# Statistics and Computing



Springer Science+Business Media, LLC

# Statistics and Computing

*Gentle:* Numerical Linear Algebra for Applications in Statistics.
*Gentle:* Random Number Generation and Monte Carlo Methods.
*Härdle/Klinke/Turlach:* XploRe: An Interactive Statistical Computing Environment.
*Krause/Olson:* The Basics of S and S-PLUS.
*Lange:* Numerical Analysis for Statisticians.
*Loader:* Local Regression and Likelihood.
*Ó Ruanaidh/Fitzgerald:* Numerical Bayesian Methods Applied to Signal Processing.
*Pannatier:* VARIOWIN: Software for Spatial Data Analysis in 2D.
*Venables/Ripley:* Modern Applied Statistics with S-PLUS, 3rd edition.
*Wilkinson:* The Grammar of Graphics

W.N. Venables
B.D. Ripley

# Modern Applied Statistics with S-PLUS

Third Edition

With 144 Figures

Springer

W.N. Venables
CSIRO Marine Laboratories
PO Box 120
Cleveland, Qld, 4163
Australia
Bill.Venables@cmis.csiro.au

B.D. Ripley
Professor of Applied Statistics
University of Oxford
1 South Parks Road
Oxford OX1 3TG
UK
ripley@stats.ox.ac.uk



Library of Congress Cataloging-in-Publication Data
Venables, W.N. (William N.)
Modern applied statistics with S-PLUS / W.N. Venables, B.D. Ripley. – [3rd ed.]
p. cm. — (Statistics and computing)
Includes bibliographical references and index.

1. S-Plus. 2. Statistics—Data processing. 3. Mathematical statistics—Data processing. I. Ripley, Brian D., 1952–
II. Title. III. Series.
QA276.4.V46 1999
005.369–dc21 99-18388

Printed on acid-free paper.


Originally published by Springer-Verlag New York, Inc in 1999.

Softcover reprint of the hardcover 3rd edition 1999


Production managed by Robert Bruni; manufacturing supervised by Joe Quatela.
Photocomposed copy prepared from the authors' PostScript files.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4757-3123-1 ISBN 978-1-4757-3121-7 (eBook)
DOI 10.1007/978-1-4757-3121-7

# Preface

S-PLUS is a system for data analysis from the Data Analysis and Products Division of MathSoft, an enhanced version of the S environment for data analysis developed at Bell Laboratories (of AT&T and now Lucent Technologies). S-PLUS has become the statistician's calculator for the 1990s, allowing easy access to the computing power and graphical capabilities of modern workstations and personal computers.

The first edition of this book appeared in 1994 and a second in 1997. The S statistical system has continued to grow rapidly, and users have contributed an ever-growing wealth of software to implement the latest statistical methods in S. This book concentrates on using the current systems to do statistics; there is a companion volume which discusses programming in the S language in much greater depth.

Several different implementations of S have appeared. There are currently two 'engines' in use: S-PLUS 3.x and 4.x are based on version 3 of the S language whereas S-PLUS 5.x is based on S version 4. Furthermore, since 1997 the Windows version of S-PLUS has had a graphical user interface in the style of widespread Windows packages. Our aim is that this book should be usable with all these versions, but we have given lower priority to S-PLUS 3.x. Some of the more specialized functionality is covered in the *on-line complements* (see page 467 for sites) which will be updated frequently. The datasets and S functions that we use are available on-line, and help greatly in making use of the book.

This is not a text in statistical theory, but does cover modern statistical methodology. Each chapter summarizes the methods discussed, in order to set out the notation and the precise method implemented in S. (It will help if the reader has a basic knowledge of the topic of the chapter, but several chapters have been successfully used for specialized courses in statistical methods.) Our aim is rather to show how we analyse datasets using S-PLUS. In doing so we aim to show both how S can be used and how the availability of a powerful and graphical system has altered the way we approach data analysis and allows penetrating analyses to be performed routinely. Once calculation became easy, the statistician's energies could be devoted to understanding his or her dataset.

The core S language is not very large, but it is quite different from most other statistics systems. We describe the language in some detail in the early chapters, but these are probably best skimmed at first reading; Chapter 1 contains the most basic ideas, and each of Chapters 2 and 3 are divided into 'basic' and 'advanced' sections. Once the philosophy of the language is grasped, its consistency and logical design will be appreciated.

The chapters on applying S to statistical problems are largely self-contained, although Chapter 6 describes the language used for linear models that is used in several later chapters. We expect that most readers will want to pick and choose among the later chapters.

This book is intended both for would-be users of S-PLUS as an introductory guide and for class use. The level of course for which it is suitable differs from country to country, but would generally range from the upper years of an undergraduate course (especially the early chapters) to Masters' level. (For example, almost all the material is covered in the M.Sc. in Applied Statistics at Oxford.) Exercises are provided, but these should not detract from the best exercise of all, using S to study datasets with which the reader is familiar. Our library provides many datasets, some of which are not used in the text but are there to provide source material for exercises. (Further exercises and answers to selected exercises are available from our WWW pages.)

Both authors take responsibility for the whole book, but Bill Venables was the lead author for Chapters 1–4 and 6–8, and Brian Ripley for Chapters 5 and 9–14. The authors may be contacted by electronic mail at

```
Bill.Venables@cmis.csiro.au
ripley@stats.ox.ac.uk
```

and would appreciate being informed of errors and improvements to the contents of this book.

To avoid any confusion, S-PLUS is a commercial product, details of which may be obtained from `http://www.mathsoft.com/splus/`.

*Acknowledgements:*

This book would not be possible without the S environment which has been principally developed by Rick Becker, John Chambers and Allan Wilks, with substantial input from Doug Bates, Bill Cleveland, Trevor Hastie and Daryl Pregibon. The code for survival analysis is the work of Terry Therneau. The S-PLUS code is the work of a much larger team acknowledged in the manuals for that system.

We are grateful to the many people who have read and commented on draft material and who have helped us test the software, as well as to those whose problems have contributed to our understanding and indirectly to examples and exercises. We cannot name them all, but in particular we would like to thank Doug Bates, Adrian Bowman, Bill Dunlap, Sue Clancy, David Cox, Anthony Davison, Peter Diggle, Matthew Eagle, Nils Hjort, Stephen Kaluzny, Francis Marriott, Joseé Pinheiro, Brett Presnell, Charles Roosen, David Smith, Patty Solomon and Terry Therneau. We thank MathSoft DAPD and CSIRO DMS for early access to versions of S-PLUS and for access to platforms for testing.

Bill Venables
Brian Ripley
January 1999

# Contents

**Preface** v

**Typographical Conventions** xi

**1 Introduction** 1

1.1 A quick overview of S . . . . . 3
1.2 Using S-PLUS . . . . . 4
1.3 An introductory session . . . . . 5
1.4 What next? . . . . . 11

**2 The S Language** 13

2.1 A concise description of S objects . . . . . 13
2.2 Calling conventions for functions . . . . . 22
2.3 Arithmetical expressions . . . . . 24
2.4 Reading data . . . . . 30
2.5 Model formulae . . . . . 33
2.6 Character vector operations . . . . . 34
2.7 Finding S objects . . . . . 36
2.8 Indexing vectors, matrices and arrays . . . . . 39
2.9 Input/Output facilities . . . . . 45
2.10 Exercises . . . . . 52

**3 Graphics** 53

3.1 Graphics devices . . . . . 53
3.2 Basic plotting functions . . . . . 58
3.3 Enhancing plots . . . . . 62
3.4 Fine control of graphics . . . . . 67
3.5 Trellis graphics . . . . . 74
3.6 Exercises . . . . . 91

**4 Programming in S** **93**
4.1 Control structures . . . . . 93
4.2 More on character strings . . . . . 96
4.3 Matrix operations . . . . . 98
4.4 Vectorized calculations and loop avoidance functions . . . . . 103
4.5 Introduction to object orientation . . . . . 111

**5 Univariate Statistics** **113**
5.1 Probability distributions . . . . . 113
5.2 Generating random data . . . . . 116
5.3 Data summaries . . . . . 117
5.4 Classical univariate statistics . . . . . 122
5.5 Robust summaries . . . . . 126
5.6 Density estimation . . . . . 132
5.7 Bootstrap and permutation methods . . . . . 142
5.8 Exercises . . . . . 148

**6 Linear Statistical Models** **149**
6.1 An analysis of covariance example . . . . . 149
6.2 Model formulae and model matrices . . . . . 154
6.3 Regression diagnostics . . . . . 161
6.4 Safe prediction . . . . . 166
6.5 Robust and resistant regression . . . . . 167
6.6 Bootstrapping linear models . . . . . 174
6.7 Factorial designs and designed experiments . . . . . 176
6.8 An unbalanced four-way layout . . . . . 180
6.9 Predicting computer performance . . . . . 188
6.10 Multiple comparisons . . . . . 189
6.11 Random and mixed effects . . . . . 192
6.12 Exercises . . . . . 208

**7 Generalized Linear Models** **211**
7.1 Functions for generalized linear modelling . . . . . 215
7.2 Binomial data . . . . . 218
7.3 Poisson and multinomial models . . . . . 226
7.4 A negative binomial family . . . . . 233
7.5 Exercises . . . . . 236

**8 Non-linear Models** **241**
8.1 An introductory example . . . . . 241
8.2 Fitting non-linear regression models . . . . . 242
8.3 Non-linear fitted model objects and method functions . . . . . 247
8.4 Confidence intervals for parameters . . . . . 251
8.5 Profiles . . . . . 257
8.6 Constrained non-linear regression . . . . . 258
8.7 General optimization and maximum likelihood estimation . . . . . 261
8.8 Non-linear mixed effects models . . . . . 270
8.9 Exercises . . . . . 277

**9 Smooth Regression** **281**
9.1 Additive models and scatterplot smoothers . . . . . 281
9.2 Projection-pursuit regression . . . . . 289
9.3 Response transformation models . . . . . 294
9.4 Neural networks . . . . . 296
9.5 Conclusions . . . . . 302
9.6 Exercises . . . . . 302

**10 Tree-based Methods** **303**
10.1 Partitioning methods . . . . . 304
10.2 Implementation in `rpart` . . . . . 310
10.3 Implementation in `tree` . . . . . 319

**11 Multivariate Analysis and Pattern Recognition** **329**
11.1 Graphical methods . . . . . 330
11.2 Cluster analysis . . . . . 336
11.3 Correspondence analysis . . . . . 342
11.4 Discriminant analysis . . . . . 344
11.5 Classification theory . . . . . 349
11.6 Other classification methods . . . . . 353
11.7 Two extended examples . . . . . 356
11.8 Calibration plots . . . . . 364
11.9 Exercises . . . . . 365

**12 Survival Analysis** **367**
12.1 Estimators of survivor curves . . . . . 369
12.2 Parametric models . . . . . 373

12.3 Cox proportional hazards model . . . . . 380
12.4 Further examples . . . . . 386

**13 Time Series Analysis** **401**
13.1 Second-order summaries . . . . . 404
13.2 ARIMA models . . . . . 412
13.3 Seasonality . . . . . 418
13.4 Nottingham temperature data . . . . . 422
13.5 Regression with autocorrelated errors . . . . . 427
13.6 Exercises . . . . . 431

**14 Spatial Statistics** **433**
14.1 Spatial interpolation and smoothing . . . . . 433
14.2 Kriging . . . . . 439
14.3 Point process analysis . . . . . 444
14.4 Exercises . . . . . 447

## Appendices

**A Getting Started** **449**
A.1 Using S-PLUS under UNIX . . . . . 449
A.2 Using S-PLUS under Windows . . . . . 452
A.3 Customizing your S-PLUS environment . . . . . 453

**B The GUI in Version 4.x** **457**
B.1 Subwindows . . . . . 457
B.2 Graphics in the GUI . . . . . 460
B.3 Statistical analysis *via* the GUI . . . . . 465

**C Datasets, Software and Libraries** **467**
C.1 Our libraries . . . . . 468
C.2 Using libraries . . . . . 469

**References** **473**

**Index** **487**

## Typographical Conventions

Throughout this book S language constructs and commands to the operating system are set in a monospaced typewriter font `like this`. The character ~ may appear as ˜ on your keyboard, screen or printer.

We often use the prompts `$` for the operating system (it is the standard prompt for the UNIX Bourne shell) and `>` for S-PLUS. However, we do *not* use prompts for continuation lines, which are indicated by indentation. One reason for this is that the length of line available to use in a book column is less than that of a standard terminal window, so we have had to break lines that were not broken at the terminal.

Some of the S-PLUS output has been edited. Where complete lines are omitted, these are usually indicated by

```
....
```

in listings; however most *blank* lines have been silently removed. Much of the S-PLUS output was generated with the options settings

```
options(width=65, digits=5)
```

in effect, whereas the defaults are around `80` and `7`. Not all functions consult these settings, so on occasion we have had to manually reduce the precision to more sensible values.