Statistics and Computing

Statistics and Computing

Brusco/Stahl: Branch-and-Bound Applications in Combinatorial Data Analysis.
Dalgaard: Introductory Statistics with R.
Gentle: Elements of Computational Statistics.
Gentle: Numerical Linear Algebra for Applications in Statistics.
Gentle: Random Number Generation and Monte Carlo Methods, 2nd Edition.
Härdle/Klinke/Turlach: XploRe: An Interactive Statistical Computing Environment.
Krause/Olson: The Basics of S-PLUS, 4th Edition.
Lange: Numerical Analysis for Statisticians.
Lemmon/Schafer: Developing Statistical Software in Fortran 95
Loader: Local Regression and Likelihood.
Ó Ruanaidh/Fitzgerald: Numerical Bayesian Methods Applied to Signal Processing.
Pannatier: VARIOWIN: Software for Spatial Data Analysis in 2D.
Pinheiro/Bates: Mixed-Effects Models in S and S-PLUS.
Venables/Ripley: Modern Applied Statistics with S, 4th Edition.
Venables/Ripley: S Programming.
Wilkinson: The Grammar of Graphics.

Andreas Krause
Melvin Olson

The Basics of S-PLUS

Fourth Edition

Springer

Andreas Krause
Novartis Pharma AG
Basel 4002
Switzerland

Melvin Olson
Novartis Pharma AG
Basel 4002
Switzerland

Library of Congress Control Number: 2005926768

ISBN-10: 0-387-26109-5
ISBN-13: 978-0387-26109-6

e-ISBN: 0-387-22708-3

Printed on acid-free paper.

Printed in the United States of America. (SBA)

9 8 7 6 5 4 3 2 1

springeronline.com

Preface

This is now the fourth edition of "The Basics of S-Plus" since 1997. S-PLUS saw a steady growth in popularity, and it established itself in many educational and business places as a major data analysis tool. S-PLUS is valued for its modern, interactive data analysis environment, whether it is the primary system or a complement to other standards like SAS (the latter is in particular true for the industry we work in, pharmaceuticals).

We have followed the various releases with new editions of our book, introducing over time major changes like the incorporation of S Version 4 (the underlying language), Trellis graphs, a graphical user interface, in particular for the Windows operating system, and a chapter on R and its differences to S-PLUS (that are minor for the material covered in this book). This edition is an update from edition 3 to cover new functions and features of S-PLUS Version 7.0 (working from the beta release for MS Windows and Linux), adding more practical tips and examples, and correcting a few mistakes.

We are very grateful to all our readers, in particular those sending us suggestions, comments, and any other kind of feedback. You will see some of these reflected in the book.

The number of new books appearing on S-PLUS and its counterpart, R, continues to grow substantially each year. If you want to get an impression of the books available, Insightful maintains a list of S-PLUS references on their web site, http://www.insightful.com/. The list currently comprises about 50 books and manuscripts. You can also search for the string "Splus" or "S-Plus" (without quotes) in an online bookstore, such as http://www.amazon.com/.

Some of the books on S-Plus are "Analyzing Medical Data Using S and S-Plus," by Everitt and Rabe-Hesketh; "Regression Modeling Strategies," by Harrell; "Bootstrap Methods and Permutation Tests" by Hesterberg, Monaghan, et.al., "Applied Statistics in the Pharmaceutical Industry," edited by Millard and Krause; "Mixed Effects Models in S and S-Plus," by Pinheiro and Bates; "Modeling Survival Data: Extending the Cox Model," by Therneau and Grambsch; and the fourth edition of "Modern Applied Statistics with S-Plus" as well as the second edition of "S Programming," both by Venables and Ripley.

This book is intended to provide a head-start into the world of S-Plus (and serves as well as an introduction to R). We are going to introduce you to S-Plus by taking you through interactive sessions, entering commands and later analyzing data, pointing out tips and pitfalls, and complementing each session with exercises and detailed solutions.

If you have never worked with S-Plus before and want to get a first impression, take a look at "A First Session," "A Second Session," and "Exploring Data" to get an idea of the philosophy of the programming language. If you intend to work with point-and-click graphical user interfaces, take a look at "Graphical User Interfaces." We are convinced though that if you want to work permanently with S-Plus, you will need to use the command line and appreciate it soon.

The book originates from lectures such that each chapter can serve as the basis for a 90-minute lecture. The exercises reinforce the material covered and sometimes point out a few more details. It should be interesting to work out a solution to an exercise before comparing it to the solution given here. There's never a single solution.

Finally, we would like to thank all those who have directly or indirectly contributed to this edition and previous ones. For this edition, Chyi-Hung Hsu, Michael Merz and José Pinheiro gave very helpful comments on the manuscript. John Kimmel at Springer–Verlag is simply a great person to work with, dedicated and professional. David Smith, Michael O'Connell and others at Insightful, Inc. (Seattle) continue to be supportive partners by providing us with alpha and beta releases of upcoming S-Plus versions. Following the discussion forum s-news provides an insight into frequently asked questions and problems that S-Plus users face. And last but not least, various people from all over the world supplied comments on earlier editions of the book, pointed out possible improvements and errors and provided a good motivation to revise and update the book with the current edition 4 – which you happen to be reading right now.

We are always interested in any kind of comment that you might have about the book. You are welcome to contact us by sending an E-mail to Andreas.Krause@Novartis.com and Melvin.Olson@Novartis.com. Note though that we will not be able to provide general support, like answering programming questions. For these purposes please refer to the sections where we describe how to get help and support (pp. 413ff. and 422ff.). You might also want to check if your institution has a contract with Insightful that covers support.

Finally, a Web page is set up to accompany this book:

http://www.elmo.ch/doc/splus-book/

It provides information about the book, links to relevant Internet pages, and a list of errata (which we are confident will be fairly short).

Basel, Switzerland
April 2005

Andreas Krause and Melvin Olson

Contents

Figures

Tables

1 Introduction

Over the years, the S language and S-Plus have undergone many changes. Since its development in the mid-1970s, 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, such as 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 modern and flexible language and high quality graphics generation.

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 1970s after the release of Tukey's book (1977) on the subject. Most of the statistical functionality was added later, and many statistics routines, such as 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 1970s and early 1980s, mainly by Rick Becker and John Chambers. Some years later, Allan Wilks joined the core team. Several other people have been involved in the project. Becker (1994) describes in great detail the foundation and development of S until 1994 (S Version 3). S Version 4 was developed by John Chambers at Bell Laboratories, Lucent Technologies.

The year 1976 can perhaps be viewed as the founding year of S, the year in which the 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. From 1981 to 1993, the source code was available to interested people outside Bell Labs.

The next years revealed a strongly 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 such as 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). StatSci became the Data Analysis Products Division of MathSoft in 1994. When the company became independent again in 2001, it received a new name: Insightful Corporation.

The company 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 such as 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 PCs 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.

Insightful released S-Plus Version 6 for UNIX in 2000 and for Windows in 2001. For the first time, the two systems are built from the same code base. The UNIX system has a graphical user interface based on Java and supports some Windows-specific elements such as reading Excel data files and writing graphs in WMF (Windows MetaFile) format.

Version 6.2 introduced an important feature for automated batch runs, the so–called verbose logging. By calling S-Plus with the switch /SBATCH an exhaustive log file is created that facilitates scanning for possible problems in batch runs.

Version 7.0 of S-Plus is about to be released by the time of writing this edition of the book (early 2005). We have an alpha and a beta version of S-Plus 7.0 for Windows and a beta version for Linux available. The new major version includes a library for handling big data sets without keeping it entirely in memory. The advertised limits are “no limits in the number of rows” and “tens of thousands of columns.” The library comprises functions for extracting simple statistics from columns and for calculating statistical models such as linear regression, generalized linear models, principal components, and clustering on big data sets. The GUI is revised and now based on a new toolkit, Eclipse. It comes with a workbench to develop S-Plus programs, track changes, and source control facilities. These new functionalities are only available for the newly introduced "Enterprise Developer" edition though and not part of the "Professional Developer" edition.

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 publicly available. In the remainder of the book, we will use S-Plus as the standard reference.

In addition to S-Plus, the R project started in 1997. R is a software that is "not unlike S" and freely available. It has become increasingly popular in recent years since a core developer team continuously improved and developed the system further. A fairly large group of supporters around the world contributes libraries, in very much the way S started out a long time before. This group of people includes many of the contributors of S-Plus libraries. We will point out some of the similarities and differences between the two systems in a separate chapter.

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 internal workings of S-Plus can be found in Sections 13.1 and 13.2.

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

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 Chapter 15 (page 417).

S-Plus has minimum requirement specifications regarding main memory and hard disk size. The requirements depend on the particular version and operating system. For a full installation, 300MB will be sufficient (including PDF manuals, etc.). For reasonable performance, 128MB of main memory (RAM) is a lower limit (per simultaneous user). Large data sets or intensive computations will require more memory.

Table 1.1. S-Plus system requirements.

Operating systems (hardware)	System requirements (recommended minimum)
Windows 2000, Windows XP Home Edition, Windows XP Professional Edition, Windows 2003 Server (on Intel platforms)	Pentium III, 512MB RAM, 450MB disk space
Solaris 2.8, 2.9 (SPARC processors)	160 MB RAM, 340 MB hard disk space
Red Hat Enterprise Linux WS 3.0, kernel 2.4.21, gcc version 3.2.3, glibc version 2.3.2 (Intel/AMD x86, 32 and 64 bit)	160 MB RAM, 340 MB hard disk space
AIX 5.1, 5.2 (POWER processors, IBM)	160 MB RAM, 340 MB hard disk space
HP-UX 11.0, 11i (PA-RISC processor, HP)	360 MB RAM, 340 MB hard disk space

The systems listed here are the officially supported systems for version 7.0. S-Plus may work other systems as well (for example, other Linux distributions). See the Insightful web page, http://www.insightful.com/, for further information.
The minimum RAM requirements for Solaris, Linux, AIX, and HP-UX can be reduced to about 60 MB if S-Plus is used from the command line and the Java GUI is not used.

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 speed-up 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 (such as Word and Powerpoint) is smoothly integrated. Graphs can be put into Word or Powerpoint documents by using copy and paste, and if a large number of graphs is generated, the Powerpoint Wizard creates a Powerpoint presentation of all S-Plus graphs with a few mouse clicks.

S-Plus for UNIX has a graphical user interface, too, but it differs slightly from the look and feel under Windows (due to its implementation in Java). It also lacks the toolbars, and graphs cannot be edited.

In essence, it is straightforward to generate S-Plus graphs on Windows and include them in Windows document, and the same is true for S-Plus graphs on Unix and inclusion in documents that reside on the Unix machine. It requires a little extra effort if S-Plus runs on one system and the documents are edited on the other system. Unless you create a larger number of graphs it should not be too much extra work though. We will cover graph generation and inclusion in documents later on in more detail.

Under S-Plus for Windows, the standard graphics device is a so-called graph sheet. The graph sheet can hold multiple pages. If some S-Plus code generates multiple graphs, they are shown on different pages and the user can switch between them by mouse clicks. S-Plus for UNIX does not have the graph sheet device. If a new graph page is created on the motif device, the page is cleared and the graph lost. The newly introduced Java device has that capability of holding multiple pages, too.

S-Plus for UNIX allows easy integration of C and Fortran code by basically automatic compilation of a “shared object file” that is linked to the system at runtime. S-Plus for Windows allows one to dynamically link libraries (DLLs) with some more manual effort.

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

S-Plus for UNIX has a special release called “Application Server.” This new implementation allows access to S-Plus via the Java interface (including the command line accessible from a Java interface that looks similar to the original command line). The application server allows users to access S-Plus using a Web browser.

There are more differences, but most of them are not directly visible to the user. With very few exceptions, S-Plus code that runs under S-Plus for Windows will also run under S-Plus for UNIX, and vice versa.

1.3 Notational Conventions

Before starting let us introduce a few notational conventions 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: `>`. If a single command extends over one line of input, S-Plus will determine that the command is incomplete and ask for completion. The prompt changes to the plus sign (+) and you can enter the remainder of the command. A preview of this is shown below.

```
> 3+(4*2
+ )
```

We mostly omit the continuation prompt, as the code is easier to read and so that the + sign is not accidentally typed in as a “plus” operator. For longer program codes, we omit the prompt, too, as we do not assume interactive input but the use of an editor instead. S-Plus objects or expressions are written in a special font as with the example `print`.

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

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 (displayed by S-Plus but omitted in this monograph)
`Commands`	S-Plus commands are shown in typewriter font
No prompt	For calls to the UNIX or DOS shell
`#`	Comment symbol indicating start of a comment
placeholders	Need to be replaced by an appropriate expression, such as *filename*, which needs to be replaced by a valid file name
Ctrl-C	Press and hold down the Ctrl key and press C
Menu	Menu entries and buttons are referred to in this font
Note	Notes point out something important, such as an example, an application, or an exception. The end of a note is marked by the symbol ◁

Finally, if S-Plus prints an object with dimensions – a vector or a matrix, for example – each row is preceded by the index of the following element, as in the following example where `[1]` indicates that the first element, 6, has the index 1.

```
> 6:9
    [1] 6 7 8 9
```

We mostly omit these indices to facilitate reading the output.

2
Graphical User Interface

2.1 Introduction

S-Plus has an excellent graphical user interface (GUI) that allows the user to access its functionalities by pointing and clicking with the mouse. This convenient way of working with S-Plus was previously only available for the Windows platform but has now been added for the UNIX/Linux platform as well.

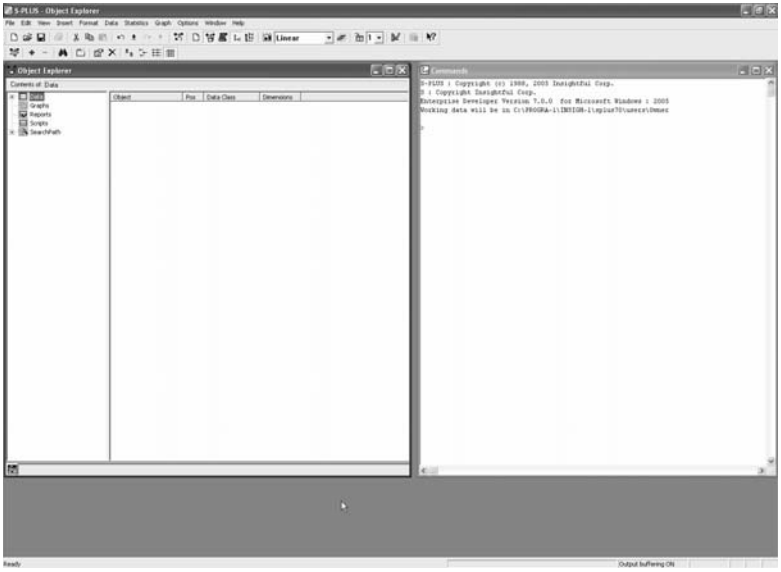
The advantage of the GUI for the novice of S-Plus is that you don't have to know the syntax of S-Plus to get started. All you need is a little familiarity with typical Windows software and a data set in some sort of standard format.

The descriptions in this chapter are, for the most part, based on S-Plus for Windows. Users of UNIX/Linux will benefit from this chapter, despite it being aimed at the Windows user of S-Plus, as we will show typical applications and examples and point out major differences between the two operating systems. The chapter will end with a section dedicated solely to the GUI available with UNIX/Linux.

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 by 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 but 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]



```
S-PLUS : Copyright (c) 1988, 2005 Insightful Corp.
S : Copyright Insightful Corp.
Enterprise Developer Version 7.0.0  for Microsoft Windows : 2005
Working data will be in C:\PROGRA~1\INSIGH~1\splus70\users\Owner

>
```

Figure 2.1. The S-Plus screen and its components: the Object Explorer, the Commands window, and the toolbars.

The general layout of the S-Plus system is similar to that of 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.

[1] According to the version of S-Plus being used, some components might not be present or look slightly different.

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 Data***, ***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., ***2D Plots***) 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. Clicking once with the right mouse button opens a context menu that 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