

Statistics and Computing

Series Editors:

John Chambers

David J. Hand

Wolfgang K. Härdle

Statistics and Computing

- Brusco/Stahl*: Branch and Bound Applications in Combinatorial Data Analysis
Chambers: Software for Data Analysis: Programming with R
Dalgaard: Introductory Statistics with R, 2nd ed.
Gentle: Elements of Computational Statistics
Gentle: Numerical Linear Algebra for Applications in Statistics
Gentle: Random Number Generation and Monte Carlo Methods, 2nd ed.
Härdle/Klinke/Turlach: XploRe: An Interactive Statistical Computing Environment
Hörmann/Leydold/Derflinger: Automatic Nonuniform Random Variate Generation
Krause/Olson: The Basics of S-PLUS, 4th ed.
Lange: Numerical Analysis for Statisticians
Lemmon/Schafer: Developing Statistical Software in Fortran 95
Loader: Local Regression and Likelihood
Marasinghe/Kennedy: SAS for Data Analysis: Intermediate Statistical Methods
Muenchen: R for SAS and SPSS Users
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
Unwin/Theus/Hofmann: Graphics of Large Datasets: Visualizing a Million
Venables/Ripley: Modern Applied Statistics with S, 4th ed.
Venables/Ripley: S Programming
Wilkinson: The Grammar of Graphics, 2nd ed.

Robert A. Muenchen

R for SAS and SPSS Users

 Springer

Robert A. Muenchen
University of Tennessee
Knoxville, TN, USA
muenchen.bob@gmail.com

S-PLUS® is a registered trademark of the Insightful Corporation.

SAS® is a registered trademark of SAS Institute.

SPSS® is a registered trademark of SPSS Inc.

Stata® is a registered trademark of Statacorp, Inc.

MATLAB® is a registered trademark of The Mathworks, Inc.

Windows Vista® and Windows XP® are registered trademarks of Microsoft, Inc.

Macintosh® and Mac OS® are registered trademarks of Apple, Inc.

Copyright © 2006, 2007, 2009 Robert A. Muenchen. All rights reserved.

ISBN: 978-0-387-09417-5 e-ISBN: 978-0-387-09418-2
DOI 10.1007/978-0-387-09418-2

Library of Congress Control Number: 2008931588

© Springer Science+Business Media, LLC 2009

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.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 be able to use R to:

- Read data from various types of text files and SAS/SPSS datasets.
- Manage your data through transformations or recodes, as well as splitting, merging and restructuring data sets.
- 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 cover much more complex methods.

Who Is This Book For?

This book is, perhaps obviously, for people who already know a statistics package. While aimed at SAS and SPSS users, many statistics packages share their main attributes, especially the use of a rectangular dataset as their only data structure. I expect users of most other statistics packages could benefit from this book. An audience I did not expect to serve is R users wanting to learn SAS or SPSS. 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. However, 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 Is This Book Not For?

I make no effort to teach statistics or graphics. Although I briefly state the goal and assumptions of each analysis, 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 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 John Chamber's *Software for Data Analysis: Programming with R* [1]. However, if you know SAS or SPSS, reading this book should be a nice leisurely step to take before diving into a book like that.

Practice Datasets and Programs

All the programs, datasets and files that we use in this book are available for download at <http://RforSASandSPSSusers.com>. Also check that site for updates and/or corrections to this book.

Acknowledgments

I am very grateful for the many people who have helped make this book possible, including the developers of the S language upon which R is based, Rick Becker, John Chambers, 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 the r-help participants who have kindly answered so many questions. Virtually all 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: Patrick Burns, Peter Flom, Martin Gregory, Ralph O'Brien, Charilaos Skiadas, Phil Spector, Michael Wexler, and seven anonymous reviewers who provided pages of invaluable advice to a neophyte.

A special thanks goes to Hadley Wickham, who provided much guidance on his `ggplot2` graphics package. Thanks to Gabor Grothendieck, Lauri Nikkinen, and Marc Schwarz and for the r-help discussion that led to Sect. 14.15. Thanks to Gabor Grothendieck also for a detailed discussion that led to Sect. 14.4. Thanks to Patrick Burns for his assistance with the glossary of terms in Appendix A.

My thanks also go to these people who helped compile *Appendix B: A Comparison of SAS and SPSS Products with R Packages and Functions* including: Thomas E. Adams, Jonathan Baron, Roger Bivand, Jason Burke, Patrick Burns, David L. Cassell, Chao Gai, Dennis Fisher, Bob Green, Frank E. Harrell Jr., Max Kuhn, Paul Murrell, Charilaos Skiadas, Antony Unwin, and Tobias Verbeke. Thanks to Henrique Dallazuanna for the code to count packages presented in that appendix.

I also thank SPSS Inc., especially Jon Peck for his helpful review of this book and his SPSS expertise which appears in the programs for extracting the first/last observation per group and generating data.

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

About the Author

Robert A. Muenchen is a consulting statistician with 28 years of experience. He is currently the manager of the Statistical Consulting Center at the University of Tennessee. He holds a B.A. in Psychology and an M.S. in Statistics. Bob has conducted research for a variety of public and private organizations and has assisted on more than 1,000 graduate theses and dissertations. He has coauthored over 40 articles published in scientific journals and conference proceedings.

Bob has served on the advisory boards of 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

1	Introduction	1
1.1	Why Learn R?	1
1.2	Is R Accurate?	2
1.3	What About Tech Support?	3
2	The Five Main Parts of SAS and SPSS	5
3	Programming Conventions	7
4	Typographic Conventions	9
5	Installing and Updating R	11
5.1	Installing Add-on Packages	11
5.2	Loading an Add-on Package	13
5.3	Updating Your Installation	15
5.4	Uninstalling R	16
5.5	Choosing Repositories	17
5.6	Accessing Data in Packages	18
6	Running R	21
6.1	Running R Interactively on Windows	21
6.2	Running R Interactively on Macintosh	23
6.3	Running R Interactively on Linux or UNIX	25
6.4	Running Programs that Include Other Programs	27
6.5	Running R in Batch Mode	27
6.6	Running R from SPSS	28
6.7	Graphical User Interfaces	32
6.7.1	R Commander	33
6.7.2	Rattle for Data Mining	34
6.7.3	JGR Java GUI for R	36

- 7 Help and Documentation** 41
 - 7.1 Help Files 41
 - 7.2 Starting Help 41
 - 7.3 Help Examples 42
 - 7.4 Help for Functions that Call Other Functions. 44
 - 7.5 Help for Packages. 44
 - 7.6 Help for Datasets 45
 - 7.7 Books and Manuals 45
 - 7.8 E-mail Lists. 45
 - 7.9 Searching the Web 46
 - 7.10 Vignettes. 47

- 8 Programming Language Basics** 49
 - 8.1 Simple Calculations 50
 - 8.2 Data Structures. 51
 - 8.2.1 Vectors. 51
 - 8.2.2 Factors. 53
 - 8.2.3 Data Frames 55
 - 8.2.4 Matrices. 57
 - 8.2.5 Arrays 59
 - 8.2.6 Lists 59
 - 8.3 Saving Your Work So Far 60
 - 8.4 Comments to Document Your Programs 62
 - 8.5 Controlling Functions (Procedures). 62
 - 8.5.1 Controlling Functions with Arguments 62
 - 8.5.2 Controlling Functions with Formulas. 64
 - 8.5.3 Controlling Functions with an Object’s Class. 65
 - 8.5.4 Controlling Functions with Extractor
Functions – ODS, OMS 67
 - 8.5.5 How Much Output Is There? 69
 - 8.5.6 Writing Your Own Functions (Macros) 73

- 9 Data Acquisition** 77
 - 9.1 The R Data Editor 77
 - 9.2 Reading Delimited Text Files. 79
 - 9.3 Reading Text Data Within a Program (Datalines, Cards,
Begin Data. . .) 84
 - 9.4 Reading Data from the Keyboard 86
 - 9.5 Reading Fixed-Width Text Files, One Record per Case 87
 - 9.5.1 Macro Substitution 90
 - 9.6 Reading Fixed-Width Text Files, Two or More Records per
Case 92
 - 9.7 Importing Data from SAS 95
 - 9.8 Importing Data from SPSS 96

- 9.9 Exporting Data 97
 - 9.9.1 Viewing an External Text File 98

- 10 Selecting Variables – Var, Variables =** 103
 - 10.1 Selecting Variables in SAS and SPSS 103
 - 10.2 Selecting All Variables 104
 - 10.3 Selecting Variables by Index Number 104
 - 10.4 Selecting Variables by Column Name 107
 - 10.5 Selecting Variables Using Logic 108
 - 10.6 Selecting Variables by String Search (varname: or varname1-varnameN) 110
 - 10.7 Selecting Variables Using \$ Notation 112
 - 10.8 Selecting Variables by Simple Name: attach and with . . 113
 - 10.9 Selecting Variables with the subset Function (varname1-varnameN) 114
 - 10.10 Selecting Variables by List 115
 - 10.11 Generating Indexes A to Z from Two Variable Names . . . 115
 - 10.12 Saving Selected Variables to a New Dataset 116
 - 10.13 Example Programs for Variable Selection 116

- 11 Selecting Observations – Where, If, Select If, Filter** 123
 - 11.1 Selecting Observations in SAS and SPSS 123
 - 11.2 Selecting All Observations 124
 - 11.3 Selecting Observations by Index Number 124
 - 11.4 Selecting Observations by Row Name 127
 - 11.5 Selecting Observations Using Logic 128
 - 11.6 Selecting Observations by String Search 132
 - 11.7 Selecting Observations with the subset Function 133
 - 11.8 Generating Indexes from A to Z from Two Row Names . . 134
 - 11.9 Variable Selection Methods with No Counterpart for Selecting Observations 135
 - 11.10 Saving Selected Observations to a New Data Frame 135
 - 11.11 Example Programs for Selecting Observations 135

- 12 Selecting Both Variables and Observations** 141

- 13 Converting Data Structures** 143
 - 13.1 Converting from Logical to Index and Back 146

- 14 Data Management** 147
 - 14.1 Transforming Variables 147
 - 14.2 Procedures or Functions? The apply Function Decides. . . 152
 - 14.2.1 Applying the mean Function 152
 - 14.2.2 Finding N or NVALID 155
 - 14.3 Conditional Transformations 158

- 14.4 Multiple Conditional Transformations 162
- 14.5 Missing Values 165
 - 14.5.1 Substituting Means for Missing Values. 166
 - 14.5.2 Finding Complete Observations 167
 - 14.5.3 When “99” Has Meaning. 168
- 14.6 Renaming Variables (. . . and Observations). 171
- 14.7 Renaming Variables – Advanced Examples. 174
 - 14.7.1 Renaming by Index 174
 - 14.7.2 Renaming by Column Name. 175
 - 14.7.3 Renaming Many Sequentially Numbered Variable Names 176
 - 14.7.4 Renaming Observations 177
- 14.8 Recoding Variables. 180
 - 14.8.1 Recoding a Few Variables. 180
 - 14.8.2 Recoding Many Variables 181
- 14.9 Keeping and Dropping Variables. 185
- 14.10 Stacking/Concatenating/Adding Datasets 186
- 14.11 Joining/Merging Data Frames 190
- 14.12 Creating Summarized or Aggregated Datasets 194
 - 14.12.1 The aggregate Function 195
 - 14.12.2 The tapply Function. 196
 - 14.12.3 Merging Aggregates with Original Data 198
 - 14.12.4 Tabular Aggregation 200
 - 14.12.5 The reshape Package 201
- 14.13 By or Split File Processing 204
 - 14.13.1 Comparing Summarization Methods 208
 - 14.13.2 Example Programs for By or Split File Processing 208
- 14.14 Removing Duplicate Observations. 210
- 14.15 Selecting First or Last Observations per Group. 213
- 14.16 Reshaping Variables to Observations and Back 217
- 14.17 Sorting Data Frames 221

- 15 Value Labels or Formats (and Measurement Level) 225**
 - 15.1 Character Factors. 226
 - 15.2 Numeric Factors. 227
 - 15.3 Making Factors of Many Variables 229
 - 15.4 Converting Factors into Numeric or Character Variables. . 232
 - 15.5 Dropping Factor Levels 233

- 16 Variable Labels 239**

- 17 Generating Data. 245**
 - 17.1 Generating Numeric Sequences 245
 - 17.2 Generating Factors. 246
 - 17.3 Generating Repetitious Patterns (not factors) 247

- 17.4 Generating Integer Measures 248
- 17.5 Generating Continuous Measures 249
- 17.6 Generating a Data Frame. 251
- 18 How R Stores Data 259**
- 19 Managing Your Files and Workspace 261**
 - 19.1 Loading and Listing Objects 261
 - 19.2 Understanding Your Search Path 264
 - 19.3 Attaching Data Frames 264
 - 19.4 Attaching Files 266
 - 19.5 Removing Objects from Your Workspace 267
 - 19.6 Minimizing Your Workspace. 268
 - 19.7 Setting Your Working Directory 268
 - 19.8 Saving Your Workspace. 269
 - 19.9 Saving Your Programs and Output 271
 - 19.10 Saving Your History (Journal). 271
- 20 Graphics Overview 273**
 - 20.1 SAS/GRAPH 273
 - 20.2 SPSS Graphics 274
 - 20.3 R Graphics 274
 - 20.4 The Grammar of Graphics. 275
 - 20.5 Other Graphics Packages 276
 - 20.6 Graphics Procedures Versus Graphics Systems 277
 - 20.7 Graphics Devices 277
 - 20.8 Practice Data: Mydata100 278
- 21 Traditional Graphics 281**
 - 21.1 Barplots 281
 - 21.1.1 Barplots of Counts. 281
 - 21.1.2 Barplots for Subgroups of Counts. 285
 - 21.1.3 Barplots of Means 286
 - 21.2 Adding Titles, Labels, Colors, and Legends. 288
 - 21.3 Graphics Parameters and Multiple Plots on a Page. 290
 - 21.4 Pie Charts 292
 - 21.5 Dotcharts 293
 - 21.6 Histograms 293
 - 21.6.1 Basic Histograms. 294
 - 21.6.2 Histograms Overlaid 297
 - 21.7 Normal QQ Plots 299
 - 21.8 Strip Charts. 301
 - 21.9 Scatterplots. 303
 - 21.9.1 Scatterplots with Jitter. 304
 - 21.9.2 Scatterplots with Large Datasets. 305

- 21.9.3 Scatterplots with Lines 307
- 21.9.4 Scatterplots with Linear Fit by Group 308
- 21.9.5 Scatterplots by Group or Level (Coplots) 309
- 21.9.6 Scatterplots with Confidence Ellipse 311
- 21.9.7 Scatterplots with Confidence and Prediction
Intervals 312
- 21.9.8 Plotting Labels Instead of Points 316
- 21.9.9 Scatterplot Matrices 318
- 21.10 Dual Axes Plots 320
- 21.11 Boxplots 322
- 21.12 Error Bar and Interaction Plots 324
- 21.13 Adding Equations and Symbols to Graphs 324
- 21.14 Summary of Graphics Elements and Parameters 325
- 21.15 Plot Demonstrating Many Modifications 328
- 21.16 Example Traditional Graphics Programs 330

- 22 Graphics with `ggplot2` (GPL)** 341
 - 22.1 Overview `qplot` and `ggplot` 342
 - 22.2 Bar Charts 344
 - 22.3 Pie Charts 347
 - 22.4 Bar Charts with Subgroups 348
 - 22.5 Plots by Group or Level 349
 - 22.6 Pre-summarized Data 351
 - 22.7 Dotcharts 352
 - 22.8 Adding Titles and Labels 353
 - 22.9 Histograms 354
 - 22.10 Normal QQ Plots 359
 - 22.11 Strip Plots 360
 - 22.12 Scatterplots 361
 - 22.13 Scatterplots with Jitter 363
 - 22.13.1 Scatterplots with Large Datasets 364
 - 22.14 Scatterplots with Fit Lines 367
 - 22.15 Scatterplots with Reference Lines 368
 - 22.16 Plotting Labels Instead of Points 371
 - 22.17 Changing Plot Symbols by Group 372
 - 22.18 Adding Linear Fits by Group 373
 - 22.19 Scatterplots Faceted by Groups 374
 - 22.20 Scatterplot Matrix 374
 - 22.21 Boxplots 376
 - 22.22 Error Barplots 380
 - 22.23 Logarithmic Axes 381
 - 22.24 Aspect Ratio 382
 - 22.25 Multiple Plots on a Page 382
 - 22.26 Saving `ggplot2` Graphs to a File 385

22.27 An Example Specifying All Defaults 385

22.28 Summary of Graphic Elements and Parameters 386

23 Statistics 403

23.1 Scientific Notation 403

23.2 Descriptive Statistics. 404

23.3 Cross-Tabulation 408

23.4 Correlation 413

23.5 Linear Regression. 417

 23.5.1 Plotting Diagnostics 420

 23.5.2 Comparing Models 421

 23.5.3 Making Predictions with New Data 422

23.6 t-Test – Independent Groups 422

23.7 Equality of Variance. 424

23.8 *t*-Test – Paired or Repeated Measures 424

23.9 Wilcoxon Mann–Whitney Rank Sum Test – Independent
Groups 425

23.10 Wilcoxon Signed-Rank Test – Paired Groups 426

23.11 Analysis of Variance. 427

23.12 Sums of Squares 431

23.13 Kruskal–Wallis Test 432

24 Conclusion 441

Appendix A A Glossary of R Jargon 443

**Appendix B A Comparison of SAS and SPSS Products with R Packages
and Functions 449**

Appendix C Automating Your Settings 453

**Appendix D A comparison of the major attributes of
SAS and SPSS to R 457**

Bibliography 459

Index 463