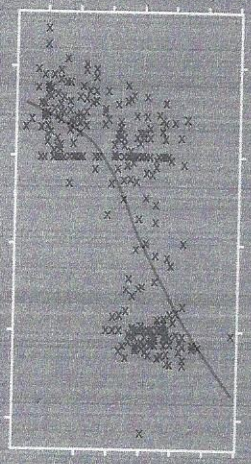
Statistics and Computing

Andreas Krause
Melvin Olson

The Basics of S and S-PLUS

Second Edition



Springer

Statistics and Computing

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Andreas Krause
Melvin Olson

The Basics of S and S-PLUS

Second Edition

With 94 Illustrations

Springer

Andreas Krause
Novartis Pharma AG
4002 Basel
Switzerland
Andreas.Krause@pharma.novartis.com

Melvin Olson
Allergan, Inc. (eff. 5/2000)
2525 Dupont Drive
P.O. Box 19534
Irvine, CA 92623-9534
USA
olson_melvin@allergan.com

Library of Congress Cataloging-in-Publication Data
Krause, Andreas.
The basics of S and S-PLUS / Andreas Krause, Melvin Olson.—2nd ed.
p. cm. — (Statistics and Computing)
Includes bibliographical references and index.
ISBN 0-387-98961-7 (softcover)
1. S-PLUS. 2. Mathematical statistics—Data processing. I. Olson, Melvin. II. Title.
III. Series.
QA276.4.K73 2000
519.5'0285'53—dc21
99-056074

Printed on acid-free paper.

Production managed by A. Orrantia; manufacturing supervised by Joe Quatela.
Photocomposed pages prepared from the authors' LaTeX files.
Printed and bound by R.R. Donnelley and Sons, Harrisonburg, VA.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-98961-7 Springer-Verlag New York Berlin Heidelberg SPIN 10750178

Contents

1
Introduction

Over the years, the S language and S-PLUS have undergone many changes. Since its development in the mid-seventies, the three main authors of S, Rick Becker, John Chambers, and Allan Wilks, have enhanced the entire language considerably. All their work was done at Bell Labs with the original goal of defining a language to make it easier to do repetitive tasks in data analysis, like calculating a linear model.

In the following years, many people contributed to the S project in one form or another. People outside Bell Labs also became aware of the interesting development and took part in it, and this is to a great extent the way S and S-PLUS are still developed today. A very lively user community works with and on S/S-PLUS, and they especially appreciate the S style of working in an integrated environment. Special strengths are the extremely modern and flexible language, which has many elements of an interactive C, Lisp, and APL, and good graphics capabilities.

It is noteworthy that the authors do not consider S as a primarily statistical system. The system was developed to be flexible and interactive, especially designed for doing data analysis of an exploratory nature, which began to boom in the late seventies after the release of Tukey's book (1977) on the subject. Most of the statistical functionality was added later, and many statistics routines like estimation, regression, and testing were incorporated by the S-PLUS team.

This chapter describes the development of S and S-PLUS over the years, clarifies the differences between the two, and points to some further references.

1.1 The History of S and S-PLUS

The S Language, and some years later the NEW S, were developed at AT&T Bell Labs in the late seventies and early eighties, mainly by Rick Becker and John Chambers. Some years later, Allan Wilks joined the core team. Since then, several other people have been involved in the project. Becker (1994) describes in great detail the foundation and development of S and points out some future directions.

The year 1976 can perhaps be viewed as the founding year of S. First concepts were developed and implemented. At first, the "system" consisted of a library of routines together with an interface, such that the kernel code itself could be kept unmodified. In 1981, the S team decided to rewrite the system in C and port it to the UNIX operating system. Since 1981, the source code has been available for interested people outside Bell Labs.

The next years revealed a strong increasing interest among statisticians and data analysts in using the system, which was called S by then. It is remarkable that the important steps in the development of S were all marked by books, such that S users today talk about the days of the *Brown Book*, the *Blue Book*, and the *White Book*. The *Green Book* marks the most recent milestone.

In 1984, as the interest in S began to rise, a real manual was necessary. The first book, today referred to as the *Brown Book*, was written by Becker and Chambers (1984). This version of S is now referred to as "Old S," as no version numbers existed at the time.

The QPE (Quantitative Programming Environment) developed by John Chambers set a milestone in the development of S. In 1988, it introduced the function concept (replacing the former macros), and new programming concepts were added. This work is described in the *Blue Book* (Becker, Chambers, and Wilks, 1988).

During all these years, the user community added substantial functionality to S, and many sophisticated techniques like tree regression, nonparametric smoothing, survival analysis, object-oriented programming, and new model formulation became a part of S. This step in the development was manifested and accompanied by the *White Book* (Chambers and Hastie, 1992).

Version 4 of S came out in 1998 and is described in full detail in Chambers (1998), the *Green Book*. The object-oriented paradigm forms the basis of the entire language, documentation is integrated into an object, and the general paradigm "everything is an object" is followed throughout the language.

In 1987, Douglas Martin at the University of Washington, Seattle, founded a small company to make S more popular. He realized that the major drawback of S was the need of professional support for end-users. Hence, he started the company Statistical Sciences, Inc. (StatSci), a division of MathSoft, Inc., since 1994. StatSci added more functionality to S, ported

it to other hardware platforms, and provided the necessary support for technical and statistical problems. The enhanced version of S received a new name: S-PLUS.

S-PLUS helped popularize S among nontechnical people. StatSci ported S-PLUS to the only non-UNIX platform, releasing S-PLUS for DOS in 1989 and S-PLUS for Windows in 1993. Up to version 3, S-PLUS for DOS/Windows and UNIX provided essentially the same functionality.

S-PLUS Version 4 was only released for the Windows platform. It was enhanced by a new graphical user interface (GUI) and a graphics system based on the Axum package. It adopts the Windows standard and gives a menu-based interface to S-PLUS. Windows-specific functionality like the creation of PowerPoint slides or direct data exchange with other programs via DLLs was added. Much of the S-PLUS functionality is available via menus and buttons, and the graphics are shown in an editable graph sheet. A major strength is that all functionality that is accessible via the menus can also be called from the command line. S-PLUS comes with a script window where the corresponding command line input is shown.

In fall 1998, S-PLUS for Windows was released in version 4.5, which is split into a "Professional Edition" that comprises the full functionality and a menu-only version called "S-PLUS Standard."

Version 5 of S-PLUS was only released on UNIX systems in late 1998 and 1999. It is based on S Version 4 as described in Chambers (1998). For the first time, the Linux system for Intel PC is supported.

S-PLUS 2000 for Windows systems (analogous to Office 2000 by Microsoft) was released in mid-1999, still based on S Version 3. The next release for Windows will (probably) be based on S Version 4.

S is still the heart of the system, and the core S team continues to work on the S system. The whole S functionality is incorporated in S-PLUS and enhanced, and today the S system is no longer publically available. In the remainder of the book, we will use S-PLUS as the standard reference.

1.2 S-PLUS on Different Operating Systems

Sometimes it is important to know about differences in software on various hardware systems or under different operating systems. This can be the case if you work on more than one computer system with S-PLUS (and therefore need to exchange data files) or if you want to be informed about the differences before deciding in favor of a specific system. In this section, we discuss some details about different systems supported by S-PLUS.

In addition, the chapter provides some basic information about the general setup of files and structures in S-PLUS. More information on the S-PLUS internal workings can be found in Sections 12.1 and 12.2.

At present, S-PLUS supports two major operating systems: UNIX (with most of its variants) and Windows (with its variants). Table 1.1 summarizes the currently supported hardware and operating systems.

Table 1.1. Systems supported by S-PLUS

AIX 4.3.1	DEC UNIX 4.0	HPUX 10.2, 11.0
Intel-based Linux	IRIX 6.2, 6.5	Solaris 2.5, 2.6 (SPARC)
Windows 95, 98, NT, 2000		

As the S source code is no longer available, machines not binary compatible to the ones supported are not able to run S or S-PLUS. In those cases, the R system offers an alternative. We will get to the details in Section 14.3 (page 367).

S-PLUS has minimum requirement specifications regarding main memory and hard disk size. The disk space needed depends on the version and operating system, but is typically between 50 and 100 Megabytes.

As a side note, S-PLUS consumes and releases memory dynamically during a session, depending on the needs. Therefore, it does not run out of memory until there is no more main memory and swap space available. If S-PLUS runs out of main memory (RAM), the operating system assigns virtual memory (i.e., hard disk space) as a substitute. As this slows down the execution time dramatically, the machine should be equipped with a reasonable amount of memory. For improving performance, main memory is the first speedup factor. If you are not satisfied with the performance, watch for permanent hard disk access while executing commands, or use a monitoring tool (like "top" under UNIX or the "Task Manager" under Windows) to track swapping activity.

Differences Between Versions

S-PLUS has some differences in its implementation between the UNIX and the Windows versions. The most visible difference is the user interface. S-PLUS for Windows has a typical Windows-like menu-based user interface. Many graphical functions can be started using the toolbar interface. Graphics are editable using point and click. Data and graphics transfer from and to other Windows applications (like Excel and Powerpoint) is part of the system.

S-PLUS for UNIX allows easy integration of C and Fortran code. S-PLUS for Windows allows one to dynamically link libraries (DLLs).

S-PLUS for Windows comes in two editions: "Professional" and "Standard." The standard version does not provide command line access.

1.3 Notational Conventions

By now you must be eager to get started, but it might be worth reviewing these few notational conventions first, as they are used throughout the book.

To begin with, you must be aware that when running S-PLUS, you will be asked for a new command with the greater than sign: >. We use the > in the book to denote the S-PLUS prompt. Also, if a single command extends over one line of input, the prompt changes to the plus sign: +. A preview of this is shown below.

```
> This is where an S-PLUS command would appear and
+ notice that the prompt changes on the second line.
```

If a longer listing or function is shown, the prompt is typically omitted, as we do not assume interactive input, and reading the code is easier.

There are occasional examples of commands to either the UNIX or DOS shells. For these examples, no prompt is used.

All commands in S-PLUS are actually calls to functions. To highlight the occurrence of a function in S-PLUS, or one of the parameters to it, we have written them in a special font as with the example, `print`.

When presenting commands, we sometimes include descriptive text. The descriptive text is written in S-PLUS syntax for comments. Anything after the number sign # until the end of the line is treated as a comment and not interpreted as a command.

A summary of these conventions is found in Table 1.2.

Table 1.2. Notational conventions

Convention	Explanation
>	S-PLUS prompt
+	Command has continued onto next line
`Commands`	Typewriter font
No prompt	For calls to UNIX or DOS shell
#	Comment symbol indicating start of a comment
placeholders	Italic. You need to replace them by an appropriate expression, like *filename*, which needs to be replaced by a valid file name
Menu	Menu entries and buttons are referred to in this font
Note	Notes point out something important, like a practical example, an application, or an exception. They end with the symbol ◁

2

Windows User Interface

2.1 Introduction

It has been fashionable lately for statistical software packages to become interactive and/or "point-and-click." S-PLUS has always been interactive but has only recently added a Windows-based point-and-click interface. The Windows user now has the best of both worlds available, the ease of the graphical user interfact (GUI) combined with the detailed commands and control offered by the command line environment.

The advantage of the GUI for the novice of S-PLUS is that you don't have to know the syntax of S-PLUS to get started. All you need is a little familiarity with typical Windows software and a data set in some sort of standard format.

The point-and-click approach to using S-PLUS is not available on UNIX platforms. With these platforms, the Commands Window is the system which does not rely on any menus or toolbars. As the complete functionality of S-PLUS is only available through the Commands Window anyway (or its counterpart, the Script Window), using S-PLUS on these platforms represents no loss of performance, although perhaps a small sacrifice of ease of use. Users of UNIX or Linux should skip to the next chapter which is where the more detailed treatment of the Commands Window begins.

The approach we will take in this chapter is to quickly introduce the S-PLUS system design under Windows, show the briefest of explanations of how it functions, and finish with describing in detail the various tasks that will be needed to complete a data analysis, from data input to printing

and saving the results. It is by no means intended to be an extensive or exhaustive exploration of the GUI, merely a way of familiarizing you with its structure, where to find things, and, most importantly, where and what to try for more detailed options.

2.2 System Overview

When you open S-PLUS under Windows by double-clicking on the S-PLUS icon, you are greeted by a screen layout as shown in Figure 2.1. This may vary slightly according to the version of S-PLUS you are using. The main elements that are visible include the Object Explorer, the Commands Window, the menus, and the toolbar. Optionally, a graphics window can be opened. A short description of each of these components is given in the next several subsections.[1]

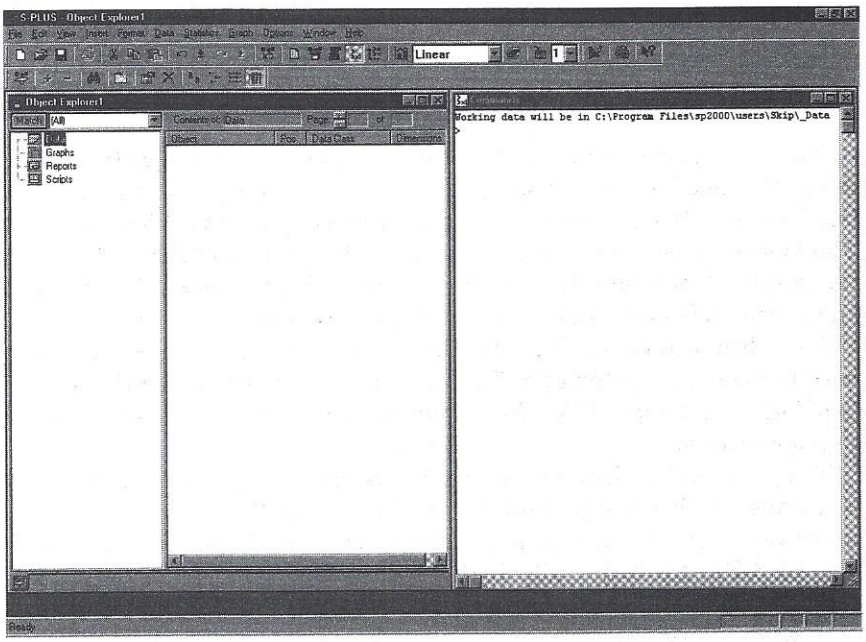


Figure 2.1. The S-PLUS screen and its components, the Object Explorer, the Commands Window, and the toolbars.

[1] According to the version of S-PLUS being used, some components might not be present.

The general layout of the S-PLUS system is similar to many popular windows systems in that it has pull-down menus at the top and toolbars just below the menus. To use such a system it is useful to be a little familiar with basic point-and-click operations and how to use a mouse.

For those who are not that comfortable with window- and icon-based software, the following subsections provide a crash introduction to the essentials. The pull-down menus across the top are used to group categories of commands or options to be set. Under the ***File*** menu, we find actions relating to files (***Open***, ***Close***, ***Import***, ***Save***) as well as to exiting the system. The toolbars below the menus contain buttons that are convenient shortcuts to commands found through the layers in the menus. Some of the toolbar buttons (e.g., ***plots2d***) open a palette containing a myriad of options to complete your task.

An online help facility is available through the main menu. The typical approach to handling the Help system (i.e., search by index or topic) has been used. The big advantage of the Help system in S-PLUS is that it includes online manuals. We encourage liberal use of the help system.

2.2.1 Using a Mouse

Using a mouse efficiently is important to get the most out of the system. Clicking once on the left mouse button is usually used to highlight an item in a list (e.g., a file out of a list of files) or to select a menu heading or button from a toolbar. You will not always be able to guess at the function of a toolbar button merely by looking at the icon, but if you are at a loss, simply position the mouse over the button in question and a short text description will appear below it. Double-clicking on the left mouse button is used to select and execute. Examples include double-clicking on a file name to select it and start the function, or on a part of a graph to select and edit it. After having selected an item by clicking once with the left mouse button, the right mouse button opens a context menu, which changes depending on the item selected.

2.2.2 Object Explorer

The Object Explorer is used to get an overview of what is available on the system, including data, functions, and graphs. It operates in much the same way as the Windows Explorer in that there is a tree-like structure in the left pane and the details are displayed in the right pane. The Object Explorer can be opened either from the menu or from a toolbar button. Not only can it show what objects exist, but with data, for example, the Object Explorer is used to view (browse) it and even edit it. Double-click on the data you want to view or edit and a spreadsheet will open containing the selected data. Once the spreadsheet is open, the data can be edited.

2.2.3 Commands Window

The Commands Window is actually the heart of the S-PLUS system. Every command that is performed via menus and buttons can be issued as a command directly from the prompt in the Commands Window. In addition, there are many functions that can only be run in the Commands Window. As examples, the functions of programming, personalized functions, data subsets, and logical operations are only available, or more extensively available, through the use of the Commands Window.

2.2.4 Toolbars

The main toolbar contains many familiar commands, including ***Open***, ***Save***, ***Print***, ***Copy***, ***Paste***, and ***Undo*** and is shown in Figure 2.2. In S-PLUS, however, one may also open the Commands Window and Object Explorer, open a new Graph Sheet, open the History Log, create 2D and 3D graphs, and so forth.

Figure 2.2. The S-PLUS main toolbar.

An additional feature of the toolbars in S-PLUS is that they are context-sensitive (smart). Open a Graph Sheet, for example, and extra buttons will appear in the toolbar area that are specific to Graph Sheets.

2.2.5 Graph Sheets

Graphs are drawn in windows referred to as Graph Sheets. Starting with Version 4.0, components of graphs can be edited and redefined by double-clicking on the component of interest. Labels can be changed, colors modified, axes redefined, and more, all through menus and dialog boxes available simply by clicking in the Graph Sheet. The ***Insert*** menu is useful for adding features and components to graphs which already exist in a Graph Sheet, and the ***Format*** menu is useful for changing the design of many components of an existing graph.

2.2.6 Script Window

The Script Window also consists of two panes, one on top and one on bottom. The top pane can be used as a development space in which to create or fine-tune a section of code. When the commands in the top pane are run, the output from them appears in the bottom pane. In function, the Script Window is similar to the Commands Window, the difference

being that the former runs commands in segments, whereas the latter is completely interactive and executes commands one at a time as they are input. More details of the layout of the Script Window and how it operates are provided in Section 2.16.

2.3 Getting Started with the Interface

If you want to use S-PLUS having had no introduction, use the menus and toolbar buttons. For the most part, you should find them to be self-explanatory. However, we have described a few key functions in a bit more detail to get you on your way.

2.3.1 Importing Data

You probably have your own data that you want to analyze, so the first thing you have to know is how to import it into S-PLUS. There is an ***Import Data*** facility located in the ***File*** menu. In the ***Import Data*** dialog, you will be asked for the name you want for your data (which may be more than eight characters) and have the usual Windows boxes for specifying the name and location of your data file. Pay careful attention to the type of data (***Files of Type*** pull-down menu) that you have and properly define it in the corresponding box. The data file types available are shown in Table 2.1.

Table 2.1. File types available with the import facility

File	Format		
ASCII	ASCII formatted	dBase	Excel
FoxPro	Gauss	Lotus 1-2-3	MatLab
MS Access	Paradox	QuattroPro	SAS
SAS Transport	SigmaPlot	S-PLUS	SPSS
SPSS Export	STATA	Systat	

When the spreadsheet containing newly imported data is closed, the new data automatically appear as a new entry in the Object Explorer.

2.3.2 Graphs

Graphs are created by connecting the data shown in the Object Explorer to a graphical function represented by a button in a palette. In the Object Explorer, the data are displayed with its full name, and after a click on the data object's name, the elements of the selected data set are displayed in the right-hand pane of the window. By clicking on these elements, the

set of variables to be plotted can be selected. To select a second and third variable, hold down the <Ctrl> key and click on the element.

Before actually displaying the data graphically, the graph palettes need to be opened, if they are not open yet. Open the 2D or 3D graph palette by using the appropriate icon in the toolbar. The 2D graph palette appears in Figure 2.3 and shows all of the types of 2D graphs available.

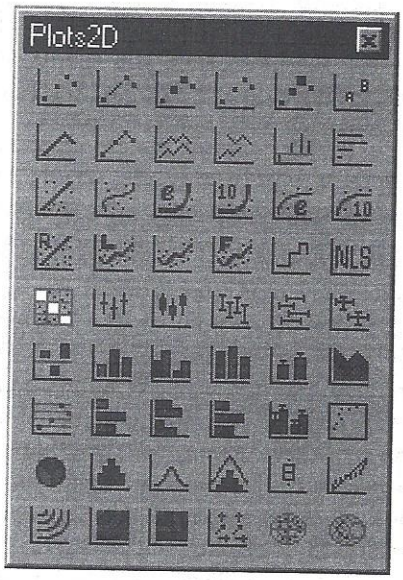


Figure 2.3. The 2D graph palette.

Once the palette is open, you might want to move the mouse over the different icons slowly. If the mouse stops for a second, the name of the method represented by the button is displayed. Select the plot type by clicking on it. S-PLUS executes the graphical method right away by using the selected data.

Note For creating a graph, the order of selection of variables is important. The variable selected first becomes the x-variable, the second variable selected becomes the y-variable, and if a third variable is selected, it becomes the z-variable. ◁

Once data have been selected, another graphical method can be applied by simply clicking on a different graph button in the palette.

Note If you click on a graphics button and a graph is set up but not drawn, this probably means that the data selected and the graph type

chosen are not compatible. Selecting two variables and clicking on a 3D graphics method is an example of such a situation. ◁

Once a basic graph has been created, it can be edited. Double-click on the element to modify, for example, the axis. A window opens and offers the possibility of modifying all the components like range, tickmarks, color, width, label, and much more.

Using the annotate palette that appears once a Graph Sheet is open, further elements such as text or graphics symbols like rectangles and arrows can be inserted into the graph.

A Graph Sheet may be printed using ***File - Print Graph Sheet...*** or saved using ***File - Save***. All that is left is to specify the destination file name and directory.

2.3.3 Data and Statistics

Data can be summarized by using graphical display techniques, but also by statistics like the mean, median, minimum, and maximum values.

There are ***Data*** and ***Statistics*** menus in the top menu bar that contain a large set of routines for manipulating and transforming a data set (***Data***) and routines to process the data (***Statistics***).

First, select the statistical method to use, then specify the data set (or vice versa) and the options you want in the menu fields. The method is applied to the data by clicking on the ***OK*** button and the results are displayed in a Report Window.

In this way, data can be summarized by looking at minimum and maximum values, at quantiles, or at the correlation between variables. Data can be processed by carrying out a t-test, a regression model, or any other method, and tick boxes can be ticked or not to choose options like paired or unpaired or the degree of a polynomial to fit.

2.3.4 Customizing the Toolbars

The toolbars and palettes can be customized to some extent. It is possible to select which palettes show up on the screen. Select ***View*** from the top menu and click on ***Toolbars***. A window opens which lets you choose which toolbars should be shown. Clicking on any of the selection fields immediately pops up or removes the corresponding palette on the desktop. These settings are stored and reused when S-PLUS is started again. Be careful which ones are chosen, remember that the Graph toolbar will open when a graph is created, and so forth.

The shapes of the palettes might not be satisfactory, depending on your screen resolution, personal preferences, and more. Clicking on one of the four edges allows you to resize the palettes, which is automatically set such that all icons fit.

Furthermore, the palette buttons can be displayed in palette form or in toolbar form. Click on the palette and drag it to the top where the primary toolbar is located. Drop it and the palette becomes a horizontal toolbar itself. Dragging the bar away from the top recreates a palette.

Adding New Palettes and Buttons

The palettes and menu buttons are open for extension. You can add menu buttons to an existing palette and create new palettes. You can generate new toolbars via the ***View*** menu by selecting the ***Toolbars*** entry and clicking on ***New***. This produces a toolbar dialog box that can be used to create a new toolbar.

You can add a new button by selecting the desired toolbar and opening the context menu with the right mouse button. The entry ***New Button*** opens up a window where you can specify the name of the button, the text to show when the mouse is over the button, and the S-PLUS function to carry out. (You will learn more about functions as you proceed.)

You can modify an existing button by selecting the button and opening its context menu (right mouse click). A property dialog displays the current settings.

2.3.5 Workspaces

S-PLUS gives you the ability to save files from different projects into different areas called workspaces. Using this feature, it is then very easy to switch between the workspaces for your various projects. All elements open during a session, such as Object Explorer and graph sheet, can be saved to a workspace such that on reopening the workspace, everything is as you left it. Since S-PLUS keeps all data items ever created, we highly recommend the use of workspaces as an easy way of organizing your work and switching between project files.

The actual operation of workspaces is quite straightforward. Under ***File-Workspaces*** you have the options of ***New***, ***Open***, and ***Save***. The available workspaces are listed beneath these options and can be clicked directly as opposed to using the ***Open*** option.

Create a new workspace by choosing the ***New*** option, and specifying the directory where it should be located. There is a box to specify ***Databases to Detach*** where the default is ticked. Leave it this way to ensure that the previously opened database is detached before the new one is opened. S-PLUS will create several subdirectories that it needs, including a _Data directory where data are stored. The new workspace is completely empty with not even a blank Object Explorer open.

Open a new Object Explorer using ***File-New-Object Explorer-OK***. We want to create some data here so that when we switch between

workspaces we can verify that they actually perform the way we want them to and keep our projects separate.

Creating Data

- Click on ***Data***, then on ***Random Numbers...***
- For ***Data Set:***, specify `test`
- For ***Sample Size:***, specify `50`
- Choose ***OK***
  – The first entry should appear in the Object Explorer.
- Click on ***Data***, then on ***Transform***
- For ***Data Set:***, specify `test`
- For ***Target Column:***, specify `Square`
- For ***Variable:***, specify `Sample`
- For ***Function:***, specify `X^2`
- Click on ***Add***
  – `Sample^2` appears in ***Expression:***
- Choose ***OK***
  – Can see that the data sheet has been opened and contains the variables `Sample` (the original Gaussian distributed data) and `Square` (the square of the variable `Sample`).

Save the new workspace (***File-Workspace-Save...***) accepting the default directories and file names as prompted (these have already been specified by you earlier). The name for the workspace is taken from the name of the directory where it was originally created. For example, if the workspace was created in the directory `c:\sbook`, then its name would be `sbook`.

Switch back to the default workspace (named ***Workspace***) by choosing it from the list under ***File-Workspace***. Does the newly created data set (`test`) appear in the Object Explorer of this workspace? Switch back to the new workspace. Is `test` the only data set that appears in the Object Explorer here?

Note Warning messages may appear about duplicate databases, detaching, attaching, and position numbers. These messages may be ignored but serve to point out that the database for workspace A is not actually removed when you switch to workspace B, but is simply moved further down the

search path. This action has two implications. Data sets from a non-active workspace cannot be seen through the Object Explorer but can be accessed. If a data set in the active workspace has the same name as one in a non-active workspace, the one in the active workspace will be accessed. ◁

In summary, the workspace feature is a convenient way to help keep yourself organized. We highly recommend that you begin using it right from the start. For practice with workspaces, try creating one for each chapter of the book and keeping chapter-specific data sets in the respective workspaces.

2.4 Detailed Use of the Windows Interface

The previous section provides a quick introduction to some of the basic concepts involved with using the graphical user interface (GUI) developed for the Windows version of S-PLUS. Brief descriptions of how to perform several tasks are given, but without much in the way of specifics. The details of how to understand and work with the GUI are given throughout this section.

We cannot emphasize enough that to really profit from this book, you need to be running S-PLUS and following the examples while reading the book. Only by trying something on your own will you really learn it to the extent that you could do it on your own. We have left a lot of tasks to be done by the reader (you!), so as to build on the concepts immediately.

2.5 Object Explorer

The Object Explorer plays the same role in S-PLUS as the Explorer does in Windows in that the "contents" are displayed in a tree-like diagram on the left-hand side of the screen and the details on the right-hand side. Similar or related objects can be stored in folders, giving structure to the way one works. When an object is put into a folder, a link (shortcut) is created such that the actual location of the object is not changed, just its virtual location. Folders can be dragged and dropped between Object Explorers, between Explorer pages within an Object Explorer, or onto the toolbar as a button.

Objects can be classified into three types: computational engine, interface, and document. Computational engine objects include data frames, matrices, lists, functions, and others, and are objects containing data or the functions that are used to process the data. Interface objects are a type that relate to communication with the system and include search paths, menu items, toolbars, dialogs, and the like. Document objects refer to out-

put from the system and include Graph Sheets, Reports, Scripts, and the like.

All of the object types mentioned can be saved in folders in the Object Explorer. Folders can be organized according to object type, data set, project, model class, or in any way that seems to be helpful.

Open the Object Explorer if it is not already open by clicking the ***Object Explorer*** button on the standard toolbar. It should look like the one in Figure 2.4.

If you have not yet worked with S-PLUS the Object Explorer will be empty and if you have, you will see different object types in the left pane and their details in the right pane.

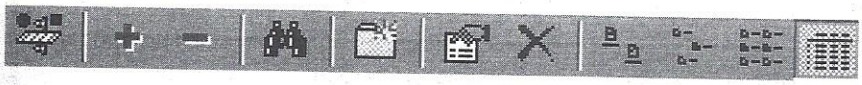


Figure 2.4. The Object Explorer toolbar.

We want you to follow an example and to get practice exporting a data set so we will take a built-in S-PLUS data set, export it, and then import it before we begin analyzing it. First, we have to use the Object Explorer to find the data set.

S-PLUS uses a system of "positions" that are searched sequentially until the desired object is found. A message that an object has not been found is only issued once all the positions have been searched. Position number 1 is the current working directory. By default, the Object Explorer is used to show the objects in the current working directory. Other positions are used to store general functions, statistical functions, other types of data sets, and the like. The data we want to use are a few positions down in the list because the built-in data sets are not meant to be worked on in a working directory, merely read from a safer location.

The search path lists all the positions currently available to S-PLUS. By examining the search path, we can find the built-in data sets, locate the one we want, and write it to the hard disk.

S-Plus Data Sets

- Click beneath the object types in the left pane of the Object Explorer (or else you will get a subfolder and not a folder)
- Create a new folder by clicking on the ***New Folder*** button on the Object Explorer toolbar
- Type in the name for the folder (use "Examples") and hit ***Return***
- Right-click on the folder

- Choose ***Folder...***
- In the ***Data Objects*** section, tick ***Data***
- Choose tab for ***Advanced*** (see Figure 2.5)
  – File locations in the ***Databases:*** section specify the Search Path. Notice that the first database matches the file location that appears in the Commands Window as the current working directory.
- In the ***Databases:*** section, choose *SPLUSpath*\STAT_DATASET where *SPLUSpath* is the path where S-PLUS is installed (typically C:\Program Files\sp2000)
  – This choice should be in position 6 (i.e., sixth in the list)
- Choose ***OK***
  – A "+" appears next to the ***Examples*** folder and the appearance of icon changes
  – List of all the data sets appears in right-hand pane (see Figure 2.6)

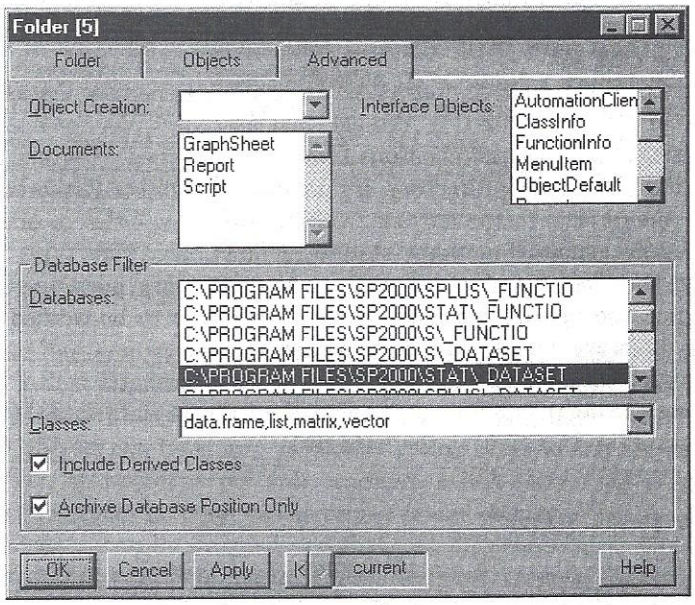


Figure 2.5. Filtering data sets for Object Explorer.

We have just used the ***Object Explorer*** to look at the search path and put the contents of one of the positions into a folder so that we can directly access the contents. The Object Explorer can also be used to find more information about the contents of a particular folder. The data set

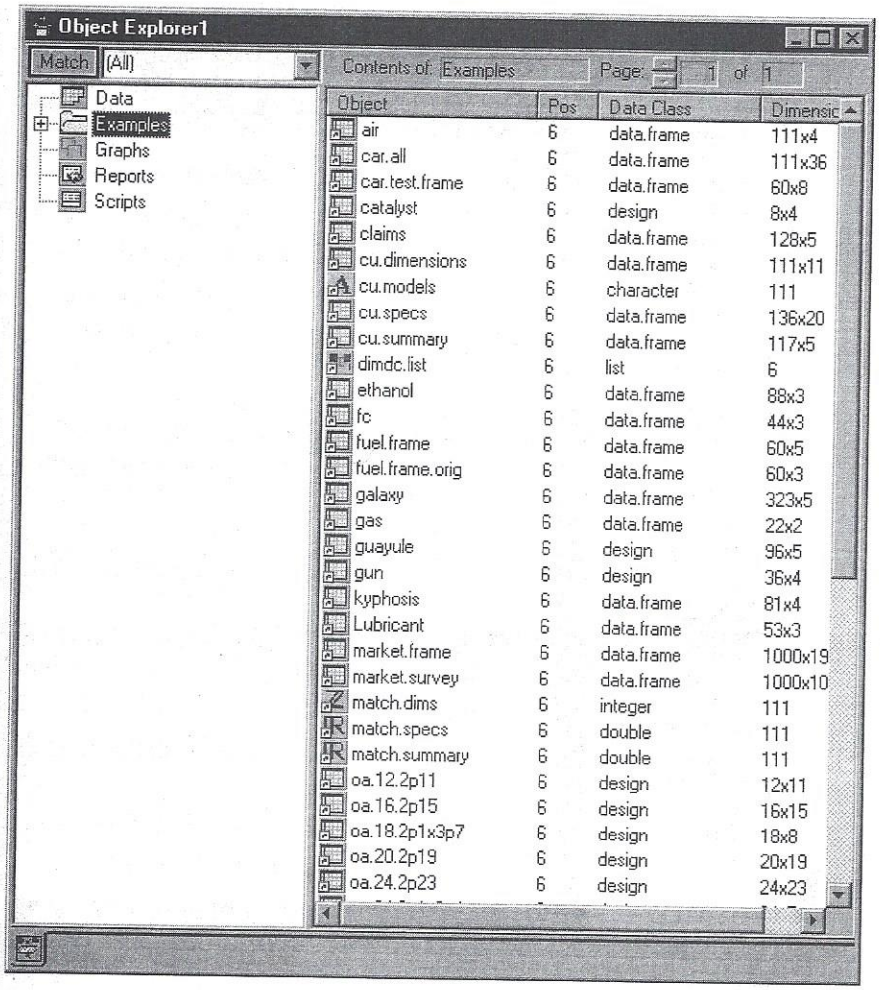


Figure 2.6. List of data sets displayed in the Object Explorer.

we want to work with is called Puromycin, can you locate it using the Object Explorer? What happens if you now double-click on Puromycin in the ***Object Explorer***? Do you see the similarity of the ***Object Explorer*** to the Explorer in Windows?

Double-clicking on Puromycin runs the application which, in this case, opens a spreadsheet-like editor that can be used to view and edit the data. By single-clicking on Puromycin in the left-hand pane of the ***Object Explorer***, a list of its variables is shown in the right-hand pane. The Puromycin data set contains the three variables, conc=concentration, vel=velocity, and state, as shown in Figure 2.7.

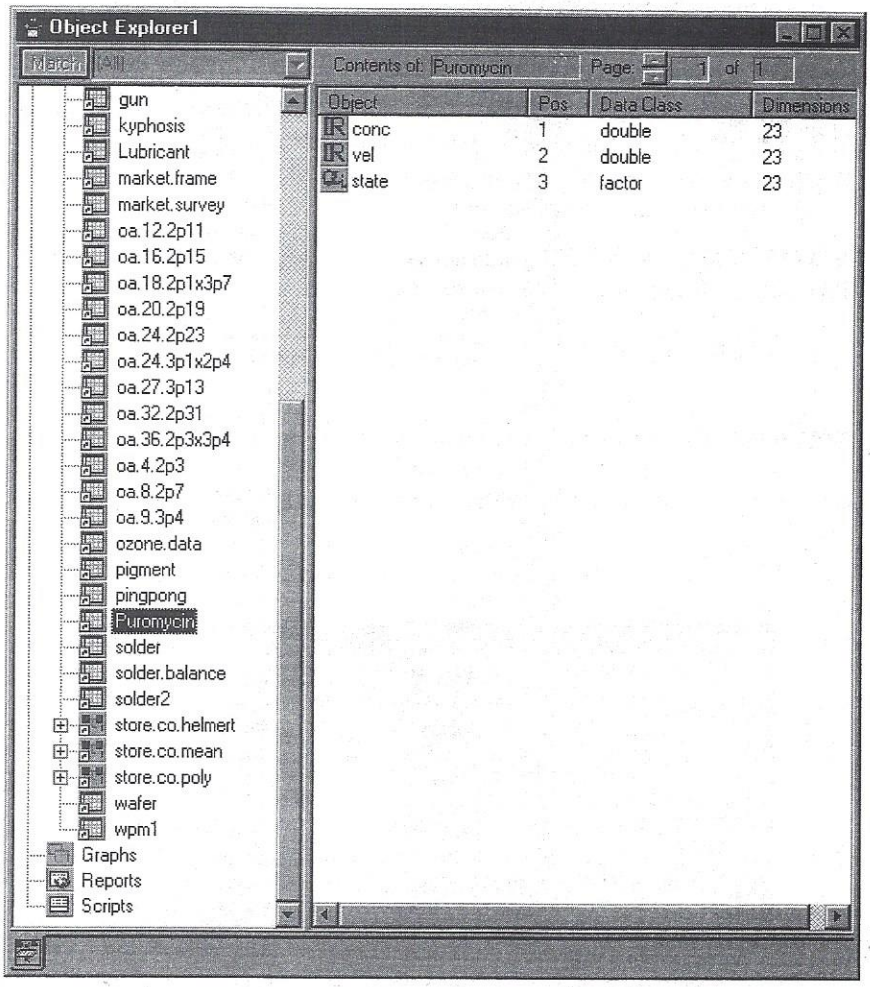


Figure 2.7. Details of the Puromycin data set.

2.6 Help

At some point you will want to use (or have to use) the Help system in S-PLUS and we have reached a point where it is possible to explain this simple procedure. Suppose we feel that the three variable names in the Puromycin data set and their corresponding data types are not terribly informative and we want to know more about it. We can invoke the Help system to find out more.

Using the Help System

- Click on the Help main menu
- Choose the ***S-Plus Help*** category
- Choose ***Index*** tab
- Type "pur"
  – Puromycin is highlighted
- Choose ***Display*** at bottom
  – Window explaining the Puromycin data set appears

The Help system can be used in the same way to find information on (almost) any other topic.

2.7 Data Export

As promised, we're going to export the Puromycin data, import it again, and then work with it. The process is all menu-driven.

Exporting Data

- Click on Puromycin in the left-hand pane of the ***Object Explorer***
- Click on the ***File*** main menu
- Choose ***Export Data***
- Choose ***To File*** ...
- Specify location for file to go using ***Save in:***
- Specify the file name (e.g., puro) using ***File name:***
- Specify file type using ***Save as type:***
  – We chose ***SAS Files (*.sd2)*** to make another point later on and suggest that you do as well (you don't need SAS)
  – Puromycin is already highlighted in ***Export From Data Set:*** (see Figure 2.8)
- Choose ***OK***

The file should now be created with the name (puro) and type (.sd2) that we specified, residing in the directory of our choice. Use the Windows

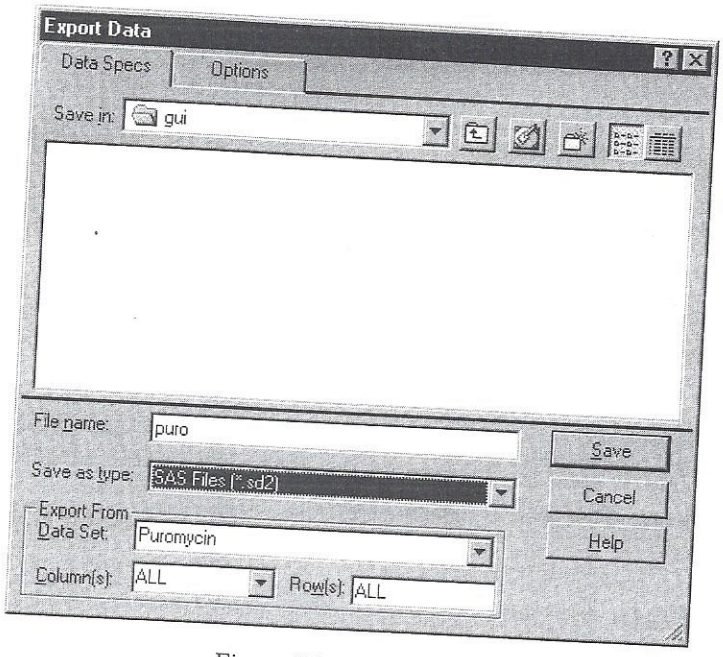


Figure 2.8. Data export.

Explorer to verify that the desired data set has been created. Save the same data in different formats and open them with the appropriate program, for example, using Excel for *.xls files. as an Excel file and open the file in Excel.

2.8 Working Directory

When working in a Windows environment, it is often convenient to set up a different "workspace" for each separate study on which you are working. Workspaces can be set up in two different ways: as shown in Section 2.3.5, or by defining multiple S-PLUS icons on the desktop. The advantage of the former approach is that you never have to leave S-PLUS to switch between project specific directories, whereas the advantage of the latter approach is that the danger of inadvertently working in the wrong directory is minimized.

The method of defining and switching between workspaces is easy, flexible, and orderly (see p. 14), but, as people work in different manners, the method of defining a working directory and creating an S-PLUS icon will be shown here.

Defining a Working Directory

Create a copy of the S-PLUS shortcut icon by:

- Right-click on the ***Start*** button in the Windows desktop (lower left)
- Choose ***Open***
- Double-click on ***Programs***
- Double-click on the S-PLUS folder
- Depress the `<Ctrl>` key, Click on the S-PLUS icon, and Drag it to the Windows desktop
- Release the mouse and the `<Ctrl>` key and you have a copy of the icon
- Right-click on new icon
- Choose ***Properties***
- Choose tab for ***Shortcut*** sheet
  - ***Target:*** is highlighted and shows the command for running S-PLUS
- After the existing text that is already in the box for ***Target:***, add ***S_PROJ=****pathname*
  - Use the path where you saved the Puromycin data.
  - The *pathname* must be enclosed with double quotes if there is a space in it.
- ***Start in:*** should be blank (see Figure 2.9)
- Choose ***OK***

You can now start S-PLUS by double-clicking on the new icon, which will have the effect of working in the directory specified in the path name. If this is the first time S-PLUS has used this directory, you will get the message, "Invalid Path - The path found for s_proj 'pathname' does not contain folders 'Data' and/or 'Prefs' for project related files. Specify another path or accept this one and the necessary folders will be created." At this point, you can specify a new path, or accept the choice and these two folders will be created for you (Choose ***OK***). Open the ***Object Explorer*** and verify that the new S-PLUS icon is using the database (empty) defined by the new path name (we did this earlier).

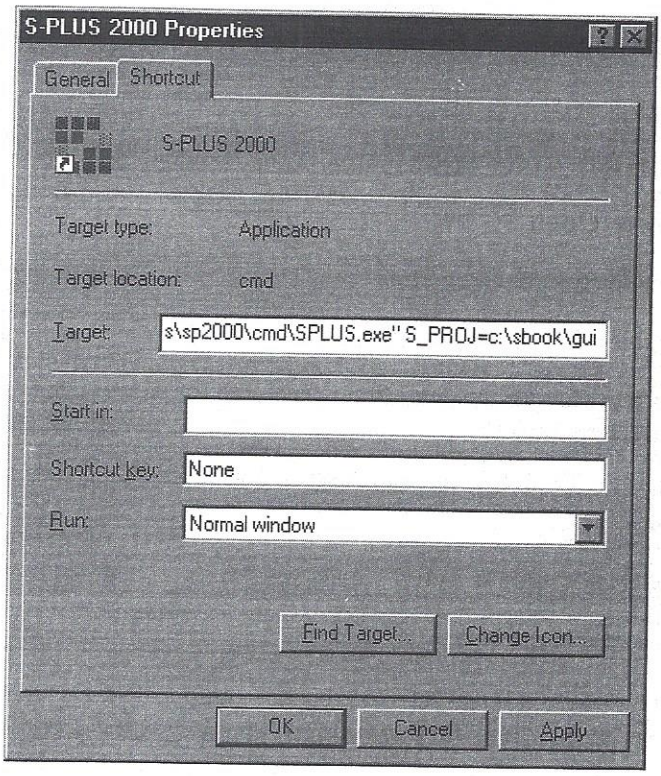


Figure 2.9. Defining a working directory.

2.9 Data Import

After this preparation, we are finally ready to look at how to read a data set into S-PLUS. The procedure is essentially the same as with exporting data and works in the following manner.

Importing Data

- Choose ***File*** from the main menu
- Choose ***Import Data***
- Choose ***From File*** ...
- Define ***Files of type:*** to be ***SAS Files (*.sd2)***
  – Check that location (path name) and file name appear as desired
- Click on file name to select it (highlighted)
  – ***Import to Data Frame:*** has puro in it automatically so that the data set will be stored in the data frame puro

- Choose ***Open***

As soon as ***Open*** has been chosen, a data window should open containing the data set displayed in typical spreadsheet fashion as in Figure 2.10. It contains 23 rows and 4 columns (variables), 1 column more than the original data frame! What has happened? In addition, the variable names are now in all capitals and S-PLUS is case-sensitive.

puro

	1	2	3	4
	ROWNAMES	CONC	VEL	STATE
1	1	0.02	76.00	treated
2	2	0.02	47.00	treated
3	3	0.06	97.00	treated
4	4	0.06	107.00	treated
5	5	0.11	123.00	treated
6	6	0.11	139.00	treated
7	7	0.22	159.00	treated
8	8	0.22	152.00	treated
9	9	0.56	191.00	treated
10	10	0.56	201.00	treated
11	11	1.10	207.00	treated
12	12	1.10	200.00	treated
13	13	0.02	67.00	untreated
14	14	0.02	51.00	untreated
15	15	0.06	84.00	untreated
16	16	0.06	86.00	untreated
17	17	0.11	98.00	untreated
18	18	0.11	115.00	untreated
19	19	0.22	131.00	untreated
20	20	0.22	124.00	untreated
21	21	0.56	144.00	untreated
22	22	0.56	158.00	untreated

Figure 2.10. Data window from Importing Data.

You'll note that the first column is called ROWNAMES and that the values go sequentially from 1 to 23. This column corresponds to the internal row counter in SAS (_N_) which was not part of the original data frame. The extra column we encountered is a special case of SAS, but there may be other file types which insert other special columns. This is a nice place

to remind you that you should always check the data to make sure that there are no unpleasant surprises. We now want to remove the extraneous column.

Deleting Data Column(s)

- Click on the first column of the Data Sheet (the one to be deleted)
- Choose ***Data*** from main menu
- Choose ***Remove***
- Choose ***Column*** ...
- Define the column(s) to be removed using the pull-down menu for ***Columns:***
  - Remember to use ***Ctrl***-Click to select noncontiguous column names
- Choose ROWNAMES (see Figure 2.11)
- Choose ***OK***
- Close ***Data Window***

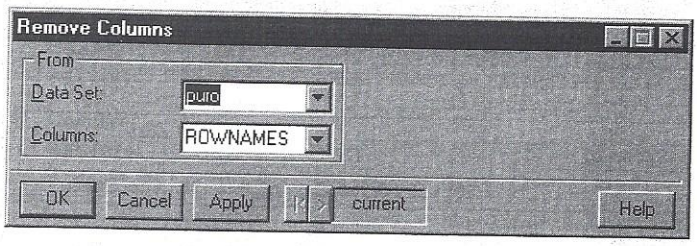


Figure 2.11. Deleting a data column.

Use the ***Object Explorer*** to open the data frame again to verify that only the three original variables are in it. Data frame is the term used by S-PLUS to denote a rectangular collection data, possibly including variables with different data types (e.g., character and integer). Many data sets typically encountered are actually data frames.

2.10 Data Summaries

Now that we have a data frame in the ***Object Explorer***, we want to "look" at the data in some sense. One of the easiest ways of getting a sense

for your data is to generate simple summary statistics (mean, standard deviation, etc.). The following command box shows how this is done.

Generating Summary Statistics

- Click on puro in left-hand pane of ***Object Explorer*** to select it
- Choose ***Statistics*** from the main menu
- Choose ***Data Summaries***
  - Options are ***Summary Statistics***, ***Crosstabulations***, and ***Correlations***
- Choose ***Summary Statistics*** ...
  - Summary statistics dialog opens to a data window.
  - Puro should be chosen as the default data frame (see Figure 2.12).

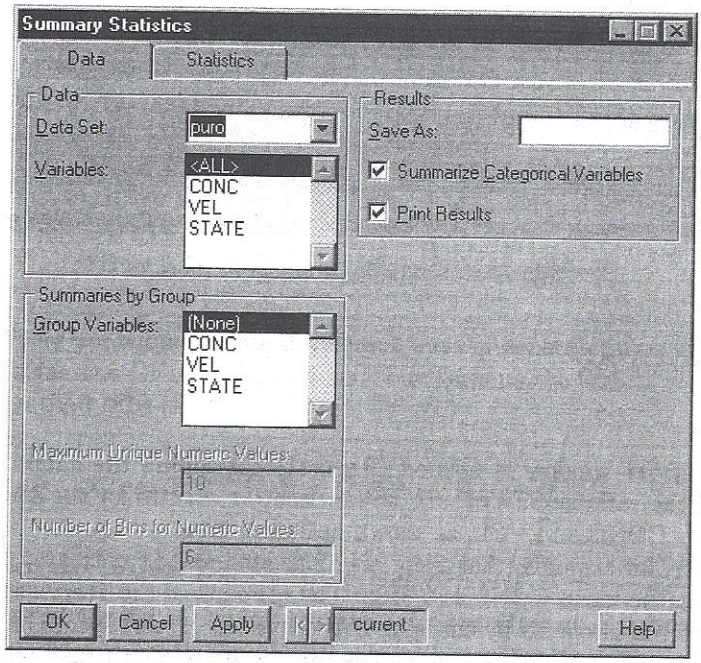


Figure 2.12. Summary Statistics data dialog.

- Choose ***All Variables*** (default)
- Choose ***No Grouping Variables*** (default)
  - Can tick or untick various statistics to customize output.
  - Can save results by supplying a name using ***Save As***.

- Choose ***Statistics*** tab at the top of dialog window.
  - Can choose specific statistics to be calculated (see Figure 2.13).

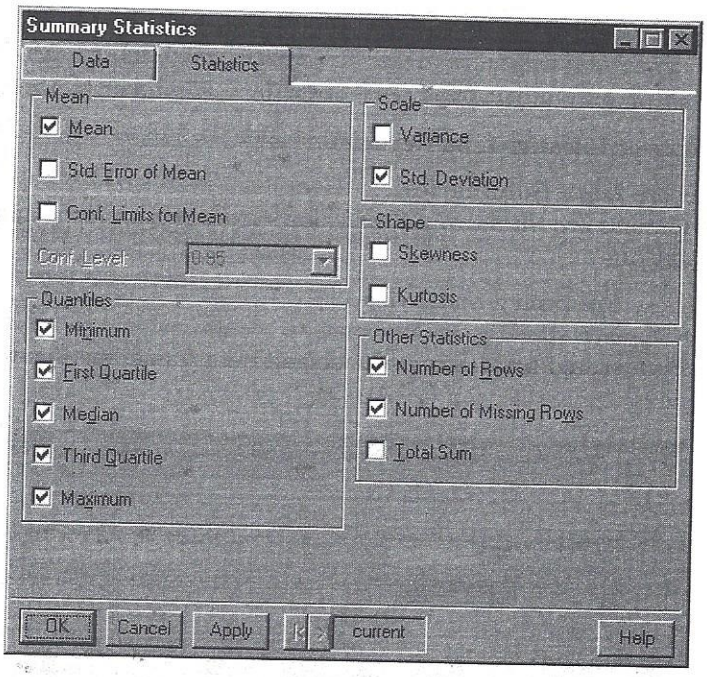


Figure 2.13. Summary Statistics dialog.

- Choose ***OK***

A ***Report*** window is opened (Report1) containing the count of each level of the variable STATE and the summary statistics for the continuous variables CONC and VEL, as shown in Figure 2.14.

We see, for example, that there are 12 treated and 11 untreated patients and that the mean CONC is 0.3122. The contents of the ***Report*** window can be saved into a file by choosing ***File-Save As*** ... and specifying a path and file name. The default file type is .srp (S-PLUS report file), which can be opened by any standard editor or from within S-PLUS.

2.11 Graphs

The problem with data summaries is that you don't actually "see" the data itself. For a more visual summary of our data, we need to construct a

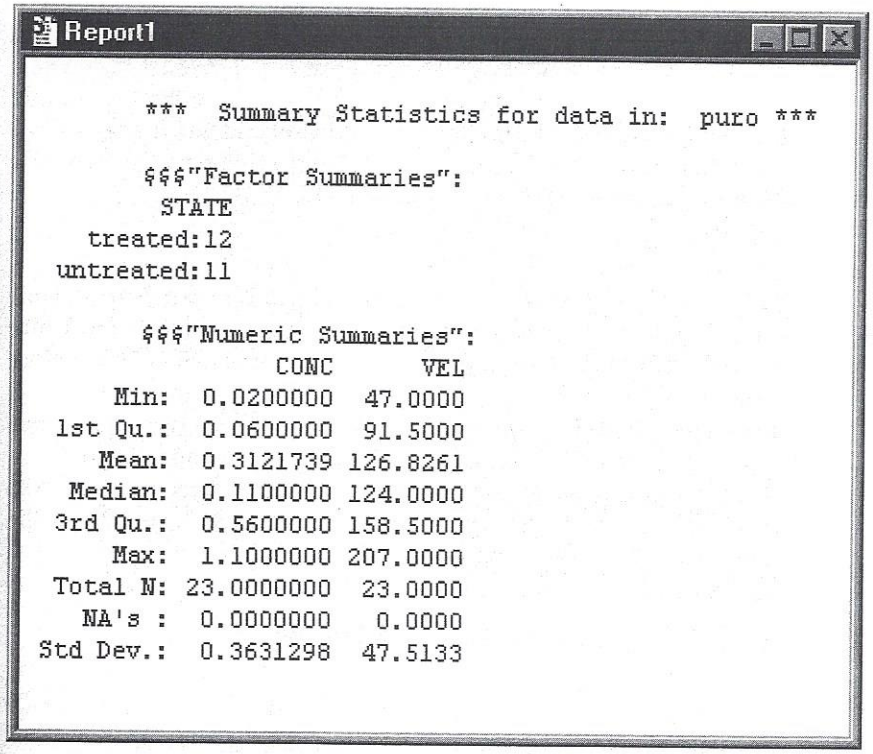


Figure 2.14. Summary Statistics Report window.

graph. There are actually several different ways of constructing a graph in S-PLUS using the GUI (menu buttons), but perhaps the easiest and most flexible method is to use the palettes.

Creating a Graph Using a Palette

- Open the ***Object Explorer*** if it is not already open
- Expand the list in the left-hand pane of the ***Object Explorer*** such that the three variables of puro are visible in the right-hand pane
- Click on CONC to highlight it
- ***CTRL***-Click on VEL so that both variables are highlighted
  – The order in which the variables are chosen (clicked) is important. The first will be used for the x-(horizontal) axis and the second for the y-(vertical) axis.
- Click on the 2D palette button on the main toolbar

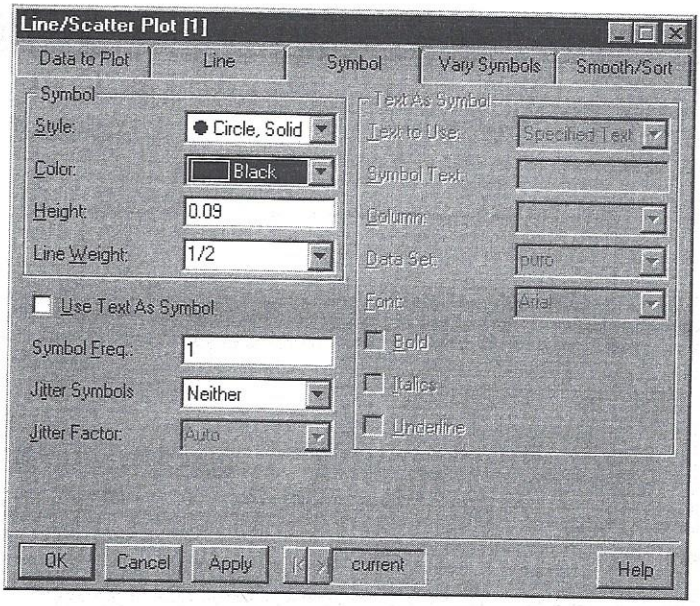


Figure 2.15. 2D Plot dialog (symbol).

Changing Axis Labels

- Click on the default x-axis label (CONC)
  - It is now surrounded by green knobs.
- Choose one of the following methods:
  - (A) One more Click allows you to change the label by simply typing over the default (Default is automatically taken as the column name and is denoted '@Auto')
  - (B1) Right-click on the default label (dialog appears)
  - (B2) Choose ***Edit In-place...*** (see Figure 2.17)
- Change text to "Concentration"
- Click outside of box to finish without hitting ***Return***

The label has now been changed and the text is better now; however, the font is rather small. We can easily change this.

Changing the Font

- Right-click on axis label text

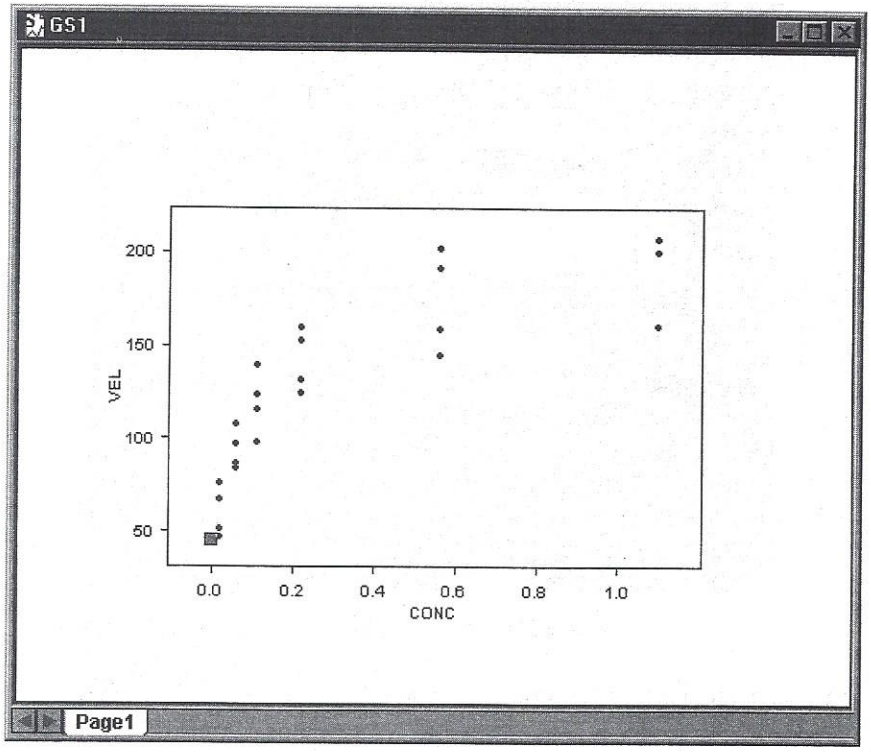


Figure 2.16. Scatterplot of Puromycin data.

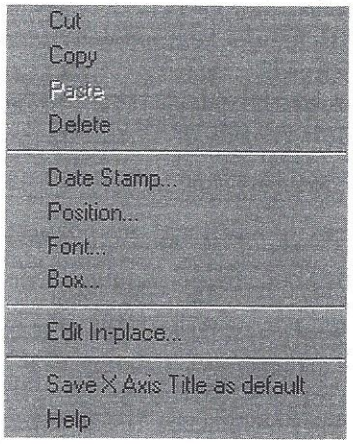


Figure 2.17. Axis dialog box.