Statistics and Computing

Series Editors:
J. Chambers
D. Hand
W. Härdle

For further volumes:
http://www.springer.com/series/3022

Robert A. Muenchen

R for SAS and SPSS Users

Second Edition

Springer

Robert A. Muenchen
University of Tennessee
Research Computing Support
109 Hoskins Library
1400 W. Cumberland
Knoxville, TN 37996-4005
USA
muenchen.bob@gmail.com

ISSN 1431-8784
ISBN 978-1-4614-0684-6 e-ISBN 978-1-4614-0685-3
DOI 10.1007/978-1-4614-0685-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011933470

Preface

While SAS and SPSS have many things in common, R is very different. My goal in writing this book is to help you translate what you know about SAS or SPSS into a working knowledge of R as quickly and easily as possible. I point out how they differ using terminology with which you are familiar, and show you which add-on packages will provide results most like those from SAS or SPSS. I provide many example programs done in SAS, SPSS, and R so that you can see how they compare topic by topic.

When finished, you should know how to:

- Install R, choose a user interface, and choose and install add-on packages.
- Read data from various sources such as text or Excel files, SAS or SPSS data sets, or relational databases.
- Manage your data through transformations, recodes, and combining data sets from both the add-cases and add-variables approaches and restructuring data from wide to long formats and vice versa.
- Create publication-quality graphs including bar, histogram, pie, line, scatter, regression, box, error bar, and interaction plots.
- Perform the basic types of analyses to measure strength of association and group differences, and be able to know where to turn to learn how to do more complex methods.

Who This Book Is For

This book teaches R requiring no prior knowledge of statistical software. However, you know SAS or SPSS this book will make learning R as easy as possible by using terms and concepts that you already know. If you do not know SAS or SPSS, then you will learn R along with how it compares to the two most popular commercial packages for data analysis. Stata users would be better off reading *R for Stata Users* [41].

An audience I did not expect to serve is R users wanting to learn SAS or SPSS. However, I have heard from quite a few of them who have said that by explaining the differences, it helped them learn in the reverse order I had anticipated. Keep in mind that I explain none of the SAS or SPSS programs, only the R ones and how the packages differ, so it is not ideal for that purpose.

Who This Book Is Not For

I make no effort to teach statistics or graphics. Although I briefly state the goal and assumptions of each analysis along with how to interpret their output, I do not cover their formulas or derivations. We have more than enough to discuss without tackling those topics too.

This is also not a book about writing complex R functions, it is about using the thousands that already exist. We will write only a few very short functions. If you want to learn more about writing functions, I recommend Jones et al.'s *Introduction to Scientific Programming and Simulation Using R* [31]. However, reading this book should ease your transition to more complex books like that one.

Practice Data Sets and Programs

All of the programs, data sets, and files that we use in this book are available for download at http://r4stats.com. A file containing corrections and clarifications is also available there.

Regarding the Second Edition

As the first edition went to press, I began planning the second edition with the main goal of adding more statistical methods. However, my readers quickly let me know that they needed far more information about the basics. There are many wonderful books devoted to statistics in R. I recommend some in Chap. 17. The enhancements to this edition include the following:

1. Programming code has been updated throughout.
2. It is easier to find reference material using the new list of tables and list of figures.
3. It is easier to find topics using the index, which now has four times as many entries.
4. The glossary defines more R terms.
5. There is a new Sect. 3.6, "Running R in SAS and WPS," including *A Bridge to R* and *IML Studio*.
6. There is a new Sect. 3.9, "Running R from within Text Editors."

7. There is a new Sect. 3.8, "Running R in Excel," complete with R Commander menus.
8. There is a new Sect. 3.10 on integrated development environments, including RStudio.
9. There is a new Sect. 3.11.1 on the `Deducer` user interface and its Plot Builder (similar to IBM SPSS Visualization Designer).
10. New Sect. 3.11.4 on Red-R, a flowchart user interface like SAS Enterprise Miner or IBM SPSS Modeler (Clementine).
11. Chapter 5, "Programming Language Basics," has been significantly enhanced, including additional examples and explanations.
12. There is a new Sect. 5.3.4 on matrix algebra with table of basic matrix algebra functions.
13. There is a new Sect. 5.6, "Comments to Document Your Objects."
14. Chapter 6, "Data Acquisition," includes improved examples of reading SAS and SPSS data files.
15. There is a new Sect. 6.2.3, "Reading Text from a Web Site."
16. There is a new Sect. 6.2.4, "Reading Text from the Clipboard."
17. There is a new Sect. 6.2.6, "Trouble with Tabs," on common problems when reading tab-delimited files.
18. Section 6.3, "Reading Text Data Within a Program," now includes a simpler approach using the `stdin` function.
19. There is a new Sect. 6.4 "Reading Multiple Observations per Line."
20. There are new sections on reading/writing Excel files.
21. There is a new Sect. 6.9, "Reading Data from Relational Databases.
22. There is a new Sect. 7.11.1, "Selecting Numeric or Character Variables," (like VAR A-numeric-Z; or A-character-Z).
23. There is a new Sect. 8.4, "Selecting Observations using Random Sampling."
24. Chapter 9, "Selecting Variables and Observations," has many more examples, and they are presented in order from most widely used to least.
25. There is a new Table 10.2, "Basic Statistical Functions."
26. There is a new Sect. 10.2.3 "Standardizing and Ranking Variables."
27. Section 10.14, "Removing Duplicate Observations," now includes an example for eliminating observations that are duplicates only on key variables (i.e., PROC SORT NODUPKEY).
28. There is a new Sect. 10.16, "Transposing or Flipping Data Sets" (tricky with character variables).
29. There is a new Sect. 10.20, "Character String Manipulations," using the `stringr` package.
30. There is a new Sect. 10.21, "Dates and Times," which covers date/time manipulations using the `lubridate` package.
31. The new Chap. 11, "Enhancing Your Output," covers how to get publication quality tables from R into word processors, Web pages or LaTeX.
32. The new Sect. 12.4, "Generating Values for Reading Fixed-Width Files," shows how to generate repetitive patterns of variable names and matching widths for reading complex text files.

33. There is a new Sect. 16.15, which shows how to make geographic maps.
34. There is a new Sect. 17.11 "Sign Test: Paired Groups."
35. Appendix B, "A Comparison of SAS and SPSS Products with R Packages and Functions," is now far more comprehensive and changes so frequently that I have moved it from the appendix to http://r4stats.com.

Acknowledgments

I am very grateful for the many people who have helped make this book possible, including the developers of the S language on which R is based, John Chambers, Douglas Bates, Rick Becker, Bill Cleveland, Trevor Hastie, Daryl Pregibon and Allan Wilks; the people who started R itself, Ross Ihaka and Robert Gentleman; the many other R developers for providing such wonderful tools for free and all of the R-help participants who have kindly answered so many questions. Most of the examples I present here are modestly tweaked versions of countless posts to the R-help discussion list, as well as a few SAS-L and SPSSX-L posts. All I add is the selection, organization, explanation, and comparison to similar SAS and SPSS programs.

I am especially grateful to the people who provided advice, caught typos, and suggested improvements, including Raymond R. Balise, Patrick Burns, Glenn Corey, Peter Flom, Chun Huang, Richard Gold, Martin Gregory, Warren Lambert, Matthew Marler, Paul Miller, Ralph O'Brien, Wayne Richter, Denis Shah, Charilaos Skiadas, Andreas Stefik, Phil Spector, Joseph Voelkel, Michael Wexler, Graham Williams, Andrew Yee, and several anonymous reviewers.

My special thanks go to Hadley Wickham, who provided much guidance on his `ggplot2` graphics package, as well as a few of his other handy packages. Thanks to Gabor Grothendieck, Lauri Nikkinen, and Marc Schwarz for the R-Help discussion that led to Sect. 10.15: "Selecting First or Last Observations per Group." Thanks to Gabor Grothendieck also for a detailed discussion that lead to Sect. 10.4, "Multiple Conditional Transformations." Thanks to Garrett Grolemund for his help in understanding dates, times and his time-saving `lubridate` package. Thanks to Frank Harrell, Jr. for helping me elucidate the discussion of object orientation in final chapter.

I also thank SPSS, Inc. especially Jon Peck, for the helpful review of this book and Jon's SPSS expertise, which benefited several areas including the programs for extracting the first/last observation per group, formatting date–time variables, and generating data. He not only improved quite a few of the SPSS programs, but found ways to improve several of the R ones as well!

At The University of Tennessee, I am thankful for the many faculty, staff, and students who have challenged me to improve my teaching and data analysis skills. My colleagues Michael Newman, Michael O'Neil, Virginia Patterson, Ann Reed, Sue Smith, Cary Springer, and James Schmidhammer have been

a source of much assistance and inspiration. Michael McGuire, provided assistance with all things Macintosh.

Finally, I am grateful to my wife, Carla Foust, and sons Alexander and Conor, who put up with many lost weekends while I wrote this book.

Robert A. Muenchen
muenchen.bob@gmail.com
Knoxville, Tennessee

About the Author

Robert A. Muenchen is a consulting statistician and, with Joseph Hilbe, author of the book *R for Stata Users* [41]. He is currently the manager of Research Computing Support (formerly the Statistical Consulting Center) at the University of Tennessee. Bob has conducted research for a variety of public and private organizations and has coauthored over 50 articles in scientific journals and conference proceedings.

Bob has served on the advisory boards of the SAS Institute, SPSS, Inc. the Statistical Graphics Corporation, and *PC Week Magazine*. His suggested improvements have been incorporated into SAS, SPSS, JMP, STATGRAPHICS, and several R packages.

His research interests include statistical computing, data graphics and visualization, text analysis, data mining, psychometrics, and resampling.

Contents

List of Tables

List of Figures