

# **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](mailto:muenchen.bob@gmail.com)

***Series Editors:***

J. Chambers  
Department of Statistics  
Sequoia Hall  
390 Serra Mall  
Stanford University  
Stanford, CA 94305-4065

D. Hand  
Department of Mathematics  
Imperial College London,  
South Kensington Campus  
London SW7 2AZ  
United Kingdom

W. Härdle  
C.A.S.E. Centre for Applied  
Statistics and Economics  
School of Business and  
Economics  
Humboldt-Universität zu Berlin  
Unter den Linden 6  
10099 Berlin  
Germany

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

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

---

## 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 L<sup>A</sup>T<sub>E</sub>X.
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 led to Sect. 10.4, “Multiple Conditional Transformations.” Thanks to Garrett Grolemond 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.

Linux<sup>®</sup> is the registered trademark of Linus Torvalds.

MATLAB<sup>®</sup> is a registered trademark of The Mathworks, Inc.

Macintosh<sup>®</sup> and Mac OS<sup>®</sup> are registered trademarks of Apple, Inc.

Oracle<sup>®</sup> and Oracle Data Mining are registered trademarks of Oracle, Inc.

R-PLUS<sup>®</sup> is a registered trademark of XL-Solutions, Inc.

RStudio<sup>®</sup> is a registered trademark of RStudio, Inc.

Revolution R<sup>®</sup> and Revolution R Enterprise<sup>®</sup> are registered trademarks of Revolution Analytics, Inc.

SAS<sup>®</sup>, SAS<sup>®</sup>, AppDev Studio<sup>™</sup>, SAS<sup>®</sup> Enterprise Guide<sup>®</sup>, SAS<sup>®</sup>

Enterprise Miner<sup>™</sup>, and SAS/IML<sup>®</sup> Studio are registered trademarks of the SAS Institute.

SPSS<sup>®</sup>, IBM SPSS Statistics<sup>®</sup>, IBM SPSS Modeler<sup>®</sup>, IBM SPSS Visualization Designer<sup>®</sup>, and Clementine<sup>®</sup>, are registered trademarks of SPSS, Inc., an IBM company.

Stata<sup>®</sup> is a registered trademark of Statacorp, Inc.

Tibco Spotfire S+<sup>®</sup> is a registered trademark of Tibco, Inc.

UNIX<sup>®</sup> is a registered trademark of The Open Group.

Windows<sup>®</sup>, Windows Vista<sup>®</sup>, Windows XP<sup>®</sup>, Windows XP<sup>®</sup>, Excel<sup>®</sup>,  
and Microsoft Word<sup>®</sup> are registered trademarks of Microsoft, Inc.

World Programming System<sup>®</sup> and WPS<sup>®</sup> are registered trademarks of  
World Programming, Ltd.

Copyright © 2006, 2007, 2008, 2011 by Robert A. Muenchen. All rights  
reserved.

---

# Contents

<b>1</b>	<b>Introduction</b>	1
1.1	Overview	1
1.2	Why Learn R?	2
1.3	Is R Accurate?	3
1.4	What About Tech Support?	4
1.5	Getting Started Quickly	5
1.6	The Five Main Parts of SAS and SPSS	5
1.7	Our Practice Data Sets	7
1.8	Programming Conventions	8
1.9	Typographic Conventions	9
<b>2</b>	<b>Installing and Updating R</b>	11
2.1	Installing Add-on Packages	11
2.2	Loading an Add-on Package	13
2.3	Updating Your Installation	15
2.4	Uninstalling R	17
2.5	Uninstalling a Package	17
2.6	Choosing Repositories	18
2.7	Accessing Data in Packages	18
<b>3</b>	<b>Running R</b>	21
3.1	Running R Interactively on Windows	21
3.2	Running R Interactively on Macintosh	24
3.3	Running R Interactively on Linux or UNIX	26
3.4	Running Programs That Include Other Programs	28
3.5	Running R in Batch Mode	29
3.6	Running R in SAS and WPS	30
3.6.1	SAS/IML Studio	30
3.6.2	A Bridge to R	31
3.6.3	The SAS X Command	31
3.6.4	Running SAS and R Sequentially	32

3.6.5	Example Program Running R from Within SAS . . . . .	32
3.7	Running R in SPSS . . . . .	33
3.7.1	Example Program Running R from Within SPSS . . . . .	37
3.8	Running R in Excel . . . . .	37
3.9	Running R from Within Text Editors . . . . .	39
3.10	Integrated Development Environments . . . . .	40
3.10.1	Eclipse . . . . .	40
3.10.2	JGR . . . . .	41
3.10.3	RStudio . . . . .	42
3.11	Graphical User Interfaces . . . . .	42
3.11.1	Deducer . . . . .	43
3.11.2	R Commander . . . . .	46
3.11.3	rattle . . . . .	48
3.11.4	Red-R . . . . .	51
<b>4</b>	<b>Help and Documentation . . . . .</b>	<b>53</b>
4.1	Starting Help . . . . .	53
4.2	Examples in Help Files . . . . .	55
4.3	Help for Functions That Call Other Functions . . . . .	57
4.4	Help for Packages . . . . .	57
4.5	Help for Data Sets . . . . .	58
4.6	Books and Manuals . . . . .	58
4.7	E-mail Lists . . . . .	58
4.8	Searching the Web . . . . .	59
4.9	Vignettes . . . . .	60
4.10	Demonstrations . . . . .	60
<b>5</b>	<b>Programming Language Basics . . . . .</b>	<b>61</b>
5.1	Introduction . . . . .	61
5.2	Simple Calculations . . . . .	62
5.3	Data Structures . . . . .	63
5.3.1	Vectors . . . . .	63
5.3.2	Factors . . . . .	68
5.3.3	Data Frames . . . . .	74
5.3.4	Matrices . . . . .	78
5.3.5	Arrays . . . . .	82
5.3.6	Lists . . . . .	83
5.4	Saving Your Work . . . . .	88
5.5	Comments to Document Your Programs . . . . .	90
5.6	Comments to Document Your Objects . . . . .	91
5.7	Controlling Functions (Procedures) . . . . .	92
5.7.1	Controlling Functions with Arguments . . . . .	92
5.7.2	Controlling Functions with Objects . . . . .	95
5.7.3	Controlling Functions with Formulas . . . . .	96
5.7.4	Controlling Functions with an Object's Class . . . . .	96

5.7.5	Controlling Functions with Extractor Functions . . . . .	99
5.8	How Much Output There? . . . . .	100
5.9	Writing Your Own Functions (Macros) . . . . .	105
5.10	Controlling Program Flow . . . . .	107
5.11	R Program Demonstrating Programming Basics . . . . .	108
<b>6</b>	<b>Data Acquisition</b> . . . . .	<b>115</b>
6.1	Manual Data Entry Using the R Data Editor . . . . .	115
6.2	Reading Delimited Text Files . . . . .	117
6.2.1	Reading Comma-Delimited Text Files . . . . .	118
6.2.2	Reading Tab-Delimited Text Files . . . . .	120
6.2.3	Reading Text from a Web Site . . . . .	121
6.2.4	Reading Text from the Clipboard . . . . .	122
6.2.5	Missing Values for Character Variables . . . . .	122
6.2.6	Trouble with Tabs . . . . .	124
6.2.7	Skipping Variables in Delimited Text Files . . . . .	125
6.2.8	Reading Character Strings . . . . .	126
6.2.9	Example Programs for Reading Delimited Text Files . . . . .	126
6.3	Reading Text Data Within a Program . . . . .	129
6.3.1	The Easy Approach . . . . .	130
6.3.2	The More General Approach . . . . .	131
6.3.3	Example Programs for Reading Text Data Within a Program . . . . .	132
6.4	Reading Multiple Observations per Line . . . . .	134
6.4.1	Example Programs for Reading Multiple Observations per Line . . . . .	136
6.5	Reading Data from the Keyboard . . . . .	138
6.6	Reading Fixed-Width Text Files, One Record per Case . . . . .	138
6.6.1	Reading Data Using Macro Substitution . . . . .	141
6.6.2	Example Programs for Reading Fixed-Width Text Files, One Record per Case . . . . .	142
6.7	Reading Fixed-Width Text Files, Two or More Records per Case . . . . .	143
6.7.1	Example Programs to Read Fixed-Width Text Files with Two Records per Case . . . . .	145
6.8	Reading Excel Files . . . . .	146
6.8.1	Example Programs for Reading Excel Files . . . . .	147
6.9	Reading from Relational Databases . . . . .	148
6.10	Reading Data from SAS . . . . .	149
6.10.1	Example Programs to Write Data from SAS and Read It into R . . . . .	150
6.11	Reading Data from SPSS . . . . .	151
6.11.1	Example Programs for Reading Data from SPSS . . . . .	152

6.12	Writing Delimited Text Files . . . . .	153
6.12.1	Example Programs for Writing Delimited Text Files .	154
6.13	Viewing a Text File . . . . .	156
6.14	Writing Excel Files . . . . .	156
6.14.1	Example Programs for Writing Excel Files . . . . .	157
6.15	Writing to Relational Databases . . . . .	158
6.16	Writing Data to SAS and SPSS . . . . .	158
6.16.1	Example Programs to Write Data to SAS and SPSS .	159
<b>7</b>	<b>Selecting Variables . . . . .</b>	<b>161</b>
7.1	Selecting Variables in SAS and SPSS . . . . .	161
7.2	Subscripting . . . . .	162
7.3	Selecting Variables by Index Number . . . . .	163
7.4	Selecting Variables by Column Name . . . . .	166
7.5	Selecting Variables Using Logic . . . . .	167
7.6	Selecting Variables by String Search (varname: or varname1-varnameN) . . . . .	169
7.7	Selecting Variables Using \$ Notation . . . . .	172
7.8	Selecting Variables by Simple Name . . . . .	172
7.8.1	The <code>attach</code> Function . . . . .	173
7.8.2	The <code>with</code> Function . . . . .	174
7.8.3	Using Short Variable Names in Formulas . . . . .	174
7.9	Selecting Variables with the <code>subset</code> Function . . . . .	175
7.10	Selecting Variables by List Subscript . . . . .	176
7.11	Generating Indices A to Z from Two Variable Names . . . . .	176
7.11.1	Selecting Numeric or Character Variables . . . . .	177
7.12	Saving Selected Variables to a New Data Set . . . . .	180
7.13	Example Programs for Variable Selection . . . . .	180
7.13.1	SAS Program to Select Variables . . . . .	181
7.13.2	SPSS Program to Select Variables . . . . .	181
7.13.3	R Program to Select Variables . . . . .	182
<b>8</b>	<b>Selecting Observations . . . . .</b>	<b>187</b>
8.1	Selecting Observations in SAS and SPSS . . . . .	187
8.2	Selecting All Observations . . . . .	188
8.3	Selecting Observations by Index Number . . . . .	189
8.4	Selecting Observations Using Random Sampling . . . . .	191
8.5	Selecting Observations by Row Name . . . . .	193
8.6	Selecting Observations Using Logic . . . . .	194
8.7	Selecting Observations by String Search . . . . .	198
8.8	Selecting Observations with the <code>subset</code> Function . . . . .	200
8.9	Generating Indices A to Z from Two Row Names . . . . .	200
8.10	Variable Selection Methods with No Counterpart for Selecting Observations . . . . .	201

8.11	Saving Selected Observations to a New Data Frame . . . . .	201
8.12	Example Programs for Selecting Observations . . . . .	202
8.12.1	SAS Program to Select Observations . . . . .	202
8.12.2	SPSS Program to Select Observations . . . . .	203
8.12.3	R Program to Select Observations . . . . .	203
<b>9</b>	<b>Selecting Variables and Observations . . . . .</b>	<b>209</b>
9.1	The <code>subset</code> Function . . . . .	209
9.2	Subscripting with Logical Selections and Variable Names . . . . .	211
9.3	Using Names to Select Both Observations and Variables . . . . .	212
9.4	Using Numeric Index Values to Select Both Observations and Variables . . . . .	213
9.5	Using Logic to Select Both Observations and Variables . . . . .	213
9.6	Saving and Loading Subsets . . . . .	214
9.7	Example Programs for Selecting Variables and Observations . . . . .	215
9.7.1	SAS Program for Selecting Variables and Observations . . . . .	215
9.7.2	SPSS Program for Selecting Variables and Observations . . . . .	215
9.7.3	R Program for Selecting Variables and Observations . . . . .	216
<b>10</b>	<b>Data Management . . . . .</b>	<b>219</b>
10.1	Transforming Variables . . . . .	219
10.1.1	Example Programs for Transforming Variables . . . . .	223
10.2	Procedures or Functions? The <code>apply</code> Function Decides . . . . .	225
10.2.1	Applying the <code>mean</code> Function . . . . .	225
10.2.2	Finding N or NVALID . . . . .	229
10.2.3	Standardizing and Ranking Variables . . . . .	231
10.2.4	Applying Your Own Functions . . . . .	233
10.2.5	Example Programs for Applying Statistical Functions . . . . .	234
10.3	Conditional Transformations . . . . .	237
10.3.1	The <code>ifelse</code> Function . . . . .	237
10.3.2	Cutting Functions . . . . .	241
10.3.3	Example Programs for Conditional Transformations . . . . .	242
10.4	Multiple Conditional Transformations . . . . .	246
10.4.1	Example Programs for Multiple Conditional Transformations . . . . .	248
10.5	Missing Values . . . . .	250
10.5.1	Substituting Means for Missing Values . . . . .	252
10.5.2	Finding Complete Observations . . . . .	253
10.5.3	When “99” Has Meaning . . . . .	254
10.5.4	Example Programs to Assign Missing Values . . . . .	255
10.6	Renaming Variables (and Observations) . . . . .	258
10.6.1	Advanced Renaming Examples . . . . .	260
10.6.2	Renaming by Index . . . . .	261

10.6.3	Renaming by Column Name . . . . .	262
10.6.4	Renaming Many Sequentially Numbered Variable Names . . . . .	263
10.6.5	Renaming Observations . . . . .	264
10.6.6	Example Programs for Renaming Variables . . . . .	264
10.7	Recoding Variables . . . . .	268
10.7.1	Recoding a Few Variables . . . . .	269
10.7.2	Recoding Many Variables . . . . .	269
10.7.3	Example Programs for Recoding Variables . . . . .	272
10.8	Indicator or Dummy Variables . . . . .	274
10.8.1	Example Programs for Indicator or Dummy Variables . . . . .	277
10.9	Keeping and Dropping Variables . . . . .	279
10.9.1	Example Programs for Keeping and Dropping Variables . . . . .	280
10.10	Stacking/Concatenating/Adding Data Sets . . . . .	281
10.10.1	Example Programs for Stacking/Concatenating/Adding Data Sets . . . . .	283
10.11	Joining/Merging Data Sets . . . . .	285
10.11.1	Example Programs for Joining/Merging Data Sets . . . . .	288
10.12	Creating Summarized or Aggregated Data Sets . . . . .	290
10.12.1	The <code>aggregate</code> Function . . . . .	290
10.12.2	The <code>tapply</code> Function . . . . .	292
10.12.3	Merging Aggregates with Original Data . . . . .	294
10.12.4	Tabular Aggregation . . . . .	296
10.12.5	The <code>plyr</code> and <code>reshape2</code> Packages . . . . .	298
10.12.6	Comparing Summarization Methods . . . . .	298
10.12.7	Example Programs for Aggregating/Summarizing Data . . . . .	299
10.13	By or Split-File Processing . . . . .	302
10.13.1	Example Programs for By or Split-File Processing . . . . .	306
10.14	Removing Duplicate Observations . . . . .	308
10.14.1	Completely Duplicate Observations . . . . .	308
10.14.2	Duplicate Keys . . . . .	311
10.14.3	Example Programs for Removing Duplicates . . . . .	311
10.15	Selecting First or Last Observations per Group . . . . .	314
10.15.1	Example Programs for Selecting Last Observation per Group . . . . .	317
10.16	Transposing or Flipping Data Sets . . . . .	319
10.16.1	Example Programs for Transposing or Flipping Data Sets . . . . .	322
10.17	Reshaping Variables to Observations and Back . . . . .	324
10.17.1	Summarizing/Aggregating Data Using <code>reshape2</code> . . . . .	328
10.17.2	Example Programs for Reshaping Variables to Observations and Back . . . . .	330
10.18	Sorting Data Frames . . . . .	333



10.18.1	Example Programs for Sorting Data Sets . . . . .	336
10.19	Converting Data Structures . . . . .	338
10.19.1	Converting from Logical to Numeric Index and Back . . . . .	341
10.20	Character String Manipulations . . . . .	342
10.20.1	Example Programs for Character String Manipulation . . . . .	349
10.21	Dates and Times . . . . .	354
10.21.1	Calculating Durations . . . . .	358
10.21.2	Adding Durations to Date–Time Variables . . . . .	362
10.21.3	Accessing Date–Time Elements . . . . .	362
10.21.4	Creating Date–Time Variables from Elements . . . . .	363
10.21.5	Logical Comparisons with Date–Time Variables . . . . .	364
10.21.6	Formatting Date–Time Output . . . . .	364
10.21.7	Two-Digit Years . . . . .	365
10.21.8	Date–Time Conclusion . . . . .	366
10.21.9	Example Programs for Dates and Times . . . . .	366
<b>11</b>	<b>Enhancing Your Output . . . . .</b>	<b>375</b>
11.1	Value Labels or Formats (and Measurement Level) . . . . .	375
11.1.1	Character Factors . . . . .	376
11.1.2	Numeric Factors . . . . .	378
11.1.3	Making Factors of Many Variables . . . . .	380
11.1.4	Converting Factors to Numeric or Character Variables . . . . .	383
11.1.5	Dropping Factor Levels . . . . .	384
11.1.6	Example Programs for Value Labels . . . . .	385
11.1.7	R Program to Assign Value Labels and Factor Status . . . . .	386
11.2	Variable Labels . . . . .	389
11.2.1	Other Packages That Support Variable Labels . . . . .	393
11.2.2	Example Programs for Variable Labels . . . . .	393
11.3	Output for Word Processing and Web Pages . . . . .	395
11.3.1	The <code>xtable</code> Package . . . . .	396
11.3.2	Other Options for Formatting Output . . . . .	398
11.3.3	Example Program for Formatting Output . . . . .	398
<b>12</b>	<b>Generating Data . . . . .</b>	<b>401</b>
12.1	Generating Numeric Sequences . . . . .	402
12.2	Generating Factors . . . . .	403
12.3	Generating Repetitious Patterns (Not Factors) . . . . .	404
12.4	Generating Values for Reading Fixed-Width Files . . . . .	405
12.5	Generating Integer Measures . . . . .	406
12.6	Generating Continuous Measures . . . . .	408
12.7	Generating a Data Frame . . . . .	409
12.8	Example Programs for Generating Data . . . . .	411
12.8.1	SAS Program for Generating Data . . . . .	411
12.8.2	SPSS Program for Generating Data . . . . .	412
12.8.3	R Program for Generating Data . . . . .	413

<b>13 Managing Your Files and Workspace</b> .....	417
13.1 Loading and Listing Objects .....	417
13.2 Understanding Your Search Path .....	421
13.3 Attaching Data Frames .....	422
13.4 Loading Packages .....	424
13.5 Attaching Files .....	426
13.6 Removing Objects from Your Workspace .....	427
13.7 Minimizing Your Workspace .....	430
13.8 Setting Your Working Directory .....	430
13.9 Saving Your Workspace .....	431
13.9.1 Saving Your Workspace Manually .....	431
13.9.2 Saving Your Workspace Automatically .....	431
13.9.3 Getting Operating Systems to Show You .RData Files ..	432
13.9.4 Organizing Projects with Windows Shortcuts .....	432
13.10 Saving Your Programs and Output .....	433
13.11 Saving Your History .....	433
13.12 Large Data Set Considerations .....	435
13.13 Example R Program for Managing Files and Workspace .....	435
<b>14 Graphics Overview</b> .....	441
14.1 Dynamic Visualization .....	441
14.2 SAS/GRAPH .....	442
14.3 SPSS Graphics .....	442
14.4 R Graphics .....	443
14.5 The Grammar of Graphics .....	444
14.6 Other Graphics Packages .....	445
14.7 Graphics Archives .....	445
14.8 Graphics Demonstrations .....	445
14.9 Graphics Procedures and Graphics Systems .....	447
14.10 Graphics Devices .....	448
<b>15 Traditional Graphics</b> .....	451
15.1 The <code>plot</code> Function .....	451
15.2 Bar Plots .....	453
15.2.1 Bar Plots of Counts .....	453
15.2.2 Bar Plots for Subgroups of Counts .....	457
15.2.3 Bar Plots of Means .....	458
15.3 Adding Titles, Labels, Colors, and Legends .....	459
15.4 Graphics Parameters and Multiple Plots on a Page .....	462
15.5 Pie Charts .....	465
15.6 Dot Charts .....	466
15.7 Histograms .....	466
15.7.1 Basic Histograms .....	467
15.7.2 Histograms Stacked .....	469

15.7.3	Histograms Overlaid	470
15.8	Normal QQ Plots	475
15.9	Strip Charts	476
15.10	Scatter and Line Plots	480
15.10.1	Scatter Plots with Jitter	483
15.10.2	Scatter Plots with Large Data Sets	483
15.10.3	Scatter Plots with Lines	486
15.10.4	Scatter Plots with Linear Fit by Group	487
15.10.5	Scatter Plots by Group or Level (Coplots)	489
15.10.6	Scatter Plots with Confidence Ellipse	489
15.10.7	Scatter Plots with Confidence and Prediction Intervals	490
15.10.8	Plotting Labels Instead of Points	496
15.10.9	Scatter Plot Matrices	498
15.11	Dual-Axis Plots	500
15.12	Box Plots	502
15.13	Error Bar Plots	505
15.14	Interaction Plots	505
15.15	Adding Equations and Symbols to Graphs	505
15.16	Summary of Graphics Elements and Parameters	507
15.17	Plot Demonstrating Many Modifications	507
15.18	Example Traditional Graphics Programs	508
15.18.1	SAS Program for Traditional Graphics	510
15.18.2	SPSS Program for Traditional Graphics	510
15.18.3	R Program for Traditional Graphics	511
<b>16</b>	<b>Graphics with ggplot2</b>	<b>521</b>
16.1	Introduction	521
16.1.1	Overview of <code>qplot</code> and <code>ggplot</code>	522
16.1.2	Missing Values	524
16.1.3	Typographic Conventions	525
16.2	Bar Plots	526
16.3	Pie Charts	528
16.4	Bar Plots for Groups	530
16.5	Plots by Group or Level	531
16.6	Presummarized Data	532
16.7	Dot Charts	534
16.8	Adding Titles and Labels	535
16.9	Histograms and Density Plots	536
16.9.1	Histograms	536
16.9.2	Density Plots	537
16.9.3	Histograms with Density Overlaid	538
16.9.4	Histograms for Groups, Stacked	539
16.9.5	Histograms for Groups, Overlaid	540
16.10	Normal QQ Plots	540
16.11	Strip Plots	541

16.12	Scatter Plots and Line Plots	544
16.12.1	Scatter Plots with Jitter	547
16.12.2	Scatter Plots for Large Data Sets	548
16.12.3	Scatter Plots with Fit Lines	553
16.12.4	Scatter Plots with Reference Lines	555
16.12.5	Scatter Plots with Labels Instead of Points	557
16.12.6	Changing Plot Symbols	559
16.12.7	Scatter Plot with Linear Fits by Group	560
16.12.8	Scatter Plots Faceted by Groups	561
16.12.9	Scatter Plot Matrix	562
16.13	Box Plots	564
16.14	Error Bar Plots	567
16.15	Geographic Maps	568
16.15.1	Finding and Converting Maps	573
16.16	Logarithmic Axes	574
16.17	Aspect Ratio	575
16.18	Multiple Plots on a Page	575
16.19	Saving <code>ggplot2</code> Graphs to a File	577
16.20	An Example Specifying All Defaults	578
16.21	Summary of Graphics Elements and Parameters	579
16.22	Example Programs for Grammar of Graphics	580
16.22.1	SPSS Program for Graphics Production Language	580
16.22.2	R Program for <code>ggplot2</code>	583
<b>17</b>	<b>Statistics</b>	<b>599</b>
17.1	Scientific Notation	599
17.2	Descriptive Statistics	600
17.2.1	The <code>Deducer</code> <code>frequencies</code> Function	600
17.2.2	The <code>Hmisc</code> <code>describe</code> Function	601
17.2.3	The <code>summary</code> Function	603
17.2.4	The <code>table</code> Function and Its Relatives	604
17.2.5	The <code>mean</code> Function and Its Relatives	606
17.3	Cross-Tabulation	607
17.3.1	The <code>CrossTable</code> Function	607
17.3.2	The <code>table</code> and <code>chisq.test</code> Functions	608
17.4	Correlation	612
17.4.1	The <code>cor</code> Function	614
17.5	Linear Regression	616
17.5.1	Plotting Diagnostics	620
17.5.2	Comparing Models	621
17.5.3	Making Predictions with New Data	622
17.6	t-Test: Independent Groups	622
17.7	Equality of Variance	624
17.8	t-Test: Paired or Repeated Measures	625

17.9	Wilcoxon–Mann–Whitney Rank Sum: Independent Groups . . . . .	626
17.10	Wilcoxon Signed-Rank Test: Paired Groups . . . . .	627
17.11	Sign Test: Paired Groups . . . . .	628
17.12	Analysis of Variance . . . . .	630
17.13	Sums of Squares . . . . .	633
17.14	The Kruskal–Wallis Test . . . . .	635
17.15	Example Programs for Statistical Tests . . . . .	637
	17.15.1 SAS Program for Statistical Tests . . . . .	637
	17.15.2 SPSS Program for Statistical Tests . . . . .	639
	17.15.3 R Program for Statistical Tests . . . . .	641
<b>18</b>	<b>Conclusion . . . . .</b>	<b>647</b>
	<b>References . . . . .</b>	<b>663</b>
	<b>Index . . . . .</b>	<b>669</b>



---

## List of Tables

3.1	R table transferred to SPSS .....	37
5.1	Matrix functions .....	82
5.2	Modes and classes of various R objects .....	97
10.1	Mathematical operators and functions .....	220
10.2	Basic statistical functions .....	232
10.3	Logical operators .....	238
10.4	Comparison of summarization functions .....	298
10.5	Data conversion functions .....	340
10.6	Date–time format conversion specifications .....	367
11.1	Data printed in $\text{\LaTeX}$ .....	395
11.2	Linear model results formatted by <code>xtable</code> .....	398
13.1	Workspace management functions .....	434
14.1	Comparison of R’s three main graphics packages .....	444
15.1	Graphics arguments for high-level functions .....	478
15.2	Graphics parameters for <code>par</code> .....	479
15.3	Graphics functions to add elements .....	480
16.1	Comparison of <code>qplot</code> and <code>ggplot</code> functions .....	525
17.1	Example formulas in SAS, SPSS, and R .....	619





---

## List of Figures

3.1	The R graphical user interface in Windows .....	22
3.2	The R graphical user interface on Macintosh .....	24
3.3	R and R Commander both integrated into Excel .....	39
3.4	JGR's program editor .....	42
3.5	JGR offering a list of arguments .....	43
3.6	JGR's Package Manager .....	44
3.7	JGR's Object Browser .....	44
3.8	The RStudio integrated development environment .....	45
3.9	Deducer's graphical user interface .....	45
3.10	Deducer's data viewer/editor .....	46
3.11	Deducer's Descriptive Statistics dialog box .....	47
3.12	Deducer's Plot Builder .....	48
3.13	R Commander user interface in use .....	49
3.14	rattle's user interface for data mining .....	50
3.15	Red-R flowchart-style graphical user interface .....	52
4.1	R's main help window .....	54
6.1	Adding a new variable in the R data editor .....	115
6.2	The R data editor with practice data entered .....	116
10.1	Renaming a variable using R's data editor .....	258
10.2	Renaming variables using the <code>edit</code> function .....	260
12.1	Bar plots of generated data .....	408
12.2	Histograms of generated data .....	410
14.1	Napoleon's march to Moscow .....	446
15.1	Default plots from <code>plot</code> function .....	452
15.2	Bar plot .....	453
15.3	Bar plot on unsummarized variable q4 .....	454

15.4	Bar plot improved	455
15.5	Bar plot of gender	455
15.6	Horizontal bar plot of workshop	456
15.7	Stacked bar plot of workshop	456
15.8	Bar plot of workshop split by gender	457
15.9	Mosaic plot of workshop by gender using <code>plot</code>	458
15.10	Mosaic plot of workshop by gender using <code>mosaicplot</code>	459
15.11	Mosaic plot of three variables	460
15.12	Bar plot of means	461
15.13	Bar plot of q1 means by workshop and gender	462
15.14	Bar plot with title, label, legend and shading	463
15.15	Bar plots of counts by workshop and gender	466
15.16	Pie chart of workshop attendance	467
15.17	Dot chart of workshop within gender	468
15.18	Histogram of <code>posttest</code>	469
15.19	Histogram of <code>posttest</code> with density and “rug”	470
15.20	Histogram of <code>posttest</code> for males only	471
15.21	Multiframe histograms of all and just males	472
15.22	Histogram of males overlaid on all	473
15.23	Histogram with matching break points	474
15.24	Normal quantile plot	476
15.25	Strip chart demonstrating jitter and stack	477
15.26	Strip chart of <code>posttest</code> by workshop	478
15.27	Scatter plot of <code>pretest</code> and <code>posttest</code>	481
15.28	Scatter plots of various types	482
15.29	Scatter plots with jitter on Likert data	483
15.30	Scatter plots on large data sets	484
15.31	Hexbin plot	485
15.32	<code>smoothScatter</code> plot	486
15.33	Scatter plot with lines, legend and title	487
15.34	Scatter plot with symbols and fits by gender	488
15.35	Scatter coplot by categorical variable	490
15.36	Scatter coplot by continuous variable	491
15.37	Scatter plot with 95% confidence ellipse	492
15.38	Scatter plot foundation for confidence intervals	493
15.39	Scatter plot with simulated confidence intervals	494
15.40	Scatter plot with actual confidence intervals	496
15.41	Scatter plot using characters as group symbols	497
15.42	Scatter plot with row names as labels	498
15.43	Scatter plot matrix	499
15.44	Scatter plot matrix with smoothed fits	501
15.45	Scatter plot with double $y$ -axes and grid	502
15.46	Box plot of <code>posttest</code> by workshop	503
15.47	Various box plots	504
15.48	Error bar plot	506

15.49	Interaction plot	507
15.50	Plot demonstrating many embellishments	509
16.1	Default plots from <code>qplot</code> function	523
16.2	Bar plot done with <code>ggplot2</code> package	526
16.3	Horizontal bar plot	528
16.4	Stacked bar plot of workshop	529
16.5	Pie chart of workshop	530
16.6	Bar plot types of stack, fill, and dodge	531
16.7	Bar plots of workshop for each gender	533
16.8	Bar plot of presummarized data	534
16.9	Dot chart of workshop by gender	536
16.10	Bar plot demonstrating titles and labels	537
16.11	Histogram of <code>posttest</code>	538
16.12	Histogram with smaller bins	539
16.13	Density plot of <code>posttest</code>	540
16.14	Histogram with density curve and rug	541
16.15	Histograms of <code>posttest</code> by gender	542
16.16	Histogram with bars filled by gender	543
16.17	Normal quantile plot	544
16.18	Strip chart with jitter	545
16.19	Strip chart by workshop	546
16.20	Scatter plots demonstrating various line types	547
16.21	Scatter plots showing effect of jitter	548
16.22	Scatter plot showing transparency	550
16.23	Scatter plot with contour lines	551
16.24	Scatter plot with density shading	552
16.25	Hexbin plot of <code>pretest</code> and <code>posttest</code>	553
16.26	Scatter plot with regression line and confidence band	554
16.27	Scatter plot with regression line but no confidence band	555
16.28	Scatter plot with $y=x$ line added	556
16.29	Scatter plot with vertical and horizontal reference lines	557
16.30	Scatter plot with multiple vertical reference lines	558
16.31	Scatter plot using labels as points	559
16.32	Scatter plot with point shape determined by gender	560
16.33	Scatter plot showing regression fits determined by gender	561
16.34	Scatter plots with regression fits by workshop and gender	562
16.35	Scatter plot matrix with lowess fits and density curves	563
16.36	Box plot of <code>posttest</code>	565
16.37	Box plot of <code>posttest</code> by group with jitter	566
16.38	Box plot of <code>posttest</code> by workshop and gender	567
16.39	Error bar plot of <code>posttest</code> by workshop	568
16.40	Map of USA using <code>path</code> geom	570
16.41	Map of USA using <code>polygon</code> geom	573
16.42	Multiframe demonstration plot	577

16.43	Scatter plot programmed several ways .....	579
17.1	Diagnostic plots for linear regression .....	620
17.2	Plot of Tukey HSD test .....	634