Chapter 1
Introduction

What Is Stat Studio?

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

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

Figure 1.1. The Stat Studio Interface
In addition, Stat 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
- dynamically linked graphics for exploratory data analysis

The programming language in Stat Studio, which is called IMLPlus, is an enhanced version of the IML programming language. IMLPlus extends IML to provide new language features such as the ability to create and manipulate statistical graphics and to call SAS procedures.

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

Related Software and Documentation

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

A second book, Stat Studio for SAS/STAT Users, is intended for SAS/STAT programmers. In it, you learn how to use Stat Studio in conjunction with SAS/STAT 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 by using the SUBMIT statement.

The third source of documentation is the Stat 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.

Stat Studio is closely related to the SAS/IML software. The language used to write programs in Stat Studio is called IMLPlus. This language consists of IML functions and subroutines, plus additional syntax to support the creation and manipulation of statistical graphics. The Stat Studio program windows color-code keywords in the IMLPlus language.

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

Exploratory 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?

Stat 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), Stat 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.

Stat 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 Stat 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.
Chapter 1. Introduction

Summary of Features

Stat 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 IML, C, FORTRAN, or Java

Comparison with SAS/INSIGHT

Stat Studio and SAS/INSIGHT 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 Stat Studio. You can see that the analyses are qualitatively similar.
Figure 1.2. A SAS/INSIGHT Analysis

Figure 1.3. A Comparable Stat Studio Analysis
However, there are three major differences between the two products. The first is that Stat 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 Stat Studio. In contrast, SAS/INSIGHT runs on the same computer on which SAS is installed.

A second major difference is that Stat Studio is programmable, and therefore extensible. SAS/INSIGHT contains standard statistical analyses that are commonly used in data analysis, but you cannot create new analyses. In contrast, you can write programs in Stat 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, Stat Studio is often referred to as the “programmable successor to SAS/INSIGHT.”

A third major difference is that the Stat 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.

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

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

The following list presents features of Stat Studio data views (tables and plots) that are not included in SAS/INSIGHT:

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

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

- Stat Studio can be programmed to call any licensed SAS analytical procedure and any IML function or subroutine.
Chapter 1. Introduction

- Stat Studio detects outliers in contaminated Gaussian data.
- Stat Studio fits robust regression models and detects outliers and high-leverage observations.
- Stat Studio supports the generalized linear model with a multinomial response.
- Stat Studio creates graphical results for the analysis of logistic models with one continuous effect and a small number of levels for classification effects.
- Stat Studio provides parametric and nonparametric methods of discriminant analysis.
- Stat Studio provides common factor analysis for interval variables.
- Stat Studio provides correspondence analysis for nominal variables.

Features of SAS/INSIGHT that are not included in Stat Studio are presented in Appendix B, “SAS/INSIGHT Features Not Available in Stat Studio.”

## Typographical Conventions

This documentation uses some special symbols and typefaces.

- Field names, menu items, and other items associated with the graphical user interface are in bold; for example, a menu item is written as File ➤ Open ➤ Server Data Set. A field in a dialog box is written as the Anchor tick field.
- Names of Windows files, folders, and paths are in bold; for example, C:\Temp\MyData.sas7bdat.
- SAS librefs, data sets, and variable names are in Helvetica; for example, the age variable in the work.MyData data set.
- Keywords in SAS or in the IMLPlus language are in all capitals; for example, the SUBMIT statement or the ORDER= option.

This documentation is full of examples. Each step in an example appears in bold.

⇒ This symbol and typeface indicates a step in an example.
References


