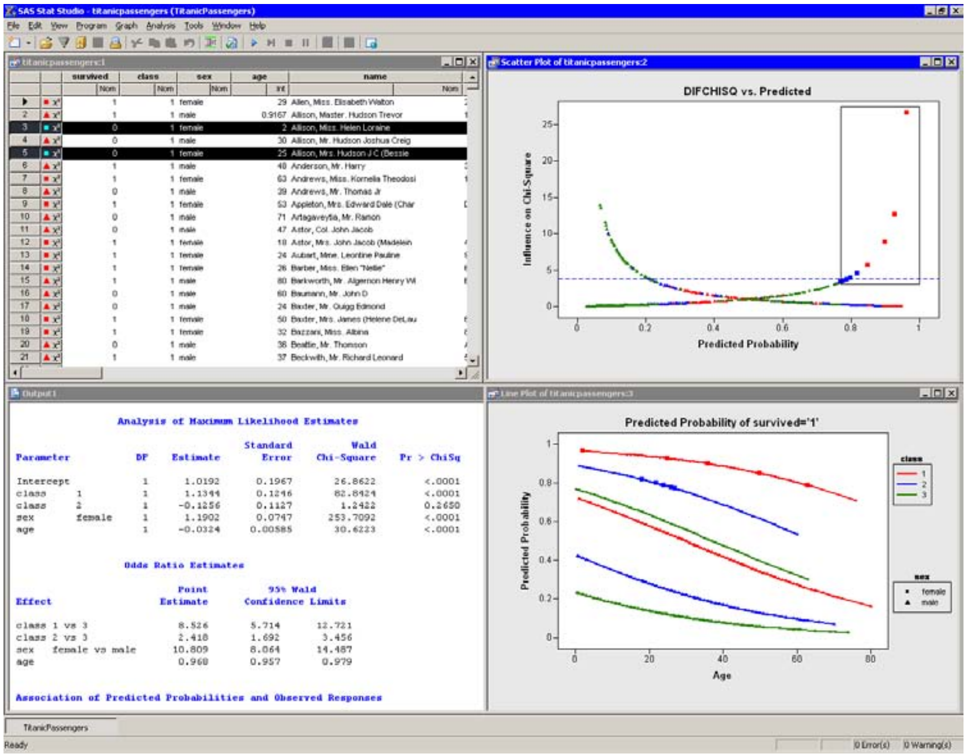
Chapter 1
Introduction

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



		survived	class	sex	age	name
		Nom	Nom	Nom	Int	Nom
▶	■	1	1	female	29	Allen, Miss. Elisabeth Walton
2	▲	1	1	male	0.9167	Allison, Master. Hudson Trevor
3	■	0	1	female	2	Allison, Miss. Helen Loraine
4	▲	0	1	male	30	Allison, Mr. Hudson Joshua Creig
5	■	0	1	female	25	Allison, Mrs. Hudson J C (Bessie
6	▲	1	1	male	48	Anderson, Mr. Harry
7	■	1	1	female	63	Andrews, Miss. Kornelia Theodosi
8	▲	0	1	male	39	Andrews, Mr. Thomas Jr
9	■	1	1	female	53	Appleton, Mrs. Edward Dale (Char
10	▲	0	1	male	71	Artagaveytia, Mr. Ramon
11	▲	0	1	male	47	Astor, Col. John Jacob
12	■	1	1	female	18	Astor, Mrs. John Jacob (Madelein
13	■	1	1	female	24	Aubart, Mme. Leontine Pauline
14	■	1	1	female	26	Barber, Miss. Ellen "Nellie"
15	▲	1	1	male	80	Barkworth, Mr. Algernon Henry Wi
16	▲	0	1	male	60	Baumann, Mr. John D
17	▲	0	1	male	24	Baxter, Mr. Quigg Edmond
18	■	1	1	female	50	Baxter, Mrs. James (Helene DeLau
19	■	1	1	female	32	Bazzani, Miss. Albina
20	▲	0	1	male	36	Beattie, Mr. Thomson
21	▲	1	1	male	37	Beckwith, Mr. Richard Leonard

```
                 Analysis of Maximum Likelihood Estimates

                                       Standard          Wald
Parameter          DF      Estimate       Error    Chi-Square    Pr > ChiSq

Intercept           1        1.0192      0.1967       26.8622        <.0001
class     1         1        1.1344      0.1246       82.8424        <.0001
class     2         1       -0.1256      0.1127        1.2422        0.2650
sex       female    1        1.1902      0.0747      253.7092        <.0001
age                 1       -0.0324     0.00585       30.6223        <.0001


                     Odds Ratio Estimates

                              Point          95% Wald
Effect                     Estimate      Confidence Limits

class 1 vs 3                  8.526       5.714      12.721
class 2 vs 3                  2.418       1.692       3.456
sex    female vs male        10.809       8.064      14.487
age                           0.968       0.957       0.979


Association of Predicted Probabilities and Observed Responses
```

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.

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 Stat Studio **Help** menu.

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.

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

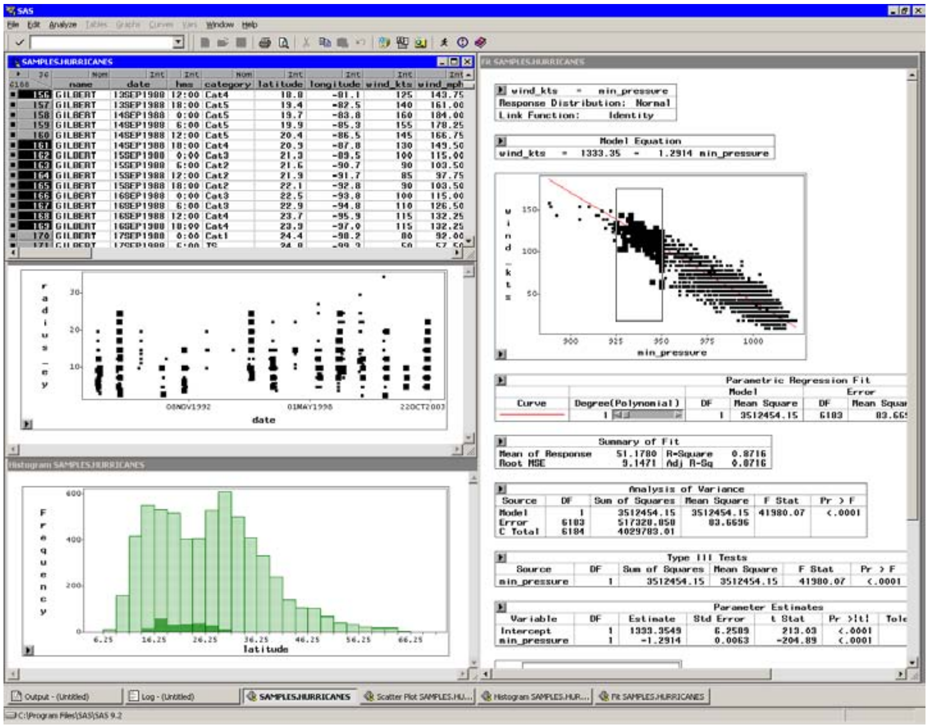
To extend the set of available analyses, you can do the following:

- 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.

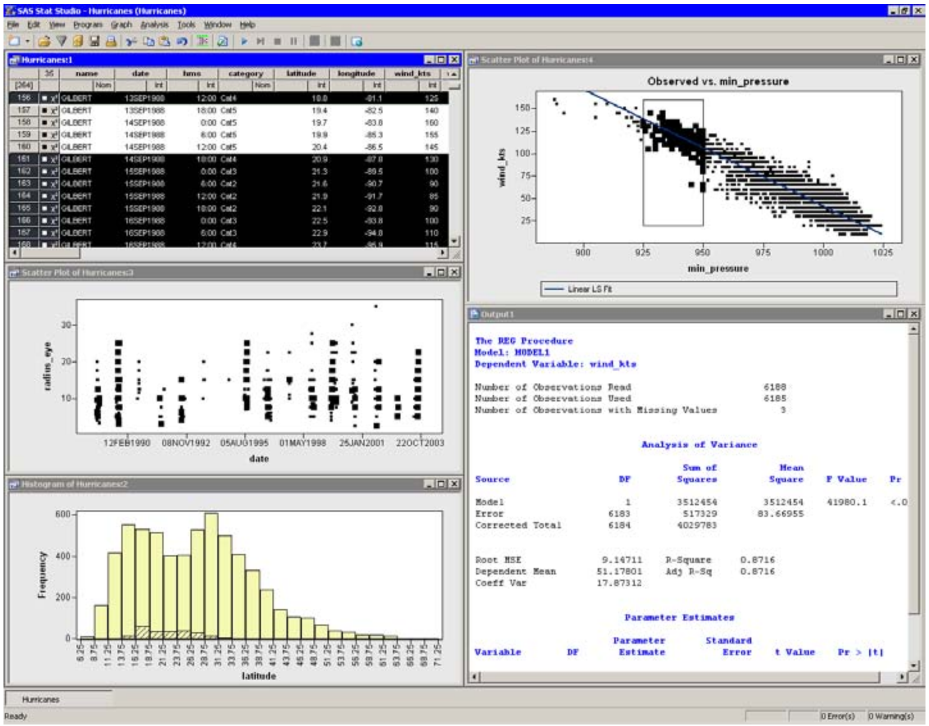
Figure 1.2 shows a typical SAS/INSIGHT analysis. Figure 1.3 shows the same analysis performed in Stat Studio. You can see that the analyses are qualitatively similar.



36	Nom	Int	Int	Nom	Int	Int	Int	Int
6188	name	date	hms	category	latitude	longitude	wind_kts	wind_mph
156	GILBERT	13SEP1988	12:00	Cat4	18.8	-81.1	125	143.75
157	GILBERT	13SEP1988	18:00	Cat5	19.4	-82.5	140	161.00
158	GILBERT	14SEP1988	0:00	Cat5	19.7	-83.8	160	184.00
159	GILBERT	14SEP1988	6:00	Cat5	19.9	-85.3	155	178.25
160	GILBERT	14SEP1988	12:00	Cat5	20.4	-86.5	145	166.75
161	GILBERT	14SEP1988	18:00	Cat4	20.9	-87.8	130	149.50
162	GILBERT	15SEP1988	0:00	Cat3	21.3	-89.5	100	115.00
163	GILBERT	15SEP1988	6:00	Cat2	21.6	-90.7	90	103.50
164	GILBERT	15SEP1988	12:00	Cat2	21.9	-91.7	85	97.75
165	GILBERT	15SEP1988	18:00	Cat2	22.1	-92.8	90	103.50
166	GILBERT	16SEP1988	0:00	Cat3	22.5	-93.8	100	115.00
167	GILBERT	16SEP1988	6:00	Cat3	22.9	-94.8	110	126.50
168	GILBERT	16SEP1988	12:00	Cat4	23.7	-95.9	115	132.25
169	GILBERT	16SEP1988	18:00	Cat4	23.9	-97.0	115	132.25
170	GILBERT	17SEP1988	0:00	Cat1	24.4	-98.2	80	92.00

```
wind_kts  =  min_pressure
Response Distribution:  Normal
Link Function:      Identity

                 Model Equation
wind_kts  =  1333.35  -  1.2914 min_pressure

                                  Parametric Regression Fit
                                        Model                  Error
   Curve      Degree(Polynomial)   DF   Mean Square     DF   Mean Squar
                      1             1   3512454.15     6183      83.66

             Summary of Fit
Mean of Response   51.1780  R-Square   0.8716
Root MSE            9.1471  Adj R-Sq   0.8716

                       Analysis of Variance
Source     DF   Sum of Squares  Mean Square    F Stat    Pr > F
Model       1      3512454.15   3512454.15   41980.07    <.0001
Error    6183      517328.858      83.6696
C Total  6184      4029783.01

                               Type III Tests
   Source         DF   Sum of Squares  Mean Square    F Stat    Pr > F
min_pressure       1       3512454.15   3512454.15   41980.07   <.0001

                                 Parameter Estimates
  Variable       DF     Estimate    Std Error     t Stat    Pr >|t|   Tole
Intercept         1    1333.3549       6.2509     213.03    <.0001
min_pressure      1      -1.2914       0.0063    -204.89    <.0001
```

Figure 1.2. A SAS/INSIGHT Analysis



	36	name	date	hms	category	latitude	longitude	wind_kts
[264]		Nom	Int	Int	Nom	Int	Int	Int
156	■ x²	GILBERT	13SEP1988	12:00	Cat4	18.8	-81.1	125
157	■ x²	GILBERT	13SEP1988	18:00	Cat5	19.4	-82.5	140
158	■ x²	GILBERT	14SEP1988	0:00	Cat5	19.7	-83.8	160
159	■ x²	GILBERT	14SEP1988	6:00	Cat5	19.9	-85.3	155
160	■ x²	GILBERT	14SEP1988	12:00	Cat5	20.4	-86.5	145
161	■ x²	GILBERT	14SEP1988	18:00	Cat4	20.9	-87.8	130
162	■ x²	GILBERT	15SEP1988	0:00	Cat3	21.3	-89.5	100
163	■ x²	GILBERT	15SEP1988	6:00	Cat2	21.6	-90.7	90
164	■ x²	GILBERT	15SEP1988	12:00	Cat2	21.9	-91.7	85
165	■ x²	GILBERT	15SEP1988	18:00	Cat2	22.1	-92.8	90
166	■ x²	GILBERT	16SEP1988	0:00	Cat3	22.5	-93.8	100
167	■ x²	GILBERT	16SEP1988	6:00	Cat3	22.9	-94.8	110

```
The REG Procedure
Model: MODEL1
Dependent Variable: wind_kts

Number of Observations Read                    6188
Number of Observations Used                    6185
Number of Observations with Missing Values        3


                       Analysis of Variance

                                  Sum of          Mean
Source               DF          Squares        Square    F Value    Pr

Model                 1          3512454       3512454    41980.1    <.0
Error              6183           517329      83.66955
Corrected Total    6184          4029783


Root MSE              9.14711    R-Square     0.8716
Dependent Mean       51.17801    Adj R-Sq     0.8716
Coeff Var            17.87312


                      Parameter Estimates

                     Parameter       Standard
Variable       DF     Estimate          Error    t Value    Pr > |t|
```

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.

- 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

Gelman, A. (2004), “Exploratory Data Analysis for Complex Models,” *Journal of Computational and Graphical Statistics*, 13(4), 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(4), 281–295.