SAS/IML® Studio 13.1
User's Guide

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2013. *SAS/IML® Studio 13.1: User's Guide*. Cary, NC: SAS Institute Inc.

SAS/IML® Studio 13.1: User's Guide

Copyright © 2013, SAS Institute Inc., Cary, NC, USA

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.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a) and DFAR 227.7202-4 and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

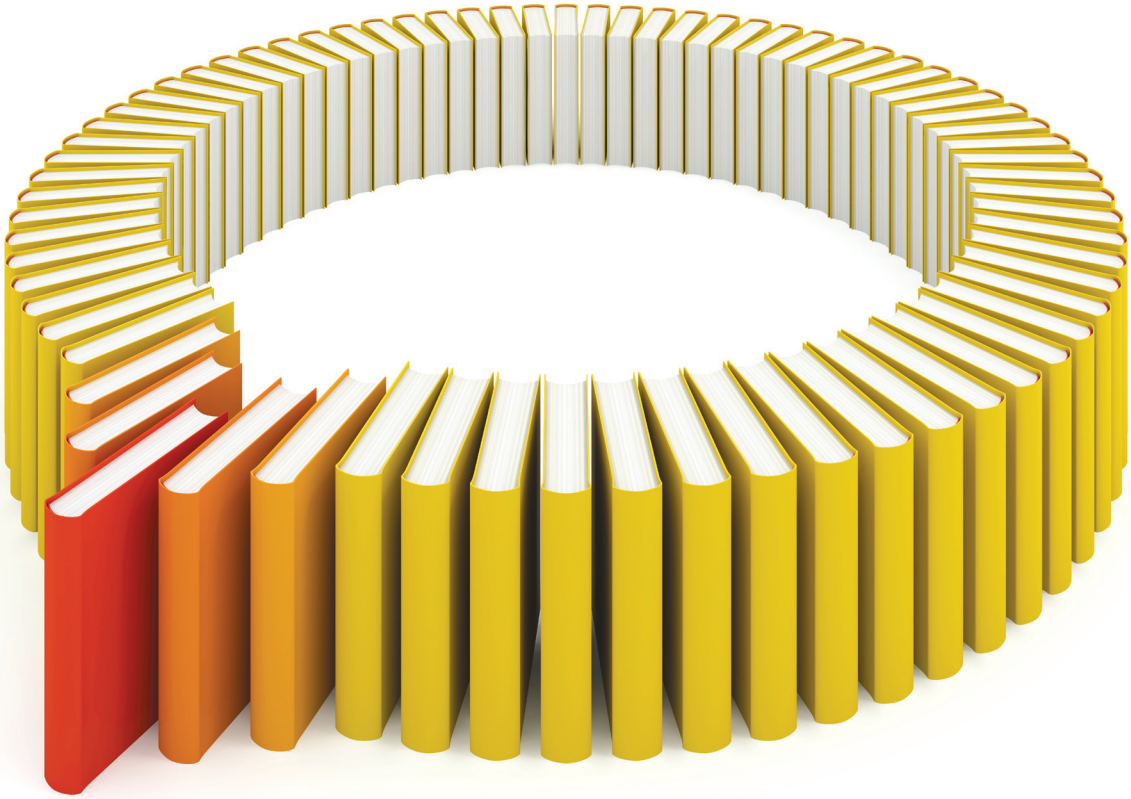
SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

December 2013

SAS provides a complete selection of books and electronic products to help customers use SAS® software to its fullest potential. For more information about our offerings, visit **support.sas.com/bookstore** 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 trademarks of their respective companies.



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.

sas
THE POWER TO KNOW.®

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 trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0613

Contents

Chapter 1. Introduction to SAS/IML Studio . . . 1
Chapter 2. Getting Started with SAS/IML Studio . . . 11
Chapter 3. Creating and Editing Data . . . 29
Chapter 4. Interacting with the Data Table . . . 37
Chapter 5. Exploring Data in One Dimension . . . 59
Chapter 6. Exploring Data in Two Dimensions . . . 77
Chapter 7. Exploring Data in Three Dimensions . . . 103
Chapter 8. Interacting with Plots . . . 131
Chapter 9. General Plot Properties . . . 145
Chapter 10. Axis Properties . . . 165
Chapter 11. Techniques for Exploring Data . . . 173
Chapter 12. Plotting Subsets of Data . . . 199
Chapter 13. Distribution Analysis: Descriptive Statistics . . . 217
Chapter 14. Distribution Analysis: Location and Scale Statistics . . . 225
Chapter 15. Distribution Analysis: Distributional Modeling . . . 233
Chapter 16. Distribution Analysis: Frequency Counts . . . 247
Chapter 17. Distribution Analysis: Outlier Detection . . . 255
Chapter 18. Data Smoothing: Loess . . . 263
Chapter 19. Data Smoothing: Thin-Plate Spline . . . 279
Chapter 20. Data Smoothing: Polynomial Regression . . . 289
Chapter 21. Model Fitting: Linear Regression . . . 301
Chapter 22. Model Fitting: Robust Regression . . . 323
Chapter 23. Model Fitting: Logistic Regression . . . 339
Chapter 24. Model Fitting: Generalized Linear Models . . . 361
Chapter 25. Multivariate Analysis: Correlation Analysis . . . 391
Chapter 26. Multivariate Analysis: Principal Component Analysis . . . 403
Chapter 27. Multivariate Analysis: Factor Analysis . . . 421
Chapter 28. Multivariate Analysis: Canonical Correlation Analysis . . . 441
Chapter 29. Multivariate Analysis: Canonical Discriminant Analysis . . . 451
Chapter 30. Multivariate Analysis: Discriminant Analysis . . . 469
Chapter 31. Multivariate Analysis: Correspondence Analysis . . . 481
Chapter 32. Variable Transformations . . . 495
Chapter 33. Running Custom Analyses . . . 527
Chapter 34. Configuring the SAS/IML Studio Interface . . . 535
Appendix A. Sample Data Sets . . . 555
Appendix B. SAS/INSIGHT Features Not Available in SAS/IML Studio . . . 567

Index **569**

Release Notes

The following release notes pertain to SAS/IML® Studio 13.1:

- For a complete list of new features in this release of SAS/IML Studio, see the SAS/IML Studio online Help: **Help ► Help Topics**. The following list highlights a few new features:
  - SAS/IML Studio supports a new type of workspace called a SAS Workspace. A SAS Workspace runs standard SAS programs, as opposed to IMLPlus programs.
  - The SAS/IML Studio program editor highlights matching braces, brackets, and parentheses. The editor supports improved color coding and automatically highlights syntax errors.
  - New keyboard shortcuts include CTRL+G to go to a line in a program editor, and CTRL+E to clear the HTML output window or the Log window.
  - The TEMPLATE procedure now works in SAS/IML Studio.
  - The online documentation has been restructured.
- SAS/IML Studio includes documentation for Base SAS 9.4, SAS/IML 13.1, and SAS/STAT 13.1. See `http://support.sas.com/documentation/` for current versions of SAS documentation.
- If you need to open a data set that contains Chinese, Japanese, or Korean characters, it is important that you configure the "Regional and Language Options" in the Windows Control Panel for the appropriate country. It is not necessary to change the Windows setting called "Language for non-Unicode programs," which is also referred to as the *system locale*.
- The SAS/IML Studio user interface is available in the following languages: English, Japanese, Korean, and Simplified Chinese. When you run SAS/IML Studio on a Windows system configured for a language other than English, you can still use English fonts by defining the *IMLStudio_ForceEnglishUI* environment variable. For more information, refer to the "Configuration" section of the chapter "The IML Studio Enviroment" in the SAS/IML Studio online Help.
- SAS/IML Studio uses the Microsoft Access Database Engine to import Microsoft Excel worksheets. The 32-bit edition of SAS/IML Studio requires the 32-bit edition of the Access Database Engine. Similarly, the 64-bit edition of SAS/IML Studio requires the 64-bit edition of the Access Database Engine.

  Windows does not permit you to install the 32-bit and the 64-bit editions of Access Database Engine at the same time. If your Windows system does not have the appropriate edition of Access Database Engine installed, you will be unable to import Microsoft Excel worksheets by using the Open File dialog box or the IMLPlus method DataObject.CreateFromExcelFile. In this situation, you should use the IMPORT procedure, which is part of SAS/ACCESS® Interface to PC Files.

Chapter 1
Introduction to SAS/IML Studio

Contents

What Is SAS/IML Studio? 1
Related Software and Documentation 2
Exploratory and Confirmatory Data Analysis 3
How Many Observations Can You Analyze? 3
Summary of Features 4
Comparison with SAS/INSIGHT Software 6
References 9

What Is SAS/IML Studio?

SAS/IML Studio is a tool for data exploration and analysis. Figure 1.1 shows a typical SAS/IML Studio analysis. You can use SAS/IML Studio to do the following:

- explore data through graphs linked across multiple windows
- subset data
- analyze univariate distributions
- fit explanatory models
- investigate multivariate relationships

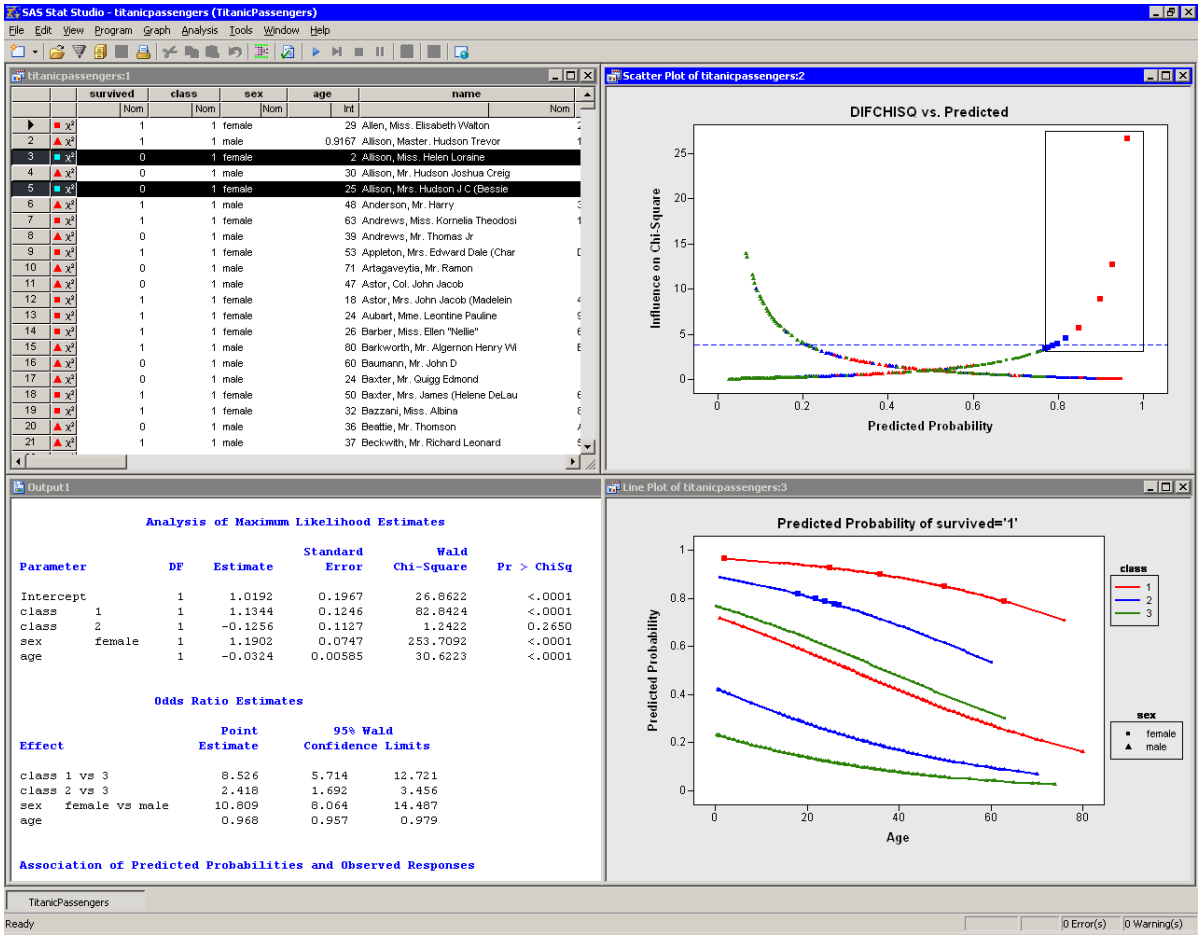
In addition, SAS/IML Studio provides an integrated development environment that enables you to write, debug, and execute programs that combine the following:

- the flexibility of the SAS/IML® matrix language
- the analytical power of SAS/STAT® procedures
- the data manipulation capabilities of Base SAS® software
- the dynamically linked graphics of SAS/IML Studio
- the functions and user-contributed packages of the open-source R language

The programming language in SAS/IML Studio, which is called *IMLPlus*, is an enhanced version of the SAS/IML programming language. The "Plus" part of the name refers to new features that extend the SAS/IML language, including the ability to create and manipulate statistical graphs, to call SAS procedures, and to call functions in the R programming language.

SAS/IML Studio requires that you have a license for Base SAS, SAS/STAT, and SAS/IML software. SAS/IML Studio runs on a PC in the Microsoft Windows operating environment.

Figure 1.1 The SAS/IML Studio Interface



Related Software and Documentation

This book is one of three documents about SAS/IML Studio. In this book you learn how to use the SAS/IML Studio GUI to conduct exploratory data analysis and standard statistical analyses.

A second book, *SAS/IML Studio for SAS/STAT Users*, is intended for SAS/STAT programmers. In it, you learn how to use SAS/IML Studio in conjunction with SAS/STAT software in order to explore data and visualize statistical models. In particular, you learn to call procedures in other SAS products such as SAS/STAT or Base SAS software by using the SUBMIT statement.

The third source of documentation is the SAS/IML Studio online Help. You can display the online Help by selecting **Help ►Help Topics** from the main menu. The online Help includes documentation for all IMLPlus classes and associated methods.

SAS/IML Studio is part of SAS/IML software. The language used to write programs in SAS/IML Studio is called *IMLPlus*. This language contains SAS/IML functions and statements implemented in the IML procedure and documented in the *SAS/IML User's Guide*. The IML procedure runs entirely on a SAS workspace server, whereas IMLPlus switches dynamically between a SAS server (for computations) and the PC client (for graphics). In short, the IMLPlus language consists of SAS/IML functions and subroutines "plus" additional syntax to support the creation and manipulation of statistical graphics. The SAS/IML Studio program windows uses color coding to display keywords in the IMLPlus language.

Most SAS/IML programs run without modification in the IMLPlus environment. The SAS/IML Studio online Help includes a list of differences between the SAS/IML language and IMLPlus.

For your convenience in referencing related SAS software, the *SAS/IML User's Guide*, the *SAS/STAT User's Guide*, and the *Base SAS Procedures Guide* are available from the SAS/IML Studio **Help** menu.

Exploratory and Confirmatory Data Analysis

Data analysis often falls into two phases: exploratory and confirmatory. The exploratory phase "isolates patterns and features of the data and reveals these forcefully to the analyst" (Hoaglin, Mosteller, and Tukey 1983). If a model is fit to the data, exploratory analysis finds patterns that represent deviations from the model. These patterns lead the analyst to revise the model, and the process is repeated.

In contrast, confirmatory data analysis "quantifies the extent to which [deviations from a model] could be expected to occur by chance" (Gelman 2004). Confirmatory analysis uses the traditional statistical tools of inference, significance, and confidence.

Exploratory data analysis is sometimes compared to detective work: it is the process of gathering evidence. Confirmatory data analysis is comparable to a court trial: it is the process of evaluating evidence. Exploratory analysis and confirmatory analysis "can—and should—proceed side by side" (Tukey 1977).

How Many Observations Can You Analyze?

SAS/IML Studio provides the data analyst with interactive and dynamic statistical graphics. By definition, interactive graphics must respond quickly to the changes and manipulations of the analyst. This quick response restricts the size of data sets that can be handled while still maintaining interactivity.

Wegman (1995) points out that the number of observations you can analyze depends on the algorithmic complexity of the statistical algorithms you are using. For example, if you have n observations, computing a mean and variance is $O(n)$, sorting is $O(n \log n)$, and solving a least squares regression on p variables is $O(np^2)$. Furthermore, visualization of individual observations is limited by the number of pixels that can be represented on a display device.

Wegman's conclusion is that "visualization of data sets say of size 10^6 or more is clearly a wide open field." More recently, Unwin, Theus, and Hofmann (2006) discuss the challenges of "visualizing a million," including a chapter dedicated to interactive graphics.

On a typical PC (for example, a 1.8 GHz CPU with 512 MB of RAM), SAS/IML Studio can help you analyze dozens of variables and tens of thousands of observations. Visualization of data with graphics such as histograms and box plots remains feasible for hundreds of thousands of observations, although the interactive graphics become less responsive. Scatter plots of this many observations suffer from overplotting.

SAS/IML Studio uses the RAM on your PC to facilitate interaction and linking between plots and data tables. If you routinely analyze large data sets, increasing the RAM on your PC might increase SAS/IML Studio's interactivity. For example, if you routinely examine hundreds of thousands of observations in dozens of variables, 1 GB of RAM is preferable to 512 MB.

Summary of Features

SAS/IML Studio provides tools for exploring data, analyzing distributions, fitting parametric and nonparametric regression models, and analyzing multivariate relationships. In addition, you can extend the set of available analyses by writing programs.

To explore data, you can do the following:

- identify observations in plots
- select observations in linked data tables, bar charts, box plots, contour plots, histograms, line plots, mosaic plots, and two- and three-dimensional scatter plots
- exclude observations from graphs and analyses
- search, sort, subset, and extract data
- transform variables
- change the color and shape of observation markers based on the value of a variable

To analyze distributions, you can do the following:

- compute descriptive statistics
- create quantile-quantile plots
- create mosaic plots of cross-classified data
- fit parametric and kernel density estimates for distributions
- detect outliers in contaminated Gaussian data

To fit parametric and nonparametric regression models, you can do the following:

- smooth two-dimensional data by using polynomials, loess curves, and thin-plate splines
- add confidence bands for mean and predicted values
- create residual and influence diagnostic plots
- fit robust regression models and detect outliers and high-leverage observations
- fit logistic models
- fit the general linear model with a wide variety of response and link functions
- include classification effects in logistic and generalized linear models

To analyze multivariate relationships, you can do the following:

- calculate correlation matrices and scatter plot matrices with confidence ellipses for relationships among pairs of variables
- reduce dimensionality with principal component analysis
- examine relationships between a nominal variable and a set of interval variables with discriminant analysis
- examine relationships between two sets of interval variables with canonical correlation analysis
- reduce dimensionality by computing common factors for a set of interval variables with factor analysis
- reduce dimensionality and graphically examine relationships between categorical variables in a contingency table with correspondence analysis

To extend the set of available analyses, you can do the following:

- write, debug, and execute IMLPlus programs in an integrated development environment
- add legends, curves, maps, or other custom features to statistical graphics
- create new static graphics
- animate graphics
- execute SAS procedures or DATA steps from within your IMLPlus programs
- develop interactive data analysis programs that use dialog boxes
- call computational routines written in C, FORTRAN, Java, R, or the SAS/IML language

Comparison with SAS/INSIGHT Software

SAS/IML Studio and SAS/INSIGHT® Software have the same goal: to be a tool for data exploration and analysis. Both have dynamically linked statistical graphics. Both come with pre-written statistical analyses for analyzing distributions, regression models, and multivariate relationships.

Figure 1.2 shows a typical SAS/INSIGHT analysis. Figure 1.3 shows the same analysis performed in SAS/IML Studio. You can see that the analyses are qualitatively similar.

Figure 1.2 A SAS/INSIGHT Analysis

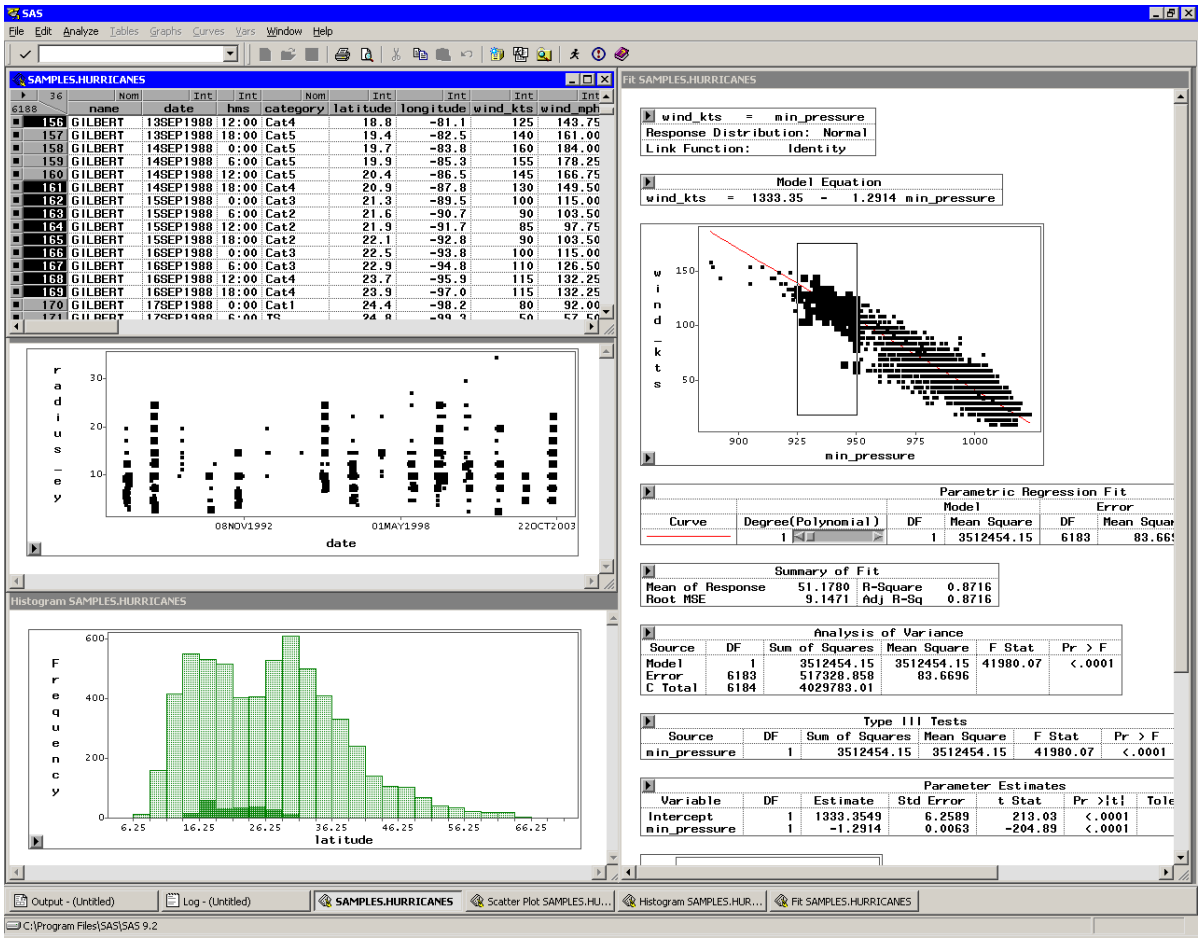
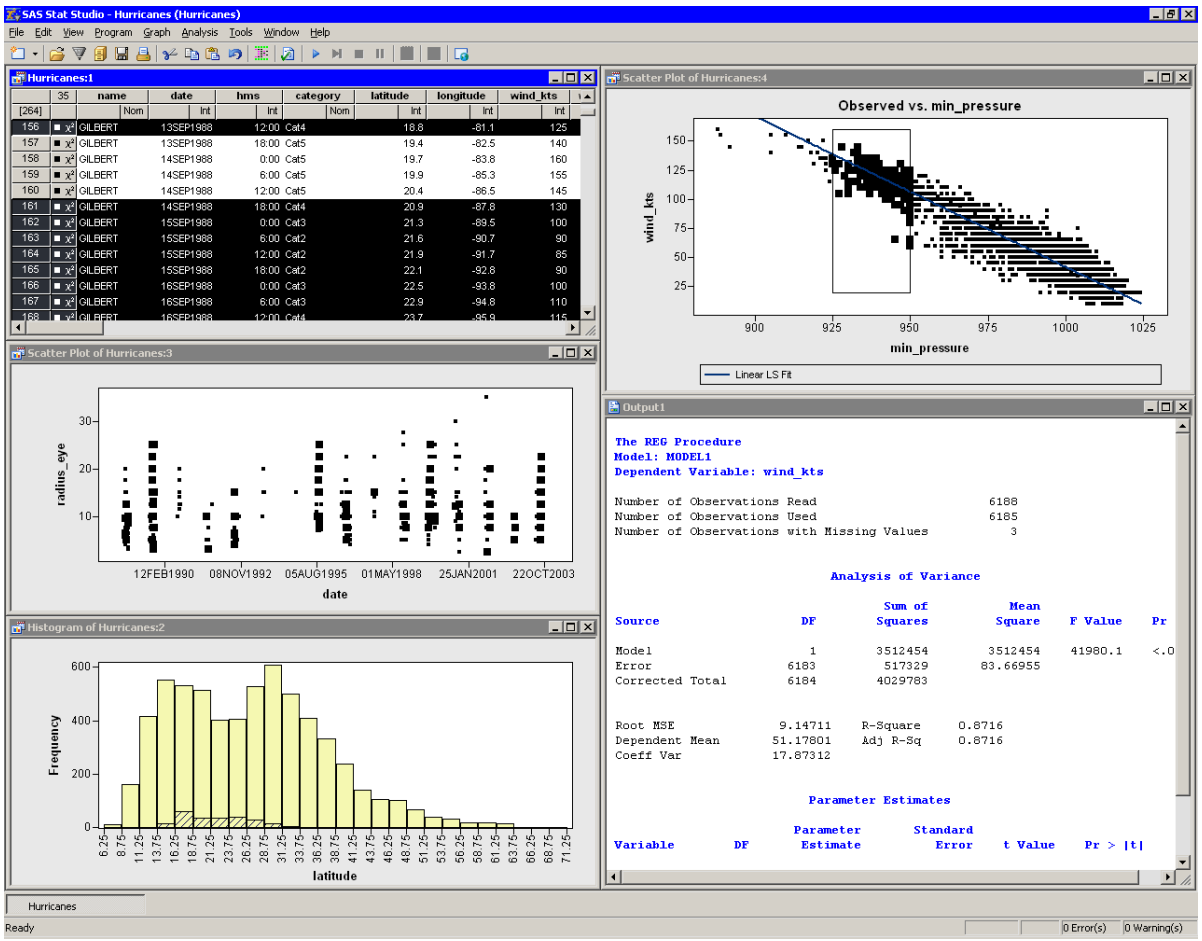


Figure 1.3 A Comparable SAS/IML Studio Analysis



However, there are three major differences between the two products. The first is that SAS/IML Studio runs on a PC in the Microsoft Windows operating environment. It is *client* software that can connect to SAS servers. The SAS server might be running on a different computer than SAS/IML Studio. In contrast, SAS/INSIGHT software runs on the same computer on which the SAS software is installed.

A second major difference is that SAS/IML Studio is programmable, and therefore extensible. SAS/INSIGHT software contains standard statistical analyses that are commonly used in data analysis, but you cannot create new analyses. In contrast, you can write programs in SAS/IML Studio that call any licensed SAS procedure, and you can include the results of that procedure in graphics, tables, and data sets. Because of this, SAS/IML Studio is sometimes referred to as the "programmable successor to SAS/INSIGHT software."

A third major difference is that the SAS/IML Studio statistical graphics are programmable. You can add legends, curves, and other features to the graphics in order to better analyze and visualize your data.

SAS/IML Studio contains many features that are not available in SAS/INSIGHT software. General features that are unique to SAS/IML Studio include the following:

- SAS/IML Studio can connect to multiple SAS servers simultaneously.
- SAS/IML Studio can run multiple programs simultaneously in different threads; each program has its own WORK library.
- SAS/IML Studio sessions can be driven by a program and rerun.

SAS/IML Studio provides the following features of data views (tables and plots) which are not included in SAS/INSIGHT software:

- modern dialog boxes with a native Windows look and feel
- a line plot in which the lines can be defined by specifying a single X variable and a single Y variable, and one or more grouping variables
- a polygon plot that can be used to build interactive regions such as maps
- programmatic methods to draw legends, curves, or other decorations on any plot
- programmatic methods to attach a menu to any plot. After the menu is selected, a user-specified program is run.
- arbitrary unions and intersections of observations selected in different views

SAS/IML Studio also provides the following analyses and options that are not included in SAS/INSIGHT software:

- a programming language that can call any licensed SAS analytical procedure and any SAS/IML function or subroutine.
- outlier detection in contaminated Gaussian data
- robust regression models and detection of outliers and high-leverage observations
- the generalized linear model with a multinomial response
- graphical results for the analysis of logistic models with one continuous effect and a small number of levels for classification effects
- parametric and nonparametric methods of discriminant analysis
- common factor analysis for interval variables
- correspondence analysis for nominal variables

Features of SAS/INSIGHT software that are not included in SAS/IML Studio are presented in Appendix B, "SAS/INSIGHT Features Not Available in SAS/IML Studio."

References

Gelman, A. (2004), "Exploratory Data Analysis for Complex Models," *Journal of Computational and Graphical Statistics*, 13, 755–779.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., eds. (1983), *Understanding Robust and Exploratory Data Analysis*, Wiley Series in Probability and Mathematical Statistics, New York: John Wiley & Sons.

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

Unwin, A., Theus, M., and Hofmann, H. (2006), *Graphics of Large Datasets*, New York: Springer.

Wegman, E. J. (1995), "Huge Data Sets and the Frontiers of Computational Feasibility," *Journal of Computational and Graphical Statistics*, 4, 281–295.

Chapter 2
Getting Started with SAS/IML Studio

Contents

Overview of SAS/IML Studio 11
Overview of the Sample Data 12
Open the Data Set 12
Create a Bar Chart 14
Exclude Observations 16
Create a Histogram 17
Create a Box Plot 20
Create a Scatter Plot 23
Model Variable Relationships 24
References 28

Overview of SAS/IML Studio

SAS/IML Studio provides a powerful programming environment that enables you to combine SAS/IML statements with calling SAS procedures, and also enables you to create and manipulate the attributes of dynamically linked statistical graphics. SAS/IML Studio also provides a GUI that enables you to visualize the results of statistical analyses. Furthermore, SAS/IML Studio provides several prewritten analyses (all implemented in IMLPlus, the SAS/IML Studio programming language) that you can access from the **Analysis** menu.

This chapter describes how you can use the SAS/IML Studio GUI for exploratory data analysis. The example in this chapter uses a sample data set, Hurricanes, that is distributed with SAS/IML Studio. The example covers the following activities:

1. Opening a data set. When you open a data set, the data are displayed in a data table. Features of the data table are described in Chapter 4, "Interacting with the Data Table."

2. Creating graphical views of the data, such as a bar chart, a histogram, a box plot, and a scatter plot. SAS/IML Studio plots and data tables are collectively known as *data views*. All data views are *dynamically linked*, which means that observations that you select in one data view are displayed as selected in all other views of the same data. Several chapters of this book are devoted to describing the SAS/IML Studio plots and how you can interact with them. Especially relevant to this example are Chapter 5, "Exploring Data in One Dimension," and Chapter 6, "Exploring Data in Two Dimensions."

3. Modeling relationships between variables. The example uses the correlation analysis and the polynomial regression analysis. These analyses are described further in Chapter 20, "Data Smoothing: Polynomial Regression," and Chapter 25, "Multivariate Analysis: Correlation Analysis."

Overview of the Sample Data

This example shows how you can use SAS/IML Studio to explore data about North Atlantic tropical cyclones. (A *cyclone* is a large system of winds that rotate about a center of low atmospheric pressure.) The data were recorded by the U.S. National Hurricane Center at six-hour intervals during the years 1988 to 2003.

The example analyzes the following variables:

category	indicator variable that corresponds to the Saffir-Simpson wind intensity scale
latitude	latitude of observation, in degrees north latitude
min_pressure	minimum central sea-level pressure, in hPa
radius_eye	radius of eye (if an eye exists), in nautical miles
wind_kts	maximum low-level sustained wind speed, in knots

The category variable is a measure of wind intensity, corresponding to the Saffir-Simpson wind intensity scale in Table 2.1.

Table 2.1 The Saffir-Simpson Intensity Scale

Category	Description	Wind Speed (Knots)
TD	Tropical depression	22–33
TS	Tropical storm	34–63
Cat1	Category 1 hurricane	64–82
Cat2	Category 2 hurricane	83–95
Cat3	Category 3 hurricane	96–113
Cat4	Category 4 hurricane	114–134
Cat5	Category 5 hurricane	135 or greater

The analysis presented in this chapter is based on Mulekar and Kimball (2004) and Kimball and Mulekar (2004). A full description of the Hurricanes data set is included in Chapter A, "Sample Data Sets."

Open the Data Set

This chapter analyzes the Hurricanes data set, which is distributed with SAS/IML Studio.

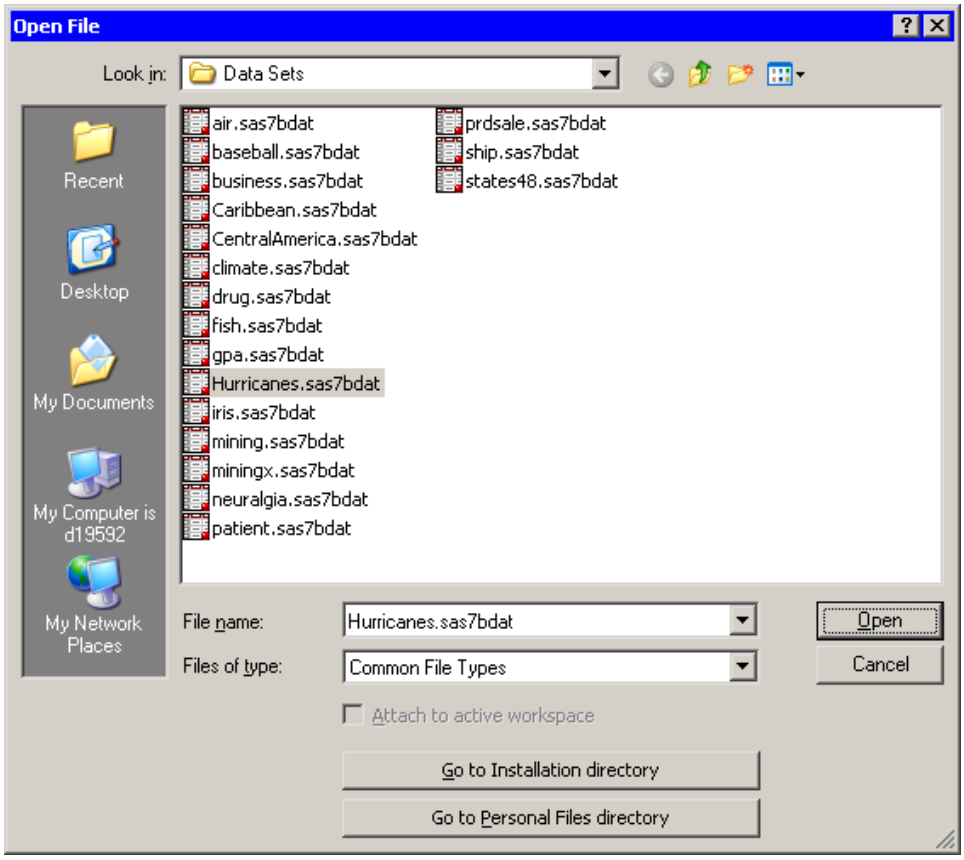
To use the GUI to open the data set:

1. Select **File ►Open ►File** from the main menu. The Open File dialog box appears. (See Figure 2.1.)
2. Click **Go to Installation directory** near the bottom of the dialog box.

3 Double-click the `Data Sets` folder.

4 Select the *Hurricanes.sas7bdat* file.

Figure 2.1 Opening a Sample Data Set



5 Click **Open**.

The data table in Figure 2.2 appears.

Figure 2.2 The Hurricanes Data

Hurricanes.sas7bdat

	34	name	date	hms	category	latitude	longitude	wind_kts
6188		Nom	Int	Int	Nom	Int	Int	Int
1	χ²	ALBERTO	05AUG1988	18:00		32	-77.5	20
2	χ²	ALBERTO	06AUG1988	0:00		32.8	-76.2	20
3	χ²	ALBERTO	06AUG1988	6:00		34	-75.2	20
4	χ²	ALBERTO	06AUG1988	12:00	TD	35.2	-74.6	25
5	χ²	ALBERTO	06AUG1988	18:00	TD	37	-73.5	25
6	χ²	ALBERTO	07AUG1988	0:00	TD	38.7	-72.4	25
7	χ²	ALBERTO	07AUG1988	6:00	TD	40	-70.8	30
8	χ²	ALBERTO	07AUG1988	12:00	TS	41.5	-69	35
9	χ²	ALBERTO	07AUG1988	18:00	TS	43	-67.5	35
10	χ²	ALBERTO	08AUG1988	0:00	TS	45	-65.5	35
11	χ²	ALBERTO	08AUG1988	6:00	TS	47	-63	35
12	χ²	ALBERTO	08AUG1988	12:00	TD	49	-60	30
13	χ²	ALBERTO	08AUG1988	18:00	TD	51	-56	25
14	χ²	BERYL	08AUG1988	0:00	TD	30.4	-90.3	25
15	χ²	BERYL	08AUG1988	6:00	TD	29.7	-89.7	30
16	χ²	BERYL	08AUG1988	12:00	TS	29.7	-89.4	35
17	χ²	BERYL	08AUG1988	18:00	TS	29.4	-89.2	40
18	χ²	BERYL	09AUG1988	0:00	TS	29.3	-89.1	45
19	χ²	BERYL	09AUG1988	6:00	TS	29.6	-89.5	45
20	χ²	BERYL	09AUG1988	12:00	TS	30.1	-90.4	45
21	χ²	BERYL	09AUG1988	18:00	TS	30.1	-90.9	40

The row heading of the data table includes two special cells for each observation: one that shows the location of the observation in the data set, and the other that shows the status of the observation in analyses and plots. The status of each observation is indicated by the presence or absence of a marker and a χ^2 symbol. The presence of a marker (by default, a filled square) indicates that the observation is included in plots; observations that are excluded from plots do not display a marker. Similarly, the χ^2 symbol indicates that the observation is included in analyses. The Hurricanes data initially has all observations included in plots and analyses. See Chapter 4, "Interacting with the Data Table," for more information about the data table symbols.

Create a Bar Chart

To create a bar chart of the category variable:

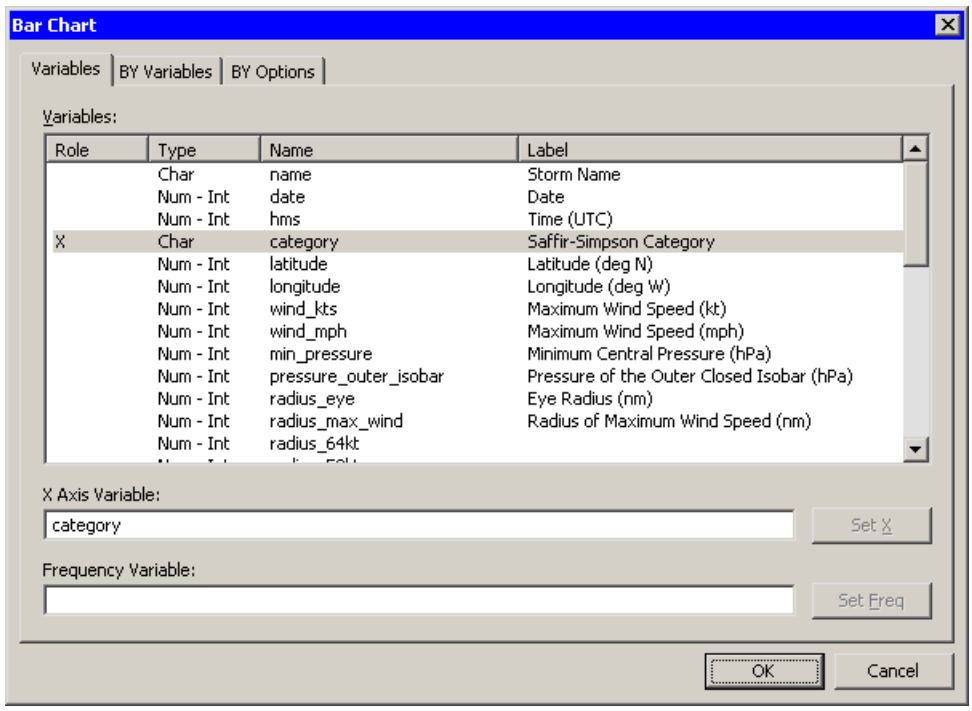
1 Select **Graph ►Bar Chart** from the main menu.

The Bar Chart dialog box appears. (See Figure 2.3.)

2 Select the variable category, and click **Set X**.

NOTE: In most dialog boxes, double-clicking a variable name adds the variable to the next appropriate field.

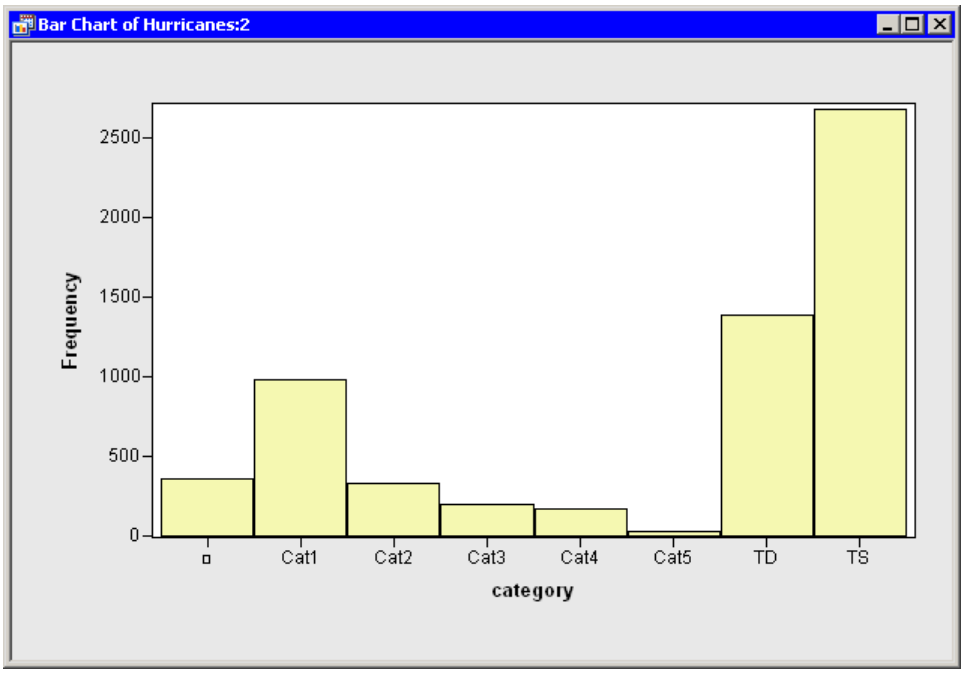
Figure 2.3 Bar Chart Dialog Box



3 Click **OK**.

The bar chart in Figure 2.4 appears. The bar chart shows the number of observations for storms in each Saffir-Simpson intensity category.

Figure 2.4 A Bar Chart



Exclude Observations

To exclude observations of less than tropical storm intensity (wind speeds less than 34 knots):

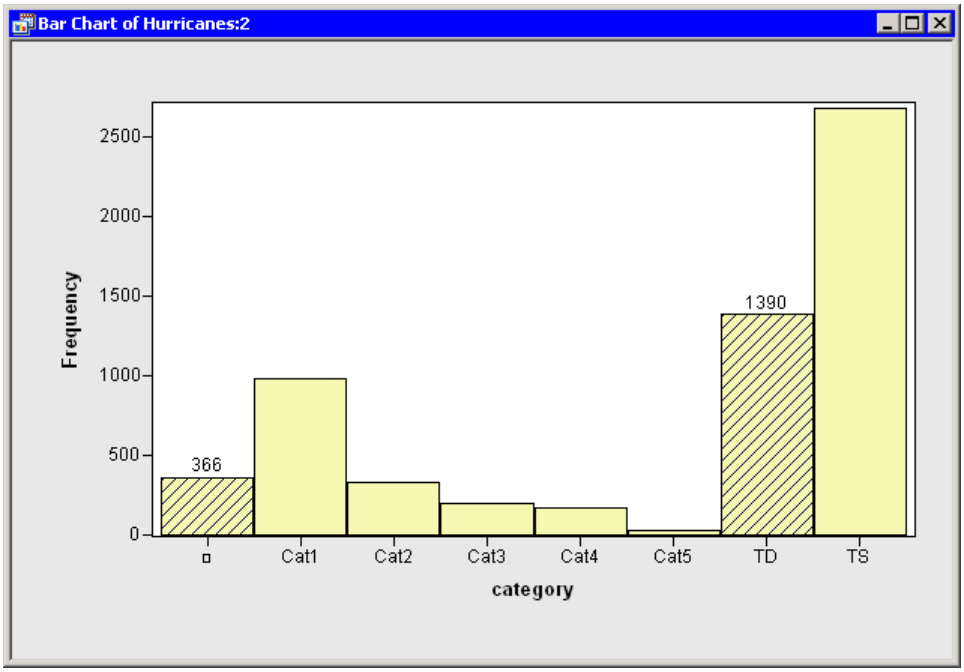
1 In the bar chart, click the bar labeled with the symbol □.

This selects observations for which the category variable has a missing value. For these data, "missing" is equivalent to an intensity of less than tropical depression strength (wind speeds less than 22 knots).

2 Hold down the CTRL key and click the bar labeled "TD."

When you hold down the CTRL key and click, you *extend* the set of selected observations. In this example, you select observations with tropical depression strength (wind speeds of 22–34 knots) without deselecting previously selected observations. The bars that contain selected observations are shown as crosshatched in Figure 2.5.

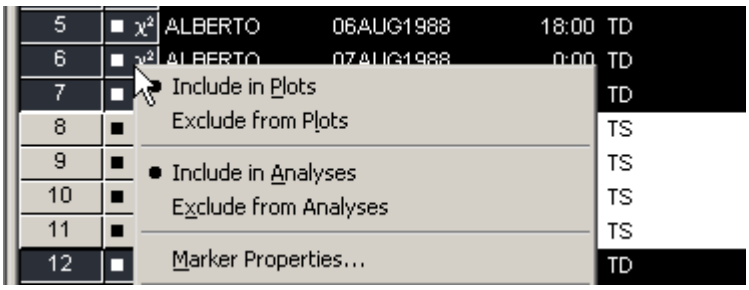
Figure 2.5 A Bar Chart with Selected Observations



3 In the data table, right-click in the row heading (to the left) of any selected observation, and select **Exclude from Plots** from the pop-up menu (shown in Figure 2.6).

Notice that the bar chart redraws itself to reflect that all observations being displayed in the plots now have at least 34-knot winds. Notice also that the square symbol in the data table is removed from observations with wind speeds less than 34 knots.

Figure 2.6 Data Table Pop-up Menu



4 In the data table, right-click in the row heading of any selected observation, and select **Exclude from Analyses** from the pop-up menu.

Notice that the χ^2 symbol is removed from observations with wind speeds less than 34 knots. Future analysis (for example, correlation analysis and regression analysis) will not use the excluded observations.

5 Click any data table cell to clear the selected observations.

NOTE: You can also exclude selected observations by using a keyboard shortcut. Select a plot and press the 'e' key to exclude selected observations from plots and from analyses. Additional keyboard shortcuts are described in Chapter 8, "Interacting with Plots."

Create a Histogram

In this section you create a histogram of the latitude variable and examine relationships between the category and latitude variables. The figures in this section assume that you have excluded observations with low wind speeds as described in the section "Exclude Observations" on page 16.

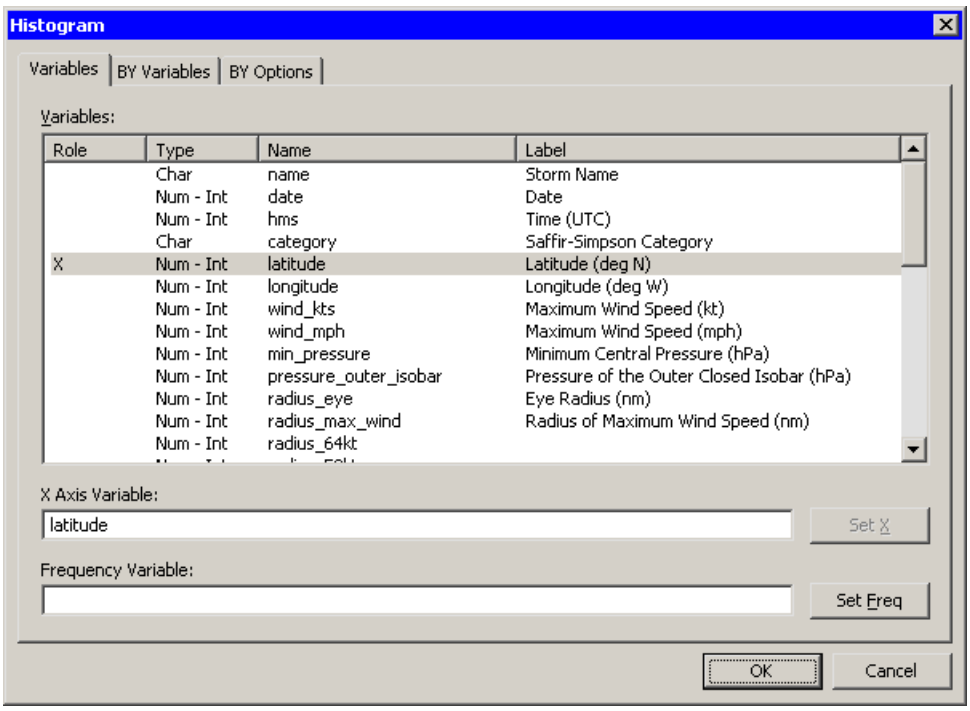
To create a histogram:

1 Select **Graph ►Histogram** from the main menu.

The Histogram dialog box appears. (See Figure 2.7.)

2 Select the variable latitude, and click **Set X**.

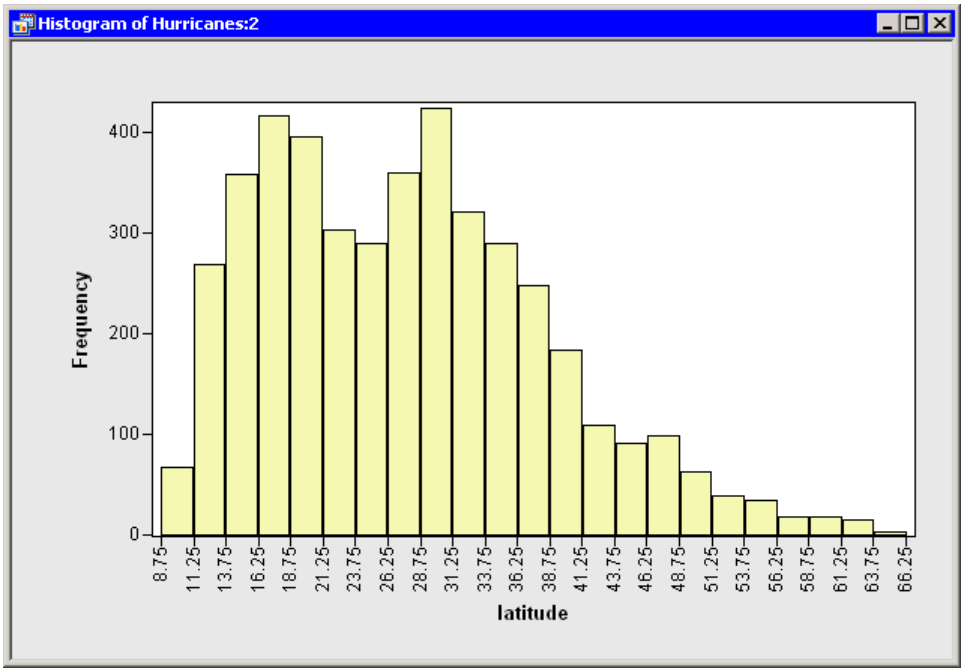
Figure 2.7 Histogram Dialog Box



3 Click **OK**.

A histogram (Figure 2.8) appears, which shows the distribution of the latitude variable for the storms that are included in the plots. Move the histogram so that it does not cover the bar chart or data table.

Figure 2.8 Histogram of Latitudes of Storms

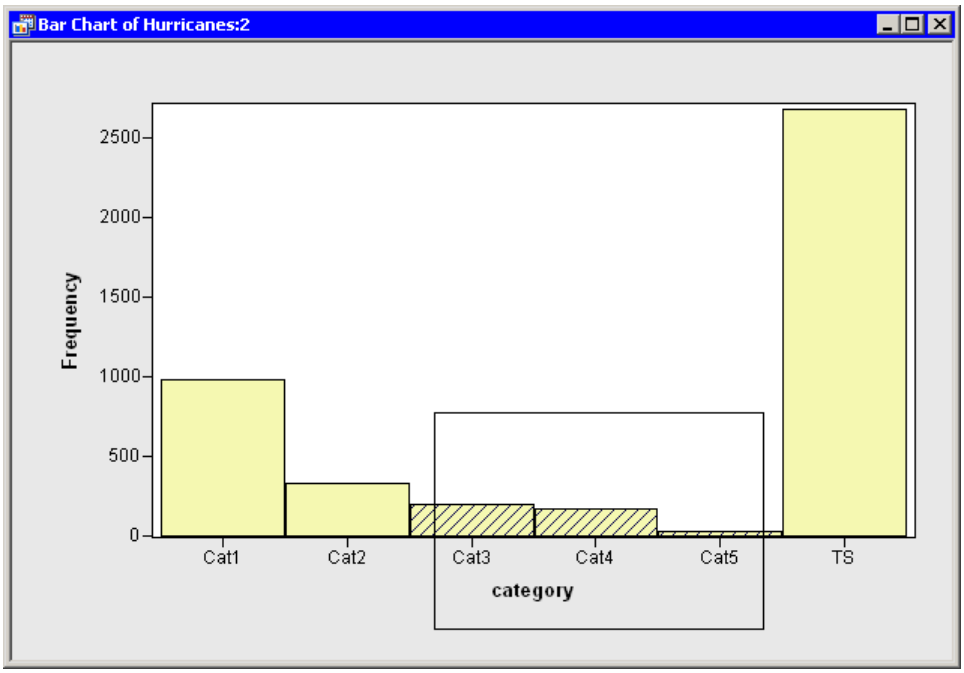


You have seen that you can select observations in a plot by clicking bars or observation markers. You can also select observations by drawing a *selection rectangle.* To draw a selection rectangle, click in a graph and hold down the left mouse button while you move the mouse pointer to a new location.

4 Draw a selection rectangle in the bar chart to select all storms of category 3, 4, and 5.

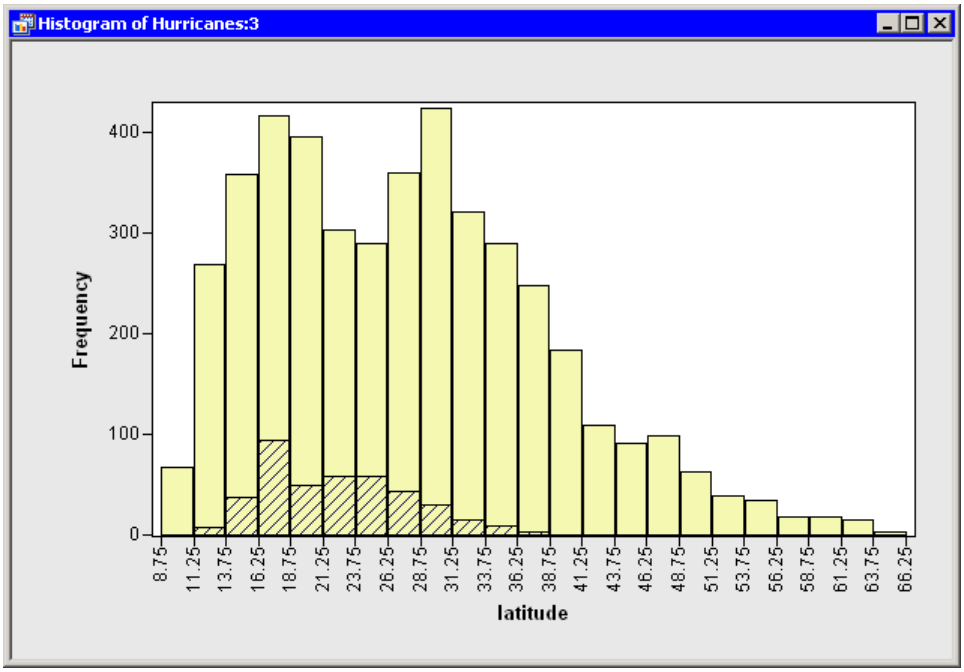
The bar chart looks like the one in Figure 2.9.

Figure 2.9 Selecting the Most Intense Storms



Note that these selected observations are also shown in the histogram in Figure 2.10. The histogram shows the conditional distribution of latitude, given that a storm is greater than or equal to category 3 intensity. The conditional distribution shows that very strong hurricanes tend to occur between 11 and 37 degrees north latitude, with a median latitude of about 22 degrees. If these data are representative of all Atlantic hurricanes, you might conjecture that it would be relatively rare for a category 3 hurricane to strike north of the North Carolina-Virginia border (roughly 36.5° north latitude).

Figure 2.10 Latitudes of Intense Storms



Create a Box Plot

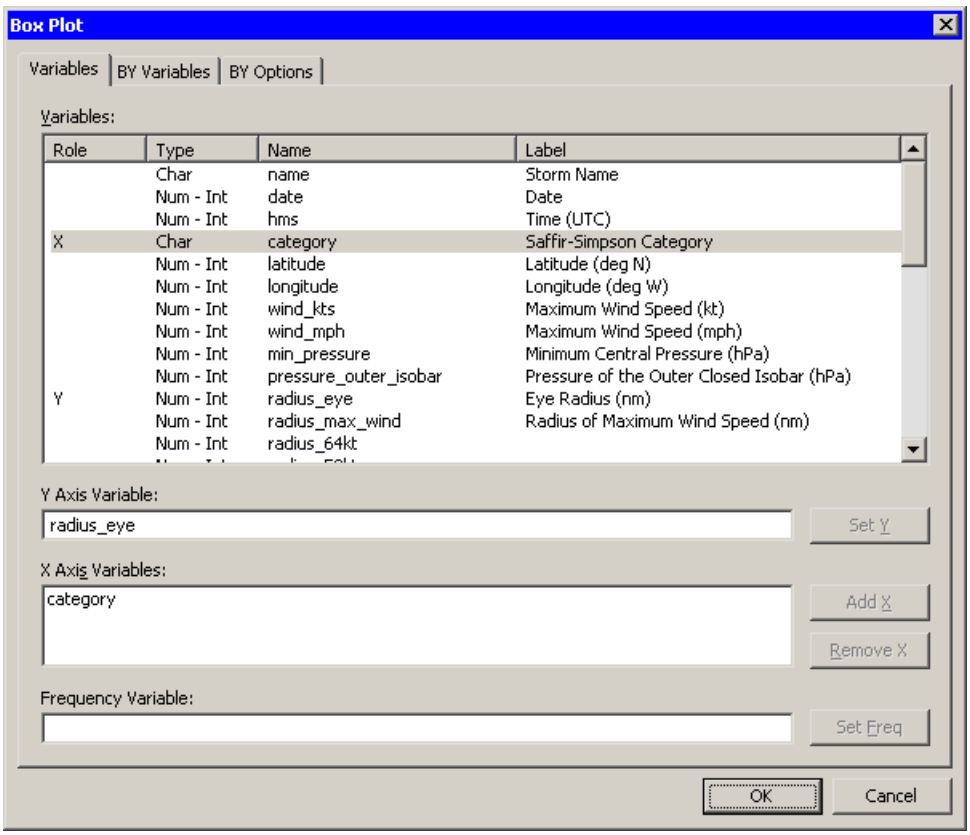
The data set contains several variables that measure the size of a tropical cyclone. One of these is the radius_eye variable, which contains the radius of a cyclone's eye in nautical miles. (The eye of a cyclone is a calm, relatively cloudless central region.) The radius_eye variable has many missing values, because not all storms have well-defined eyes.

The following steps create a box plot that shows how the radius of a cyclone's eye varies with the Saffir-Simpson category. The figures in this section assume that you have excluded observations with low wind speeds as described in the section "Exclude Observations" on page 16.

1 Select **Graph ►Box Plot** from the main menu.

The Box Plot dialog box appears. (See Figure 2.11.)

Figure 2.11 Box Plot Dialog Box



2 Select the variable radius_eye, and click **Set Y**.

3 Select the variable category, and click **Add X**.

4 Click **OK**.

A box plot appears as in Figure 2.12. Move the box plot so that it does not cover the data table or other plots.

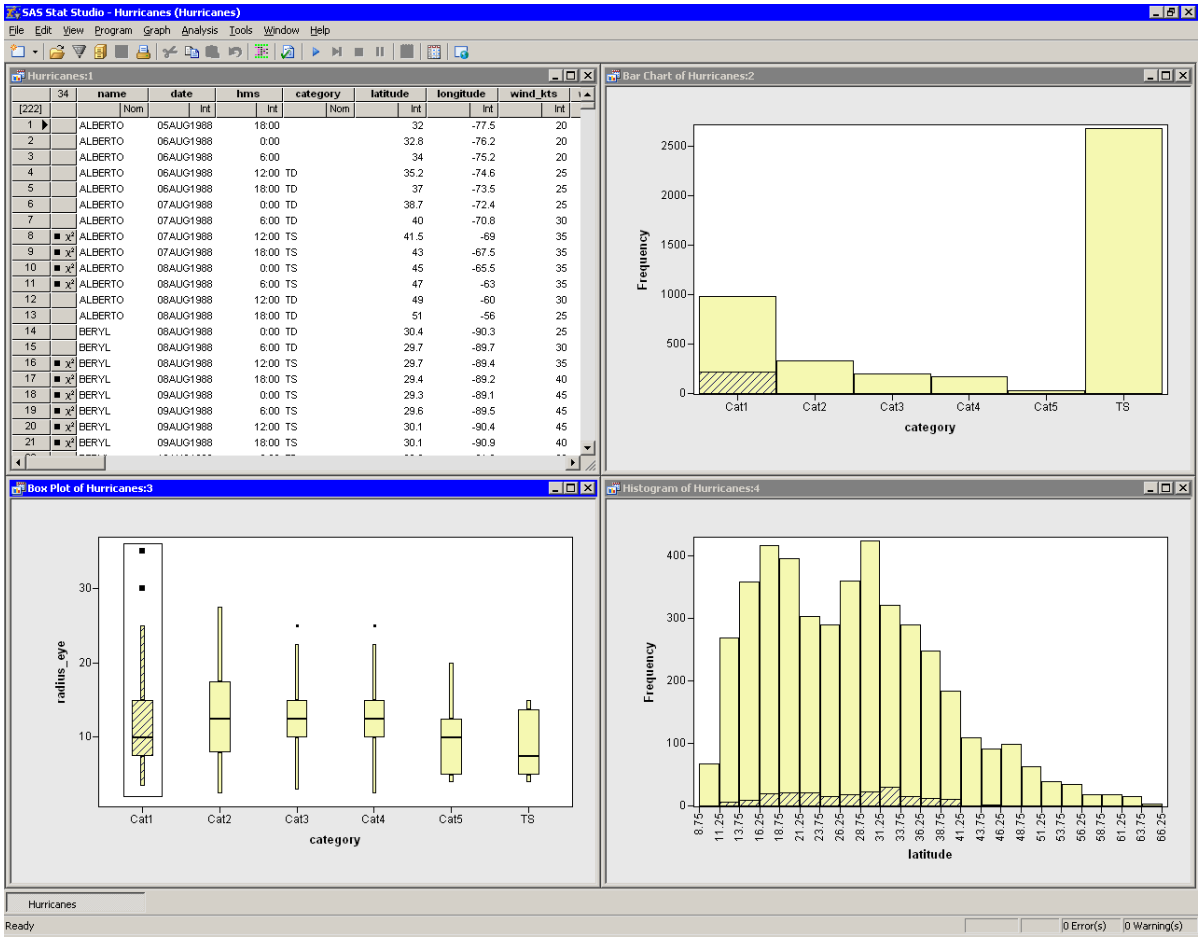
The box plot summarizes the distribution of eye radii for each Saffir-Simpson category. The plot indicates that the median eye radius tends to increase with storm intensity for tropical storms, category 1, and category 2 hurricanes. Category 2–4 storms have similar distributions, while the most intense hurricanes (category 5) in this data set tend to have eyes that are small and compact. The box plot also indicates considerable spread in the radii of eyes.

Recall that the radius_eye variable contains many missing values. These missing values are not displayed by the box plot. You might wonder what percentage of all storms of a given Saffir-Simpson intensity have well-defined eyes. You can determine this percentage by selecting all observations in one of the box plots and noting the proportion of observations that are selected in the bar chart.

5 Draw a selection rectangle in the box plot around the category 1 storms.

In the bar chart in Figure 2.12, note that approximately 25% of the bar for category 1 storms is displayed as selected, which means that approximately one quarter of the category 1 storms in this data set have nonmissing measurements for radius_eye.

Figure 2.12 Proportion of Category 1 Storms with Well-Defined Eyes



6 Drag the selection rectangle to select eye radii in other categories.

The selected observations displayed in the bar chart reveal the proportion of storms in each Saffir-Simpson category that have nonmissing values for radius_eye. Note in particular that very few tropical storms have eyes, whereas almost all category 4 and 5 storms have well-defined eyes.

7 Click outside the plot area in any plot to deselect all observations.

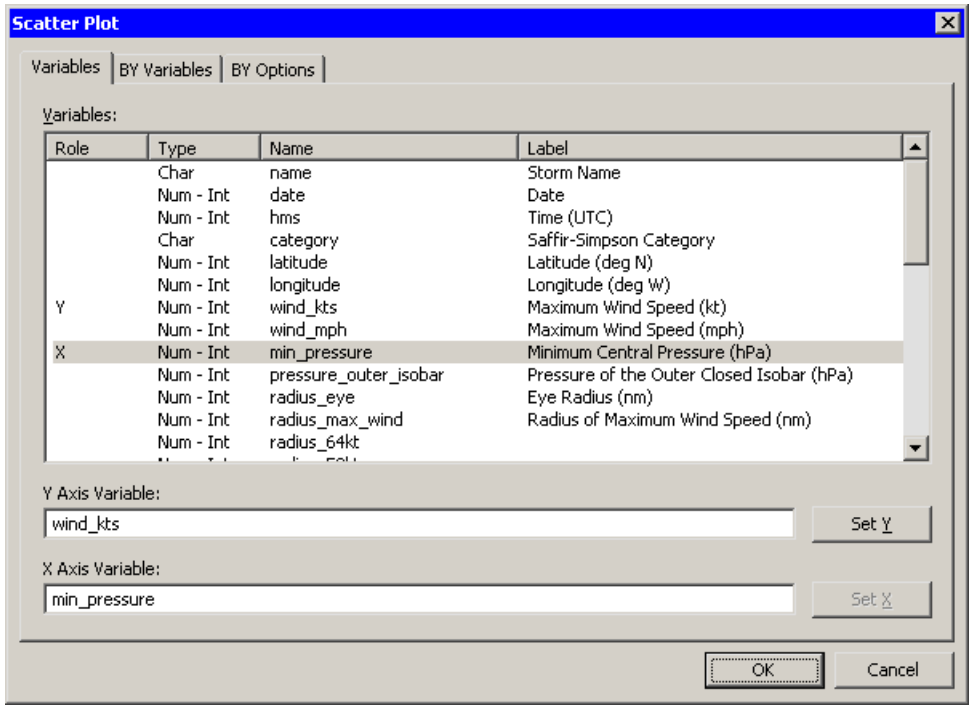
Create a Scatter Plot

The following steps examine the relationship between wind speed and atmospheric pressure for tropical cyclones. The National Hurricane Center routinely reports both of these quantities as indicators of a storm's intensity. The figures in this section assume that you have excluded observations with low wind speeds as described in the section "Exclude Observations" on page 16.

1 Select **Graph ►Scatter Plot** from the main menu.

The Scatter Plot dialog box appears. (See Figure 2.13.)

Figure 2.13 Scatter Plot Dialog Box



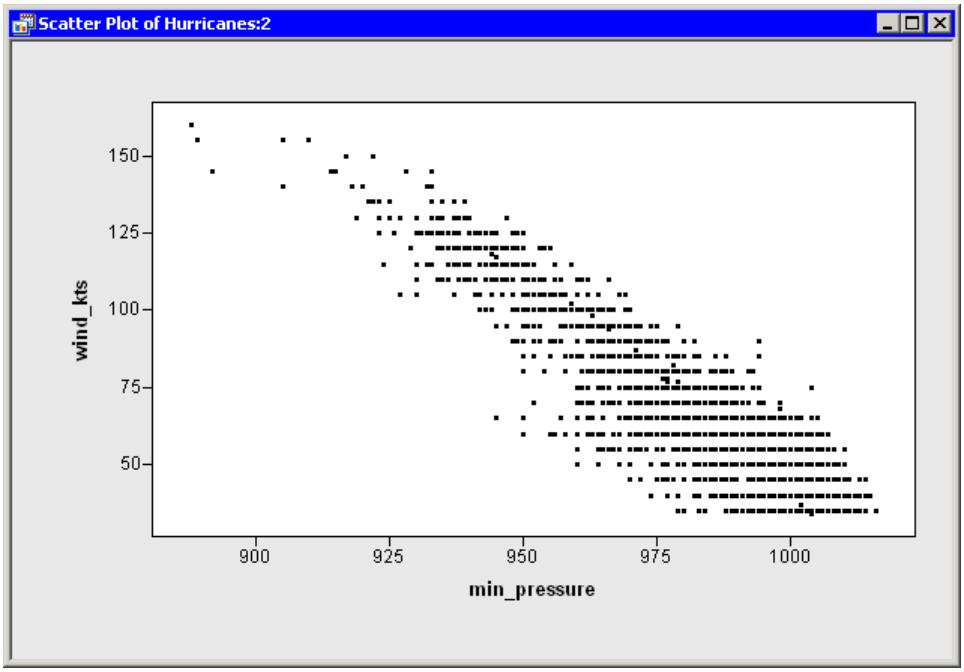
2 Select the variable wind_kts, and click **Set Y**.

3 Select the variable min_pressure, and click **Set X**.

4 Click **OK**.

A scatter plot appears as in Figure 2.14.

Figure 2.14 Wind Speed versus Minimum Pressure



Model Variable Relationships

In this section you model the relationship between wind speed and atmospheric pressure for tropical cyclones. The scatter plot in Figure 2.14 shows a strong negative correlation between wind speed and pressure. To compute the correlation between these variables, you can run SAS/IML Studio's correlation analysis. The results in this section assume that you have excluded observations with low wind speeds as described in the section "Exclude Observations" on page 16.

NOTE: You can select from the **Analysis** or **Graph** menu only when the *active window* is a data table or a graph. Click a window's title bar to make it the active window.

To run an analysis in SAS/IML Studio:

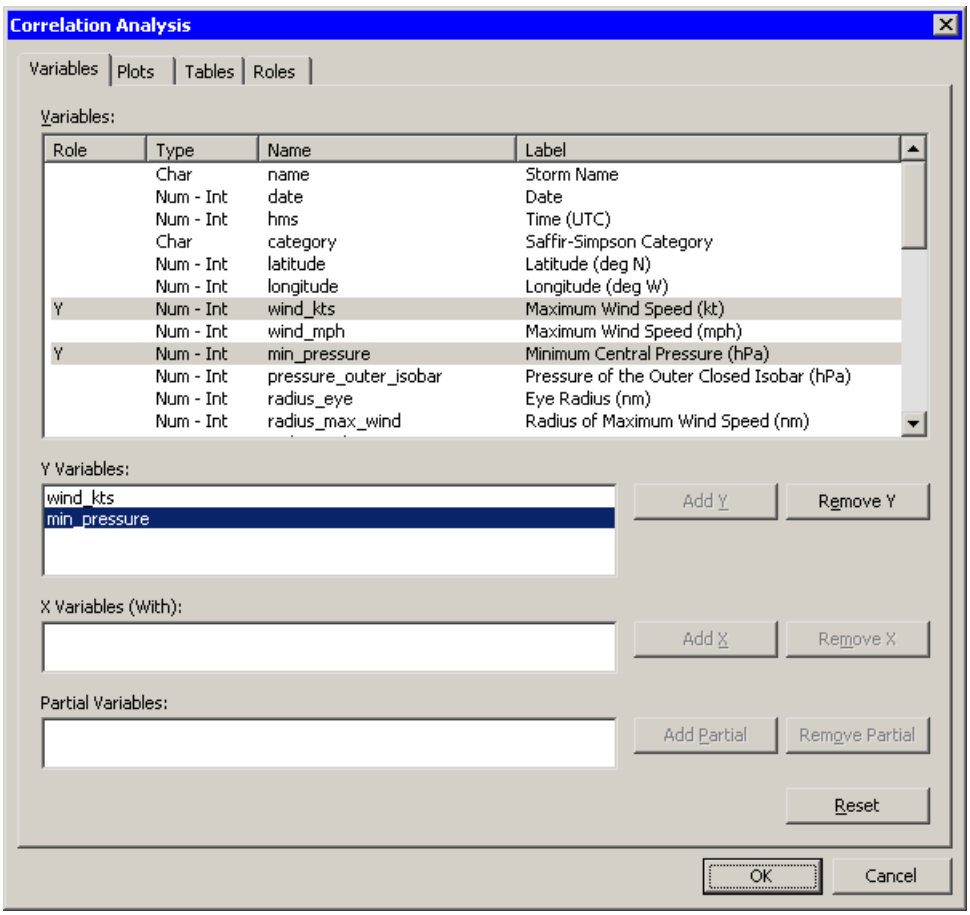
1 Select **Analysis ►Multivariate Analysis ►Correlation Analysis** from the main menu.

The Correlation Analysis dialog box appears. (See Figure 2.15.)

2 Click the wind_kts variable. Hold down the CTRL key and click the min_pressure variable. Click **Add Y**.

Both variables are added to the list of Y variables.

Figure 2.15 Correlations Analysis Dialog Box



3 Click the **Plots** tab.

4 Clear the **Pairwise correlation plot** check box.

5 Click **OK**.

See Chapter 25, "Multivariate Analysis: Correlation Analysis," for more information about the correlations analysis.

An output window appears (Figure 2.16), which shows the results from the CORR procedure. The output shows that the Pearson correlation between wind_kts and min_pressure is –0.92533.

Figure 2.16 Output from the CORR Procedure

Output1

```
            Pearson Correlation Coefficients
               Prob > |r| under H0: Rho=0
                  Number of Observations

                                                   min_
                                       wind_kts  pressure

wind_kts                                1.00000  -0.92533
Maximum Wind Speed (kt)                            <.0001
                                           4432      4431

min_pressure                           -0.92533   1.00000
Minimum Central Pressure (hPa)           <.0001
                                           4431      4431
```

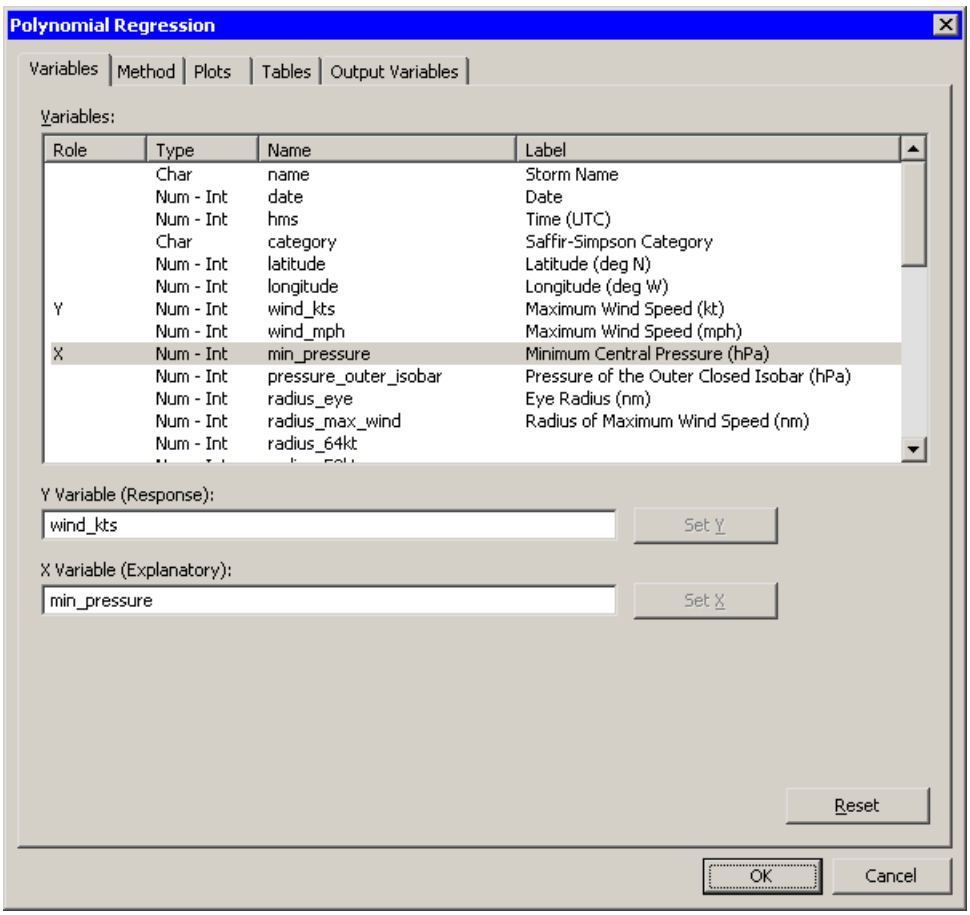
Suppose you want to compute a linear model that relates wind_kts to min_pressure. Several choices of parametric and nonparametric models are available from the **Analysis ►Model Fitting** menu. If you are interested in a response due to a single explanatory variable, you can also choose from models available from the **Analysis ►Data Smoothing** menu.

NOTE: If the scatter plot of wind_kts versus min_pressure is the active window when you select an analysis from the **Analysis ►Data Smoothing** menu, then the data smoother is added to the existing scatter plot. Otherwise, a new scatter plot is created by the analysis.

6 Activate the scatter plot of wind_kts versus min_pressure. Select **Analysis ►Data Smoothing ►Polynomial Regression** from the main menu.

The Polynomial Regression dialog box appears. (See Figure 2.17.)

Figure 2.17 Polynomial Smoother Dialog Box



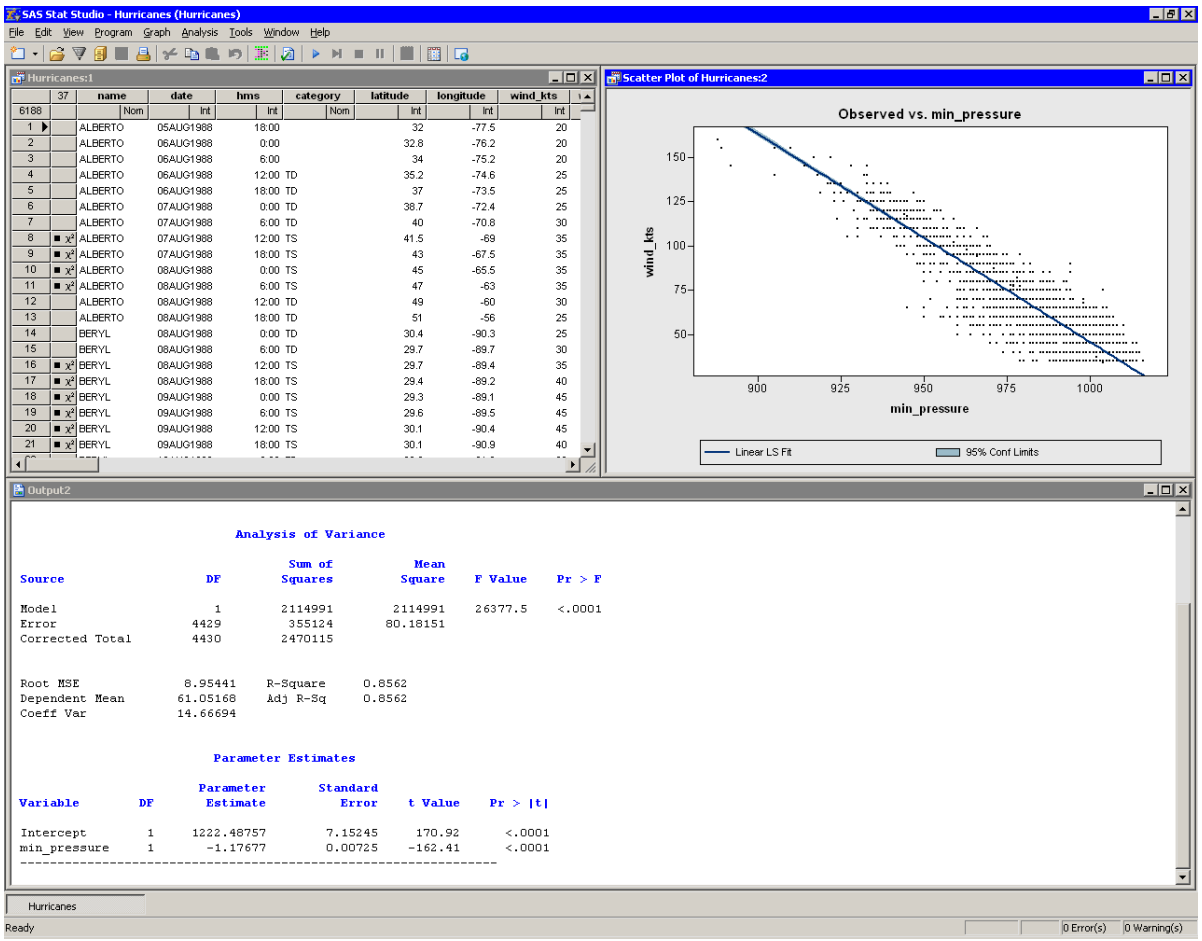
7 Select the variable wind_kts, and click **Set Y**.

8 Select the variable min_pressure, and click **Set X**.

9 Click **OK**.

A scatter plot appears (Figure 2.18), and output from the REG procedure is added at the bottom of the output window.

Figure 2.18 Least Squares Regression



The output from the REG procedure indicates an R-square value of 0.8562 for the line of least squares given approximately by wind_kts$= 1222 - 1.177\times$min_pressure. The scatter plot shows this line and a 95% confidence band for the predicted mean. The confidence band is very thin, which indicates high confidence in the means of the predicted values.

References

Kimball, S. K. and Mulekar, M. S. (2004), "A 15-Year Climatology of North Atlantic Tropical Cyclones, Part I: Size Parameters," *Journal of Climatology*, 17, 3555–3575.

Mulekar, M. S. and Kimball, S. K. (2004), "The Statistics of Hurricanes," *STATS*, 39, 3–8.

Chapter 3
Creating and Editing Data

Contents

Overview of Creating and Entering Data **29**
Entering Data **29**
Example: Create a Small Data Set 30
Adding Variables 33
Adding and Editing Observations 35

Overview of Creating and Entering Data

The SAS/IML Studio data table displays data in a tabular view. You can create small data sets by entering data into the table. You can edit cells to examine “what-if” scenarios. You can add new variables or observations, and you can cut and paste between cells of the data table and the Microsoft Windows clipboard.

Entering Data

This section describes how you can use the data table to enter small data sets. You learn how to do the following:

- enter new variables
- enter or edit observations
- copy, cut, and paste to and from the Windows clipboard

Example: Create a Small Data Set

The following steps describe how to enter data into a data table. The data in this example are quarterly sales for two employees, June and Bob.

1 Create a new data set by selecting **File ►New ►Data Set** from the main menu.

The New Data Set dialog box appears so that you can create the first variable.

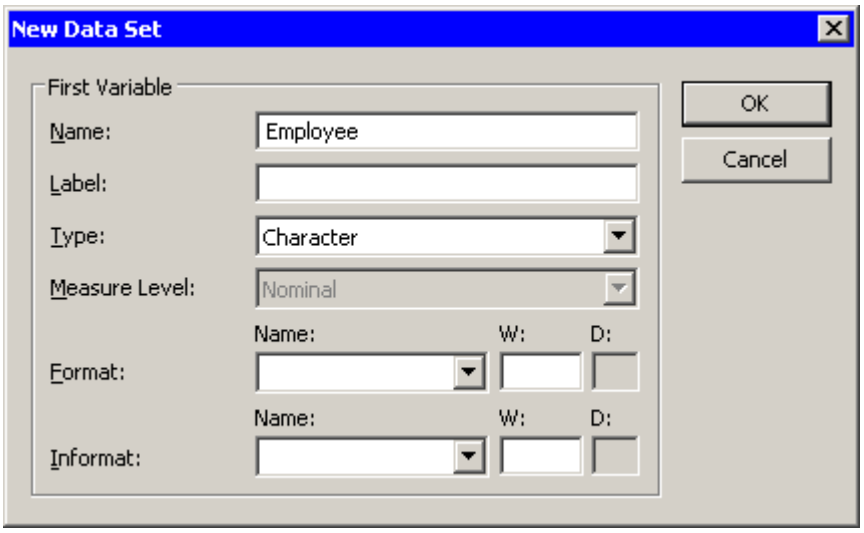
The first variable will contain the name of the sales staff, so you must specify a valid SAS variable name. Fill in the dialog box as follows (see Figure 3.1):

a In the **Name** field, type `Employee`.

b In the **Type** field, select **Character**.

c Click **OK**.

Figure 3.1 Creating a Character Variable



2 Create a new variable by selecting **Edit ►Variables ►New Variable** from the main menu.

The second variable will indicate the quarter of the financial year for which sales are recorded. Because the only valid values for this numeric variable are the discrete integers 1–4, you specify the measure level as nominal.

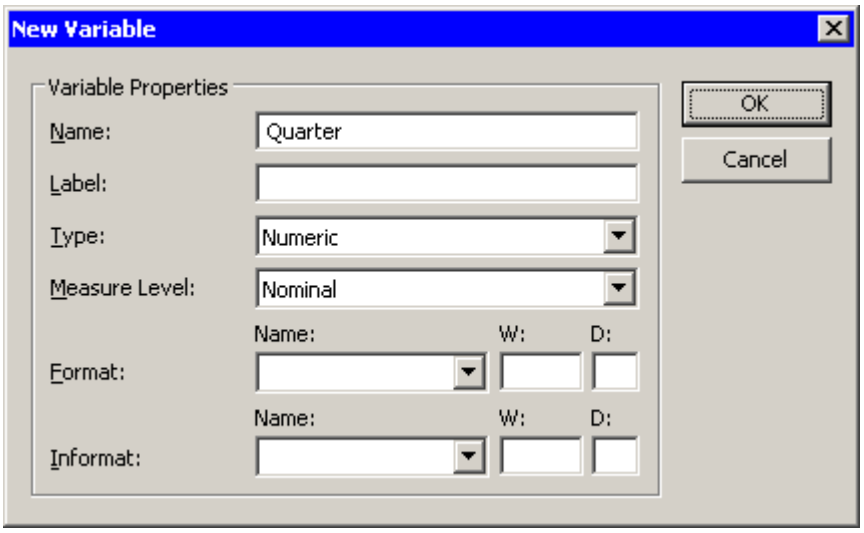
Fill in the dialog box as follows (see Figure 3.2):

a Type `Quarter` in the **Name** field.

b Select **Nominal** from the **Measure Level** menu.

c Click **OK**.

Figure 3.2 Creating a Nominal Numeric Variable



3 Create a third variable by selecting **Edit ►Variables ►New Variable** from the main menu.

The third variable will contain the revenue, in thousands of dollars, for each salesperson for each financial quarter.

Fill in the dialog box as follows (see Figure 3.3):

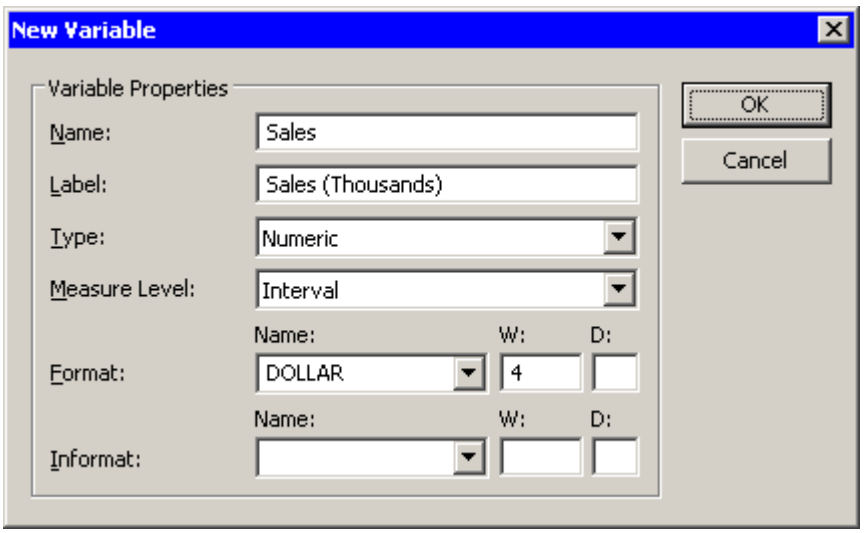
a Type `Sales` in the **Name** field.

b In the **Label** field, type `Sales (Thousands)`.

c In the **Format** list, select **DOLLAR**. Type `4` in the **W** field.

d Click **OK**.

Figure 3.3 Creating a Numeric Variable with a Format



4 Now you can enter the data shown in Table 3.1 as observations for each variable. Notice that the new data set was created with one observation that contains a missing value for each variable. (A missing values for a numerical variable is displayed as a dot.) Type the first observation in the first row.

When you enter data in the data table row marked with an asterisk (∗), a new row is created. When you are entering (or editing) data, the ENTER key takes you down to the next observation. The TAB key moves the active cell to the right, whereas holding down the SHIFT key and pressing TAB moves the active cell to the left. You can also use the keyboard arrow keys to navigate the cells of the data table.

Table 3.1 Sample Data

Employee	Quarter	Sales
June	1	34
Bob	1	29
June	2	24
Bob	2	18
June	3	28
Bob	3	25
June	4	45
Bob	4	32

NOTE: When you enter the data for the Sales variable, *do not* type the dollar sign. The actual data is $\{34, 29, \ldots, 32\}$, but because the variable has a DOLLAR4. format, the data table displays a dollar sign in each cell.

The data table looks like the table in Figure 3.4.

Figure 3.4 New Data Set

sales.sas7bdat

	3	Employee	Quarter	Sales
8		Nom	Nom	Int
1	χ²	June	1	$34
2	χ²	Bob	1	$29
3	χ²	June	2	$24
4	χ²	Bob	2	$18
5	χ²	June	3	$28
6	χ²	Bob	3	$25
7	χ²	June	4	$45
8	χ²	Bob	4	$32
*				

At this point you can save your data.

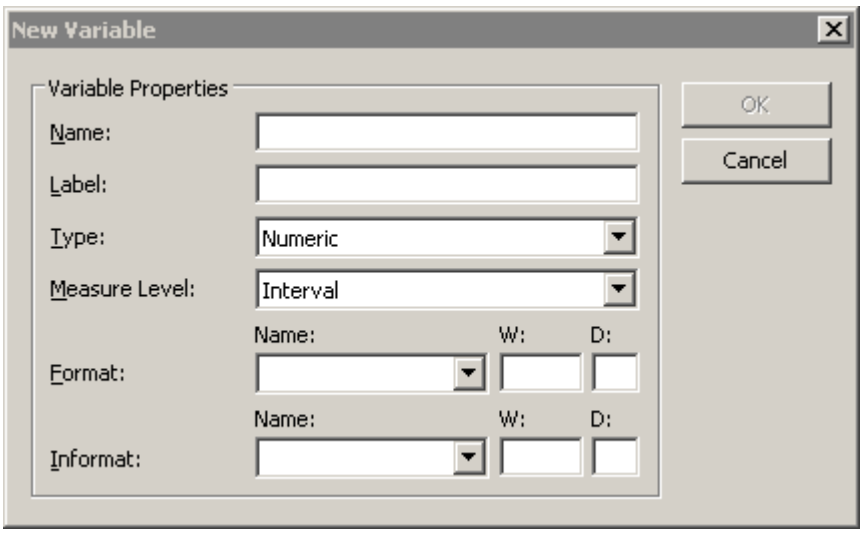
5 Select **File ►Save as File** from the main menu. Navigate to the `Data Sets` subdirectory of your personal files directory and save the file as *sales.sas7bdat*.

NOTE: The default location of the *personal files directory* is given in the "The Personal Files Directory" section in Chapter 34, "Configuring the SAS/IML Studio Interface." When you want to open your data later, you can select **File ►Open ►File** from the main menu. The dialog box that appears has a button near the bottom that says **Go to Personal Files directory**. For this reason, it is convenient to save data in your personal files directory.

Adding Variables

You can add a new variable by selecting **Edit ►Variables ►New Variable** from the main menu. Alternatively, you can right-click anywhere in the variable heading row. The New Variable dialog box appears. (See Figure 3.5.)

Figure 3.5 The New Variable Dialog Box



The New Variable dialog box enables you to define the variable properties. The following list describes each field in the dialog box.

Name

specifies the name of the new variable. This must be a valid SAS variable name. This means the name must satisfy the following conditions:

- must be at most 32 characters
- must begin with an English letter or underscore
- cannot contain blanks
- cannot contain special characters other than an underscore

Label

specifies the label for the variable.

Type

specifies the type of variable: numeric or character.

Measure Level

specifies the variable's *measure level*. The measure level determines the way a variable is used in graphs and analyses. A character variable is always nominal. For numeric variables, you can choose from two measure levels:

Interval The variable contains values that vary across a continuous range. For example, a variable that measures temperature would likely be an interval variable.

Nominal The variable contains a discrete set of values. For example, a variable that indicates gender would be a nominal variable.

Format
specifies the SAS format for the variable. For many formats you also need to specify values for the **W** (width) and **D** (decimal) fields that are associated with the format. For more information about formats, see the *SAS Language Reference: Dictionary*.

Informat
specifies the SAS informat for the variable. For many informats you also need to specify values for the **W** (width) and **D** (decimal) fields that are associated with the format. For more information about informats, see the *SAS Language Reference: Dictionary*.

NOTE: You can type the name of a format into the **Format** or **Informat** field, even if the name does not appear in the list.

Adding and Editing Observations

To add a new observation, type data into any cell in the last data table row. This row is marked with an asterisk (∗).

When you are entering (or editing) data, the ENTER key takes you down to the next observation. The TAB key moves the active cell to the right, whereas holding down the SHIFT key and pressing TAB moves the active cell to the left. You can also use the keyboard arrow keys to navigate the cells of the data table.

It is possible to perform operations on a range of cells. If you select a range of cells, then you can do the following:

- Delete the contents of the cells with the DELETE key.
- Cut or copy the contents of the range of cells to the Windows clipboard, in tab-delimited format. This makes the contents of the cells available to all Windows applications (Excel, Word, and so on).
- Paste from the Windows clipboard into the selected range of cells, provided that the data on the clipboard is in tab-delimited format. You can paste numeric data into cells in a character variable (the data are converted to text), but you cannot paste character data into cells in a numeric variable.

Typing in a cell changes the data for that cell. Graphs that use that observation will update to reflect the new data.

NOTE: If you change data after an analysis has been run, you need to rerun the analysis; the analysis does not automatically rerun to reflect the new data.

Chapter 4
Interacting with the Data Table

Contents

Overview of the Data Table **37**
Data Table Menus **37**
The Variables Menu **38**
The _OBSTAT_ Variable 42
Using the _OBSTAT_ Variable in SAS Procedures 42
Sorting Observations **43**
Selecting Observations **45**
The Observations Menu **46**
Changing Marker Properties 47
Changing Observation Labels 48
Including and Excluding Observations 49
Examining Data **50**
Finding Observations 50
Examining Selected Observations 54
Copying Selected Data 54
Saving Data **55**
Properties of Data Tables **56**
Keyboard Shortcuts in Data Tables **57**

Overview of the Data Table

The SAS/IML Studio data table displays data in a tabular view. You can use the data table to change properties of a variable, such as a variable's name, label, or format. You can also change properties of observations, including the shape and color of markers used to represent observations in graphs. You can also control which observations are visible in graphs and which are used in statistical analyses.

Data Table Menus

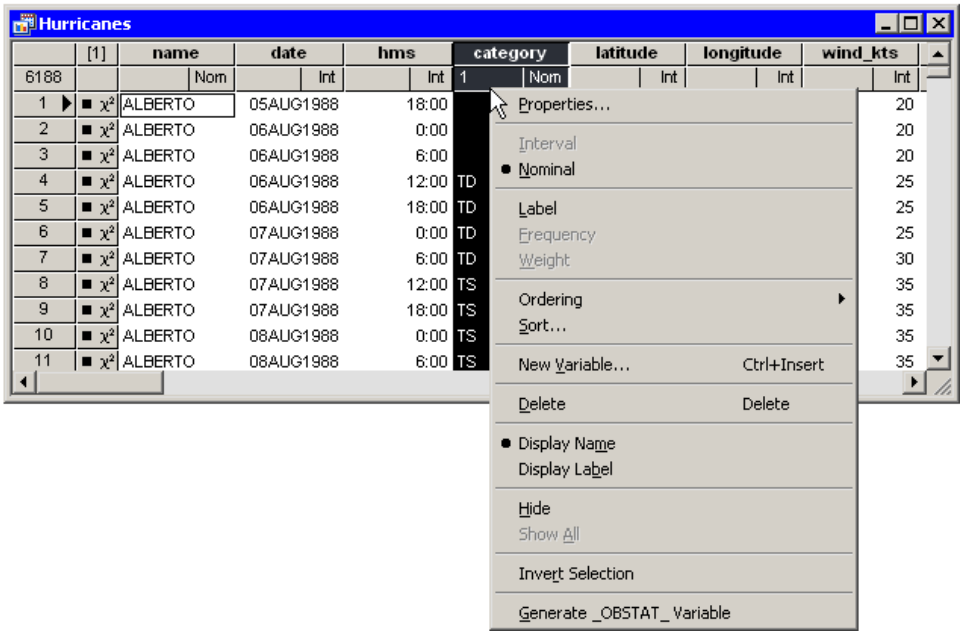
The first two rows of the data table are column headings (also called variable headings). The first row displays the variable's name or label. The second row indicates the variable's measure level (nominal or interval), the default role the variable plays, and, if the variable is selected, in what order it was selected. Subsequent rows contain observations.

The first two columns of the data table are row headings (also called observation headings). The first column displays the observation number (or some other label variable). The second column indicates whether the observation is included in plots and analyses.

The effect of selecting a cell of the data table depends on the location of the cell. To select a variable, click the column heading. To select an observation, click the row heading.

You can display a context menu as in Figure 4.1 by right-clicking a column heading or row heading. A context menu means that you see different menus depending on where the mouse pointer is when you right-click. For the data table, the **Variables** menu differs from the **Observations** menu.

Figure 4.1 Data Table with the Variables Menu



The Variables Menu

You can access the **Variables** menu (shown in Figure 4.2) by clicking a column heading and selecting **Edit ►Variables** from the main menu. Alternatively, right-clicking a variable heading (see Figure 4.1) selects that variable and displays the same menu.

You can use the **Variables** menu to do the following:

- change properties of existing variables
- create a new variable
- change the set of variables that are displayed in the data table
- change the set of selected and unselected variables
- set the *role* of an existing variable. You can assign three default roles:

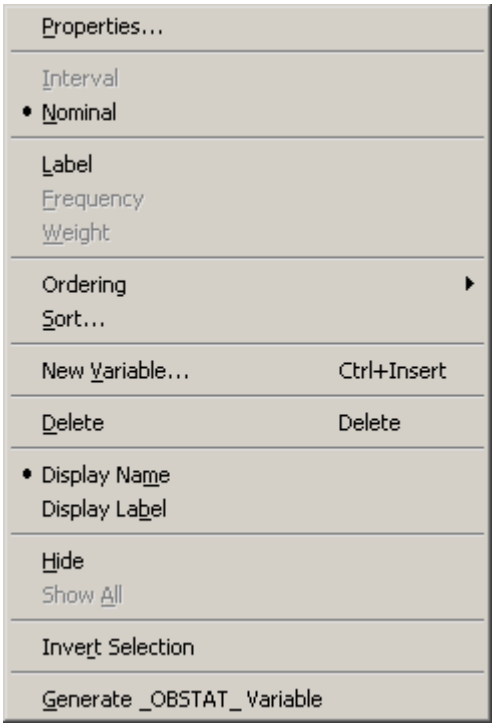
Label The values of the variable are used to label the markers in a plot. Only the markers that you have clicked are labeled.

Frequency The values of the variable are used as the frequency of occurrence for each observation. If you assign a variable to a Frequency role, then that variable is automatically added to dialog boxes for analyses and graphs that support a frequency variable.

Weight The values of the variable are used as weights for each observation. If you assign a variable to a Weight role, then that variable is automatically added to dialog boxes for analyses and graphs that support a weight variable.

All roles are optional; you do not need to specify any roles. A variable can play multiple roles, but there can be at most one variable assigned to each role.

Figure 4.2 The Variables Menu



The following list describes each item on the Variables menu.

Properties
displays the Variable Properties dialog box, described in the section "Adding Variables" on page 33 in Chapter 3, "Creating and Editing Data." The dialog box enables you to change most properties for the selected variable. However, you cannot change the type (character or numeric) of an existing variable.

Interval/Nominal
changes the measure level of the selected numeric variable. A character variable cannot be interval.

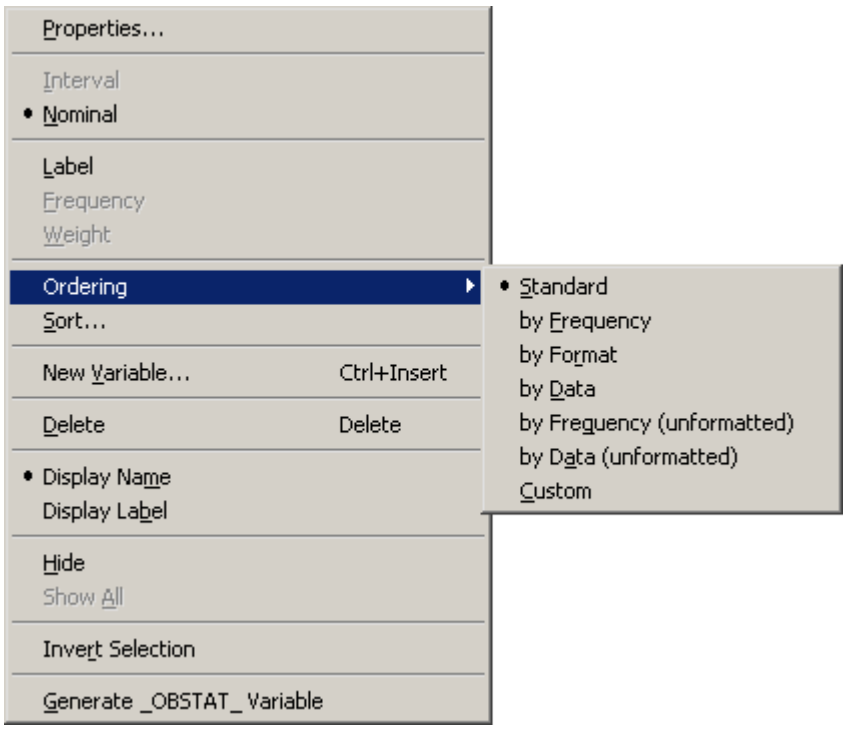
Label
makes the selected variable the label variable for plots. Only one variable can have this role.

Frequency
makes the selected variable the frequency variable for analyses and plots that support a frequency variable. Only a numeric variable can have a Frequency role.

Weight
makes the selected variable the weight variable for analyses and plots that support a weight variable. Only a numeric variable can have a Weight role.

Ordering
specifies how nominal variables are ordered. This affects the way that a variable is sorted and the order of categories in plots. If a variable has missing values, they are always ordered first. See the "Ordering Categories of a Nominal Variable" section in Chapter 11 for further details. The **Ordering** submenu is shown in Figure 4.3.

Figure 4.3 The Ordering Menu



You can order a variable in the following ways:

Standard specifies that categories be arranged in ASCII order by their unformatted values. In ASCII order, numerals precede uppercase letters, which precede lowercase letters.

by Frequency specifies that categories be arranged according to the descending frequency count of formatted values in each category.

by Format specifies that categories be arranged in ASCII order by their formatted values.

by Data specifies that categories be arranged according to the *data order* of formatted values. The data order is determined by traversing the values of a variable, starting from the first observation. The first (nonmissing) value you encounter is ordered first, the next unique (nonmissing) value of the variable is ordered second, and so on. Sorting the data table does not affect this ordering; the ordering is based on the original sequence of observations.

by Frequency (unformatted) specifies that categories be arranged according to the descending frequency count of unformatted values in each category.

by Data (unformatted) specifies that categories be arranged according to the data order of unformatted values. Sorting the data table does not affect this ordering; the ordering is based on the original sequence of observations.

Custom specifies that this variable be ordered by calling the DataObject.SetVarValueOrder method. See the SAS/IML Studio online Help for details about this method.

Sort

displays the Sort dialog box. The Sort dialog box is described in the section "Sorting Observations" on page 43.

New Variable

displays the New Variable dialog box to create a new variable as described in the section "Adding Variables" on page 33 in Chapter 3, "Creating and Editing Data." (See Figure 3.5.)

Delete

deletes the selected variables.

Display Name/Display Label

toggles whether the column heading displays the names of variables or displays their labels.

Hide

hides the selected variables. The variables can be displayed at a later time by selecting **Show All**. Hidden variables cannot be selected.

Show All

displays all variables, including variables that were hidden.

Invert Selection

changes the set of selected variables. Deselected variables become selected, and selected variables become deselected.

Generate _OBSTAT_ Variable

creates a new character variable called _OBSTAT_ that encodes the current state of each observation. The values of the _OBSTAT_ variable are described in the next section.

The _OBSTAT_ Variable

The _OBSTAT_ variable is a character variable of length 20. It was introduced in SAS/INSIGHT software as a way to capture the state of observations, including the color and shape of markers and whether an observation is selected. The first few characters encode the state of binary options such as whether an observation is selected. A character is '1' if the corresponding property is true and '0' if the related property is false. The properties are described in the following list:

Character 1 stores whether the observation is selected.

Character 2 stores whether the observation is included in plots.

Character 3 stores whether the observation is included in analyses.

Character 4 stores whether the observation has a label.

Character 5 stores the marker shape for an observation. This is a value between 1 and 8 that corresponds to a shape, as given in the following table:

Value	Shape
1	□
2	+
3	○
4	◇
5	×
6	△
7	▽
8	★

Characters 6–20 store the RGB value of the fill color for an observation marker. The RGB color model represents colors as combinations of the colors red, green, and blue.

Each component is a five-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.

If you read a data set for which there is no associated DMM file and if that data set contains a variable named _OBSTAT_, then the state of each observation is determined by the corresponding value of the _OBSTAT_ variable.

If an _OBSTAT_ variable already exists when you select **Generate _OBSTAT_ Variable** from the variable menu, then the values of the variable are updated with the current state of the observations.

Using the _OBSTAT_ Variable in SAS Procedures

The _OBSTAT_ variable is often used in conjunction with a SAS procedure to analyze observations that satisfy certain criteria. For example, you might want to perform a linear regression only on observations that have the Include in Analysis property. Or you might want to compute a correlation matrix only for observations that are represented by a square marker shape.

The _OBSTAT_ variable contains information about the state of observations in SAS/IML Studio. It is often convenient to use the DATA step to split the single _OBSTAT_ variable into several indicator variables so that it is easier to use a WHERE clause to choose only observations that have a desired property.

To use the _OBSTAT_ variable to select observations for analysis by a SAS procedure:

1. Create an _OBSTAT_ variable by selecting **Generate _OBSTAT_ Variable** from the variable menu.
2. Save the augmented data to a SAS data set such as SASUSER.MyData.
3. Use the following DATA step to extract each observation property into its own variable:

```
/* Create numerical variables from an _OBSTAT_ variable. */
data MyData;
set sasuser.MyData;
ObsIsSelected   = inputn(substr(_obstat_,  1, 1), 1.);
ObsIsInPlots    = inputn(substr(_obstat_,  2, 1), 1.);
ObsIsInAnalysis = inputn(substr(_obstat_,  3, 1), 1.);
ObsIsLabeled    = inputn(substr(_obstat_,  4, 1), 1.);
ObsMarkerShape  = inputn(substr(_obstat_,  5, 1), 1.);
ObsMarkerRed    = inputn(substr(_obstat_,  6, 5), 5.);
ObsMarkerGreen  = inputn(substr(_obstat_, 11, 5), 5.);
ObsMarkerBlue   = inputn(substr(_obstat_, 16, 5), 5.);
run;
```

4. Use a WHERE clause to analyze only observations with a given set of properties. For example, the following statements compute a correlation matrix for observations that are represented in SAS/IML Studio by a marker shape:

```
data Subset;
set MyData(where=(ObsMarkerShape=1);
run;

proc corr data=Subset(drop=Obs:);
run;
```

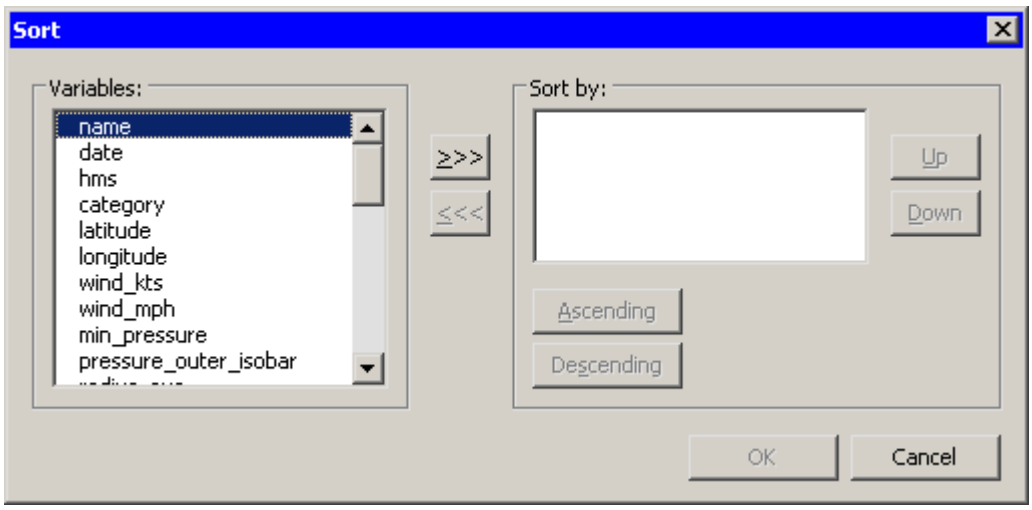
Sorting Observations

This section describes how to sort a data table by one or more variables.

To open the Sort dialog box, you can select **Edit ►Variables ►Sort** from the main menu. Alternatively, you can right-click a variable heading to display the **Variables** menu (shown in Figure 4.2), and then select **Sort**. The Sort dialog box is shown in Figure 4.4.

The first time the Sort dialog box is created, any variables that are selected are automatically placed in the **Sort by** list. Subsequently, the Sort dialog box remembers the **Sort by** list from the last sort.

Figure 4.4 The Sort Dialog Box



The following list describes each item in the Sort dialog box.

Variables
lists the variables in the data set that are not yet in the **Sort by** list. Select variables in this list to transfer them to the **Sort by** list.

⋙
transfers the selected variables from the **Variables** list to the **Sort by** list.

⋘
removes selected variables from the **Sort by** list.

Sort by
lists the variables to sort by.

Up
moves a selected variable up one space in the **Sort by** list.

Down
moves a selected variable down one space in the **Sort by** list.

Ascending
marks the selected variables in the **Sort by** list to be sorted in ascending order.

Descending
marks the selected variables in the **Sort by** list to be sorted in descending order.

To carry out the sort operation, click **OK**.

As described in the section "The Variables Menu" on page 38, a nominal variable can be ordered in different ways. If a variable has an ordering different from the standard ordering, then the sort dialog box indicates that fact by marking the variable name with an asterisk.

Selecting Observations

You can select observations in a data table by clicking the row heading on the left side of the data table. You can drag down or up to select contiguous observations. You can click while holding down the CTRL key to select new observations without losing the ones already selected. Figure 4.5 shows selected observations.

NOTE: Highlighting a range of cells in the data table does not select the observations. The section "Adding and Editing Observations" on page 35 in Chapter 3, "Creating and Editing Data," lists operations that you can perform on a range of cells.

Figure 4.5 Selected Observations

Hurricanes

	34	name	date	hms	category
[9]		Nom	Int	Int	Nom
1	■ χ^2	ALBERTO	05AUG1988	18:00	
2	■ χ^2	ALBERTO	06AUG1988	0:00	
3	■ χ^2	ALBERTO	06AUG1988	6:00	
4	■ χ^2	ALBERTO	06AUG1988	12:00	TD
5	■ χ^2	ALBERTO	06AUG1988	18:00	TD
6	■ χ^2	ALBERTO	07AUG1988	0:00	TD
7	■ χ^2	ALBERTO	07AUG1988	6:00	TD
8	■ χ^2	ALBERTO	07AUG1988	12:00	TS
9	■ χ^2	ALBERTO	07AUG1988	18:00	TS
10	■ χ^2	ALBERTO	08AUG1988	0:00	TS
11	■ χ^2	ALBERTO	08AUG1988	6:00	TS
12	■ χ^2	ALBERTO	08AUG1988	12:00	TD
13	■ χ^2	ALBERTO	08AUG1988	18:00	TD

The four cells in the upper left corner of the data table are different from the other row headings, as described in the following list:

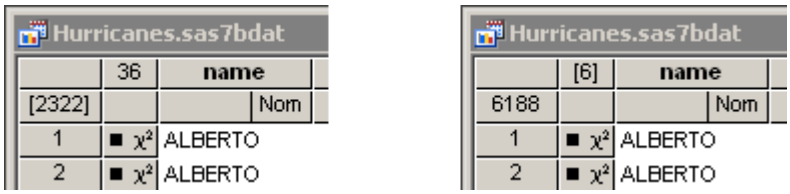
- Right-click in any of the four cells to display the **Observations** menu. The **Observations** menu is described in the section "The Observations Menu" on page 46. Consequently, this is a safe place to right-click when you want to change properties of the selected observations, but no selected observations are currently visible.
- Click in the upper left or lower right cell to deselect all observations and variables.
- Click in the upper right cell to deselect all observations and select all variables.
- Click in the lower left cell to deselect all variables and select all observations.

If no observations are selected, the lower left cell displays the total number of observations in the data table. If observations are selected, the lower left cell displays (in brackets) the number of selected observations.

If no variables are selected, the upper right cell displays the total number of variables in the data table. If variables are selected, the upper right cell displays (in brackets) the number of selected variables.

Figure 4.6 illustrates two possibilities. The left portion of the figure indicates a data table that has 2,322 selected observations; none of the 36 variables are selected. The right portion of the figure indicates that 6 variables are selected, but none of the 6,188 observations are selected.

Figure 4.6 Indicating Selected Observations (Left) and Variables (Right)

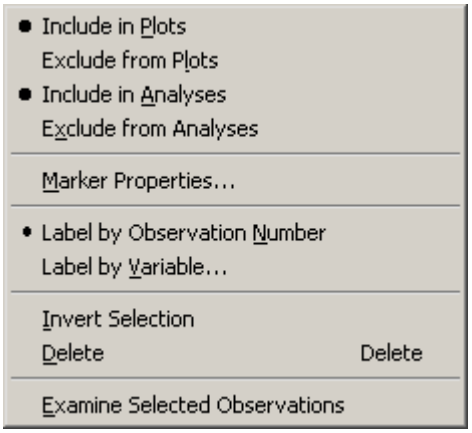


The Observations Menu

The row heading on the left side of the data table gives the status of each observation. The heading indicates whether an observation is selected, which shape and color is used to represent the observation in plots, and whether the observation is included in analyses.

You can change the properties of selected observations by using the **Observations** menu. You can access the **Observations** menu by selecting **Edit ►Observations** from the main menu. Alternatively, right-clicking the row heading of a selected observation displays the same **Observations** menu, shown in Figure 4.7.

Figure 4.7 The Observations Menu



The following list describes each item on the **Observations** menu.

Include in Plots
includes the selected observations in graphs.

Exclude from Plots
excludes the selected observations from graphs.

Include in Analyses
includes the selected observations in statistical analyses.

Exclude from Analyses
excludes the selected observations from statistical analyses.

Marker Properties
displays the Marker Properties dialog box. The Marker Properties dialog box is described in section "Changing Marker Properties" on page 47.

Label by Observation Number
sets the label that is displayed in the left-most column of the data table to be the observation number. The observation number is also set as the default label that is displayed when you click an observation marker in a graph.

Label by Variable
displays the Label by Variable dialog box. The Label by Variable dialog box is described in section "Changing Observation Labels" on page 48.

Invert Selection
changes the set of selected observations. Deselected observations become selected, and selected observations become deselected.

Delete
deletes the selected observations.

Examine Selected Observations
displays the Examine Selected Observations dialog box. You can use this dialog box to view and compare the selected observations. The Examine Selected Observations dialog box is described in section "Examining Selected Observations" on page 54.

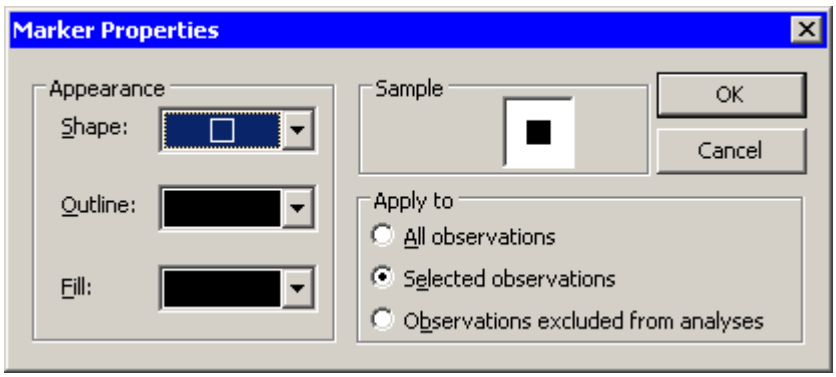
Changing Marker Properties

You can change the markers used to represent observations. You can use marker shapes and colors to represent observations that share common properties.

Marker shapes are often used to discriminate observations with different values of a categorical variable (for example, male versus female). Marker colors can also be used for this purpose, or they can represent a continuous variable. Chapter 9, "General Plot Properties," describes coloring markers by a continuous variable.

Select **Edit ►Observations ►Marker Properties** from the main menu to open the Marker Properties dialog box. (See Figure 4.8.)

Figure 4.8 The Marker Properties Dialog Box



The Marker Properties dialog box contains the following UI controls:

Shape
sets the marker shape for the observations.

Outline
sets the marker outline color for the observations.

Fill
sets the marker fill color for the observations.

Sample
shows what the marker with the specified shape and colors looks like.

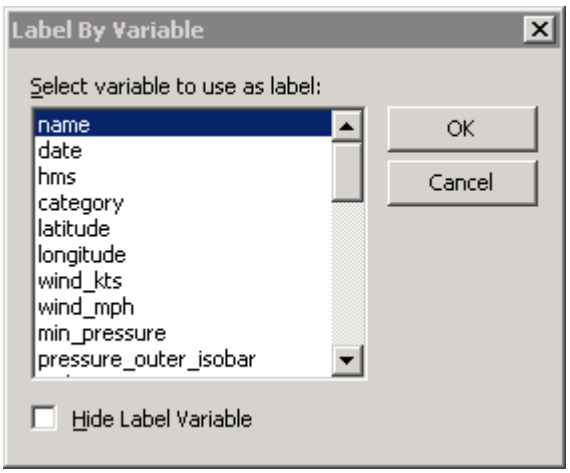
Apply to
specifies the set of observations whose markers will change. By default, changes are applied to only the selected observations.

Changing Observation Labels

You can change the label displayed in the left-most column of the data table. Observation numbers are shown by default.

You can select **Edit ►Observations ►Label by Variable** from the main menu to open the Label by Variable dialog box. (See Figure 4.9.) You can use this dialog box to select the variable whose values are displayed in the left-most column of the data table. The variable is also set as the default label that is displayed when you click an observation marker in a graph.

Figure 4.9 The Label by Variable Dialog Box



The **Hide Label Variable** check box hides the label variable. This is especially useful if the label variable is one of the first variables in the data table.

Including and Excluding Observations

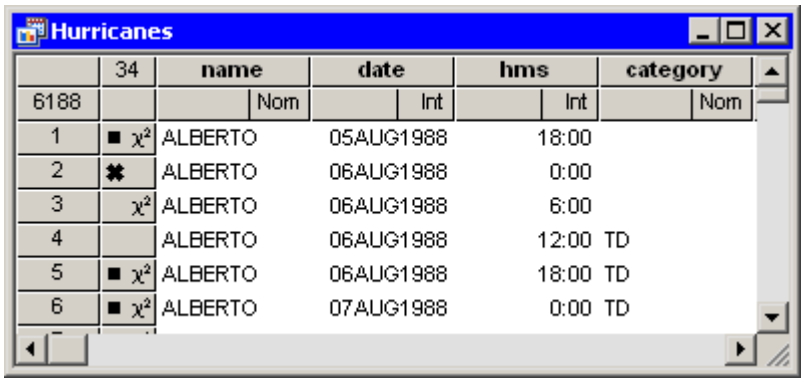
You can choose which observations appear in plots and which are used in analyses.

To include or exclude observations, first select the observations. From the **Edit ►Observations** menu, you can then select **Include in Plots**, **Exclude from Plots**, **Include in Analyses**, or **Exclude from Analyses**.

The row heading of the data table shows the status of an observation in analyses and plots. A marker symbol indicates that the observation is included in plots; observations excluded from plots do not have a marker symbol shown in the data table. Similarly, the χ^2 symbol is present if and only if the observation is included in analyses. If an observation is excluded from analyses but included in plots, then the marker symbol changes to the $\times$ symbol.

For example, Figure 4.10 shows what the data table would look like if you excluded some observations. In this example, the second observation is included in plots but excluded from analyses. The third observation is excluded from plots but included in analyses. The fourth observation is excluded from both plots and analyses.

Figure 4.10 Excluded Observations



Examining Data

This section describes how to do the following:

- find observations that satisfy certain conditions
- examine selected observations
- copy selected observations into a separate data set

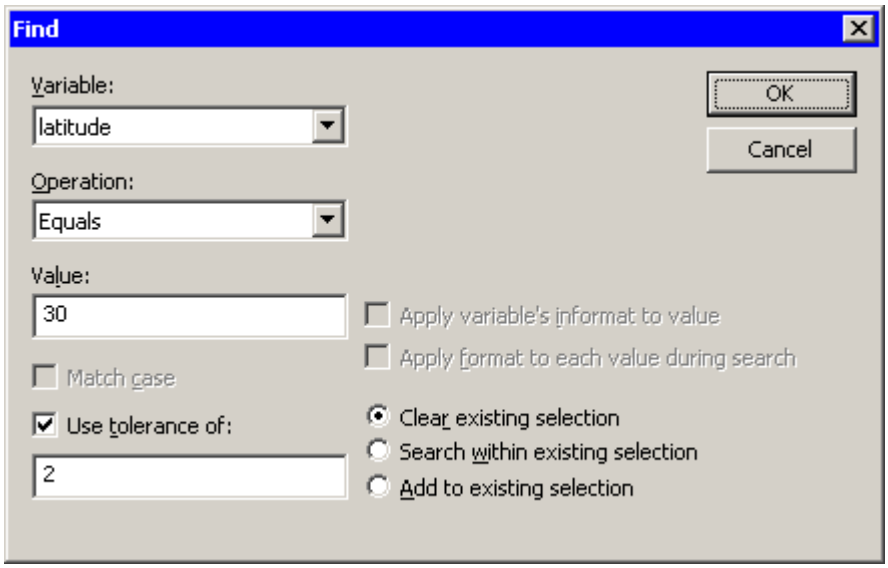
In analyzing data, you might want to find observations that satisfy certain conditions. For example, you might want to select all sales to a particular company. Or you might want to select all patients with high blood pressure.

After you have found the observations, you can examine the observations or copy them to a new data set.

Finding Observations

You can select observations in the data table by using the Find dialog box. (For a way to graphically and interactively select observations that satisfy multiple constraints, see Chapter 11, "Techniques for Exploring Data.") You can open the Find dialog box (shown in Figure 4.11) by selecting **Edit ►Find** from the main menu.

Figure 4.11 The Find Dialog Box



The Find dialog box contains the following UI controls:

Variable
chooses the variable whose values are examined. The list includes each variable in the data set.

Operation
selects the logical operation used to compare each observation with the contents of the **Value** field.

Value
specifies the value used to select observations.

Apply variable's informat to value
applies the variable's informat to the contents of the **Value** field. If the variable does not have an informat, then this item is inactive.

Apply format to each value during search
applies the variable's format to the variable and then compares the formatted data to the contents of the **Value** field. If the variable does not have a format, then this item is inactive.

Match case
specifies that each observation be compared to the contents of the **Value** field in a case-sensitive manner. If the variable is numeric, then this item is inactive.

Use tolerance of
specifies that a tolerance, ϵ, be used in comparing each observation to the contents of the **Value** field. Table 4.1 specifies how ϵ is used. If the chosen variable is a character variable, then this item is inactive.

Clear existing selection
specifies that all observations be searched, but only the observations that match the search criterion be selected.

Search within existing selection
specifies that only the observations that are selected be searched. You can use this option to perform logical AND operations.

Add to existing selection
specifies that all observations be searched, but observations that were selected prior to the search remain selected. You can use this option to perform logical OR operations.

For numeric variables, let v be the value of the **Value** field and let ϵ be the value of the **Use tolerance of** field. (If you are not using a tolerance, then $\epsilon = 0$.) Table 4.1 specifies whether an observation with value x for the chosen variable matches the query.

Table 4.1 Find Operations for Numeric Variables

Operation	Values Found	Missing Selected?
Equals	$x \in [v - \epsilon, v + \epsilon]$	No
Less than	$x < v + \epsilon$	Yes
Greater than	$x > v - \epsilon$	No
Not equals	$x \notin [v - \epsilon, v + \epsilon]$	Yes
Less than or equals	$x \leq v + \epsilon$	Yes
Greater than or equals	$x \geq v - \epsilon$	No
Is missing	x is missing	Yes

To remember whether missing values match the query, recall that SAS missing values are represented as large negative numbers. Table 4.1 is consistent with the WHERE clause in the SAS DATA step.

For character variables, comparisons are performed according to the ASCII order of characters. In particular, all uppercase letters [A–Z] precede lowercase characters [a–z]. Let v be the value of the **Value** field and let $v \prec x$ indicate that v precedes x in ASCII order. Table 4.2 specifies whether an observation with value x for the chosen variable matches the query.

Table 4.2 Find Operations for Character Variables

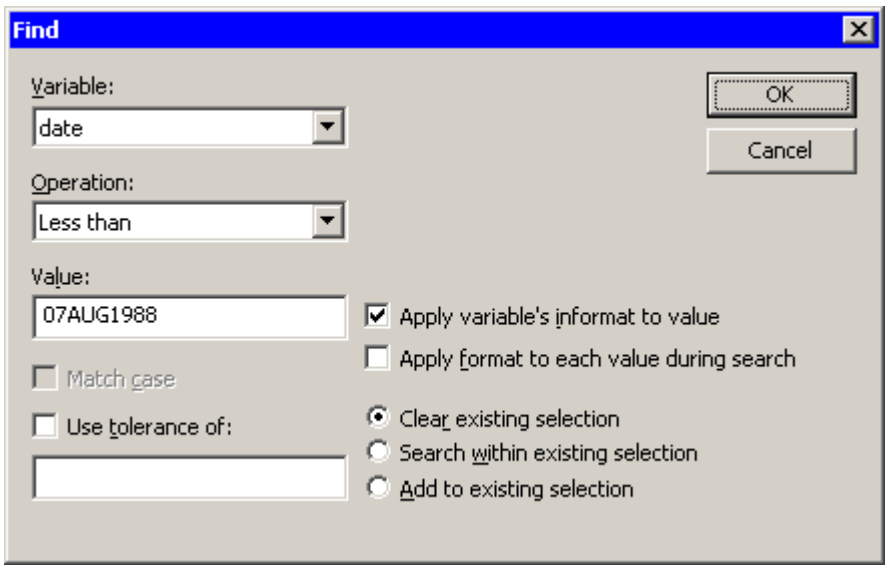
Operation	Values Found	Missing Selected?
Equals	$x = v$	No
Less than	$x \prec v$	Yes
Greater than	$v \prec x$	No
Not equals	$x \neq v$	Yes
Less than or equals	$x \preceq v$	Yes
Greater than or equals	$v \preceq x$	No
Is missing	x is missing	Yes
Contains	x contains v	No
Does not contain	x does not contain v	Yes
Begins with	x begins with v	No

To help remember whether character missing values match the query, think of the character missing value as being a zero-length string that contain no characters. Table 4.2 is consistent with the WHERE clause in the SAS DATA step.

As a first example, Figure 4.11 shows how to find observations in the Hurricanes data set whose latitude variable is contained in the interval [28, 32]. This is a quick way to find observations with latitudes between 28 and 32 in a single search.

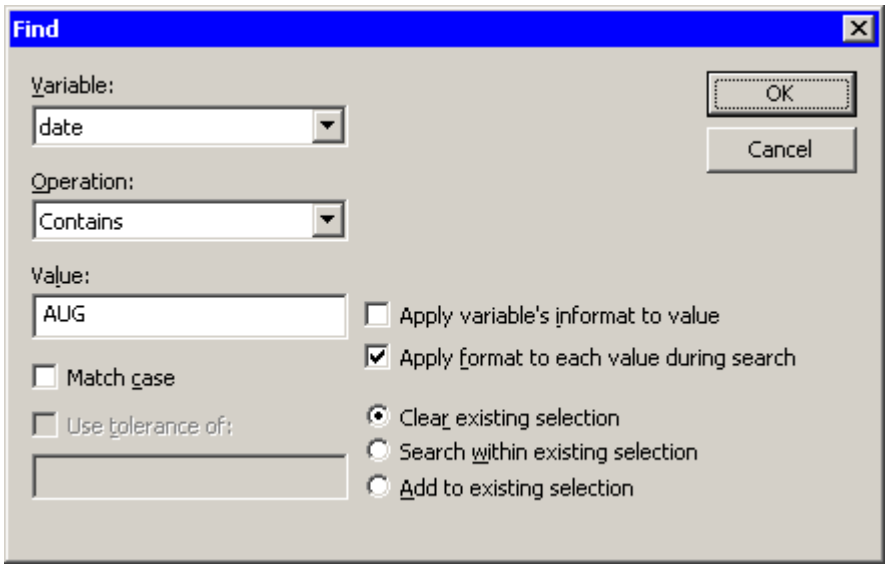
A second example is shown in Figure 4.12. This search finds observations for which the date variable strictly precedes 07AUG1988. The date variable has a DATE9. informat, so you can use that informat to make it more convenient to input the contents of the **Value** field. (Without the informat, you would need to search for the value 10445, the SAS date value that corresponds to 06AUG1988.) Recall that the date variable is a numeric variable, even though the formatted values appear as text.

Figure 4.12 Searching for Dates



A related example is shown in Figure 4.13. This search finds all observations for which the date variable contains the text "AUG". To perform this search you must check **Apply format to each value during search**. This forces the Find dialog box to apply the DATE9. format to the date variable, which means comparing strings (character data) instead of numbers (numeric data). You can then select **Contains** from the **Operation** list. Each formatted string is searched for the value "AUG".

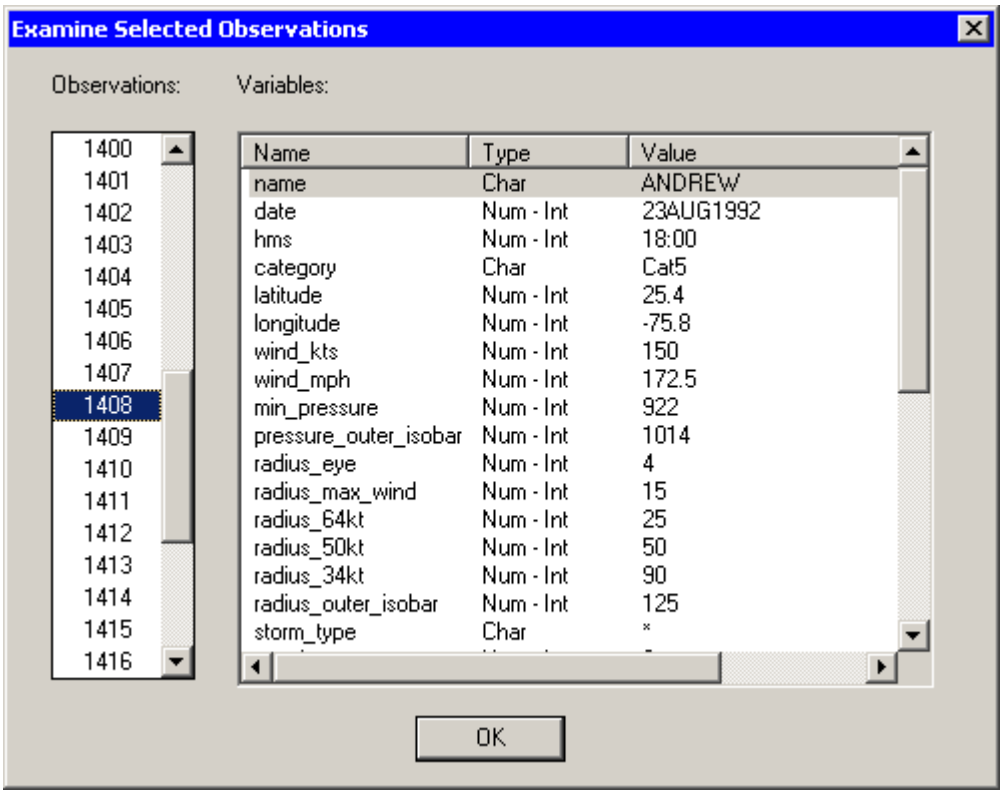
Figure 4.13 Matching Text in a Formatted Variable



Examining Selected Observations

You can examine the values of selected observations. To do this, select **Edit ►Observations ►Examine Selected Observations** from the main menu. Figure 4.14 shows the dialog box that appears. By clicking observation numbers in the list on the left (or by using the UP and DOWN arrow keys), you can examine each selected observation in turn.

Figure 4.14 Examining Selected Observations

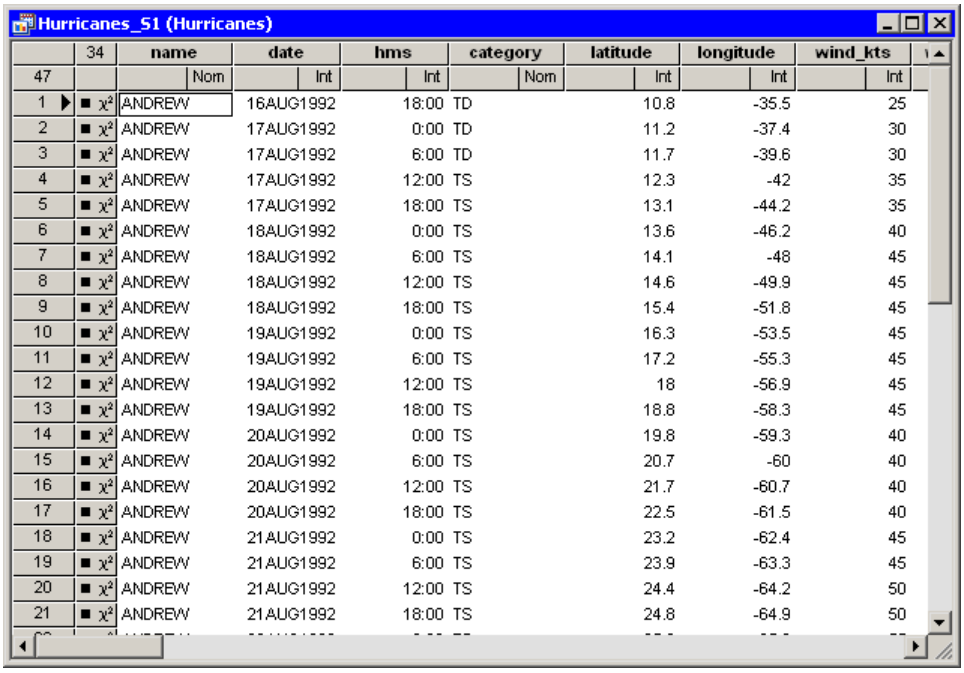


Copying Selected Data

You can subset your data by copying selected observations or variables to a separate data set. (You can select variables without losing selected observations by holding down the CTRL key while you click.) You can then analyze or save this new data set.

If no variables are selected, all variables are copied. If no observations are selected, all observations are copied. After you have selected observations or variables or both, select **File ►New ►Data Set from Selected Data** from the main menu. A new data table (Figure 4.15) appears, which contains only the selected subset of the original data.

Figure 4.15 Copying Selected Data



Hurricanes_51 (Hurricanes)

	34	name	date	hms	category	latitude	longitude	wind_kts
47		Nom	Int	Int	Nom	Int	Int	Int
1	ANDREW		16AUG1992	18:00	TD	10.8	-35.5	25
2	ANDREW		17AUG1992	0:00	TD	11.2	-37.4	30
3	ANDREW		17AUG1992	6:00	TD	11.7	-39.6	30
4	ANDREW		17AUG1992	12:00	TS	12.3	-42	35
5	ANDREW		17AUG1992	18:00	TS	13.1	-44.2	35
6	ANDREW		18AUG1992	0:00	TS	13.6	-46.2	40
7	ANDREW		18AUG1992	6:00	TS	14.1	-48	45
8	ANDREW		18AUG1992	12:00	TS	14.6	-49.9	45
9	ANDREW		18AUG1992	18:00	TS	15.4	-51.8	45
10	ANDREW		19AUG1992	0:00	TS	16.3	-53.5	45
11	ANDREW		19AUG1992	6:00	TS	17.2	-55.3	45
12	ANDREW		19AUG1992	12:00	TS	18	-56.9	45
13	ANDREW		19AUG1992	18:00	TS	18.8	-58.3	45
14	ANDREW		20AUG1992	0:00	TS	19.8	-59.3	40
15	ANDREW		20AUG1992	6:00	TS	20.7	-60	40
16	ANDREW		20AUG1992	12:00	TS	21.7	-60.7	40
17	ANDREW		20AUG1992	18:00	TS	22.5	-61.5	40
18	ANDREW		21AUG1992	0:00	TS	23.2	-62.4	45
19	ANDREW		21AUG1992	6:00	TS	23.9	-63.3	45
20	ANDREW		21AUG1992	12:00	TS	24.4	-64.2	50
21	ANDREW		21AUG1992	18:00	TS	24.8	-64.9	50

Saving Data

If you save data after changing variable or observation properties, then the changes are saved as well. Most variable properties (for example, formats) are saved with the SAS data set, whereas observation properties (for example, marker shapes) are saved in a separate *metadata file*. The metadata file is stored on the client PC and has the same name as the data set, but with a *dmm* extension.

For example, if you save a data set named MyData to your PC, then a file named *MyData.dmm* is also created in the same Windows folder as the *MyData.sas7bdat* file.

If you have changed the data and try to exit SAS/IML Studio, you are prompted to save the data set if you have done any of the following actions:

- edited cells in the data table
- changed a variable's properties (name, label, format, informat)
- changed a variable's measure level (nominal, interval)
- sorted a data set
- added or deleted a variable
- included or excluded observations
- changed an observation's marker properties (shape, color)
- added or deleted an observation

Properties of Data Tables

When a data table is the active window, you can do the following:

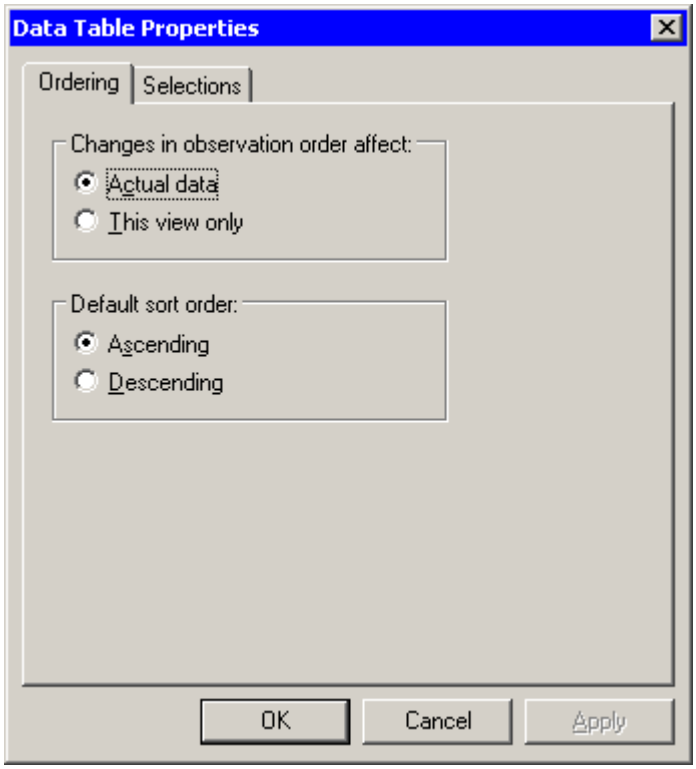
- create additional copies of the data table
- change the default properties of data tables in the current workspace

You can select **Windows ►New Window** from the main menu to create a copy of the current data table. (The new table might appear on top of the existing data table, so drag it to a new location if necessary.) This second data table can be scrolled independently from the first. This is useful, for example, if you are interested in examining several variables or observations whose positions in the data table vary widely. You can examine different subsets of the data simultaneously by using two or more tabular views of the same data.

By default, if you sort one data table, then other data tables that view the same data are also sorted in the same order. This is because a sort typically changes the order of the underlying data. (As mentioned in the section "Saving Data" on page 55, when you exit SAS/IML Studio you are prompted to save the data if you have sorted it.) However, there might be instances when it is useful to view the same data, but sorted in a different order. To accomplish this, you can *locally sort* a data table.

To locally sort a data table, select **Edit ►Properties** from the main menu, which displays the dialog box shown in Figure 4.16.

Figure 4.16 Data Table Ordering Properties



The **Ordering** tab contains the following UI controls:

Changes in observation order affect

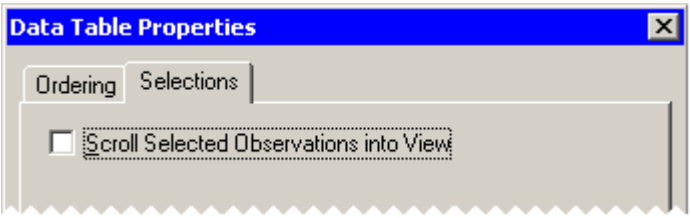
gives you two choices. If you select **Actual data**, then sorting the data table results in a global sort that reorders the observation in all views of the data. If you select **This view only**, then sorting the data table results in a local sort that does not reorder the observations but only changes the view of the data in the current data table.

Default sort order

gives you two choices. Your selection of **Ascending** or **Descending** determines the default order in which variables are sorted.

The **Selections** tab has a single item, as shown in Figure 4.17. If you select **Scroll selected observations into view**, then the data table automatically scrolls to a selected observation each time an observation is selected. To manually scroll a selected item into view, use the F3 key.

Figure 4.17 Data Table Selection Properties



Keyboard Shortcuts in Data Tables

When a data table is active, some keys are associated with certain actions, as shown in Table 4.3.

Table 4.3 Keys and Actions in Data Tables

Key	Action
ESC	When editing data, aborts the current edit and deselect cells.
ESC	Deselects any selected observations and variables.
F1	Displays the online Help system.
F3	Moves the active cell to the row of the next selected observation.
SHIFT+F3	Moves the active cell to the row of the previous selected observation.
F10	If observations are selected, displays the **Observations** menu. If variables are selected, displays the **Variables** menu. If observations and variables are selected, displays the **Observations** menu followed by the **Variables** menu.
TAB	Moves the active cell to the right.
SHIFT+TAB	Moves the active cell to the left.
ENTER	Moves the active cell down one row.
ALT+RIGHT ARROW ALT+LEFT ARROW	Toggles selection of a variable without changing the active cell.
ALT+DOWN ARROW ALT+UP ARROW	Toggles selection of an observation without changing the active cell.

Table 4.3 *continued*

Key	Action
SHIFT+ALT+RIGHT ARROW SHIFT+ALT+LEFT ARROW	Toggles selection of a variable and moves the active cell to the next or previous variable.
SHIFT+ALT+DOWN ARROW SHIFT+ALT+UP ARROW	Toggles selection of an observation and moves the active cell to the next or previous observation.
SHIFT+RIGHT ARROW SHIFT+LEFT ARROW	Extends the selection of a range of cell columns.
SHIFT+DOWN ARROW SHIFT+UP ARROW	Extends the selection of a range of cell rows.
HOME	Edits the active cell and places the cursor at the beginning of the cell.
END	Edits the active cell and places the cursor at the end of the cell.
CTRL+SPACEBAR	Clears selected observations and variables.
CTRL+HOME	Sets the active cell to the first row and first column.
CTRL+END	Sets the active cell to the last row and last column.
CTRL+INSERT	Displays the New Variable dialog box.
DELETE	If observations or variables are selected, deletes the selected variables or observations. If cells are selected, deletes the contents of the selected cells.

In addition, the data table supports the arrow keys for navigating cells, and it supports the standard Microsoft control sequences shown in Table 4.4.

Table 4.4 Standard Control Sequences in Data Tables

Key	Action
CTRL+A	Selects all observations.
CTRL+C	Copies contents of selected cells to Windows clipboard.
CTRL+F	Displays the Find dialog box.
CTRL+P	Prints the data table.
CTRL+V	Pastes contents of Windows clipboard to cells.
CTRL+X	Cuts contents of selected cells and paste to Windows clipboard.
CTRL+Y	Redoes last undo.
CTRL+Z	Undoes last operation.

Chapter 5
Exploring Data in One Dimension

Contents

Overview of Exploring Data in One Dimension **59**
Bar Charts **59**
Example: Create a Bar Chart 60
Bar Chart Properties 62
Bar Charts of Selected Variables 64
Histograms **64**
Example: Create a Histogram 64
Histogram Properties 66
Histograms of Selected Variables 68
Example: Change the Positions of Histograms Bins 68
Interactive Histogram Binning 70
Box Plots **71**
Example: Create a Box Plot 72
Box Plot Properties 75
Box Plots of Selected Variables 76
References **76**

Overview of Exploring Data in One Dimension

This chapter describes how to use SAS/IML Studio to examine univariate distributions. You can explore the distributions of nominal variables by using bar charts. You can explore the univariate distributions of interval variables by using histograms and box plots.

Bar Charts

This section describes how to use a bar chart to visualize the distribution of a nominal variable. A bar chart shows the relative frequency of unique values of a variable. The height of each bar is proportional to the number of observations with each given value.

Example: Create a Bar Chart

In this section you create a bar chart of the category variable of the Hurricanes data set. The category variable gives the Saffir-Simpson wind intensity category for each observation.

The category variable is encoded according to the value of wind_kts, as shown in Table 5.1.

Table 5.1 The Saffir-Simpson Intensity Scale

Category	Description	Wind Speed (Knots)
TD	Tropical depression	22–33
TS	Tropical storm	34–63
Cat1	Category 1 hurricane	64–82
Cat2	Category 2 hurricane	83 –95
Cat3	Category 3 hurricane	96 –113
Cat4	Category 4 hurricane	114 –134
Cat5	Category 5 hurricane	135 or greater

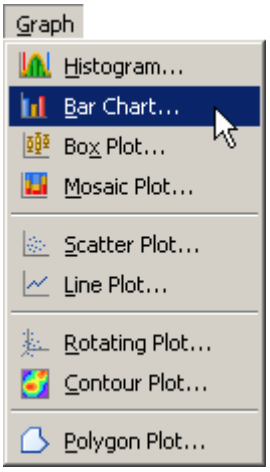
The category variable also has missing values, which represent weak intensities (wind speed less than 22 knots).

To create a bar chart:

1 Open the Hurricanes data set.

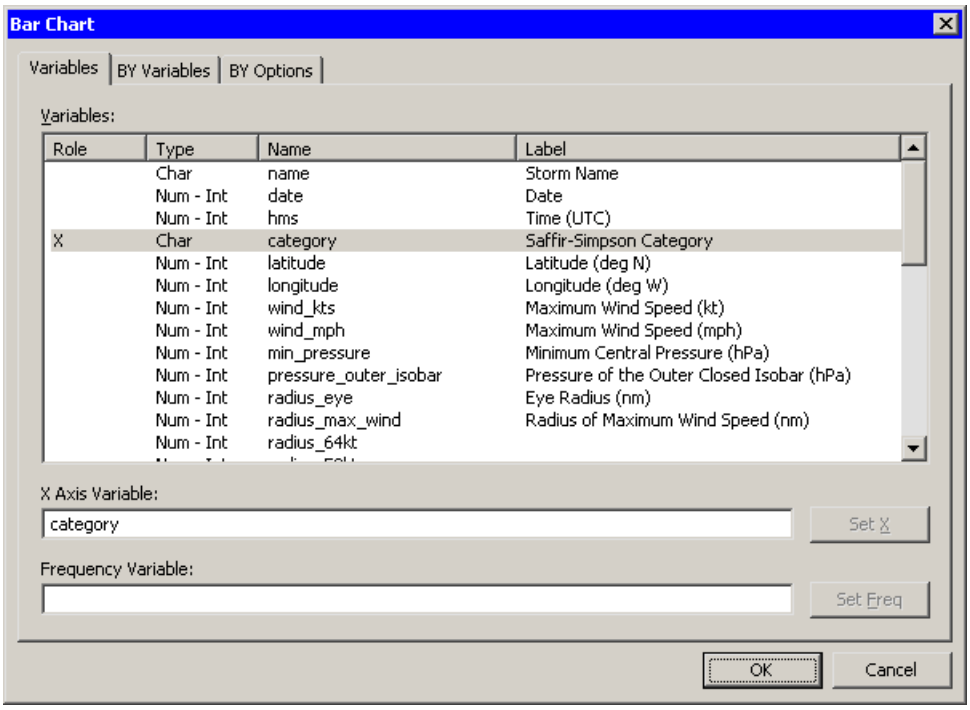
2 Select **Graph ►Bar Chart** from the main menu, as shown in Figure 5.1.

Figure 5.1 Selecting a Bar Chart



The Bar Chart dialog box appears. (See Figure 5.2.)

Figure 5.2 The Bar Chart Dialog Box



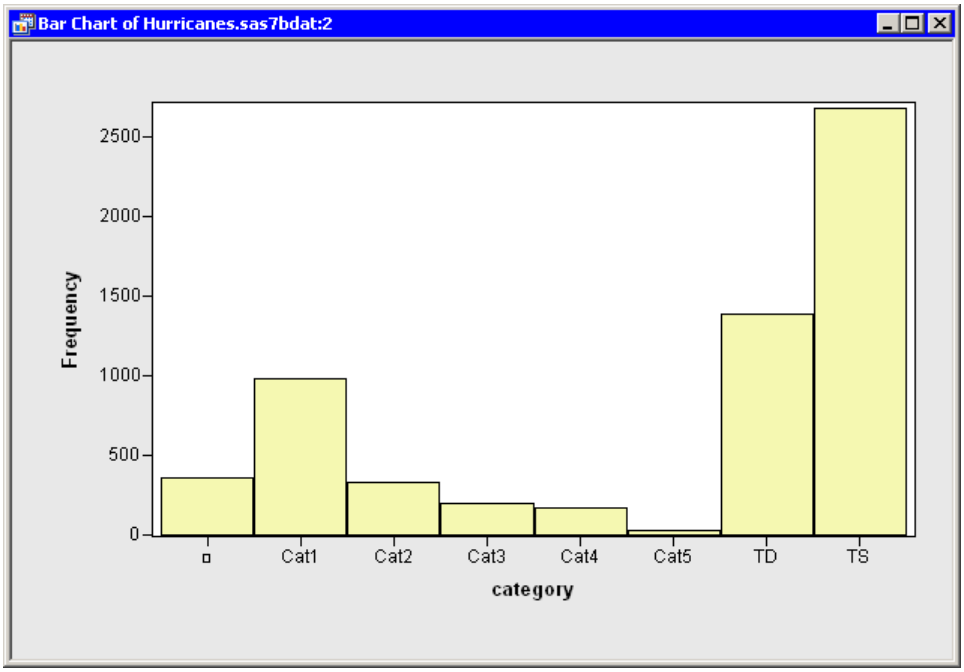
3 Select the category variable, and click **Set X**.

4 Click **OK**.

NOTE: The bar chart also supports an optional frequency variable.

A bar chart appears (Figure 5.3), which shows the unique values of the category variable. The chart shows that most of the observations in the data set are for tropical storms and tropical depressions. There are relatively few category 5 hurricanes.

Figure 5.3 A Bar Chart



The category variable has missing values. The set of missing values are grouped together and represented by a bar that is labeled with the □ symbol.

You can click a bar to select the observations contained in that bar. You can click while holding down the CTRL key to select observations in multiple bars. You can draw a selection rectangle to select observations in contiguous bars.

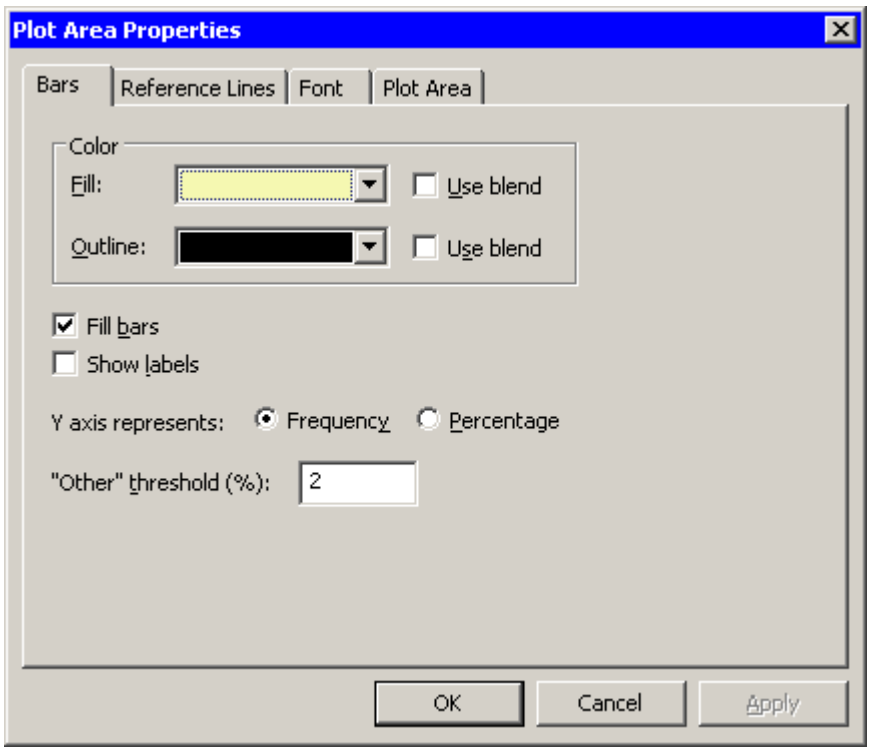
You can create bar charts of any nominal variable, numeric or character.

Bar Chart Properties

This section describes the **Bars** tab that is associated with a bar chart. To access the bar chart properties, right-click near the center of a plot, and select **Plot Area Properties** from the pop-up menu.

The **Bars** tab controls attributes of the bar chart. The **Bars** tab is shown in Figure 5.4.

Figure 5.4 Plot Area Properties for a Bar Chart



The **Bars** tab contains the following UI controls:

Fill
sets the fill color for each bar.

Fill: Use blend
sets the fill color for each bar according to a color gradient.

Outline
sets the outline color for each bar.

Outline: Use blend
sets the outline color for each bar according to a color gradient.

Fill bars
specifies whether each bar is filled with a color. When not selected, only the outline of the bar is shown.

Show labels
specifies whether each bar is labeled with the height of the bar.

Y axis represents
specifies whether the vertical scale represents frequency counts or percentage.

"Other" threshold (%)
sets a cutoff value for determining which observations are placed into an "Others" category.

For a discussion of the remaining tabs, see Chapter 9, "General Plot Properties."

Bar Charts of Selected Variables

If one or more nominal variables are selected in a data table when you select **Graph ►Bar Chart**, then the Bar Chart dialog box does not appear. Instead bar charts are created of the selected nominal variables.

You can also select nominal *and* interval variables and select **Graph ►Bar Chart**. A bar chart appears for each nominal variable; a histogram appears for each interval variable.

If you create a matrix of plots from selected variables, you can close the matrix by pressing the F11 key while any plot is active and selecting from the pop-up menu. Alternatively, you can use the Workspace Explorer to quickly close plots. (See the section "Workspace Explorer" on page 190.)

If a variable in the data table has a Frequency role, it is automatically used as the frequency variable for the plots; the frequency variable should not be one of the selected variables.

Variables with a Weight role are ignored when you are creating bar charts. For more information about the Frequency and Weight roles, see the section "The Variables Menu" on page 38.

Histograms

This section describes how to use a histogram to visualize the distribution of a continuous (interval) variable. A histogram is an estimate of the density of data. The range of the variable is divided into a certain number of subintervals, or bins. The height of the bar in each bin is proportional to the number of data points that have values in that bin. A histogram is determined not only by the bin width, but also by the choice of an anchor (or origin).

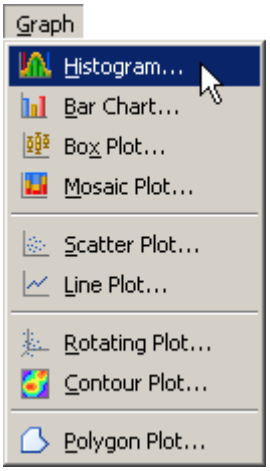
Example: Create a Histogram

In this section you create a histogram of the latitude variable of the Hurricanes data set. The latitude variable gives the latitude of the center of each tropical cyclone observation.

To create a histogram:

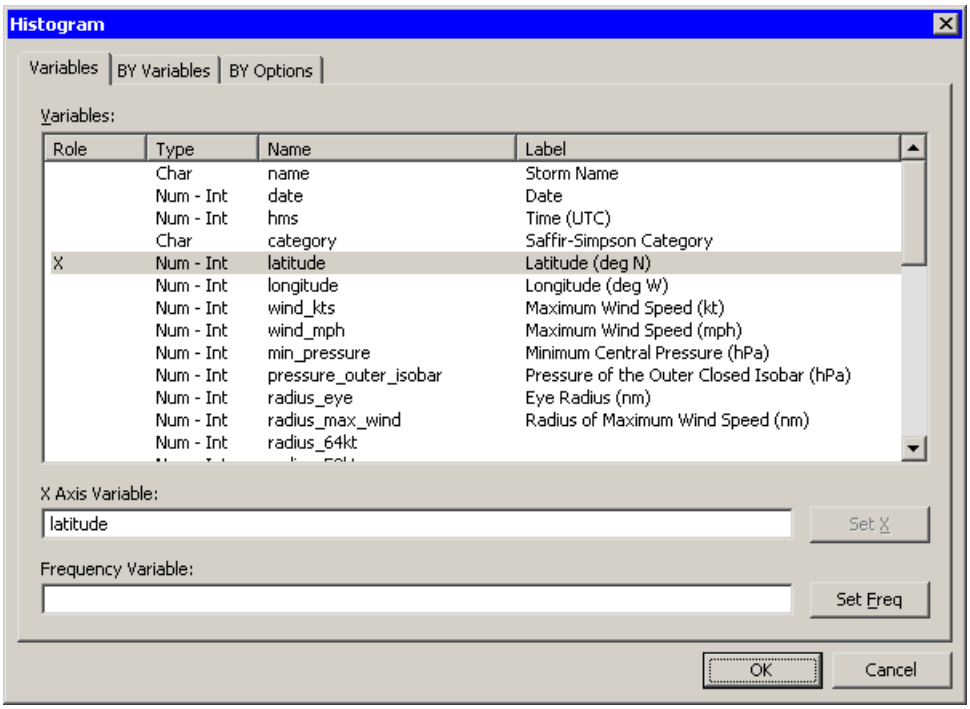
1. Open the Hurricanes data set.
2. Select **Graph ►Histogram** from the main menu, as shown in Figure 5.5.

Figure 5.5 Selecting a Histogram



The Histogram dialog box appears. (See Figure 5.6.)

Figure 5.6 The Histogram Dialog Box



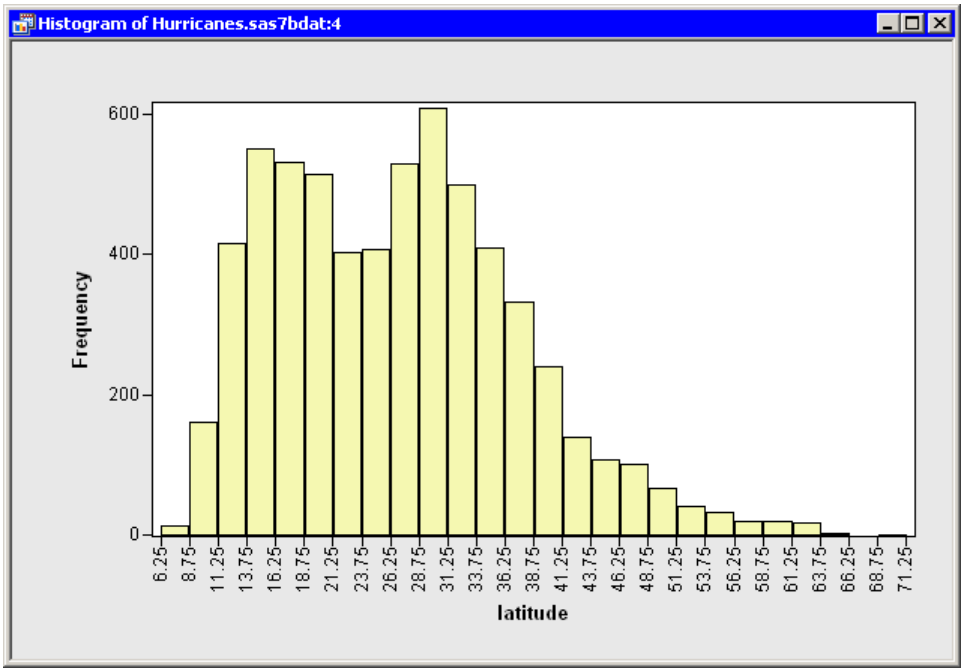
3 Select the latitude variable, and click **Set X**.

4 Click **OK**.

NOTE: The histogram also supports an optional frequency variable.

A histogram appears (Figure 5.7), which shows the distribution of latitudes for the tropical cyclones in this data set. The histogram shows that most Atlantic tropical cyclones occur between 10 and 40 degrees north latitude. The data distribution looks bimodal: one mode near 15 degrees and the other near 30 degrees of latitude.

Figure 5.7 A Histogram



If a variable has missing values, those values are not included in the histogram.

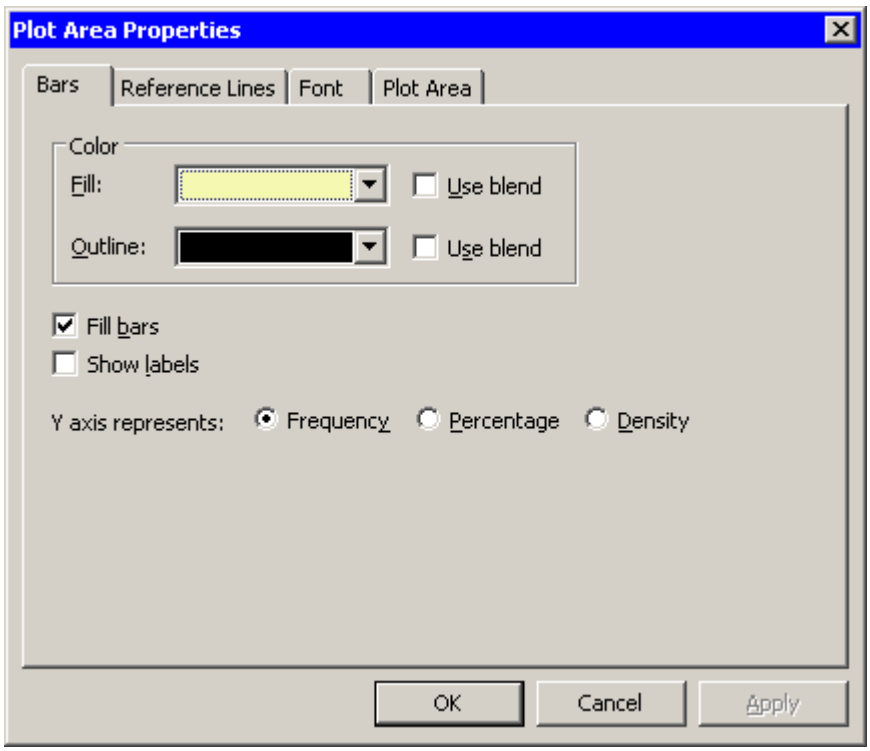
You can click a histogram bar to select the observations contained in that bin. You can click while holding down the CTRL key to select observations in multiple bins. You can draw a selection rectangle to select observations in contiguous bins.

Histogram Properties

This section describes the **Bars** tab that is associated with a histogram. To access the histogram properties, right-click near the center of a plot, and select **Plot Area Properties** from the pop-up menu.

The **Bars** tab controls attributes of the histogram. The **Bars** tab is shown in Figure 5.8.

Figure 5.8 Plot Area Properties for a Histogram



The **Bars** tab contains the following UI controls:

Fill
sets the fill color for each bar.

Fill: Use blend
sets the fill color for each bar according to a color gradient.

Outline
sets the outline color for each bar.

Outline: Use blend
sets the outline color for each bar according to a color gradient.

Fill bars
specifies whether each bar is filled with a color. When not selected, only the outline of the bar is shown.

Show labels
specifies whether each bar is labeled with the height of the bar.

Y axis represents
specifies whether the vertical scale represents frequency counts, percentage, or density.

For a discussion of the remaining tabs, see Chapter 9, "General Plot Properties."

Histograms of Selected Variables

If one or more interval variables are selected in a data table when you select **Graph ►Histogram**, then the Histogram dialog box does not appear. Instead histograms are created of the selected interval variables.

You can also select nominal *and* interval variables and select **Graph ►Histogram**. A bar chart appears for each nominal variable; a histogram appears for each interval variable.

If a variable has a Frequency role, it is automatically used as the frequency variable for the plots; the frequency variable does not need to be selected.

If you create a matrix of plots from selected variables, you can close the matrix by pressing the F11 key while any plot is active and selecting from the pop-up menu. Alternatively, you can use the Workspace Explorer to quickly close plots. (See the section "Workspace Explorer" on page 190.)

Example: Change the Positions of Histograms Bins

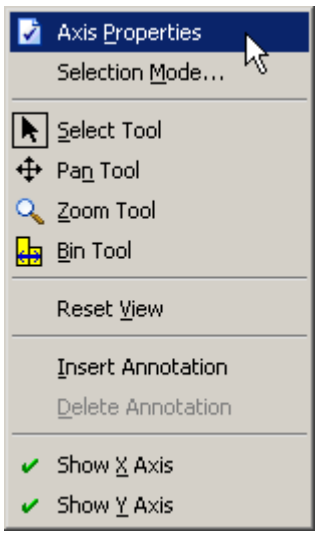
By default, SAS/IML Studio produces histograms with an anchor location and bin width chosen according to an algorithm by Terrell and Scott (1985). This section describes how you can choose a different anchor location or bin width for a histogram. The example in this section is a continuation of the example in "Example: Create a Histogram" on page 64, in which you created a histogram of the latitude variable in the Hurricanes data set.

For a histogram, the major tick unit is also the width of the histogram bins. For example, the tick marks for the histogram in Figure 5.7 are anchored at 6.25 and have a tick unit of 2.5. You can change the location of the histogram ticks so that the bins show the frequency of observations in the intervals 5–10, 10–15, 15–20, and so on.

To change the location of the histogram ticks:

1 Right-click anywhere on the horizontal axis of the histogram, and select **Axis Properties** from the pop-up menu, as shown in Figure 5.9.

Figure 5.9 The Axis Pop-up Menu

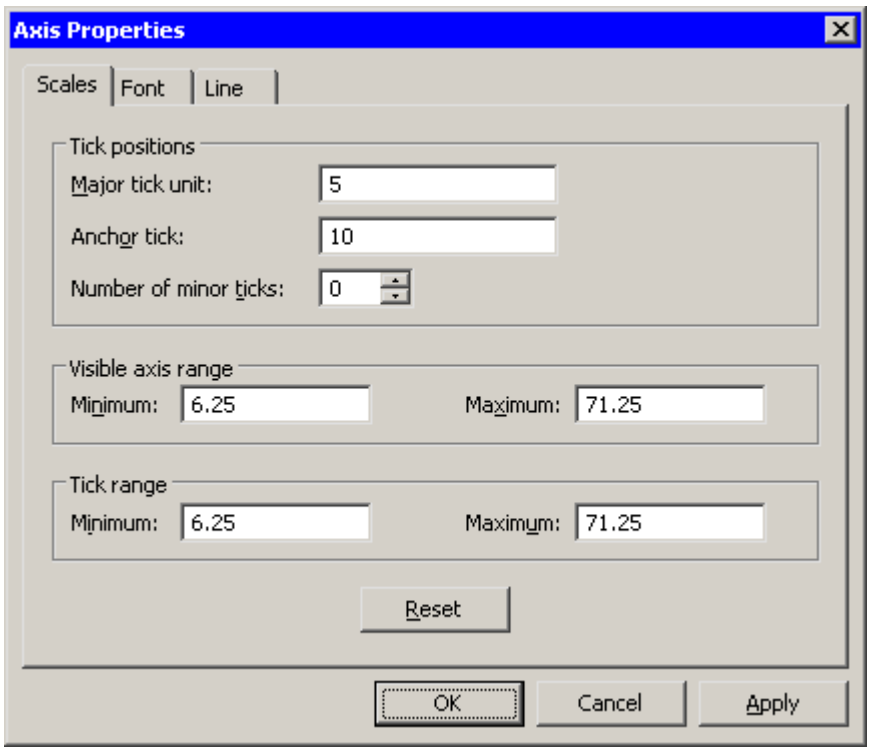


The Axis Properties dialog box appears as in Figure 5.10. This is a quick way to determine the anchor location, tick unit, and tick range for an axis.

2 Change the **Major tick unit** value to `5`.

3 Change the **Anchor tick** value to `10`.

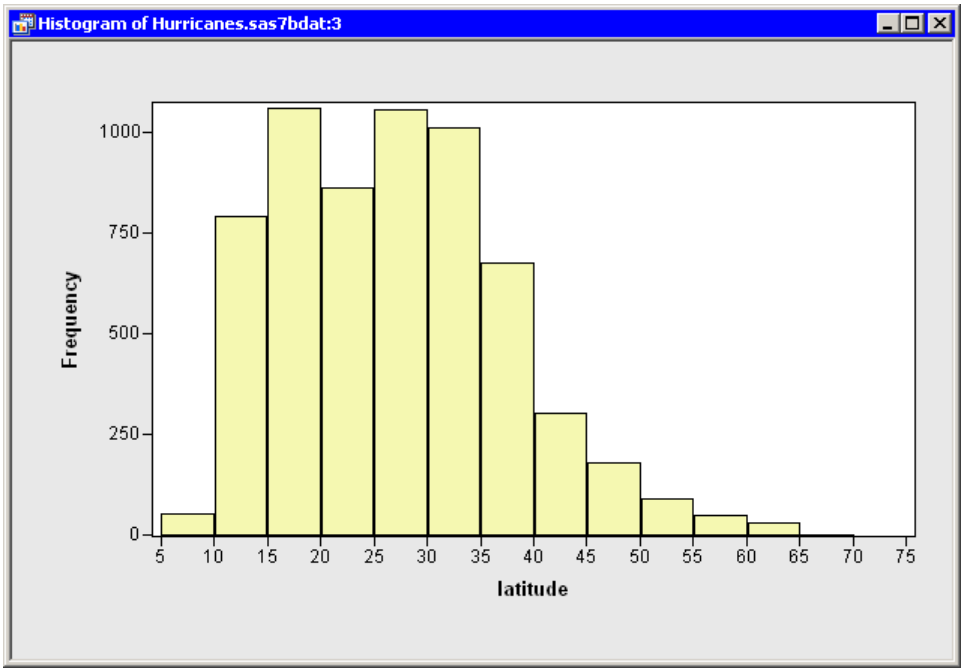
Figure 5.10 Dialog Box for Specifying Histogram Bins



4 Click **OK**.

The histogram updates to reflect the new histogram bin locations. The revised histogram is shown in Figure 5.11. The **Tick Range** field shown in Figure 5.10 is automatically widened, if necessary, so that all data are contained in bins.

Figure 5.11 Histogram with Customized Bins

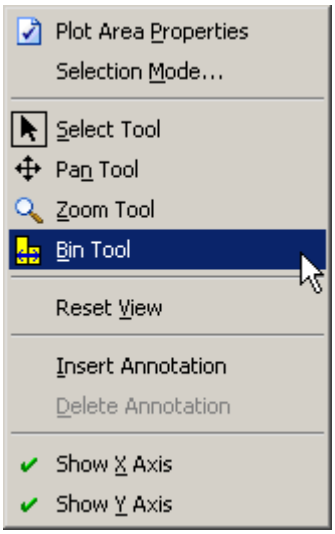


Interactive Histogram Binning

Sometimes it is useful to explore how the shape of a histogram varies with different combinations of anchor locations and bin widths. Interactively changing the histogram can help you determine whether apparent modes in the data are real or are an artifact of a specific binning.

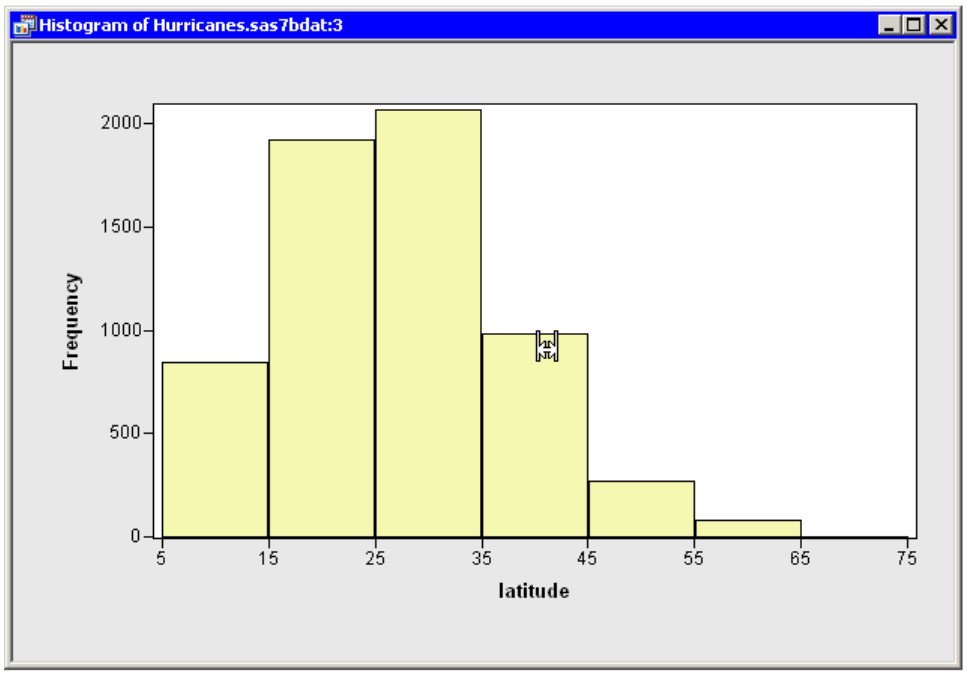
To interactively change the anchor location and bin width, right-click in the middle of the histogram and select **Bin Tool** from the pop-up menu, as shown in Figure 5.12.

Figure 5.12 The Histogram Pop-up Menu



The mouse pointer changes its shape, as shown in Figure 5.13. If you drag the pointer around in the plot area, then the histogram rebins. Dragging the pointer horizontally changes the anchor position. Dragging the pointer vertically changes the bin width. When the pointer is near the top of the plot area, the bin widths are relatively small; when the pointer is near the bottom, the bin widths are larger.

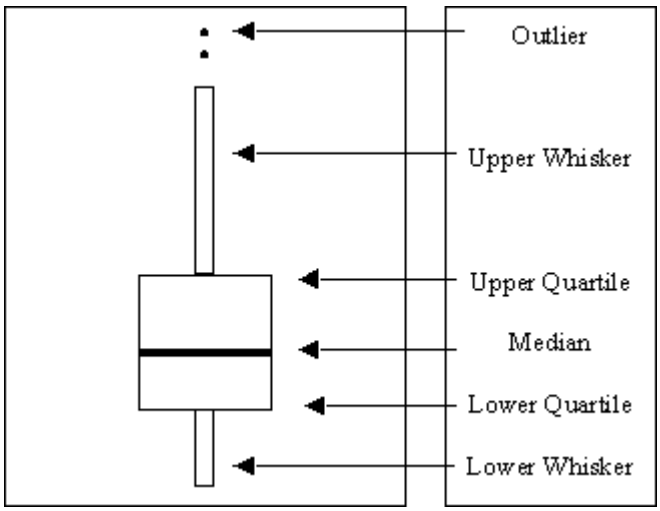
Figure 5.13 Interactively Rebinning a Histogram



Box Plots

A box plot summarizes the distribution of data sampled from a continuous numeric variable. The central line in a box plot indicates the median of the data, while the edges of the box indicate the first and third quartiles (that is, the 25th and 75th percentiles). Extending from the box are whiskers that represent data that are a certain distance from the median. Beyond the whiskers are outliers: observations that are relatively far from the median. These features are shown in Figure 5.14.

Figure 5.14 Schematic Description of a Box Plot



This section describes how to use a box plot to visualize the distribution of a continuous (interval) variable. You can also use box plots to see how the distribution changes across levels of one or more nominal variables.

Example: Create a Box Plot

In this section you create a box plot of the latitude variable of the Hurricanes data set, grouped by levels of the category variable. The latitude variable gives the latitude of the center of each tropical cyclone observation. The category variable gives the Saffir-Simpson wind intensity category for each observation.

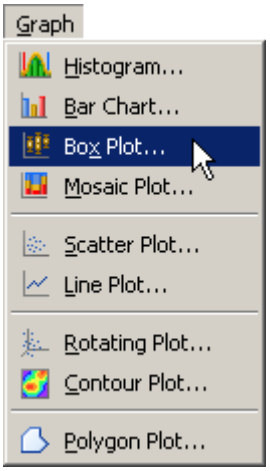
The category variable also has missing values, which represent weak intensities (wind speed less than 22 knots).

To create a box plot:

1 Open the Hurricanes data set.

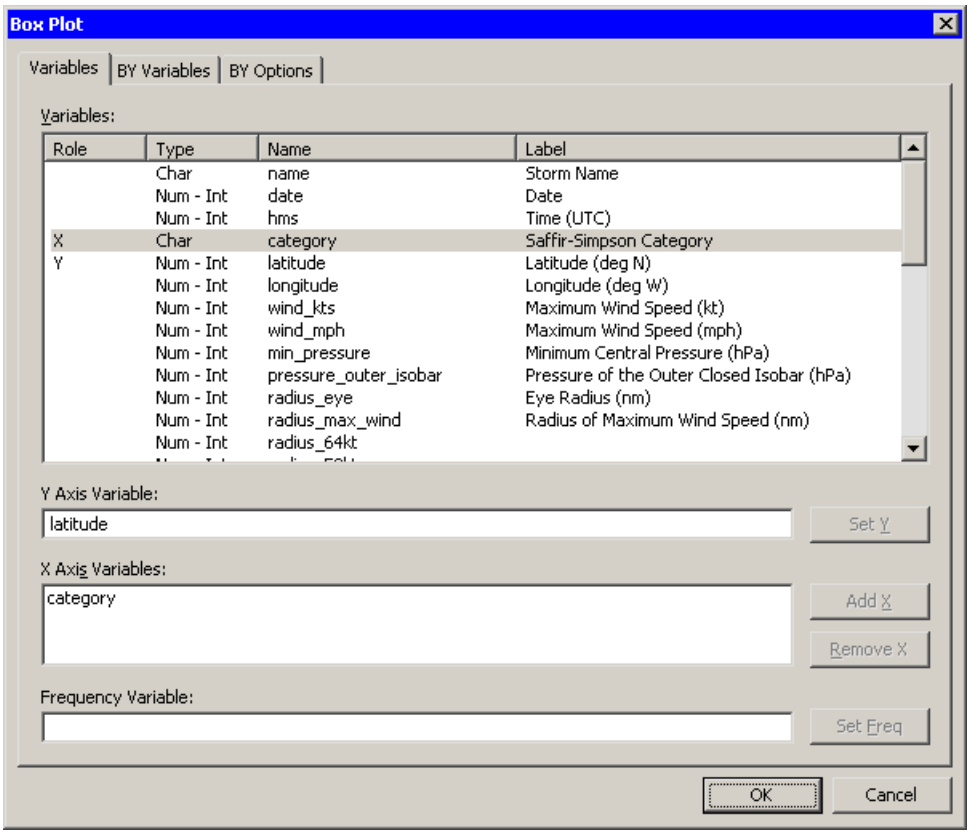
2 Select **Graph ►Box Plot** from the main menu, as shown in Figure 5.15.

Figure 5.15 Selecting a Box Plot



The Box Plot dialog box appears as in Figure 5.16.

Figure 5.16 The Box Plot Dialog Box



3 Select the latitude variable, and click **Set Y**.

4 Select the category variable, and click **Add X**.

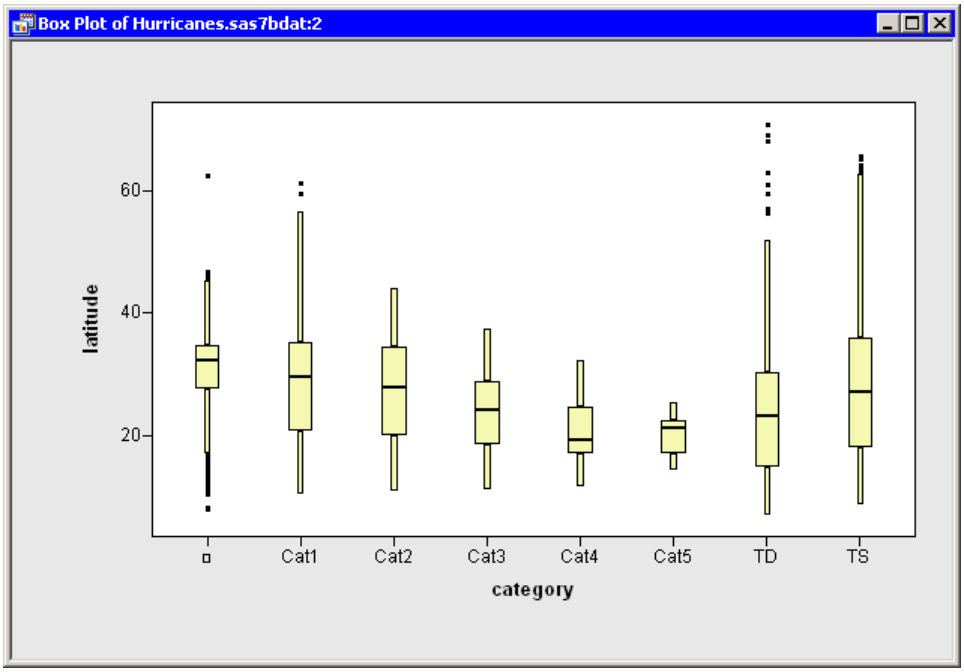
5 Click **OK**.

NOTE: X variables are optional. If you do not select an X variable, you get a box plot of the Y variable. Only nominal variables can be selected as an X variable.

NOTE: The box plot also supports an optional frequency variable.

A box plot appears (Figure 5.17), which shows the distribution of the latitude variable for each unique value of the category variable. The plot shows that the most intense hurricanes occur in a relatively narrow band of southern latitudes. Intense hurricanes have median latitudes that are farther south than weaker hurricanes. There is also less variance in the latitudes of the intense hurricanes. Tropical storms and tropical depressions do not follow these general trends, and they have the largest spread in latitude.

Figure 5.17 A Box Plot



The category variable has missing values. The set of missing values are grouped together and represented by a bar labeled with the □ symbol.

You can click any box, whisker, or outlier to select the observations contained in that box. You can click while holding down the CTRL key to select observations in multiple boxes. You can draw a selection rectangle to select observations in adjacent boxes.

Box Plot Properties

This section describes the **Boxes** tab that is associated with a box plot. To access the box plot properties, right-click near the center of a plot, and select **Plot Area Properties** from the pop-up menu.

The **Boxes** tab controls attributes of the box plot. The **Boxes** tab is shown in Figure 5.18.

The **Boxes** tab contains the following UI controls:

Box: Whisker length
: sets the length of the whiskers. A length of w means that whiskers are drawn from the quartiles to the farthest observation not more than w times the interquartile distance (Q3–Q1).

Box: with serifs
: specifies whether each whisker is capped with a horizontal line segment.

Box: with notches
: specifies whether each box is drawn with notches. The medians of two box plots are significantly different at approximately the 0.05 level if the corresponding notches do not overlap.

Mean: with one standard deviation
: specifies whether each box is drawn with mean markers that extend one standard deviation from the mean. The central line of the mean marker indicates the mean. The upper and lower extents of the mean marker indicate the mean plus or minus one standard deviation.

Mean: with two standard deviations
: specifies whether each box is drawn with mean markers that extend two standard deviation from the mean.

Mean: Shape
: specifies whether the mean markers are drawn as a diamond or an ellipse.

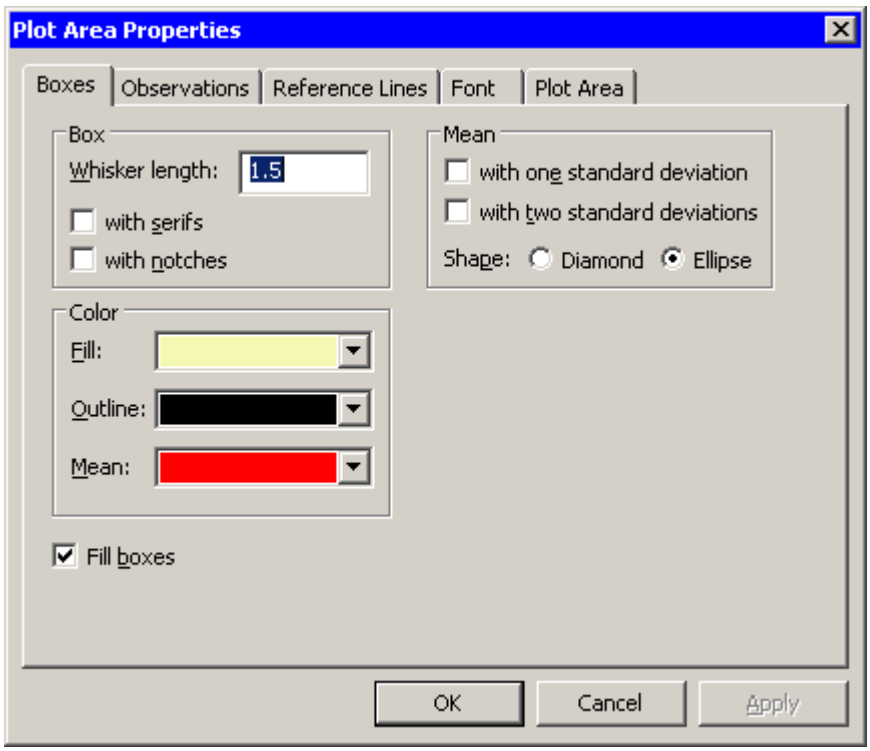
Color: Fill
: sets the fill color for each box.

Color: Outline
: sets the outline color for each box.

Color: Mean
: sets the color for mean markers.

Fill boxes
: specifies whether each box is filled with a color. When not selected, only the outline of the box is shown.

Figure 5.18 Plot Area Properties for a Box Plot



For a discussion of the **Observations** tab, see Chapter 6, "Exploring Data in Two Dimensions." For a discussion of the remaining tabs, see Chapter 9, "General Plot Properties."

Box Plots of Selected Variables

If one or more interval variables are selected in a data table when you select **Graph ►Box Plot**, then the Box Plot dialog box does not appear. Instead box plots are created for each selected interval variable.

You can also select nominal *and* interval variables and select **Graph ►Box Plot**. A box plot appears for each interval variable; nominal variables are assigned to the X axis.

If a variable has a Frequency role, it is automatically used as the frequency variable for the plots; the frequency variable does not need to be selected.

If you create a matrix of plots from selected variables, you can close the matrix by pressing the F11 key while any plot is active and selecting from the pop-up menu. Alternatively, you can use the Workspace Explorer to quickly close plots. (See the section "Workspace Explorer" on page 190.)

References

Terrell, G. R. and Scott, D. W. (1985), "Oversmoothed Nonparametric Density Estimates," *Journal of the American Statistical Association*, 80, 209–214.

Chapter 6
Exploring Data in Two Dimensions

Contents

Overview of Exploring Data in Two Dimensions **77**
Mosaic Plots **78**
Example: Create a Mosaic Plot 78
Mosaic Plot Properties 81
Mosaic Plots of Selected Variables 82
Scatter Plots **83**
Example: Create a Scatter Plot 83
Scatter Plot Properties 85
Scatter Plots of Selected Variables 86
Line Plots **87**
Example: Create a Line Plot from Multiple Y Variables 88
Example: Create a Line Plot from a Group Variable 92
Line Plot Properties 96
Line Plots of Selected Variables 97
Polygon Plots **98**
Example: Create a Polygon Plot 98
Polygon Plot Properties 101
Polygon Plots of Selected Variables 102

Overview of Exploring Data in Two Dimensions

This chapter describes how to use SAS/IML Studio to examine relationships between pairs of variables.

You can explore the relationship between two (or more) nominal variables by using a mosaic chart. You can explore the relationship between two variables by using a scatter plot. Usually the variables in a scatter plot are interval variables.

If you have a time variable, you can observe the behavior of one or more variables over time with a line plot. You can also use line plots to visualize a response variable (and, optionally, fitted curves and confidence bands) versus values of an explanatory variable.

You can create and explore maps with a polygon plot.

Mosaic Plots

This section describes how to use a mosaic plot to visualize the cells of a contingency table. A mosaic plot displays the frequency of data with respect to multiple nominal variables.

A mosaic plot is a set of adjacent bar plots formed first by dividing the horizontal axis according to the proportion of observations in each category of the first variable and then by dividing the vertical axis according to the proportion of observations in the second variable. For more than two nominal variables, this process can be continued by further horizontal or vertical subdivision. The area of each block is proportional to the number of observations it represents.

Example: Create a Mosaic Plot

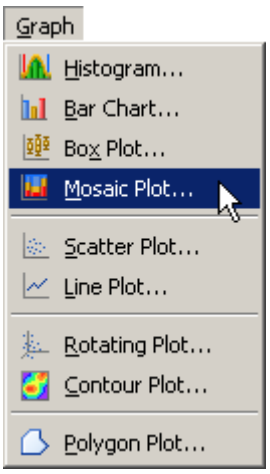
In this section you create a mosaic plot of the nation and industry variables of the Business data set. The nation variable gives the nation of each business listed in the data set, and the industry variable assigns each business to a category that describes the business.

To create a mosaic plot:

1 Open the Business data set.

2 Select **Graph ►Mosaic Plot** from the main menu, as shown in Figure 6.1.

Figure 6.1 Selecting a Mosaic Plot



The Mosaic Plot dialog box appears. (See Figure 6.2.)

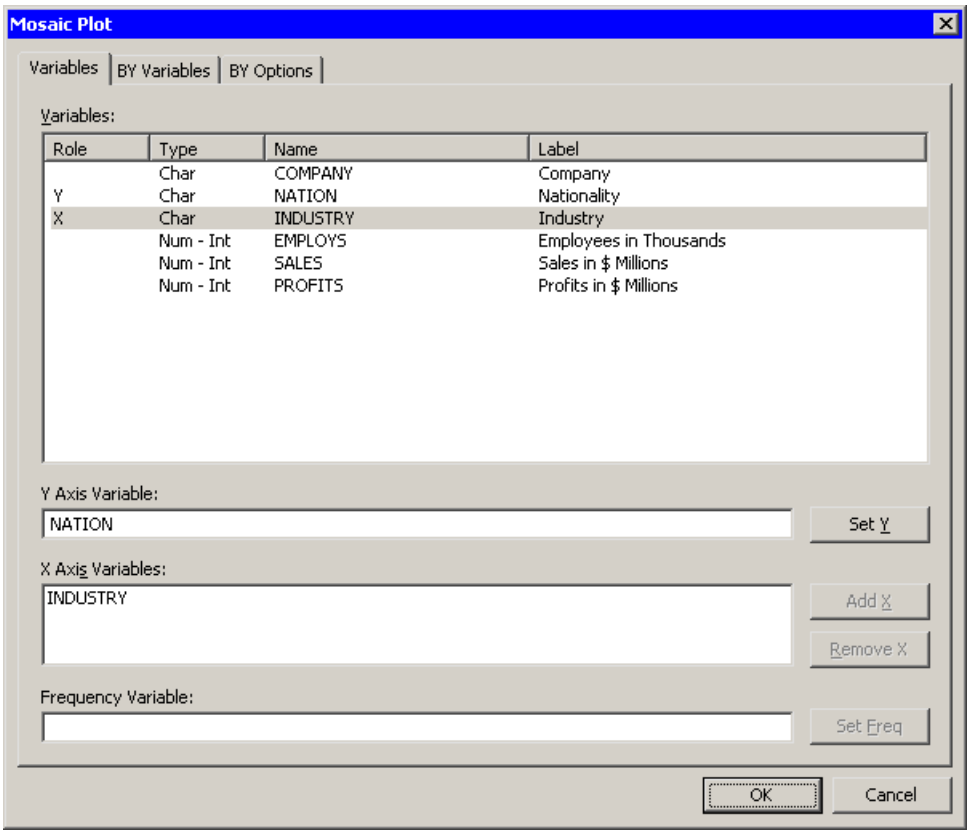
3 Select the nation variable, and click **Set Y**.

4 Select the industry variable, and click **Add X**.

5 Click **OK**.

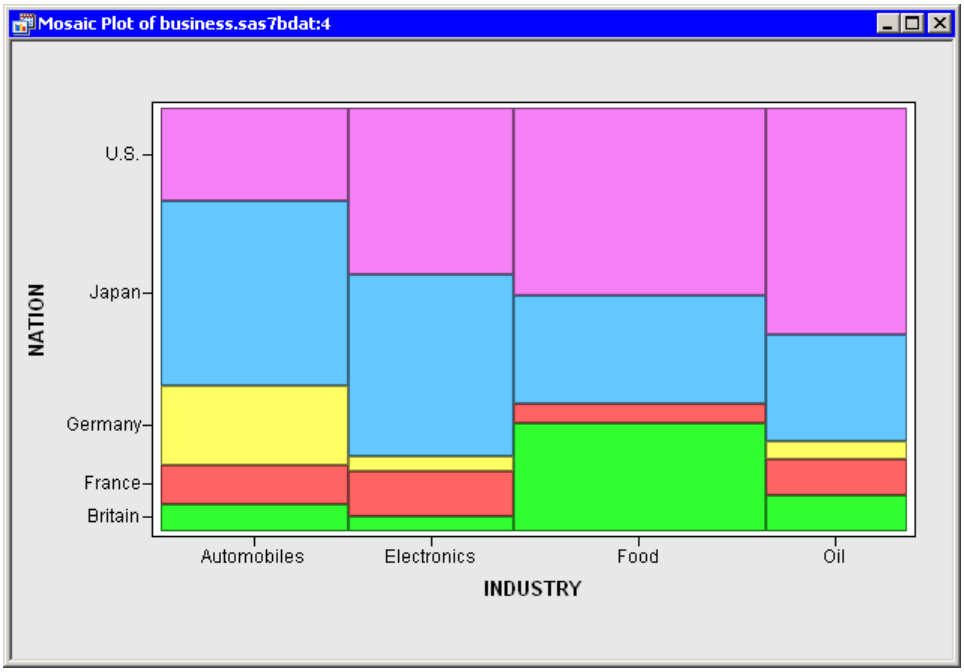
NOTE: The mosaic also supports an optional frequency variable.

Figure 6.2 The Mosaic Plot Dialog Box



A mosaic plot appears (Figure 6.3), which shows the relative proportions of businesses in this data set as grouped by nation and industry. The mosaic plot shows that the U.S. food companies make up the largest subset, because that cell has the largest area. Other large cells include Japanese automobile companies, Japanese electronics companies, and U.S. oil companies. The plot also shows that there are no German food companies in the data set.

Figure 6.3 A Mosaic Plot



You can click a cell to select the observations contained in that cell. Clicking a cell also shows you the number of observations in that cell. You can click while holding down the CTRL key to select observations in multiple cells. You can draw a selection rectangle to select observations in contiguous cells.

You can create mosaic plots of any nominal variables, numeric or character. However, the variables should have a small to moderate number of levels.

The cells in this mosaic plot represent the count (number of observations) of businesses in each nation and industry. However, you might be more interested in comparing the revenue generated by these businesses. You can make this comparison by re-creating the mosaic plot and adding sales as a frequency variable.

6 Select **Graph ►Mosaic Plot** from the main menu.

The Mosaic Plot dialog box appears.

7 Select the nation variable, and click **Set Y**.

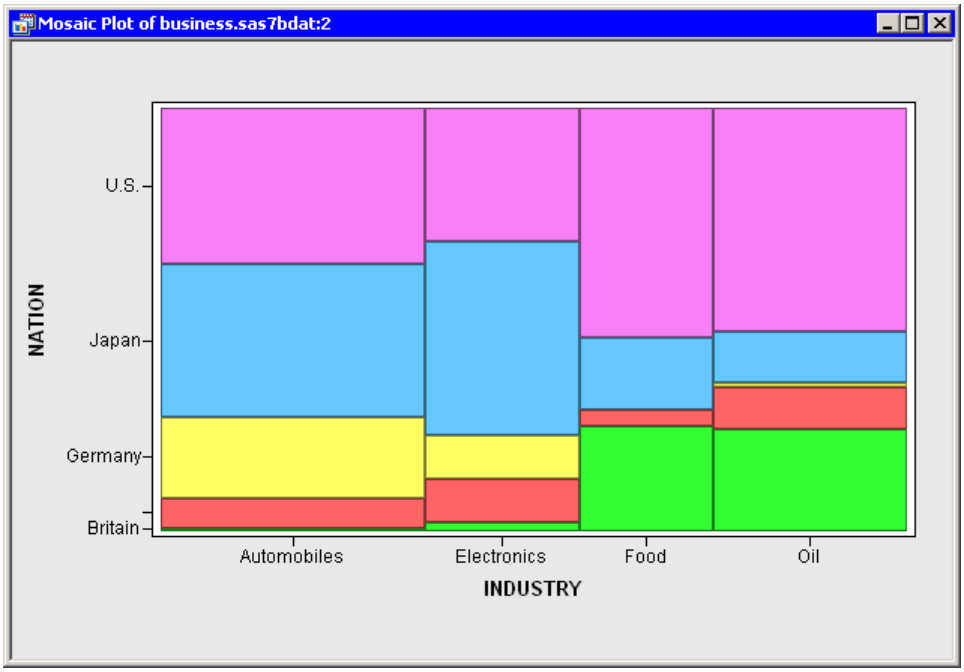
8 Select the industry variable, and click **Add X**.

9 Select the sales variable, and click **Set Freq**.

10 Click **OK**.

A mosaic plot appears (Figure 6.4), which shows the relative proportions of sales for each nation and industry. The mosaic plot shows that the U.S. oil companies generate the most revenue, followed by the U.S. and Japanese automobile companies. Companies from the U.S. and Japan account for over two thirds of the sales.

Figure 6.4 A Mosaic Plot with a Frequency Variable



Similarly, if you were interested in comparing the number of employees in these businesses, you could use employees as a frequency variable. However, note that you could not compare profits in this way, because some profits are negative and the mosaic plot ignores any observation whose frequency is negative. You should also make sure that the frequency variable contains integers; noninteger values are truncated.

Mosaic Plot Properties

This section describes the **Mosaic** tab that is associated with a mosaic plot. To access the mosaic plot properties, right-click near the center of a plot, and select **Plot Area Properties** from the pop-up menu.

The **Mosaic** tab controls attributes of the mosaic plot. The **Mosaic** tab is shown in Figure 6.5.

The **Mosaic** tab contains the following UI controls:

"Other" threshold (%)
sets a cutoff value for determining which observations are placed into an "Others" category.

Layout
sets the method by which cells are formed from the X and Y variables.

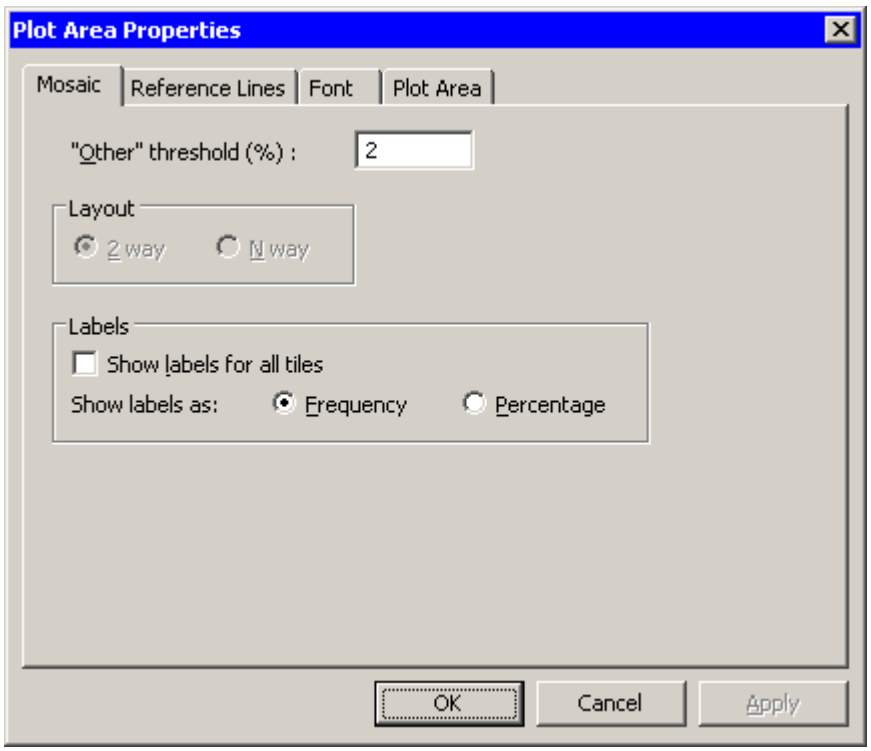
2 way In this layout scheme, the X variables determine groups, and the mosaic plot displays a stacked bar chart of the Y variable for each group.

N way This layout scheme is available only if there are exactly two X variables. In this layout scheme, the plot subdivides in the horizontal direction by the first X variable, then subdivides in the vertical direction by the Y variable, and finally subdivides in the horizontal direction by the second X variable.

Show labels for all tiles
specifies whether each cell is labeled with the proportion it represents.

Show labels as
specifies whether a cell represents frequency or percentage.

Figure 6.5 Plot Area Properties for a Mosaic Plot



For a discussion of the remaining tabs, see Chapter 9, "General Plot Properties."

Mosaic Plots of Selected Variables

If one or more nominal variables are selected in a data table when you select **Graph ►Mosaic Plot**, then the Mosaic Plot dialog box does not appear. Instead mosaic plots are created for each pair of the selected nominal variables.

If you create a matrix of plots from selected variables, you can close the matrix by pressing the F11 key while any plot is active and selecting from the pop-up menu. Alternatively, you can use the Workspace Explorer to quickly close plots. (See the section "Workspace Explorer" on page 190.)

If a variable in the data table has a Frequency role, it is automatically used as the frequency variable for the plots; the frequency variable should not be one of the selected variables.

Variables with a Weight role are ignored when you are creating mosaic plots.

Scatter Plots

This section describes how to use a scatter plot to visualize the relationship between two variables. Usually each variable is continuous (interval), but that is not a requirement.

Example: Create a Scatter Plot

In this section you create a scatter plot of the wind_kts and min_pressure variables of the Hurricanes data set. The wind_kts variable is the wind speed in knots; the min_pressure variable is the minimum central pressure for each observation.

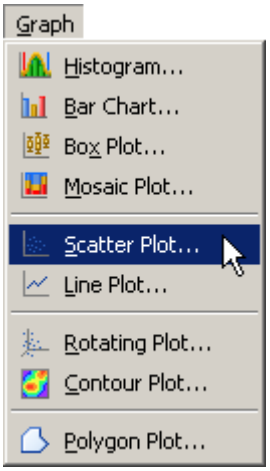
The min_pressure variable has a few missing values; those observations are not included in the scatter plot.

To create a scatter plot:

1 Open the Hurricanes data set.

2 Select **Graph ►Scatter Plot** from the main menu, as shown in Figure 6.6.

Figure 6.6 Selecting a Scatter Plot



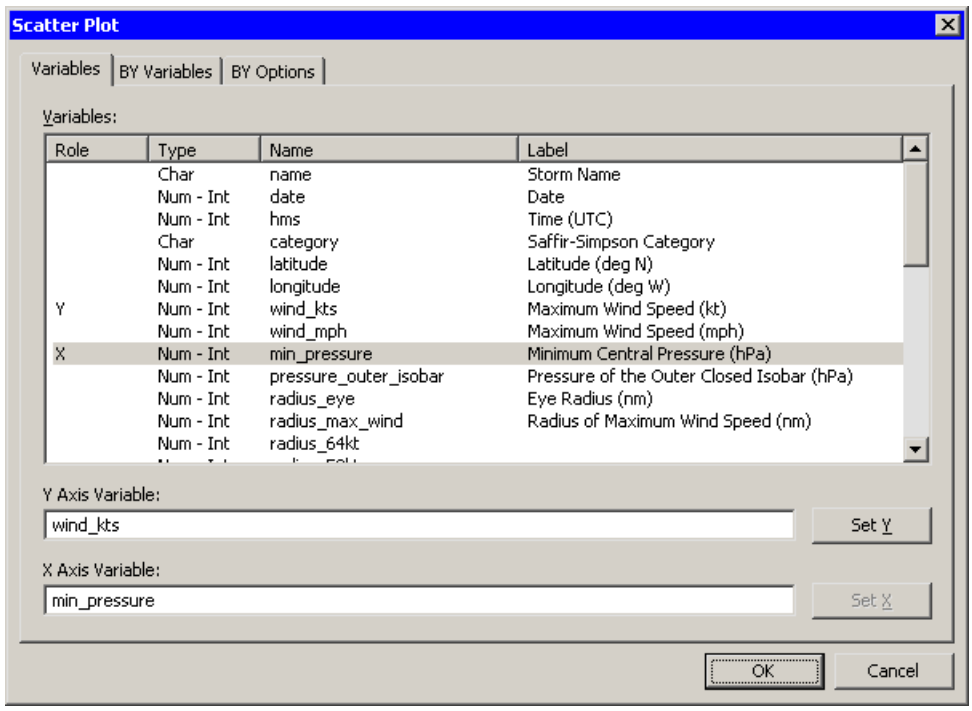
The Scatter Plot dialog box appears. (See Figure 6.7.)

3 Select the variable wind_kts, and click **Set Y**.

4 Select the variable min_pressure, and click **Set X**.

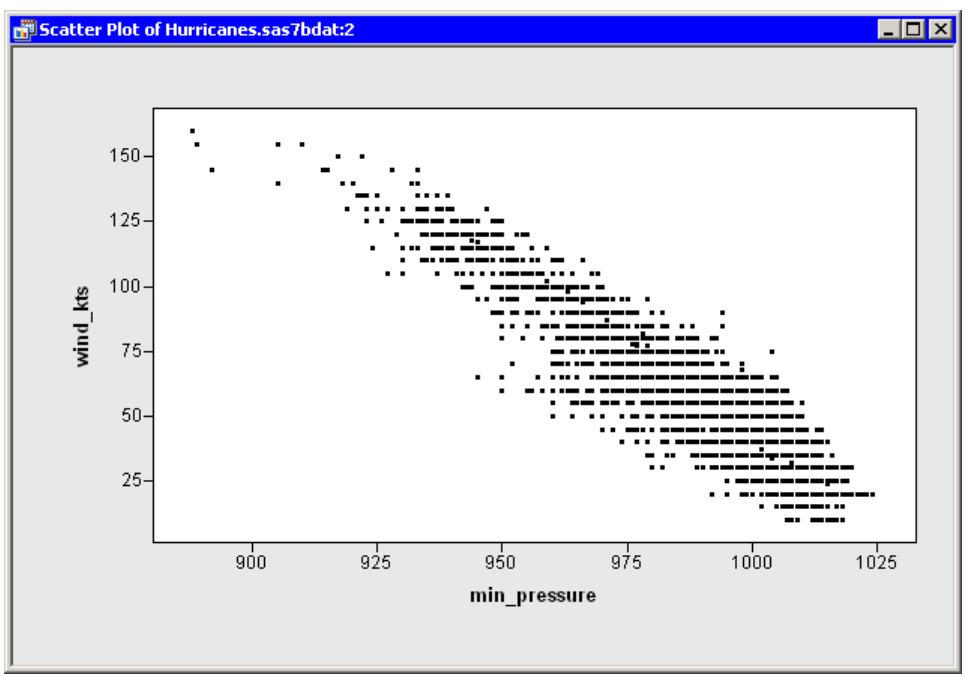
5 Click **OK**.

Figure 6.7 The Scatter Plot Dialog Box



A scatter plot appears (Figure 6.8) that shows the bivariate data. The plot shows a strong negative correlation ($\rho = -0.93$) between wind speed and pressure. The plot also shows that most, although not all, wind speeds are rounded to the nearest 5 knots.

Figure 6.8 A Scatter Plot



You can click any observation marker to select the observation. You can click while holding down the CTRL key to select multiple observations. You can draw a selection rectangle to select a group of observations.

Scatter Plot Properties

This section describes the **Observations** tab that is associated with a scatter plot. To access the scatter plot properties, right-click near the center of a plot, and select **Plot Area Properties** from the pop-up menu.

The **Observations** tab controls attributes of the scatter plot. The **Observations** tab is shown in Figure 6.9.

The **Observations** tab contains the following UI controls:

Marker Attributes: Shape
sets the shape of the marker for each observation.

Marker Attributes: Outline
specifies the color of the marker boundary. If the **Blend** list is set to **None**, the **Outline** list enables you to specify the outline color of observation markers. If the **Blend** list is not set to **None**, the **Outline** list enables you to specify the color blend to be used to color the outlines of observation markers.

Marker Attributes: Blend (Outline)
sets the variable whose values should be used to perform color blending for the outline colors of observation markers. If this value is set to **None**, color blending is not performed.

Marker Attributes: Fill
specifies the color of the marker interior. If the **Blend** list is set to **None**, the **Fill** list enables you to specify the fill color of observation markers. If the **Blend** list is not set to **None**, the **Fill** list enables you to specify the color blend to be used to color the interiors of observation markers.

Marker Attributes: Blend (Fill)
sets the variable whose values should be used to perform color blending for the fill colors of observation markers. If this value is set to **None**, color blending is not performed.

Marker Attributes: Apply to
specifies whether marker shape and color changes are applied to all observations, or just to the ones currently selected.

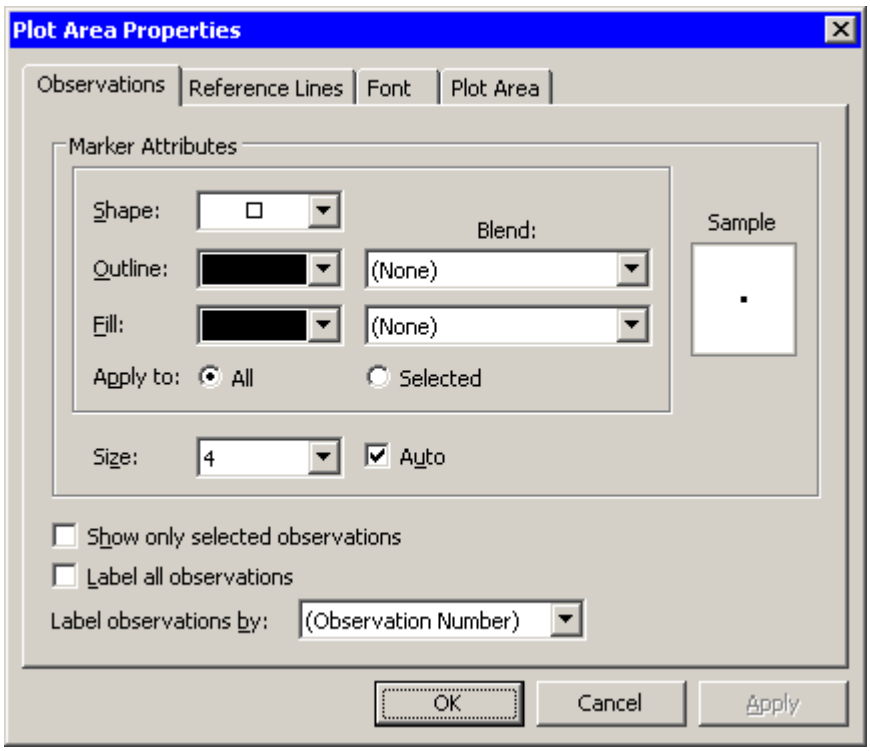
Marker Attributes: Size
specifies the size of observation markers. All observation markers in a plot are drawn at the same size. Selecting **Auto** causes the size of markers to change according to the size of the plot.

Show only selected observations
specifies whether observation markers are shown only for selected observations.

Label all observations
specifies whether labels are displayed next to each observation marker.

Label observations by
specifies the variable to use to label observations.

Figure 6.9 Plot Area Properties for a Scatter Plot



For a discussion of the remaining tabs, see Chapter 9, "General Plot Properties."

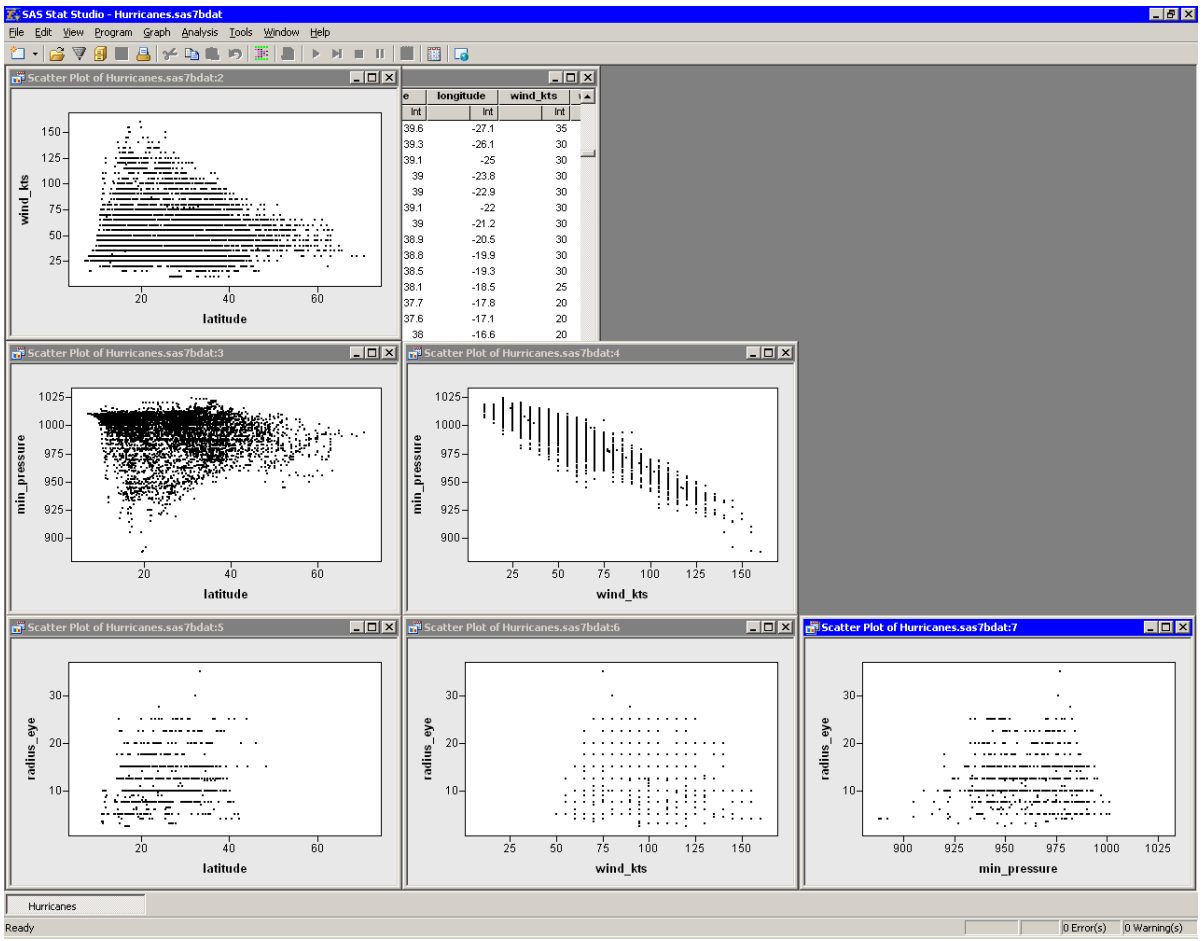
Scatter Plots of Selected Variables

If one or more variables are selected in a data table when you select **Graph ►Scatter Plot**, then the Scatter Plot dialog box does not appear. Instead, a scatter plot matrix is created that shows each pair of the selected variables. (See Figure 6.10.)

If you create a matrix of plots from selected variables, you can close the matrix by pressing the F11 key while any plot is active and selecting from the pop-up menu. Alternatively, you can use the Workspace Explorer to quickly close plots. (See the section "Workspace Explorer" on page 190.)

Variables with a Frequency or Weight role are ignored when you are creating scatter plots.

Figure 6.10 A Matrix of Scatter Plots



Line Plots

This section describes how to use a line plot to observe the behavior of one or more variables over time. You can also use line plots to visualize a response variable (and, optionally, fitted curves and confidence bands) versus values of an explanatory variable.

You can create line plots when your data are in one of two configurations. The first configuration (Table 6.1) is when you have an X variable and one or more Y variables. Each Y variable has the same number of observations as the X variable. (Some of the Y values might be missing.) In this configuration there are as many lines in the plot as there are Y variables.

Table 6.1 A Data Configuration for a Line Plot

X	Y1	Y2
1	1	4
2	3	3
3	2	3
4	4	2
5	5	1

In the second configuration (Table 6.2), there is a single X and a single Y variable, but there are one or more *group* variables that specify which line each observation belongs to. In this configuration there are as many lines in the plot as there are unique values of the group variables.

Table 6.2 An Alternative Data Configuration for a Line Plot

X	Y	Group
1	1	A
1	4	B
2	3	A
2	3	B
3	2	A
3	3	B
4	4	A
4	2	B
5	5	A
5	1	B

The X variable does not need to be sorted in either configuration. Any data arranged in the first configuration can be rewritten in the second. For example, Table 6.2 represents the same data as Table 6.1. The second configuration is more useful if you have different values of the X variable for each group.

Example: Create a Line Plot from Multiple Y Variables

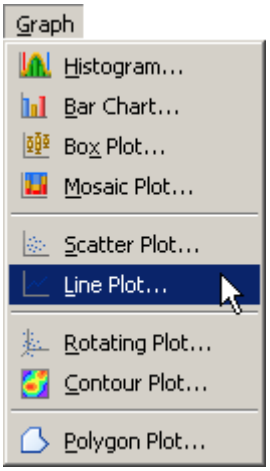
In this section you create a line plot of the co and wind variables versus the datetime variable of the Air data set. The co variable is a measurement of carbon monoxide. The wind variable is a measurement of wind speed. The datetime variable is the hour and date of each measurement.

To create a line plot:

1 Open the Air data set.

2 Select **Graph ►Line Plot** from the main menu, as shown in Figure 6.11.

Figure 6.11 Selecting a Line Plot



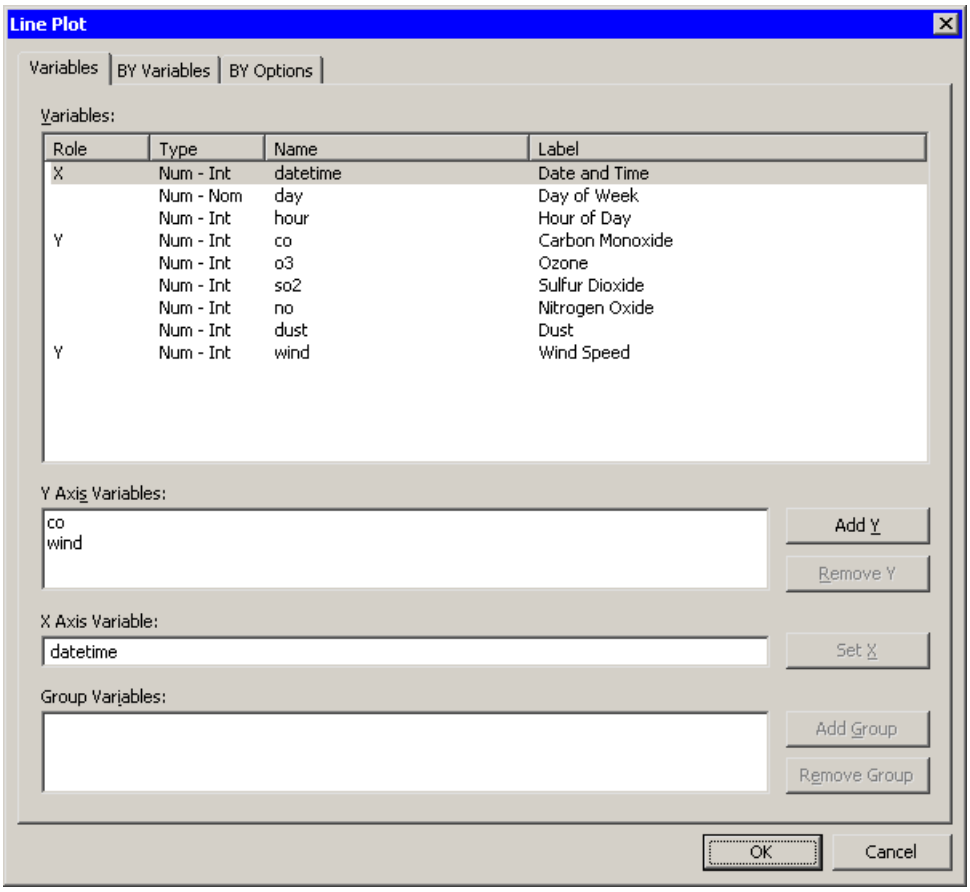
The Line Plot dialog box appears. (See Figure 6.12.)

3 Select the co variable. Hold down the CTRL key and select the wind variable. Click **Add Y**.

4 Select the variable datetime, and click **Set X**.

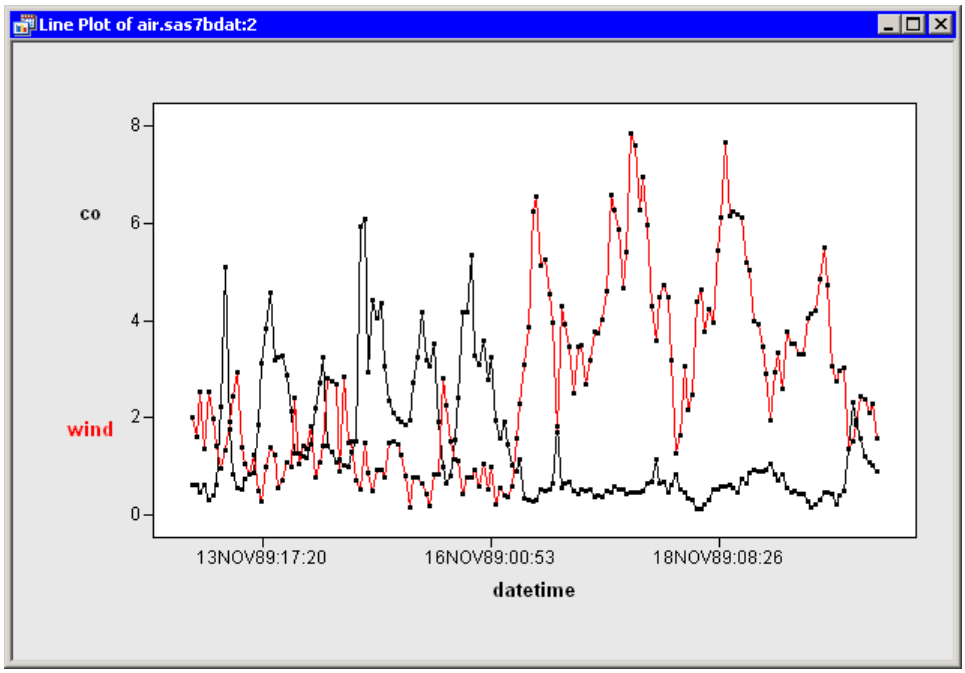
5 Click **OK**.

Figure 6.12 The Line Plot Dialog Box



A line plot appears (Figure 6.13), which shows the carbon monoxide and wind measurements for each hour of a seven-day period. By default, the two lines are displayed in different colors. You can change the color and line style of the lines, as shown in the remainder of this example.

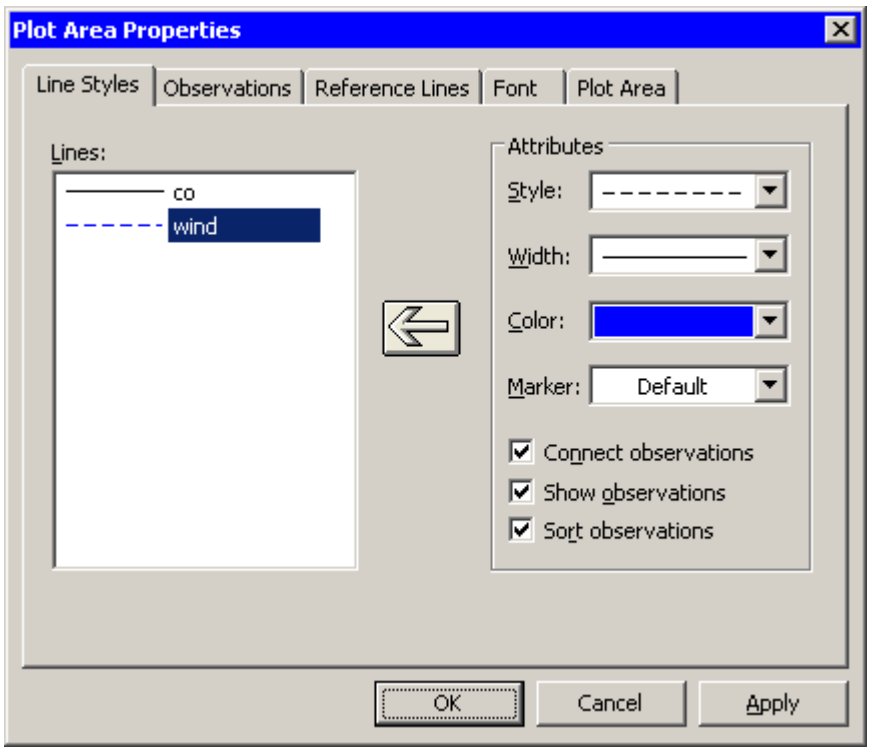
Figure 6.13 A Line Plot



6 Right-click near the center of the plot, and select **Plot Area Properties** from the pop-up menu.

The Plot Area Properties dialog box appears. (See Figure 6.14.)

Figure 6.14 Plot Area Properties for a Line Plot



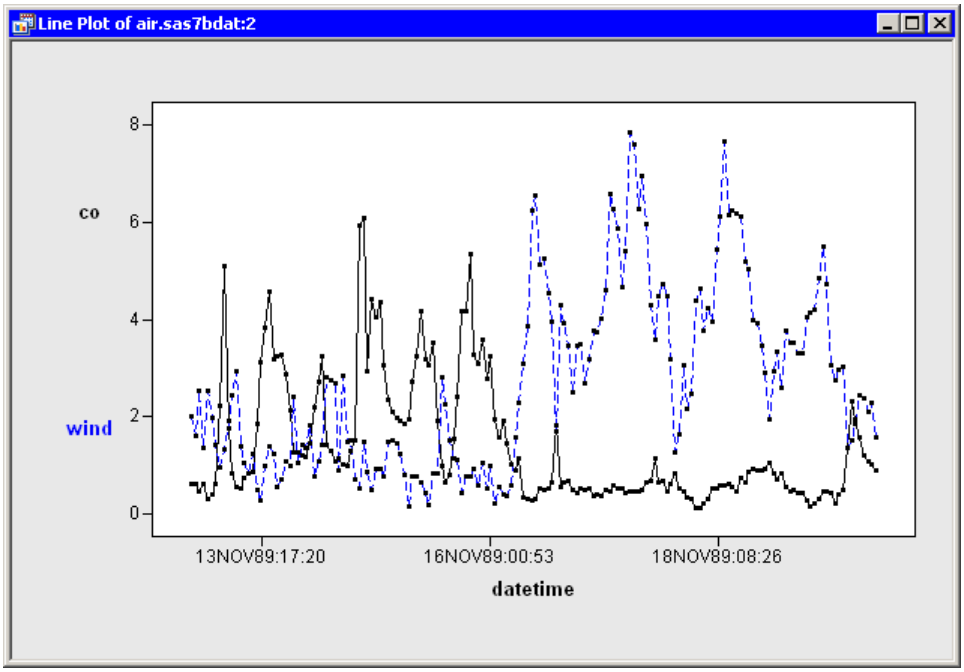
7 Select wind in the **Lines** list.

8 Change the line style to dashed and the color to blue.

9 Click the large left arrow in the center of the dialog box to apply the changes to the wind line.

10 Click **OK**.

Figure 6.15 A Line Plot with Line Colors and Styles



The line plot now looks like the plot in Figure 6.15. The carbon monoxide line shows periodic behavior for the first half of the week, followed by extremely low values for the second half of the week. The wind values are low for the first half of the week, but much stronger for the second half. These data might indicate that sufficiently strong winds can blow away carbon monoxide.

You can click any observation marker to select the observation. You can click while holding down the CTRL key to select multiple observations. You can draw a selection rectangle to select a group of observations. You can also select the lines themselves by clicking a line segment that is away from any observation. If you open the dialog box shown in Figure 6.14, selected lines in the line plot are also selected in the **Lines** list in the dialog box.

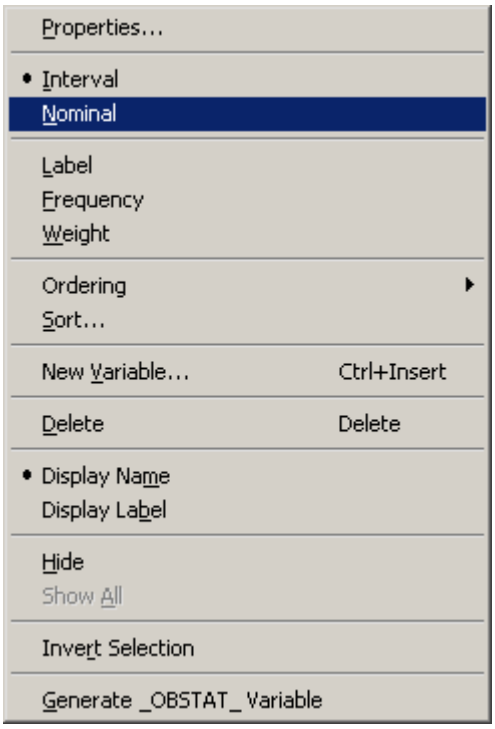
NOTE: If you plot multiple Y variables, then an observation in the data table is represented by multiple markers in the line plot. Clicking any marker in the plot selects the entire corresponding observation.

Example: Create a Line Plot from a Group Variable

The following steps use the same data set as the previous example, but this time plot the co variable over a 24-hour period for each day of the week.

1 In the data table, right-click the day variable, and select **Nominal** from the pop-up menu, as shown in Figure 6.16.

Figure 6.16 Changing the Role of a Variable



Nominal variables can be used as *group* variables in the construction of a line plot.

2 Press the ESC key to deselect the day variable.

3 Select **Graph ►Line Plot** from the main menu.

The Line Plot dialog box appears. (See Figure 6.17.)

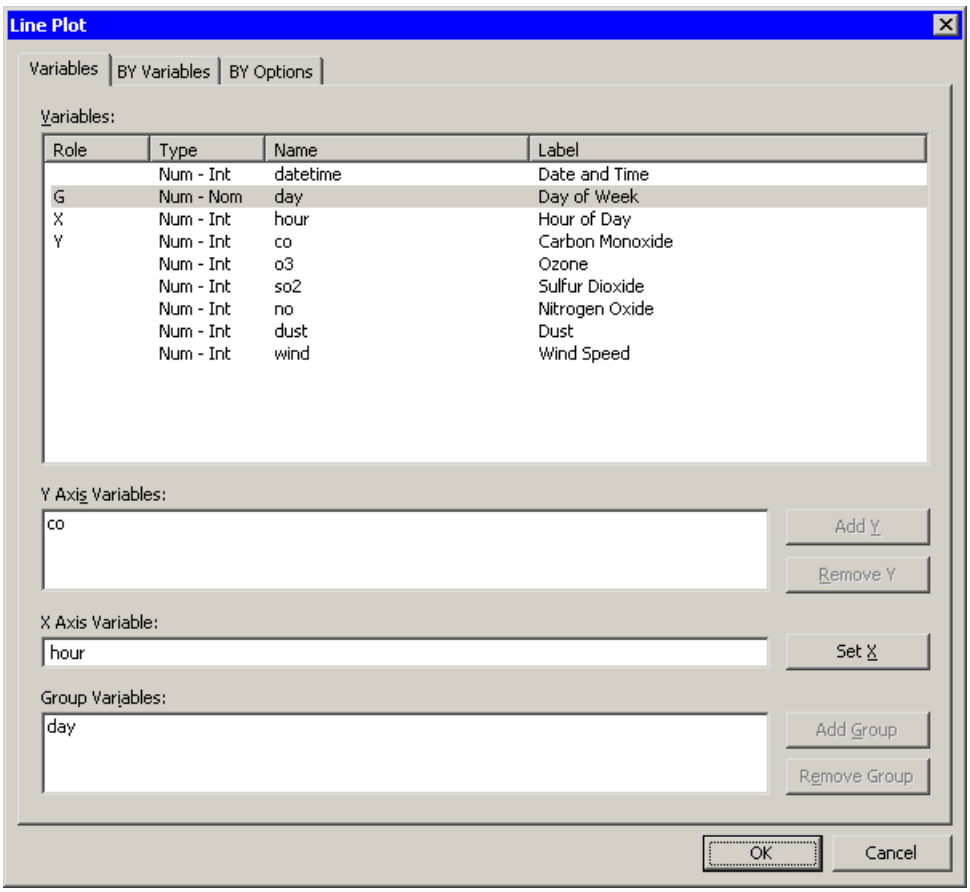
4 Select the variable co, and click **Add Y**.

5 Select the variable hour, and click **Set X**.

6 Select the variable day, and click **Add Group**.

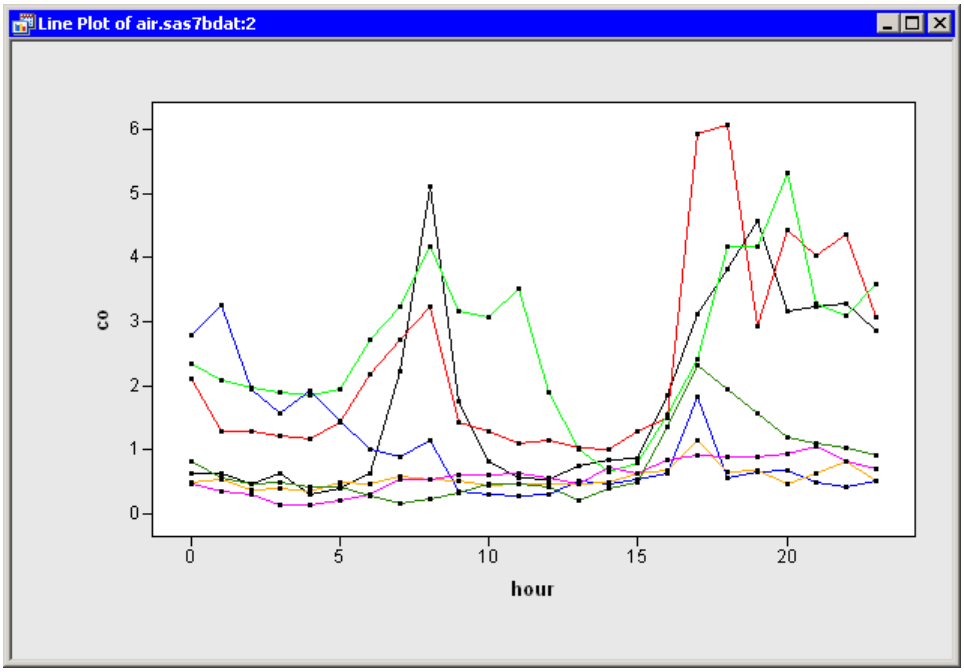
7 Click **OK**.

Figure 6.17 Specifying a Group Variable



The line plot that appears (Figure 6.18) has seven lines, one for each day of the week. For several days early in the week, the daily carbon monoxide peaked during the commuting times in the morning and evening: roughly 8 a.m. and 6–7 p.m.

Figure 6.18 A Line Plot with a Group Variable



To better visualize each day's carbon dioxide, you can use a bar chart to select each day individually.

8 Select **Graph ►Bar Chart** from the main menu.

The Bar Chart dialog box appears.

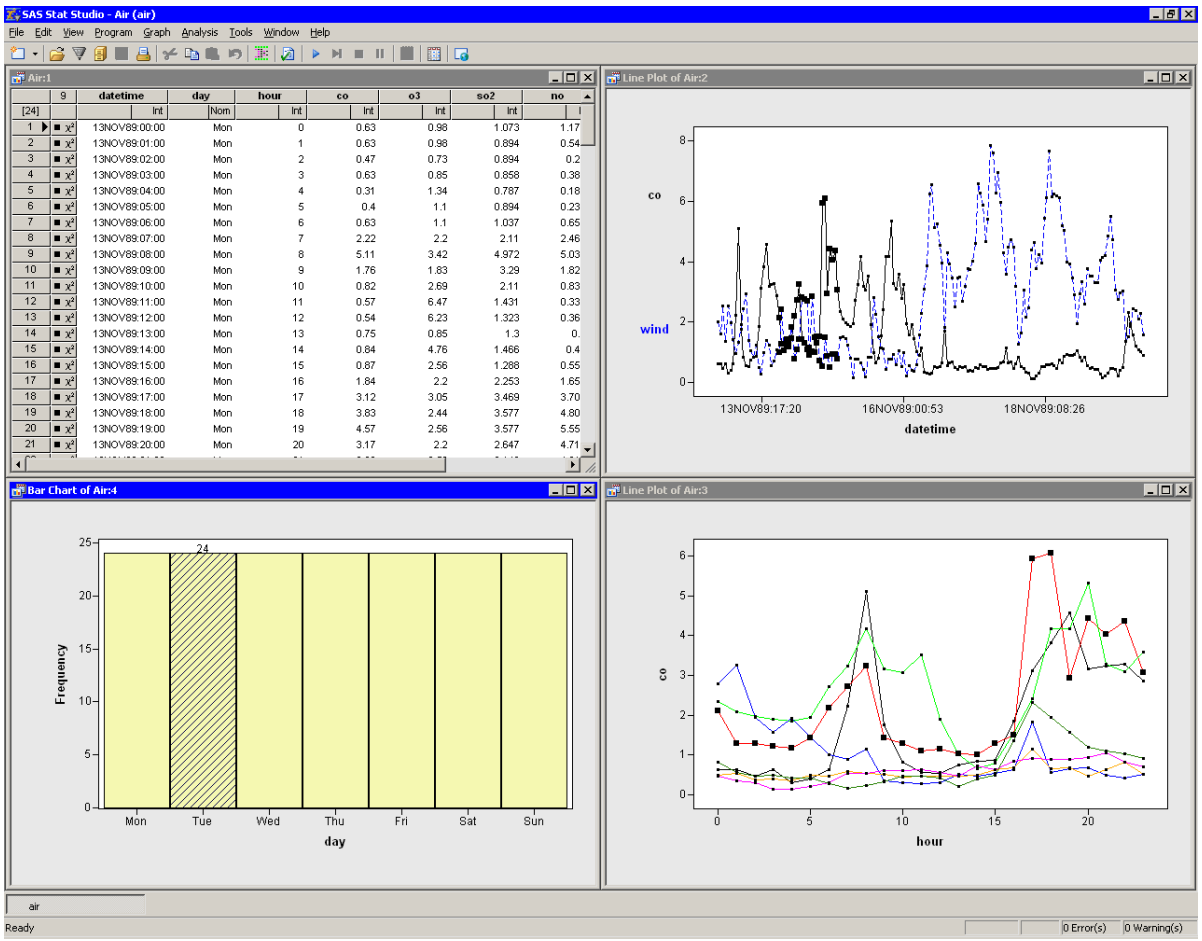
9 Select the day variable, and click **Set X**.

10 Click **OK**.

The resulting plots are shown in Figure 6.19.

You can now select each day of the week and examine the observations for that day.

Figure 6.19 Exploring Data for Each Day



Line Plot Properties

This section describes the **Line Styles** tab that is associated with a line plot. To access the line plot properties, right-click near the center of a plot, and select **Plot Area Properties** from the pop-up menu.

The **Line Styles** tab controls attributes of the lines displayed on a line plot. The **Line Styles** tab is shown in Figure 6.14.

The **Line Styles** tab contains the following UI controls:

Lines
displays each line in the plot. You can select one or more items in the list to change their properties.

⟸ (large left arrow)
applies the current set of properties to the lines selected in the **Lines** list. You must click the large left arrow to transfer the line attributes to the selected items in the **Lines** list.

Attributes: Style
sets the line style.

Attributes: Width
sets the line width.

Attributes: Color
sets the line color.

Attributes: Marker
sets the markers for the line. The default marker is the marker shown in the data table for each observation. Line markers are independent from observation markers.

Attributes: Connect observations
specifies whether the line connects adjacent observations with a line segment.

Attributes: Show observations
specifies whether observations are shown along the line.

Attributes: Sort observations
specifies whether observations along the line are sorted according to the value of the X variable.

For a discussion of the **Observations** tab, see the section "Scatter Plot Properties" on page 85. For a discussion of the remaining tabs, see Chapter 9, "General Plot Properties."

Line Plots of Selected Variables

If one or more variables are selected in a data table when you select **Graph ►Line Plot**, then the Line Plot dialog box does not appear. Instead, a line plot is created. The rules for constructing the line plot are as follows:

1. If one variable is selected, create a line plot of Y versus Y.
2. If exactly two variables are selected, the first is used as the Y variable and the second as the X variable.
3. If $k > 2$ variables are selected, then count the number of selected nominal variables.
   a) If no nominal variables are selected, create a line plot of $(Y_1, Y_2, \ldots, Y_{k-1})$ versus Y_k.
   b) If there are nominal variables selected, then count the number of selected interval variables.
      i. If no interval variables are selected, then plot the first selected variable as Y, plot the second selected variable as X, and use the remaining selected variables as group variables.
      ii. If there is exactly one interval variable, then plot it as Y if it was chosen first, and otherwise plot it as X. The first nominal variable is assigned to the X or Y role, and the remaining selected variables are used as group variables.
      iii. If there are exactly two interval variables, then plot the first selected interval variable as Y, plot the second as X, and use the remaining selected variables as group variables.
      iv. If there are more than two interval variables, then ignore the nominal variables and plot the interval variables as in rule 1.

Variables with a Frequency or Weight role are ignored when you are creating line plots.

Polygon Plots

This section describes how to use a polygon plot to visualize map data. A polygon plot displays polygons that are linked to levels of one or more categorical variables.

The polygon plot can display arbitrary polylines and polygons. To create a polygon plot, you need to specify at least three variables. The coordinates of vertices of each polygon (or vertices of a piecewise-linear polyline) are specified with X and Y variables. The polygon is drawn in the order in which the coordinates are specified. A third nominal variable specifies an identifier to which each coordinate belongs.

In some instances, a polygon is composed of subpolygons. For example, a continent is composed of countries, a country is composed of individual provinces or states, and some of those states are composed of disconnected landmasses (such as islands). The polygon plot supports this hierarchical structure by allowing multiple nominal variables that identify the continent, state, and island to which each coordinate pair belongs.

Example: Create a Polygon Plot

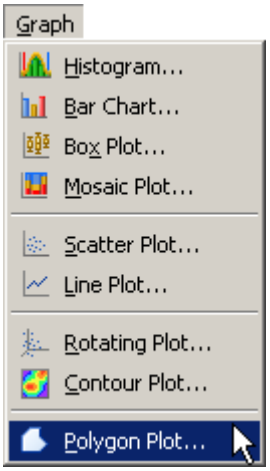
In this section you create a polygon plot of the lat and lon variables of the States48 data set. The lat variable gives the latitude of state boundaries for the lower 48 contiguous United States. The lon variable gives the corresponding longitude.

To create a polygon plot:

1 Open the States48 data set.

2 Select **Graph ►Polygon Plot** from the main menu, as shown in Figure 6.20.

Figure 6.20 Selecting a Polygon Plot



The Polygon Plot dialog box appears. (See Figure 6.21.)

3 Select the lon variable, and click **Set X**.

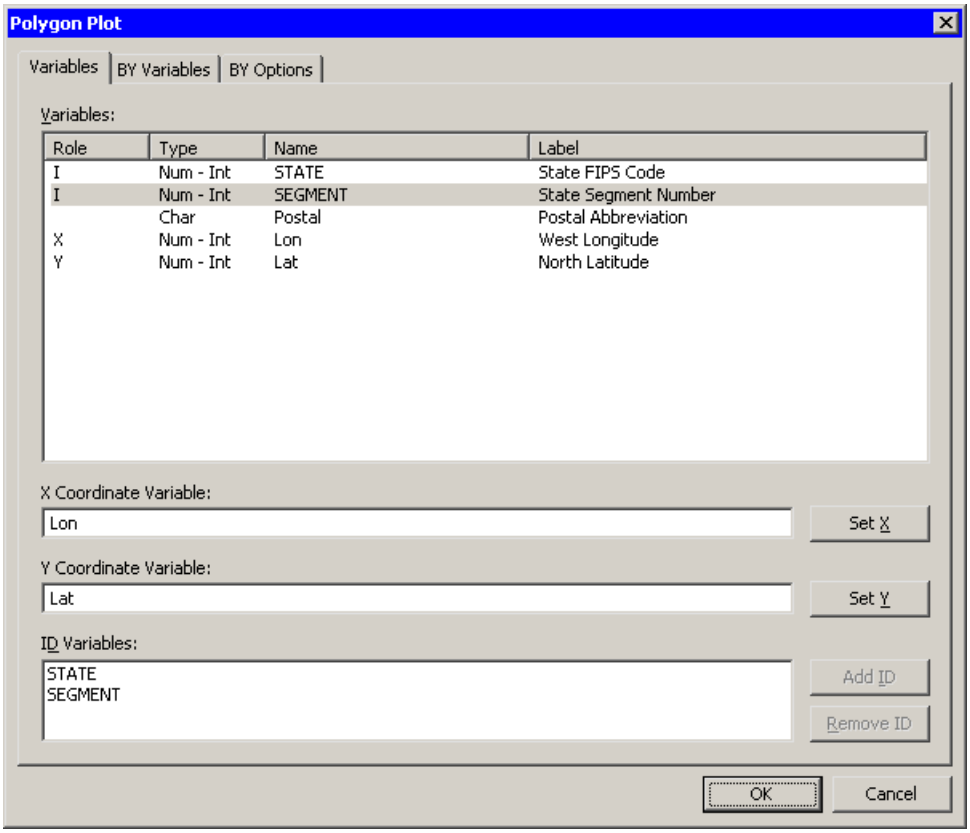
4 Select the lat variable, and click **Set Y**.

5 Select the state variable. Hold down the CTRL key and select the segment variable. Click **Add ID**.

6 Click **OK**.

NOTE: The order of the ID variables is important. The second variable should be nested in the first variable.

Figure 6.21 The Polygon Plot Dialog Box



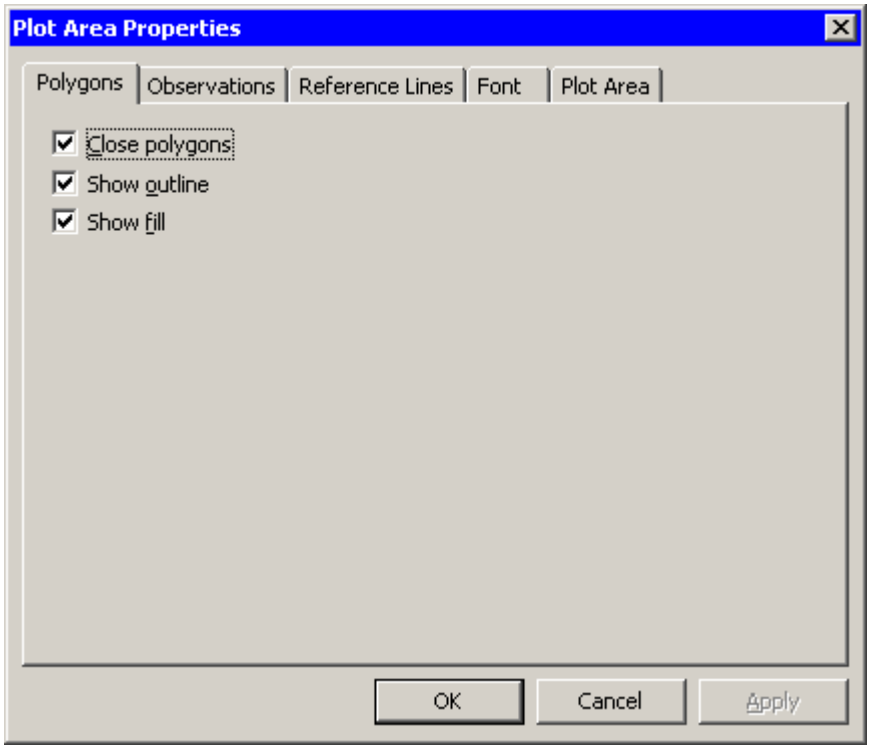
A polygon plot appears (similar to Figure 6.24) that shows the contiguous 48 United States. The color of a region (in this example, a state) is determined by the first observation encountered for that region. The observation's fill color determines the color of the interior of the polygon; the outline color determines the color of the region's outline.

For these data, the observations are all black. To make the polygon plot look more like a map, you can color observations by the value of the state variable.

7 Right-click near the center of the plot, and select **Plot Area Properties** from the pop-up menu.

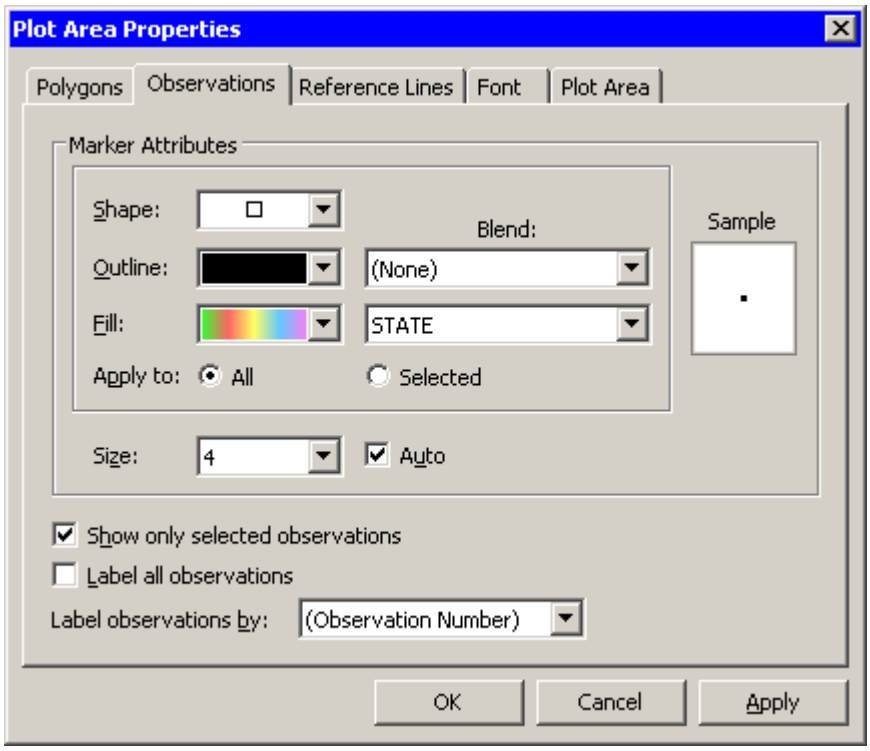
The Plot Area Properties dialog box appears. (See Figure 6.22.)

Figure 6.22 Plot Area Properties for a Polygon Plot



8 Click the **Observations** tab, as shown in Figure 6.23.

Figure 6.23 The Observations Tab



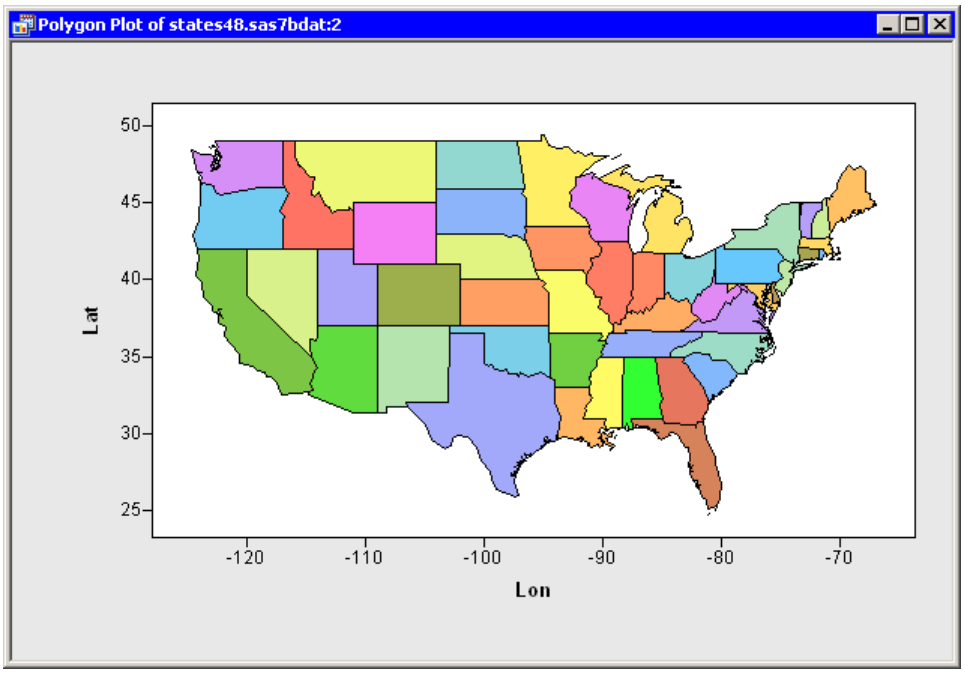
9 Select state from the **Fill: Blend** menu.

10 Select a gradient color map from the **Fill** menu.

11 Click **OK**.

The polygon plot is now colored according to your choice of color map. (See Figure 6.24.)

Figure 6.24 A Polygon Plot



The polygon plot supports the selection of polygonal regions. For example, you can click a state to select the observations that define the boundary of that state. You can click while holding down the CTRL key to select observations that define multiple states. You can also draw a selection rectangle to select observations that define contiguous states.

If a state is composed of two or more components, you can click each component independently. For example, you can select just the upper peninsula of Michigan, or select only Long Island, New York. You can also color each region independently.

Polygon Plot Properties

This section describes the **Polygons** tab associated with a polygon plot. To access the polygon plot properties, right-click near the center of a plot, and select **Plot Area Properties** from the pop-up menu.

The **Polygons** tab controls attributes of the polygon plot. The **Polygons** tab is shown in Figure 6.22.

The **Polygons** tab contains the following UI controls:

Close polygons
specifies whether a line segment is drawn from the last observation in each region to the first observation in that region.

Show outline
specifies whether the outline of a region is displayed.

Show fill
specifies whether the interior of a region is displayed.

For a discussion of the **Observations** tab, see the section "Scatter Plot Properties" on page 85. For a discussion of the remaining tabs, see Chapter 9, "General Plot Properties."

Polygon Plots of Selected Variables

If variables are selected in a data table when you select **Graph ►Polygon Plot**, then the polygon plot dialog box does not appear. The first selected interval variable is used for the X variable; the second is used for the Y variable. Any interval variables after the second variable are ignored. Any nominal variables are assigned the ID role in the order in which they were selected.

Variables with a Frequency or Weight role are ignored when you are creating polygon plots.

Chapter 7
Exploring Data in Three Dimensions

Contents

Overview of Exploring Data in Three Dimensions **103**
Rotating Plots **103**
Example: Create a Rotating Scatter Plot 103
Example: Create a Rotating Surface Plot 110
Rotating Plot Properties 113
Rotating Plots of Selected Variables 116
Contour Plots **117**
Example: Create a Contour Plot 118
Contour Plot Properties 126
Contour Plots of Selected Variables 129

Overview of Exploring Data in Three Dimensions

This chapter describes how to use SAS/IML Studio to examine relationships among three variables.

You can explore the relationships among three variables by using a rotating scatter plot. Often the three variables are interval variables.

If one of the variables can be modeled as a function of the other two variables, then you can add a response surface to the rotating plot. Similarly, you can visualize contours of the response variable by using a contour plot.

Rotating Plots

This section describes how to use a rotating plot to visualize the relationships among three variables. Often each variable is continuous (interval), but that is not a requirement.

Example: Create a Rotating Scatter Plot

In this section you create a rotating plot to explore the relationships among the wind_kts, latitude, and longitude variables of the Hurricanes data set. The wind_kts variable gives the wind speed in knots for each observation.

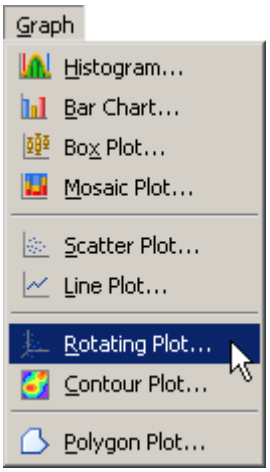
None of the variables in this example have missing values. If an observation has a missing value for any of the three variables in the rotating plot, that observation is not plotted.

To create a rotating plot:

1 Open the Hurricanes data set.

2 Select **Graph ►Rotating Plot** from the main menu, as shown in Figure 7.1.

Figure 7.1 Selecting a Rotating Plot



The Rotating Plot dialog box appears. (See Figure 7.2.)

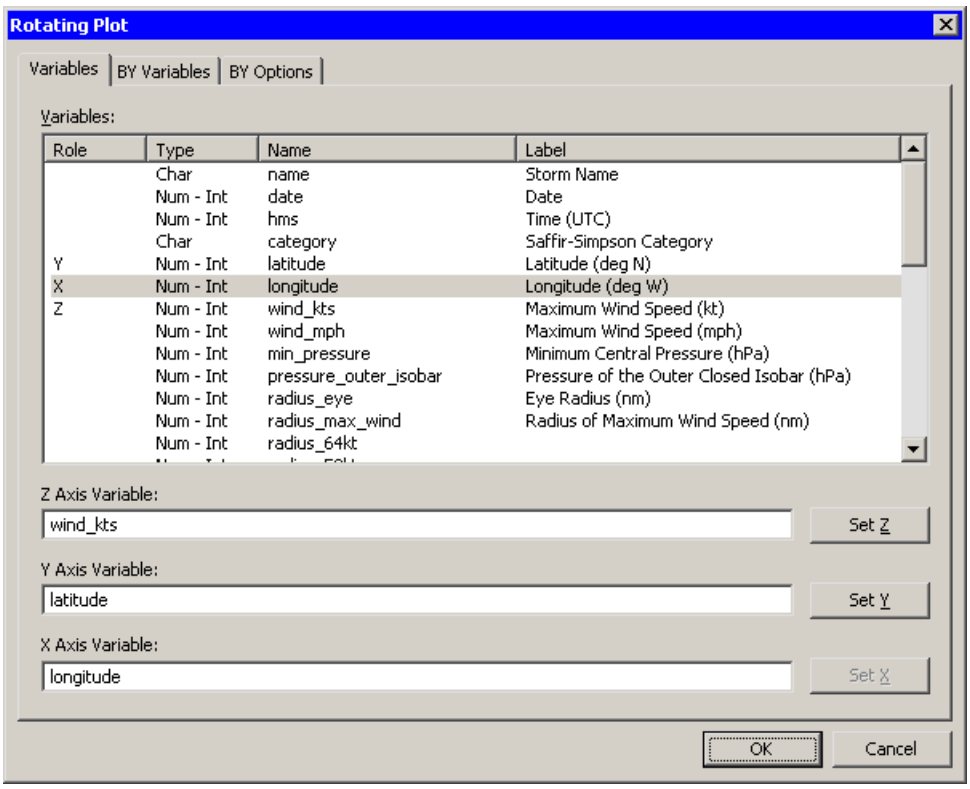
3 Select the wind_kts variable, and click **Set Z**.

4 Select the latitude variable, and click **Set Y**.

5 Select the longitude variable, and click **Set X**.

6 Click **OK**.

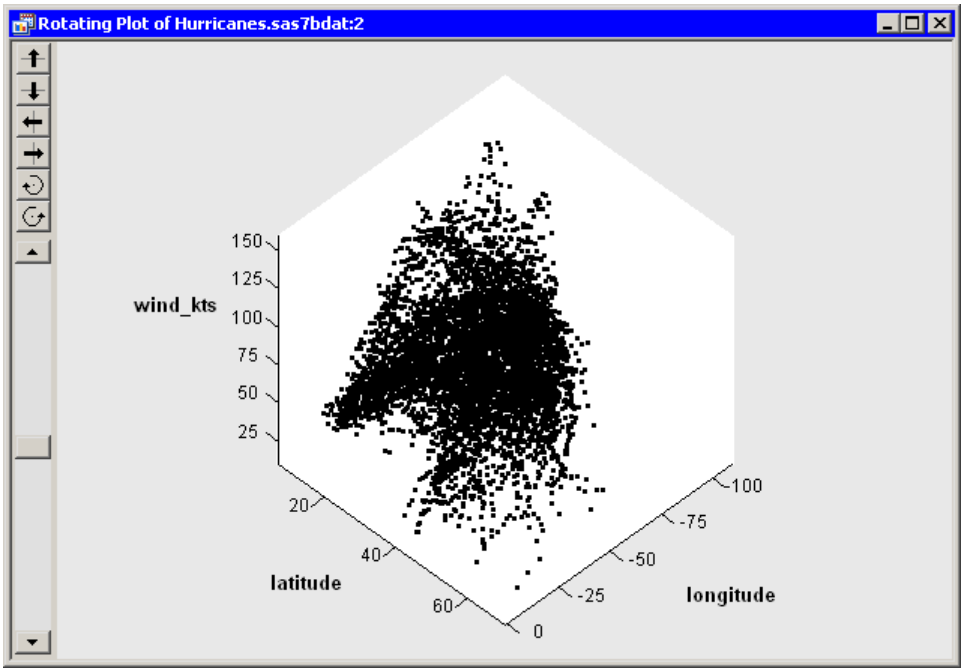
Figure 7.2 The Rotating Plot Dialog Box



A rotating plot appears (Figure 7.3), which shows a cloud of points. You can rotate the plot by clicking the icons on the left side of the plot. The top two buttons rotate the plot about a horizontal axis. The next two buttons rotate the plot about a vertical axis. The last two buttons rotate the plot clockwise and counterclockwise. The slider below the buttons controls the speed of rotation.

Alternatively, you can rotate the plot by moving the mouse pointer into a corner of the plot until the pointer changes (to ↺). You can interactively rotate the plot by holding down the left mouse button while you move the mouse.

Figure 7.3 A Rotating Plot



You can click an observation in a rotating plot to select the observation. You can click while holding down the CTRL key to select multiple observations. You can also draw a selection rectangle to select multiple observations.

You can create rotating plots of any variables, numeric or character.

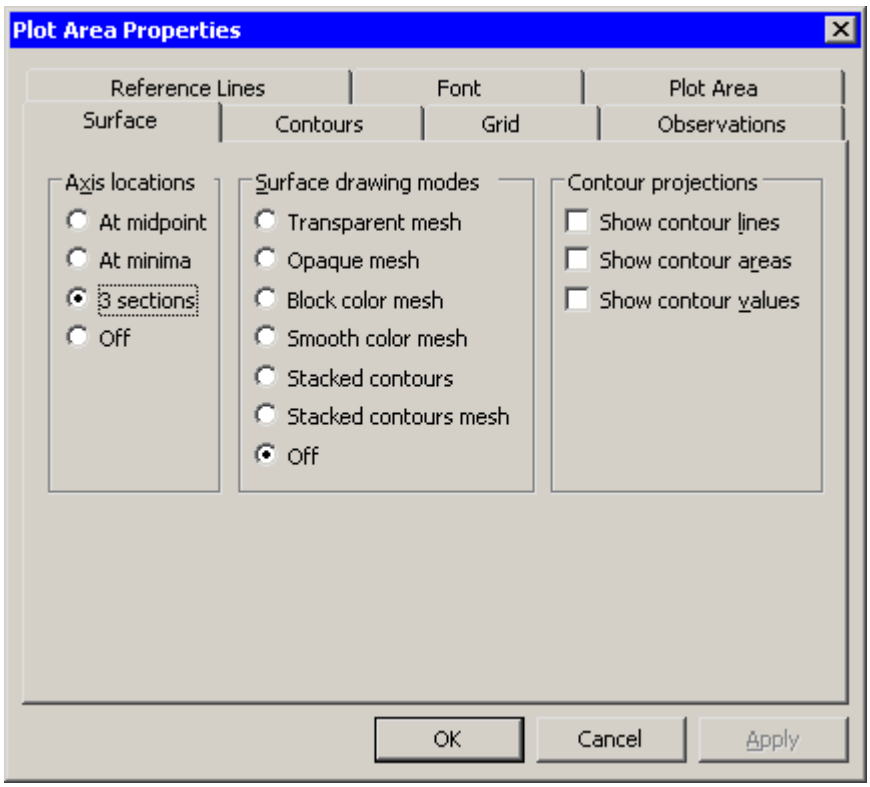
Because there are so many observations in the rotating plot, some observations obscure others—a phenomenon known as *overplotting*. It also can be difficult to discern the coordinates of observations as they are positioned in three-dimensional space. That is, which observations are "closer" to the viewer?

A visualization technique that sometimes helps distinguish observations with similar projected coordinates is to color the observations. For these data, you can color the observations according to the wind_kts variable.

7 Right-click near the center of the plot, and select **Plot Area Properties** from the pop-up menu.

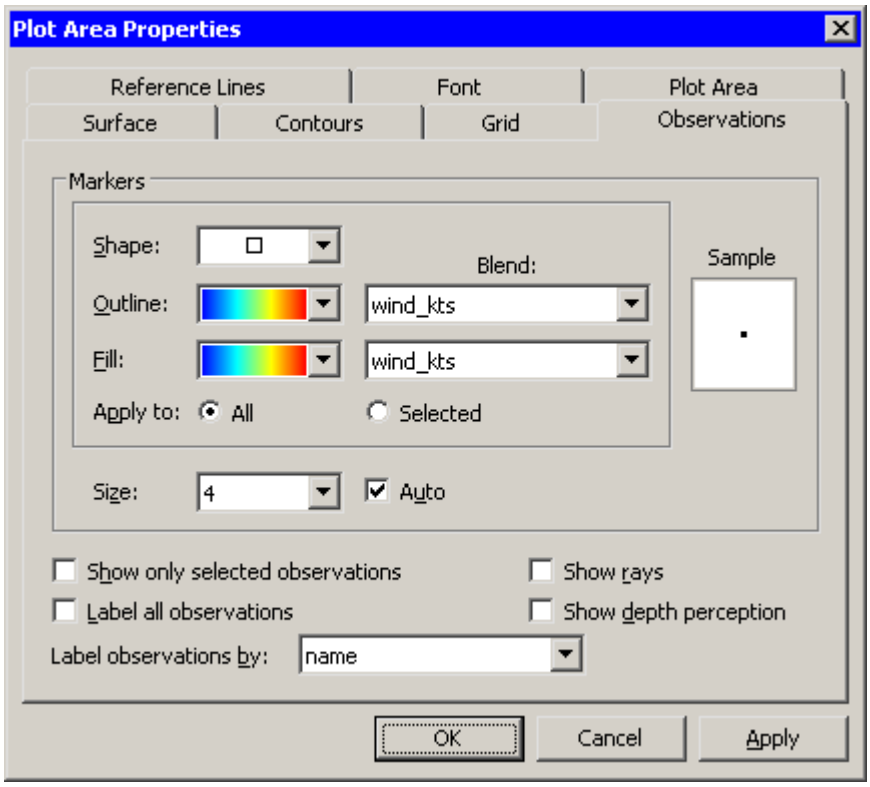
The dialog box shown in Figure 7.4 appears.

Figure 7.4 Plot Area Properties for a Rotating Plot



8 Click the **Observations** tab, as shown in Figure 7.5.

Figure 7.5 Observations Tab for a Rotating Plot



9 Select wind_kts from the **Outline: Blend** list.

10 Select a gradient color map from the **Outline** list.

11 Select the same options for the **Fill: Blend** and **Fill** lists.

12 Select name from the **Label observations by** list.

This step specifies that the name of the cyclone should appear when you click an observation. By default the observation number is used as a label.

13 Click **Apply** to update the plot with the options you have selected so far.

You can optionally use two additional features to aid in visualizing these data.

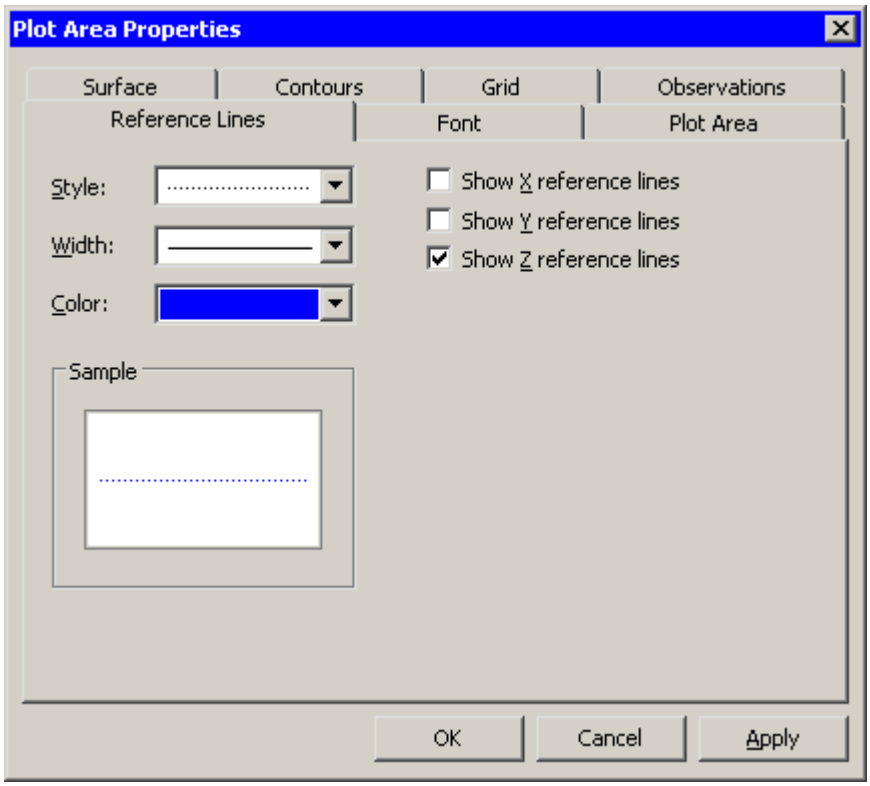
14 Click the **Reference Lines** tab, shown in Figure 7.6.

15 Select **Show Z reference lines**.

16 Click **Apply**.

When you click **Apply**, the plot updates to show reference lines at each tick on the axis for the Z variable (in this case, wind_kts). The reference lines are displayed in Figure 7.8.

Figure 7.6 Reference Lines Tab for a Rotating Plot



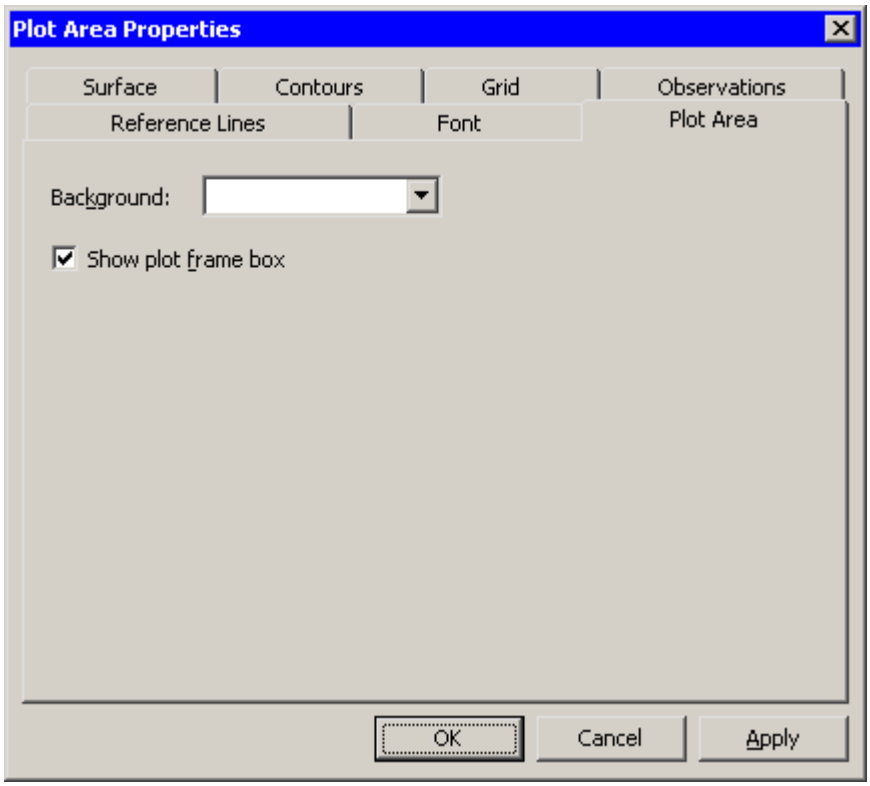
17 Click the **Plot Area** tab, shown in Figure 7.7.

18 Select **Show plot frame box**.

19 Click **OK**.

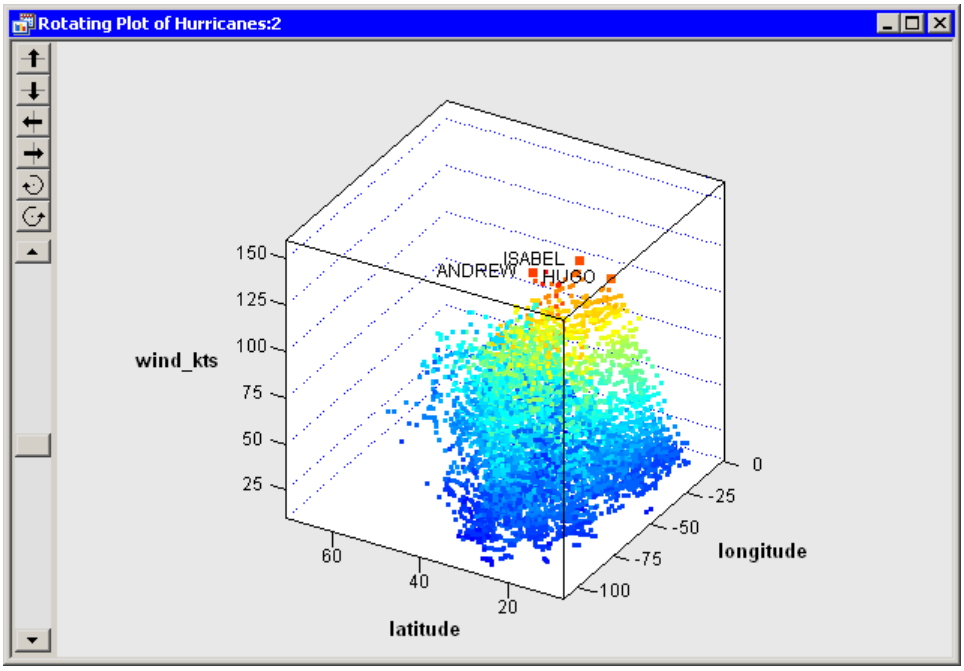
The rotating plot updates to reflect the options you selected. (See Figure 7.8.) You can rotate the plot to observe how wind speeds in these tropical cyclones vary according to latitude and longitude. You can click interesting observations and see the name of the storms they represent.

Figure 7.7 Plot Area Tab for a Rotating Plot



You can see that the storms with the strongest winds tend to occur west of 45 degrees west latitude, and roughly between 12 and 32 degrees north latitude. You can also see that many cyclones begin in southern latitudes where they move west or northwest, then later they turn north and northeast as they approach higher latitudes. The wind speed along a track tends to increase over warm water and decrease over land or cooler water.

Figure 7.8 A Rotating Plot with Selected Observations



Example: Create a Rotating Surface Plot

In the previous example you created a rotating scatter plot. A rotating scatter plot does not presume any relationship between the Z variable and the X and Y variables.

In this section you create a rotating plot in which you assume that the Z variable is functionally related to the X and Y variables. That is, the Z variable can be modeled as a response variable of X and Y.

A typical use of the rotating surface plot is to visualize the response surface for a regression model of two continuous variables. If you model a response variable by using an analysis chosen from the **Analysis ►Model Fitting** menu, you can add the predicted values of the model to the data table. Then you can plot the predicted values as a function of the two regressor variables.

In this example you examine three variables in the Climate data set. You explore the functional relationship between the elevationFeet variable and the latitude and longitude variables. The elevationFeet variable gives the elevation in feet above mean sea level for each of 40 cities in the continental United States.

To create a rotating plot and add a surface:

1 Open the Climate data set.

2 Select **Graph ►Rotating Plot** from the main menu.

The Rotating Plot dialog box appears. (See Figure 7.9.)

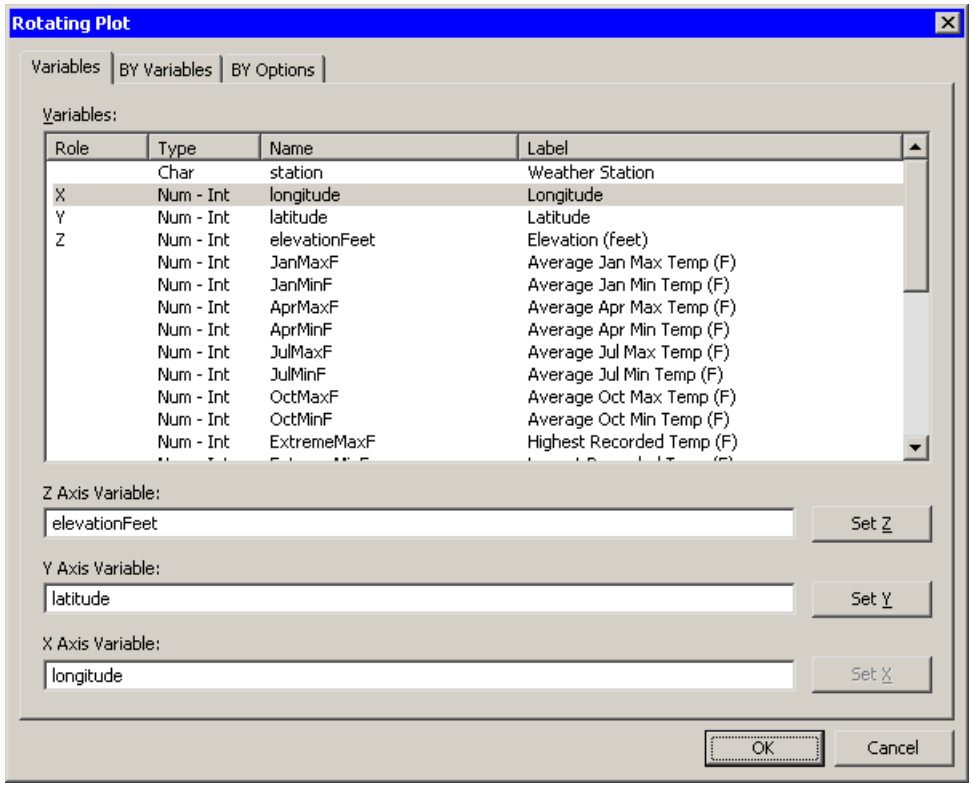
3 Select the elevationFeet variable, and click **Set Z**.

4 Select the latitude variable, and click **Set Y**.

5 Select the longitude variable, and click **Set X**.

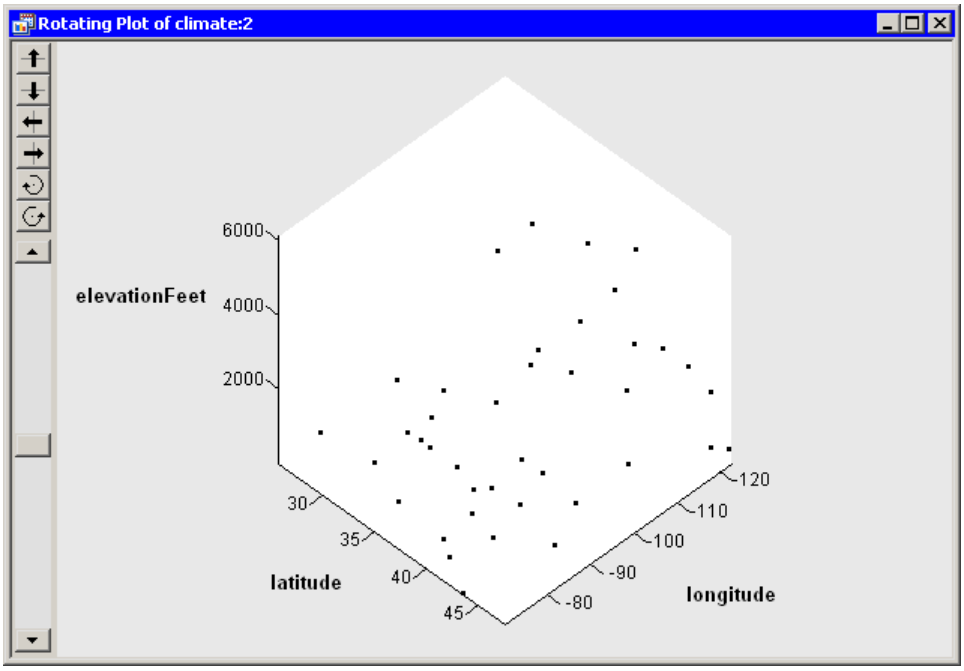
6 Click **OK**.

Figure 7.9 The Rotating Plot Dialog Box



A rotating plot appears (Figure 7.10), which shows a cloud of points. You can rotate the plot as explained in the previous example.

Figure 7.10 A Rotating Plot



You can visualize elevation as a function over longitude and latitude by adding a surface to these data.

7 Right-click near the center of the plot, and select **Plot Area Properties** from the pop-up menu.

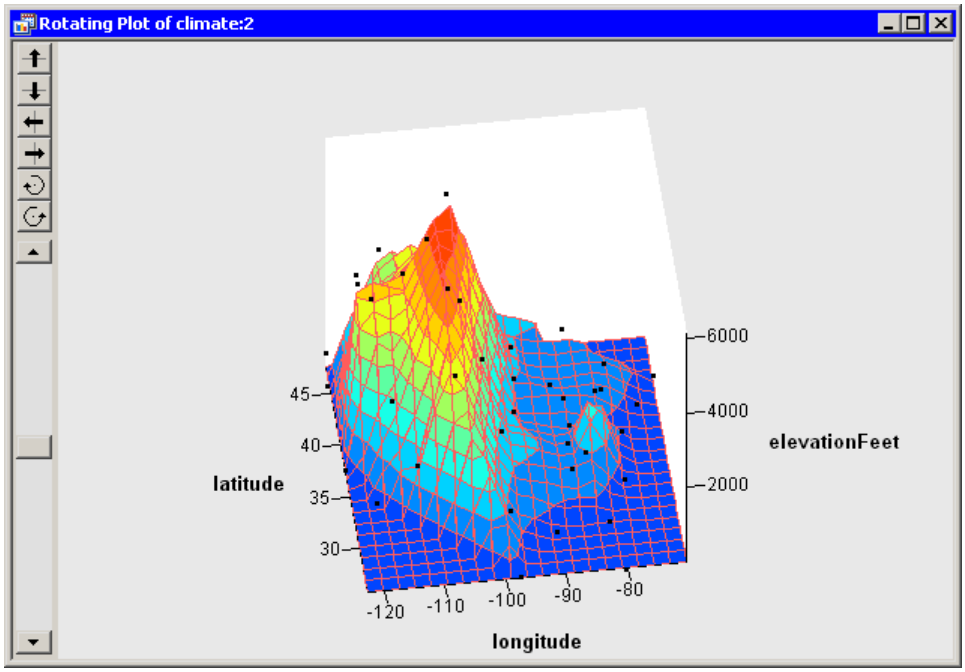
8 Select **Smooth color mesh** from the group of radio buttons labeled **Surface drawing modes**.

9 Click **OK**.

The rotating plot updates to show a rough approximation to an elevation map of the continental United States. (See Figure 7.11.) There are only 40 data points in the plot, so the surface map is understandably coarse. Having more data points distributed uniformly across the country would result in a surface that is a better approximation of actual elevations.

Nevertheless, the surface helps you to identify cities near the Rocky Mountains with high elevations (Cheyenne, WY, and Albuquerque, NM), one city in the Appalachian Mountains (Asheville, NC), and the coastal cities.

Figure 7.11 A Rotating Plot



NOTE: You can add a surface to any rotating scatter plot, but you should first determine whether it is appropriate to do so. Surface plots might not be appropriate for data with replicated measurements. Surface plots of highly correlated data can be degenerate.

Rotating Plot Properties

This section describes the property tabs that are associated with a rotating plot. To access the rotating plot properties, right-click near the center of a plot, and select **Plot Area Properties** from the pop-up menu.

Tabs not discussed in this chapter are discussed in Chapter 9, "General Plot Properties."

Surface Tab

The **Surface** tab controls attributes of the rotating plot. (See Figure 7.4.) You can use the **Surface** tab to control the placement of axes, the type of surface that is drawn, and whether contours of the data are shown in the (X,Y) plane. The **Surface** tab contains the following UI controls:

Axis Locations
specifies the location of axes.

At midpoint specifies that the origin of each axis is placed at the midpoint of the range of the variable for that axis.

At minima specifies that the origin of each axis is placed at the minimum value of the variable for that axis.

3 sections specifies that each axis is placed on an edge of the bounding cube that surrounds the data so that the axis interferes as little as possible with viewing the data.

Off specifies that no axes are displayed.

Surface Drawing Modes
specifies the attributes of the surface that is added to the rotating plot.

Transparent mesh specifies that the surface is drawn as a wire mesh without removing hidden lines.

Opaque mesh specifies that the surface is drawn as a wire mesh and hidden lines are removed.

Block color mesh specifies that the surface is drawn as a patch of rectangles in which each rectangle is a single color.

Smooth color mesh specifies that the surface is drawn as a patch of triangles in which each triangle is a single color.

Stacked contours specifies that the surface is not drawn, but that contour levels are drawn.

Stacked contours mesh specifies that the surface is drawn as for the **Opaque mesh** option and also that contour levels are added as for the **Stacked contours** option.

Off specifies that no surface is displayed.

Contour Projections
specifies whether contours of the data are shown in the (X,Y) plane.

Show contour lines specifies that contours for the surface are shown projected onto the (X,Y) plane.

Show contour areas specifies that region between projected contours are filled with color.

Show contour values specifies whether projected contour lines are labeled by the value of the Z axis variable.

Contours Tab

The **Contours** tab controls attributes of the projected contours for the surface. The **Contours** tab is described in the section "Contour Plot Properties" on page 126.

Grid Tab

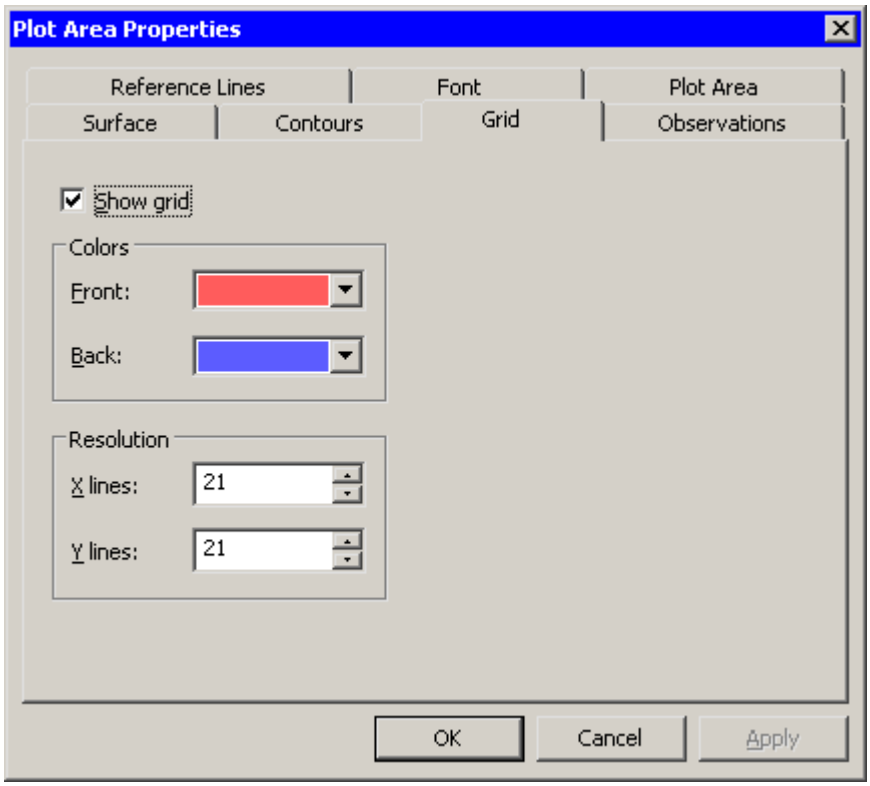
The **Grid** tab controls the size and color of the grid used to construct a surface and to compute contours for the surface. (See Figure 7.12.) The **Grid** tab contains the following UI controls:

Show grid
specifies whether to display a grid on the displayed surface.

Colors
specifies the color of the grid when seen from the front (positive Z direction) or back (negative Z direction).

Resolution
specifies the resolution of the computational grid used to fit a surface to the data. The algorithm that computes the surface uses a grid superimposed on the (X,Y) plane. This grid consists of evenly spaced subdivisions along the X and Y axes. Generally, having more subdivisions results in a smoother surface, whereas having fewer subdivisions results in a rougher surface.

Figure 7.12 The Grid Tab



Observations Tab

The **Observations** tab controls the attributes of markers in the rotating plot. (See Figure 7.5.)

The **Observations** tab for the rotating plot contains all the controls documented in the section "Scatter Plot Properties" on page 85. In addition, the **Observations** tab for the rotating plot includes the following check boxes:

Show rays
: specifies whether lines are drawn from the center of the bounding cube to each observation marker.

Show depth perception
: specifies whether observation markers are drawn in varying sizes to indicate their distance from the viewer.

Reference Lines Tab

The **Reference Lines** tab controls the attributes of reference lines in the rotating plot. (See Figure 7.6.)

The **Reference Lines** tab for the rotating plot contains all the controls documented in the section "Reference Lines Tab" on page 158. In addition, the **Reference Lines** tab for the rotating plot includes the following check box:

Show Z reference lines
: specifies whether to show reference lines for the Z axis.

Plot Area Tab

The **Plot Area** tab controls the attributes of plot area in the rotating plot. (See Figure 7.7.) The **Plot Area** tab contains the following UI controls:

Background
: specifies the color of the background of the plot area.

Show plot frame box
: specifies whether to display a framing box surrounding the plot area.

Rotating Plots of Selected Variables

If one or more variables are selected in a data table when you select **Graph ►Rotating Plot**, then the Rotating Plot dialog box does not appear. Instead rotating plots are created of the selected variables.

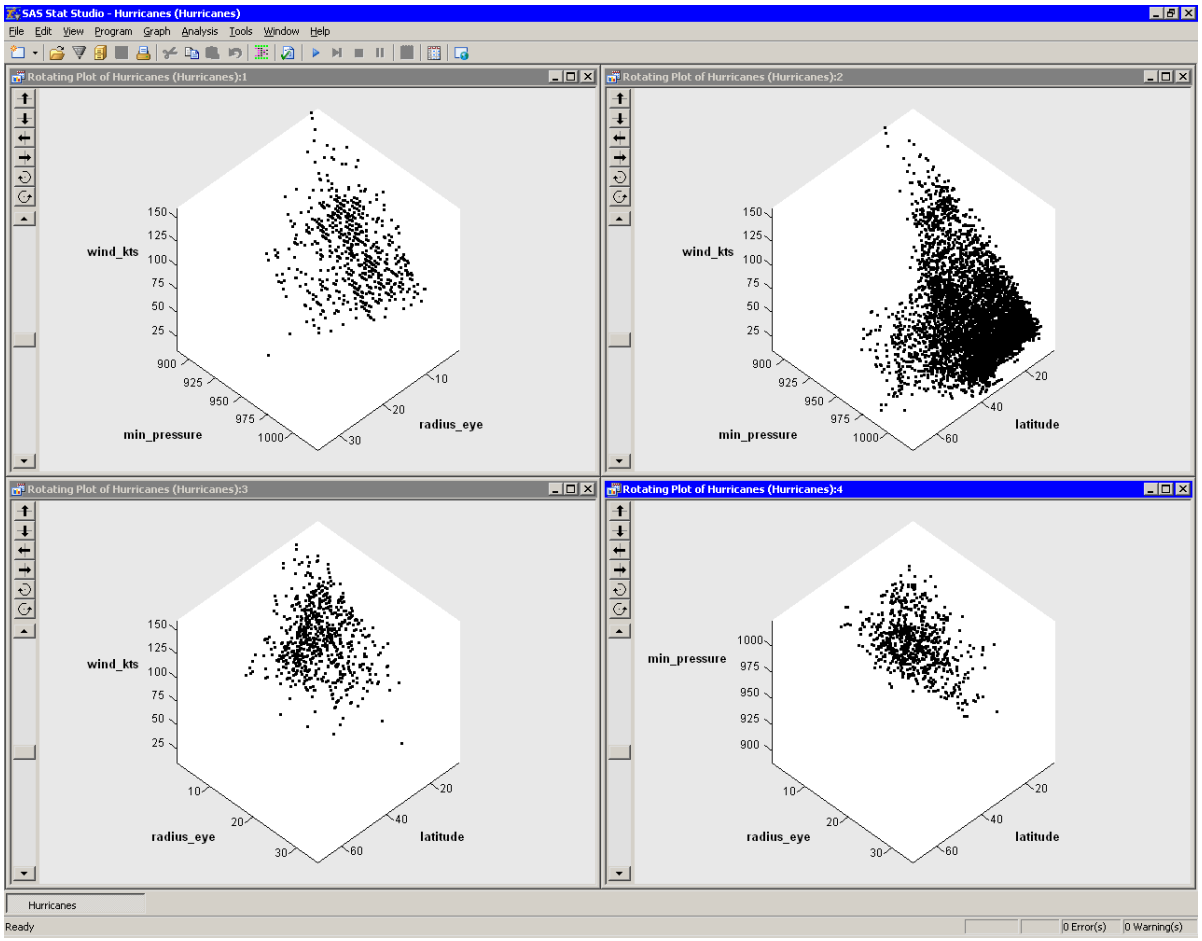
All threefold combinations of the selected variables are plotted. That is, if you select $n \geq 3$ variables, then you see a matrix of $\binom{n}{3}$ rotating plots.

If you create a matrix of plots from selected variables, you can close the matrix by pressing the F11 key while any plot is active and selecting from the pop-up menu. Alternatively, you can use the Workspace Explorer to quickly close plots. (See the section "Workspace Explorer" on page 190.)

Variables with a Frequency or Weight role are ignored when you are creating rotating plots.

Figure 7.13 shows a matrix of rotating plots for four selected variables: wind_kts, min_pressure, radius_eye, and latitude.

Figure 7.13 A Matrix of Rotating Plots



Contour Plots

In this section you create a contour plot. A contour plot assumes that the Z variable is functionally related to the X and Y variables. That is, the Z variable can be modeled as a response variable of X and Y.

A typical use of a contour plot is to visualize the response for a regression model of two continuous variables. If you model a response variable by using an analysis chosen from the **Analysis ►Model Fitting** menu, you can add the predicted values of the model to the data table. Then you can create a contour plot of the predicted values as a function of the two regressor variables.

Contour plots are most useful when the X and Y variables are nearly uncorrelated. The contour plot fits a piecewise-linear surface to the data, which models Z as a response function of X and Y. The contours are level curves of the response function. By default, the minimum and maximum values of the Z variable are used to compute the contour levels.

The three variables in a contour plot must be interval variables.

Example: Create a Contour Plot

In this example you examine three variables in the Climate data set. You explore the functional relationship between the elevationFeet variable and the latitude and longitude variables. The elevationFeet variable gives the elevation in feet above mean sea level for each of 40 cities in the continental United States.

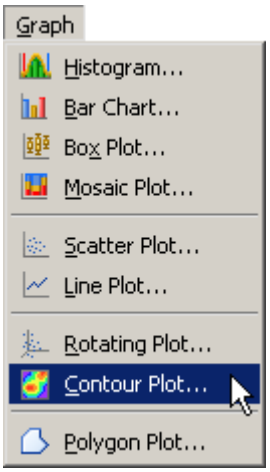
None of the variables in this example have missing values. If an observation has a missing value for any of the three variables in the contour plot, that observation is not plotted.

To create a contour plot:

1 Open the Climate data set.

2 Select **Graph ►Contour Plot** from the main menu, as shown in Figure 7.14.

Figure 7.14 Selecting a Contour Plot



The Contour Plot dialog box appears. (See Figure 7.15.)

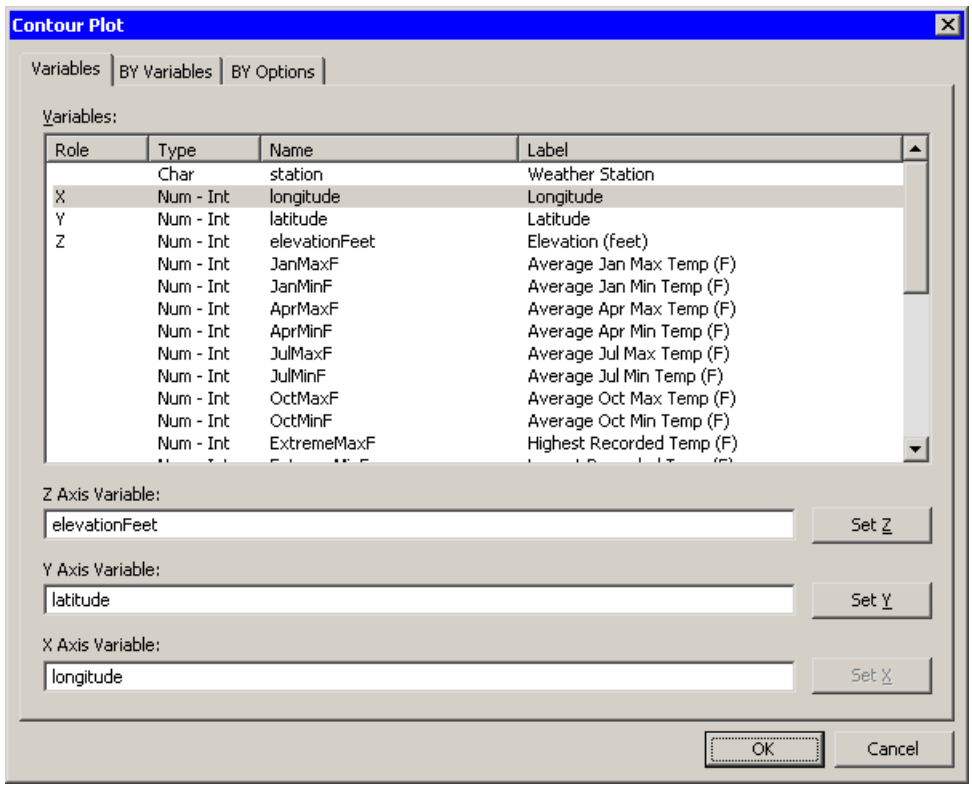
3 Select the elevationFeet variable, and click **Set Z**.

4 Select the latitude variable, and click **Set Y**.

5 Select the longitude variable, and click **Set X**.

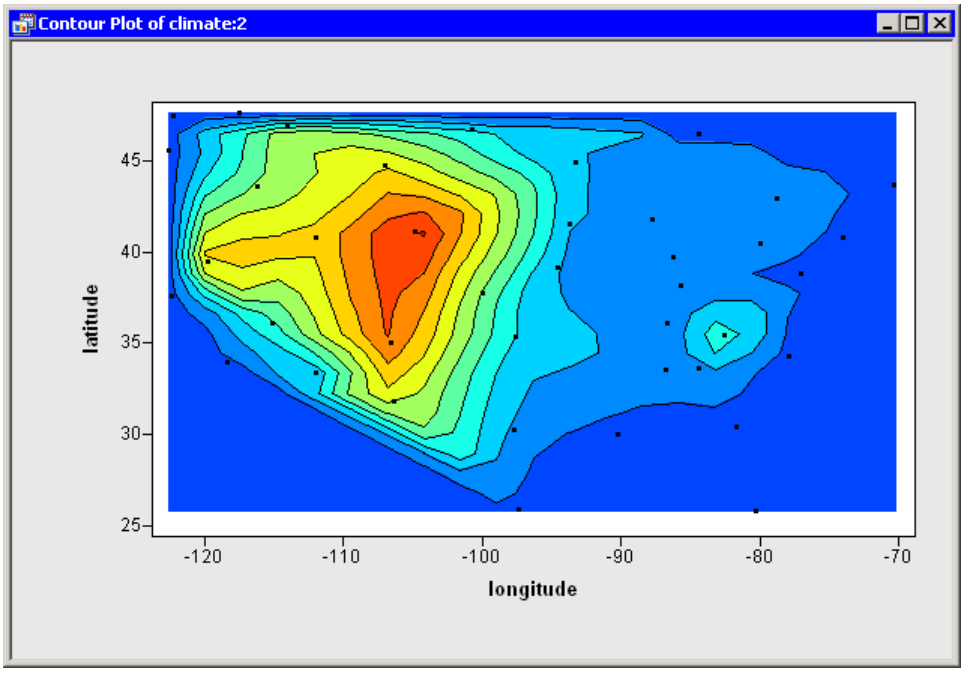
6 Click **OK**.

Figure 7.15 A Contour Plot Dialog Box



A contour plot appears (Figure 7.16), which shows a scatter plot of the longitude and latitude variables. Contours of the elevationFeet variable are shown overlaid on the scatter plot.

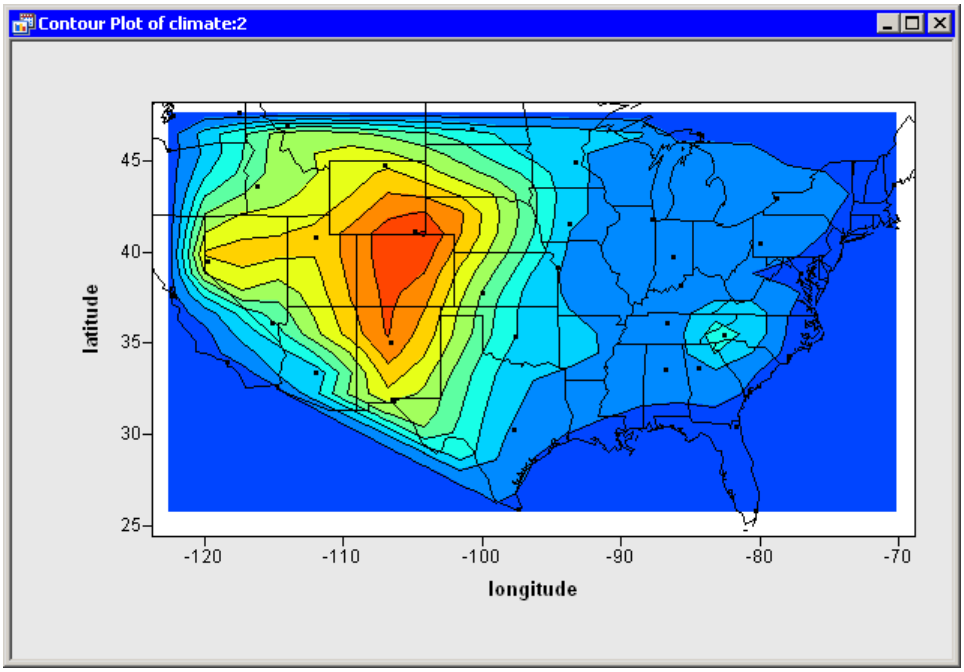
Figure 7.16 A Contour Plot



You can double-click an observation to display the variable values that are associated with that observation. (See the section "Observation Inspector" on page 139 for further details.) In this way, you can identify cities and find out their exact elevations.

It is somewhat difficult to guess where the state boundaries are in Figure 7.16, so Figure 7.17 overlays the outline of the continental United States onto the contour plot. The figure was created by using the DrawPolygonsByGroups module, which is documented in the SAS/IML Studio online Help chapter titled "IMLPlus Module Reference."

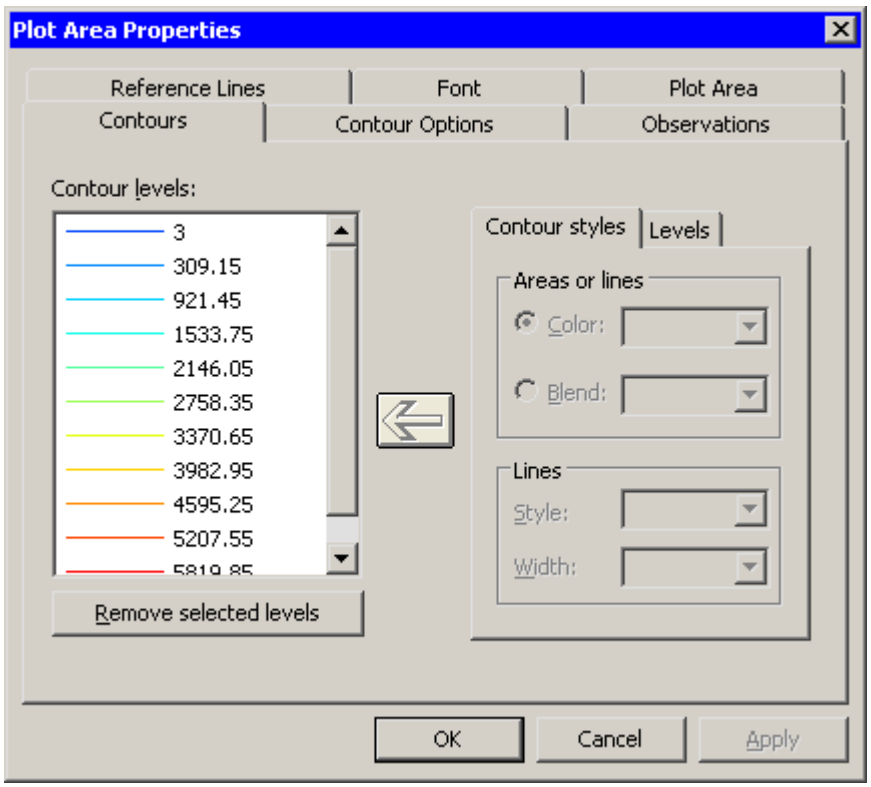
Figure 7.17 A Contour Plot



NOTE: You can create a contour plot of any three continuous variables, but you should first determine whether it is appropriate to do so. Contour plots might not be appropriate for data with replicated measurements or for data with highly correlated X and Y variables.

If you display the contour plot property dialog box, you can examine the values that are associated with each contour. (To display plot properties, right-click near the center of a plot, and select **Plot Area Properties** from the pop-up menu.) Figure 7.18 shows that there are 10 evenly spaced contours in the range of the elevationFeet variable. The minimum and maximum values of elevationFeet are 3 and 6126.

Figure 7.18 Default Contours



Example: Change Contour Values and Appearance

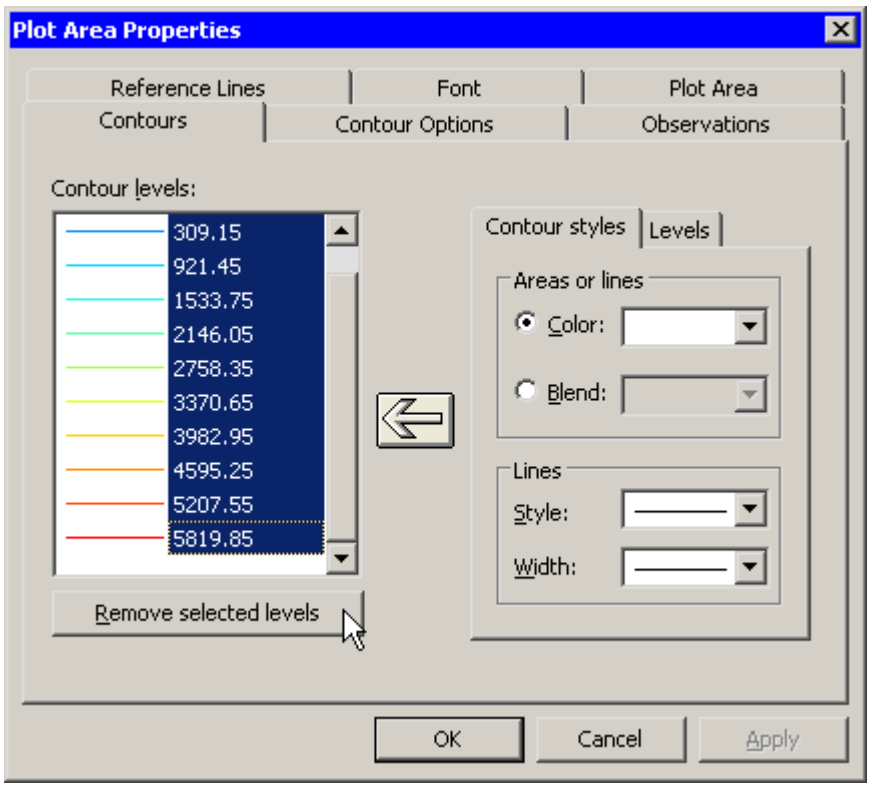
The default contours are usually adequate for obtaining a qualitative feel for the response surface. However, sometimes you might want to manually specify the levels of the contours. You might need to conform to some standard (for example, 50-meter contour intervals) or include a critical level (for example, a control limit).

Suppose you decide that you want the contour levels of elevationFeet to be "round numbers," such as multiples of 100. You can change the set of contours by removing the old contours, adding new contours, and coloring the new contours.

To remove the old contours:

1. Select the first contour (labeled "3"). Scroll the **Contour Levels** list to the last contour. Hold down the SHIFT key while clicking the last contour (labeled "5819.85") to select all contours in the list.
2. Click **Remove selected levels**, as shown in Figure 7.19.

Figure 7.19 Removing Contours



To add a new set of uniformly spaced contours, do the following:

3 Click the **Levels** subtab. (See Figure 7.20.)

4 Type `10` in the **Number** field.

5 Type `0` in the **Minimum** field.

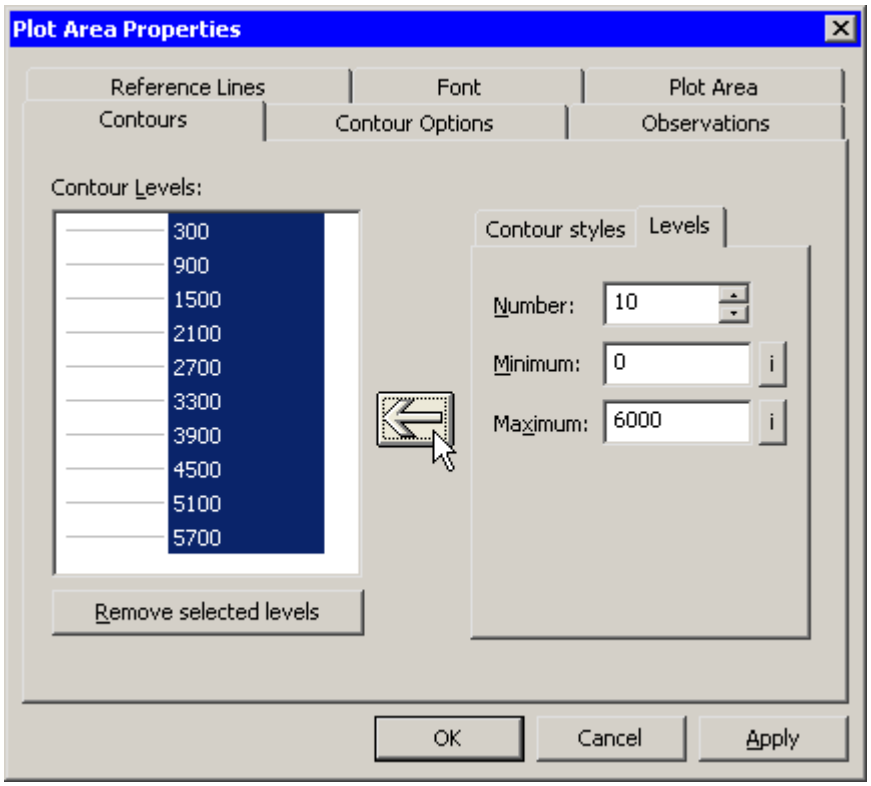
The value for this field is typically a “round number” near the minimum value of the Z variable.

6 Type `6000` in the **Maximum** field.

The value for this field is typically a “round number” near the maximum value of the Z variable.

7 Click the large left arrow (⟸) to create the contours, as shown in Figure 7.20.

Figure 7.20 Adding Evenly Spaced Contours



The **Contour Levels** list is filled with the values 300, 900, . . . , 5700. These values do not include the minimum and maximum specified values (0 and 6000), because contours at the extreme values are often degenerate.

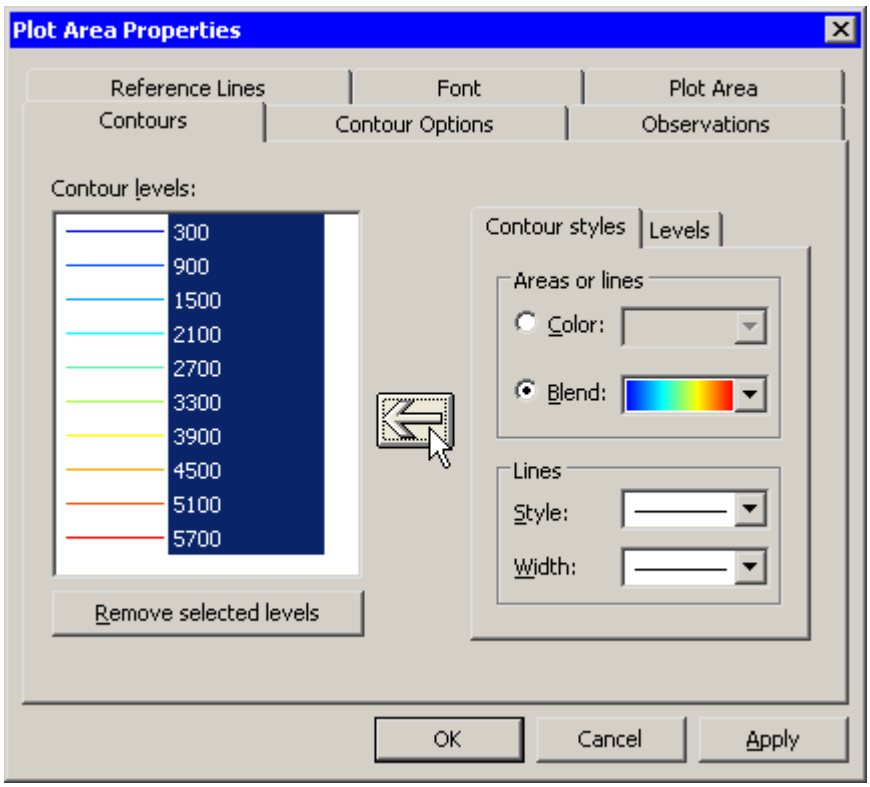
By default, the region between the new contours is gray. You can change the colors of contours by doing the following:

8 Click the **Contour Styles** subtab. (See Figure 7.21.)

9 Select a gradient color map from the **Blend** list.

10 Click the large left arrow (⟸) to color the selected contours according to the gradient color map, as shown in Figure 7.21.

Figure 7.21 Coloring Contours



11 Click **Apply** to update the contour plot.

You can also add individual contours for specific levels. For example, some investigators might want to see the "sea level" contour, Z = 0. Adding an individual contour is similar to adding a set of contours:

12 Click the **Levels** subtab.

13 Type `1` in the **Number** field.

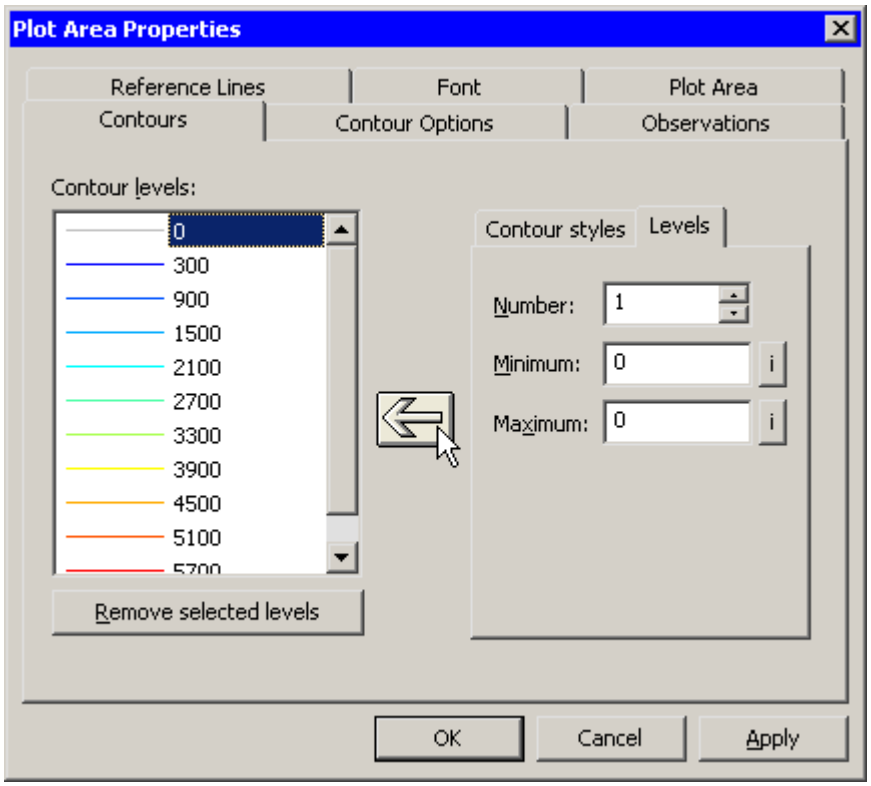
14 Type `0` in the **Minimum** field.

15 Type `0` in the **Maximum** field.

If the minimum and maximum values are the same, then a single contour is created at the common value.

16 Click the large left arrow (⟸) to create the contour, as shown in Figure 7.22.

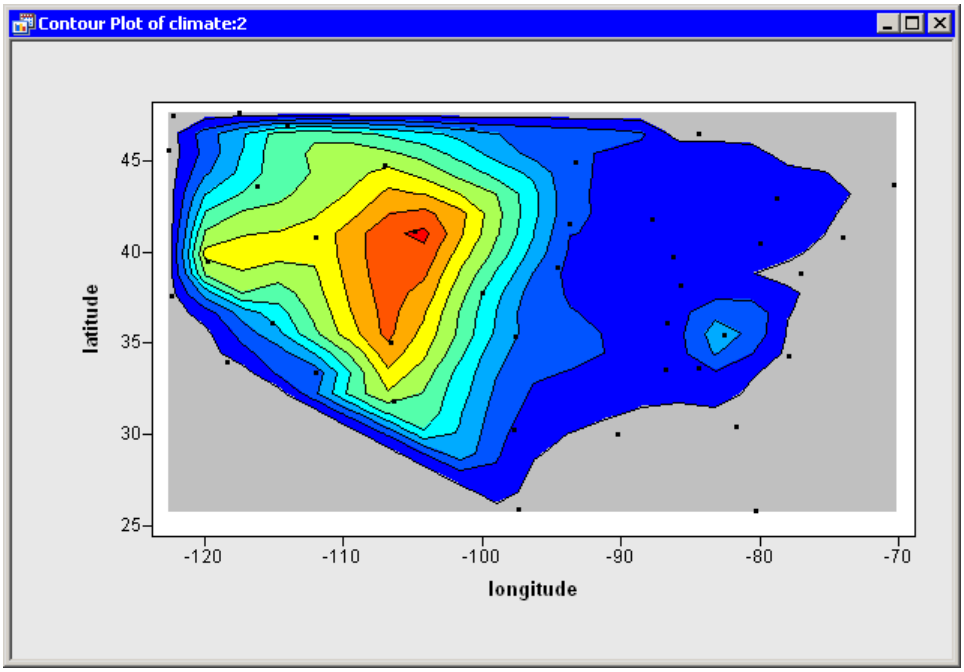
Figure 7.22 Adding a Single Contour



17 Click **OK** to apply the changes.

The contour plot looks like the plot in Figure 7.23. Note that the contour plot has not qualitatively changed from Figure 7.16. The new contour values are within a few hundred feet of their previous values, so the new contour curves are close to the previous contours. The primary change is that the new contours correspond to "round numbers" of elevationFeet. The colors are also slightly different.

Figure 7.23 A Plot with Custom Contours



NOTE: In this example you added a single contour at Z = 0. While SAS/IML Studio permits you to add contours at any level of the Z variable, you should usually choose evenly spaced levels. A standard usage of contour maps is to locate regions in which the contours are densely packed. These regions correspond to places where the gradient of Z is large; that is, the function is changing rapidly in these regions. If you add contours that are not evenly spaced in the range of Z, then you risk creating contours that are close together even though the gradient of Z is not large.

Contour Plot Properties

This section describes the property tabs that are associated with a contour plot. To access the rotating plot properties, right-click near the center of a plot, and select **Plot Area Properties** from the pop-up menu.

For a discussion of the **Observations** tab, see Chapter 6, "Exploring Data in Two Dimensions." For a discussion of the remaining tabs, see Chapter 9, "General Plot Properties."

Contours Tab

The **Contours** tab controls attributes of the contours. You can use this tab to remove contours, create contours, and change the color or line styles of contours. The **Contours** tab is shown in Figure 7.18.

The **Contours** tab has two subtabs: the **Contour Styles** subtab and the **Levels** subtab. The **Contours** tab contains the following UI controls:

Contour Levels
displays each contour in the plot. The contours are labeled by values of the Z variable. You can select one or more items in the list to change their properties or to remove them from the list.

Remove selected levels
removes contours that are selected in the **Contour Levels** list.

⟸ (large left arrow)
applies the current set of properties to the contours selected in the **Contour Levels** list. You must click the large left arrow to transfer the contour attributes to the selected items in the **Contour Levels** list, or to create new contours.

Contour Styles
specifies the contour colors and line styles. These attributes are not applied until you click the large left arrow (⟸). The **Contour Styles** subtab contains the following UI controls:

Color specifies a single color for the selected contour levels.

Blend specifies a gradient color map for a range of contour levels.

Style specifies a line style for the selected contours.

Width specifies a line width for the selected contours.

Levels
specifies the number and range of contour levels. These contours are created when you click the large left arrow (⟸). The **Levels** subtab contains the following UI controls:

Number specifies the number of contours to create.

Minimum specifies a value z_L used in the creation of new contours.

Maximum specifies a value z_R used in the creation of new contours.

You can create a set of contours by using the **Levels** subtab, as shown in Figure 7.20. Let n be the value in the **Number** field, and let z_L and z_R be the values in the **Minimum** and **Maximum** fields. These values specify that the interval between contours is $\delta = (z_R - z_L)/n$.

When you click the large left arrow (⟸), contours are created for the levels $z_i = z_L + \delta/2 + \delta i$, for $i = 0, \ldots, (n-1)$. This implies that the first level is $z_L + \delta/2$ and the last level is $z_R - \delta/2$. No contours appear for the z_L and z_R levels because levels for extreme values are often degenerate. (For example, if $z = x^2 + y^2$ on the domain $[-1, 1] \times [-1, 1]$, then the minimum value of z is 0, and the contour for that level is a single point.)

If, instead, you know that you want the first contour to be at the level z_0 and you want the contour interval to be δ, then it is straightforward to compute values of n, z_L, and z_R that satisfy those conditions. You can choose $z_L = z_0 - \delta/2$ and $z_R = z_L + n\delta$, where n is an integer.

If you want the contours to encompass all of your data, then you can compute $n = \lceil (z_{\max} - z_L)/\delta \rceil$, where $z_{\max}$ is the largest data value for the Z variable and $\lceil x \rceil$ is the least integer greater than x. You should also choose z_0 so that $|z_0 - z_{\min}| \leq \delta/2$. For example, if the range of your data is $[3, 97]$, and you want a contour interval of $\delta = 10$ with the first contour at $z_0 = 5$, then you can choose $z_L = 5 - 10/2 = 0$, $n = \lceil (97 - 0)/10 \rceil = 10$, and $z_R = 0 + 10 * 10 = 100$.

Contour Options Tab

The **Contour Options** tab controls attributes of the contour plot. You can also use this tab to control the size of the grid used to construct contours. The **Contour Options** tab is shown in Figure 7.24. The **Contour Options** tab contains the following UI controls:

Grid Sizes

specifies the resolution of the computational grid used to construct contours from the data. The algorithm that computes the surface uses a grid superimposed on the (X,Y) plane. This grid consists of evenly spaced subdivisions along the X and Y axes. Generally, having more subdivisions results in smoother contours, whereas having fewer subdivisions results in a rougher contours.

Show contour lines

specifies whether contours are shown.

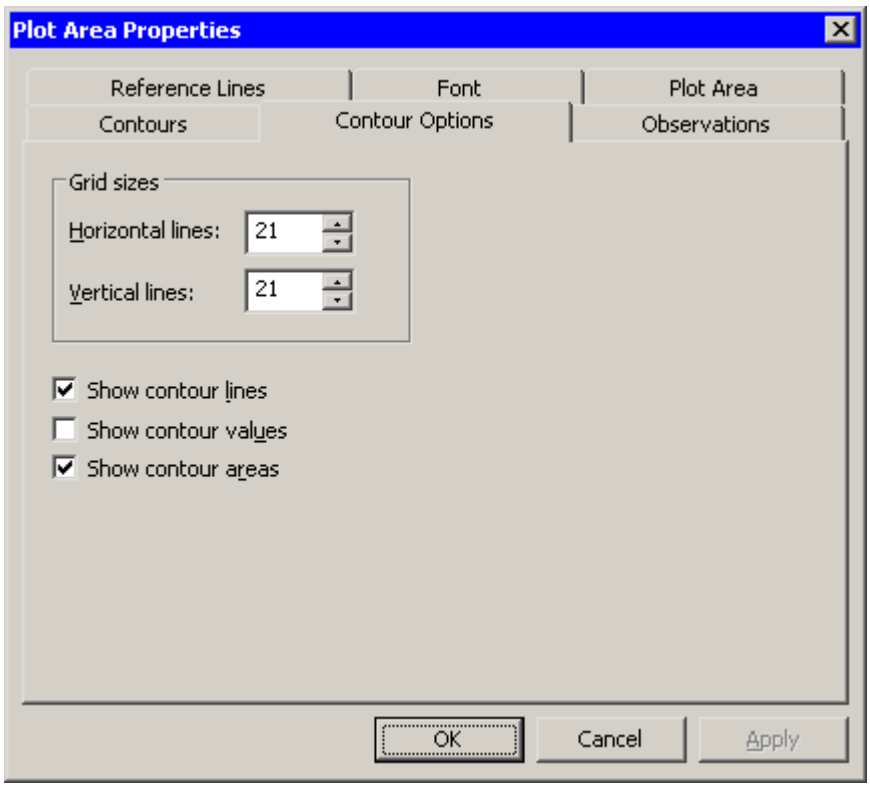
Show contour values

specifies whether contour lines are labeled by the value of the Z axis variable.

Show contour areas

specifies whether the region between contours is filled with color.

Figure 7.24 The Contour Options Tab



Contour Plots of Selected Variables

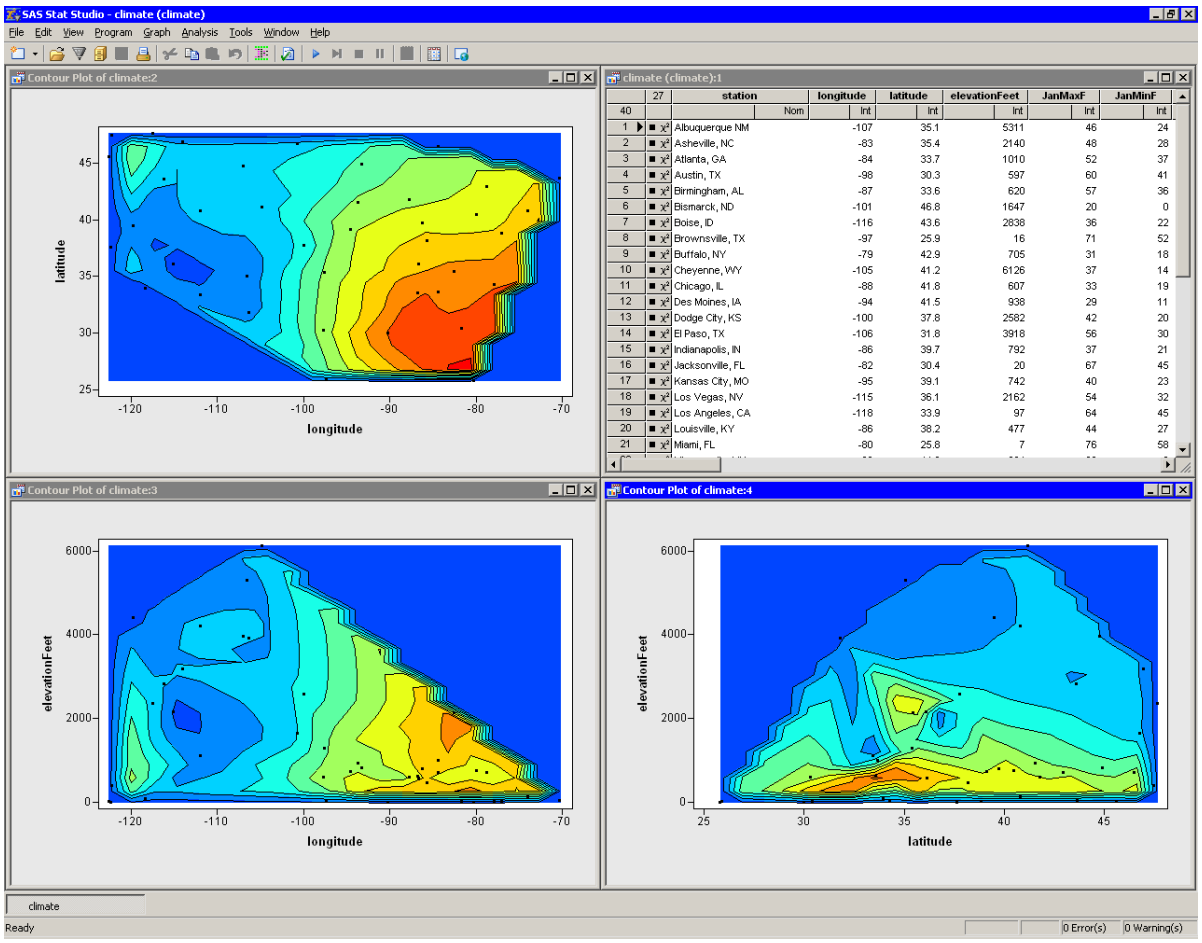
If one or more interval variables are selected in a data table when you select **Graph ►Contour Plot**, then the Contour Plot dialog box does not appear. Instead, the first selected variable is used as the Z variable. Contour plots are created for Z as a function of each pair of remaining interval variables.

If you create a matrix of plots from selected variables, you can close the matrix by pressing the F11 key while any plot is active and selecting from the pop-up menu. Alternatively, you can use the Workspace Explorer to quickly close plots. (See the section "Workspace Explorer" on page 190.)

Variables with a Frequency or Weight role are ignored when you are creating contour plots.

Figure 7.25 shows a matrix of contour plots for four selected variables. The TotalAvePrecipIn variable is plotted as a function of longitude, latitude, and elevationFeet.

Figure 7.25 A Matrix of Contour Plots



Chapter 8
Interacting with Plots

Contents

Overview of Interacting with Plots **131**
Interaction Tools **131**
Select Tool 132
Pan Tool 133
Zoom Tool 134
Spin Tool 134
Bin Tool 134
Level Tool 134
Resetting the Plot View **135**
Inserting Annotations **135**
Example: Insert an Annotation 135
Annotation Properties 138
Adjusting Graph Area Margins **139**
Observation Inspector **139**
Copying Plots to the Windows Clipboard **140**
Keyboard Shortcuts in Plots **140**

Overview of Interacting with Plots

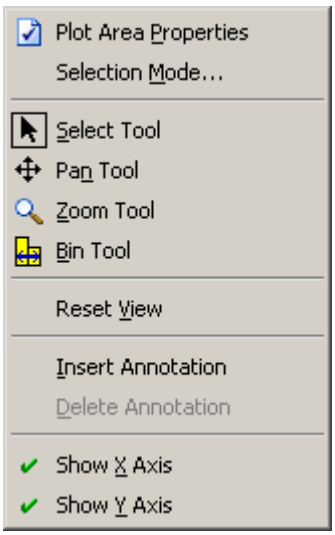
In this chapter you learn how you can interact with plots. These interactions include selecting observations, panning, and zooming. You learn how to display text on a plot and how to adjust the margins in a plot. You also learn about the *observation inspector*, a window that displays values of all variables for an observation.

Interaction Tools

The simplest way to interact with plots is by using the mouse to click or drag in the plot. Each plot supports tools that control the way that clicking or dragging affects the plot.

You can see the interaction tools for a plot by right-clicking in a plot. For example, Figure 8.1 shows the tools available for a histogram. Selecting a tool item from the pop-up menu changes the shape of the mouse pointer and determines how the plot interprets a mouse click.

Figure 8.1 Some Available Tools



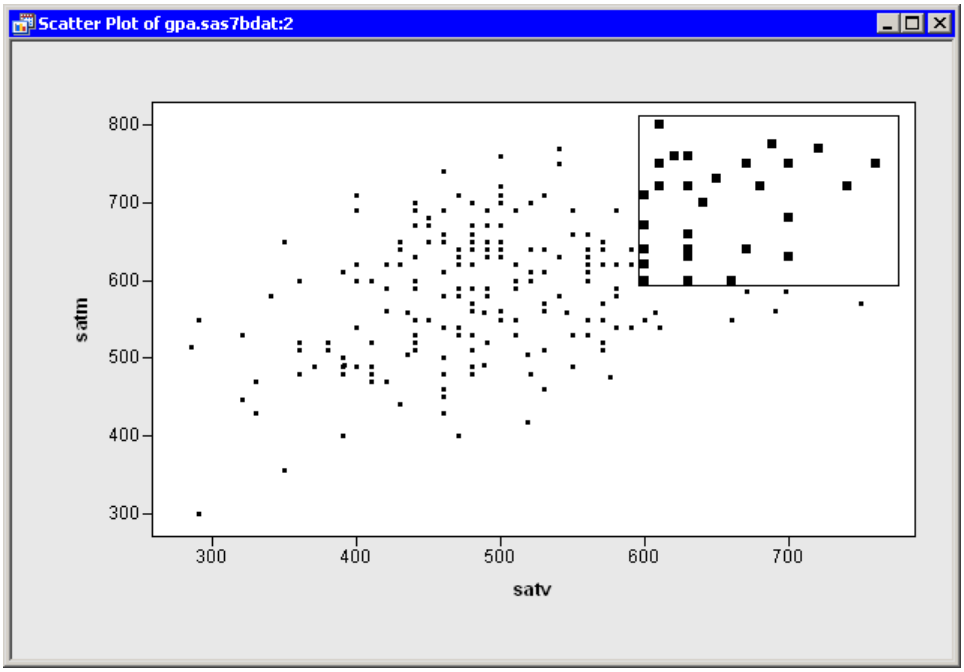
The default tool for all plots is the select tool. The various tools and their effects on the plots are summarized in the following sections.

Select Tool

When you choose the select tool, the mouse pointer looks like a diagonally pointing arrow. Clicking a plot marker selects the corresponding observation. Clicking a bar in a histogram or bar chart selects all observations that are represented by that bar. Clicking a box plot quartile or whisker selects all observations in that quartile or whisker. By holding down the SHIFT or CTRL key, you can select multiple graphical elements.

Dragging a rectangle selects all observations within that rectangle. The rectangle is also known as a *brush*. After a brush is created, you can move it by placing the mouse pointer inside the rectangle and dragging it to a new location. As the brush passes over observations, those observations are automatically selected, as shown in Figure 8.2. If you hold down the CTRL key while moving the brush, observations outside the brush are not deselected.

Figure 8.2 Selecting Observations with a Brush



It is also possible to *throw the brush*. Release the mouse button while dragging the brush, and the brush begins moving freely in the direction in which you last dragged it. The brush bounces off the sides of the graph area. Throwing the brush can be computationally intensive when you are working with large data sets.

NOTE: If you click an observation, it is labeled in the plot. Details are given in the section "Labeling Observations" on page 156. Observations that are selected by using a brush are not labeled.

Pan Tool

When you choose the pan tool, the mouse pointer looks like four arrows that meet at right angles. By dragging the pointer, you can translate the contents of the plot. The rotating plot does not support the pan tool.

Zoom Tool

When you choose the zoom tool, the mouse pointer looks like a magnifying glass. Clicking in a plot fixes the relative position of the pointer and expands the scale of the plot by a factor of 1.5. Clicking while holding down the SHIFT key shrinks the scale of the plot by a factor of 1.5.

If you draw a rectangle with the zoom tool, the region inside the rectangle expands to fill the plot area. If you draw a rectangle with the zoom tool while holding down the SHIFT key, the plot area is shrunk down into the rectangle.

The rotating plot does not support the zoom tool.

Spin Tool

When you choose the spin tool, the mouse pointer looks like a circular arrow (↺). Only the rotating plot supports the spin tool.

Clicking in the plot causes the plot to rotate toward the pointer by an amount that is proportional to the distance between the pointer and the center of the plot. Dragging the pointer rotates the plot. If you release the mouse button while the pointer is in motion, the plot rotates freely. Click anywhere in the plot to stop the rotation.

Bin Tool

When you choose the bin tool, the mouse pointer looks like a double-headed arrow between a pair of lines. Only the histogram supports the bin tool.

Clicking or dragging the bin tool shifts the location of the histogram bins. Clicking near the horizontal axis reduces the number of bins and makes the bars wider. Clicking near the top of the plot increases the number of bins and makes the bars narrower. Dragging the mouse pointer horizontally does not change the number of bins but changes the position at which the bins start.

When the pointer is at the left edge of the histogram, the bins start at an integral multiple of the bin width. When you move the pointer toward the right, the bins are offset by an amount that is proportional to the distance between the pointer and the left edge of the histogram.

Level Tool

When you choose the level tool, the mouse pointer looks like a pencil. Only the contour plot supports the level tool.

Clicking and dragging the level tool near a contour changes the value of the Z variable that is associated with the contour. You can insert a new contour by clicking the level tool away from existing contours.

Resetting the Plot View

In many cases, you can reset a plot to its original view of the data. Right-click in the plot and select **Reset View** from the pop-up menu to reset changes to a plot that were made by the pan tool, zoom tool, or spin tool. Changes made by the bin tool or level tool are not affected by **Reset View**.

Inserting Annotations

In this section you learn how to display text on a plot. For example, you might want to draw attention to an outlier or display statistics that are associated with the plot. To add text to a plot, right-click in the plot and select **Insert Annotation** from the pop-up menu, as shown in Figure 8.3.

Example: Insert an Annotation

The following steps insert text that displays certain statistics that are related to a scatter plot of two variables:

1 Open the Hurricanes data set, and create a scatter plot of wind_kts versus min_pressure.

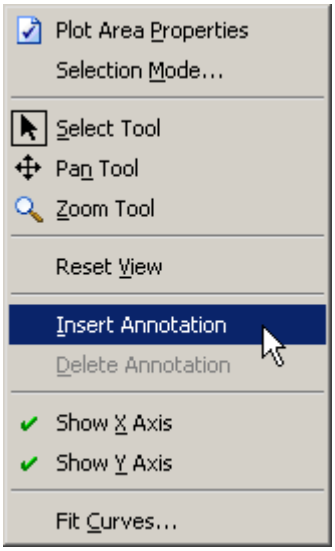
The scatter plot shows a strong negative correlation between wind speed and pressure. (See Figure 8.4.) A correlation analysis reveals the following:

- There are 6,185 observations for which both variables are nonmissing.
- The correlation between these two variables is approximately –0.93.

You can display these statistics on the plot.

2 Right-click in the plot, and select **Insert Annotation** from the pop-up menu.

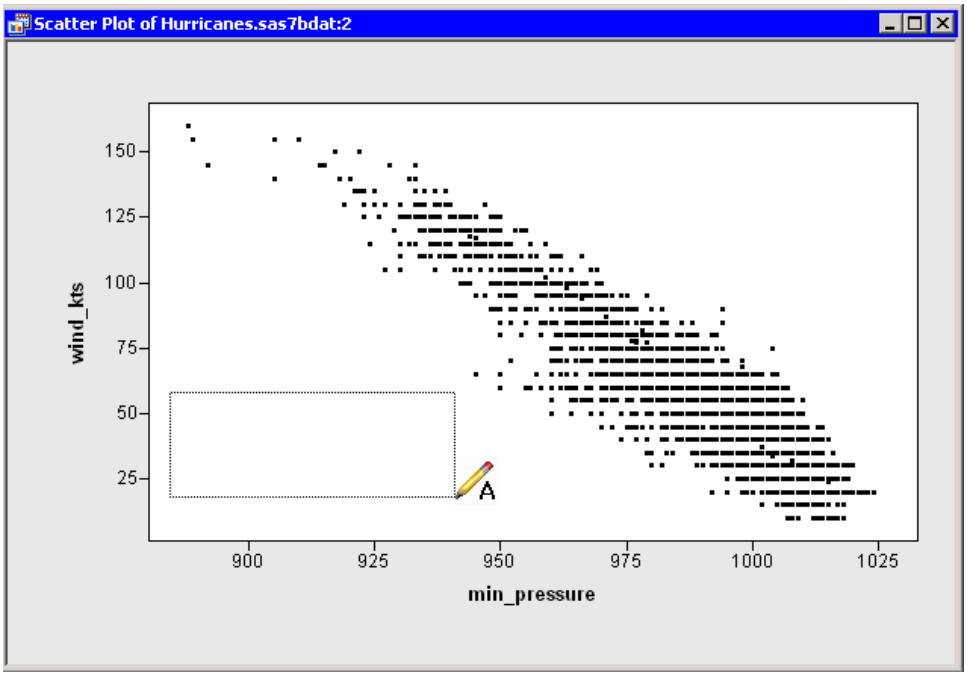
Figure 8.3 Creating an Annotation



The mouse pointer changes its shape. It looks like a pencil with the letter “A” next to it.

3 Draw a rectangle with the mouse pointer, as shown in Figure 8.4.

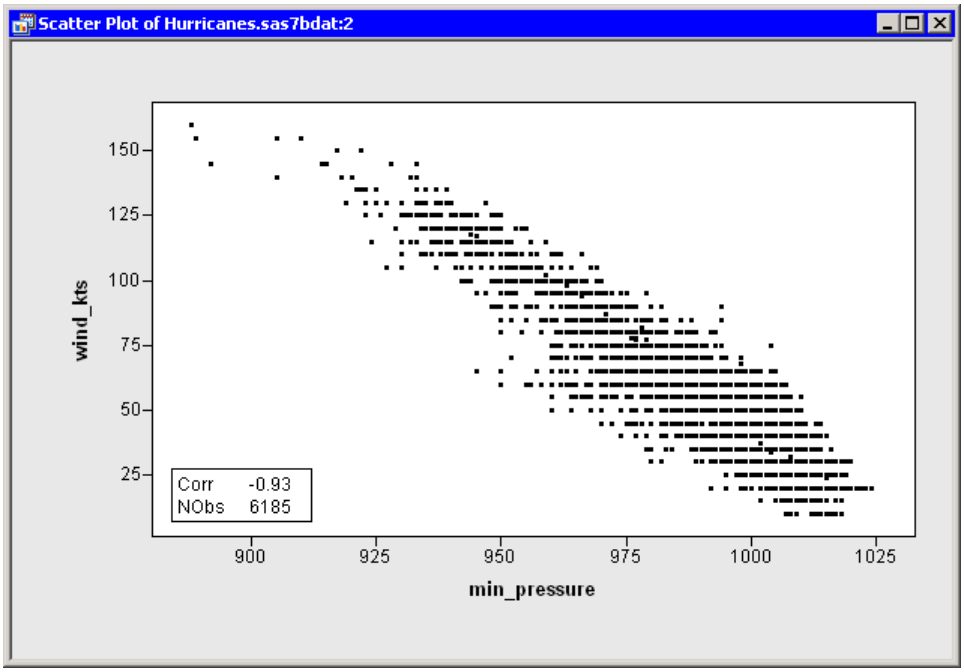
Figure 8.4 Inserting an Annotation



4 Type text into the text box, as shown in Figure 8.5. When you are finished editing the text, click outside the text box.

You can resize or move the text rectangle after it is created, if necessary. You can also right-click the text box to change properties of the text or the text box. For example, in Figure 8.5 the text box is displayed with a border around it. The annotation properties are discussed in the section “Annotation Properties” on page 138.

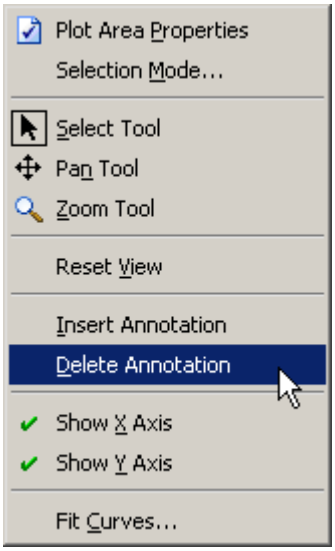
Figure 8.5 An Inset That Contains Statistics



If you decide to delete the annotation, click the text box to select it. Then right-click *outside the text box*, and select **Delete Annotation** from the pop-up menu, as shown in Figure 8.6.

NOTE: If you right-click *inside* the text box, you get a different menu, as discussed in the following section.

Figure 8.6 Deleting an Annotation



Annotation Properties

You can change properties of an annotation. Click the annotation text box to select it. Right-click inside the text box, and select **Properties** from the pop-up menu.

The Text Properties dialog box appears. (See Figure 8.7.) The dialog box has two tabs. You can use the **Font** tab to set attributes of the font that is used to display an annotation. The **Font** tab is described in "Common Plot Properties" on page 158.

You can use the **Text Editor** tab to set attributes of the text box. The **Text Editor** tab has the following UI controls:

Text Alignment
: specifies the alignment of the text within the text box.

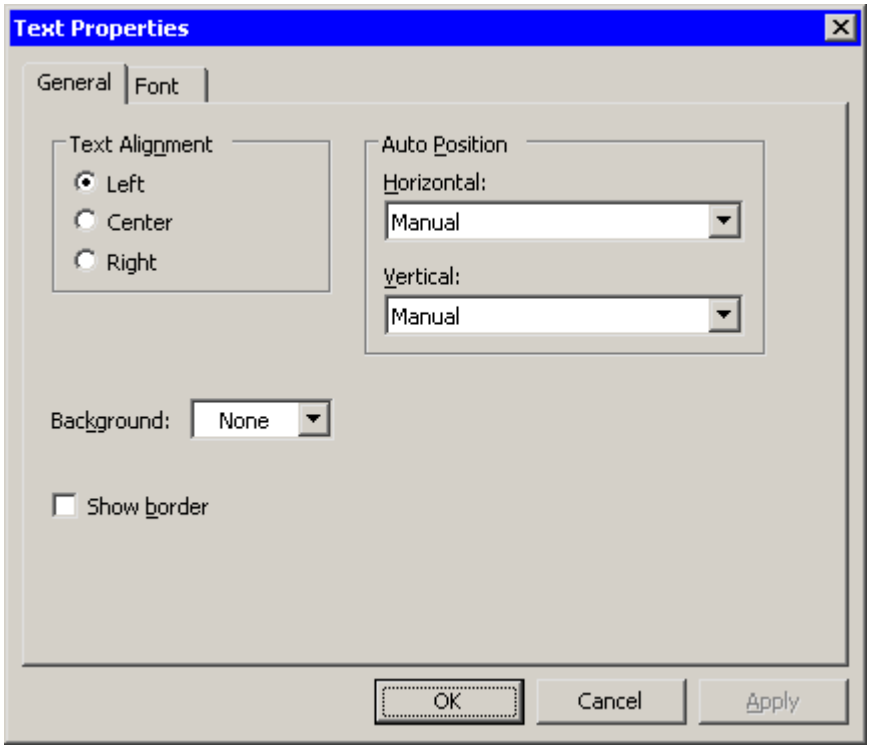
Horizontal
: specifies the horizontal position of the text box within the graph area or plot area.

Vertical
: specifies the vertical position of the text box within the graph area or plot area.

Background
: specifies the color of the text box background.

Show border
: specifies whether to display a border around the text box.

Figure 8.7 Text Editor Tab



Adjusting Graph Area Margins

You can interactively resize the plot area when the select tool is active. Rest the mouse pointer at the edge of the plot area until the pointer changes to a double-headed arrow. Then drag the plot area to resize it. When you resize the plot area, you are actually changing the graph area margins, as described in the section "Common Graph Area Properties" on page 162.

You cannot adjust the graph area margins if the plot has a fixed aspect ratio.

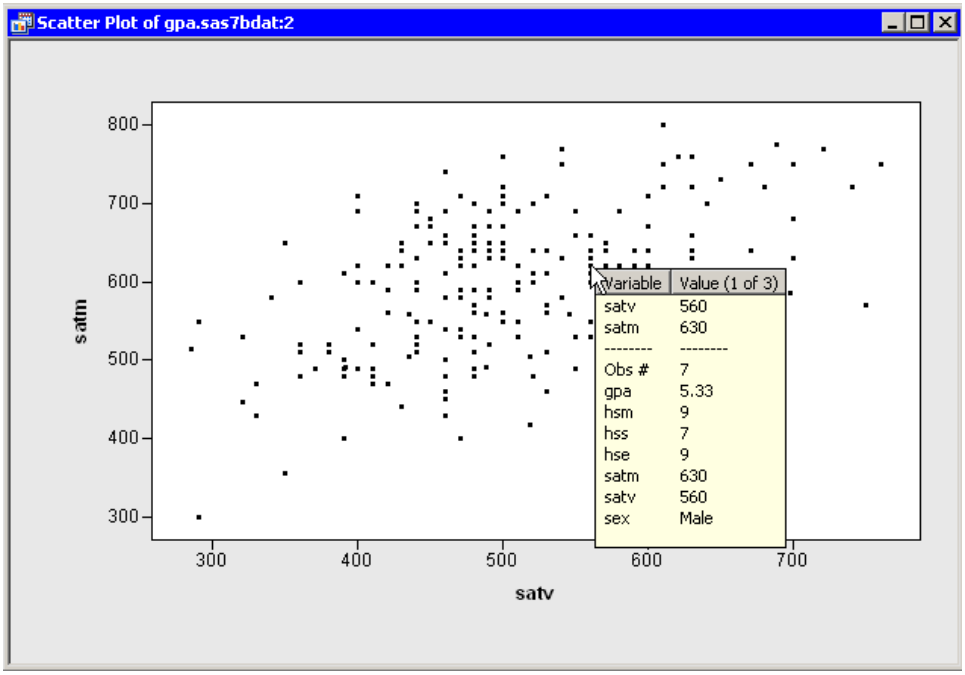
Observation Inspector

You can interactively query plots to display the values of variables for the observations beneath the mouse pointer. The discussion in this section applies to plots that show individual markers for each observation.

The observation inspector window displays the values of all the variables for a particular observation. (See Figure 8.8.) You can display the observation inspector window in one of three ways:

- Hold down the F2 key. The observation inspector window appears for any observations beneath the mouse pointer.
- Press SHIFT+F2. You are now in observation inspector mode. If you rest the mouse pointer over an observation, the observation inspector window appears. To exit observation inspector mode, press the ESC key while the observation inspector is active, or press SHIFT+F2 a second time.
- Double-click an observation.

Figure 8.8 The Observation Inspector Window



The top portion of the observation inspector window displays the variables that are used by the plot. For example, the observation inspector window for a scatter plot displays the X- and Y-axis variables first. If observations are labeled by some variable, that label variable also appears. The observation inspector window next displays a horizontal line, followed by the observation number (in the current sort order), followed by all variables in the order in which they appear in the data set.

If there are many variables, it is possible that not all of the variables fit into the observation inspector window. You can scroll the observation inspector window by using the HOME, END, PAGE UP, PAGE DOWN, UP ARROW, and DOWN ARROW keys.

If there are multiple observation markers near the mouse pointer (as in Figure 8.8), the observation inspector creates a list of all the nearby observations and displays the text "Value (1 of *N*)" in its column heading. You can display the next observation in the list by pressing the RIGHT ARROW key. You can go back to a previous observation by pressing the LEFT ARROW key. Pressing RIGHT ARROW or LEFT ARROW while holding down the SHIFT key causes a jump forward or backward to the observation that is approximately *N*/5 entries away in the list.

Copying Plots to the Windows Clipboard

You can copy a plot to the Windows clipboard by selecting **Edit ►Copy** from the main menu when the plot is active. Alternatively, you can press CTRL+C.

You can paste plots into the SAS/IML Studio output document or into other applications such as Microsoft Word and PowerPoint.

Keyboard Shortcuts in Plots

All plots support the standard Microsoft Windows control sequences listed in Table 8.1.

Table 8.1 Standard Control Sequences in Plots

Key	Action
CTRL+A	Selects all observations that are included in the plots.
CTRL+C	Copies the plot to Windows clipboard.
CTRL+P	Prints the plot.

All plots support the keyboard shortcuts listed in Table 8.2.

Table 8.2 Keys and Actions in All Plots

Key	Action
0, 1–9	Sets the color of selected observations to the color specified in Table 8.3. If no observations are selected, sets the marker color of all observations.
A	Selects all observations that are included in the plots.
B	Applies a color blend according to values of the X variable. (Plots with multiple X variables ignore this key.) Bar charts and histograms also color each bin according to the color blend.
C	Selects the complement of the selected observations.
E	Excludes selected observations from plots and analyses.
G	Toggles a reference line grid.
I	Includes selected observations in plots and analyses.
L	Toggles labels on bars or observations.
X, Y, Z	Displays axis property dialog box for the corresponding axis.

Ten predefined colors are associated with number keys. Table 8.3 lists the color that is associated with each digit 0–9.

Table 8.3 Keys and Colors in Plots

Key	Color
0	Black
1	Red
2	Green
3	Blue
4	Gold
5	Magenta
6	Olive
7	Steel
8	Brown
9	Violet

Area plots (histograms, bar charts, and mosaic plots) support the keyboard shortcuts listed in Table 8.4.

Table 8.4 Keys and Actions in Area Plots

Key	Action
H	Toggles filling the bars.
P	Cycles through displaying frequency, percentages, and (for histograms) density on the Y axis.
CTRL+*digit*	Sets the percentage threshold for the "Others" category. For example, CTRL+4 sets the threshold to 4%, whereas CTRL+0 sets the threshold to 0% and therefore turns off the "Others" category. (The histogram ignores this key.)

Point plots (any plot that displays individual observations) support the keyboard shortcuts listed in Table 8.5.

Table 8.5 Keys and Actions in Point Plots

Key	Action
F2	Displays the observation inspector for any observations beneath the mouse pointer.
SHIFT+F2	Toggles observation inspector mode, as described in the section "Observation Inspector" on page 139.
H	Toggles the "show only selected observations" option.
`[` or `]` (square bracket)	Toggles fixed aspect ratio. (The box plot ignores this key.)
CTRL+UP ARROW	Increases the size of markers.
CTRL+DOWN ARROW	Decreases the size of markers.
ALT+UP ARROW	Increases the size difference between selected and unselected markers.
ALT+DOWN ARROW	Decreases the size difference between selected and unselected markers.

Box plots support the keyboard shortcuts listed in Table 8.6.

Table 8.6 Keys and Actions in Box Plots

Key	Action
M	Toggles displaying the mean and standard deviation.
N	Toggles displaying the notches that measure the significance of the difference between two medians.
HYPHEN	Toggles displaying serifs.

Line plots support the keyboard shortcuts listed in Table 8.7.

Table 8.7 Keys and Actions in Line Plots

Key	Action
0–9	Sets the color of the selected lines. The colors are listed in Table 8.3.
ESC	Deselects all lines.
CTRL+PAGE UP, CTRL+PAGE DOWN	Selects the previous or next line. (Selects the first line if no line is selected.)
CTRL+UP ARROW	Increases the width of selected lines.
CTRL+DOWN ARROW	Decreases the width of selected lines.
CTRL+LEFT ARROW, CTRL+RIGHT ARROW	Cycles through line styles for the selected lines.

Rotating plots support the keyboard shortcuts listed in Table 8.8.

Table 8.8 Keys and Actions in Rotating Plots

Key	Action
UP ARROW	Rotates up.
DOWN ARROW	Rotates down.
LEFT ARROW	Rotates left.
RIGHT ARROW	Rotates right.
PAGE UP, PAGE DOWN	Rotates about an axis that is perpendicular to the computer monitor.
CTRL+B	Toggles displaying the frame box.
CTRL+D	Toggles depth perception.
CTRL+G	Toggles displaying the surface graph.
CTRL+R	Toggles displaying rays from the origin.

Polygon plots support the keyboard shortcut listed in Table 8.9.

Table 8.9 Keys and Actions in Polygon Plots

Key	Action
CTRL+ALT+F	Toggles filling the polygons.

Chapter 9
General Plot Properties

Contents

Overview of Plot Properties **145**
Context Areas **146**
Changing Marker Shapes **147**
Example: Change Marker Shapes 147
Changing Marker Colors **150**
Example: Change Marker Colors 150
Displaying Only Selected Observations **152**
Example: Display Only Selected Observations 153
Labeling Observations **156**
Example: Label Observations 156
Common Plot Properties **158**
Reference Lines Tab 158
Font Tab 160
Plot Area Tab 161
Common Graph Area Properties **162**
Graph Area Tab 163

Overview of Plot Properties

In this chapter you learn about basic properties of plots. Knowing how to change the default plot properties enables you to better visualize and explore your data.

In this chapter you learn how to do the following:

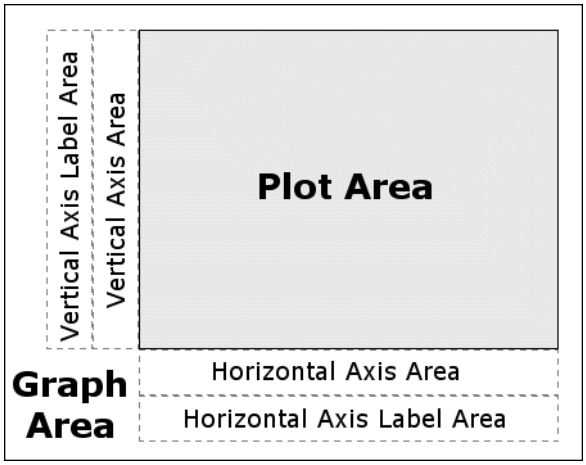
- display different menus by clicking in different regions of a plot
- change the shape of markers
- change the color of markers
- display only selected observations
- label observations
- change common plot properties such as reference lines, fonts, and plot margins
- change common graph properties such as margins, titles, and footnotes

Context Areas

Right-clicking inside a plot window pops up a *context menu*, whose contents depend on the location of the mouse pointer when you right-click.

Figure 9.1 shows the six nonoverlapping regions in a two-dimensional plot: the plot area, the graph area, two axis areas, and two axis label areas. An example of a pop-up context menu is shown in Figure 8.1, which shows the context menu for the plot area. The context menus for the other areas in Figure 9.1 look similar; only the first menu item differs among the context menus.

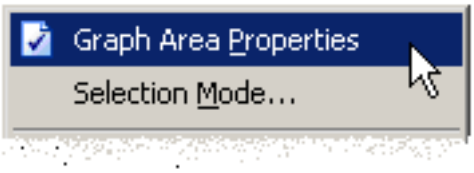
Figure 9.1 Context Areas for a Two-Dimensional Plot



If you right-click in the plot area and select **Plot Area Properties**, the Plot Area Properties dialog box appears. The dialog box for the plot area controls the appearance of the plot. The tabs on the dialog box vary according to the plot type (histogram, scatter plot, box plot, and so on). Properties that are common to all plots are discussed in this chapter. Plot-specific properties are discussed in Chapter 5, "Exploring Data in One Dimension," Chapter 6, "Exploring Data in Two Dimensions," Chapter 7, "Exploring Data in Three Dimensions," and Chapter 8, "Interacting with Plots."

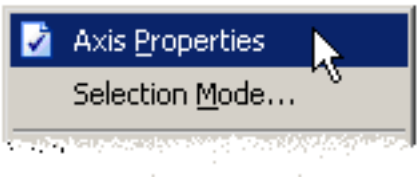
If you right-click in the graph area and select **Graph Area Properties** (see Figure 9.2), the Graphs Area Properties dialog box appears. You can use the dialog box to change general properties that affect the layout of the plot within the larger graph area. This dialog box is discussed in the section "Common Graph Area Properties" on page 162.

Figure 9.2 Context Menu for the Graph Area



If you right-click in the axis area and select **Axis Properties** (see Figure 9.3), the Axis Properties dialog box appears. You can use the dialog box to control the scale, font, and placement of tick marks for that axis. Similarly, if you right-click in the axis label area and select **Axis Label Properties**, the Axis Label Properties dialog box appears. You can use the dialog box to control the font and text used to label that axis. These dialog boxes are discussed in Chapter 10, "Axis Properties."

Figure 9.3 Context Menu for an Axis Area



The context areas for a rotating plot are slightly different than shown in Figure 9.1. A rotating plot lacks the "axis area" regions. The rotating plot behaves differently because the position of the axes changes as the plot rotates.

Changing Marker Shapes

Not every plot shows individual observations. Some plots, such as histograms, bar charts, and mosaic plots, aggregate observations into a group and represent that group with a bar or box. The discussion in this section applies to plots that show individual markers.

When a graph is printed on a gray-scale printer, it is often easier to discern observations that have different marker shapes than it is to discern markers that have different colors. Even on a computer screen, marker shape is sometimes preferred for classifying markers according to a small number of discrete values. For example, marker shape is an ideal way to encode gender.

You can change the marker shape for all observations, or just for observations that are selected. You can select observations by using graphical techniques or by using the Find dialog box in a data table, as discussed in the section "Finding Observations" on page 50.

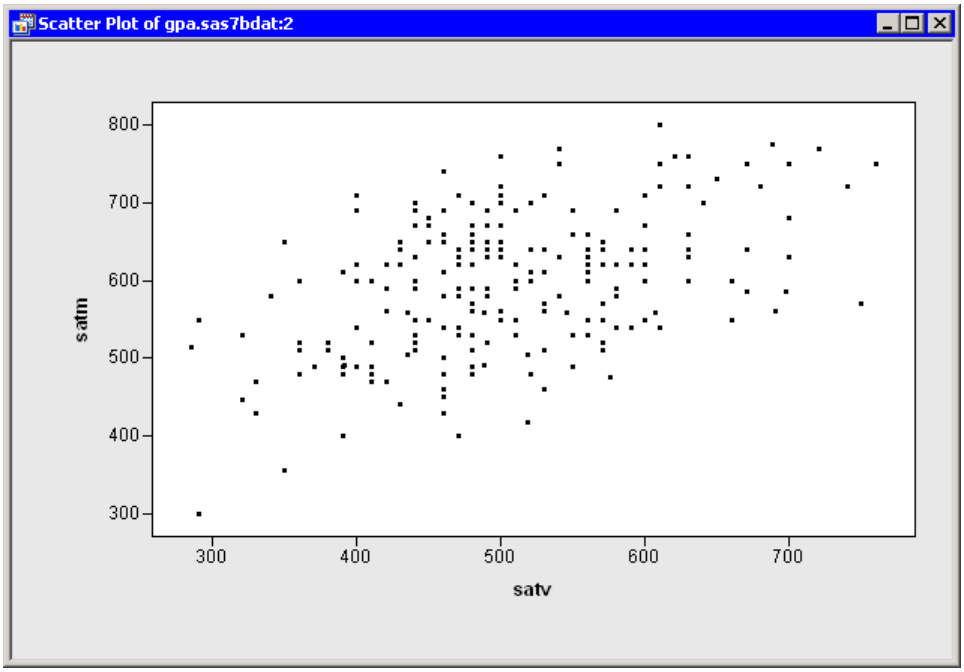
Example: Change Marker Shapes

In this example, you use a bar chart of a categorical variable to select observations, and you change the marker shape of the selected observations.

1 Open the GPA data set, and create a scatter plot of satm versus satv.

The scatter plot appears. (See Figure 9.4.)

Figure 9.4 A Scatter Plot



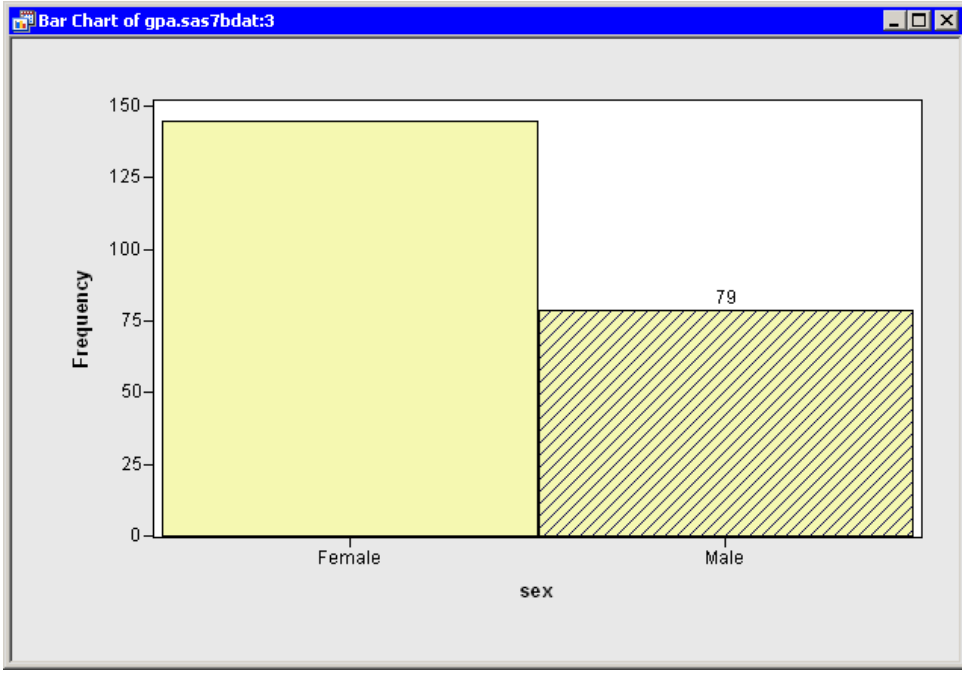
Each observation in this data set represents a student. You can use marker shape to indicate each student's gender.

2 Create a bar chart of the sex variable.

If necessary, move the bar chart so that it does not overlap the scatter plot.

3 Select all the male students in the bar chart, as shown in Figure 9.5.

Figure 9.5 A Bar Chart with Male Students Selected

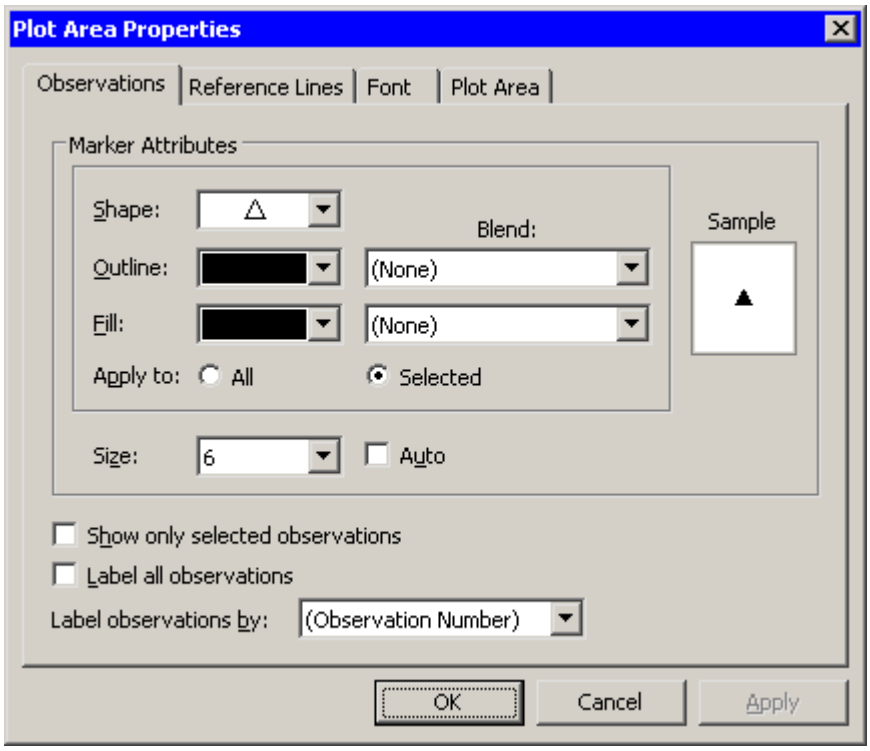


While the bar chart is convenient for selecting all the male students, you need to return to the scatter plot in order to change the marker shapes of the selected observations.

4 Right-click near the center of the scatter plot, and select **Plot Area Properties** from the pop-up menu.

The Plot Area Properties dialog box appears. (See Figure 9.6.) You can use the **Observations** tab to change marker shapes, colors, and sizes. The section "Scatter Plot Properties" on page 85 gives a complete description of the options available on the **Observations** tab.

Figure 9.6 The Observations Tab



5 Select a triangle (△) from the **Shape** list. When there are selected observations, **Apply to** defaults to **Selected**. This means that the **Shape**, **Outline**, and **Fill** options are applied only to the selected observations. (You can, of course, override this default and apply changes to all observations.)

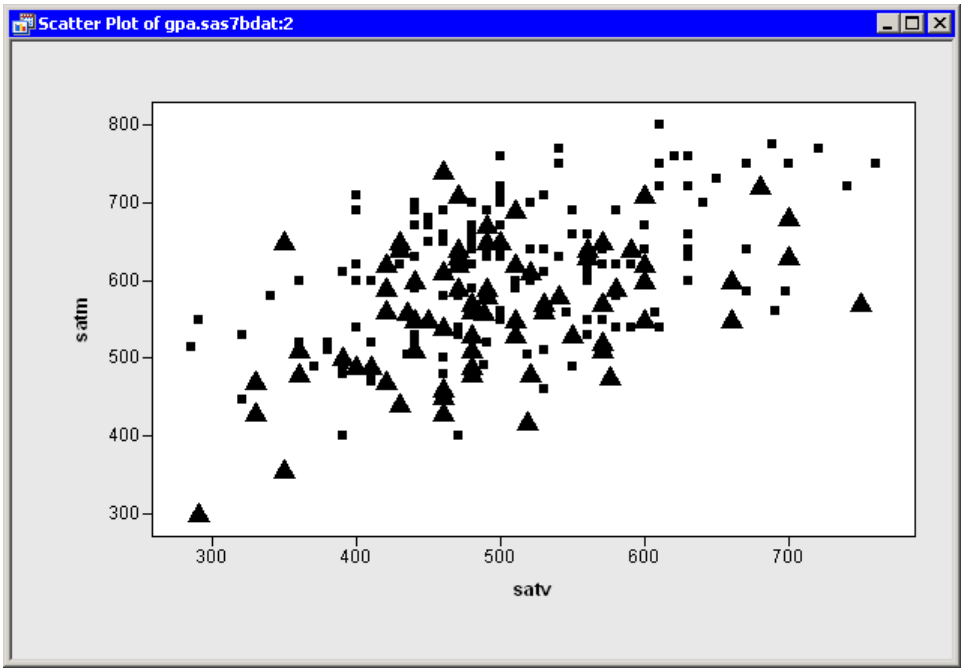
6 Select **6** from the **Size** list.

Note that the **Size** list is *not* in the same group box as **Apply to**. All markers in a plot have a common scale; size differences are used to distinguish between selected and unselected observations. When a plot is active, you can increase the size difference between selected and unselected markers by pressing the UP ARROW key while holding down the ALT key.

7 Click **OK**.

The scatter plot updates, as shown in Figure 9.7. The SAT scores of male students are represented by triangles; SAT scores of female students are represented by squares.

Figure 9.7 Using Marker Shape to Indicate Gender



Changing Marker Colors

You can use the color of markers to indicate observations of interest (for example, outliers) or to color observations according to the value of some variable. The discussion in this section applies to plots that show individual markers.

The simplest use of color is to assign a color to one or more selected observations. For example, you can repeat the example of the section "Changing Marker Shapes" on page 147, but use color to indicate the male students.

You can color markers according to values of a nominal or interval variable. In the next example you color markers according to an interval variable. This technique is sometimes useful for visualizing trivariate data by using a scatter plot to visualize two variables and using color to visualize the third.

Example: Change Marker Colors

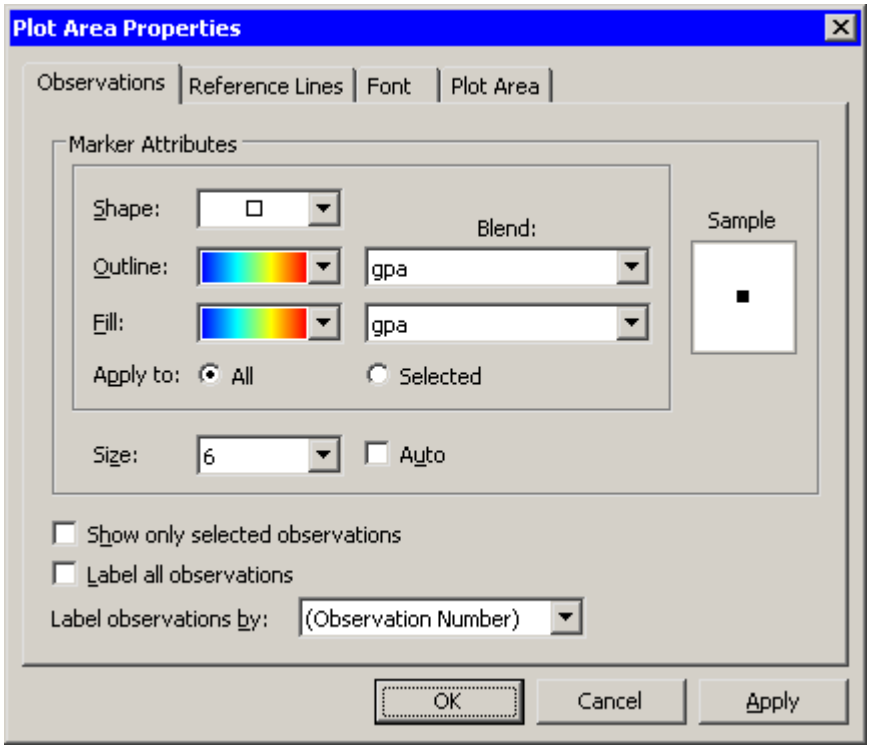
1 Open the GPA data set, and create a scatter plot of satm versus satv.

The scatter plot appears. (See Figure 9.4.) You can use color to visualize the grade point average (GPA) for each student.

2 Right-click near the center of the plot, and select **Plot Area Properties** from the pop-up menu.

The Plot Area Properties dialog box appears. (See Figure 9.8.) You can use the **Observations** tab to change marker shapes, colors, and sizes. The section "Scatter Plot Properties" on page 85 gives a complete description of the options available on the **Observations** tab.

Figure 9.8 The Observation Tab



3 Select gpa from the **Outline: Blend** and **Fill: Blend** lists. Select a gradient color map (the same one) from the **Outline** and **Fill** lists.

Make sure that **Apply to** is set to **All**.

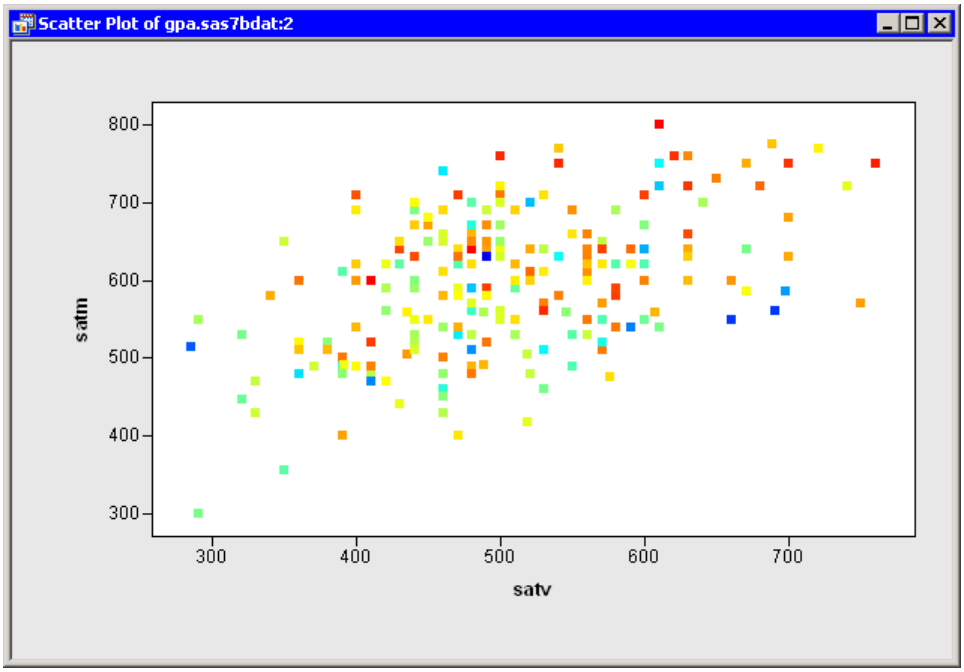
4 Select **6** from the **Size** list.

Note that the **Size** list is *not* in the same group box as **Apply to**. All markers in a plot have a common scale; size differences are used to distinguish between selected and unselected observations.

5 Click **OK**.

The scatter plot updates, as shown in Figure 9.9. These data do not seem to indicate a strong relationship between a student's college grade point average and SAT scores.

Figure 9.9 Using Color to Indicate Grade Point Average



Displaying Only Selected Observations

The discussion in this section applies to plots that show individual markers.

The default SAS/IML Studio behavior is to show all observations in a plot. Selected observations are displayed at a larger size than unselected observations. You can choose instead to display only selected observations. This is useful when there are so many points in a plot that the selected observations are not distinguishable (a phenomenon known as *overplotting*).

You can also examine subsets of the data by displaying only selected observations. This technique is called *slicing*. You can slice dynamically to explore multivariate relationships.

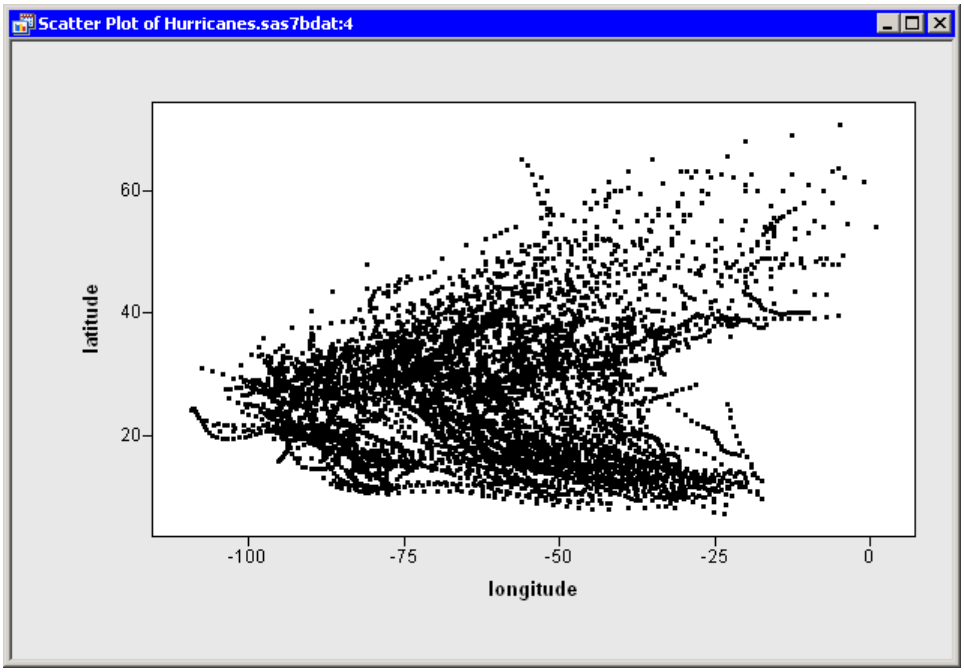
Example: Display Only Selected Observations

In this example, you visualize the distribution of points in a scatter plot, as subset by values of a categorical variable. This is sometimes called a *conditional plot*.

1 Open the Hurricanes data set, and create a scatter plot of latitude versus longitude.

The scatter plot appears. (See Figure 9.10.) The plot shows the position of Atlantic cyclones during a 16-year period. There is considerable overplotting in this scatter plot, particularly along a path between the Cape Verde Islands (lower right corner of the plot) and the Caribbean Sea (near the coordinates $(-75, 20)$).

Figure 9.10 A Scatter Plot

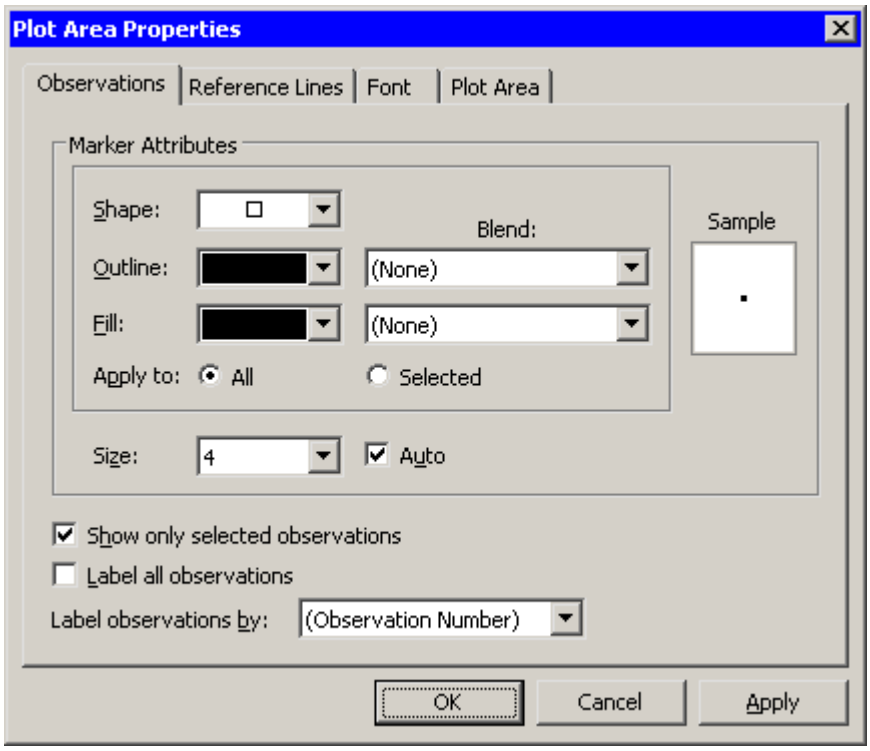


The overplotting prevents the clear examination of rare events such as category 4 and category 5 hurricanes. You can modify the scatter plot so that it displays only selected observations. This makes it easier to examine these storms.

2 Right-click near the center of the plot, and select **Plot Area Properties** from the pop-up menu.

The Plot Area Properties dialog box appears. (See Figure 9.11.)

Figure 9.11 The Observations Tab



3 Select **Show only selected observations**.

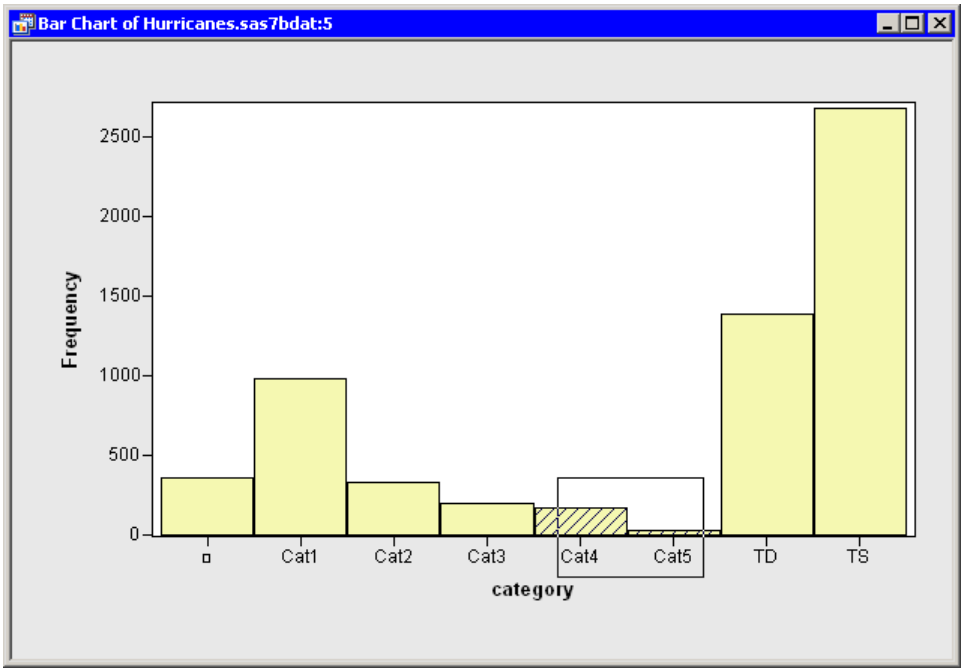
4 Click **OK**.

The scatter plot updates. All of the observations disappear because none are selected. You can use another plot or the data table's Find dialog box (see the section "Finding Observations" on page 50) to select data of interest.

5 Create a bar chart of the category variable.

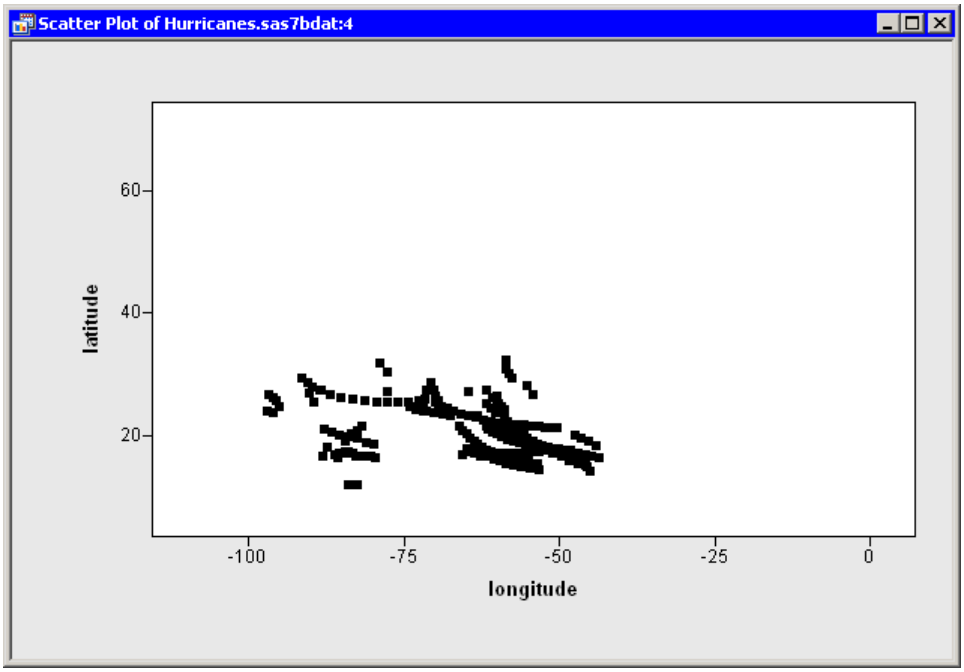
6 Select all category 4 and 5 hurricanes in the bar chart, as shown in Figure 9.12.

Figure 9.12 A Bar Chart with Category 4 and 5 Hurricanes Selected



The selected observations appear in the scatter plot, as shown in Figure 9.13. Most of the selected storms appear in the Gulf of Mexico, the Caribbean Sea, and the Atlantic Ocean east of the Greater Antilles.

Figure 9.13 Displaying Only Selected Observations



Labeling Observations

The discussion in this section applies to plots that show individual markers.

If you click an observation in a plot, a label appears near the selected observation. By default, the label is the observation number (position in the data table). You can choose instead to label observations by the value of any variable, called the *label variable*. You can set a default label variable that is used for all plots, or you can set a label variable for a particular plot that overrides the default label variable.

Example: Label Observations

In this example, you label observations in a scatter plot according to values of a third variable.

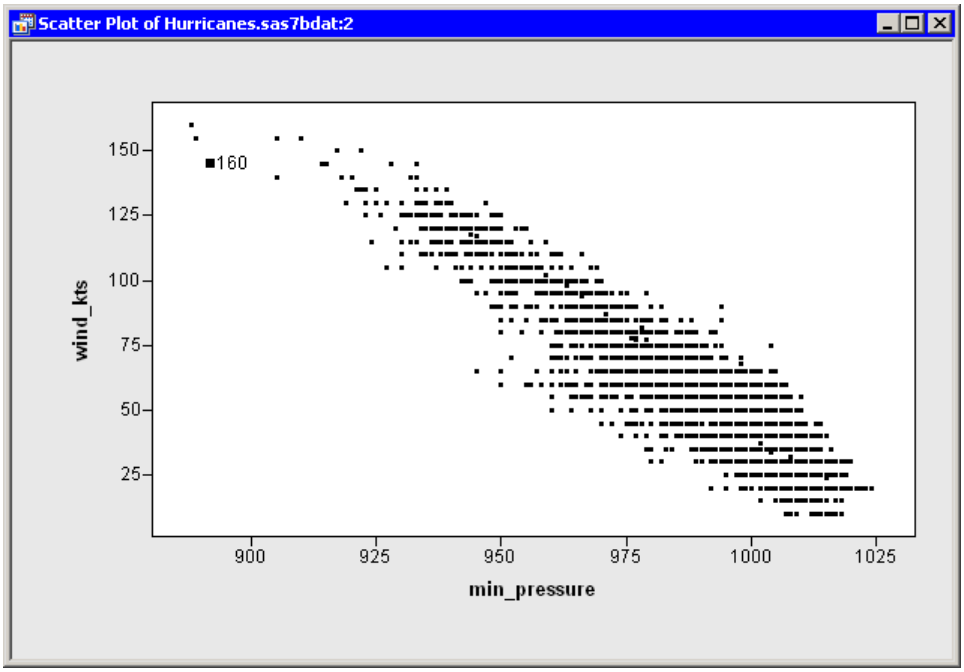
1 Open the Hurricanes data set and create a scatter plot of wind_kts versus min_pressure.

The scatter plot appears. (See Figure 9.14.)

2 Click an observation.

The selected observation is labeled by its position in the data table.

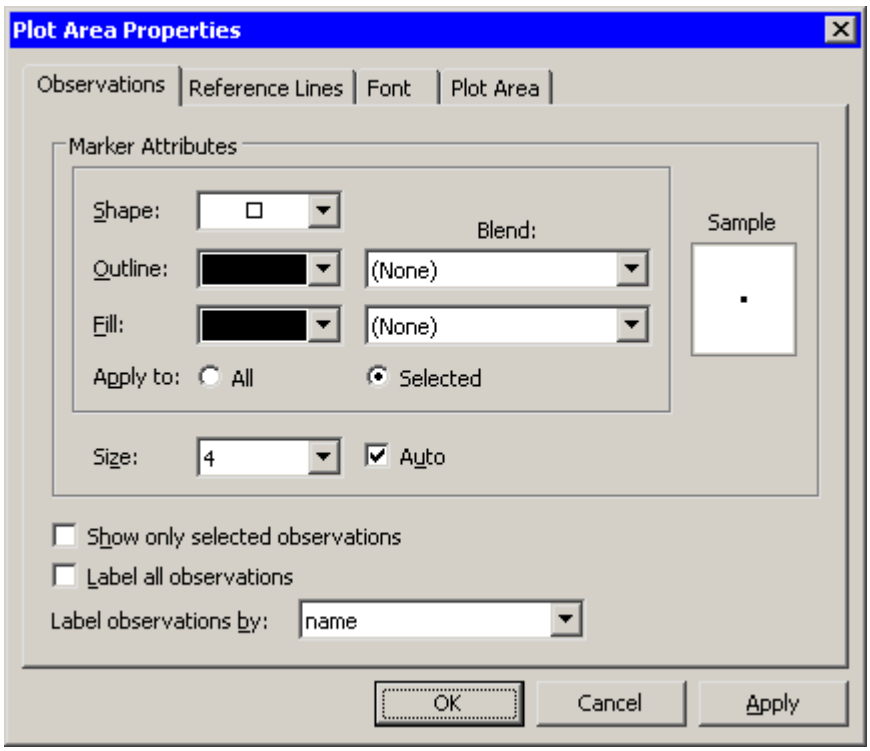
Figure 9.14 A Scatter Plot



3 Right-click near the center of the plot, and select **Plot Area Properties** from the pop-up menu.

The Plot Area Properties dialog box appears. (See Figure 9.15.)

Figure 9.15 The Observations Tab

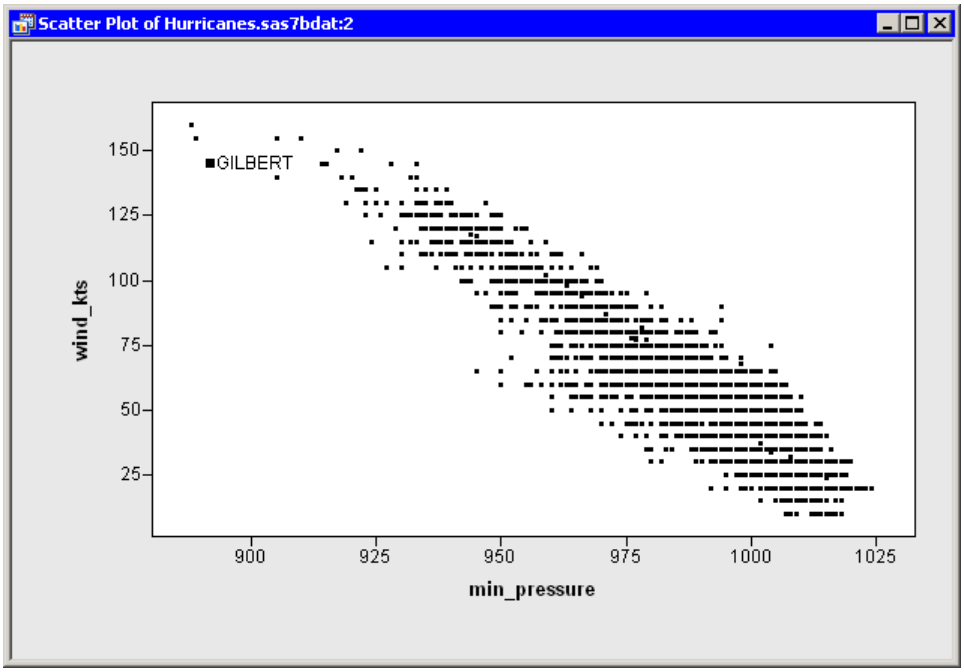


4 Select **name** from the **Label observations by** list.

5 Click **OK**.

The label for the selected observation updates, as shown in Figure 9.16. If you click subsequent observations, each label displays a storm name.

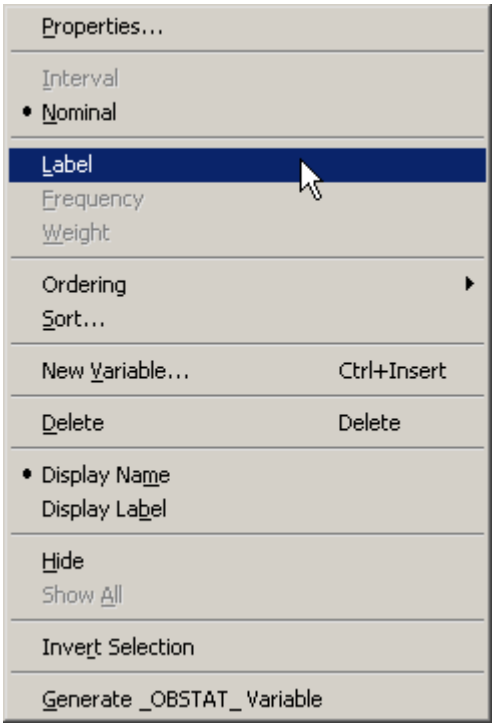
Figure 9.16 Labeling Only Selected Observations



NOTE: When you select **Label observations by** on the Plot Area Properties dialog box, it only affects the scatter plot to which the dialog box belongs. If you create a second plot, that new plot defaults to using observation numbers to label observations.

You can also set a *default label variable* that is used for all plots. In the data table, right-click a variable heading. Select **Label** from the pop-up menu, as shown in Figure 9.17. The values of the selected variable are displayed when you click observations in a plot (unless that plot overrides the default).

Figure 9.17 The Variables Menu



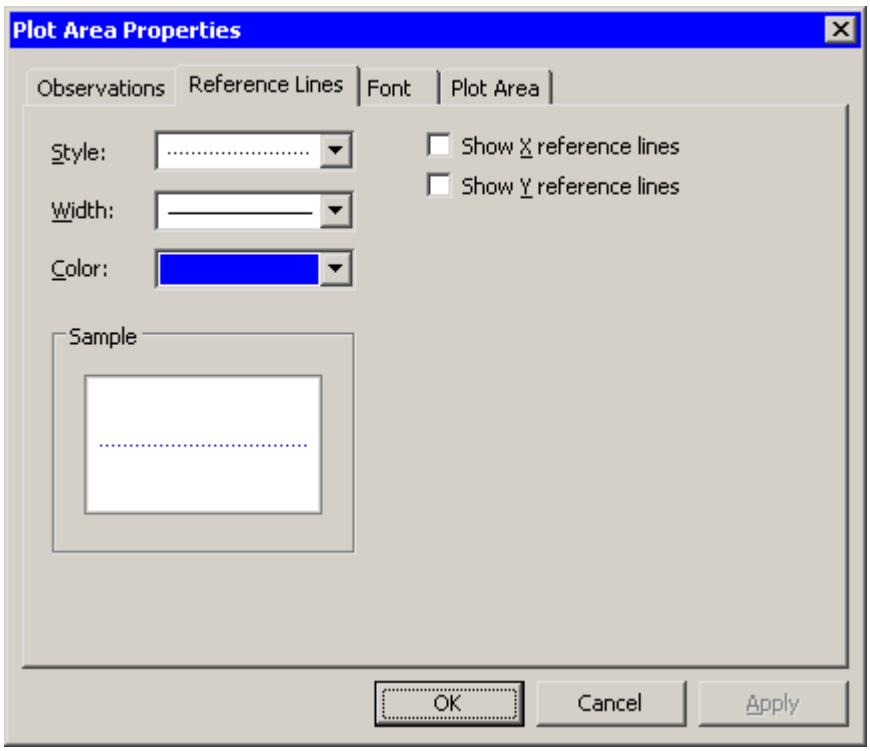
Common Plot Properties

This section presents plot properties that are common to multiple plots. These properties are found in the Plot Area Properties dialog box. You can access the properties by right-clicking near the center of the plot and selecting **Plot Area Properties** from the pop-up menu.

Reference Lines Tab

You can use the **Reference Lines** tab (Figure 9.18) to set attributes of reference lines that are displayed in the background of a plot.

Figure 9.18 The Reference Lines Tab



The **Reference Lines** tab contains the following UI controls:

Style
specifies the style of the line used for reference lines.

Width
specifies the width of the line used for reference lines.

Color
specifies the color of the line used for reference lines.

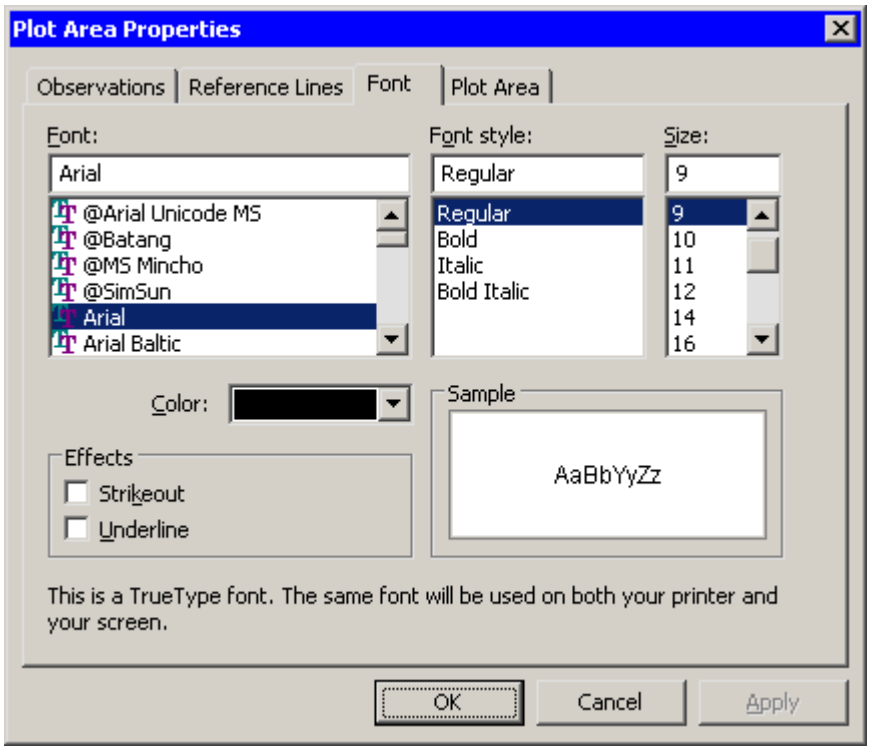
Show X reference lines
specifies whether to show reference lines for the X axis. These are vertical lines that originate at each tick mark on the X axis.

Show Y reference lines
specifies whether to show reference lines for the Y axis. These are horizontal lines that originate at each tick mark on the Y axis.

Font Tab

You can use the **Font** tab (Figure 9.19) to set attributes of the font that is used to display observation labels in plots. The section "Labeling Observations" on page 156 discusses observation labels.

Figure 9.19 The Font Tab



The **Font** tab contains the following UI controls:

Font
: specifies the font used for text in the plot area.

Font style
: specifies the style of the font used for text in the plot area.

Size
: specifies the point size of the text in the plot area.

Color
: specifies the color of the text in the plot area.

Sample
: shows what text with the specified properties looks like.

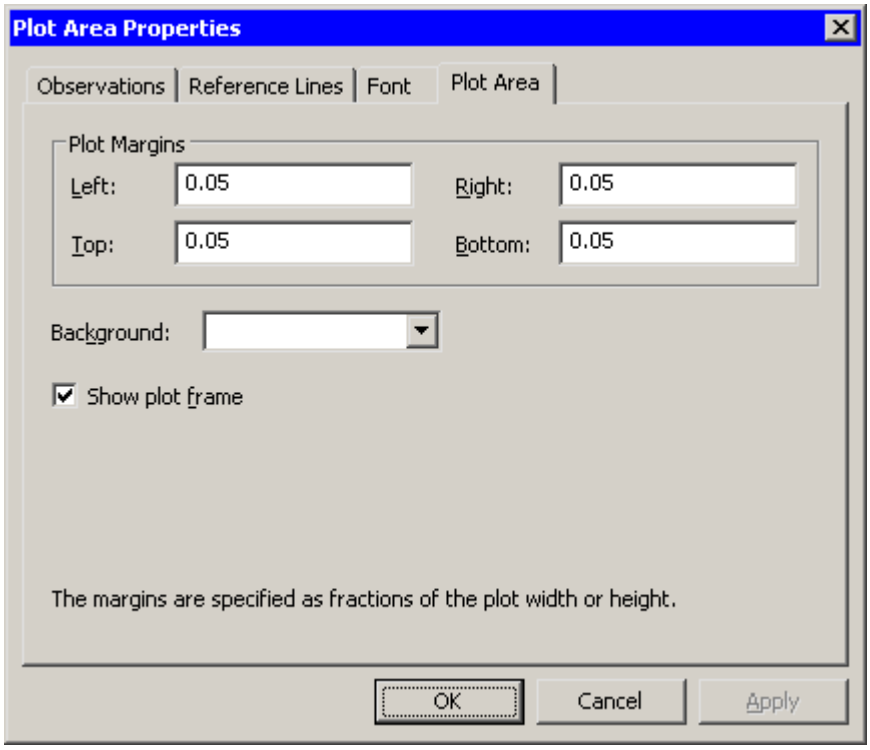
Strikeout
: specifies whether a line is drawn through text in the plot area.

Underline
specifies whether a line is drawn below text in the plot area.

Plot Area Tab

You can use the **Plot Area** tab (Figure 9.20) to set attributes of the plot area.

Figure 9.20 The Plot Area Tab



The **Plot Area** tab contains the following UI controls:

Left
specifies the distance between the left edge of the plot area and the minimum value of the visible axis range for the X axis. The distance is specified as a fraction of the plot area's width. The value must be in the range 0 to 0.8.

Right
specifies the distance between the right edge of the plot area and the maximum value of the visible axis range for the X axis. The distance is specified as a fraction of the plot area's width. The value must be in the range 0 to 0.8.

Top
specifies the distance between the top edge of the plot area and the maximum value of the visible axis range for the Y axis. The distance is specified as a fraction of the plot area's height. The value must be in the range 0 to 0.8.

Bottom
: specifies the distance between the bottom edge of the plot area and the minimum value of the visible axis range for the Y axis. The distance is specified as a fraction of the plot area's height. The value must be in the range 0 to 0.8.

Background
: specifies the background color of the plot area.

Show plot frame
: specifies whether the plot area's frame is displayed.

NOTE: Because the plot area has margins, the edges of the plot area do not correspond to the minimum and maximum values of the axis. Let x_L and x_R be the minimum and maximum values of the horizontal axis. Let m_L and m_R be the left and right margin fractions.

Define $s = (x_R - x_L)/(1 - m_L - m_R)$. Then the left edge of the plot area is located at $x_L - sm_L$, and the right edge of the plot area is located at $x_R + sm_R$.

For example, if $x_L = 0$, $x_R = 1$, $m_L = 1/20$, and $m_R = 1/10$, then $s = 20/17$. The left edge of the plot area is located at $-1/17 \approx -0.0588$, while the right edge is located at $19/17 \approx 1.118$.

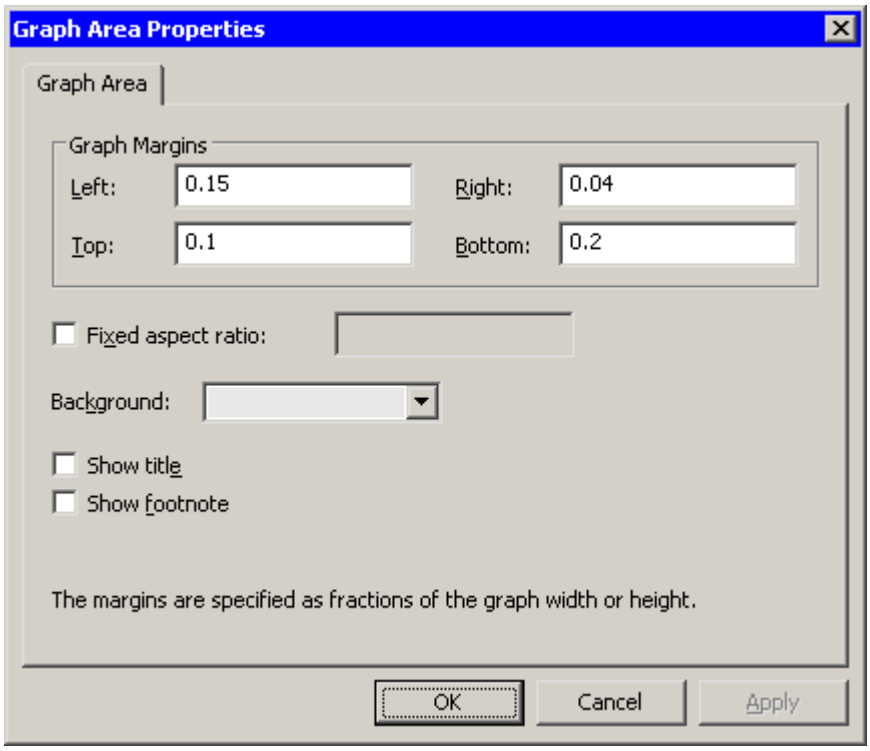
Common Graph Area Properties

This section presents graph area properties that are common to multiple plots. These properties are found in the Graph Area Properties dialog box. You can access the properties by right-clicking near a corner of the graph area and selecting **Graph Area Properties** from the pop-up menu.

Graph Area Tab

You can use the **Graph Area** tab (Figure 9.21) to set attributes of the graph area.

Figure 9.21 The Graph Area Tab



The **Graph Area** tab contains the following UI controls:

Left
specifies the distance between the left edge of the graph and the left edge of the plot area. The distance is specified as a fraction of the graph's width. The value must be in the range 0 to 1.

Right
specifies the distance between the right edge of the graph and the right edge of the plot area. The distance is specified as a fraction of the graph's width. The value must be in the range 0 to 1.

Top
specifies the distance between the top edge of the graph and the top edge of the plot area. The distance is specified as a fraction of the graph's height. The value must be in the range 0 to 1.

Bottom
specifies the distance between the bottom edge of the graph and the bottom edge of the plot area. The distance is specified as a fraction of the graph's height. The value must be in the range 0 to 1.

Fixed aspect ratio
specifies a fixed ratio between units on the Y axis and units on the X axis. When you select this check box, you can specify the ratio. If a plot has a fixed aspect ratio, then the **Graph Margins** are not active.

Background
specifies the background color of the graph area.

Show title
specifies whether the graph's title is displayed.

Show footnote
specifies whether the graph's footnote is displayed.

If you select **Show title**, the graph initially displays a default title. Click the title to edit it. You can also change the title's font or position by right-clicking the title and selecting **Properties** from the pop-up menu. The section "Annotation Properties" on page 138 describes the dialog box that appears.

A default footnote appears when you select **Show footnote**. To edit the footnote, follow the preceding instructions.

If you do not want to display a plot's title or footnote, open the Graph Area Properties dialog box, and clear the appropriate check boxes on the **Graph Area** tab.

Chapter 10
Axis Properties

Contents

Overview of Axis Properties **165**
Adjusting Axes and Ticks **165**
Example: Change Positions of Axis Tick Marks 165
Axis Properties Dialog Box 168
Changing an Axis Label **169**
Suppressing the Display of Axes **171**

Overview of Axis Properties

In this chapter you learn about basic properties of axes. You learn how to change view ranges, tick marks, and labels on axes.

Adjusting Axes and Ticks

In this section you learn how to change the axis range and tick marks for plots.

The section "Example: Change the Positions of Histograms Bins" on page 68 discusses adjusting tick marks for a histogram. For a histogram, the major tick unit is also the width of each histogram bin. Therefore, changing the major tick unit is equivalent to rebinning.

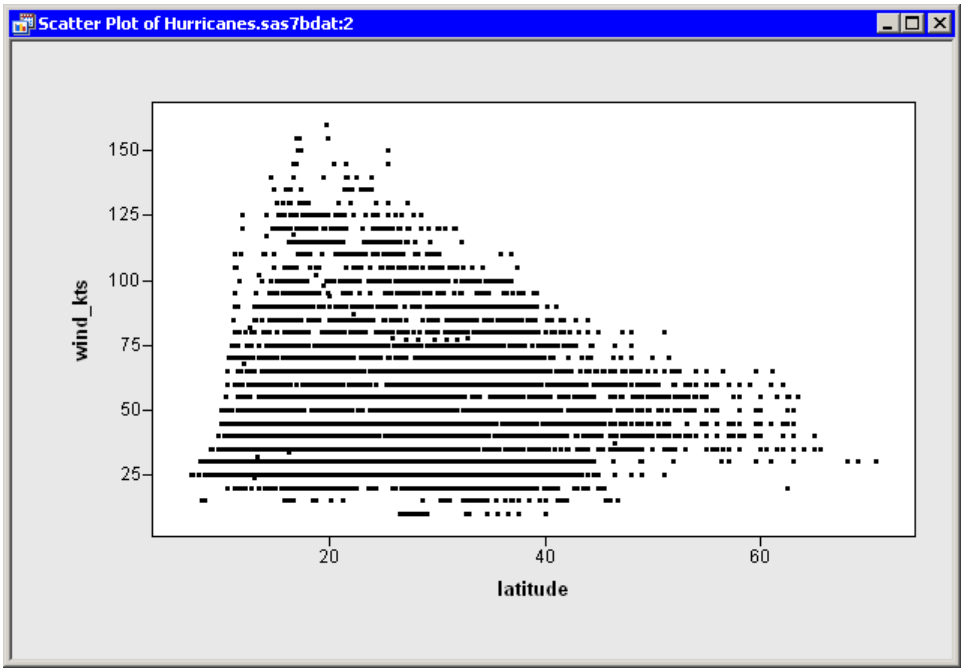
Example: Change Positions of Axis Tick Marks

To change the default tick marks for the axis of an interval variable:

1 Open the Hurricanes data set, and create a scatter plot of wind_kts versus latitude.

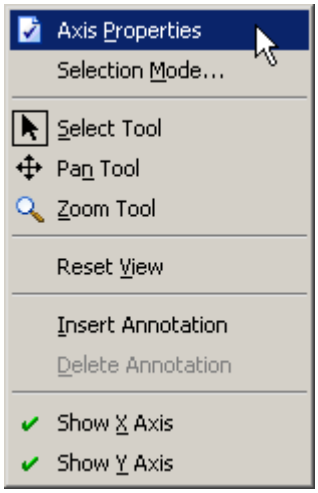
The scatter plot appears. (See Figure 10.1.) Note that the latitude axis has only a few tick marks. You might decide to add a few additional tick marks.

Figure 10.1 A Scatter Plot



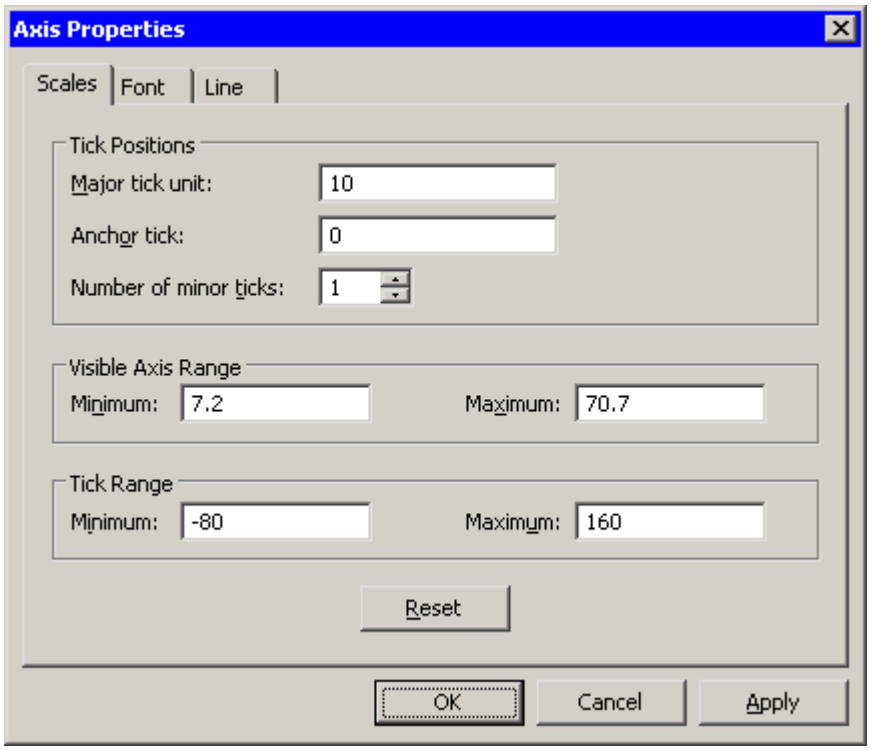
2 Right-click the horizontal axis of the plot, and select **Axis Properties** from the pop-up menu, as shown in Figure 10.2.

Figure 10.2 The Axis Pop-up Menu



The Axis Properties dialog box appears, as shown in Figure 10.3. This is a quick way to determine the anchor location, tick unit, and tick range for an axis.

Figure 10.3 Axis Properties Dialog Box



3 Change the **Major tick unit** value to `10`.

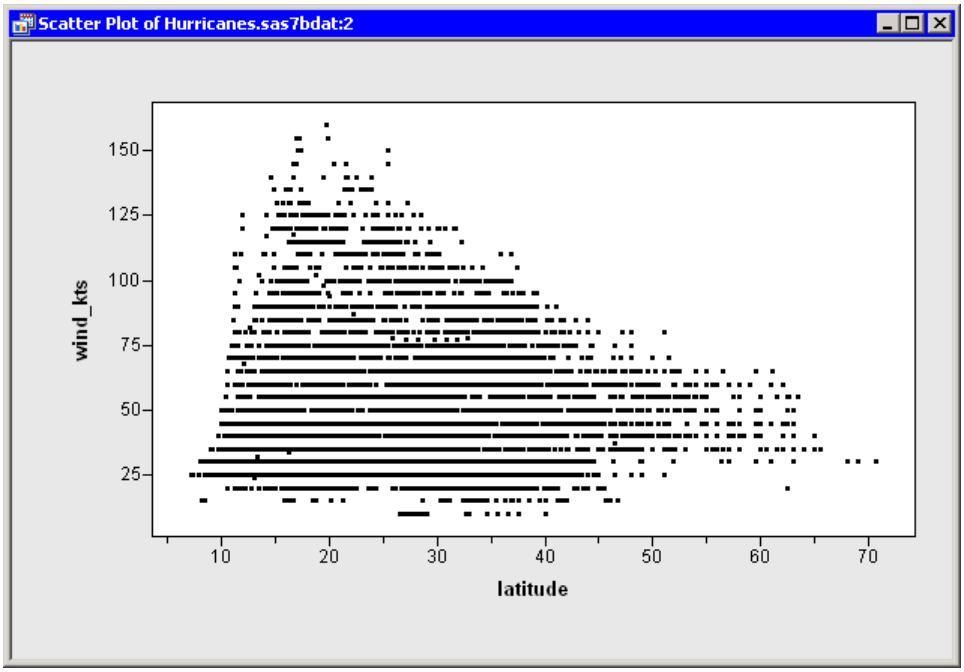
4 Change the **Anchor tick** value to `0`.

5 Change the **Number of minor ticks** value to `1`.

6 Click **OK**.

The latitude axis updates, as shown in Figure 10.4.

Figure 10.4 A Scatter Plot with Custom Tick Marks



Axis Properties Dialog Box

The Axis Properties dialog box controls the appearance of an axis. For an interval variable, major tick marks are placed on an axis within the interval $[L, R]$ at locations $x_0 \pm i\delta$ for integer i. The value x_0 is called the *anchor tick*. The positive quantity δ is called the *major tick unit*. The interval $[L, R]$ is called the *tick range*.

The Axis Properties dialog box has the following tabs: **Scales**, **Font**, and **Line**. The **Scales** tab (Figure 10.3) appears only for interval variables. You can use the **Scales** tab to set tick marks. The **Scales** tab has the following fields:

Major tick unit
sets the distance between tick marks.

Anchor tick
sets the value of one tick mark from which the positions of other ticks are computed.

Number of minor ticks
sets the number of unlabeled tick marks to appear between consecutive major ticks.

Visible Axis Range: Minimum
sets the minimum value of the axis range.

Visible Axis Range: Maximum
sets the maximum value of the axis range.

Tick Range: Minimum
sets the minimum value of a tick mark. Ticks with values less than this value are not displayed.

Tick Range: Maximum
sets the maximum value of a tick mark. Ticks with values greater than this value are not displayed.

NOTE: The minimum and maximum values for **Visible Axis Range** do not necessarily correspond to the edges of the plot area. The plot area also has plot area margins. The computation to find the edges of the plot area is described in the section "Plot Area Tab" on page 161.

The **Font** tab is used to change the font and size of labels on an axis. The **Font** tab is described in section "Common Plot Properties" on page 158.

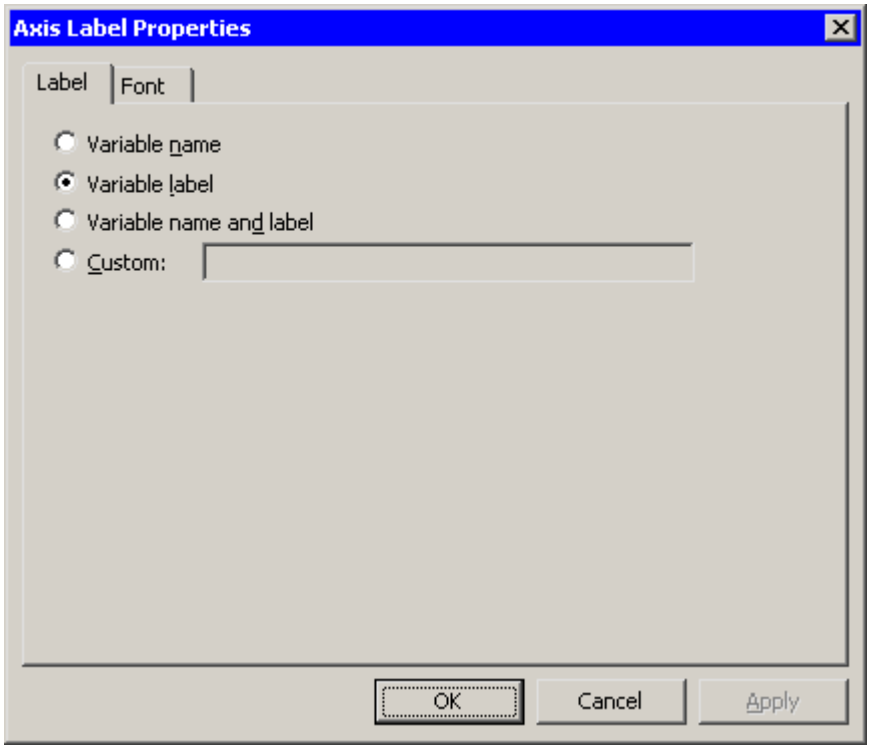
The **Line** tab is used to set the line styles for an axis. The **Line** tab is similar to the **Reference Lines** tab, which is also described in "Common Plot Properties" on page 158.

Changing an Axis Label

An axis label is text near an axis that identifies the axis variable. You can change the axis label. By default, plots display the name of a variable as the label. However, you might prefer that the plot display a variable's label instead of its name. Or you might prefer to customize the axis label in some other way.

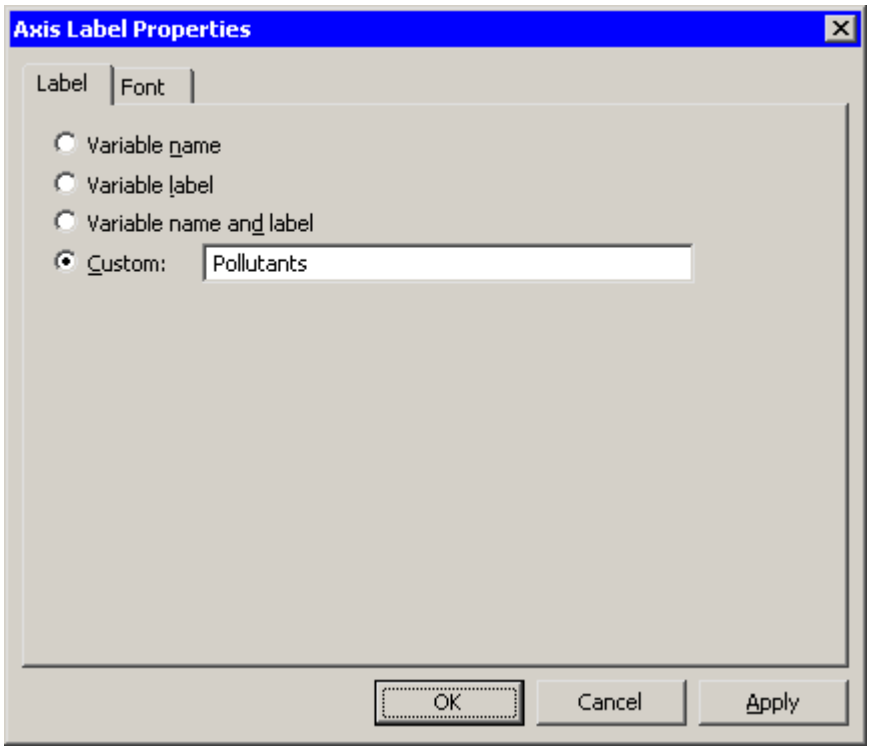
To change the axis label properties, right-click while the mouse pointer is on the axis label. You can then select **Axis Label Properties** from the pop-up menu. The Axis Label Properties dialog box appears, as shown in Figure 10.5.

Figure 10.5 Axis Label Properties Dialog Box



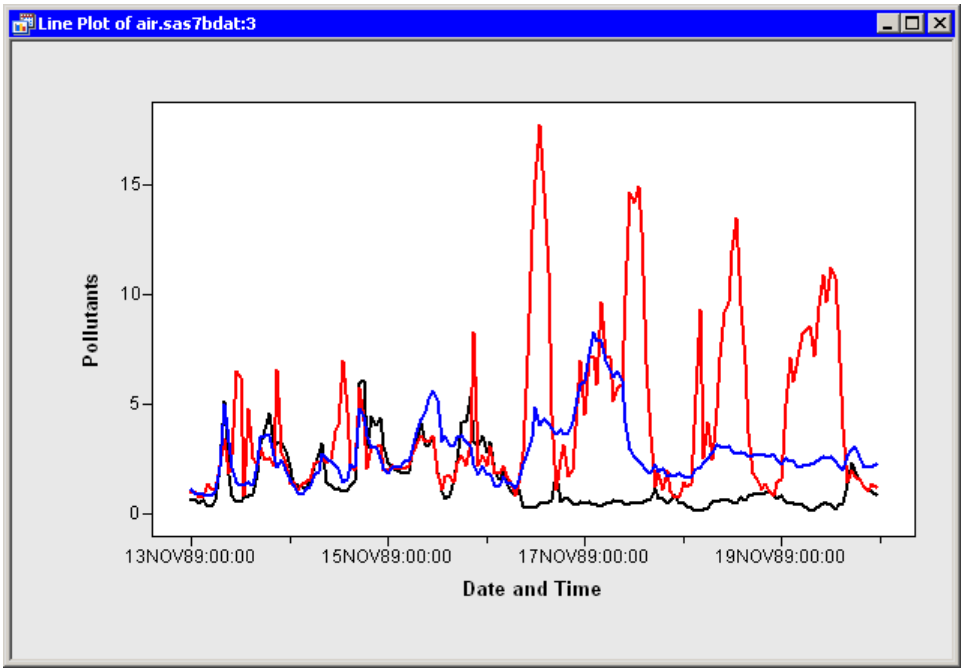
You can display a variable's label instead of the variable's name by selecting **Variable label**. If the variable does not have a label defined, or if you prefer to display a different label, you can select **Custom** and type your own label. This is shown in Figure 10.6.

Figure 10.6 Specifying a Custom Label



One instance in which you might want to define your own label is for a line plot that has multiple Y variables. If the Y variables all measure different aspects of a single quantity, you can replace the multiple Y labels with a single custom label. For example, Figure 10.7 shows a line plot of the co, o3, and so2 variables versus datetime for the Air data set. Each of the Y variables is a kind of pollutant, so the three Y labels are replaced with a single custom label.

Figure 10.7 A Custom Label for the Y Axis



Suppressing the Display of Axes

By default, axes are shown for all plots. However, you can suppress the display of all axes. Right-click in a plot and select **Show X Axis** or **Show Y Axis** from the pop-up menu to toggle the display of an axis and the variable label.

Chapter 11
Techniques for Exploring Data

Contents

Overview of Techniques for Exploring Data **173**
Subsetting Data **174**
Excluding Observations **176**
Ordering Categories of a Nominal Variable **178**
Example: Order Categories in a Bar Chart 179
Local Selection Mode **183**
Overview of Global and Local Selection Modes 183
Example: Select Observations That Satisfy Multiple Criteria 184
Details 189
Workspace Explorer **190**
Example: Manage Multiple Plot Windows 190
Copying Plots to the Windows Clipboard **198**

Overview of Techniques for Exploring Data

This chapter describes some useful techniques for analyzing data in SAS/IML Studio. The following techniques are presented in this chapter:

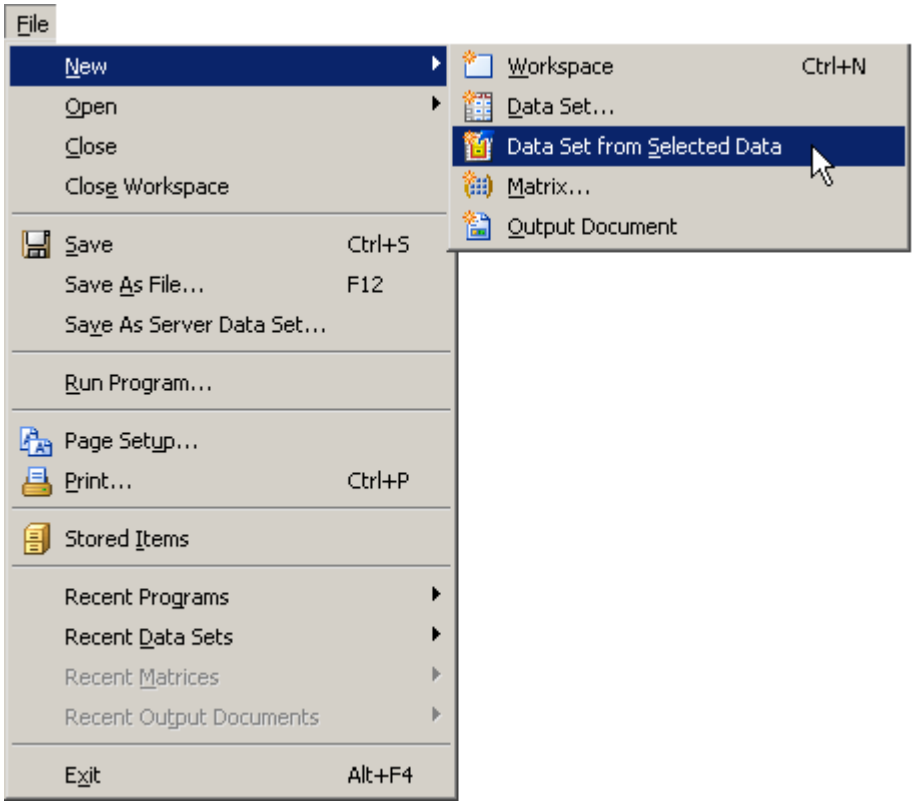
- copying selected observations or variables to a new data table
- excluding observations from plots or analyses
- ordering categories of a nominal variable
- graphically selecting observations that satisfy complex criteria
- managing graphs and workspaces with the Workspace Explorer
- copying plots to the Window clipboard and pasting them to another application, such as Microsoft Word or PowerPoint

Subsetting Data

This section describes how to copy selected observations or variables to a new data table. The new data table is not dynamically linked to the original data. The original data are not changed.

You can copy selected data by selecting **File ►New ►Data Set from Selected Data** from the main menu, as shown in Figure 11.1.

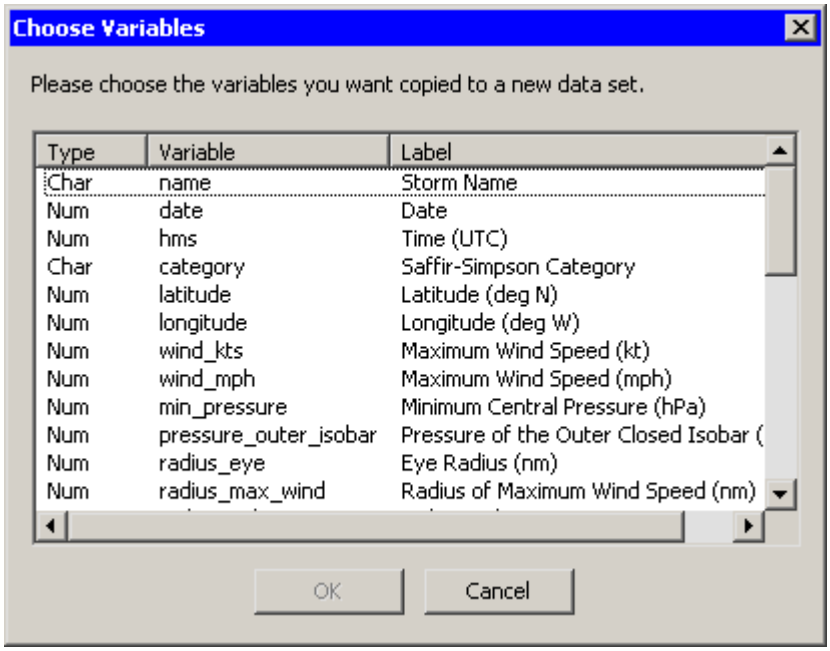
Figure 11.1 Creating a New Data Table from Selected Data



When you select **File ►New ►Data Set from Selected Data**, SAS/IML Studio performs one of the following actions:

- If no variables or observations are selected, the Choose Variables dialog box opens and prompts you to select one or more variables. (See Figure 11.2.) When you click **OK**, the selected variables are copied to a new data table. The variables are copied in the order in which they appear in the original data table.

Figure 11.2 The Choose Variables Dialog Box



- If no variables are selected, but there are selected observations, the selected observations (for all variables) are copied to a new data table. You can use this technique to copy data that satisfy certain conditions.

- If variables are selected, but there are no selected observations, the selected variables are copied to a new data table. The variables are copied in the order in which they were selected. You can use this technique to reorder variables.

- If both variables and observations are selected, the selected observations for the selected variables are copied to a new data table. The variables are copied in the order in which they were selected. For example, in Figure 11.3 the variables were selected in the order longitude, latitude, and category. (Note that the column headings display numbers that indicate the order in which you selected the variables.) If you copy the data to a new data table, the new data table contains 12 observations for the longitude, latitude, and category variables, in that order.

Figure 11.3 Selected Variables and Observations

Hurricanes

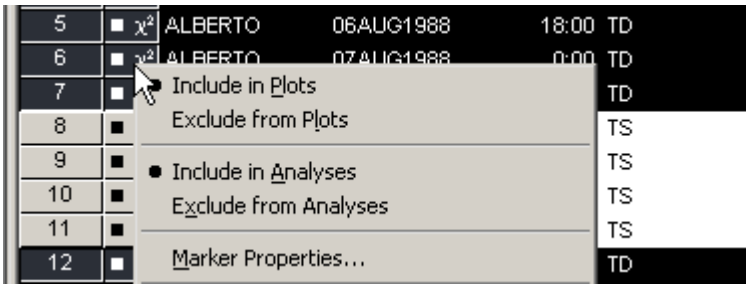
	[3]	name	date	hms	category	latitude	longitude	wind_kts
[12]		Nom	Int	Int	3 Nom	2 Int	1 Int	Int
1	χ²	ALBERTO	05AUG1988	18:00		32	-77.5	20
2	χ²	ALBERTO	06AUG1988	0:00		32.8	-76.2	20
3	χ²	ALBERTO	06AUG1988	6:00		34	-75.2	20
4	χ²	ALBERTO	06AUG1988	12:00	TD	35.2	-74.6	25
5	χ²	ALBERTO	06AUG1988	18:00	TD	37	-73.5	25
6	χ²	ALBERTO	07AUG1988	0:00	TD	38.7	-72.4	25
7	χ²	ALBERTO	07AUG1988	6:00	TD	40	-70.8	30
8	χ²	ALBERTO	07AUG1988	12:00	TS	41.5	-69	35
9	χ²	ALBERTO	07AUG1988	18:00	TS	43	-67.5	35
10	χ²	ALBERTO	08AUG1988	0:00	TS	45	-65.5	35
11	χ²	ALBERTO	08AUG1988	6:00	TS	47	-63	35
12	χ²	ALBERTO	08AUG1988	12:00	TD	49	-60	30
13	χ²	ALBERTO	08AUG1988	18:00	TD	51	-56	25
14	χ²	BERYL	08AUG1988	0:00	TD	30.4	-90.3	25
15	χ²	BERYL	08AUG1988	6:00	TD	29.7	-89.7	30

Excluding Observations

This section describes how to exclude selected observations from plots and from statistical analyses. The data table must be the active window in order for you to exclude observations. Select **Edit ►Observations ►Exclude from Plots** from the main menu to exclude selected observations from plots. Select **Edit ►Observations ►Exclude from Analyses** to exclude selected observations from analyses.

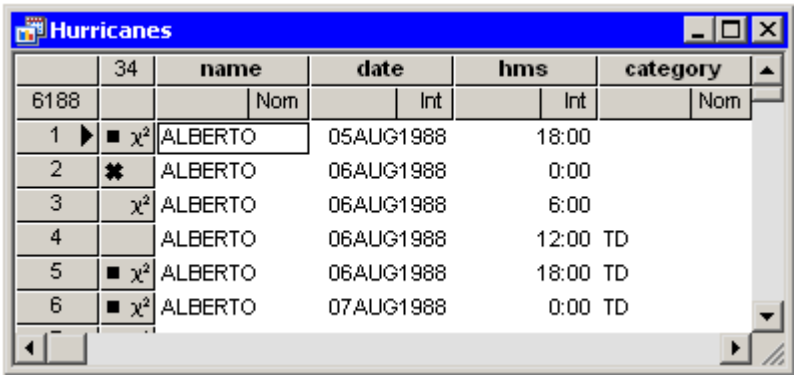
Alternatively, you can right-click the row heading of any selected observation in the data table and select **Exclude from Plots** or **Exclude from Analyses** from the pop-up menu, as shown in Figure 11.4.

Figure 11.4 Data Table Pop-up Menu



The row heading of the data table shows the status of an observation in analyses and plots. A marker symbol indicates that the observation is included in plots; observations excluded from plots do not have a marker symbol shown in the data table. Similarly, the χ^2 symbol is present if and only if the observation is included in analyses. For example, the first, fifth, and sixth observations in Figure 11.5 are included in plots and analyses.

Figure 11.5 Excluded Observations



Hurricanes

	34	name	date	hms	category
6188		Nom	Int	Int	Nom
1	■ χ^2	ALBERTO	05AUG1988	18:00	
2	✖	ALBERTO	06AUG1988	0:00	
3	χ^2	ALBERTO	06AUG1988	6:00	
4		ALBERTO	06AUG1988	12:00	TD
5	■ χ^2	ALBERTO	06AUG1988	18:00	TD
6	■ χ^2	ALBERTO	07AUG1988	0:00	TD

If you exclude observations from plots, all plots linked to the current data table automatically redraw themselves. (For example, excluding an extreme value might result in a new range for an axis.) The row headings for the excluded observations no longer show the observation marker. For example, the third and fourth observations in Figure 11.5 are excluded from plots.

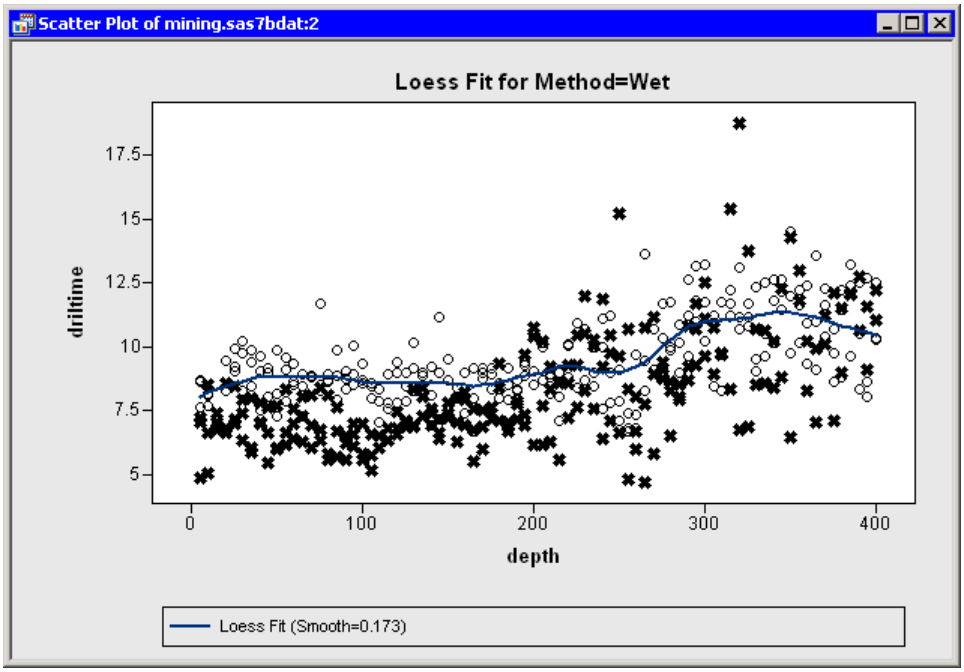
If you exclude observations from analyses, the row headings for the excluded observations no longer show the χ^2 symbol. For example, the second and fourth observations in Figure 11.5 are excluded from analyses.

NOTE: If you change the observations that are included in analyses, previously run analyses and statistics are *not* automatically rerun.

If an observation is excluded from analyses but included in plots, then the marker symbol changes to the $\times$ symbol. This combination is useful if you want to fit a regression model to data but also want to exclude outliers or high-leverage observations prior to modeling. The regression model does not use the excluded observations, but the observations show up (as $\times$) on diagnostic plots for the regression.

An example of including some observations in plots but not in analyses is shown in Figure 11.6. The figure shows data from the Mining data set—the results of an experiment to determine whether drilling time was faster for wet drilling or dry drilling. The plot shows the time required to drill the last five feet of a hole plotted against the depth of the hole. A loess fit is plotted only for the wet drilling trials (open circles). This is accomplished by excluding the observations for dry drilling (markers with the $\times$ shape) before running the loess analysis.

Figure 11.6 Loess Fit of a Subset of Data



Although SAS/IML Studio analyses do not support BY-group processing, you can restrict an analysis to a single BY group by excluding all other BY groups. For data with many BY groups, this is tedious to do using the SAS/IML Studio GUI, but you can write an IMLPlus program to automate the processing of BY groups.

You easily restore all observations into plots and analyses:

1. Activate the data table. Press CTRL+A. This selects all observations in the table.
2. Select **Edit ►Observations ►Include in Plots** from the main menu.
3. Select **Edit ►Observations ►Include in Analyses** from the main menu.

Ordering Categories of a Nominal Variable

This section describes how to specify the order of categories for a nominal variable. You cannot change the order of values for interval variables.

By default, numeric nominal variables are ordered numerically, whereas character nominal variables are arranged in ASCII order. In ASCII order, numerals precede uppercase letters, which precede lowercase letters. Even if a variable has a SAS format, SAS/IML Studio determines the default order of categories by using the ASCII order of the *unformatted* values.

When the data table is active, you can select **Edit ►Variables ►Ordering** to change the order of categories for a nominal variable. You can order nominal variables according to the ASCII order of values, the frequency count of values, or the data order of values. For each ordering, you can specify whether to base the order on formatted or unformatted values. Therefore, there are six possible ways to order a nominal variable. Four of these orderings are the same as provided by the ORDER= option of the FREQ procedure. An ordering determines the order of categories in a plot (for example, a bar chart) and also the order of sorted observations when sorting a variable in a data table.

As an example, consider the data presented in Table 11.1.

Table 11.1 Sample Data

Observation	Y
1	C
2	B
3	C
4	a
5	a
6	a

The Y variable has three categories: `a`, `B`, and `C`. The ASCII order of this data is {`B`, `C`, `a`} because uppercase letters precede lowercase letters. The data order is {`C`, `B`, `a`} because as you traverse the data from top to bottom, `C` is the first value you encounter, followed by `B`, followed by `a`. The order by frequency count is {`a`, `C`, `B`}, because there are three observations with the value `a`, two with the value `C`, and one with the value `B`.

If you specify an ordering based on formatted values when the variable does not have a SAS format, then SAS/IML Studio applies either a BEST12. format (for numeric variables) or a $*w*. format (for character variables).

When a variable has missing values, the missing values are always ordered first.

Example: Order Categories in a Bar Chart

To create a bar chart of the category variable in the Hurricanes data set:

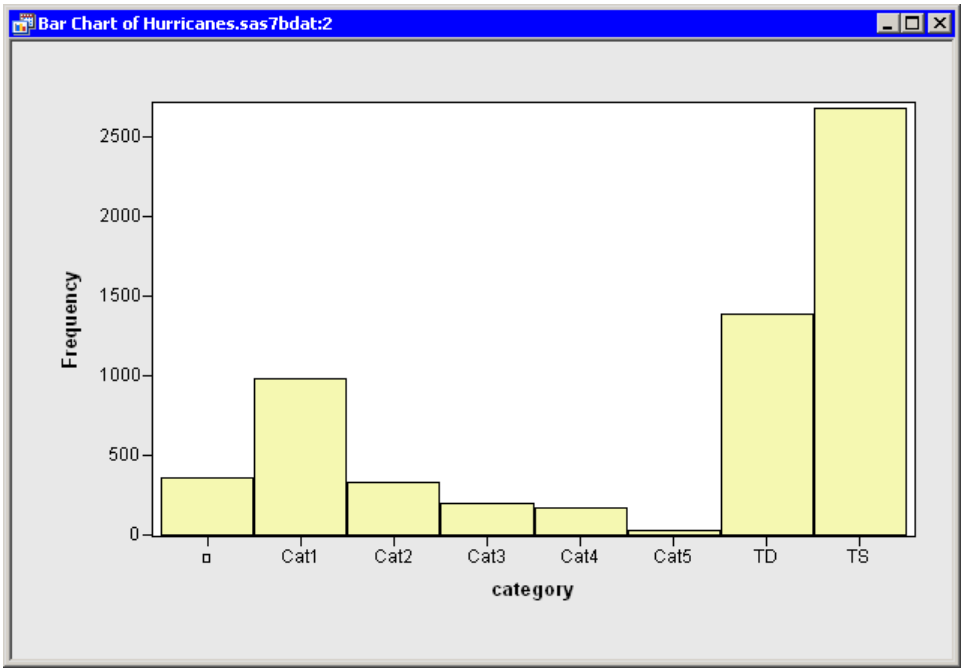
1. Open the Hurricanes data set.

   Note that the column heading for the category variable displays **Nom** to indicate that the variable is nominal.

2. Create a bar chart of the category variable.

   The bar chart is shown in Figure 11.7. Note that the first category consists of missing values, and the other categories appear in standard ASCII order.

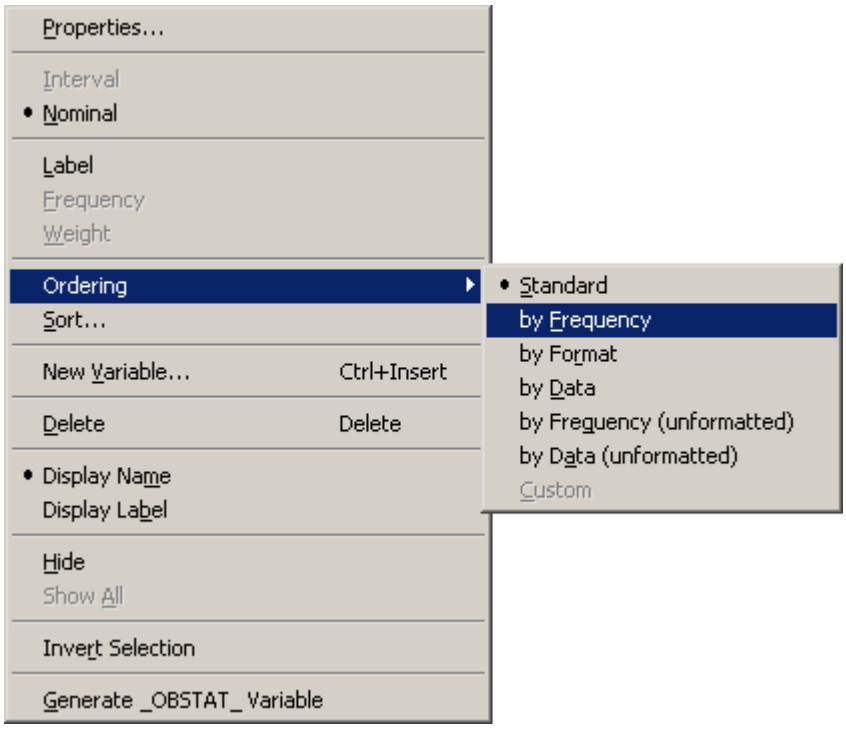
Figure 11.7 Standard Ordering of the Category Data



When you explore data, it is useful to be able to reorder data categories. The next step arranges the bar chart categories according to frequency counts.

3 Right-click in the data table on the column heading for the category variable. Select **Ordering ►by Frequency**, as shown in Figure 11.8.

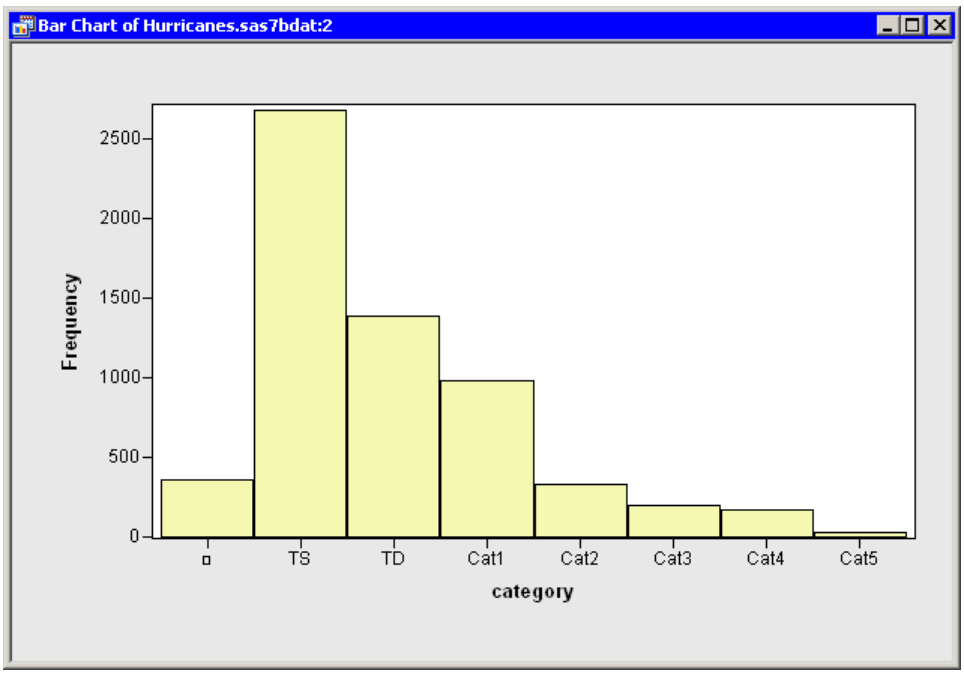
Figure 11.8 Ordering by Frequency Count



The bar chart automatically updates, as shown in Figure 11.9. Note that the first bar still represents missing values, but that the remaining bars are ordered by their frequency counts. This presentation of the plot makes it easier to compare the relative frequencies of categories.

Note that the column heading for the category variable now displays **Ord** to indicate that this variable has a nonstandard ordering.

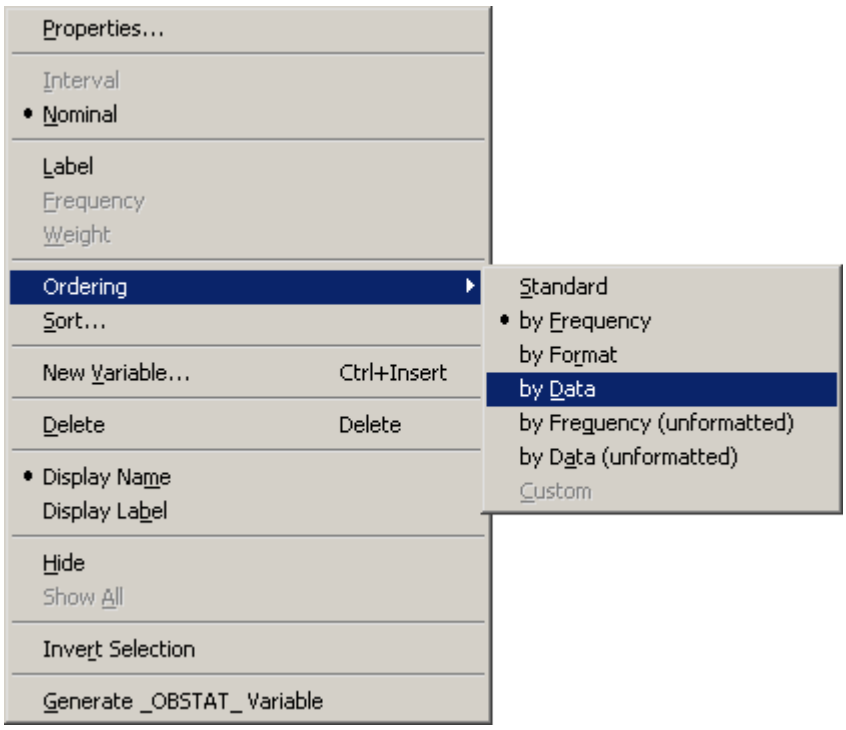
Figure 11.9 The Category Data Ordered by Frequency Count



The next step arranges the bar chart categories according to the data order of the seven nonmissing categories.

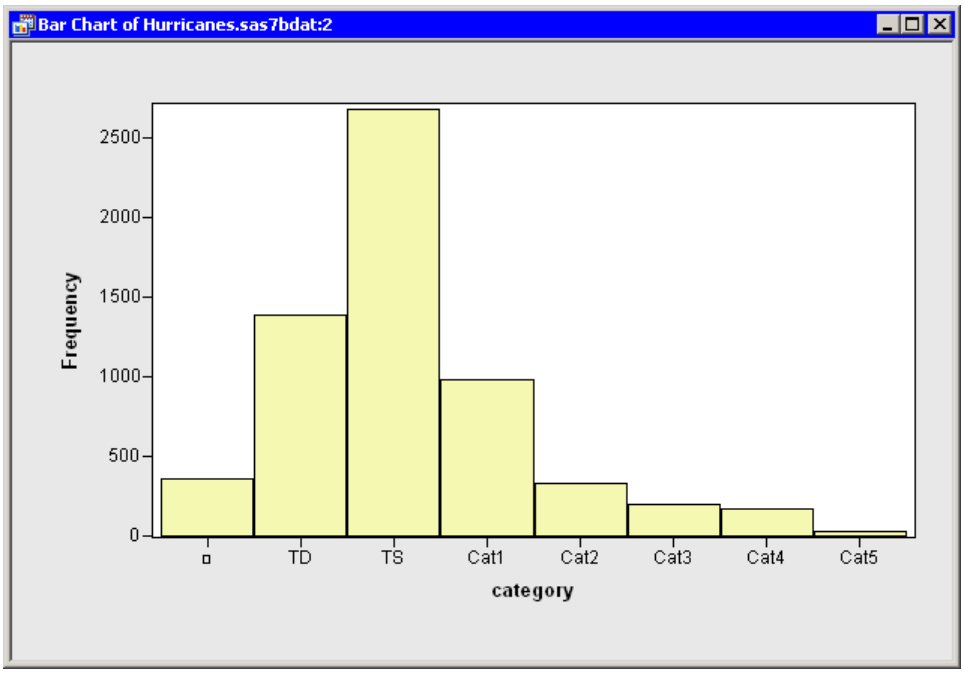
4 Right-click in the data table on the column heading for the category variable. Select **Ordering ►by Data**, as shown in Figure 11.10.

Figure 11.10 Ordering by Data Set Position



The bar chart automatically updates, as shown in Figure 11.11. As always, the first bar represents missing values. The TD category is ordered next, because TD is the first nonmissing value for the category variable. The next category is TS, because as you traverse the data (starting from the first observation) the next unique value you encounter is TS (the eighth observation). The remaining categories are Cat1 (the 72nd observation), Cat2 (the 148th observation), Cat3 (the 149th observation), Cat4 (the 155th observation), and Cat5 (the 157th observation).

Figure 11.11 The Category Data Ordered by Data Set Position



Arranging values by their data order is sometimes useful when the values are inherently ordered. For example, suppose you have a variable Y with the values Low, Medium, and High. The ASCII order for these categories is {High, Low, Medium}. A plot that displays the categories in this order might be confusing.

To deal with this problem:

1. Create a numerical indicator variable with the values {1, 2, 3} that corresponds to observations with the values {Low, Medium, High} for Y. The section "Custom Transformations" on page 518 describes how to create an indicator variable.
2. Sort the data by the indicator variable.
3. Save the sorted data.
4. Close your workspace.
5. Open the sorted data.
6. Right-click the column heading for the variable, and select **Ordering ►by Data**.

Plots of the Y variable display the categories in the order {Low, Medium, High}.

Although you can use the previous steps to order any single variable, you might not be able to order multiple variables simultaneously using this method. In that case, you should consult the online Help and read about the DataObject.SetVarValueOrder method.

Local Selection Mode

This section describes how to use graphical methods to visualize observations that satisfy multiple conditions simultaneously.

Overview of Global and Local Selection Modes

SAS/IML Studio supports two techniques for selecting observations:

- *Global selection mode* is the traditional selection technique used in SAS/INSIGHT and other products. This is the default selection mode in SAS/IML Studio. In global selection mode, all *data views* (that is, plots or data tables) share a common selection state for observations. When you select an observation in one view, that observation is treated as selected in all other views.

  Global selection mode enables you to graphically subset data by interacting with a single data view. For example, if you have three plots called A, B, and C, selecting observations in plot A causes plots B and C to display those same observations as selected.

- In contrast, *local selection mode* enables you to subset data by interacting with multiple data views. In local selection mode, you specify each data view to be either a *selector* or an *observer*. You configure an observer to display either the union or the intersection of the selected observations in all selector

views. For example, if you have three plots called A, B, and C, you can configure plot C to be the "observer of the intersection" of the other plots. This means that an observation is selected in plot C only if it is selected in both plot A and plot B.

You can manually select observations in selector views. You cannot manually select observations in an observer view. An observer view displays an observation as selected based on the observation's selection state in the selector views. An "observer of the union" displays an observation as selected if the observation is selected in *any* of the selector views. An "observer of the intersection" displays an observation as selected if the observation is selected in *all* of the selector views.

Example: Select Observations That Satisfy Multiple Criteria

In this section, you create several plots of variables in the Hurricanes data set. You use local selection mode to display the wind speed and pressure of tropical cyclones that satisfy certain spatial and temporal criteria.

To use local selection mode:

1. Open the Hurricanes data set.
2. Create a histogram of the latitude variable.

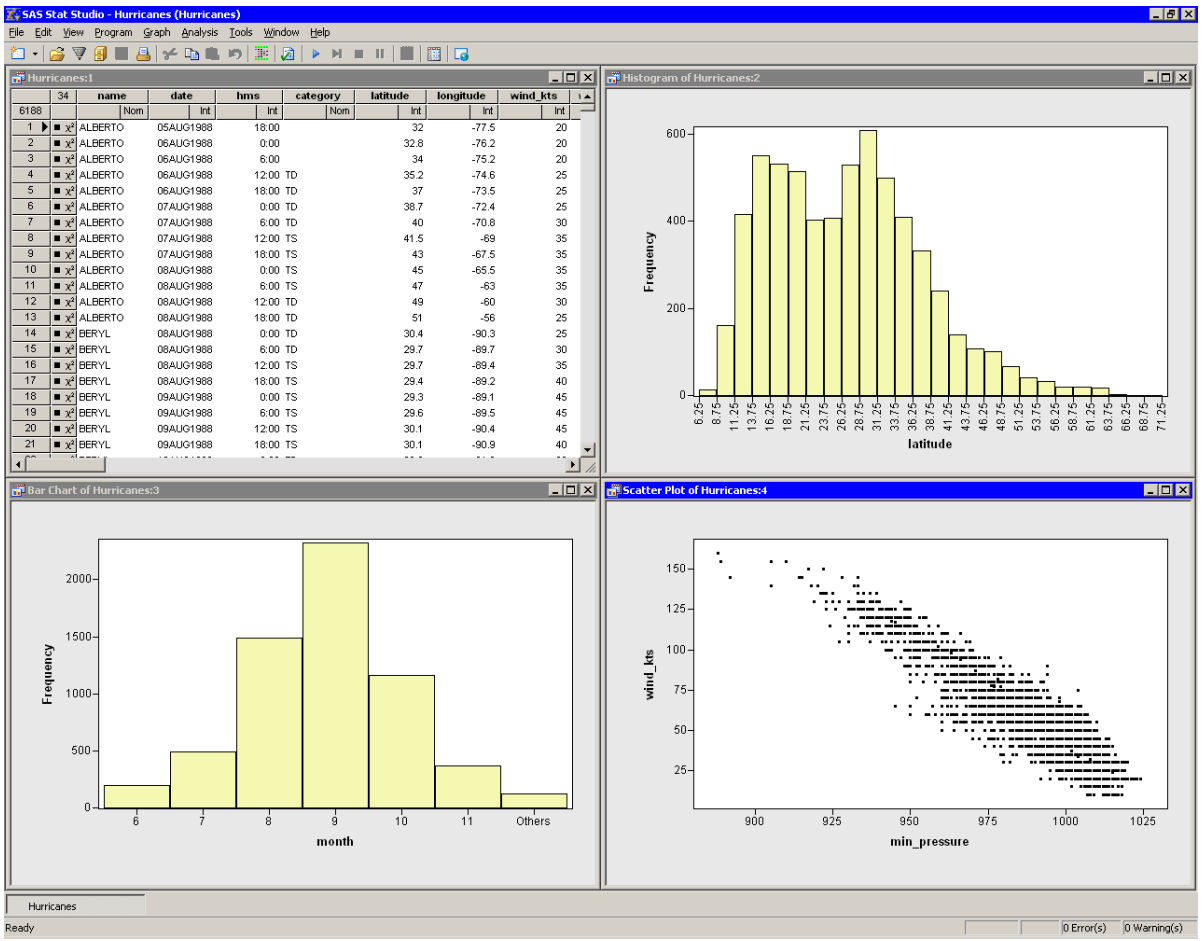
   The histogram becomes one of the selector views.

   The next plot to create is a bar chart of the month variable. By default, the month variable is an interval (continuous) variable. In order to create a bar chart, you first need to change the measure level from interval to nominal.
3. Scroll the data table horizontally until you see the month variable.
4. Right-click the heading of the month column, and select **Nominal** from the pop-up menu.
5. Create a bar chart of the month variable. Move the bar chart so that it does not overlap other data views.

   The bar chart becomes a second selector view.
6. Create a scatter plot of wind_kts versus min_pressure. Move the plot so that it does not overlap other data views.

   The scatter plot becomes an observer view. The workspace now looks like Figure 11.12.

Figure 11.12 Global Selection Mode

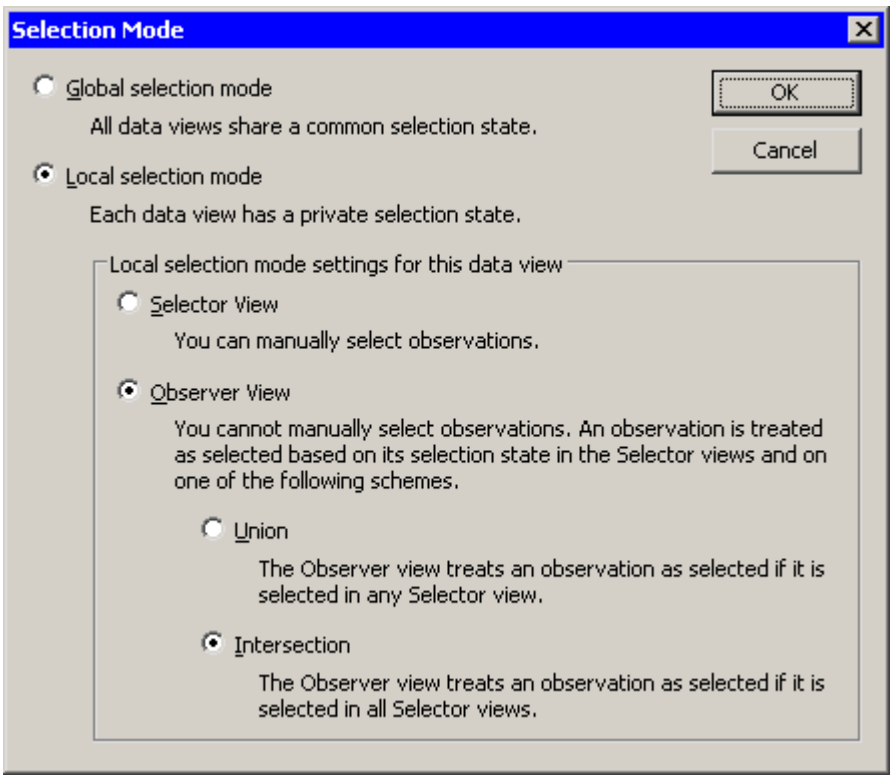


7 Close the data table.

8 Right-click the plot area in the scatter plot. Select **Selection Mode** from the pop-up menu.

The dialog box shown in Figure 11.13 appears.

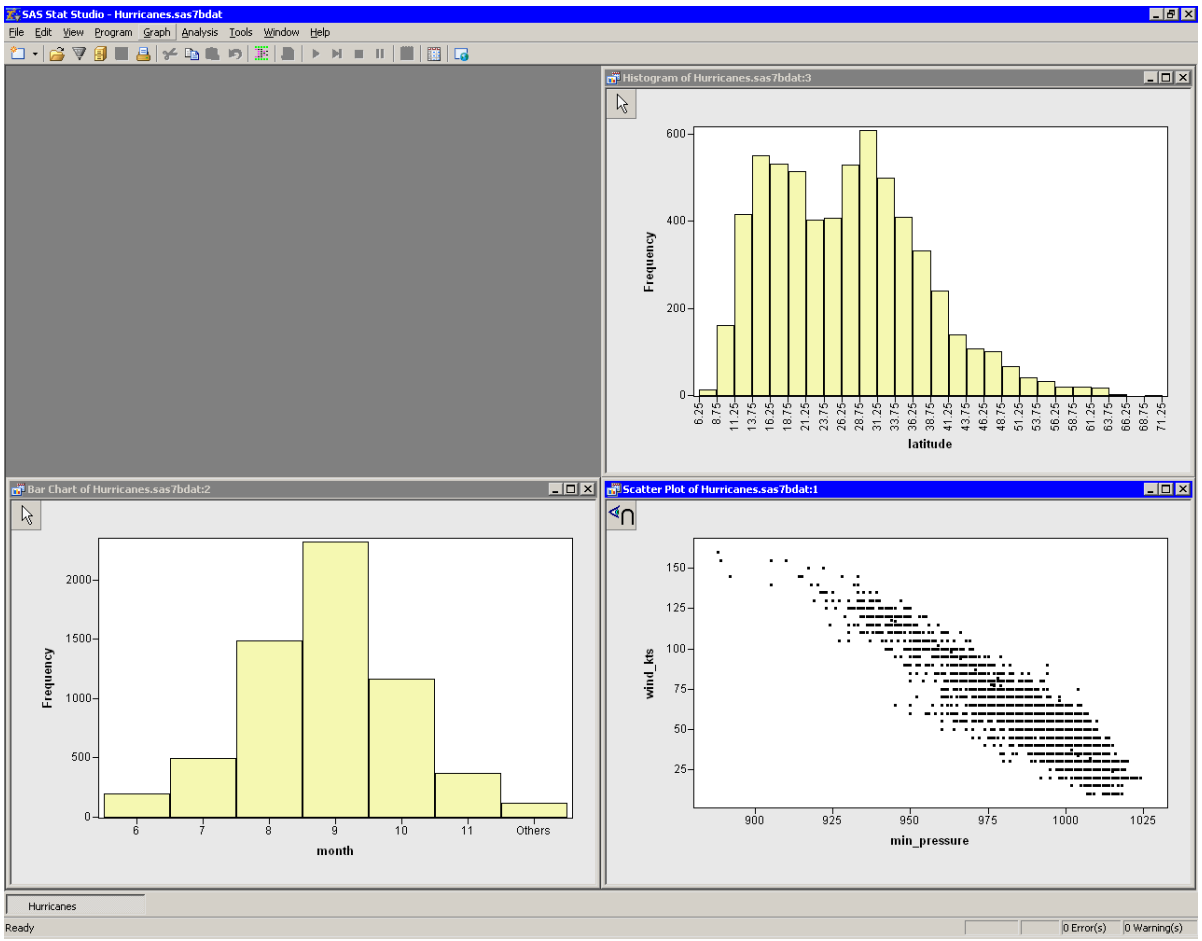
Figure 11.13 Selection Mode Dialog Box



9 Click **Local Selection Mode**, **Observer View**, and **Intersection**. Click **OK**.

The workspace now looks like Figure 11.14. The scatter plot is an observer view. All of the other data views were set to be selector views when you entered local selection mode. Note that selector views are indicated by an arrow icon in the upper left corner of the view. Observer views are indicated by an icon that looks like an eye looking at the mathematical symbol for intersection (or union).

Figure 11.14 Local Selection Mode



NOTE: If you forget to close the data table, then it, too, is a selector view. A common error is to leave the data table open. If the data table is left open, then no observations are selected in the observer scatter plot unless they are selected in *all* other selector views, including data tables.

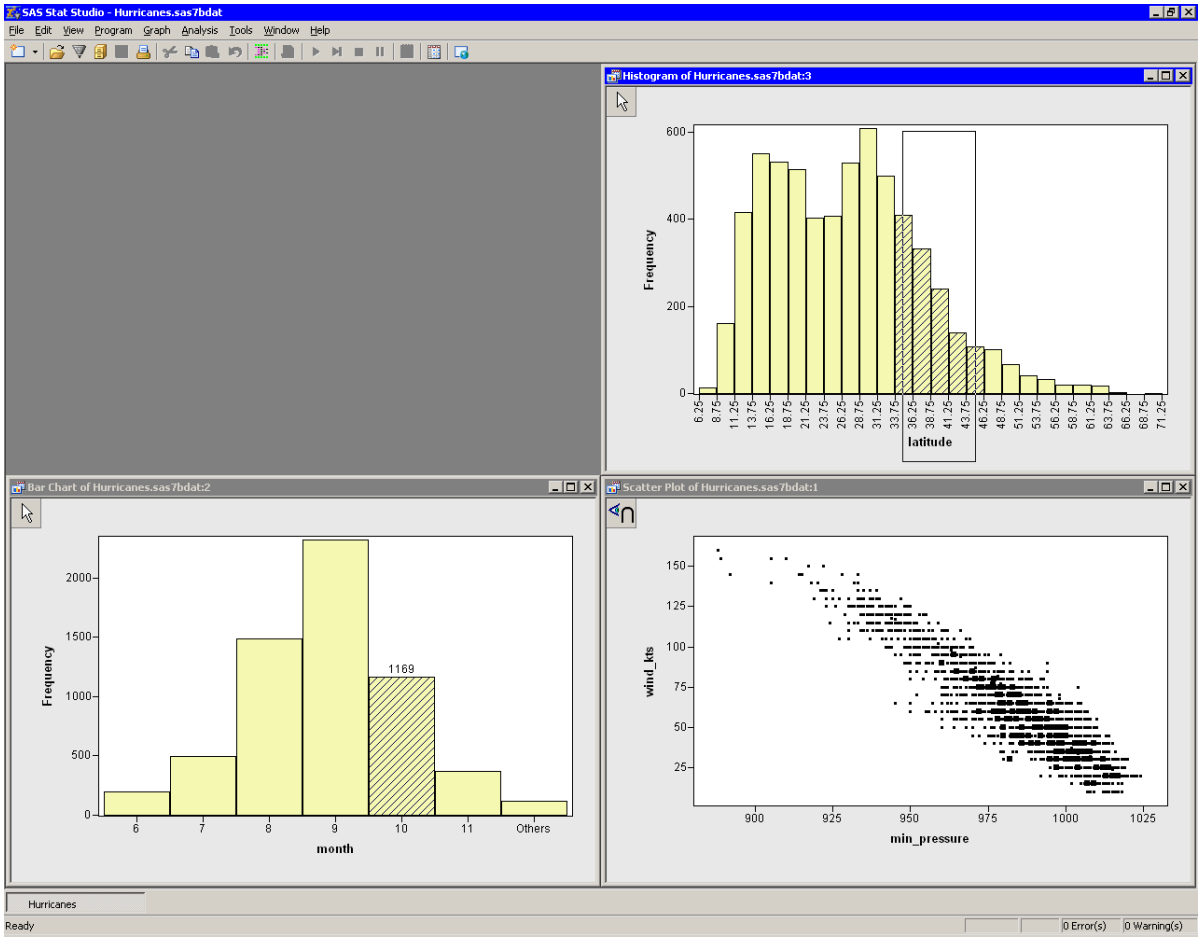
10 In the bar chart, click the bar labeled “10” to select observations that correspond to the tenth month (October).

The histogram does not display any observations because it is a selector view. The scatter plot does not display any observations because it is an observer view; it displays observations as selected only if they are selected in all selector views.

11 Create a selection rectangle in histogram. Move it around the plot.

The workspace now looks like Figure 11.15.

Figure 11.15 Displaying the Intersection of Multiple Selector Views



The observations displayed as selected in the scatter plot are those that are selected in both the bar chart and the histogram. The selected observations in the scatter plot in Figure 11.15 are those tropical storms that occurred in October (month = 10) of any year and whose position was between 33.75 and 46.25 degrees north latitude.

Details

This section describes the Selection Mode dialog box, shown in Figure 11.13. To open the Selection Mode dialog box, right-click on a plot or data table, and select **Selection Mode** from the pop-up menu. Alternatively, click a data view's title bar to activate it, and select **Edit ►Selection Mode** from the main menu.

The Selection Mode dialog box has the following fields:

Global selection mode
sets the selection mode to be global selection mode.

Local selection mode
sets the selection mode to be local selection mode. The active window becomes either a selector view or an observer view. All other data views linked to the active window become selector views.

Selector View
sets the active window to be a selector view.

Observer View
sets the active window to be an observer view.

Union
sets the active window to be an observer of the union of selector views. An observation is displayed as selected if it is selected in any selector view.

Intersection
sets the active window to be an observer of the intersection of selector views. An observation is displayed as selected if it is selected in all selector views.

The following list presents a few additional details about using local selection mode:

- There is a limit of 31 selector views that can be linked to an observer view. There is no limit to the number of observer views.
- It is often useful to have multiple selector views but only one observer view. In this case it is quickest to activate the plot that is to become the observer view, and then to select **Edit ►Selection Mode** from the main menu. Configure that plot as a local observer view, and click **OK**. All of the other data views are automatically changed to selector views. This technique was used in the example.
- If the observer view is a plot that displays individual observation markers (for example, a scatter plot), it is often useful to configure the plot to show only the selected observations. See the section "Displaying Only Selected Observations" on page 152 for details. This technique is sometimes called *graphical filtering*, because selected observations do not "reach" the observer view until they have passed through all of the "filters" (criteria) imposed by the selector views.

Workspace Explorer

In SAS/IML Studio, it is easy to generate a large number of plots. Keeping track of the plots that are associated with an analysis can be a challenge. Manually closing or minimizing a large number of plots can be tedious. Finding a particular plot from among a large number of plots can be cumbersome. The Workspace Explorer helps solve all of these potential problems.

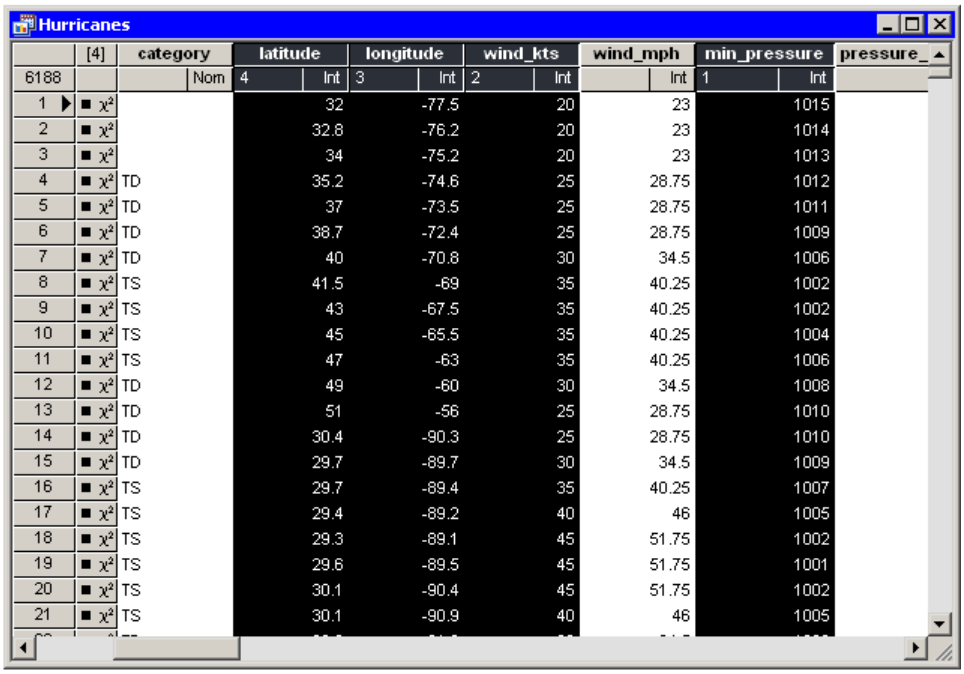
Example: Manage Multiple Plot Windows

In this section, you create many plots of variables in the Hurricanes data set. The following steps use the Workspace Explorer to manage the display of plots:

1. Open the Hurricanes data set.
2. Scroll the data table horizontally until the min_pressure variable appears. Hold down the CTRL key while you select the min_pressure, wind_kts, longitude, and latitude variables, in that order.

   Figure 11.16 shows the selected variables. Note that the column headings display numbers that indicate the order in which you selected the variables.

Figure 11.16 Selecting Variables



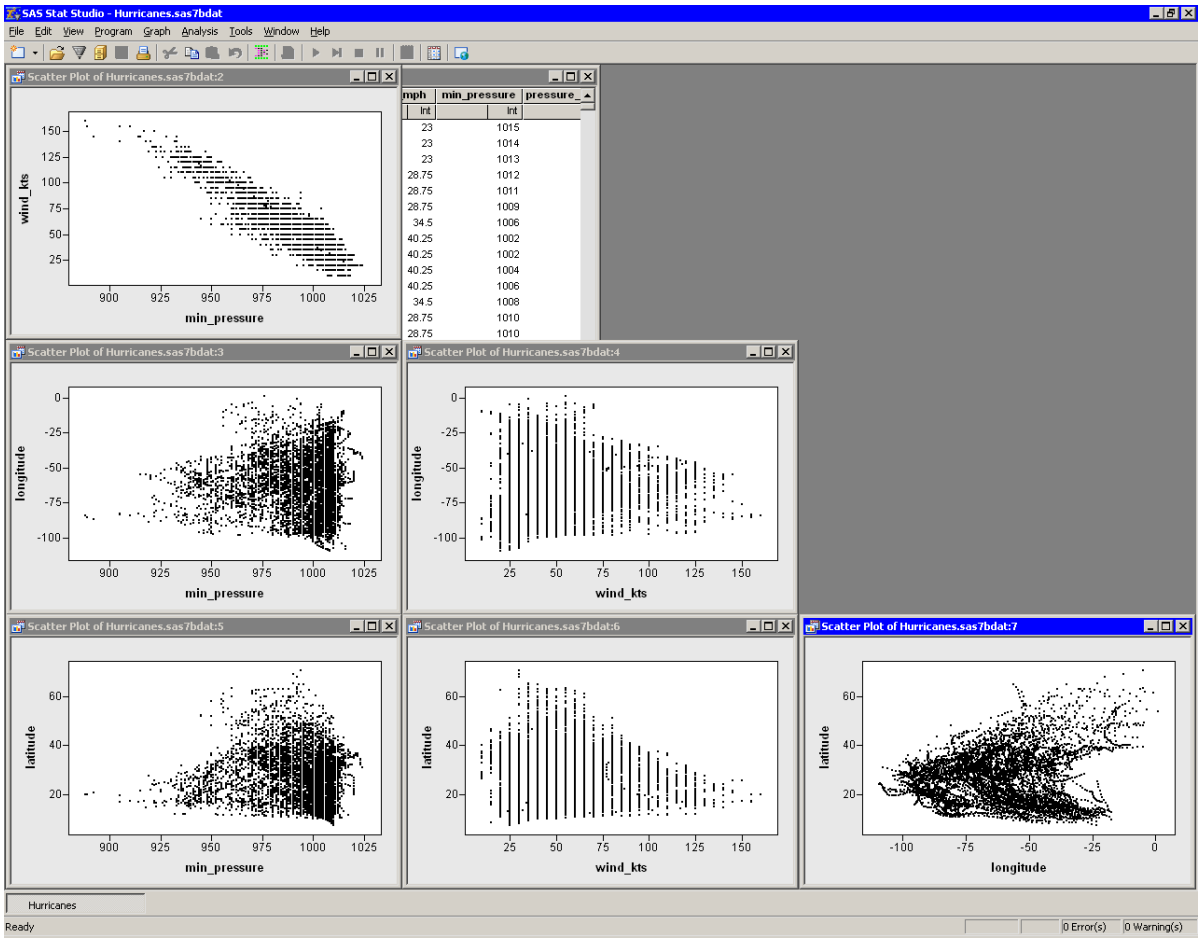
Hurricanes

	[4]	category	latitude	longitude	wind_kts	wind_mph	min_pressure	pressure_
6188		Nom	4 Int	3 Int	2 Int	Int	1 Int	
1	χ^2		32	-77.5	20	23	1015	
2	χ^2		32.8	-76.2	20	23	1014	
3	χ^2		34	-75.2	20	23	1013	
4	χ^2	TD	35.2	-74.6	25	28.75	1012	
5	χ^2	TD	37	-73.5	25	28.75	1011	
6	χ^2	TD	38.7	-72.4	25	28.75	1009	
7	χ^2	TD	40	-70.8	30	34.5	1006	
8	χ^2	TS	41.5	-69	35	40.25	1002	
9	χ^2	TS	43	-67.5	35	40.25	1002	
10	χ^2	TS	45	-65.5	35	40.25	1004	
11	χ^2	TS	47	-63	35	40.25	1006	
12	χ^2	TD	49	-60	30	34.5	1008	
13	χ^2	TD	51	-56	25	28.75	1010	
14	χ^2	TD	30.4	-90.3	25	28.75	1010	
15	χ^2	TD	29.7	-89.7	30	34.5	1009	
16	χ^2	TS	29.7	-89.4	35	40.25	1007	
17	χ^2	TS	29.4	-89.2	40	46	1005	
18	χ^2	TS	29.3	-89.1	45	51.75	1002	
19	χ^2	TS	29.6	-89.5	45	51.75	1001	
20	χ^2	TS	30.1	-90.4	45	51.75	1002	
21	χ^2	TS	30.1	-90.9	40	46	1005	

3 Select **Graph ►Scatter Plot** from the main menu.

A matrix of scatter plots appears, as shown in Figure 11.17.

Figure 11.17 A Matrix of Scatter Plots

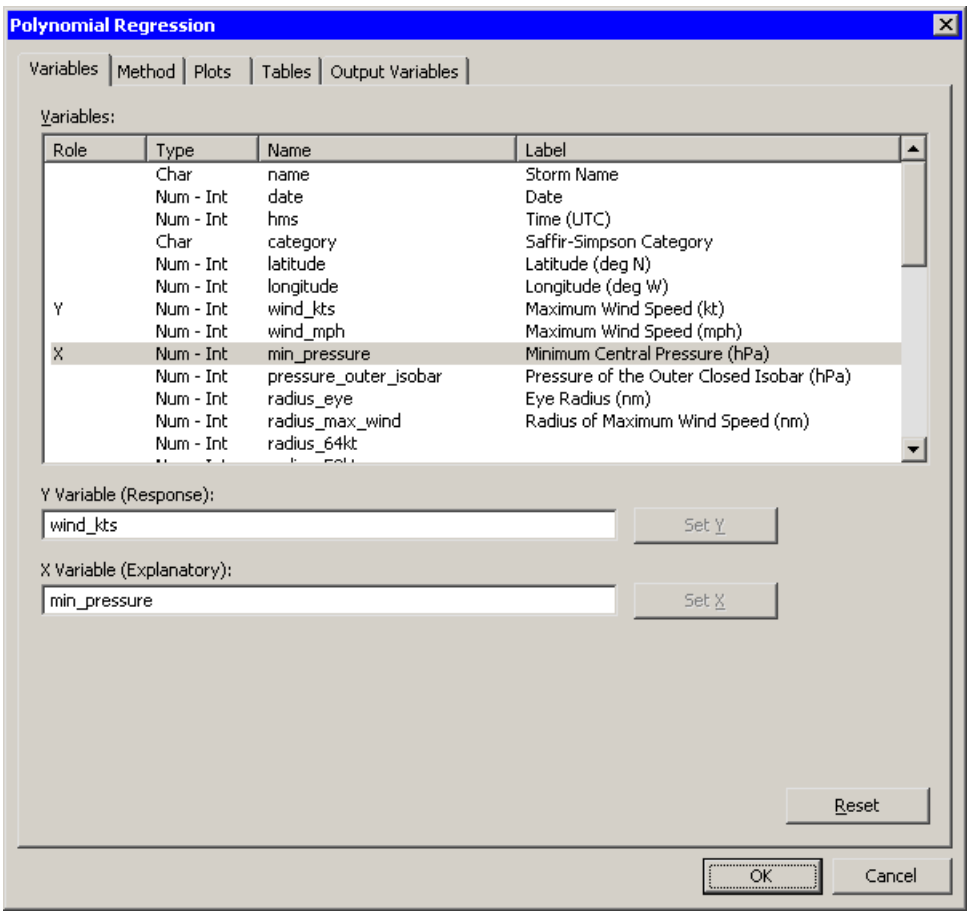


The scatter plot of wind_kts versus min_pressure show a strong negative correlation ($\rho = -0.93$) between wind speed and pressure. In the following steps, you model the linear relationship between these two variables and create plots of the fit residuals.

4 Select **Analysis ►Data Smoothing ►Polynomial Regression** from the main menu.

The dialog box shown in Figure 11.18 appears.

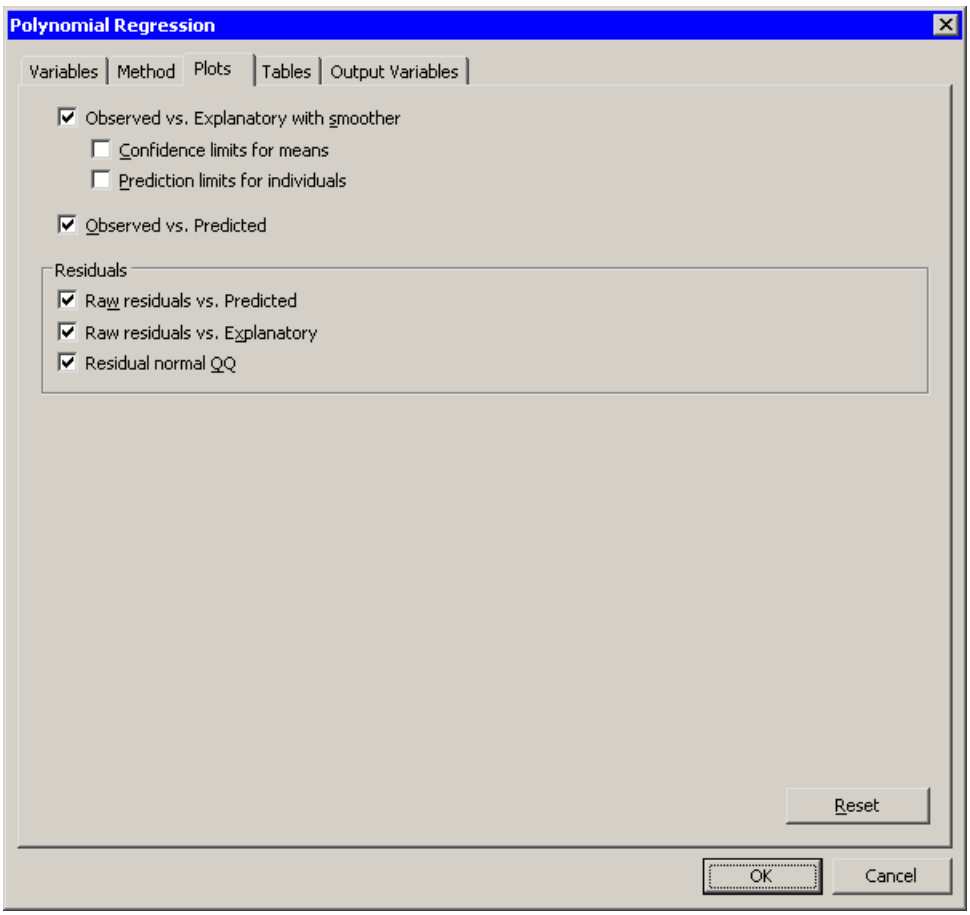
Figure 11.18 The Polynomial Regression Dialog Box



5 Select the variable wind_kts, and click **Set Y**. Select the variable min_pressure, and click **Set X**.

6 Click the Plots tab, as shown Figure 11.19.

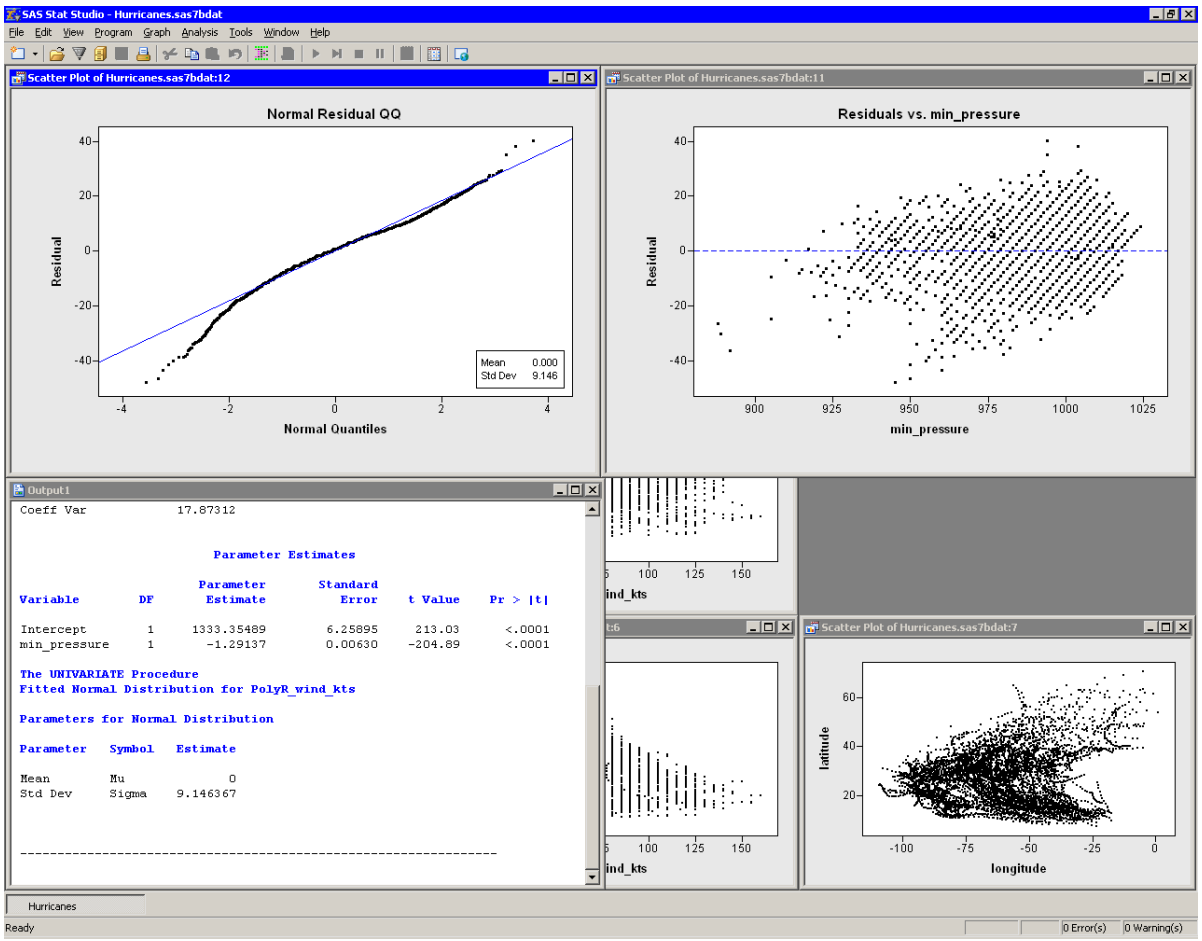
Figure 11.19 The Plots Tab



7 Select all plots. Clear the check boxes **Confidence limits for means** and **Prediction limits for individuals**. Click **OK**.

The analysis creates the five requested plots and an output window, as shown in Figure 11.20. Some of the plots produced by the analysis might be hidden beneath other plots.

Figure 11.20 Output and Plots from Polynomial Regression

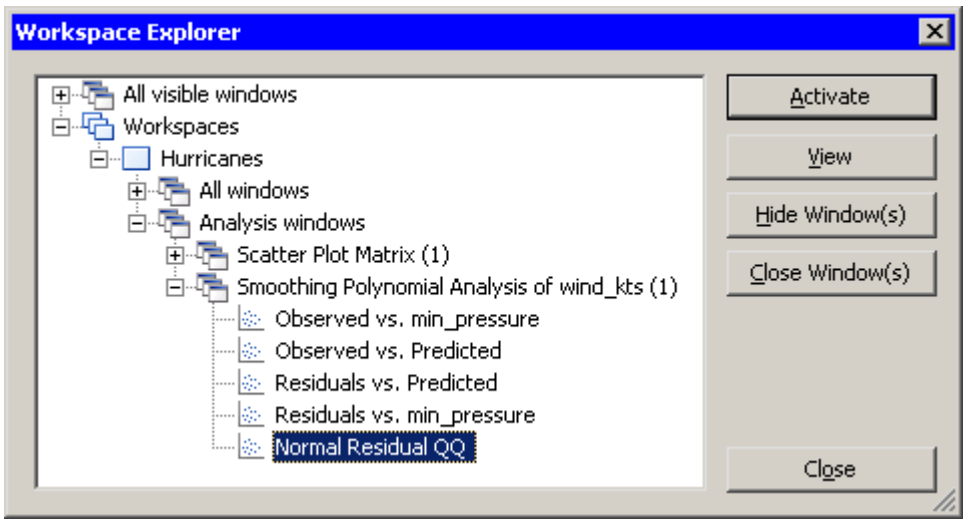


Your workspace now has a data table, a matrix of six scatter plots, five plots that are associated with an analysis, and an output window, for a total of 13 windows. The Workspace Explore enables you to manage these windows.

8 Press ALT+X to open the Workspace Explorer.

The Workspace Explorer is shown in Figure 11.21.

Figure 11.21 The Workspace Explorer



You can use the Workspace Explorer to do the following:

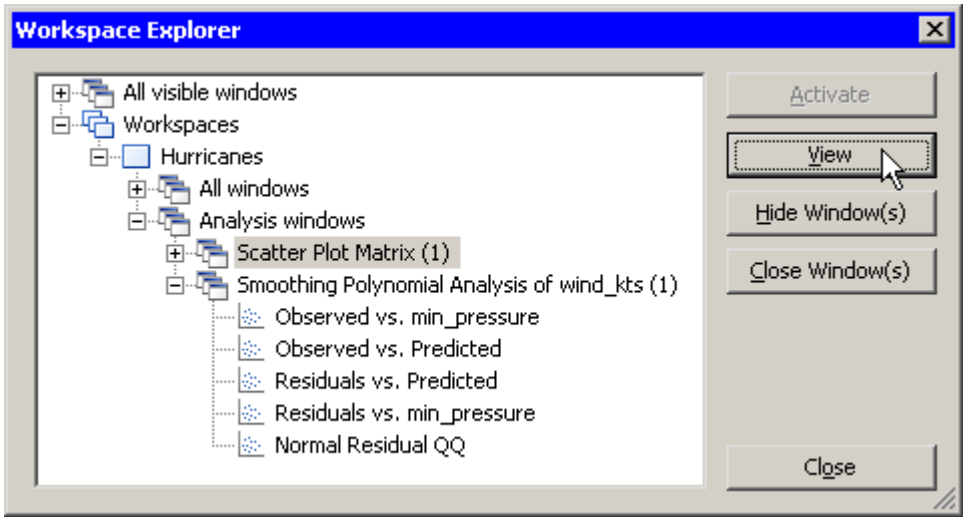
- bring a window or group of windows to the front of other windows
- hide a window or group of windows
- close a window or group of windows

For example, if you want to see all of the windows that are associated with the scatter plot matrix, you can do the following steps.

9 Click the node labeled **Scatter Plot Matrix**, and click **View**.

This step is shown in Figure 11.22. The matrix of scatter plots becomes visible.

Figure 11.22 Viewing a Group of Windows



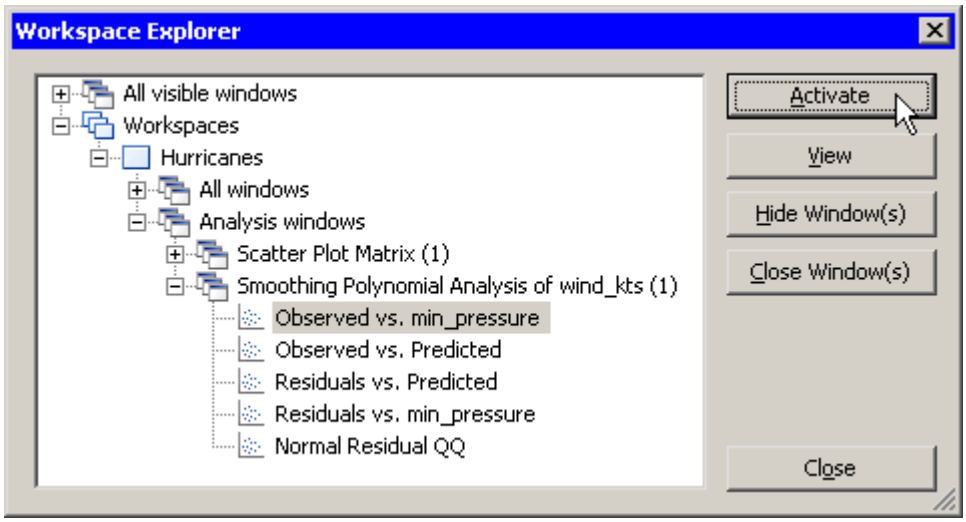
You also can view a particular plot. For example, the following steps activate the plot that contains the least squares line.

10 In the Workspace Explorer, expand the node labeled **Smoothing Polynomial Analysis of wind_kts**, if it is not already expanded.

11 Click the item labeled **Observed vs. min_pressure**.

This step is shown in Figure 11.23. The icon to the left of the plot name indicates that the plot is a scatter plot. The icons in the Workspace Explorer match the icons on the **Graph** main menu.

Figure 11.23 Activating a Window



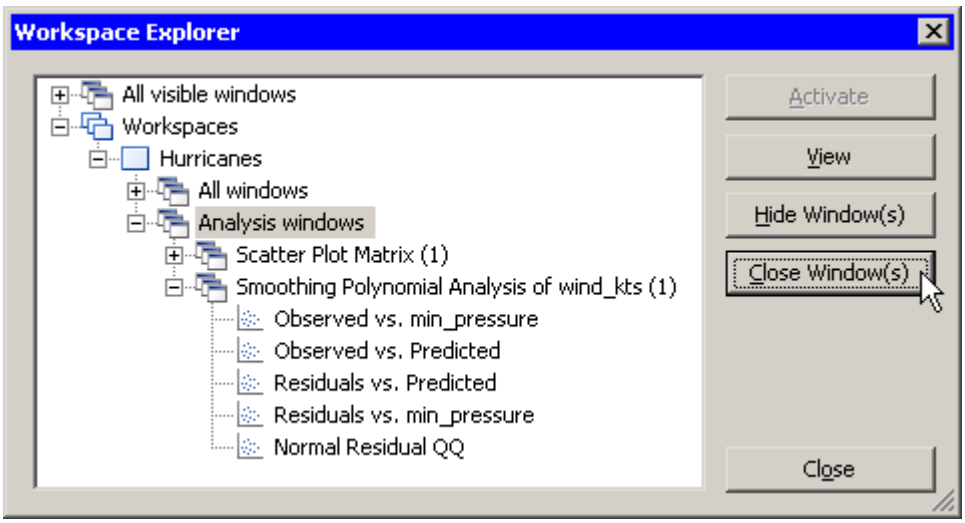
Note that the **Activate** button is now active, whereas it was previously inactive. This is because the selected item is an individual window instead of a group of windows. **Activate** behaves similarly to **View**, but it also closes the Workspace Explorer and makes the selected window the active window.

12 Click **Activate**.

13 When you are finished viewing a group of plots, the Workspace Explorer makes it easy to close them. Press ALT+X to open the Workspace Explorer. Click the node labeled **Analysis windows**. Click **Close Window(s)**.

This step is shown in Figure 11.24. SAS/IML Studio closes all of the plots created in this example. You can close workspaces in the same way.

Figure 11.24 Closing a Group of Windows



In summary, the Workspace Explorer enables you to view (or hide) windows. The following list describes each button in the Workspace Explorer.

Activate
makes the selected window visible and active. Selecting this button also closes the Workspace Explorer.

View
makes the selected window or group of windows visible.

Hide Window(s)
hides the selected window or group of windows.

Close Window(s)
closes the selected window or group of windows. You can also press the DELETE key to close the selected window or group of windows.

Close
closes the Workspace Explorer.

Copying Plots to the Windows Clipboard

It is easy to copy a plot to the Windows clipboard, and to paste the plot from the clipboard to the SAS/IML Studio output document window or to another application, such as Microsoft Windows or PowerPoint.

To copy a plot to the clipboard, activate the plot and select **Edit ►Copy** or press CTRL+C. You can paste to most applications by selecting **Edit ►Paste** or pressing CTRL+V.

SAS/IML Studio places the plot on the clipboard in one of the following three graphics formats:

Windows Enhanced Metafile Format (EMF)
stores the image as a series of 32-bit Windows drawing commands. This is the best format for exporting plots from SAS/IML Studio, because the file size is small and the image is faithful to the original. However, not all Windows applications support the EMF format. Specifically, the SAS/IML Studio output document window does not support the EMF format. Microsoft Word and PowerPoint do support the EMF format.

Windows Metafile Format (WMF)
stores the image as a series of 16-bit Windows drawing commands. This format is supported by virtually all Windows applications. However, the WMF format does not support *wide patterned lines*—lines that are not solid and have a width greater than one pixel. The WMF format represents a wide patterned line as a solid line of the same width.

Windows Device Independent Bitmap Format (BMP)
stores the image as a bitmap. This format is supported by virtually all Windows applications. Plots stored in the BMP format require much more memory than those stored in either the EMF or WMF format.

NOTE: When you paste a plot from the clipboard to a SAS/IML Studio output document window, SAS/IML Studio uses the BMP format to paste the plot. If the plot you are pasting does not make use of wide patterned lines, you can save memory by selecting **Edit ►Paste Special** which uses the WMF format to paste the plot.

Chapter 12
Plotting Subsets of Data

Contents

Overview of Plotting Subsets of Data **199**
Visualizing Data Features across Subsets of Data **200**
Example: Create a Scatter Plot for Each Value of a BY Variable **201**
Example: Set Marker Attributes for Each BY Group **205**
Part 1: Create an Indicator Variable 206
Part 2: Change Marker Properties 208
Part 3: Create Plots for Each BY Group 210
Techniques for Managing BY Group Plots **211**
BY Options Properties **214**

Overview of Plotting Subsets of Data

When your data contain categorical variables, you might be interested in comparing subsets of data that are defined by values of those variables. For example, if your data contains a gender variable, you might want to compare the characteristics of males with those of females.

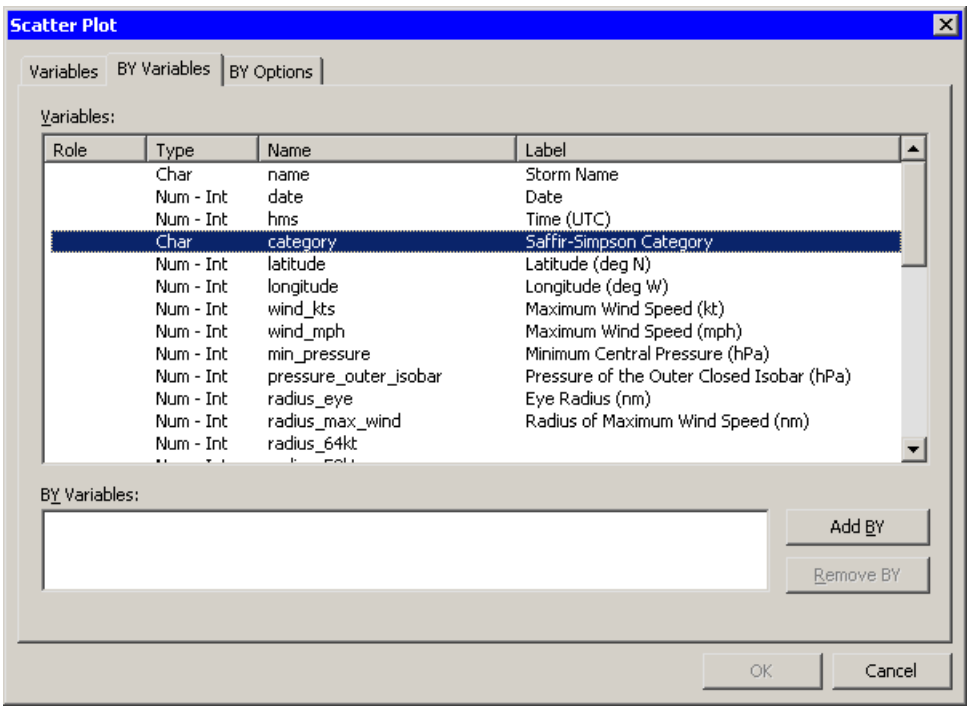
In SAS/IML Studio you can create plots of subsets of data that are defined by values of one or more categorical variables. The variables whose values define the subsets are called *BY variables* in SAS, and the subsets are known as *BY groups*. The BY groups are, by definition, mutually disjoint. Consequently, these plots are not dynamically linked to each other. In SAS/IML Studio, these plots are also not linked to the original data.

When you select any graph from the main **Graph** menu, a dialog box appears that has multiple tabs. (See Figure 12.1.) You can use the **Variables** tab to define the variable used by the plot. If you click **OK**, the plot is created on the full data and is linked to other plots and views of that data.

Alternatively, you can click the **BY Variables** tab and define one or more BY variables. (See Figure 12.1.) When you click **OK**, the data are split into BY groups, and a plot is created for each BY group. (The BY variables are usually nominal variables.)

You can specify options for the BY-group plots on the **BY Options** tab.

Figure 12.1 A Plot Dialog Box



Visualizing Data Features across Subsets of Data

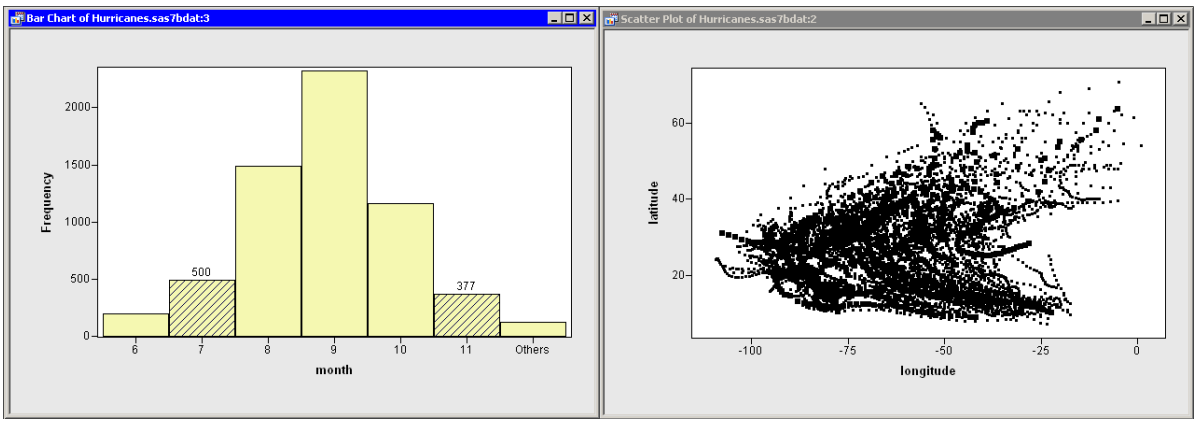
Suppose that you are interested in visualizing the location of tropical cyclones for each month (irrespective of the year). That is, you want to examine a scatter plot that shows the location of all April cyclones, another that shows the locations of May cyclones, and so on. There are at least two methods to accomplish this:

- One approach is to create a bar chart of months, select a bar (that is, a particular month) in the bar chart, and look at the selected observations in a scatter plot of wind_kts versus latitude. This technique is illustrated in Figure 12.2.

  This works well for many data sets. However, the selected observations might not be visible when the scatter plot suffers from overplotting (as in Figure 12.2), or when the number of selected observations is small relative to the total number of observations. A variation of this technique is to show only the selected observations. See the section "Displaying Only Selected Observations" on page 152 for a complete example that illustrates this approach.

  Overplotting can also make it difficult to compare features of the data across months. For example, in Figure 12.2, do early-summer cyclones originate in the same regions as autumn cyclones? Does the general shape of cyclone trajectories vary by month?

Figure 12.2 Selecting Cyclones in Certain Months



- A second visualization approach, known as BY-group processing, attempts to circumvent these problems by abandoning the concept of viewing all of the data in one plot. The idea behind BY-group processing is simple: instead of using a single scatter plot linked to a bar chart, you subset the data into mutually exclusive BY groups and make a scatter plot for each subset. This enables you to see each month's data in isolation, rather than superimposed on a single plot.

Example: Create a Scatter Plot for Each Value of a BY Variable

In this example, you create scatter plots of the latitude and longitude variables of the Hurricanes data set. The scatter plots are made for subsets of the hurricane data that correspond to the nine values of the month variable. (The data set does not contain any cyclones for January, February, or March.)

To create a scatter plot for each value of a BY variable:

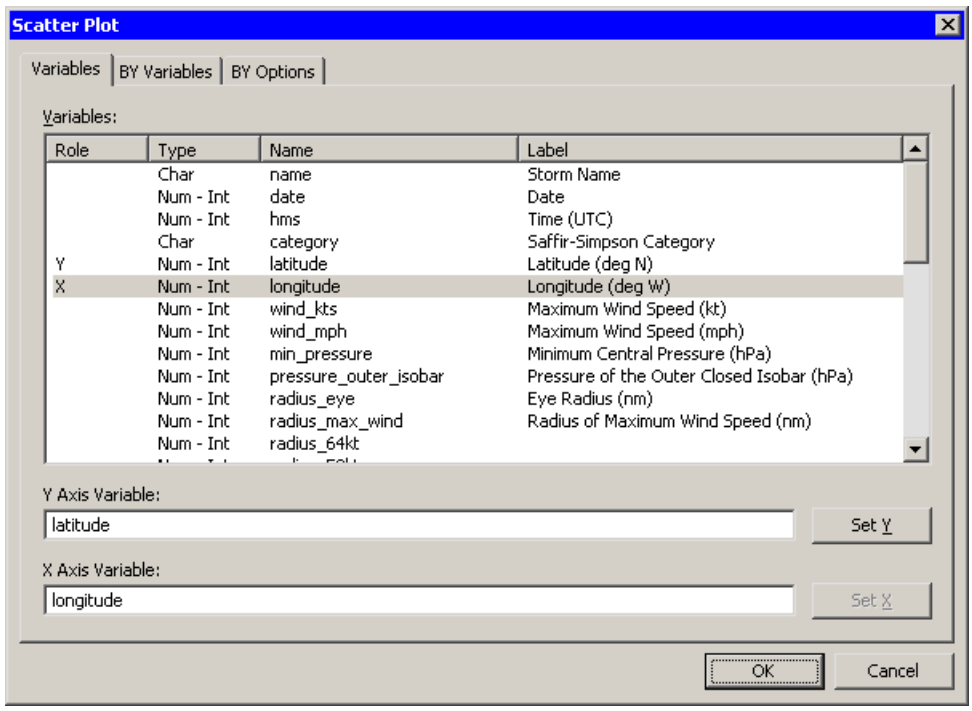
1 Open the Hurricanes data set.

2 Select **Graph ►Scatter Plot** from the main menu.

The Scatter Plot dialog box appears. (See Figure 12.3.)

3 Select the latitude variable, and click **Set Y**. Select the longitude variable, and click **Set X**.

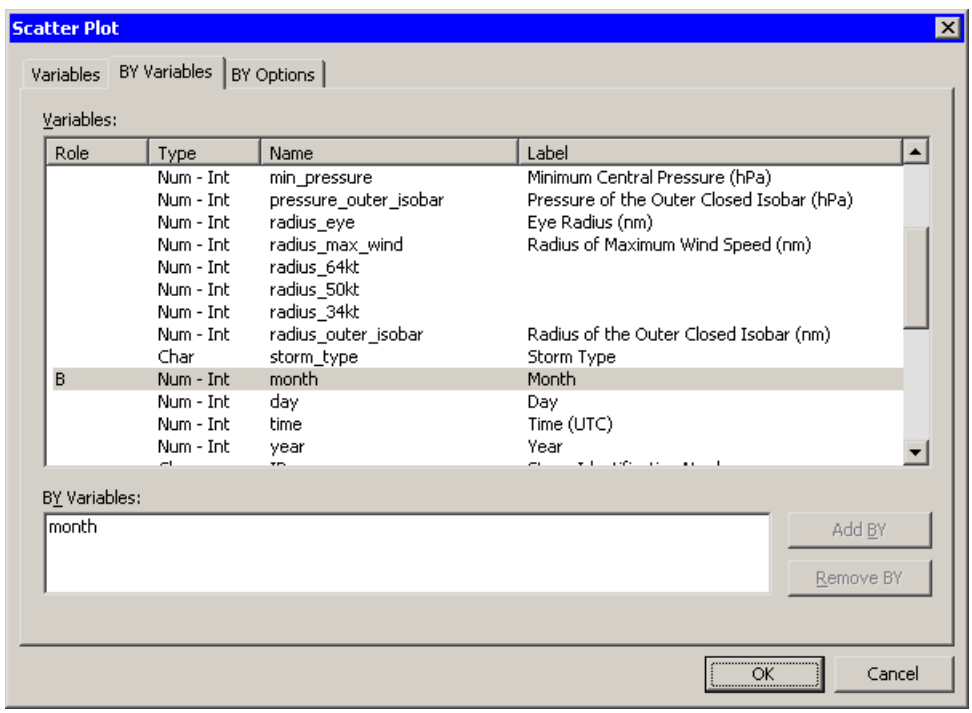
Figure 12.3 Selecting Scatter Plot Variables



4 Click the **BY Variables** tab.

The **BY Variables** tab is shown in Figure 12.4.

Figure 12.4 Selecting BY Variables

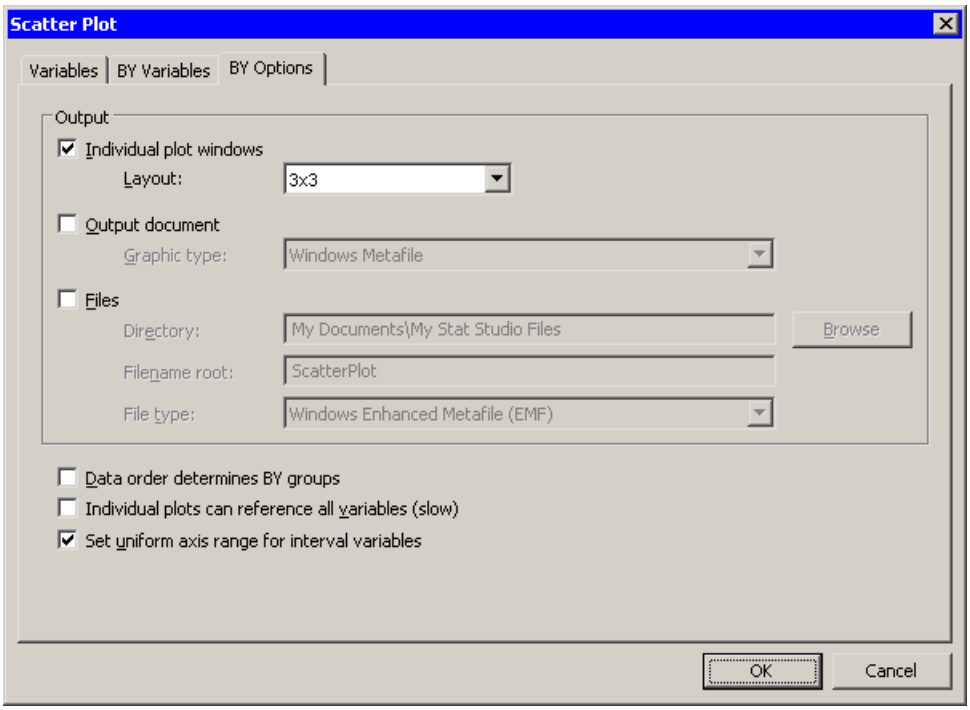


5 Scroll down in the list of variables and select the month variable. Click **Add BY**.

6 Click the **BY Options** tab.

The **BY Options** tab is shown in Figure 12.5.

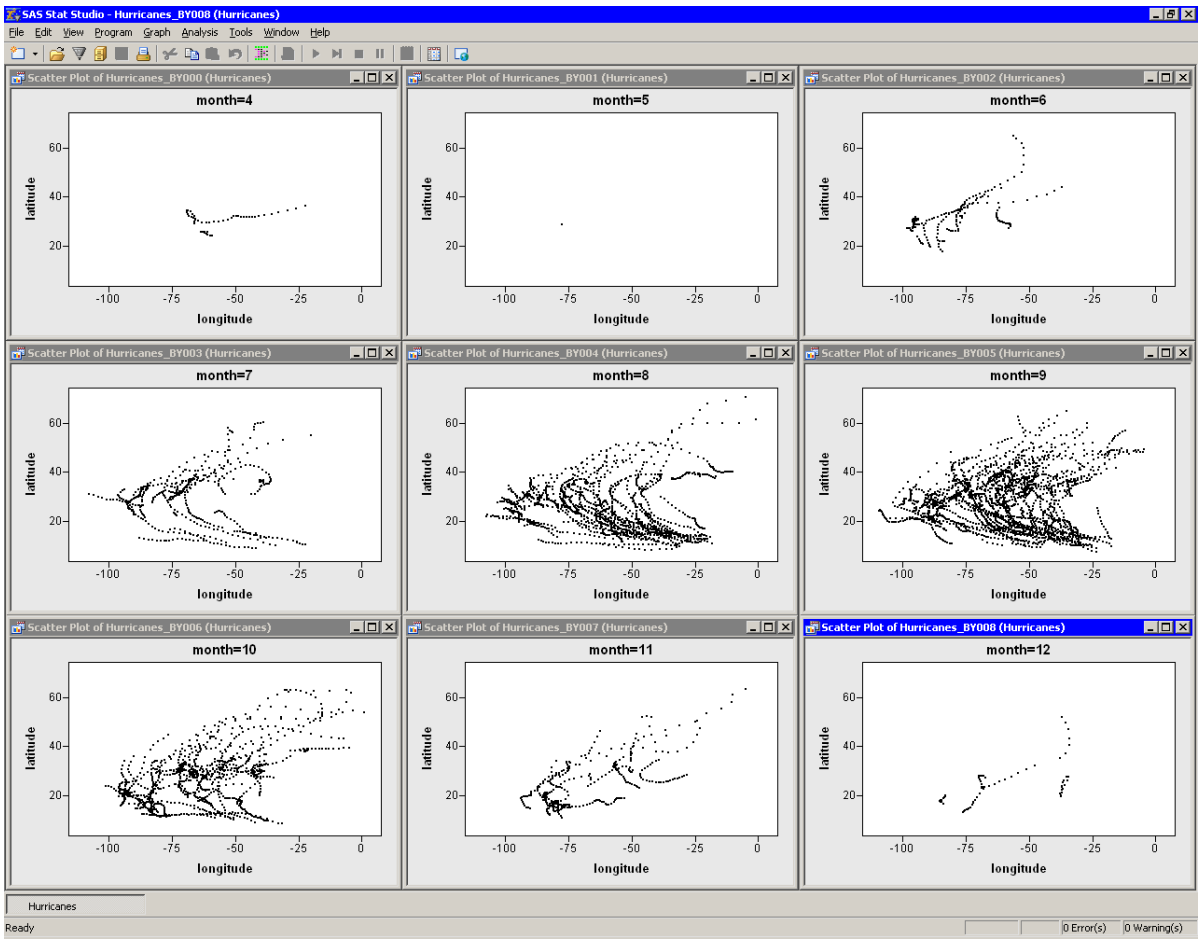
Figure 12.5 Subsetting Data and Plotting BY Groups



7 Select **3x3** for the **Layout** option. Click **OK**.

Nine scatter plots appear, one for each month 4–12, as shown in Figure 12.6.

Figure 12.6 Scatter Plots of Location by Month



Note that the X and Y axes are all set to a common range. This makes it easier to compare data characteristics across BY groups. If you want each plot to scale its axes independently, you can clear **Set uniform axis range for interval variables** on the **BY Options** tab.

A few features of the data are apparent.

- Many tropical cyclones occur in September (month=9).
- There is no apparent relationship between month and the shape of cyclone trajectories.

It is not clear from this display whether the origin of cyclones varies with the month. Perhaps storms in May (month=6) originate farther west than September storms (month=9), but more investigation is needed. The next example continues this investigation.

Example: Set Marker Attributes for Each BY Group

This example illustrates the fact that observation properties (such as the color and shape of markers) are copied to each BY group during the subsetting of the data. One way to visualize the location in which tropical cyclones originate is to mark the origin of each storm with a special symbol.

Figure 12.7 shows the first few observations of the Hurricanes data set. Observations 1–13 correspond to a time series for Tropical Storm Alberto. Observations 14–25 correspond to Beryl. Observations 26–63 correspond to Chris, and so on. The values of the latitude and longitude variables for observations 1, 14, 26, 64, . . . , are the origins of the cyclones. It would be useful to mark these observations so that they are noticeable in the BY-group plots.

Figure 12.7 Hurricane Data

Hurricanes.sas7bdat

	36	name	date	hms	category	latitude	longitude	wind_kts
6188		Nom	Int	Int	Nom	Int	Int	Int
1	χ^2	ALBERTO	05AUG1988	18:00		32	-77.5	20
2	χ^2	ALBERTO	06AUG1988	0:00		32.8	-76.2	20
3	χ^2	ALBERTO	06AUG1988	6:00		34	-75.2	20
4	χ^2	ALBERTO	06AUG1988	12:00	TD	35.2	-74.6	25
5	χ^2	ALBERTO	06AUG1988	18:00	TD	37	-73.5	25
6	χ^2	ALBERTO	07AUG1988	0:00	TD	38.7	-72.4	25
7	χ^2	ALBERTO	07AUG1988	6:00	TD	40	-70.8	30
8	χ^2	ALBERTO	07AUG1988	12:00	TS	41.5	-69	35
9	χ^2	ALBERTO	07AUG1988	18:00	TS	43	-67.5	35
10	χ^2	ALBERTO	08AUG1988	0:00	TS	45	-65.5	35
11	χ^2	ALBERTO	08AUG1988	6:00	TS	47	-63	35
12	χ^2	ALBERTO	08AUG1988	12:00	TD	49	-60	30
13	χ^2	ALBERTO	08AUG1988	18:00	TD	51	-56	25
14	χ^2	BERYL	08AUG1988	0:00	TD	30.4	-90.3	25
15	χ^2	BERYL	08AUG1988	6:00	TD	29.7	-89.7	30
16	χ^2	BERYL	08AUG1988	12:00	TS	29.7	-89.4	35
17	χ^2	BERYL	08AUG1988	18:00	TS	29.4	-89.2	40
18	χ^2	BERYL	09AUG1988	0:00	TS	29.3	-89.1	45
19	χ^2	BERYL	09AUG1988	6:00	TS	29.6	-89.5	45
20	χ^2	BERYL	09AUG1988	12:00	TS	30.1	-90.4	45
21	χ^2	BERYL	09AUG1988	18:00	TS	30.1	-90.9	40

This example has three parts. The first part creates an indicator variable that enumerates the observations for each cyclone. In particular, an observation for which the indicator variable is ‘1’ represents the origin of the storm. The second part of the example assigns a special marker property to the origins. The third part creates plots of BY group, as in the previous example.

Part 1: Create an Indicator Variable

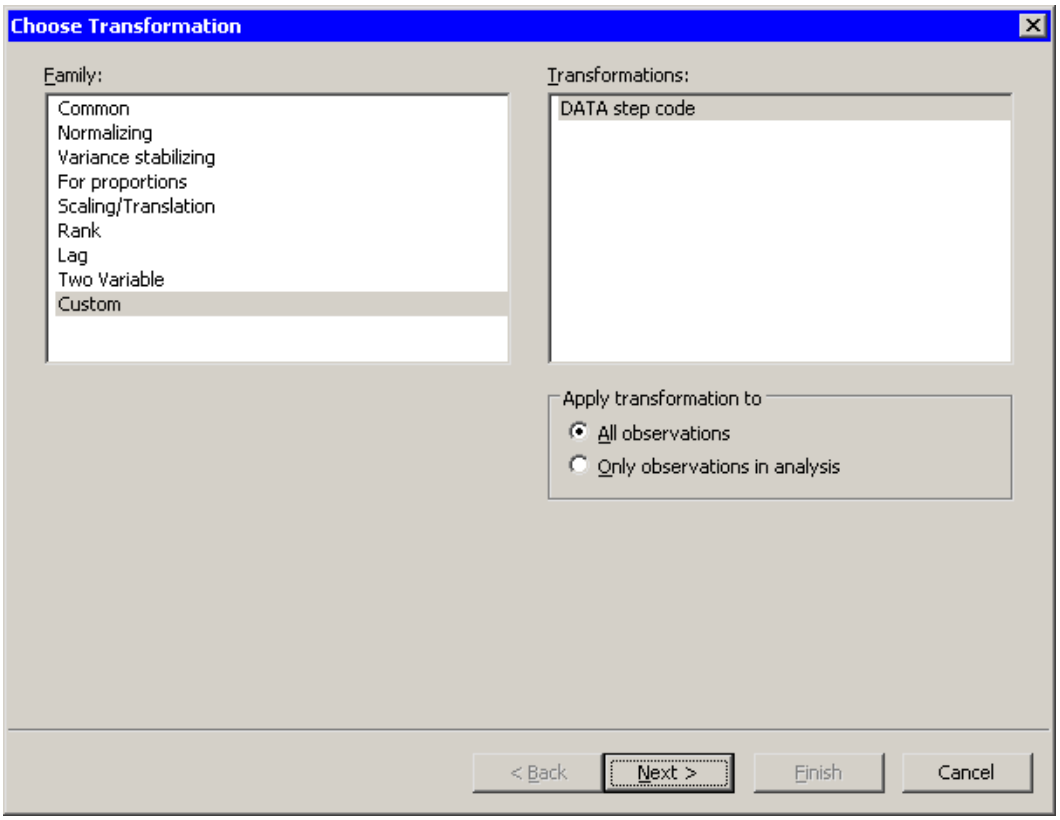
Using the DATA step is an easy way to create a variable that enumerates the observations for each cyclone. The following steps use the Variable Transformation Wizard to create the indicator variable. See "Custom Transformations" on page 518 for details about the Variable Transformation Wizard.

1 If you have not already done so, open the Hurricanes data set.

2 Select **Analysis ►Variable Transformation** from the main menu.

The Variable Transformation Wizard appears, as shown in Figure 12.8.

Figure 12.8 Selecting a Custom Transformation



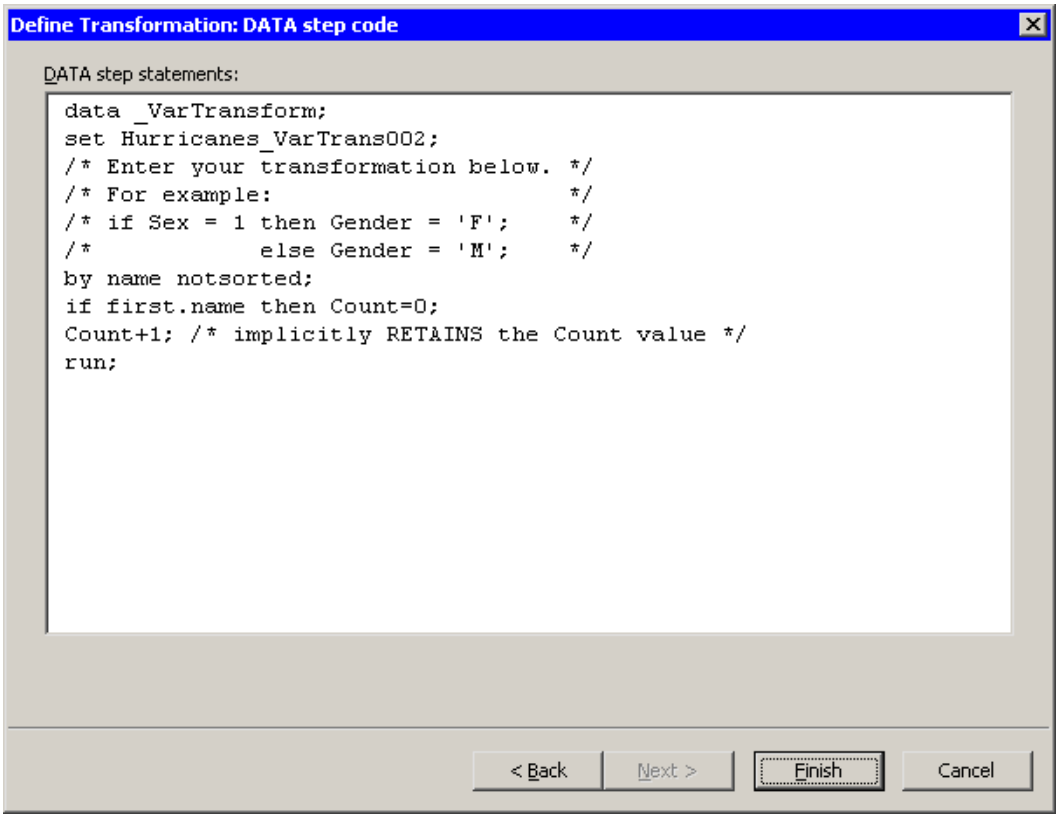
3 Select **Custom** from the **Family** list and click **Next**.

The second page of the wizard provides a window where you can enter DATA step statements.

4 Type in the following DATA step statements, prior to the RUN statement, as shown in Figure 12.9.

```
by name notsorted;
if first.name then Count=0;
Count+1; /* implicitly RETAINS the Count value */
```

Figure 12.9 Entering DATA Step Code



5 Click **Finish**.

A new variable, Count, is added to the data table. The variable enumerates the observations for each cyclone. In particular, Count=1 indicates the first observation for each cyclone. Figure 12.10 shows the new variable. (Some variables in the table are hidden.)

Figure 12.10 Hurricane Data With a New Variable

Hurricanes.sas7bdat

	36	name	date	hms	category	latitude	longitude	Count
6188		Nom	Int	Int	Nom	Int	Int	Int
1	■ χ^2	ALBERTO	05AUG1988	18:00		32	-77.5	1
2	■ χ^2	ALBERTO	06AUG1988	0:00		32.8	-76.2	2
3	■ χ^2	ALBERTO	06AUG1988	6:00		34	-75.2	3
4	■ χ^2	ALBERTO	06AUG1988	12:00	TD	35.2	-74.6	4
5	■ χ^2	ALBERTO	06AUG1988	18:00	TD	37	-73.5	5
6	■ χ^2	ALBERTO	07AUG1988	0:00	TD	38.7	-72.4	6
7	■ χ^2	ALBERTO	07AUG1988	6:00	TD	40	-70.8	7
8	■ χ^2	ALBERTO	07AUG1988	12:00	TS	41.5	-69	8
9	■ χ^2	ALBERTO	07AUG1988	18:00	TS	43	-67.5	9
10	■ χ^2	ALBERTO	08AUG1988	0:00	TS	45	-65.5	10
11	■ χ^2	ALBERTO	08AUG1988	6:00	TS	47	-63	11
12	■ χ^2	ALBERTO	08AUG1988	12:00	TD	49	-60	12
13	■ χ^2	ALBERTO	08AUG1988	18:00	TD	51	-56	13
14	■ χ^2	BERYL	08AUG1988	0:00	TD	30.4	-90.3	1
15	■ χ^2	BERYL	08AUG1988	6:00	TD	29.7	-89.7	2
16	■ χ^2	BERYL	08AUG1988	12:00	TS	29.7	-89.4	3
17	■ χ^2	BERYL	08AUG1988	18:00	TS	29.4	-89.2	4
18	■ χ^2	BERYL	09AUG1988	0:00	TS	29.3	-89.1	5
19	■ χ^2	BERYL	09AUG1988	6:00	TS	29.6	-89.5	6
20	■ χ^2	BERYL	09AUG1988	12:00	TS	30.1	-90.4	7
21	■ χ^2	BERYL	09AUG1988	18:00	TS	30.1	-90.9	8
22	■ χ^2	BERYL	10AUG1988	0:00	TD	30.3	-91.6	9
23	■ χ^2	BERYL	10AUG1988	6:00	TD	30.7	-92.2	10
24	■ χ^2	BERYL	10AUG1988	12:00	TD	31.2	-92.6	11
25	■ χ^2	BERYL	10AUG1988	18:00		31.7	-93.2	12
26	■ χ^2	CHRIS	21AUG1988	12:00	TD	14.9	-43.3	1

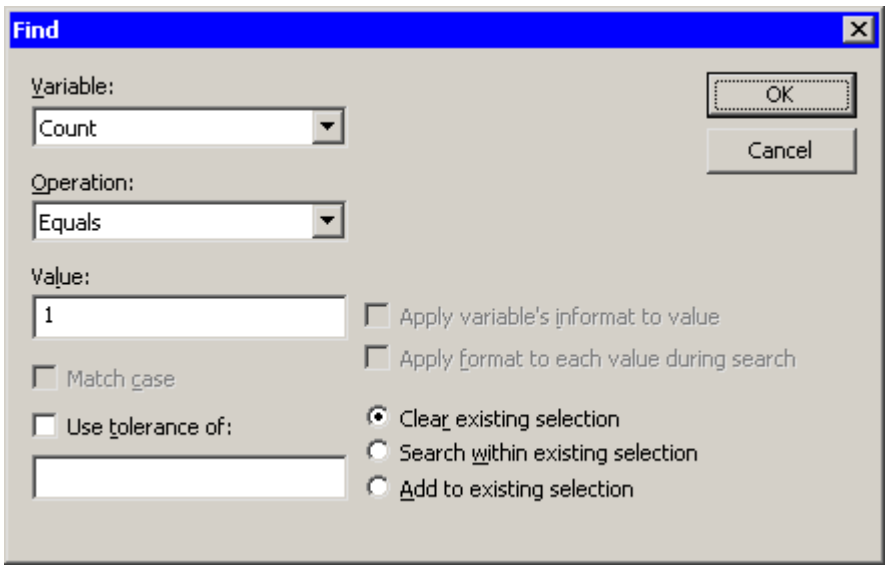
Part 2: Change Marker Properties

To select observations for which Count=1 and change the shape and color of those observations:

1 Select **Edit ►Find** from the main menu.

2 Fill out the dialog box to find observations for which Count equals 1, as shown in Figure 12.11. Click **OK**.

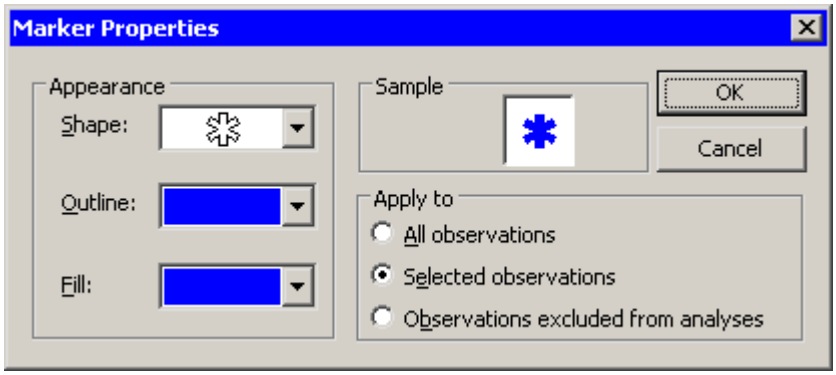
Figure 12.11 The Find Dialog Box



3 Select **Edit ►Observations ►Marker Properties** from the main menu.

The Marker Properties dialog box appears, as shown in Figure 12.12.

Figure 12.12 The Marker Properties Dialog Box



4 Change **Shape** to a star (∗). Change the **Outline** and **Fill** to blue. Click **OK**.

The observations with Count=1 are now selected and represented by blue star-shaped markers.

Part 3: Create Plots for Each BY Group

To create a scatter plot for each value of a BY variable:

1 Select **Graph ►Scatter Plot** from the main menu.

2 Select the latitude variable, and click **Set Y**. Select the longitude variable, and click **Set X**.

3 Click the **BY Variables** tab.

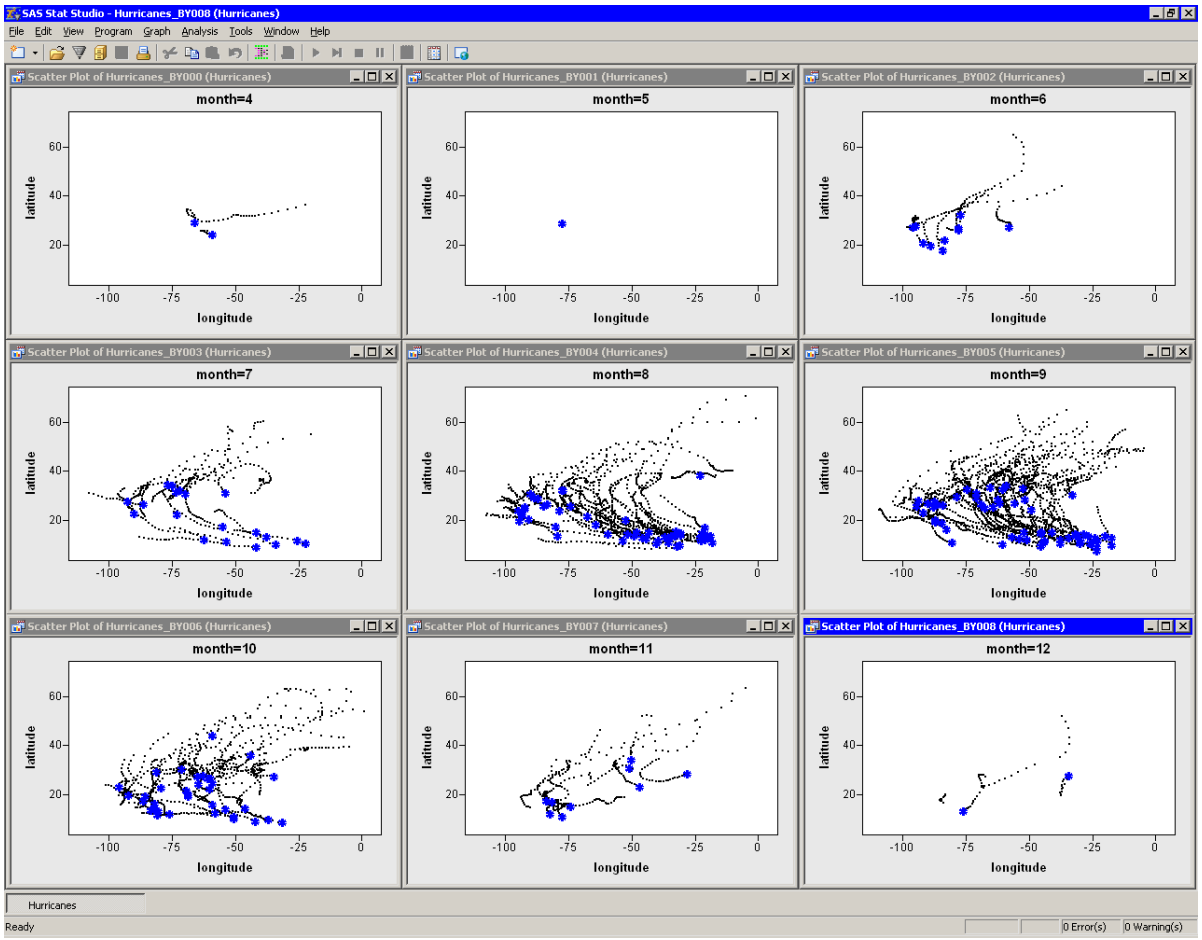
4 Scroll down in the list of variables and select the month variable. Click **Add BY**.

The **BY Options** tab should be populated with your choices from the previous example.

5 Click **OK**.

Nine scatter plots appear, one for each month 4–12, as shown in Figure 12.13.

Figure 12.13 Scatter Plots of Location by Month



Note that marker properties such as color, shape, and selected status are copied to each of the BY groups. In particular, the selected blue stars enable you to see the origin of each cyclone.

A few new features of the data are apparent.

- The origin of cyclones varies with the month.
- Cyclones early in the season (May–June) and late in the season (October–November) often originate in the Gulf of Mexico (81–98 degrees west longitude and 18–30 degrees north latitude) or in the Caribbean Sea.
- In August and September quite a few Cape Verde cyclones are apparent. Cape Verde cyclones originate between the Cape Verde islands (23 degrees west longitude and 15 degrees north latitude) and the Lesser Antilles (60 degrees west longitude).
- A large number of cyclones originate in the mid-Atlantic (25–35 degrees north latitude) in September, although mid-Atlantic origins are also seen in other months.

The next section describes how you can use the Workspace Explorer to view, hide, close, and compare BY-group plots.

Techniques for Managing BY Group Plots

You can use BY-group plots more effectively if you understand a few details about the way BY-group plots are implemented in SAS/IML Studio.

When you create BY-group plots, the following steps occur:

1. A new variable, _ObsNum_, is added to the current data table.
2. The observations that correspond to each BY groups are identified.
3. The observations in each BY group are copied to a new DataObject. (See *SAS/IML Studio for SAS/STAT Users* for details about the DataObject class.) The variables that are copied depend on the **Individual plots can reference all variables** option on the **BY Options** tab, shown in Figure 12.5.
4. The plots are created.

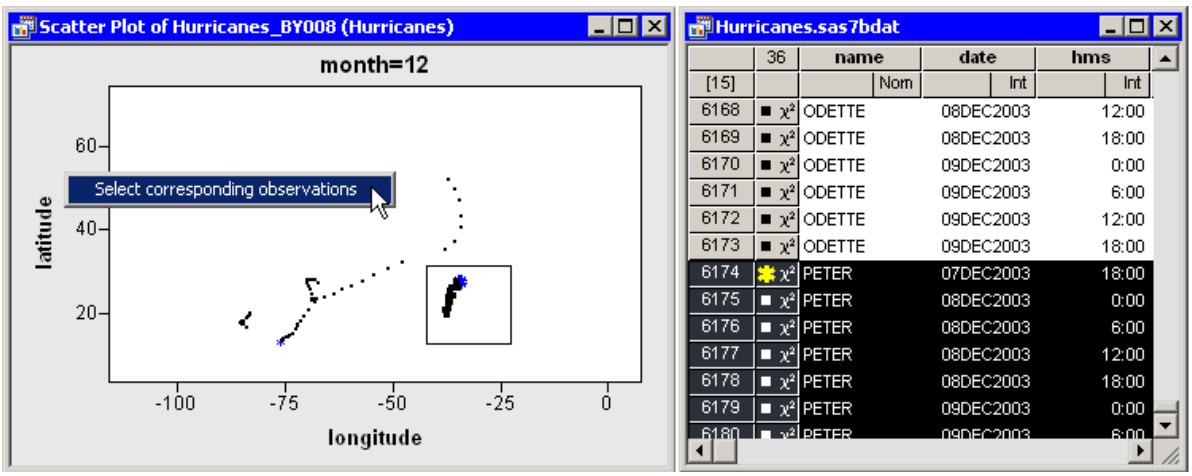
If all observations in a BY group are excluded from plots, the BY group is not copied and no plot is created.

The BY-group plots are *not* dynamically linked to the original data. Consequently, selections made to the original data are not reflected in the BY groups. However, you can use an *action menu* to select observations in the original data that correspond to selected observations in a BY-group plot. See the online Help for a description of action menus.

Figure 12.14 illustrates the action menu. Press the F11 key to display the action menu in a BY-group plot. When you select the action menu item, SAS/IML Studio looks at the values of the _ObsNum_ variable for the selected observations. SAS/IML Studio then selects observations in the original data that contain the same values of _ObsNum_, as shown in the right-hand portion of Figure 12.14.

Using the action menu to select observations is a cumulative process: if an observation in the original data was selected prior to this action, it remains selected after the action. You can clear selections in the data table the usual way: press the ESC key or click in the upper left cell of the data table.

Figure 12.14 Selecting Observations from a BY Group Plot

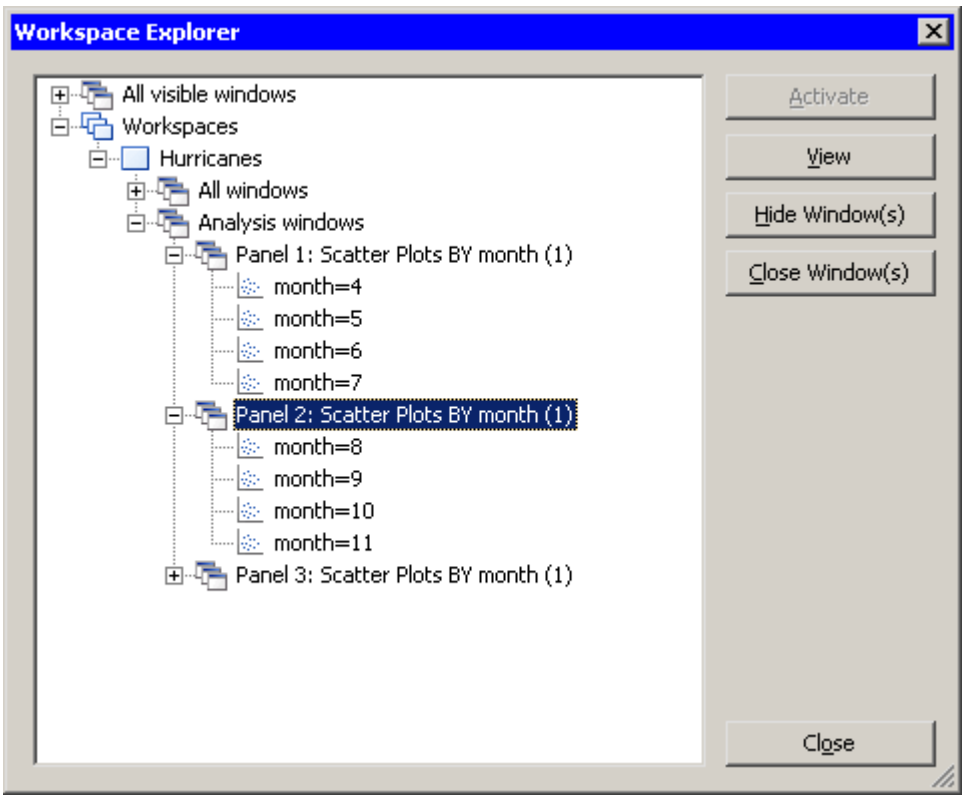


The **Layout** field shown in Figure 12.5 determines how many BY-group plots are displayed on the screen. If you create more BY-group plots than can fit on the screen, then the remaining plots are created as hidden windows.

You can use the Workspace Explorer to manage BY-group plots. The Workspace Explorer is described in "Workspace Explorer" on page 190.

For example, if you recreate the previous example, but select **2x2** for the **Layout** field, then only the first four plots are displayed. You can select **Windows ►Workspace Explorer** from the main menu to display the Workspace Explorer, as shown in Figure 12.15. You can select "Panel 2" and click **View** to see the next four plots. You can also hide an entire panel by clicking **Hide Window(s)**. Finally, you can compare plots that belong to different panels by selecting each individual plot and clicking **View**.

Figure 12.15 Managing BY Group Plots with Workspace Explorer



NOTE: The number of plots that you can display on the screen at one time is limited by Windows resources. The number of plots you can create depends on characteristics of your PC; a typical PC can create a few hundred. SAS/IML Studio prevents you from creating more than 128 BY-group plots on the screen. If you need to create more plots than this limit, use the options on the **BY Options** tab to write the plots to the output document or to send the plots to files.

BY Options Properties

This section describes the **BY Options** tab that is associated with plots.

The **BY Options** tab controls how data are divided into subsets and how the plots are displayed. The **BY Options** tab (shown in Figure 12.16) contains the following UI controls:

Individual plot windows
specifies whether to display plots on the screen.

Layout
specifies how plots are arranged on the screen.

Output document
specifies whether to copy plots to the output document.

Graphic type
specifies the image type for plots copied to the output document.

Files
specifies whether to write plots to files on the client (or a networked drive).

Directory
specifies the directory for writing plots to files.

Filename root
specifies the prefix used for writing plots to files. The plots are named *Root*001, *Root*002, and so on. The suffix of each file corresponds to an enumeration of the BY groups. Existing files with the same name are overwritten.

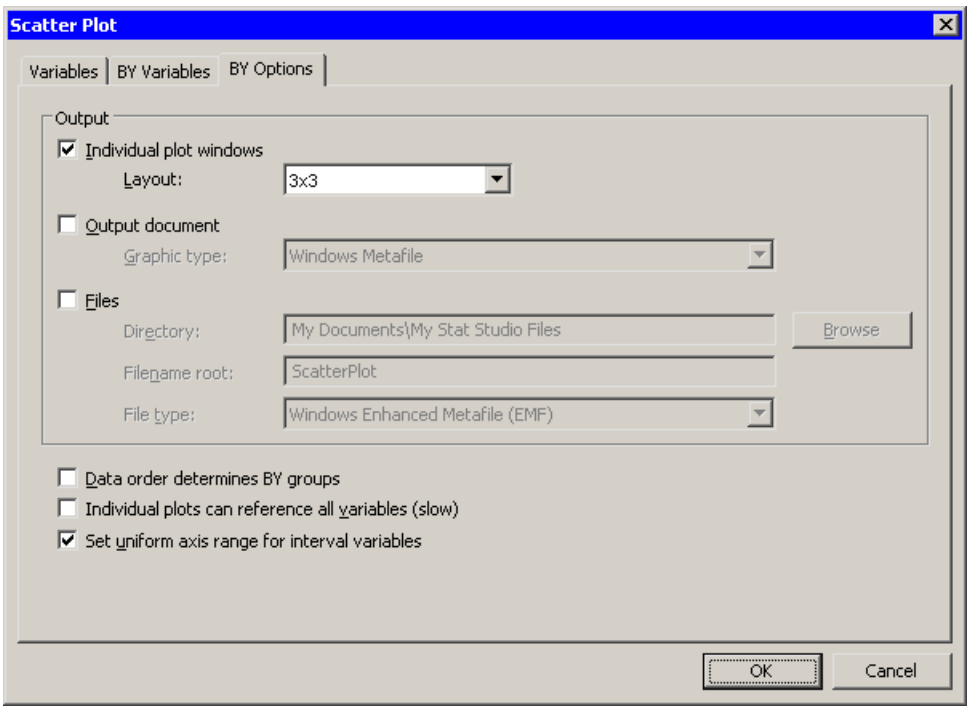
File type
specifies the image type for plots written to files.

Data order determines BY groups
This option corresponds to the NOTSORTED option in the BY statement in SAS procedures. If this option is selected, then no sorting is done prior to forming the BY groups. If this option is not selected, then the BY variables are internally sorted and the BY groups consist of observations that correspond to the unique values of the BY variables.

Individual plots can reference all variables (slow) If this option is selected, then all variables are copied when forming BY groups. If this option is not selected, then the BY groups contain only the variables that are specified on the **Variables** and **BY Variables** tabs. This option is available only when **Individual plot windows** is selected.

Set uniform axis range for interval variables If this option is selected, then the axes of interval variables are set to a common range. If this option is not selected, each axis is scaled individually according to the data in each BY group. This option is ignored for a rotating plot and for nominal axes. This option does not affect the frequency axis for histograms or bar charts.

Figure 12.16 BY Group Options



Chapter 13
Distribution Analysis: Descriptive Statistics

Contents

Overview of the Descriptive Statistics Analysis **217**
Example: Compute Descriptive Statistics **217**
Specifying the Descriptive Statistics Analysis **221**
Variables Tab 221
Plots Tab 221
Tables Tab 222
Roles Tab 223
Analysis of Selected Variables **223**
References **223**

Overview of the Descriptive Statistics Analysis

You can use the Descriptive Statistics analysis to compute descriptive statistics for a numeric variable. You can compute basic statistics such as the mean, median, variance, and interquartile range for the selected variable. You can also compute quantiles and extreme values. Finally, you can produce a histogram and box plot that are dynamically linked to the data.

You can run a Descriptive Statistics analysis by selecting **Analysis ►Distribution Analysis ►Descriptive Statistics** from the main menu. When you request descriptive statistics, SAS/IML Studio calls the UNIVARIATE procedure in Base SAS software. See the UNIVARIATE procedure documentation in the *Base SAS Procedures Guide* for additional details.

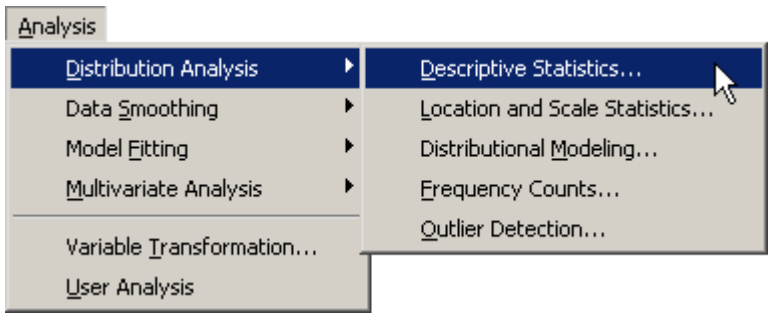
Example: Compute Descriptive Statistics

In this example, you generate descriptive statistics for the pressure_outer_isobar variable of the Hurricanes data set. The Hurricanes data set contains 6,188 observations of tropical cyclones in the Atlantic basin. The pressure_outer_isobar variable gives the sea-level atmospheric pressure for the outermost closed isobar of a cyclone. This is a measure of the atmospheric pressure at the outermost edge of the storm.

To create descriptive statistics:

1. Open the Hurricanes data set.
2. Select **Analysis ►Distribution Analysis ►Descriptive Statistics** from the main menu, as shown in Figure 13.1.

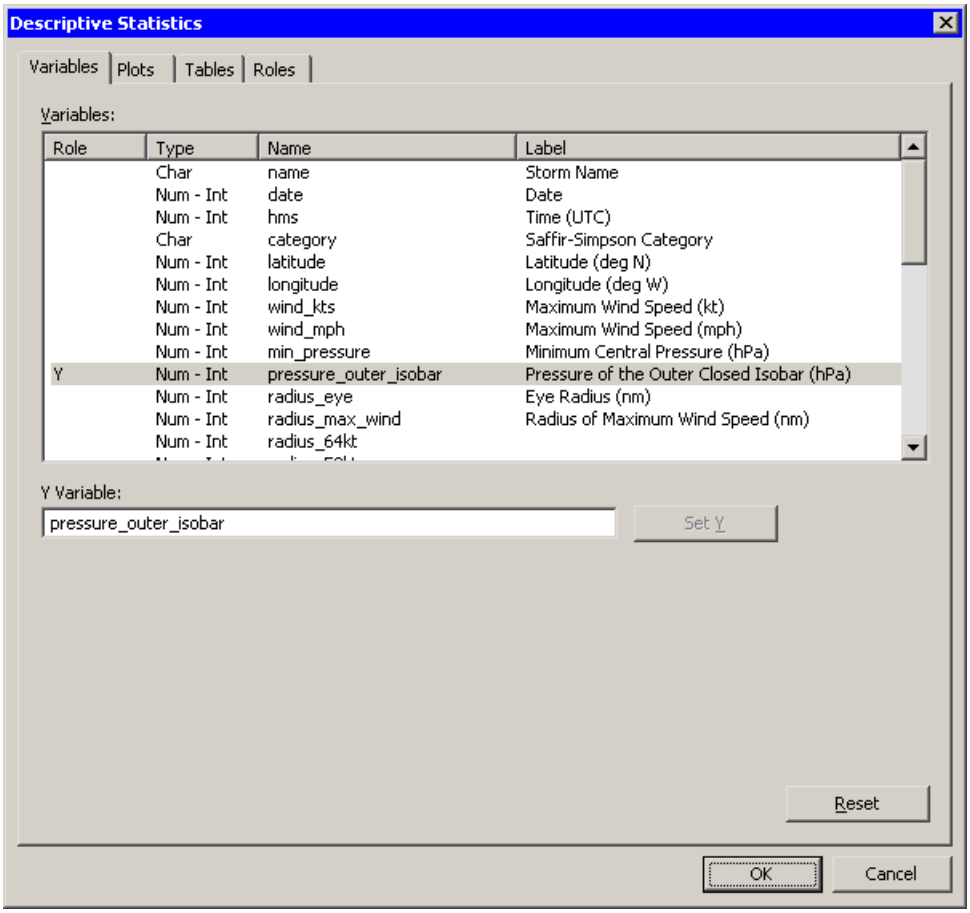
Figure 13.1 Selecting the Descriptive Statistics Analysis



The Descriptive Statistics dialog box appears. (See Figure 13.2.) You can select a variable for the univariate analysis by using the **Variables** tab.

3 Select the variable pressure_outer_isobar, and click **Set Y**.

Figure 13.2 Selecting a Variable



4 Click the **Tables** tab.

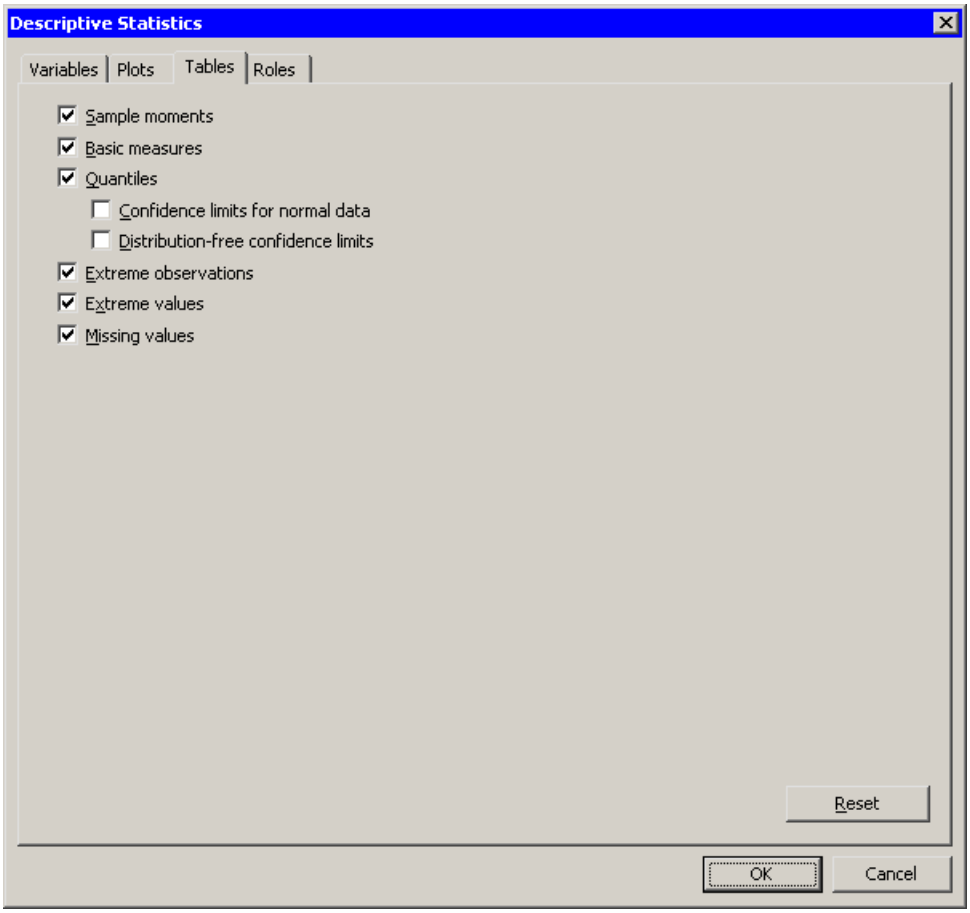
The **Tables** tab becomes active. (See Figure 13.3.)

5 Select **Extreme Values**.

6 Select **Missing Values**.

7 Click **OK**.

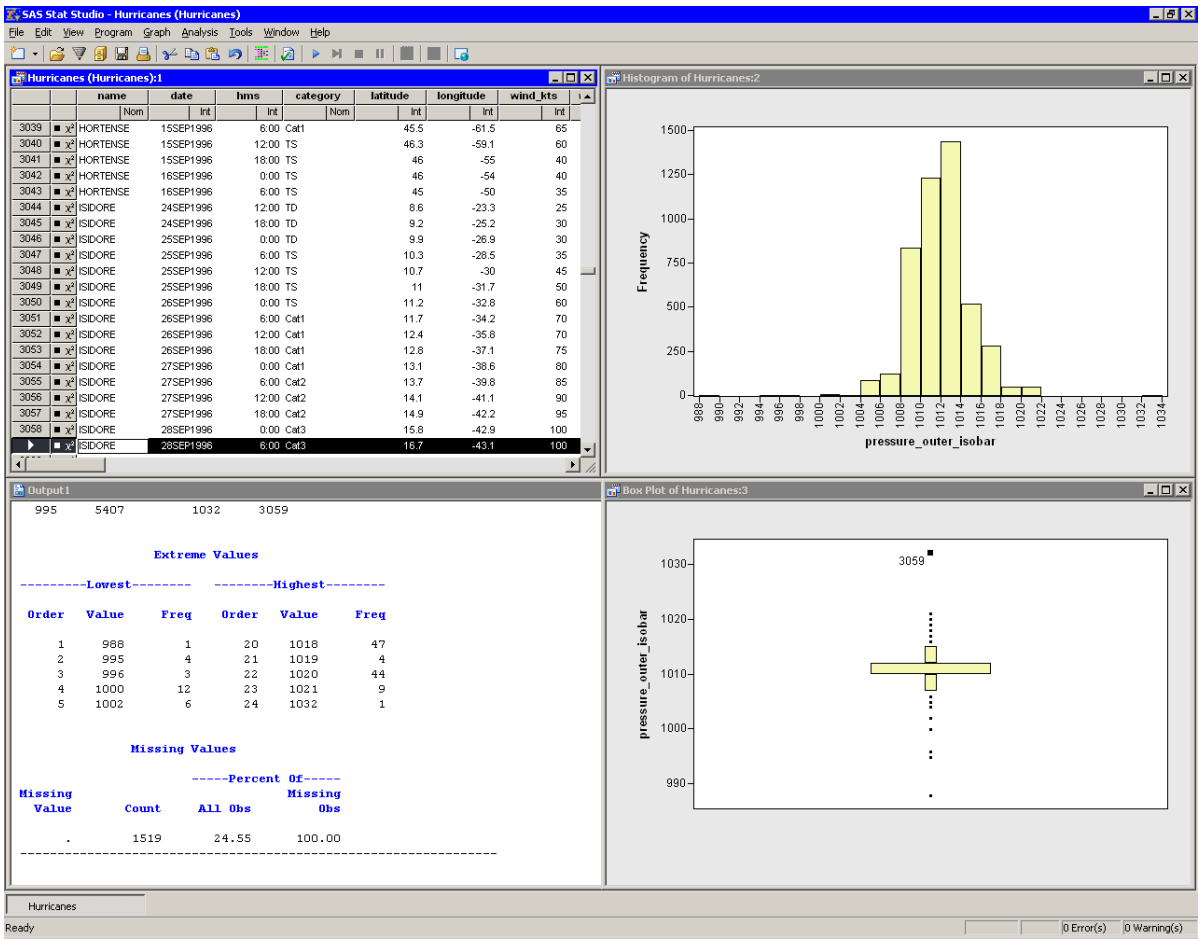
Figure 13.3 Selecting Tables



The analysis calls the UNIVARIATE procedure, which uses the options specified in the dialog box. The procedure displays tables in the output document, as shown in Figure 13.4. In addition to displaying basic statistics such as the mean, median, and standard deviation, the tables also display a few extreme values that seem incongruous. The "Extreme Values" table shows that there is one low value (998) and one high value (1032) that require investigation. The "Missing Values" table reveals that almost 25% of the values for this variable are missing.

Two plots are created. One plot shows a histogram of the selected variable; the other shows a box plot. One plot might be hidden beneath the other.

Figure 13.4 Output from a Descriptive Statistics Analysis



For the pressure_outer_isobar variable, the box plot and the "Extreme Values" table reveal many outliers. It is often useful to investigate outliers to determine whether they are spurious or miscoded data, or to better understand the extreme limits of the data.

8 In the box plot, click the outlier with the highest value of pressure_outer_isobar.

This selects the observation in all views of the data, including the data table. You can use the F3 key to scroll through the data table to the next selected observations.

9 Activate the data table by clicking the title bar. Use the F3 key to scroll the selected observation into view.

The selected observation corresponds to Hurricane Isadore, September 28, 1996. Scrolling through the data table reveals that the observations before and after the selected observation had a value of 1012 hPa for pressure_outer_isobar. This might indicate that the outlier value of 1032 hPa is a misrecorded value.

You can examine other outliers similarly.

10 In the box plot, click the outlier with the lowest value of pressure_outer_isobar.

11 Activate the data table by clicking its title bar. Use the F3 key to scroll the selected observation into view.

This selected observation corresponds to a pressure of 988 hPa for the outermost closed isobar of Hurricane Hugo, September 23, 1989. The data table shows that the observations before the selected observation had considerably larger values of pressure_outer_isobar. Furthermore, the value of min_pressure for the selected observation is 990 hPa, which is larger than the value being investigated. This violates the fact that for a low pressure system, the minimum central pressure should be less than the pressure of the outermost closed isobar. Therefore, the 988 hPa value is most likely misrecorded.

You can exclude misrecorded observations by using the **Exclude from Plots** and **Exclude from Analysis** features of the data table. For details, see Chapter 4, "Interacting with the Data Table." Excluding an observation affects *all* variables. You can also exclude a single misrecorded value by doing the following: replace the erroneous value with a missing value by typing "." (or " " for a character variable) into the data table cell. Save the data if you want to make the change permanent.

Specifying the Descriptive Statistics Analysis

This section describes the dialog box options that are associated with the Descriptive Statistics analysis. The Descriptive Statistics analysis calls the UNIVARIATE procedure in Base SAS software.

Variables Tab

You can use the **Variables** tab to specify the variable for the analysis. Only a single variable can be analyzed at a time. The **Variables** tab is shown in Figure 13.2.

Plots Tab

You can use the **Plots** tab to create a histogram and a box plot of the chosen variable. (See Figure 13.5.)

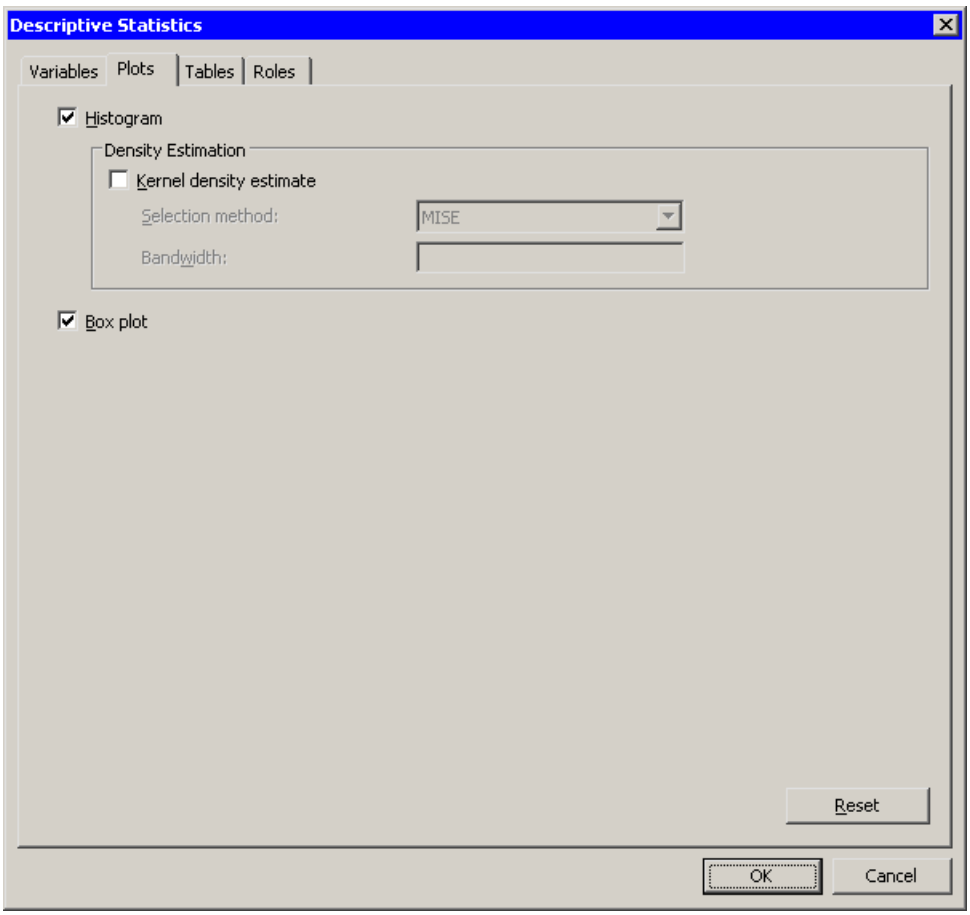
The histogram can include a kernel density estimate. You can determine the bandwidth for the kernel density method by selecting an option from the **Selection method** list. The options are as follows:

MISE
specifies that the kernel bandwidth is chosen to minimize an approximate mean integrated square error.

Sheather-Jones
specifies that the kernel bandwidth is chosen by a plug-in formula of Sheather and Jones (Jones, Marron, and Sheather 1996).

Manual
sets the kernel bandwidth to the value of the **Bandwidth** field.

Figure 13.5 Selecting Plots



NOTE: SAS/IML Studio adds a kernel density estimate to an *existing* histogram when both of the following conditions are satisfied:

- The histogram is the active window when you select the analysis.
- The histogram variable and the analysis variable are the same.

Tables Tab

You can use the **Tables** tab to display tables that summarize the results of the univariate analysis. The **Tables** tab is shown in Figure 13.3. You can choose from the following tables:

Sample moments
displays sample moments and related statistics, including the mean, variance, skewness, and kurtosis.

Basic measures
displays statistics that are related to the central location and the spread of the data.

Quantiles
displays quantile information.

Confidence limits for normal data
adds confidence limits to the "Quantiles" table, based on the assumption that the data are normally distributed.

Distribution-free confidence limits
adds confidence limits to the "Quantiles" table, based on order statistics.

Extreme observations
displays the observations with the highest and lowest values for the selected variable.

Extreme values
displays the extreme values (highest and lowest) for the selected variable.

Missing values
displays the frequency and percentage of missing values for the selected variable.

NOTE: The observation numbers in the "Extreme Observations" table reflect the observations that are included in the analysis. If you exclude observations from the analysis, the observation numbers reported in the "Extreme Observations" table might not correspond to the same observations in the data table.

Roles Tab

You can use the **Roles** tab to specify a frequency variable for the analysis. A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, each observation represents n observations, where n is the value of the frequency variable.

Analysis of Selected Variables

If an interval variable is selected in a data table when you run the analysis, then that variable is automatically entered in the **Y Variable** field of the **Variables** tab.

If any variable in the data table has a Frequency role, it is automatically entered in the **Frequency Variable** field of the **Roles** tab.

References

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996), "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, 91, 401–407.

Chapter 14
Distribution Analysis: Location and Scale Statistics

Contents

Overview of the Location and Scale Statistics Analysis **225**
Example: Compute Location and Scale Statistics **225**
Specifying the Location and Scale Statistics Analysis **230**
Variables Tab 230
Tables Tab 230
Roles Tab 231
Analysis of Selected Variables **231**

Overview of the Location and Scale Statistics Analysis

Univariate data are often summarized by computing statistics that estimate location and scale. The mean, median, mode, trimmed mean, and Winsorized mean are all statistics that describe the *location* (or central tendency) of data. Statistics that describe the *scale* (or variability) include the standard deviation, interquartile range, Gini's mean difference, and median absolute deviation from the median (MAD). You can use the Location and Scale Statistics analysis to compute location and scale estimates for a single numeric variable. You can also test the hypothesis that the population mean equals a particular value.

You can run a Location and Scale Statistics analysis by selecting **Analysis ►Distribution Analysis ►Location and Scale Statistics** from the main menu. When you request location and scale estimates, SAS/IML Studio calls the UNIVARIATE procedure in Base SAS software. See the UNIVARIATE procedure documentation in the *Base SAS Procedures Guide* for additional details.

Example: Compute Location and Scale Statistics

In this example, you compute statistics that estimate the location and scale for the pressure_outer_isobar variable of the Hurricanes data set. The Hurricanes data set contains 6,188 observations of tropical cyclones in the Atlantic basin. The pressure_outer_isobar variable gives the sea-level atmospheric pressure for the outermost closed isobar of a cyclone. This is a measure of the atmospheric pressure at the outermost edge of the storm. The pressure_outer_isobar variable contains 4,669 nonmissing values.

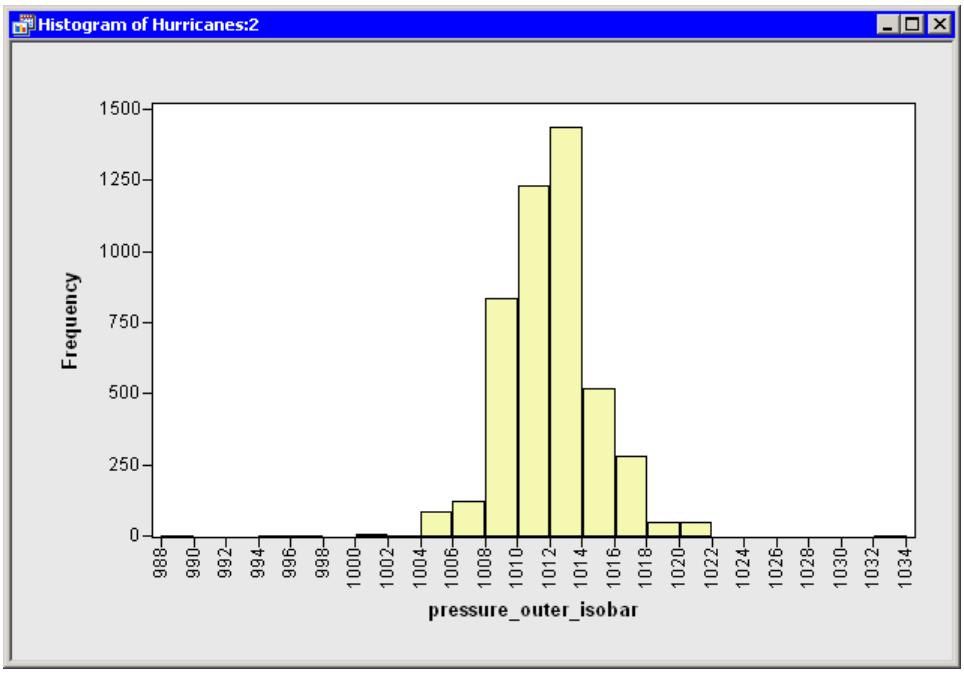
To compute estimates for the location and scale parameters:

1 Open the Hurricanes data set.

2 Create a histogram of the pressure_outer_isobar variable.

A histogram appears, as shown in Figure 14.1.

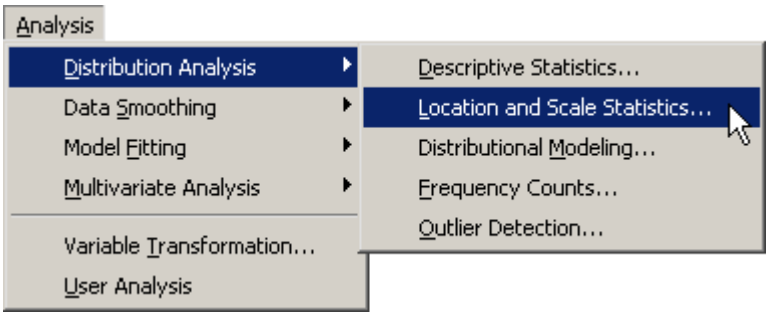
Figure 14.1 A Histogram



The histogram indicates that there are outliers in these data. Consequently, you might decide to compute robust estimates of location and scale for this variable, in addition to traditional estimates.

3 Select **Analysis ►Distribution Analysis ►Location and Scale Statistics** from the main menu, as shown in Figure 14.2.

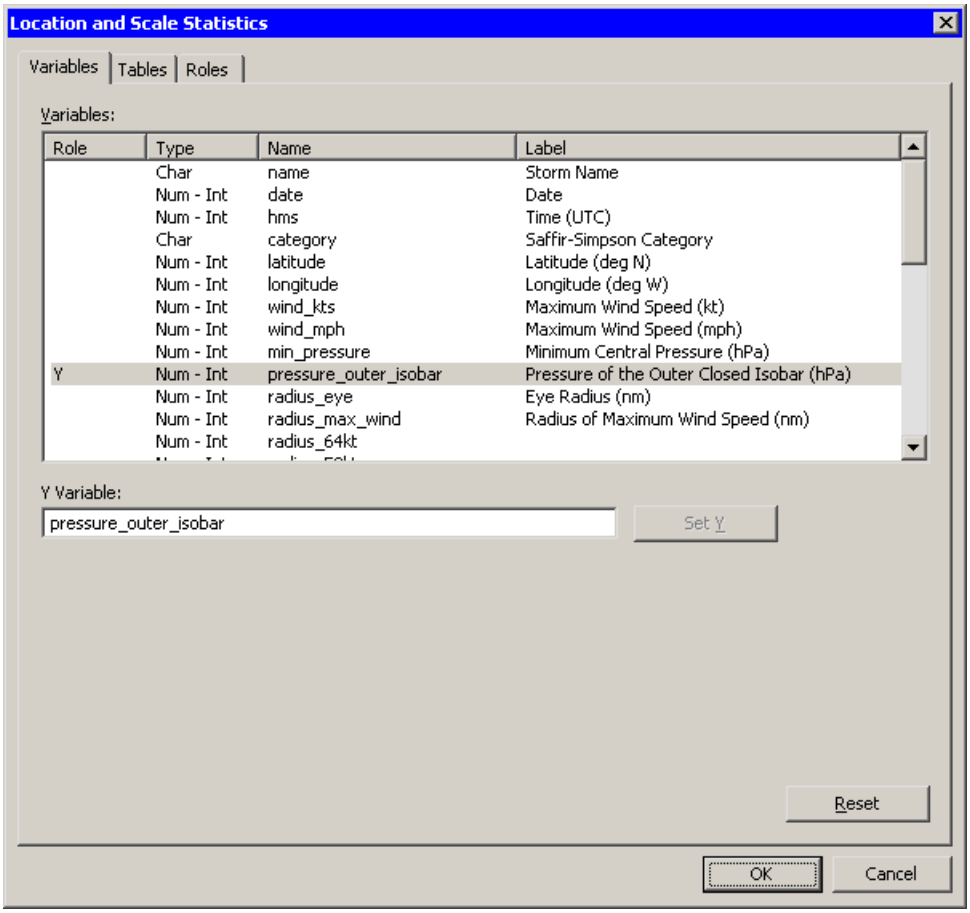
Figure 14.2 Selecting the Location and Scale Statistics Analysis



The Location and Scale Statistics dialog box appears. (See Figure 14.3.) You can select a variable for the univariate analysis on the **Variables** tab.

4 Select the variable pressure_outer_isobar, and click **Set Y**.

Figure 14.3 Selecting a Variable



5 Click the **Tables** tab.

The **Tables** tab becomes active. (See Figure 14.4.)

6 Select **Modes**.

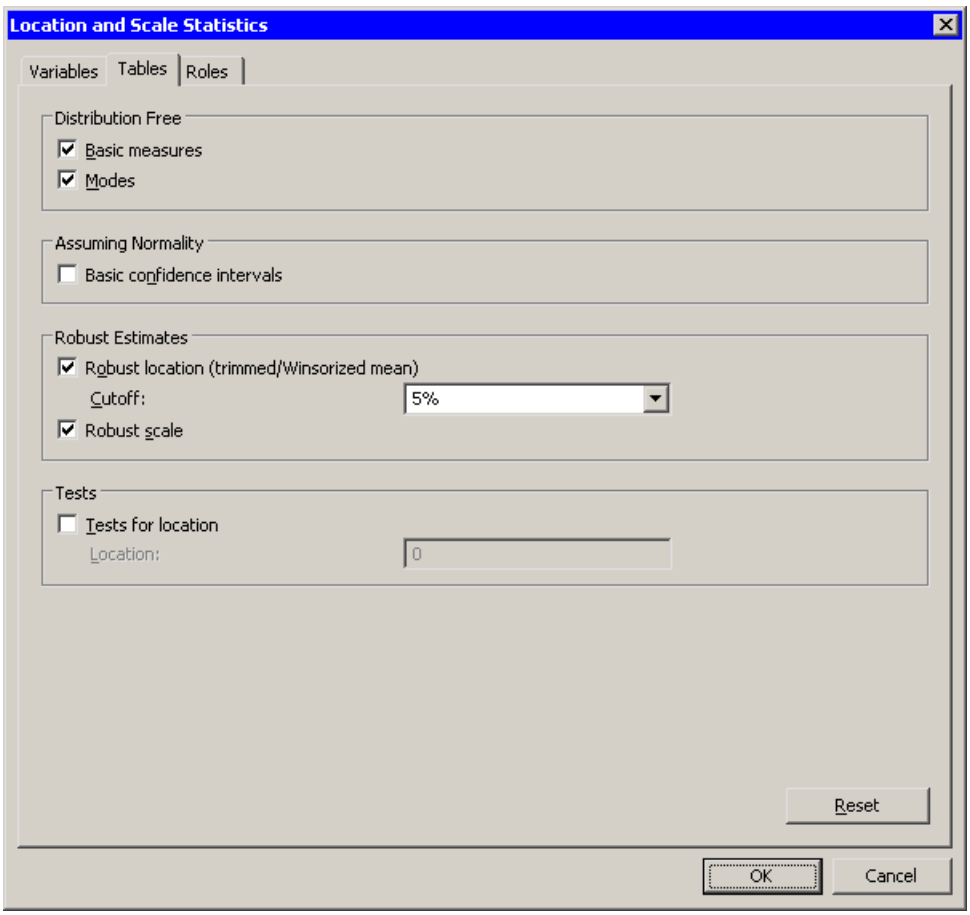
The following steps compute robust estimates for the location and scale of these data:

7 Select **Robust location (trimmed/Winsorized mean)**.

8 Select **Robust scale**.

9 Click **OK**.

Figure 14.4 Selecting Tables



The analysis calls the UNIVARIATE procedure, which uses the options specified in the dialog box. The procedure displays tables in the output document, as shown in Figure 14.5.

Figure 14.5 Output from a Location and Scale Statistics Analysis

Output1

```
The UNIVARIATE Procedure
Variable:  pressure_outer_isobar  (Pressure of the Outer Closed Isobar (hPa))

                Basic Statistical Measures

     Location                     Variability

Mean     1011.173     Std Deviation            2.97572
Median   1012.000     Variance                 8.85493
Mode     1012.000     Range                   44.00000
                      Interquartile Range      2.00000


      Modes

 Mode     Count

 1012      1212


                                        Trimmed Means

Percent    Number                Std Error
Trimmed   Trimmed     Trimmed     Trimmed     95% Confidence                t for H0:
in Tail   in Tail        Mean        Mean          Limits              DF    Mu0=0.00    Pr > |t|

   5.01       234    1011.181    0.040541   1011.102  1011.261     4200    24942.19      <.0001


                                       Winsorized Means

   Percent      Number                Std Error
Winsorized  Winsorized  Winsorized  Winsorized    95% Confidence             t for H0:
   in Tail     in Tail        Mean        Mean         Limits            DF   Mu0=0.00  Pr > |t|

      5.01         234    1011.213    0.040541  1011.134  1011.293    4200   24942.68    <.0001


          Robust Measures of Scale

                                     Estimate
Measure                     Value   of Sigma

Interquartile Range      2.000000   1.482602
Gini's Mean Difference   3.187969   2.825264
MAD                      2.000000   2.965200
Sn                       2.385200   2.385660
Qn                       2.221900   2.221234
```

For the pressure_outer_isobar variable, the location statistics are in the range of 1011–1012 hPa. Most of the scale statistics are in the range of 2–3 hPa.

The mean is a nonrobust statistic, whereas the median, trimmed mean, and Winsorized mean are robust. Note that there is not much difference between the nonrobust and robust statistics of location for these data. The pressure_outer_isobar variable has outliers with extreme high *and* extreme low values. Therefore, the outliers did not appreciably change the mean. In general, the mean *is* affected by outliers.

Robust statistics of scale are listed in the "Robust Measures of Scale" table (not shown in Figure 14.5). The table has two columns. The first column lists the value of each robust statistic, and the second column scales the statistics to estimate the normal standard deviation *under the assumption that the data are from a normal sample.* The "Details" section of the UNIVARIATE procedure documentation presents details about the statistics in this table.

The values of the interquartile range and the MAD statistics should be interpreted with caution for these data because the values of the pressure_outer_isobar variable are discrete integers. More important, meteorologists traditionally display on weather maps only the isobars that correspond to even values. For these data, more than 81% of the nonmissing data are even integers.

Specifying the Location and Scale Statistics Analysis

This section describes the dialog box tabs that are associated with the Location and Scale analysis. The Location and Scale Statistics analysis calls the UNIVARIATE procedure in Base SAS software.

Variables Tab

You can use the **Variables** tab to specify the variable for the analysis. Only a single variable can be analyzed at a time. The **Variables** tab is shown in Figure 14.3.

Tables Tab

You can use the **Tables** tab to display tables that summarize the location and scale estimates. The **Tables** tab is shown in Figure 14.4.

The following list describes the tables that can be displayed by the analysis:

Basic measures
: displays statistics that are related to the central location and the spread of the data.

Modes
: displays the most frequently occurring value or values.

Basic confidence intervals
: displays confidence limits for the mean, standard deviation, and variance, under the assumption that the data are normally distributed.

Robust location (trimmed/Winsorized mean)
: displays information and statistics for a two-sided trimmed mean and a two-sided Winsorized mean. You can use the **Cutoff** field to enter the percentage or number of observations to trim or Winsorize.

Robust scale
displays various robust scale statistics.

Tests for location
displays various tests for the hypothesis that the mean or median is equal to a given value. You can use the **Location** field to specify the value. The value is also used in the tables for the trimmed and Winsorized means.

Roles Tab

You can use the **Roles** tab to specify a frequency variable for the analysis. A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents *n* observations, where *n* is the value of the frequency variable.

Analysis of Selected Variables

If an interval variable is selected in a data table when you run the analysis, then that variable is automatically entered in the **Y Variable** field of the **Variables** tab.

If any variable in the data table has a Frequency role, it is automatically entered in the **Frequency Variable** field of the **Roles** tab.

Chapter 15
Distribution Analysis: Distributional Modeling

Contents

Overview of the Distributional Modeling Analysis **233**
Example: Fit a Normal Distribution to Data **234**
Are the Data Normal? 238
How Do the Data Deviate from Normality? 238
What Proportion of the Data Satisfies Certain Conditions? 238
Example: Specify Multiple Density Curves **239**
Specifying the Distributional Modeling Analysis **242**
Variables Tab 242
Estimators Tab 242
Plots Tab 243
Tables Tab 245
Roles Tab 245
Analysis of Selected Variables **245**
References **246**

Overview of the Distributional Modeling Analysis

You can use the Distributional Modeling analysis to fit parametric distributions to univariate data. You can estimate parameters for the fitted distributions, compute goodness-of-fit statistics, and display quantiles of the fitted distributions.

You can use this analysis to create a histogram that is overlaid with up to five density curves. You can create a quantile-quantile (Q-Q) plot to help you determine how well a given distribution fits the data. You can also create a plot of the empirical cumulative distribution function.

You can run a Distributional Modeling analysis by selecting **Analysis ►Distribution Analysis ►Distributional Modeling** from the main menu. When you request distributional modeling, SAS/IML Studio calls the UNIVARIATE procedure in Base SAS software. See the UNIVARIATE procedure documentation in the *Base SAS Procedures Guide* for additional details.

Example: Fit a Normal Distribution to Data

In this example, you fit a normal distribution to the pressure_outer_isobar variable of the Hurricanes data set. The Hurricanes data set contains 6,188 observations of tropical cyclones in the Atlantic basin. The pressure_outer_isobar variable gives the sea-level atmospheric pressure for the outermost closed isobar of a cyclone. This is a measure of the atmospheric pressure at the outermost edge of the storm.

The plots and statistics in the Distributional Modeling analysis can help you answer questions such as the following:

- Can these data be modeled by a parametric distribution? For example, are the data normally distributed?
- If not, which characteristics of the data depart from the fitted distribution? For example, is the data distribution long-tailed? Is it skewed?
- What proportion of the data is within a given range of values?

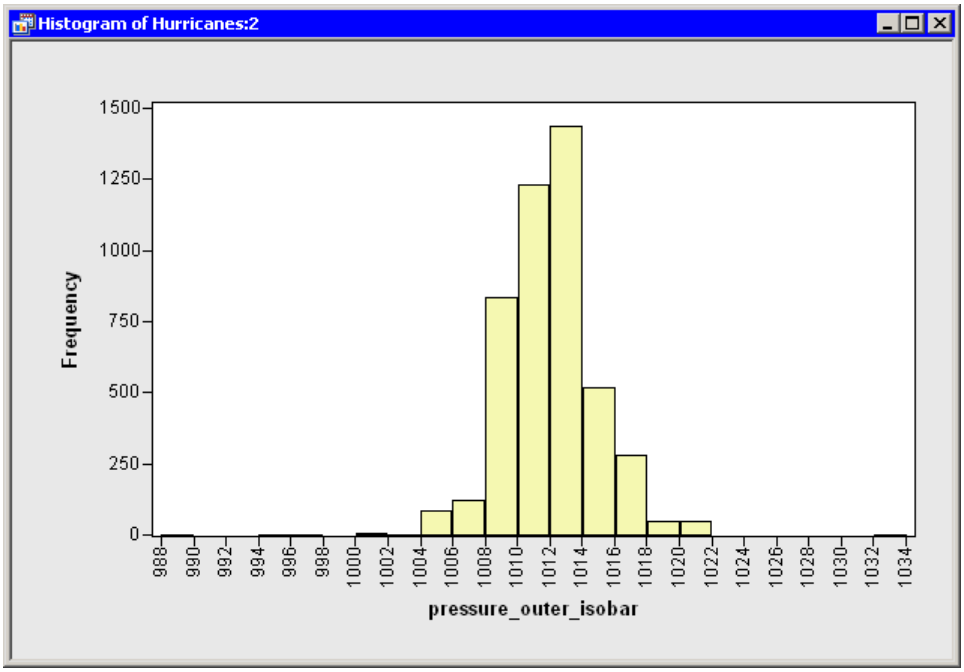
Answers to these questions for the pressure_outer_isobar variable appear at the end of this example.

1 Open the Hurricanes data set.

2 Create a histogram of the pressure_outer_isobar variable.

A histogram appears, as shown in Figure 15.1.

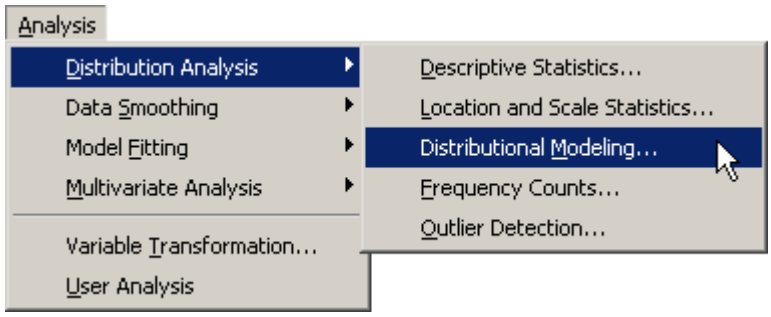
Figure 15.1 A Histogram



From the shape of the histogram you might wonder if the data distribution can be modeled by a normal distribution. If not, how do these data deviate from normality? The following steps add a normal curve to the histogram and create other plots and statistics.

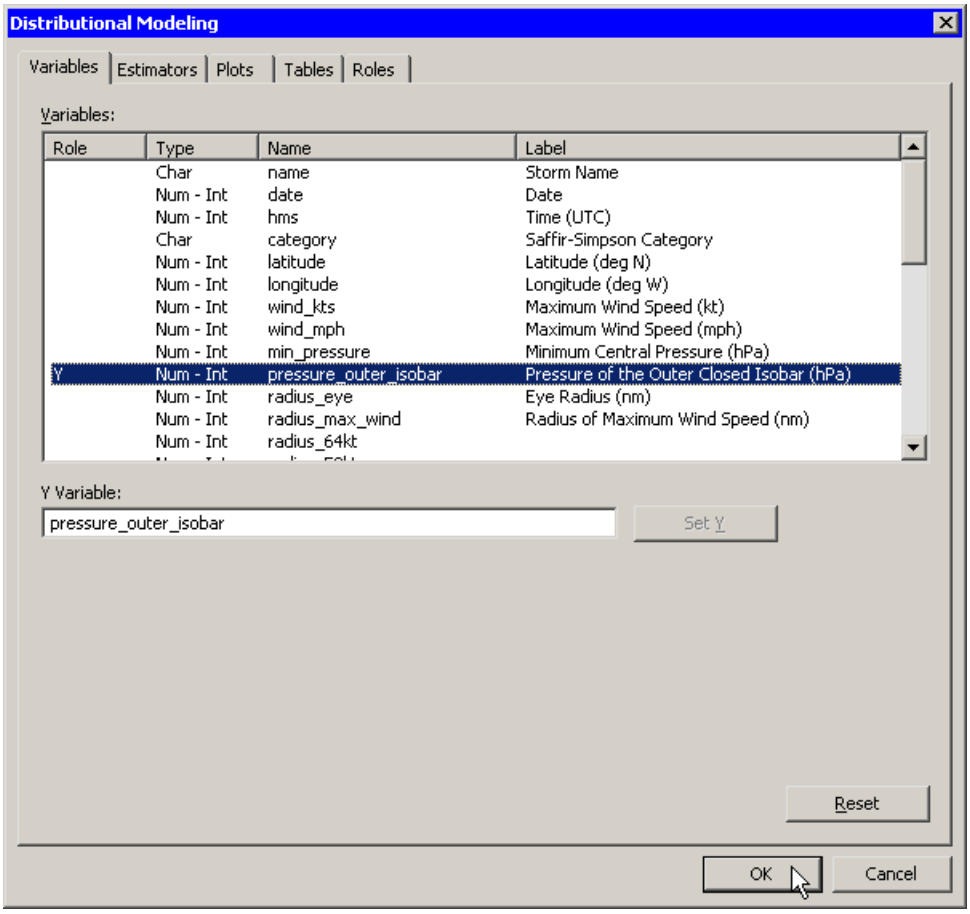
3 Select **Analysis ►Distribution Analysis ►Distributional Modeling** from the main menu, as shown in Figure 15.2.

Figure 15.2 Selecting the Distributional Modeling Analysis



The Distrbutional Modeling dialog box appears. (See Figure 15.3.) You can select a variable for the univariate analysis by using the **Variables** tab.

Figure 15.3 Selecting a Variable

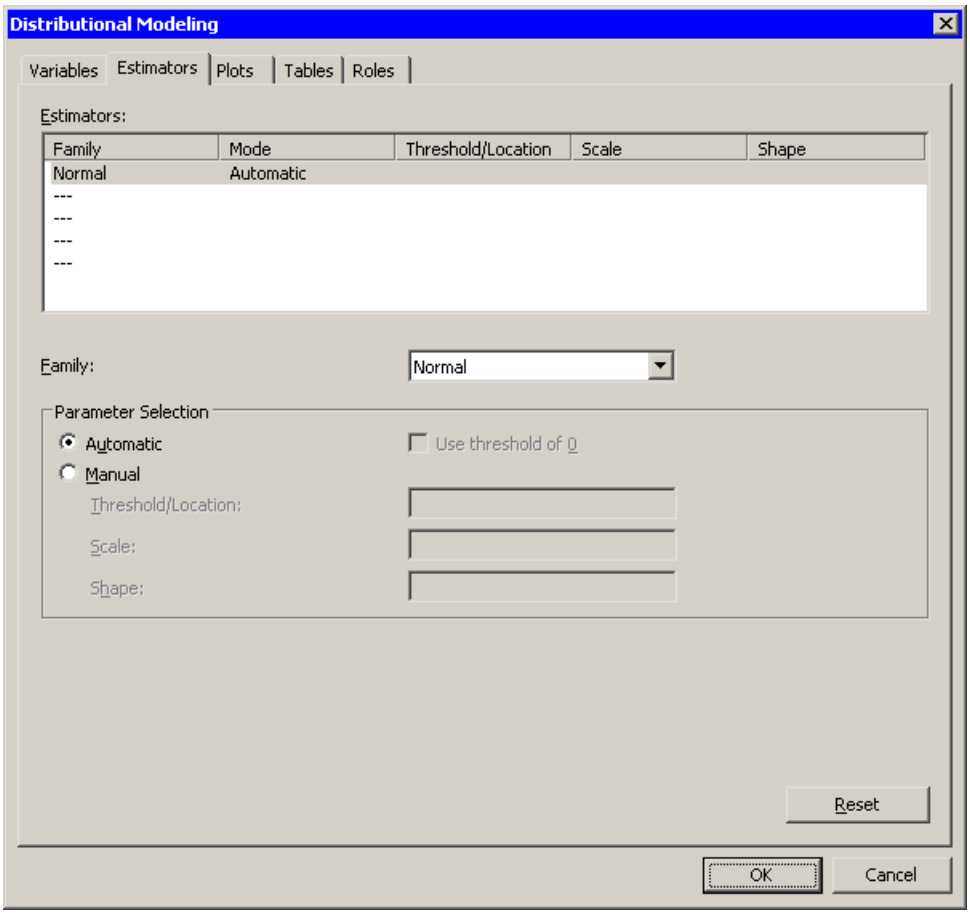


4 Select the variable pressure_outer_isobar, and click **Set Y**.

5 Click the **Estimators** tab.

The **Estimators** tab is shown in Figure 15.4.

Figure 15.4 Selecting a Distribution Family



The **Estimators** tab enables you to select distributions to fit to the data. For each distribution, you can enter known parameters or indicate that the parameters should be estimated by maximum likelihood.

The section "Example: Specify Multiple Density Curves" on page 239 describes how to create a histogram overlaid with more than one density curve. For this example, you select a single distribution to fit to the data.

The normal distribution appears in the **Estimators** list by default. Also by default, the **Automatic** radio button is selected. This specifies that the location and scale parameters for the normal distribution be determined by using maximum likelihood estimation.

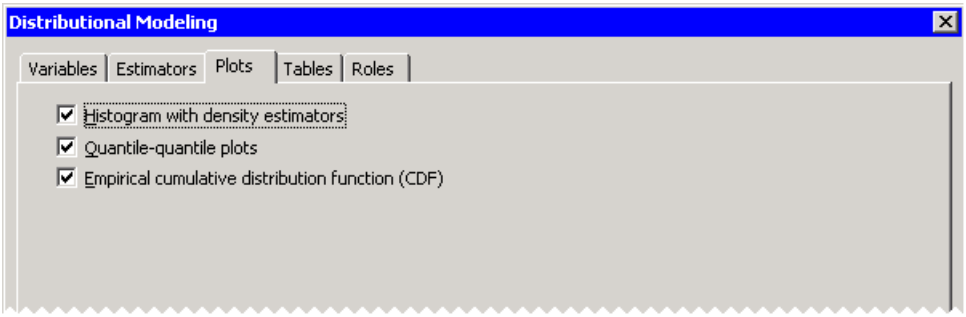
Accept these defaults and proceed to the next tab.

6 Click the **Plots** tab.

7 Select all plots, as shown in Figure 15.5.

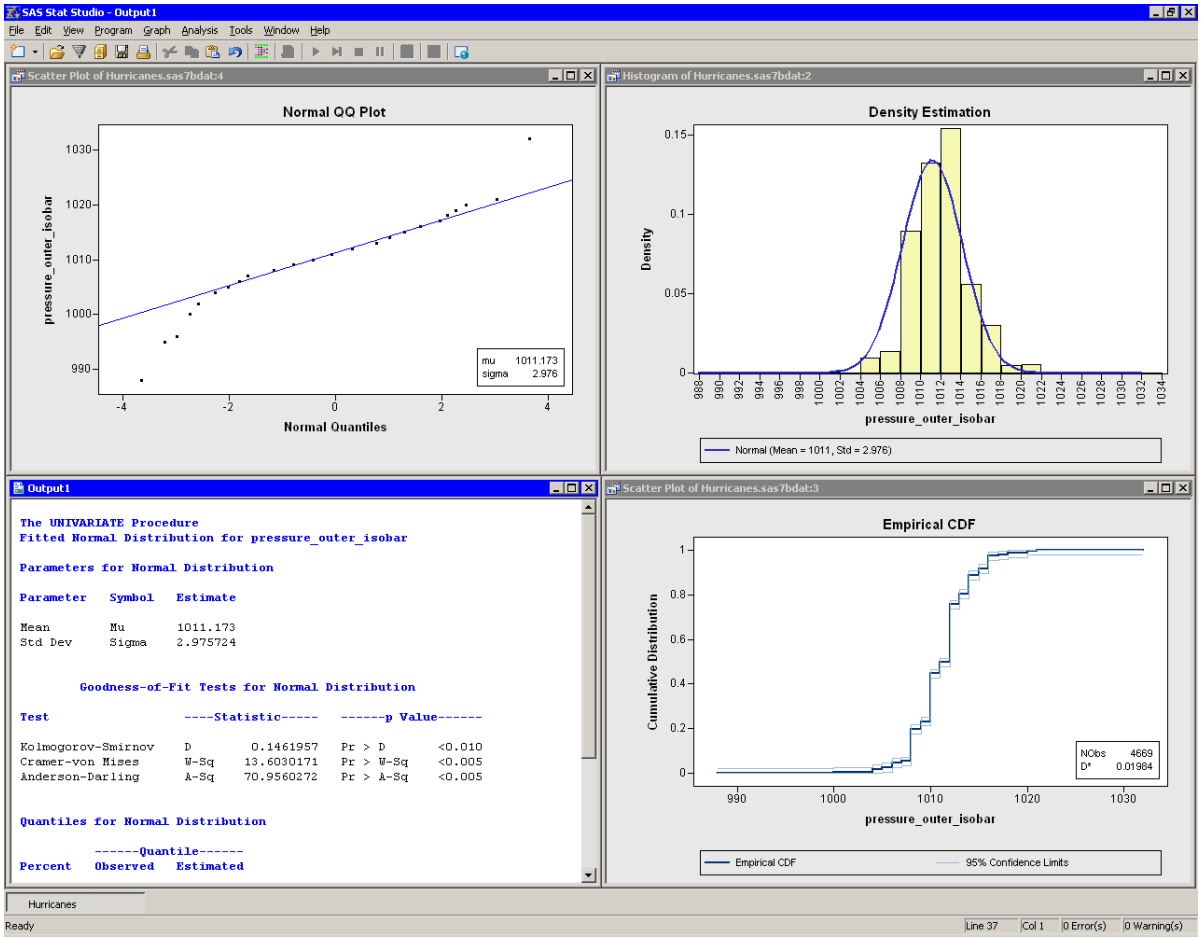
8 Click **OK**.

Figure 15.5 Selecting Plots



The analysis calls the UNIVARIATE procedure, which uses the options specified in the dialog box. The procedure displays tables in the output document, as shown in Figure 15.6.

Figure 15.6 Output from a Distributional Modeling Analysis



Several plots are created. These plots can help answer the questions posed earlier.

Are the Data Normal?

The histogram (the upper right plot in Figure 15.6) is overlaid with a normal density curve. The curve does not fit the data in several locations. The curve predicts more observations in the [1006, 1008] bin than actually occur, and underestimates the count in the [1012, 1014] bin.

How Do the Data Deviate from Normality?

A normal Q-Q plot appears as the upper left plot in Figure 15.6. A Q-Q plot graphically indicates whether there is agreement between quantiles of the data and quantiles of a theoretical distribution. The Q-Q plot for the normal distribution shows several points to the left that are below the diagonal line. These points indicate that the data distribution has a longer left tail than would be expected from normally distributed data. The point to the right that is above the line might indicate an outlier in the data. Table 15.1 describes how to interpret common features of a Q-Q plot.

The goodness-of-fit table in the output document shows that the p-values for the goodness-of-fit tests are very small. The null hypothesis for the goodness-of-fit tests is that the data are from a specified theoretical distribution. The smaller the p-value, the stronger the evidence against the null hypothesis. The small p-values in this example indicate that the normal distribution is not an adequate model to describe these data.

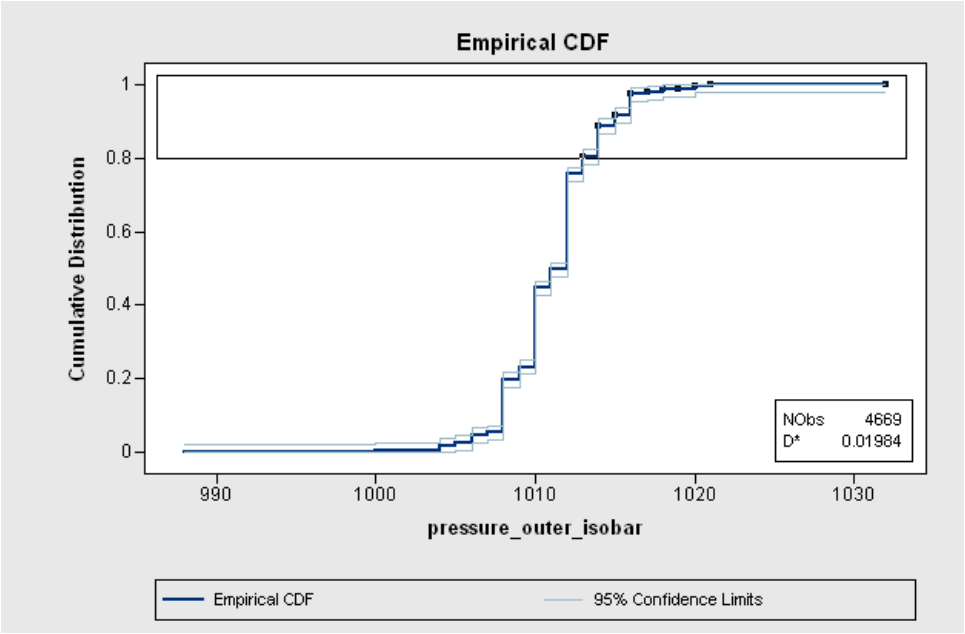
NOTE: The pressure_outer_isobar variable contains 4,669 nonmissing values. For a sample of this size, the goodness-of-fit tests can detect small departures from normality, so it is not surprising that these tests reject the null hypothesis.

What Proportion of the Data Satisfies Certain Conditions?

A CDF plot appears as the lower right plot in Figure 15.6. The CDF plot shows a graph of the empirical cumulative distribution function. You can use the CDF plot to examine relationships between data values and data proportions.

For example, Figure 15.7 graphically answers the question, “What observations are contained in the upper quintile (20%) of the data?” The selected observations show that the answer to the question is, “Data values greater than or equal to 1013 hPa.” Similarly, you can ask a converse question, “What percentage of the data has values less than or equal to 1000 hPa?” The answer (0.4%) can also be obtained by interacting with the CDF plot.

Figure 15.7 A CDF Plot



The CDF plot also shows how data are distributed. For example, the long vertical jumps in the CDF that occur at even values (1008, 1010, and 1012 hPa) indicate that there are many observations with these values. In contrast, the short vertical jumps at odd values (for example, 1009, 1011, and 1013 hPa) indicate that there are not many observations with these values. This fact is not apparent from the histogram, because the default bin width is 2 hPa.

Example: Specify Multiple Density Curves

You can overlay two (or more) density curves on a single histogram. The curves can be different distributions from the same family or distributions from different families.

In this section, you fit a lognormal distribution and a Weibull distribution to data in the radius_eye variable. The radius_eye variable gives the radius of a cyclone's eye (if an eye exists), in nautical miles. (The eye of a cyclone is a calm, relatively cloudless central region.)

NOTE: There are often scientific or engineering considerations that lead to the choice of either a lognormal or a Weibull model. This example does not have a scientific basis; it merely illustrates how you can add multiple curves to a histogram.

To fit parametric distributions to the data:

1 Select **Analysis ►Distribution Analysis ►Distributional Modeling** from the main menu.

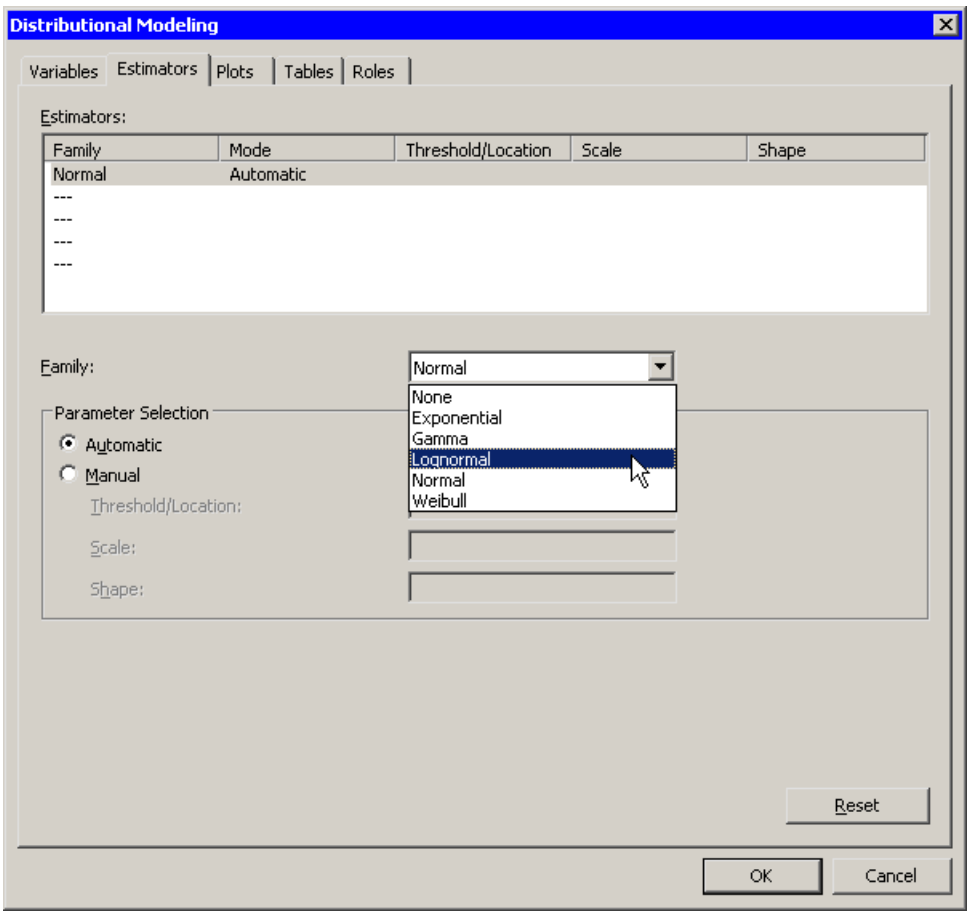
2 Select the variable radius_eye, and click **Set Y**.

3 Click the **Estimators** tab.

The normal distribution appears in the **Estimators** list. The next step changes this item to a lognormal distribution.

4 Select the first item ("Normal") in the **Estimators** list. Select **Lognormal** from the **Family** list, as shown in Figure 15.8.

Figure 15.8 Selecting a Lognormal Distribution



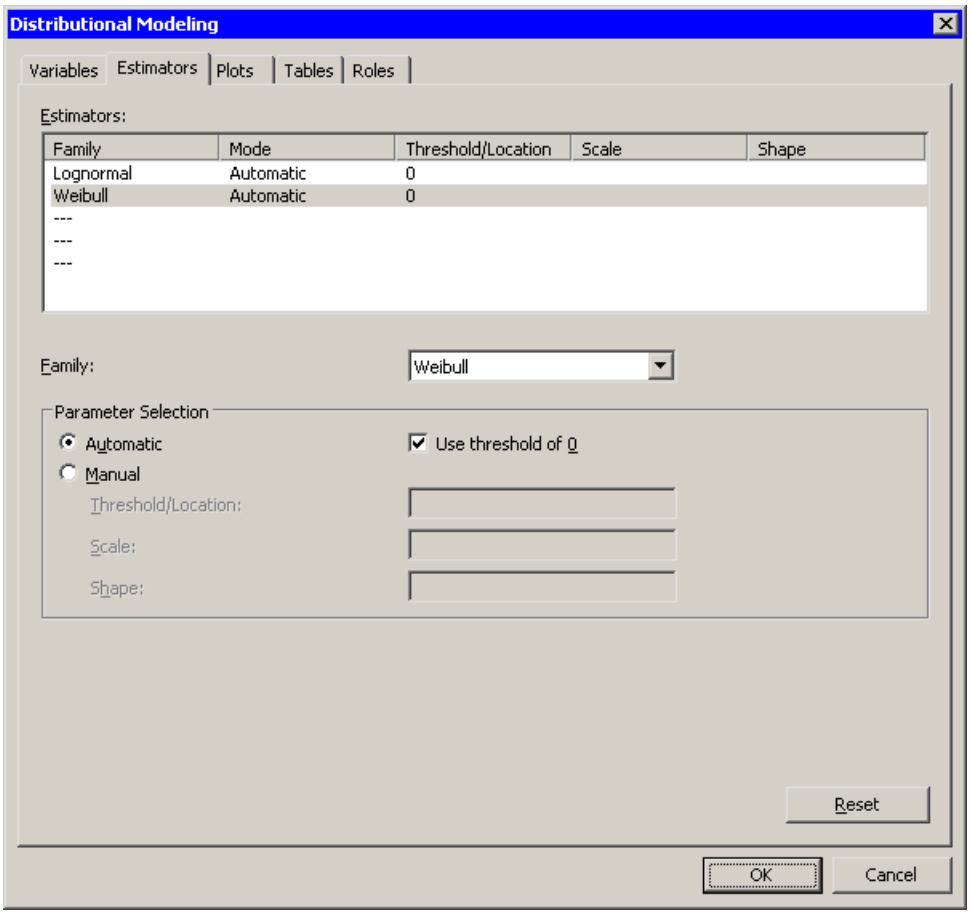
The lognormal distribution has three parameters. By default, the *thresholdparameter* is set to zero, and the *scale* and *shape* parameters are estimated by maximum likelihood.

The next two steps add a Weibull distribution to the **Estimators** list.

5 Select the second item (a dashed line) in the **Estimators** list.

6 Select **Weibull** from the **Family** list.

Figure 15.9 Selecting Multiple Distributions

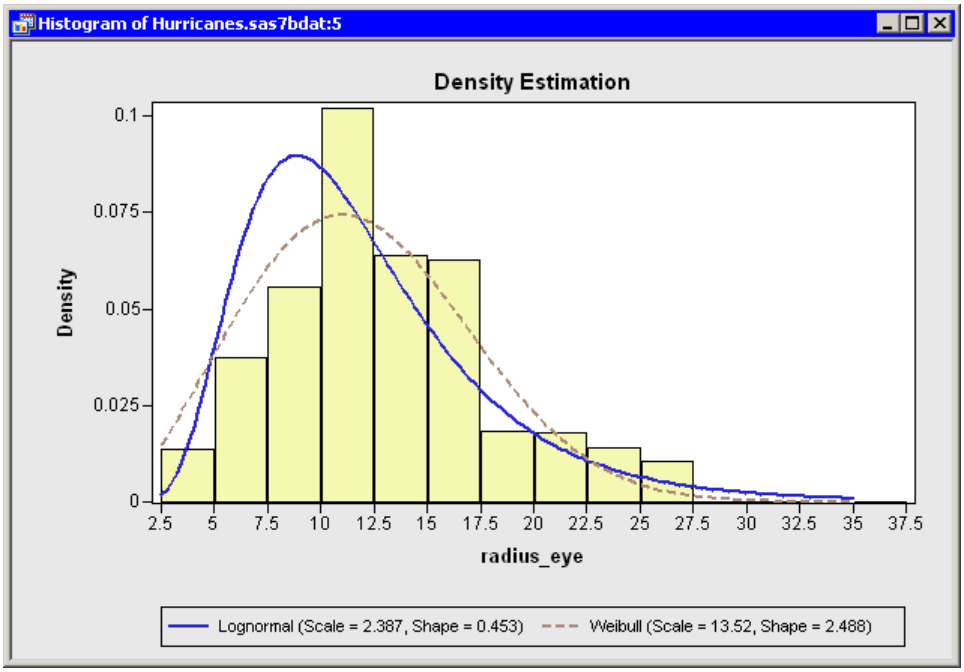


The Weibull distribution also has three parameters. Again, the threshold parameter defaults to zero, whereas the other parameters are estimated by maximum likelihood. Accept these defaults, as shown in Figure 15.9.

7 Click **OK**.

Two density curves are added to the histogram, as shown in Figure 15.10. If these were competing scientific models, you could analyze and compare the merits of the models.

Figure 15.10 Multiple Density Curves



Specifying the Distributional Modeling Analysis

This section describes the dialog box options that are associated with the Distributional Modeling analysis. The Distributional Modeling analysis calls the UNIVARIATE procedure in Base SAS software.

Variables Tab

You can use the **Variables** tab to specify the variable for the analysis. Only a single variable can be analyzed at a time. The **Variables** tab is shown in Figure 15.3.

Estimators Tab

You can use the **Estimators** tab to specify parametric distributions to fit to the data. (See Figure 15.4.) The options for the **Estimators** tab correspond to options for the HISTOGRAM statement in the UNIVARIATE procedure. See the documentation in the *Base SAS Procedures Guide* for details.

For each distribution, you can enter values for one or more parameters, and estimate the remaining parameters with maximum likelihood estimation (MLE). The analysis typically creates a histogram overlaid with density curves, one for each specified distribution.

To add a new distribution to the **Estimators** list, click a blank item and select a distribution from the **Family** list.

To delete a distribution from the **Estimators** list, click an existing distribution and select **None** from the **Family** list.

To change a distribution in the **Estimators** list, click the distribution and select a new distribution from the **Family** list.

Threshold parameters are set to zero unless you clear the **Use threshold of 0** check box, in which case the threshold parameter is estimated by MLE. Other parameters in a distribution are estimated from the data by using MLE, unless you select **Manual** parameter selection.

The **Estimator** tab contains the following UI controls:

Estimators
: displays a list of distributions that are fitted to the data. Clicking an item in this list enables you to change the distribution or to specify parameters for the distribution. You can specify up to five distributions.

Family
: specifies the distribution for the selected item in the **Estimators** list.

Parameter Selection
: specifies how to determine parameters of the selected distribution in the **Estimators** list. If **Automatic** is selected, then parameters are estimated by using MLE. If **Manual** is selected, then you can enter one or more known parameters. Unspecified parameters are estimated by using MLE.

Use threshold of 0
: specifies whether the threshold parameter is set to zero for the current distribution. If you clear this check box, then the threshold parameter is estimated by using MLE.

NOTE: Maximum likelihood estimation of two parameters does not always converge. Three-parameter estimation *often* does not converge. Three-parameter estimation is attempted if you clear the **Use threshold of 0** check box while **Automatic** is selected.

Plots Tab

You can use the **Plots** tab to create the following plots:

Histogram with density estimators
: creates a histogram overlaid with density curves for the parametric distributions that are specified on the **Estimators** tab.

Quantile-quantile plots
: creates one Q-Q plot for each parametric distribution that is specified on the **Estimators** tab.

Empirical cumulative distribution function (CDF)
: creates a plot of the empirical cumulative distribution function.

NOTE: SAS/IML Studio adds a density curve to an *existing* histogram when both of the following conditions are satisfied:

- The histogram is the active window when you select the analysis.
- The histogram variable and the analysis variable are the same.

Q-Q Plots

A Q-Q plot graphically indicates whether there is agreement between quantiles of the data and quantiles of a theoretical distribution. If the quantiles of the theoretical and data distributions agree, the plotted points fall along a straight line. For most distributions, the slope of the line is the value of the scale parameter, and the intercept of the line is the value of the threshold or location parameter. (For the lognormal distribution, the slope is e^{ζ}, where ζ is the value of the scale parameter.) The parameter estimates for the distribution that best fits the data appear in an inset in the Q-Q plot.

Table 15.1 presents reasons why the points in a Q-Q plot might not be linear.

Table 15.1 Interpretation of Q-Q Plots

Description of Point Pattern	Possible Interpretation
All but a few points fall on a line.	There are outliers in the data.
Left end of pattern is below the line; right end of pattern is above the line.	There are long tails at both ends of the data distribution.
Left end of pattern is above the line; right end of pattern is below the line.	There are Short tails at both ends of the data distribution.
Curved pattern with slope that increase from left to right	Data distribution is skewed to the right.
Curved pattern with slope that decreases from left to right	Data distribution is skewed to the left.
Most points are not near line $ax + b$ with scale parameter a and location parameter b.	Data do not fit the theoretical distribution.

NOTE: When the variable being graphed has repeated values, the Q-Q plot produced by SAS/IML Studio is different from the Q-Q plot produced by the UNIVARIATE procedure. The UNIVARIATE procedure arbitrarily ranks the repeated values and assigns a quantile for the theoretical distribution based on the ranks. Two observations with the same value are assigned different quantiles. If a variable has many repeated values, the Q-Q plot produced by the UNIVARIATE procedure looks like a staircase. However, SAS/IML Studio (and SAS/INSIGHT) averages the ranks of repeated values. Two observations with the same value are assigned the same quantiles for the theoretical distribution.

CDF Plots

A CDF plot shows the empirical cumulative distribution function. You can use the CDF plot to examine relationships between data values and data proportions. For example, you can determine whether a given percentage of your data is below some upper control limit. You can also determine what percentage of the data has values within a given range of values.

The inset for the CDF plot displays two statistics. The first is the number of nonmissing observations for the plotted variable. The second is labeled D^*. If D is the 95% quantile for Kolmogorov's D distribution ($D \approx 1.36$) and N is the number of nonmissing observations, then (D'Agostino and Stephens 1986)

$$D^* = D/\left(\sqrt{N} + 0.12 + 0.11/\sqrt{N}\right)$$

The 95% confidence limits in the CDF plot are obtained by adding and subtracting D^* from the empirical CDF. They form a confidence band around the estimate for the cumulative distribution function.

Tables Tab

You can use the **Tables** tab to display the following tables that summarize the results of the univariate analysis:

Parameter estimates
: displays parameter estimates for the specified theoretical distribution.

Goodness-of-fit tests
: displays goodness-of-fit statistics that test whether the data come from the specified theoretical distribution.

Quantiles of fitted distribution
: displays quantile information for the data and theoretical distributions.

Roles Tab

You can use the **Roles** tab to specify a frequency variable for the analysis. A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents n observations, where n is the value of the frequency variable.

Analysis of Selected Variables

If an interval variable is selected in a data table when you run the analysis, then that variable is automatically entered in the **Y Variable** field of the **Variables** tab.

If any variable in the data table has a Frequency role, it is automatically entered in the **Frequency Variable** field of the **Roles** tab.

References

D'Agostino, R. B. and Stephens, M., eds. (1986), *Goodness-of-Fit Techniques*, New York: Marcel Dekker.

Chapter 16
Distribution Analysis: Frequency Counts

Contents

Overview of the Frequency Counts Analysis **247**
Example: Display Frequency Counts **247**
Specifying the Frequency Counts Analysis **252**
Variables Tab 252
Plots Tab 252
Tables Tab 253
Roles Tab 253
Analysis of Selected Variables **254**

Overview of the Frequency Counts Analysis

You can use the Frequency Counts analysis to produce one-way frequency tables and compute chi-square statistics to test for equal proportions.

You can use the analysis to tabulate the number of observations in each category of a variable. For nominal variables, you can also create a bar chart of the variable.

You can run a Frequency Counts analysis by selecting **Analysis ►Distribution Analysis ►Frequency Counts** from the main menu. When you request a one-way frequency table and associated statistics, SAS/IML Studio calls the FREQ procedure in Base SAS software.

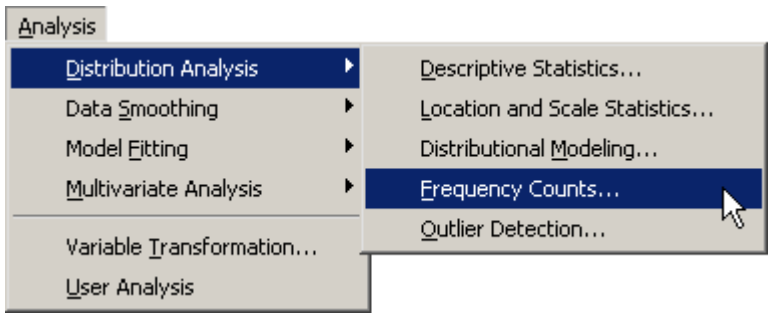
Example: Display Frequency Counts

In this example, you create a one-way frequency table for the category variable of the Hurricanes data set. The Hurricanes data set contains 6,188 observations of tropical cyclones in the Atlantic basin. The category variable gives the Saffir-Simpson category of the tropical cyclone for each observation. A missing value of the category variable means that the storm had an intensity of less than tropical depression strength (wind speeds less than 22 knots) at the time of observation.

To create a one-way frequency table:

1. Open the Hurricanes data set.
2. Select **Analysis ►Distribution Analysis ►Frequency Counts** from the main menu, as shown in Figure 16.1.

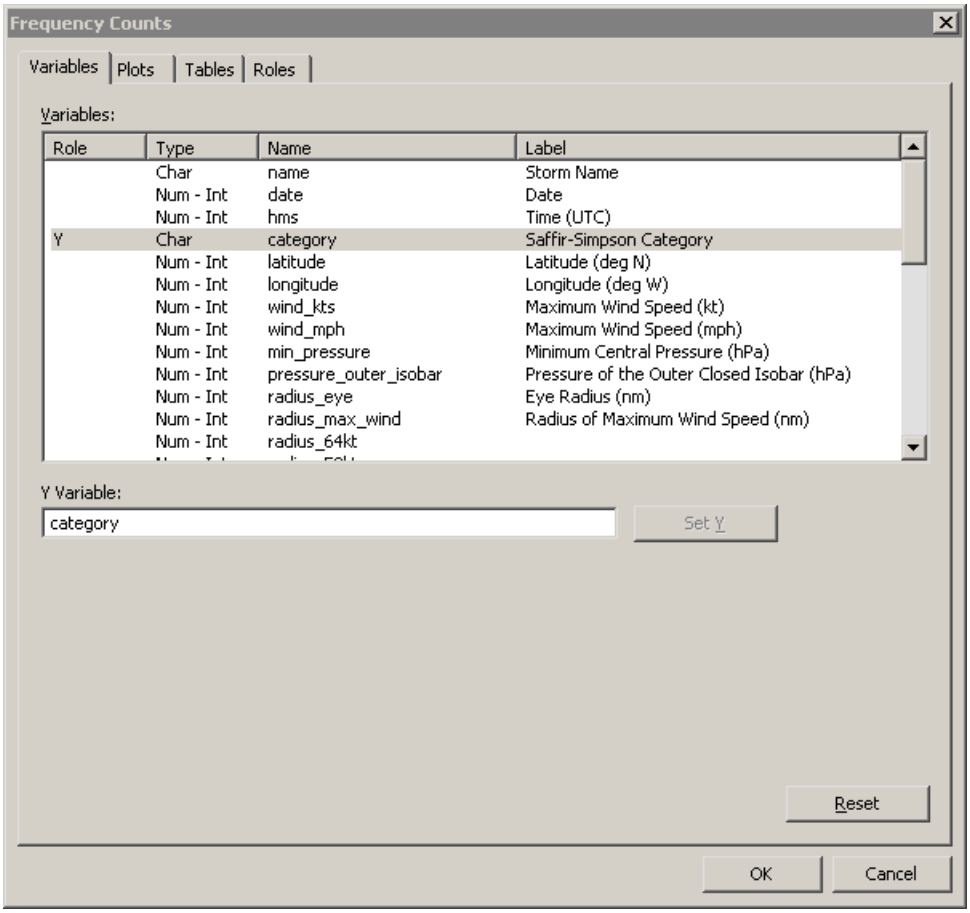
Figure 16.1 Selecting the Frequency Counts Analysis



The Frequency Counts dialog box appears. (See Figure 16.2.) You can select a variable for the analysis on the **Variables** tab.

3 Select the variable category, and click **Set Y**.

Figure 16.2 Specifying a Variable



For nominal variables, you can produce a bar chart of the categories of the chosen variable.

4 Click the **Plots** tab.

The **Plots** tab becomes active. (See Figure 16.3.)

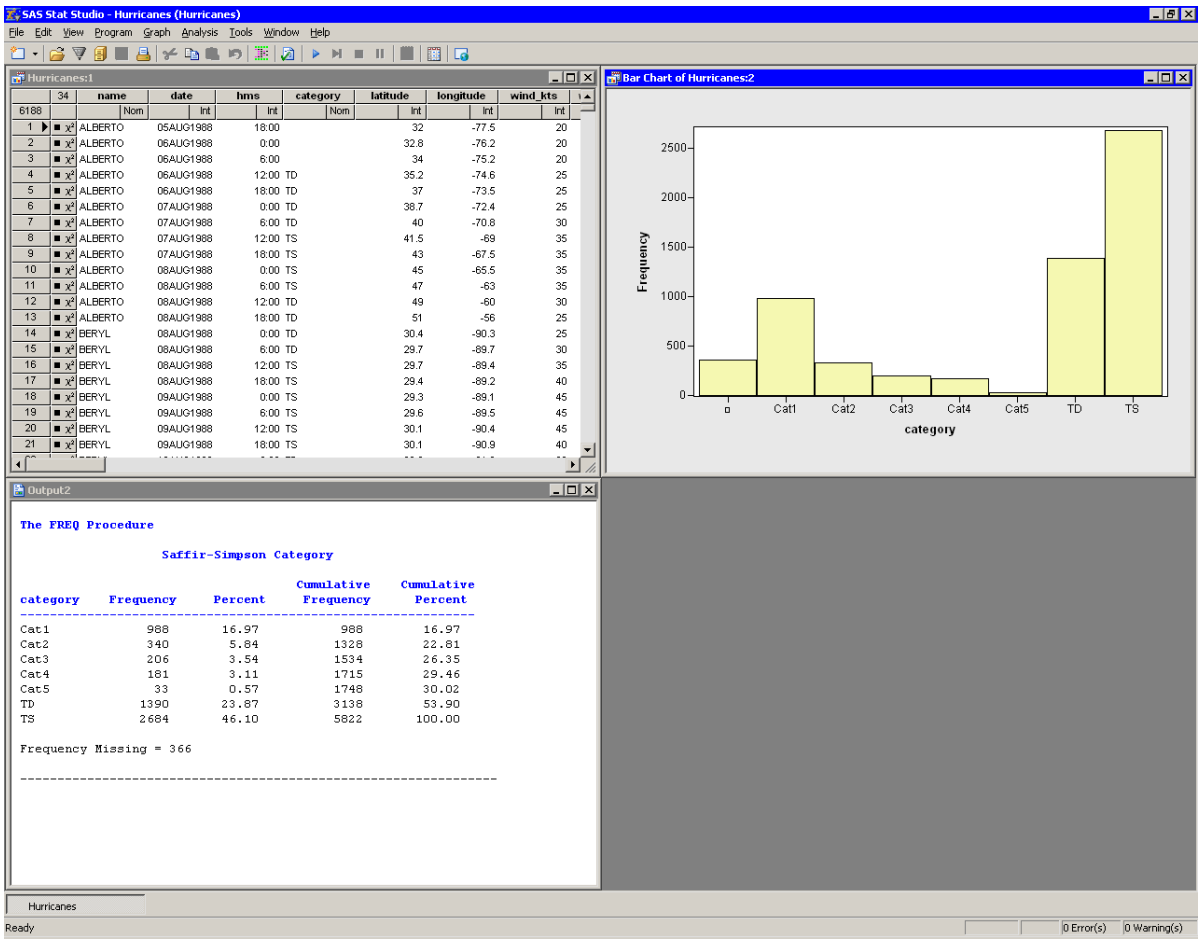
5 Select **Bar chart**.

6 Click **OK**.

Figure 16.3 Selecting Plots

Figure 16.4 shows the results of this analysis. The analysis calls the FREQ procedure, which uses the options specified in the dialog box. The procedure displays a frequency table in the output document. The table shows the frequency and percent of each Saffir-Simpson category for these data. Hurricanes of category 3 or higher account for only 7% of the nonmissing data, whereas almost half of the observations are classified as tropical storms.

Figure 16.4 Output from a Frequency Counts Analysis



The bar chart shows a graphical view of the category variable. You can create a graphical version of the output table by labeling the bars in the bar chart with their frequencies or percentages. To add labels to the bar chart, do the following:

7 Right-click near the center of the plot area. Select **Plot Area Properties** from the pop-up menu.

The Plot Area Properties dialog box appears. (See Figure 16.5.) The **Bars** tab controls attributes of the bar chart.

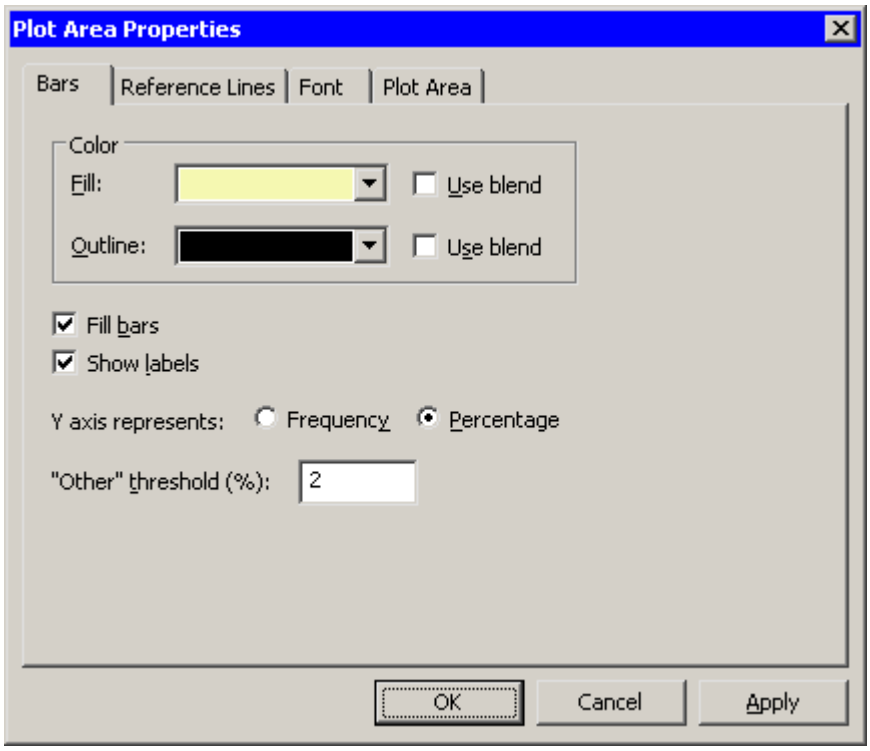
8 Click **Show labels**.

9 Click **Y axis represents: Percentage**.

10 Click **OK**.

NOTE: You can also label the bar chart by using keyboard shortcuts. Activate the bar chart. Press the "l" key (lowercase "L") to toggle labels. Press the "p" key to alternate between displaying frequency and percentage.

Figure 16.5 Bar Chart Properties



The percentages displayed on the bar chart do not match the percentages in the one-way frequency table. That is because the bar chart includes the 366 missing observations in the total number of observations, whereas the analysis does not include those observations by default. (The counts for each bar *do* match the counts in the table; only the percentages differ.)

To exclude missing values from the bar chart:

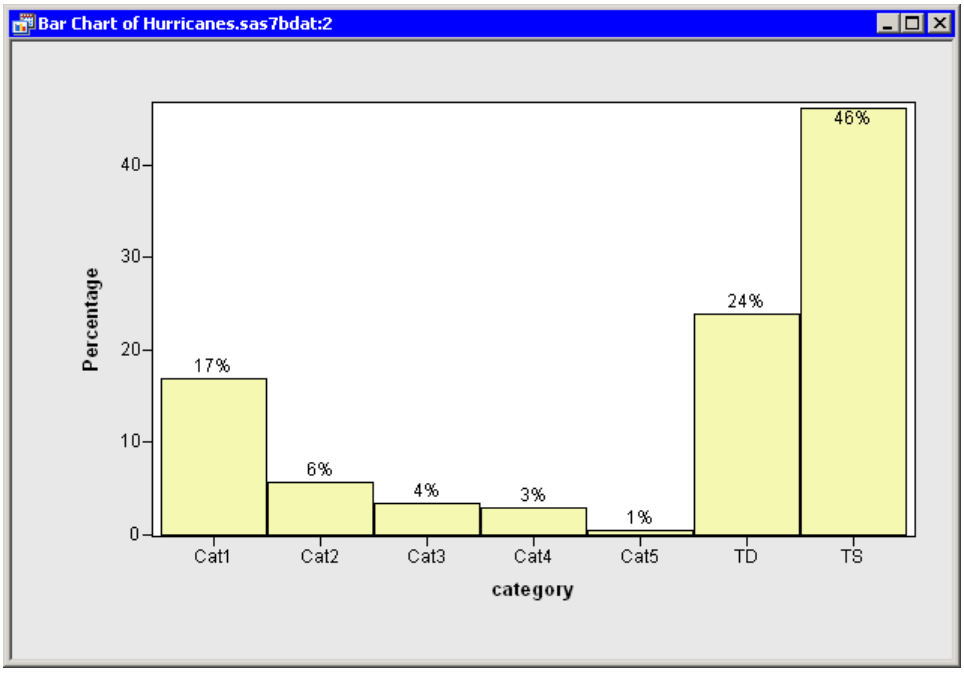
1 Select the missing observations by clicking the first bar in the bar chart.

2 Select the data table to make it the active window.

3 Select **Edit ►Observations ►Exclude from Plots**.

The bar chart now omits the missing values as shown in Figure 16.6.

Figure 16.6 The Bar Chart without Missing Values



Alternatively, if you want to include missing values as a valid category, then you can specify that the one-way table should include a category of missing values.

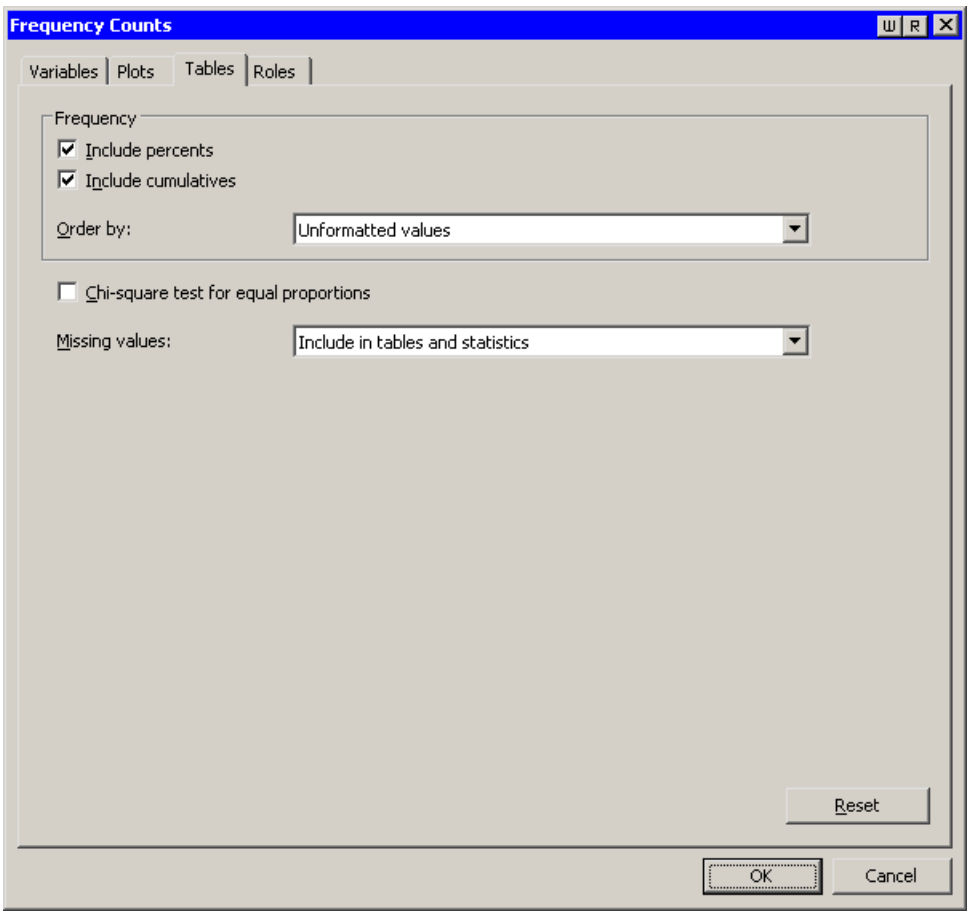
To specify options for the Frequency Counts analysis:

1 Click the **Tables** tab, as shown in Figure 16.7.

2 In the **Missing values** list, select the option **Include in tables and statistics**.

This option specifies that missing values should be regarded as a valid category. If you run (or rerun) the analysis with this option, the one-way table includes missing values as a valid category. The frequency table produced with this option agrees with the default bar chart.

Figure 16.7 The Tables Tab



Specifying the Frequency Counts Analysis

This section describes the dialog box tabs that are associated with the Frequency Counts analysis. The Frequency Counts analysis calls the FREQ procedure in Base SAS software to compute counts and percentages of each unique value of a variable.

Variables Tab

You can use the **Variables** tab to specify the variable for the TABLES statement of the FREQ procedure. Only a single variable can be analyzed at a time. The **Variables** tab is shown in Figure 16.2.

Plots Tab

You can use the **Plots** tab to create a bar chart if the chosen variable is nominal. (See Figure 16.3.) If the chosen variable is not nominal, the analysis prints a warning message to the log. (You can convert an interval

variable to nominal. In the data table, right-click the variable's column heading and select **Nominal** from the pop-up menu.)

Tables Tab

You can use the **Tables** tab, shown in Figure 16.7, to specify the options used to produce the one-way frequency table. Each of these options corresponds to an option in the FREQ procedure, as indicated in the following list.

Include percents
specifies that a column of percentages be included in the one-way frequency table.

Include cumulatives
specifies that a column of cumulative percentages be included in the one-way frequency table.

Order by
specifies the order in which the values of the variable appear in the frequency table. This corresponds to the ORDER= option in the PROC FREQ statement.

Chi-square test for equal proportions
requests a chi-square goodness-of-fit test for equal proportions. This corresponds to the CHISQ option in the TABLES statement.

Missing values
specifies the treatment of missing values. The following options are supported:

Exclude from tables and statistics specifies that missing values be excluded from the analysis.

Include in tables; Exclude from statistics specifies that missing value frequencies be displayed, even though the frequencies are not used in the calculation of statistics. This corresponds to the MISSPRINT option in the TABLES statement.

Include in tables and statistics specifies that missing values be treated the same as nonmissing values: they are included in calculations of percentages and other statistics. This corresponds to the MISSING option in the TABLES statement.

Roles Tab

You can use the **Roles** tab to specify a weight variable for the analysis. The weight variable in the FREQ procedure is a numeric variable whose value represents the frequency of the observation. If you use a weight variable, the FREQ procedure assumes that each observation represents n observations, where n is the value of the weight variable. For further information, see the documentation for the FREQ procedure in the *SAS/STAT User's Guide*.

Analysis of Selected Variables

If a variable is selected in a data table when you run the analysis, then that variable is automatically entered in the **Y Variable** field of the **Variables** tab.

If any variable in the data table has a Weight role, it is entered in the **Weight Variable** field of the **Roles** tab.

Chapter 17
Distribution Analysis: Outlier Detection

Contents

Overview of the Outlier Detection Analysis **255**
Example: Detect Univariate Outliers **256**
Specifying the Outlier Detection Analysis **260**
Variables Tab 260
Method Tab 260
Plots Tab 261
Output Variables Tab 262
Roles Tab 262
Analysis of Selected Variables **262**

Overview of the Outlier Detection Analysis

The Outlier Detection analysis computes outliers in contaminated normally distributed data. This analysis defines outliers as values that are sufficiently far from an estimate of the central tendency of the data.

More formally, suppose the data are normally distributed with location parameter μ and scale parameter σ. Let $\hat{\mu}$ be an estimate of the location parameter. Let $\hat{\sigma}$ be an estimate of the scale parameter. Then a value x is considered an outlier if

$$|x - \hat{\mu}| > c\hat{\sigma}$$

where c is a constant that you can specify. The constant c is called the *scale multiplier*.

The basic idea is that if the data are normally distributed, then about 99% of the data are within three standard deviations of the mean. Therefore, if you can accurately estimate the mean (location parameter) and standard deviation (scale parameter), you can identify values in the tails of the distribution. However, if the data contain outliers, then you need to use robust estimators of the location and scale parameters. Robust estimates are described in the "Details" section of the documentation for the UNIVARIATE procedure in the *Base SAS Procedures Guide*.

You can use the analysis to specify traditional or robust estimates of location and scale parameters for a numerical variable. You can create a histogram with a normal curve overlaid. You can create an indicator variable that has the value 1 for observations that are sufficiently far from the location estimate.

You can run an Outlier Detection analysis by selecting **Analysis ►Distribution Analysis ►Outlier Detection** from the main menu. When you request outlier detection, SAS/IML Studio calls the UNIVARIATE procedure in Base SAS software to compute location and scale estimates. SAS/IML statements are then used to compute the outliers.

Example: Detect Univariate Outliers

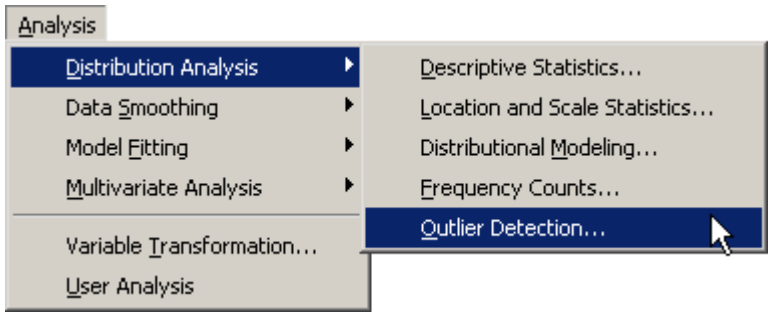
In this example, you detect outliers for the pressure_outer_isobar variable of the Hurricanes data set. The Hurricanes data set contains 6,188 observations of tropical cyclones in the Atlantic basin. The pressure_outer_isobar variable gives the sea-level atmospheric pressure for the outermost closed isobar of a cyclone. This is a measure of the atmospheric pressure at the outermost edge of the storm. There are 4,662 nonmissing values of pressure_outer_isobar.

To find outliers in univariate data:

1 Open the Hurricanes data set.

2 Select **Analysis ►Distribution Analysis ►Outlier Detection** from the main menu, as shown in Figure 17.1.

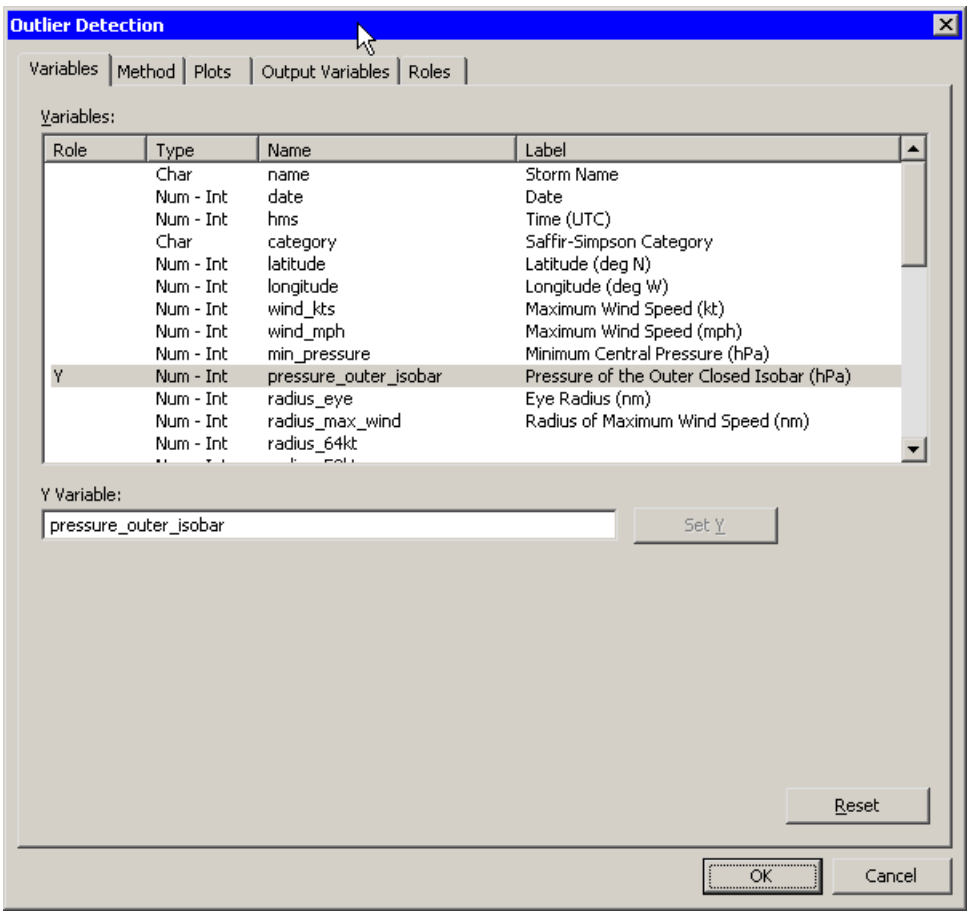
Figure 17.1 Selecting the Outlier Detection Analysis



The Outlier Detection dialog box appears. (See Figure 17.2.) You can select a variable for the analysis by using the **Variables** tab.

3 Select the variable pressure_outer_isobar, and click **Set Y**.

Figure 17.2 Specifying a Variable



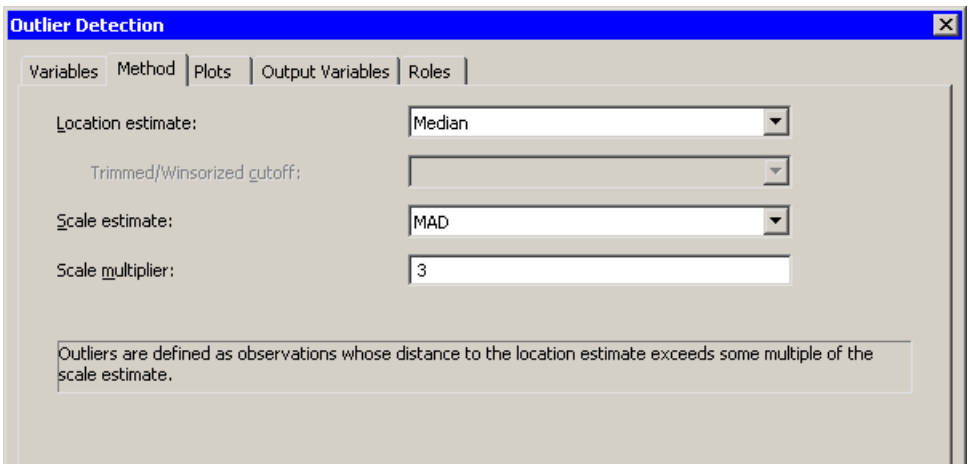
You can specify how the location and scale parameters are estimated by using the **Method** tab.

4 Click the **Method** tab.

The **Method** tab becomes active. (See Figure 17.3.) The default is to estimate the location with the median of the data, and to estimate the scale with the median absolute deviation from the median (MAD). Each estimate is described in the documentation for the UNIVARIATE procedure in the *Base SAS Procedures Guide*. The default scale multiplier is 3.

You can accept the default method parameters for this example.

Figure 17.3 Specifying the Method



5 Click the **Plots** tab.

The **Plots** tab becomes active. (See Figure 17.4.)

6 Select **Normal quantile-quantile plot**.

7 Click **OK**.

Figure 17.4 Selecting Plots

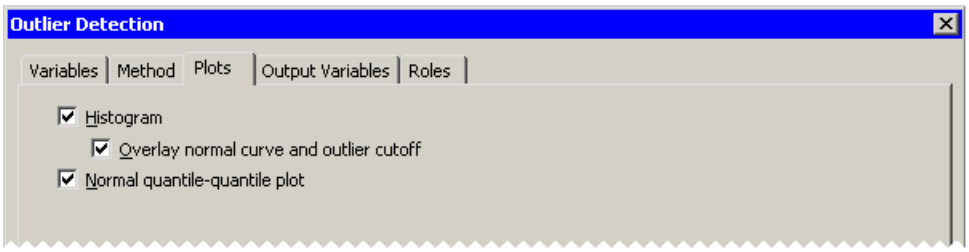
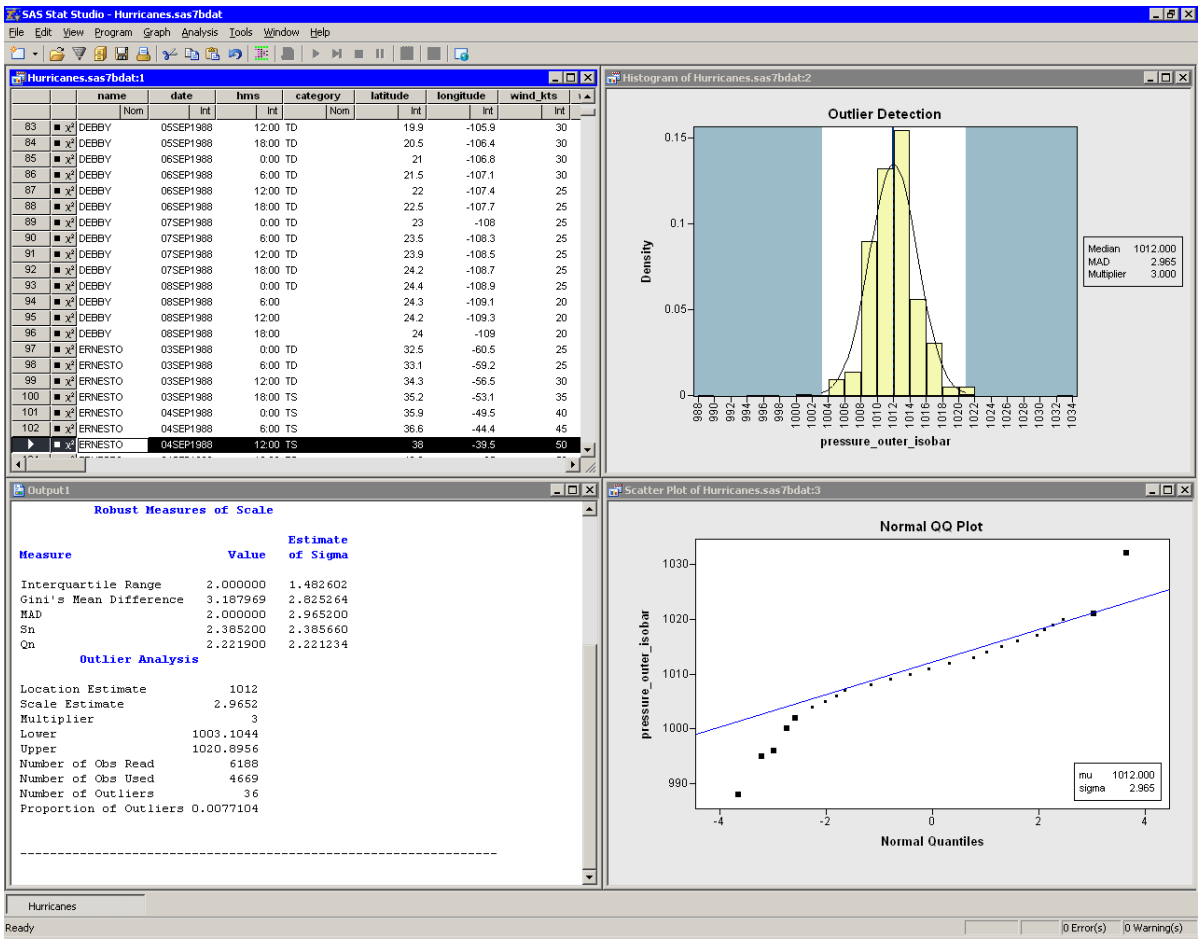


Figure 17.5 shows the results of this analysis. The analysis calls the UNIVARIATE procedure, which uses the options specified in the dialog box. The procedure displays tables in the output document. The tables show several estimates of the location and scale parameters. For this example, the median is 1012 hPa with a scale estimate of 2.965. SAS/IML statements are then used to read in the specified estimates and to compute values of pressure_outer_isobar that are more than $3 \times 2.965 = 8.895$ units away from 1012.

Two plots are created. One shows a histogram of the selected variable. The histogram is overlaid with a normal curve with $\mu = 1012$ and $\sigma = 2.965$. A vertical line at 1012 indicates the location estimate, and shading indicates regions that are more than 8.965 units from 1012. The other plot is a normal Q-Q plot of the data.

Figure 17.5 Output from an Outlier Detection Analysis



By default, the analysis adds an indicator variable to the data table. The indicator variable is named Outlier_*Y*, where *Y* is the name of the chosen variable. You can select all observations that are marked as outliers by doing the following:

8 Select the data table window to make it active.

9 Select **Edit ►Find** from the main menu.

The Find dialog box appears as in Figure 17.6.

10 Select Outlier_pressure_outer_isobar from the **Variable** list.

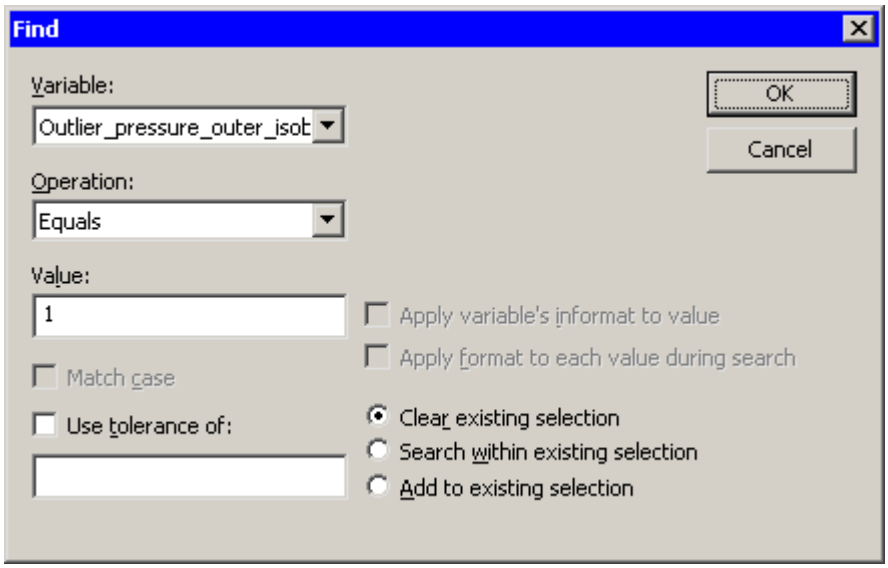
11 Select **Equals** from the **Operation** list.

12 Type `1` in the **Value** field.

13 Click **OK**.

There are 36 observations marked as outliers. If the data table is active, you can use the F3 key to advance to the next selected observation. (Alternatively, you can use **Edit ►Observations ►Examine Selected Observations** to examine each selected observation in turn.) The normal Q-Q plot shows that the quantiles of the unselected observations fall along a straight line, which indicates that those observations appear to be normally distributed. (See Figure 17.5.) The selected observations (the outliers) deviate from the line.

Figure 17.6 Finding Outliers



Specifying the Outlier Detection Analysis

This section describes the dialog box tabs that are associated with the Outlier Detection analysis. The Outlier Detection analysis calls the UNIVARIATE procedure in Base SAS software to compute estimates of the location and scale. SAS/IML statements are then used to determine which values are sufficiently far from the location estimate.

Variables Tab

You can use the **Variables** tab to specify the variable for the analysis. Only a single variable can be analyzed at a time. The **Variables** tab is shown in Figure 17.2.

Method Tab

You can use the **Method** tab to specify the following options for estimating the location and scale parameters for the data, and for specifying the scale multiple. The **Method** tab is shown in Figure 17.3.

The **Method** tab contains the following UI controls:

Location estimate
lists statistics that are used to estimate the location parameter for the data. Each statistic is described in the "Details" section of the UNIVARIATE procedure documentation in the *Base SAS Procedures Guide*. The statistics are as follows:

Mean estimates the location parameter by using the mean of the data. (**NOTE:** The mean is not a robust statistic; it is influenced by outliers.)

Median estimates the location parameter by using the median of the data.

Trimmed mean estimates the location parameter by using the trimmed mean of the data.

Winsorized mean estimates the location parameter by using the Winsorized mean of the data.

Trimmed/Winsorized cutoff
specifies the number of observations or proportion of observations used to estimate a trimmed or Winsorized mean.

Scale estimate lists the statistics for estimating the scale parameter for the (uncontaminated) data. The statistics are as follows:

Standard deviation estimates the scale parameter by using the standard deviation of the data. (**NOTE:** The standard deviation is not a robust statistic; it is influenced by outliers.)

MAD estimates the scale parameter by using 1.4826 times the median absolute deviation from the median of the data.

Sn estimates the scale parameter by using a specified constant times the robust statistic S_n of the data.

Qn estimates the scale parameter by using a specified constant times the robust statistic Q_n of the data.

Interquartile range estimates the scale parameter by using the interquartile range of the data divided by 1.34898.

Gini's mean difference estimates the scale parameter by using $\sqrt{\pi}/2$ times Gini's mean difference.

Scale multiplier
specifies the constant used to multiply the scale estimate. The resulting product, d, determines outliers: all values whose distance to the location estimate is greater than d are labeled as outliers.

Plots Tab

You can use the **Plots** tab (Figure 17.4) to create a histogram and a normal Q-Q plot of the chosen variable.

If you select **Overlay normal curve and outlier cutoff**, then the histogram includes an overlaid normal curve. (See Figure 17.5.) The parameters for the normal curve are the location and scale estimates of the data. A vertical reference line in the histogram indicates the location estimate, and shading indicates regions more than $c\hat{\sigma}$ units from the location estimate, where c is the scale multiplier and $\hat{\sigma}$ is the scale estimate.

Output Variables Tab

You can use the **Output Variables** tab to add an indicator variable to the data table. The indicator variable is named Outlier_*Y*, where *Y* is the name of the chosen variable. The indicator variable is 1 for observations that are classified as outliers.

Roles Tab

You can use the **Roles** tab to specify a frequency variable for the analysis. A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents *n* observations, where *n* is the value of the frequency variable.

Analysis of Selected Variables

If an interval variable is selected in a data table when you run the analysis, then that variable is automatically entered in the **Y Variable** field of the **Variables** tab.

If any variable in the data table has a Frequency role, it is automatically entered in the **Frequency Variable** field of the **Roles** tab.

Chapter 18
Data Smoothing: Loess

Contents

Overview of the Loess Analysis **263**
Example: Fit a Loess Curve to Data **264**
Example: Compare Smoothers **268**
Removing a Smoother 272
Specifying the Loess Analysis **274**
Variables Tab 274
Method Tab 274
Plots Tab 275
Tables Tab 276
Output Variables Tab 276
Analysis of Selected Variables **277**
References **277**

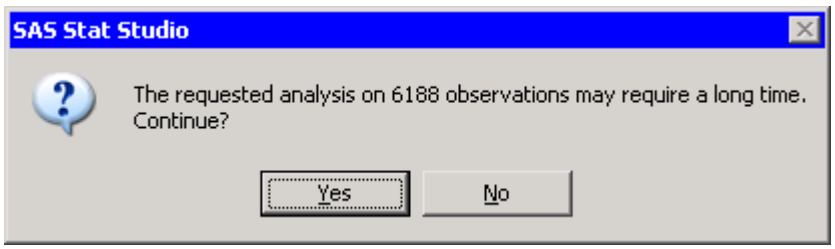
Overview of the Loess Analysis

The Loess analysis is intended for scatter plot smoothing. Given bivariate data $(x_i, y_i), i = 1, \ldots, n$, the Loess analysis fits a regression function f whose value at a point x is obtained by evaluating a *local regression function* at that point. This function is constructed based on data within a neighborhood of x. Although the fit in each local neighborhood is parametric, the construction of the function f depends on many neighborhoods. Consequently, the resulting function is nonparametric.

You can run a Loess analysis by selecting **Analysis ►Data Smoothing ►Loess** from the main menu. The computation of the loess regression function, confidence limits, and related statistics is implemented by calling the LOESS procedure in SAS/STAT software. See the LOESS procedure documentation in the *SAS/STAT User's Guide* for additional details.

NOTE: Fitting a loess curve to data sets with more than several thousand observations might require you to wait a while for the computation to finish, especially if you are computing confidence limits or performing an exhaustive search to find the optimal value of the smoothing parameter. Because of this, the Loess analysis presents a warning message (shown in Figure 18.1) when your data contain more than 5,000 observations. A similar warning appears if you are performing an exhaustive search and there are more than 1,000 observations.

Figure 18.1 A Warning Message



Example: Fit a Loess Curve to Data

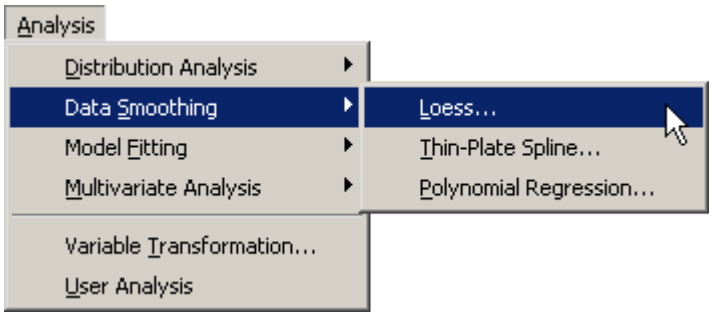
In this example, you fit a loess curve to data in the miningx data set. The miningx data set contains 80 observations that correspond to a single test hole in the mining data set. The driltime variable is the time that is required to drill the last five feet of the current depth, in minutes; the current depth is recorded in the depth variable.

To fit a loess curve:

1 Open the miningx data set.

2 Select **Analysis ►Data Smoothing ►Loess** from the main menu, as shown in Figure 18.2.

Figure 18.2 Selecting the Loess Analysis

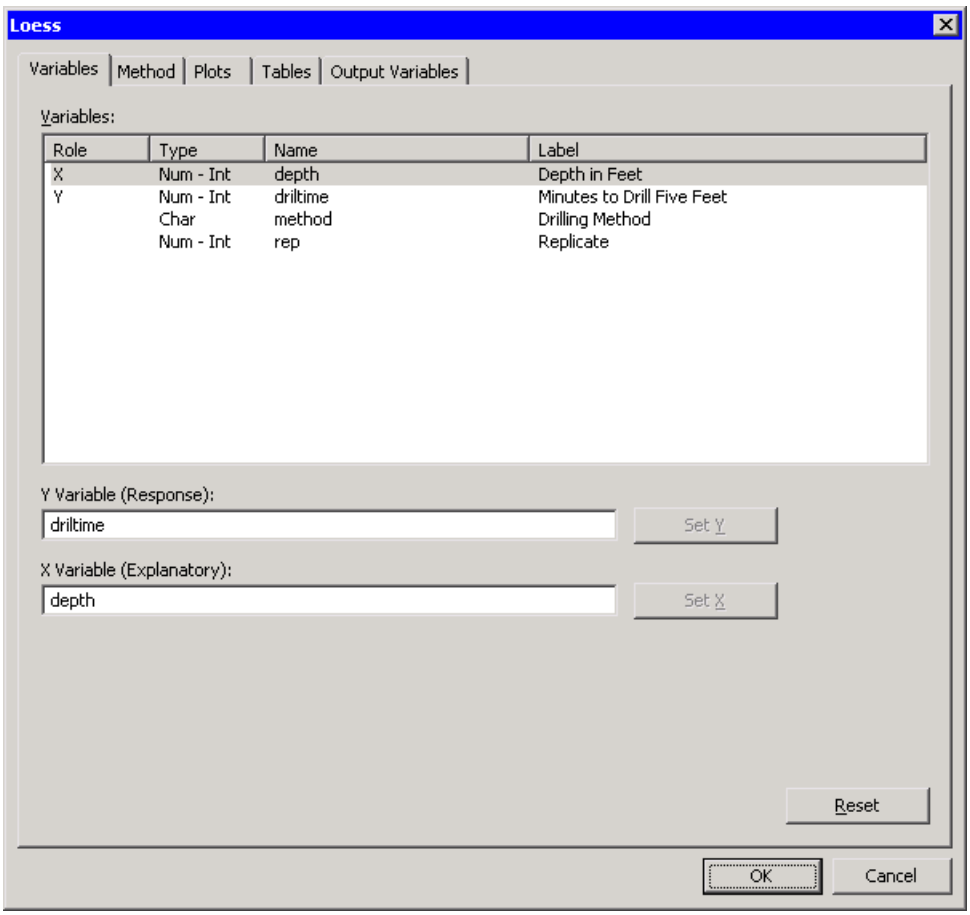


The Loess dialog box appears. You can select variables for the analysis by using the **Variables** tab, shown in Figure 18.3.

3 Select the variable driltime, and click **Set Y**.

4 Select the variable depth and click **Set X**.

Figure 18.3 Selecting Variables



5 Click the **Plots** tab.

The **Plots** tab becomes active. (See Figure 18.4.) You can use this tab to request additional plots.

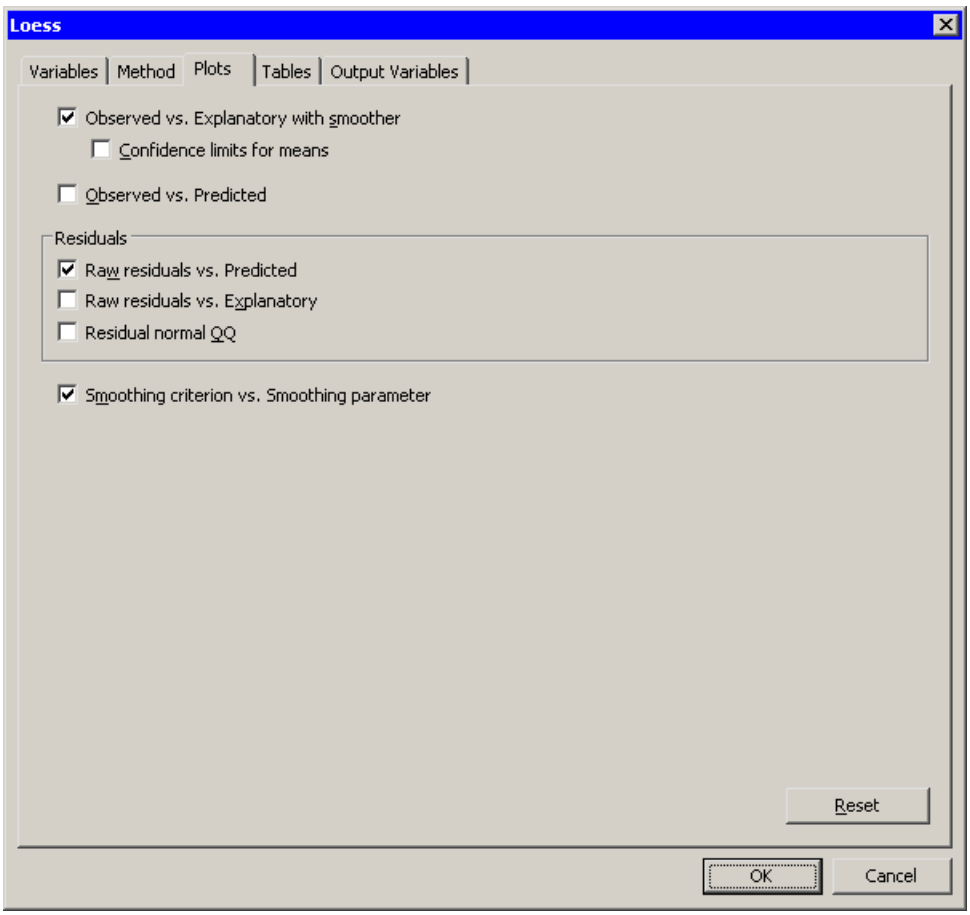
6 Select **Raw residuals vs. Explanatory**.

For this example it is useful to request a plot of the smoothing criterion versus the smoothing parameter. The loess smoothing parameter determines the percentage of observations used to fit a weighted regression in each local neighborhood. Small values of the smoothing parameter often correspond to undersmoothed curves with many undulations; large values correspond to oversmoothed curves with few undulations. The parameter value that minimizes the smoothing criterion represents a compromise between model fit and model complexity.

7 Select **Smoothing criterion vs. Smoothing parameter**.

8 Click **OK**.

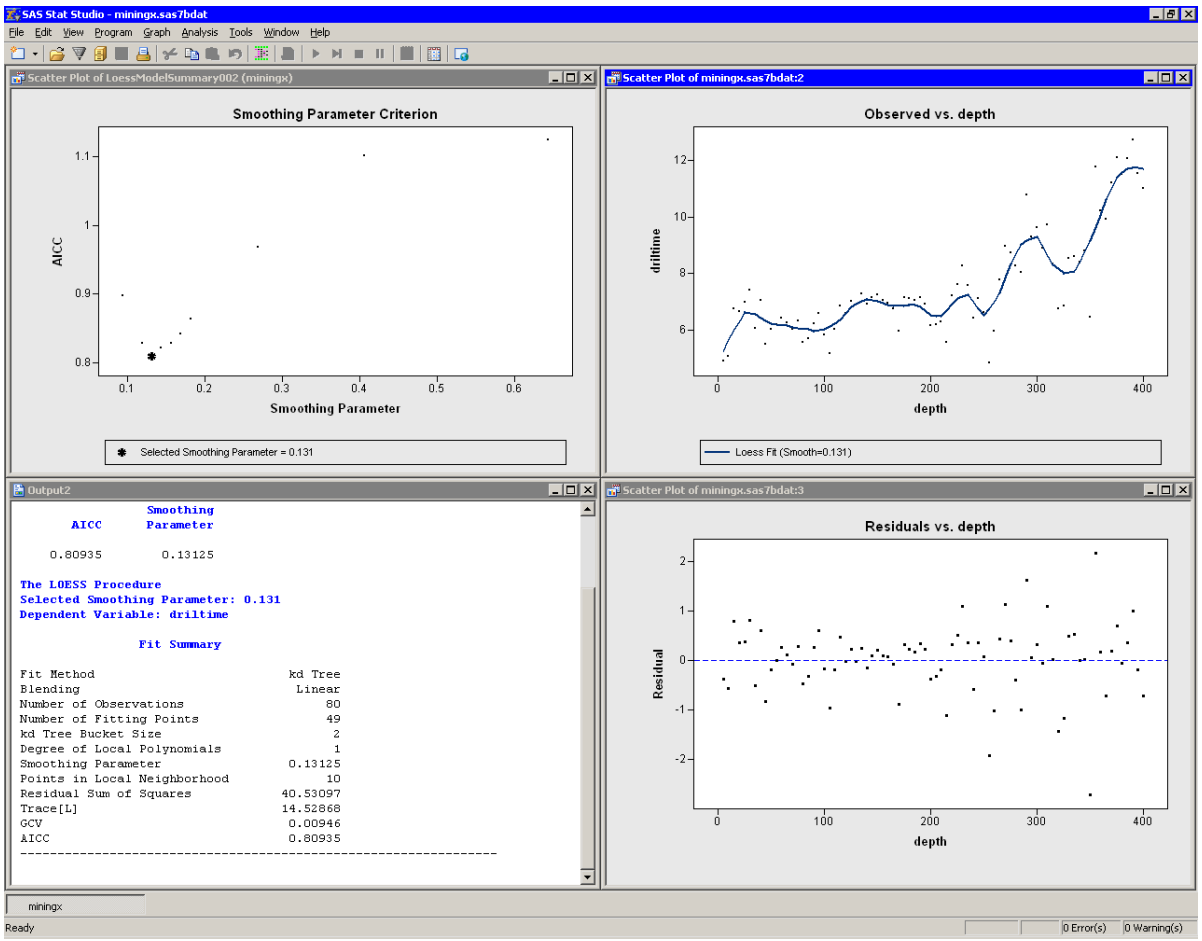
Figure 18.4 Selecting Plots



The Loess analysis calls the LOESS procedure with the options specified in the dialog box. The procedure displays two tables in the output document, as shown in Figure 18.5. The first table shows that the minimum value of the bias-corrected Akaike's information criterion (AICC) was achieved for a smoothing parameter of 0.13125. The second table summarizes the options used by the LOESS procedure and also summarizes the loess fit.

Three plots are created. Some plots might be hidden beneath others. If so, move the plots so that the workspace looks like Figure 18.5.

Figure 18.5 Output from a Loess Analysis



One plot (upper left in Figure 18.5) shows the AICC for each value of the smoothing parameter evaluated by the LOESS procedure. Note that the selected smoothing parameter is the one that minimizes the AICC.

A second plot (upper right in Figure 18.5) shows a scatter plot of driltime versus depth, with a loess smoother overlaid. The undulations in the smoother might correspond to depths at which variations in rock hardness affect the drilling time. In particular, it is known that the decrease in drilling time at 250 feet is due to encountering a layer of soft copper-nickel ore (Penner and Watts 1991).

The third plot (lower right in Figure 18.5) shows the residuals versus depth. The spread of the residuals suggests that the variance of the drilling time is a function of the depth of the hole being drilled.

The next example creates a second curve that smooths out some of the undulations. This is accomplished by restricting the smoothing parameter to relatively large values. Specifically, the next example specifies that at least 50% of the points in the data set should be used for each local weighted regression.

Example: Compare Smoothers

The "Details" section of the LOESS procedure documentation describes how the LOESS procedure computes predicted values. The predicted value at a point x is determined by a weighted average of observations near x. The number of observations used to form the predicted value depends on the smoothing parameter.

Recall that the response variable in the previous example is the length of time required to drill the last five feet of a hole that is depth feet deep. For these data, the optimal smoothing parameter was approximately 0.131. This value results in a smoother that varies with the hardness of the underlying rock strata.

However, you might want to average out the variations in rock hardness to get a better indication of how the drilling time varies with depth. While 0.131 is a *global minimum* of the AICC function, there might be a *local minimum* at a larger value of the smoothing parameter. Using a larger value results in a smoother that is less sensitive to local variation in rock hardness.

This example computes another possible loess fit and compares it to the smoother with the parameter 0.131. The example assumes you have completed the previous example and your workspace looks like Figure 18.5.

Recall that SAS/IML Studio adds a smoother to an *existing* scatter plot when both of the following conditions are satisfied:

- The scatter plot is the active window when you select the analysis.
- The scatter plot variables match the analysis variables.

To compute a second loess fit and compare the two models:

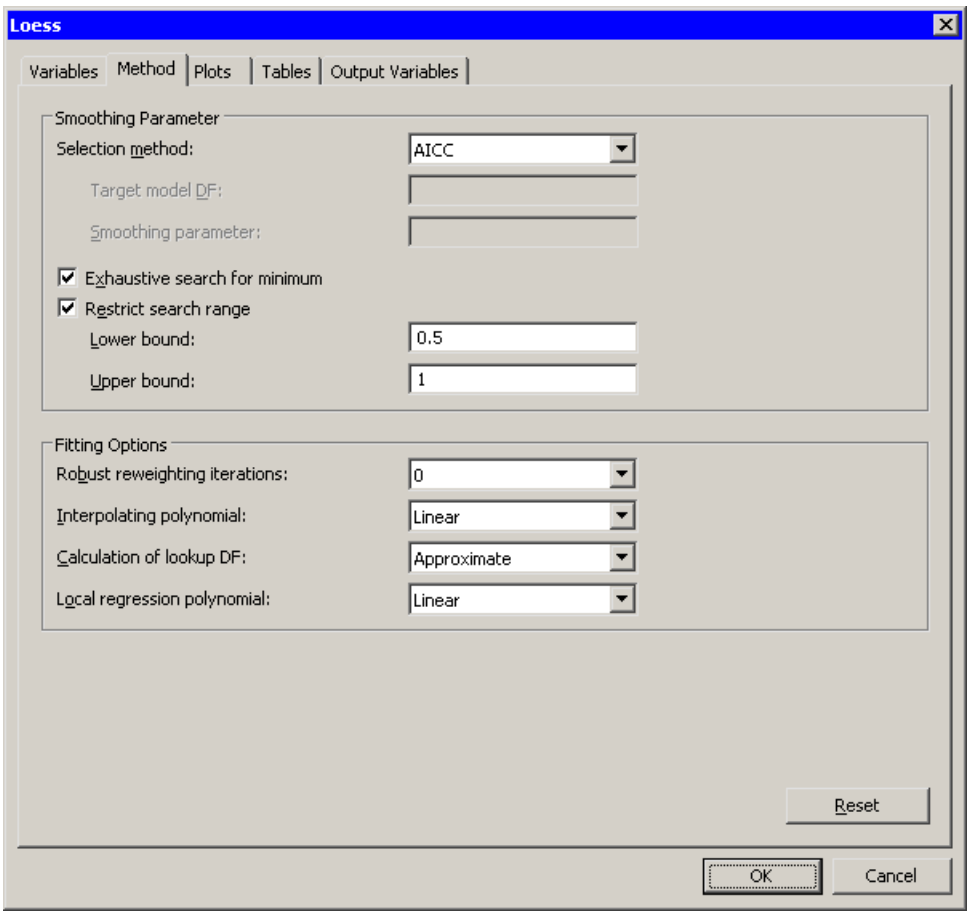
1. Click the scatter plot of driltime versus depth to activate that window.
2. Select **Analysis ►Data Smoothing ►Loess** from the main menu.

   The loess dialog box appears. The dialog box remembers the variables you used in the last analysis.
3. Make sure that driltime is selected as the Y variable and depth is selected as the X variable.

   By examining the AICC plot from the previous example (upper left in Figure 18.5), you might guess that the AICC is an increasing function of the smoothing parameter on the interval [0.131, 0.5]. Thus, if there is a local minimum for AICC at a larger value of the smoothing parameter, it must occur in the interval [0.5, 1]. In the following steps you search for a local minimum of AICC restricted to this interval.
4. Click the **Method** tab.

   The **Method** tab is activated, as shown in Figure 18.6.

Figure 18.6 The Method Tab



5 Click **Exhaustive search for minimum**.

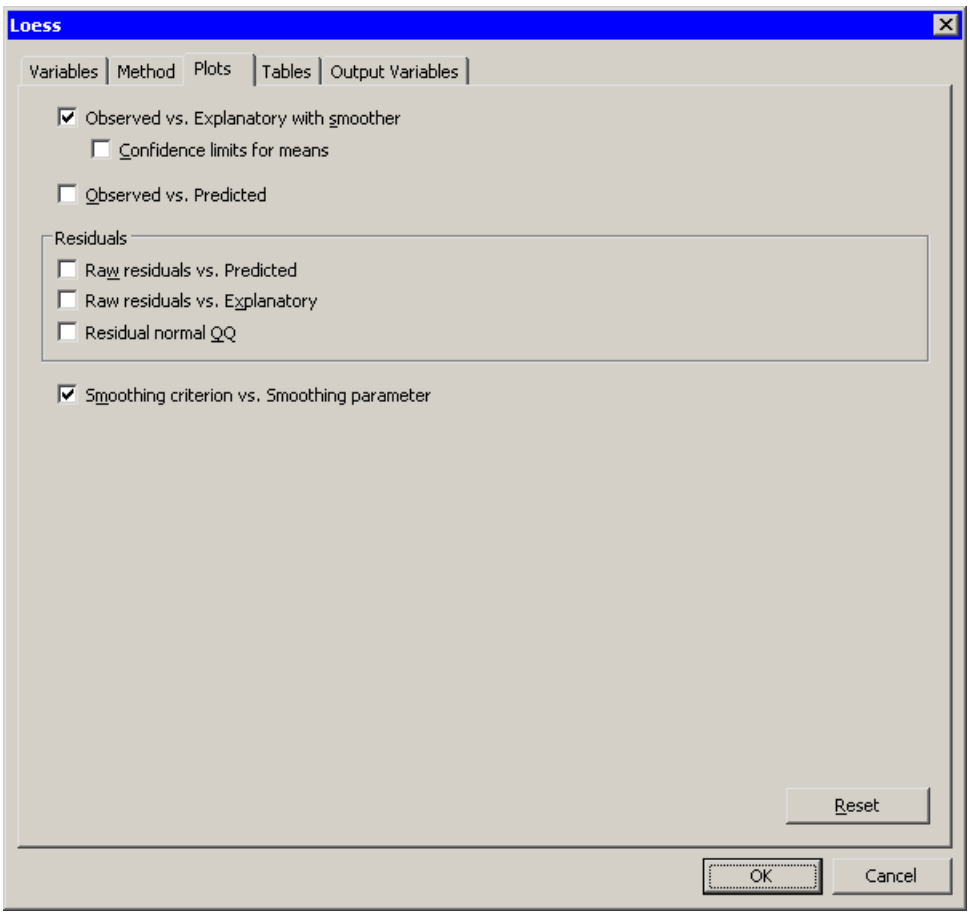
6 Click **Restrict search range** and type `0.5` for the **Lower bound**.

NOTE: The **Exhaustive search for minimum** option is computationally expensive. It corresponds to the GLOBAL modifier of the SELECT= option in the LOESS MODEL statement. For the current example, which has 80 observations, the option results in evaluating loess models with at least 40 (0.5×80) points in the local neighborhoods. Thus, this option causes the LOESS procedure to evaluate many separate models: one with 40 points in the local neighborhoods, one with 41 points, and so on, up to 80 points. For a data set with 10,000 observations, the same options would result in evaluating up to 5,000 models.

7 Click the **Plots** tab.

The **Plots** tab is activated, as shown in Figure 18.7.

Figure 18.7 Selecting Plots

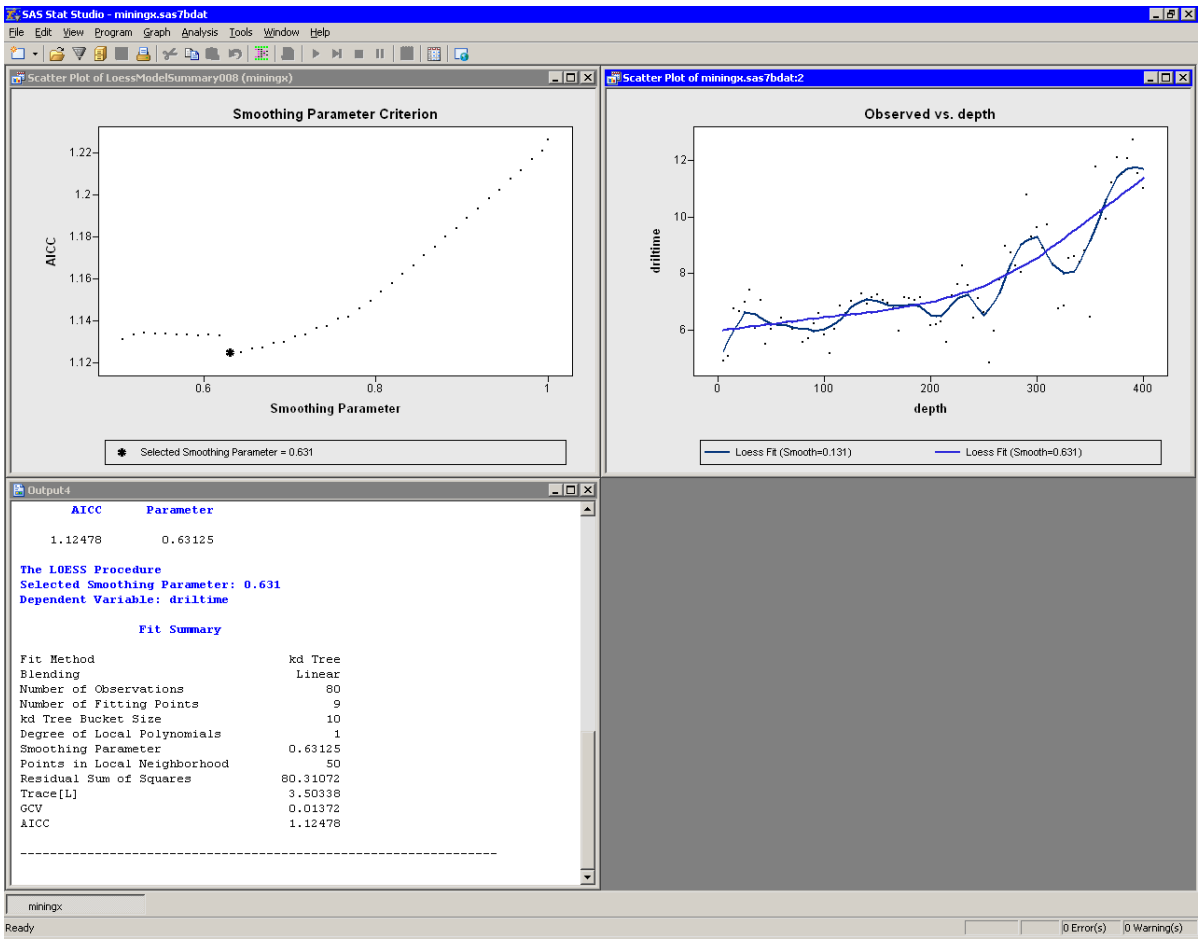


8 Clear **Raw residuals vs. Explanatory**.

9 Click **OK**.

As shown in Figure 18.8, the scatter plot of driltime versus depth updates to display the new loess smoother. The AICC plot now shows that the chosen smoothing parameter is approximately 0.631, which corresponds to using 50 ($\approx 0.631 \times 80$) points in the local neighborhoods.

Figure 18.8 Example: Rerun a Loess Analysis



NOTE: This second Loess analysis creates a predicted value variable named LoessP_driltime. This variable overwrites the variable of the same name that was created by the first Loess analysis. If you want to compare the predicted values for these two models, you need to rename the first variable prior to running the second analysis.

Removing a Smoother

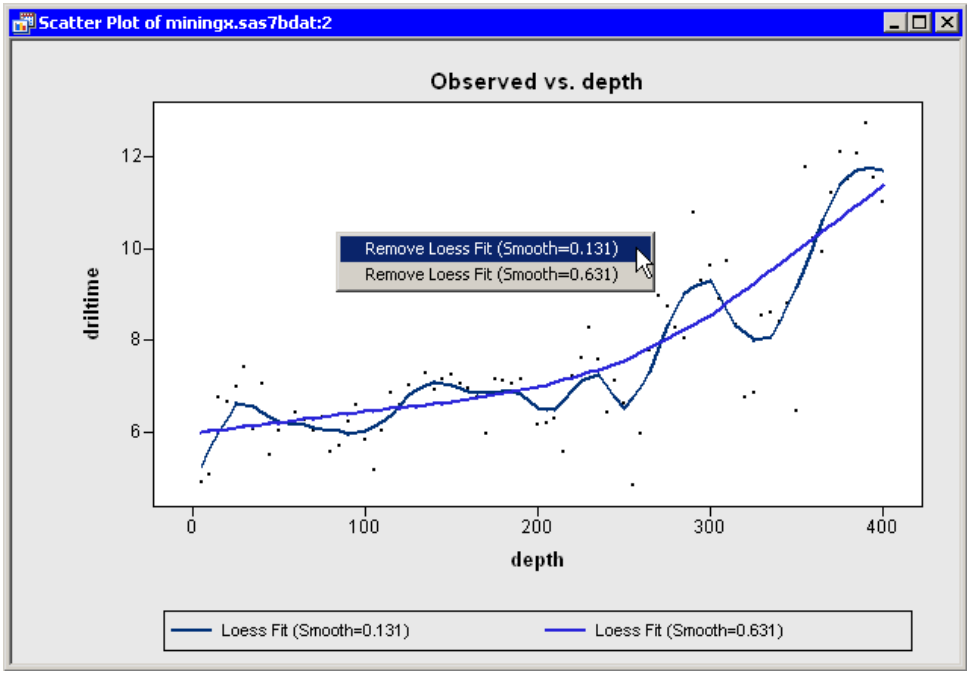
If you are trying to determine the relationship between drilling time and depth while averaging out variations in the rock strata, you might prefer the second smoother to the first. If so, you might want to remove the first smoother.

When SAS/IML Studio adds a smoother, it also adds an *action menu* to remove that smoother. The following steps access this menu by pressing the F11 key while the plot is active.

1 Activate the scatter plot of driltime versus depth and press the F11 key.

An action menu appears, as shown in Figure 18.9.

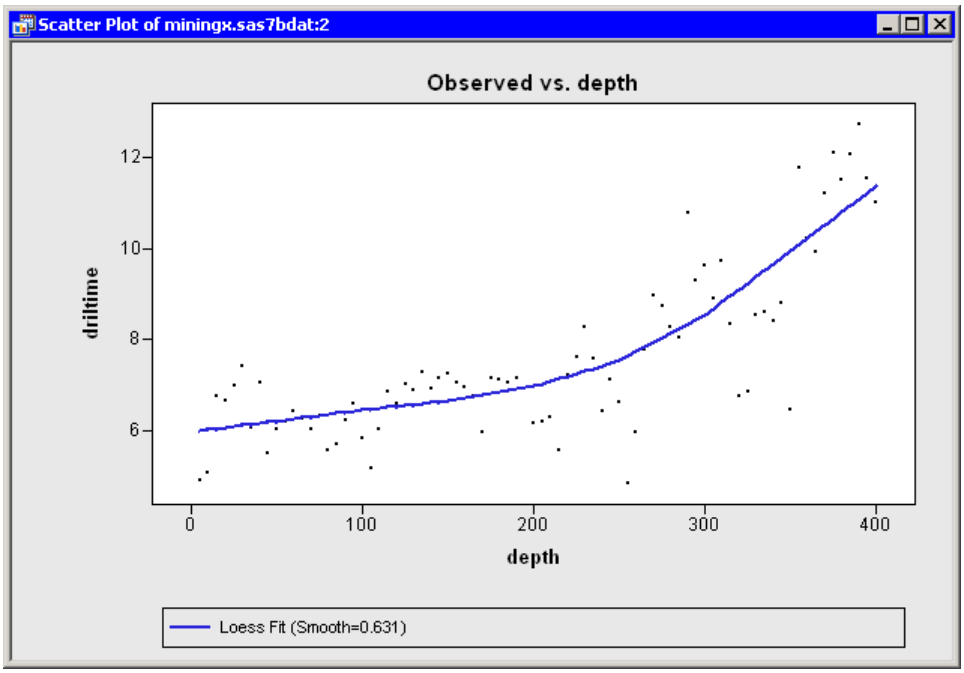
Figure 18.9 Removing a Smoother



2 Select the first menu item: **Remove Loess Fit (Smooth=0.131)**.

The first smoother vanishes. The plot now looks like the one in Figure 18.10.

Figure 18.10 A Single Smoother



The new loess smoother indicates that the drilling time varies roughly linearly at depths between 0 and 200 feet, and linearly (with a different slope) at depths greater than 300 feet. Between 200 and 300 feet, the response varies nonlinearly. Penner and Watts (1991) suggest that air forced through the drill shaft is able to expel debris from the hole at depths less than 200 feet, but at greater depths more and more debris falls back into the hole, thus reducing the drill's efficiency.

You can use the techniques presented in this example to compare the loess model to other smoothers. For example, you might decide to compare the loess curve to a quadratic polynomial. If the predictions are nearly the same, you might favor the simpler model.

Specifying the Loess Analysis

This section describes the Loess analysis dialog box options. The Loess analysis calls the LOESS procedure in SAS/STAT software. See the LOESS procedure documentation in the *SAS/STAT User's Guide* for details.

Variables Tab

You can use the **Variables** tab to specify the response and explanatory variables for the LOESS MODEL statement. The Y variable specifies the dependent (response) variable, and the X variable specifies the independent (explanatory) variable. The SAS/IML Studio Loess analysis supports only a single dependent variable and a single smoothing variable.

Method Tab

You can use the **Method** tab to specify options for the loess algorithm. The **Method** tab contains the following UI controls:

Selection method
specifies how to choose the loess smoothing parameter. This option corresponds to the SELECT= option in the MODEL statement.

AICC
selects the smoothing parameter that minimizes the corrected Akaike information criterion.

GCV
selects the smoothing parameter that minimizes the generalized cross validation criterion.

Approx. model DF
selects the smoothing parameter for which the trace of the prediction matrix is closest to the **Target model DF**. This corresponds to the SELECT=DF1 option in the MODEL statement.

Manual
enables you to specify the value in the **Smoothing parameter** field.

Exhaustive search for minimum
specifies that a global minimum be found within the range of smoothing parameter values examined. This corresponds to the GLOBAL modifier to the SELECT= option in the MODEL statement. This option is computationally expensive.

Restrict search range
specifies that only smoothing parameters greater than or equal to **Lower bound** and less than or equal to **Upper bound** be examined.

Robust reweighting iterations
specifies the number of iterative reweighting steps. SAS/IML Studio counts the initial fit as the 0th *reweighting* iteration. This differs from the LOESS procedure, which counts the initial fit as the first iteration. Thus if you type n in this field, the option corresponds to ITERATIONS=$n + 1$ in the MODEL statement.

Interpolating polynomial
specifies whether the interpolating polynomial is linear or cubic. This corresponds to the INTERP= option in the MODEL statement.

Calculation of lookup DF
specifies the method to use for calculating the "lookup" degrees of freedom used in performing statistical inference. This corresponds to the DFMETHOD= option in the MODEL statement.

Local regression polynomial
specifies the degree of the local polynomial to use for each local regression. The choice is linear or quadratic. This corresponds to the DEGREE= option in the MODEL statement.

Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the Loess analysis. (See Figure 18.4.) The raw residuals are computed as $Y - \hat{Y}$, where $\hat{Y}$ indicates the variable that contains the predicted values of the response.

Creating a plot often adds one or more variables to the data table. The following plots are available:

Observed vs. Explanatory with smoother
creates a scatter plot of the X and Y variables, overlaid with a smoother.

Confidence limits for means
specifies whether to add 95% upper and lower confidence limits to the Observed vs. Explanatory plot.

Observed vs. Predicted
creates a scatter plot of the Y variable versus the predicted values.

Raw residuals vs. Predicted
creates a scatter plot of the residuals versus the predicted values.

Raw residuals vs. Explanatory
creates a scatter plot of the residuals versus the X variable.

Residual normal QQ
creates a normal Q-Q plot of the residuals.

Smoothing criterion vs. Smoothing parameter
creates a scatter plot of the smoothing criterion (for example, AICC) versus the smoothing parameter value for all smoothing parameter values examined in the selection process. The value that minimizes the criterion is indicated by a star-shaped marker.

NOTE: SAS/IML Studio adds a smoother to an *existing* scatter plot when both of the following conditions are satisfied:

- The scatter plot is the active window when you select the analysis.
- The scatter plot variables match the analysis variables.

Tables Tab

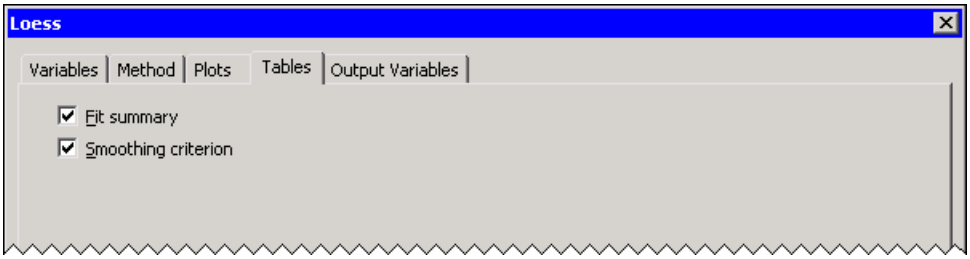
You can use the **Tables** tab to display tables that summarize the results of the analysis.

The **Tables** tab is shown in Figure 18.11. The following tables are available:

Fit summary
summarizes the fit and the fit parameters.

Smoothing criterion
displays the selected smoothing parameter and the corresponding criterion value.

Figure 18.11 The Tables Tab



Output Variables Tab

You can use the **Output Variables** tab to add analysis variables to the data table. (See Figure 18.12.) If you request a plot that uses one of the output variables, then that variable is automatically created even if you did not explicitly select the variable on the **Output Variables** tab.

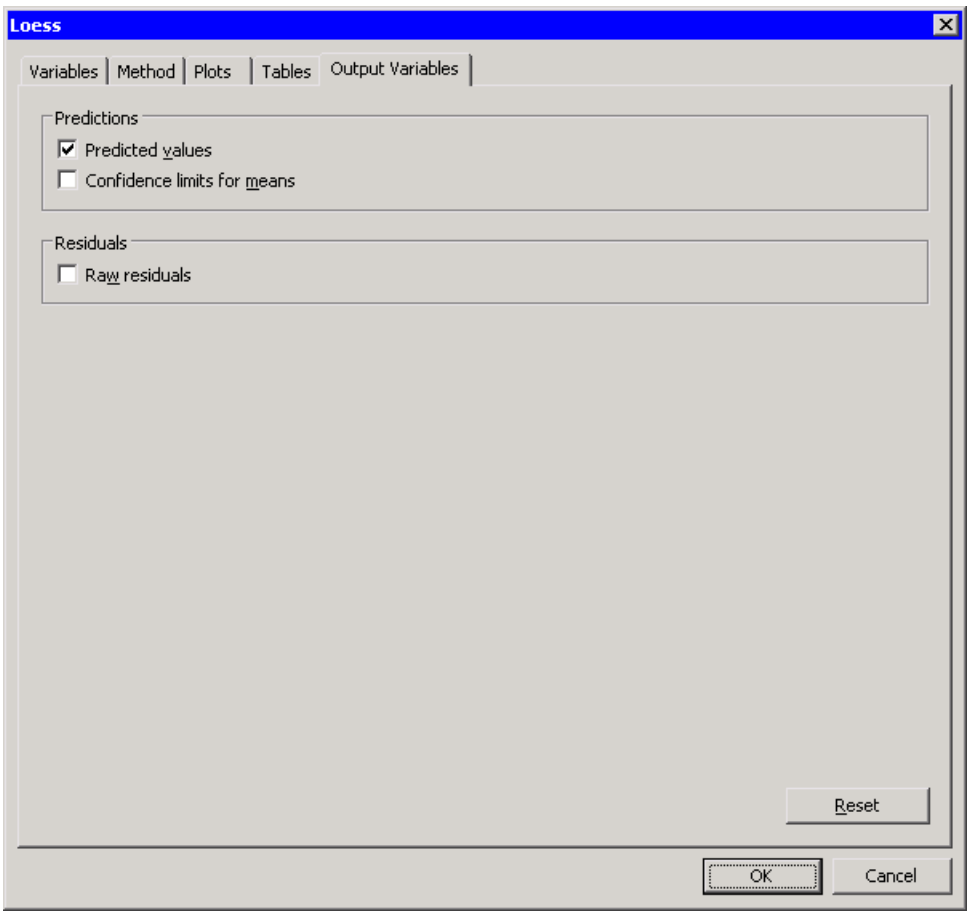
The following list describes each output variable and indicates how it is named. Y represents the name of the response variable.

Predicted values
adds predicted values. The variable is named LoessP_Y.

Confidence limits for means
adds 95% confidence limits for the expected value (mean). The variables are named LoessLclm_Y and LoessUclm_Y.

Raw residuals
adds residuals, which are calculated as observed values minus predicted values. The variable is named LoessR_Y.

Figure 18.12 The Output Variables Tab



Analysis of Selected Variables

If one or more interval variables are selected in a data table when you run the analysis, then the following occur:

- The first selected interval variable is automatically entered in the **Y Variable** field of the **Variables** tab.
- The second selected interval variable is automatically entered in the **X Variable** field.

No role variables are used for this analysis.

References

Penner, R. and Watts, D. G. (1991), "Mining Information," *American Statistician*, 45, 4–9.

Chapter 19
Data Smoothing: Thin-Plate Spline

Contents

Overview of the Thin-Plate Spline Analysis **279**
Example: Fit a Thin-Plate Spline Curve to Data **279**
Specifying the Thin-Plate Spline Analysis **283**
Variables Tab 283
Method Tab 283
Plots Tab 284
Tables Tab 285
Output Variables Tab 286
Analysis of Selected Variables **287**

Overview of the Thin-Plate Spline Analysis

The Thin-Plate Spline analysis is intended for scatter plot smoothing. The Thin-Plate Spline analysis uses a penalized least squares method to fit a nonparametric regression model. You can use the generalized cross validation (GCV) function to select the amount of smoothing.

You can run the Thin-Plate Spline analysis by selecting **Analysis ►Data Smoothing ►Thin-Plate Spline** from the main menu. The computation of the fitted spline function, confidence limits, and related statistics is implemented by calling the TPSPLINE procedure in SAS/STAT software. See the TPSPLINE procedure documentation in the *SAS/STAT User's Guide* for additional details.

Example: Fit a Thin-Plate Spline Curve to Data

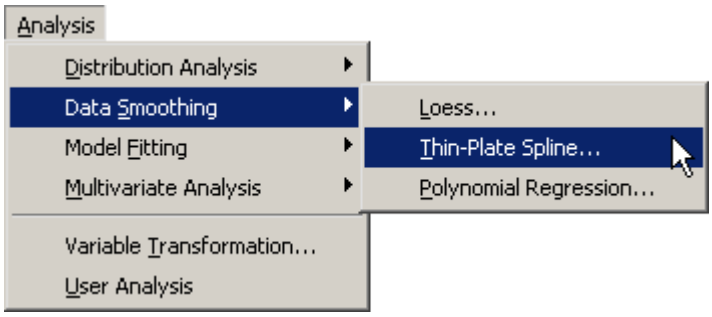
In this example, you fit a thin-plate spline curve to data in the miningx data set. These data are discussed in Chapter 18, "Data Smoothing: Loess." The miningx data set contains 80 observations that correspond to a single test hole in the mining data set. The driltime variable is the time that is required to drill the last five feet of the current depth, in minutes; the hole depth is recorded in the depth variable.

To fit a thin-plate spline curve:

1 Open the miningx data set.

2 Select **Analysis ►Data Smoothing ►Thin-Plate Spline** from the main menu, as shown in Figure 19.1.

Figure 19.1 Selecting the Thin-Plate Spline Analysis

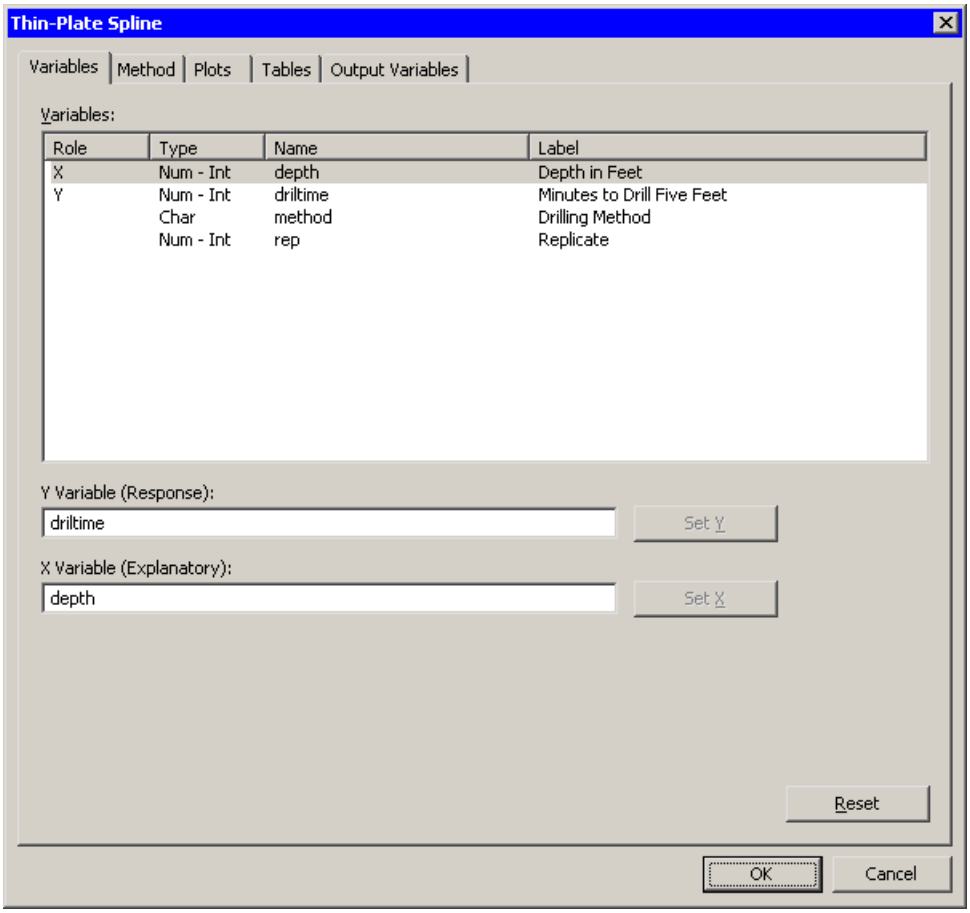


The Thin-Plate Spline dialog box appears. You can select variables for the analysis by using the **Variables** tab, shown in Figure 19.2.

3 Select the variable driltime, and click **Set Y**.

4 Select the variable depth, and click **Set X**.

Figure 19.2 Selecting Variables



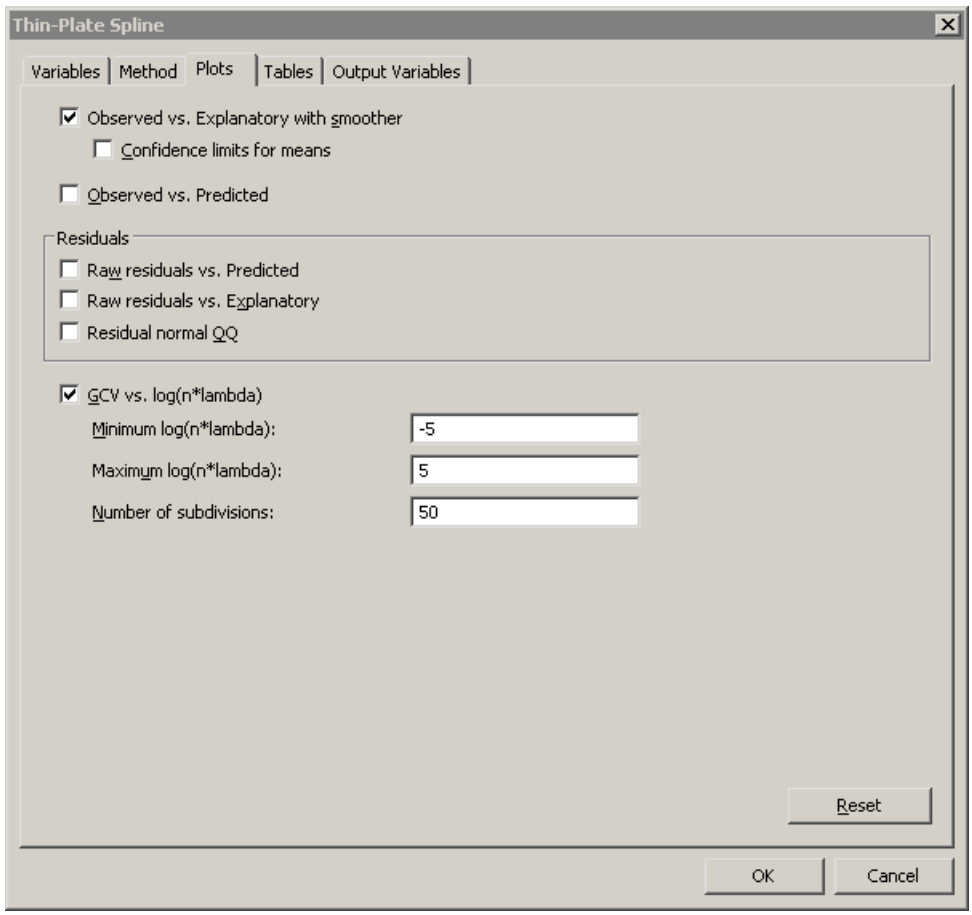
5 Click the **Plots** tab.

The **Plots** tab (Figure 19.3) becomes active. By default, the analysis creates a scatter plot of Y versus X with the smoother overlaid. The smoothing penalty parameter is chosen to minimize the generalized cross validation (GCV) criterion.

6 Select **GCV vs. log(n*lambda)** to visualize how the smoothing parameter affects the GCV criterion.

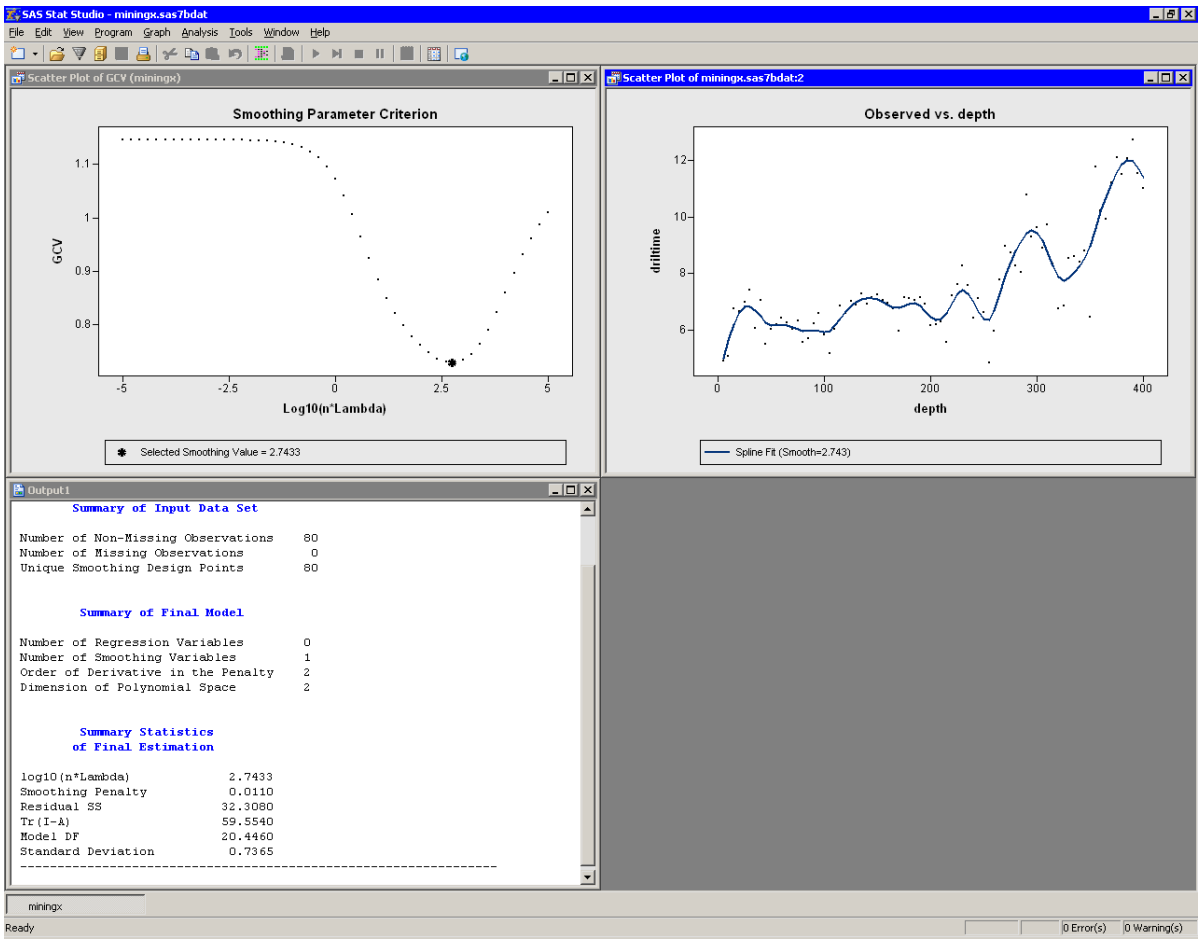
7 Click **OK**.

Figure 19.3 Selecting Plots



The Thin-Plate Spline analysis calls the TPSPLINE procedure with the options specified in the dialog box. The procedure displays three tables in the output document, as shown in Figure 19.4. The first table shows information about the number of observations. The second table summarizes model options used by the TPSPLINE procedure. The third table summarizes the fit, including the smoothing value (2.7433) that is chosen to optimize the selection criterion.

Figure 19.4 Output from a Loess Analysis



Two plots are created, as shown in Figure 19.4. The upper left plot in Figure 19.4 shows the GCV criterion for a range of smoothing parameter values. Note that the selected smoothing parameter (2.7433) is the one that minimizes the GCV. A second plot overlays a scatter plot of drilltime versus depth with a thin-plate smoother. As discussed in Chapter 18, "Data Smoothing: Loess," the undulations in the smoother correspond to geological variations in the rock strata. Chapter 18, "Data Smoothing: Loess," also discusses how to display multiple smoothers in a single scatter plot and how to remove smoothers from a scatter plot.

Specifying the Thin-Plate Spline Analysis

This section describes the Thin-Plate Spline analysis dialog box options. The Thin-Plate Spline analysis calls the TPSPLINE procedure in SAS/STAT software. See the TPSPLINE procedure documentation in the *SAS/STAT User's Guide* for details.

Variables Tab

You can use the **Variables** tab to specify the response and explanatory variables for the TPSPLINE MODEL statement. The Y variable specifies the dependent (response) variable, and the X variable specifies the independent (smoothing) variable. The Thin-Plate Spline analysis supports a single dependent variable and a single smoothing variable. Semiparametric fits are not supported: if your data have a polynomial trend, you should subtract the trend and use thin-plate splines to model the residuals.

Method Tab

You can use the **Method** tab to specify options for the thin-plate spline algorithm. (See Figure 19.5.) The **Method** tab contains the following UI controls:

Selection method
specifies how to choose the smoothing penalty parameter. This option corresponds to the SELECT= option in the MODEL statement.

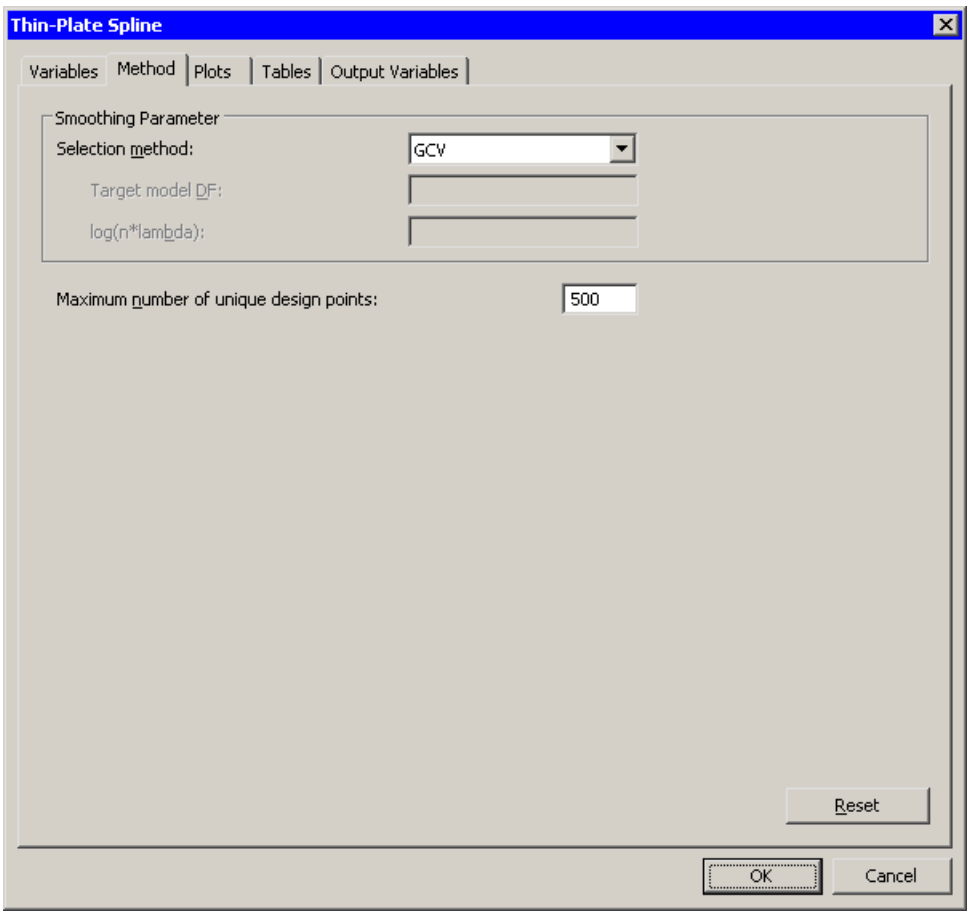
GCV
selects the smoothing parameter that minimizes the generalized cross validation criterion.

Approx. model DF
selects the smoothing parameter for which the trace of the prediction matrix is closest to the **Target model DF**. This corresponds to SELECT=DF option.

Manual
enables you to specify the value in the **log(n*lambda)** field. This corresponds to the LOGNLAMBDA= option.

Maximum number of unique design points
specifies a limit on the number of unique design points, N_x, in the model. This option corresponds to the DISTANCE= option in the MODEL statement in the following way: the value in this field is used to compute a value for the DISTANCE= option so that there are at most N_x design points. This option is useful for large data sets, since the TPSPLINE procedure is computationally expensive.

Figure 19.5 The Method Tab



Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the Thin-Plate Spline analysis. (See Figure 19.3.) The raw residuals are computed as $Y - \hat{Y}$, where $\hat{Y}$ indicates the variable that contains the predicted values of the response.

Creating a plot often adds one or more variables to the data table. The following plots are available:

Observed vs. Explanatory with smoother
creates a scatter plot of the X and Y variables, overlaid with a smoother.

Confidence limits for means
specifies whether to add 95% upper and lower confidence curves to the Observed vs. Explanatory plot. The meaning of the curves is described in the section "Computational Formulas" in the TPSPLINE documentation.

Observed vs. Predicted
creates a scatter plot of the Y variable versus the predicted values.

Raw residuals vs. Predicted
creates a scatter plot of the residuals versus the predicted values.

Raw residuals vs. Explanatory
creates a scatter plot of the residuals versus the X variable.

Residual normal QQ
creates a normal Q-Q plot of the residuals.

GCV vs. log(n*lambda)
creates a scatter plot of the GCV criterion versus the smoothing parameter value for a range of smoothing parameter values.

Minimum log(n*lambda)
specifies the minimum value of the smoothing parameter to consider.

Maximum log(n*lambda)
specifies the maximum value of the smoothing parameter to consider.

Number of subdivisions
specifies the number of smoothing parameters to consider. The value in this field is combined with the values in the previous two fields to form a list of values for the LOGNLAMBDA= option.

NOTE: SAS/IML Studio adds a smoother to an *existing* scatter plot when both of the following conditions are satisfied:

- The scatter plot is the active window when you select the analysis.
- The scatter plot variables match the analysis variables.

Chapter 18, "Data Smoothing: Loess," discusses how to display multiple smoothers in a single scatter plot and how to remove smoothers from a scatter plot.

Tables Tab

You can use the **Tables** tab to display tables that summarize the results of the analysis.

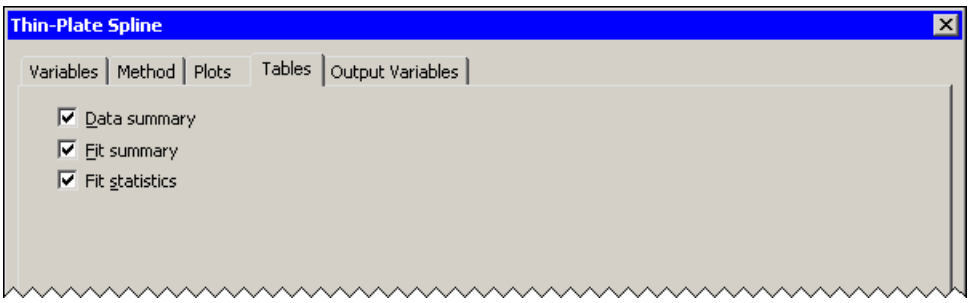
The **Tables** tab is shown in Figure 19.6. The following tables are available:

Data summary
summarizes information about the number of observations.

Fit summary
summarizes the model parameters.

Fit statistics
summarizes the fit, including the smoothing value that optimizes the selection criterion.

Figure 19.6 The Tables Tab



Output Variables Tab

You can use the **Output Variables** tab (Figure 19.7) to add analysis variables to the data table. If you request a plot that uses one of the output variables, then that variable is automatically created even if you did not explicitly select the variable on the **Output Variables** tab.

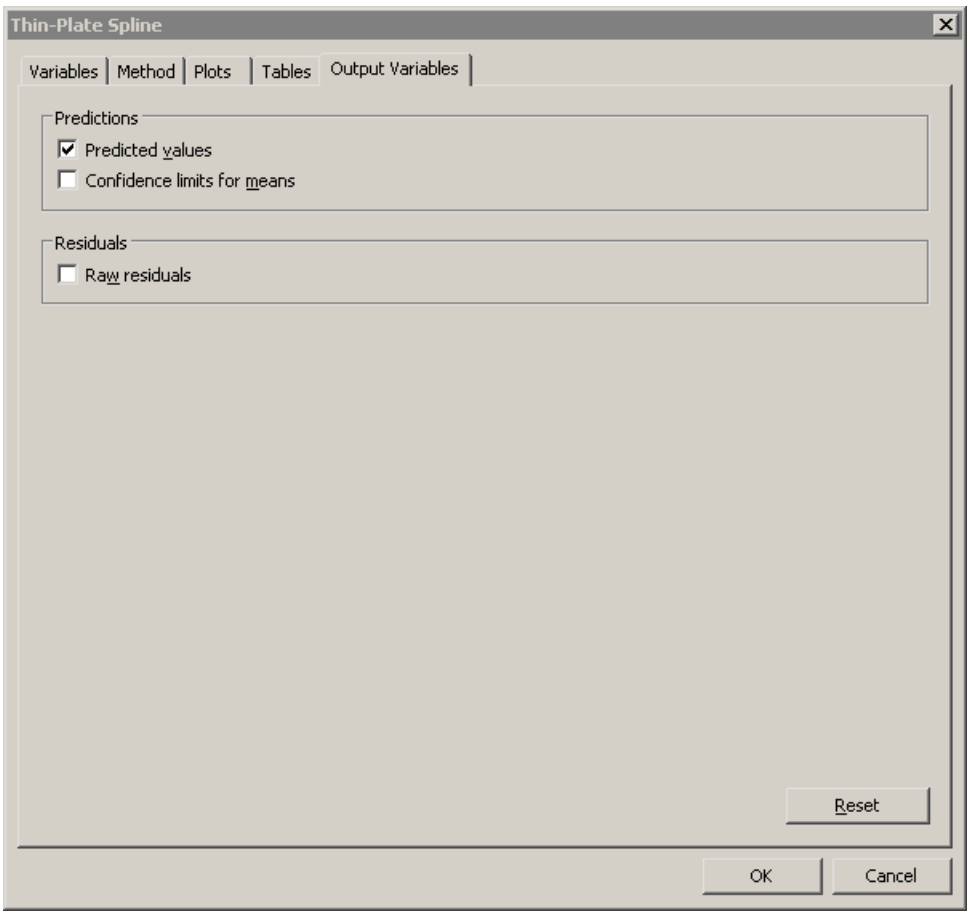
The following list describes each output variable and indicates how it is named. Y represents the name of the response variable.

Predicted values
adds predicted values. The variable is named TPSplP_Y.

Confidence limits for means
adds 95% confidence limits for the expected value (mean). The variables are named TPSplLclm_Y and TPSplUclm_Y.

Raw residuals
adds residuals, which are calculated as observed values minus predicted values. The variable is named TPSplR_Y.

Figure 19.7 The Output Variables Tab



Analysis of Selected Variables

If one or more interval variables are selected in a data table when you run the analysis, then the following occurs:

- The first selected interval variable is automatically entered in the **Y Variable** field of the **Variables** tab.
- The second selected interval variable is automatically entered in the **X Variable** field.

No role variables are used for this analysis.

Chapter 20
Data Smoothing: Polynomial Regression

Contents

Overview of the Polynomial Regression Analysis **289**
Example: Fit a Polynomial Curve to Data **289**
Specifying the Polynomial Regression Analysis **294**
Variables Tab 294
Method Tab 294
Plots Tab 295
Tables Tab 296
Output Variables Tab 298
Analysis of Selected Variables **300**

Overview of the Polynomial Regression Analysis

The Polynomial Regression analysis fits a low-order polynomial regression function to bivariate data by using ordinary least squares. This is a *global* parametric fit, whereas the other SAS/IML Studio smoothers are modern *local* nonparametric smoothers.

You can run a Polynomial Regression analysis by selecting **Analysis ►Data Smoothing ►Polynomial Regression** from the main menu. The computation of the regression function, confidence limits, and related statistics is implemented by calling the REG procedure in SAS/STAT software. See the documentation for the REG procedure in the *SAS/STAT User's Guide* for additional details.

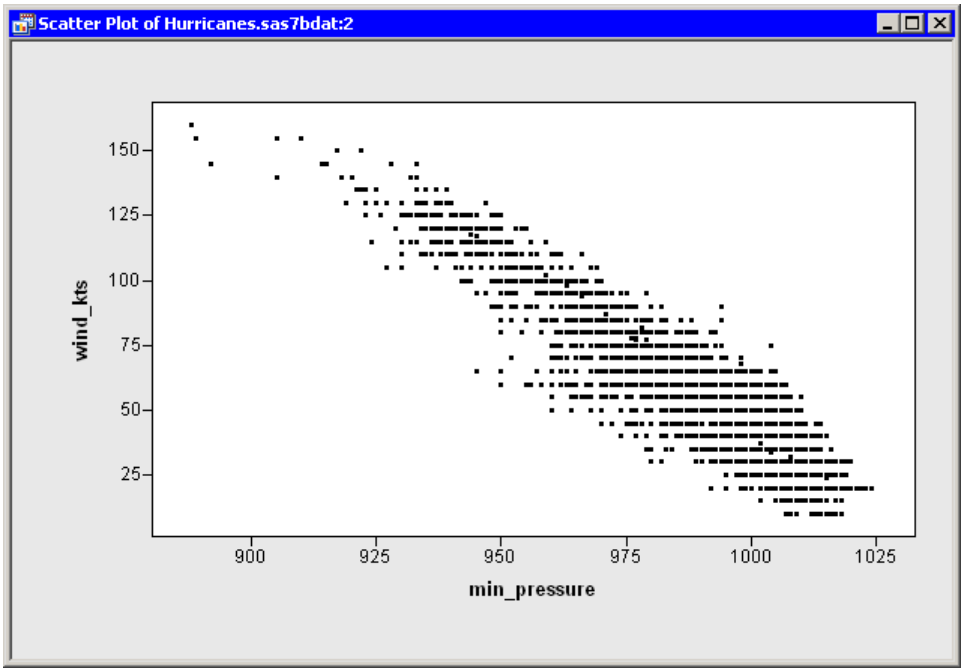
General multivariate regression is available by selecting **Analysis ►Model Fitting ►Linear Regression** from the main menu.

Example: Fit a Polynomial Curve to Data

In this example, you create a polynomial regression analysis of wind_kts as a function of min_pressure in the Hurricanes data set. The wind_kts variable is the wind speed in knots; the min_pressure variable is the minimum central pressure for each observation.

A scatter plot of these variables indicates that the relationship between these variables is approximately linear, as shown in Figure 20.1, so this example fits a line to the data.

Figure 20.1 Linearly Related Variables

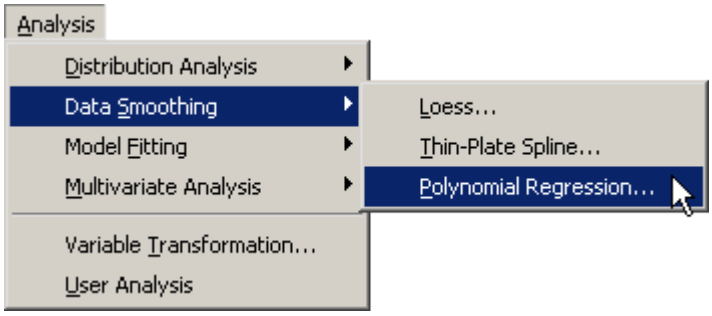


To create a polynomial regression analysis:

1 Open the Hurricanes data set.

2 Select **Analysis ►Data Smoothing ►Polynomial Regression** from the main menu, as shown in Figure 20.2.

Figure 20.2 Selecting Variables

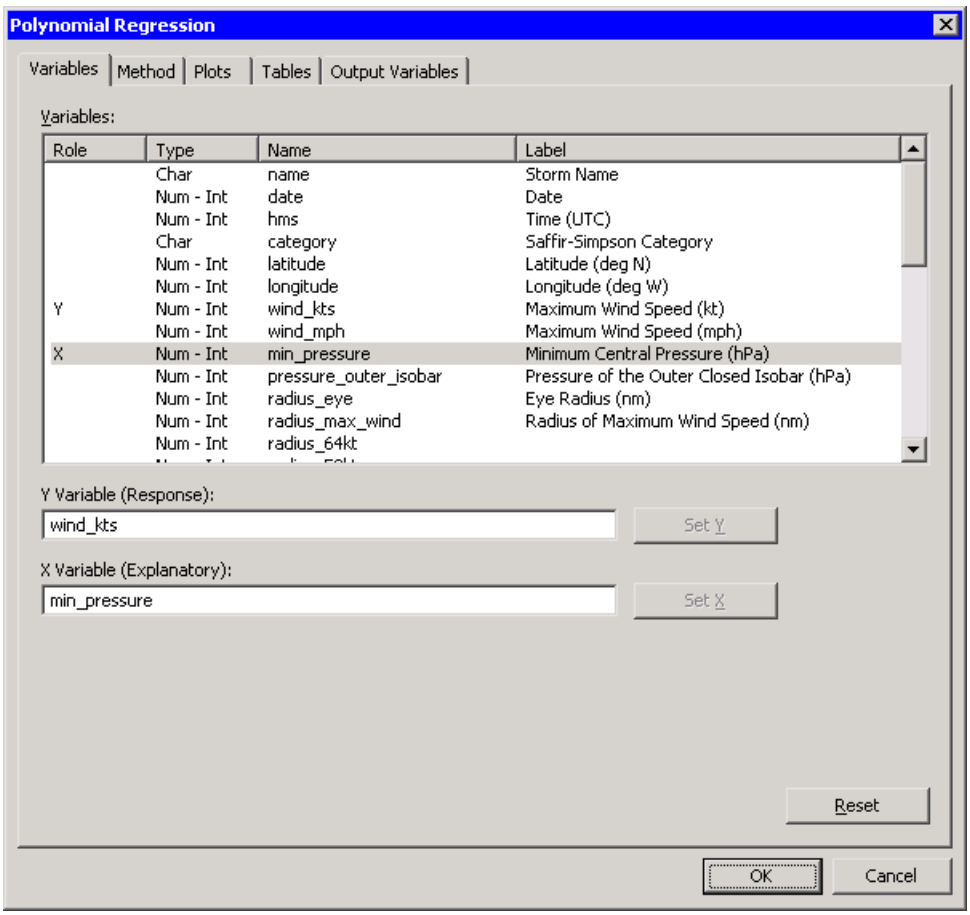


The Polynomial Regression dialog box appears. (See Figure 20.3.)

3 Select the variable wind_kts, and click **Set Y**.

4 Select the variable min_pressure, and click **Set X**.

Figure 20.3 The Variables Tab



5 Click the **Plots** tab.

The **Plots** tab becomes active. This tab controls which graphs are produced by the analysis, and the options for each graph (for example, whether to display confidence limits).

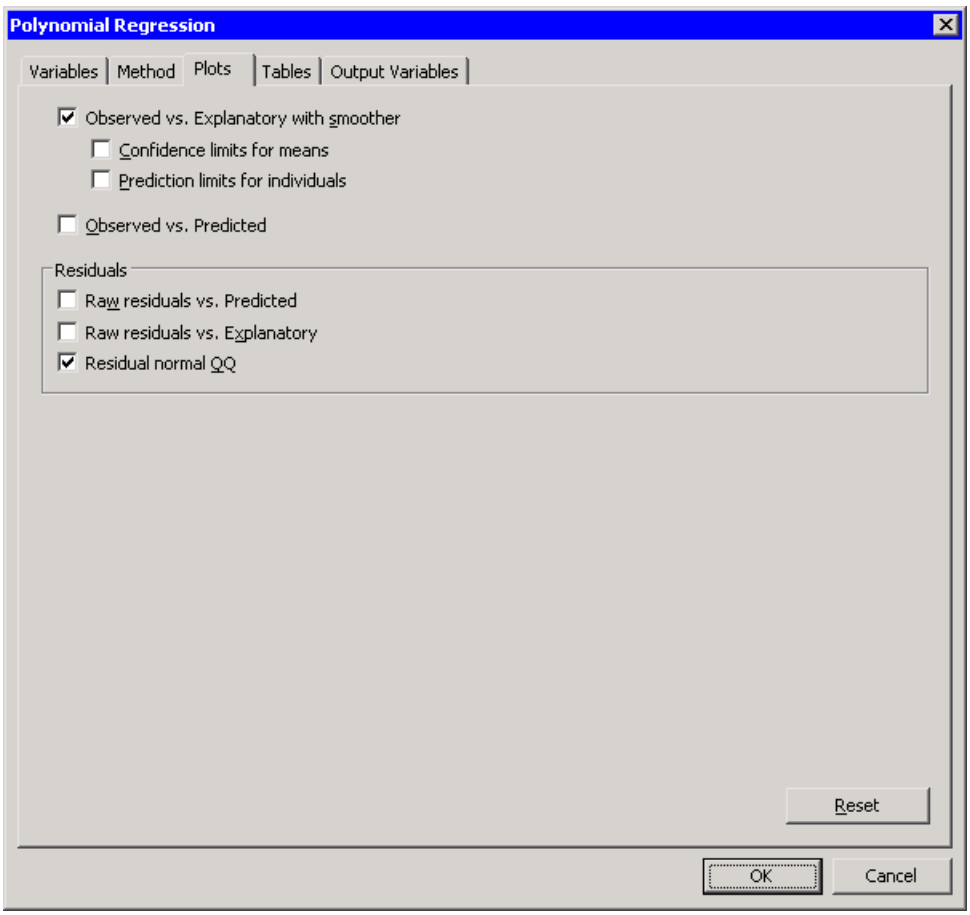
By default, the analysis creates a scatter plot of the observed X and Y variables, with a smoother added. You can decide whether or not to plot confidence limits.

6 Clear **Confidence limits for means**.

7 Select **Residual normal QQ**.

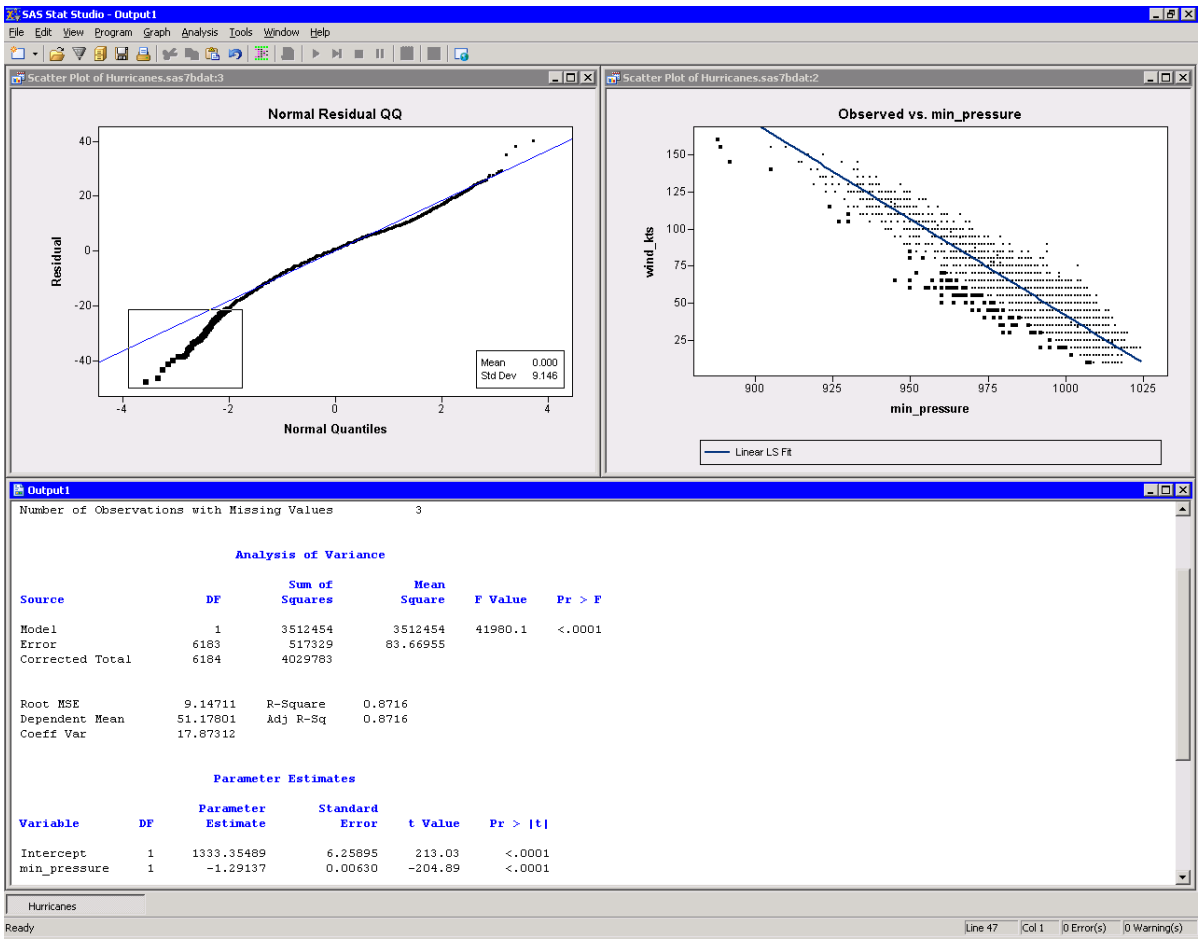
8 Click **OK**.

Figure 20.4 The Plots Tab



Several plots appear, along with output from the REG procedure. (See Figure 20.5.) A scatter plot shows the bivariate data and the requested linear smoother. The analysis also creates a normal Q-Q plot of the residuals. The Q-Q plot indicates that quite a few observations have wind speeds that are substantially lower than would be expected by assuming a linear model with normally distributed errors. In Figure 20.5 these observations are selected, and the corresponding markers in the scatter plot are highlighted.

Figure 20.5 Results from the Polynomial Regression Analysis



Output from the REG procedure appears in the output document. The output informs you that min_pressure has three missing values; those observations are not included in the analysis. The parameter estimates table indicates that when the central atmospheric pressure of a cyclone decreases by 1 hPa, you can expect the wind speed to increase by about 1.3 knots.

Specifying the Polynomial Regression Analysis

This section describes the dialog box tabs that are associated with the Polynomial Regression analysis. The Polynomial Regression analysis calls the REG procedure in SAS/STAT software. See the REG procedure documentation in the *SAS/STAT User's Guide* for details.

Variables Tab

You can use the **Variables** tab to specify the variables for the Polynomial Regression analysis.

The **Variables** tab is shown in Figure 20.3. The Y variable is the response variable. The dialog box supports a single X (explanatory) variable. To analyze a response that depends on multiple explanatory variables, you can use the Linear Regression (see Chapter 21, "Model Fitting: Linear Regression") or the Generalized Linear Models (see Chapter 24, "Model Fitting: Generalized Linear Models") analysis.

Method Tab

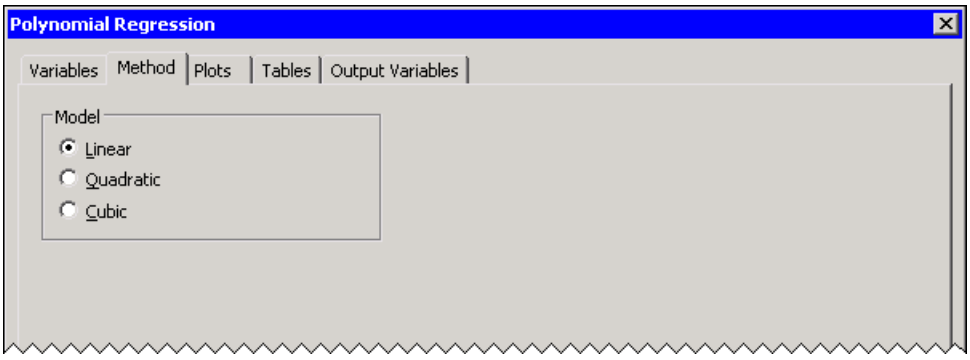
You can use the **Method** tab to specify the degree of the polynomial used to model the data. (See Figure 20.6.) The **Method** tab contains the following UI controls:

Linear
specifies a first-degree (linear) polynomial.

Quadratic
specifies a second-degree polynomial.

Cubic
specifies a third-degree polynomial.

Figure 20.6 The Method Tab



Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 20.4.) The raw residuals are computed as $Y - \hat{Y}$, where $\hat{Y}$ indicates the variable that contains the predicted values of the response.

Creating a plot often adds one or more variables to the data table. The following plots are available:

Observed vs. Explanatory with smoother
creates a scatter plot of the X and Y variables, overlaid with the polynomial smoother.

Confidence limits for means
specifies whether to add 95% upper and lower confidence limits to the Observed vs. Explanatory plot.

Prediction limits for individuals
specifies whether to add 95% upper and lower individual prediction limits to the Observed vs. Explanatory plot.

Observed vs. Predicted
creates a scatter plot of the Y variable versus the predicted values.

Raw residuals vs. Predicted
creates a scatter plot of the residuals versus the predicted values.

Raw residuals vs. Explanatory
creates a scatter plot of the residuals versus the X variable.

Residual normal QQ
creates a normal Q-Q plot of the residuals.

NOTE: SAS/IML Studio adds a smoother to an *existing* scatter plot when both of the following conditions are satisfied:

- The scatter plot is the active window when you select the analysis.
- The scatter plot variables match the analysis variables.

Chapter 18, "Data Smoothing: Loess," discusses how to display multiple smoothers in a single scatter plot, and how to remove smoothers from a scatter plot.

Tables Tab

You can use the **Tables** tab to display tables that summarize the results of the analysis.

The **Tables** tab is shown in Figure 20.7. The following tables are available:

Analysis of variance
displays an ANOVA table.

Summary of fit
displays a table of model fit statistics.

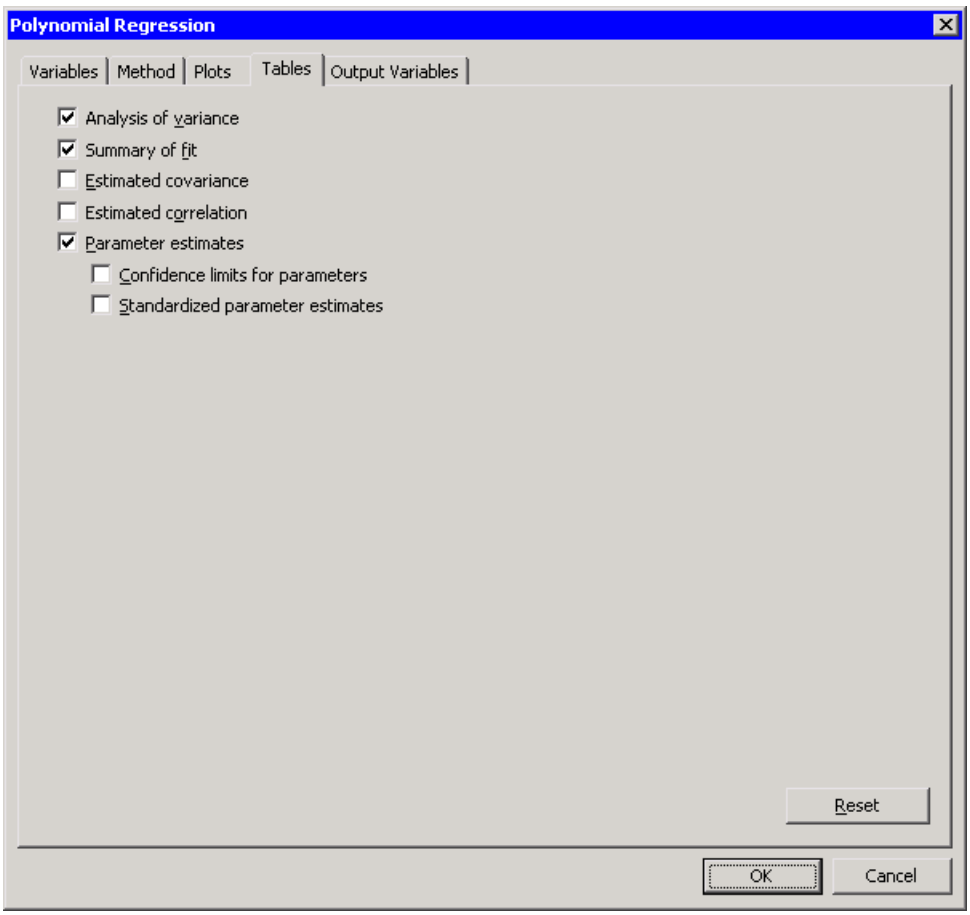
Estimated covariance
displays the covariance of the parameter estimates.

Estimated correlation
displays the correlation of the parameter estimates.

Parameter estimates
displays estimates for the model parameters.

- **Confidence limits for parameters**
  adds 95% confidence limits for the parameter estimates.
- **Standardized parameter estimates**
  adds standardized parameter estimates.

Figure 20.7 The Tables Tab



Output Variables Tab

You can use the **Output Variables** tab to add analysis variables to the data table. (See Figure 20.8.) If you request a plot that uses one of the output variables, then that variable is automatically created even if you did not explicitly select the variable on the **Output Variables** tab.

The following list describes each output variable and indicates how it is named. Y represents the name of the response variable.

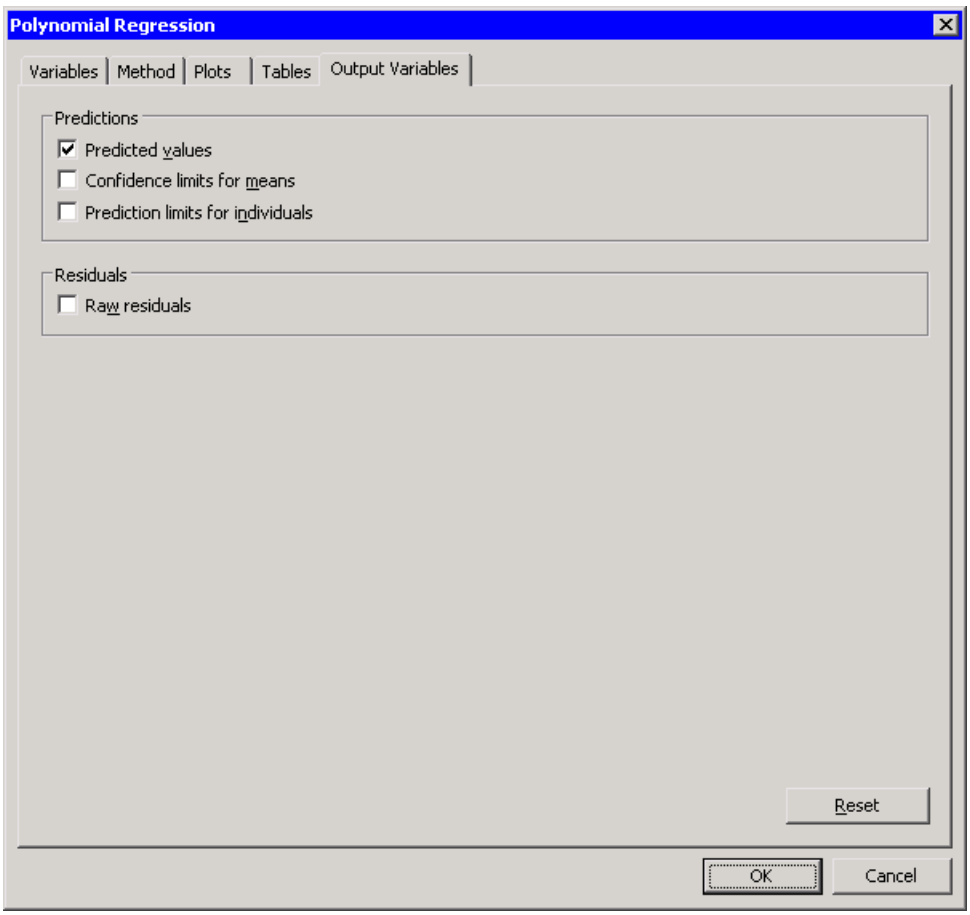
Predicted values
adds predicted values. The variable is named PolyP_Y.

Confidence limits for means
adds 95% confidence limits for the expected value (mean). The variables are named PolyLclm_Y and PolyUclm_Y.

Prediction limits for individuals
adds 95% confidence limits for an individual prediction. The variables are named PolyLcli_Y and PolyUcli_Y.

Raw residuals
adds residuals, which are calculated as observed values minus predicted values. The variable is named PolyR_Y.

Figure 20.8 The Output Variables Tab



Analysis of Selected Variables

If one or more interval variables are selected in a data table when you run the analysis, then the following occurs:

- The first selected interval variable is automatically entered in the **Y Variable** field of the **Variables** tab.
- The second selected interval variable is automatically entered in the **X Variable** field.

No role variables are used for this analysis.

Chapter 21
Model Fitting: Linear Regression

Contents

Overview of the Linear Regression Analysis **301**
Example: Fit a Linear Regression Model **301**
Part 1: Transform the Response Variable 302
Part 2: Select a Variable that Identifies Observations 305
Part 3: Model the Response Variable 306
Part 4: Interpret the Plots 309
Specifying the Linear Regression Analysis **315**
Variables Tab 316
Plots Tab 316
Tables Tab 317
Output Variables Tab 319
Roles Tab 321
Analysis of Selected Variables **321**
References **321**

Overview of the Linear Regression Analysis

The Linear Regression analysis fits a linear regression model by using ordinary least squares. You can write the multiple linear regression equation for a model with p explanatory variables as

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

where Y is the response variable, the X_i's are explanatory variables, and the b_i's are regression coefficients.

You can run a Linear Regression analysis by selecting **Analysis ►Model Fitting ►Linear Regression** from the main menu. The computation of the regression function, confidence limits, and related statistics is implemented by calling the REG procedure in SAS/STAT software. See the documentation for the REG procedure in the *SAS/STAT User's Guide* for additional details.

Example: Fit a Linear Regression Model

In this example you fit a linear regression model to predict the 1987 salaries of Major League Baseball players as a function of several explanatory variables in the Baseball data set. The response variable is salary. The example examines three explanatory variables: two measures of hitting performance and one measure of longevity. The explanatory variables are described in the following list:

- no_hits, the number of hits in 1986
- no_home, the number of home runs in 1986
- yr_major, the number of years that the player had been in the major leagues as of 1987

The example has four major steps:

1. Apply a logarithmic transformation to the response variable.
2. Set name to be the variables whose values are used to label observations.
3. Run the Linear Regression analysis.
4. Interpret the various plots that the analysis can produce.

Part 1: Transform the Response Variable

The salary variable ranges from 67.5 to 2,460 (measured in thousands of dollars). Since the variation of salaries is much greater for the higher salaries, it is appropriate to apply a logarithmic transformation to the salaries before fitting the model. The following steps use the Variable Transformation Wizard to transform the salary variable. (This wizard is described in further detail in Chapter 32, "Variable Transformations.")

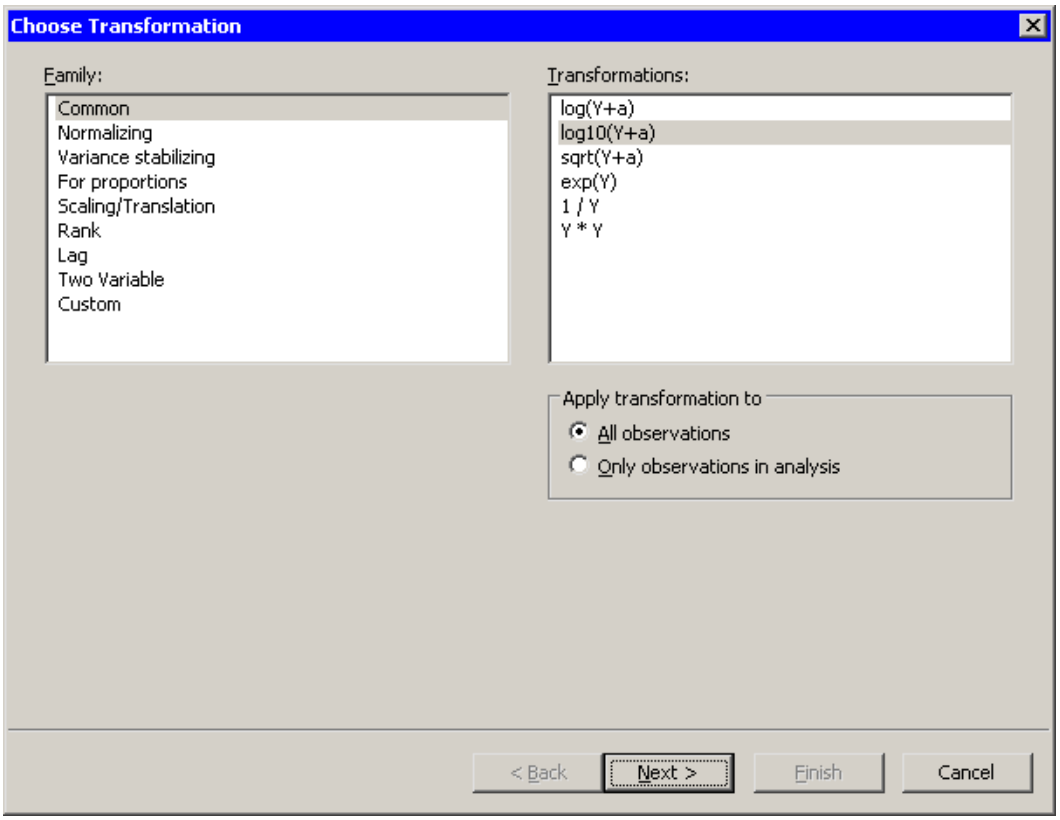
1 Open the Baseball data set.

2 Select **Analysis ►Variable Transformation** from the main menu.

The Variable Transformation Wizard in Figure 21.1 appears.

3 Select the **log10(Y+a)** transformation from the **Transformations** list.

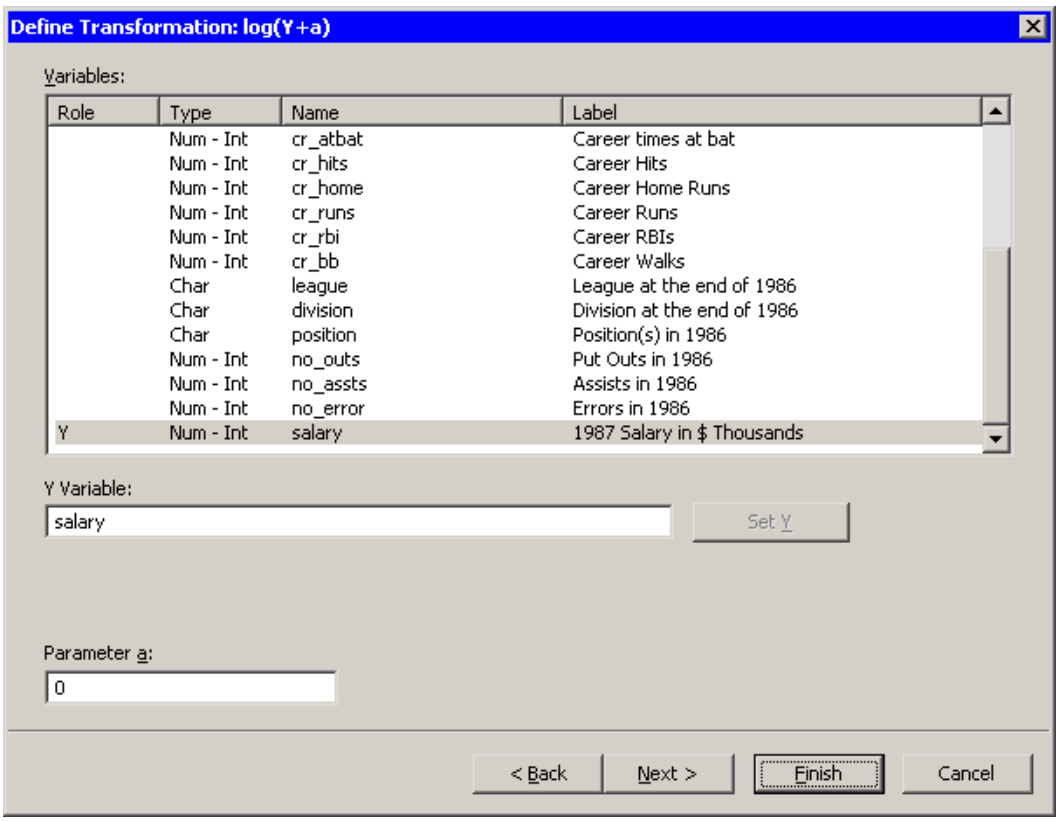
Figure 21.1 Selecting a Log10 Transformation



4 Click **Next**.

The wizard displays the page shown in Figure 21.2.

Figure 21.2 Selecting a Variable and Parameters



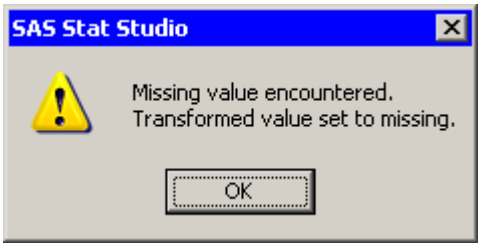
5 Scroll to the end of the variable list. Select the salary variable, and click **Set Y**.

The parameter *a* is an offset that is useful if your variable contains nonpositive values. For these data, you can accept the default value of 0.

6 Click **Finish**.

Because there are missing values for the salary variable, a warning message appears (Figure 21.3) that informs you that the transformed values for these observations are set to missing values.

Figure 21.3 A Warning Message



7 Click **OK** to dismiss the warning message.

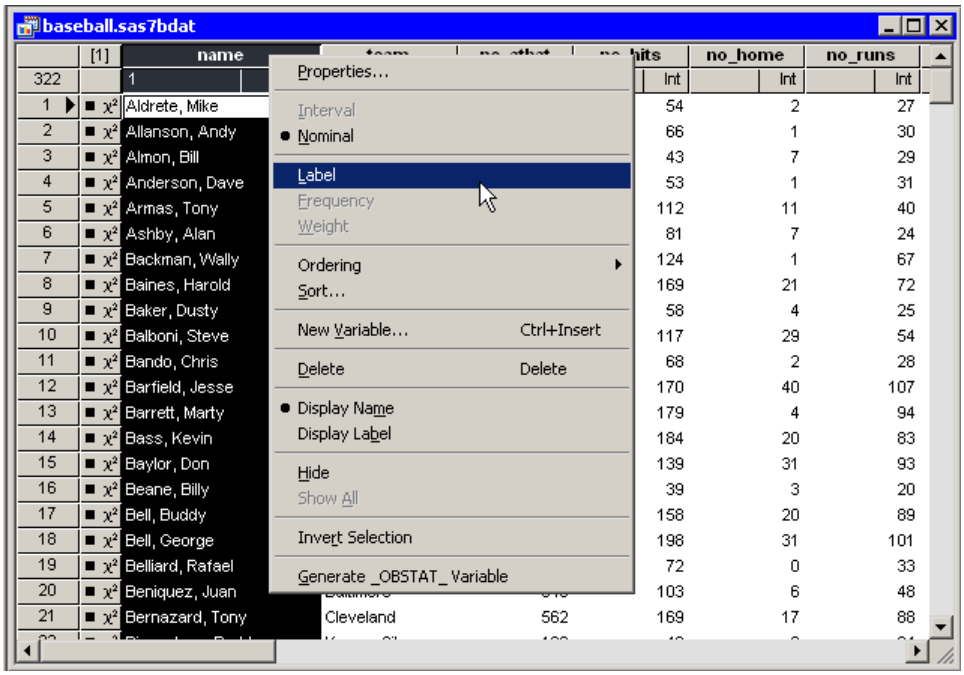
SAS/IML Studio adds the new variable, Log10_salary, as the last variable in the data set.

Part 2: Select a Variable that Identifies Observations

For these data, each observation represents a player. It is convenient to use the name of each player to identify observations in residual plots and diagnostic plots. The following step sets the value of the name variable to be the label you see when you click an observation.

1 Right-click the variable heading for name to display the **Variables** menu. Select **Label**, as shown in Figure 21.4.

Figure 21.4 Selecting a Variable Used to Label Observations

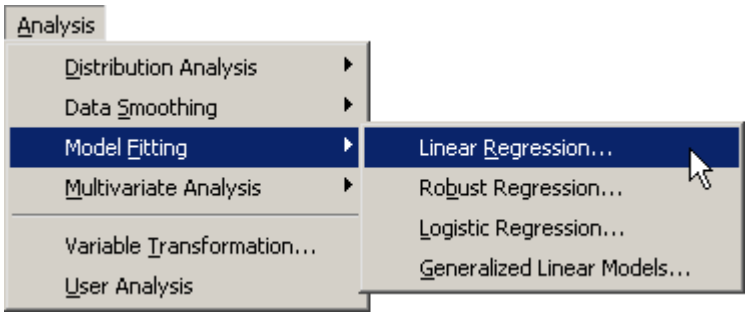


Part 3: Model the Response Variable

To model Log10_salary as a function of three explanatory variables:

1 Select **Analysis ►Model Fitting ►Linear Regression** from the main menu, as shown in Figure 21.5.

Figure 21.5 Selecting a Linear Regression

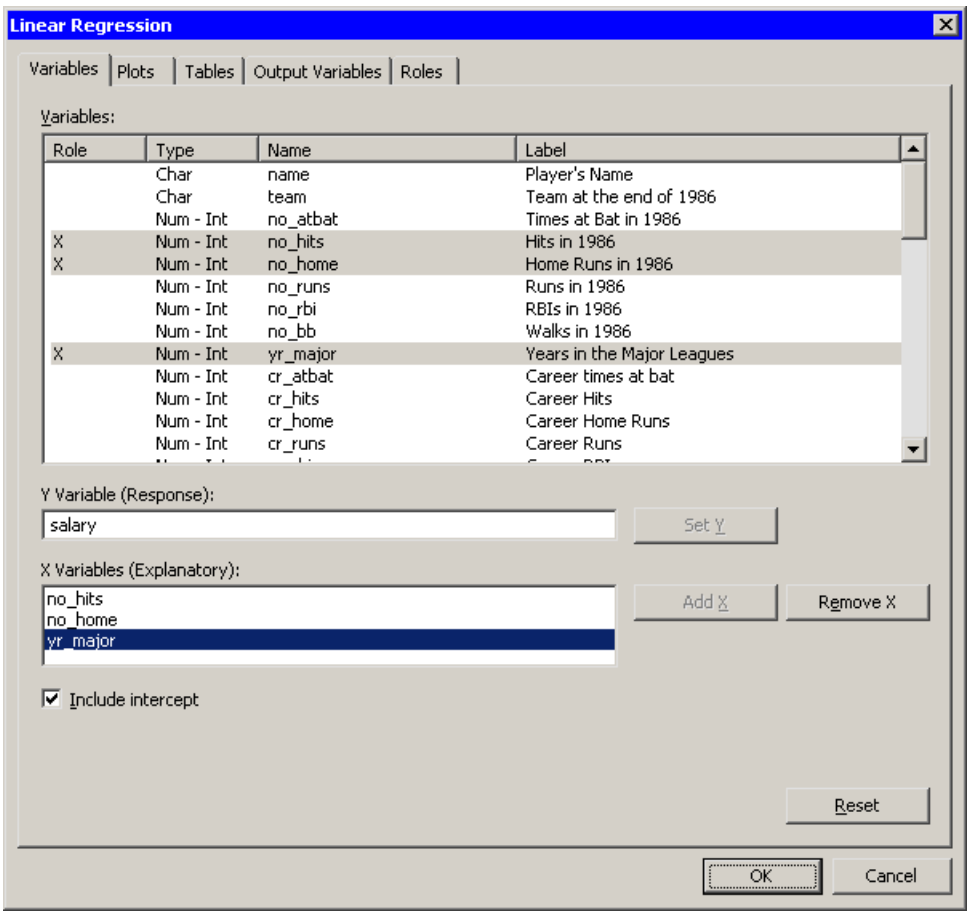


The Linear Regression dialog box appears. (See Figure 21.6.)

2 Scroll to the end of the variable list. Select Log10_salary, and click **Set Y**.

3 Select no_hits. While holding down the CTRL key, select no_home, and yr_major. Click **Add X**.

Figure 21.6 The Variables Tab

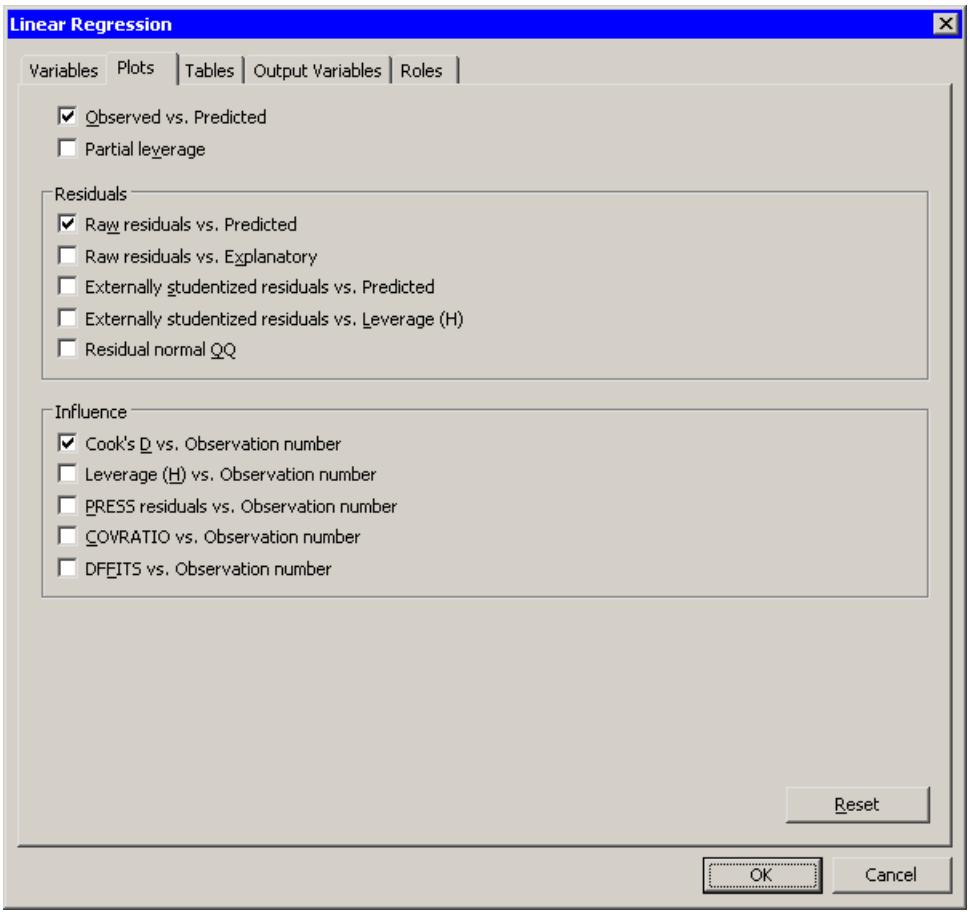


4 Click the **Plots** tab.

The **Plots** tab becomes active, as shown in Figure 21.7. This tab controls which graphs are produced by the analysis.

5 Select **Cook's D vs. Observation number**.

Figure 21.7 The Plots Tab

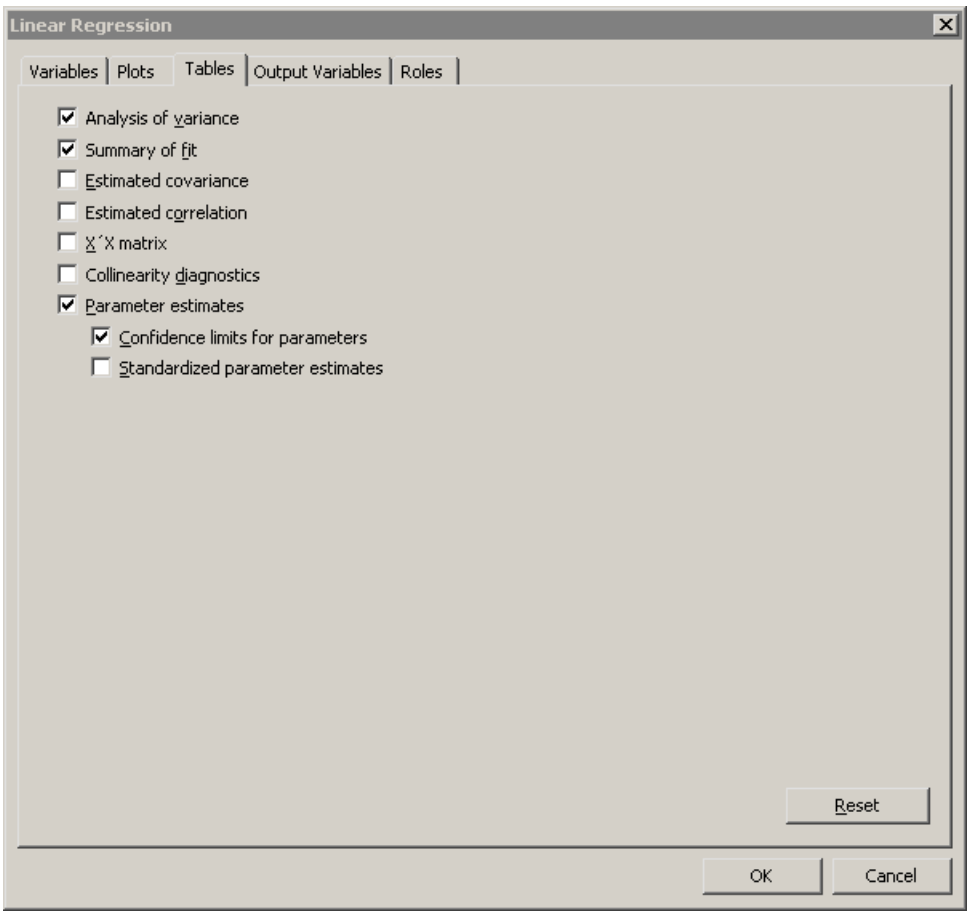


6 Click the **Tables** tab.

The **Tables** tab becomes active, as shown in Figure 21.8.

7 Click **Confidence limits for parameters**.

Figure 21.8 The Tables Tab



8 Click **OK**.

Several plots appear, along with output from the REG procedure. Some plots might be hidden beneath others. Move the windows so that they are arranged as in Figure 21.9.

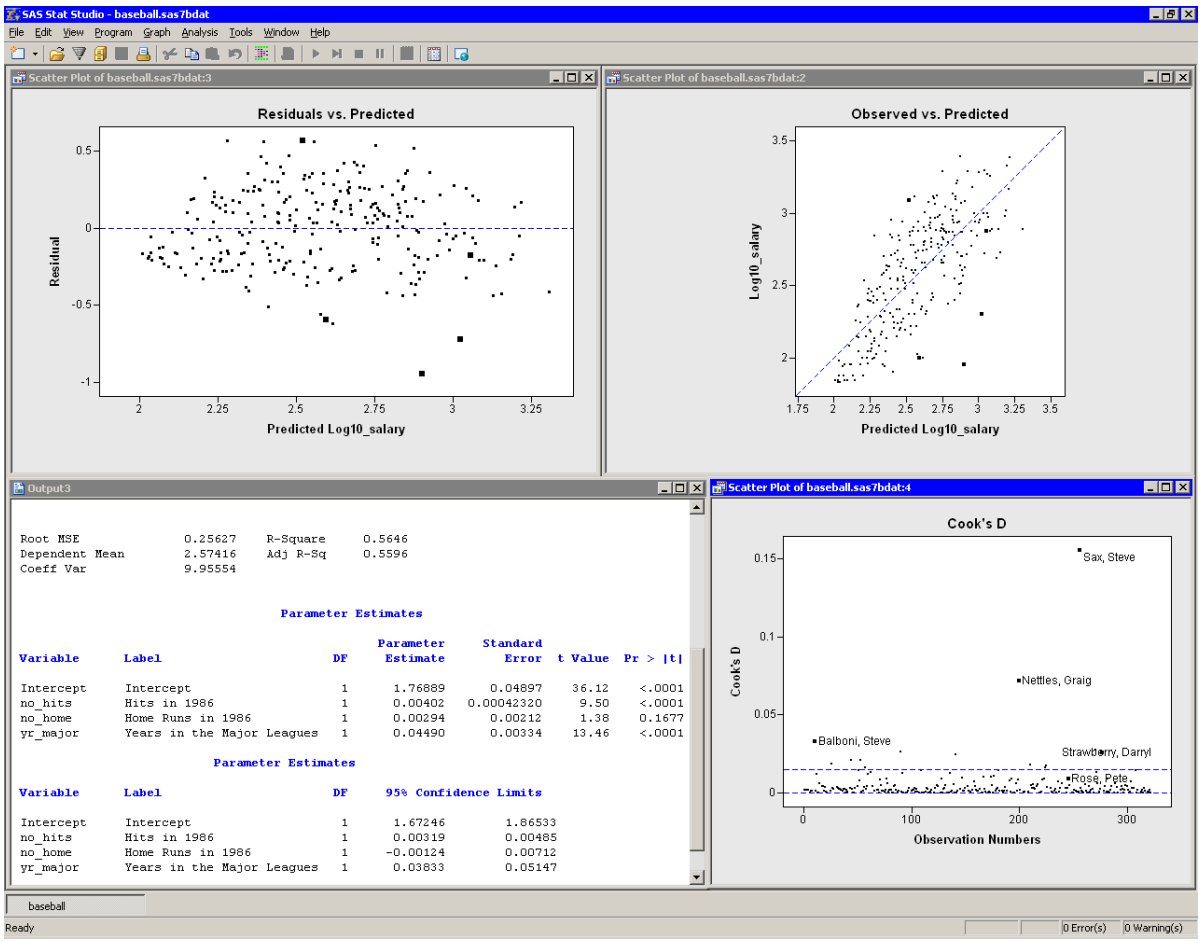
The Residuals vs. Predicted plot does not show any obvious trends in the residuals, although possibly the residuals are slightly higher for predicted values near the middle of the predicted range. The Observed vs. Predicted plot shows a reasonable fit, with a few exceptions.

In the output window you can see that R square is 0.5646, which means that the model accounts for 56% of the variation in the data. The no_home term is not significant ($t = 1.38$, $p = 0.1677$) and thus can be removed from the model. This is also seen by noting that the 95% confidence limits for the coefficient of no_home include zero.

The Cook's D plot shows how deleting any one observation would change the parameter estimates. (Cook's D and other influence statistics are described in the "Influence Diagnostics" section of the documentation for the REG procedure.) A few influential observations have been selected in the plot of Cook's D; these observations are seen highlighted in the other plots. Three players (Steve Sax, Graig Nettles, and Steve Balboni) with high Cook's D values also have large negative residuals which indicates that they were paid less than the model predicts.

Two other players (Darryl Strawberry and Pete Rose) are also highlighted. These players are discussed in the next section.

Figure 21.9 Results from the Linear Regression Analysis



Part 4: Interpret the Plots

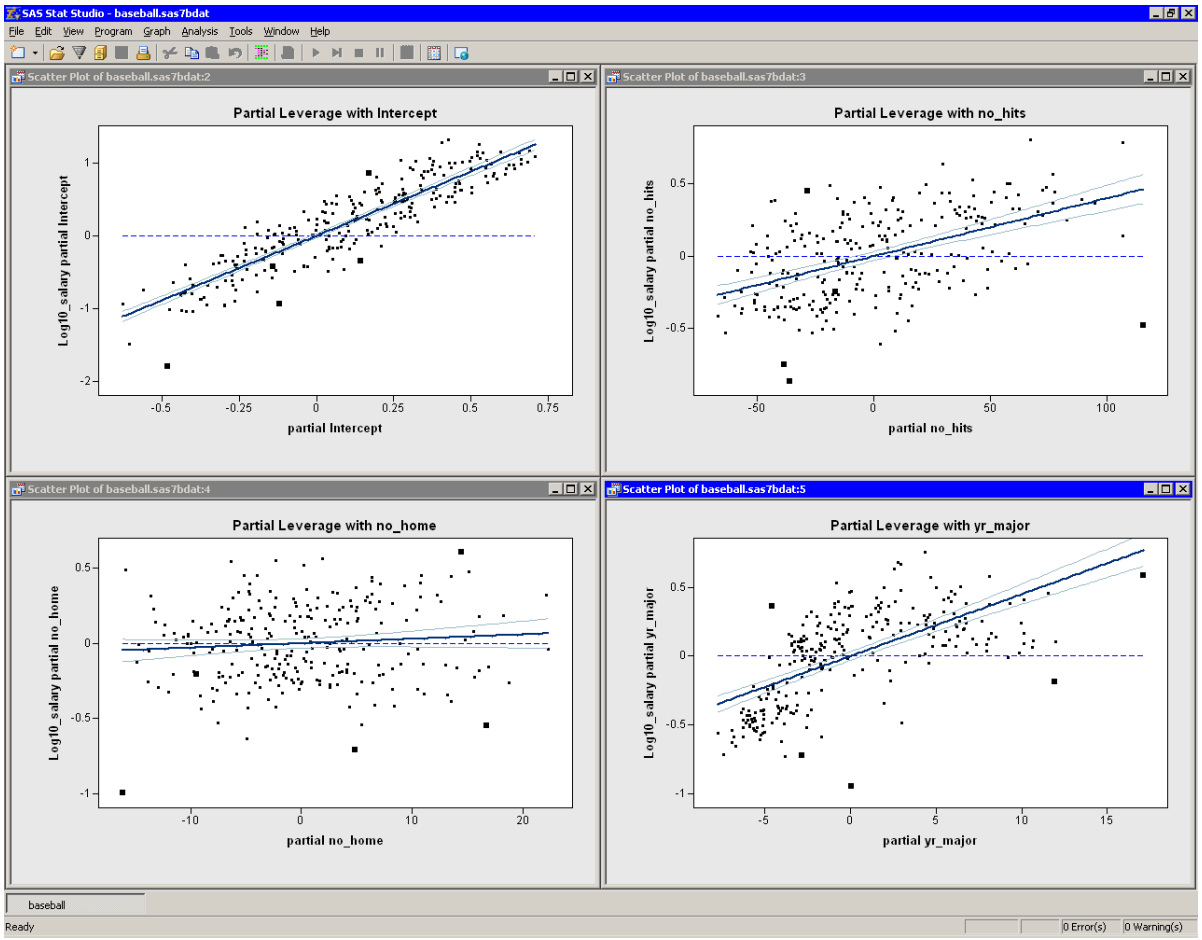
You can use the Linear Regression analysis to create a variety of residual and diagnostic plots, as indicated by Figure 21.7. This section briefly presents the types of plots that are available. To provide common reference points, the same five observations are selected in each set of plots.

Partial Leverage Plots

Partial leverage plots are an attempt to isolate the effects of a single variable on the residuals (Rawlings, Pantula, and Dickey 1998, p. 359). A partial regression leverage plot is the plot of the residuals for the dependent variable against the residuals for a selected regressor, where the residuals for the dependent variable are calculated with the selected regressor omitted and the residuals for the selected regressor are calculated from a model in which the selected regressor is regressed on the remaining regressors. A line fit to the points has a slope that is equal to the parameter estimate in the full model. Confidence limits for each regressor are related to the confidence limits for parameter estimates (Sall 1990).

Partial leverage plots for the previous example are shown in Figure 21.10. The lower left plot shows residuals of no_home. The confidence bands in this plot contain the horizontal reference line, which indicates that the coefficient of no_home is not significantly different from zero.

Figure 21.10 Partial Leverage Plots

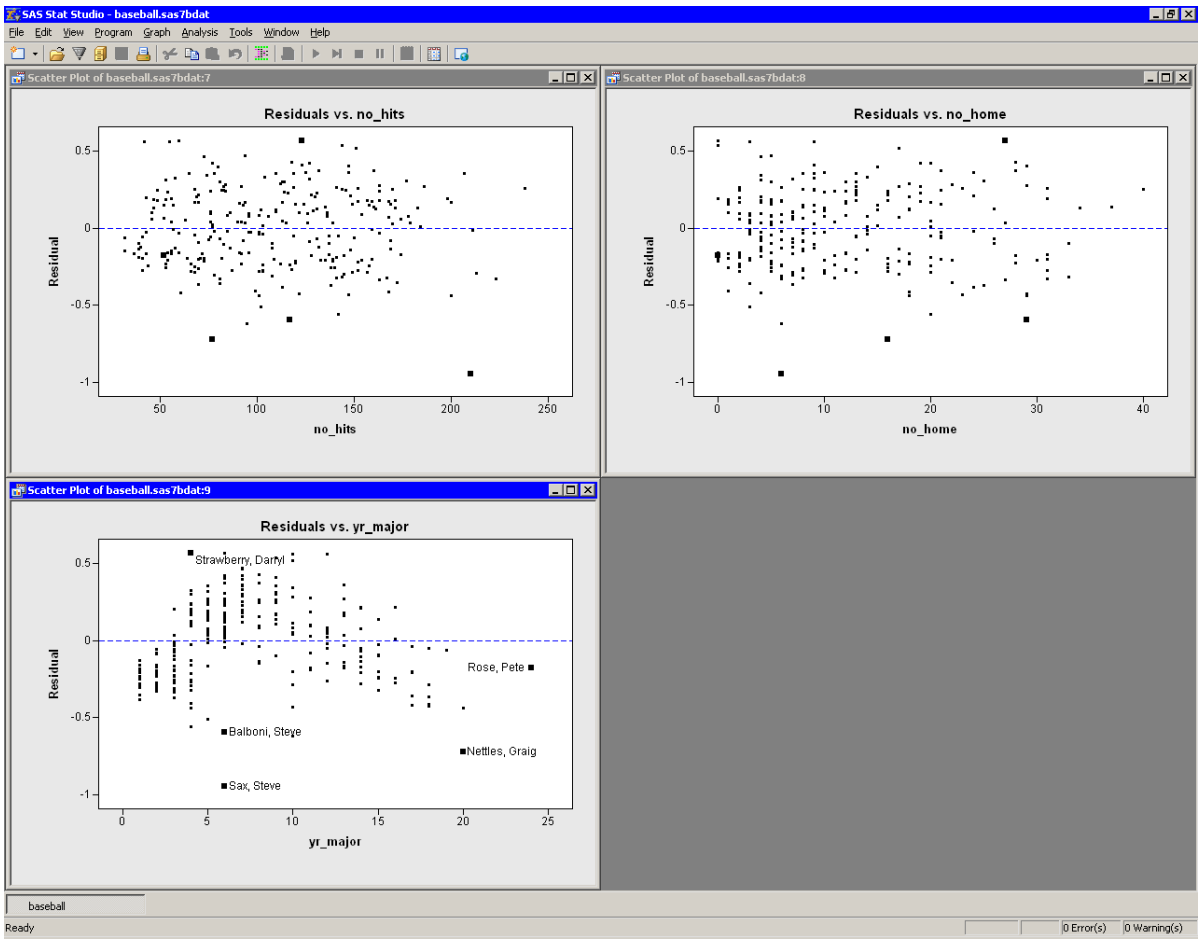


Plots of Residuals versus Explanatory Variables

Figure 21.11 shows the residuals plotted against the three explanatory variables in the model. Note that the Residuals vs. yr_major plot shows a distinct pattern. The plot indicates that players who have recently joined the major leagues earn less money, on average, than their veteran counterparts with 5–10 years of experience. The mean salary for players with 10–20 years of experience is comparable to the salary that new players make.

This pattern of residuals suggests that the example does not correctly model the effect of the yr_major variable. Perhaps it is more appropriate to model log10_salary as a nonlinear function of yr_major. Also, the low salaries of Steve Sax, Graig Nettles, and Steve Balboni might be unduly influencing the fit.

Figure 21.11 Residual versus Explanatory Plots



More Residual Plots

Figure 21.12 shows several residual plots.

The Q-Q plot (upper left in Figure 21.12) shows that the residuals are approximately normally distributed. Three players with large negative residuals (Steve Sax, Graig Nettles, and Steve Balboni) are highlighted below the diagonal line in the plot. These players seem to be outliers for this model.

The Residuals vs. Predicted plot is located in the upper right corner of Figure 21.12. As noted in the example, the residuals show a slight "bend" when plotted against the predicted value. Figure 21.13 makes the trend easier to see by adding a loess smoother to the residual plot. (See Chapter 18, "Data Smoothing: Loess," for more information about adding loess curves.) As discussed in the previous section, this trend might indicate the need for a nonlinear term that involves yr_major. Alternatively, excluding or downweighting outliers might lead to a better fit.

Figure 21.12 Residual Plots

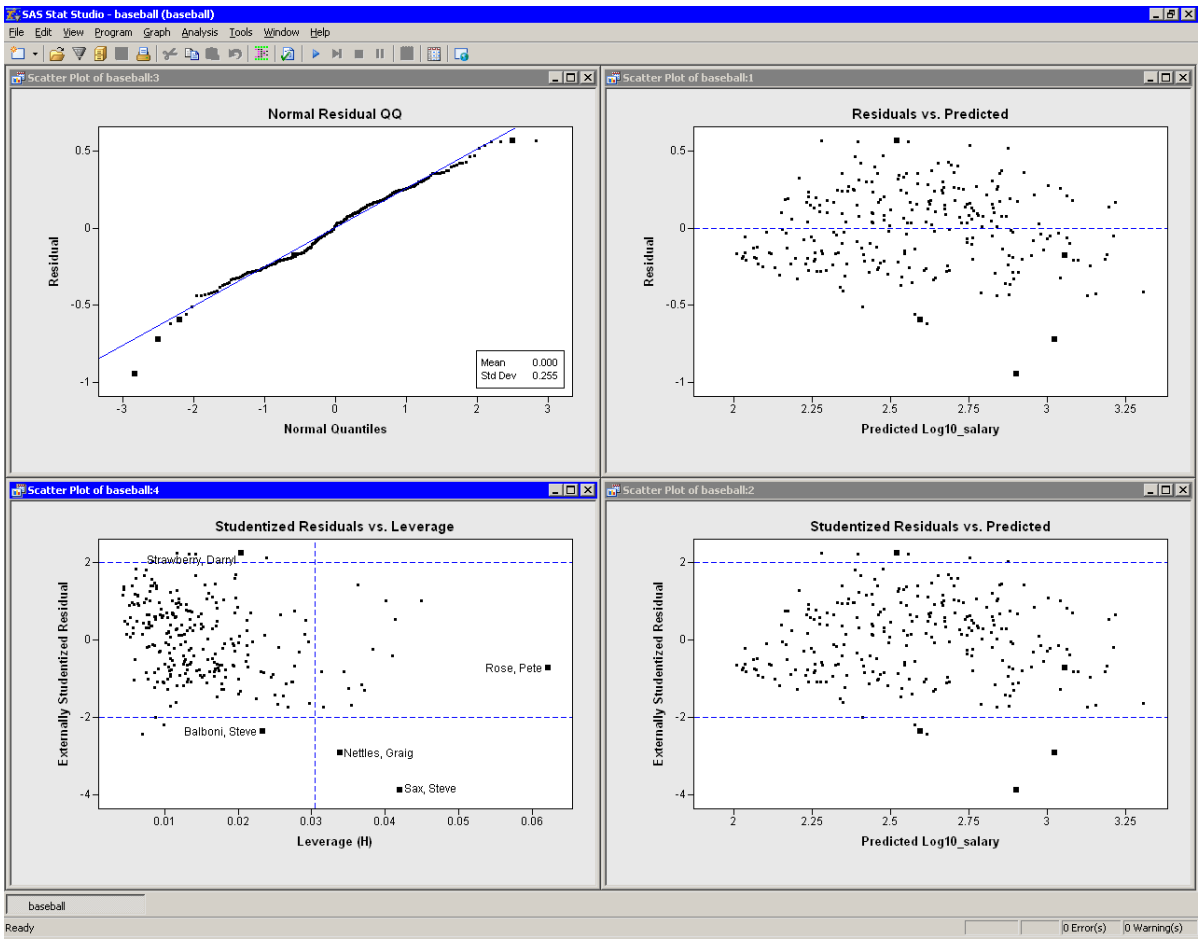
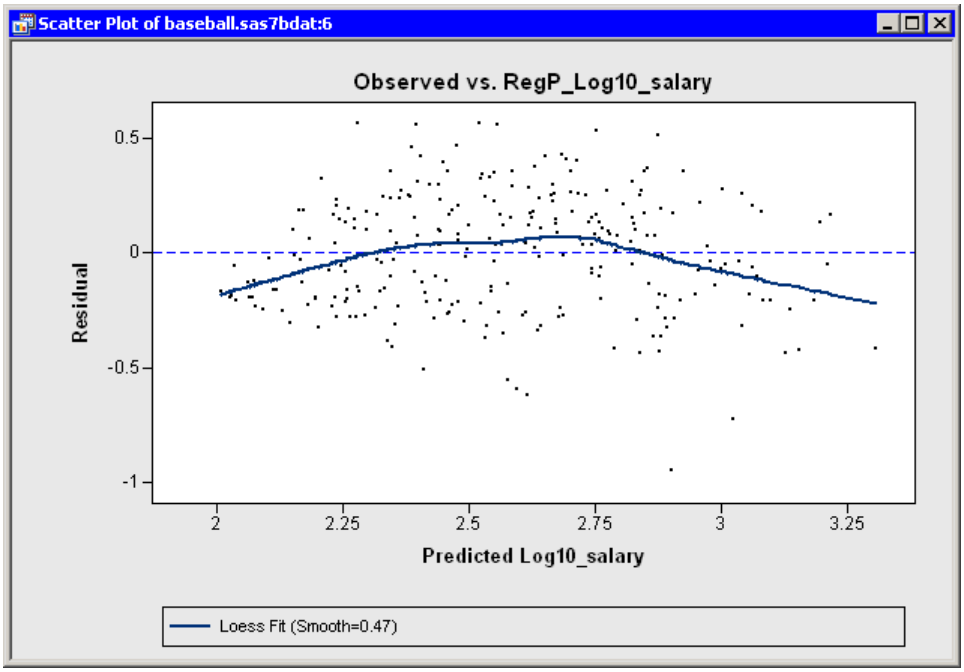


Figure 21.13 A Loess Smoother of the Residuals



The lower right plot in Figure 21.12 is a graph of externally studentized residuals versus predicted values. The externally studentized residual (known as RSTUDENT in the documentation of the REG procedure) is a studentized residual in which the error variance for the ith observation is estimated without including the ith observation. You should examine an observation when the absolute value of the studentized residual exceeds 2.

The lower left plot in Figure 21.12 is a graph of (externally) studentized residuals versus the *leverage statistic*. The leverage statistic for the ith observation is also the ith element on the diagonal of the *hat matrix*. The leverage statistic indicates how far an observation is from the centroid of the data in the space of the explanatory variables. Observations far from the centroid are potentially influential in fitting the regression model.

Observations whose leverage values exceed $2p/n$ are called *high leverage points* (Belsley, Kuh, and Welsch 1980). Here p is the number of parameters in the model (including the intercept) and n is the number of observations used in computing the least squares estimates. For the example, $n = 263$ observations are used. There are three regressors in addition to the intercept, so $p = 4$. The cutoff value is therefore 0.0304.

The Studentized Residuals vs. Leverage plot has a vertical line that indicates high leverage points and two horizontal lines that indicate potential outliers. In Figure 21.12, Pete Rose is an observation with high leverage (due to his 24 years in the major leagues), but not an outlier. Graig Nettles and Steve Sax are outliers and leverage points. Steve Balboni is an outlier because of a low salary relative to the model, whereas Darryl Strawberry's salary is high relative to the prediction of the model.

You should be careful in interpreting results when there are high leverage points. It is possible that Pete Rose fits the model precisely because he *is* a high leverage point. Chapter 22, "Model Fitting: Robust Regression," describes a robust technique for identifying high leverage points and outliers.

Influence Diagnostic Plots

Previous sections discussed plots that included Cook's D statistic and the leverage statistic. Both of these statistics are *influence diagnostics*. (See Rawlings, Pantula, and Dickey 1998, p. 361, for a summary of influence statistics.) Figure 21.14 show other plots that are designed to identify observations that have a large influence on the parameter estimates in the model. For each plot, the horizontal axis is the observation number.

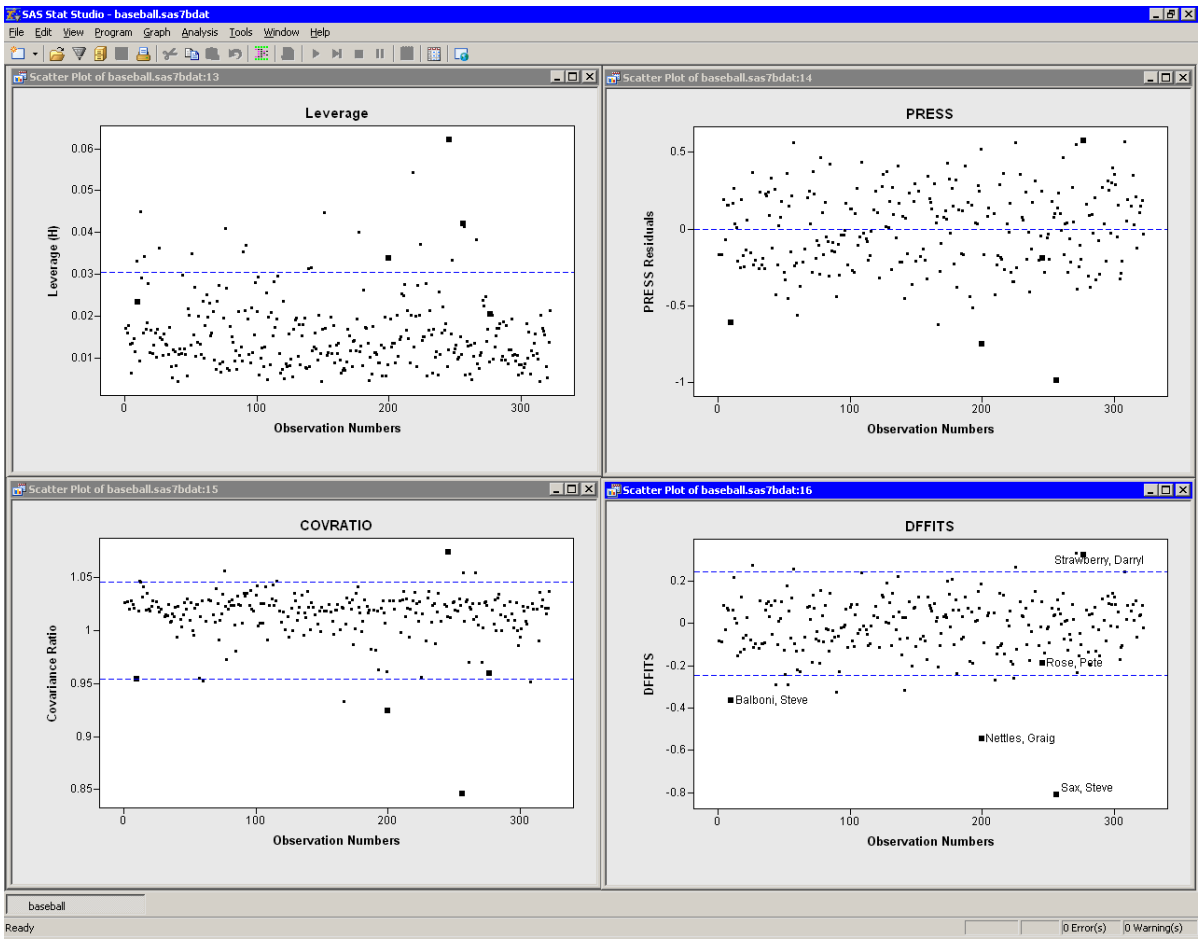
The upper left plot displays the leverage statistic along with the cutoff $2p/n$.

The upper right plot displays the PRESS residuals. The PRESS residual for observation i is the residual that would result if you fit the model without using the ith observation. A large press residual indicates an influential observation. Pete Rose does not have a large PRESS residual.

The lower left plot displays the *covariance ratio*. The covariance ratio measures the change in the determinant of the covariance matrix of the estimates by deleting the ith observation. Influential observations have $|c - 1| \geq 3p/n$, where c is the covariance ratio (Belsley, Kuh, and Welsch 1980). Horizontal lines on the plot mark the critical values. Pete Rose has the largest value of the covariance ratio.

The lower right plot displays the DFFIT statistic, which is similar to Cook's D. The observations outside of $\pm\sqrt{p/n}$ are influential (Belsley, Kuh, and Welsch 1980). Pete Rose is not influential by this measure.

Figure 21.14 Influence Diagnostics Plots



Specifying the Linear Regression Analysis

This section explains the dialog box tabs that are associated with the Linear Regression analysis. The Linear Regression analysis calls the REG procedure in SAS/STAT software. See the REG documentation in the *SAS/STAT User's Guide* for details.

Variables Tab

You can use the **Variables** tab to specify the variables for the Linear Regression analysis.

The **Variables** tab is shown in Figure 21.6. The Y variable is the response variable. The dialog box supports multiple X (explanatory) variables. All X and Y variables must be interval variables.

The Linear Regression analysis does not support nominal classification variables, nor does it support specifying interaction effects such as X1*X2 or higher-order polynomial effects such as $X1^2$. You can create models with these features by using the Generalized Linear Models analysis, as described in Chapter 24, "Model Fitting: Generalized Linear Models."

Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 21.15.) There are plots that help you to visualize the fit, the residuals, and various influence diagnostics.

Creating a plot often adds one or more variables to the data table. The following plots are available:

Observed vs. Predicted
creates a scatter plot of the Y variables versus the predicted values, overlaid with the diagonal line that represents a perfect fit.

Partial leverage
creates a partial leverage plot for each regressor and for the intercept. A line in the plot has a slope equal to the parameter estimate in the full model. Confidence limits for each regressor are related to the confidence limits for parameter estimates.

Raw residuals vs. Predicted
creates a scatter plot of the residuals versus the predicted values.

Raw residuals vs. Explanatory
creates scatter plots of the residuals versus the X variables.

Externally studentized residuals vs. Predicted
creates a scatter plot of the studentized residuals versus the predicted value.

Externally studentized residuals vs. Leverage (H)
creates a scatter plot of the studentized residuals versus the leverage statistic.

Residual normal QQ
creates a normal Q-Q plot of the residuals.

Cook's D vs. Observation number
creates a scatter plot of Cook's D statistic for each observation.

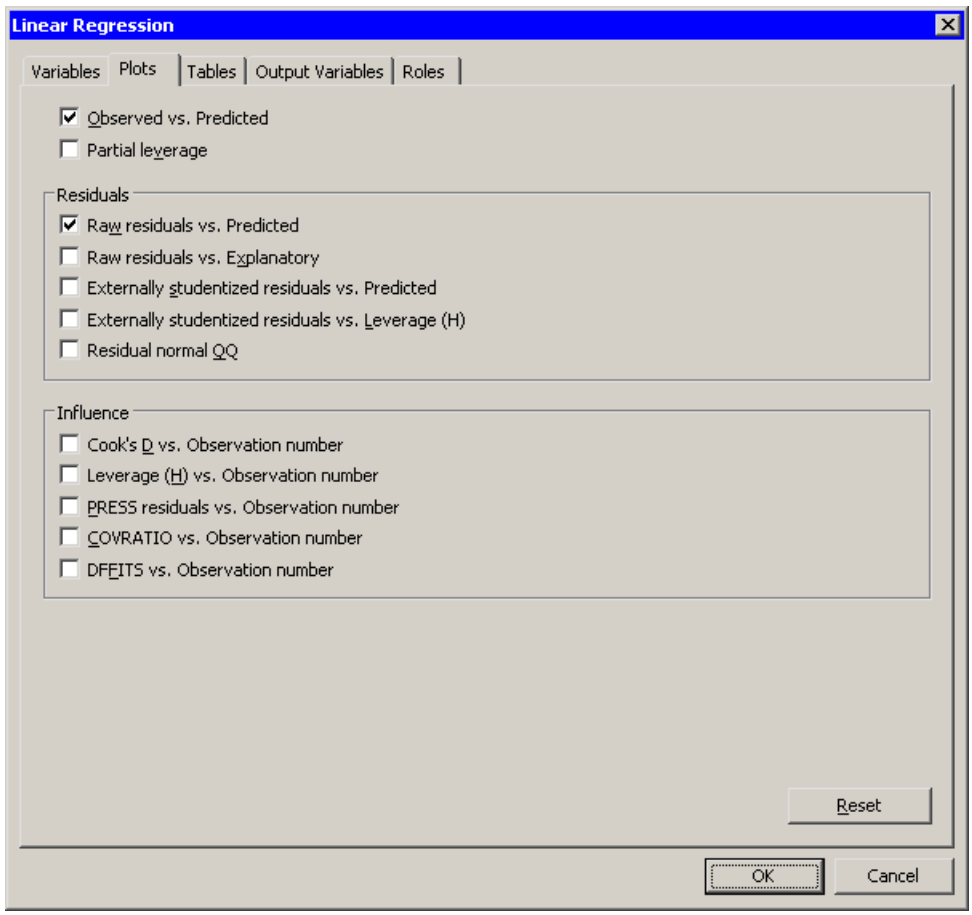
Leverage (H) vs. Observation number
creates a scatter plot of the leverage statistic for each observation.

PRESS residuals vs. Observation number
creates a scatter plot of the PRESS residual for each observation.

COVRATIO vs. Observation number
creates a scatter plot of the covariance ratio for each observation.

DFFITS vs. Observation number
creates a scatter plot of the DFFIT statistic for each observation.

Figure 21.15 The Plots Tab



Tables Tab

The **Tables** tab is shown in Figure 21.16. You can use the **Tables** tab to display the following tables that summarize the results of the analysis:

Analysis of variance
displays an ANOVA table.

Summary of fit
displays a table of model fit statistics.

Estimated covariance
displays the covariance of the parameter estimates.

Estimated correlation
displays the correlation of the parameter estimates.

X'X matrix
displays the X'X crossproducts matrix for the model. The crossproducts matrix is bordered by the X'Y and Y'Y matrices.

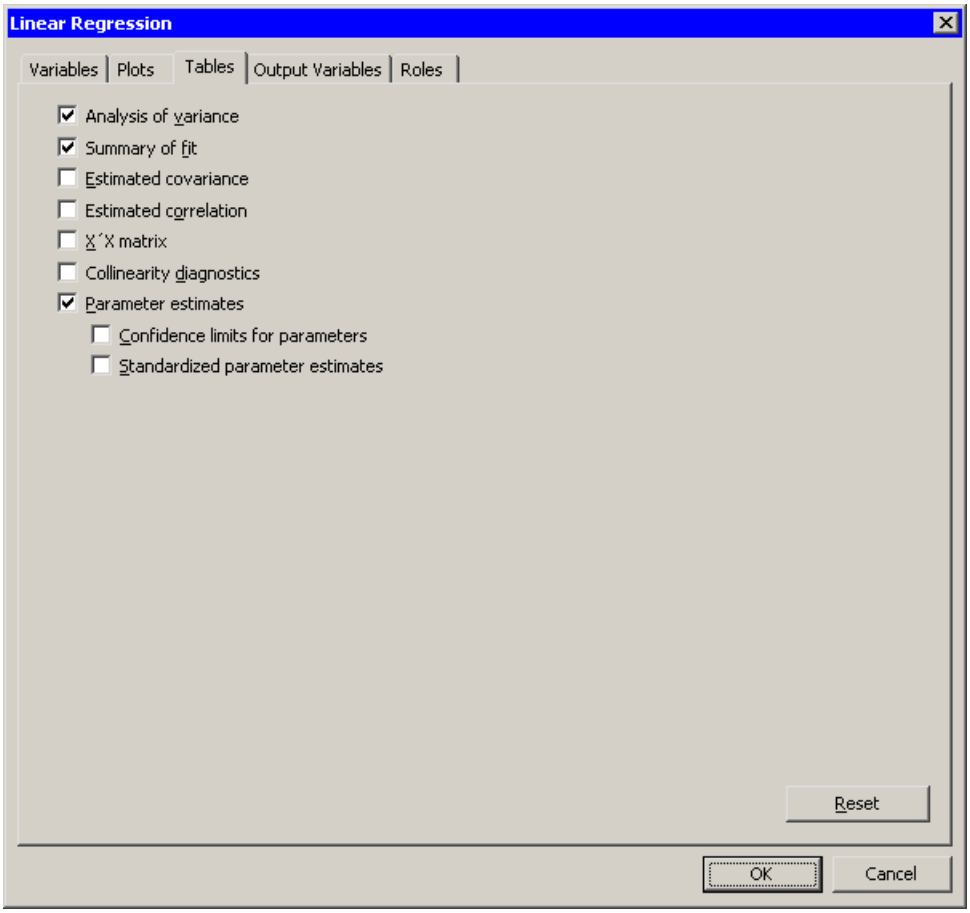
Collinearity diagnostics
displays a detailed analysis of collinearity among the regressors.

Parameter estimates
displays estimates for the model parameters.

Confidence limits for parameters
adds 95% confidence limits for the parameter estimates.

Standardized parameter estimates
adds standardized parameter estimates.

Figure 21.16 The Tables Tab



Output Variables Tab

You can use the **Output Variables** tab to add analysis variables to the data table. (See Figure 21.17.) If you request a plot that uses one of the output variables, then that variable is automatically created even if you did not explicitly select the variable on the **Output Variables** tab.

The following list describes each output variable and indicates how the output variable is named. Y represents the name of the response variable.

Predicted values
adds predicted values. The variable is named RegP_Y.

Confidence limits for means
adds 95% confidence limits for the expected value (mean). The variables are named RegLclm_Y and RegUclm_Y.

Prediction limits for individuals
adds 95% confidence limits for an individual prediction. The variables are named RegLcli_Y and RegUcli_Y.

Raw residuals
adds residuals, which are calculated as observed values minus predicted values. The variable is named RegR_Y.

Internally studentized residuals
adds internally studentized residuals, which are the residuals divided by their standard errors. (These correspond to the STUDENT= option in the OUTPUT statement.) The variable is named RegIntR_Y.

Externally studentized residuals
adds externally studentized residuals, which are studentized residuals with the current observation deleted. (These correspond to the RSTUDENT= option in the OUTPUT statement.) The variable is named RegExtR_Y.

Cook's D
adds Cook's D influence statistic. The variable is named RegCooksD_Y.

Leverage (H)
adds the leverage statistic. The variable is named RegH_Y.

PRESS residuals
adds the PRESS residuals. This is the ith residual divided by $1 - h$, where h is the leverage and where the model has been refit without the ith observation. The variable is named RegPRESS_Y.

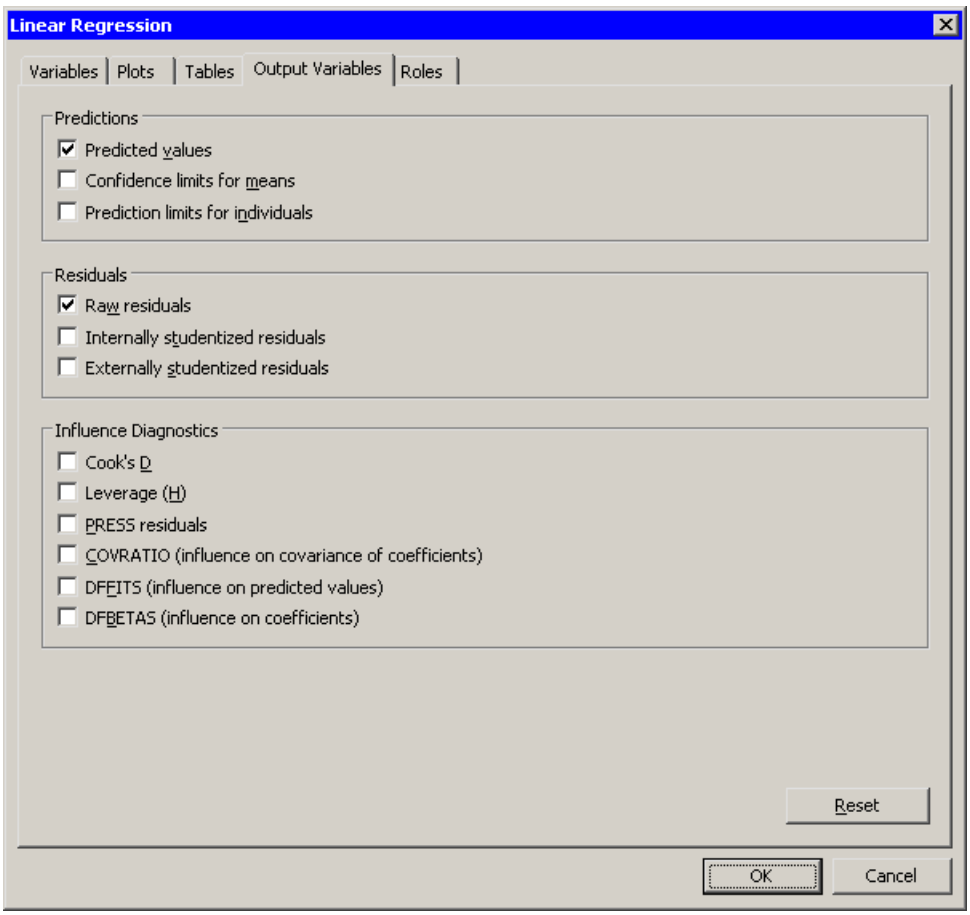
COVRATIO (influence on covariance of coefficients)
adds the covariance ratio. This is the ith residual divided by $1 - h$, where h is the leverage and where the model has been refit without the ith observation. The variable is named RegCovRatio_Y.

DFFITS (influence on predicted values)
adds the standard influence of observation on the predicted value. The variable is named RegDFFITS_Y.

DFBETAS (influence on coefficients)

adds p variables, where p is the number of parameters in the model. The variables are scaled measures of the change in each parameter estimate and are calculated by deleting the ith observation. Large values of DFBETAS indicate observations that are influential in estimating a given parameter. Belsley, Kuh, and Welsch (1980) recommend $2/\sqrt{n}$ as a size-adjusted cutoff. The variables are named DFB_X_j, where X_j is the name of the jth regressor (including the intercept).

Figure 21.17 The Output Variables Tab



Roles Tab

You can use the **Roles** tab to specify a frequency variable or weight variable for the analysis.

A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents n observations, where n is the value of the frequency variable.

A weight variable is a numeric variable with values that are relative weights for a weighted least squares fit.

Analysis of Selected Variables

If one or more interval variables are selected in a data table when you run the analysis, then the following occurs:

- The first selected interval variable is automatically entered in the **Y Variable** field of the **Variables** tab.
- The remaining selected interval variables are automatically entered in the **X Variable** field of the **Variables** tab.

Any variable in the data table with a Frequency or Weight role is automatically entered in the appropriate field of the **Roles** tab.

References

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: John Wiley & Sons.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998), *Applied Regression Analysis: A Research Tool*, 2nd Edition, New York: Springer-Verlag.

Sall, J. P. (1990), "Leverage Plots for General Linear Hypotheses," *American Statistician*, 44, 308–315.

Chapter 22
Model Fitting: Robust Regression

Contents

Overview of the Robust Regression Analysis **323**
Example: Fit a Robust Regression Model **323**
Using the Results of Robust Regression 329
Specifying the Robust Regression Analysis **331**
Variables Tab 332
Method Tab 332
Plots Tab 333
Tables Tab 334
Output Variables Tab 336
Roles Tab 337
Analysis of Selected Variables **337**

Overview of the Robust Regression Analysis

The Robust Regression analysis fits a linear regression model that is robust in the presence of outliers and high leverage points. You can use robust regression to identify observations that are outliers and high leverage points. Once these observations are identified, they can be reweighted or excluded from nonrobust analyses.

You can run a Robust Regression analysis by selecting **Analysis ►Model Fitting ►Robust Regression** from the main menu. The computation of the robust regression function and the identification of outliers and leverage points are implemented by calling the ROBUSTREG procedure in SAS/STAT software. See the documentation for the ROBUSTREG procedure in the *SAS/STAT User's Guide* for additional details.

Example: Fit a Robust Regression Model

The example in Chapter 21, "Model Fitting: Linear Regression," models 1987 salaries of Major League Baseball players as a function of several explanatory variables in the Baseball data set by using ordinary least squares regression. In that example, two conclusions are reached:

- no_home, the number of home runs is not a significant variable in the model.
- Several players are high leverage points. Pete Rose has the highest leverage because of his 25 years in the major leagues. Graig Nettles and Steve Sax are leverage points and also outliers.

However, the model fitted by using ordinary least squares is influenced by high leverage points and outliers. Robust regression is a preferable method of detecting influential observations. This example uses the Robust Regression analysis to identify leverage points and outliers in the Baseball data.

To model the logarithm of salary by using no_hits and yr_major as explanatory variables:

1 Open the Baseball data set.

The following two steps are the same as for the example in the "Example: Fit a Linear Regression Model" section in Chapter 21:

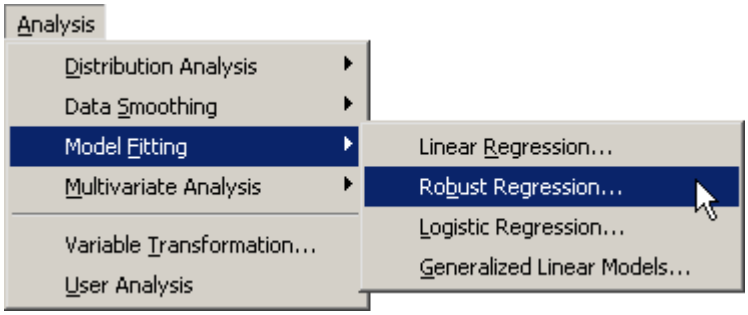
2 Use the Variable Transformation Wizard to create a new variable, Log10_salary, which contains the logarithmic transformation of the salary variable.

3 Choose name to be the label variable for these data.

The following steps model Log10_salary as a function of two explanatory variables.

4 Select **Analysis ►Model Fitting ►Robust Regression** from the main menu, as shown in Figure 22.1.

Figure 22.1 Selecting a Robust Regression

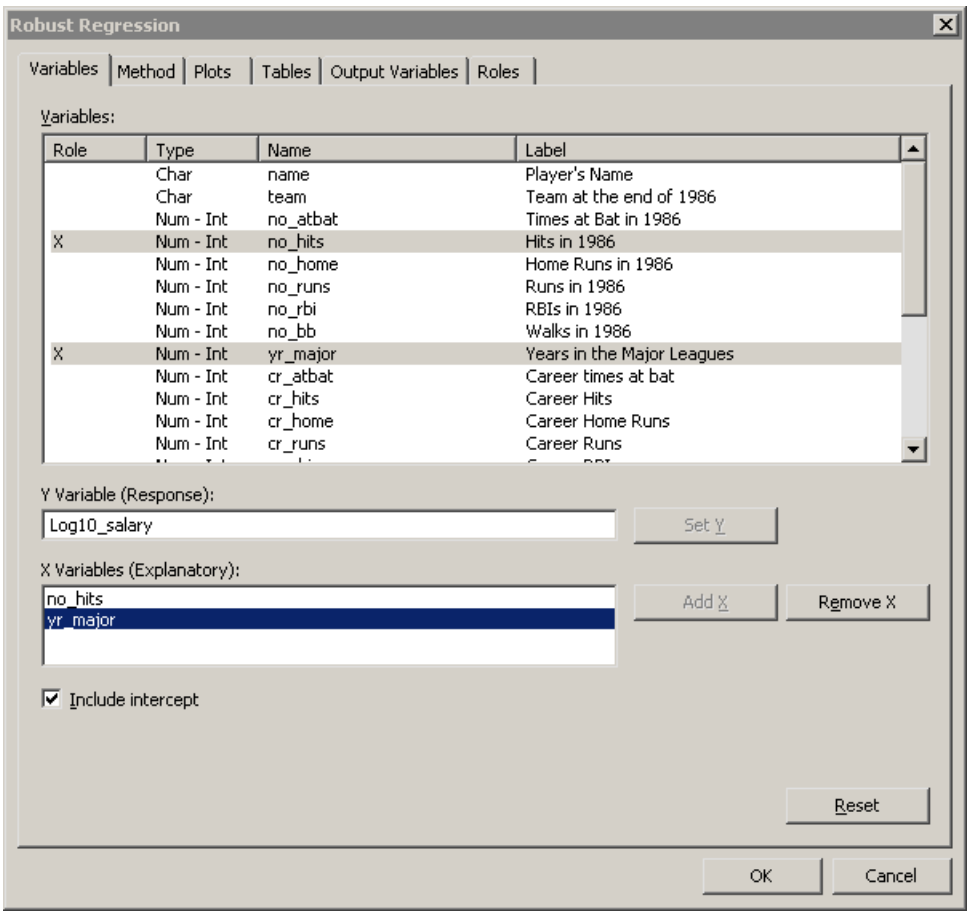


The Robust Regression dialog box appears. (See Figure 22.2.)

5 Scroll to the end of the variable list. Select the Log10_salary, and click **Set Y**.

6 Select no_hits. While holding down the CTRL key, select yr_major. Click **Add X**.

Figure 22.2 The Variables Tab



7 Click the **Method** tab.

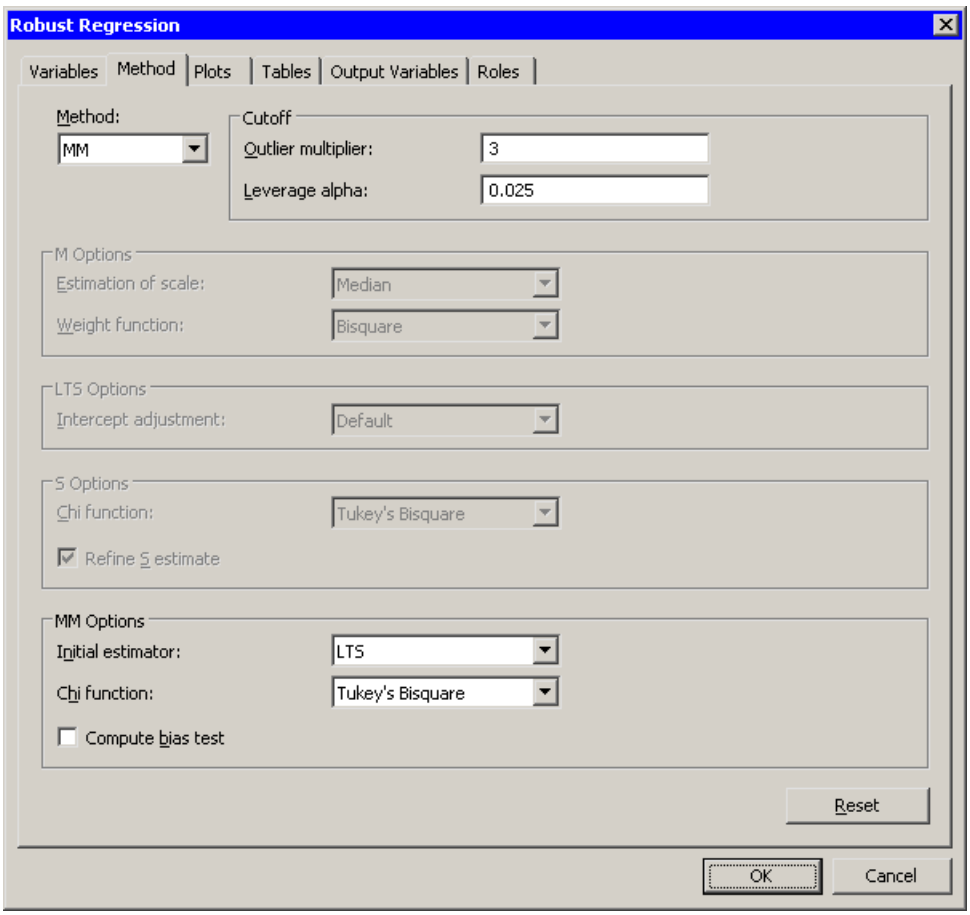
The **Method** tab becomes active, as shown in Figure 22.3. There are four robust estimation methods. The default method, known as *M estimation*, is not robust in the presence of high leverage points. The LTS and MM methods are better suited for handling high leverage points.

8 Select **MM** for the method.

NOTE: If you use M estimation on data that contain leverage points, the ROBUSTREG procedure prints the following message to the error log:

> WARNING: The data set contains one or more high leverage points, for which M estimation is not robust. It is recommended that you use METHOD=LTS or METHOD=MM for this data set.

Figure 22.3 The Method Tab



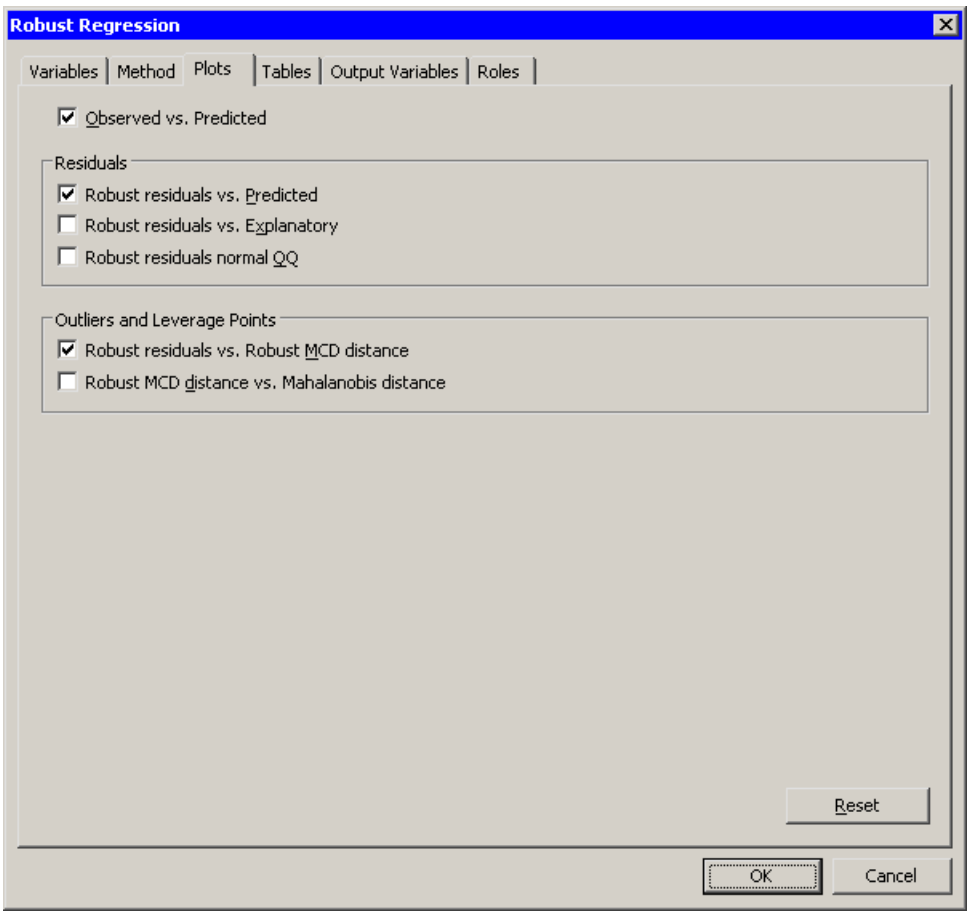
9 Click the **Plots** tab.

The **Plots** tab becomes active, as shown in Figure 22.4. This tab controls which graphs are produced by the analysis. One plot is selected by default. For this example, select the following additional plots:

10 Select **Observed vs. Predicted**.

11 Select **Robust residuals vs. Predicted**.

Figure 22.4 The Plots Tab



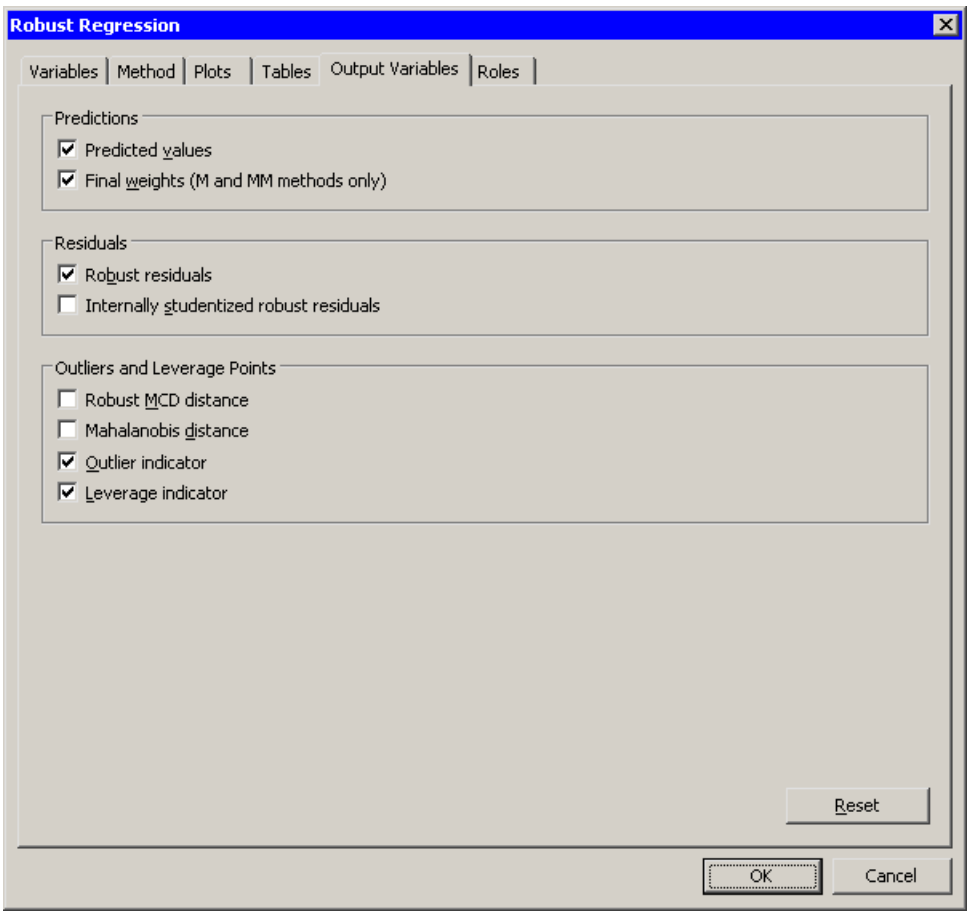
12 Click the **Output Variables** tab.

The **Output Variables** tab becomes active, as shown in Figure 22.5. This tab controls which analysis variables are added to the data table.

13 Select **Final Weights (M and MM methods only)**.

Note that the **Outlier indicator** and **Leverage indicator** options are selected by default. These options create indicator variables in the data table that you can use to identify outliers and leverage points.

Figure 22.5 The Output Variables Tab



14 Click **OK** to run the analysis.

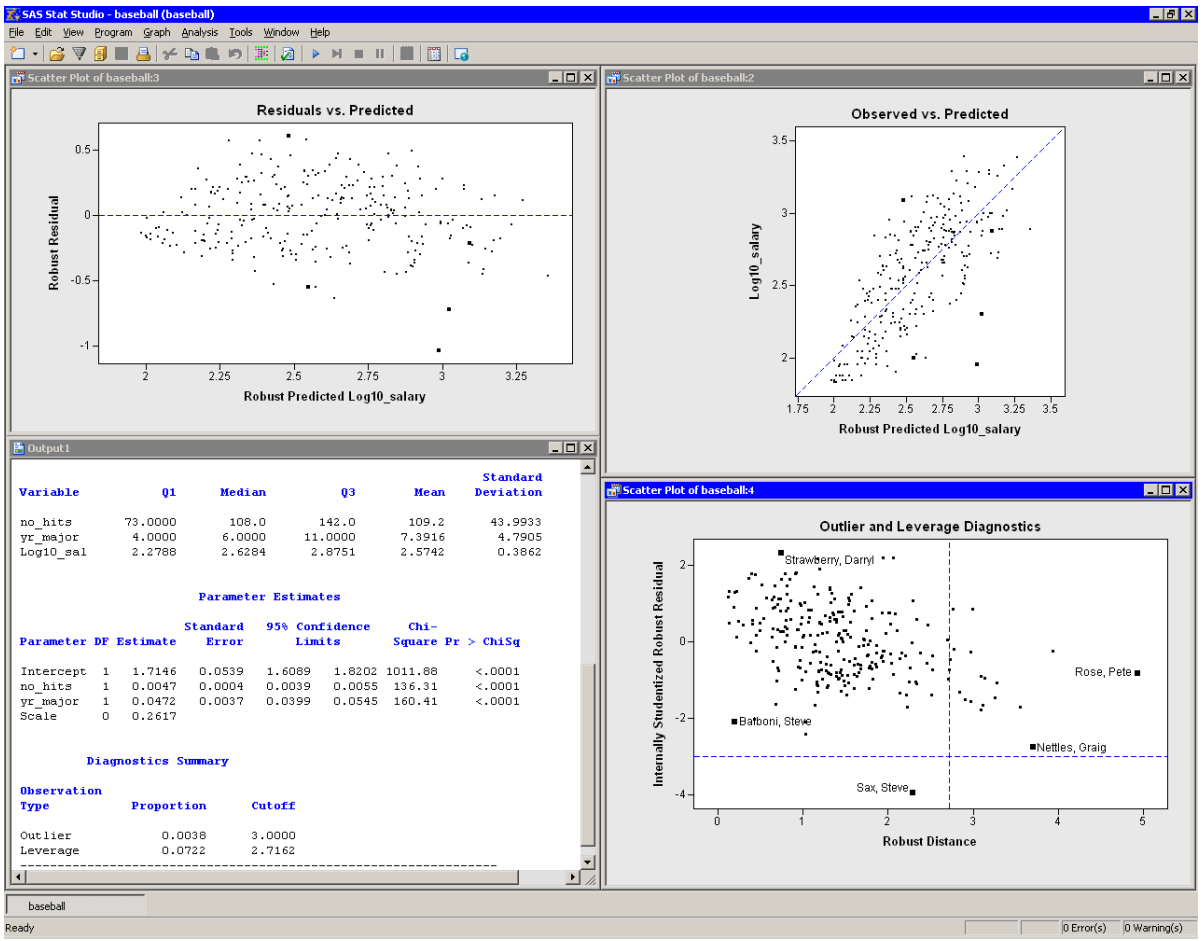
Several plots appear, along with output from the ROBUSTREG procedure. Some plots might be hidden beneath others. Move the windows so that they are arranged as in Figure 22.6. In the figure, five players are selected to facilitate comparison with Figure 21.9 and Figure 21.12.

The plots that display predicted values are similar to those in Figure 21.9. The Residuals vs. Predicted plot does not show any obvious trends. The Observed vs. Predicted plot shows a reasonable fit, with a few exceptions.

The plot of (internally) studentized robust residuals versus robust distance (known as an *RD plot*) identifies which observations are outliers and which are high leverage points. Observations outside the horizontal lines at ± 3 are outliers; observations to the right of the vertical line at 2.7162 are leverage points. The values of the outlier and leverage cutoffs are displayed in the "Diagnostics Summary" table in the output window. You can control these values from the **Method** tab.

The robust regression model identifies Steve Sax as an outlier and identifies 19 other players (including Pete Rose and Graig Nettles) as leverage points. As displayed in the "Diagnostics Summary" table, these 19 players represent 7.2% of the 263 observations used in the analysis. (For comparison, the analysis in Chapter 21, "Model Fitting: Linear Regression," suggests 11 outliers and 16 leverage points.)

Figure 22.6 Results from the Robust Regression Analysis



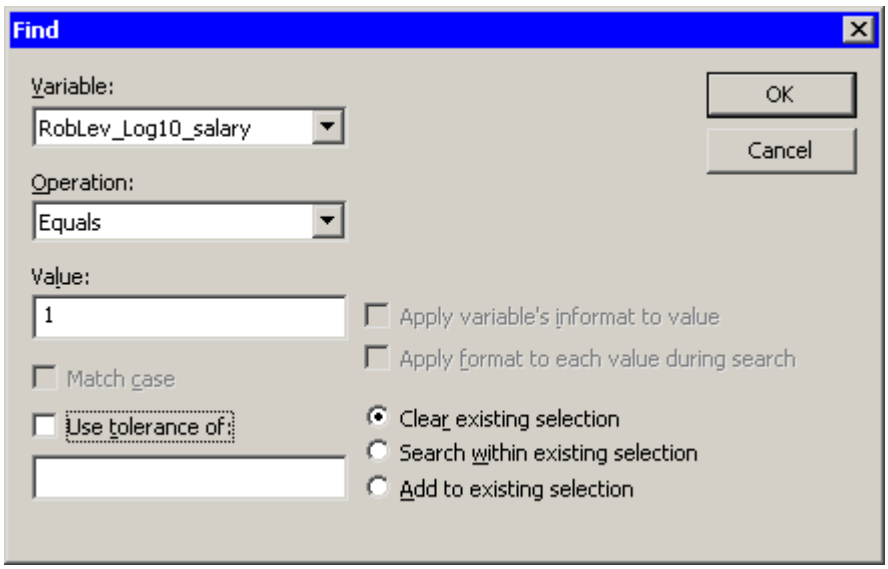
Using the Results of Robust Regression

Frequently, robust regression is used to identify outliers and leverage points.

You can easily select outliers and leverage points by using the mouse to select observations in the RD plot, or by using the Find dialog box. (You can display the Find dialog box by selecting **Edit ►Find** from the main menu.) The analysis added two indicator variables to the data table. The variable RobLev_Log10_salary has the value 1 for observations that are high leverage points. The variable RobOut_Log10_salary has the value 1 for the single observations that is an outlier.

Figure 22.7 shows how you can select all of the leverage points. After the observations are selected, you can examine their values, exclude them, change the shapes of their markers, or otherwise give them special treatment.

Figure 22.7 Finding Leverage Points



Similarly, you can select outliers.

To indicate a typical analysis of data that are contaminated with outliers:

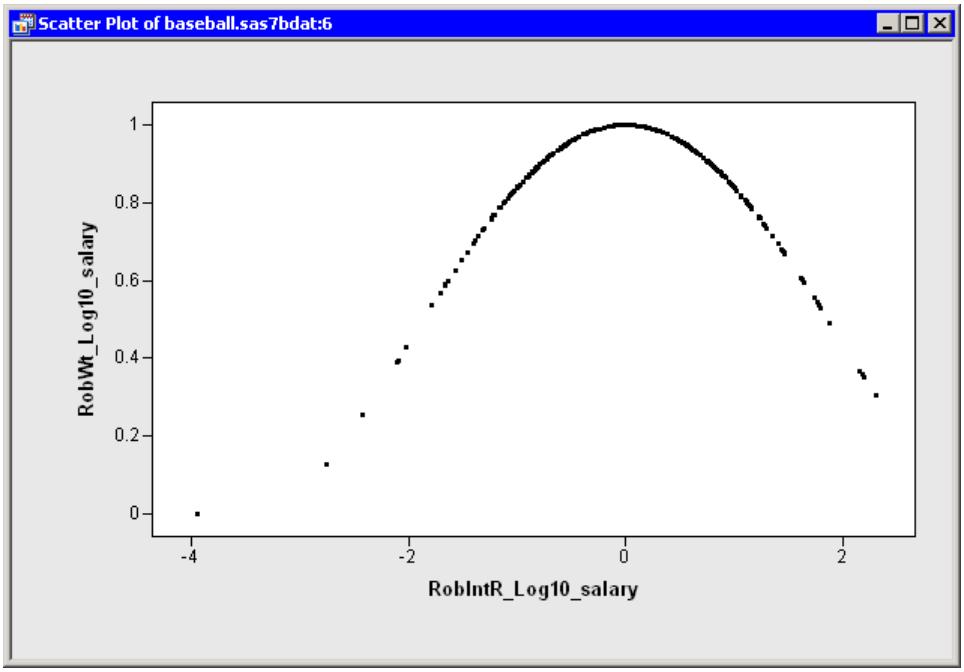
1. Examine the outliers.
2. If it makes sense to exclude the observation from future analyses, select **Edit ►Observations ►Exclude from Analyses** from the main menu.
3. Use ordinary least squares regression to model the data without the presence of outliers.

NOTE: You can select **Final least squares estimates after excluding outliers** on the **Tables** tab. The parameter estimates in this table are the ordinary least squares estimates after excluding outliers.

A second approach involves using the "Final Weights" variable that you requested on the **Output Variables** tab. The MM method uses an iteratively reweighted least squares algorithm to compute the final estimate, and the RobWt_Log10_salary variable contains the final weights.

Figure 22.8 shows the relationship between the weights and the studentized residuals. The graph shows that observations with large residuals (in absolute value) receive little or no weight during the reweighted least squares algorithm. In particular, Steve Sax receives no weight, and so his salary was not used in computing the final estimate. For this example, Tukey's bisquare function was used for the χ function in the MM method; if you use the Yohai function instead, Figure 22.8 looks different.

Figure 22.8 Weights versus Studentized Residuals



You can use the final weights to duplicate the parameter estimates by using ordinary least squares regression. For example, if you run the REG procedure on the Baseball data and use RobWt_Log10_salary as a WEIGHT variable, you get approximately the same parameter estimates table as displayed by the ROBUSTREG procedure:

$$\log_{10}(\text{salary}) = 1.7146 + 0.0047\ \text{no_hits} + 0.0472\ \text{yr_major}$$

Specifying the Robust Regression Analysis

This section explains the dialog box tabs that are associated with the Robust Regression analysis. The Robust Regression analysis calls the ROBUSTREG procedure in SAS/STAT software. See the ROBUSTREG documentation in the *SAS/STAT User's Guide* for details.

Variables Tab

You can use the **Variables** tab to specify the variables for the Robust Regression analysis.

The **Variables** tab is shown in Figure 22.2. The Y variable is the response variable. The dialog box supports multiple X (explanatory) variables. All X and Y variables must be interval variables: the analysis does not support choosing a nominal classification variable.

Method Tab

You can use the **Method** tab to specify options for one of four robust regression algorithms:

The **Method** tab is shown in Figure 22.3. Each of the following options corresponds to an option in the ROBUSTREG procedure.

Method
specifies the algorithm used for the robust regression. The choices are M, LTS, S, and MM. This corresponds to the METHOD= option in the PROC ROBUSTREG statement.

Outlier multiplier
specifies the multiplier of the robust estimate of scale to use for outlier detection. This corresponds to the CUTOFF= option in the MODEL statement.

Leverage alpha
specifies a cutoff value for leverage-point detection. This corresponds to the CUTOFFALPHA= suboption of the LEVERAGE option in the MODEL statement.

The various methods each have options that are associated with them. When you select a method, the relevant options become active.

Options with Method=M

With METHOD=M, you can specify the following additional suboptions:

Estimation of scale
specifies a method for estimating the scale parameter. This corresponds to the SCALE= option.

Weight function
specifies the weight function used for the M estimate. This corresponds to the WF= option.

Options with Method=LTS

With METHOD=LTS, you can specify the following additional suboptions:

Intercept adjustment
specifies the intercept adjustment method in the LTS algorithm. Choosing “Default” corresponds to omitting the IADJUST= option. The other choices correspond to IADJUST=ALL or IADJUST=NONE.

Options with Method=S

With METHOD=S, you can specify the following additional suboptions:

Chi function
specifies the choice of the χ function for the S estimator. This corresponds to the CHIF= option.

Refine S estimate
specifies whether to refine for the S estimate. This corresponds to the NOREFINE option.

Options with Method=MM

With METHOD=MM, you can specify the following additional suboptions:

Initial estimator
specifies the initial estimator for the MM estimator. This corresponds to the INITEST= option.

Chi function
specifies the choice of the χ function for the MM estimator. This corresponds to the CHIF= option.

Compute bias test
specifies whether to display the bias test for the final MM estimate. This corresponds to the BIASTEST option.

Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 22.4.) There are plots that help you to visualize the fit, the residuals, and various influence diagnostics.

Creating a plot often adds one or more variables to the data table. The following plots are available:

Observed vs. Predicted
creates a scatter plot of the Y variables versus the predicted values, overlaid with the diagonal line that represents a perfect fit.

Robust residuals vs. Predicted
creates a scatter plot of the residuals versus the predicted values.

Robust residuals vs. Explanatory
creates scatter plots of the residuals versus the X variables.

Residual normal QQ
creates a normal Q-Q plot of the residuals.

Robust residuals vs. Robust MCD distance
creates a scatter plot of the internally studentized residuals versus the *robust distance*. The robust distance is a measure of the distance between an observation and a robust estimate of location. The distance function uses robust estimates of scale and location computed by the minimum covariance determinant (MCD) method.

Robust MCD distance vs. Mahalanobis distance
creates a scatter plot of the robust distance versus the Mahalanobis distance.

Tables Tab

The **Tables** tab is shown in Figure 22.9. You can use the **Tables** tab to display the following tables that summarize the results of the analysis:

Summary statistics
displays summary statistics for model variables. The statistics include robust estimates of the location and scale for each variable.

Parameter estimates
displays estimates for the model parameters.

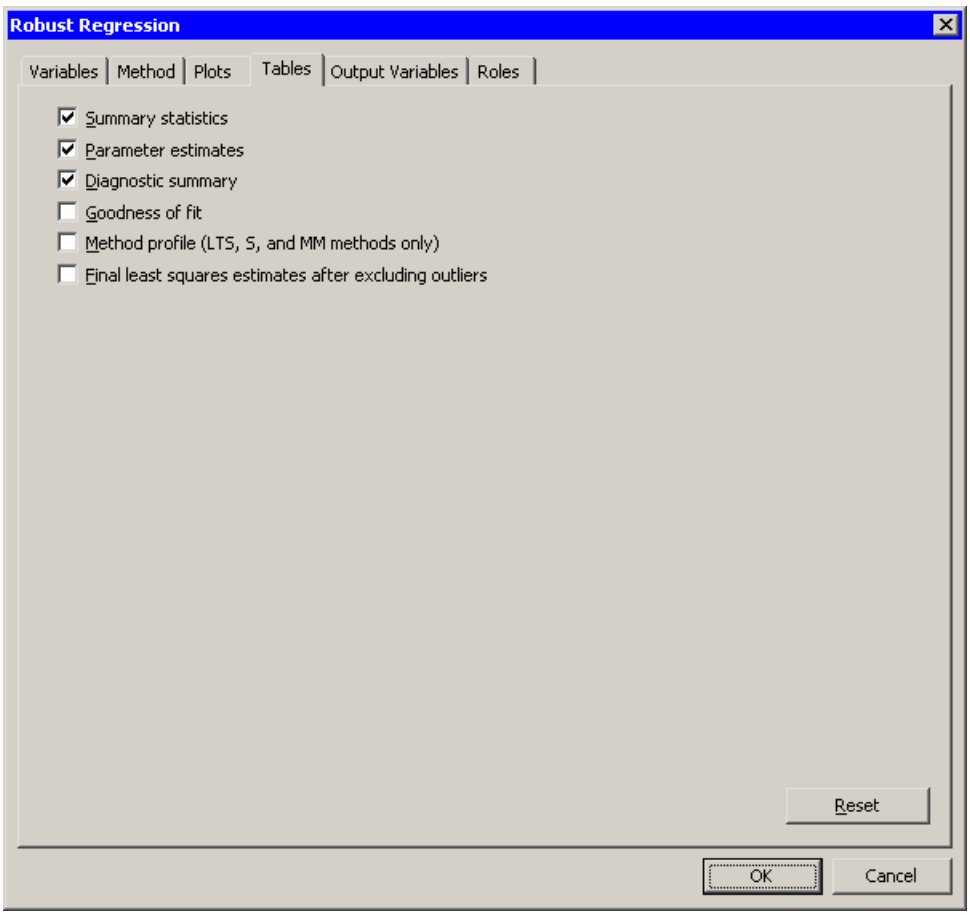
Diagnostics summary
displays a summary of the outlier and leverage diagnostics.

Goodness of fit
displays goodness-of-fit statistics.

Method profile (LTS, S, and MM methods only)
displays a summary of the options used by the method.

Final least squares estimates after excluding outliers
displays least squares estimates computed after deleting the detected outliers. This corresponds to the FWLS option in the PROC ROBUSTREG statement. The parameter estimates reported in this table are the same as the estimates you get if you exclude the outliers reported by ROBUSTREG and then run the REG procedure on the remaining observations.

Figure 22.9 The Tables Tab



Output Variables Tab

You can use the **Output Variables** tab to add analysis variables to the data table. (See Figure 22.5.) If you request a plot that uses one of the output variables, then that variable is automatically created even if you did not explicitly select the variable on the **Output Variables** tab.

The following list describes each output variable and indicates how the output variable is named. Y represents the name of the response variable.

Predicted values
adds predicted values. The variable is named RobP_Y.

Final weights (M and MM methods only)
adds the final weights used in the iteratively reweighted least squares algorithm. The variable is named RobWt_Y.

Robust residuals
adds residuals, which are calculated as observed values minus predicted values. The variable is named RobR_Y.

Internally studentized robust residuals
adds internally studentized residuals, which are the residuals divided by their standard errors. The variable is named RobIntR_Y.

Robust MCD distance
adds a robust measure of distance between an observation and a robust estimate of location. The variable is named RobRD_Y.

Mahalanobis distance
adds the Mahalanobis distance between an observation and the multivariate mean of the data. The variable is named RobMD_Y.

Outlier indicator
adds an indicator variable for outliers. The variable is named RobOut_Y.

Leverage indicator
adds an indicator variable for leverage points. The variable is named RobLev_Y.

Roles Tab

You can use the **Roles** tab to specify a weight variable for the analysis.

A weight variable is a numeric variable with values that are relative weights for the regression.

Analysis of Selected Variables

If one or more interval variables are selected in a data table when you run the analysis, then the following occurs:

- The first selected interval variable is automatically entered in the **Y Variable** field of the **Variables** tab.
- The remaining selected interval variables are automatically entered in the **X Variable** field of the **Variables** tab.

Any variable in the data table with a Weight role is automatically entered in the appropriate field of the **Roles** tab.

Chapter 23
Model Fitting: Logistic Regression

Contents

Overview of the Logistic Regression Analysis **339**
Example: Fit a Logistic Regression Model **340**
Specifying the Logistic Regression Analysis **345**
Variables Tab 345
Effects Tab 346
Method Tab 353
Plots Tab 353
Tables Tab 355
Output Variables Tab 356
Roles Tab 358
Analysis of Selected Variables **359**

Overview of the Logistic Regression Analysis

The Logistic Regression analysis fits a logistic regression model by using the method of maximum likelihood estimation.

If X_i are explanatory variables and p is the response probability to be modeled, the logistic model has the form

$$\log(p/(1-p)) = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m$$

where the b_i are regression coefficients.

The explanatory variables in the Logistic Regression analysis can be interval variables or nominal variables (also known as *classification variables*). You can also specify more complex model terms such as interactions and nested terms. Any term specified in the model is referred to as an *effect*, whether it is the main effect of a variable, or a classification variable, or an interaction, or a nested term.

You can run a Logistic Regression analysis by selecting **Analysis ►Model Fitting ►Logistic Regression** from the main menu. The computation of the estimated regression coefficients, confidence limits, and related statistics is implemented by calling the LOGISTIC procedure in SAS/STAT software. See the documentation for the LOGISTIC procedure in the *SAS/STAT User's Guide* for additional details.

Example: Fit a Logistic Regression Model

Neuralgia is pain that follows the path of specific nerves. Neuralgia is most common in elderly persons, but it can occur at any age. In this example, you use a logistic model to compare the effects of two test treatments and a placebo on a dichotomous response: whether or not the patient reported pain after the treatment. In particular, the example examines three explanatory variables:

- Treatment, the administered treatment. This variable has three values: A and B represent the two test treatments, and P represents the placebo treatment.
- Sex, the patient gender
- Age, the patient's age, in years, when treatment began

Some questions that you might ask regarding these data include the following:

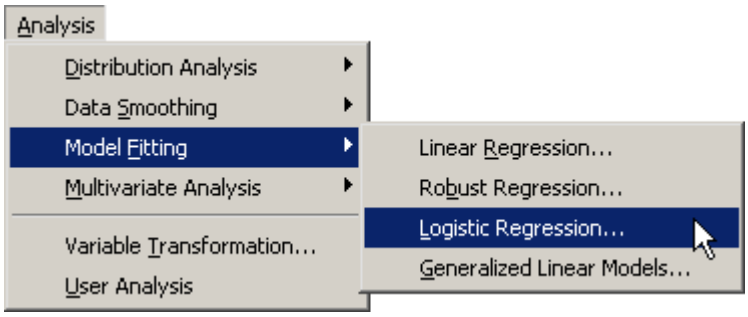
- Is either treatment better than the placebo at reducing neuralgia?
- How does age or gender affect the results?

The following steps answer these questions:

1 Open the Neuralgia data set.

2 Select **Analysis ►Model Fitting ►Logistic Regression** from the main menu, as shown in Figure 23.1.

Figure 23.1 Selecting a Logistic Regression



The Logistic Regression dialog box appears. (See Figure 23.2.)

You can model the probability that a patient reports no pain after treatment in order to determine whether the treatments are effective.

3 Select Pain, and click **Add Y**.

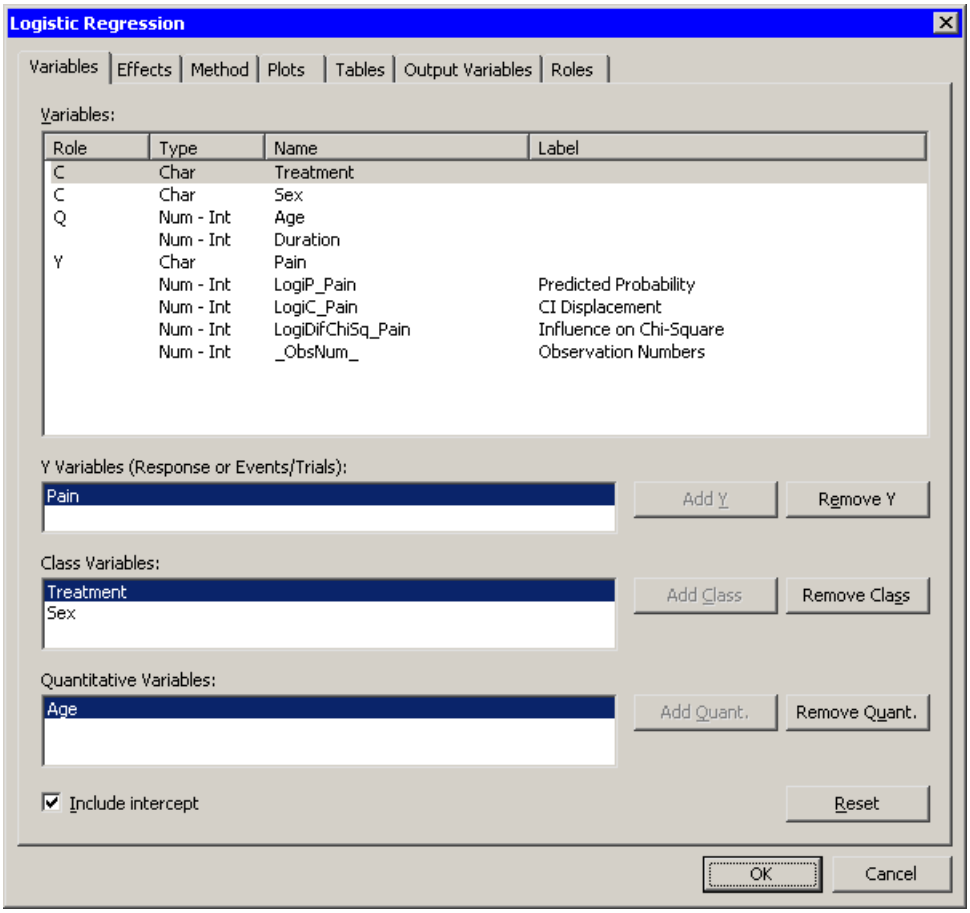
The Treatment and Sex variables are both classification variables, whereas Age is a quantitative (that is, interval) variable.

4 Select Treatment. While holding down the CTRL key, select Sex. Click **Add Class**.

5 Select Age, and click **Add Quant.**

NOTE: Alternatively, you can double-click a variable to automatically add it as an explanatory variable. Nominal variables are automatically added as classification variables; interval variables are automatically added as quantitative variables.

Figure 23.2 The Variables Tab



6 Click the **Method** tab.

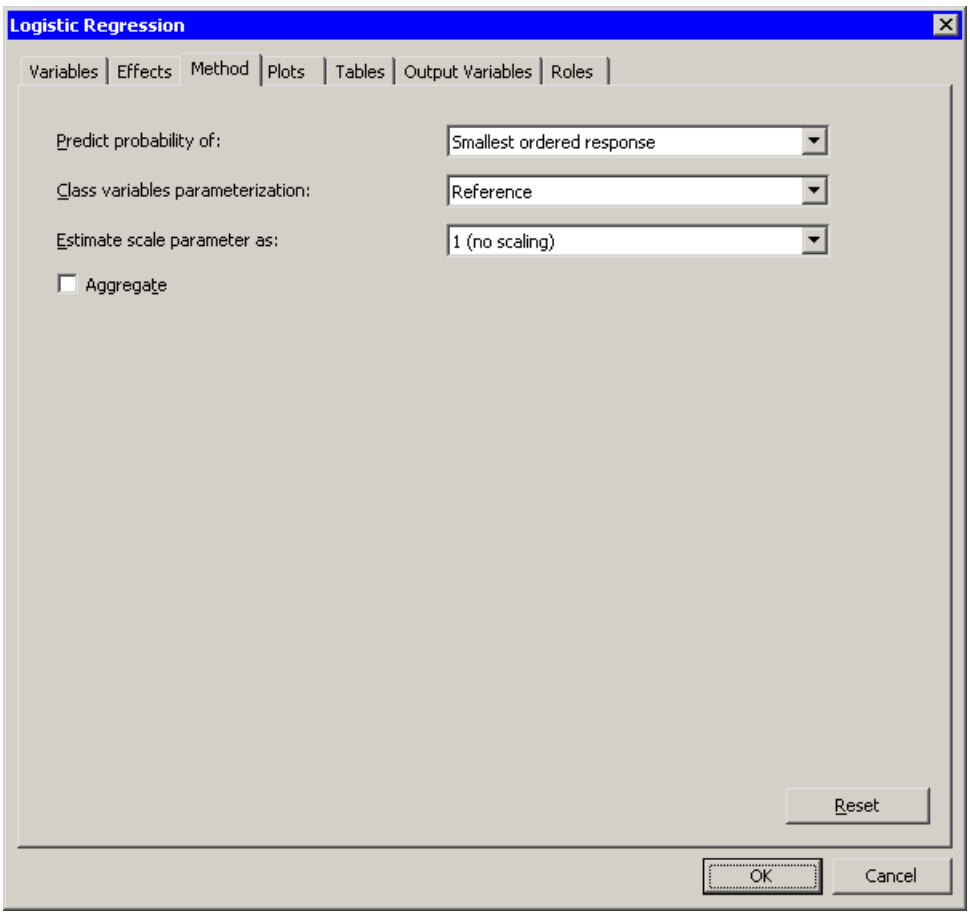
The **Method** tab becomes active, as shown in Figure 23.3. You can use this tab to set options for the analysis.

The first option on this tab specifies whether the analysis predicts the probability of the smallest ordered response. The responses for this example are "Yes" and "No." Since "No" precedes "Yes" in alphabetical ordering, the smaller ordered response is "No." This example predicts the probability that a patient will report no pain.

This example includes data for a placebo treatment. It is easier to interpret the parameters of the model if you choose a reference parameterization for the coding of the classification variable. (For further details on parameterizations, see the section "CLASS Variable Parameterization" in the "Details" section of the documentation for the LOGISTIC procedure.)

7 Select **Reference** for the **Classification variables parameterization** option.

Figure 23.3 The Method Tab



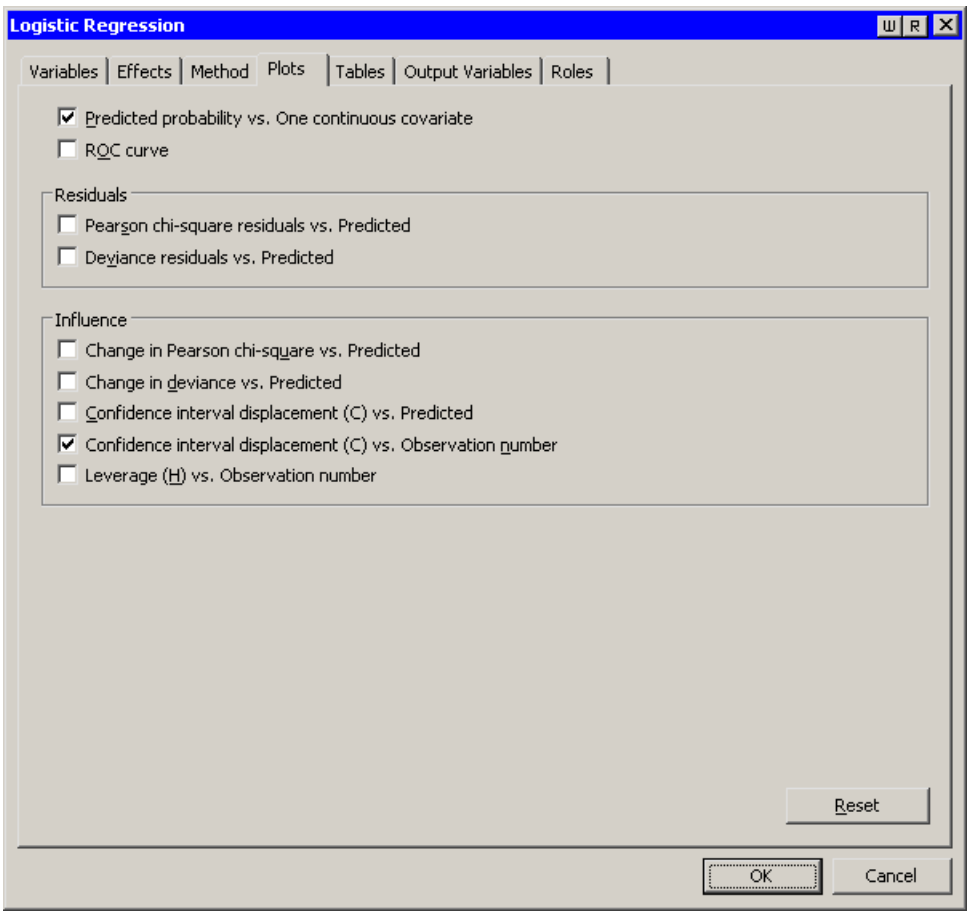
8 Click the **Plots** tab.

The **Plots** tab becomes active, as shown in Figure 23.4. This tab controls which graphs are produced by the analysis.

By default, the analysis creates three plots.

9 Clear **Change in Pearson chi-square residuals vs. Predicted** to reduce the number of plots that the analysis creates.

Figure 23.4 The Plots Tab



10 Click **OK**.

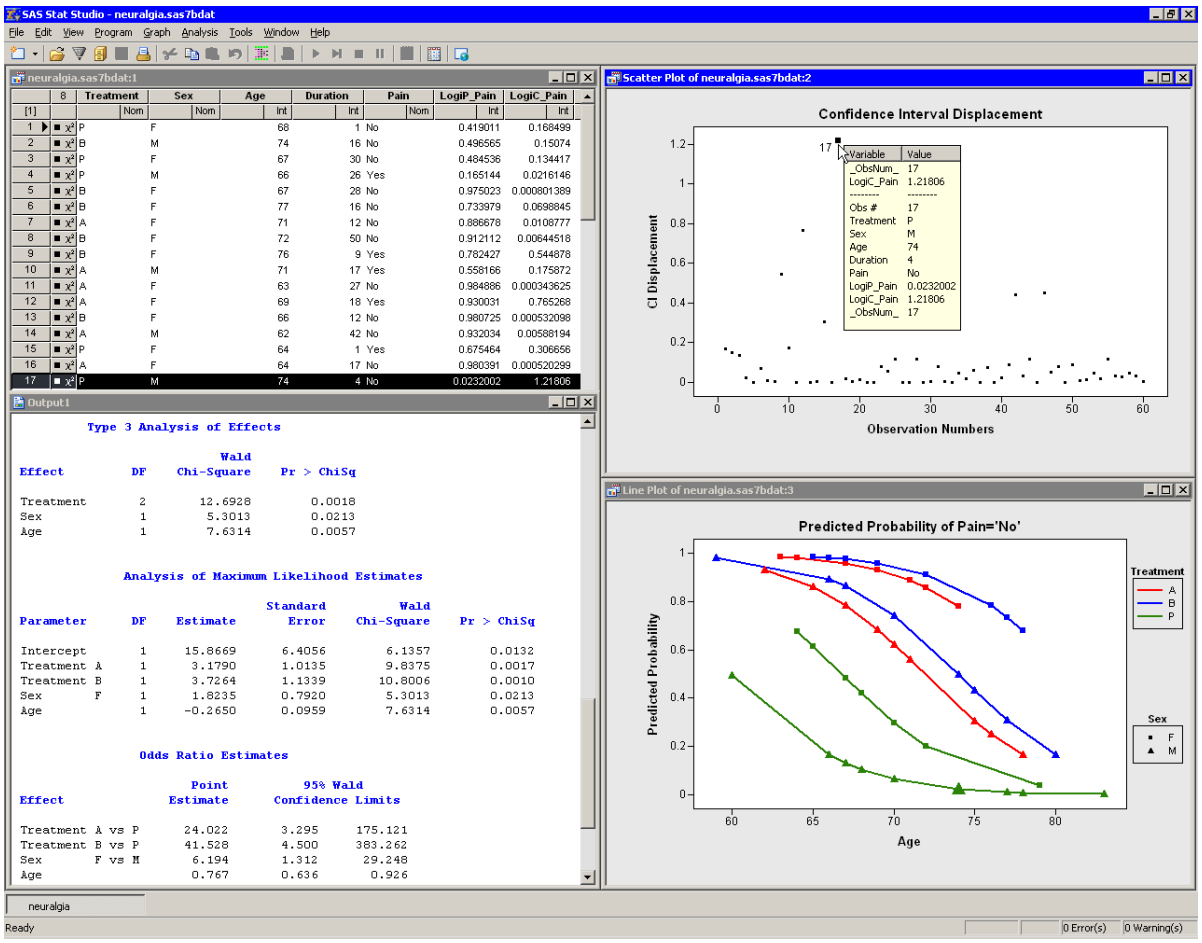
Two plots appear, along with output from the LOGISTIC procedure. One plot might be hidden beneath the other. Move the plots so that they are arranged as in Figure 23.5.

The tables created by the LOGISTIC procedure appear in the output window. The "Model Fit Statistics" table indicates that the model with the specified explanatory variables is preferable to an intercept-only model. The "Type 3 Analysis of Effects" table indicates that all explanatory variables in this model are significant.

The "Analysis of Maximum Likelihood Estimates" table displays estimates for the parameters in the logistic model. The p-values for Treatment A and B (0.0017 and 0.0010, respectively) indicate that these treatments are significantly better at treating neuralgia than the placebo. The negative estimate for the age effect indicates that older patients in the study responded less favorably to treatment than younger patients.

The "Odds Ratio Estimate" table enables you to quantify how changes in an explanatory variable affect the likelihood of the response outcome, assuming the other variables are fixed.

Figure 23.5 Results from the Logistic Regression Analysis



For an interval explanatory variable, the odds ratio approximates how much a unit change in the explanatory variable affects the likelihood of the outcome. For example, the estimate for the odds ratio for Age is 0.767. This indicates that the outcome of eliminating neuralgia occurs only 77% as often among patients of age $x + 1$, as compared with those of age x. In other words, neuralgia in older patients is less likely to go away than neuralgia in younger patients.

For a categorical explanatory variable, the odds ratio compares the odds for the outcome between one level of the explanatory variable and the reference level. The estimate of the odds ratio for treatment A is 24.022. This means that eliminating neuralgia occurs 24 times as often among patients that receive treatment A as among those receiving the placebo. Similarly, eliminating neuralgia occurs more than 41 times as often in patients that receive treatment B, compared to the placebo patients. In the same way, eliminating pain occurs six times more often in females than in males. For a detailed description of how to interpret the odds ratio,

including a discussion of various parameterization schemes, see the "Odds Ratio Estimation" section of the documentation for the LOGISTIC procedure.

The results of the analysis are summarized by the line plot of predicted probability versus Age. Each line corresponds to a joint level of Treatment and Sex. The line colors indicate levels of Treatment; marker shapes indicate gender.

The line plot graphically illustrates a few conclusions from the "Analysis of Maximum Likelihood Estimates" table:

- Given a gender and an age, treatment A and treatment B are better at treating neuralgia than the placebo.
- Given a treatment and an age, females tend to report less pain than males.
- The efficacy of the treatments decreases with the age of the patient.

This analysis did not include an interaction term between treatment and gender, so no conclusions are possible regarding whether the treatments affect pain differently in men and women. Also, this analysis did not compare treatment A with treatment B.

The other graph in Figure 23.5 plots the confidence interval (CI) displacement diagnostic versus the observation numbers. The CI displacement measures the influence of individual observations on the regression estimates. Observations with large CI displacement values are influential to the prediction. Often these observations are outliers for the model.

For example, the observation with the largest CI displacement value is selected in Figure 23.5. (You can double-click an observation to display the observation inspector, described in Chapter 8, "Interacting with Plots.") This patient is a 74-year-old male who was given a placebo. He reported no pain after the treatment, in spite of the fact that the model predicts only a 2% probability that this would happen. The patient with the next largest CI displacement value (not selected in the figure) was a 69-year-old female receiving treatment A. She reported that her pain persisted, although the model predicted a 93% probability that she would not report pain.

Specifying the Logistic Regression Analysis

This section describes the dialog box tabs that are associated with the Logistic Regression analysis. The Logistic Regression analysis calls the LOGISTIC procedure in SAS/STAT software. See the LOGISTIC documentation in the *SAS/STAT User's Guide* for details.

Variables Tab

You can use the **Variables** tab to specify the variables for the Logistic Regression analysis. The **Variables** tab is shown in Figure 23.2.

The analysis handles two types of models. For *single-trial syntax*, you specify a single binary variable as the response variable. This variable can be character or numeric. For *events/trials syntax*, you specify two

numeric variables that contain count data for a binomial experiment. The value of the first variable is the number of positive responses (or *events*). The value of the second variable is the number of *trials*.

The dialog box supports multiple explanatory variables. You can include nominal variables in the model by adding them to the **Classification variables** list. You can include interval variables in the model by adding them to the **Quantitative variables** list.

When you add an explanatory variable, that main effect is added to the **Effects** tab. You can add interaction effects and nested effects by using the **Effects** tab.

Effects Tab

You can use the **Effects** tab to add several different types of effects to your model. All effects appear in the **Effects in Model** list. You can specify the following types of effects:

- main effects
- crossed effects
- nested effects

You can also use the tab to quickly create certain standard effects: factorial effects, polynomial effects, and multivariate polynomial effects.

The notation for an effect consists of variable names, asterisks, and at most one pair of parentheses. The asterisks denote interactions; the parentheses denote nested effects. There are two rules to follow when specifying effects:

- A nominal variable can appear in an effect at most once.
- An interval variable cannot appear inside parentheses.

The following text describes how to specify effects on the **Effects** tab. In the descriptions, assume that A, B, and C are classification variables and that X and Y are interval variables.

Specifying Main Effects

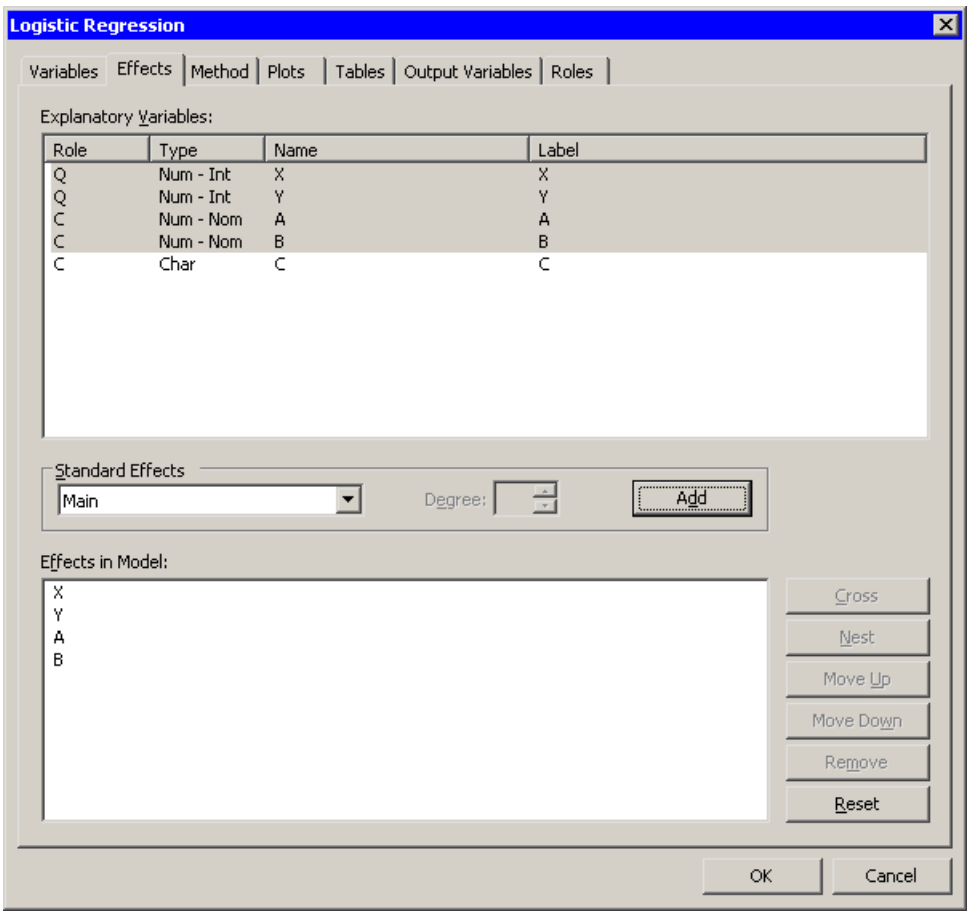
The notation for a main effect is just the name of the variable itself.

To specify a main effect:

1. Select **Main** from the **Standard Effects** list.
2. Select one or more variables from the **Explanatory Variables** list.
3. Click **Add**.

The effects are added to the **Effects in Model** list, as shown in Figure 23.6. Each main effect appears on a line by itself in the **Effects in Model** list. Because main effects are automatically added to this list when you select a variable on the **Variables** tab, you usually do not need to add main effects.

Figure 23.6 Specifying Main Effects



Specifying Crossed Effects

The notation for a crossed effect is two or more variable names that are joined by asterisks. A crossed effect can involve one or more interval variables (such as X*X and X*Y) or two or more nominal variables (such as A*B, B*C, and A*B*C). You cannot cross a nominal variable with itself, but you can create crossed effects that involve both interval variables and nominal variables, such as X*A.

To specify a crossed effect in which each variable appears once (such as X*Y):

1 Select **Cross** from the **Standard Effects** list.

2 Select two or more variables from the **Explanatory Variables** list.

3 Click **Add**.

For example, the preceding steps were used to create the X*Y effect shown in Figure 23.7.

To cross variables with effects already in the model:

1 Select **Cross** from the **Standard Effects** list.

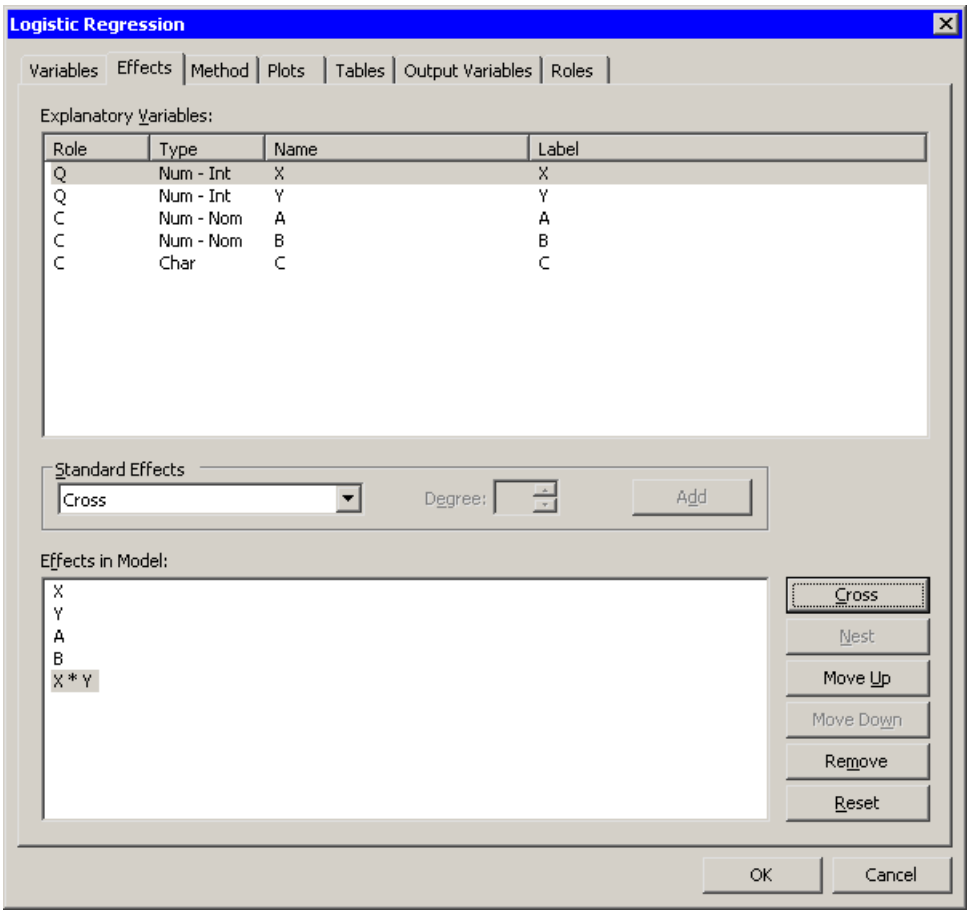
2 Select one or more variables from the **Explanatory Variables** list.

3 Select one or more effects from the **Effects in Model** list.

4 Click **Cross**, located to the right of the **Effects in Model** list.

For example, Figure 23.7 shows one way to create the effect X*X*Y. You can select the X variable from the **Explanatory Variables** list and the X*Y effect from the **Effects in Model** list. The X*X*Y effect is created when you click **Cross**.

Figure 23.7 Specifying Crossed Effects



Specifying Nested Effects

The notation for a nested effect contains two parts. The first part is a main effect or crossed effect. The second part consists of a classification variable or an interaction between classification variables. The second part is enclosed in parentheses. The main effect or crossed effect is said to be "nested within" the effects in parentheses. For example, A(B*C) means "effect A is nested within the levels of the factors B and C." The

Standard Effects value is ignored when you specify nested effects.

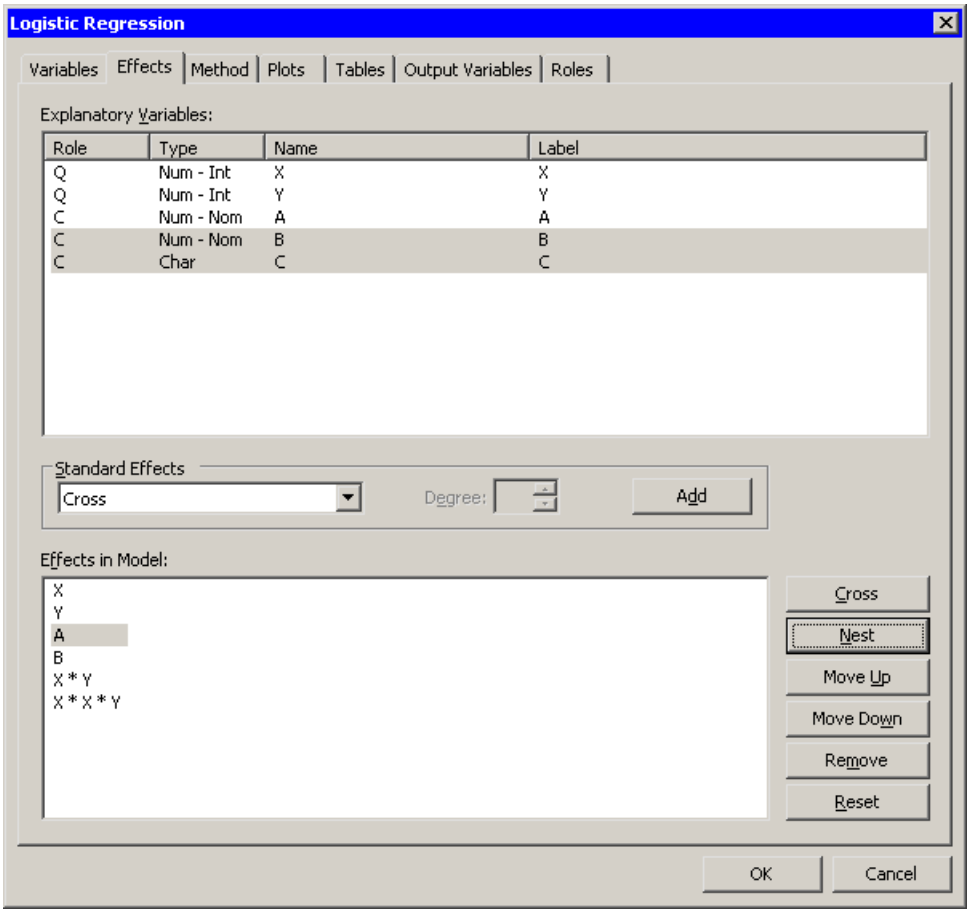
To create a nested effect, the effect outside the parentheses must already be specified in the **Effects in Model** list.

To create a nested effect:

1. Select one or more nominal variables from the **Explanatory Variables** list. These variables will appear inside the parentheses.
2. Select one or more effects from the **Effects in Model** list. These variables will appear outside the parentheses. Make sure that the nominal variables selected in the **Explanatory Variables** list do not appear in any of the effects selected in the **Effects in Model** list.
3. Click **Nest**, located to the right of the **Effects in Model** list.
4. The effects in the **Effects in Model** list are replaced with the nested effects.

For example, Figure 23.8 shows one way to create the effect A(B*C). Select the B and C variables from the **Explanatory Variables** list, and select the A main effect from the **Effects in Model** list. The A(B*C) effect is created when you click **Nest**. It replaces the A effect that is currently in the list.

Figure 23.8 Specifying Nested Effects



Specifying Factorial Effects

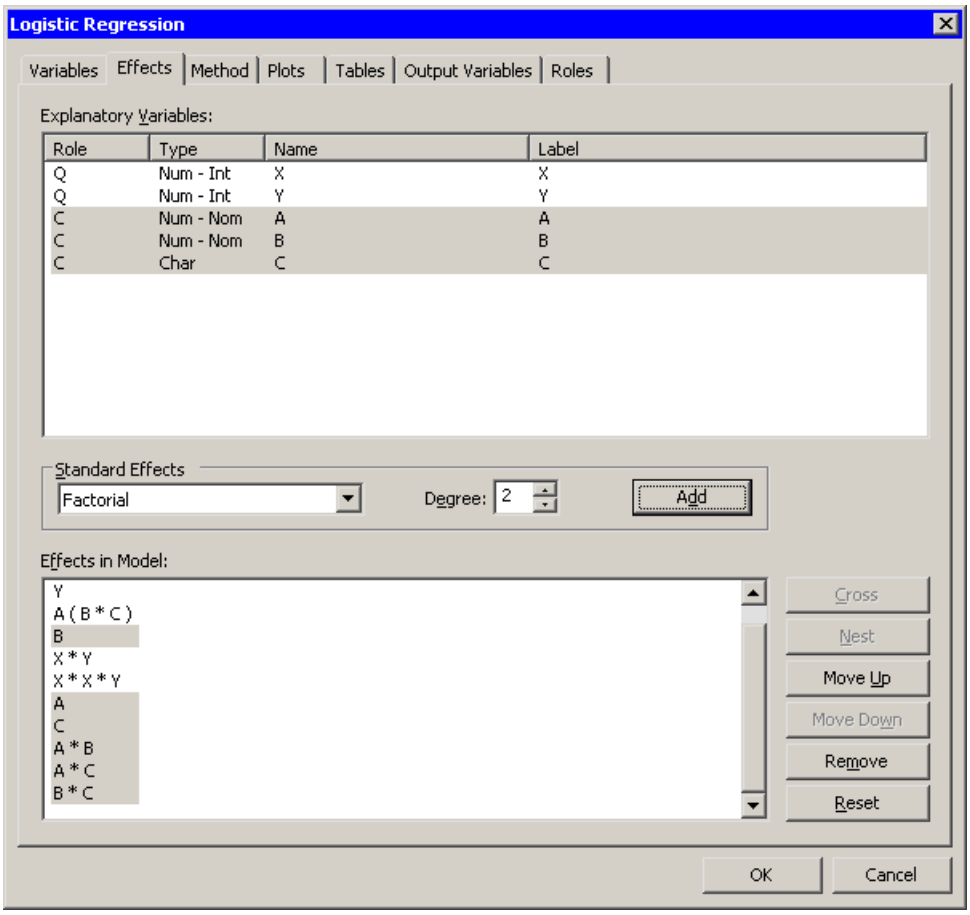
Factorial effects are k-way interactions between a set of variables.

To create factorial effects:

1. Select **Factorial** from the **Standard Effects** list.
2. Enter the **Degree** of the model.
3. Select two or more variables from the **Explanatory Variables** list.
4. Click **Add**.
5. The factorial effects are added to the **Effects in Model** list. Any effects already in the model (for example, main effects) are highlighted, although their position in the **Effects in Model** list does not change.

For example, Figure 23.9 shows how to create a full three-way factorial model with the variables A, B, and C. The following effects are added to the **Effects in Model** list: A, B, C, A*B, A*C, B*C, and A*B*C.

Figure 23.9 Specifying Factorial Interaction Effects



Specifying Polynomial Effects

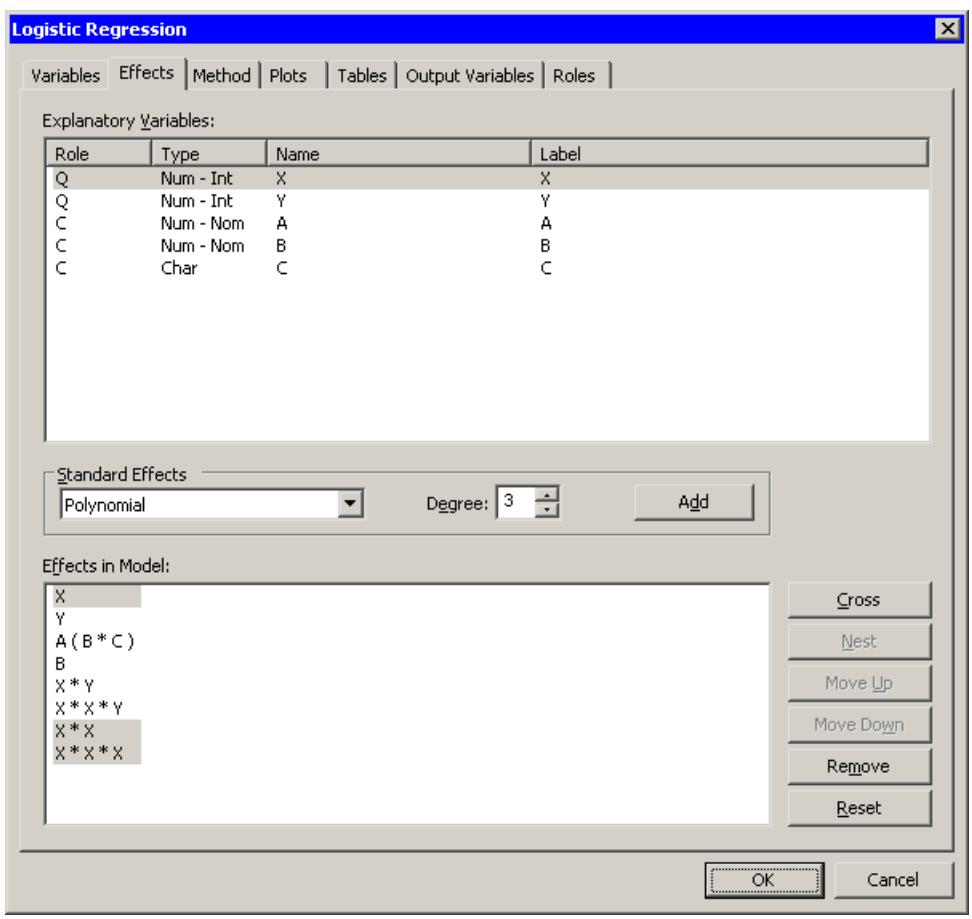
Interactions of an interval variable with itself are called *polynomial effects*. Each term is a monomial in one variable.

To create polynomial effects:

1. Select **Polynomial** from the **Standard Effects** list.
2. Enter the **Degree** of the model. (The maximum degree is 10.)
3. Select one or more variables from the **Explanatory Variables** list.
4. Click **Add**.
5. The polynomial effects are added to the **Effects in Model** list. Any effects already in the model (for example, main effects) are highlighted, although their position in the **Effects in Model** list does not change.

For example, Figure 23.10 shows how to create all terms in a degree-three polynomial in the variable X. The following effects are added to the **Effects in Model** list: X, X*X, and X*X*X.

Figure 23.10 Specifying Polynomial Effects



Specifying Multivariate Polynomial Effects

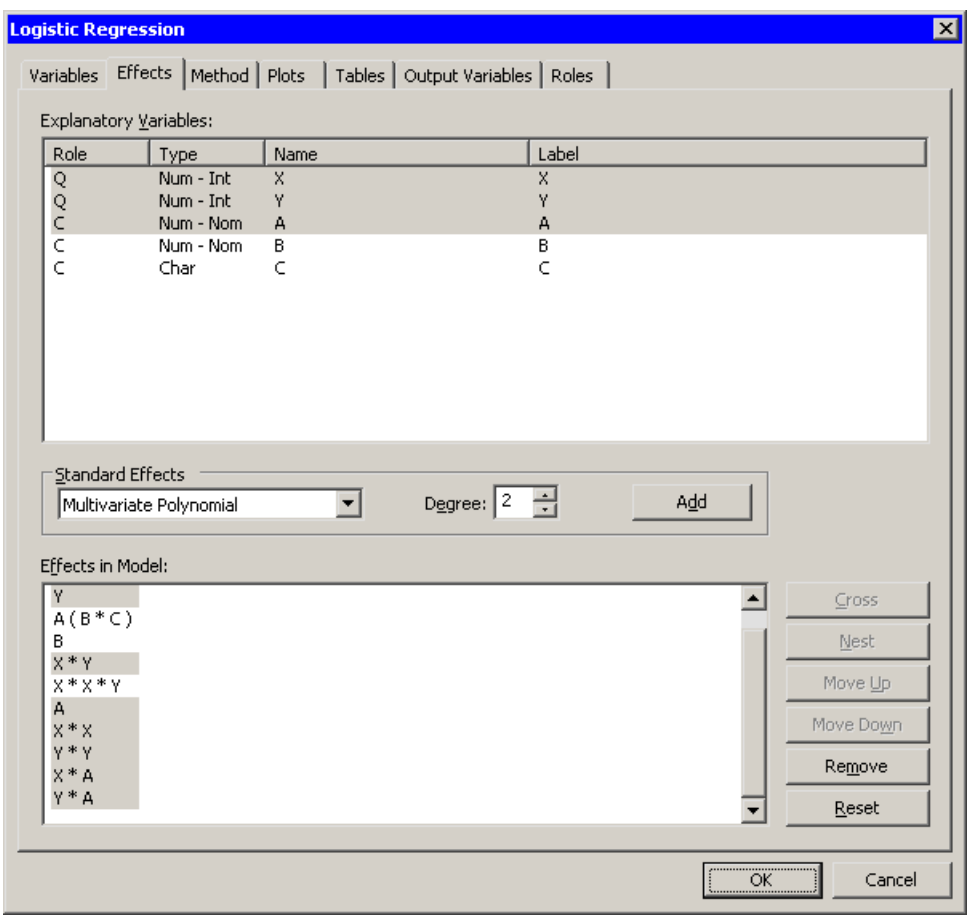
Multivariate polynomial effects are polynomial and interaction effects among a group of variables. If you select m variables and request effects from a degree-d multivariate polynomial, then each term is a multivariate monomial, with degree at most $\min(k, d)$.

To create multivariate polynomial interaction effects:

1. Select **Multivariate Polynomial** from the **Standard Effects** list.
2. Enter the **Degree** of the model. (The maximum degree is 4.)
3. Select one or more variables from the **Explanatory Variables** list.
4. Click **Add**.
5. The polynomial effects are added to the **Effects in Model** list. Any effects already in the model (for example, main effects) are highlighted, although their position in the **Effects in Model** list does not change.

For example, Figure 23.11 shows how to create all main effects and valid two-way interactions among the three variables X, Y, and A. The following effects are added to the **Effects in Model** list: X, Y, A, X*X, Y*Y, X*Y, X*A, and Y*A. The term A*A is not created because A is a classification variable.

Figure 23.11 Specifying Polynomial Interaction Effects



Reordering Effects

You can reorder and remove effects in the **Effects in Model** list. The order in which effects appear in the list is the order in which the effects appear in the MODEL statement of SAS/STAT procedures.

Move Up
moves selected effects up one position in the **Effects in Model** list.

Move Down
moves selected effects down one position in the **Effects in Model** list.

Remove
removes the selected effects from the **Effects in Model** list.

Reset
deletes all effects and then adds main effects to the **Effects in Model** list.

Method Tab

You can use the **Method** tab to set the options for the analysis. (See Figure 23.3.)

The **Method** tab contains the following UI controls:

Predict probability of
specifies whether to model the probability of the first or last level of the response variable. For example, if the response variable has levels 0 and 1, then you select **Largest ordered response** to model the probability of 1. This corresponds to the DESCENDING option in the PROC LOGISTIC statement.

Classification variables parameterization
specifies the parameterization method for the classification variables. This corresponds to the PARAM= option in the CLASS statement. The dialog box supports the GLM, effect, and reference coding schemes.

Estimate scale parameter as
specifies the method for estimating the dispersion parameter. This corresponds to the SCALE= option in the MODEL statement.

Aggregate
specifies the subpopulations on which certain test statistics are calculated. This corresponds to the AGGREGATE option in the MODEL statement.

Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 23.4.) There are plots that help you to visualize the fit, the residuals, and various influence diagnostics.

Creating a plot often adds one or more variables to the data table. The following plots are available:

Predicted probability vs. One continuous covariate

creates a line plot of the predicted probability versus the continuous explanatory variable. This plot is created only if the following conditions are satisfied:

- There is exactly one continuous explanatory variable.
- There are three or fewer classification variables.
- There are 12 or fewer joint levels of the classification variables.

ROC curve

creates a line plot that shows the trade-off between sensitivity and specificity. Models that fit the data well correspond to a receiver operating characteristic (ROC) curve that has an area close to unity. A completely random predictor would produce an ROC curve that is close to the diagonal and has an area close to 0.5.

Pearson chi-square residuals vs. Predicted

creates a scatter plot of the Pearson chi-square residuals versus the predicted probabilities.

Deviance residuals vs. Predicted

creates a scatter plot of the deviance residuals versus the predicted probabilities.

Change in Pearson chi-square vs. Predicted

creates a scatter plot of the deletion chi-square goodness-of-fit (DIFCHISQ) statistic versus the predicted probabilities.

Change in deviance vs. Predicted

creates a scatter plot of the deletion deviance (DIFDEV) statistic versus the predicted probabilities.

Confidence interval displacement (C) vs. Predicted

creates a scatter plot of the confidence interval displacement diagnostic (C) versus the predicted probabilities.

Confidence interval displacement (C) vs. Observation number
creates a scatter plot of the confidence interval displacement diagnostic (C) for each observation.

Leverage (H) vs. Observation number
creates a scatter plot of the leverage statistic for each observation.

Tables Tab

The **Tables** tab is shown in Figure 23.12. You can use the **Tables** tab to display the following tables that summarize the results of the analysis:

Simple descriptive statistics
displays a table of summary statistics for the explanatory variables.

Model fit statistics
displays a table of model fit statistics.

Generalized R-square
displays generalized R-square statistics.

Parameter estimates
displays estimates for the model parameters.

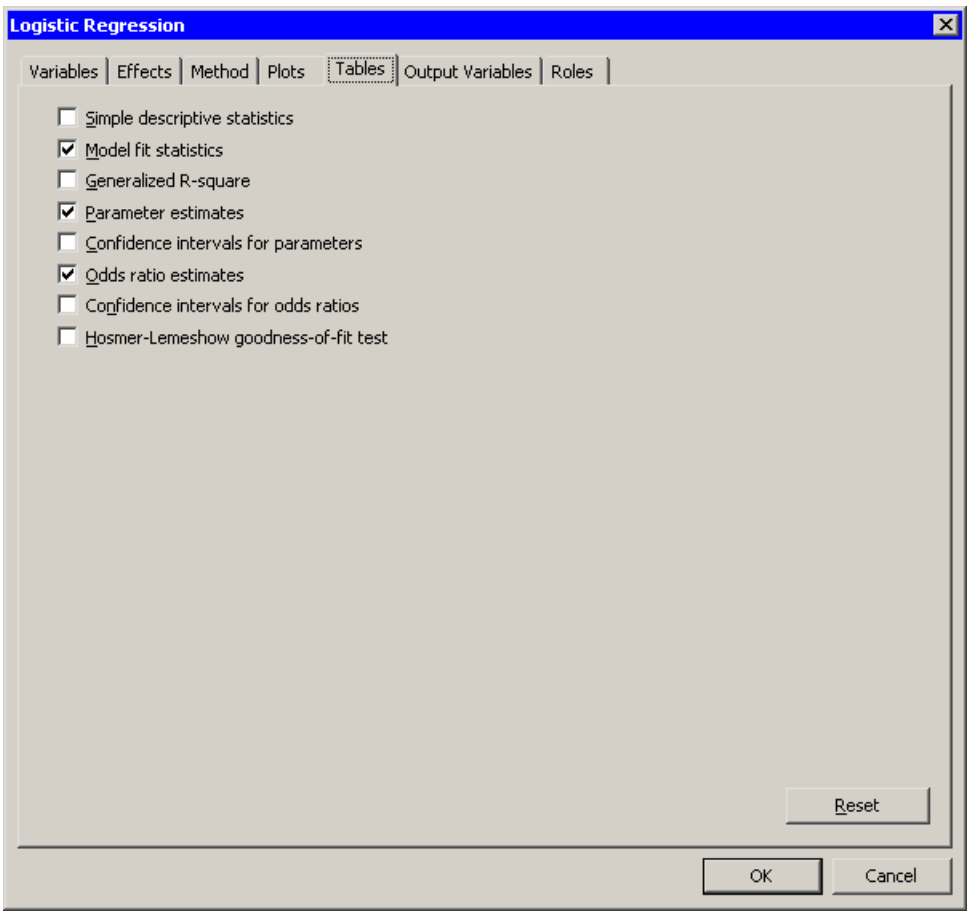
Confidence intervals for parameters
displays estimates of 95% confidence intervals for the model parameters.

Odds ratios estimates
displays the odds ratio estimates.

Confidence intervals for odds ratios
displays estimates of 95% confidence intervals for the odds ratios.

Hosmer-Lemeshow goodness-of-fit test
displays partition information and statistics for the Hosmer-Lemeshow goodness-of-fit test.

Figure 23.12 The Tables Tab



Output Variables Tab

You can use the **Output Variables** tab to add analysis variables to the data table. (See Figure 23.13.) If you request a plot that uses one of the output variables, then that variable is automatically created even if you did not explicitly select the variable on the **Output Variables** tab.

The following list describes each output variable and indicates how the output variable is named. Y represents the name of the response variable. If you use events/trials syntax, then Y represents the name of the events variable.

Proportions for events/trials
adds a variable named Proportion_ET, where E is the name if the events variable and T is the name of the trials variable. The value of the variable is the ratio E/T. This variable is added only when you use events/trials syntax.

Predicted probabilities
adds predicted probabilities. The variable is named LogiP_Y.

Confidence limits for predicted probabilities
adds 95% confidence limits for the predicted probabilities. The variables are named LogiLclm_Y and LogiUclm_Y.

Linear predictor (log odds)
adds the linear predictor values. The variable is named LogiXBeta_Y.

Pearson chi-square residuals
adds the Pearson chi-square residuals. The variable is named LogiChiSqR_Y.

Deviance residuals
adds the deviance residuals. The variable is named LogiDevR_Y.

Confidence interval displacement (C)
adds the confidence interval displacement diagnostic, C. The variable is named LogiC_Y.

Scaled confidence interval displacement (CBAR)
adds the confidence interval displacement diagnostic, $\bar{C}$. The variable is named LogiCBar_Y.

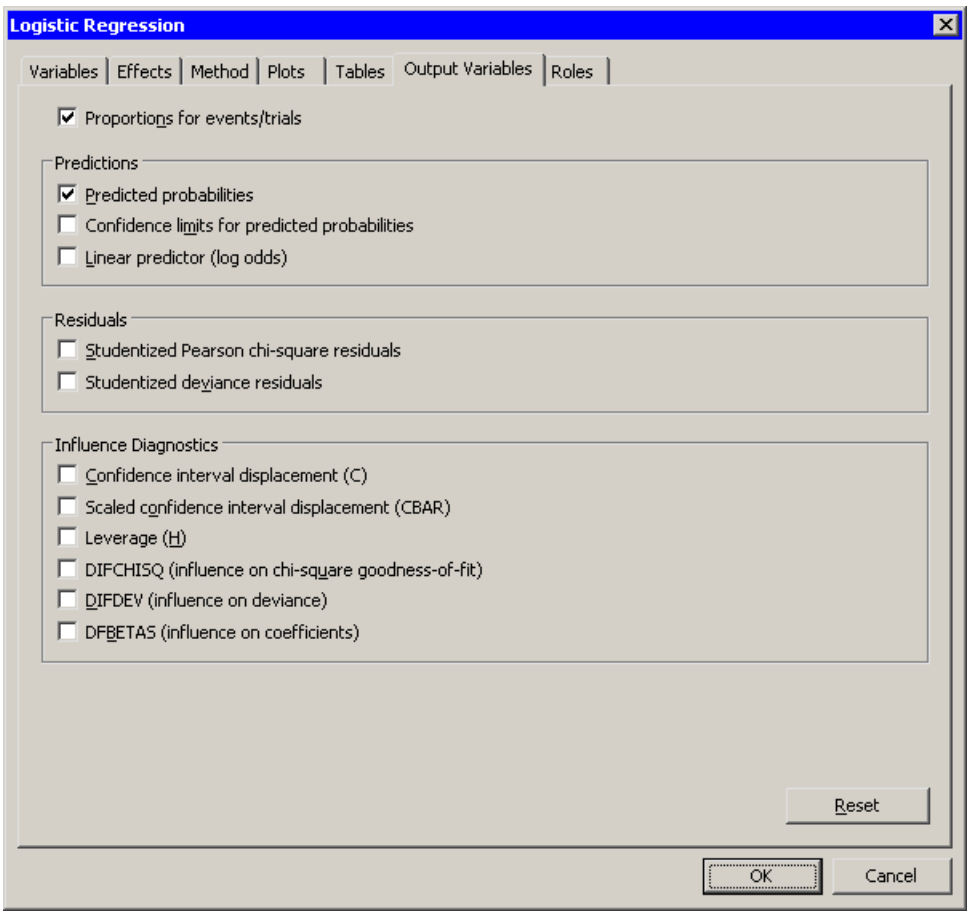
Leverage (H)
adds the leverage statistic. The variable is named LogiH_Y.

DIFCHISQ (influence on chi-square goodness-of-fit)
adds the change in the chi-square goodness-of-fit statistic that is attributed to deleting the individual observation. The variable is named LogiDifChiSq_Y.

DIFDEV (influence on deviance)
adds the change in the deviance that is attributed to deleting the individual observation. The variable is named LogiDifDev_Y.

DFBETAS (influence on coefficients)
adds m variables, where m is the number of parameters in the model. The variables are scaled measures of the change in each parameter estimate and are calculated by deleting the ith observation. Large probabilities of DFBETAS indicate observations that are influential in estimating a given parameter. The variables are named DFBETA_X, where X is the name of an interval regressor (including the intercept). For classification variables, the variables are named DFBETA_CL, where C is the name of the variable and L represents a level.

Figure 23.13 The Output Variables Tab



Roles Tab

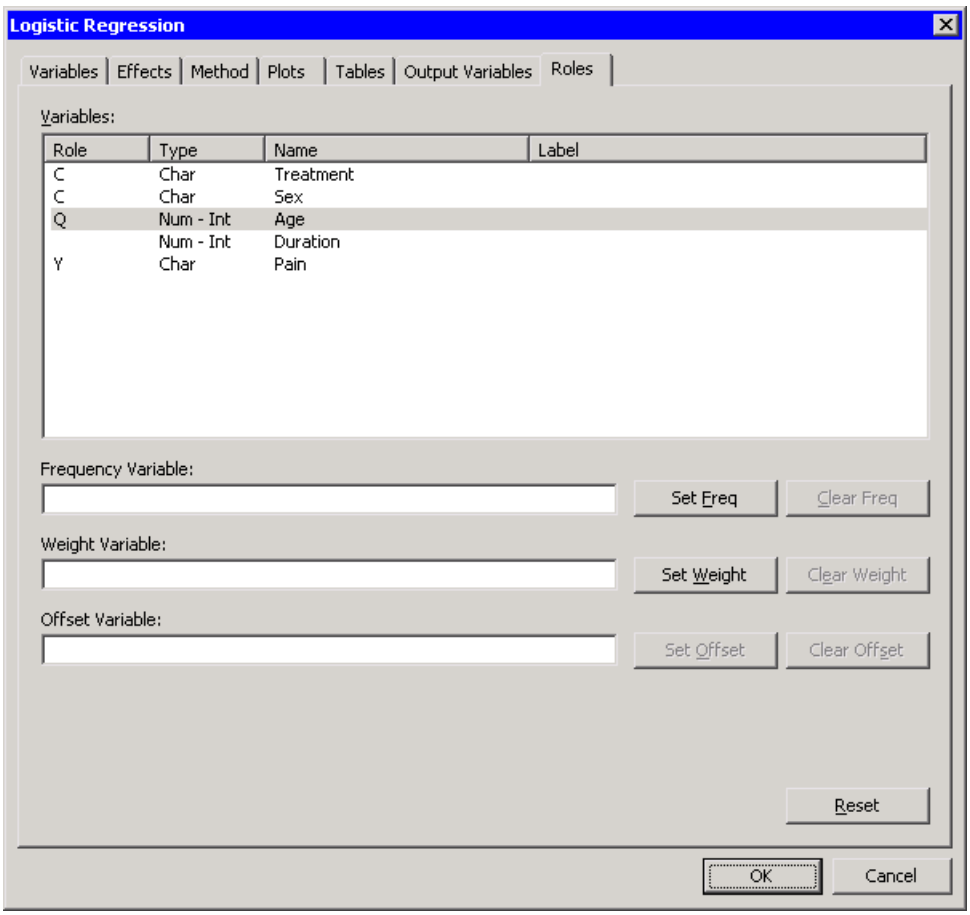
You can use the **Roles** tab to specify a frequency variable or weight variable for the analysis. (See Figure 23.14.) You can also specify an *offset variable*.

A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents n observations, where n is the value of the frequency variable.

A weight variable is a numeric variable with values that weigh each observation in the regression.

An offset variable is a special explanatory variable. The regression coefficient for this variable is fixed at 1. This corresponds to the OFFSET= option in the MODEL statement.

Figure 23.14 The Roles Tab



Analysis of Selected Variables

If one or more interval variables are selected in a data table when you run the analysis, then the following occurs:

- The first selected nominal variable is automatically entered in the **Y Variables** field of the **Variables** tab.
- Subsequent selected nominal variables are automatically entered in the **Classification Variables** field.
- Selected interval variables are automatically entered in the **Quantitative Variables** field.

Any variable in the data table with a Frequency or Weight role is automatically entered in the appropriate field of the **Roles** tab.

Chapter 24
Model Fitting: Generalized Linear Models

Contents

Overview of the Generalized Linear Models Analysis **361**
Example 1: Fit a Linear Regression Model with Classification Variables **362**
Explore the Data 363
Create an Initial Model 366
Revise the Model 371
Example 2: Fit a Poisson Regression Model **374**
Explore the Data 374
Create the Offset Variable 376
Model the Data 378
Model Overdispersion 381
Specifying the Generalized Linear Models Analysis **383**
Variables Tab 383
Effects Tab 384
Method Tab 384
Plots Tab 385
Tables Tab 386
Output Variables Tab 387
Roles Tab 389
Analysis of Selected Variables **389**
References **390**

Overview of the Generalized Linear Models Analysis

The generalized linear model is a generalization of the traditional linear model. It differs from a linear model in that it assumes that the response distribution is related to the linear predictor through a function called the *link function*.

Specifically, a generalized linear model has a linear component

$$\boldsymbol{\eta} = \boldsymbol{\eta_0} + \boldsymbol{X\beta}$$

and a monotonic differentiable function, g, that links the expected response mean, $\boldsymbol{\mu}$, to the linear predictor $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu})$$

The response $\boldsymbol{y}$ is assumed to have a distribution from the exponential family (for example, normal, gamma, Poisson, binomial, and so on). The vector $\boldsymbol{\eta_0}$ is called an *offset variable*. As in least squares regression, $\boldsymbol{X}$ is the design matrix and $\boldsymbol{\beta}$ is a vector of unknown parameters.

The explanatory variables in the Generalized Linear Models analysis can be interval variables or nominal variables (also known as *classification variables*). You can also specify more complex model terms such as interactions and nested effects.

As mentioned in Chapter 21, "Model Fitting: Linear Regression," the Linear Regression analysis in SAS/IML Studio does not support classification variables. You can use the Generalized Linear Models analysis to fit a linear regression with classification variables by specifying that the response variable is normally distributed and that the link function is the identity function. The first example in this chapter demonstrates this technique. The second example in this chapter fits a Poisson regression model. The link function for this example is the log function.

You can run a Generalized Linear Models analysis by selecting **Analysis ►Model Fitting ►Generalized Linear Models** from the main menu. The computation of the regression function and related statistics is implemented by calling the GENMOD procedure in SAS/STAT software. See the documentation for the GENMOD procedure in the *SAS/STAT User's Guide* for additional details.

Example 1: Fit a Linear Regression Model with Classification Variables

In this example you use the Generalized Linear Models analysis to fit a linear regression model with classification variables and an interaction term. In particular, you model how two variables affect the change in blood pressure in a designed experiment.

The Drug data set contains results of an experiment that is carried out to evaluate the effect of four drugs with three experimentally induced diseases. Each drug-by-disease combination was applied to six randomly selected dogs. The response variable, chang_bp, is the increase in systolic blood pressure due to the treatment. The variables drug and disease are classification variables: their values identify distinct levels or groups.

To fit a generalized linear model:

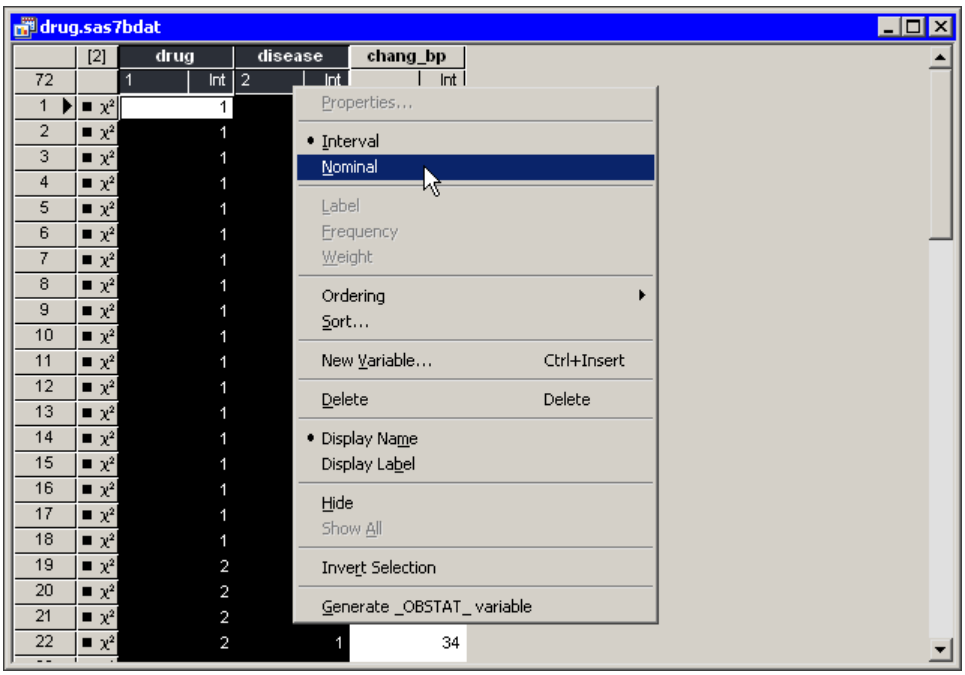
1. Open the Drug data set.

   You need to specify that the drug and disease variables are nominal in order to model them as classification variables. The "Data Table Menus" section in Chapter 4 describes measure levels for variables. The following steps change the measure level of these variables from interval to nominal:

2. Select the drug and disease variables by holding down the CTRL key while you click the column heading for each variable.

3. Right-click the column heading for either variable and select **Nominal** from the pop-up menu, as shown in Figure 24.1.

Figure 24.1 Changing the Measure Level for Variables



4 Clear the selected variables by clicking the blank cell in the upper left corner of the data table.

Explore the Data

The following steps describe how to use box plots to explore the relationship between blood pressure and the levels of drug and disease. The section "Box Plots" on page 71 describes how to create a box plot.

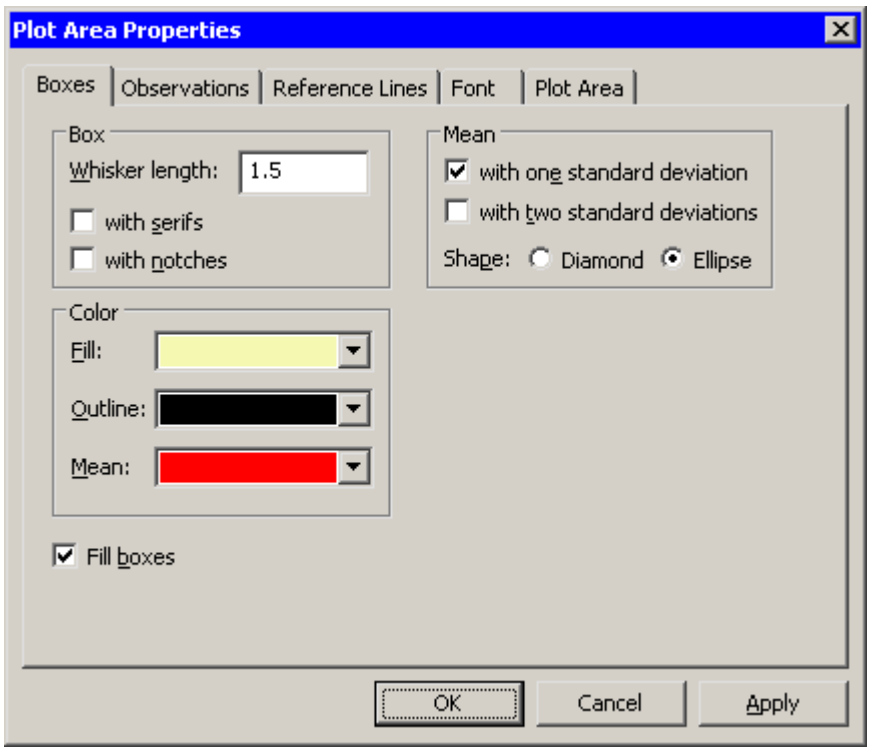
1 Select **Graph ►Box Plot** from the main menu. Create a box plot of chang_bp versus drug.

The following steps add an indicator of the mean and standard deviation of each group to the box plot.

2 Right-click near the center of the scatter plot, and select **Plot Area Properties** from the pop-up menu.

The Generalized Linear Models dialog box appears. (See Figure 24.2.) You can use the **Boxes** tab to change attributes of the box plot.

Figure 24.2 The Box Plot Dialog Box



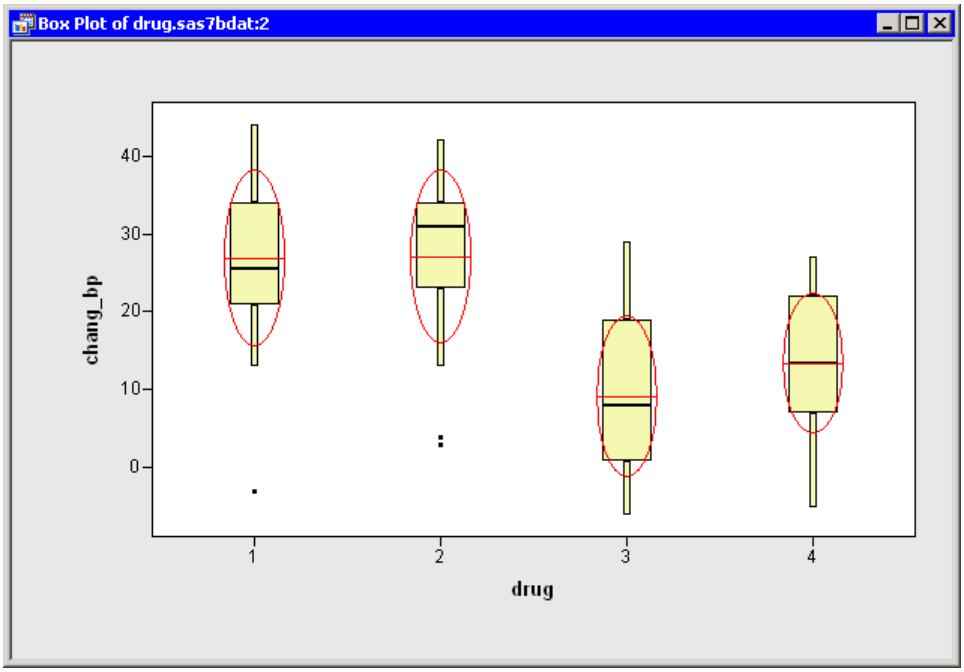
3 Select **Mean: with one standard deviation**.

4 Click **OK**.

NOTE: As a shortcut to the previous three steps, you can press the "m" key while the box plot window is active to toggle the display of means and standard deviations.

The box plot is shown in Figure 24.3. The mean change in blood pressure for Drug 1 and Drug 2 is higher than the mean change for Drug 3 and Drug 4 (averaged over all three levels of disease). This difference might indicate that the main effect for drug should be included in a model for predicting chang_bp.

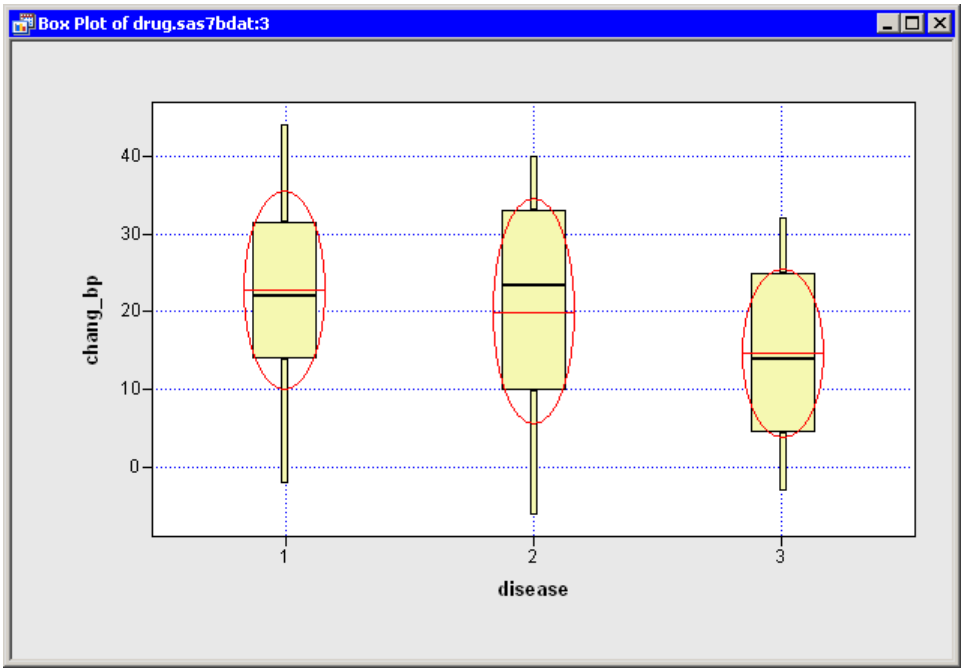
Figure 24.3 Blood Pressure Grouped by Drug



5 Repeat the previous steps to create a box plot of chang_bp versus disease. Add means and standard deviations to the plot.

A box plot that groups the response by disease is shown in Figure 24.4. The means for these groups vary according to the values of disease. The differences between the three disease levels are not as pronounced as those observed for drug. Still, the plot indicates that disease might be a factor in predicting chang_bp.

Figure 24.4 Blood Pressure Grouped by Disease



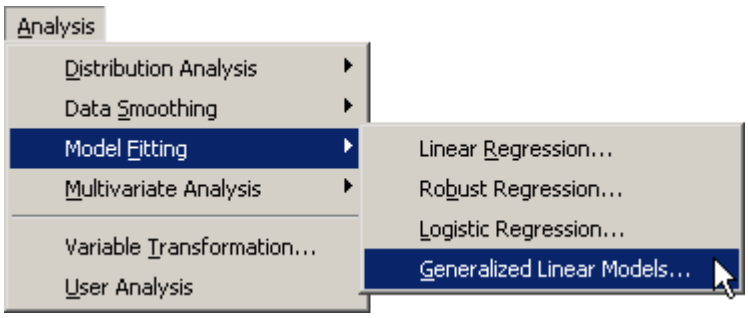
Create an Initial Model

The two box plots indicate that both drug and disease affect the change in blood pressure in the experimental subjects. Kutner (1974) proposed a two-way analysis of variance model for these data. You can use the Generalized Linear Models analysis to determine which effects are significant and to estimate parameters in the model. However, note that the analysis does not create an ANOVA table, since the GENMOD procedure does not produce ANOVA tables.

To begin the analysis:

1 Select **Analysis ►Model Fitting ►Generalized Linear Models** from the main menu, as shown in Figure 24.5.

Figure 24.5 Selecting a Generalized Linear Models Analysis



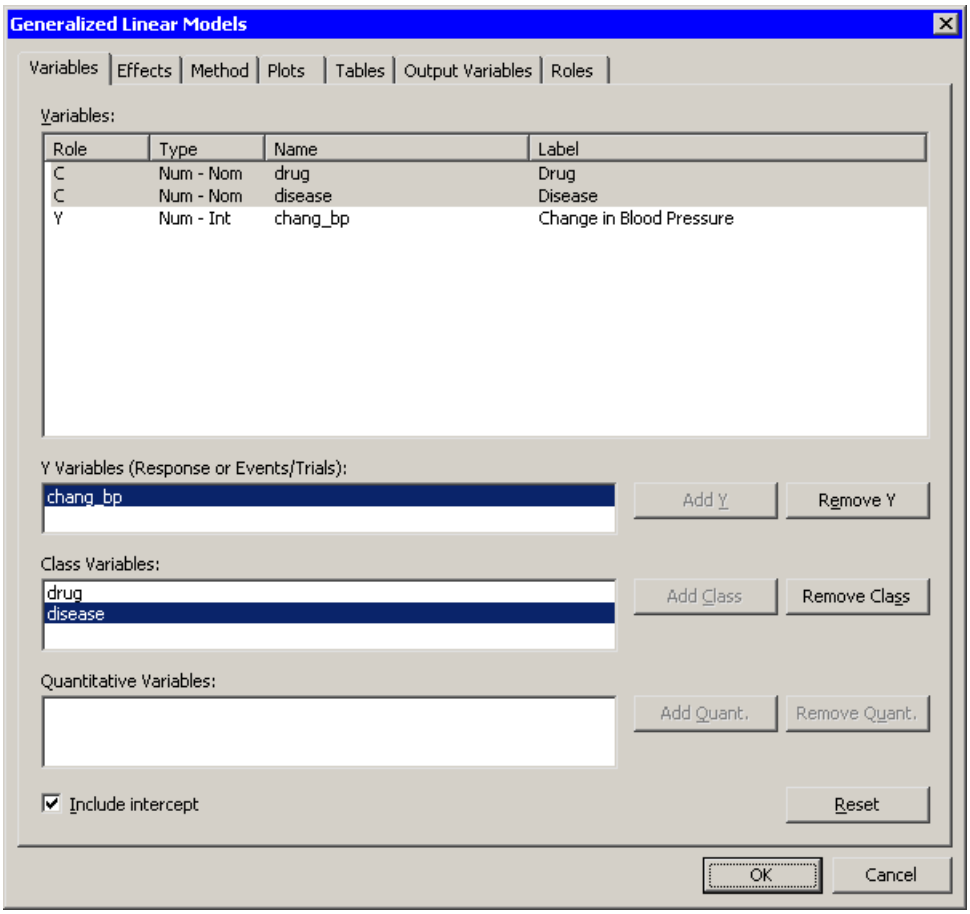
The Generalized Linear Models dialog box appears. (See Figure 24.6.)

2 Select chang_bp, and click **Add Y**.

3 Select drug. While holding down the CTRL key, select disease. Click **Add Class**.

NOTE: Alternatively, you can double-click a variable to automatically add it as an explanatory variable. Nominal variables are automatically added as classification variables; interval variables are automatically added as quantitative variables.

Figure 24.6 The Variables Tab



When you add explanatory variables to the model by using the **Variables** tab, the main effects for those variables are automatically added to the **Effects** tab. It is not clear from the box plots whether drug and disease interact. By adding an interaction term, you can determine whether the level of drug affects the change in blood pressure differently for different levels of disease.

To add an interaction term to the model:

4 Click the **Effects** tab.

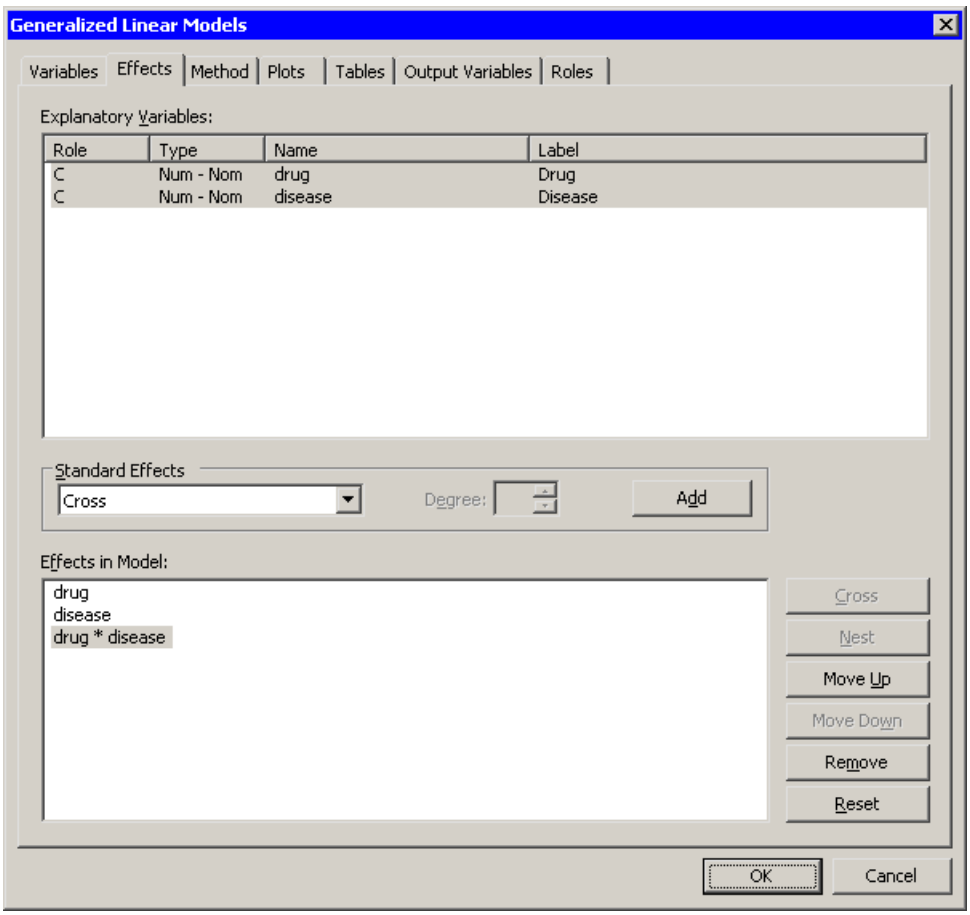
5 Select drug and disease from the **Explanatory Variables** list.

6 Select **Cross** from the **Standard Effects** list, if it is not already selected.

7 Click **Add**.

The interaction term drug*disease is added to the **Effects in Model** list, as shown in Figure 24.7.

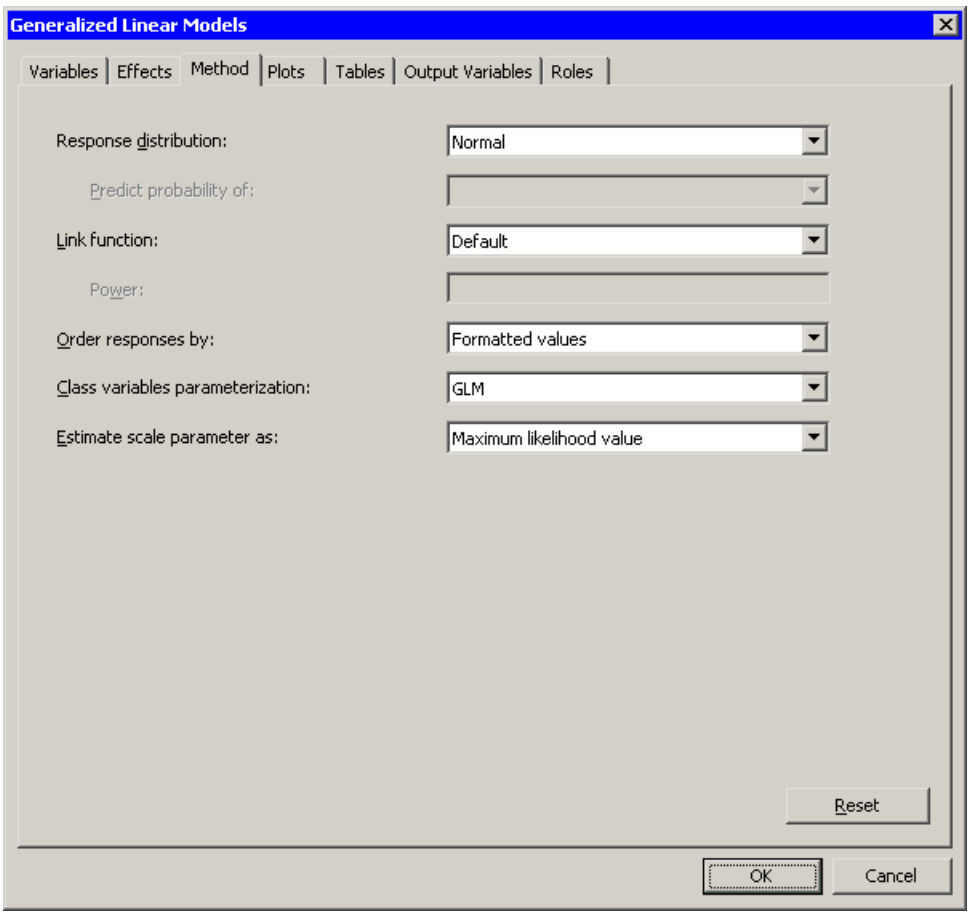
Figure 24.7 The Effects Tab



8 Click the **Method** tab.

The **Method** tab enables you to specify aspects of the generalized linear model such as the response distribution and the link function. (See Figure 24.8.) The default distribution for the response is the normal distribution, and the default link function is the identity function. You do not need to modify this tab since these choices are appropriate for the current analysis.

Figure 24.8 The Method Tab



9 Click the **Tables** tab.

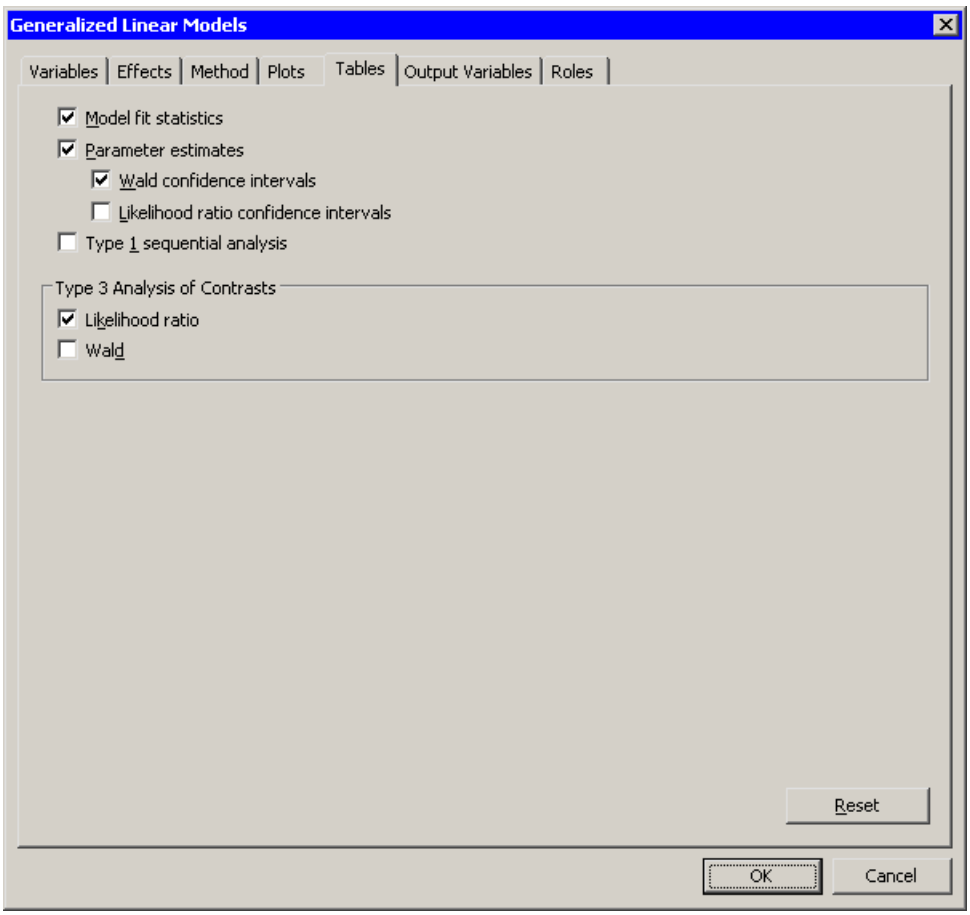
The **Tables** tab becomes active, as shown in Figure 24.9. This tab controls which tables are produced by the analysis.

By default, the analysis displays Type 3 Wald statistics for the significance of effects. The Wald statistics require less computational time than the Type 3 likelihood ratio statistics, but they can be less accurate. For this example, select the more accurate likelihood ratio statistics.

10 Clear **Wald** in the Type 3 Analysis of Contrasts group box.

11 Select **Likelihood Ratio** to request statistics for Type 3 contrasts.

Figure 24.9 The Tables Tab



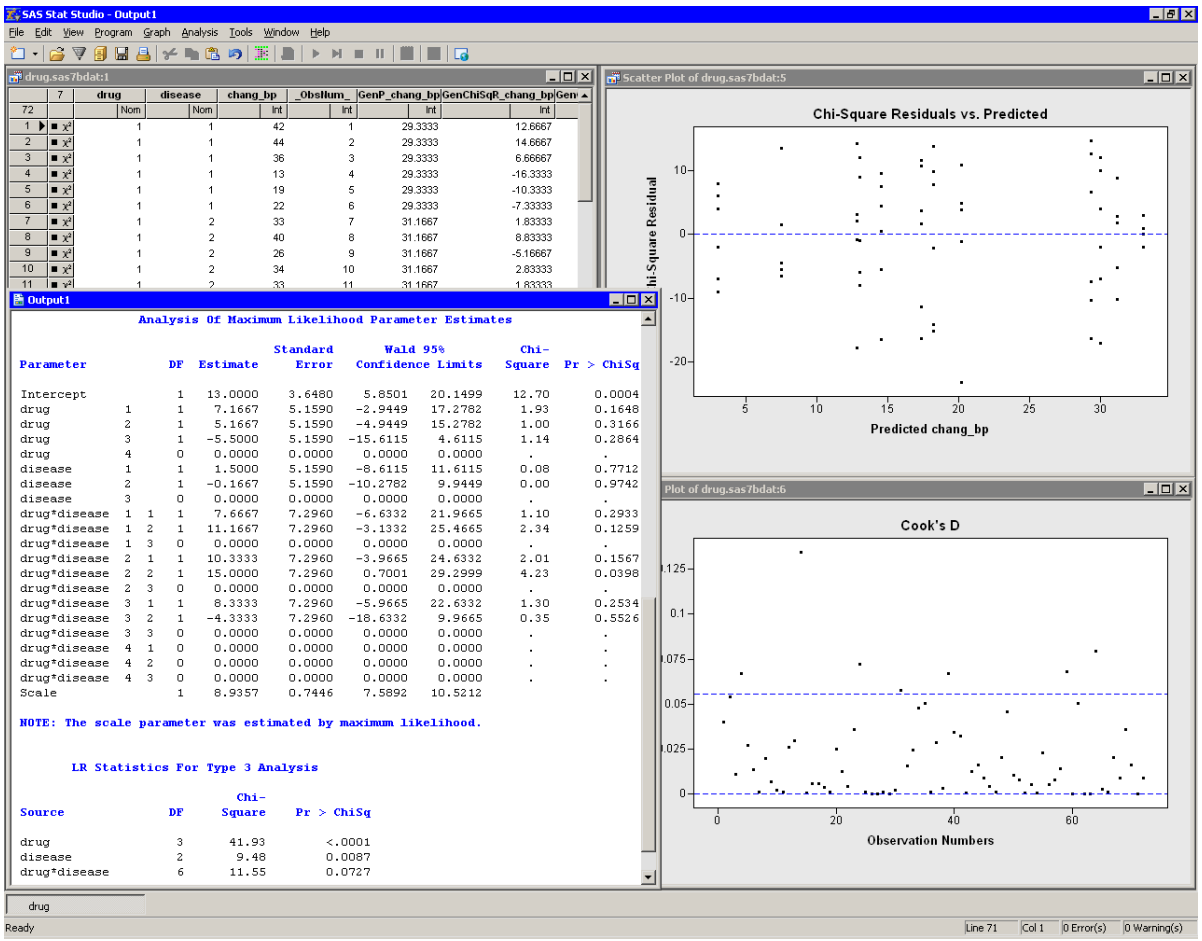
12 Click **OK**.

The analysis creates plots, along with output from the GENMOD procedure. Move the plots so that they are arranged as in Figure 24.10.

The tables created by the GENMOD procedure appear in the output window. The "LR Statistics For Type 3 Analysis" table indicates which effects in the model are significant. The Type 3 chi-square value for an effect tests the contribution due to that effect, after correcting for the other effects in the model. For example, the chi-square value for the interaction term drug*disease compares the log likelihood for the full model with the log likelihood for the model with only main effects. The value of the Type 3 likelihood ratio statistic for the interaction term is 11.55. The associated p-value indicates that this term is not significant in predicting the change in blood pressure at the 0.05 significance level. The main effects for drug and disease are significant.

Since the interaction effect is not significant, the parameter estimates in the "Analysis Of Maximum Likelihood Parameter Estimates" table are not useful. You should rerun the model without the interaction effect before examining the parameter estimates. The next section shows you how to delete the interaction effect and rerun the analysis.

Figure 24.10 Preliminary Generalized Linear Models Analysis



Revise the Model

To remove the interaction effect from the previous model and refit the data:

1 Select **Analysis ►Model Fitting ►Generalized Linear Models** to redisplay the dialog box for this analysis.

NOTE: The items on the **Analysis** menu are not available if the output window is active. If the menu is not enabled, you should activate a graphical or tabular view of the data before clicking on the **Analysis** menu.

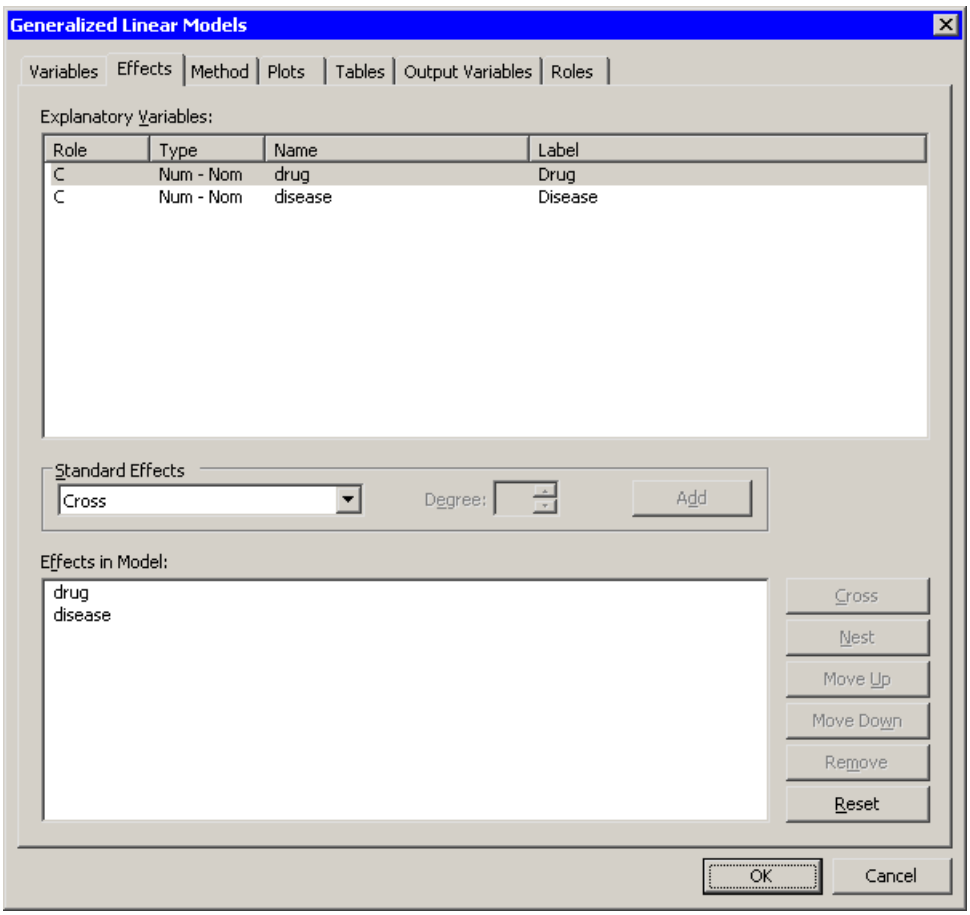
2 Click the **Effects** tab.

3 Select **drug * disease** from the **Effects in Model** list.

4 Click **Remove**.

The interaction term is removed from the list of effects, as shown in Figure 24.11.

Figure 24.11 Revising the Model



5 Click **OK**.

Move the workspace windows so that they are arranged as in Figure 24.12. The "LR Statistics For Type 3 Analysis" table indicates that both main effects are significant.

The "Analysis Of Maximum Likelihood Parameter Estimates" table displays parameter estimates for the model. You can use these values to determine the predicted mean response for each experimental group. The interpretation of the parameter estimates depends on the parameterization used to encode the classification variables in the model design matrix. This example used the GLM coding (see Figure 24.8). For this parameterization, the predicted response for a subject is obtained by adding the estimate for the intercept to the parameter estimates for the groups to which the subject belongs. For example, the predicted change in blood pressure in a subject with drug=1 and disease=2 is $8.9861 + 13.4444 + 5.2917 \approx 27.7$.

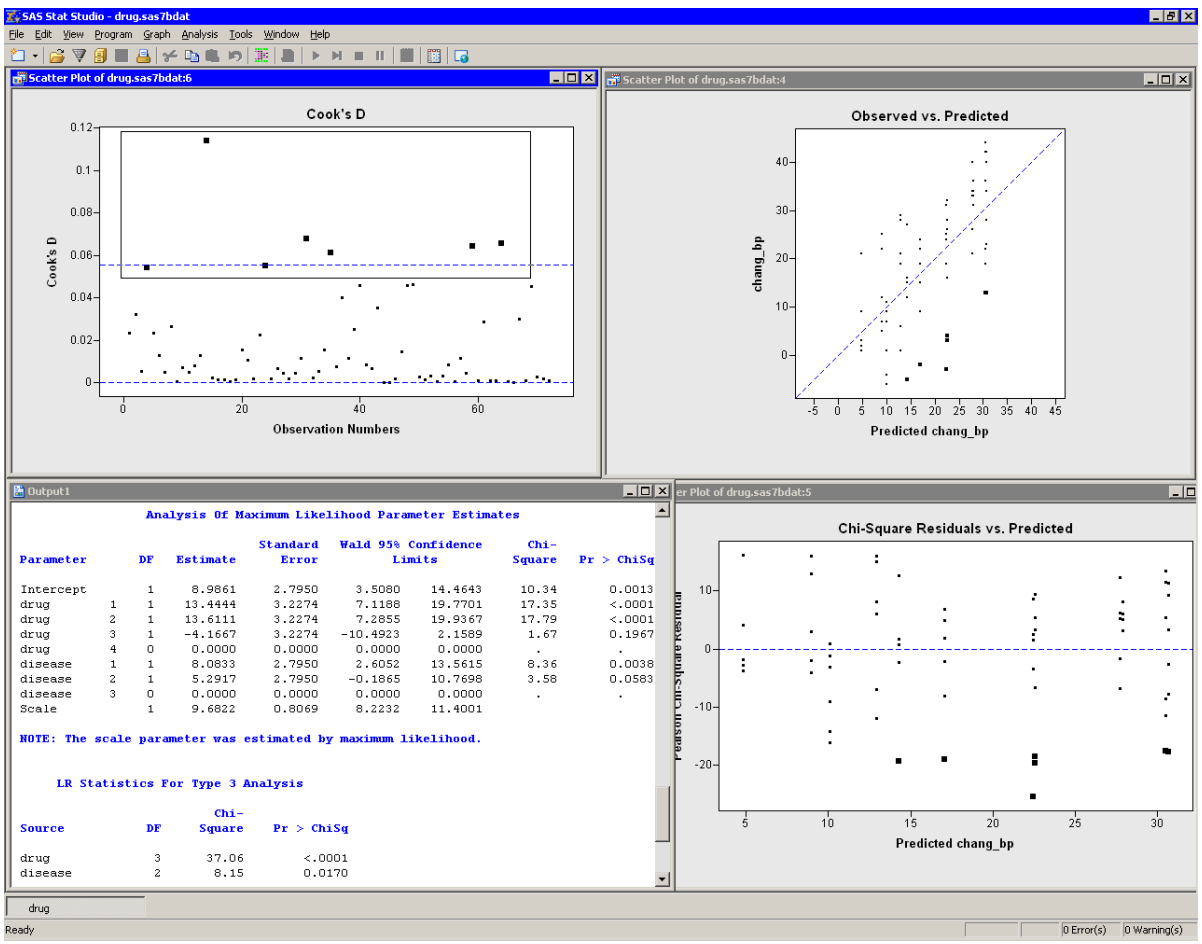
For a given level, the parameter estimate represents the difference between that level and the last level. For example, the estimate of the difference between the parameters for Drug 1 and Drug 4 is 13.4444, and this estimate is significantly different from zero (as indicated by the p-value in the "Pr > ChiSq" column). In contrast, the difference in the coefficients between Drug 3 and Drug 4 is –4.1667, but this estimate is not significantly different from zero. Similarly, the estimate of the difference between Disease 2 and Disease 3 is (marginally) not significant.

The parameter estimates table also estimates the scale parameter. For a normally distributed response, the scale parameter is the standard deviation of the response. See the documentation for the GENMOD procedure in the *SAS/STAT User's Guide* for additional details.

There are three plots in Figure 24.12. The Observed vs. Predicted plot (upper right in Figure 24.12) shows how well the model fits the data. Since this model assumes a normally distributed response with an identity link, the Chi-Square Residuals vs. Predicted plot (lower right in Figure 24.12) is just an ordinary residual plot (see the "Residuals" section of the documentation for the GENMOD procedure). The observations fall along vertical lines because all observations with the *i*th drug and the *j*th disease have the same predicted value.

The scatter plot of Cook's D (upper left in Figure 24.12) indicates which observations have a large influence on the parameter estimates. Influential observations (that is, those with relatively large values of Cook's D) are selected in the figure. The selected observations are highlighted in the other plots. Each observation corresponds to a large negative residual, which indicates that the observed change in blood pressure for these subjects was substantially less than the model predicts.

Figure 24.12 A Revised Generalized Linear Models Analysis



Example 2: Fit a Poisson Regression Model

In this example, you examine another example of a generalized linear model: Poisson regression. A Poisson regression analysis might be appropriate when the response variable represents counts or rates. If your explanatory variables are all nominal (that is, you can write a contingency table that contains the data), then the Poisson model is often called a *log-linear model*.

Counts are always nonnegative, but a linear model can predict negative values for the response. Consequently, it is common to choose a logarithmic link function for the response. That is, if the response variable is Y and the expected value of Y is μ, a Poisson regression finds parameters that best fit the data to the model $\log(\mu) = \mathbf{X}\boldsymbol{\beta}$.

Sometimes the counts represent the number of events that occurred during an observed time period. Some counts might correspond to longer time periods than others do. In this situation, you want to model the rate at which the events occur. When you model a rate, you are modeling the number of events, Y, per unit of time, T. The expected value of the rate is μ/T, where μ is the expected value of Y. In this case, the Poisson model is $\log(\mu/T) = \mathbf{X}\boldsymbol{\beta}$. By using the fact that $\log(\mu/T) = \log(\mu) - \log(T)$, this equation can be rewritten as

$$\log(\mu) = \log(T) + \mathbf{X}\boldsymbol{\beta}$$

The term $\log(T)$ is called an *offset variable*.

The example in this section fits a Poisson model to data in the Ship data set. The data and analysis are from McCullagh and Nelder (1989). The response variable, Y, is the number of damage incidents that occurred during the number of months that ship was in service (contained in the months variable). As discussed in the previous paragraph, the quantity $\log$(months) is an offset variable for this model. The three classification variables are as follows:

- the ship type (type), which contains five levels, a–e
- the year of construction (year), which contains four levels: 1960–64, 1965–69, 1970–74, and 1975–79
- the period of operation (period), which contains two levels: 1960–74 and 1975–79

Explore the Data

To use box plots to explore the data:

1 Open the Ship data set.

You can use box plots to explore how the ratio of Y to months varies according to the levels of the classification variables.

Figure 24.13 shows plots that indicate how the number of damage incidents per month varies with the explanatory variables. The Variable Transformation Wizard is used to create a new variable, IncidentsPerMonth, as the ratio of Y and months. The new variable was created by using the **Y / X** transformation from the **Two**

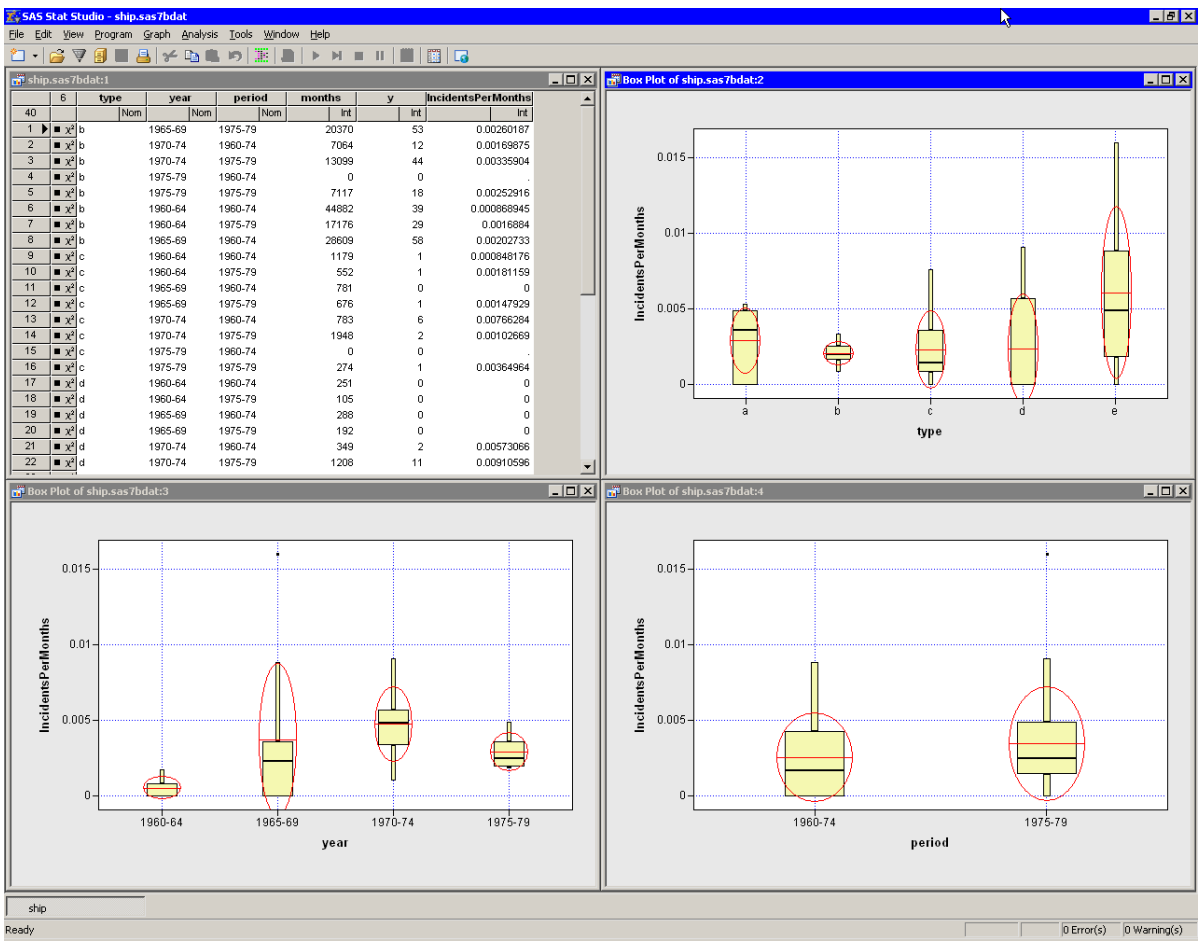
Variable family of transformations. The Variable Transformation Wizard is described in further detail in Chapter 32, "Variable Transformations."

The three box plots indicate that the mean of IncidentsPerMonth is as follows:

- highest for ships of Type e, and low for the other types
- highest for ships constructed in the years 1970–74, and lowest for ships constructed in the years 1960–64
- highest for ships that operated in the 1975–79 period, and lowest for ships that operated in the 1960–74 period

This preliminary analysis indicates that the main effects of type, year, and period are important in predicting IncidentsPerMonth. The next section creates a generalized linear model with these effects.

Figure 24.13 Incidents per Month, Grouped by Classification Variables



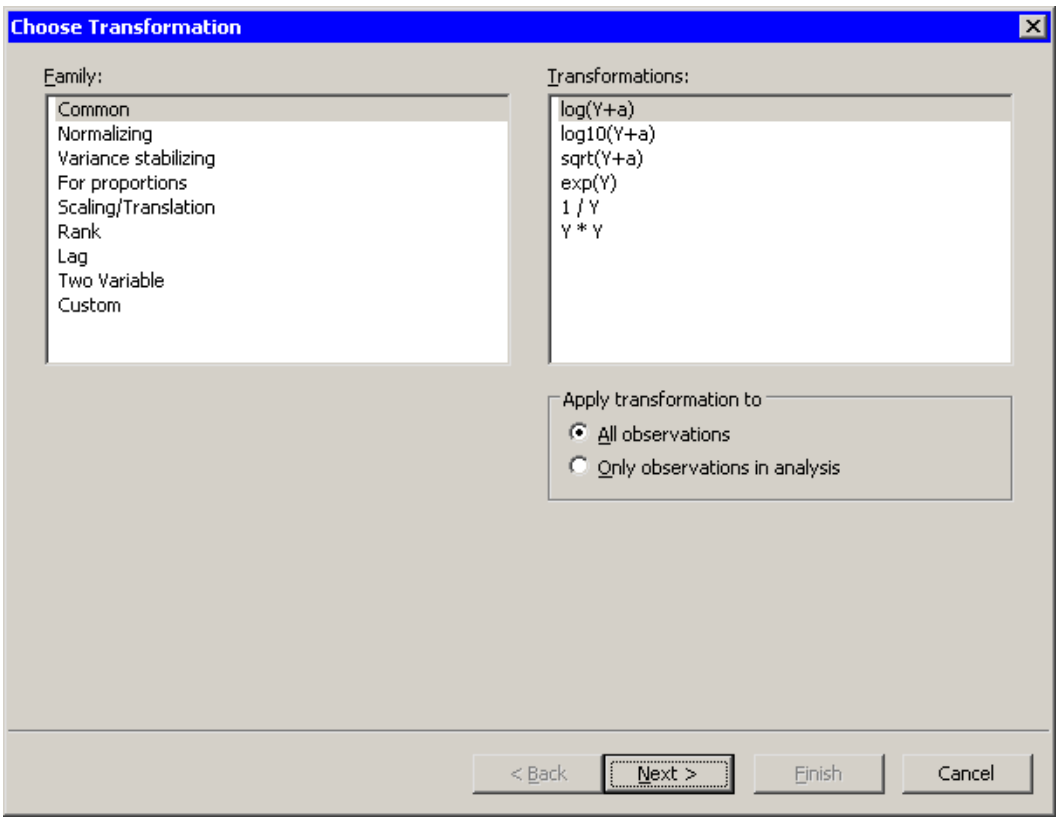
Create the Offset Variable

As discussed earlier in this example, the quantity log(months) is an offset variable for this model. The following steps use the Variable Transformation Wizard to create this variable. (This wizard is described in further detail in Chapter 32, "Variable Transformations.")

1 Select **Analysis ►Variable Transformation** from the main menu.

The Variable Transformation Wizard in Figure 24.14 appears.

Figure 24.14 Selecting a Transformation

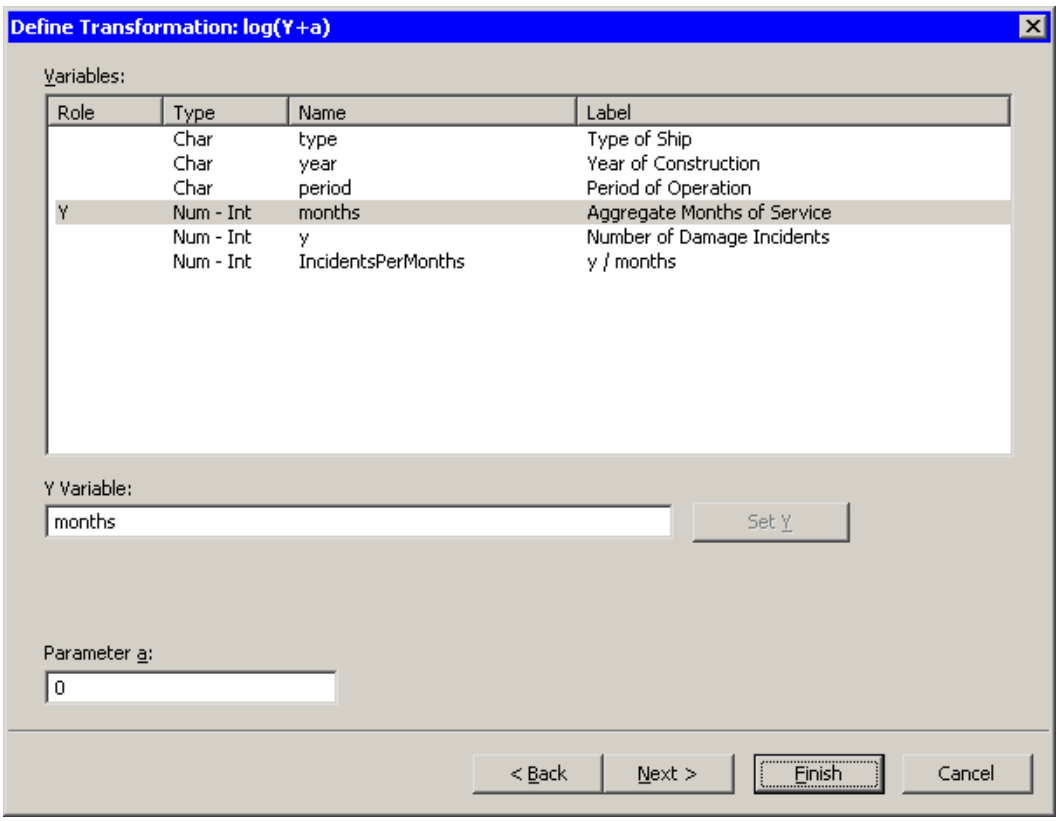


The transformation **log(Y+a)** is highlighted by default. Since this is the desired transformation, you can proceed to the next page of the wizard.

2 Click **Next**.

The wizard displays the page shown in Figure 24.15. Note that the transformation appears in the page's title bar.

Figure 24.15 Selecting a Variable and a Parameter

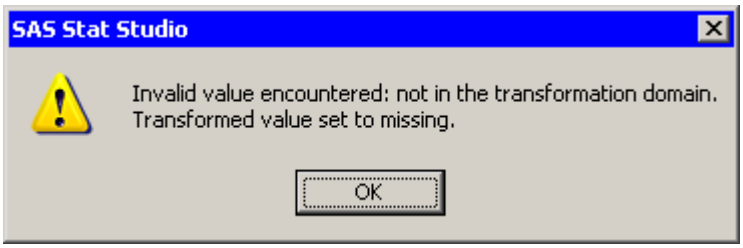


3 Select the months variable, and click **Set Y**.

4 Click **Finish**.

Because there are six observations for which months=0, a warning message appears (Figure 24.16) that informs you that the transformed values for these observations are set to missing values.

Figure 24.16 A Warning Message



5 Click **OK** to dismiss the warning message.

The new variable is named Log_months. It contains six missing values. Observations with missing values for the explanatory variables (including the offset variable) or the response variable are not used in fitting the model.

Model the Data

The previous sections describe the Poisson model and create an offset variable for this model. The following steps describe how to specify the model.

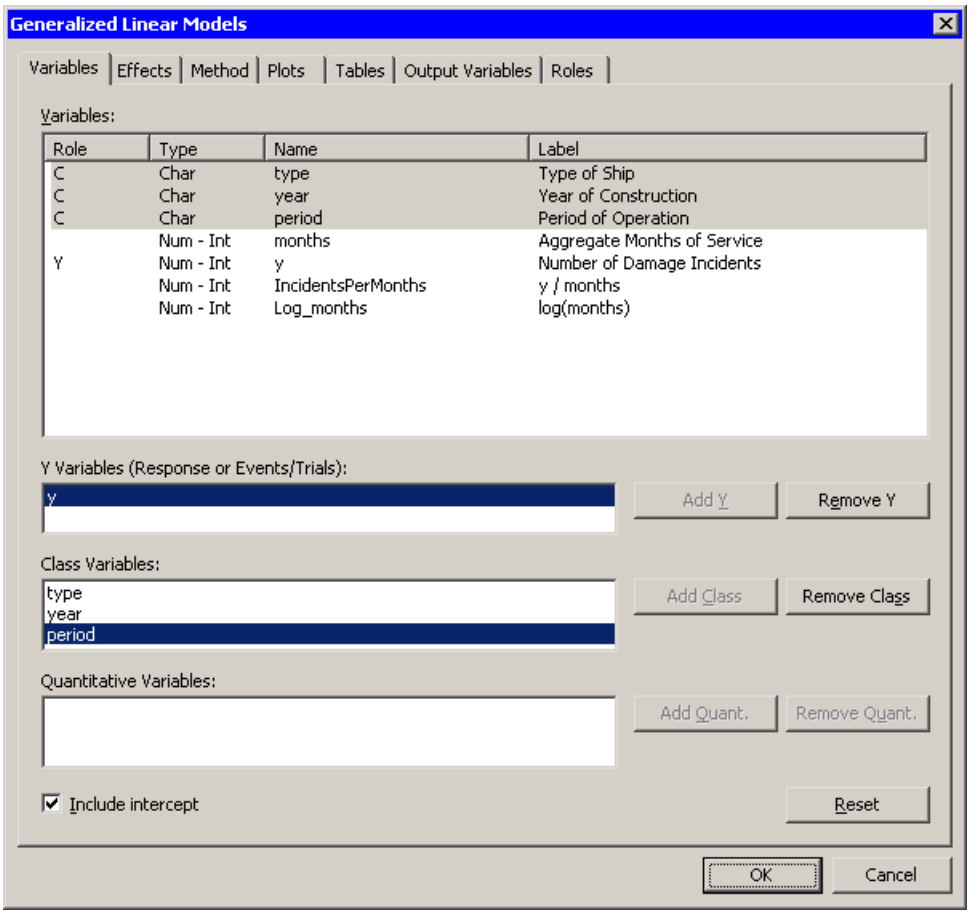
1 Select **Analysis ►Model Fitting ►Generalized Linear Models** from the main menu.

The Generalized Linear Models dialog box appears. (See Figure 24.17.)

2 Select y, and click **Add Y**.

3 Select type. While holding down the CTRL key, select year and period. Click **Add Class**.

Figure 24.17 The Variables Tab



Recall that when you add a variable on the **Variables** tab, the main effect for that variable is added to the **Effects** tab. This model includes only the main effects, so you do not need to click the **Effects** tab.

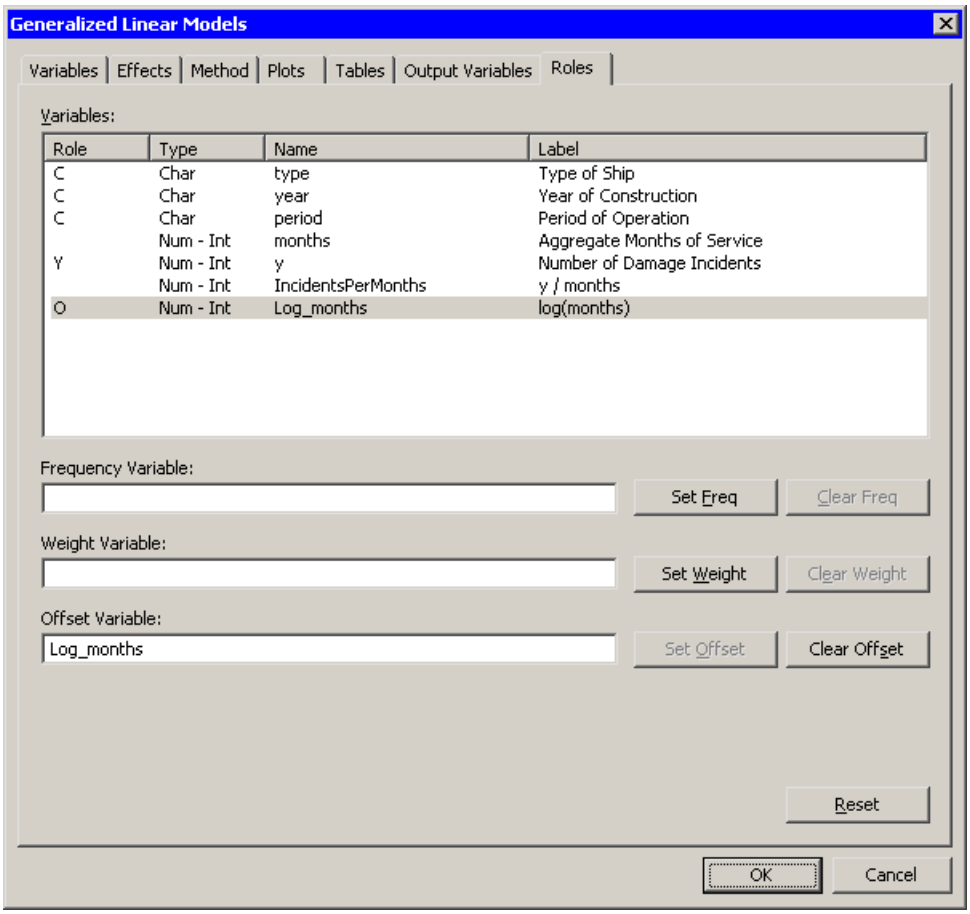
There is one more variable to specify. The following steps specify Log_months as the offset variable:

4 Click the **Roles** tab.

The **Roles** tab appears, as shown in Figure 24.18.

5 Select Log_months, and click **Set Offset**.

Figure 24.18 The Roles Tab



You have specified the variables in the model. The next steps specify the response distribution and the link function for a Poisson regression:

6 Click the **Method** tab.

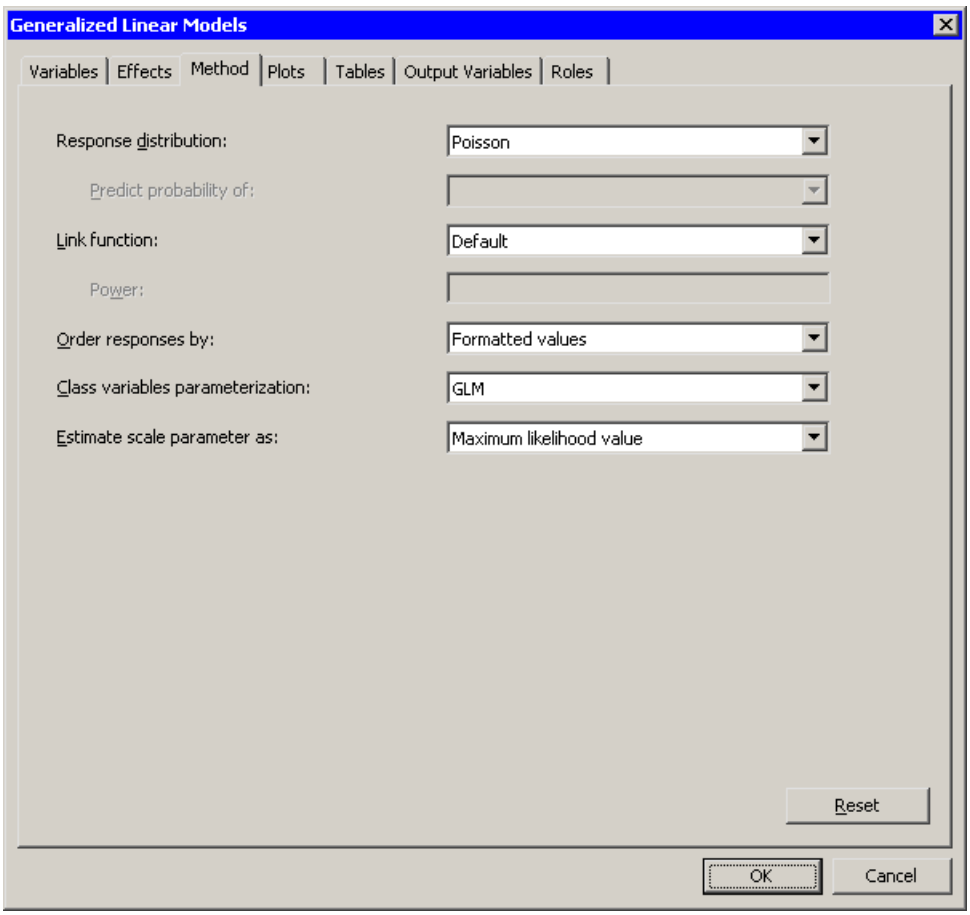
The **Method** tab appears as in Figure 24.19.

7 Select **Poisson** for **Response Distribution**.

This specifies that the values of y have a probability distribution that is Poisson. (This also implies that the variance of y is proportional to the mean.)

When a response distribution is Poisson, the default link function is the natural log. Consequently, you do not need to change the **Link function** value.

Figure 24.19 The Method Tab



8 Click the **Tables** tab.

The **Tables** tab becomes active, as shown in Figure 24.9. This tab controls which tables are produced by the analysis.

9 Clear **Wald** in the Type 3 Analysis of Contrasts group box.

10 Select **Likelihood Ratio** to request statistics for Type 3 contrasts.

11 Click **OK** to run the analysis.

The results of the analysis are shown in Figure 24.20. Move the workspace windows so that they are arranged as in the figure.

The "LR Statistics For Type 3 Analysis" table indicates that all main effects are significant, although period is the weakest of the three.

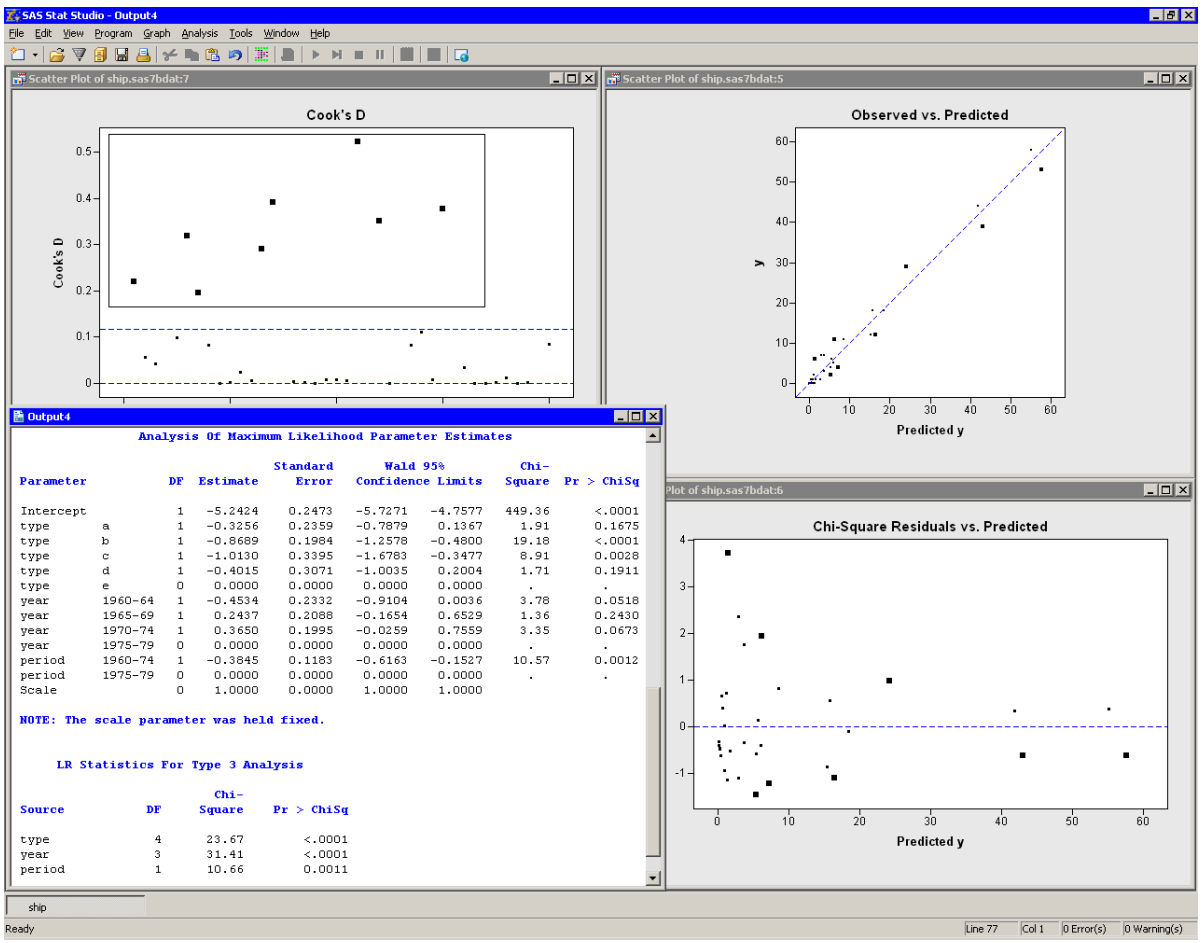
The "Analysis Of Maximum Likelihood Parameter Estimates" table displays parameter estimates for each level of the effects. The Parameter Estimates column indicates that ships of Type b and Type c have the lowest risk and ships of Type e have the highest. The oldest ships (built from 1960 to 1964) have the lowest risk, and ships built from 1965 to 1974 have the highest risk. However, the estimates of the difference between

the older ships and the newer ships are not significantly different from zero (as indicated by the Pr > ChiSq column). Ships operated from 1960 to 1974 have a lower risk than ships operated from 1975 to 1979.

The GENMOD procedure displays a note that indicates that the scale parameter is fixed—that is, not estimated by the iterative fitting process.

There are three plots in Figure 24.20. The scatter plot of Cook's D (upper left in Figure 24.20) indicates which observations have a large influence on the parameter estimates. Influential observations are highlighted in all plots. Note that the influential observations are not necessarily those with the largest residual values.

Figure 24.20 A Poisson Regression Analysis



Model Overdispersion

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. (Recall that the Poisson variance equals the response mean: $\mathrm{Var}(y) = \mu$.) To account for the overdispersion that might occur in the Ship data, you can specify a method for estimating the overdispersion.

To estimate overdispersion:

1 Select **Analysis ►Model Fitting ►Generalized Linear Models** from the main menu.

Each tab of the dialog box initializes with the values from the previous analysis of these data.

2 Click the **Method** tab.

3 Select **Pearson chi-square/DF** for the field **Estimate scale parameter as** (shown in Figure 24.21).

4 Click **OK**.

Figure 24.21 Modeling Overdispersion

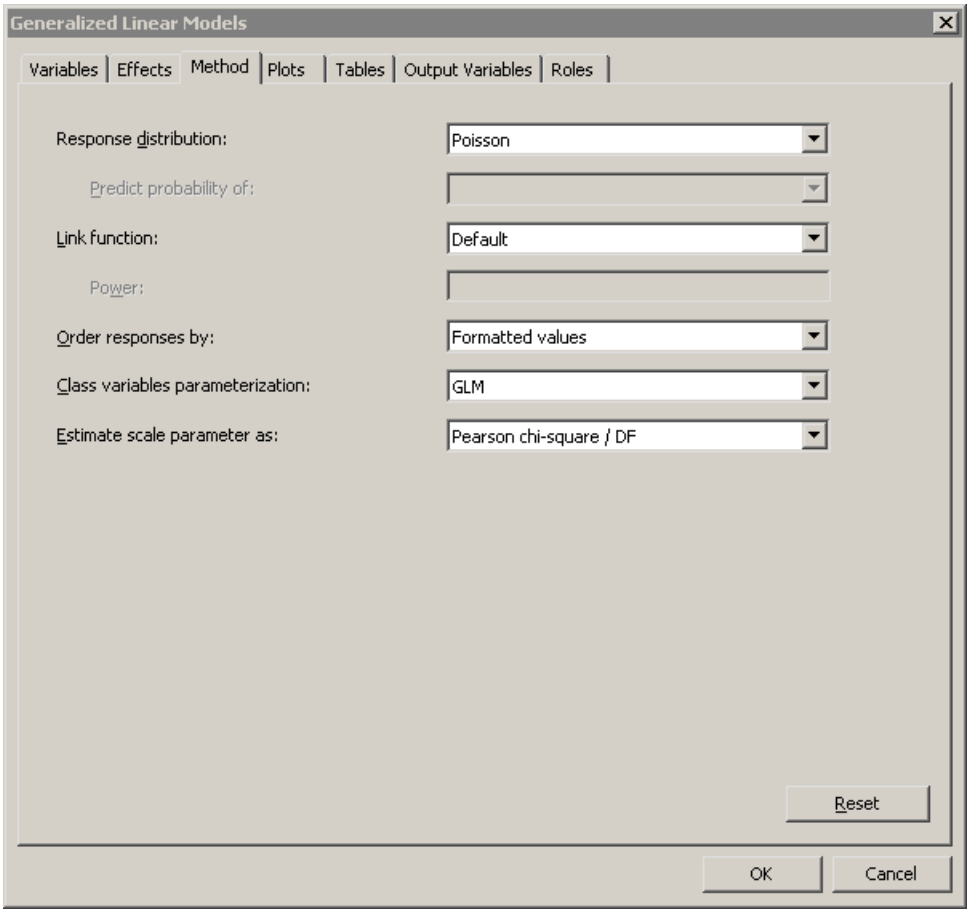


Figure 24.22 shows the output for the analysis.

A note states that "the scale parameter was estimated by the square root of Pearson's Chi-Square/DOF." The scale value reported in the "Analysis Of Maximum Likelihood Parameter Estimates" table is greater than 1, which suggests that overdispersion exists in the model.

Note that the parameter estimates are unchanged by the dispersion estimate. However, the estimate does affect the covariance matrix, standard errors, and log likelihoods used in likelihood ratio tests. A comparison of Figure 24.20 with Figure 24.22 shows multiple differences in the output statistics.

Although the estimate of the dispersion parameter is often used to indicate overdispersion or underdispersion, this estimate might also indicate other problems, such as an incorrectly specified model or outliers in the data.

See the subsection "Generalized Linear Models Theory" in the "Details" section of the documentation for the GENMOD procedure for a discussion of the dispersion parameter and overdispersion.

Figure 24.22 Estimating the Overdispersion Parameter

Output4

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-5.2424	0.3216	-5.8727	-4.6121	265.73	<.0001
type	a	1	-0.3256	0.3067	-0.9268	0.2756	1.13	0.2885
type	b	1	-0.8689	0.2580	-1.3746	-0.3632	11.34	0.0008
type	c	1	-1.0130	0.4414	-1.8781	-0.1478	5.27	0.0217
type	d	1	-0.4015	0.3994	-1.1843	0.3812	1.01	0.3147
type	e	0	0.0000	0.0000	0.0000	0.0000	.	.
year	1960-64	1	-0.4534	0.3032	-1.0477	0.1409	2.24	0.1348
year	1965-69	1	0.2437	0.2715	-0.2883	0.7758	0.81	0.3693
year	1970-74	1	0.3650	0.2594	-0.1434	0.8734	1.98	0.1594
year	1975-79	0	0.0000	0.0000	0.0000	0.0000	.	.
period	1960-74	1	-0.3845	0.1538	-0.6859	-0.0830	6.25	0.0124
period	1975-79	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.3004	0.0000	1.3004	1.3004		

NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
type	4	25	3.50	0.0212	14.00	0.0073
year	3	25	6.19	0.0027	18.57	0.0003
period	1	25	6.30	0.0189	6.30	0.0120

Specifying the Generalized Linear Models Analysis

This section describes the dialog box tabs that are associated with the Generalized Linear Models analysis. The Generalized Linear Models analysis calls the GENMOD procedure in SAS/STAT software. See the documentation for the GENMOD procedure in the *SAS/STAT User's Guide* for details.

Variables Tab

You can use the **Variables** tab to specify the variables for the Generalized Linear Models analysis. The **Variables** tab is shown in Figure 24.6.

For most response distributions, you only need to specify a single response variable in the **Y Variables** list. If you specify two numeric variables, the analysis assumes that the variables contain count data for a binomial experiment. The value of the first variable is the number of positive responses (or *events*). The value of the second variable is the number of *trials*. In this case, the response distribution is automatically set to binomial.

The dialog box supports multiple explanatory variables. You can include nominal variables in the model by adding them to the **Classification variables** list. You can include interval variables in the model by adding them to the **Quantitative variables** list.

When you add an explanatory variable, that main effect is added to the **Effects** tab. You can add interaction effects and nested effects by using the **Effects** tab.

Effects Tab

You can use the **Effects** tab to add several different types of effects to your model. All effects appear in the **Effects in Model** list. The section "Effects Tab" on page 346 in Chapter 23, "Model Fitting: Logistic Regression," describes how to use the **Effects** tab to specify effects.

Method Tab

You can use the **Method** tab to specify aspects of the generalized linear model such as the response distribution and the link function. (See Figure 24.8.)

The **Method** tab contains the following UI controls:

Response distribution
specifies the distribution of the response variable. This corresponds to the DIST= option in the MODEL statement.

Predict probability of
specifies whether to model the probability of the first or last level of the response variable. This item is available only when the response distribution is binomial or multinomial. This corresponds to the DESCENDING option in the PROC GENMOD statement.

Link function
specifies the link function. This corresponds to the LINK= option in the MODEL statement.

The following table specifies the default link function for each response distribution.

Table 24.1 Default Link Functions

Distribution	Default Link Function
Binomial	Logit
Gamma	Inverse (power(–1))
Inverse gaussian	Inverse squared (power(–2))
Multinomial	Cumulative logit
Negative binomial	Log
Normal	Identity
Poisson	Log

When the choice of response distribution is multinomial, the choice of link functions is limited to the cumulative logit, the cumulative probit, and the cumulative complementary log-log.

Power
specifies the number to use for a power link function. This item is available only when the link function is the power function.

Order response by
specifies how to order the response variable. This corresponds to the RORDER= option in the PROC GENMOD statement.

Classification variables parameterization
specifies the parameterization method for the classification variables. This corresponds to the PARAM= option in the CLASS statement. The dialog box supports the GLM, effect, and reference coding schemes.

Estimate scale parameter as
specifies the method for estimating the dispersion parameter. This corresponds to the SCALE= option in the MODEL statement.

Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 24.23.) There are plots that help you to visualize the fit, the residuals, and various influence diagnostics.

Creating a plot often adds one or more variables to the data table. For a multinomial response, residuals and influence diagnostics are not available, so the only possible plot for multinomial data is the predicted response plot.

The following plots are available:

Observed vs. Predicted
creates a scatter plot of the Y variables versus the predicted values, overlaid with the diagonal line that represents a perfect fit.

Predicted response plot
creates a line plot of the predicted probability versus the continuous explanatory variable. This plot is created only if the following conditions are satisfied:

- There is exactly one continuous explanatory variable.
- There are three or fewer classification variables.
- There are 12 or fewer joint levels of the classification variables.

If the response distribution is multinomial, there are $k - 1$ plots, where k is the number of response levels.

Pearson chi-square residuals vs. Predicted
creates a scatter plot of the residuals versus the predicted probabilities.

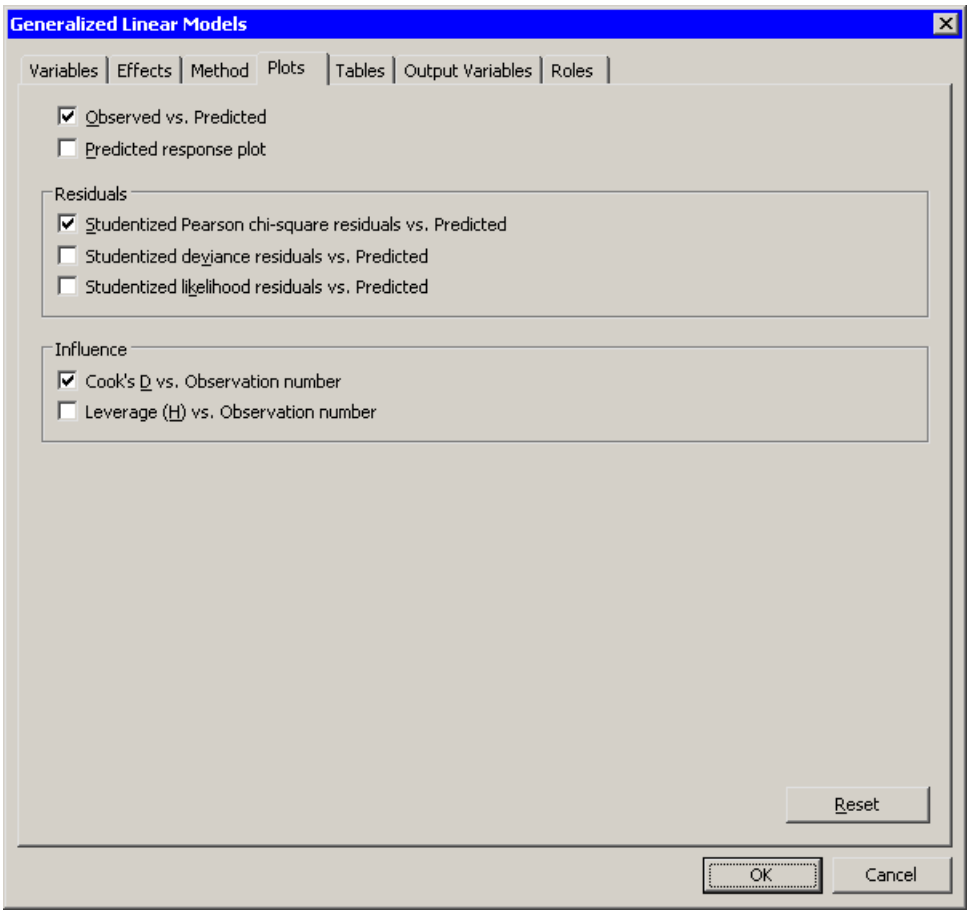
Deviance residuals vs. Predicted
creates a scatter plot of the deviance residuals versus the predicted probabilities.

Likelihood residuals vs. Predicted
creates a scatter plot of the likelihood residuals versus the predicted probabilities.

Cook's D vs. Observation number
creates a scatter plot of Cook's D statistic for each observation.

Leverage (H) vs. Observation number
creates a scatter plot of the leverage statistic for each observation.

Figure 24.23 The Plots Tab



Tables Tab

The **Tables** tab is shown in Figure 24.9. You can use the **Tables** tab to display the following tables that summarize the results of the analysis:

Model fit statistics
displays a table of model fit statistics.

Parameter estimates
displays estimates for the model parameters and the scale parameter.

Wald confidence intervals
displays estimates of 95% Wald confidence intervals for the model, based on the asymptotic normality of the parameter estimators. This corresponds to the WALDCI option in the MODEL statement. **NOTE:** The GENMOD procedure displays the Wald confidence limits by default. Consequently, Wald confidence limits appear in the parameter estimates table even if you clear both of the check boxes for confidence limits in the dialog box.

Likelihood ratio confidence intervals
displays estimates of 95% confidence intervals for the model parameters, based on the profile likelihood function. This corresponds to the LRCI option in the MODEL statement.

Type 1 sequential analysis specifies that a Type 1 sequential analysis be displayed. This corresponds to the TYPE1 option in the MODEL statement.

Likelihood ratio
specifies that Type 3 likelihood statistics be displayed. This corresponds to the TYPE3 option in the MODEL statement.

Wald
specifies that a Type 3 Wald statistics be displayed. This corresponds to the TYPE3WALD option in the MODEL statement.

Output Variables Tab

You can use the **Output Variables** tab to add analysis variables to the data table. (See Figure 24.24.) If you request a plot that uses one of the output variables, then that variable is automatically created even if you did not explicitly select the variable on the **Output Variables** tab.

For a multinomial response, residuals and influence diagnostics are not available.

The following list describes each output variable and indicates how the output variable is named. Y represents the name of the response variable. If you use events/trials syntax, then Y represents the name of the events variable.

Proportions for events/trials
adds a variable named Proportion_ET, where E is the name of the events variable and T is the name of the trials variable. The value of the variable is the ratio E/T. This variable is added only when you use events/trials syntax.

Predicted values
adds predicted values. The variable is named GenP_Y.

Confidence limits for predicted values
adds 95% confidence limits for the predicted values. The variables are named GenLclm_Y and GenUclm_Y.

Linear predictor
adds the linear predictor values. The variable is named GenXBeta_Y.

Raw residuals

adds residuals, which are calculated as observed values minus predicted values. The variable is named GenR_Y.

Pearson chi-square residuals

adds the Pearson chi-square residuals. The variable is named GenChiSqR_Y.

Deviance residuals

adds the deviance residuals. The variable is named GenDevR_Y.

Likelihood residuals

adds the likelihood residuals. The variable is named GenLikR_Y.

Cook's D

adds Cook's D influence statistic. The variable is named GenCooksD_Y.

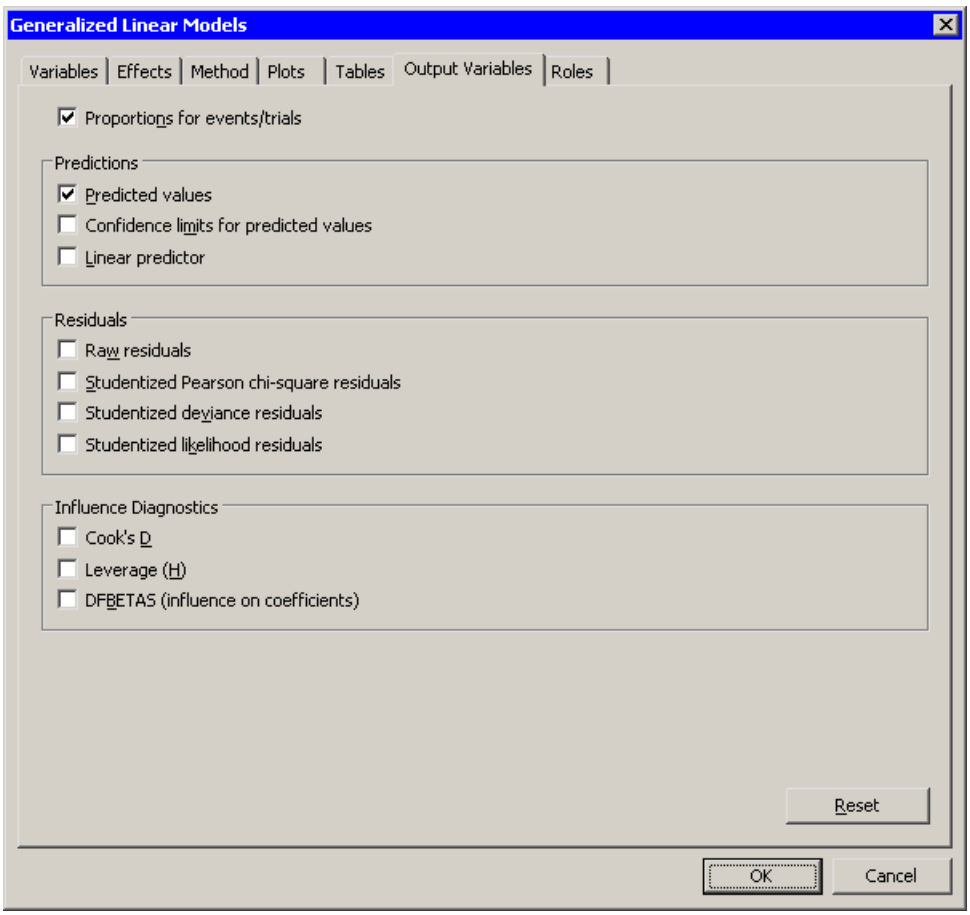
Leverage (H)

adds the leverage statistic. The variable is named GenH_Y.

DFBETAS (influence on coefficients)

adds p variables, where p is the number of parameters in the model. A classification variable with k levels counts as k parameters. The variables are scaled measures of the change in each parameter estimate and are calculated by deleting the ith observation. Large values of DFBETAS indicate observations that are influential in estimating a given parameter. Belsley, Kuh, and Welsch (1980) recommend $2/\sqrt{n}$ as a size-adjusted cutoff. The variables are named DFBeta$_j$, for $j = 1 \ldots p$.

Figure 24.24 The Output Variables Tab



Roles Tab

You can use the **Roles** tab to specify a frequency variable or weight variable for the analysis. (See Figure 24.18.) You can also specify an offset variable.

A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents n observations, where n is the value of the frequency variable.

A weight variable is a numeric variable with values that are relative weights for a weighted regression.

An offset variable is a variable used as a vector of constants in the regression. Its regression coefficient is set to 1. This corresponds to the OFFSET= option in the MODEL statement.

Analysis of Selected Variables

If one or more interval variables are selected in a data table when you run the analysis, then the following occurs:

- The first selected nominal variable is automatically entered in the **Y Variables** field of the **Variables** tab.
- Subsequent selected nominal variables are automatically entered in the **Classification Variables** field.
- Selected interval variables are automatically entered in the **Quantitative Variables** field.

Any variable in the data table with a Frequency or Weight role is automatically entered in the appropriate field of the **Roles** tab.

References

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: John Wiley & Sons.

Kutner, M. H. (1974), "Hypothesis Testing in Linear Models (Eisenhart Model)," *American Statistician*, 28, 98–100.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd Edition, London: Chapman & Hall.

Chapter 25
Multivariate Analysis: Correlation Analysis

Contents

Overview of the Correlation Analysis **391**
Example: Examine Correlations between Variables **391**
Use the Workspace Explorer to View All Plots 395
Specifying the Correlation Analysis **398**
Variables Tab 399
Plots Tab 399
Tables Tab 400
Roles Tab 402
Analysis of Selected Variables **402**

Overview of the Correlation Analysis

The Correlation analysis can help you to understand and visualize relationships between pairs of variables. You can use correlation coefficients to measure the strength of the linear association between two numerical variables. You can also use prediction ellipses in scatter plots as a visual test for bivariate normality and an indication of the strength of the correlation.

You can run the Correlation analysis by selecting **Analysis ►Multivariate Analysis ►Correlation Analysis** from the main menu. The analysis is implemented by calling the CORR procedure in Base SAS software. See the CORR procedure documentation in the *Base SAS Procedures Guide* for additional details.

Example: Examine Correlations between Variables

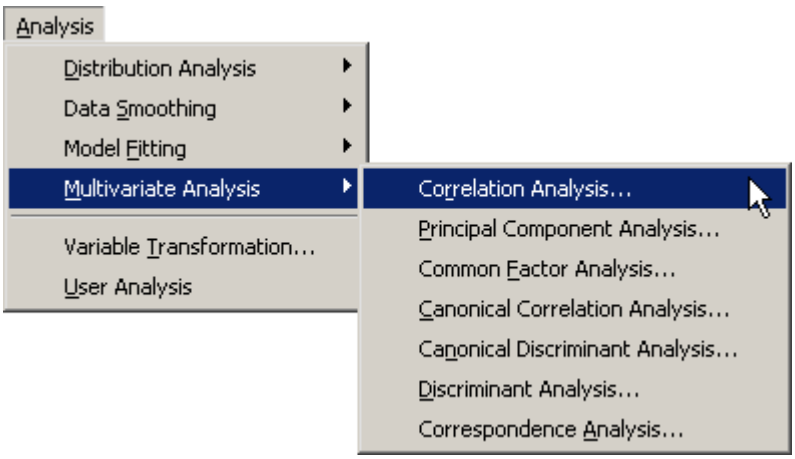
In this example, you explore correlations and bivariate relationships between variables in the Hurricanes data set. The data are for North Atlantic tropical cyclones from 1988 to 2003. The data set includes information about each storm's latitude (in the latitude variable), its sustained low-level winds (wind_kts), its central atmospheric pressure (min_pressure), and the size of its eye (radius_eye). A full description of the Hurricanes data set is included in Chapter A, "Sample Data Sets."

To run a correlation analysis:

1 Open the Hurricanes data set.

2 Select **Analysis ►Multivariate Analysis ►Correlation Analysis** from the main menu, as shown in Figure 25.1.

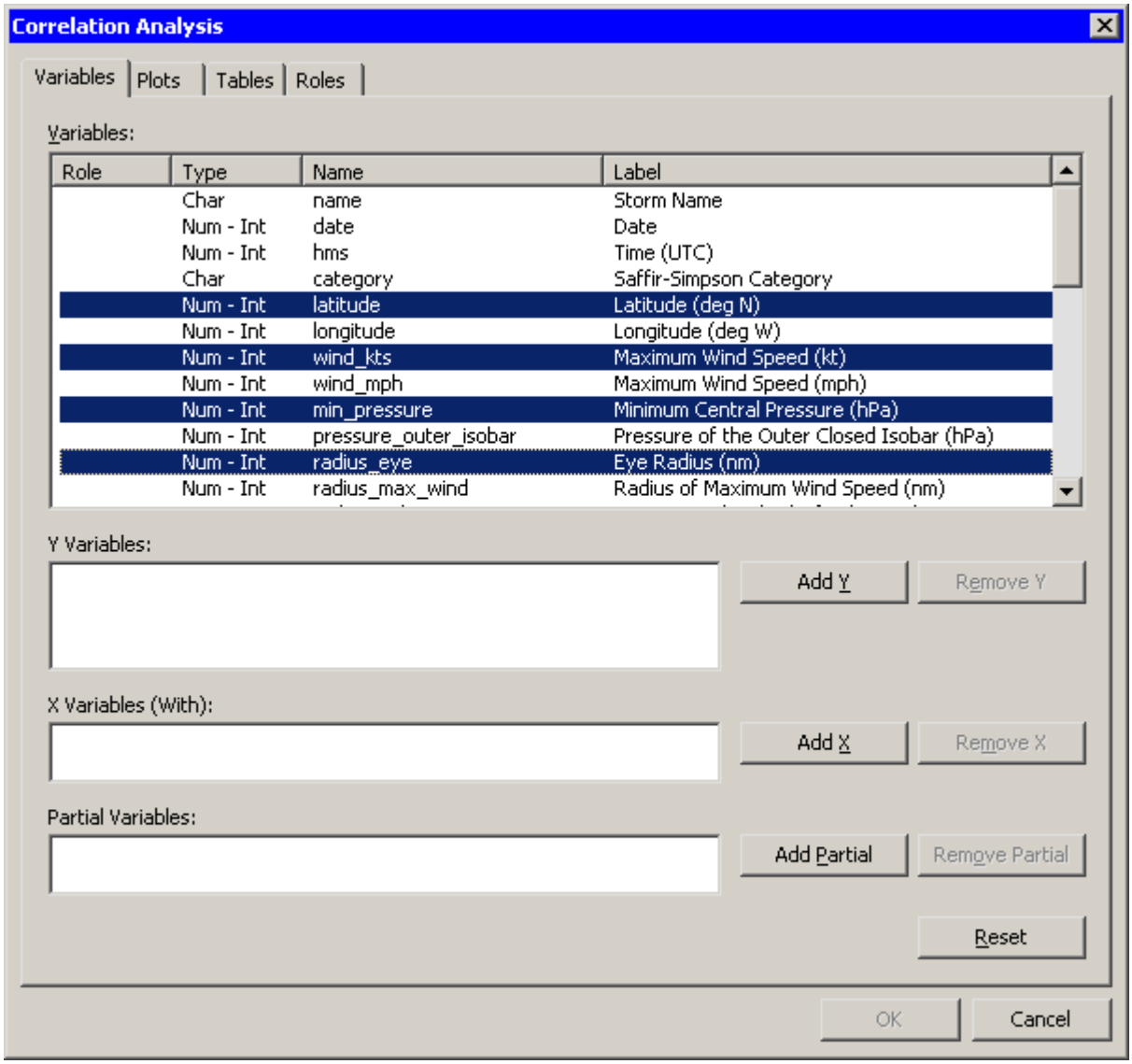
Figure 25.1 Selecting the Correlation Analysis



The Correlation Analysis dialog box appears. (See Figure 25.2.) You can select variables for the analysis by using the **Variables** tab.

3 Select latitude. While holding down the CTRL key, select wind_kts, min_pressure, and radius_eye, and click **Add Y**.

Figure 25.2 The Variables Tab



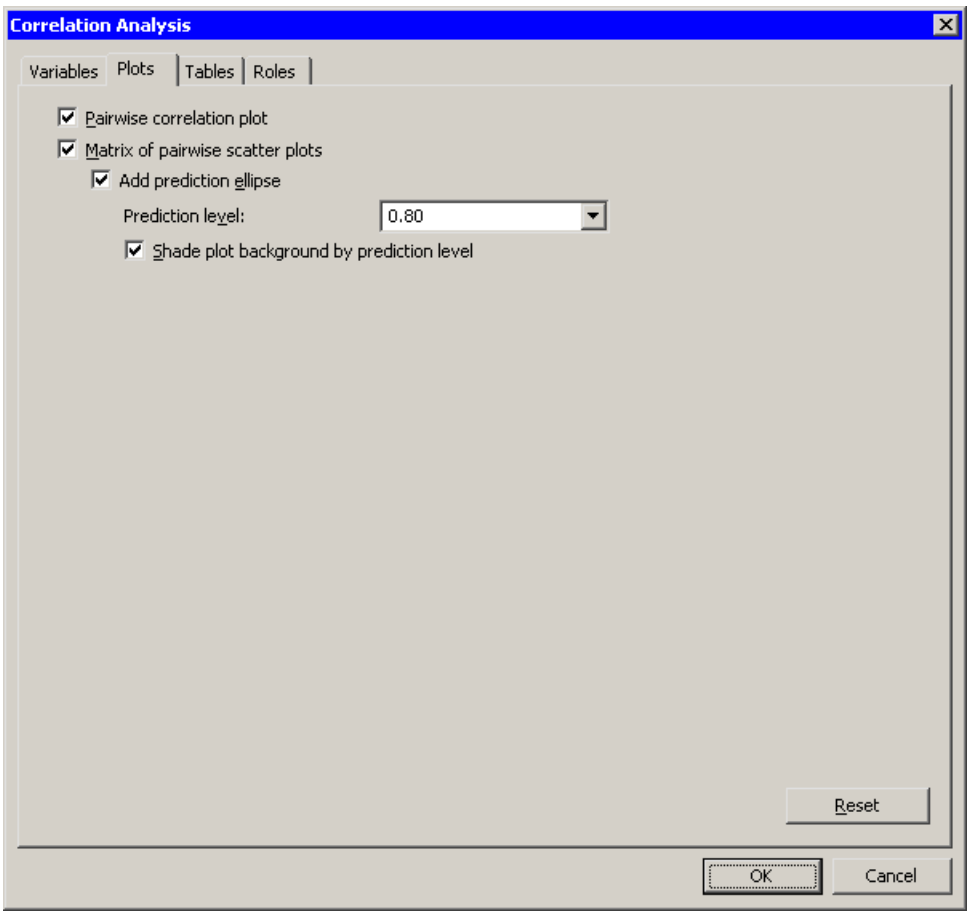
4 Click the **Plots** tab.

The **Plots** tab becomes active. (See Figure 25.3.)

5 Select **Matrix of pairwise scatter plots**.

6 Click **OK**.

Figure 25.3 The Plots Tab



The analysis calls the CORR procedure, which uses the options specified in the dialog box. The procedure displays tables in the output document, as shown in Figure 25.4.

The "Simple Statistics" table (not shown in the figure) displays basic statistics such as the mean, standard deviation, and range of each variable.

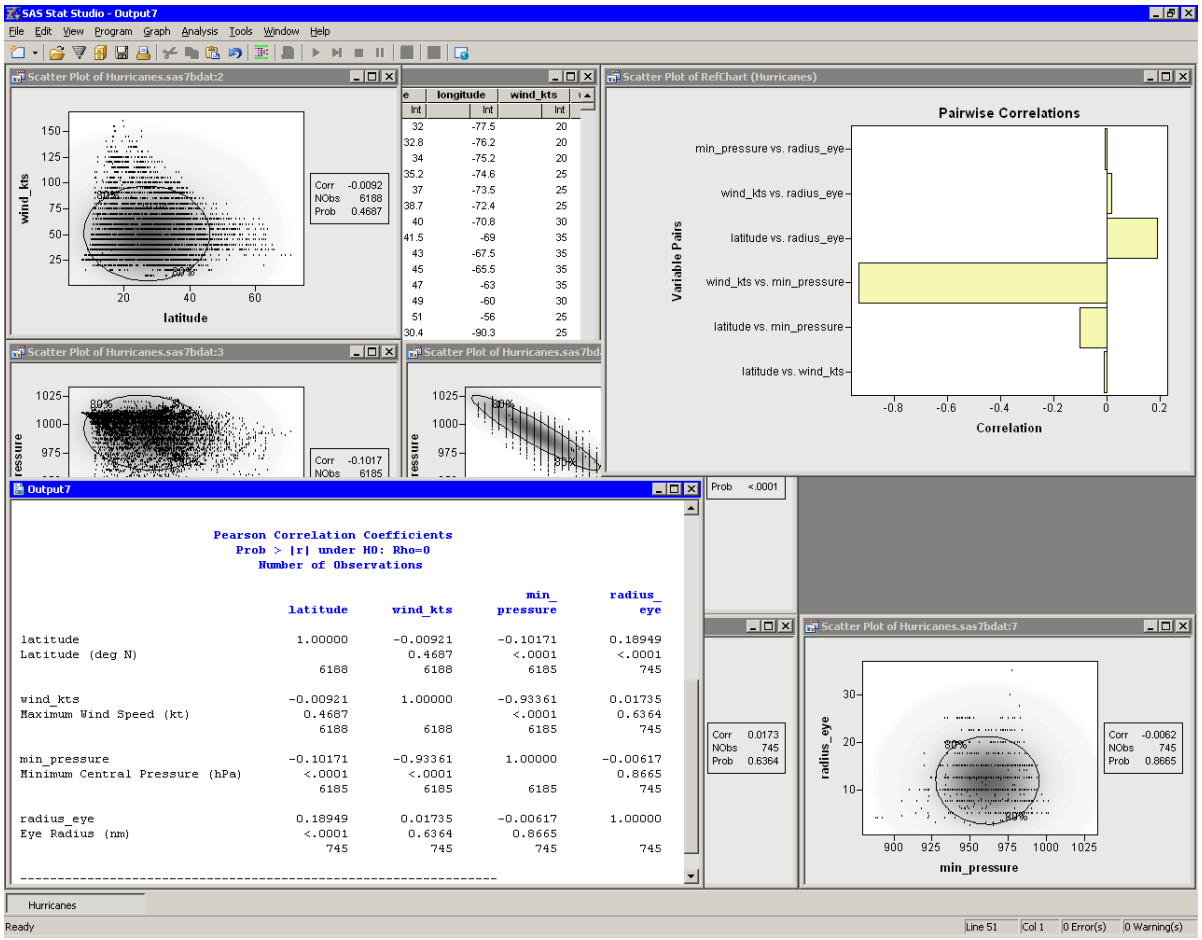
The "Pearson Correlation Coefficients" table displays the correlation coefficients between pairs of variables. In addition, the table gives the number of nonmissing observations for each pair of variables, and tests the hypothesis that the coefficient is zero.

Note that the number of observations used to compute the correlation coefficients can vary. For example, there are no missing values in the latitude of wind_kts variables, so the correlation coefficient for this pair is computed using all 6,188 observations in the data set. In contrast, only 745 values for radius_eye are nonmissing, reflecting the fact that not all cyclones have well-defined eyes.

For these data, the correlation between min_pressure and wind_kts is strong and negative, with a value near −0.93. This is not surprising, since winds are determined by a pressure gradient. Although not as strong, there is also negative correlation between latitude and min_pressure. In contrast, the correlation between latitude and radius_eye is positive. The correlation between the following pairs of variables is not significantly different from zero: latitude and wind_kts, radius_eye and wind_kts, and radius_eye and min_pressure.

These results are graphically summarized in the pairwise correlations plot, shown in the upper right corner of Figure 25.4. This plot is not linked to the original data set because it has a different number of observations. However, you can view the data for this plot by pressing the F9 key when the plot is active.

Figure 25.4 Output from a Correlation Analysis



Use the Workspace Explorer to View All Plots

Partly visible in Figure 25.4 is the matrix of pairwise scatter plots between the variables. Some of these plots are hidden by the output window and the pairwise correlation plot.

To use the Workspace Explorer to view all the scatter plots:

1. Close the pairwise correlation plot.
2. Press ALT+X to open the Workspace Explorer.

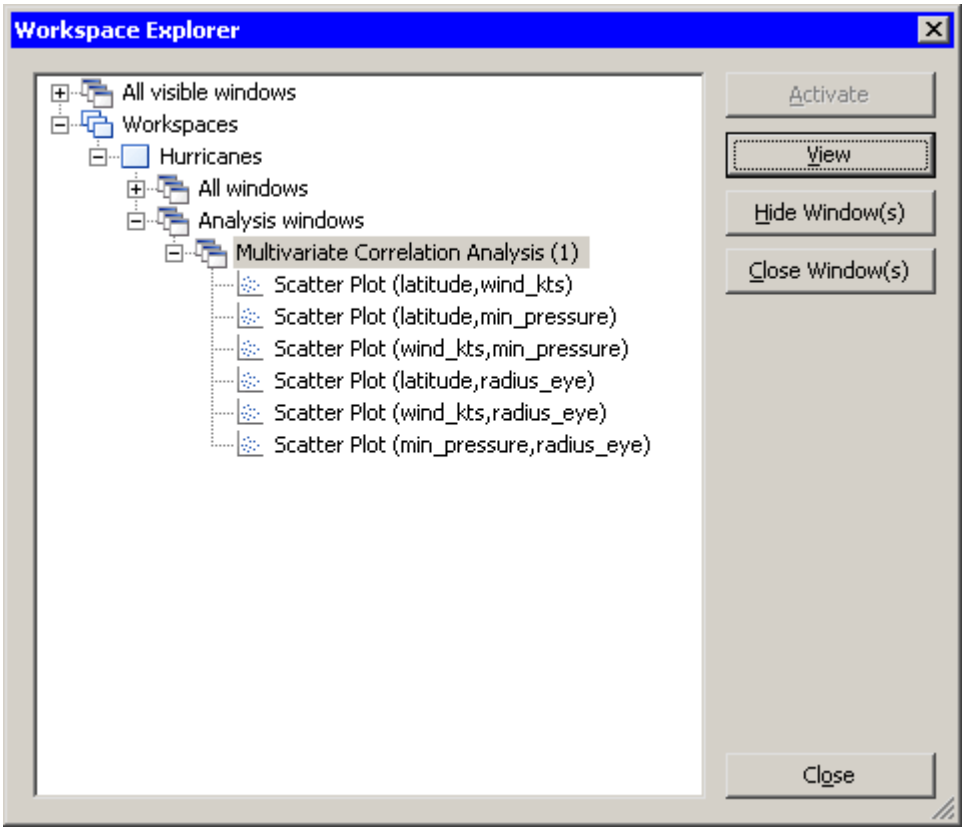
   You can use the Workspace Explorer to manage the display of plots. The Workspace Explorer is described in the section "Workspace Explorer" on page 190 of Chapter 11, "Techniques for Exploring Data."
3. Select the entry in the Workspace Explorer labeled **Multivariate Correlation Analysis**, as shown in Figure 25.5.

4 Click **View**.

The scatter plots that are associated with the analysis appear in front of other windows.

5 Click **Close** to close the Workspace Explorer.

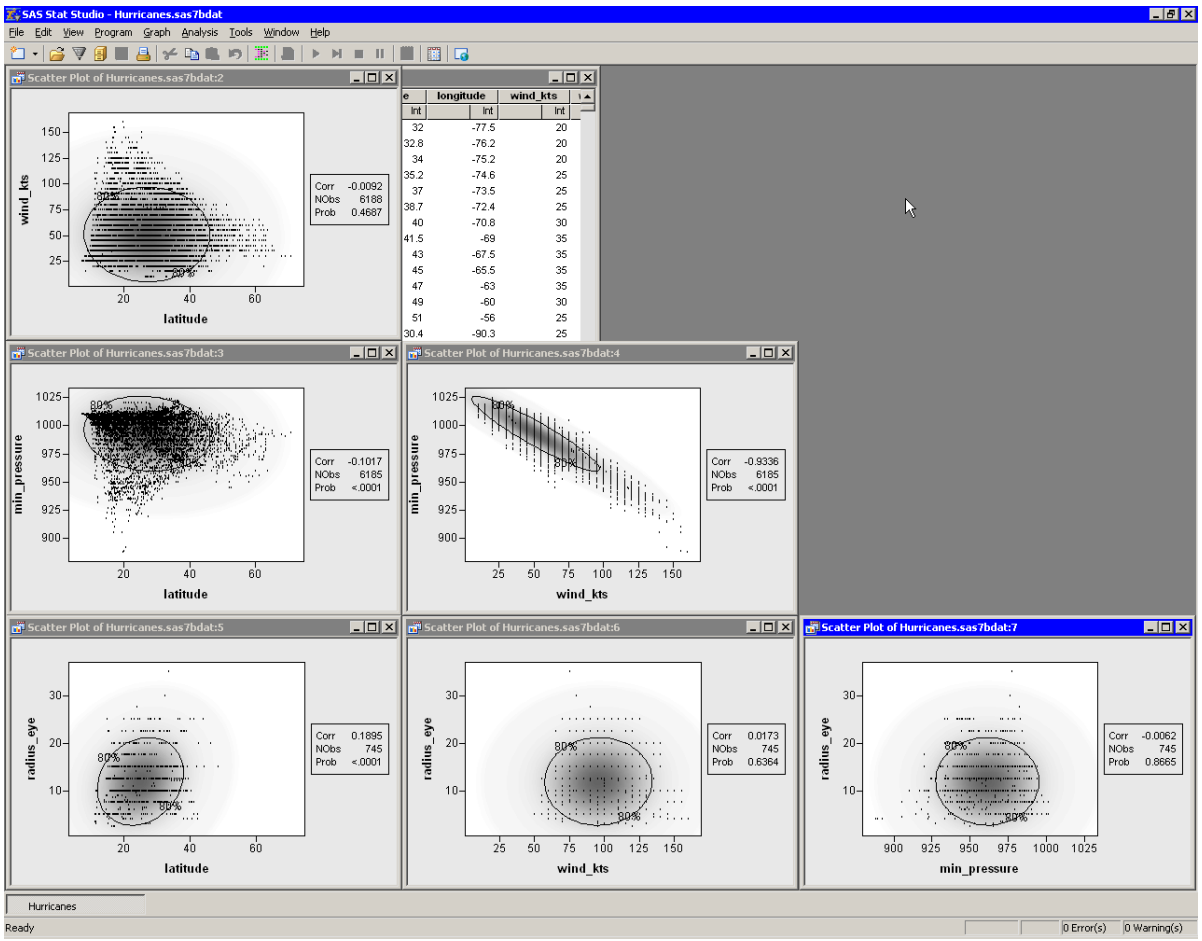
Figure 25.5 Selecting a Group of Plots



The workspace is now arranged as shown in Figure 25.6. 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 level. The plots also contain a gradient shading that indicates a nested sequence of ellipses. The darkest shading occurs at the bivariate means for each pair of variables. The lightest shading corresponds to 0.9999 probability.

Variables that are bivariate normal have most of their observations close to the bivariate mean and have a bivariate density that is proportional to the gradient shading. The plot of wind_kts versus latitude shows that these two variables are not bivariate normal. Similarly, min_pressure and latitude are not bivariate normal.

Figure 25.6 A Matrix of Scatter Plots



The variables wind_kts and min_pressure are highly correlated and linearly related. In contrast, wind_kts is not correlated with latitude or radius_eye, although you can still notice certain relationships:

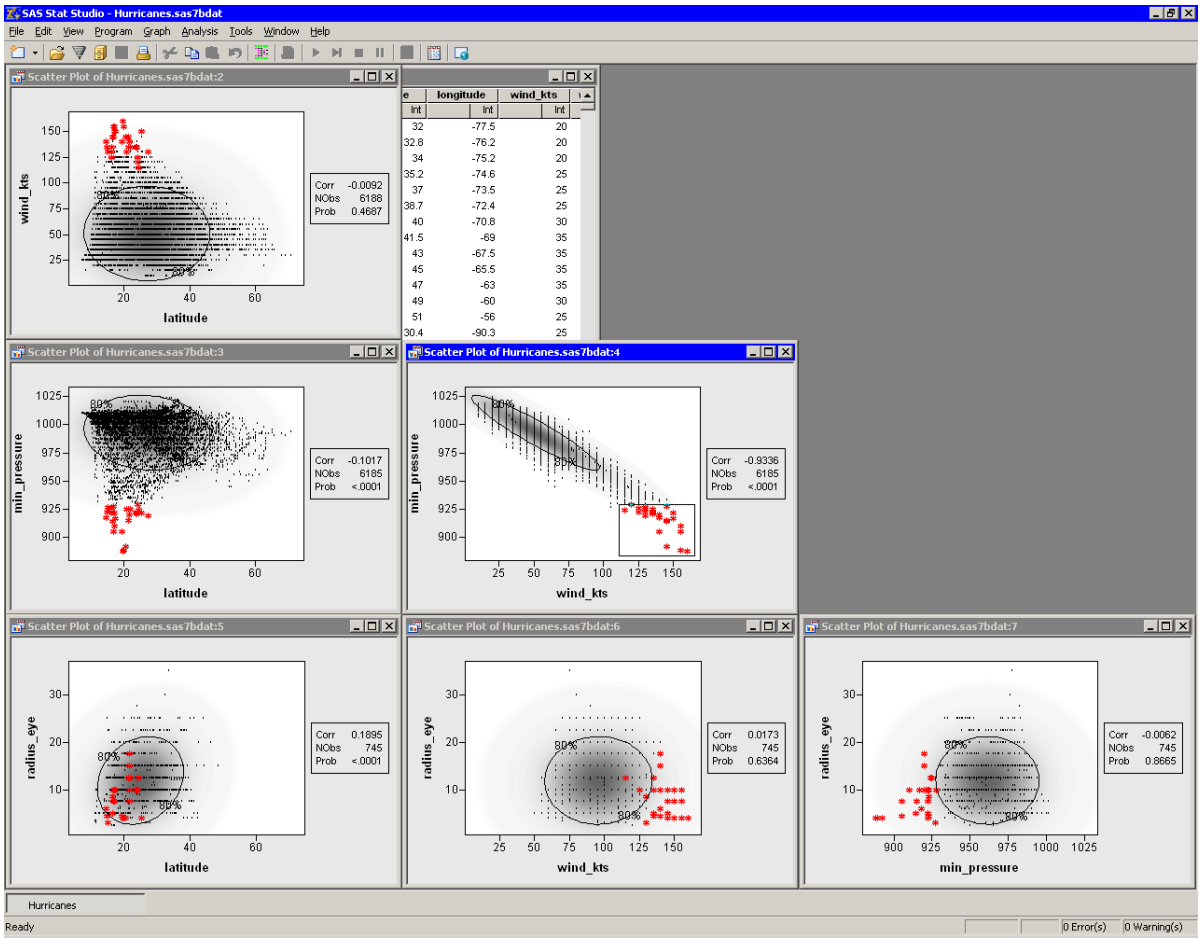
- Cyclones with high wind speeds occur only at lower latitudes.
- Cyclones north of 43 degrees of latitude tend to have wind speeds less than 75 knots.
- The size of a cyclone's eye seems to be unrelated to the speed of its winds.

You can observe similar relationships between min_pressure and the latitude and radius_eye variables.

The matrix of scatter plots also reveals an aspect of the data that might not be apparent from univariate plots. The plots that display wind_kts or radius_eye show a granular appearance that indicates the data are rounded. Most of the wind speed measurements are rounded to the nearest five knots, whereas the values for the eye radius are rounded to the nearest 2.5 nautical miles. (You can also find observations for these variables that are not rounded.)

Figure 25.7 shows another use of the scatter plot matrix. Some observations with extreme values of min_pressure and wind_kts are selected. The marker shape and color for these observations were changed to make them more noticeable. You can use this technique to investigate whether outliers for one pair of variables are, in fact, multivariate outliers with respect to multivariate normality. Most of the selected data in Figure 25.7 are inside the 80% ellipse for the radius_eye versus latitude scatter plot. This indicates that these data are not far from the mean in those variables. However, a few observations (corresponding to Hurricane Hugo when it was category 5) do appear to be multivariate outliers in these variables.

Figure 25.7 Selecting Bivariate Outliers



Specifying the Correlation Analysis

This section describes the dialog box tabs that are associated with the Correlation analysis. The Correlation analysis calls the CORR procedure in Base SAS software. See the CORR procedure documentation in the *Base SAS Procedures Guide* for additional details.

Variables Tab

You can use the **Variables** tab to specify the numerical variables for the analysis. The **Variables** tab is shown in Figure 25.2.

The variables in the **Y Variables** list correspond to variables in the VAR statement of the CORR procedure. The variables in the **X Variables (With)** list correspond to variables in the WITH statement of the CORR procedure.

The simplest way to analyze correlations is to add the variables of interest to the **Y Variables** list, as in the example earlier in this chapter.

If the **X Variables (With)** list is empty, the correlation matrix is symmetric. If you request a matrix of pairwise scatter plots (on the **Plots** tab), you will get plots for pairs of variables in the lower triangular portion of the matrix.

If the **X Variables (With)** list is not empty, the correlation matrix is not symmetric. If you specify $C_1, \ldots C_m$ as the Y variables and $R_1, \ldots R_n$ as the WITH variables, then the ijth cell of the correlation matrix will be the correlation of R_i with C_j. If you request a matrix of pairwise scatter plots, you will get nm plots, arranged in n rows and m columns.

The **Partial** list is rarely used. The variables in this list correspond to variables in the PARTIAL statement of the CORR procedure. A partial correlation measures the strength of a relationship between two variables, while controlling the effect of other variables. The Pearson partial correlation between two variables, after controlling for variables in the PARTIAL statement, is equivalent to the Pearson correlation between the residuals of the two variables after regression on the controlling variables.

If there are variables in the **Partial** list, then the following conditions hold:

- You cannot request Hoeffding's D correlation statistic.
- Observations with missing values are excluded from the analysis.

Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 25.3.) These plots do not add any variables to the data table.

The following plots are available:

Pairwise correlation plot
creates a bar chart that shows the Pearson correlation between pairs of variables.

Matrix of pairwise scatter plots
creates a matrix of scatter plots that shows bivariate data for pairs of variables. If you do not specify any X variables in the **X Variables (With)** list on the **Variables** tab, then you will get a lower triangular array of plots. If you do specify X variables, then you will get a rectangular array of plots. The inset added to each plot contains the following:

- the Pearson correlation coefficient
- the number of nonmissing observations for each pair of variables
- the p-value under the null hypothesis of zero correlation

Add prediction ellipse
adds a prediction ellipse to the scatter plot. The ellipse is calculated under the assumption that the data are bivariate normal. A prediction ellipse is a region for predicting a new observation in the population. It also approximates a region that contains a specified percentage of the population.

Confidence level
specifies the confidence level for the prediction ellipse.

Shade plot background by confidence level
specifies that the background of each scatter plot be shaded according to a nested family of prediction ellipses.

Tables Tab

The **Tables** tab is shown in Figure 25.8. You can use the **Tables** tab to display the following tables that summarize the results of the analysis:

Pearson's product-moment
displays a table of Pearson correlation coefficients. Selecting this field corresponds to the PEARSON option in the PROC CORR statement. Clearing this field corresponds to the NOCORR option in the PROC CORR statement.

Hoeffding's D
displays a table of Hoeffding's D statistic. This statistic is not available if you specify variables in the **Partial** list on the **Variables** tab. This corresponds to the HOEFFDING option in the PROC CORR statement.

Kendall's tau-b
displays a table of Kendall's tau-b statistic. This corresponds to the KENDALL option in the PROC CORR statement.

Spearman's rho
displays a table of Spearman's rank-order correlation. This corresponds to the SPEARMAN option in the PROC CORR statement.

Show significance probabilities for H0: correlation=0
displays p-values under the null hypothesis of zero correlation. Clearing this field corresponds to the NOPROB option in the PROC CORR statement.

Simple descriptive statistics

displays descriptive statistics for the variables in the analysis. Clearing this field corresponds to the NOSIMPLE option in the PROC CORR statement.

Covariances

displays the covariance matrix for the variables in the analysis. This corresponds to the COV option in the PROC CORR statement.

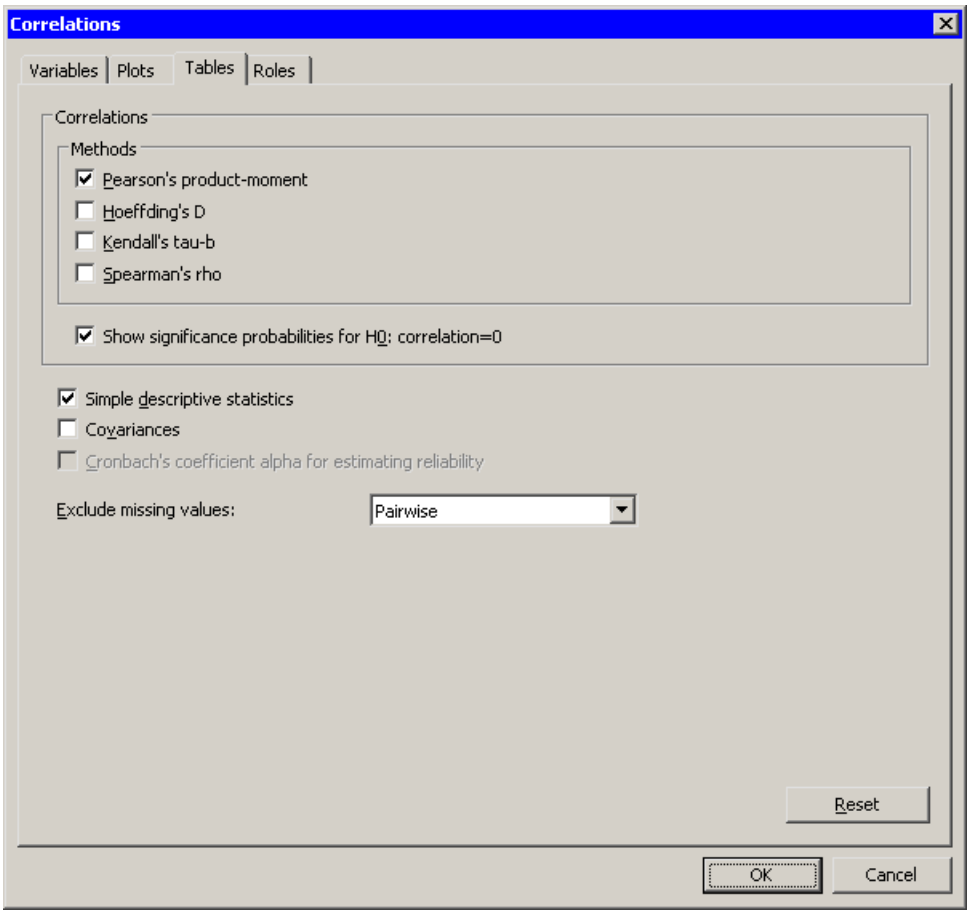
Cronbach's coefficient alpha for estimating reliability

displays Cronbach's coefficient alpha for the variables in the analysis. This corresponds to the ALPHA option in the PROC CORR statement. This statistic is not available if you specify variables in the **X Variables (With)** list on the **Variables** tab. This statistic is not available unless you select **Listwise** for **Exclude missing values**.

Exclude missing values

specifies how to treat missing values in the analysis. If you select **Listwise**, then observations with missing values are excluded from the analysis. This corresponds to the NOMISS option in the PROC CORR statement. Otherwise, statistics are computed using all of the nonmissing pairs of variables.

Figure 25.8 The Tables Tab



Roles Tab

You can use the **Roles** tab to specify a frequency variable or weight variable for the analysis.

A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents n observations, where n is the value of the frequency variable.

A weight variable is a numeric variable with values that are relative weights for the analysis.

Analysis of Selected Variables

If any numeric variables are selected in a data table when you run the analysis, these variables are automatically entered in the **Y Variables** field of the **Variables** tab.

Any variable in the data table with a Frequency or Weight role is automatically entered in the appropriate field of the **Roles** tab.

Chapter 26
Multivariate Analysis: Principal Component Analysis

Contents

Overview of the Principal Component Analysis **403**
Example: Reduce Dimensionality through Principal Component Analysis **404**
Biplots **413**
Specifying the Principal Component Analysis **415**
Variables Tab 415
Method Tab 416
Plots Tab 416
Tables Tab 417
Output Variables Tab 419
Roles Tab 420
Analysis of Selected Variables **420**
References **420**

Overview of the Principal Component Analysis

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 fewer dimensions. Each dimension is called a *principal component* and represents a linear combination of the original variables. The first principal component accounts for as much variation in the data as possible. Each subsequent principal component accounts for as much of the remaining variation as possible and is orthogonal to all of the previous principal components.

You can examine principal components to understand the sources of variation in your data. You can also use them in forming predictive models. If most of the variation in your data exists in a low-dimensional subset, you might be able to model your response variable in terms of the principal components. You can use principal components to reduce the number of variables in regression, clustering, and other statistical techniques.

You can run the Principal Component analysis by selecting **Analysis ►Multivariate Analysis ►Principal Component Analysis** from the main menu. The analysis is implemented by calling the PRINCOMP procedure in SAS/STAT software. See the PRINCOMP procedure documentation in the *SAS/STAT User's Guide* for additional details.

Example: Reduce Dimensionality through Principal Component Analysis

In this example, you compute principal components of several variables in the Baseball data set. The Baseball data set contains performance measures for major league baseball players in 1986. A full description of the Baseball data is included in Chapter A, "Sample Data Sets."

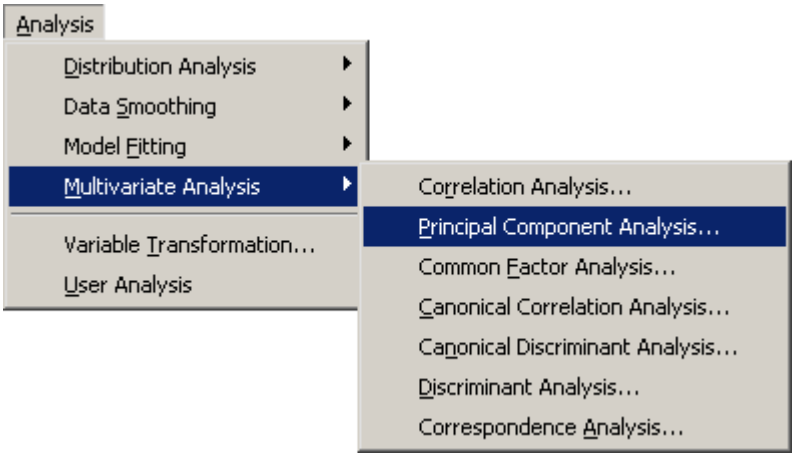
Suppose you are interested in exploring the sources of variation in players' performances during the 1986 season. There are six measures of players' batting performance: no_atbat, no_hits, no_home, no_runs, no_rbi, and no_bb. There are three measures of players' fielding performance: no_outs, no_assts, and no_error. These data form a nine-dimensional space. The goal of this example is to use principal component analysis to capture most of the variance of these data in a low-dimensional subspace—preferably in two or three dimensions. The subspace will be formed by the span of the first few principal components. (Recall that the *span* of a set of vectors is the vector space consisting of all linear combinations of the vectors.)

To run a principal component analysis:

1 Open the Baseball data set.

2 Select **Analysis ►Multivariate Analysis Principal Component Analysis** from the main menu, as shown in Figure 26.1.

Figure 26.1 Selecting the Principal Component Analysis



The Principal Component Analysis dialog box appears. (See Figure 26.2.) You can select variables for the analysis by using the **Variables** tab.

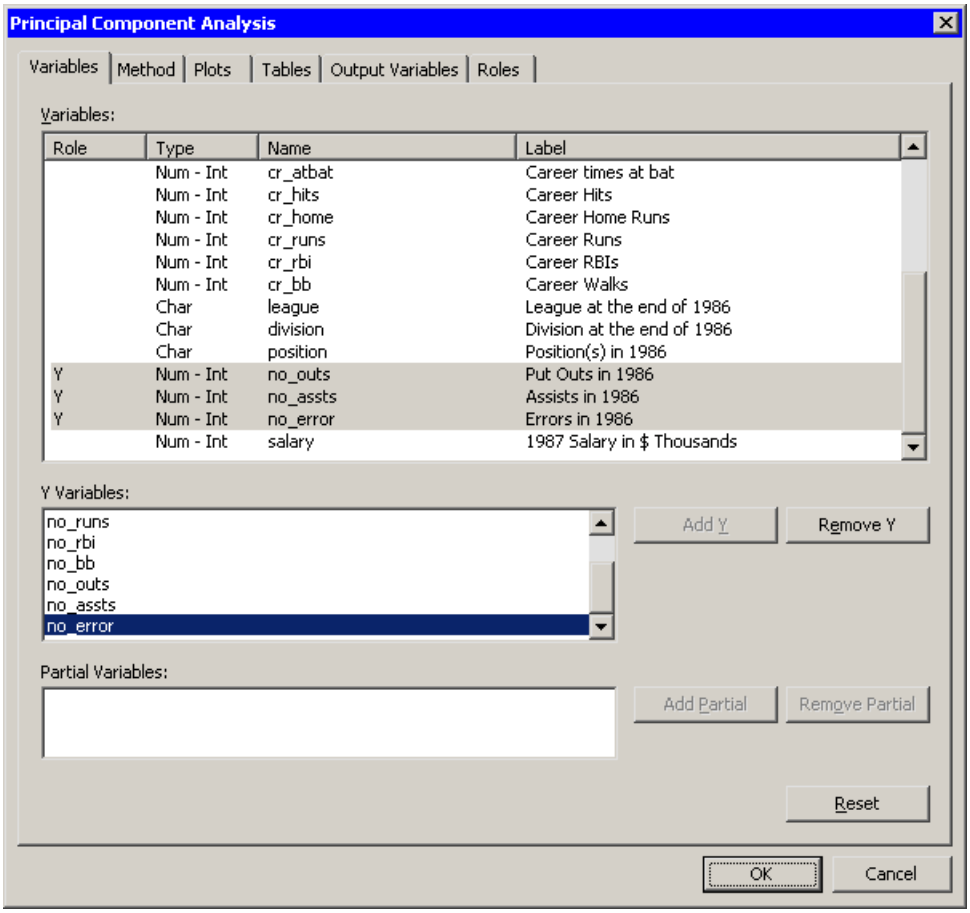
3 Select no_atbat. While holding down the CTRL key, select no_hits, no_home, no_runs, no_rbi, and no_bb. Click **Add Y**.

NOTE: Alternately, you can select the variables by using *contiguous selection*: click the first item, hold down the SHIFT key, and click the last item. All items between the first and last item are selected and can be added by clicking **Add Y**.

The three measures of fielding performance are located near the end of the list of variables.

4 Scroll to the end of the variable list. Select no_outs. While holding down the CTRL key, select no_assts and no_error. Click **Add Y**.

Figure 26.2 The Variables Tab



5 Click the **Method** tab.

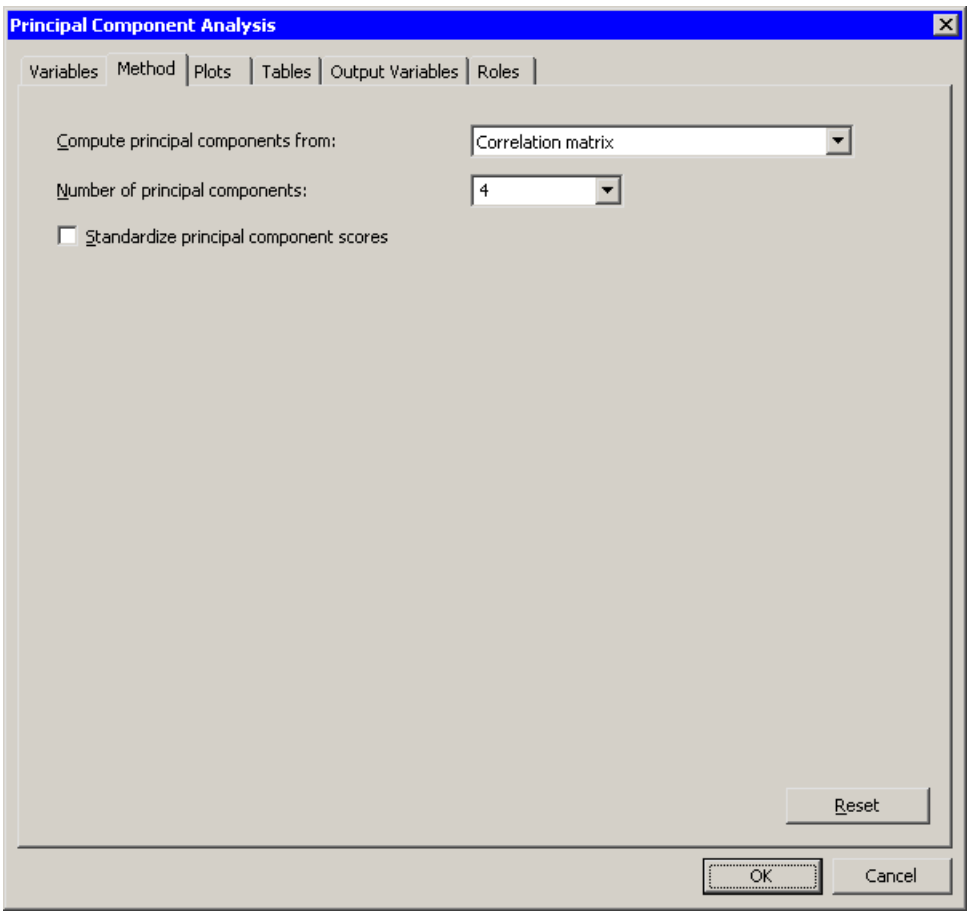
The **Method** tab becomes active. (See Figure 26.3.) You can use the **Method** tab to set options in the analysis.

By default, the analysis is carried out on the correlation matrix. The alternative is to use the covariance matrix. The covariance matrix is recommended only when all the variables are measured in comparable units. For this example, the correlation matrix is appropriate.

By default, the analysis computes all p principal components for the p variables selected in the **Variables** tab. It is often sufficient to compute a smaller number of principal components.

6 Set **Number of principal components** to `4`.

Figure 26.3 The Method Tab



7 Click the **Plots** tab.

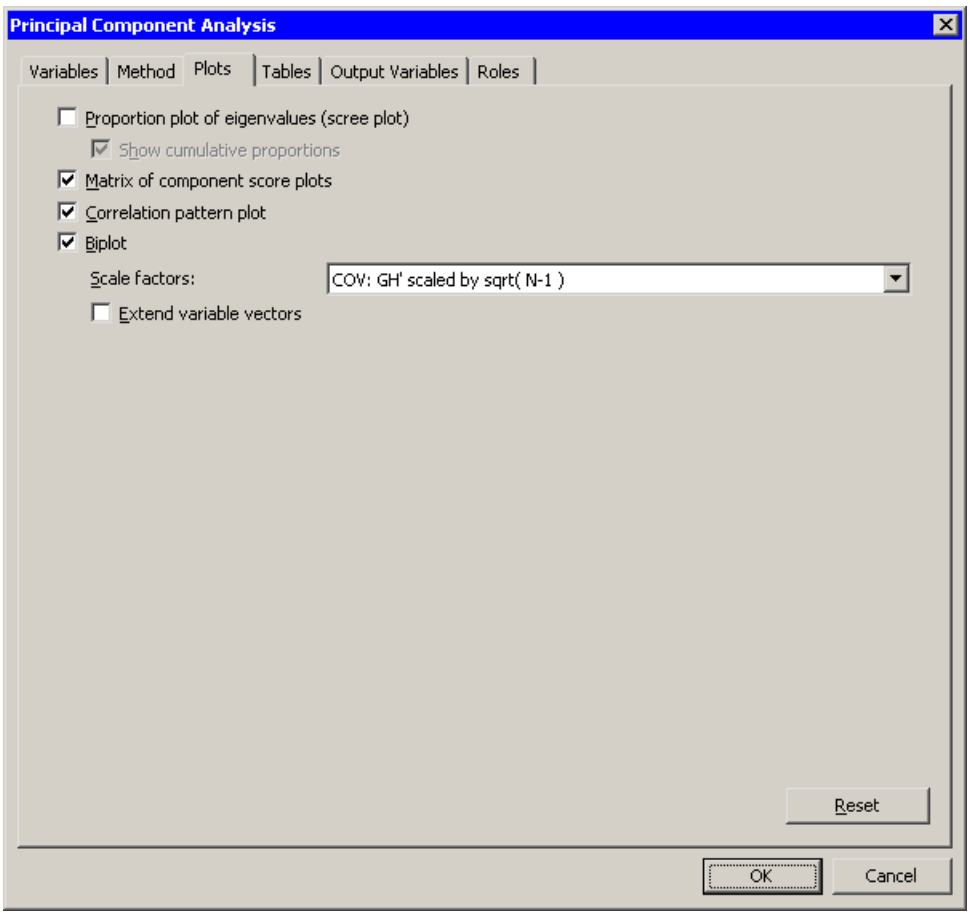
The **Plots** tab becomes active. (See Figure 26.4.)

8 Clear **Proportion plot of eigenvalues (scree plot).**

9 Select **Matrix of component score plots**.

10 Click **OK**.

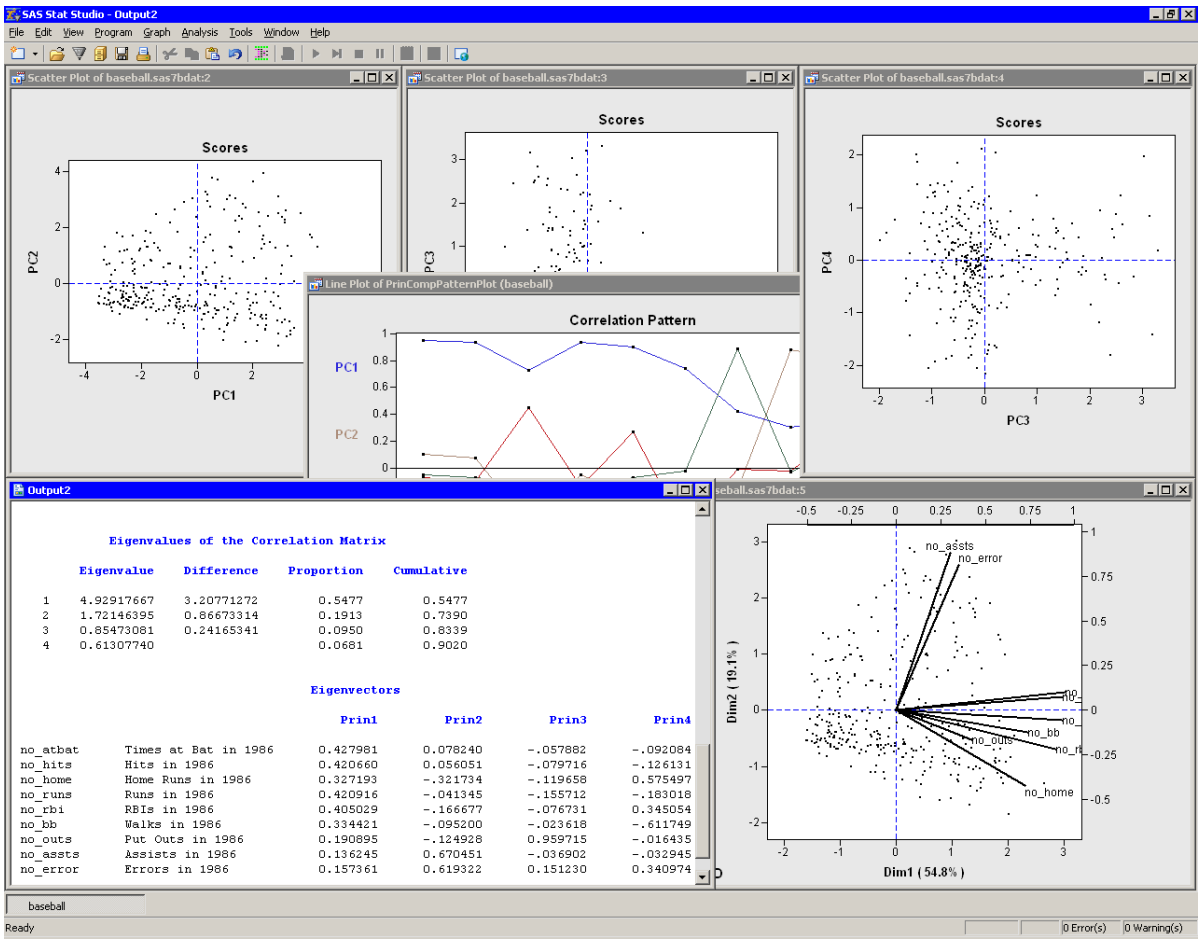
Figure 26.4 The Plots Tab



The analysis calls the PRINCOMP procedure, which uses the options specified in the dialog box. The procedure displays tables in the output document, as shown in Figure 26.5. The "Simple Statistics" table displays the mean and standard deviation for each variable. (The "Simple Statistics" table is not visible in Figure 26.5. You can scroll through the output window to view it.) The "Correlation Matrix" table (also not shown) displays the correlation between each pair of variables.

The "Eigenvalues of the Correlation Matrix" table 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. The eigenvalues correspond to 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 three principal components account for over 83% of the variation; the first four account for 90%.

Figure 26.5 Output from a Principal Component Analysis



The "Eigenvectors" table contains the first four eigenvectors of the correlation matrix. The eigenvectors are principal component vectors. The first column of the table corresponds to the first principal component, the second column to the second principal component, and so on. Each principal component is a linear combination of the Y variables. For example, the first principal component corresponds to the linear combination

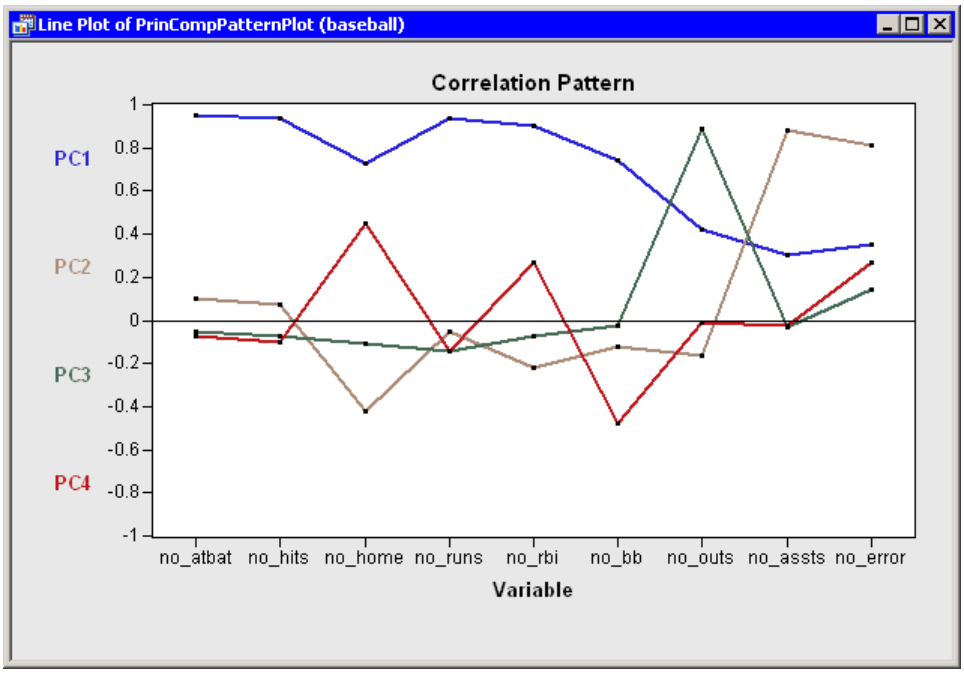
$$\mathrm{PC}_1 = 0.42798\ \text{no_atbat} + 0.42066\ \text{no_hits} + \ldots + 0.15736\ \text{no_error}$$

The first principal component (PC1) appears to be a weighted measure of the players' overall performance, as seen by the relative magnitudes of the coefficients. More weight is given to batting performance (the batting coefficients are in the range 0.33–0.43) than to fielding performance (the fielding coefficients are in the range 0.14–0.19). The second principal component (PC2) is primarily related to the no_assts and no_error variables. Players with large values of PC2 have many assists, but also relatively many errors. The third component (PC3) is primarily related to the no_outs variable. The fourth component is a contrast between no_home and no_bb (that is, between home runs and walks). This component separates players with many home runs and few walks from the players who often walk and rarely hit a home run.

You can use the correlation pattern plot to examine correlations between the principal components and the original variables. (See Figure 26.6.)

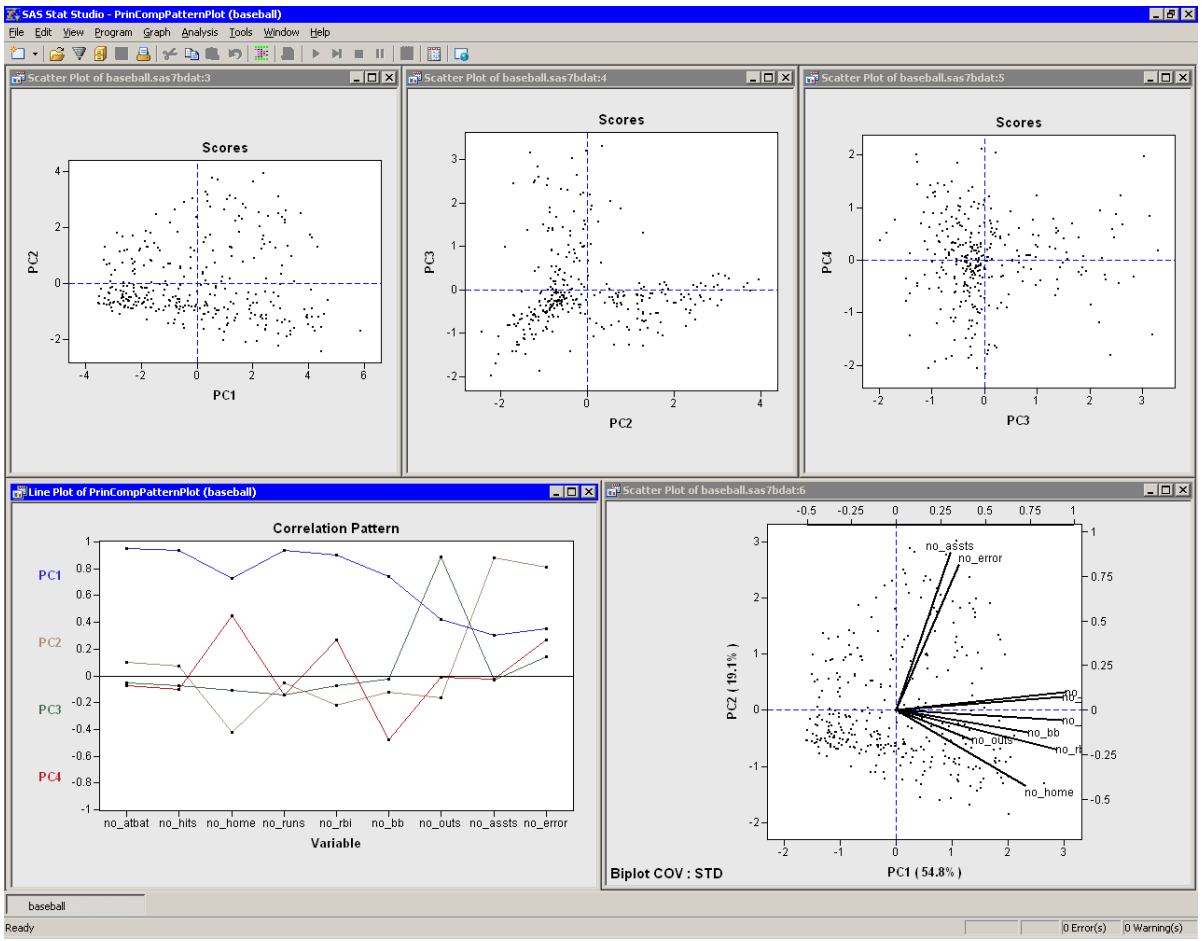
For example, the first principal component (PC1) is positively correlated with all of the original variables. It is correlated more with batting performance than with the fielding variables.

Figure 26.6 Correlation Pattern Plot



The relationship between the original variables and observations is shown in the biplot, at the lower right of Figure 26.7. The line segments represent the projection of a vector in the direction of each original variable onto a two-dimensional subspace. The points in the biplot are the projection of the observations onto the same two-dimensional subspace. The section "Biplots" on page 413 discusses biplots in further detail.

Figure 26.7 Graphs from a Principal Component Analysis



The plots tiled across the top of Figure 26.7 are called *score plots*. These are plots of the observations in the coordinate system defined by the principal components. For these data, each observation represents a player.

To set the value of the name variable to be the label you see when you click an observation:

11 Click the score plot of PC2 versus PC1 to activate it.

12 Press the F9 key to display the data table that is associated with this plot.

13 Right-click the variable heading for name to display the **Variables** menu. Select **Label**.

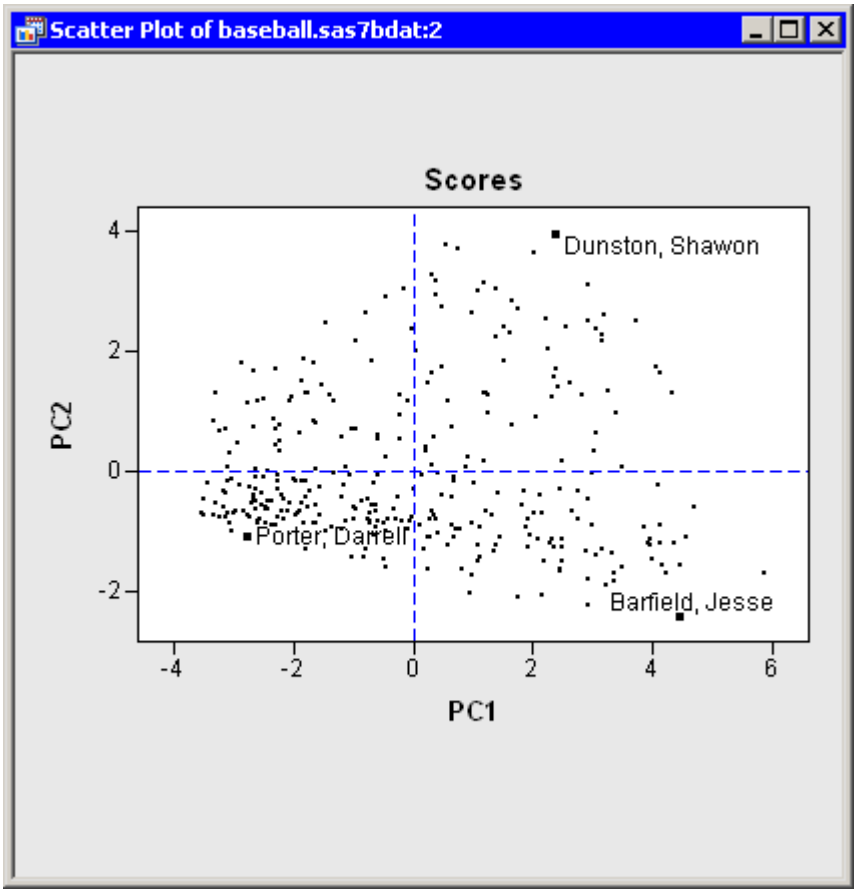
14 Click in the upper left cell of the data table to deselect the variable.

15 Close the data table.

16 Click some observations in the score plot of PC2 versus PC1, as shown in Figure 26.8.

The first principal component measures a player's hitting performance during the 1986 season. Consequently, players to the right (such as Jesse Barfield) had strong hitting statistics, whereas players to the left (such as Darrell Porter) had weaker statistics. The second principal component primarily measures the number of assists (and errors) for each player. Consequently, players near the top of the plot (such as Shawon Dunston) have many assists, whereas players near the bottom (such as Jesse Barfield) have few.

Figure 26.8 Score Plot of First Two Principal Components



The score plot of the second and third principal components is interesting because it compares two different measures of fielding performance. Also, there are few players in the first quadrant of the plot. Recall that the third principal component primarily measures the no_outs variable. This variable records *putouts*. Common situations leading to a putout include tagging or forcing out a base runner, catching a fly ball, or (for catchers) catching a third strike. The opportunities for a player to get a putout or an assist are highly dependent on the player's position.

Figure 26.9 shows the score plot for the positions of second base, third base, and shortstop. Note that these observations primarily lie in the fourth quadrant. These players have many assists because they often field ground balls and throw to first base, but they have relatively few opportunities to put out runners themselves. In contrast, Figure 26.10 shows the score plot for outfielders and designated hitters. These observations lie in the third quadrant. These players have few assists and relatively few putouts. (The outfielders are credited with a putout when they catch a fly ball, but there are many fewer fly balls than ground balls in a typical game.) Catchers and first basemen (not shown) have scores primarily in the second quadrant of the plot, corresponding to many putouts but few assists.

Figure 26.9 Fielding Scores for Some Infielders

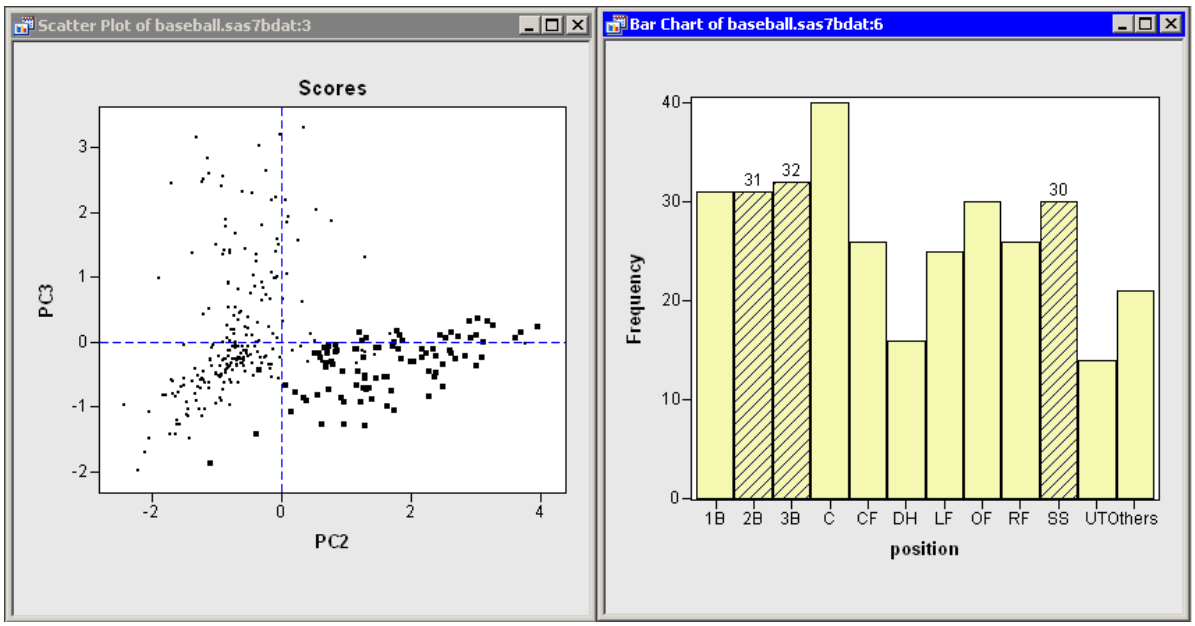
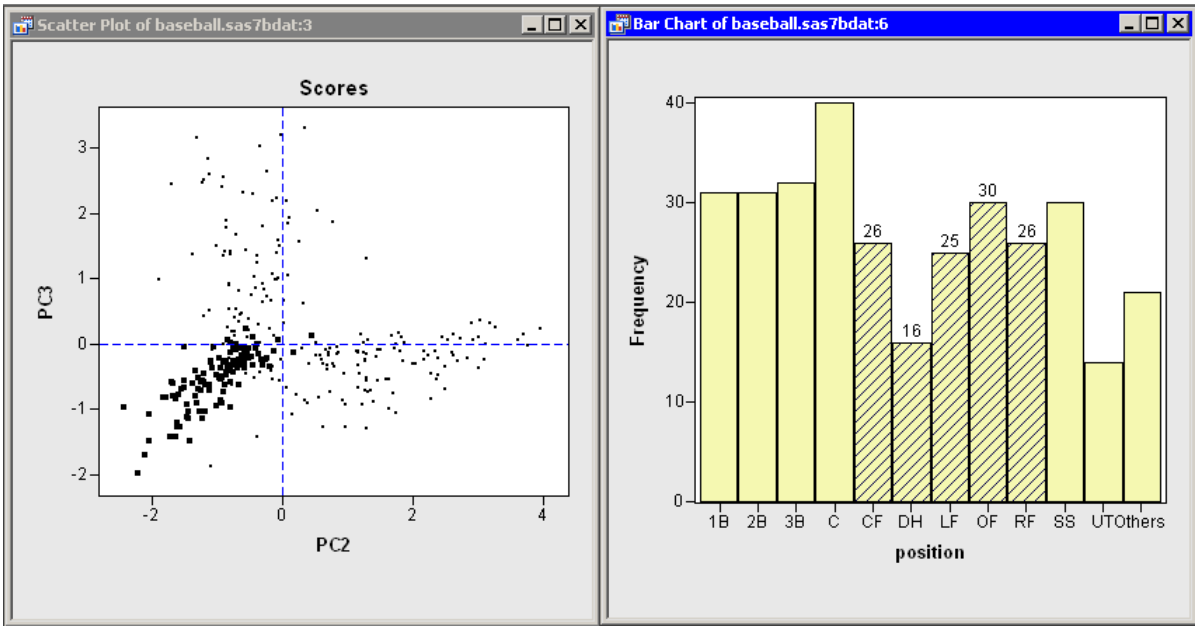


Figure 26.10 Fielding Scores for Outfielders and Designated Hitters



In summary, the analysis shows that most of the variation in these data occurs in the first principal component: an overall measure of batting performance. The next two principal components incorporate variation due to fielding performance. Figure 26.9 and Figure 26.10 show that the source of this fielding variation is differences in player positions. Together, these three components account for 83% of the variation in the nine-dimensional space of the original variables.

Principal components can also be used as explanatory variables in regression. For example, you could examine how well overall batting performance in 1986 predicts a player's salary by using PC1 as an explanatory variable in a regression model.

Biplots

A *biplot* is a display that attempts to represent both the observations and variables of multivariate data in the same plot. SAS/IML Studio provides biplots as part of the Principal Component analysis.

The computation of biplots in SAS/IML Studio follows the presentation given in Friendly (1991) and Jackson (1991). Detailed discussions of how to compute and interpret biplots are available in Gabriel (1971) and Gower and Hand (1996).

The computation of a biplot begins with the data matrix. If you choose to compute principal components from the covariance matrix (on the **Method** tab; see Figure 26.3), then the data matrix is centered by subtracting the mean of each column. Otherwise, it is standardized so that each variable has zero mean and unit standard deviation.

In either case, let X denote the resulting $N \times p$ matrix. The singular value decomposition (SVD) of X is the factorization

$$\begin{aligned} X &= ULV' \\ &= \left(UL^{\alpha}\right)\left(L^{1-\alpha}V'\right) \\ &= GH' \end{aligned}$$

where L is the diagonal matrix of singular values. If you replace G and H with their first two columns, then an approximate relationship exists: $X \approx GH'$. This is a rank-two approximation of X. In fact, it is the closest rank-two approximation to X in a least squares sense (Golub and Van Loan 1989).

In a biplot, the rows of the $N \times 2$ matrix G are plotted as points, which correspond to observations. The rows of the $p \times 2$ matrix H are plotted as vectors, which correspond to variables.

The choice of α determines the scaling of the observations and vectors in the biplot. In general, it is impossible to accurately represent the variables and observations in only two dimensions, but you can choose values of α that preserve certain properties of the high-dimensional data. Common choices are $\alpha = 0, 1/2$, and 1. SAS/IML Studio implements four different versions of the biplot:

GH′ This factorization uses $\alpha = 0$. This biplot attempts to preserve relationships between variables. This biplot has two useful properties:

- The length of a vector (a row of H) is proportional to the variance of the corresponding variable.
- The Euclidean distance between the ith and jth rows of G is proportional to the Mahalanobis distance between the ith and jth observations in the data set.

JK′ This factorization uses $\alpha = 1$. This biplot attempts to preserve the distance between observations. This biplot has two useful properties:

- The positions of the points in the biplot are identical to the score plot of first two principal components.
- The Euclidean distance between the ith and jth rows of G is equal to the Euclidean distance between the ith and jth observations in the data set.

SYM This factorization uses $\alpha = 1/2$. This biplot treats observations and variables symmetrically. This biplot attempts to preserve the values of observations.

COV This factorization uses $\alpha = 0$, but also multiplies G by $\sqrt{N-1}$ and divides H by the same quantity. This biplot has two useful properties:

- The length of a vector (a row of H) is equal to the variance of the corresponding variable.
- The Euclidean distance between the ith and jth rows of G is equal to the Mahalanobis distance between the ith and jth observations in the data set.

The axes at the bottom and left of the biplot are the coordinate axes for the observations. The axes at the top and right of the biplot are the coordinate axes for the vectors.

If the data matrix X is not well approximated by a rank-two matrix, then the visual information in the biplot is not a good approximation to the data. In this case, you should not try to interpret the biplot. However, if X is close to a rank-two matrix, then you can interpret a biplot in the following ways:

- The cosine of the angle between a vector and an axis indicates the importance of the contribution of the corresponding variable to the axis dimension.
- The cosine of the angle between vectors indicates correlation between variables. Highly correlated variables point in the same direction; uncorrelated variables are at right angles to each other.
- Points that are close to each other in the biplot represent observations with similar values.
- You can approximate the coordinates of an observation by projecting the point onto the variable vectors within the biplot.

For example, in Figure 26.11 the two principal components account for approximately 74% of the variance in the data. This means that the biplot is a fair (but not good) approximation to the data. The footnote in the plot indicates that the biplot is based on the COV factorization and that the data matrix was standardized (STD).

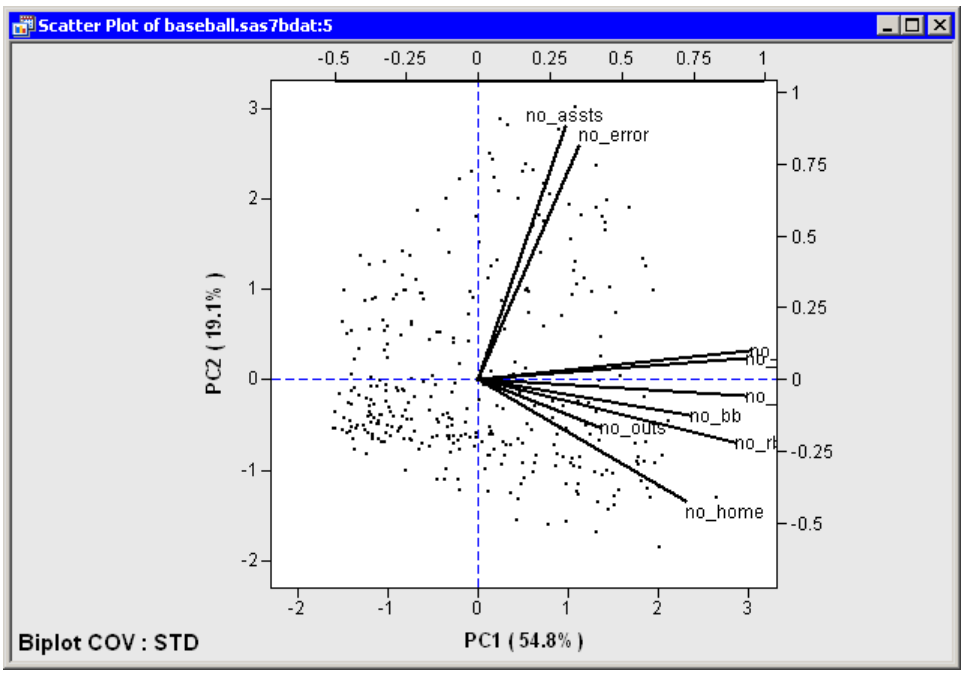
The variables are grouped: the hitting variables point primarily in the direction of the horizontal axis; no_assts and no_error point primarily in the direction of the vertical axis. The no_outs vector is much shorter than the other vectors, which often indicates that the vector does not lie near the span of the two biplot dimensions.

The hitting variables are strongly correlated with each other. The variables no_assts and no_error are correlated with each other, but they are not correlated with the hitting variables or with no_outs.

Because the biplot is only a moderately good approximation to the data, the following statements are *approximately* true:

- The first and fourth quadrants contain players who tend to be strong hitters. The other quadrants contain weak hitters.
- The first and second quadrants contain players who tend to have many assists and errors. The other quadrants contain players with few assists and errors.

Figure 26.11 Biplot for Baseball Data



Specifying the Principal Component Analysis

This section describes the dialog box tabs that are associated with the Principal Component analysis. The Principal Component analysis calls the PRINCOMP procedure in SAS/STAT software. See the PRINCOMP procedure documentation in the *SAS/STAT User's Guide* for additional details.

Variables Tab

You can use the **Variables** tab to specify the numerical variables for the analysis. The **Variables** tab is shown in Figure 26.2. The variables in the **Y Variables** list correspond to variables in the VAR statement of the PRINCOMP procedure.

The **Partial** list is rarely used. The variables in this list correspond to variables in the PARTIAL statement of the PRINCOMP procedure. The PRINCOMP procedure computes the principal components of the residuals from the prediction of the VAR variables by the PARTIAL variables.

Method Tab

You can use the **Method** tab to set options in the analysis. (See Figure 26.3.) Each UI control in the tab corresponds to an option in the PRINCOMP procedure. The **Rotation** tab contains the following controls:

Compute principal components from
specifies whether the principal components are computed for the correlation matrix or the covariance matrix. This corresponds to the COV option in the PROC PRINCOMP statement.

Number of principal components
specifies how many principal components to compute. This corresponds to the N= option in the PROC PRINCOMP statement. You can type in this field. If you want five principal components, you can type `5` even though this is not an option in the list.

Standardize principal component scores
specifies whether to standardize the principal component score. This corresponds to the STANDARD option in the PROC PRINCOMP statement. If you clear this option, the scores have variance equal to the corresponding eigenvalue.

Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 26.4.)

Creating a plot often adds one or more variables to the data table. The following plots are available:

Proportion plot of eigenvalues (scree plot)
creates a plot that summarizes the eigenvalues of the correlation or covariance matrix.

Show cumulative proportions
adds cumulative proportions of eigenvalues to the proportion plot.

Matrix of component score plots
creates a matrix of scatter plots that shows scores for consecutive pairs of principal components.

Correlation pattern plot
creates a line plot that shows the correlations between principal components and the original variables.

Biplot

creates a biplot. A biplot shows relationships between observations and variables in a single plot.

Scale factors

specifies how to scale and factor the SVD of the data matrix. The scaling determines the values for the biplot. The methods are described in the section "Biplots" on page 413.

Extend variable vectors

specifies whether to extend the vectors to the edge of the biplot. This is useful for visualizing the direction of short vectors.

Tables Tab

The **Tables** tab is shown in Figure 26.12. You can use the **Tables** tab to display the following tables that summarize the results of the analysis:

Simple descriptive statistics

specifies whether to display the mean and standard deviation for each variable.

Correlation or covariance matrix

specifies whether to display the correlation or covariance matrix, as selected on the **Method** tab.

Eigenvalues

specifies whether to display the eigenvalues of the correlation or covariance matrix, as well as the difference between successive eigenvalues, the proportion of variance explained by each eigenvalue, and the cumulative proportion of variance explained.

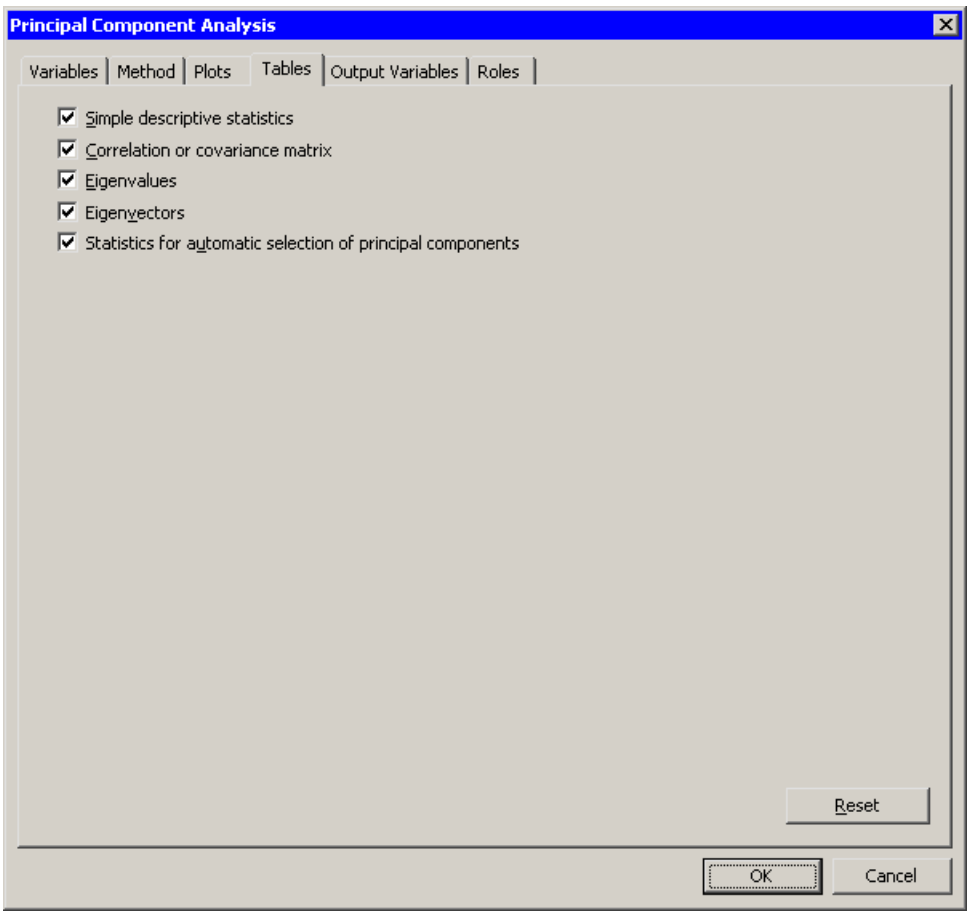
Eigenvectors

specifies whether to display the eigenvectors of the correlation or covariance matrix. The eigenvectors are used to form the principal components.

Statistics for automatic selection of principal components

specifies whether to display statistics that indicate how many principal components are needed to represent the p-dimensional data. This table is displayed only if you request at least as many principal components as there are variables.

Figure 26.12 The Tables Tab



A primary use of principal component analysis is to represent p-dimensional data in $k < p$ dimensions. In practice, it is often difficult to determine the best choice for k. The "Automatic Selection of Principal Components" table, shown in Figure 26.13, is provided to help you choose k. Numerous papers have been written comparing various methods for choosing k, but no method has shown itself to be superior. The following list briefly describes each method reported in the table. Jackson (1991, p. 41–51) gives further details.

Parallel Analysis
generates random data sets with N observations and p variables. The variables are normally distributed and uncorrelated. The method chooses k to be the largest integer for which the scree plot of the original data lies above the graph of the upper 95 percentiles of the eigenvalues of the random data.

Broken Stick
retains components that explain more variance than would be expected by randomly dividing the variance into p parts.

Average Root
keeps components that explain more variance than the mean of the eigenvalues.

0.7 * Average Root
keeps components that explain more variance than 0.7 times the mean of the eigenvalues.

Imbedded Error

chooses k to be the value that minimizes a certain function of the eigenvalues.

Velicer's MAP

chooses k to minimizes a certain function that involves partial correlations. This method is called Velicer's minimum average partial (MAP) test or Velicer's partial correlation procedure.

Figure 26.13 How Many Principal Components Are Needed?

Output12

	Prin5	Prin6	Prin7	Prin8	Prin9
no_atbat	-.332978	-.043523	-.146560	0.622494	-.526435
no_hits	-.452628	-.069845	-.154869	0.028800	0.750384
no_home	0.329397	0.173334	0.382778	0.341589	0.205348
no_runs	-.164848	-.192787	0.630909	-.489081	-.248577
no_rbi	0.062365	0.165752	-.601056	-.493536	-.216819
no_bb	0.681834	-.049954	-.145221	0.090375	0.087489
no_outs	-.059937	0.118362	0.088145	-.035222	-.004084
no_assts	0.082391	0.708188	0.135287	-.052652	0.009641
no_error	0.263455	-.616853	-.039259	-.005483	0.032510

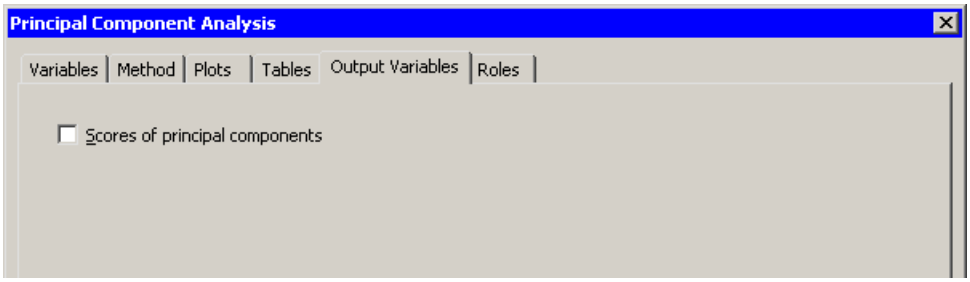
Automatic Selection of Principal Components

Method	Num Suggested PCs
Parallel Analysis	2
Broken Stick	1
Average Root	2
0.7 * Average Root	3
Imbedded Error	8
Velicer's MAP	1

Output Variables Tab

You can use the **Output Variables** tab to add principal component scores to the data table. (See Figure 26.14.) The options on the **Method** tab determine the number of scores and whether the scores are standardized.

Figure 26.14 The Output Variables Tab



Roles Tab

You can use the **Roles** tab to specify a frequency variable or weight variable for the analysis.

A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents n observations, where n is the value of the frequency variable.

A weight variable is a numeric variable with values that are relative weights for the analysis.

Analysis of Selected Variables

If any numeric variables are selected in a data table when you run the analysis, these variables are automatically entered in the **Y Variables** field of the **Variables** tab.

Any variable in the data table with a Frequency or Weight role is automatically entered in the appropriate field of the **Roles** tab.

References

Friendly, M. (1991), *SAS System for Statistical Graphics*, SAS Series in Statistical Applications, Cary, NC: SAS Institute Inc.

Gabriel, K. R. (1971), "The Biplot Graphical Display of Matrices with Applications to Principal Component Analysis," *Biometrika*, 58, 453–467.

Golub, G. H. and Van Loan, C. F. (1989), *Matrix Computations*, 2nd Edition, Baltimore: Johns Hopkins University Press.

Gower, J. C. and Hand, D. J. (1996), *Biplots*, London: Chapman & Hall.

Jackson, J. E. (1991), *A User's Guide to Principal Components*, New York: John Wiley & Sons.

Chapter 27
Multivariate Analysis: Factor Analysis

Contents

Overview of Factor Analysis **421**
Example: Reduce Dimensionality through Common Factor Analysis **423**
Specifying the Factor Analysis **434**
Variables Tab 435
Method Tab 435
Rotation Tab 436
Plots Tab 436
Tables Tab 437
Output Variables Tab 438
Roles Tab 438
Analysis of Selected Variables **438**
References **439**

Overview of Factor Analysis

Like principal component analysis, *common factor analysis* is a technique for reducing the complexity of high-dimensional data. (For brevity, this chapter refers to common factor analysis as simply "factor analysis.") However, the techniques differ in how they construct a subspace of reduced dimensionality. Jackson (1981, 1991) provides an excellent comparison of the two methods.

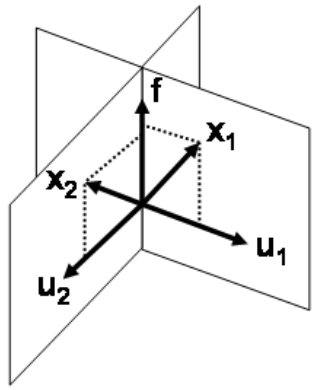
Principal component analysis chooses a coordinate system for the vector space spanned by the variables. (Recall that the *span* of a set of vectors is the vector space consisting of all linear combinations of the vectors.) The first principal component points in the direction of maximum variation in the data. Subsequent components account for as much of the remaining variation as possible while being orthogonal to all of the previous principal components. Each principal component is a linear combination of the original variables. Dimensional reduction is achieved by ignoring dimensions that do not explain much variation.

While principal component analysis explains variability, factor analysis explains correlation. Suppose two variables, $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, are correlated, but not collinear. Factor analysis assumes the existence of an unobserved variable (often called a *latent variable*) that is linearly related to $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, and explains the correlation between them. The goal of factor analysis is to estimate this latent variable from the structure of the original variables. An estimate of the unobserved variable is called a *common factor*.

The geometry of the relationship between the original variables and the common factor is illustrated in Figure 27.1. (The figure is based on a similar figure in Wickens (1995), as is the following description of the geometry.) The correlated variables $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are shown schematically in the figure. Each vector is

decomposed into a linear combination of a common factor and a *unique factor*. That is, $\boldsymbol{x_i} = c_i \boldsymbol{f} + d_i \boldsymbol{u_i}$, $i = 1, 2$. The unique factors, $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$, are uncorrelated with the common factor, $\boldsymbol{f}$, and with each other. Note that $\boldsymbol{f}$, $\boldsymbol{u_1}$, and $\boldsymbol{u_2}$ are mutually orthogonal in the figure.

Figure 27.1 The Geometry of Factor Analysis



In contrast to principal components, a factor is not, in general, a linear combination of the original variables. Furthermore, a principal component analysis depends only on the data, whereas a factor analysis requires fitting the theoretical structure in the previous paragraph to the observed data.

If there are p variables and you postulate the existence of m common factors, then each variable is represented as a linear combination of the m common factors and a single unique factor. Since the unique factors are uncorrelated with the common factors and with each other, factor analysis requires $m + p$ dimensions. (Figure 27.1 illustrates the case $p = 2$ and $m = 1$.) However, the orthogonality of the unique factors means that the geometry is readily understood by projecting the original variables onto the span of the m factors (called the *factor space*). A graph of this projection is called a *pattern plot*. In Figure 27.1, the pattern plot is the two points on $\boldsymbol{f}$ obtained by projecting $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ onto $\boldsymbol{f}$.

The length of the projection of an original variable $\boldsymbol{x}$ onto the factor space indicates the proportion of the variability of $\boldsymbol{x}$ that is shared with the other variables. This proportion is called the *communality*. Consequently, the variance of each original variable is the sum of the common variance (represented by the communality) and the variance of the unique factor for that variable. In a pattern plot, the communality is the squared distance from the origin to a point.

In factor analysis, the common factors are not unique. Typically an initial orthonormal set of common factors is computed, but then these factors are rotated so that the factors are more easily interpreted in terms of the original variables. An orthogonal rotation preserves the orthonormality of the factors; an oblique transformation introduces correlations among one or more factors.

You can run the Factor analysis in SAS/IML Studio by selecting **Analysis ►Multivariate Analysis ►Factor Analysis** from the main menu. The analysis is implemented by calling the FACTOR procedure in SAS/STAT software. See the FACTOR procedure documentation in the *SAS/STAT User's Guide* for additional details.

The FACTOR procedure provides several methods of estimating the common factors and the communalities. Since an $(m + p)$-dimensional model is fit by using the original p variables, you should interpret the results with caution. The following list describes special issues that can occur:

- Some of the eigenvalues of the *reduced correlation matrix* might be negative. A reduced correlation matrix is the correlation matrix of the original variables, except that the 1's on the diagonal are replaced

by prior communality estimates. These estimates are less than 1, and so the reduced correlation matrix might not be positive definite. In this case, the factors that correspond to the largest eigenvalues might account for more than 100% of the common variance.

- The communalities are the proportions of the variance of the original variables that can be attributed to the common factors. As such, the communalities should be in the interval [0, 1]. However, factor analyses that use iterative fitting estimate the communality at each iteration. For some data, the estimate might equal (or exceed) 1 before the analysis has converged to a solution. This is known as a Heywood (or an ultra-Heywood) case, and it implies that one or more unique factor has a nonpositive variance. When this occurs, the factor analysis stops iterating and reports an error.

These and other issues are described in the section "Heywood Cases and Other Anomalies about Communality Estimates" in the documentation for the FACTOR procedure.

You can use many different methods to perform a factor analysis. Two popular methods are the principal factor method and the maximum likelihood method. The principal factor method is computationally efficient and has similarities to principal component analysis. The maximum likelihood (ML) method is an iterative method that is computationally more demanding and is prone to Heywood cases, nonconvergence, and multiple optimal solutions. However, the ML method also provides statistics such as standard errors and confidence limits that help you to assess how well the model fits the data, and to interpret factors. Consequently, the ML method is often favored by statisticians.

In addition to these various methods of factor analysis, you can use SAS/IML Studio to compute various component analyses: principal component analysis, Harris component analysis, and image component analysis.

Example: Reduce Dimensionality through Common Factor Analysis

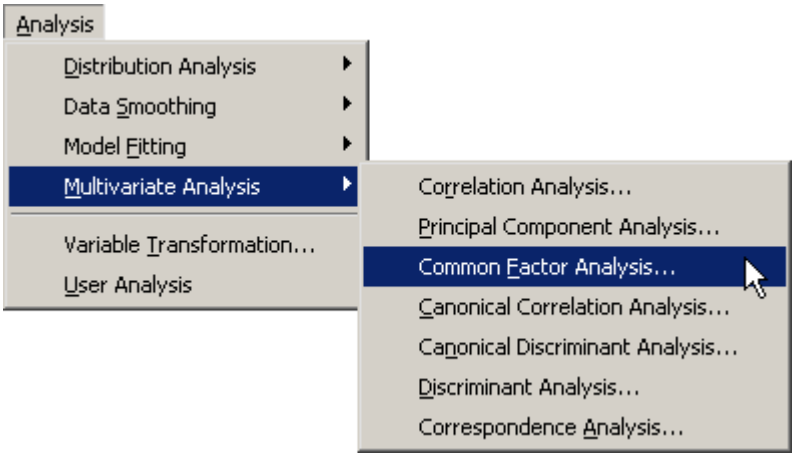
This example investigates factors that explain several variables in the Baseball data set. The Baseball data set contains performance measures for major league baseball players in 1986. A full description of the Baseball data is included in Chapter A, "Sample Data Sets."

Suppose you postulate the existence of latent variables that explain the hitting and fielding performance of players' performances during the 1986 season. (An example of a latent variable in the context of baseball is "quickness," which could explain correlation between a player's runs, stolen bases, and fielding statistics.) There are six variables that measure a player's batting performance: no_atbat, no_hits, no_home, no_runs, no_rbi, and no_bb. There are three variables that measure a player's fielding performance: no_outs, no_assts, and no_error.

To form a low-dimensional factor space that explains the relationships among these nine variables:

1. Open the Baseball data set.
2. Select **Analysis ►Multivariate Analysis ►Factor Analysis** from the main menu, as shown in Figure 27.2.

Figure 27.2 Selecting the Factor Analysis



The Factor Analysis dialog box appears. (See Figure 27.3.) You can select variables for the analysis by using the **Variables** tab.

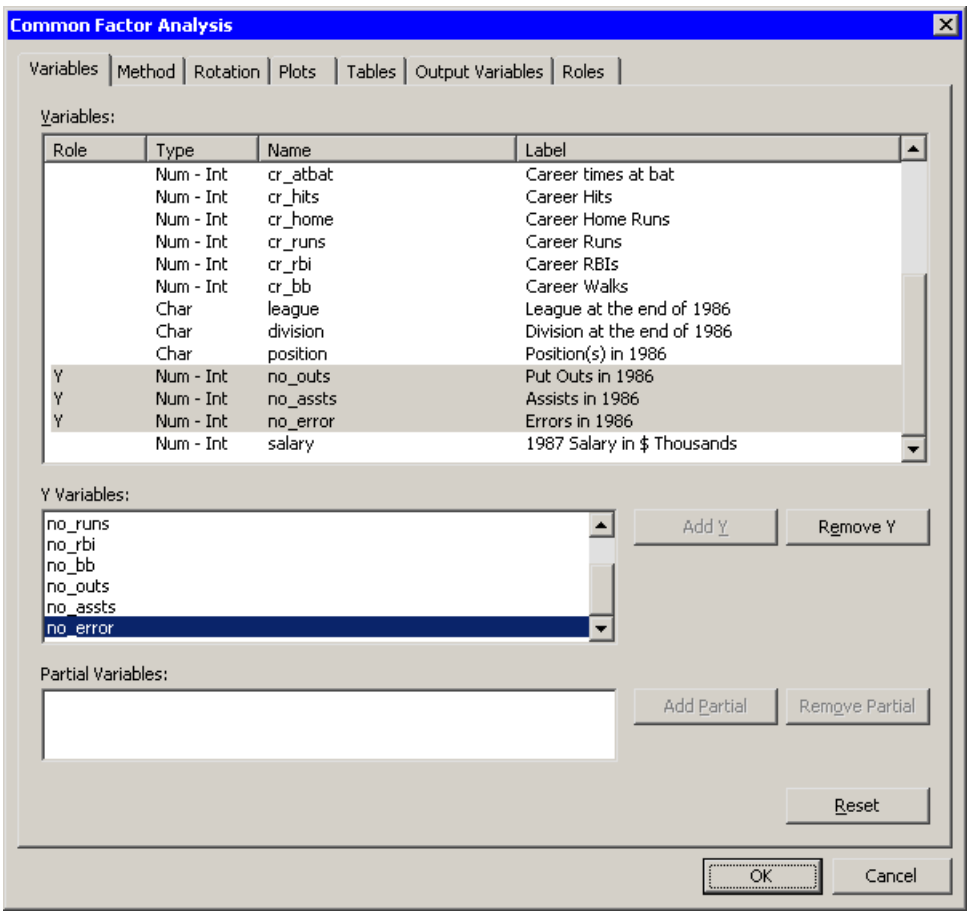
3 Select no_atbat. While holding down the CTRL key, select no_hits, no_home, no_runs, no_rbi, and no_bb. Click **Add Y**.

NOTE: Alternately, you can select the variables by using *contiguous selection*: click the first item, hold down the SHIFT key, and click the last item. All items between the first and last item are selected and can be added by clicking **Add Y**.

The three measures of fielding performance are located near the end of the list of variables.

4 Scroll to the end of the variable list. Select no_outs. While holding down the CTRL key, select no_assts and no_error. Click **Add Y**.

Figure 27.3 The Variables Tab



5 Click the **Method** tab.

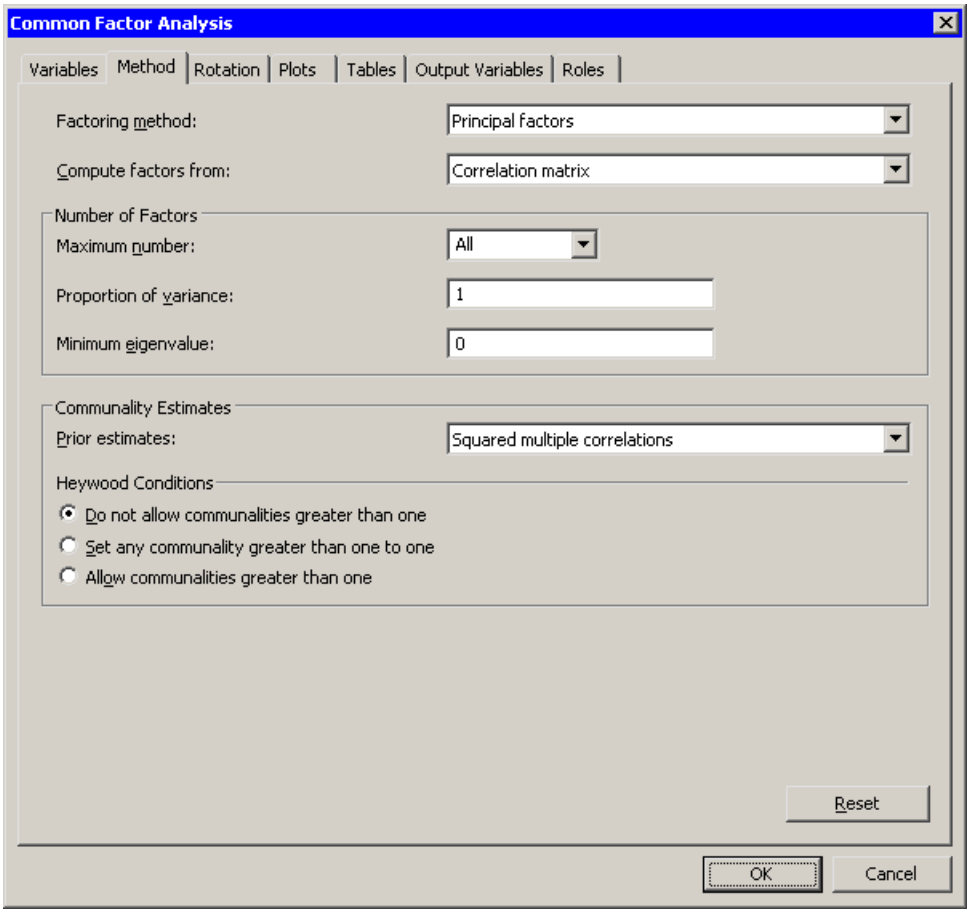
The **Method** tab becomes active. (See Figure 27.4.) You can use the **Method** tab to set options in the analysis.

The default method is principal factor analysis. However, the default method of estimating the prior communalities is to set all prior communalities to 1. This would result in a principal component analysis rather than a factor analysis.

6 Set **Prior estimates** to **Squared multiple correlations**.

The preceding step sets the prior communality estimate for each variable to its squared multiple correlation with all other variables.

Figure 27.4 The Method Tab

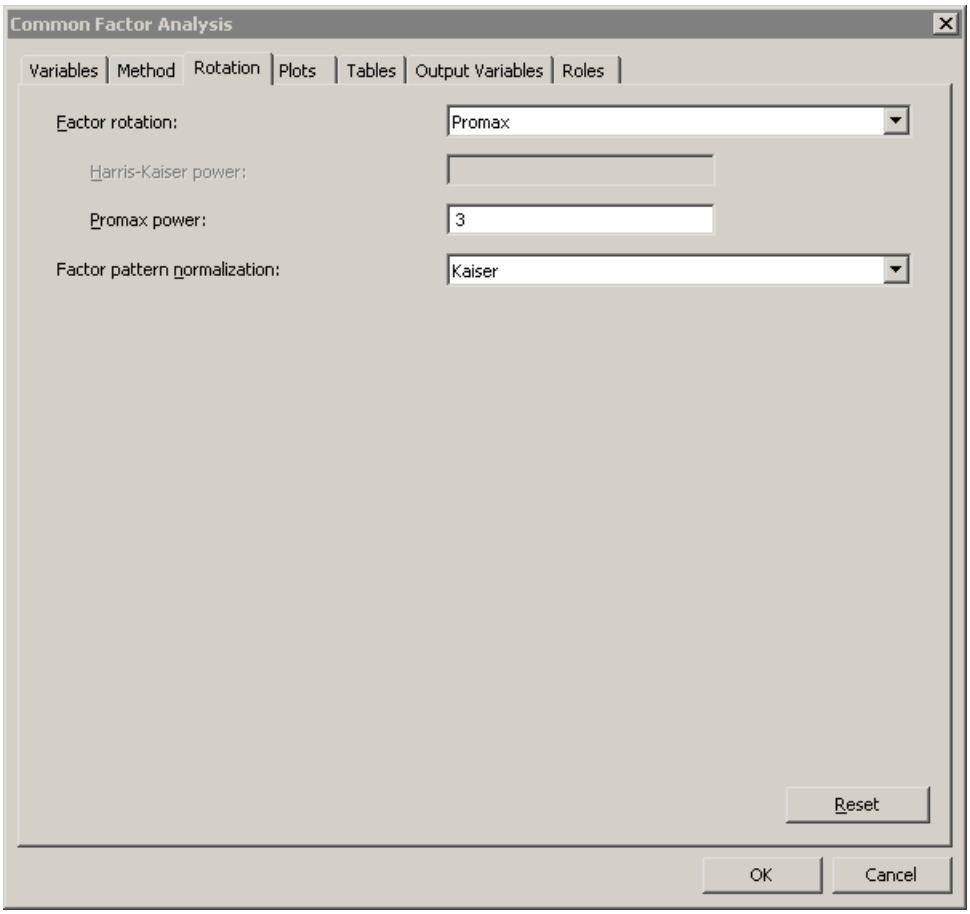


7 Click the **Rotation** tab.

The **Rotation** tab becomes active. (See Figure 27.5.) The default behavior is to leave factors unrotated. This example requests that an oblique transformation be applied to the factors in order to illustrate how rotated factors can sometimes be more interpretable.

8 Select **Promax** for the **Factor rotation** option.

Figure 27.5 The Rotation Tab



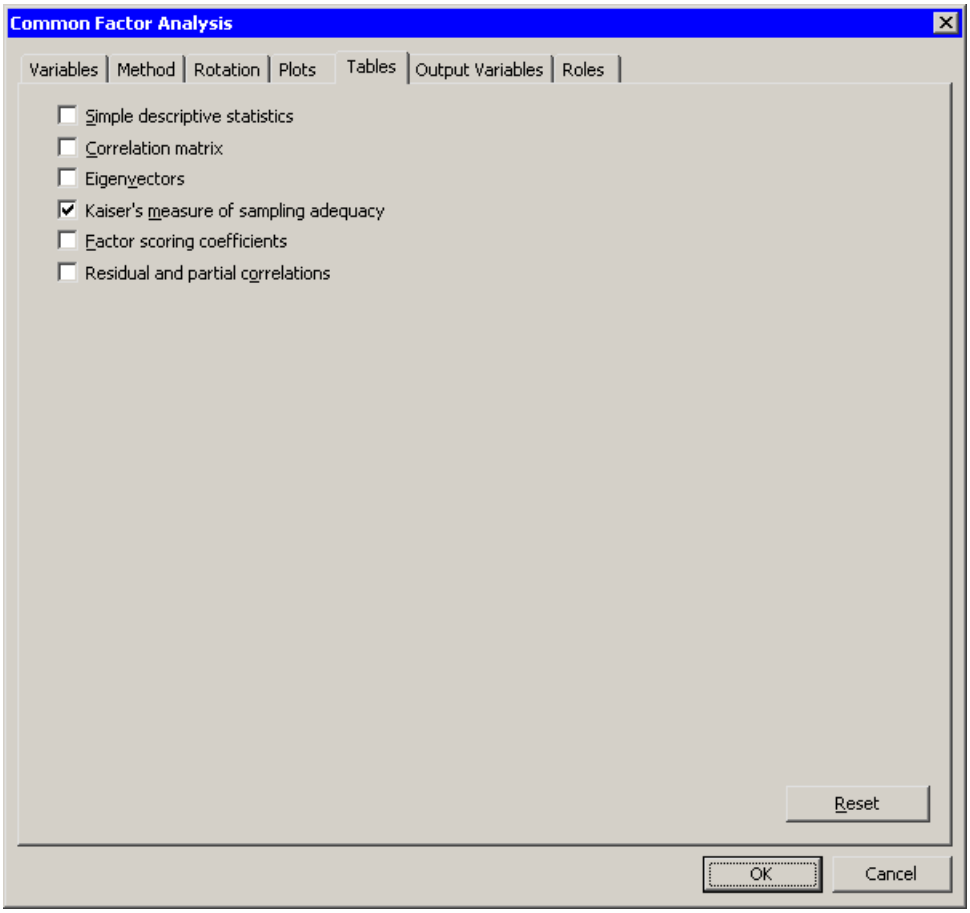
9 Click the **Tables** tab.

The **Tables** tab becomes active. (See Figure 27.6.) To help determine whether the data are appropriate for the common factor model, you can request Kaiser's measure of sampling adequacy (MSA).

10 Select **Kaiser's measure of sampling adequacy**.

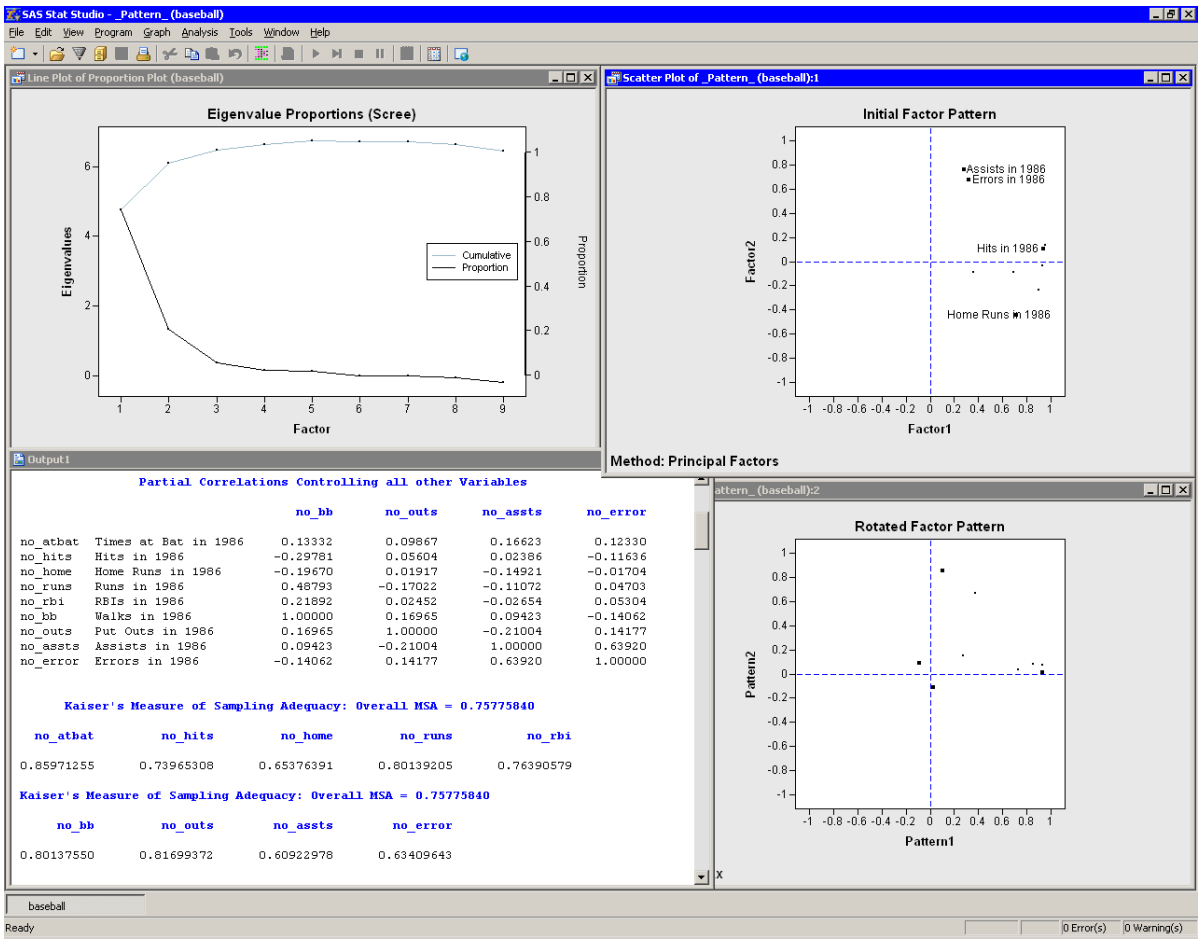
11 Click **OK**.

Figure 27.6 The Tables Tab



The analysis calls the FACTOR procedure, which uses the options specified in the dialog box. The procedure displays tables in the output document, as shown in Figure 27.7. As is discussed subsequently, the Factor analysis extracts three principal factors for these data. Three plots also appear.

Figure 27.7 Output from a Factor Analysis



The eigenvalue plot shows the eigenvalues of the reduced correlation matrix, along with the cumulative proportion of common variance accounted for by the factors. The first two factors account for almost 95% of the common variance, and the first three factors account for 101%. The reduced correlation matrix for these data has negative eigenvalues, which explains why the factors that correspond to the largest eigenvalues account for more than 100% of the common variance.

The initial factor pattern plot shows the projection of the original variables onto the subspace spanned by the first two factors. As shown in Figure 27.7, you can click a point in order to identify the corresponding variable. The points with high values of Factor1 are all hitting variables, including no_hits. The points with the highest values of Factor2 are two of the fielding variables: no_assts and no_error. The third fielding variable (no_outs) is closest to the origin in this plot. The initial factor pattern plot indicates that the first (unrotated) factor correlates highly with the hitting variables, whereas the second correlates with assists and errors.

NOTE: If you want to visualize the third extracted factor, you can color the observations according to the value of the Factor3 variable or create a three-dimensional scatter plot of the three factors. You can view the data for this plot by pressing the F9 key when the plot is active.

The rotated factor pattern plot in Figure 27.7 shows the projection of the original variables onto the subspace spanned by the first two rotated factors. A promax transformation is used to transform the original factors (which are orthogonal to each other) to new factors that, in many cases, are easier to interpret in terms of the original variables. This transformation does not change the common factor space or the communality estimates.

In the rotated factor pattern plot, the cluster of points with high values of Pattern1 are the variables no_atbat, no_hits, no_runs, and no_bb. (These points are not labeled in Figure 27.7, but they are labeled in Figure 27.8.) Players with high values of these variables get on base often, so you might interpret the first (rotated) factor to be "Getting on Base." The two points with high values of Pattern2 are the variables no_home and no_rbi. Players who have high values of these variables contribute many runs to their teams' scores, so you might interpret the second (rotated) factor as "Scoring."

In the rotated factor pattern plot, the fielding variables are positioned near the origin, which indicates that these variables are not strongly correlated with the first two rotated factors. Figure 27.8 shows a three-dimensional scatter plot that visualizes the three rotated factors. The plot shows that no_assts and no_error are highly correlated with the third rotated factor, while no_outs is not strongly correlated with any of the first three factors. The third rotated factor identifies players who make many assists and many errors. These are typically infielders who play second base, shortstop, or third base. Consequently, you might interpret the third rotated factor as a "Fielding Position" factor.

Figure 27.8 Plot of Obliquely Transformed Factors

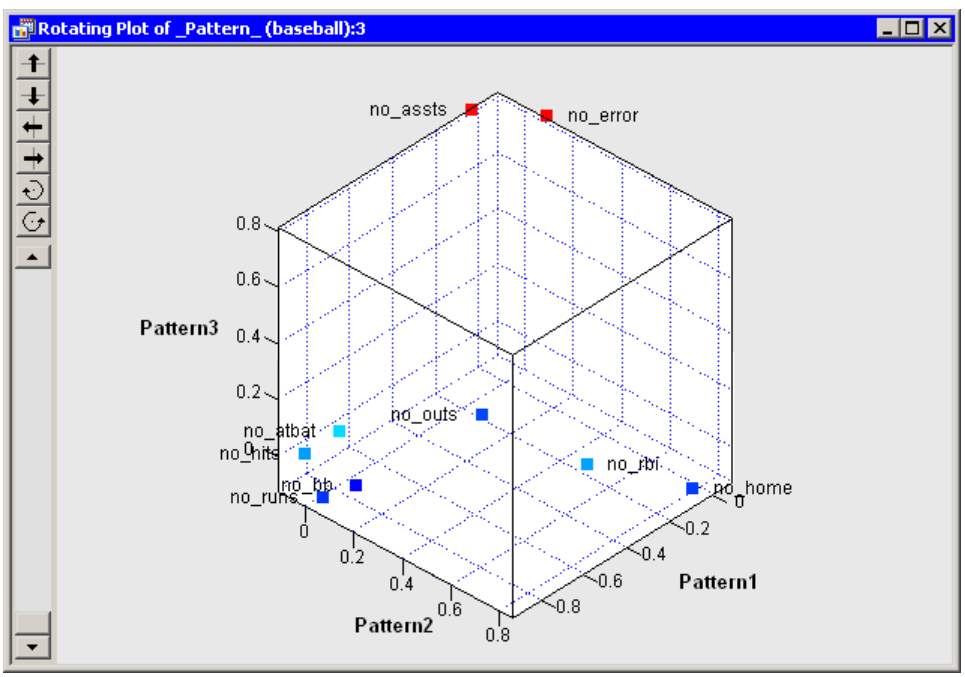


Figure 27.7 shows part of the partial correlations matrix for the original variables. If the data are appropriate for the common factor model, the partial correlations (controlling the other variables) should be small compared to the original correlations. Recall that the partial correlation between two variables, controlling for the variables $X_1, \ldots, X_k$, is the correlation between the residuals of the two variables after regression on the X_i.

Figure 27.7 also shows the MSA statistics. Kaiser's MSA (Kaiser 1970) is a summary, for each variable and for all variables together, of how much smaller the partial correlations are than the original correlations. Values of 0.8 or 0.9 are considered good, while MSAs less than 0.5 are unacceptable. The no_assts and no_error variables have the poorest MSAs. The overall MSA of 0.76 is adequate for proceeding with the factor analysis; an overall MSA lower than 0.6 often indicates that the data are not likely to factor well.

Figure 27.9 shows additional output. The prior communality estimates indicate that the variance of no_outs might not be well explained by the three common factors. The table of eigenvalues displays the eigenvalues for the reduced correlation matrix, which is the correlation matrix of the original variables, except that the 1's on the diagonal are replaced by the prior communality estimates. A note is printed below this table which indicates that three factors are retained because they account for (at least) 100% of the common variance.

Figure 27.9 Output from a Factor Analysis

Output1

Prior Communality Estimates: SMC

no_atbat	no_hits	no_home	no_runs	no_rbi
0.94485664	0.95394685	0.83396584	0.90490622	0.89771472

no_bb	no_outs	no_assts	no_error
0.57442507	0.19852778	0.59437919	0.53574570

Eigenvalues of the Reduced Correlation Matrix: Total = 6.438468 Average = 0.71538533

	Eigenvalue	Difference	Proportion	Cumulative
1	4.77064168	3.43453194	0.7410	0.7410
2	1.33610974	0.95839702	0.2075	0.9485
3	0.37771272	0.23341569	0.0587	1.0071
4	0.14429704	0.02721227	0.0224	1.0296
5	0.11708477	0.13530602	0.0182	1.0477
6	-.01822126	0.00328724	-0.0028	1.0449
7	-.02150850	0.05041550	-0.0033	1.0416
8	-.07192400	0.12380019	-0.0112	1.0304
9	-.19572419		-0.0304	1.0000

3 factors will be retained by the PROPORTION criterion.

Figure 27.10 shows additional output from the FACTOR procedure. The "Factor Pattern" table shows the relationship between the unrotated factors and the original Y variables. Each Y variable is a linear combinations of the common factor and a unique factor. For example, no_atbat corresponds to the linear combination

$$\text{no_atbat} = 0.95565\ \text{Factor1} + 0.13507\ \text{Factor2} - 0.12293\ \text{Factor3} + u_1$$

If you decide not to rotate the factors, you can attempt to interpret these factors by looking at the relative magnitudes of the coefficients. For example, the first unrotated factor appears to measure a player's overall performance. More weight is given to getting on base (coefficients in the range 0.89–0.96), less weight is given to scoring runs (coefficients in the range 0.68–0.72), and little weight is given to the fielding statistics. The figure also shows the common variance explained by each factor and the final communality estimates.

Figure 27.10 Unrotated Factors

Output1

Factor Pattern

		Factor1	Factor2	Factor3
no_atbat	Times at Bat in 1986	0.95565	0.13507	-0.12293
no_hits	Hits in 1986	0.94194	0.10661	-0.19118
no_home	Home Runs in 1986	0.71580	-0.44501	0.36945
no_runs	Runs in 1986	0.93298	-0.03777	-0.18521
no_rbi	RBIs in 1986	0.89648	-0.23324	0.24696
no_bb	Walks in 1986	0.68917	-0.08826	-0.17231
no_outs	Put Outs in 1986	0.36013	-0.09163	0.00035
no_assts	Assists in 1986	0.27909	0.76365	0.10686
no_error	Errors in 1986	0.31796	0.67327	0.23055

Variance Explained by Each Factor

Factor1	Factor2	Factor3
4.7706417	1.3361097	0.3777127

Final Communality Estimates: Total = 6.484464

no_atbat	no_hits	no_home	no_runs	no_rbi
0.94661815	0.93516499	0.84690232	0.90617263	0.91905937

no_bb	no_outs	no_assts	no_error
0.51244408	0.13808830	0.67247223	0.60754208

Whereas Figure 27.10 displays information about the unrotated factors, Figure 27.11 displays information about the rotated factors. The promax transformation is the composition of two transformations: an orthogonal varimax rotation and an oblique Procrustean transformation. Figure 27.11 displays information about the factors after the orthogonal varimax rotation. You can also visualize the pattern of the rotated factors as follows: view the data for a factor pattern plot by pressing the F9 key when the factor pattern plot is active, and then create scatter plots of the variables named Prerotat*n*. The Prerotat*n* variables correspond to the columns of the "Rotated Factor Pattern Table."

Figure 27.11 Orthogonally Rotated Factors

Output1

The FACTOR Procedure
Prerotation Method: Varimax

Orthogonal Transformation Matrix

	1	2	3
1	0.83182	0.49225	0.25645
2	-0.04219	-0.40462	0.91351
3	-0.55344	0.77069	0.31580

Rotated Factor Pattern

		Factor1	Factor2	Factor3
no_atbat	Times at Bat in 1986	0.85726	0.32103	0.32964
no_hits	Hits in 1986	0.88483	0.27320	0.27857
no_home	Home Runs in 1986	0.40972	0.81715	-0.10629
no_runs	Runs in 1986	0.88016	0.33181	0.14626
no_rbi	RBIs in 1986	0.61887	0.72600	0.09482
no_bb	Walks in 1986	0.67236	0.24216	0.04169
no_outs	Put Outs in 1986	0.30324	0.21462	0.00876
no_assts	Assists in 1986	0.14080	-0.08924	0.80292
no_error	Errors in 1986	0.10849	0.06178	0.76939

Variance Explained by Each Factor

Factor1	Factor2	Factor3
3.4189825	1.5990851	1.4663966

Final Communality Estimates: Total = 6.484464

no_atbat	no_hits	no_home	no_runs	no_rbi
0.94661815	0.93516499	0.84690232	0.90617263	0.91905937

no_bb	no_outs	no_assts	no_error
0.51244408	0.13808830	0.67247223	0.60754208

Figure 27.12 displays information about the obliquely transformed factors. The Procrustean transformation is displayed, followed by the matrix used to transform the unrotated factors into the factors displayed in the "Rotated Factor Pattern (Standardized Regression Coefficients)" table. The factor loadings shown in this table are shown graphically in the rotated factor pattern plot. (See Figure 27.7.) An oblique transformation introduces correlations between the factors, and the "Inter-Factor Correlations" table shows those correlations. You can convert the correlations into angles between the factors by applying the arccosine function. For example, the angle between the first and second factors is $\cos^{-1}(0.59222)$, or approximately 53.7 degrees, whereas the second and third factors are almost orthogonal.

The output contains additional tables (not shown) that display further correlations, structures, and variances. The "Displayed Output" section of the FACTOR procedure documentation describes all of the tables.

Figure 27.12 Obliquely Rotated Factors

Output1

Procrustean Transformation Matrix

	1	2	3
1	1.41484316	-0.4285994	-0.4055101
2	-0.608408	1.34374299	0.29671601
3	-0.278359	0.07344782	1.2818418

Normalized Oblique Transformation Matrix

	1	2	3
1	0.74595392	0.30286241	0.11653946
2	-0.0627432	-0.428906	0.90540075
3	-1.2399938	1.21228394	0.72728323

Inter-Factor Correlations

	Factor1	Factor2	Factor3
Factor1	1.00000	0.59222	0.41007
Factor2	0.59222	1.00000	-0.00910
Factor3	0.41007	-0.00910	1.00000

Rotated Factor Pattern (Standardized Regression Coefficients)

		Factor1	Factor2	Factor3
no_atbat	Times at Bat in 1986	0.85683	0.08247	0.14425
no_hits	Hits in 1986	0.93301	0.00779	0.06726
no_home	Home Runs in 1986	0.10375	0.85554	-0.05080
no_runs	Runs in 1986	0.92798	0.07424	-0.06017
no_rbi	RBIs in 1986	0.37713	0.67093	0.07291
no_bb	Walks in 1986	0.73330	0.03769	-0.12492
no_outs	Put Outs in 1986	0.27396	0.14879	-0.04074
no_assts	Assists in 1986	0.02777	-0.11346	0.80165
no_error	Errors in 1986	-0.09094	0.08702	0.81431

Specifying the Factor Analysis

This section describes the dialog box tabs that are associated with the Factor analysis. The Factor analysis calls the FACTOR procedure in SAS/STAT software. See the FACTOR procedure documentation in the *SAS/STAT User's Guide* for additional details.

Variables Tab

You can use the **Variables** tab to specify the numerical variables for the analysis. The **Variables** tab is shown in Figure 27.3.

The variables in the **Y Variables** list correspond to variables in the VAR statement of the FACTOR procedure.

The **Partial** list is rarely used. The variables in this list correspond to variables in the PARTIAL statement of the FACTOR procedure. The FACTOR procedure computes the factors for the residuals of the Y variables after regression on the PARTIAL variables. Equivalently, the factors are determined by the partial correlation matrix between the Y variables, controlling for the PARTIAL variables.

Method Tab

You can use the **Method** tab to set options in the analysis. (See Figure 27.4.) Each UI control in the tab corresponds to an option in the FACTOR procedure. The **Method** tab contains the following controls:

Factoring method
specifies the method used to extract factors or specifies a component analysis. This corresponds to the METHOD= option in the PROC FACTOR statement.

Compute factors from
specifies whether the factors are computed for the correlation matrix or the covariance matrix. This corresponds to the COV option in the PROC PRINCOMP statement. **NOTE:** Some methods require a correlation matrix.

Number of Factors
The number of factors retained is determined by the minimum number that satisfies the next three criteria.

Maximum number
specifies how many factors to compute. This corresponds to the N= option in the PROC FACTOR statement. You can type into the **Maximum number** field; if you want five factors, you can enter `5` even though this is not an option on the list.

Proportion of variance
specifies the proportion of common variance in the retained factors. This value is in the range $(0, 1]$. The option corresponds to the PROPORTION= option in the PROC FACTOR statement.

Minimum eigenvalue
specifies the smallest eigenvalue for which a factor is retained. This corresponds to the MINEIGEN= option in the PROC FACTOR statement.

Prior estimates
specifies a method for computing prior communality estimates. This corresponds to the PRIORS= option in the PROC FACTOR statement. The default method for the principal factor method is to set all priors equal to 1. This results in a principal *component* analysis. If you want a principal *factor* analysis, you should select a different method for estimating the prior communalities, as illustrated in the section "Example: Reduce Dimensionality through Common Factor Analysis" on page 423.

Heywood Conditions

specifies how the factor analysis behaves if a communality is greater than 1. The section "Heywood Cases and Other Anomalies about Communality Estimates" in the documentation for the FACTOR procedure describes why this situation might occur.

Do not allow communalities greater than one

specifies that an analysis should stop processing if it encounters a communality greater than one.

Set any communality greater than one to one

specifies that an analysis should set any communality greater than one to one, and then continue. This corresponds to the HEYWOOD option in the PROC FACTOR statement.

Allow communalities greater than one

specifies that an analysis should allow any communality. This corresponds to the ULTRAHEYWOOD option in the PROC FACTOR statement.

Rotation Tab

You can use the **Rotation** tab to transform the factors by orthogonal or oblique rotations. (See Figure 27.5.) Orthogonal rotations rigidly rotate the factors; oblique transformations introduce correlations between the factors. Transformed factors are often more interpretable in terms of the original variables.

The **Rotation** tab contains the following UI controls:

Factor rotation

specifies the rotation method. You can select from a set of common orthogonal or oblique transformations. This corresponds to the ROTATE= option in the PROC FACTOR statement.

Harris-Kaiser power

specifies the power of the square roots of the eigenvalues used to rescale the eigenvectors for Harris-Kaiser orthoblique transformation. This corresponds to the HKPOWER= option in the PROC FACTOR statement.

Promax power

specifies the power for forming the target Procrustean matrix. This corresponds to the POWER= option in the PROC FACTOR statement.

Factor pattern normalization

specifies the method for normalizing the rows of the factor pattern for rotation. This corresponds to the NORM= option in the PROC FACTOR statement.

Plots Tab

You can use the **Plots** tab to create plots that display results of the analysis. (See Figure 27.13.)

The plots for the Factor analysis are not linked to the original data table. The scree plot has its own data table; the two factor pattern plots (also called *factor loading plots*) are linked to each other. You can view the data for a plot by pressing the F9 key when the plot is active.

The following plots are available:

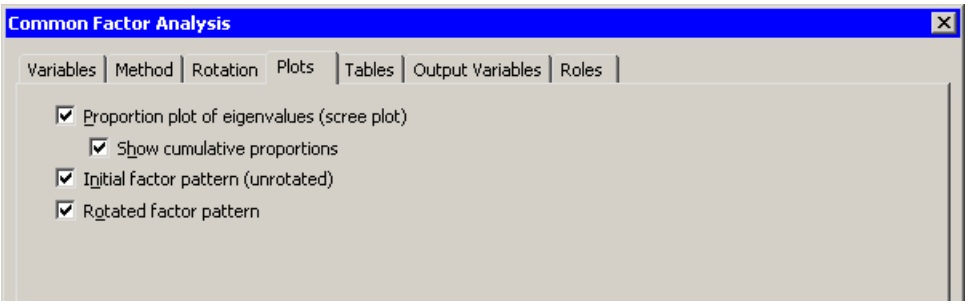
Proportion plot of eigenvalues (scree plot)
creates a plot that summarizes the eigenvalues of the reduced correlation or reduced covariance matrix.

Show cumulative proportions
adds cumulative proportions of eigenvalues to the proportion plot.

Initial factor pattern (unrotated)
creates a plot that shows the relationships between the initial (unrotated) factors and the original variables.

Rotated factor pattern
creates a plot that shows the relationships between the final rotated factors and the original variables. This plot is created only if you specify a rotation on the **Rotation** tab.

Figure 27.13 The Plots Tab



Tables Tab

The **Tables** tab is shown in Figure 27.6. You can use the **Tables** tab to display the following tables that summarize the results of the analysis:

Simple descriptive statistics
specifies whether to display the mean and standard deviation for each variable. This corresponds to the SIMPLE option in the PROC FACTOR statement.

Correlation matrix
specifies whether to display the correlation matrix. This corresponds to the CORR option in the PROC FACTOR statement.

Eigenvectors
specifies whether to display the eigenvectors of the reduced correlation matrix. This corresponds to the EIGENVECTORS option in the PROC FACTOR statement.

Kaiser's measure of sampling adequacy
specifies whether to display partial correlations between each pair of variables (controlling for all other variables), and Kaiser's measure of sampling adequacy. This corresponds to the MSA option in the PROC FACTOR statement.

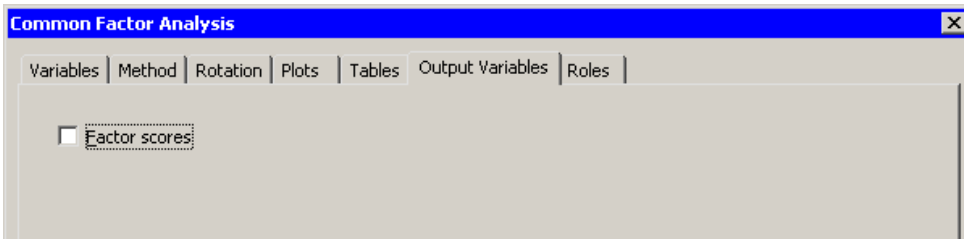
Factor scoring coefficients
specifies whether to display the factor scoring coefficients. This corresponds to the SCORE option in the PROC FACTOR statement.

Residual and partial correlations
specifies whether to display the residual correlation matrix and the associated partial correlation matrix. This corresponds to the RESIDUALS option in the PROC FACTOR statement.

Output Variables Tab

You can use the **Output Variables** tab to add estimated factor scores to the data table. (See Figure 27.14.) Each estimated factor score is computed as a linear combination of the standardized values of the variables that are factored. The names of the variables are of the form FACi, where $i = 1 \ldots k$, and k is the number of retained factors.

Figure 27.14 The Output Variables Tab



Roles Tab

You can use the **Roles** tab to specify a frequency variable or weight variable for the analysis.

A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents n observations, where n is the value of the frequency variable.

A weight variable is a numeric variable with values that are relative weights for the analysis.

Analysis of Selected Variables

If any numeric variables are selected in a data table when you run the analysis, these variables are automatically entered in the **Y Variables** field of the **Variables** tab.

Any variable in the data table with a Frequency or Weight role is automatically entered in the appropriate field of the **Roles** tab.

References

Jackson, J. E. (1981), "Principal Components and Factor Analysis, Part 3: What Is Factor Analysis?" *Journal of Quality Technology*, 13, 125–130.

Jackson, J. E. (1991), *A User's Guide to Principal Components*, New York: John Wiley & Sons.

Kaiser, H. F. (1970), "A Second Generation Little Jiffy," *Psychometrika*, 35, 401–415.

Wickens, T. D. (1995), *The Geometry of Multivariate Statistics*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Chapter 28
Multivariate Analysis: Canonical Correlation Analysis

Contents

Overview of Canonical Correlation Analysis **441**
Example: Analyze the Relationship between Two Groups of Variables **442**
Specifying the Canonical Correlation Analysis **448**
Variables Tab 448
Method Tab 448
Plots Tab 449
Tables Tab 449
Output Variables Tab 450
Roles Tab 450
Analysis of Selected Variables **450**

Overview of Canonical Correlation Analysis

Canonical correlation analysis is a technique for analyzing the relationship between two sets (or groups) of variables. Each set can contain multiple variables.

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 first canonical correlation is at least as large as the multiple correlation between any variable and the opposite set of variables. 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.

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. However, the canonical variables do not represent jointly perpendicular directions through the space of the original variables.

You can run the Canonical Correlation analysis by selecting **Analysis ►Multivariate Analysis ►Canonical Correlation Analysis** from the main menu. The analysis is implemented by calling the CANCORR procedure

in SAS/STAT software. See the CANCORR procedure documentation in the *SAS/STAT User's Guide* for additional details.

Example: Analyze the Relationship between Two Groups of Variables

In this example, you examine canonical correlations between sets of variables in the GPA data set. The GPA data set contains average high school grades in mathematics, science, and English for students that applied to a university computer science program. The data also contains the students' scores on the mathematics and verbal sections of the SAT, which is a standardized test to measure aptitude.

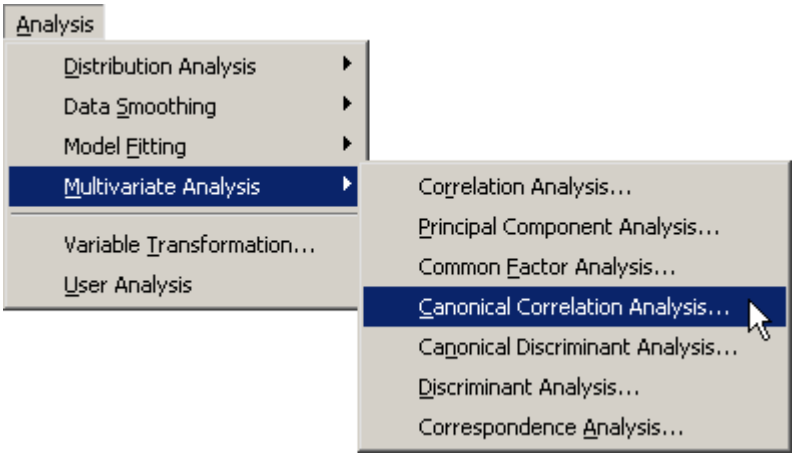
Suppose you are interested in the relationship between the variables that represent analytical thinking and those that represent verbal thinking. You can group the following variables into the analytical set: hsm (high school math average), hss (high school science average), and satm (SAT math score). You can group the following variables into the verbal set: hse (high school English average) and satv (SAT verbal score).

To run a canonical correlation analysis:

1 Open the GPA data set.

2 Select **Analysis ►Multivariate Analysis ►Canonical Correlation Analysis** from the main menu, as shown in Figure 28.1.

Figure 28.1 Selecting the Canonical Correlation Analysis

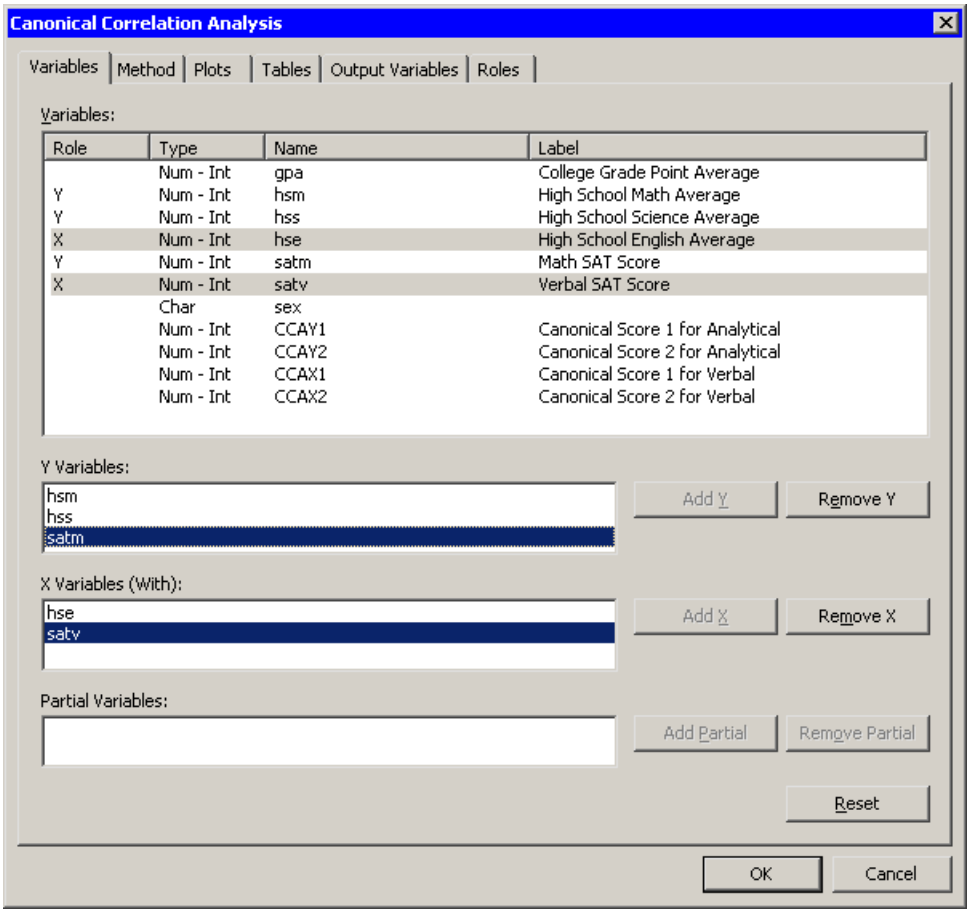


The Canonical Correlation Analysis dialog box appears. (See Figure 28.2.) You can select variables for the analysis by using the **Variables** tab.

3 Select hsm. While holding down the CTRL key, select hss and satm. Click **Add Y**.

4 Select hse. While holding down the CTRL key, select satv. Click **Add X**.

Figure 28.2 The Variables Tab



5 Click the **Tables** tab.

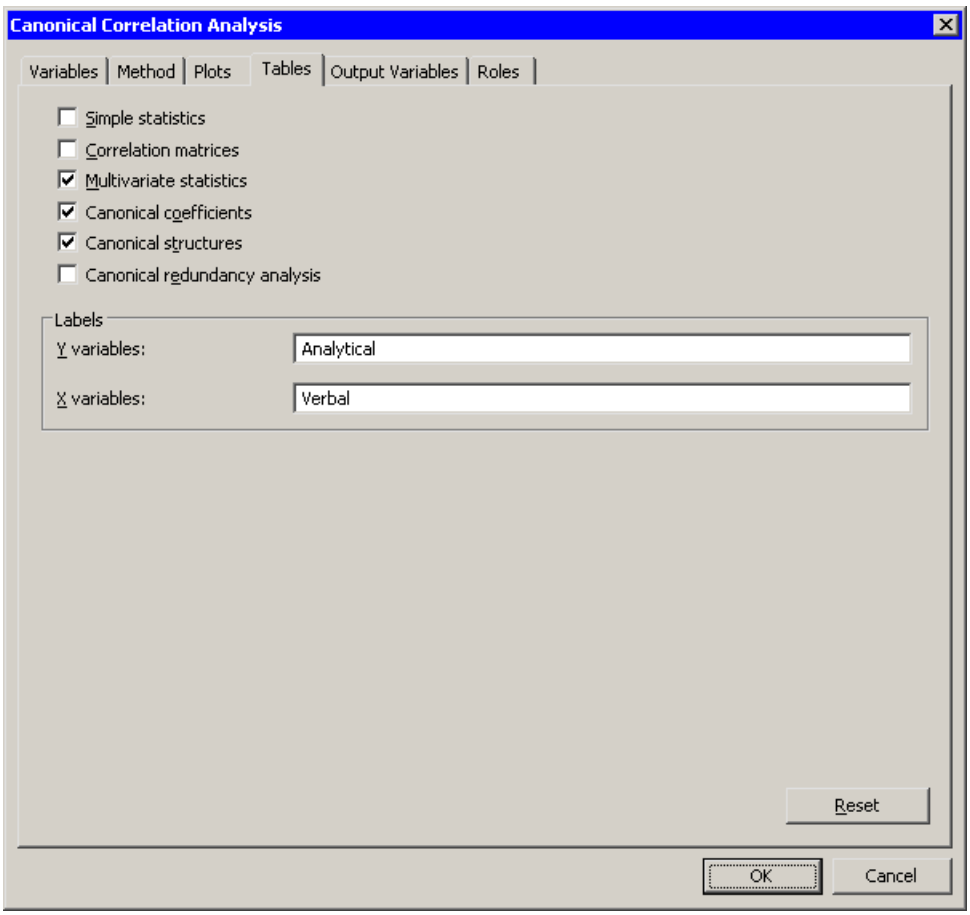
The **Tables** tab becomes active. (See Figure 28.3.) You can use the **Tables** tab to display statistics that are associated with the analysis, and to specify labels that identify the two sets of variables.

For this example, you can label the first set of variables as the "Analytical" set and the second set as the "Verbal" set.

6 Type `Analytical` into the **Y variables** field.

7 Type `Verbal` into the **X variables** field.

Figure 28.3 The Tables Tab



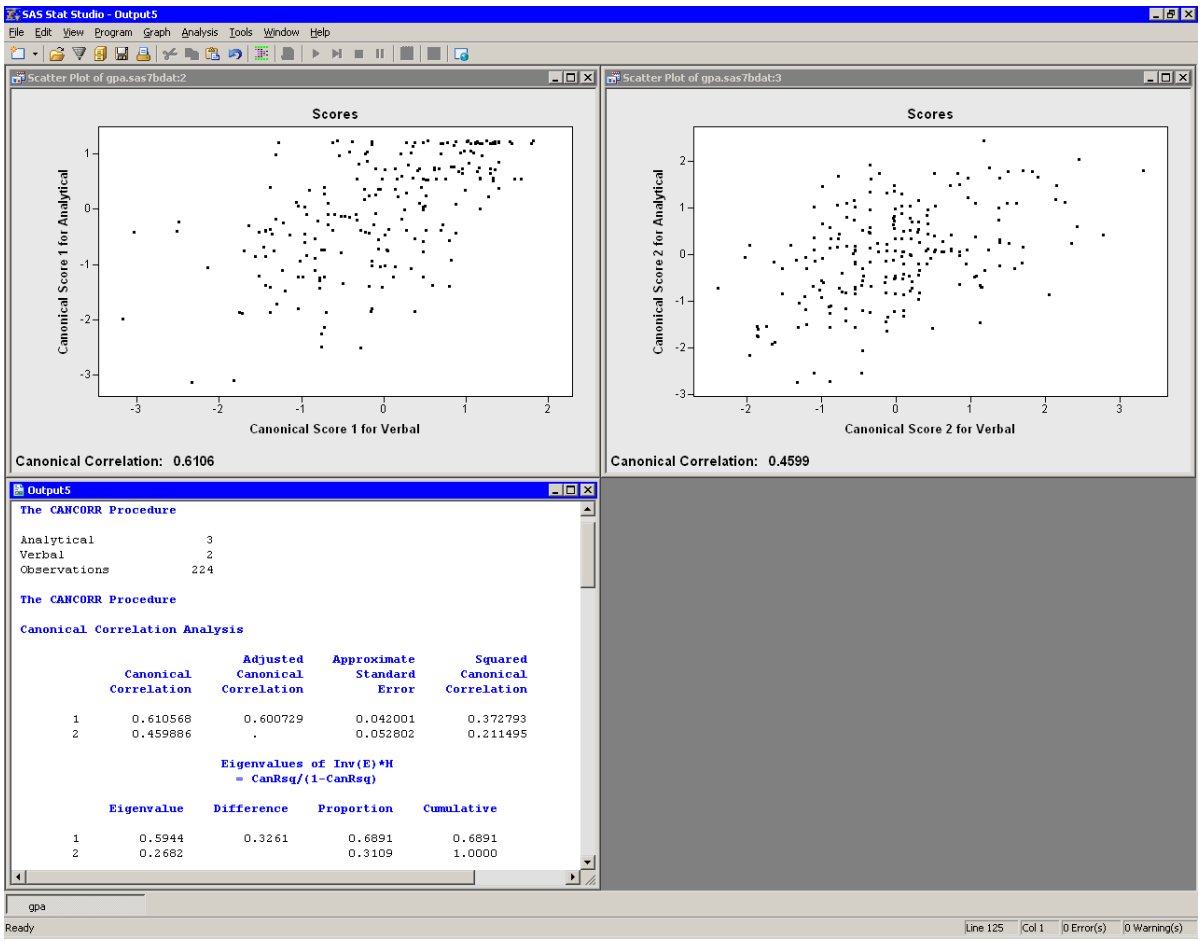
8 Click **OK**.

The analysis calls the CANCORR procedure, which uses the options specified in the dialog box. The procedure displays tables in the output document, as shown in Figure 28.4. Two plots are also created.

The plot of the first canonical variables shows the strength of the relationship between the set of analytical variables and the set of verbal variables. The second plot shows the second canonical variables. The footnote of these plots displays the canonical correlations. Note that the correlation between the second pair of canonical variables is less than the correlation between the first pair.

The output window in Figure 28.4 displays the canonical correlation, which is the correlation between the first pair of canonical variables. The value 0.6106 represents the highest possible correlation between any linear combination of the analytical variables and any linear combination of the verbal variables.

Figure 28.4 Output from a Canonical Correlation Analysis



The output window contains additional tables, as shown in Figure 28.5. The figure displays the likelihood ratios and associated statistics for testing the hypothesis that the canonical correlations in the current row and all that follow are zero. The first approximate F value of 30.80 corresponds to the test that all canonical correlations are zero. Since the p-value is small, you can reject the null hypothesis at the 95% level. Similarly, the second approximate F value of 29.50 corresponds to the test that the second canonical correlation is zero. This test also rejects the hypothesis.

Several multivariate statistics and F test approximations are also provided. These statistics test the null hypothesis that all canonical correlations are zero. The small p-values for these tests (< 0.0001) are evidence for rejecting the null hypothesis.

Figure 28.5 Testing Whether Canonical Correlations Are Zero

Output5

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.49455557	30.80	6	438	<.0001
2	0.78850512	29.50	2	220	<.0001

Multivariate Statistics and F Approximations

S=2 M=0 N=108.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.49455557	30.80	6	438	<.0001
Pillai's Trace	0.58428834	30.27	6	440	<.0001
Hotelling-Lawley Trace	0.86259372	31.41	6	290.23	<.0001
Roy's Greatest Root	0.59437114	43.59	3	220	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

The analysis creates canonical variables and adds them to the data table. The canonical variables for the analytical group are named CCAY1 and CCAY2. The canonical variables for the verbal group are named CCAX1 and CCAX2. The canonical variables are linear combinations of the original variables, so you can sometimes interpret the meaning of the canonical variables in terms of the original variables.

To interpret the variables, inspect the standardized coefficients of the canonical variables and the correlations between the canonical variables and their original variables. These statistics are shown in Figure 28.6. For example, the first canonical variables are represented by

$$\begin{aligned} \text{CCAY1} &= 0.0249\,\text{satm} + 0.8009\,\text{hss} + 0.2836\,\text{hsm} \\ \text{CCAX1} &= 0.9129\,\text{hse} + 0.2424\,\text{satv} \end{aligned}$$

The standardized canonical coefficients show that the first canonical variable for the analytical group is a weighted sum of the variables hss (with coefficient 0.8009) and hsm (0.2836), with the emphasis on the science grade. The coefficient for the variable satm is close to zero. The second canonical variable for the analytical group is a contrast between the variables satm (1.1208) and hsm (–0.4752), with more weight given to the SAT math score.

The coefficients for the verbal variables show that hse contributes heavily to the CCAX1 canonical variable (0.9129), whereas CCAX2 is heavily influenced by satv (1.0022).

Figure 28.6 Canonical Coefficients

Output1

Standardized Canonical Coefficients for the Analytical

		CCAY1	CCAY2
hsm	High School Math Average	0.2836	-0.4752
hss	High School Science Average	0.8009	-0.0295
satm	Math SAT Score	0.0249	1.1208

Standardized Canonical Coefficients for the Verbal

		CCAX1	CCAX2
hse	High School English Average	0.9129	-0.4793
satv	Verbal SAT Score	0.2424	1.0022

Figure 28.7 displays the table of correlations between the canonical variables and the original variables. These univariate correlations must be interpreted with caution, since they do not indicate how the original variables contribute jointly to the canonical analysis. However, they are often useful in the interpretation of the canonical variables.

The first canonical variable for the analytical group is strongly correlated with hsm and hss, with correlations 0.7560 and 0.9702, respectively. The second canonical variable for the analytical group is strongly correlated with satm, with a correlation of 0.8982.

The first canonical variable for the verbal group is strongly correlated with hse, with a correlation of 0.9720. The second canonical variable for the verbal group is strongly correlated with satv, with a correlation of 0.8854.

Figure 28.7 Correlations between Canonical and Original Variables

Output1

Correlations Between the Analytical and Their Canonical Variables

		CCAY1	CCAY2
hsm	High School Math Average	0.7560	0.0161
hss	High School Science Average	0.9702	-0.0336
satm	Math SAT Score	0.3461	0.8982

Correlations Between the Verbal and Their Canonical Variables

		CCAX1	CCAX2
hse	High School English Average	0.9720	-0.2351
satv	Verbal SAT Score	0.4649	0.8854

Correlations Between the Analytical and the Canonical Variables of the Verbal

		CCAX1	CCAX2
hsm	High School Math Average	0.4616	0.0074
hss	High School Science Average	0.5923	-0.0154
satm	Math SAT Score	0.2113	0.4131

Correlations Between the Verbal and the Canonical Variables of the Analytical

		CCAY1	CCAY2
hse	High School English Average	0.5935	-0.1081
satv	Verbal SAT Score	0.2838	0.4072

In summary, the analytical and verbal variables are moderately correlated with each other, with a canonical correlation of 0.6106. The first canonical variables are close to the linear subspace spanned by the variables that measure a student's high school grades. The second canonical variables are close to the linear subspace spanned by the SAT variables. (Recall that the *span* of a set of vectors is the vector space consisting of all linear combinations of the vectors.)

Specifying the Canonical Correlation Analysis

This section describes the dialog box tabs that are associated with the Canonical Correlation analysis. The Canonical Correlation analysis calls the CANCORR procedure in SAS/STAT software. See the CANCORR procedure documentation in the *SAS/STAT User's Guide* for additional details.

Variables Tab

You can use the **Variables** tab to specify the numerical variables for the analysis. The **Variables** tab is shown in Figure 28.2.

The variables in the **Y Variables** list correspond to variables in the VAR statement of the CANCORR procedure. The variables in the **X Variables (With)** list correspond to variables in the WITH statement of the CANCORR procedure.

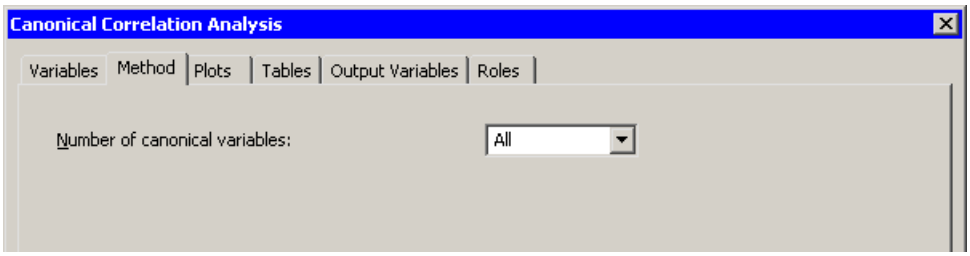
The **Partial** list is rarely used. The variables in this list correspond to variables in the PARTIAL statement of the CANCORR procedure. The CANCORR procedure computes the canonical correlations of the residuals from the prediction of the VAR and WITH variables by the PARTIAL variables.

Method Tab

You can use the **Method** tab to set options in the analysis. (See Figure 28.8.)

You can use the **Number of canonical variables** option to specify the number of canonical variables displayed in the output window. This option corresponds to the NCAN= option in the PROC CANCORR statement.

Figure 28.8 The Method Tab



Plots Tab

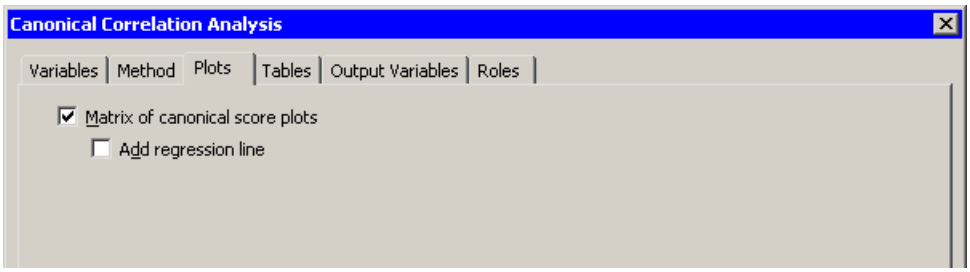
You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 28.9.)

Creating a plot adds canonical variables to the data table. The following plots are available:

Matrix of canonical score plots
creates a plot for each pair of canonical variables that summarizes the strength of the relationship between the variables.

Add regression line
adds a least squares regression line to each score plot. The regression line predicts the ith canonical variable in the second group from the ith canonical variable in the first group.

Figure 28.9 The Plots Tab



Tables Tab

The **Tables** tab is shown in Figure 28.3. You can use the **Tables** tab to display the following tables that summarize the results of the analysis:

Simple statistics
specifies whether to display the mean and standard deviation for each variable. This option corresponds to the SIMPLE option in the PROC CANCORR statement.

Correlation matrices
specifies whether to display the correlation matrices for each set of variables. This option corresponds to the CORR option in the PROC CANCORR statement.

Multivariate statistics
specifies whether to display a table of multivariate statistics and F approximations.

Canonical coefficients
specifies whether to display the raw and standardized canonical coefficients for each set of variables.

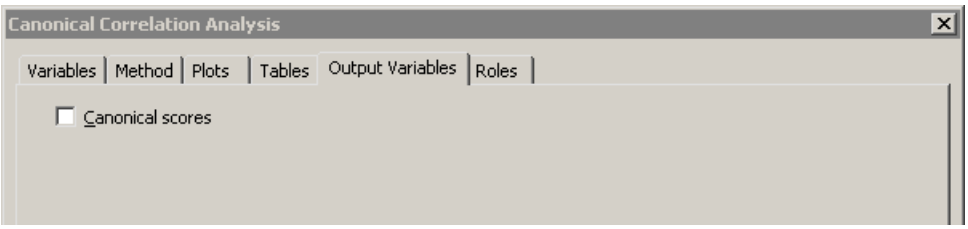
Canonical structures
specifies whether to display correlations between the canonical variables and the original variables.

Canonical redundancy analysis
specifies whether to display a canonical redundancy analysis. This option corresponds to the REDUNDANCY option in the PROC CANCORR statement.

Output Variables Tab

You can use the **Output Variables** tab to add canonical variables (also called canonical scores) to the data table. (See Figure 28.10.) The option on the **Method** tab determines how many variables are added.

Figure 28.10 The Output Variables Tab



Roles Tab

You can use the **Roles** tab to specify a frequency variable or weight variable for the analysis.

A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents n observations, where n is the value of the frequency variable.

A weight variable is a numeric variable with values that are relative weights for the analysis.

Analysis of Selected Variables

If any numeric variables are selected in a data table when you run the analysis, these variables are automatically entered in the **Y Variables** field of the **Variables** tab.

Any variable in the data table with a Frequency or Weight role is automatically entered in the appropriate field of the **Roles** tab.

Chapter 29
Multivariate Analysis: Canonical Discriminant Analysis

Contents

Overview of Canonical Discriminant Analysis **451**
Example: Construct Linear Subspaces that Discriminate between Categories **452**
Specifying the Canonical Discriminant Analysis **461**
Variables Tab 461
Method Tab 461
The Plots Tab 462
Tables Tab 464
Output Variables Tab 466
Roles Tab 466
Analysis of Selected Variables **467**

Overview of Canonical Discriminant Analysis

Canonical discriminant analysis is a dimension-reduction technique that is related to principal component analysis and canonical correlation. Given a nominal 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.

Canonical discriminant analysis is equivalent to canonical correlation analysis between the quantitative variables and a set of dummy variables coded from the classification variable.

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 maximum multiple correlation is called the first canonical correlation. The coefficients of the linear combination are the *canonical coefficients*. The variable defined by the linear combination is the first canonical variable. 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. Canonical variables are also called *canonical components*.

You can run the Canonical Discriminant analysis by selecting **Analysis ►Multivariate Analysis ►Canonical Discriminant Analysis** from the main menu. The analysis is implemented by calling the DISCRIM procedure with the CANONICAL option in SAS/STAT software. See the documentation for the DISCRIM and CANDISC procedures in the *SAS/STAT User's Guide* for additional details.

The analysis calls the DISCRIM procedure (rather than the CANDISC procedure) because the DISCRIM procedure produces a discriminant function that can be used to classify current or future observations.

Example: Construct Linear Subspaces that Discriminate between Categories

In this example, you examine measurements of 159 fish caught in Finland's Lake Laengelmavesi. The fish are one of seven species: bream, parkki, perch, pike, roach, smelt, and whitefish. Associated with each fish are physical measurements of weight, length, height, and width. A full description of the Fish data is included in Chapter A, "Sample Data Sets."

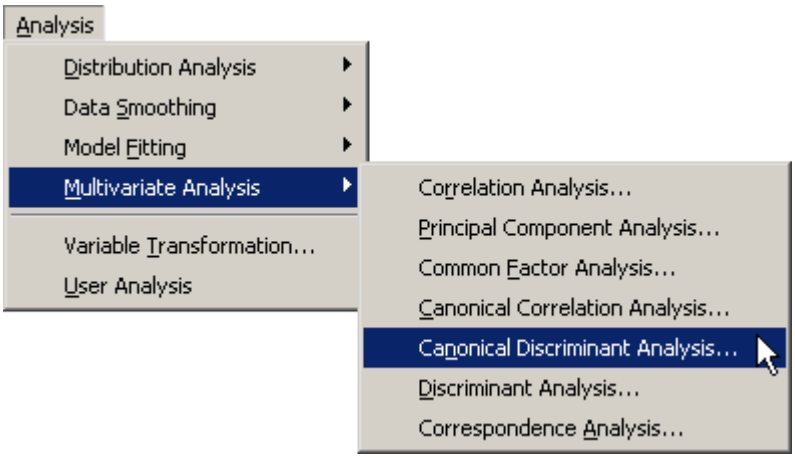
The goal of this example is to use canonical discriminant analysis to construct linear combinations of the size and weight variables that best discriminate between the species. By looking at the coefficients of the linear combinations, you can determine which physical measurements are most important in discriminating between groups. You can also determine whether there are two or more groups that cannot be discriminated using these measurements.

To run a canonical discriminant analysis:

1 Open the Fish data set.

2 Select **Analysis ►Multivariate Analysis ►Canonical Discriminant Analysis** from the main menu, as shown in Figure 29.1.

Figure 29.1 Selecting the Canonical Discriminant Analysis



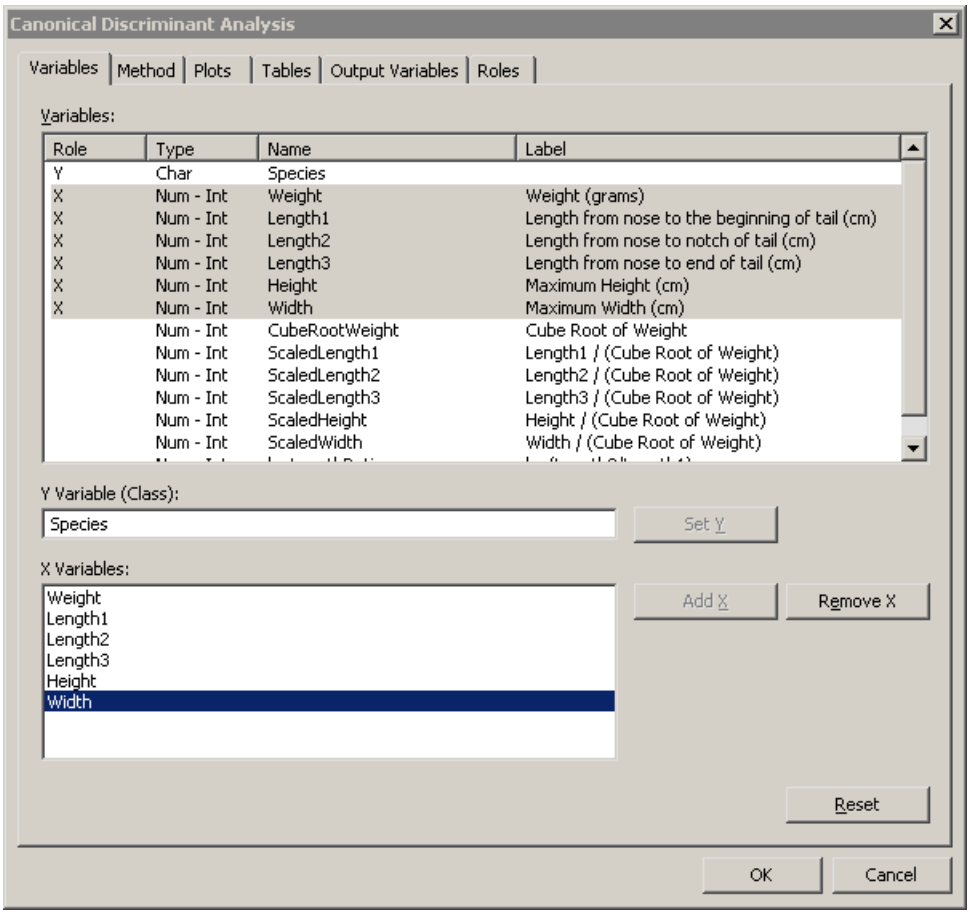
The Canonical Discriminant Analysis dialog box appears. (See Figure 29.2.) You can select variables for the analysis by using the **Variables** tab.

3 Select Species and click **Set Y**.

4 Select Weight. While holding down the CTRL key, select Length1, Length2, Length3, Height, and Width. Click **Add X**.

NOTE: Alternately, you can select the variables by using *contiguous selection*: click the first variable (Weight), hold down the SHIFT key, and click the last variable (Width). All variables between the first and last item are selected and can be added by clicking **Add X**.

Figure 29.2 The Variables Tab



5 Click the **Method** tab.

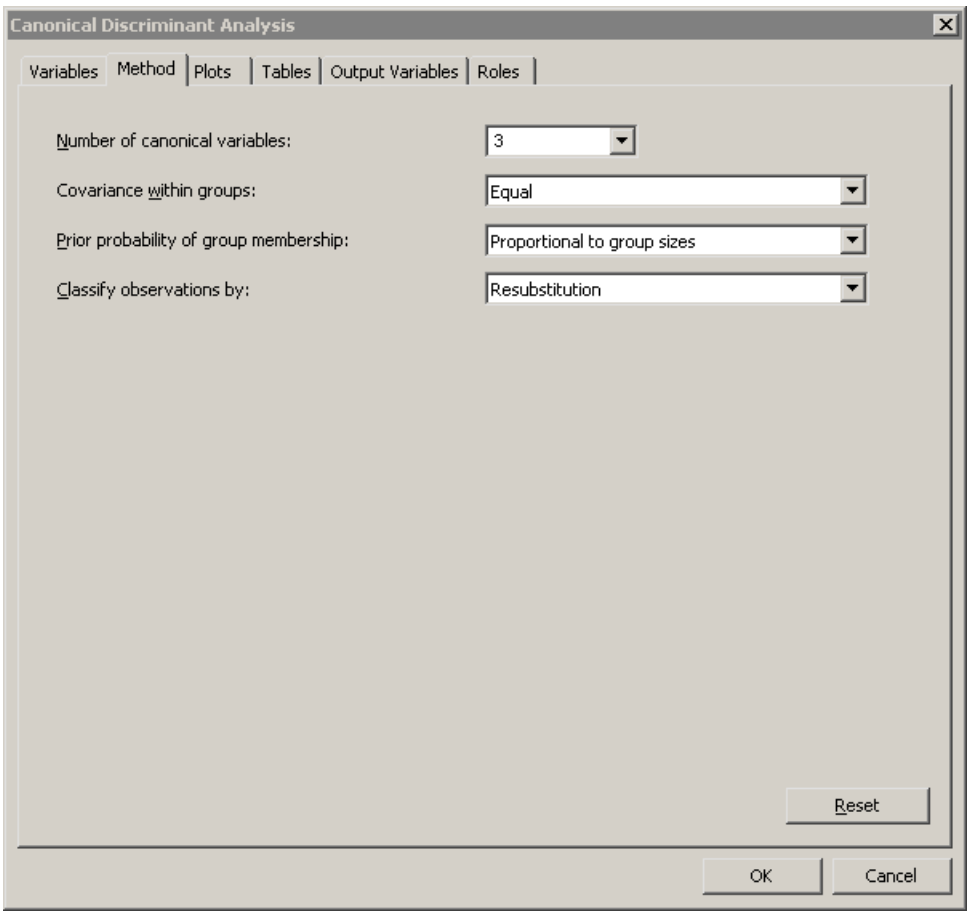
The **Method** tab becomes active. (See Figure 29.3.) You can use the **Method** tab to set options in the analysis.

6 Select **3** for **Number of canonical variables**.

The number of fish in any lake varies by species. That is, there is no reason to suspect that the number of whitefish in the lake is the same as the number of perch or bream. In the absence of prior knowledge about the distribution of fish species, you can assume that the number of fish of each species in the lake is proportional to the number in the sample.

7 Select **Proportional to group sizes** for **Prior probability of group membership**.

Figure 29.3 The Method Tab



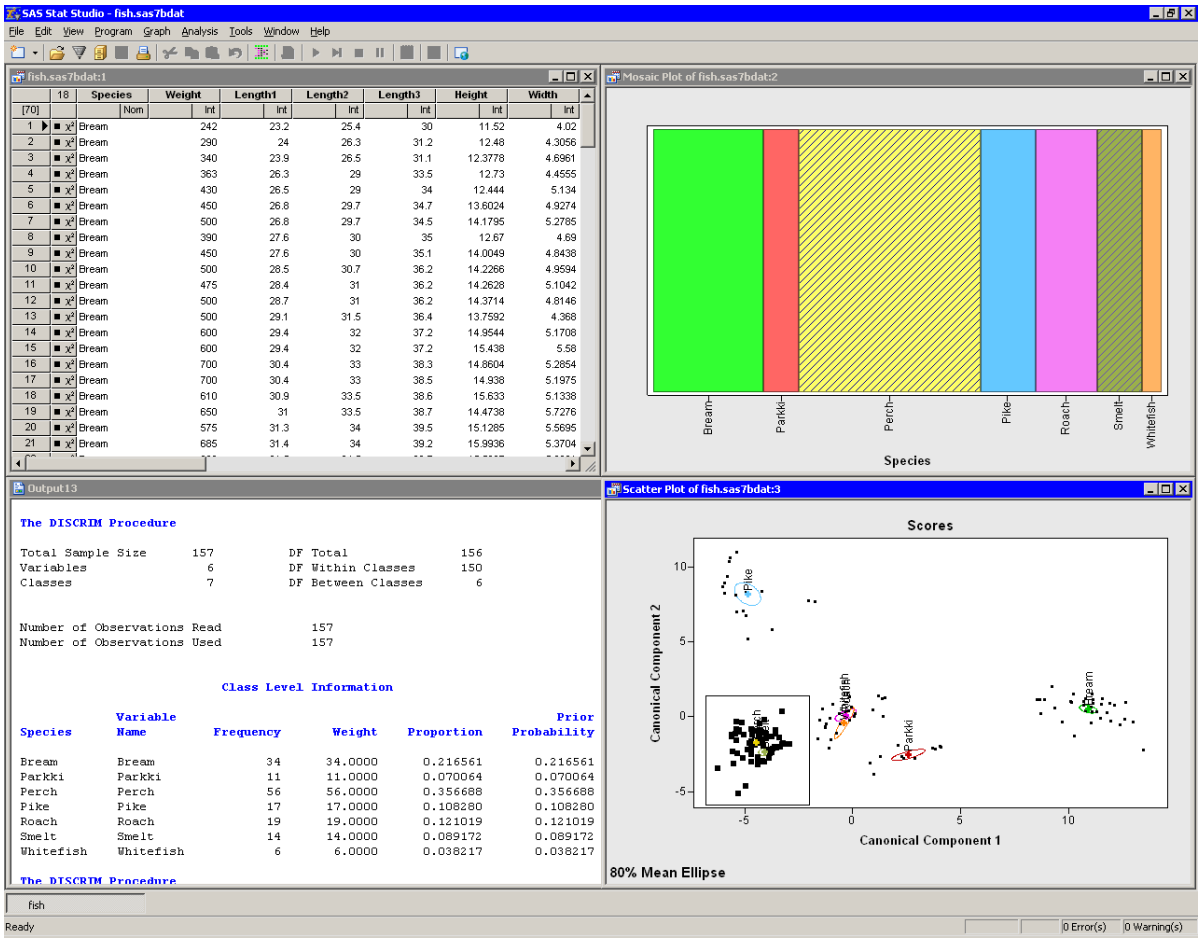
8 Click **OK**.

The analysis calls the DISCRIM procedure with the CANONICAL option. The procedure uses the options specified in the dialog box. The procedure displays tables in the output document, as shown in Figure 29.4. Two plots are also created.

The plot of the first two canonical components shows how well the first two canonical variables discriminate between the species of fish. The first canonical component differentiates among four groups: the pike-perch-smelt group, the roach-whitefish group, the parkki group, and the bream group. The second canonical component differentiates the pike groups from the other groups. Thus, the first two canonical components cannot differentiate between perch and smelt, nor between roach and whitefish. In Figure 29.4, a cloud of observations is selected. You can see from the linked bar chart that these observations consist of perch and smelt.

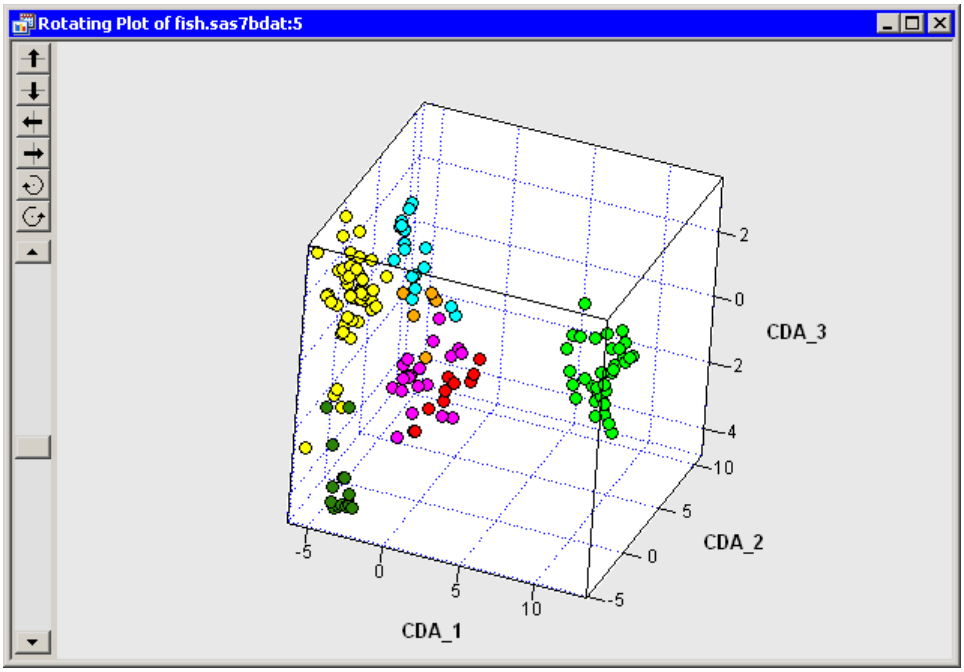
The location of the multivariate means for each species is indicated in the plot of the first two canonical components, along with an 80% confidence ellipse for the mean. The means of the perch and smelt groups are close to each other, as are the means of the roach and whitefish.

Figure 29.4 Output from a Canonical Discriminant Analysis



NOTE: The third canonical component helps to differentiate between perch and smelt, and between roach and whitefish. The canonical variables were added to the data table by the analysis, so you can create a scatter plot of the second and third canonical variables (CDA_3 versus CDA_2) or create a rotating plot of all three canonical components, as shown in Figure 29.5.

Figure 29.5 A Rotating Plot of the Canonical Components



The output window contains many tables of statistics. Figure 29.4 shows a summary of the model, as well as the frequency and proportion of each species.

Recall that canonical discriminant analysis is equivalent to canonical correlation analysis between the quantitative variables and a set of dummy variables coded from the classification variable (in this case, Species). Figure 29.6 displays statistics that are related to the canonical correlations. The multivariate statistics and F approximations test the null hypothesis that all canonical correlations are zero. The small p-values for these tests (< 0.0001) are evidence for rejecting the null hypothesis that all canonical correlations are zero. The table of canonical correlations shows that the first three canonical components are all highly correlated with the classification variable.

Figure 29.6 Canonical Correlations

Output9

Multivariate Statistics and F Approximations

S=6 M=-0.5 N=71.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00036339	90.09	36	639.5	<.0001
Pillai's Trace	3.10039600	26.73	36	900	<.0001
Hotelling-Lawley Trace	52.11364485	208.02	36	410.72	<.0001
Roy's Greatest Root	39.17299881	979.32	6	150	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

The DISCRIM Procedure
Canonical Discriminant Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.987475	0.986679	0.001993	0.975108
2	0.952409	0.950143	0.007439	0.907083
3	0.839159	0.833037	0.023684	0.704188
4	0.633761	0.624395	0.047906	0.401653
5	0.335157	0.324303	0.071070	0.112330
6	0.005855		0.080061	0.000034

The portion of the output window shown in Figure 29.7 shows the canonical structure. These are tables of correlations between the canonical variables and the original variables. The canonical variables are linear combinations of the original variables, so you can sometimes interpret the canonical variables in terms of the original variables.

The "Total Canonical Structure" table displays the correlations without regard for group membership. Since these correlations do not account for the groups, they can sometimes be misleading.

The "Between Canonical Structure" table removes the within-class variability before computing the correlations. For each variable X, define the *group mean vector of* X to be the vector whose ith element is the mean of all values of X that belong to the same group as X_i. The values in the "Between Canonical Structure" table are the correlations between the group mean vectors of the canonical variables and the group mean vectors of the original variables.

The "Pooled Within Canonical Structure" table removes the between-class variability before computing the correlations. The values in this table are the correlations between the residuals of the original and canonical variables, after regressing them onto the group variable.

For this example, the "Total Canonical Structure" table and the "Between Canonical Structure" table have similar interpretations: the first canonical component is strongly correlated with Height. The second canonical variable is strongly correlated with the length variables, and also with Weight. The third canonical component is a weighted average of all the variables, with slightly more weight given to Width.

Figure 29.7 Canonical Structure

Output9

Total Canonical Structure

Variable	Label	Can1	Can2	Can3
Weight	Weight (grams)	0.230928	0.421330	0.407874
Length1	Length from nose to the beginning of tail (cm)	0.102126	0.670718	0.490648
Length2	Length from nose to notch of tail (cm)	0.119342	0.664838	0.506407
Length3	Length from nose to end of tail (cm)	0.222490	0.665768	0.484026
Height	Maximum Height (cm)	0.763514	0.130039	0.484129
Width	Maximum Width (cm)	0.240478	0.272013	0.694825

Between Canonical Structure

Variable	Label	Can1	Can2	Can3
Weight	Weight (grams)	0.374275	0.658620	0.561771
Length1	Length from nose to the beginning of tail (cm)	0.130134	0.824317	0.531306
Length2	Length from nose to notch of tail (cm)	0.151063	0.811667	0.544732
Length3	Length from nose to end of tail (cm)	0.277461	0.800776	0.512953
Height	Maximum Height (cm)	0.867799	0.142552	0.467608
Width	Maximum Width (cm)	0.342914	0.374107	0.841981

Pooled Within Canonical Structure

Variable	Label	Can1	Can2	Can3
Weight	Weight (grams)	0.045947	0.161964	0.279757
Length1	Length from nose to the beginning of tail (cm)	0.025493	0.323481	0.422219
Length2	Length from nose to notch of tail (cm)	0.030096	0.323927	0.440241
Length3	Length from nose to end of tail (cm)	0.057477	0.332292	0.431047
Height	Maximum Height (cm)	0.243284	0.080054	0.531781
Width	Maximum Width (cm)	0.052592	0.114934	0.523834

The first canonical variable separates the species most effectively. An examination of the "Raw Canonical Coefficients" table (Figure 29.8) shows that the first canonical variable is the following linear combination of the centered variables:

$$\text{Can}_1 = -0.0006\,\text{Weight} - 0.328\,\text{Length1} + \ldots - 1.44\,\text{Width}$$

The coefficients are standardized so that the canonical variables have zero mean and a pooled within-class variance equal to one.

The second canonical variable provides the greatest difference between group means while being uncorrelated with the first canonical variable.

Figure 29.8 also shows the coordinates of the group means in terms of the canonical variables. For example, the mean of the bream species projected onto the span of the first two canonical components is (10.91, 0.51). (Recall that the *span* of a set of vectors is the vector space consisting of all linear combinations of the vectors.) This agrees with the graph shown in Figure 29.4. The means of the perch and smelt groups are close to each other when projected onto the span of the first two canonical components. However, the third canonical component separates these means.

Figure 29.8 Canonical Coefficients and Group Means

Output9

Raw Canonical Coefficients

Variable	Label	Can1	Can2	Can3
Weight	Weight (grams)	-0.000625155	-0.005179382	-0.005657421
Length1	Length from nose to the beginning of tail (cm)	-0.328362122	-0.621556479	-2.913223091
Length2	Length from nose to notch of tail (cm)	-2.489821235	-0.702083723	4.037406197
Length3	Length from nose to end of tail (cm)	2.598449063	1.807327493	-1.148401859
Height	Maximum Height (cm)	1.114092550	-0.718015251	0.291026348
Width	Maximum Width (cm)	-1.439398520	-0.898812533	0.723585755

Class Means on Canonical Variables

Species	Can1	Can2	Can3
Bream	10.90666501	0.51280830	0.23380442
Parkki	2.56821869	-2.54947525	-0.47951075
Perch	-4.46929721	-1.70826770	1.28505496
Pike	-4.87623835	8.19805469	-0.15625411
Roach	-0.33684741	0.11460424	-1.14283441
Smelt	-4.07966832	-2.33972040	-4.03918108
Whitefish	-0.39747712	-0.41943132	1.04681646

Figure 29.9 displays a table that summarizes how many fish are classified (or misclassified) into each species. If the canonical components capture most of the between-class variation of the data, then the elements on the table's main diagonal are large, compared to the off-diagonal elements. For these data, two smelt are misclassified as perch, but no other fish are misclassified. This indicates that the first three canonical components are good discriminators for Species.

NOTE: If you choose different options on the **Method** tab, the classification of observations will be different.

Figure 29.9 Classification of Observations into Groups

Output13

Number of Observations and Percent Classified into Species

From Species	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish	Total
Bream	34	0	0	0	0	0	0	34
	100.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
Parkki	0	11	0	0	0	0	0	11
	0.00	100.00	0.00	0.00	0.00	0.00	0.00	100.00
Perch	0	0	54	0	0	2	0	56
	0.00	0.00	96.43	0.00	0.00	3.57	0.00	100.00
Pike	0	0	0	17	0	0	0	17
	0.00	0.00	0.00	100.00	0.00	0.00	0.00	100.00
Roach	0	0	0	0	19	0	0	19
	0.00	0.00	0.00	0.00	100.00	0.00	0.00	100.00
Smelt	0	0	0	0	0	14	0	14
	0.00	0.00	0.00	0.00	0.00	100.00	0.00	100.00
Whitefish	0	0	0	0	0	0	6	6
	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
Total	34	11	54	17	19	16	6	157
	21.66	7.01	34.39	10.83	12.10	10.19	3.82	100.00
Priors	0.21656	0.07006	0.35669	0.10828	0.12102	0.08917	0.03822	

Error Count Estimates for Species

	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish	Total
Rate	0.0000	0.0000	0.0357	0.0000	0.0000	0.0000	0.0000	0.0127
Priors	0.2166	0.0701	0.3567	0.1083	0.1210	0.0892	0.0382	

In summary, it is possible to use canonical discriminant analysis to discriminate between these species of fish by using three canonical components that are linear combinations of physical measurements. Trying to discriminate by using only two canonical components leads to classification errors, because the projection onto the span of the first two canonical components does not separate the perch group from the smelt group, nor does it separate the roach group from the whitefish group.

Specifying the Canonical Discriminant Analysis

This section describes the dialog box tabs that are associated with the Canonical Discriminant analysis. The Canonical Discriminant analysis calls the DISCRIM procedure with the CANONICAL option. See the DISCRIM procedure documentation in the *SAS/STAT User's Guide* for additional details.

Variables Tab

You can use the **Variables** tab to specify the variables for the analysis. The **Variables** tab is shown in Figure 29.2.

The variable in the **Y Variable (Classification)** list corresponds to the variable in the CLASS statement of the DISCRIM procedure. This variable must be nominal.

The variables in the **X Variables** list correspond to variables in the VAR statement of the DISCRIM procedure.

Method Tab

You can use the **Method** tab to set options in the analysis. (See Figure 29.3.) The **Method** tab contains the following UI controls:

Number of canonical variables
: specifies the number of canonical variables. This option corresponds to the NCAN= option in the PROC DISCRIM statement.

Covariance within groups
: specifies assumptions about the homogeneity of within-group covariances. This option corresponds to the POOL= option in the PROC DISCRIM statement.

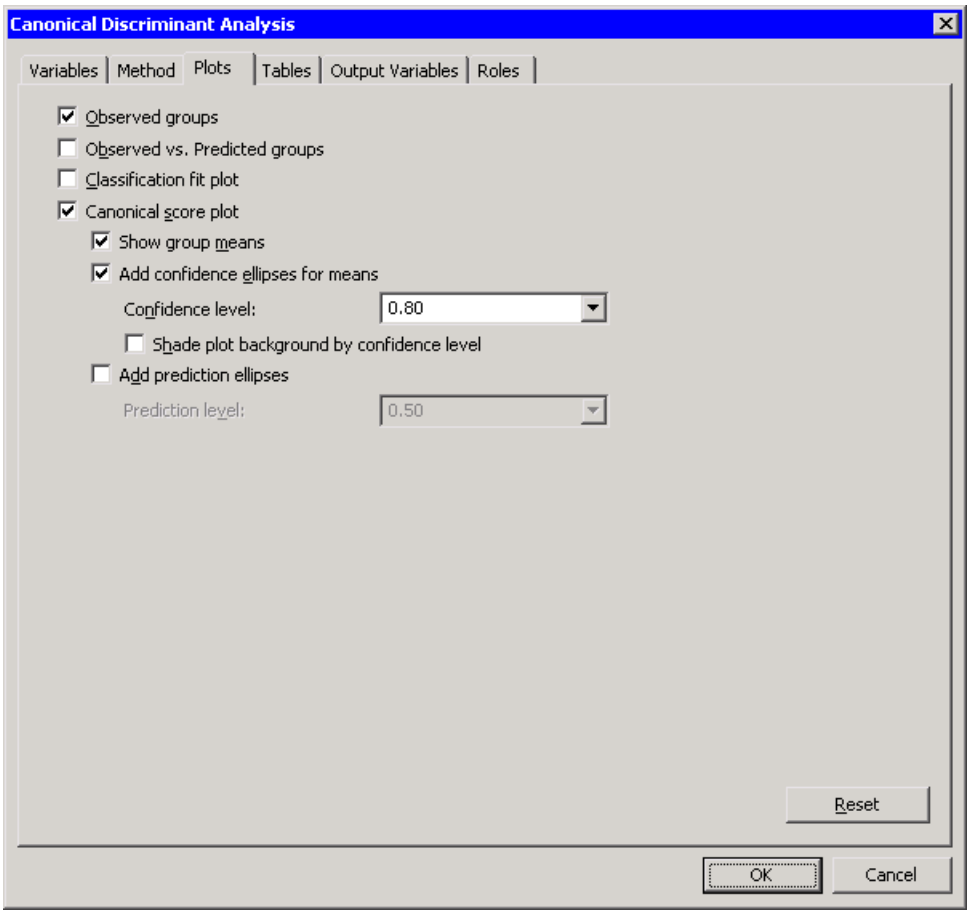
Prior probability of group membership
: specifies assumptions about the prior probabilities of group membership. This option corresponds to choosing either the EQUAL or PROPORTIONAL option in the PRIORS statement.

Classify observations by
: specifies a method of classifying observations based on their canonical scores. This option corresponds to the CROSSVALIDATE option in the PROC DISCRIM statement.

The Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 29.10.)

Figure 29.10 The Plots Tab



Creating a plot often adds one or more variables to the data table. The following plots are available:

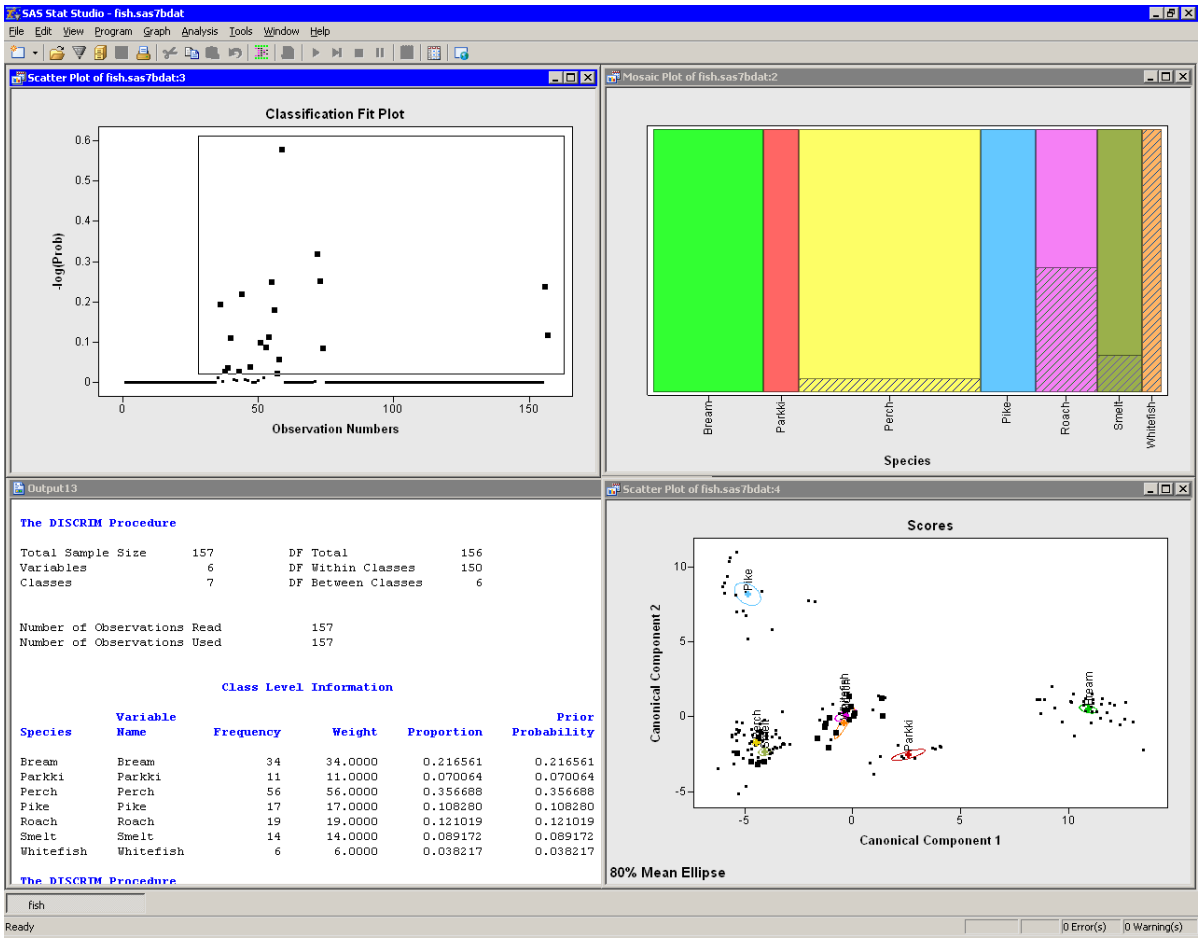
Observed groups
creates a spine plot (a one-dimensional mosaic plot) of the groups for the Y variable.

Observed vs. Predicted groups
creates a mosaic plot of the groups for the Y variable versus the group as classified by a discriminant function. Each observation is placed in the group that minimizes the generalized squared distance between the observation and the group mean.

Classification fit plot
creates a plot that indicates how well each observation is classified by the discriminant function. This plot is shown in Figure 29.11. Observations that are close to two or more group means are selected in the plot.

For each observation, PROC DISCRIM computes posterior probabilities for membership in each group. Let m_i be the maximum posterior probability for the ith observation. The classification fit plot is a plot of $-\log(m_i)$ versus i.

Figure 29.11 A Classification Fit Plot



Canonical score plot

creates a plot of the first two canonical variables. (If there is only one canonical variable, then a histogram of that variable is created instead.)

Show group means

displays the mean of each group in the score plot.

Add confidence ellipses for means

displays a confidence ellipse for the mean of each group in the score plot.

Confidence level

specifies the probability level for the confidence ellipse.

Shade plot background by confidence level

specifies that the background of each scatter plot be shaded according to a nested family of confidence ellipses.

Add prediction ellipses
displays a prediction ellipse for the mean of each group in the score plot, assuming multivariate normality within each group.

Prediction level
specifies the probability level for the prediction ellipse.

Tables Tab

The **Tables** tab is shown in Figure 29.12. You can use the **Tables** tab to display the following tables that summarize the results of the analysis. For more information, see the "Displayed Output" subsection of the "Details" section in the documentation for the DISCRIM procedure.

Simple statistics
specifies whether to display descriptive statistics for the total sample and within each group. This option corresponds to the SIMPLE option in the PROC DISCRIM statement.

Univariate ANOVA
specifies whether to display univariate statistics for testing the hypothesis that the population group means are equal for each variable. This option corresponds to the ANOVA option in the PROC DISCRIM statement.

Multivariate ANOVA
specifies whether to display multivariate statistics for testing the hypothesis that the population group means are equal for each variable. This option corresponds to the MANOVA option in the PROC DISCRIM statement.

Squared distances between group means
specifies whether to display the squared Mahalanobis distances (and associated statistics) between the group means. This option corresponds to the DISTANCE option in the PROC DISCRIM statement.

Standardized group means
specifies whether to display total-sample and pooled within-group standardized group means. This option corresponds to the STDMEAN option in the PROC DISCRIM statement.

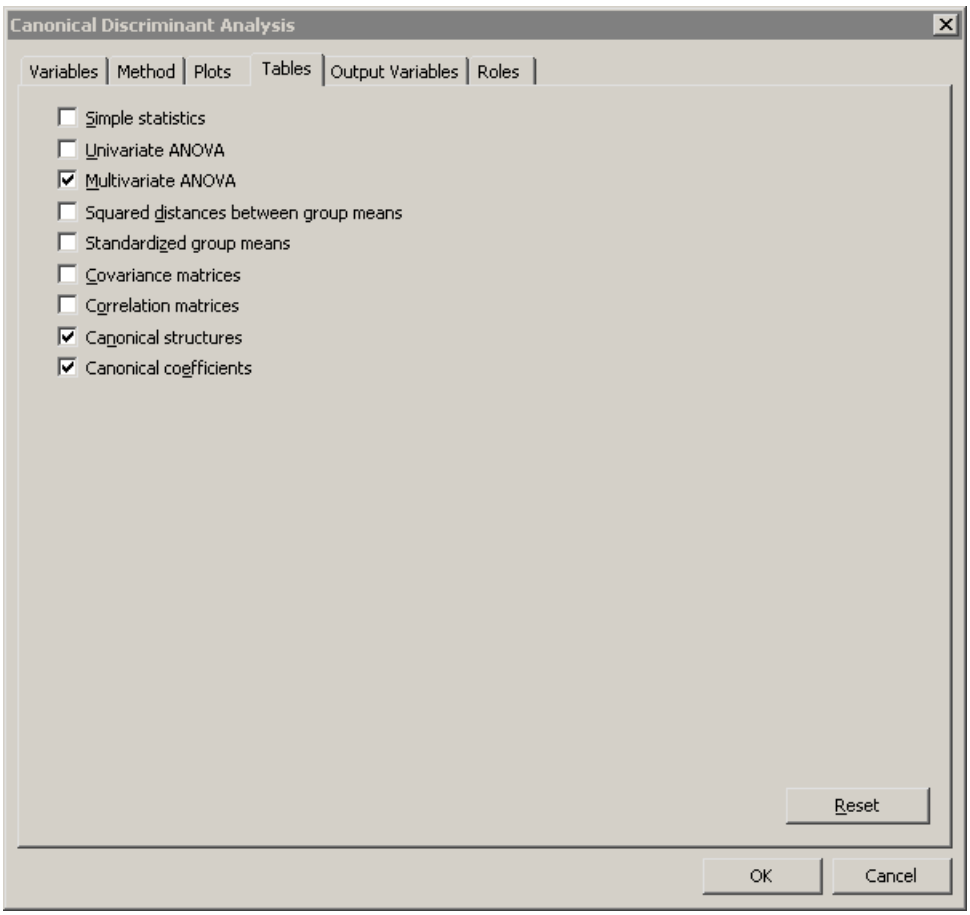
Covariance matrices
specifies whether to display the correlation matrices for each set of variables. This option corresponds to the BCOV, PCOV, TCOV, and WCOV options in the PROC DISCRIM statement.

Correlation matrices
specifies whether to display the correlation matrices for each set of variables. This option corresponds to the BCORR, PCORR, TCORR, and WCORR options in the PROC DISCRIM statement.

Canonical structures
specifies whether to display correlations between the canonical variables and the original variables.

Canonical coefficients
specifies whether to display the raw and standardized canonical coefficients for each set of variables.

Figure 29.12 The Tables Tab



In addition to the previous optional tables, the Canonical Discriminant analysis always creates the following tables. The name of the table refers to the ODS table name.

Counts
corresponds to the "Counts" table.

Class level information
corresponds to the "Levels" table.

Canonical correlations
corresponds to the "CanCorr" table. **NOTE:** This table looks like three tables: canonical correlations, eigenvalues of $E^{-1}H$, and tests for hypothesis that the canonical coefficients equal zero.

Class means on canonical variables
corresponds to the "CanonicalMeans" table.

Linear discriminant function
corresponds to the "LinearDiscFunc" table. This table is displayed only for the linear parametric classification method.

Number of observations and percent classified
corresponds to the "ClassifiedResub" or "ClassifiedCrossVal" table.

Error count estimates
corresponds to the "ErrorResub" or "ErrorCrossVal" table.

Output Variables Tab

You can use the **Output Variables** tab to add analysis variables to the data table. (See Figure 29.13.) If you request a plot that uses one of the output variables, then that variable is automatically created even if you did not explicitly select the variable on the **Output Variables** tab.

The following list describes each output variable added to the data table and indicates how the output variable is named. Y represents the name of the classification variable.

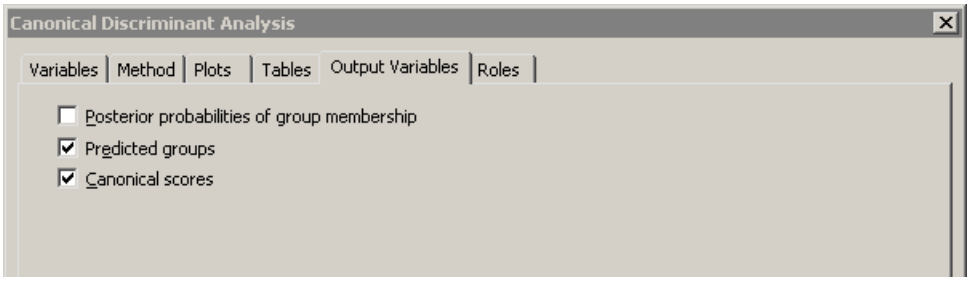
Posterior probabilities of group membership
adds variables named CDAProb_X, where X is the name of an X variable.

Predicted groups
adds a variable named CDAPred_Y that contains the name of the group to which each observation is assigned.

Canonical scores
adds variables named CDA_1 through CDA_k, where k is the number of canonical components.

If a classification fit plot is requested on the **Plots** tab, then a variable named CDALogProb_Y is created, as described in the section "The Plots Tab" on page 462.

Figure 29.13 The Output Variables Tab



Roles Tab

You can use the **Roles** tab to specify a frequency variable or weight variable for the analysis.

A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents n observations, where n is the value of the frequency variable.

A weight variable is a numeric variable with values that are relative weights for the analysis.

Analysis of Selected Variables

If a nominal variable is selected in a data table when you run the analysis, this variable is automatically entered in the **Y Variable (Classification)** field of the **Variables** tab.

Any selected interval variables are automatically entered in the **X Variables** field of the **Variables** tab.

Any variable in the data table with a Frequency or Weight role is automatically entered in the appropriate field of the **Roles** tab.

Chapter 30
Multivariate Analysis: Discriminant Analysis

Contents

Overview of Discriminant Analysis **469**
Example: Construct a Discriminant Function that Classifies Categories **470**
Specifying the Discriminant Analysis **475**
Variables Tab 475
The Method Tab 475
The Plots Tab 476
Tables Tab 477
Output Variables Tab 479
Roles Tab 479
Analysis of Selected Variables **479**

Overview of Discriminant Analysis

For a set of observations that contains one or more interval variables and also a classification variable that defines groups of observations, *discriminant analysis* derives a discriminant criterion function to classify each observation into one of the groups.

When the distribution within each group is assumed to be multivariate normal, a parametric method can be used to develop a discriminant function. The discriminant function, also known as a *classification criterion*, is determined by a generalized squared distance. The classification criterion can be based on either the individual within-group covariance matrices (yielding a quadratic function) or the pooled covariance matrix (yielding a linear function). It also takes into account the prior probabilities of the groups.

When no assumptions can be made about the distribution within each group, or when the distribution is not assumed to be multivariate normal, nonparametric methods can be used to estimate the group-specific densities. These methods include the kernel and k-nearest-neighbor methods.

You can run the Discriminant analysis by selecting **Analysis ►Multivariate Analysis ►Discriminant Analysis** from the main menu. The analysis is implemented by calling the DISCRIM procedure in SAS/STAT software. See the documentation for the DISCRIM procedure in the *SAS/STAT User's Guide* for additional details.

Example: Construct a Discriminant Function that Classifies Categories

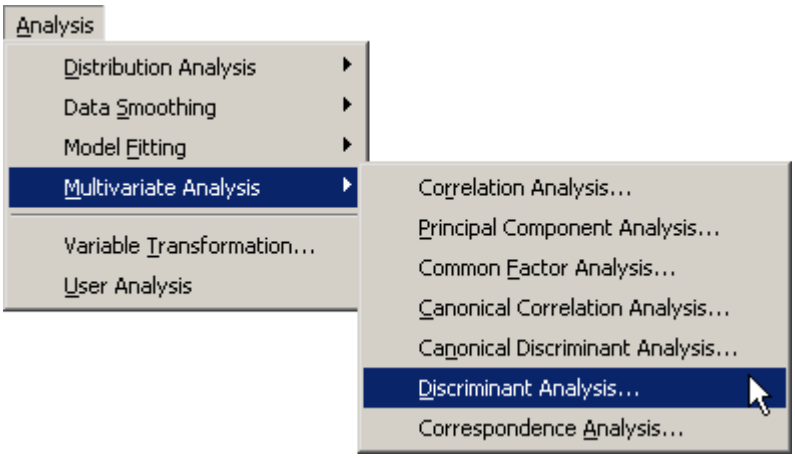
In this example, you examine measurements of 159 fish caught in Finland's Lake Laengelmavesi. The fish are one of seven species: bream, parkki, perch, pike, roach, smelt, and whitefish. Associated with each fish are physical measurements of weight, length, height, and width.

To construct a discriminant function that classifies species based on physical measurements:

1 Open the Fish data set.

2 Select **Analysis ►Multivariate Analysis ►Discriminant Analysis** from the main menu, as shown in Figure 30.1.

Figure 30.1 Selecting the Discriminant Analysis



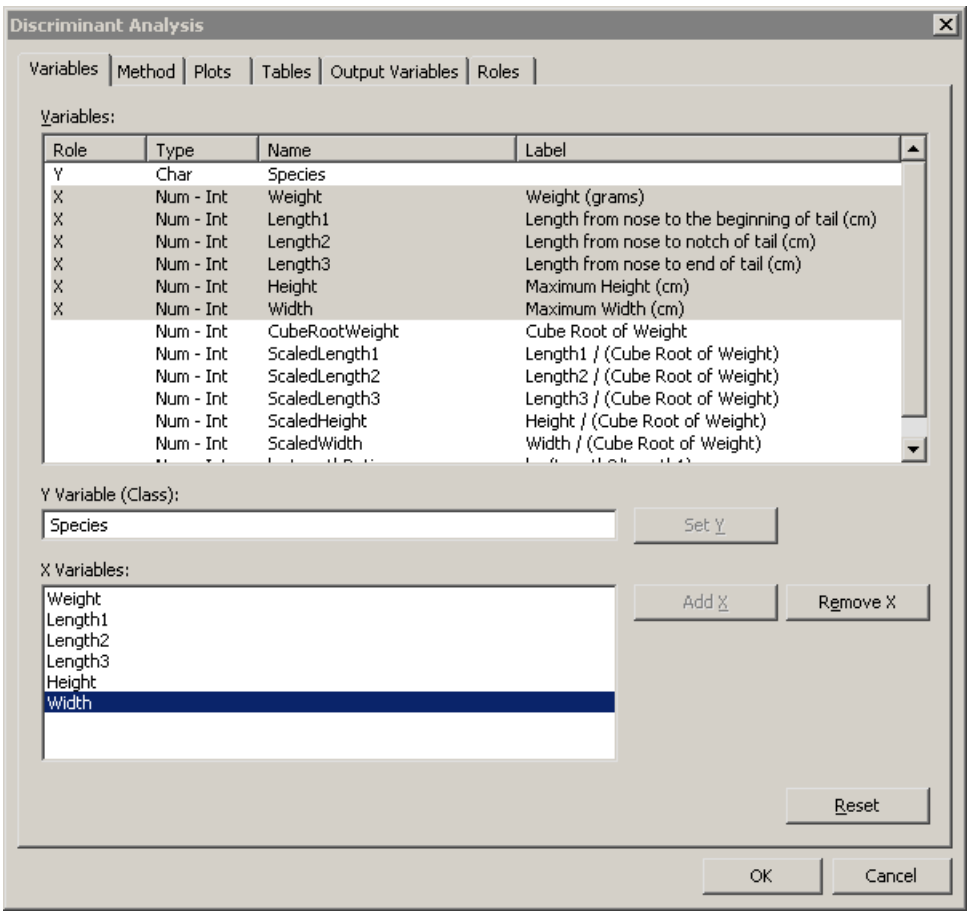
The Discriminant Analysis dialog box appears. (See Figure 30.2.) You can select variables for the analysis by using the **Variables** tab.

3 Select Species and click **Set Y**.

4 Select Weight. While holding down the CTRL key, select Length1, Length2, Length3, Height, and Width. Click **Add X**.

NOTE: Alternately, you can select the variables by using *contiguous selection*: click the first variable (Weight), hold down the SHIFT key, and click the last variable (Width). All variables between the first and last item are selected and can be added by clicking **Add X**.

Figure 30.2 The Variables Tab



5 Click the **Method** tab.

The **Method** tab becomes active. (See Figure 30.3.) You can use the **Method** tab to set options in the analysis.

6 Select **Kernel density** for **Classification method**.

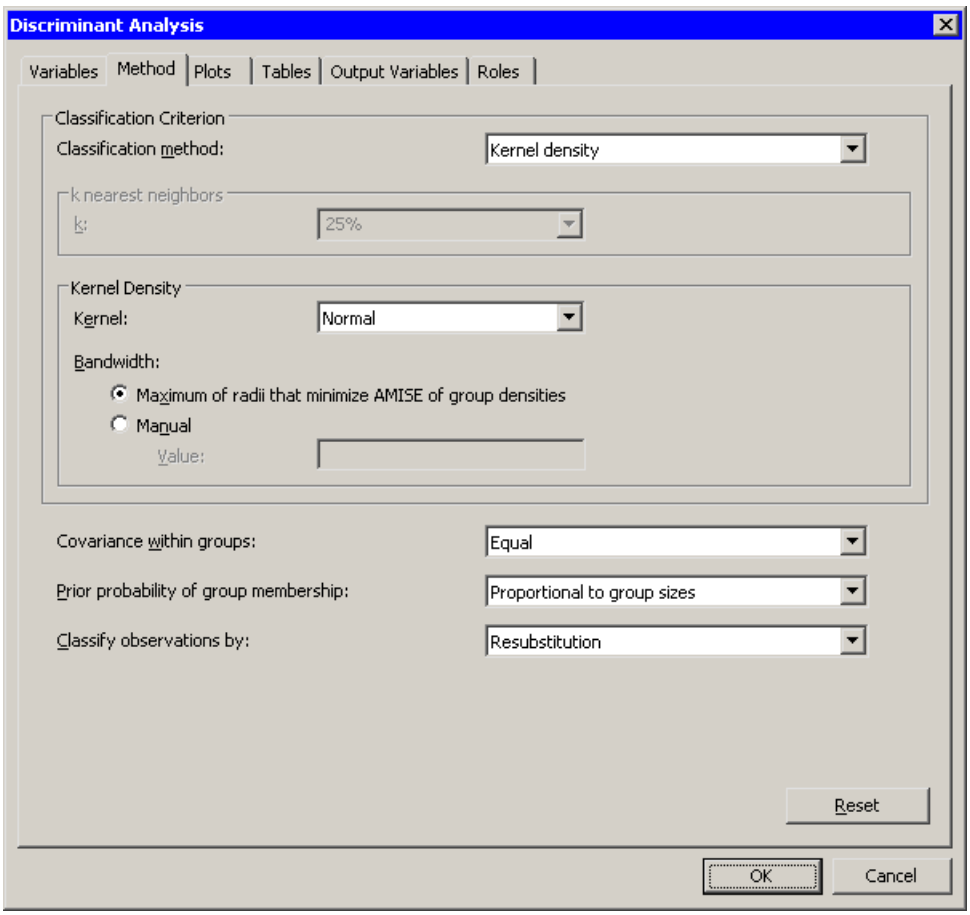
The options that are associated with the kernel density classification method become active.

7 Select **Normal** for **Kernel**.

The number of fish in the lake probably varies by species. That is, there is no reason to suspect that the number of whitefish in the lake is the same as the number of perch or bream. In the absence of prior knowledge about the distribution of fish species, you can assume that the number of fish of each species in the lake is proportional to the number in the sample.

8 Select **Proportional to group sizes** for **Prior probability of group membership**.

Figure 30.3 The Method Tab

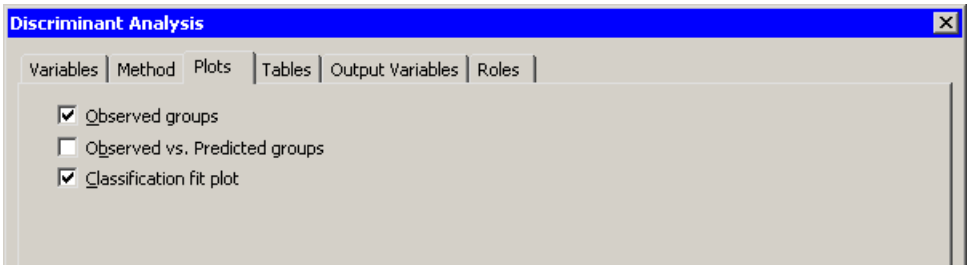


9 Click the **Plots** tab.

The **Plots** tab becomes active. (See Figure 30.4.)

10 Select **Classification fit plot**.

Figure 30.4 The Plots Tab



11 Click **OK**.

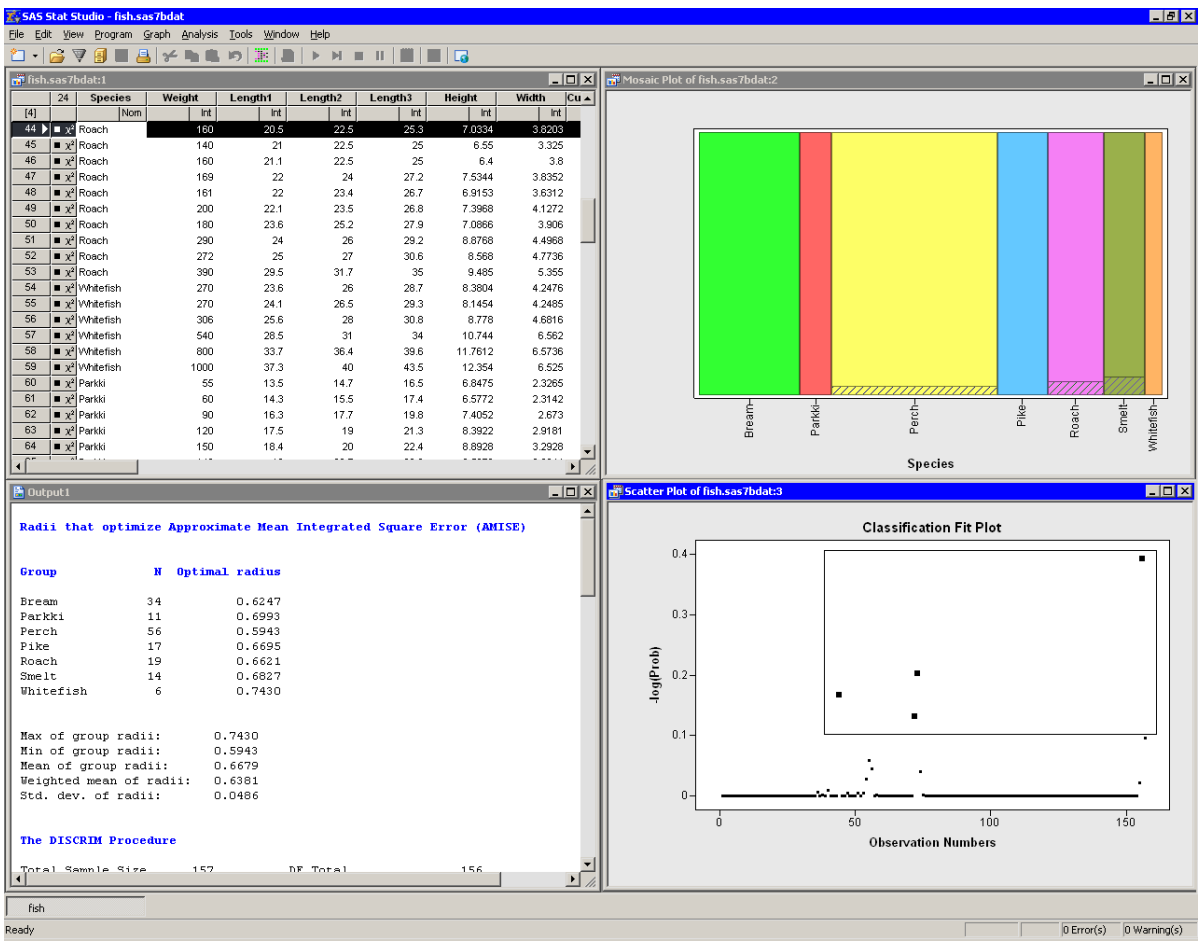
The analysis calls the DISCRIM procedure. The procedure uses the options specified in the dialog box. The procedure displays tables in the output document, as shown in Figure 30.5. Two plots are also created.

Move the classification fit plot so that the workspace is arranged as in Figure 30.5.

The classification fit plot indicates how well each observation is classified by the discriminant function. For each observation, PROC DISCRIM computes posterior probabilities for membership in each group. Let m_i be the maximum posterior probability for the ith observation. The classification fit plot is a plot of $-\log(m_i)$ versus i. In Figure 30.5, the selected observations are those with $-\log(m_i) \geq 0.1$. Equivalently, the maximum posterior probability for membership for the selected observations is less than $\exp(-0.1) \approx 0.9$. The selected fish are those with relatively large probabilities of misclassification. Conversely, selecting the bream, parkki, and pike species in the spine plot (the upper right plot in Figure 30.5) shows that the classification criterion discriminates between these species quite well. A *spine plot* is a one-dimensional mosaic plot in which the width of a bar represent the number of observations in a category.

NOTE: If there are k groups, then the maximum posterior probability of membership is at least $1/k$, so the vertical axis of the classification fit plot is bounded above by $\log(k)$.

Figure 30.5 Output from a Discriminant Analysis



The output window contains many tables of statistics. The first table in Figure 30.5 is produced by SAS/IML Studio. It is associated with a heuristic method of choosing the bandwidth for the kernel density classification method. This table is described in the section "The Method Tab" on page 475.

Figure 30.6 displays a table that summarizes how many fish are classified (or misclassified) into each species. If the discriminant function correctly classifies most observations, then the elements on the table's main diagonal are large compared to the off-diagonal elements. For this example, the nonparametric discriminant function correctly classified all fish into the species to which they belong.

NOTE: The classification in this example was performed using resubstitution. This estimate of the error rate is optimistically biased. You can obtain a less biased estimate by using cross validation. You can select cross validation for the **Classify observations by** option on the **Method** tab.

Figure 30.6 Classification of Observations into Groups

Output1

Number of Observations and Percent Classified into Species

From Species	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish	Total
Bream	34	0	0	0	0	0	0	34
	100.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
Parkki	0	11	0	0	0	0	0	11
	0.00	100.00	0.00	0.00	0.00	0.00	0.00	100.00
Perch	0	0	56	0	0	0	0	56
	0.00	0.00	100.00	0.00	0.00	0.00	0.00	100.00
Pike	0	0	0	17	0	0	0	17
	0.00	0.00	0.00	100.00	0.00	0.00	0.00	100.00
Roach	0	0	0	0	19	0	0	19
	0.00	0.00	0.00	0.00	100.00	0.00	0.00	100.00
Smelt	0	0	0	0	0	14	0	14
	0.00	0.00	0.00	0.00	0.00	100.00	0.00	100.00
Whitefish	0	0	0	0	0	0	6	6
	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
Total	34	11	56	17	19	14	6	157
	21.66	7.01	35.67	10.83	12.10	8.92	3.82	100.00
Priors	0.21656	0.07006	0.35669	0.10828	0.12102	0.08917	0.03822	

Error Count Estimates for Species

	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish	Total
Rate	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Priors	0.2166	0.0701	0.3567	0.1083	0.1210	0.0892	0.0382	

In summary, the nonparametric discriminant function in this example does an excellent job of discriminating among these species of fish.

Specifying the Discriminant Analysis

This section describes the dialog box tabs that are associated with the Discriminant analysis. The Discriminant analysis calls the DISCRIM procedure option. See the DISCRIM procedure documentation in the *SAS/STAT User's Guide* for additional details.

Variables Tab

You can use the **Variables** tab to specify the variables for the analysis. The **Variables** tab is shown in Figure 30.2.

The variable in the **Y Variable (Classification)** list corresponds to the variable in the CLASS statement of the DISCRIM procedure. This variable must be nominal.

The variables in the **X Variables** list correspond to variables in the VAR statement of the DISCRIM procedure.

The Method Tab

You can use the **Method** tab to set options in the analysis. (See Figure 30.3.) The **Method** tab contains the following UI controls:

Classification method
specifies the method used to construct the discriminant function.

Parametric
specifies that a parametric method based on a multivariate normal distribution within each group be used to derive a linear or quadratic discriminant function. This corresponds to the METHOD=NORMAL option in the PROC DISCRIM statement.

k nearest neighbors
specifies that a nonparametric classification method be used. An observation is classified into a group based on the information from the k nearest neighbors of the observation. This corresponds to the METHOD=NPAR K= option in the PROC DISCRIM statement.

Kernel density
specifies that a nonparametric classification method be used. An observation is classified into a group based on the information from observations within a given radius of the observation. This corresponds to the METHOD=NPAR R= option in the PROC DISCRIM statement.

k
specifies the number of nearest neighbors for the **k nearest neighbors** method. You can select a fixed number of observations, or a proportion of the total number of observations. You can type a value in this field or choose from a set of standard values. This option corresponds to the K= or KPROP= option in the PROC DISCRIM statement.

Kernel
specifies the shape of the kernel function for the **Kernel density** method. You can specify a uniform,

Epanechnikov (quadratic), or normal kernel function. This corresponds to the KERNEL= option in the PROC DISCRIM statement.

Bandwidth
specifies the bandwidth for the kernel density classification method. This corresponds to the R= option in the PROC DISCRIM statement. There are two options for choosing the bandwidth:

Maximum of radii that minimizes AMISE of group densities
This option uses a heuristic to automatically choose a bandwidth. The "Background" subsection of the "Details" section in the documentation for the DISCRIM procedure presents formulas for the bandwidths that minimize an approximate mean integrated square error of the estimated density within each group. The formulas assume the data within each group are multivariate normal.

The optimal radius for each group is determined for each group, as shown in Figure 30.5. Descriptive statistics of the radii are also displayed, including the mean of the radii weighted by the number of observations in each group. The bandwidth used for the R= option in the PROC DISCRIM statement is the maximum of the radii.

Manual
sets the kernel bandwidth to the value in the **Value** field.

Covariance within groups
specifies assumptions about the homogeneity of within-group covariances. This option corresponds to the POOL= option in the PROC DISCRIM statement. For the parametric classification method, the assumption of equal covariances results in a linear discriminant function. The assumption of unequal covariances results in a quadratic discriminant function.

Prior probability of group membership
specifies assumptions about the prior probabilities of group membership. This option corresponds to the EQUAL and PROPORTIONAL options in the PRIORS statement.

Classify observations by
specifies a method of classifying observations based on their canonical scores. This option corresponds to the CROSSVALIDATE option in the PROC DISCRIM statement.

The Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 30.4.)

Creating a plot often adds one or more variables to the data table. The following plots are available:

Observed groups
creates a spine plot (a one-dimensional mosaic plot) of the groups for the Y variable.

Observed vs. Predicted groups
creates a mosaic plot of the groups for the Y variable versus the group as classified by a discriminant function. Each observation is placed in the group that minimizes the generalized squared distance between the observation and the group mean.

Classification fit plot

creates a plot that indicates how well each observation is classified by the discriminant function. This plot is shown in Figure 30.5. The observations selected in the plot have a low posterior probability of group membership.

For each observation, PROC DISCRIM computes posterior probabilities for membership in each group. Let m_i be the maximum posterior probability for the ith observation. The classification fit plot is a plot of $-\log(m_i)$ versus i.

Tables Tab

The **Tables** tab is shown in Figure 30.7. You can use the **Tables** tab to display the following tables that summarize the results of the analysis. For more information, see the "Displayed Output" subsection of the "Details" section in the documentation for the DISCRIM procedure.

Simple statistics

specifies whether to display descriptive statistics for the total sample and within each group. This option corresponds to the SIMPLE option in the PROC DISCRIM statement.

Univariate ANOVA

specifies whether to display univariate statistics for testing the hypothesis that the population group means are equal for each variable. This option corresponds to the ANOVA option in the PROC DISCRIM statement.

Multivariate ANOVA

specifies whether to display multivariate statistics for testing the hypothesis that the population group means are equal for each variable. This option corresponds to the MANOVA option in the PROC DISCRIM statement.

Squared distances between group means

specifies whether to display the squared Mahalanobis distances (and associated statistics) between the group means. This option corresponds to the DISTANCE option in the PROC DISCRIM statement.

Standardized group means

specifies whether to display total-sample and pooled within-group standardized group means. This option corresponds to the STDMEAN option in the PROC DISCRIM statement.

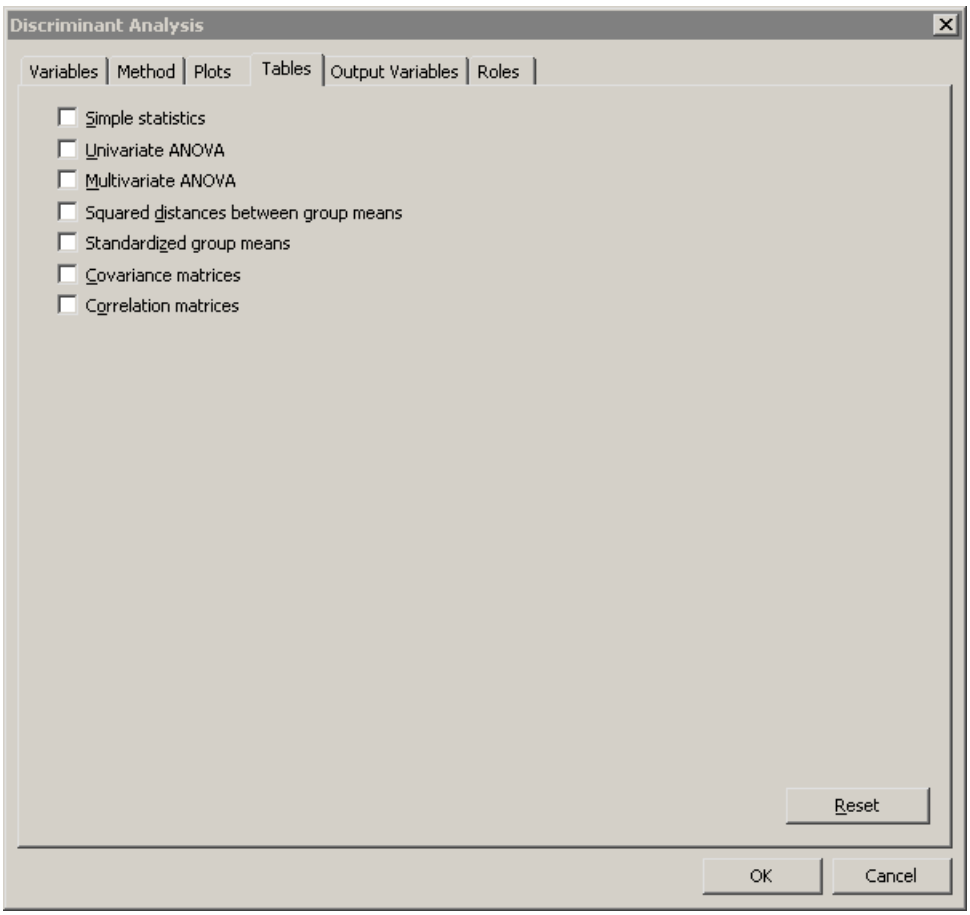
Covariance matrices

specifies whether to display the correlation matrices for each set of variables. This option corresponds to the BCOV, PCOV, TCOV, and WCOV options in the PROC DISCRIM statement.

Correlation matrices

specifies whether to display the correlation matrices for each set of variables. This option corresponds to the BCORR, PCORR, TCORR, and WCORR options in the PROC DISCRIM statement.

Figure 30.7 The Tables Tab



In addition to the previous optional tables, the Discriminant analysis always creates the following tables. The name of the table refers to the ODS table name.

Counts
corresponds to the "Counts" table.

Class level information
corresponds to the "Levels" table.

Linear discriminant function
corresponds to the "LinearDiscFunc" table. This table is displayed only for the linear parametric classification method.

Number of observations and percent classified
corresponds to the "ClassifiedResub" or "ClassifiedCrossVal" table.

Error count estimates
corresponds to the "ErrorResub" or "ErrorCrossVal" table.

Output Variables Tab

You can use the **Output Variables** tab to add analysis variables to the data table. (See Figure 30.8.) If you request a plot that uses one of the output variables, then that variable is automatically created even if you did not explicitly select the variable on the **Output Variables** tab.

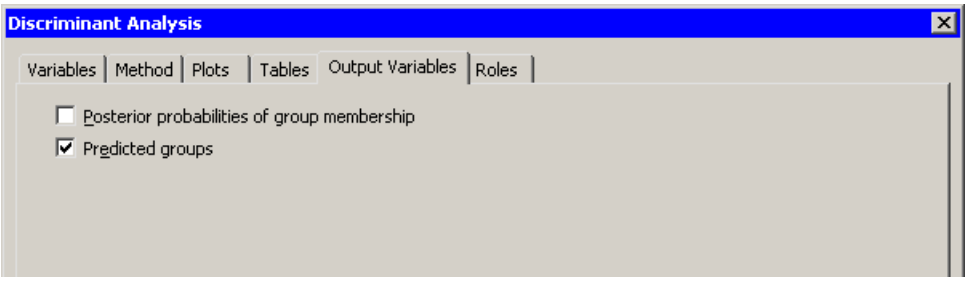
The following list describes each output variable and indicates how the output variable is named. Y represents the name of the classification variable.

Posterior probabilities of group membership
adds variables named DiscProb_X, where X is the name of an X variable.

Predicted groups
adds a variable named DiscPred_Y that contains the name of the group to which each observation is assigned.

If a classification fit plot is requested on the **Plots** tab, then a variable named DiscLogProb_Y is created, as described in the section "The Plots Tab" on page 476.

Figure 30.8 The Output Variables Tab



Roles Tab

You can use the **Roles** tab to specify a frequency variable or weight variable for the analysis.

A frequency variable is a numeric variable whose value represents the frequency of the observation. If you use a frequency variable, the underlying procedure assumes that each observation represents n observations, where n is the value of the frequency variable.

A weight variable is a numeric variable with values that are relative weights for the analysis.

Analysis of Selected Variables

If a nominal variable is selected in a data table when you run the analysis, this variable is automatically entered in the **Y Variable (Classification)** field of the **Variables** tab.

Any selected interval variables are automatically entered in the **X Variables** field of the **Variables** tab.

Any variable in the data table with a Frequency or Weight role is automatically entered in the appropriate field of the **Roles** tab.

Chapter 31
Multivariate Analysis: Correspondence Analysis

Contents

Overview of the Correspondence Analysis **481**
Example: Summarize the Association between Categories **482**
Specifying the Correspondence Analysis **489**
Variables Tab 489
The Method Tab 489
Plots Tab 490
Tables Tab 491
Roles Tab 492
Analysis of Selected Variables **493**
References **493**

Overview of the Correspondence Analysis

The Correspondence analysis performs simple correspondence analysis, which you can use to analyze frequency data and associations between two or more nominal variables. The correspondence analysis finds a low-dimensional representation of the rows and columns of a contingency table consisting of the counts for the variables.

While principal component analysis constructs directions in the space of variables that explain variance, correspondence analysis constructs directions (sometimes called *principal coordinates*) that explain *inertia*. Inertia is the total chi-square statistic divided by the total number of observations. Correspondence analysis computes directions that best explain deviations from expected values (assuming no association). The analysis graphically represents each row and column by a point in a *configuration plot*.

You can run the Correspondence analysis by selecting **Analysis ►Multivariate Analysis ►Correspondence Analysis** from the main menu. The analysis is implemented by calling the CORRESP procedure in SAS/STAT software. See the documentation for the CORRESP procedure in the *SAS/STAT User's Guide* for additional details. For a general introduction to correspondence analysis, see Friendly (2000).

Example: Summarize the Association between Categories

In this example, you examine data from 1991 about 127 companies from five nations in four industries. The companies are from Britain, France, Germany, Japan, and the United States. The companies are in the following industries: automobiles, electronics, food, and oil.

To run a correspondence analysis:

1 Open the Business data set.

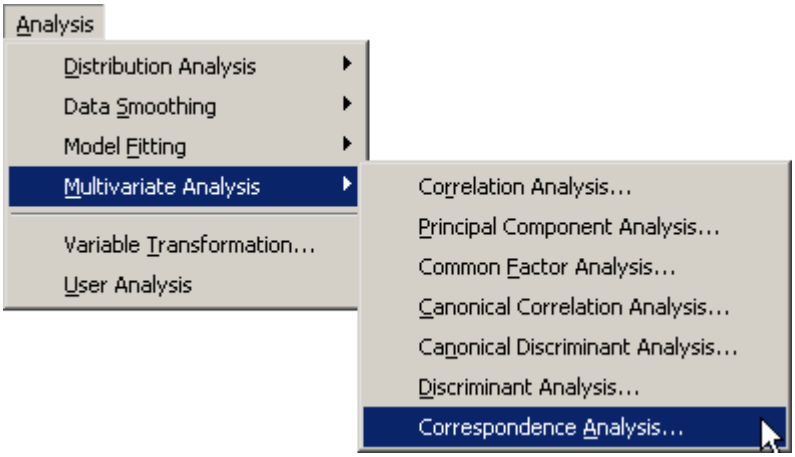
Table 31.1 shows a contingency table of the number of companies in each Industry for each Nation. The goal of this example is to use correspondence analysis to examine relationships between and among the Nation and Industry variables.

Table 31.1 Contingency Table of Industry and Nation

	Nation				
Industry	**Britain**	**France**	**Germany**	**Japan**	**U.S.**
Automobiles	2	3	5	14	7
Electronics	1	3	1	12	11
Food	11	2	0	11	19
Oil	2	2	1	5	13

2 Select **Analysis ►Multivariate Analysis ►Correspondence Analysis** from the main menu, as shown in Figure 31.1.

Figure 31.1 Selecting the Correspondence Analysis

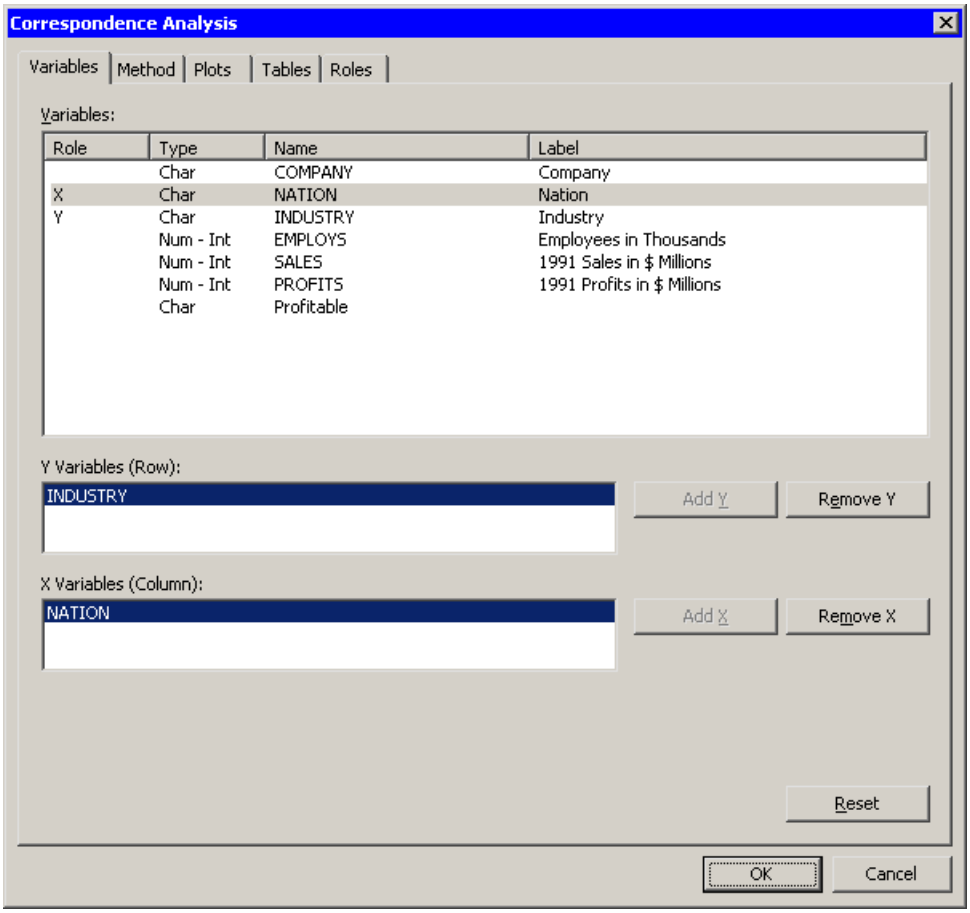


The Correspondence Analysis dialog box appears. (See Figure 31.2.) You can select variables for the analysis by using the **Variables** tab. In Table 31.1, the levels of Industry specify the rows of the table and are displayed along the vertical dimension of the table. Thus Industry is the Y variable whose values determine the rows. Similarly, Nation is the X variable whose values determine the columns.

3 Select Industry and click **Add Y**.

4 Select Nation and click **Add X**.

Figure 31.2 The Variables Tab

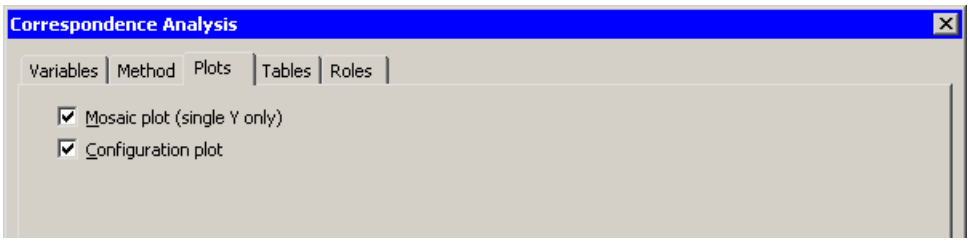


5 Click the **Plots** tab.

The **Plots** tab becomes active. (See Figure 31.3.)

6 Select **Mosaic plot (single Y only)**.

Figure 31.3 The Plots Tab

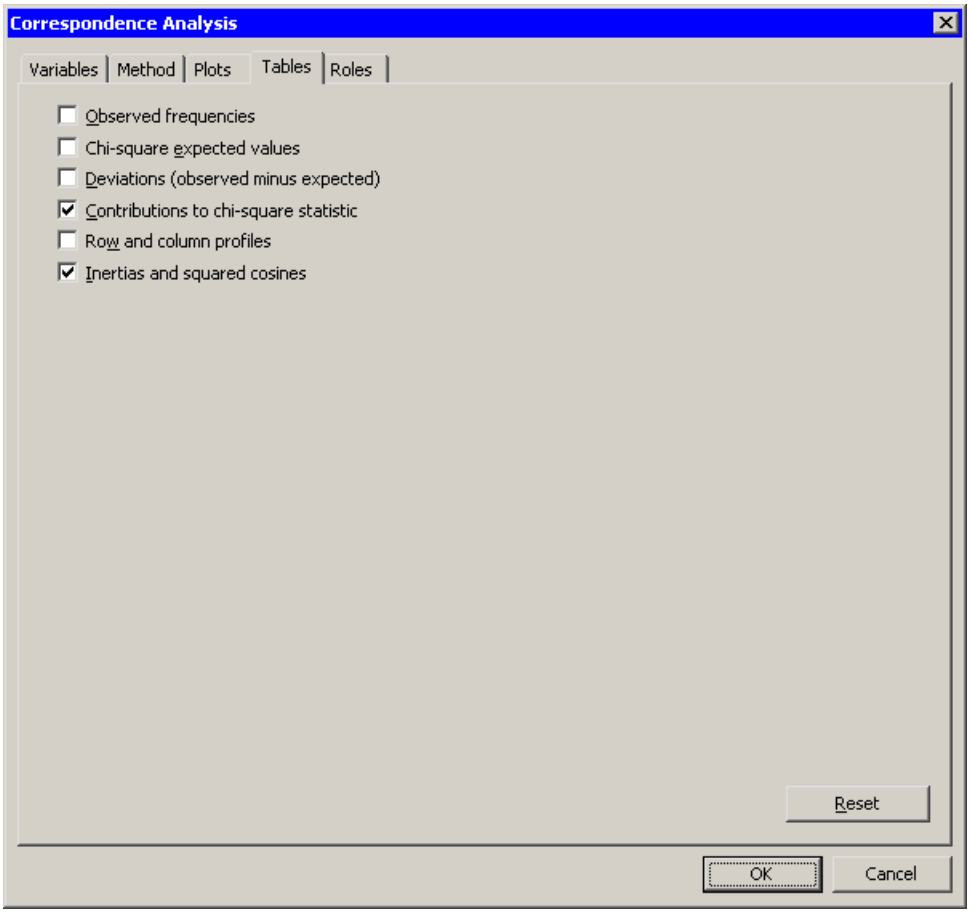


7 Click the **Tables** tab.

The **Tables** tab becomes active. (See Figure 31.4.) For this example, it is informative to see how each cell, column, and row of Table 31.1 contributes to the chi-square association statistic for the table.

8 Select **Contributions to chi-square statistic**.

Figure 31.4 The Tables Tab



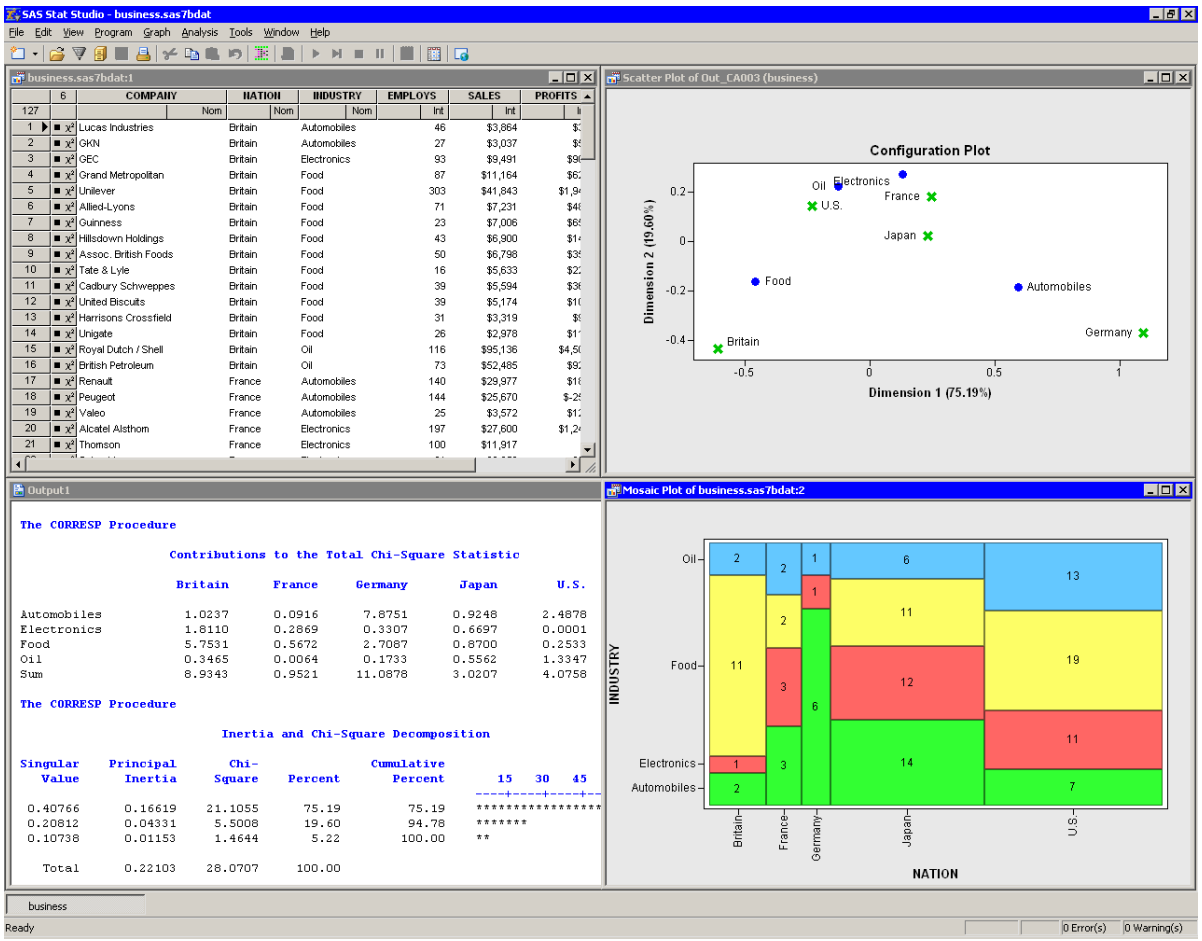
9 Click **OK**.

The analysis calls the CORRESP procedure. The procedure uses the options specified in the dialog box. The procedure displays tables in the output document, as shown in Figure 31.5. Two plots are also created.

The mosaic plot indicates the frequency count for each cell in the contingency table. You can add labels to the cells of the mosaic plot to make the frequency count more evident.

10 Activate the mosaic plot. Press the “l” key (lowercase “L”) to toggle labels.

Figure 31.5 Output from a Correspondence Analysis



The mosaic plot shows several interesting facts. The British companies are not evenly divided among industries; many British companies in these data are food companies. Similarly, the lack of German food companies is evident, as is the preponderance of German automobile companies. The United States has the largest proportion of oil companies.

Correspondence analysis plots all the categories in a Euclidean space. The first two dimensions of this space are plotted in a *configuration plot*, shown in the upper right corner of Figure 31.5. As indicated by the labels for the axes, the first principal coordinate accounts for 75% of the inertia, while the second accounts for almost 20%. Thus, these two principal coordinates account for almost 95% of the inertia in this example. The plot should be thought of as two different overlaid plots, one for each categorical variable. Distances between points within a variable have meaning, but distances between points from different variables do not.

The configuration plot summarizes association between categories, and indicates the contribution to the chi-square statistic from each cell. To interpret the plot, start by interpreting the row points: the categories of Industry. The points for food and automobiles are farthest from the origin, so these industries contribute the most to the chi-square statistic. Oil and electronics contribute relatively less to the chi-square statistic.

For the column points, the points for the United States, France, and Japan are near the origin, so these countries contribute a relatively small amount to the chi-square statistic. The points for Britain and Germany are far from the origin; they make relatively large contributions to the chi-square statistic.

The "Contributions to the Total Chi-Square Statistic" table in Figure 31.6 displays the contributions to the chi-square statistic for each industry and country. The last column summarizes the contributions for industry. Automobiles (12.4) and food (10.15) contribute the most, a fact apparent from the configuration plot. Similarly, the last row summarizes the contributions for countries. Britain and Germany make the largest contributions.

The "Inertia and Chi-Square Decomposition" table summarizes the chi-square decomposition. The first two components account for almost 95% of the chi-square association.

Figure 31.6 Contributions to the Chi-Square Statistic

Output1

The CORRESP Procedure

Contributions to the Total Chi-Square Statistic

	Britain	France	Germany	Japan	U.S.	Sum
Automobiles	1.0237	0.0916	7.8751	0.9248	2.4878	12.4029
Electronics	1.8110	0.2869	0.3307	0.6697	0.0001	3.0984
Food	5.7531	0.5672	2.7087	0.8700	0.2533	10.1523
Oil	0.3465	0.0064	0.1733	0.5562	1.3347	2.4171
Sum	8.9343	0.9521	11.0878	3.0207	4.0758	28.0707

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	15 30 45 60 75
0.40766	0.16619	21.1055	75.19	75.19	**********
0.20812	0.04331	5.5008	19.60	94.78	*******
0.10738	0.01153	1.4644	5.22	100.00	**
Total	0.22103	28.0707	100.00		

Degrees of Freedom = 12

The next series of tables summarize the correspondence analysis for the row variable (Industry). These tables are shown in Figure 31.7.

The "Row Coordinates" table displays the coordinates of the various industries in the configuration plot. The "Summary Statistics" table displays various statistics, including the so-called *quality* of the representation. Categories with low quality values (for example, oil) are not well represented by the two principal coordinates. The quality statistic is equal to the sum of the squared cosines, which are displayed in the last table of Figure 31.7. The squared cosines are the square of the cosines of the angles between each axis and a vector from the origin to the point. Thus, points with a squared cosine near 1 are located near a principal coordinate axis, and so have high quality.

The "Partial Contributions to Inertia" table indicates how much of the total inertia is accounted for by each category in each dimension. This table corresponds to the spread of the points in the configuration plot in the horizontal and vertical dimensions. In the first principal coordinate, automobiles and food contribute the most. In the second principal coordinate, electronics contributes the most, although the contributions are more evenly spread across categories.

For further details, see the "Algorithm and Notation" and "Displayed Output" sections of the documentation for the CORRESP procedure.

Figure 31.7 Output from a Correspondence Analysis

Output1

Row Coordinates

	Dim1	Dim2
Automobiles	0.5938	-0.1854
Electronics	0.1305	0.2714
Food	-0.4566	-0.1632
Oil	-0.1259	0.2230

Summary Statistics for the Row Points

	Quality	Mass	Inertia
Automobiles	0.9984	0.2520	0.4418
Electronics	0.8194	0.2205	0.1104
Food	0.9959	0.3386	0.3617
Oil	0.6510	0.1890	0.0861

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
Automobiles	0.5346	0.1999
Electronics	0.0226	0.3749
Food	0.4248	0.2082
Oil	0.0180	0.2169

Indices of the Coordinates that Contribute Most to Inertia for the Row Points

	Dim1	Dim2	Best
Automobiles	1	0	1
Electronics	0	2	2
Food	1	1	1
Oil	0	2	2

Squared Cosines for the Row Points

	Dim1	Dim2
Automobiles	0.9097	0.0887
Electronics	0.1538	0.6656
Food	0.8831	0.1128
Oil	0.1573	0.4937

The analysis of the countries is similar. Figure 31.8 shows a partial view of the related statistics. The quality statistic helps explain a seeming discrepancy in the configuration plot. (See Figure 31.5.) From the configuration plot (and from the "Column Coordinates" table), it is apparent that the marker that represents Japan is closer to the origin than the marker that represents France. It is tempting to conclude that Japan contributes less to the chi-square statistic than France. But the "Contributions to the Total Chi-Square Statistic" table in Figure 31.6 and the "Partial Contributions to Inertia" table in Figure 31.8 show that the opposite is true.

The contradictory evidence can be resolved by noticing that the quality statistic for Japan is only 0.787. That value is the sum of the squared cosines for each dimension. The squared cosine for the second dimension is nearly zero, which indicates that Japan's position is almost completely determined by the first dimension.

Figure 31.8 Output from a Correspondence Analysis

Output1

Column Coordinates

	Dim1	Dim2
Britain	-0.6066	-0.4351
France	0.2472	0.1814
Germany	1.0939	-0.3711
Japan	0.2340	0.0228
U.S.	-0.2316	0.1427

Summary Statistics for the Column Points

	Quality	Mass	Inertia
Britain	0.9979	0.1260	0.3183
France	0.9877	0.0787	0.0339
Germany	0.9627	0.0630	0.3950
Japan	0.7865	0.3386	0.1076
U.S.	0.9077	0.3937	0.1452

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
Britain	0.2789	0.5506
France	0.0290	0.0598
Germany	0.4536	0.2003
Japan	0.1115	0.0041
U.S.	0.1270	0.1852

You cannot compare row points with column points in the configuration plot. For example, you cannot compare the distance from the origin for electronics to the distance for Japan and draw any meaningful conclusions.

However, you can interpret associations between rows and columns. For example, the first principal coordinate shows a greater association with being British and being a food company than would be expected if these two categories were independent. Similarly, the association between being German and being an automobile company is greater than expected under the assumption of independence.

Specifying the Correspondence Analysis

This section describes the dialog box tabs that are associated with the Correspondence analysis. The Correspondence analysis calls the CORRESP procedure option. See the CORRESP procedure documentation in the *SAS/STAT User's Guide* for additional details.

Variables Tab

You can use the **Variables** tab to specify the variables for the analysis. The **Variables** tab is shown in Figure 31.2.

The variables in the **Y Variables (Row)** list corresponds to the row variables in the TABLE statement of the CORRESP procedure. These variables must be nominal.

The variables in the **X Variables (Col)** list corresponds to the column variables in the TABLE statement of the CORRESP procedure. These variables must be nominal.

These variables are used to construct the rows and columns of a contingency table. You can specify a Weight variable on the **Roles** tab to read category frequencies.

The Method Tab

You can use the **Method** tab to set options in the analysis. (See Figure 31.9.) The tab supports the following options:

Cross levels of row variables
specifies that each combination of levels for all row variables become a row label. When not selected, each level of every row variable becomes a row label. This corresponds to the CROSS=ROW option in the PROC CORRESP statement.

Cross levels of column variables
specifies that each combination of levels for all column variables become a column label. When not selected, each level of every column variable becomes a column label. This corresponds to the CROSS=COL option in the PROC CORRESP statement.

Selecting both of the previous options corresponds to specifying the CROSS=BOTH option in the PROC CORRESP statement; clearing both of the previous options corresponds to specifying the CROSS=NONE option.

Missing values
specifies whether to include observations with missing values in the analysis.

Exclude from analysis specifies that observations with missing values be excluded from the analysis.

Use as category levels specifies that missing values be treated as a distinct level of each categorical variable. This corresponds to the MISSING option in the PROC CORRESP statement.

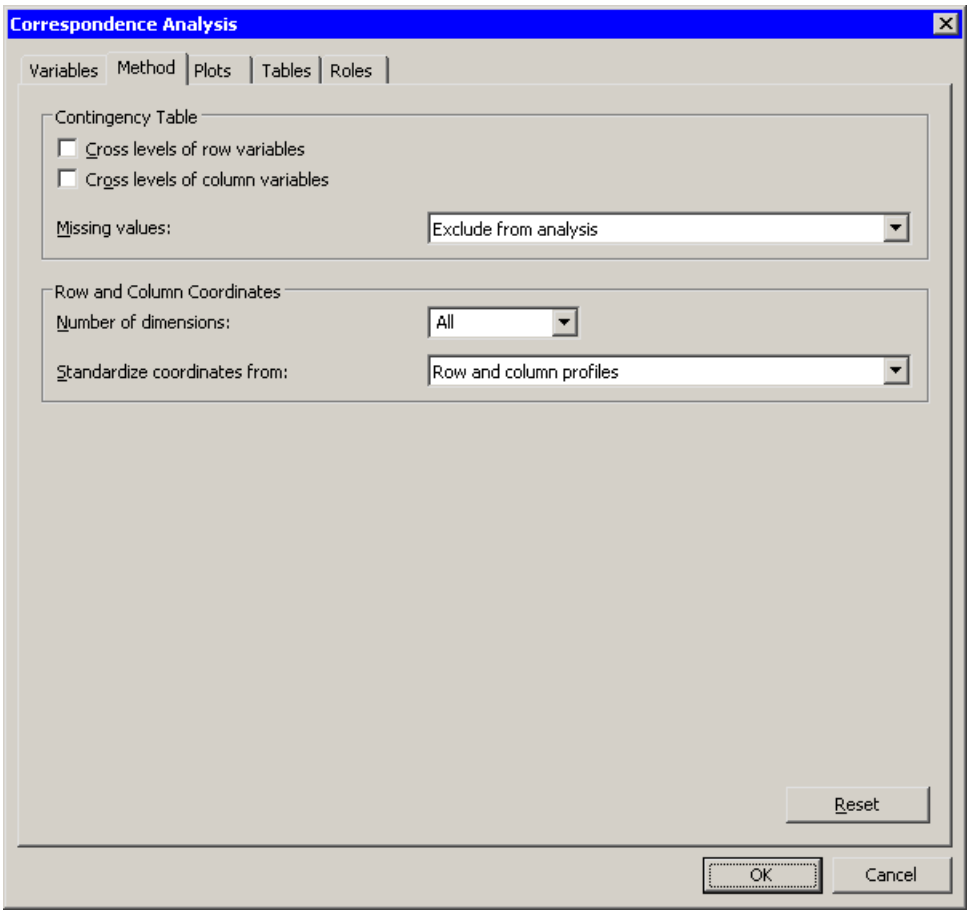
Number of dimensions

specifies the number of principal coordinates to use for the analysis. You can type a value in this field. If your contingency table is an $R \times C$ table, the number of dimensions in the correspondence analysis is at most $\min(R-1, C-1)$. This corresponds to the DIMENS= option in the PROC CORRESP statement.

Standardize coordinates from

specifies the standardization for the row and column coordinates. This corresponds to the PROFILE= option in the PROC CORRESP statement.

Figure 31.9 The Method Tab



Plots Tab

You can use the **Plots** tab to create plots that graphically display results of the analysis. (See Figure 31.3.) The following plots are available:

Mosaic plot (single Y only)

creates a mosaic plot of a single Y variable versus the X variables. The mosaic plot is a graphical representation of the contingency table for the data.

Configuration plot
creates a plot of the first two principal coordinates. These directions account for the greatest deviation from independence. The row and column categories are plotted in these coordinates.

NOTE: The configuration plot is not linked to the original data set because it has a different number of observations. However, you can view the data for this plot by pressing the F9 key when the plot is active. The data are created by the combination of the SOURCE and OUTC= options in the PROC CORRESP statement.

Tables Tab

The **Tables** tab is shown in Figure 31.4. You can use the **Tables** tab to display the following tables that summarize the results of the analysis. For more information, see the "Displayed Output" subsection of the "Details" section in the documentation for the CORRESP procedure.

Observed frequencies
specifies whether to display the contingency table of observed frequencies. This option corresponds to the OBSERVED option in the PROC CORRESP statement.

Chi-square expected values
specifies whether to display the expected frequencies for the contingency table. This option corresponds to the EXPECTED option in the PROC CORRESP statement.

Deviations (observed minus expected)
specifies whether to display the difference between the observed and expected frequencies for the contingency table. This option corresponds to the DEVIATIONS option in the PROC CORRESP statement.

Contributions to chi-square statistic
specifies whether to display contributions to the total chi-square test statistic, including the row and column marginals and the total chi-square statistic. This option corresponds to the CELLCHI2 option in the PROC CORRESP statement.

Row and column profiles
specifies whether to display row and column profiles. The row profile is the matrix of row-conditional probabilities. The column profile is the matrix of column-conditional probabilities. This option corresponds to the RP and CP options in the PROC CORRESP statement.

Inertias and squared cosines
specifies whether to display statistics that are related to inertia and squared cosines. The names of the ODS tables displayed by this option are "Inertias," "ColBest," "ColContr," "ColQualMassIn," "ColSqCos," "RowBest," "RowContr," "RowQualMassIn," and "RowSqCos."

In addition to the previous optional tables, the Correspondence analysis always creates the following tables:

Row coordinates
corresponds to the "RowCoors" table.

Column coordinates
corresponds to the "ColCoors" table.

Roles Tab

You can use the **Roles** tab to specify a weight variable or supplementary variables for the analysis. (See Figure 31.10.)

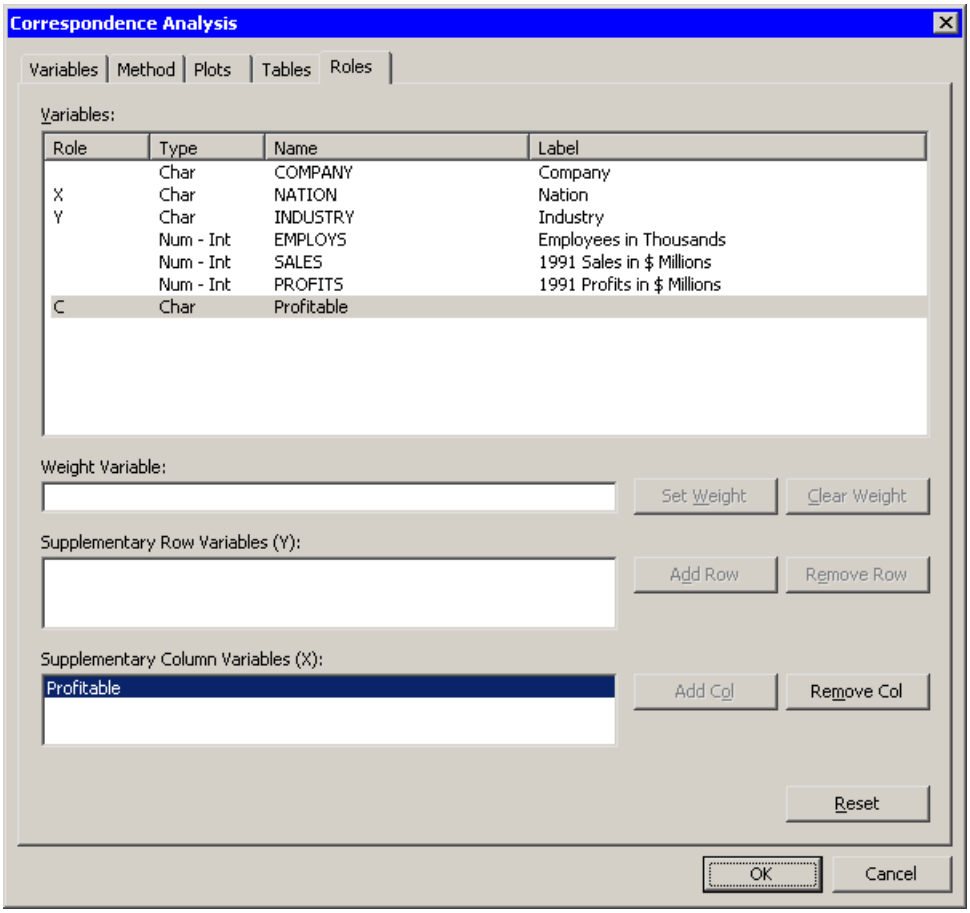
A weight variable is a numeric variable that represents category frequencies. In the absence of a weight variable, each observation contributes a value of 1 to the frequency count for its category. That is, each observation represents one subject. When you specify a weight variable, each observation contributes the value of the weighting variable for that observation. For example, a weight of 3 means that the observation represents three subjects.

Supplementary variables are displayed as points in the configuration plot, but these variables are not used in computing the correspondence analysis. In other words, a supplementary variable is projected onto the principal coordinate directions, but it is not used to compute the principal coordinates.

NOTE: In the CORRESP procedure, supplementary variables must be listed in the TABLE statement in addition to being listed in the SUPPLEMENTARY statement. In SAS/IML Studio, you should not specify supplementary variables on the **Variables** tab.

As an example of using supplementary variables, suppose you use the Variable Transformation Wizard to create a nominal variable that indicates whether a company is profitable. You can display the levels of this variable in the configuration plot by adding the variable to a supplementary variable list, as shown in Figure 31.10.

Figure 31.10 The Roles Tab



Analysis of Selected Variables

If a nominal variable is selected in a data table when you run the analysis, this variable is automatically entered in the **Y Variables (Row)** field of the **Variables** tab.

Any variable in the data table with a Weight role is automatically entered in the appropriate field of the **Roles** tab.

References

Friendly, M. (2000), *Visualizing Categorical Data*, Cary, NC: SAS Institute Inc.

Chapter 32
Variable Transformations

Contents

Overview of Variable Transformations **495**
Example: Apply a Logarithmic Transformation **496**
Example: Apply a Box-Cox Transformation **500**
Common Transformations **504**
Normalizing Transformations **506**
Variance Stabilizing Transformations **508**
Transformations for Proportion Variables **509**
The Folded Power Transformation 510
The Guerrero-Johnson Transformation 511
The Aranda-Ordaz Transformation 511
Scaling and Translation Transformations **511**
Rank Transformations **513**
Lag Transformations **514**
Two-Variable Transformations **516**
Custom Transformations **518**
Example: Define a Custom Transformation 520
Applying Normalizing Transformations **523**
Translating Data 523
Skewness 524
References **526**

Overview of Variable Transformations

Transforming data is an important technique in exploratory data analysis. Centering and scaling are simple examples of transforming data.

More complex transformations are useful for a variety of purposes. A variable that violates the assumptions of a statistical technique can sometimes be transformed to fit the assumptions better. For example, a variable that is not normally distributed can be transformed in an attempt to improve normality; a variable with nonhomogeneous variance can be transformed in an attempt to improve homogeneity of variance.

You can create new variables in a data set by transforming existing variables. SAS/IML Studio provides a Variable Transformation Wizard that enables you to quickly apply standard transformations to your data. These include normalizing transformations (such as logarithmic and power transformations), logit and probit transformations, affine transformations (including centering and standardizing), and rank transformations.

You can create your own transformations within the Variable Transformation Wizard by using SAS DATA step syntax and functions. These enable you to recode variables, to create variables with simulated values from known distributions, and to use arbitrarily complex formulas and logical statements to define new variables.

Most SAS/IML Studio transformations create a new numerical variable from an existing numerical variable. You can define custom DATA step transformations that use and create variables of any type.

You can apply transformations to all observations, or you can apply the transformation only to observations that are included in analyses.

Example: Apply a Logarithmic Transformation

Many statistical analyses assume that the data are normally distributed. If a variable is not normally distributed, it is often possible to improve normality by using an appropriate transformation of the variable. The three transformations used most often for this purpose are the logarithmic, square root, and inverse transformations.

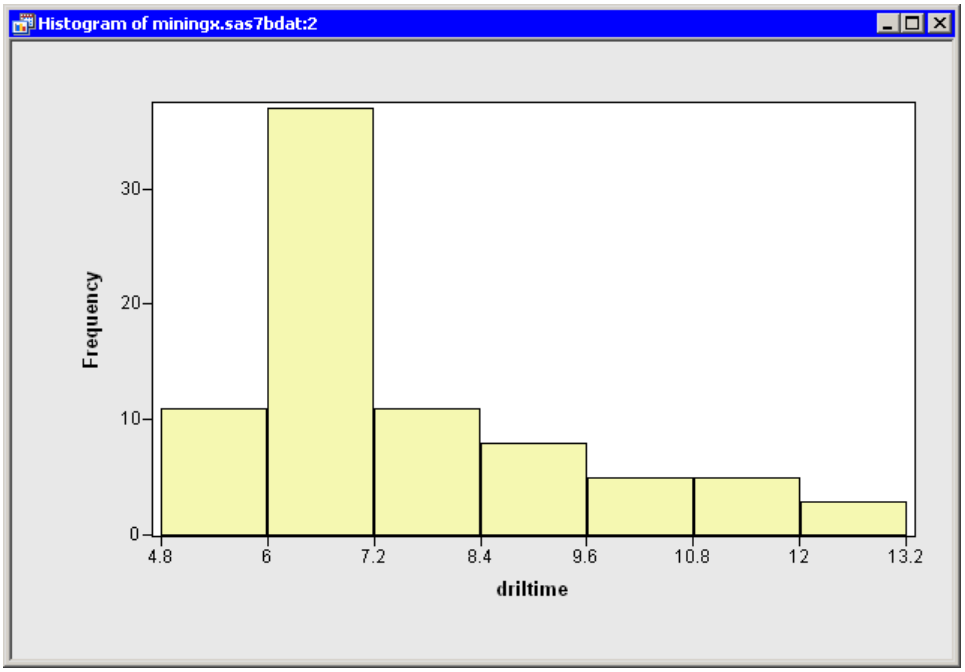
The following steps apply a logarithmic transformation to the driltime variable of the Miningx data set. Because the driltime variable is nonnegative, a logarithmic transformation is well-defined.

1 Open the Miningx data set.

2 Create a histogram of the driltime variable.

The histogram is shown in Figure 32.1.

Figure 32.1 Histogram of Drilling Time



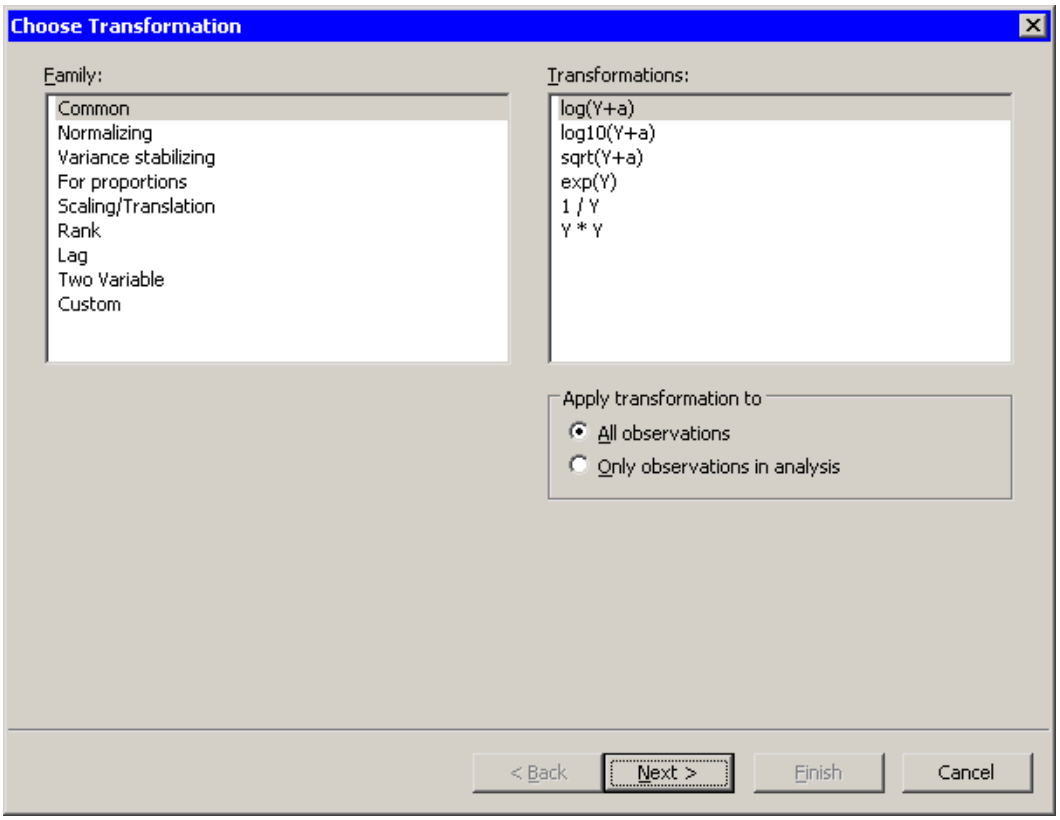
Clearly, the driltime variable is not normally distributed. You might explore whether some transformation

of driltime is approximately normal. To begin, you might try a logarithmic transformation.

3 Select **Analysis ►Variable Transformation** from the main menu.

The Variable Transformation Wizard in Figure 32.2 appears.

Figure 32.2 Selecting a Transformation



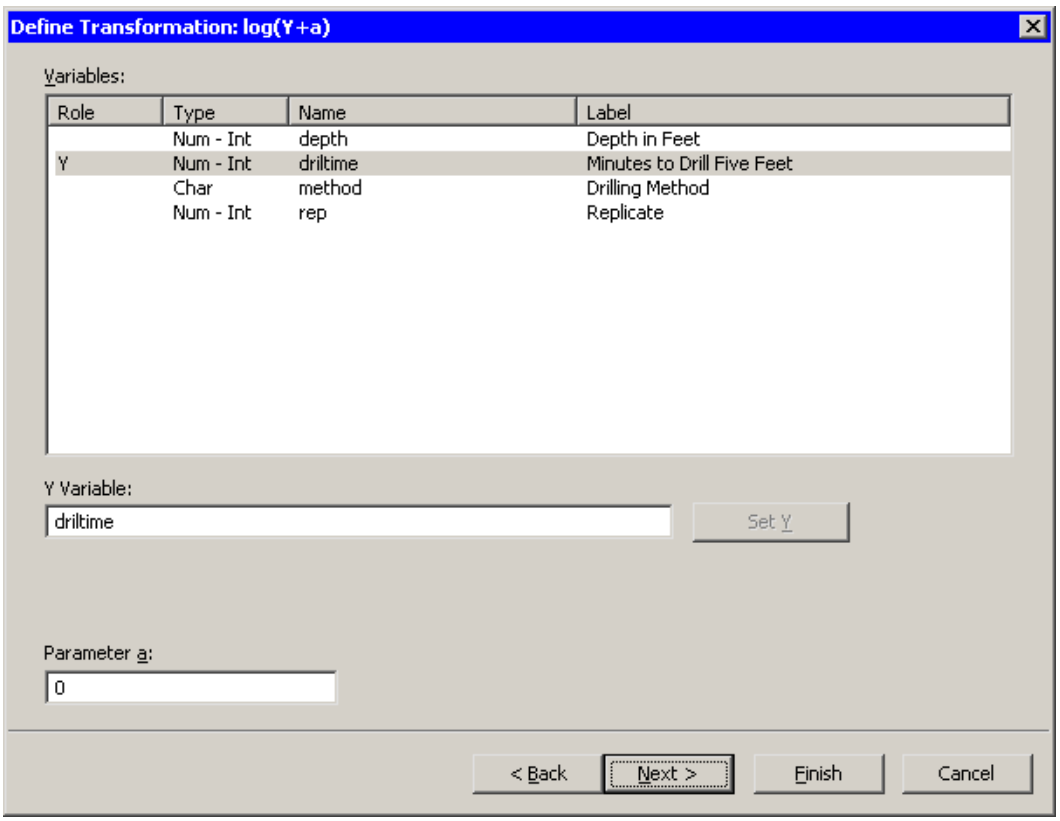
The first page of the wizard enables you to select a transformation family and a specific transformation within that family. The logarithmic transformation is available from several items in the **Family** list, including the **Common** family. This transformation is of the form $\log(y + a)$, so you need to specify the variable y and the parameter a.

The transformation **log(Y+a)** is highlighted by default. Since this is the desired transformation, you can proceed to the next page of the wizard.

4 Click **Next**.

The wizard displays the page shown in Figure 32.3. Note that the transformation appears on the page's title bar.

Figure 32.3 Selecting a Variable and a Parameter



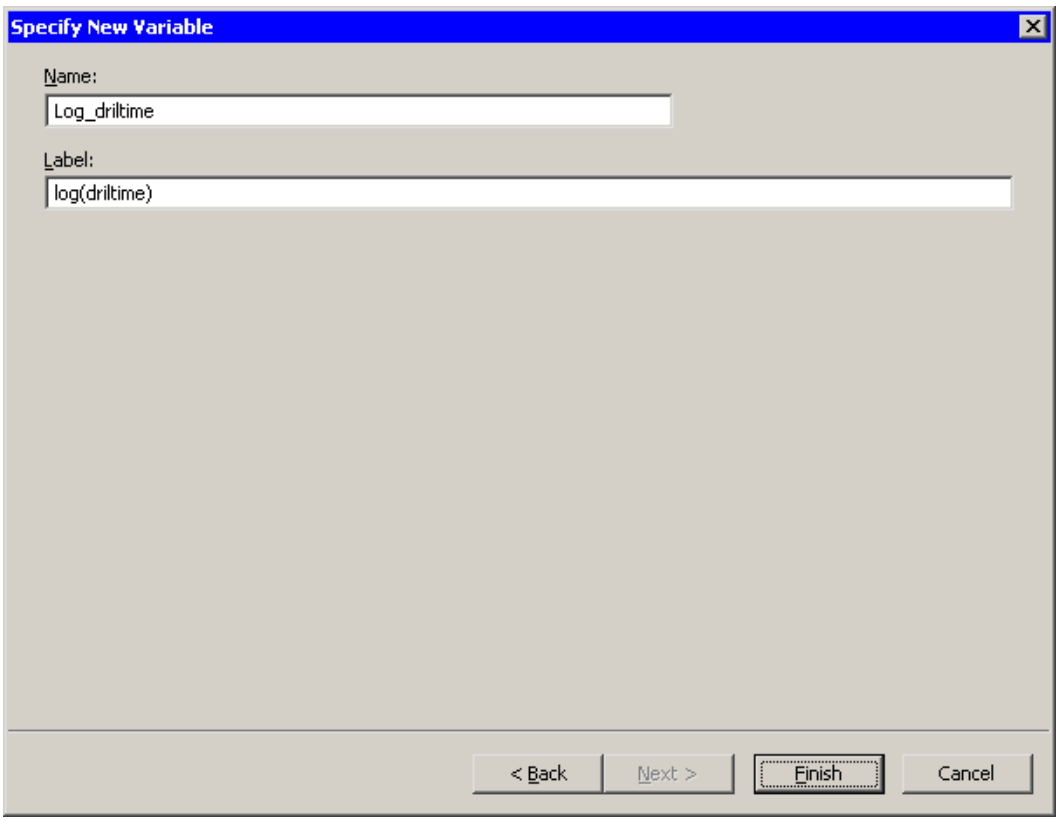
5 Select the driltime variable, and click **Set Y**.

The parameter *a* is an offset that is useful if your variable contains nonpositive values. For these data, you can accept the default value of 0.

6 Click **Next**.

The wizard displays the page shown in Figure 32.4. You can use this page to specify a variable name (and, optionally, a label) for the new variable.

Figure 32.4 Specifying the Variable Name and Label



For this example, you can accept the default variable name.

7 Click **Finish**.

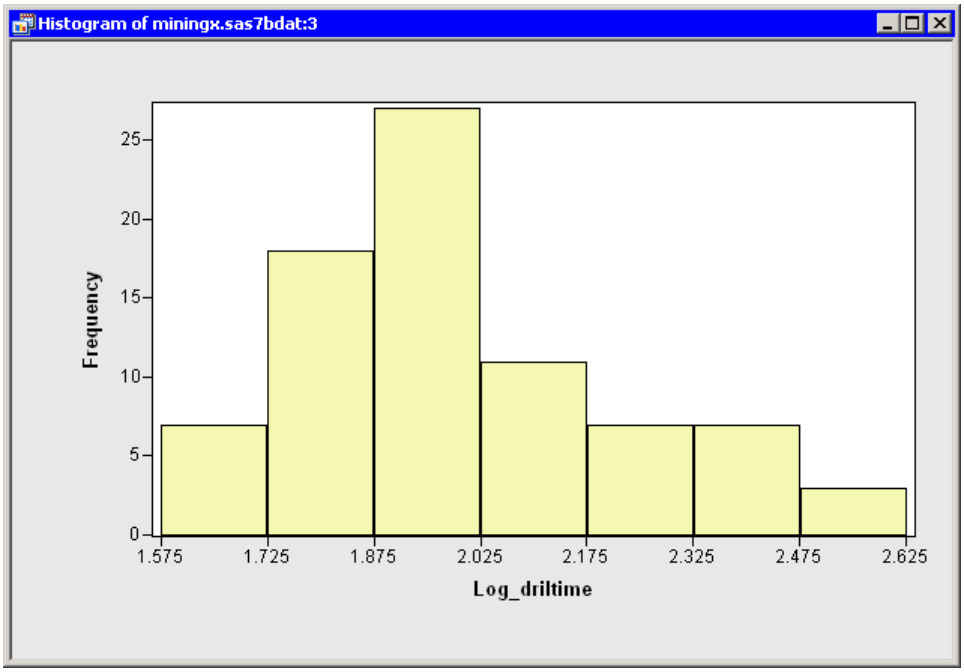
SAS/IML Studio adds the new variable, Log_driltime, as the last variable in the data set. You can horizontally scroll through the data table to see the variable.

To complete this example, you can visualize the distribution of the new variable.

8 Create a histogram of the Log_driltime variable.

The histogram shows improved normality, but the transformed data distribution is still skewed to the right. (See Figure 32.5.)

Figure 32.5 A Histogram of the Transformed Data



Example: Apply a Box-Cox Transformation

This example is a continuation of the previous example. The goal is the same: to normalize the driltime variable in the Miningx data set.

In the previous example, you tried a logarithmic transformation. Unfortunately, it is often not clear which transformation most improves normality. One strategy is to consider a family of transformations, and to select the transformation within the family for which the transformed data are "most normal." The Box-Cox family (Box and Cox 1964) is a family of power transformations that includes the logarithmic transformation as a limiting case:

$$\mathrm{BC}(y;\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases}$$

The parameter λ can be chosen by maximizing a log-likelihood function. For details see the section "Normalizing Transformations" on page 506.

NOTE: The Box-Cox parameter is traditionally denoted by λ, as in the previous formula and in the plot in Figure 32.8. However, the Variable Transformation Wizard uses a as a generic notation for a transformation parameter, as shown in Figure 32.7.

To apply a Box-Cox transformation:

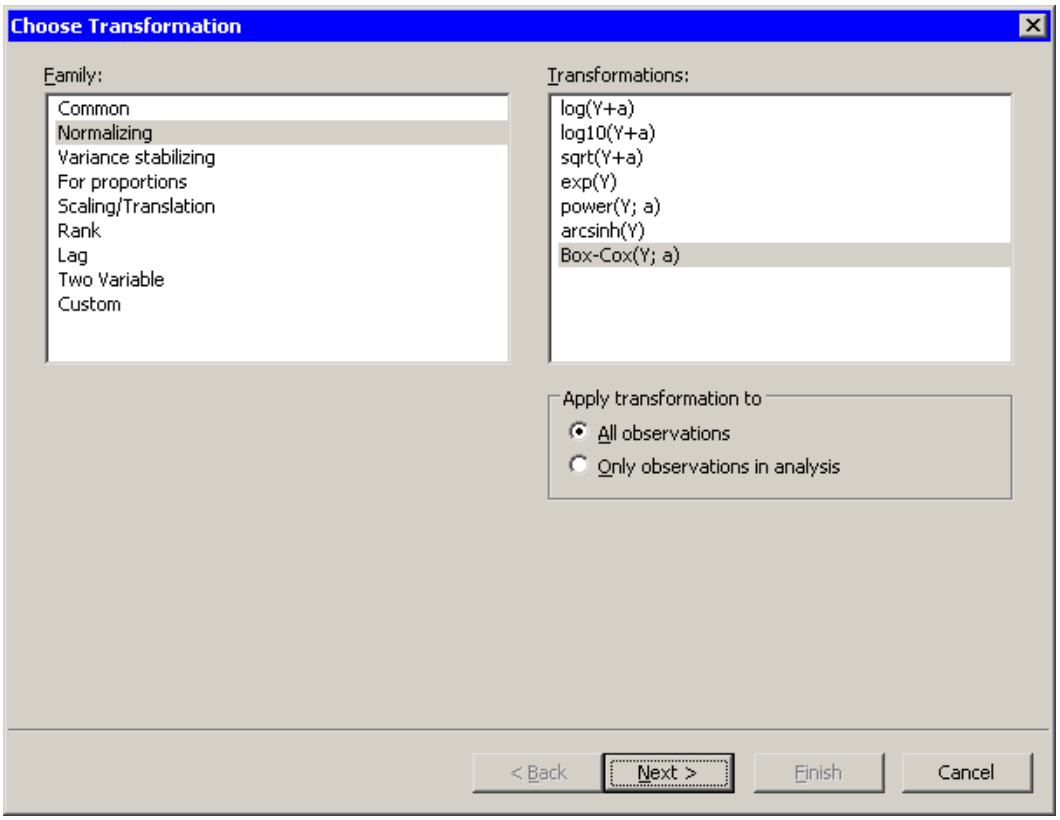
1 Open the Miningx data set, if it is not already open.

2 Select **Analysis ►Variable Transformation** from the main menu.

3 Select **Normalizing** from the **Family** list.

4 Select the **Box-Cox(Y;a)** transformation from the **Transformations** list, as shown in Figure 32.6.

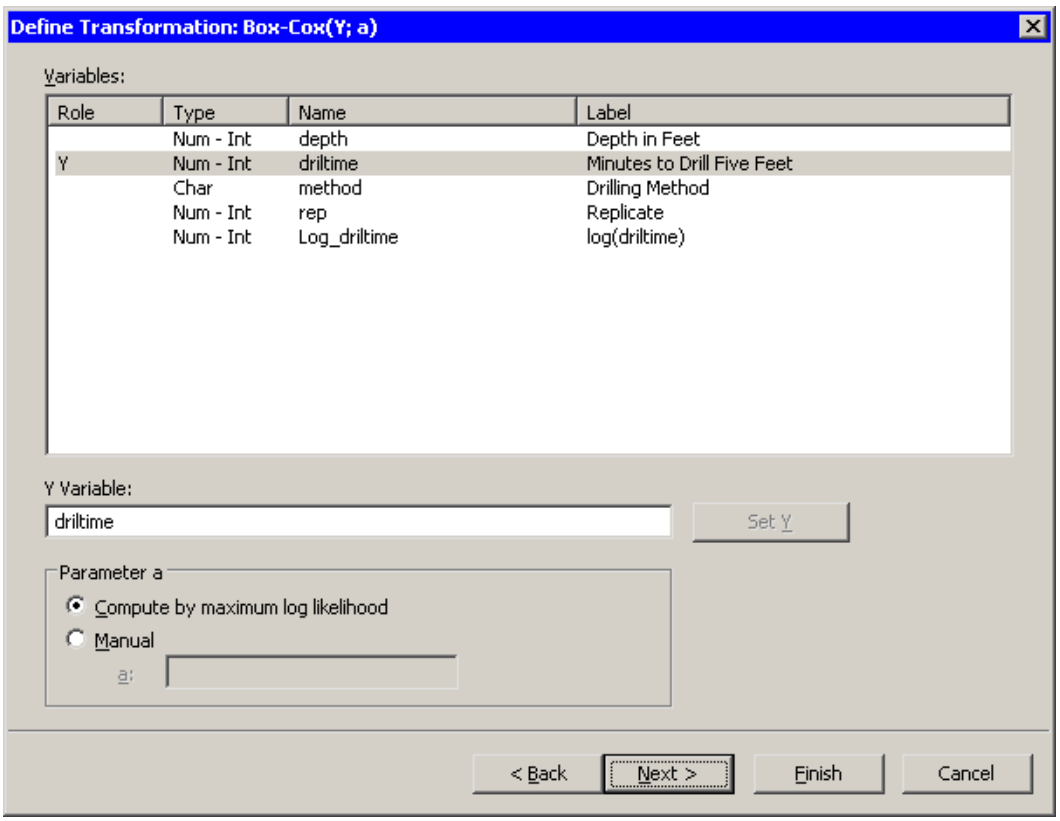
Figure 32.6 Selecting a Box-Cox Transformation



5 Click **Next**.

The wizard displays the page shown in Figure 32.7.

Figure 32.7 Selecting a Variable and Parameters



6 Select the driltime variable, and click **Set Y**.

By default, the Box-Cox parameter is estimated by maximum likelihood estimation. Alternatively, you can manually specify the parameter. For this example, accept the default method.

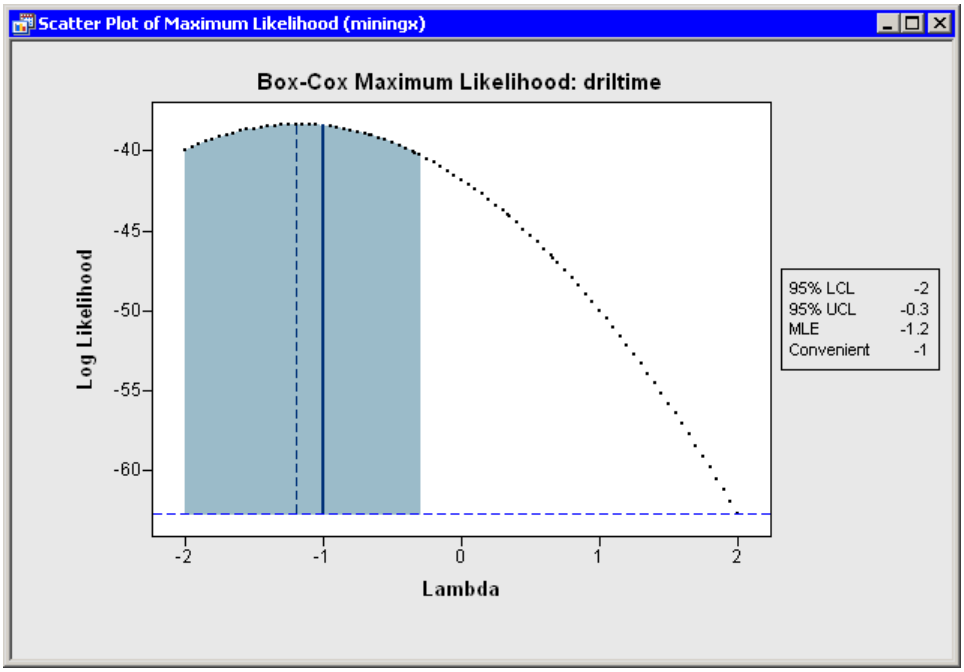
You could proceed to the next page of the wizard if you wanted to change the default name for the new variable. (The default name is BC_driltime.) For this example, accept the default name and skip the last page of the wizard.

7 Click **Finish**.

A graph appears (Figure 32.8) that plots the log-likelihood function as a function of the parameter. An inset gives the lower and upper 95% confidence limits for the maximum log-likelihood estimate, the maximum likelihood estimate (MLE), and a *convenient estimate*. A convenient estimate is a fraction with a small denominator (such as an integer, a half integer, or an integer multiple of $1/3$ or $1/4$) that is within the 95% confidence limits about the MLE. Using a convenient estimate sometimes results in a Box-Cox transformation that is more interpretable in terms of the original variable.

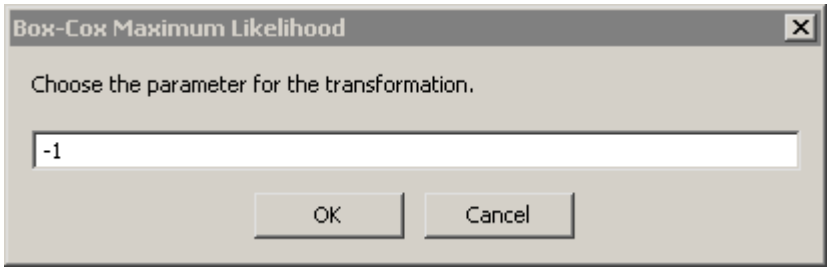
NOTE: If there is no convenient estimate within the 95% confidence limits, then the inset does not include this information.

Figure 32.8 Plot of Log Likelihood



A dialog box (see Figure 32.9) also appears that prompts you for a parameter value to use for the Box-Cox transformation. For this example, you are prompted to accept the convenient estimate of –1, even though the MLE estimate is approximately –1.2.

Figure 32.9 Setting the Box-Cox Parameter



8 Click **OK** to accept the value of –1.

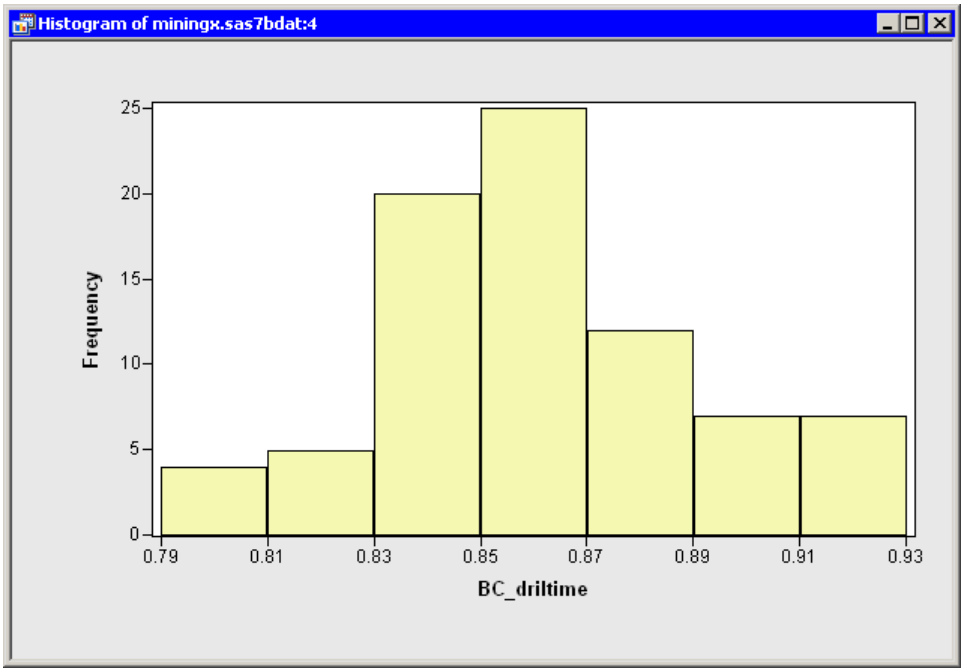
The parameter –1 specifies the Box-Cox transformation as $\mathrm{BC}(y, -1) = 1 - y^{-1}$, which is essentially an inverse transformation followed by a reflection and translation.

To complete this example, you can visualize the distribution of the new variable.

9 Create a histogram of the BC_driltime variable.

The histogram is shown in Figure 32.10. The transformed data show improved normality: the distribution is more symmetric and the tails are not as long.

Figure 32.10 Histogram of the Box-Cox Transformed Data



Common Transformations

Figure 32.11 shows the transformations that are available when you select **Common** from the **Family** list. Equations for these transformations are given in Table 32.1.

Figure 32.11 Common Transformations

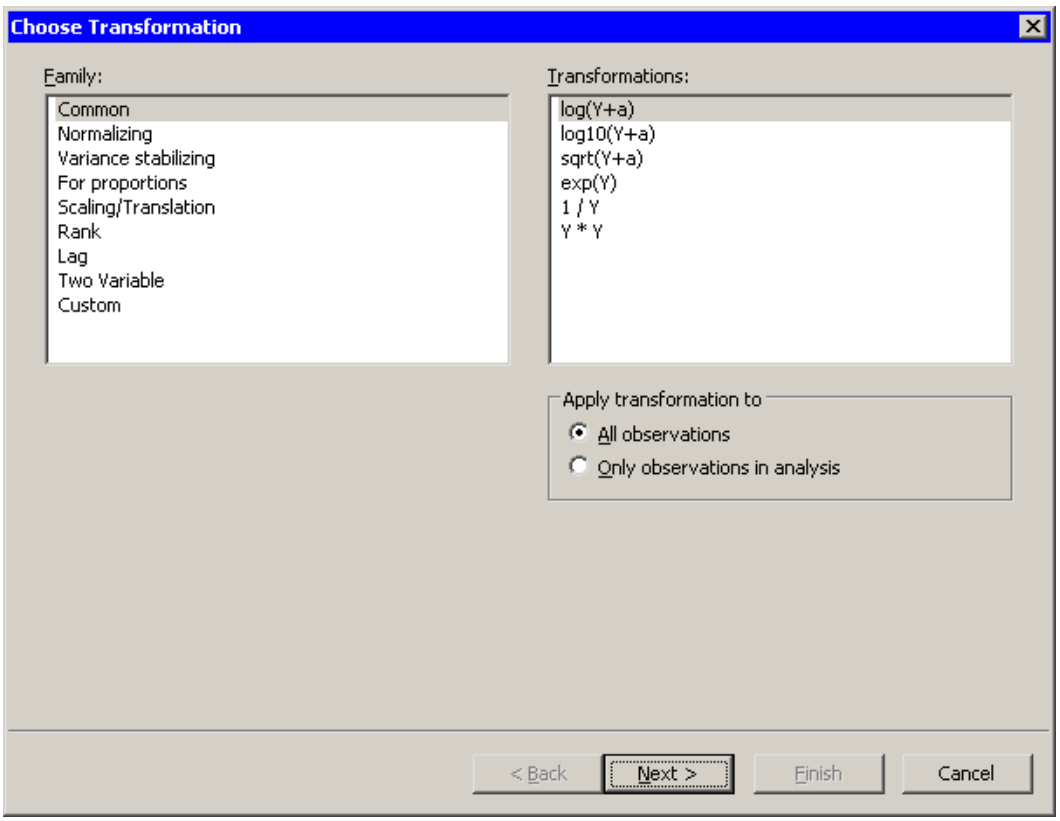


Table 32.1 Description of Common Transformations

Transformation	Default Parameter	Name of New Variable	Equation
log(Y+a)	$a = 0$	Log_Y	$\log(Y + a), \quad Y + a > 0$
log10(Y+a)	$a = 0$	Log10_Y	$\log_{10}(Y + a), \quad Y + a > 0$
sqrt(Y+a)	$a = 0$	Sqrt_Y	$\sqrt{Y + a}, \quad Y + a > 0$
exp(Y)		Exp_Y	$\exp(Y)$
1 / Y		Inv_Y	$1/Y, \quad Y \neq 0$
Y * Y		Squared_Y	Y^2

The logarithmic transformations are often used when the scale of the data range exceeds an order of magnitude. The square root transformation is often used when your data are counts. The inverse transformation is often used to transform waiting times.

Normalizing Transformations

Figure 32.12 shows the transformations that are available when you select **Normalizing** from the **Family** list. These transformations are often used to improve the normality of a variable. Equations for these transformations are given in Table 32.2.

Figure 32.12 Normalizing Transformations

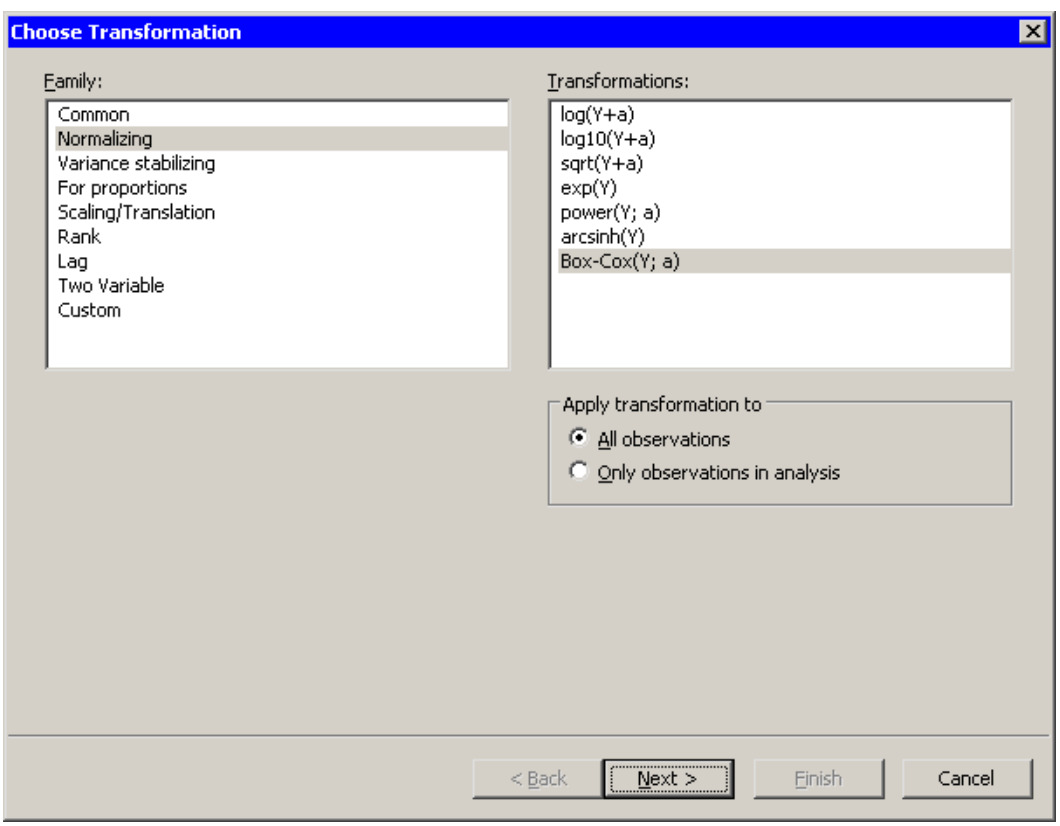


Table 32.2 Description of Normalizing Transformations

Transformation	Default Parameter	Name of New Variable	Equation
log(Y+a)	$a = 0$	Log_Y	$\log(Y + a), \quad Y + a > 0$
log10(Y+a)	$a = 0$	Log10_Y	$\log_{10}(Y + a), \quad Y + a > 0$
sqrt(Y+a)	$a = 0$	Sqrt_Y	$\sqrt{Y + a}, \quad Y + a > 0$
exp(Y)		Exp_Y	$\exp(Y)$
power(Y;a)	$a = 1$	Pow_Y	$Y^a, \quad Y > 0$ if a is not integral
arcsinh(Y)		Arcsinh_Y	$\log(Y + \sqrt{Y^2 + 1})$
Box-Cox(Y;a)	MLE	BC_Y	See text.

The Box-Cox transformation (Box and Cox 1964) is a one-parameter family of power transformations that includes the logarithmic transformation as a limiting case. For $Y > 0$,

$$\mathrm{BC}(y;\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases}$$

You can specify the parameter, λ, for the Box-Cox transformation, but typically you choose a value for λ that maximizes (or nearly maximizes) a log-likelihood function.

SAS/IML Studio plots the log-likelihood function versus the parameter, as shown in Figure 32.8. An inset gives the lower and upper 95% confidence limits for the maximum log-likelihood estimate, the MLE estimate, and a *convenient estimate*. A convenient estimate is a fraction with a small denominator (such as an integer, a half integer, or an integer multiple of $1/3$ or $1/4$) that is within the 95% confidence limits about the MLE. Although the value of the parameter is not bounded, SAS/IML Studio graphs the log-likelihood function restricted to the interval $[-2, 2]$.

A dialog box (see Figure 32.9) also appears that prompts you to enter the parameter value to use for the Box-Cox transformation.

The log-likelihood function for the Box-Cox transformation is defined as follows. Write the normalized Box-Cox transformation, $\boldsymbol{z}$, as

$$\boldsymbol{z}(\lambda; y) = \begin{cases} \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda - 1}} & \text{if } \lambda \neq 0 \\ \dot{y} \log y & \text{if } \lambda = 0 \end{cases}$$

where $\dot{y}$ is the geometric mean of y. Let N be the number of nonmissing values, and define

$$R(\lambda; \boldsymbol{z}) = \boldsymbol{z}'\boldsymbol{z} - (\Sigma z_i)^2 / N$$

The log-likelihood function is (Atkinson 1985, p. 87)

$$L(\lambda; \boldsymbol{z}) = -(N/2) \log(R(\lambda; \boldsymbol{z})/(N - 1))$$

Variance Stabilizing Transformations

Figure 32.13 shows the transformations that are available when you select **Variance stabilizing** from the **Family** list. Variance stabilizing transformations are often used to transform a variable whose variance depends on the value of the variable. For example, the variability of a variable Y might increase as Y increases. Equations for these transformations are given in Table 32.3.

Figure 32.13 Variance Stabilizing Transformations

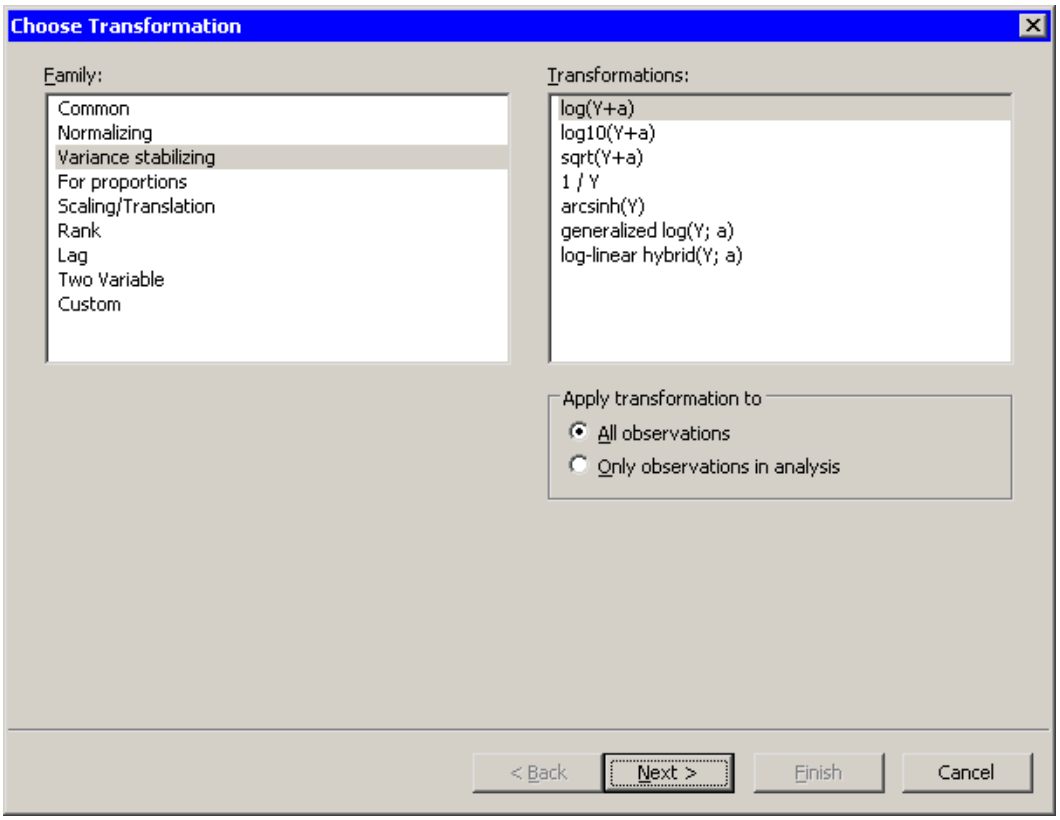


Table 32.3 Description of Variance Stabilizing Transformations

Transformation	Default Parameter	Name of New Variable	Equation
log(Y+a)	$a = 0$	Log_Y	$\log(Y + a), \quad Y + a > 0$
log10(Y+a)	$a = 0$	Log10_Y	$\log_{10}(Y + a), \quad Y + a > 0$
sqrt(Y+a)	$a = 0$	Sqrt_Y	$\sqrt{Y + a}, \quad Y + a > 0$
1 / Y		Inv_Y	$1/Y, \quad Y \neq 0$
arcsinh(Y)		Arcsinh_Y	$\log(Y + \sqrt{Y^2 + 1})$
generalized log(Y;a)	$a = 0$	GLog_Y	$\log((Y + \sqrt{Y^2 + a^2})/2)$
log-linear hybrid(Y;a)	$a = 1$	LogLin_Y	See text.

The log-linear hybrid transformation is defined for $a > 0$ as follows:

$$H(y;a) = \begin{cases} y/a + \log(a) - 1 & \text{if } y < a \\ \log y & \text{if } y \geq a \end{cases}$$

The function is linear for $y < a$, logarithmic for $y > a$, and continuously differentiable.

The generalized log and the log-linear hybrid transformations were introduced in the context of gene-expression microarray data by Rocke and Durbin (2003).

Transformations for Proportion Variables

Figure 32.14 shows the transformations that are available when you select **For proportions** from the **Family** list. These transformations are intended for variables that represent proportions. That is, the Y variable must take values between 0 and 1. You can also use these transformations for percentages if you first divide the percentages by 100.

Chapter 7 of Atkinson (1985) is devoted to transformations of proportions. Equations for these transformations are given in Table 32.4.

Figure 32.14 Transformations for Proportions

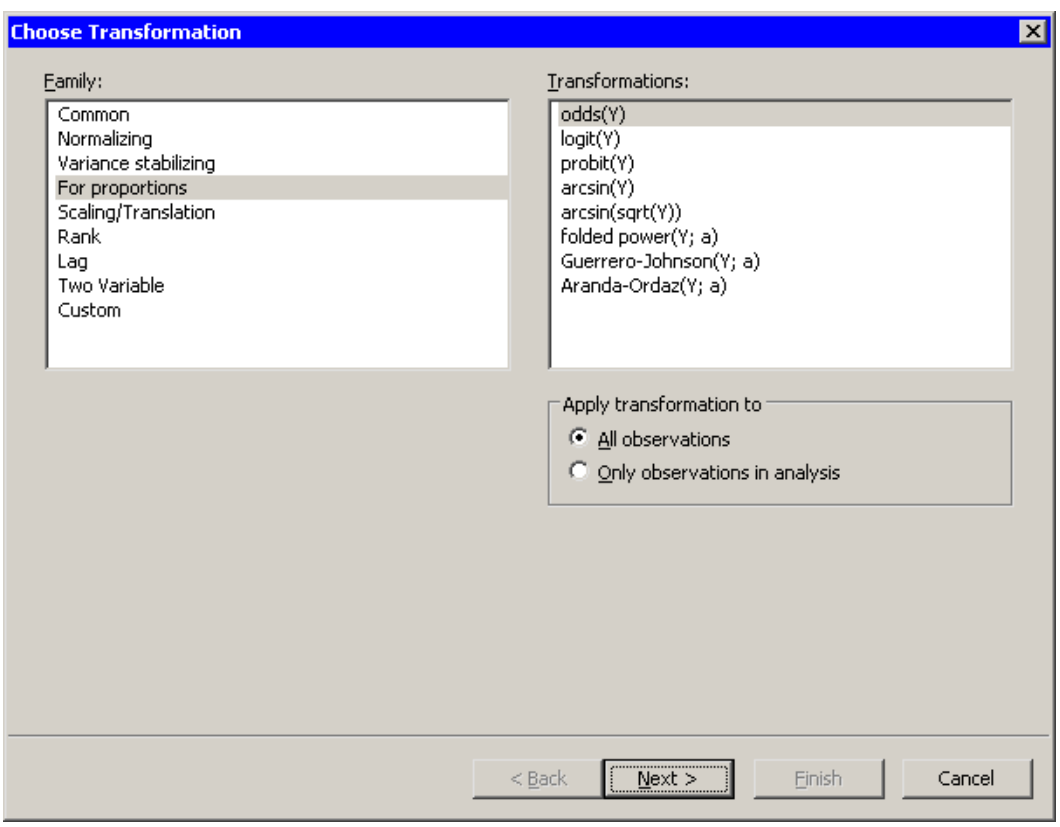


Table 32.4 Description of Transformations for Proportions $Y \in [0, 1)$

Transformation	Default Parameter	Name of New Variable	Equation
odds(Y)		Odds_Y	$Y/(1-Y)$
logit(Y)		Logit_Y	$\log(Y/(1-Y))$
probit(Y)		Probit_Y	$\text{probit}(Y)$
arcsin(Y)		Arcsin_Y	$\arcsin(Y)$
arcsin(sqrt(Y))		Angular_Y	$\arcsin(\sqrt{Y})$
folded power(Y;a)	MLE	FPow_Y	See text.
Guerrero-Johnson(Y;a)	MLE	GJ_Y	See text.
Aranda-Ordaz(Y;a)	MLE	AO_Y	See text.

The probit function is the quantile function of the standard normal distribution.

The last three transformations in the list are similar to the Box-Cox transformation described in the section "Normalizing Transformations" on page 506. The parameter for each transformation is in the unit interval: $a \in [0, 1]$. Typically, you choose a parameter that maximizes (or nearly maximizes) a log-likelihood function.

The log-likelihood function is defined as follows. Let N be the number of nonmissing values, and let $G(\cdot)$ be the geometric mean function. Each transformation has a corresponding normalized transformation $\boldsymbol{z}(\lambda; y)$, to be defined later. Define

$$R(\lambda; \boldsymbol{z}) = \boldsymbol{z}'\boldsymbol{z} - (\Sigma z_i)^2 / N$$

and define the log-likelihood function as

$$L(\lambda; \boldsymbol{z}) = -(N/2) \log(R(\lambda; \boldsymbol{z})/(N-1))$$

The following sections define the normalized transformation for the folded power, Guerrero-Johnson, and Aranda-Ordaz transformations. In each section, $p = y/(1-y)$.

The Folded Power Transformation

The folded power transformation is defined as

$$\mathrm{f}(y; \lambda) = \begin{cases} \frac{y^\lambda - (1-y)^\lambda}{\lambda} & \text{if } \lambda \neq 0 \\ \log(p) & \text{if } \lambda = 0 \end{cases}$$

The normalized folded power transformation is defined as (Atkinson 1985, p. 139)

$$\boldsymbol{z}_\mathrm{f}(\lambda; y) = \begin{cases} \frac{y^\lambda - (1-y)^\lambda}{\lambda G_\mathrm{f}(\lambda)} & \text{if } \lambda \neq 0 \\ \log(p) G(y(1-y)) & \text{if } \lambda = 0 \end{cases}$$

where $G_\mathrm{f}(\lambda) = G(y^{\lambda-1} + (1-y)^{\lambda-1})$. When you select the folded power transformation, a plot of $L(\lambda; \boldsymbol{z}_\mathrm{f})$ appears. You should choose a value close to the MLE value.

The Guerrero-Johnson Transformation

The Guerrero-Johnson transformation is defined as

$$\mathrm{GJ}(y;\lambda) = \begin{cases} \frac{p^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(p) & \text{if } \lambda = 0 \end{cases}$$

The normalized Guerrero-Johnson transformation is defined as (Atkinson 1985, p. 145)

$$z_{\mathrm{GJ}}(\lambda;y) = \begin{cases} \frac{p^{\lambda}-1}{\lambda G_{\mathrm{GJ}}(\lambda)} & \text{if } \lambda \neq 0 \\ \log(p)G(y(1-y)) & \text{if } \lambda = 0 \end{cases}$$

where $G_{\mathrm{GJ}}(\lambda) = G(y^{\lambda-1}/(1-y)^{\lambda+1})$. When you select the Guerrero-Johnson transformation, a plot of $L(\lambda; z_{\mathrm{GJ}})$ appears. You should choose a value close to the MLE value.

The Aranda-Ordaz Transformation

The Aranda-Ordaz transformation is defined as

$$\mathrm{AO}(y;\lambda) = \begin{cases} \frac{2(p^{\lambda}-1)}{\lambda(p^{\lambda}+1)} & \text{if } \lambda \neq 0 \\ \log(p) & \text{if } \lambda = 0 \end{cases}$$

The normalized Aranda-Ordaz transformation is defined as (Atkinson 1985, p. 149)

$$z_{\mathrm{AO}}(\lambda;y) = \begin{cases} \frac{p^{\lambda}-1}{\lambda(p^{\lambda}+1)G_{\mathrm{AO}}(\lambda)} & \text{if } \lambda \neq 0 \\ \log(p)G(y(1-y)) & \text{if } \lambda = 0 \end{cases}$$

where $G_{\mathrm{AO}}(\lambda) = G(2p^{\lambda-1}(1+p)^2/(p^{\lambda}+1)^2)$. When you select the Aranda-Ordaz transformation, a plot of $L(\lambda; z_{\mathrm{AO}})$ appears. You should choose a value close to the MLE value.

Scaling and Translation Transformations

Figure 32.15 shows the transformations that are available when you select **Scaling/Translation** from the **Family** list. These transformations are used to center and scale a variable. Equations for these transformations are given in Table 32.5.

Figure 32.15 Scaling and Translation Transformations

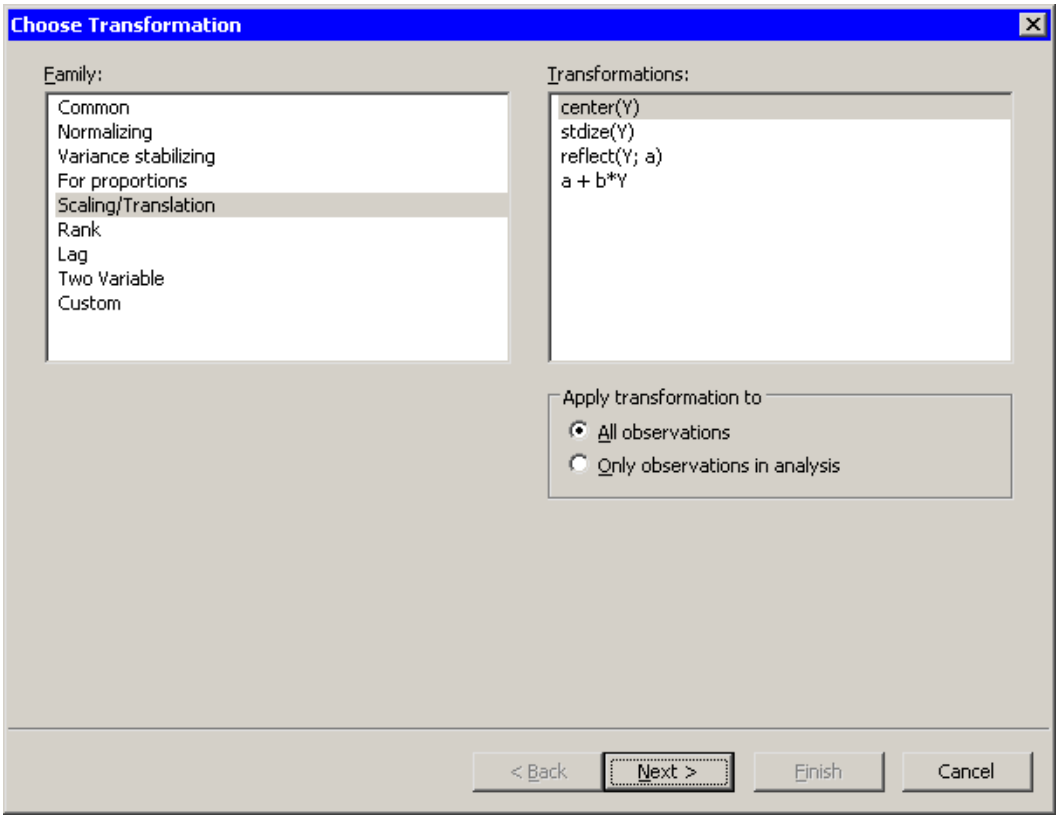


Table 32.5 Description of Scaling and Translation Transformations

Transformation	Default Parameter	Name of New Variable	Equation
center(Y)		Center_Y	$Y - \text{mean}(Y)$
stdize(Y)		Stdize_Y	See text.
reflect(Y;a)	$a = 0$	Reflect_Y	$2a - Y$
a+b*Y	$a = 0, b = 1$	Linear_Y	$a + bY$

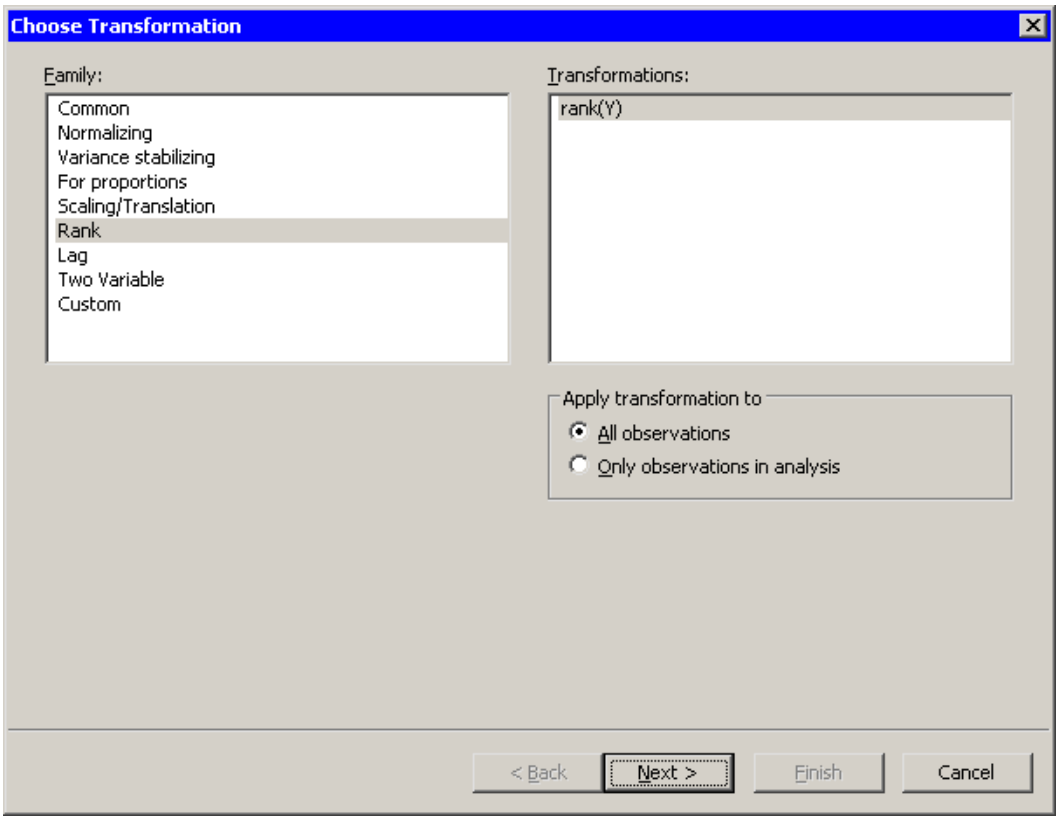
The **stdize(Y)** transformation transforms the data to have zero mean and unit variance.

The **reflect(Y)** transformation reflects the data about the value $Y = a$.

Rank Transformations

Figure 32.16 shows the transformations that are available when you select **Rank** from the **Family** list. The rank transformation of a variable Y is a new variable that contains the ranks of the corresponding values of Y.

Figure 32.16 Rank Transformations



There are actually four different rank functions, depending on the options you select on the second page of the wizard. (See Figure 32.17.) If you select **Assign arbitrarily** as the **Rank of Ties** option, then the SAS/IML RANK function is used to compute ranks. If you select **Assign to average**, then the SAS/IML RANKTIE function is used. This is summarized in Table 32.6.

Figure 32.17 Rank Transformations

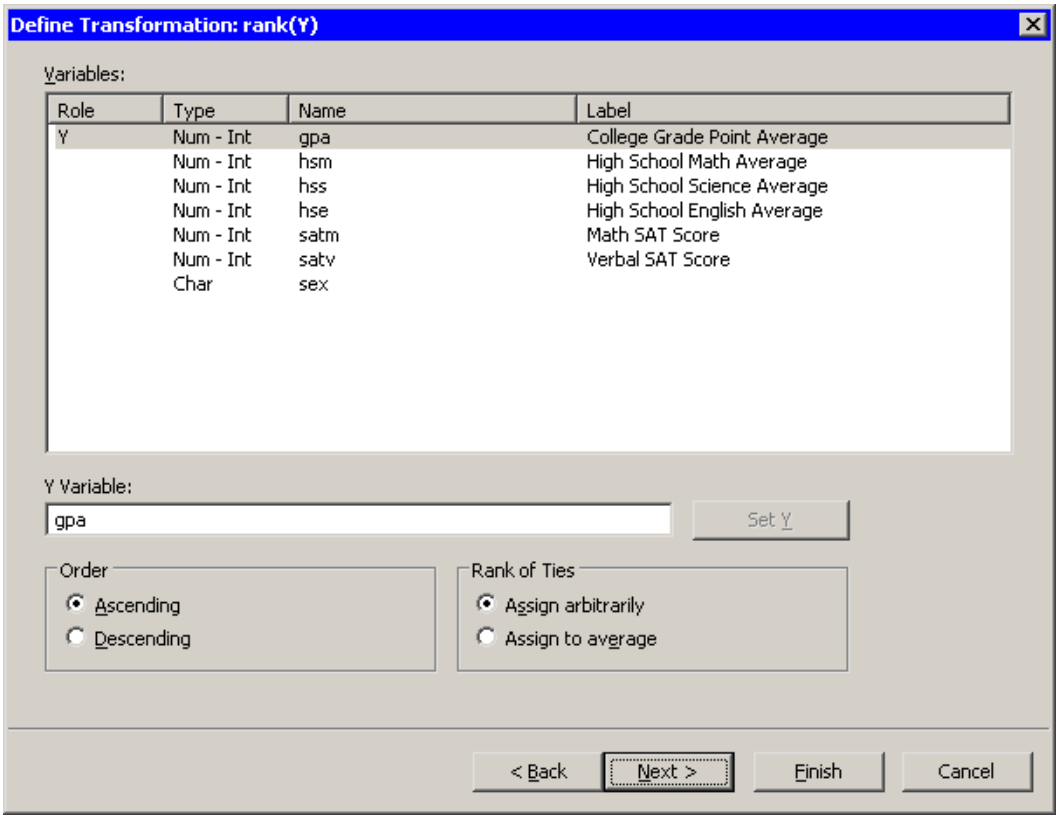


Table 32.6 Description of Rank Transformations

Transformation	Order	Rank of Ties	Name of New Variable	Equation
rank(Y)	Ascending	Arbitrary	Rank_*Y*	rank(Y)
	Descending	Arbitrary	Rank_*Y*	rank(-Y)
	Ascending	Average	Rank_*Y*	ranktie(Y)
	Descending	Average	Rank_*Y*	ranktie(-Y)

Lag Transformations

Figure 32.18 shows the transformations that are available when you select **Lag** from the **Family** list. These transformations are used to compute lagged transformations of a variable's value. Equations for these transformations are given in Table 32.7.

Figure 32.18 Lag Transformations

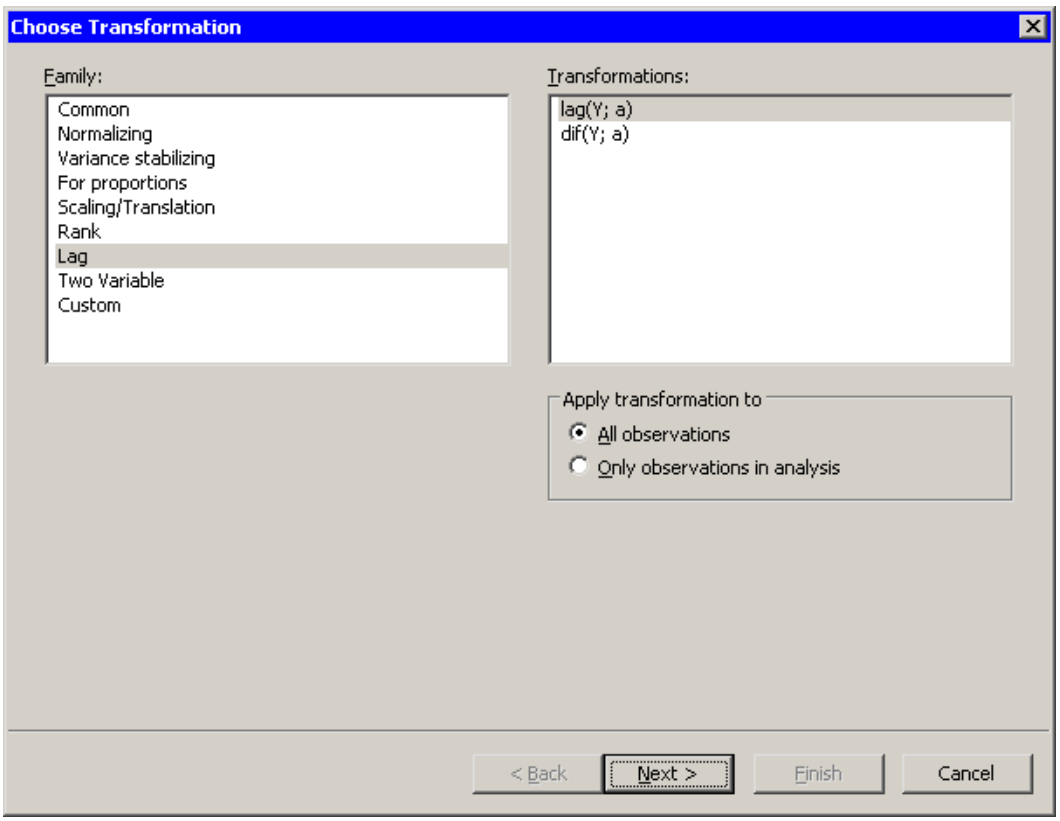


Table 32.7 Description of Lag Transformations

Transformation	**Default** TheadParameter	**Name of New Variable**	**Equation**
lag(Y;a)	$a = 1$	Lag_Y	$\text{lag}(Y, a)$
dif(Y;a)	$a = 1$	Dif_Y	$\text{dif}(Y, a)$

The **lag(Y;a)** transformation creates a new variable whose ith value is equal to Y_{i-a} for $i > a$. For $i \leq a$, the new variable contains missing values. See the documentation for the LAG function in Base SAS software for further details.

The **dif(Y;a)** transformation creates a new variable whose ith value is equal to $Y_i - Y_{i-a}$ for $i > a$. For $i \leq a$, the new variable contains missing values. If either Y_i or Y_{i-a} is missing, then so is their difference. See the documentation for the DIF function in Base SAS software for further details.

If some observations are excluded from analyses and you select **Only observations in analysis**, shown in Figure 32.18, then the lag transformations use only the observations included in analyses. Figure 32.19 presents an example of how these transformations behave when some observations are excluded. In the data table, Y has values 1–5, but observations 3 and 4 are excluded from analyses.

The Lag_Y variable is the result of the **lag(Y;1)** transformation. The third and fourth values are missing because these observations are excluded from analyses. The fifth value of Lag_Y is 2, the previous value of Y that is included in analyses.

The Dif_Y variable is the result of the **dif(Y;1)** transformation. The values are the difference between the first and second columns.

Figure 32.19 Transformations with Excluded Observations

Data Set1

	3	Y		Lag_Y		Dif_Y	
5			Int		Int		Int
1	■ χ^2		1		.		.
2	■ χ^2		2		1		1
3	✖		3		.		.
4	✖		4		.		.
5	■ χ^2		5		2		3
*							

Two-Variable Transformations

Figure 32.20 shows the transformations that are available when you select **Two Variable** from the **Family** list. The two-variable transformations are used to compute a new variable from standard arithmetic operations on two variables. The arithmetic is performed for each observation. Equations for these transformations are given in Table 32.8.

Figure 32.20 Two-Variable Transformations

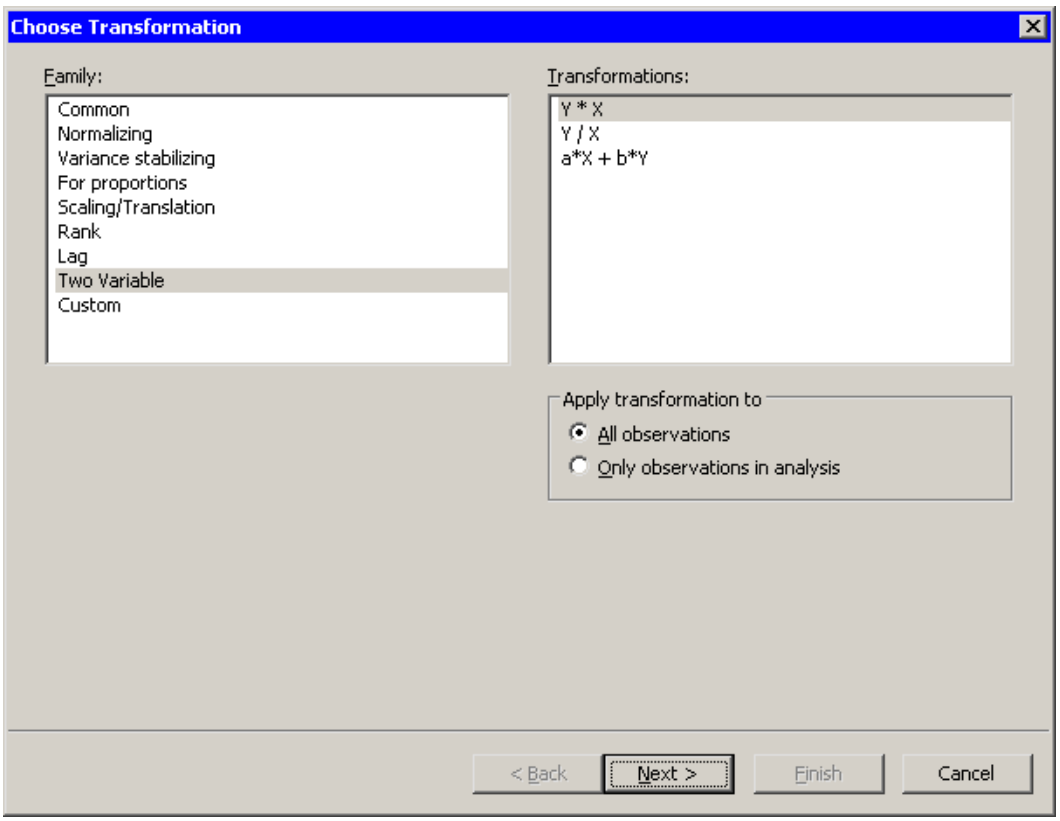


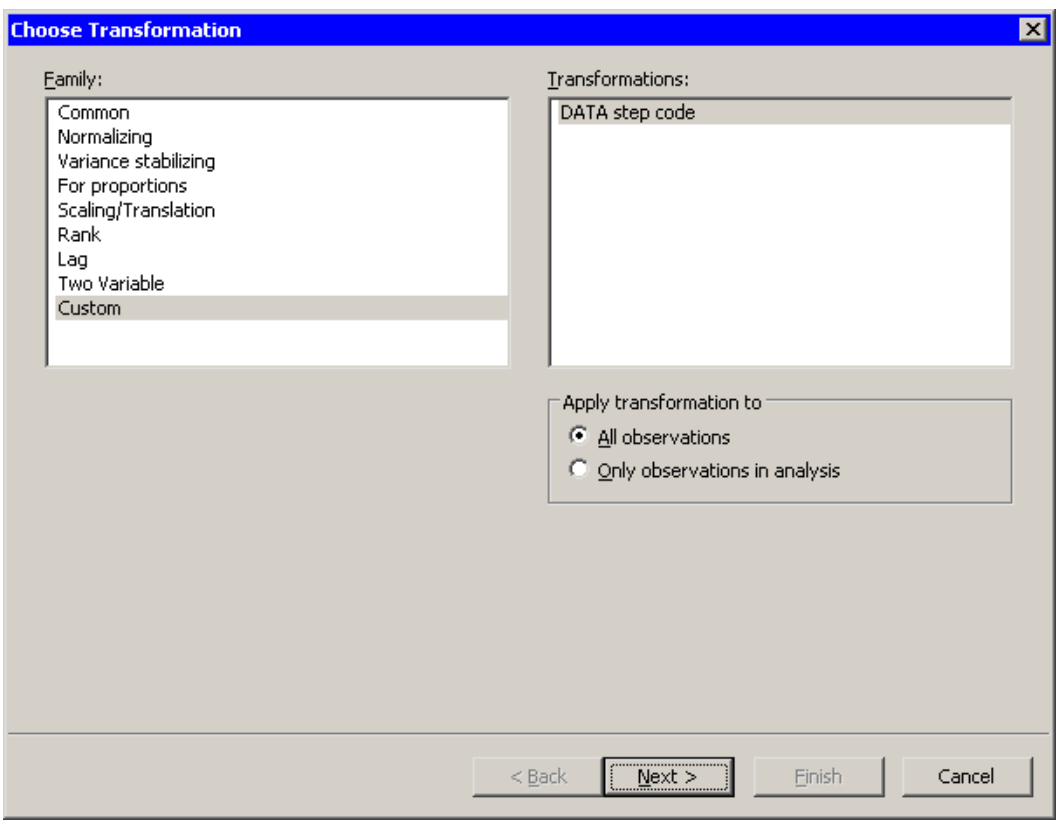
Table 32.8 Description of Two-Variable Transformations

Transformation	Default Parameter	Name of New Variable	Equation
Y*X		Mult_Y_X	YX
Y/X		Div_Y_X	Y/X
a*X+b*Y	$a = 1, b = 1$	Linear_Y_X	$aX + bY$

Custom Transformations

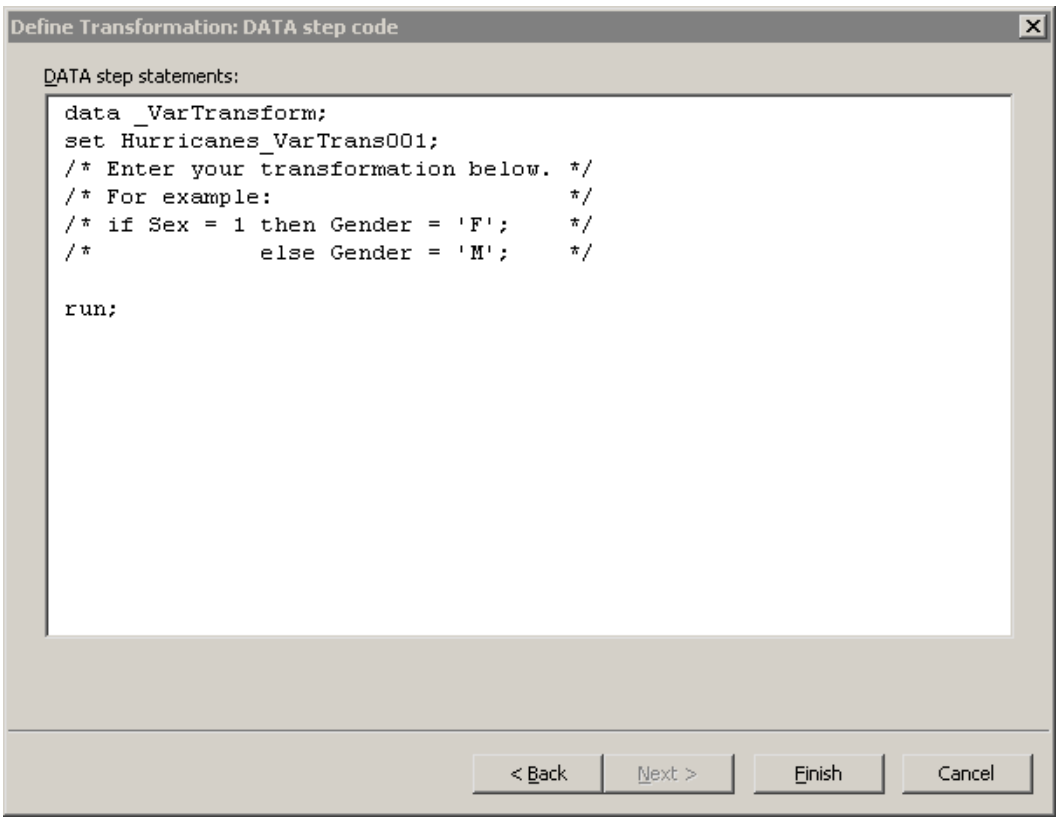
While SAS/IML Studio provides many standard transformations, the most powerful feature of the Variable Transformation Wizard is that you can use the SAS DATA step to create new variables defined by arbitrarily complex formulas. You can define custom transformations after selecting **Custom** from the **Family** list in the Variable Transformation Wizard. (See Figure 32.21.)

Figure 32.21 Selecting a Custom Transformation



The second page of the wizard provides a window where you can enter DATA step statements. The wizard displays the page shown in Figure 32.22.

Figure 32.22 A Window Where You Can Enter DATA Step Statements



You can enter any valid DATA step statements into this window, with the following conditions:

- The statements must begin with a DATA statement.
- The statements must include a SET statement.
- The statements must end with a RUN statement.
- The statements must create an output data set that contains the same number of observations as the data table (or the same number as are included in analyses).

The data set specified in the SET statement is called the *input data set*. The data set specified in the DATA statement is called the *output data set*.

The DATA step template shown in Figure 32.22 satisfies the first three conditions in the previous list. It is up to you to satisfy the last condition by inserting statements before the RUN statement.

The name of the output data set defaults to _VarTransform; the name of the input data set is automatically generated based on the name of your data table. You can accept these default data set names, or you can enter different names.

When you click **Finish**, the following steps occur:

1. SAS/IML Studio scans the text in the window. If the names of any variables in the current data table are found in the text, then these variables are written to the input data set on the SAS server.
2. The DATA step is executed on the server. This creates the output data set.
3. The variables in the output data set are compared with the variables in the input data set.
   a) Any variables in the output data set that are not in the input data set are copied from the server and added to the current data table.
   b) Any variables common to the input and output data sets are compared. If the DATA step changed any values, the new values are copied to the current data table.
4. The input and output data sets are deleted from the server.

Each workspace remembers the last custom transformation you entered. If there is an error in your DATA step statements, you can again select MenuitemAnalysis ►**Variable Transformation** from the main menu and attempt to correct your error. Custom transformations are not remembered between SAS/IML Studio sessions.

Example: Define a Custom Transformation

This example illustrates how to define a custom transformation by using the Variable Transformation Wizard.

NOTE: This example is intended for SAS programmers who are comfortable writing DATA step statements.

Kimball and Mulekar (2004) analyze the *intensification tendency* of Atlantic cyclones. This example is based on their analysis and graphics.

In this example, you use the Variable Transformation Wizard to write DATA step statements that creates a character variable, Tendency, that encodes whether a storm is strengthening or weakening. The Tendency variable is computed by transforming a numeric variable for wind speed. For each observation of each storm, the Tendency variable has the value "Intensifying" when the wind speed is stronger than it was for the previous observation, "Steady" when the wind speed stays the same, and "Weakening" when the wind speed is less than it was for the previous observation.

To transform a variable with a DATA step:

1 Open the Hurricanes data set.

The wind speed is contained in the wind_kts variable. Note that the values of the wind_kts variable are rounded to the nearest 5 knots. The name of each storm is contained in the name variable.

The data are grouped according to storm name, so an algorithm for creating the Tendency variable is as follows.

```
For each named storm:

   Compute the difference between the current wind speed and the
      previous wind speed by using the DIF function in Base SAS software.

   Specify a value for the tendency variable according to whether
      the difference in wind speed is less than zero, exactly
      zero, or greater than zero.
```

If you were to write a DATA step to create the Tendency variable in a data set, you might write statements like the following. The DATA step creates two new variables: a numeric variable called dif_wind_kts and a character variable of length 12 called Tendency. The BY statement is used to loop through the names of cyclones; the NOTSORTED option specifies that the Name variable in the input data set is not sorted in alphabetic order.

```
data WindTendency;
set Hurricanes;
by name notsorted;
length Tendency $12;
dif_wind_kts = dif(wind_kts);
if first.name then do;
   Tendency = "Intensifying";
   dif_wind_kts = .;
end;
else do;
   if dif_wind_kts < 0 then
      Tendency = "Weakening";
   else if dif_wind_kts > 0 then
      Tendency = "Intensifying";
   else
      Tendency = "Steady";
end;
run;
```

The Tendency variable is assigned to “Intensifying” for the first observation of each storm because the storm system was weaker six hours earlier. The dif_wind_kts variable is assigned a missing value for the first observation of each storm because the previous wind speed is unknown.

For subsequent storm observations, the dif_wind_kts variable is assigned the results of the DIF function, which computes the difference between the current and previous values of wind_kts.

Submitting this DATA step in the Variable Transformation Wizard is easy. No changes are required.

2 Select **Analysis ►Variable Transformation** from the main menu.

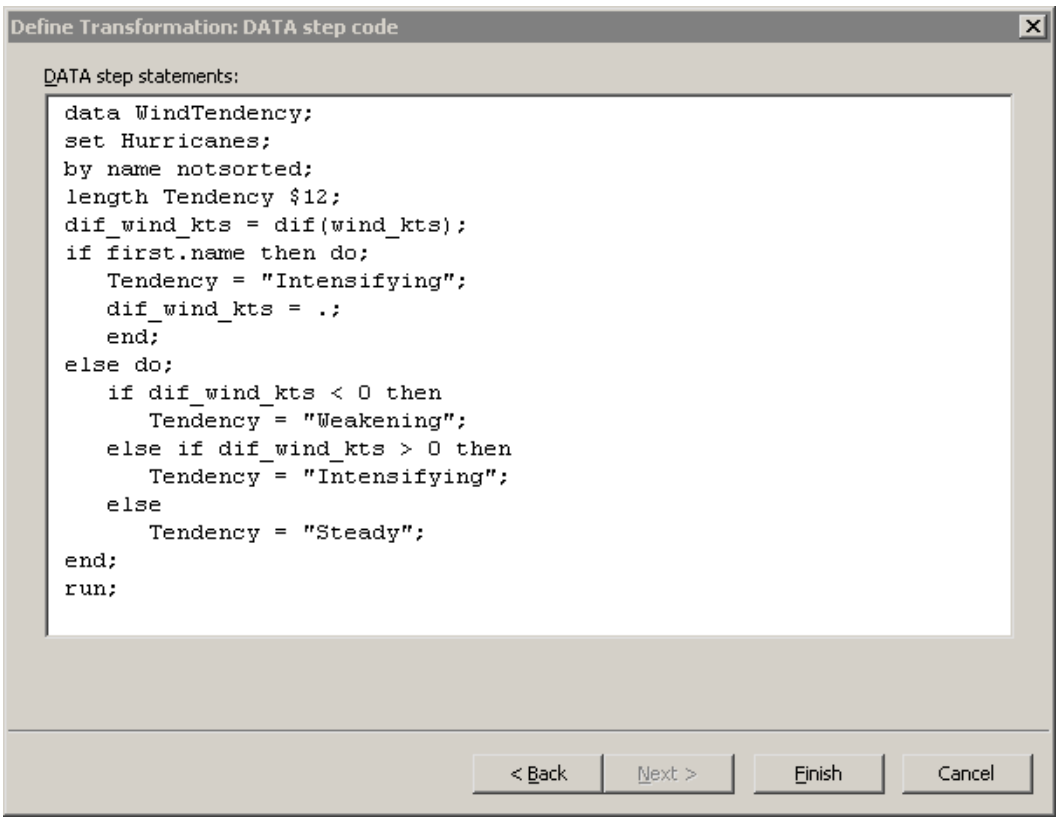
3 Select **Custom** from the **Family** list on the left side of the page, as shown in Figure 32.21.

4 Click **Next**.

The wizard displays the page shown in Figure 32.22.

5 Type the DATA step into the Variable Transformation Wizard, as shown in Figure 32.23.

Figure 32.23 A Custom Transformation



6 Click **Finish**.

SAS/IML Studio scans the contents of the window and determines that the name and wind_kts variables are needed by the DATA step. The input data set, Hurricanes, is created in the WORK library. The input data set contains the name and wind_kts variables.

Next, the DATA step executes on the SAS server. The DATA step creates the output data set, WindTendency, which contains the dif_wind_kts and Tendency variables. The dif_wind_kts and Tendency variables are copied from the output data set to the SAS/IML Studio data table.

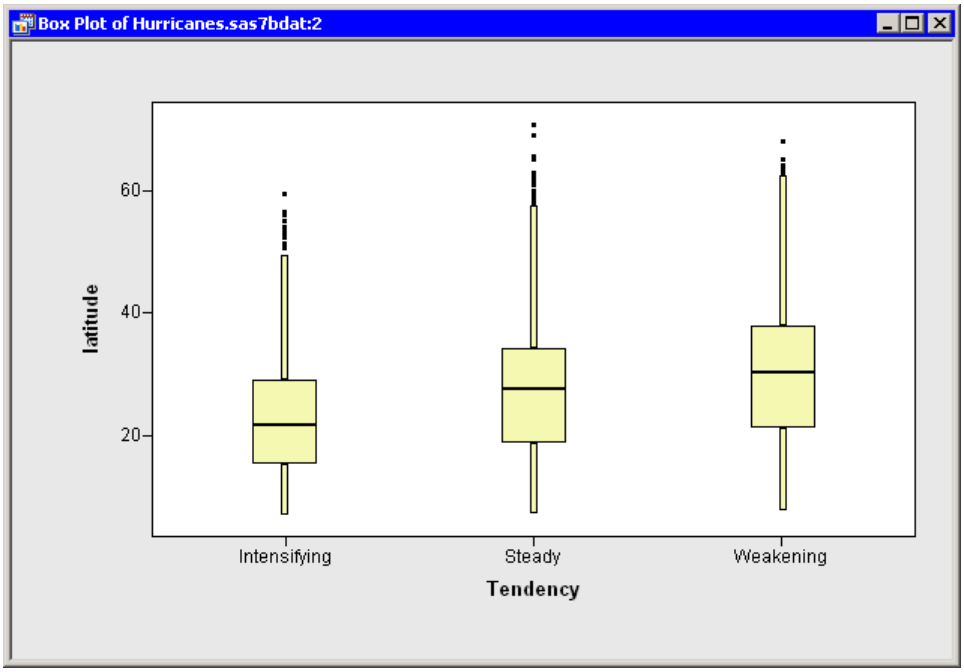
7 Scroll the data table to the extreme right to see the newly created variables.

You can now investigate the relationship between the Tendency variable and other variables of interest.

8 Create a box plot of latitude versus Tendency.

The box plot in Figure 32.24 shows the distribution of latitudes for intensifying, steady, and weakening storms. Intensifying storms tend to occur at more southerly latitudes, whereas weakening storms tend to occur at more northerly latitudes.

Figure 32.24 Latitude Stratified by Intensification Tendency



Applying Normalizing Transformations

This section describes some issues to consider when you are applying normalizing transformations.

Translating Data

The logarithmic and square root transformations are typically most effective at normalizing data that have a minimum value near 1 and have a range that is at most a few orders of magnitude. If a variable consists entirely of large positive values, the transformed data do not show improved normality.

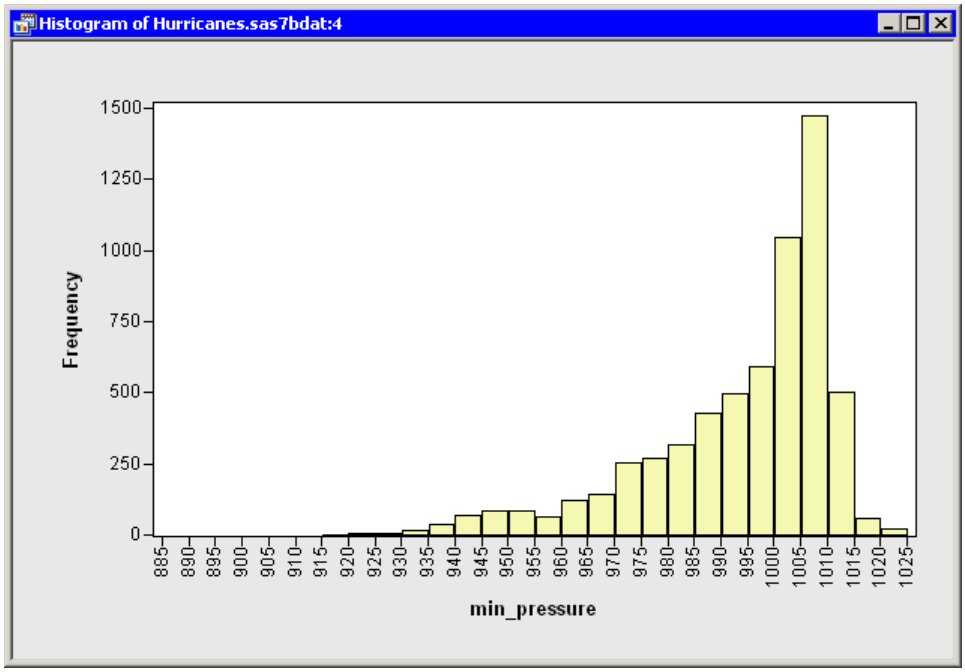
For example, if the minimum value of your data is m, you might want to subtract $m-1$ from your data as a first step so that the new minimum value is 1. You can translate (and scale) data by using the **a+b*Y** transformation in the **Scaling/Translation** family. Alternatively, the square root and logarithmic transformations are defined as **log(Y+a)** and **sqrt(Y+a)**, so you can specify negative values for the a parameter in these transformations. An example of this is presented in the next section.

Skewness

Data can be positively or negatively skewed. The transformations commonly used to improve normality compress the right side of the distribution more than the left side. Consequently, they improve the normality of positively skewed distributions.

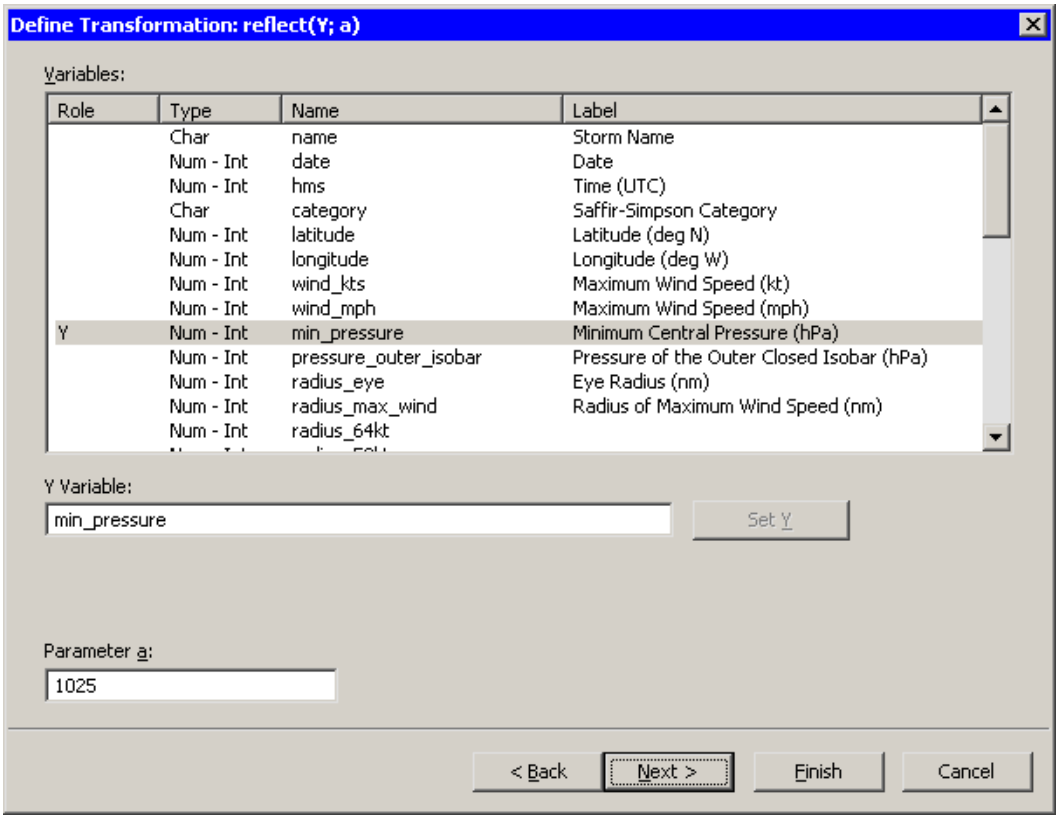
For example, look at the histogram of the min_pressure variable in the Hurricanes data, shown in Figure 32.25. The data are negatively skewed.

Figure 32.25 A Negatively Skewed Variable



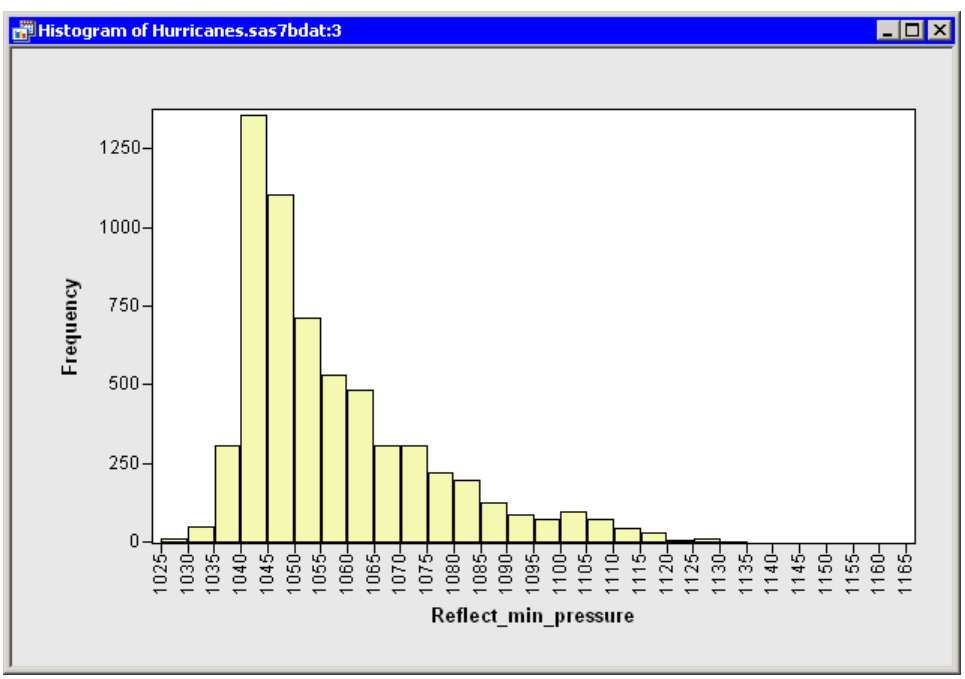
To improve the normality of these data, you first need to reflect the distribution to make it positively skewed. You can reflect data by using the **Reflect(Y;a)** transformation in the **Scaling/Translation** family. Reflecting the data about any point accomplishes the goal of reversing the sign of the skewness. The transformation shown in Figure 32.26 uses $a = 1025$.

Figure 32.26 Defining a Reflection Transformation



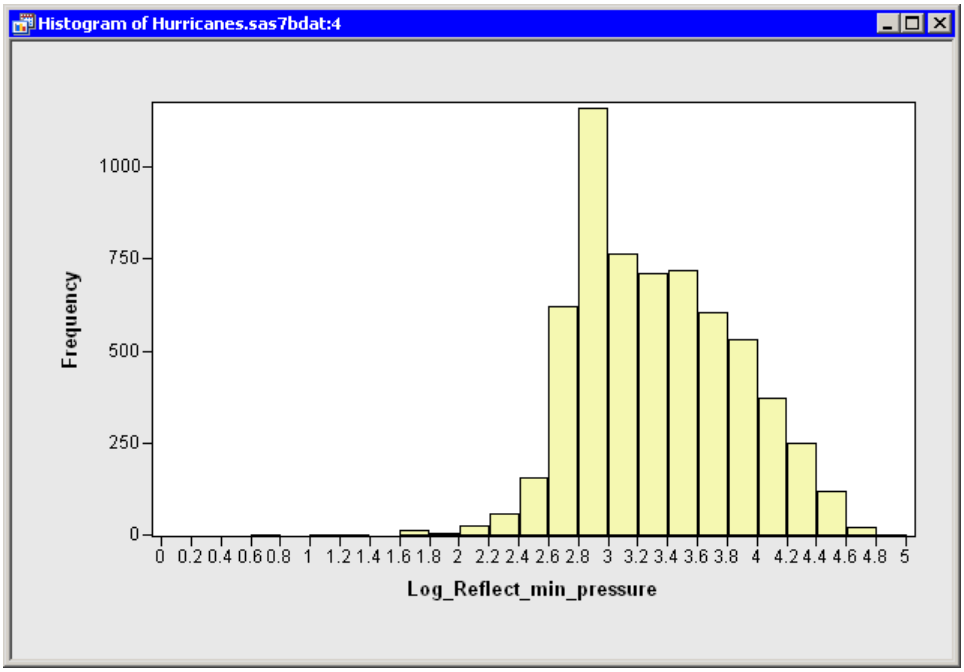
A histogram of the reflected data is shown in Figure 32.27.

Figure 32.27 A Histogram of Reflected Data



You can now apply a normalizing transformation to the Reflect_min_pressure variable. The minimum value of this variable is 1026. As described in the section "Translating Data" on page 523, you can translate and apply a logarithmic transformation in a single step: select the **log(Y+a)** transformation with $a = -1025$. A histogram for the logarithmically transformed variable shows improved normality, but it is still far from normal. (See Figure 32.28.)

Figure 32.28 A Histogram of the Logarithm of Reflected Data



Alternatively, you could transform the Reflect_min_pressure variable in two steps: use the **a+b*Y** transformation with $a = -1025$ and $b = 1$, and then apply a normalizing transformation. This technique is recommended for transformations (such as the Box-Cox family) that do not have a built-in translation parameter.

References

Atkinson, A. C. (1985), *Plots, Transformations, and Regression*, New York: Oxford University Press.

Box, G. E. P. and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, Series B*, 26, 211–234.

Kimball, S. K. and Mulekar, M. S. (2004), "A 15-Year Climatology of North Atlantic Tropical Cyclones, Part I: Size Parameters," *Journal of Climatology*, 17, 3555–3575.

Rocke, D. M. and Durbin, B. P. (2003), "Approximate Variance-Stabilizing Transformations for Gene-Expression Microarray Data," *Bioinformatics*, 19, 966–972.

Chapter 33
Running Custom Analyses

Contents

Overview of Running a Custom Analysis **527**
Example: Run Sample Programs **527**
Example: Run a User Analysis from the Main Menu **528**
Example: Modify the UserAnalysis Module **531**
Example: Create an Action Menu **533**

Overview of Running a Custom Analysis

The programming language in SAS/IML Studio, which is called *IMLPlus*, is an enhanced version of the SAS/IML programming language. The "Plus" part of the name refers to new features that extend the SAS/IML language, including the ability to create and manipulate statistical graphics and to call SAS procedures.

You can write programs in IMLPlus to perform analyses not included in SAS/IML Studio. The analyses can be quite complex. In fact, when you use the SAS/IML Studio GUI to select an analysis from the **Analysis** menu, SAS/IML Studio actually calls an IMLPlus program, so you have already seen examples of what you can accomplish by running IMLPlus programs.

Example: Run Sample Programs

SAS/IML Studio is distributed with samples of programs written in IMLPlus. To open these programs:

1. Select **File ►Open ►File** from the main menu.
2. Click **Go to Installation directory** near the bottom of the dialog box.
3. Double-click the `Programs` folder.
4. Double-click one of the subfolders: `Demos`, `Doc`, or `Samples`. Navigate additional subfolders as necessary.
5. Select a file with an *.sx* extension.
6. Click **Open**.

The `Demos` folder contains advanced programs that demonstrate some of the capabilities of the IMLPlus language. The `Doc` folder contains introductory programs that are described in *SAS/IML Studio for SAS/STAT Users*. The `Samples` folder contains elementary programs that demonstrate how to perform simple tasks in IMLPlus. You can refer to these sample programs as you write more sophisticated programs.

Example: Run a User Analysis from the Main Menu

You can create your own custom analyses by writing an IMLPlus program. An introduction to IMLPlus programming is described in *SAS/IML Studio for SAS/STAT Users* and in the SAS/IML Studio online Help. You can display the online Help by selecting **Help ►Help Topics** from the main menu.

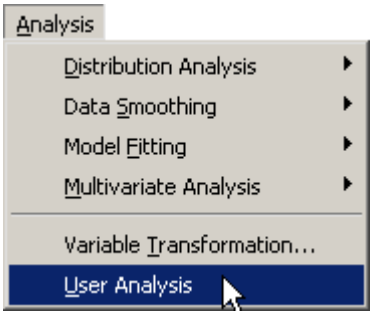
When you select **Analysis ►User Analysis** from the main menu, SAS/IML Studio calls a module called UserAnalysis. SAS/IML Studio distributes a sample UserAnalysis module as an example of the sort of analyses that you can write. You can copy and modify the UserAnalysis module to execute your own IMLPlus programs.

To run the sample UserAnalysis module:

1 Open the Baseball data set.

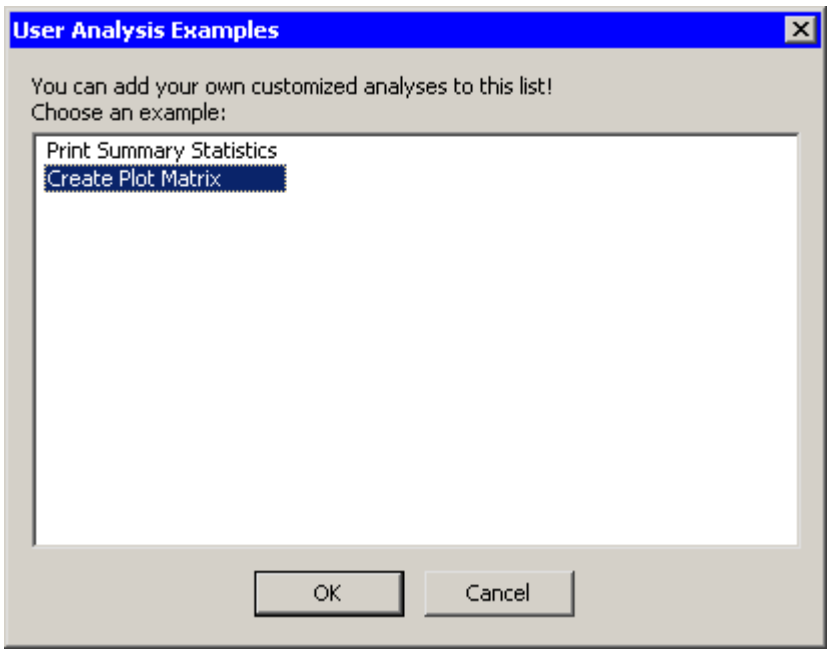
2 Select **Analysis ►User Analysis** from the main menu, as shown in Figure 33.1.

Figure 33.1 Running a User Analysis



The sample UserAnalysis module displays a simple dialog box that contains a list of analyses that you can run on the data. (See Figure 33.2.) The dialog box displays a list of two analyses.

Figure 33.2 Selecting from a List of Analyses

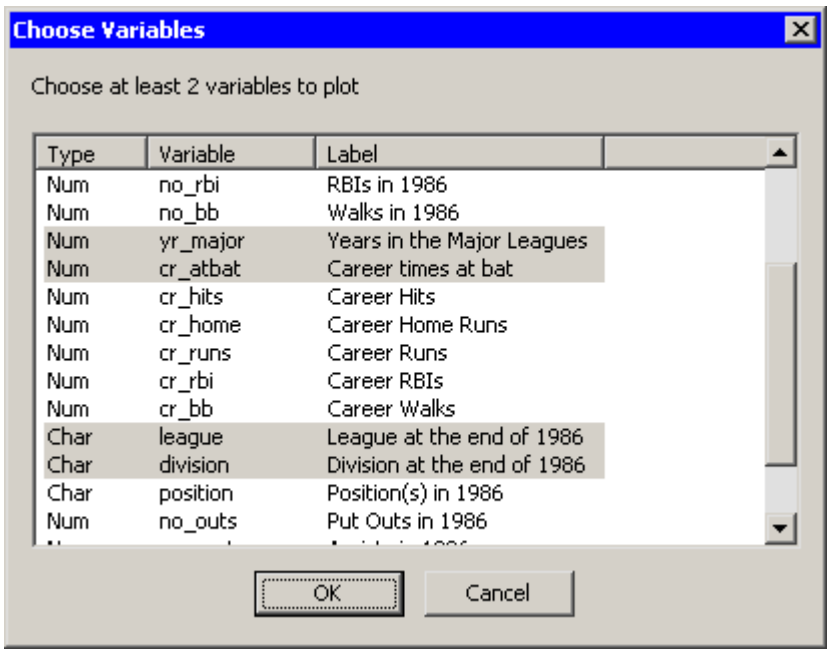


3 Select **Create Plot Matrix** and click **OK**.

The **Create Plot Matrix** analysis demonstrates one way to query information from the person that is running the analysis. In this case, the program prompts you to select several variables to plot. If you select n variables from this list, the variables will be plotted against each other in an $(n-1) \times (n-1)$ lower-triangular array of plots of the pairwise combination of variables.

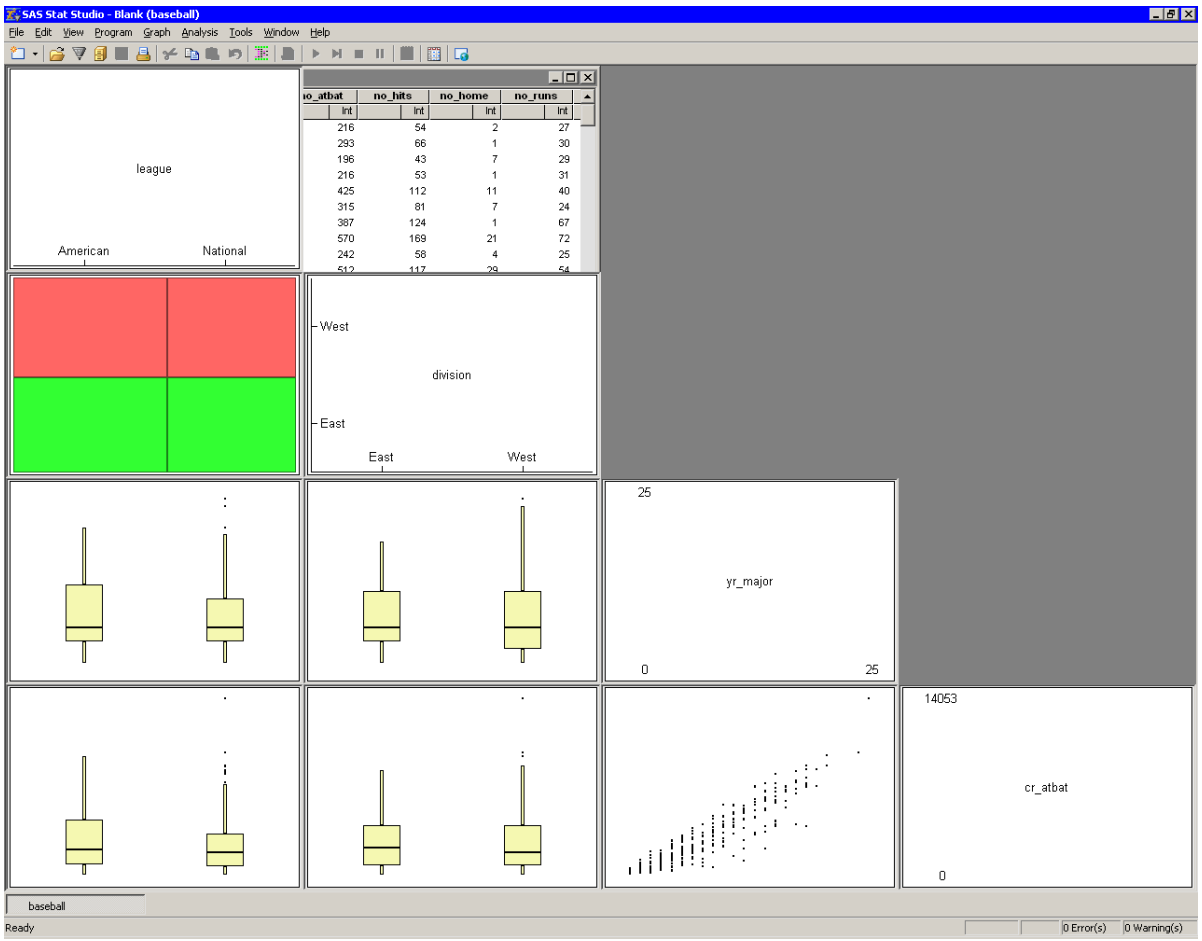
4 Hold down the CTRL key and select yr_major, cr_atbat, league, and division, as shown in Figure 33.3. Click **OK**.

Figure 33.3 Selecting Variables



These four variables are plotted in pairwise combinations, as shown in Figure 33.4. Three different plots are created. Mosaic plots display the relationship between pairs of nominal variables. Box plots are used to plot an interval variable against a nominal variable. Scatter plots display the relationship between pairs of interval variables. Windows along the diagonal display variable names and values of each axis.

Figure 33.4 The Results of the Analysis



Example: Modify the UserAnalysis Module

You can modify the UserAnalysis module to call your own analyses from the **Analysis ►User Analysis** menu item.

To create your own UserAnalysis module:

1. Copy the UserAnalysis.sxs file distributed with SAS/IML Studio to your personal modules directory. The UserAnalysis.sxs file is distributed in the `Modules\System` subdirectory of the SAS/IML Studio installation directory. Your personal modules directory is usually the `Modules` subdirectory of your personal files directory. (See the "The Personal Files Directory" section in Chapter 34 for more information about the personal files directory.)

2. Edit your personal copy of the UserAnalysis.sxs file. Modify the body of the UserAnalysis module so that it performs an analysis of your choosing.

3. Save the file.

4 Select **Program ►Run** to store the module.

5 Open any data set, and select **Analysis ►User Analysis** to run the module.

The UserAnalysis module must take a DataObject variable as its single argument. When you select **Analysis ►User Analysis**, the module is called. The currently active DataObject is used as the argument to the module.

Table 33.1 lists a few of the methods in the DataObject class. You might find these methods useful in writing your analyses. These and other IMLPlus class methods are documented in the SAS/IML Studio online Help, in the "DataObject" section of the "IMLPlus Class Reference" chapter.

Table 33.1 Frequently Used DataObject Methods

Method	Description
AddAnalysisVar	Adds a new variable to the DataObject.
GetNumObs	Returns the number of observations in the DataObject.
GetSelectedObsNumbers	Gets the index of observations selected in the DataObject.
GetSelectedVarNames	Gets the names of variables selected in the DataObject.
GetVarData	Gets the data for a variable in the DataObject.
IsNominal	Returns **true** if the named variable is nominal.
IsNumeric	Returns **true** if the named variable is numeric.
SelectObs	Selects observations in the DataObject.
SetMarkerColor	Sets the color of observation markers.
SetMarkerShape	Sets the shape of observation markers.

For example, you could modify the body of the UserAnalysis module to include the following statements. If you select a nominal variable from a data table and then select **Analysis ►User Analysis**, these statements assign a distinct marker shape to each unique value of the nominal variable. (If there are more unique values than marker shapes, the shapes are reused.) The NCOL, UNIQUE, LOC, and MOD functions are all part of the SAS/IML language, as are the IF and DO statements.

```
start UserAnalysis(DataObject dobj);
 dobj.GetSelectedVarNames(VarName); /* get selected var name */
 if ncol(VarName) = 0 then return;  /* return if no selected variable */
 if dobj.IsNominal(VarName) then do;/* if it is nominal... */
   shapes = MARKER_SQUARE || MARKER_PLUS || MARKER_CIRCLE ||
            MARKER_DIAMOND ||MARKER_X || MARKER_TRIANGLE ||
            MARKER_INVTRIANGLE || MARKER_STAR;
   dobj.GetVarData(VarName, x);    /* get the data */
   ux = unique(x);                 /* find the unique values */
   do i = 1 to ncol(ux);           /* for each unique value... */
      idx = loc(x = ux[<i]);        /* find obs with that value */
      iShape = 1 + mod(i-1, 8);    /* choose next shape (mod 8) */
      /* set the shape of the relevant observations */
      dobj.SetMarkerShape(idx, shapes[iShape]);
   end;
 end;
finish;
store module=UserAnalysis;
```

Example: Create an Action Menu

You can create a custom menu for a plot and associate one or more IMLPlus statements with each item on the menu. Such a menu is referred to as an *action menu*. To display a plot's action menu, press F11 while the plot's window is active. Selecting an item on the menu executes the IMLPlus statements that are associated with that item.

Several previous chapters use action menus to run an analysis on a plot. For example, Figure 12.14 and Figure 18.9 show action menus attached to plots.

Action menus are described in the SAS/IML Studio online Help, in the section called "The Action Menu" in the chapter titled "The Plots."

As an example, the following statements create a histogram and attach an action menu to the plot. When the menu item is selected, the module PrintMean is executed. If the X variable is numeric, then the PrintMean module gets the data that are associated with the X variable of the plot and computes the mean value of these data.

```
x = normal( j(100,1,1) );
declare Histogram plot;
plot = Histogram.Create("Histogram", x);
plot.AppendActionMenuItem("Print Mean", "run PrintMean();");
/* Press F11 in the plot window and select the menu item. */

/* module to run when menu item is selected */
start PrintMean();
   declare Plot plot;
   plot = DataView.GetInitiator();  /* get the active plot */
   plot.GetVars(ROLE_X, VarName);   /* get the X var name */

   declare DataObject dobj;
   dobj = plot.GetDataObject();     /* get the DataObject */
   if dobj.IsNumeric(VarName) then do;
      dobj.GetVarData(VarName, x);  /* get the X values */
      mean = x[:];                  /* compute the mean */
      print "The mean X value is " mean;
   end;
finish;
```

Chapter 34
Configuring the SAS/IML Studio Interface

Contents

Overview of Configuring SAS/IML Studio **535**
SAS/IML Studio Window Types **536**
General Options **538**
Program Editor Options **540**
Output Options **542**
Runtime Options **542**
Server Options **543**
Windows Options **545**
Directory and Search Path Options **548**
Example: Change the Search Path for Data Files 550
The Personal Files Directory **553**
Example: Change the Personal Files Directory 554

Overview of Configuring SAS/IML Studio

You can configure many aspects of SAS/IML Studio, including the following:

- the appearance of GUI items, such as toolbars
- the behavior of the program editor
- the default SAS server
- the default positions of SAS/IML Studio windows, such as graphs, data tables, and output documents
- the directories that SAS/IML Studio searches when trying to locate Java classes, data files, matrices, and modules
- the location of your personal files directory

This chapter describes configuring SAS/IML Studio by using the Options dialog box. You can open the Options dialog box by selecting **Tools ►Options** from the main menu.

If you change options in the Options dialog box, the changes apply to all workspaces. Some changes affect only new workspaces.

SAS/IML Studio Window Types

SAS/IML Studio provides the following different types of windows.

Program Window

A program window is an editor for IMLPlus programs. For each program window, SAS/IML Studio creates a *workspace*. There is always a one-to-one correspondence between a program window and a workspace. It is not possible to have two program windows share a single workspace, nor is it possible to have a single program window connected to more than one workspace.

Program windows provide the following features:

- color coding of IMLPlus keywords, string literals, comments, and constants
- automatic indention of program statements
- drag-and-drop text editing
- positioning the cursor at the source of a program error
- following errors into SAS/IML modules
- multilevel undo and redo
- bookmarks
- finding and replacing text

Error Log Window

An error log window reports warnings and errors from analyses, and programming errors that occur when you run a program.

Output Document Window

An output document window displays output from analyses, output from the SAS/IML PRINT statement, and output from programs that you run. The output window supports the Microsoft rich text format (RTF), so you can paste graphical objects (including SAS/IML Studio graphics) from the Windows clipboard into the output document.

Data View Window

A data view is a generic name for a data table or a plot. Data views that display common data are linked together, which means that selections made in one view are displayed in all views of the same data.

Auxiliary Input Window

An auxiliary input window is a secondary programming window that is linked to the main program window. The IMLPlus PAUSE statement pauses the main program, creates an auxiliary input window, and waits until you click **Resume**. You can use the auxiliary input window as a debugging tool or to prompt for user input.

For example, the following program prompts for user input:

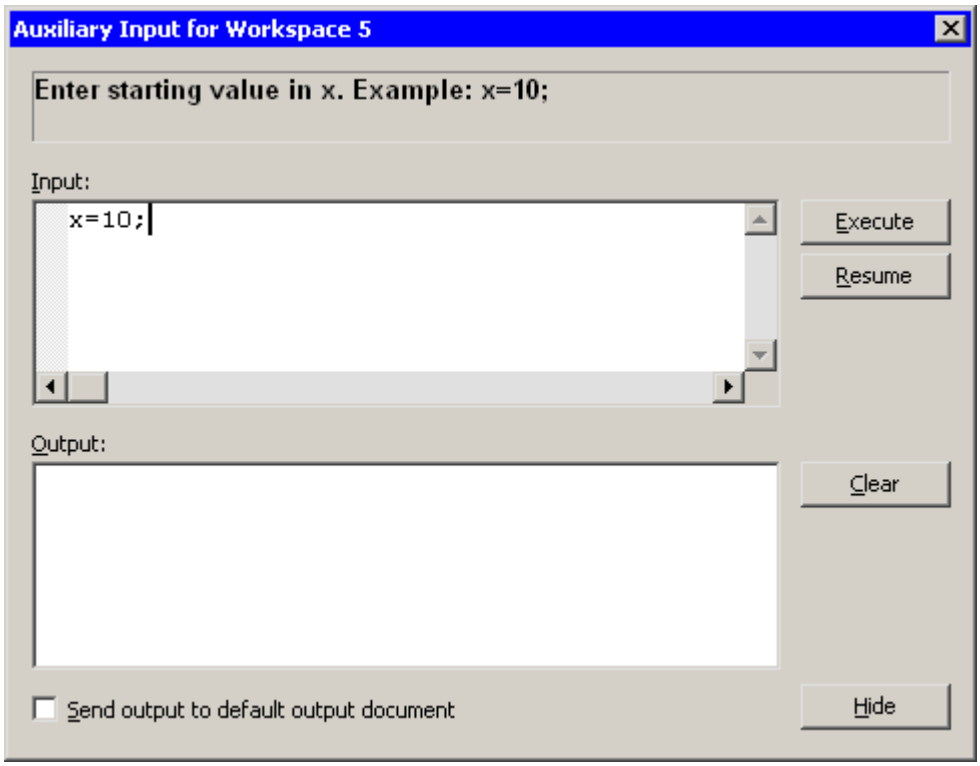
```
pause "Enter starting value in x. Example: x=10;";
do while ( x > 0 );
   print x;
   x = x - 1;
end;
```

When this program is executed, the auxiliary input window appears with the PAUSE statement's message displayed, as shown in Figure 34.1. You can then type the statement

```
x=10;
```

into the **Input** box, and click **Resume**. IMLPlus executes the statement to create the matrix x, and then resumes execution of the main program from the line following the PAUSE statement.

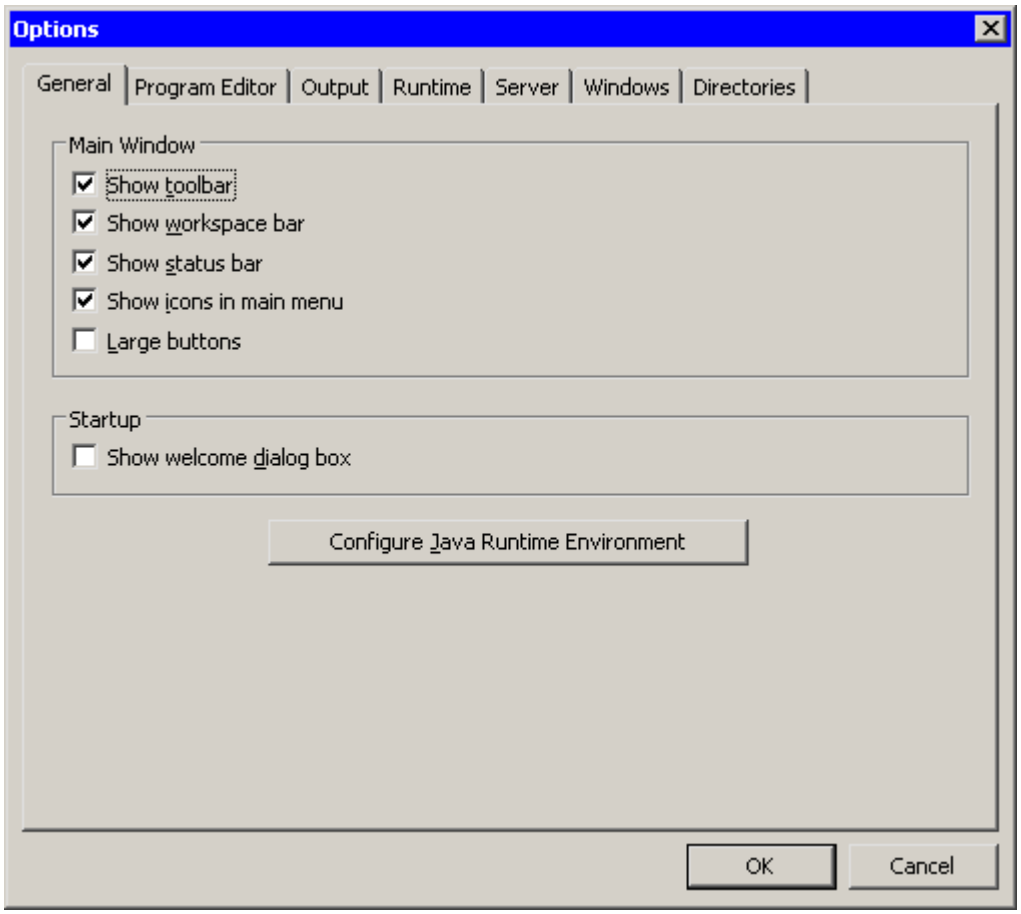
Figure 34.1 The Auxiliary Input Window



General Options

You can configure aspects of the SAS/IML Studio GUI. If you select **Tools ►Options** from the main menu, the Options dialog box appears. By default, the **General** tab is active, as shown in Figure 34.2.

Figure 34.2 The General Tab



The **General** tab has the following fields:

Show toolbar
specifies whether to display the toolbar below the main menu. You can use the toolbar to initiate commonly used actions.

Show workspace bar
specifies whether to display the workspace bar at the bottom of SAS/IML Studio's main window. You can use the workspace bar to switch between different SAS/IML Studio workspaces.

Show status bar
specifies whether to display the status bar at the bottom of SAS/IML Studio's main window. The status bar displays a short message, such as an error message or a description of a menu item.

Show icons in main menu
specifies whether to display icons on the main SAS/IML Studio menus (**File**, **Edit**, **View**, and so on).

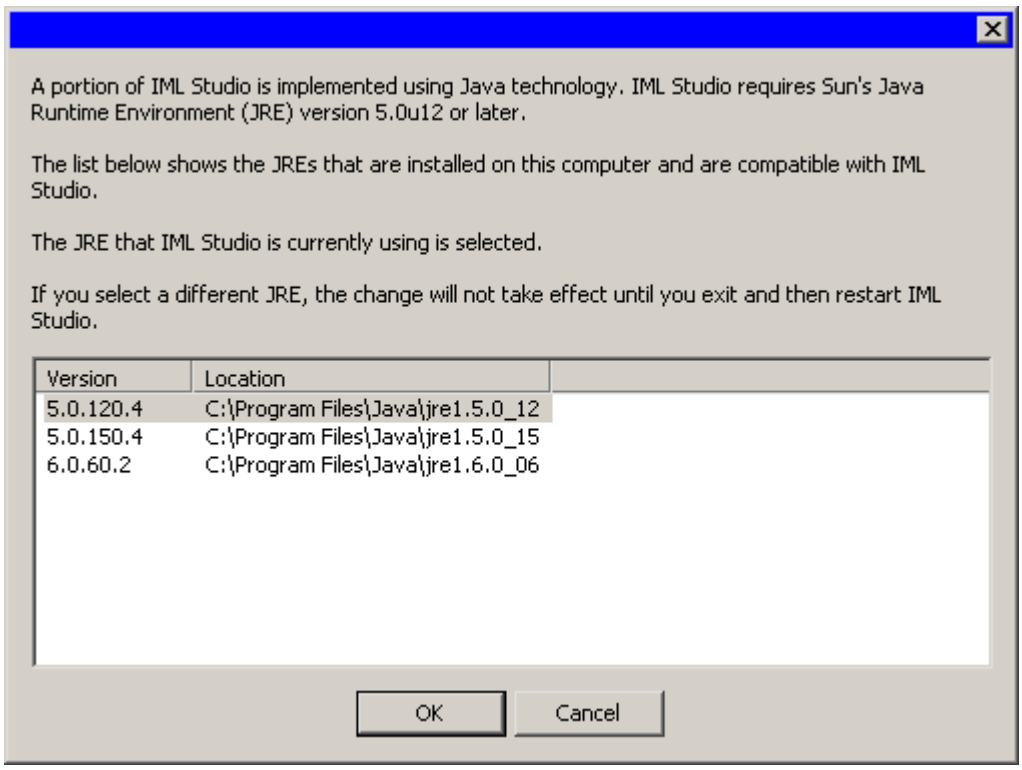
Large buttons
specifies whether to display the buttons on the main toolbar in a large size.

Show welcome dialog box
specifies whether to display the Welcome dialog box, shown in Figure 34.3, when you start SAS/IML Studio.

Configure Java Runtime Environment
enables you to select the Java runtime environment for SAS/IML Studio. If you click this button, the dialog box in Figure 34.4 appears.

Figure 34.3 The Welcome Dialog Box

Figure 34.4 Configuring the Java Runtime Environment

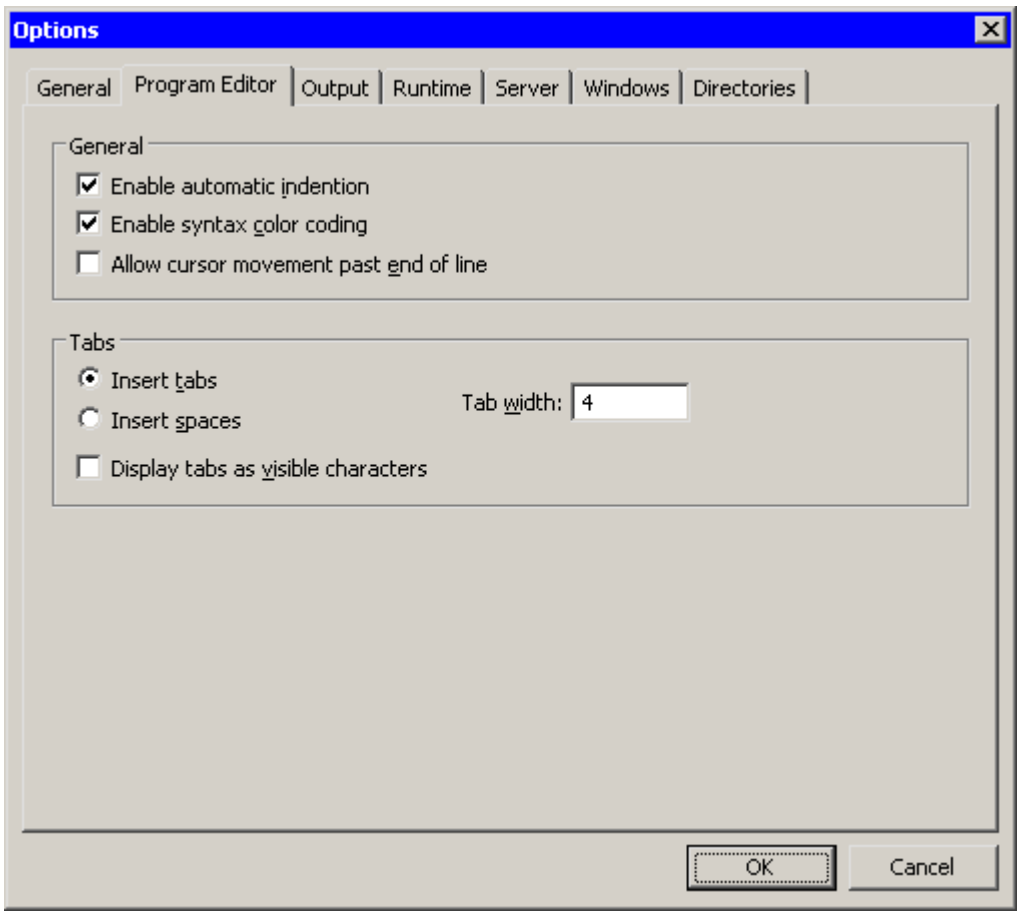


Program Editor Options

You can configure aspects of the SAS/IML Studio program editor. The program editor is used to write and debug IMLPlus programs. IMLPlus programming is described in *SAS/IML Studio for SAS/STAT Users* and in the SAS/IML Studio online Help. You can display the online Help by selecting **Help ►Help Topics** from the main menu.

To display the **Program Editor** tab (shown in Figure 34.5), select **Tools ►Options** from the main menu, and click **Program Editor**.

Figure 34.5 The Program Editor Tab



The **Program Editor** tab has the following fields:

Enable automatic indention
: specifies whether the program editor automatically indents new lines to match the indention of the previous line.

Enable syntax color coding
: specifies whether the program editor color-codes keywords, string literals, comments, and predefined IMLPlus constants.

Allow cursor movement past end of line
specifies whether you can move the cursor beyond an end-of-line character in the program editor.

Insert tabs/spaces
specifies whether the program editor inserts a tab character or space characters when you press the TAB key, and when the program editor automatically indents a line.

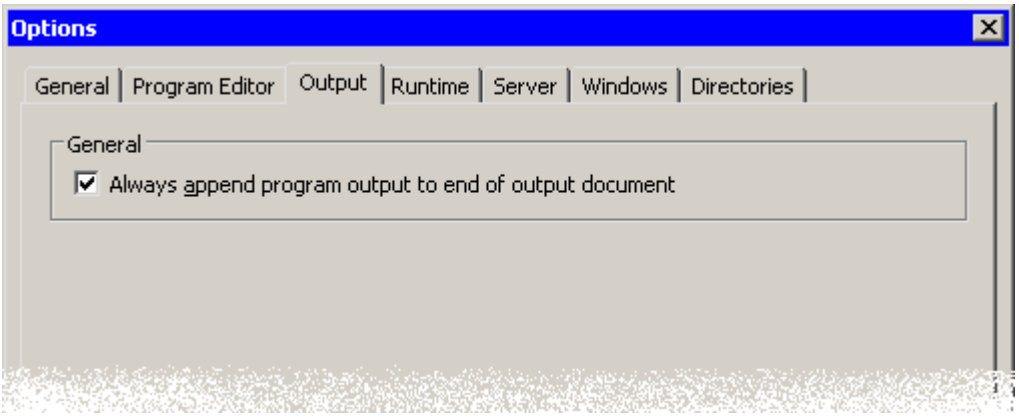
Tab width
specifies the width (in characters) of the tab positions.

Display tabs as visible characters
specifies whether the program editor displays each tab character as the symbol ≫.

Output Options

You can configure aspects of the way that SAS/IML Studio displays output in the output document. Output from SAS procedures is sent to the output document when you run analyses. To display the **Output** tab (shown in Figure 34.6), select **Tools ►Options** from the main menu, and click **Output**.

Figure 34.6 The Output Tab



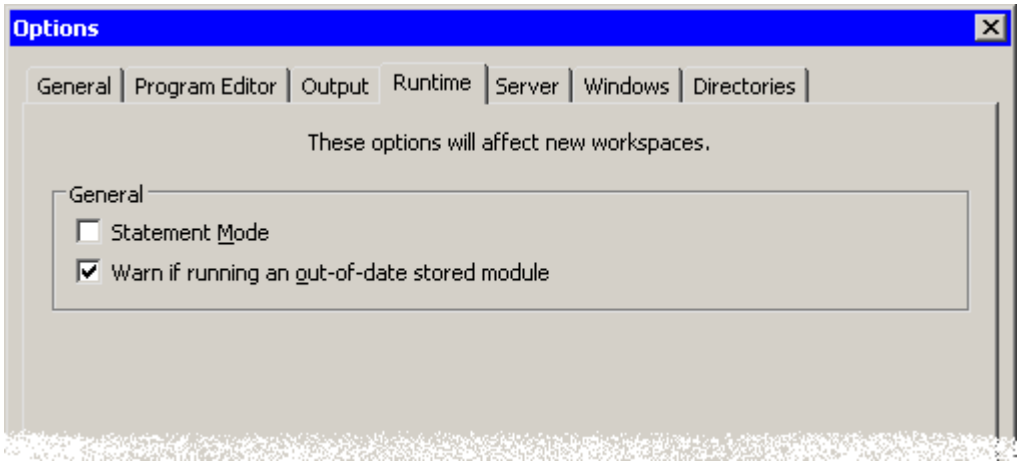
The **Output** tab has a single option. If you select **Always append program output to end of output document**, then output from SAS procedures and IMLPlus programs is always added at the bottom of the output document. If you clear this option, then output is inserted into the output document at the current cursor position.

Runtime Options

You can configure aspects of the SAS/IML Studio programming environment.

To configure default options for new program windows, select **Tools ►Options** from the main menu, and click the **Runtime** tab. This tab is shown in Figure 34.7.

Figure 34.7 The Runtime Tab



The **Runtime** tab has the following fields:

Statement Mode
specifies that the program environment defaults to Statement Mode. For information about Statement Mode, see the SAS/IML Studio online Help. You can display the online Help by selecting **Help ►Help Topics** from the main menu.

Warn if running an out-of-date stored module
specifies that a warning message is printed to the error log when an IMLPlus program executes an out-of-date module. An out-of-date module is one whose source code has been changed since the module was last stored by using the SAS/IML STORE statement.

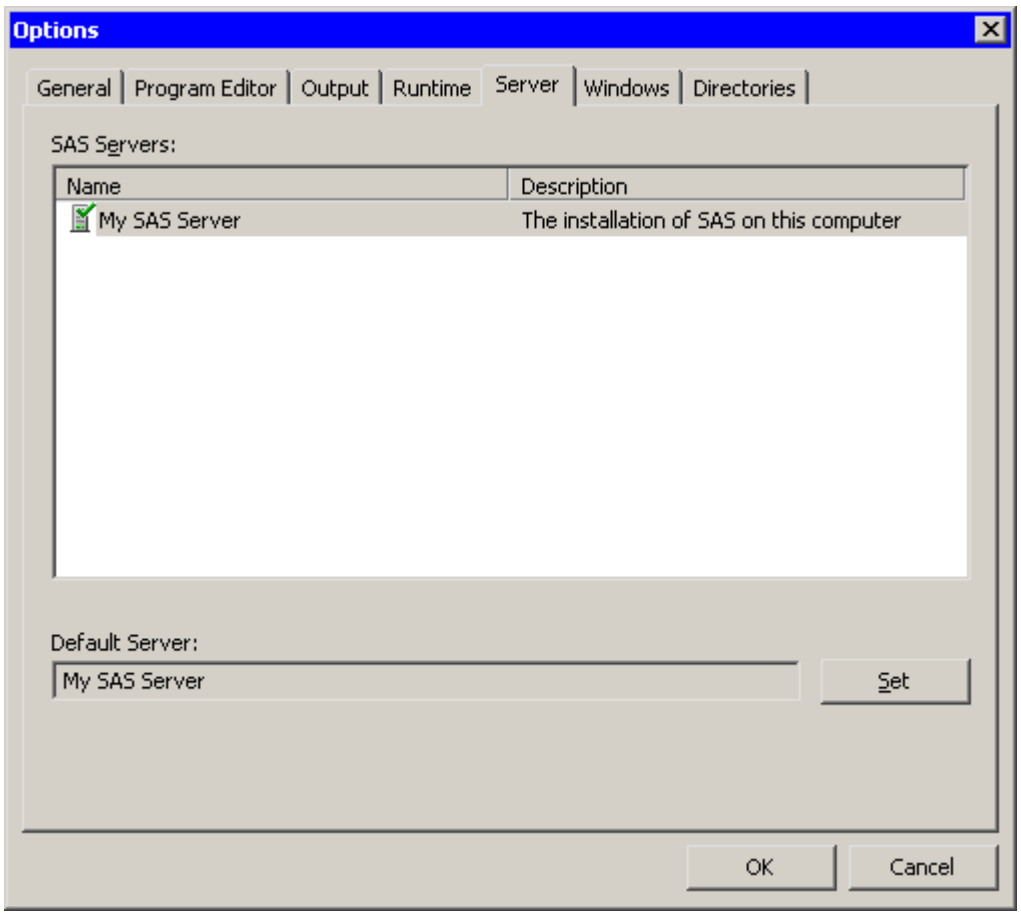
To change these options for a currently open workspace, select **Program ►Configure** from the main menu, and click the **Runtime** tab.

Server Options

The PC that runs SAS/IML Studio is called the *client*. The computer that runs SAS is called the SAS *server*. You can specify the default SAS server that SAS/IML Studio should use. Different workspaces can be connected to different servers.

To configure the default server for new workspaces, select **Tools ►Options** from the main menu, and click the **Server** tab. This tab is shown in Figure 34.8.

Figure 34.8 The Server Tab



The **Server** tab enables you to specify which SAS server is the default server for new workspaces. After you select a server, click the **Set** button.

To change the SAS server for a currently open workspace, select **Program ►Configure** from the main menu, and click the **Server** tab.

Windows Options

You can configure the default positioning of each SAS/IML Studio window type. SAS/IML Studio provides the following types of windows:

- program windows
- error log windows
- output document windows
- data view windows (plots and data tables)

SAS/IML Studio assigns two properties to each type of window. These properties are as follows:

Auto Position
specifies a default window position.

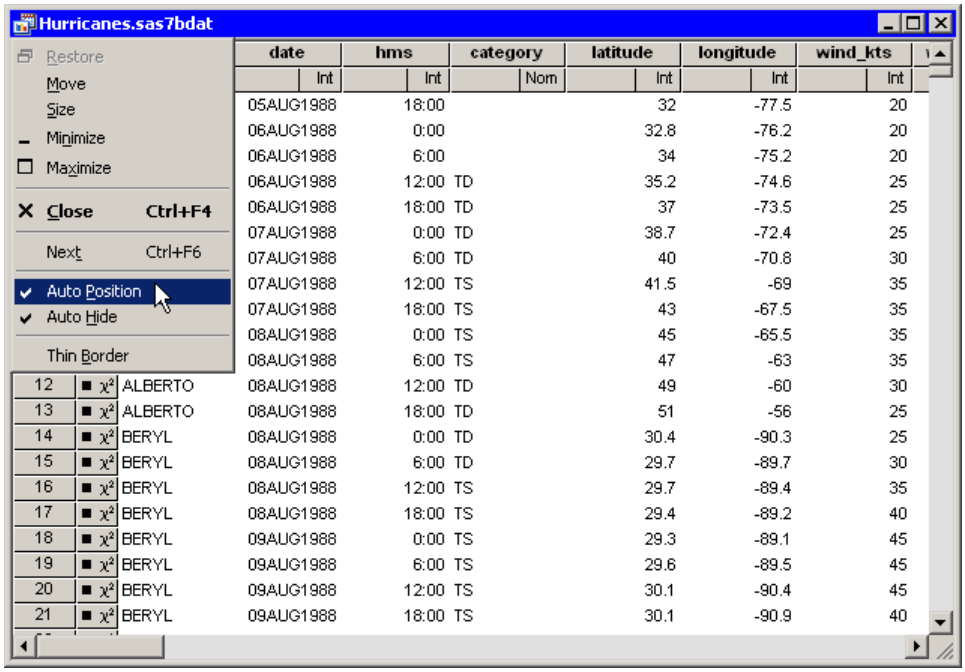
Auto Hide
specifies that the window is hidden when not attached to the active workspace. Error log windows always have the Auto Hide property.

In addition, output document windows have a third property:

Auto Close
specifies that an output document window is automatically closed when the last associated workspace is closed. (Output document windows can be attached to multiple workspaces.)

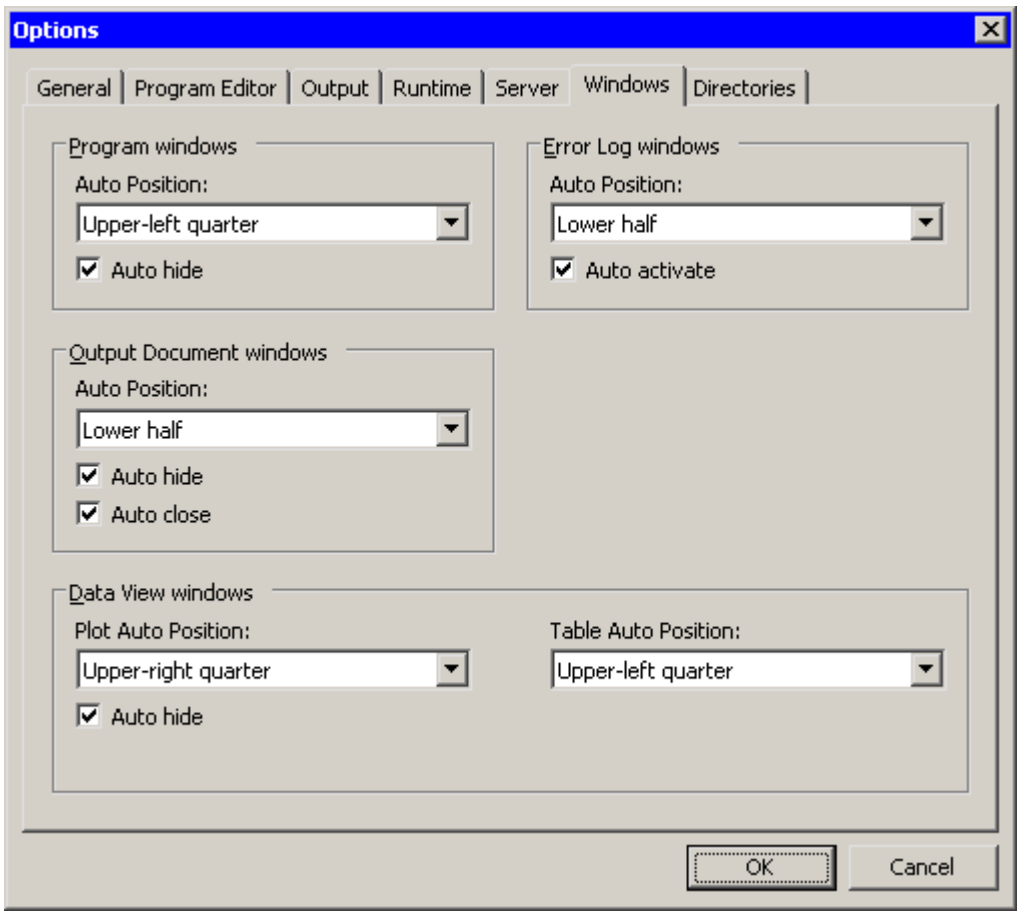
To change a property for an existing window, click the icon in the window's title bar. This displays the **Control** menu, as shown in Figure 34.9. (You can also display the **Control** menu for the active window by pressing ALT+HYPHEN.) You can use this menu to toggle the **Auto Position**, **Auto Hide**, and (for an output document window) **Auto Close** properties.

Figure 34.9 A Control Menu



You can configure the default window properties for each type of window. Select **Tools ►Options** from the main menu, and click the **Windows** tab. This tab is shown in Figure 34.10.

Figure 34.10 The Windows Tab



You can select an **Auto Position** location for all window types. This specifies the default location for a window.

NOTE: If you create multiple windows of the same type (for example, two graphs), then the second window is positioned on top of the first. Move the topmost window to reveal the window hidden beneath. You can select **Auto hide** for all window types except error log windows. A window with this property is hidden when it is not attached to the current workspace. This means that if you change to a different workspace, the windows that are associated with the previous workspace disappear from view. Error log windows always have this property; they appear only in the workspace to which they are attached.

You can select **Auto activate** for error log windows. This causes the error log window to open and become the active window when an error occurs.

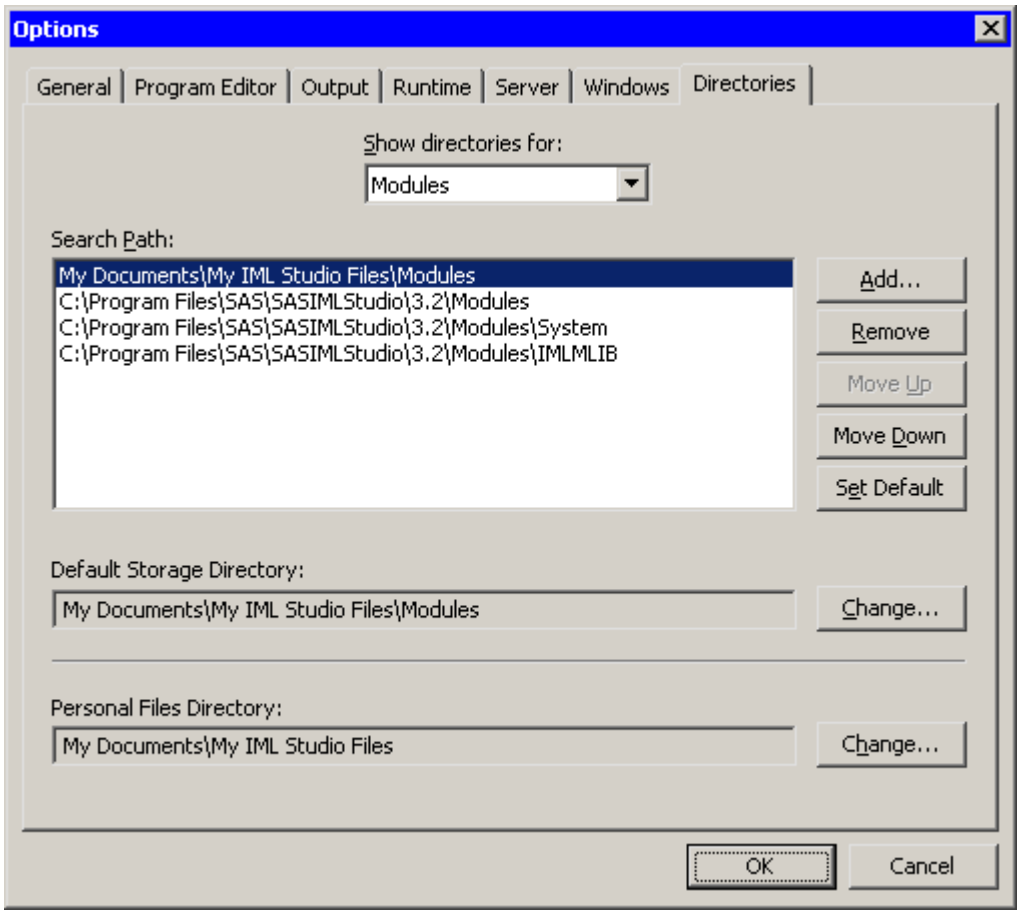
You can select **Auto close** for output document windows. This causes the output document window to close when you close the last workspace to which it is attached. (Output document windows can be attached to multiple workspaces.)

Directory and Search Path Options

You can configure the directories that SAS/IML Studio searches when trying to locate Java classes, data files, matrices, and modules.

Select **Tools ►Options** from the main menu, and click the **Directories** tab. This tab is shown in Figure 34.11.

Figure 34.11 The Directories Tab



The **Directories** tab has the following fields:

Show directories for
specifies the type of file (Java classes, data files, matrices, or modules) that the search path applies to.

Search Path specifies the directories to search when SAS/IML Studio tries to find the indicated type of file. The directories are searched in the order listed.

Add
opens the Browse for Folder dialog box. (See Figure 34.12.) When you select a directory, the directory name is added to the **Search Path** list.

Remove
removes the selected directory from the **Search Path** list.

Move Up
moves the selected directory up one position in the **Search Path** list. The directories in the list are searched in order, from top to bottom, so to reduce search time you should position frequently used directories near the top of the list. **CAUTION:** Do not change the relative positions of the four standard entries.

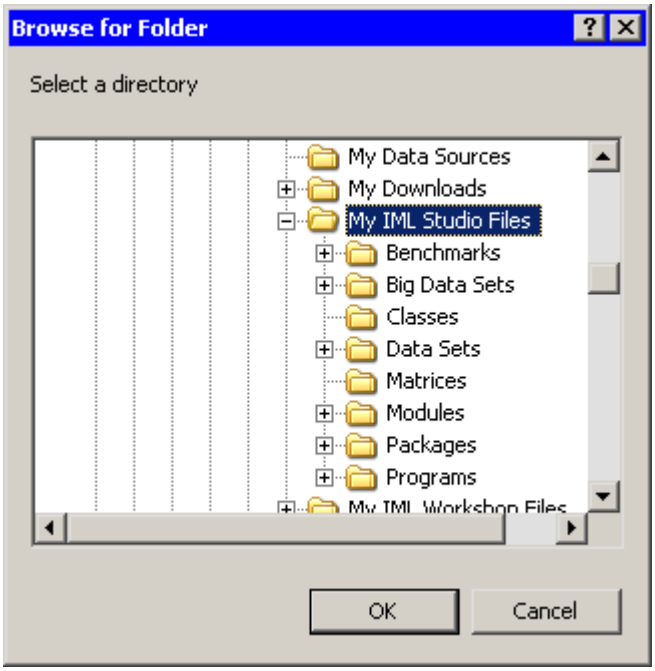
Move Down
moves the selected directory down one position in the **Search Path** list.

Set Default
copies the selected directory into the **Default Storage Directory** field.

Default Storage Directory
specifies the directory in which to store modules or matrices when an IMLPlus program executes a STORE statement. To change this field, click **Change** or **Set Default**.

Personal Files Directory
specifies the personal files directory. To change this field, click **Change**. The personal files directory is described in the section "The Personal Files Directory" on page 553.

Figure 34.12 The Browse for Folder Dialog Box



Example: Change the Search Path for Data Files

In this section, you add a new directory to the search path for data files. Data files include SAS data sets (with extensions *sd6* or *sas7bdat*) and Microsoft Excel files (with extension *xls*). When you try to load an IMLPlus matrix (with extension *imx*), SAS/IML Studio searches the directories in the search path for matrices. If the file is not found, SAS/IML Studio searches the directories in the search path for data files.

Assume that you have SAS data sets in a directory on your PC. The following steps add this directory to the beginning of the search path for data sets.

1. Select **Tools ►Options** from the main menu, and click the **Directories** tab.

   The **Directories** tab is shown in Figure 34.11.

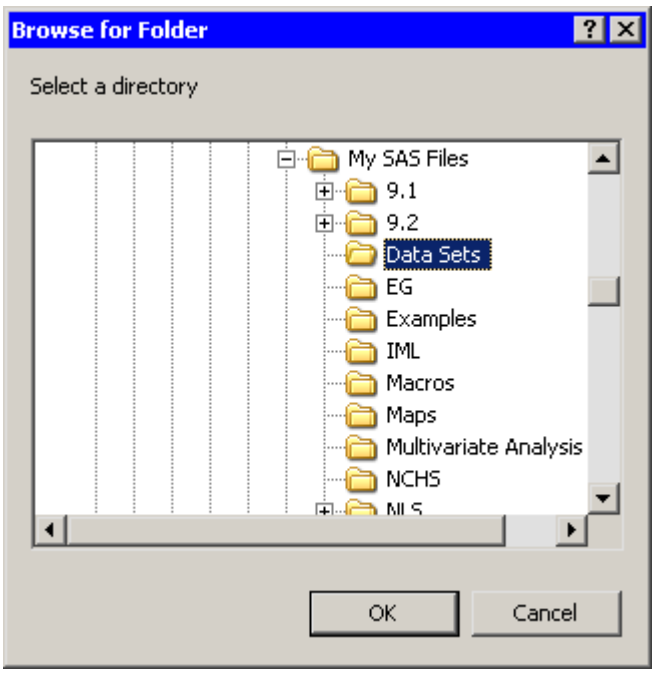
2. Select Data Files from the **Show directories for** list.

3. Click **Add**.

   The Browse for Folder dialog box appears.

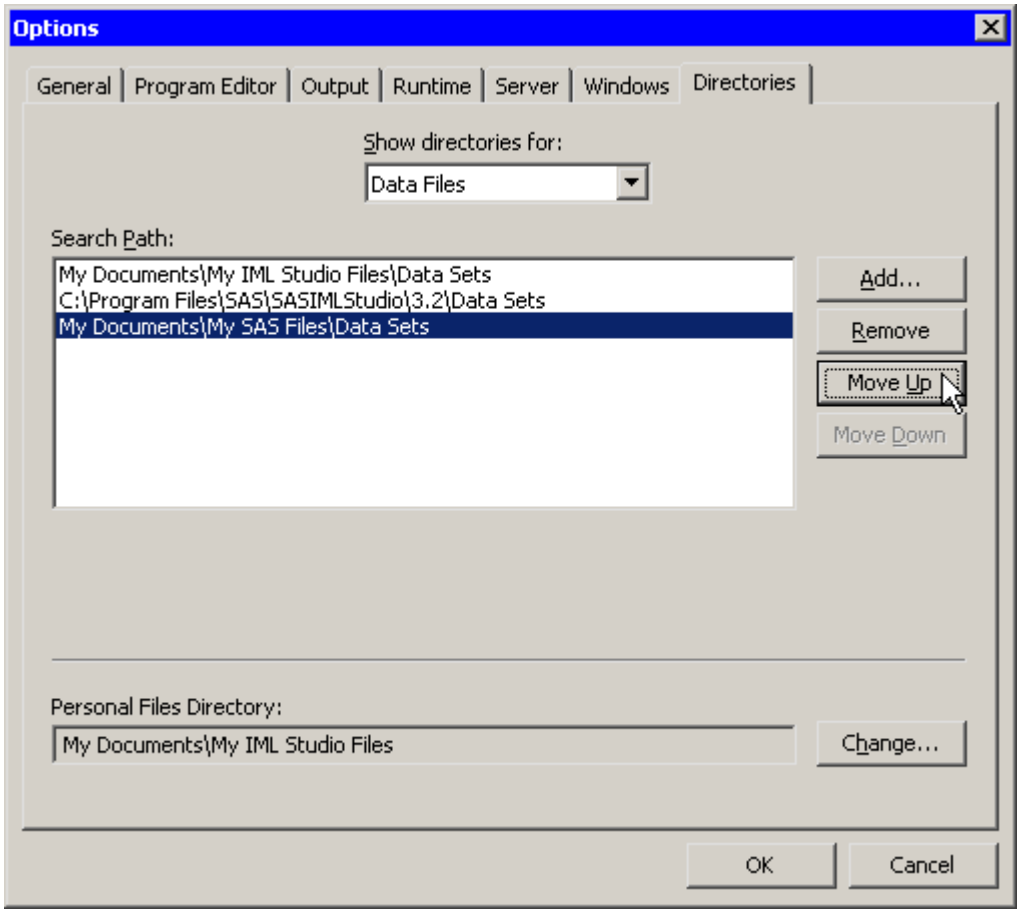
4. Navigate to the directory that contains your data, as shown in Figure 34.13. Click **OK**.

Figure 34.13 Changing the Search Path



The directory is appended to the end of the **Search Path** list, as shown in Figure 34.14.

Figure 34.14 Adding a New Directory

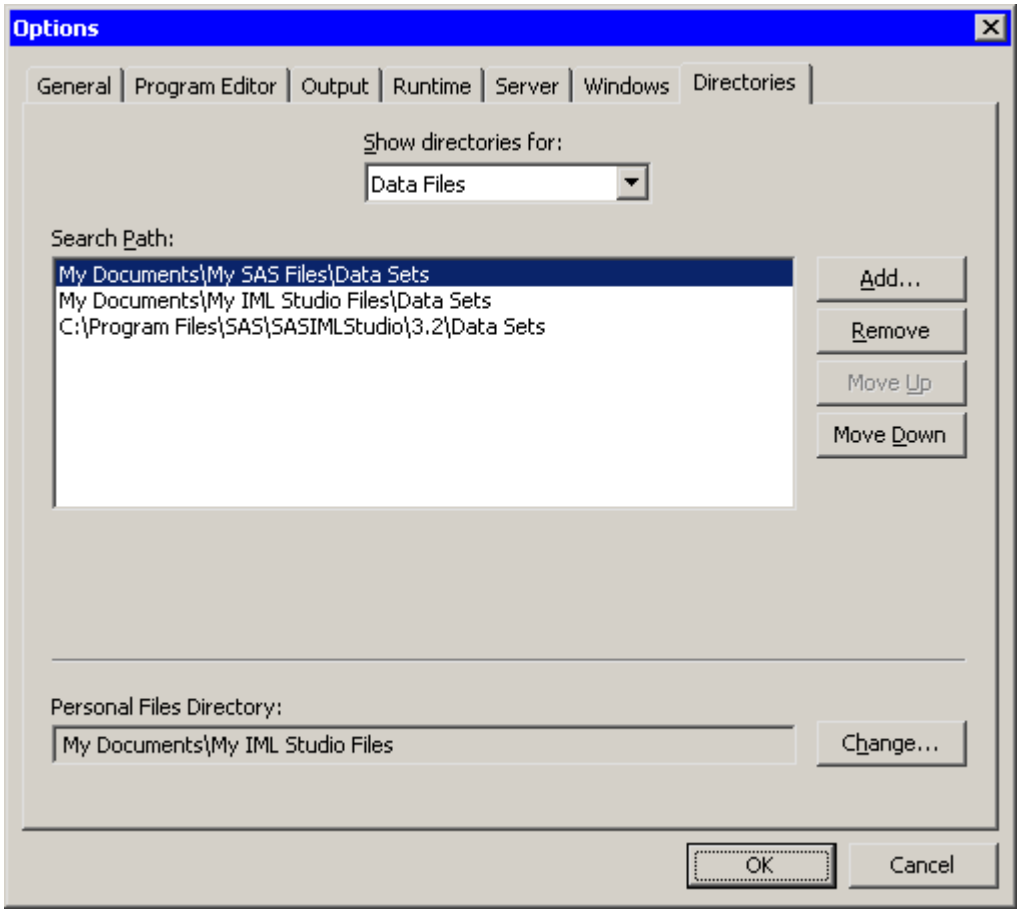


5 Click **Move Up** twice.

The directory is now at the beginning of the **Search Path** list, as shown in Figure 34.15.

6 Click **OK** to apply the changes.

Figure 34.15 The New Search Path



The Personal Files Directory

The first time you run SAS/IML Studio, a *personal files directory* called *My IML Studio Files* is created. By default, the personal files directory corresponds to the Windows directory shown in Table 34.1.

Table 34.1 The Personal Files Directory

Windows XP	**C:\Documents and Settings***userid***\My Documents\My IML Studio Files**
Windows Vista	**C:\Users***userid***\Documents\My IML Studio Files**

It is recommended that you store the files you create with SAS/IML Studio in subdirectories of the personal files directory. This provides the following advantages:

- Each person who logs on to the computer has a unique personal files directory.
- The personal files directory keeps your files separate from files distributed with SAS/IML Studio.
- If all your SAS/IML Studio files are in subdirectories of the personal files directory, it is easier for you to back up your files.
- When you open a file by selecting **File ►Open ►File** from the main menu, the dialog box contains a button that lets you navigate directly to the personal files directory.

In the personal files directory, SAS/IML Studio creates the following subdirectories:

Classes directory for user-written Java classes

Data Sets directory for SAS data sets

Matrices directory for IMLPlus matrices stored on the client computer

Modules directory for IMLPlus modules

Programs directory for IMLPlus programs

Example: Change the Personal Files Directory

To change the location of your personal files directory:

1 Select **Tools ►Options** from the main menu, and click the **Directories** tab.

The **Directories** tab is shown in Figure 34.11.

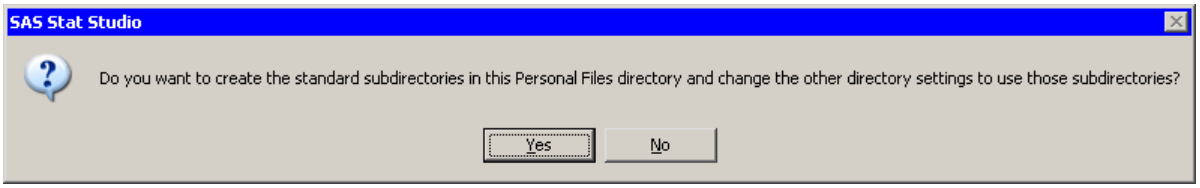
2 Click **Change** next to the **Personal Files Directory** field.

The Browse for Folder dialog box appears, as shown in Figure 34.12.

3 Select the directory you want to become your new personal files directory, and click **OK**.

A message box appears, as shown in Figure 34.16. You are asked whether you want to create the standard subdirectories in this new personal files directory.

Figure 34.16 A Message Box



4 Usually, you will want to respond to this prompt by clicking **Yes**.

5 Click **OK** to close the Options dialog box.

NOTE: When you change the location of the personal files directory, SAS/IML Studio does not move files from the previous personal files directory location. You must move the files yourself.

Appendix A
Sample Data Sets

Contents

Overview of Sample Data Sets 555
Air Data 556
Baseball Data 556
Business Data 557
Caribbean Data 558
Central America Data 558
Climate Data 559
Drug Data 560
Fish Data 560
GPA Data 561
Hurricanes Data 561
Iris Data 563
Mining Data 564
Miningx Data 564
Neuralgia Data 564
Patient Data 564
PRDSALE Data 565
Ship Data 565
States48 Data 566
References 566

Overview of Sample Data Sets

SAS/IML Studio is distributed with several data sets. These data sets are used in this documentation to demonstrate the capabilities and features of SAS/IML Studio.

To open any data sets described in this section, do the following:

1. Select **File ►Open ►File** from the main menu. The Open File dialog box appears.
2. Click **Go to Installation directory** near the bottom of the dialog box.
3. Double-click the *Data Sets* folder.

4. Select a data set.
5. Click **Open**.

The following sections describe the SAS/IML Studio sample data sets.

Air Data

The Air data set contains measurements of pollutant concentrations from a city in Germany during a week in November 1989.

The following list describes each variable.

datetime	date and hour, in SAS datetime format
day	day of the week
hour	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, in knots

Baseball Data

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 (Reichler 1987). There is one observation per hitter.

The following list describes each variable.

name	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 putouts (in 1986)
no_assts	number of assists (in 1986)
no_error	number of errors (in 1986)
salary	salary, in thousands of dollars (in 1986)

The position variable in the Baseball data set is encoded as follows:

Table A.1 Values for the Position Variable

13	First base and third base	CS	Center field and shortstop
1B	First base	DH	Designated hitter
1O	First base and outfield	DO	Designated hitter and outfield
23	Second base and third base	LF	Left field
2B	Second base	O1	Outfield and first base
2S	Second base and shortstop	OD	Outfield and designated hitter
32	Third base and second base	OF	Outfield
3B	Third base	OS	Outfield and shortstop
3O	Third base and outfield	RF	Right field
3S	Third base and shortstop	S3	Shortstop and third base
C	Catcher	SS	Shortstop
CD	Center field and designated hitter	UT	Utility
CF	Center field		

Business Data

The Business data set contains information about publicly held German, Japanese, and U.S. companies in the automotive, chemical, electronics, and oil refining industries in 1991. There is one observation for each company.

The following list describes each variable.

nation	nationality of the company
industry	principal business of the company
employs	number of employees
sales	sales for 1991, in millions of dollars
profits	profits for 1991, in millions of dollars

Caribbean Data

The Caribbean data set contains geographical data for countries in the western Atlantic Ocean. The data are used to create a map of the Caribbean islands. To create a map, plot lat versus lon, and select ID and segment as ID (grouping) variables.

The following list describes each variable.

ID	country code identifier
segment	segment code identifier for a country
lon	longitude of each point of a country segment
lat	latitude of each point of a country segment

Central America Data

The CentralAmerica data set contains geographical data for countries in Central America. The data are used to create a map of Central America. To create a map, plot lat versus lon, and select ID and segment as ID (grouping) variables.

The following list describes each variable.

ID	country code identifier
segment	segment code identifier for a country
lon	longitude of each point of a country segment
lat	latitude of each point of a country segment

Climate Data

The Climate data set contains geographical and meteorological data for certain cities in the 48 contiguous states of the United States.

The following list describes each variable.

station	name of city that contains the weather station
longitude	longitude of city
latitude	latitude of city
elevationFeet	elevation of city, in feet above mean sea level
JanMaxF	average maximum temperature in January, in degrees Fahrenheit
JanMinF	average minimum temperature in January, in degrees Fahrenheit
AprMaxF	average maximum temperature in April, in degrees Fahrenheit
AprMinF	average minimum temperature in April, in degrees Fahrenheit
JulMaxF	average maximum temperature in July, in degrees Fahrenheit
JulMinF	average minimum temperature in July, in degrees Fahrenheit
OctMaxF	average maximum temperature in October, in degrees Fahrenheit
OctMinF	average minimum temperature in October, in degrees Fahrenheit
extremeMaxF	highest recorded temperature, in degrees Fahrenheit
extremeMinF	lowest recorded temperature, in degrees Fahrenheit
JanAvePrecipIn	average precipitation in January, in inches
FebAvePrecipIn	average precipitation in February, in inches
MarAvePrecipIn	average precipitation in March, in inches
AprAvePrecipIn	average precipitation in April, in inches
MayAvePrecipIn	average precipitation in May, in inches
JunAvePrecipIn	average precipitation in June, in inches
JulAvePrecipIn	average precipitation in July, in inches
AugAvePrecipIn	average precipitation in August, in inches
SepAvePrecipIn	average precipitation in September, in inches
OctAvePrecipIn	average precipitation in October, in inches
NovAvePrecipIn	average precipitation in November, in inches
DecAvePrecipIn	average precipitation in December, in inches
totalAvePrecipIn	average total annual precipitation, in inches

Drug Data

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.

The following list describes each variable.

drug	drug used in treatment
disease	disease identifier
chang_bp	change in systolic blood pressure due to treatment

Fish Data

The Fish data set contains measurements of 159 fish caught in Finland's Lake Laengelmavesi (Puranen 1917).

The following list describes each variable.

species	species of fish
weight	weight of the fish, in grams
length1	length of the fish from the nose to the beginning of the tail, in centimeters
length2	length of the fish from the nose to the notch of the tail, in centimeters
length3	length of the fish from the nose to the end of the tail, in centimeters
height	maximum height of the fish, in centimeters
width	maximum width of the fish, in centimeters

In addition to these variables, the data set contains the following transformed variables.

cubeRootWeight	cube root of the weight
scaledLength1	the ratio length1 / cubeRootWeight
scaledLength2	the ratio length2 / cubeRootWeight
scaledLength3	the ratio length3 / cubeRootWeight
scaledHeight	the ratio height / cubeRootWeight
scaledWidth	the ratio width / cubeRootWeight
logLengthRatio	logarithm of the ratio length3 / length1

GPA Data

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.

The following list describes each variable.

gpa	grade point average of students in the computer science program
hsm	average high school grade in mathematics
hse	average high school grade in English
hss	average high school grade in science
satm	score on the mathematics section of the SAT
satv	score on the verbal section of the SAT
sex	student's gender

Hurricanes Data

The U.S. National Hurricane Center records intensity and track information for tropical cyclones at six-hour intervals. The Hurricanes data set is an "extended best-track" (EBT) data set that adds six measured size parameters to the best-track data. The data were prepared by DeMaria, Pennington, and Williams (2004). The cyclones from 1988 to 2003 are included.

The version distributed with SAS/IML Studio is Version 1.6, released February 2004. An earlier version of the EBT data was analyzed in Mulekar and Kimball (2004) and Kimball and Mulekar (2004).

The data as assembled by DeMaria include the following variables.

name	storm name
date	date of observation, in SAS date format
hms	time of observation (UTC), in SAS time format
latitude	latitude of observation, in degrees north latitude
longitude	longitude of observation. **NOTE:** DeMaria encodes this variable as degrees west longitude. For ease of plotting, this variable is recoded as a (usually) negative value in degrees east longitude.
wind_kts	maximum low-level sustained wind speed, in knots
min_pressure	minimum central sea-level pressure, in hPa
pressure_outer_isobar	pressure of outer closed isobar, in hPa
radius_eye	radius of eye (if an eye exists), in nautical miles. **NOTE:** A nautical mile is one minute of latitude, or approximately 1.15 statute miles.

radius_max_wind	radius at which maximum wind speed was measured, in nautical miles
radius_64kt	average radius of 64-knot (hurricane strength) winds, in nautical miles
radius_50kt	average radius of 50-knot winds, in nautical miles
radius_34kt	average radius of 34-knot (tropical storm strength) winds, in nautical miles
radius_outer_isobar	radius of outer closed isobar, in nautical miles
storm_type	indicator of whether the system was purely tropical, subtropical, or extra-tropical
month	month
day	day of the month
time	time of day (UTC)
year	year
ID	storm identification number
radius_34kt_ne	radius of 34-knot (tropical storm strength) winds northeast of the storm's center, in nautical miles
radius_34kt_se	radius of 34-knot (tropical storm strength) winds southeast of the storm's center, in nautical miles
radius_34kt_sw	radius of 34-knot (tropical storm strength) winds southwest of the storm's center, in nautical miles
radius_34kt_nw	radius of 34-knot (tropical storm strength) winds northwest of the storm's center, in nautical miles
radius_50kt_ne	radius of 50-knot winds northeast of the storm's center, in nautical miles
radius_50kt_se	radius of 50-knot winds southeast of the storm's center, in nautical miles
radius_50kt_sw	radius of 50-knot winds southwest of the storm's center, in nautical miles
radius_50kt_nw	radius of 50-knot winds northwest of the storm's center, in nautical miles
radius_64kt_ne	radius of 64-knot (hurricane strength) winds northeast of the storm's center, in nautical miles
radius_64kt_se	radius of 64-knot (hurricane strength) winds southeast of the storm's center, in nautical miles
radius_64kt_sw	radius of 64-knot (hurricane strength) winds southwest of the storm's center, in nautical miles
radius_64kt_nw	radius of 64-knot (hurricane strength) winds northwest of the storm's center, in nautical miles

The storm_type variable is encoded as follows:

*	Tropical system
W	Tropical wave
D	Tropical disturbance
S	Subtropical storm
E	Extra-tropical storm
L	Remnant low

In addition to these variables, the data set contains the following variables, suggested in the analyses of Mulekar and Kimball (2004) and Kimball and Mulekar (2004). Missing values were converted to the SAS missing value.

category	indicator variable that corresponds to the Saffir-Simpson wind intensity scale
wind_mph	maximum low-level sustained wind speed, in miles per hour. This variable is computed as wind_kts times 1.15.
radius_64kt	average of nonmissing values of the 64-knot radii in the northeast, southeast, southwest, and northwest directions
radius_50kt	average of nonmissing values of the 50-knot radii in the northeast, southeast, southwest, and northwest directions
radius_34kt	average of nonmissing values of the 34-knot radii in the northeast, southeast, southwest, and northwest directions

The category variable is encoded according to the value of wind_kts (wind speed) as in Table A.2.

Table A.2 The Saffir-Simpson Intensity Scale

Category	Description	Wind Speed (Knots)
TD	Tropical depression	22–33
TS	Tropical storm	34–63
Cat1	Category 1 hurricane	64–82
Cat2	Category 2 hurricane	83–95
Cat3	Category 3 hurricane	96–113
Cat4	Category 4 hurricane	114–134
Cat5	Category 5 hurricane	135 or greater

Iris Data

The Iris data set is Fisher's iris data (Fisher 1936). Sepal and petal size were measured for 50 specimens from each of three species of iris. There is one observation per specimen.

The following list describes each variable.

sepallen	sepal length, in millimeters
sepalwid	sepal width, in millimeters
petallen	petal length, in millimeters
petalwid	petal width, in millimeters
species	species of iris

Mining Data

The Mining data set contains the 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.

The following list describes each variable.

depth	depth of the hole, in feet
driltime	time to drill the last five feet of the current depth, in minutes
method	drilling method, wet or dry
rep	replicate number

Miningx Data

The Miningx data set is a subset of the Mining data set. It contains data from only one of the test holes.

Neuralgia Data

Neuralgia is pain that follows the path of specific nerves. Neuralgia is most common in elderly persons, but it can occur at any age. The Neuralgia data set contains data on 60 patients. These data are hypothetical, but they are similar to data reported by Layman, Agyras, and Glynn (1986).

Two test treatments and a placebo are compared. The response variable is Pain, which has the value "No" if the patient reports no pain or a substantial lessening of pain, and the value "Yes" if the patient still experienced pain after treatment.

The explanatory variables are as follows:

Treatment	treatment administered. "A" and "B" represent the two test treatments. "P" represents the placebo treatment.
Sex	gender of the patient
Age	age of the patient, in years, when treatment began
Duration	duration of complaint, in months, before the treatment began

Patient Data

The Patient data set contains data collected on cancer patients (Lee 1974). There is one observation per patient.

The response variable is remiss, which has the value 1 if the patient experienced cancer remission, and 0 otherwise.

The explanatory variables are the results from blood tests and physiological measurements on each patient. The variables are rescaled. The explanatory variables are cell, smear, infil, li, blast, and temp.

PRDSALE Data

The PRDSALE data set is also distributed in the SASHELP library. The data are artificial; the data set is typically used for resolving technical support issues.

The following list describes each variable.

actual	revenue from the sale of an item of furniture, in dollars
predict	predicted revenue from the sale, in dollars
country	country in which the item was sold
region	region in which the item was sold
prodtype	product type
product	item of furniture
quarter	quarter of year in which the item was sold
year	year in which the item was sold
month	month in which the item was sold

Ship Data

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 hull construction. There is one observation per ship.

The following list describes each variable.

type	type of ship
year	year of construction
period	period of operation
months	aggregate months of service
y	number of damage incidents

States48 Data

The States48 data set contains geographical data for the 48 contiguous states in the United States. The data are used to create a map of the continental United States. To create a map, plot lat versus lon, and select state and segment as ID (grouping) variables.

The following list describes each variable.

state	state code identifier
segment	segment code identifier for a state
postal	postal code identifier for a state
lon	longitude of each point of a state segment, in degrees west longitude
lat	latitude of each point of a state segment, in degrees north latitude

References

Afifi, A. A. and Azen, S. P. (1972), *Statistical Analysis: A Computer-Oriented Approach*, New York: Academic Press.

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.

DeMaria, M., Pennington, J., and Williams, K. (2004), "Description of the Extended Best Track File," Version 1.6, ftp://ftp.cira.colostate.edu/demaria/ebtrk/ (accessed March 1, 2004).

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

Kimball, S. K. and Mulekar, M. S. (2004), "A 15-Year Climatology of North Atlantic Tropical Cyclones, Part I: Size Parameters," *Journal of Climatology*, 17, 3555–3575.

Layman, P. R., Agyras, E., and Glynn, C. J. (1986), "Iontophoresis of Vincristine versus Saline in Post-herpetic Neuralgia: A Controlled Trial," *Pain*, 25, 165–170.

Lee, E. T. (1974), "A Computer Program for Linear Logistic Regression Analysis," *Computer Programs in Biomedicine*, 4, 80–92.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd Edition, London: Chapman & Hall.

Mulekar, M. S. and Kimball, S. K. (2004), "The Statistics of Hurricanes," *STATS*, 39, 3–8.

Penner, R. and Watts, D. G. (1991), "Mining Information," *American Statistician*, 45, 4–9.

Puranen, J. (1917), "Fish Catch data set (1917)," Journal of Statistics Education Data Archive, accessed May 22, 2009.
URL http://www.amstat.org/publications/jse/datasets/fishcatch.txt

Reichler, J. L., ed. (1987), *The 1987 Baseball Encyclopedia Update*, New York: Macmillan.

Appendix B
SAS/INSIGHT Features Not Available in SAS/IML Studio

Contents

The following list presents general features of SAS/INSIGHT that are not included in SAS/IML Studio:

- SAS/INSIGHT can be launched from SAS DMS mode (from the SAS program editor, from the **Solutions ►Analysis** menu, or from the SAS command line).
- SAS/INSIGHT shares the libraries and catalogs defined in DMS mode.
- SAS/INSIGHT automatically recomputes analyses (including curves on graphs) and statistics if data are changed.
- SAS/INSIGHT supports recording an interactive session for later playback.

The following list presents features of SAS/INSIGHT data views (tables and plots) that are not included in SAS/IML Studio.

- multiple plots in a single window
- “renewing” a plot or analysis
- GUI support for animation
- changing the orientation of plots
- changing the formats of table cells after the table is created
- saving tables to data sets after they are created
- changing the attributes of a curve after it is created
- user-defined formats
- a “Tools window” for rapidly changing attributes of markers and curves
- a mechanism to set a common view range for all plots that display a given variable
- multiple plots (for example, BY-group plots and scatter plot matrices) in a single window

The following list presents features of SAS/INSIGHT analyses that are not included in SAS/IML Studio.

- adding or deleting curves, graphs, variables, and tables from existing analyses without explicitly rerunning the analysis
- "group" variables for the analysis of BY-groups
- "freezing" an analysis for easy comparison with subsequent analyses
- sliders for interactively varying parameters in models
- creating a parametric CDF
- a kernel smoother for scatter plot smoothing
- maximum redundancy analysis
- biplots for many multivariate analyses

Index

OBSTAT variable, 41
ObsNum variable, 211

action menu, 211, 272
action menus, 533
active window, 24
AddAnalysisVar method, 532
adding
    observations, 35
    variables, 33
aggregate, 353
Air data set, 88, 556
Akaike information criterion, 274
analysis menu, 217
    not enabled, 371
animation, 567
annotations
    deleting, 137
    inserting, 135
    properties, 138
ANOVA, 317, 464, 477
AppendActionMenuItem method, 533
ASCII order, 52, 179
aspect ratio, 139, 142, 164
auto close property, 545
auto hide property, 545
auto position property, 545
auxiliary input window, 536
axes
    changing range, 165
    changing tick marks, 165
    labels, 169
    location, 114
    properties, 168
    setting common view range, 214
axis area, 146
axis label area, 146

bar charts, 14, 59
    properties, 62
Baseball data set, 301, 323, 404, 423, 528, 556
bin tool, 70, 134
biplots, 413, 417, 568
box plots, 20, 71
    displaying means, 142
    displaying notches, 142
    displaying serifs, 142
    properties, 75
Business data set, 78, 482, 557

BY groups, 178, 199, 201
BY variables, 199
BY-group analysis, 568
BY-group plots, 211
    copying to output doc, 214
    layout, 214
    not linked to original data, 211
    writing to files, 214

CANCORR procedure, 442
CANDISC procedure, 452
canonical components, 451
canonical correlation analysis, 441
canonical discriminant analysis, 451
canonical variables, 441
Caribbean data set, 558
CDF plot
    parametric, 568
CDF plots, 238, 243, 245
CentralAmerica data set, 558
changing contours, 121
chi-square residuals, 354
chi-squared (χ^2) symbol, 177
classification criterion, 469
classification fit plots, 462, 477
classification variables, 339, 346, 362, 384
client, 543
Climate data set, 111, 118, 559
closing windows, 195
color blend, 85, 141
colors
    of lines, 90
    of markers, 47, 85, 150
    predefined, 141
column headings, 37
column variables, 489
common factors, 421
communality, 422
comparing smoothers, 268
complement of selected observations, 141
confidence ellipses, 463
confidence interval displacement diagnostic, 354
confidence intervals, 230, 387
confidence levels, 400
confidence limits for means, 275, 284, 295
confidence limits for parameters, 318
configuration plots, 485, 491
configuring SAS/IML Studio, 535

confirmatory data analysis, 3
context areas, 146
context menus, 37, 146
contiguous selection, 404, 424, 453, 470
contingency tables, 78
contour plots, 117
    properties, 126
contours
    changing, 121
    levels, 127
    styles, 127
control menu, 545
convenient estimate, 502
Cook's *D* statistic, 308, 316, 373, 386
copying
    data, 54
    plots, 140, 198
CORR procedure, 391
correlation, 24, 84
    pairwise, 399
    partial, 399
correlation analysis, 391
correlation matrix
    in correlation analysis, 400
    in factor analysis, 435
    in principal component analysis, 405
    reduced, 422
correlation pattern plots, 409, 416
CORRESP procedure, 481
correspondence analysis, 481
covariance matrix
    in correlation analysis, 401
    in factor analysis, 435
    in principal component analysis, 405
covariance ratio, 314, 317
creating data, 29
curve attributes, 567
custom analysis, 528
cyclones, 12

data
    copying, 54
    creating, 29
    editing, 29
    saving, 33, 55
    subsetting, 54
data analysis, 1
data smoothing
    loess, 263
    polynomial regression, 289
    thin-plate spline, 279
data tables, 37
    creating new from selected data, 174
    properties, 56
data views, 11
DataObject methods
    AddAnalysisVar, 532
    GetNumObs, 532
    GetSelectedObsNumbers, 532
    GetSelectedVarNames, 532
    GetVarData, 532
    IsNominal, 532
    IsNumeric, 532
    SelectObs, 532
    SetMarkerColor, 532
    SetMarkerShape, 532
DataObject.SetVarValueOrder method, 183
DataView methods
    AppendActionMenuItem , 533
    GetDataObject, 533
    GetInitiator, 533
default label variables, 158
delete annotations, 137
design points, 283
deviance residuals, 354
DFBETAS, 357, 388
DFFIT statistic, 317
DIFCHISQ statistic, 354
DIFDEV statistic, 354
DISCRIM procedure, 452, 469
discriminant analysis, 469
discriminant function, 475
dispersion, 385
distribution analysis
    descriptive statistics, 217
    distributional modeling, 233
    frequency counts, 247
    location and scale statistics, 225
    outlier detection, 255
dmm file, 55
Drug data set, 362, 560
dynamically linked, 1, 11

editing
    data, 29
    observations, 35
effects, 339, 384
    crossed, 347
    factorial, 350
    main, 346
    multivariate polynomial, 352
    nested, 348
    polynomial, 351
    reordering, 353
    specifying, 346
eigenvalues, 407, 416, 417, 429, 435
eigenvectors, 408, 417, 437
error log window, 536

events, 346, 383
events/trials syntax, 346
examining selected observations, 54, 260
exclude from analyses, 47, 49, 141
exclude from plots, 16, 46, 49, 141
excluding observations, 176
    analyses not rerun, 177
    plots recomputed, 177
explanatory variables, 301
exploratory data analysis, 3, 11
extended selection, 16, 75

factor analysis, 421
factor plots, 422
FACTOR procedure, 422
factor spaces, 422
finding observations, 50
Fish data set, 452, 470, 560
font, 160
footnote, 164
format, 31, 53
freezing an analysis, 568
FREQ procedure, 247
frequency role, 39
frequency variables, 40

generalized cross validation, 274, 283
generalized squared distance, 469
GENMOD procedure, 362
GetDataObject method, 533
GetInitiator method, 533
GetNumObs method, 532
GetSelectedObsNumbers method, 532
GetSelectedVarNames method, 532
GetVarData method, 532
GetVars method, 533
Gini's mean difference, 261
global selection mode, 183, 189
goodness-of-fit test, 253
GPA data set, 442, 561
gradient color map, 101
graph area, 146
    margins, 139, 164
    properties, 163
graphical filtering, 189
group mean vector, 457
group variables, 88, 92

hat matrix, 313
Help ►Help Topics, 3
Heywood case, 423, 436
hiding windows, 195
high leverage points, 314
HISTOGRAM statement, 242
histograms, 17, 64
    anchor, 68
    bin tool, 70
    bin width, 68
    binning, 68, 70
    properties, 66
Hurricanes data set, 12, 60, 64, 72, 83, 104, 201, 217, 226, 234, 247, 256, 290, 392, 561

IMLPlus, 2, 527
include in analyses, 47, 49, 141
include in plots, 16, 46, 49, 141
including observations, 178
inertia, 481
influence diagnostics, 314
informat, 52
input data set, 519
insert annotations, 135
interaction tools, 131
interquartile range, 261
Iris data set, 563
IsNominal method, 532
IsNumeric method, 532
iterative reweighting, 274

kernel bandwidth, 221
kernel density estimate, 221
kernel smoother, 568
keyboard shortcuts
    in data tables, 57
    in plots, 140
kurtosis, 222

label role, 39
label variables, 156
labeling observations, 156
labels, 141
large left arrow, 96, 123, 125, 127
layout, 203, 214
level tool, 134
leverage points, 323
leverage statistic, 313, 316, 355
line plots, 87
    changing line properties, 143
    properties, 96
    selecting line, 143
    setting line color, 143
lines
    colors, 90
    selecting, 92
    styles, 90
link function, 361, 384
local regression, 263
local selection mode, 183, 189
local sorting, 56
location estimates, 230, 261

location parameter, 255
LOESS procedure, 263
log-linear model, 374
LOGISTIC procedure, 339

MAD, *see* median absolute deviation, *see* median absolute deviation
Mahalanobis distance, 334, 414
markers
  attributes, 205
  changing size, 142, 149
  changing size difference, 142, 149
  coloring, 141
  colors, 47, 85, 150
  properties, 47
  shapes, 47, 85, 147
  sizes, 85
maximum likelihood estimate, 502
maximum likelihood estimation, 242, 339
maximum redundancy analysis, 568
mean, 222
measure level, 34, 40
median absolute deviation, 225, 261
metadata, 55
Mining data set, 564
Miningx data set, 264, 280, 496, 501, 564
missing values, 52, 223, 251, 377, 394, 489
  in bar charts, 16, 62
  in box plots, 21, 74
MLE, *see* maximum likelihood estimation
model fitting
  generalized linear models, 361
  linear regression, 301
  logistic regression, 339
  robust regression, 323
modes, 230
mosaic plots, 78
  properties, 81
multivariate analysis
  canonical correlation analysis, 441
  canonical discriminat analysis, 451
  correlation analysis, 391
  correspondence analysis, 481
  discriminant analysis, 469
  factor analysis, 421
  principal component analysis, 403

Neuralgia data set, 340
normal density, 238
normalizing transformations, 495
notches, 75

oblique rotations, 436
observation inspector, 139
  multiple observations, 140
  scrolling, 140
observation inspector mode, 139
observations
  adding, 35
  editing, 35
  excluding, 176
  finding, 50
  including, 178
  labeling, 156
  labels, 48, 160
  properties, 46
  selecting, 45
  sorting, 43
observations menu, 46
observer view, 184
  of the intersection, 184
  of the union, 184
offset variables, 358, 362, 374, 376, 389
ordering, 178
  by data, 179, 181
  by frequency count, 179, 181
  missing values, 179
  nominal variables, 40
ordinary least squares regression, 301
orientation of plots, 567
orthogonal rotations, 436
Other threshold, 63, 81
Others category, 142
outliers, 255, 323
output data set, 519
output document, 214, 542
output document window, 536
overdispersion, 381
overplotting, 106, 152, 200

pairwise correlation, 399
pan tool, 133
parameter estimates, 318, 334, 386
parameterization, 353, 385
parametric distributions, 242, 243
partial correlation, 399
partial leverage, 316
partial leverage plots, 309
partial variables, 399, 415, 435, 448
pasting plots, 140
Patient data set, 564
pattern plots, 437
PAUSE statement, 536
personal files directory, 549, 553
  changing the location, 554
players, 556
plot area, 146
  margins, 161, 162
  properties, 161

values at edges, 162
Plot methods
GetVars, 533
plots
copying, 140, 198
not linked to original data, 395, 491
pasting, 140
regions, 146
Poisson regression, 374
pollutants, 556
polygon plots, 98
coloring regions, 99
filling polygons, 143
properties, 101
power transformations, 500
PRDSALE data set, 565
prediction ellipses, 396, 400, 464
prediction limits, 295
PRESS residuals, 314, 316
principal component analysis, 403
principal components, 403
automatic selection, 417
principal coordinates, 481
PRINCOMP procedure, 403
prior probability, 461
program editor, 540
program window, 536
programming language, 527

Q-Q plots, 238, 243, 244, 275, 285, 295, 311, 316, 333
quantiles, 223

RANK function, 513
RANKTIE function, 513
RD plots, 328
rebinning, 134
reduced correlation matrix, 422
reference lines, 141, 158
REG procedure, 289, 301
removing smoothers, 272
renewing a plot, 567
reset plot view, 135
residual plots, 275, 285, 295, 310, 316, 333, 354, 385
response distribution, 384
response variables, 301
robust distance, 333
robust regression algorithm, 332
ROBUSTREG procedure, 323
ROC curve, 354
role
frequency, 39
label, 39
weight, 39
rotating buttons, 105
rotating plots, 103
properties, 113, 143
rotating, 143
row headings, 38
row variables, 489

Saffir-Simpson Intensity Scale, 12, 60
sample programs, 527
SAS servers, 8, 543
SAS/INSIGHT, 6, 42
saving
data, 33, 55
plots, 214
saving tables, 567
scale estimates, 230, 261
scale multiplier, 255, 261
scale parameter, 240, 255
scatter plot smoothers
comparing, 268
loess, 268
removing, 272
scatter plots, 23, 83
matrix, 395, 400
properties, 85
score plots, 410, 416, 449, 463
scree plots, 416, 437
scrolling selected observations into view, 57
search path, 549
select tool, 132
selecting
lines, 92
observations, 45
selection rectangle, 19, 75
SelectObs method, 532
selector view, 184, 189
limit, 189
serifs, 75
server, 7, 8, 543
SetMarkerColor method, 532
SetMarkerShape method, 532
shape parameter, 240
Ship data set, 374, 565
show only selected observations, 85, 142, 152, 200
single-trial syntax, 345
singular value decomposition, 413
skewness, 222
slicing, 152
sliders, 568
smoothing criterion, 275
sorting observations, 43
span, 404, 421, 448, 459
spin tool, 134
spine plots, 462, 473, 476
standard deviation, 261

statement mode, 543
States48 data set, 566
status bar, 539
STORE statement, 549
studentized residuals, 313, 316, 333
subsetting data, 54, 174
supplementary variables, 492
surface drawing modes, 114
surface plots, 110

TABLES statement, 253
testing for normality, 238
threshold parameters, 240
ticks
    adjusting, 69
    anchor, 168
    major, 168
    minor, 168
    range of, 168
title, 164
tolerance, 51
tool bar, 538
tools window, 567
TPSPLINE procedure, 279
transformations
    Aranda-Ordaz, 511
    Box-Cox, 500
    common, 504
    custom, 518
    folded power, 510
    for proportion variables, 509
    Guerrero-Johnson, 511
    inverse, 506
    issues to consider, 523
    lag, 514
    logarithmic, 496
    normalizing, 495, 506
    rank, 513
    scaling and translation, 511
    square root, 506
    two-variable, 516
    variance stabilizing, 508
trials, 346, 383
trimmed mean, 230
Type 1 sequential analysis, 387
Type 3 statistic, 387

unicode characters, vii
unique factors, 422
UNIVARIATE procedure, 217, 225, 233, 255
user analysis, 528
user-defined formats, 567
UserAnalysis module, 528

variable transformation wizard, 495
variables
    adding, 33
    BY, 199
    canonical, 441
    classification, 339, 346, 362, 384
    explanatory, 301
    frequency, 40
    group, 88, 92
    label, 156
    offset, 358, 362, 374, 376, 389
    partial, 399, 415, 435, 448
    properties, 38
    response, 301
    roles, 38
    supplementary, 492
    weight, 40
    WITH, 399, 448
variables menu, 38
variance, 222

weight role, 39
weight variables, 40
welcome dialog, 539
whiskers, 71, 75
windows clipboard, 140, 198
Windows Device Independent Bitmap Format (BMP), 198
Windows Enhanced Metafile Format (EMF), 198
Winsorized mean, 230
WITH variables, 399, 448
workspace, 536
workspace bar, 538
workspace explorer, 190, 212, 395

zoom tool, 134