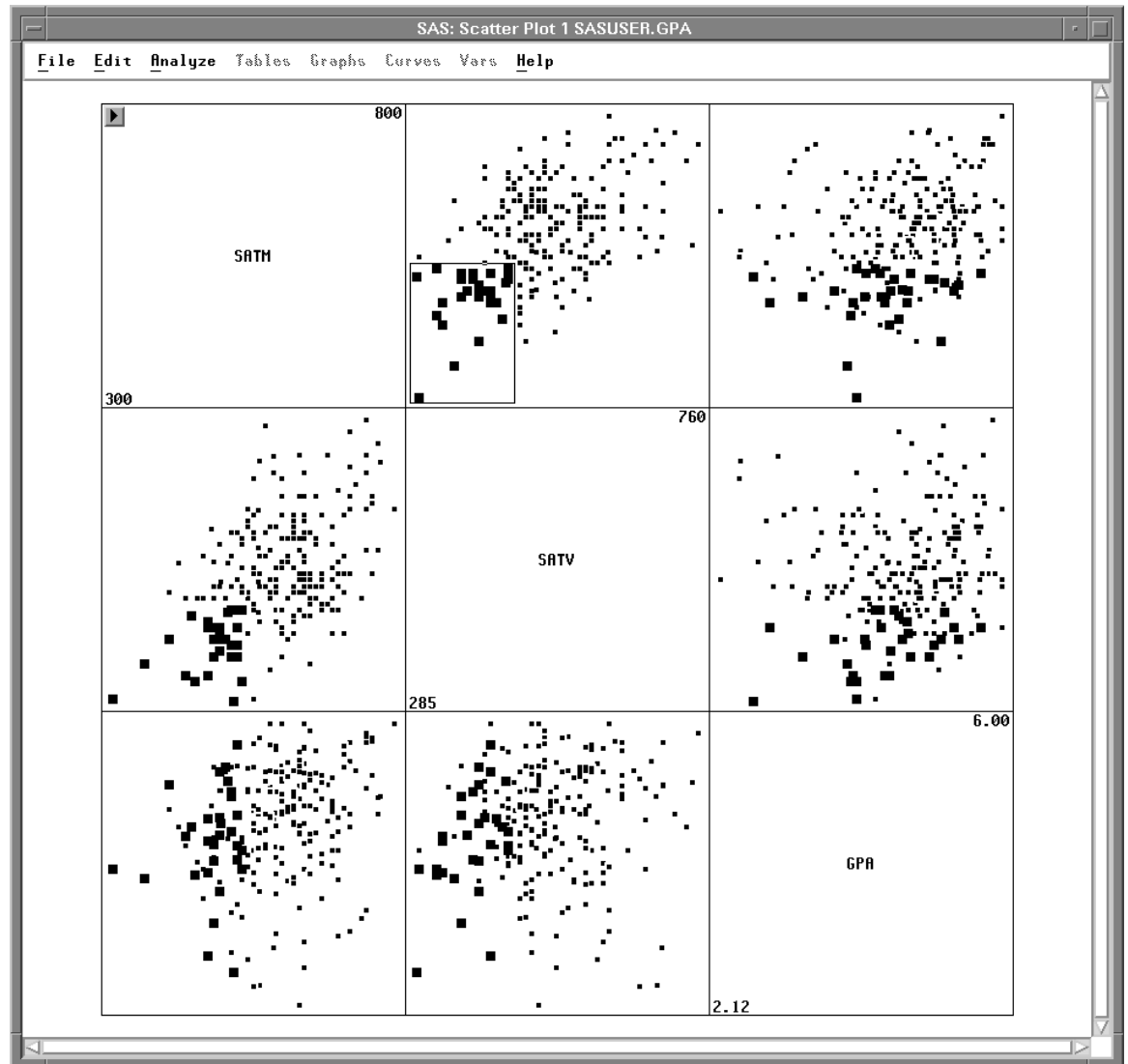


Figure 5.15. Brushing in a Scatter Plot Matrix

Examine one of the scatter plots involving **GPA**. Several of the selected observations have **GPA** values of 4 or above, indicating that SAT scores are not always good indicators of success in the school's computer science program.

You can change the size of your brush to select different observations.

- ⇒ **Place the cursor on the *corner* of the brush and drag the cursor.**
 The brush changes size as you drag until you release the mouse button.

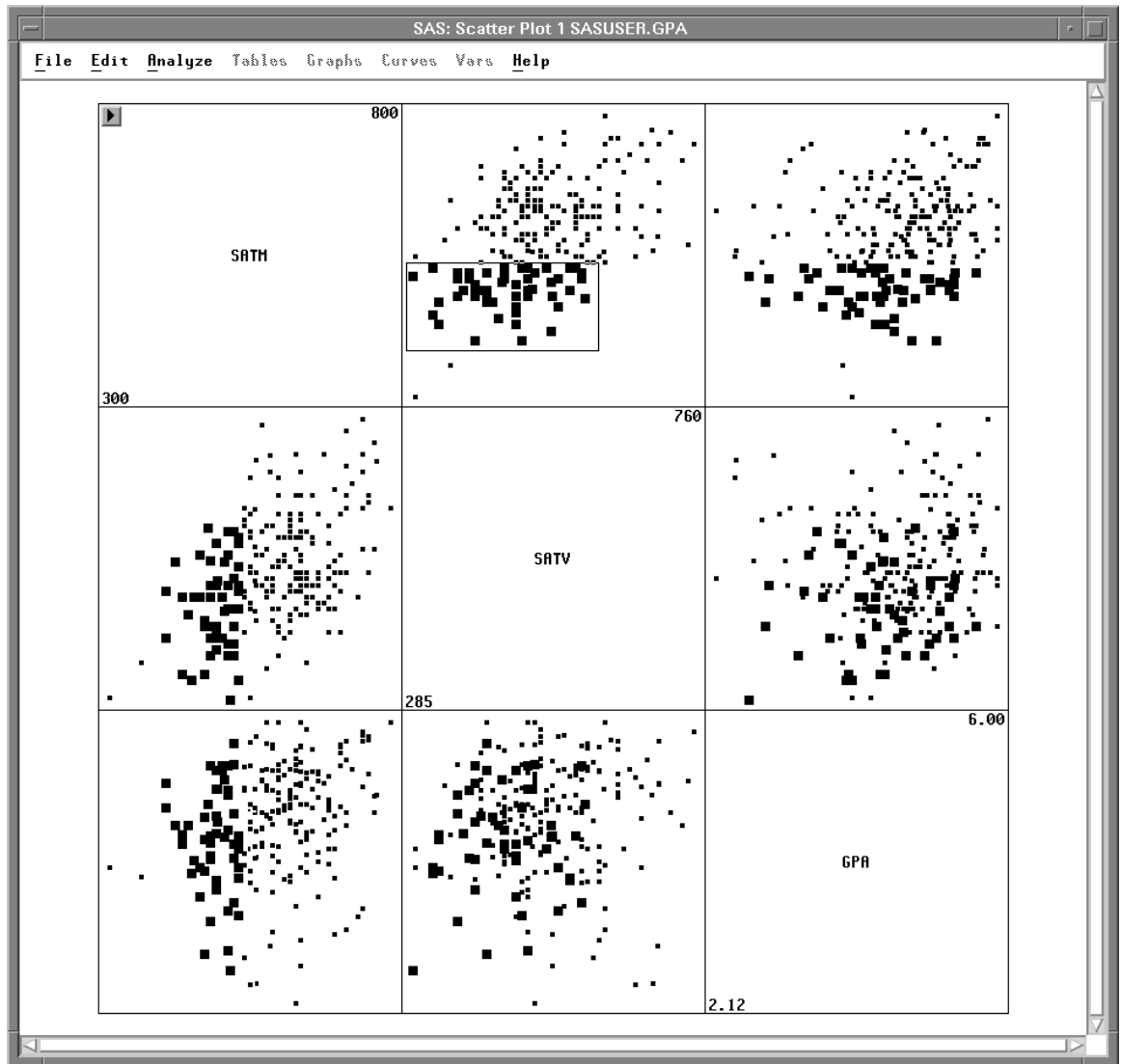


Figure 5.16. Changing the Size of a Brush

You can move the brush to select observations dynamically.

⇒ **Place the cursor in the brush and drag the brush across the plot.**

As observations enter the brush they become selected, and as they leave they are deselected. The corresponding observations in all the other scatter plots are also selected and deselected as you move the brush.

If you release the mouse button while you are moving the brush, the brush continues to move. *Throwing* the brush in this way removes the burden of eye-hand coordination, enabling you to take your eyes off the brush and more easily see its effect in other plots.

You can also brush with extended selection. This is a convenient way to select a set of observations that does not fit the rectangular shape of the brush. Extended selection,

described in Chapter 1, uses the **Shift** key on most hosts.

⇒ **Using extended selection, create another brush.**

The observations that were in the previous brush remain selected.

⇒ **Using extended selection, move the brush.**

Observations become selected as they enter the brush, but they are not deselected when they leave the brush, as illustrated in [Figure 5.17](#).

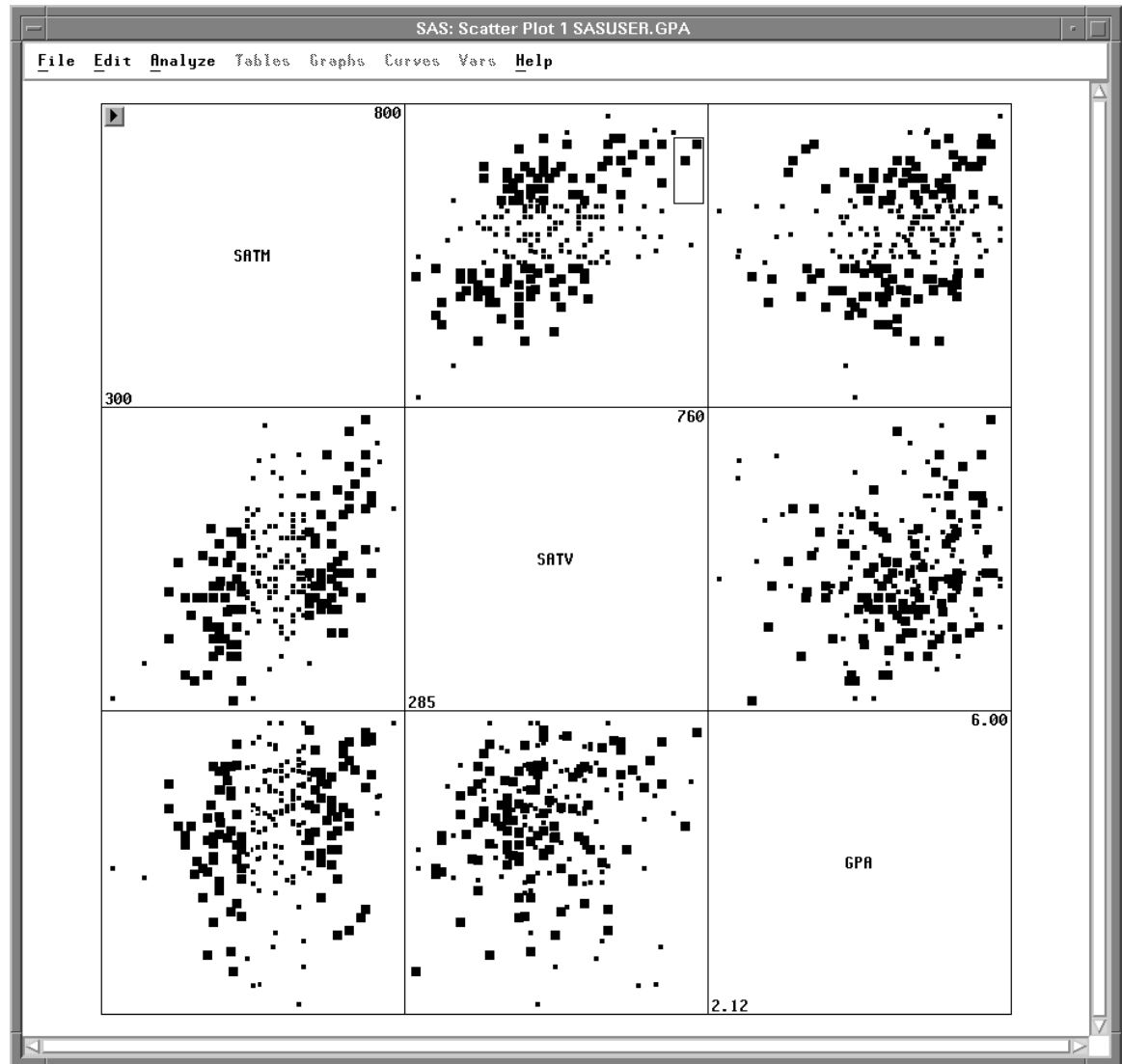


Figure 5.17. Brushing with Extended Selection

⇒ **To remove the brush, click in any empty area of the window.**

Clicking on nothing deselects all selected objects.

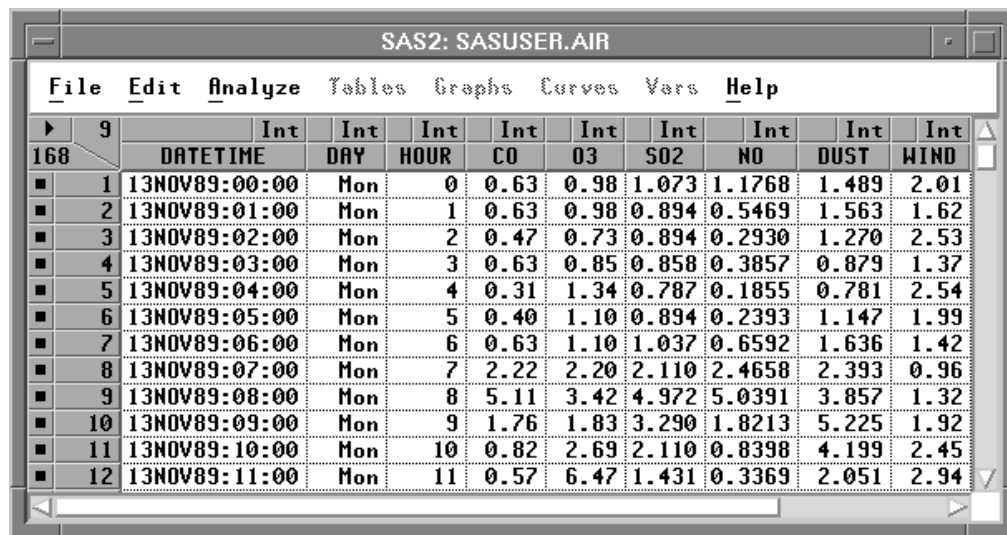
⊕ **Related Reading:** Scatter Plots, [Chapter 35](#).

Line Plots

Line plots are often used to show trends over time. For example, you can explore the patterns in pollutant concentrations in the **AIR** data set by following these steps.

⇒ **Open the AIR data set.**

This data set contains measurements of air quality as indicated by concentrations of various pollutants. Among the pollutants are carbon monoxide (**CO**), ozone (**O3**), sulfur dioxide (**SO2**), nitrogen oxide (**NO**), and **DUST**.



	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
168	DATETIME	DAY	HOUR	CO	O3	SO2	NO	DUST	WIND	
1	13NOV89:00:00	Mon	0	0.63	0.98	1.073	1.1768	1.489	2.01	
2	13NOV89:01:00	Mon	1	0.63	0.98	0.894	0.5469	1.563	1.62	
3	13NOV89:02:00	Mon	2	0.47	0.73	0.894	0.2930	1.270	2.53	
4	13NOV89:03:00	Mon	3	0.63	0.85	0.858	0.3857	0.879	1.37	
5	13NOV89:04:00	Mon	4	0.31	1.34	0.787	0.1855	0.781	2.54	
6	13NOV89:05:00	Mon	5	0.40	1.10	0.894	0.2393	1.147	1.99	
7	13NOV89:06:00	Mon	6	0.63	1.10	1.037	0.6592	1.636	1.42	
8	13NOV89:07:00	Mon	7	2.22	2.20	2.110	2.4658	2.393	0.96	
9	13NOV89:08:00	Mon	8	5.11	3.42	4.972	5.0391	3.857	1.32	
10	13NOV89:09:00	Mon	9	1.76	1.83	3.290	1.8213	5.225	1.92	
11	13NOV89:10:00	Mon	10	0.82	2.69	2.110	0.8398	4.199	2.45	
12	13NOV89:11:00	Mon	11	0.57	6.47	1.431	0.3369	2.051	2.94	

Figure 5.18. AIR Data

⇒ **Choose Analyze:Line Plot (Y X).**

This displays the line plot variables dialog.

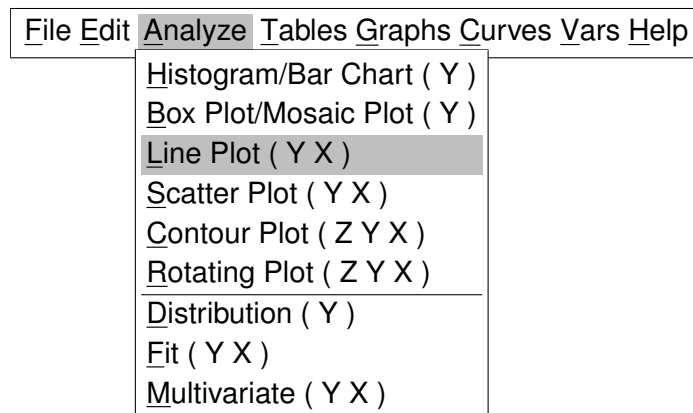


Figure 5.19. Creating a Line Plot

⇒ **Assign CO and SO2 the Y role, and DATETIME the X role.**

⇒ **Assign DATETIME the Label role also. Then click OK.**

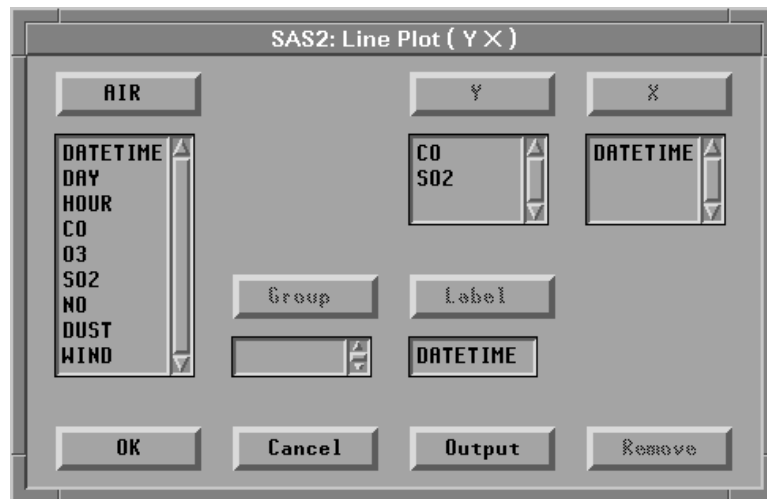


Figure 5.20. Assigning Line Plot Variables

This creates a line plot with one line for each **Y** variable.

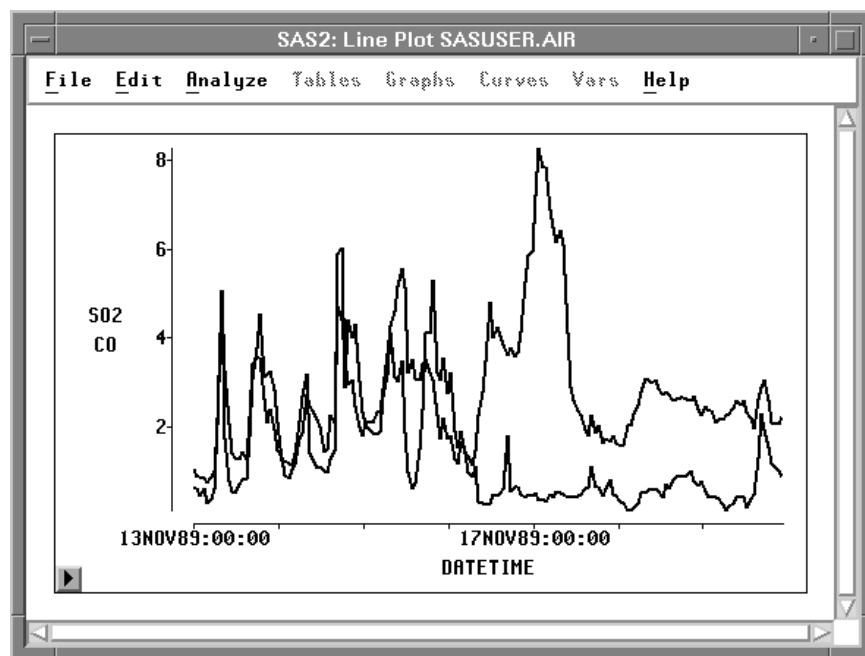


Figure 5.21. Line Plot

To associate lines with variables, simply select the variable.

⇒ **Click on the SO2 variable.**

This highlights both the variable and the corresponding line.

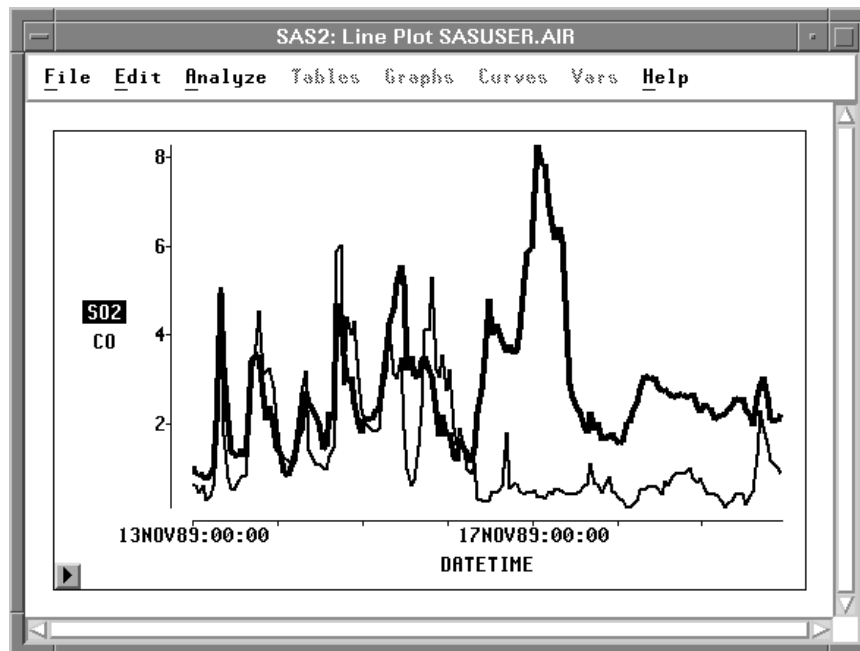


Figure 5.22. SO2 Selected

By clicking on the variables, you can see that the **SO2** concentration rises to a peak on the 17th of November and then falls. The **CO** concentration shows a regular pattern of peaks and valleys up until the 16th; then it falls also.

To show more information, you can add observation markers to the line plot.

⇒ **Click on the menu button in the lower left corner of the plot. Choose Observations.**

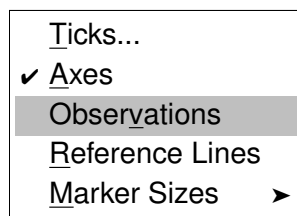


Figure 5.23. Line Plot Pop-up Menu

This displays the line plot with observation markers.

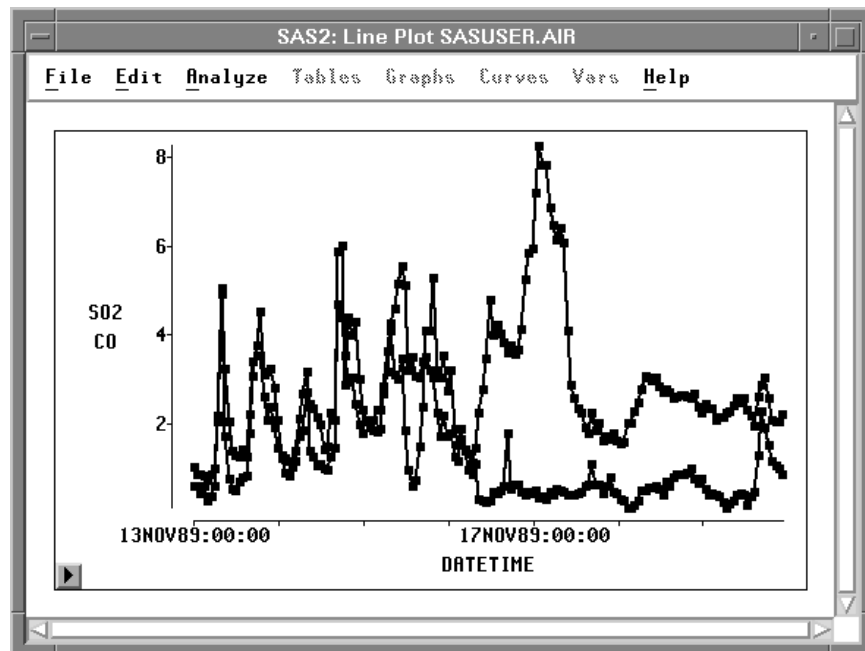


Figure 5.24. Line Plot with Observations

⇒ Point and click to identify observations with the highest pollutant concentrations.

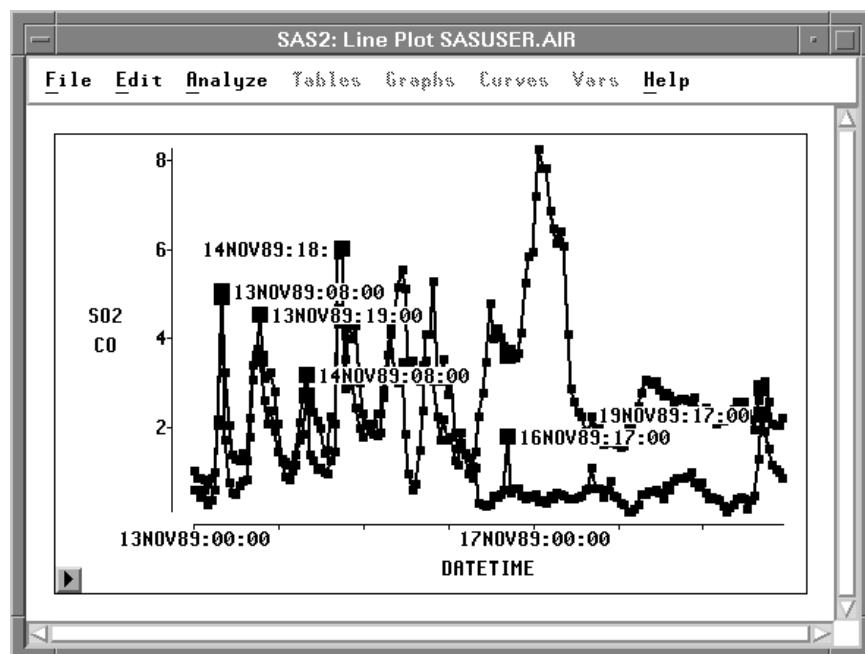


Figure 5.25. Identifying Observations

Techniques ♦ Exploring Data in Two Dimensions

Most of the peaks for **CO** occur in the morning and evening, around hours 08:00 or 18:00. Carbon monoxide pollution is often caused by automobiles, so these peaks might be caused by rush-hour traffic.

The **SO2** concentration follows a different pattern. Sulfur dioxide is a pollutant given off by power plants. Perhaps there was a peak demand for electricity on the 17th.

The drop in pollutants after the 17th can be partly explained by noting that the 18th and 19th were Saturday and Sunday. The weekend eliminates rush-hour traffic patterns. However, the **CO** level dropped on the 16th also, which was Thursday. There is an additional factor at work here.

⇒ Choose **Edit:Windows:Renew** to re-create the line plot.

⇒ Add **WIND** to the **Y** variable list. Then click **OK**.

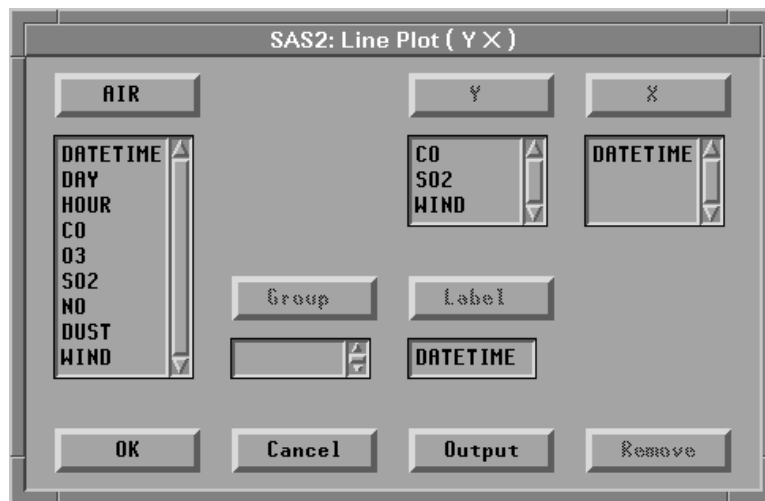


Figure 5.26. Adding **WIND** Variable

⇒ In the line plot, click on the **WIND** variable.

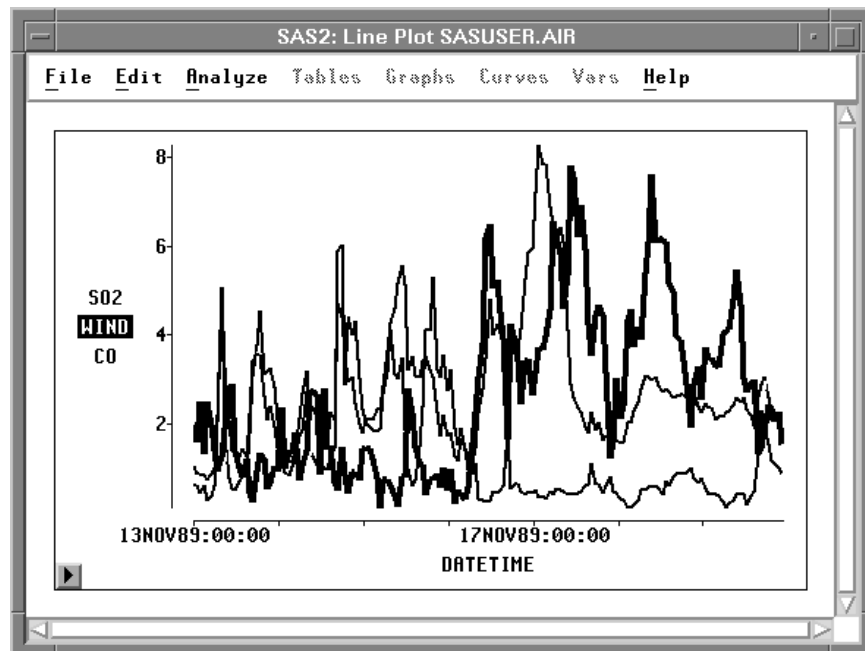


Figure 5.27. WIND Speed

Not only were the 18th and 19th a weekend, but there were high winds on the 16th, 17th, 18th, and 19th. These winds cleared much of the pollutants from the local atmosphere.

- ⊕ **Related Reading:** Mosaic Plots, [Chapter 33](#).
- ⊕ **Related Reading:** Scatter Plots, [Chapter 35](#).
- ⊕ **Related Reading:** Line Plots, [Chapter 34](#).

References

Becker, R.A., Cleveland, W.S., and Wilks, A.R. (1987), “Dynamic Graphics for Data Analysis,” *Statistical Science*, 2 (4), 355–382.

Chapter 6

Exploring Data in Three Dimensions

Chapter Contents

ROTATING PLOTS	110
ROTATING PLOT WITH FITTED SURFACE	116
CONTOUR PLOTS	118

Chapter 6

Exploring Data in Three Dimensions

SAS/INSIGHT software provides rotating plots, surface plots, and contour plots for exploring data in three dimensions. A *rotating plot* is a three-dimensional scatter plot, so it shows a graphic representation of the relationship among three interval variables. A *surface plot* is a rotating plot with a surface that models a third variable as a function of two other variables. A *contour plot* shows how the values of one variable may depend on the values of two other variables.

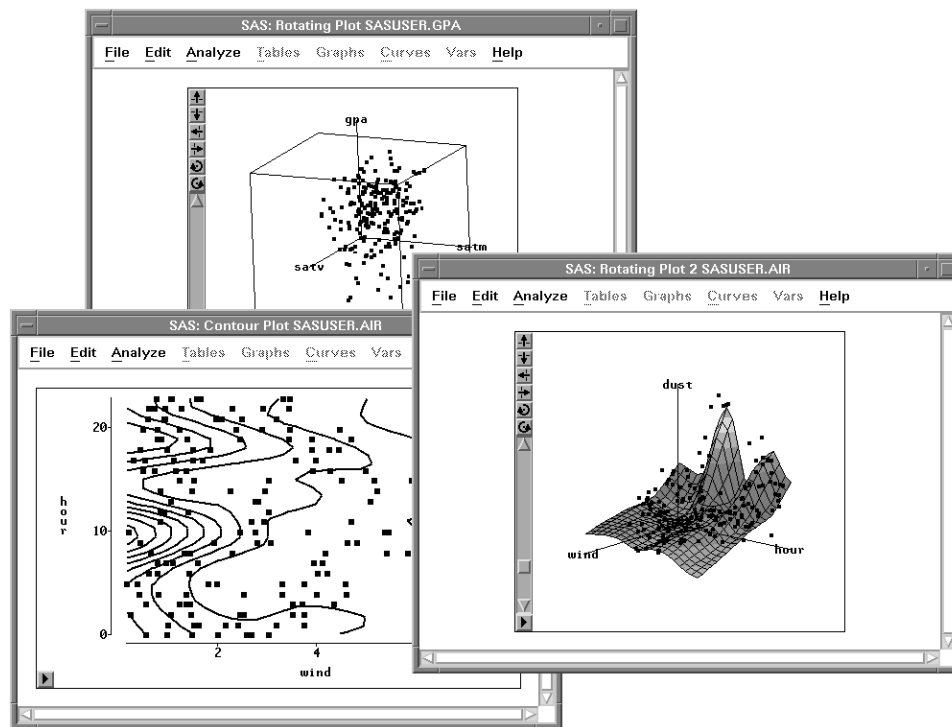


Figure 6.1. A Rotating Plot, Surface Plot, and Contour Plot

Rotating Plots

Using rotation you can obtain unique views into the data that can reveal structure not visible with static plots or not detectable with analytic methods.

Follow these steps to explore the relationships among students' SAT verbal scores, SAT math scores, and college grade point averages.

⇒ **Open the GPA data set.**

⇒ **Choose Analyze:Rotating Plot (Z Y X).**

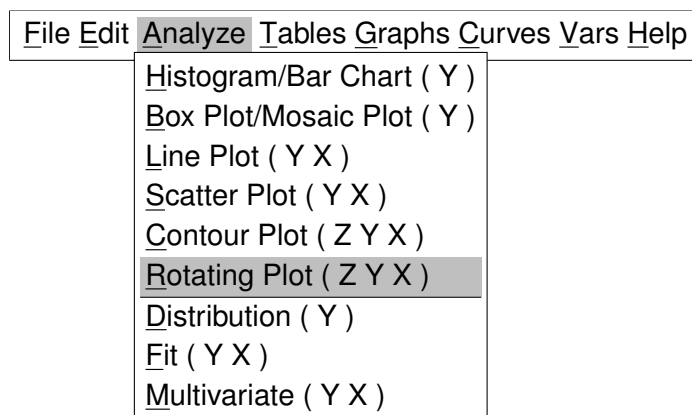


Figure 6.2. Creating a Rotating Plot

A rotating plot variables dialog appears, as shown in [Figure 6.3](#). The (Z Y X) in the menu indicates that **Z**, **Y**, and **X** variables are *required* to create the rotating plot.

⇒ **Select GPA in the variables list at the left. Then click Z.**

This assigns the **Z** role to the **GPA** variable. Using the same method, assign **SATM** the **Y** role and **SATV** the **X** role.

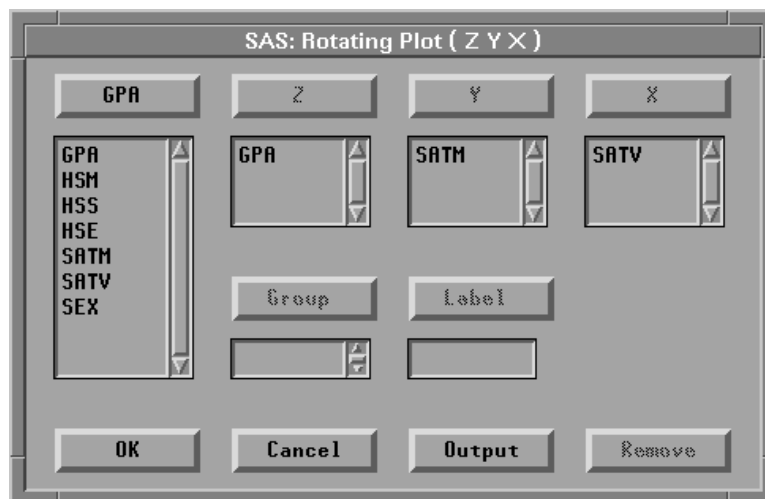


Figure 6.3. Rotating Plot Variables Dialog

⇒ **Click OK to create a rotating plot.**

The **GPA** axis is not visible when the rotating plot first appears on the display because the **Z** dimension is projected into the **X-Y** plane.

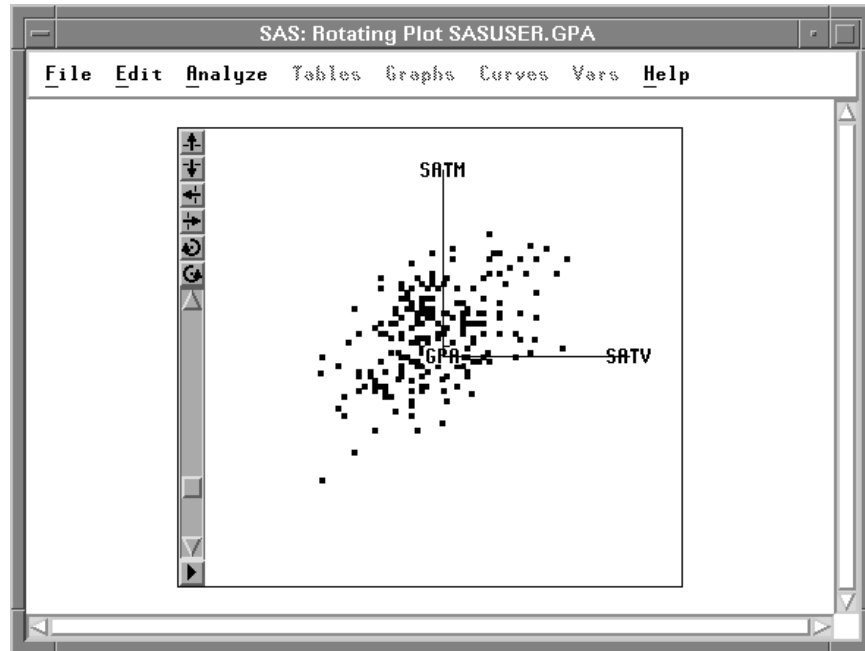


Figure 6.4. Rotating Plot

SAS/INSIGHT software provides both control buttons and a hand tool to rotate the plot. First, examine the control buttons at the left of the plot. The top two buttons rotate the plot up and down. The next two buttons rotate the plot left and right. The last two buttons rotate the plot clockwise and counter-clockwise. You can use these buttons by clicking, pressing, Shift-clicking, and Ctrl-clicking.

⇒ **Click the top rotation button and release it.**

The plot rotates a small increment and stops when you release the button.

⇒ **Press the clockwise rotation button and hold it down.**

The plot rotates clockwise as long as you hold the button down.

⇒ **Press the Shift key and click any of the buttons.**

The plot rotates continuously until you click another button.

⇒ **Press the Ctrl key and click any of the buttons.**

This also rotates the plot continuously until you click another button.

Below the directional buttons is a slider to control the speed of rotation. When the slider is at the top, rotation is at maximum speed.

⇒ **Drag the slider, then try the control buttons again to rotate at different speeds.**

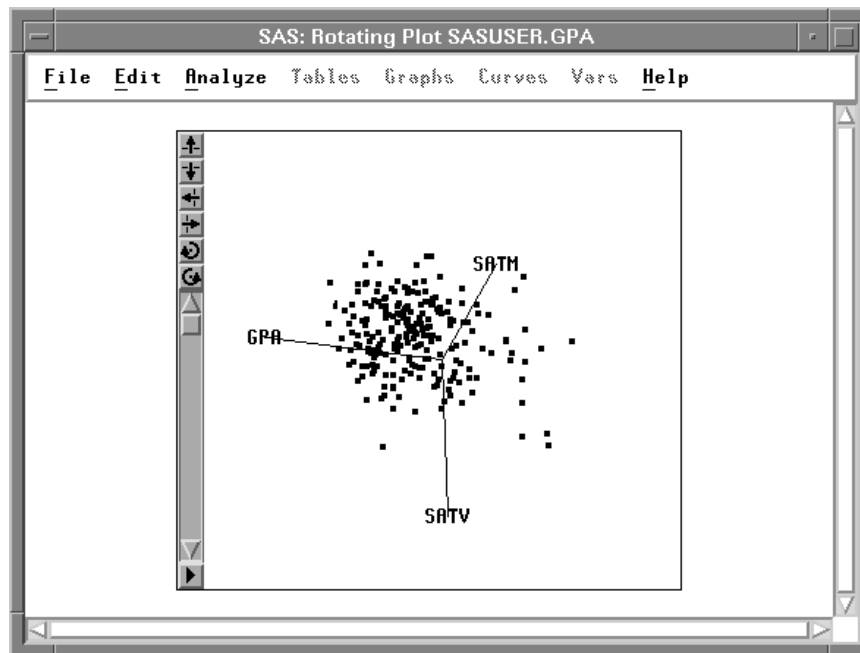


Figure 6.5. Slider at Maximum

The buttons offer precise control of rotation, but the hand tool offers greater flexibility. Using the hand tool, you can rotate about any axis.

⇒ Choose **Edit:Windows:Tools** to display the tools window.

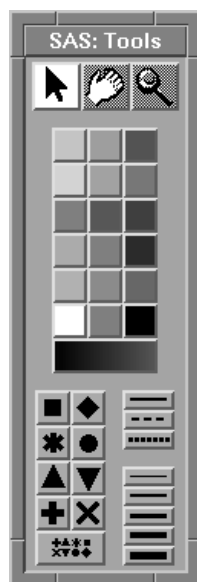


Figure 6.6. Tools Window

⇒ Click the **Hand** tool at the top of the Tools window.

The cursor changes to a hand.

⇒ **Click and drag the hand in the rotating plot.**

When you use the hand tool, the plot acts as a freely rotating sphere. When you click with the hand, the plot rotates a small increment. When you drag the hand, the plot follows your motion. The plot rotates as long as you press the mouse button and hold it down. If you release the button while you are dragging the hand, the plot continues rotating in the direction you were dragging.

You can use the hand without displaying the Tools window. The hand is active in each corner of the plot.

⇒ **Click the Arrow tool at the top of the Tools window.**

The cursor changes to an arrow.

⇒ **Move the Arrow tool to any corner of the rotating plot.**

The cursor changes to a hand. Click or drag the hand to rotate the plot.

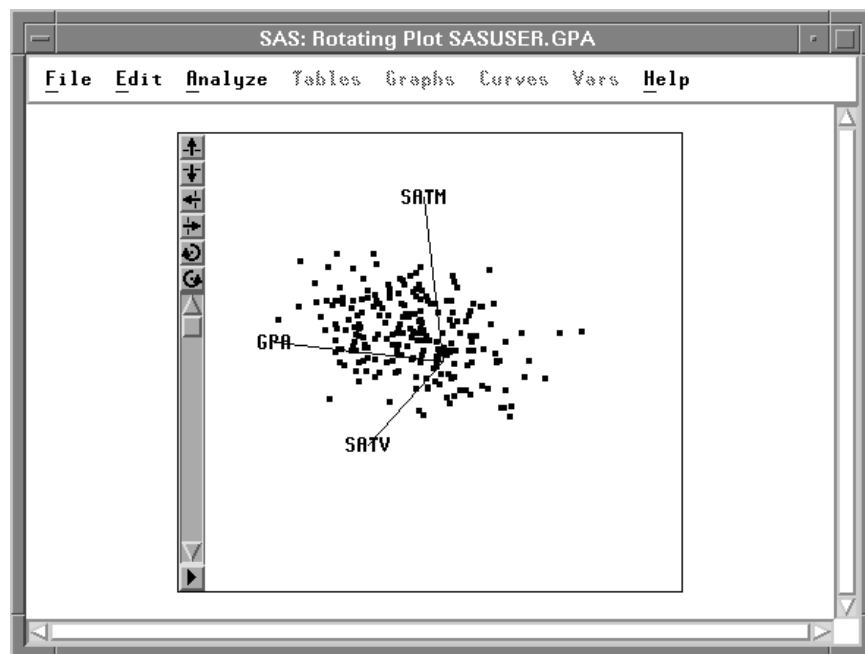


Figure 6.7. Hand Tool

⇒ **Click on the button in the lower left corner of the plot.**

This calls up the rotating plot pop-up menu. You can customize the appearance of the rotating plot with the choices on this menu.

⇒ **Choose Cube.**

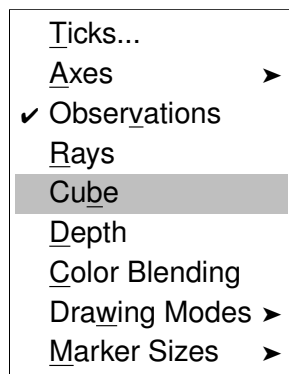


Figure 6.8. Rotating Plot Pop-up Menu

This draws a cube around the point cloud. The cube shows the range of the data and aids in maintaining visual orientation.

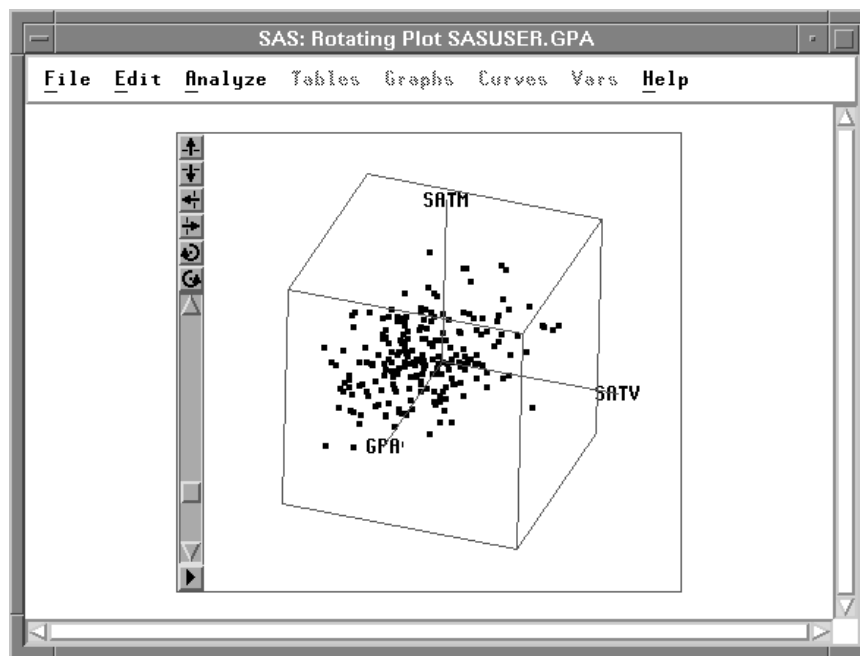


Figure 6.9. Rotating Plot with Bounding Cube

⇒ **Choose Depth from the pop-up menu.**

This draws distant markers smaller than near markers to serve as a visual cue for depth perception.

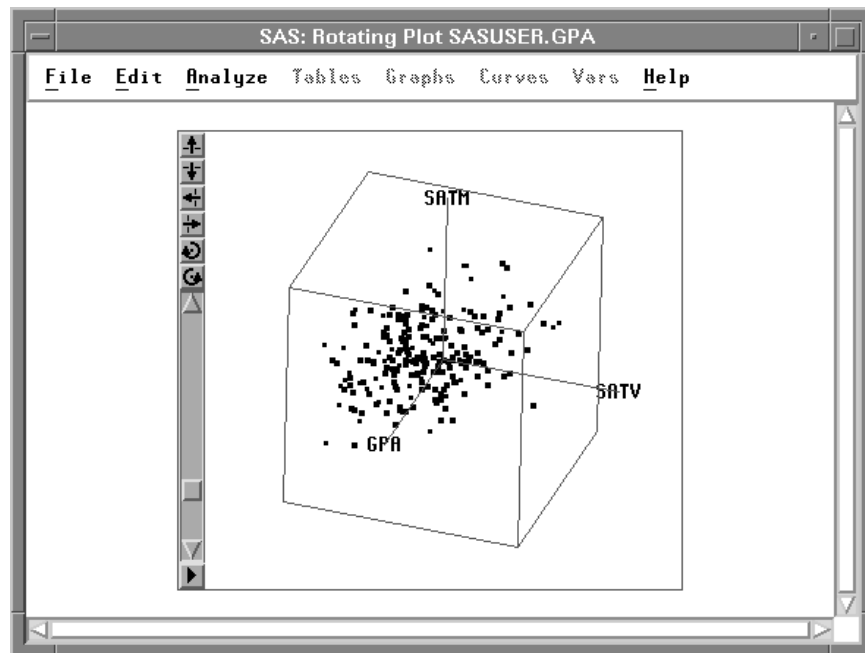


Figure 6.10. Depth Cueing

Both the **Cube** and the **Depth** choices serve as toggles, so you can choose them again to remove the cube or to return all markers to the same size. You can use other choices on the pop-up menu to toggle the display of observations and rays and to set ticks, axes, and marker sizes.

You can create a matrix of rotating plots just as you created a matrix of scatter plots in the preceding chapter. If you select more than three variables in the data window and then choose **Analyze:Rotating Plot (Z Y X)**, you create a matrix containing one rotating plot for every unique combination of three variables.

You can also identify observations in rotating plots just as in other plots. Click once on an observation marker to select it and to see its label. Double-click on an observation marker to display the examine observations dialog.

⊕ **Related Reading:** Rotating Plots, [Chapter 37](#).

Rotating Plot with Fitted Surface

When you suspect that the values of one variable may be predicted by the values of two other variables, you can choose to fit a *response surface* to your data.

Follow these steps to explore how dust concentration varies with the wind speed and with the time of day in the **AIR** data set.

⇒ **Open the AIR data set.**

⇒ **Choose Analyze:Rotating Plot (Z Y X).**

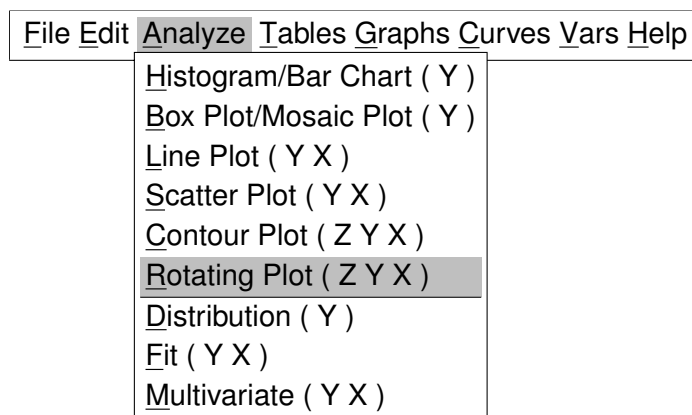


Figure 6.11. Creating a Rotating Plot with Fitted Surface

A rotating plot variables dialog appears, as shown in [Figure 6.12](#).

⇒ **Select DUST in the variables list at the left. Then click Z.**

This assigns the **Z** role to the **DUST** variable. Similarly, assign **HOUR** the **Y** role and **WIND** the **X** role.

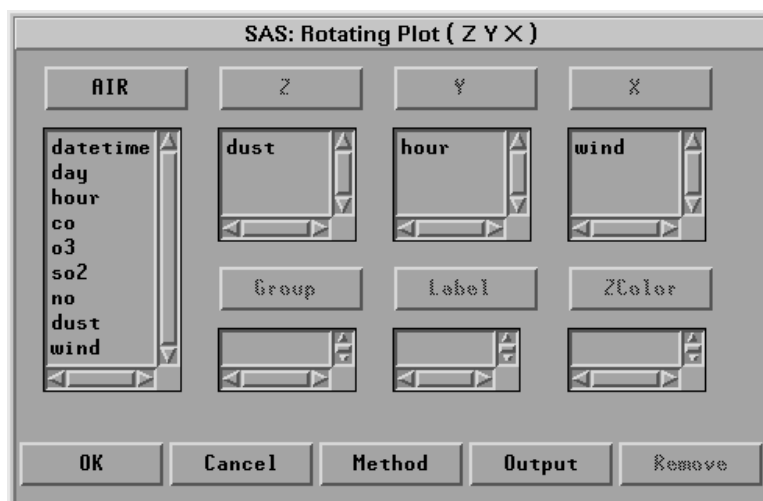


Figure 6.12. Rotating Plot Variables Dialog

⇒ **Click Output to display the Output dialog, as shown in [Figure 6.13](#).**

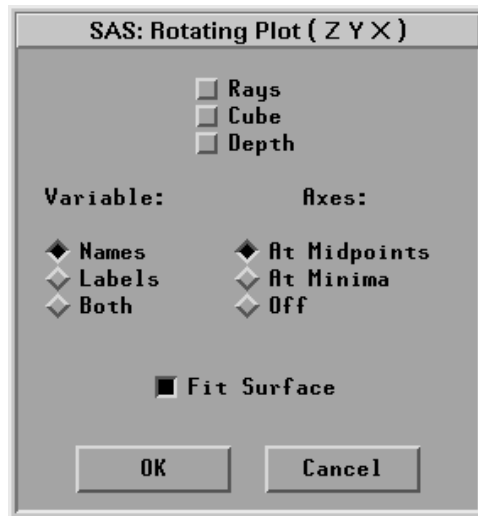


Figure 6.13. Output Dialog for Rotating Plot

⇒ **Select Fit Surface and click OK.**

⇒ **Click Method to display the Method dialog, as shown in Figure 6.14.**

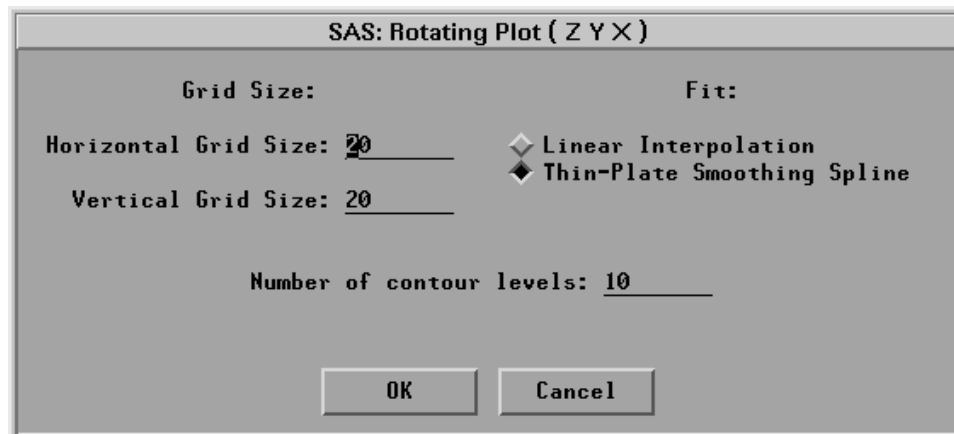


Figure 6.14. Method Dialog for Rotating Plot

⇒ **Select Fit:Thin-Plate Smoothing Spline and click OK.**

⇒ **Click OK to create a surface plot.**

⇒ **Click on the menu button in the lower left corner of the plot.**
 Choose **Drawing Modes:Smooth Color** and **Axes:At Minima**.

⇒ **Rotate the plot as described in the previous section.**

You see a surface that models the response of dust concentration as a function of the wind speed and the time of day.

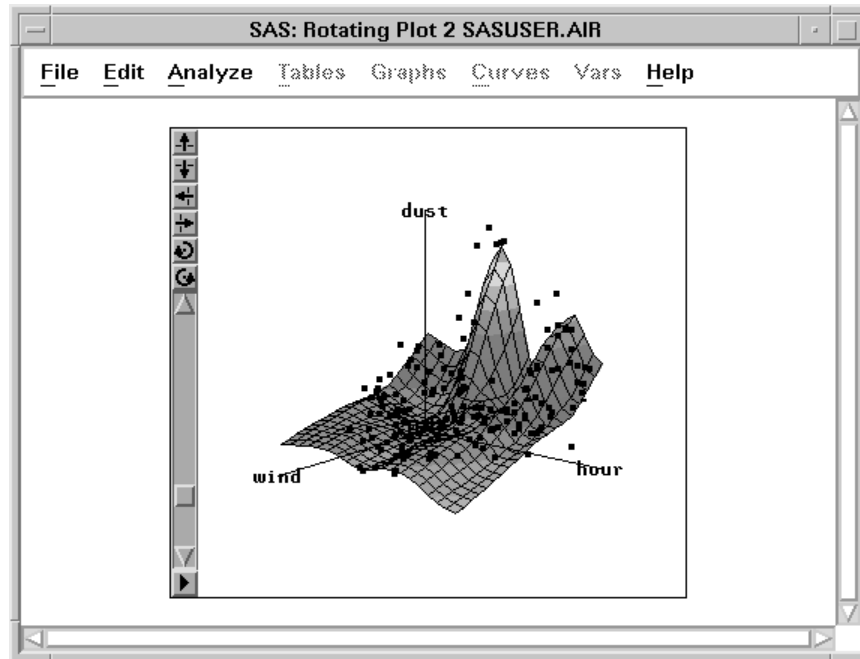


Figure 6.15. Rotating Plot with Fitted Surface

Contour Plots

The contour plot provides an alternative graphical method for examining the variations of a response surface. The contour plot displays the geometric features of the response surface as a family of contours or *level sets* lying in the domain of the predictor variables.

If the **AIR** data set is not already open, open it now.

⇒ **Choose Analyze:Contour Plot (Z Y X).**

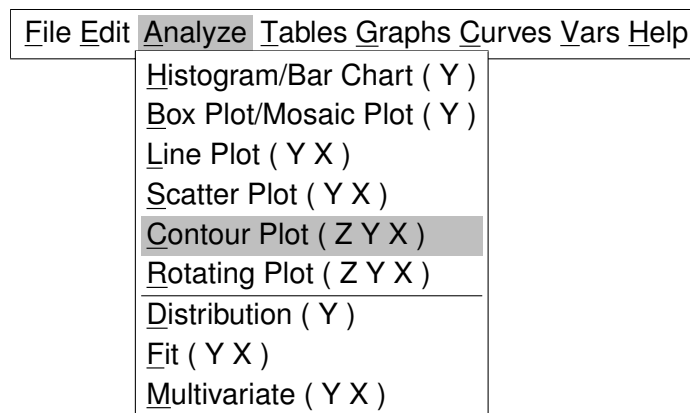


Figure 6.16. Creating a Contour Plot

A contour plot variables dialog appears, as shown in [Figure 6.17](#).

- ⇒ Assign the **Z** role to the **DUST** variable, assign **HOUR** the **Y** role, and assign **WIND** the **X** role.

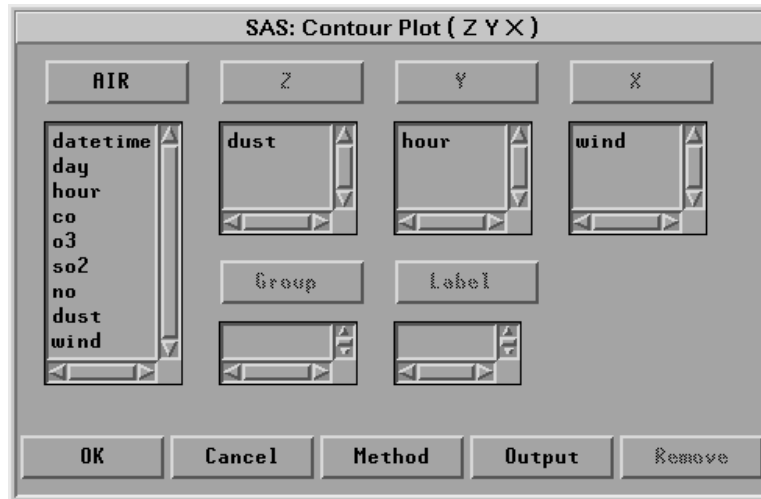


Figure 6.17. Contour Plot Variables Dialog

- ⇒ Click **Method** to display the Method dialog.
This dialog looks exactly like the Method dialog for the rotating plot, as shown in [Figure 6.14](#).
- ⇒ Select **Fit:Thin-Plate Smoothing Spline** and click **OK**.
- ⇒ Click **OK** to create a contour plot.
- ⇒ Click on the menu button in the lower left corner of the plot. Choose **Observations**.

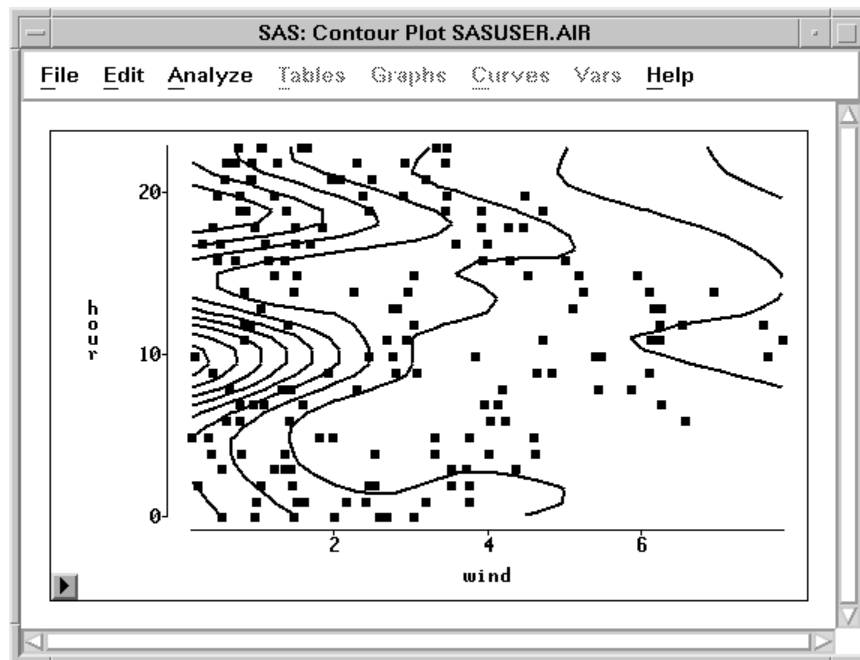


Figure 6.18. Contour Plot

By default, the contour lines of the response surface are evenly spaced in the units of the response variable. For this example, each contour represents about 1.3 units of change in the dust concentration. Note that regions where the contour lines are close together indicate regions in which small changes in the wind speed or the time of day will lead to relatively large changes in the modeled response for dust.

The response model indicates that peak dust concentrations for this data primarily occur when there are only gentle winds during the mid-morning and late afternoon. To see if this prediction qualitatively fits the **AIR** data set, you can examine the observations with high dust values.

⇒ **Select Edit:Observations:Find .**

The **Find Observations** dialog appears.

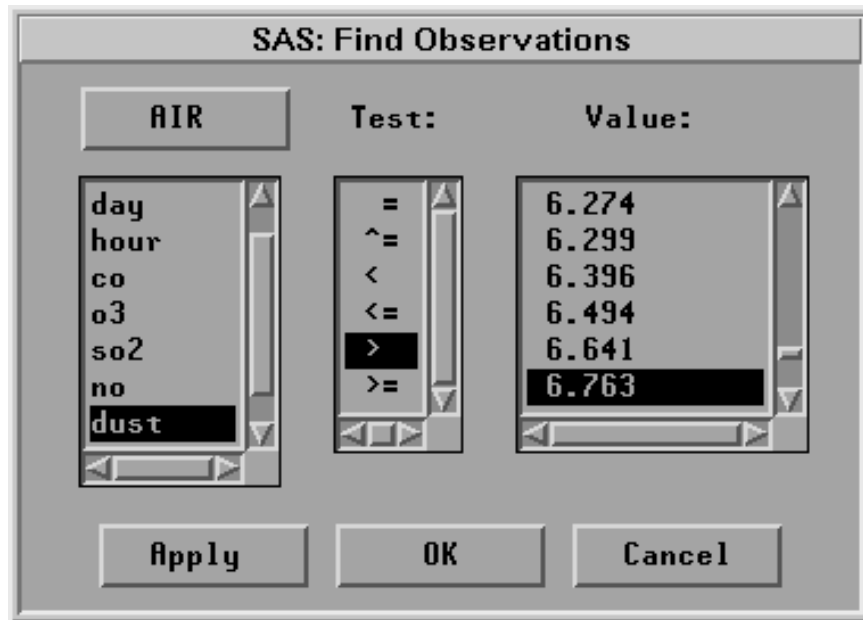


Figure 6.19. Find Observations dialog

- ⇒ Select **DUST** in the left-hand column, the greater-than test (>) in the middle column, and the value 6.763 in the right-hand column.
 This selects all observations that have dust values greater than 6.763.

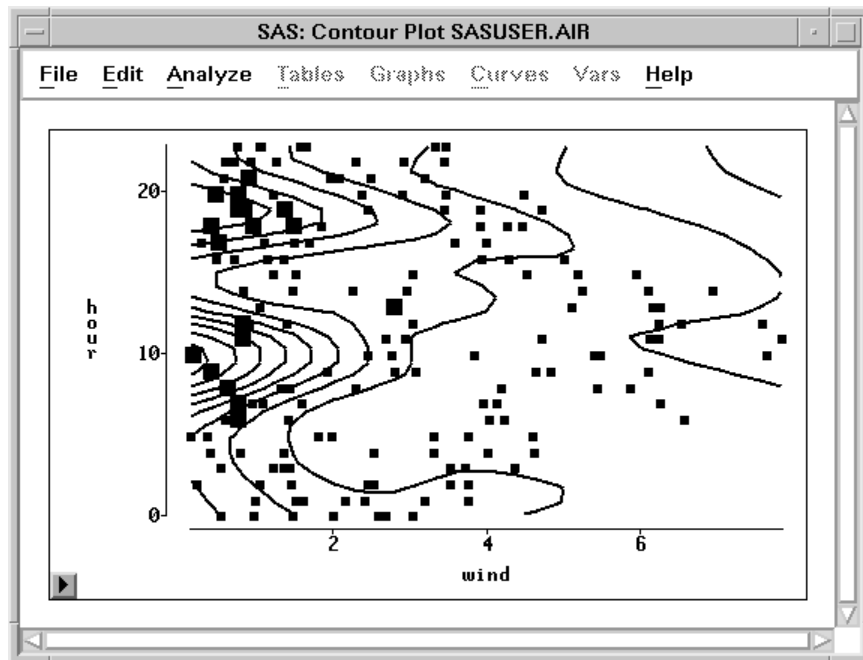


Figure 6.20. Selecting High **DUST** Values

Techniques ♦ *Exploring Data in Three Dimensions*

All but one of the selected observations occur in the mid-morning or late afternoon on days with light winds. However, note that there are also observations in those regions that have small dust concentration values.

Consult [Chapter 39, “Fit Analyses,”](#) to determine whether a model response surface provides a good quantitative fit to your data.

- ⊕ **Related Reading:** Contour Plots, [Chapter 36](#).
- ⊕ **Related Reading:** Fit Analysis, [Chapter 39](#).

Chapter 7

Adjusting Axes and Ticks

Chapter Contents

ADJUSTING TICKS	126
ADJUSTING 2D AXES	129
ADJUSTING 3D AXES	131

Chapter 7

Adjusting Axes and Ticks

With SAS/INSIGHT software, you have control over the appearance of axes. In all graphs, you can specify major and minor tick marks. In two-dimensional graphs, you can adjust axis position dynamically. In three-dimensional graphs, you can place axes at the center or the minimum of the data range.

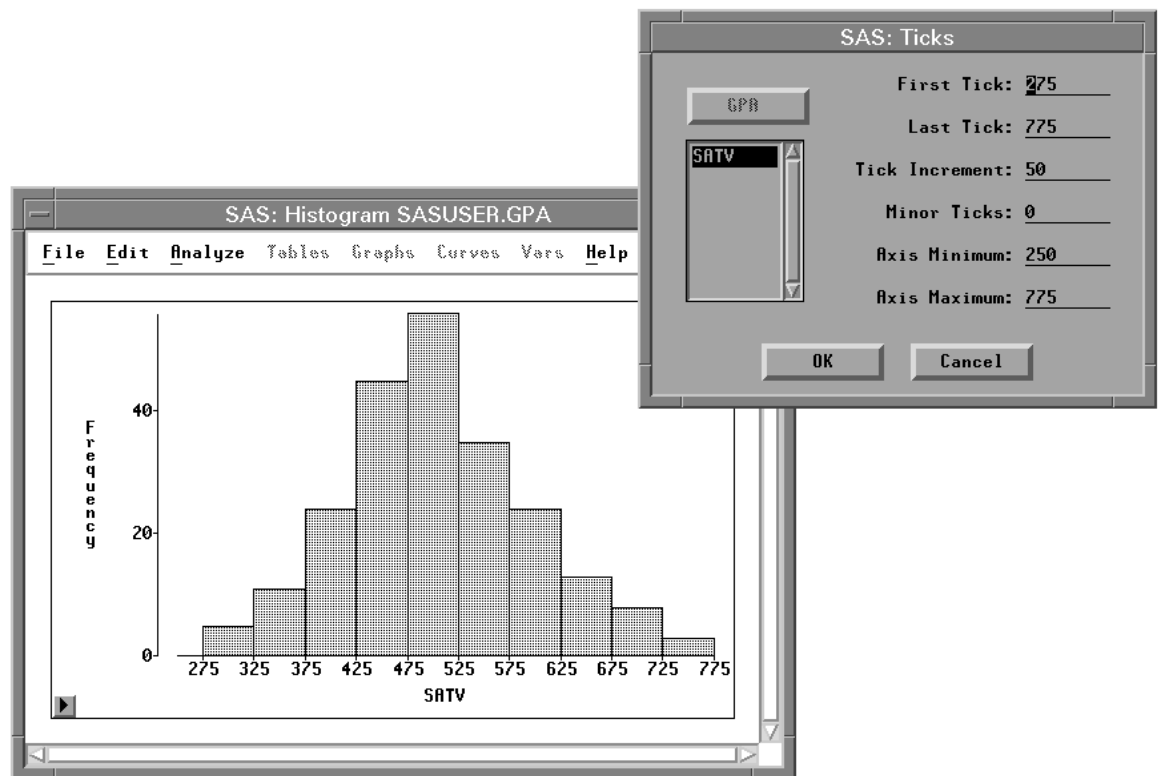


Figure 7.1. Adjusting Histogram Ticks

Adjusting Ticks

Major tick marks have an associated tick label, if space permits. *Minor* tick marks are smaller marks evenly spaced between the major tick marks. By default, the number of minor tick marks is 0.

You can change the default tick marks in a histogram of verbal SAT scores by following these steps.

- ⇒ **Open the GPA data set and create a histogram of verbal SAT scores.**
- ⇒ **Select the variable on the axis of interest.**

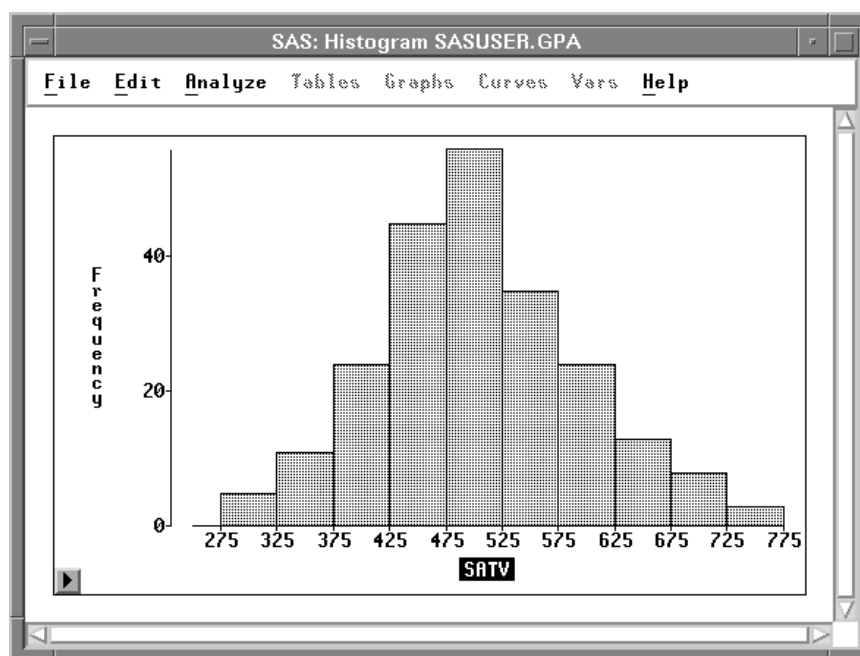


Figure 7.2. Selecting Variable **SATV**

- ⇒ **Click on the button in the lower left corner to display the histogram pop-up menu.**
Choose **Ticks** from the pop-up menu to display the Ticks dialog.

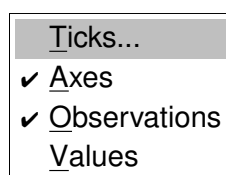


Figure 7.3. Histogram Pop-up Menu

Figure 7.4 shows the Ticks dialog for the **SATV** axis in the histogram.



Figure 7.4. Ticks Dialog

⇒ **Change the values in the Ticks dialog.**

Set the first tick to 200, the last tick to 800, the axis minimum to 175, and the axis maximum to 825.

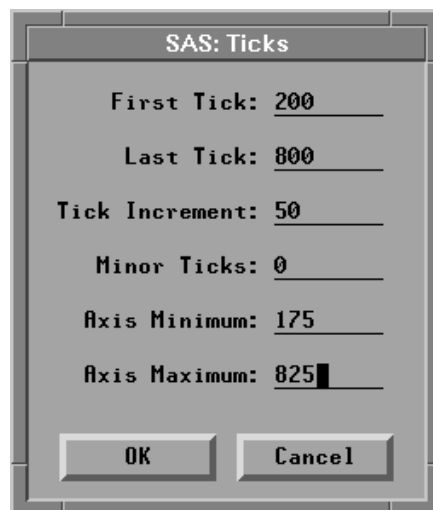


Figure 7.5. Changing Ticks

⇒ **Click OK to redraw the histogram with the new tick specifications.**

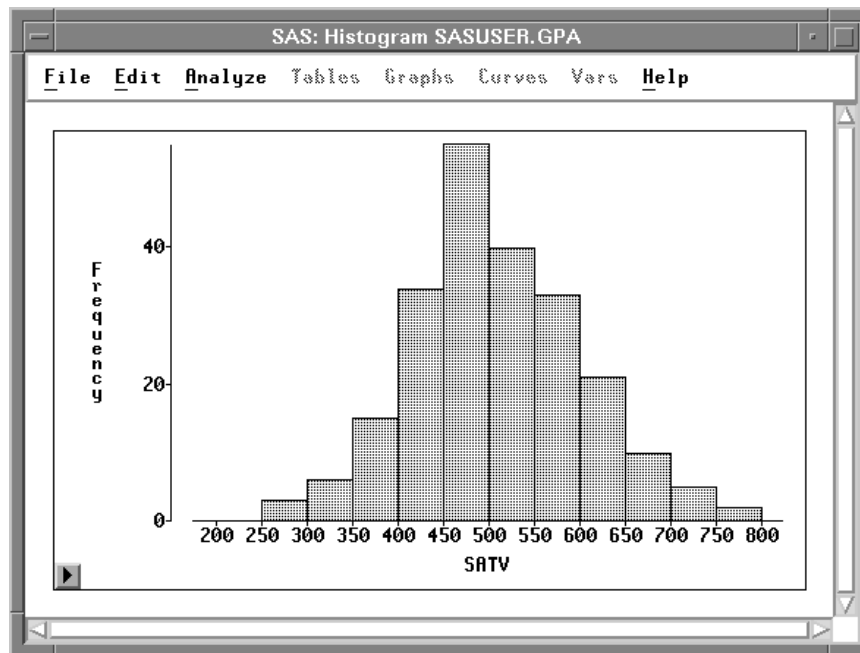


Figure 7.6. Histogram with New Ticks

You can use the **Ticks** dialog similarly to scale axes in all other two-dimensional and three-dimensional graphs.

Adjusting 2D Axes

You can adjust horizontal and vertical axes in all two-dimensional graphs. For example, [Figure 7.7](#) shows tick labels truncated because the axis does not have space to show them completely. To increase the axis space, point to the axis with the mouse. Note that the cursor changes to a hand when it is positioned over the axis.

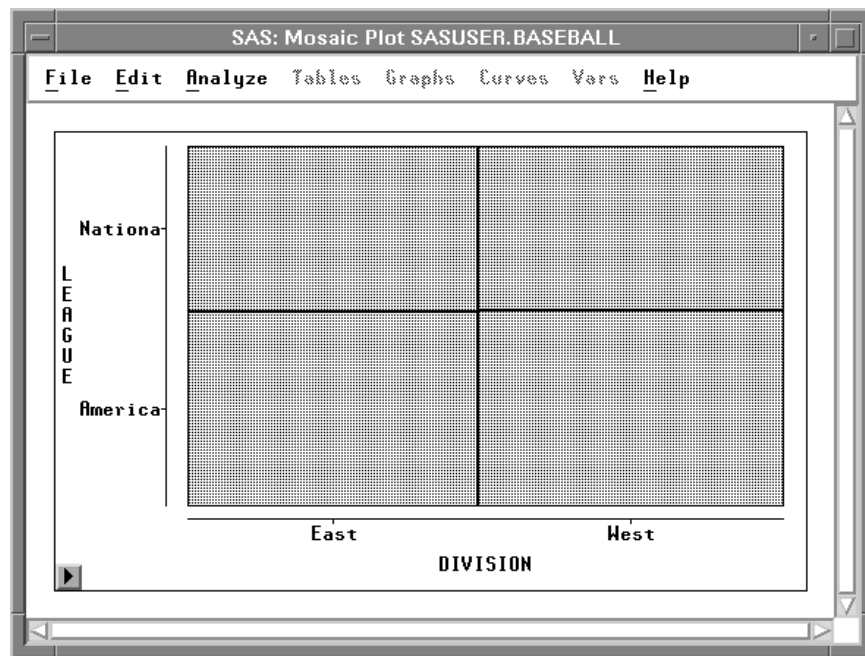


Figure 7.7. Adjusting an Axis

Press the mouse button and drag the axis to a new position. When you release the mouse button, the axis moves to its new position.

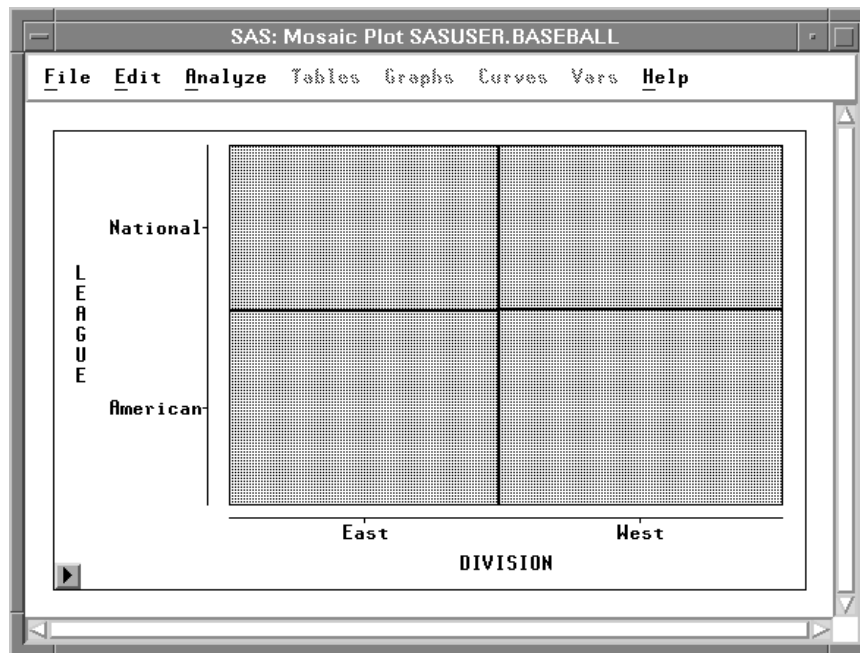


Figure 7.8. Axis at New Position

Adjusting 3D Axes

The rotating plot pop-up menu provides control over the position of the axes. Display the pop-up menu and choose from the **Axes** submenu.

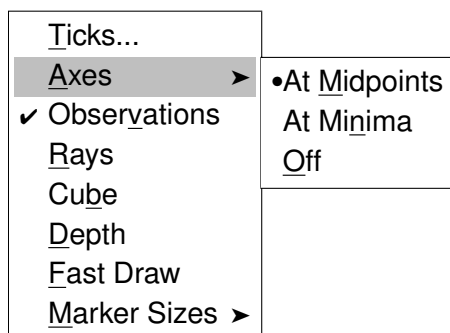


Figure 7.9. Rotating Plot Pop-up Menu

If you are doing exploratory work and are primarily interested in the shape of the point cloud, choose **Axes:At Midpoints** to display the axes centered in the plot. This display minimizes interference of the axes with your view of the data, in part because tick marks and tick labels are not displayed.

Choose **Axes:At Minima** to display axes at the minimum data values if you have spatial data and are interested in observation positions. These axes span the range of the data. All tick marks and tick labels are also displayed.

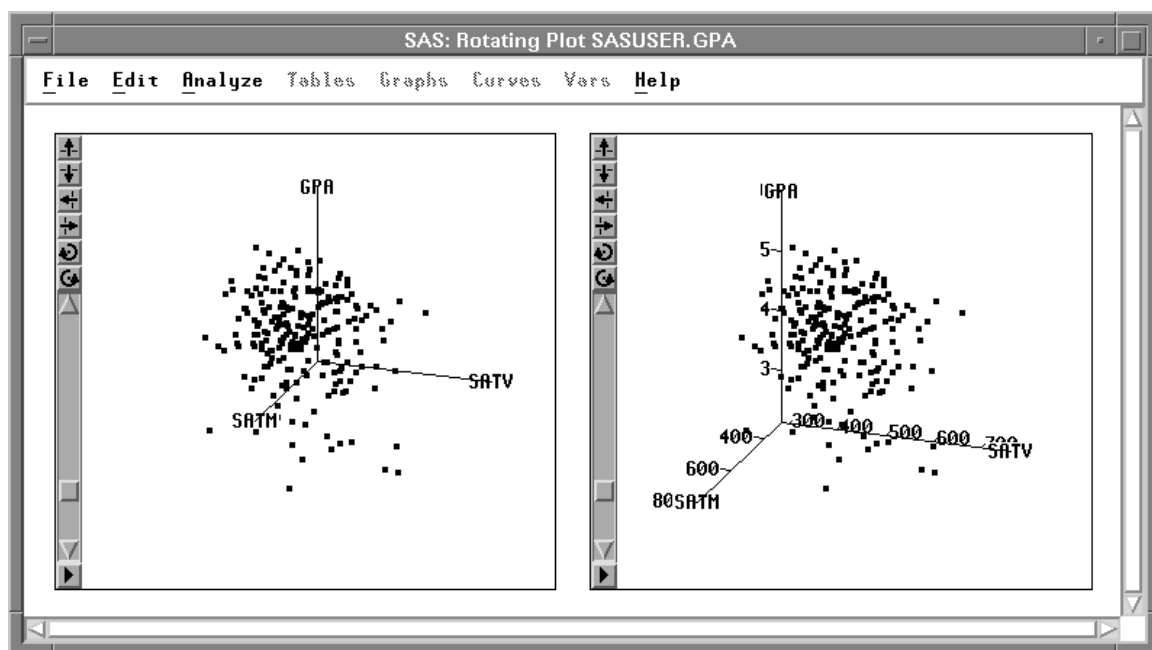


Figure 7.10. Axes at Midpoints and at Minima

Axes:At Midpoints is the default setting. To change the default, click the **Output**

Techniques ♦ *Adjusting Axes and Ticks*

button in the Rotating Plot Variables dialog and set the **Axes:At Minima** option. Choose **File:Save:Options** to save your options.

Chapter 8

Labeling Observations

Chapter Contents

TEMPORARY AND PERMANENT LABELS	136
USING LABEL VARIABLES	139
SETTING A DEFAULT LABEL VARIABLE	141

Chapter 8

Labeling Observations

Labels identify observations in plots. You can label observations by number or by the value of a variable. You can assign temporary or permanent labels.

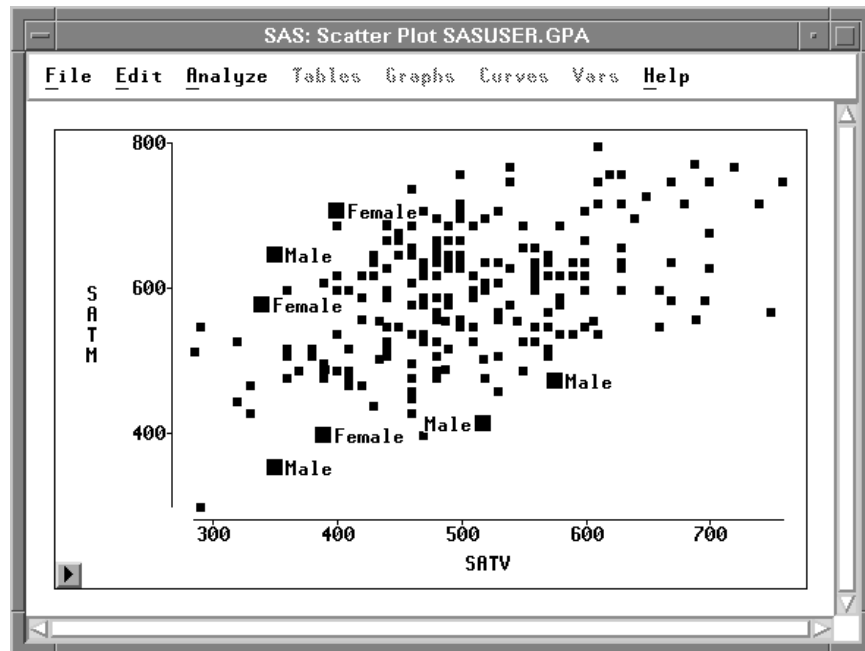


Figure 8.1. Labeling Observations

Temporary and Permanent Labels

When you click on an observation, you display its temporary label. To see this, follow these steps.

- ⇒ **Open the GPA data set.**
- ⇒ **Choose `Analyze:Scatter Plot (Y X)`.**
This displays a scatter plot variables dialog, as shown in [Figure 8.2](#).
- ⇒ **Select `SATM` and `SATV` as `X` variables and `GPA` as the `Y` variable.**

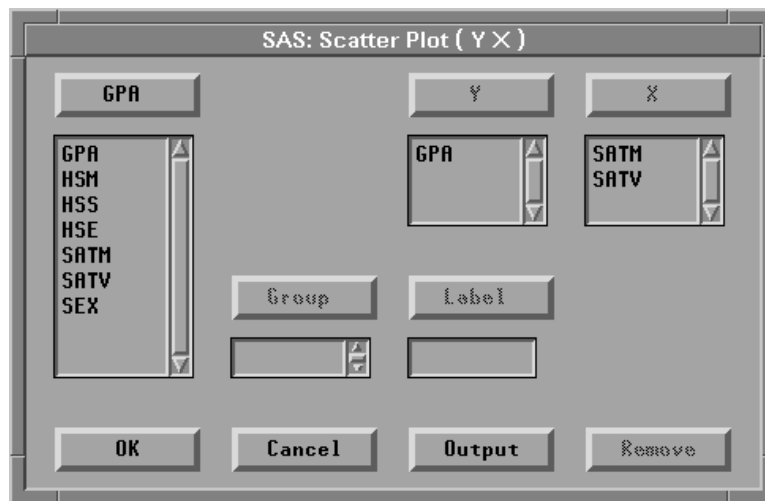


Figure 8.2. Scatter Plot Variables Dialog

- ⇒ **Click the `OK` button.**
This creates two scatter plots, as shown in [Figure 8.3](#).
- ⇒ **Click on an observation in one of the plots.**
The observation is highlighted in both plots, and a label appears beside the observation in the plot in which you clicked. This label is temporary; it disappears when you deselect the observation.

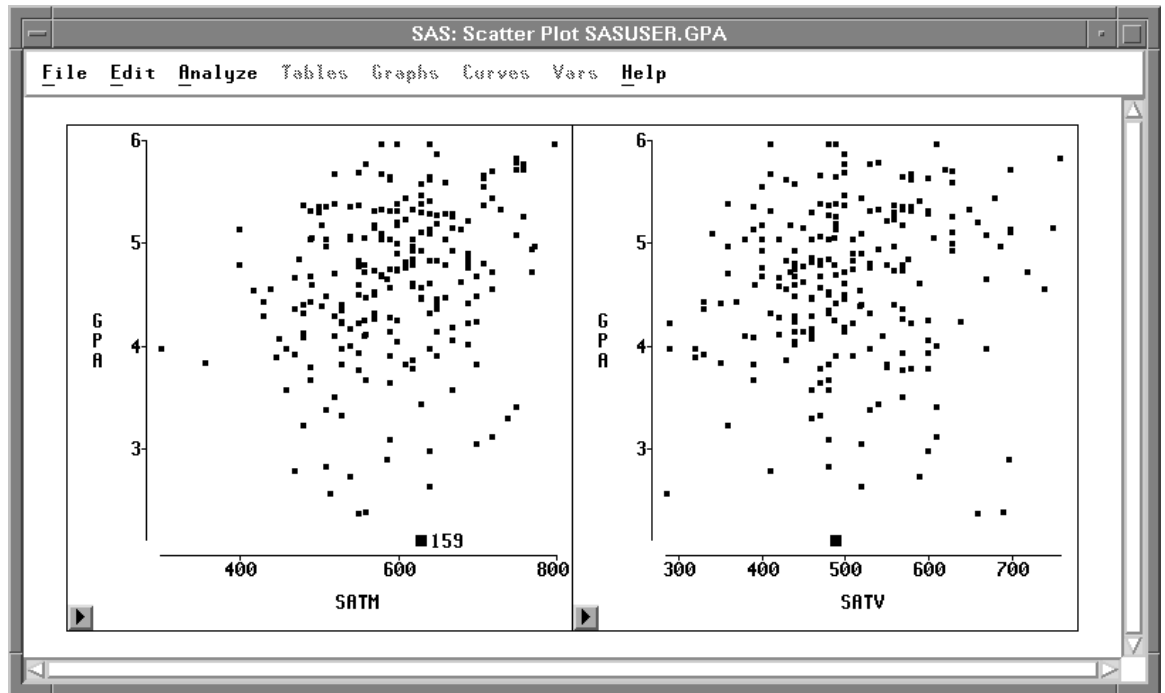


Figure 8.3. Temporary Label

You can turn this label into a permanent label.

⇒ **Choose Edit:Observations:Label in Plots.**

This labels the observation in all plots, and the label remains if you deselect the observation.

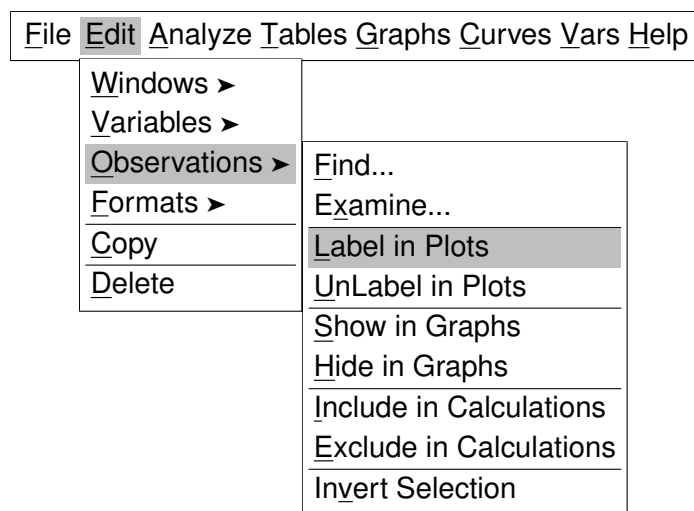


Figure 8.4. Edit: Observations Menu

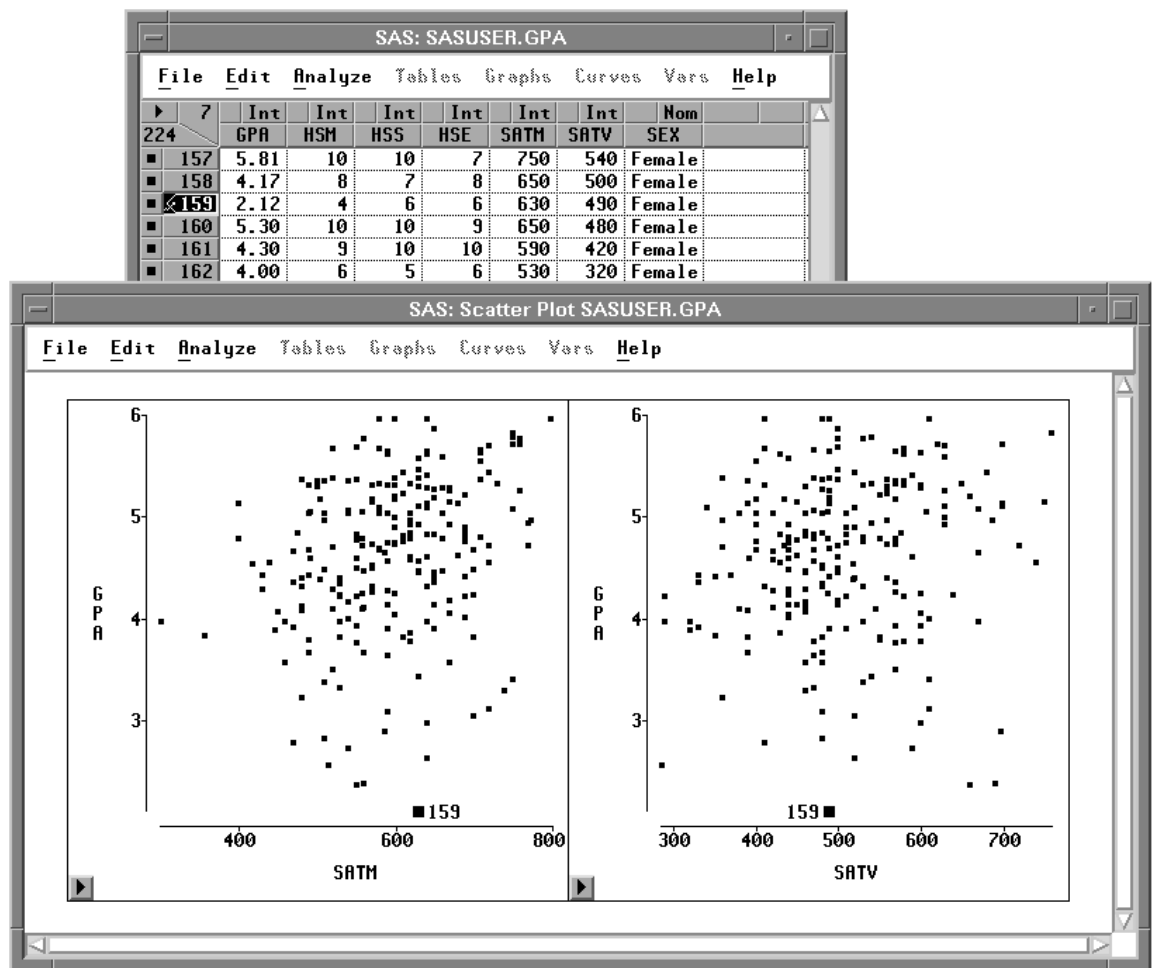


Figure 8.5. Permanently Labeled Observations

Notice in the data window that the observation is displayed with a picture of a label. This indicates that a label will always be displayed for this observation in all plots.

If you change your mind, you can remove the permanent label by choosing **Edit:Observations:UnLabel in Plots**.

Using Label Variables

SAS/INSIGHT software shows the observation number as the label by default. You can choose a variable to supply the label text by specifying a *label variable*.

⇒ Choose **Edit:Windows:Renew** to redisplay the scatter plot variables dialog.

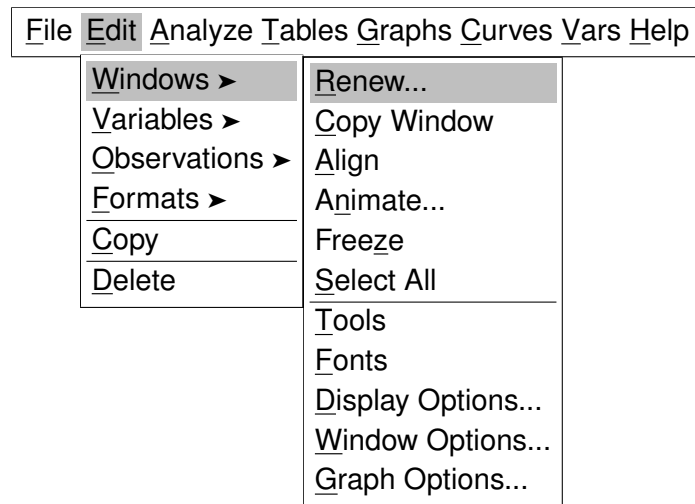


Figure 8.6. Edit:Windows Menu

⇒ In the dialog, select **SEX** and then click the **Label** button.

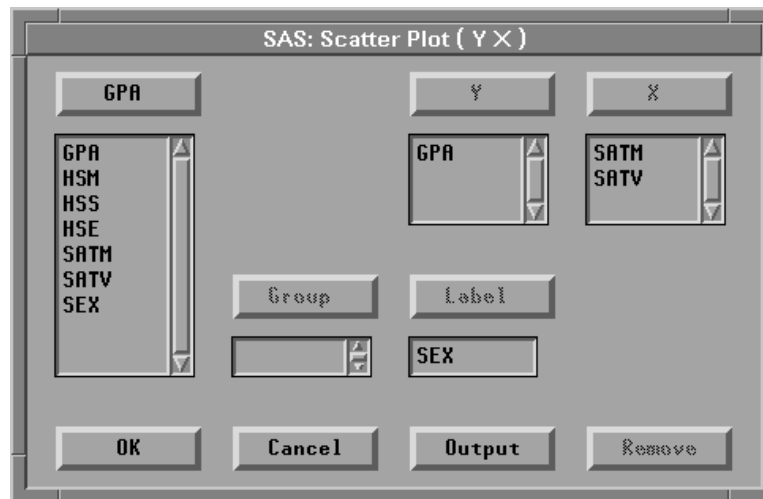


Figure 8.7. Assigning **Label** Role

Techniques ♦ Labeling Observations

⇒ Click the **OK** button.

Now the value of **SEX**, instead of the observation number, labels the observation.

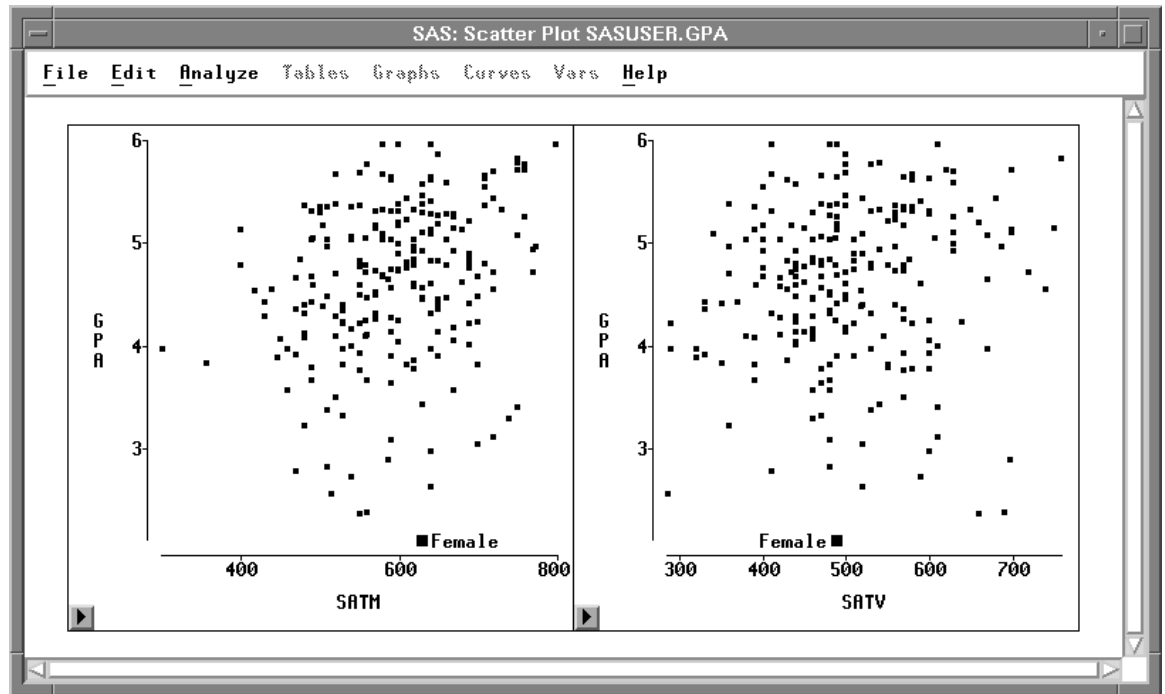


Figure 8.8. Observation Labeled by **SEX**

Setting a Default Label Variable

In addition to specifying label variables for individual plots, you can specify a label variable that will automatically be used in all future plots.

⇒ **Click on the upper left corner of the variable **SEX** in the data window.**
This displays a pop-up menu. Choose **Label** from the pop-up menu.

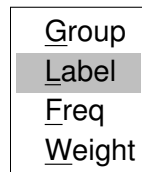


Figure 8.9. Variable Role Pop-up Menu

Now **SEX** is the default label variable, and future plots will use **SEX** for labels. In the data window, the word **Label** appears above the variable name, as shown in [Figure 8.10](#).

SAS: SASUSER.GPA

File Edit Analyze Tables Graphs Curves Vars Help

7	Int	Int	Int	Int	Int	Int	Label	Nom	
224	GPA	HSM	HSS	HSE	SATM	SATV	SEX		
1	5.32	10	10	10	670	600	Female		
2	5.14	9	9	10	630	700	Male		
3	3.84	9	6	6	610	390	Female		
4	5.34	10	9	9	570	530	Male		
5	4.26	6	8	5	700	640	Female		
6	4.35	8	6	8	640	530	Female		
7	5.33	9	7	9	630	560	Male		
8	4.85	10	8	8	610	460	Male		
9	4.76	10	10	10	570	570	Male		
10	5.72	7	8	7	550	500	Female		
11	4.08	9	10	7	670	600	Female		
12	5.38	8	9	8	540	580	Female		

Figure 8.10. Label Variable Role

⊕ **Related Reading:** Variable Roles, [Chapter 31, “Data Windows.”](#)

Chapter 9

Hiding Observations

Chapter Contents

HIDING INDIVIDUAL OBSERVATIONS	146
TOGGLING THE DISPLAY OF OBSERVATIONS	149
SLICING	153

Chapter 9

Hiding Observations

You can *hide* observations to prevent them from appearing in graphs. You can *toggle* the display of observations to keep them from appearing in a graph unless they are selected. You can *slice* observations by dynamically toggling their display. These techniques are useful for adjusting the range of data displayed and for showing subsets of your data.

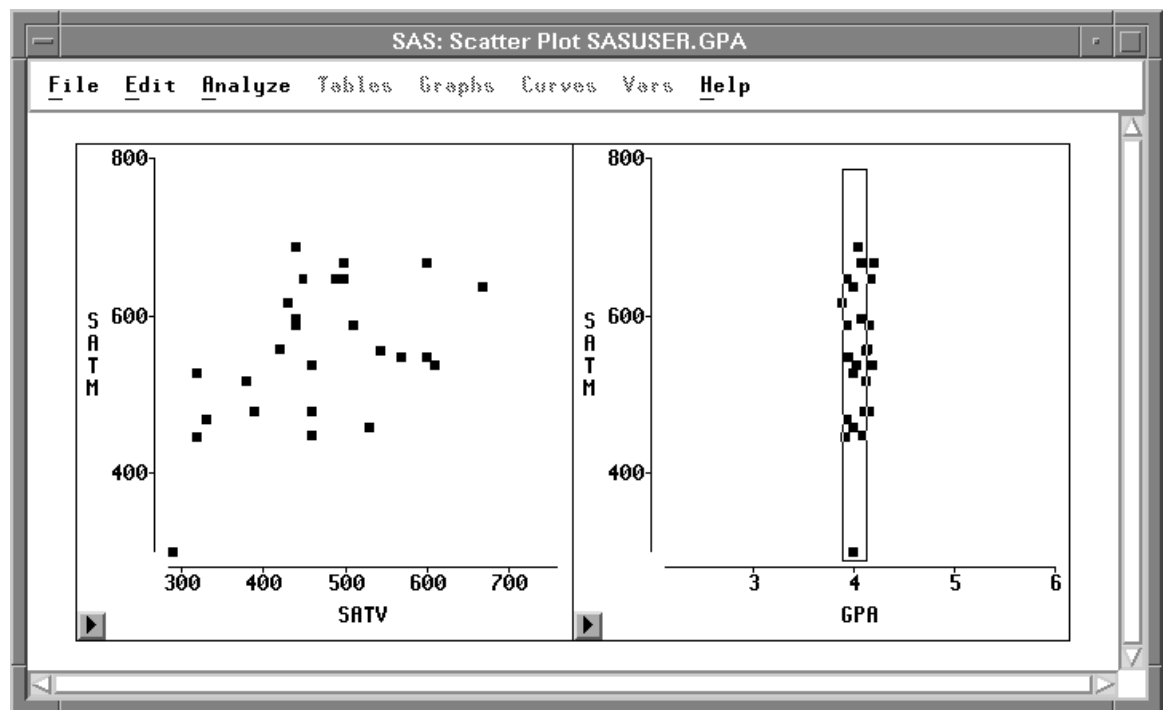


Figure 9.1. Slicing Observations

Hiding Individual Observations

You can adjust the range of data displayed and show subsets of your data by hiding observations.

† **Note:** Hiding observations in graphs does not exclude them from calculations. To exclude observations from calculations, see [Chapter 21, “Comparing Analyses.”](#)

⇒ **Open the GPA data set.**

⇒ **Create a scatter plot of SATM versus SATV.**

Use the techniques described in [Chapter 5, “Exploring Data in Two Dimensions.”](#)

⇒ **Select the two observations with values of SATM below 400.**

Use extended selection or drag a rectangle around both observations.

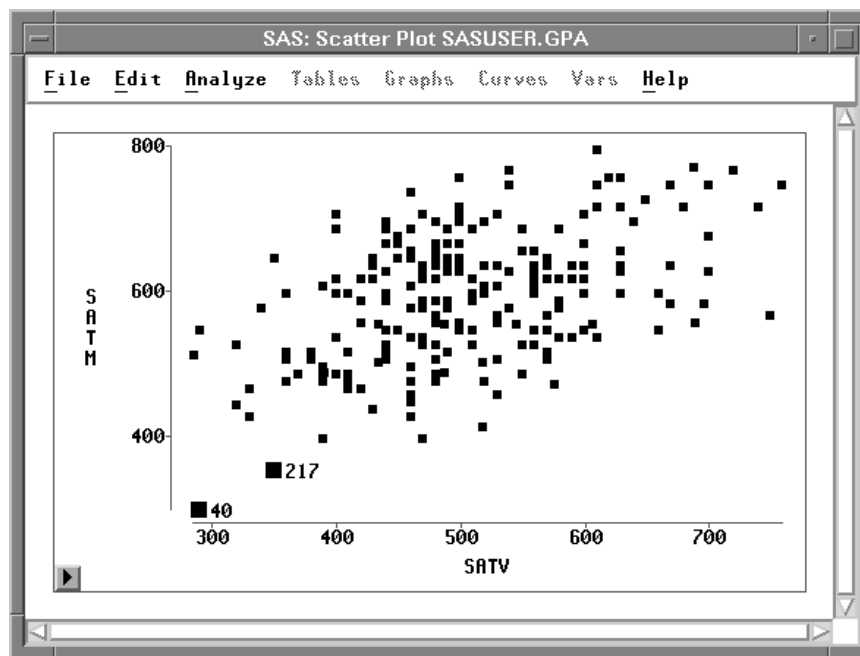


Figure 9.2. Observations Selected

⇒ **Choose Edit:Observations:Hide in Graphs.**

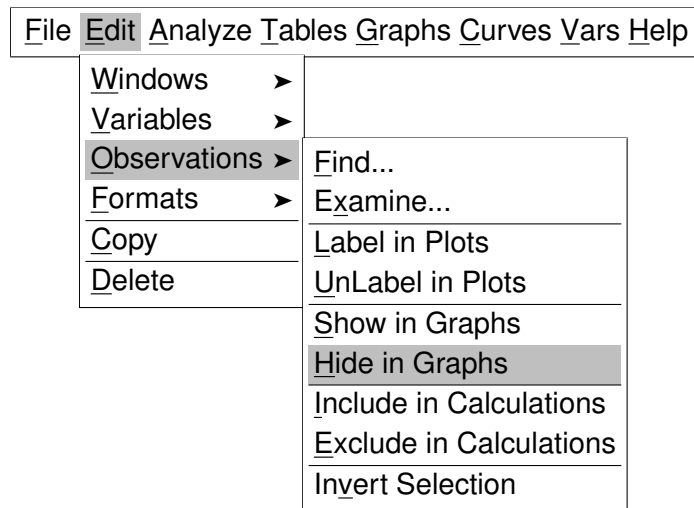


Figure 9.3. Edit: Observations Menu

This causes the selected observations to disappear from the graph. The graph rescales automatically. The new **SATM** axis starts at 400.

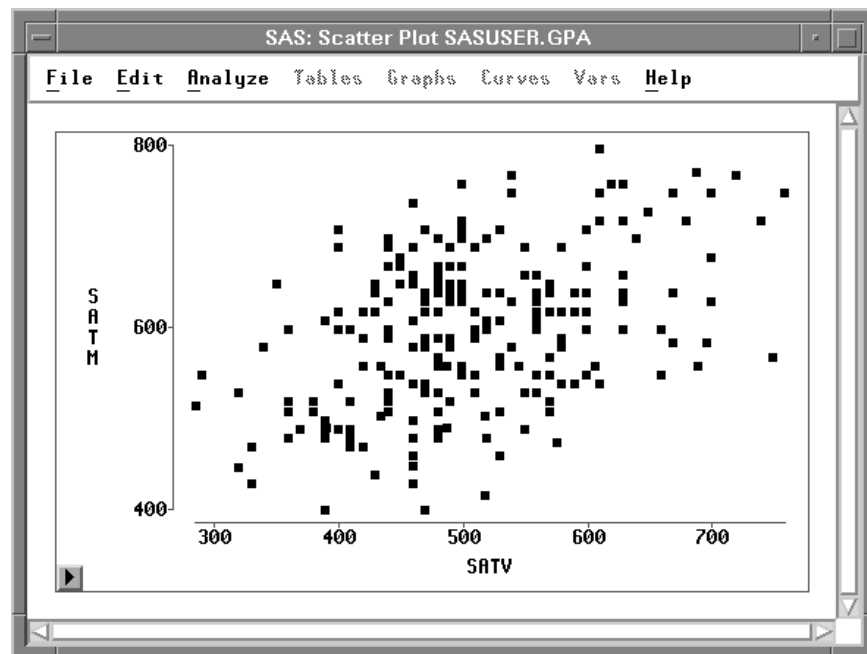


Figure 9.4. Observations Hidden

⇒ **Choose Find Next from the data window pop-up menu.**

This scrolls to the next selected observation and shows that the hidden observation has no marker. The absence of the marker in the data window indicates that the observation is hidden in *all* graphs.

SAS: SASUSER.GPA									
File		Edit	Analyze	Tables	Graphs	Curves	Vars	Help	
▶	7	Int	Int	Int	Int	Int	Int	Nom	
224		GPA	HSM	HSS	HSE	SATM	SATV	SEX	
	40	4.00	2	4	6	300	290	Male	
■	41	3.43	10	9	9	750	610	Female	
■	42	4.48	8	9	6	650	460	Female	
■	43	5.73	10	10	9	720	630	Female	
■	44	4.43	7	10	10	530	560	Female	
■	45	3.69	7	6	7	560	480	Male	
■	46	5.80	10	10	9	760	500	Female	
■	47	5.18	10	10	10	570	750	Male	
■	48	6.00	9	10	10	640	480	Female	
■	49	6.00	9	9	8	800	610	Female	
■	50	4.00	9	6	5	640	670	Female	
■	51	5.06	9	10	9	590	420	Male	

Figure 9.5. Data Window after Hiding Observations

⇒ Choose **Edit:Observations:Show in Graphs**.

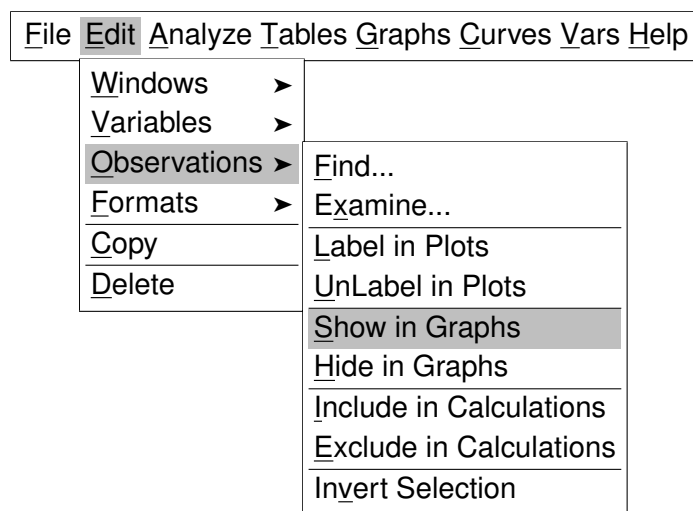


Figure 9.6. Edit: Observations Menu

This makes the observations visible again. The scatter plot rescales.

Toggling the Display of Observations

You can show subsets of your data by toggling the display of observations. This causes observations to be displayed only when they are selected.

- ⇒ **Deselect all observations by clicking in any open area of a graph.**
- ⇒ **Choose **Edit:Windows:Renew** to redisplay the scatter plot variables dialog.**

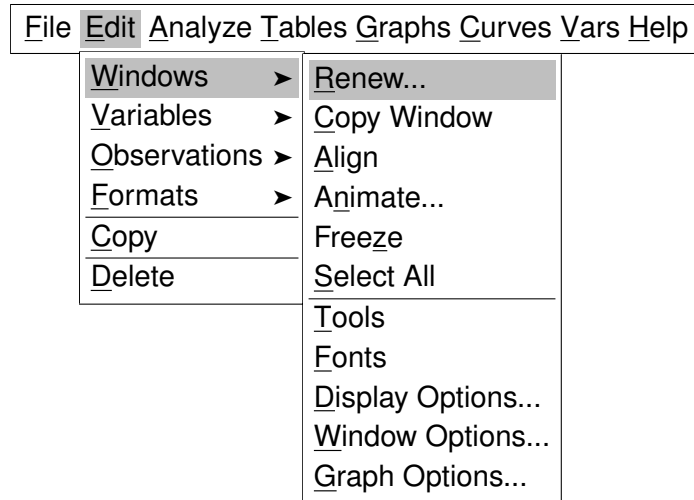


Figure 9.7. Edit:Windows Menu

- ⇒ **Click on **GPA** in the variables list and then click on the **X** button.**
This adds **GPA** to the **X** variables list.

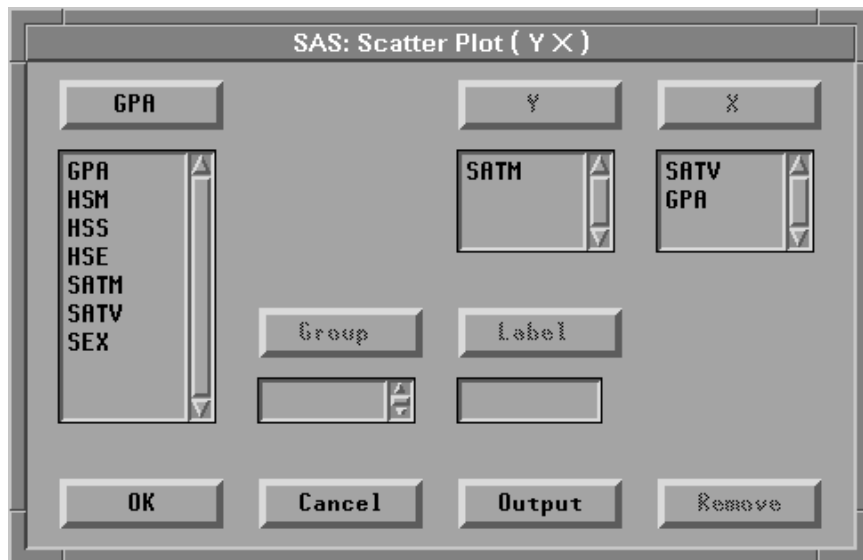


Figure 9.8. Variable Roles Assigned

⇒ **Click the OK button.**

This creates two scatter plots, as shown in [Figure 9.9](#).

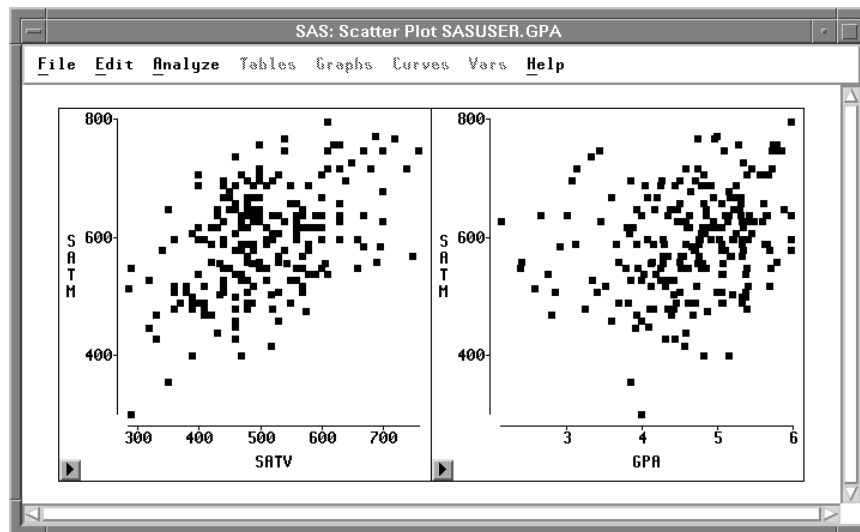


Figure 9.9. Scatter Plots

⇒ **Click on the button at the lower left to display the scatter plot pop-up menu.**

Choose **Observations** to turn off the display of observations in the scatter plot.

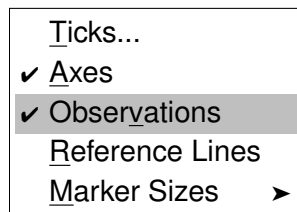


Figure 9.10. Scatter Plot Pop-up Menu

Do the same thing for the scatter plot on the right side. All the observation markers disappear, as shown in [Figure 9.11](#).

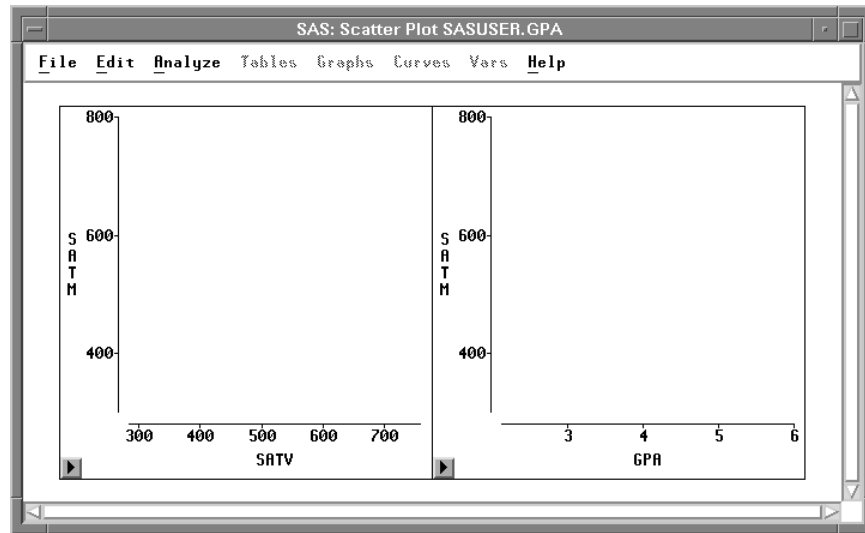


Figure 9.11. Turning Off Observations Display

⇒ **Choose Edit:Observations:Find**

This displays the Find Observations dialog. Select the variable **SEX**. With the default values in the other lists, this creates a test for **SEX = Female**.



Figure 9.12. Find Observations Dialog

⇒ **Click the OK button.**

This selects all **Female** observations and displays them in the scatter plots.

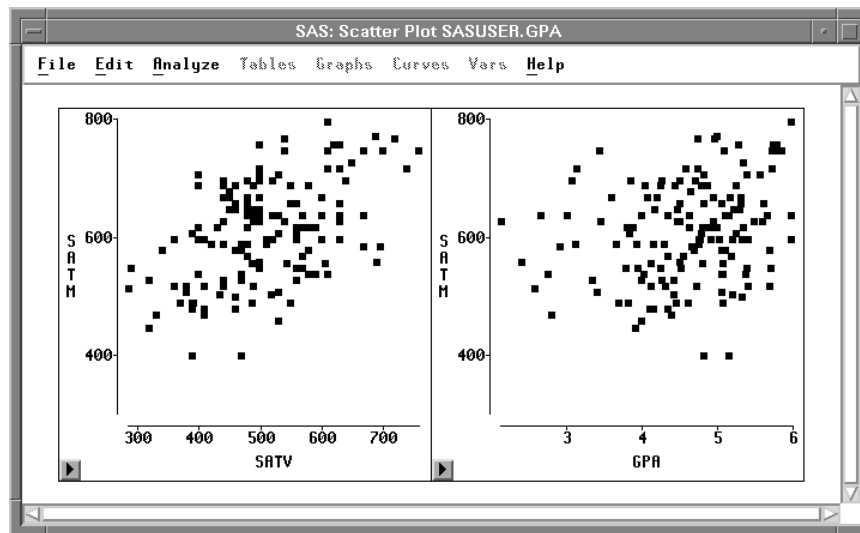


Figure 9.13. Female Observations

⇒ Choose **Edit:Observations:Invert Selection**.

Invert Selection deselects all selected observations and selects all deselected observations. Now the scatter plots show all observations where **SEX** is **Male**.

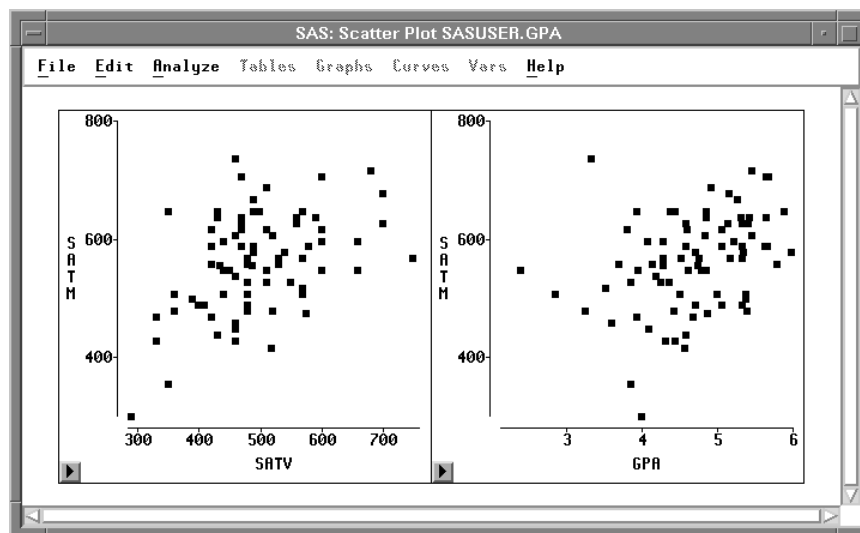


Figure 9.14. Male Observations

Toggling observations in the scatter plots shows there are more females than males in these data. The female students appear to have slightly higher scores on the mathematics portion of the SAT exam.

Slicing

Slicing is a dynamic technique for subsetting your data based on a range of values for one variable. You can create a brush both to restrict the range of values in one plot and to select observations in all plots. You can slice dynamically to explore relationships in more than two dimensions.

Follow these steps to see how **GPA** is related to the two SAT scores.

- ⇒ **Drag a rectangle with the mouse in the scatter plot of SATM versus GPA.**
This selects the observations within the rectangle and creates a rectangular *brush*.
- ⇒ **Move the brush by dragging with the mouse inside the brush.**
Observations that are selected by the brush become visible in both scatter plots. The second plot shows the conditional distribution of the data as restricted by the position of the brush in the first plot.

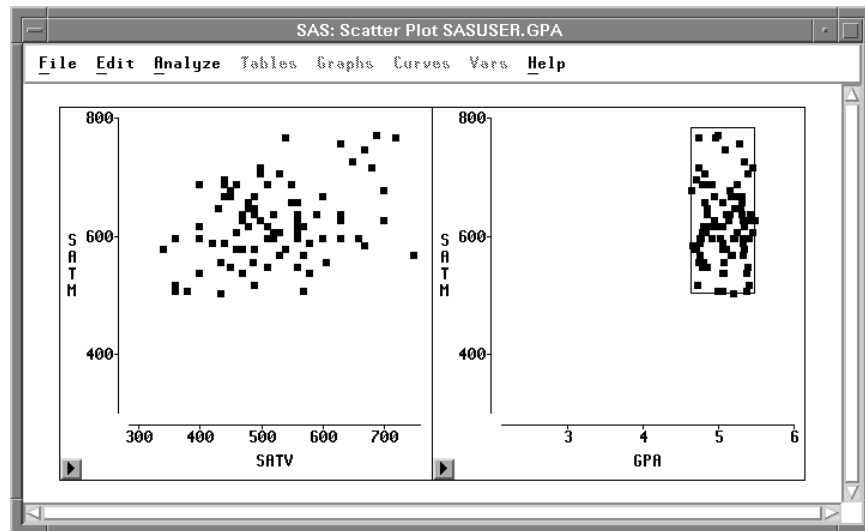


Figure 9.15. Brushing Invisible Observations

- ⇒ **Drag the corners of the brush to make it tall and thin.**
This restricts selected observations to a narrow range of values for **GPA**.
- ⇒ **Move the brush to the left and right.**
The scatter plot of **SATM** versus **SATV** in [Figure 9.16](#) shows the joint distribution of the two SAT scores when **GPA** is near 4.0. By sliding the brush, you can see whether the distributions change significantly as **GPA** increases or decreases.

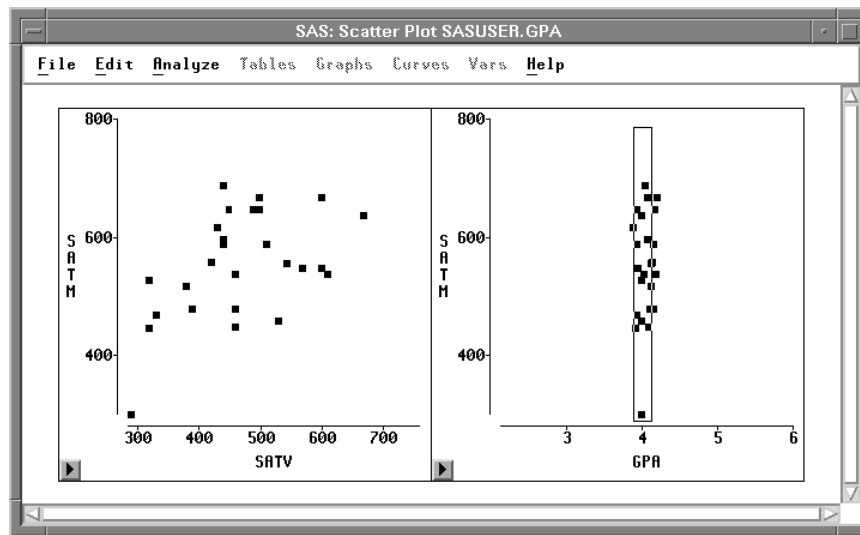


Figure 9.16. Slicing Observations

⇒ Use the scatter plot pop-up menu to make observations visible again.

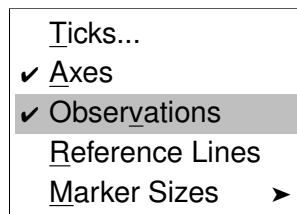


Figure 9.17. Scatter Plot Pop-up Menu

Chapter 10

Marking Observations

Chapter Contents

MARKING INDIVIDUAL OBSERVATIONS	158
MARKING BY NOMINAL VARIABLE	160
MARKING BY INTERVAL VARIABLE	161
ADJUSTING MARKER SIZE	162

Chapter 10

Marking Observations

You can assign markers to use for displaying observations in box plots, scatter plots, and rotating plots. The markers appear with each observation in the data window. You can assign markers for observations you select, and you can let SAS/INSIGHT software assign markers automatically based on the value of a variable. You can control the size of the markers in any plot.

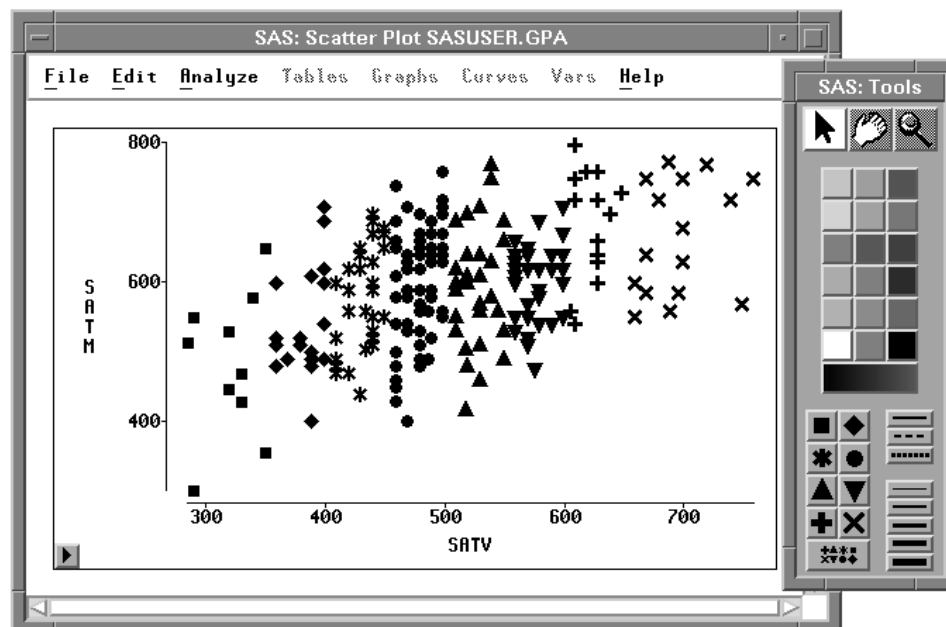


Figure 10.1. Marking Observations

Marking Individual Observations

You can set the marker shape for any observations you select.

⇒ **Open the GPA data set.**

⇒ **Create a scatter plot of SATM versus SATV.**

Use the techniques described in [Chapter 5, “Exploring Data in Two Dimensions.”](#)

⇒ **Click on an observation to select it.**

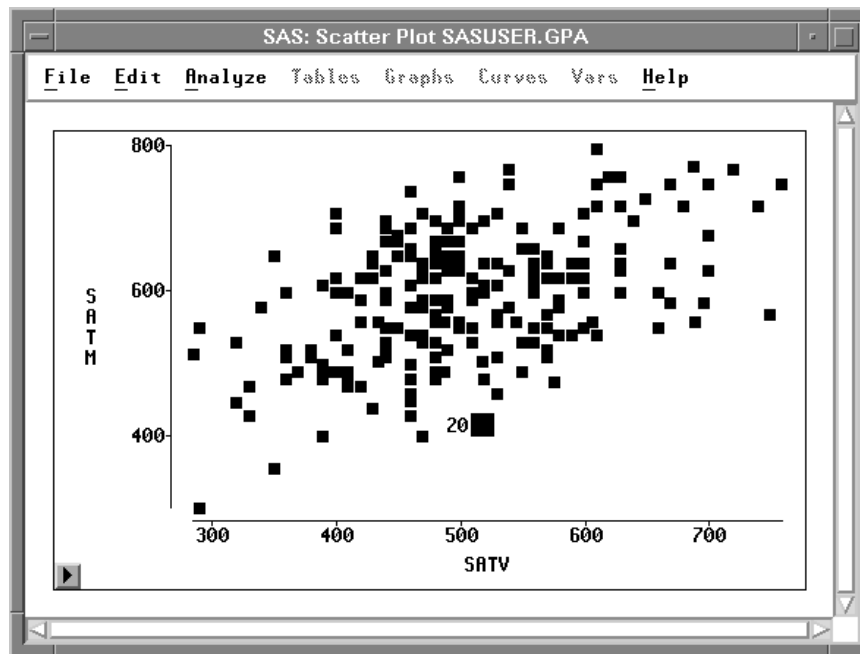


Figure 10.2. Scatter Plot

⇒ **Choose Edit:Windows:Tools.**

This toggles the display of the tools window, as shown in [Figure 10.4](#).

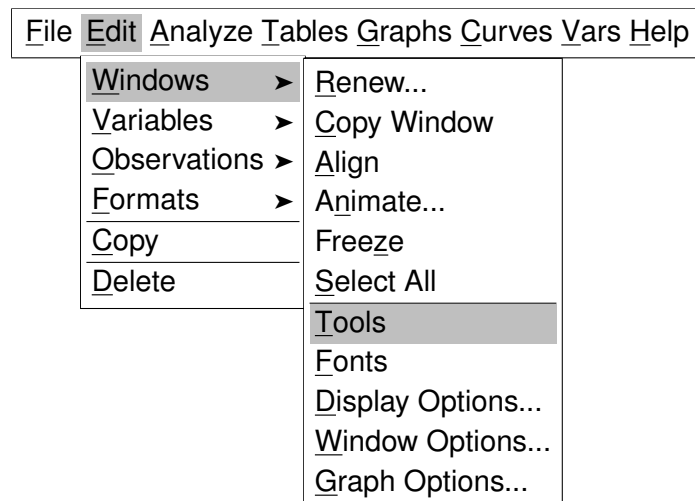


Figure 10.3. Edit:Windows Menu

⇒ **Click on the upward-pointing triangle in the tools window.**

This changes the marker for the selected observation from a square to a triangle. The marker also changes to a triangle in the data window and in any other windows.

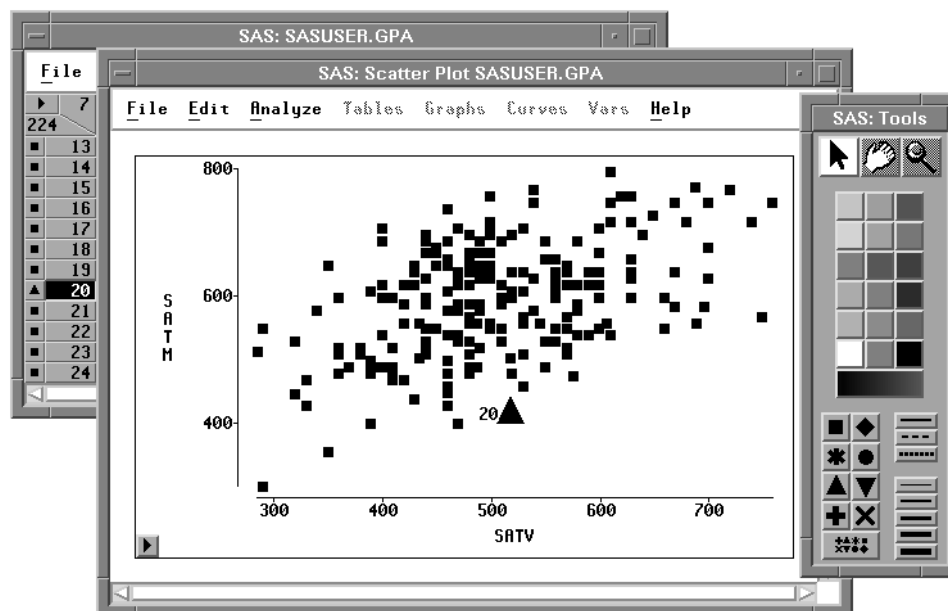


Figure 10.4. Changing a Marker

Similarly, you can select a group of observations in a brush and assign markers for the group. Markers provide a convenient way to track observations across multiple windows. They also enable you to keep track of observations when they are deselected.

Marking by Nominal Variable

You can assign markers automatically based on the value of a nominal variable. This is a good way to distinguish quickly between groups of observations.

⇒ **Select **SEX** in the data window.**

⇒ **Click on the multiple markers button at the bottom of the markers window.**

SAS/INSIGHT software assigns a different marker for each value of the nominal variable. In this case, observations with a value of **MALE** are displayed with crosses, and observations with a value of **FEMALE** are displayed with squares.

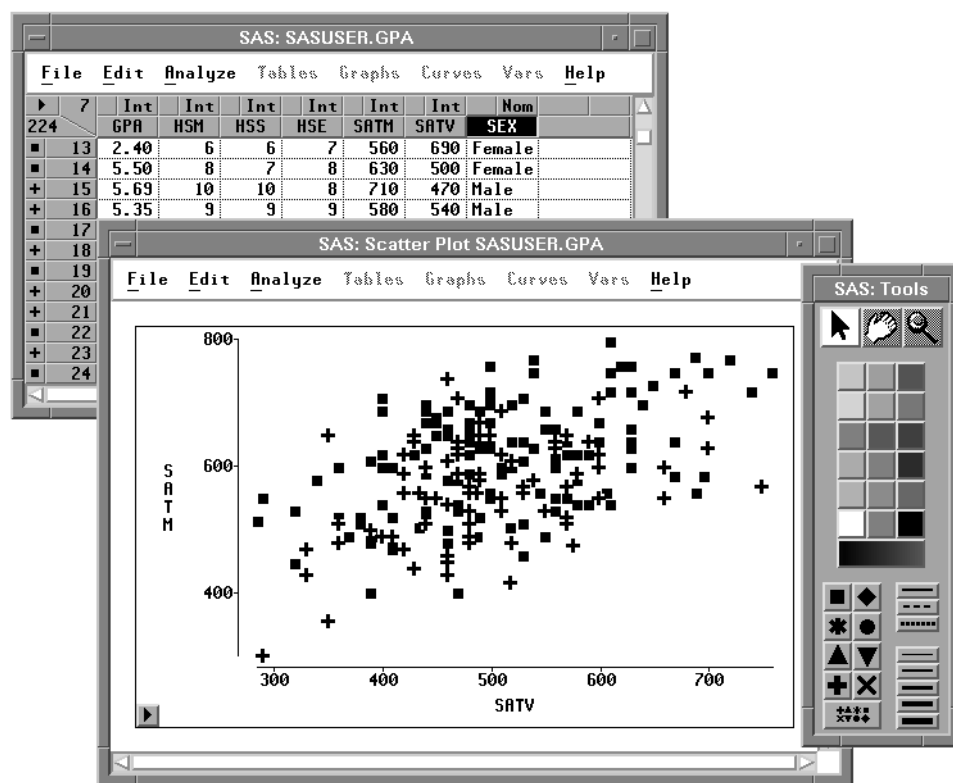


Figure 10.5. Assigning Markers by **SEX**

Marking by Interval Variable

You can also assign markers based on the value of an interval variable.

⇒ Select **GPA** in the data window.

⇒ Click on the **multiple markers button** at the bottom of the markers window.

SAS/INSIGHT software assigns three markers to the observations depending on the value of **GPA** for that observation. Observations with values in the upper third of the range of **GPA** are assigned upward-pointing triangles. Observations with values in the middle third of the range of **GPA** are assigned squares. Observations with values in the lower third of the range of **GPA** are assigned downward-pointing triangles. These markers show a rough picture of the correlation between grade point average and SAT scores.

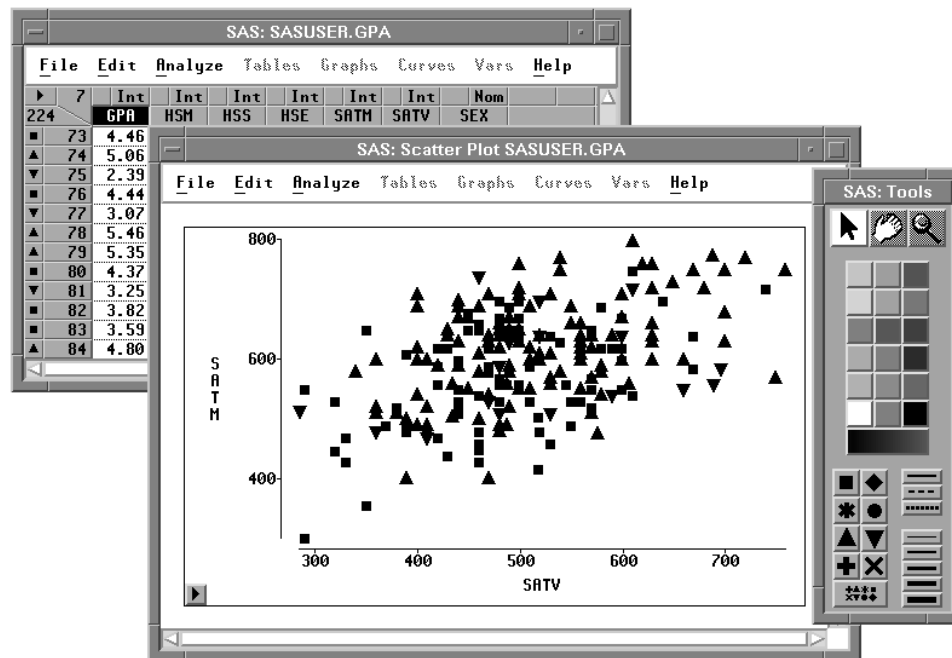


Figure 10.6. Assigning Markers by GPA

Adjusting Marker Size

You can adjust marker size by using the scatter plot pop-up menu.

- ⇒ Click on the button in the lower left corner of the scatter plot.
Choose **Marker Sizes:1**. This assigns markers their minimum size.

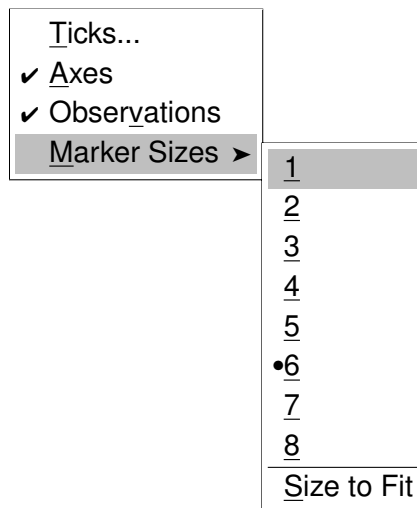


Figure 10.7. Marker Sizes Menu

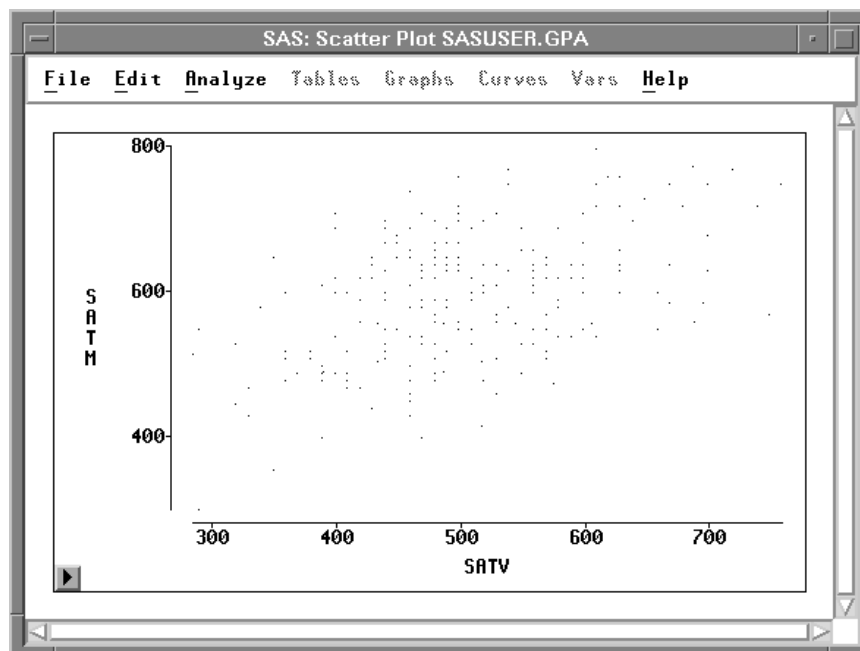


Figure 10.8. Markers at Minimum Size

- ⇒ Choose **Marker Sizes:8** from the pop-up menu.
This assigns markers their maximum size.

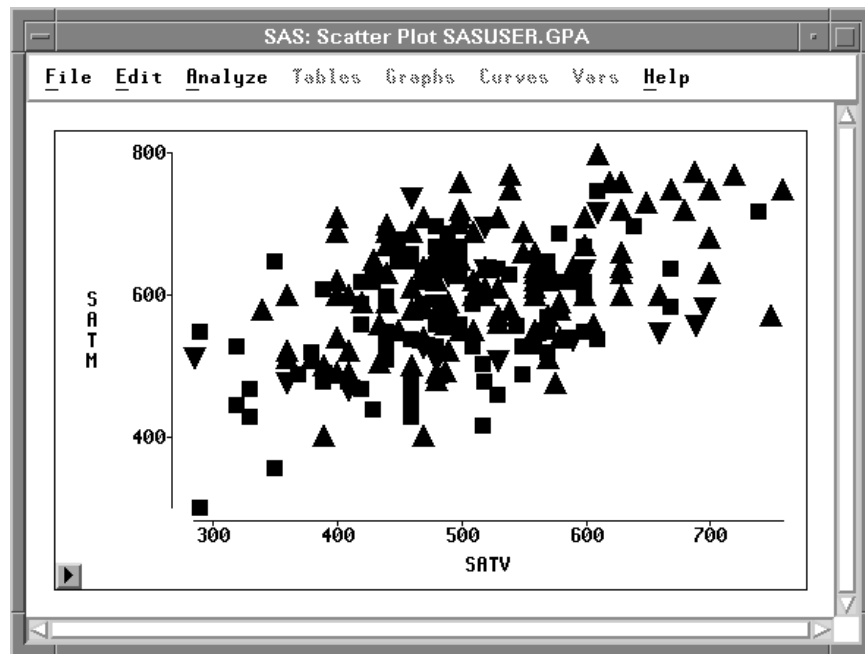


Figure 10.9. Markers at Maximum Size

- ⇒ Choose **Marker Sizes:Size to Fit** from the pop-up menu.
This assigns markers their default size.

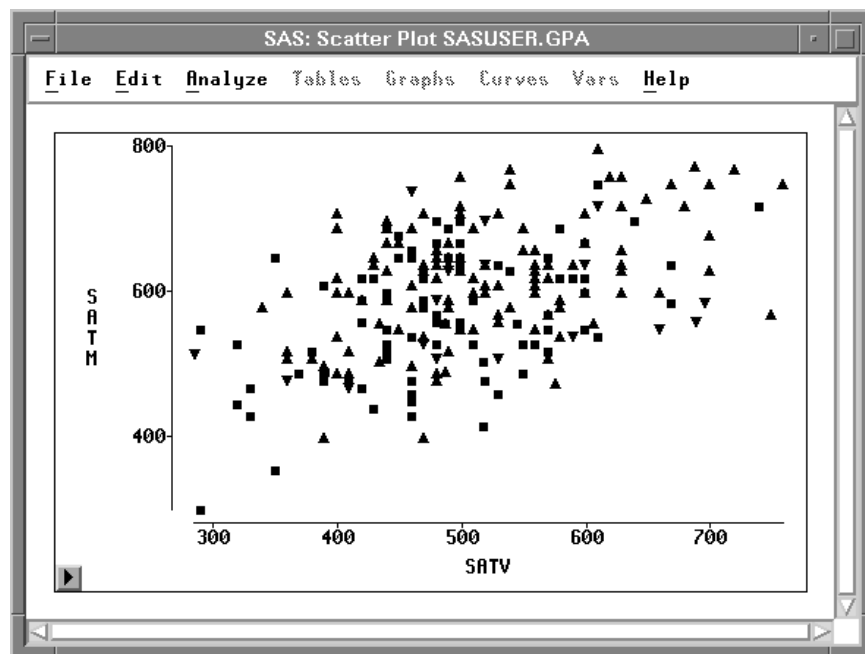


Figure 10.10. Default Marker Size

The default marker size is determined by the size of your graph, the resolution of your display, and the setting of the **Marker Size** option. You can set the **Marker Size** option as described in [Chapter 29, “Configuring SAS/INSIGHT Software.”](#)

† **Note:** For large data sets, markers require plenty of memory. If your data set contains hundreds of observations and your host has insufficient memory, you can improve performance by using the default square marker for all observations.

If you have a color display, it is often clearer to distinguish observations by color. Turn to the next chapter to see how to assign colors.

Chapter 11

Coloring Observations

Chapter Contents

COLORING INDIVIDUAL OBSERVATIONS	170
COLORING BY NOMINAL VARIABLE	172
COLORING BY INTERVAL VARIABLE	173
MULTIPLE COLOR BLENDS	174

Chapter 11

Coloring Observations

You can assign the colors for displaying observations in plots. You can assign colors for the observations you select, and you can let SAS/INSIGHT software assign colors automatically based on the value of a variable.

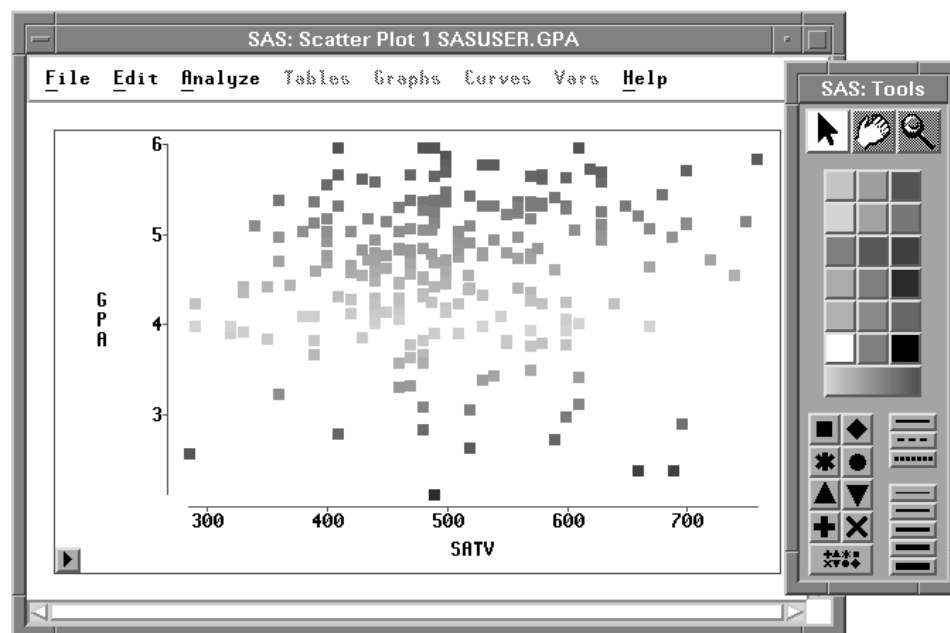


Figure 11.1. Coloring Observations

Coloring Individual Observations

You can set the color for any observations you select.

⇒ **Open the GPA data set.**

⇒ **Create a scatter plot of SATM versus SATV.**

Use the techniques described in [Chapter 5, “Exploring Data in Two Dimensions.”](#)

⇒ **Click on an observation to select it.**

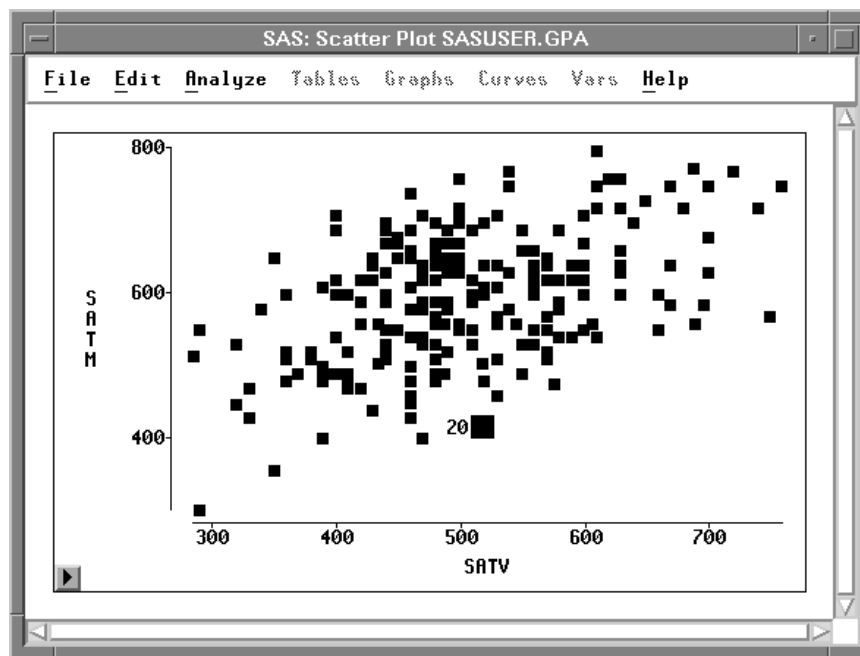


Figure 11.2. Scatter Plot

⇒ **Choose Edit:Windows:Tools.**

This toggles the display of the tools window, shown in [Figure 11.4](#).

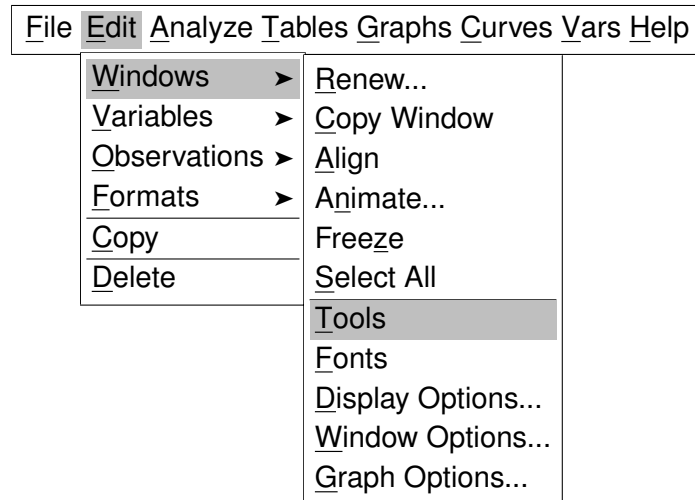


Figure 11.3. Edit:Windows Menu

⇒ **Click on the red button in the tools window.**

This causes the selected observation to turn red. The marker also becomes red in the data window and in any other windows.

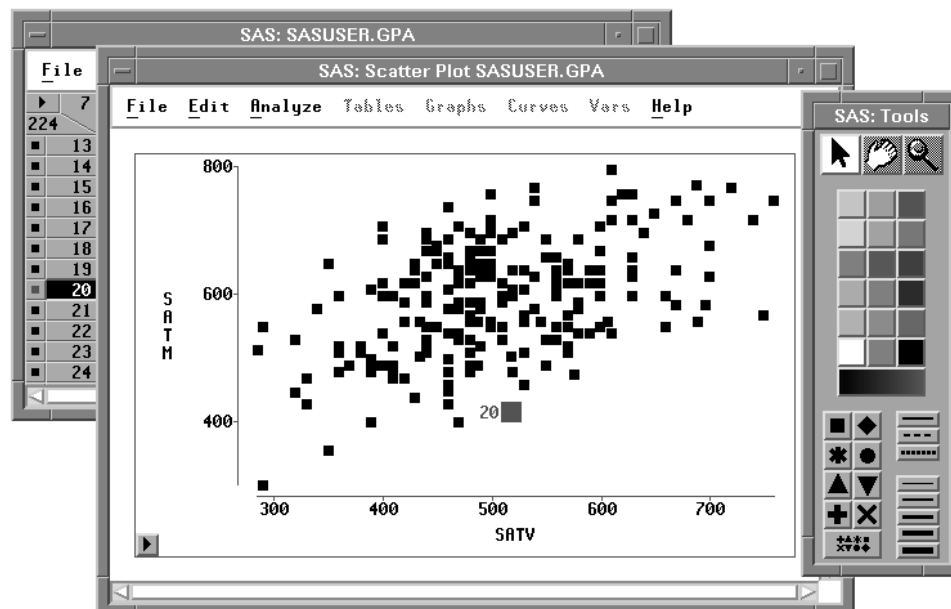


Figure 11.4. Changing a Color

You can similarly select a group of observations in a brush and assign colors for the group. Colors, like markers, provide a convenient way to track observations through multiple windows.

Coloring by Nominal Variable

You can set observation colors based on the value of a nominal variable. This is a good way to display subsets of the data.

⇒ Click on **SEX** in the data window.

⇒ Click on the large multiple colors button in the tools window.

SAS/INSIGHT software automatically assigns a different color for each value of the nominal variable.

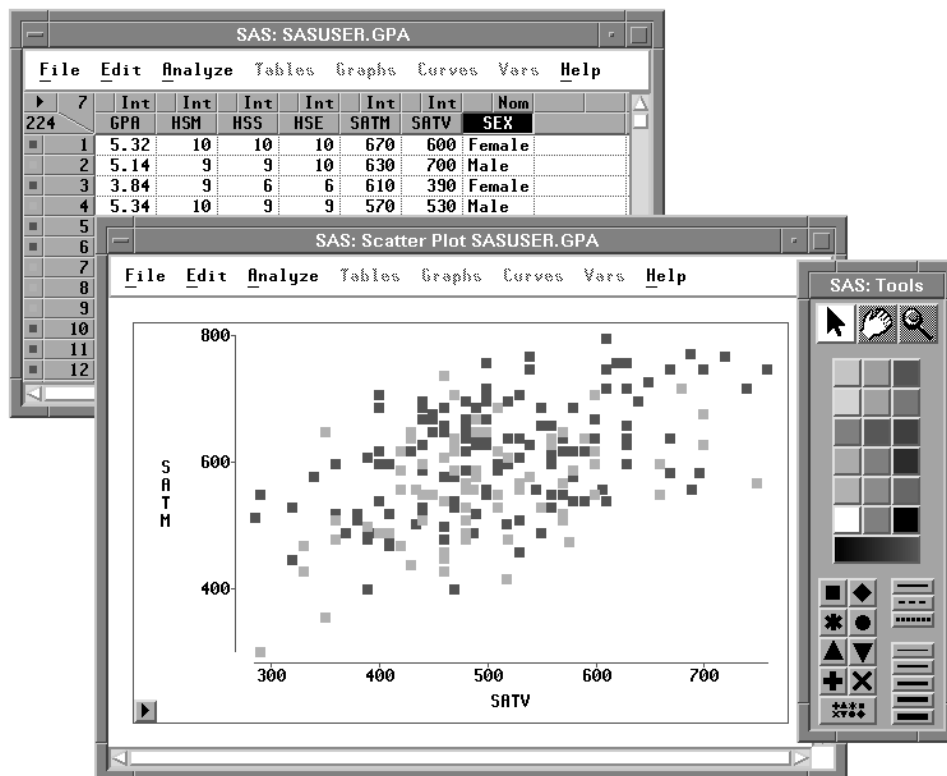


Figure 11.5. Assigning Colors by **SEX**

Coloring by Interval Variable

You can also set the marker colors based on the value of an interval variable.

⇒ Click on **GPA** in the data window.

⇒ Click on the large multiple colors button in the tools window.

SAS/INSIGHT software assigns a color to each observation depending on the value of **GPA** for that observation. The color varies smoothly between the two colors at the ends of the button. This use of color adds an extra dimension to the plot.

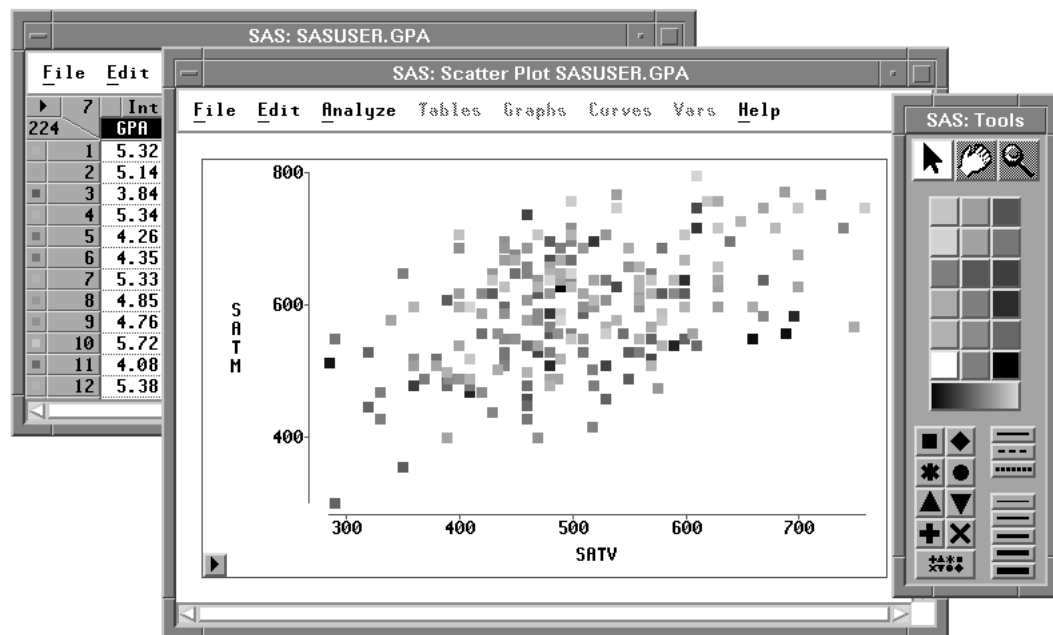


Figure 11.6. Assigning Colors by **GPA**

† **Note:** Some hosts do not support color blending. On these hosts, SAS/INSIGHT software uses a discrete set of colors instead of a smooth blend. You may also see this behavior when running multiple applications that do not share color resources. When the host does not support blending, or insufficient colors are available, the multiple colors button shows discrete bands of colors instead of a smooth blend.

On hosts that support color blending, you can choose the range over which the color varies. The left end of the multiple colors button defaults to white or black, whichever contrasts with the background color. The right end of the multiple colors button defaults to red. To use a range from blue to red, follow these steps.

⇒ Place the cursor on the blue button in the tools window.

⇒ Drag the blue color down to the left end of the large button.

Then release the mouse button. The colors in the button change to a smooth blend between blue and red.

You can also drag colors to the right side of the button to make other blends. This lets you choose colors that have meaning for your data, for example, blue-to-red for cold-to-hot or brown-to-green for arid-to-tropical.

Multiple Color Blends

Color blending applies to all observations if none are selected. If observations are selected, color blending applies only to the selected observations. This enables you to assign multiple color blends for a single variable.

- ⇒ **Create a scatter plot of GPA versus SATV.**
- ⇒ **Create a blue-to-yellow blend in the tools window.**
Drag the blue color to the left end of the multiple colors button, and drag the yellow color to the right end.
- ⇒ **Select observations with values of GPA less than or equal to 4.**

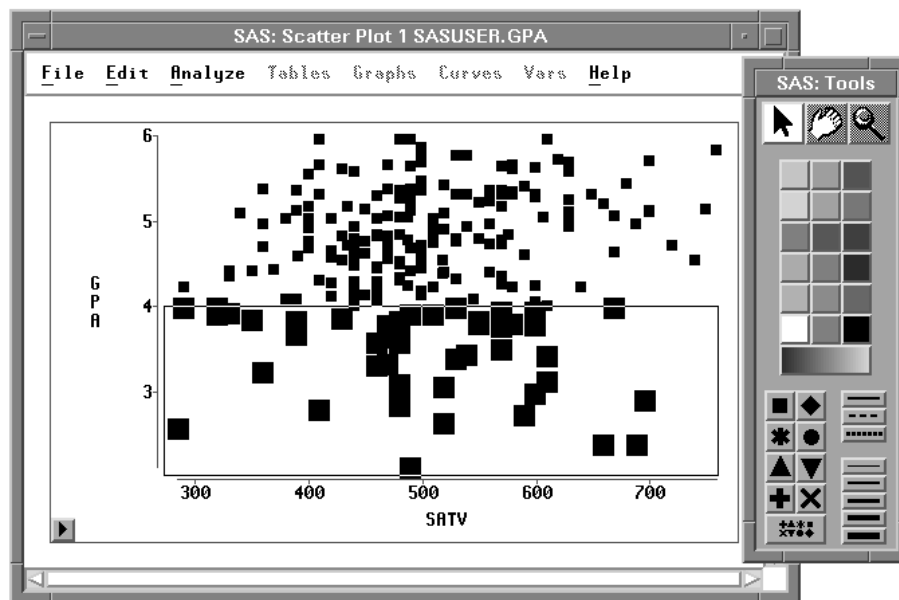


Figure 11.7. Selecting Observations Where $GPA \leq 4$

- ⇒ **Click the multiple colors button.**
This displays a variables dialog, as shown in [Figure 11.8](#).



Figure 11.8. Variables Dialog

⇒ **In the variables dialog, select GPA, then click OK.**

This assigns the blue-to-yellow blend to observations with values of **GPA** less than or equal to 4.

You can use similar steps to assign a yellow-to-red blend to all observations with values of **GPA** greater than 4. To save time, select both observations and variables using extended selection instead of using the variables dialog.

⇒ **Create a yellow-to-red blend in the tools window.**

Drag the yellow color to the left end of the multiple colors button, and drag the red color to the right end.

⇒ **Select observations with values of GPA greater than or equal to 4.**

⇒ **Using extended selection, select the variable GPA.**

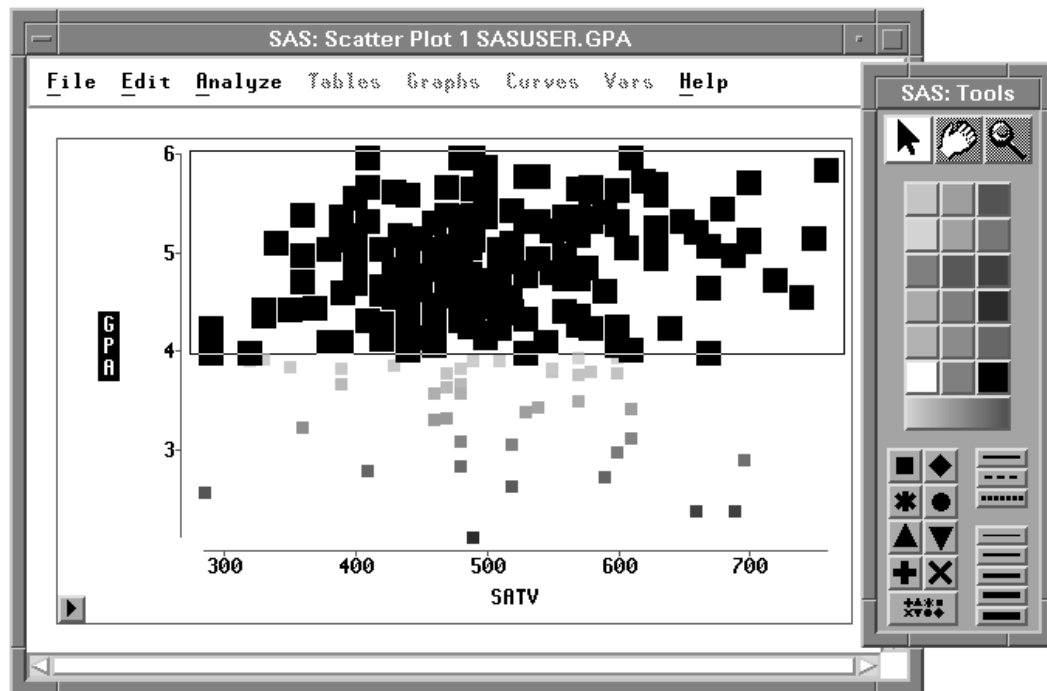


Figure 11.9. Selecting Variable **GPA** and Observations Where **GPA** ≥ 4

⇒ **Click the multiple colors button.**

This assigns the yellow-to-red blend to observations with values of **GPA** greater than or equal to 4. Now all observations are assigned a color based on their value for **GPA**, with colors smoothly blended from blue through yellow to red.

† **Note:** In addition to the two-color blends described above, you can create a blended color strip based on the interpolation of up to five colors.
To do this, follow these steps:

- Bring up the tools window by using **Edit:Window:Tools**.
- Choose a color in the tools window and place the cursor over that color button. For the sake of this example, choose the white button.
- Hold down the **shift** key.
- Shift-drag the white button onto the large multiple colors button.
- Release the mouse button while the cursor is in the middle of the multiple colors button. One of the existing colors that make up the multiple color button is replaced by white.
- You can further modify the color strip by shift-dragging other color buttons to varying positions along the length of the multiple color button.

Chapter 12

Examining Distributions

Chapter Contents

CREATING THE DISTRIBUTION ANALYSIS	180
Box Plot	183
Histogram	185
Moments and Quantiles Tables	188
ADDING DENSITY ESTIMATES	189
Normal Density Curve	189
Kernel Density Curve	191
TESTING DISTRIBUTIONS	194
REFERENCES	197

Chapter 12

Examining Distributions

In Chapter 4, “Exploring Data in One Dimension,” you examined distributions using bar charts and box plots. In this chapter, you examine the distribution of an interval variable using graphs and statistical tables.

You can examine box plots and histograms of the data along with **Moments** and **Quantiles** tables. You can superimpose density curves on the histogram. You can carry out tests to determine whether the data are from specific parametric distributions, such as normal or lognormal.

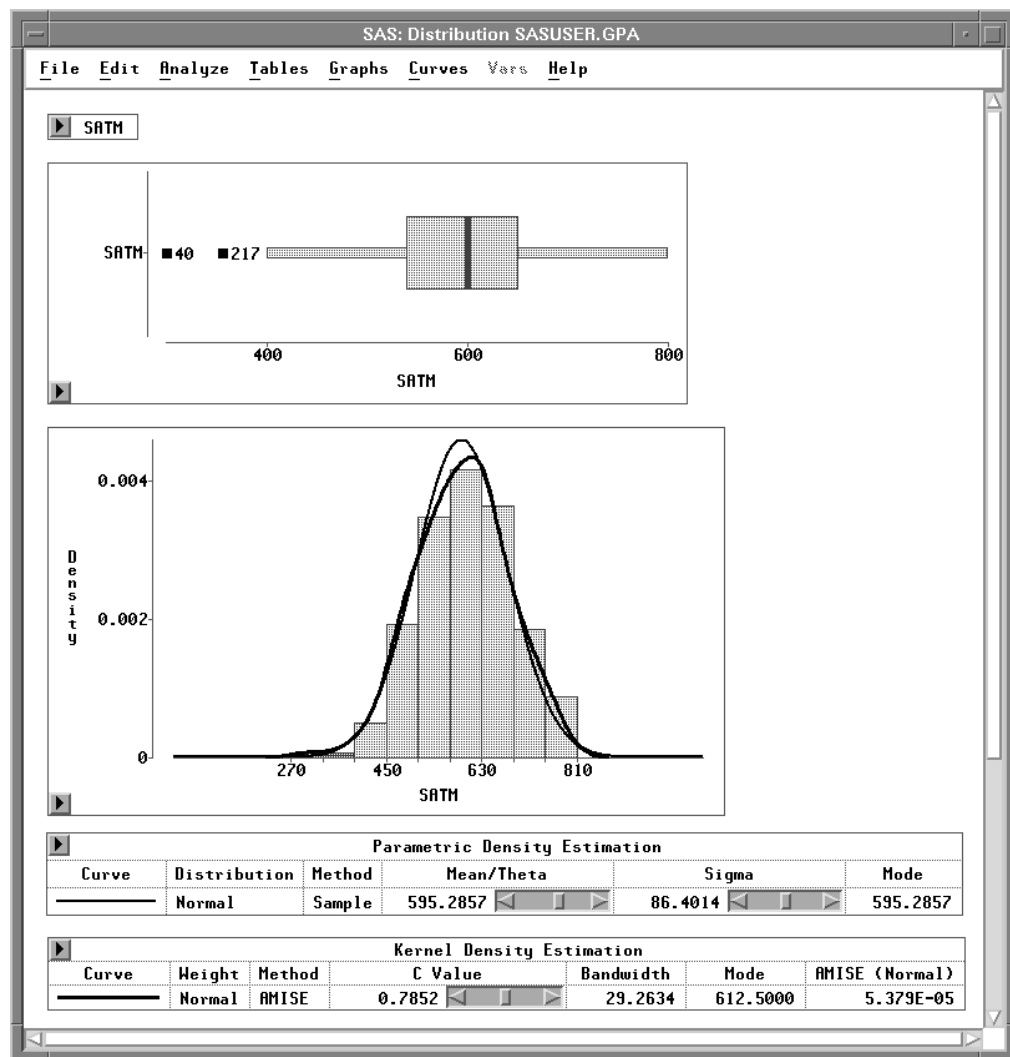


Figure 12.1. Distribution Analysis

Creating the Distribution Analysis

The *distribution* of a variable is the pattern of variation of its numerical values (Moore and McCabe 1989). In this example, you examine a distribution of scores on the mathematics portion of the SAT exam.

⇒ Open the **GPA** data set.

⇒ Select the variable **SATM** by clicking on its name in the data window.

SAS: SASUSER.GPA

File Edit Analyze Tables Graphs Curves Vars Help

7	Int	Int	Int	Int	Int	Int	Nom		
224	GPA	HSM	HSS	HSE	SATM	SATV	SEX		
1	5.32	10	10	10	670	600	Female		
2	5.14	9	9	10	630	700	Male		
3	3.84	9	6	6	610	390	Female		
4	5.34	10	9	9	570	530	Male		
5	4.26	6	8	5	700	640	Female		
6	4.35	8	6	8	640	530	Female		
7	5.33	9	7	9	630	560	Male		
8	4.85	10	8	8	610	460	Male		
9	4.76	10	10	10	570	570	Male		

Figure 12.2. Data Window with **SATM** Selected

⇒ Choose **Analyze:Distribution (Y)**.

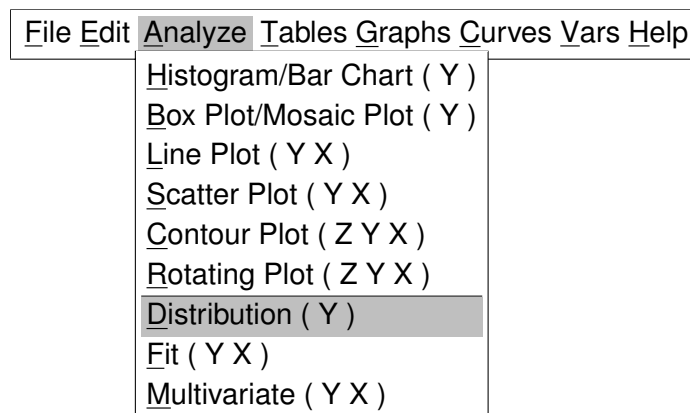


Figure 12.3. Analyze Menu

This creates a distribution window, as shown in [Figure 12.4](#). A box plot, histogram, **Moments** table, and **Quantiles** table appear by default. With these graphs and tables, you can examine important features of a distribution.

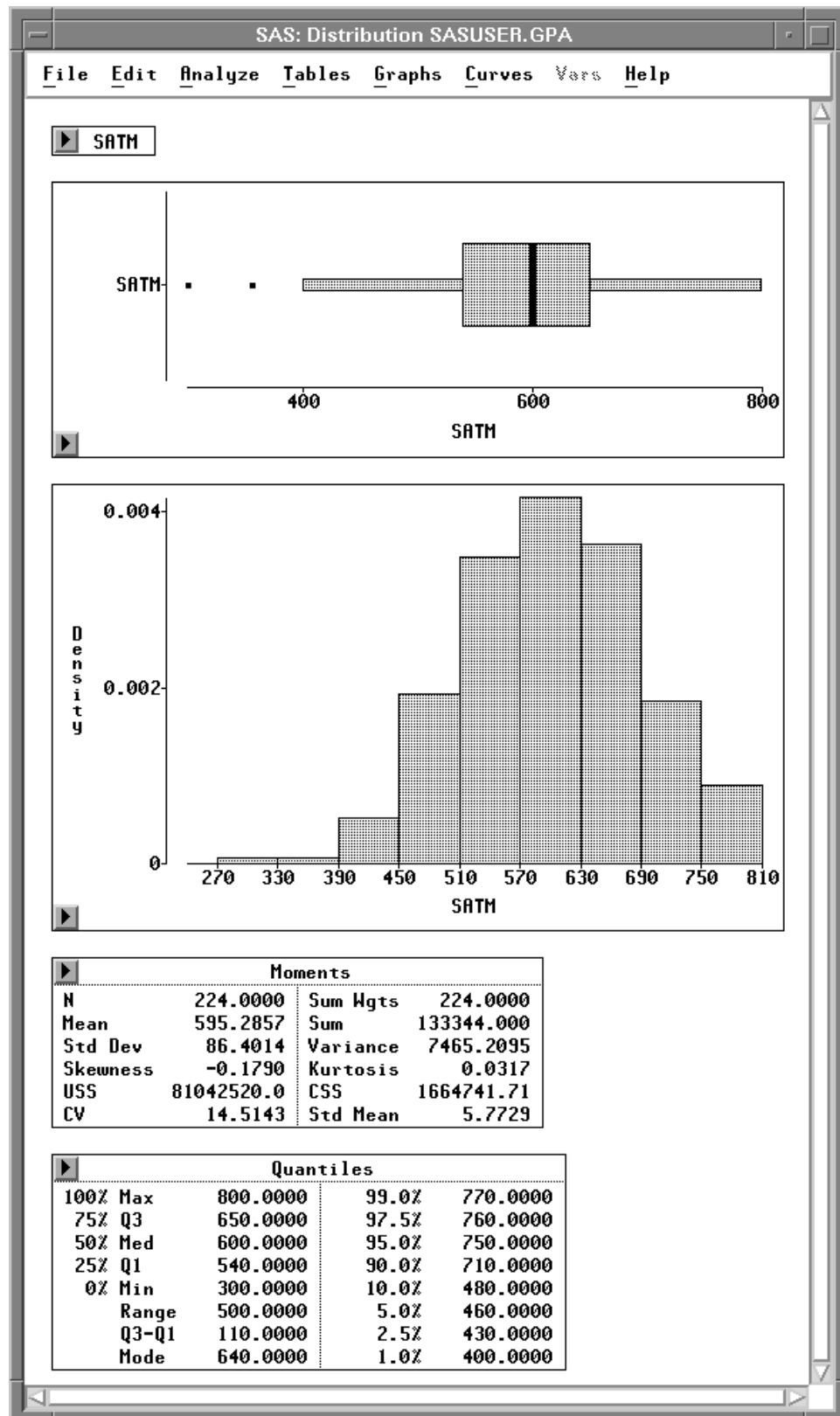


Figure 12.4. Distribution Analysis

Box Plot

A box plot is a schematic representation of a distribution. The vertical lines in the box mark the 25th, 50th, and 75th percentiles of the data. The p th percentile of a distribution is the value such that p percent of the observations fall at or below it. The 50th percentile is also called the *median*, and the 25th and 75th percentiles are called *quartiles*.

The narrow boxes extending to the left and right are called *whiskers*. Whiskers extend from the quartiles to the farthest observation not farther than 1.5 times the distance between the quartiles (the *interquartile range*). Beyond the whiskers, extreme observations are plotted individually.

The box plot gives a concise picture of the distribution and emphasizes any extreme values. This particular box plot appears fairly symmetric, with median around 600. You can see two extreme values.

⇒ **Identify the extreme observations by clicking on them.**

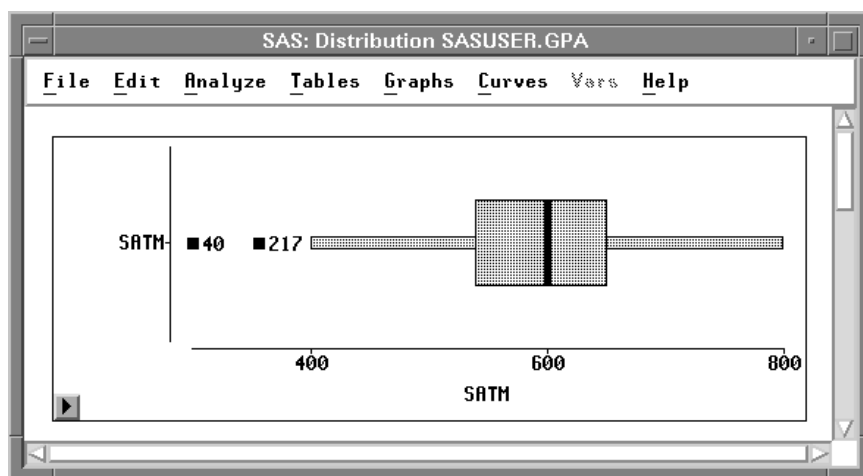


Figure 12.5. Identifying Extreme Observations

These are observations 40 and 217. When you click on them, the observations are selected in the box plot, the histogram, and the data window as well.

⊕ **Related Reading:** Box Plots, [Chapter 33](#).

⇒ **Click in the upper left corner of the data window.**

This displays the data pop-up menu.

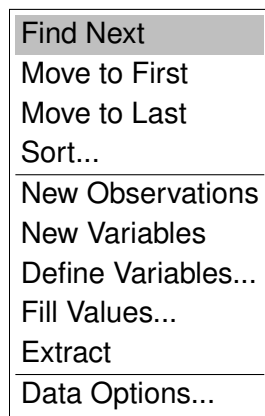


Figure 12.6. Data Pop-up Menu

⇒ **Choose Find Next from the pop-up menu.**

This scrolls the data window to the next selected observation, as shown in [Figure 12.7](#). By choosing **Find Next** again, you can examine all values for the extreme observations.

A screenshot of the SAS Data Window titled 'SAS: SASUSER.GPA'. The window displays a table of data with columns for GPA, HSM, HSS, HSE, SATM, SATV, and SEX. The table is sorted by GPA in descending order. The observation with the lowest GPA (4.00) is selected, and the window is scrolled to show the bottom of the data set, including observations 40 through 48.

	7	Int	Int	Int	Int	Int	Int	Nom
		GPA	HSM	HSS	HSE	SATM	SATV	SEX
224	40	4.00	2	4	6	300	290	Male
	41	3.43	10	9	9	750	610	Female
	42	4.48	8	9	6	650	460	Female
	43	5.73	10	10	9	720	630	Female
	44	4.43	7	10	10	530	560	Female
	45	3.69	7	6	7	560	480	Male
	46	5.80	10	10	9	760	500	Female
	47	5.18	10	10	10	570	750	Male
	48	6.00	9	10	10	640	480	Female

Figure 12.7. Extreme Observation in Data Window

Histogram

A *histogram* is a bar chart of an interval variable. In a histogram, the interval represented by a bar is called a *bin*. Instead of a frequency axis, histograms in a distribution analysis use a *density* axis to measure the fractional distribution over a given interval.

Examine the histogram of **SATM**. The shape of the distribution is fairly symmetric except for slight skewing in the left tail. The distribution's center is around 600.

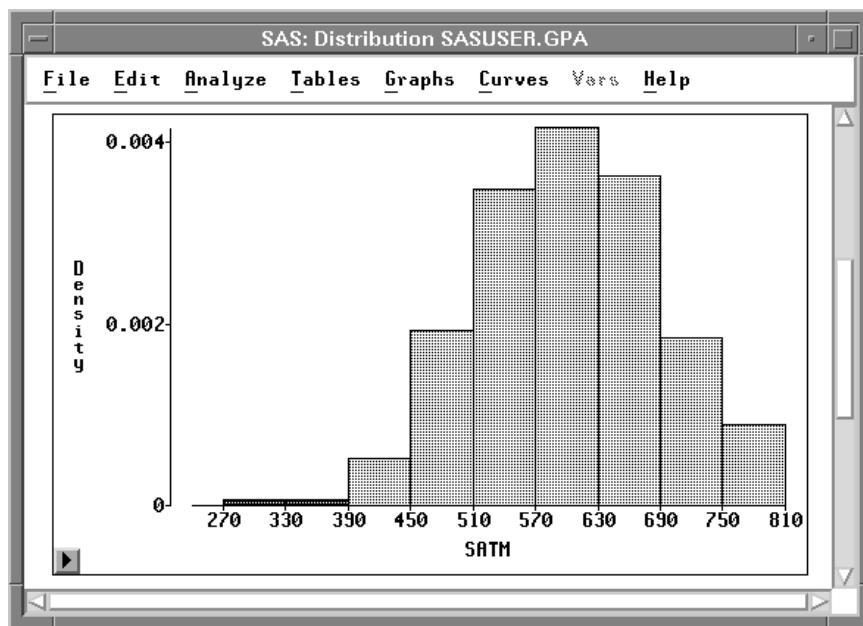


Figure 12.8. Histogram of **SATM**

A histogram is a good tool for visually examining the distribution. However, changes in the width and position of the bars can greatly affect your perception of the shape of the distribution. The histogram illustrated in [Figure 12.8](#) is only one representation of the distribution of **SATM**. It is easy to change the bar widths and positions with SAS/INSIGHT software to explore many different histograms.

⇒ **Choose Edit:Windows:Tools.**

This displays the tools window, as shown in [Figure 12.9](#).

⇒ **Click on the hand in the tools window.**

The cursor changes shape from an arrow to a hand.

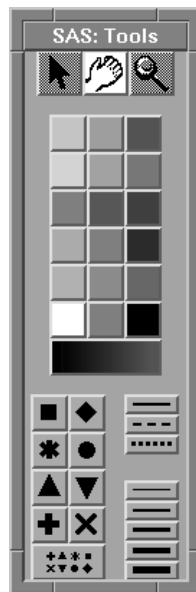


Figure 12.9. Tools Window

⇒ **Move the cursor back to the distribution window and click on the histogram.**
 This changes the width of the bars in proportion to the distance of the hand tool from the base of the bars. If the hand tool is close to the base of the bars, the bars are wide, as shown in [Figure 12.10](#).

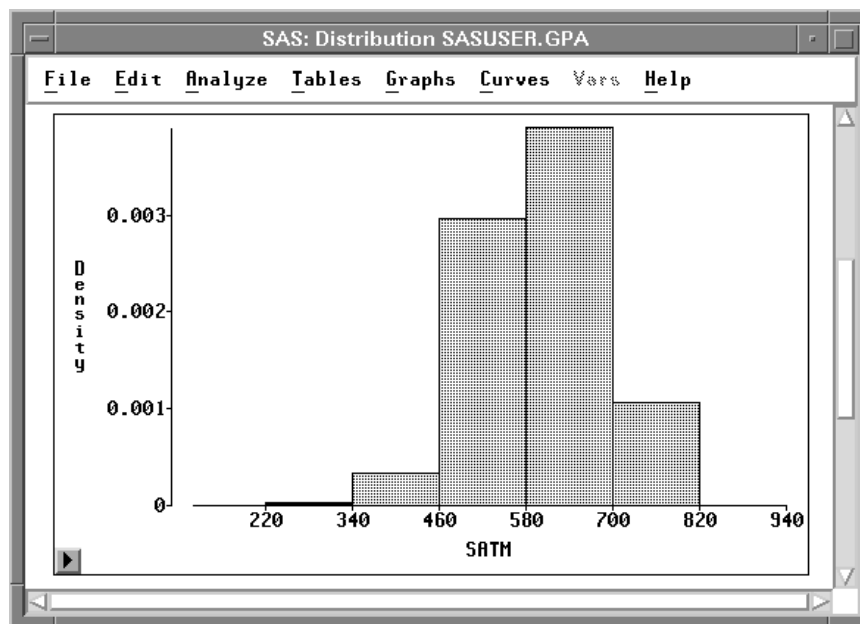


Figure 12.10. Clicking Close to the Base of the Bars

If the hand tool is far from the base of the bars, clicking makes the bars narrow, as

shown in [Figure 12.11](#).

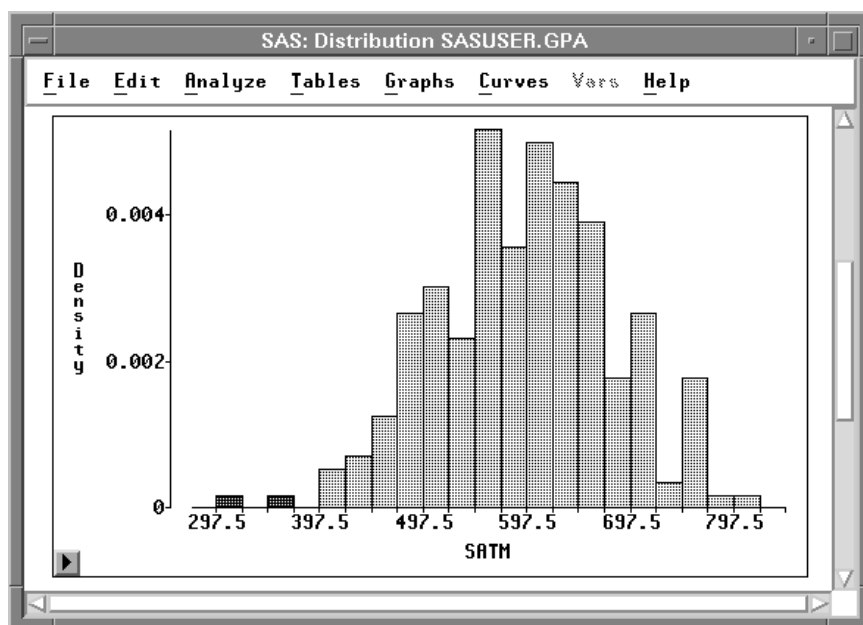
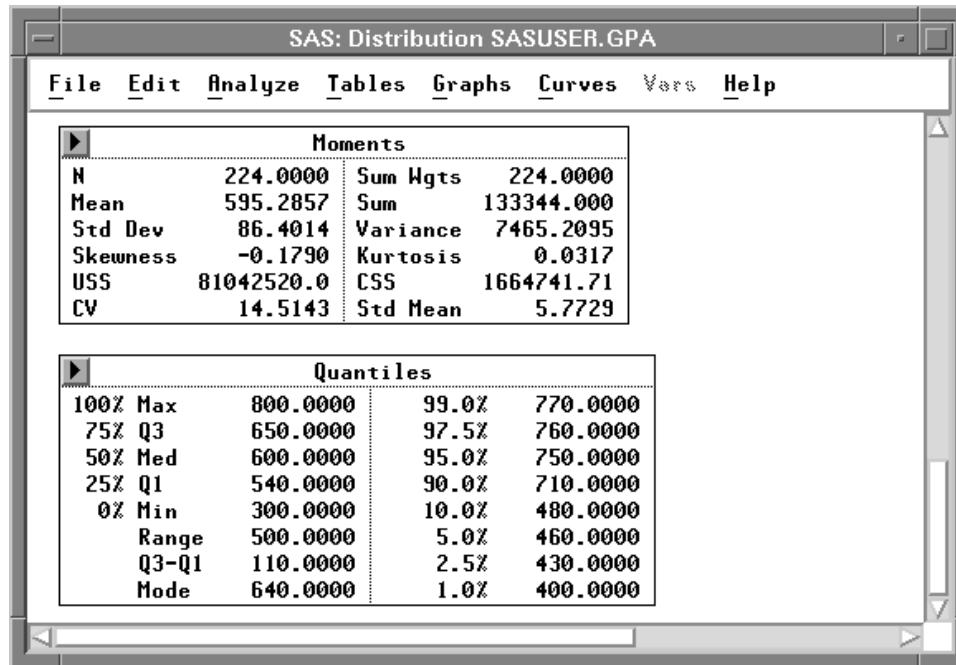


Figure 12.11. Clicking Far from the Base of the Bars

- ⇒ **Press the mouse button and hold it down as you move horizontally over the bars.**
Notice how the histogram changes as you move the hand. As you move horizontally, the bin width does not change, but the bins start at different locations. When the hand is at the left of the histogram, the bins start at an integral multiple of the bin width. When the hand moves toward the right, the bins are *offset* an amount proportional to the distance of the hand across the histogram.
 - ⇒ **Drag the hand horizontally and vertically in the histogram.**
Release the mouse button when you find a histogram that captures the dominant shape of the distribution.
 - ⇒ **Click on the arrow in the tools window before proceeding.**
- ⊕ **Related Reading:** Bar Charts, [Chapter 32](#).

Moments and Quantiles Tables

The **Moments** and **Quantiles** tables give descriptive information that quantifies what you observe in the box plot and histogram.



The screenshot shows the SAS Distribution window for SASUSER.GPA. It contains two tables: Moments and Quantiles.

Moments			
N	224.0000	Sum Wgts	224.0000
Mean	595.2857	Sum	133344.000
Std Dev	86.4014	Variance	7465.2095
Skewness	-0.1790	Kurtosis	0.0317
USS	81042520.0	CSS	1664741.71
CV	14.5143	Std Mean	5.7729

Quantiles			
100% Max	800.0000	99.0%	770.0000
75% Q3	650.0000	97.5%	760.0000
50% Med	600.0000	95.0%	750.0000
25% Q1	540.0000	90.0%	710.0000
0% Min	300.0000	10.0%	480.0000
Range	500.0000	5.0%	460.0000
Q3-Q1	110.0000	2.5%	430.0000
Mode	640.0000	1.0%	400.0000

Figure 12.12. Moments and Quantiles Tables

In the **Moments** table, **N** is the number of nonmissing observations, **Mean** is the arithmetic mean, **Std Dev** is the standard deviation, and **Variance** is the variance. **Skewness** and **Kurtosis** are both measures of the shape of the distribution.

Skewness is a measure of the tendency of the deviations from the mean to be larger in one direction than in the other. A positive value for **Skewness** indicates that the data are skewed to the right. A negative value indicates that the data are skewed to the left. The distribution of **SATM** is skewed slightly to the left, as you observed previously; thus, the value for **Skewness** is negative.

Kurtosis is primarily a measure of the heaviness of the tails of a distribution. Large values of **Kurtosis** indicate that the distribution has heavy tails. This statistic is standardized so that a normal distribution has a kurtosis of 0.

The **Quantiles** table gives information about the variability in the data as well as about the center of the data. Two distributions having the same center can look quite different if the variability in the two distributions is different. This variability is shown by the percentiles in the **Quantiles** table. The **Quantiles** table also shows the **Range** of the data, the interquartile range **Q3-Q1**, and the **Mode**.

Adding Density Estimates

A *cumulative distribution function* gives the proportion of the data less than each possible value. A *density function* is the derivative of the cumulative distribution function. *Density estimation* is the construction of an estimate of the density function from the observed data.

Histograms are one type of density estimation. You can also plot the density function to construct density curves. Density curves are sometimes preferred because they do not contain the discontinuous steps present in histograms.

Distribution (Y) provides two types of density estimation: parametric and kernel. In parametric estimation, the data are assumed to be from a known parametric family of distributions. The normal distribution is one of the most commonly used parametric distributions. Others include lognormal, exponential, and Weibull.

In kernel estimation, little is assumed about the functional form of the data. The data more completely determine the shape of the density curve. Kernel estimation is a type of nonparametric estimation.

Normal Density Curve

Begin by adding a normal density curve.

⇒ **Choose Curves:Parametric Density.**

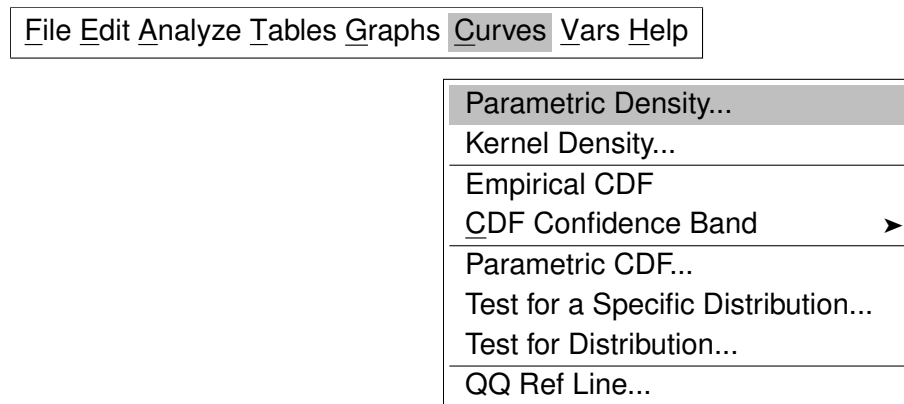


Figure 12.13. Normal Density Menu

This displays the parametric density estimation dialog in [Figure 12.14](#). You can select one of four distribution families, and you can use sample parameter estimates or you can specify your own.

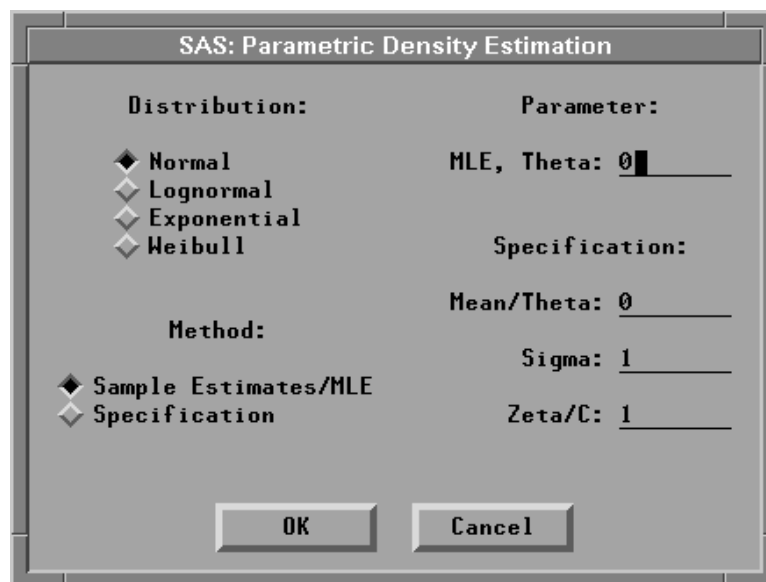


Figure 12.14. Parametric Density Estimation Dialog

⇒ Click OK in the dialog.

This requests the default density estimate: a normal distribution using the sample estimates as parameter values. The density curve is superimposed on the histogram, as illustrated in [Figure 12.15](#).

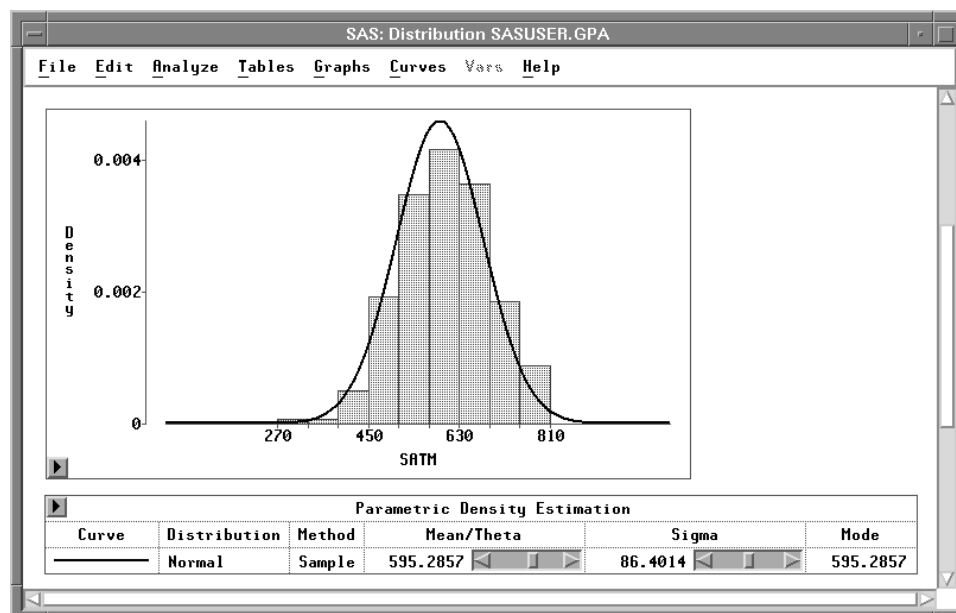


Figure 12.15. Parametric Normal Density Estimation

In addition, a **Parametric Density Estimation** table that contains parameter information appears in the window. You can change the specified parameters and the corresponding curve using the sliders next to the parameter values.

Note that the values of **Mean / Theta** and **Sigma** are equal to the sample **Mean** and **Std Dev** displayed in the **Moments** table illustrated in [Figure 12.12](#). The density curve follows the shape of the distribution fairly well.

⇒ **Select the density curve.**

You can select the curve by clicking on either the curve in the histogram or the legend on the table. Both the curve and the legend become highlighted.

⇒ **Choose Edit:Delete.**

The selected curve and its associated table are deleted from the window.

Kernel Density Curve

A kernel density curve may follow the shape of the distribution more closely. To construct a normal kernel density curve, one parameter is required: the bandwidth λ . The value of λ determines the degree of smoothing in the estimate of the density function. You can either specify a value of λ , or you can let SAS/INSIGHT software find a value based on minimizing an estimate of the mean integrated square error (MISE).

⇒ **Choose Curves:Kernel Density.**

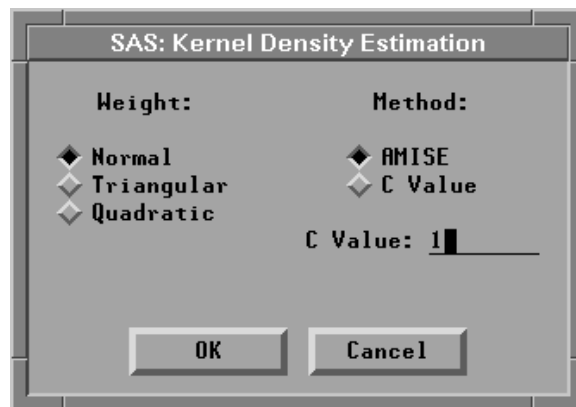


Figure 12.16. Kernel Density Estimation Dialog

⇒ **Click OK in the dialog.**

The kernel density curve is constructed with a bandwidth based on the approximated mean integrated square error (**AMISE**), and it provides a good visual representation of the distribution, as illustrated in [Figure 12.17](#). A table containing the bandwidth and the **AMISE** is also added to the window.

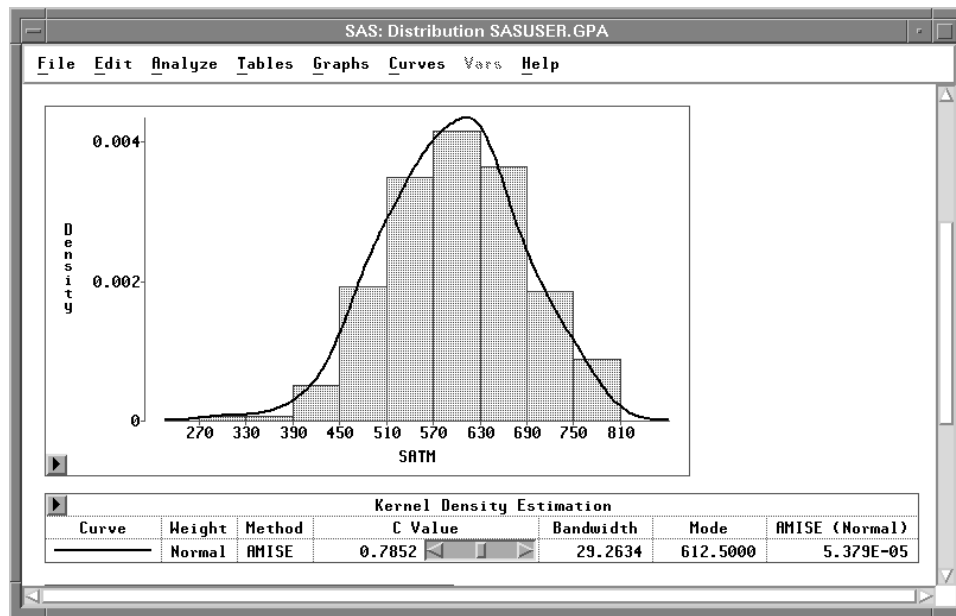


Figure 12.17. Kernel Density Estimate

The **C Value** slider in the table can be used to change the **C** value of the kernel estimate. You can use the slider in three ways:

- click the arrow buttons
- click within the slider
- drag within the slider

⇒ **Click the left arrow button in the slider.**

This decreases the **C** value by half. As the **C** value decreases, the density estimate becomes less smooth, as illustrated in [Figure 12.18](#).

⇒ **Click within the slider, just to the right of the slider control.**

This moves the slider control to the position where you click. The **C** value is set to a value proportional to the slider position. On most personal computers, clicking within the slider is the fastest way to adjust a curve.

⇒ **Drag the slider control left and right.**

When you drag the slider, its speed depends on the number of data points, the type of curve, and the speed of your host. Depending on your host, you may be able to improve the speed of the dynamic graphics with an alternate drawing algorithm. To try this, choose **Edit:Windows:Graph Options**, and set the **Fast Draw** option.

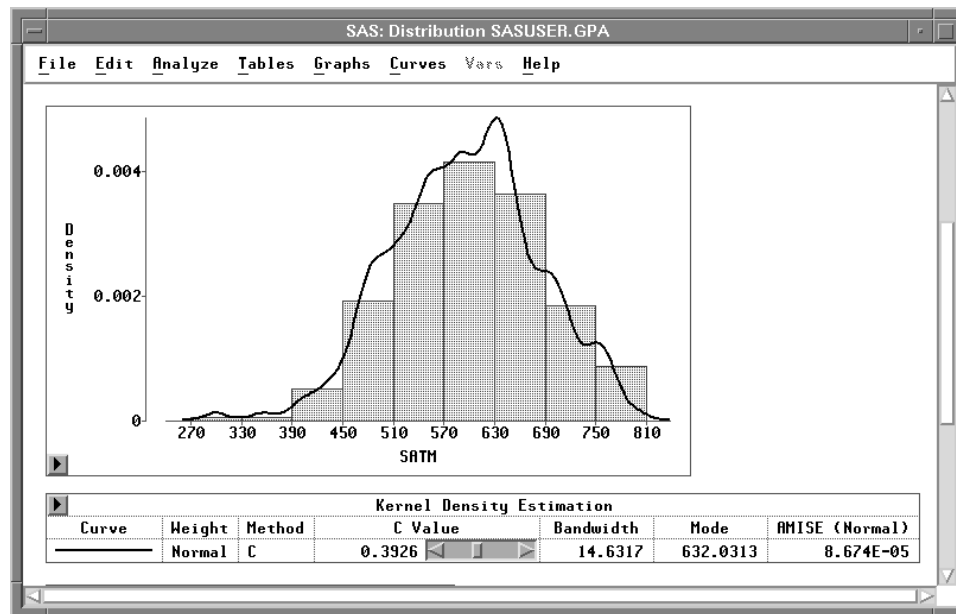


Figure 12.18. Kernel Density Estimate with a Smaller C Value

Testing Distributions

You can add a graph to examine the cumulative distribution function, and you can test for distributions by using the Kolmogorov statistic.

⇒ Choose **Curves:CDF Confidence Band:95%**.

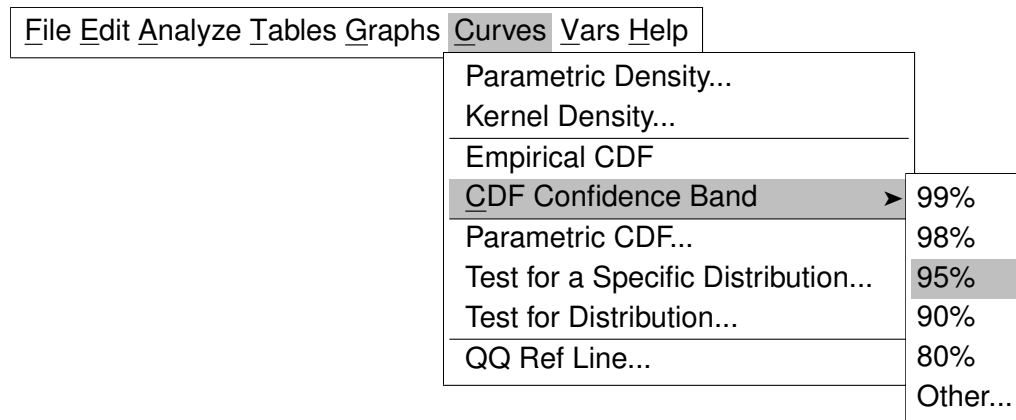


Figure 12.19. Confidence Band Menu

This adds a graph of the cumulative distribution function with 95% confidence bands, as illustrated in [Figure 12.20](#).

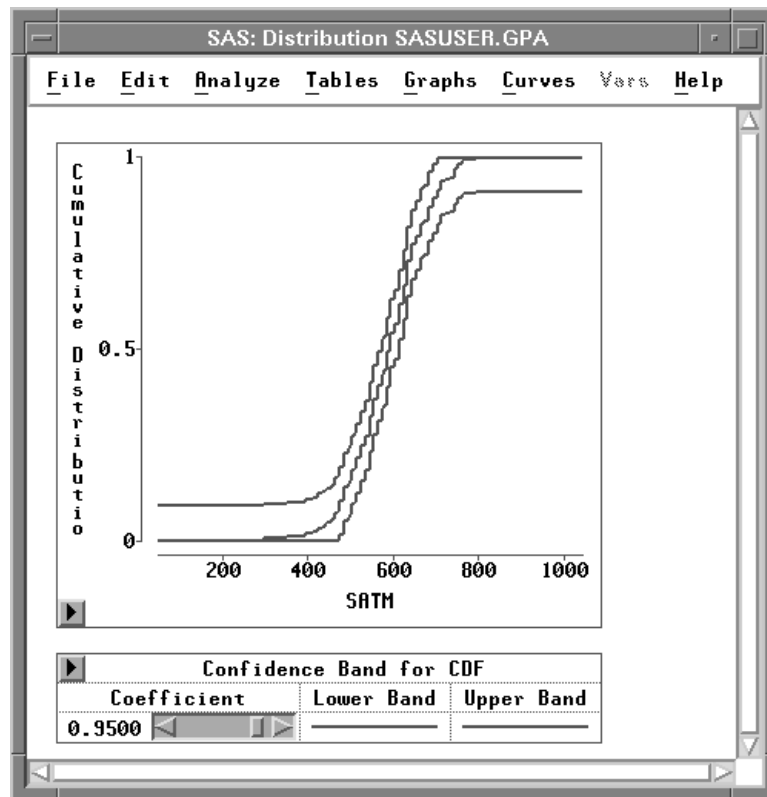


Figure 12.20. Cumulative Distribution Function

⇒ **Choose Curves:Test for Distribution.**

This displays the test for distribution dialog. The default settings test whether the data are from a normal distribution.

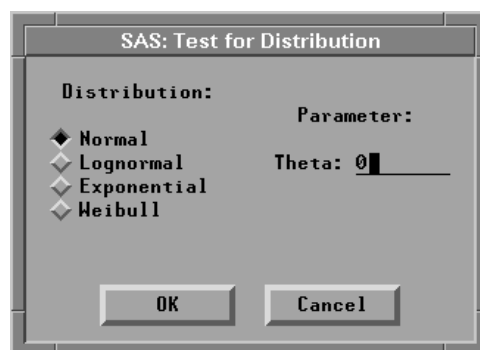


Figure 12.21. Test for Distribution Dialog

⇒ **Click OK in the dialog.**

This adds a curve to the graph and a **Test for Distribution** table to the window, as illustrated in [Figure 12.22](#).

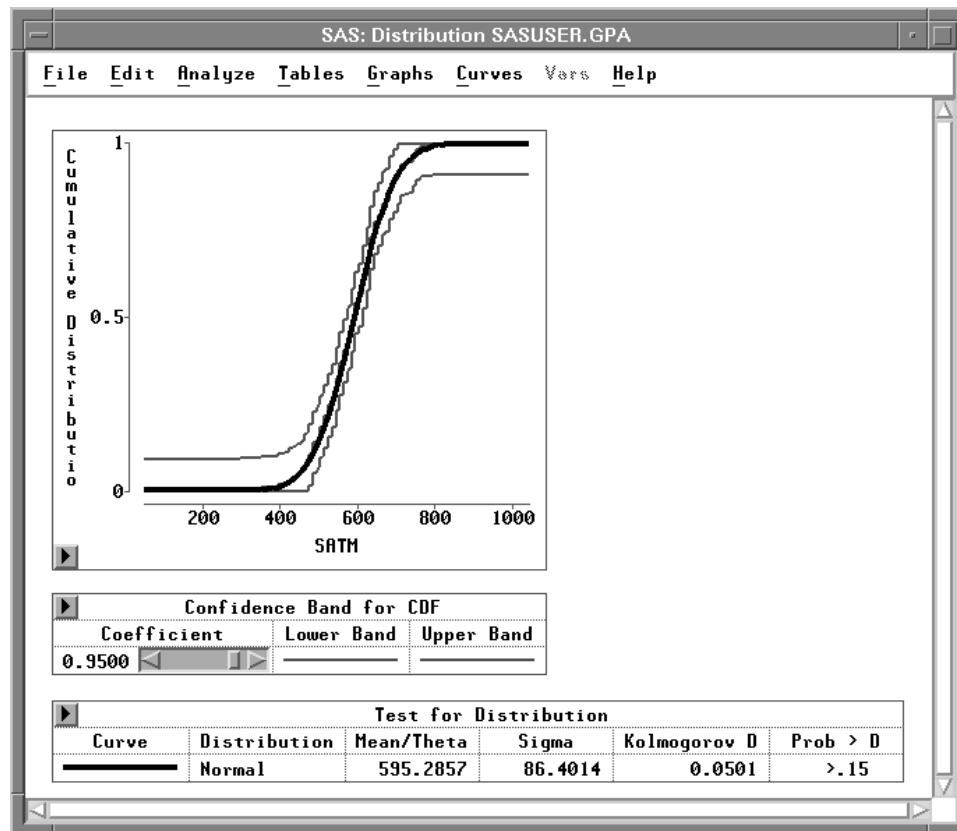


Figure 12.22. Test for Normal Distribution

The smooth curve in the graph represents the fitted normal distribution. It lies quite close to the irregular curve representing the empirical distribution function. The **Test for Distribution** table contains the mean (**Mean / Theta**) and standard deviation (**Sigma**) for the data along with the results of Kolmogorov's test for normality. This tests the null hypothesis that the data come from a normal distribution with unknown mean and variance. The p -value (**Prob > D**), also referred to as the *probability value* or *observed significance level*, is the probability of obtaining a D statistic greater than the computed D statistic when the null hypothesis is true. The smaller the p -value, the stronger the evidence against the null hypothesis. The computed p -value is large (**>0.15**), so there is no reason to conclude that these data are not normally distributed.

⊕ **Related Reading:** Distributions, [Chapter 38](#).

References

Moore, D.S. and McCabe, G.P. (1989), *Introduction to the Practice of Statistics*, New York: W.H. Freeman and Company.

Chapter 13

Fitting Curves

Chapter Contents

PARAMETRIC REGRESSION FITS	202
Changing the Polynomial Degree	204
Adding Curves	207
Line Colors, Patterns, and Widths	208
NONPARAMETRIC FITS	210
Normal Kernel Fit	211
Loess Smoothing	213
REFERENCES	215

Chapter 13

Fitting Curves

You can use **Fit (Y X)** to fit curves when you have one **X** variable. Curve-fitting helps you identify trends and relationships in two-dimensional data. SAS/INSIGHT software offers both parametric and nonparametric methods to fit curves. You can generate confidence ellipses, fit parametric polynomials with confidence curves, and fit nonparametric curves using spline, kernel, and loess estimators.

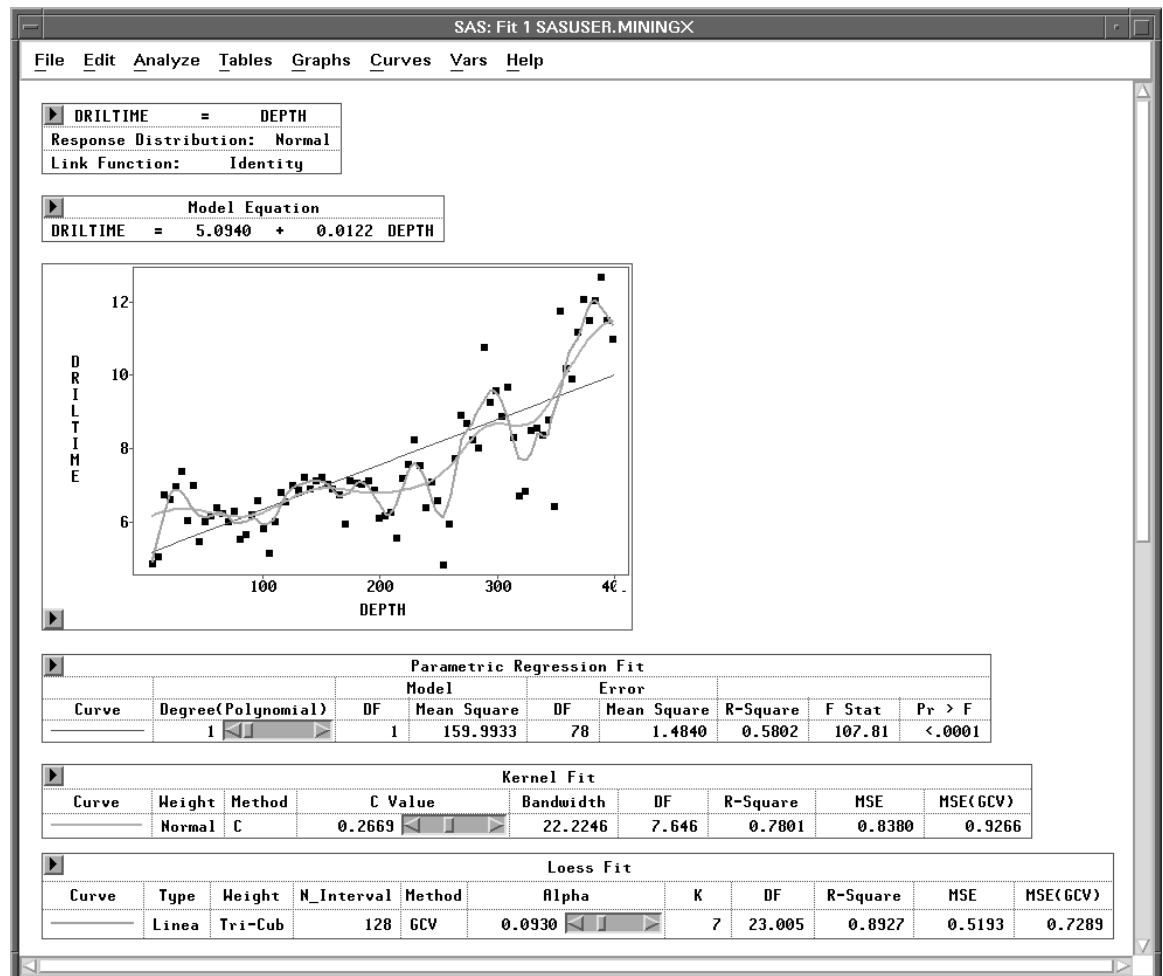


Figure 13.1. Fit Window with Several Curves

Parametric Regression Fits

Fitting a curve produces a visual display that reflects the systematic variation of the data. In this section, you will fit polynomial curves using a subset of the **MINING** data set described in [Chapter 1, “Getting Started.”](#)

⇒ **Open the MININGX data set.**

⇒ **Choose Analyze:Fit (Y X).**

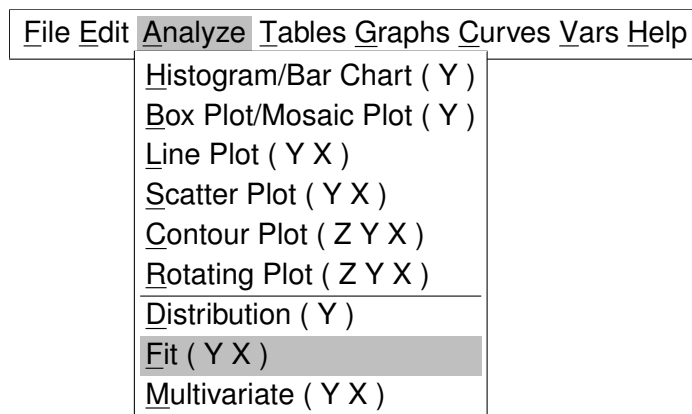


Figure 13.2. Analyze Menu

The fit variables dialog appears, as shown in [Figure 13.3.](#)

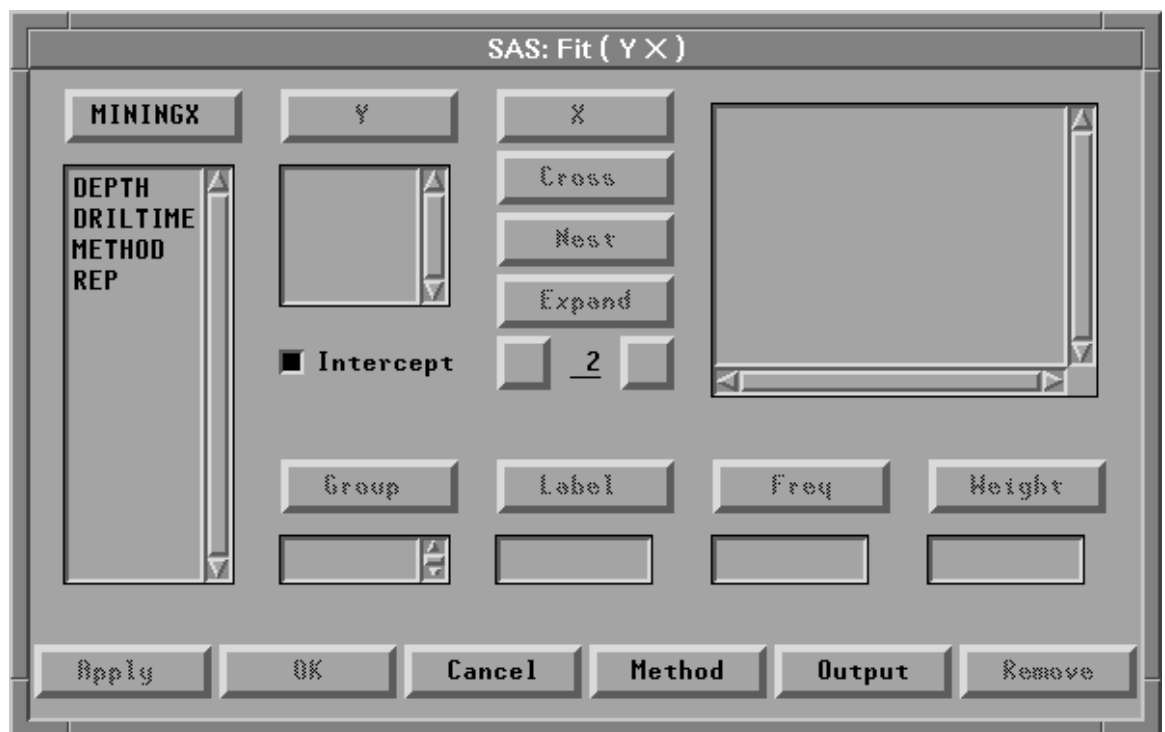


Figure 13.3. Fit Variables Dialog

⇒ Select the variable **DRILTIME**, then click the **Y** button.

DRILTIME appears in the **Y** variables list.

⇒ Select the variable **DEPTH**, then click the **X** button.

DEPTH appears in the **X** variables list.

⇒ Click the **Output** button.

The fit output options dialog, shown in [Figure 13.4](#), appears on your display.

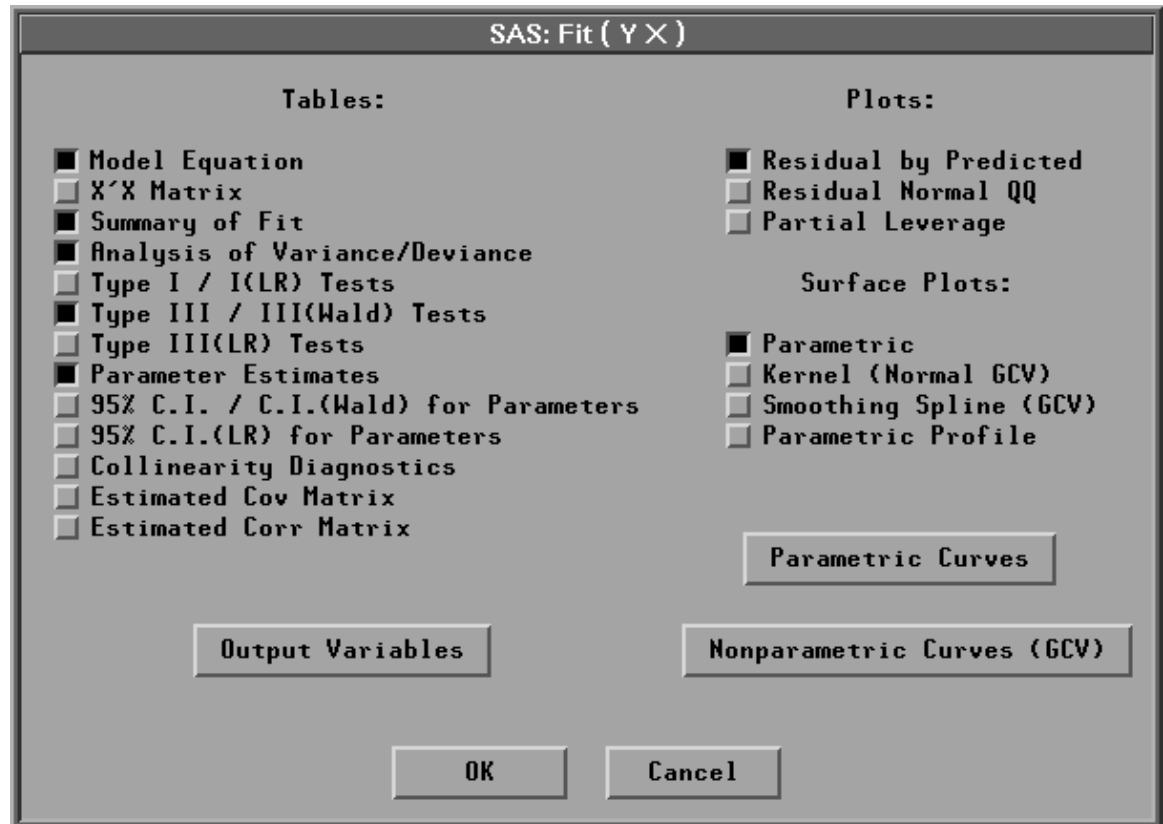


Figure 13.4. Fit Output Options Dialog

In the output options dialog, you specify which curves and tables will appear in the fit window. The default curve is a polynomial of degree one, that is, a line. The options set by default in this dialog are appropriate aids to a careful modeling of the data. They are not needed here where the purpose is to produce a visual display that reflects the trend of the data.

⇒ Turn off all check boxes by clicking on any that are highlighted.

⇒ Click the **OK** button in all dialogs.

A fit window appears, as shown in [Figure 13.5](#).

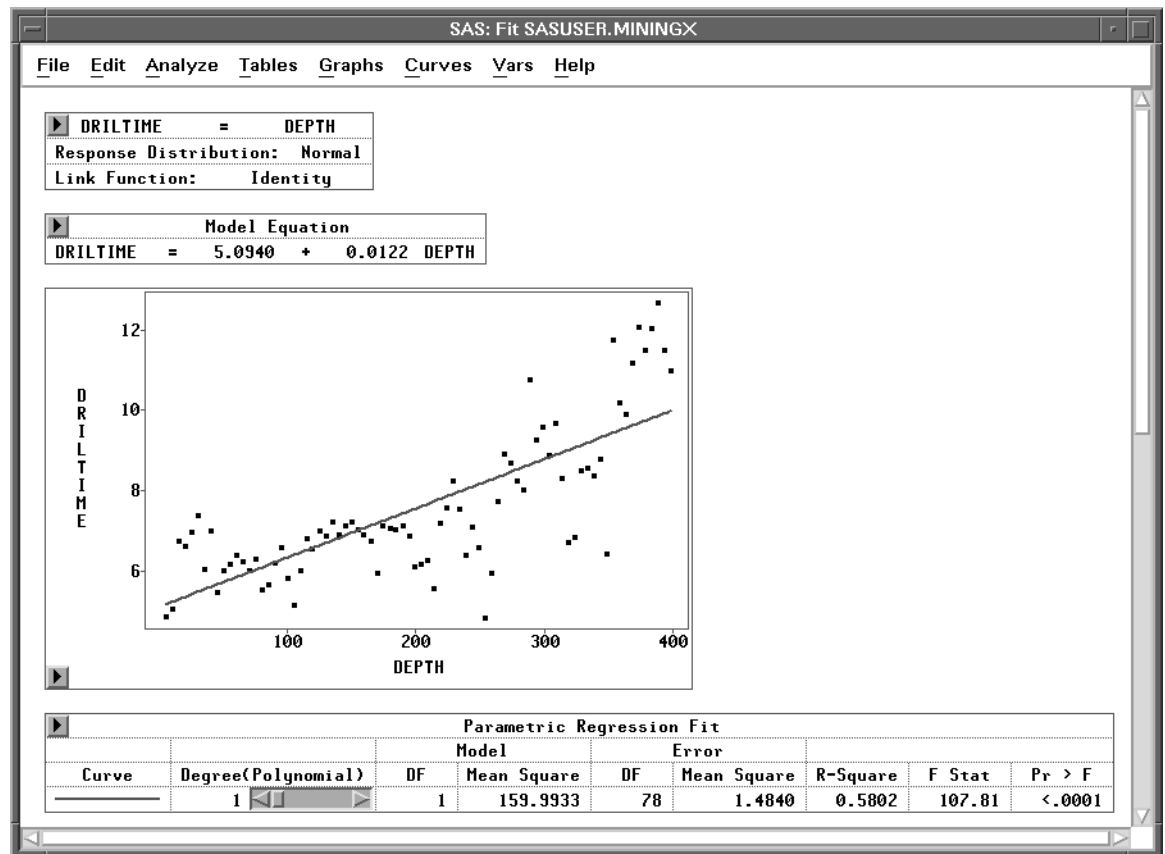


Figure 13.5. Fit Window with Line

The fit window contains a plot of **DRILTIME** by **DEPTH** along with a table summarizing the fit. A simple regression line is superimposed on the plot; it follows the *linear* trend of the data. Notice, though, that the plot shows curvature that a straight line cannot follow.

First examine the **Parametric Regression Fit** table corresponding to these data. The **R-Square** value is **0.5802**, which means that 58% of the variation in drilling times is explained by **DEPTH**. The rest of this table contains statistics pertinent to hypothesis testing, and they are discussed in [Chapter 14, “Multiple Regression.”](#)

Changing the Polynomial Degree

Examine the **Parametric Regression Fit** table in [Figure 13.6](#). Note that next to the polynomial degree is a slider that enables you to change the degree of polynomial fit to try to account for the curvature in the plot not explained by the straight line.

You can use the slider in three ways to adjust curves:

- click the arrow buttons
- click within the slider

- drag within the slider

⇒ **Click the left arrow button in the slider.**

This decreases the degree of the polynomial to zero. A zero-degree polynomial fit is just a mean line.

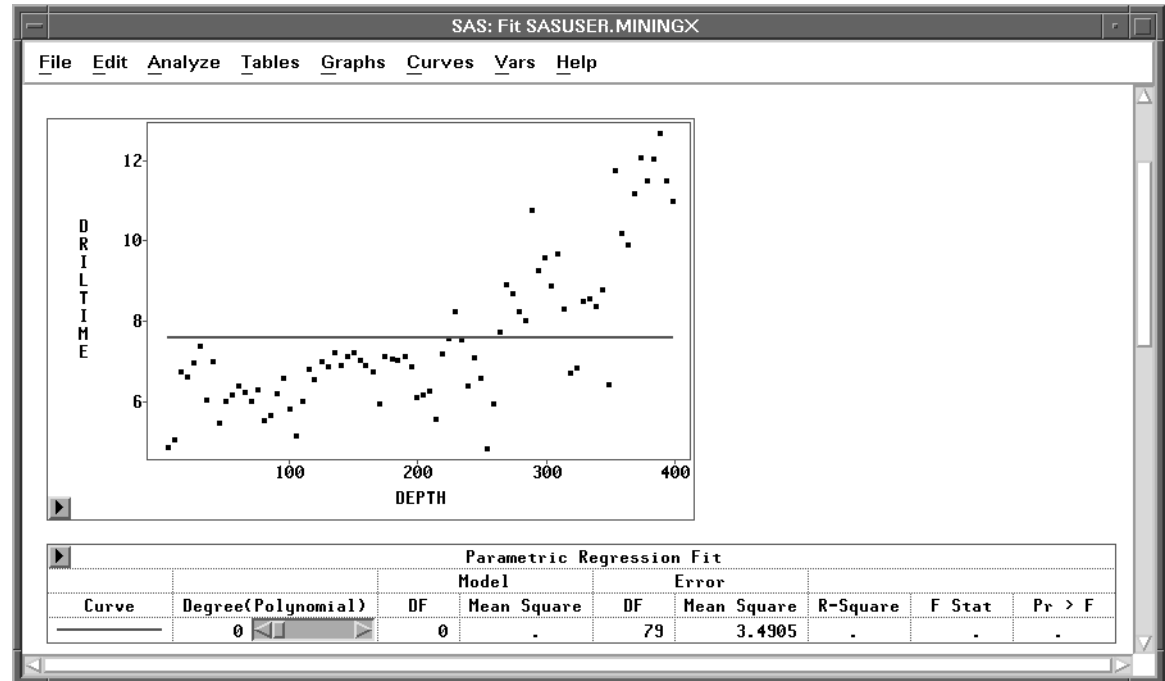


Figure 13.6. Fit Window with Mean Line

⇒ **Click twice on the right arrow button in the slider.**

This increases the polynomial degree to **2**, a quadratic fit, as shown in [Figure 13.7](#). The quadratic fit does a much better job accounting for the curvature in the plot. Note also that the **R-Square** value for the quadratic polynomial has increased to over 70%. You can fit successively higher-degree polynomials that continue to increase the **R-Square** value; but beyond a certain degree, small increases in **R-Square** do not compensate for the intuitive appeal in fitting a low degree polynomial.

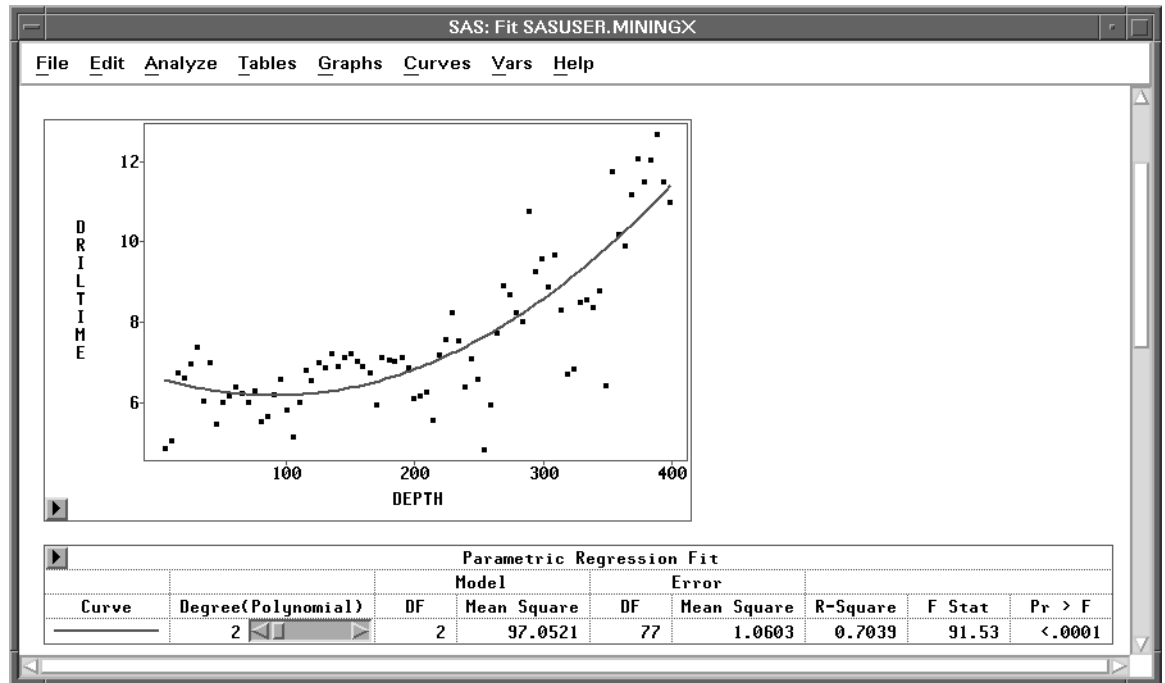


Figure 13.7. Quadratic Fit

⇒ **Click within the slider, just to the right of the slider control.**

This moves the slider control to the position where you click. The polynomial degree is set to a value proportional to the slider position. On most personal computers, clicking within the slider is the fastest way to adjust a curve.

⇒ **Drag the slider control left and right.**

When you drag the slider, its speed depends on the number of data points, the type of curve, and the speed of your host. Depending on your host, you may be able to improve the speed of the dynamic graphics with an alternate drawing algorithm. To try this, choose **Edit:Windows:Graph Options**, and set the **Fast Draw** option.

† **Note:** The **Degree(Polynomial)** is the degree being specified in the polynomial fit, and the **Model DF** is the polynomial degree actually fitted.

To avoid unnecessary computation, the maximum degree that can be actually fitted is not calculated, and the maximum **Degree(Polynomial)** in the slider is set to be the number of unique **X** variable values minus 1. When a polynomial term for the **X** variable in the specified polynomial fit is a linear combination of its lower polynomial terms, the **Degree(Polynomial)** will be greater than the **Model DF**; that is, the degree actually fitted is less than the degree specified in these cases..

Adding Curves

You can add curves to a scatter plot in the fit window in two ways. You can choose from the **Curves** menu or you can select **Edit:Windows:Renew** to reset the fit output options. When you add a curve from the **Curves** menu, SAS/INSIGHT adds either a new table entry or a whole new table that contains a summary of the new curve fit. Suppose you want to compare polynomial fits of different degree directly on the scatter plot. Begin by adding a second polynomial fit to the plot.

⇒ Choose **Curves:Polynomial**.

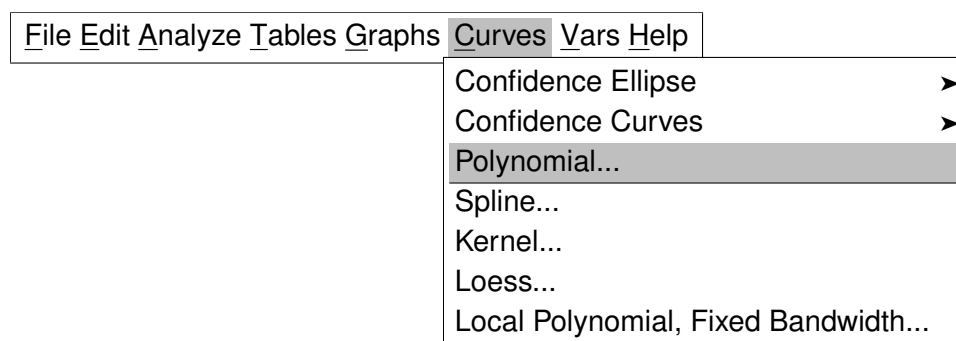


Figure 13.8. Curves Menu

This displays the polynomial fit dialog shown in [Figure 13.9](#).

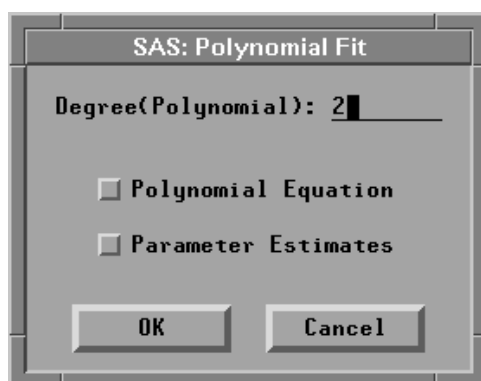


Figure 13.9. Polynomial Fit Dialog

⇒ Set the degree for the new polynomial to **3** and click **OK**.

This adds a cubic polynomial fit to the scatter plot, as shown in [Figure 13.10](#).

Now you have two polynomial fits in the window. Note that an entry for the cubic polynomial has been added to the **Parametric Regression Fit** table. Each entry in the table has its own slider so that you can adjust the degree of either polynomial to compare any pair of fits.

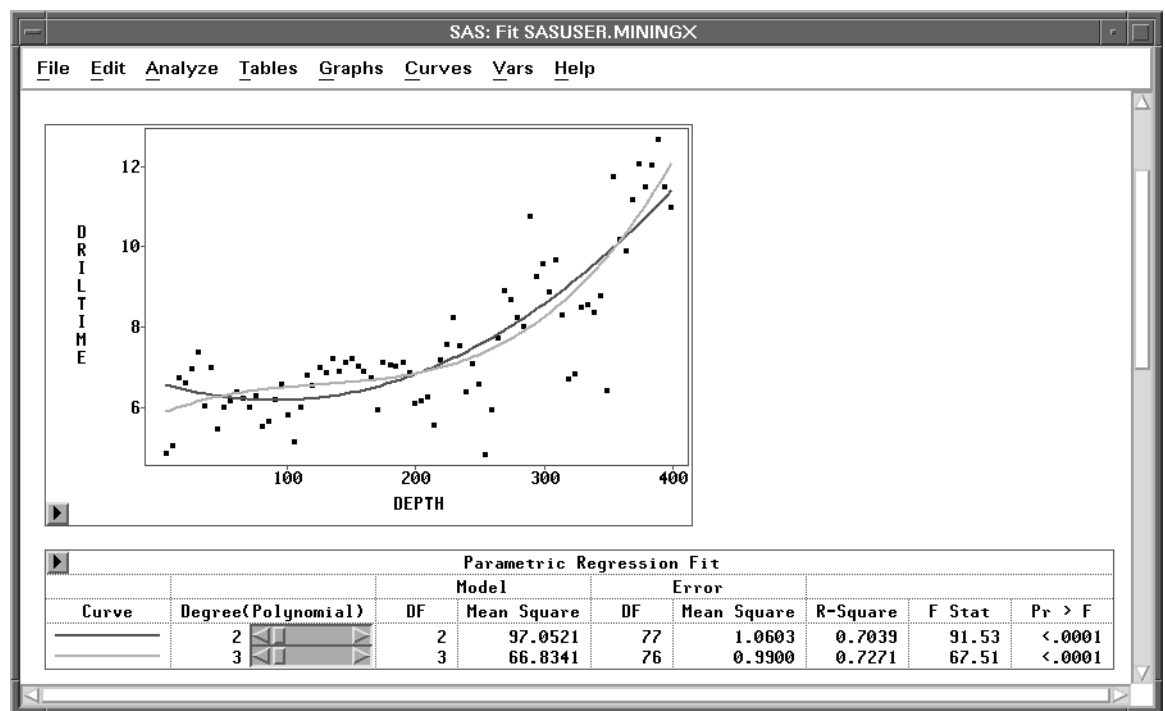


Figure 13.10. Fit Window with Two Polynomial Fits

Line Colors, Patterns, and Widths

Notice in Figure 13.10 that it is difficult to distinguish the two polynomial curves. On color displays, curve colors are chosen by default to contrast with the window background color and with existing curves. Curves are always drawn as solid lines by default. You can set default curve widths with display options. You can use the **Tools** window to change any of these curve features.

⇒ **Choose Edit:Windows:Tools to display the tools window.**

The tools window displays a palette of colors, three line patterns, and five curve widths that you can choose for the selected curve, as shown in Figure 13.11

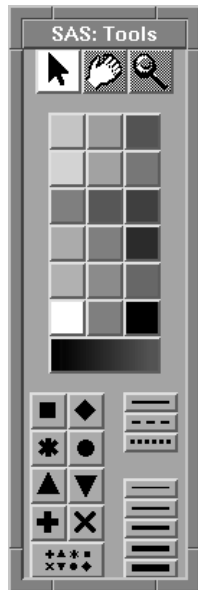


Figure 13.11. Tools Window

- ⇒ Click on the cubic fit curve legend to select the curve.
Clicking on either the legend or the curve highlights both the legend and the curve.

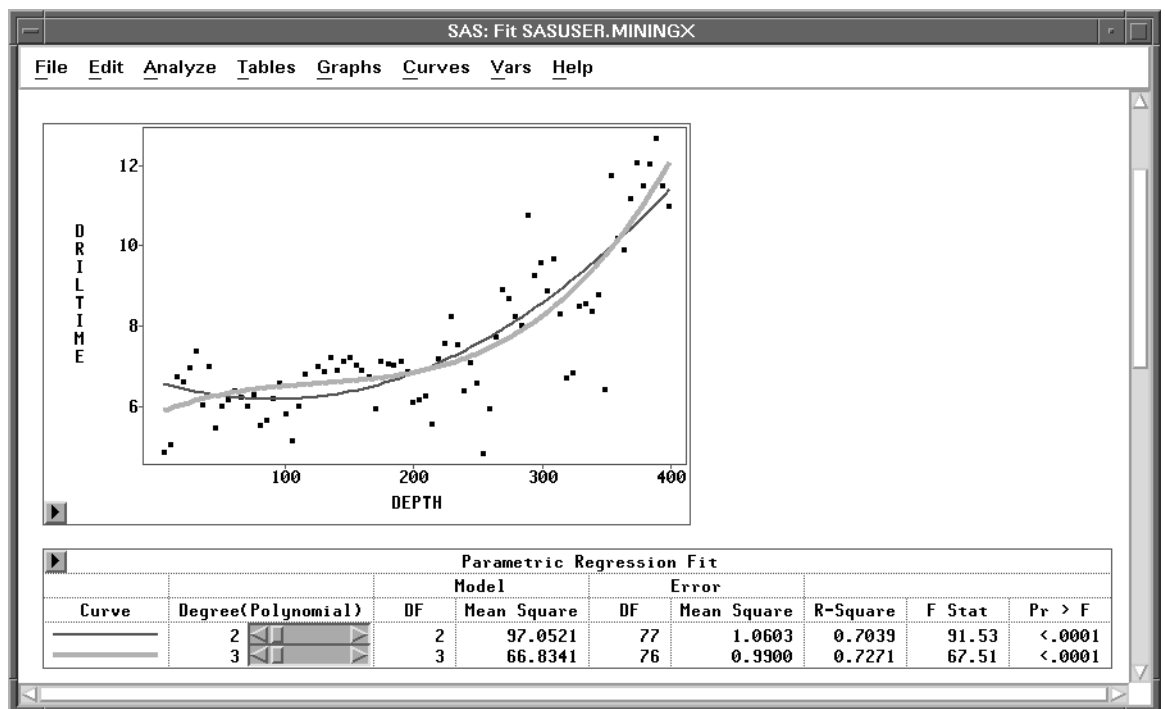


Figure 13.12. Cubic Fit Curve Selected

- ⇒ In the Tools window, click on the dotted line pattern.

Again note that the legend in the table matches the new curve pattern.

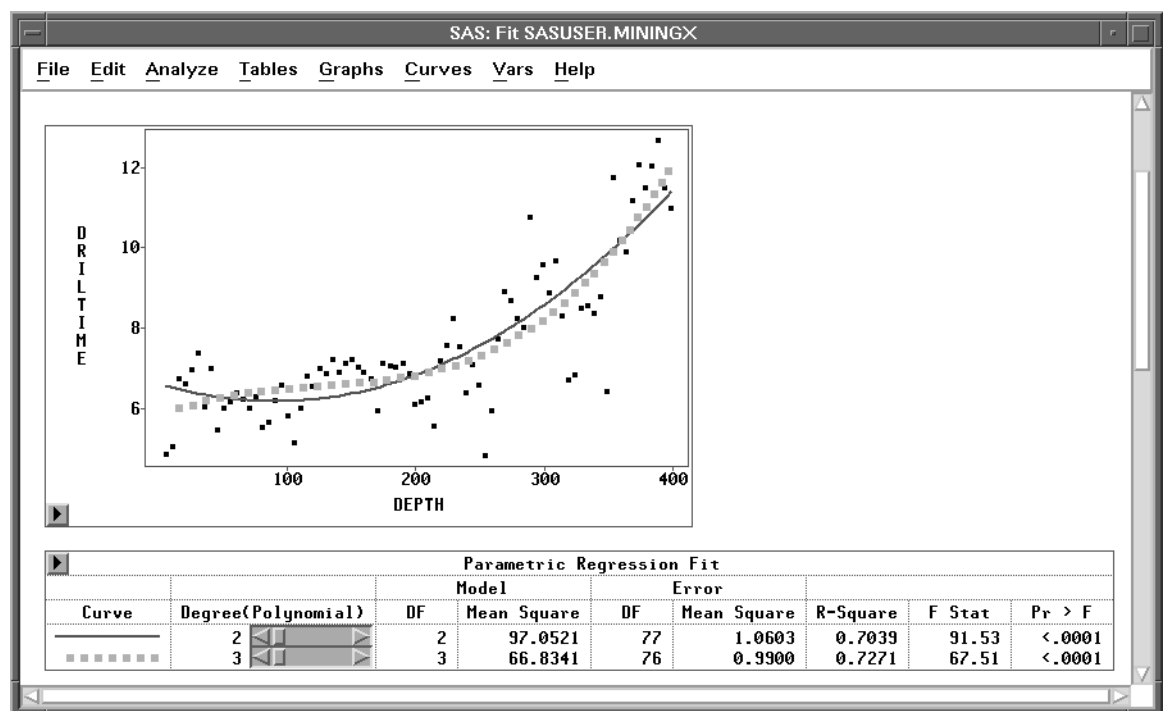


Figure 13.13. New Pattern for Cubic Fit

- ⇒ **Click in any blank area of the fit window to deselect the curve.**
You can select a curve again and try various colors, patterns, or widths.
- ⇒ **Select the Parametric Regression Fit table.**
- ⇒ **Choose Edit:Delete.**
The selected parametric regression fit table and its associated curves are deleted from the window.

Nonparametric Fits

SAS/INSIGHT software provides nonparametric curve-fitting estimates from smoothing spline, kernel, loess, and fixed bandwidth local polynomial estimators that are alternatives to fitting polynomials. Because nonparametric methods allow more flexibility for the functional dependence of Y on X than a typical parametric model does, nonparametric methods are well suited for situations where little is known about the process under study.

To carry out a nonparametric regression, you need first to determine the smoothness of the fit. With SAS/INSIGHT software, you can specify a particular value for a smoothing parameter, specify a particular degrees of freedom for a smoother, or request a default best fit. The data are then smoothed to estimate the regression curve. This is in contrast to the parametric regression where the degree of the polynomial controls the complexity of the fit. For the polynomial, additional complexity

can result in inappropriate global behavior. Nonparametric methods allow local use of additional complexity and thus are better tools to capture complex behavior than polynomials.

Normal Kernel Fit

To add a normal kernel estimate in the **MININGX** fit window from the preceding section, follow these steps.

⇒ **Choose Curves:Kernel.**

This displays the kernel fit dialog, as shown in [Figure 13.14](#).

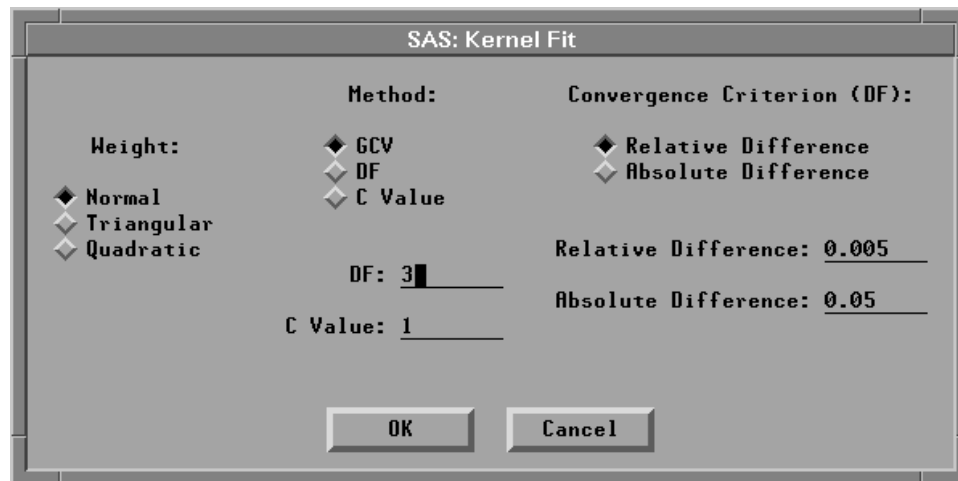


Figure 13.14. Kernel Fit Dialog

⇒ Click on **OK** in the dialog to display the kernel fit, as shown in [Figure 13.15](#).

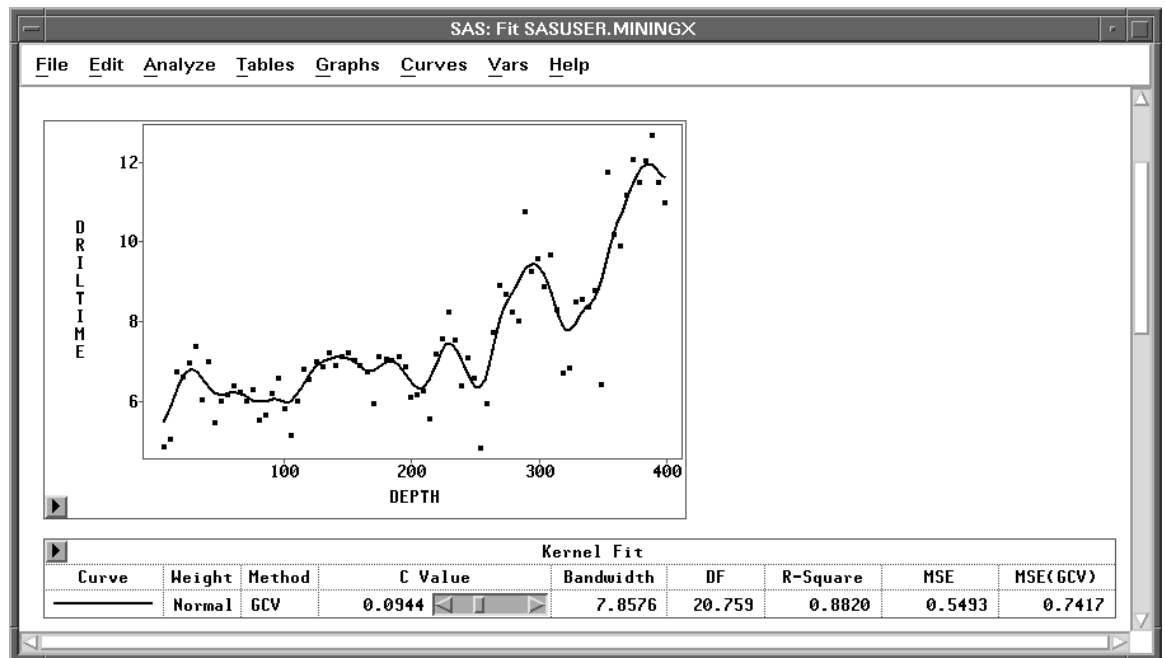


Figure 13.15. Normal Kernel Fit

By default, the optimal kernel smoothness is estimated based on mean square error using *generalized cross validation* (GCV). Cross validation leaves out points (x_i, y_i) one at a time and computes the kernel regression at x_i based on the remaining $n-1$ observations. Generalized cross validation is a weighted version of cross validation and is easier to compute. This estimation is carried out for a number of different values of the smoothing parameter, and the value that minimizes the estimated mean square error is selected (Hastie and Tibshirani 1990). This technique is described in detail in [Chapter 39, “Fit Analyses.”](#) Note that in [Figure 13.15](#), the **Kernel Fit** table shows the **Method** as **GCV**.

You can change the degree of smoothness by using the slider in the **Kernel Fit** table to change the value of c . Higher values of c result in smoother curves closer to a straight line; smaller values produce more flexible curves. It is often necessary to experiment with several values before finding one that fits your data well. See [Chapter 39, “Fit Analyses,”](#) for detailed information about kernels and the c parameter. Note that if you use the slider to change the value of c , the **Method** entry also changes.

The **Kernel Fit** table contains several statistics for comparing the kernel fit to other fits. The table contains the bandwidth or smoothing parameter of the kernel that corresponds to the value of c . The column labeled **DF** gives the approximate degrees of freedom for the kernel fit. Smoother curves have fewer degrees of freedom and result in lower values of R^2 and possibly higher values of mean square error. **R-square** measures the proportion of the total variation accounted for by the kernel fit. **MSE(GCV)** is an estimate of the mean square error using generalized cross validation. These statistics are also discussed in [Chapter 39, “Fit Analyses.”](#)

This kernel tracks the data fairly well. The fit requires 20.759 degrees of freedom, in-

dicating that the model may still be under-smoothed. The generalized cross validation method often results in under-smoothed fits, particularly with small data sets (Hastie and Tibshirani 1990). In this case, the data were collected from a single drilling hole, and this can lead to spurious cyclical patterns in the data caused by autocorrelation. The curve may be tracking these cycles. A smoother fit is probably desirable.

⇒ **Click three times on the right arrow in the slider.**

This results in a smoother kernel fit, as shown in [Figure 13.16](#).

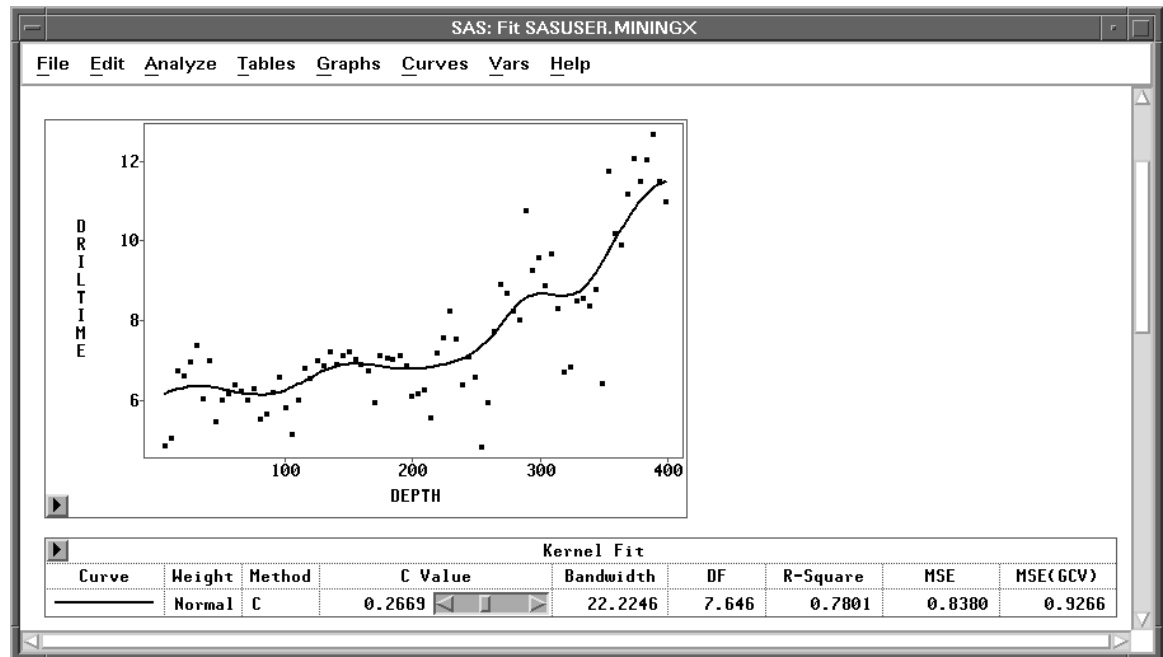


Figure 13.16. Normal Kernel Fit Made Smoother

Loess Smoothing

Loess smoothing is a curve-fitting technique based on local regression (Cleveland 1993). To fit a loess curve to the mining data, follow these steps:

⇒ **Choose **Curves:Loess** to display the loess fit dialog.**



Figure 13.17. Loess Fit Dialog

⇒ **Click on OK in the dialog to display the loess fit, as shown in Figure 13.18.**

As with the kernel fit, the best fit for loess smoothing is determined by generalized cross validation (GCV). GCV and other aspects of curve-fitting are described in Chapter 39, “Fit Analyses.”

You can also output predicted values from fitted curves. To output predicted values from the preceding loess fit, do the following:

⇒ **Choose Vars:Predicted Curves:Loess.**

This displays the same loess fit dialog as shown in Figure 13.17.

⇒ **Click on OK in the dialog to output the predicted values from the loess fit**

A new variable, **PL_DRILT**, should now be added to the data window.

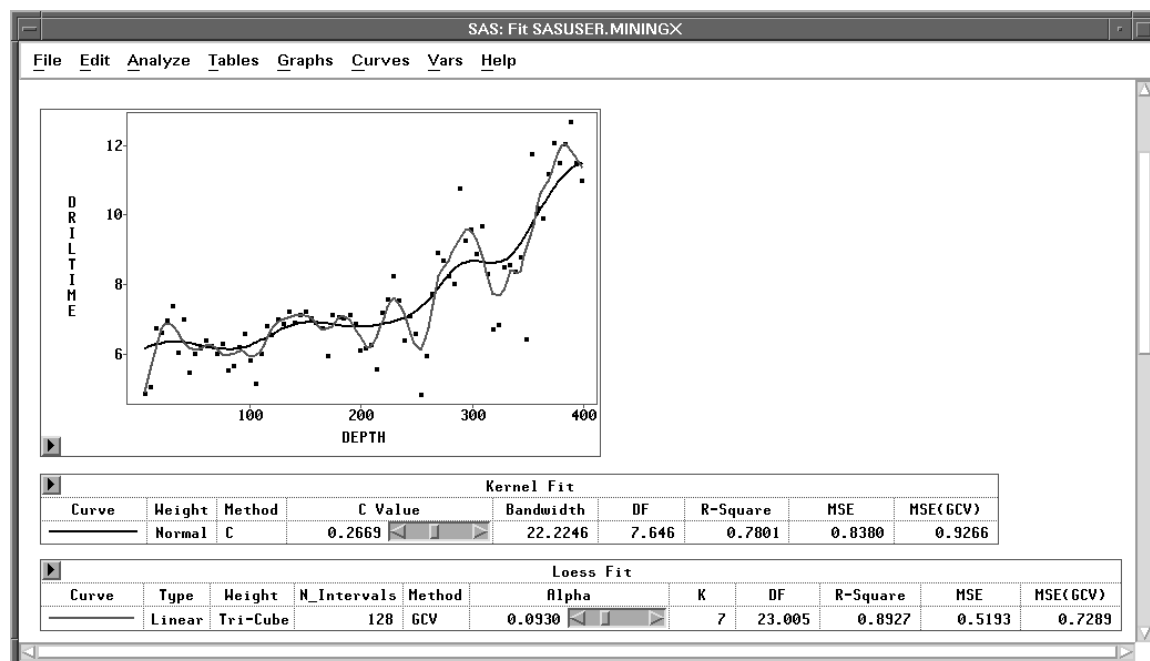


Figure 13.18. Loess Fit

You can use the slider control to adjust the loess curve just as with other curves. For loess, the slider controls the α value for the fit. The greater the α value, the smoother the fit.

On rare occasions, you may want to fit a curve for α values outside the bounds of the slider. For loess and other curves, the bounds of the slider are chosen for best fit in most cases. If you need to fit a curve with unusual parameter values, you can specify these values in the curve dialog.

⊕ **Related Reading:** Fit Curves, [Chapter 39](#).

References

- Cleveland, W.S. (1993), *Visualizing Data*, Summit, New Jersey: Hobart Press.
- Hastie, Y.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.
- McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman and Hall.
- Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.

Chapter 14

Multiple Regression

Chapter Contents

CREATING THE ANALYSIS	220
Model Information	225
Summary of Fit	225
Analysis of Variance	225
Type III Tests	226
Parameter Estimates	226
Residuals-by-Predicted Plot	227
ADDING TABLES AND GRAPHS	228
Collinearity Diagnostics Table	228
Partial Leverage Plots	229
Residual-by-Hat Diagonal Plot	230
MODIFYING THE MODEL	235
SAVING THE RESIDUALS	238
REFERENCES	239

Chapter 14

Multiple Regression

You can create multiple regression models quickly using the fit variables dialog. You can use diagnostic plots to assess the validity of the models and identify potential outliers and influential observations. You can save residuals and other output variables from your models for future analysis.

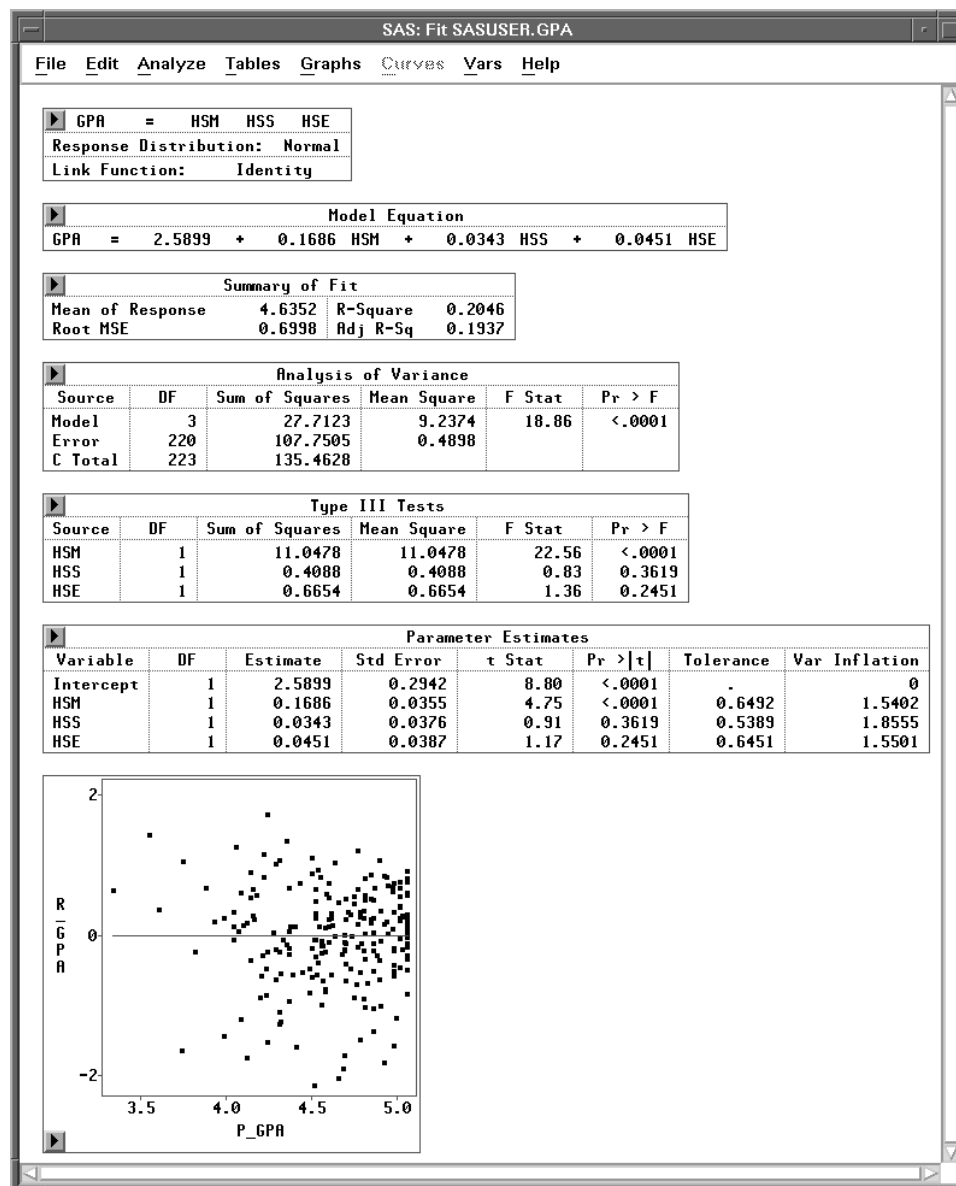


Figure 14.1. Multiple Regression Analysis

Creating the Analysis

The **GPA** data set contains data collected to determine which applicants at a large midwestern university were likely to succeed in its computer science program. The variable **GPA** is the measure of success of students in the computer science program, and it is the response variable. A *response variable* measures the outcome to be explained or predicted.

Several other variables are also included in the study as possible explanatory variables or predictors of **GPA**. An *explanatory variable* may explain variation in the response variable. Explanatory variables for this example include average high school grades in mathematics (**HSM**), English (**HSE**), and science (**HSS**) (Moore and McCabe 1989).

To begin the regression analysis, follow these steps.

⇒ **Open the GPA data set.**

⇒ **Choose Analyze:Fit (Y X).**

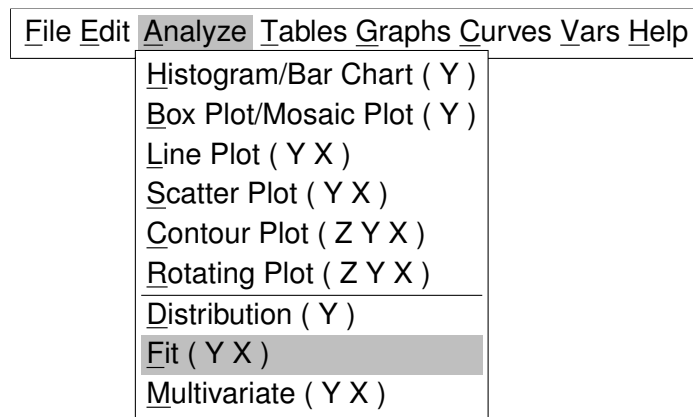


Figure 14.2. Analyze Menu

The fit variables dialog appears, as shown in [Figure 14.3](#). This dialog differs from all other variables dialogs because it can remain visible even after you create the fit window. This makes it convenient to add and remove variables from the model. To make the variables dialog stay on the display, click on the **Apply** button when you are finished specifying the model. Each time you modify the model and use the **Apply** button, a new fit window appears so you can easily compare models. Clicking on **OK** also displays a new fit window but closes the dialog.

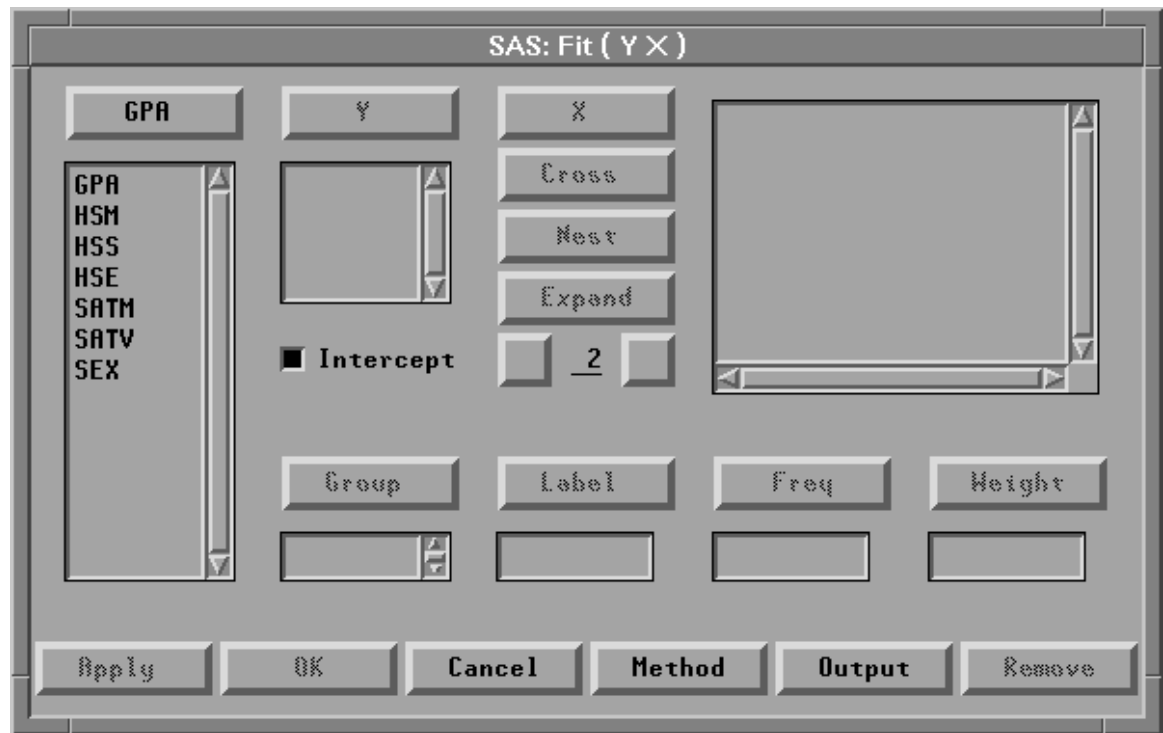


Figure 14.3. Fit Variables Dialog

- ⇒ Select the variable **GPA** in the list on the left, then click the **Y** button.
GPA appears in the **Y** variables list.
- ⇒ Select the variables **HSM**, **HSS**, and **HSE**, then click the **X** button.
HSM, **HSS**, and **HSE** appear in the **X** variables list.



Figure 14.4. Variable Roles Assigned

⇒ **Click the Apply button.**

A fit window appears, as shown in [Figure 14.5](#).

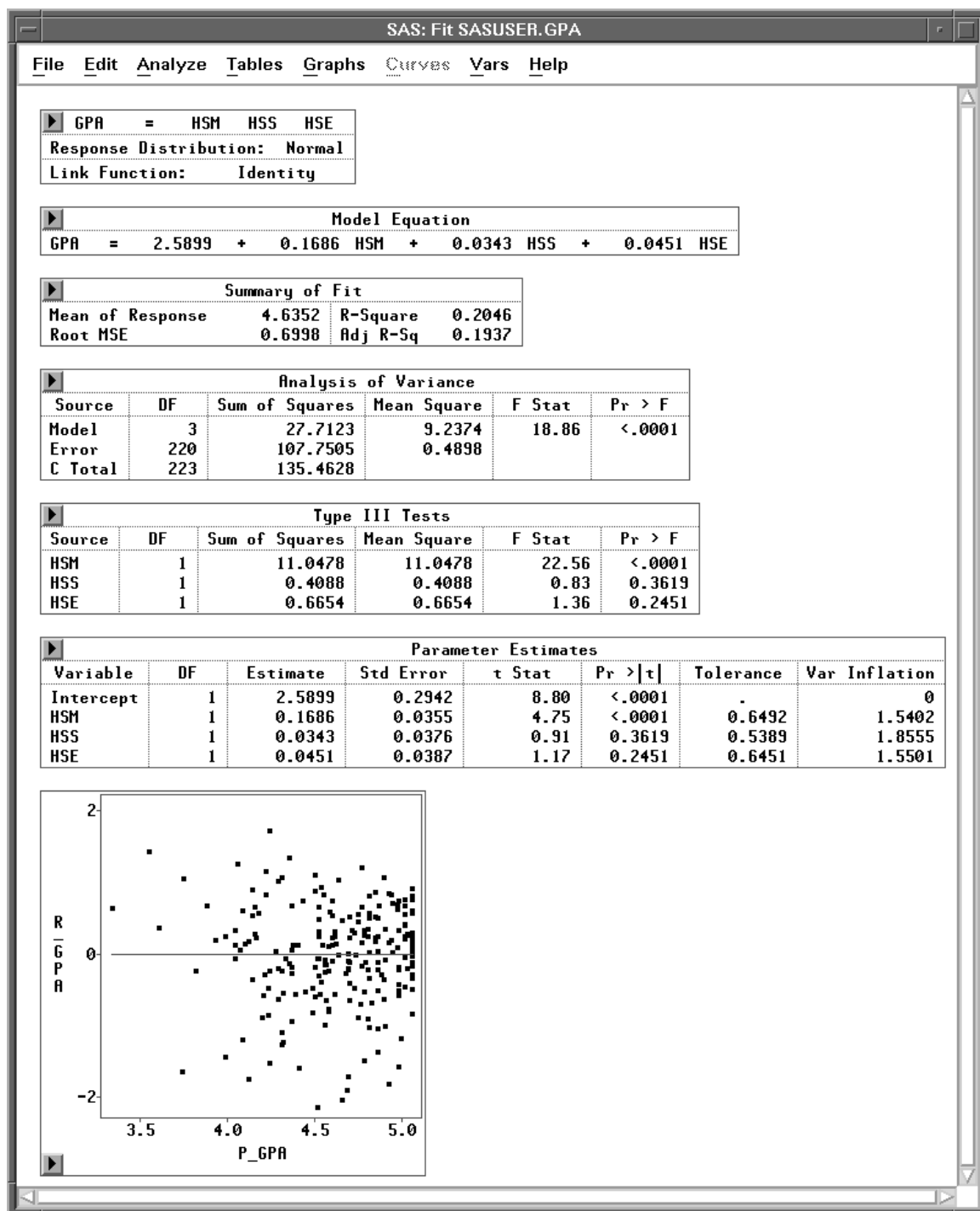


Figure 14.5. Fit Window

This window shows the results of a regression analysis of **GPA** on **HSM**, **HSS**, and **HSE**. The regression model for the i th observation can be written as

$$\text{GPA}_i = \beta_0 + \beta_1 \text{HSM}_i + \beta_2 \text{HSS}_i + \beta_3 \text{HSE}_i + \epsilon_i$$

Techniques ♦ *Multiple Regression*

where GPA_i is the value of GPA; β_0 to β_3 are the regression coefficients (parameters); HSM_i , HSS_i , and HSE_i are the values of the explanatory variables; and ϵ_i is the random error term. The ϵ_i 's are assumed to be uncorrelated, with mean 0 and variance σ^2 .

By default, the fit window displays tables for model information, **Model Equation**, **Summary of Fit**, **Analysis of Variance**, **Type III Tests**, and **Parameter Estimates**, and a residual-by-predicted plot, as illustrated in Figure 14.5. You can display other tables and graphs by clicking on the **Output** button on the fit variables dialog or by choosing menus as described in the section “Adding Tables and Graphs” later in this chapter.

Model Information

Model information is contained in the first two tables in the fit analysis. The first table displays the model specification, the response distribution, and the link function. The **Model Equation** table writes out the fitted model using the estimated regression coefficients $\hat{\beta}_0$ to $\hat{\beta}_3$:

$$\hat{\text{GPA}} = 2.5899 + 0.1686 \text{ HSM} + 0.0343 \text{ HSS} + 0.0451 \text{ HSE}$$

Summary of Fit

The **Summary of Fit** table contains summary statistics including **Root MSE** and **R-Square**. The **Root MSE** value is **0.6998** and is the square root of the mean square error given in the **Analysis of Variance** table. **Root MSE** is an estimate of σ in the preceding regression model.

The **R-Square** value is **0.2046**, which means that 20% of the variation in **GPA** scores is explained by the fitted model. The **Summary of Fit** table also contains an adjusted R-square value, **Adj R-Sq**. Because **Adj R-Sq** is adjusted for the number of parameters in the model, it is more comparable over models involving different numbers of parameters than **R-Square**.

Analysis of Variance

The **Analysis of Variance** table summarizes information about the sources of variation in the data. **Sum of Squares** represents variation present in the data. These values are calculated by summing squared deviations. In multiple regression, there are three sources of variation: **Model**, **Error**, and **C Total**. **C Total** is the total sum of squares corrected for the mean, and it is the sum of **Model** and **Error**. Degrees of Freedom, **DF**, are associated with each sum of squares and are related in the same way. **Mean Square** is the **Sum of Squares** divided by its associated **DF** (Moore and McCabe 1989).

If the data are normally distributed, the ratio of the **Mean Square** for the **Model** to the **Mean Square** for **Error** is an *F statistic*. This *F* statistic tests the null hypothesis that *none* of the explanatory variables has any effect (that is, that the regression coefficients β_1 , β_2 , and β_3 are all zero). In this case the computed *F* statistic (labeled **F Stat**) is 18.8606. You can use the *p*-value (labeled **Pr > F**) to determine whether to reject the null hypothesis. The *p*-value, also referred to as the *probability* value or *observed* significance level, is the probability of obtaining, by chance alone, an *F* statistic greater than the computed *F* statistic when the null hypothesis is true. The smaller the *p*-value, the stronger the evidence against the null hypothesis.

In this example, the p -value is so small that you can clearly reject the null hypothesis and conclude that at least one of the explanatory variables has an effect on **GPA**.

Type III Tests

The **Type III Tests** table presents the Type III sums of squares associated with the estimated coefficients in the model. Type III sums of squares are commonly called partial sums of squares (for a complete discussion, refer to the chapter titled “The Four Types of Estimable Functions” in the *SAS/STAT User’s Guide*). The Type III sum of squares for a particular variable is the increase in the model sum of squares due to adding the variable to a model that already contains all the other variables in the model. Type III sums of squares, therefore, do not depend on the order in which the explanatory variables are specified in the model. Furthermore, they do not yield an additive partitioning of the **Model** sum of squares unless the explanatory variables are uncorrelated (which they are not for this example).

F tests are formed from this table as explained previously in the “[Analysis of Variance](#)” section. Note that when **DF = 1**, the Type III F statistic for a given parameter estimate is equal to the square of the t statistic for the same parameter estimate. For example, the **T Stat** value for **HSM** given in the **Parameter Estimates** table is **4.7494**. The corresponding **F Stat** value in the **Type III Tests** table is **22.5569**, which is **4.7494** squared.

Parameter Estimates

The **Parameter Estimates** table, as shown in [Figure 14.5](#), displays the parameter estimates and the corresponding degrees of freedom, standard deviation, t statistic, and p -values. Using the parameter estimates, you can also write out the fitted model:

$$\hat{\text{GPA}} = 2.5899 + 0.1686\text{HSM} + 0.0343\text{HSS} + 0.0451\text{HSE}.$$

The t statistic is used to test the null hypothesis that a parameter is 0 in the model. In this example, only the coefficient for **HSM** appears to be statistically significant ($p \leq 0.0001$). The coefficients for **HSS** and **HSE** are not significant, partly because of the relatively high correlations among the three explanatory variables. Once **HSM** is included in the model, adding **HSS** and **HSE** does not substantially improve the model fit. Thus, their corresponding parameters are not statistically significant.

Two other statistics, tolerance and variance inflation, also appear in the **Parameter Estimates** table. These measure the strength of interrelationships among the explanatory variables in the model. Tolerances close to 0 and large variance inflation factor values indicate strong linear association or collinearity among the explanatory variables (Rawlings 1988, p. 277). For the **GPA** data, these statistics signal no problems of collinearity, even for **HSE** and **HSS**, which are the two most highly correlated variables in the data set.

Residuals-by-Predicted Plot

SAS/INSIGHT software provides many diagnostic tools to help you decide if your regression model fits well. These tools are based on the *residuals* from the fitted model. The residual for the i th observation is the observed value minus the predicted value:

$$\text{GPA}_i - \hat{\text{GPA}}_i.$$

The plot of the residuals versus the predicted values is a classical diagnostic tool used in regression analysis. The plot is useful for discovering poorly specified models or heterogeneity of variance (Myers 1986, pp. 138–139). The plot of **R_GPA** versus **P_GPA** in [Figure 14.5](#) indicates no such problems. The observations are randomly scattered above and below the zero line, and no observations appear to be outliers.

Adding Tables and Graphs

The menus at the top of the fit window enable you to add tables and graphs to the fit window and output variables to the data window. When there is only one **X** variable, you can also fit curves as described in [Chapter 13, “Fitting Curves.”](#)

Following are some examples of tables and graphs you can add to a fit window.

Collinearity Diagnostics Table

⇒ Choose **Tables:Collinearity Diagnostics.**

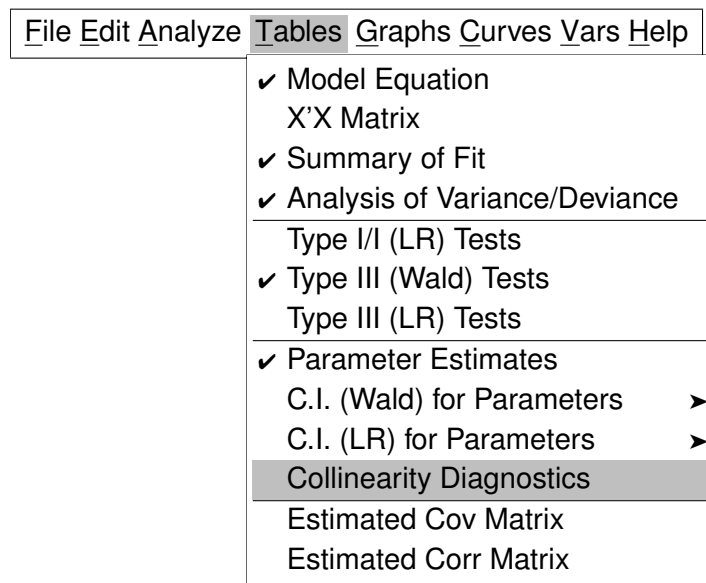


Figure 14.6. Tables Menu

This displays the table shown in [Figure 14.7.](#)

The screenshot shows the SAS Fit window titled 'SAS: Fit SASUSER.GPA'. The 'Tables' menu is open, and the 'Collinearity Diagnostics' table is displayed. The table has 7 columns: Number, Eigenvalue, Condition Index, Intercept, HSM, HSS, and HSE. The data is as follows:

Number	Eigenvalue	Condition Index	Variance Proportion			
			Intercept	HSM	HSS	HSE
1	3.9453	1.0000	0.0016	0.0015	0.0014	0.0014
2	0.0216	13.5089	0.6378	0.1018	0.3579	0.0217
3	0.0195	14.2404	0.0443	0.6337	0.0979	0.4110
4	0.0136	17.0416	0.3163	0.2630	0.5428	0.5660

Figure 14.7. Collinearity Diagnostics Table

When an explanatory variable is nearly a linear combination of other explanatory variables in the model, the estimates of the coefficients in the regression model are unstable and have high standard errors. This problem is called *collinearity*. The **Collinearity Diagnostics** table is calculated using the eigenstructure of the $X'X$ matrix. See [Chapter 13, “Fitting Curves,”](#) for a complete explanation.

A collinearity problem exists when a component associated with a high condition index contributes strongly to the variance of two or more variables. The highest condition number in this table is **17.0416**. Belsley, Kuh, and Welsch (1980) propose that a condition index of 30 to 100 indicates moderate to strong collinearity.

Partial Leverage Plots

Another diagnostic tool available in the fit window is partial leverage plots. When there is more than one explanatory variable in a model, the relationship of the residuals to one explanatory variable can be obscured by the effects of other explanatory variables. Partial leverage plots attempt to reveal these relationships (Rawlings 1988, pp. 265–266).

⇒ Choose **Graphs:Partial Leverage**.

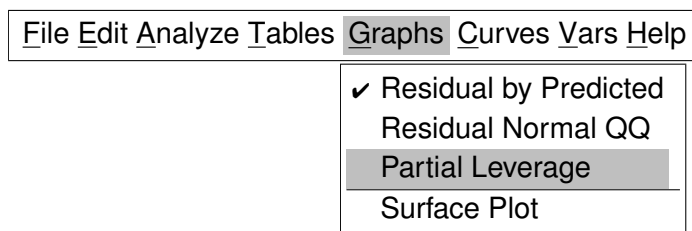


Figure 14.8. Graphs Menu

This displays the graphs shown in [Figure 14.9](#).

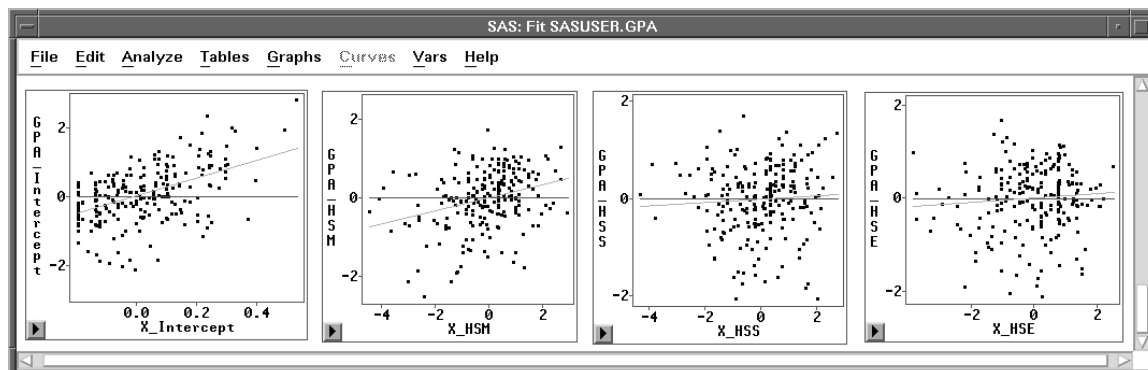


Figure 14.9. Partial Leverage Plots

In each plot in [Figure 14.9](#), the x-axis represents the residuals of the explanatory variable from a model that regresses that explanatory variable on the remaining explanatory variables. The y-axis represents the residuals of the response variable calculated with the explanatory variable omitted.

Two reference lines appear in each plot. One is the horizontal line $Y=0$, and the other is the fitted regression line with slope equal to the parameter estimate of the corresponding explanatory variable from the original regression model. The latter line shows the effect of the variable when it is added to the model last. An explanatory variable having little or no effect results in a line close to the horizontal line $Y=0$.

Examine the slopes of the lines in the partial leverage plots. The slopes for the plots representing **HSS** and **HSE** are nearly 0. This is not surprising since the coefficients for the parameter estimates of these two explanatory variables are nearly 0. You will examine the effect of removing these two variables from the model in the section “[Modifying the Model](#)” later in this chapter.

Curvilinear relationships not already included in the model may also be evident in a partial leverage plot (Rawlings 1988). No curvilinearity is evident in any of these plots.

Residual-by-Hat Diagonal Plot

The fit window contains additional diagnostic tools for examining the effect of observations. One such tool is the residual-by-hat diagonal plot. *Hat diagonal* refers to the diagonal elements of the hat matrix (Rawlings 1988). Hat diagonal measures the leverage of each observation on the predicted value for that observation.

Choosing **Fit (Y X)** does not automatically generate the residual-by-hat diagonal plot, but you can easily add it to the fit window. First, add the hat diagonal variable to the data window.

⇒ **Choose Vars:Hat Diag.**

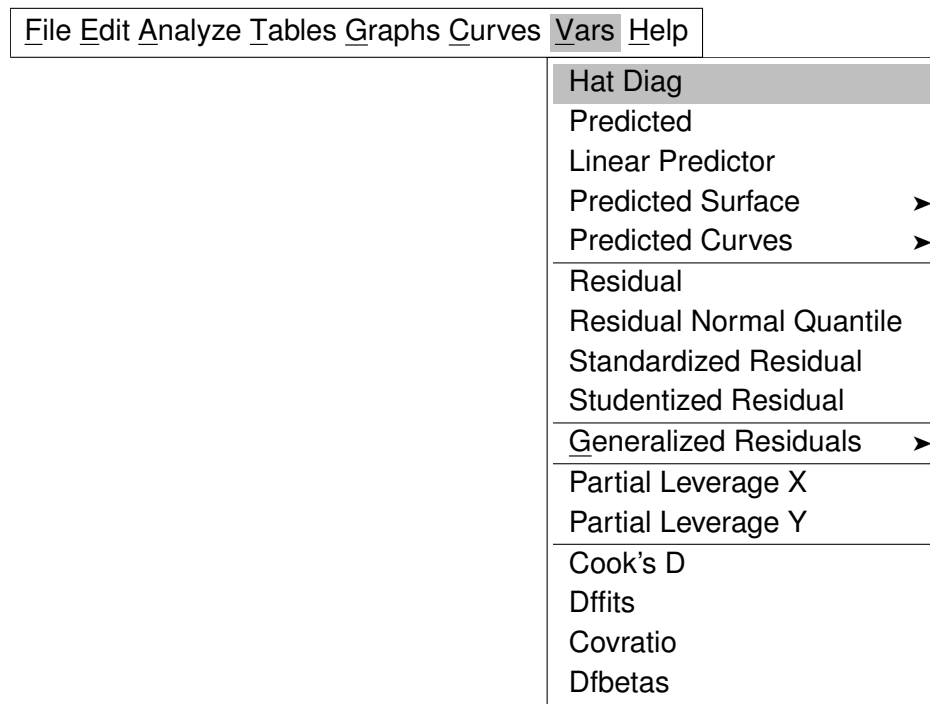


Figure 14.10. Vars Menu

This adds the variable **H_GPA** to the data window, as shown in Figure 14.11. (The residual variable, **R_GPA**, is added when a residual-by-predicted plot is created.)

	18	Int	Int	Int	Int	Int
	X_HSS	GPA_HSS	X_HSE	GPA_HSE	H_GPA	
1	0.3568	0.2625	0.8277	0.2876	0.0132	
2	-0.2327	0.2652	1.4111	0.3368	0.0119	
3	-1.4180	-0.7921	-1.3066	-0.8024	0.0249	
4	-0.1895	0.3432	0.2551	0.3612	0.0094	
5	2.2673	0.2365	-2.6935	0.0372	0.0395	
6	-1.9148	-0.2208	0.8494	-0.1168	0.0152	
7	-1.7790	0.5159	1.2660	0.6340	0.0153	
8	-0.7359	-0.0861	-0.3174	-0.0752	0.0122	
9	0.3568	-0.2975	0.8277	-0.2724	0.0132	
10	0.9494	1.3925	-0.8495	1.3216	0.0107	

Figure 14.11. GPA Data Window with H_GPA Added

⇒ Drag a rectangle in the fit window to select an area for the new plot.

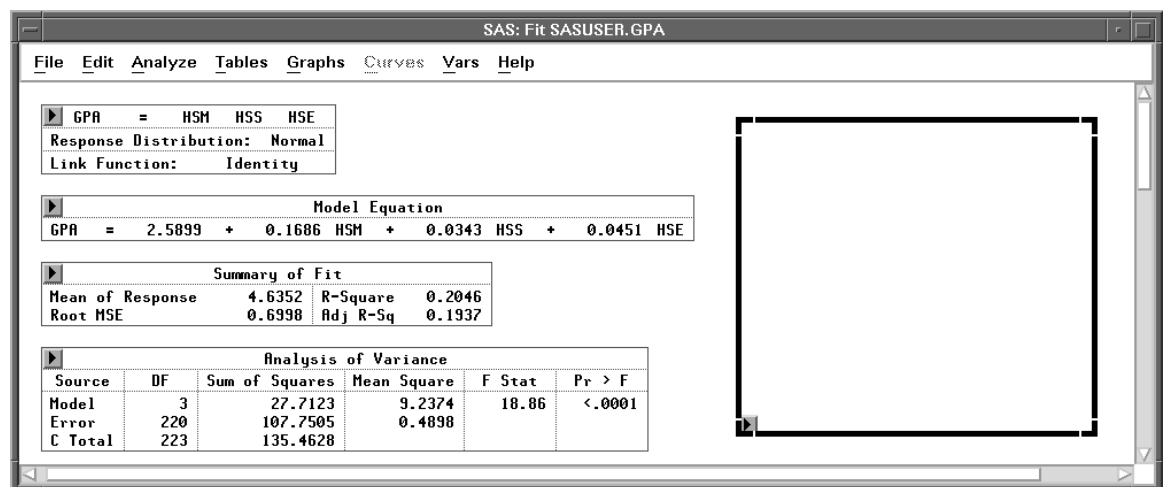


Figure 14.12. Selecting an Area

⇒ Choose **Analyze:Scatter Plot (Y X)**.

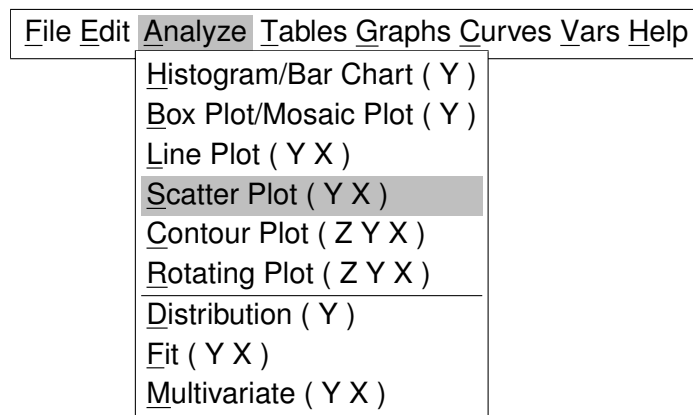


Figure 14.13. Analyze Menu

This displays the scatter plot variables dialog.

⇒ Assign **R_GPA** the **Y** role and **H_GPA** the **X** role, then click on **OK**.

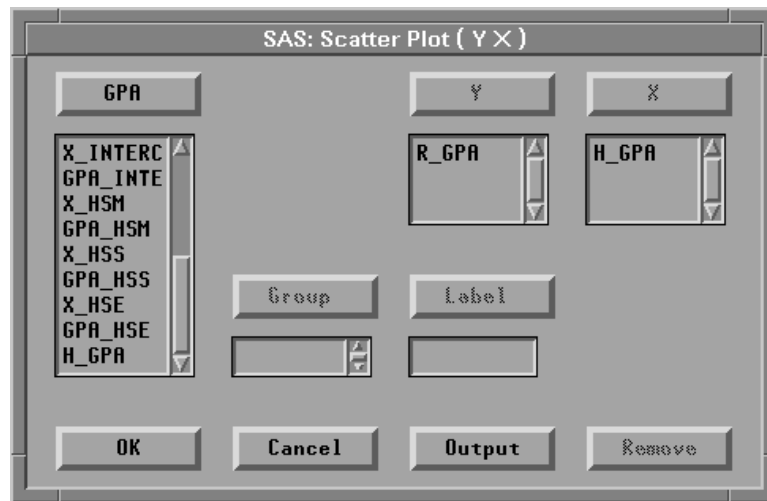


Figure 14.14. Scatter Plot Variables Dialog

The plot appears in the fit window in the area you selected.

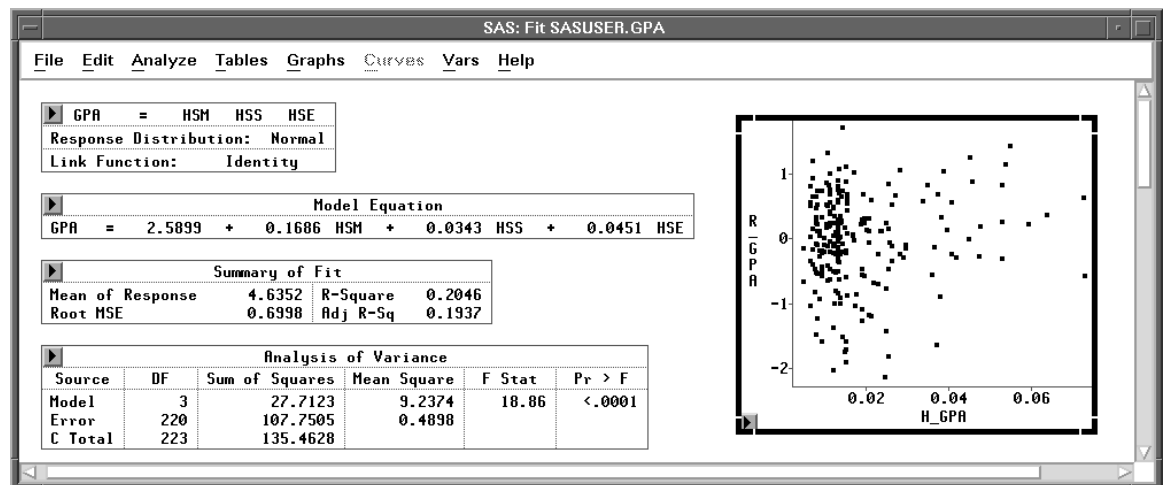


Figure 14.15. Residual by Hat Diagonal Plot

Belsley, Kuh, and Welsch (1980) propose a cutoff of $2p/n$ for the hat diagonal values, where n is the number of observations used to fit the model and p is the number of parameters in the model. Observations with values above this cutoff should be investigated. For this example, **H_GPA** values over 0.036 should be investigated. About 15% of the observations have values above this cutoff.

There are other measures you can use to determine the influence of observations. These include Cook's D, Dffits, Covratio, and Dfbetas. Each of these measures examines some effect of deleting the i th observation.

⇒ **Choose Vars:Dffits.**

A new variable, **F_GPA**, that contains the Dffits values is added to the data window.

Large absolute values of Dffits indicate influential observations. A general cutoff to consider is 2. It is, thus, useful in this example to identify those observations where **H_GPA** exceeds 0.036 and the absolute value of **F_GPA** is greater than 2. One way to accomplish this is by examining the **H_GPA** by **F_GPA** scatter plot.

⇒ **Choose Analyze:Scatter Plot (Y X).**

This displays the scatter plot variables dialog.

⇒ **Assign H_GPA the Y role and F_GPA the X role, then click on OK.**

This displays the **H_GPA** by **F_GPA** scatter plot.

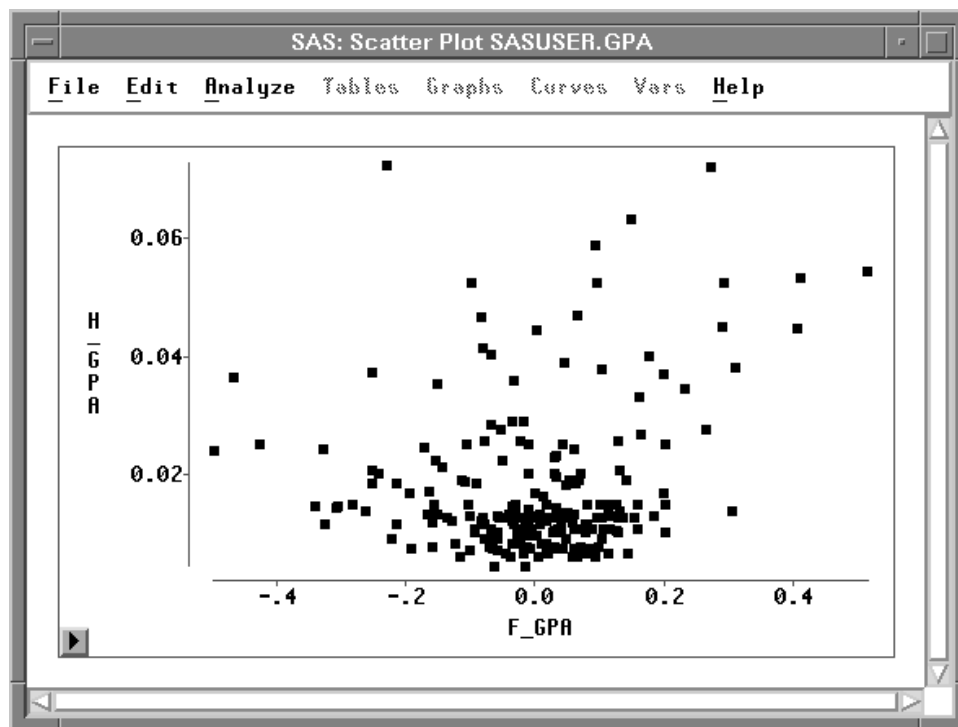


Figure 14.16. H_GPA by F_GPA Scatter Plot

None of the observations identified as potential influential observations (**H_GPA** > **0.036**) are, in fact, influential for this model using the criterion $|F_GPA| > 2$.

Modifying the Model

It may be possible to simplify the model without losing explanatory power. The change in the adjusted R-square value is one indicator of whether you are losing explanatory power by removing a variable. The estimate for **HSS** has the largest p -value, **0.3619**. Remove **HSS** from the model and see what effect this has on the adjusted R-square value.

From the fit variables dialog, follow these steps to request a new model with **HSS** removed. Remember, if you click **Apply** in the variables dialog, the dialog stays on the display so you can easily modify the regression model. You may need to rearrange the windows on your display if the fit variables dialog is not visible.

- ⇒ **Select HSS in the X variables list, then click the Remove button.**
This removes **HSS** from the model.



Figure 14.17. Removing the Variable **HSS**

- ⇒ **Click the Apply button.**
A new fit window appears, as shown in [Figure 14.18](#).

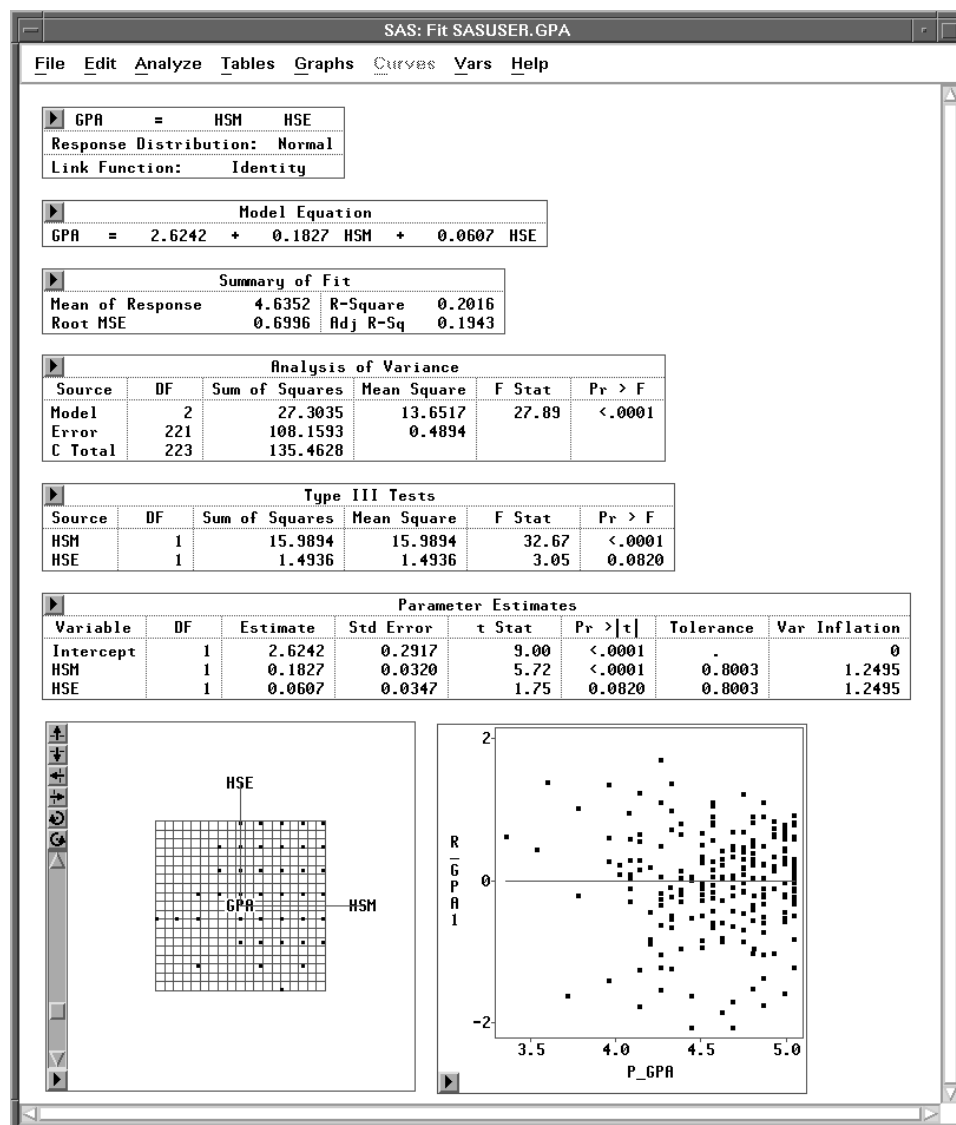


Figure 14.18. Fit Window with HSM and HSE as Explanatory Variables

Reposition the two fit windows so you can compare the two models. Notice that the adjusted R-square value has actually increased slightly from 0.1937 to 0.1943. Little explanatory power is lost by removing **HSS**. Notice that within this model the *p*-value for **HSE** is a modest 0.0820. You can remove **HSE** from the new fit window without creating a third fit window.

⇒ **Select HSE in the second fit window.**

⇒ **Choose Edit:Delete in the second fit window.**

This recomputes the second fit using only **HSM** as an explanatory variable.

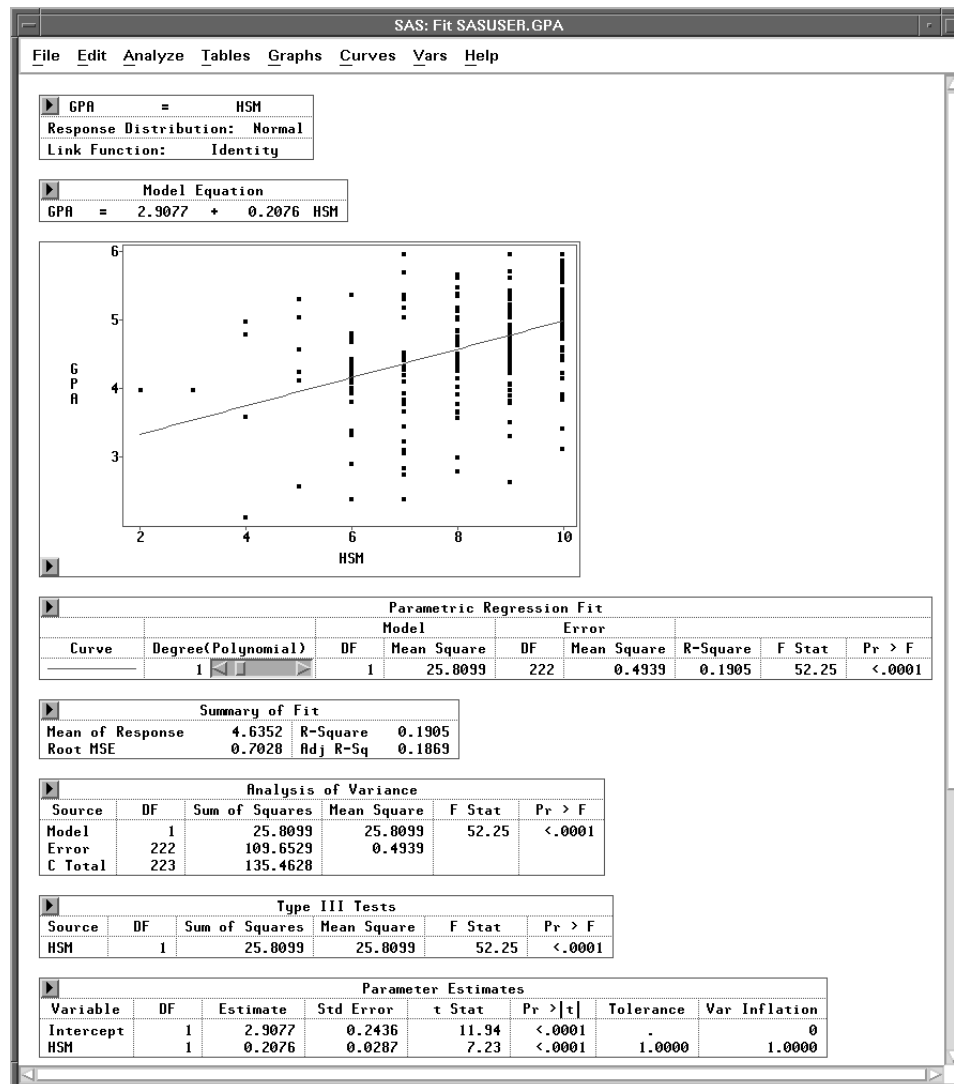


Figure 14.19. Fit Window with HSM as Explanatory Variable

The adjusted R-square value drops only slightly to **0.1869**. Removing **HSE** from the model also appears to have little effect. So, of the three explanatory variables you considered, only **HSM** appears to have strong explanatory power.

Saving the Residuals

One of the assumptions made in carrying out hypothesis tests in regression analysis is that the errors are normally distributed (Myers 1986). You can use residuals to check assumptions about errors. For this example, the *studentized* residuals are used because they are somewhat better than ordinary residuals for assessing normality, especially in the presence of outliers (Weisberg 1985). You can create a distribution window to check the normality of the residuals, as described in [Chapter 12](#), “Examining Distributions.”

⇒ **Choose Vars:Studentized Residual.**

A variable called **RT_GPA_1** is placed in the data window, as shown in [Figure 14.20](#).

SAS: SASUSER.GPA

File Edit Analyze Tables Graphs Curves Vars Help

22	Int	Int	Int	Int	Int					
224	GPA_HSE	H_GPA	R_GPA_1	P_GPA_1	RT_GPA_1					
1	0.2876	0.0132	0.3363	4.9837	0.4799					
2	0.3368	0.0119	0.3639	4.7761	0.5183					
3	-0.8024	0.0249	-0.9361	4.7761	-1.3378					
4	0.3612	0.0094	0.3563	4.9837	0.5085					
5	0.0372	0.0395	0.1067	4.1533	0.1525					
6	-0.1168	0.0152	-0.2185	4.5685	-0.3110					
7	0.6340	0.0153	0.5539	4.7761	0.7895					
8	-0.0752	0.0122	-0.1337	4.9837	-0.1907					
9	-0.2724	0.0132	-0.2237	4.9837	-0.3191					
10	1.3216	0.0107	1.3591	4.3609	1.9533					

Figure 14.20. GPA Data Window with RT_GPA_1 Added

Notice the names of the last three variables. The number you see at the end of the variable names corresponds to the number of the fit window that generated the variables. See [Chapter 39](#), “Fit Analyses,” for detailed information about how generated variables are named.

⊕ **Related Reading:** Linear Models, Residuals, [Chapter 39](#).

References

- Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics*, New York: John Wiley and Sons, Inc.
- Freedman, D., Pisani, R., and Purves, R. (1978), *Statistics*, New York: W.W. Norton & Company, Inc.
- Moore, D.S. and McCabe, G.P. (1989), *Introduction to the Practice of Statistics*, New York: W.H. Freeman and Company.
- Myers, R.H. (1986), *Classical and Modern Regression with Applications*, Boston, MA: Duxbury Press.
- Rawlings, J.O. (1988), *Applied Regression Analysis: A Research Tool*, Pacific Grove, CA: Wadsworth and Brooks/Cole Advanced Books and Software.
- Weisberg, S. (1985), *Applied Linear Regression, Second Edition*, New York: John Wiley and Sons, Inc.

Chapter 15

Analysis of Variance

Chapter Contents

ASSIGNING MEASUREMENT LEVELS	245
CREATING THE ANALYSIS OF VARIANCE	247
Model Information	250
Summary of Fit	251
Analysis of Variance	252
Type III Tests	252
Parameter Estimates	253
Residuals-by-Predicted Plot	254
EXAMINING THE MEANS	255
REFERENCES	259

Chapter 15

Analysis of Variance

In this chapter, you consider analyses that use least-squares methods to fit the general linear model. Such analyses include regression, analysis of variance, and analysis of covariance. You can choose **Analyze:Fit (Y X)** to carry out an analysis of variance.

You can use box plots to examine individual group means.

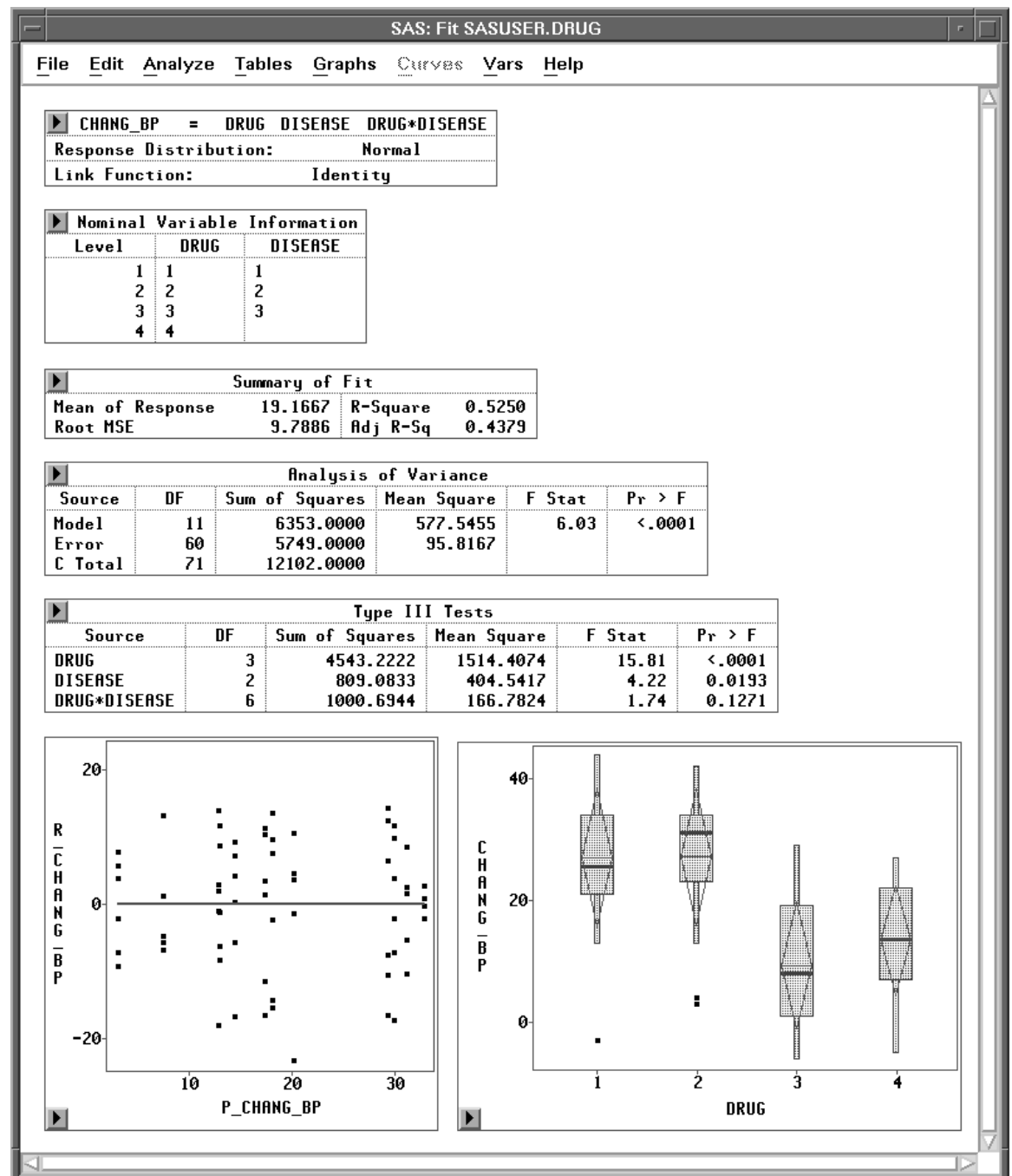


Figure 15.1. Fit Window with Analysis of Variance Results

Assigning Measurement Levels

The **DRUG** data set contains results of an experiment carried out to evaluate the effect of four drugs (**DRUG**) with three experimentally induced diseases (**DISEASE**). Each drug by disease combination was applied to six randomly selected dogs. The response variable is the increase in systolic blood pressure (**CHANG_BP**) due to the drug treatment. **DRUG** and **DISEASE** are *classification* or class variables; that is, variables that identify distinct levels or groups. **DRUG** contains four levels or classes and **DISEASE** contains three.

⇒ Open the **DRUG** data set.

	Int	Int	Int
	DRUG	DISEASE	CHANG_BP
1	1	1	42
2	1	1	44
3	1	1	36
4	1	1	13
5	1	1	19
6	1	1	22
7	1	2	33
8	1	2	40
9	1	2	26
10	1	2	34

Figure 15.2. Data Window

A variable's *measurement level* determines the way it is treated in analyses. In the data window, measurement levels appear above the variable names, in the upper right portion of the column header. SAS/INSIGHT software supports two measurement levels: interval (**Int**) and nominal (**Nom**).

Interval variables contain values that vary across a continuous range. In this data set, the change in blood pressure (**CHANG_BP**) is an interval variable.

Nominal variables contain a discrete set of values. In this data set, both **DRUG** and **DISEASE** contain a discrete set of values. However, since these are numeric variables, by default they have interval measurement levels (**Int**).

You need to assign both these variables the nominal measurement level (**Nom**) in order to treat them as classification variables. To do so, use the data measurement level pop-up menu.

- ⇒ Click on the **Int** measurement level indicator for the variable **DRUG**.
This displays a pop-up menu.

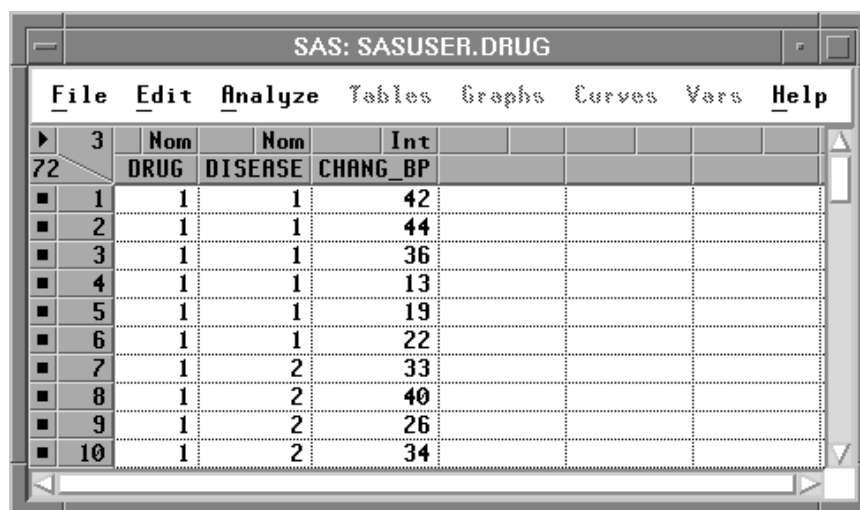
•Interval
Nominal

Figure 15.3. Measurement Levels Menu

The radio mark beside **Interval** shows the current measurement level. Because **DRUG** is a numeric variable, it can use either an interval or a nominal measurement level.

- ⇒ Choose **Nominal** in the pop-up menu to change **DRUG**'s measurement level.
⇒ Repeat these steps to change the measurement level for **DISEASE**.

Check the measurement levels for **DRUG** and **DISEASE** in the data window. Both have **Nom** measurement levels.



		Nom	Nom	Int					
		DRUG	DISEASE	CHANG_BP					
1	1	1	1	42					
2	1	1	1	44					
3	1	1	1	36					
4	1	1	1	13					
5	1	1	1	19					
6	1	1	1	22					
7	1	2	2	33					
8	1	2	2	40					
9	1	2	2	26					
10	1	2	2	34					

Figure 15.4. Data with Nominal Variables **DRUG** and **DISEASE**

Creating the Analysis of Variance

Consider the two-way analysis of variance model Kutner (1974) proposed for these data:

$$\text{CHANG_BP}_{ijk} = \mu + \gamma_i + \tau_j + (\gamma\tau)_{ij} + \epsilon_{ijk}$$

where μ is the overall mean effect, γ_i is the effect of the i th level of **DRUG**, τ_j is the effect of the j th level of **DISEASE**, $(\gamma\tau)_{ij}$ is the joint effect of the i th level of **DRUG** with the j th level of **DISEASE**, and ϵ_{ijk} is the random error term for the k th observation in the i th level of **DRUG** and j th level of **DISEASE**. The ϵ_{ijk} 's are assumed to be normally distributed and uncorrelated and to have mean 0 and common variance σ^2 .

The effects for **DRUG** and **DISEASE** are often referred to as the *main effects* in the model and the **DRUG*DISEASE** effect as an *interaction effect*. The interaction effect enables you to determine whether the level of **DRUG** affects the change in blood pressure differently for different levels of **DISEASE**.

To begin the analysis of variance, follow these steps.

- ⇒ **Choose Analyze:Fit (Y X).**
- ⇒ **Select CHANG_BP in the variables list on the left, then click the Y button.**
CHANG_BP appears in the **Y** variables list and is now defined as the response variable.
- ⇒ **Select DRUG and DISEASE, then click the Expand button.**
Your variables dialog should now appear, as shown in [Figure 15.5](#).



Figure 15.5. Fit Variables Dialog with Variable Roles Assigned

The **Expand** button provides a convenient way to specify interactions of any order. The degree of expansion is controlled by the value below the **Expand** button. The order **2** is the default, so clicking **Expand** constructs all possible effects from the selected variables up to second-order effects. This adds **DRUG**, **DISEASE**, and **DRUG*DISEASE** to the effects list.

† **Note:** You could have added the same effects by using the **X** and **Cross** buttons, but the **Expand** button is faster. There is also a **Nest** button for specifying nested effects. For more information on the effects buttons, see [Chapter 39, “Fit Analyses.”](#)

⇒ **Click the OK button.**

A fit window appears, as shown in [Figure 15.6](#).

You can control which tables and graphs the fit window contains by clicking the **Output** button in the fit variables dialog or by choosing from the **Tables** and **Graphs** menus. By default, the fit window contains tables for model specification, **Nominal Variable Information**, **Parameter Information**, **Model Equation**, **Summary of Fit**, **Analysis of Variance**, **Type III Tests**, and **Parameter Estimates**, as well as a residual-by-predicted plot.

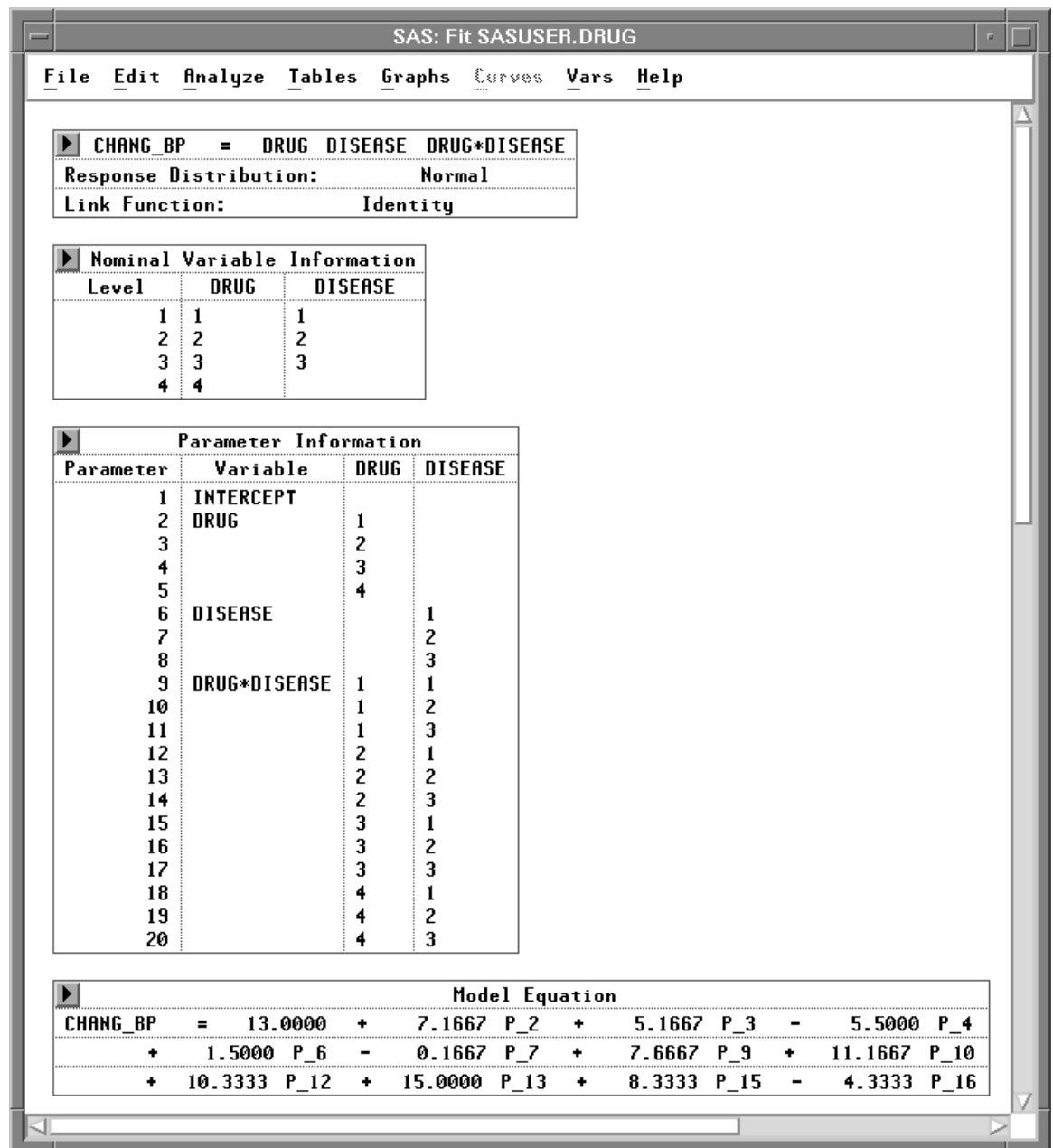


Figure 15.6. Fit Window - Model Information

Model Information

The first four tables in the fit analysis contain model information. The first table displays the model specification, the response distribution, and the link function. The **Nominal Variable Information** table shows the levels of the nominal variables. The levels are determined from the formatted values of the nominal variables.

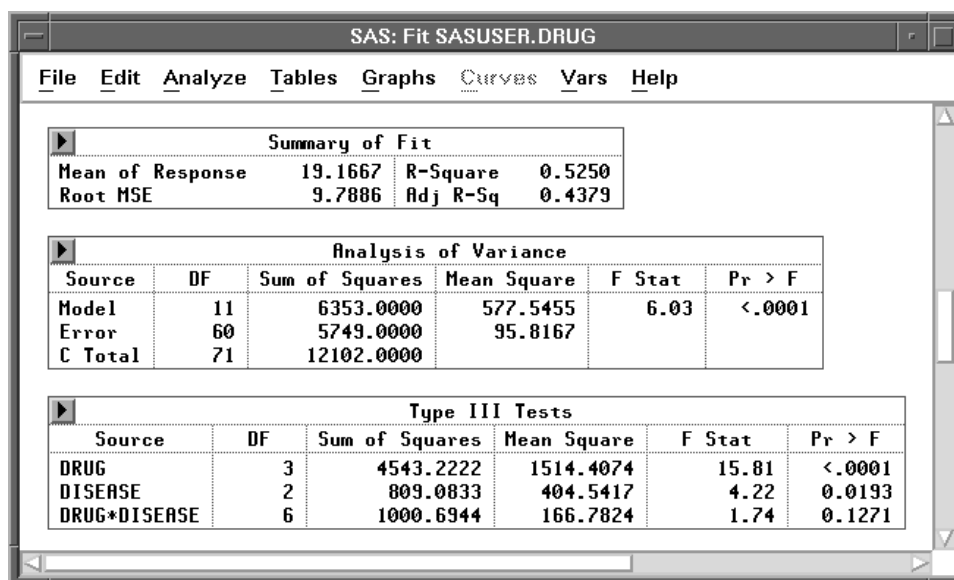
An additional **Parameter Information** table shows the variable indices for the parameters in the model equation, the X'X matrix, the estimated covariance matrix,

and the estimated correlation matrix. The **Model Equation** table gives the fitted equation for the model.

Summary of Fit

The **Summary of Fit** table, as shown in Figure 15.7, contains summary statistics. The **Mean of Response 19.1667** is the overall mean of **CHANG_BP**. The **Root MSE 9.7886** is the square root of the mean square error given in the **Analysis of Variance** table. **Root MSE** is an estimate of σ in the preceding analysis of variance model.

The **R-Square** value is **0.5250**, which means that 52% of the variation in **CHANG_BP** is explained by the fitted model. **Adj R-Sq** is an alternative to **R-Square**, adjusted for the number of parameters in the model.



The screenshot shows the SAS Fit window titled "SAS: Fit SASUSER.DRUG". It contains three tables: "Summary of Fit", "Analysis of Variance", and "Type III Tests".

Summary of Fit			
Mean of Response	19.1667	R-Square	0.5250
Root MSE	9.7886	Adj R-Sq	0.4379

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	11	6353.0000	577.5455	6.03	<.0001
Error	60	5749.0000	95.8167		
C Total	71	12102.0000			

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
DRUG	3	4543.2222	1514.4074	15.81	<.0001
DISEASE	2	809.0833	404.5417	4.22	0.0193
DRUG*DISEASE	6	1000.6944	166.7824	1.74	0.1271

Figure 15.7. Fit Window - Summary of Fit

Analysis of Variance

The **Analysis of Variance** table summarizes the information related to the sources of variation in the data. **Sum of Squares** measures variation present in the data. It is calculated by summing squared deviations. There are three sources of variation: **Model**, **Error**, and **C Total**. The **Model** row in the table corresponds to the variation *among* class means. The **Error** row corresponds to ϵ in the model and represents variation *within* class means. **C Total** is the total sum of squares corrected for the mean, and it is the sum of **Model** and **Error**. Degrees of Freedom, **DF**, are associated with each sum of squares and are related in the same way. **Mean Square** is the **Sum of Squares** divided by its associated **DF** (Moore and McCabe 1989, p.685).

If the data are normally distributed, the ratio of the **Mean Square** for the **Model** to the **Mean Square** for **Error** is an *F statistic*. This *F* statistic tests the null hypothesis that all the class means are the same against the alternative hypothesis that the means are not all equal. Think of the ratio as a comparison of the variation *among* class means to variation *within* class means. The larger the ratio, the more evidence that the means are not the same. The computed *F* statistic (labeled **F Stat**) is **6.0276**. You can use the *p*-value (labeled **Pr > F**) to determine whether to reject the null hypothesis. The *p*-value, also referred to as the *probability value* or *observed significance level*, is the probability of obtaining (by chance alone) an *F* statistic greater than the computed *F* statistic when the null hypothesis is true. The smaller the *p*-value, the stronger the evidence against the null hypothesis.

In this example, the *p*-value is so small that you can clearly reject the null hypothesis and conclude that at least one of the class means is different. At this point, you have demonstrated statistical significance but cannot make statements about which class means are different.

Type III Tests

The **Type III Tests** table is a further breakdown of the variation due to **MODEL**. The **Sum of Squares** and **DF** for **Model** are broken down into terms corresponding to the main effect for **DRUG**, the main effect for **DISEASE**, and the interaction effect for **DRUG*DISEASE**. The sum of squares for each term represents the variation among the means for the different levels of the factors.

The **Type III Tests** table presents the Type III sums of squares associated with the effects in the model. The Type III sum of squares for a particular effect is the amount of variation in the response due to that effect after correcting for all other terms in the model. Type III sums of squares, therefore, do not depend on the order in which the effects are specified in the model. Refer to the chapter on “The Four Types of Estimable Functions,” in the *SAS/STAT User’s Guide* for a complete discussion of Type I–IV sums of squares.

F tests are formed from this table in the same fashion that was explained previously in the section “Analysis of Variance.” In this case, there are three null hypotheses being tested: class means are all the same for the main effect **DRUG**, the main effect **DISEASE**, and the interaction effect **DRUG*DISEASE**. Begin by examining the test for the interaction effect since a strong interaction makes the interpretation of main effects difficult if not impossible. The computed F statistic is **1.7406** with a p -value of **0.1271**. This gives little evidence for an interaction effect. Now examine the main effects. The computed F statistic for **DRUG** is **15.8053** with a p -value less than or equal to 0.0001. The computed F statistic for **DISEASE** is **4.2220** with a p -value of 0.0193. While both effects are significant, the **DRUG** effect appears to be stronger.

Now you have more information about which means are significantly different. The results of the F test in the **Analysis of Variance** table indicated only that *at least one* of the class means is different from the others. Now you know that the difference in means can be associated with the different levels of the main effects, **DRUG** and **DISEASE**.

Parameter Estimates

Parameter estimates resulting from analysis of variance models where the effects are all classification variables are different from those observed in a regression model. They represent a non-unique solution to the normal equations, and thus the individual elements in the table are not as easily interpretable as they are in multiple regression. For a complete discussion of parameter estimates involving classification variables, refer to the chapter “Details of the Linear Model: Understanding GLM Concepts,” in *SAS System for Linear Models, Third Edition*.

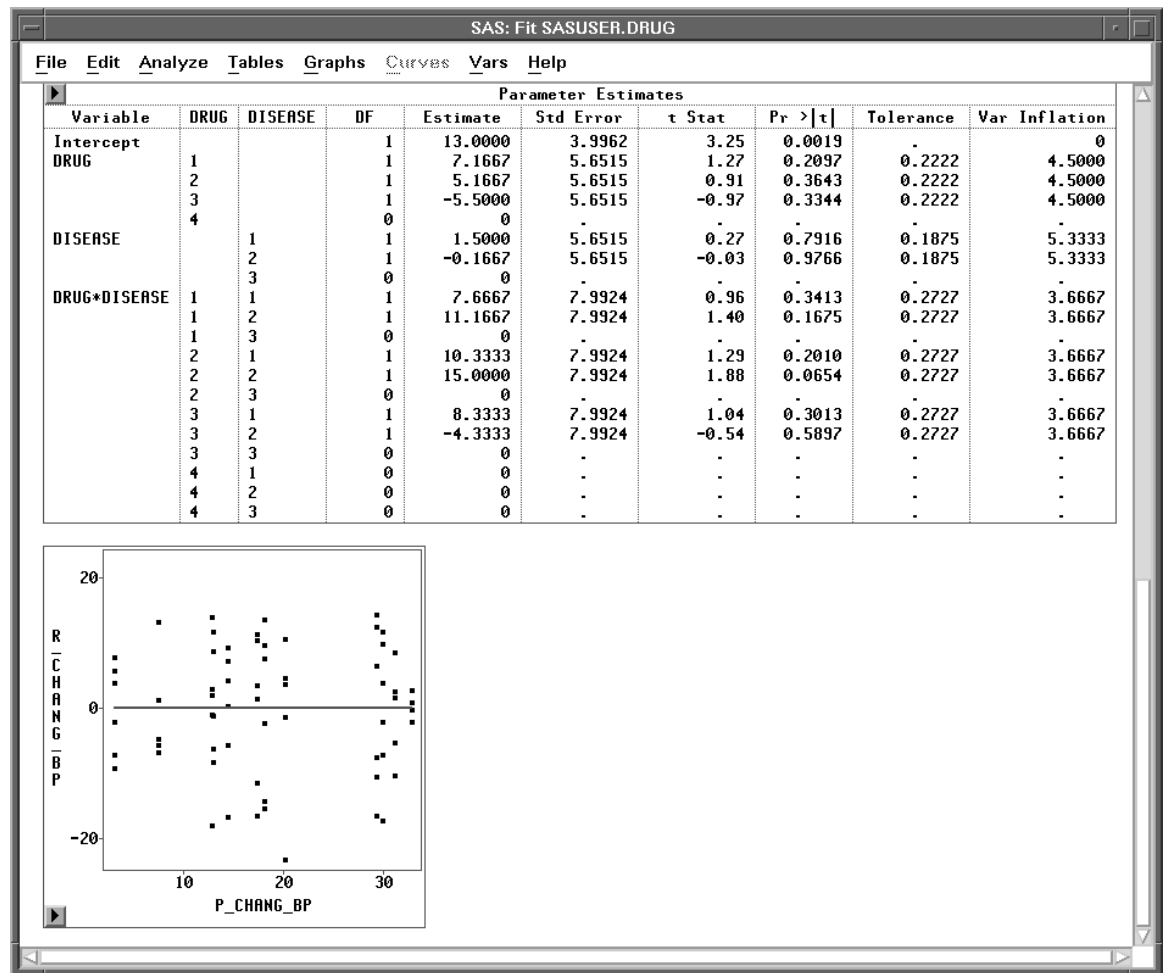


Figure 15.8. Fit Window - Parameter Estimates

Residuals-by-Predicted Plot

It is appropriate to examine the residuals from the fitted model for analysis of variance just as you did with the multiple regression model you fit in [Chapter 14, “Multiple Regression.”](#) The residuals-by-predicted graph illustrated in [Figure 15.8](#), along with several other diagnostic plots, are available for examining residuals. Since this topic is discussed in [Chapter 14](#), residual plots are not examined here.

Examining the Means

Before you can interpret the results for the significant main effects you observed in the **Type III Tests** table, you need to examine the means for the different levels of these effects. Box plots are an excellent tool for displaying means because means and standard deviations for each level of a variable can be placed side-by-side for easy comparison.

Follow these steps to add box plots for each level of **DRUG** to the **Fit(Y X)** window.

⇒ **Select an area for the box plot.**

Drag the cursor until you have a rectangle of suitable size.

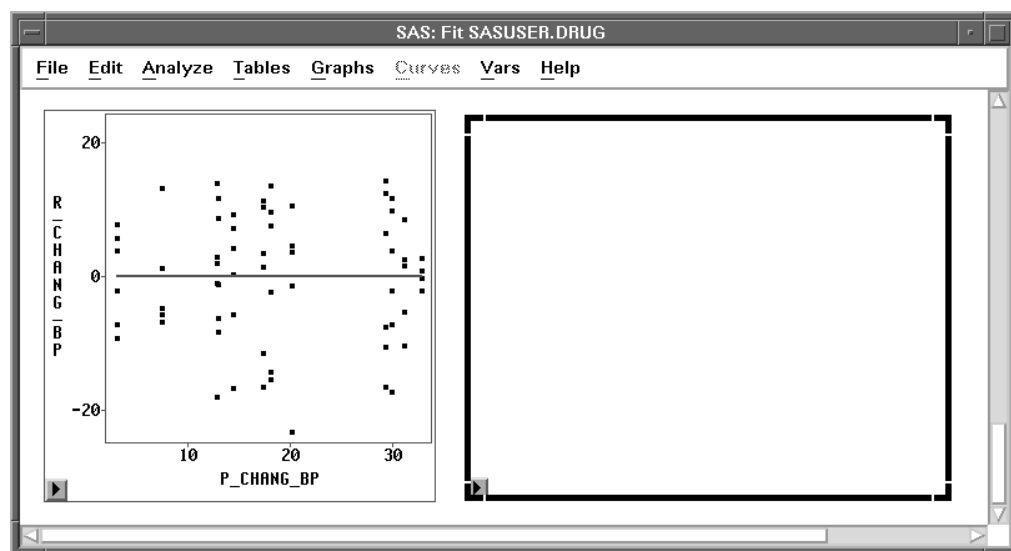


Figure 15.9. Selecting an Area

⇒ **Choose Analyze:Box Plot/Mosaic Plot (Y).**

⇒ **Select CHANG_BP in the list at the left, then click the Y button.**

This assigns the **Y** role to this variable.

⇒ **Select DRUG in the list at the left, then click the X button.**

This assigns the **X** role to this variable and requests a separate box plot for each level of **DRUG**. Your variables dialog should now appear, as shown in [Figure 15.10](#).

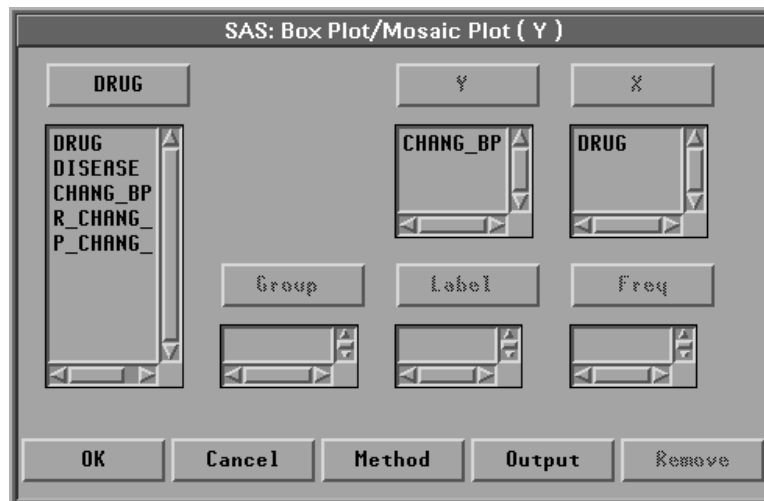


Figure 15.10. Box Plot Variables Dialog with Variable Roles Assigned

⇒ **Click the Output button.**

The output options dialog shown in [Figure 15.11](#) appears on your display. In this dialog, you can specify options to determine the output produced by the box plot.

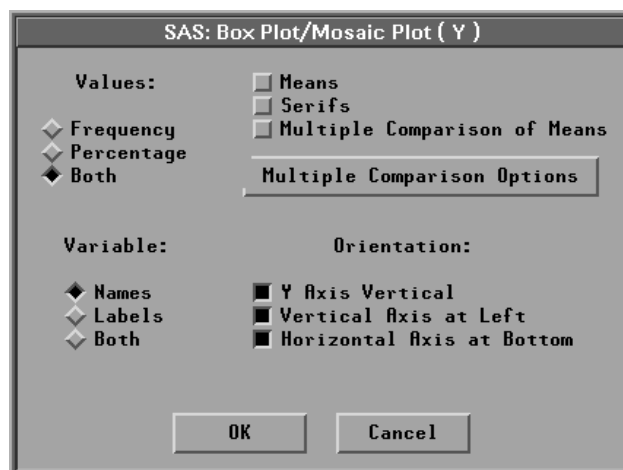


Figure 15.11. Box Plot Output Options Dialog

⇒ **Click on Means.**

Means displays mean diamonds for all boxes. The central line in the mean diamond marks the mean; the size of the mean diamond is two standard deviations, one above and one below the mean.

⇒ **Click OK in both dialogs to create the Box Plots.**

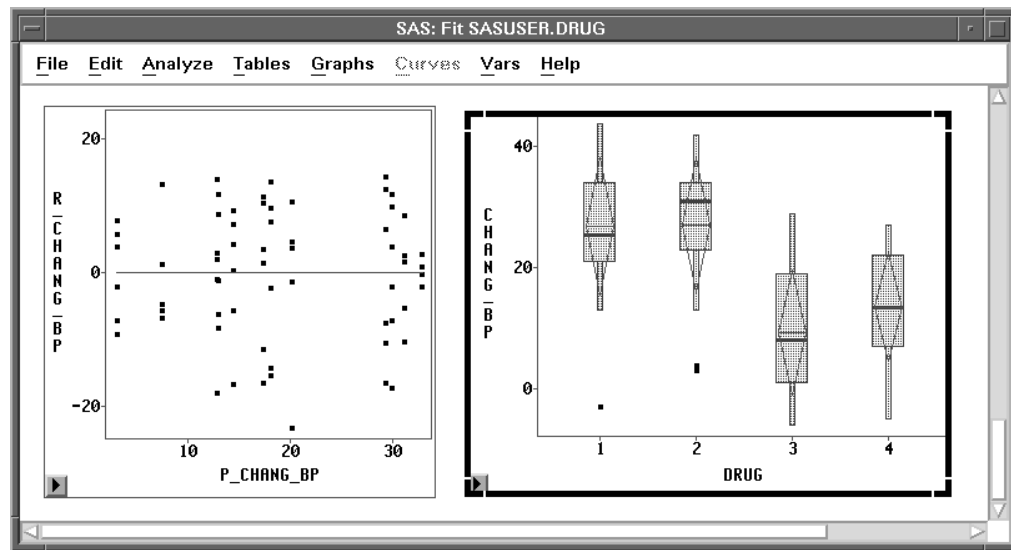


Figure 15.12. Box Plots for different **DRUG** Levels

Examine the box plot representing the four levels of **DRUG**. Recall that the central line in each mean diamond marks the mean while the height of the mean diamond shows one standard deviation on either side of the mean. The box and whiskers display percentiles for the data. (See [Chapter 4, “Exploring Data in One Dimension,”](#) for a complete description of the parts of the box plot.)

Follow these steps to hide the display of box and whiskers in order to display the means and standard deviations better.

⇒ **Click on Observations in the box plot pop-up menu.**

This toggles the display of observations and thus turns off the display of the box, whiskers, and individual observations in the box plot.

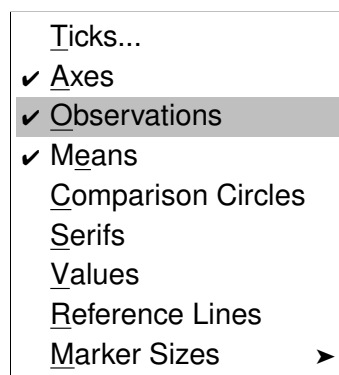


Figure 15.13. Box Plot Pop-up Menu

⇒ **Click on Values in the box plot pop-up menu.**

This toggles the display of values of the mean for each box plot.

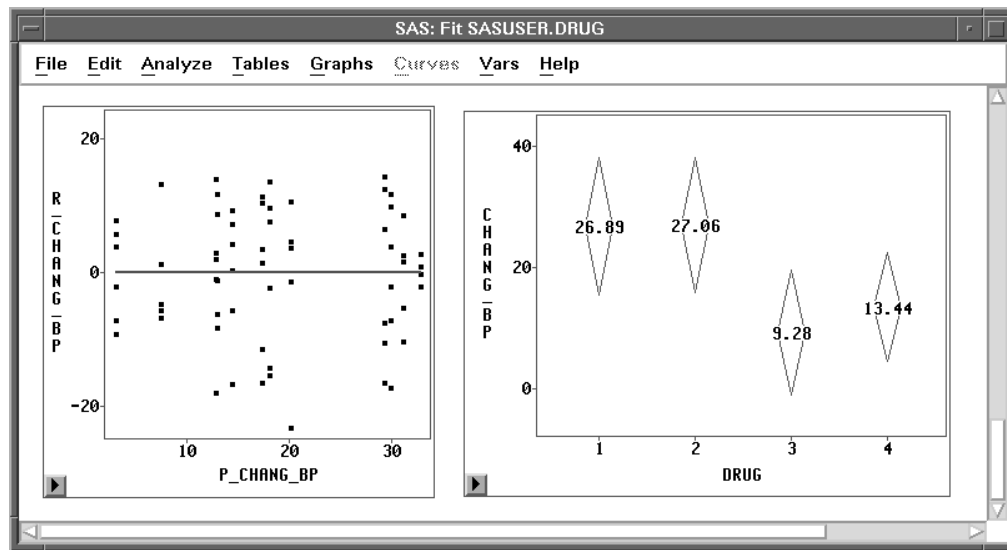


Figure 15.14. Mean Diamonds for **DRUG**

The largest effect noted in these plots is that drugs **1** and **2** have a higher average increase in systolic blood pressure than drugs **3** and **4** (averaged over all three levels of **DISEASE**). This difference resulted in the significant main effect for **DRUG** that was observed in the **Type III Tests** table.

⇒ Repeat the preceding steps and display box plots for the levels of **DISEASE**.

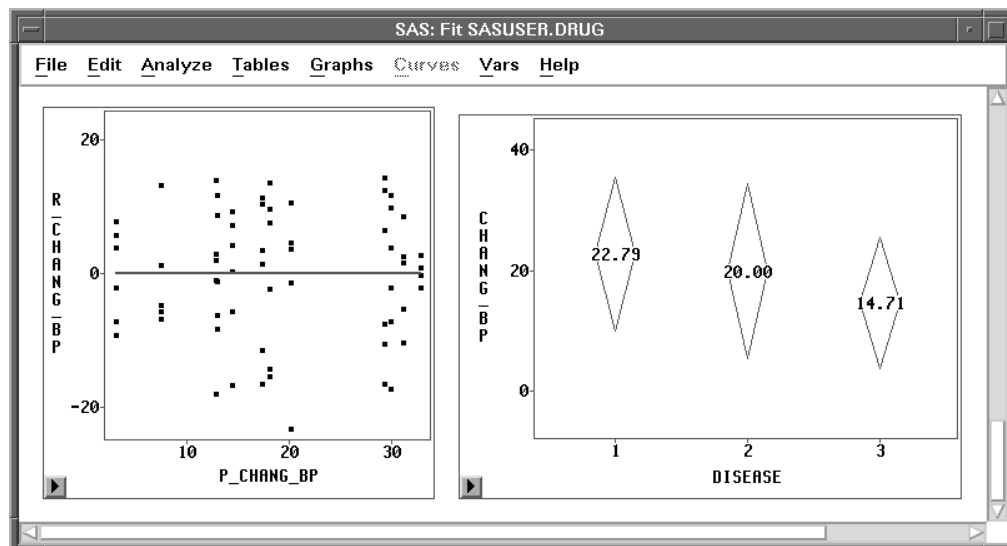


Figure 15.15. Mean Diamonds for **DISEASE**

The differences between the three **DISEASE** levels are not as pronounced as those observed for **DRUG**. Disease **3** is associated with a lower average increase in systolic blood pressure than the other two diseases (averaged over all four levels of **DRUG**).

The smaller p -value observed for the **DRUG** main effect is more evidence that the mean differences for **DISEASE** are not as pronounced as those for **DRUG**.

This example illustrates one way to use **Analyze:Fit** to fit the general linear model. Turn to the next chapter to see how to fit the generalized linear model.

⊕ **Related Reading:** Box Plots, [Chapter 33](#).

⊕ **Related Reading:** Linear Models, [Chapter 39](#).

References

Kutner, M.H. (1974), “Hypothesis Testing in Linear Models (Eisenhart Model I),” *The American Statistician*, 28 (3), 98.

Moore, D.S. and McCabe, G.P. (1989), *Introduction to the Practice of Statistics*, New York: W.H. Freeman and Company.

Chapter 16

Logistic Regression

Chapter Contents

DISPLAYING THE LOGISTIC REGRESSION ANALYSIS	265
Model Equation	269
Summary of Fit	270
Analysis of Deviance	270
Type III (Wald) Tests	270
Parameter Estimates Table	270
Residuals-by-Predicted Plot	270
MODIFYING THE MODEL	271
REFERENCES	275

Chapter 16

Logistic Regression

In the last two chapters, you used least-squares methods to fit linear models. In this chapter, you use maximum-likelihood methods to fit *generalized* linear models. You can choose **Analyze:Fit (Y X)** to carry out a logistic regression analysis. You can use the fit variables and method dialogs to specify generalized linear models and to add and delete variables from the model.

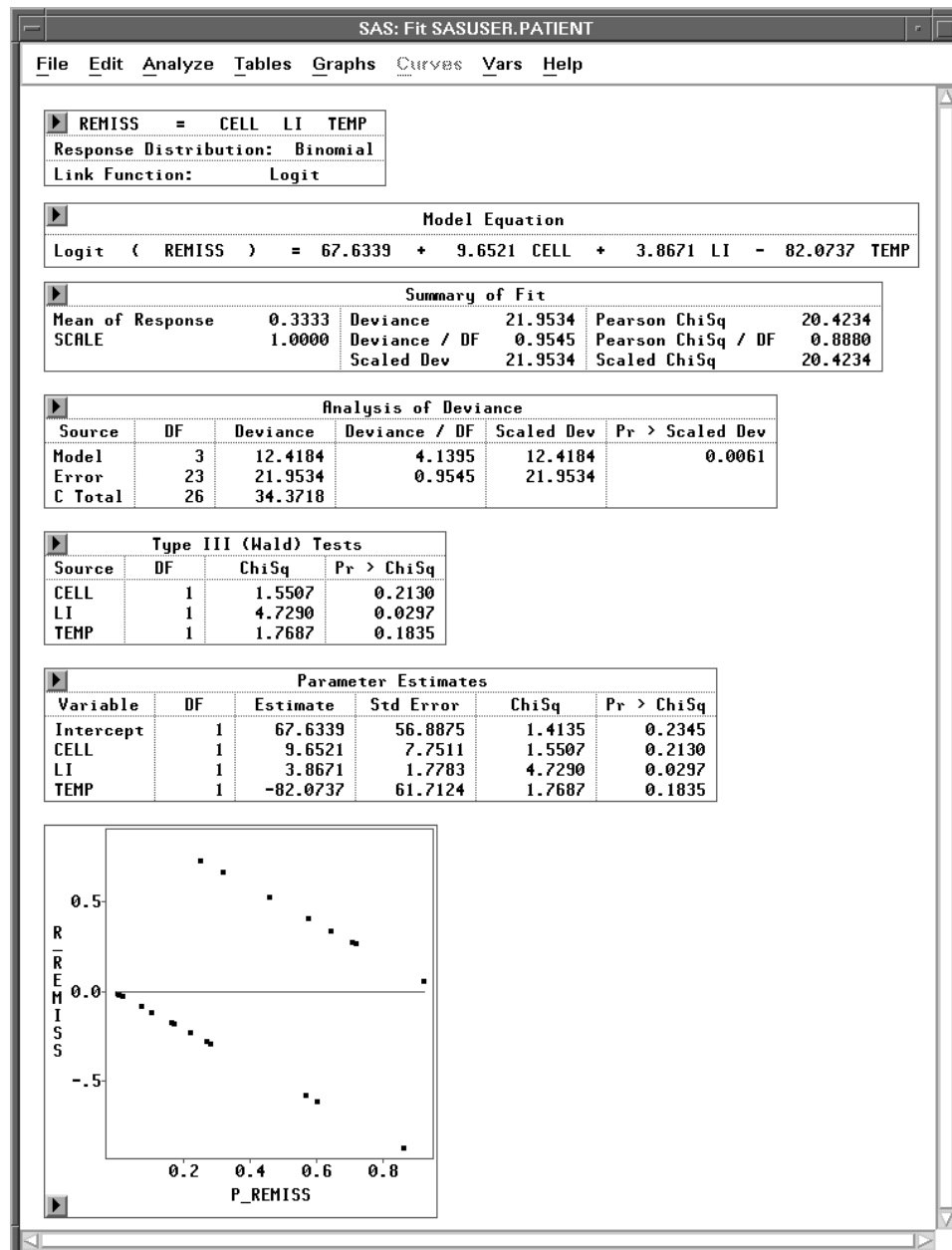


Figure 16.1. Logistic Regression Analysis

Displaying the Logistic Regression Analysis

The **PATIENT** data set, described by Lee (1974), contains data collected on 27 cancer patients. The response variable, **REMISS**, is binary and indicates whether cancer remission occurred:

REMISS = 1 indicates success (remission occurred)

REMISS = 0 indicates failure (remission did not occur)

Several other variables containing patient characteristics thought to affect cancer remission were also included in the study. For this example, consider the following three explanatory variables: **CELL**, **LI**, and **TEMP**. (You may want to carry out a more complete analysis on your own.)

⇒ Open the **PATIENT** data set.

	Int	Int	Int	Int	Int	Int	Int
	REMISS	CELL	SMEAR	INFIL	LI	BLAST	TEMP
1	1	0.80	0.83	0.66	1.9	1.100	0.996
2	1	0.90	0.36	0.32	1.4	0.740	0.992
3	0	0.80	0.88	0.70	0.8	0.176	0.982
4	0	1.00	0.87	0.87	0.7	1.053	0.986
5	1	0.90	0.75	0.68	1.3	0.519	0.980
6	0	1.00	0.65	0.65	0.6	0.519	0.982
7	1	0.95	0.97	0.92	1.0	1.230	0.992
8	0	0.95	0.87	0.83	1.9	1.354	1.020
9	0	1.00	0.45	0.45	0.8	0.322	0.999
10	0	0.95	0.36	0.34	0.5	0.000	1.038

Figure 16.2. Data Window

The generalized linear model has three components:

- a linear predictor function constructed from explanatory variables. For this example, the function is

$$\theta_i = \beta_0 + \beta_1 \text{CELL}_i + \beta_2 \text{LI}_i + \beta_3 \text{TEMP}_i$$

where $\beta_0, \beta_1, \beta_2$ and β_3 are coefficients (parameters) for the linear predictor, and $\text{CELL}_i, \text{LI}_i$, and TEMP_i are the values of the explanatory variables.

- a distribution or probability function for the response variable that depends on the mean μ and sometimes other parameters as well. For this example, the probability function is binomial.

Techniques ♦ Logistic Regression

- a link function, $g(\cdot)$, that relates the mean to the linear predictor function. For logistic regression, the link function is the logit

$$g(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \theta_i$$

where $p_i = \Pr(\text{REMISS}=1 \mid x_i)$ is the response probability to be modeled, and x_i is the set of explanatory variables for the i th patient.

You can specify these three components to fit a generalized linear model by following these steps.

- ⇒ Choose **Analyze:Fit (Y X)** to display the fit variables dialog.
- ⇒ Select **REMISS** in the list at the left, then click the **Y** button.
- ⇒ Select **CELL**, **LI**, and **TEMP** in the variables list, then click the **X** button.

Your variables dialog should now appear, as shown in [Figure 16.3](#).

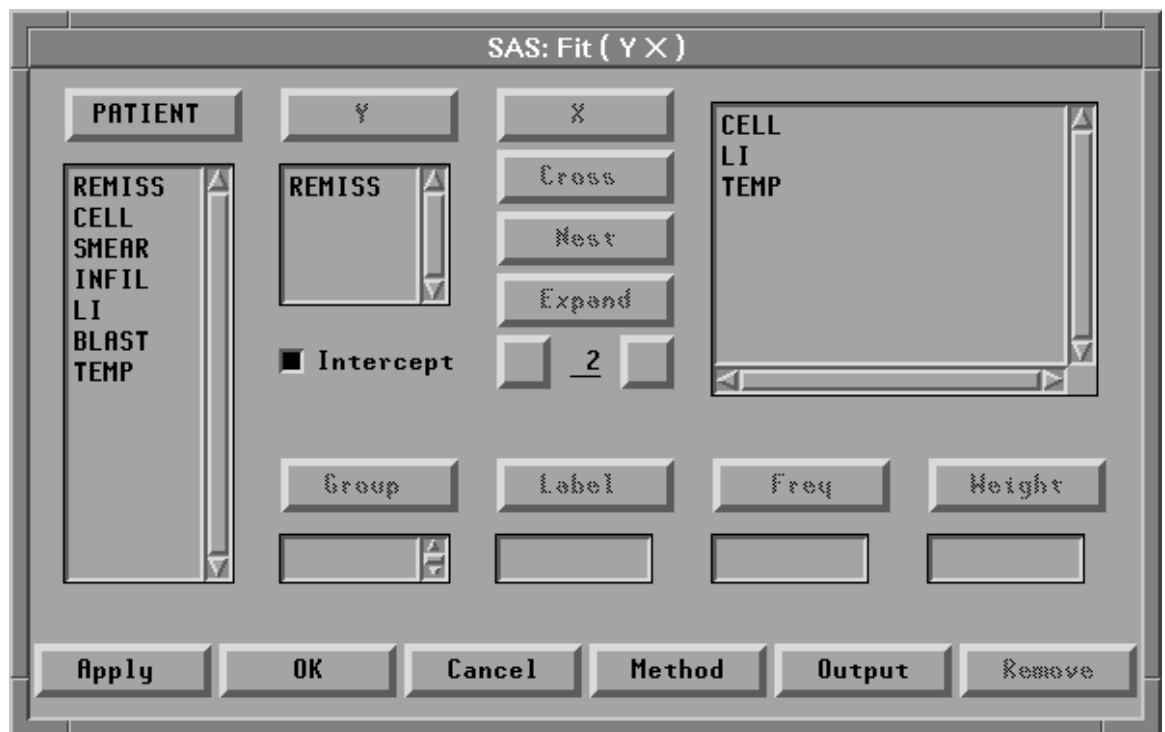


Figure 16.3. Fit Variables Dialog with Variable Roles Assigned

To specify the probability distribution for the response variable and the link function, follow these steps.

- ⇒ Click the **Method** button in the variables dialog to display the method dialog.



Figure 16.4. Fit Method Dialog

- ⇒ Click on **Binomial** under **Response Dist** to specify the probability distribution.

You do not need to specify a **Link Function** for this example. **Canonical**, the default, allows **Fit (Y X)** to choose a link dependent on the probability distribution. For the binomial distribution, as in this example, it is equivalent to choosing **Logit**, which yields a logistic regression.

- ⇒ Click the **OK** button to close the method dialog.

- ⇒ Click the **Apply** button in the variables dialog.

This creates the analysis shown in [Figure 16.5](#). Recall that the **Apply** button causes the variables dialog to stay on the screen after the fit window appears. This is convenient for adding and deleting variables from the model.

By default, the fit window displays tables for model information, **Model Equation**, **Summary of Fit**, **Analysis of Deviance**, **Type III (Wald) Tests**, and **Parameter Estimates**, and a residual-by-predicted plot. You can control the tables and graphs displayed by clicking on the **Output** button in the fit variables dialog or by choosing from the **Tables** and **Graphs** menus.

The first table displays the model information. The first line gives the model specification. The second and third lines give the error distribution and the link function you specified in the Method dialog.

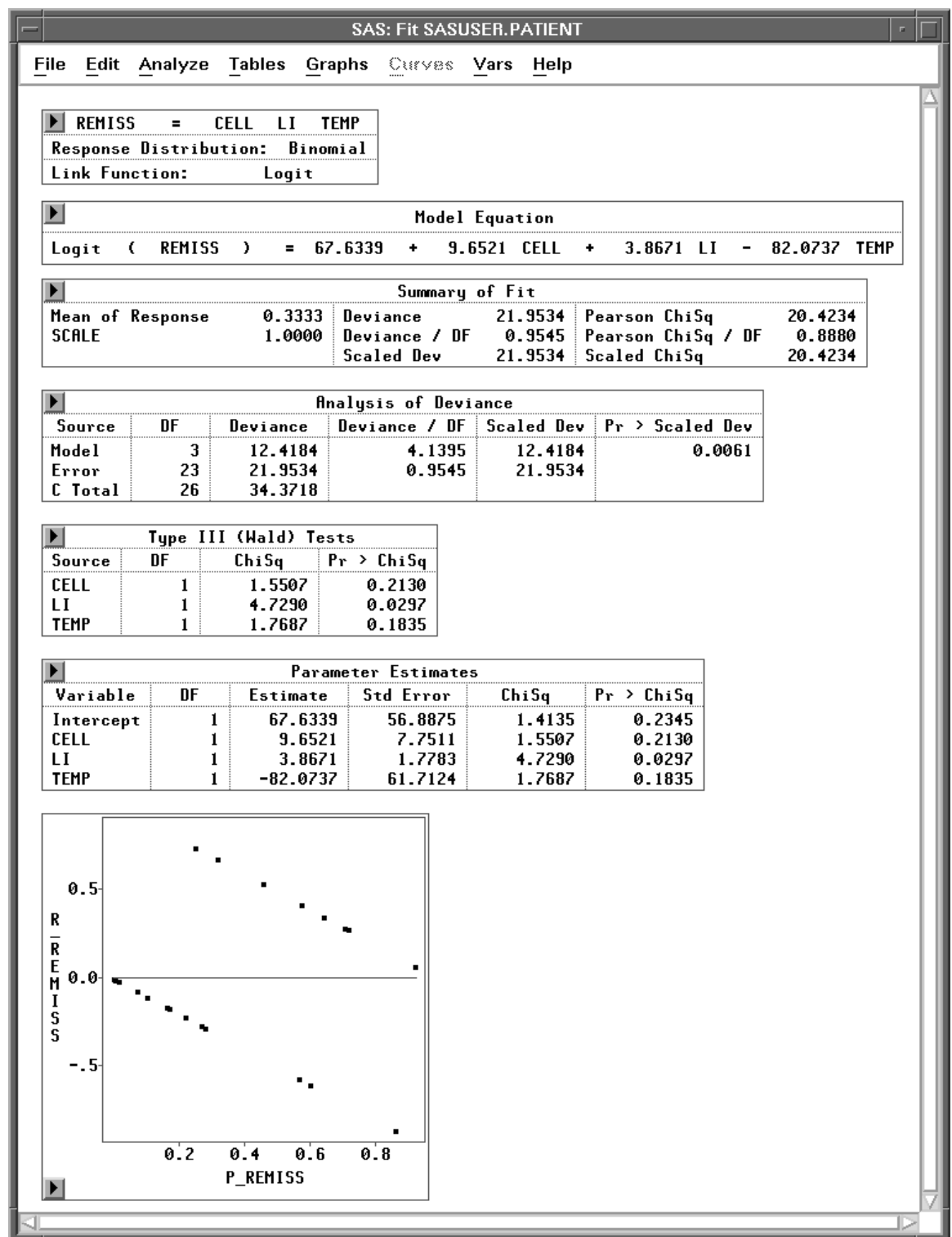


Figure 16.5. Fit Window

Model Equation

The **Model Equation** table writes out the fitted model using the estimated regression coefficients:

$$\begin{aligned} &\text{logit}(\text{Prob}(\text{REMISS} = 1)) \\ &= 67.6399 + 9.6521 * \text{CELL} + 3.8671 * \text{LI} - 82.0737 * \text{TEMP} \end{aligned}$$

Summary of Fit

The **Summary of Fit** table contains summary statistics for the fit of the model including values for **Deviance** and **Pearson's Chi-Squared** statistics. These values contrast the fit of your model to that of a saturated model that allows a different fit for each observation. If the data are sparse in the sense that most observations have a different set of explanatory variables, as in this set of data, then the quality of these measures is likely to be poor. Inferences drawn from these measures should be treated cautiously.

Analysis of Deviance

The **Analysis of Deviance** table summarizes information about the variation in the response for the set of data. Some of the variation can be explained by the **Model**. The **Error** is the remainder that is not systematically explained. **C Total** (the total corrected or adjusted for the mean) is the sum of **Model** and **Error**. The probability values give a measure of whether the amount of variation is consistent with chance alone or whether there is evidence of additional variation. In this case the **Deviance** associated with the **Model** shows a significant effect for the model, ($p = 0.0061$).

Type III (Wald) Tests

Wald tests are Chi-square statistics that test the null hypothesis that a parameter is 0; in other words, that the corresponding variable has no effect given that the other variables are in the model. These are approximate tests that are more accurate with larger sample sizes. In this example, only the coefficient for **LI** is statistically significant ($p = 0.0297$).

Parameter Estimates Table

The **Parameter Estimates** table shows the estimate, standard error, Chi-square statistic and associated degrees of freedom, and p -value for each of the parameters estimated.

Residuals-by-Predicted Plot

In the diagnostic plot of residuals versus predicted values, you can examine residuals for the model. You can point and click to identify individual observations. Because the observed response must either be 0 or 1, the plot of the residuals versus predicted values must lie along two straight lines. Plots of residuals versus the independent variables and other possible explanatory variables may be more useful. You can create scatter plots by selecting the response and explanatory variables in the data window and choosing **Analyze:Scatter Plot (Y X)**.

Modifying the Model

Plots of the residuals against other variables may suggest extensions of the model. Alternatively you may be able to remove some variables and thus simplify the model without losing explanatory power. The **Type III (Wald) Tests** table or the possibly more accurate **Type III (LR) Tests** table contains statistics that can help you decide whether to remove an effect. If the p -value associated with the test is large, then there is little evidence for any explanatory value of the corresponding variable.

⇒ Choose **Tables:Type III (LR) Tests**.

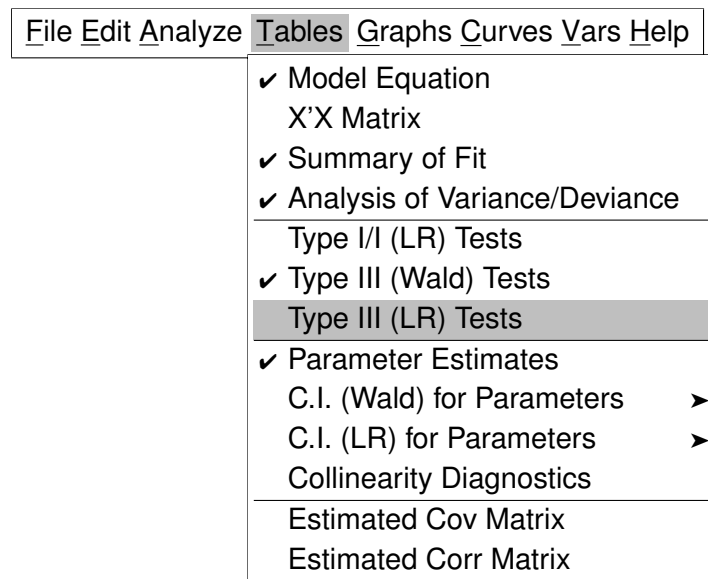


Figure 16.6. Tables Menu

This displays the table shown in [Figure 16.7](#).

Type III (LR) Tests			
Source	DF	ChiSq	Pr > ChiSq
CELL	1	2.6945	0.1007
LI	1	8.8752	0.0029
TEMP	1	2.3874	0.1223

Figure 16.7. Likelihood Ratio Type III Tests

Techniques ♦ *Logistic Regression*

The p -values for **TEMP** and **CELL** are relatively large, suggesting these effects could be removed. Although the numbers are different, the same conclusions would be reached from the corresponding **Wald** tests. In the Fit Variables dialog, follow these steps to request a new model with **TEMP** removed.

- ⇒ **Select TEMP in the effects list, then click the Remove button.**
TEMP disappears from the effects list.
- ⇒ **Click on Apply, and a new fit window appears, as shown in [Figure 16.8](#).**

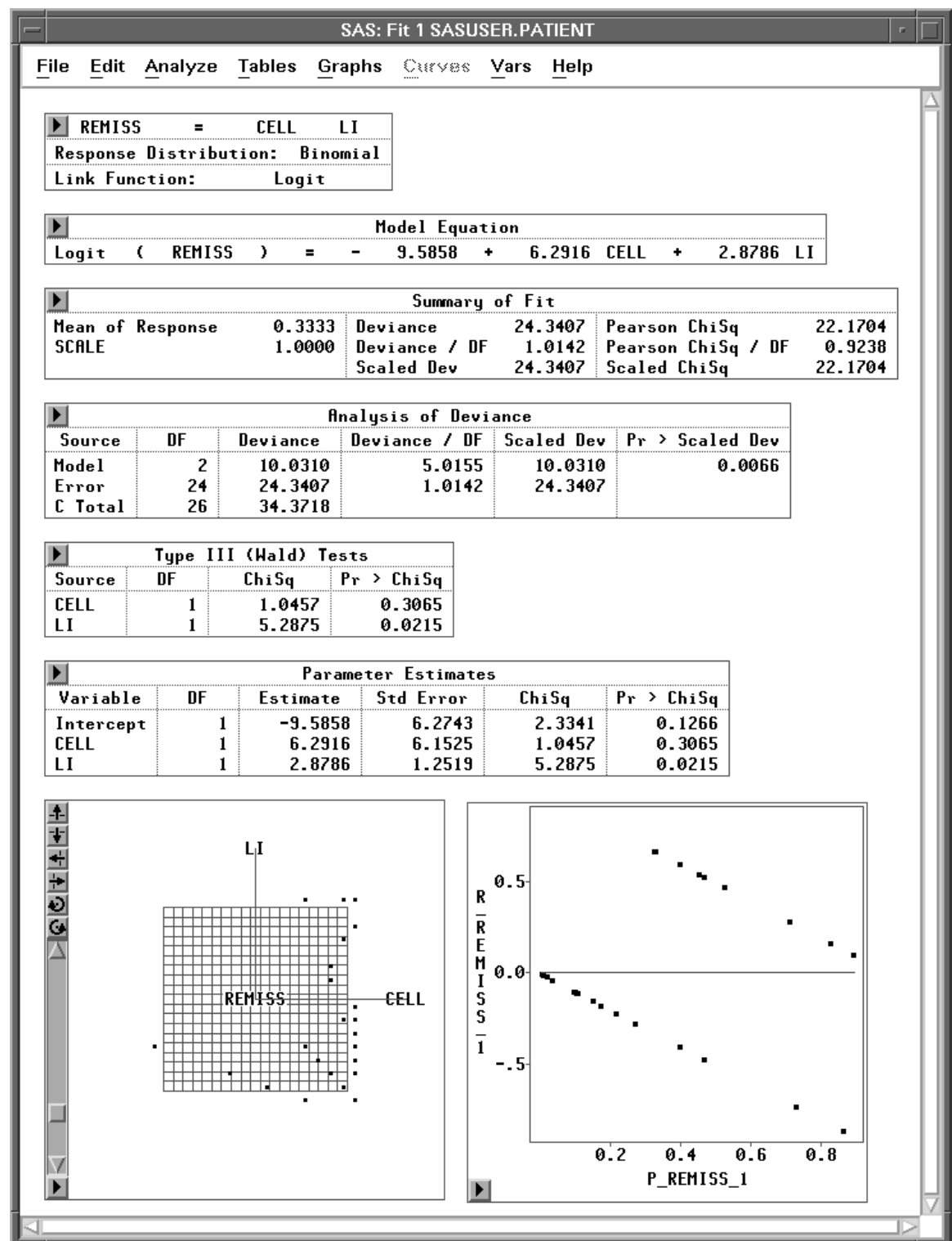
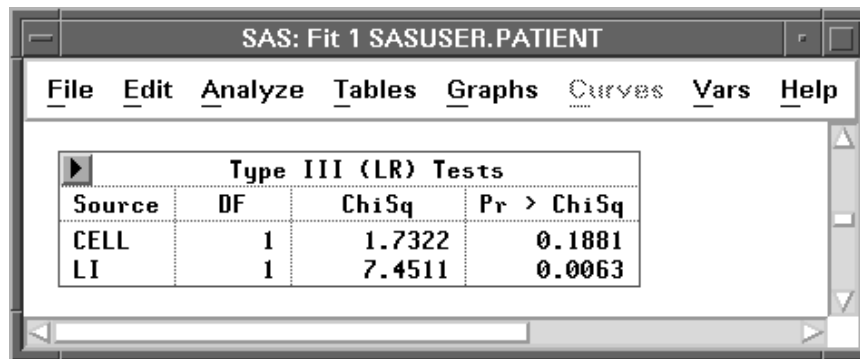


Figure 16.8. Fit Window with CELL and LI as Explanatory Variables

- ⇒ Choose **Tables:Type III (LR) Tests** in the new fit window.
This displays a **Type III (LR) Tests** table in the window.



The screenshot shows a SAS window titled "SAS: Fit 1 SASUSER.PATIENT". The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. The main area displays a table titled "Type III (LR) Tests".

Source	DF	ChiSq	Pr > ChiSq
CELL	1	1.7322	0.1881
LI	1	7.4511	0.0063

Figure 16.9. Likelihood Ratio Type III Tests

The p -value for **CELL** in the LR test suggests that this effect could also be removed.

- ⇒ Click on the variable **CELL** in the effects list in the Fit Dialog.
Then click on **Remove**. **CELL** disappears from the effects list.
- ⇒ Click on **Apply**, and a new Fit window appears, as shown in [Figure 16.10](#).
Since the new model contains only one **X** variable, the fit window displays a plot of **REMISS** versus **CELL**.

Using the **Apply** button, you have quickly created three logistic regression models. Logistic regression is only one special case of the generalized linear model. Another case, Poisson regression, is described in the next chapter.

- ⊕ **Related Reading:** Generalized Linear Models, [Chapter 39](#).

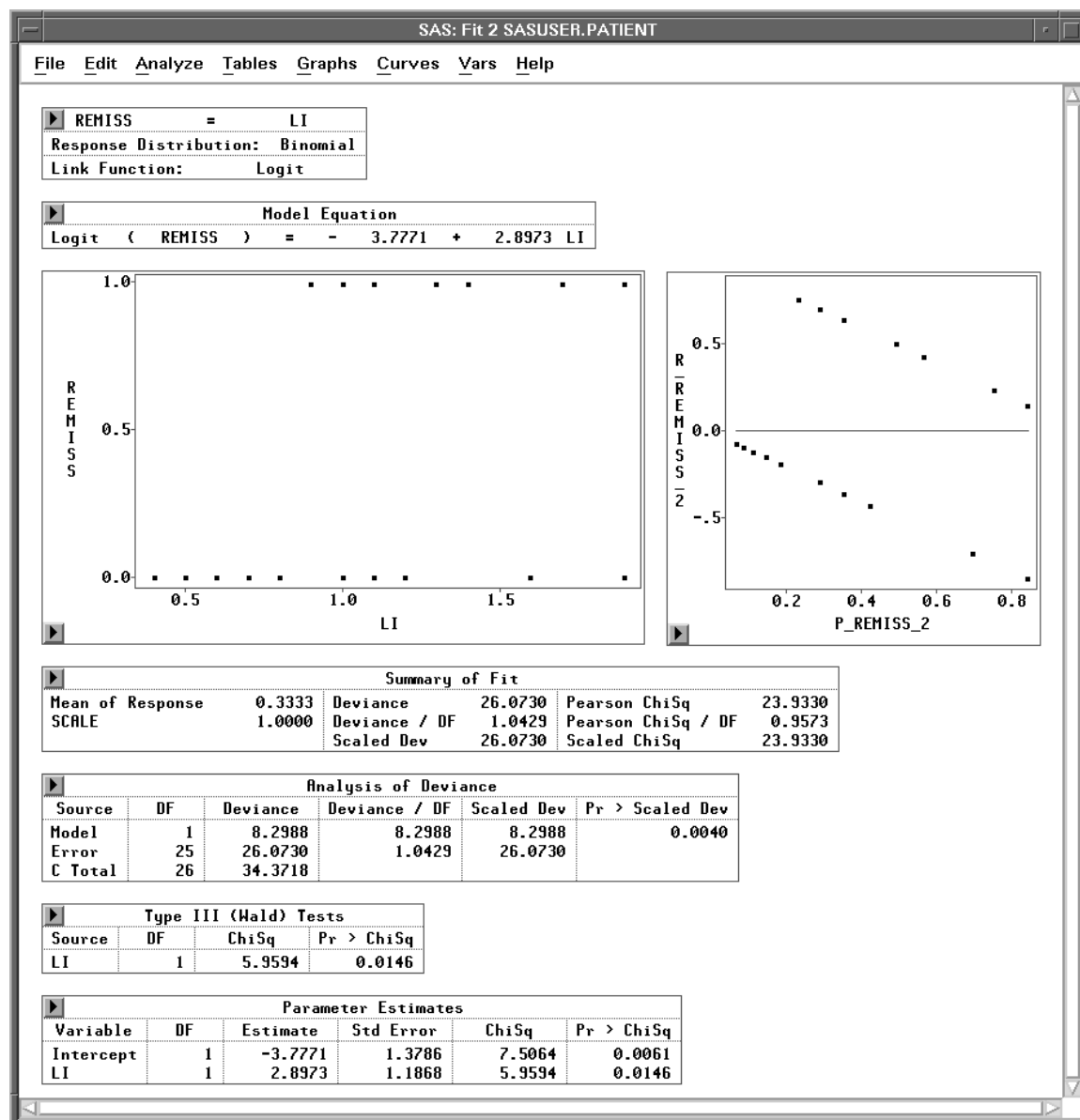


Figure 16.10. Fit Window with LI as the Only Explanatory Variable

References

- Lee, E.T. (1974), "A Computer Program for Linear Logistic Regression Analysis," *Computer Programs in Biomedicine*, 80–92.
- McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman and Hall.

Chapter 17

Poisson Regression

Chapter Contents

DISPLAYING THE POISSON REGRESSION ANALYSIS	281
Model Information	287
Summary of Fit	287
Analysis of Deviance	287
Type III (Wald) Tests	287
MODIFYING THE MODEL	288
Parameter Estimates	290
REFERENCES	291

Chapter 17

Poisson Regression

In [Chapter 16, “Logistic Regression,”](#) you examined logistic regression as an example of a generalized linear model.

In this chapter, you will examine another example of a generalized linear model, Poisson regression. You can choose **Analyze:Fit (Y X)** to carry out a Poisson regression analysis when the response variable represents counts. You can use the fit variables and methods dialogs to specify this generalized linear model.

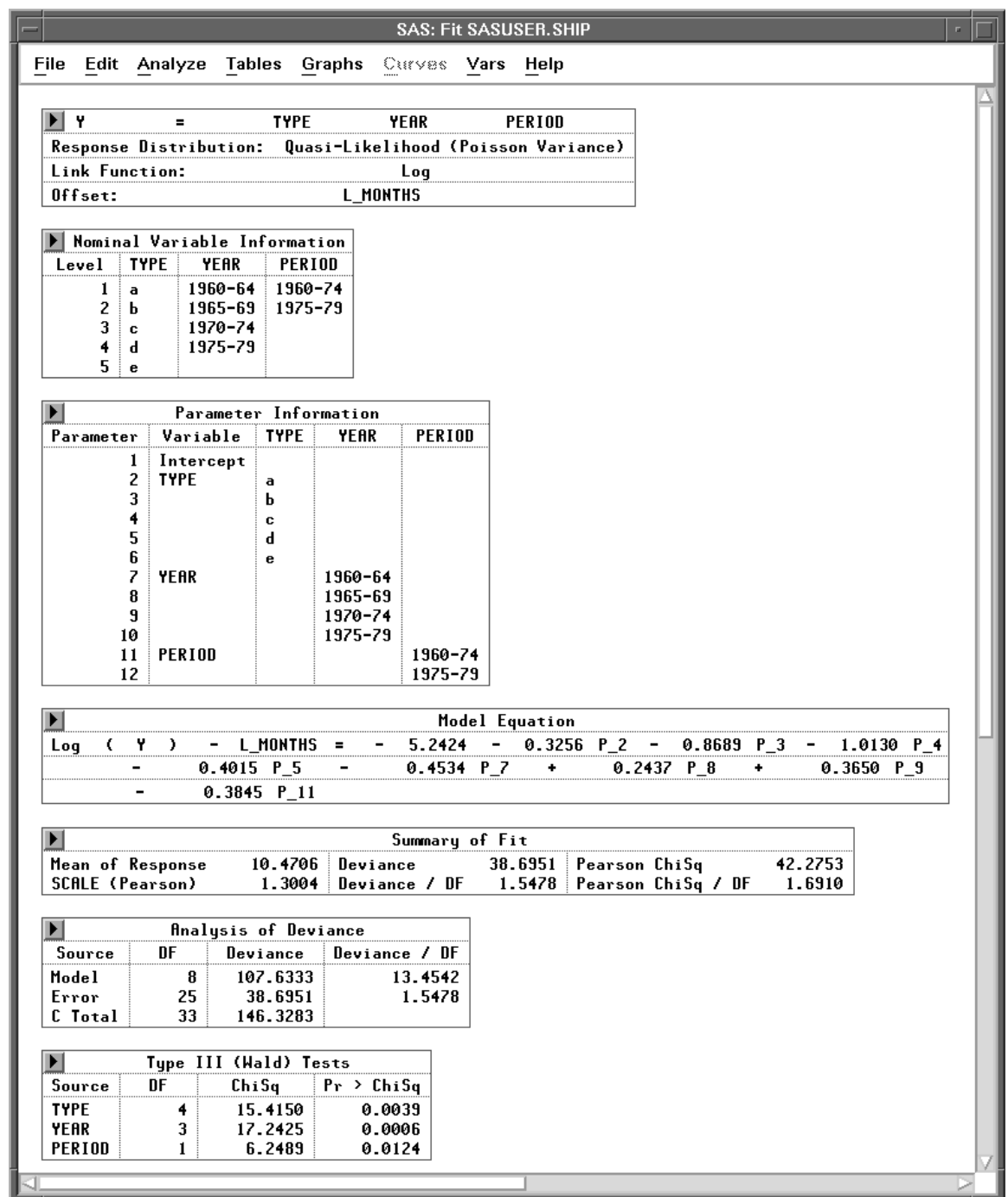


Figure 17.1. Poisson Regression Analysis

Displaying the Poisson Regression Analysis

The **SHIP** data shown in Figure 17.2 represent damage caused by waves to the forward section of certain cargo-carrying vessels. The purpose of the investigation was to set standards for future hull construction. In order to do so, the investigators needed to know the risk of damage associated with five ship types (**TYPE**), year of construction (**YEAR**), and period of operation (**PERIOD**). These three variables are the classification variables. **MONTHS** is the aggregate number of months in service and is an explanatory variable. **Y** is the response variable and represents the number of damage incidents (McCullagh and Nelder 1989).

	5	Nom	Nom	Nom	Int	Int
40	TYPE	YEAR	PERIOD	MONTHS	Y	
1	b	1965-69	1975-79	20370	53	
2	b	1970-74	1960-74	7064	12	
3	b	1970-74	1975-79	13099	44	
4	b	1975-79	1960-74	0	0	
5	b	1975-79	1975-79	7117	18	
6	b	1960-64	1960-74	44882	39	
7	b	1960-64	1975-79	17176	29	
8	b	1965-69	1960-74	28609	58	
9	c	1960-64	1960-74	1179	1	
10	c	1960-64	1975-79	552	1	

Figure 17.2. SHIP Data Set

Recall from Chapter 16 that the generalized linear model has three basic components:

- a linear function of explanatory variables. For this example, the function is

$$\beta_0 + \beta_1 \log(\text{MONTHS}) + \gamma_i + \tau_j + \delta_k + (\gamma\tau)_{ij} + (\gamma\delta)_{ik} + (\tau\delta)_{jk}$$

where $\log(\text{MONTHS})$ is a variable whose coefficient β_1 is believed to be 1. An effect such as this is commonly referred to as an *offset*. γ_i is the effect of the i th level of **TYPE**, τ_j is the effect of the j th level of **YEAR**, δ_k is the effect of the k th level of **PERIOD**, $(\gamma\tau)_{ij}$ is the effect of the ij th level of the **TYPE** by **YEAR** interaction, $(\gamma\delta)_{ik}$ is the effect of the ik th level of the **TYPE** by **PERIOD** interaction, and $(\tau\delta)_{jk}$ is the effect of the jk th level of the **YEAR** by **PERIOD** interaction.

- a probability function for the response variable that depends on the mean and sometimes other parameters as well. For this example, the probability function of the response variable is Poisson.

- a link function that relates the mean to the linear function of explanatory variables. For this example, the link function is the log

$\log(\text{expected number of damage incidents})$

$$= \beta_0 + \beta_1 \log(\text{MONTHS}) + \gamma_i + \tau_j + \delta_k + (\gamma\tau)_{ij} + (\gamma\delta)_{ik} + (\tau\delta)_{jk}$$

⇒ **Open the SHIP data set.**

Recall from the previous equation that **Y** is assumed to be directly proportional to **MONTHS**. Since $\log(Y)$ is being modeled, you need to carry out a log transformation on **MONTHS**. Follow these steps to create a new variable that represents the log of **MONTHS**.

⇒ **Select MONTHS in the data window.**

⇒ **Choose Edit:Variables:log(Y).**

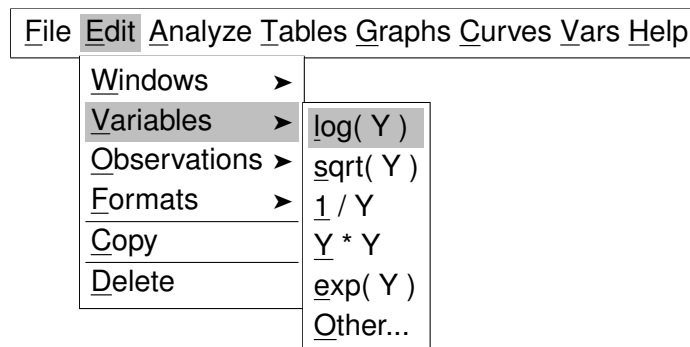


Figure 17.3. Edit:Variables Menu

A new variable, **L_MONTHS**, now appears in the data window.

SAS: SASUSER.SHIP							
File Edit Analyze Tables Graphs Curves Vars Help							
6	Nom	Nom	Nom	Int	Int	Int	
40	TYPE	YEAR	PERIOD	MONTHS	Y	L_MONTHS	
1	b	1965-69	1975-79	20370	53	9.9218	
2	b	1970-74	1960-74	7064	12	8.8628	
3	b	1970-74	1975-79	13099	44	9.4803	
4	b	1975-79	1960-74	0	0	.	
5	b	1975-79	1975-79	7117	18	8.8702	
6	b	1960-64	1960-74	44882	39	10.7118	
7	b	1960-64	1975-79	17176	29	9.7513	
8	b	1965-69	1960-74	28609	58	10.2615	
9	c	1960-64	1960-74	1179	1	7.0724	
10	c	1960-64	1975-79	552	1	6.3135	

Figure 17.4. Data Window with **L_MONTHS** Added

⇒ **Deselect L_MONTHS in the data window.** Some values of **MONTHS** are **0**, meaning that this kind of ship has not seen service. You need to restrict these observations from entering into the model fit. The log transformation does this automatically since $\log(\mathbf{MONTHS})$ becomes a missing value for the observations with a value of **0** for **MONTH**. Observations with missing values for the explanatory variables or the response variable are not used in the model fit.

Now you are ready to begin the analysis.

⇒ **Choose Analyze:Fit (Y X) to display the fit variables dialog**

⇒ **Select Y in the list at the left, then click the Y button.**

Y appears in the **Y** variables list.

⇒ **Select TYPE, YEAR, and PERIOD, then click the Expand button.**

TYPE, **YEAR**, and **PERIOD**, along with all two-way interaction effects, appear in the **X** variables list. Your variables dialog should now appear as shown in [Figure 17.5](#).



Figure 17.5. Fit Variables Dialog with Variable Roles Assigned

The **Expand** button provides a convenient way to specify interactions of any order. The order **2** is the default. You can change the order by entering a different value to replace the **2** or by clicking on the buttons to the right or left of the **2** to increase or decrease the order, respectively.

⇒ **Click the Method button to display the fit method dialog**

This dialog enables you to specify the probability function or the quasi-likelihood function for the response variable and the link function.

Overdispersion is a phenomenon that occurs occasionally with binomial and Poisson data. For Poisson data, it occurs when the variance of the response Y exceeds the Poisson variance $\text{Var}(y)=\mu$. To account for the overdispersion that might occur in the **SHIP** data set, a quasi-likelihood function with variance function $\text{Var}(\mu)=\mu$ (Poisson variance) will be used for the response variable. The variance is given by

$$\text{Var}(y) = \sigma^2 \mu$$

where σ^2 is the dispersion parameter with value greater than 1 for overdispersion.

⇒ **Select the check box for Quasi-Likelihood.**

⇒ **Click on Poisson under Response Dist.**

This uses the Poisson variance function $\text{Var}(\mu) = \mu$ for the quasi-likelihood function.

⇒ **Click on Pearson under Scale.**

This uses the scale parameter based on the Pearson χ^2 statistic.

⇒ **Select L_MONTHS in the list at the left, then click the Offset button.**

L_MONTHS appears in the **Offset** variables list. Your method dialog should now appear as shown in [Figure 17.6](#).



Figure 17.6. Fit Method Dialog

It is not necessary to specify a **Link Function**. **Canonical** is the default and allows

Fit (Y X) to choose an appropriate link. For this example, it is equivalent to choosing **Log** as the **Link Function**.

⇒ Click the **OK** button to close both dialogs and display the analysis.

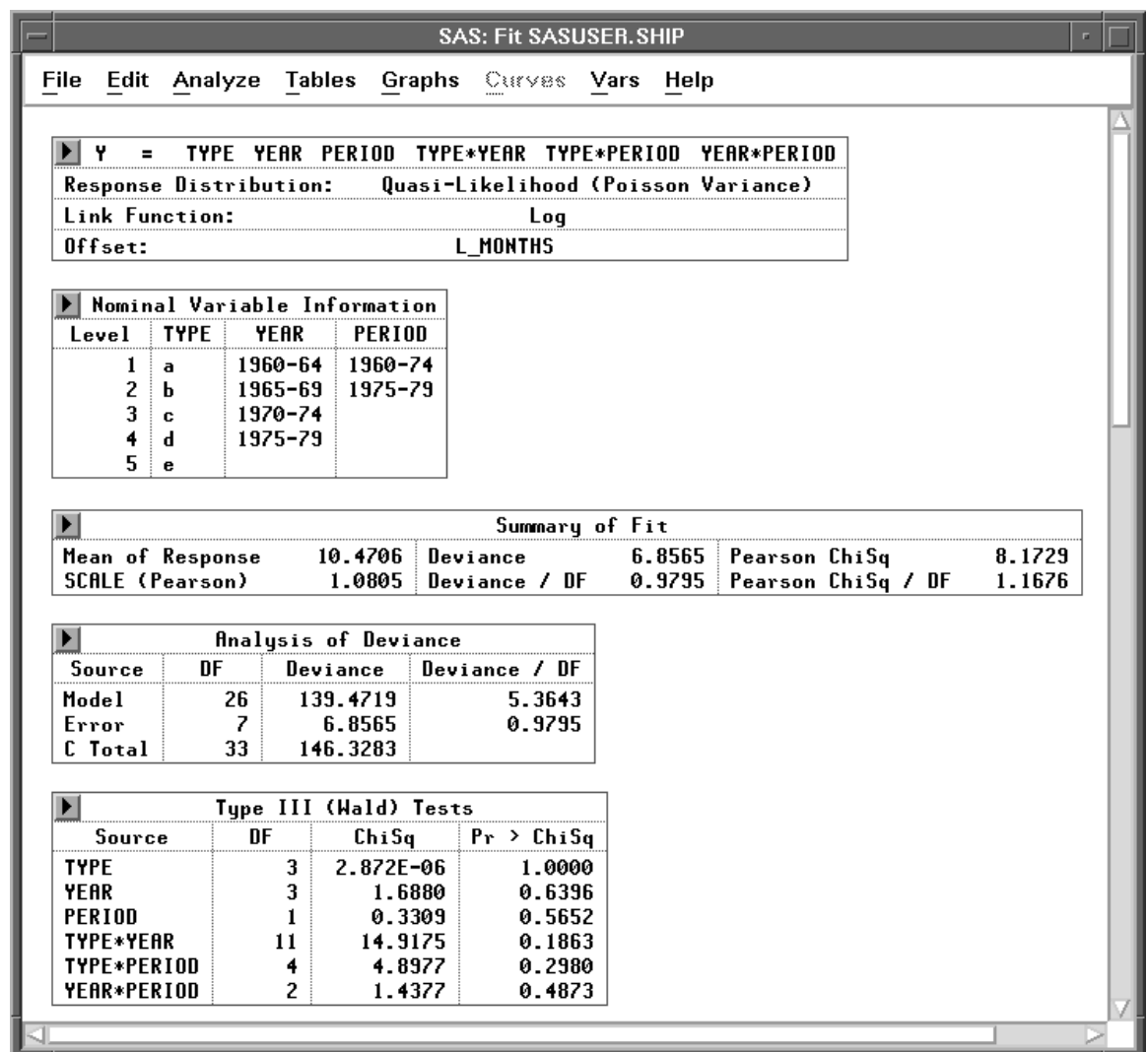


Figure 17.7. Fit Window

By default, the window includes many tables, but only a few are shown in [Figure 17.7](#). These tables are described in the following sections. For more information about the other tables and graphs in the window, see [Chapter 39, “Fit Analyses.”](#)

† **Note:** A warning message—The negative of the Hessian is not positive definite. The convergence is questionable—appears when the specified model does not converge, as in this example. The output tables, graphs, and variables are based on the results from the last iteration.

Model Information

Begin by examining the table at the top of the fit window that describes the model. The first line gives the effects in the model. The second line gives the response distribution from which the variance function used in the quasi-likelihood function is obtained. The third line gives the link function of **Y**. When an **Offset** variable is also specified in the fit method dialog, the fourth line gives the offset in the model.

The **Nominal Variable Information** table contains the levels of the nominal variables. The **Parameter Information** table, as displayed in [Figure 17.1](#), shows the variable indices for the parameters.

Summary of Fit

The **Summary of Fit** table contains summary statistics including **Mean of Response**, **Deviance**, and **Pearson Chi-Square**. **SCALE (Pearson)** gives the scale parameter estimated from the Pearson χ^2 statistic.

Analysis of Deviance

The **Analysis of Deviance** table summarizes the information related to the sources of variation in the data. **Deviance** represents variation present in the data. **Error** gives the deviance for the current model, and **C Total**, corrected for an overall mean, is the deviance for the model with intercept only. **Model** gives the variation modeled by the explanatory variables, and it is the difference between **C Total** and **Error** deviances.

Type III (Wald) Tests

The **Type III (Wald) Tests** table in this example is a further breakdown of the variation due to **MODEL**. The **DF** for **Model** are broken down into terms corresponding to the main effects for **YEAR**, **TYPE**, and **PERIOD**, and the interaction effects for **TYPE*YEAR**, **YEAR*PERIOD**, and **TYPE*PERIOD**. The composite explanatory power of the set of parameters associated with each effect is measured by the **Chi-Square** statistic. The *p*-value corresponding to each **Chi-Square** statistic is the probability of observing a statistic of equal or greater value, given that the corresponding parameters are all 0.

Modifying the Model

For this model and this set of data, there does not appear to be sufficient explanatory power in the **YEAR*PERIOD** effect to include it in the model.

⇒ Click on **YEAR*PERIOD** in the fit window.

⇒ Choose **Edit:Delete** from the menu.

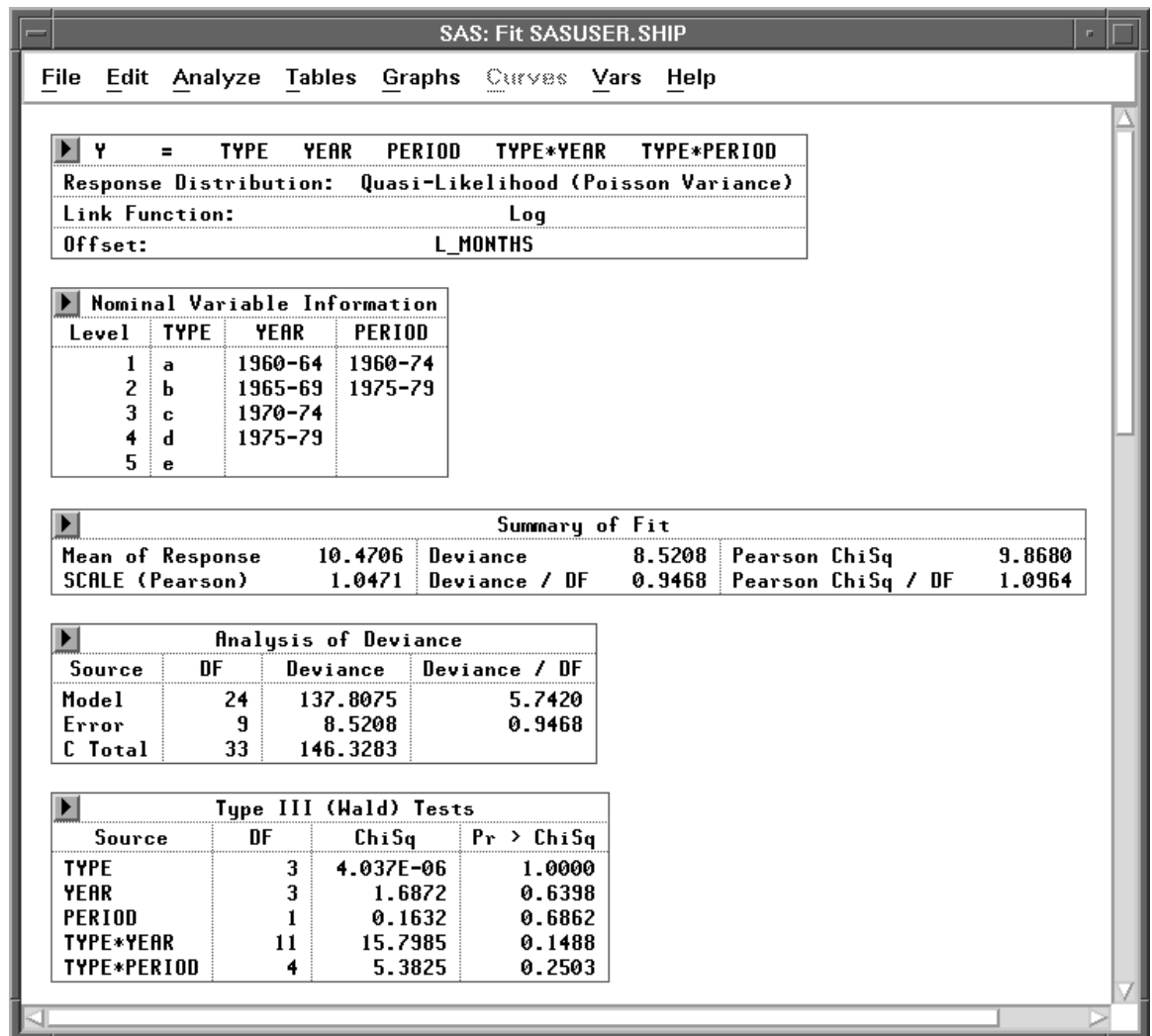


Figure 17.8. Modified Fit Model

Follow the previous steps to remove the other two interaction terms from the model. The resulting main effects model is shown in [Figure 17.9](#).

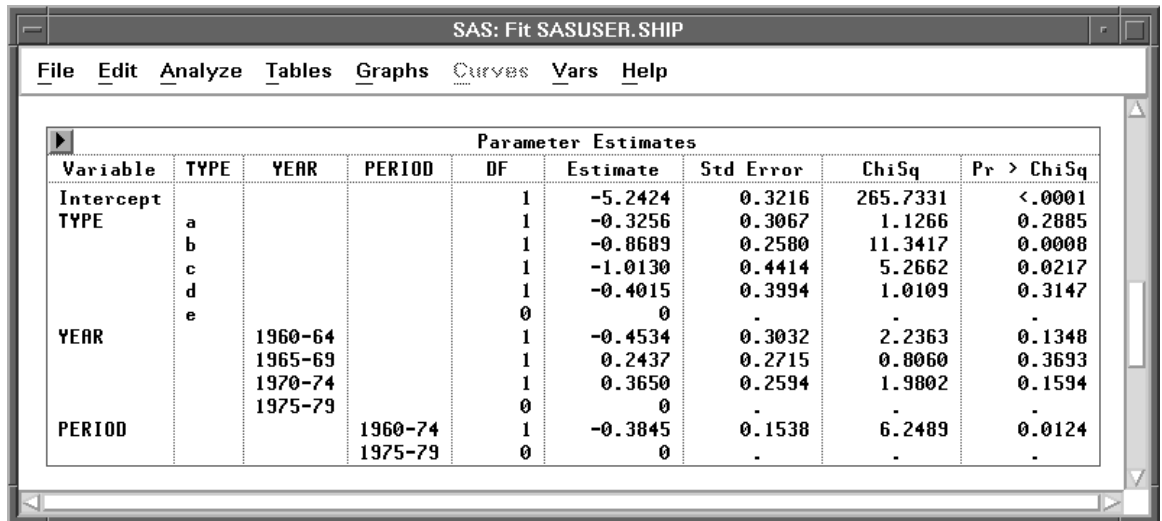
SAS: Fit SASUSER.SHIP																																
File Edit Analyze Tables Graphs Curves Vars Help																																
<table><tr><td>Y</td><td>=</td><td>TYPE</td><td>YEAR</td><td>PERIOD</td></tr><tr><td colspan="5">Response Distribution: Quasi-Likelihood (Poisson Variance)</td></tr><tr><td colspan="5">Link Function: Log</td></tr><tr><td colspan="5">Offset: L_MONTHS</td></tr></table>					Y	=	TYPE	YEAR	PERIOD	Response Distribution: Quasi-Likelihood (Poisson Variance)					Link Function: Log					Offset: L_MONTHS												
Y	=	TYPE	YEAR	PERIOD																												
Response Distribution: Quasi-Likelihood (Poisson Variance)																																
Link Function: Log																																
Offset: L_MONTHS																																
<table><tr><th colspan="4">Nominal Variable Information</th></tr><tr><th>Level</th><th>TYPE</th><th>YEAR</th><th>PERIOD</th></tr><tr><td>1</td><td>a</td><td>1960-64</td><td>1960-74</td></tr><tr><td>2</td><td>b</td><td>1965-69</td><td>1975-79</td></tr><tr><td>3</td><td>c</td><td>1970-74</td><td></td></tr><tr><td>4</td><td>d</td><td>1975-79</td><td></td></tr><tr><td>5</td><td>e</td><td></td><td></td></tr></table>					Nominal Variable Information				Level	TYPE	YEAR	PERIOD	1	a	1960-64	1960-74	2	b	1965-69	1975-79	3	c	1970-74		4	d	1975-79		5	e		
Nominal Variable Information																																
Level	TYPE	YEAR	PERIOD																													
1	a	1960-64	1960-74																													
2	b	1965-69	1975-79																													
3	c	1970-74																														
4	d	1975-79																														
5	e																															
<table><tr><th colspan="5">Summary of Fit</th></tr><tr><td>Mean of Response</td><td>10.4706</td><td>Deviance</td><td>38.6951</td><td>Pearson ChiSq</td><td>42.2753</td></tr><tr><td>SCALE (Pearson)</td><td>1.3004</td><td>Deviance / DF</td><td>1.5478</td><td>Pearson ChiSq / DF</td><td>1.6910</td></tr></table>					Summary of Fit					Mean of Response	10.4706	Deviance	38.6951	Pearson ChiSq	42.2753	SCALE (Pearson)	1.3004	Deviance / DF	1.5478	Pearson ChiSq / DF	1.6910											
Summary of Fit																																
Mean of Response	10.4706	Deviance	38.6951	Pearson ChiSq	42.2753																											
SCALE (Pearson)	1.3004	Deviance / DF	1.5478	Pearson ChiSq / DF	1.6910																											
<table><tr><th colspan="4">Analysis of Deviance</th></tr><tr><th>Source</th><th>DF</th><th>Deviance</th><th>Deviance / DF</th></tr><tr><td>Model</td><td>8</td><td>107.6333</td><td>13.4542</td></tr><tr><td>Error</td><td>25</td><td>38.6951</td><td>1.5478</td></tr><tr><td>C Total</td><td>33</td><td>146.3283</td><td></td></tr></table>					Analysis of Deviance				Source	DF	Deviance	Deviance / DF	Model	8	107.6333	13.4542	Error	25	38.6951	1.5478	C Total	33	146.3283									
Analysis of Deviance																																
Source	DF	Deviance	Deviance / DF																													
Model	8	107.6333	13.4542																													
Error	25	38.6951	1.5478																													
C Total	33	146.3283																														
<table><tr><th colspan="4">Type III (Wald) Tests</th></tr><tr><th>Source</th><th>DF</th><th>ChiSq</th><th>Pr > ChiSq</th></tr><tr><td>TYPE</td><td>4</td><td>15.4150</td><td>0.0039</td></tr><tr><td>YEAR</td><td>3</td><td>17.2425</td><td>0.0006</td></tr><tr><td>PERIOD</td><td>1</td><td>6.2489</td><td>0.0124</td></tr></table>					Type III (Wald) Tests				Source	DF	ChiSq	Pr > ChiSq	TYPE	4	15.4150	0.0039	YEAR	3	17.2425	0.0006	PERIOD	1	6.2489	0.0124								
Type III (Wald) Tests																																
Source	DF	ChiSq	Pr > ChiSq																													
TYPE	4	15.4150	0.0039																													
YEAR	3	17.2425	0.0006																													
PERIOD	1	6.2489	0.0124																													

Figure 17.9. Main Effects Model

The estimate of the dispersion parameter $\phi = \sigma^2 = 1.6910$ suggests that overdispersion exists in the model. **Type III (Wald) Tests** table shows that all of the main effects are significant.

Parameter Estimates

Analyses where some effects are classification variables yield different parameter estimates from those observed in a regression setting. They represent a different additive contribution for each level value (or combination of level values for interaction effects), and thus the individual elements in the table are not as easily interpretable as they are in multiple regression.



The screenshot shows the SAS Fit window titled "SAS: Fit SASUSER.SHIP". The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. The main display area shows the "Parameter Estimates" table with the following data:

Variable	TYPE	YEAR	PERIOD	DF	Estimate	Std Error	ChiSq	Pr > ChiSq
Intercept				1	-5.2424	0.3216	265.7331	<.0001
TYPE	a			1	-0.3256	0.3067	1.1266	0.2885
	b			1	-0.8689	0.2580	11.3417	0.0008
	c			1	-1.0130	0.4414	5.2662	0.0217
	d			1	-0.4015	0.3994	1.0109	0.3147
	e			0	0	.	.	.
YEAR		1960-64		1	-0.4534	0.3032	2.2363	0.1348
		1965-69		1	0.2437	0.2715	0.8060	0.3693
		1970-74		1	0.3650	0.2594	1.9802	0.1594
		1975-79		0	0	.	.	.
PERIOD			1960-74	1	-0.3845	0.1538	6.2489	0.0124
			1975-79	0	0	.	.	.

Figure 17.10. Parameter Estimates Table

Because the overall level is set by the **INTERCEPT** parameter, the set of parameters associated with an effect is redundant. This shows up in the **Parameter Estimates** table as parameters with degrees of freedom (DF) that are **0** and estimates that are **0**. An example of this is the parameter for the **e** level of the **TYPE** variable.

From the **Parameter Estimates** table, ships of types **b** and **c** have the lowest risk, and ships of type **e** the highest. The oldest ships (built between 1960 and 1964) have the lowest risk and ships built between 1965 and 1974 have the highest risk. Ships operated between 1960 to 1974 have a lower risk than ships operated between 1975 to 1979.

The analysis provides a table for the complete fitted model, but you should not use these parameter estimates and their associated statistics individually to determine which parameters have an effect. For further information on parameter estimates and other features of the Fit window, see [Chapter 39, “Fit Analyses.”](#)

⊕ **Related Reading:** Generalized Linear Models, [Chapter 39](#).

References

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman and Hall.

Chapter 18

Examining Correlations

Chapter Contents

CREATING THE ANALYSIS	296
Correlation Matrix	299
Confidence Ellipses	299
REFERENCES	301

Chapter 18

Examining Correlations

In this chapter you examine relationships between pairs of variables by looking at correlations.

You can use correlation coefficients to measure the strength of the linear association between two variables. You can also use confidence ellipses in scatter plots as a visual test for bivariate normality and an indication of the strength of the correlation.

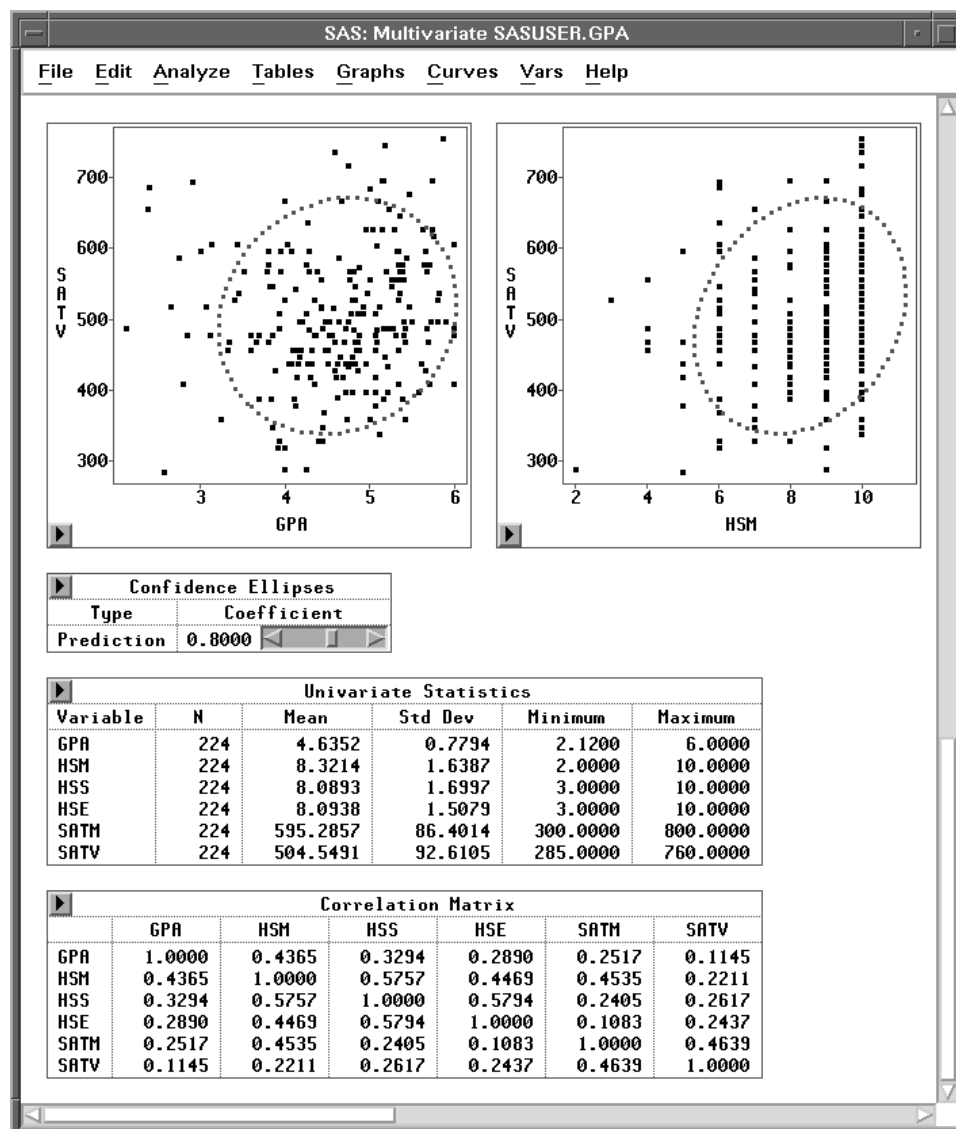


Figure 18.1. Multivariate Window with Correlation Analysis

Creating the Analysis

The **GPA** data set contains information collected to determine which applicants at a university were likely to succeed in its computer science program. The variable **GPA** is the grade point average; **HSM**, **HSS**, and **HSE** are average high school grades in mathematics, science, and English; and **SATM** and **SATV** are scores on the mathematics and verbal portion of the SAT exam (Moore and McCabe 1989).

Follow these steps to create a correlation analysis of the **GPA** data.

⇒ **Open the GPA data set.**

SAS: SASUSER.GPA

File Edit Analyze Tables Graphs Curves Vars Help

7	Int	Int	Int	Int	Int	Int	Nom		
224	GPA	HSM	HSS	HSE	SATM	SATV	SEX		
1	5.32	10	10	10	670	600	Female		
2	5.14	9	9	10	630	700	Male		
3	3.84	9	6	6	610	390	Female		
4	5.34	10	9	9	570	530	Male		
5	4.26	6	8	5	700	640	Female		
6	4.35	8	6	8	640	530	Female		
7	5.33	9	7	9	630	560	Male		
8	4.85	10	8	8	610	460	Male		
9	4.76	10	10	10	570	570	Male		
10	5.72	7	8	7	550	500	Female		

Figure 18.2. GPA Data

⇒ **Choose Analyze:Multivariate (Y's).**

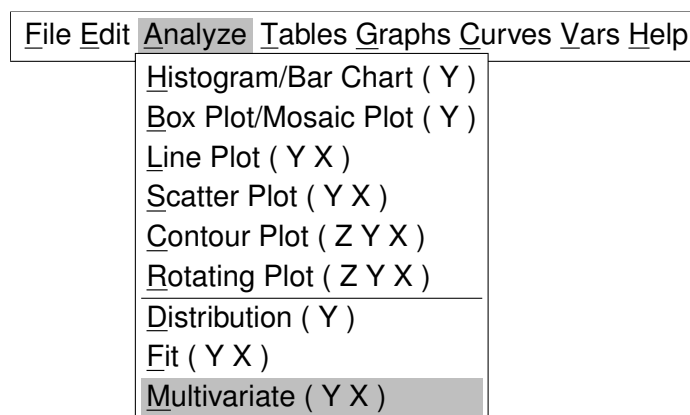


Figure 18.3. Analyze Menu

⇒ Select **GPA, HSM, HSS, HSE, SATM, and SATV**. Then click the **Y** button to assign these variables the **Y** role.

Your variables dialog should now appear, as shown in [Figure 18.4](#).



Figure 18.4. Multivariate Variables Dialog

⇒ Click **OK** to create the multivariate window.

By default, the multivariate window contains tables of **Univariate Statistics** and the **Correlation Matrix**.

SAS: Multivariate SASUSER.GPA						
File Edit Analyze Tables Graphs Curves Vars Help						
▶ GPA HSM HSS HSE SATM SATV						
▶ Univariate Statistics						
Variable	N	Mean	Std Dev	Minimum	Maximum	
GPA	224	4.6352	0.7794	2.1200	6.0000	
HSM	224	8.3214	1.6387	2.0000	10.0000	
HSS	224	8.0893	1.6997	3.0000	10.0000	
HSE	224	8.0938	1.5079	3.0000	10.0000	
SATM	224	595.2857	86.4014	300.0000	800.0000	
SATV	224	504.5491	92.6105	285.0000	760.0000	
▶ Correlation Matrix						
	GPA	HSM	HSS	HSE	SATM	SATV
GPA	1.0000	0.4365	0.3294	0.2890	0.2517	0.1145
HSM	0.4365	1.0000	0.5757	0.4469	0.4535	0.2211
HSS	0.3294	0.5757	1.0000	0.5794	0.2405	0.2617
HSE	0.2890	0.4469	0.5794	1.0000	0.1083	0.2437
SATM	0.2517	0.4535	0.2405	0.1083	1.0000	0.4639
SATV	0.1145	0.2211	0.2617	0.2437	0.4639	1.0000

Figure 18.5. Multivariate Window

Correlation Matrix

Examine the **Correlation Matrix** table. The *correlation coefficient* is a numerical measure that quantifies the strength of linear relationships. **GPA**, the grade point average, shows a correlation of 0.4365 with **HSM**, the high school math average. This is not surprising since you would expect the more successful computer science majors to have stronger quantitative skills.

GPA is not as strongly correlated with the other variables and shows a correlation of only 0.1145 with **SATV**. The verbal portion of the SAT exam does not measure the quantitative skills needed by computer science majors.

Confidence Ellipses

To learn more about correlations in the data, add a scatter plot matrix with confidence ellipses for all of the variables under consideration.

⇒ Choose **Curves:Confidence Ellipse:Prediction: 80%**.

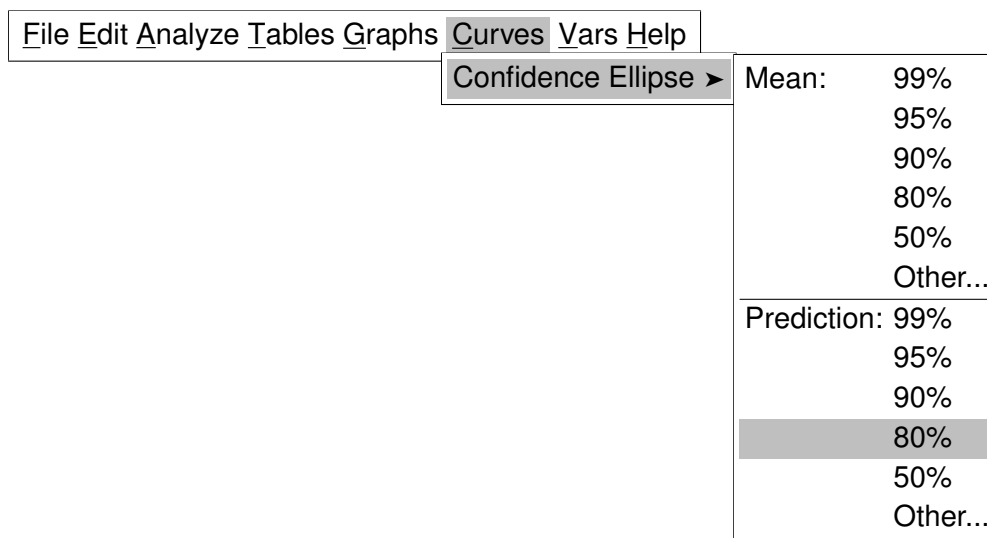


Figure 18.6. Curves Menu

The lower half of the scatter plot matrix for the six variables appears on your display with the 80% prediction confidence ellipses drawn, as shown in [Figure 18.7](#).

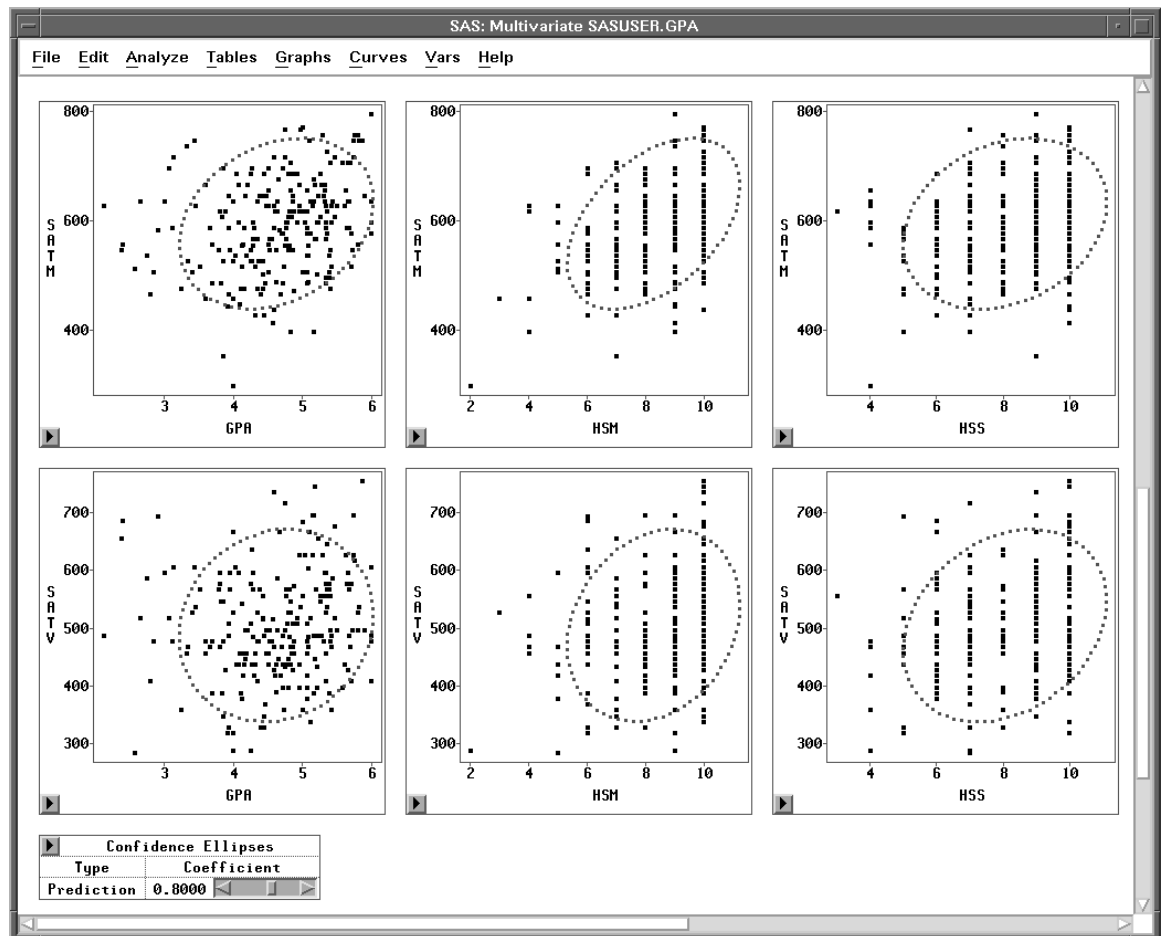


Figure 18.7. Multivariate Window with Confidence Ellipses

There are two ways to interpret the ellipses: as confidence curves for bivariate normal distributions and as indicators of correlation.

As confidence curves, the ellipses show where the specified percentage of the data should lie, assuming a bivariate normal distribution. Under bivariate normality, the percentage of observations falling inside the ellipse should closely agree with the specified confidence level. You can examine the effect of increasing or decreasing the confidence level by adjusting the slider in the **Confidence Ellipses** table below the scatter plot matrix.

Confidence ellipses can also serve as visual indicators of correlations. The confidence ellipse collapses diagonally as the correlation between two variables approaches 1 or -1. The confidence ellipse is more circular when two variables are uncorrelated.

In this case the scatter plots for high school scores (**HSM**, **HSS**, and **HSE**) show a granular appearance that indicates the data are not continuous. These scatter plots clearly do not follow a bivariate normal distribution; therefore, it is not appropriate to interpret confidence ellipses.

The confidence ellipses for **GPA**, **SATM**, and **SATV** can be interpreted. These confidence ellipses contain observations appropriate to the 80% confidence level you specified. The nearly circular appearance of the confidence ellipse in the plot of **GPA** versus **SATV** reflects the small correlation you observed in the **Correlation Matrix** table. The ellipse in the plot of **GPA** versus **SATM** is somewhat more elongated, reflecting a higher correlation.

† **Note:** Visual interpretation of correlations can be subjective because changes in scale affect your perception (Moore and McCabe 1989). When examining correlations, you should use correlation coefficients as well as confidence ellipses.

⊕ **Related Reading:** Correlation Coefficients, Confidence Ellipses, [Chapter 40](#).

References

Moore, D.S. and McCabe, G.P. (1989), *Introduction to the Practice of Statistics*, New York: W.H. Freeman and Company, 179–199.

Chapter 19

Calculating Principal Components

Chapter Contents

CALCULATING PRINCIPAL COMPONENTS	306
Principal Component Tables	311
Principal Component Plots	313
PLOTTING AGAINST ORIGINAL VARIABLES	314
SAVING PRINCIPAL COMPONENTS	316

Chapter 19

Calculating Principal Components

Principal component analysis is a technique for reducing the complexity of high dimensional data. You can use principal component analysis to approximate high dimensional data with a few dimensions so you can examine them visually. In SAS/INSIGHT software you can calculate principal components, store them, and plot them in two and three dimensions.

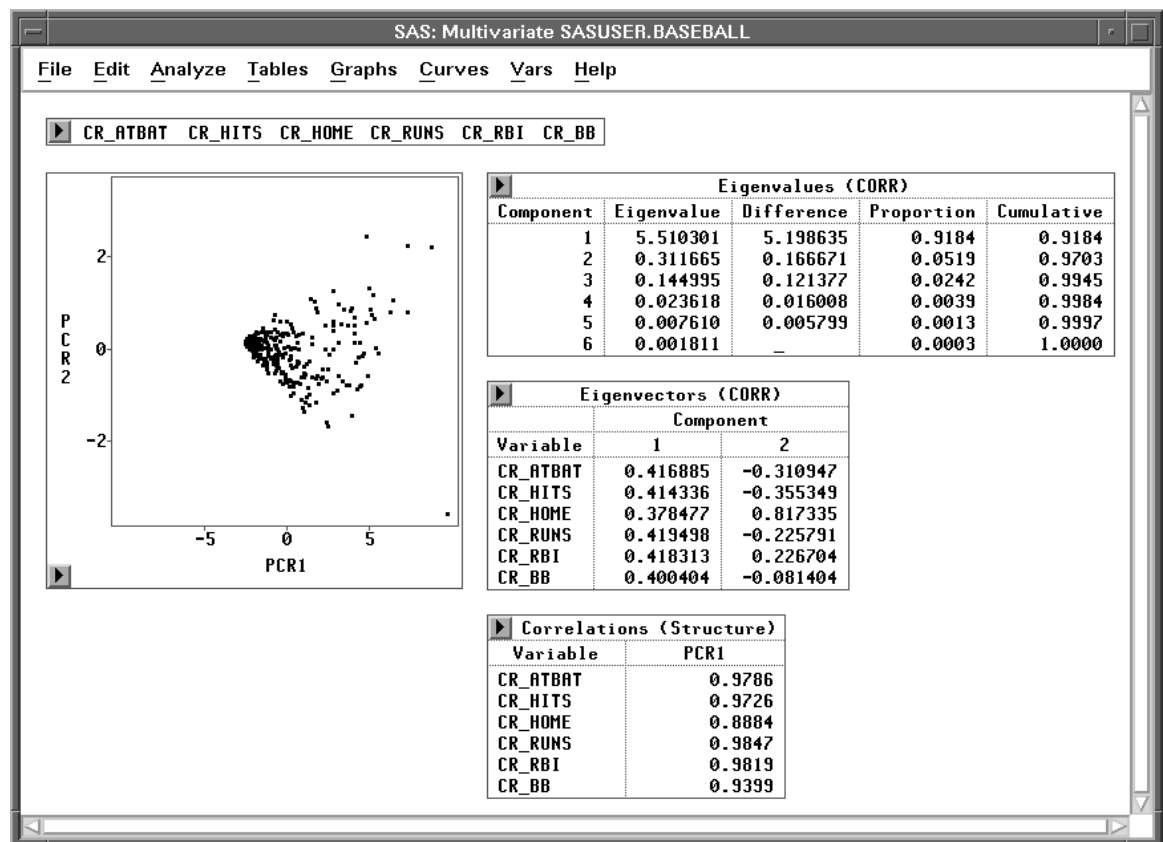


Figure 19.1. Principal Component Analysis

Calculating Principal Components

Principal component analysis summarizes high dimensional data into a few dimensions. Each dimension is called a *principal component* and represents a linear combination of the variables. The first principal component accounts for as much variation in the data as possible. Each succeeding principal component accounts for as much of the variation unaccounted for by preceding principal components as possible.

Consider the **BASEBALL** data set. These data contain performance measures and salary levels for regular hitters and leading substitute hitters in the major leagues in 1986. Suppose you are interested in exploring the relationship between players' performances and their salaries.

If you can first reduce the six career hitting and fielding variables into two or three dimensions—that is, two or three linear combinations of these variables—then graphing these against the **SALARY** variable would be useful. You can then look for relationships between performance and salary.

To create the principal component analysis, follow these steps.

- ⇒ **Open the BASEBALL data set.**
- ⇒ **Choose Analyze:Multivariate (Y's).**

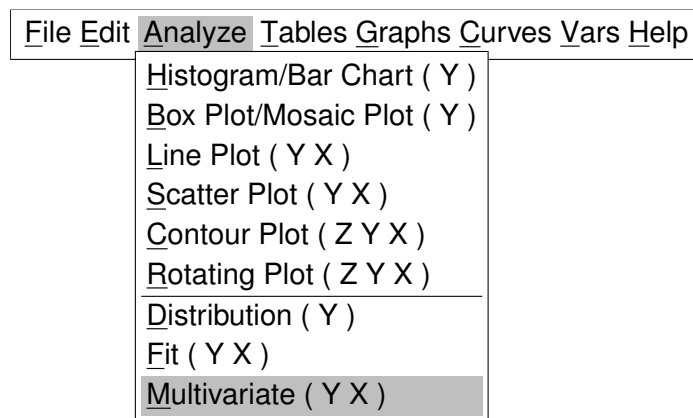


Figure 19.2. Analyze Menu

- ⇒ **Select the six career hitting variables in the list at the left.**
These are **CR_ATBAT**, **CR_HITS**, **CR_HOME**, **CR_RUNS**, **CR_RBI**, and **CR_BB**. Click the **Y** button. The selected variables appear in the **Y** variables list.
- ⇒ **Select NAME in the list at the left, then click the Label button.**
NAME appears in the **Label** variables list. Your variables dialog should now appear as shown in [Figure 19.3](#).

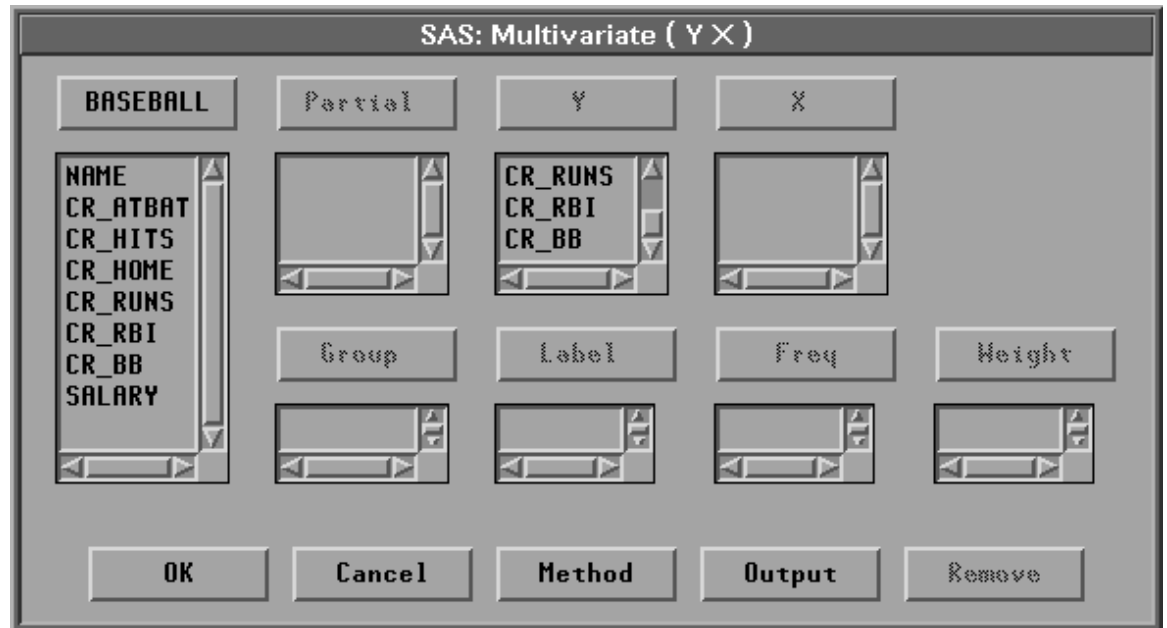


Figure 19.3. Variables Dialog with Variable Roles Assigned

⇒ **Click the Output button.**

The output options dialog appears.

⇒ **Click the Principal Component Analysis check box in the output options dialog**

This requests a principal component analysis. Your output options dialog should now appear as shown in [Figure 19.4](#).

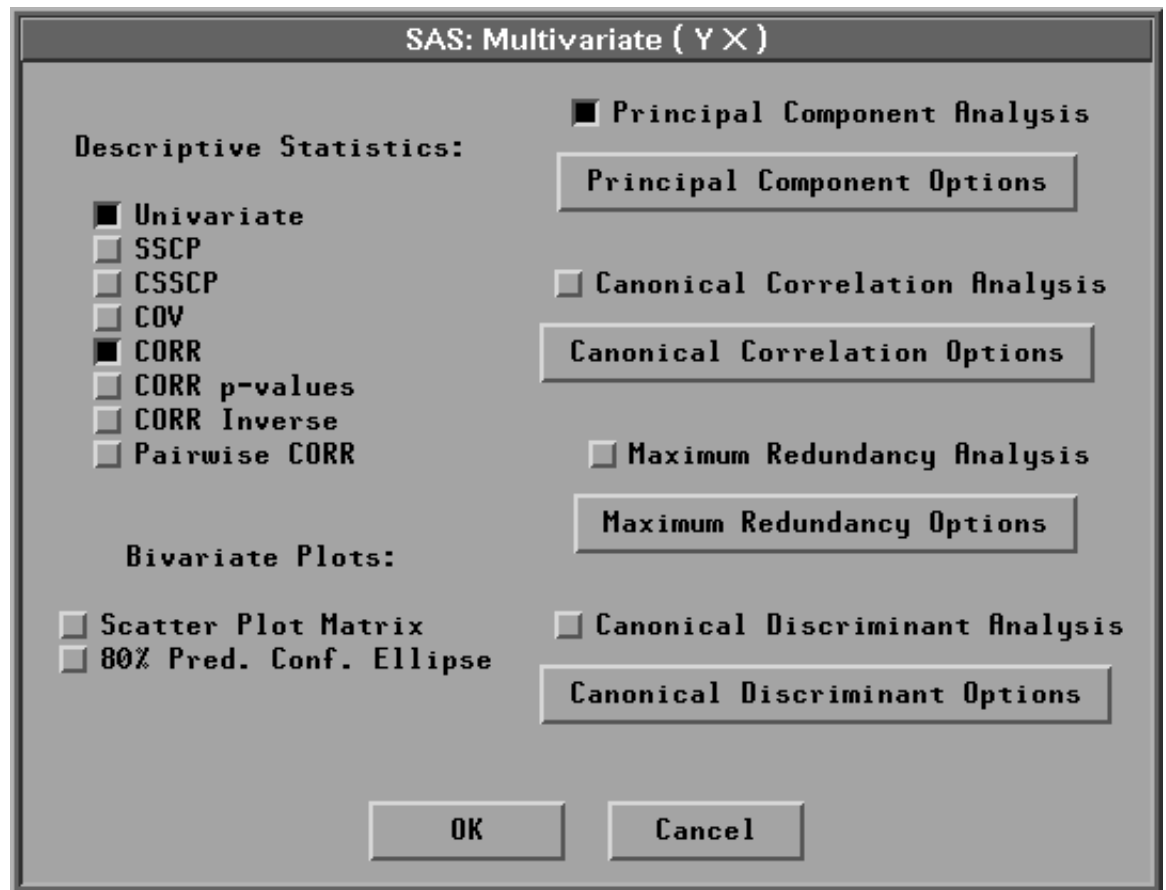


Figure 19.4. Multivariate Output Options Dialog

- ⇒ Click the **Principal Component Options** button in the output options dialog
 A principal component options dialog should now appear as shown in [Figure 19.5](#).

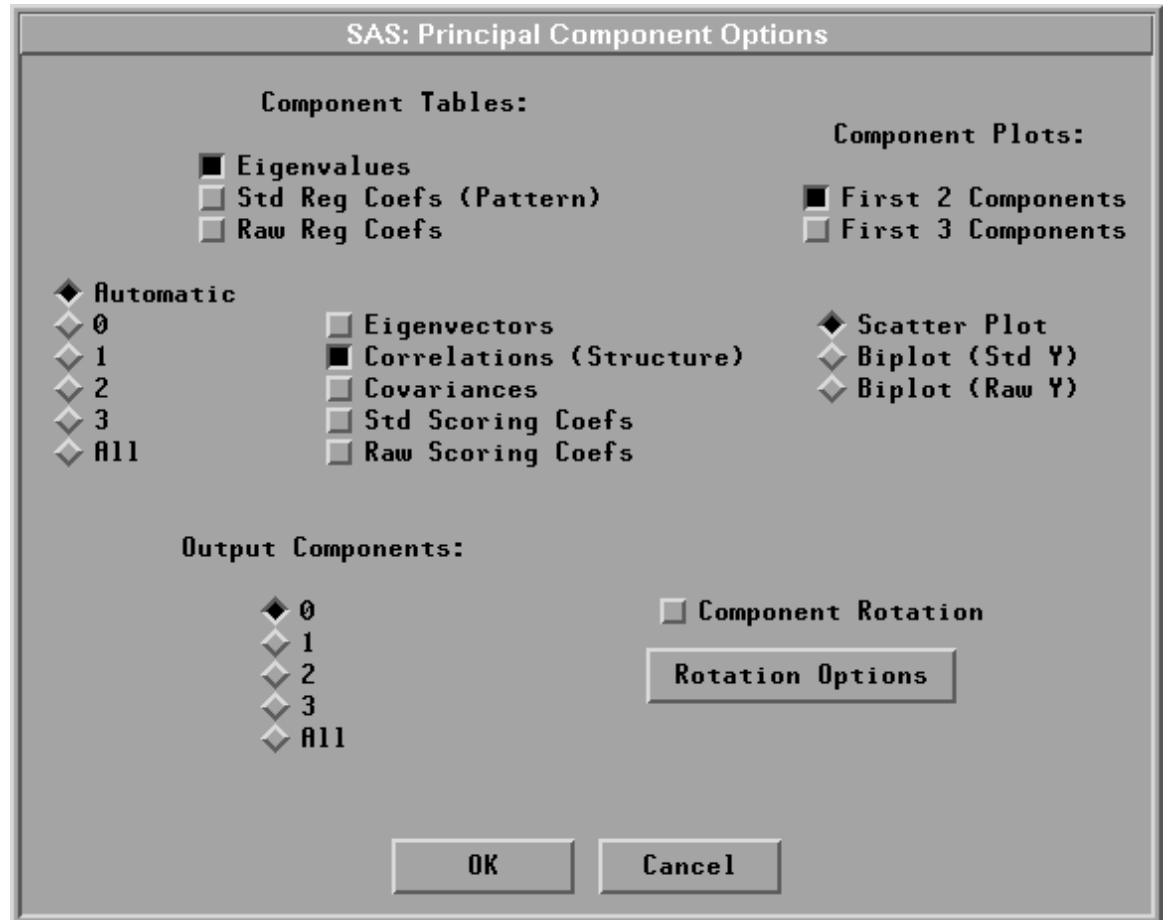


Figure 19.5. Principal Component Options Dialog

- ⇒ Click the **Eigenvectors** check box in the principal component options dialog
- ⇒ Click the radio mark **2** in the options dialog
 This requests that the first two principal components are used for tables of eigenvectors and correlations.
- † **Note:** By default, the analysis is carried out on the correlation matrix. You can use the covariance matrix instead by setting options with the **Method** button in the Multivariate variables dialog. The covariance matrix is recommended only when all the variables are measured in comparable units.
- ⇒ Click **OK** in all dialogs.
 A multivariate window appears. At the bottom of the window is the principal component analysis, as shown in [Figure 19.6](#).

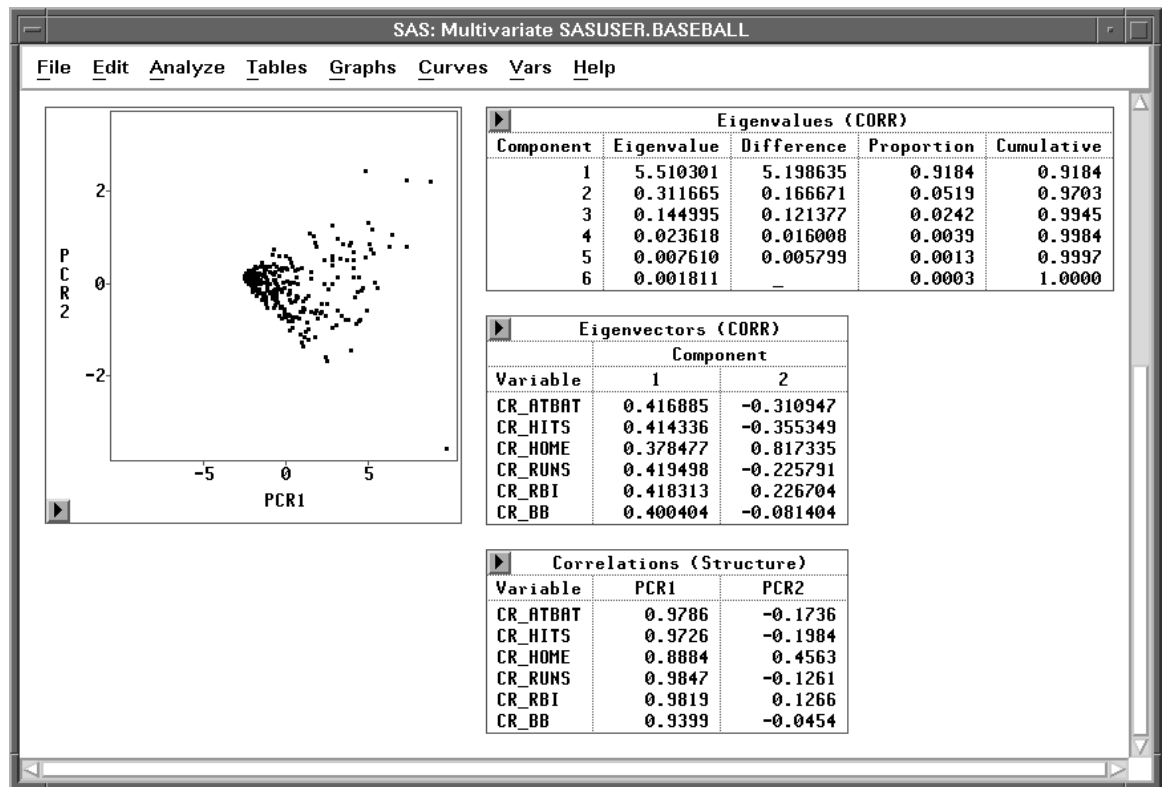


Figure 19.6. Multivariate Window

Principal Component Tables

The **Eigenvalues (CORR)** table illustrated in [Figure 19.7](#) contains all the eigenvalues of the correlation matrix, differences between successive eigenvalues, the proportion of variance explained by each eigenvalue, and the cumulative proportion of the variance explained. Eigenvalues correspond to each of the principal components and represent a partitioning of the total variation in the sample. Because correlations are used, the sum of all the eigenvalues is equal to the number of variables. The first row of the table corresponds to the first principal component, the second row to the second principal component, and so on. In this example, the first two principal components account for over 97% of the variation.

SAS: Multivariate SASUSER.BASEBALL					
File Edit Analyze Tables Graphs Curves Vars Help					
Eigenvalues (CORR)					
Component	Eigenvalue	Difference	Proportion	Cumulative	
1	5.510301	5.198635	0.9184	0.9184	
2	0.311665	0.166671	0.0519	0.9703	
3	0.144995	0.121377	0.0242	0.9945	
4	0.023618	0.016008	0.0039	0.9984	
5	0.007610	0.005799	0.0013	0.9997	
6	0.001811	—	0.0003	1.0000	
Eigenvectors (CORR)					
	Component				
Variable	1	2			
CR_ATBAT	0.416885	-0.310947			
CR_HITS	0.414336	-0.355349			
CR_HOME	0.378477	0.817335			
CR_RUNS	0.419498	-0.225791			
CR_RBI	0.418313	0.226704			
CR_BB	0.400404	-0.081404			
Correlations (Structure)					
Variable	PCR1	PCR2			
CR_ATBAT	0.9786	-0.1736			
CR_HITS	0.9726	-0.1984			
CR_HOME	0.8884	0.4563			
CR_RUNS	0.9847	-0.1261			
CR_RBI	0.9819	0.1266			
CR_BB	0.9399	-0.0454			

Figure 19.7. Principal Component Tables

The **Eigenvectors (CORR)** table illustrated in [Figure 19.7](#) contains the first two eigenvectors of the correlation matrix. Eigenvectors correspond to each of the eigenvalues and associated principal components and are used to form linear combinations of the Y variables. The first column of the table corresponds to the first principal component, and the second column to the second principal component.

Now examine the coefficients making up the eigenvectors. The first component (**PCR1**) appears to be a measure of the player's overall performance as is evidenced

by approximately the same magnitude of the coefficients corresponding to all six variables.

Next examine the coefficients making up the eigenvector for the second principal component (**PCR2**). Only the coefficients associated with the variables **CR_HOME** and **CR_RBI** are positive, and the remaining coefficients are negative. The coefficient with the variable **CR_HOME** is considerably larger than any of the other coefficients. This indicates a measure of career home runs performance versus other performance for 1986.

One way to quantify the strength of the linear relationship between the original Y variables and principal components is through the **Correlations (Structure)** table, as shown in [Figure 19.7](#). This correlation matrix contains the correlations between the Y variables and the principal components.

Eigenvector coefficients of a relatively large magnitude translate into larger correlations and vice versa. For example, **PCR2** has one coefficient substantially larger than other coefficients in the same eigenvector, **CR_HOME**. The correlation of the variable with this **PCR2** is also large.

Principal Component Plots

Examine the scatter plot of the first two principal components shown in [Figure 19.6](#). Each marker on the plot represents two principal component scores. The output component scores are a linear combination of the standardized Y variables with coefficients equal to the eigenvectors of the correlation matrix.

⇒ **Click on the observations with the four highest values for PCR1.**

The resulting scatter plot should now appear as shown in [Figure 19.8](#).

These four observations correspond to Mike Schmidt, Reggie Jackson, Tony Perez, and Pete Rose. The label for Mike Schmidt is not shown because the observation is too close to Reggie Jackson. This is not unexpected since the first principal component is a measure of the player's overall career performance.

Now examine observations in the second principal component direction on the scatter plot. Recall that the second component appeared to be a measure of the combined performance of home runs and runs batted in versus other career performance. The observations with large values of **PCR2** correspond to Mike Schmidt and Reggie Jackson. As one might expect, both players have high career-long home runs and runs batted in.

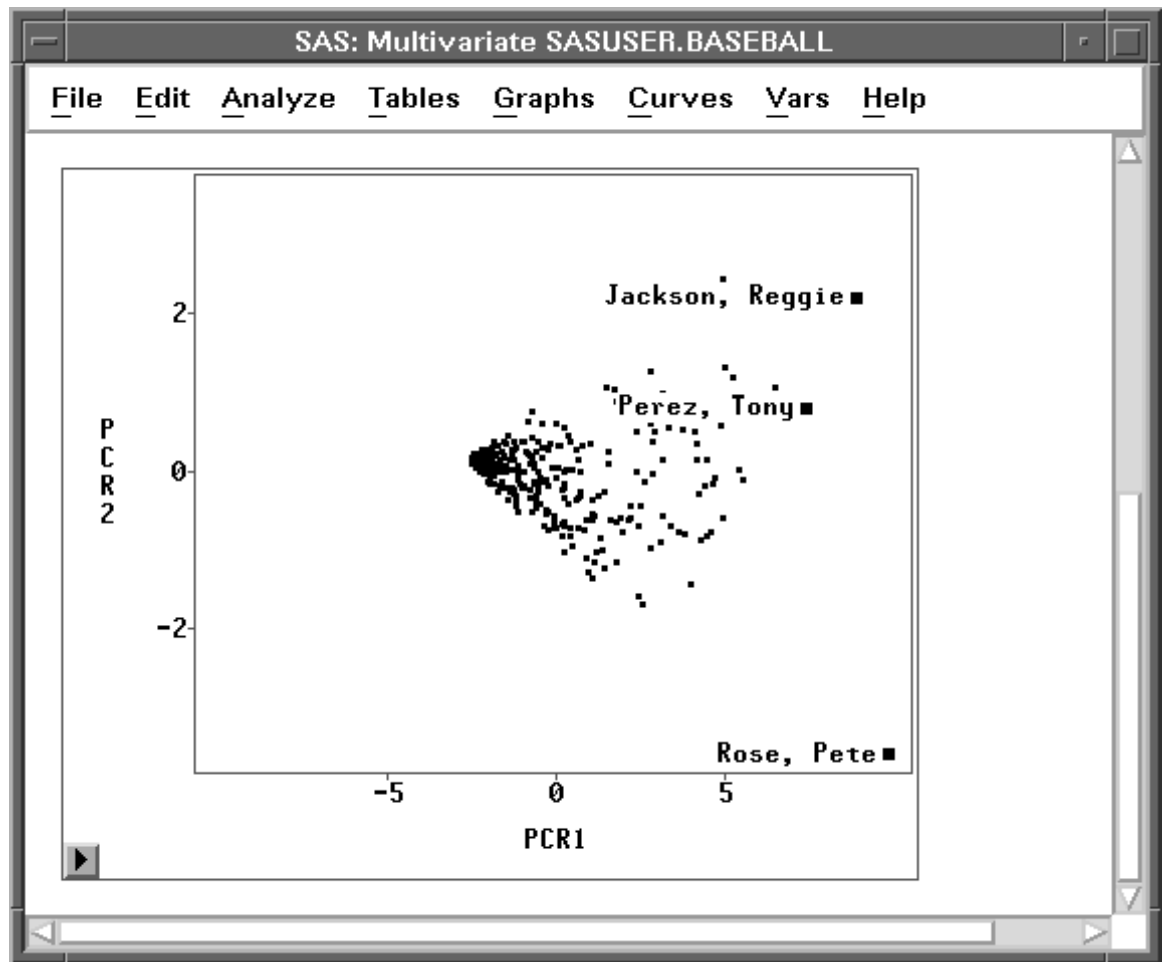


Figure 19.8. Scatter Plot of First Two Principal Components

Plotting Against Original Variables

Now that you have reduced the dimensionality of the career performance variables to two dimensions, you can easily examine scatter plots of these principal components versus the **SALARY** variable. The two principal component scores are automatically stored in the data window.

- ⇒ **Choose Analyze:Scatter Plot (Y X).**
This displays the scatter plot variables dialog.
- ⇒ **Select SALARY in the list at the left, then click the Y button.**
SALARY appears in the **Y** variables list.
- ⇒ **Select PCR1 and PCR2, then click the X button.**
PCR1 and **PCR2** appear in the **X** variables list.
- ⇒ **Select NAME in the list at the left, then click the Label button.**
NAME appears in the **LABEL** variables list.

A scatter plot variables dialog should now appear as in [Figure 19.9](#).

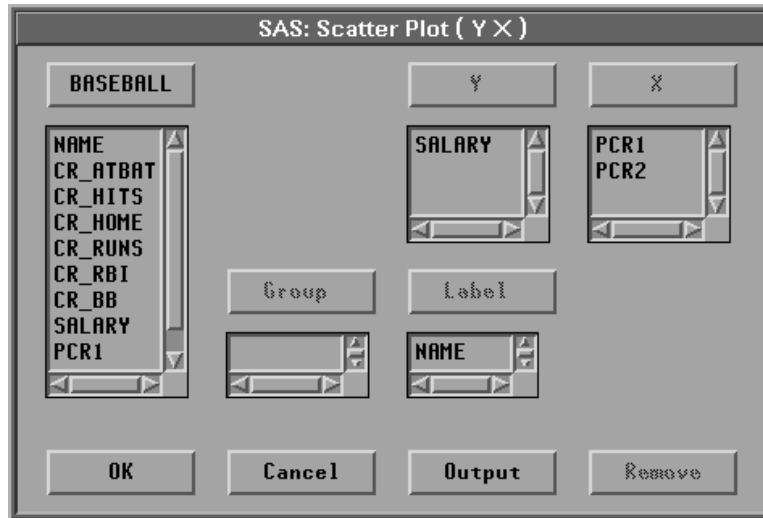


Figure 19.9. Variable Roles Assigned

⇒ Click the **OK** button.

A scatter plot window appears, as shown in [Figure 19.10](#).

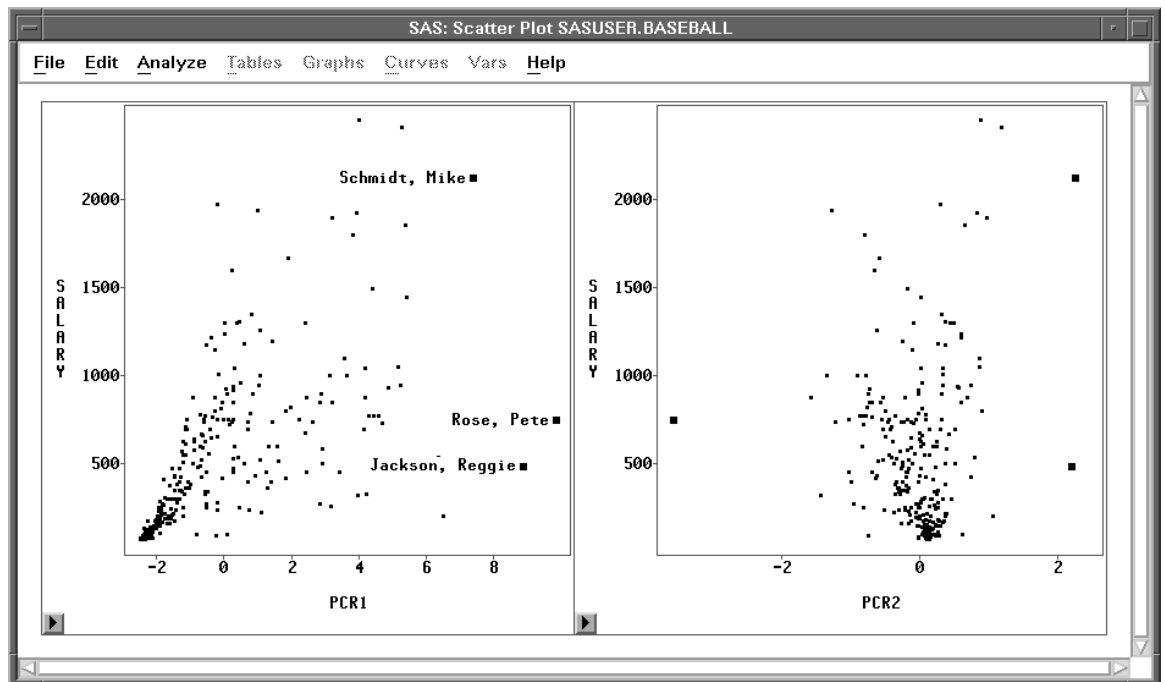


Figure 19.10. SALARY versus First Two Principal Components

Examine the scatter plot of **SALARY** versus **PCR1**, recalling that **PCR1** is highly associated with overall career performance. The linear trend evident in the plot indicates a strong linear relationship between a player's salary and his overall perfor-

mance. On the other hand, if you examine the scatter plot of **SALARY** versus **PCR2** (which is the contrast between the combined performance of career home runs and runs batted in versus the other performance), you can see that there is no evident relationship.

You can also examine these scatter plots for potential outliers. Click on the observations with large values of **PCR1** in the scatter plot of **SALARY** versus **PCR1**. These observations correspond to players who have had outstanding careers.

Saving Principal Components

This completes the principal component analysis. You began with a high dimensional set of data (six variables) and reduced it to two dimensions (two variables representing principal component scores) that accounted for over 95% of the variation. You were then able to plot the principal component scores against the variable of interest, **SALARY**.

At this point, you may want to save the principal component scores for use in subsequent analyses.

⇒ **Choose Vars:Principal Components:2.**

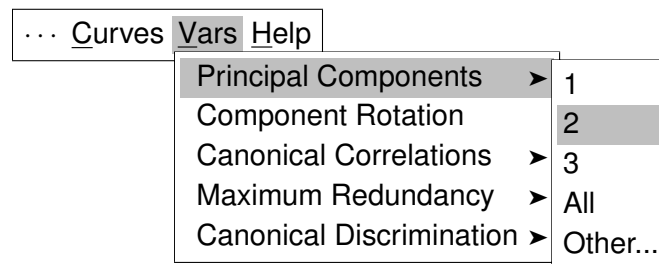


Figure 19.11. Vars Menu

This causes the two variables, **PCR1** and **PCR2**, to be retained in the data window even after you delete the multivariate window. You can then include these variables in later analyses.

⊕ **Related Reading:** Principal Components, [Chapter 40](#).

Chapter 20

Transforming Variables

Chapter Contents

COMMON TRANSFORMATIONS	320
OTHER TRANSFORMATIONS	329
REFERENCES	335

Chapter 20

Transforming Variables

A *transformation* generates a new variable from existing variables according to a mathematical formula. SAS/INSIGHT software provides a variety of variable transformations. The most commonly used transformations are available from the **Edit:Variables** menu. You can perform other more complex transformations using the Edit Variables dialog.

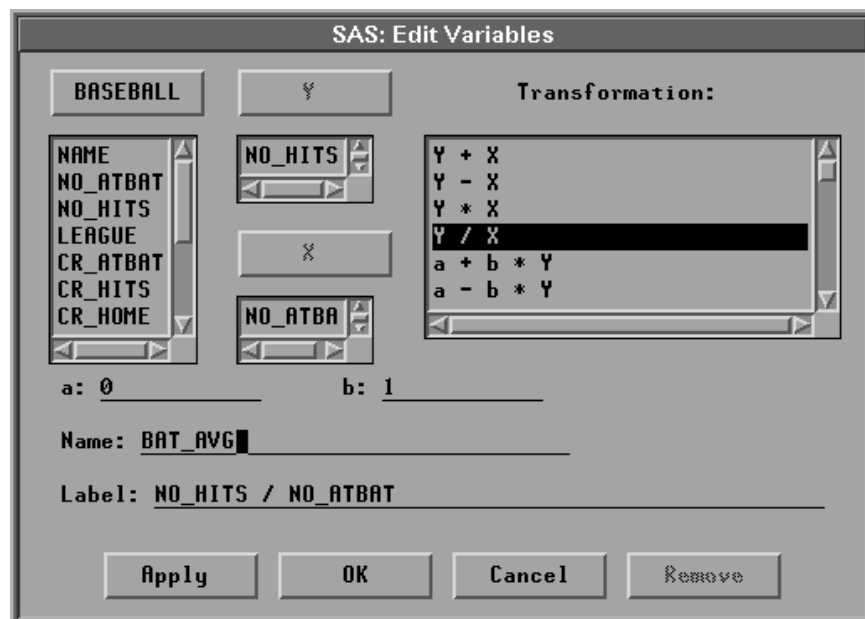


Figure 20.1. Edit Variables Dialog

Common Transformations

The most common transformations are available in the **Edit:Variables** menu. For example, log transformations are commonly used to linearize relationships, stabilize variances, or reduce skewness. Perform a log transformation in a fit window by following these steps:

- ⇒ Open the **BASEBALL** data set.
- ⇒ Create a fit analysis of **SALARY** versus **CR_HOME**.

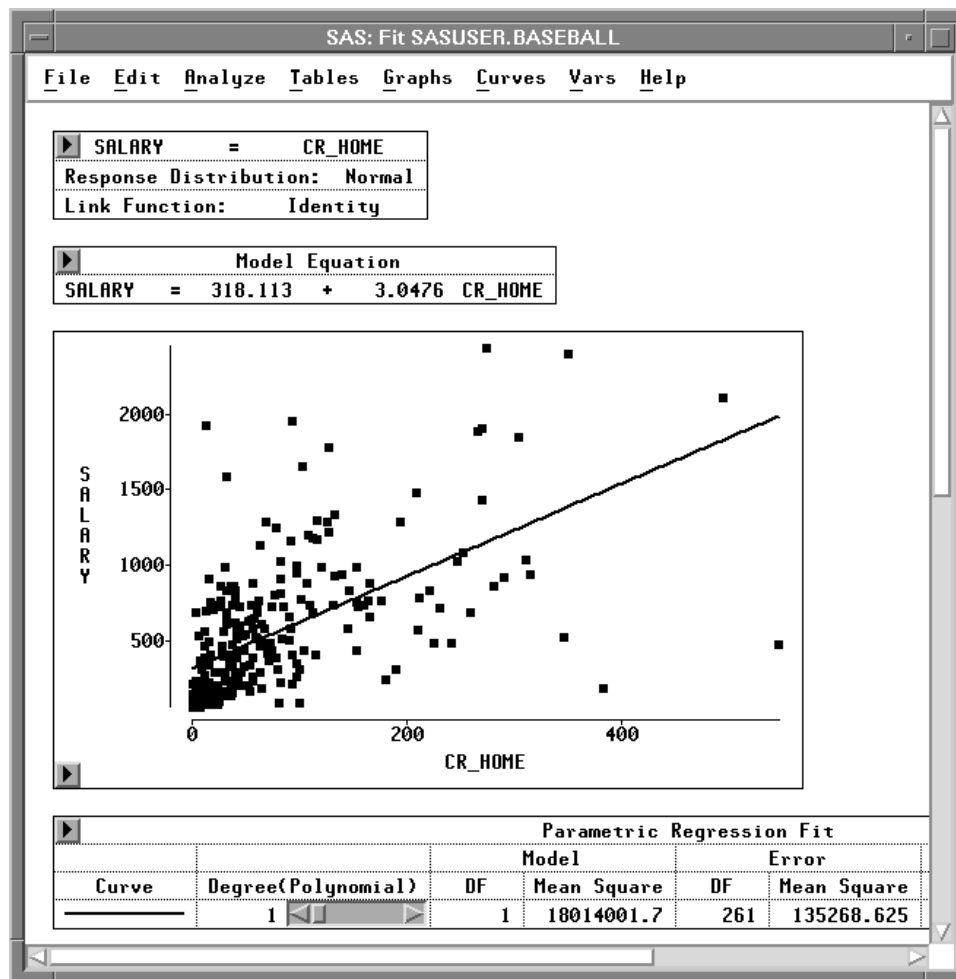


Figure 20.2. Fit Analysis of **SALARY** versus **CR_HOME**

You might expect players who hit many home runs to receive high salaries. However, most players do not hit many home runs, and most do not have high salaries. This obscures the relationship between **SALARY** and **CR_HOME**. Most of the observations appear in the lower left corner of the scatter plot, and the regression line does not fit the data well. To make the relationship clearer, apply a logarithmic transformation.

- ⇒ **Select both variables in the scatter plot.**
 Use your host's method for noncontiguous selection.

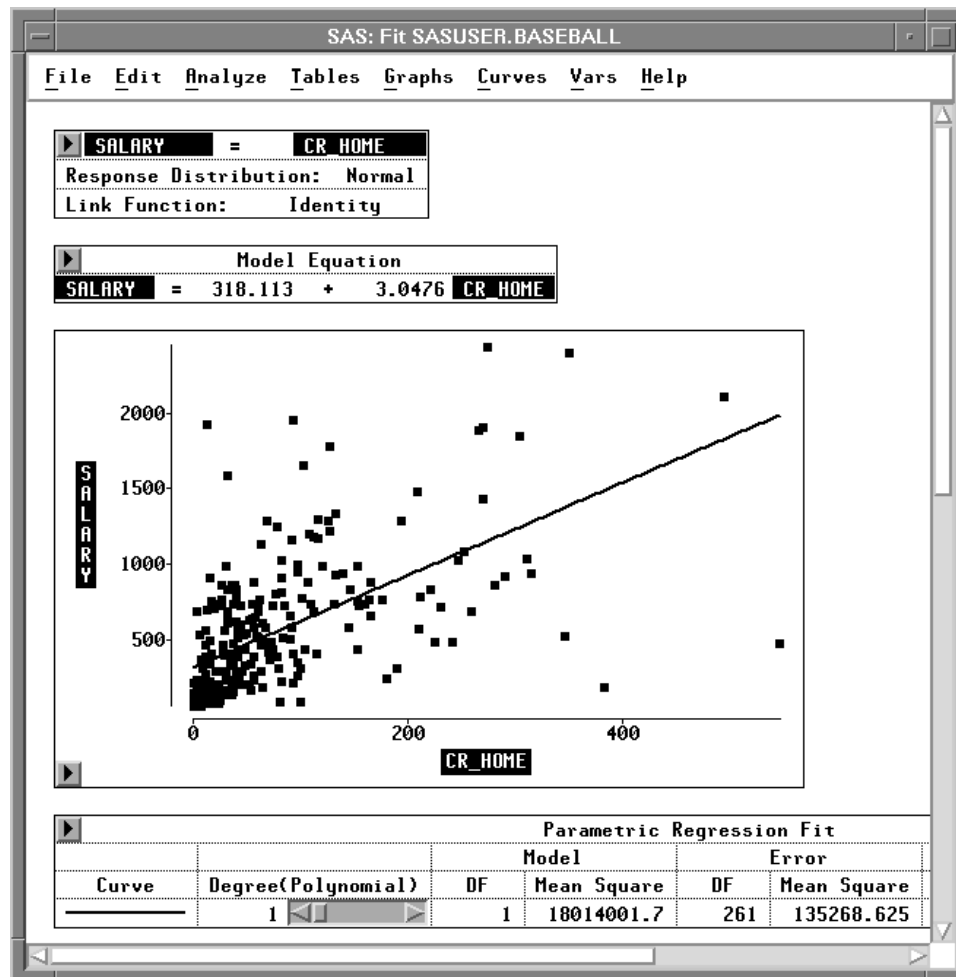


Figure 20.3. SALARY and CR_HOME Selected

- ⇒ Choose **Edit:Variables:log(Y)**.

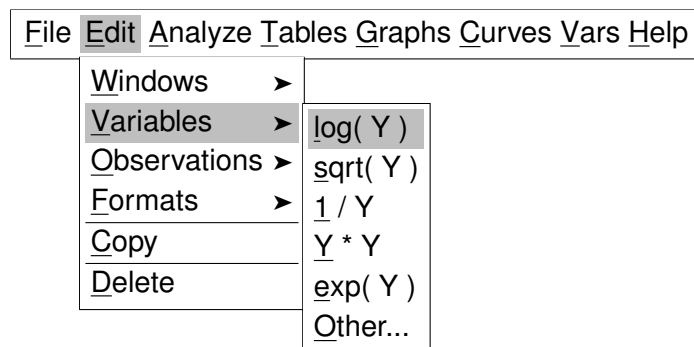


Figure 20.4. Edit:Variables Menu

This performs a log transformation on both **SALARY** and **CR_HOME** and transforms the scatter plot to a log-log plot. Now the regression fit is improved, and the relationship between salary and home run production is clearer.

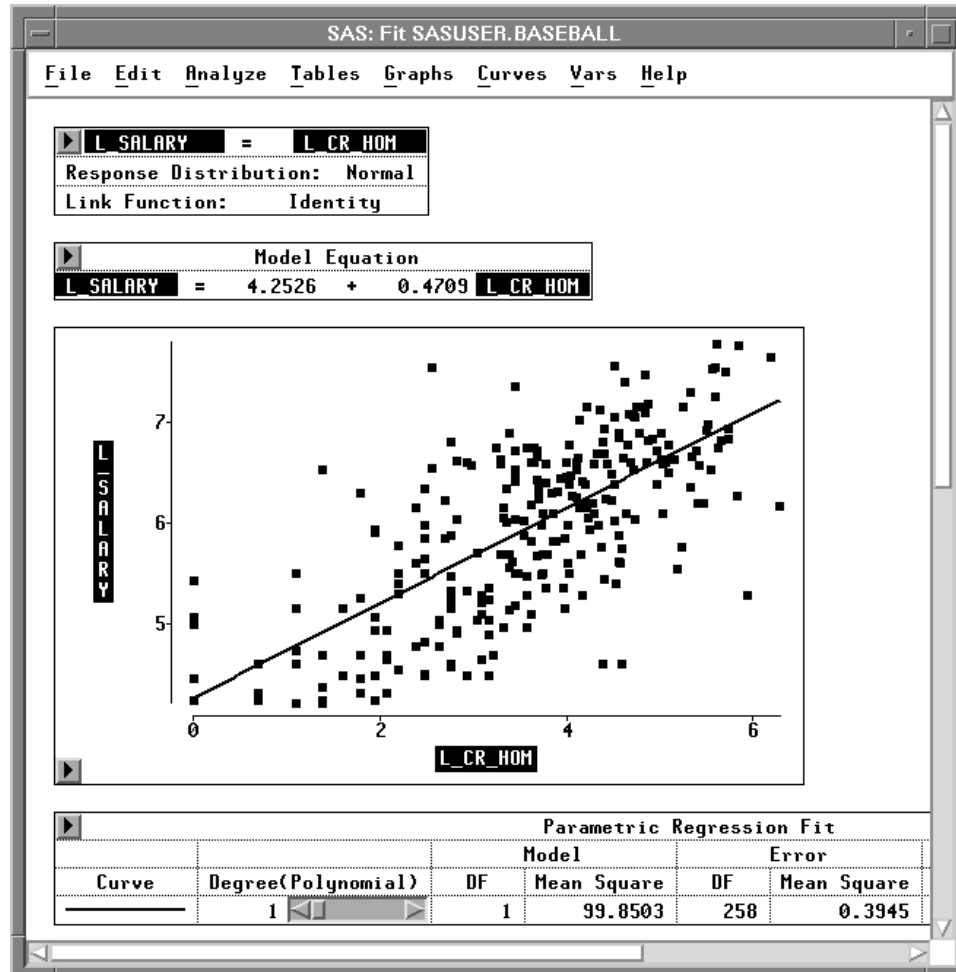


Figure 20.5. Fit Analysis of **L_SALARY** versus **L_CR_HOM**

The degrees of freedom (**DF**) is reduced from 261 to 258. This is due to missing values resulting from the log transformation, described in the following step.

⇒ **Scroll the data window to display the last four variables.**

Notice that in addition to residual and predicted values from the regression, the log transformations created two new variables: **L_SALARY** and **L_CR_HOM**.

	Int	Int	Int	Int	Int
	SALARY	L_SALARY	L_CR_HOM	R_L_SALA	P_L_SALA
1	75.000	4.3175	0.6931	-0.26147	4.5790
2	.	.	0.0000	.	4.2526
3	240.000	5.4806	3.5835	-0.45940	5.9400
4	225.000	5.4161	2.1972	0.12887	5.2872
5	.	.	5.4116	.	6.8009
6	475.000	6.1633	4.2341	-0.08308	6.2464
7	550.000	6.3099	1.7918	1.21362	5.0963
8	950.000	6.8565	4.9416	0.27688	6.5796
9	.	.	5.4889	.	6.8373
10	100.000	4.6052	4.6052	-1.81596	6.4211
11	305.000	5.7203	3.0445	0.03409	5.6862
12	1237.500	7.1208	4.8520	0.58347	6.5374

Figure 20.6. New Variables

The log transformation is useful in many cases. However, the result of $\log(Y)$ is undefined where Y is less than or equal to 0. In such cases, SAS/INSIGHT software cannot transform the value, so a missing value (.) is generated. To see this, sort the data in the data window.

⇒ Select **L_CR_HOM** in the data window, and choose **Sort** from the data pop-up menu.

	Int	Int	Int	Int	Int
	SALARY	L_SALARY	L_CR_HOM	R_L_SALA	P_L_SALA
1	75.000	4.3175	.	.	.
2	130.000	4.8675	.	.	.
3
4	100.000	4.6052	.	.	.
5	150.000	5.0106	0.0000	0.75808	4.2526
6	160.000	5.0752	0.0000	0.82262	4.2526
7	230.000	5.4381	0.0000	1.18552	4.2526
8	87.500	4.4716	0.0000	0.21908	4.2526
9	70.000	4.2485	0.0000	-0.00406	4.2526
10	.	.	0.0000	.	4.2526
11	.	.	0.0000	.	4.2526
12	75.000	4.3175	0.6931	-0.26147	4.5790

Figure 20.7. Missing Values in Log Transformation

Missing values in the SAS System are considered to be less than any other value, so they appear first in the sorted variable. These values represent players who have never hit home runs. Their value for **CR_HOME** is 0, so the log of this value cannot be calculated. This means the log transformation has removed data from the fit analysis.

The following steps circumvent this problem.

⇒ Select **CR_HOME** in the data window.

SAS: SASUSER.BASEBALL										
File		Edit	Analyze	Tables	Graphs	Curves	Vars	Help		
▶	26		Int	Int	Int	Int	Int	Int	Int	Int
322		NO_RUNS	NO_RBI	NO_BB	YR_MAJOR	CR_ATBAT	CR_HITS	CR_HOME	CR	
■	1	31	26	30	1	279	64	0		
■	2	33	31	26	5	354	82	0		
■	3	20	13	17	1	166	34	0		
■	4	24	24	7	3	509	108	0		
■	5	44	36	65	4	711	148	1		
■	6	94	29	60	2	1236	309	1		
■	7	27	15	11	4	1115	270	1		
■	8	46	24	29	4	618	129	1		
■	9	23	8	21	2	214	42	1		
■	10	30	29	14	1	293	66	1		
■	11	34	12	14	1	241	61	1		
■	12	13	9	16	3	196	44	2		

Figure 20.8. CR_HOME Selected

⇒ Choose **Edit:Variables:Other**.

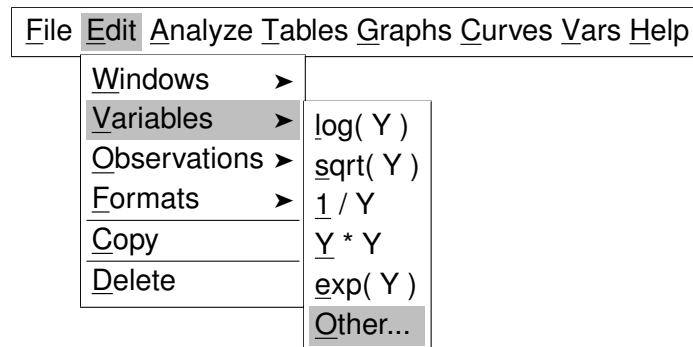


Figure 20.9. Edit:Variables Menu

This displays the Edit Variables dialog shown in [Figure 20.10](#). In the dialog you can see that the variable **CR_HOME** is already assigned as the **Y** variable.

⇒ Scroll down the transformation window, and select $\log(Y + a)$.

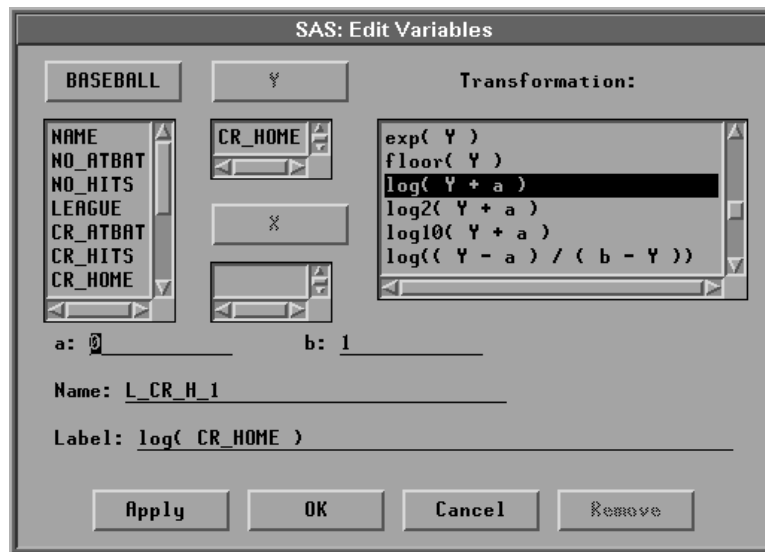


Figure 20.10. Edit Variables Dialog

- ⇒ In the field for **a** enter the value **1**, then press the Return key.
 Notice that the **Label** value changes from **log(CR_HOME)** to **log(CR_HOME + 1)** to reflect the new value of **a**. Setting **a** to **1** avoids the problem of generating missing values because **(CR_HOME + 1)** is greater than zero in all cases for this data.

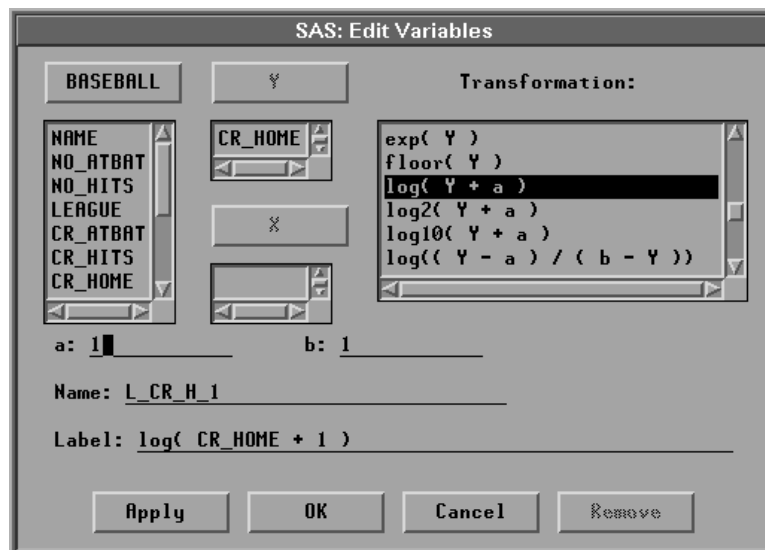


Figure 20.11. Edit Variables Dialog

- ⇒ Click **OK** to perform the transformation.
 ⇒ Scroll all the way to the right to see the new variable, **L_CR_H_1**.
 Notice that the new variable contains no missing values.

	Int	Int	Int	Int	Int
	L_SALARY	L_CR_HOM	R_L_SALA	P_L_SALA	L_CR_H_1
1	4.3175	.	.	.	0.0000
2	4.8675	.	.	.	0.0000
3	0.0000
4	4.6052	.	.	.	0.0000
5	5.0106	0.0000	0.75808	4.2526	0.6931
6	5.0752	0.0000	0.82262	4.2526	0.6931
7	5.4381	0.0000	1.18552	4.2526	0.6931
8	4.4716	0.0000	0.21908	4.2526	0.6931
9	4.2485	0.0000	-0.00406	4.2526	0.6931
10	.	0.0000	.	4.2526	0.6931
11	.	0.0000	.	4.2526	0.6931
12	4.3175	0.6931	-0.26147	4.5790	1.0986

Figure 20.12. New Variable

⇒ Select **L_SALARY** and **L_CR_H_1**, then choose **Analyze:Fit (Y X)**.

At the lower left corner of the scatter plot, you can see observations that were not used in the previous fit analysis. Also note that the degrees of freedom (**DF**) is back to 261.

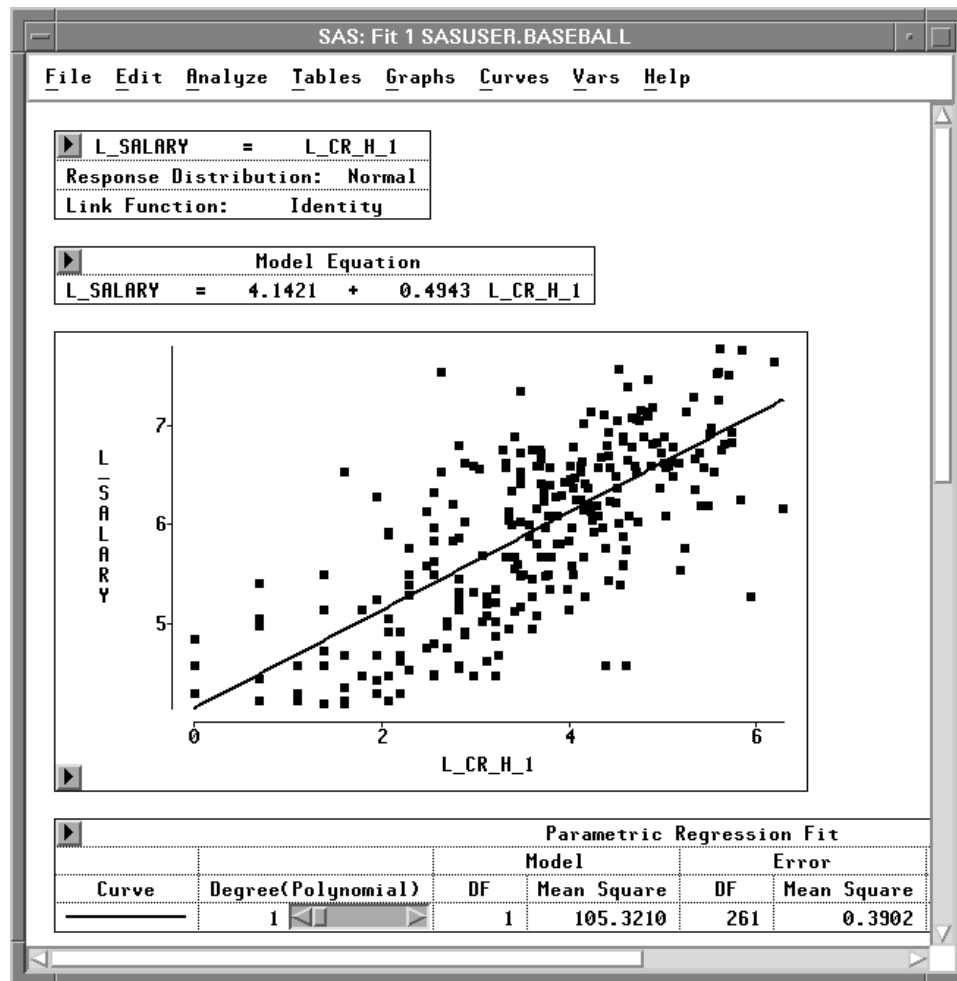


Figure 20.13. New Fit Analysis

⊕ **Related Reading:** Linear Models, [Chapter 39](#).

Other Transformations

You can use the Edit Variables dialog to create other types of transformations. Most transformations require one selected variable, as in the previous example. Here is an example using two variables. Suppose you are interested in batting averages, that is, the number of hits per batting opportunity. Calculate batting averages by following these steps.

- ⇒ Choose **Edit:Variables:Other** to display the Edit Variables dialog
- ⇒ Assign **NO_HITS** the **Y** role and **NO_ATBAT** the **X** role.



Figure 20.14. Edit Variables Dialog

- ⇒ Click on the **Y / X** transformation.
Notice that the **Label** value is now **NO_HITS / NO_ATBAT**. You might want to enter a more mnemonic value for **Name**.
- ⇒ Enter **BAT_AVG** in the **Name** field.

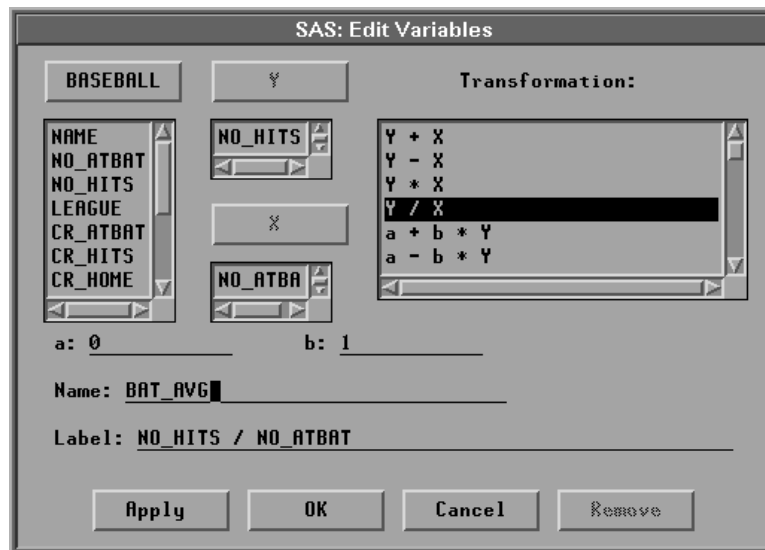


Figure 20.15. Creating the Transformation

⇒ Click the **OK** button to calculate the batting average.
The new **BAT_AVG** variable appears at the last position in the data window.

The screenshot shows the SAS data window titled "SAS: SASUSER.BASEBALL". The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. The data table has 12 rows and 6 columns. The columns are: an unlabeled column with values 23 and 322, and five columns labeled NO_OUTS, NO_ASSTS, NO_ERROR, SALARY, and BAT_AVG. The BAT_AVG column contains values ranging from 0.2253 to 0.2886.

	Int	Int	Int	Int	Int
	NO_OUTS	NO_ASSTS	NO_ERROR	SALARY	BAT_AVG
1	317	36	1	75.000	0.2500
2	446	33	20	.	0.2253
3	80	45	8	240.000	0.2194
4	73	152	11	225.000	0.2454
5	247	4	8	.	0.2635
6	632	43	10	475.000	0.2571
7	186	290	17	550.000	0.3204
8	295	15	5	950.000	0.2965
9	90	4	0	.	0.2397
10	1236	98	18	100.000	0.2285
11	359	30	4	305.000	0.2677
12	368	20	3	1237.500	0.2886

Figure 20.16. New **BAT_AVG** Variable

Now look at the distribution of batting averages for each league by creating a box plot.

⇒ Choose **Analyze:Box Plot/Mosaic Plot (Y)**.
Specify **BAT_AVG** as the **Y** variable, **LEAGUE** as the **X** variable, and **NAME** for the **Label** role in the box plot variables dialog. Then click on **OK**.

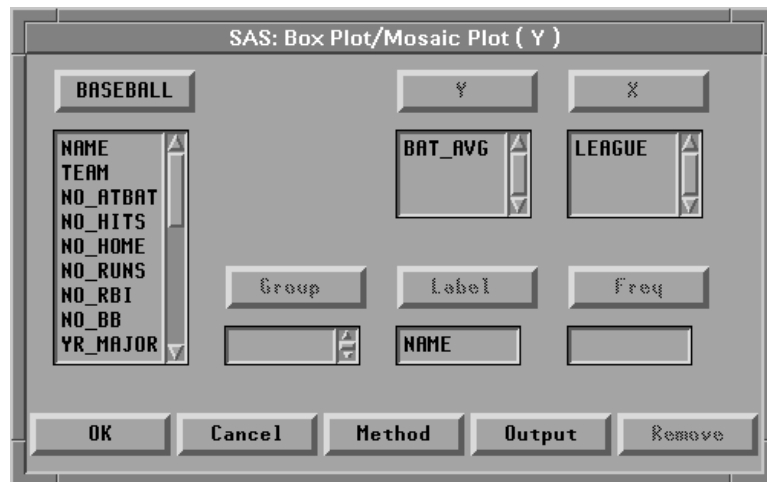


Figure 20.17. Box Plot Dialog

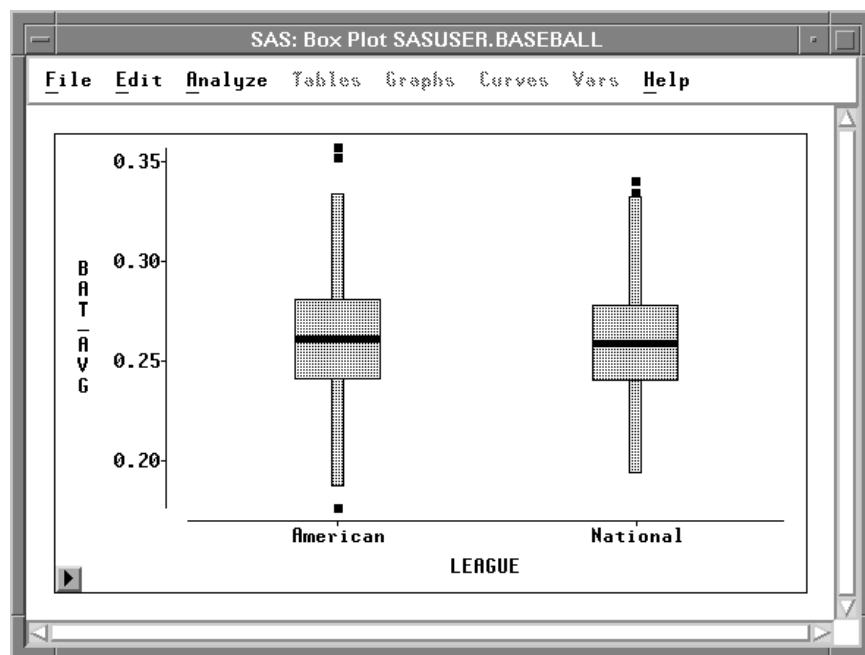


Figure 20.18. Box Plot of Batting Averages

Most players are batting between .200 and .300. There are, however, a few extreme observations.

⇒ **Select the upper extreme observations for each league.**

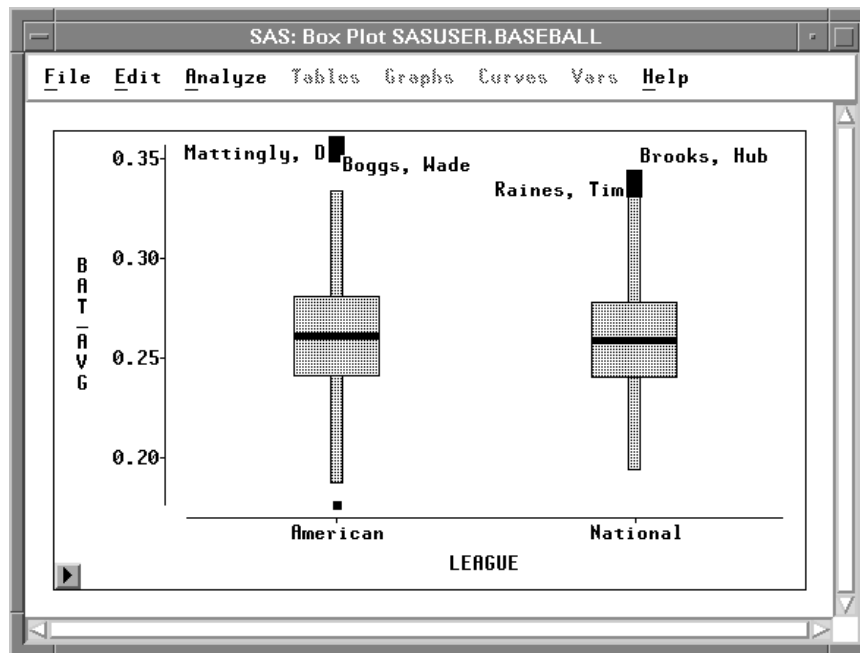


Figure 20.19. Examining the Extreme Observations

Don Mattingly and Wade Boggs led the American League in batting, while Tim Raines and Hubie Brooks led the National League.

The **Edit:Variables** menu and dialog offer many other transformations. Here is the complete list of transformations in the **Edit:Variables** menu:

log(Y)	calculates the natural logarithm of the Y variable.
sqrt(Y)	calculates the square root of the Y variable.
1 / Y	calculates the reciprocal of the Y variable.
Y * Y	calculates the square of the Y variable.
exp(Y)	raises e (2.718...) to the power given by the Y variable.

Here is the complete list of transformations in the **Edit:Variables** dialog:

Y + X	These four transformations perform addition, subtraction, multiplication, and division on the specified Y and X variables.
Y - X	
Y * X	
Y / X	
a + b * Y	These four transformations create linear transformations of the Y variable. Using the default values a =0 and b =1, the second and third transformations create additive and multiplicative inverses -Y and 1 / Y .
a - b * Y	
a + b / Y	
a - b / Y	

$Y ** b$	is the power transform. b can be positive or negative.
$((Y + a) ** b - 1) / b$	is the Box-Cox transformation. This transformation raises the sum of the Y variable plus a to the power b , then subtracts 1 and divides by b .
$a <= Y <= b$	creates a variable with value 1 when the value of Y is between a and b inclusively, and value 0 for all other values of Y . Values for a and b can be character or numeric; character values should not be in quotations. You can use this transformation to create indicator variables for subsetting your data.
$(Y - \text{mean}(Y)) / \text{std}(Y)$	standardizes the Y variable by subtracting its mean and dividing by its standard deviation. Standardizing changes the mean of the variable to 0 and its standard deviation to 1.
$\text{abs}(Y)$	calculates the absolute value of Y .
$\text{arccos}(Y)$	calculates the arccosine (inverse cosine) of Y . The value is returned in radians.
$\text{arcsin}(Y)$	calculates the arcsine (inverse sine) of Y . The value is returned in radians.
$\text{arcsin}(\text{sqrt}(Y))$	calculates the arcsine of the square root of Y . The value is returned in radians.
$\text{arctan}(Y)$	calculates the arctangent (inverse tangent) of Y . The value is returned in radians.
$\text{ceil}(Y)$	calculates the smallest integer greater than or equal to Y .
$\text{cos}(Y)$	calculates the cosine of Y .
$\text{exp}(Y)$	raises <i>e</i> (2.718...) to the power given by the Y variable.
$\text{floor}(Y)$	calculates the largest integer less than or equal to Y .
$\text{log}(Y + a)$	calculates the natural logarithm of the Y variable plus an offset a .
$\text{log2}(Y + a)$	calculates the logarithm base 2 of the Y variable plus an offset a .
$\text{log10}(Y + a)$	calculates the logarithm base 10 of the Y variable plus an offset a .
$\text{log}((Y - a) / (b - Y))$	calculates the natural logarithm of the quotient of the Y variable minus a divided by b minus the Y variable. When a = 0 and b = 1, this is a logit transformation.

ranbin(a, b)	generates a binomial random variable containing values either 0 or 1. a is the seed value for the random transformation. b is the probability that the generated value will be 1. If a is less than or equal to 0, the time of day is used. This is a special case of the SAS function RANBIN where <i>n</i> , the number of trials, is 1.
ranexp(a)	generates a random variable from an exponential distribution. a is the seed value for the random transformation. If a is less than or equal to 0, the time of day is used.
rangam(a, b)	generates a random variable from a gamma distribution. a is the seed value for the random transformation, and b is the shape parameter. If a is less than or equal to 0, the time of day is used.
rannor(a)	generates a random variable from a normal distribution with mean 0 and variance 1. a is the seed value for the random transformation. If a is less than or equal to 0, the time of day is used.
ranpoi(a, b)	generates a random variable from a Poisson distribution. a is the seed value for the random transformation, and b is the mean parameter. If a is less than or equal to 0, the time of day is used.
ranuni(a)	generates a uniform random variable containing values between 0 and 1. a is the seed value for the random transformation. If a is less than or equal to 0, the time of day is used.
round(Y)	calculates the nearest integer to Y .
sin(Y)	calculates the sine of Y .
sqrt(Y + a)	calculates the square root of the Y variable plus an offset a .
tan(Y)	calculates the tangent of Y .

If your work requires other transformations that do not appear in the **Edit:Variables** menu or in the **Edit Variables** dialog, you can perform many kinds of transformations using the SAS DATA step. For more complete descriptions of the **ranbin**, **ranexp**, **rangam**, **rannor**, **ranpoi**, and **ranuni** transformations and for complete information on the DATA step, refer to *SAS Language Reference: Dictionary*.

References

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth International Group.

Chapter 21

Comparing Analyses

Chapter Contents

COMPARING ANALYSES OF DIFFERENT OBSERVATIONS	340
Extracting Observations	340
Excluding Observations	344
COMPARING ANALYSES OF DIFFERENT VARIABLES	349
Deleting Variables	349
Transforming Variables	352

Chapter 21

Comparing Analyses

You can compare analyses that use different observations or variables. For example, you can exclude certain observations from a model and see how that affects the fit. You can delete and transform variables to create and compare different models.

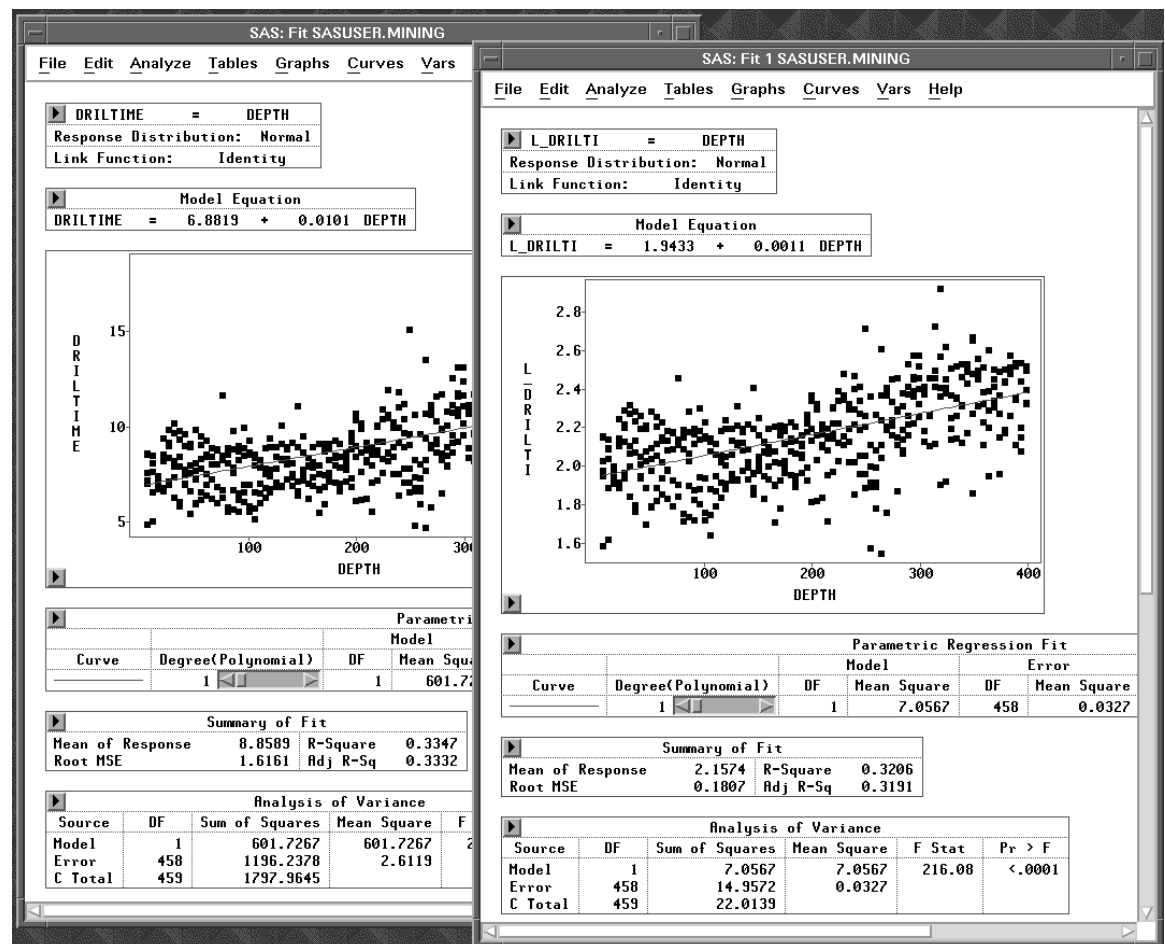


Figure 21.1. Comparing two Regression Analyses

Comparing Analyses of Different Observations

There are two ways to compare analyses that use different observations. You can *extract* observations or you can *exclude* them.

Extracting Observations

You can compare analyses made with different observations by *extracting* a subset, that is, by creating a new data set that contains a subset of observations from the original data set. Then you can request separate analyses for each data set.

Consider the **MINING** data. This data set contains results of an experiment to examine drilling times (**DRILTIME**) for different drilling methods (**METHOD**). As it turned out, the experimenters encountered difficulties due to changing rock types after a depth of about 200 feet. It might be worthwhile to compare the distribution of **DRILTIME** for depths greater than 200 feet to the distribution of **DRILTIME** for the entire data set. To compare the two distributions, you need to select the observations where **DEPTH** is greater than 200 feet and extract them into a new data window.

⇒ Open the **MINING** data set.

	4	Int	Int	Nom	Int				
480		DEPTH	DRILTIME	METHOD	REP				
1	1	5	7.61	Wet	1				
2	2	5	8.68	Wet	2				
3	3	5	8.61	Wet	3				
4	4	5	7.25	Dry	1				
5	5	5	7.07	Dry	2				
6	6	5	4.90	Dry	3				
7	7	10	8.16	Wet	1				
8	8	10	8.13	Wet	2				
9	9	10	7.71	Wet	3				
10	10	10	8.55	Dry	1				
11	11	10	6.62	Dry	2				
12	12	10	5.07	Dry	3				

Figure 21.2. MINING Data

⇒ Choose **Edit:Observations:Find**.

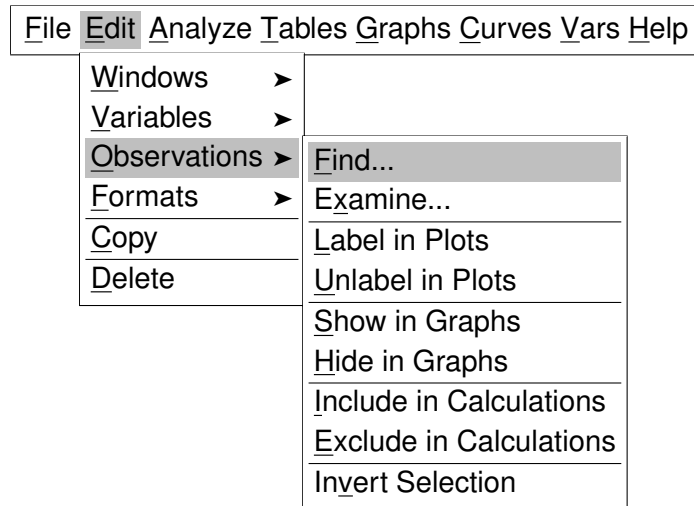


Figure 21.3. Finding Observations

This displays the **Find Observations** dialog.



Figure 21.4. Find Observations Dialog

⇒ Select > in the **Test** list and 200 in the **Value** list.



Figure 21.5. Finding **DEPTH** > 200

⇒ Click the **OK** button.

This selects all observations where **DEPTH** is greater than 200 feet. To see the selected observations, either choose **Find Next** from the data pop-up menu or scroll down using the vertical scroll bar on the right (as indicated by the arrow in the figure).

	Int	Int	Nom	Int
	DEPTH	DRILTIME	METHOD	REP
238	200	10.42	Dry	1
239	200	10.76	Dry	2
240	200	6.15	Dry	3
241	205	9.97	Wet	1
242	205	8.71	Wet	2
243	205	10.19	Wet	3
244	205	7.67	Dry	1
245	205	10.24	Dry	2
246	205	6.19	Dry	3
247	210	8.19	Wet	1
248	210	8.80	Wet	2
249	210	8.95	Wet	3

Figure 21.6. Observations where **DEPTH** > 200

⇒ Choose **Extract** from the data pop-up menu.

A new data set containing observations where **DEPTH** is greater than 200 feet appears, as shown in [Figure 21.7](#). The new data window is named automatically by adding a subscript to the original name. You may have to scroll to the top of the data window to duplicate the next figure.

	4	Int	Int	Nom	Int
240	DEPTH	DRILTIME	METHOD	REP	
1	205	9.97	Wet	1	
2	205	8.71	Wet	2	
3	205	10.19	Wet	3	
4	205	7.67	Dry	1	
5	205	10.24	Dry	2	
6	205	6.19	Dry	3	
7	210	8.19	Wet	1	
8	210	8.80	Wet	2	
9	210	8.95	Wet	3	
10	210	8.32	Dry	1	
11	210	9.22	Dry	2	
12	210	6.29	Dry	3	

Figure 21.7. MINING1 Data

Now create distribution analyses for both data sets.

⇒ **Select DRILTIME in the MINING data window.**

⇒ **Choose Analyze:Distribution (Y).**

A distribution analysis using all the observations appears on your display.

⇒ **Select DRILTIME in the MINING1 data window.**

⇒ **Choose Analyze:Distribution (Y).**

A distribution analysis using the subset of observations appears on your display.

⇒ **Move the two analysis windows side-by-side to compare the distributions.**

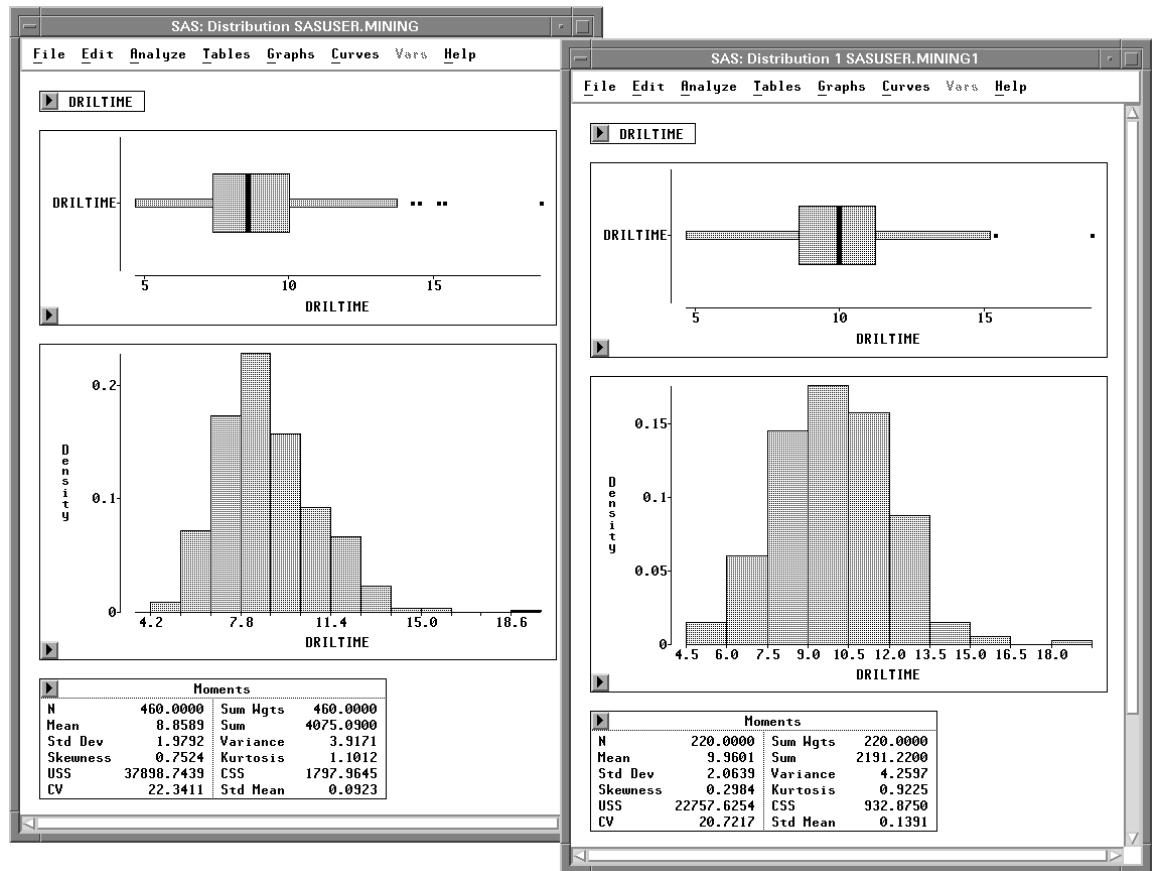


Figure 21.8. Comparing Two Distribution Analyses

The mean drilling time at depths greater than 200 feet was 9.9601, while the mean overall was only 8.8589. The drills may have found harder rock at greater depths. You may want to create an additional analysis to compare depths greater than 200 feet with depths less than or equal to 200 feet.

⇒ **Choose File:End** to delete **MINING1** and the two analysis windows.

† **Note:** Sometimes you will want to compare analyses that use different subsets of observations based on the values of some variable. If this is the case, you can assign the variable the **Group** role, as described in [Chapter 22, “Analyzing by Groups.”](#)

⊕ **Related Reading:** Distributions, [Chapter 38](#).

Excluding Observations

Another way to compare analyses using different observations is to *exclude* observations, that is, to remove them from calculations in the analysis. The observations still appear in graphs. To illustrate this technique, consider a simple linear regression model with **DRILTIME** as the response variable and **DEPTH** as the explanatory variable.

- ⇒ Select **DRILTIME**, then **DEPTH**, then choose **Analyze:Fit (Y X)**.
This displays a fit window.

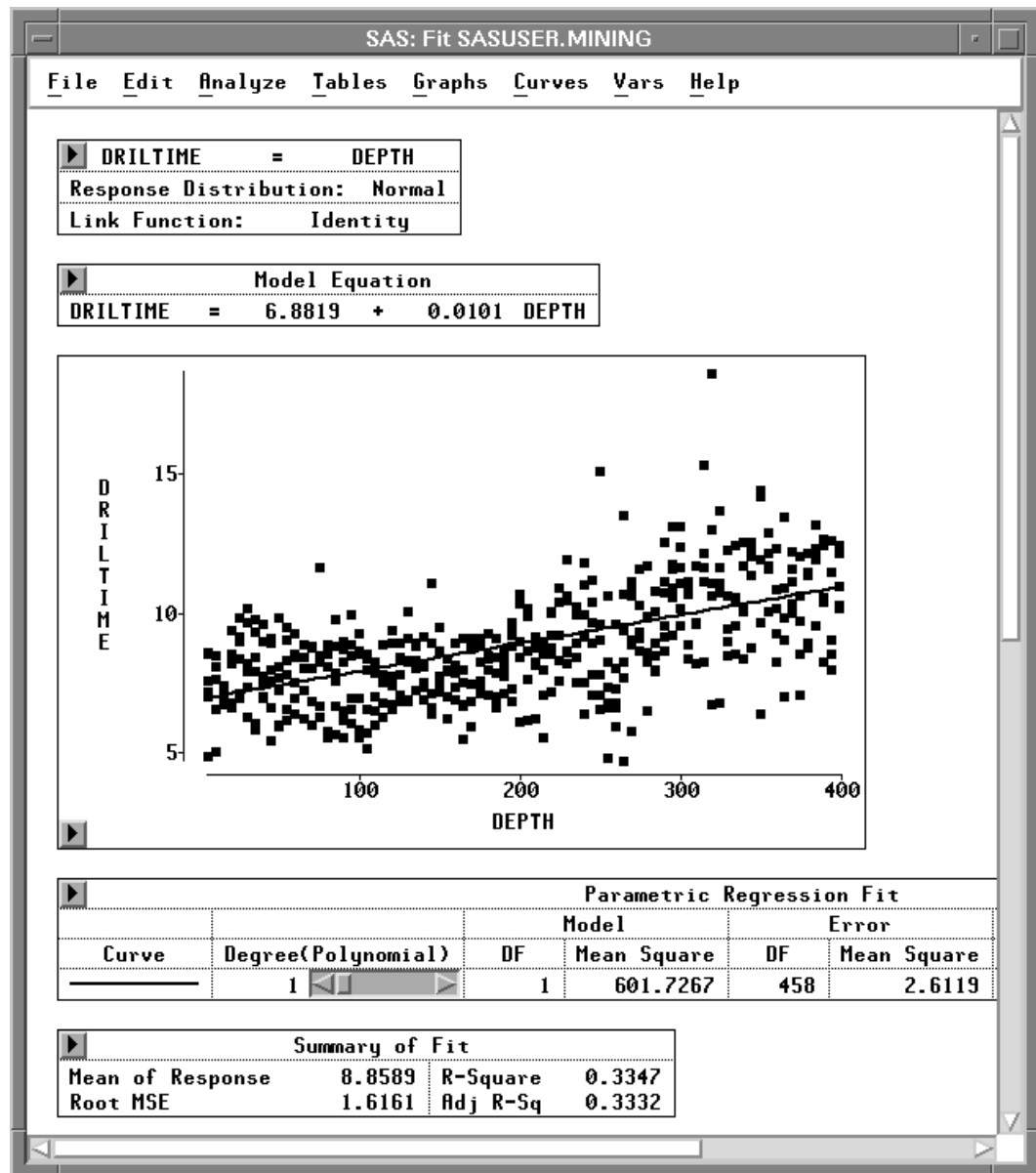


Figure 21.9. Fit Window

- ⇒ Choose **Edit:Windows:Copy Window** in the fit window.
This creates a copy of the fit window.

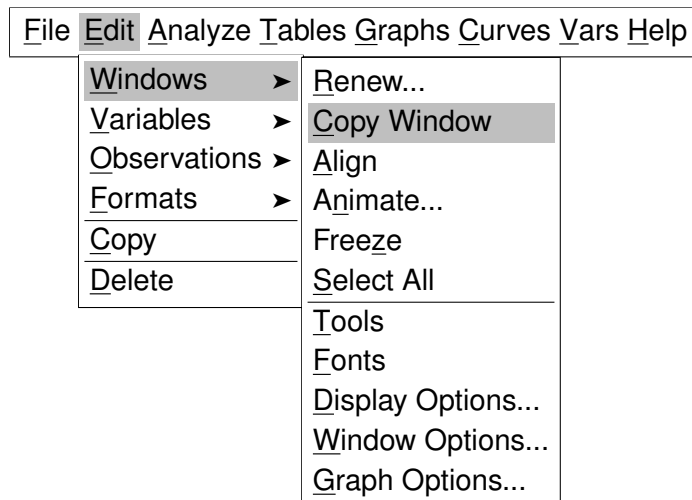


Figure 21.10. Edit:Windows Menu

⇒ Move the two fit windows side by side.

⇒ Choose **Edit:Windows:Freeze** in the fit window on the left.

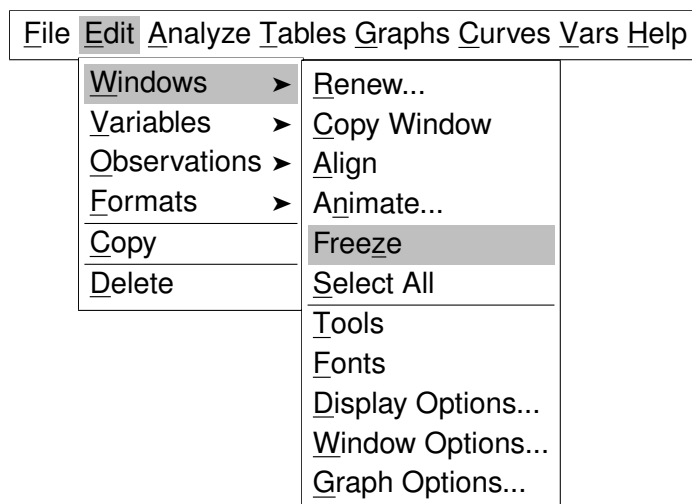


Figure 21.11. Edit:Windows Menu

This freezes the window, as indicated by the frost in the corners of the window. *Freezing* a window converts the window to a static image that ignores any changes to the data. Normally, all SAS/INSIGHT windows are linked to their data, and any changes to the data are automatically reflected in all analyses. By freezing a window, you can compare windows using different observations without creating additional data sets.

Techniques ♦ Comparing Analyses

This recalculates the fit analysis without the selected observations. Normally, both windows would be recalculated, but since the window on the left is frozen, it does not change. Now you can compare the two fit windows.

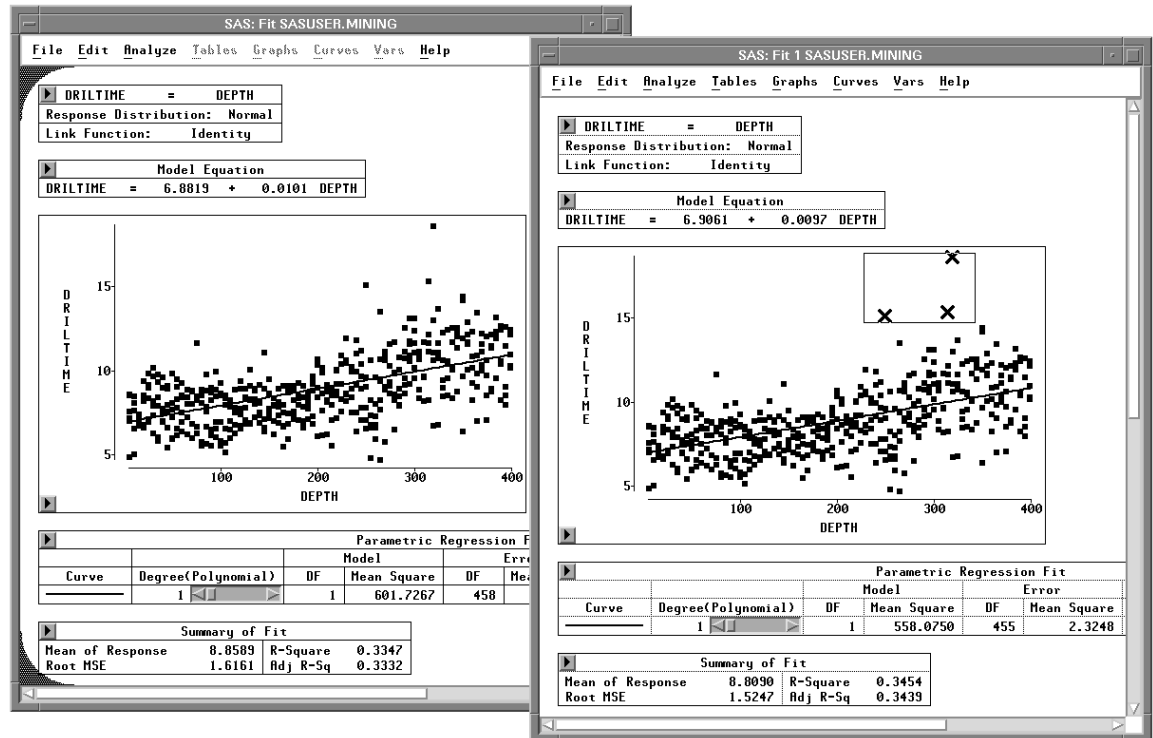


Figure 21.14. Comparing Two Fit Windows

To thaw a frozen window, follow these steps.

- ⇒ **Choose Edit:Windows:Freeze again.**
This recalculates the frozen window and restores its dynamic behavior.
- ⇒ **Close all analysis windows before proceeding to the next section.**

Comparing Analyses of Different Variables

You have already seen one easy way to compare analyses using different variables. The **Apply** button, discussed in [Chapter 14, “Multiple Regression,”](#) and [Chapter 16, “Logistic Regression,”](#), enables you to create models quickly with different effects.

In this section, you will see two additional ways to compare analyses using different variables. In any analysis, you can *delete* variables or you can *transform* them.

Deleting Variables

You can delete any effect in a fit analysis. To see this, do the following:

- ⇒ Select **DRILTIME**, then **DEPTH**, then **METHOD** in the data window.
- ⇒ Choose **Analyze:Fit (Y X)**.

A fit window appears, as shown in [Figure 21.15](#).

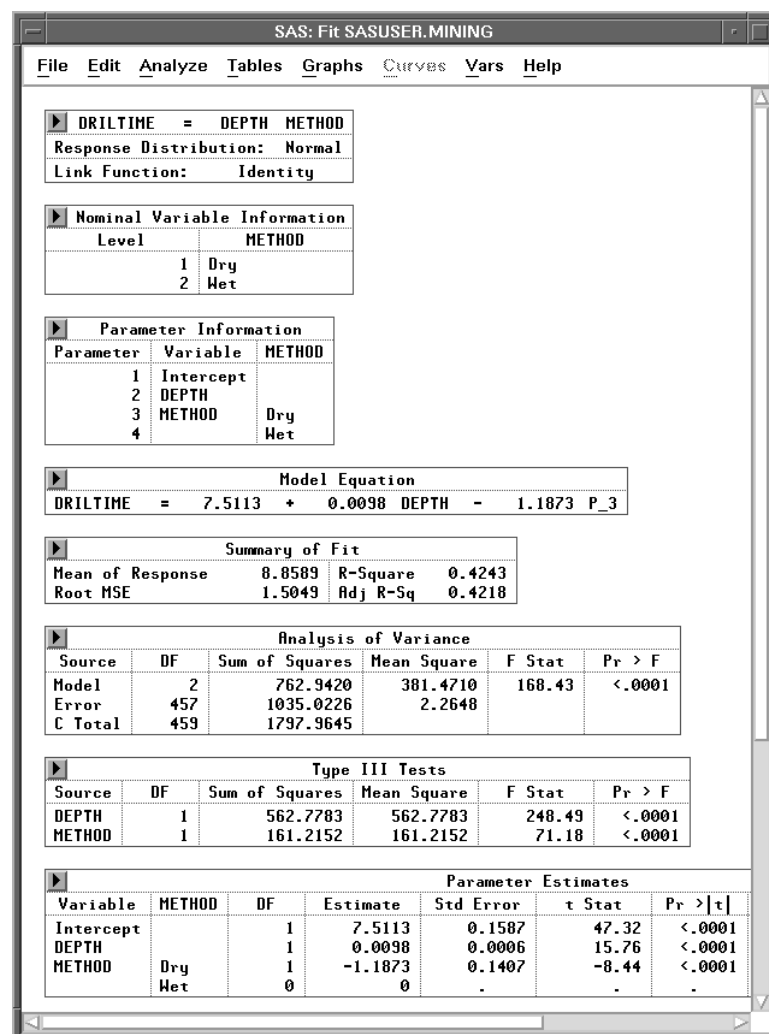


Figure 21.15. Fit Window

⇒ Choose **Edit:Windows:Copy Window**.

Now you have two identical fit windows.

⇒ Select **METHOD** in one of the fit windows.

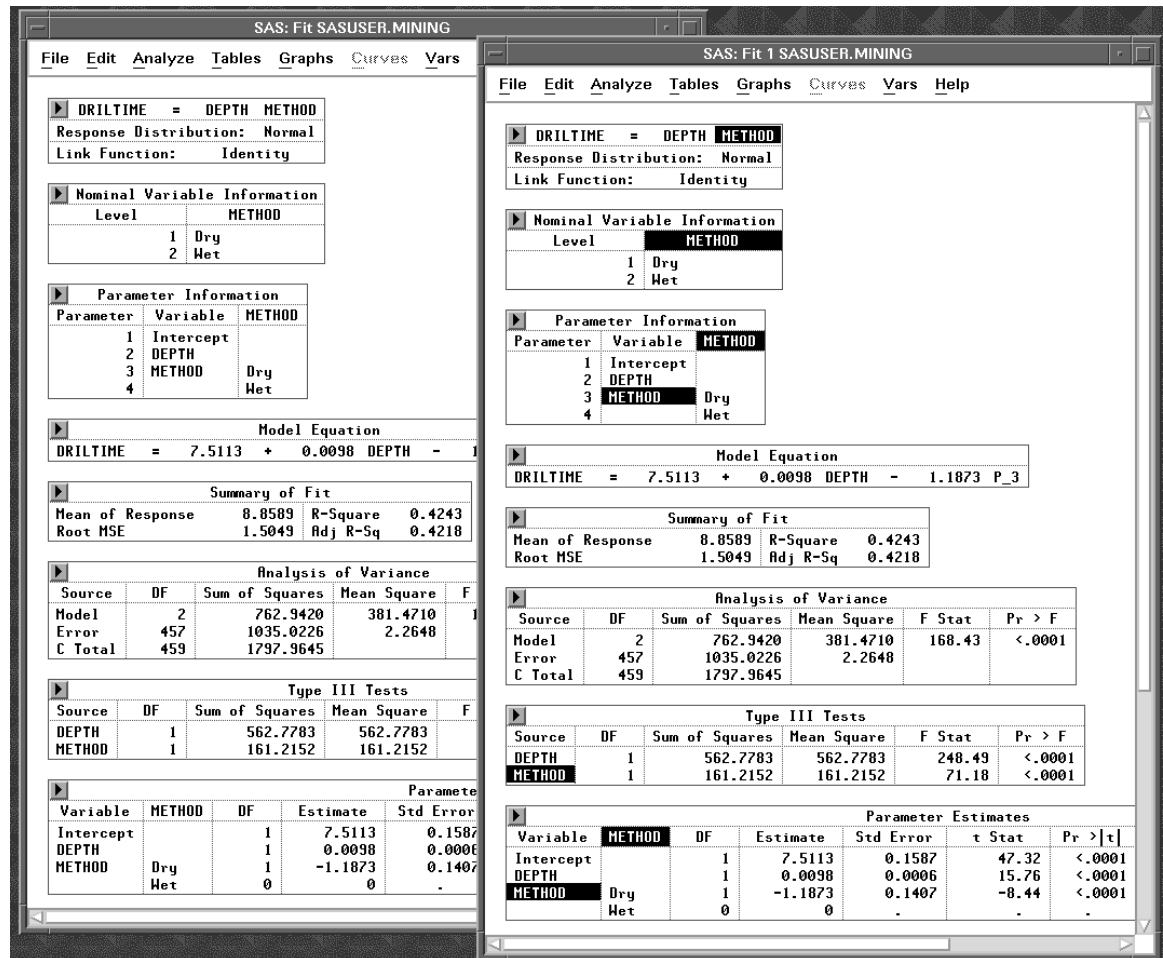


Figure 21.16. Two Fit Windows, **METHOD** Selected in One

⇒ Choose **Edit:Delete**.

This recalculates the fit window without the effect you deleted. Now you have two fit windows for two different models.

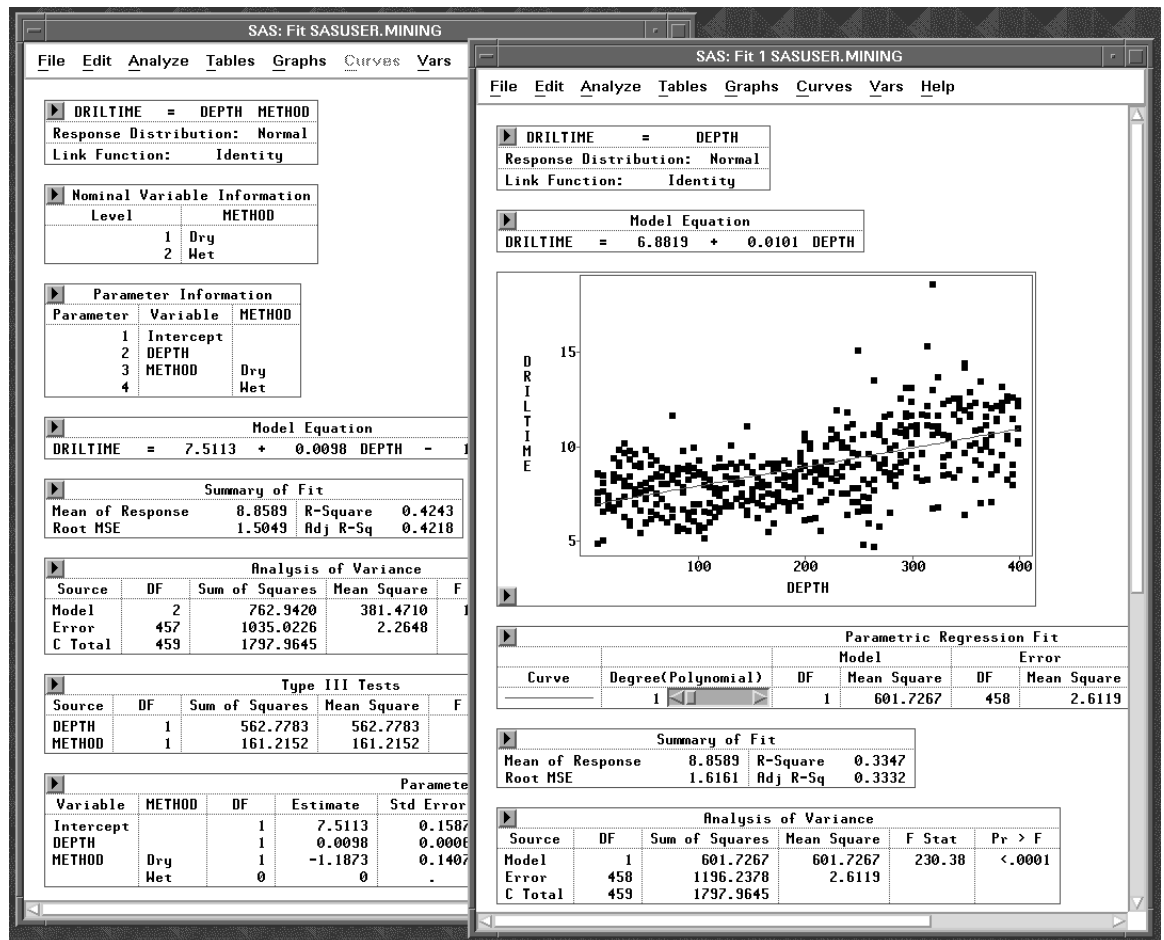


Figure 21.17. Comparing Two Models

Deleting **METHOD** caused the adjusted R-square value to drop from 0.4218 to 0.3332. It was expected that different drilling methods might produce different drilling times.

Transforming Variables

You can compare analyses by transforming variables in any window.

⇒ Create identical fit windows for **DRILTIME = DEPTH**.

Either delete **METHOD** from the first window or choose **Edit:Windows:Copy Window** in the second window.

⇒ Select **DRILTIME** in one of the fit windows.

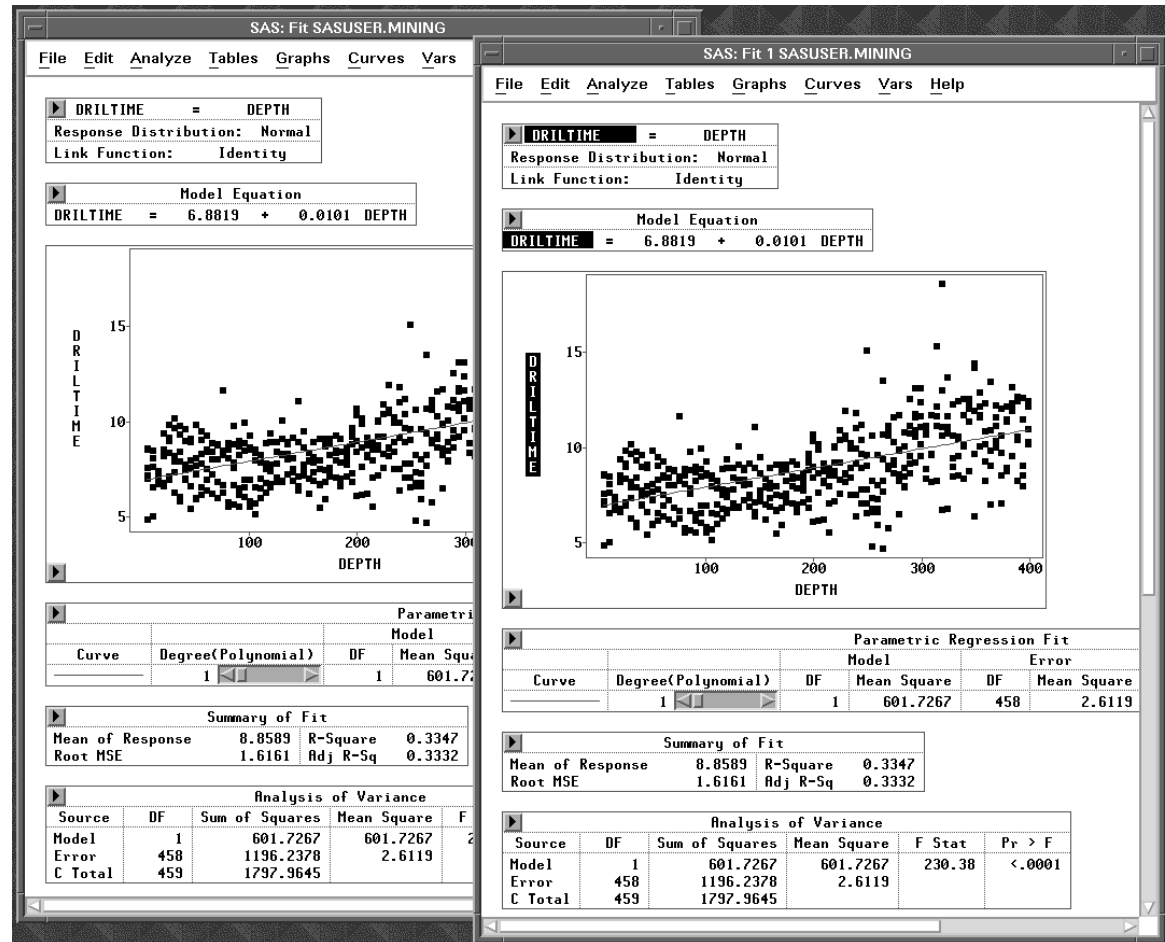


Figure 21.18. Two Fit Windows, **DRILTIME** Selected

⇒ Choose **Edit:Variables:log(Y)**.

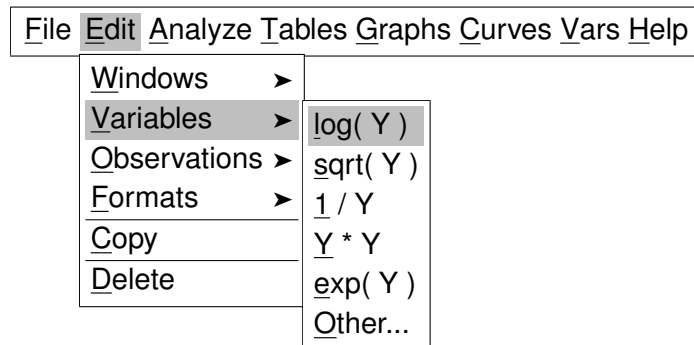


Figure 21.19. Edit:Variables Menu

This recalculates the fit window using the log of the response variable (**L_DRILTI**). Now you have two fit windows for two different models.

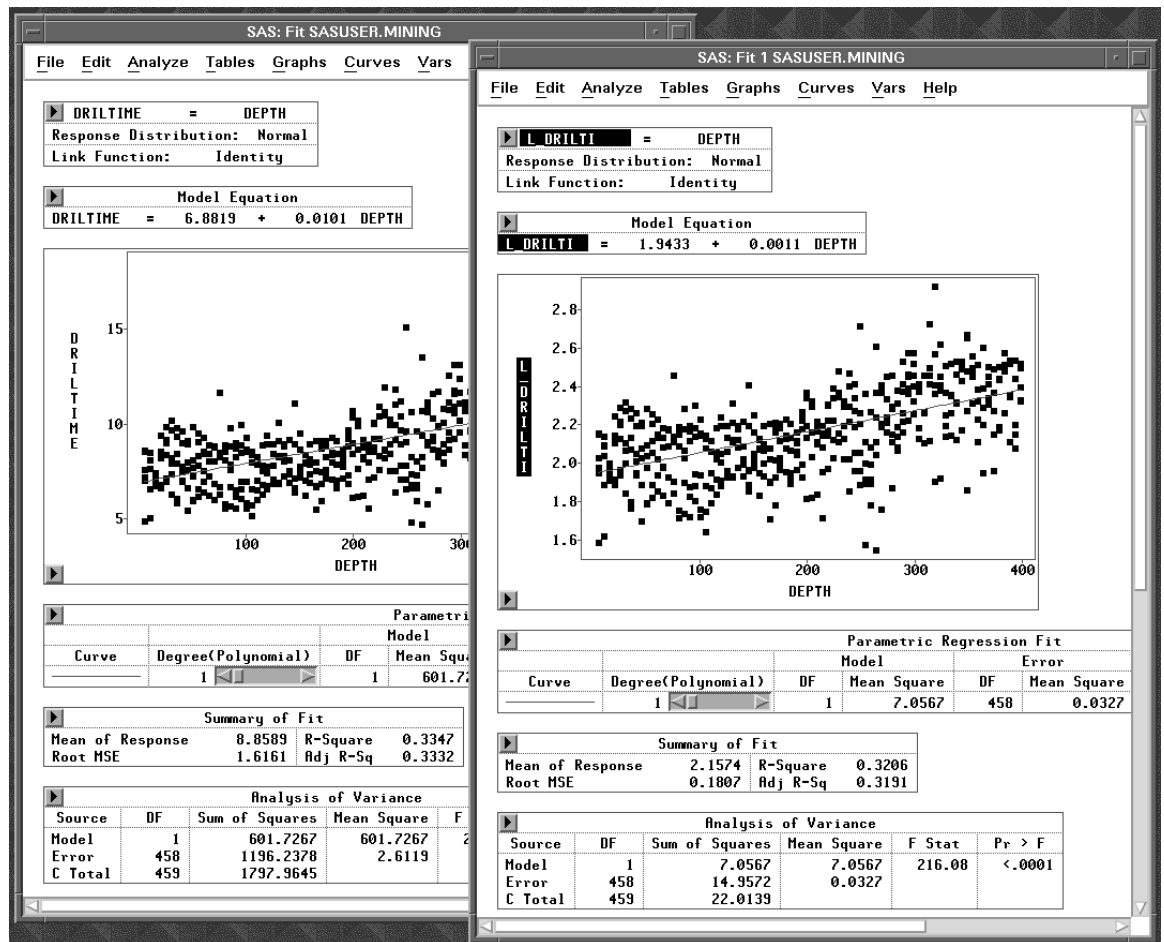


Figure 21.20. Comparing Two Fit Analyses

In this case, the log transform did not improve the fit. To undo the log transform, you can choose **Edit:Windows:Renew**.

In this chapter you have seen how to compare analysis windows that use different observations by extracting and excluding. You have also compared analyses using different variables by deleting and transforming. In the next chapter, you will see how to compare analyses using **Group** variables.

- ⊕ **Related Reading:** Transformations, [Chapter 20](#).
- ⊕ **Related Reading:** Linear Models, [Chapter 39](#).

Chapter 22

Analyzing by Groups

Chapter Contents

USING GROUP VARIABLES	358
COMPARING GROUPS BY COPYING WINDOWS	360
SETTING DEFAULT GROUP VARIABLES	363
FORMATTING GROUP VARIABLES	366

Chapter 22

Analyzing by Groups

In SAS/INSIGHT software, you can use a *group variable* to process your data separately for each value of the group variable. You can use multiple group variables to process your data separately for each unique combination of grouping values.

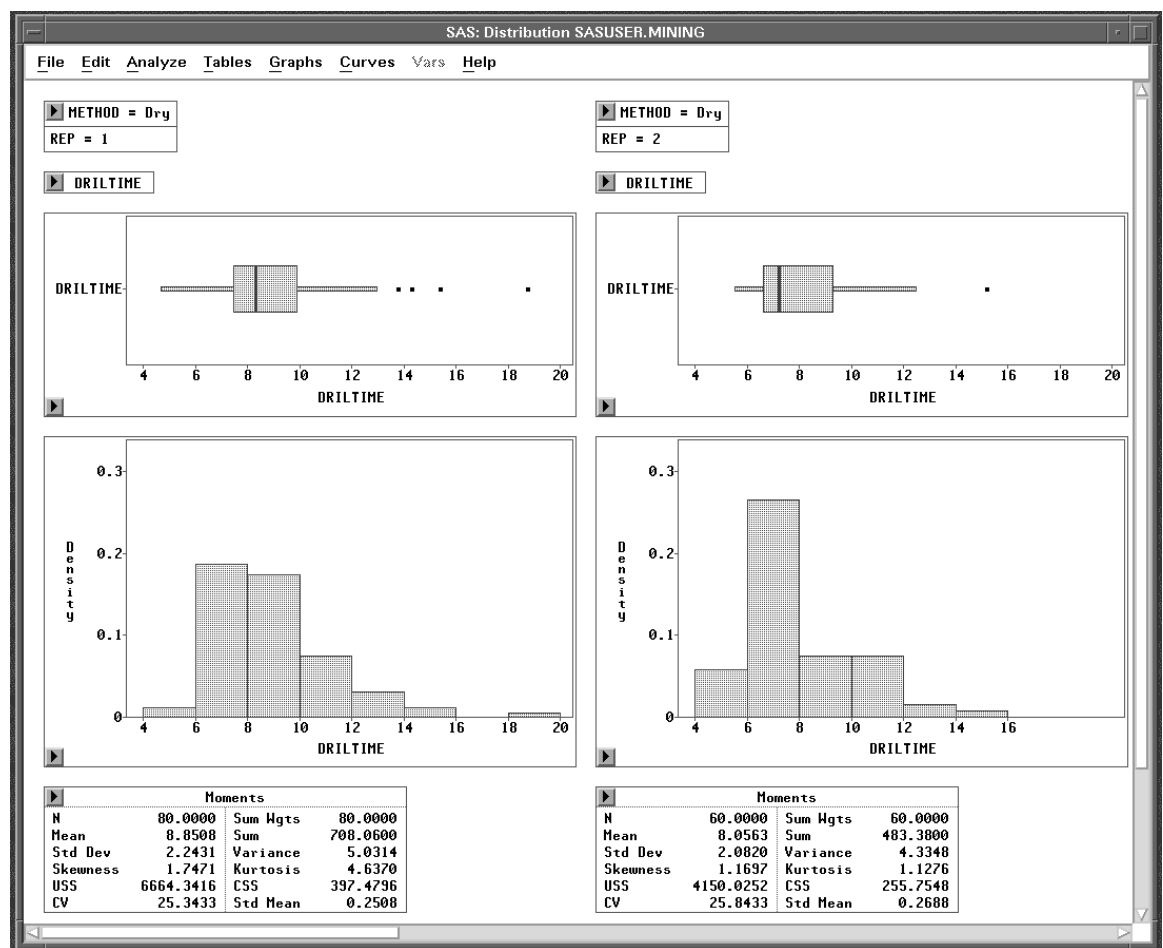


Figure 22.1. Analyzing by Groups

Using Group Variables

You can learn more about the distribution of drilling times by constructing a distribution analysis using group variables.

⇒ Choose **Analyze:Distribution (Y)**.

This displays the distribution variables dialog.

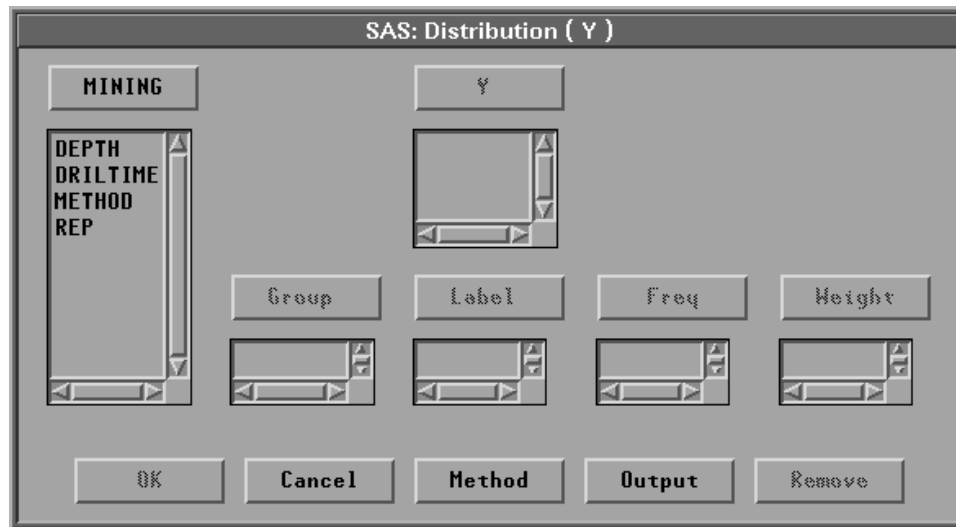


Figure 22.2. Distribution Variables Dialog

⇒ In the dialog, select **DRILTIME**, then click the **Y** button.

This assigns **DRILTIME** the required **Y** role.

⇒ Select **METHOD** and **REP**, and click the **Group** button.

This assigns **METHOD** and **REP** the **Group** role. You can scroll the **Group** list to see both variables. Because there are two values for **Method** and three values for **Rep**, this produces six groups.

⇒ Click **OK** to create the distribution window, as shown in [Figure 22.3](#).

The distribution window shows detailed information on the distributions, including box plots, histograms, moments, and quantiles. At the top of the distribution window is a table indicating the unique combination of values of the two group variables. You can scroll the distribution window to the right to see other levels.

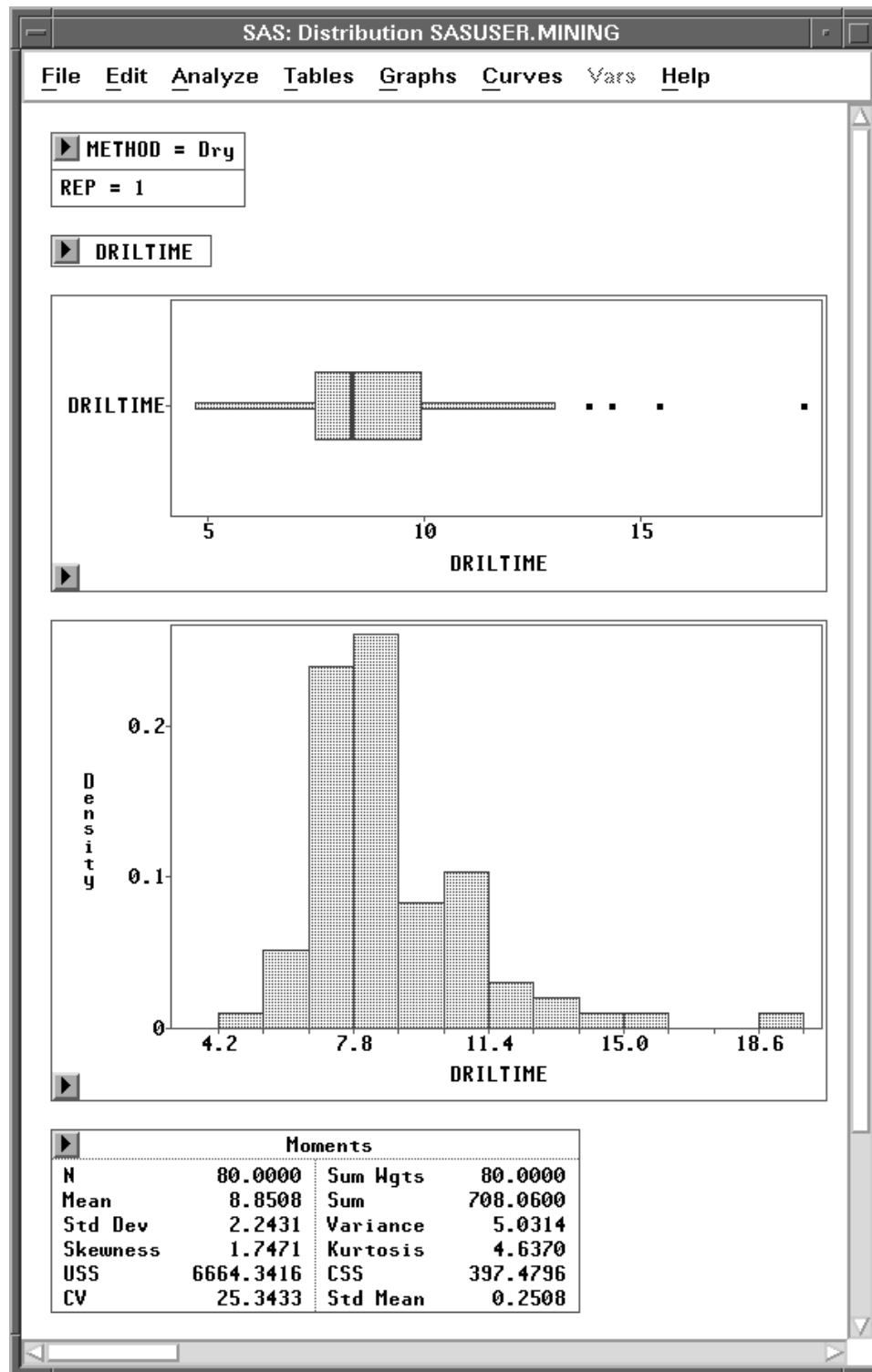


Figure 22.3. Distribution Window with Group Variables

Comparing Groups by Copying Windows

Because there are six groups, it is difficult to compare two groups side by side. Also, the axes are scaled to fit the data, so by default graphs use different axes.

To compare two groups side by side using the same axes, you can create a copy of the distribution analysis, set tick marks, and align the axes.

⇒ **Choose Edit:Windows:Copy Window in the Distribution analysis.**

This creates a copy of the distribution analysis.

⇒ **Move the two analyses side by side.**

Now you can scroll the windows horizontally to compare any two groups. [Figure 22.4](#) shows the first and last groups side by side.

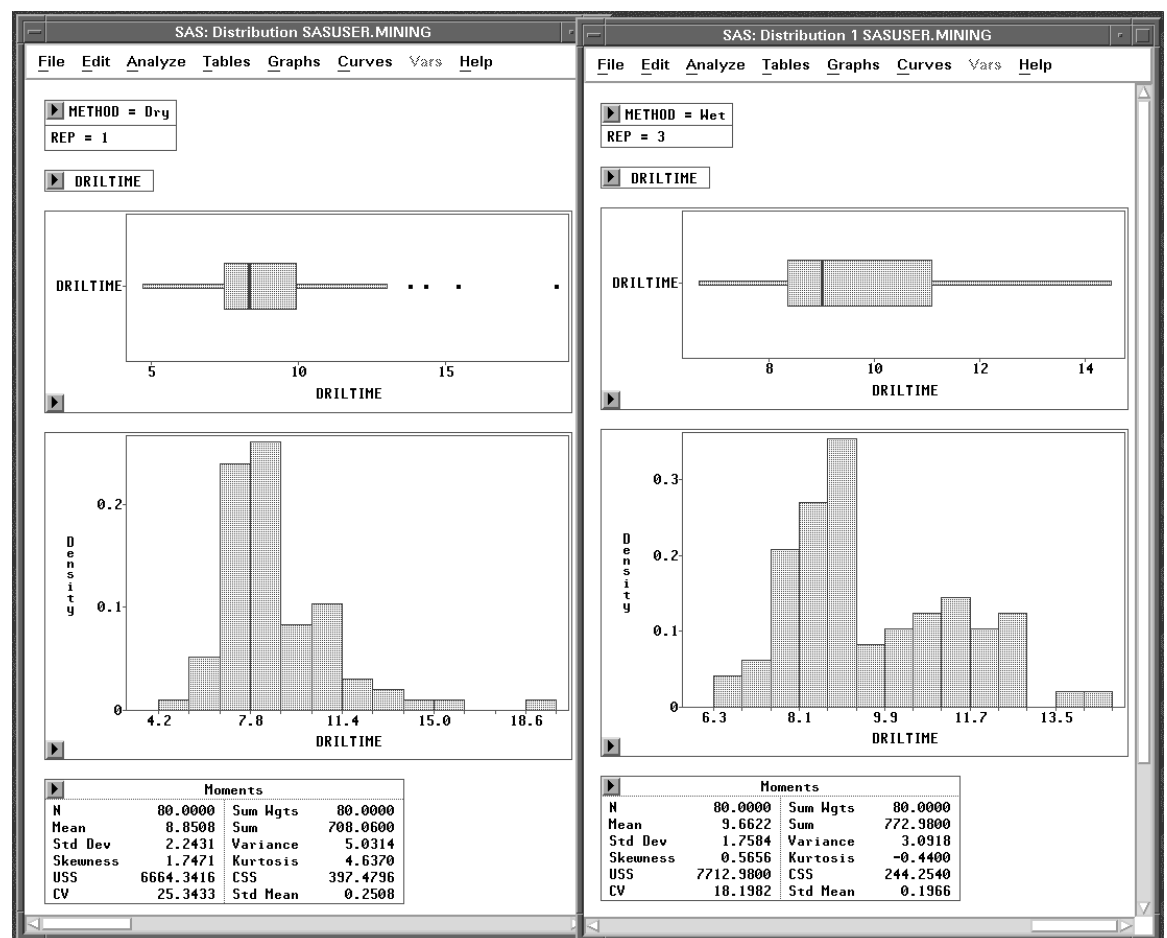


Figure 22.4. Comparing Distribution Analyses

The **Moments** and **Quantiles** tables below the histograms present statistics you can compare. The box plots and histograms, however, are difficult to compare because they use different axes. You can customize the axes with the following steps.

- ⇒ Select **DRILTIME** in the first distribution window.
- ⇒ Choose **Ticks** from the histogram pop-up menu in the first window.
This displays the **Ticks** dialog.
- ⇒ Make the adjustments shown in the following figure, and click the **OK** button.
This scales the **DRILTIME** axis for all histograms.

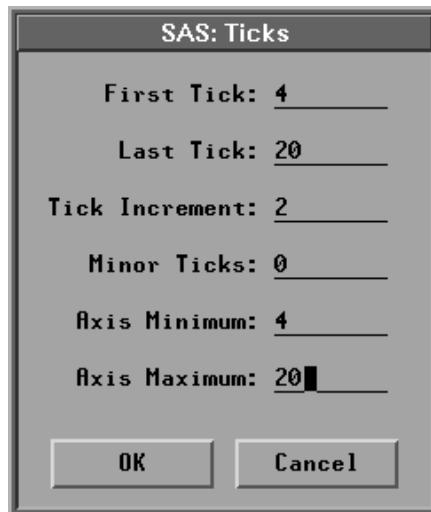


Figure 22.5. Ticks Dialog

- ⇒ Repeat these steps for the box plots in the first window.
This scales the **DRILTIME** axis for all box plots.
- ⇒ Repeat these steps for the second window.
Now you can compare box plots and histograms in both windows.

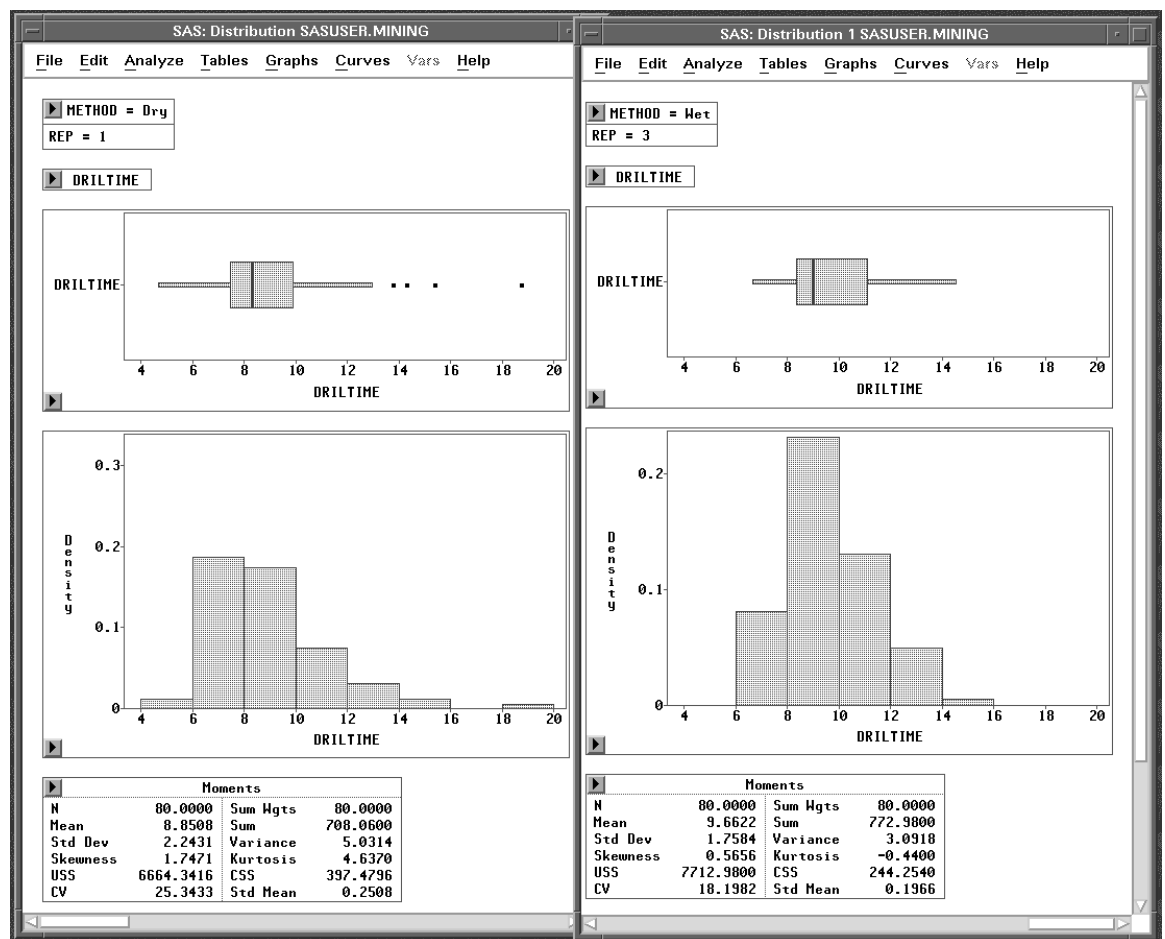


Figure 22.6. Comparing Distribution Analyses

† **Note:** By default, axes in SAS/INSIGHT software are scaled to fit the data. You can choose **Edit:Windows:Align** in any analysis window to align axes that use the same variable. Aligning affects only the axis scale, not the tick marks. When aligning histogram axes as in the preceding example, you should use the Ticks dialog to give histogram bars the same width and position.

⊕ **Related Reading:** Distributions, [Chapter 38](#).

Setting Default Group Variables

Often you will want to assign **Group** roles to the same group variables throughout a SAS/INSIGHT session. You can save time by setting default **Group** roles in the data window so that you do not have to set them in every variables dialog.

To set default **Group** roles for **SASUSER.MINING**, follow these steps.

- ⇒ **Choose Define Variables from the data pop-up menu.**
This displays the Define Variables dialog.
- ⇒ **In the dialog, click on METHOD, then click on Group under Default Role.**
This assigns the **Group** role to the **METHOD** variable.

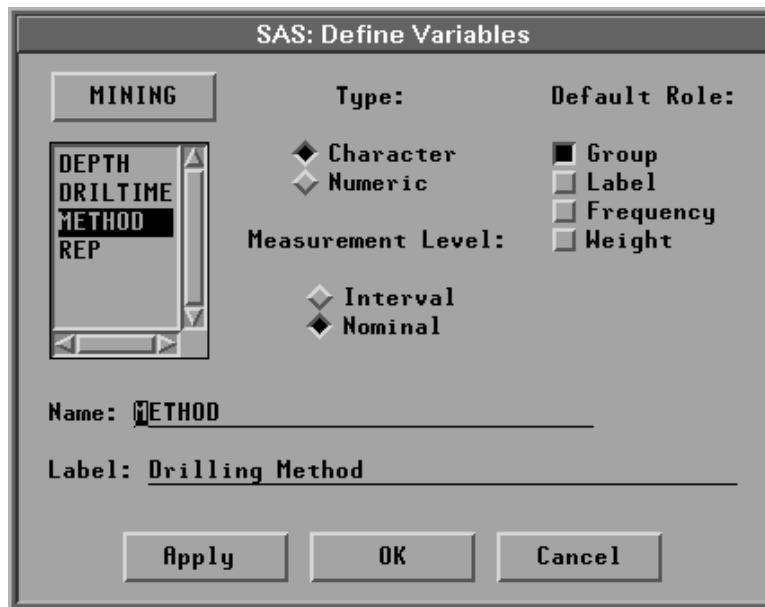
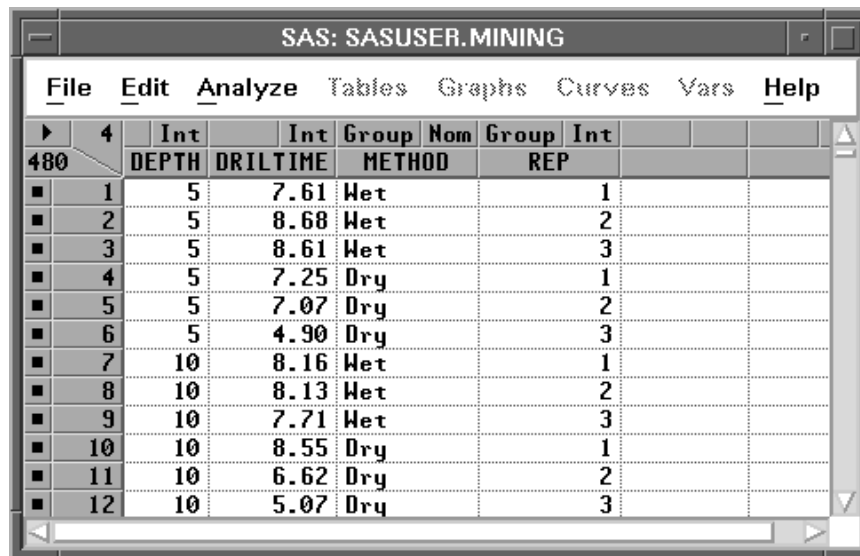


Figure 22.7. Assigning Group Role

- ⇒ **Click the Apply button.**
This assigns the **Group** role to **METHOD** but leaves the Define Variables window open so that you can assign roles to other variables as well.
- ⇒ **Click on REP, then click on Group under Default Role.**
This assigns the **Group** role to the **REP** variable as well.
- ⇒ **Click the OK button to close the dialog**
The **Group** role now appears above both **METHOD** and **REP** in the data window.



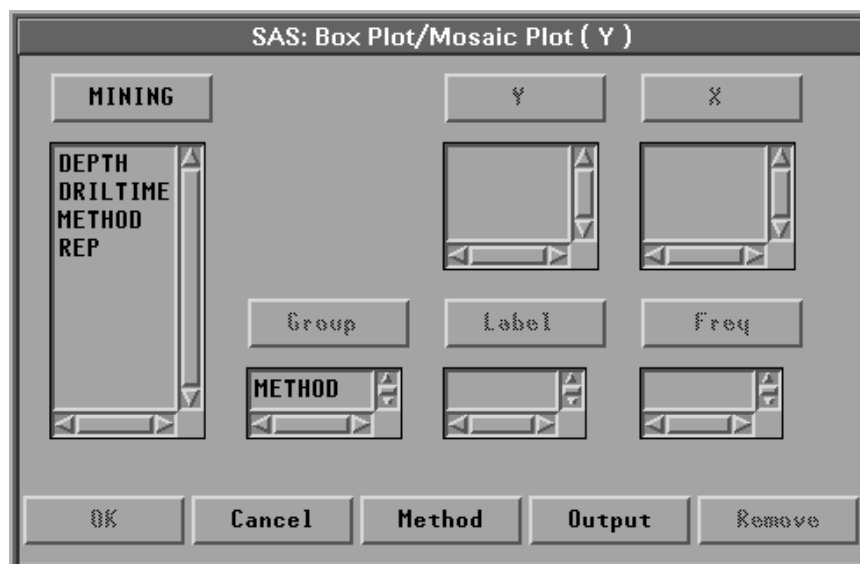
	4	Int	Int	Group	Nom	Group	Int
	480	DEPTH	DRILTIME	METHOD		REP	
■	1	5	7.61	Wet		1	
■	2	5	8.68	Wet		2	
■	3	5	8.61	Wet		3	
■	4	5	7.25	Dry		1	
■	5	5	7.07	Dry		2	
■	6	5	4.90	Dry		3	
■	7	10	8.16	Wet		1	
■	8	10	8.13	Wet		2	
■	9	10	7.71	Wet		3	
■	10	10	8.55	Dry		1	
■	11	10	6.62	Dry		2	
■	12	10	5.07	Dry		3	

Figure 22.8. Two Group Roles Assigned

† **Note:** Order is significant. The order in which you assign roles is the order in which your group variables are used in analyses.

⇒ Choose **Analyze:Box Plot/Mosaic Plot (Y)**.

Notice that the **Group** roles are already assigned. Only **METHOD** is visible, but you can scroll the **Group** list to see **REP**.



SAS: Box Plot/Mosaic Plot (Y)

MINING

Y

X

DEPTH
DRILTIME
METHOD
REP

Group

Label

Freq

METHOD

OK Cancel Method Output Remove

Figure 22.9. Box Plot Variables Dialog

Now every analysis you create will use the default **Group** roles you assigned in the data window. If you want to create an analysis without these variables, you can select

them in the variables dialog and click the **Remove** button.

Formatting Group Variables

Usually, SAS formats in SAS/INSIGHT software determine only how data are visually displayed. Group variables, however, can use SAS formats to combine different values into a larger group. For example, suppose you are interested only in approximate depths, not in the exact values of **DEPTH**. You can use a format to combine the values of **DEPTH** into three groups:

- $\text{DEPTH} \leq 100$
- $100 < \text{DEPTH} \leq 300$
- $300 < \text{DEPTH}$

Once you have assigned this format to **DEPTH**, you can assign **DEPTH** a **Group** role and use it as described earlier in this section. Each use of **DEPTH** creates three groups containing values in the three ranges you specified.

⊕ **Related Reading:** Formats, [Chapter 24](#).

Chapter 23

Animating Graphs

Chapter Contents

ANIMATING SELECTION OF OBSERVATIONS	369
ANIMATING SELECTED GRAPHS	373

Chapter 23

Animating Graphs

SAS/INSIGHT software provides two ways to *animate* graphs.

You can animate selected observations in all graphs simultaneously. This produces the same visual effect as brushing but gives you precise control over the display.

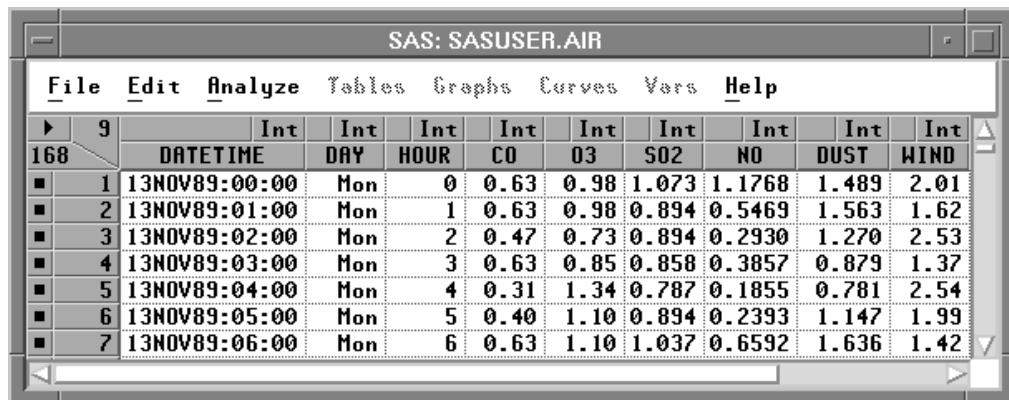
You can animate selected graphs individually. This restricts the animation to one graph and animates observations and other graph features.

Animating Selection of Observations

To animate the selection of observations, follow these steps.

⇒ **Open the AIR data set.**

This data set contains measurements of carbon monoxide and sulfur dioxide in city air over various times and dates. Since these data are time-dependent, they are a good subject for animation.



	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
	DATETIME	DAY	HOUR	CO	O3	SO2	NO	DUST	WIND	
1	13NOV89:00:00	Mon	0	0.63	0.98	1.073	1.1768	1.489	2.01	
2	13NOV89:01:00	Mon	1	0.63	0.98	0.894	0.5469	1.563	1.62	
3	13NOV89:02:00	Mon	2	0.47	0.73	0.894	0.2930	1.270	2.53	
4	13NOV89:03:00	Mon	3	0.63	0.85	0.858	0.3857	0.879	1.37	
5	13NOV89:04:00	Mon	4	0.31	1.34	0.787	0.1855	0.781	2.54	
6	13NOV89:05:00	Mon	5	0.40	1.10	0.894	0.2393	1.147	1.99	
7	13NOV89:06:00	Mon	6	0.63	1.10	1.037	0.6592	1.636	1.42	

Figure 23.1. AIR Data

⇒ **Select CO, then SO2 in the data window using extended selection.**

⇒ **Choose Analyze:Scatter Plot (Y X).**

This creates a scatter plot of **CO** versus **SO2**.

⇒ **Choose Edit:Windows:Animate.**

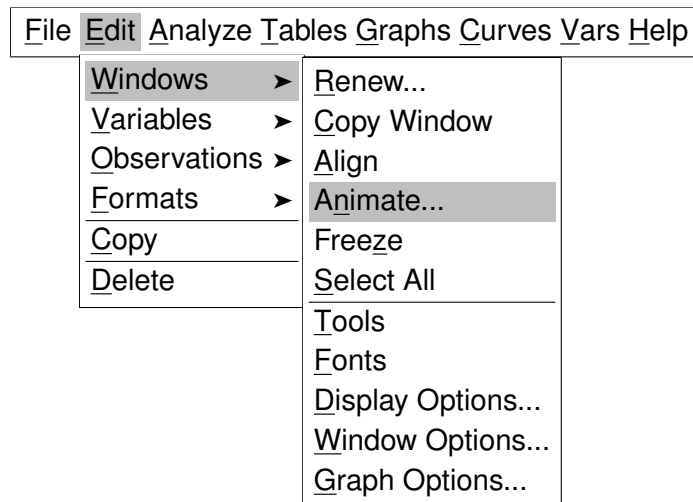


Figure 23.2. Edit:Windows Menu

This displays the animation dialog. The animation dialog contains a list of variables, a list of values, and a slider to control speed.



Figure 23.3. Animation Dialog

⇒ Select **DAY** in the list of variables, then click the **Apply** button.

This animates the selection of observations over all values of **DAY** in the order in

which they are displayed in the animation dialog. Observations are selected in both the scatter plot and the data window, and the current value is selected in the animation dialog.

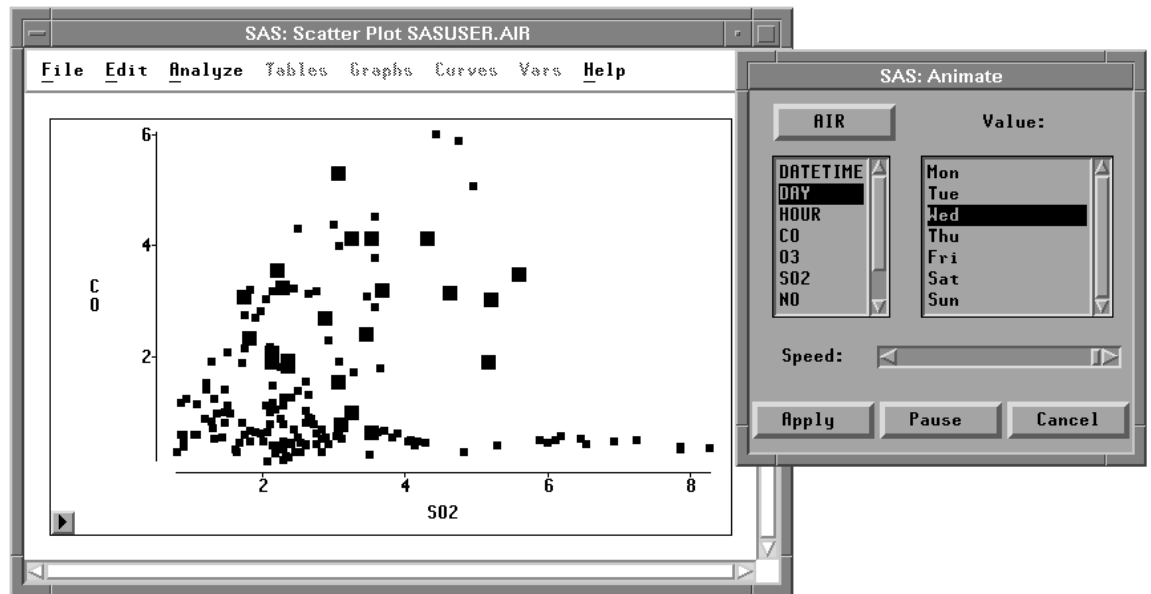


Figure 23.4. Animating Selection of Observations

⇒ **Adjust speed by clicking or dragging on the slider.**

When the slider is at the extreme left, speed is slowest; at the right, speed is fastest. Animation speed also depends on the speed of your host, the number of observations in your data set, and the number of graphs displayed.

⇒ **Click the Pause button to stop the animation.**

You can make the pattern of animation clearer by toggling the display of observations.

⇒ **Choose Observations from the scatter plot pop-up menu.**

This turns off the display of all deselected observations.

⇒ **Click the Apply button to restart the animation.**

You should begin to see the conditional distributions of **CO** and **SO2** as **DAY** varies over the day of the week.

⇒ **Click in the Value list in the animation dialog**

This enables you to stop the animation on particular values. You can click in the **Value** list to compare pollutant concentrations on different days.

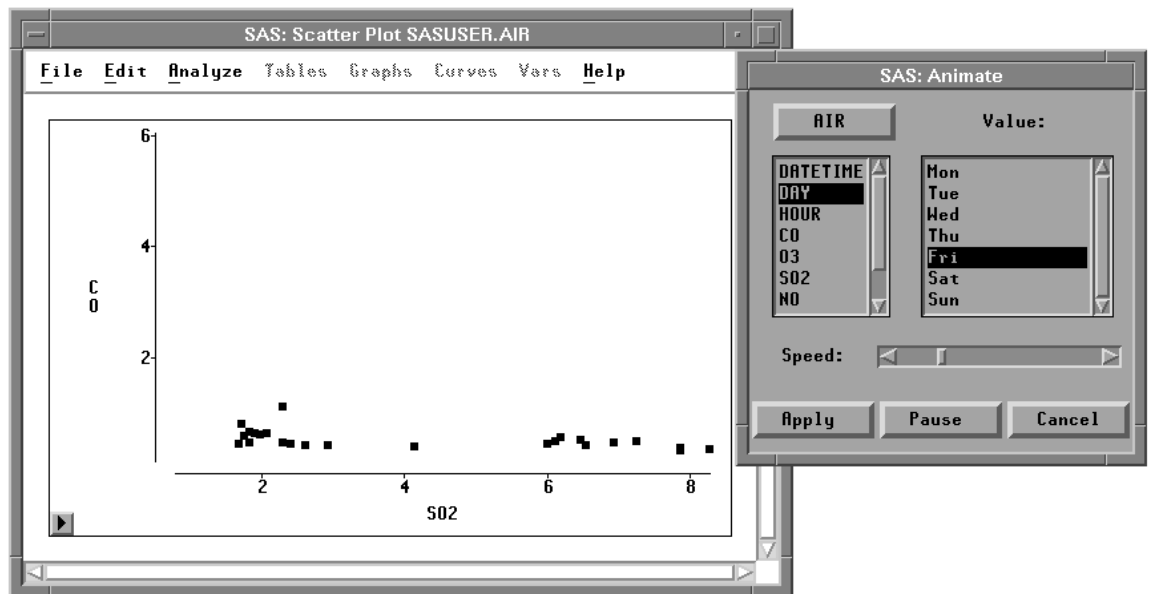


Figure 23.5. Animating Only Selected Observations

The concentrations of CO and SO₂ vary widely through most of the days but are much lower on Saturday and Sunday. Carbon monoxide is produced primarily by automobile exhaust, and automobile traffic appears to be reduced on the weekends. Sulfur dioxide concentrations are also lower; this pollutant is produced by power plants that operate at a reduced rate on weekends.

Animating Selected Graphs

Line plots are an effective way to look at time-dependent data. You can animate line plots and other graphs by selecting them before using the animation dialog. This animates lines and other features in the graph, not just selected observations.

⇒ **Select CO, then SO2, then HOUR in the data window.**

The last variable you select, **HOUR**, will receive the **X** role in the line plot.

⇒ **Choose Analyze:Line Plot (Y X).**

This creates a plot with two overlaid lines. The lines are jagged because the data contain seven observations for each hour.

⇒ **Select the line plot by clicking on any edge.**

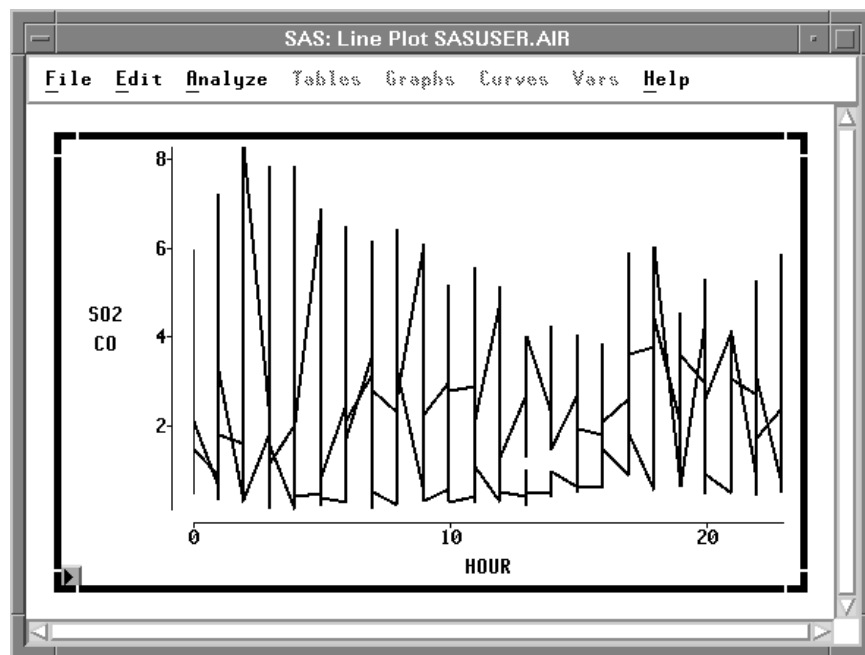


Figure 23.6. Selected Line Plot

⇒ **Select DAY in the animation dialog, then click the Apply button.**

This animates the line plot, showing pollutant concentrations for each day of the week.

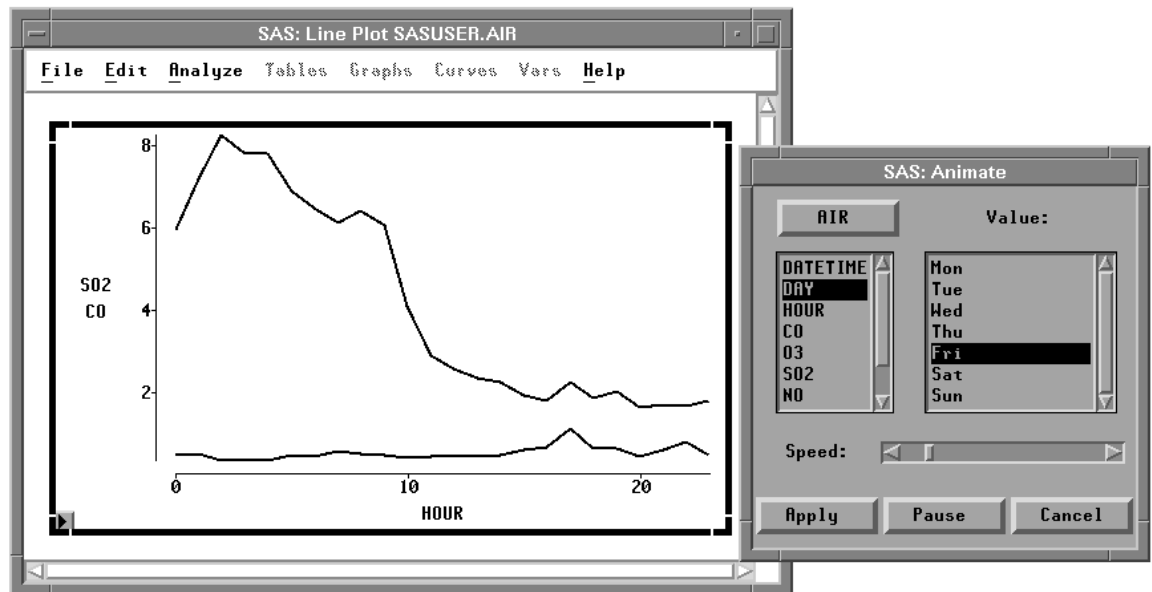


Figure 23.7. Animating a Line Plot

Notice the peak CO concentrations on weekday mornings and afternoons. These might be caused by increased automobile emissions during rush-hour traffic.

⇒ **When you are finished, click **Cancel** to close the animation dialog**

Chapter 24

Formatting Variables and Values

Chapter Contents

ASSIGNING FORMATS	378
CREATING FORMATS	385

Chapter 24

Formatting Variables and Values

Formats determine how variables and values are displayed. In group variables and model effects, formats can also determine how values are used in calculations.

You can use formats to set the width of displayed values, the number of decimal points displayed, the handling of blanks, zeroes, and commas, and other details. The SAS System provides many standard formats for displaying character and numeric values.

In addition, you can use the FORMAT procedure to create your own formats.

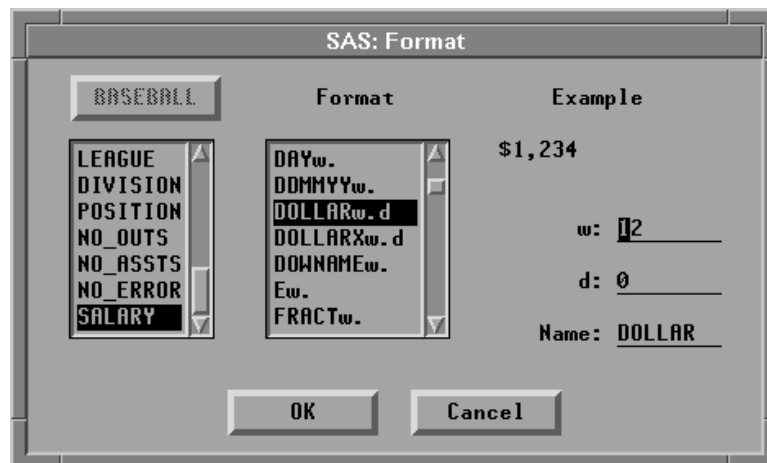


Figure 24.1. Assigning a Format

Assigning Formats

By default, SAS/INSIGHT software displays each variable using the format supplied in your SAS data set. If your data set contains numeric variables with no formats, SAS/INSIGHT software chooses a format based on that variable's values. When you save the data, formats chosen by SAS/INSIGHT software are not automatically saved, but any formats you assign are saved.

You can assign formats by using the **Edit:Formats** menu.

⇒ **Open the BASEBALL data set.**

This data set contains statistics and salaries of major league baseball players.

⇒ **Select the variable SALARY.**

		Nom	Int	Int	Int	Int
322	POSITION	NO_OUTS	NO_ASSSTS	NO_ERROR	SALARY	
1	10	317	36	1	75.000	
2	C	446	33	20	.	
3	UT	80	45	8	240.000	
4	3S	73	152	11	225.000	
5	CF	247	4	8	.	
6	C	632	43	10	475.000	
7	2B	186	290	17	550.000	
8	RF	295	15	5	950.000	
9	OF	90	4	0	.	
10	1B	1236	98	18	100.000	
11	C	359	30	4	305.000	
12	RF	368	20	3	1237.500	

Figure 24.2. SALARY Selected

⇒ **Choose Edit:Formats:9.1.**

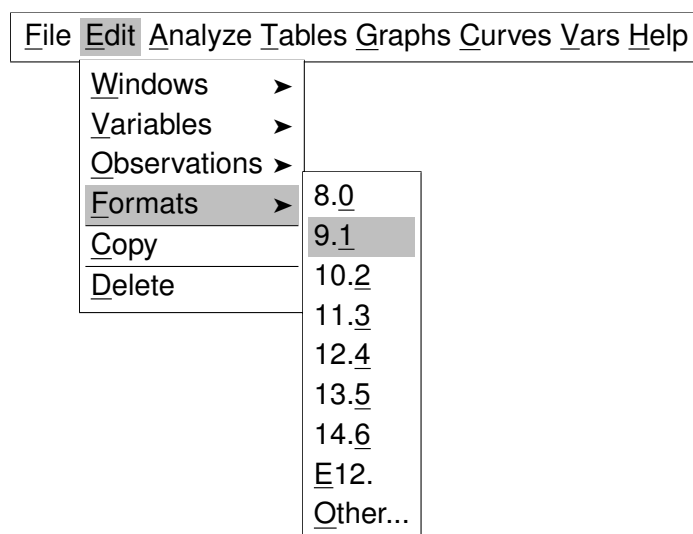
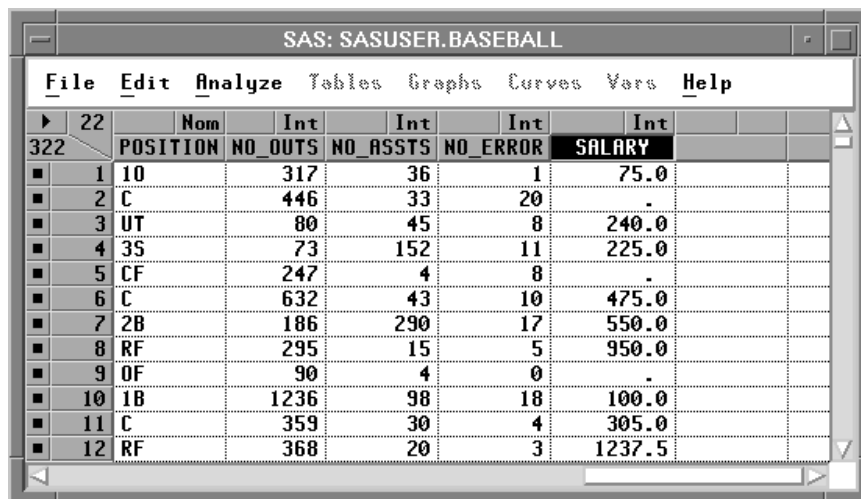


Figure 24.3. Edit:Formats Menu

This gives the variable **SALARY** a width of nine character positions, including the decimal and one position after the decimal. The actual data values for **SALARY** continue to be stored with double precision.



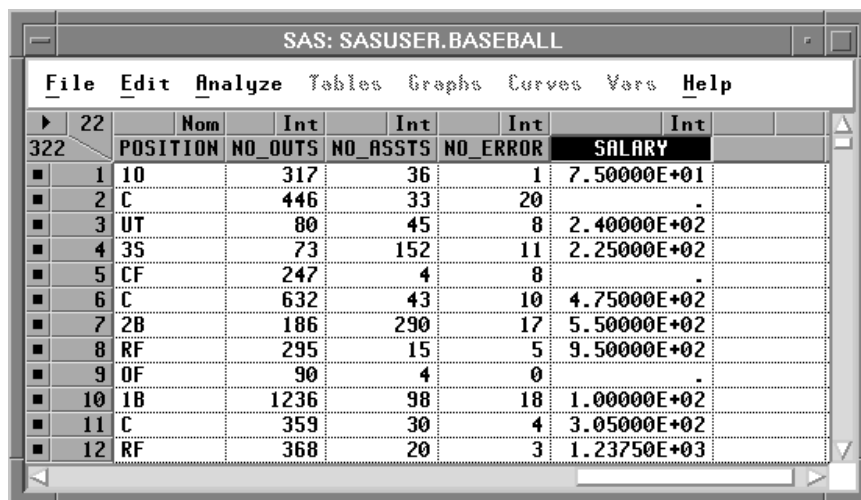
The screenshot shows the SAS window titled 'SAS: SASUSER.BASEBALL'. The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. The data grid displays 12 rows of baseball player statistics. The columns are labeled: POSITION (Nom), NO_OUTS (Int), NO_ASSTS (Int), NO_ERROR (Int), and SALARY (Int). The SALARY column is formatted with Format 9.1, showing values like 75.0, 240.0, 225.0, 475.0, 550.0, 950.0, 100.0, 305.0, and 1237.5.

	POSITION	NO_OUTS	NO_ASSTS	NO_ERROR	SALARY
1	10	317	36	1	75.0
2	C	446	33	20	.
3	UT	80	45	8	240.0
4	3S	73	152	11	225.0
5	CF	247	4	8	.
6	C	632	43	10	475.0
7	2B	186	290	17	550.0
8	RF	295	15	5	950.0
9	OF	90	4	0	.
10	1B	1236	98	18	100.0
11	C	359	30	4	305.0
12	RF	368	20	3	1237.5

Figure 24.4. Format 9.1

⇒ Choose **Edit:Formats:E12**.

This gives the variable **SALARY** a width of 12 character positions and expresses each value in exponential notation.



The screenshot shows the same SAS window as Figure 24.4, but now the SALARY column is formatted with Format E12. The values are displayed in exponential notation, such as 7.50000E+01, 2.40000E+02, 2.25000E+02, 4.75000E+02, 5.50000E+02, 9.50000E+02, 1.00000E+02, 3.05000E+02, and 1.23750E+03.

	POSITION	NO_OUTS	NO_ASSTS	NO_ERROR	SALARY
1	10	317	36	1	7.50000E+01
2	C	446	33	20	.
3	UT	80	45	8	2.40000E+02
4	3S	73	152	11	2.25000E+02
5	CF	247	4	8	.
6	C	632	43	10	4.75000E+02
7	2B	186	290	17	5.50000E+02
8	RF	295	15	5	9.50000E+02
9	OF	90	4	0	.
10	1B	1236	98	18	1.00000E+02
11	C	359	30	4	3.05000E+02
12	RF	368	20	3	1.23750E+03

Figure 24.5. Format E12.

The **Edit:Formats** menu provides quick access to frequently used formats. There are many other standard formats provided by the SAS System.

⇒ Choose **Edit:Formats:Other**.

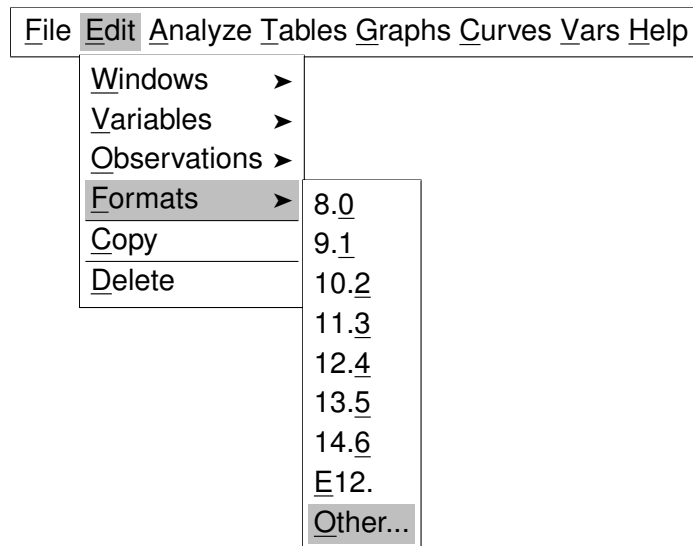


Figure 24.6. Edit:Formats Menu

This displays the Format dialog. In the dialog, the fields **w** and **d** specify the width and decimal places to be used by the formats. Note that the **SALARY** variable and the **E12.** format are currently selected.

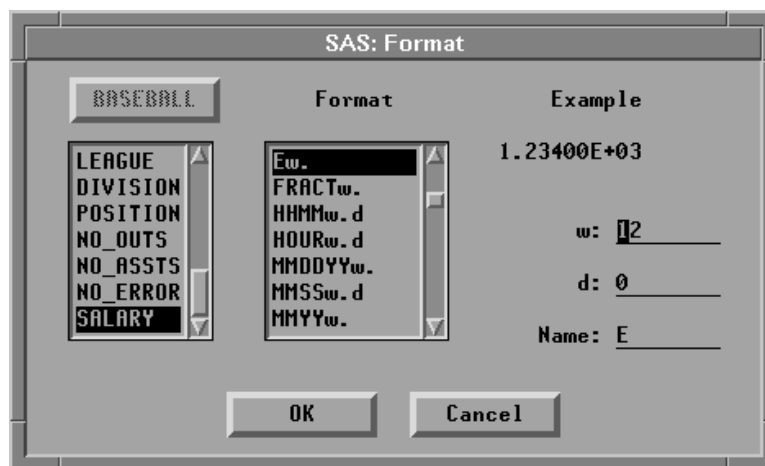


Figure 24.7. Format Dialog

⇒ **Select DOLLARw.d in the Format list.**

Formats are listed alphabetically, so the **DOLLARw.d** format is above the **Ew.** format.



Figure 24.8. Format **DOLLARw.d**

The example in the upper right corner of the dialog illustrates the format you have selected. **DOLLAR** is the standard format for display of currency in the United States. There is also a **DOLLARX** format sometimes preferred in European countries.

⇒ Click **OK** to set the format you prefer.

The image shows the 'SAS: SASUSER.BASEBALL' data window. The table has columns: POSITION, NO_OUTS, NO_ASSTS, NO_ERROR, and SALARY. The 'SALARY' column is formatted with the DOLLARw.d format, showing values like \$75, \$240, \$225, \$475, \$550, \$950, \$100, \$305, and \$1,238.

	POSITION	NO_OUTS	NO_ASSTS	NO_ERROR	SALARY
1	10	317	36	1	\$75
2	C	446	33	20	.
3	UT	80	45	8	\$240
4	3S	73	152	11	\$225
5	CF	247	4	8	.
6	C	632	43	10	\$475
7	2B	186	290	17	\$550
8	RF	295	15	5	\$950
9	OF	90	4	0	.
10	1B	1236	98	18	\$100
11	C	359	30	4	\$305
12	RF	368	20	3	\$1,238

Figure 24.9. **SALARY** Formatted

Now the variable **SALARY** uses the format you assigned. By default, this format is also used for axes in subsequent analyses. You can modify the axes, however, to use other formats.

⇒ Choose **Analyze:Distribution (Y)**.

This creates a distribution analysis of **SALARY**. The box plot and histogram axes use the format you assigned to the **SALARY** variable in the data window.

⇒ Select **SALARY** in the distribution window.

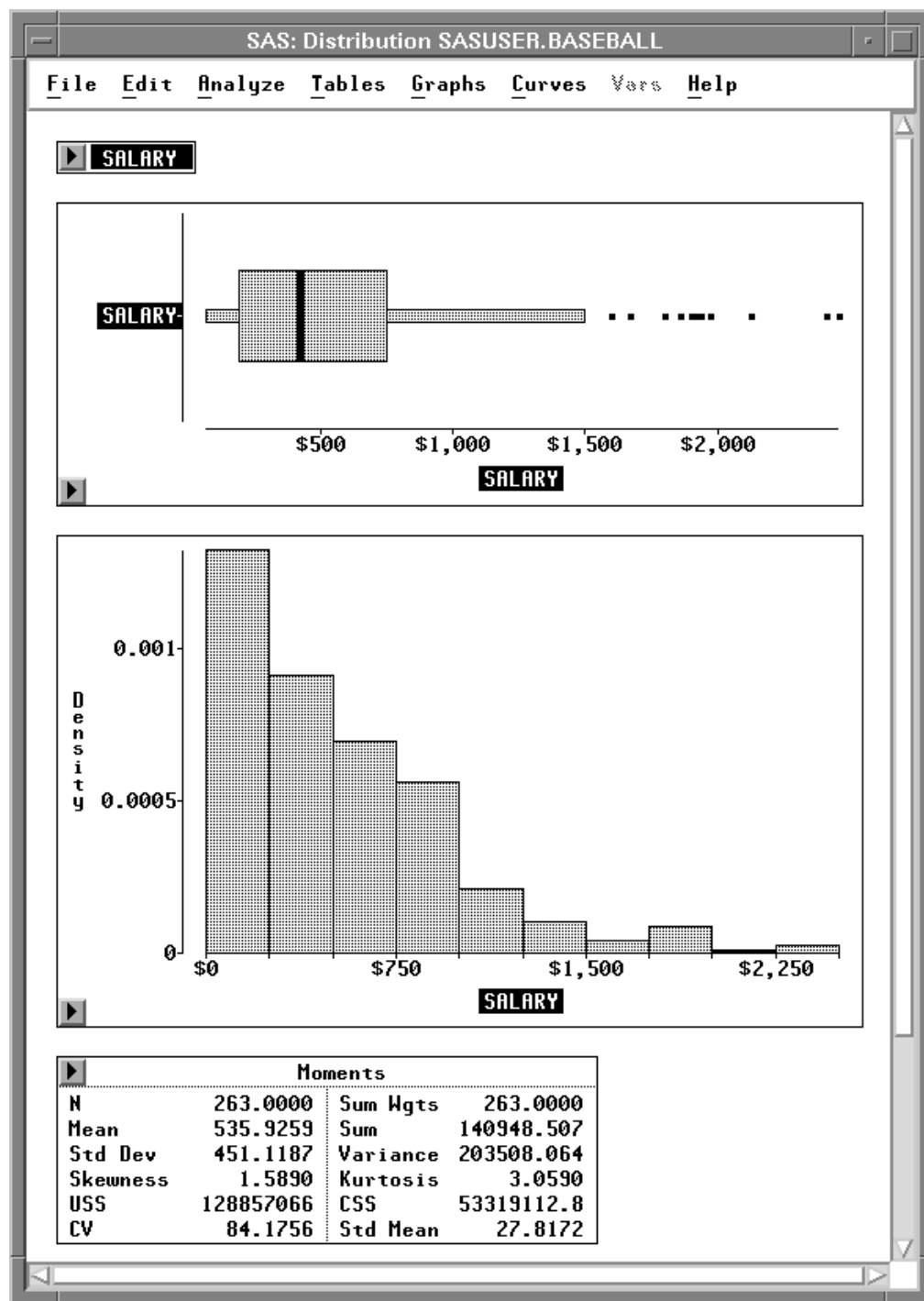


Figure 24.10. Distribution Analysis, **SALARY** Selected

⇒ Choose **Edit:Formats:8.0**.

This assigns the **8.0** format to **SALARY** on axes in the distribution window. In the data window, **SALARY** continues to use the **DOLLAR** format.

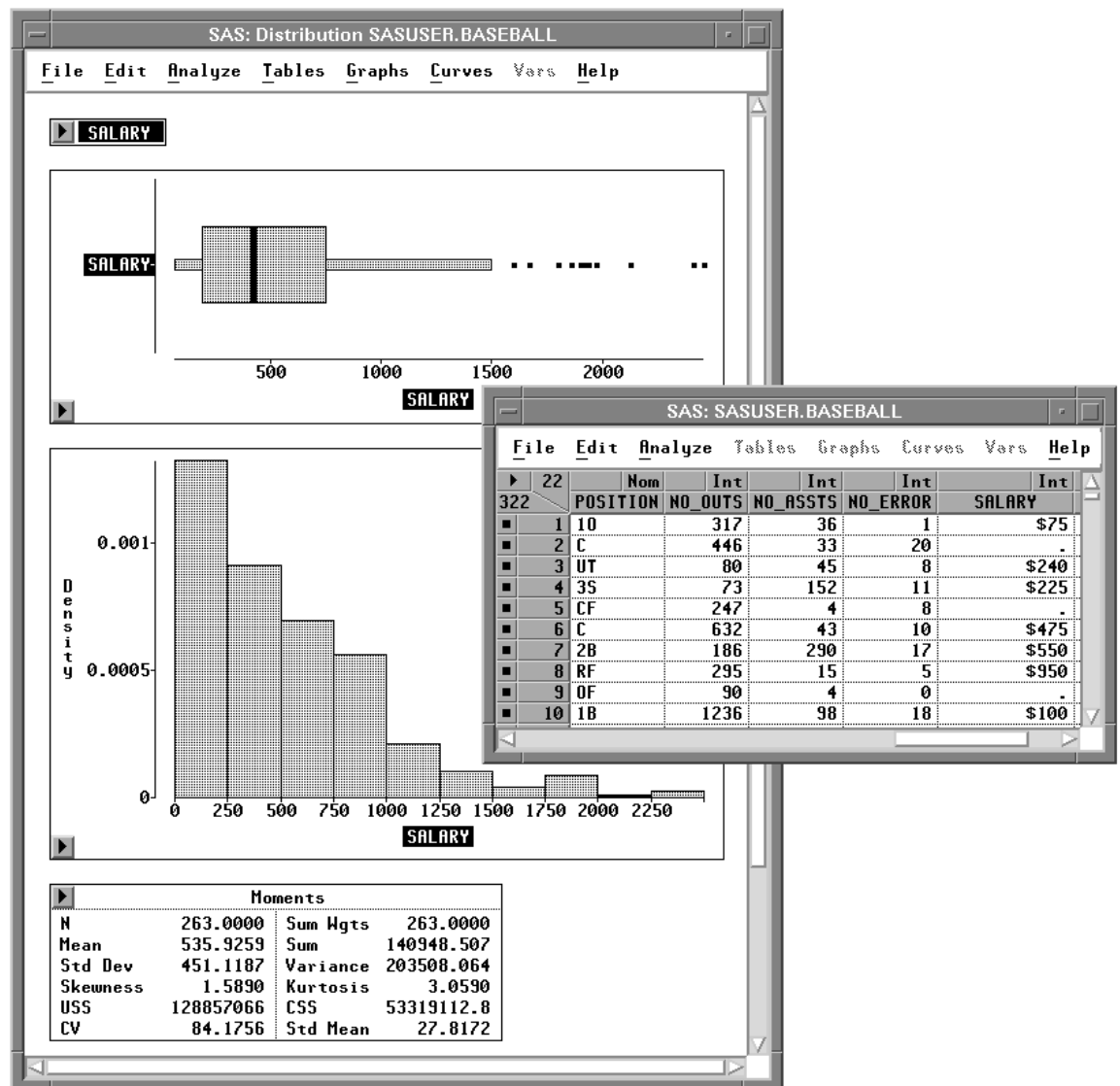
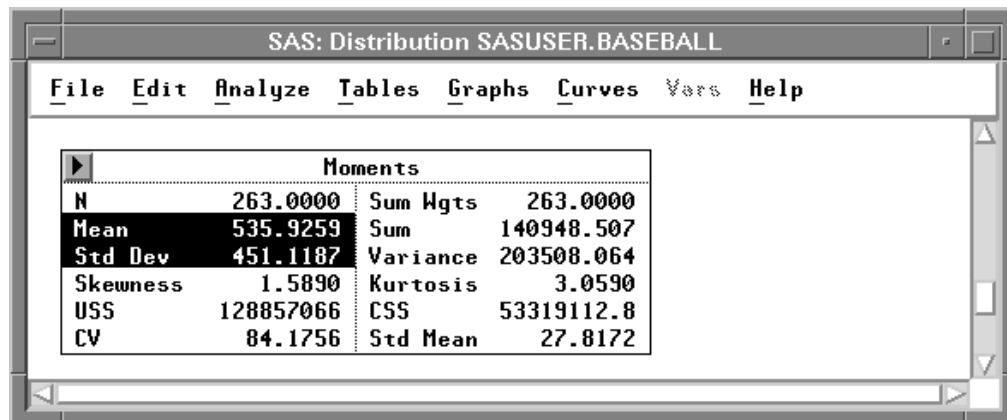


Figure 24.11. SALARY Axes Formatted

You can also format individual values in analysis tables. For example, suppose you need to see greater precision for the mean and standard deviation.

⇒ Select the values for **Mean** and **Std Dev** in the **Moments** table.



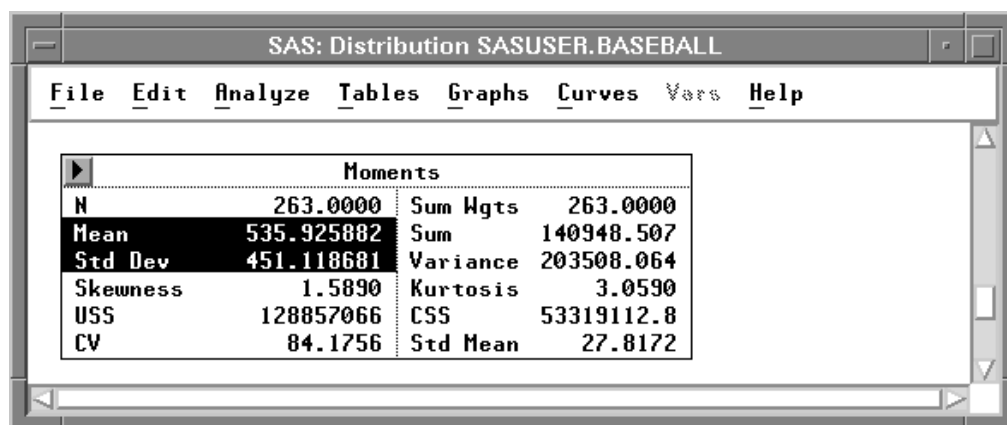
The screenshot shows the SAS Distribution window for SASUSER.BASEBALL. The 'Moments' table is displayed with the following data:

Moments			
N	263.0000	Sum Wgts	263.0000
Mean	535.9259	Sum	140948.507
Std Dev	451.1187	Variance	203508.064
Skewness	1.5890	Kurtosis	3.0590
USS	128857066	CSS	53319112.8
CV	84.1756	Std Mean	27.8172

Figure 24.12. Moments Table, Values Selected

⇒ Choose **Edit:Formats:14.6**.

Now the mean and standard deviation show six digits after the decimal.



The screenshot shows the same SAS Distribution window, but the 'Mean' and 'Std Dev' values have been formatted to show six digits after the decimal point.

Moments			
N	263.0000	Sum Wgts	263.0000
Mean	535.925882	Sum	140948.507
Std Dev	451.118681	Variance	203508.064
Skewness	1.5890	Kurtosis	3.0590
USS	128857066	CSS	53319112.8
CV	84.1756	Std Mean	27.8172

Figure 24.13. Moments Table, Values Formatted

Creating Formats

Although there are many formats available in the SAS System, occasionally you will want to create your own. To do this, use the FORMAT procedure.

For example, suppose you want to consider certain groupings of baseball players based on the length of their careers. You can combine the values of **YR_MAJOR** into four groups, as follows.

⇒ **Enter PROC FORMAT statements in the Program Editor.**

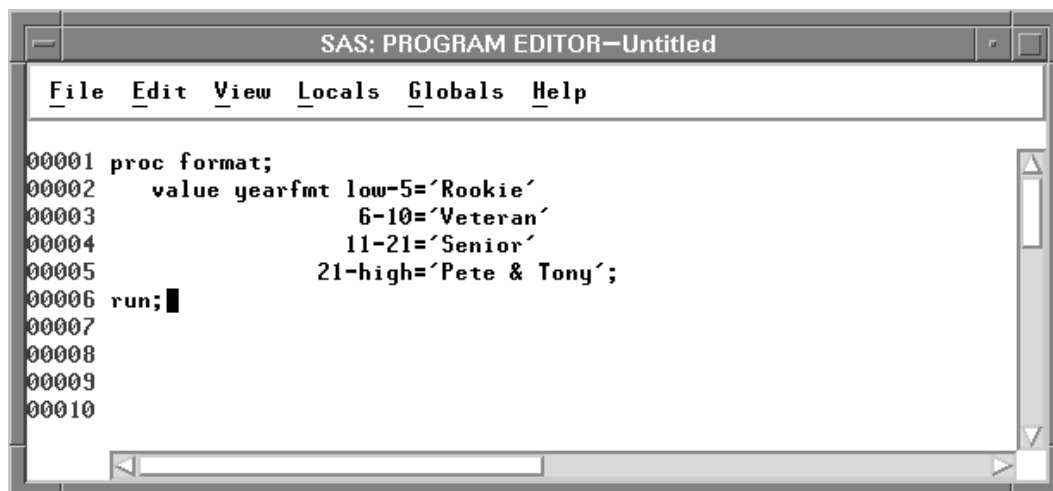


Figure 24.14. Program Editor

⇒ **Choose Run:Submit.**

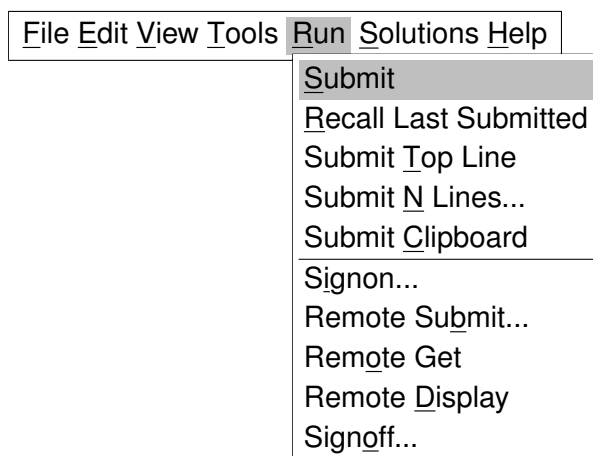


Figure 24.15. Run Menu

⇒ **Select YR_MAJOR.**

	Int	Int	Int	Int	Int	Int	Int
	NO_HOME	NO_RUNS	NO_RBI	NO_BB	YR_MAJOR	CR_ATBAT	CR_HITS
1	2	27	25	33	1	216	54
2	1	30	29	14	1	293	66
3	7	29	27	30	13	3231	825
4	1	31	15	22	4	926	210
5	11	40	58	24	11	4513	1134
6	7	24	38	39	14	3449	835
7	1	67	27	36	7	1775	506
8	21	72	88	38	7	3754	1077
9	4	25	19	27	19	7117	1981
10	29	54	88	43	6	1750	412
11	2	28	26	22	6	999	236
12	40	107	108	69	6	2325	634

Figure 24.16. YR_MAJOR Selected

⇒ Choose **Edit:Formats:Other**.
This displays the Format Dialog.

SAS: Format

BASEBALL

Format

YR_MAJOR
CR_ATBAT
CR_HITS
CR_HOME
CR_RUNS
CR_RBI
CR_BB

w.d
BESTw.
COMMAw.d
COMMAXw.d
DATEw.
DATETIMEw.d
DAYw.

Example

12

w: 2
d: 0
Name: w.d

OK Cancel

Figure 24.17. Format Dialog

⇒ Enter **YEARFMT** in the **Name** field.

⇒ Enter **12** in the **w** field, then press the **Return** key.

Now the example in the upper right of the dialog shows a value formatted with **YEARFMT**.

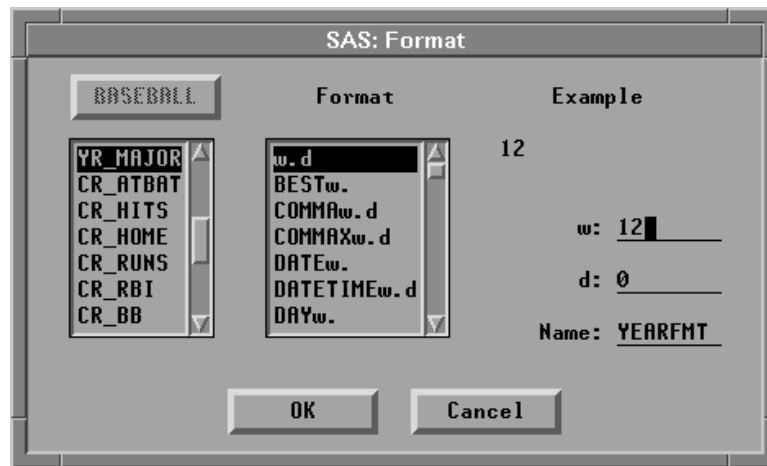


Figure 24.18. YEARFMT Entered

⇒ Click **OK** to close the Format dialog

Now **YEARFMT** is used to display the values of **YR_MAJOR**.

	Int	Int	Int	Int	Int	Int
	NO_HOME	NO_RUNS	NO_RBI	NO_BB	YR_MAJOR	CR_ATBAT
1	2	27	25	33	Rookie	216
2	1	30	29	14	Rookie	293
3	7	29	27	30	Senior	3231
4	1	31	15	22	Rookie	926
5	11	40	58	24	Senior	4513
6	7	24	38	39	Senior	3449
7	1	67	27	36	Veteran	1775
8	21	72	88	38	Veteran	3754
9	4	25	19	27	Senior	7117
10	29	54	88	43	Veteran	1750
11	2	28	26	22	Veteran	999
12	40	107	108	69	Veteran	2325

Figure 24.19. YEARFMT Assigned

By default, the new format is used to display values wherever you use **YR_MAJOR**. Formats are not used in calculations except for nominal variables in model effects or for group variables. In these cases, the format is used to determine the groups. You can see this use of formats by creating a box plot.

⇒ Deselect **YR_MAJOR** in the data window.

⇒ Choose **Analyze:Box Plot/Mosaic Plot (Y)**.

This displays the box plot variables dialog.

⇒ Assign **YR_MAJOR** the X role and **CR_HITS** the Y role.



Figure 24.20. Box Plot Variables Dialog

⇒ Click the **OK** button to create the box plot.

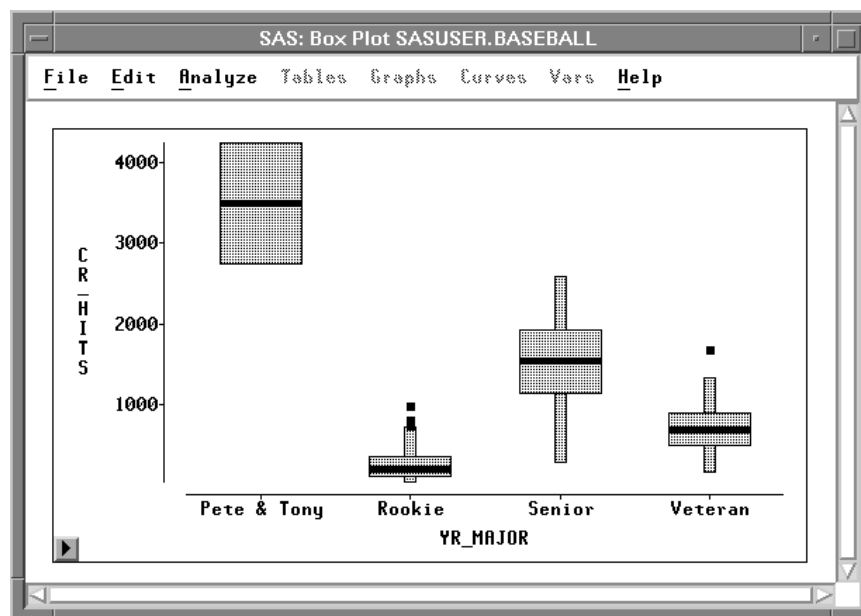


Figure 24.21. Box Plot of **CR_HITS** by **YR_MAJOR**

Since **YEARFMT** defines four formatted values, there are four boxes in the box plot. One of the boxes has no whiskers because it represents only two observations. Pete Rose and Tony Perez, ballplayers of exceptional hitting ability and longevity, are in a class by themselves.

To learn more about SAS formats, refer to *SAS Language Reference: Dictionary*. To learn more about creating your own formats with PROC FORMAT, refer to the *SAS Procedures Guide*.

⊕ **Related Reading:** Box Plots, [Chapter 33](#).

Chapter 25

Editing Windows

Chapter Contents

ZOOMING WINDOWS	394
RENEWING WINDOWS	401
ADDING AND DELETING	404
MOVING AND SIZING	411
ALIGNING GRAPHS	417

Chapter 25

Editing Windows

SAS/INSIGHT software provides many ways to edit the contents of your analysis windows. You can zoom in and out to see more or less detail. You can move, resize, add, and delete graphs and tables. You can align graphs. If you change your mind about your window layout, you can renew any window to restore its original state.

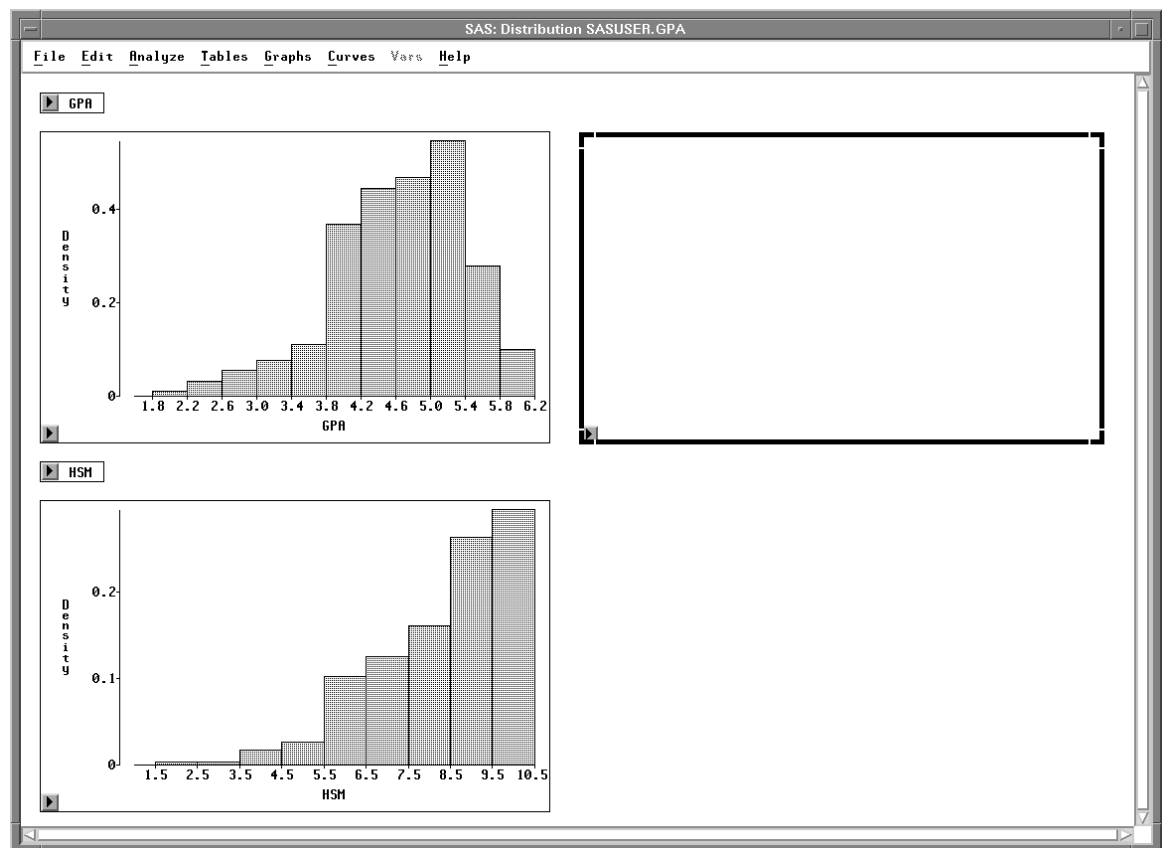


Figure 25.1. Adding a Graph to a Window

Zooming Windows

Zooming a window means adjusting the focus to make objects in the window larger or smaller. Zooming is most useful when you want to see more detail. For example, you may use zooming to explore data in a scatter plot matrix.

⇒ **Open the GPA data set.**

This data set contains college grade point averages, high school math, science, and English averages, and SAT scores of first-year college students.

⇒ **Select all the variables.**

Click on the variables count in the upper left corner.

SAS: SASUSER.GPA

File Edit Analyze Tables Graphs Curves Vars Help

7	Int	Int	Int	Int	Int	Int	Nom		
224	GPA	HSM	HSS	HSE	SATM	SATV	SEX		
1	5.32	10	10	10	670	600	Female		
2	5.14	9	9	10	630	700	Male		
3	3.84	9	6	6	610	390	Female		
4	5.34	10	9	9	570	530	Male		
5	4.26	6	8	5	700	640	Female		
6	4.35	8	6	8	640	530	Female		
7	5.33	9	7	9	630	560	Male		
8	4.85	10	8	8	610	460	Male		
9	4.76	10	10	10	570	570	Male		
10	5.72	7	8	7	550	500	Female		
11	4.08	9	10	7	670	600	Female		
12	5.38	8	9	8	540	580	Female		

Figure 25.2. Selecting All Variables

⇒ **Choose Analyze:Scatter Plot (Y X).**

This creates a seven-by-seven scatter plot matrix.

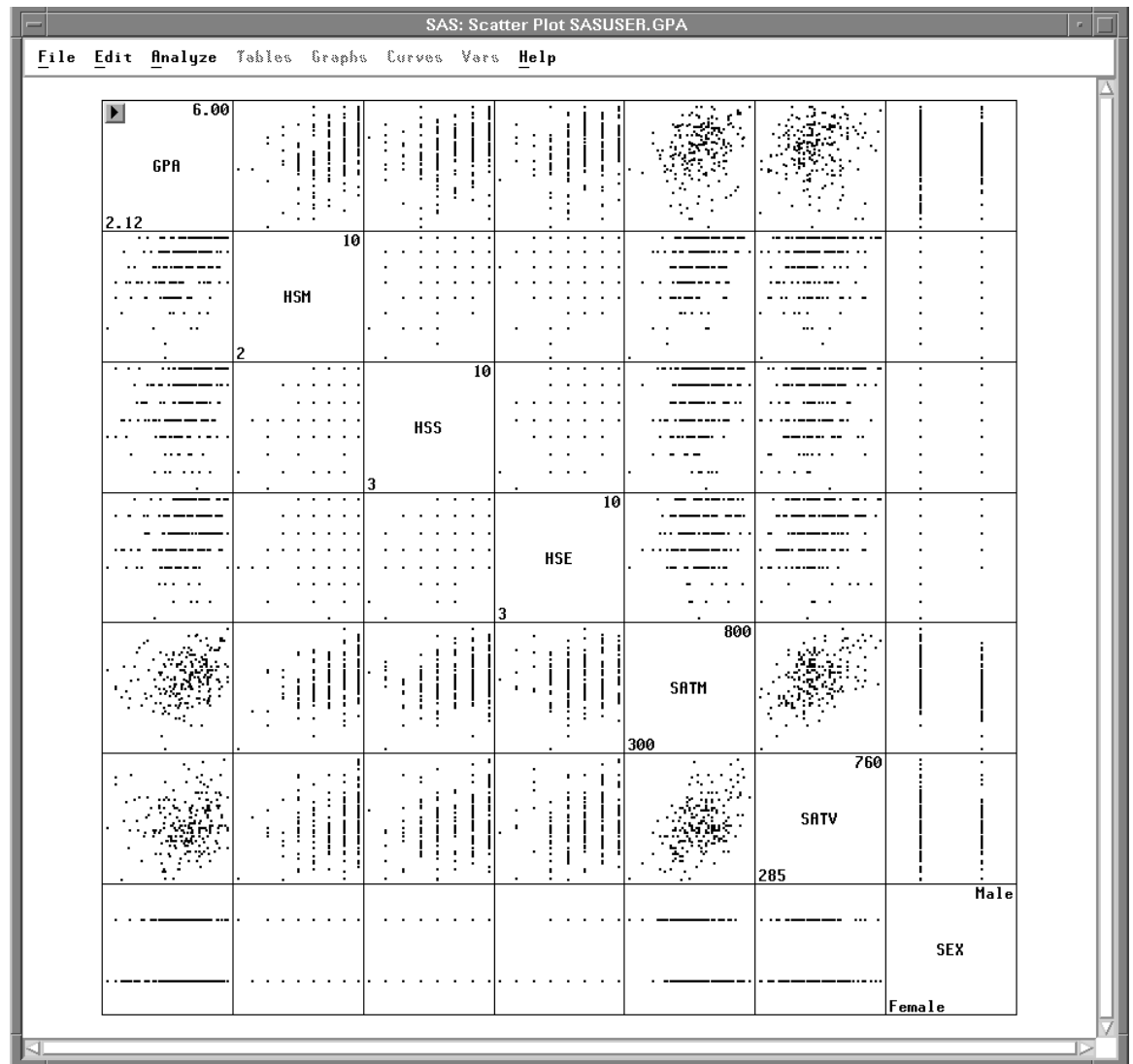


Figure 25.3. Scatter Plot Matrix

Some of these plots show interesting patterns. However, it is difficult to see the plots when they are so small. To change the size of the plots, follow these steps.

⇒ Choose **Edit:Windows:Tools**.

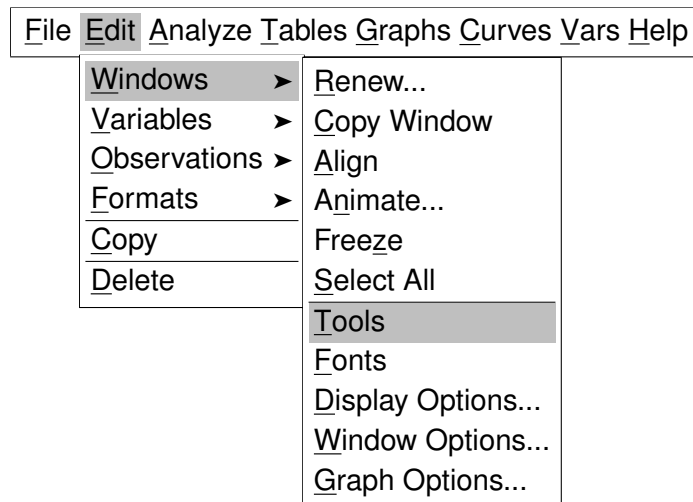


Figure 25.4. Edit:Windows Menu

This displays the Tools window. At the top, the window contains three tools, each indicating a different mode of operation. To select and identify objects, use the arrow. To manipulate objects, use the hand. To zoom, use the magnifying glass.

⇒ **Click on the magnifying glass in the Tools window.**

Now the magnifying glass in the window is highlighted, and the cursor changes from an arrow to a magnifying glass.

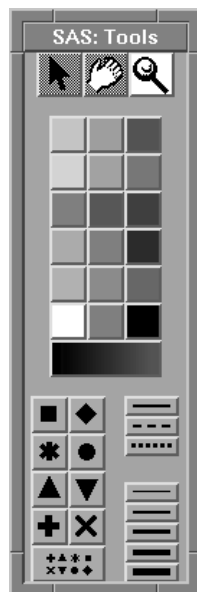


Figure 25.5. Tools Window

⇒ **Move the magnifying glass to the center of the window and click several times.**

When it is near the center of the window, the magnifying glass is large.

Clicking near the center makes objects larger.

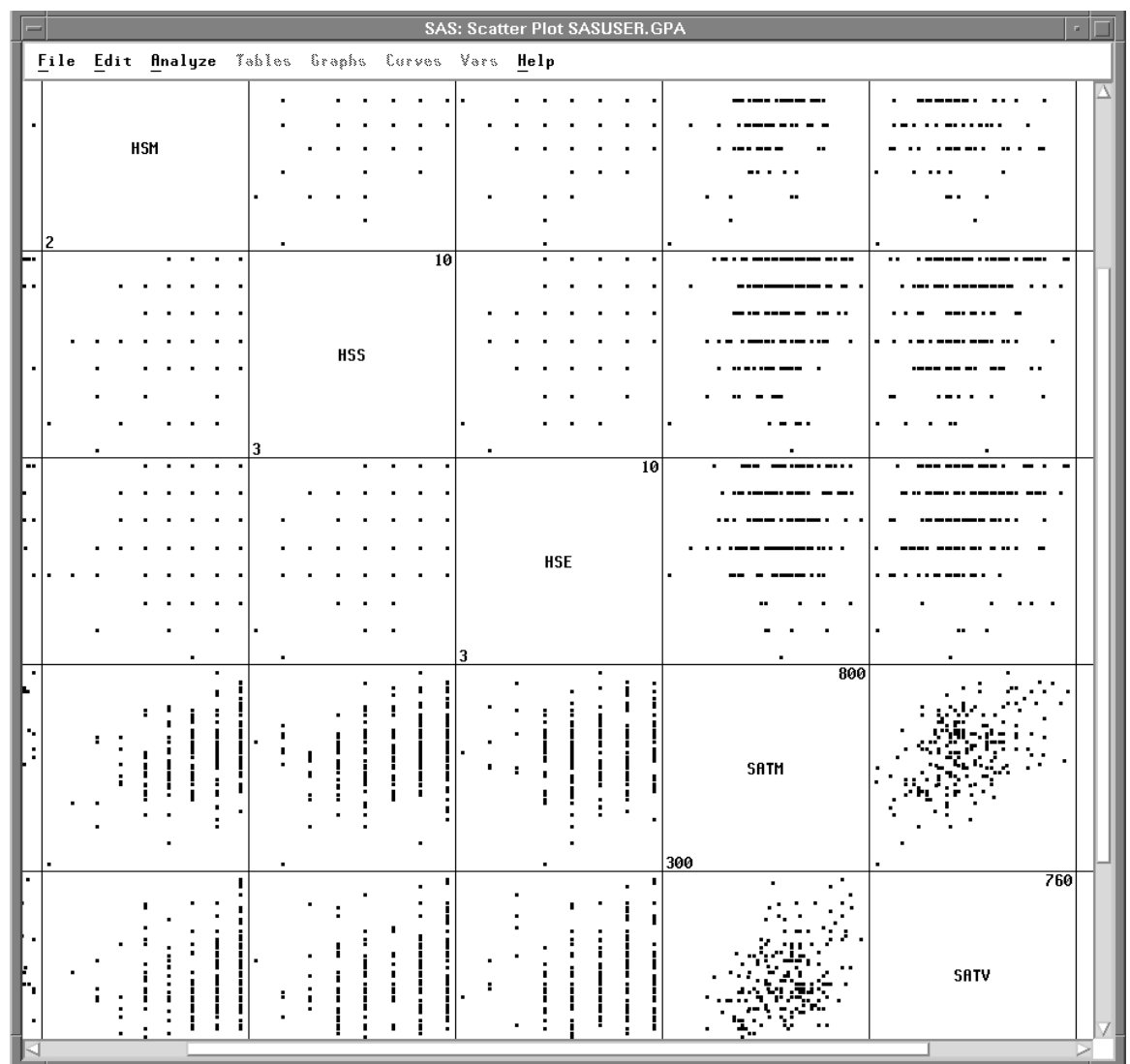


Figure 25.6. Zooming In

⇒ **Move the magnifying glass to the edge of the window and click several times.**

When it is near the edge of the window, the magnifying glass is small.

Clicking near the edge makes objects smaller until all objects fit in the window.

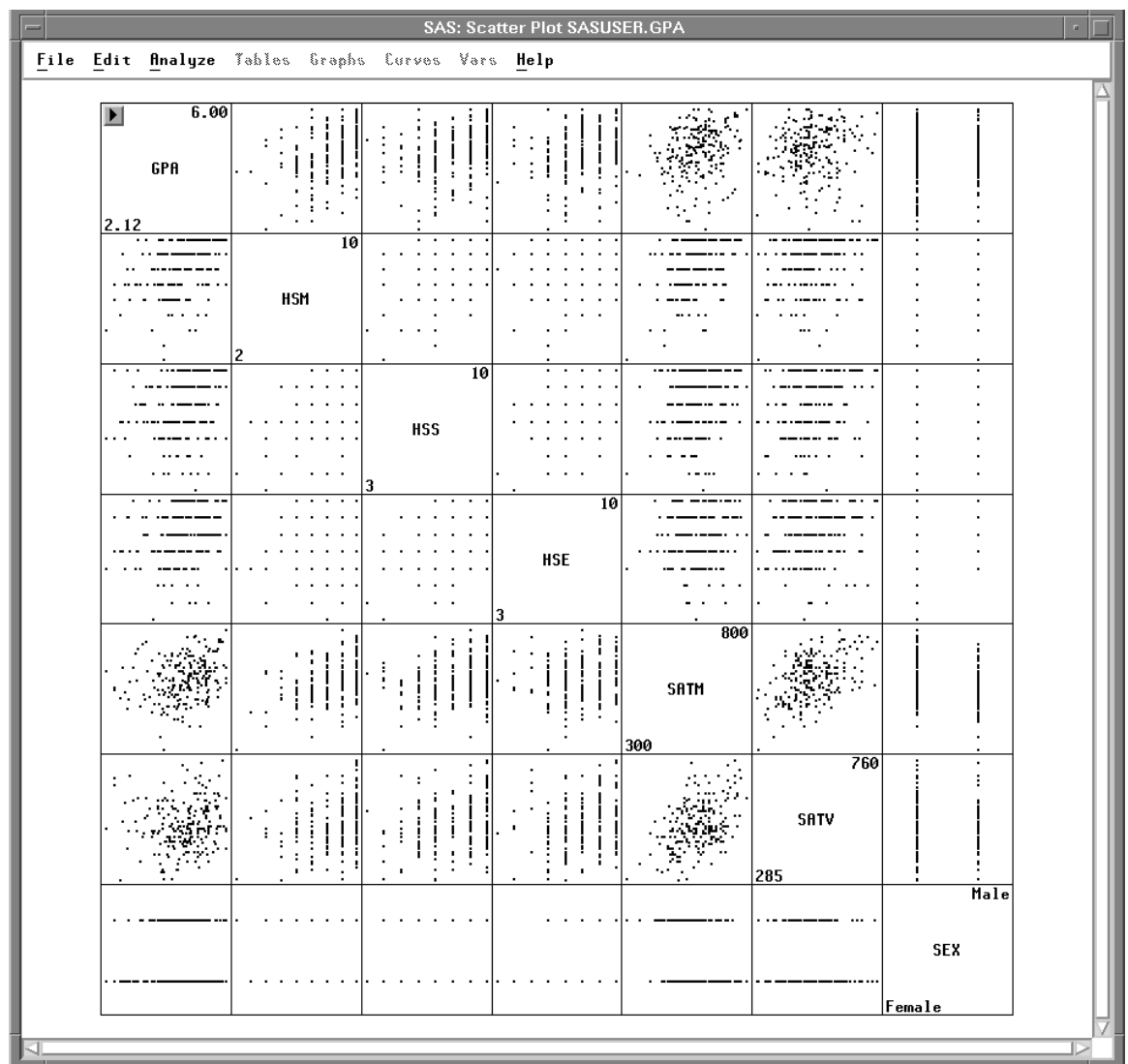


Figure 25.7. Zooming Out

⇒ **Click several times between the center and the edge of the window.**

The degree of magnification is proportional to the distance of your cursor from the center or the edge of the window. Clicking between the center and the edge makes fine adjustments. By clicking in this area, you can give the plots exactly the size you want.

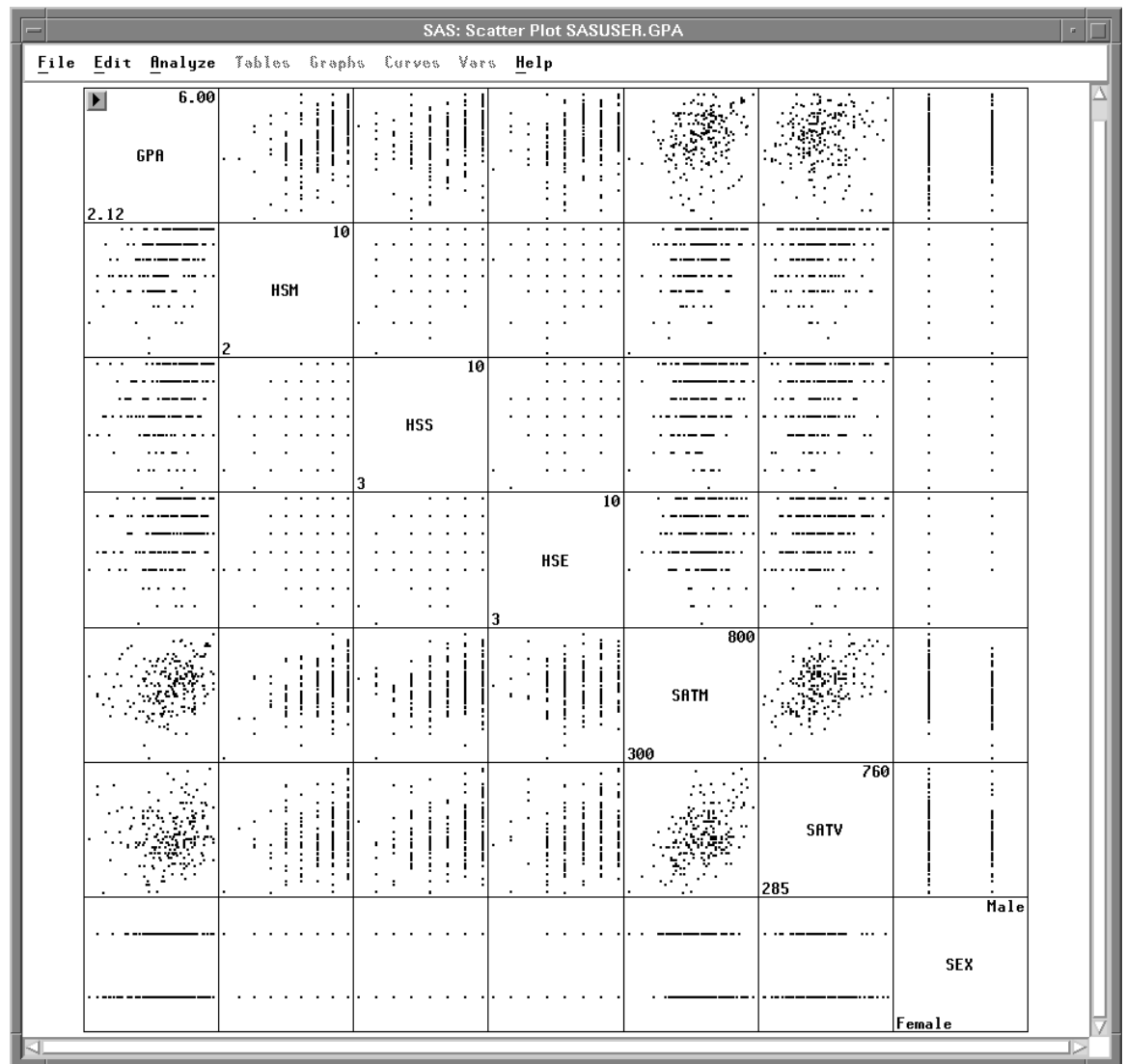


Figure 25.8. Making Fine Adjustments

To zoom in on a specific area, you can drag a rectangle with the magnifying glass.

⇒ **Drag a rectangle around the plot of GPA versus HSM.**

On some hosts, to drag a rectangle it is necessary to begin moving the mouse as soon as you depress the mouse button.

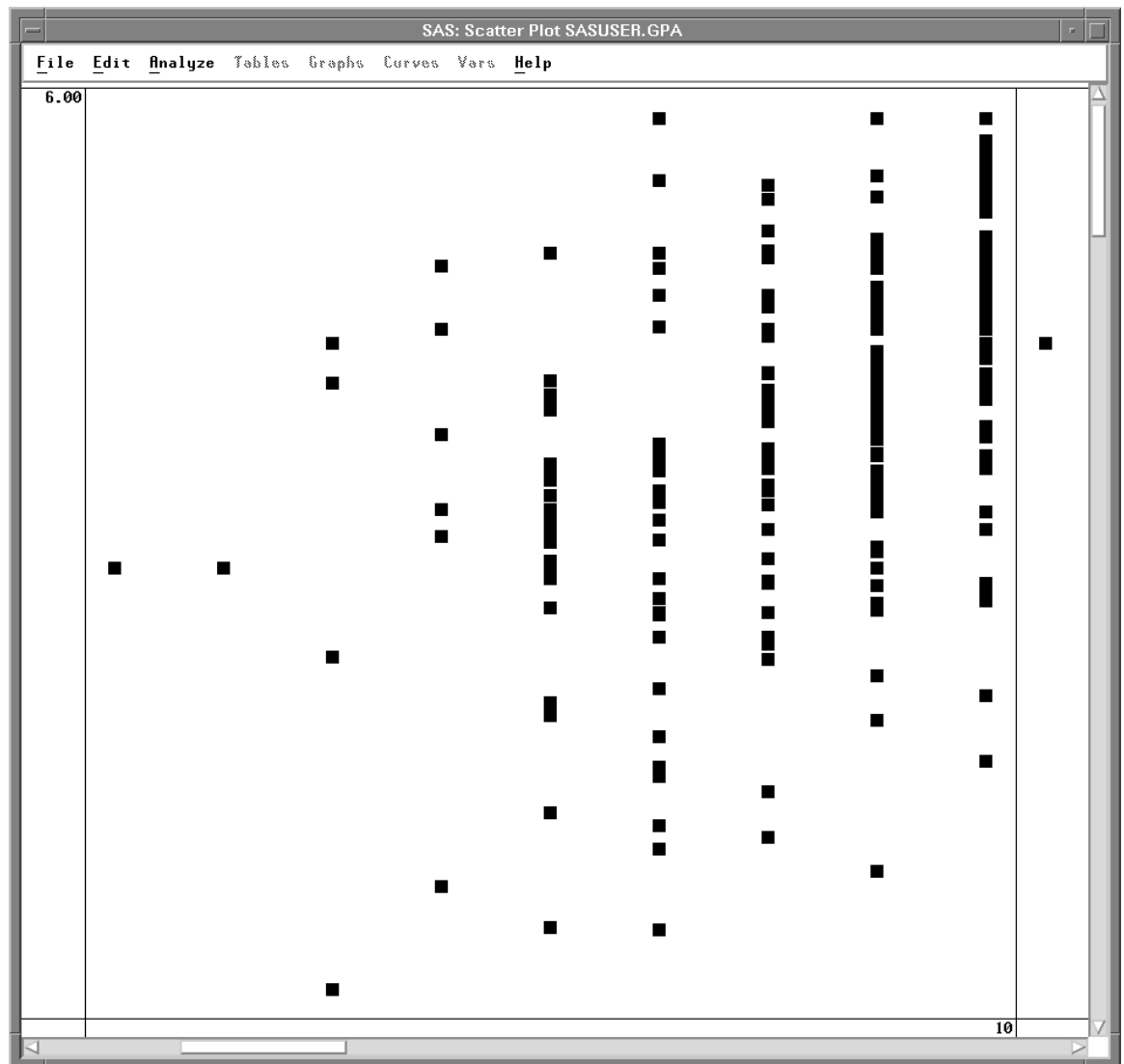


Figure 25.9. Zooming in on **GPA** versus **HSM**

You can restore the original size of the plots by clicking repeatedly near the edge of the window. If you prefer, instead of clicking repeatedly, you can press the mouse button down and hold it down. On most hosts, holding has the same effect as repeated clicks.

When you have zoomed in far, you may find it easier to **Renew** the window, as described in the next section.

Renewing Windows

Renewing restores the original state of the window. *Renewing* also gives you the opportunity to change the variables and options used to create the window.

⇒ **Restore the arrow tool by clicking on the arrow button in the Tools window.**

⇒ **Choose **Edit:Windows:Renew**.**

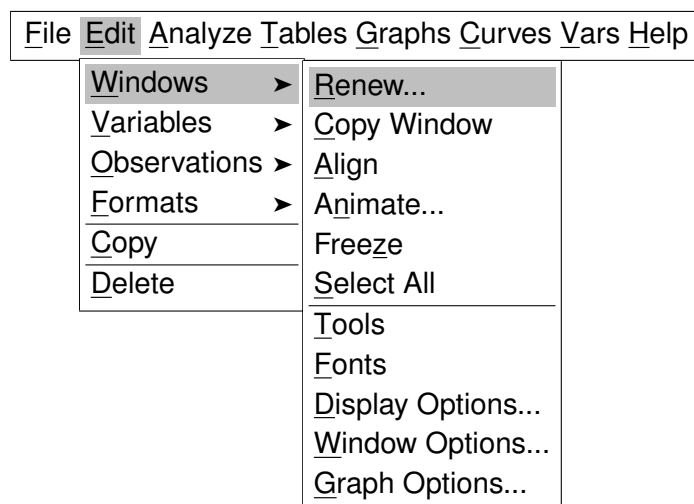


Figure 25.10. Edit:Windows Menu

This displays the Scatter Plot variables dialog used to create the window.

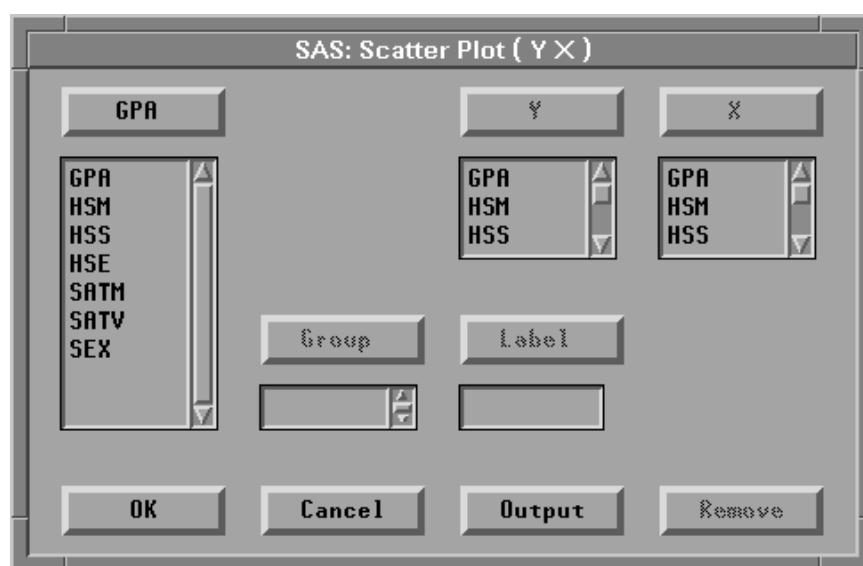


Figure 25.11. Scatter Plot Variables Dialog

⇒ Click **OK** to re-create the scatter plot matrix at its original size, as shown in [Figure 25.3](#).

You can also use **Edit:Windows:Renew** to adjust variables and options associated with your window.

⇒ Choose **Edit:Windows:Renew** again to display the variables dialog

⇒ In the dialog, select **SATM**, **SATV**, and **SEX** in both **Y** and **X** lists.

⇒ Click **Remove** to remove these variables.

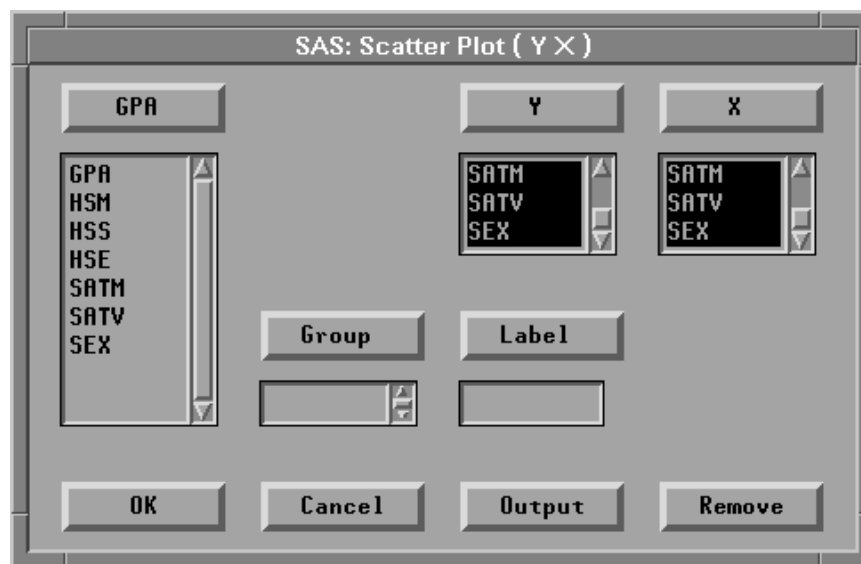


Figure 25.12. Removing Variables

⇒ Click **Output** to display the output options dialog

⇒ In the options dialog, click on the **Labels** button to display variable labels.

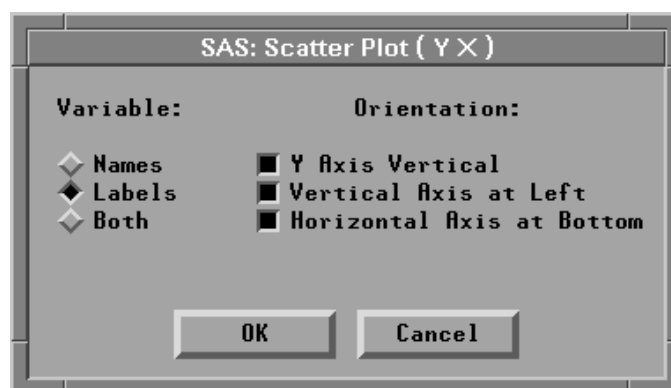


Figure 25.13. Setting Variable Labels

⇒ Click **OK** in both dialogs to renew the window.

The matrix that was seven-by-seven is now four-by-four, and it displays variable labels instead of names.

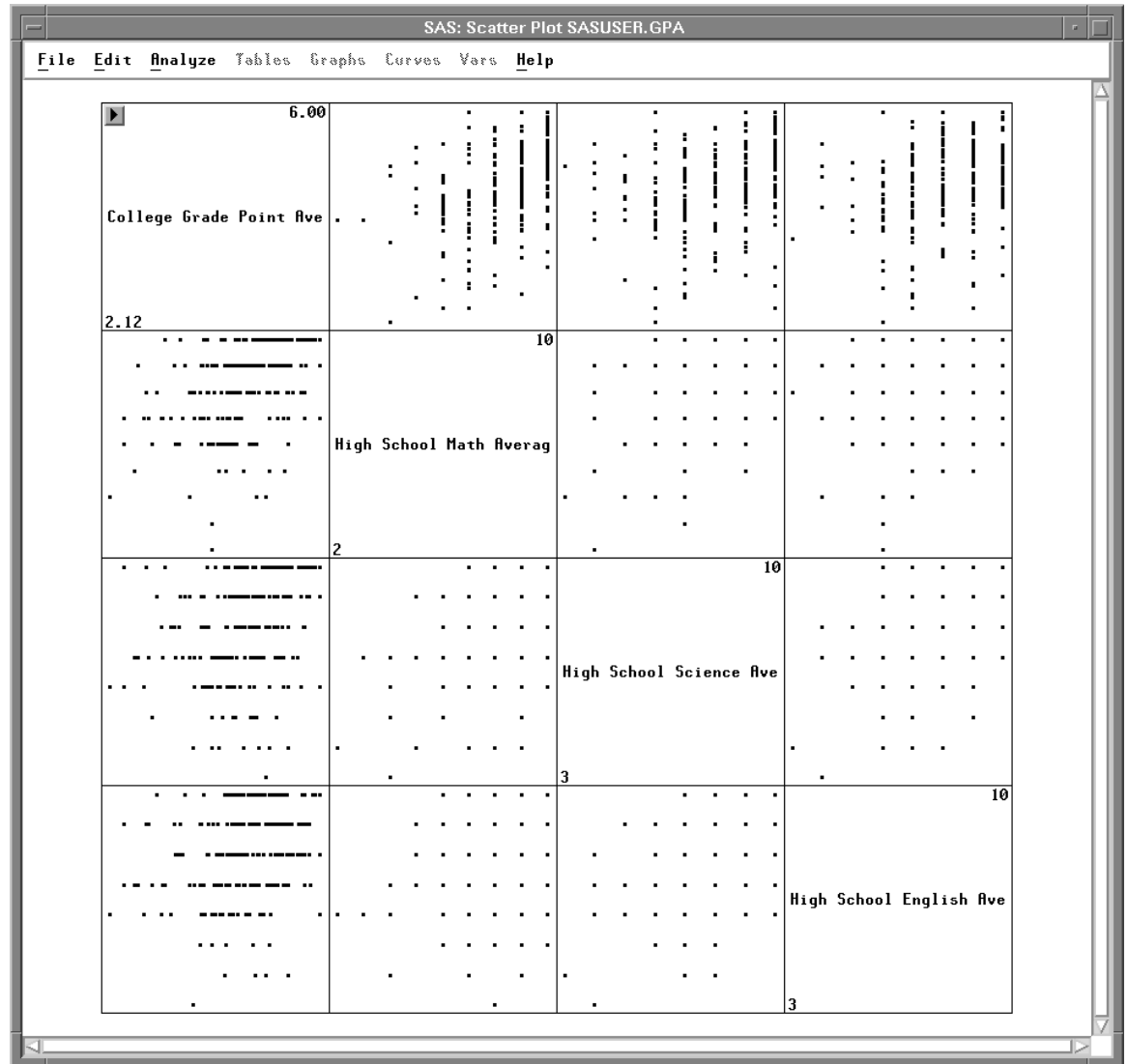


Figure 25.14. Renewed Window

To reset the scatter plot output to display variable names again, follow the same steps to display the scatter plot options dialog, then click on the **Names** button under **Variable:** in the dialog.

⊕ **Related Reading:** Scatter Plot Matrix, [Chapter 5](#), [Chapter 35](#).

Adding and Deleting

Many windows contain **Graphs** and **Tables** menus that enable you to add the most commonly used graphs and tables to any window. For example, in the Fit window you can add residual plots; in the Distribution window you can add tests for distributions.

If a graph you need is not listed in the **Graphs** menu, you can use the **Analyze** menu to add any graph to any window. For example, suppose you want to create a scatter plot with marginal histograms. To create this combination of graphs, first create a distribution analysis on two variables.

⇒ Choose **Analyze:Distribution (Y)**.

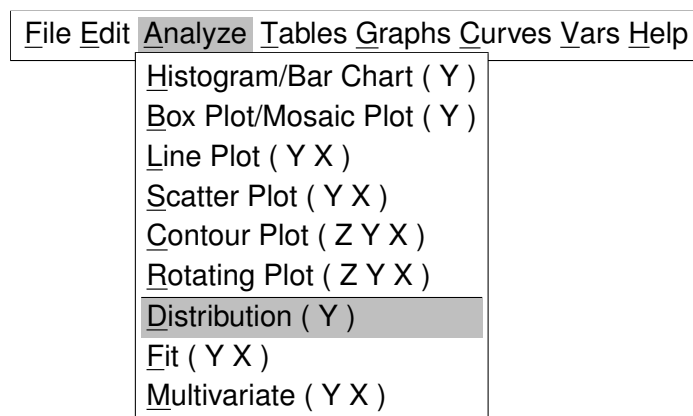


Figure 25.15. Analyze Menu

This displays the Distribution variables dialog.

⇒ Select **GPA** and **HSM**, then click the **Y** button.

This assigns **GPA** and **HSM** the **Y** role in the Distribution analysis.

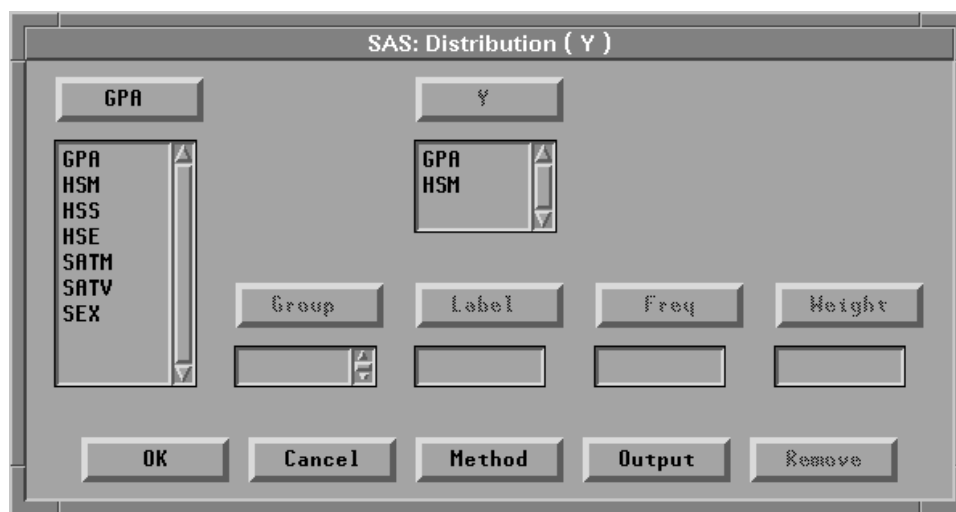


Figure 25.16. Distribution Variables Dialog

⇒ Click the **Output** button.

This displays the output options dialog.

⇒ In the output dialog, turn off all options except **Histogram/Bar Chart**.

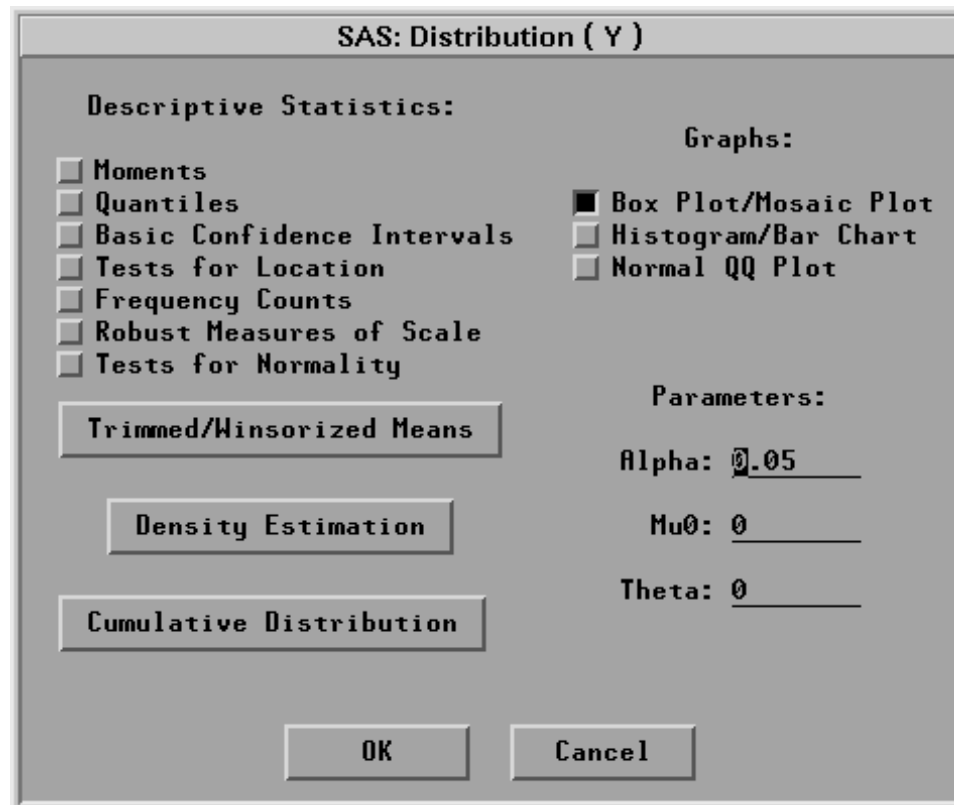


Figure 25.17. Output Options Dialog

⇒ Click **OK** in both dialogs to create the distribution analysis.

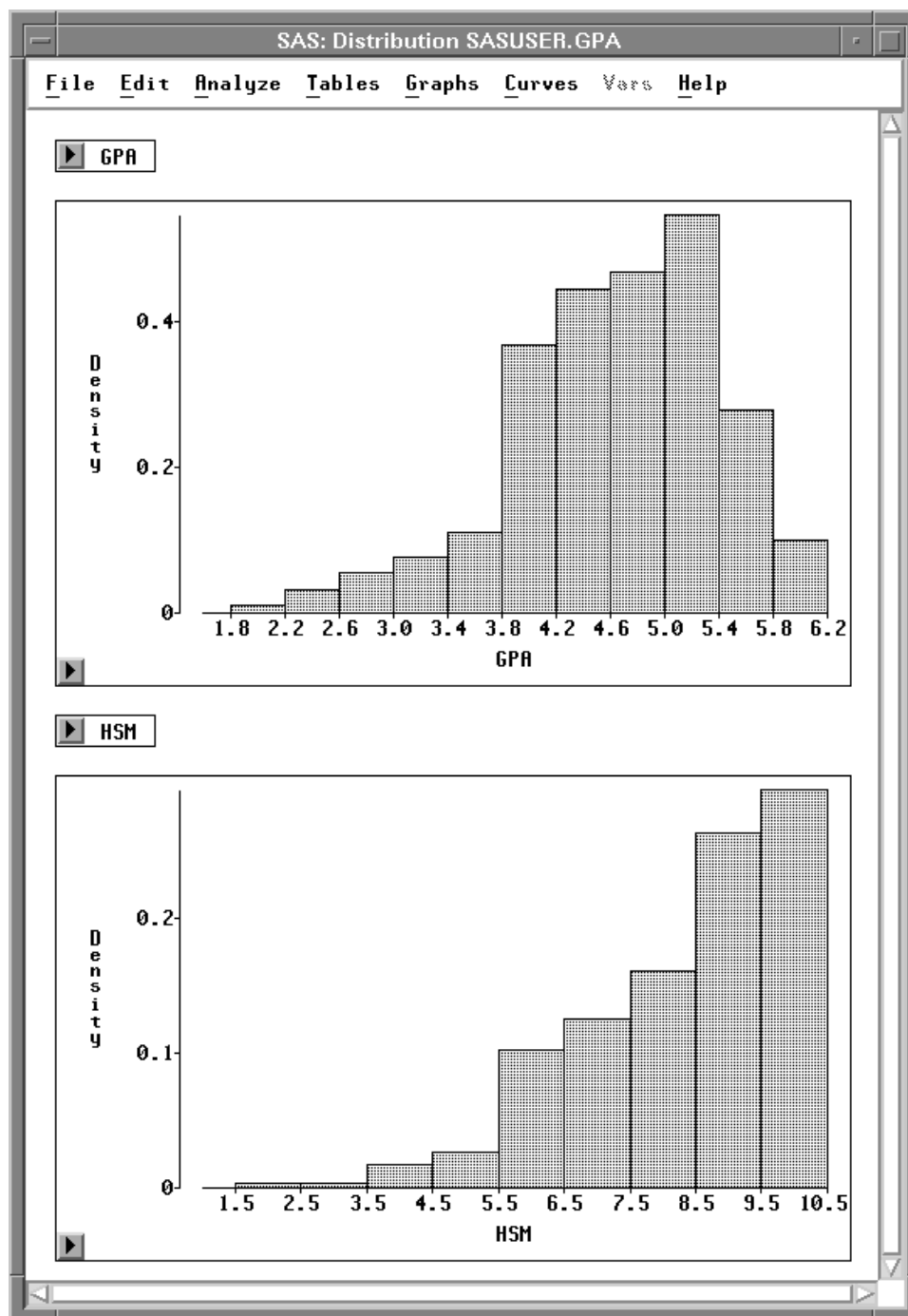


Figure 25.18. Distribution Window

Now you have a distribution window with two histograms. To add a scatter plot of both variables, follow these steps.

⇒ Drag the bottom right corner of the window to the right.

This increases the window size to provide blank space to the right of the histograms.

⇒ **Drag a rectangle to select an area in the window.**

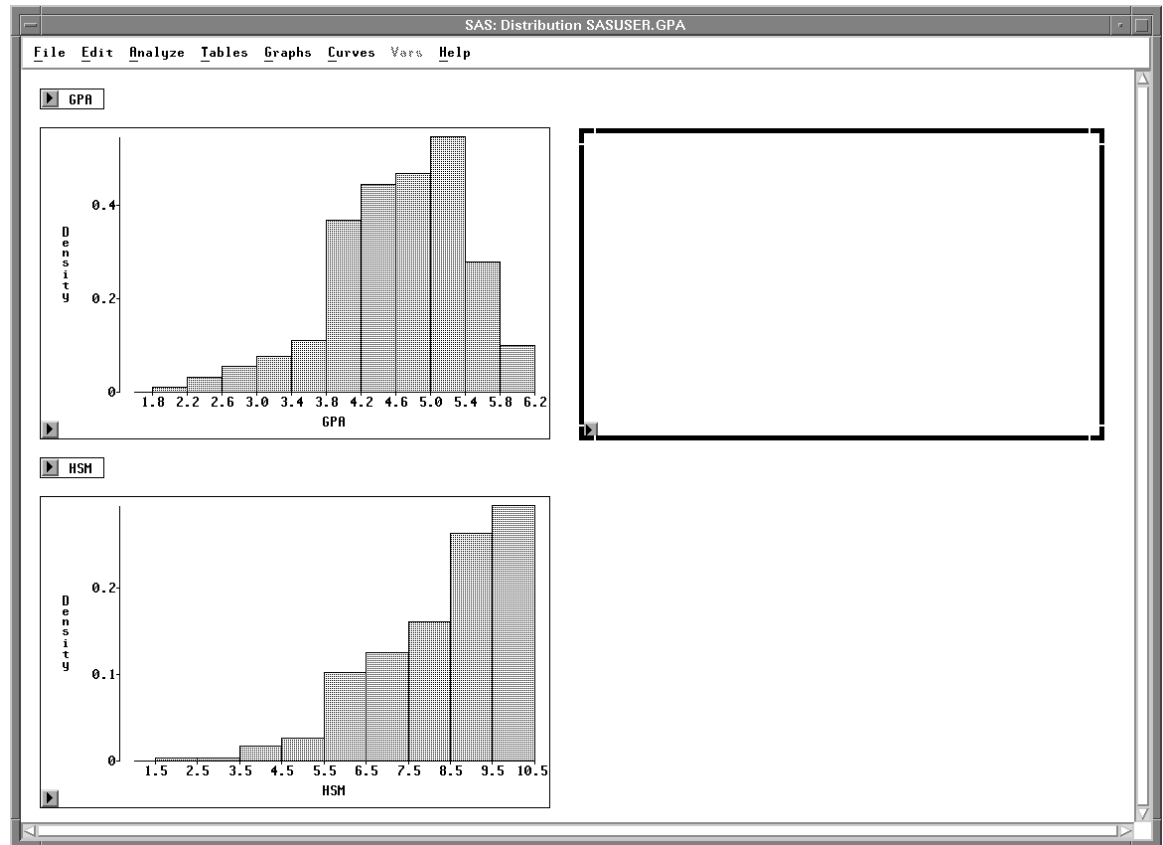


Figure 25.19. Area Selected

⇒ **Choose Analyze:Scatter Plot (Y X).**

This displays the scatter plot variables dialog.

⇒ **In the dialog, assign GPA the Y role, and HSM the X role.**

⇒ **Click OK to add the scatter plot to the distribution window.**

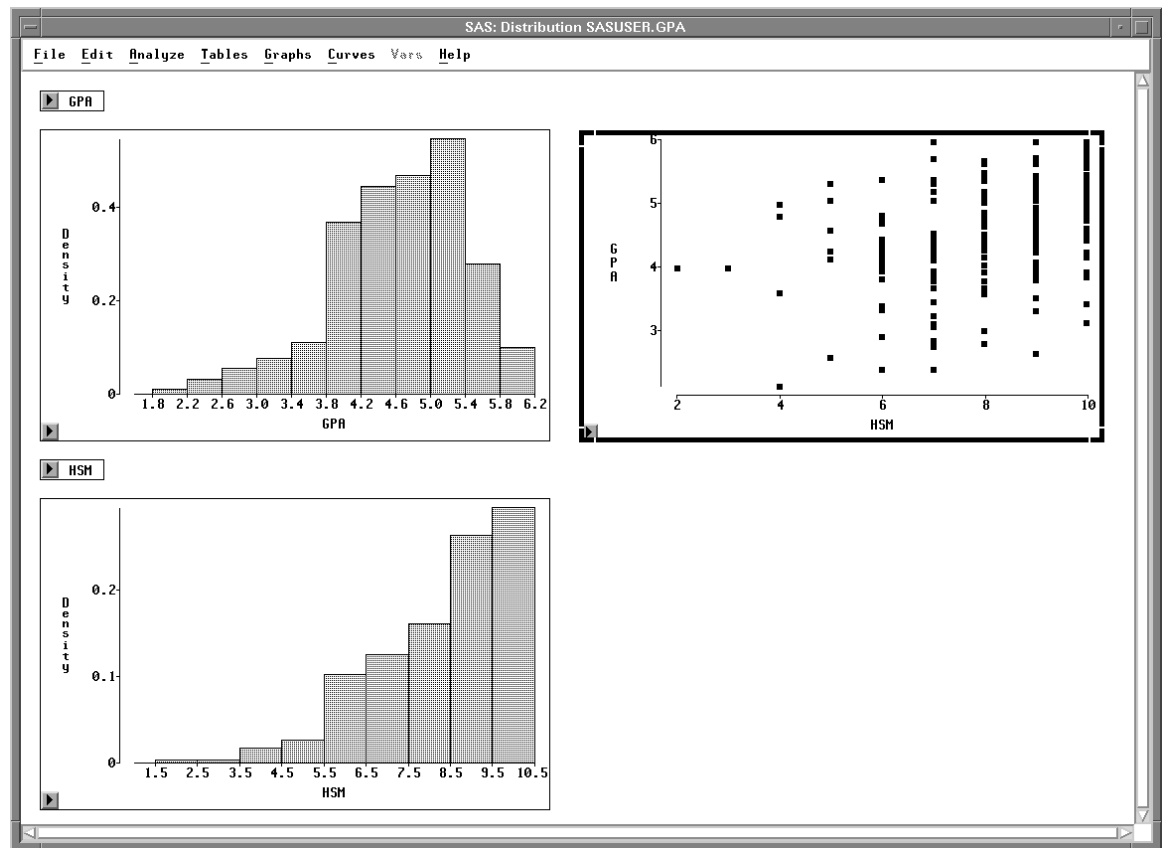


Figure 25.20. Distribution Window with Scatter Plot

You can delete any graph or table in the distribution window. For example, in this window the two small tables that contain variable names are not needed.

- ⇒ Click on any edge of the **GPA** table to select it.
- ⇒ Use extended selection to select the **HSM** table also.

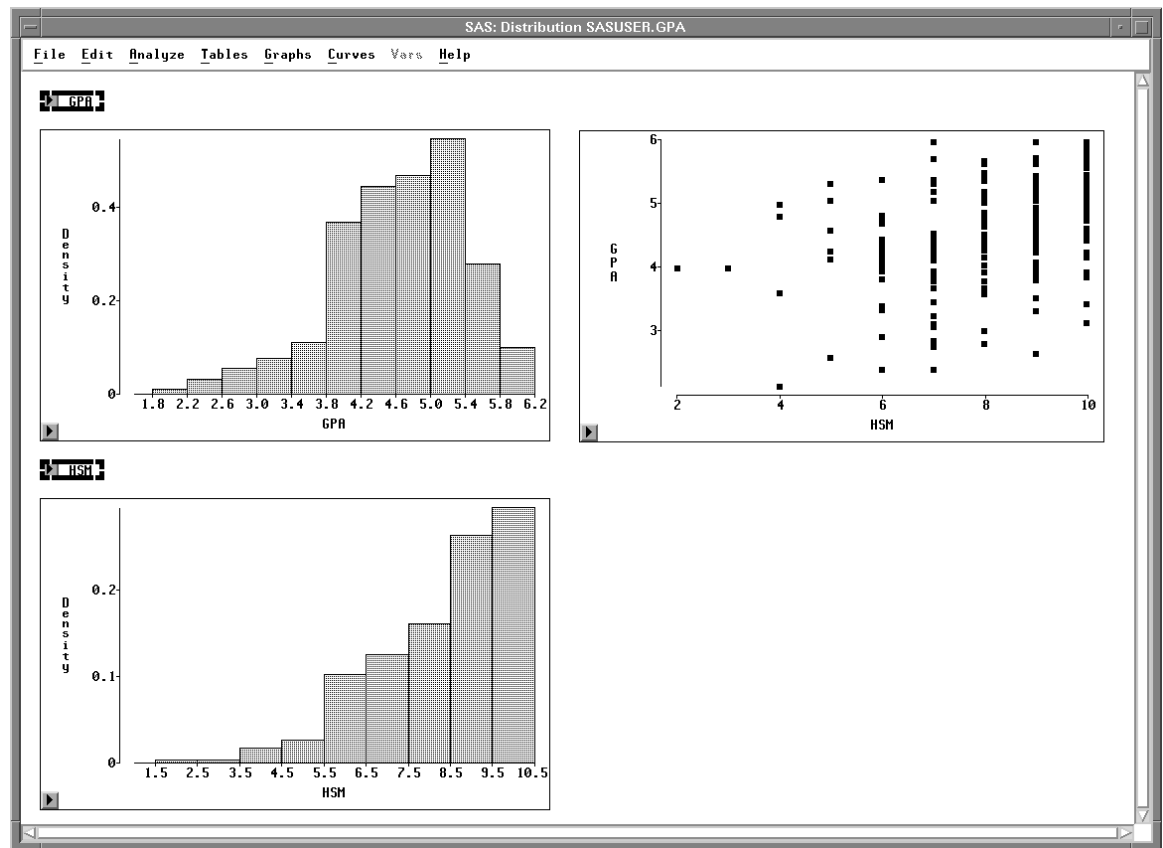


Figure 25.21. Tables Selected

⇒ Choose **Edit:Delete** to delete the tables.

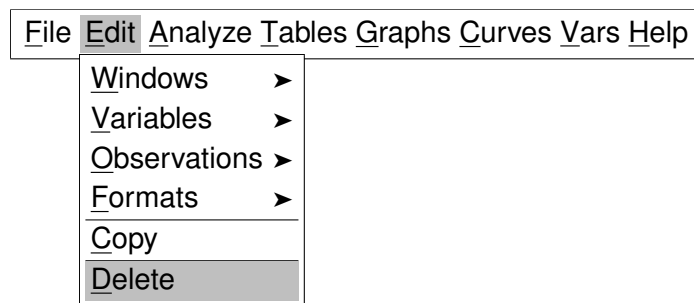


Figure 25.22. Edit:Windows Menu

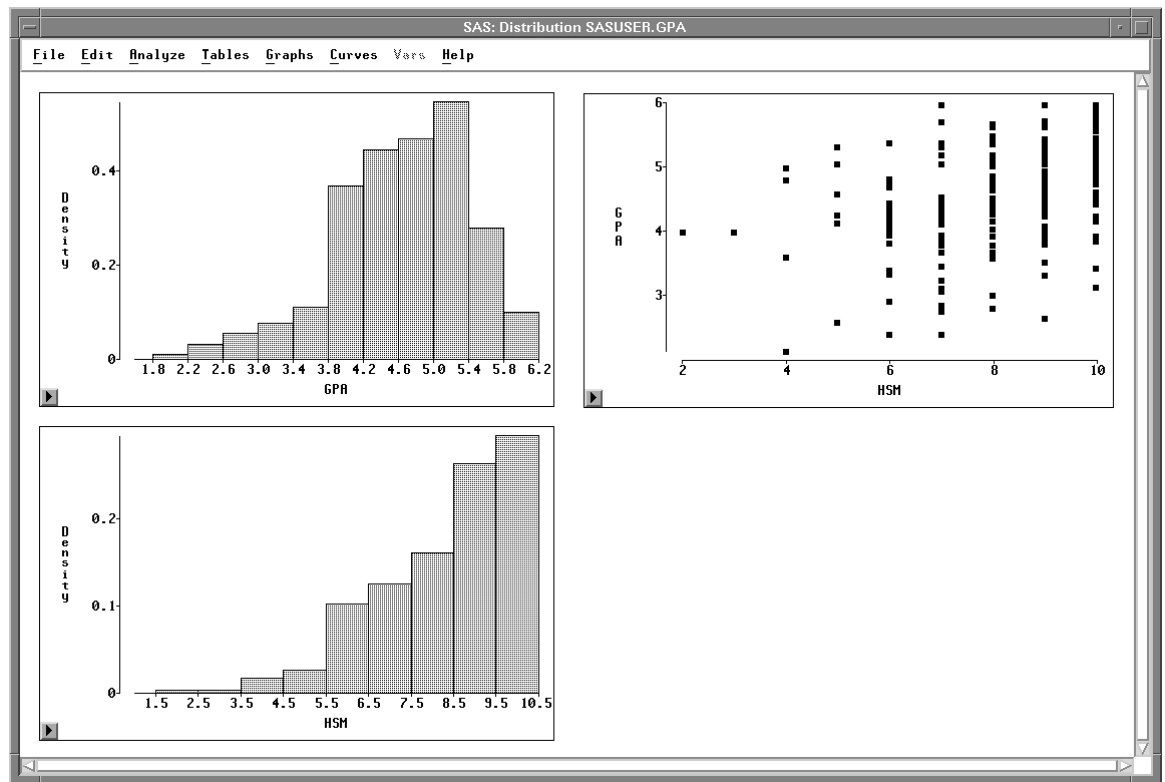


Figure 25.23. Tables Deleted

By choosing from the **Analyze** menu and choosing **Edit:Delete**, you have created a window containing one scatter plot and two histograms. In the same manner, you can add any graph and delete any graph or table in a window.

Moving and Sizing

Now you have a window containing one scatter plot and two histograms. To make marginal histograms, you should position the graphs so that common axes are parallel.

You can move any graph or table by dragging on its side.

⇒ **Drag the HSM histogram below the scatter plot.**

Press the mouse button down on any side of the histogram. Move the mouse to the right. Release the mouse button when you have the histogram positioned below the scatter plot.

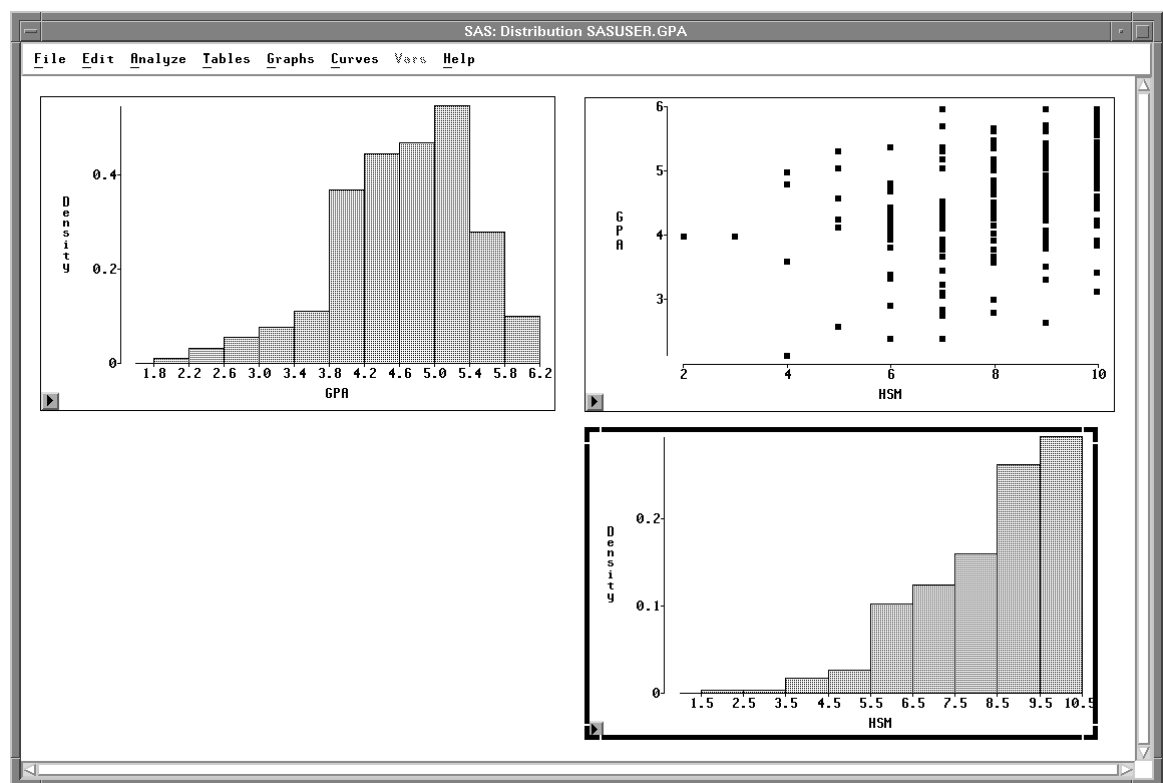


Figure 25.24. Histogram Moved

Now the histogram is in approximately the right place, but it is too large and its orientation is wrong. A marginal histogram should be smaller and the bars should be pointing downward.

You can resize and reorient any graph by dragging on a corner.

⇒ **Drag the lower right corner of the HSM histogram upward.**

Press the mouse button down on the lower right corner. Move the mouse upward. Release the mouse button when the histogram is about half its original size.

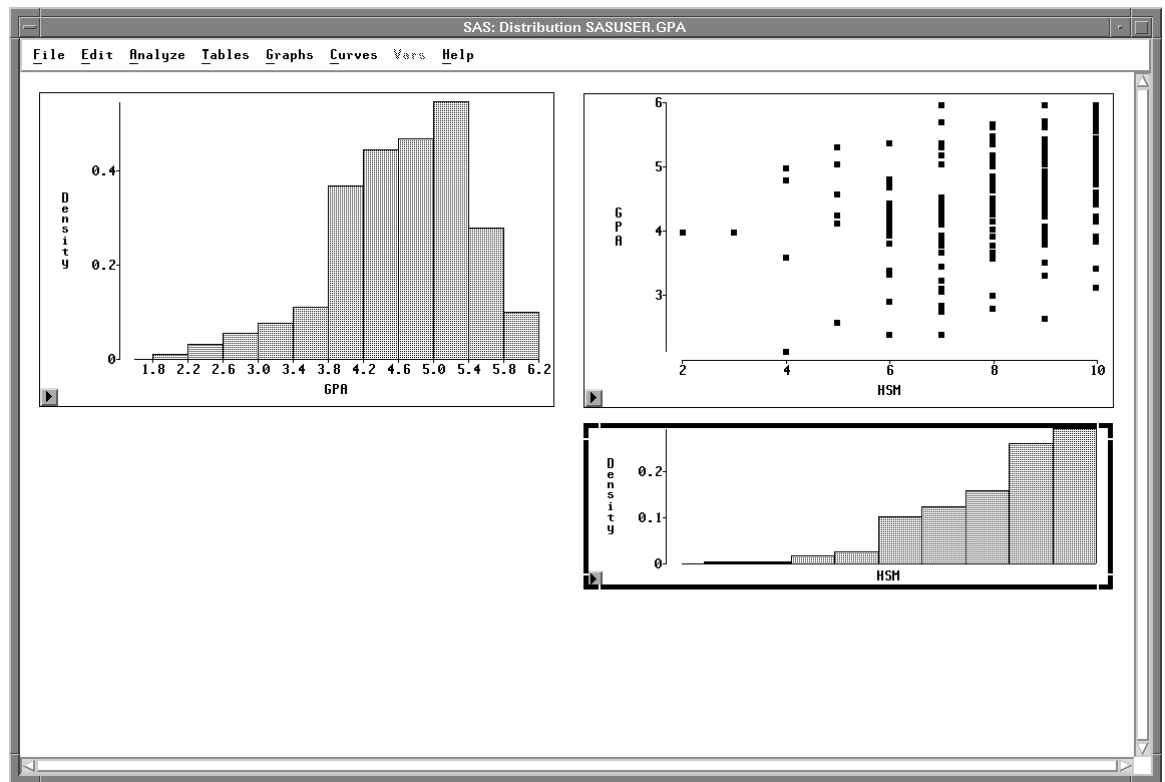


Figure 25.25. Histogram Resized

To change the orientation of the histogram, you can flip it over by dragging one corner across another.

- ⇒ **Drag the upper right corner down past the lower right corner.**
This flips the histogram so that the bars are pointing downward.

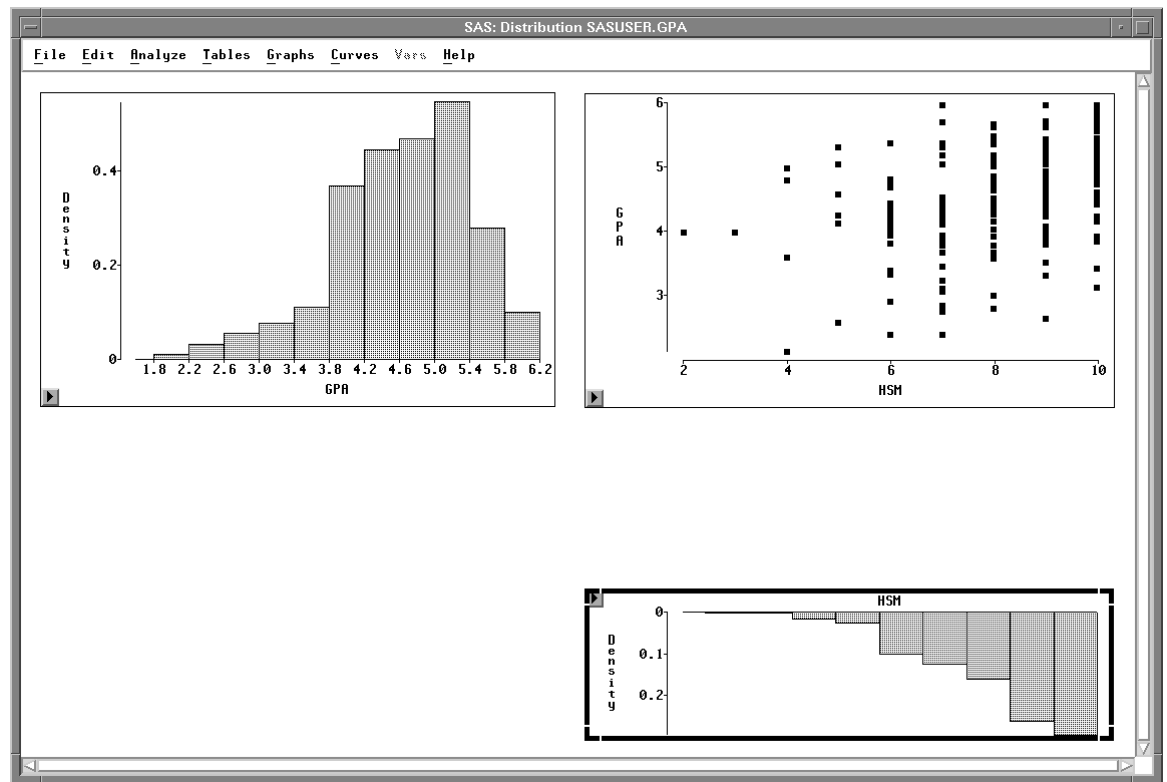


Figure 25.26. Histogram Reoriented

Now you have a scatter plot and one marginal histogram. To orient the other histogram correctly requires two flips.

- ⇒ **Drag the upper left corner of the GPA histogram past the lower right corner.**
 This flips the histogram across its diagonal. The bars that were vertical are now horizontal.

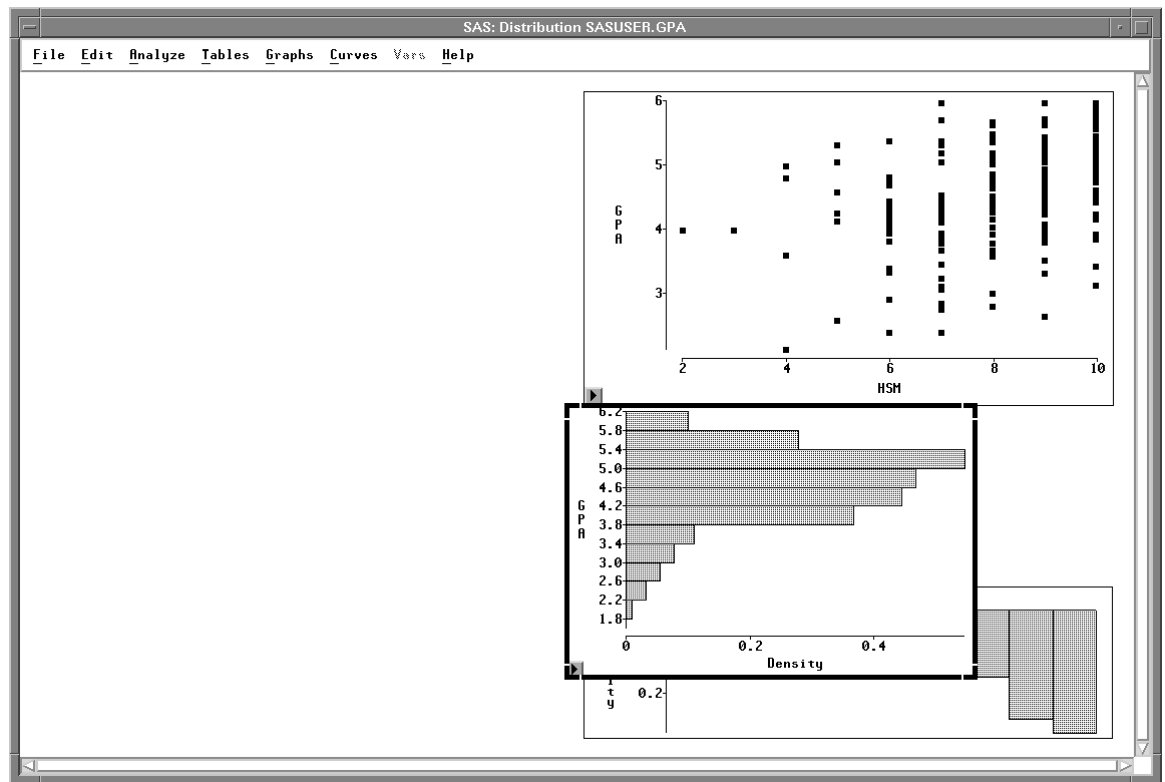


Figure 25.27. Histogram Reoriented

⇒ **Drag the upper right corner left past the upper left corner.**
 This flips the histogram so that the bars are pointing to the left.

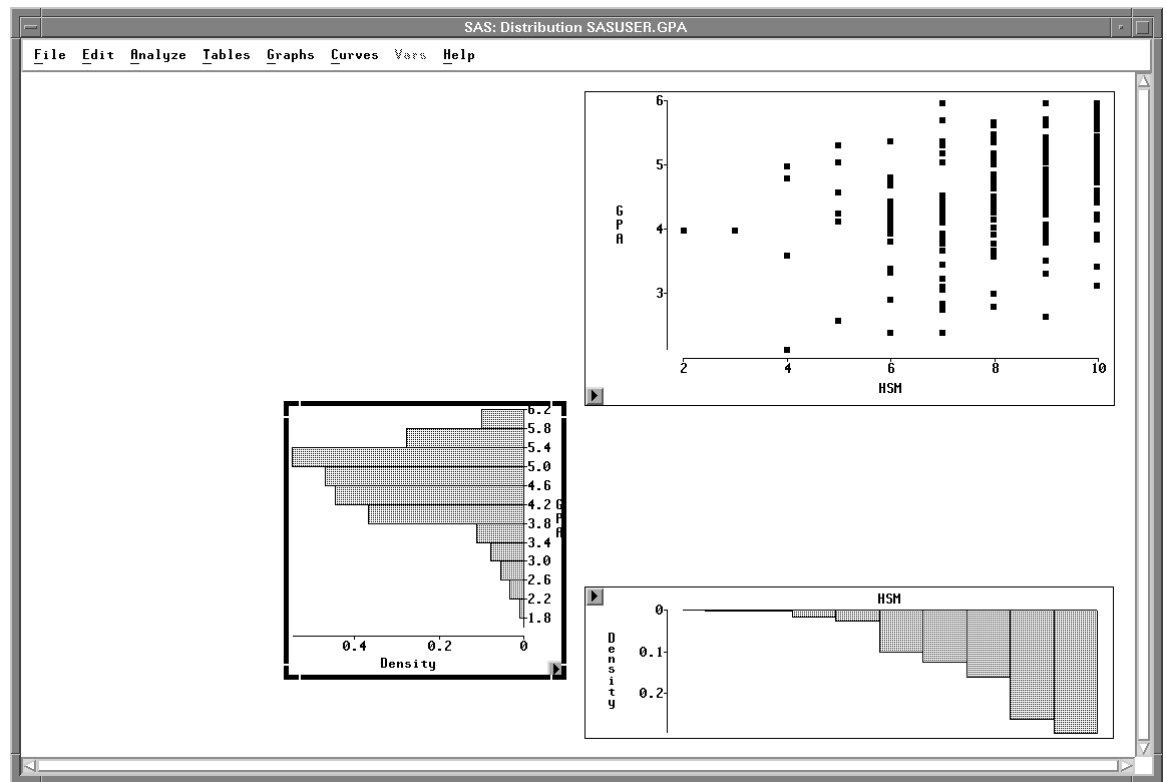


Figure 25.28. Histogram Reoriented

⇒ **Size and move both histograms to the margins of the scatter plot.**

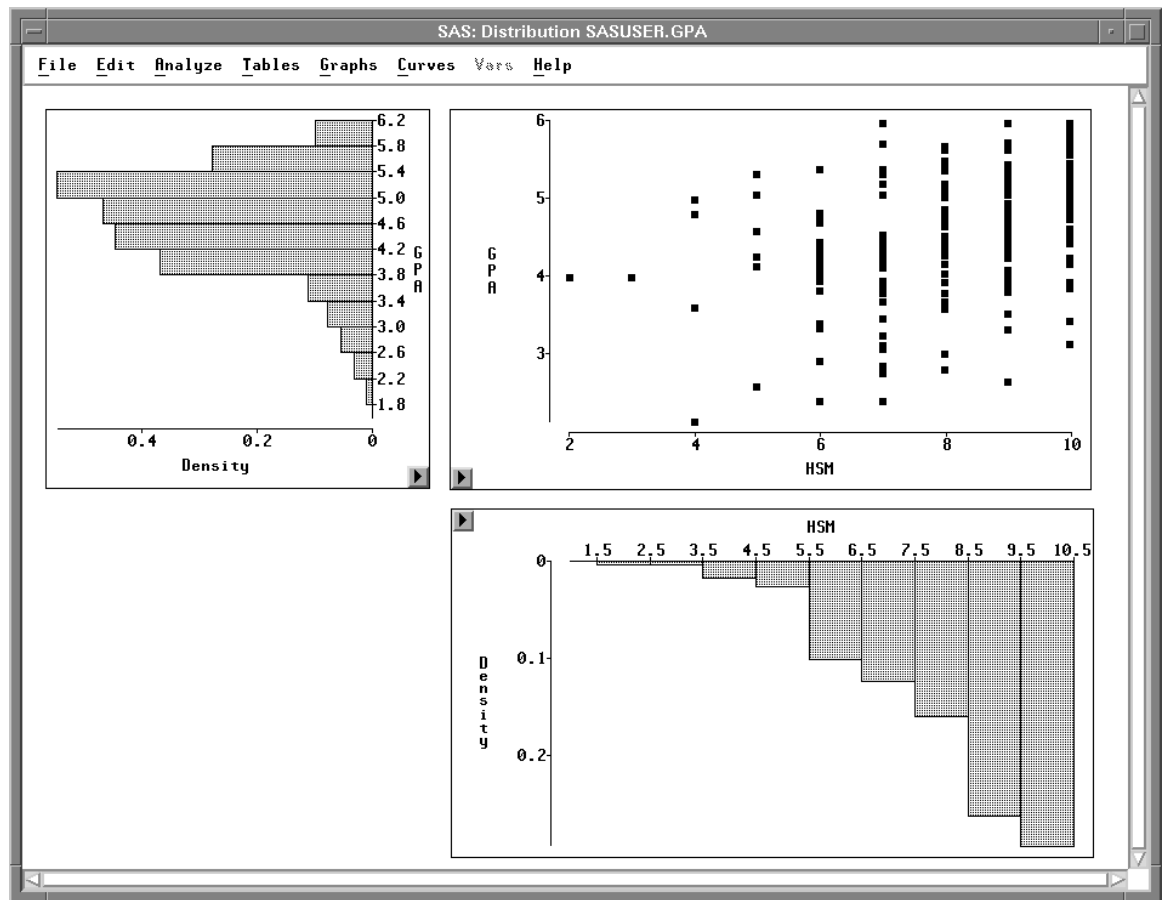


Figure 25.29. Scatter Plot with Marginal Histograms

Now both histograms are correctly oriented and placed at the margins of the scatter plot.

Aligning Graphs

Now that you have created a scatter plot with marginal histograms, you may notice that the axes are not perfectly aligned. For example, the tick label 1.5 in the **HSM** histogram appears to the right of the tick label 2 in the scatter plot. Similarly, the tick label 6.20 in the **GPA** histogram appears below the tick label 6.00 in the scatter plot. This occurs because, by default, axes are chosen to maximize the display of the data. You can override this behavior to align axes in different graphs.

⇒ **Click once in any empty area to deselect the histogram.**

⇒ **Choose Edit:Windows:Align.**

This aligns the **HSM** and **GPA** axes in all graphs.

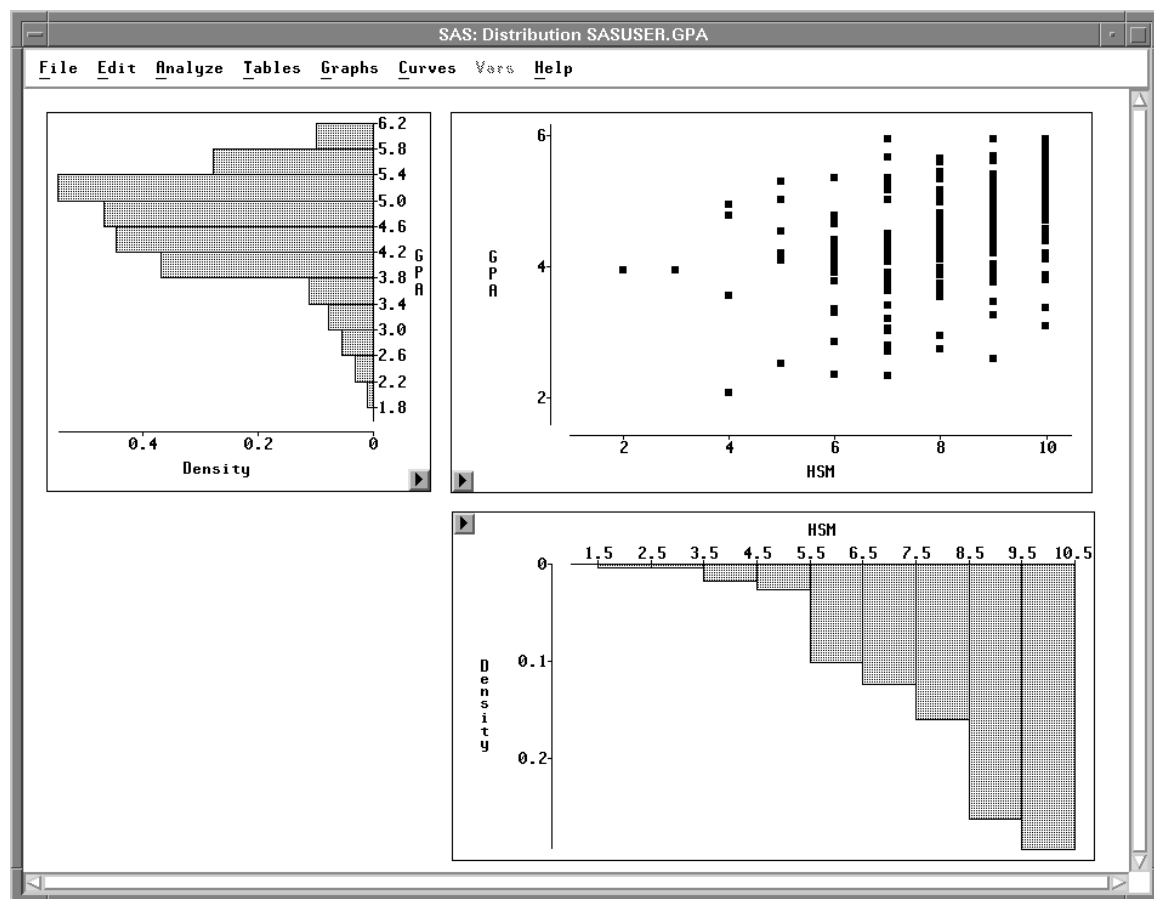


Figure 25.30. Graphs Aligned

You can align any axes that display the same variable. When you do not want to align all axes in a window, select the axes of interest before choosing **Edit:Windows:Align**.

Once you have moved, sized, added, deleted, and aligned objects in your windows, you will often want to save and print them. The next three chapters describe how to save and print data, graphs, and tables.

Chapter 26

Saving and Printing Data

Chapter Contents

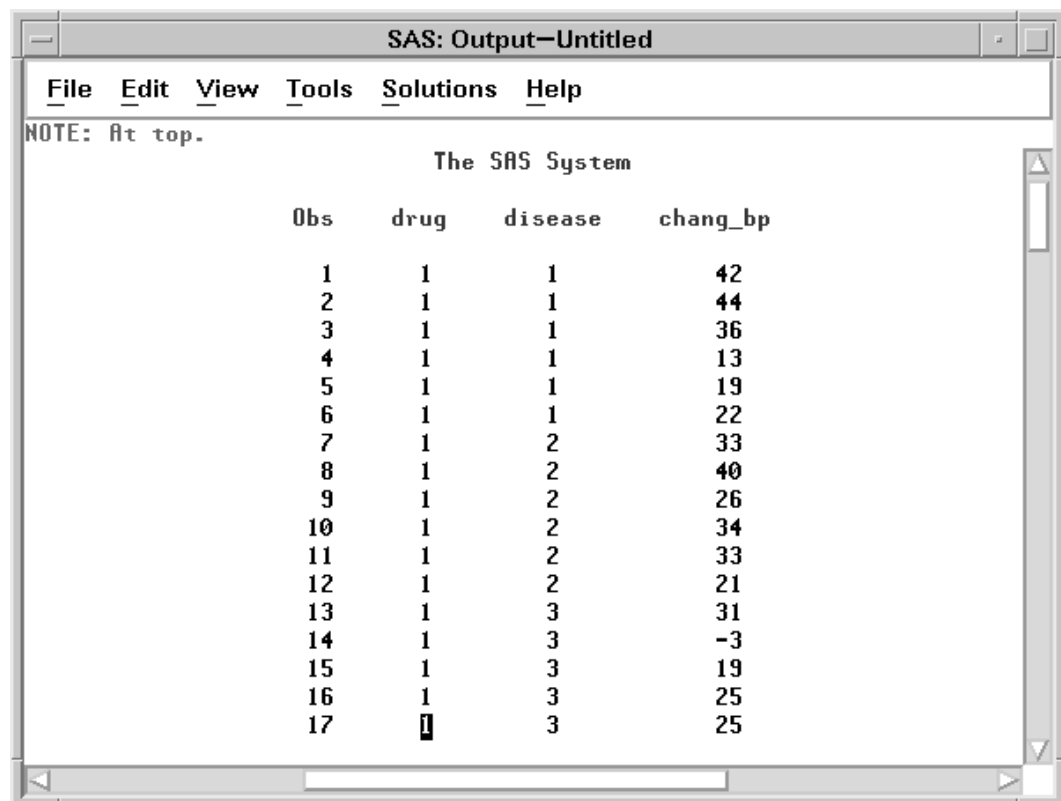
SAVING DATA	422
PRINTING DATA	425

Chapter 26

Saving and Printing Data

Saving a data set means copying the data from a SAS/INSIGHT data window to a SAS data set. SAS/INSIGHT software automatically saves observation colors, markers, and other states as a variable in the SAS data set.

You can print the contents of any SAS data set to the SAS Output window and to a hardcopy device by using the PRINT procedure.



NOTE: At top.

The SAS System

Obs	drug	disease	chang_bp
1	1	1	42
2	1	1	44
3	1	1	36
4	1	1	13
5	1	1	19
6	1	1	22
7	1	2	33
8	1	2	40
9	1	2	26
10	1	2	34
11	1	2	33
12	1	2	21
13	1	3	31
14	1	3	-3
15	1	3	19
16	1	3	25
17	1	3	25

Figure 26.1. PROC PRINT Output

Saving Data

All data analysis in SAS/INSIGHT software uses a copy of a SAS data set stored in memory. Since your original SAS data set is not stored in memory, it is not affected by changes you make in the data window.

When you save the data, you copy the data in memory to a SAS data set stored on disk. Saving the data makes a copy of

- all data values, including any you have edited with the **Data:Fill** menu
- all variables and observations, including any you have created
- measurement levels for up to 250 variables
- all observation states, including color, marker shape, show/hide, include/exclude, label/nolabel, and select states

Observation states are stored in a special variable **_OBSTAT_** that is automatically read in the next time you open the data set. Thus, if you have colored, marked, hidden, excluded, and labeled observations, you can save all these states, exit SAS/INSIGHT software, and invoke SAS/INSIGHT software again later without losing your work. You can also set the values of the **_OBSTAT_** variable to initialize observation states. For an example of this, see [Chapter 30, “Working with Other SAS Products.”](#)

The following steps illustrate how to save data to a SAS data set.

⇒ **Open the DRUG data set.**

	Int	Int	Int
72	DRUG	DISEASE	CHANG_BP
1	1	1	42
2	1	1	44
3	1	1	36
4	1	1	13
5	1	1	19
6	1	1	22
7	1	2	33
8	1	2	40
9	1	2	26
10	1	2	34

Figure 26.2. DRUG data

⇒ **Choose File:Save:Data.**

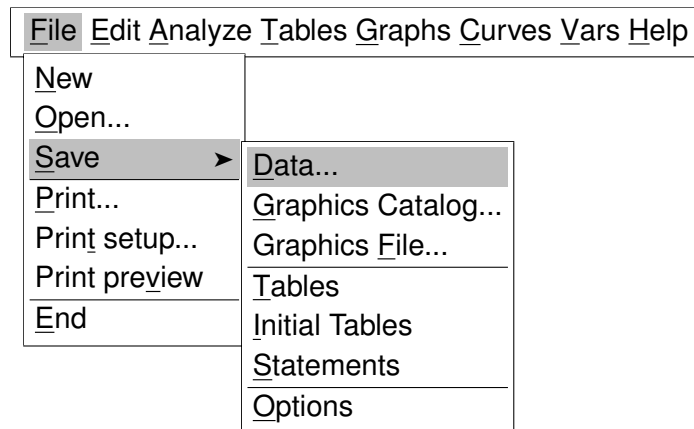


Figure 26.3. File:Save Menu

This displays a dialog. By default, the data set you save to has the same name as the data window in your SAS/INSIGHT session. If you prefer, you can select another library and enter another data set name in the dialog.

⇒ **Click OK to save the data.**

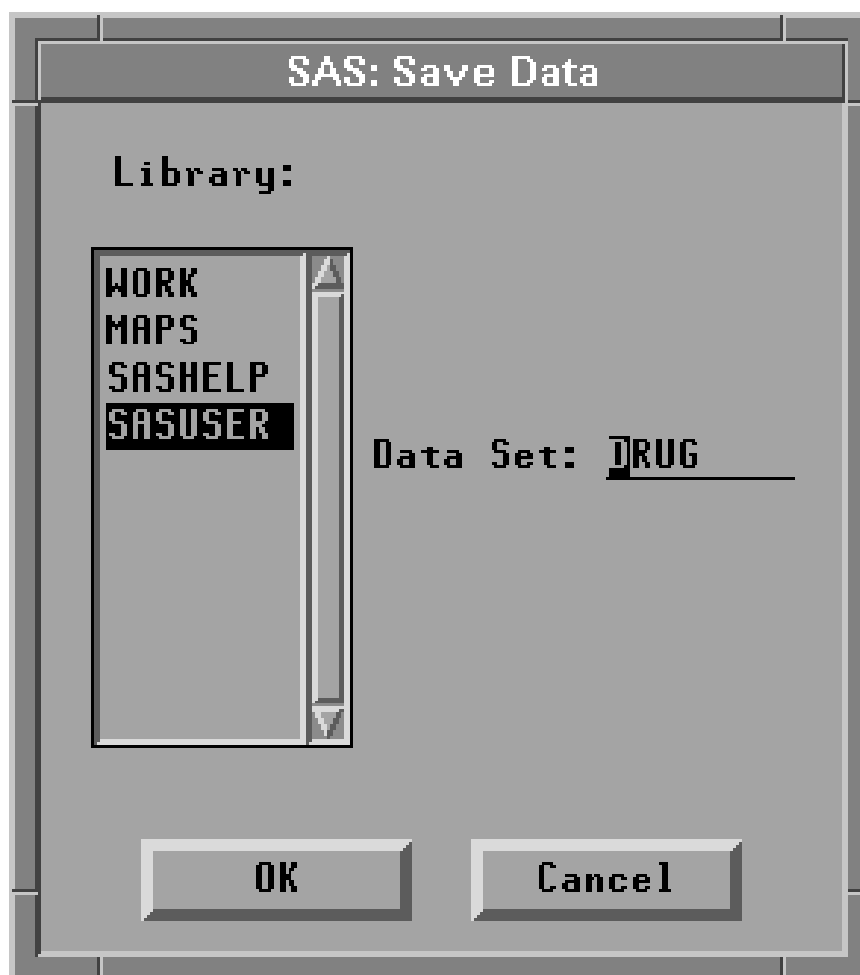


Figure 26.4. Save Dialog

Printing Data

You can print the contents of the data window by saving it as a SAS data set and using the PRINT procedure. PROC PRINT sends its output to the Output window. You can send the contents of the Output window to a file or printer.

⇒ **Enter a PROC PRINT statement in the Program Editor.**

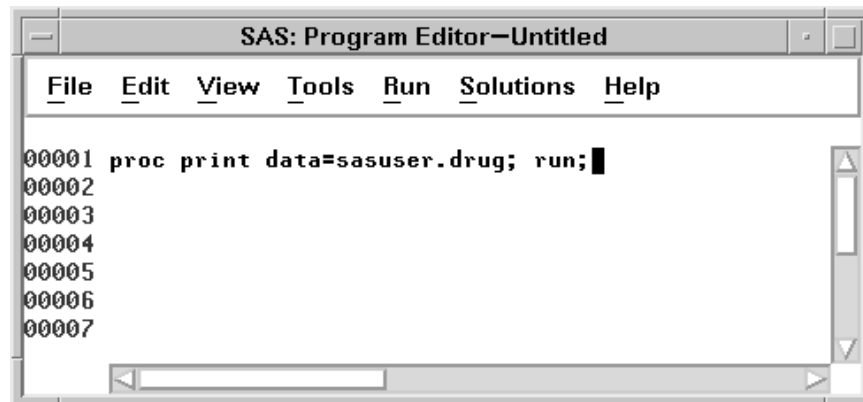


Figure 26.5. Program Editor

⇒ **Choose Run:Submit.**

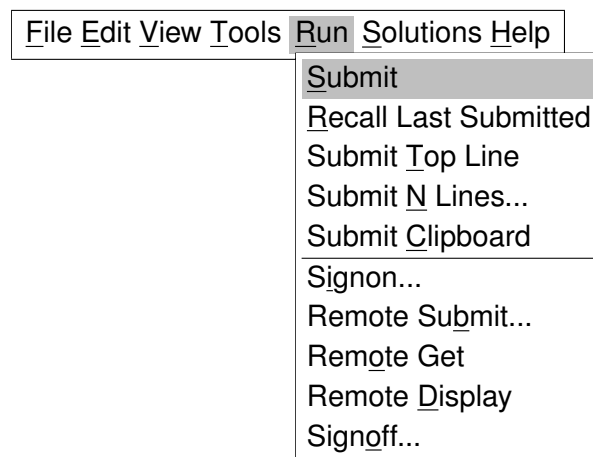
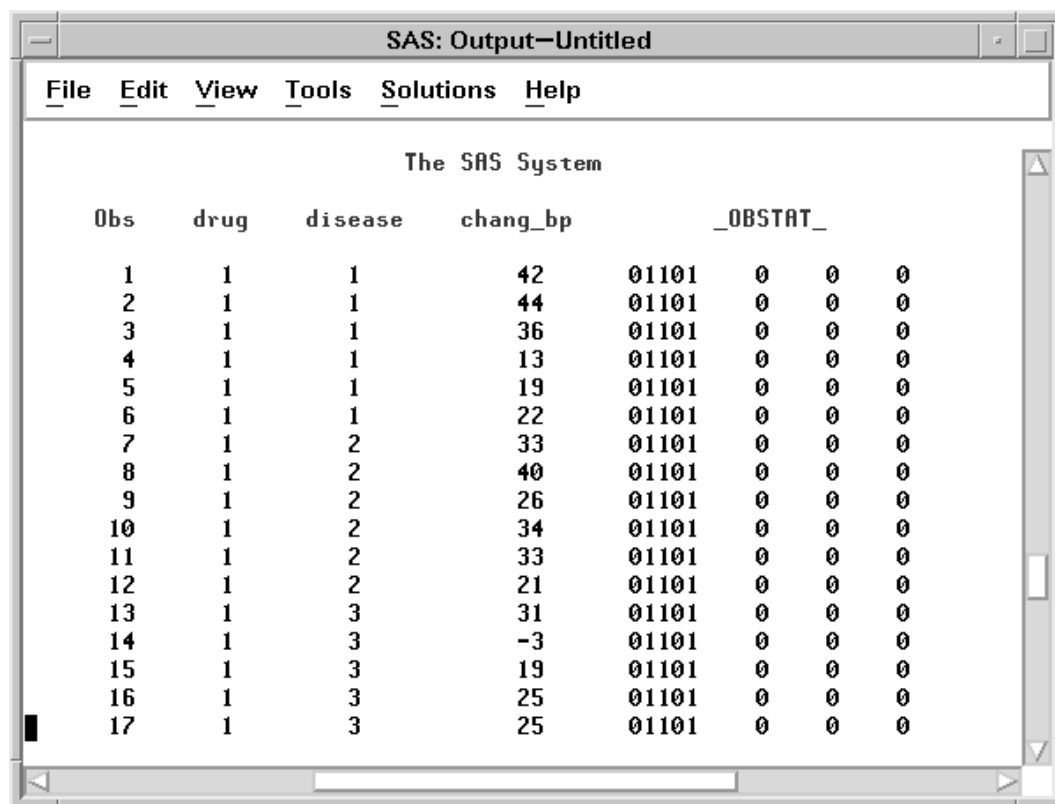


Figure 26.6. Run Menu

This displays the observations in the Output window.



The SAS System				
Obs	drug	disease	chang_bp	_OBSTAT_
1	1	1	42	01101
2	1	1	44	01101
3	1	1	36	01101
4	1	1	13	01101
5	1	1	19	01101
6	1	1	22	01101
7	1	2	33	01101
8	1	2	40	01101
9	1	2	26	01101
10	1	2	34	01101
11	1	2	33	01101
12	1	2	21	01101
13	1	3	31	01101
14	1	3	-3	01101
15	1	3	19	01101
16	1	3	25	01101
17	1	3	25	01101

Figure 26.7. Output Window

You can send the contents of the Output window to a file or to a printer by choosing **File:Print** in the Output window. On many hosts, the SAS System is installed so that this menu sends the contents of the Output window to a default printer. You can also choose this menu to save the window contents to a file and later route them to a printer using appropriate host commands.

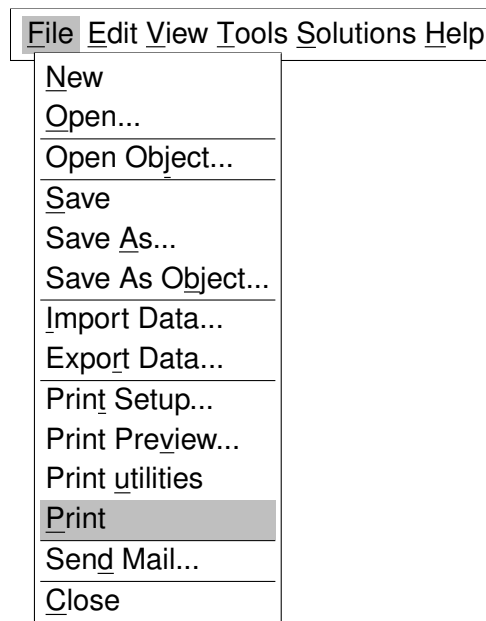


Figure 26.8. File Menu

Techniques ♦ *Saving and Printing Data*

Alternatively, you can redirect SAS System output from the Output window to a text file by using the PRINTTO procedure.

For more information on printing from the Output window, refer to the SAS companion for your host. For more information on PROC PRINT and PROC PRINTTO, refer to the *SAS Procedures Guide*.

Chapter 27

Saving and Printing Graphics

Chapter Contents

CHOOSING FONTS	432
SETTING DISPLAY OPTIONS	435
SAVING GRAPHICS	436
Saving Graphics Catalogs	436
Saving Graphics Files	437
PRINTING	439
Printing from the Display	439
Printing from the Clipboard	439
Printing from the Window	440

Chapter 27

Saving and Printing Graphics

If you have SAS/GRAPH software installed, you can save any SAS/INSIGHT window to a graphics catalog. You can modify graphics using the Graphics Editor and print them on any SAS/GRAPH device. You can save graphics files in bitmap formats including GIF, TIFF, and PostScript™.

On Windows and OS/2 hosts, SAS/INSIGHT software prints using host printing facilities. On other hosts, you can print using SAS/GRAPH software or host-provided screen-dumping utilities.

To improve your output, you can choose proportional fonts and set display options.

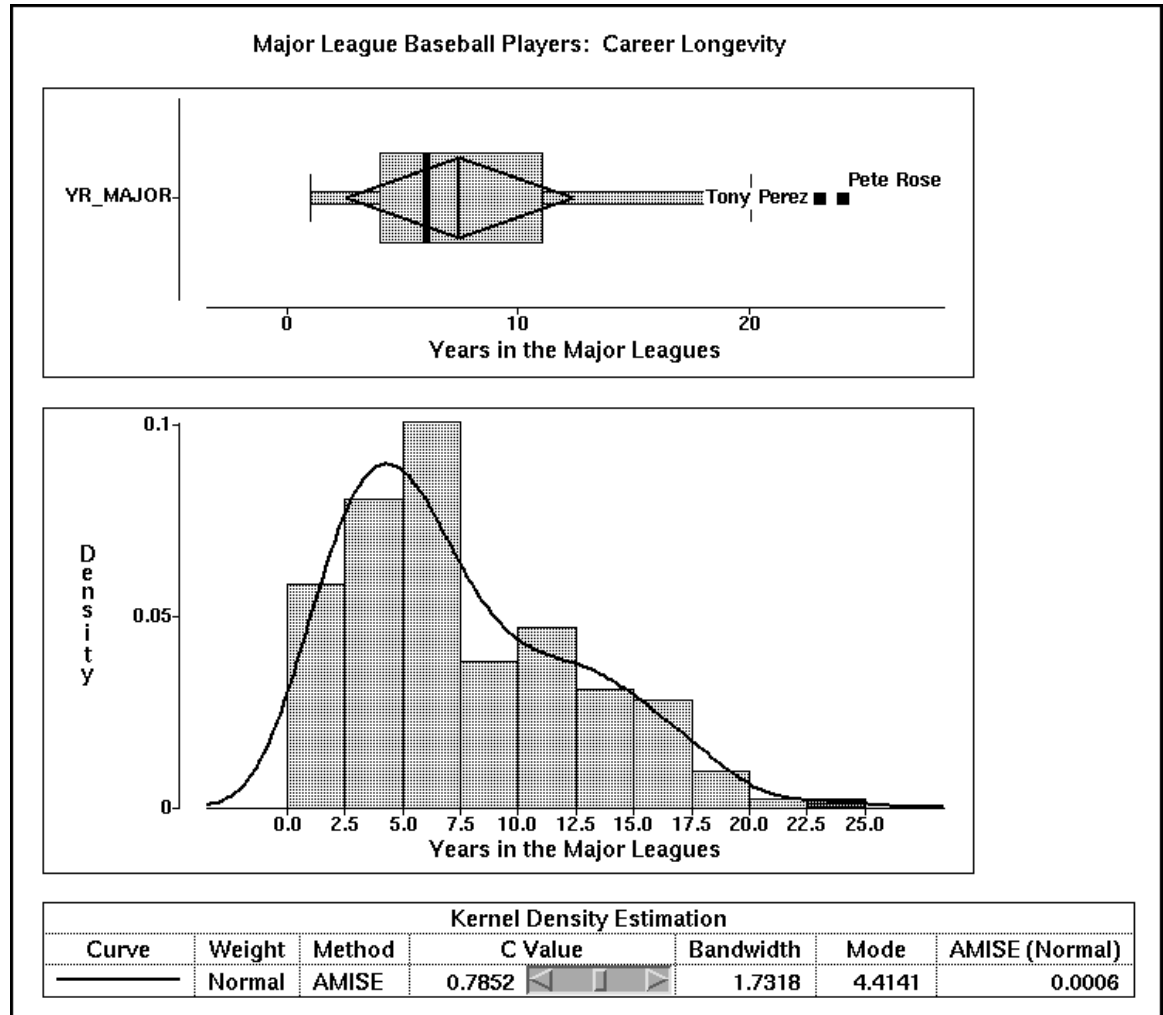


Figure 27.1. Printed Output with Title

Choosing Fonts

Proportional fonts make your output more readable. Choose **Edit:Windows:Fonts** to display the fonts dialog.

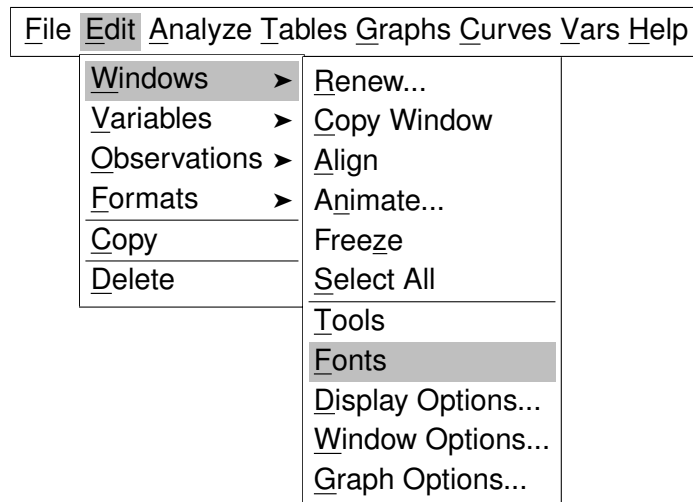


Figure 27.2. Edit:Windows Menu

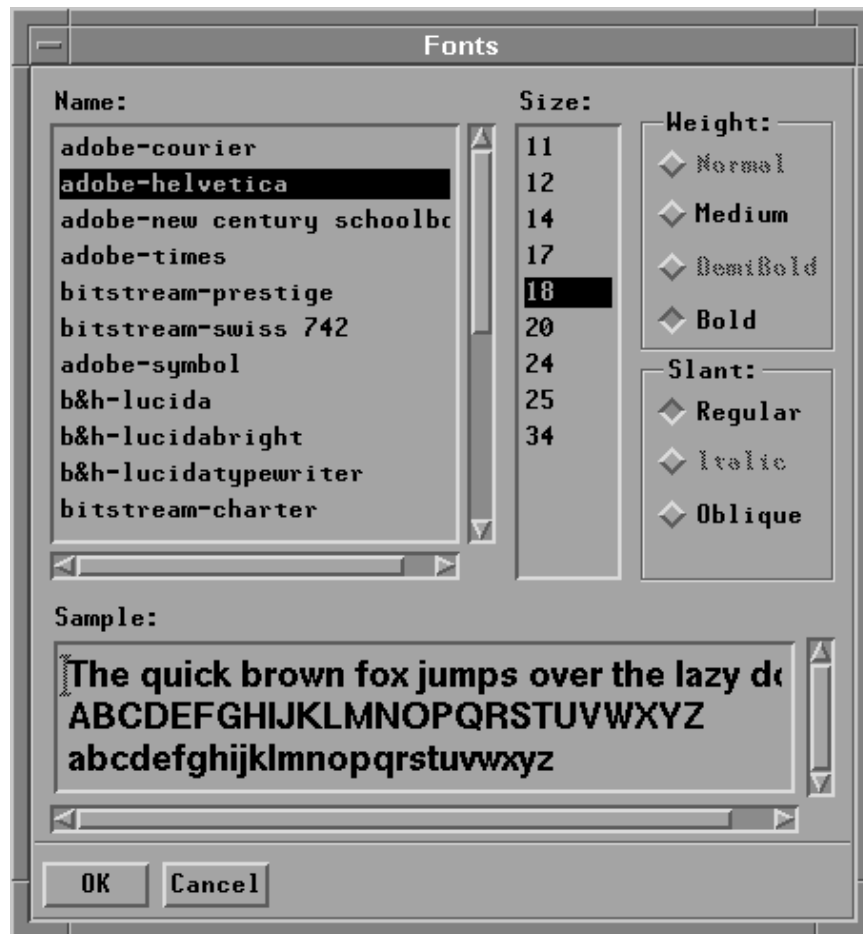


Figure 27.3. Fonts Dialog

Techniques ♦ Saving and Printing Graphics

The appearance of the fonts dialog depends on your host, and its contents depend on the fonts you have installed. On most hosts, you can simply click on a font name, click on other settings if desired, then click **OK** to set the font.

The font you choose is used to display tables, data values, and axis labels in graphs.

Tick labels in graphs use a slightly smaller font from the same font family.

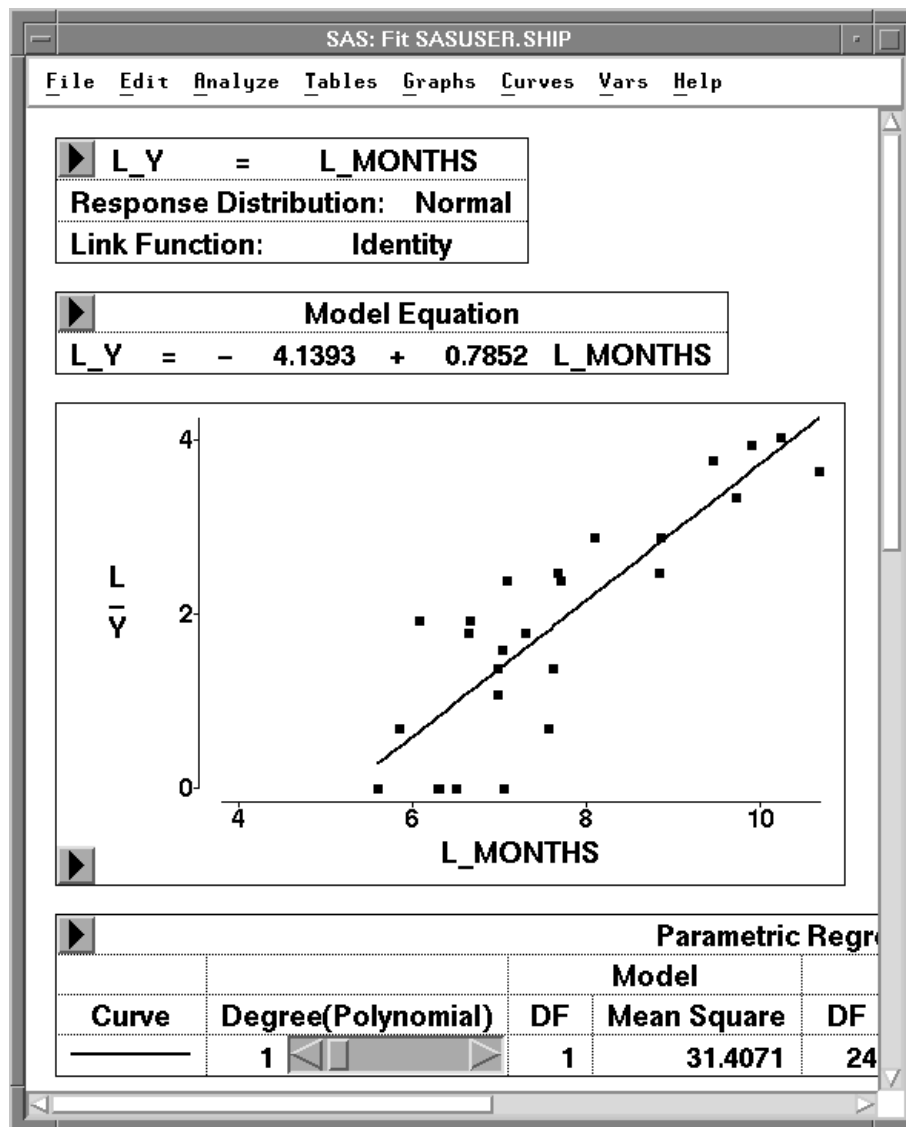


Figure 27.4. Fit Analysis with Proportional Font

Setting Display Options

To improve presentation output, SAS/INSIGHT software provides display options. Choose **Edit:Windows:Display Options** to produce the display options dialog.

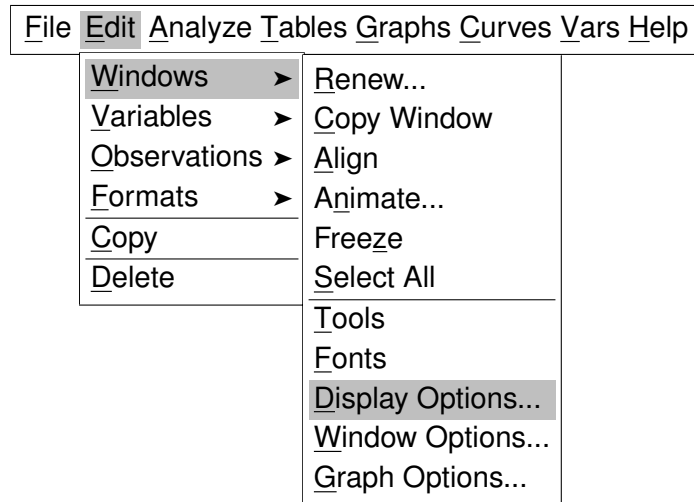


Figure 27.5. Edit:Windows Menu

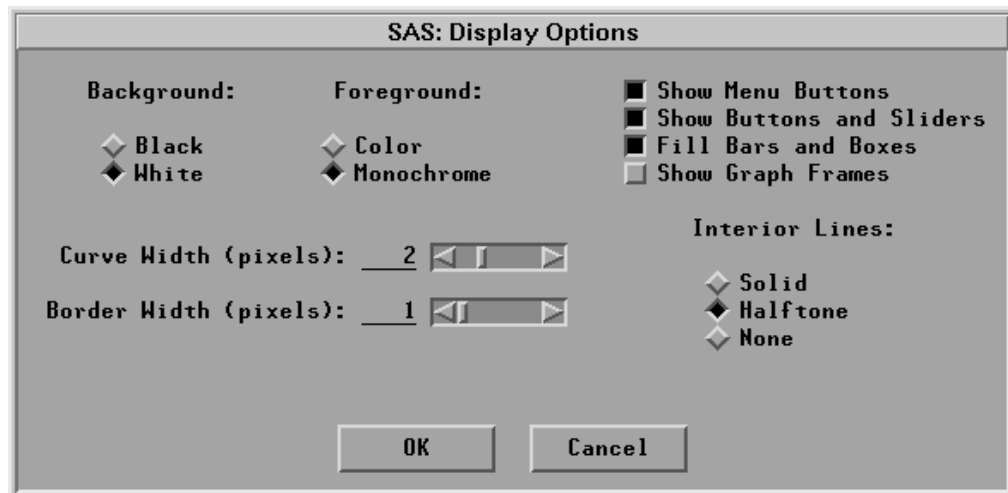


Figure 27.6. Display Options Dialog

The **Background** option enables you to choose a **Black** or **White** background. Because they provide maximum contrast, black and white are the best background colors for exploratory data analysis.

Printing on black-and-white printers may translate colors to shades of gray. If gray shades do not reproduce well on your printer, choose **Foreground:Monochrome** to improve your output. The figures in this book are set as in [Figure 27.6](#).

The remaining display options are described in detail in [Chapter 29, “Configuring SAS/INSIGHT Software.”](#) You can choose **File:Save:Options** to save all option settings to use as defaults in subsequent SAS/INSIGHT sessions.

Saving Graphics

If you have SAS/GRAPH software installed, you can save graphics catalogs in either **Graph** or **Image** format. You can use SAS/GRAPH software to save graphics files in a variety of bitmap formats.

Saving Graphics Catalogs

To save SAS/GRAPH catalogs from SAS/INSIGHT software, follow these steps.

⇒ **Select any graphs or tables you want to save.**

If no graphs or tables are selected, you will save all objects visible in the active window. To save all objects in the window, visible or not, choose **Edit:Windows:Select All**. Choosing this menu selects all graphs and tables in the active window.

⇒ **Choose File:Save:Graphics Catalog.**

This calls up the save graphics catalog dialog.

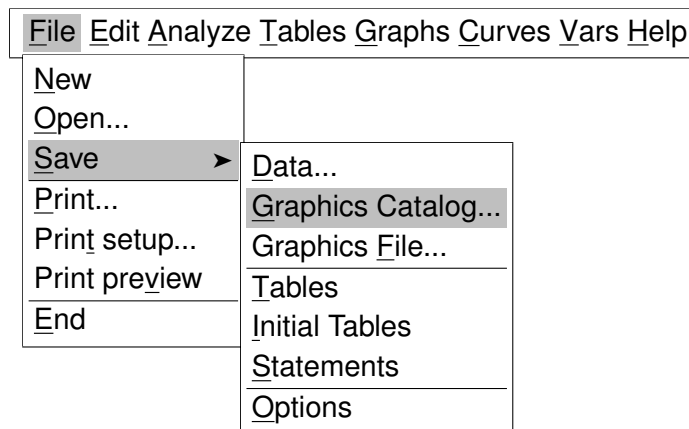


Figure 27.7. File:Save Menu

⇒ **Click the Graph or Image radio button to set your preference.**



Figure 27.8. Graphics Catalog Dialog

You can also specify catalog, entry, and description for your graphics output.

Set the **One Per Entry** option if you want to store each graph and table as a separate catalog entry. Entry names are derived from the name of the graph or table.

Set the **Titles and Footnotes** option if you want to use SAS titles and footnotes.

If you set both **One Per Entry** and **Titles and Footnotes** options, and if your window contains group variables, an additional title is generated to show the group. The group title is similar to the BY-group title in SAS/GRAPH output.

⇒ Click **OK** to save the catalog

Saving Graphics Files

You can use SAS/GRAPH software to save graphics files in a variety of bitmap formats. To save bitmaps, follow these steps.

⇒ **Select any graphs or tables you want to save.**

If no graphs or tables are selected, you will save all objects visible in the active window.

⇒ **Choose File:Save:Graphics File to display the graphics file dialog**

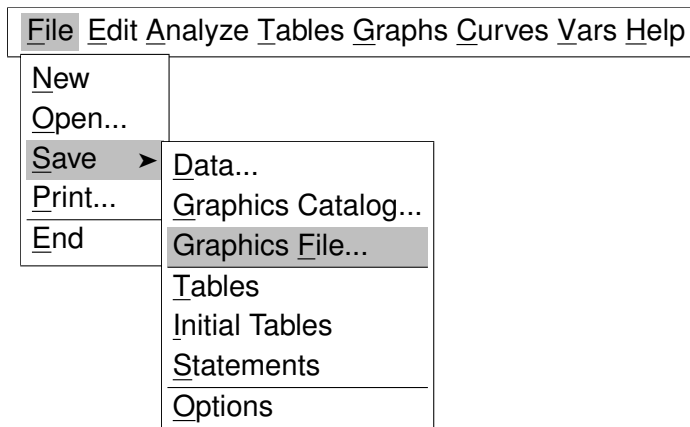


Figure 27.9. File:Save Menu

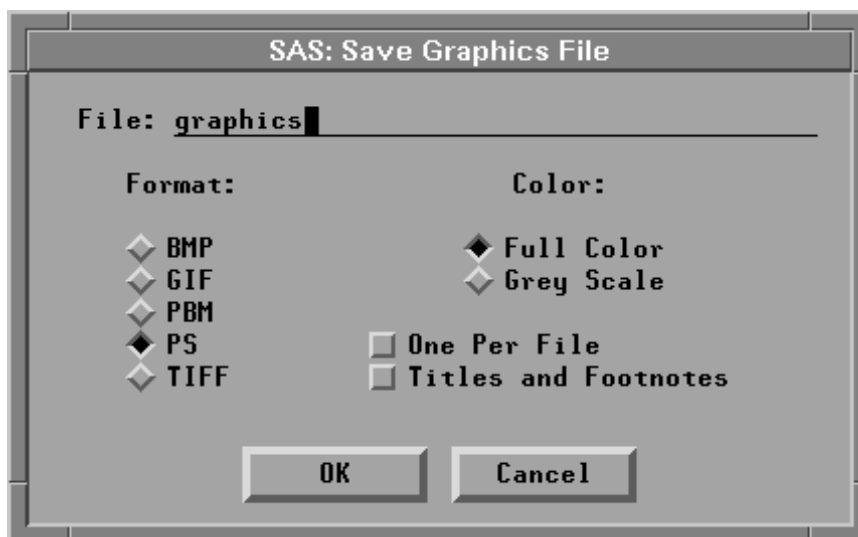


Figure 27.10. Graphics File Dialog

⇒ **Enter your file name, choose a format, and set additional options.**

Use the **Full Color** or **Grey Scale** options to control the colors stored in graphics files. Usually **Grey Scale** produces smaller files for faster printing.

Set the **One Per File** option if you want to store each graph and table in a separate file. If you set this option, the directory name is derived from the name you enter. Eight-character file names are derived from the name of the graph or table; for example, “scatter” for scatter plots, or “parametr” for parameter estimates.

Set the **Titles and Footnotes** option if you want to use SAS titles and footnotes.

If you set both **One Per File** and **Titles and Footnotes** options, and if your window contains group variables, an additional title is generated to show the group. The group title is similar to the BY-group title in SAS/GRAPH output.

⇒ **Click OK to save the graphics file.**

† **Note:** Clicking **OK** overwrites any files with the same file name.

For more information on saving graphics in bitmap formats, refer to the chapter on “Exporting SAS/Graph Output” in *SAS/GRAPH Software: Reference*.

Printing

Methods of printing vary greatly among different hosts. This section describes briefly the typical steps in printing on most personal computers and workstations. For more information on printing, refer to your host documentation and to the SAS companion for your host. See also the host changes and enhancements reports for Releases 6.10 and 6.11, as several hosts have improved printing in these releases.

Briefly, SAS/INSIGHT supports three ways of printing. If your host provides screen-dumping utilities, you can print anything that is visible on the display. Alternatively, on many hosts you can copy graphs and tables to the clipboard and then print the clipboard. Finally, you can use host printing facilities or SAS/GRAPH software to print directly from SAS/INSIGHT windows.

Printing from the Display

Many hosts provide tools or interfaces to print directly from the display. On UNIX hosts, tools such as **xwd** and **xv** deliver high-quality output. On Windows hosts, you can print the active window directly from the display by following these steps.

⇒ **Choose File:Print.**

⇒ **Set the Print as Bitmap check box.**

⇒ **Click OK.**

Printing from the display restricts you to printing objects that are visible. For more flexibility, you can print from the clipboard.

Printing from the Clipboard

Windows hosts support printing from the clipboard. To print graphs and tables from the clipboard, follow these steps.

⇒ **Select any graphs or tables you wish to print**

Drag a rectangle through the graphs and tables, or click on their edges. If no graphs or tables are selected, you will print all objects visible in the active window.

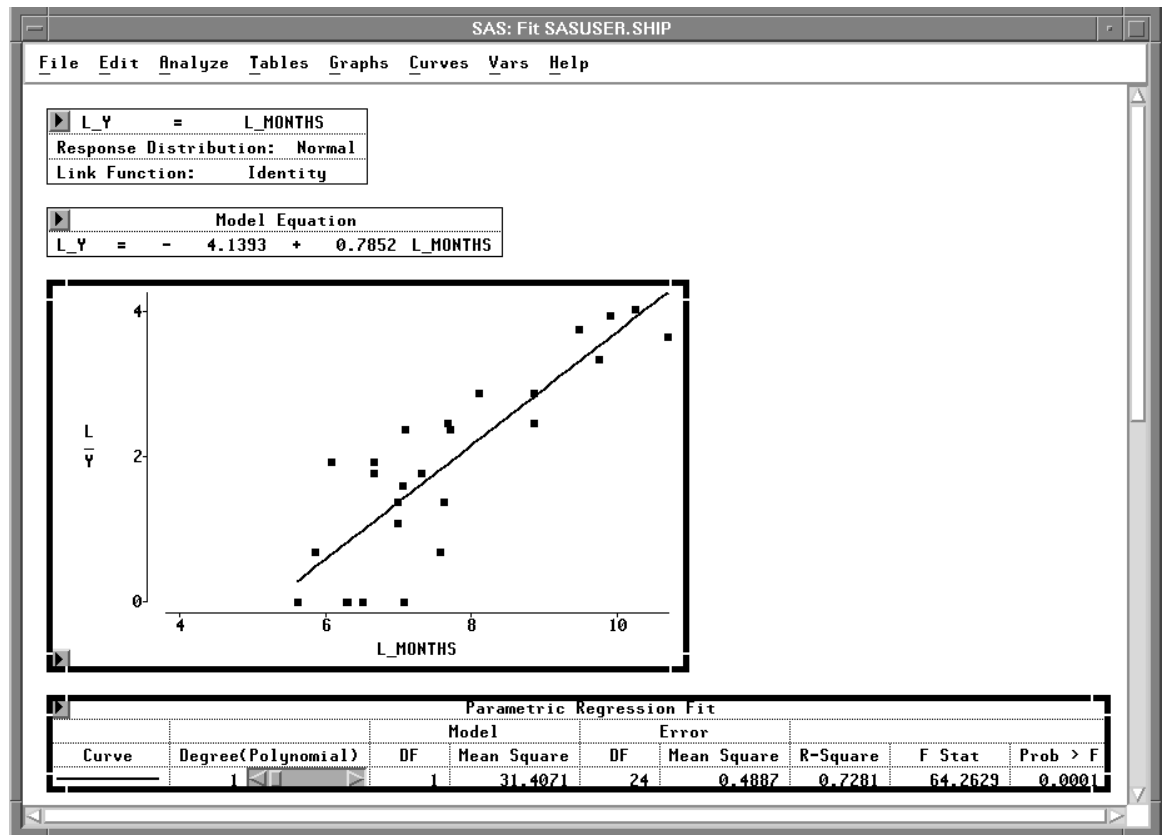


Figure 27.11. Selected Graphs and Tables

- ⇒ Choose **Edit:Copy** to copy selected objects to the clipboard.
- ⇒ Choose **File:Print**.
- ⇒ Set the **Print as Bitmap** check box.
- ⇒ Set the **Contents of list to Clipboard (bitmap)**.
- ⇒ Click **OK**.

Printing from the clipboard is not supported on all hosts. For a more general way of printing, you can print directly from a SAS/INSIGHT window.

Printing from the Window

Printing from the window is the most flexible way to print. To print from a SAS/INSIGHT window, follow these steps.

- ⇒ **Select any graphs or tables you wish to print**
If no graphs or tables are selected, you will print all objects visible in the active window. To print all objects in the window, visible or not, choose **Edit:Windows>Select All** to select all graphs and tables in the window.

⇒ **Choose File:Print.**

On Windows and OS/2, this displays a host Print dialog, with options such as the **Print as Bitmap** option in the preceding sections. If you receive a host Print dialog, click **OK**. This displays the SAS/INSIGHT Print dialog.



Figure 27.12. SAS/INSIGHT Print Dialog

In the SAS/INSIGHT Print dialog, the **Fill Page** option expands your output to fill the area of the page. The **One Per Page** option prints each graph and table on a separate page. The **Titles and Footnotes** option prints using SAS titles and footnotes.

If you set both **One Per Page** and **Titles and Footnotes** options, and if your window contains group variables, an additional title is generated to show the group. The group title is similar to the BY-group title in SAS/GRAPH output. An example of the group title for histograms of **YR_MAJOR** by **LEAGUE** is shown in [Figure 27.13](#).

⇒ **Set options as needed, then click OK in the Print dialog**

Clicking **OK** in the Print dialog routes your printing through host printing facilities if they are provided. Windows and OS/2 provide such facilities, and they are documented in SAS companions and host changes and enhancements reports.

If your host does not support host printing, your printing is routed through SAS/GRAPH software. You will be prompted for an output device if you have not specified one with the GOPTIONS TARGETDEVICE= option.

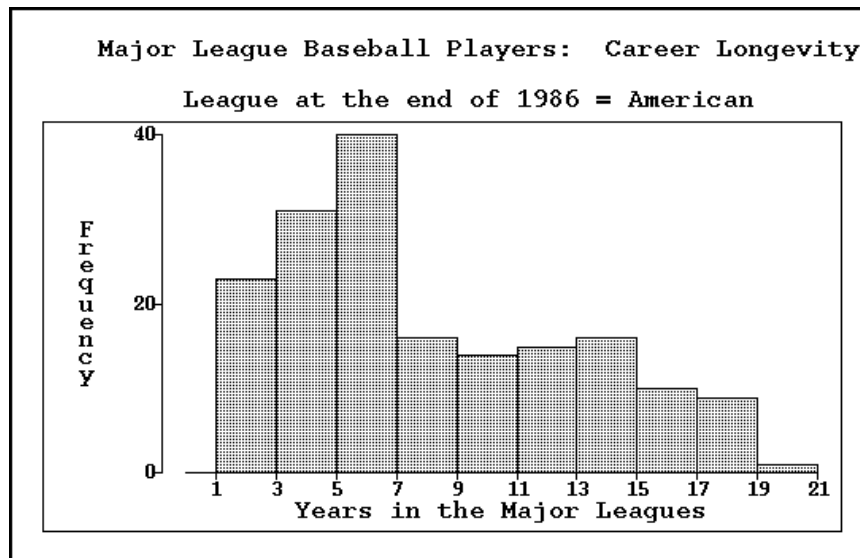


Figure 27.13. Output with Title and Group Title

An alternative way of printing is to save your graphics to catalogs and print them from SAS/GRAPH software. This enables you to edit your output before printing. SAS/GRAPH printing and graphics catalogs are described in *SAS/GRAPH Software: Reference*.

Chapter 28

Saving and Printing Tables

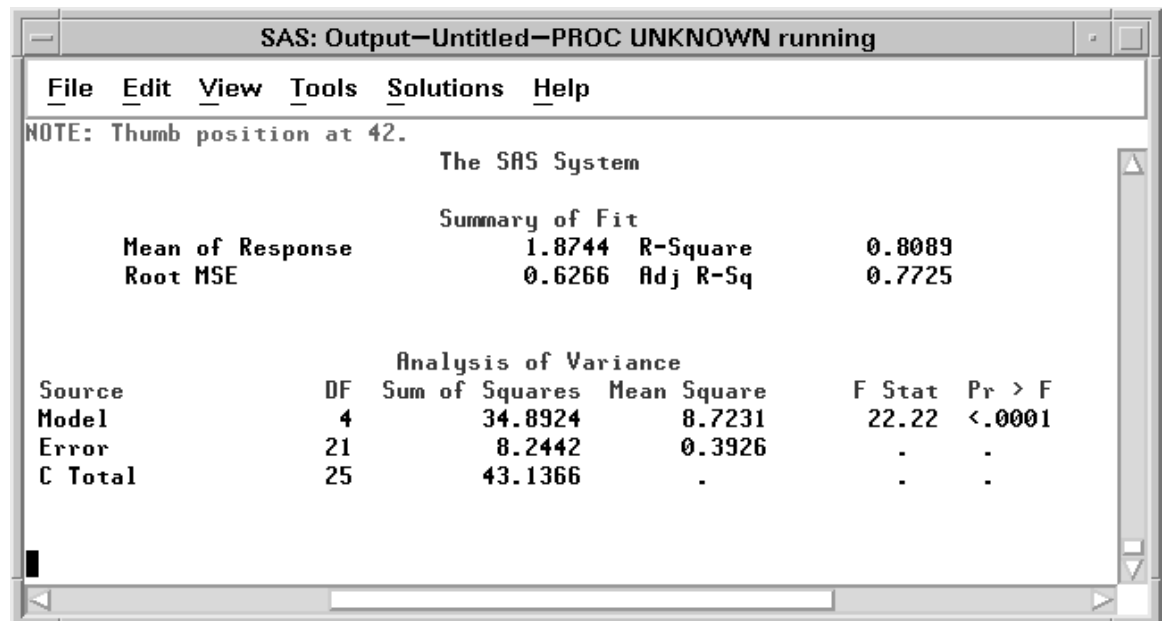
Chapter Contents

SAVING AND PRINTING TABLES AS OUTPUT OBJECTS	446
OUTPUT OBJECTS	450

Chapter 28

Saving and Printing Tables

SAS/INSIGHT software uses the Output Delivery System (ODS) to save tables. Thus you can save and print analysis tables to keep records of your SAS/INSIGHT session. You can also save tables as SAS data sets to use them as input for further analysis.



The screenshot shows a SAS window titled "SAS: Output—Untitled—PROC UNKNOWN running". The menu bar includes File, Edit, View, Tools, Solutions, and Help. The output text includes a note about thumb position, followed by "The SAS System" and a "Summary of Fit" table. Below that is an "Analysis of Variance" table.

NOTE: Thumb position at 42.

The SAS System

Summary of Fit			
Mean of Response	1.8744	R-Square	0.8089
Root MSE	0.6266	Adj R-Sq	0.7725

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	4	34.8924	8.7231	22.22	<.0001
Error	21	8.2442	0.3926	.	.
Total	25	43.1366	.	.	.

Figure 28.1. Output Tables

Saving and Printing Tables as Output Objects

SAS/INSIGHT software saves and prints tables using the Output Delivery System. The Output Delivery System enables you to save tables as *output objects*. You can edit and manipulate output objects using the OUTPUT procedure, and you can save output objects as text files, catalogs, or SAS data sets.

- ⇒ **Invoke SAS/INSIGHT software, create analyses, and select any tables of interest.**
To select tables, drag a rectangle across the tables or click on their edges. If you have no tables selected, you will save or print all tables in the window.

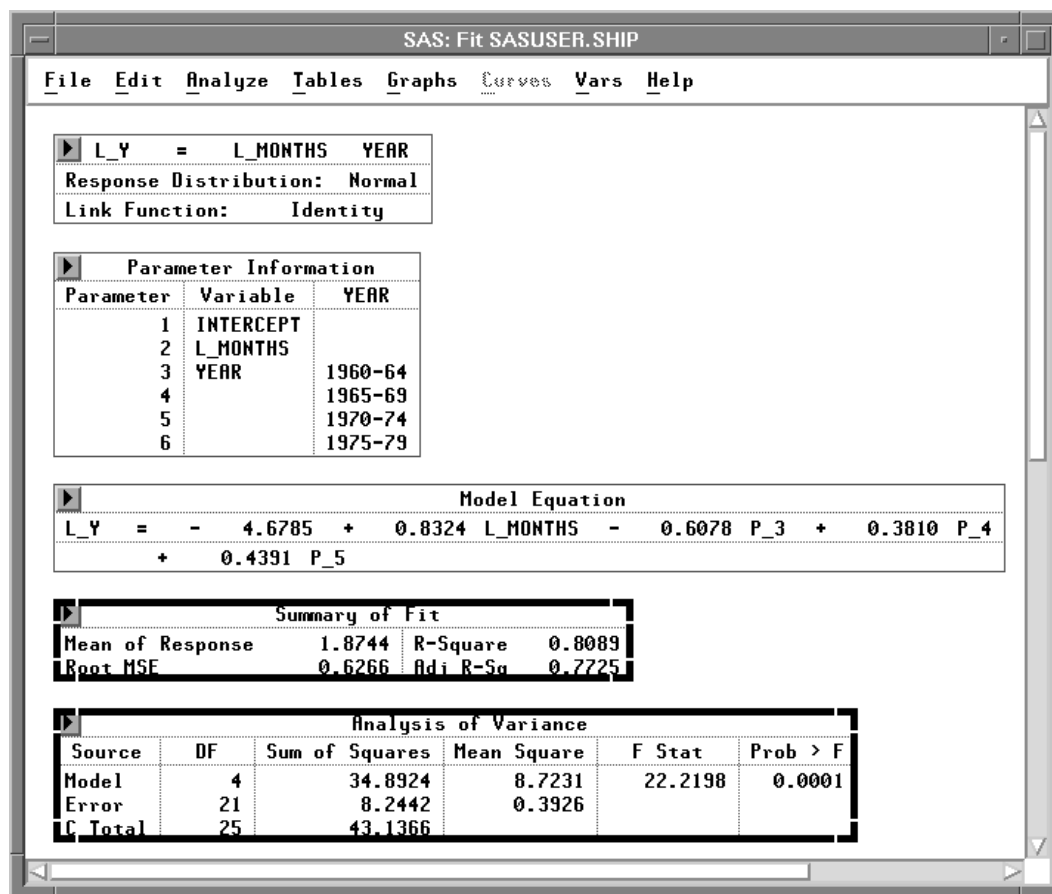


Figure 28.2. Tables Selected

- ⇒ **Choose File:Save:Tables.**

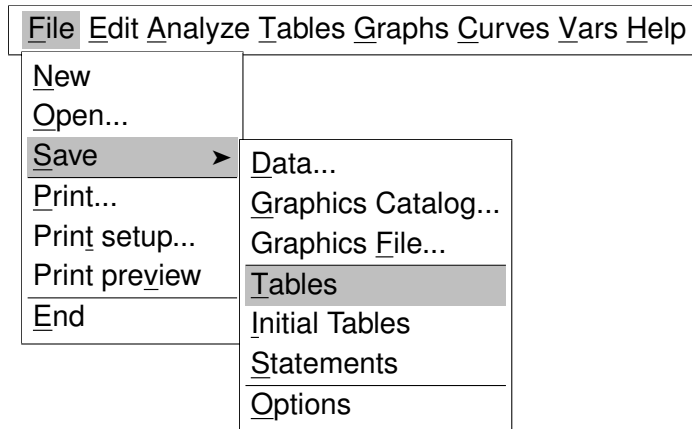


Figure 28.3. File:Save Menu

⇒ From the Program Editor menu, select **View:Results** to create the Results Window.

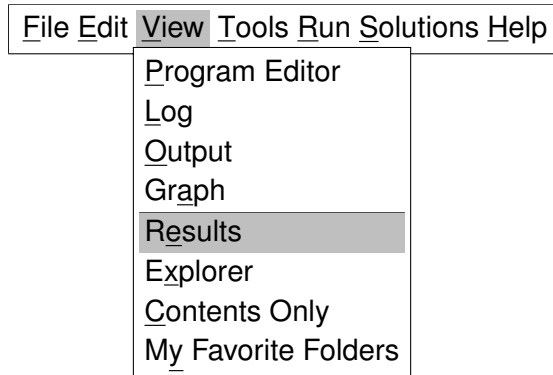


Figure 28.4. View Menu

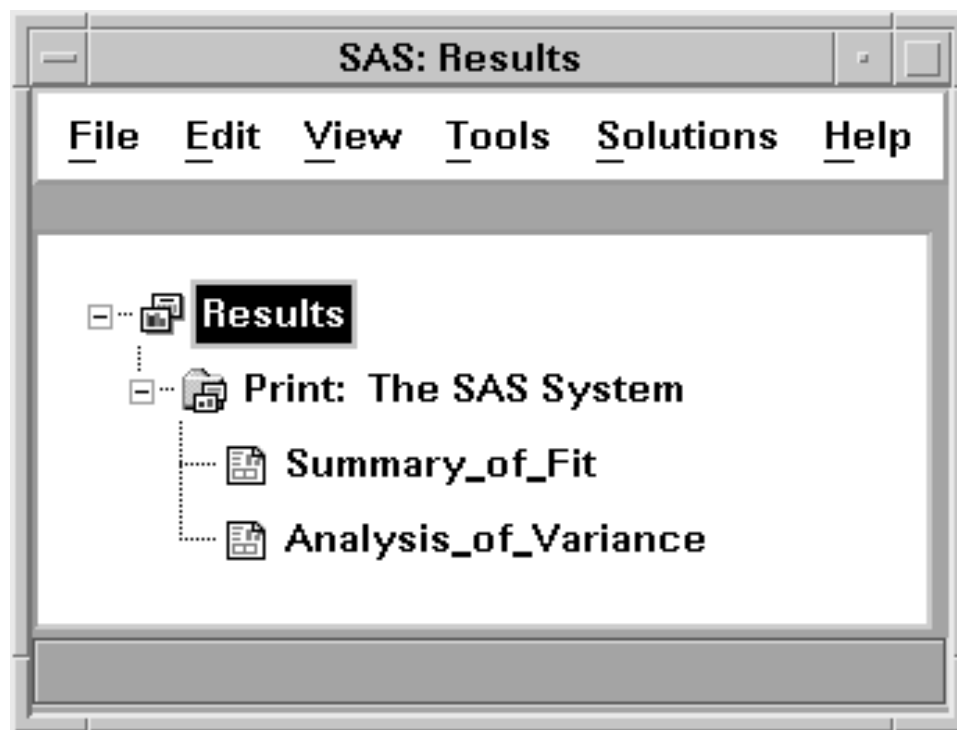


Figure 28.5. Results Window

⇒ Selecting the name of a table in the results window displays that table in the Output window.

You can save all tables at the creation of each analysis by choosing **File:Save:Initial Tables**. This menu is a toggle; choosing it again turns off the automatic saving of tables.

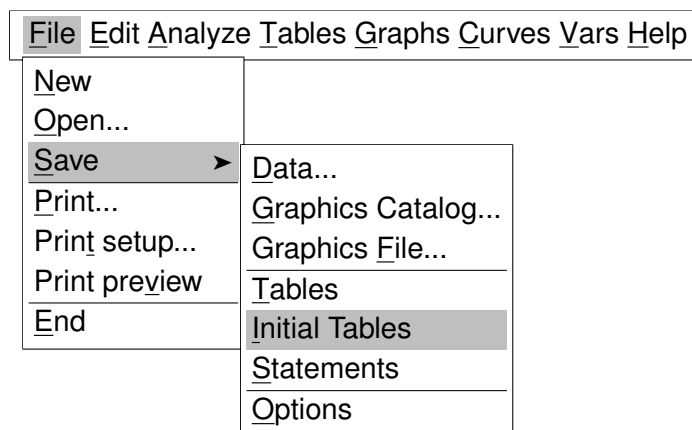


Figure 28.6. File:Save Menu

Also, each table has a pop-up menu to save just that table. Click on the menu button

at the upper left of the table to display the pop-up menu.

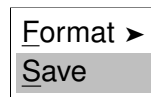


Figure 28.7. Table Pop-up Menu

Saving tables to the Output Delivery System converts your tables to output objects. Variables in output objects have names derived from the table headers. Where conflicts occur, a new unique name is generated. Variables in the output object are assigned formats derived from the tables.

You can send the contents of the Output window to a file or printer by choosing **File:Print** in the Output window. On many hosts, the SAS System is installed so that this menu sends the contents of the Output window to a default printer. You can also choose this menu to save the window contents to a file and later route them to a printer using appropriate host commands.

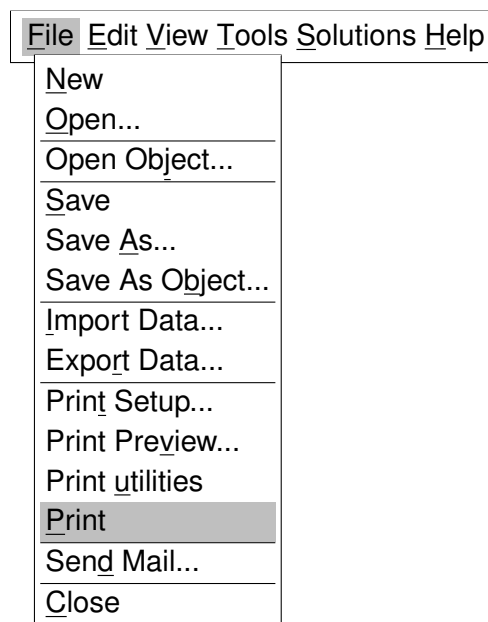


Figure 28.8. File Menu

Alternatively, you can redirect SAS System output from the Output window to a text file by using the PRINTTO procedure.

For more information on printing from the Output window, refer to the SAS companion for your host. For more information on PROC PRINTTO, refer to the *SAS Procedures Guide*.

Output Objects

By default, tables are saved in text format. It is also possible to save SAS/INSIGHT tables as SAS data sets, or in other formats such as HTML.

For example, the following steps illustrate how to create a data set from the Moments table in a distribution analysis.

⇒ **In the Program Editor, submit the following ODS command:**

```
ods output Moments = MOMENTS;
```

This command instructs ODS to create a SAS data set called **MOMENTS** from a table named “Moments”.

⇒ **Create a Moments table.**

One way to do this is to open the **DRUG** data set, select the **CHANG_BP** variable, and select **Analyze:Distribution (Y)** to obtain a distribution analysis. The Moments table is generated by default.

⇒ **Open the MOMENTS data set .**

Select **File:Open**, and look under the **WORK** library. Select the **MOMENTS** data set and click **Open**.

† **Note:** You can find out the name of any table created in SAS/INSIGHT. To do this, submit the following ODS command in the Program Editor prior to creating the table.

```
ods trace output;
```

When you create a table, the name of that table is printed to the Log window.

You can also redirect all of your SAS/INSIGHT tables to an HTML file. Prior to creating any tables, submit an ODS command such as

```
ods html body="tables.htm";
```

Any tables you now save are written as HTML. When you are finished saving tables, submit the ODS command

```
ods html close;
```

To view the table’s values, select **View:Results** from the Program Editor menu. Then select the name of a table to view.

For more information on the Output Delivery System, refer to the chapter on “Using the Output Delivery System” in the *SAS/STAT User’s Guide* or refer to *The Complete Guide to the SAS Output Delivery System*.

Chapter 29

Configuring SAS/INSIGHT Software

Chapter Contents

SETTING OPTIONS	454
Setting Method and Output Options	454
Setting Display, Window, and Graph Options	458
SAVING OPTIONS	466
SETTING HOST RESOURCES	467

Chapter 29

Configuring SAS/INSIGHT Software

You can configure SAS/INSIGHT software in two ways. You can tailor SAS/INSIGHT software to the way you work by saving option settings for future use. You can also set host resources to improve SAS/INSIGHT software's performance on your host.

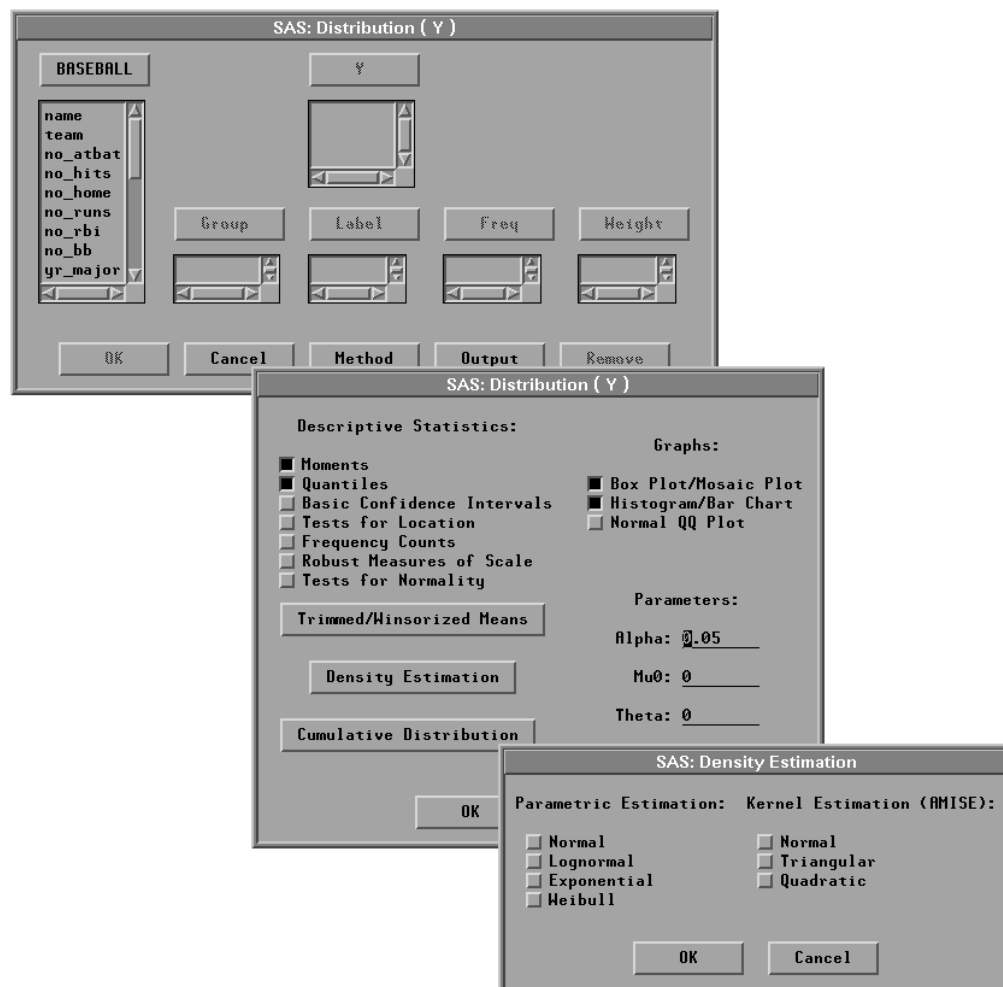


Figure 29.1. Setting Output Options

Setting Options

With SAS/INSIGHT software, you can set options in two ways. You can set options in an analysis window that affect the calculations and output displayed only in that window. Alternatively, you can set options that affect the display of all windows.

Setting Method and Output Options

Method options and *output options* affect only the individual analysis window for which they are set. You can set *method options* to determine how SAS/INSIGHT software performs calculations for a particular analysis. You can set *output options* to control the output produced in a graph or analysis. To modify method and output options for a box plot, follow these steps.

- ⇒ **Open the BASEBALL data set.**
- ⇒ **Choose Analyze:Box Plot/Mosaic Plot (Y).**

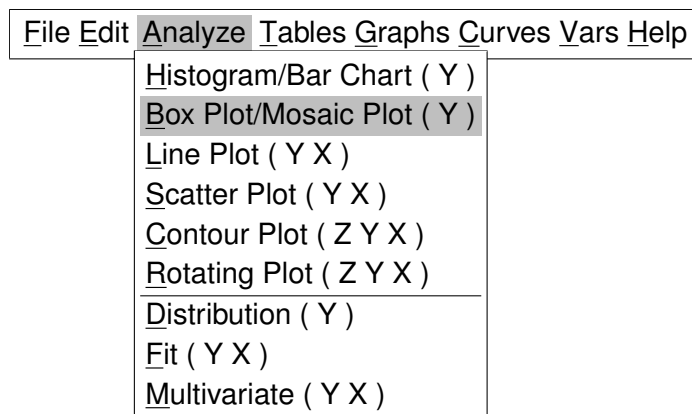


Figure 29.2. Analyze Menu

This displays the box plot variables dialog, as shown in [Figure 29.3](#). Note that both a **Method** and an **Output** button are displayed in this dialog. You can set **Output** options for each of the choices in the **Analyze** menu in [Figure 29.2](#). You can set **Method** options for each of these choices except for line plots, scatter plots, and rotating plots. You can find details on options for each analysis in the reference chapters.

- ⇒ **Assign NO_RBI the Y role by clicking on NO_RBI, then on Y.**



Figure 29.3. Box Plot Variables Dialog

⇒ Click the **OK** button to create the box plot.

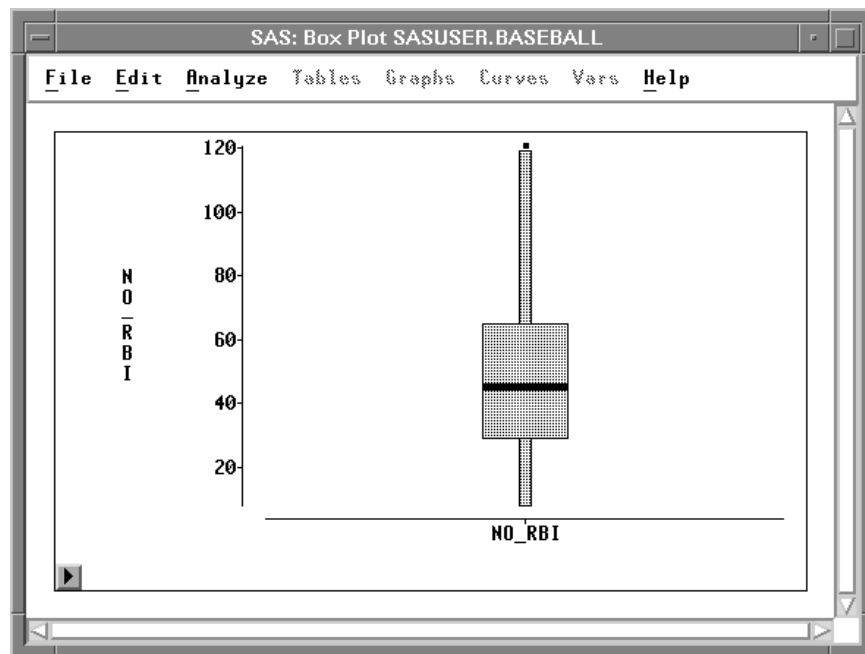


Figure 29.4. Box Plot

⇒ Choose **Edit:Windows:Renew** in the box plot window.
This redisplay the box plot variables dialog.

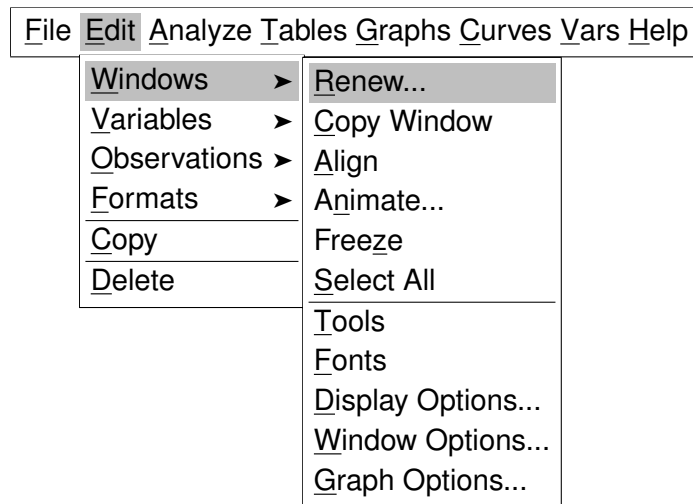


Figure 29.5. Edit:Windows Menu

⇒ Click on the **Method** button to display the box plot method dialog

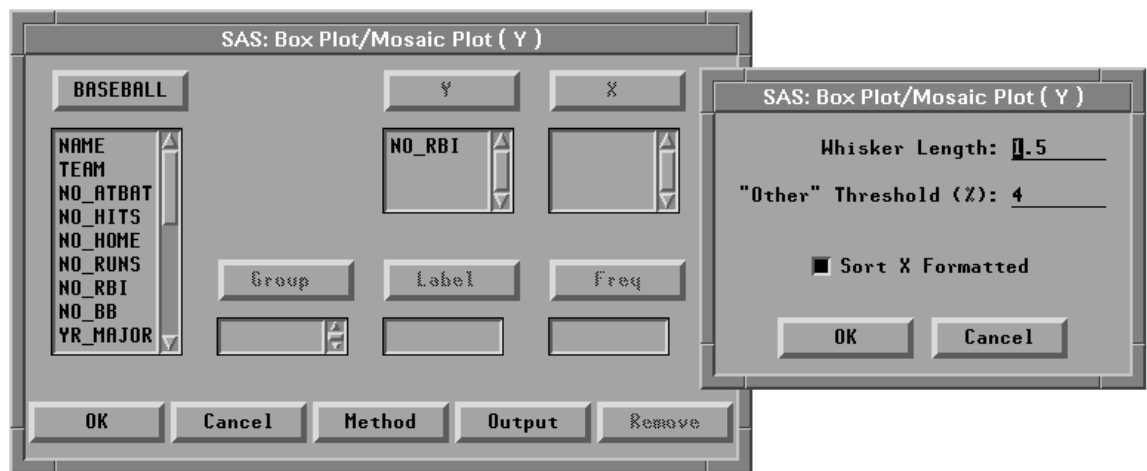


Figure 29.6. Variables and Method Dialogs

⇒ Change the whisker length to 1.0 and click the **OK** button in the method dialog

⇒ Click the **Output** button to display the box plot output dialog

⇒ Click the **Means**, **Labels**, and **Y Axis Vertical** buttons.

The **Means** and **Y Axis Vertical** buttons are toggles. The display of a means diamond is now on, and the **Y** axis is set to be displayed horizontally instead of vertically. The **Labels** button is a state indicator showing that variable labels are set to be displayed.

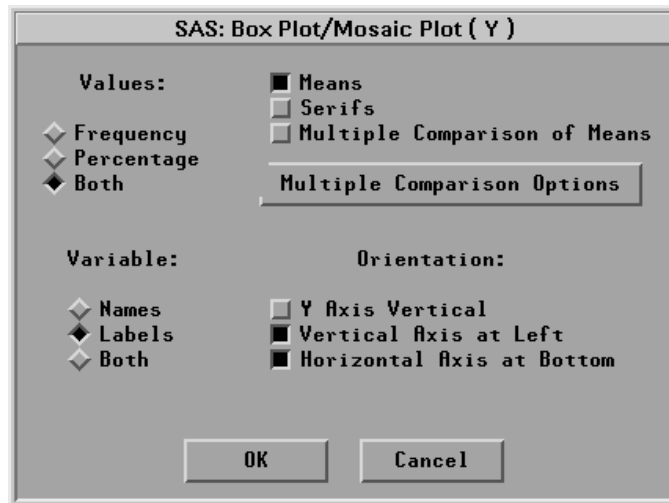


Figure 29.7. Box Plot Output Dialog

⇒ **Click OK in both the output dialog and the variables dialog**

This displays the new box plot in [Figure 29.8](#). Note that the box plot is displayed horizontally with a mean diamond. The upper whisker is now only the same length as the box, showing more points as individual outliers. Also, the RBI axis shows the variable label instead of the variable name.

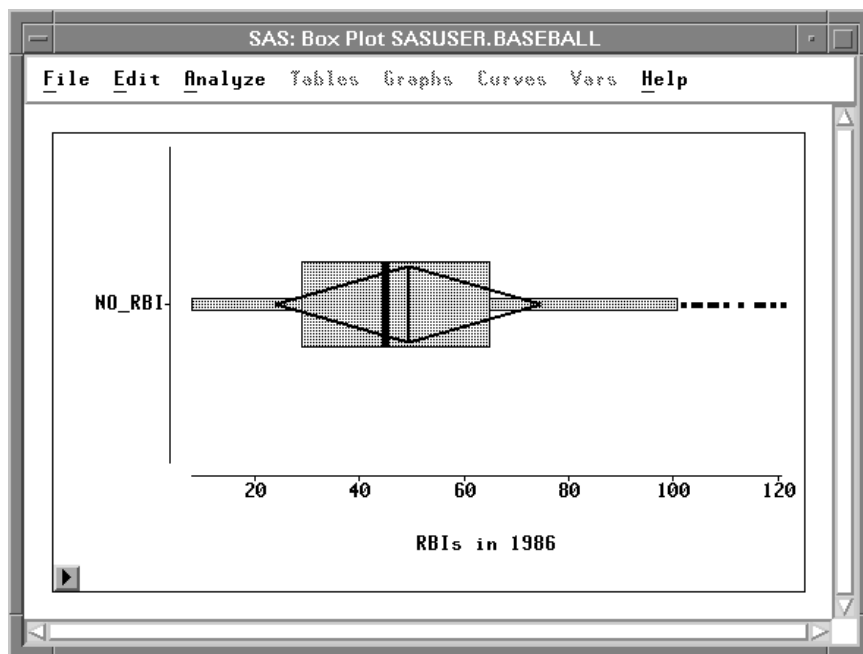


Figure 29.8. Modified Box Plot

Setting Display, Window, and Graph Options

Display options, *window options*, and *graph options* modify aspects of the software that affect every analysis. To set *display options*, choose **Edit:Windows:Display Options**. Note that you also set *window options* and *graph options* from the **Edit:Windows** menu.

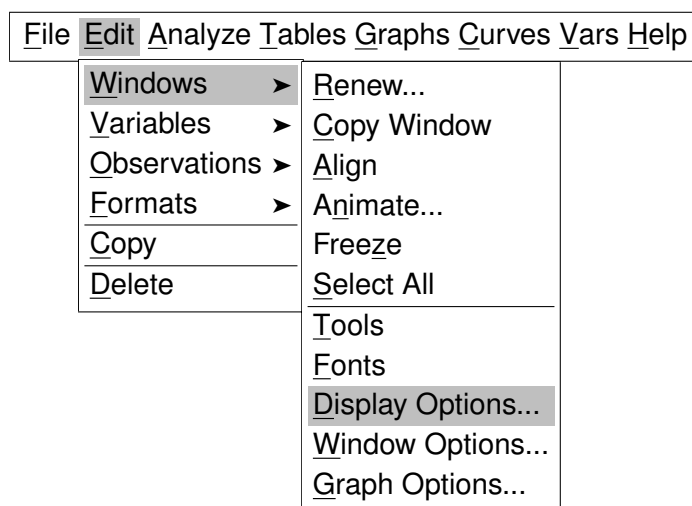


Figure 29.9. Edit:Windows Menu

This displays the display options dialog, as shown in [Figure 29.10](#).

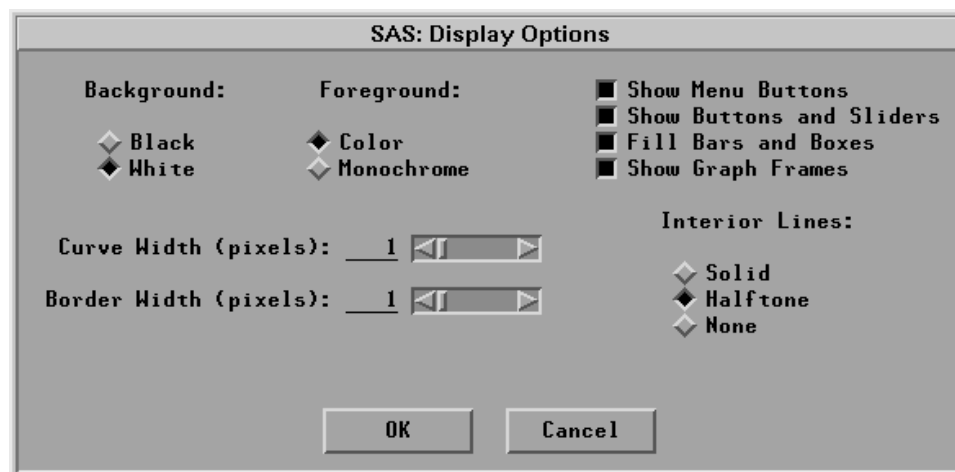


Figure 29.10. Display Options Dialog

The dialog contains the following options:

Background	specifies either Black or White background.
Foreground	specifies either Color or Monochrome foreground. Monochrome display improves printed output by removing shades of gray used to approximate color.
Show Menu Buttons	governs the display of pop-up menu buttons in all windows. Turn this option off to remove menu buttons.
Show Buttons and Sliders	governs the display of all buttons and sliders except menu buttons. Turn this option off to remove buttons and sliders.
Fill Bars and Boxes	specifies the use of pattern fill in bar charts, box plots, and mosaic plots. Turn this option off to display empty bars and boxes. On slower hosts, turning this option off improves display speed as well as printed output.
Show Graph Frames	In nonrotating plots, this option specifies whether the two axes are displayed as two disjoint line segments or are joined together as part of a frame.
Curve Width	sets the default width of curves in pixels. On most hosts, a width of 1 pixel maximizes display speed.
Border Width	sets the default width of graph and table borders in pixels. When you are printing with a black background, increasing border width improves the display of graphs and tables.
Interior Lines	sets the display of lines within the data window and analysis tables. Solid produces solid lines; Halftone produces a dimmer line; None removes interior lines. Solid and None settings improve display speed on personal computers.

The figures in this book are produced with **Foreground** set to **Monochrome** and **Curve Width** set to **2** pixels. Most figures have **Show Graph Frames** turned off.

To set *window options*, choose **Edit:Windows:Window Options**. This displays the window options dialog.



Figure 29.11. Window Options

The dialog contains the following options:

- Layout** sets the algorithm for positioning windows. **Spread** spreads the windows so that the maximum number of tables and graphs are visible. **Cascade** causes each window to be offset a small distance from the previous window. On some hosts, the effect of this option is overridden by the host window manager.
- Show Tools at Startup** causes the Tools window to display automatically when you invoke SAS/INSIGHT software.
- Zoom/Scroll Speed (%)** sets the speed of the zoom tool and the speed of automatic scrolling when you drag a selection past the window border. The speed is a percentage value between 0 and 100. Some hosts override this option.
- Default Margin (mm.)** sets the spacing in millimeters between graphs and tables in analysis windows. If your display is small, reduce this value to maximize the display of information.
- Number of Groups** sets the number of groups you can use in an analysis without getting a request for confirmation.

Zoom/Scroll Speed, **Default Margin**, and **Number of Groups** can be controlled by sliders to the right of the option. To set these options, either click or drag on the sliders or type in the entry field.

To set graph options, choose **Edit:Windows:Graph Options**. This displays the graph options dialog.

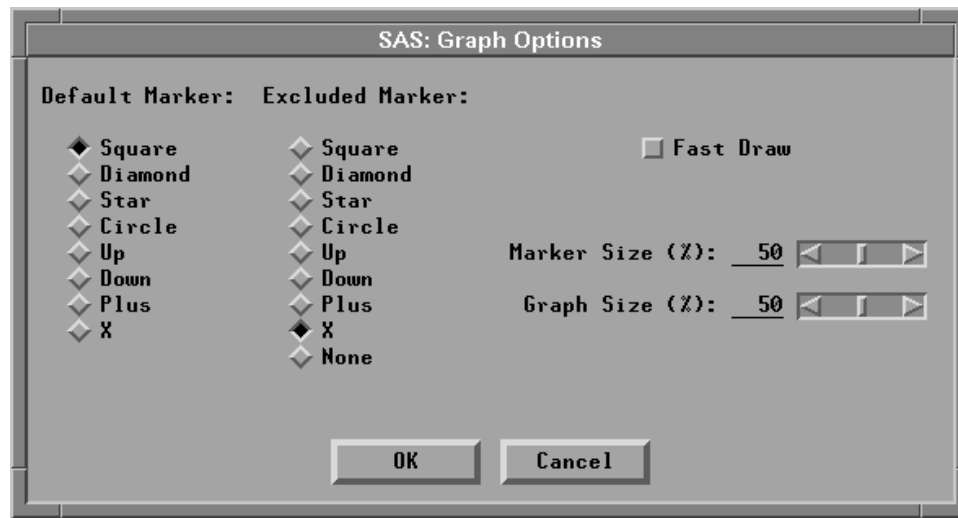


Figure 29.12. Graph Options

The dialog contains the following options:

Default Marker sets the default marker shape. On personal computers, **Square** and **Plus** are the best choices; these markers are the fastest to display. On fast workstations, **Circle** is preferable to minimize interference between plotted observations.

Excluded Marker sets the marker shape for observations that are excluded from calculations. **X** is the default. If you choose **None**, marker shape is not affected by exclusion.

Fast Draw sets display algorithms for rotation, brushing, manipulation of histograms, and dynamic curve fitting. By default, this option is off, which produces slower but smoother dynamic effects. If this option is on, speed is improved but, on some hosts, the display may flicker. The better choice of algorithms depends on your host, the size of your graphs, and the number of observations.

Marker Size (%) sets the default size of markers in plots. This is the marker size used when you choose **Marker Sizes:Size to Fit**. This is a percentage value between 0 and 100.

Graph Size (%) sets the default size of windows and graphs. This is a percentage value between 0 and 100. If your display is small, reduce this value to display more graphs.

To see the effects of various display, window, and graph options, follow these steps.

⇒ **Create a fit analysis for the model $NO_RBI = NO_HITS$.**

Use the techniques described in [Chapter 13, “Fitting Curves.”](#) This creates the fit analysis shown in [Figure 29.13](#).

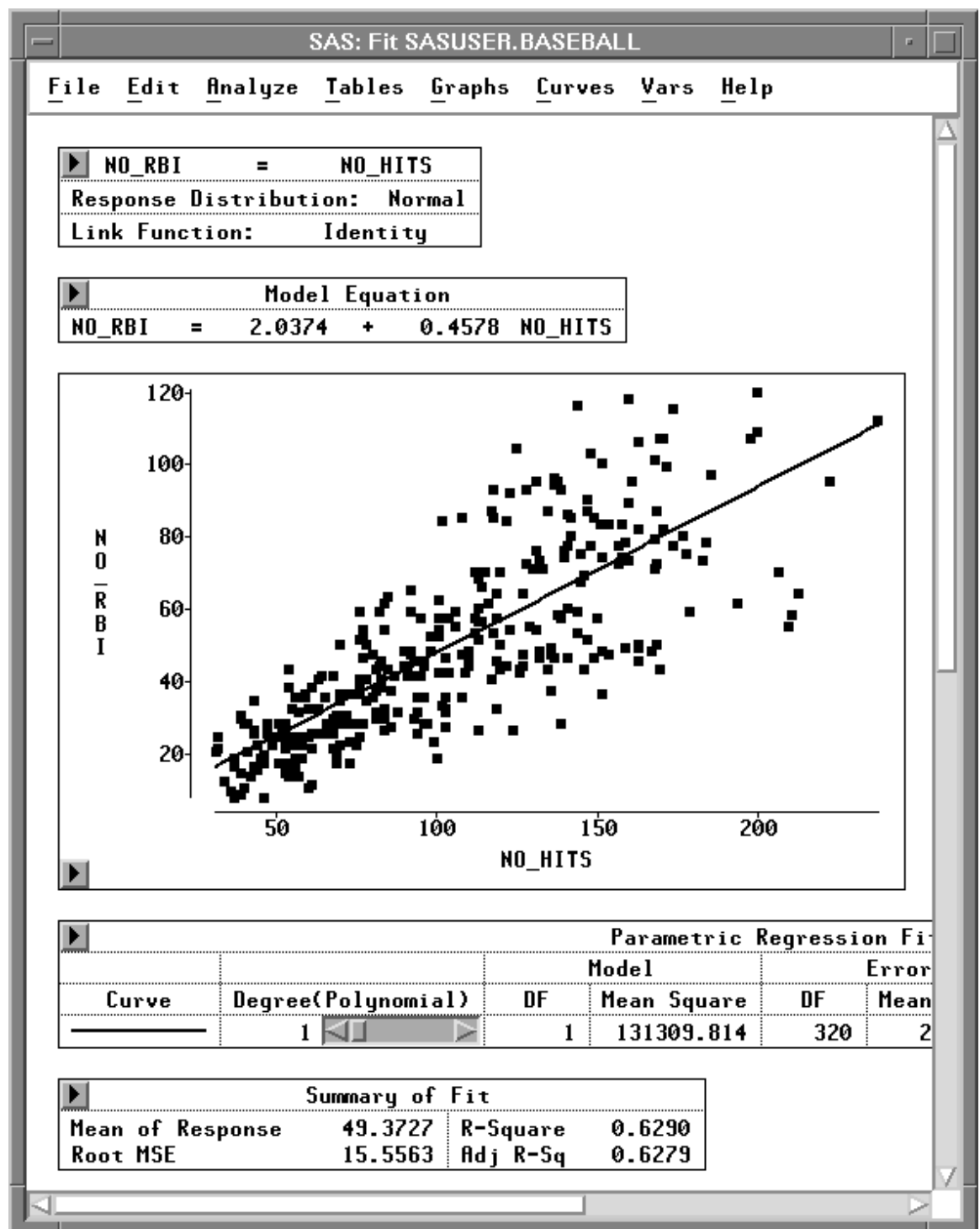


Figure 29.13. Fit Analysis

- ⇒ Choose **Edit:Windows:Display Options** to display the display options dialog
- ⇒ Click on the toggle button for **Show Menu Buttons**.

Recall that the figures here already have **Foreground** set to **Monochrome** and **Curve Width** set to **2** pixels.

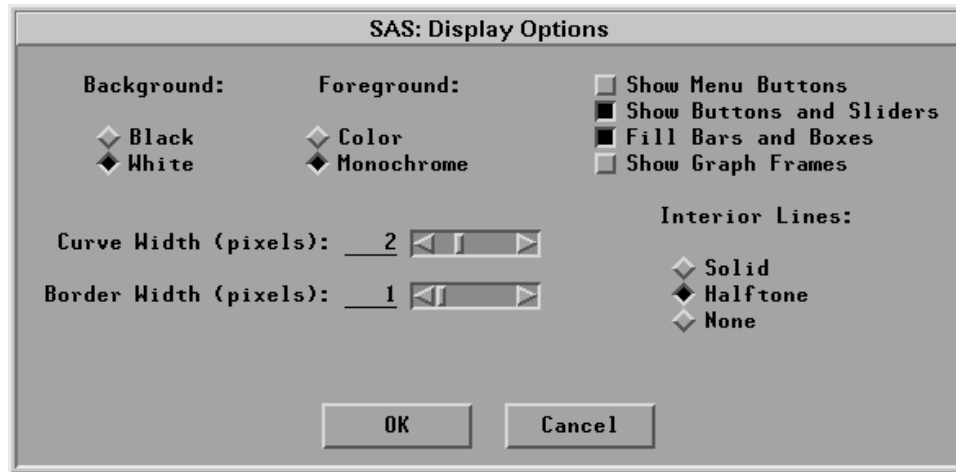


Figure 29.14. Setting Display Options

- ⇒ Click **OK** to set the display options and close the dialog
- ⇒ Choose **Edit:Windows:Window Options** to display the window options dialog
- ⇒ Set the **Default Margin** to **1 mm**.



Figure 29.15. Setting Window Options

- ⇒ Click **OK** to set the window options and close the dialog
- ⇒ Choose **Edit:Windows:Graph Options** to display the graph options dialog
- ⇒ Set the **Marker Size** to **100%**.

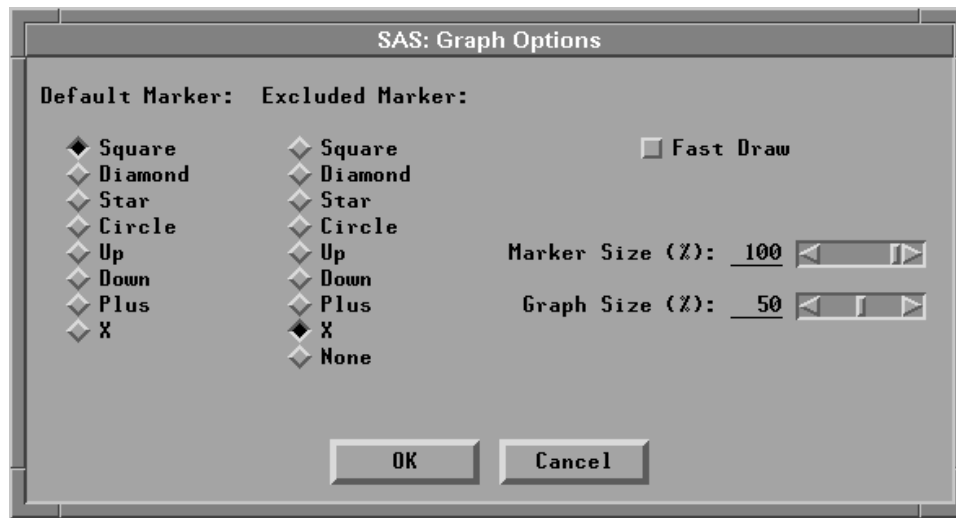


Figure 29.16. Setting Graph Options

- ⇒ Click **OK** to set the graph options and close the dialog
- ⇒ Choose **Edit:Windows:Renew** in the fit analysis window.
This displays the fit analysis variables dialog.

⇒ Click **OK** in the variables dialog

This redisplay the fit analysis with the modified option settings. Contrast [Figure 29.17](#) with [Chapter 39](#). Note that the menu buttons are no longer displayed, the space between the tables and graphs is reduced, and the marker size is increased.

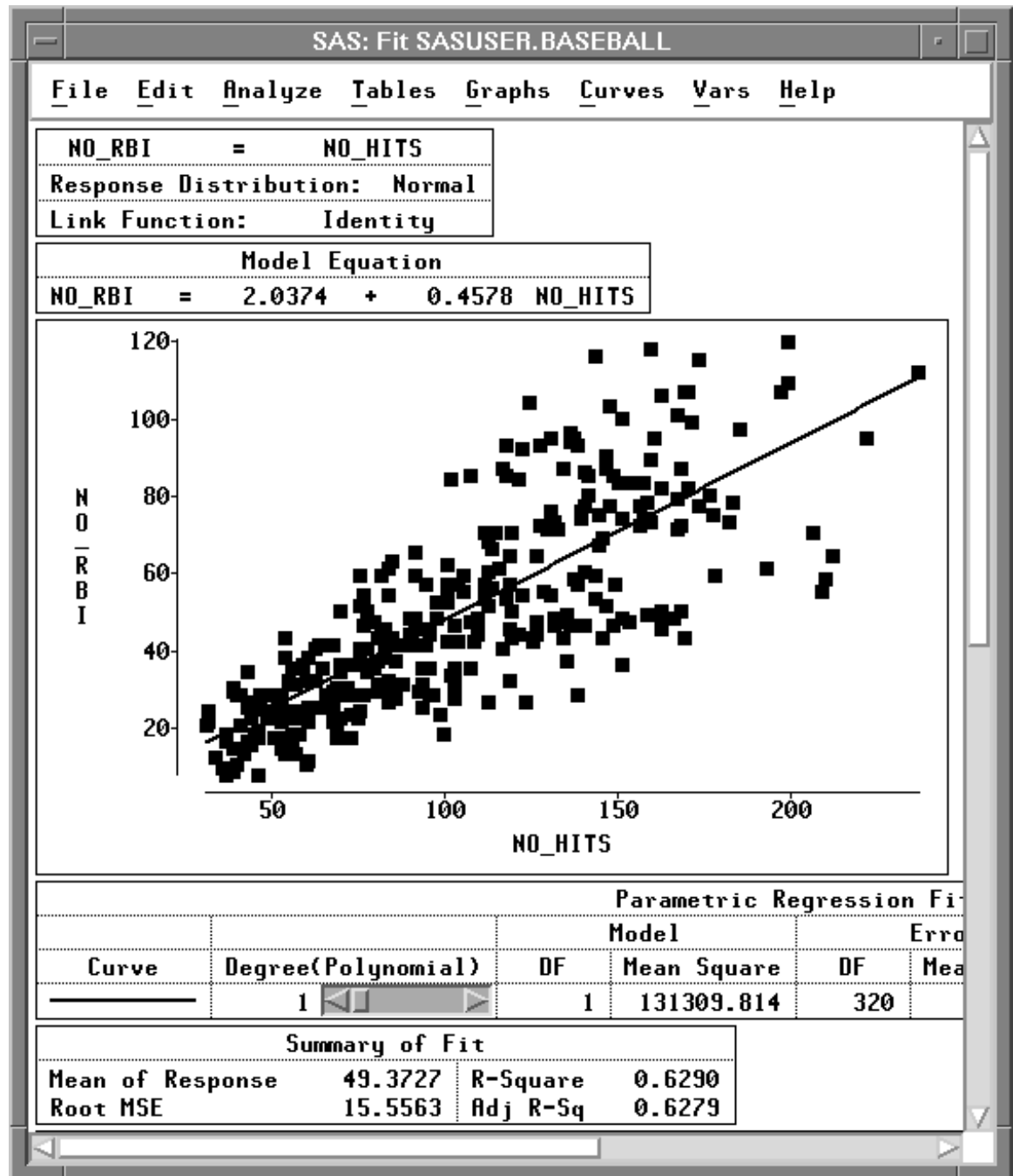


Figure 29.17. Modified Fit Analysis

Saving Options

Once you set any option, it remains in effect for the rest of your SAS/INSIGHT session. You can also save options so they become the default for future SAS/INSIGHT sessions by choosing **File:Save:Options**.

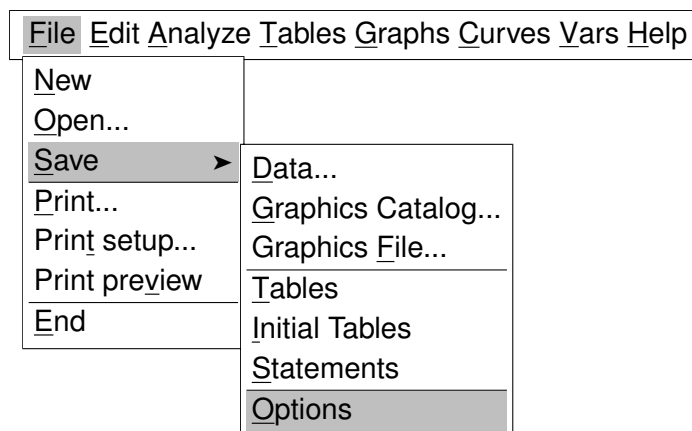


Figure 29.18. File:Save Menu

This saves options for all graphs and analyses, as well as display, window, and graph options, and stores these options in your SASUSER.PROFILE catalog. Option settings are read from SASUSER.PROFILE.INSIGHT and used as default settings the next time you invoke SAS/INSIGHT software. This enables you to tailor SAS/INSIGHT software to the way you work.

Setting Host Resources

You can modify the operation and appearance of SAS/INSIGHT software in ways that are specific to your host by setting *host resources*. For details on host resources, refer to the SAS companion for your host.

If you are on a UNIX host running X Windows, the behavior of the SAS System is determined by X resources. The following X resources improve the performance of SAS/INSIGHT software.

```
# SAS resources
SAS.windowUnitType:      percentage
SAS.windowHeight:       90
SAS.windowWidth:        100
SAS.maxWindowHeight:    90
SAS.maxWindowWidth:     100
SAS.sessionGravity:     NorthWestGravity

# Motif resources
Mwm*IconPlacement:      right bottom
Mwm*InteractivePlacement: false
Mwm*ClientAutoPlace:    false
Mwm*KeyboardFocusPolicy: pointer
```

These SAS resources and Motif resources enable the SAS System to use 90% of the display and enable SAS/INSIGHT software to place windows efficiently when you set the **Window Layout:Spread** option. If your host does not use the Motif window manager, it may use another window manager with similarly named resources.

Resource names are case-sensitive. You can load X resources at system initialization or use the UNIX **xrdb** command. For more information on X resources, refer to the SAS companion for the UNIX environment or your host documentation.

Chapter 30

Working with Other SAS Products

Chapter Contents

VIEWING RESULTS FROM SAS/STAT SOFTWARE	472
SUBMITTING SAS/INSIGHT STATEMENTS	478
RECORDING SAS/INSIGHT STATEMENTS	481
REFERENCES	482

Chapter 30

Working with Other SAS Products

This chapter illustrates how to use SAS/INSIGHT software with other components of the SAS System.

A typical usage is to create an analysis in another SAS product and then view the results using SAS/INSIGHT software. For example, you can use SAS/STAT software to create an analysis and use SAS/INSIGHT software to display its results. This enables you to take advantage of the strengths of both products.

You can also use grammar statements to drive SAS/INSIGHT software from other SAS products. This enables you to save time by automating repetitive tasks.

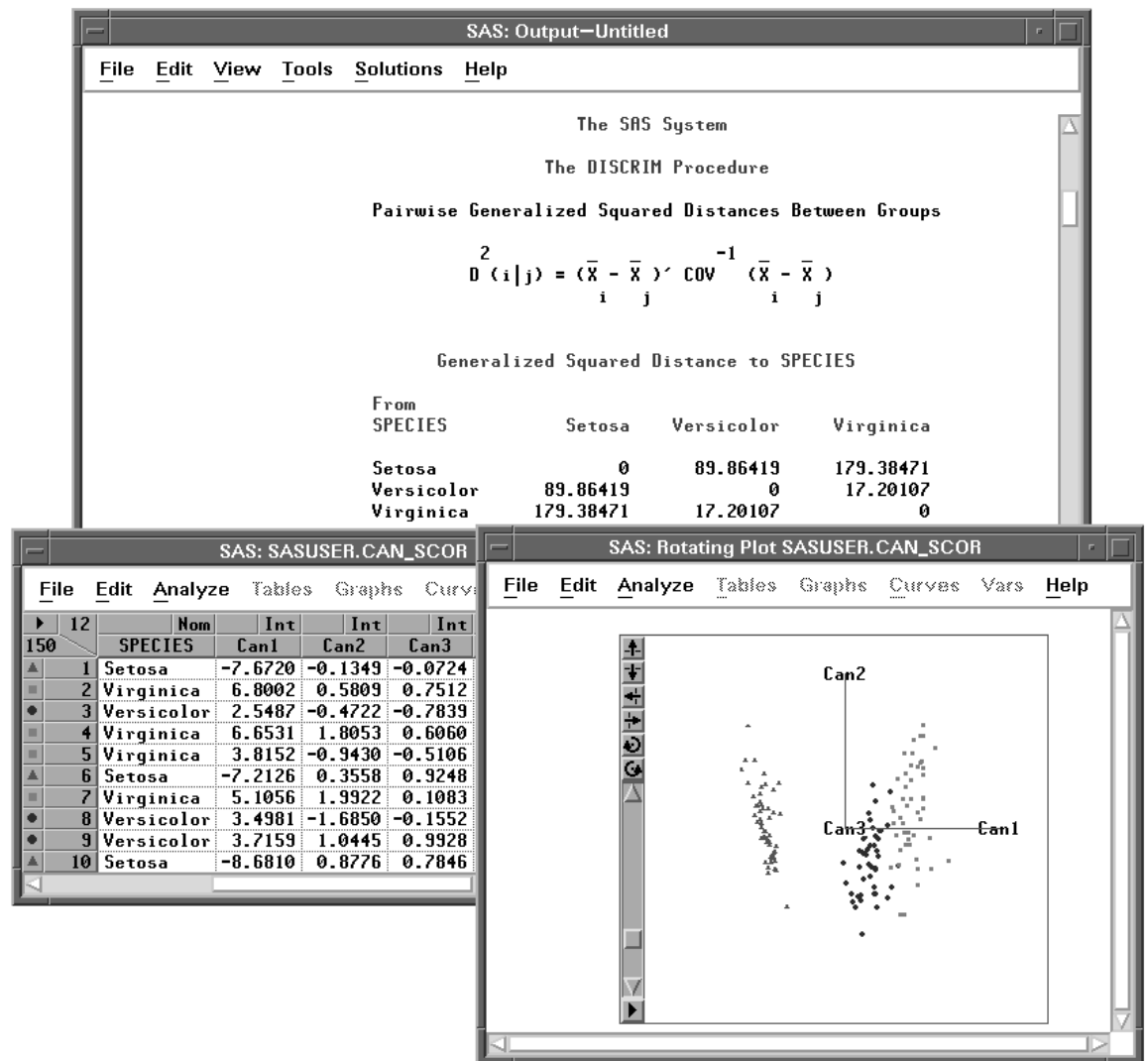


Figure 30.1. Viewing Results from SAS/STAT Software

Viewing Results from SAS/STAT Software

The **IRIS** data, published by Fisher (1936), have been used widely for examples in discriminant analysis. The goal of the analysis is to find functions of a set of quantitative variables that best summarize the differences among groups of observations determined by the classification variable. The **IRIS** data contain four quantitative variables measured on 150 specimens of iris plants. These include sepal length (**SEPALLEN**), sepal width (**SEPALWID**), petal length (**PETALLEN**), and petal width (**PETALWID**). The classification variable, **SPECIES**, represents the species of iris from which the measurements were taken. There are three species in the data: *Iris setosa*, *Iris versicolor*, and *Iris virginica*.

SAS: SASUSER.IRIS															
File		Edit		Analyze		Tables		Graphs		Curves		Vars		Help	
▶	5	Int	Int	Int	Int	Nom									
150		SEPALLEN	SEPALWID	PETALLEN	PETALWID	SPECIES									
■	1	50	33	14	2	Setosa									
■	2	64	28	56	22	Virginica									
■	3	65	28	46	15	Versicolor									
■	4	67	31	56	24	Virginica									
■	5	63	28	51	15	Virginica									
■	6	46	34	14	3	Setosa									
■	7	69	31	51	23	Virginica									
■	8	62	22	45	15	Versicolor									
■	9	59	32	48	18	Versicolor									
■	10	46	36	10	2	Setosa									

Figure 30.2. IRIS Data Set

Linear combinations of the four measurement variables best summarize the differences among the three species, assuming multivariate normality with covariance constant among groups. This requires a canonical discriminant analysis that is available in both SAS/INSIGHT software and SAS/STAT software. The following steps illustrate how to create an output data set that contains scores on the canonical variables in SAS/STAT software and how to use SAS/INSIGHT software to plot them.

⇒ **If you are running the SAS System in interactive line mode, exit the SAS System and reenter under the display manager.**

You must invoke SAS/INSIGHT software from a command line or from the **Solutions** menu to use SAS/INSIGHT software and the Program Editor concurrently.

⇒ **In the Program Editor, enter the statements shown in Figure 30.3.**

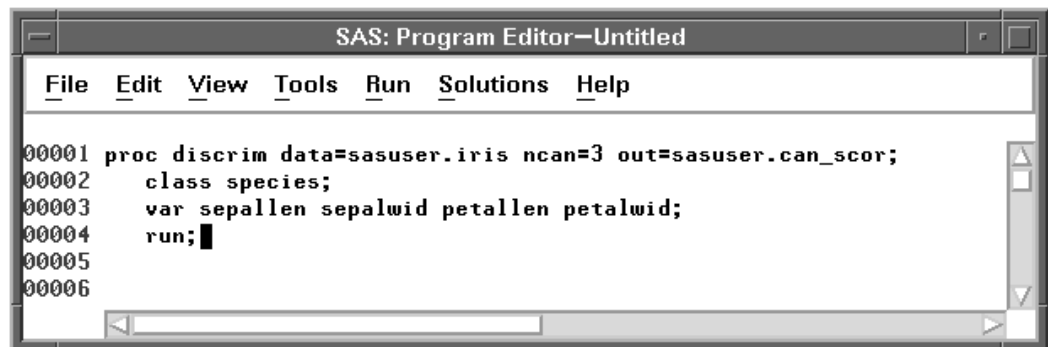


Figure 30.3. Program Editor with PROC Statement

The **OUT=** option in the **PROC DISCRIM** statement puts the scores and the original variables in the **SASUSER** library in a data set called **CAN_SCOR**. For complete documentation on the **DISCRIM** procedure, refer to the chapter titled “The DISCRIM Procedure,” in the *SAS/STAT User’s Guide*.

⇒ **In the Program Editor, enter the statements in Figure 30.4.**

These statements create the **_OBSTAT_** variable, which stores observation colors, shapes, and other states. If you create the **_OBSTAT_** variable as shown, **SETOSA** observations will be red triangles, **VERSICOLOR** observations will be blue circles, and **VIRGINICA** observations will be magenta squares.

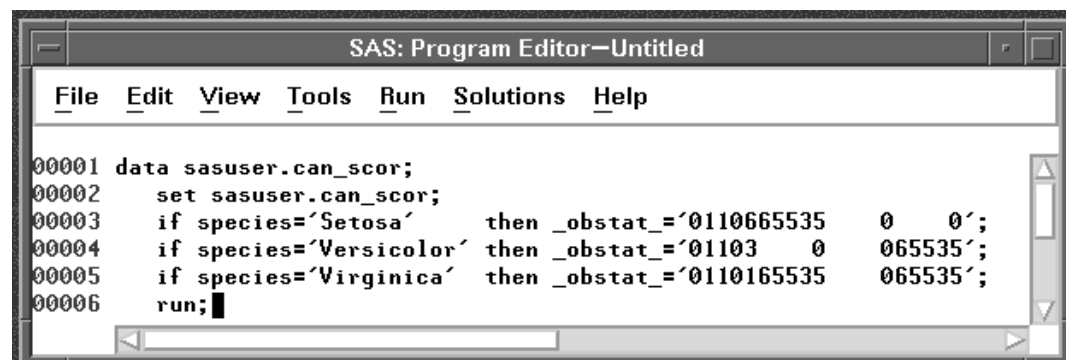


Figure 30.4. Program Editor with DATA Step

OBSTAT is a character variable. You can use it to set other observation states in addition to color and shape. The format of the **_OBSTAT_** variable is as follows.

- | | |
|-------------|--|
| Character 1 | stores the observation’s selection state. It is '1' for selected observations and '0' for observations that are not selected. |
| Character 2 | stores the observation’s Show/Hide state. It is '1' for observations that are displayed in graphs and '0' for observations that are not displayed in graphs. |

- | | | | | | | | | | | | | | | | | | |
|-----------------|--|---|--------|---|------|---|--------|---|---------|---|---|---|-------------|---|---------------|---|------|
| Character 3 | stores the observation's Include/Exclude state. It is '1' for observations that are included in calculations and '0' for observations that are excluded from calculations. | | | | | | | | | | | | | | | | |
| Character 4 | stores the observation's Label/UnLabel state. It is '1' for observations whose label is displayed by default, and '0' for observations whose label is not displayed by default. | | | | | | | | | | | | | | | | |
| Character 5 | stores the observation's marker shape, a value between '1' and '8': <table border="0" style="margin-left: 40px;"> <tr><td>1</td><td>Square</td></tr> <tr><td>2</td><td>Plus</td></tr> <tr><td>3</td><td>Circle</td></tr> <tr><td>4</td><td>Diamond</td></tr> <tr><td>5</td><td>X</td></tr> <tr><td>6</td><td>Up Triangle</td></tr> <tr><td>7</td><td>Down Triangle</td></tr> <tr><td>8</td><td>Star</td></tr> </table> | 1 | Square | 2 | Plus | 3 | Circle | 4 | Diamond | 5 | X | 6 | Up Triangle | 7 | Down Triangle | 8 | Star |
| 1 | Square | | | | | | | | | | | | | | | | |
| 2 | Plus | | | | | | | | | | | | | | | | |
| 3 | Circle | | | | | | | | | | | | | | | | |
| 4 | Diamond | | | | | | | | | | | | | | | | |
| 5 | X | | | | | | | | | | | | | | | | |
| 6 | Up Triangle | | | | | | | | | | | | | | | | |
| 7 | Down Triangle | | | | | | | | | | | | | | | | |
| 8 | Star | | | | | | | | | | | | | | | | |
| Characters 6–20 | store the observation's color as Red-Green-Blue (RGB) components. The RGB color model represents colors as combinations of the colors red, green, and blue. You can obtain intermediate colors by varying the proportion of these primary colors.

Each component is a 5-digit decimal number between 0 and 65535. Characters 6–10 store the red component. Characters 11–15 store the green component. Characters 16–20 store the blue component. | | | | | | | | | | | | | | | | |

The **_OBSTAT_** variable can be used to create color blends as well as discrete colors. For an example of this usage, refer to Robinson (1995).

⇒ **Choose Run:Submit to submit the SAS statements.**

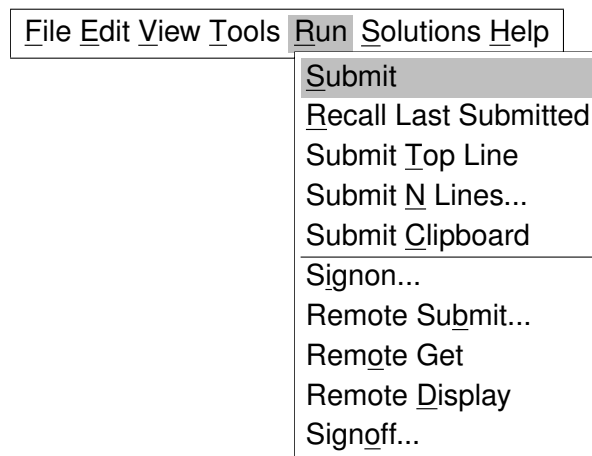
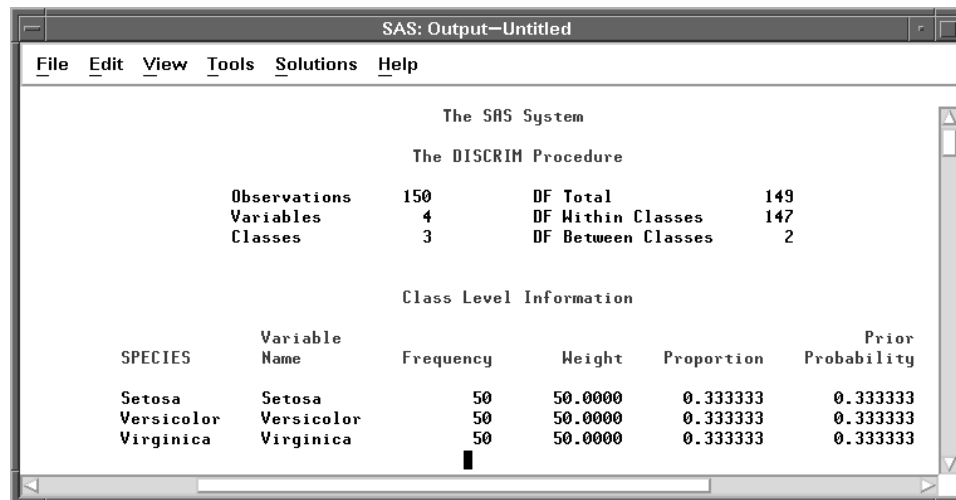


Figure 30.5. Run Menu

This produces the PROC DISCRIM output shown in Figure 30.6 and creates the **CAN_SCOR** data set.



The screenshot shows the SAS Output window titled "SAS: Output—Untitled". It displays the results of the PROC DISCRIM procedure. The output is organized into sections: "The SAS System", "The DISCRIM Procedure", and "Class Level Information".

The DISCRIM Procedure

	Observations	150	DF Total	149
Variables	4		DF Within Classes	147
Classes	3		DF Between Classes	2

Class Level Information

SPECIES	Variable Name	Frequency	Height	Proportion	Prior Probability
Setosa	Setosa	50	50.0000	0.333333	0.333333
Versicolor	Versicolor	50	50.0000	0.333333	0.333333
Virginica	Virginica	50	50.0000	0.333333	0.333333

Figure 30.6. PROC DISCRIM Output

- ⇒ Invoke SAS/INSIGHT software, and open the **CAN_SCOR** data set.
- ⇒ Scroll to the right to see the canonical variables **CAN1**, **CAN2**, and **CAN3**.
These variables represent the linear combinations of the four measurement variables that summarize the differences among the three species.

SAS: SASUSER.CAN_SCOR

FileEditAnalyzeTablesGraphsCurvesVarsHelp

▶	12		Int		Int		Nom		Int		Int		Int	
150			PETALLEN		PETALWID		SPECIES		Can1		Can2		Can3	
▲	1		14		2		Setosa		-7.6720		-0.1349		-0.0724	
■	2		56		22		Virginica		6.8002		0.5809		0.7512	
●	3		46		15		Versicolor		2.5487		-0.4722		-0.7839	
■	4		56		24		Virginica		6.6531		1.8053		0.6060	
■	5		51		15		Virginica		3.8152		-0.9430		-0.5106	
▲	6		14		3		Setosa		-7.2126		0.3558		0.9248	
■	7		51		23		Virginica		5.1056		1.9922		0.1083	
●	8		45		15		Versicolor		3.4981		-1.6850		-0.1552	
●	9		48		18		Versicolor		3.7159		1.0445		0.9928	
▲	10		10		2		Setosa		-8.6810		0.8776		0.7846	

Figure 30.7. CAN_SCOR Data

By plotting the canonical variables, you can visualize how well the variables discriminate among the three groups. Canonical variables, having more discriminatory power, show more separation among the groups in their associated axes on a plot, while variables having little discriminatory power show little separation among groups.

⇒ Choose **Analyze:Rotating Plot (Z Y X)**. Assign **CAN3** the Z role, **CAN2** the Y role, and **CAN1** the X role.

This produces a plot with the **CAN3** axis pointing toward you, showing clear separation of the species.

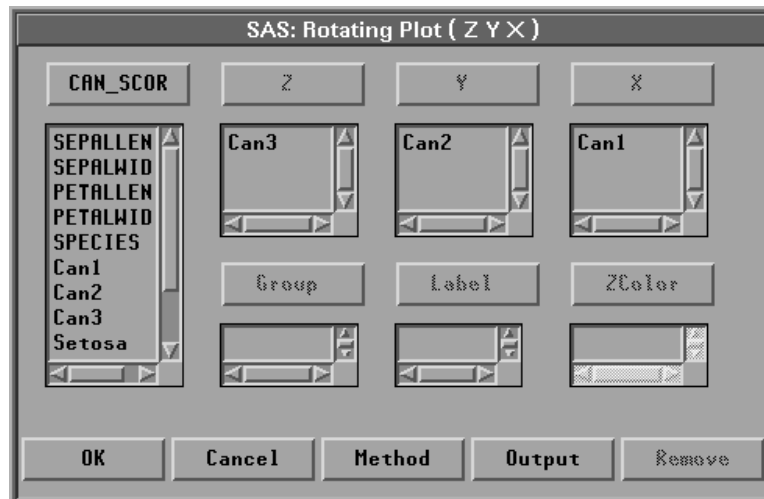


Figure 30.8. Rotating Plot Dialog

⇒ Click **OK** in the dialog to create the rotating plot.

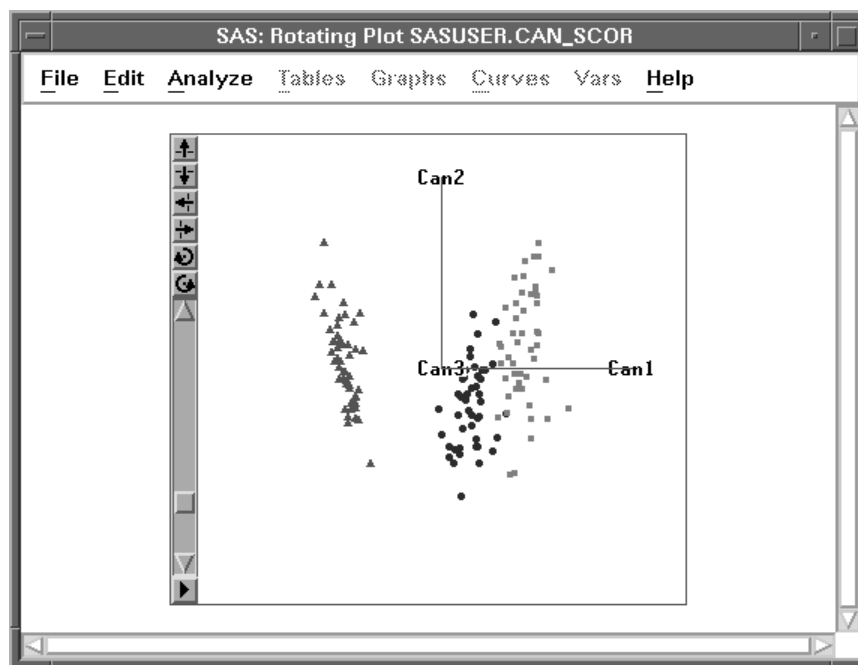


Figure 30.9. Rotating Plot, **CAN3** Toward Viewer

⇒ **Rotate the plot so the axis representing CAN1 points toward you.**

Refer to [Chapter 6, “Exploring Data in Three Dimensions,”](#) for information on how to rotate plots. This orientation shows little, if any, differentiation among species. This is because **CAN2** and **CAN3** contribute little information towards separating the groups.

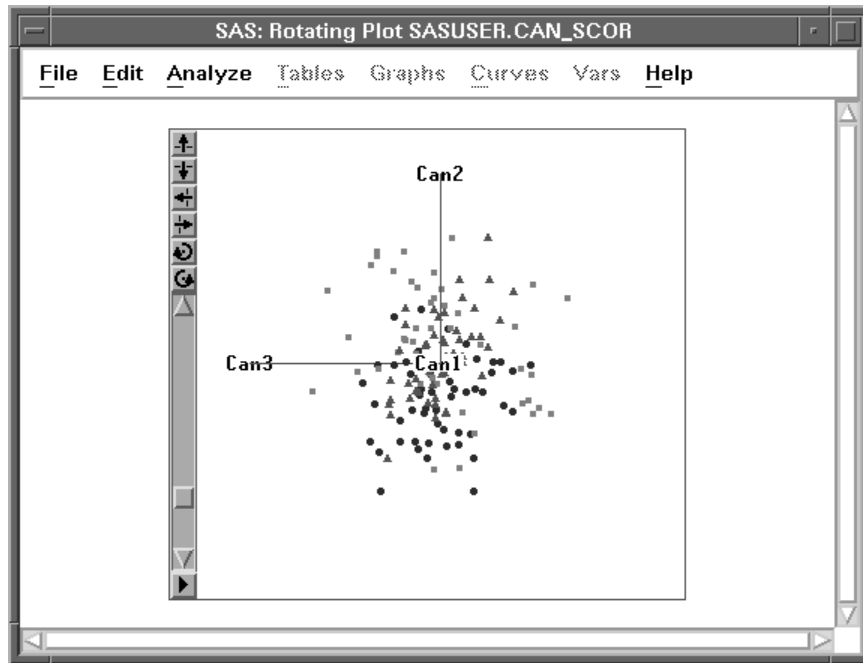


Figure 30.10. Rotating Plot, **CAN1** Toward Viewer

Another way of illustrating this would be to create a scatter plot matrix of **CAN1**, **CAN2**, and **CAN3**. Only plots involving **CAN1** would show much group differentiation. The **CAN2**-by-**CAN3** plot would show little or no group differentiation.

⊕ **Related Reading:** Rotating Plots, [Chapter 6](#), [Chapter 37](#).

Submitting SAS/INSIGHT Statements

If this analysis were a task you perform frequently, you could save time by automating the creation of the rotating plot. To do this, you can submit SAS/INSIGHT statements in the Program Editor.

You can submit statements when SAS/INSIGHT is executing either as a procedure or as a task. To submit statements to the procedure, do the following.

- ⇒ **Choose File:End in the data window to exit SAS/INSIGHT.**
- ⇒ **In the Program Editor, enter the statements shown in Figure 30.11.**

The DATA option opens the **CAN_SCOR** data set. The ROTATE statement creates the rotating plot.

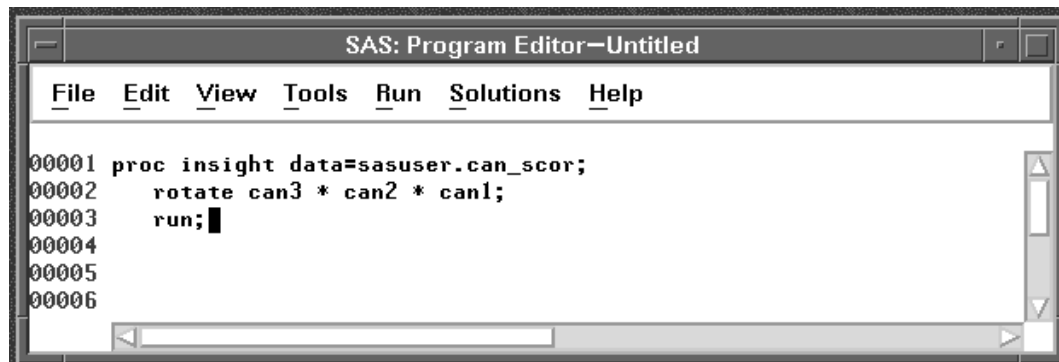


Figure 30.11. SAS/INSIGHT Statements in Program Editor

- ⇒ **Choose Run:Submit to submit the SAS statements.**
This opens the data set and creates the plot.

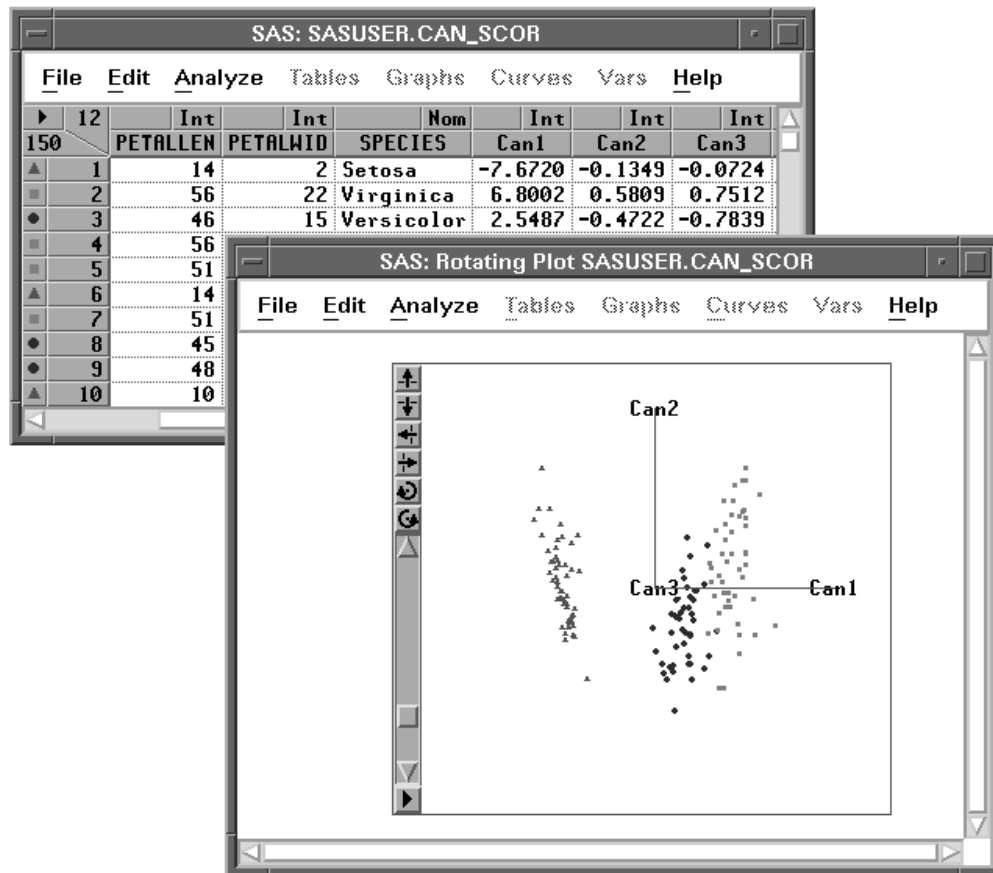


Figure 30.12. Data Window and Rotating Plot

It is often preferable to invoke SAS/INSIGHT as a task instead of a procedure. It is sometimes preferable to open a data set without displaying it. To invoke SAS/INSIGHT as a task and display a rotating plot without a data window, follow these steps.

⇒ **Store the following three statements in a text file called myfile.**

```
open sasuser.can_scor / nodisplay;
rotate can3 * can2 * can1;
run;
```

⇒ **In the Program Editor, enter the FILENAME statement shown in Figure 30.13.**
The **FILENAME** statement assigns a fileref.

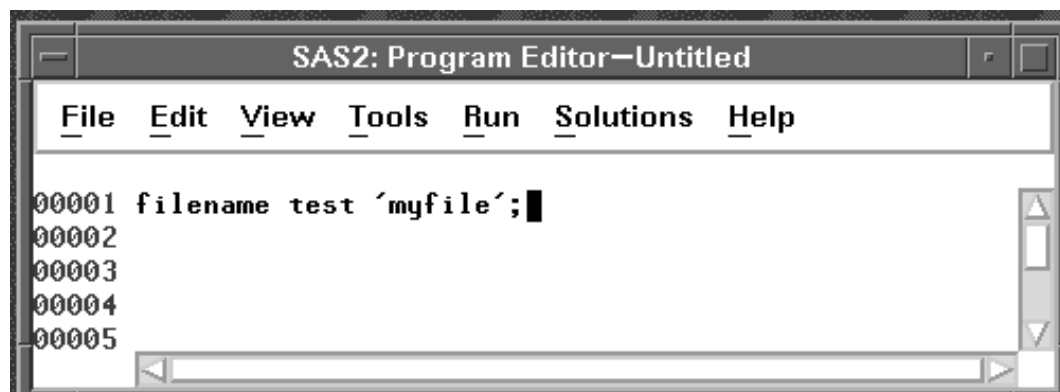


Figure 30.13. Submitting a FILENAME statement

⇒ Choose **Run:Submit** to submit the statement

⇒ Invoke SAS/INSIGHT as a task with the **INFILE=** option.

You can invoke SAS/INSIGHT on the command line with the statement

```
insight infile=test
```

This opens the data set **SASUSER.CAN_SCORE** without displaying it and then creates a rotating plot of **CAN3** versus **CAN2** versus **CAN1**.

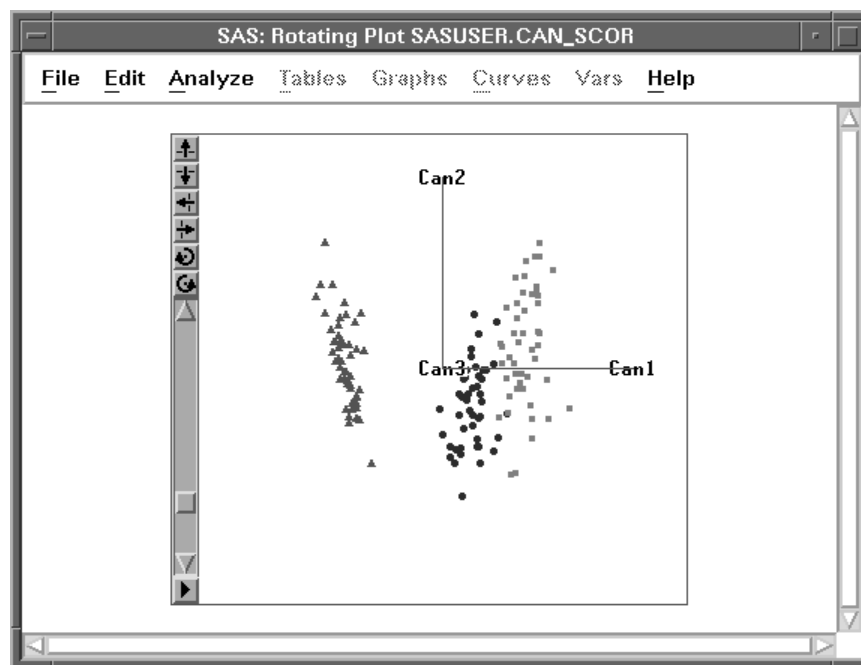


Figure 30.14. Rotating Plot

You can use grammar statements such as these to drive SAS/INSIGHT software from SAS/AF software. For portability, statements can be stored in catalog entries by using a FILENAME statement with the keyword LIBRARY. For example, if you stored statements in a catalog entry **sasuser.insight.test.source**, you could assign the fileref with the statement

```
filename test library 'sasuser.insight.test.source';
```

For SAS/AF applications, you can improve the display of SAS/INSIGHT windows by suppressing the display of menus, buttons, and confirmation dialogs. You can also save options to configure your graphs and analyses. These techniques are described in [Chapter 41, “SAS/INSIGHT Statements,”](#) and [Chapter 29, “Configuring SAS/INSIGHT Software.”](#)

Recording SAS/INSIGHT Statements

SAS/INSIGHT statements also provide a record of the analyses you create, including model equations. You can record your SAS/INSIGHT session using the **File:Save:Statements** menu or the FILE= option.

To create a record of your SAS/INSIGHT session, follow these steps.

- ⇒ **Invoke SAS/INSIGHT and open the BUSINESS data set.**
- ⇒ **Choose File:Save:Statements.**
This toggles the recording of statements to the SAS log.

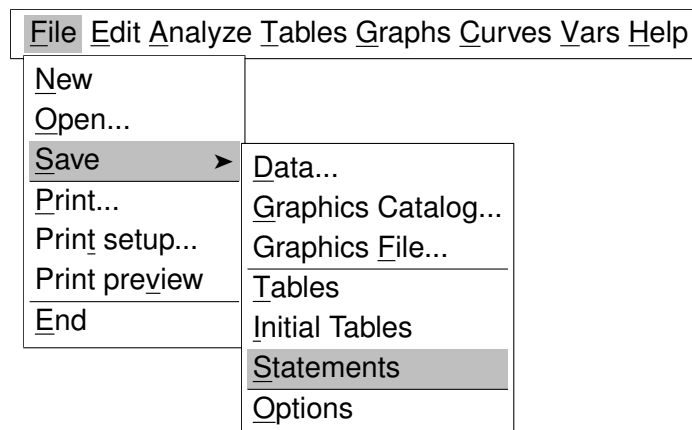


Figure 30.15. File:Save Menu

- ⇒ **Create graphs and analyses as you like.**
The Log window displays a record of your actions. For example, a record of three model fits might look like the following.

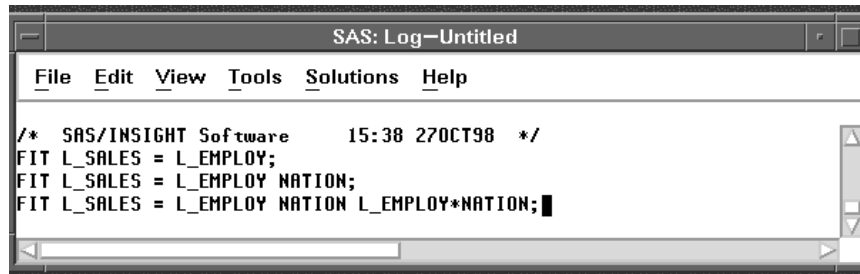


Figure 30.16. Log Window

Recorded output uses the same syntax as statement input, so you can replay the statements you record. However, intermediate events such as transformation of variables, exclusion of observations, and data entry are not recorded. Therefore, replaying will not always reproduce the original analysis.

As an alternative to the **File:Save:Statements** menu, you can use the FILE= option when you invoke SAS/INSIGHT. The FILE= option and other options are described in [Chapter 41, “SAS/INSIGHT Statements.”](#)

References

- Fisher, R.A. (1936), “The Use of Multiple Measurements in Taxonomic Problems,” *Annals of Eugenics*, 7, 179–188.
- Robinson, H. (1995), “Batch Processing in SAS/INSIGHT Software,” *Proceedings of the 20th Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., 560.

Part 3

Reference

Contents

Chapter 31. Data Windows	485
Chapter 32. Histograms and Bar Charts	497
Chapter 33. Box Plots and Mosaic Plots	505
Chapter 34. Line Plots	519
Chapter 35. Scatter Plots	525
Chapter 36. Contour Plot	533
Chapter 37. Rotating Plot	543
Chapter 38. Distribution Analyses	553
Chapter 39. Fit Analyses	611
Chapter 40. Multivariate Analyses	705
Chapter 41. SAS/INSIGHT Statements	777

Reference

Chapter 31

Data Windows

Chapter Contents

OPENING A DATA WINDOW	488
VARIABLES	489
OBSERVATIONS	491
THE DATA MENU	493

Chapter 31

Data Windows

A *data window* displays a SAS data set as a table, with columns of the table containing variables and rows containing observations.

In a data window, you can sort, search, edit, and extract subsets of your data. You can also assign measurement levels and default roles that determine how your variables are used in graphs and analyses.

	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME
1	Aldrete, Mike	SanFrancisco	216	54	2
2	Allanson, Andy	Cleveland	293	66	1
3	Almon, Bill	Pittsburgh	196	43	7
4	Anderson, Dave	LosAngeles	216	53	1
5	Armas, Tony	Boston	425	112	11
6	Ashby, Alan	Houston	315	81	7
7	Backman, Wally	NewYork	387	124	1
8	Baines, Harold	Chicago	570	169	21
9	Baker, Dusty	Oakland	242	58	4
10	Balboni, Steve	KansasCity	512	117	29
11	Bando, Chris	Cleveland	254	68	2
12	Barfield, Jesse	Toronto	589	170	40
13	Barrett, Marty	Boston	625	179	4
14	Bass, Kevin	Houston	591	184	20
15	Baylor, Don	Boston	585	139	31
16	Beane, Billy	Minneapolis	183	39	3
17	Bell, Buddy	Cincinnati	568	158	20
18	Bell, George	Toronto	641	198	31
19	Belliard, Rafael	Pittsburgh	309	72	0
20	Beniquez, Juan	Baltimore	343	103	6
21	Bernazard, Tony	Cleveland	562	169	17
22	Biancalana, Buddy	KansasCity	190	46	2
23	Bilardello, Dann	Montreal	191	37	4
24	Bochte, Bruce	Oakland	407	104	6
25	Bochy, Bruce	SanDiego	127	32	8

Figure 31.1. Data Window

Opening a Data Window

You can open data windows in several ways. One way is to specify a data set with the `DATA=` option when you invoke SAS/INSIGHT software. If you do not specify a data set, a data set dialog appears.



Figure 31.2. Data Set Dialog

This dialog displays two lists: **Library** and **Data Set**. A *library* is a location where data sets are stored. The **Library** list always contains the standard libraries **WORK**, **MAPS**, **SASHELP**, and **SASUSER**. You can define other libraries using the `LIBNAME` statement. For more information on the `LIBNAME` statement, refer to *SAS Language Reference: Dictionary*.

By default, **SASUSER** is selected in the **Library** list. To see the data sets in any other library, click on the library's name. This causes the **Data Set** list to display all data sets in that library. For information on how to create SAS data sets, see [Chapter 2, "Entering Data."](#)

By default, the first data set in the **Data Set** list is selected. To select another data set, click on its name. Then click on **OK** to display the data window. On many hosts, instead of clicking on the data set name, then on **OK**, you can *double-click* on the data set name to open the data set and close the dialog.

The **Options** button on the dialog enables you to enter `WHERE` clauses and other SAS data set options. For information on data set options, refer to *SAS Language Reference: Dictionary*.

You can also open a data window with the **File:Open** menu.

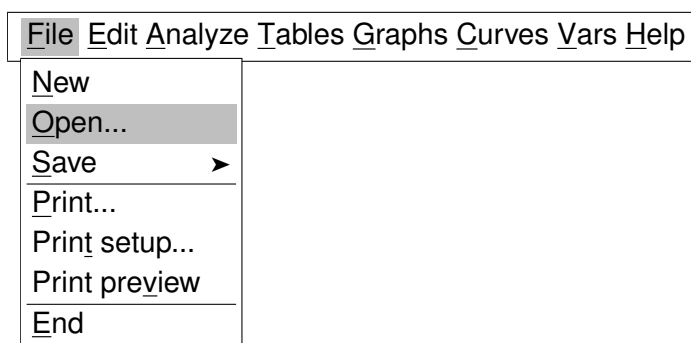


Figure 31.3. File Menu

This displays the data set dialog as described previously.

You can open any number of data windows on different data sets, but you can open only one data window on each data set.

Variables

The column headings in a data window give information on each variable, including the name, label, default roles, and measurement level. The number of variables appears in the upper left corner of the data window.

Label	Nom	Group	Nom	Int	Int	Int
NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME		
1 Aldrete, Mike	SanFrancisco	216	54	2		
2 Allanson, Andy	Cleveland	293	66	1		
3 Almon, Bill	Pittsburgh	196	43	7		
4 Anderson, Dave	LosAngeles	216	53	1		
5 Armas, Tony	Boston	425	112	11		
6 Ashby, Alan	Houston	315	81	7		
7 Backman, Hally	NewYork	387	124	1		
8 Baines, Harold	Chicago	570	169	21		

Figure 31.4. Variables

A variable's *default role* assigns the role a variable plays by default in graphs and analyses. Click in the upper left portion of the variable header to display a pop-up menu of variable roles.

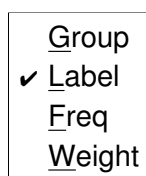


Figure 31.5. Variable Roles Pop-up Menu

You can assign four default roles:

Group	enables you to process your data by groups. You can use multiple group variables to process your data by groups for each unique combination of values of the group variables.
Label	labels observations in scatter plots, rotating plots, and box plots.
Frequency	represents the frequency of occurrence for other values in each observation.
Weight	supplies weights for each observation.

You can assign **Freq**, **Weight**, and **Label** roles to only one variable at a time. You can assign the **Group** role to more than one variable. The order in which you assign the group role determines the order in which the variables are used to define groups.

A variable's *measurement level* determines the way it is treated in graphs and analyses.

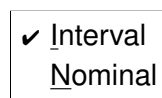


Figure 31.6. Measurement Levels Pop-up Menu

You can assign two measurement levels:

Interval	contains values that vary across a continuous range. For example, a variable measuring temperature would likely be an interval variable. Numeric variables default to the interval measurement level but can be changed to nominal.
Nominal	contains a discrete set of values. For example, a variable indicating gender would be a nominal variable. Character variables can use only the nominal measurement level.

Up to 250 variable measurement levels can be stored with a data set.

Default roles and measurement levels are displayed in the column headings above the variable names. The default role appears at the upper left of the column heading and the measurement level appears at the upper right. If a variable has more than one default role, then only the first character of each role appears.

In [Figure 31.4](#), **NAME** has a label default role, and **TEAM** has a group default role. **NAME** and **TEAM** both have a nominal measurement level, while the remaining variables have an interval measurement level.

† **Note:** Up to 250 measurement levels can be stored in the SAS data set. You can use the data pop-up menu to create new variables or to change the default role or measurement level of existing variables. For more information, see the section “Data Menu” later in this chapter.

You can use the **Edit:Variables** menu to create new variables that are transformations of existing variables. See [Chapter 20, “Transforming Variables,”](#) for more information.

Observations

The row headings in a data window give information on each observation, including the observation states and observation number. The total number of observations appears in the upper left corner of the data window.

	22	Label	Nom	Group	Nom	Int	Int	Int
	322	NAME		TEAM		NO_ATBAT	NO_HITS	NO_HOME
■	1	Aldrete, Mike		SanFrancisco		216	54	2
■	2	Allanson, Andy		Cleveland		293	66	1
■	3	Almon, Bill		Pittsburgh		196	43	7
■	4	Anderson, Dave		LosAngeles		216	53	1
■	5	Armas, Tony		Boston		425	112	11
■	6	Ashby, Alan		Houston		315	81	7
■	7	Backman, Wally		NewYork		387	124	1
■	8	Baines, Harold		Chicago		570	169	21

Figure 31.7. Observations

SAS/INSIGHT software supports the following observation states:

Marker	shows the shape of the marker used in scatter plots, rotating plots, and box plots.
Color	shows the color of the observation.
Label/UnLabel	tells whether a label is displayed by default.
Show/Hide	tells whether an observation is displayed in graphs.
Include/Exclude	tells whether an observation is included in calculations for curves and analysis tables.
Select	tells whether an observation is selected.

An observation's marker and color appear at the left side of the row heading, as shown in [Figure 31.7](#).

An observation's Label/UnLabel state is shown by a picture of a label around the observation number if the observation's label is displayed by default. In [Figure 31.7](#), observations **2**, **4**, and **8** are labeled.

An observation's Show/Hide state is shown by whether or not a marker is displayed in the row heading. In [Figure 31.7](#), observations **2**, **3**, and **6** are hidden.

An observation's Include/Exclude state is shown by the way the observation number is displayed. The observation number is grayed-out for observations that are excluded from calculations. In [Figure 31.7](#), observations **5** and **6** are excluded.

An observation's select state is shown by whether the row heading is highlighted or not. In [Figure 31.7](#), observations **1**, **2**, **6**, and **8** are selected.

You can use the **Edit:Observations** menu to set all of these observation states. This menu also enables you to find observations meeting a specific search criterion or to examine observations in detail.

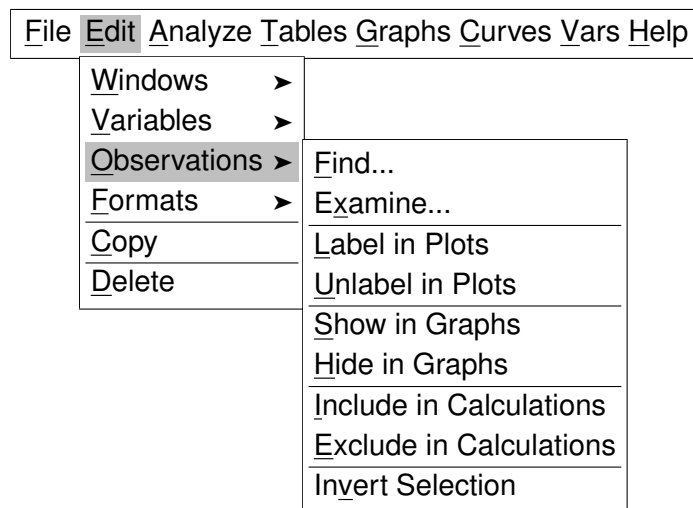


Figure 31.8. Edit Observations Menu

You can also use the observation pop-up menu to set observation states. To see this menu for a particular observation, click on the observation's marker.

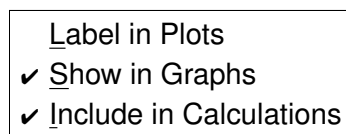


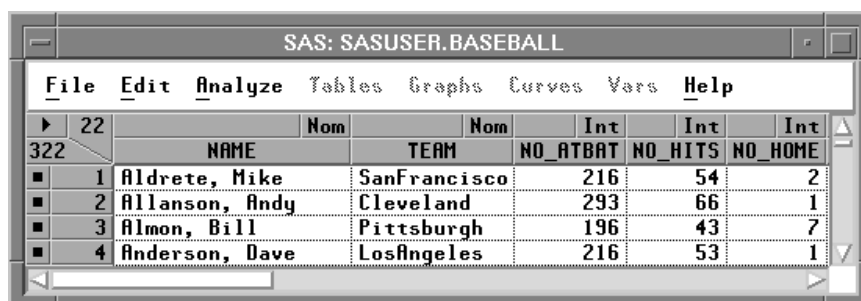
Figure 31.9. Observation Pop-up Menu

† **Note:** SAS/INSIGHT software saves observation states when you save a data set and restores them when you read a data set.

- ⊕ **Related Reading:** Label/Unlabel, [Chapter 8](#).
- ⊕ **Related Reading:** Show/Hide, [Chapter 9](#).
- ⊕ **Related Reading:** Include/Exclude, [Chapter 21](#).
- ⊕ **Related Reading:** Saving Observation States, [Chapter 30](#).

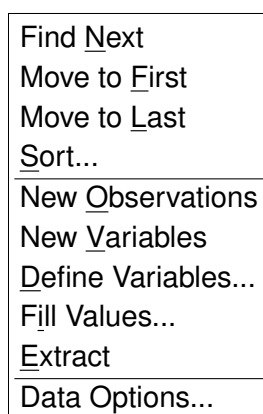
The Data Menu

The data pop-up menu provides a variety of ways to manipulate your data. Display the data pop-up menu by clicking on the button in the upper left corner of the data window.



		Nom	Nom	Int	Int	Int
	NAME	TEAM	NO_ATBAT	NO_HITS	NO_HOME	
1	Aldrete, Mike	SanFrancisco	216	54	2	
2	Allanson, Andy	Cleveland	293	66	1	
3	Almon, Bill	Pittsburgh	196	43	7	
4	Anderson, Dave	LosAngeles	216	53	1	

Figure 31.10. Displaying the Data Pop-up Menu



Find <u>N</u> ext
Move to <u>F</u> irst
Move to <u>L</u> ast
<u>S</u> ort...
New <u>O</u> bservations
New <u>V</u> ariables
<u>D</u> efine Variables...
<u>F</u> ill Values...
<u>E</u> xtract
<u>D</u> ata Options...

Figure 31.11. Data Pop-up Menu

Choose **Find Next** to scroll the data window to the next selected observation. If no observations are selected, it scrolls the data window one observation.

Choose **Move to First** to move selected observations to the top of the data window and to move selected variables to the left side of the data window.

Choose **Move to Last** to move selected observations to the bottom of the data window and to move selected variables to the right side of the data window.

† **Note:** In addition to **Move to First** and **Move to Last**, you can use the hand tool to move variables and observations. Drag on the column or row heading, then release the mouse at the new location.

Choose **Sort** to sort observations on one or more variables. If any variables are selected, your data are sorted in ascending order on the unformatted values of those variables. If no variables are selected, you are prompted with a dialog to select some.



Figure 31.12. Sort Dialog

In the dialog, select variables and click the **Y** button to assign variables to the sort list. You can select variables in the sort list and click the **Asc/Des** and **Unf/For** buttons to toggle the sort order and formatting. If you select multiple variables for the sort, they are used in the order in which you select them.

Choose **New Observations** to add space to enter values for new observations.

Choose **New Variables** to add space To enter values for new variables.

Choose **Define Variables** to display the dialog in [Figure 31.13](#). Use this dialog to set variable type, default roles, measurement level, name, and label.

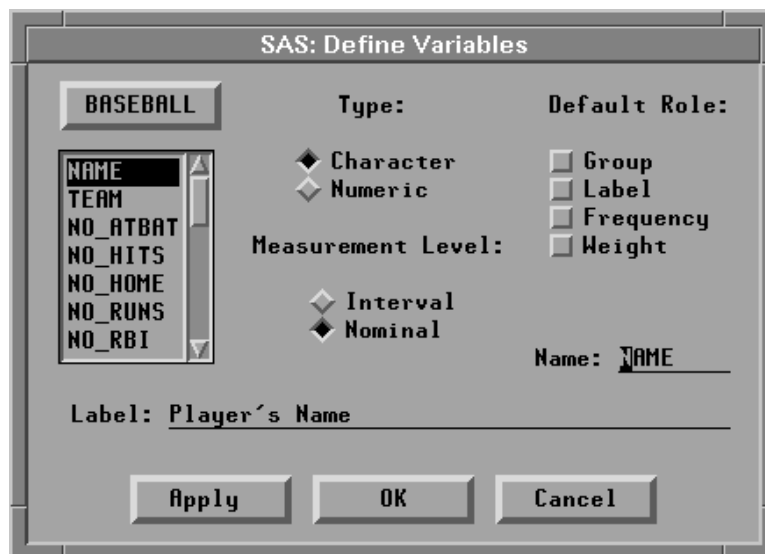


Figure 31.13. Define Variables Dialog

Choose **Fill Values** to modify data values in the data window.

Choose **Extract** to create a new data window from an existing data window. You can **Extract** any subset of your data. If you have variables, observations, or values selected, your selections are extracted to fill the new data window. If you have no selections, you are prompted to select variables.

Choose **Data Options** to set options that control the appearance and operation of the data window.

- ⊕ **Related Reading:** Fill Values, Data Options, [Chapter 2](#).
- ⊕ **Related Reading:** Find, Move to First, Sort, [Chapter 3](#).
- ⊕ **Related Reading:** Define Variables, [Chapter 8](#), [Chapter 15](#), [Chapter 22](#).
- ⊕ **Related Reading:** Extract, [Chapter 21](#).

Chapter 32

Histograms and Bar Charts

Chapter Contents

VARIABLES	500
METHOD	501
OUTPUT	502
REFERENCES	504

Chapter 32

Histograms and Bar Charts

Bar charts are pictorial representations of the distribution of values of a variable.

You can use bar charts to show distributions of interval or nominal variables. Bar charts of interval variables are also called *histograms*.

You can label the heights of the bars in a bar chart, control the orientation, and control the information shown on the axes. For bar charts of interval variables, you can also control the width and offset of the bars.

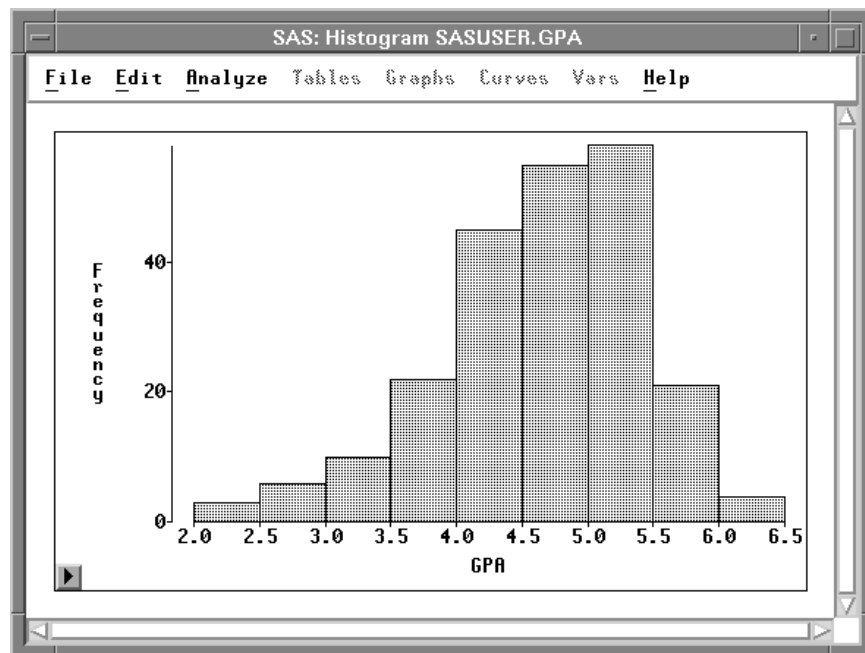


Figure 32.1. Bar Chart

Variables

To create a bar chart, choose **Analyze:Histogram/Bar Chart (Y)**. Bar charts require a **Y** variable. If you have already selected one or more variables, they are assigned the **Y** variable role, and a bar chart is created for each selected variable. If you have not selected any variables, a variables dialog appears.



Figure 32.2. Bar Chart Variables Dialog

In the dialog, select at least one **Y** variable. A separate bar chart is created for each **Y** variable you select.

You can select one or more **Group** variables if you have grouped data. This creates one bar chart for each group.

You can select a **Freq** variable. If you select a **Freq** variable, each observation is assumed to represent n observations, where n is the value of the **Freq** variable.

Method

Observations with missing values for **Y** variables are not used. Observations with **Freq** values that are missing or that are less than or equal to 0 are not used. Only the integer part of **Freq** values is used.

For nominal variables, values that represent less than 4% of the total frequency are grouped together in an “**Other**” category by default. Clicking on the **Method** button in the variables dialog displays the dialog in [Figure 32.3](#). This dialog enables you to change the threshold at which values are grouped into the **Other** category.



Figure 32.3. Bar Chart Method Options Dialog

For interval variables, values that fall on the boundary between two bars are added to the upper bar. For example, if two bars span ranges (1 to 2) and (2 to 3), the value 2 is considered to fall in the range (2 to 3).

By default, bar width and offset are calculated using an algorithm developed from Terrell and Scott (1985). *Bar width* is the distance along the **Y** axis represented by one bar. *Bar offset* is the distance from the start of the bar to the nearest multiple of the bar width. For example, if a bar starts at 1.2 and has a width of 1, then the offset is 0.2.

Output

For nominal variables, bars are distinguished by different colors. For interval variables, all bars have the same color.

To view or modify output options associated with your bar chart, click on the **Output** button of the variables dialog. This displays the options dialog shown in [Figure 32.4](#).

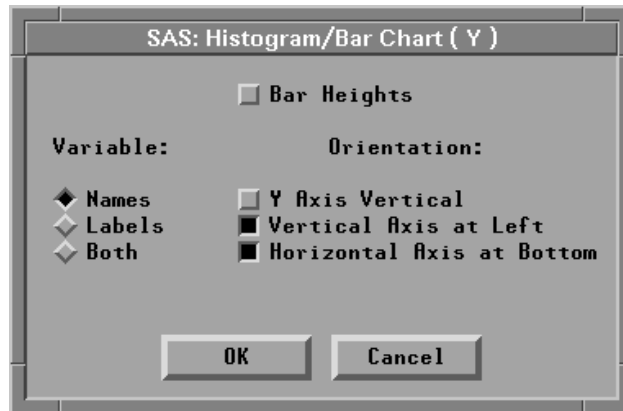


Figure 32.4. Bar Chart Output Options Dialog

Bar Heights	labels all bars with their heights.
Variable:Names	labels the Y axis with variable names.
Variable:Labels	labels the Y axis with variable labels.
Variable:Both	labels the Y axis with both names and labels.
Orientation: Y Axis Vertical	draws the axis for the Y variable vertically. If this option is turned off, the Y axis is horizontal.
Orientation: Vertical Axis at Left	places the vertical axis at the left side of the chart. If this option is turned off, the vertical axis is at the right side of the chart.
Orientation: Horizontal Axis at Bottom	places the horizontal axis at the bottom of the chart. If this option is turned off, the horizontal axis is at the top of the chart.

You can modify other aspects of the bar chart using the bar chart pop-up menu. Click on the button at the lower left corner of the bar chart to display the pop-up menu.

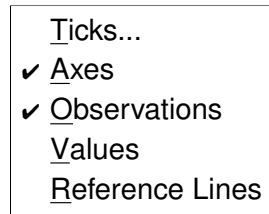


Figure 32.5. Bar Chart Pop-up Menu

- Ticks...** displays the dialog in [Figure 32.6](#) to set tick values for the variable being charted. In histograms, you can use this menu to set bar width and offset. You can set tick values for the frequency axis by clicking on the **Frequency** label before selecting **Ticks** from the pop-up menu.
- Axes** toggles the display of axes.
- Observations** toggles the display of observations (bars). When this menu is toggled off, observations are displayed only if selected.
- Values** toggles the display of values for bar heights.
- Reference Lines** toggles the display of lines that indicate the position of major ticks on the frequency axis. This option is not available unless the axes are visible.

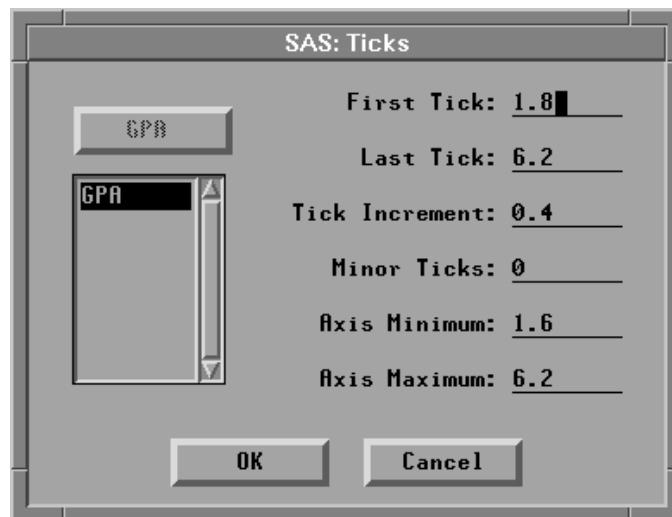


Figure 32.6. Bar Chart Ticks Dialog

You can adjust bar width and offset interactively with the Hand Tool, as described in [Chapter 12, “Examining Distributions.”](#) You can also add density curves to bar charts in distribution analyses, as described in [Chapter 38, “Distribution Analyses.”](#)

⊕ **Related Reading:** Bar Charts, [Chapter 4](#).

⊕ **Related Reading:** Distributions, [Chapter 12](#), [Chapter 38](#).

References

Terrell, G.R. and Scott, D.W. (1985), “Oversmoothed Nonparametric Density Estimates,” *Journal of the American Statistical Association*, 80 (389), 209–214.

Chapter 33

Box Plots and Mosaic Plots

Chapter Contents

VARIABLES	509
METHOD	511
OUTPUT	512
Multiple Comparison Options	514
Multiple Comparison Circles	516
REFERENCES	517

Chapter 33

Box Plots and Mosaic Plots

Box plots are pictorial representations of the distribution of values of a variable. The central line in each box marks the median value and the edges of the box mark the first and third quartiles.

The *median* value of a distribution is the 50th *percentile*: It is the value less than and greater than 50% of the data. The first and third *quartiles* are the 25th and 75th percentiles. By combining these three values in a schematic diagram and plotting individual markers for extreme data values, the box plot provides a concise display of a distribution (Tukey 1977).

Mosaic plots are pictorial representations of frequency counts of a single nominal variable or cross-classified nominal variables. Because mosaic plots display the frequencies graphically, they are easier to understand than crosstabulations. You can select and brush mosaic plots to explore dependencies between variables.

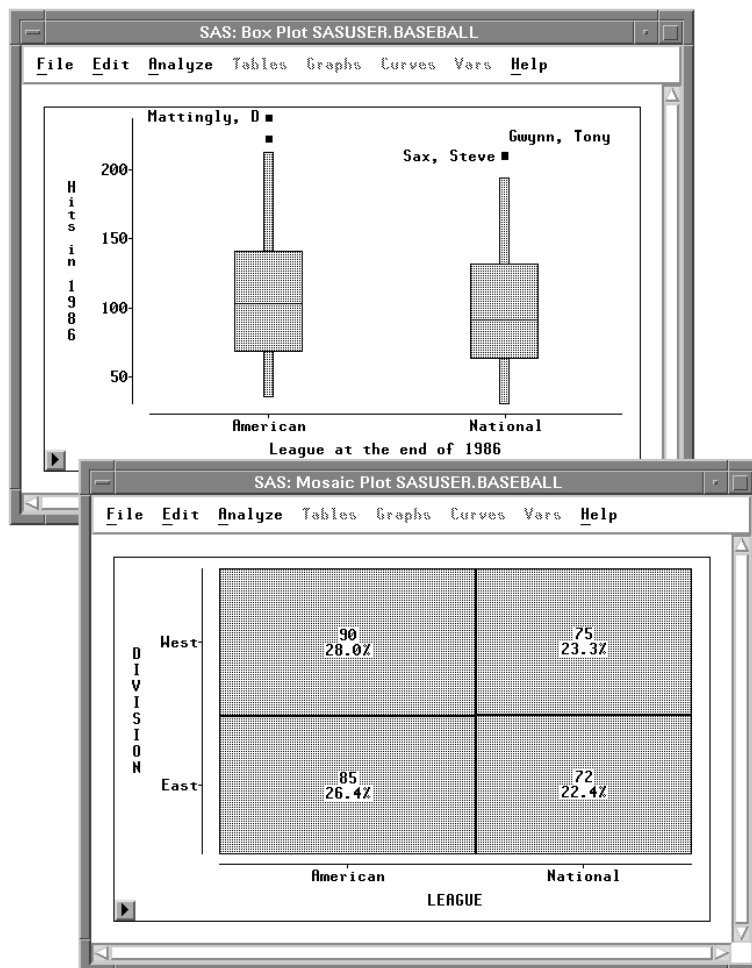


Figure 33.1. Box Plot and Mosaic Plot

Variables

To create a box plot or mosaic plot, choose **Analyze:Box Plot/Mosaic Plot (Y)**. If you have previously selected one or more variables, they are assigned the required **Y** variable role. A single plot is created containing a separate schematic diagram for each **Y** variable selected. For interval **Y** variables, box plots are created. For nominal **Y** variables, mosaic plots are created.

If you have not selected any variables, a variables dialog appears.



Figure 33.2. Box Plot/Mosaic Plot Variables Dialog

In the dialog, select at least one **Y** variable.

You can select one or more **X** variables to compare distributions. If you do not select **X** variables, you get one plot containing one schematic diagram for each **Y** variable. If you select **X** variables, you get one plot for each **Y** variable, and each plot contains one schematic diagram for each combination of **X** values. For example, [Figure 33.3](#) shows the box plot created using the **BASEBALL** data set with **NO_HITS** as the **Y** variable and **LEAGUE** as the **X** variable.

You can select one or more **Group** variables if you have grouped data. This creates a separate box or mosaic plot for each group. For example, [Figure 33.4](#) shows the box plots created using the **BASEBALL** data set with **NO_HITS** as the **Y** variable and **LEAGUE** as the **Group** variable.

You can select a **Label** variable to label extreme values in box plots.

If you select a **Freq** variable, each observation is assumed to represent n observations, where n is the value of the **Freq** variable.

You can identify extreme values in the box plot and display the *mean* or average value. You can also control the marker size of extreme values and the information shown in the box plot axes.

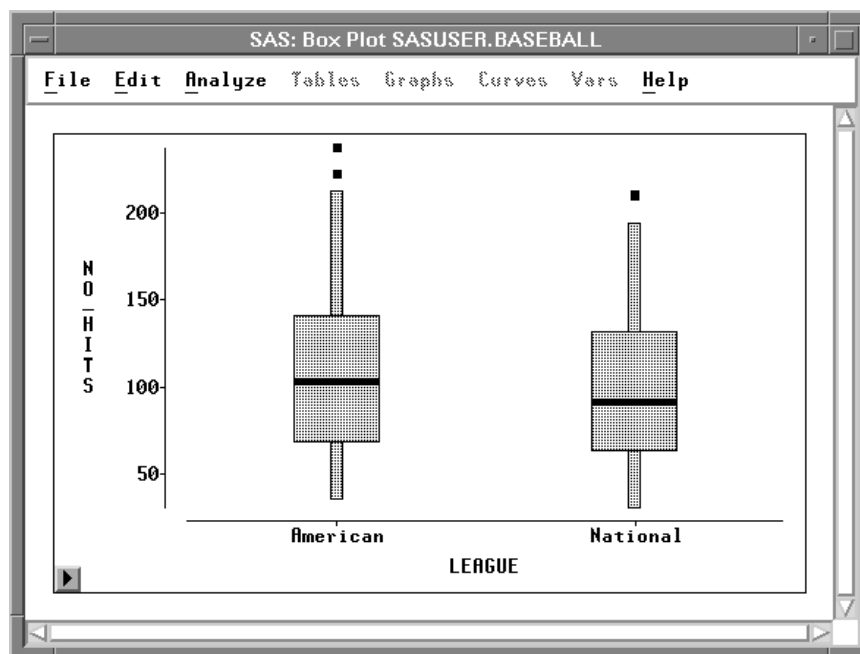


Figure 33.3. Box Plot Using **X** Variable

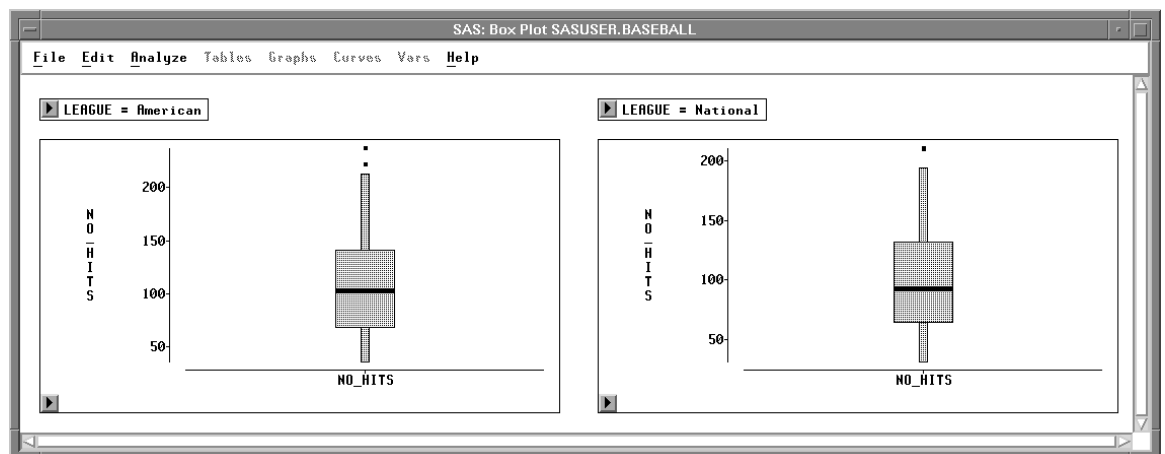


Figure 33.4. Box Plot Using **Group** Variable

Method

Observations with missing values for **Y** variables are not used. Observations with **Freq** values that are missing or that are less than or equal to 0 are not used. Only the integer part of **Freq** values is used.

The following method is used to compute the median and quartiles. Let

n be the number of data values

y_1, y_2, \dots, y_n be the data values listed in increasing order

p be the desired percentile (25, 50, or 75)

i be the integer part, and f the fractional part, of the ordinal of the desired percentile:

$$i + f = n * p / 100$$

Then the value of the desired percentile is

$$\begin{array}{ll} (y_i + y_{i+1})/2 & \text{if } f = 0 \\ y_{i+1} & \text{if } f > 0 \end{array}$$

You can adjust three calculation methods by clicking on the **Method** button in the variables dialog. This displays the method options dialog.



Figure 33.5. Box Plot/Mosaic Plot Method Options Dialog

By default, *whiskers* on the box plot are drawn from the quartiles to the farthest observation not farther than 1.5 times the distance between the quartiles. Type your preferred whisker length factor in the entry field. The figures in this chapter were created using whisker lengths that were 1.0 times the distance between the quartiles; this results in more observations being classified as outliers.

By default, for variables in mosaic plots, values that represent less than 4% of the total frequency are grouped together in an “**Other**” category. The Method dialog enables you to change the threshold at which values are grouped in the **Other** category.

By default, **X** variable values are sorted by their formatted value. Turn off the **Sort X Formatted** check box to sort **X** variable values by their unformatted value.

Output

To view or modify output options associated with your plot, click on the **Output** button of the variables dialog. This displays the output options dialog.

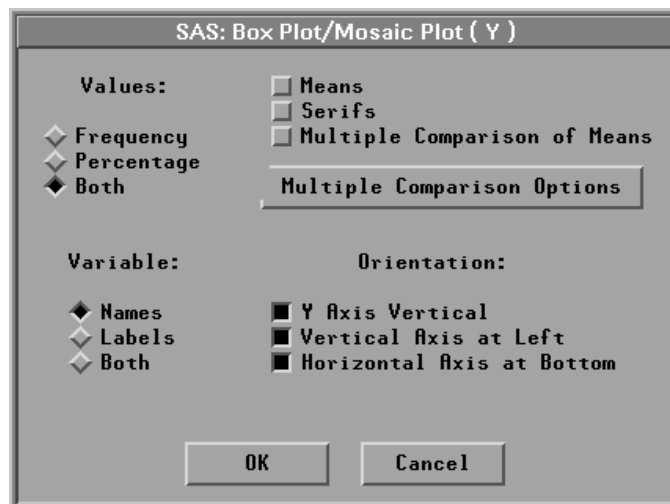


Figure 33.6. Box Plot/Mosaic Plot Output Options Dialog

Values:Frequency	labels mosaic boxes with the frequency of observations represented in each box.
Values:Percentage	labels mosaic boxes with the percentage of observations represented in each box.
Values:Both	labels mosaic boxes with both frequency and percentage.
Means	displays mean diamonds on box plots. The central line in the diamond marks the mean. The size of the diamond is two standard deviations, one on either side of the mean.
Serifs	displays serifs at the ends of box plot whiskers.
Multiple Comparison of Means	displays a <i>comparison circle</i> (Sall 1992) for each box. The center of each circle marks the mean of each box. The color and line style of each circle indicates how the mean value of one box compares with the means of other boxes. A selected circle is highlighted and is drawn in red on color monitors. Circles corresponding to categories whose mean values are significantly different from a selected group are drawn in cyan on color monitors. Circles corresponding to categories whose mean values are not different are drawn with a dashed line and are red on color monitors. See the section “ Multiple Comparison Circles ” later in this chapter.

Multiple Comparison Options	displays the Multiple Comparison Options dialog window.
Variable:Names	labels the axes with variable names.
Variable:Labels	labels the axes with variable labels.
Variable:Both	labels the axes with both names and labels.
Orientation: Y Axis Vertical	draws the axis for the Y variable vertically. If this option is off, the Y axis is horizontal.
Orientation: Vertical Axis at Left	places the vertical axis at the left side of the plot. If this option is off, the vertical axis is at the right side.
Orientation: Horizontal Axis at Bottom	places the horizontal axis at the bottom of the plot. If this option is off, the horizontal axis is at the top.

You can modify other aspects of box and mosaic plots with the pop-up menu.

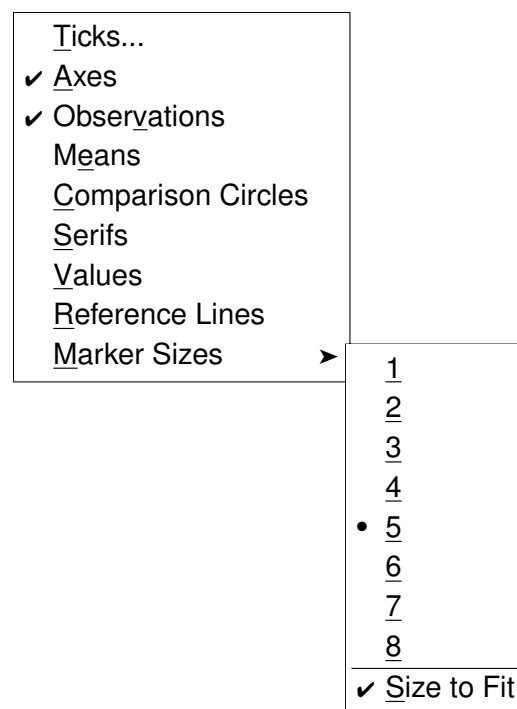


Figure 33.7. Box Plot/Mosaic Plot Pop-up Menu

Ticks...	specifies tick labels on the Y axis.
Axes	toggles the display of axes.
Observations	toggles the display of observations (boxes and extreme values). When this menu is toggled off, observations are displayed only if selected.

Means	toggles the display of mean diamonds in box plots.
Comparison Circles	toggles the display of comparison circles in box plots.
Serifs	toggles the display of serifs at the ends of box plot whiskers.
Values	toggles the display of values for means, medians, quartiles, and ends of whiskers in box plots. Toggles the display of frequency and percentage counts in mosaic plots.
Reference Lines	toggles the display of lines that indicate the position of major ticks on the Y axis. This option is not available unless the axes are visible.
Marker Sizes	sets the size of markers that display extreme values in box plots.

Multiple Comparison Options

Box plots enable you to examine means in different groups. Statistical questions you might have about the group means include

- Which underlying group means are likely to be different?
- Which group means are better than the mean of a standard group?
- Which group means are statistically indistinguishable from the best?

From the **Multiple Comparison Options** dialog, you can select a multiple comparison of means test and a confidence level for the test. Multiple comparison tests enable you to infer differences between means and also to construct simultaneous confidence intervals for these differences.

All of the tests implemented in SAS/INSIGHT software are constructed assuming that the displayed variables are independent and normally distributed with identical variance. For details, refer to Hsu (1996).

Each of the tests available in SAS/INSIGHT software is described below. In the descriptions that follow, k is the number of categories (that is, the number of boxes in the box plot), n_i is the number of observations for the i th category, μ_i is the true mean for the i th category, $\hat{\mu}_i$ is the sample mean for the i th category, $\nu = \sum_{i=1}^k (n_i - 1)$ is the total degrees of freedom, and $\hat{\sigma}$ is the root mean square error, also known as the pooled standard deviation. Each test creates a table showing $100(1 - \alpha)\%$ confidence intervals for the difference $\hat{\mu}_i - \hat{\mu}_j$, $i \neq j$, $i = 1 \dots k$.

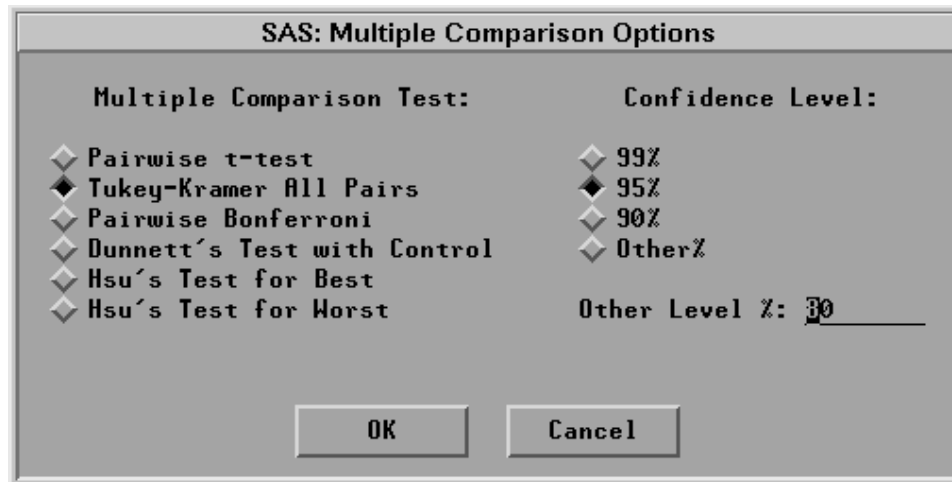


Figure 33.8. Multiple Comparison Options

The **Pairwise t -test** is not a true simultaneous comparison test, but rather uses a pairwise t test to provide confidence intervals about the difference between two means. These intervals have a half-width equal to $t_{\alpha/2, \nu} \hat{\sigma} \sqrt{n_i^{-1} + n_j^{-1}}$. Although each confidence interval was computed at the $100(1 - \alpha)\%$ level, the probability that all of your confidence intervals are correct *simultaneously* is less than $100(1 - \alpha)\%$. The actual simultaneous confidence for the t -based intervals is approximately $100(1 - k\alpha)\%$. For example, for five groups the actual simultaneous confidence for the t -based intervals is approximately only 75%.

The **Tukey-Kramer** method is a true “multiple comparison” test, appropriate when all pairwise comparisons are of interest; it is the default test used. The test is an exact α -level test if the sample sizes are the same, and it is slightly conservative for unequal sample sizes. The confidence interval around the point-estimate $\hat{\mu}_i - \hat{\mu}_j$ has half-width $q^* \hat{\sigma} \sqrt{n_i^{-1} + n_j^{-1}}$. It is a common convention to report the quantity $\sqrt{2}q^*$ as the Tukey-Kramer quantile, rather than just q^* .

The **Pairwise Bonferroni** method is also appropriate when all pairwise comparisons are of interest. It is conservative; that is, Bonferroni tests performed at a nominal significance level of α actually have a somewhat greater level of significance. The Bonferroni method uses the t distribution, like the pairwise t test, but returns smaller intervals with half-width $t_{\alpha/(k(k-1)), \nu} \hat{\sigma} \sqrt{n_i^{-1} + n_j^{-1}}$. Note that the t probability ($\alpha/2$, since this is a two-sided test) is divided by the total number of pairwise comparisons ($k(k-1)/2$). The Bonferroni test produces wider confidence intervals than the Tukey-Kramer test.

Dunnett's Test with Control is a two-sided multiple comparison method used to compare a set of categories to a control group. The quantile that scales the confidence interval is usually denoted $|d|$. If the i th confidence interval does not include zero, you may infer that the i th group is significantly different from the control. A control group may be a placebo or null treatment, or it may be a standard treatment. While

the interactive nature of SAS/INSIGHT enables you to select any category to use as the basis of comparison in Dunnett's test, you should select a category only if it truly is the control group. To select a category, click on the corresponding comparison circle.

Hsu's Test for Best can be used to screen out group means that are statistically less than the (unknown) largest true mean. It forms *nonsymmetric* confidence intervals around the difference between the largest sample mean and each of the others. If an interval does not properly contain zero in its interior, then you may infer that the associated group is not among the best.

Similarly, **Hsu's Test for Worst** can be used to screen out group means that are statistically greater than the (unknown) smallest true mean. If an interval does not properly contain zero in its interior, then you may infer that the true mean of that group is not equal to the (unknown) smallest true mean.

Multiple Comparison Circles

In addition to a table that summarizes the statistics for simultaneous multiple comparison of means, SAS/INSIGHT software provides a graphical technique to help visualize which groups are significantly different from a selected group. Each test is accompanied by a *comparison circles* plot that graphically illustrates the comparisons (Sall 1992).

There is a circle next to the box plot and centered at each category's sample mean. The radius of the i th circle is $q\hat{\sigma}/\sqrt{n_i}$, where q is a quantile used to scale the circles according to the test being used. For details on how each quantile is computed, see refer to Hsu (1996).

If the j th group is selected (by clicking on its circle), then its circle is highlighted. This circle is red on color monitors. You can determine whether another group is significantly different than the selected group based on how much their corresponding circles overlap. If their circles are nested or nearly overlap so that the external angle of intersection is greater than 90 degrees, then you cannot claim that the means of the two groups are different. If, however, the two circles are disjoint or just barely overlap so that their external angle of intersection is less than 90 degrees, then you can conclude that the means of the two groups are significantly different at the given confidence level.

Circles corresponding to categories that are significantly different from the selected group are drawn in cyan on color monitors. Circles corresponding to categories that are not different are drawn with a dashed line and are red on color monitors.

The geometry behind comparison circles is based on the Pythagorean Theorem: since the radius of the i th circle is $r_i = q\hat{\sigma}/\sqrt{n_i}$, and since the circle is centered at $\hat{\mu}_i$, then if the two circles meet at right angles, the distance between centers is the hypotenuse of the right triangle formed by the circles' radii. Therefore, when the circles meet at right angles, $|\hat{\mu}_i - \hat{\mu}_j| = q\hat{\sigma}\sqrt{n_i^{-1} + n_j^{-1}}$. Statistically, this geometry corresponds to the critical case in which zero happens to fall on the boundary of the confidence interval about $\hat{\mu}_i - \hat{\mu}_j$. If $|\hat{\mu}_i - \hat{\mu}_j| > q\hat{\sigma}\sqrt{n_i^{-1} + n_j^{-1}}$, then the external intersection

of the circles is less than 90 degrees, and zero is not contained in the confidence interval about $\hat{\mu}_i - \hat{\mu}_j$. Thus the circles are significantly different.

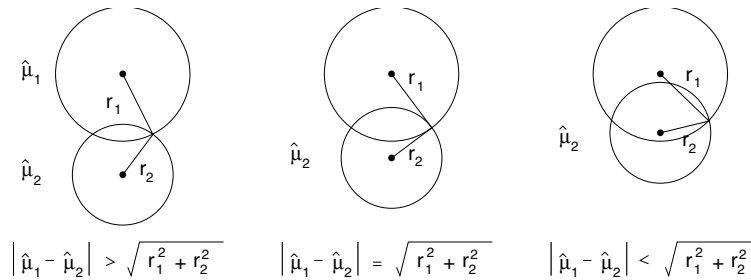


Figure 33.9. The Geometry of Multiple Comparison Circles

The statistics for Hsu's Test for Best and Hsu's Test for Worst are computed differently from the other tests. First, the comparison circles are not selectable. The Test for Best automatically selects the category with the largest sample mean; the Test for Worst selects the category with the smallest sample mean. Second, the quantile used to scale the comparison circles is the maximum of the quantiles computed by running Dunnett's one-sided test $k - 1$ times, with each "non-best" (or "non-worst") group serving in turn as the "control" for Dunnett's test.

Because Hsu's Test for Best does not provide symmetric intervals about $\hat{\mu}_i - \hat{\mu}_j$, the comparison circle technique must be modified. While the statistical table reports exactly which groups can be inferred not to be the best, the comparison circles are more conservative because the quantile used to scale the circle radii is the maximum of all quantiles encountered during Hsu's test. The same is true for Hsu's Test for Worst.

⊕ **Related Reading:** Box Plots, [Chapter 4](#).

⊕ **Related Reading:** Mosaic Plots, [Chapter 5](#).

⊕ **Related Reading:** Distributions, [Chapter 12](#).

References

- Hartigan, J.A. and Kleiner, B. (1984), "A Mosaic of Television Ratings," *The American Statistician*, 38, 32–35.
- Hsu, J.C. (1996), *Multiple Comparisons: Theory and Methods*, London: Chapman & Hall.
- Sall, J. (1992), "Graphical Comparison of Means," *Statistical Computing and Statistical Graphics Newsletter*, 3, 27–32.
- Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Chapter 34

Line Plots

Chapter Contents

VARIABLES	522
METHOD	522
OUTPUT	523

Chapter 34

Line Plots

You can create *line plots* to show the path of a variable over time. You can control the orientation of the plot, the information shown on the axes, and the color of the lines.

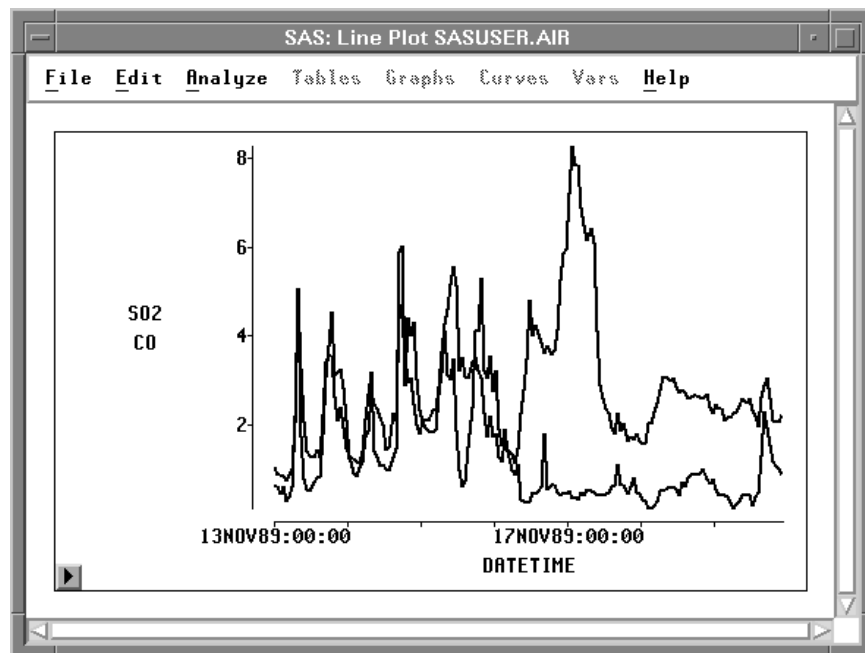


Figure 34.1. Line Plot

Variables

To assign variables for a line plot, choose **Analyze:Line Plot (Y X)**. If you have already selected two or more variables, you obtain a line plot. The last variable you selected is assigned the **X** role, and all other variables are assigned the **Y** role.

If you have not selected any variables, a variables dialog appears.

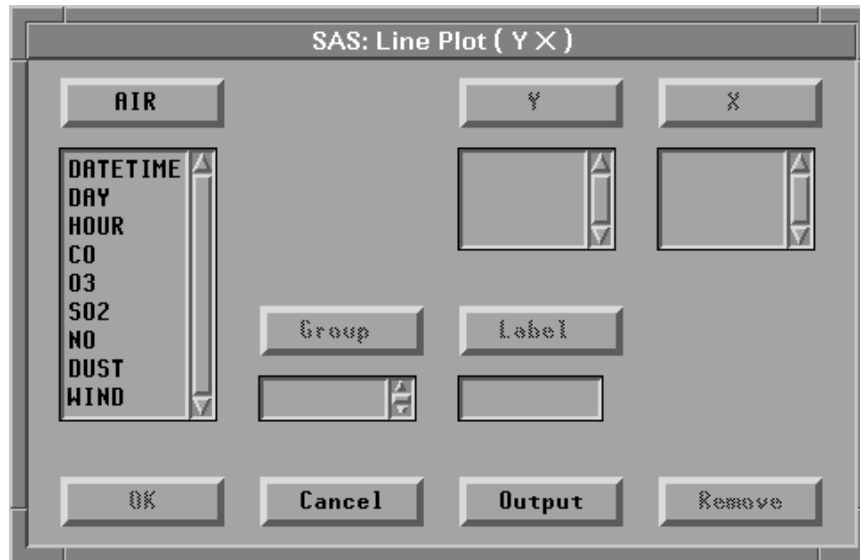


Figure 34.2. Line Plot Variables Dialog

In the dialog, select at least one **Y** variable and at least one **X** variable. You will obtain one line plot for each **X** variable, while multiple **Y** variables are represented on each plot as multiple lines.

You can select one or more **Group** variables if you have grouped data. This creates line plots for each group.

You can select a **Label** variable to label observations in the plots.

Method

Observations with missing values for **X** variables are not used. Observations with missing values for a **Y** variable are not used in the line for that **Y** variable but are used in lines for other **Y** variables.

Output

To view or modify output options associated with your line plot, click on the **Output** button of the variables dialog. This displays the options dialog.

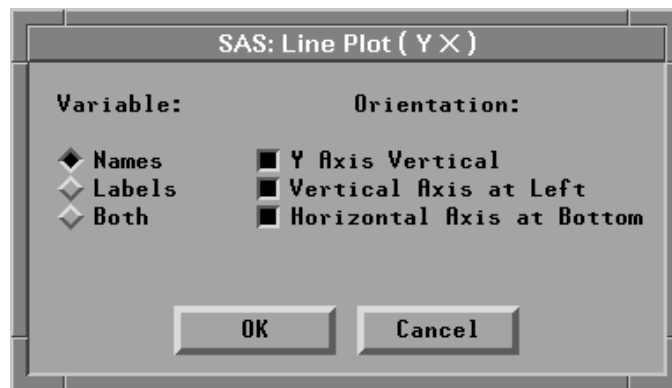


Figure 34.3. Line Plot Output Options Dialog

Variable:Names	labels the axes with variable names.
Variable:Labels	labels the axes with variable labels.
Variable:Both	labels the axes with both names and labels.
Orientation: Y Axis Vertical	draws the axis for the Y variable vertically. If this option is turned off, the Y axis is horizontal.
Orientation: Vertical Axis at Left	places the vertical axis at the left side of the plot. If this option is turned off, the vertical axis appears at the right side of the plot.
Orientation: Horizontal Axis at Bottom	places the horizontal axis at the bottom of the plot. If this option is turned off, the horizontal axis appears at the top of the plot.

You can modify other aspects of the line plot by using the pop-up menu.

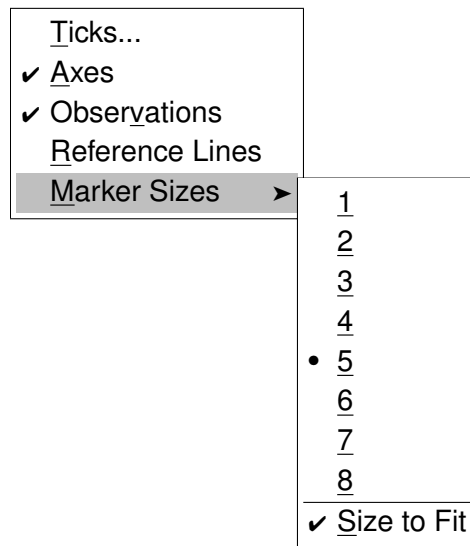


Figure 34.4. Line Plot Pop-up Menu

Ticks...	specifies tick labels on either axis.
Axes	toggles the display of axes.
Observations	toggles the display of observations. When this menu is toggled off, observations are displayed only if selected.
Reference Lines	toggles the display of lines that indicate the position of major ticks on the axes. This option is not available unless the axes are visible.
Marker Sizes	sets the size of markers used to display observations.

You can select and brush observations in the line plot even when they are not visible. If you click on a line at the location of an observation, you select that observation. If you click on a line between two observations, you select the line.

Lines in the plot are linked to variables on the Y axis. Click either on the line or on a Y variable to select both the line and its associated variable.

Finally, you can set colors, patterns, and widths of lines the same way you set these attributes for curves. See [Chapter 13, “Fitting Curves,”](#) for examples of setting patterns, widths, and colors.

Chapter 35

Scatter Plots

Chapter Contents

VARIABLES	528
METHOD	528
OUTPUT	529

Chapter 35

Scatter Plots

A *scatter plot* is a graphic representation of the relationship between two variables.

You can identify and label observations in the scatter plot, control the orientation of the plot, and control the information shown on the axes. You can explore multivariate data in a scatter plot matrix.

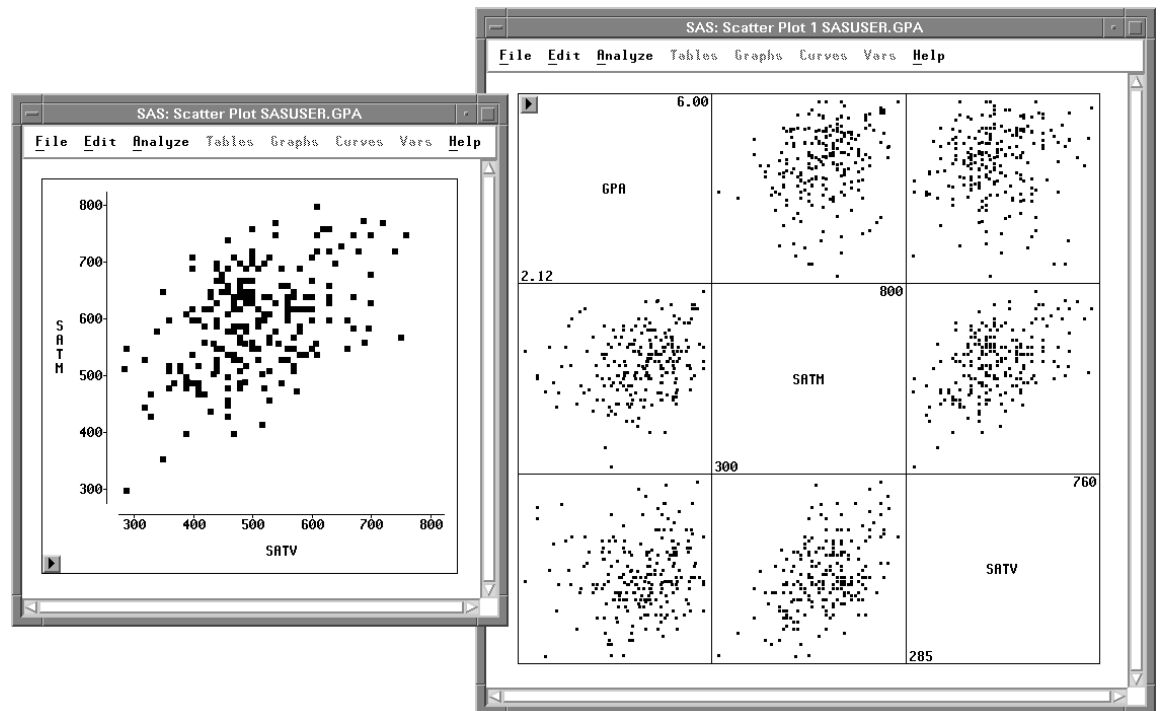


Figure 35.1. Scatter Plot and Scatter Plot Matrix

Variables

To create a scatter plot, choose **Analyze:Scatter Plot (Y X)**. If you have already selected two or more variables, you obtain a *scatter plot matrix*. A scatter plot matrix consists of all pairwise scatter plots of the selected variables. If you assign **Y** and **X** roles to the same set of variables, variable names and minimum and maximum values appear in the diagonal panels.

If you have not selected any variables, a variables dialog appears.

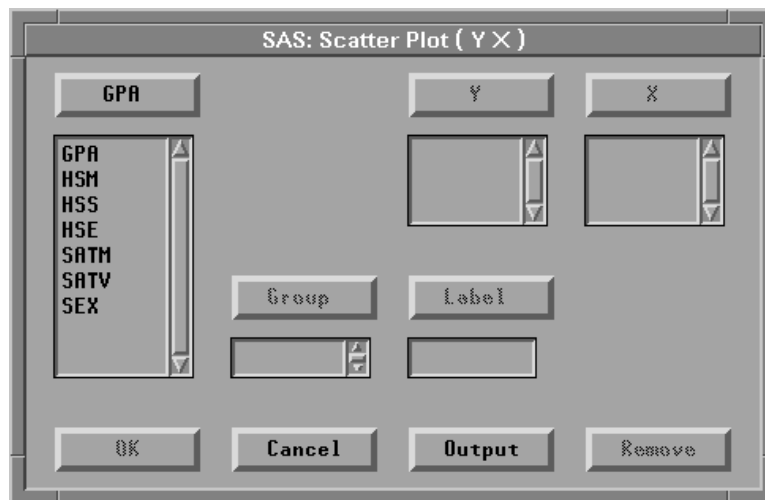


Figure 35.2. Scatter Plot Variables Dialog

In the dialog, select at least one **Y** variable and at least one **X** variable.

You can select one or more **Group** variables if you have grouped data. This creates scatter plots for each group.

You can select a **Label** variable to label observations in the plots.

Method

Observations with missing values for **Y** or **X** variables are not used.

Output

To view or modify output options associated with your scatter plot, click on the **Output** button of the variables dialog. This displays the options dialog shown in Figure 35.3.

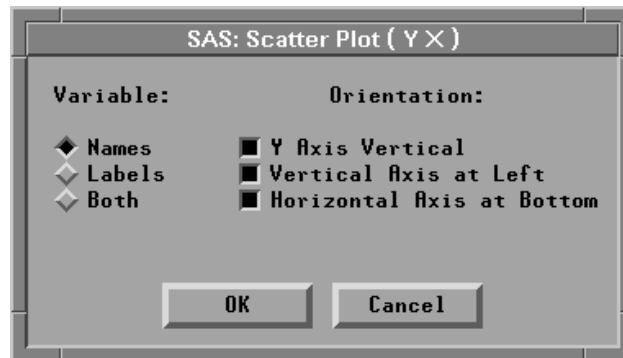


Figure 35.3. Scatter Plot Output Options Dialog

Variable:Names	labels the axes with variable names.
Variable:Labels	labels the axes with variable labels.
Variable:Both	labels the axes with both names and labels.
Orientation: Y Axis Vertical	draws the axis for the Y variable vertically. If this option is turned off, the Y axis is horizontal.
Orientation: Vertical Axis at Left	places the vertical axis at the left side of the plot. If this option is turned off, the vertical axis is at the right side of the plot.
Orientation: Horizontal Axis at Bottom	places the horizontal axis at the bottom of the plot. If this option is turned off, the horizontal axis is at the top of the plot.

You can modify other aspects of a scatter plot or scatter plot matrix using the pop-up menu. For scatter plots, the pop-up menu has the following choices.

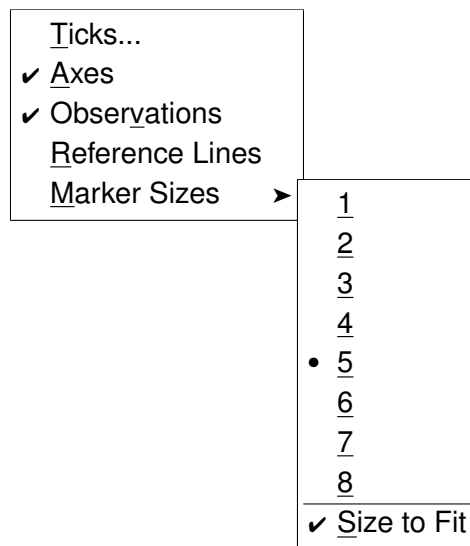


Figure 35.4. Scatter Plot Pop-up Menu

Ticks...	specifies tick labels on either axis.
Axes	toggles the display of axes.
Observations	toggles the display of observations. When this menu is toggled off, observations are displayed only if selected.
Reference Lines	toggles the display of lines that indicate the position of major ticks on the axes. This option is not available unless the axes are visible.
Marker Sizes	sets the size of markers used to display observations.

When **Marker Sizes:Size to Fit** is checked, marker sizes are chosen to fit the graph.

You can manipulate square scatter plot matrices as a unit. For example, you can resize the entire matrix by dragging a corner. Pop-up menus act on all plots in the matrix.

If you have created a brush, an additional pop-up menu is available, as shown in [Figure 35.5](#). (See [Chapter 5, “Exploring Data in Two Dimensions,”](#) for more information on brushing.)

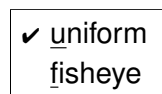


Figure 35.5. Scatter Plot Lens Pop-up Menu

uniform	specifies that observations beneath the brush are seen as if the brush were a typical camera lens. The relative positions of brushed observations are not distorted by the presence of the brush.
----------------	---

fisheye

specifies that observations beneath the brush are seen as if the brush were a fisheye camera lens. The relative positions of brushed observations are transformed so that observations near the center of the brush are magnified, whereas observations away from the center appear small. The fisheye lens may be useful for discerning individual observations within densely clustered data.

- ⊕ **Related Reading:** Scatter Plots, [Chapter 5](#).
- ⊕ **Related Reading:** Fitting Curves, [Chapter 13](#).
- ⊕ **Related Reading:** Confidence Ellipses, [Chapter 18](#).

Chapter 36

Contour Plot

Chapter Contents

VARIABLES	537
METHOD	538
OUTPUT	539

Chapter 36

Contour Plot

A *contour plot* is a graphic representation of the relationships among three numeric variables in two dimensions. Two variables are for X and Y axes, and a third variable Z is for contour levels. The contour levels are plotted as curves; the area between curves can be color coded to indicate interpolated values.

You can interactively identify, label, color, and move contour levels, and change the resolutions of rectangular grids to get better contouring quality and performance. You can choose linear interpolation or thin-plate smoothing spline to fit contour surface functions.

You can also toggle, identify and label observations in the contour plot, control the orientation of the plot, and control the information shown on the axes.

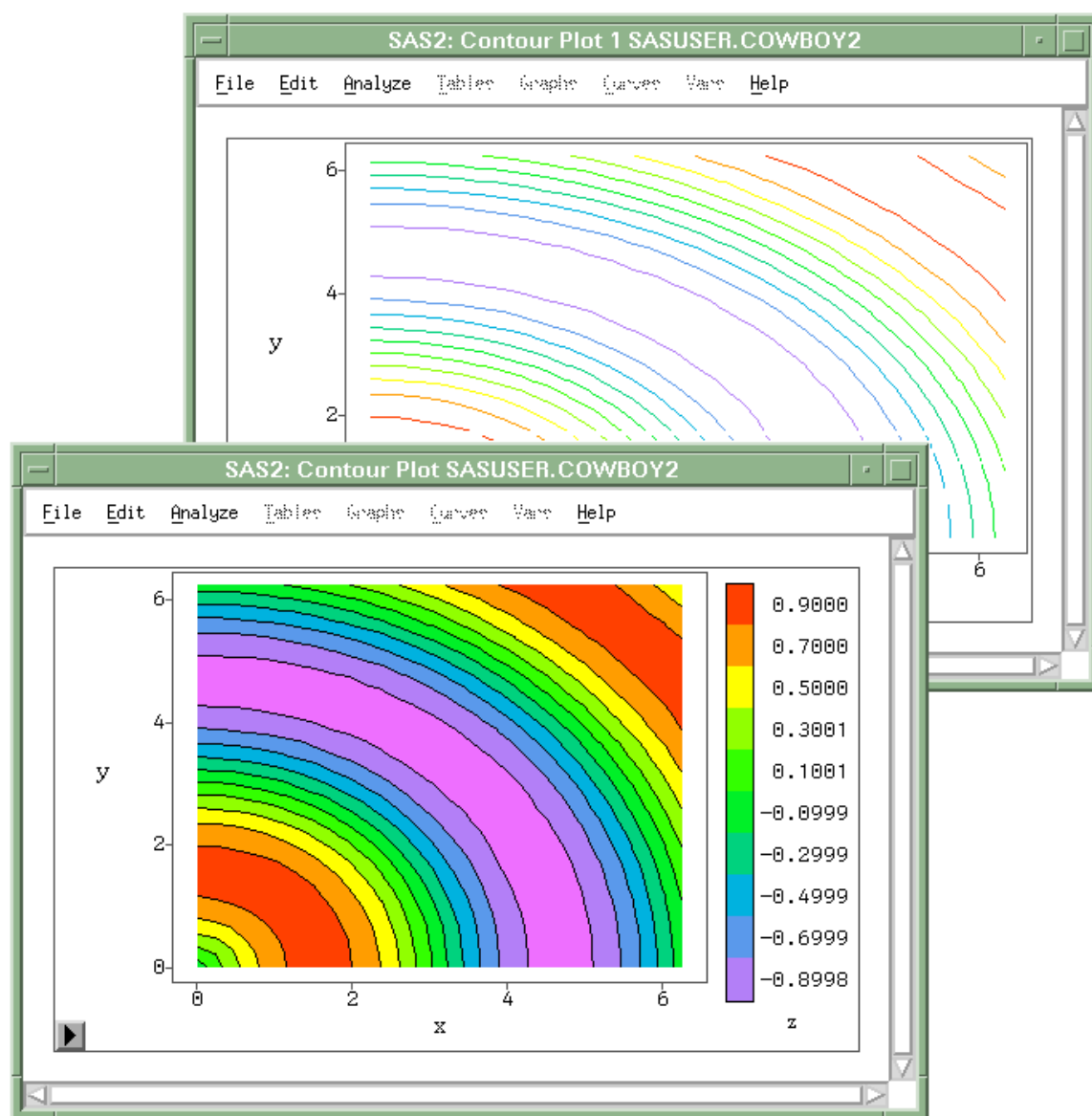


Figure 36.1. Contour Plot

Variables

To create a contour plot, choose **Analyze:Contour Plot (Z Y X)**. If you have already selected three or more numeric variables, a contour plot for each unique triplet of variables appears. If you have not selected any variables, a variables dialog appears.

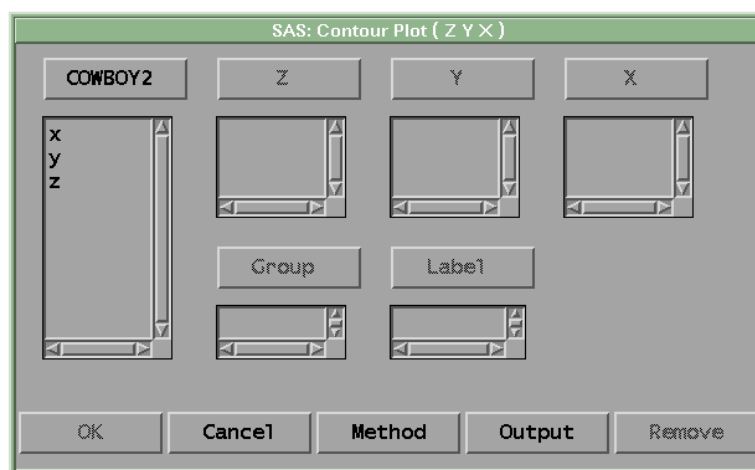


Figure 36.2. Contour Plot Variables Dialog

In the dialog, select at least one **Z**, **Y**, **X** variable. If you select more than three variables, you obtain a matrix of contour plots. If the **X** variable and **Y** variable are the same, you get a plot without contours.

You can select one or more **Group** variables if you have grouped data. This creates contour plots for each group.

You can select a **Label** variable for labeling observations in the plots.

Method

Observations that have missing values for any of the **Z**, **Y**, **X** variables are not used.

If two or more observations have the same (x, y) values, their mean Z value is used as the Z value at point (x, y) .

Clicking on the **Method** button in the variable dialog displays the dialog in [Figure 36.3](#).

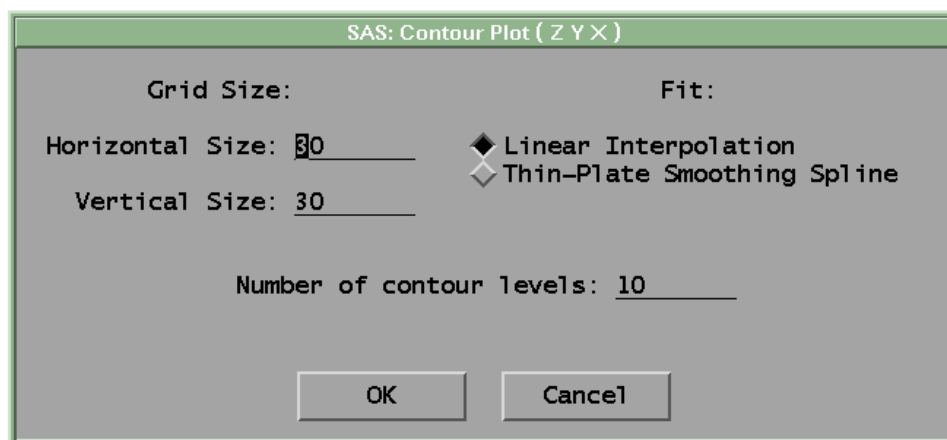


Figure 36.3. Contour Plot Method Dialog

Grid Size: Horizontal Size	specifies the horizontal resolution of the rectangular grid over which contour function is evaluated.
Grid Size: Vertical Size	specifies the vertical resolution of the rectangular grid over which contour function is evaluated.
Fit:Linear Interpolation	linearly interpolates contour function across rectangular grid cells.
Fit:Thin-Plate Smoothing Spline	fits contour function over rectangular grid using thin-plate smoothing spline fitting. The process may be much slower than linear interpolation. It usually produces very smooth contours. See “Smoothing Spline Surface Plot” in Chapter 39 , “Fit Analyses,” for more information on thin-plate splines.
Number of Contour Levels	specifies the number of contour levels to be drawn in a contour plot. The contour levels are initially spaced evenly within the range of the Z variable.

Output

To view or modify output options associated with your contour plot, click on the **Output** button of the rotating plot variables dialog. This displays the options dialog in [Figure 36.4](#).

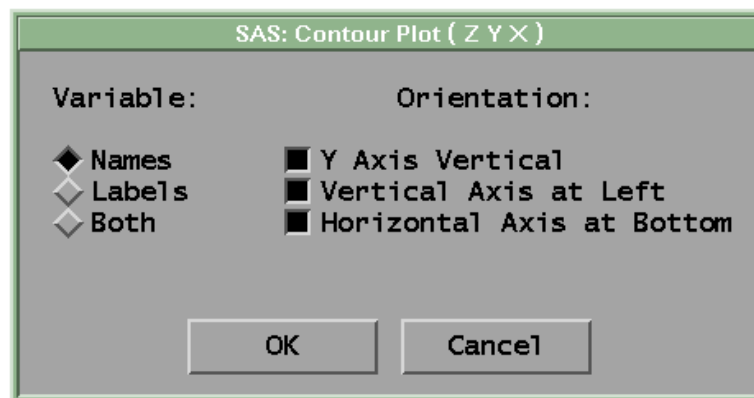


Figure 36.4. Contour Plot Output Options Dialog

Variable:Names	labels the axes with variable names.
Variable:Labels	labels the axes with variable labels.
Variable:Both	labels the axes with both names and labels.
Orientation: Y Axis Vertical	draws the axis for the Y variable vertically. If this option is turned off, the Y axis is horizontal.
Orientation: Vertical Axis at Left	places the vertical axis at the left side of the plot. If this option is turned off, the vertical axis appears at the right side of the plot.
Orientation: Horizontal Axis at Bottom	places the horizontal axis at the bottom of the plot. If this option is turned off, the horizontal axis appears at the top of the plot.

You can modify other aspects of a contour plot by using the pop-up menu.

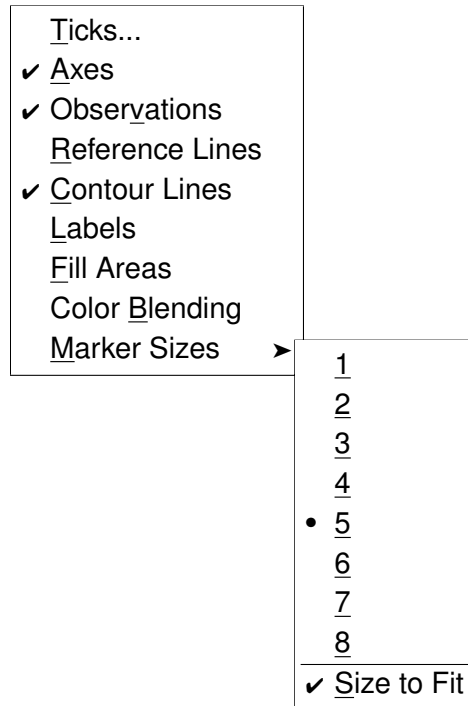


Figure 36.5. Contour Plot Pop-up Menu

Ticks...	specifies tick labels on either axis.
Axes	toggles the display of axes.
Observations	toggles the display of observations. When this menu is toggled off, observations are displayed only if selected.
Reference Lines	toggles the display of lines that indicate the position of major ticks on the axes. This option is not available unless the axes are visible.
Contour Lines	toggles the display of contours (level curves).
Labels	toggles the display of contour level labels.
Fill Areas	toggles the display of filled contour areas. When this menu item is toggled on, an area between two adjacent contour levels is filled in with the color of the lower level.
Color Blending	applies color blending to all contour levels. The color blend in the tools window is used.
Marker Sizes	sets the size of markers used to display observations.

You can select and brush observations in the contour plot even when they are not visible. If you click on a curve at the location of an observation, you select that observation. If you click on a contour curve between two observations, you select the curve.

You can use the hand tool to add contour curves at new locations. To add a new level curve, click at some (x, y) position; the level curve that passes through that location is computed and displayed. To move a contour level, drag on the level curve, then release the mouse at a new location (x', y') . Mathematically, this process results in seeing the level set that passes through (x', y') .

Finally, you can set colors, patterns, and widths of contour lines the same way you set these attributes for curves. See [Chapter 13, “Fitting Curves,”](#) for examples of setting patterns, widths, and colors. See also [Chapter 11, “Coloring Observations,”](#) for instructions on color blending.

Chapter 37

Rotating Plot

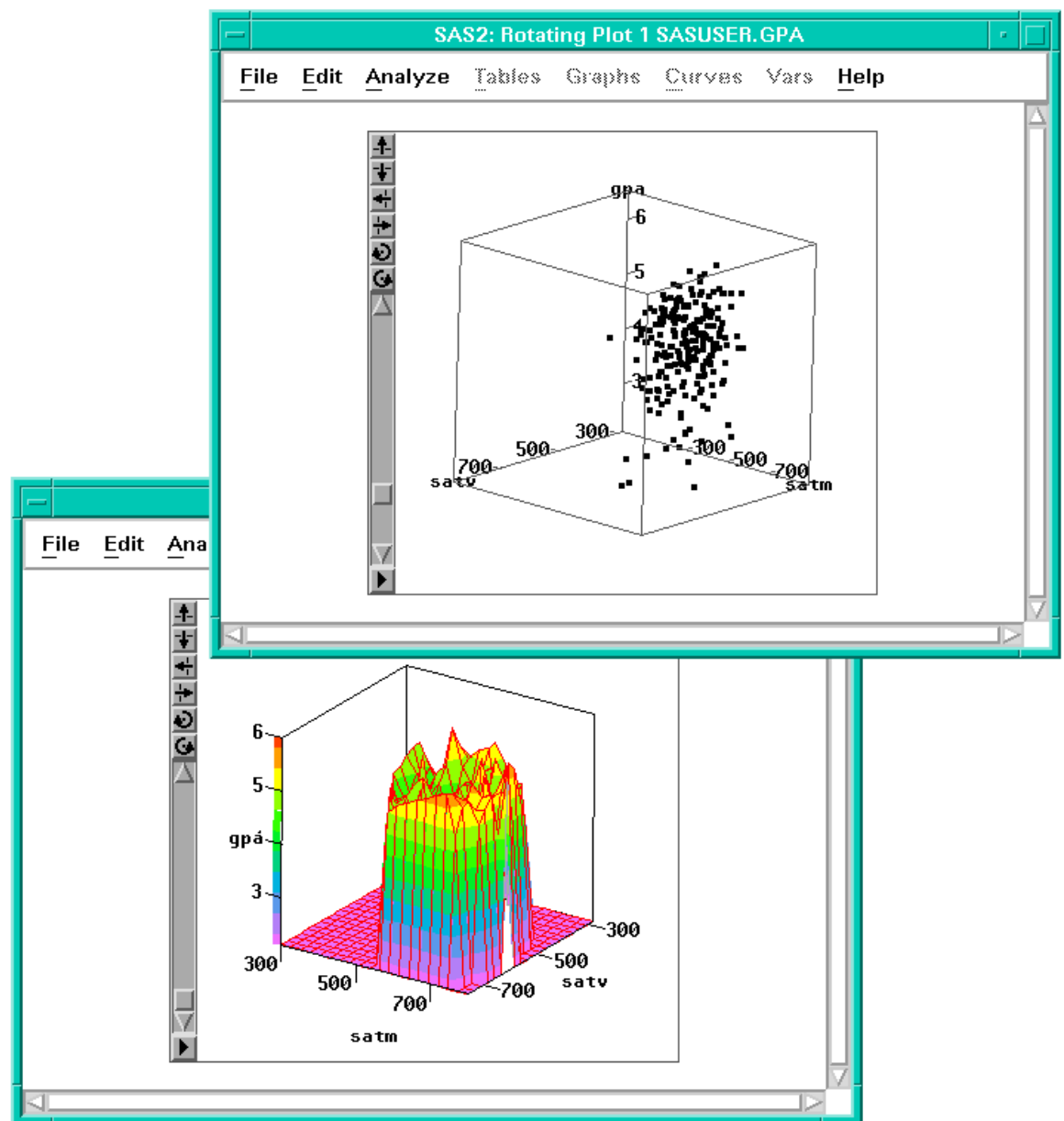
Chapter Contents

VARIABLES	547
METHOD	548
OUTPUT	549
REFERENCES	552

Chapter 37

Rotating Plot

A *rotating plot* is a graphic representation of the relationships among three variables. Rotating plots enable you to see structure in the data that is not apparent in two-dimensional scatter plots. Surface characteristics and general dependencies of one variable on the other two variables can be brought out by the three-dimensional representation (Becker, Cleveland, and Weil 1989).



Reference ♦ *Rotating Plot*

Figure 37.1. Rotating Plot

A surface plot is a rotating plot with a fit surface. It is a graphic representation of the relationships among three or four variables. A fourth variable can be used to color surface contours along **Z** direction in three-dimensional space. You can use linear interpolation or a thin-plate smoothing spline to fit surface functions.

Various drawing modes are provided to view a surface. For example, you can interactively color contour levels, and you can control the resolution of the rectangular grid used to compute a fitted surface.

You can toggle the display of axes and rays in any rotating plot. You can add a bounding cube to the display to show the range of the data and to provide perspective to the axes. You can adjust parameters that control depth cueing, the use of color, and the algorithm used for rotation.

Variables

To create a rotating plot, choose **Analyze:Rotating Plot (Z Y X)**. If you have already selected three or more variables, a rotating plot for each unique triplet of variables appears. If you have not selected any variables, a variables dialog appears.

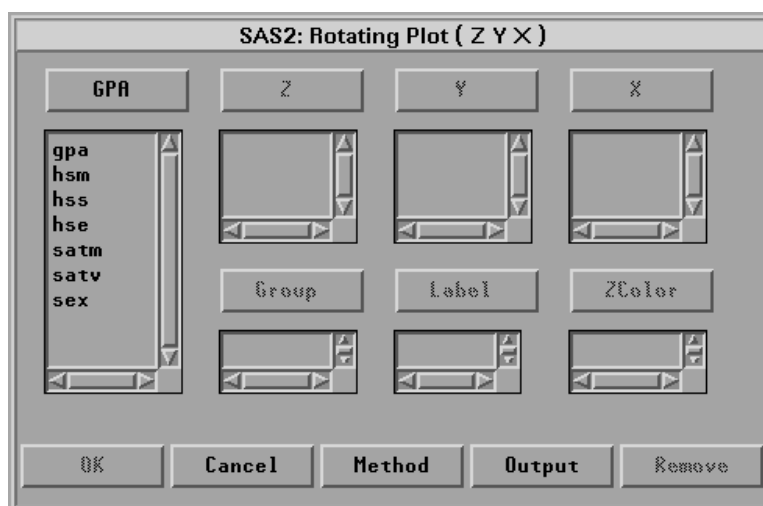


Figure 37.2. Rotating Plot Variables Dialog

In the dialog, select at least one **Z**, **Y**, **X** variable. If you select more than three variables, you obtain a matrix of rotating plots.

You can select one or more **Group** variables if you have grouped data. This creates rotating plots for each group.

You can select a **Label** variable for labeling observations in the plots.

To create a surface plot, select the **Fit Surface** option in the **Output** dialog as shown in [Figure 37.3](#). If the **X** variable and **Y** variable are the same, you get a rotating plot without surface.

You can select one or more **ZColor** variables to color surfaces. This creates surface plots for each color variable. The hues in the multiple colors button in the tools window are applied to the surface, according to interpolated values of the **ZColor** variable.

Method

Observations with missing values for **Z**, **Y**, **X** variables are not used.

If there are observations that all share the same values for the X and Y variables, then the mean Z value of the set is used for the purpose of fitting a surface to the data set.

Clicking on the **Method** button in the variables dialog displays the dialog in [Figure 37.3](#).

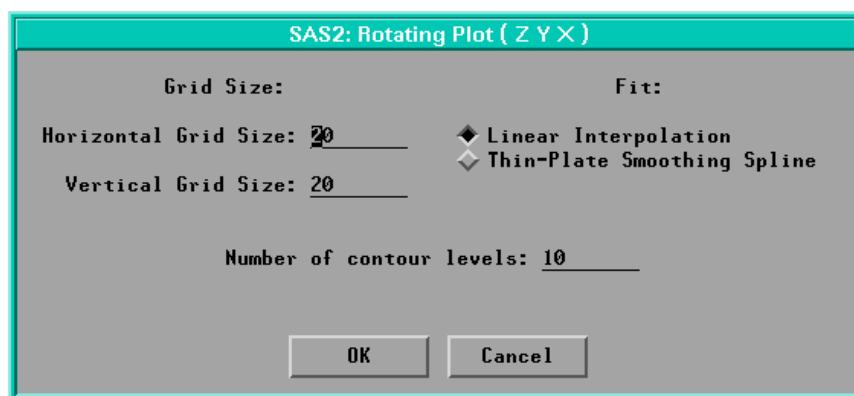


Figure 37.3. Rotating Plot Methods Dialog

Grid Size: Horizontal Size	specifies the horizontal resolution of the rectangular grid over which surface and ZColor functions are evaluated.
Grid Size: Vertical Size	specifies the vertical resolution of the rectangular grid over which surface and ZColor functions are evaluated.
Fit:Linear Interpolation	linearly interpolates surface and ZColor functions across rectangular grid cells.
Fit:Thin-Plate Smoothing Spline	fits surface and ZColor functions over the rectangular grid using thin-plate smoothing spline fitting. The process may be much slower than linear interpolation. It usually produces very smooth surfaces and colors.
Number of Contour Levels	specifies the number of contour levels to be drawn on the surface. The contour levels are spaced evenly within the range of the ZColor variable, or the range of the Z variable if no ZColor variable is specified.

Output

To view or modify output options associated with your rotating plot, click on the **Output** button of the rotating plot variables dialog. This displays the options dialog in [Figure 37.4](#).



Figure 37.4. Rotating Plot Output Options Dialog

Rays	draws a line segment from the center of the plot to each observation. These segments may help show the structure of the data.
Cube	displays a perspective cube around the observations to show the range of the data.
Depth	displays observations in two sizes (larger for near observations and smaller for distant observations) to aid three-dimensional visualization. If the marker size is 1 while Depth is in effect, only near observations are displayed.
Variable:Names	labels the axes with variable names.
Variable:Labels	labels the axes with variable labels.
Variable:Both	labels the axes with both names and labels.
Axes:At Midpoints	positions axes at the midpoints of the data, with no ticks. This is the best position for exploratory data analysis, as it minimizes interference of the axes with the point cloud.
Axes:At Minima	positions axes at the minima of the data, with ticks. This is the best position for viewing spatial or volumetric data.
Axes:Off	removes axes from the rotating plot.

Fit Surface fits a surface in the rotating plot.

You can modify other aspects of the rotating plot by using the rotating plot pop-up menu. Click the menu button at the lower left corner of the plot to display the pop-up menu.

The pop-up menu for a rotating plot without surface is shown in [Figure 37.5](#).

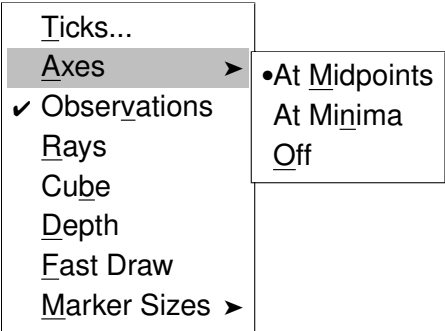


Figure 37.5. Rotating Plot without Surface Pop-up Menu

Ticks...	specifies tick labels on any axis.
Axes, Rays, Cube, Depth	set the display of axes, observation vectors, perspective cube, and depth cueing as described in the previous section on output options.
Observations	toggles the display of observations. When this menu item is toggled off, observations are displayed only if selected.
Fast Draw	toggles the use of drawing algorithms that may be faster, depending on your host. The effect of these algorithms also depends on the size of your data set. On some hosts, this menu improves rotation speed for large data sets.
Marker Sizes	sets the size of markers used to display observations.

The pop-up menu of a rotating plot with a fitted surface is shown in [Figure 37.6](#).

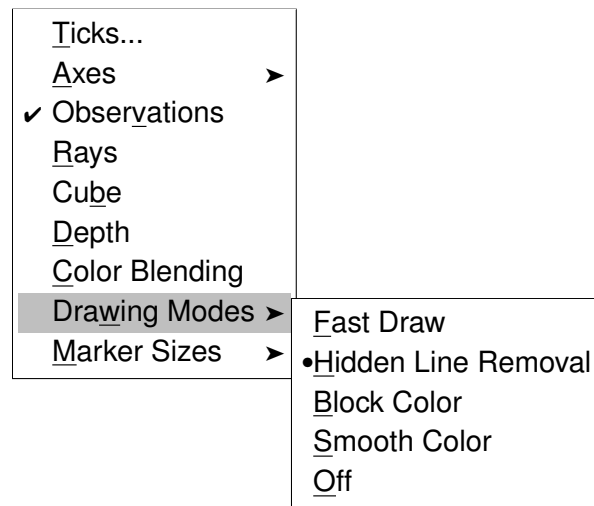


Figure 37.6. Rotating Plot with Surface Pop-up Menu

In addition to the menu items shown in [Figure 37.5](#), the following items are specific for the surface plot.

**Axes:Three
Sections**

positions axes, with ticks, on the edges of a bounding cube surrounding the data and fitted surface. The axes are placed so that the tick labels minimally interfere with viewing the data.

Color Blending

applies color blending to all contour levels. The color blends in the tools window are used. The surface is colored when the **Block Color** or **Smooth Color** display modes are on.

**Drawing Modes:
Fast Draw**

toggles the use of drawing algorithms that may be faster, depending on your host. The effect of these algorithms also depends on the size of your data set. On some hosts, this menu improves rotation speed for large data sets.

**Drawing Modes:
Hidden Line
Removal**

draws the surface in wireframe with hidden line removal. The front and back faces are in two different colors.

**Drawing Modes:
Block Color**

fills each surface grid cell with a color block by using color interpolation at the grid cell level.

**Drawing Modes:
Smooth Color**

fills the surface by using smooth color interpolation at the screen pixel level.

**Drawing Modes:
Off**

toggles the display of the fitted surface.

† **Note:** In color drawing modes, a color legend bar is drawn along the Z axis in 3D space if no **ZColor** variable is specified. Otherwise, a 2D color bar is drawn at the right side of the plot for the **ZColor** variable.

† **Note:** You can create a blended color strip based on the interpolation of up to five colors, as described in [Chapter 11, “Coloring Observations.”](#)

With large data sets, rotation speed can be slow. The most reliable ways to optimize rotation speed are as follows:

- Use only square observation markers.
- Use only one color for observations.
- Use a small marker size, 1 if possible.
- Use **Fast Draw** or **Hidden Line Removal** drawing modes for surface.

When modeling with two explanatory variables, you may want to display a fitted plane in the rotating plot. You can write SAS statements to add planes and surfaces to the data set and rotate them with the original data. Muenchen (1992) has developed and documented a flexible set of SAS statements for this purpose.

References

Becker, R.A., Cleveland, W.S., and Weil, G. (1989), “An Interactive System for Multivariate Data Display,” *Proceedings of the 11th Conference on Probability and Statistics*, Boston: American Meteorological Society.

Muenchen, R.A. (1992), “INSIGHT into Multiple Regression,” *Proceedings of the Seventeenth Annual SAS Users Group International Conference*, 17, 1407–1410.

Chapter 38

Distribution Analyses

Chapter Contents

PARAMETRIC DISTRIBUTIONS	556
Normal Distribution	556
Lognormal Distribution	556
Exponential Distribution	557
Weibull Distribution	557
VARIABLES	558
METHOD	559
OUTPUT	563
TABLES	568
Moments	568
Quantiles	570
Basic Confidence Intervals	571
Tests for Location	572
Frequency Counts	574
Robust Measures of Scale	576
Tests for Normality	578
Trimmed and Winsorized Means	580
GRAPHS	584
Box Plot/Mosaic Plot	584
Histogram/Bar Chart	584
QQ Plot	585
CURVES	589
Parametric Density	590
Kernel Density	592
Empirical CDF	594
CDF Confidence Band	595
Parametric CDF	597
Test for a Specific Distribution	599
Test for Distribution	601
QQ Ref Line	603
ANALYSIS FOR NOMINAL VARIABLES	605

REFERENCES 608

Chapter 38

Distribution Analyses

Choosing **Analyze:Distribution (Y)** gives you access to a variety of *distribution analyses*. For nominal **Y** variables, you can generate bar charts, mosaic plots, and frequency counts tables.

For interval variables, you can generate univariate statistics, such as moments, quantiles, confidence intervals for the mean, standard deviation, and variance, tests for location, frequency counts, robust measures of the scale, tests for normality, and trimmed and Winsorized means.

You can use parametric estimation based on normal, lognormal, exponential, or Weibull distributions to estimate density and cumulative distribution functions and to generate quantile-quantile plots. You can also generate nonparametric density estimates based on normal, triangular, or quadratic kernels.

You can use Kolmogorov statistics to generate confidence bands for the cumulative distribution and to test the hypothesis that the data are from a completely specified distribution with known parameters. You can also test the hypothesis that the data are from a specific family of distributions but with unknown parameters.

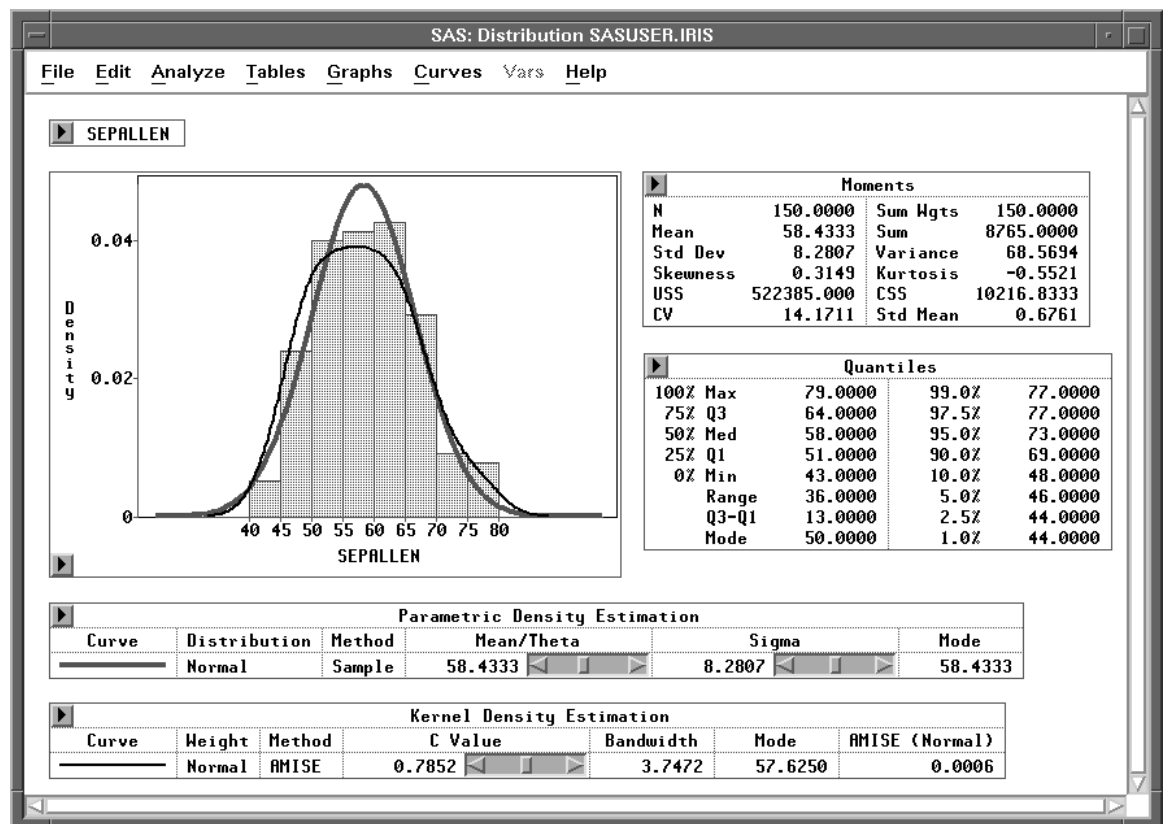


Figure 38.1. Distribution Analysis

Parametric Distributions

A parametric family of distributions is a collection of distributions with a known form that is indexed by a set of quantities called *parameters*. Methods based on parametric distributions of normal, lognormal, exponential, and Weibull are available in a distribution analysis. This section describes the details of each of these distributions. Use of these distributions is described in the sections “[Graphs](#)” and “[Curves](#)” later in this chapter.

You can use both the density function and the cumulative distribution function to identify the distribution. The density function is often more easily interpreted than the cumulative distribution function.

Normal Distribution

The normal distribution has the probability density function

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) \quad \text{for } -\infty < y < \infty$$

where μ is the mean and σ is the scale parameter.

The cumulative distribution function is

$$F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right)$$

where the function Φ is the cumulative distribution function of the standard normal variable: $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp(-u^2/2) du$

Lognormal Distribution

The lognormal distribution has the probability density function

$$f(y) = \frac{1}{y-\theta} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\log(y-\theta)-\zeta}{\sigma}\right)^2\right) \quad \text{for } y > \theta$$

where θ is the threshold parameter, ζ is the scale parameter, and σ is the shape parameter.

The cumulative distribution function is

$$F(y) = \Phi\left(\frac{\log(y-\theta)-\zeta}{\sigma}\right) \quad \text{for } y > \theta$$

Exponential Distribution

The exponential distribution has the probability density function

$$f(y) = \frac{1}{\sigma} \exp\left(-\frac{y - \theta}{\sigma}\right) \quad \text{for } y > \theta$$

where θ is the threshold parameter and σ is the scale parameter.

The cumulative distribution function is

$$F(y) = 1 - \exp\left(-\frac{y - \theta}{\sigma}\right) \quad \text{for } y > \theta$$

Weibull Distribution

The Weibull distribution has the probability density function

$$f(y) = \frac{c}{\sigma} \left(\frac{y - \theta}{\sigma}\right)^{c-1} \exp\left(-\left(\frac{y - \theta}{\sigma}\right)^c\right) \quad \text{for } y > \theta, c > 0$$

where θ is the threshold parameter, σ is the scale parameter, and c is the shape parameter.

The cumulative distribution function is

$$F(y) = 1 - \exp\left(-\left(\frac{y - \theta}{\sigma}\right)^c\right) \quad \text{for } y > \theta$$

Variables

To create a distribution analysis, choose **Analyze:Distribution (Y)**. If you have already selected one or more variables, a distribution analysis for each selected variable appears. If you have not selected any variables, a variables dialog appears.

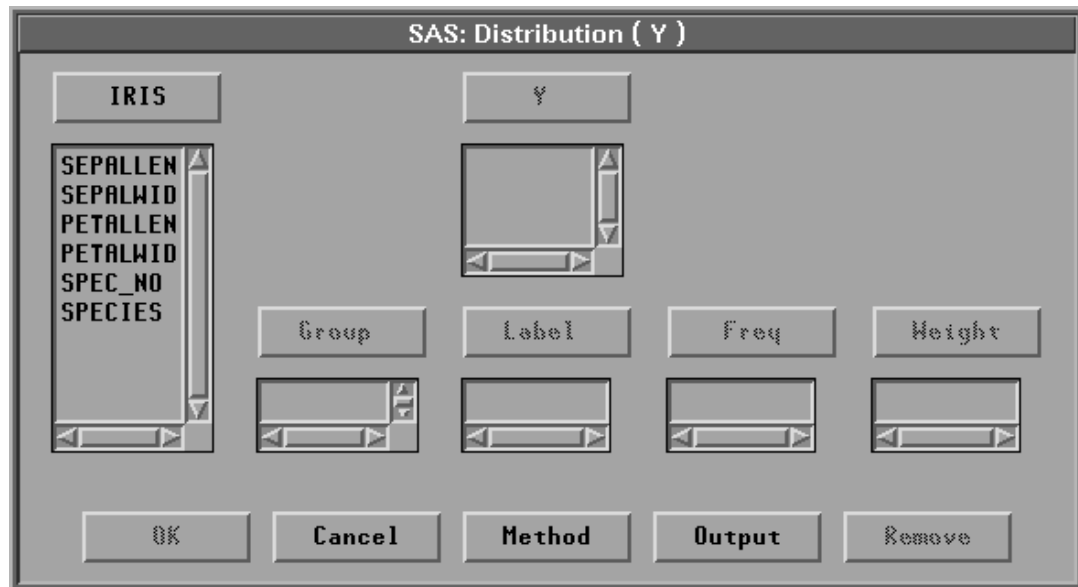


Figure 38.2. Distribution Variables Dialog

Select at least one **Y** variable for each distribution analysis.

You can select one or more **Group** variables if you have grouped data. This creates one distribution analysis for each group.

You can select a **Label** variable to label observations in the plots.

You can select a **Freq** variable. If you select a **Freq** variable, each observation is assumed to represent n observations, where n is the value of the **Freq** variable.

You can select a **Weight** variable to specify relative weights for each observation in the analysis. The details of weighted analyses are explained in the individual sections of this chapter.

Method

Observations with missing values for a **Y** variable are not used in the analysis for that variable. Observations with **Weight** or **Freq** values that are missing or that are less than or equal to zero are not used. Only the integer part of **Freq** values is used.

The following notation is used in the rest of this chapter:

- n is the number of nonmissing values.
- y_i is the i th observed nonmissing value.
- $y_{(i)}$ is the i th ordered nonmissing value, $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$.
- \bar{y} is the sample mean, $\sum_i y_i / n$.
- d is the variance divisor.
- s^2 is the sample variance, $\sum_i (y_i - \bar{y})^2 / d$.
- z_i is the standardized value, $(y_i - \bar{y}) / s$.

The summation \sum_i represents a summation of $\sum_{i=1}^n$.

Based on the variance definition, vardef, the variance divisor d is computed as

- $d = n - 1$ for vardef=**DF**, degrees of freedom
- $d = n$ for vardef=**N**, number of observations

The skewness is a measure of the tendency of the deviations from the mean to be larger in one direction than in the other. The sample skewness is calculated as

- $g_1 = c_{3n} \sum_i z_i^3$ for vardef=**DF**
- $g_1 = \frac{1}{n} \sum_i z_i^3$ for vardef=**N**

where $c_{3n} = \frac{n}{(n-2)} \frac{1}{(n-1)}$.

The kurtosis is primarily a measure of the heaviness of the tails of a distribution. The sample kurtosis is calculated as

- $g_2 = c_{4n} \sum_i z_i^4 - 3c_n$ for vardef=**DF**
- $g_2 = \frac{1}{n} \sum_i z_i^4 - 3$ for vardef=**N**

where $c_{4n} = \frac{n(n+1)}{(n-2)(n-3)} \frac{1}{(n-1)}$ and $c_n = \frac{(n-1)^2}{(n-2)(n-3)}$.

When the observations are independently distributed with a common mean and unequal variances, $\sigma_i^2 = \sigma^2/w_i$, where w_i are individual weights, weighted analyses may be appropriate. You select a **Weight** variable to specify relative weights for each observation in the analysis.

The following notation is used in weighted analyses:

- w_i is the weight associated with y_i .
- $w_{(i)}$ is the weight associated with $y_{(i)}$.
- \bar{w} is the average observation weight, $\sum_i w_i/n$.
- \bar{y}_w is the weighted sample mean, $\sum_i w_i y_i / \sum_i w_i$.
- s_w^2 is the weighted sample variance, $\sum_i w_i (y_i - \bar{y}_w)^2 / d$.
- z_{wi} is the standardized value, $(y_i - \bar{y}_w) / (s_w / \sqrt{w_i})$.

In addition to vardef=DF and vardef=N, the variance divisor is also computed as

- $d = \sum_i w_i - 1$ for vardef=**WDF**, sum of weights minus 1
- $d = \sum_i w_i$ for vardef=**WGT**, sum of weights

With $Var(y_i) = \sigma_i^2 = \sigma^2/w_i$, $Var(\bar{y}_w) = \sigma^2 / \sum_i w_i$ and the expected value

$$E \left(\sum_i w_i (y_i - \bar{y}_w)^2 \right) = E \left(\sum_i w_i (y_i - \mu)^2 - \sum_i w_i (\bar{y}_w - \mu)^2 \right) = (n - 1) \sigma^2$$

† **Note:** The use of vardef=WDF/WGT may not be appropriate since it is the weighted average of individual variances, σ_i^2 , which have unequal expected values.

For vardef=**DF/N**, s_w^2 is the variance of observations with unit weight and may not be informative in the weighted plots of parametric normal distributions. SAS/INSIGHT software uses the weighted sample variance for an observation with average weight, $s_a^2 = s_w^2 / \bar{w}$, to replace s_w^2 in the plots.

The weighted skewness is computed as

- $g_{w1} = c_{3n} \sum_i z_3^{wi} = c_{3n} \sum_i w_i^{\frac{3}{2}} \left(\frac{y_i - \bar{y}}{s_w} \right)^3$ for **DF**
- $g_{w1} = \frac{1}{n} \sum_i z_3^{wi} = \frac{1}{n} \sum_i w_i^{\frac{3}{2}} \left(\frac{y_i - \bar{y}}{s_w} \right)^3$ for **N**

The weighted kurtosis is computed as

- $g_{w2} = c_{4n} \sum_i z_4^{wi} - 3c_n = c_{4n} \sum_i w_i^2 \left(\frac{y_i - \bar{y}}{s_w} \right)^4 - 3c_n$ for **DF**
- $g_{w2} = \frac{1}{n} \sum_i z_4^{wi} - 3 = \frac{1}{n} \sum_i w_i^2 \left(\frac{y_i - \bar{y}}{s_w} \right)^4 - 3$ for **N**

The formulations are invariant under the transformation $w_i^* = cw_i, c > 0$. The sample skewness and kurtosis are set to missing if vardef=**WDF** or vardef=**WGT**.

To view or change the divisor d used in the calculation of variances, or to view or change the use of observations with missing values, click on the **Method** button from the variables dialog to display the method options dialog.

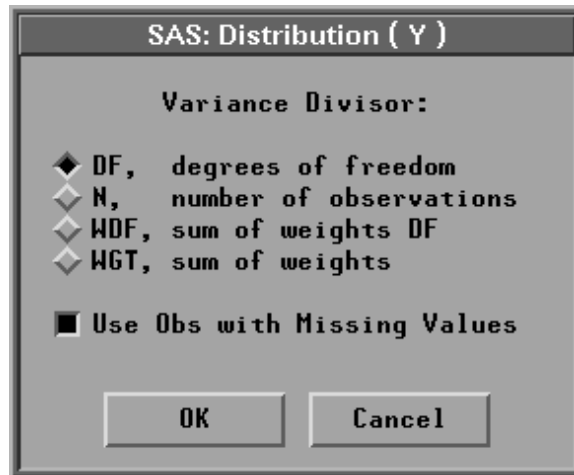


Figure 38.3. Distribution Method Options Dialog

By default, SAS/INSIGHT software uses vardef=**DF, degrees of freedom** to compute the variance divisor.

When multiple **Y** variables are analyzed, and some **Y** variables have missing values, the **Use Obs with Missing Values** option uses all observations with nonmissing values for the **Y** variable being analyzed. If the option is turned off, observations with missing values for *any* **Y** variable are not used for any analysis.

Output

To view or change the options associated with your distribution analysis, click on the **Output** button from the variables dialog. This displays the output options dialog.

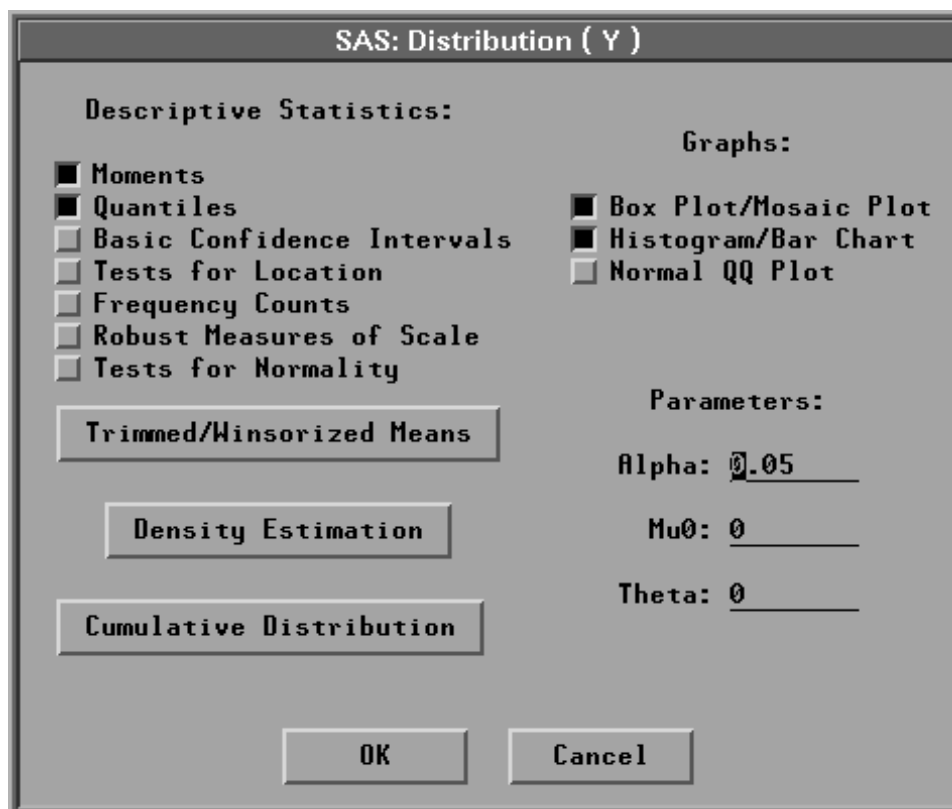


Figure 38.4. Distribution Output Options Dialog

The options you set in this dialog determine which tables and graphs appear in the distribution window. A distribution analysis can include descriptive statistics, graphs, density estimates, and cumulative distribution function estimates. By default, SAS/INSIGHT software displays a moments table, a quantiles table, a box plot, and a histogram. Individual tables and graphs are described following this section.

You can specify the α coefficient in the **Parameters:Alpha:** entry field. The $100(1 - \alpha)\%$ confidence level is used in the basic confidence intervals and the trimmed/Winsorized means tables. You can specify μ_0 in the **Parameters: Mu0:** entry field. μ_0 is used in the tests for location and the trimmed/Winsorized means tables. You can also specify θ in the **Parameters: Theta:** entry field. The parameter θ is used in the parametric density estimation and cumulative distribution for lognormal, exponential, and Weibull distributions.

If you select a **Weight** variable, tables of weighted moments, weighted quantiles, weighted confidence intervals, weighted tests for location, and weighted frequency counts can be generated. Robust measures of scale, tests for normality,

Reference ♦ *Distribution Analyses*

and trimmed/Winsorized means are not computed. Graphs of weighted box plot, weighted histogram, and weighted normal QQ plot can also be generated.

The **Trimmed/Winsorized Means** button enables you to view or change the options associated with trimmed and Winsorized means. Click on **Trimmed/Winsorized Means** to display the **Trimmed/Winsorized Means** dialog.

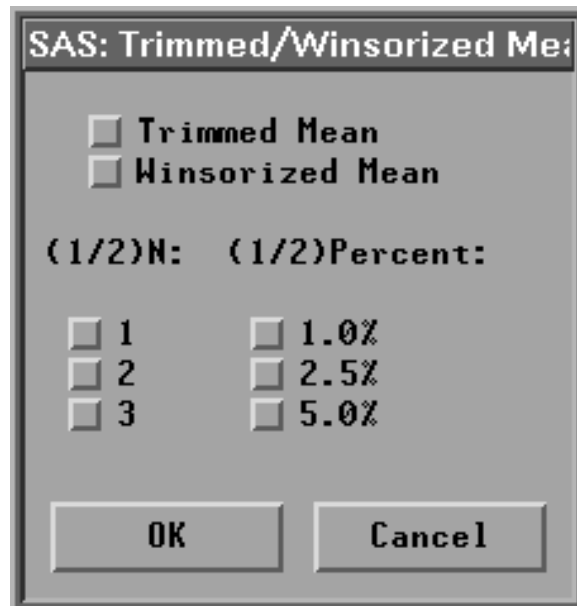


Figure 38.5. Trimmed / Winsorized Means Dialog

In the dialog, you choose the number of observations trimmed or Winsorized in each tail in $(1/2)N$ and the percent of observations trimmed or Winsorized in each tail in $(1/2)\text{Percent}$. If you specify a percentage, the smallest integer greater than or equal to np is trimmed or Winsorized.

The **Density Estimation** button enables you to set the options associated with both parametric density and nonparametric kernel density estimation. Click on **Density Estimation** to display the **Density Estimation** dialog.

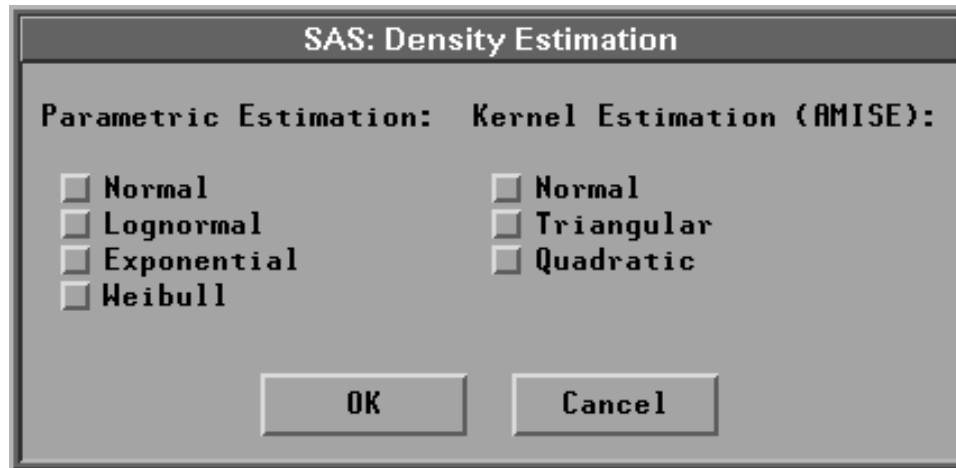


Figure 38.6. Density Estimation Dialog

If you select **Parametric Estimation:Normal**, a normal distribution with the sample mean and standard deviation is created. For the lognormal, exponential, and Weibull distributions, you specify the threshold parameter θ in the **Parameters:Theta:** entry field in the distribution output options dialog, as shown in [Figure 38.4](#), and have the remaining parameters estimated by the maximum-likelihood estimates.

If you select a **Weight** variable, the weighted parametric normal density and weighted kernel density are generated. The parametric lognormal, exponential, and Weibull density are not computed.

The **Cumulative Distribution** button enables you to set the options associated with cumulative distribution estimation. Click on **Cumulative Distribution** to display the **Cumulative Distribution** dialog.

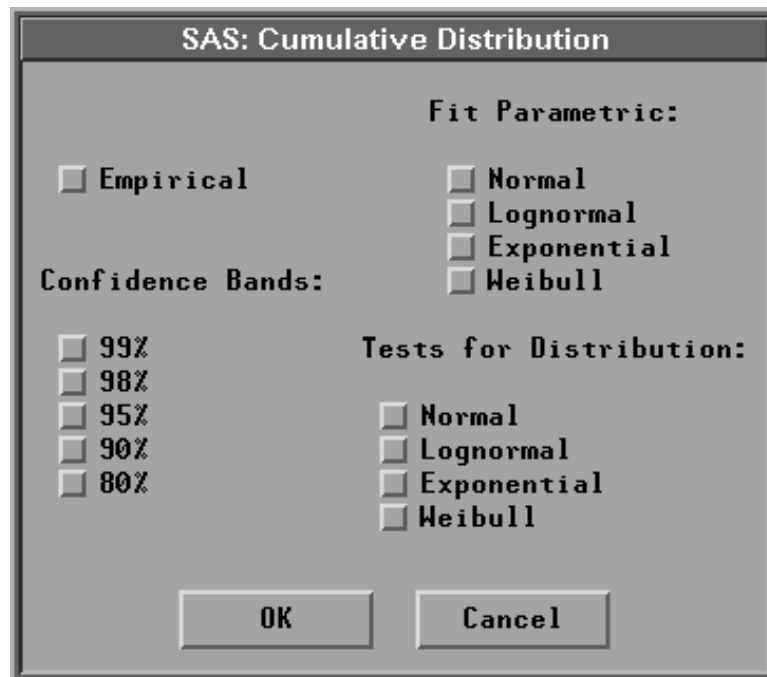


Figure 38.7. Cumulative Distribution Dialog

If you select **Fit Parametric:Normal**, a normal distribution with the sample mean and standard deviation is created. For the lognormal, exponential, and Weibull distributions, you specify the threshold parameter θ in the **Parameters:Theta:** entry field in the distribution output options dialog, as shown in [Figure 38.4](#), and have the remaining parameters estimated by the maximum-likelihood estimates.

If you select a **Weight** variable, weighted empirical and normal cumulative distribution functions can be generated. The confidence bands, the parametric lognormal, exponential, and Weibull cumulative distributions, and tests for distribution are not computed.

Click on **OK** to close the dialogs and create your distribution analysis.

Tables

You can generate distribution tables by setting the options in the output options dialog or by choosing from the **Tables** menu.

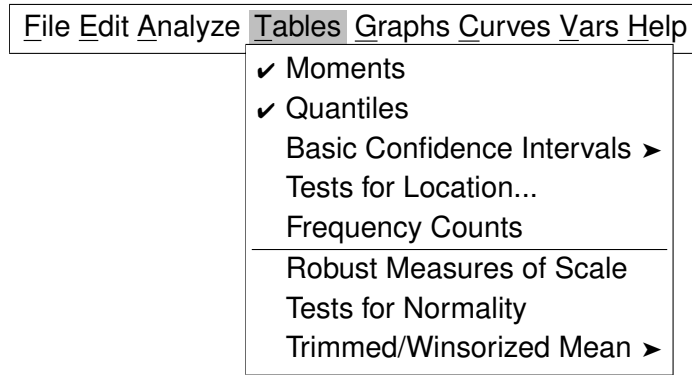


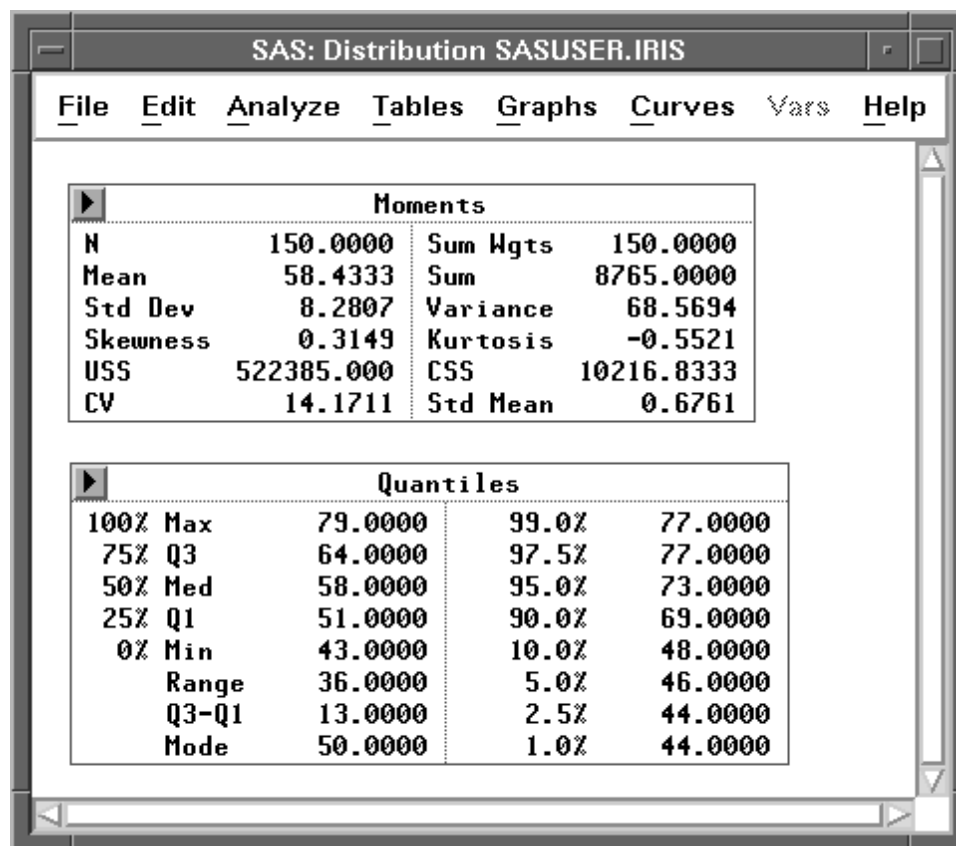
Figure 38.8. Tables Menu

The tables of robust measures of scale, tests for normality, and trimmed/Winsorized mean are not created for weighted analyses.

Moments

The **Moments** table, as shown in [Figure 38.9](#), includes the following statistics:

- **N** is the number of nonmissing values, n .
- **Sum Wgts** is the sum of weights and is equal to n if no **Weight** variable is specified.
- **Mean** is the sample mean, \bar{y} .
- **Sum** is the variable sum, $\sum_i y_i$.
- **Std Dev** is the standard deviation, s .
- **Variance** is the variance, s^2 .
- **Skewness** is the sample skewness, g_1 .
- **Kurtosis** is the sample kurtosis, g_2 .
- **USS** is the uncorrected sum of squares, $\sum_i y_i^2$.
- **CSS** is the sum of squares corrected for the mean, $\sum_i (y_i - \bar{y})^2$.
- **CV** is the percent coefficient of variation, $100s/\bar{y}$.
- **Std Mean** is the standard error of the mean, s/\sqrt{n} . The value is set to missing if vardef \neq **DF**.



The screenshot shows the SAS Distribution window for the variable IRIS. The title bar reads 'SAS: Distribution SASUSER.IRIS'. The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. Two tables are displayed: 'Moments' and 'Quantiles'.

N	150.0000	Sum Wgts	150.0000
Mean	58.4333	Sum	8765.0000
Std Dev	8.2807	Variance	68.5694
Skewness	0.3149	Kurtosis	-0.5521
USS	522385.000	CSS	10216.8333
CV	14.1711	Std Mean	0.6761

100% Max	79.0000	99.0%	77.0000
75% Q3	64.0000	97.5%	77.0000
50% Med	58.0000	95.0%	73.0000
25% Q1	51.0000	90.0%	69.0000
0% Min	43.0000	10.0%	48.0000
Range	36.0000	5.0%	46.0000
Q3-Q1	13.0000	2.5%	44.0000
Mode	50.0000	1.0%	44.0000

Figure 38.9. Moments and Quantiles Tables

For weighted analyses, the **Weighted Moments** table includes the following statistics:

- **N** is the number of nonmissing values, n .
- **Sum Wgts** is the sum of weights, $\sum_i w_i$.
- **Mean** is the weighted sample mean, \bar{y}_w .
- **Sum** is the weighted variable sum, $\sum_i w_i y_i$.
- **Std Dev** is the weighted standard deviation, s_w .
- **Variance** is the weighted variance, s_w^2 .
- **Skewness** is the weighted sample skewness, g_{w1} .
- **Kurtosis** is the weighted sample kurtosis, g_{w2} .
- **USS** is the uncorrected weighted sum of squares, $\sum_i w_i y_i^2$.
- **CSS** is the weighted sum of squares corrected for the mean, $\sum_i w_i (y_i - \bar{y}_w)^2$.
- **CV** is the percent coefficient of variation, $100s_w/\bar{y}_w$.
- **Std Mean** is the standard error of the weighted mean, $s_w/\sum_i w_i$.

The value is set to missing if vardef \neq DF.

Quantiles

It is often convenient to subdivide the area under a density curve so that the area to the left of the dividing value is some specified fraction of the total unit area. For a given value of p between 0 and 1, the p th quantile (or 100 p th percentile) is the value such that the area to the left of it is p .

The p th quantile is computed from the empirical distribution function with averaging:

$$y = \begin{cases} \frac{1}{2}(y_{(i)} + y_{(i+1)}) & \text{if } f = 0 \\ y_{(i+1)} & \text{if } f > 0 \end{cases}$$

where i is the integer part and f is the fractional part of $np = i + f$.

If you specify a **Weight** variable, the p th quantile is computed as

$$y = \begin{cases} \frac{1}{2}(y_{(i)} + y_{(i+1)}) & \text{if } \sum_{j=1}^i w_{(j)} = p \sum_{j=1}^n w_{(j)} \\ y_{(i+1)} & \text{if } \sum_{j=1}^i w_{(j)} < p \sum_{j=1}^n w_{(j)} < \sum_{j=1}^{i+1} w_{(j)} \end{cases}$$

When each observation has an identical weight, the weighted quantiles are identical to the unweighted quantiles.

The **Quantiles** table, as shown in [Figure 38.9](#), includes the following statistics:

- **100% Max** is the maximum, $y_{(n)}$.
- **75% Q3** is the upper quartile (the 75th percentile).
- **50% Med** is the median.
- **25% Q1** is the lower quartile (the 25th percentile).
- **0% Min** is the minimum, $y_{(1)}$.
- **99%, 97.5%, 95%, 90%, 10%, 5%, 2.5%, and 1%** give the corresponding percentiles.
- **Range** is the range, $y_{(n)} - y_{(1)}$.
- **Q3-Q1**, the interquartile range, is the difference between the upper and lower quartiles.
- **Mode** is the most frequently occurring value. When there is more than one mode, the lowest mode is displayed. When all the distinct values have frequency one, the value is set to missing.

Basic Confidence Intervals

Assuming that the population is normally distributed, the **Confidence Intervals** table gives confidence intervals for the mean, standard deviation, and variance at the confidence coefficient specified. You specify the confidence intervals either in the distribution output options dialog or from the **Tables** menu.

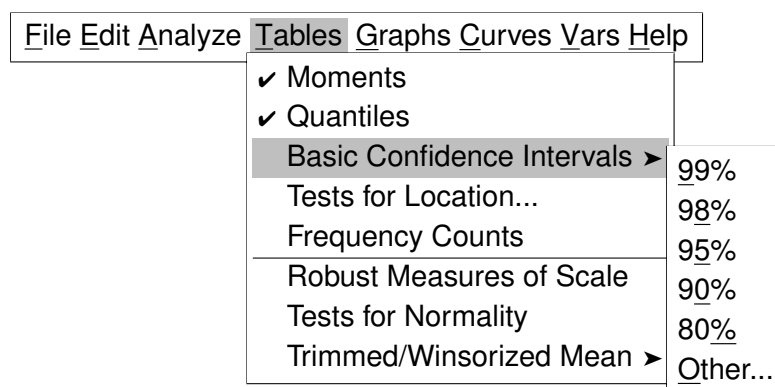


Figure 38.10. Basic Confidence Intervals Menu

The $100(1 - \alpha)\%$ confidence interval for the mean has upper and lower limits

$$\bar{y} \pm t_{(1-\alpha/2)} \frac{s}{\sqrt{n}}$$

where $t_{(1-\alpha/2)}$ is the $(1 - \alpha/2)$ critical value of the Student's t statistic with $n - 1$ degrees of freedom.

For weighted analyses, the limits are

$$\bar{y}_w \pm t_{(1-\alpha/2)} \frac{s_w}{\sqrt{\sum_i w_i}}$$

For large values of n , $t_{(1-\alpha/2)}$ acts as $z_{(1-\alpha/2)}$, the $(1 - \alpha/2)$ critical value of the standard normal distribution.

The $100(1 - \alpha)\%$ confidence interval for the standard deviation has upper and lower limits

$$s \sqrt{\frac{n-1}{c_{\alpha/2}}} \text{ and } s \sqrt{\frac{n-1}{c_{(1-\alpha/2)}}}$$

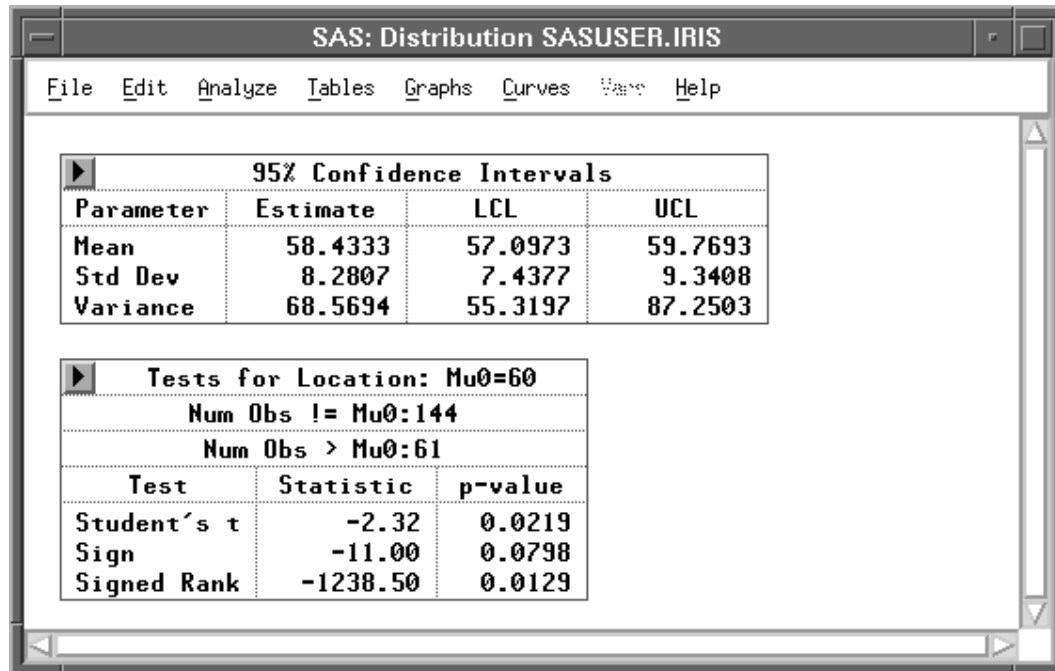
where $c_{\alpha/2}$ and $c_{(1-\alpha/2)}$ are the $\alpha/2$ and $(1 - \alpha/2)$ critical values of the chi-square distribution with $n - 1$ degrees of freedom.

For weighted analyses, the limits are

$$s_w \sqrt{\frac{n-1}{c_{\alpha/2}}} \text{ and } s_w \sqrt{\frac{n-1}{c_{(1-\alpha/2)}}}$$

The $100(1 - \alpha)\%$ confidence interval for the variance has upper and lower limits equal to the squares of the corresponding upper and lower limits for the standard deviation.

Figure 38.11 shows a table of the 95% confidence intervals for the mean, standard deviation, and variance.



The screenshot shows the SAS: Distribution SASUSER.IRIS window. It contains two tables. The first table, titled '95% Confidence Intervals', has four columns: Parameter, Estimate, LCL, and UCL. The second table, titled 'Tests for Location: Mu0=60', has three columns: Test, Statistic, and p-value. It also includes summary statistics for the number of observations not equal to and greater than the hypothesized mean.

95% Confidence Intervals			
Parameter	Estimate	LCL	UCL
Mean	58.4333	57.0973	59.7693
Std Dev	8.2807	7.4377	9.3408
Variance	68.5694	55.3197	87.2503

Tests for Location: Mu0=60		
Num Obs != Mu0:144		
Num Obs > Mu0:61		
Test	Statistic	p-value
Student's t	-2.32	0.0219
Sign	-11.00	0.0798
Signed Rank	-1238.50	0.0129

Figure 38.11. Basic Confidence Intervals and Tests for Location Tables

† **Note:** The confidence intervals are set to missing if vardef≠DF.

Tests for Location

The location tests include the Student's t , sign, and signed rank tests of the hypothesis that the mean/median is equal to a given value μ against the two-sided alternative that the mean/median is not equal to μ . The Student's t test is appropriate when the data are from an approximately normal population; otherwise, nonparametric tests such as the sign or signed rank test should be used.

The **Student's t** gives a Student's t statistic

$$t = \frac{\bar{y} - \mu_0}{s / \sqrt{n}}$$

For weighted analyses, the t statistic is computed as

$$t = \frac{\bar{y}_w - \mu_0}{s_w / \sqrt{\sum_i w_i}}$$

Assuming that the null hypothesis (H_0 : mean = μ) is true and the population is normally distributed, the t statistic has a Student's t distribution with $n - 1$ degrees of freedom. The p -value is the probability of obtaining a Student's t statistic greater in absolute value than the absolute value of the observed statistic t .

† **Note:** The t statistic and p -value are set to missing if vardef \neq DF.

The **Sign** statistic is

$$M = \frac{1}{2}(n^+ - n^-)$$

where n^+ is the number of observations with values greater than μ , and n^- is the number of observations with values less than μ .

Assuming that the null hypothesis (H_0 : median = μ_0) is true, the p -value for the observed statistic M is

$$\text{Prob}\{|M| \geq |M|\} = \left(\frac{1}{2}\right)^{n_t-1} \sum_{i=0}^{\min(n^+, n^-)} \binom{n_t}{i}$$

where $n_t = n^+ + n^-$ is the number of y_i values not equal to μ_0 .

The **Signed Rank** test assumes that the distribution is symmetric. The signed rank statistic is computed as $S = \sum r_i^+ - n_t(n_t + 1)/4$ where r_i^+ is the rank of $|y_i - \mu_0|$ after discarding y_i values equal to μ_0 , and the sum is calculated for values of $y_i > \mu_0$. Average ranks are used for tied values.

The p -value is the probability of obtaining a signed rank statistic greater in absolute value than the absolute value of the observed statistic S . If $n_t \leq 20$, the p -value of the statistic S is computed from the exact distribution of S . When $n_t > 20$, the significance level of S is computed by treating

$$\sqrt{n_t - 1} \frac{S}{\sqrt{n_t V - S^2}}$$

as a Student's t variate with $n_t - 1$ degrees of freedom, where V is computed as

$$V = \frac{1}{24} \{n_t(n_t + 1)(2n_t + 1) - \frac{1}{2} \sum_{j=1}^n t_j(t_j + 1)(t_j - 1)\}.$$

The sum is calculated over groups tied in absolute value, and t_j is the number of tied values in the j th group (Iman 1974, Lehmann 1975).

You can specify location tests either in the distribution output options dialog or in the **Location Tests** dialog after choosing **Tables:Tests for Location** from the menu.

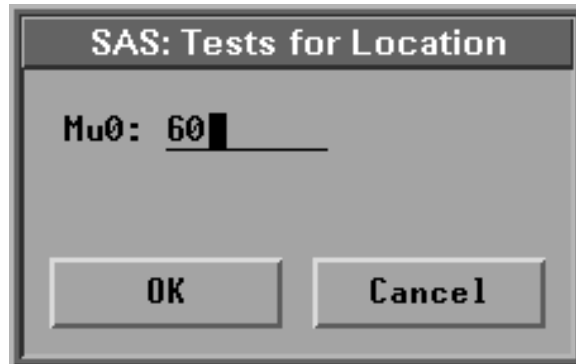


Figure 38.12. Location Tests Dialog

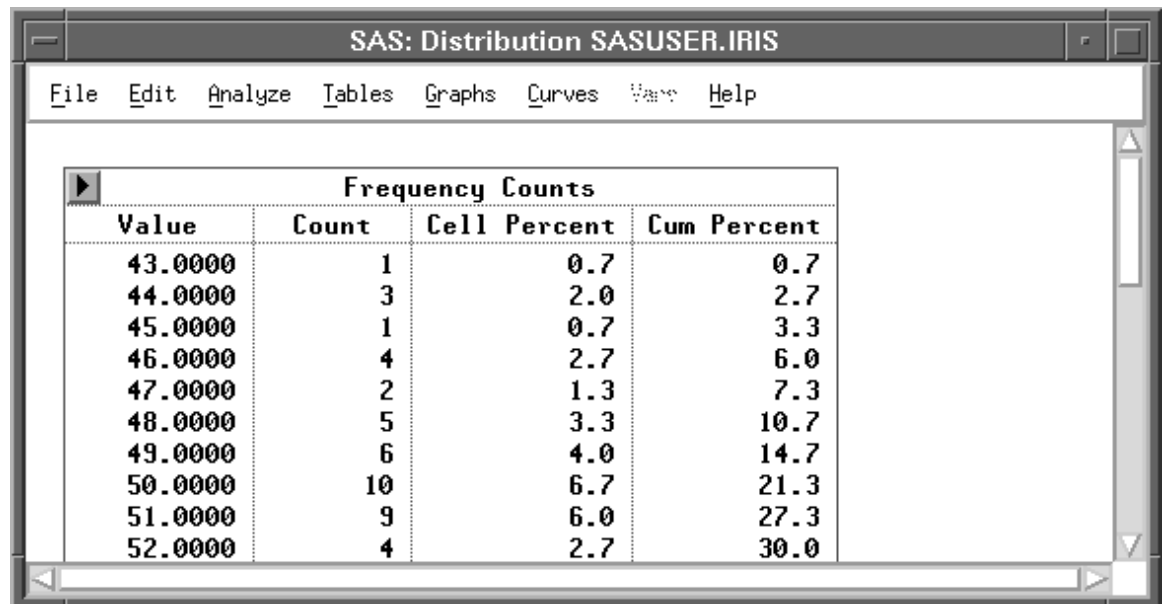
In the dialog, you can specify the parameter μ_0 . [Figure 38.11](#) shows a table of the three location tests for $\mu_0 = 60$. Here, **Num Obs != Mu0** is the number of observations with values not equal to μ_0 , and **Num Obs > Mu0** is the number of observations with values greater than μ_0 .

For weighted analyses, the sign and signed rank tests are not generated.

Frequency Counts

The **Frequency Counts** table, a portion of which is shown in [Figure 38.13](#), includes the variable values, counts, percentages, and cumulative percentages. You can generate frequency tables for both interval and nominal variables.

If you specify a **Weight** variable, the table also includes the weighted counts. These weighted counts are used to compute the percentages and cumulative percentages.



The screenshot shows the SAS Distribution window for the variable SASUSER.IRIS. The window has a menu bar with File, Edit, Analyze, Tables, Graphs, Curves, View, and Help. The main area displays a table titled 'Frequency Counts' with four columns: Value, Count, Cell Percent, and Cum Percent. The data is as follows:

Value	Count	Cell Percent	Cum Percent
43.0000	1	0.7	0.7
44.0000	3	2.0	2.7
45.0000	1	0.7	3.3
46.0000	4	2.7	6.0
47.0000	2	1.3	7.3
48.0000	5	3.3	10.7
49.0000	6	4.0	14.7
50.0000	10	6.7	21.3
51.0000	9	6.0	27.3
52.0000	4	2.7	30.0

Figure 38.13. Frequency Counts Table

Robust Measures of Scale

The sample standard deviation is a commonly used estimator of the population scale. However, it is sensitive to outliers and may not remain bounded when a single data point is replaced by an arbitrary number. With robust scale estimators, the estimates remain bounded even when a portion of the data points are replaced by arbitrary numbers.

A simple robust scale estimator is the interquartile range, which is the difference between the upper and lower quartiles. For a normal population, the standard deviation σ can be estimated by dividing the interquartile range by 1.34898.

Gini's mean difference is also a robust estimator of the standard deviation σ . It is computed as

$$G = \frac{1}{\binom{n}{2}} \sum_{i < j} |y_i - y_j|$$

If the observations are from a normal distribution, then $\sqrt{\pi}G/2$ is an unbiased estimator of the standard deviation σ .

A very robust scale estimator is the median absolute deviation (*MAD*) about the median (Hampel 1974).

$$MAD = \text{med}_i(|y_i - \text{med}_j(y_j)|)$$

where the inner median, $\text{med}_j(y_j)$, is the median of the n observations and the outer median, med_i , is the median of the n absolute values of the deviations about the median.

For a normal distribution, 1.4826 MAD can be used to estimate the standard deviation σ .

The *MAD* statistic has low efficiency for normal distributions and it may not be appropriate for symmetric distributions. Rousseeuw and Croux (1993) proposed two new statistics as alternatives to the *MAD* statistic, S_n and Q_n .

$$S_n = 1.1926 \text{ med}_i(\text{med}_j(|y_i - y_j|))$$

where the outer median, med_i , is the median of the n medians of

$$\{|y_i - y_j|; j = 1, 2, \dots, n\}.$$

To reduce small-sample bias, $c_{sn}S_n$ is used to estimate the standard deviation σ , where c_{sn} is a correction factor (Croux and Rousseeuw 1992).

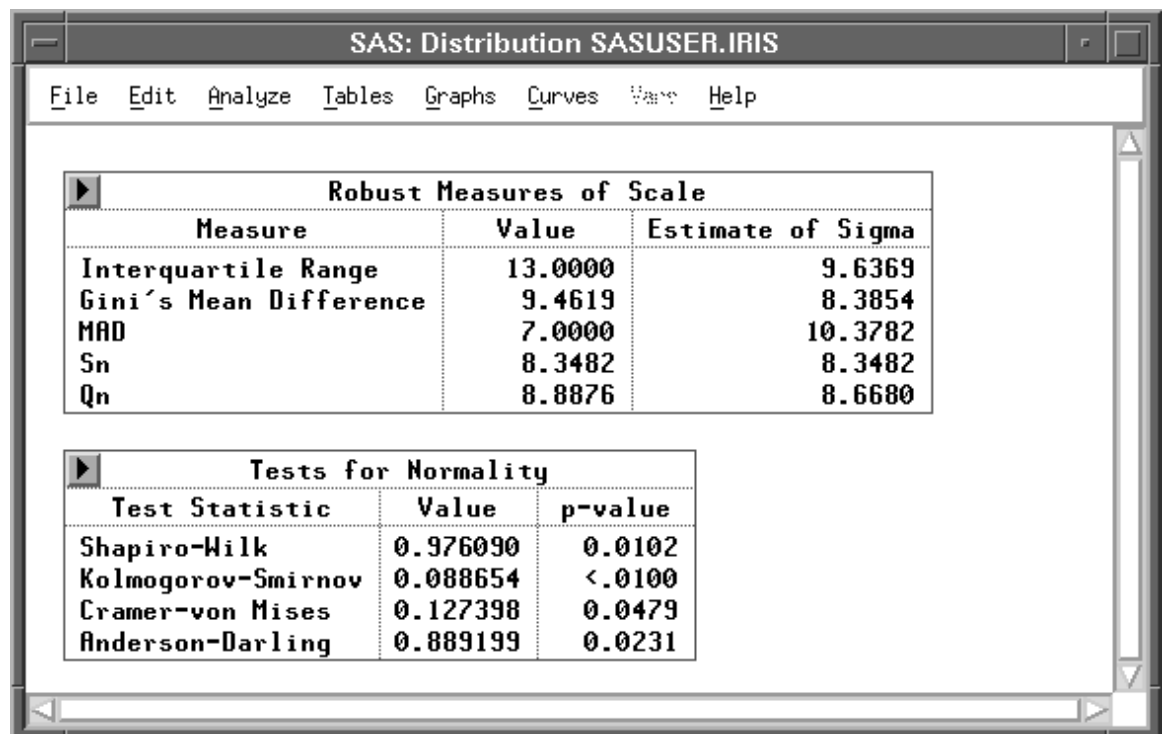
The second statistic is computed as

$$Q_n = 2.2219\{|y_i - y_j|; i < j\}_{(k)}$$

where $k = \binom{h}{2}$, $h = [n/2] + 1$ and $[n/2]$ is the integer part of $n/2$. That is, Q_n is 2.2219 times the k th order statistic of the $\binom{n}{2}$ distances between data points.

The bias-corrected statistic $c_{qn}Q_n$ is used to estimate the standard deviation σ , where c_{qn} is the correction factor.

A **Robust Measures of Scale** table includes the interquartile range, Gini's mean difference, MAD , S_n , and Q_n , with their corresponding estimates of σ , as shown in Figure 38.14.



The screenshot shows the SAS Distribution window for the variable IRIS. It contains two tables: 'Robust Measures of Scale' and 'Tests for Normality'.

Robust Measures of Scale		
Measure	Value	Estimate of Sigma
Interquartile Range	13.0000	9.6369
Gini's Mean Difference	9.4619	8.3854
MAD	7.0000	10.3782
S_n	8.3482	8.3482
Q_n	8.8876	8.6680

Tests for Normality		
Test Statistic	Value	p-value
Shapiro-Wilk	0.976090	0.0102
Kolmogorov-Smirnov	0.088654	<.0100
Cramer-von Mises	0.127398	0.0479
Anderson-Darling	0.889199	0.0231

Figure 38.14. Robust Measures of Scale and Tests for Normality

Tests for Normality

SAS/INSIGHT software provides tests for the null hypothesis that the input data values are a random sample from a normal distribution. These test statistics include the Shapiro-Wilk statistic (W) and statistics based on the empirical distribution function: the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling statistics.

The Shapiro-Wilk statistic is the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance. W must be greater than zero and less than or equal to one, with small values of W leading to rejection of the null hypothesis of normality. Note that the distribution of W is highly skewed. Seemingly large values of W (such as 0.90) may be considered small and lead to the rejection of the null hypothesis.

The W statistic is computed when the sample size is less than or equal to 2000. When the sample size is greater than three, the coefficients for computing the linear combination of the order statistics are approximated by the method of Royston (1992).

With a sample size of three, the probability distribution of W is known and is used to determine the significance level. When the sample size is greater than three, simulation results are used to obtain the approximate normalizing transformation (Royston 1992)

$$Z_n = \begin{cases} (-\log(\gamma - \log(1 - W_n)) - \mu)/\sigma & \text{if } 4 \leq n \leq 11 \\ (\log(1 - W_n) - \mu)/\sigma & \text{if } 12 \leq n \leq 2000 \end{cases}$$

where γ , μ , and σ are functions of n , obtained from simulation results, and Z_n is a standard normal variate with large values indicating departure from normality.

The Kolmogorov statistic assesses the discrepancy between the empirical distribution and the estimated hypothesized distribution. For a test of normality, the hypothesized distribution is a normal distribution function with parameters μ and σ estimated by the sample mean and standard deviation. The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values given by Stephens (1974).

The Cramer-von Mises statistic (W^2) is defined as

$$W^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dF(x)$$

and it is computed as

$$W^2 = \sum_{i=1}^n \left(U_{(i)} - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}$$

where $U_{(i)} = F(y_{(i)})$ is the cumulative distribution function value at $y_{(i)}$, the i th ordered value. The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values given by Stephens (1974).

The Anderson-Darling statistic (A^2) is defined as

$$A^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \{F(x)(1 - F(x))\}^{-1} dF(x)$$

and it is computed as

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^n \{(2i-1)(\log(U_{(i)}) + \log(1 - U_{(n+1-i)}))\}$$

The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values in D'Agostino and Stephens (1986).

A **Tests for Normality** table includes the Shapiro-Wilk, Kolmogorov, Cramer-von Mises, and Anderson-Darling test statistics, with their corresponding p -values, as shown in [Figure 38.14](#).

Trimmed and Winsorized Means

When outliers are present in the data, trimmed and Winsorized means are robust estimators of the population mean that are relatively insensitive to the outlying values. Therefore, trimming and Winsorization are methods for reducing the effects of extreme values in the sample.

The k -times trimmed mean is calculated as

$$\bar{y}_{tk} = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} y_{(i)}$$

The trimmed mean is computed after the k smallest and k largest observations are deleted from the sample. In other words, the observations are trimmed at each end.

The k -times Winsorized mean is calculated as

$$\bar{y}_{wk} = \frac{1}{n} \{ (k+1)y_{(k+1)} + \sum_{i=k+2}^{n-k-1} y_{(i)} + (k+1)y_{(n-k)} \}$$

The Winsorized mean is computed after the k smallest observations are replaced by the $(k+1)$ st smallest observation, and the k largest observations are replaced by the $(k+1)$ st largest observation. In other words, the observations are Winsorized at each end.

For a symmetric distribution, the symmetrically trimmed or Winsorized mean is an unbiased estimate of the population mean. But the trimmed or Winsorized mean does not have a normal distribution even if the data are from a normal population.

The Winsorized sum of squared deviations is defined as

$$s_{wk}^2 = (k+1)(y_{(k+1)} - \bar{y}_{wk})^2 + \sum_{i=k+2}^{n-k-1} (y_{(i)} - \bar{y}_{wk})^2 + (k+1)(y_{(n-k)} - \bar{y}_{wk})^2$$

A robust estimate of the variance of the trimmed mean \bar{y}_{tk} can be based on the Winsorized sum of squared deviations (Tukey and McLaughlin 1963). The resulting trimmed t test is given by

$$t_{tk} = \frac{\bar{y}_{tk}}{\text{STDERR}(\bar{y}_{tk})}$$

where $\text{STDERR}(\bar{y}_{tk})$ is the standard error of \bar{y}_{tk} :

$$\text{STDERR}(\bar{y}_{tk}) = \frac{s_{wk}}{\sqrt{(n-2k)(n-2k-1)}}$$

A Winsorized t test is given by

$$t_{wk} = \frac{\bar{y}_{wk}}{\text{STDERR}(\bar{y}_{wk})}$$

where $\text{STDERR}(\bar{y}_{wk})$ is the standard error of \bar{y}_{wk} :

$$\text{STDERR}(\bar{y}_{wk}) = \frac{n-1}{n-2k-1} \frac{s_{wk}}{\sqrt{n(n-1)}}$$

When the data are from a symmetric distribution, the distribution of the trimmed t statistic t_{tk} or the Winsorized t statistic t_{wk} can be approximated by a Student's t distribution with $n - 2k - 1$ degrees of freedom (Tukey and McLaughlin 1963, Dixon and Tukey 1968).

You can specify the number or percentage of observations to be trimmed or Winsorized from each end either by using the **Trimmed/Winsorized Means** options dialog or by using the **Trimmed/Winsorized Means** dialog after choosing **Tables:Trimmed/Winsorized Mean:(1/2)N** or **Tables:Trimmed/Winsorized Mean:(1/2)Percent** from the menus.

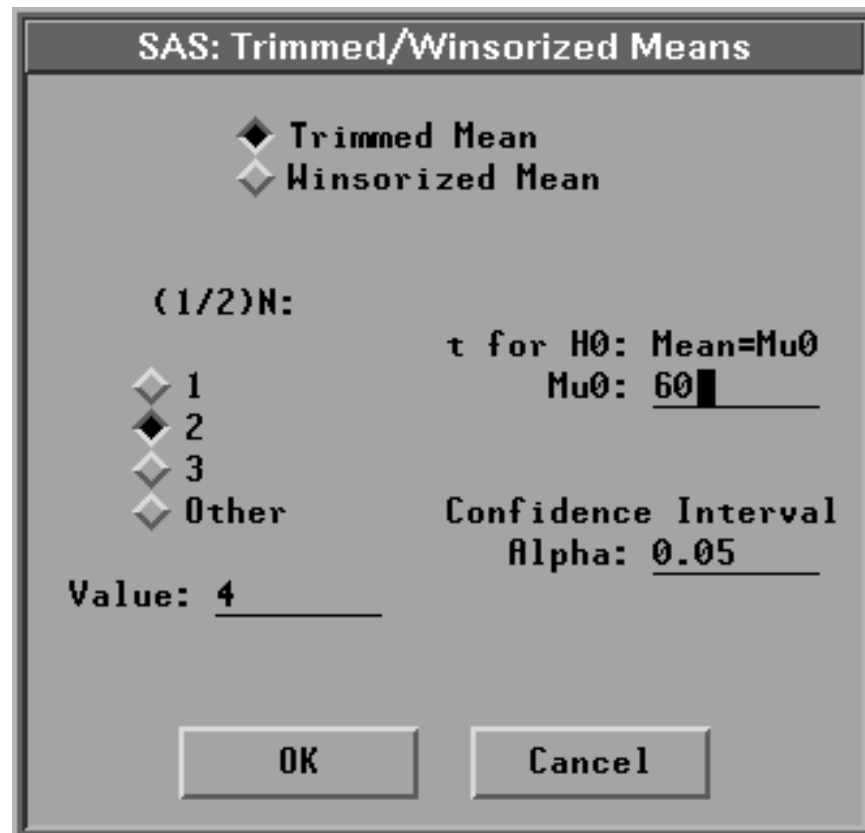


Figure 38.15. (1/2)N Menu

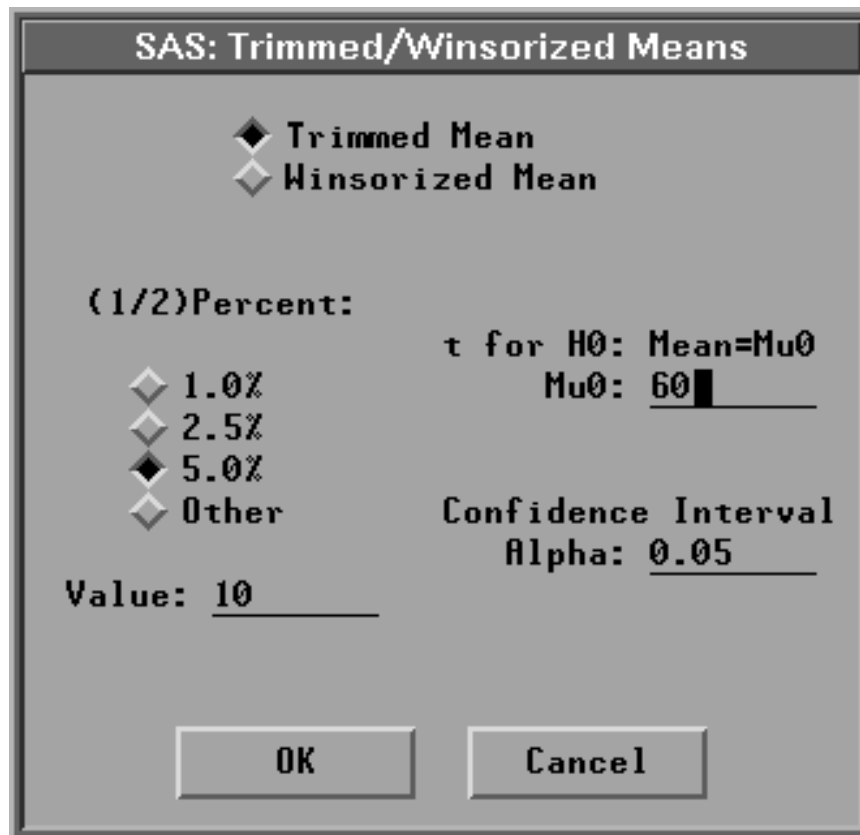
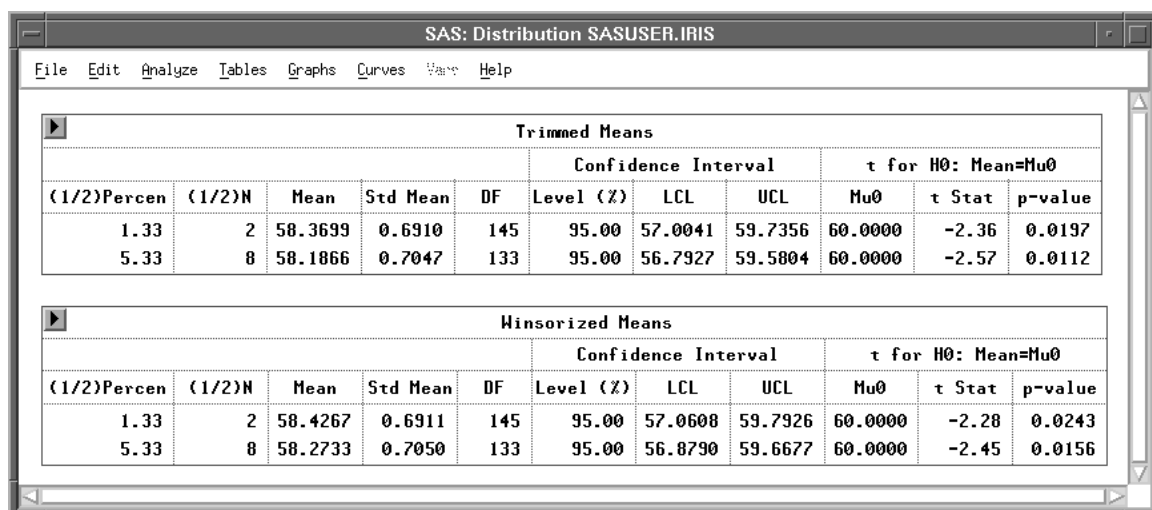


Figure 38.16. (1/2)Percent Menu

If you specify a percentage, $100p\%$, $0 < p < 1$, the smallest integer greater than or equal to np is trimmed or Winsorized from each end.

The **Trimmed Mean** and **Winsorized Mean** tables, as shown in Figure 38.17, contain the following statistics:

- **(1/2)Percent** is the percentage of observations trimmed or Winsorized at each end.
- **(1/2)N** is the number of observations trimmed or Winsorized at each end.
- **Mean** is the trimmed or Winsorized mean.
- **Std Mean** is the standard error of the trimmed or Winsorized mean.
- **DF** is the degrees of freedom used in the Student's t test for the trimmed or Winsorized mean.
- **Confidence Interval** includes **Level (%)**: the confidence level, **LCL**: lower confidence limit, and **UCL**: upper confidence limit.
- **t for H0: Mean=Mu0** includes **Mu0**: the location parameter μ_0 , **t Stat**: the trimmed or Winsorized t statistic for testing the hypothesis that the population mean is μ_0 , and **p-value**: the approximate p -value of the trimmed or Winsorized t statistic.



The screenshot shows the SAS Distribution SASUSER.IRIS window. It contains two tables: 'Trimmed Means' and 'Winsorized Means'. Both tables have columns for (1/2)Percent, (1/2)N, Mean, Std Mean, DF, Level (%), LCL, UCL, Mu0, t Stat, and p-value. The data is presented for two different percent values: 1.33 and 5.33.

Trimmed Means										
(1/2)Percent	(1/2)N	Mean	Std Mean	DF	Confidence Interval			t for H0: Mean=Mu0		
					Level (%)	LCL	UCL	Mu0	t Stat	p-value
1.33	2	58.3699	0.6910	145	95.00	57.0041	59.7356	60.0000	-2.36	0.0197
5.33	8	58.1866	0.7047	133	95.00	56.7927	59.5804	60.0000	-2.57	0.0112

Winsorized Means										
(1/2)Percent	(1/2)N	Mean	Std Mean	DF	Confidence Interval			t for H0: Mean=Mu0		
					Level (%)	LCL	UCL	Mu0	t Stat	p-value
1.33	2	58.4267	0.6911	145	95.00	57.0608	59.7926	60.0000	-2.28	0.0243
5.33	8	58.2733	0.7050	133	95.00	56.8790	59.6677	60.0000	-2.45	0.0156

Figure 38.17. Trimmed Means and Winsorized Means Tables

Graphs

You can generate a histogram, a box plot, or a quantile-quantile plot in the distribution output options dialog or from the **Graphs** menu.

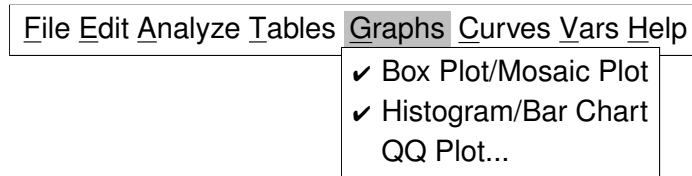


Figure 38.18. Graphs Menu

If you select a **Weight** variable, a weighted box plot/mosaic plot, a weighted histogram/bar chart, and a weighted normal QQ plot can be generated.

Box Plot/Mosaic Plot

The *box plot* is a stylized representation of the distribution of a variable, and it is shown in [Figure 38.19](#). You can also display mosaic plots for nominal variables, as shown in [Figure 38.37](#).

In a box plot, the sample mean and sample standard deviation computed with `vardef=DF` are used in the construction of the mean diamond, as shown in [Figure 38.19](#).

If you select a **Weight** variable, a weighted box plot based on weighted quantiles is created. The weighted sample mean and the weighted sample standard deviation of an observation with average weight for `vardef=DF` is used in the construction of the mean diamond.

⊕ **Related Reading:** Box Plots, [Chapter 33](#).

Histogram/Bar Chart

The histogram is the most widely used density estimator, and it is shown in [Figure 38.19](#). You can also display bar charts for nominal variables, as shown in [Figure 38.37](#).

⊕ **Related Reading:** Bar Charts, [Chapter 32](#).

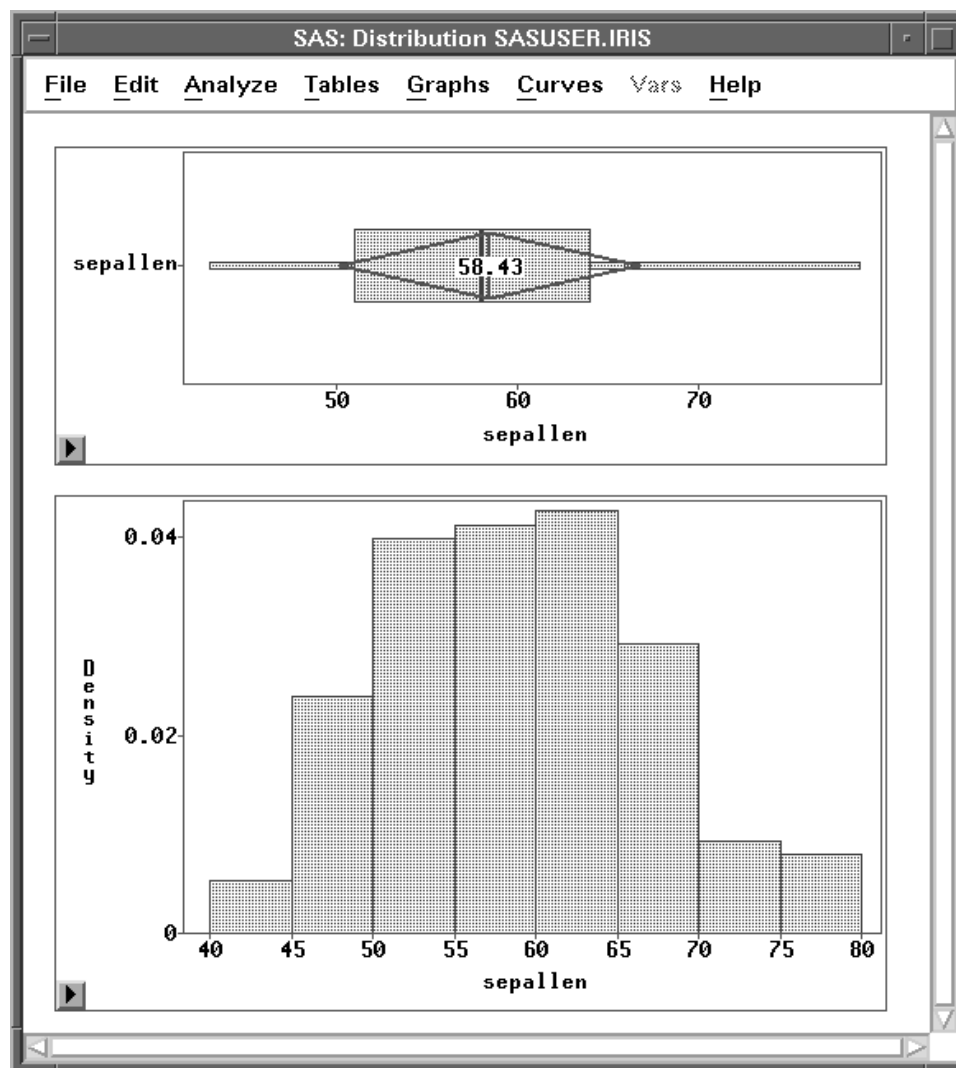


Figure 38.19. Box Plot and Histogram

QQ Plot

A *quantile-quantile plot* (QQ plot) compares ordered values of a variable with quantiles of a specific theoretical distribution. If the data are from the theoretical distribution, the points on the QQ plot lie approximately on a straight line. The normal, lognormal, exponential, and Weibull distributions can be used in the plot.

You can specify the type of QQ plot from the **QQ Plot** dialog after choosing **Graphs:QQ Plot** from the menu.

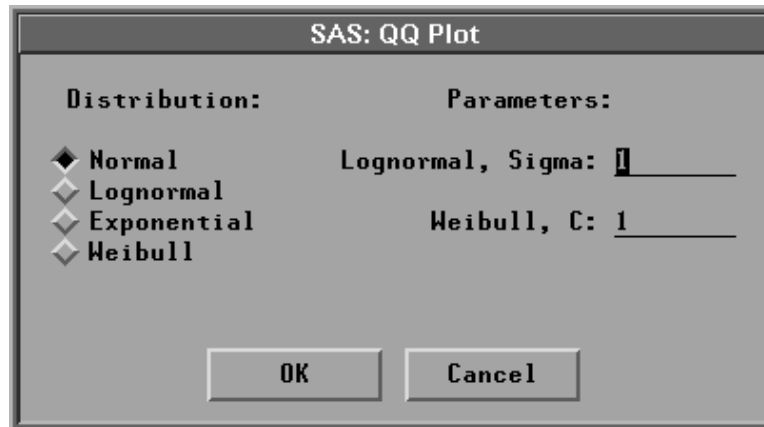


Figure 38.20. QQ Plot Dialog

In the dialog, you must specify a shape parameter for the lognormal or Weibull distribution. The normal QQ plot can also be generated with the graphs options dialog. As described later in this chapter, you can also add a reference line to the QQ plot from the **Curves** menu.

The following expression is used in the discussion that follows:

$$v_i = \frac{i - 0.375}{n + 0.25} \quad \text{for } i = 1, 2, \dots, n$$

where n is the number of nonmissing observations.

For the normal distribution, the i th ordered observation is plotted against the normal quantile $\Phi^{-1}(v_i)$, where Φ^{-1} is the inverse standard cumulative normal distribution. If the data are normally distributed with mean μ and standard deviation σ , the points on the plot should lie approximately on a straight line with intercept μ and slope σ . The normal quantiles are stored in variables named **N_name** for each variable, where **name** is the **Y** variable name.

For the lognormal distribution, the i th ordered observation is plotted against the lognormal quantile $\exp(\sigma\Phi^{-1}(v_i))$ for a given shape parameter σ . If the data are lognormally distributed with parameters θ , σ , and ζ , the points on the plot should lie approximately on a straight line with intercept θ and slope $\exp(\zeta)$. The lognormal quantiles are stored in variables named **L_name** for each variable, where **name** is the **Y** variable name.

For the exponential distribution, the i th ordered observation is plotted against the exponential quantile $-\log(1 - v_i)$. If the data are exponentially distributed with parameters θ and σ , the points on the plot should lie approximately on a straight line with intercept θ and slope σ . The exponential quantiles are stored in variables named **E_name** for each variable, where **name** is the **Y** variable name.

For the Weibull distribution, the i th ordered observation is plotted against the Weibull quantile $(-\log(1 - v_i))^{\frac{1}{c}}$ for a given shape parameter c . If the data are from a Weibull

distribution with parameters θ , σ , and c , the points on the plot should lie approximately on a straight line with intercept θ and slope σ . The Weibull quantiles are stored in variables named **W_name** for each variable, where **name** is the **Y** variable name.

A normal QQ plot is shown in Figure 38.21. You can also add a reference line to the QQ plot from the **Curves** menu. You specify the intercept and slope for the reference line from the **Curves** menu.

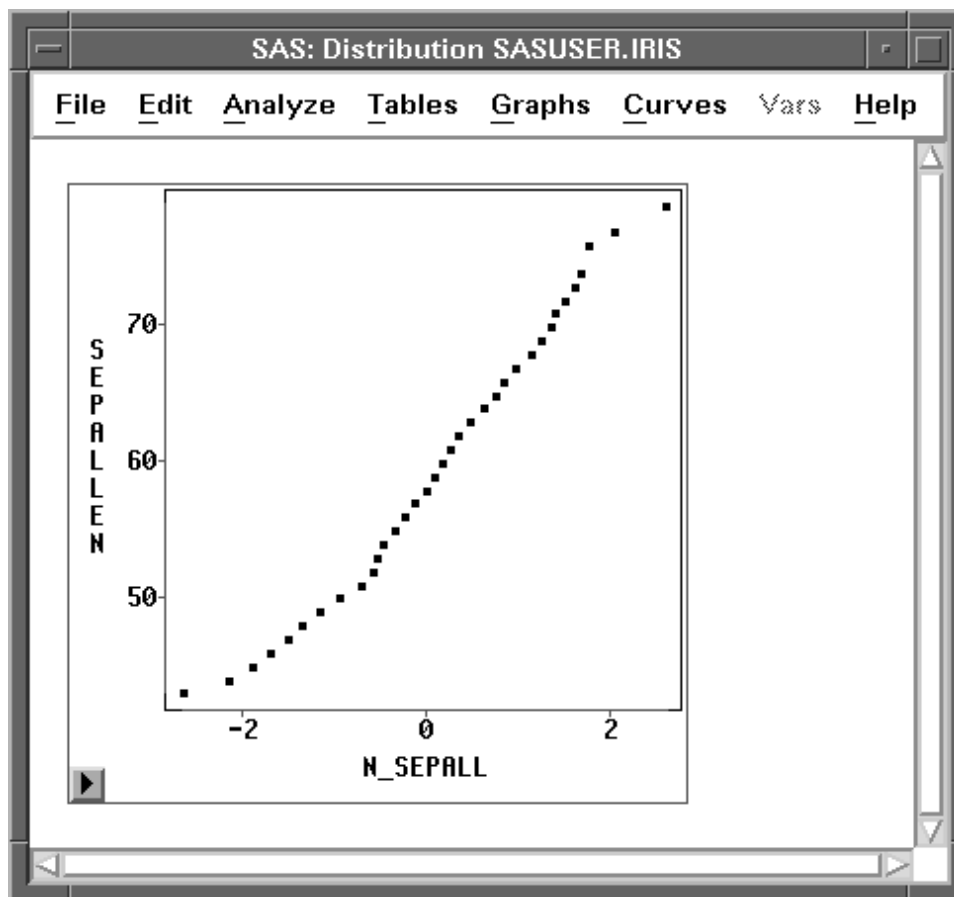


Figure 38.21. Normal QQ Plot

Further information on interpreting quantile-quantile plots can be found in Chambers et al. (1983).

If you select a **Weight** variable, a weighted normal QQ plot can be generated. Lognormal, exponential, and Weibull QQ plots are not computed.

For a weighted normal QQ plot, the i th ordered observation is plotted against the normal quantile $\Phi^{-1}(v_i)$, where

$$v_i = \frac{(\sum_{j=1}^i w_{(j)})(1 - 0.375/i)}{W(1 + 0.25/n)}$$

When each observation has an identical weight, $w_{(j)} = w_0$, the formulation reduces to the usual expression in the unweighted normal probability plot

$$v_i = \frac{i - 0.375}{n + 0.25}$$

If the data are normally distributed with mean μ and standard deviation σ and if each observation has approximately the same weight (w_0), then, as in the unweighted normal QQ plot, the points on the plot should lie approximately on a straight line with intercept μ and slope σ for vardef=WDF/WGT and with slope $\sigma/\sqrt{w_0}$ for vardef=DF/N.

Curves

Density estimation is the construction of an estimate of the density function from the observed data. The methods provided for univariate density estimation include parametric estimators and kernel estimators.

Cumulative distribution analyses include the empirical and the parametric cumulative distribution function. The empirical distribution function is a nonparametric estimator of the cumulative distribution function. You can fit parametric distribution functions if the data are from a known family of distributions, such as the normal, lognormal, exponential, or Weibull.

You can use the Kolmogorov statistic to construct a confidence band for the unknown distribution function. The statistic also tests the hypotheses that the data are from a completely specified distribution or from a specified family of distributions with unknown parameters.

You can generate density estimates and cumulative distribution analysis in the output options dialog, as described previously in the section “Output,” or by choosing from the **Curves** menu, as shown in [Figure 38.22](#). You can also generate QQ reference lines from the **Curves** menu.

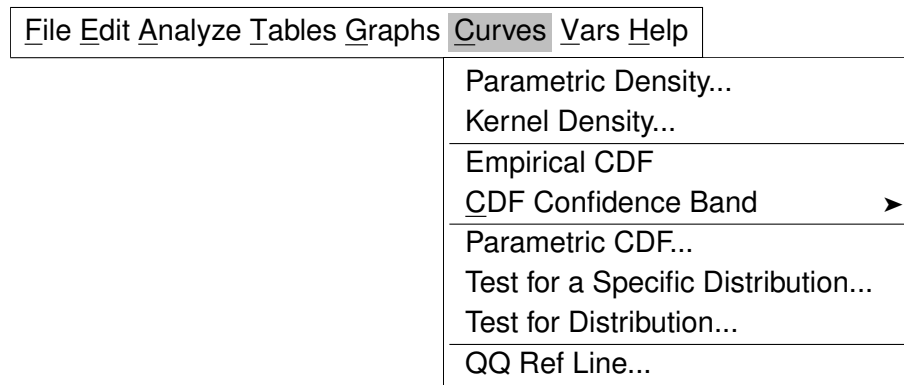


Figure 38.22. Curves Menu

If you select a **Weight** variable, curves of parametric weighted normal density, weighted kernel density, weighted empirical CDF, parametric weighted normal CDF, and weighted QQ reference line (based on weighted least squares) can be generated. CDF confidence band, test for a specific distribution, and test for distribution are not computed.

Parametric Density

Parametric density estimation assumes that the data are from a known family of distributions, such as the normal, lognormal, exponential, and Weibull. After choosing **Curves:Parametric Density** from the menu, you specify the family of distributions in the **Parametric Density Estimation** dialog, as shown in [Figure 38.23](#).

Figure 38.23. Parametric Density Dialog

The default uses a normal distribution with the sample mean and standard deviation as estimates for μ and σ . You can also specify your own μ and σ parameters for the normal distribution by choosing **Method:Specification** in the dialog.

For the lognormal, exponential, and Weibull distributions, you can specify your own threshold parameter θ in the **Parameter:MLE, Theta** entry field and have the remaining parameters estimated by the maximum-likelihood estimates (MLE) by choosing **Method:Sample Estimates/MLE**. Otherwise, you can specify all the parameters in the **Specification** fields and choose **Method:Specification** in the dialog.

If you select a **Weight** variable, only normal density can be created. For **Method:Sample Estimates/MLE**, \bar{y}_w and s_w are used to display the density with vardef=WDF/WGT; \bar{y}_w and s_a are used with vardef=DF/N. For **Method:Specification**, the values in the entry fields **Mean/Theta** and **Sigma** are used to display the density with vardef=WDF/WGT; the values of **Mean/Theta** and **Sigma**/ \sqrt{w} are used with vardef=DF/N.

Figure 38.24 displays a normal density estimate with $\mu = 58.4333$ (the sample mean) and $\sigma = 8.2807$ (the sample standard deviation). It also displays a lognormal density estimate with $\theta = 30$ and with σ and ζ estimated by the MLE.

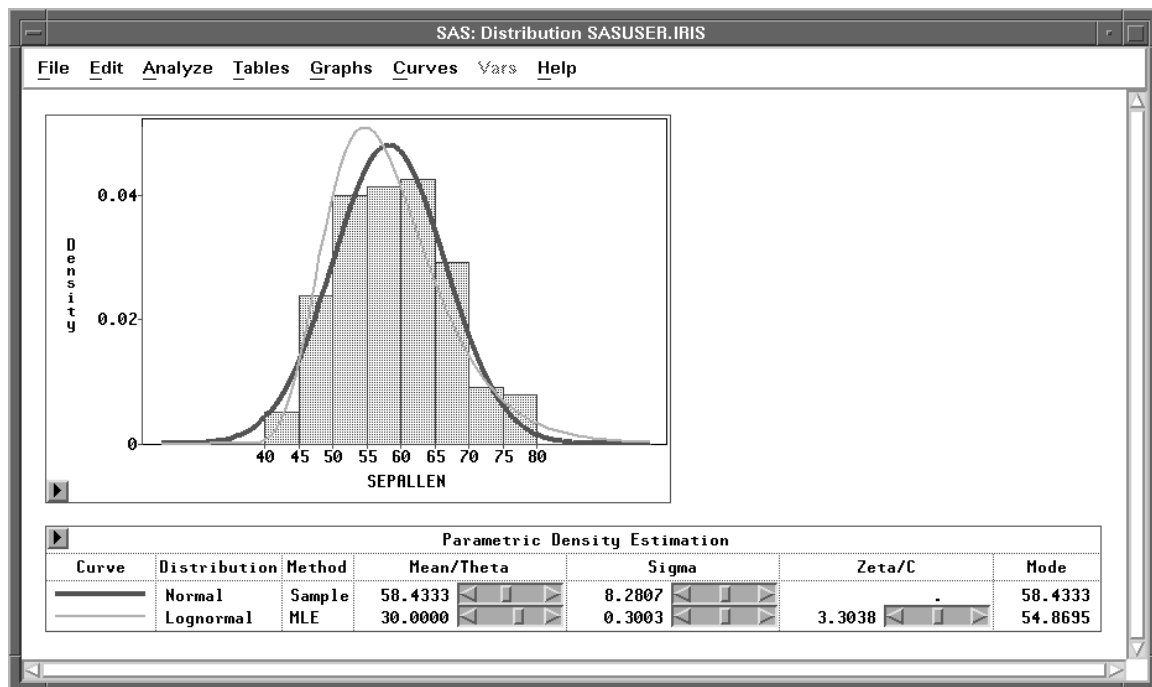


Figure 38.24. Parametric Density Estimation

The **Mode** is the point with the largest estimated density. Use sliders in the table to change the density estimate. When MLE is used for the lognormal, exponential, and Weibull distributions, changing the value of θ in the **Mean/Theta** slider also causes the remaining parameters to be estimated by the MLE for the new θ .

Kernel Density

Kernel density estimation provides normal, triangular, and quadratic kernel density estimators. The general form of a kernel estimator is

$$\hat{f}_\lambda(y) = \frac{1}{n\lambda} \sum_{i=1}^n K_0\left(\frac{y - y_i}{\lambda}\right)$$

where K_0 is a kernel function and λ is the bandwidth.

Some symmetric probability density functions commonly used as kernel functions are

- *Normal* $K_0(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$ for $-\infty < t < \infty$
- *Triangular* $K_0(t) = \begin{cases} 1 - |t| & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- *Quadratic* $K_0(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$

Both theory and practice suggest that the choice of a kernel function is not crucial to the statistical performance of the method (Epanechnikov 1969). With a specific kernel function, the value of λ determines the degree of averaging in the estimate of the density function and is called a *smoothing parameter*. You select a bandwidth λ for each kernel estimator by specifying c in the formula

$$\lambda = n^{-\frac{1}{5}} Q c$$

where Q is the sample interquartile range of the \mathbf{Y} variable. This formulation makes c independent of the units of \mathbf{Y} .

For a specific kernel function, the discrepancy between the density estimator $\hat{f}_\lambda(y)$ and the true density $f(y)$ can be measured by the mean integrated square error

$$\text{MISE}(\lambda) = \int_{\mathbf{y}} \{E(\hat{f}_\lambda(y)) - f(y)\}^2 dy + \int_{\mathbf{y}} \text{Var}(\hat{f}_\lambda(y)) dy$$

which is the sum of the integrated square bias and the integrated variance.

An approximate mean integrated square error based on the bandwidth λ is

$$\text{AMISE}(\lambda) = \frac{1}{4} \lambda^4 \left(\int_{\mathbf{t}} t^2 K(t) dt \right)^2 \int_{\mathbf{y}} (f''(y))^2 dy + \frac{1}{n\lambda} \int_{\mathbf{t}} K(t)^2 dt$$

If $f(y)$ is assumed normal, then a bandwidth based on the sample mean and variance can be computed to minimize AMISE. The resulting bandwidth for a specific kernel is used when the associated kernel function is selected in the density estimation options dialog. This is equivalent to choosing **MISE** from the normal, triangular, or quadratic kernel menus. If $f(y)$ is not roughly normal, this choice may not be appropriate.

SAS/INSIGHT software divides the range of the data into 128 evenly spaced intervals, then approximates the data on this grid and uses the fast Fourier transformation (Silverman 1986) to estimate the density.

If you select a **Weight** variable, the kernel estimator is modified to include the individual observation weights.

$$\hat{f}_\lambda(y) = \frac{1}{\sum_i w_i \lambda} \sum_{i=1}^n w_i K_0 \left(\frac{y - y_i}{\lambda} \right)$$

You can specify the kernel function in the density estimation options dialog or from the **Curves** menu. When you specify the kernel function in the density estimation options dialog, **AMISE** is used. After choosing **Curves:Kernel Density** from the menu, you can specify the kernel function and use either **AMISE** or a specified **C** value in the **Kernel Density Estimation** dialog.

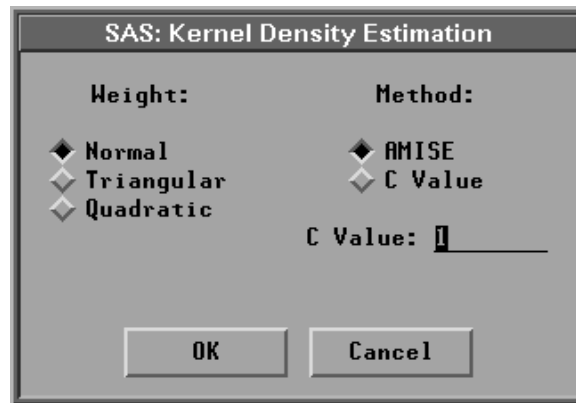


Figure 38.25. Kernel Density Dialog

The default uses a normal kernel density with a c value that minimizes the AMISE. [Figure 38.26](#) displays normal kernel estimates with $c = 0.7852$ (the AMISE value) and $c = 0.25$. Small values of c (and hence small values of the smoothing parameter λ) provide jagged estimates as the curve more closely follows the data points. Large values of c provide smoother estimates. The **Mode** is the point with the largest estimated density. Use the slider to change the smoothing parameter, c .

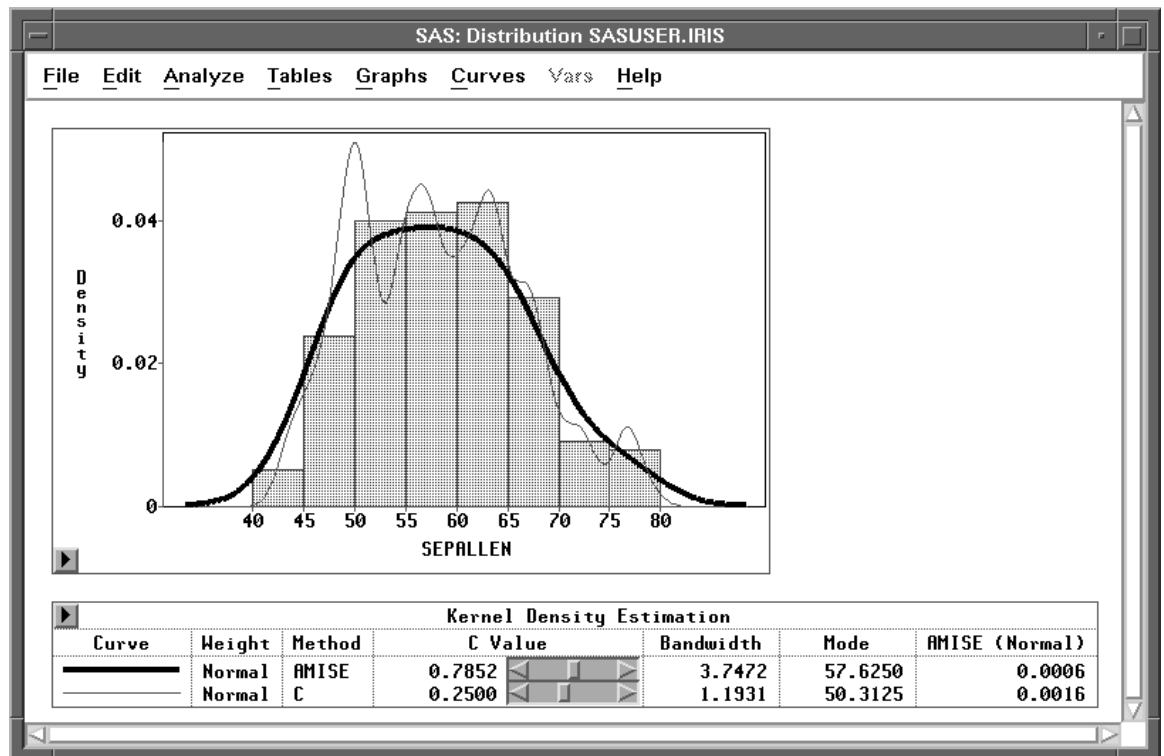


Figure 38.26. Kernel Density Estimation

Empirical CDF

The *empirical distribution function* of a sample, $F_n(y)$, is the proportion of observations less than or equal to y .

$$F_n(y) = \frac{1}{n} \sum_{i=1}^n I(y_i \leq y)$$

where n is the number of observations, and $I(y_i \leq y)$ is an indicator function with value 1 if $y_i \leq y$ and with value 0 otherwise.

The Kolmogorov statistic D is a measure of the discrepancy between the empirical distribution and the hypothesized distribution.

$$D = \text{Max}_y |F_n(y) - F(y)|$$

where $F(y)$ is the hypothesized cumulative distribution function. The statistic is the maximum vertical distance between the two distribution functions. The Kolmogorov statistic can be used to construct a confidence band for the unknown distribution function, to test for a hypothesized completely known distribution, and to test for a specific family of distributions with unknown parameters.

If you select a **Weight** variable, the weighted empirical distribution function is the proportion of observation weights for observations less than or equal to y .

$$F_w(y) = \frac{1}{\sum_i w_i} \sum_{i=1}^n w_i I(y_i \leq y)$$

CDF Confidence Band

The *confidence band* gives a confidence region for the population distribution. The critical values given by Feller (1948) for the completely specified hypothesized distribution are used to generate the confidence band. All parameters in the hypothesized distribution are known. The null hypothesis that the population distribution is equal to a given completely specified distribution is rejected if the hypothesized distribution falls outside the confidence band at any point.

You specify the confidence coefficient in the cumulative distribution options dialog or by choosing **Curves:CDF Confidence Band**.

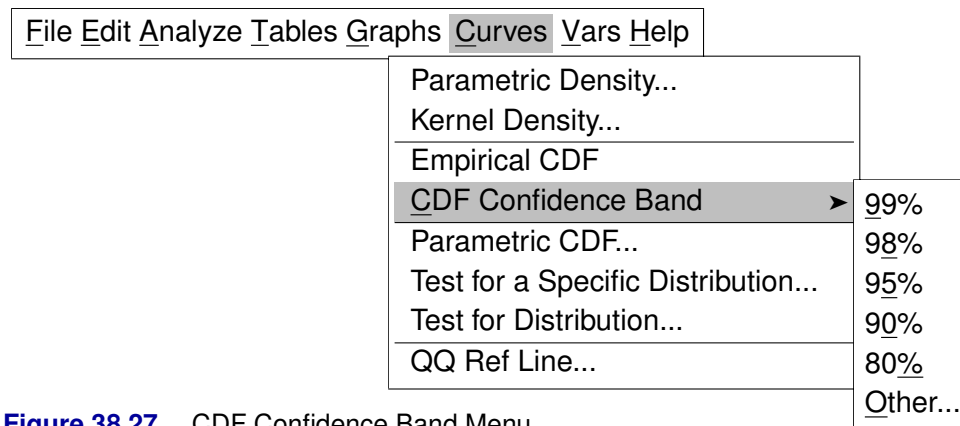


Figure 38.27. CDF Confidence Band Menu

Figure 38.28 displays an empirical distribution function and a 95% confidence band for the cumulative distribution function. Use the **Coefficient** slider to change the coefficient for the confidence band.

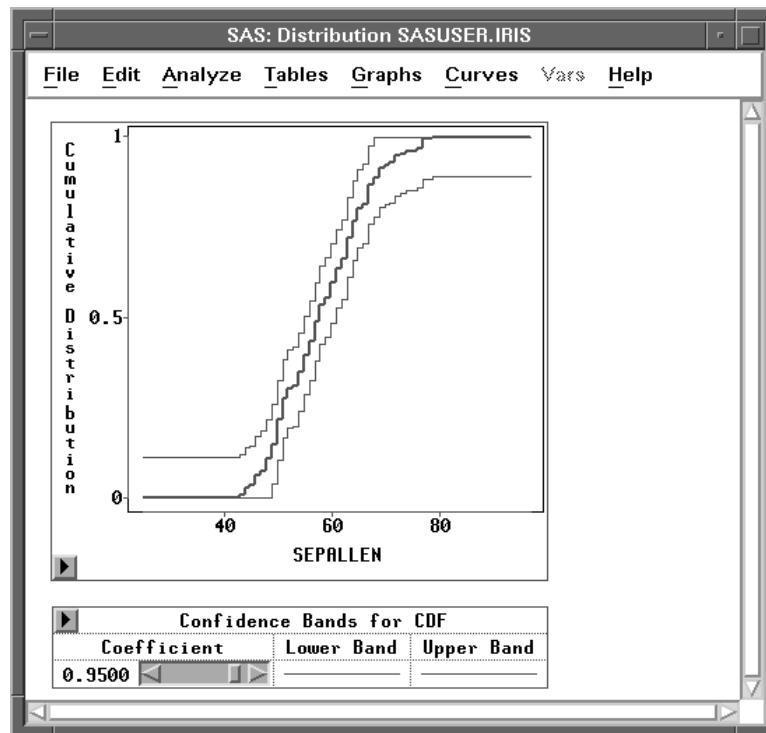


Figure 38.28. CDF Confidence Band

Parametric CDF

You can fit the normal, lognormal, exponential, and Weibull distributions to your data. You specify the family of distributions either in the cumulative distribution options dialog or from the **Parametric CDF Estimation** dialog after choosing **Curves:Parametric CDF** from the menu.

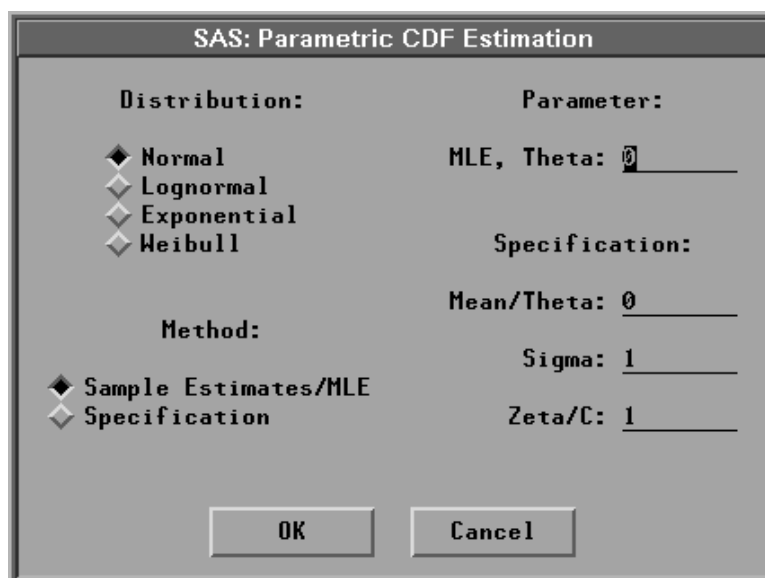


Figure 38.29. Parametric CDF Dialog

For the normal distribution, you can specify your own μ and σ parameters from the **Fit Parametric** menu. Otherwise, you can use the sample mean and standard deviation as estimates for μ and σ by selecting **Fit Parametric:Normal** in the cumulative distribution options dialog or by choosing **Distribution:Normal** and **Method:Sample Estimates/MLE** in the **Parametric CDF Estimation** dialog.

For the lognormal, exponential, and Weibull distributions, you can specify your own threshold parameter θ and have the remaining parameters estimated by the maximum-likelihood method, or you can specify all the distribution parameters in the **Parametric CDF Estimation** dialog. Otherwise, you can have the threshold parameter set to 0 and the remaining parameters estimated by the maximum-likelihood method. To do this, select **Lognormal**, **Exponential**, or **Weibull** in the Cumulative Distribution Output dialog or choose **Method:Sample Estimates/MLE** and **Parameter:MLE, Theta:0** in the **Parametric CDF Estimation** dialog.

If you select a **Weight** variable, only normal CDF can be created. For **Method:Sample Estimates/MLE**, \bar{y}_w and s_w are used to display the cumulative distribution function with vardef=WDF/WGT; \bar{y}_w and s_a are used with vardef=DF/N. For **Method:Specification**, the values in the entry fields **Mean/Theta** and **Sigma** are used to display the cumulative distribution function with vardef=WDF/WGT; the values of **Mean/Theta** and **Sigma**/ \sqrt{w} are used with vardef=DF/N.

Figure 38.30 displays a normal distribution function with $\mu = 58.4333$ (the sample mean) and $\sigma = 8.2807$ (the sample standard deviation); it also displays a lognormal distribution function with $\theta = 30$ and σ and ζ estimated by the MLE.

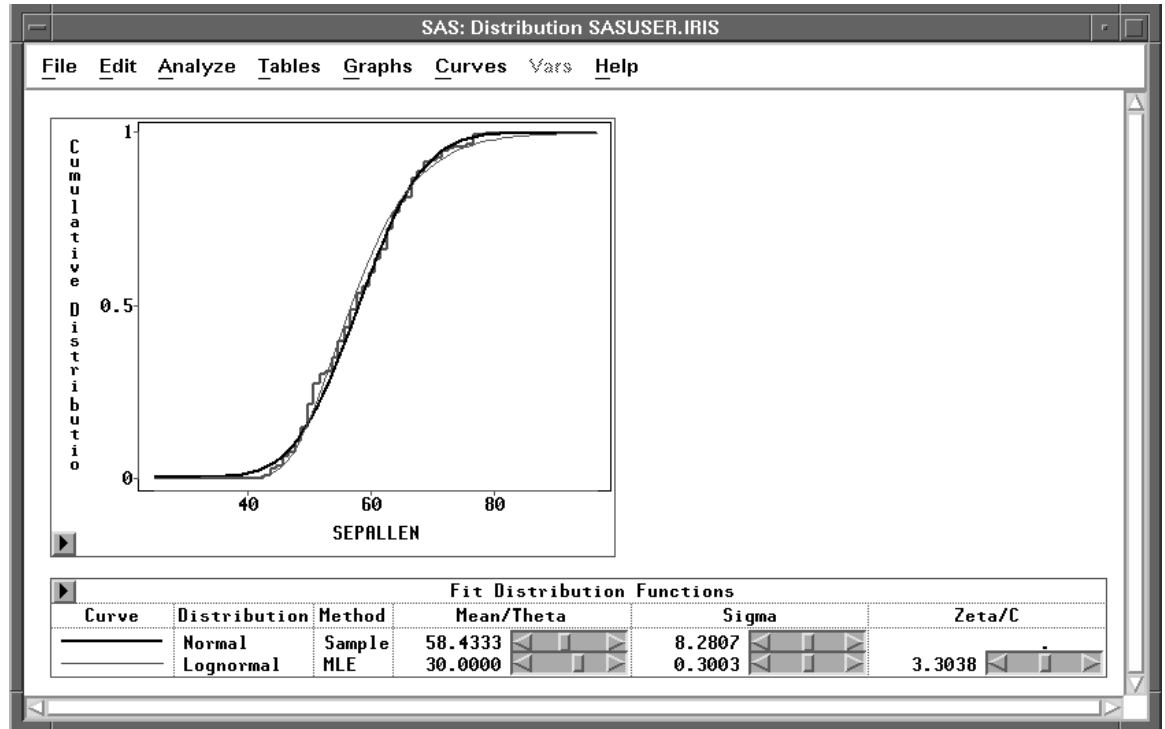


Figure 38.30. Parametric CDF

Use sliders to change the CDF estimate. When MLE is used for the lognormal, exponential, and Weibull distributions, changing the value of θ in the slider also causes the remaining parameters to be estimated by the MLE for the new θ .

Test for a Specific Distribution

You can test whether the data are from a specific distribution with known parameters by using the Kolmogorov statistic. The probability of a larger Kolmogorov statistic is given in Feller (1948). After choosing **Curves:Test for a Specific Distribution** from the menu, you can specify the distribution and its parameters in the **Test for a Specific Distribution** dialog.

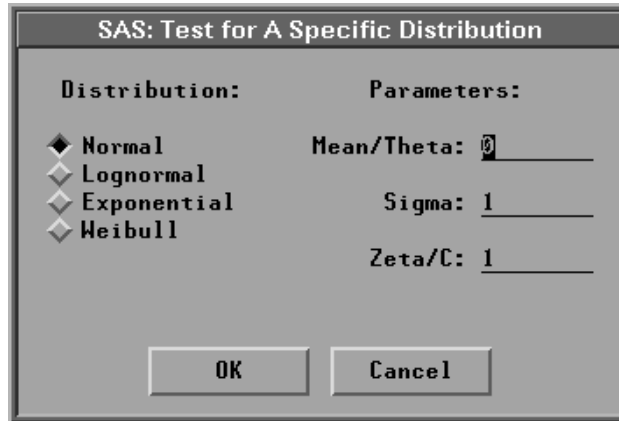


Figure 38.31. Test for a Specific Distribution Dialog

The default tests that the data are from a normal distribution with $\mu = 0$ and $\sigma = 1$. [Figure 38.32](#) shows a test for a specified normal distribution ($\mu = 60$, $\sigma = 10$). Use sliders to change the distribution parameters and have the test results updated accordingly.

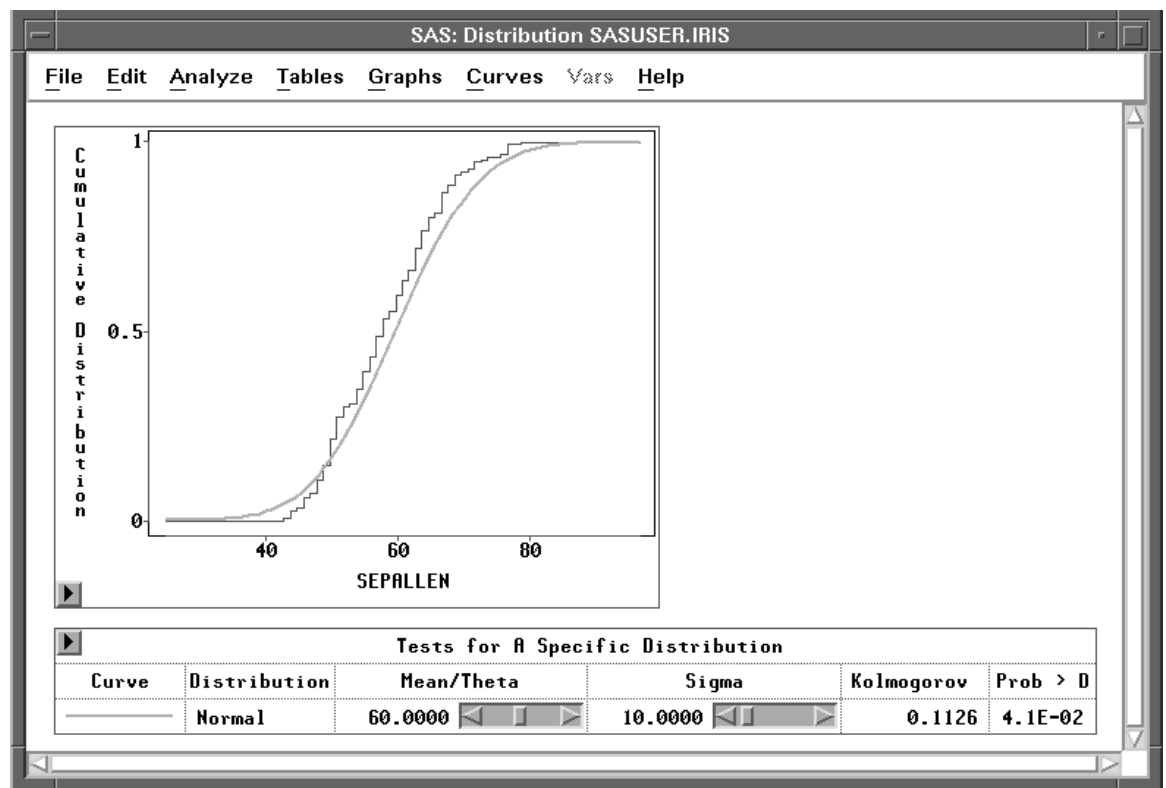


Figure 38.32. Test for a Specific Distribution

Test for Distribution

You can test that the data are from a specific family of distributions, such as the normal, lognormal, exponential, or Weibull distributions. You do not need to specify the distribution parameters except the threshold parameters for the lognormal, exponential, and Weibull distributions. The Kolmogorov statistic assesses the discrepancy between the empirical distribution and the estimated hypothesized distribution F .

For a test of normality, the hypothesized distribution is a normal distribution function with parameters μ and σ estimated by the sample mean and standard deviation. The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values given by Stephens (1974).

For a test of whether the data are from a lognormal distribution, the hypothesized distribution is a lognormal distribution function with a given parameter θ and parameters ζ and σ estimated from the sample after the logarithmic transformation of the data, $\log(y - \theta)$. The sample mean and standard deviation of the transformed sample are used as the parameter estimates. The test is therefore equivalent to the test of normality on the transformed sample.

For a test of exponentiality, the hypothesized distribution is an exponential distribution function with a given parameter θ and a parameter σ estimated by $\bar{y} - \theta$. The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values given by Stephens (1974).

For a test of whether the data are from a Weibull distribution, the hypothesized distribution is a Weibull distribution function with a given parameter θ and parameters c and σ estimated by the maximum-likelihood method. The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values given by Chandra, Singpurwalla, and Stephens (1981).

You specify the distribution in the cumulative distribution options dialog or in the **Test for Distribution** dialog after choosing **Curves:Test for Distribution** from the menu, as shown in Figure 38.33. You can also specify a threshold parameter other than zero for lognormal, exponential, and Weibull distributions.

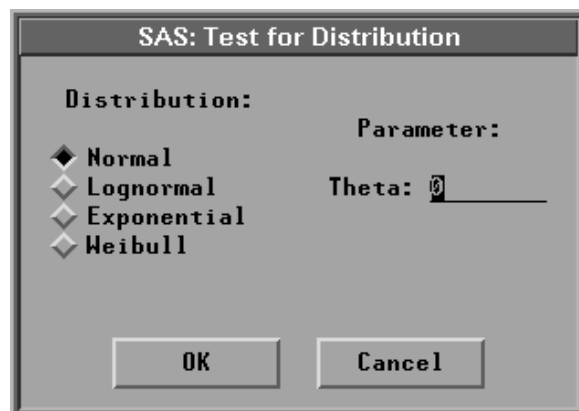


Figure 38.33. Test for Distribution Dialog

Reference ♦ Distribution Analyses

The default tests that the data are from a normal distribution. A test for normality and a test for lognormal distribution with $\theta = 30$ are given in [Figure 38.34](#). You can use the **Mean/Theta** slider to adjust the threshold parameter, θ , for lognormal, exponential, and Weibull distributions.

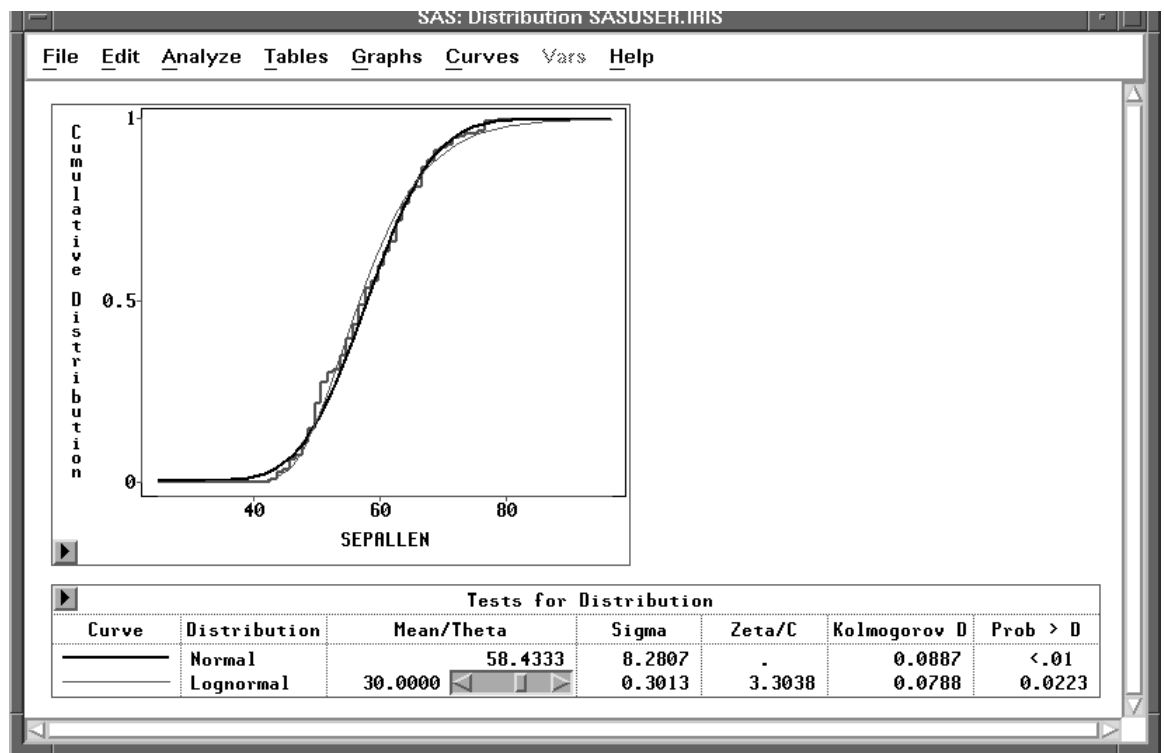


Figure 38.34. Tests for Distribution

QQ Ref Line

After choosing **Curves:QQ Ref Line**, you can use the **QQ Ref Line** dialog to add distribution reference lines to QQ plots.

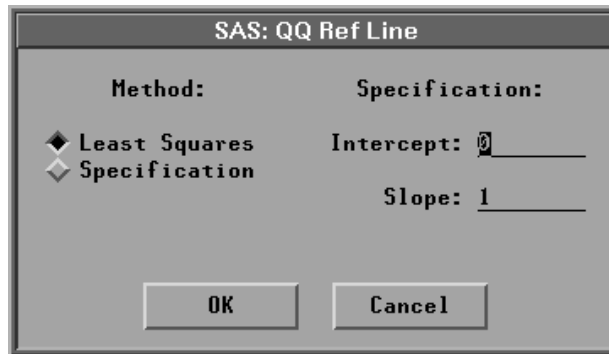


Figure 38.35. QQ Ref Line Dialog

The default adds a least squares regression line. You can also specify your own reference line by choosing **Method:Specification** and specifying both the intercept and slope.

If you select a **Weight** variable, you can add a weighted least squares regression line to the normal QQ plot. If the data are normally distributed with mean μ and standard deviation σ and if each observation has approximately the same weight (w_0), then the least squares regression line has approximately intercept μ and slope σ for vardef=WDF/WEIGHT and slope $\sigma/\sqrt{w_0}$ for vardef=DF/N.

A normal QQ plot with a least squares reference line is shown in [Figure 38.36](#). Use the sliders to change the intercept and slope of the reference line.

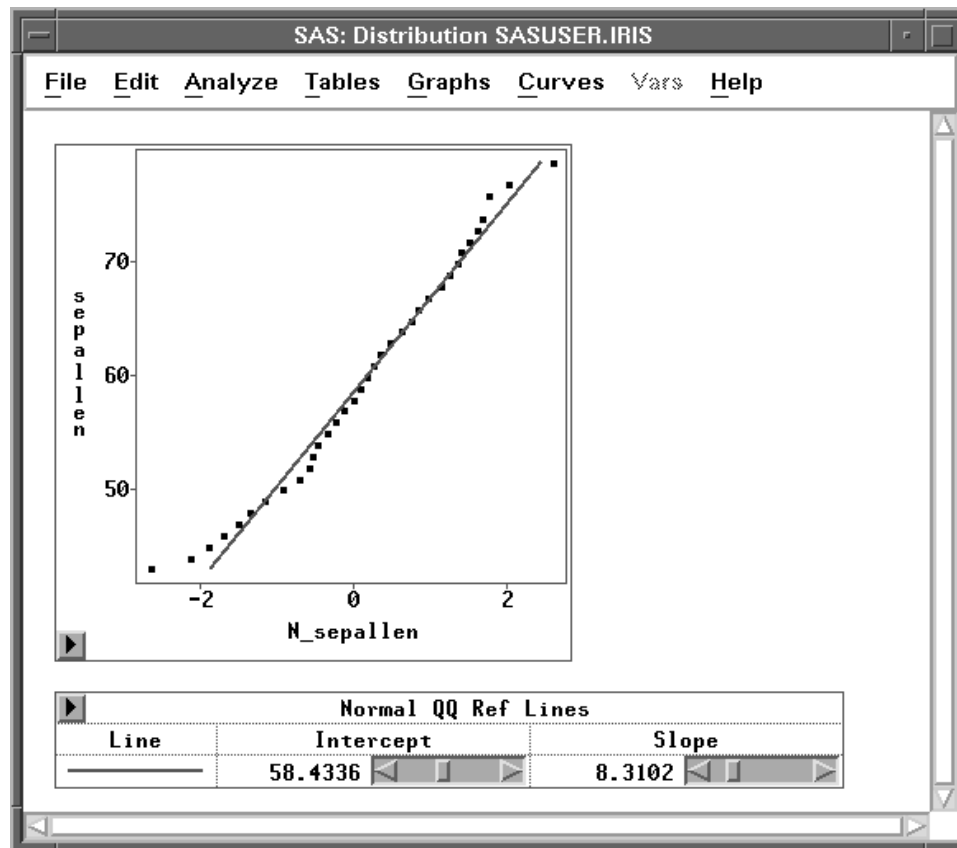


Figure 38.36. Normal QQ Plot with a Reference Line

Analysis for Nominal Variables

You can generate a frequency table, display a bar chart, and display a mosaic plot for each nominal variable in the distribution analysis, as shown in [Figure 38.37](#).

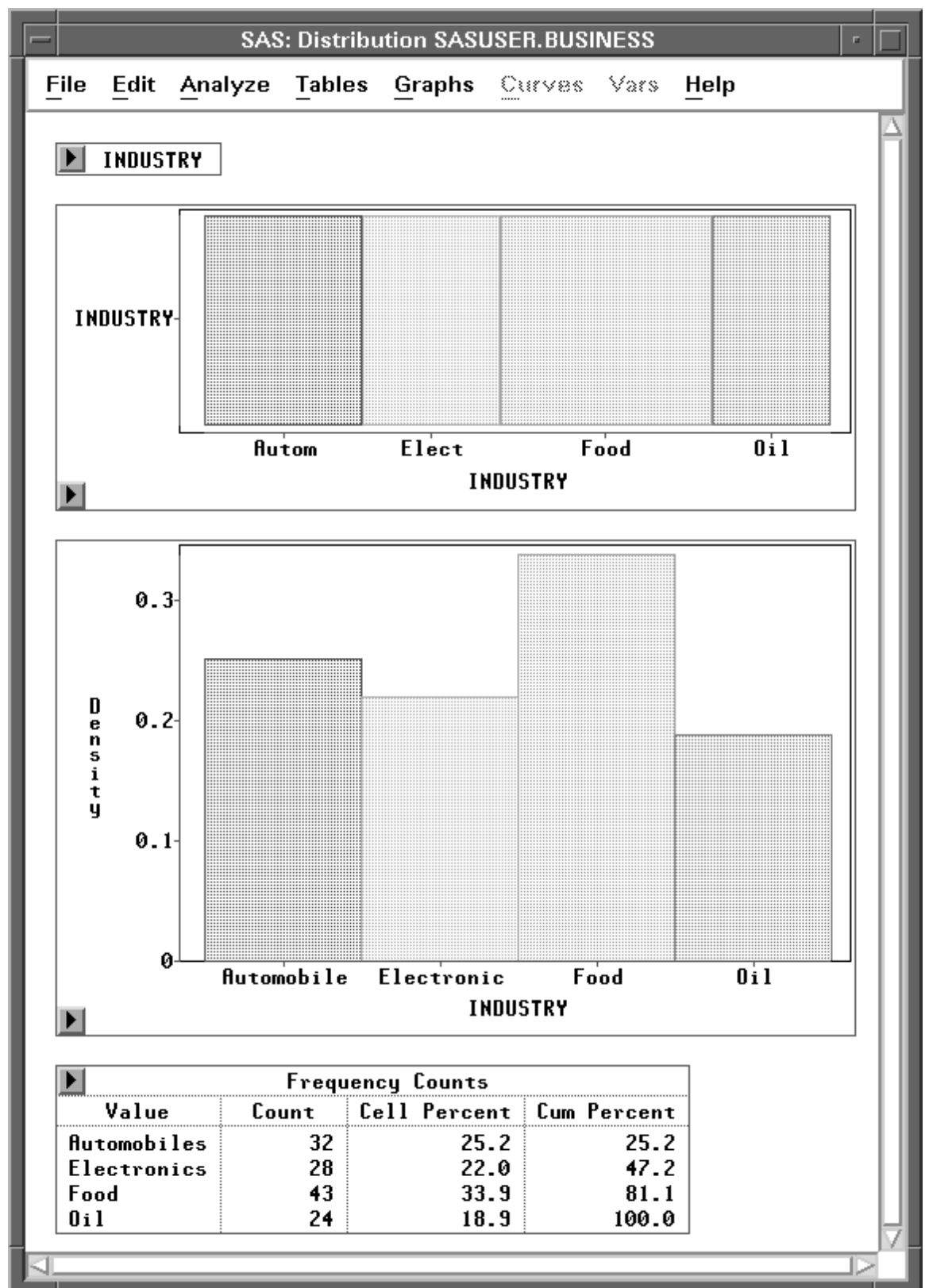


Figure 38.37. Nominal Variable Output

- ⊕ **Related Reading:** Bar Charts, [Chapter 32](#).
- ⊕ **Related Reading:** Mosaic Plots, [Chapter 33](#).

References

- Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth International Group.
- Chandra, M., Singpurwalla, N.D., and Stephens, M.A. (1981), "Kolmogorov Statistics for Tests of Fit for the Extreme-Value and Weibull Distributions," *Journal of the American Statistical Association*, 76, 729–731.
- Conover, W.J. (1980), *Practical Nonparametric Statistics*, Second Edition, New York: John Wiley & Sons, Inc.
- Croux, C. and Rousseeuw, P.J. (1992), "Time-Efficient Algorithms for Two Highly Robust Estimators of Scale," *Computational Statistics*, Volume 1, 411–428.
- D'Agostino, R.B. and Stephens, M.A., Eds. (1986), *Goodness-of-Fit Techniques*, New York: Marcel Dekker, Inc.
- Dixon, W.J. and Tukey, J.W. (1968), "Approximate Behavior of the Distribution of Winsorized t (Trimming/Winsorization 2)," *Technometrics*, 10, 83–98.
- Epanechnikov, V.A. (1969), "Nonparametric Estimation of a Multivariate Probability Density," *Theory of Probability and Its Applications*, 14, 153–158.
- Feller, W. (1948), "On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions," *Annals of Math. Stat.*, 19, 177–189.
- Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.
- Hahn, G.J. and Meeker, W.Q. (1991), *Statistical Intervals: A Guide for Practitioners*, New York: John Wiley & Sons, Inc.
- Hampel, F.R. (1974), "The Influence Curve and its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.
- Iman, R.L. (1974), "Use of a t -statistic as an Approximation to the Exact Distribution of the Wilcoxon Signed Ranks Test Statistic," *Communications in Statistics*, 3, 795–806.
- Johnson, N.L. and Kotz, S. (1970), *Continuous Univariate Distributions —I*, New York: John Wiley & Sons, Inc.
- Lehmann, E.L. (1975), *Nonparametric: Statistical Methods Based on Ranks*, San Francisco: Holden-Day, Inc.
- Rosenberger, J.L. and Gasko, M. (1983), "Comparing Location Estimators: Trimmed Means, Medians, and Trimean," in *Understanding Robust and Exploratory Data Analysis*, eds. D.C. Hoaglin, F. Mosteller, and J.W. Tukey, New York: John Wiley & Sons, Inc., 297–338.

- Rousseeuw, P.J. and Croux, C. (1993), “Alternatives to the Median Absolute Deviation,” *Journal of the American Statistical Association*, 88, 1273–1283.
- Royston, P. (1992), “Approximating the Shapiro-Wilk W-Test for non-normality,” *Statistics and Computing*, 2, 117–119.
- Silverman, B.W. (1982), “Kernel Density Estimation using the Fast Fourier Transform,” *Applied Statistics*, 31, 93–99.
- Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.
- Smirnov, N. (1948) “Table for Estimating the Goodness of Fit of Empirical Distributions,” *Annals of Math. Stat.*, 19, 279.
- Stephens, M.A. (1974), “EDF Statistics for Goodness of Fit and Some Comparisons,” *Journal of the American Statistical Association*, 69, 730–737.
- Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.
- Tukey, J.W. and McLaughlin, D.H. (1963), “Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1,” *Sankhya A*, 25, 331–352.

Chapter 39

Fit Analyses

Chapter Contents

STATISTICAL MODELS	614
LINEAR MODELS	615
GENERALIZED LINEAR MODELS	618
The Exponential Family of Distributions	618
Link Function	619
The Likelihood Function and Maximum-Likelihood Estimation	620
Scale Parameter	622
Goodness of Fit	622
Quasi-Likelihood Functions	623
NONPARAMETRIC SMOOTHERS	626
Smoother Degrees of Freedom	627
Smoother Generalized Cross Validation	628
VARIABLES	629
METHOD	631
OUTPUT	634
TABLES	638
Model Information	638
Model Equation	638
X'X Matrix	639
Summary of Fit for Linear Models	640
Summary of Fit for Generalized Linear Models	642
Analysis of Variance for Linear Models	643
Analysis of Deviance for Generalized Linear Models	644
Type I Tests	644
Type III Tests	646
Parameter Estimates for Linear Models	649
Parameter Estimates for Generalized Linear Models	651
C.I. for Parameters	652
Collinearity Diagnostics	657
Estimated COV Matrix and Estimated CORR Matrix	658
RESIDUAL AND SURFACE PLOTS	659

Residual-by-Predicted Plot	659
Residual Normal QQ Plot	661
Partial Leverage Plots	661
Parametric Surface Plot	662
Smoothing Spline Surface Plot	663
Kernel Surface Plot	667
Parametric Profile Surface Plot	670
FIT CURVES	671
Parametric Curves: Confidence Ellipses	671
Parametric Curves: Polynomial	674
Parametric Curves: Confidence Curves	677
Nonparametric Smoothing Spline	679
Nonparametric Kernel Smoother	682
Nonparametric Local Polynomial Smoother	684
OUTPUT VARIABLES	691
Hat Matrix Diagonal	692
Predicted Values	693
Linear Predictor	693
Residuals	693
Residual Normal Quantiles	693
Predicted Surfaces	694
Predicted Curves	695
Standardized and Studentized Residuals	696
Deviance Residuals	697
Pearson Residuals	697
Anscombe Residuals	698
Partial Leverage Variables	699
Cook's D	700
Dffits	700
Covratio	701
Dfbetas	701
WEIGHTED ANALYSES	702
REFERENCES	703

Chapter 39

Fit Analyses

Choosing **Analyze:Fit (Y X)** gives you access to a variety of techniques for fitting models to data. These provide methods for examining the relationship between a response (dependent) variable and a set of explanatory (independent) variables.

You can use least-squares methods for simple and multiple linear regression with various diagnostic capabilities when the response is normally distributed.

You can use generalized linear models to analyze the data when the response is from a distribution of the exponential family and a function can be used to link the response mean to a linear combination of the explanatory variables.

You can use spline and kernel smoothers for nonparametric regression when the model has one or two explanatory variables.

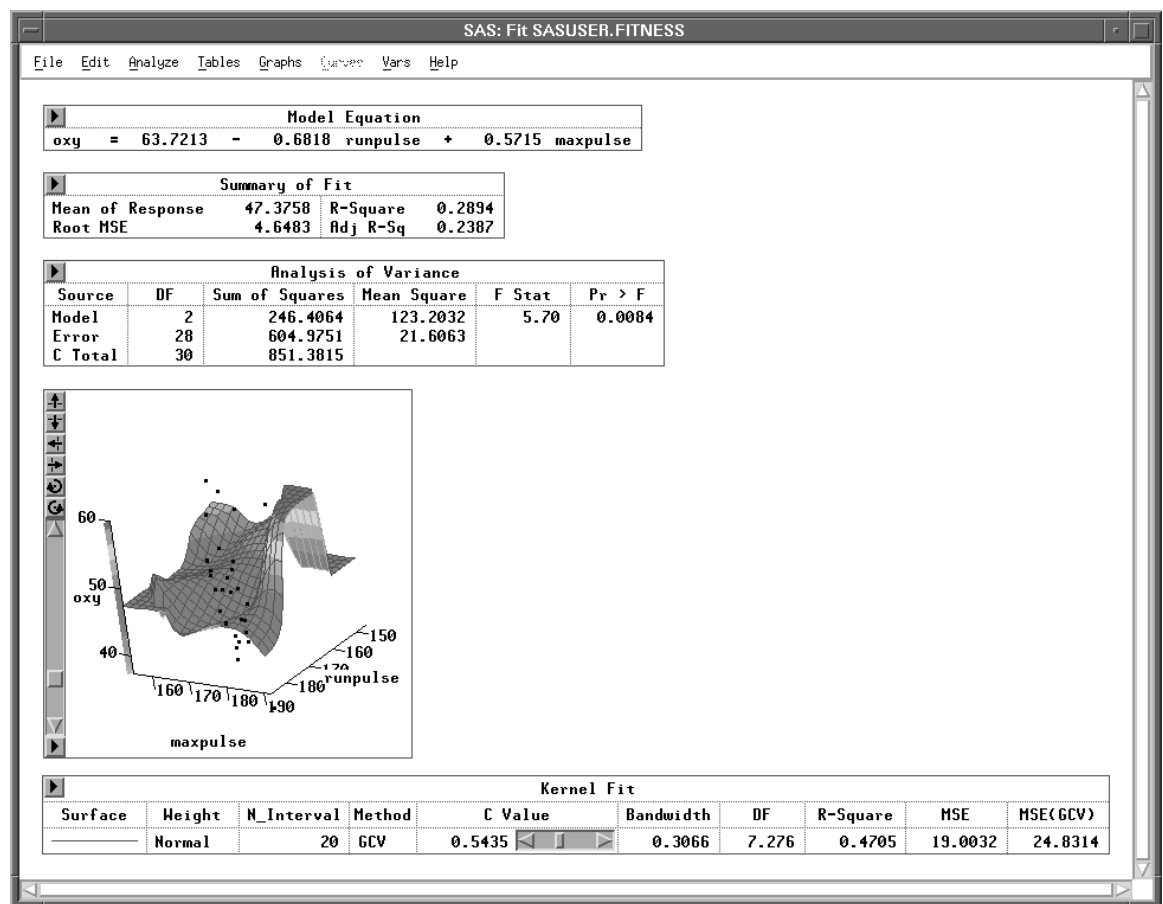


Figure 39.1. Fit Analysis

Statistical Models

The relationship between a response variable and a set of explanatory variables can be studied through a regression model

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

where y_i is the i th observed response value, \mathbf{x}_i is the i th vector of explanatory values, and ε_i 's are uncorrelated random variables with zero mean and a common variance.

If the form of the regression function f is known except for certain parameters, the model is called a *parametric regression model*. Furthermore, if the regression function is linear in the unknown parameters, the model is called a *linear model*.

In the case of linear models with the error term ε_i assumed to be normally distributed, you can use classical linear models to explore the relationship between the response variable and the explanatory variables.

A *nonparametric model* generally assumes only that f belongs to some infinite-dimensional collection of functions. For example, f may be assumed to be differentiable with a square-integrable second derivative.

When there is only one explanatory X variable, you can use nonparametric smoothing methods, such as smoothing splines, kernel estimators, and local polynomial smoothers. You can also request confidence ellipses and parametric fits (mean, linear regression, and polynomial curves) with a linear model. These are added to a scatter plot generated from \mathbf{Y} by a single \mathbf{X} and are described in the “[Fit Curves](#)” section.

When there are two explanatory variables in the model, you can create parametric and nonparametric (kernel and thin-plate smoothing spline) response surface plots. With more than two explanatory variables in the model, a parametric profile response surface plot with two selected explanatory variables can be created.

When the response y_i has a distribution from the exponential family (normal, inverse Gaussian, gamma, Poisson, binomial), and the mean μ_i of the response variable y_i is assumed to be related to a linear predictor through a monotone function g

$$g(\mu_i) = \mathbf{x}_i' \beta$$

where β is a vector of unknown parameters, you can explore the relationship by using generalized linear models.

Linear Models

SAS/INSIGHT fit analysis provides the traditional parametric regression analysis assuming that the regression function is linear in the unknown parameters. The relationship is expressed as an equation that predicts a response variable from a linear function of explanatory variables.

Besides the usual estimators and test statistics produced for a regression, a fit analysis can produce many diagnostic statistics. Collinearity diagnostics measure the strength of the linear relationship among explanatory variables and how this affects the stability of the estimates. Influence diagnostics measure how each individual observation contributes to determining the parameter estimates and the fitted values.

In matrix algebra notation, a linear model is written as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where \mathbf{y} is the $n \times 1$ vector of responses, \mathbf{X} is the $n \times p$ design matrix (rows are observations and columns are explanatory variables), β is the $p \times 1$ vector of unknown parameters, and ϵ is the $n \times 1$ vector of unknown errors.

Each effect in the model generates one or more columns in a design matrix \mathbf{X} . The first column of \mathbf{X} is usually a vector of 1's used to estimate the intercept term. In general, no-intercept models should be fit only when theoretical justification exists. Refer to the chapter on the GLM procedure in the *SAS/STAT User's Guide* for a description of the model parameterization.

The classical theory of linear models is based on some strict assumptions. Ideally, the response is measured with all the explanatory variables controlled in an experimentally determined environment. If the explanatory variables do not have experimentally fixed values but are stochastic, the conditional distribution of \mathbf{y} given \mathbf{X} must be normal in the appropriate form.

Less restrictive assumptions are as follows:

- The form of the model is correct (all important \mathbf{X} variables have been included).
- Explanatory variables are measured without error.
- The expected value of the errors is 0.
- The variance of the errors (and thus the response variable) is constant across observations (denoted by σ^2).
- The errors are uncorrelated across observations.

If all the necessary assumptions are met, the least-squares estimates of β are the best linear unbiased estimates (BLUE); in other words, the estimates have minimum variance among the class of estimators that are unbiased and are linear functions of the responses. In addition, when the error term is assumed to be normally distributed, sampling distributions for the computed statistics can be derived. These sampling distributions form the basis for hypothesis tests on the parameters.

Reference ♦ Fit Analyses

The method used to estimate the parameters is to minimize the sum of squares of the differences between the actual response values and the values predicted by the model. An estimator \mathbf{b} for β is generated by solving the resulting normal equations

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

yielding

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Let \mathbf{H} be the projection matrix for the space spanned by \mathbf{X} , sometimes called the hat matrix,

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Then the predicted mean vector of the n observation responses is

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{y}$$

The sum of squares for error is

$$SSE = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \sum_{i=1}^n (y_i - \mathbf{x}_i\mathbf{b})^2$$

where \mathbf{x}_i is the i th row of the \mathbf{X} matrix.

Assume that \mathbf{X} is of full rank. The variance σ^2 of the error is estimated by the mean square error

$$s^2 = MSE = \frac{SSE}{n - p}$$

The parameter estimates are unbiased:

$$E(\mathbf{b}) = \beta$$

$$E(s^2) = \sigma^2.$$

The covariance matrix of the estimates is

$$\text{Var}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

The estimate of the covariance matrix, $\widehat{\text{Var}}(\mathbf{b})$, is obtained by replacing σ^2 with its estimate, s^2 , in the preceding formula:

$$\widehat{\text{Var}}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}s^2$$

The correlations of the estimates,

$$\mathbf{S}^{-1/2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{S}^{-1/2}$$

are derived by scaling to one on the diagonal, where $\mathbf{S} = \text{diag} ((\mathbf{X}'\mathbf{X})^{-1})$.

If the model is not full rank, the matrix $\mathbf{X}'\mathbf{X}$ is singular. A generalized (g2) inverse (Pringle and Raynor 1971), denoted as $(\mathbf{X}'\mathbf{X})^-$, is then used to solve the normal equations, as follows:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{Y}$$

However, this solution is not unique, and there are an infinite number of solutions using different generalized inverses. In SAS/INSIGHT software, the fit analysis chooses a basis of all variables that are linearly independent of previous variables and a zero solution for the remaining variables.

⊕ **Related Reading:** Multiple Regression, [Chapter 14](#).

⊕ **Related Reading:** Analysis of Variance, [Chapter 15](#).

Generalized Linear Models

Generalized linear models assume that the response y_i has a distribution from the exponential family (normal, inverse Gaussian, gamma, Poisson, binomial) and a function can be used to link the expected response mean and a linear function of the \mathbf{X} effects. In SAS/INSIGHT software, a generalized linear model is written as

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

$$\eta = g(\boldsymbol{\mu}) = \boldsymbol{\eta}_0 + \mathbf{X}\boldsymbol{\beta}$$

where \mathbf{y} is the $n \times 1$ vector of responses, $\boldsymbol{\mu}$ is the $n \times 1$ expected response means, and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of unknown errors.

The monotone function g links the response mean $\boldsymbol{\mu}$ with a linear predictor η from the effects, and it is called the *link function*. The $n \times 1$ vector $\boldsymbol{\eta}_0$ is the offset, \mathbf{X} is the $n \times p$ design matrix, and $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown parameters. The design matrix is generated the same way as for linear models.

You specify the response distribution, the link function, and the offset variable in the fit method options dialog.

The Exponential Family of Distributions

The distribution of a random variable \mathbf{Y} belongs to the exponential family if its probability (density) function can be written in the form

$$f(y; \theta, \phi) = \exp \left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right)$$

where θ is the natural or canonical parameter, ϕ is the dispersion parameter, and a , b and c are specific functions.

The mean and variance of \mathbf{Y} are then given by (McCullagh and Nelder 1989)

$$E(y) = \mu = b'(\theta)$$

$$\text{Var}(y) = a(\phi)b''(\theta)$$

The function $b''(\theta)$ can be expressed as a function of μ , $b''(\theta) = V(\mu)$, and it is called the *variance function*. Different choices of the function $b(\theta)$ generate different distributions in the exponential family. For a binomial distribution with m trials, the function $a(\phi) = \phi/m$. For other distributions in the exponential family, $a(\phi) = \phi$.

SAS/INSIGHT software includes normal, inverse Gaussian, gamma, Poisson, and binomial distributions for the response distribution. For these response distributions, the density functions $f(y)$, the variance functions $V(\mu)$, and the dispersion parameters ϕ with function $a(\phi)$ are

Normal
$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) \quad \text{for } -\infty < y < \infty$$

$$V(\mu) = 1$$

$$a(\phi) = \phi = \sigma^2$$

Inverse Gaussian
$$f(y) = \frac{1}{\sqrt{2\pi y^3}\sigma} \exp\left(-\frac{1}{2\mu^2 y}\left(\frac{y-\mu}{\sigma}\right)^2\right) \quad \text{for } y > 0$$

$$V(\mu) = \mu^3$$

$$a(\phi) = \phi = \sigma^2$$

Gamma
$$f(y) = \frac{1}{y\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^\nu \exp\left(-\frac{\nu y}{\mu}\right) \quad \text{for } y > 0$$

$$V(\mu) = \mu^2$$

$$a(\phi) = \phi = \nu^{-1}$$

Poisson
$$f(y) = \frac{\mu^y e^{-\mu}}{y!} \quad \text{for } y = 0, 1, 2, \dots$$

$$V(\mu) = \mu$$

$$a(\phi) = \phi = 1$$

Binomial
$$f(y) = \binom{m}{r} \mu^r (1-\mu)^{m-r} \quad \text{for } y = r/m, r = 0, 1, 2, \dots, m$$

$$V(\mu) = \mu(1-\mu)$$

$$a(\phi) = \phi/m = 1/m$$

Link Function

The link function links the response mean μ to the linear predictor η . SAS/INSIGHT software provides six types of link functions:

Identity	$g(\mu) = \mu$
Log	$g(\mu) = \log(\mu)$
Logit	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
Probit	$g(\mu) = \Phi^{-1}(\mu)$
Comp. Log-log	$g(\mu) = \log(-\log(1-\mu))$
Power	$g(\mu) = \mu^\lambda$ where λ is the value in the Power entry field.

For each response distribution in the exponential family, there exists a special link function, the canonical link, for which $\theta = \eta$. The canonical links expressed in terms of the mean parameter μ are

Normal	$g(\mu) = \mu$
Inverse Gaussian	$g(\mu) = \mu^{-2}$
Gamma	$g(\mu) = \mu^{-1}$
Poisson	$g(\mu) = \log(\mu)$
Binomial	$g(\mu) = \log(\frac{\mu}{1-\mu})$

† **Note:** Some links are not appropriate for all distributions. For example, logit, probit, and complementary log-log links are only appropriate for the binomial distribution.

The Likelihood Function and Maximum-Likelihood Estimation

The log-likelihood function

$$l(\theta, \phi; y) = \log f(y; \theta, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

can be expressed in terms of the mean μ and the dispersion parameter ϕ :

Normal	$l(\mu, \phi; y) = -\frac{1}{2} \log(\phi) - \frac{1}{2\phi}(y - \mu)^2 \quad \text{for } -\infty < y < \infty$
---------------	---

Inverse Gaussian	$l(\mu, \phi; y) = -\log(y^3\phi) - \frac{(y-\mu)^2}{2y\mu^2\phi} \quad \text{for } y > 0$
-------------------------	--

Gamma	$l(\mu, \phi; y) = -\log(y\Gamma(\frac{1}{\phi})) + \frac{1}{\phi} \log(\frac{y}{\mu\phi}) - \frac{y}{\mu\phi} \quad \text{for } y > 0$
--------------	---

Poisson	$l(\mu, \phi; y) = y \log(\mu) - \mu \quad \text{for } y = 0, 1, 2, \dots$
----------------	--

Binomial	$l(\mu, \phi; y) = r \log(\mu) + (m - r) \log(1 - \mu)$ for $y = r/m, r = 0, 1, 2, \dots, m$
-----------------	---

† **Note:** Some terms in the density function have been dropped in the log-likelihood function since they do not affect the estimation of the mean and scale parameters.

SAS/INSIGHT software uses a ridge stabilized Newton-Raphson algorithm to maximize the log-likelihood function $l(\mu, \phi; y)$ with respect to the regression parameters. On the r th iteration, the algorithm updates the parameter vector \mathbf{b} by

$$\mathbf{b}_{(r)} = \mathbf{b}_{(r-1)} - \mathbf{H}_{(r-1)}^{-1} \mathbf{u}_{(r-1)}$$

where \mathbf{H} is the Hessian matrix and \mathbf{u} is the gradient vector, both evaluated at $\beta = \mathbf{b}_{(r-1)}$.

$$\mathbf{H} = (h_{jk}) = \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right)$$

$$\mathbf{u} = (u_j) = \left(\frac{\partial l}{\partial \beta_j} \right).$$

The Hessian matrix \mathbf{H} can be expressed as

$$\mathbf{H} = -\mathbf{X}'\mathbf{W}_o\mathbf{X}$$

where \mathbf{X} is the design matrix, \mathbf{W}_o is a diagonal matrix with i th diagonal element

$$w_{oi} = w_{ei} + (y_i - \mu_i) \frac{V_i g_i'' + V_i' g_i'}{V_i^2 (g_i')^3 a_i(\phi)}$$

$$w_{ei} = E(w_{oi}) = \frac{1}{a_i(\phi) V_i (g_i')^2}$$

where g_i is the link function, V_i is the variance function, and the primes denote derivatives of g and V with respect to μ . All values are evaluated at the current mean estimate μ_i . $a_i(\phi) = \phi/w_i$, where w_i is the prior weight for the i th observation.

SAS/INSIGHT software uses either the full Hessian matrix $\mathbf{H} = -\mathbf{X}'\mathbf{W}_o\mathbf{X}$ or the Fisher's scoring method in the maximum-likelihood estimation. In the Fisher's scoring method, \mathbf{W}_o is replaced by its expected value \mathbf{W}_e with i th element w_{ei} .

$$\mathbf{H} = \mathbf{X}'\mathbf{W}_e\mathbf{X}$$

The estimated variance-covariance matrix of the parameter estimates is

$$\hat{\Sigma} = -\mathbf{H}^{-1}$$

where \mathbf{H} is the Hessian matrix evaluated at the model parameter estimates.

The estimated correlation matrix of the parameter estimates is derived by scaling the estimated variance-covariance matrix to 1 on the diagonal.

† **Note:** A warning message appears when the specified model fails to converge. The output tables, graphs, and variables are based on the results from the last iteration.

Scale Parameter

A scale parameter is related to the dispersion parameter ϕ and is given by

Normal	$\sigma = \sqrt{\phi}$
Inverse Gaussian	$\sigma = \sqrt{\phi}$
Gamma	$\nu = 1/\phi$
Poisson	1
Binomial	1

The scale parameter is 1 for Poisson and binomial distributions. SAS/INSIGHT software provides different scale parameter estimates for normal, inverse Gaussian, and gamma distributions:

MLE	the maximum-likelihood estimate
Deviance	the mean deviance
Pearson	the mean Pearson χ^2
Constant	the value in the Constant entry field

When maximum-likelihood estimation is used, the Hessian **H** and the gradient **u** also include the term for the scale parameter.

† **Note:** You can request an exponential distribution for the response variable by specifying a gamma distribution with scale parameter set to 1.

Goodness of Fit

The log-likelihood can be expressed in terms of the mean parameter μ and the log-likelihood-ratio statistic is the scaled deviance

$$D^*(y; \hat{\mu}) = -2(l(\hat{\mu}; y) - l(\hat{\mu}_{max}; y))$$

where $l(\hat{\mu}; y)$ is the log-likelihood under the model and $l(\hat{\mu}_{max}; y)$ is the log-likelihood under the maximum achievable (saturated) model.

For generalized linear models, the scaled deviance can be expressed as

$$D^*(y; \hat{\mu}) = \frac{1}{\phi} D(y; \hat{\mu})$$

where $D(y; \hat{\mu})$ is the residual deviance for the model and is the sum of individual deviance contributions.

The forms of the individual deviance contributions, d_i , are

Normal	$(y - \hat{\mu})^2$
Inverse Gaussian	$(y - \hat{\mu})^2 / (\hat{\mu}^2 y)$
Gamma	$-2 \log(y / \hat{\mu}) + 2(y - \hat{\mu}) / \hat{\mu}$
Poisson	$2y \log(y / \hat{\mu}) - 2(y - \hat{\mu})$
Binomial	$2(r \log(y / \hat{\mu}) + (m - r) \log((1 - y) / (1 - \hat{\mu})))$
where $y = r/m$, r is the number of successes in m trials.	

For a binomial distribution with m_i trials in the i th observation, the Pearson χ^2 statistic is

$$\chi^2 = \sum_{i=1}^n m_i \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

For other distributions, the Pearson χ^2 statistic is

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

The scaled Pearson χ^2 statistic is χ^2 / ϕ . Either the mean deviance $D(y; \hat{\mu}) / (n - p)$ or the mean Pearson χ^2 statistic $\chi^2 / (n - p)$ can be used to estimate the dispersion parameter ϕ . The χ^2 approximation is usually quite accurate for the differences of deviances for nested models (McCullagh and Nelder 1989).

Quasi-Likelihood Functions

For binomial and Poisson distributions, the scale parameter has a value of 1. The variance of \mathbf{Y} is $\text{Var}(y) = \mu(1 - \mu)/m$ for the binomial distribution and $\text{Var}(y) = \mu$ for the Poisson distribution. *Overdispersion* occurs when the variance of \mathbf{Y} exceeds the $\text{Var}(y)$ above. That is, the variance of \mathbf{Y} is $\sigma^2 V(\mu)$, where $\sigma > 1$.

With overdispersion, methods based on quasi-likelihood can be used to estimate the parameters β and σ . A quasi-likelihood function

$$Q(\mu; y) = \int_y^\mu \frac{y - t}{\sigma^2 V(t)} dt$$

is specified by its associated variance function.

SAS/INSIGHT software includes the quasi-likelihoods associated with the variance functions $V(\mu) = 1, \mu, \mu^2, \mu^3$, and $\mu(1 - \mu)$. The associated distributions (with the same variance function), the quasi-likelihoods $Q(\mu; y)$, the canonical links $g(\mu)$, and the scale parameters σ and ν for these variance functions are

$V(\mu) = 1$	<p>Normal</p> $\sigma^2 Q(\mu; y) = -\frac{1}{2}(y - \mu)^2 \quad \text{for } -\infty < y < \infty$ $g(\mu) = \mu$ $\sigma = \sqrt{\phi}$
$V(\mu) = \mu$	<p>Poisson</p> $\sigma^2 Q(\mu; y) = y \log(\mu) - \mu \quad \text{for } \mu > 0, y \geq 0$ $g(\mu) = \log \mu$ $\sigma = \sqrt{\phi}$
$V(\mu) = \mu^2$	<p>Gamma</p> $\sigma^2 Q(\mu; y) = -y/\mu - \log(\mu) \quad \text{for } \mu > 0, y \geq 0$ $g(\mu) = \mu^{-1}$ $\nu = \phi^{-1}$
$V(\mu) = \mu^3$	<p>Inverse Gaussian</p> $\sigma^2 Q(\mu; y) = -y/(2\mu^2) + 1/\mu \quad \text{for } \mu > 0, y \geq 0$ $g(\mu) = \mu^{-2}$ $\sigma = \sqrt{\phi}$
$V(\mu) = \mu(1 - \mu)$	<p>Binomial</p> $\sigma^2 Q(\mu; y) = r \log(\mu) + (m - r) \log(1 - \mu)$ <p>for $0 < \mu < 1, y = r/m, r = 0, 1, 2, \dots, m$</p> $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ $\sigma = \sqrt{\phi}$

SAS/INSIGHT software uses the mean deviance, the mean Pearson χ^2 , or the value in the **Constant** entry field to estimate the dispersion parameter ϕ . The conventional estimate of ϕ is the mean Pearson χ^2 statistic.

Maximum quasi-likelihood estimation is similar to ordinary maximum-likelihood estimation and has the same parameter estimates as the distribution with the same variance function. These estimates are not affected by the dispersion parameter ϕ , but ϕ is used in the variance-covariance matrix of the parameter estimates. However, the likelihood-ratio based statistics, such as **Type I (LR)**, **Type III (LR)**, and **C.I.(LR) for Parameters** tables, are not produced in the analysis.

⊕ **Related Reading:** Logistic Regression, [Chapter 16](#).

⊕ **Related Reading:** Poisson Regression, [Chapter 17](#).

Nonparametric Smoothers

For a simple regression model with one or two explanatory variables,

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

a *smoother* $\hat{f}_\lambda(\mathbf{x})$ is a function that summarizes the trend of \mathbf{Y} as a function of \mathbf{X} . It can enhance the visual perception of either a \mathbf{Y} -by- \mathbf{X} scatter plot or a rotating plot. The smoothing parameter λ controls the smoothness of the estimate.

With one explanatory variable in the model, $\hat{f}_\lambda(\mathbf{x})$ is called a *scatter plot smoother*. SAS/INSIGHT software provides nonparametric curve estimates from smoothing spline, kernel, loess (nearest neighbors local polynomial), and fixed bandwidth local polynomial smoothers.

For smoothing spline, kernel, and fixed bandwidth local polynomial smoothers, SAS/INSIGHT software derives the smoothing parameter λ from a constant c that is independent of the units of \mathbf{X} . For a loess smoother, the smoothing parameter λ is a positive constant α .

With two explanatory variables in the model, $\hat{f}_\lambda(\mathbf{x})$ is called a *surface smoother*. SAS/INSIGHT software provides nonparametric surface estimates from thin-plate smoothing spline and kernel smoothers. The explanatory variables are scaled by their corresponding sample interquartile ranges. The smoothing parameter λ is derived from a constant c and both are independent of the units of \mathbf{X} .

Similar to parametric regression, the R^2 value for an estimate is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{f}_\lambda(\mathbf{x}_i))^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

You can use the following methods to choose the λ value:

DF	uses the λ value that makes the resulting smoothing estimate have the specified degrees of freedom (df).
GCV	uses the λ value that minimizes the generalized cross validation (GCV) mean squared error.
C Value	uses the λ value derived from the specified c value for nonparametric smoothers other than the loess smoother.
Alpha	uses the specified α value for the loess estimator.

If you specify a **DF** value for a smoother, an iterative procedure is used to find the estimate with the specified df . You can choose a convergence criterion γ based on either the relative difference or the absolute difference. A smoother satisfying the following conditions is then created:

$$\frac{|df(\text{fitted}) - df(\text{specified})|}{df(\text{specified})} < \gamma \quad \text{for relative difference}$$

$$|df(\text{fitted}) - df(\text{specified})| < \gamma \quad \text{for absolute difference}$$

Smoother Degrees of Freedom

For a nonparametric smoother with a parameter λ , the fitted values can be written as

$$\hat{\mathbf{y}} = \mathbf{H}_\lambda \mathbf{y}$$

where \mathbf{y} is the $n \times 1$ vector of observed responses y_i , $\hat{\mathbf{y}}$ is the $n \times 1$ vector of fitted values $\hat{y}_i = \hat{f}_\lambda(x_i)$, and the smoother matrix \mathbf{H}_λ is an $n \times n$ matrix that depends on the value of λ .

The degrees of freedom, or the effective number of parameters, of a smoother can be used to compare different smoothers and to describe the flexibility of the smoother. SAS/INSIGHT software defines the degrees of freedom of a smoother as

$$df_\lambda = \text{trace}(\mathbf{H}_\lambda)$$

which is the sum of the diagonal elements of \mathbf{H}_λ .

† **Note:** Two other popular definitions of degrees of freedom for a smoother are $\text{trace}(\mathbf{H}_\lambda \mathbf{H}'_\lambda)$ and $\text{trace}(2\mathbf{H}_\lambda - \mathbf{H}_\lambda \mathbf{H}'_\lambda)$ (Hastie and Tibshirani 1990).

Smoother Generalized Cross Validation

With the degrees of freedom of an estimate df_λ , the mean squared error is given as

$$\text{MSE}(\lambda) = \frac{1}{n - df_\lambda} \sum_{i=1}^n (y_i - \hat{f}_\lambda(\mathbf{x}_i))^2$$

Cross-validation (CV) estimates the response at each x_i from the smoother that uses only the remaining $n - 1$ observations. The resulting cross validation mean squared error is

$$\text{MSE}_{\text{CV}}(\lambda) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}_{\lambda(i)}(\mathbf{x}_i))^2$$

where $\hat{f}_{\lambda(i)}(\mathbf{x}_i)$ is the fitted value at x_i computed without the i th observation.

The cross validation mean squared error can also be written as

$$\text{MSE}_{\text{CV}}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{f}_\lambda(\mathbf{x}_i)}{1 - h_{\lambda i}} \right)^2$$

where $h_{\lambda i}$ is the i th diagonal element of the \mathbf{H}_λ matrix (Hastie and Tibshirani 1990).

Generalized cross validation replaces $h_{\lambda i}$ by its average value, $\frac{1}{n} df_\lambda$. The generalized cross validation mean squared error is

$$\text{MSE}_{\text{GCV}}(\lambda) = \frac{1}{n(1 - df_\lambda/n)^2} \sum_{i=1}^n (y_i - \hat{f}_\lambda(\mathbf{x}_i))^2$$

† **Note:** The function $\text{MSE}_{\text{GCV}}(\lambda)$ may have multiple minima, so the value estimated by SAS/INSIGHT software may be only a local minimum, not the global minimum.

Variables

To create a fit analysis, choose **Analyze:Fit (Y X)**. If you have already selected one or more variables, the first variable selected is the response or dependent variable, and it is assigned the **Y** variable role. The remaining variables are explanatory or independent variables, and they are assigned the **X** variable role. If you do not select any **X** effects, a model with only an intercept term (mean) is fit.

If you have not selected any variables, a variables dialog appears.

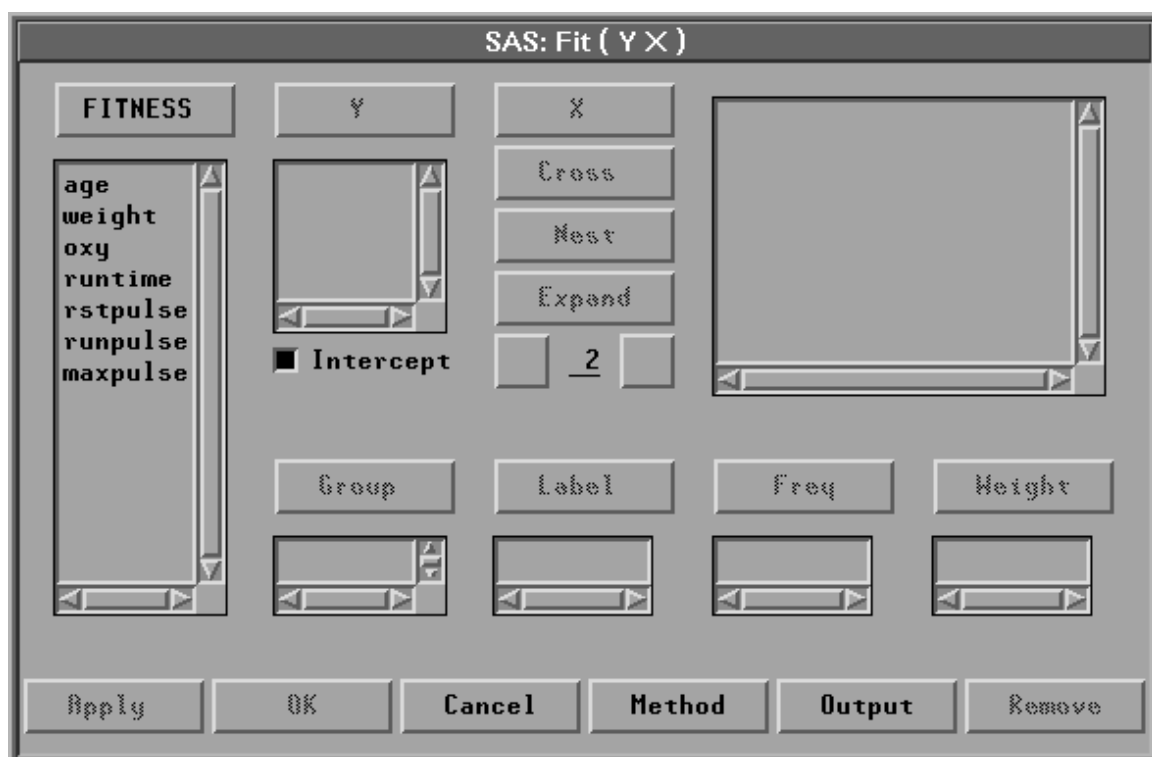


Figure 39.2. Fit Variables Dialog

In the dialog, select one **Y** variable for each fit analysis. Create **X** effects in the model by using the **X**, **Cross**, **Nest**, and **Expand** buttons. An *effect* is a variable or combination of variables that constitutes a term in the model. There are four ways to specify effects in SAS/INSIGHT software. In the following discussion, assume that **X1** and **X2** are interval variables and **A** and **B** are nominal variables.

You can use the **X** button to create regressor effects of the interval variables and main effects of the nominal variables. Select any variable, then click the **X** button. For example, selecting **A** and then clicking the **X** button adds **A** to the effects list.

You can use the **Cross** button to create crossed effects. These include polynomial effects of the interval variables, interactions of the nominal variables, and interaction effects of interval and nominal variables. Select two or more variables, then click

the **Cross** button. For example, selecting **X1** and **X2** and then clicking the **Cross** button generates the crossed effect **X1*X2**.

You can use the **Nest** button to create nested effects. In a nested effect, a variable or crossed effect is nested within the effects of one or more nominal variables. Select a variable or crossed effect and one or more nominal variables, then click the **Nest** button. For example, selecting **X1*X2**, **A**, and **B** and then clicking the **Nest** button generates the nested effect **X1*X2(A B)**.

You can use the **Expand** button and the associated entry field to create expanded effects. These include response surface effects for interval variables and factorial effects for nominal variables. The **Expand** button expands all possible effects to the degree of expansion specified in the entry field below the **Expand** button. The value **2** is the default degree of expansion. You can click the right button of the entry field to increase the expansion degree by 1 or the left button to decrease it by 1.

Choose the degree of expansion, then select variables or effects and click the **Expand** button. For example, with degree of expansion 2 and variables **A** and **B** selected, clicking the **Expand** button generates three effects

A B A*B

With degree of expansion 2 and variables **X1** and **X2** selected, clicking the **Expand** button generates five effects

X1 X2 X1*X1 X1*X2 X2*X2

Intercept is checked by default to include the intercept term in the model. As a general rule, no-intercept models should be fit only when theoretical justification exists.

You can select one or more **Group** variables if you have grouped data. This creates a fit analysis for each group.

You can select a **Label** variable to label observations in the plots.

You can select a **Freq** variable. If you select a **Freq** variable, each observation is assumed to represent n observations, where n is the value of the **Freq** variable.

You can select a **Weight** variable to assign relative weights for each observation in the analysis. The details of weighted analyses are explained in the “[Weighted Analyses](#)” section at the end of this chapter.

The fit variables dialog provides an **Apply** button. The **Apply** button displays the fit window without closing the fit variables dialog. This makes it easy to modify the model by adding or removing variables. Each time you modify the model using the **Apply** button, a *new* fit window is displayed so you can easily compare models. The **OK** button also displays a new fit window but closes the dialog.

Method

Observations with missing values for **Y**, **X**, **Weight**, or **Freq** variables are not used. Observations with nonpositive **Weight** or **Freq** values are not used. Only the integer part of **Freq** values is used.

To view or change the response distribution and link function, click the **Method** button in the variables dialog. This displays the dialog shown in [Figure 39.3](#).



Figure 39.3. Fit Method Options Dialog

You can choose the response distribution and link function of the **Y** variables. If you choose a binomial distribution, specify either

- a **Y** variable with values 1 or 0 indicating success or failure
- a **Y** variable giving the number of successes in a certain number of trials, and a **Binomial** variable to give the corresponding number of trials

If you choose a power link function, specify the power value in the **Power** entry field.

If you select an **Offset** variable, it is treated as an **X** variable with coefficient fixed at 1.0.

You can choose the scale parameter for the response distribution. If you choose a **Constant** scale, specify the constant value in the **Constant** entry field.

Reference ♦ *Fit Analyses*

With overdispersion in the model, you can specify the **Quasi-Likelihood** option to fit the generalized linear model using the quasi-likelihood functions.

If you choose a normal response distribution with a canonical link (identity for normal distributions), you can specify the **Exact Distribution** option to fit the linear model using the usual exact distributions for the test statistics.

You can specify the **Fisher's Scoring** option to use the Fisher's scoring method in the maximum-likelihood estimation for the regression parameters.

By default, SAS/INSIGHT software uses the **Normal** response distribution and **Canonical** link with the **Exact Distribution** option to perform a fit analysis for the linear model.

Output

To view or change the options associated with your fit analysis, click the **Output** button in the variables dialog. This displays the output options dialog shown in [Figure 39.4](#).

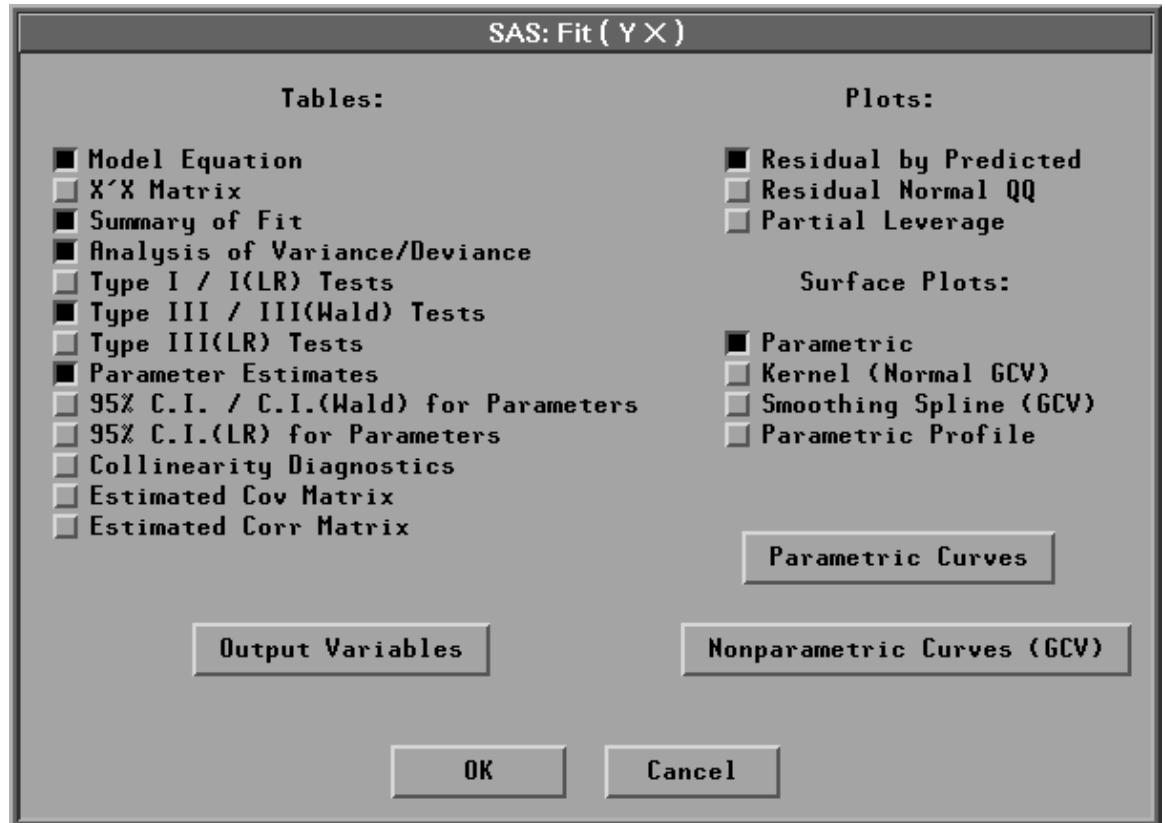


Figure 39.4. Fit Output Options Dialog

The options you set in this dialog determine the tables and graphs that appear in the fit window. Provided by default are tables of the model equation, summary of fit, analysis of variance or deviance, type III or type III (Wald) tests, and parameter estimates and a plot of residuals by predicted values.

When there are two explanatory variables in the model, a parametric response surface plot is created by default. You can also generate a nonparametric kernel or a thin-plate smoothing spline response surface plot. With more than two explanatory variables in the model, a parametric profile response surface plot with the first two explanatory variables can be created. The values of the remaining explanatory variables are set to their corresponding means in the plot. You can use the sliders to change these values of the remaining explanatory variables.

Click on the **Output Variables** button in the fit dialog to display the **Output Variables** dialog shown in Figure 39.5. The **Output Variables** dialog enables you to specify variables that can be saved in the data window. Output variables include predicted values and several influence diagnostic variables based on the model you fit.

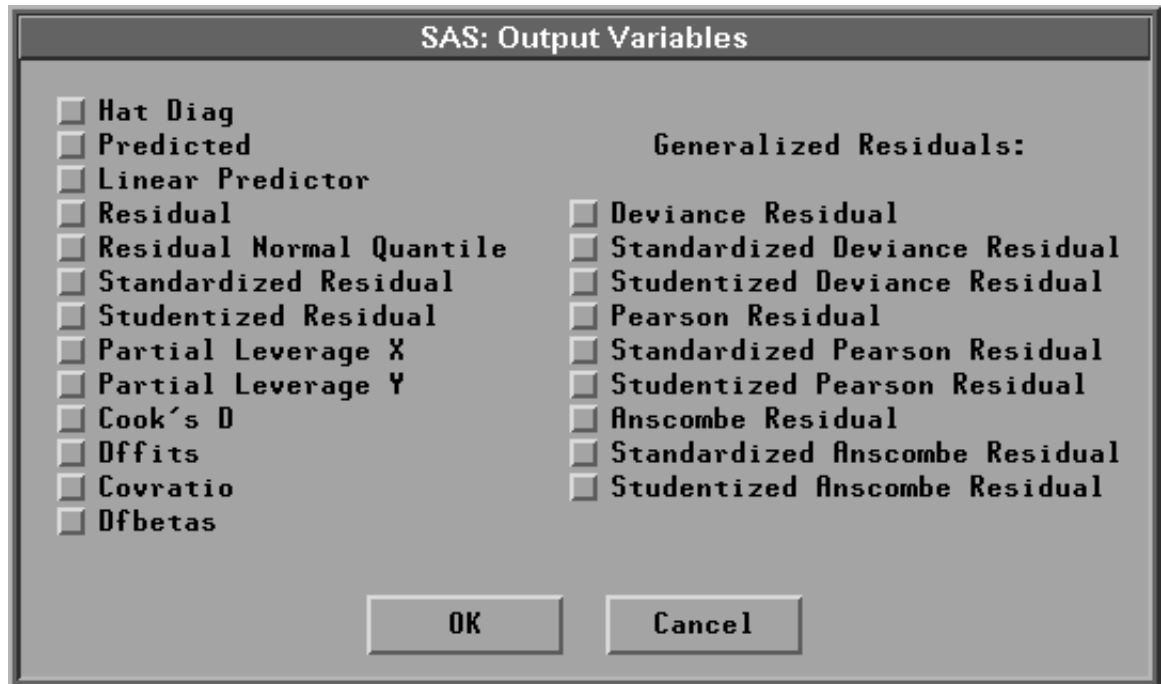


Figure 39.5. Output Variables Dialog

When there is only one explanatory variable in the model, a **Y-by-X** scatter plot is generated. The **Parametric Curves** and **Nonparametric Curves (GCV)** buttons display dialogs that enable you to fit parametric and nonparametric curves to this scatter plot.

Click on **Parametric Curves** to display the **Parametric Curves** dialog.

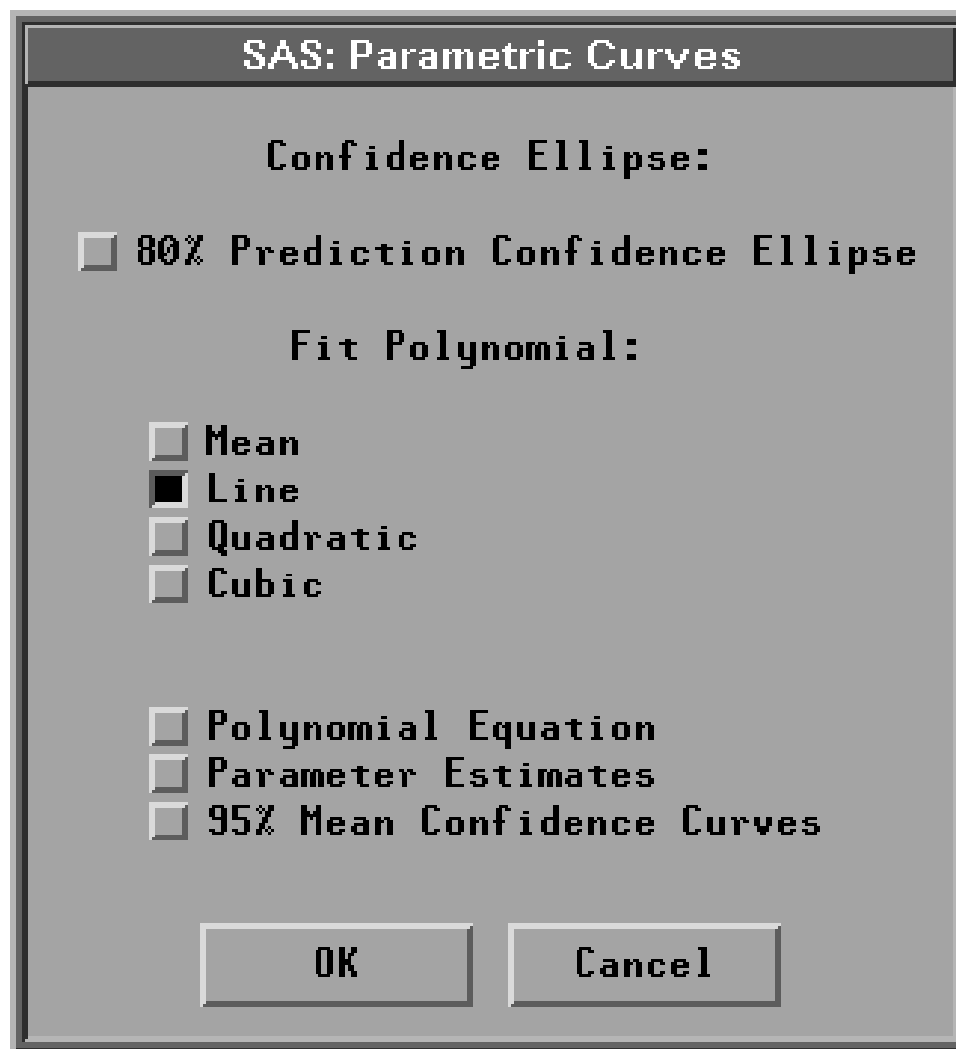
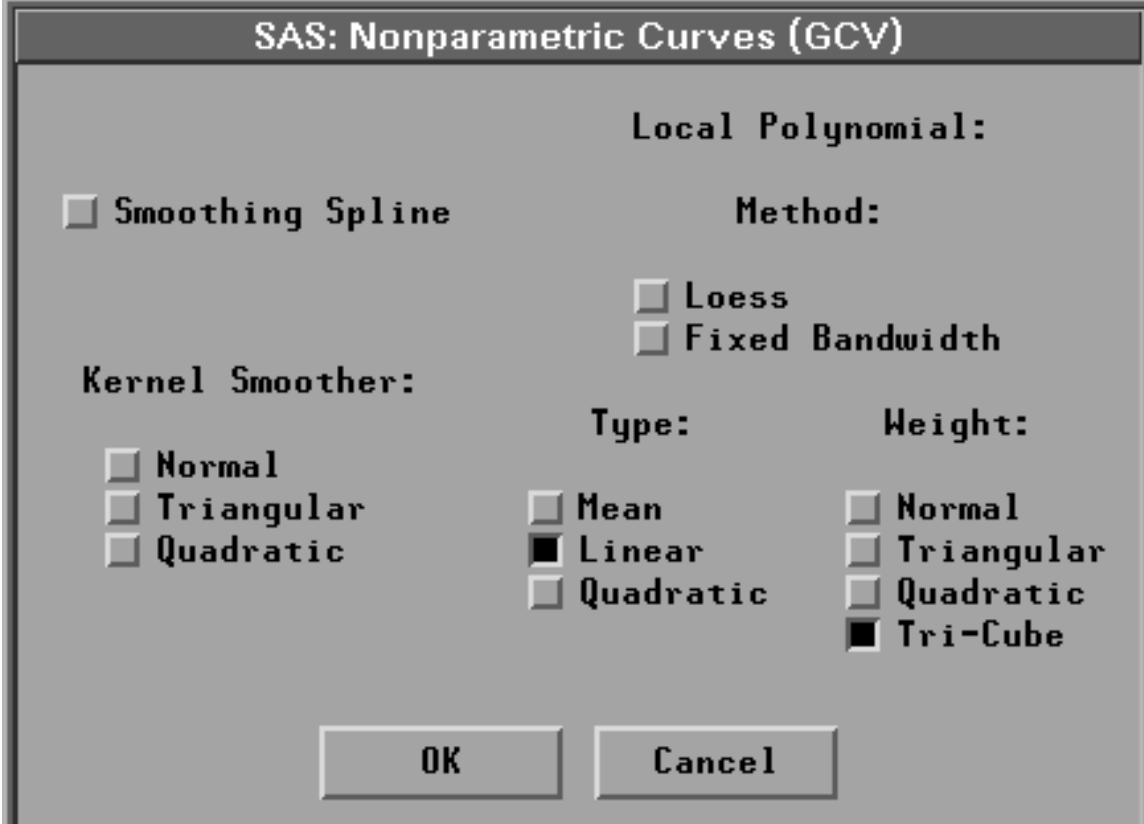


Figure 39.6. Parametric Curves Dialog

A regression line fit is provided by default. You can request an 80% prediction ellipse and other polynomial fits in the dialog. You can also request polynomial equation tables, parameter estimates tables, and 95% mean confidence curves for fitted polynomials.

The **Nonparametric Curves (GCV)** dialog in [Figure 39.7](#) includes a smoothing spline, a kernel smoother, and a local polynomial smoother. You must specify the method, regression type, and weight function for a local polynomial fit.



The dialog box is titled "SAS: Nonparametric Curves (GCV)". It contains three main sections for selecting smoothing methods:

- Smoothing Spline:** A single checkbox labeled "Smoothing Spline".
- Kernel Smoother:** Three checkboxes labeled "Normal", "Triangular", and "Quadratic".
- Local Polynomial:** This section includes:
 - Method:** Two checkboxes labeled "Loess" and "Fixed Bandwidth".
 - Type:** Three checkboxes labeled "Mean", "Linear" (which is selected with a black square), and "Quadratic".
 - Weight:** Four checkboxes labeled "Normal", "Triangular", "Quadratic", and "Tri-Cube" (which is selected with a black square).

At the bottom of the dialog are two buttons: "OK" and "Cancel".

Figure 39.7. Nonparametric Curves Dialog

Tables

You can generate tables that present the results of a model fit and diagnostics for assessing how well the model fits the data. Set options in the output dialog as described in the “Output” section or choose from the **Tables** menu.

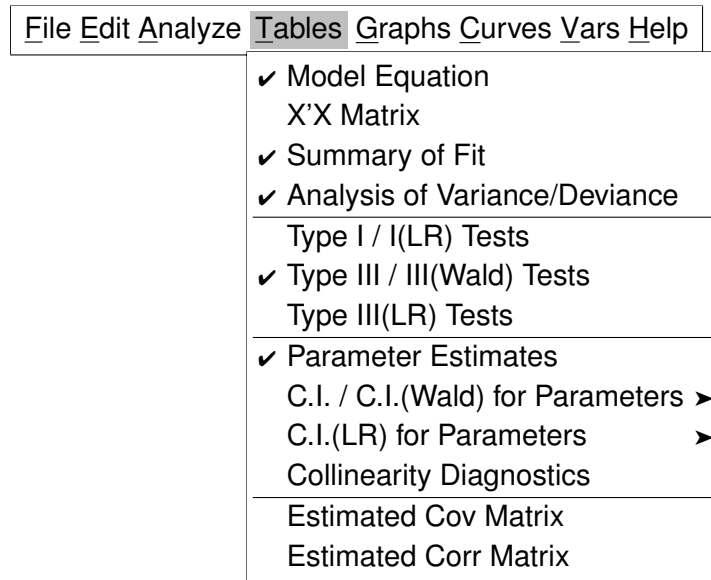


Figure 39.8. Tables Menu

Model Information

The first table in the fit analysis contains the model specification, the response distribution, and the link function, as illustrated in [Figure 39.9](#).

When the model contains nominal variables in its effects, the levels of the nominal variables are displayed in the **Nominal Variable Information** table, as shown in [Figure 39.9](#). The levels are determined from the formatted values of the nominal variables. An additional **Parameter Information** table, as illustrated in [Figure 39.9](#), shows the variable indices for the parameters in the model equation, the $X'X$ matrix, the estimated covariance matrix, and the estimated correlation matrix.

Model Equation

The model equation table gives the fitted equation for the model. [Figure 39.9](#) shows an equation for a model with nominal variables, and [Chapter 39](#) shows an equation for a model without nominal variables

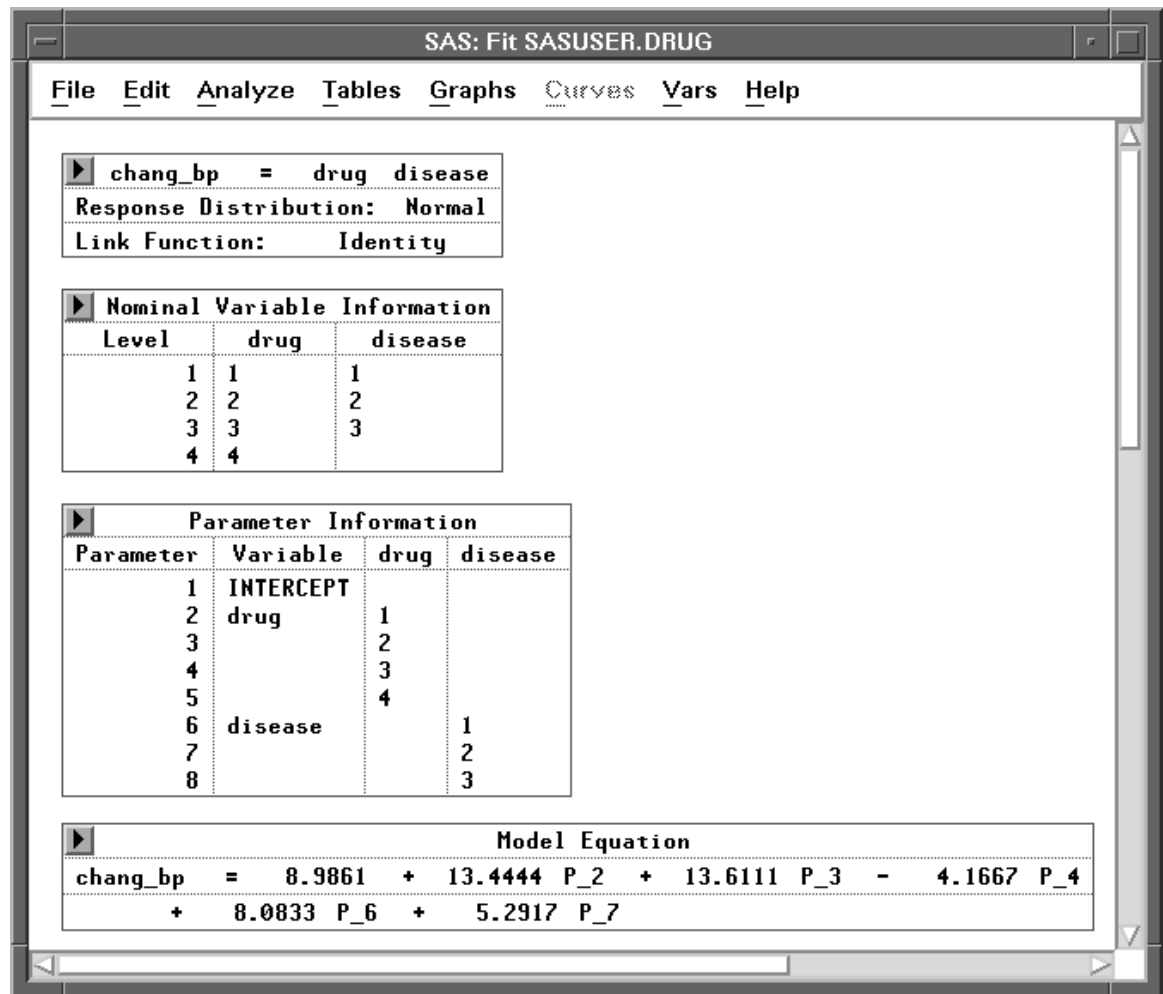


Figure 39.9. Model Information Tables

X'X Matrix

The X'X matrix table, as illustrated by Figure 39.10, contains the X'X crossproducts matrix for the model.

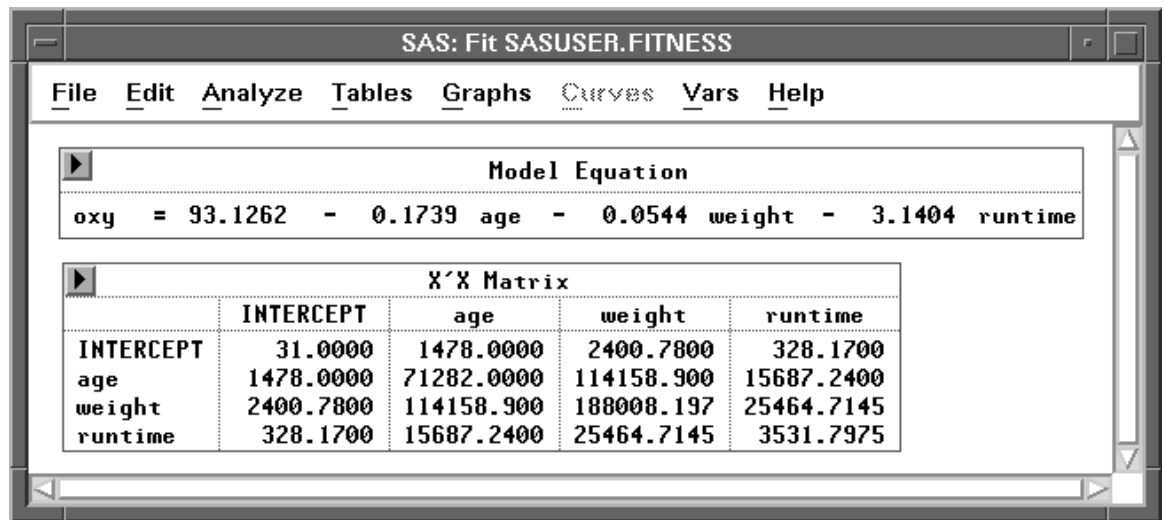


Figure 39.10. X'X Matrix for Linear Models

Summary of Fit for Linear Models

The **Summary of Fit** table for linear models, shown in Figure 39.11, includes the following:

Mean of Response	is the sample mean, \bar{y} , of the response variable.
Root MSE	is the estimate of the standard deviation of the error term. It is calculated as the square root of the mean square error.
R-Square	R^2 , with values between 0 and 1, indicates the proportion of the (corrected) total variation attributed to the fit.
Adj R-Sq	An adjusted R^2 is a version of R^2 that has been adjusted for degrees of freedom.

The screenshot shows a SAS window titled "SAS: Fit SASUSER.FITNESS". The menu bar includes File, Edit, Analyze, Tables, Graphs, Output, Vars, and Help. The main content area displays two tables. The first table, "Summary of Fit", shows the Mean of Response as 47.3758, Root MSE as 2.6882, R-Square as 0.7708, and Adj R-Sq as 0.7454. The second table, "Analysis of Variance", shows the following values: Model (DF=3, Sum of Squares=656.2709, Mean Square=218.7570, F Stat=30.27, Pr > F<.0001), Error (DF=27, Sum of Squares=195.1106, Mean Square=7.2263), and C Total (DF=30, Sum of Squares=851.3815).

Mean of Response	47.3758	R-Square	0.7708
Root MSE	2.6882	Adj R-Sq	0.7454

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	3	656.2709	218.7570	30.27	<.0001
Error	27	195.1106	7.2263		
C Total	30	851.3815			

Figure 39.11. Summary of Fit, Analysis of Variance Tables for Linear Models

With an intercept term in the model, R^2 is defined as

$$R^2 = 1 - (SSE/CSS)$$

where $CSS = \sum_{i=1}^n (y_i - \bar{y})^2$ is the corrected sum of squares and $SSE = \sum_{i=1}^n (y_i - \hat{y})^2$ is the sum of squares for error.

The R^2 statistic is also the square of the multiple correlation, that is, the square of the correlation between the response variable and the predicted values.

The adjusted R^2 statistic, an alternative to R^2 , is adjusted for the degrees of freedom of the sums of squares associated with R^2 . It is calculated as

$$AdjR^2 = 1 - \frac{SSE/(n-p)}{CSS/(n-1)} = 1 - \frac{n-1}{n-p}(1 - R^2)$$

Without an intercept term in the model, R^2 is defined as

$$R^2 = 1 - (SSE/TSS)$$

where $TSS = \sum_{i=1}^n y_i^2$ is the uncorrected total sum of squares.

The adjusted R^2 statistic is then calculated as

$$AdjR^2 = 1 - \frac{SSE/(n-p)}{TSS/n} = 1 - \frac{n}{n-p}(1 - R^2)$$

† **Note:** Other definitions of R^2 exist for models with no intercept term. Care should be taken to ensure that this is the definition desired.

Summary of Fit for Generalized Linear Models

For generalized linear models, the **Summary of Fit** table, as illustrated by [Figure 39.12](#), includes the following:

Mean of Response	is the sample mean, \bar{y} , of the response variable.
SCALE	is the constant scale parameter specified in the method dialog or a value of 1.0 for maximum-likelihood estimation for Poisson or binomial distributions.
SCALE (MLE)	is the maximum-likelihood estimate of the scale parameter for normal, gamma, and inverse Gaussian distributions.
SCALE (Deviance)	is the scale parameter estimated by the mean error deviance.
SCALE (Pearson)	is the scale parameter estimated by the mean Pearson χ^2 .
Deviance	is the error deviance.
Deviance/DF	is the mean error deviance, the error deviance divided by its associated degrees of freedom.
Pearson ChiSq	is the Pearson χ^2 statistic.
Pearson ChiSq / DF	is the mean Pearson χ^2 , the Pearson χ^2 divided by its associated degrees of freedom.

When the scale parameter is a constant specified in the method dialog, or when the response has a Poisson or binomial distribution, the table also contains the scaled deviance and the scaled Pearson χ^2 :

Scaled Dev	is the error deviance divided by the dispersion parameter.
Scaled ChiSq	is the Pearson χ^2 divided by the dispersion parameter.

The screenshot shows the SAS: Fit SASUSER.PATIENT window. The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. The main area displays the following information:

remiss = cell smear infil li blast temp
 Response Distribution: Binomial
 Link Function: Logit

Summary of Fit

Mean of Response	0.3333	Deviance	21.7507	Pearson ChiSq	19.4781
SCALE	1.0000	Deviance / DF	1.0875	Pearson ChiSq / DF	0.9739
		Scaled Dev	21.7507	Scaled ChiSq	19.4781

Analysis of Deviance

Source	DF	Deviance	Deviance / DF	Scaled Dev	Pr > Scaled Dev
Model	6	12.6211	2.1035	12.6211	0.0495
Error	20	21.7507	1.0875	21.7507	
C Total	26	34.3718			

Figure 39.12. Summary of Fit and Analysis of Deviance Tables for Generalized Linear Models

Analysis of Variance for Linear Models

The **Analysis of Variance** table for linear models, shown in [Figure 39.11](#), includes the following:

Source	indicates the source of the variation. Sources include Model for the fitted regression and Error for the residual error. C Total is the sum of the Model and Error components, and it is the total variation after correcting for the mean. When the model does not have an intercept term, the uncorrected total variation (U Total) is displayed.
DF	is the degrees of freedom associated with each source of variation.
Sum of Squares	is the sum of squares for each source of variation.
Mean Square	is the sum of squares divided by its associated degrees of freedom.
F Stat	is the F statistic for testing the null hypothesis that all parameters are 0 except for the intercept. This is formed by dividing the mean square for model by the mean square for error.
Pr > F	is the probability of obtaining a greater F statistic than that observed if the null hypothesis is true. This quantity is also called a p -value. A small p -value is evidence for rejecting the null hypothesis.

Analysis of Deviance for Generalized Linear Models

The **Analysis of Deviance** table for generalized linear models, as illustrated by Figure 39.12, includes the following:

Source	indicates the source of the variation. Sources include Model for the fitted regression and Error for the residual error. C Total is the sum of the Model and Error components, and it is the total variation after correcting for the mean. When the model does not have an intercept term, the uncorrected total variation (U Total) is printed.
DF	is the degrees of freedom associated with each source of variation.
Deviance	is the deviance for each source of variation.
Deviance/DF	is the deviance divided by its associated degrees of freedom.

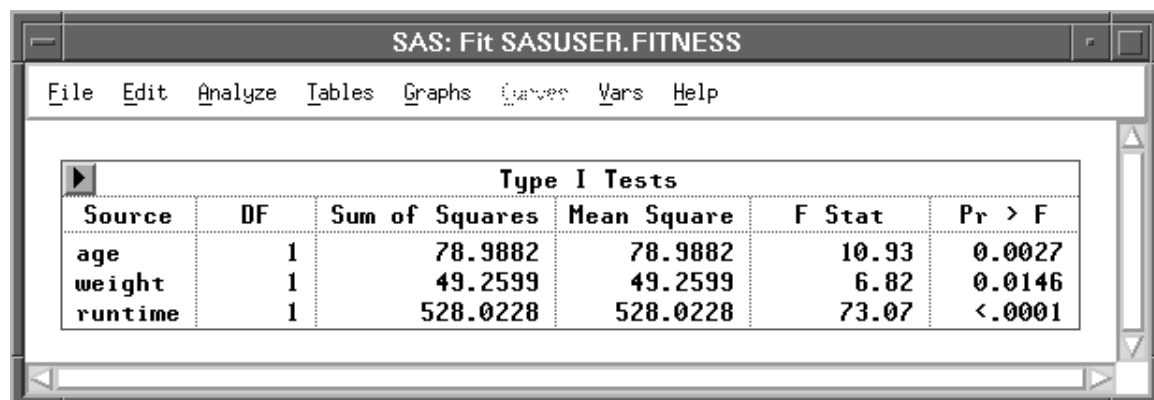
When the scale parameter is a constant specified in the method dialog, or when the response has a Poisson or binomial distribution, the table also contains the following:

Scaled Dev	is the deviance divided by the dispersion parameter.
Pr>Scaled Dev	is the probability of obtaining a greater scaled deviance statistic than that observed if the null hypothesis is true. Under the null hypothesis, all parameters are 0 except for the intercept, and the scaled deviance has an approximate χ^2 distribution.

Type I Tests

Type I tests examine the sequential incremental improvement in the fit of the model as each effect is added. They can be computed by fitting the model in steps and recording the difference in error sum of squares (linear models) and log-likelihood statistics (generalized linear models). The **Type I Tests** table for linear models, as illustrated by Figure 39.13, includes the following:

Source	is the name for each effect.
DF	is the degrees of freedom associated with each effect.
Sum of Squares	is the incremental error sum of squares for the model as each effect is added.
Mean Square	is the sum of squares divided by its associated degrees of freedom.
F Stat	is the F statistic for testing the null hypothesis that the parameters for the added effect are 0. This is formed by dividing the mean square for the effect by the mean square for error from the complete model.
Pr > F	is the probability of obtaining a greater F statistic than that observed if the null hypothesis is true.



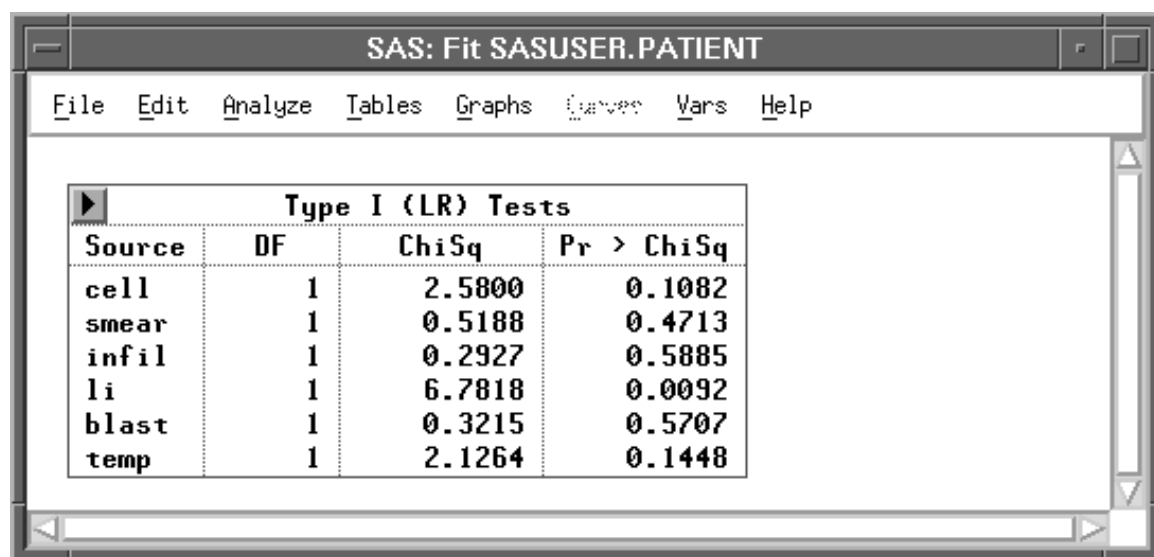
The screenshot shows a SAS window titled "SAS: Fit SASUSER.FITNESS". The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. The main content area displays a table titled "Type I Tests".

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
age	1	78.9882	78.9882	10.93	0.0027
weight	1	49.2599	49.2599	6.82	0.0146
runtime	1	528.0228	528.0228	73.07	<.0001

Figure 39.13. Type I Tests Table

The **Type I (LR) Tests** table for generalized linear models, as illustrated by [Figure 39.14](#), includes the following:

Source	is the name for each effect.
DF	is the degrees of freedom associated with each effect.
ChiSq	is the χ^2 value for testing the null hypothesis that the parameters for the added effect are 0. This is evaluated as twice the incremental log-likelihood for the model as each effect is added, and it has an asymptotic χ^2 distribution under the null hypothesis.
Pr > ChiSq	is the probability of obtaining a greater χ^2 statistic than that observed, if the null hypothesis is true.



The screenshot shows a SAS window titled "SAS: Fit SASUSER.PATIENT". The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. The main content area displays a table titled "Type I (LR) Tests".

Source	DF	ChiSq	Pr > ChiSq
cell	1	2.5800	0.1082
smear	1	0.5188	0.4713
infil	1	0.2927	0.5885
li	1	6.7818	0.0092
blast	1	0.3215	0.5707
temp	1	2.1264	0.1448

Figure 39.14. Type I Likelihood Ratio Tests

Type III Tests

Type III tests examine the significance of each partial effect, that is, the significance of an effect with all the other effects in the model. They are computed by constructing a type III hypothesis matrix \mathbf{L} and then computing statistics associated with the hypothesis $\mathbf{L}\beta = 0$. Refer to the chapter titled “The Four Types of Estimable Functions,” in the *SAS/STAT User’s Guide* for the construction of the matrix \mathbf{L} .

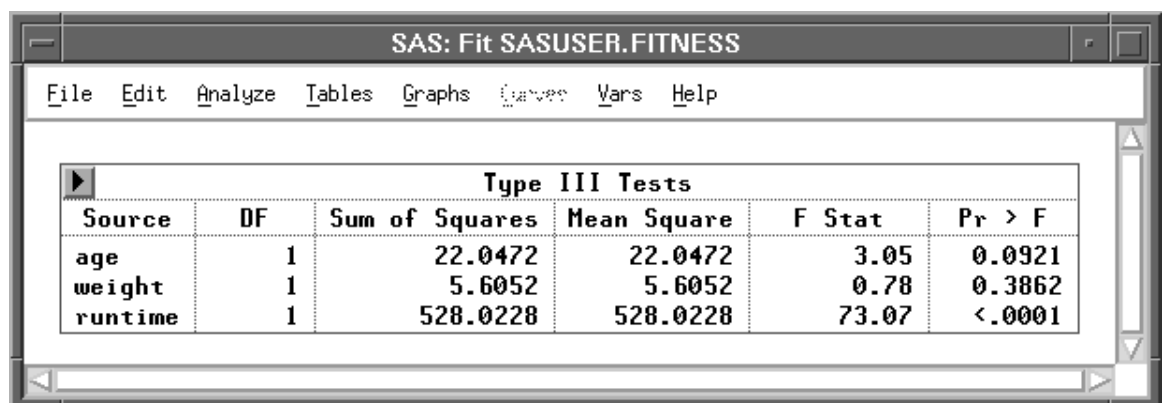
For linear models, the type III or partial sum of squares

$$(\mathbf{Lb})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{Lb})$$

is used to test the hypothesis $\mathbf{L}\beta = 0$.

The **Type III Tests** table for linear models, as illustrated by Figure 39.15, includes the following:

Source	is the name for each effect.
DF	is the degrees of freedom associated with each effect.
Sum of Squares	is the partial sum of squares for each effect in the model.
Mean Square	is the sum of squares divided by its associated degrees of freedom.
F Stat	is the F statistic for testing the null hypothesis that the linear combinations of parameters described previously for the hypothesis matrix \mathbf{L} are 0. This is formed by dividing the mean square for the hypothesis matrix \mathbf{L} by the mean square for error from the complete model.
Pr > F	is the probability of obtaining a greater F statistic than that observed if the null hypothesis is true.



The screenshot shows the SAS window titled "SAS: Fit SASUSER.FITNESS". The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. A table titled "Type III Tests" is displayed, showing the following data:

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
age	1	22.0472	22.0472	3.05	0.0921
weight	1	5.6052	5.6052	0.78	0.3862
runtime	1	528.0228	528.0228	73.07	<.0001

Figure 39.15. Type III Tests Table for Linear Models

For generalized linear models, either the Wald statistic or the likelihood-ratio statistic can be used to test the hypothesis $\mathbf{L}\beta = 0$. For the linear model, the two tests are equivalent.

The Wald statistic is given by

$$(\mathbf{Lb})'(\mathbf{L}\widehat{\text{Var}}(\mathbf{b})\mathbf{L}')^{-1}(\mathbf{Lb})$$

where $\widehat{\text{Var}}(\mathbf{b})$ is the estimated covariance matrix of the parameters. The likelihood-ratio statistic is computed as twice the difference between the maximum log-likelihood achievable under the unconstrained model and the maximum log-likelihood for the model under the restriction or constraint $\mathbf{L}\beta = 0$. Both the Wald statistic and the likelihood-ratio statistic have an asymptotic χ^2 distribution.

The **Type III (Wald) Tests** and **Type III (LR) Tests** tables, as illustrated by [Figure 39.16](#), include the following:

Source	is the name for each effect.
DF	is the degrees of freedom associated with each effect.
ChiSq	is the Wald statistic for the Wald tests or the likelihood-ratio statistic for the LR tests of the null hypothesis that the parameters for the effect are 0. This has an asymptotic χ^2 distribution.
Pr > ChiSq	is the probability of obtaining a greater χ^2 statistic than that observed, if the null hypothesis is true.

The screenshot shows the SAS window titled "SAS: Fit SASUSER.PATIENT". The menu bar includes File, Edit, Analyze, Tables, Graphs, Output, Vars, and Help. Two tables are displayed:

Source	DF	ChiSq	Pr > ChiSq
cell	1	0.2658	0.6062
smear	1	0.1108	0.7392
infil	1	0.1010	0.7507
li	1	2.7789	0.0955
blast	1	0.0044	0.9471
temp	1	1.6742	0.1957

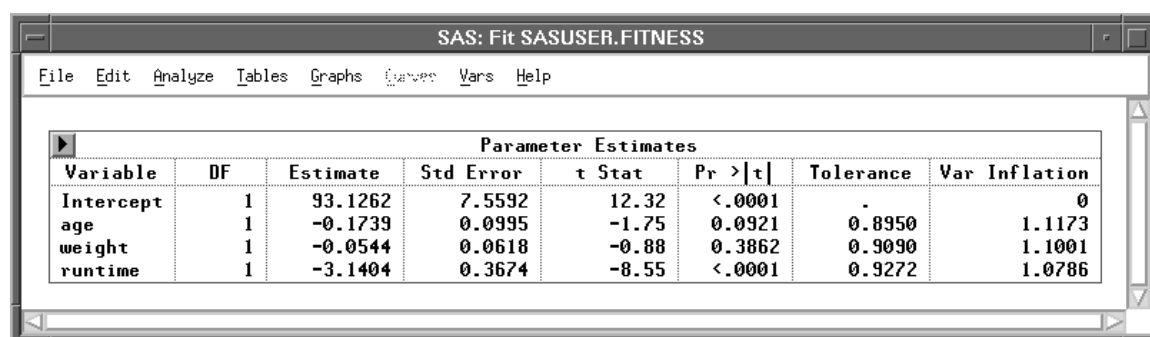
Source	DF	ChiSq	Pr > ChiSq
cell	1	0.3202	0.5715
smear	1	0.1180	0.7312
infil	1	0.1065	0.7442
li	1	4.3446	0.0371
blast	1	0.0044	0.9472
temp	1	2.1264	0.1448

Figure 39.16. Type III Tests Tables for Generalized Linear Models

Parameter Estimates for Linear Models

The **Parameter Estimates** table for linear models, as illustrated by Figure 39.17, includes the following:

Variable	names the variable associated with the estimated parameter. The name INTERCEPT represents the estimate of the intercept parameter.
DF	is the degrees of freedom associated with each parameter estimate. There is one degree of freedom unless the model is not of full rank. In this case, any parameter whose definition is confounded with previous parameters in the model has its degrees of freedom set to 0.
Estimate	is the parameter estimate.
Std Error	is the standard error, the estimate of the standard deviation of the parameter estimate.
t Stat	is the t statistic for testing that the parameter is 0. This is computed as the parameter estimate divided by the standard error.
Pr > t 	is the probability of obtaining (by chance alone) a t statistic greater in absolute value than that observed given that the true parameter is 0. This is referred to as a two-sided p -value. A small p -value is evidence for concluding that the parameter is not 0.
Tolerance	is the tolerance of the explanatory variable on the other variables.
Var Inflation	is the variance inflation factor of the explanatory variable.



The screenshot shows the SAS window titled "SAS: Fit SASUSER.FITNESS". The menu bar includes File, Edit, Analyze, Tables, Graphs, Output, Vars, and Help. The main display area shows the "Parameter Estimates" table with the following data:

Variable	DF	Estimate	Std Error	t Stat	Pr > t	Tolerance	Var Inflation
Intercept	1	93.1262	7.5592	12.32	<.0001	.	0
age	1	-0.1739	0.0995	-1.75	0.0921	0.8950	1.1173
weight	1	-0.0544	0.0618	-0.88	0.3862	0.9090	1.1001
runtime	1	-3.1404	0.3674	-8.55	<.0001	0.9272	1.0786

Figure 39.17. Parameter Estimates Table for Linear Models

The standard error of the j th parameter estimate b_j is computed using the equation

$$\text{STDERR}(b_j) = \sqrt{(\mathbf{X}'\mathbf{X})_{jj}^{-1}s^2}$$

where $(\mathbf{X}'\mathbf{X})_{jj}^{-1}$ is the j th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

Under the hypothesis that β_j is 0, the ratio

$$t = \frac{b_j}{\text{STDERR}(b_j)}$$

is distributed as Student's t with degrees of freedom equal to the degrees of freedom for the mean square error.

When an explanatory variable is nearly a linear combination of other explanatory variables in the model, the affected estimates are unstable and have high standard errors. This problem is called *collinearity* or *multicollinearity*. A fit analysis provides several methods for detecting collinearity.

Tolerances (TOL) and *variance inflation factors (VIF)* measure the strength of interrelationships among the explanatory variables in the model. Tolerance is $1 - R^2$ for the R^2 that results from the regression of the explanatory variable on the other explanatory variables in the model. Variance inflation factors are diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$ after $\mathbf{X}'\mathbf{X}$ is scaled to correlation form. The variance inflation measures the inflation in the variance of the parameter estimate due to collinearity between the explanatory variable and other variables. These measures are related by $VIF = 1 / TOL$.

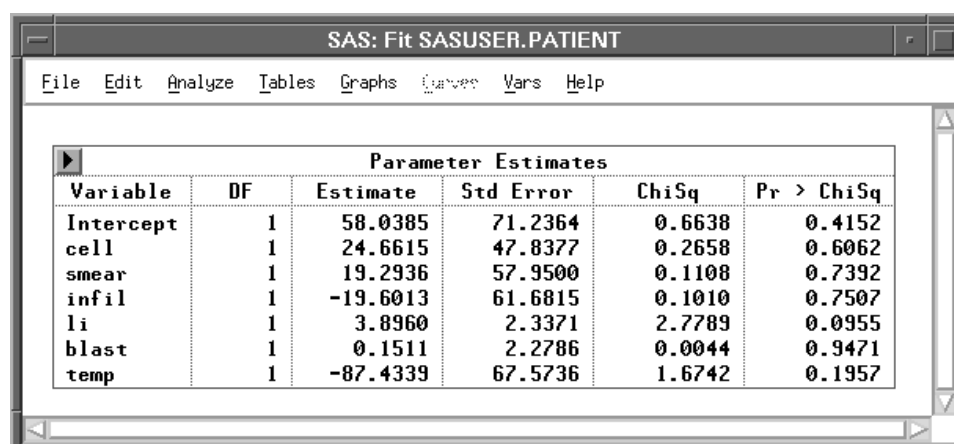
If all variables are orthogonal to each other, both tolerance and variance inflation are 1. If a variable is closely related to other variables, the tolerance goes to 0 and the variance inflation becomes large.

When the $\mathbf{X}'\mathbf{X}$ matrix is singular, least-squares solutions for the parameters are not unique. An estimate is 0 if the variable is a linear combination of previous explanatory variables. The degrees of freedom for the zeroed estimates are reported as 0. The hypotheses that are not testable have t tests printed as missing.

Parameter Estimates for Generalized Linear Models

The **Parameter Estimates** table for generalized linear models, as illustrated by Figure 39.18, includes the following:

Variable	names the variable associated with the estimated parameter. The name INTERCEPT represents the estimate of the intercept parameter.
DF	is the degrees of freedom associated with each parameter estimate. There is one degree of freedom unless the model is not full rank. In this case, any parameter that is confounded with previous parameters in the model has its degrees of freedom set to 0.
Estimate	is the parameter estimate.
Std Error	is the estimated standard deviation of the parameter estimate.
ChiSq	is the χ^2 test statistic for testing that the parameter is 0. This is computed as the square of the ratio of the parameter estimate divided by the standard error.
Pr > ChiSq	is the probability of obtaining an χ^2 statistic greater than that observed given that the true parameter is 0. A small p -value is evidence for concluding that the parameter is not 0.



The screenshot shows a SAS window titled "SAS: Fit SASUSER.PATIENT". The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. The main display area shows a table titled "Parameter Estimates" with the following data:

Variable	DF	Estimate	Std Error	ChiSq	Pr > ChiSq
Intercept	1	58.0385	71.2364	0.6638	0.4152
cell	1	24.6615	47.8377	0.2658	0.6062
smear	1	19.2936	57.9500	0.1108	0.7392
infil	1	-19.6013	61.6815	0.1010	0.7507
li	1	3.8960	2.3371	2.7789	0.0955
blast	1	0.1511	2.2786	0.0044	0.9471
temp	1	-87.4339	67.5736	1.6742	0.1957

Figure 39.18. Parameter Estimates Table for Generalized Linear Models

C.I. for Parameters

The **C.I. for Parameters** table gives a confidence interval for each parameter for each confidence coefficient specified. You choose the confidence interval for parameters either in the fit output options dialog or from the **Tables** menu, as shown in Figure 39.19.

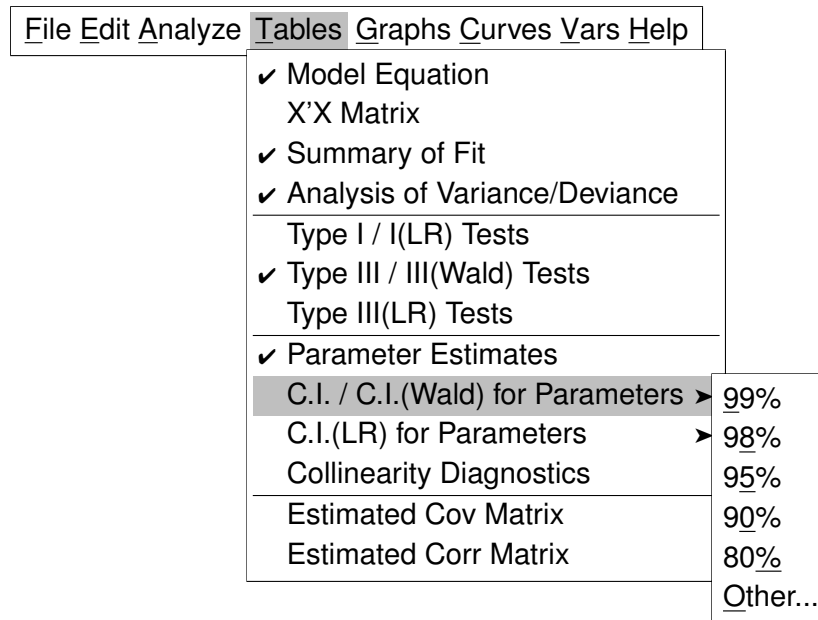
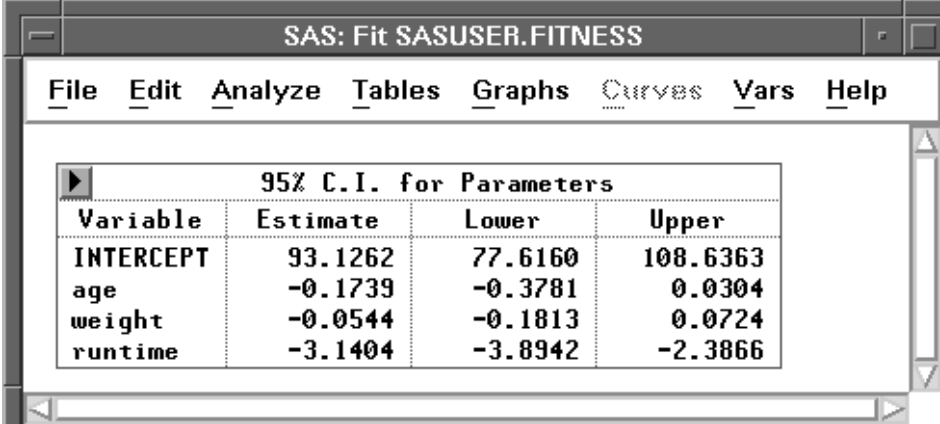


Figure 39.19. C.I. for Parameters Menu

Selecting **95% C.I. / C.I.(Wald) for Parameters** or **95% C.I.(LR) for Parameters** in the fit output options dialog produces a table with a 95% confidence interval for the parameters. This is the equivalent of choosing **Tables:C.I. / C.I.(Wald) for Parameters:95%** or **Tables:C.I.(LR) for Parameters:95%** from the **Tables** menu. You can also choose other confidence coefficients from the **Tables** menu. Figure 39.20 illustrates a 95% confidence intervals table for the parameters in a linear model.



Variable	Estimate	Lower	Upper
INTERCEPT	93.1262	77.6160	108.6363
age	-0.1739	-0.3781	0.0304
weight	-0.0544	-0.1813	0.0724
runtime	-3.1404	-3.8942	-2.3866

Figure 39.20. C.I. for Parameters Table

Reference ♦ Fit Analyses

For linear models, a $100(1 - \alpha)\%$ confidence interval has upper and lower limits

$$b_j \pm t_{(1-\alpha/2)} s_j$$

where $t_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ critical value of the Student's t statistic with degrees of freedom $n-p$, used in computing s_j , the estimated standard deviation of the parameter estimate b_j .

For generalized models, you can specify the confidence interval based on either a Wald type statistic or the likelihood function.

A $100(1 - \alpha)\%$ Wald type confidence interval is constructed from

$$\left(\frac{\beta^j - b_j}{s_j} \right)^2 \leq \chi_{(1-\alpha),1}^2$$

where $\chi_{(1-\alpha),1}^2$ is the $(1 - \alpha)$ critical value of the χ^2 statistic with one degree of freedom, and s_j is the estimated standard deviation of the parameter estimate b_j .

Thus, $100(1 - \alpha)\%$ upper and lower limits are

$$b_j \pm z_{(1-\alpha/2)} s_j$$

where $z_{(1-\alpha/2)}$ is the $(1 - \alpha/2)$ critical value of the standard normal statistic.

A table of 95% Wald type confidence intervals for the parameters is shown in [Figure 39.21](#).

SAS: Fit SASUSER.PATIENT			
File Edit Analyze Tables Graphs Curves Vars Help			
95% C.I. (Wald) for Parameters			
Variable	Estimate	Lower	Upper
INTERCEPT	58.0385	-81.5824	197.6593
cell	24.6615	-69.0987	118.4218
smear	19.2936	-94.2864	132.8736
infil	-19.6013	-140.4948	101.2923
li	3.8960	-0.6847	8.4766
blast	0.1511	-4.3148	4.6170
temp	-87.4339	-219.8756	45.0078
95% C.I. (LR) for Parameters			
Variable	Estimate	Lower	Upper
INTERCEPT	58.0385	-70.9470	222.1757
cell	24.6615	-27.4212	138.3904
smear	19.2936	-60.2651	152.1511
infil	-19.6013	-159.7391	67.3877
li	3.8960	0.1943	9.5266
blast	0.1511	-4.5238	4.7145
temp	-87.4339	-244.7432	24.9519

Figure 39.21. C.I. for Parameters Tables

The likelihood ratio test statistic for the null hypothesis

$$H_0: \beta_j = \beta_{j0}$$

where β_{j0} is a specified value, is

$$\lambda = -2(l(\hat{\beta}_0) - l(\hat{\beta}))$$

where $l(\hat{\beta}_0)$ is the maximized log likelihood under H_0 and $l(\hat{\beta})$ is the maximized log likelihood over all β .

In large samples, the hypothesis is rejected at level α if the test statistic λ is greater than the $(1 - \alpha)$ critical value of the chi-squared statistic with one degree of freedom.

Thus a $100(1 - \alpha)\%$ likelihood-based confidence interval is constructed using restricted maximization to find upper and lower limits satisfying

$$l(\hat{\beta}_0) = l(\hat{\beta}) - \frac{1}{2}\chi_{(1-\alpha),1}^2$$

An iterative procedure is used to obtain these limits. A 95% likelihood-based confidence interval table for the parameters is illustrated in [Figure 39.21](#).

Collinearity Diagnostics

The **Collinearity Diagnostics** table is illustrated by Figure 39.22.

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	Variance Proportion			
			INTERCEPT	age	weight	runtime
1	3.9718	1.0000	0.0003	0.0006	0.0006	0.0010
2	0.0139	16.8921	0.0005	0.3254	0.3924	0.0138
3	0.0114	18.6347	0.0213	0.1416	0.0405	0.9753
4	0.0029	37.2156	0.9779	0.5323	0.5665	0.0100

Figure 39.22. Collinearity Diagnostics Table

- Number** is the eigenvalue number.
- Eigenvalue** gives the eigenvalues of the $\mathbf{X}'\mathbf{X}$ matrix.
- Condition Index** is the square root of the ratio of the largest eigenvalue to the corresponding eigenvalue.
- Variance Proportion** is the proportion of the variance of each estimate accounted for by each component.

Detailed collinearity diagnostics use the eigenstructure of $\mathbf{X}'\mathbf{X}$, which can be written as

$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}'$ where \mathbf{V} is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{X}'\mathbf{X}$, and \mathbf{D}^2 is a diagonal matrix of eigenvalues

$$d_1^2 \geq d_2^2 \geq \dots \geq d_p^2$$

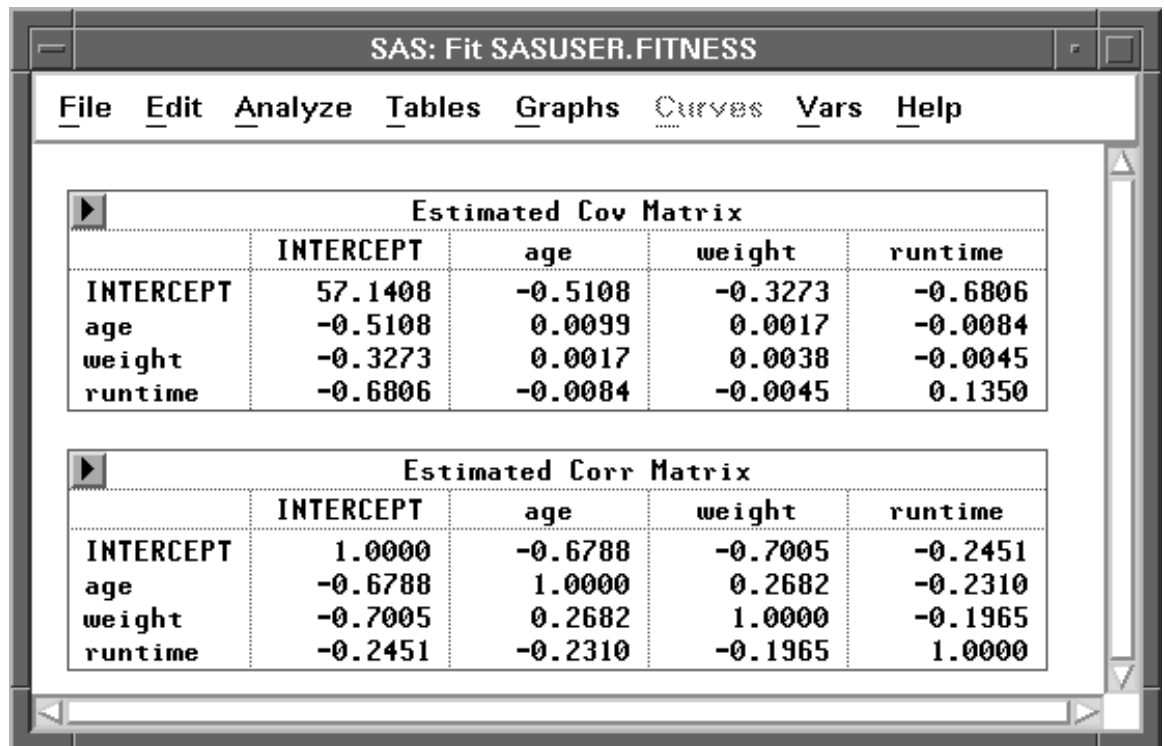
After scaling $(\mathbf{X}'\mathbf{X})$ to correlation form, Belsley, Kuh, and Welsch (1980) construct the condition indices as the square roots of the ratio of the largest eigenvalue to each individual eigenvalue, d_1/d_j , $j = 1, 2, \dots, p$.

The *condition number* of the \mathbf{X} matrix is defined as the largest condition index, d_1/d_p . When this number is large, the data are said to be *ill conditioned*. A condition index of 30 to 100 indicates moderate to strong collinearity.

For each variable, the proportion of the variance of its estimate accounted for by each component d_j can be evaluated. A collinearity problem occurs when a component associated with a high condition index contributes strongly to the variance of two or more variables. Thus, for a high condition index (>30), the corresponding row should be examined to see which variables have high values. Those would indicate near-linear dependence.

Estimated COV Matrix and Estimated CORR Matrix

The **Estimated COV Matrix** table contains the estimated variance-covariance matrix of the parameters. The **Estimated CORR Matrix** table contains the estimated correlation matrix of the parameters. Sample tables are shown in [Figure 39.23](#).



	INTERCEPT	age	weight	runtime
INTERCEPT	57.1408	-0.5108	-0.3273	-0.6806
age	-0.5108	0.0099	0.0017	-0.0084
weight	-0.3273	0.0017	0.0038	-0.0045
runtime	-0.6806	-0.0084	-0.0045	0.1350

	INTERCEPT	age	weight	runtime
INTERCEPT	1.0000	-0.6788	-0.7005	-0.2451
age	-0.6788	1.0000	0.2682	-0.2310
weight	-0.7005	0.2682	1.0000	-0.1965
runtime	-0.2451	-0.2310	-0.1965	1.0000

Figure 39.23. Estimated COV and CORR Matrices

Residual and Surface Plots

Residual plots provide visual displays for assessing how well the model fits the data, for evaluating the distribution of the residuals, and for identifying influential observations. Surface plots are three-dimensional displays of continuous response surfaces on the regular grids of the explanatory variables. They are much easier to comprehend than rotating plots.

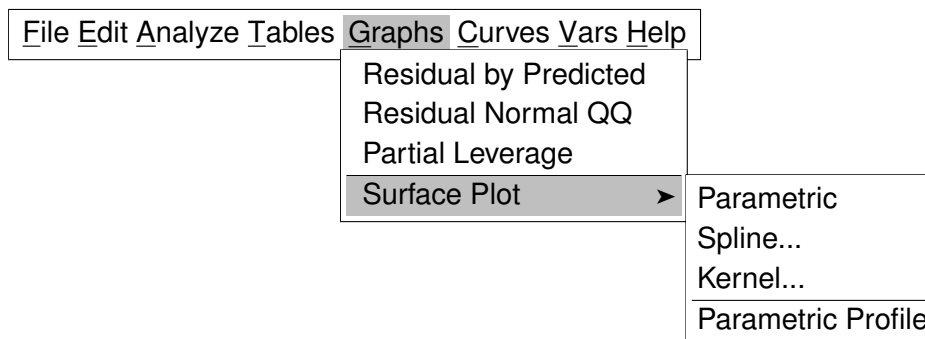


Figure 39.24. Graphs Menu

Residual-by-Predicted Plot

A residual-by-predicted plot is commonly used to diagnose nonlinearity or nonconstant error variance. It is also used to find outliers. A residual-by-predicted plot, as illustrated by the plot on the left in [Figure 39.25](#), is a plot of residuals versus predicted response for each observation. See the “[Predicted Values](#)” and “[Residuals](#)” sections for a further explanation of the axis variables.

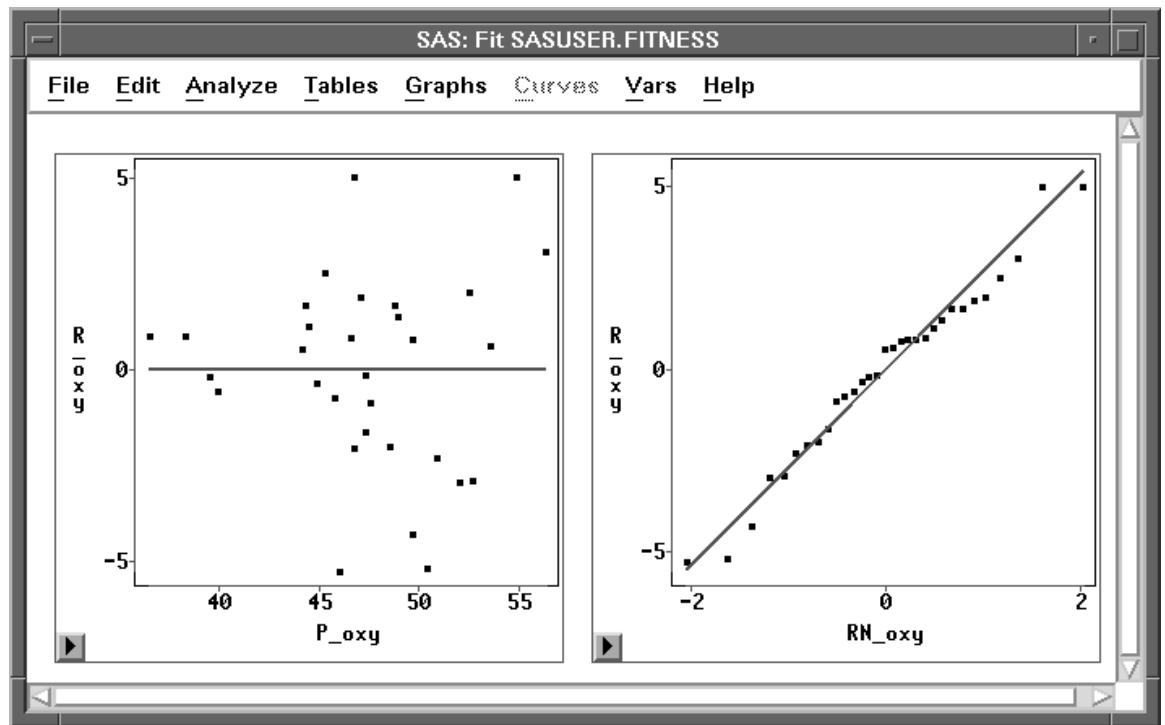


Figure 39.25. Residual-by-Predicted and Residual Normal QQ Plots

Residual Normal QQ Plot

A normal quantile-quantile plot of residuals is illustrated by the plot on the right in [Figure 39.25](#). See the “[Residual Normal Quantiles](#)” section for an explanation of the **X** axis variable.

The empirical quantiles are plotted against the quantiles of a standard normal distribution. If the residuals are from a normal distribution with mean 0, the points tend to fall along the reference line that has an intercept of 0 and a slope equal to the estimated standard deviation.

Partial Leverage Plots

For linear models, the partial leverage plot for a selected explanatory variable can be obtained by plotting the residuals for the response variable against the residuals for the selected explanatory variable. The residuals for the response variable are calculated from a model having the selected explanatory variable omitted, and the residuals for the selected explanatory variable are calculated from a model where the selected explanatory variable is regressed on the remaining explanatory variables.

Let $\mathbf{X}_{[j]}$ be the $n \times (p-1)$ matrix formed from the design matrix \mathbf{X} by removing the j th column, \mathbf{X}_j . Let $\mathbf{r}_{y[j]}$ be the partial leverage **Y** variable containing the residuals that result from regressing \mathbf{y} on $\mathbf{X}_{[j]}$ and let $\mathbf{r}_{x[j]}$ be the partial leverage **X** variable containing the residuals that result from regressing \mathbf{X}_j on $\mathbf{X}_{[j]}$. Then a partial leverage plot is a scatter plot of $\mathbf{r}_{y[j]}$ against $\mathbf{r}_{x[j]}$. Partial leverage plots for two explanatory variables are illustrated by [Figure 39.26](#).

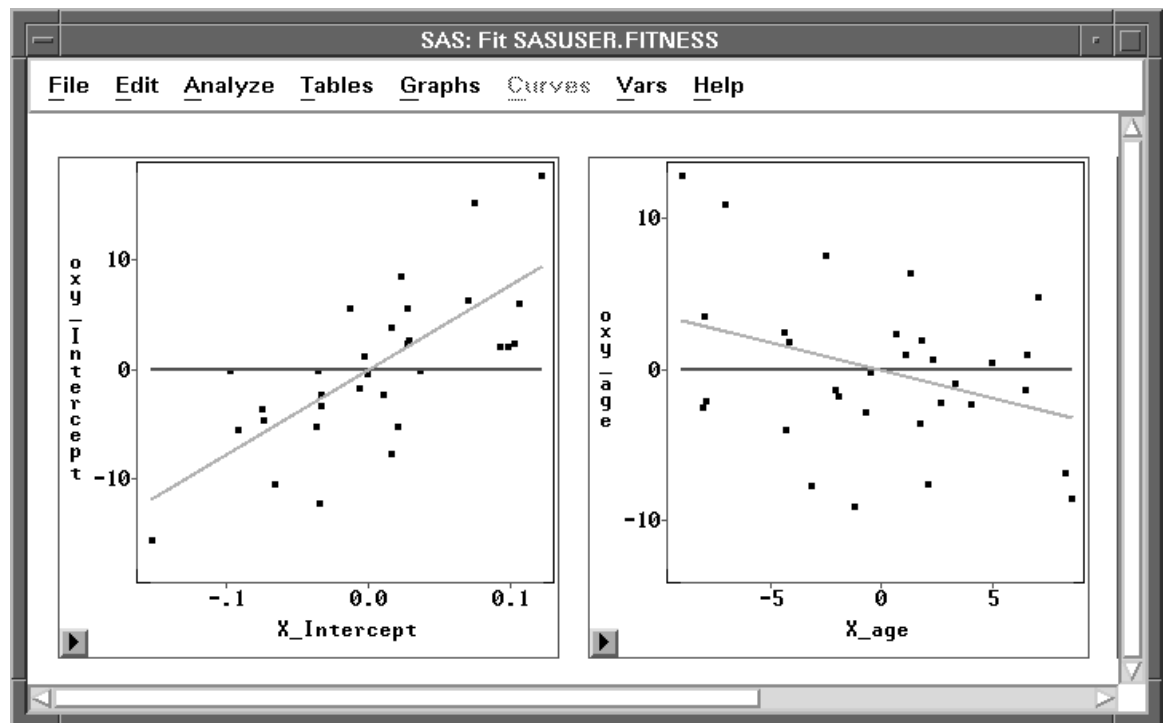


Figure 39.26. Partial Leverage Plots

In a partial leverage plot, the partial leverage \mathbf{Y} variable $\mathbf{r}_{y[j]}$ can also be computed as

$$r_{y[j]i} = r_{x[j]i}b_j + (y_i - \hat{\mu}_i)$$

For generalized linear models, the partial leverage \mathbf{Y} is also computed as

$$r_{y[j]i} = r_{x[j]i}b_j + (y_i - \hat{\mu}_i)g'(\hat{\mu}_i)$$

Two reference lines are also displayed in the plots. One is the horizontal line of $\mathbf{Y} = 0$, and the other is the fitted regression of $\mathbf{r}_{y[j]}$ against $\mathbf{r}_{x[j]}$. The latter has an intercept of 0 and a slope equal to the parameter estimate associated with the explanatory variable in the model. The leverage plot shows the changes in the residuals for the model with and without the explanatory variable. For a given data point in the plot, its residual without the explanatory variable is the vertical distance between the point and the horizontal line; its residual with the explanatory variable is the vertical distance between the point and the fitted line.

Parametric Surface Plot

With two explanatory interval variables in the model, a parametric surface plot is a continuous surface plot of the predicted responses from the fitted parametric model on a set of regular grids of the explanatory variables. [Figure 39.27](#) shows a response surface plot of **oxy** as a quadratic function of **age** and **weight**.

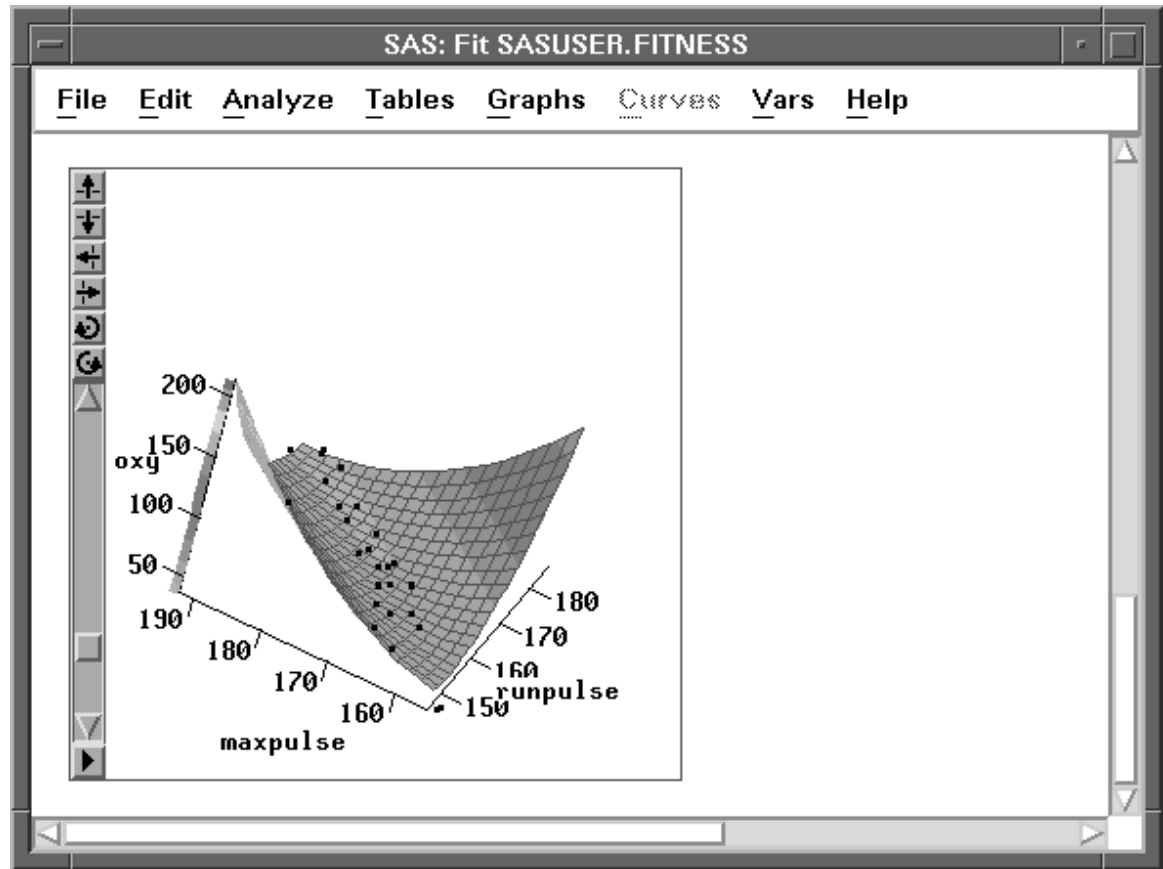


Figure 39.27. Parametric Surface Plot

The response surface is displayed with options **Drawing Modes:Smooth Color** and **Axes:Three Sections**.

Smoothing Spline Surface Plot

Two criteria can be used to select an estimator \hat{f}_λ for the function f :

- goodness of fit to the data
- smoothness of the fit

A standard measure of goodness of fit is the mean residual sum of squares

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}_\lambda(\mathbf{x}_i))^2$$

A measure of the smoothness of a fit is an integrated squared second derivative

$$J_2(f_\lambda) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\left(\frac{\partial^2 f_\lambda}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f_\lambda}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f_\lambda}{\partial x_2^2} \right)^2 \right) dx_1 dx_2$$

A single criterion that combines the two criteria is then given by

$$S(\lambda) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}_\lambda(\mathbf{x}_i))^2 + \lambda J_2(f_\lambda)$$

where \hat{f}_λ belongs to the set of all continuously differentiable functions with square integrable second derivatives, and λ is a positive constant.

The estimator that results from minimizing $S(\lambda)$ is called a *thin-plate smoothing spline estimator*. Wahba and Wendelberger (1980) derived a closed form expression for the thin-plate smoothing spline estimator.

† **Note:** The computations for a thin-plate smoothing spline are time intensive, especially for large data sets.

The smoothing parameter λ controls the amount of smoothing; that is, it controls the trade-off between the goodness of fit to the data and the smoothness of the fit. You select a smoothing parameter λ by specifying a constant c in the formula

$$\lambda = c/100$$

The values of the explanatory variables are scaled by their corresponding interquartile ranges before the computations. This makes the computations independent of the units of X_1 and X_2 .

After choosing **Graphs:Surface Plot:Spline** from the menu, you specify a smoothing parameter selection method in the **Spline Fit** dialog.

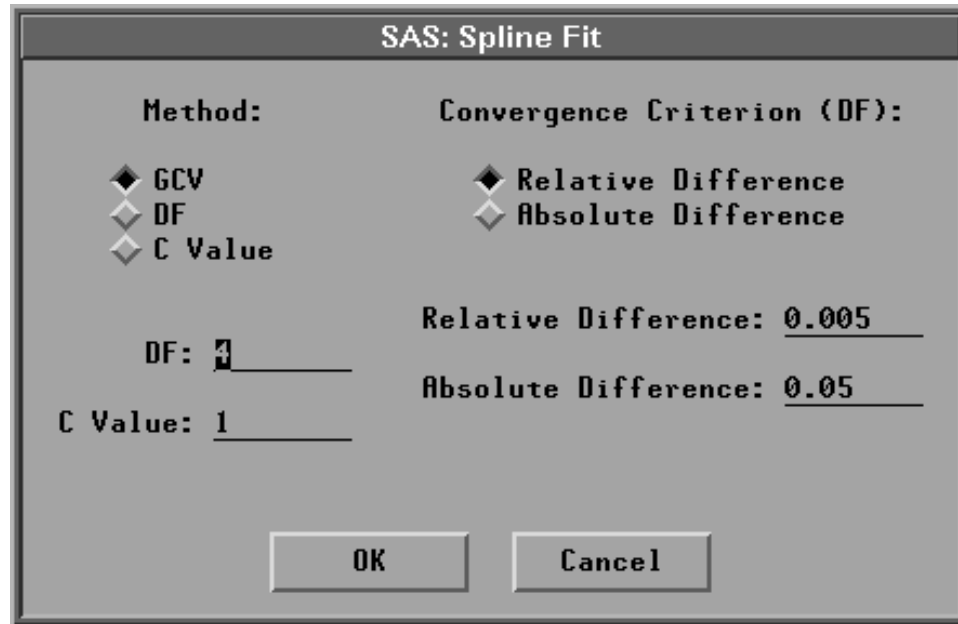


Figure 39.28. Spline Surface Fit Dialog

The default **Method:GCV** uses a c value that minimizes the generalized cross validation mean squared error $MSE_{GCV}(\lambda)$. [Figure 39.29](#) displays smoothing spline estimates with c values of 0.0000831 (the GCV value) and 0.4127 (DF=6). Use the slider in the table to change the c value of the spline fit.

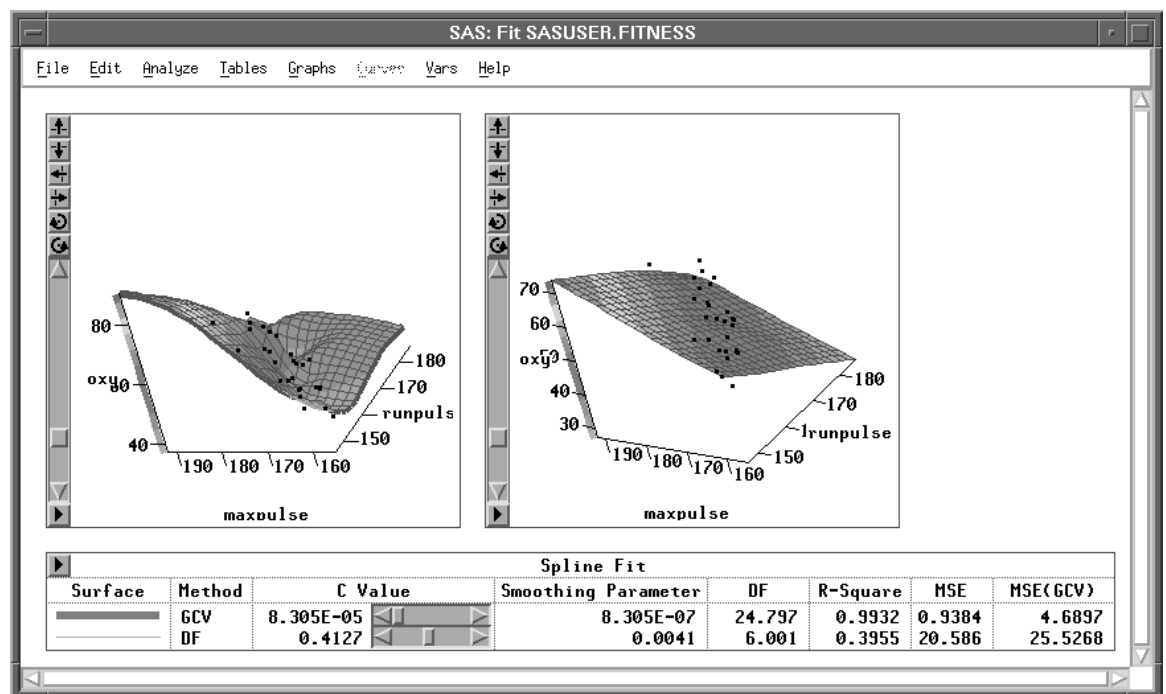


Figure 39.29. Smoothing Spline Surface Plot

Kernel Surface Plot

A *kernel estimator* uses an explicitly defined set of weights at each point \mathbf{x} to produce the estimate at \mathbf{x} . The kernel estimator of f has the form

$$\hat{f}_\lambda(\mathbf{x}) = \sum_{i=1}^n W(\mathbf{x}, \mathbf{x}_i; \lambda, \mathbf{V}_\mathbf{x}) y_i$$

where $W(\mathbf{x}, \mathbf{x}_i; \lambda, \mathbf{V}_\mathbf{x})$ is the weight function that depends on the smoothing parameter λ and the diagonal matrix $\mathbf{V}_\mathbf{x}$ of the squares of the sample interquartile ranges.

The weights are derived from a single function that is independent of the design

$$W(\mathbf{x}, \mathbf{x}_i; \lambda, \mathbf{V}_\mathbf{x}) = \frac{K_0((\mathbf{x} - \mathbf{x}_i)/\lambda, \mathbf{V}_\mathbf{x})}{\sum_{j=1}^n K_0((\mathbf{x} - \mathbf{x}_j)/\lambda, \mathbf{V}_\mathbf{x})}$$

where K_0 is a kernel function and λ is the bandwidth or smoothing parameter. The weights are nonnegative and sum to 1.

Symmetric probability density functions commonly used as kernel functions are

- *Normal* $K_0(\mathbf{t}, \mathbf{V}) = \frac{1}{2\pi} \exp(-\frac{1}{2}\mathbf{t}'\mathbf{V}^{-1}\mathbf{t})$ for all \mathbf{t}
- *Quadratic* $K_0(\mathbf{t}, \mathbf{V}) = \begin{cases} \frac{2}{\pi}(1 - \mathbf{t}'\mathbf{V}^{-1}\mathbf{t}) & \text{for } \mathbf{t}'\mathbf{V}^{-1}\mathbf{t} \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- *Biweight* $K_0(\mathbf{t}, \mathbf{V}) = \begin{cases} \frac{3}{\pi}(1 - \mathbf{t}'\mathbf{V}^{-1}\mathbf{t})^2 & \text{for } \mathbf{t}'\mathbf{V}^{-1}\mathbf{t} \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- *Triweight* $K_0(\mathbf{t}, \mathbf{V}) = \begin{cases} \frac{4}{\pi}(1 - \mathbf{t}'\mathbf{V}^{-1}\mathbf{t})^3 & \text{for } \mathbf{t}'\mathbf{V}^{-1}\mathbf{t} \leq 1 \\ 0 & \text{otherwise} \end{cases}$

You select a bandwidth λ for each kernel estimator by specifying c in the formula

$$\lambda = n^{-\frac{1}{6}} c$$

where n is the sample size. Both λ and c are independent of the units of \mathbf{X} .

SAS/INSIGHT software divides the range of each explanatory variable into a number of evenly spaced intervals, then estimates the kernel fit on this grid. For a data point \mathbf{x}_i that lies between two grid points, a linear interpolation is used to compute the predicted value. For \mathbf{x}_i that lies inside a square of grid points, a pair of points that lie on the same vertical line as \mathbf{x}_i and each lying between two grid points can be found. A linear interpolation of these two points is used to compute the predicted value.

After choosing **Graphs:Surface Plot:Kernel** from the menu, you specify a kernel and smoothing parameter selection method in the **Kernel Fit** dialog.

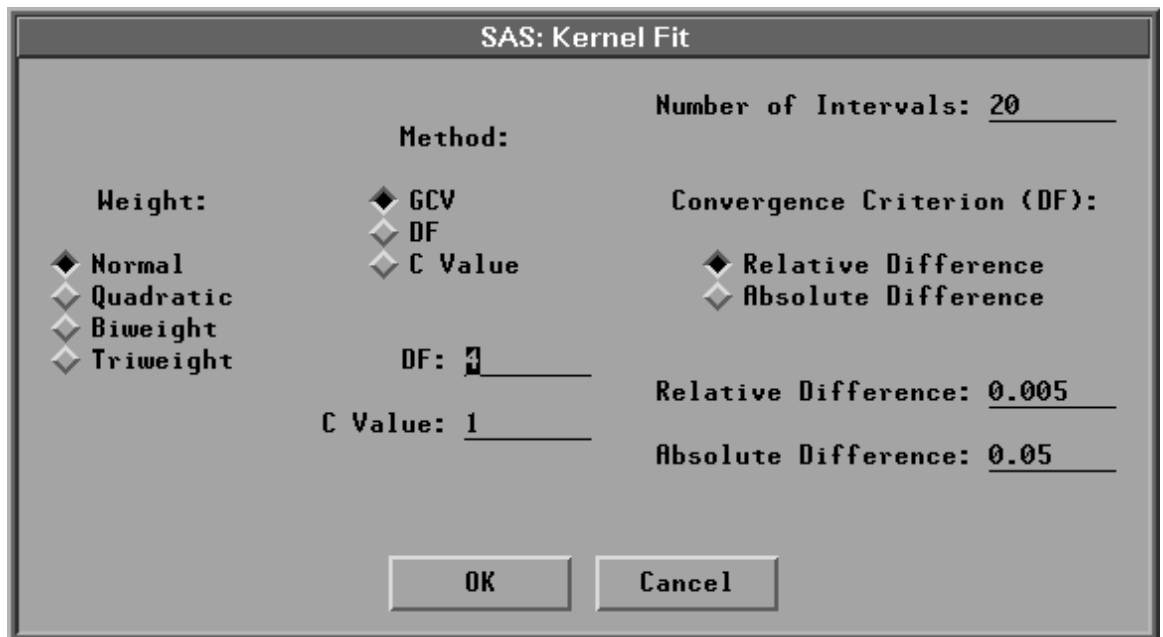


Figure 39.30. Kernel Surface Fit Dialog

By default, SAS/INSIGHT software divides the range of each explanatory variable into 20 evenly spaced intervals, uses a normal weight, and uses a c value that minimizes $MSE_{GCV}(\lambda)$. [Figure 39.31](#) illustrates normal kernel estimates with c values of 0.5435 (the GCV value) and 1.0. Use the slider to change the c value of the kernel fit.

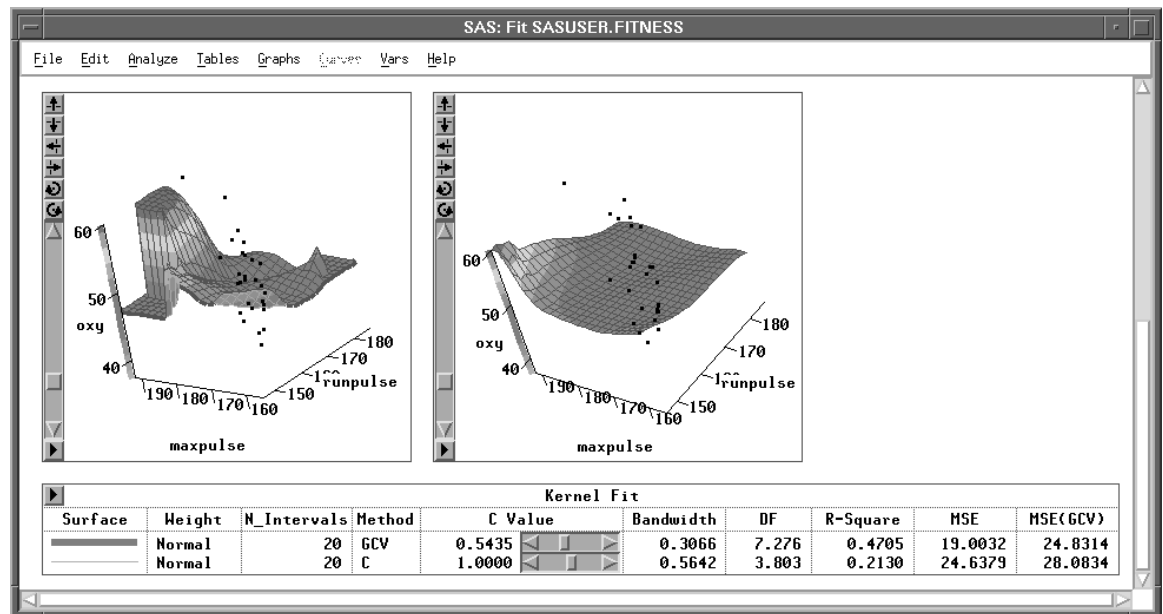


Figure 39.31. Kernel Surface Plot

Parametric Profile Surface Plot

With more than two explanatory interval variables in the model, a parametric profile surface plot is a continuous surface plot of the predicted responses from the fitted parametric model on a set of regular grids of a pair of explanatory variables. The values of the remaining explanatory variables are initially set at their means and can be adjusted with the sliders.

By default, the first two explanatory variables are used in the surface plot. You can also create profile surface plots for other explanatory variables by selecting the two variables before choosing **Graphs:Surface Plot:Parametric profile**. Figure 39.32 shows a parametric profile surface plot of **oxy** as a quadratic function of **runpulse** and **maxpulse** with **rstpulse** = 53.4516.

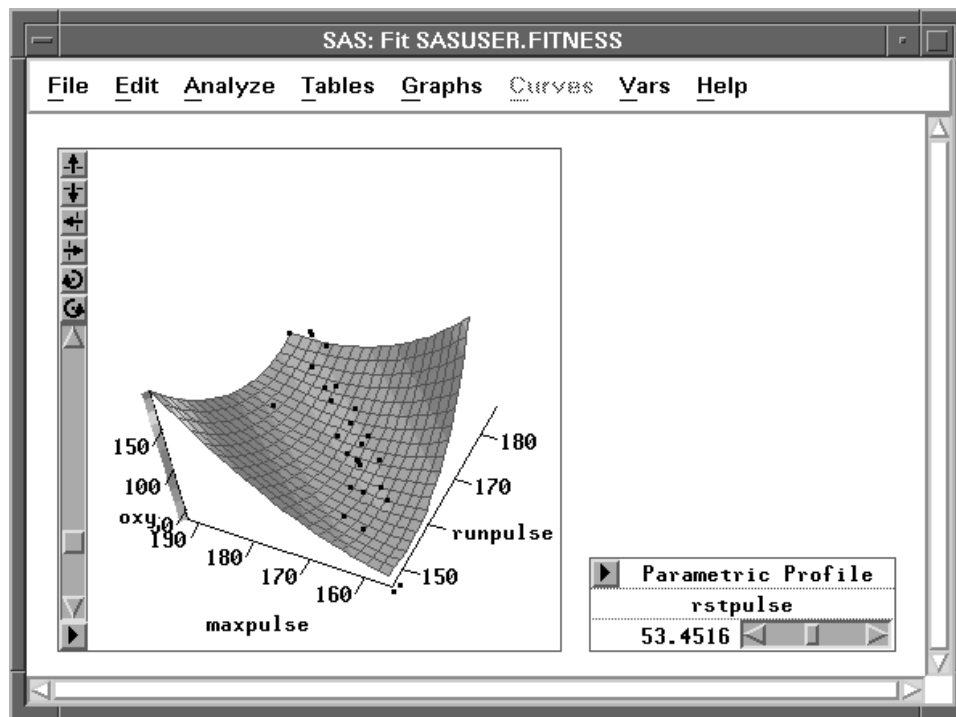


Figure 39.32. Parametric Profile Surface Plot

Fit Curves

When you are working with one explanatory variable, you can fit curves to the **Y**-by-**X** scatter plot generated when the analysis is first created. Use the output dialog (see [Figure 39.4](#), [Figure 39.6](#), and [Figure 39.7](#)) or the **Curves** menu in [Figure 39.33](#) to fit curves to the scatter plot.

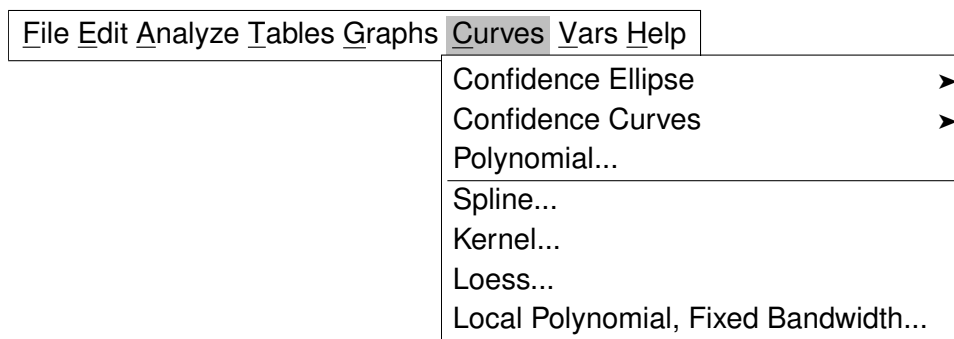


Figure 39.33. Curves Menu

There are two types of fitting techniques: parametric and nonparametric. Parametric techniques enable you to add confidence ellipses, fit regression polynomials, and add confidence curves of fitted polynomials to the **Y**-by-**X** scatter plot. Nonparametric techniques enable you to add spline, kernel, and local polynomial fits to the **Y**-by-**X** scatter plot.

Parametric Curves: Confidence Ellipses

SAS/INSIGHT software provides two types of confidence ellipses for each pair of **X** and **Y** variables assuming a bivariate normal distribution. One is a confidence ellipse for the population mean, and the other is a confidence ellipse for prediction.

Let $\bar{\mathbf{Z}}$ and \mathbf{S} be the sample mean and the unbiased estimate of the covariance matrix of a random sample of size n from a bivariate normal distribution with mean μ and covariance matrix Σ .

The variable $\bar{\mathbf{Z}} - \mu$ is distributed as a bivariate normal variate with mean 0 and covariance $n^{-1}\Sigma$, and it is independent of \mathbf{S} . The confidence ellipse for μ is based on Hotelling's T^2 statistic:

$$T^2 = n(\bar{\mathbf{Z}} - \mu)' \mathbf{S}^{-1} (\bar{\mathbf{Z}} - \mu)$$

A $100(1 - \alpha)\%$ confidence ellipse for μ is defined by the equation

$$(\bar{\mathbf{Z}} - \mu)' \mathbf{S}^{-1} (\bar{\mathbf{Z}} - \mu) = \frac{2(n-1)}{n(n-2)} F_{2,n-2}(1 - \alpha)$$

where $F_{2,n-2}(1 - \alpha)$ is the $(1 - \alpha)$ critical value of an F variate with degrees of freedom 2 and $n - 2$.

A confidence ellipse for prediction is a confidence region for predicting a new observation in the population. It also approximates a region containing a specified percentage of the population.

Consider \mathbf{Z} as a bivariate random variable for a new observation. The variable $\mathbf{Z} - \bar{\mathbf{Z}}$ is distributed as a bivariate normal variate with mean 0 and covariance $(1 + 1/n)\mathbf{\Sigma}$, and it is independent of \mathbf{S} .

A $100(1 - \alpha)\%$ confidence ellipse for prediction is then given by the equation

$$(\mathbf{Z} - \bar{\mathbf{Z}})' \mathbf{S}^{-1} (\mathbf{Z} - \bar{\mathbf{Z}}) = \frac{2(n+1)(n-1)}{n(n-2)} F_{2,n-2}(1-\alpha)$$

The family of ellipses generated by different F critical values has a common center (the sample mean) and common major and minor axes.

The ellipses graphically indicate the correlation between two variables. When the variable axes are standardized (by dividing the variables by their respective standard deviations), the ratio of the two axis lengths (in Euclidean distances) reflects the magnitude of the correlation between the two variables. A ratio of 1 between the major and minor axes corresponds to a circular confidence contour and indicates that the variables are uncorrelated. A larger value of the ratio indicates a larger positive or negative correlation between the variables.

You can choose the level of the confidence region from the **Confidence Ellipse** menus, as illustrated by [Figure 39.34](#).

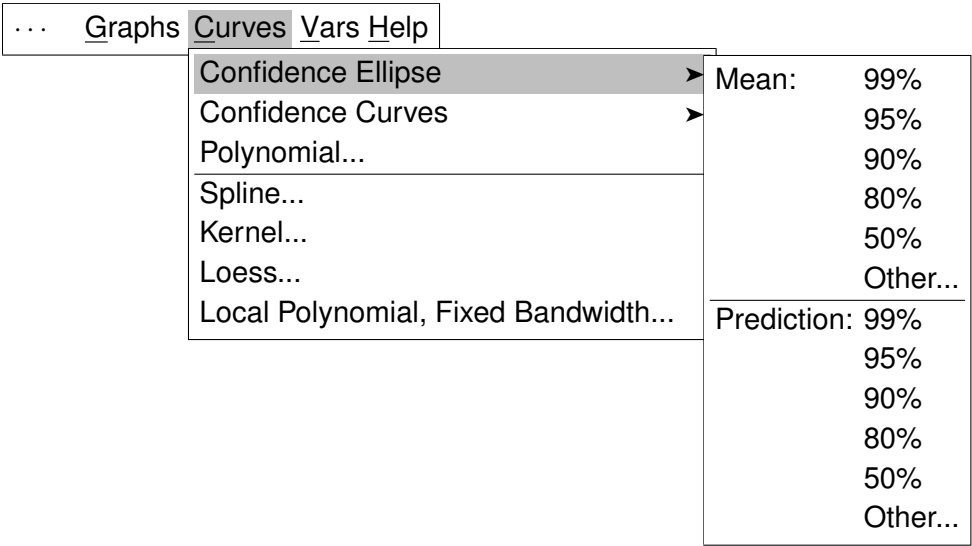


Figure 39.34. Confidence Ellipse Menu

A confidence ellipse for the population mean is displayed with dashed lines, and a confidence ellipse for prediction is displayed with dotted lines. [Figure 39.35](#) displays a scatter plot with 50% and 80% confidence ellipses for prediction. Use the sliders in the **Confidence Ellipses** table to change the coefficient of the confidence ellipses.

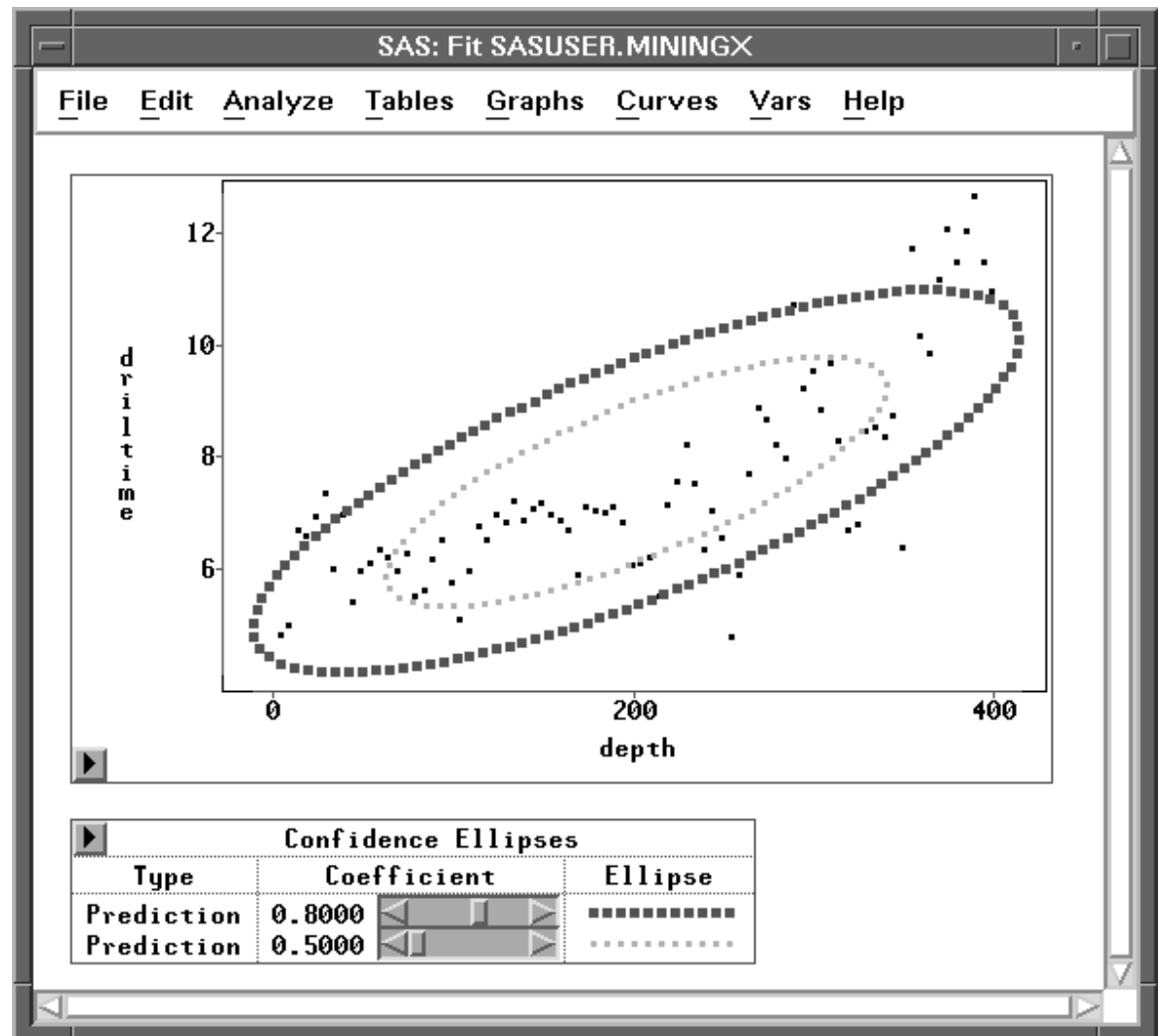


Figure 39.35. Confidence Ellipses for Prediction

Parametric Curves: Polynomial

Choose **Curves:Polynomial** from the menu to add a polynomial regression fit to the **Y-by-X** scatter plot. This displays the Polynomial Fit dialog in [Figure 39.36](#).

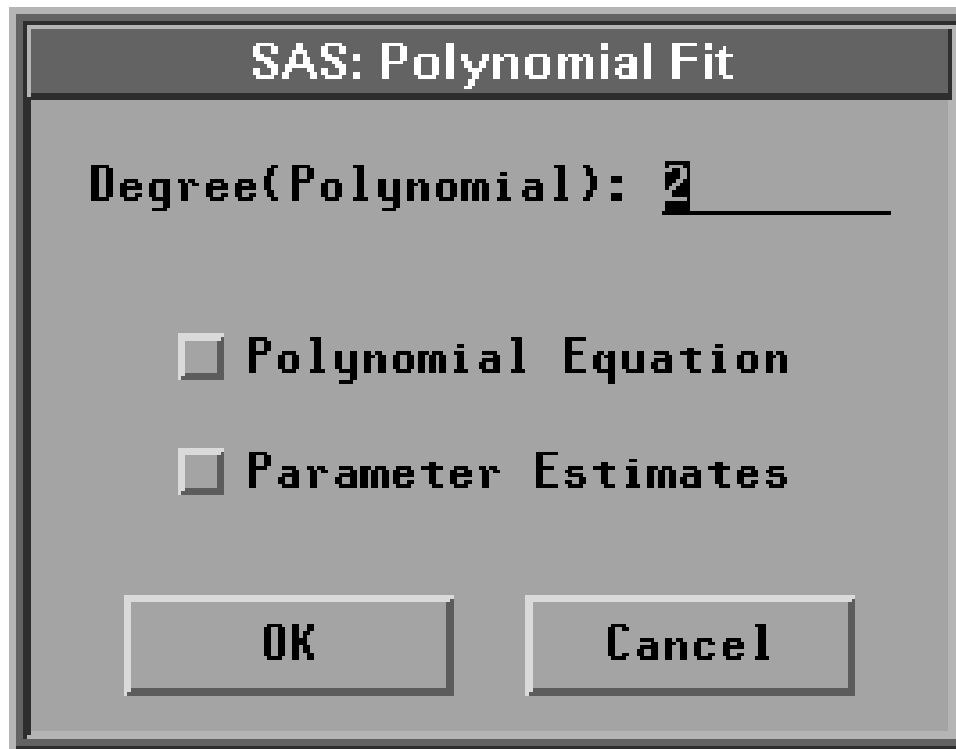


Figure 39.36. Polynomial Fit Dialog

In the **Polynomial Fit** dialog, you enter the degree for the polynomial fit. Select the **Polynomial Equation** or **Parameter Estimates** options to create a **Polynomial Equation** or **Parameter Estimates** table for the fitted curve.

Information about the polynomial fit is displayed in a table, as illustrated by [Figure 39.37](#). The information includes the R^2 value and an F statistic and its associated p -value for testing the null hypothesis that all parameters are 0 except for the intercept. A parametric regression fit table includes the following:

Curve	is the curve in the Y-by-X scatter plot.
Degree(Polynomial)	is the degree for the polynomial fit.
Model DF	is the degrees of freedom for model.
Model Mean Square	is the mean square for model.
Error DF	is the degrees of freedom for error.
Error Mean Square	is the mean square for error.
R-Square	is the proportion of the (corrected) total variation attributed to the fit.

F Stat	is the F statistic for testing the null hypothesis that all parameters are zero except for the intercept. This is formed by dividing the mean square for model by the mean square for error.
Pr > F	is the probability under the null hypothesis of obtaining a greater F statistic than that observed.

Figure 39.37 displays a quadratic polynomial fit with **Polynomial Equation** and **Parameter Estimates** tables.

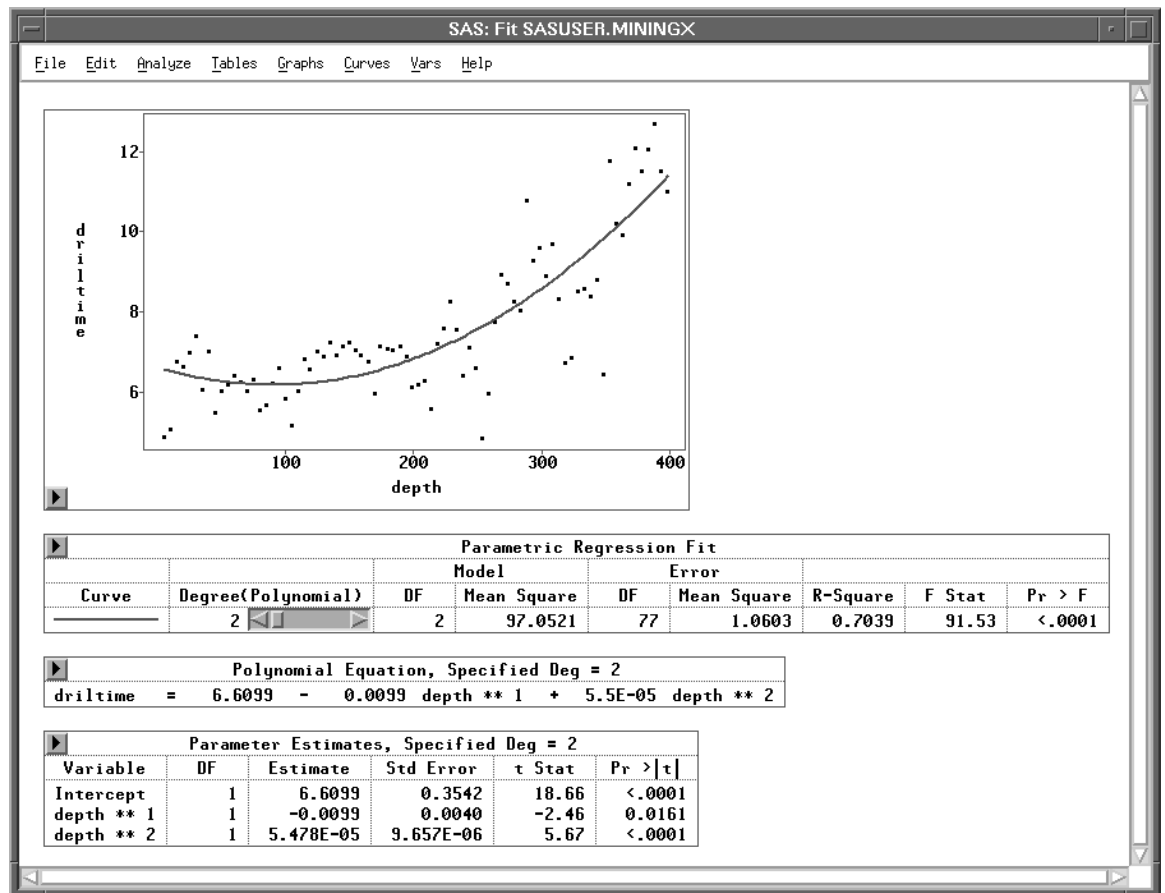


Figure 39.37. A Quadratic Polynomial Fit

You can use the **Degree(Polynomial)** slider in the **Parametric Regression Fit** table to change the degree of the polynomial curve fit. However, these will not change the **Polynomial Equation** and **Parameter Estimates** tables. You can produce a new **Polynomial Equation** or **Parameter Estimates** table by selecting the **Polynomial Equation** or **Parameter Estimates** option from the **Polynomial Fit** dialog.

Parametric Curves: Confidence Curves

You can add two types of confidence curves for the predicted values. One curve is for the mean value of the response, and the other is for the prediction of a new observation.

For the i th observation, a confidence interval that covers the expected value of the response with probability $1 - \alpha$ has upper and lower limits

$$\mathbf{x}_i \mathbf{b} \pm t_{(1-\alpha/2)} \sqrt{h_i s}$$

where $t_{(1-\alpha/2)}$ is the $(1 - \alpha/2)$ critical value of the Student's t statistic with degrees of freedom equal to the degrees of freedom for the mean squared error and h_i is the i th diagonal element of the hat matrix \mathbf{H} . The hat matrix \mathbf{H} is described in the section “Output Variables” later in this chapter.

The $100(1 - \alpha)\%$ upper and lower limits for prediction are

$$\mathbf{x}_i \mathbf{b} \pm t_{(1-\alpha/2)} \sqrt{1 + h_i s}$$

You can generate confidence curves for a parametric regression fit by choosing the confidence coefficient from the **Curves:Confidence Curves** menu.

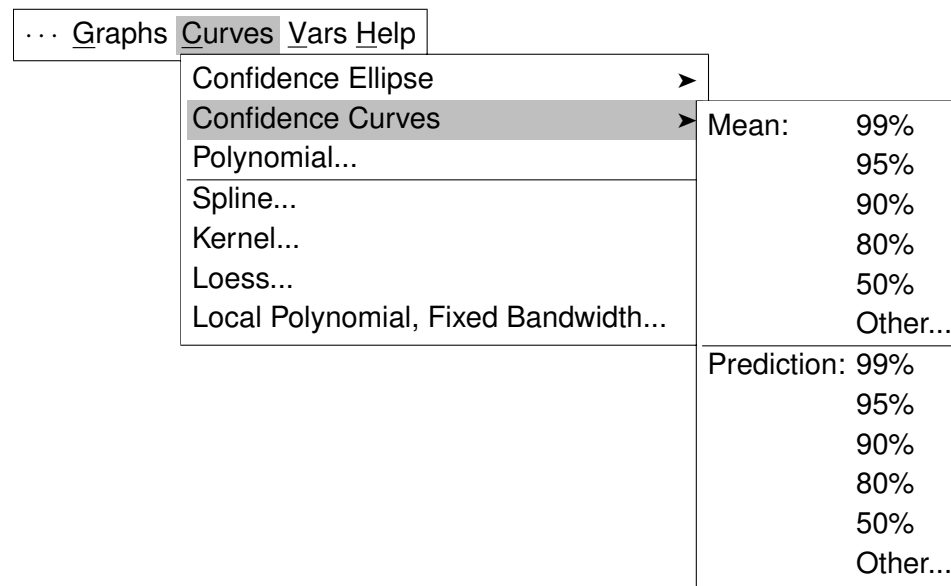


Figure 39.38. Confidence Curves Menu

Reference ♦ Fit Analyses

Figure 39.39 displays a quadratic polynomial fit with 95% mean confidence curves for the response. Use the **Coefficient** slider to change the confidence coefficient.

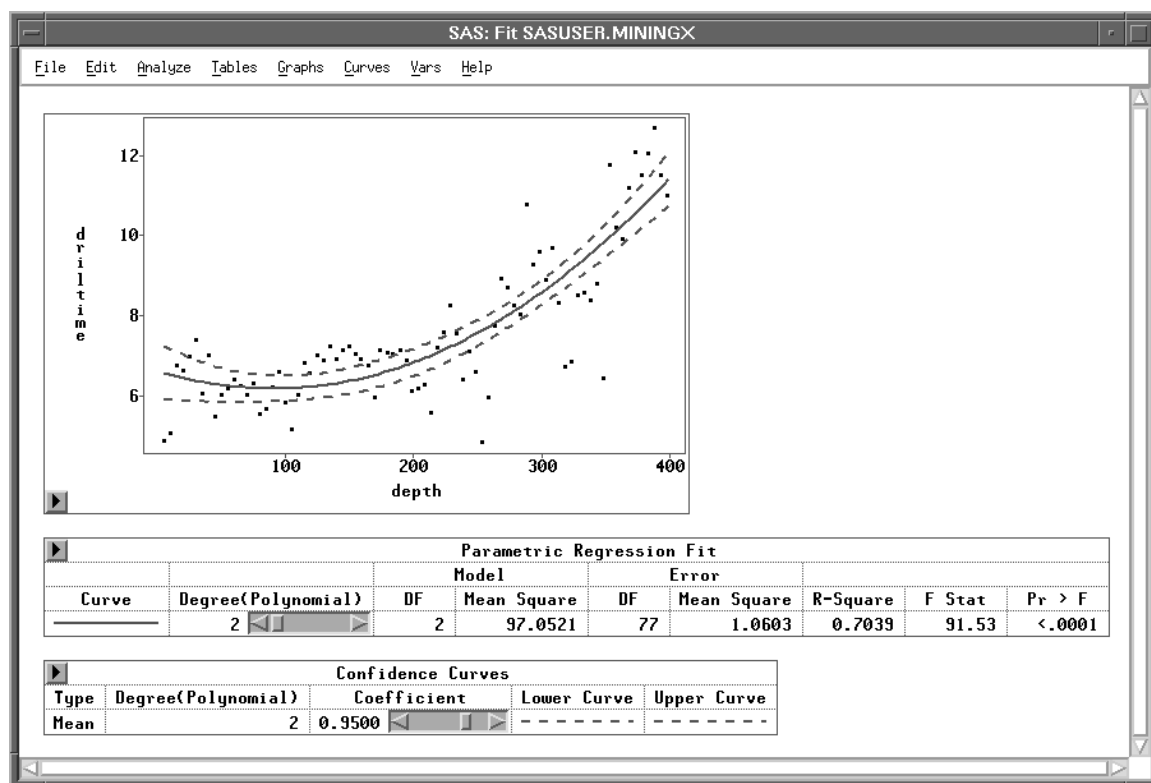


Figure 39.39. A Quadratic Polynomial Fit with 99% Mean Confidence Curves

Nonparametric Smoothing Spline

Two criteria can be used to select an estimator \hat{f}_λ for the function f :

- goodness of fit to the data
- smoothness of the fit

A standard measure of goodness of fit is the mean residual sum of squares

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}_\lambda(x_i))^2$$

A measure of the smoothness of a fit is the integrated squared second derivative

$$\int_{-\infty}^{\infty} (\hat{f}_\lambda''(x))^2 dx$$

A single criterion that combines the two criteria is then given by

$$S(\lambda) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}_\lambda(x_i))^2 + \lambda \int_{-\infty}^{\infty} (\hat{f}_\lambda''(x))^2 dx$$

where \hat{f}_λ belongs to the set of all continuously differentiable functions with square integrable second derivatives, and λ is a positive constant.

The estimator that results from minimizing $S(\lambda)$ is called the *smoothing spline estimator*. This estimator fits a cubic polynomial in each interval between points. At each point x_i , the curve and its first two derivatives are continuous (Reinsch 1967).

The smoothing parameter λ controls the amount of smoothing; that is, it controls the trade-off between the goodness of fit to the data and the smoothness of the fit. You select a smoothing parameter λ by specifying a constant c in the formula

$$\lambda = (Q/10)^3 c$$

where Q is the interquartile range of the explanatory variable. This formulation makes c independent of the units of \mathbf{X} .

After choosing **Curves:Spline**, you specify a smoothing parameter selection method in the **Spline Fit** dialog.

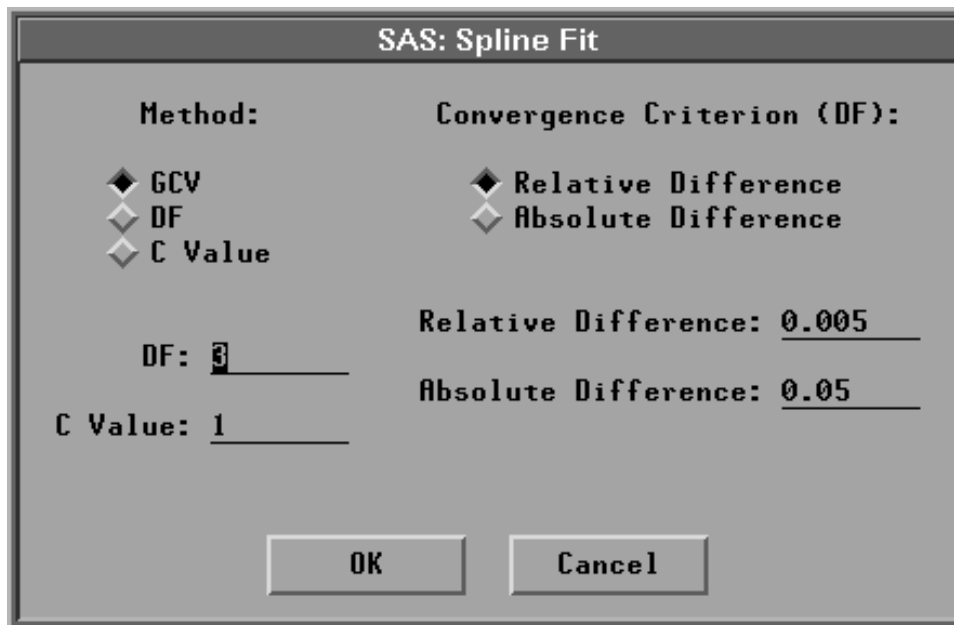


Figure 39.40. Spline Fit Dialog

The default **Method:GCV** uses a c value that minimizes the generalized cross validation mean squared error $MSE_{GCV}(\lambda)$. [Figure 39.41](#) displays smoothing spline estimates with c values of 0.0017 (the GCV value) and 15.2219 (DF=3). Use the slider in the table to change the c value of the spline fit.

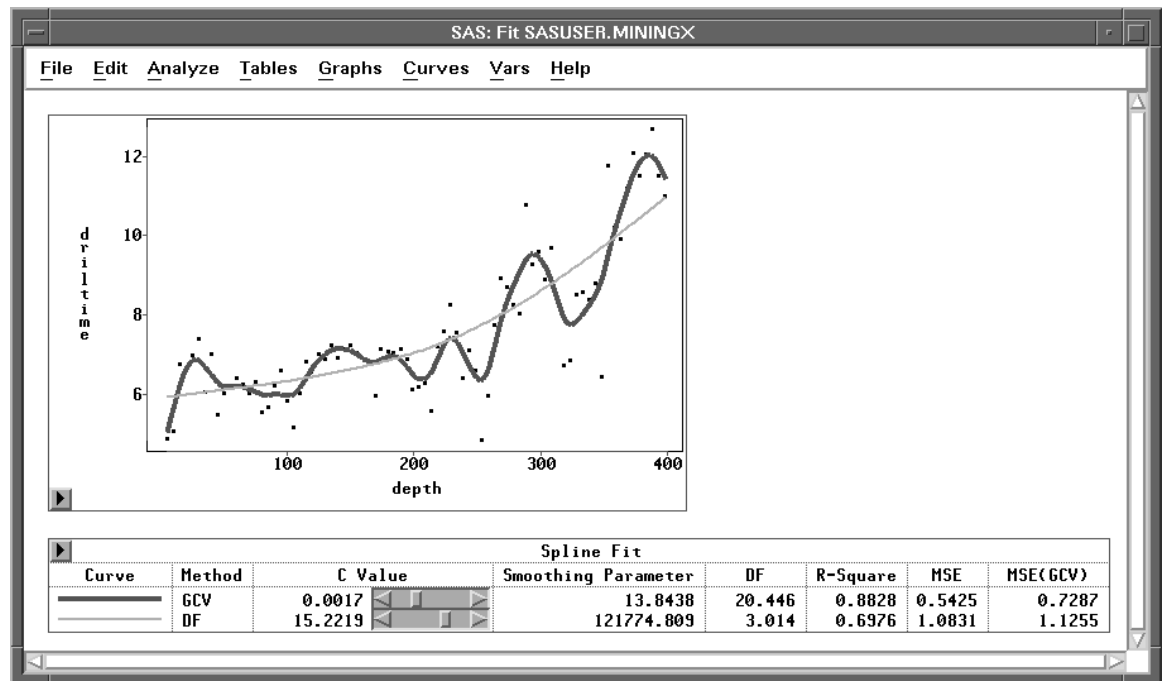


Figure 39.41. Smoothing Spline Estimates

Nonparametric Kernel Smoother

A *kernel estimator* uses an explicitly defined set of weights at each point x to produce the estimate at x . The kernel estimator of f has the form

$$\hat{f}_\lambda(x) = \sum_{i=1}^n W(x, x_i; \lambda) y_i$$

where $W(x, x_i; \lambda)$ is the weight function that depends on the smoothing parameter λ .

The weights are derived from a single function that is independent of the design

$$W(x, x_i; \lambda) = \frac{K_0\left(\frac{x-x_i}{\lambda}\right)}{\sum_{j=1}^n K_0\left(\frac{x-x_j}{\lambda}\right)}$$

where K_0 is a kernel function and λ is the bandwidth or smoothing parameter. The weights are nonnegative and sum to 1.

Symmetric probability density functions commonly used as kernel functions are

- *Normal* $K_0(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$ for $-\infty < t < \infty$
- *Triangular* $K_0(t) = \begin{cases} 1 - |t| & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- *Quadratic* $K_0(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$

You select a bandwidth λ for each kernel estimator by specifying c in the formula

$$\lambda = n^{-\frac{1}{5}} Q c$$

where Q is the sample interquartile range of the explanatory variable and n is the sample size. This formulation makes c independent of the units of \mathbf{X} .

SAS/INSIGHT software divides the range of the explanatory variable into 128 evenly spaced intervals, then approximates the data on this grid and uses the fast Fourier transformation (Silverman 1986) to estimate the kernel fit on this grid. For a data point x_i that lies between two grid points, a linear interpolation is used to compute the predicted value. A small value of λ (relative to the width of the interval) may give unstable estimates of the kernel fit.

After choosing **Curves:Kernel**, you specify a kernel and smoothing parameter selection method in the **Kernel Fit** dialog.

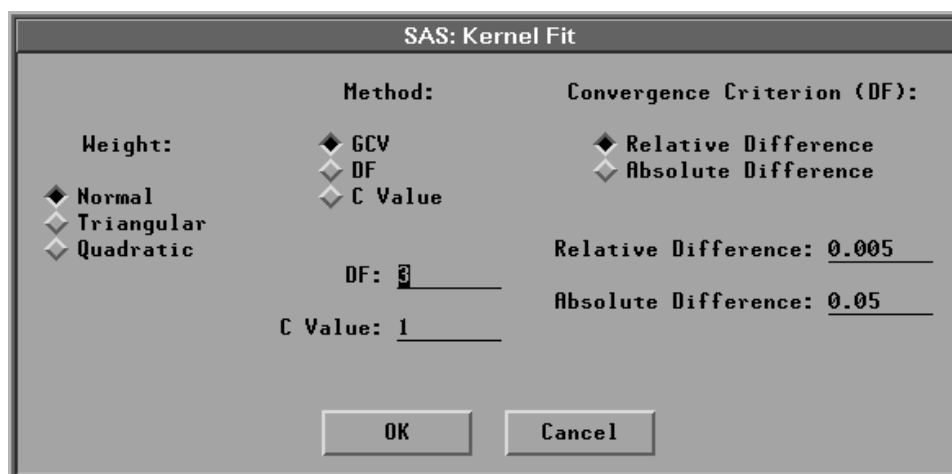


Figure 39.42. Kernel Fit Dialog

The default **Weight:Normal** uses a normal weight, and **Method:GCV** uses a c value that minimizes $MSE_{GCV}(\lambda)$. [Figure 39.43](#) illustrates normal kernel estimates with c values of 0.0944 (the GCV value) and 0.7546 (DF=3). Use the slider to change the c value of the kernel fit.

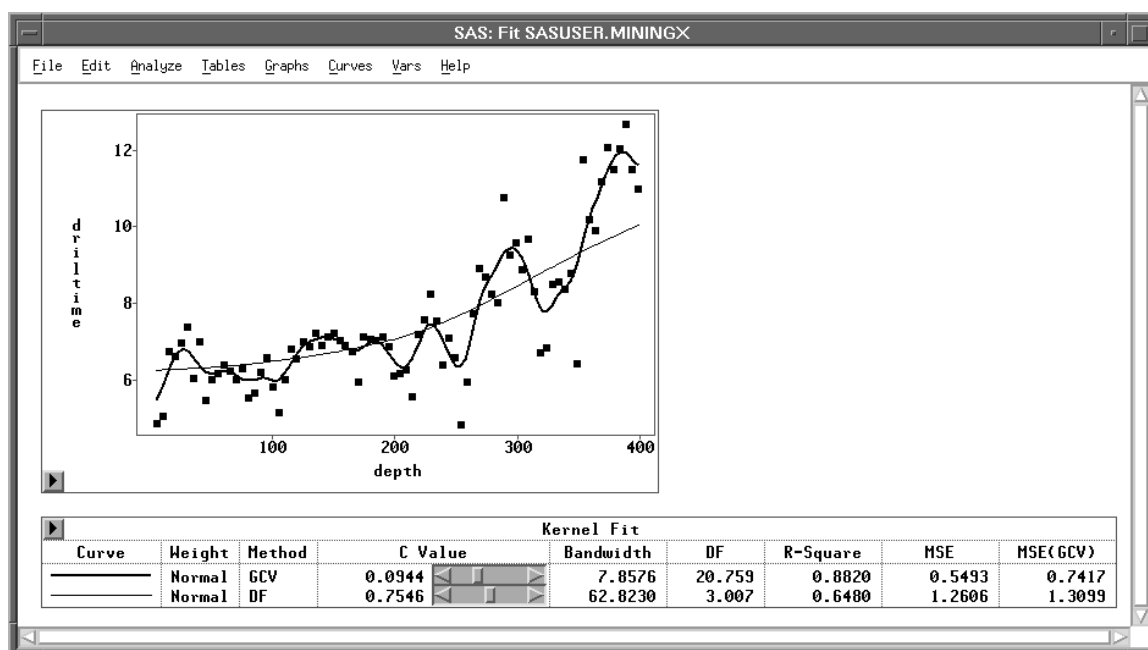


Figure 39.43. Kernel Estimates

Nonparametric Local Polynomial Smoother

The kernel estimator fits a local mean at each point x and thus cannot even estimate a line without bias (Cleveland, Cleveland, Devlin and Grosse 1988). An estimator based on locally-weighted regression lines or locally-weighted quadratic polynomials may give more satisfactory results.

A local polynomial smoother fits a locally-weighted regression at each point x to produce the estimate at x . Different types of regression and weight functions are used in the estimation.

SAS/INSIGHT software provides the following three types of regression:

- *Mean* a locally-weighted mean
- *Linear* a locally-weighted regression line
- *Quadratic* a locally-weighted quadratic polynomial regression

The weights are derived from a single function that is independent of the design

$$W(x, x_i; \lambda_i) = K_0\left(\frac{x - x_i}{\lambda_i}\right)$$

where K_0 is a weight function and λ_i is the local bandwidth at x_i .

SAS/INSIGHT software uses the following weight functions:

- *Normal* $K_0(t) = \begin{cases} \exp(-t^2/2) & \text{for } |t| \leq 3.5 \\ 0 & \text{otherwise} \end{cases}$
- *Triangular* $K_0(t) = \begin{cases} 1 - |t| & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- *Quadratic* $K_0(t) = \begin{cases} 1 - t^2 & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- *Tri - Cube* $K_0(t) = \begin{cases} (1 - |t|^3)^3 & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$

† **Note:** The normal weight function is proportional to a truncated normal density function.

SAS/INSIGHT software provides two methods to compute the local bandwidth λ_i . The loess estimator (Cleveland 1979; Cleveland, Devlin and Grosse 1988) evaluates λ_i based on the furthest distance from k nearest neighbors. A fixed bandwidth local polynomial estimator uses a constant bandwidth λ at each x_i .

For a loess estimator, you select k nearest neighbors by specifying a positive constant α . For $\alpha \leq 1$, k is αn truncated to an integer, where n is the number of observations. For $\alpha > 1$, k is set to n .

The local bandwidth λ_i is then computed as

$$\lambda_i = \begin{cases} d_{(k)}(x_i) & \text{for } 0 < \alpha \leq 1 \\ \alpha d_{(n)}(x_i) & \text{for } \alpha > 1 \end{cases}$$

where $d_{(k)}(x_i)$ is the furthest distance from x_i to its k nearest neighbors.

† **Note:** For $\alpha \leq 1$, the local bandwidth λ_i is a function of k and thus a step function of α .

For a fixed bandwidth local polynomial estimator, you select a bandwidth λ by specifying c in the formula

$$\lambda = n^{-\frac{1}{5}} Q c$$

where Q is the sample interquartile range of the explanatory variable and n is the sample size. This formulation makes c independent of the units of \mathbf{X} .

† **Note:** A fixed bandwidth local mean estimator is equivalent to a kernel smoother.

By default, SAS/INSIGHT software divides the range of the explanatory variable into 128 evenly spaced intervals, then it fits locally-weighted regressions on this grid. A small value of c or α may give the local polynomial fit to the data points near the grid points only and may not apply to the remaining points.

For a data point x_i that lies between two grid points $x_{i[j]} \leq x_i < x_{i[j+1]}$, the predicted value is the weighted average of the two predicted values at the two nearest grid points:

$$(1 - d_{ij})\hat{y}_{i[j]} + d_{ij}\hat{y}_{i[j+1]}$$

where $\hat{y}_{i[j]}$ and $\hat{y}_{i[j+1]}$ are the predicted values at the two nearest grid points and

$$d_{ij} = \frac{x_i - x_{i[j]}}{x_{i[j+1]} - x_{i[j]}}$$

A similar algorithm is used to compute the degrees of freedom of a local polynomial estimate, $df_\lambda = \text{trace}(\mathbf{H}_\lambda)$. The i th diagonal element of the matrix \mathbf{H}_λ is

$$(1 - d_{ij})h_{i[j]} + d_{ij}h_{i[j+1]}$$

where $h_{i[j]}$ and $h_{i[j+1]}$ are the i th diagonal elements of the projection matrices of the two regression fits.

After choosing **Curves:Loess** from the menu, you specify a loess fit in the **Loess Fit** dialog.

The dialog box is titled "SAS: Loess Fit". It contains the following settings:

- Type:**
 - ☐ Mean
 - ☒ Linear
 - ☐ Quadratic
- Method:**
 - ☒ GCV
 - ☐ DF
 - ☐ Alpha
- Number of Intervals:** 128
- Convergence Criterion (DF):**
 - ☒ Relative Difference
 - ☐ Absolute Difference
- Weight:**
 - ☐ Normal
 - ☐ Triangular
 - ☐ Quadratic
 - ☒ Tri-Cube
- DF:** 3
- Alpha:** 0.5
- Relative Difference:** 0.005
- Absolute Difference:** 0.05

At the bottom are "OK" and "Cancel" buttons.

Figure 39.44. Loess Fit Dialog

In the dialog, you can specify the number of intervals, the regression type, the weight function, and the method for choosing the smoothing parameter. The default **Type:Linear** uses a linear regression, **Weight:Tri-Cube** uses a tri-cube weight function, and **Method:GCV** uses an α value that minimizes $MSE_{GCV}(\lambda)$.

Figure 39.45 illustrates loess estimates with **Type=Linear**, **Weight=Tri-Cube**, and α values of 0.0930 (the GCV value) and 0.7795 (DF=3). Use the slider to change the α value of the loess fit.

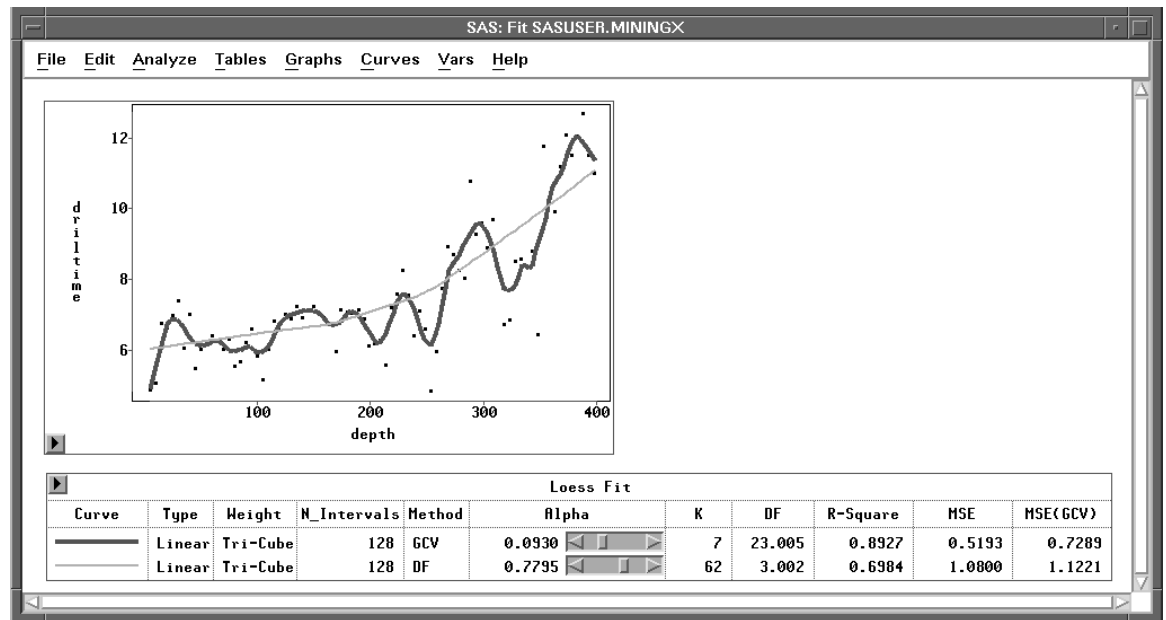


Figure 39.45. Loess Estimates

The loess degrees of freedom is a function of local bandwidth λ_i . For $\alpha \leq 1$, λ_i is a step function of α and thus the loess df is a step function of α . The convergence criterion applies only when the specified df is less than $df_{(\alpha=1)}$, the loess df for $\alpha = 1$. When the specified df is greater than $df_{(\alpha=1)}$, SAS/INSIGHT software uses the α value that has its df closest to the specified df .

Similarly, you can choose **Curves:Local Polynomial, Fixed Bandwidth** from the menu to specify a fixed bandwidth local polynomial fit.

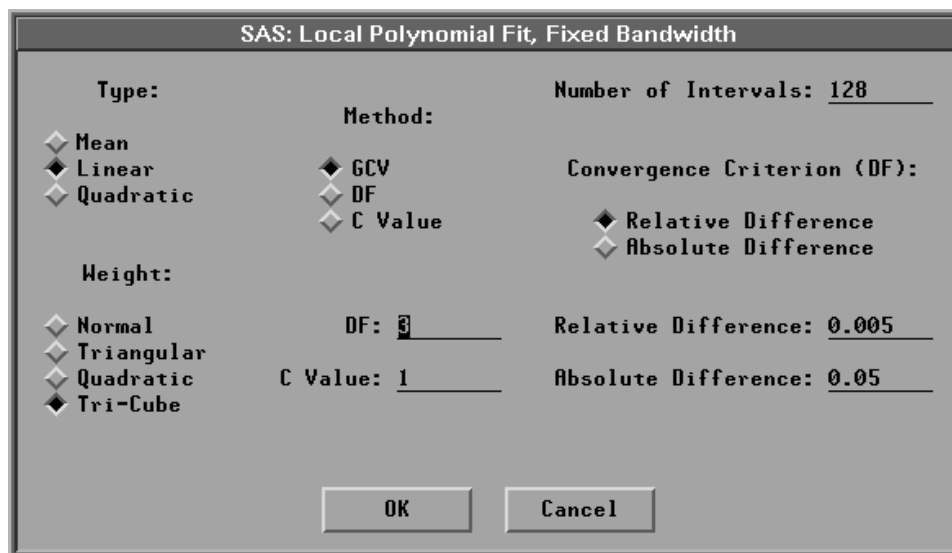


Figure 39.46. Fixed Bandwidth Local Polynomial Fit Dialog

Figure 39.47 illustrates fixed bandwidth local polynomial estimates with **Type=Linear**, **Weight=Tri-Cube**, and c values of 0.2026 (the GCV value) and 2.6505 (DF=3). Use the slider to change the c value of the local polynomial fit.

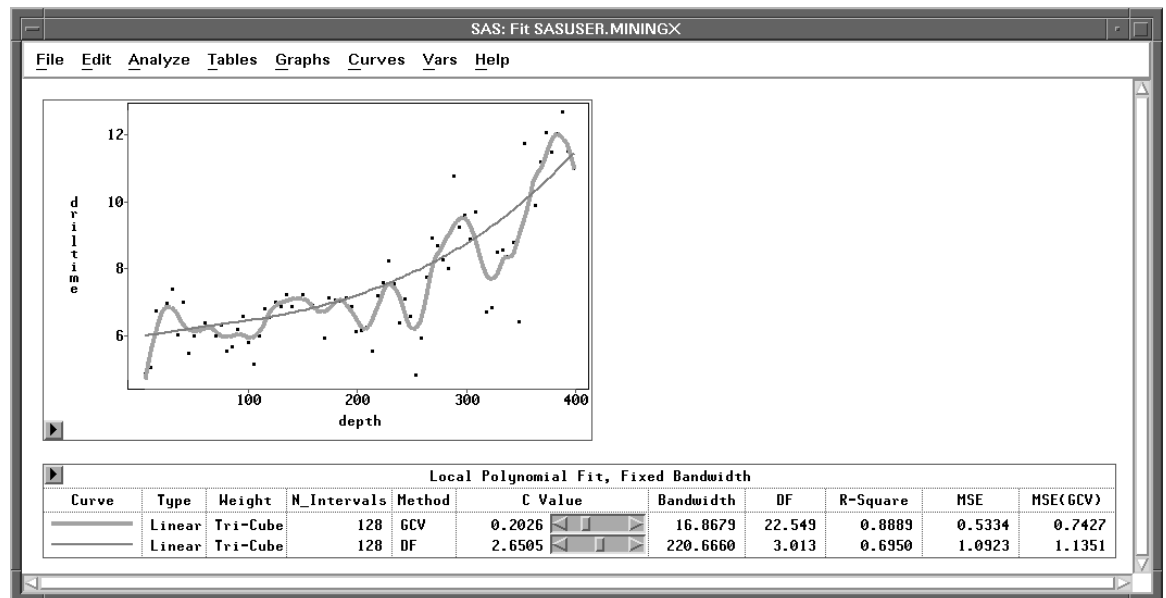


Figure 39.47. Fixed Bandwidth Local Polynomial Estimates

Output Variables

Output variables based on the model you fit can be saved in the data window. From the data window, you can store these variables in a SAS data set. This enables you, for example, to perform additional analyses using SAS/STAT software.

Axis variables in residual plots are automatically saved in the data window used to create the analysis. For example, when you create a residual-by-predicted plot, residual and predicted variables are always generated. These variables are deleted when you close the analysis window.

You can save variables permanently by using the fit output options dialog or the **Vars** menu shown in [Figure 39.48](#). Such variables remain stored in the data window after you close the analysis window.

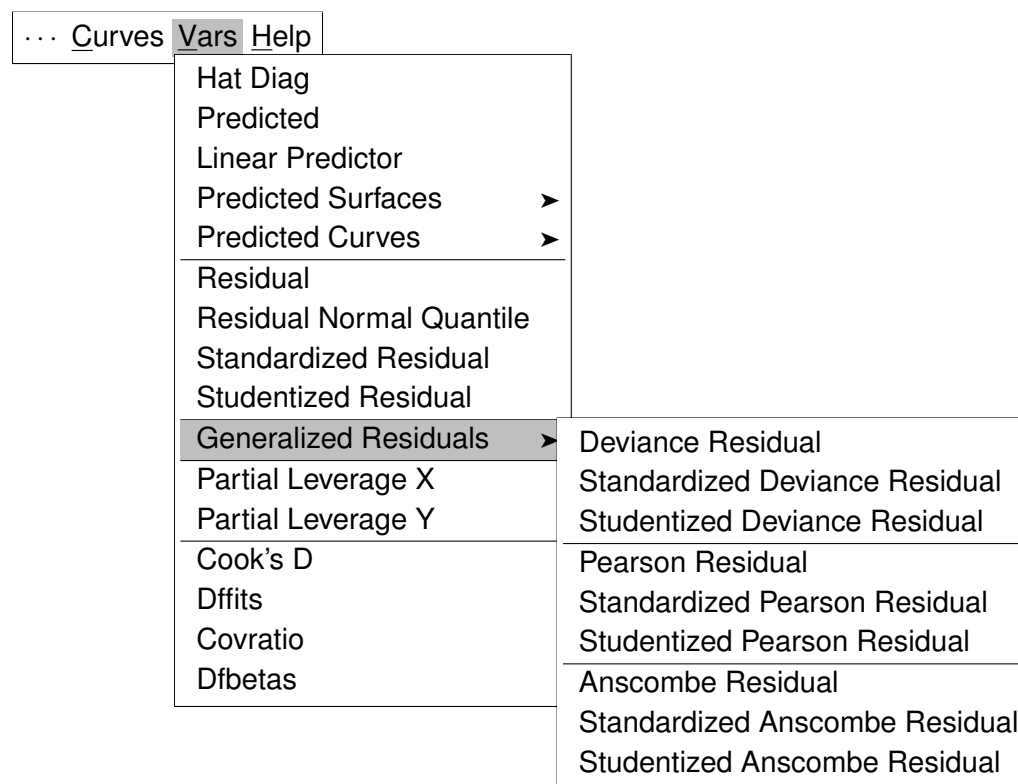


Figure 39.48. Vars Menu

SAS/INSIGHT software provides predicted and residual variables, a linear predictor, a residual normal quantile variable, partial leverage **X** and **Y** variables, and influence diagnostic variables.

Influence diagnostics are measures of the influence of each observation on the parameter estimates. These diagnostics include the hat diagonal values, standardized residuals, and studentized residuals. Cook's D, Dffits, Covratio, and Dfbetas also measure the effect of deleting observations.

Some influence diagnostics require a refit of the model after excluding each observation. For generalized linear models, numerical iterations are used for the fits, and

the process can be expensive. One-step methods are used to approximate these diagnostics after each fit. The process involves doing one iteration of the fit without the excluded observation, starting with the final parameter estimates and weights from the complete fit.

You can also create generalized residuals such as Pearson, deviance, and Anscombe residuals with generalized linear models. These residuals are applicable to the non-normal response distributions.

Generated variables use the naming conventions described later in this section. If a resulting variable name has more than 32 characters, only the first 32 characters are used. Generated variables also follow the same numbering convention as the analysis window when you create more than one fit analysis from the same data window. If the generated variable name is longer than 32 characters, the original variable name is truncated to the necessary length.

Hat Matrix Diagonal

Data points that are far from the centroid of the X -space are potentially influential. A measure of the distance between a data point, x_i , and the centroid of the X -space is the data point's associated diagonal element h_i in the hat matrix. Belsley, Kuh, and Welsch (1980) propose a cutoff of $2p/n$ for the diagonal elements of the hat matrix, where n is the number of observations used to fit the model, and p is the number of parameters in the model. Observations with h_i values above this cutoff should be investigated.

For linear models, the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

can be used as a projection matrix. The hat matrix diagonal variable contains the diagonal elements of the hat matrix

$$h_i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$$

For generalized linear models, an approximate projection matrix is given by

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$$

where $\mathbf{W} = \mathbf{W}_o$ when the full Hessian is used and $\mathbf{W} = \mathbf{W}_e$ when Fisher's scoring method is used.

The values of h_i are stored in a variable named **H_yname**, where **yname** is the response variable name.

Predicted Values

After the model has been fit, the predicted values are calculated from the estimated regression equation.

For linear models, the predicted mean vector of the n observation responses is

$$\hat{\mu} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{y}$$

$$\hat{\mu}_i = \mathbf{x}_i\mathbf{b}$$

For generalized linear models,

$$\hat{\mu}_i = g^{-1}(\eta_{0i} + \mathbf{x}_i\mathbf{b})$$

where η_{0i} is the offset for the i th observation.

The predicted values are stored in variables named **P_yname** for each response variable, where **yname** is the response variable name.

Linear Predictor

The *linear predictor* values are the linear function values, $\mathbf{x}_i\mathbf{b}$, in the predicted values. The linear predictor values are stored in variables named **LP_yname** for each response variable, where **yname** is the response variable name.

Residuals

The *residuals* are calculated as actual response minus predicted value,

$$r_i = y_i - \hat{\mu}_i$$

The residuals are stored in variables named **R_yname** for each response variable, where **yname** is the response variable name.

Residual Normal Quantiles

The *normal quantile* of the i th ordered residual is computed as

$$\Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right)$$

where Φ^{-1} is the inverse standard cumulative normal distribution.

If the residuals are normally distributed, the points on the residual normal quantile-quantile plot should lie approximately on a straight line with residual mean as the intercept and residual standard deviation as the slope.

The normal quantiles of the residuals are stored in variables named **RN_yname** for each response variable, where **yname** is the response variable name.

Predicted Surfaces

You can output predicted values from fitted kernel and thin-plate smoothing spline surfaces by choosing **Vars:Predicted Surfaces** from the menu.

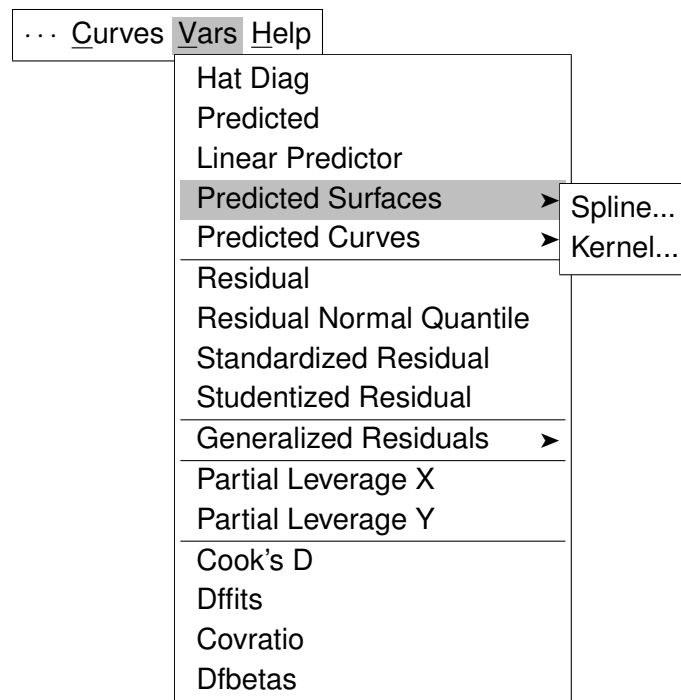


Figure 39.49. Predicted Surfaces Menu

For predicted values from a spline or kernel fit, you specify the surface fit in the dialogs, as shown in [Figure 39.28](#) or [Figure 39.30](#), respectively.

The predicted values for each response variable are stored in variables named **PS_yname** for spline and **PK_yname** for kernel, where **yname** is the response variable name.

Predicted Curves

You can output predicted values from fitted curves by choosing **Vars:Predicted Curves** from the menu.

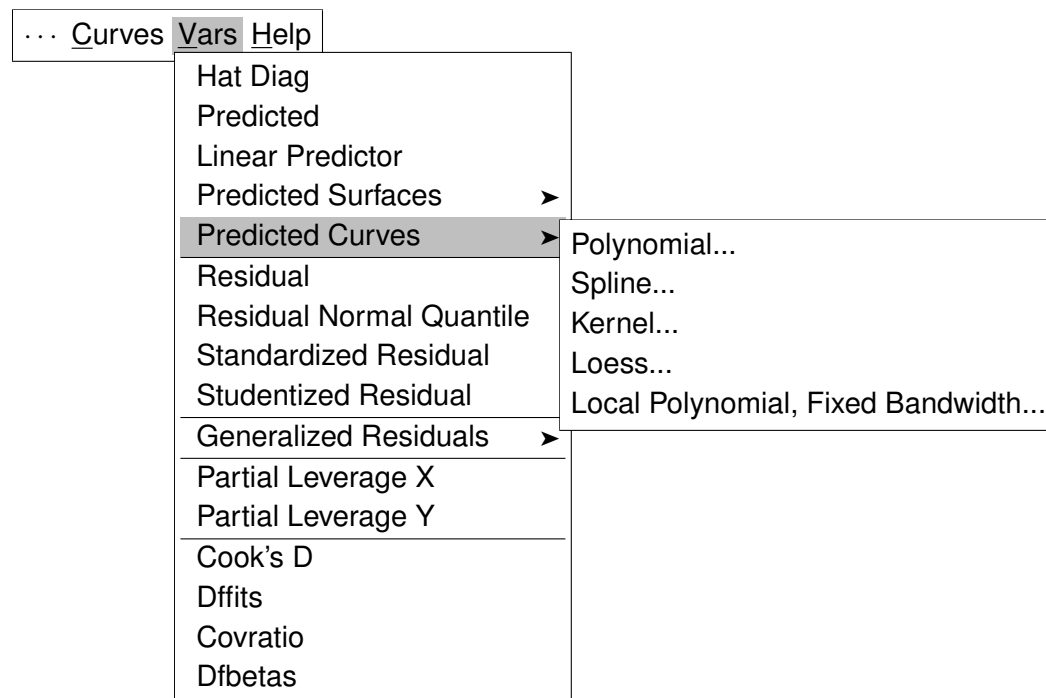


Figure 39.50. Predicted Curves Menu

After choosing **Vars:Predicted Curves:Polynomial** from the menu, you can specify the degree of polynomial in the **Polynomial Fit** dialog.

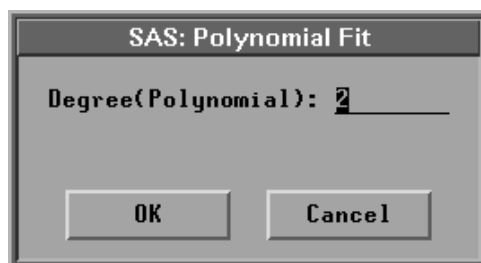


Figure 39.51. Predicted Polynomial Fit Dialog

For predicted values from a spline, kernel, loess, or fixed bandwidth local polynomial fit, you specify the curve fit in the dialogs, as shown in [Figure 39.40](#), [Figure 39.42](#), [Figure 39.44](#), or [Figure 39.46](#), respectively.

The predicted values for each response variable are stored in variables named **PP_yname** for polynomial, **PS_yname** for spline, **PK_yname** for kernel, and **PL_yname** for loess and fixed bandwidth local polynomial, where **yname** is the response variable name.

Standardized and Studentized Residuals

For linear models, the variance of the residual r_i is

$$\text{Var}(r_i) = \sigma^2(1 - h_i)$$

and an estimate of the standard error of the residual is

$$\text{STDERR}(r_i) = s\sqrt{1 - h_i}$$

Thus, the residuals can be modified to better detect unusual observations. The ratio of the residual to its standard error, called the *standardized residual*, is

$$r_{si} = \frac{r_i}{s\sqrt{1 - h_i}}$$

If the residual is standardized with an independent estimate of σ^2 , the result has a Student's t distribution if the data satisfy the normality assumption. If you estimate σ^2 by $s_{(i)}^2$, the estimate of σ^2 obtained after deleting the i th observation, the result is a studentized residual:

$$r_{ti} = \frac{r_i}{s_{(i)}\sqrt{1 - h_i}}$$

Observations with $|r_{ti}| > 2$ may deserve investigation.

For generalized linear models, the standardized and studentized residuals are

$$r_{si} = \frac{r_i}{\sqrt{\hat{\phi}(1 - h_i)}}$$

$$r_{ti} = \frac{r_i}{\sqrt{\hat{\phi}_{(i)}(1 - h_i)}}$$

where $\hat{\phi}$ is the estimate of the dispersion parameter ϕ , and $\hat{\phi}_{(i)}$ is a one-step approximation of ϕ after excluding the i th observation.

The standardized residuals are stored in variables named **RS_yname** and the Studentized residuals are stored in variables named **RT_yname** for each response variable, where **yname** is the response variable name.

Deviance Residuals

The *deviance residual* is the measure of deviance contributed from each observation and is given by

$$r_{Di} = \text{sign}(r_i) \sqrt{d_i}$$

where d_i is the individual deviance contribution.

The deviance residuals can be used to check the model fit at each observation for generalized linear models. These residuals are stored in variables named **RD_ynname** for each response variable, where **ynname** is the response variable name.

The standardized and studentized deviance residuals are

$$r_{Dsi} = \frac{r_{Di}}{\sqrt{\hat{\phi}(1 - h_i)}}$$

$$r_{Dti} = \frac{r_{Di}}{\sqrt{\hat{\phi}_{(i)}(1 - h_i)}}$$

The standardized deviance residuals are stored in variables named **RDS_ynname** and the studentized deviance residuals are stored in variables named **RDT_ynname** for each response variable, where **ynname** is the response variable name.

Pearson Residuals

The *Pearson residual* is the raw residual divided by the square root of the variance function $V(\mu)$.

The Pearson residual is the individual contribution to the Pearson χ^2 statistic. For a binomial distribution with m_i trials in the i th observation, it is defined as

$$r_{Pi} = \sqrt{m_i} \frac{r_i}{\sqrt{V(\hat{\mu}_i)}}$$

For other distributions, the Pearson residual is defined as

$$r_{Pi} = \frac{r_i}{\sqrt{V(\hat{\mu}_i)}}$$

The Pearson residuals can be used to check the model fit at each observation for generalized linear models. These residuals are stored in variables named **RP_ynname** for each response variable, where **ynname** is the response variable name.

The standardized and studentized Pearson residuals are

$$r_{Psi} = \frac{r_{Pi}}{\sqrt{\hat{\phi}(1 - h_i)}}$$

$$r_{Pti} = \frac{r_{Pi}}{\sqrt{\hat{\phi}_{(i)}(1 - h_i)}}$$

The standardized Pearson residuals are stored in variables named **RPS_ynname** and the studentized Pearson residuals are stored in variables named **RPT_ynname** for each response variable, where **ynname** is the response variable name.

Anscombe Residuals

For nonnormal response distributions in generalized linear models, the distribution of the Pearson residuals is often skewed. Anscombe proposed a residual using a function $A(y)$ in place of y in the residual derivation (Anscombe 1953, McCullagh and Nelder 1989). The function $A(y)$ is chosen to make the distribution of $A(y)$ as normal as possible and is given by

$$A(\mu) = \int_{-\infty}^{\mu} V^{-1/3}(t) dt$$

where $V(t)$ is the variance function.

For a binomial distribution with m_i trials in the i th observation, the *Anscombe residual* is defined as

$$r_{Ai} = \sqrt{m_i} \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i) \sqrt{V(\hat{\mu}_i)}}$$

For other distributions, the Anscombe residual is defined as

$$r_{Ai} = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i) \sqrt{V(\hat{\mu}_i)}}$$

where $A'(\mu)$ is the derivative of $A(\mu)$.

For the response distributions used in the fit analysis, Anscombe residuals are

Normal	$r_{Ai} = y_i - \hat{\mu}_i$
Inverse Gaussian	$r_{Ai} = (\log(y_i) - \log(\hat{\mu}_i)) / \hat{\mu}_i^{1/2}$
Gamma	$r_{Ai} = 3((y_i / \hat{\mu}_i)^{1/3} - 1)$
Poisson	$r_{Ai} = \frac{3}{2}(y_i^{2/3} \hat{\mu}_i^{-1/6} - \hat{\mu}_i^{1/2})$
Binomial	$r_{Ai} = \sqrt{m_i} \left(B(y_i, \frac{2}{3}, \frac{2}{3}) - B(\hat{\mu}_i, \frac{2}{3}, \frac{2}{3}) \right) (\hat{\mu}_i(1 - \hat{\mu}_i))^{-1/6}$ where $B(z, a, b) = \int_0^z t^{a-1}(1-t)^{b-1} dt$

You can save Anscombe residuals to your data set by using the **Output Variables** dialog, as shown in [Figure 39.5](#), or the **Vars** menu, as shown in [Figure 39.48](#). These residuals are stored in variables named **RA_ynname** for each response variable, where **ynname** is the response variable name.

The standardized and studentized Anscombe residuals are

$$r_{Asi} = \frac{r_{Ai}}{\sqrt{\hat{\phi}(1 - h_i)}}$$

$$r_{Ati} = \frac{r_{Ai}}{\sqrt{\hat{\phi}_{(i)}(1 - h_i)}}$$

where $\hat{\phi}$ is the estimate of the dispersion parameter ϕ , and $\hat{\phi}_{(i)}$ is a one-step approximation of ϕ after excluding the i th observation.

The standardized Anscombe residuals are stored in variables named **RAS_ynname** and the studentized Anscombe residuals are stored in variables named **RAT_ynname** for each response variable, where **ynname** is the response variable name.

Partial Leverage Variables

The *partial leverage output variables* are variables used in the partial leverage plots. For each interval **X** variable, the corresponding partial leverage **X** variable is named **X_xname**, where **xname** is the **X** variable name. For each pair of **Y** and **X** variables, the corresponding partial leverage **Y** variable is named **ynname_xname**, where **ynname** is the **Y** variable name and **xname** is the **X** variable name. Up to the first three characters of the response variable name are used to create the new variable name.

Cook's D

Cook's D measures the change in the parameter estimates caused by deleting each observation. For linear models,

$$D_i = \frac{1}{ps^2}(\mathbf{b} - \mathbf{b}_{(i)})'(\mathbf{X}'\mathbf{X})(\mathbf{b} - \mathbf{b}_{(i)})$$

where $\mathbf{b}_{(i)}$ is the vector of parameter estimates obtained after deleting the i th observation.

Cook (1977) suggests comparing D_i to the F distribution with p and $n - p$ degrees of freedom.

For generalized linear models,

$$D_i = \frac{1}{p\hat{\phi}}(\mathbf{b} - \mathbf{b}_{(i)})'(\mathbf{X}'\mathbf{W}\mathbf{X})(\mathbf{b} - \mathbf{b}_{(i)})$$

where $\mathbf{W} = \mathbf{W}_o$ when the full Hessian is used and $\mathbf{W} = \mathbf{W}_e$ when Fisher's scoring method is used.

Cook's D statistics are stored in variables named **D_yname** for each response variable, where **yname** is the response variable name.

Dffits

The *Dffits statistic* is a scaled measure of the change in the predicted value for the i th observation. For linear models,

$$F_i = \frac{\hat{\mu}_i - \hat{\mu}_{(i)}}{s_{(i)}\sqrt{h_i}}$$

where $\hat{\mu}_{(i)}$ is the i th value predicted without using the i th observation.

Large absolute values of F_i indicate influential observations. A general cutoff to consider is 2; a recommended size-adjusted cutoff is $2\sqrt{p/n}$.

For generalized linear models,

$$F_i = \frac{\hat{\mu}_i - \hat{\mu}_{(i)}}{\sqrt{\hat{\phi}_{(i)}h_i}}$$

The Dffits statistics are stored in variables named **F_yname** for each response variable, where **yname** is the response variable name.

Covratio

Covratio measures the effect of observations on the covariance matrix of the parameter estimates. For linear models,

$$C_i = \frac{|s_{(i)}^2(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}|}{|s^2(\mathbf{X}'\mathbf{X})^{-1}|}$$

where $\mathbf{X}_{(i)}$ is the \mathbf{X} matrix without the i th observation.

Values of C_i near 1 indicate that the observation has little effect on the precision of the estimates. Observations with $|C_i - 1| \geq 3p/n$ suggest a need for further investigation.

For generalized linear models,

$$C_i = \frac{|\hat{\phi}_{(i)}(\mathbf{X}'_{(i)}\mathbf{W}_{(i)}\mathbf{X}_{(i)})^{-1}|}{|\hat{\phi}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}|}$$

where $\mathbf{W}_{(i)}$ is the \mathbf{W} matrix without the i th observation, $\mathbf{W} = \mathbf{W}_o$ when the full Hessian is used, and $\mathbf{W} = \mathbf{W}_e$ when Fisher's scoring method is used.

The Covratio statistics are stored in variables named **C_ynname** for each response variable, where **ynname** is the response variable name.

Dfbetas

Dfbetas is a normalized measure of the effect of observations on the estimated regression coefficients. For linear models,

$$B_{j,i} = \frac{b_j - b_{j(i)}}{s_{(i)}\sqrt{(\mathbf{X}'\mathbf{X})_{jj}^{-1}}}$$

where $(\mathbf{X}'\mathbf{X})_{jj}^{-1}$ is the j th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Values of $B_{j,i} > 2$ indicate observations that are influential in estimating a given parameter. A recommended size-adjusted cutoff is $2/\sqrt{n}$.

For generalized linear models,

$$B_{j,i} = \frac{b_j - b_{j(i)}}{\sqrt{\hat{\phi}_{(i)}(\mathbf{X}'\mathbf{W}\mathbf{X})_{jj}^{-1}}}$$

where $\mathbf{W} = \mathbf{W}_o$ when the full Hessian is used and $\mathbf{W} = \mathbf{W}_e$ when the Fisher's scoring method is used.

The dfbetas statistics are stored in variables named **Bynname_xname** for each pair of response and explanatory variables, where **ynname** is the response variable name and **xname** is the explanatory variable name. Up to the first two characters of the response variable name are used to create the new variable name.

Weighted Analyses

If the errors ε_i do not have a common variance in the regression model

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

a weighted analysis may be appropriate. The observation weights are the values of the **Weight** variable you specified.

In parametric regression, the linear model is given by

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Let \mathbf{W} be an $n \times n$ diagonal matrix consisting of weights $w_1 > 0, w_2 > 0, \dots$, and $w_n > 0$ for the observations, and let $\mathbf{W}^{1/2}$ be an $n \times n$ diagonal matrix with diagonal elements $w_1^{1/2}, w_2^{1/2}, \dots$, and $w_n^{1/2}$.

The weighted fit analysis is equivalent to the usual (unweighted) fit analysis of the transformed model

$$\mathbf{y}^* = \mathbf{X}^*\beta + \epsilon^*$$

where $\mathbf{y}^* = \mathbf{W}^{1/2}\mathbf{y}$, $\mathbf{X}^* = \mathbf{W}^{1/2}\mathbf{X}$, and $\epsilon^* = \mathbf{W}^{1/2}\epsilon$.

The estimate of β is then given by

$$\mathbf{b}_w = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

For nonparametric weighted regression, the minimizing criterion in spline estimation is given by

$$S(\lambda) = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \{y_i - \hat{f}_\lambda(x_i)\}^2 + \lambda \int_{-\infty}^{\infty} \{\hat{f}_\lambda''(x)\}^2 dx$$

In kernel estimation, individual weights are

$$W(x, x_i; \lambda) = \frac{w_i K_0\left(\frac{x-x_i}{\lambda}\right)}{\sum_{j=1}^n w_j K_0\left(\frac{x-x_j}{\lambda}\right)}$$

For generalized linear models, the function $a_i(\phi) = \phi/(m_i w_i)$ for binomial distribution with m_i trials in the i th observation, $a_i(\phi) = \phi/w_i$ for other distributions. The function $a_i(\phi)$ is used to compute the likelihood function and the diagonal matrices \mathbf{W}_o and \mathbf{W}_e .

The individual deviance contribution d_i is obtained by multiplying the weight w_i by the unweighted deviance contribution. The deviance is the sum of these weighted deviance contributions.

The Pearson χ^2 statistic is

$$\chi^2 = \sum_{i=1}^n w_i m_i (y_i - \mu_i)^2 / V(\mu_i)$$

for binomial distribution with m_i trials in the i th observation,

$$\chi^2 = \sum_{i=1}^n w_i (y_i - \mu_i)^2 / V(\mu_i)$$

for other distributions.

References

- Anscombe, F.J. (1953), "Contribution to the Discussion of H. Hotelling's Paper," *Journal of the Royal Statistical Society, Series B*, 15, 229–230.
- Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics*, New York: John Wiley & Sons, Inc.
- Cleveland, W.S. (1979), "Robust Locally-Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.
- Cleveland, W.S., Devlin, S.J., and Grosse, E. (1988), "Regression by Local Fitting: Methods, Properties, and Computational Algorithms," *Journal of Econometrics*, 37, 87–114.
- Cleveland, W.S. and Grosse, E. (1991), "Computational Methods for Local Regression," *Journal of Statistics and Computing*, 1, 47–62.
- Cleveland, W.S. (1993), *Visualizing Data*, Summit, New Jersey: Hobart Press.
- Cook, R.D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.
- Cook, R.D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, New York: Chapman and Hall.
- Cox, D.R. and Snell, E.J. (1968), "A General Definition of Residuals," *Journal of the Royal Statistical Society, Series B*, 30, 248–275.
- Dobson, A.J. (1990), *An Introduction to Generalized Linear Models*, New York: Chapman and Hall.
- Eubank, R.L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker, Inc.
- Hastie, Y.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

- Hinkley, D.V., Reid, N., and Snell, E.J. (1991), *Statistical Theory and Modelling*, New York: Chapman and Hall.
- Hoaglin, D.C. and Welsch, R.E. (1978), “The Hat Matrix in Regression and ANOVA,” *The American Statistician*, 32, 17–22.
- Kvalseth, T.O. (1985), “Cautionary Note About R^2 ,” *The American Statistician*, 39, 279.
- McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman and Hall.
- Pringle, R.M. and Raynor, A.A. (1971), *Generalized Inverse Matrices with Applications to Statistics*, New York: Hafner Publishing Co.
- Rawlings, J.O. (1988), *Applied Regression Analysis: A Research Tool*, Pacific Grove, CA: Wadsworth & Brooks.
- Reinsch, C. (1967), “Smoothing by Spline Functions,” *Numerische Mathematik*, 10, 177–183.
- Silverman, B.W. (1982), “Kernel Density Estimation using the Fast Fourier Transform,” *Applied Statistics*, 31, 93–99.
- Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.
- Velleman, P.F. and Welsch, R.E. (1981), “Efficient Computing of Regression Diagnostics,” *The American Statistician*, 35, 234–242.
- Wahba, G. and Wendelberger, J.G. (1980), “Some New Mathematical Methods for Variational Objective Analysis using Splines and Cross Validation,” *Monthly Weather Review*, 108, 1122–1143.

Chapter 40

Multivariate Analyses

Chapter Contents

VARIABLES	708
METHOD	710
Principal Component Analysis	713
Principal Component Rotation	715
Canonical Correlation Analysis	717
Maximum Redundancy Analysis	718
Canonical Discriminant Analysis	718
OUTPUT	720
Principal Component Analysis	722
Principal Component Rotation	723
Canonical Correlation Analysis	724
Maximum Redundancy Analysis	725
Canonical Discriminant Analysis	726
TABLES	727
Univariate Statistics	727
Sums of Squares and Crossproducts	727
Corrected Sums of Squares and Crossproducts	728
Covariance Matrix	728
Correlation Matrix	729
P-Values of the Correlations	729
Inverse Correlation Matrix	731
Pairwise Correlations	732
Principal Component Analysis	733
Principal Components Rotation	737
Canonical Correlation Analysis	740
Maximum Redundancy Analysis	745
Canonical Discriminant Analysis	749
GRAPHS	753
Scatter Plot Matrix	753
Principal Component Plots	754
Component Rotation Plots	758
Canonical Correlation Plots	760
Maximum Redundancy Plots	763

Reference ♦ Multivariate Analyses

Canonical Discrimination Plots	765
CONFIDENCE ELLIPSES	768
Scatter Plot Confidence Ellipses	769
Canonical Discriminant Confidence Ellipses	770
OUTPUT VARIABLES	771
Principal Components	772
Principal Component Rotation	772
Canonical Variables	772
Maximum Redundancy	772
Canonical Discriminant	772
WEIGHTED ANALYSES	773
REFERENCES	774

Chapter 40

Multivariate Analyses

Choosing **Analyze:Multivariate (Y X)** gives you access to a variety of *multivariate analyses*. These provide methods for examining relationships among variables and between two sets of variables.

You can calculate correlation matrices and scatter plot matrices with confidence ellipses to explore relationships among pairs of variables. You can use principal component analysis to examine relationships among several variables, canonical correlation analysis and maximum redundancy analysis to examine relationships between two sets of interval variables, and canonical discriminant analysis to examine relationships between a nominal variable and a set of interval variables.

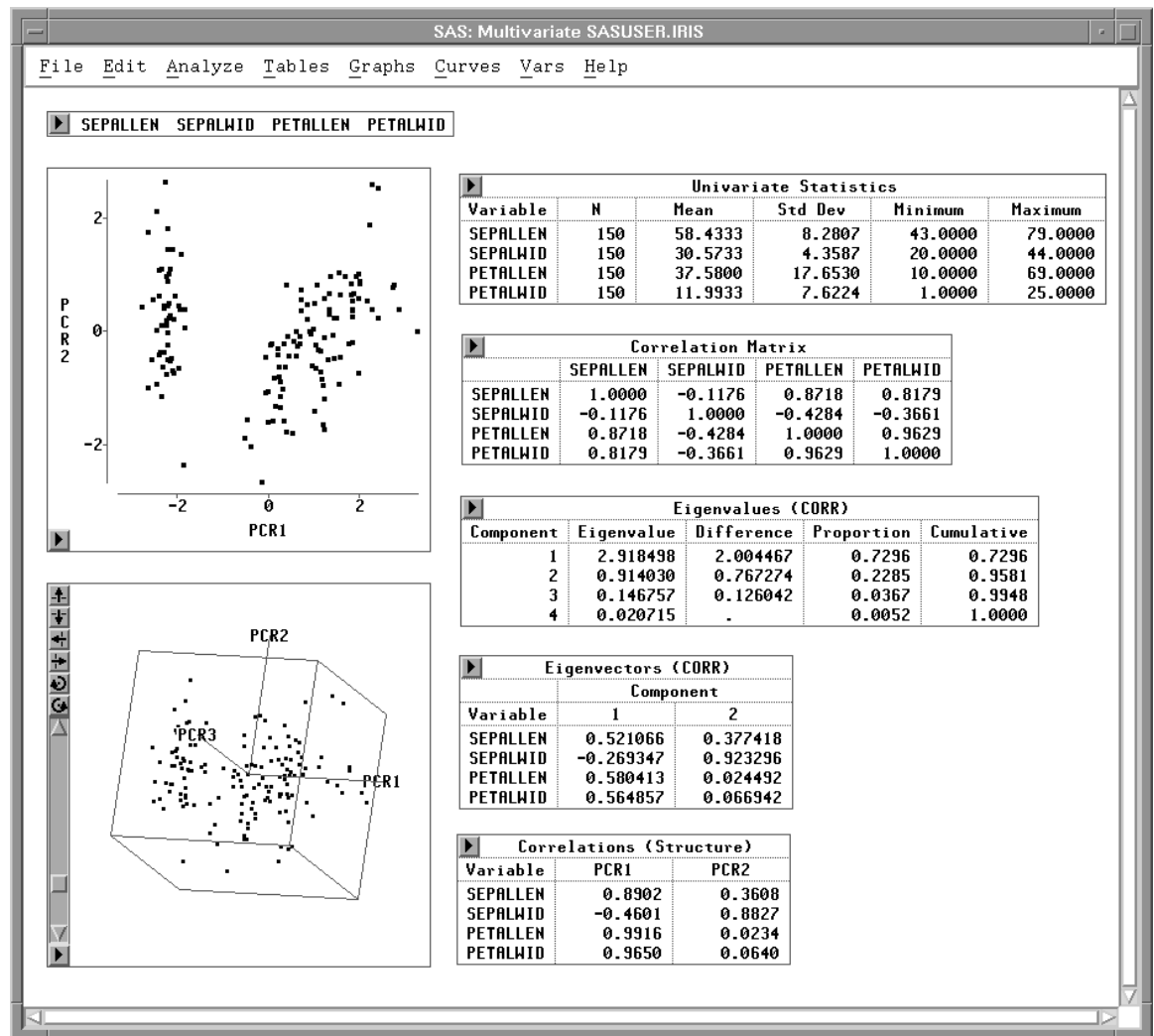


Figure 40.1. Multivariate Analysis

Variables

To create a multivariate analysis, choose **Analyze:Multivariate (Y's)**. If you have already selected one or more interval variables, these selected variables are treated as **Y** variables and a multivariate analysis for the variables appears. If you have not selected any variables, a variables dialog appears.

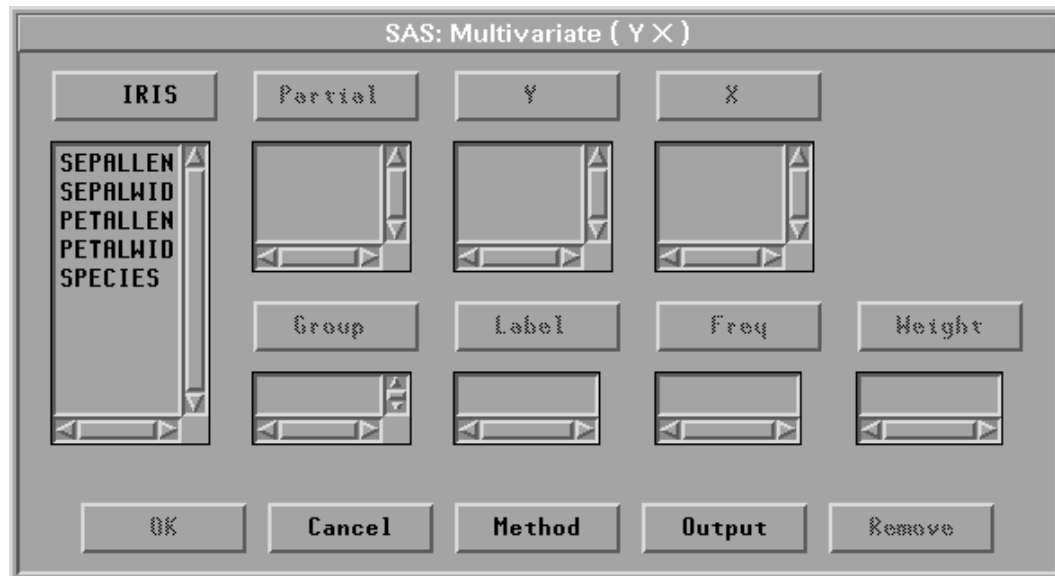


Figure 40.2. Multivariate Variables Dialog

Select at least one **Y** variable. With canonical correlation analysis and maximum redundancy analysis, you need to select a set of **X** variables. With canonical discriminant analysis, you need to select a nominal **Y** variable and a set of **X** variables.

Without **X** variables, sums of squares and crossproducts, corrected sums of squares and crossproducts, covariances, and correlations are displayed as symmetric matrices with **Y** variables as both the row variables and the column variables. With a nominal **Y** variable, these statistics are displayed as symmetric matrices with **X** variables as both the row variables and the column variables. When both interval **Y** variables and interval **X** variables are selected, these statistics are displayed as rectangular matrices with **Y** variables as the row variables and **X** variables as the column variables.

You can select one or more **Partial** variables. The multivariate analysis analyzes **Y** and **X** variables using their residuals after partialling out the **Partial** variables.

You can select one or more **Group** variables, if you have grouped data. This creates one multivariate analysis for each group. You can select a **Label** variable to label observations in the plots.

You can select a **Freq** variable. If you select a **Freq** variable, each observation is assumed to represent n_i observations, where n_i is the value of the **Freq** variable.

You can select a **Weight** variable to specify relative weights for each observation in the analysis. The details of weighted analyses are explained in the “[Method](#)” section, which follows, and the “[Weighted Analyses](#)” section at the end of this chapter.

Method

Observations with missing values for any of the **Partial** variables are not used. Observations with **Weight** or **Freq** values that are missing or that are less than or equal to 0 are not used. Only the integer part of **Freq** values is used.

Observations with missing values for **Y** or **X** variables are not used in the analysis except for the computation of pairwise correlations. Pairwise correlations are computed from all observations that have nonmissing values for any pair of variables.

The following notation is used in this chapter:

- n is the number of nonmissing observations.
- n_p , n_y , and n_x are the numbers of **Partial**, **Y**, and **X** variables.
- d is the variance divisor.
- w_i is the i th observation weight (values of the **Weight** variable).
- \mathbf{y}_i and \mathbf{x}_i are the i th observed nonmissing **Y** and **X** vectors.
- $\bar{\mathbf{y}}$ and $\bar{\mathbf{x}}$ are the sample mean vectors, $\sum_{i=1}^n \mathbf{y}_i/n$, $\sum_{i=1}^n \mathbf{x}_i/n$.

The sums of squares and crossproducts of the variables are

- $\mathbf{U}_{yy} = \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i'$
- $\mathbf{U}_{yx} = \sum_{i=1}^n \mathbf{y}_i \mathbf{x}_i'$
- $\mathbf{U}_{xx} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$

The corrected sums of squares and crossproducts of the variables are

- $\mathbf{C}_{yy} = \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$
- $\mathbf{C}_{yx} = \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{x}_i - \bar{\mathbf{x}})'$
- $\mathbf{C}_{xx} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$

If you select a **Weight** variable, the sample mean vectors are

$$\bar{\mathbf{y}} = \sum_{i=1}^n w_i \mathbf{y}_i / \sum_{i=1}^n w_i \quad \bar{\mathbf{x}} = \sum_{i=1}^n w_i \mathbf{x}_i / \sum_{i=1}^n w_i$$

The sums of squares and crossproducts with a **Weight** variable are

- $\mathbf{U}_{yy} = \sum_{i=1}^n w_i \mathbf{y}_i \mathbf{y}_i'$
- $\mathbf{U}_{yx} = \sum_{i=1}^n w_i \mathbf{y}_i \mathbf{x}_i'$
- $\mathbf{U}_{xx} = \sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i'$

The corrected sums of squares and crossproducts with a **Weight** variable are

- $C_{yy} = \sum_{i=1}^n w_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$
- $C_{yx} = \sum_{i=1}^n w_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{x}_i - \bar{\mathbf{x}})'$
- $C_{xx} = \sum_{i=1}^n w_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$

The covariance matrices are computed as

$$S_{yy} = C_{yy}/d \quad S_{yx} = C_{yx}/d \quad S_{xx} = C_{xx}/d$$

To view or change the variance divisor d used in the calculation of variances and covariances, or to view or change other method options in the multivariate analysis, click on the **Method** button from the variables dialog to display the method options dialog.

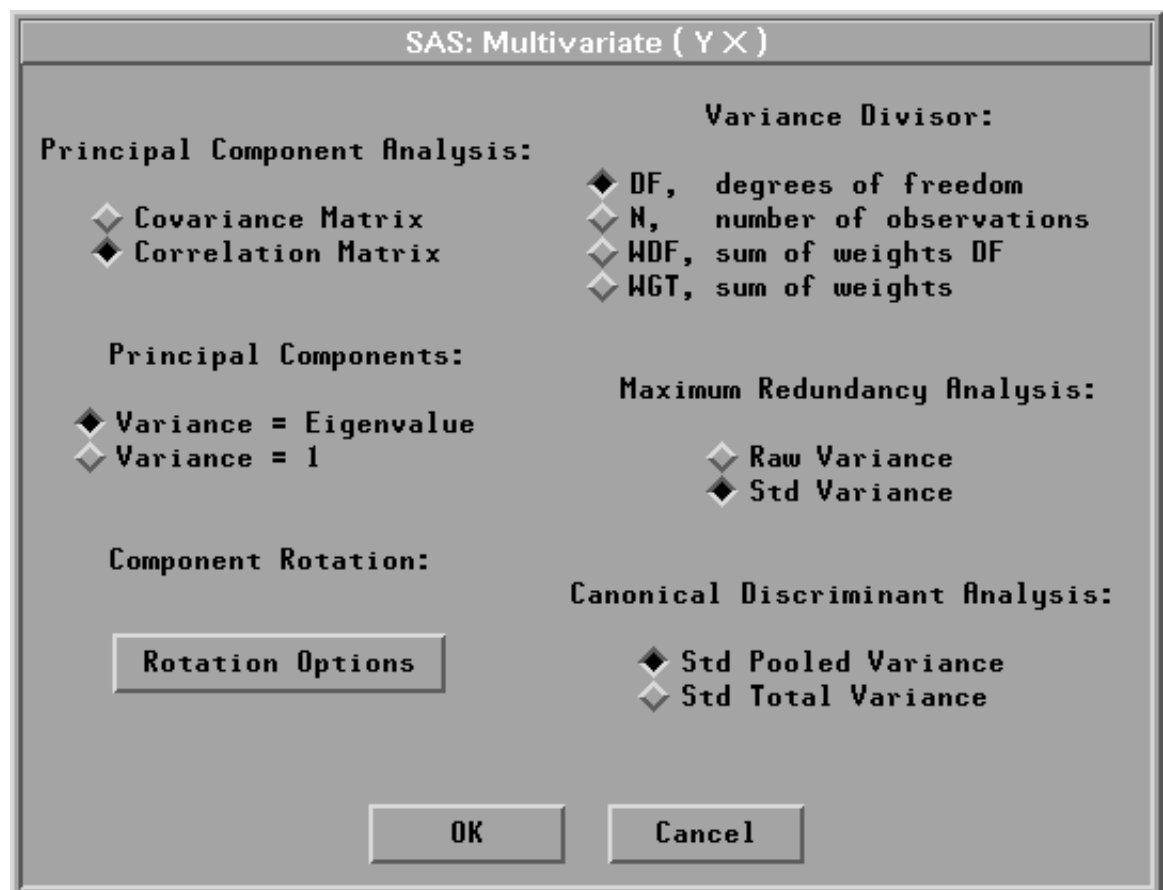


Figure 40.3. Multivariate Method Options Dialog

The variance divisor d is defined as

Reference ♦ *Multivariate Analyses*

- $d = n - n_p - 1$ for vardef=**DF**, degrees of freedom
- $d = n$ for vardef=**N**, number of observations
- $d = \sum_i w_i - n_p - 1$ for vardef=**WDF**, sum of weights minus number of partial variables minus 1
- $d = \sum_i w_i$ for vardef=**WGT**, sum of weights

By default, SAS/INSIGHT software uses **DF, degrees of freedom**.

The correlation matrices \mathbf{R}_{yy} , \mathbf{R}_{yx} , and \mathbf{R}_{xx} , containing the Pearson product-moment correlations of the variables, are derived by scaling their corresponding covariance matrices:

- $\mathbf{R}_{yy} = \mathbf{D}_{yy}^{-1} \mathbf{S}_{yy} \mathbf{D}_{yy}^{-1}$
- $\mathbf{R}_{yx} = \mathbf{D}_{yy}^{-1} \mathbf{S}_{yx} \mathbf{D}_{xx}^{-1}$
- $\mathbf{R}_{xx} = \mathbf{D}_{xx}^{-1} \mathbf{S}_{xx} \mathbf{D}_{xx}^{-1}$

where \mathbf{D}_{yy} and \mathbf{D}_{xx} are diagonal matrices whose diagonal elements are the square roots of the diagonal elements of \mathbf{S}_{yy} and \mathbf{S}_{xx} :

- $\mathbf{D}_{yy} = (\text{diag}(\mathbf{S}_{yy}))^{1/2}$
- $\mathbf{D}_{xx} = (\text{diag}(\mathbf{S}_{xx}))^{1/2}$

Principal Component Analysis

Principal component analysis was originated by Pearson (1901) and later developed by Hotelling (1933). It is a multivariate technique for examining relationships among several quantitative variables. Principal component analysis can be used to summarize data and detect linear relationships. It can also be used for exploring polynomial relationships and for multivariate outlier detection (Gnanadesikan 1997).

Principal component analysis reduces the dimensionality of a set of data while trying to preserve the structure. Given a data set with n_y \mathbf{Y} variables, n_y eigenvalues and their associated eigenvectors can be computed from its covariance or correlation matrix. The eigenvectors are standardized to unit length.

The principal components are linear combinations of the \mathbf{Y} variables. The coefficients of the linear combinations are the eigenvectors of the covariance or correlation matrix. Principal components are formed as follows:

- The first principal component is the linear combination of the \mathbf{Y} variables that accounts for the greatest possible variance.
- Each subsequent principal component is the linear combination of the \mathbf{Y} variables that has the greatest possible variance and is uncorrelated with the previously defined components.

For a covariance or correlation matrix, the sum of its eigenvalues equals the *trace* of the matrix, that is, the sum of the variances of the n_y variables for a covariance matrix, and n_y for a correlation matrix. The principal components are sorted by descending order of their variances, which are equal to the associated eigenvalues.

Principal components can be used to reduce the number of variables in statistical analyses. Different methods for selecting the number of principal components to retain have been suggested. One simple criterion is to retain components with associated eigenvalues greater than the average eigenvalue (Kaiser 1958). SAS/INSIGHT software offers this criterion as an option for selecting the numbers of eigenvalues, eigenvectors, and principal components in the analysis.

Principal components have a variety of useful properties (Rao 1964; Kshirsagar 1972):

- The eigenvectors are orthogonal, so the principal components represent jointly perpendicular directions through the space of the original variables.
- The principal component scores are jointly uncorrelated. Note that this property is quite distinct from the previous one.
- The first principal component has the largest variance of any unit-length linear combination of the observed variables. The j th principal component has the largest variance of any unit-length linear combination orthogonal to the first $j - 1$ principal components. The last principal component has the smallest variance of any linear combination of the original variables.
- The scores on the first j principal components have the highest possible generalized variance of any set of unit-length linear combinations of the original variables.
- In geometric terms, the j -dimensional linear subspace spanned by the first j principal components gives the best possible fit to the data points as measured by the sum of squared perpendicular distances from each data point to the subspace. This is in contrast to the geometric interpretation of least squares regression, which minimizes the sum of squared vertical distances. For example, suppose you have two variables. Then, the first principal component minimizes the sum of squared perpendicular distances from the points to the first principal axis. This is in contrast to least squares, which would minimize the sum of squared vertical distances from the points to the fitted line.

SAS/INSIGHT software computes principal components from either the correlation or the covariance matrix. The covariance matrix can be used when the variables are measured on comparable scales. Otherwise, the correlation matrix should be used. The new variables with principal component scores have variances equal to corresponding eigenvalues (**Variance=Eigenvalues**) or one (**Variance=1**). You specify the computation method and type of output components in the method options dialog, as shown in [Figure 40.3](#). By default, SAS/INSIGHT software uses the correlation matrix with new variable variances equal to corresponding eigenvalues.

Principal Component Rotation

Orthogonal transformations can be used on principal components to obtain factors that are more easily interpretable. The principal components are uncorrelated with each other, the rotated principal components are also uncorrelated after an orthogonal transformation. Different orthogonal transformations can be derived from maximizing the following quantity with respect to γ :

$$\sum_{j=1}^{n_f} \left(\sum_{i=1}^{n_y} b_{ij}^4 - \frac{\gamma}{n_y} \left(\sum_{i=1}^{n_y} b_{ij}^2 \right)^2 \right)$$

where n_f is the specified number of principal components to be rotated (number of factors), $b_{ij}^2 = r_{ij}^2 / \sum_{k=1}^{n_f} r_{ik}^2$, and r_{ij} is the correlation between the i th **Y** variable and the j th principal component.

SAS/INSIGHT software uses the following orthogonal transformations:

Equamax	$\gamma = \frac{n_f}{2}$
Orthomax	γ
Parsimax	$\gamma = \frac{n_y(n_f-1)}{(n_y+n_f-2)}$
Quartimax	$\gamma = 0$
Varimax	$\gamma = 1$

To view or change the principal components rotation options, click on the **Rotation Options** button in the method options dialog shown in [Figure 40.3](#) to display the Rotation Options dialog.

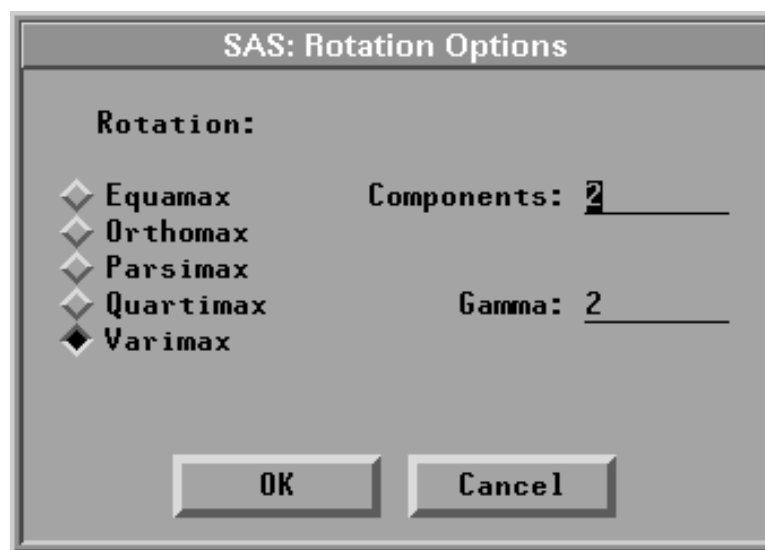


Figure 40.4. Rotation Options Dialog

Reference ♦ *Multivariate Analyses*

You can specify the type of rotation and number of principal components to be rotated in the dialog. By default, SAS/INSIGHT software uses **Varimax** rotation on the first two components. If you specify **Orthomax**, you also need to enter the γ value for the rotation in the **Gamma:** field.

Canonical Correlation Analysis

Canonical correlation was developed by Hotelling (1935, 1936). Its application is discussed by Cooley and Lohnes (1971), Kshirsagar (1972), and Mardia, Kent, and Bibby (1979). It is a technique for analyzing the relationship between two sets of variables. Each set can contain several variables. Multiple and simple correlation are special cases of canonical correlation in which one or both sets contain a single variable, respectively.

Given two sets of variables, canonical correlation analysis finds a linear combination from each set, called a canonical variable, such that the correlation between the two canonical variables is maximized. This correlation between the two canonical variables is the first canonical correlation. The coefficients of the linear combinations are canonical coefficients or canonical weights. It is customary to normalize the canonical coefficients so that each canonical variable has a variance of 1.

The first canonical correlation is at least as large as the multiple correlation between any variable and the opposite set of variables. It is possible for the first canonical correlation to be very large while all the multiple correlations for predicting one of the original variables from the opposite set of canonical variables are small.

Canonical correlation analysis continues by finding a second set of canonical variables, uncorrelated with the first pair, that produces the second highest correlation coefficient. The process of constructing canonical variables continues until the number of pairs of canonical variables equals the number of variables in the smaller group.

Each canonical variable is uncorrelated with all the other canonical variables of either set except for the one corresponding canonical variable in the opposite set. The canonical coefficients are not generally orthogonal, however, so the canonical variables do not represent jointly perpendicular directions through the space of the original variables.

The canonical correlation analysis includes tests of a series of hypotheses that each canonical correlation and all smaller canonical correlations are zero in the population. SAS/INSIGHT software uses an F approximation (Rao 1973; Kshirsagar 1972) that gives better small sample results than the usual χ^2 approximation. At least one of the two sets of variables should have an approximately multivariate normal distribution in order for the probability levels to be valid.

Canonical redundancy analysis (Stewart and Love 1968; Cooley and Lohnes 1971; van den Wollenberg 1977) examines how well the original variables can be predicted from the canonical variables. The analysis includes the proportion and cumulative proportion of the variance of the set of \mathbf{Y} and the set of \mathbf{X} variables explained by their own canonical variables and explained by the opposite canonical variables. Either raw or standardized variance can be used in the analysis.

Maximum Redundancy Analysis

In contrast to canonical redundancy analysis, which examines how well the original variables can be predicted from the canonical variables, maximum redundancy analysis finds the variables that can best predict the original variables.

Given two sets of variables, maximum redundancy analysis finds a linear combination from one set of variables that best predicts the variables in the opposite set. SAS/INSIGHT software normalizes the coefficients of the linear combinations so that each maximum redundancy variable has a variance of 1.

Maximum redundancy analysis continues by finding a second maximum redundancy variable from one set of variables, uncorrelated with the first one, that produces the second highest prediction power for the variables in the opposite set. The process of constructing maximum redundancy variables continues until the number of maximum redundancy variables equals the number of variables in the smaller group.

Either raw variances (**Raw Variance**) or standardized variances (**Std Variance**) can be used in the analysis. You specify the selection in the method options dialog as shown in [Figure 40.3](#). By default, standardized variances are used.

Canonical Discriminant Analysis

Canonical discriminant analysis is a dimension-reduction technique related to principal component analysis and canonical correlation. Given a classification variable and several interval variables, canonical discriminant analysis derives *canonical variables* (linear combinations of the interval variables) that summarize between-class variation in much the same way that principal components summarize total variation.

Given two or more groups of observations with measurements on several interval variables, canonical discriminant analysis derives a linear combination of the variables that has the highest possible multiple correlation with the groups. This maximal multiple correlation is called the first canonical correlation. The coefficients of the linear combination are the canonical coefficients or canonical weights. The variable defined by the linear combination is the first canonical variable or canonical component. The second canonical correlation is obtained by finding the linear combination uncorrelated with the first canonical variable that has the highest possible multiple correlation with the groups. The process of extracting canonical variables can be repeated until the number of canonical variables equals the number of original variables or the number of classes minus one, whichever is smaller.

The first canonical correlation is at least as large as the multiple correlation between the groups and any of the original variables. If the original variables have high within-group correlations, the first canonical correlation can be large even if all the multiple correlations are small. In other words, the first canonical variable can show substantial differences among the classes, even if none of the original variables does.

For each canonical correlation, canonical discriminant analysis tests the hypothesis that it and all smaller canonical correlations are zero in the population. An F approximation is used that gives better small-sample results than the usual χ^2 approximation. The variables should have an approximate multivariate normal distribution within each class, with a common covariance matrix in order for the probability levels to be valid.

The new variables with canonical variable scores in canonical discriminant analysis have either pooled within-class variances equal to one (**Std Pooled Variance**) or total-sample variances equal to one (**Std Total Variance**). You specify the selection in the method options dialog as shown in [Figure 40.3](#). By default, canonical variable scores have pooled within-class variances equal to one.

Output

To view or change the output options associated with your multivariate analysis, click on the **Output** button from the variables dialog. This displays the output options dialog.

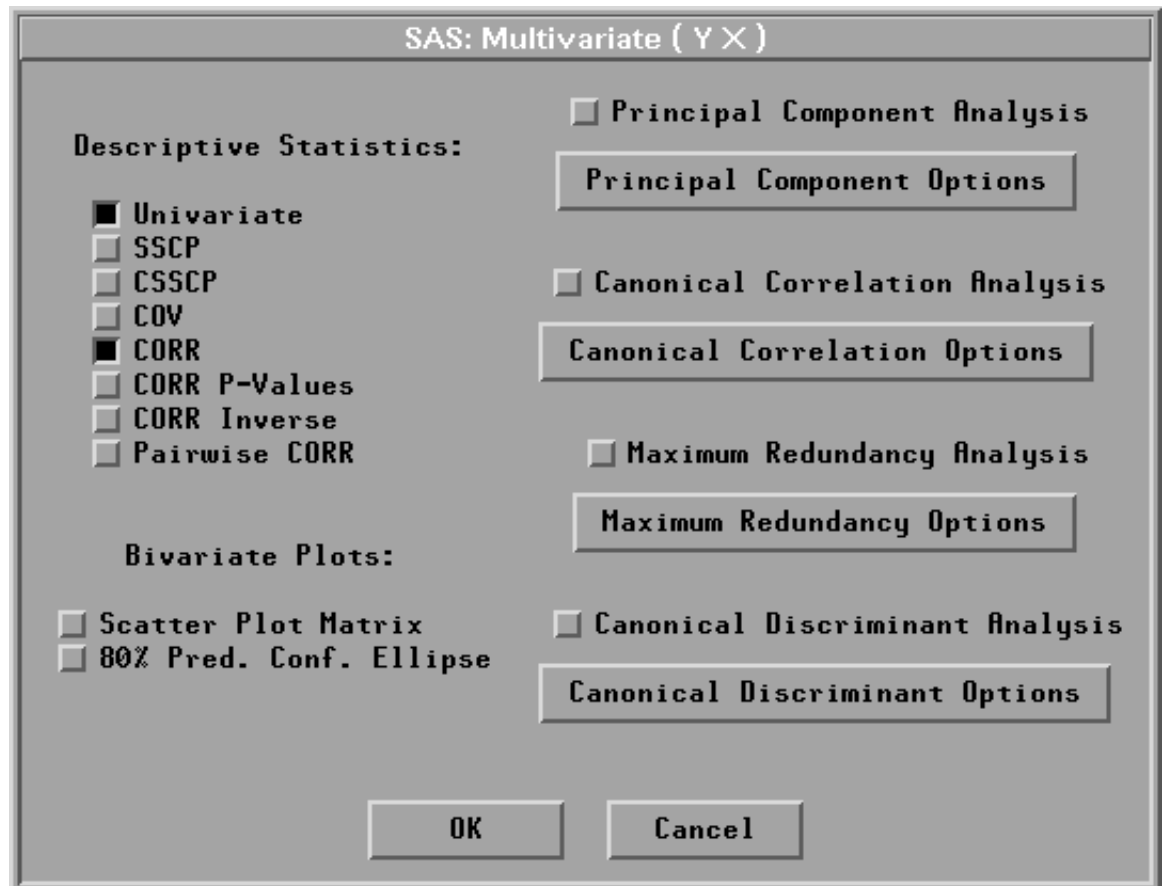


Figure 40.5. Multivariate Output Options Dialog

The options you set in this dialog determine which tables and graphs appear in the multivariate window. SAS/INSIGHT software provides univariate statistics and correlation matrix tables by default.

Descriptive statistics provide tables for examining the relationships among quantitative variables from univariate, bivariate, and multivariate perspectives.

Plots can be more informative than tables when you are trying to understand multivariate data. You can display a matrix of scatter plots for the analyzing variables. You can also add a bivariate confidence ellipse for mean or predicted values to the scatter plots. Using the confidence ellipses assumes each pair of variables has a bivariate normal distribution.

With appropriate variables chosen, you can generate principal component analysis (interval Y variables), canonical correlation analysis (interval Y, X variables), maxi-

mum redundancy analysis (interval Y, X variables), and canonical discriminant analysis (one nominal Y variable, interval X variables) by selecting the corresponding checkbox in the Output Options dialog.

Principal Component Analysis

Clicking the **Principal Component Options** button in the Output Options dialog shown in [Figure 40.5](#) displays the dialog shown in [Figure 40.6](#).

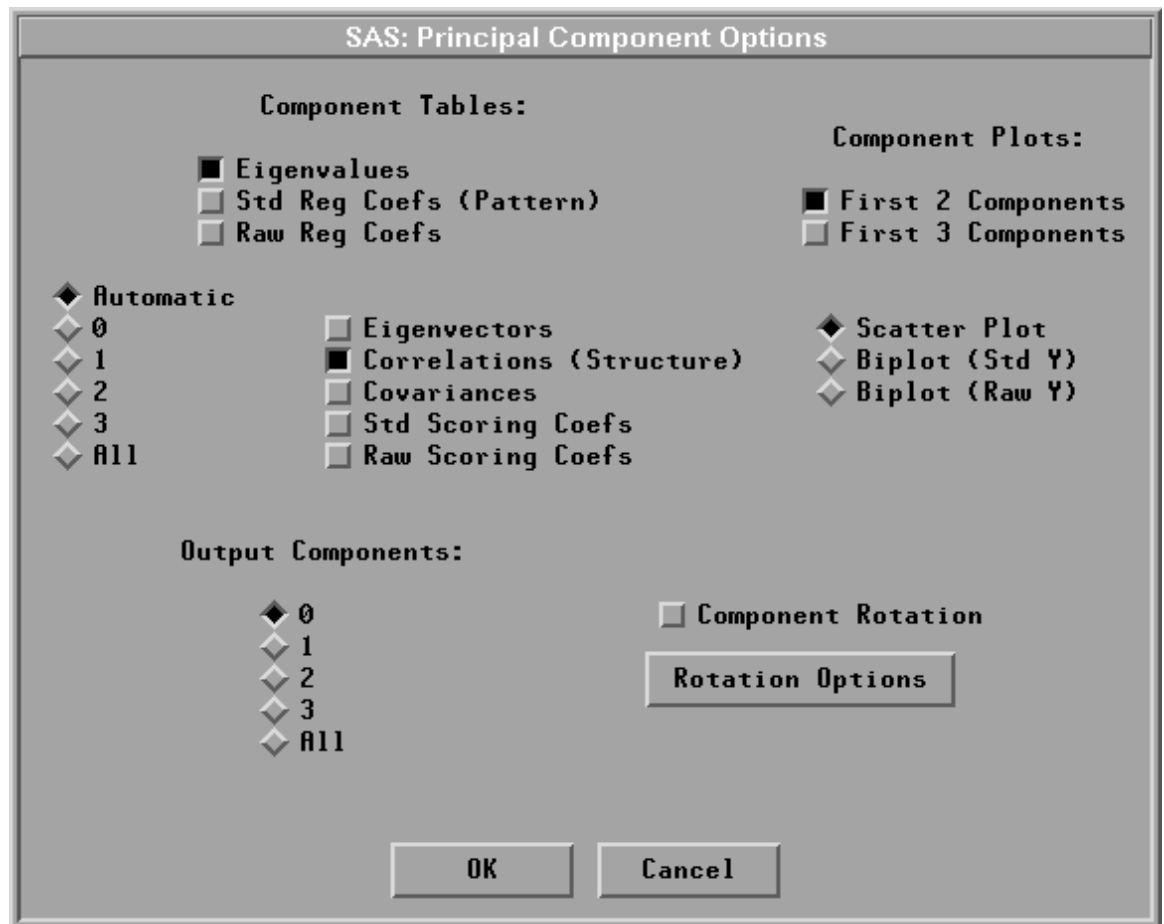


Figure 40.6. Principal Components Options Dialog

The dialog enables you to view or change the output options associated with principal component analyses and save principal component scores in the data window.

In the dialog, you need to specify the number of components when selecting tables of **Eigenvectors**, **Correlations (Structure)**, **Covariances**, **Std Scoring Coefs**, and **Raw Scoring Coefs**. **Automatic** uses principal components with corresponding eigenvalues greater than the average eigenvalue. By default, SAS/INSIGHT software displays a plot of the first two principal components, a table of all the eigenvalues, and a table of correlations between the **Y** variables and principal components with corresponding eigenvalues greater than the average eigenvalue.

You can generate principal component rotation analysis by selecting the **Component Rotation** checkbox in the dialog.

Principal Component Rotation

Clicking the **Rotation Options** button in the **Principal Components Options** dialog shown in [Figure 40.6](#) displays the **Rotation Options** dialog shown in [Figure 40.7](#).

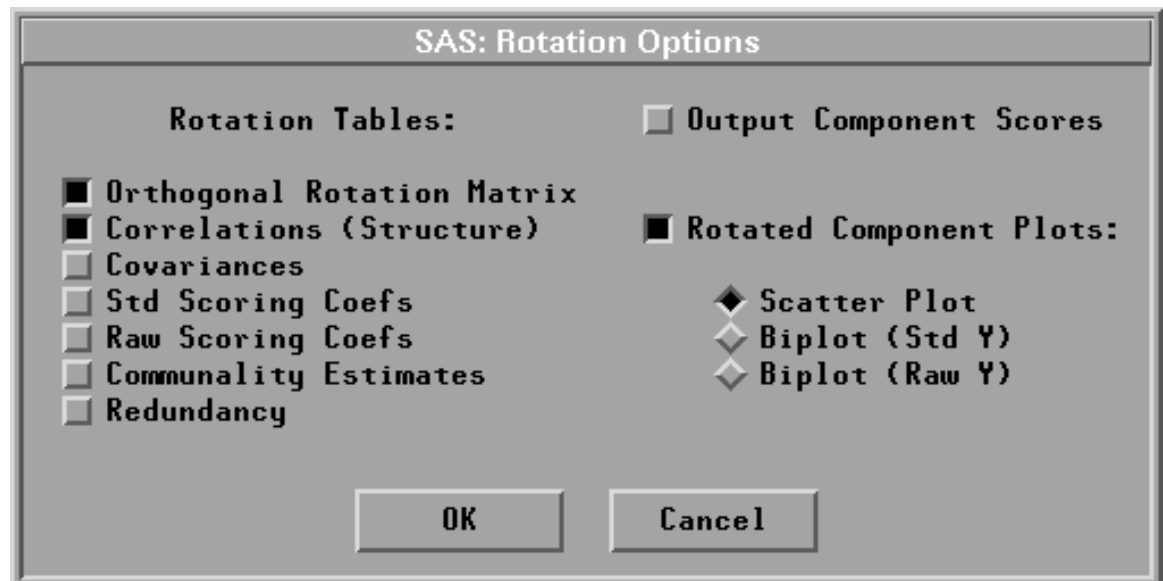


Figure 40.7. Principal Components Rotation Options Dialog

The number of components rotated is specified in the **Principal Components Rotation Options** dialog shown in [Figure 40.4](#). By default, SAS/INSIGHT software displays a plot of the rotated components (when the specified number is two or three), a rotation matrix table, and a table of correlations between the **Y** variables and rotated principal components.

Canonical Correlation Analysis

Clicking the **Canonical Correlation Options** button in the Output Options dialog shown in [Figure 40.5](#) displays the dialog shown in [Figure 40.8](#).

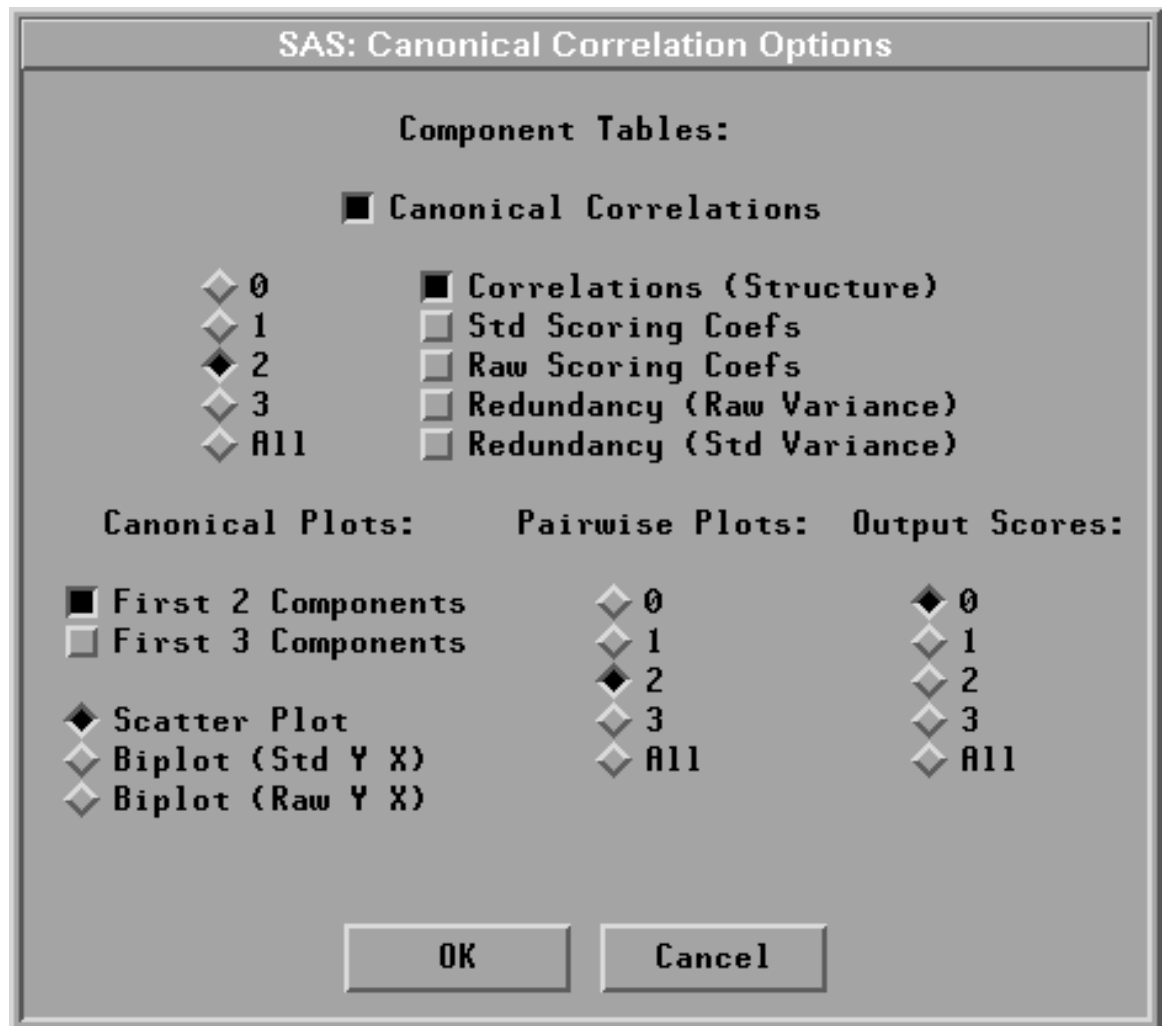


Figure 40.8. Canonical Correlation Options Dialog

This dialog enables you to view or change the options associated with canonical correlation analyses and save maximum redundancy variable scores in the data window. You specify the number of components when selecting tables of **Correlations (Structure)**, **Std Scoring Coefs**, **Raw Scoring Coefs**, **Redundancy (Raw Variance)**, and **Redundancy (Std Variance)**.

By default, SAS/INSIGHT software displays a plot of the first two canonical variables, plots of the first two pairs of canonical variables, a canonical correlations table, and a table of correlations between the **Y**, **X** variables and the first two canonical variables from both **Y** variables and **X** variables.

Maximum Redundancy Analysis

Clicking the **Maximum Redundancy Options** button in the Output Options dialog shown in [Figure 40.5](#) displays the dialog shown in [Figure 40.9](#).

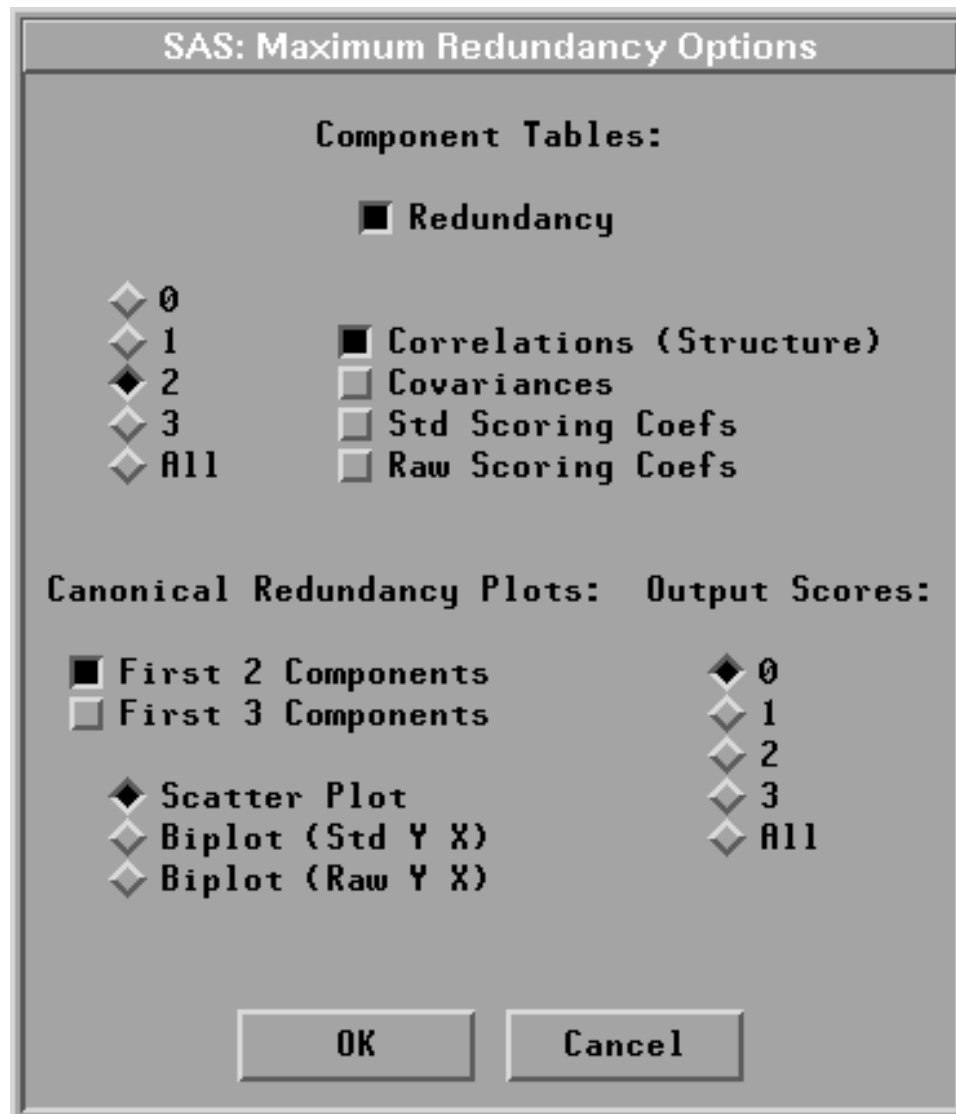


Figure 40.9. Maximum Redundancy Options Dialog

This dialog enables you to view or change the options associated with canonical correlation analyses and save maximum redundancy variable scores in the data window. You specify the number of components when selecting tables of **Correlations (Structure)**, **Covariances**, **Std Scoring Coefs**, and **Raw Scoring Coefs**.

By default, SAS/INSIGHT software displays a plot of the first two canonical redundancy variables, a canonical redundancy table, and a table of correlations between the **Y**, **X** variables and the first two canonical redundancy variables from both **Y** variables and **X** variables.

Canonical Discriminant Analysis

Clicking the **Canonical Discriminant Options** button in the Output Options dialog shown in [Figure 40.5](#) displays the dialog shown in [Figure 40.10](#).

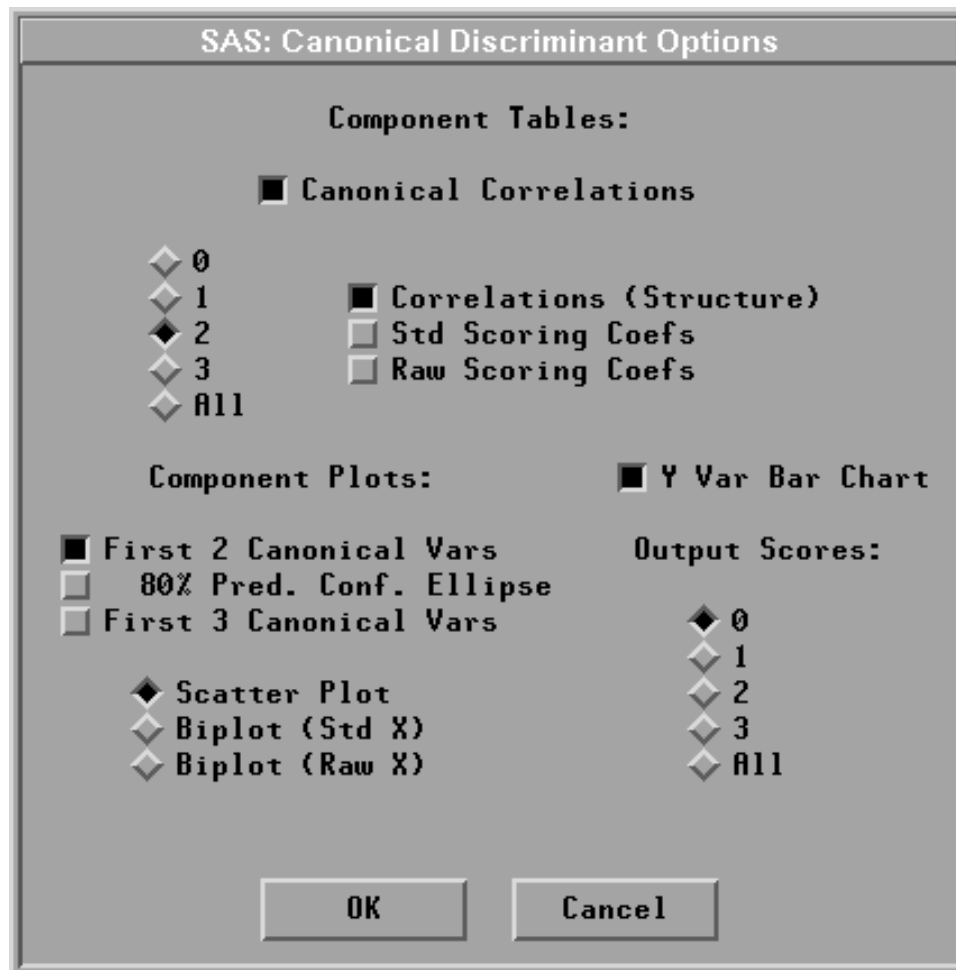


Figure 40.10. Canonical Discriminant Options Dialog

You specify the number of components when selecting tables of **Correlations (Structure)**, **Std Scoring Coefs**, and **Raw Scoring Coefs**.

By default, SAS/INSIGHT software displays a plot of the first two canonical variables, a bar chart for the nominal **Y** variable, a canonical correlation table, and a table of correlations between the **X** variables and the first two canonical variables.

Tables

You can generate tables of descriptive statistics and output from multivariate analyses by setting options in output options dialogs, as shown in [Figure 40.5](#) to [Figure 40.10](#), or by choosing from the **Tables** menu shown in [Figure 40.11](#).

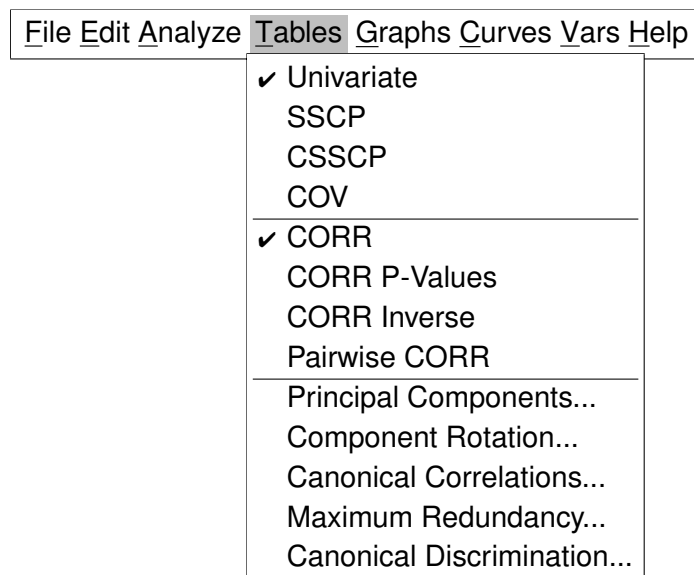


Figure 40.11. Tables Menu

Univariate Statistics

The **Univariate Statistics** table, as shown in [Figure 40.12](#) contains the following information:

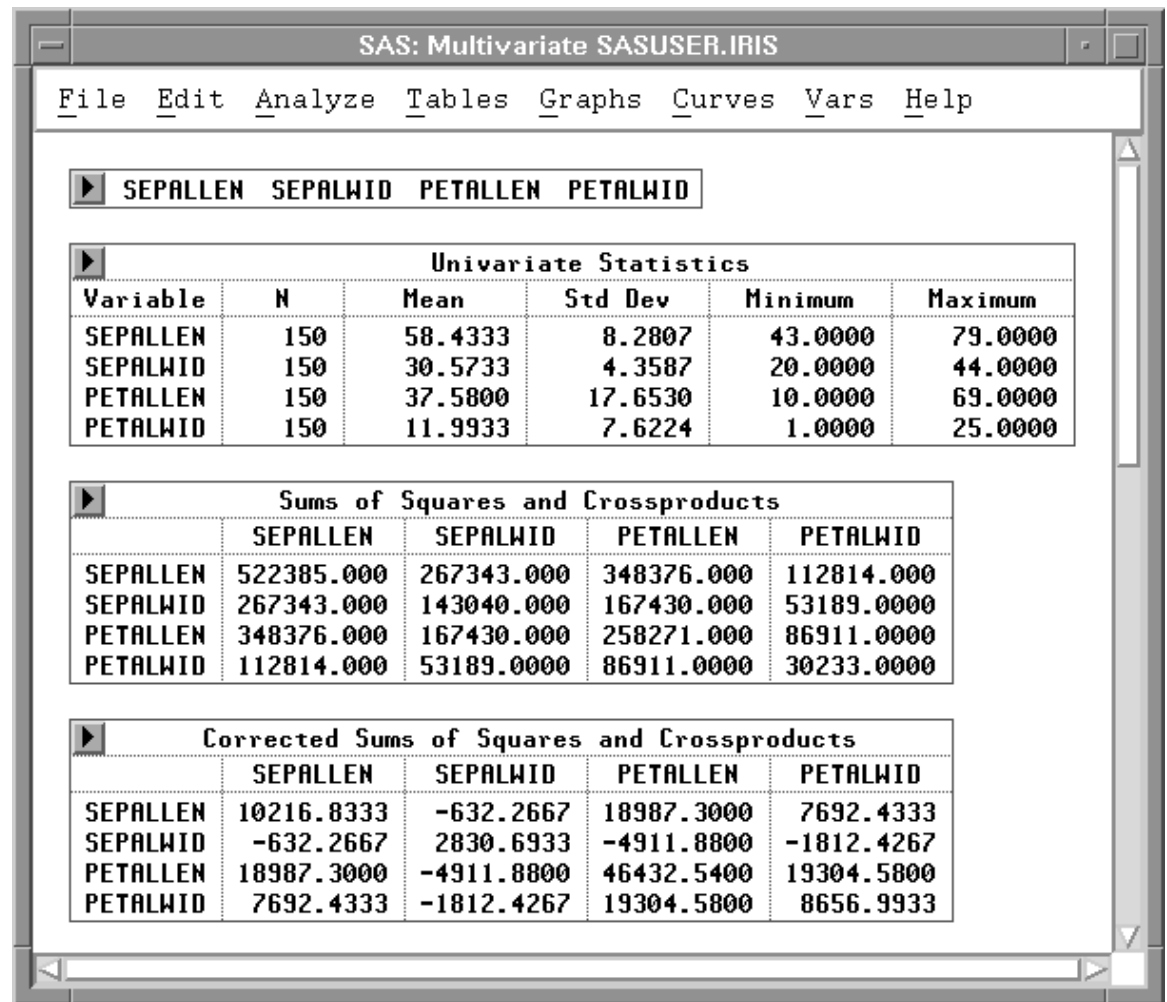
- **Variable** is the variable name.
- **N** is the number of nonmissing observations, n .
- **Mean** is the variable mean, \bar{y} or \bar{x} .
- **Std Dev** is the standard deviation of the variable, the square root of the corresponding diagonal element of S_{yy} or S_{xx} .
- **Minimum** is the minimum value.
- **Maximum** is the maximum value.
- **Partial Std Dev** (with selected **Partial** variables) is the partial standard deviation of the variable after partialling out the **Partial** variables.

Sums of Squares and Crossproducts

The **Sums of Squares and Crossproducts** (SSCP) table, as illustrated by [Figure 40.12](#), contains the sums of squares and crossproducts of the variables.

Corrected Sums of Squares and Crossproducts

The **Corrected Sums of Squares and Crossproducts** (CSSCP) table, as shown in Figure 40.12, contains the sums of squares and crossproducts of the variables corrected for the mean.



SAS: Multivariate SASUSER.IRIS

File Edit Analyze Tables Graphs Curves Vars Help

SEPALLEN SEPALWID PETALLEN PETALWID

Univariate Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
SEPALLEN	150	58.4333	8.2807	43.0000	79.0000
SEPALWID	150	30.5733	4.3587	20.0000	44.0000
PETALLEN	150	37.5800	17.6530	10.0000	69.0000
PETALWID	150	11.9933	7.6224	1.0000	25.0000

Sums of Squares and Crossproducts

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	522385.000	267343.000	348376.000	112814.000
SEPALWID	267343.000	143040.000	167430.000	53189.0000
PETALLEN	348376.000	167430.000	258271.000	86911.0000
PETALWID	112814.000	53189.0000	86911.0000	30233.0000

Corrected Sums of Squares and Crossproducts

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	10216.8333	-632.2667	18987.3000	7692.4333
SEPALWID	-632.2667	2830.6933	-4911.8800	-1812.4267
PETALLEN	18987.3000	-4911.8800	46432.5400	19304.5800
PETALWID	7692.4333	-1812.4267	19304.5800	8656.9933

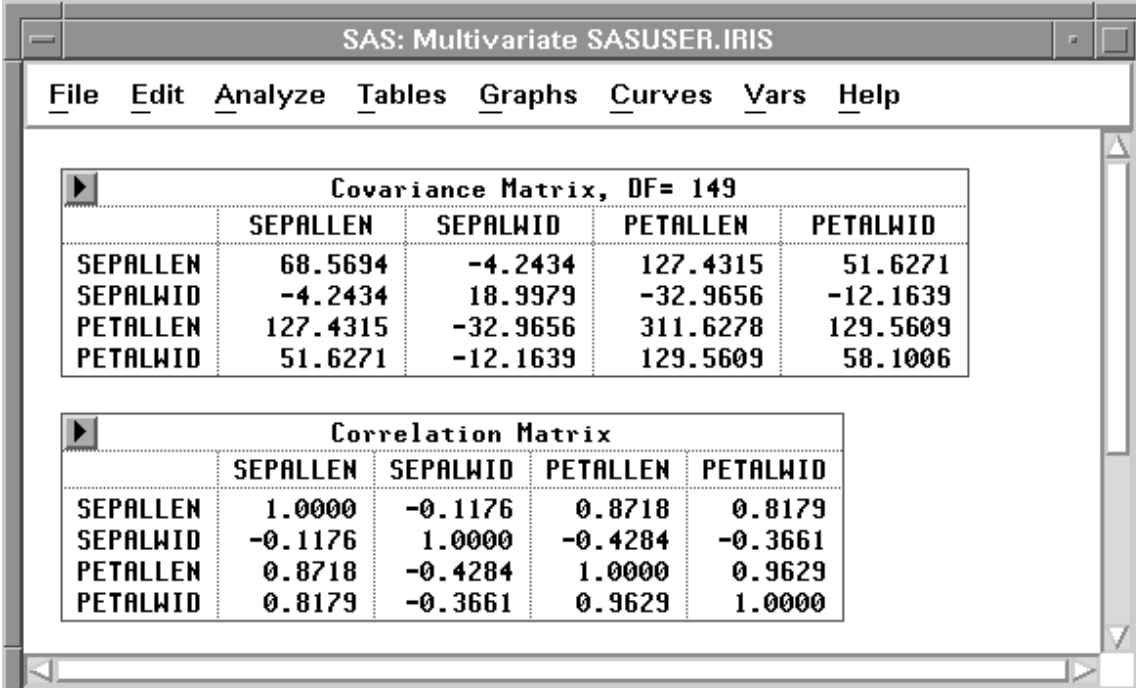
Figure 40.12. Univariate Statistics, SSCP, and CSSCP Tables

Covariance Matrix

The **Covariance Matrix** (COV) table, as shown in Figure 40.13, contains the estimated variances and covariances of the variables, with their associated degrees of freedom. The variance measures the spread of the distribution around the mean, and the covariance measures the tendency of two variables to linearly increase or decrease together

Correlation Matrix

The **Correlation Matrix** (CORR) table contains the Pearson product-moment correlations of the **Y** variables, as shown in [Figure 40.13](#). Correlation measures the strength of the linear relationship between two variables. A correlation of 0 means that there is no linear association between two variables. A correlation of 1 (-1) means that there is an exact positive (negative) linear association between the two variables.



The screenshot shows a SAS window titled 'SAS: Multivariate SASUSER.IRIS'. It contains two tables: 'Covariance Matrix, DF= 149' and 'Correlation Matrix'. Both tables have four columns: SEPALLEN, SEPALWID, PETALLEN, and PETALWID.

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	68.5694	-4.2434	127.4315	51.6271
SEPALWID	-4.2434	18.9979	-32.9656	-12.1639
PETALLEN	127.4315	-32.9656	311.6278	129.5609
PETALWID	51.6271	-12.1639	129.5609	58.1006

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.0000	-0.1176	0.8718	0.8179
SEPALWID	-0.1176	1.0000	-0.4284	-0.3661
PETALLEN	0.8718	-0.4284	1.0000	0.9629
PETALWID	0.8179	-0.3661	0.9629	1.0000

Figure 40.13. COV and CORR Tables

P-Values of the Correlations

The **P-Values of the Correlations** table contains the p -value of each correlation under the null hypothesis that the correlation is 0, assuming independent and identically distributed (unless weights are specified) observations from a bivariate distribution with at least one variable normally distributed. This table is shown in [Figure 40.14](#). Each p -value in this table can be used to assess the significance of the corresponding correlation coefficient.

The p -value of a correlation r is obtained by treating the statistic

$$t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}$$

as having a Student's t distribution with $n - 2$ degrees of freedom. The p -value of the correlation r is the probability of obtaining a Student's t statistic greater in absolute value than the absolute value of the observed statistic t .

Reference ♦ *Multivariate Analyses*

With partial variables, the p -value of a correlation is obtained by treating the statistic

$$t = \sqrt{n - n_p - 2} \frac{r}{\sqrt{1 - r^2}}$$

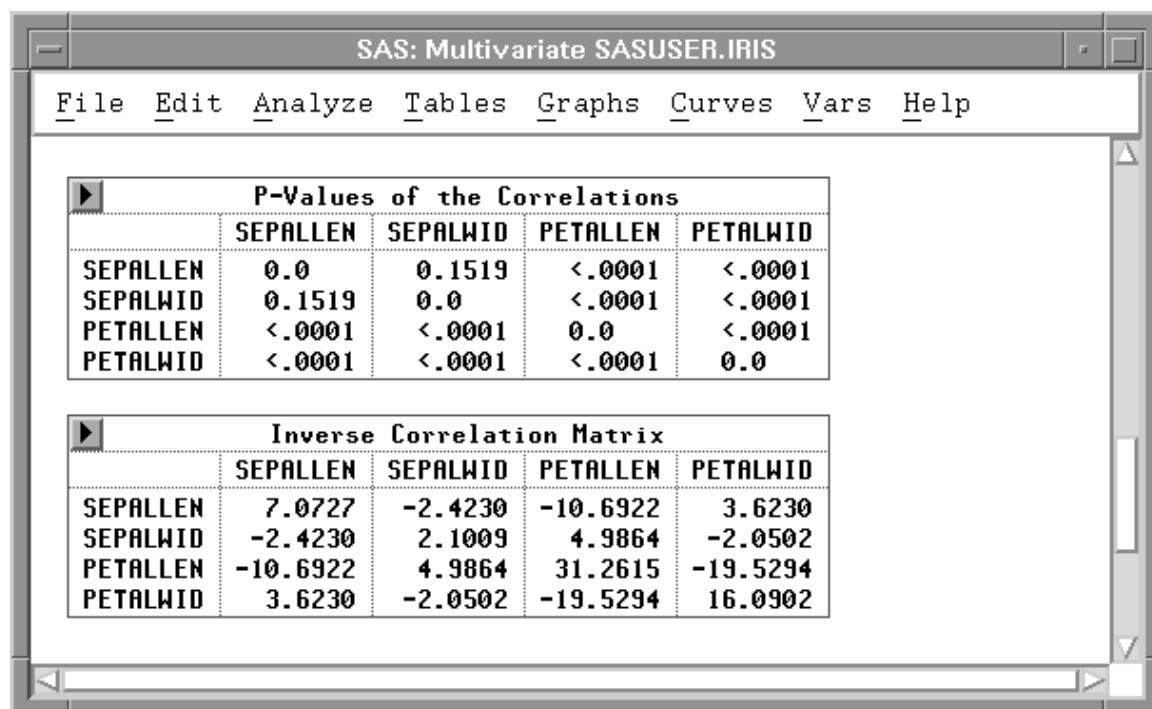
as having a Student's t distribution with $n - n_p - 2$ degrees of freedom.

Inverse Correlation Matrix

For a symmetric correlation matrix, the **Inverse Correlation Matrix** table contains the inverse of the correlation matrix, as shown in Figure 40.14.

The diagonal elements of the inverse correlation matrix, sometimes referred to as *variance inflation factors*, measure the extent to which the variables are linear combinations of other variables. The j th diagonal element of the inverse correlation matrix is $1/(1 - R_j^2)$, where R_j^2 is the squared multiple correlation of the j th variable with the other variables. Large diagonal elements indicate that variables are highly correlated.

When a correlation matrix is singular (less than full rank), some variables are linear functions of other variables, and a g2 inverse for the matrix is displayed. The g2 inverse depends on the order in which you select the variables. A value of 0 in the j th diagonal indicates that the j th variable is a linear function of the previous variables.



The screenshot shows the SAS Multivariate SASUSER.IRIS window. It contains two tables. The first table, 'P-Values of the Correlations', shows p-values for the relationships between four variables: SEPALLEN, SEPALWID, PETALLEN, and PETALWID. The second table, 'Inverse Correlation Matrix', shows the inverse of the correlation matrix for the same variables.

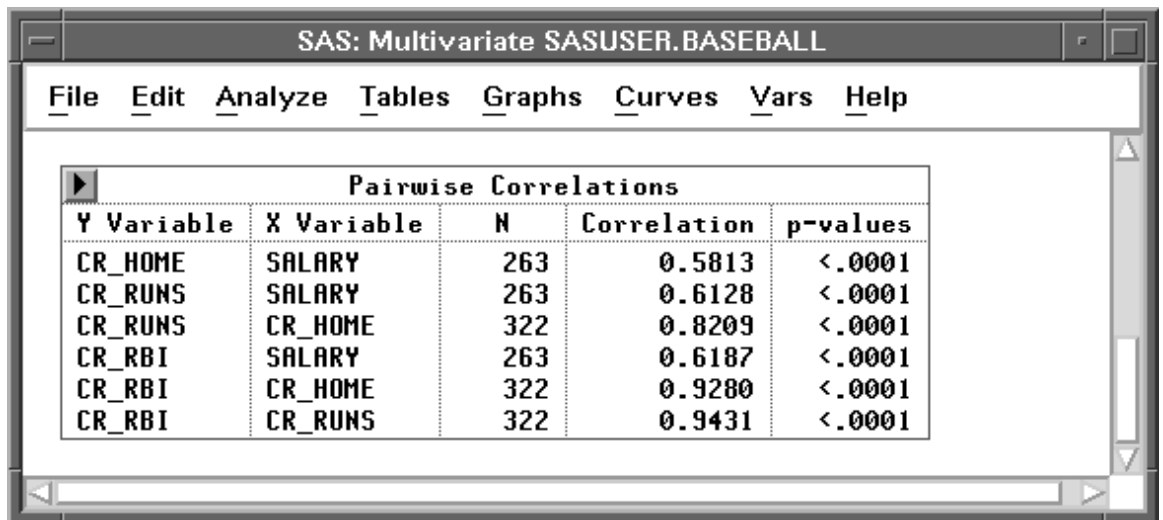
P-Values of the Correlations				
	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	0.0	0.1519	<.0001	<.0001
SEPALWID	0.1519	0.0	<.0001	<.0001
PETALLEN	<.0001	<.0001	0.0	<.0001
PETALWID	<.0001	<.0001	<.0001	0.0

Inverse Correlation Matrix				
	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	7.0727	-2.4230	-10.6922	3.6230
SEPALWID	-2.4230	2.1009	4.9864	-2.0502
PETALLEN	-10.6922	4.9864	31.2615	-19.5294
PETALWID	3.6230	-2.0502	-19.5294	16.0902

Figure 40.14. P-values of Correlations and Inverse Correlation Matrix

Pairwise Correlations

SAS/INSIGHT software drops an observation with a missing value for any variable used in the analysis from all calculations. The **Pairwise CORR** table gives correlations that are computed from all observations that have nonmissing values for any pair of variables. [Figure 40.15](#) shows a table of pairwise correlations.



The screenshot shows a SAS/INSIGHT window titled "SAS: Multivariate SASUSER.BASEBALL". The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. The main display area shows a table titled "Pairwise Correlations". The table has five columns: Y Variable, X Variable, N, Correlation, and p-values. The data rows are as follows:

Y Variable	X Variable	N	Correlation	p-values
CR_HOME	SALARY	263	0.5813	<.0001
CR_RUNS	SALARY	263	0.6128	<.0001
CR_RUNS	CR_HOME	322	0.8209	<.0001
CR_RBI	SALARY	263	0.6187	<.0001
CR_RBI	CR_HOME	322	0.9280	<.0001
CR_RBI	CR_RUNS	322	0.9431	<.0001

Figure 40.15. Pairwise CORR Table

Principal Component Analysis

You can generate tables of output from principal component analyses by setting options in the principal component options dialog shown in [Figure 40.6](#) or from the **Tables** menu shown in [Figure 40.11](#). Select **Principal Components** from the **Tables** menu to display the principal component tables dialog shown in [Figure 40.16](#).

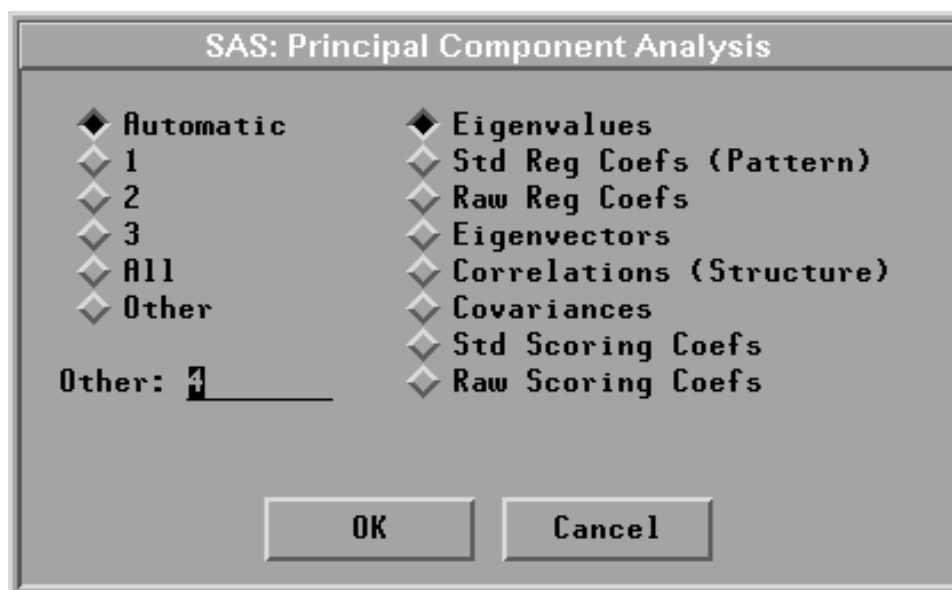


Figure 40.16. Principal Component Tables Dialog

Choose **Automatic** to display principal components with eigenvalues greater than the average eigenvalue. Selecting **1**, **2**, or **3** gives you 1, 2, or 3 principal components. **All** gives you all eigenvalues. Selecting **0** in the principal component options dialog suppresses the principal component tables.

The **Eigenvalues (COV)** or **Eigenvalues (CORR)** table includes the eigenvalues of the covariance or correlation matrix, the difference between successive eigenvalues, the proportion of variance explained by each eigenvalue, and the cumulative proportion of variance explained.

Eigenvalues correspond to each of the principal components and represent a partitioning of the total variation in the sample. The sum of all eigenvalues is equal to the sum of all variable variances if the covariance matrix is used or to the number of variables, p , if the correlation matrix is used.

The **Eigenvectors (COV)** or **Eigenvectors (CORR)** table includes the eigenvectors of the covariance or correlation matrix. Eigenvectors correspond to each of the principal components and are used as the coefficients to form linear combinations of the **Y** variables (principal components).

Figure 40.17 shows tables of all eigenvalues and eigenvectors for the first two principal components.

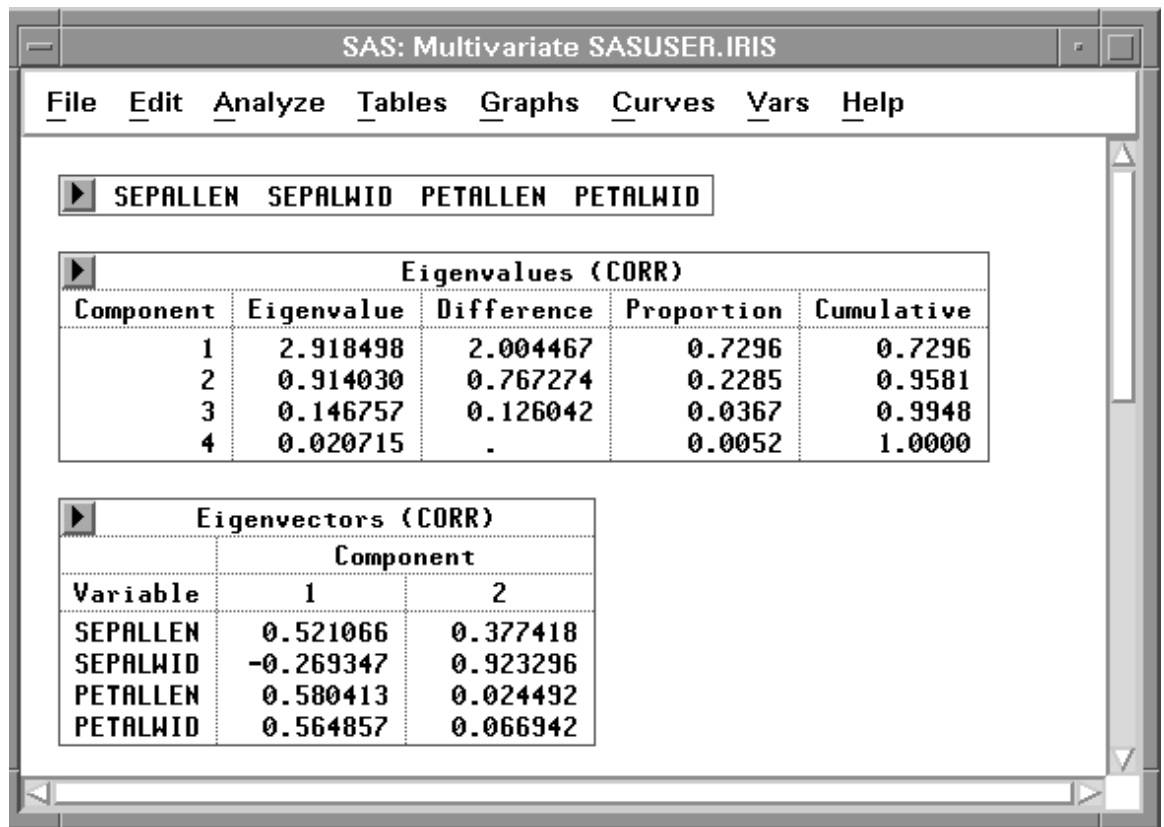
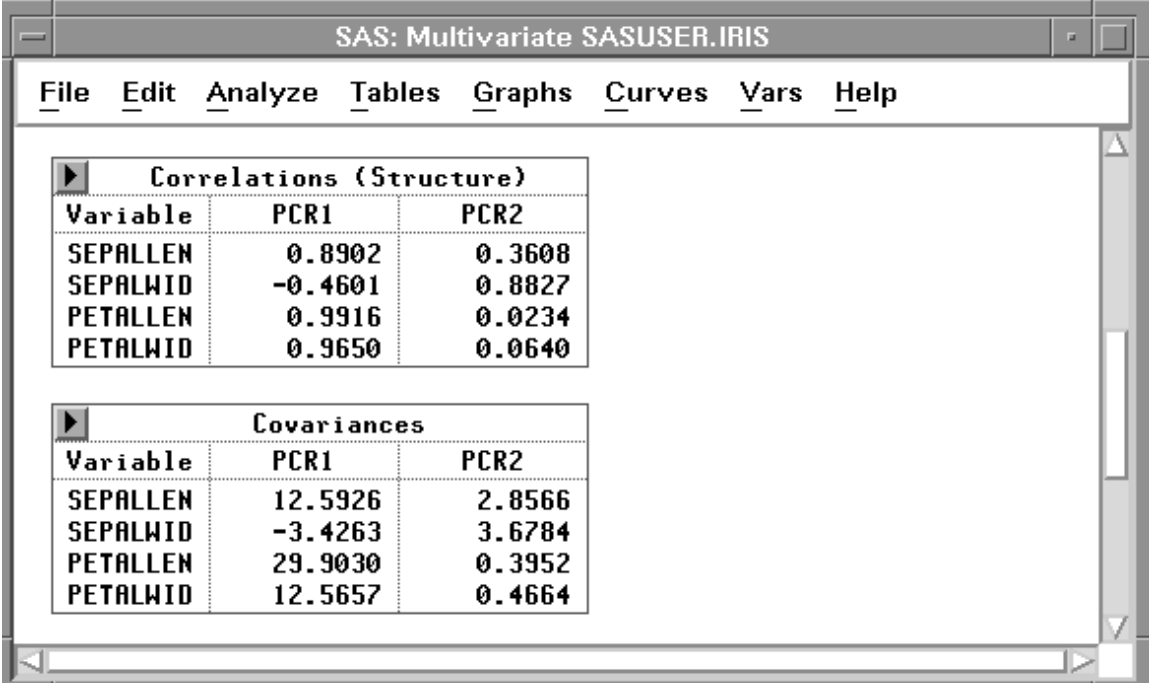


Figure 40.17. Eigenvalues and Eigenvectors Tables

The **Correlations (Structure)** and **Covariances** tables include the correlations and covariances, respectively, between the **Y** variables and principal components. The correlation and covariance matrices measure the strength of the linear relationship between the derived principal components and each of the **Y** variables. Figure 40.18 shows the correlations and covariances between the **Y** variables and the first two principal components.



The screenshot shows a SAS window titled "SAS: Multivariate SASUSER.IRIS". The menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. The main content area displays two tables. The first table, "Correlations (Structure)", shows the correlation coefficients between four variables (SEPALLEN, SEPALWID, PETALLEN, PETALWID) and two principal components (PCR1, PCR2). The second table, "Covariances", shows the covariance values for the same variables and principal components.

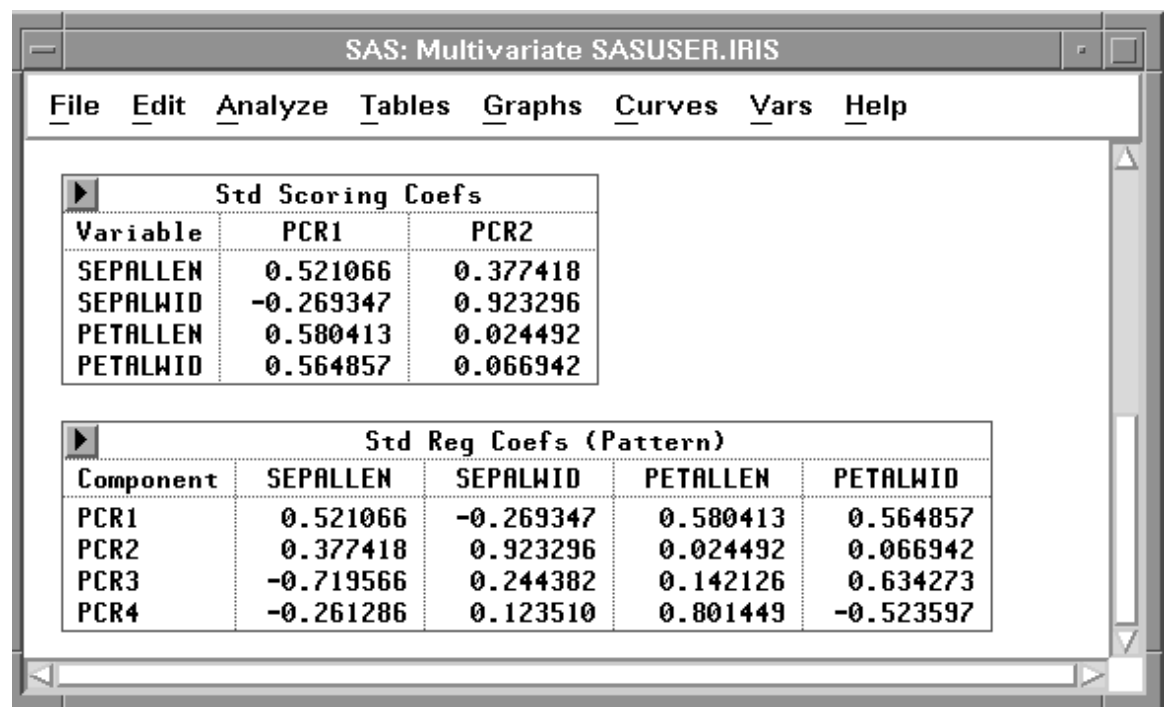
Variable	PCR1	PCR2
SEPALLEN	0.8902	0.3608
SEPALWID	-0.4601	0.8827
PETALLEN	0.9916	0.0234
PETALWID	0.9650	0.0640

Variable	PCR1	PCR2
SEPALLEN	12.5926	2.8566
SEPALWID	-3.4263	3.6784
PETALLEN	29.9030	0.3952
PETALWID	12.5657	0.4664

Figure 40.18. Correlations and Covariances Tables

The scoring coefficients are the coefficients of the **Y** variables used to generate principal components. The **Std Scoring Coefs** table includes the scoring coefficients of the standardized **Y** variables, and the **Raw Scoring Coefs** table includes the scoring coefficients of the centered **Y** variables.

The regression coefficients are the coefficients of principal components used to generate estimated **Y** variables. The **Std Reg Coefs (Pattern)** and **Raw Reg Coefs** tables include the regression coefficients of principal components used to generate estimated standardized and centered **Y** variables. Figure 40.19 shows the regression coefficients of the principal components for the standardized **Y** variables, as well as the scoring coefficients of the standardized **Y** variables for the first two principal components.



The screenshot shows a SAS window titled "SAS: Multivariate SASUSER.IRIS". It contains two tables. The first table, "Std Scoring Coefs", shows coefficients for four variables (SEPALLEN, SEPALWID, PETALLEN, PETALWID) across two principal components (PCR1, PCR2). The second table, "Std Reg Coefs (Pattern)", shows coefficients for the same four variables across four principal components (PCR1, PCR2, PCR3, PCR4).

Variable	PCR1	PCR2
SEPALLEN	0.521066	0.377418
SEPALWID	-0.269347	0.923296
PETALLEN	0.580413	0.024492
PETALWID	0.564857	0.066942

Component	SEPALLEN	SEPALWID	PETALLEN	PETALWID
PCR1	0.521066	-0.269347	0.580413	0.564857
PCR2	0.377418	0.923296	0.024492	0.066942
PCR3	-0.719566	0.244382	0.142126	0.634273
PCR4	-0.261286	0.123510	0.801449	-0.523597

Figure 40.19. Regression Coefficients and Scoring Coefficients Tables

Principal Components Rotation

You can generate tables of output from principal component rotation by setting options in the **Rotation Options** dialog shown in [Figure 40.7](#) or from the **Tables** menu shown in [Figure 40.11](#). Select **Component Rotation** from the **Tables** menu to display the principal component rotation dialog shown in [Figure 40.20](#).

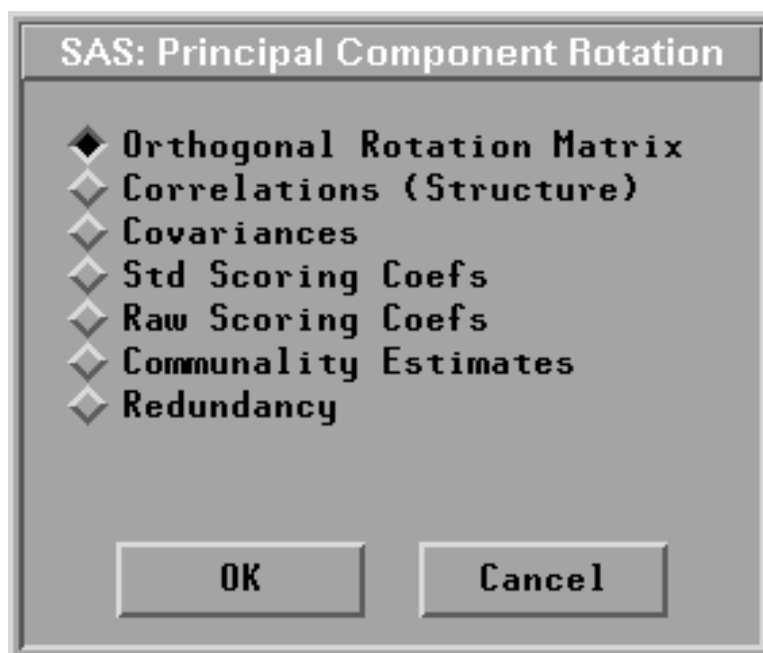


Figure 40.20. Principal Components Rotation Dialog

You specify the number of components and type of rotation in the **Rotation Options** dialog, as shown in [Figure 40.4](#).

The **Orthogonal Rotation Matrix** is the orthogonal rotation matrix used to compute the rotated principal components from the standardized principal components.

The **Correlations (Structure)** and **Covariances** tables include the correlations and covariances between the **Y** variables and the rotated principal components.

[Figure 40.21](#) shows the rotation matrix and correlations and covariances between the **Y** variables and the first two rotated principal components.

The scoring coefficients are the coefficients of the **Y** variables used to generate rotated principal components. The **Std Scoring Coefs** table includes the scoring coefficients of the standardized **Y** variables, and the **Raw Scoring Coefs** table includes the scoring coefficients of the centered **Y** variables.

The **Communality Estimates** table gives the standardized variance of each **Y** variable explained by the rotated principal components.

The **Redundancy** table gives the variances of the standardized **Y** variables explained by each rotated principal component.

Figure 40.22 shows the scoring coefficients of the standardized **Y** variables, communality estimates for the **Y** variables, and redundancy for each rotated component.

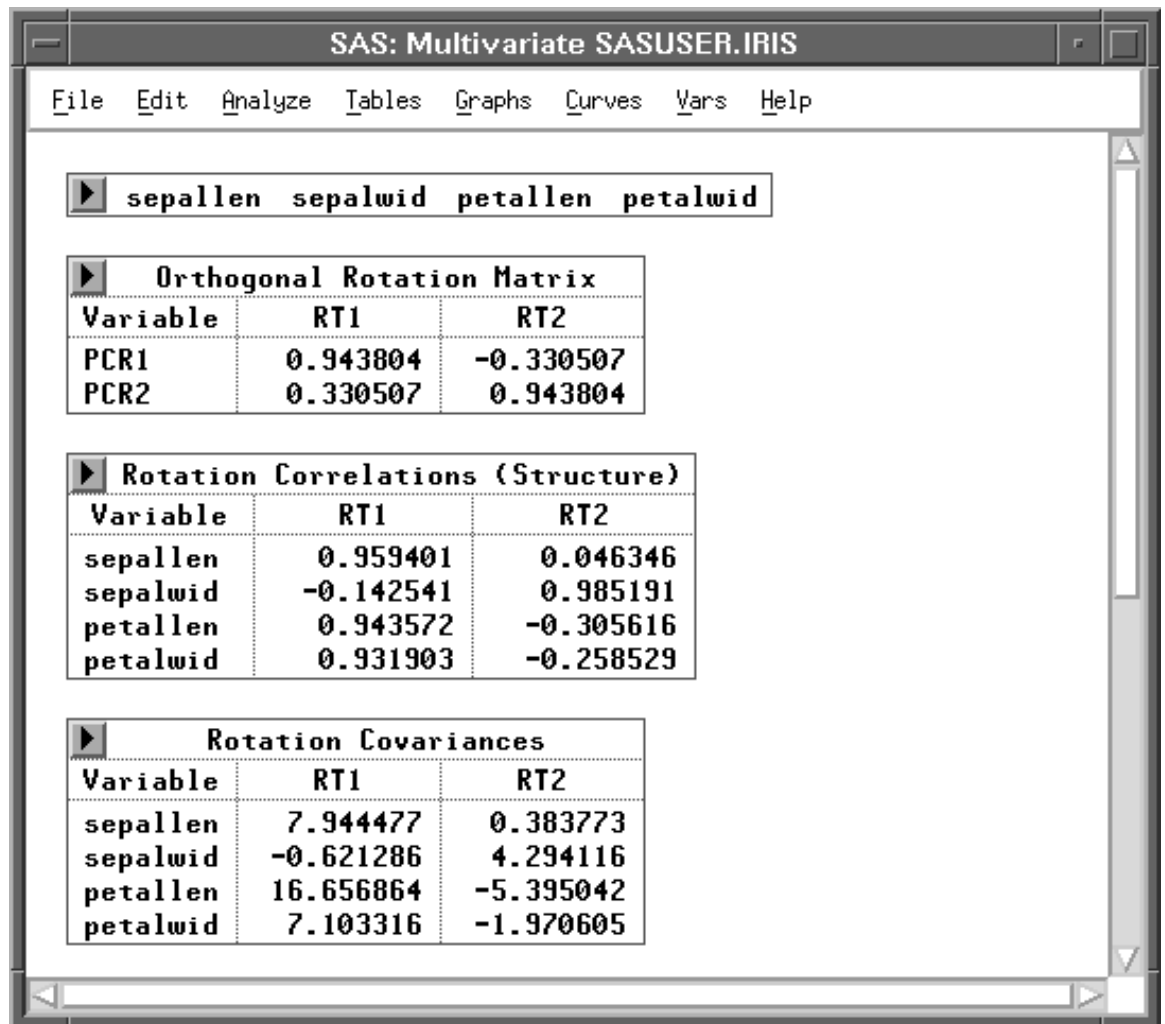


Figure 40.21. Rotation Matrix, Correlation, and Covariance Tables

The screenshot shows the SAS: Multivariate SASUSER.IRIS window. It contains three tables of results:

Rotation Std Scoring Coefs		
Variable	RT1	RT2
sepallen	0.418342	0.271776
sepalwid	0.170380	0.963578
petallen	0.329123	-0.088111
petalwid	0.335203	-0.043195

Rotation Communality Estimates	
Variable	Estimates
sepallen	0.922599
sepalwid	0.990919
petallen	0.983730
petalwid	0.935280

Rotation Redundancy	
RT1	RT2
2.699540	1.132988

Figure 40.22. Scoring Coefficients, Communality, and Redundancy Tables

Canonical Correlation Analysis

You can generate tables of output from canonical correlation analyses by setting options in the Canonical Correlation Options dialog shown in [Figure 40.8](#) or from the **Tables** menu shown in [Figure 40.11](#). Select **Canonical Correlations** from the **Tables** menu to display the canonical correlation dialog shown in [Figure 40.23](#).

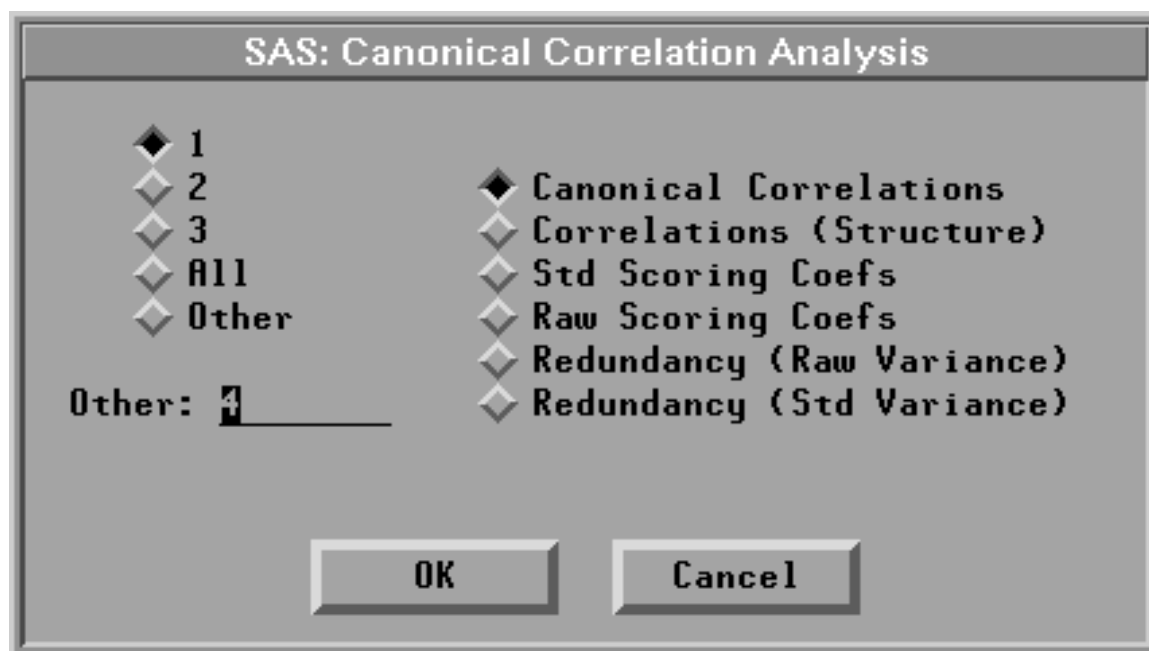


Figure 40.23. Canonical Correlation Dialog

The **Canonical Correlations** table contains the following:

- **CanCorr**, the canonical correlations, which are always nonnegative
- **Adj. CanCorr**, the adjusted canonical correlations, which are asymptotically less biased than the raw correlations and may be negative. The adjusted canonical correlations may not be computable, and they are displayed as missing values if two canonical correlations are nearly equal or if some are close to zero. A missing value is also displayed if an adjusted canonical correlation is larger than a previous adjusted canonical correlation.
- **Approx Std. Error**, the approximate standard errors of the canonical correlations
- **CanRsqr**, the squared canonical correlations
- **Eigenvalues**, the eigenvalues of the matrix $R_{yy}^{-1}R_{yx}R_{xx}^{-1}R'_{yx}$. These eigenvalues are equal to $\text{CanRsqr}/(1 - \text{CanRsqr})$, where CanRsqr is the corresponding squared canonical correlation. Also printed for each eigenvalue is the difference from the next eigenvalue, the proportion of the sum of the eigenvalues, and the cumulative proportion.

- **Test for H0: CanCorrj=0, j>=k**, the likelihood ratio for the hypothesis that the current canonical correlation and all smaller ones are zero in the population
- **Approx F** based on Rao's approximation to the distribution of the likelihood ratio
- **Num DF** and **Den DF** (numerator and denominator degrees of freedom) and **Pr > F** (probability level) associated with the F statistic

Figure 40.24 shows tables of canonical correlations.

The screenshot shows the SAS Multivariate SASUSER.FIT window. At the top, a menu bar includes File, Edit, Analyze, Tables, Graphs, Curves, Vars, and Help. Below the menu bar, a list of variables is shown: weight, waist, pulse, chins, situps, jumps. Three tables are displayed below the variable list:

Canonical Correlations

	CanCorr	Adj. CanCorr	Approx Std. Error	CanRsq
1	0.795608	0.754056	0.084197	0.632992
2	0.200556	-0.076399	0.220188	0.040223
3	0.072570	.	0.228208	0.005266

Eigenvalues

	Eigenvalue	Difference	Proportion	Cumulative
1	1.7247	1.6828	0.9734	0.9734
2	0.0419	0.0366	0.0237	0.9970
3	0.0053	.	0.0030	1.0000

Test of H0: CanCorr[j]=0, j>=K

K	L. Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.350391	2.0482	9	34.2229	0.0635
2	0.954723	0.1758	4	30.0000	0.9491
3	0.994734	0.0847	1	16.0000	0.7748

Figure 40.24. Canonical Correlations Tables

The **Correlations (Structure)** table includes the correlations between the input **Y**, **X** variables and canonical variables.

The scoring coefficients are the coefficients of the **Y** or **X** variables that are used to compute canonical variable scores. The **Std Scoring Coefs** table includes the scoring coefficients of the standardized **Y** or **X** variables and the **Raw Scoring Coefs** table includes the scoring coefficients of the centered **Y** or **X** variables.

Figure 40.25 shows a table of correlations between the **Y**, **X** variables and the first two canonical variables from the **Y** and **X** variables and the tables of scoring coefficients of the standardized **Y** and **X** variables.

The screenshot shows the SAS: Multivariate SASUSER.FIT window with a menu bar (File, Edit, Analyze, Tables, Graphs, Curves, Vars, Help). The main area displays three tables:

Variable	CY1	CY2	CX1	CX2
weight	0.6206	-0.7724	0.4938	-0.1549
waist	0.9254	-0.3777	0.7363	-0.0757
pulse	-0.3328	0.0415	-0.2648	0.0083
chins	-0.5789	0.0475	-0.7276	0.2370
situps	-0.6506	0.1149	-0.8177	0.5730
jumps	-0.1290	0.1923	-0.1622	0.9586

Variable	CY1	CY2
weight	-0.775398	-1.884367
waist	1.579347	1.180641
pulse	-0.059120	-0.231107

Variable	CX1	CX2
chins	-0.349497	-0.375544
situps	-1.054011	0.123490
jumps	0.716427	1.062167

Figure 40.25. Correlations and Scoring Coefficients Tables

The **Redundancy** table gives the canonical redundancy analysis, which includes the proportion and cumulative proportion of the raw (unstandardized) and the standardized variance of the set of **Y** and the set of **X** variables explained by their own canonical variables and explained by the opposite canonical variables. [Figure 40.26](#) shows tables of redundancy of standardized **Y** and **X** variables.

SAS: Multivariate SASUSER.FIT					
File Edit Analyze Tables Graphs Curves Vars Help					
Std Variance (Y Variables)					
	Explained by CY's		CanRsq	Explained by CX's	
	Proportion	Cumulative		Proportion	Cumulative
1	0.4508	0.4508	0.632992	0.2854	0.2854
2	0.2470	0.6978	0.040223	0.0099	0.2953
Std Variance (X Variables)					
	Explained by CY's		CanRsq	Explained by CX's	
	Proportion	Cumulative		Proportion	Cumulative
1	0.2584	0.2584	0.632992	0.4081	0.4081
2	0.0175	0.2758	0.040223	0.4345	0.8426

Figure 40.26. Redundancy Tables

Maximum Redundancy Analysis

You can generate tables of output from maximum redundancy analysis by setting options in the Maximum Redundancy Options dialog shown in [Figure 40.9](#) or from the **Tables** menu shown in [Figure 40.11](#). Select **Maximum Redundancy** from the **Tables** menu to display the maximum redundancy dialog shown in [Figure 40.27](#).

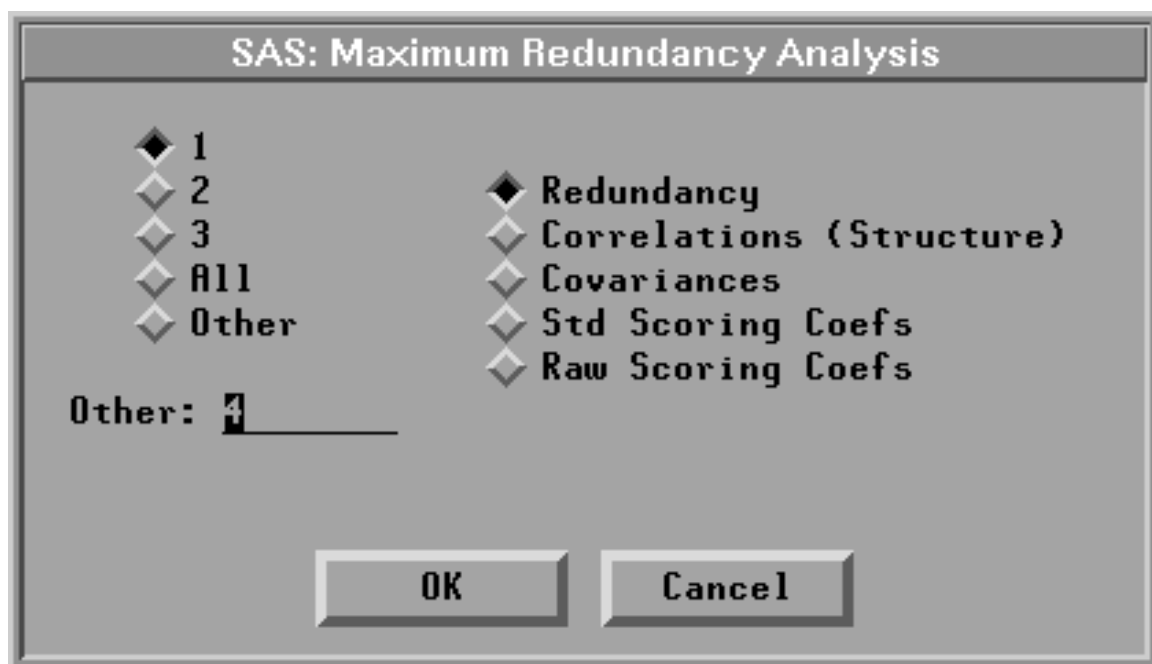


Figure 40.27. Maximum Redundancy Dialog

Either the raw (centered) or standardized variance is used in the maximum redundancy analysis, and it is specified in the Multivariate Method Options dialog in [Figure 40.3](#). The **Redundancy** table includes the proportion and cumulative proportion of the variance of the set of **Y** variables and the set of **X** variables explained by the opposite canonical variables. [Figure 40.28](#) shows tables of redundancy of the standardized **Y** and **X** variables.

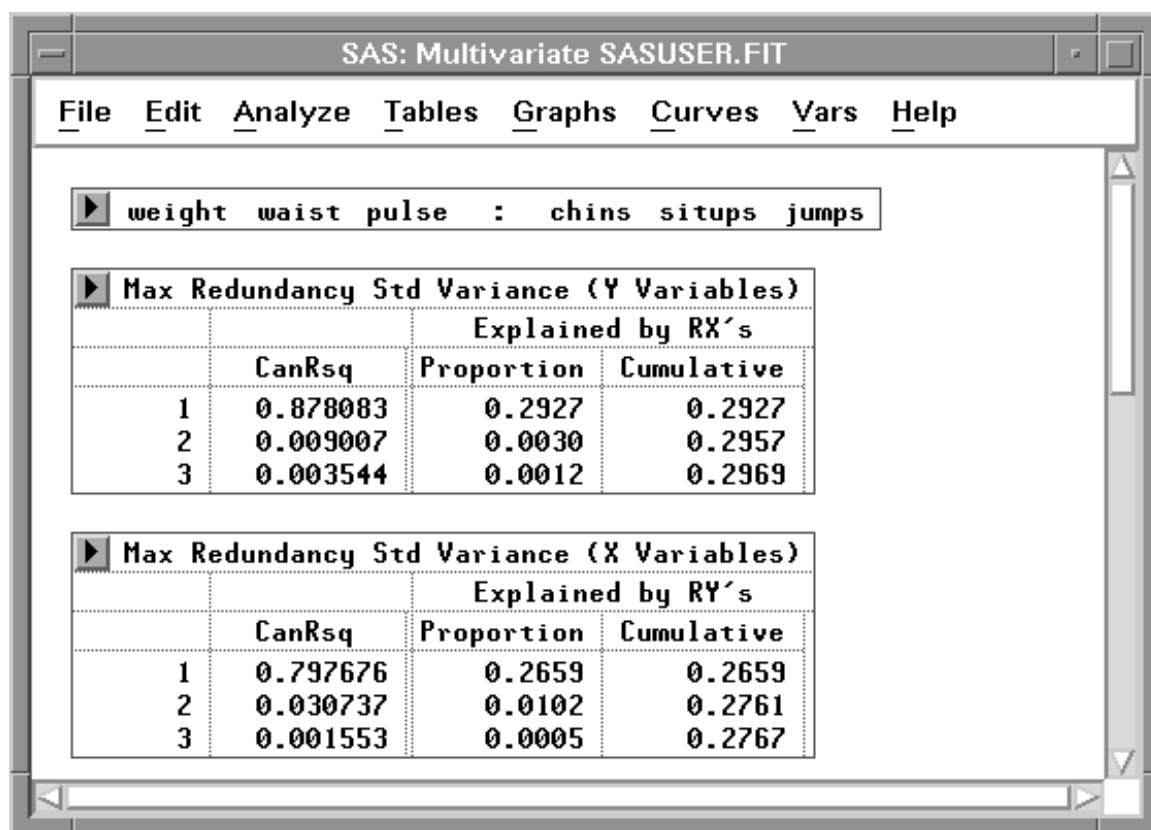
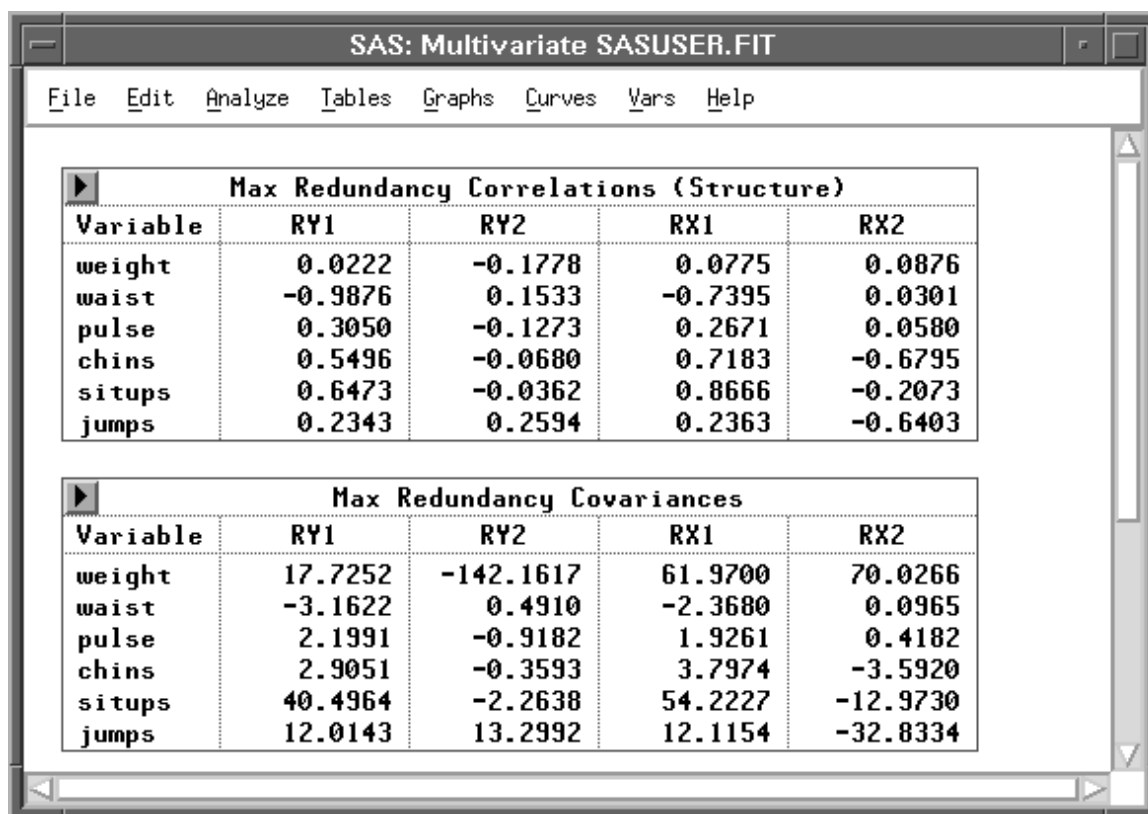


Figure 40.28. Maximum Redundancy Tables

The **Correlations (Structure)** or **Covariances** table includes the correlations or covariances between the **Y**, **X** variables and the maximum redundancy variables. [Figure 40.29](#) shows the correlations and covariances between the **Y**, **X** variables and the first two maximum redundancy variables from the **Y** variables and the **X** variables.



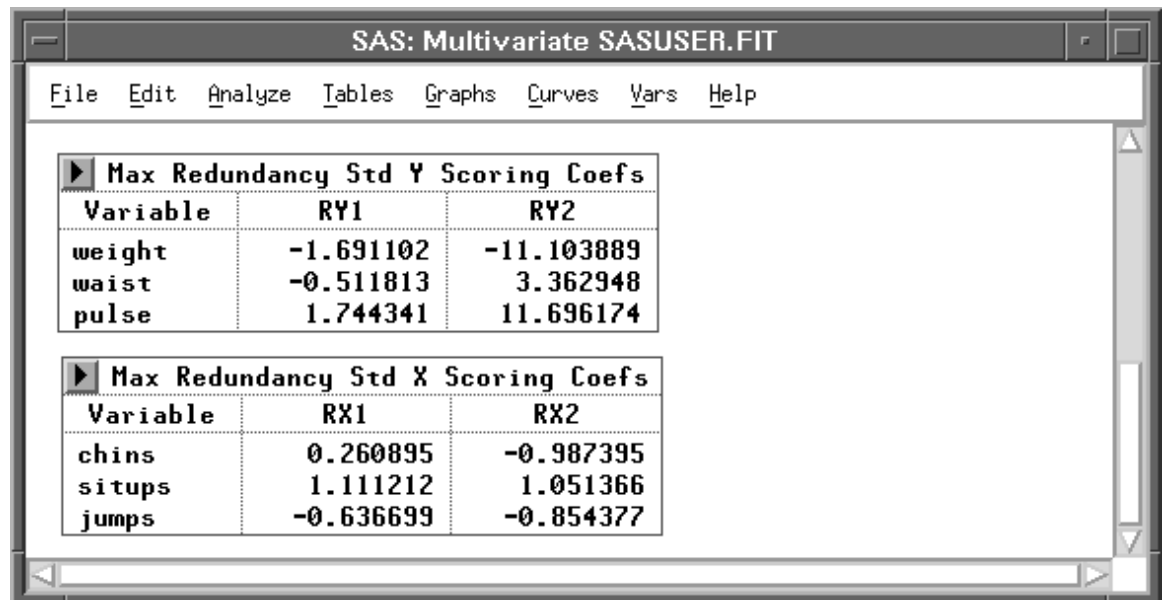
The screenshot shows the SAS: Multivariate SASUSER.FIT window. It contains two tables. The first table, 'Max Redundancy Correlations (Structure)', shows correlations between variables (weight, waist, pulse, chins, situps, jumps) and maximum redundancy variables (RY1, RY2, RX1, RX2). The second table, 'Max Redundancy Covariances', shows the corresponding covariance values for the same variables and redundancy variables.

Max Redundancy Correlations (Structure)				
Variable	RY1	RY2	RX1	RX2
weight	0.0222	-0.1778	0.0775	0.0876
waist	-0.9876	0.1533	-0.7395	0.0301
pulse	0.3050	-0.1273	0.2671	0.0580
chins	0.5496	-0.0680	0.7183	-0.6795
situps	0.6473	-0.0362	0.8666	-0.2073
jumps	0.2343	0.2594	0.2363	-0.6403

Max Redundancy Covariances				
Variable	RY1	RY2	RX1	RX2
weight	17.7252	-142.1617	61.9700	70.0266
waist	-3.1622	0.4910	-2.3680	0.0965
pulse	2.1991	-0.9182	1.9261	0.4182
chins	2.9051	-0.3593	3.7974	-3.5920
situps	40.4964	-2.2638	54.2227	-12.9730
jumps	12.0143	13.2992	12.1154	-32.8334

Figure 40.29. Correlation and Covariance Tables

The scoring coefficients are the coefficients of the **Y** or **X** variables that are used to compute maximum redundancy variables. The **Std Scoring Coefs** table includes the scoring coefficients of the standardized **Y** or **X** variables, and the **Raw Scoring Coefs** table includes the scoring coefficients of the centered **Y** or **X** variables. [Figure 40.30](#) shows tables of the scoring coefficients of the standardized **Y** and **X** variables.



The screenshot shows a SAS window titled "SAS: Multivariate SASUSER.FIT" with a menu bar (File, Edit, Analyze, Tables, Graphs, Curves, Vars, Help). The main area displays two tables of standardized scoring coefficients.

Variable	RY1	RY2
weight	-1.691102	-11.103889
waist	-0.511813	3.362948
pulse	1.744341	11.696174

Variable	RX1	RX2
chins	0.260895	-0.987395
situps	1.111212	1.051366
jumps	-0.636699	-0.854377

Figure 40.30. Standardized Scoring Coefficients Tables

Canonical Discriminant Analysis

You can generate tables of output from canonical discriminant analyses by setting options in the Canonical Discriminant Options dialog shown in [Figure 40.10](#) or from the **Tables** menu shown in [Figure 40.11](#). Select **Canonical Discrimination** from the **Tables** menu to display the canonical discriminant analysis dialog shown in [Figure 40.31](#).

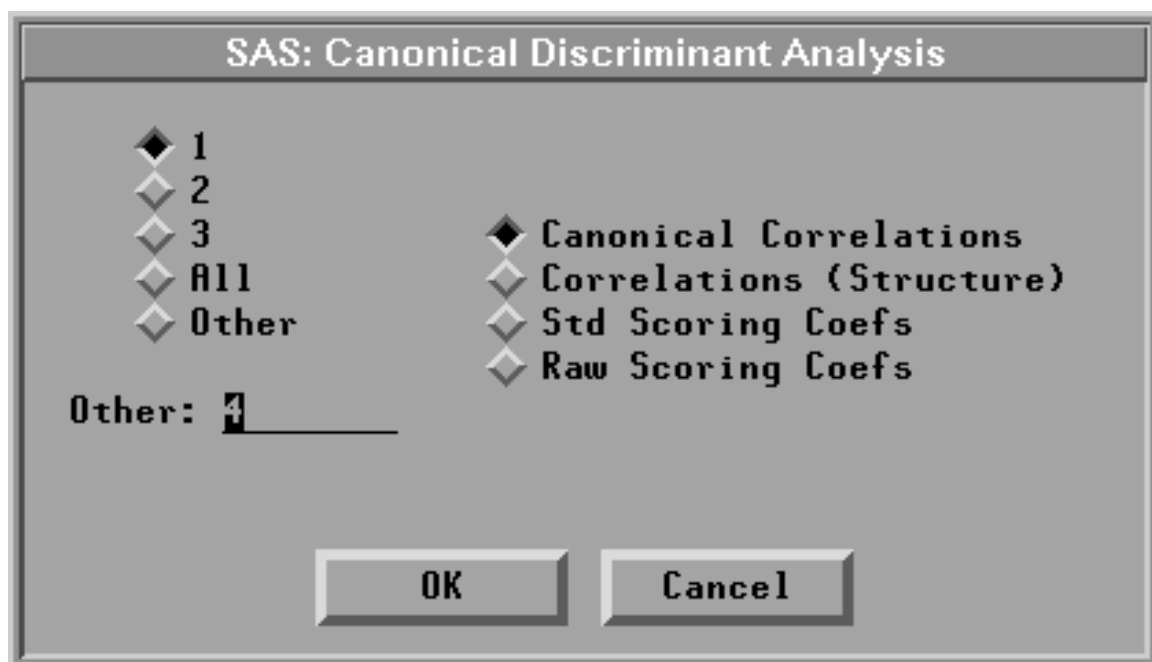


Figure 40.31. Canonical Discriminant Analysis Dialog

The **Canonical Correlations** table, as shown in [Figure 40.32](#), contains the following:

- **CanCorr**, the canonical correlations, which are always nonnegative
- **Adj. CanCorr**, the adjusted canonical correlations, which are asymptotically less biased than the raw correlations and may be negative. The adjusted canonical correlations may not be computable and are displayed as missing values if two canonical correlations are nearly equal or if some are close to zero. A missing value is also displayed if an adjusted canonical correlation is larger than a previous adjusted canonical correlation.
- **Approx Std. Error**, the approximate standard errors of the canonical correlations
- **CanRsqr**, the squared canonical correlations
- **Eigenvalues**, eigenvalues of the matrix $E^{-1}H$, where E is the matrix of the within-class sums of squares and crossproducts and H is the matrix of the between-class sums of squares and crossproducts. These eigenvalues are equal

Reference ♦ *Multivariate Analyses*

to $\text{CanRsq}/(1 - \text{CanRsq})$, where CanRsq is the corresponding squared canonical correlation. Also displayed for each eigenvalue is the difference from the next eigenvalue, the proportion of the sum of the eigenvalues, and the cumulative proportion.

- **Test for H0: CanCorrj=0, j>=k**, the likelihood ratio for the hypothesis that the current canonical correlation and all smaller ones are zero in the population
- **Approx F** based on Rao's approximation to the distribution of the likelihood ratio
- **Num DF** and **Den DF** (numerator and denominator degrees of freedom) and **Pr > F** (probability level) associated with the F statistic

The screenshot shows the SAS output window with the following tables:

SPECIES : SEPALLEN SEPALWID PETALLEN PETALWID

Canonical Correlations

	CanCorr	Adj. CanCorr	Approx Std. Error	CanRsq
1	0.984821	0.984508	0.002468	0.969872
2	0.471197	0.461445	0.063734	0.222027

Eigenvalues

	Eigenvalue	Difference	Proportion	Cumulative
1	32.1919	31.9065	0.9912	0.9912
2	0.2854	.	0.0088	1.0000

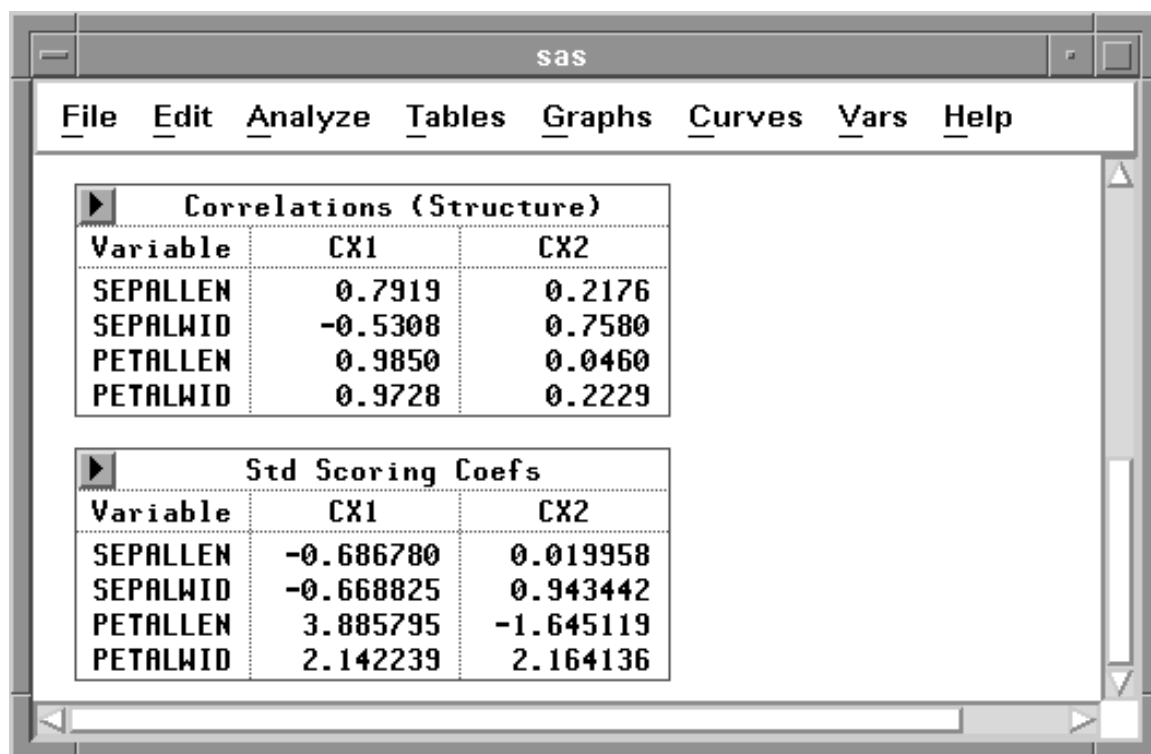
Test of H0: CanCorr[j]=0, j>=K

K	L. Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.023439	199.1453	8	288.0000	<.0001
2	0.777973	13.7939	3	145.0000	<.0001

Figure 40.32. Canonical Correlations Tables

The **Correlations (Structure)** table includes the correlations between the input **X** variables and the canonical variables. The scoring coefficients are the coefficients of the **X** variables that are used to compute canonical variable scores. The **Std Scoring Coefs** table includes the scoring coefficients of the standardized **X** variables, and the **Raw Scoring Coefs** table includes the scoring coefficients of the centered **X** variables.

Figure 40.33 shows tables of correlations between the **X** variables and the first two canonical variables, and the scoring coefficients of the standardized **X** variables.



The image shows a screenshot of the SAS software interface. The window title is 'sas'. The menu bar includes 'File', 'Edit', 'Analyze', 'Tables', 'Graphs', 'Curves', 'Vars', and 'Help'. The main display area contains two tables. The first table is titled 'Correlations (Structure)' and the second is titled 'Std Scoring Coefs'. Both tables have three columns: 'Variable', 'CX1', and 'CX2'. The first table shows correlations for variables SEPALLEN, SEPALWID, PETALLEN, and PETALWID. The second table shows the standardized scoring coefficients for the same variables.

Correlations (Structure)		
Variable	CX1	CX2
SEPALLEN	0.7919	0.2176
SEPALWID	-0.5308	0.7580
PETALLEN	0.9850	0.0460
PETALWID	0.9728	0.2229

Std Scoring Coefs		
Variable	CX1	CX2
SEPALLEN	-0.686780	0.019958
SEPALWID	-0.668825	0.943442
PETALLEN	3.885795	-1.645119
PETALWID	2.142239	2.164136

Figure 40.33. Correlations and Scoring Coefficients Tables

Graphs

You can create a scatter plot matrix and plots corresponding to various multivariate analyses by setting options in the Output Options dialogs, as shown in [Figure 40.5](#) to [Figure 40.10](#), or by choosing from the **Graphs** menu, as shown in [Figure 40.34](#).

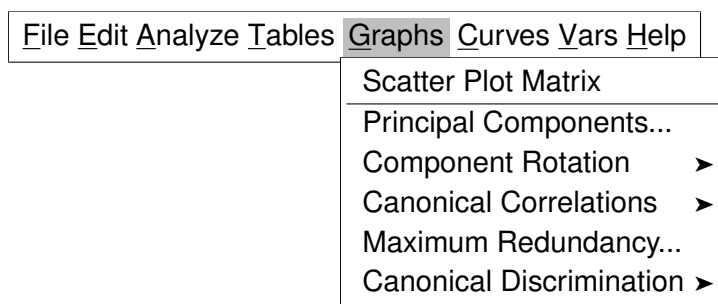


Figure 40.34. Graphs Menu

Scatter Plot Matrix

Scatter plots are displayed for pairs of variables. Without **X** variables, scatter plots are displayed as a symmetric matrix containing each pair of **Y** variables. With a nominal **Y** variable, scatter plots are displayed as a symmetric matrix containing each pair of **X** variables. When both interval **Y** variables and interval **X** variables are selected, scatter plots are displayed as a rectangular matrix with **Y** variables as the row variables and **X** variables as the column variables.

Figure 40.35 displays part of a scatter plot matrix with 80% prediction confidence ellipses.

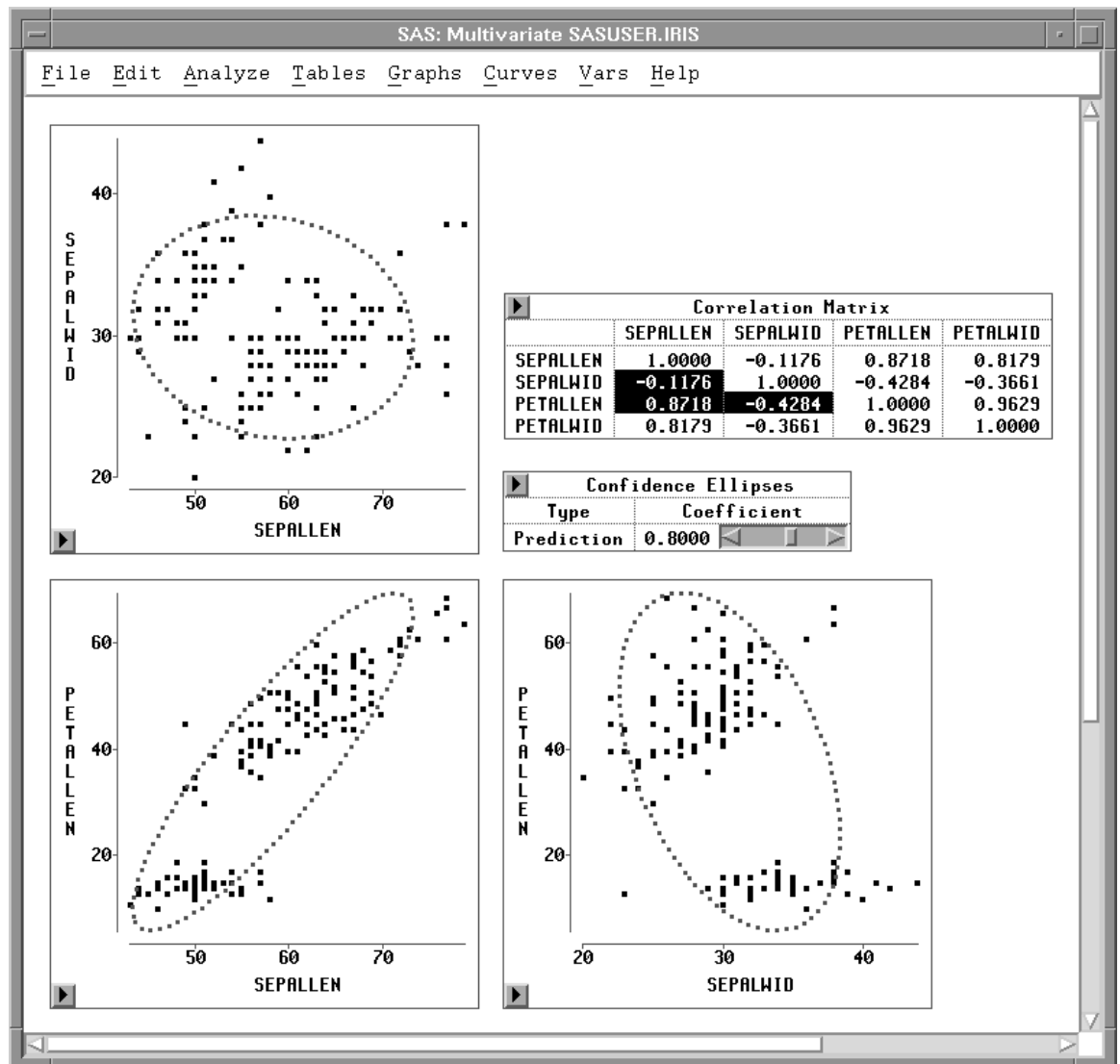


Figure 40.35. Scatter Plot Matrix with 80% Prediction Confidence Ellipses

Principal Component Plots

You can use principal component analysis to transform the **Y** variables into a smaller number of principal components that account for most of the variance of the **Y** variables. The plots of the first few components can reveal useful information about the distribution of the data, such as identifying different groups of the data or identifying observations with extreme values (possible outliers).

You can request a plot of the first two principal components or the first three principal components from the Principal Components Options dialog, shown in [Figure 40.6](#), or from the **Graphs** menu, shown in [Figure 40.34](#). Select **Principal Components** from the **Graphs** menu to display the **Principal Component Plots** dialog.

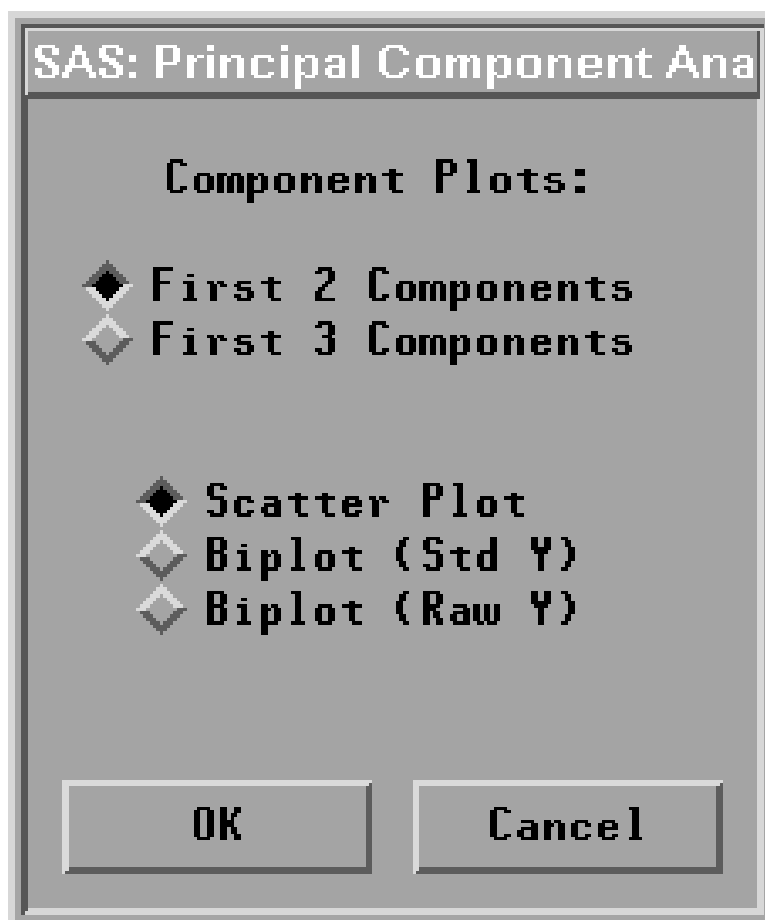


Figure 40.36. Principal Component Plots Dialog

In the dialog, you choose a principal component scatter plot (**Scatter Plot**), a principal component biplot with standardized **Y** variables (**Biplot (Std Y)**), or a principal component biplot with centered **Y** variables (**Biplot (Raw Y)**).

A *biplot* is a joint display of two sets of variables. The data points are first displayed in a scatter plot of principal components. With the approximated **Y** variable axes also displayed in the scatter plot, the data values of the **Y** variables are graphically estimated.

The **Y** variable axes are generated from the regression coefficients of the **Y** variables on the principal components. The lengths of the axes are approximately proportional to the standard deviations of the variables. A closer parallel between a **Y** variable axis and a principal component axis indicates a higher correlation between the two variables.

Reference ♦ *Multivariate Analyses*

For a **Y** variable **Y1**, the **Y1** variable value of a data point *y* in a principal component biplot is geometrically evaluated as follows:

- A perpendicular is dropped from point *y* onto the **Y1** axis.
- The distance from the origin to this perpendicular is measured.
- The distance is multiplied by the length of the **Y1** axis; this gives an approximation of the **Y1** variable value for point *y*.

Two sets of variables are used in creating principal component biplots. One set is the **Y** variables. Either standardized or centered **Y** variables are used, as specified in the Principal Component Plots dialog, shown in [Figure 40.36](#).

The other set is the principal component variables. These variables have variances either equal to one or equal to corresponding eigenvalues. You specify the principal component variable variance in the Multivariate Method Options dialog, shown in [Figure 40.3](#).

† **Note:** A biplot with principal component variable variances equal to one is called a **GH'** biplot, and a biplot with principal component variable variances equal to corresponding eigenvalues is called a **JK'** biplot.

A biplot is a useful tool for examining data patterns and outliers. [Figure 40.37](#) shows a biplot of the first two principal components from the correlation matrix and a rotating plot of the first three principal components. The biplot shows that the variable SEPALWID (highlighted axis) has a moderate negative correlation with PCR1 and a high correlation with PCR2.

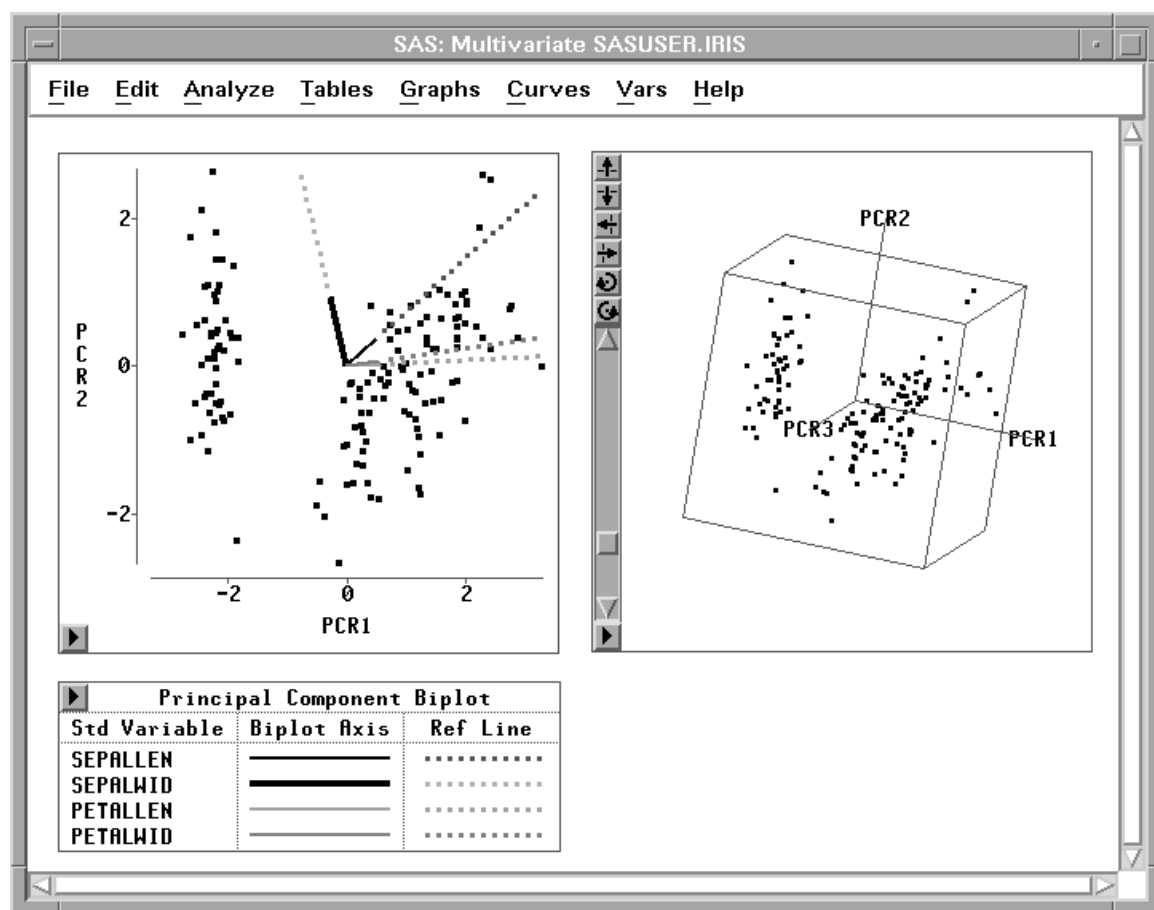


Figure 40.37. Principal Component Plots

Component Rotation Plots

You can request a plot of the rotated principal components from the Principal Components Rotation Options dialog, shown in [Figure 40.7](#), or from the **Component Rotation** menu, shown in [Figure 40.38](#).

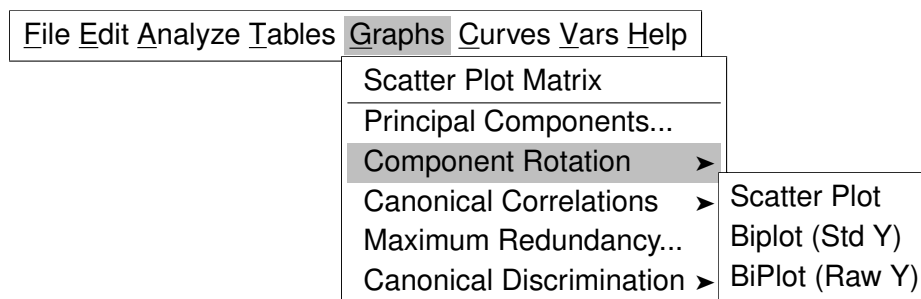


Figure 40.38. Component Rotation Menu

In the menu, you select a rotated component scatter plot (**Scatter Plot**), a rotated component biplot with standardized **Y** variables (**Biplot (Std Y)**), or a rotated component biplot with centered **Y** variables (**Biplot (Raw Y)**).

In a component rotation plot, the data points are displayed in a scatter plot of rotated principal components. With the approximated **Y** variable axes also displayed in the scatter plot, the data values of the **Y** variables are graphically estimated, as described previously in the “Principal Component Plots” section.

[Figure 40.39](#) shows a biplot of the rotated first two principal components with standardized **Y** variables. The biplot shows that the variable **SEPALWID** (highlighted axis) has a high correlation with **RT2** and that the other three **Y** variables all have high correlations with **RT1**.

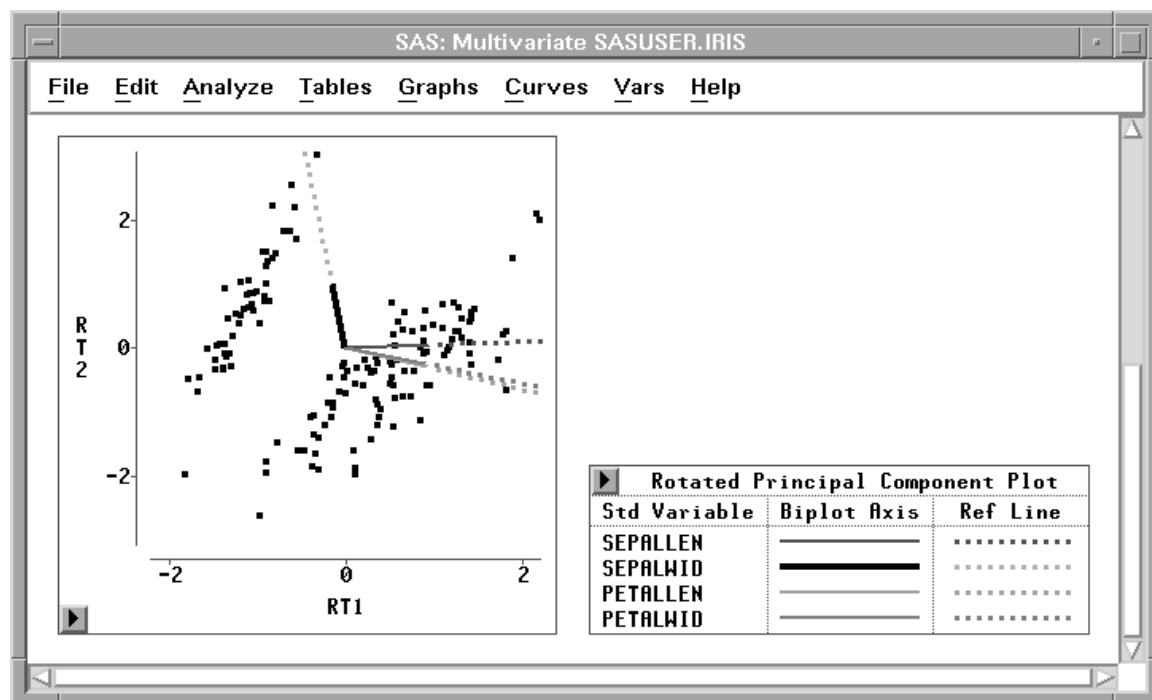


Figure 40.39. Rotated Principal Component Biplots

Canonical Correlation Plots

You can request pairwise canonical variable plots and a plot of the first two canonical variables or the first three canonical variables from each variable set from the Canonical Correlation Options dialog, shown in [Figure 40.8](#), or from the **Graphs** menu, shown in [Figure 40.40](#).

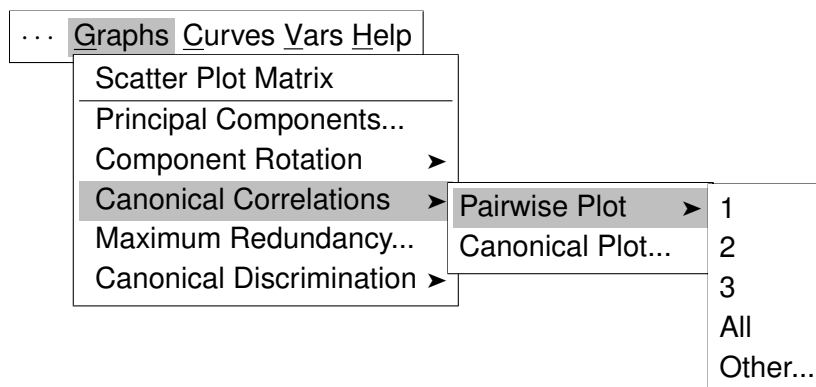


Figure 40.40. Canonical Correlations Menu

[Figure 40.41](#) shows scatter plots of the first two pairs of canonical variables. The first scatter plot shows a high canonical correlation (0.7956) between canonical variables **CX1** and **CY1** and the second scatter plot shows a low correlation (0.2005) between **CX2** and **CY2**.

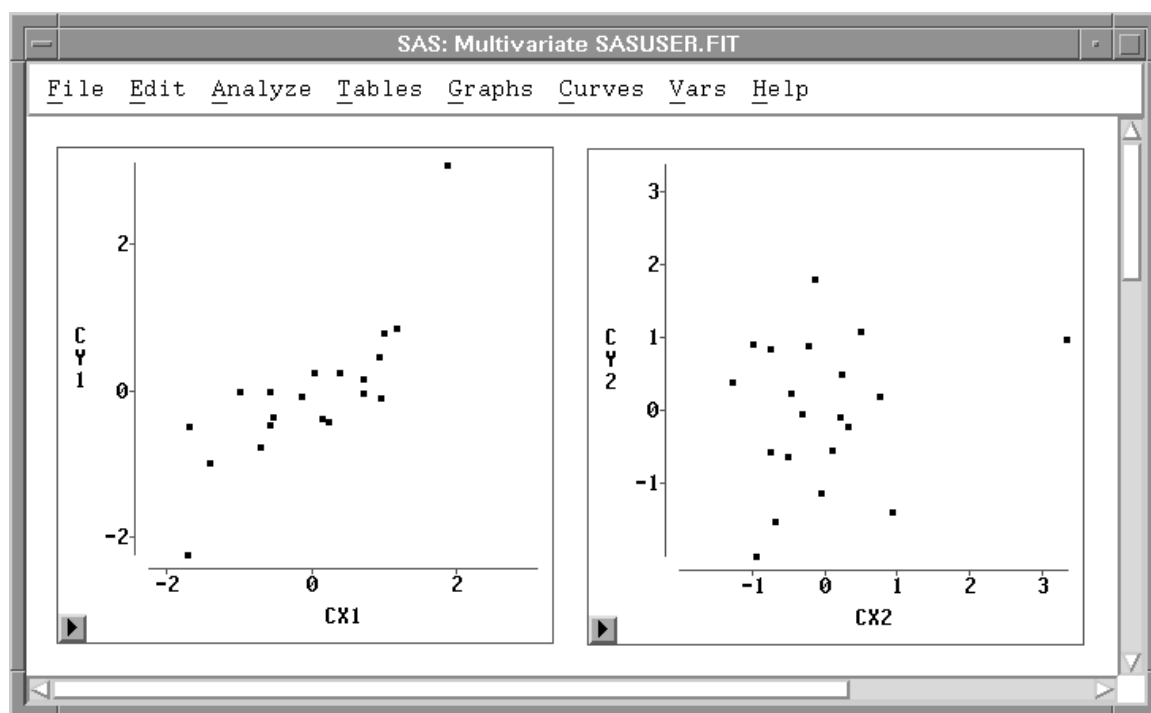


Figure 40.41. Canonical Correlation Pairwise Plots

Select **Canonical Plot** from the **Canonical Correlations** menu in [Figure 40.40](#) to display a Canonical Correlation Component Plots dialog.

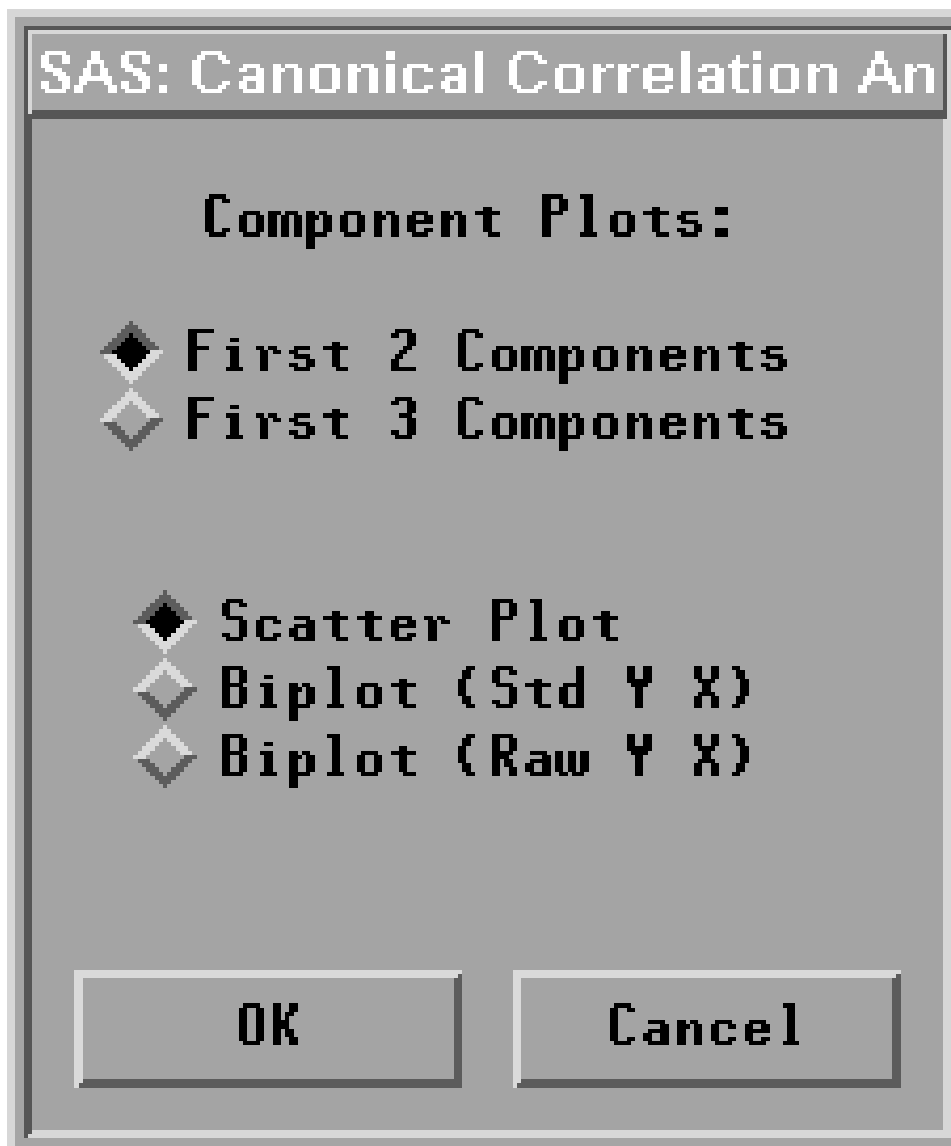


Figure 40.42. Canonical Correlation Component Plots Dialog

In the dialog, you choose a canonical correlation component scatter plot (**Scatter Plot**), a component biplot with standardized **Y** and **X** variables (**Biplot (Std Y X)**), or a component biplot with centered **Y** and **X** variables (**Biplot (Raw Y X)**).

In a canonical correlation component biplot, the data points are displayed in a scatter plot of canonical correlation components. With the approximated **Y** and **X** variable axes also displayed in the scatter plot, the data values of the **Y** and **X** variables are graphically estimated, as described previously in the "Principal Component Plots" section.

Figure 40.43 shows a biplot of the first two canonical variables from the **Y** variable sets with standardized **Y** and **X** variables. The biplot shows that the variables **WEIGHT** and **WAIST** (highlighted axes) have positive correlations with **CY1** and negative correlations with **CY2**. The other four variables have negative correlations with **CY1** and positive correlations with **CY2**.

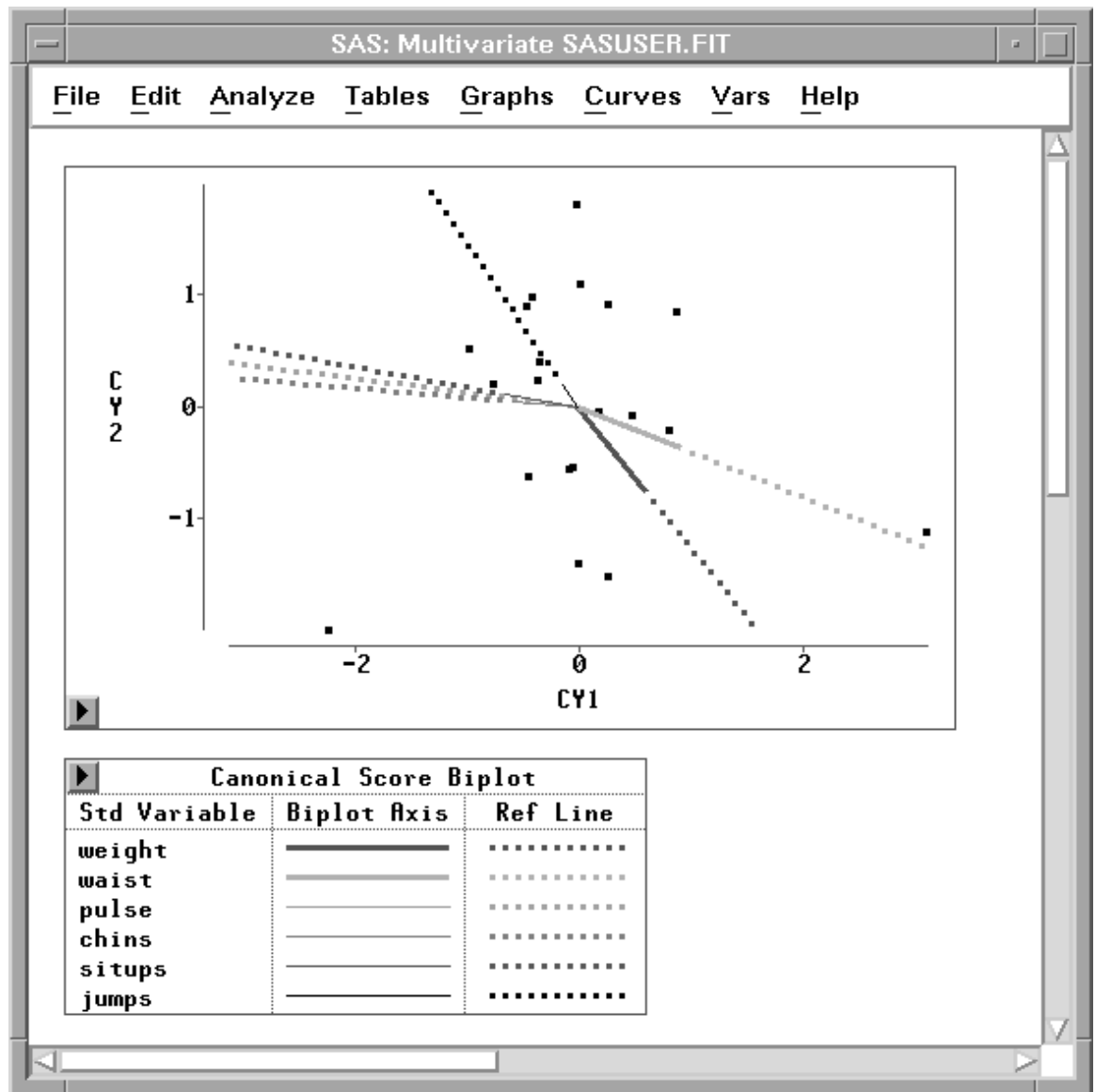


Figure 40.43. Canonical Correlation Component Biplot

Maximum Redundancy Plots

You can request a plot of the first two canonical variables or the first three canonical variables from each variable set from the Maximum Redundancy Options dialog, shown in [Figure 40.9](#), or from the **Graphs** menu, shown in [Figure 40.34](#). Select **Maximum Redundancy** from the **Graphs** menu to display a Maximum Redundancy Component Plots dialog.

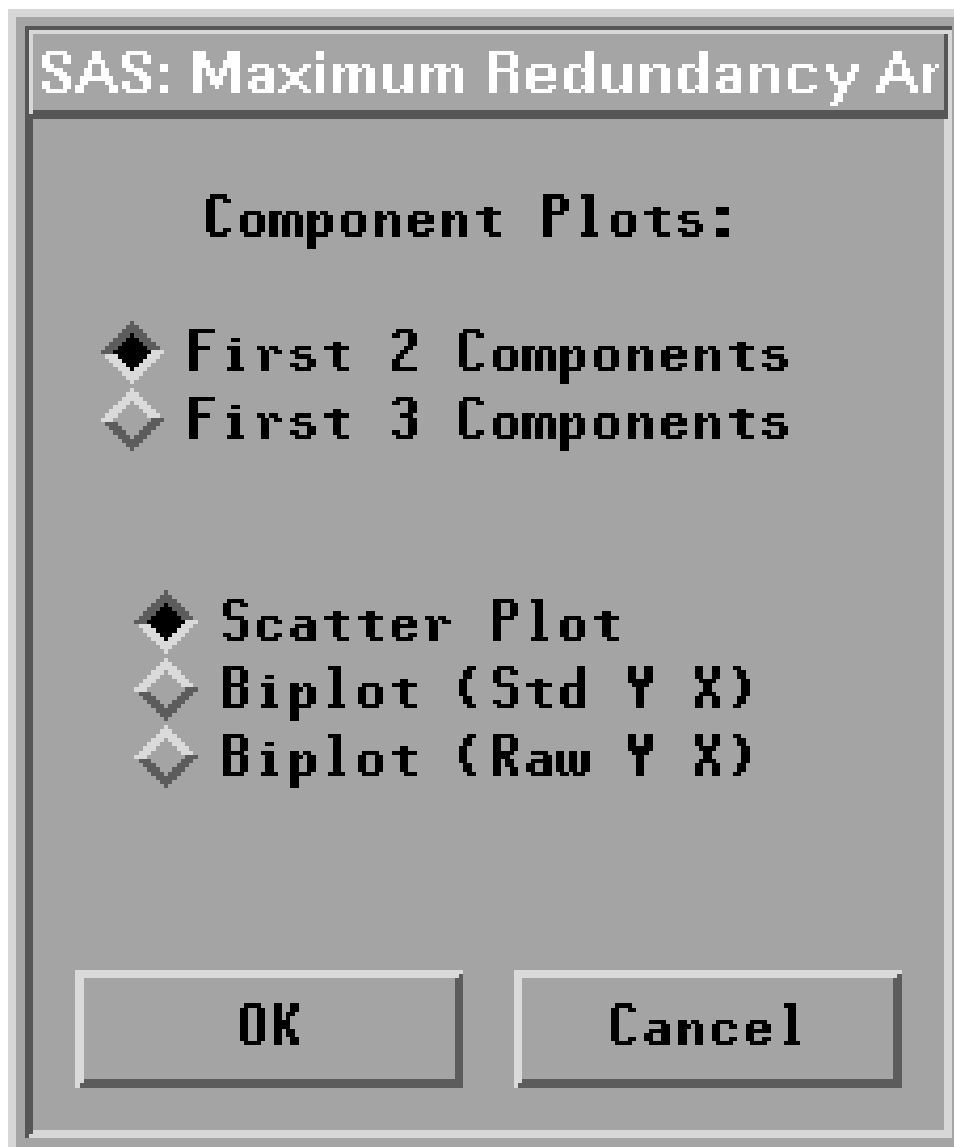


Figure 40.44. Maximum Redundancy Component Plots Dialog

In the dialog, you choose a maximum redundancy component scatter plot (**Scatter Plot**), a component biplot with standardized **Y** and **X** variables (**Biplot (Std Y X)**), or a component biplot with centered **Y** and **X** variables (**Biplot (Raw Y X)**).

In a maximum redundancy component biplot, the data points are displayed in a scatter

plot of maximum redundancy components. With the approximated **Y** and **X** variable axes also displayed in the scatter plot, the data values of the **Y** and **X** variables are graphically estimated, as described previously in the “Principal Component Plots” section.

Figure 40.45 shows scatter plots of the first two canonical variables from each set of variables. The canonical variables in each plot are uncorrelated.

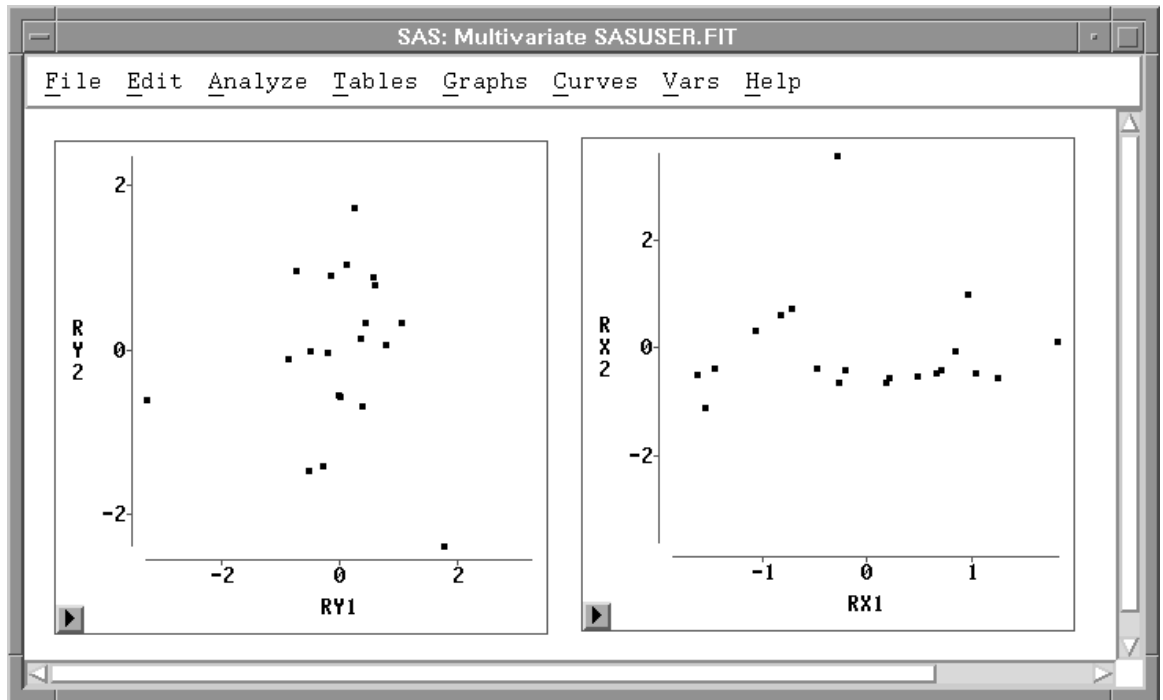


Figure 40.45. Maximum Redundancy Component Scatter Plots

Canonical Discrimination Plots

You can request a bar chart for the **Y** variable and a plot of the first two canonical variables or the first three canonical variables from the canonical discriminant options dialog, shown in [Figure 40.10](#), or from the **Graphs** menu, shown in [Figure 40.46](#).

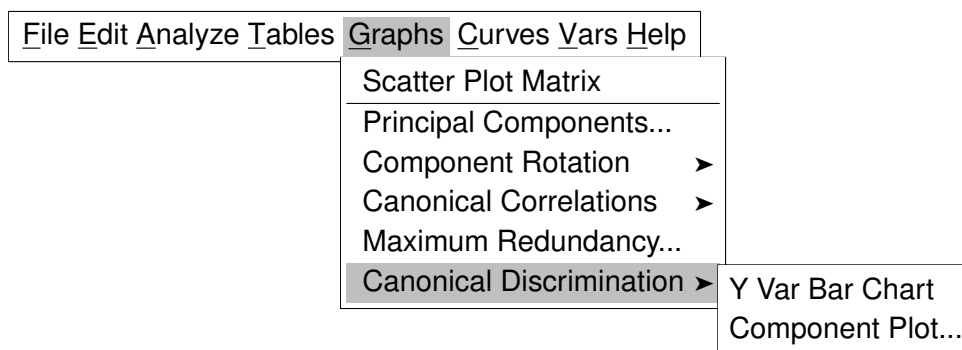


Figure 40.46. Canonical Discrimination Menu

[Figure 40.47](#) shows a bar chart for the variable **SPECIES**.

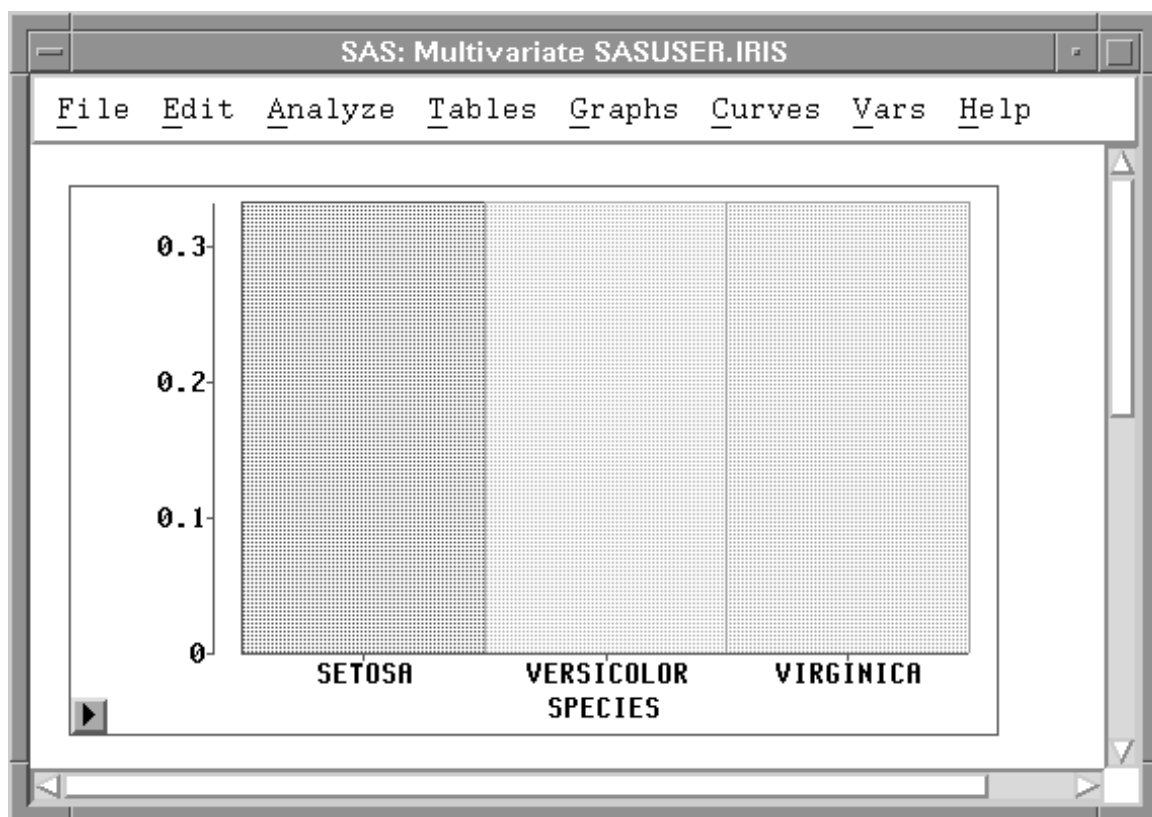


Figure 40.47. Y Var Bar Chart

Select **Component Plot** from the **Canonical Discriminant** menu in [Figure 40.48](#) to display a Canonical Discriminant Component Plots dialog.

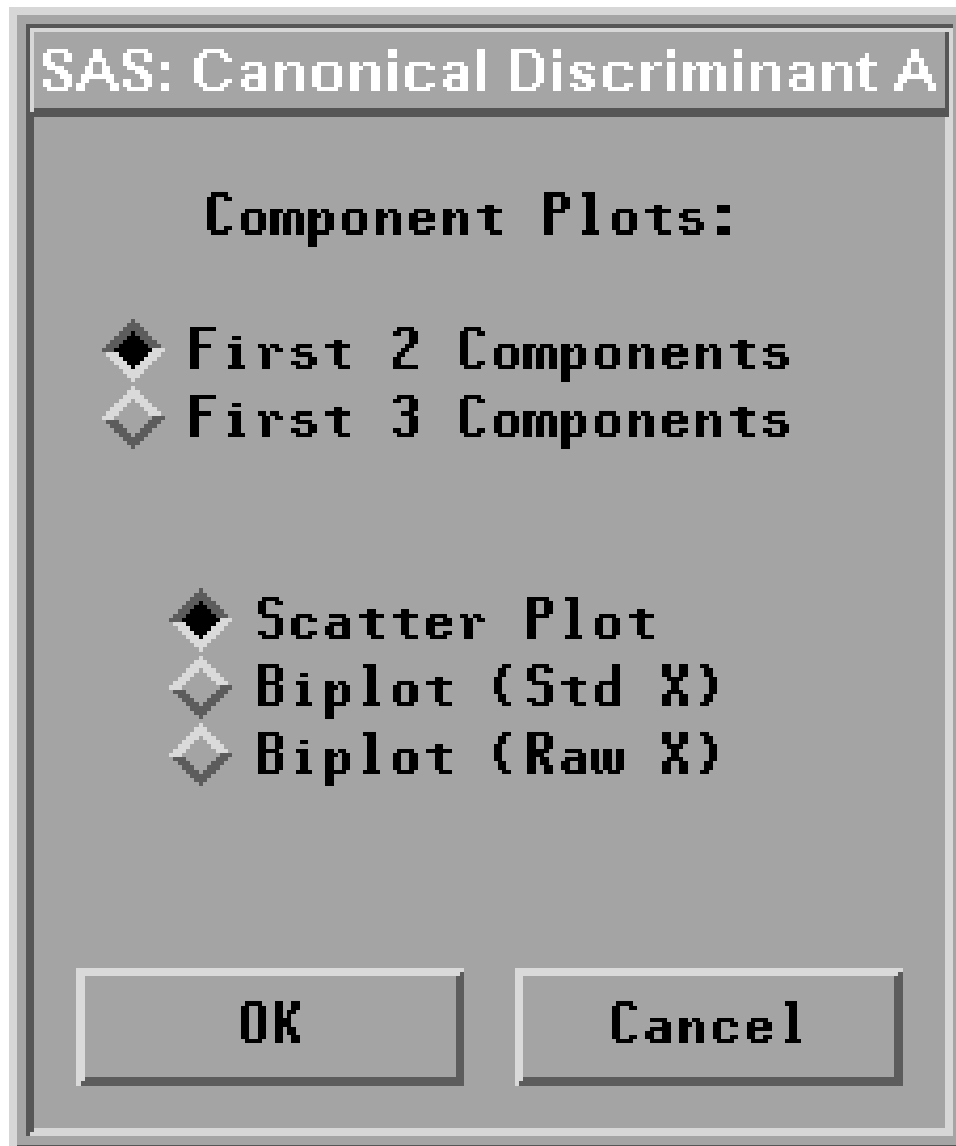


Figure 40.48. Canonical Discriminant Component Plots Dialog

In the dialog, you choose a canonical discriminant component scatter plot (**Scatter Plot**), a component biplot with standardized **X** variables (**Biplot (Std X)**), or a component biplot with centered **X** variables (**Biplot (Raw X)**).

In a canonical discriminant component biplot, the data points are displayed in a scatter plot of canonical discriminant components. With the approximated **X** variable axes also displayed in the scatter plot, the data values of the **X** variables are graphically estimated, as described previously in the "Principal Component Plots" section.

[Figure 40.49](#) shows a biplot of the first two canonical variables from the **X** vari-

able set with centered **X** variables. The biplot shows that the variable SEPALWID (highlighted axis) has a moderate negative correlation with CX1 and the other three variables have high correlation with CX1.

† **Note:** Use caution when evaluating distances in the biplot when the axes do not have comparable scales.

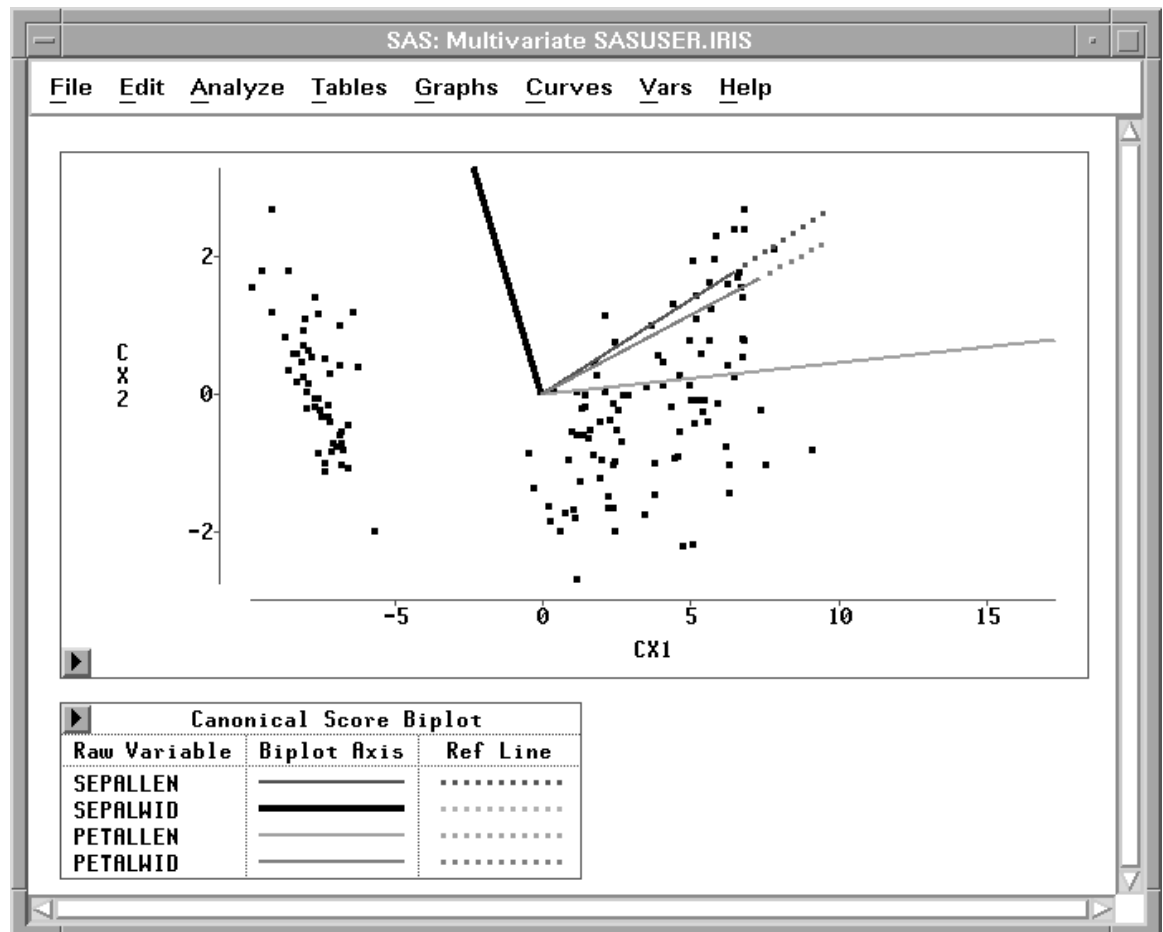


Figure 40.49. Canonical Discrimination Component Plot

Confidence Ellipses

SAS/INSIGHT software provides two types of confidence ellipses for pairs of analysis variables. One is a confidence ellipse for the population mean, and the other is a confidence ellipse for prediction. A confidence ellipse for the population mean is displayed with dashed lines, and a confidence ellipse for prediction is displayed with dotted lines.

Using these confidence ellipses assumes that each pair of variables has a bivariate normal distribution. Let $\bar{\mathbf{Z}}$ and \mathbf{S} be the sample mean and the unbiased estimate of the covariance matrix of a random sample of size n from a bivariate normal distribution with mean μ and covariance matrix Σ .

The variable $\bar{\mathbf{Z}} - \mu$ is distributed as a bivariate normal variate with mean 0 and covariance $n^{-1}\Sigma$, and it is independent of \mathbf{S} . The confidence ellipse for μ is based on Hotelling's T^2 statistic:

$$T^2 = n(\bar{\mathbf{Z}} - \mu)' \mathbf{S}^{-1} (\bar{\mathbf{Z}} - \mu)$$

A $100(1 - \alpha)\%$ confidence ellipse for μ is defined by the equation

$$(\bar{\mathbf{Z}} - \mu)' \mathbf{S}^{-1} (\bar{\mathbf{Z}} - \mu) = \frac{2(n-1)}{n(n-2)} F_{2,n-2}(1 - \alpha)$$

where $F_{2,n-2}(1 - \alpha)$ is the $(1 - \alpha)$ critical value of an F variate with degrees of freedom 2 and $n - 2$.

A confidence ellipse for prediction is a confidence region for predicting a new observation in the population. It also approximates a region containing a specified percentage of the population.

Consider \mathbf{Z} as a bivariate random variable for a new observation. The variable $\mathbf{Z} - \bar{\mathbf{Z}}$ is distributed as a bivariate normal variate with mean 0 and covariance $(1 + 1/n)\Sigma$, and it is independent of \mathbf{S} .

A $100(1 - \alpha)\%$ confidence ellipse for prediction is then given by the equation

$$(\mathbf{Z} - \bar{\mathbf{Z}})' \mathbf{S}^{-1} (\mathbf{Z} - \bar{\mathbf{Z}}) = \frac{2(n+1)(n-1)}{n(n-2)} F_{2,n-2}(1 - \alpha)$$

The family of ellipses generated by different F critical values has a common center (the sample mean) and common major and minor axes.

The ellipses graphically indicate the correlation between two variables. When the variable axes are standardized (by dividing the variables by their respective standard deviations), the ratio of the two axis lengths (in Euclidean distances) reflects the magnitude of the correlation between the two variables. A ratio of 1 between the major and minor axes corresponds to a circular confidence contour and indicates that the variables are uncorrelated. A larger value of the ratio indicates a larger positive or negative correlation between the variables.

Scatter Plot Confidence Ellipses

You can generate confidence ellipses by setting the options in the multivariate output options dialog, shown in [Figure 40.5](#), or by choosing from the **Curves** menu, shown in [Figure 40.50](#).

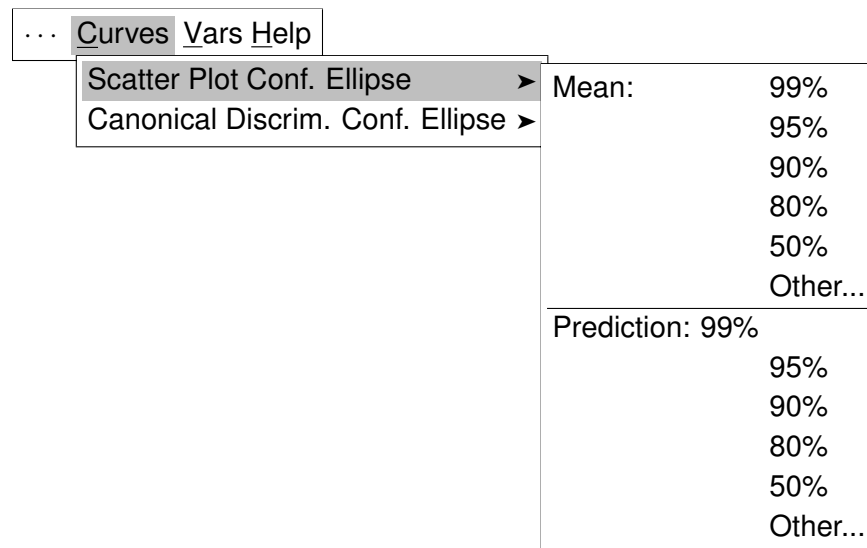


Figure 40.50. Curves Menu

Only 80% prediction confidence ellipses can be selected in the multivariate output options dialog. You must use the **Curves** menu to display mean confidence ellipses. You can use the confidence coefficient slider in the **Confidence Ellipses** table to change the coefficient for these ellipses.

[Figure 40.35](#) displays part of a scatter plot matrix with 80% prediction confidence ellipses and the **Correlation Matrix** table with corresponding correlations highlighted. The ellipses graphically show a small negative correlation (-0.1176) between variables **SEPALLEN** and **SEPALWID**, a moderate negative correlation (-0.4284) between variables **SEPALWID** and **PETALLEN**, and a large positive correlation (0.8718) between variables **SEPALLEN** and **PETALLEN**.

† **Note:** The confidence ellipses displayed in this illustration may not be appropriate since none of the scatter plots suggest bivariate normality.

Canonical Discriminant Confidence Ellipses

You can also generate class-specific confidence ellipses for the first two canonical components in canonical discriminant analysis by setting the options in the Canonical Discriminant Options dialog, shown in [Figure 40.10](#), or by choosing from the pre-deeding **Curves** menu.

[Figure 40.51](#) displays a scatter plot of the first two canonical components with class-specific 80% prediction confidence ellipses. The figure shows that the first canonical variable **CX1** has most of the discriminatory power between the two canonical variables.

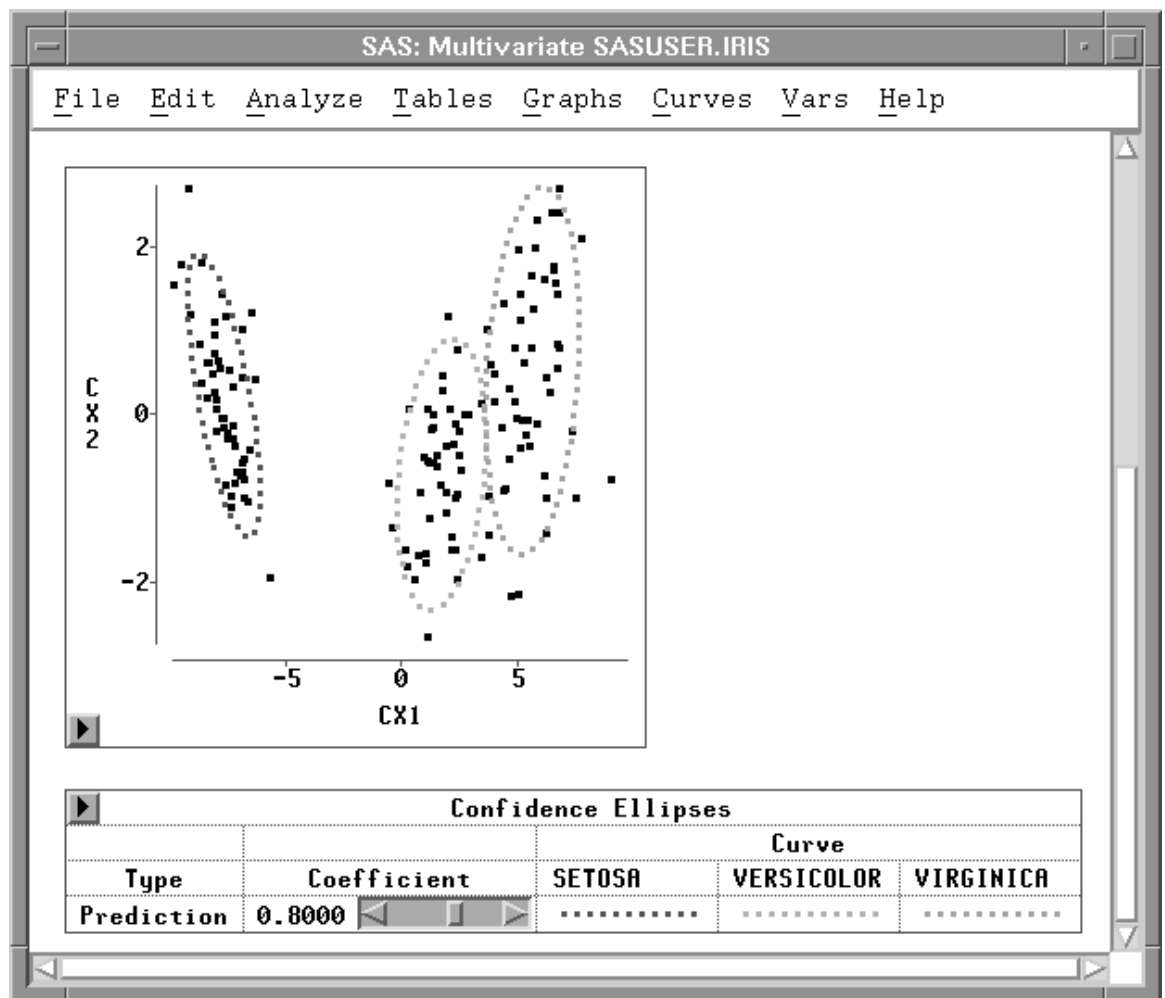


Figure 40.51. Canonical Discriminant Confidence Ellipses

Output Variables

You can save component scores from principal component analysis, component rotation, canonical correlation analysis, maximum redundancy analysis, and canonical discriminant analysis in the data window for use in subsequent analyses. For component rotation, the number of component output variables is the number of components rotated, as specified in [Figure 40.4](#). For other analyses, you specify the number of component output variables in the Output Options dialogs, shown in [Figure 40.6](#) to [Figure 40.10](#), or from the **Vars** menu, shown in [Figure 40.52](#). For component rotation, you specify the number of output rotated components in the Rotation Options dialog, shown in [Figure 40.4](#).

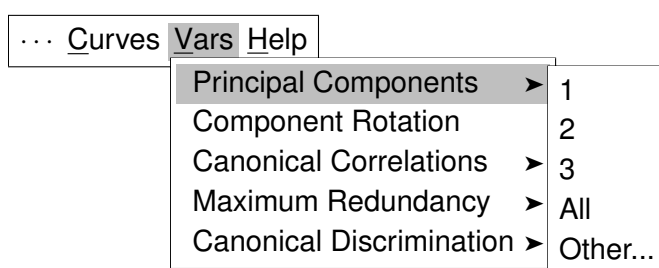


Figure 40.52. Vars Menu

Selecting **1**, **2**, or **3** gives you 1, 2, or 3 components. **All** gives you all components. Selecting **0** in the component options dialogs suppresses the output variables in the corresponding analysis. Selecting **Other** in the **Vars** menu displays the dialog shown in [Figure 40.53](#). You specify the number of components you want to save in the dialog.

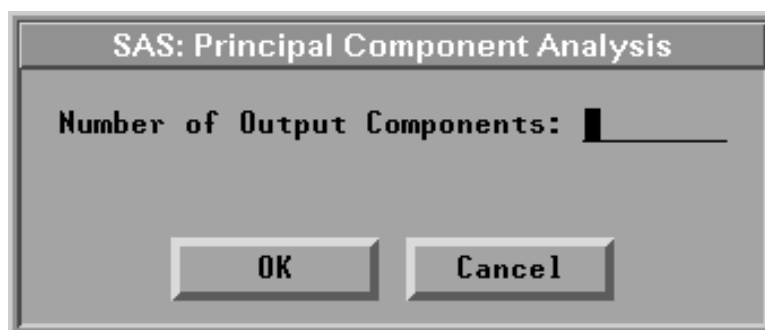


Figure 40.53. Output Components Dialog

Principal Components

For principal components from a covariance matrix, the names of the variables containing principal component scores are **PCV1**, **PCV2**, **PCV3**, and so on. The output component scores are a linear combination of the centered **Y** variables with coefficients equal to the eigenvectors of the covariance matrix.

For principal components from a correlation matrix, the names of the variables containing principal component scores are **PCR1**, **PCR2**, **PCR3**, and so on. The output component scores are a linear combination of the standardized **Y** variables with coefficients equal to the eigenvectors of the correlation matrix.

If you specify **Variance=Eigenvalues** in the multivariate method options dialog, the new variables of principal component scores have mean zero and variance equal to the associated eigenvalues. If you specify **Variance=1**, the new variables have variance equal to one.

Principal Component Rotation

The names of the variables containing rotated principal component scores are **RT1**, **RT2**, **RT3**, and so on. The new variables of rotated principal component scores have mean zero and variance equal to one.

Canonical Variables

The names of the variables containing canonical component scores are **CY1**, **CY2**, **CY3**, and so on, from the **Y** variable list, and **CX1**, **CX2**, **CX3**, from the **X** variable list. The new variables of canonical component scores have mean zero and variance equal to one.

Maximum Redundancy

The names of the variables containing maximum redundancy scores are **RY1**, **RY2**, **RY3**, and so on, from the **Y** variable list, and **RX1**, **RX2**, **RX3**, from the **X** variable list. The new variables of maximum redundancy scores have mean zero and variance equal to one.

Canonical Discriminant

The names of the variables containing canonical component scores are **CX1**, **CX2**, **CX3**, and so on. If you specify **Std Pooled Variance** in the multivariate method options dialog, the new variables of canonical component scores have mean zero and pooled within-class variance equal to one. If you specify **Std Total Variance**, the new variables have total-sample variance equal to one.

Weighted Analyses

When the observations are independently distributed with a common mean and unequal variances, a weighted analysis may be appropriate. The individual weights are the values of the **Weight** variable you specify.

The following statistics are modified to incorporate the observation weights:

- Mean \bar{y}_w, \bar{x}_w
- SSCP U_{yy}, U_{yx}, U_{xx}
- CSSCP C_{yy}, C_{yx}, C_{xx}
- COV S_{yy}, S_{yx}, S_{xx}
- CORR R_{yy}, R_{yx}, R_{xx}

The formulas for these weighted statistics are given in the “Method” section earlier in this chapter. The resulting weighted statistics are used in the multivariate analyses.

References

- Cooley, W.W. and Lohnes, P.R. (1971), *Multivariate Data Analysis*, New York: John Wiley & Sons, Inc.
- Dillon, W.R. and Goldstein, M. (1984), *Multivariate Analysis*, New York: John Wiley & Sons, Inc.
- Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.
- Gabriel, K.R. (1971), "The Biplot Graphical Display of Matrices with Application to Principal Component Analysis," *Biometrika*, 58, 453–467.
- Gnanadesikan, R. (1997), *Methods for Statistical Data Analysis of Multivariate Observations*, Second Edition, New York: John Wiley & Sons, Inc.
- Gower, J.C. and Hand, D.J. (1996), *Biplots*, New York: Chapman and Hall.
- Hotelling, H. (1933), "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, 24, 417–441, 498–520.
- Hotelling, H. (1935), "The Most Predictable Criterion," *Journal of Educational Psychology*, 26, 139–142.
- Hotelling, H. (1936), "Relations Between Two Sets of Variables," *Biometrika*, 28, 321–377.
- Jobson, J.D. (1992), *Applied Multivariate Data Analysis, Vol 2: Categorical and Multivariate Methods*, New York: Springer-Verlag.
- Kaiser, H.F. (1958), "The Varimax Criterion of Analytic Rotation in Factor Analysis," *Psychometrika*, 23, 187–200.
- Krzanowski, W.J. (1988), *Principles of Multivariate Analysis: A User's Perspective*, New York: Oxford University Press.
- Kshirsagar, A.M. (1972), *Multivariate Analysis*, New York: Marcel Dekker, Inc.
- Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis*, New York: Academic Press.
- Morrison, D.F. (1976), *Multivariate Statistical Methods*, Second Edition, New York: McGraw-Hill Book Co.
- Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, 6(2), 559–572.
- Pringle, R.M. and Raynor, A.A. (1971), *Generalized Inverse Matrices with Applications to Statistics*, New York: Hafner Publishing Co.
- Rao, C.R. (1964), "The Use and Interpretation of Principal Component Analysis in Applied Research," *Sankhya A*, 26, 329–358.
- Rao, C.R. (1973), *Linear Statistical Inference*, New York: John Wiley & Sons, Inc.

- Stewart, D.K. and Love, W.A. (1968), "A General Canonical Correlation Index," *Psychological Bulletin*, 70, 160–163.
- van den Wollenberg, A.L. (1977), "Redundancy Analysis—An Alternative to Canonical Correlation Analysis," *Psychometrika*, 42, 207–219.

Chapter 41

SAS/INSIGHT Statements

Chapter Contents

DETAILS	780
PROC INSIGHT Statement	781
WINDOW Statement	782
OPEN Statement	782
BY Statement	783
CLASS Statement	783
BAR Statement	783
BOX Statement	784
LINE Statement	784
SCATTER Statement	785
CONTOUR Statement	785
ROTATE Statement	786
DIST Statement	786
MULT Statement	786
FIT Statement	787
TABLES statement	788
RUN statement	788
QUIT statement	789

Chapter 41

SAS/INSIGHT Statements

You can submit SAS/INSIGHT statements to create graphs and analyses automatically. This saves time when you have repetitive analyses to perform or when you work with large data sets.

SAS/INSIGHT statements also provide a record of the analyses you create, including model equations. You can store statements in a text file or in the SAS log.

Included in this release are the new WINDOW statement, the OTHER= option, the MARKERSIZE= option, and axis options.

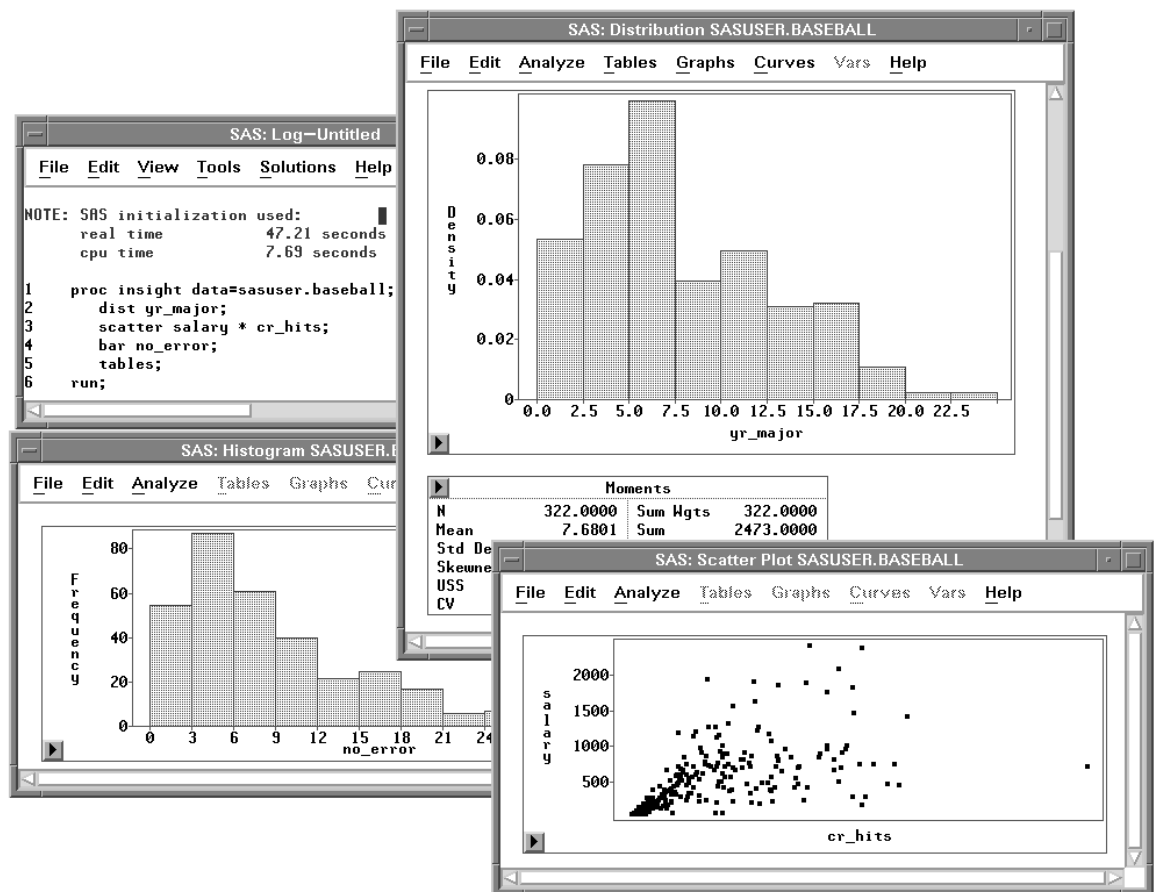


Figure 41.1. SAS/INSIGHT Statements and Output

Details

You can use the following statements when invoking SAS/INSIGHT either as a procedure or as a task. Brackets (<>) denote optional parameters.

```

PROC INSIGHT < INFILE=fileref > < FILE<=fileref> >
    < DATA=SAS-data-set > < TOOLS >
    < NOMENU > < NOBUTTON > < NOCONFIRM >;
WINDOW < x y width height > < / NOSCROLL >;
OPEN SAS-data-set < / NODISPLAY >;
BY < variable-list >;
CLASS variable-list;
BAR variable-list
    < / < FREQ=variable > < OTHER=value >
    < XAXIS=axis > < YAXIS=axis > >;
BOX variable-list < * variable-list >
    < / < FREQ=variable > < LABEL=variable >
    < OTHER=value > < < MARKERSIZE | MS >=value >
    < YAXIS=axis > >;
LINE variable-list * variable
    < / < LABEL=variable > < < MARKERSIZE | MS >=value >
    < XAXIS=axis > < YAXIS=axis > >;
SCATTER variable-list * variable-list
    < / < LABEL=variable > < < MARKERSIZE | MS >=value >
    < XAXIS=axis > < YAXIS=axis > >;
ROTATE variable-list * variable-list * variable-list
    < / < LABEL=variable > < < MARKERSIZE | MS >=value >
    < XAXIS=axis > < YAXIS=axis > < ZAXIS=axis > >;
DIST variable-list
    < / < FREQ=variable > < WEIGHT=variable >
    < LABEL=variable > >;
MULT variable-list
    < / < FREQ=variable > < WEIGHT=variable >
    < LABEL=variable > >;
FIT variable-list < = effects-list >
    < / < FREQ=variable > < WEIGHT=variable >
    < LABEL=variable > < NOINT >
    < RESP=response > < BINOM=variable >
    < OFFSET=variable > < LINK=link >
    < POWER=value > < NOEXACT > < FISHER >
    < QUASI > < SCALE=scale > < CONSTANT=value > >;
TABLES;
RUN;
QUIT;

```


Unless you override them with the options listed above, graph and analysis statements use options stored in your SASUSER.PROFILE catalog. For more information on SAS/INSIGHT options, see [Chapter 30, “Working with Other SAS Products.”](#)

The WINDOW statement and the NODISPLAY, OTHER=, MARKERSIZE=, and axis options can be used as input, but they are not recordable.

PROC INSIGHT Statement

```
PROC INSIGHT < INFILE=fileref > < FILE=<=fileref > >
               < DATA=SAS-data-set > < TOOLS >
               < NOMENU > < NOBUTTON > < NOCONFIRM >;
```

PROC INSIGHT options apply to both the procedure and the task. When invoking SAS/INSIGHT from the command line, you can follow the INSIGHT command with any of the PROC INSIGHT options.

INFILE=*fileref*

The INFILE= option directs SAS/INSIGHT software to read additional statements from the specified text file. For examples using the INFILE= option, see [Chapter 30, “Working with Other SAS Products.”](#)

FILE | **FILE=***fileref*

The FILE option directs SAS/INSIGHT software to write statements to the SAS log. FILE=*fileref* directs SAS/INSIGHT software to write statements to the text file *fileref*. For examples using the FILE option, see [Chapter 30, “Working with Other SAS Products.”](#)

DATA | **DATA=***SAS-data-set*

The DATA option opens a SAS data set and displays it in a window. If DATA is used without =*SAS-data-set*, a new data window is created. You can use either the DATA option or the OPEN statement to specify an initial data set. If you use neither, but simply enter “insight” or “proc insight; run;”, a data set dialog prompts you to choose an initial data set.

You can specify data set options in parentheses after the data set name. For example, to see all businesses that had large profits, you might enter

```
insight data=sasuser.business (where=(profits>=2000) )
```

Alternatively, you can enter data set options by pressing the **Options** button in the data set dialog. Data set options are described in *SAS Language Reference: Dictionary*.

TOOLS

The TOOLS option causes the Tools window to be displayed by default. If you use tools frequently, this option saves the step of choosing **Edit:Windows:Tools**.

NOMENU | NOMEN

The NOMENU option suppresses the display of menu bars. If your host defines a pop-up key, menu bars are still available when you press the pop-up key in an area containing no graphs or tables.

NOBUTTON | NOBUT

The NOBUTTON option suppresses the display of pop-up menu buttons. If your host defines a pop-up key, pop-up menus are still available when you press the pop-up key on graphs or tables.

NOCONFIRM | NOCON

The NOCONFIRM option suppresses the display of confirmation dialogs for potentially harmful user actions. Such actions include deleting variables, closing data windows, and exiting SAS/INSIGHT. By default, confirmation dialogs provide a chance to cancel these actions.

WINDOW Statement

WINDOW *< x y width height >* *< / NOSCROLL >*;

The WINDOW statement specifies the position of subsequently created windows. Parameters are percentage values between 0 and 100. If parameters are omitted, the next created window uses a default position.

For example, to position a window in the upper left corner, covering one quarter of the display, you might enter

```
window 0 0 50 50;
```

To restore default positioning, use

```
window;
```

You can use the NOSCROLL option to create windows without scroll bars. On most hosts, this option simplifies your display. However, it should be used only when creating single graphs for which scrolling is not needed.

OPEN Statement

OPEN *SAS-data-set* *< / NODISPLAY >*;

The OPEN statement opens a SAS data set and displays it in a window. An OPEN statement with the NODISPLAY option opens a data set without displaying a window.

You can use the OPEN statement to open multiple data sets at the same time. BY, CLASS, graph, and analysis statements apply only to the most recently opened data set.

You can specify data set options in parentheses after the data set name. For example, to see all businesses that had large profits, you might enter

```
open sasuser.business (where=(profits>=2000)) ;
```

Data set options are described in *SAS Language Reference: Dictionary*.

BY Statement

```
BY < variable-list >;
```

The BY statement assigns variables the group role in subsequent graphs and analyses.

To de-assign group roles, use the BY statement without specifying variables.

CLASS Statement

```
CLASS variable-list;
```

The CLASS statement sets the measurement level of the specified variables to nominal. Use this statement to override the default interval measurement level of numeric variables.

BAR Statement

```
BAR variable-list
```

```
< / < FREQ=variable > < OTHER=value >
```

```
< XAXIS=axis > < YAXIS=axis > >;
```

The BAR statement creates bar charts or histograms for the specified **Y** variables. You can use the FREQ= option to assign a **Frequency** variable.

Use the OTHER= option to set the “Other” threshold for nominal bar charts. The “Other” threshold is a percentage between 0 and 100.

Use the XAXIS= and YAXIS= options to specify axes for numeric variables with interval measurement level. The *axis* specification is a list of six numeric values: *First Tick*, *Last Tick*, *Tick Increment*, *Number of Minor Ticks*, *Axis Minimum*, and *Axis Maximum*.

For example, to specify tick marks ranging from 2 to 8, with tick increment 2, 1 minor tick, and Y axis ranging from 0 to 10, you could use

```
bar age / yaxis = 2 8 2 1 0 10;
```

Note that the “X” and “Y” prefixes refer to variable roles, not vertical or horizontal orientation. For the BAR statement, the YAXIS= option specifies the axis of the **Y** variable, and the XAXIS= option specifies the Frequency axis.

BOX Statement

```
BOX variable-list < * variable-list >
    < / < FREQ=variable > < LABEL=variable >
    < OTHER=value > < < MARKERSIZE | MS >=value >
    < YAXIS=axis > >;
```

The BOX statement creates box or mosaic plots. The BOX statement requires at least one list of **Y** variables, optionally followed by an asterisk (*) and a list of **X** variables. If the **Y** variables have interval measurement level, the BOX statement creates box plots. If the **Y** variables are nominal, the BOX statement creates mosaic plots.

If you use **X** variables, you get one plot for each **Y** variable, and each plot contains one schematic diagram for each combination of **X** values. If you use no **X** variables, you get one plot containing one schematic diagram for each **Y** variable.

You can use the **FREQ**= and **LABEL**= options to assign **Frequency** and **Label** variables.

Use the **OTHER**= option to set the “Other” threshold for mosaic plots. The “Other” threshold is a percentage between 0 and 100.

Use the **MARKERSIZE**= or **MS**= option to specify the size of observation markers. Marker size is a number between 1 and 8.

Use the **YAXIS**= option to specify a numeric axis for the **Y** variable. The syntax for axis options is described under the BAR statement.

LINE Statement

```
LINE variable-list * variable
    < / < LABEL=variable > < < MARKERSIZE | MS >=value >
    < XAXIS=axis > < YAXIS=axis > >;
```

The LINE statement creates overlaid line plots, with one line for each **Y** variable.

Use at least one **Y** variable, followed by an asterisk, followed by a single **X** variable. You can use the **LABEL**= option to assign a **Label** variable.

Use the **MARKERSIZE**= or **MS**= option to specify the size of observation markers. Marker size is a number between 1 and 8.

Use the **XAXIS**= and **YAXIS**= options to specify numeric axes. The syntax for axis options is described under the BAR statement.

SCATTER Statement

```
SCATTER variable-list * variable-list
      < / < LABEL=variable > < < MARKERSIZE | MS >=value >
      < XAXIS=axis > < YAXIS=axis > >;
```

The SCATTER statement creates two-dimensional scatter plots.

Use at least one **Y** variable, followed by an asterisk, followed by at least one **X** variable. Use multiple **Y** and **X** variables to create a scatter plot matrix. For example, you might use

```
scatter a b c * a b c;
```

to create a 3×3 scatter plot matrix for the variables **a**, **b**, and **c**.

You can use the LABEL= option to assign a **Label** variable.

Use the MARKERSIZE= or MS= option to specify the size of observation markers. Marker size is a number between 1 and 8.

Use the XAXIS= and YAXIS= options to specify numeric axes. The syntax for axis options is described under the BAR statement.

CONTOUR Statement

```
CONTOUR variable-list * variable-list * variable-list
      < / < LABEL=variable > < < MARKERSIZE | MS >=value >
      < XAXIS=axis > < YAXIS=axis > < ZAXIS=axis > >;
```

The CONTOUR statement creates level curves of a surface that fits the data, assuming that the **Z** variable is a function of the **X** and **Y** variables.

Use at least one **Z** variable, followed by an asterisk, followed by at least one **Y** variable, followed by an asterisk, followed by at least one **X** variable. Use multiple **Z**, **Y**, and **X** variables to create a matrix of contour plots.

You can use the LABEL= option to assign a **Label** variable.

Use the MARKERSIZE= or MS= option to specify the size of observation markers. Marker size is a number between 1 and 8.

Use the XAXIS=, YAXIS=, and ZAXIS= options to specify numeric axes. The syntax for axis options is described under the BAR statement.

ROTATE Statement

```

ROTATE variable-list * variable-list * variable-list
  < / < LABEL=variable > < < MARKERSIZE / MS >=value >
  < XAXIS=axis > < YAXIS=axis > < ZAXIS=axis > >;

```

The ROTATE statement creates three-dimensional rotating plots.

Use at least one **Z** variable, followed by an asterisk, followed by at least one **Y** variable, followed by an asterisk, followed by at least one **X** variable. Use multiple **Z**, **Y**, and **X** variables to create a rotating plot matrix. For example, you might use

```

rotate a b c d * a b c d * a b c d;

```

to create a matrix displaying all possible three-dimensional plots for the variables **a**, **b**, **c**, and **d**.

You can use the LABEL= option to assign a **Label** variable.

Use the MARKERSIZE= or MS= option to specify the size of observation markers. Marker size is a number between 1 and 8.

Use the XAXIS=, YAXIS=, and ZAXIS= options to specify numeric axes. Syntax of axis options is described under the BAR statement.

DIST Statement

```

DIST variable-list
  < / < FREQ=variable > < WEIGHT=variable >
  < LABEL=variable > >;

```

The DIST statement creates a distribution analysis of the specified **Y** variables. You can use the FREQ=, WEIGHT=, and LABEL= options to assign **Freq**, **Weight**, and **Label** variables.

MULT Statement

```

MULT variable-list
  < / < FREQ=variable > < WEIGHT=variable >
  < LABEL=variable > >;

```

The MULT statement creates a multivariate analysis of the specified **Y** variables. You can use the FREQ=, WEIGHT=, and LABEL= options to assign **Freq**, **Weight**, and **Label** variables.

FIT Statement

```

FIT variable-list < = effects-list >
  < / < FREQ=variable > < WEIGHT=variable >
  < LABEL=variable > < NOINT >
  < RESP=response > < BINOM=variable >
  < OFFSET=variable > < LINK=link >
  < POWER=value > < NOEXACT > < FISHER >
  < QUASI > < SCALE=scale > < CONSTANT=value > >;

```

The FIT statement creates a fit analysis.

You must specify at least one **Y** variable. You can follow the **Y** variables with an equal sign (=) and a list of model effects, including simple, crossed, and nested effects:

```

Y = X
Y = X1 * X2
Y = X( A B )

```

If you do not specify an effects list, a model with only an intercept term (mean) is fit.

You can use the FREQ=, WEIGHT=, and LABEL= options to assign **Freq**, **Weight**, and **Label** variables.

FIT statement options default to fit classical linear models, but you can set them to fit the generalized linear model.

NOINT

Use the NOINT option to fit a model without an intercept term.

RESP=response

For response distribution, choose NORMAL, INVGAUSS, GAMMA, POISSON, or BINOMIAL. By default, RESP= is NORMAL.

BINOM=variable

Use the BINOM= option to specify a **Binomial** variable when RESP=Binomial. When RESP is not Binomial, the BINOM= option is not used.

OFFSET=variable

Use the OFFSET= option to specify an **Offset** variable.

LINK=link

For link function, choose CANONICAL, IDENTITY, LOG, LOGIT, PROBIT, CLOGCLOG, or POWER. By default, LINK= is CANONICAL.

POWER=value

Use the POWER= option to set a value for the POWER link function. If LINK= is not set to POWER, the POWER= option is not used.

NOEXACT

Use the NOEXACT option to fit a linear model without using exact distributions for the test statistics.

FISHER

Use the FISHER option to use Fisher's scoring method in the maximum-likelihood estimation for the regression parameters.

QUASI

If overdispersion is present in the model, you can use the QUASI option to fit the generalized linear model using the quasi-likelihood functions. To use the QUASI option, you must also set the SCALE= option to a scale other than **MISCALE=scale**. For scale, choose MLE, DEVIANCE, PEARSON, or CONSTANT. By default, SCALE= is MLE (maximum-likelihood estimate).

CONSTANT=value

Use the CONSTANT= option to set a constant value when SCALE=CONSTANT. If SCALE= is not set to CONSTANT, the CONSTANT= option is not used.

TABLES statement

TABLES;

The TABLES statement saves and prints all tables in the most recent analysis, using the Output Delivery System.

You can redirect output to a text file by using the PRINTTO procedure.

For more information on PROC PRINTTO, refer to the *SAS Procedures Guide*.

RUN statement

RUN;

The RUN statement invokes SAS/INSIGHT software and executes all preceding SAS/INSIGHT statements.

Use the RUN statement if you want SAS/INSIGHT to remain available after executing your statements. You must terminate the list of statements with either a RUN or a QUIT statement.

QUIT statement

QUIT;

The QUIT statement invokes SAS/INSIGHT software, executes all preceding statements, and exits SAS/INSIGHT software.

Use the QUIT statement if you do not want SAS/INSIGHT to remain available after executing your statements. You must terminate the list of statements with either a QUIT or a RUN statement.

Index

Default

OBSTAT variable, 422, 474

A

adding

- curves, 207
- effects, 220, 247, 249, 630
- graphs, 228
- observations, 35
- tables, 228
- variables, 35

adding graphs, 404, 406, 407
to fit window, 230, 255

adding tables and graphs
multiple regression, 228

Adj R-Sq,
see adjusted R-square

adjust
speed of animation, 371

Adjusted R-Square, 640, 642

adjusted R-square, 225, 251, 352
in multiple regression, 235

adjusting
axes, 129, 131, 361
ticks, 126, 361

adjusting axes,
see aligning axes
see ticks

Afifi, A.A., 19

AIR data set, 18, 100, 116, 118, 369

algorithm,
see method

Align, 362, 417

aligning
axes, 417
graphs, 417
ticks, 417

AMISE,
see approximate mean integrated square error

analyses
comparing, 339

analysis, 5

analysis of covariance, 244

analysis of deviance, 643
logistic regression, 270
Poisson regression, 287

analysis of deviance for generalized linear models
fit analyses, 644

analysis of variance, 244

analysis of variance table, 252
assigning measurement levels, 245
examining the means, 255
multiple regression, 225
parameter estimates, 253
residuals-by-predicted plot, 254
summary of fit, 251
type III tests, 252

analysis of variance for linear models
fit analyses, 643

analysis of variance table
analysis of variance, 252

Analyze, 81, 410

Analyze:Bar Chart (Y), 500

Analyze:Box Plot (Y), 387

Analyze:Box Plot (Y), 330

Analyze:Box Plot/Mosaic Plot (Y), 80, 88,
255, 364, 454, 509

Analyze:Contour Plot (Z Y X), 118, 537

Analyze:Distribution (Y), 381, 404, 555, 558

Analyze:Distribution (Y), 343, 358

Analyze:Fit (Y X), 202, 263, 266, 279, 284, 613,
629

Analyze:Fit (Y X), 220, 244, 327, 345, 349

Analyze:Histogram/Bar Chart (Y), 72, 76

Analyze:Line Plot (Y X), 100, 522

Analyze:Multivariate (Y's), 296, 707, 708

Analyze:Multivariate (Y's), 306

Analyze:Rotating Plot (Z Y X), 110, 115,
116, 547

Analyze:Scatter Plot (Y X), 91, 94, 136, 270,
314, 394, 407, 528

Analyze:Scatter Plot (Y X), 232

analyzing by
groups, 357

analyzing by groups, 357

and group variables
formats, 366

Animate, 369

animating
graphs, 373
selections, 369

ANOVA,
see analysis of variance

Anscombe residuals
fit analyses, 698

Anscombe, F.J., 704

Apply button, 220, 267, 274
animation dialog, 371

Index

- fit analyses, 630
- approximate mean integrated square error
 - kernel estimation, 592
- arranging,
 - see moving
- arranging observations, 493
- arrow buttons
 - on scroll bar, 51
- arrow keys, 32
- arrow tool, 396
- Asc/Des** button, 494
- assigning
 - formats, 378
- assigning measurement levels
 - analysis of variance, 245
- assigning observation states, 474
- assumptions
 - in linear models, 615
- automatic
 - marker size, 163
- Axes**, 503, 549
- axes
 - adjusting, 129, 131, 361
 - aligning, 362, 417
 - default scale, 362
 - in bar chart, 503
 - in rotating plot, 131
- axis labels
 - in bar chart, 502
 - in box plots, 513
 - in contour plots, 539
 - in line plot, 523
 - in rotating plots, 549
 - in scatter plots, 529
- Azen, S.P., 19
- B**
- background, 435
- bandwidth
 - kernel estimation, 592, 667, 682
- bar chart, 499
 - distribution analyses, 584
 - graphs, 584
 - method, 501
 - orientation, 502
 - output, 502
 - variables, 500
- Bar Chart (Y)**, 500
- bar charts, 71, 72
 - bar heights in, 74
 - clicking in, 73, 78
 - features of, 74
 - labeling bars, 74
- bar heights in
 - bar charts, 74
- BAR statement, 783
- Base SAS Software, 446
- BASEBALL data set, 18, 306, 320, 378, 454
- batting averages, 329
- Becker, R.A., 96, 545
- Belsley, D.A., 233, 239, 657, 692, 704
- Bibby, J.M., 775
- BINOM option, 787
- Binomial**, 619
- binomial
 - deviance, 623
 - log-likelihood function, 620
 - response distribution, 619
- binomial deviance
 - generalized linear models, 623
- binomial distribution
 - generalized linear models, 619
- binomial log-likelihood function
 - generalized linear models, 620
- Binomial variable
 - fit analyses, 631
- bivariate plots
 - confidence ellipses, 768
 - scatter plot matrix, 753
- blending colors, 173, 174, 176
- blends
 - five colors, 176
 - two colors, 173
- Bonferroni method, 515
- Both**, 502, 513, 523, 529, 539, 549
- box,
 - see cube
- box plot
 - distribution analyses, 584
 - formatted values in, 387
 - graphs, 584
- Box Plot (Y)**, 387
- Box Plot (Y)**, 330
- Box Plot/Mosaic Plot (Y)**, 80, 88, 255, 364, 454, 509
- box plots, 71, 80, 507, 509
 - clicking in, 82
 - comparing distributions, 509
 - comparison circles, 84, 516
 - features of, 81, 83
 - mean diamonds, 83
 - method, 511
 - multiple comparison tests, 514
 - orientation, 513
 - output, 512
 - variables, 509
- BOX statement, 784
- brush, 96, 153
 - changing size of, 153
 - deleting, 99
 - dragging, 98
 - moving, 98
 - resizing, 97
 - throwing, 98
- brushing, 96, 98
 - in scatter plots, 531
 - with extended selection, 99
- brushing observations, 153

- BUSINESS data set, 19
- by groups, 357
 - comparing analyses, 360
- BY statement, 783
- by variable,
 - see group variable
- BY-group title, 437, 438
- C**
- C.I.,
 - see confidence interval
- C.I. for parameters,
 - see confidence interval for parameters
- calculation of
 - quantiles, 570
- calculations,
 - see transformations
- Campbell, P.F., 20
- Cancel** button, 67
- cancer
 - PATIENT data set, 21
- canonical
 - link function, 620
- Canonical**, 267, 285
- canonical correlation
 - multivariate analyses, 717
- Canonical Correlation Options** button, 724
- canonical discriminant
 - multivariate analyses, 718
- canonical link
 - quasi-likelihood, 624
- canonical link function
 - generalized linear models, 620
- canonical parameter
 - generalized linear models, 618
- canonical scores, 475
- canonical variables
 - components, 772
 - multivariate analyses, 772
- cascading
 - window positions, 460
- catalog
 - SAS/GRAPH, 431
- CDF,
 - see cumulative distribution function
- CDF:Confidence Band**, 595
- CDF:Fit Parametric:Normal:Sample Mean, Std Dev**, 597
- cell,
 - of matrix, see scatter plot matrix
 - of data window, see editing data values
- centroid, 692
- Chambers, J.M., 609
- Chandra, M., 601, 609
- changing,
 - see editing
- changing size of
 - brush, 153
- chart,
 - see graph
- Chi-squared statistic, 225, 251, 270, 287, 642, 651
- choosing
 - from menus, 10
 - order of variables, 81
- class,
 - see classification variable
 - see group
- CLASS statement, 783
- classification variables, 245
- Cleveland, W.S., 96, 545, 609, 704
- clicking, 8
- closing windows, 67
- clustered data, 531
- coefficient of variation
 - distribution analyses, 568, 569
- Collier Books, 18
- collinearity, 229, 657
 - fit analyses, 650
- Collinearity Diagnostics**, 228
- collinearity diagnostics
 - fit analyses, 615, 657
 - multiple regression, 228
- color
 - of curves, 208
- color blending
 - in contour plots, 540
- color blends, 173, 176
- color state, 491
- coloring curves, 208
- coloring observations, 169
- colors, 474
 - assigning by variable, 172, 175
 - background, 435
 - modifying blend, 173, 176
- columns
 - in data window, 489
- command line, 28
- Comp. Log-log**, 619
- comparing
 - analyses, 339
- comparing analyses, 339
 - by groups, 360
- comparison circles, 84, 512
- complement log-log
 - link function, 619
- complement log-log link function
 - generalized linear models, 619
- component plots, 755
 - principal components, 754
- components
 - canonical variables, 772
 - principal components, 771, 772
- condition index, 657
 - fit analyses, 657
- condition number
 - fit analyses, 657
- conditional distribution, 153, 371
- Confidence Band**, 595

Index

- confidence band
 - cumulative distribution, 595
- confidence curves
 - mean, 677
 - predicted, 677
- confidence ellipse
 - mean, 768
 - prediction, 672, 768
- Confidence Ellipses**, 672
- confidence ellipses, 295, 299
 - bivariate plots, 768
 - creating, 299
 - fit analyses, 671
 - interpreting, 300
 - mean, 673
 - multivariate analyses, 768
 - predicted, 673
- confidence interval
 - generalized linear models, 654
 - likelihood-based, 656
 - linear models, 654
 - Wald, 654
- confidence interval for mean
 - descriptive statistics, 571
- confidence interval for parameters
 - fit analyses, 652
- confidence interval for std dev
 - descriptive statistics, 571
- confidence interval for variance
 - distribution analyses, 571
- confidence intervals
 - distribution analyses, 571
- confidence levels, 514
- configuring
 - axes, 131
 - SAS/INSIGHT software, 453
- configuring SAS/INSIGHT software, 453
- Confirm dialog
 - exiting SAS/INSIGHT software, 67
- Conover, W.J., 609
- Constant**, 622, 631
 - fit analyses, 631
- constant
 - scale parameter estimates, 622
- constant for scale parameter
 - generalized linear models, 622
- CONSTANT option, 788
- continuous variable,
 - see interval variable
- Contour Plot (Z Y X)**, 118, 537
- contour plots, 118, 535
 - method, 538
 - output, 539
 - variables, 537
- CONTOUR statement, 785
- conventions
 - of user's guide, 15
- Cook's D, 233
 - fit analyses, 700
- Cook, R.D., 700, 704
- Copy Window**, 345, 351, 352, 360
- copying
 - data to disk, 422
 - windows, 345
- CORR,
 - see correlation
- corrected sums of squares and crossproducts, 728
- correlation, 295, 296, 729
 - and confidence ellipses, 672, 768
- correlation coefficient, 299
- correlation coefficients
 - in principal component analysis, 313
- Correlation Matrix, 299
- correlation matrix, 313, 773
 - descriptive statistics, 729
 - in principal component analysis, 309
 - multivariate analyses, 729
- correlation matrix of the parameter estimates
 - generalized linear models, 621
- correlations of the parameter estimates
 - fit analyses, 617
- COV,
 - see covariance
- covariance matrix, 773
 - descriptive statistics, 728
 - in principal component analysis, 309
- covariance matrix of the parameter estimates
 - fit analyses, 616
 - generalized linear models, 621
- Covratio, 233
 - fit analyses, 701
- Cox, D.R., 704
- creating
 - bar chart, 500
 - bar charts, 72
 - box plots, 80
 - confidence ellipses, 299
 - contour plots, 118
 - distribution analysis, 558
 - fit analysis, 629
 - formats, 385
 - line plots, 100
 - mosaic plots, 88
 - multivariate analysis, 708
 - rotating plots, 110
 - scatter plots, 91
 - surface plots, 116
 - variables, 319
- creating custom color blends, 176
- Cross** button, 249, 630
- cross validation, 628
 - fitting curves, 212
- crossed effects
 - fit analyses, 630
- CSS,
 - see corrected sums of squares
- distribution analyses, 568, 569
- CSSCP, 773,

- see corrected sums of squares and crossproducts
- Ctrl key, 10
- Cube**, 113, 549
- cube, 113
- cubic spline estimator, 679
- Cumulative Distribution**, 567
- cumulative distribution
 - confidence band, 595
 - empirical, 594
 - empirical distribution, 594
 - fit parametric, 597
 - Kolmogorov statistic, 594
 - parametric distribution, 597
 - test for a specific distribution, 599
 - test for distribution, 601
- cumulative distribution function, 556
 - exponential distribution, 557
 - lognormal distribution, 556
 - normal distribution, 556
 - Weibull distribution, 557
- currency format, 380
- cursor, 8
 - distance from, 460
 - shape of, 396, 397
- curve-fitting, 201
- Curves**, 207, 769
- curves, 201, 671
 - adding, 207
 - colors, 208
 - distribution analyses, 589
 - fitting, 207
 - nonparametric, 210
 - patterns, 208
 - width, 208
- Curves:Confidence Curves**, 677
- Curves:Kernel**, 668, 683
- Curves:Polynomial**, 207
- Curves:Prediction Confidence Ellipse**, 299
- Curves:Spline**, 665, 680
- CV,
 - see coefficient of variation
 - distribution analyses, 568, 569
- D**
- D,
 - see Cook's D
 - see Kolmogorov's D
 - see Kolmogorov statistic
- Data**, 422
- data
 - entering, 27
 - examining, 49
 - exploring, 71
 - extracting, 495
 - fast entry, 40
 - fill, 44
 - printing, 421, 425
 - saving, 421, 422
 - size of, 51
 - sorting, 56
 - subset of, 495
 - windows, 50
- data analysis, 5
- data exploration, 5
- DATA option, 781
- Data Set**, 488
- data set, 27, 487, 488,
 - see saving data
- data set dialog, 50, 488
- data values
 - editing, 494
- data window, 487
 - opening, 488
 - scrolling, 51
- Data:Fill**, 422
- Data:Move to Last**, 53
- Data:Sort**, 494
- DATA=, 488
- decimal format, 378
- default
 - variable role, 141
- default options,
 - see configuring SAS/INSIGHT software
- default role, 490
- default roles
 - group variables, 363
- default values,
 - see configuring SAS/INSIGHT software
- default variables
 - group, 363
- defaults
 - marker size, 165
- Define Variables**, 141, 363
- deflist,
 - see markers
- degree
 - of polynomial fit, 205
- degree of expansion, 630
- degrees of freedom, 212, 270, 323, 643, 644, 646, 647, 649, 651
- Delete**, 236, 288, 349, 351, 409, 410
- deleting
 - brush, 99
 - effects, 349, 351
 - graphs, 408, 410
 - tables, 408, 410
 - variables, 349
- density
 - parametric estimation, 590
- Density Estimation**, 565
- density estimation
 - kernel estimation, 592
- density function, 556, 619
 - exponential distribution, 557
 - lognormal distribution, 556
 - normal distribution, 556
 - Weibull distribution, 557

Index

- dependent variable,
 - see response variable
- Depth**, 114, 549
- depth cueing, 114
- descriptive statistics
 - confidence interval for mean, 571
 - confidence interval for std dev, 571
 - correlation matrix, 729
 - covariance matrix, 728
 - frequency table, 574
 - inverse correlation matrix, 731
 - location tests, 572
 - moments, 568
 - p-values of the correlations, 729
 - quantiles, 570
 - univariate statistics, 727
- deselecting, 99
- design matrix, 615, 661
- Deviance**, 622
- deviance, 225, 251, 270, 287, 642–644
 - binomial, 623
 - gamma, 623
 - generalized linear models, 622
 - inverse Gaussian, 623
 - normal, 623
 - Poisson, 623
- deviance residuals
 - fit analyses, 697
- Devlin, S.J., 704
- DF,
 - see degrees of freedom
- Dfbetas, 233
 - fit analyses, 701
- Dffits**, 233, 234
 - fit analyses, 700
- diagnostic statistics, 615
- differing means, 514
- Dillon, W.R., 775
- dimension
 - reducing, 306
- dimensionality
 - reducing, 713
- discrete variable,
 - see nominal variable
- DISCRIM procedure, 473
- discriminant analysis, 472
- disease,
 - see DRUG data set
- dispersion parameter, 619, 620
 - generalized linear models, 618, 622
 - quasi-likelihood, 625
- display, 8
 - options, 458
- Display Options**, 435
- Display options, 458
- DIST statement, 786
- distance from
 - cursor, 460
- distribution
 - of response variable, 618
- Distribution (Y)**, 381, 404, 555, 558
- Distribution (Y)**, 343, 358
- distribution analyses, 555
 - bar chart, 584
 - box plot, 584
 - coefficient of variation, 568, 569
 - confidence interval for variance, 571
 - confidence intervals, 571
 - CSS, 568, 569
 - curves, 589
 - CV, 568, 569
 - exponential distribution, 557
 - exponential quantile, 586
 - frequency table, 574
 - Gini's mean difference, 576
 - histogram, 584
 - interquartile range, 570
 - kernel estimation, 592
 - kurtosis, 559, 568, 569
 - location tests, 572
 - lognormal distribution, 556
 - lognormal quantile, 586
 - maximum, 570
 - median, 570
 - method, 559
 - minimum, 570
 - mode, 570
 - moments, 568
 - mosaic plot, 584
 - nominal variable, 605
 - normal distribution, 556
 - normal quantile, 586
 - output, 563
 - parametric density, 590
 - parametric distributions, 556
 - Q1, 570
 - Q3, 570
 - QQ plot, 585
 - QQ ref line, 603
 - quantile-quantile plot, 585
 - quantiles, 570
 - range, 570
 - skewness, 559, 568, 569
 - standard error of the mean, 568, 569
 - sum of squares corrected for the mean, 568, 569
 - tables, 568
 - test for a specific distribution, 599
 - test for distribution, 601
 - trimmed mean, 580
 - trimmed means, 580
 - trimmed t statistic, 580
 - uncorrected sum of squares, 568, 569
 - USS, 568, 569
 - variables, 558
 - Weibull distribution, 557
 - Weibull quantile, 587
 - Weight variable, 558
 - Winsorized mean, 580

- Winsorized means, 580
- Winsorized sum of squared deviations, 580
- Winsorized t statistic, 580
- distribution analysis
 - groups, 358
- distribution location tests
 - sign statistic, 573
 - signed rank statistic, 573
 - Student's t statistic, 573
- distributions
 - comparing in box plots, 509
- Dixon, W.J., 581, 609
- Dobson, A.J., 704
- DOLLAR format, 380
- double-click, 50
- double-clicking, 488
- draftsman's display,
 - see scatter plot matrix
- dragging, 9, 91
 - brush, 98
 - creating a brush, 96
- drilling
 - MINING data set, 20
- DRUG data set, 19, 245, 422
- Dunnett's test with control, 516

E

- E,
 - see exponential format
- Edit:Delete**, 236, 288, 351, 409
- Edit:Formats**, 378, 379, 382, 384
- Edit:Formats:Other**, 379, 386
- Edit:Observations:Exclude Calculations**, 347
- Edit:Observations:Find**, 59, 340
- Edit:Observations:Hide in Graphs**, 146
- Edit:Observations:Invert Selection**, 152
- Edit:Observations:Label in Plots**, 137
- Edit:Observations:Show in Graphs**, 148
- Edit:Observations:UnLabel in Plots**, 138
- Edit:Variables**, 320, 332
- Edit:Variables:log(Y)**, 282, 353
- Edit:Variables:log(Y)**, 321
- Edit:Variables:Other**, 325, 329
- Edit:Windows:Align**, 362, 417
- Edit:Windows:Animate**, 369
- Edit:Windows:Copy Window**, 345, 351, 352, 360
- Edit:Windows>Delete**, 410
- Edit:Windows:Display Options**, 435, 458
- Edit:Windows:Fonts**, 432
- Edit:Windows:Freeze**, 346
- Edit:Windows:Renew**, 139, 149, 354, 401, 402
- Edit:Windows:Tools**, 159, 171, 395
- editing
 - data values, 494
 - variables, 319
- editing formats,
 - see formats

- editing graphs,
 - see graphs
- editing marker sizes,
 - see markers
- editing observations
 - excluding, 347
 - hiding, 146
 - labeling, 137
 - showing in graphs, 148
- editing windows, 393
- effects
 - deleting, 236, 349
 - in model, 247
 - nominal, 251, 638
 - removing from model, 235, 272
 - specifying, 629
- Eigenvalue, 657
- Eigenvalues, 311
- eigenvalues, 713
- Eigenvectors, 312
- eigenvectors, 713
- ellipses
 - confidence, 299
- empirical
 - cumulative distribution, 594
- empirical distribution
 - cumulative distribution, 594
- End**, 67, 344
- entering
 - numeric data with keypad, 43
- Epanechnikov, V.A., 592, 609
- error term
 - in linear model, 614
- estimated CORR matrix
 - fit analyses, 658
- estimated COV matrix
 - fit analyses, 658
- Eubank, R.L., 704
- Exact Distribution**, 631, 633
 - fit analyses, 631, 633
- examining
 - data, 49
- examining the means
 - analysis of variance, 255
- Exclude in Calculations**, 347
- excluding observations, 347
- excluding observations from calculations, 344
- exiting SAS/INSIGHT software, 67
- Expand** button, 249, 284, 630
- expanded effects
 - fit analyses, 630
- explanatory variable, 614, 615, 629
- explanatory variables, 220
- exploration, 5
- exploring data, 71, 87, 110
- exponential
 - quantile, 586
 - test for distribution, 601
- exponential distribution

Index

- distribution analyses, 557
- fit parametric, 597
- parametric distributions, 557
- testing for, 601
- exponential family of distributions
 - fit analyses, 618
 - generalized linear models, 618
- exponential format, 379
- exponential quantile
 - distribution analyses, 586
- extended
 - selection, 9, 10
- extended selection, 10
 - and color blends, 175
 - brushing, 99
- Extract**, 342, 495
- extracted data windows
 - names of, 342
- extracting
 - observations, 340
- extracting data, 495
- F**
- F statistic
 - in analysis of variance, 252, 253
 - in multiple regression, 225, 226
- F test
 - in analysis of variance, 226, 253
- F-statistic
 - in analysis of variance, 643
 - in type I tests, 644
 - in type III tests, 646
- factorial expansion, 630
- features of
 - bar charts, 74
 - box plots, 81
 - SAS/INSIGHT software, 6
- Feller, W., 609
- FILE option, 781
- File:End**, 67, 344
- File:Open**, 67, 488
- File:Print**, 426
- File:Print:Print file**, 449
- File:Save:Data**, 422
- File:Save:Initial Tables**, 448
- File:Save:Options**, 45
- File:Save:Tables**, 446
- files
 - printing, 449
- Fill Areas**, 540
- Fill Values**, 494
- Find**, 59, 340
- Find Next**, 61, 493
- finding
 - observations, 151
- finding observations, 59, 340, 493
- First 2 Components Plot**, 755
- First 3 Components Plot**, 755
- Fisher
 - IRIS data set, 20
 - FISHER option, 788
 - Fisher's Scoring**, 633
 - fit analyses, 633
 - Fisher's scoring method
 - generalized linear models, 621
 - Fisher, R.A., 20, 472, 609, 775
 - fisheye lens, 531
 - Fit (Y X)**, 202, 263, 266, 279, 284, 613, 629
 - Fit (Y X)**, 220, 244, 327, 345, 349
 - fit analyses, 613
 - analysis of deviance for generalized linear models, 644
 - analysis of variance for linear models, 643
 - Anscombe residuals, 698
 - Apply button, 630
 - Binomial variable, 631
 - collinearity, 650
 - collinearity diagnostics, 615, 657
 - condition index, 657
 - condition number, 657
 - confidence ellipses, 671
 - confidence interval for parameters, 652
 - Constant, 631
 - Cook's D, 700
 - correlations of the parameter estimates, 617
 - covariance matrix of the parameter estimates, 616
 - Covratio, 701
 - crossed effects, 630
 - deviance residuals, 697
 - Dfbetas, 701
 - Dffits, 700
 - estimated CORR matrix, 658
 - estimated COV matrix, 658
 - Exact Distribution, 631, 633
 - expanded effects, 630
 - exponential family of distributions, 618
 - Fisher's Scoring, 633
 - fit curves, 671
 - Freq variable, 630
 - generalized linear models, 618
 - goodness of fit, 622
 - Group variables, 630
 - hat matrix, 616
 - hat matrix diagonal, 692
 - influence diagnostics, 691
 - kernel estimator, 667, 682
 - kernel function, 667, 682
 - Label variable, 630
 - leverage plots, 661
 - leverage variables, 699
 - likelihood function, 620
 - linear model, 614, 615
 - linear models, 615
 - link function, 619
 - maximum-likelihood estimation, 620
 - mean confidence curves, 677
 - mean square error, 616

- method, 631
- model equation, 251, 638
- model information, 250, 638
- multicollinearity, 650
- nested effects, 630
- nominal variable information, 251, 638
- nonparametric model, 614
- nonparametric smoothers, 626
- normal equation, 616
- normal kernel, 667, 682
- normal weight, 684
- Offset variable, 631
- output, 634
- parameter estimates for generalized linear models, 651
- parameter estimates for linear models, 649
- parameter information, 251, 638
- parametric confidence curves, 677
- parametric polynomial, 674
- parametric regression model, 614
- partial leverage plots, 661
- partial leverage variables, 699
- Pearson residuals, 697
- predicted curves, 695
- predicted mean vector, 616
- predicted surfaces, 694
- predicted values, 693
- prediction confidence curves, 677
- prediction confidence ellipses, 672, 768
- projection matrix, 616
- quadratic kernel, 667, 682
- quadratic weight, 684
- Quasi-Likelihood, 632
- quasi-likelihood functions, 623
- residual normal QQ Plot, 661
- residual normal quantiles, 693
- residual plots, 659
- residual-by-predicted plot, 659
- residuals, 693
- scale parameter, 622
- scatter plot smoother, 626
- smoother degrees of freedom, 627
- smoother generalized cross validation, 628
- smoothing spline, 663, 679
- standardized residuals, 696
- statistical models, 614
- studentized residuals, 696
- sum of squares for error, 616
- summary of fit for generalized linear models, 642
- summary of fit for linear models, 640
- tables, 638
- tolerance, 650
- tri-cube weight, 684
- triangular kernel, 667, 682
- triangular weight, 684
- type I tests, 644
- type III tests, 646
- variables, 629, 691
- variance, 616
- variance inflation factor, 650
- Weight variable, 630
- weighted analyses, 702
- X variable, 629
- X variable effects, 629
- X'X matrix, 639
- Y variable, 629
- fit curves
 - fit analyses, 671
 - kernel, 682
 - nonparametric local polynomial smoother, 684
 - nonparametric smoothers, 626
 - parameter estimates, 674
 - parametric confidence curves, 677
 - parametric confidence ellipses, 671
 - parametric polynomial, 674
 - polynomial equation, 674
 - smoother degrees of freedom, 627
 - smoother generalized cross validation, 628
 - smoothing spline, 663, 679
- fit parametric
 - cumulative distribution, 597
- Fit Parametric:Normal:Sample Mean, Std Dev**, 597
- FIT statement, 787
- fitting curves, 201
 - cross validation, 212
 - generalized cross validation, 212
 - loess smoother, 213
 - loess smoother fit, 213
 - normal kernel fit, 211
 - parametric regression, 202
 - polynomial, 202
- fitting techniques, 671
- five-color blends, 176
- flipping graphs, 412–414
- focus,
 - see zooming
- fonts
 - choosing, 432
- footnotes, 437, 438, 441
- FORMAT procedure, 377, 385
- Formats**, 378, 379, 382, 384
- formats
 - and group variables, 366
 - assigning, 378
 - creating, 385
 - currency, 380
 - decimal, 378
 - exponential, 379
 - in analysis tables, 383
 - in data window, 382
 - in groups, 387
 - of axes, 381, 382
 - of values, 383
 - scientific, 379
 - sorting by, 494
 - use in calculations, 387

Index

Formats:Other, 379, 386

formatting, 377

group variables, 366

formula,

see transformation

Freedman, D., 239

Freeze, 346

freezing windows, 346

Freq,

see frequency

FREQ option, 784, 786, 787

Freq variable

fit analyses, 630

multivariate analyses, 708

frequency role, 490

frequency table

descriptive statistics, 574

distribution analyses, 574

frequency values

in bar charts, 501

in box plots, 511

in distribution analyses, 559

in fit analyses, 631

in multivariate analyses, 710

frequency variable

in box plot, 500

in box plots, 509

in distribution analyses, 558

G

Gamma, 619

gamma

deviance, 623

log-likelihood function, 620

response distribution, 619

gamma deviance

generalized linear models, 623

gamma distribution

generalized linear models, 619

gamma log-likelihood function

generalized linear models, 620

GCV,

see generalized cross validation

general linear model, 244,

see linear model

generalized cross validation

fitting curves, 212

generalized linear model, 613, 614

components of, 265, 266, 281, 282

logistic regression, 263

Poisson regression, 279

specifying, 266, 284

generalized linear models, 618

binomial deviance, 623

binomial distribution, 619

binomial log-likelihood function, 620

canonical link function, 620

canonical parameter, 618

complement log-log link function, 619

confidence interval, 654

constant for scale parameter, 622

correlation matrix of the parameter estimates,
621

covariance matrix of the parameter estimates,
621

deviance, 622

dispersion parameter, 618, 622

exponential family of distributions, 618

Fisher's scoring method, 621

fit analyses, 618

gamma deviance, 623

gamma distribution, 619

gamma log-likelihood function, 620

goodness of fit, 622

gradient vector, 621

Hessian matrix, 621

identity link function, 619

inverse Gaussian deviance, 623

inverse Gaussian distributions, 619

inverse Gaussian log-likelihood function, 620

likelihood function, 620

linear predictor, 618

link function, 618, 619

log link function, 619

logit link function, 619

maximum quasi-likelihood estimation, 625

maximum-likelihood estimate for scale parameter,
622

maximum-likelihood estimation, 620

mean deviance, 623

mean deviance for scale parameter, 622

mean Pearson chi-squared, 623

mean Pearson chi-squared for scale parameter,
622

natural parameter, 618

normal deviance, 623

normal distribution, 619

normal log-likelihood function, 620

offset, 618

overdispersion, 623

Pearson chi-squared, 623

Poisson deviance, 623

Poisson distribution, 619

Poisson log-likelihood function, 620

power link function, 619

probit link function, 619

quasi-likelihood functions, 623

response distribution, 619

scale parameter, 622

scale parameter estimates, 622

scaled deviance, 622

scaled Pearson chi-squared, 623

variance function, 618

generalized residuals, 692

Gini's mean difference

distribution analyses, 576

robust estimation, 576

Goldstein, M., 775

goodness of fit, 622, 663, 679
 fit analyses, 622
 generalized linear models, 622
 GPA data set, 91, 110, 146, 158, 220, 296, 394
 grade point average, 19
 grabber,
 see hand tool
 grade point average
 GPA data set, 19
 gradient vector
 generalized linear models, 621
 graph
 options, 458
 Graph options, 458
 graphics
 printing, 431
 saving, 431
Graphs, 249, 267, 584, 753
 graphs
 adding, 404, 406, 407
 aligning, 417
 bar chart, 584
 box plot, 584
 deleting, 408, 410
 flipping, 412–414
 growing, 411
 histogram, 584
 margin between, 460
 mosaic plot, 584
 moving, 411
 multivariate analyses, 753
 orienting, 412–414
 QQ plot, 585
 shrinking, 411
 size of, 460
 sizing, 411
Graphs:First 2 Components Plot, 755
Graphs:First 3 Components Plot, 755
Graphs:Partial Leverage, 229
Graphs:QQ Plot, 585, 587
 Grosse, E., 704
 group
 default variables, 363
 group role, 490
 group variable
 in box plots, 509
 in line plot, 522
 in rotating plots, 547
 Group variables
 fit analyses, 630
 multivariate analyses, 708
 group variables, 357, 358
 default roles, 363
 formatting, 366
 in contour plots, 537
 in distribution analyses, 558
 in rotating plots, 528
 order of, 364
 groups

analyzing by, 357
 order of, 490

H

hand
 adjusting axes, 129
 hand tool, 54, 493
 Hastie, Y.J., 212, 213, 704
Hat Diag, 230
 hat diagonal, 230
 hat matrix
 fit analyses, 616
 hat matrix diagonal
 fit analyses, 692
 heights
 of bars, 502
Help, 15
 help
 context-sensitive, 16
 Help key, 16
 help system, 15, 18
 index, 17
 SAS/INSIGHT software, 15, 18
Help:Index, 17
Help:Introduction, 17
Help:Reference, 17
Help:Techniques, 17
 Hessian matrix
 generalized linear models, 621
Hide in Graphs, 146
 hiding observations, 145, 146
 Hinkley, D.V., 704
 histogram, 499
 distribution analyses, 584
 graphs, 584
Histogram/Bar Chart (Y), 72, 76
 Hoaglin, D.C., 704
 holding the mouse button, 400
 horizontal,
 see orientation
Horizontal Axis at Bottom, 502, 513, 523, 529, 539
 host, 8
 available colors, 173
 host resources, 453, 467
 Hotelling's T-squared statistic, 671, 768
 Hotelling, H., 775
 Hsu's test for best, 516
 Hsu's test for worst, 516
 Hsu, J. C., 514, 516
 HTML, 450
 hypothesis testing, 225, 226, 252, 253, 270, 555, 572, 589, 594, 595, 615, 643–647, 650, 656

I

identifying observations, 92, 135, 313
 in box plots, 82
Identity, 619
 identity

Index

- link function, 619
- identity link function
 - generalized linear models, 619
- ill conditioned, 657
- Iman, R.L., 573, 609
- in analysis of variance
 - mean, 255
- in distribution analyses
 - frequency variable, 558
 - group variables, 558
 - label variable, 558
- in multiple regression
 - parameter estimates, 226
- in Multivariate analysis
 - scatter plot matrix, 299
- in principal component analysis
 - correlation coefficients, 313
- Include/Exclude state, 491
- include/exclude state, 474
- independent variable,
 - see explanatory variable
- Index**, 17
- index
 - help system, 17
 - SAS/INSIGHT User's Guide, 15
- INFILE option, 781
- influence diagnostics
 - fit analyses, 691
- influential observations, 234, 692, 700, 701
- Initial Tables**, 448
- initial values,
 - see default values
- initializing,
 - see configuring SAS/INSIGHT software
- input data set,
 - see DATA= option
- INSIGHT, 23
- interaction effect, 247
- interaction effects
 - specifying, 284
- Interactive Data Analysis, 28
- Intercept**, 630
- intercept
 - QQ ref line, 603
- interpreting
 - confidence ellipses, 300
- interquartile range
 - distribution analyses, 570
- interval variable, 490
- interval variables, 51, 72
 - in analysis of variance, 245
- Introduction**, 17
- inverse correlation matrix
 - descriptive statistics, 731
 - multivariate analyses, 731
- Inverse Gaussian**, 619
- inverse Gaussian
 - deviance, 623
 - log-likelihood function, 620

- response distribution, 619
- inverse Gaussian deviance
 - generalized linear models, 623
- inverse Gaussian distributions
 - generalized linear models, 619
- inverse Gaussian log-likelihood function
 - generalized linear models, 620
- Invert Selection**, 152
- invisible observations, 148
- invoking
 - SAS/INSIGHT software, 50
- IRIS data set, 472
 - Fisher, 20

J

- Jobson, J.D., 775
- Johnson, N.L., 609
- joint distribution, 153
- journaling SAS/INSIGHT session,
 - see saving tables

K

- Kaiser, H.F., 714, 775
- Kent, J.T., 775
- kernel
 - fit curves, 682
 - normal, 211
- kernel estimation
 - approximate mean integrated square error, 592
 - bandwidth, 592, 667, 682
 - density estimation, 592
 - distribution analyses, 592
 - mean integrated square error, 592
 - normal, 592
 - normal distribution, 592
 - quadratic, 592
 - quadratic distribution, 592
 - triangular, 592
 - triangular distribution, 592
- kernel estimator
 - fit analyses, 667, 682
 - in fit analyses, 614
- kernel function
 - choice of, 592
 - fit analyses, 667, 682
 - normal, 667, 682
 - quadratic, 667, 682
 - triangular, 667, 682
- Kleiner, B., 609
- Kolmogorov statistic, 599
 - cumulative distribution, 594
- Kotz, S., 609
- Krzanowski, W.J., 775
- Kuh, E., 233, 239, 657, 692, 704
- kurtosis, 568, 569
 - distribution analyses, 559, 568, 569
- Kutner, M.H., 247
- Kvalseth, T.O., 704

L

label

- in data window, 138
- observations, 135
- permanent, 137
- removing, 138
- temporary, 136, 137

Label button, 139**Label in Plots**, 137

LABEL option, 784–787

label role, 490

Label variable

- fit analyses, 630
- multivariate analyses, 708

label variable, 139, 141

- in box plot, 500
- in box plots, 509
- in contour plots, 537
- in distribution analyses, 558
- in line plot, 522
- in rotating plots, 547

label variables

- in box plots, 81
- in rotating plots, 528

Label/UnLabel state, 491

label/unlabel state, 474

labeling observations, 92

Labels, 502, 513, 523, 529, 539, 549

labels

- bar chart axes, 502
- box plot axes, 513
- contour plot axes, 539
- line plot axes, 523
- of transformed variables, 329
- rotating plot axes, 549
- scatter plot axes, 529

lack of fit,

- see goodness of fit

layout

- scatter plot matrix, 95

learning

- SAS/INSIGHT software, 15

least-squares estimates, 615

Lee, E.T., 21

Lehmann, E.L., 573, 609

level,

- see measurement level
- see classification variable
- see group

level sets, 118

leverage plots

- fit analyses, 661

leverage variables

- fit analyses, 699

LIBNAME statement, 488

Library, 488

library, 50

likelihood function

- fit analyses, 620

generalized linear models, 620

likelihood ratio, 656

likelihood ratio test, 271

likelihood-based

- confidence interval, 656

likelihood-ratio statistic

- type III tests, 647

line fit, 204

line plot, 521

- method, 522
- output, 523
- variables, 522

Line Plot (Y X), 100, 522

line plots, 87

LINE statement, 784

linear model, 614

- fit analyses, 614, 615

linear models

- confidence interval, 654
- fit analyses, 615

linear predictor

- generalized linear models, 618

linear regression, 204

Link Function, 631

link function, 618, 631, 638

- canonical, 620
- complement log-log, 619
- fit analyses, 619
- generalized linear models, 618, 619
- identity, 619
- log, 619
- logit, 619
- power, 619
- probit, 619

LINK option, 787

linking of windows, 92, 346

local polynomial fit

- weight function, 684

locating observations,

- see finding observations

Location Tests, 574

location tests, 574

- descriptive statistics, 572
- distribution analyses, 572

loess fit

- weight function, 684

loess smoother

- fitting curves, 213

loess smoother fit

- fitting curves, 213

Log, 619

log, 353

- link function, 619

log link function

- generalized linear models, 619

log transformation, 282, 320

log(Y), 282, 353**log(Y)**, 321

log-likelihood function

Index

- binomial, 620
- gamma, 620
- inverse Gaussian, 620
- normal, 620
- Poisson, 620
- logistic regression, 263
 - analysis of deviance, 270
 - model equation, 269
 - modifying the model, 271
 - parameter estimates, 270
 - residuals-by-predicted plot, 270
 - summary of fit, 270
 - type III (LR) tests, 271
 - type III (Wald) tests, 270
- Logit**, 619
- logit
 - link function, 619
- logit link function
 - generalized linear models, 619
- lognormal
 - quantile, 586
 - test for distribution, 601
- lognormal distribution
 - distribution analyses, 556
 - fit parametric, 597
 - parametric distributions, 556
 - testing for, 601
- lognormal quantile
 - distribution analyses, 586
- LR,
 - see likelihood ratio
- M**
- magnifying glass tool, 394–396
- main effect, 247
- major ticks, 126
- manager, 471
- Mardia, K.V., 775
- margin between
 - graphs, 460
- marginal histograms, 404, 416
- marker, 51
- Marker Sizes**, 162, 163
- marker sizes
 - in bar chart, 503
- marker state, 491
- markers, 92, 157, 159, 474
 - assigning by variable, 160
 - size of, 162
- MARKERSIZE option, 784–786
- matrix
 - of rotating plots, 115
- matrix, correlation,
 - see correlation matrix
- matrix, covariance,
 - see covariance matrix
- matrix, design,
 - see design matrix
- matrix, hat,
 - see hat matrix
- matrix, Hessian,
 - see Hessian matrix
- matrix, patter,
 - see pattern matrix
- matrix, $X'X$,
 - see $X'X$ matrix
- maximum, 727
 - distribution analyses, 570
- maximum quasi-likelihood estimation
 - generalized linear models, 625
- maximum redundancy
 - multivariate analyses, 718
- maximum-likelihood estimate
 - scale parameter estimates, 622
- maximum-likelihood estimate for scale parameter
 - generalized linear models, 622
- maximum-likelihood estimation
 - fit analyses, 620
 - generalized linear models, 620
- McCabe, G.P., 20, 225, 239, 252, 296, 301
- McCullagh, P., 21, 275, 281, 618, 623, 704
- McLaughlin, D.H., 580, 581
- McLaughlin, D.H., 609
- Mean**, 256, 512
- mean, 225, 251, 270, 287, 383, 509, 568, 569, 597, 620, 640, 727, 773
 - box plot, 256
 - confidence curves, 677
 - confidence ellipse, 768
 - in analysis of variance, 255
- mean confidence curves
 - fit analyses, 677
- mean confidence ellipse
 - multivariate analyses, 768
- mean confidence ellipses, 673
- mean deviance
 - generalized linear models, 623
 - scale parameter estimates, 622
- mean deviance for scale parameter
 - generalized linear models, 622
- mean diamonds, 83, 258, 259
- mean integrated square error
 - kernel estimation, 592
- mean line fit, 205
- mean Pearson chi-squared
 - generalized linear models, 623
 - scale parameter estimates, 622
- mean Pearson chi-squared for scale parameter
 - generalized linear models, 622
- mean square error, 212
 - fit analyses, 616
- means, 83
- measurement level, 51, 245, 490
 - assigning, 246
 - variables, 490
- median, 81, 507, 511
 - distribution analyses, 570
- memory

- storing data set in, 422
- memory, optimizing, 165
- menu, 8, 10
 - pulldown, 10
- menu bar, 10
- method
 - bar chart, 501
 - box plots, 511
 - contour plots, 538
 - distribution analyses, 559
 - fit analyses, 631
 - line plot, 522
 - multivariate analyses, 710
 - options, 454, 456
 - rotating plots, 548
 - scatter plots, 528
- Method** button, 284, 511, 562, 631, 711, 715
- Method dialog
 - Fit window, 266, 284
- method options, 454, 456
- minimum, 727
 - distribution analyses, 570
- MINING data set, 340
 - drilling, 20
- MININGX data set, 20, 202
- minor ticks, 126
- MISE**, 593,
 - see mean integrated square error
- missing values, 58, 325
 - in bar charts, 501
 - in box plots, 511
 - in contour plots, 538
 - in distribution analyses, 559, 562
 - in fit analyses, 631
 - in line plots, 522
 - in multivariate analyses, 710
 - in rotating plots, 548
 - in scatter plots, 528
- MLE**, 622,
 - see maximum-likelihood estimate
- MLE, Theta=0**, 597
- mode
 - distribution analyses, 570
 - parametric density, 591
- model
 - modifying, 630
 - removing effects, 235, 272
 - specifying effects, 629
- model effects, 247
- model equation
 - fit analyses, 251, 638
 - logistic regression, 269
 - multiple regression, 225, 313
- model information
 - fit analyses, 250, 638
 - in Fit window, 267, 287
- modifying,
 - see editing
- modifying the model
 - logistic regression, 271
 - multiple regression, 235
 - poisson regression, 288
- moments, 570
 - descriptive statistics, 568
 - distribution analyses, 568
- monochrome images, 435
- Moore, D.S., 225, 239, 252, 296, 301
- Morrison, D.F., 775
- mosaic plot
 - distribution analyses, 584
 - graphs, 584
- mosaic plots, 87, 509
- Motif window manager
 - setting X resources, 467
- mouse, 8
- mouse button, 8
- Move to First**, 493
- Move to Last**, 53, 493
- moving
 - columns, 56
 - graphs, 411
 - tables, 411
- moving observations, 493
- MSE,
 - see mean square error
- Muenchen, R.A., 552
- MULT statement, 786
- multicollinearity
 - fit analyses, 650
- multiple
 - selection, 9
- multiple color blends, 174
- multiple comparison circles, 516
- Multiple Comparison of Means**, 512
- multiple comparison of means, 84
- Multiple Comparison Options**, 513, 514
- Multiple Comparison Test**, 514
- multiple comparison tests
 - Dunnett's test with control, 516
 - Hsu's test for best, 516
 - Hsu's test for worst, 516
 - pairwise Bonferroni, 515
 - pairwise t-test, 515
 - Tukey-Kramer all pairs, 515
- multiple regression, 219
 - adding tables and graphs, 228
 - analysis of variance, 225
 - collinearity diagnostics, 228
 - model equation, 225, 313
 - modifying the model, 235
 - parameter estimates, 226
 - partial leverage plots, 229
 - residual-by-hat diagonal plot, 230
 - residual-by-predicted plot, 227
 - saving the residuals, 238
 - summary of fit, 225
 - type III tests, 226
- Multivariate (Y's)**, 296, 707, 708

Index

Multivariate (Y's), 306

- multivariate analyses, 707
 - canonical correlation, 717
 - canonical discriminant, 718
 - canonical variables, 772
 - confidence ellipses, 768
 - corrected sums of squares and crossproducts, 728
 - correlation matrix, 729
 - Freq variable, 708
 - graphs, 753
 - Group variables, 708
 - inverse correlation matrix, 731
 - Label variable, 708
 - maximum redundancy, 718
 - mean confidence ellipse, 768
 - method, 710
 - output, 720
 - p-values of the correlations, 729
 - prediction confidence ellipse, 768
 - principal component plots, 754
 - principal components, 713, 771, 772
 - principal components rotation, 715
 - scatter plot matrix, 753
 - sums of squares and crossproducts, 727
 - tables, 727
 - univariate statistics, 727
 - variables, 708
 - variance divisor, 712
 - Weight variable, 709
 - weighted analyses, 773

Multivariate analysis, 305

Myers, R.H., 238, 239

N

N, 727,

- see number of observations

name mangling, 438

Names, 502, 513, 523, 529, 539, 549

names

- of data windows, 342

- of transformed variables, 329

names of

- extracted data windows, 342

names of tables, 450

naming

- catalog entries, 437, 438

- Cook's D variables, 700

- Covratio variables, 701

- data sets, 423

- dfbetas variables, 701

- Dffits variables, 700

- partial leverage variables, 699

- residual variables, 694, 696–699

- variables, 692

natural parameter

- generalized linear models, 618

navigating, 34

Nelder, J.A., 21, 275, 281, 618, 623, 704

Nest button, 630

nested effects

- fit analyses, 630

New Observations, 494

New Variables, 494

NOBUTTON option, 782

NOCONFIRM option, 782

NOEXACT option, 788

NOINT option, 787

NOMENU option, 782

nominal variable, 490

- distribution analyses, 605

nominal variable information

- fit analyses, 251, 638

nominal variables, 51, 75

- in analysis of variance, 245

noncontiguous

- selection, 10

noncontiguous selection, 10

nonparametric curves, 210

Nonparametric Curves button, 635

nonparametric local polynomial smoother

- fit curves, 684

nonparametric model

- fit analyses, 614

nonparametric regression, 211

nonparametric smoothers

- fit analyses, 626

- fit curves, 626

Normal, 619

normal

- deviance, 623

- kernel estimation, 592

- kernel function, 667, 682

- log-likelihood function, 620

- quantile, 586

- response distribution, 619

- test for distribution, 601

- weight function, 684

normal deviance

- generalized linear models, 623

normal distribution, 300

- distribution analyses, 556

- fit parametric, 597

- generalized linear models, 619

- kernel estimation, 592

- parametric distributions, 556

- testing for, 601

normal equation

- fit analyses, 616

normal kernel

- fit analyses, 667, 682

normal kernel fit

- fitting curves, 211

normal log-likelihood function

- generalized linear models, 620

normal quantile

- distribution analyses, 586

normal quantile-quantile plot

- fit analyses, 661
- normal quantiles, 694
- normal weight
 - fit analyses, 684
- NOSCROLL option, 782
- null hypothesis,
 - see hypothesis testing
- number of observations
 - as label, 137
 - as observation label, 92
 - in data window, 491
 - in Moments table, 568, 569
- number of variables
 - in data window, 489

O

- objects, 446
 - output, 450
- observation, 27
- observation number
 - as label, 137
 - as observation label, 92
- observation state, 474
- observation states, 474, 491
 - saving, 422, 492
- Observations**, 257, 503
- observations, 491,
 - number of, see number of observations
 - adding, 35
 - brushing, 96, 153
 - coloring, 169
 - deselecting, 147
 - excluding, 344
 - extracting, 340
 - finding, 59, 151, 340, 493
 - hiding, 145, 146
 - identifying, 92
 - in bar chart, 503
 - invisible, 151
 - labeling, 92
 - markers, 51, 92
 - marking, 157
 - new, 494
 - querying for, 59
 - selecting, 92
 - slicing, 145, 153
 - sorting, 56, 494
 - states, 491
 - tooggling display of, 145, 149, 257

Observations:Exclude in Calculations,
347

Observations:Hide in Graphs, 146

Observations:Invert Selection, 152

Observations:Label in Plots, 137

Observations:Show in Graphs, 148

Observations:UnLabel in Plots, 138

ODS,

- see Output Delivery System

of data windows

- names, 342
- offset
 - generalized linear models, 618
 - of bars, 501, 503
- OFFSET option, 787
- Offset variable
 - fit analyses, 631
- OK** button, 220
- Open**, 67, 488
- OPEN statement, 782
- opening
 - data set, 488
 - data window, 488
- operation of
 - SAS/INSIGHT software, 8
- optimizing memory, 165
- optional variables, 81
- Options**, 458
- options, 454, 456
 - BINOM, 787
 - box plot, 256
 - CONSTANT, 788
 - DATA, 781
 - display, 435, 458
 - distribution, 405
 - FILE, 781
 - FISHER, 788
 - FREQ, 784, 786, 787
 - graph, 458
 - grey scale graphics, 438
 - in fit analysis, 203
 - INFILE, 781
 - LABEL, 784–787
 - LINK, 787
 - MARKERSIZE, 784–786
 - method, 454, 456
 - NOBUTTON, 782
 - NOCONFIRM, 782
 - NOEXACT, 788
 - NOINT, 787
 - NOMENU, 782
 - NOSCROLL, 782
 - OFFSET, 787
 - OTHER, 783, 784
 - output, 454, 456
 - POWER, 787
 - QUASI, 788
 - RESP, 787
 - SAS/INSIGHT, 454
 - saving, 453, 466
 - SCALE, 788
 - setting default, 45
 - TOOLS, 781
 - used in this book, 435
 - WEIGHT, 786, 787
 - window, 458, 459
 - XAXIS, 783–786
 - YAXIS, 783–786
 - ZAXIS, 785, 786

Index

- order of expansion,
 - see degree of expansion
- order of observations,
 - see moving, sorting
- order of polynomial,
 - see degree of polynomial
- order of variables,
 - see moving
- orientation
 - of bar chart, 502
 - of box plots, 513
 - of contour plots, 539
 - of line plot, 523
 - of scatter plots, 529
- Orientation:Horizontal Axis at Bottom,**
 - 502, 513, 523, 529, 539
- Orientation:Vertical Axis at Left,** 502, 513,
 - 523, 529, 539
- Orientation:Y Axis Vertical,** 502, 513, 523,
 - 529, 539
- orienting graphs, 412–414
- OTHER option, 783, 784
- outlier, 580, 659, 754
- outliers, 95, 316
- output
 - bar chart, 502
 - box plots, 512
 - contour plots, 539
 - distribution analyses, 563
 - fit analyses, 634
 - line plot, 523
 - multivariate analyses, 720
 - objects, 450
 - options, 454, 456
 - rotating plots, 549
 - scatter plots, 529
- Output** button, 225, 502, 512, 523, 529, 549, 563,
 - 634, 720
- Output Components,** 771
- Output Delivery System, 446, 449
- Output Principal Components:2,** 316
- Output Variables** button, 635
- output window, 425
- overdispersion, 289
 - generalized linear models, 623
 - Poisson regression, 285
- P**
- p-value
 - for F statistic, 225, 252
- p-values
 - for likelihood ratio type III tests, 271
- p-values of the correlations
 - descriptive statistics, 729
 - multivariate analyses, 729
- pairwise Bonferroni, 515
- pairwise t-test, 515
- parameter estimates, 649
 - analysis of variance, 253
 - fit curves, 674
 - in multiple regression, 226
 - logistic regression, 270
 - multiple regression, 226
 - Poisson regression, 290
- parameter estimates for generalized linear models
 - fit analyses, 651
- parameter estimates for linear models
 - fit analyses, 649
- parameter information
 - fit analyses, 251, 638
- parametric
 - regression, 202
- parametric confidence curves
 - fit analyses, 677
 - fit curves, 677
- parametric confidence ellipses
 - fit curves, 671
- Parametric Curves** button, 635
- parametric density
 - distribution analyses, 590
 - mode, 591
- parametric distribution
 - cumulative distribution, 597
- parametric distributions
 - distribution analyses, 556
 - exponential distribution, 557
 - lognormal distribution, 556
 - normal distribution, 556
 - Weibull distribution, 557
- parametric estimation
 - density, 590
- parametric polynomial
 - fit analyses, 674
 - fit curves, 674
- parametric regression, 202, 204
 - fitting curves, 202
- parametric regression model
 - fit analyses, 614
- Partial Leverage,** 229
- partial leverage plots
 - fit analyses, 661
 - multiple regression, 229
 - residual plots, 661
- partial leverage variables
 - fit analyses, 699
- paste buffer,
 - see clipboard
- PATIENT data set, 265
 - cancer, 21
- pattern
 - of curves, 208
- pause animation, 371
- PC,
 - see principal component
- PCA,
 - see principal component analysis
- Pearson,** 622
- Pearson chi-squared

- generalized linear models, 623
- Pearson chi-squared statistic,
 - see chi-squared statistic
- Pearson product-moment correlations, 729
- Pearson residuals
 - fit analyses, 697
- Pearson, K., 775
- Penner, R., 20
- percentile, 507
- permanent
 - label, 137
- perspective,
 - see depth cueing
- Pisani, R., 239
- plane, rotating, 552
- players, 18,
 - see BASEBALL data set
- plot,
 - see graph
 - quantile-quantile, 585
- plotting symbols,
 - see markers
- pointer,
 - see cursor
- pointing, 8
- Poisson**, 619
 - deviance, 623
 - log-likelihood function, 620
 - response distribution, 619
- Poisson deviance
 - generalized linear models, 623
- Poisson distribution
 - generalized linear models, 619
- Poisson log-likelihood function
 - generalized linear models, 620
- Poisson regression, 279
 - analysis of deviance, 287
 - overdispersion, 285
 - parameter estimates, 290
 - summary of fit, 287
 - type III (Wald) tests, 287
- poisson regression
 - modifying the model, 288
- pollutants, 18
- Polynomial**, 207
- polynomial
 - fitting curves, 202
- polynomial curves, 202
- polynomial equation
 - fit curves, 674
- polynomial expansion, 630
- polynomial fit, 202
- position of
 - windows, 460
- position of windows, 460
- Power**, 619, 631
- power
 - fit analyses, 631
 - link function, 619
- power link function
 - generalized linear models, 619
- POWER option, 787
- precision
 - of formatted values, 380
- predicted
 - confidence curves, 677
- predicted confidence ellipses, 673
- predicted curves
 - fit analyses, 695
- predicted mean vector
 - fit analyses, 616
- predicted surfaces
 - fit analyses, 694
- predicted values
 - fit analyses, 693
- prediction
 - confidence ellipse, 672, 768
- prediction confidence curves
 - fit analyses, 677
- Prediction Confidence Ellipse**, 299
- prediction confidence ellipse
 - multivariate analyses, 768
- prediction confidence ellipses
 - fit analyses, 672, 768
- pressing the mouse button, 400
- principal component analysis, 305
- principal component options, 720
- Principal Component Options** button, 722
- principal component plots
 - multivariate analyses, 754
- principal components, 306
 - component plots, 754
 - components, 771, 772
 - multivariate analyses, 713, 771, 772
 - saving, 316
- principal components rotation
 - multivariate analyses, 715
- Principal Components:Output Components**, 771
- Pringle, R.M., 704, 775
- Print**, 426
- Print file**, 449
- PRINT procedure, 421, 425
- Print:Print file**, 449
- printing, 439
 - all contents of window, 436, 440
 - color images, 435
 - data, 421
 - files, 426, 449
 - from clipboard, 439
 - from window, 440
 - graphics, 431
 - selected portion of window, 436, 440
 - tables, 445
- PRINTTO procedure, 428, 449
- Probit**, 619
- probit
 - link function, 619

Index

- probit link function
 - generalized linear models, 619
 - PROC DISCRIM, 473,
 - see DISCRIM procedure
 - PROC FORMAT, 377, 385,
 - see FORMAT procedure
 - proc insight, 28
 - PROC INSIGHT statement, 781
 - PROC OUTPUT,
 - see OUTPUT procedure
 - PROC PRINT, 421, 425,
 - see PRINT procedure
 - PROC PRINTTO, 428, 449,
 - see PRINTTO procedure
 - PROFILE catalog, 466
 - program editor, 385, 425, 471, 473
 - invoking SAS/INSIGHT software from, 28
 - projection matrix
 - fit analyses, 616
 - properties,
 - see variable properties
 - pulldown
 - menu, 10
 - pulldown menu, 10
 - purpose of
 - SAS/INSIGHT Software, 5
 - Purves, R., 239
 - Pythagorean theorem, 516
- Q**
- Q1
 - distribution analyses, 570
 - Q3
 - distribution analyses, 570
 - QQ Plot,
 - see quantile-quantile plot
 - QQ plot
 - distribution analyses, 585
 - graphs, 585
 - QQ ref line, 603
 - distribution analyses, 603
 - intercept, 603
 - slope, 603
 - quadratic
 - kernel estimation, 592
 - kernel function, 667, 682
 - weight function, 684
 - quadratic distribution
 - kernel estimation, 592
 - quadratic kernel
 - fit analyses, 667, 682
 - quadratic polynomial fit, 205
 - quadratic weight
 - fit analyses, 684
 - qualitative variable,
 - see nominal variable
 - quantile
 - exponential, 586
 - lognormal, 586
 - normal, 586
 - Weibull, 587
 - quantile-quantile plot, 694
 - distribution analyses, 585
 - fit analyses, 661
 - quantiles
 - calculation of, 570
 - descriptive statistics, 570
 - distribution analyses, 570
 - quantitative variable,
 - see interval variable
 - quartiles, 81, 507, 511
 - QUASI option, 788
 - Quasi-Likelihood**, 632
 - fit analyses, 632
 - quasi-likelihood, 285, 623, 624
 - canonical link, 624
 - dispersion parameter, 625
 - scale parameter, 624
 - variance function, 624
 - quasi-likelihood functions
 - fit analyses, 623
 - generalized linear models, 623
 - querying, 493
 - querying for observations, 59
 - QUIT statement, 789
- R**
- R-Square, 640, 642
 - R-square, 204, 225, 251
 - range
 - distribution analyses, 570
 - of data displayed, 146
 - Rawlings, J.O., 230, 239, 704
 - Raynor, A.A., 704, 775
 - Rays**, 549
 - recording SAS/INSIGHT session,
 - see saving tables
 - recording statements, 481
 - recreating,
 - see Renew
 - Reference**, 17
 - reference, 15
 - Reference Lines**, 503
 - reference lines
 - in bar chart, 503
 - regression, 244, 613–615
 - linear, 204
 - multiple, 219
 - nonparametric, 211
 - parametric, 202, 204
 - simple, 204
 - Reid, N., 704
 - Reinsch, C., 679, 704
 - removing,
 - see deleting
 - removing variable from model, 272
 - removing variables from model, 235
 - Renew**, 139, 149, 354, 401, 402

- renewing windows, 401
 - repeated points
 - in contour plots, 538
 - required variables, 80, 110
 - residual, 227
 - residual normal QQ Plot
 - fit analyses, 661
 - residual plots, 661
 - residual normal quantiles
 - fit analyses, 693
 - residual plots
 - fit analyses, 659
 - partial leverage plots, 661
 - residual normal QQ Plot, 661
 - residual-by-predicted plot, 659
 - residual-by-hat diagonal plot
 - multiple regression, 230
 - residual-by-predicted plot
 - fit analyses, 659
 - multiple regression, 227
 - residual plots, 659
 - residuals, 691
 - fit analyses, 693
 - generalized, 692
 - saving, 238
 - studentized, 238
 - residuals-by-predicted plot
 - analysis of variance, 254
 - logistic regression, 270
 - resizing,
 - see sizing
 - resources, 467
 - RESP option, 787
 - Response Dist.**, 631
 - response distribution, 618, 631, 638
 - binomial, 619
 - gamma, 619
 - generalized linear models, 619
 - inverse Gaussian, 619
 - normal, 619
 - Poisson, 619
 - response surface, 116
 - response variable, 220, 614, 615, 629
 - results window, 447
 - robust estimation
 - Gini's mean difference, 576
 - trimmed means, 580
 - Winsorized means, 580
 - role, 490
 - variables, 80
 - root mean square error, 225, 251, 640
 - Root MSE,
 - see root mean square error
 - ROTATE statement, 786
 - rotating planes and surfaces, 552
 - rotating plot, 110
 - features of, 114
 - matrix of, 115
 - of principal components, 754
 - Rotating Plot (Z Y X)**, 110, 115, 116, 547
 - rotating plots, 545
 - method, 548
 - of canonical scores, 476
 - output, 549
 - variables, 547
 - rows
 - in data window, 491
 - RUN statement, 788
 - Run:Submit**, 28, 385, 425, 474
- ## S
- Sall, J., 512, 516
 - sample mean, 640, 642
 - in box plots, 83
 - Sample Mean, Std Dev**, 597
 - SAS data set, 27
 - SAS/GRAPH software, 431, 436
 - SAS/INSIGHT, 23
 - options, 454
 - SAS/INSIGHT Software
 - purpose of, 5
 - SAS/INSIGHT software
 - configuring, 453
 - exiting, 67
 - features of, 6
 - help system, 15, 18
 - invoking, 28, 50
 - learning, 15
 - operation of, 8
 - SAS/INSIGHT statements, 779
 - SAS/STAT software, 472
 - SASHELP library, 488
 - SASUSER library, 488
 - SASUSER.PROFILE catalog, 466
 - Save:Data**, 422
 - Save:Tables**, 446
 - saving
 - bitmaps, 437
 - catalogs, 436
 - colors, 474
 - data, 421, 422
 - defaults, 131
 - formats, 378
 - graphics, 431, 436
 - graphics files, 437
 - include/exclude state, 474
 - label/unlabel state, 474
 - markers, 474
 - observation states, 422, 492
 - options, 453, 466
 - principal components, 316
 - residuals, 238
 - select state, 474
 - show/hide state, 474
 - tables, 445, 449, 450
 - tables as data sets, 450
 - tables as html, 450
 - variables, 316, 691

Index

Scale, 631

scale

- of graphs, 147

SCALE option, 788

scale parameter, 642

- fit analyses, 622

- generalized linear models, 622

- quasi-likelihood, 624

scale parameter estimates

- constant, 622

- generalized linear models, 622

- maximum-likelihood estimate, 622

- mean deviance, 622

- mean Pearson chi-squared, 622

scale parameters, 624

scaled deviance

- generalized linear models, 622

scaled Pearson chi-squared

- generalized linear models, 623

scatter plot

- adding curves, 207

- confidence ellipses, 299, 768

- of principal components, 314, 754

Scatter Plot (Y X), 91, 94, 136, 270, 314, 394, 407, 528

Scatter Plot (Y X), 232

scatter plot matrix, 94, 394, 527, 528

- bivariate plots, 753

- in Multivariate analysis, 299

- layout, 95

- multivariate analyses, 753

scatter plot smoother

- fit analyses, 626

scatter plots, 87, 91, 527

- clicking in, 92

- method, 528

- output, 529

- variables, 528

- viewing brushed observations, 531

SCATTER statement, 785

schematic plot,

- see box plot

scientific format, 379

scientific notation,

- see exponential format

Scott, D.W., 501

screen,

- see display

scroll bar, 51, 52

scrolling, 52

- data window, 51

searching, 493

searching for observations, 59

seed, random,

- see random

select state, 474, 491

selecting, 8

- area, 255, 406

- comparison circles, 516

- contours, 540

- level curves, 540

- observations, 92

- tables, 446

- values in tables, 383

selection, 8

- extended, 9, 10, 99

- multiple, 9

- noncontiguous, 10

- order of, 81

Serifs, 512

set properties, 141,

- see variable properties

setting

- default window options, 45

shape,

- of observation markers, see marker

- of cursor, see cursor

shape parameter, 586

Shift key, 9

SHIP data set, 281, 282, 434

- wave damage, 21

Show in Graphs, 148

Show/Hide state, 491

show/hide state, 474

sign statistic

- distribution location tests, 573

signed rank statistic

- distribution location tests, 573

significance, 226, 252

Silverman, B.W., 593, 609, 704

simple regression, 204

simultaneous confidence intervals, 515

Singpurwalla, N.D., 601, 609

size of

- graphs, 460

size of markers, 162

Size to Fit, 165

sizing

- graphs, 411

skewness, 559, 568, 569

- distribution analyses, 559, 568, 569

- in box plots, 81

slicing

- observations, 145, 153

slider

- in scroll bar, 51

slope

- QQ ref line, 603

Smirnov, N., 609

smoother degrees of freedom

- fit analyses, 627

- fit curves, 627

smoother generalized cross validation

- fit analyses, 628

- fit curves, 628

smoothing parameter, 626

- kernel estimation, 592

- of kernel curve, 212

- smoothing spline, 538, 548, 664, 679
 - fit analyses, 663, 679
 - fit curves, 663, 679
 - smoothness of fit, 664
 - Snell, E.J., 704
 - Solutions**, 472
 - Sort**, 494
 - sorting
 - data, 56
 - observations, 56, 494
 - order of, 494
 - spinning,
 - see rotating
 - spline, 614
 - Spread**, 467
 - spreading
 - window positions, 460, 467
 - spreadsheet,
 - see data window
 - Sqrt,
 - see square root
 - SSCP, 773,
 - see sums of squares and crossproducts
 - standard deviation, 383, 512, 568, 569, 597, 651, 727
 - in box plots, 83
 - standard error, 270, 649, 650
 - trimmed mean, 580
 - Winsorized mean, 580
 - standard error of the mean
 - distribution analyses, 568, 569
 - standardized residuals
 - fit analyses, 696
 - statements
 - BAR, 783
 - BOX, 784
 - BY, 783
 - CLASS, 783
 - CONTOUR, 785
 - DIST, 786
 - FIT, 787
 - LINE, 784
 - MULT, 786
 - OPEN, 782
 - PROC INSIGHT, 781
 - QUIT, 789
 - recording, 481
 - ROTATE, 786
 - RUN, 788
 - SAS/INSIGHT, 779
 - SCATTER, 785
 - TABLES, 788
 - WINDOW, 782
 - states,
 - see observation states
 - statistical models
 - fit analyses, 614
 - statistical significance, 226, 252
 - statistics, descriptive,
 - see descriptive statistics
 - statistics, diagnostic,
 - see diagnostic statistics
 - statistics, summary,
 - see summary statistics
 - statistics, univariate,
 - see univariate statistics
 - Std Dev,
 - see standard deviation
 - Stephens, M.A., 601, 609
 - storing,
 - see saving
 - Student's t statistic
 - distribution location tests, 573
 - studentized residuals, 238
 - fit analyses, 696
 - Submit**, 28, 385, 425, 474
 - subsets
 - coloring observations, 172
 - group variables, 357
 - hiding observations, 146
 - marking observations, 160
 - of data, 495
 - of observations, 340
 - toggling display of observations, 149
 - sum, 568, 569
 - sum of squares, 225, 252, 643, 646
 - sum of squares corrected for the mean
 - distribution analyses, 568, 569
 - sum of squares for error
 - fit analyses, 616
 - sum of weights, 568, 569
 - summary of fit, 225, 251, 270, 643
 - analysis of variance, 251
 - logistic regression, 270
 - multiple regression, 225
 - Poisson regression, 287
 - summary of fit for generalized linear models
 - fit analyses, 642
 - summary of fit for linear models
 - fit analyses, 640
 - summary statistics, 225, 251, 270, 287
 - sums of squares
 - Type III, 226, 252
 - sums of squares and crossproducts, 727
 - surface plots, 116, 547
 - surface, rotating, 552
 - symbols,
 - see markers
- T**
- Tab key, 32, 44
 - Tables**, 249, 267, 446, 571, 638, 652, 727
 - tables
 - deleting, 408, 410
 - distribution analyses, 568
 - fit analyses, 638
 - html, 450
 - moving, 411
 - multivariate analyses, 727

Index

- printing, 445
 - saving, 445, 450
 - TABLES statement, 788
 - Tables:Collinearity Diagnostics**, 228
 - Tables:Location Tests**, 574
 - Tables:Type III (LR) Tests**, 271, 274
 - Techniques**, 17
 - techniques, 15
 - temporary
 - label, 137
 - Terrell, G.R., 501
 - test for a specific distribution
 - cumulative distribution, 599
 - distribution analyses, 599
 - test for distribution
 - cumulative distribution, 601
 - distribution analyses, 601
 - exponential, 601
 - lognormal, 601
 - normal, 601
 - Weibull, 601
 - tests
 - type I, 644
 - type I (LR), 645
 - type III, 645, 646
 - type III (LR), 648
 - type III (Wald), 648
 - thin-plate smoothing spline, 664
 - thin-plate splines, 538, 548
 - threshold parameter, 597
 - throwing, 98
 - Tibshirani, R.J., 212, 213, 704
 - Ticks**, 361, 503
 - ticks
 - adjusting, 126, 361
 - aligning, 417
 - font for labels, 434
 - in bar chart, 503
 - major, 126
 - minor, 126
 - size of labels, 434
 - titles, 437, 438, 441
 - toggling display of
 - observations, 145, 149
 - TOL,
 - see tolerance
 - tolerance, 649
 - fit analyses, 650
 - in multiple regression, 226
 - Tools**, 159, 171, 395
 - tools
 - magnifying glass, 395
 - windows, 54
 - TOOLS option, 781
 - tools window, 395, 396
 - trace, 713
 - transformation, 319
 - log, 282
 - transformations, 332
 - transforming variables, 319, 352, 353
 - tri-cube weight
 - fit analyses, 684
 - triangular
 - kernel estimation, 592
 - kernel function, 667, 682
 - weight function, 684
 - triangular distribution
 - kernel estimation, 592
 - triangular kernel
 - fit analyses, 667, 682
 - triangular weight
 - fit analyses, 684
 - trimmed mean
 - distribution analyses, 580
 - standard error, 580
 - trimmed means
 - distribution analyses, 580
 - robust estimation, 580
 - trimmed t statistic
 - distribution analyses, 580
 - Trimmed/Winsorized Means**, 565
 - Tukey, J.W., 507, 580, 581, 609
 - Tukey, P.A., 609
 - Tukey-Kramer method, 515
 - two-color blends, 173
 - type I tests
 - fit analyses, 644
 - Type III (LR) Tests**, 271, 274
 - type III (LR) tests
 - logistic regression, 271
 - type III (Wald) tests
 - logistic regression, 270
 - Poisson regression, 287
 - type III tests
 - analysis of variance, 252
 - fit analyses, 646
 - likelihood-ratio statistic, 647
 - multiple regression, 226
 - Wald statistic, 647
- ## U
- uncorrected sum of squares
 - distribution analyses, 568, 569
 - undo,
 - see Renew
 - Renew, 354
 - Unf/For** button, 494
 - uniform lens, 531
 - univariate statistics
 - descriptive statistics, 727
 - multivariate analyses, 727
 - UNIX operating system
 - setting X resources, 467
 - UnLabel in Plots**, 138
 - Use Obs with Missing Values**, 562
 - user's guide
 - conventions of, 15
 - using, 15

USS,
 see uncorrected sums of squares
 distribution analyses, 568, 569

V

Values, 257, 503

values

 in bar chart, 503

variable roles, 500, 509

Variable:Both, 502, 513, 523, 529, 539, 549

Variable:Labels, 502, 513, 523, 529, 539, 549

Variable:Names, 502, 513, 523, 529, 539, 549

Variables, 320

variables, 27, 489

 adding, 35

 arranging, 52

 bar chart, 500

 box plots, 509

 contour plots, 537

 default role, 141

 defining, 37

 deleting, 349

 distribution analyses, 558

 editing, 319

 explanatory, 220

 fit analyses, 629, 691

 frequency, 500, 509, 558, 630, 708

 generated, 692

 group, 509, 522, 528, 537, 547, 558, 630, 708

 in box plot, 509

 in scatter plots, 528

 influence diagnostics, 691

 interval, 51, 72

 label, 500, 509, 522, 528, 537, 547, 548, 558,
 630, 708

 line plot, 522

 measurement level, 51, 490

 moving, 52

 multivariate analyses, 708

 names of, 692

 new, 494

 nominal, 51, 75

 optional, 81

 removing from model, 235, 272

 response, 220

 role, 80, 490

 rotating plots, 547

 saving, 316, 691

 selected, 43

 selecting, 53

 transforming, 319, 352, 353

 weight, 558, 630, 709

 X, 522, 528, 537, 547

 Y, 500, 509, 522, 528, 537, 547, 558

 Z, 537, 547

Variables:log(Y), 282, 353

Variables:log(Y), 321

Variables:Other, 325, 329

variance, 559, 568, 569

 fit analyses, 616

variance divisor

 multivariate analyses, 712

variance function, 619

 generalized linear models, 618

 quasi-likelihood, 624

variance inflation

 in multiple regression, 226

variance inflation factor, 649

 fit analyses, 650

variance proportion, 657

variation

 sources of, 225, 252, 643, 644

Vars, 691

Vars:Dffits, 234

Vars:Hat Diag, 230

Vars:Output Principal Components:2, 316

Vars:Studentized Residual, 238

Velleman, P.F., 704

vertical,

 see orientation

Vertical Axis at Left, 502, 513, 523, 529, 539

View:Results, 447

viewing clustered data, 531

VIF,

 see variance inflation factor

visualization, 549

W

Wald

 confidence interval, 654

Wald statistic

 type III tests, 647

Wald tests, 270, 287

Watts, D.G., 20

wave damage

 SHIP data set, 21

Weibull

 quantile, 587

 test for distribution, 601

Weibull distribution

 distribution analyses, 557

 fit parametric, 597

 parametric distributions, 557

 testing for, 601

Weibull quantile

 distribution analyses, 587

weight function

 local polynomial fit, 684

 loess fit, 684

 normal, 684

 quadratic, 684

 triangular, 684

WEIGHT option, 786, 787

weight role, 490

weight values

 in distribution analyses, 559

 in fit analyses, 631

 in multivariate analyses, 710

Index

Weight variable
 distribution analyses, 558
 fit analyses, 630
 multivariate analyses, 709
weight variable, 773
weighted analyses
 fit analyses, 702
 multivariate analyses, 773
Weil, G., 545
Weisberg, S., 238, 239, 704
Welsch, R.E., 233, 239, 657, 692, 704
Whisker Length, 511
whiskers, 81, 511
width
 of bars, 501, 503
 of curves, 208
 of formatted values, 380
Wilks, A.R., 96
window, 8
 options, 458, 459
Window Layout:Spread, 467
Window options, 458, 459
WINDOW statement, 782
windows
 closing, 67
 copying, 345
 data, 50, 488
 editing, 393
 output, 425
 position of, 460
 printing, 440
 renewing, 401, 402
 results, 447
 tools, 54, 396
 zooming, 394
Windows:Align, 362, 417
Windows:Animate, 369
Windows:Copy Window, 345, 351, 352, 360
Windows>Delete, 410
Windows:Display Options, 435
Windows:Freeze, 346
Windows:Options, 458
Windows:Renew, 139, 149, 354, 401, 402
Windows:Tools, 159, 171, 395
Winsorized mean
 distribution analyses, 580
 standard error, 580
Winsorized means
 distribution analyses, 580
 robust estimation, 580
Winsorized sum of squared deviations
 distribution analyses, 580
Winsorized t statistic
 distribution analyses, 580
WORK library, 488
working with other SAS products, 471

X

X button, 629

X resources, 467
X variable
 fit analyses, 629
 in contour plots, 537
 in line plot, 522
 in rotating plots, 547
X variable effects
 fit analyses, 629
X variables
 in scatter plots, 528
X`X matrix, 650, 657
 fit analyses, 639
XAXIS option, 783–786

Y

Y Axis Vertical, 502, 513, 523, 529, 539
Y variable
 fit analyses, 629
 in box plot, 500
 in box plots, 509
 in contour plots, 537
 in distribution analyses, 558
 in line plot, 522
 in rotating plots, 547
Y variables
 in scatter plots, 528
YAXIS option, 783–786

Z

Z variable
 in contour plots, 537
 in rotating plots, 547
ZAXIS option, 785, 786
ZColor variable
 in rotating plots, 548
zooming, 394, 397–399

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **`yourturn@sas.com`**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **`suggest@sas.com`**.

SAS® Publishing delivers!

Whether you are new to the workforce or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart.

SAS® Press Series

Need to learn the basics? Struggling with a programming problem? You'll find the expert answers that you need in example-rich books from the SAS Press Series. Written by experienced SAS professionals from around the world, these books deliver real-world insights on a broad range of topics for all skill levels.

support.sas.com/saspress

SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information—SAS documentation. We currently produce the following types of reference documentation: online help that is built into the software, tutorials that are integrated into the product, reference documentation delivered in HTML and PDF—free on the Web, and hard-copy books.

support.sas.com/publishing

SAS® Learning Edition 4.1

Get a workplace advantage, perform analytics in less time, and prepare for the SAS Base Programming exam and SAS Advanced Programming exam with SAS® Learning Edition 4.1. This inexpensive, intuitive personal learning version of SAS includes Base SAS® 9.1.3, SAS/STAT®, SAS/GRAPH®, SAS/QC®, SAS/ETS®, and SAS® Enterprise Guide® 4.1. Whether you are a professor, student, or business professional, this is a great way to learn SAS.

support.sas.com/LE



**THE
POWER
TO KNOW®**

