Using R and RStudio for Data Management, Statistical Analysis and Graphics

Second Edition

Nicholas J. Horton
Department of Mathematics and Statistics
Amherst College
Massachusetts, U.S.A.

Ken Kleinman
Department of Population Medicine
Harvard Medical School and
Harvard Pilgrim Health Care Institute
Boston, Massachusetts, U.S.A.

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Preface to the second edition

Software systems such as R evolve rapidly, and so do the approaches and expertise of statistical analysts.

In 2009, we began a blog in which we explored many new case studies and applications, ranging from generating a Fibonacci series to fitting finite mixture models with concomitant variables. We also discussed some additions to R, the RStudio integrated development environment, and new or improved R packages. The blog now has hundreds of entries and according to Google Analytics has received hundreds of thousands of visits.

The volume you are holding is a larger format and longer than the first edition, and much of the new material is adapted from these blog entries, while it also includes other improvements and additions that have emerged in the last few years.

We have extensively reorganized the material in the book and created three new chapters. The firsts "Simulation," includes examples where data are generated from complex models such as mixed-effects models and survival models, and from distributions using the Metropolis–Hastings algorithm. We also explore interesting statistics and probability examples via simulation. The second is "Special topics," where we describe some key features, such as processing by group, and detail several important areas of statistics, including Bayesian methods, propensity scores, and bootstrapping. The last is "Case studies," where we demonstrate examples of useful data management tasks, read complex files, make and annotate maps, show how to "scrape" data from the web, mine text files, and generate dynamic graphics.

We also describe RStudio in detail. This powerful and easy-to-use front end adds innumerable features to R. In our experience, it dramatically increases the productivity of R users, and by tightly integrating reproducible analysis tools, helps avoid error-prone "cut and paste" workflows. Our students and colleagues find RStudio an extremely comfortable interface.

We used a reproducible analysis system (`knitr`) to generate the example code and output in the book. Code extracted from these files is provided on the book website. In this edition, we provide a detailed discussion of the philosophy and use of these systems. In particular, we feel that the `knitr` and `markdown` packages for R, which are tightly integrated with RStudio, should become a part of every R user's toolbox. We can't imagine working on a project without them.

The second edition of the book features extensive use of a number of new packages that extend the functionality of the system. Some of these include `dplyr` (tools for working with dataframe-like objects and databases), `ggplot2` (implementation of the Grammar of Graphics), `ggmap` (spatial mapping using `ggplot2`), `ggvis` (to build interactive graphical displays), `httr` (tools for working with URLs and HTTP), `lubridate` (date and time manipulations), `markdown` (for simplified reproducible analysis), `shiny` (to build interactive web applications), `swirl` (for learning R, in R), `tidyr` (for data manipulation), and `xtable` (to create publication-quality tables). Overall, these packages facilitate ever more sophisticated analyses.

Finally, we've reorganized much of the material from the first edition into smaller, more focused chapters. Readers will now find separate (and enhanced) chapters on data input and output, data management, statistical and mathematical functions, and programming, rather than a single chapter on "data management." Graphics are now discussed in two chapters: one on high-level types of plots such as scatterplots and histograms, and another on customizing the fine details of the plots, such as the number of tick marks and the color of plot symbols.

We're immensely gratified by the positive response the first edition elicited, and hope the current volume will be even more useful to you.

On the web

The book website at `http://www.amherst.edu/~nhorton/r2` includes the table of contents, the indices, the HELP dataset in various formats, example code, a pointer to the blog, and a list of errata.

Acknowledgments

In addition to those acknowledged in the first edition, we would like to thank J.J. Allaire and the RStudio developers, Danny Kaplan, Deborah Nolan, Daniel Parel, Randall Pruim, Romain Francois, and Hadley Wickham, plus the many individuals who have created and shared R packages. Their contributions to R and RStudio, programming efforts, comments, and guidance and/or helpful suggestions on drafts of the revision have been extremely helpful. Above all, we greatly appreciate Sara and Julia as well as Abby, Alana, Kinari, and Sam, for their patience and support.

Amherst, MA
October 2014

Preface to the first edition

R (R development core team, 2009) is a general purpose statistical software package used in many fields of research. It is licensed for free, as open-source software. The system is developed by a large group of people, almost all volunteers. It has a large and growing user and developer base. Methodologists often release applications for general use in R shortly after they have been introduced into the literature. While professional customer support is not provided, there are many resources to help support users.

We have written this book as a reference text for users of R. Our primary goal is to provide users with an easy way to learn how to perform an analytic task in this system, without having to navigate through the extensive, idiosyncratic, and sometimes unwieldy documentation or to sort through the huge number of add-on packages. We include many common tasks, including data management, descriptive summaries, inferential procedures, regression analysis, multivariate methods, and the creation of graphics. We also show some more complex applications. In toto, we hope that the text will facilitate more efficient use of this powerful system.

We do not attempt to exhaustively detail all possible ways available to accomplish a given task in each system. Neither do we claim to provide the most elegant solution. We have tried to provide a simple approach that is easy to understand for a new user, and have supplied several solutions when it seems likely to be helpful.

Who should use this book

Those with an understanding of statistics at the level of multiple-regression analysis should find this book helpful. This group includes professional analysts who use statistical packages almost every day as well as statisticians, epidemiologists, economists, engineers, physicians, sociologists, and others engaged in research or data analysis. We anticipate that this tool will be particularly useful for sophisticated users, those with years of experience in only one system, who need or want to use the other system. However, intermediate-level analysts should reap the same benefit. In addition, the book will bolster the analytic abilities of a relatively new user, by providing a concise reference manual and annotated examples.

Using the book

The book has two indices, in addition to the comprehensive table of contents. These include: 1) a detailed topic (subject) index in English; 2) an R command index, describing R syntax.

Extensive example analyses of data from a clinical trial are presented; see Table B.1 (p. 237) for a comprehensive list. These employ a single dataset (from the HELP study), described in Appendix B. Readers are encouraged to download the dataset and code from the book website. The examples demonstrate the code in action and facilitate exploration by the reader.

In addition to the HELP examples, a case studies and extended examples chapter utilizes many of the functions, idioms and code samples introduced earlier. These include explications of analytic and empirical power calculations, missing data methods, propensity score analysis, sophisticated data manipulation, data gleaning from websites, map making, simulation studies, and optimization. Entries from earlier chapters are cross-referenced to help guide the reader.

Where to begin

We do not anticipate that the book will be read cover to cover. Instead, we hope that the extensive indexing, cross-referencing, and worked examples will make it possible for readers to directly find and then implement what they need. A new user should begin by reading the first chapter, which includes a sample session and overview of the system. Experienced users may find the case studies to be valuable as a source of ideas on problem solving in R.

Acknowledgments

We would like to thank Rob Calver, Kari Budyk, Shashi Kumar, and Sarah Morris for their support and guidance at Informa CRC/Chapman and Hall. We also thank Ben Cowling, Stephanie Greenlaw, Tanya Hakim, Albyn Jones, Michael Lavine, Pamela Matheson, Elizabeth Stuart, Rebbecca Wilson, and Andrew Zieffler for comments, guidance and/or helpful suggestions on drafts of the manuscript.

Above all we greatly appreciate Julia and Sara as well as Abby, Alana, Kinari, and Sam, for their patience and support.

Northampton, MA and Amherst, MA
February, 2010