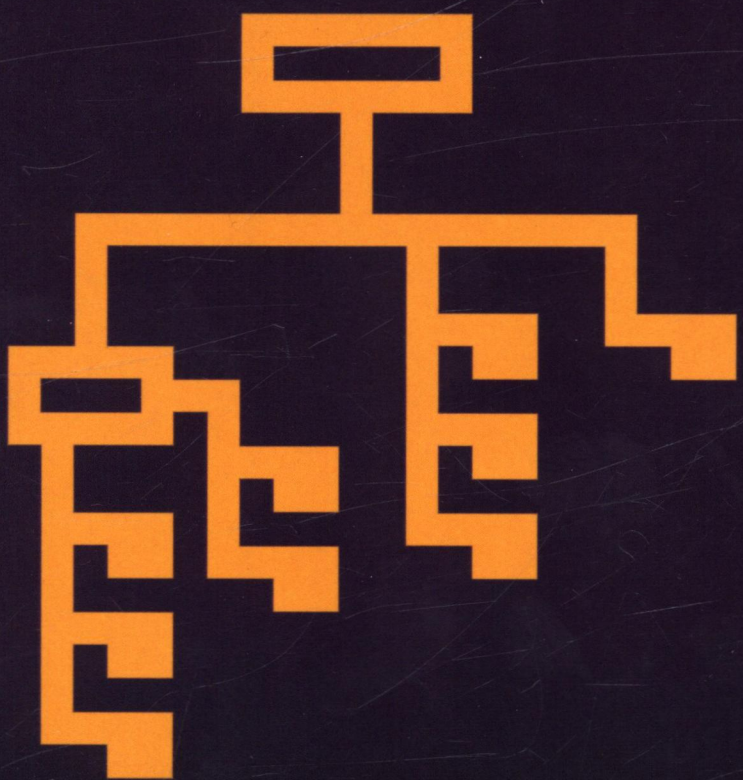
The R Series

Reproducible Research with R and RStudio

Second Edition



Christopher Gandrud

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Statistics

Reproducible Research with R and RStudio, Second Edition brings together the skills and tools needed for doing and presenting computational research. Using straightforward examples, the book takes you through an entire reproducible research workflow. This practical workflow enables you to gather and analyze data as well as dynamically present results in print and on the web.

New to the Second Edition

- The *rmarkdown* package that allows you to create reproducible research documents in PDF, HTML, and Microsoft Word formats using the simple and intuitive Markdown syntax
- Improvements to RStudio's interface and capabilities, such as its new tools for handling R Markdown documents
- Expanded *knitr* R code chunk capabilities
- The kable function in the *knitr* package and the *texreg* package for dynamically creating tables to present your data and statistical results
- An improved discussion of file organization, enabling you to take full advantage of relative file paths so that your documents are more easily reproducible across computers and systems
- The *dplyr, magrittr*, and *tidyr* packages for fast data manipulation
- Numerous modifications to R syntax in user-created packages
- Changes to GitHub's and Dropbox's interfaces

This updated book provides all the tools to combine your research with the presentation of your findings. It saves you time searching for information so that you can spend more time actually addressing your research questions. Supplementary files used for the examples and a reproducible research project are available on the author's website.

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

Second Edition

Reproducible Research with R and RStudio

Gandrud

CRC

Reproducible Research with R and RStudio

Second Edition

Christopher Gandrud
Hertie School of Governance
Berlin, Germany

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

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

First issued in hardback 2017

ISBN-13: 978-1-4987-1537-9 (pbk)
ISBN-13: 978-1-1384-6964-8 (hbk)

Reproducible Research with R and RStudio

Second Edition

Chapman & Hall/CRC
The R Series

Series Editors

John M. Chambers
Department of Statistics
Stanford University
Stanford, California, USA

Torsten Hothorn
Division of Biostatistics
University of Zurich
Switzerland

Duncan Temple Lang
Department of Statistics
University of California, Davis
Davis, California, USA

Hadley Wickham
RStudio
Boston, Massachusetts, USA

Aims and Scope

This book series reflects the recent rapid growth in the development and application of R, the programming language and software environment for statistical computing and graphics. R is now widely used in academic research, education, and industry. It is constantly growing, with new versions of the core software released regularly and more than 6,000 packages available. It is difficult for the documentation to keep pace with the expansion of the software, and this vital book series provides a forum for the publication of books covering many aspects of the development and application of R.

The scope of the series is wide, covering three main threads:

- Applications of R to specific disciplines such as biology, epidemiology, genetics, engineering, finance, and the social sciences.
- Using R for the study of topics of statistical methodology, such as linear and mixed modeling, time series, Bayesian methods, and missing data.
- The development of R, including programming, building packages, and graphics.

The books will appeal to programmers and developers of R software, as well as applied statisticians and data analysts in many fields. The books will feature detailed worked examples and R code fully integrated into the text, ensuring their usefulness to researchers, practitioners and students.

Published Titles

Stated Preference Methods Using R, *Hideo Aizaki, Tomoaki Nakatani, and Kazuo Sato*

Using R for Numerical Analysis in Science and Engineering, *Victor A. Bloomfield*

Event History Analysis with R, *Göran Broström*

Computational Actuarial Science with R, *Arthur Charpentier*

Statistical Computing in C++ and R, *Randall L. Eubank and Ana Kupresanin*

Reproducible Research with R and RStudio, Second Edition, *Christopher Gandrud*

Introduction to Scientific Programming and Simulation Using R, Second Edition, *Owen Jones, Robert Maillardet, and Andrew Robinson*

Nonparametric Statistical Methods Using R, *John Kloke and Joseph McKean*

Displaying Time Series, Spatial, and Space-Time Data with R, *Oscar Perpiñán Lamigueiro*

Programming Graphical User Interfaces with R, *Michael F. Lawrence and John Verzani*

Analyzing Sensory Data with R, *Sébastien Lê and Theirry Worch*

Parallel Computing for Data Science: With Examples in R, C++ and CUDA, *Norman Matloff*

Analyzing Baseball Data with R, *Max Marchi and Jim Albert*

Growth Curve Analysis and Visualization Using R, *Daniel Mirman*

R Graphics, Second Edition, *Paul Murrell*

Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving, *Deborah Nolan and Duncan Temple Lang*

Multiple Factor Analysis by Example Using R, *Jérôme Pagès*

Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R, *Daniel S. Putler and Robert E. Krider*

Implementing Reproducible Research, *Victoria Stodden, Friedrich Leisch, and Roger D. Peng*

Graphical Data Analysis with R, *Antony Unwin*

Using R for Introductory Statistics, Second Edition, *John Verzani*

Advanced R, *Hadley Wickham*

Dynamic Documents with R and knitr, Second Edition, *Yihui Xie*

Preface

This book has its genesis in my PhD research at the London School of Economics. I started the degree with questions about the 2008/09 financial crisis and planned to spend most of my time researching capital adequacy requirements. But I quickly realized that I would actually spend a large proportion of my time learning the day-to-day tasks of data gathering, analysis, and results presentation. After plodding through for a while with Word, Excel, and Stata, my breaking point came while reentering results into a regression table after I had tweaked one of my statistical models, yet again. Surely there was a better way to *do* research that would allow me to spend more time answering my research questions. Making research reproducible for others also means making it better organized and efficient for yourself. My search for a better way led me straight to the tools for reproducible computational research.

The reproducible research community is very active, knowledgeable, and helpful. Nonetheless, I often encountered holes in this collective knowledge, or at least had no resource organize it all together as a whole. That is my intention for this book: to bring together the skills I have picked up for actually doing and presenting computational research. Hopefully, the book, along with making reproducible research more widely used, will save researchers hours of googling, so they can spend more time addressing their research questions.

Changes to the Second Edition

The tools of reproducible research have developed rapidly since the first edition of this book was published just two years ago. The second edition has been updated to incorporate the most important of these advancements, including discussions of:

- The *rmarkdown* package, which allows you to create reproducible research documents in PDF, HTML, and Microsoft Word formats using the simple and intuitive Markdown syntax.
- Improvements and changes to RStudio's interface and capabilities, such as its new tools for handling R Markdown documents.
- Expanded *knitr* R code chunk capabilities.
- The `kable` function in the *knitr* package and the *texreg* package for dynamically creating tables to present your data and statistical results.

- An improved discussion of file organization allowing you to take full advantage of relative file paths so that your documents are more easily reproducible across computers and systems.
- The *dplyr*, *magrittr*, and *tidyr* packages for fast data manipulation.
- Numerous changes to R syntax in user-created packages.
- Changes to GitHub's and Dropbox's interfaces.

Acknowledgements

I would not have been able to write this book without many people's advice and support. Foremost is John Kimmel, acquisitions editor at Chapman and Hall. He approached me in Spring 2012 with the general idea and opportunity for this book. Other editors at Chapman and Hall and Taylor and Francis have greatly contributed to this project, including Marcus Fontaine. I would also like to thank all of the book's reviewers whose helpful comments have greatly improved it. The first edition's reviewers include:

- Jeromy Anglim, Deakin University
- Karl Broman, University of Wisconsin, Madison
- Jake Bowers, University of Illinois, Urbana-Champaign
- Corey Chivers, McGill University
- Mark M. Fredrickson, University of Illinois, Urbana-Champaign
- Benjamin Lauderdale, London School of Economics
- Ramnath Vaidyanathan, McGill University

The developer and blogging community has also been incredibly important for making this book possible. Foremost among these people is Yihui Xie. He is the main developer behind the *knitr* package, co-developer of *rmarkdown*, and also an avid blog writer and commenter. Without him the ability to do reproducible research would be much harder and the blogging community that spreads knowledge about how to do these things would be poorer. Other great contributors to the reproducible research community include Carl Boettiger, Karl Broman, Markus Gesmann (who developed *googleVis*), Rob Hyndman, and Hadley Wickham (who has developed numerous very useful R packages). Thank you also to Victoria Stodden and Michael Malecki for helpful suggestions. And, of course, thank you to everyone at RStudio (especially JJ Allaire) for creating an increasingly useful program for reproducible research.

The second edition has benefited immensely from first edition readers' comments and suggestions. For a list of their valuable contributions, please see the book's GitHub Issues page `https://GitHub.com/christophergandrud/Rep-Res-Book/issues` and the first edition's Errata page `http://christophergandrud.GitHub.io/RepResR-RStudio/errata.htm`.

My students at Yonsei University were an important part of making the first edition. One of the reasons that I got interested in using many of the tools covered in this book, like using *knitr* in slideshows, was to improve a course I taught there: Introduction to Social Science Data Analysis. I tested many of the explanations and examples in this book on my students. Their feedback has been very helpful for making the book clearer and more useful. Their experience with using these tools on Microsoft Windows computers was also important for improving the book's Windows documentation. Similarly, my students at the Hertie School of Governance inspired and tested key sections of the second edition.

The vibrant community at Stack Overflow `http://stackoverflow.com/` and Stack Exchange `http://stackexchange.com/` are always very helpful for finding answers to problems that plague any computational researcher. Importantly, the sites make it easy for others to find the answers to questions that have already been asked.

My wife, Kristina Gandrud, has been immensely supportive and patient with me throughout the writing of this book (and pretty much my entire academic career). Certainly this is not the proper forum for musing about marital relations, but I'll do a musing anyways. Having a person who supports your interests, even if they don't completely share them, is immensely helpful for a researcher. It keeps you going.

Stylistic Conventions

I use the following conventions throughout this book:

- **Abstract variables**: Abstract variables, i.e. variables that do not represent specific objects in an example, are in `ALL CAPS TYPEWRITER TEXT`.
- **Clickable buttons**: Clickable Buttons are in `typewriter text`.
- **Code**: All code is in `typewriter text`.
- **Filenames and directories**: Filenames and directories more generally are printed in *italics*. I use CamelBack for file and directory names.
- **File extensions**: Like filenames, file extensions are *italicized*.
- **Individual variable values**: Individual variable values mentioned in the text are in *italics*.
- **Objects**: Objects are printed in *italics*. I use CamelBack for object names.
- **Object columns**: Data frame object columns are printed in *italics*.
- **Packages**: **R** packages are printed in *italics*.
- **Windows and RStudio panes**: Open windows and RStudio panes are written in *italics*.
- **Variable names**: Variable names are printed in **bold**. I use CamelBack for individual variable names.

Required R Packages

In this book I discuss how to use a number of user-written R packages for reproducible research. Many of these packages are not included in the default R installation. They need to be installed separately.

Note: in general you should aim to minimize the number of packages that your research depends on. Doing so will lessen the possibility that your code will "break" when a package is updated. This book depends on relatively many packages because of its special and unusual purpose of illustrating a variety of tools that you can use for reproducible research.

To install key user-written packages discussed in this book, copy the following code and paste it into your R console:

```
install.packages(c("brew", "countrycode",
                   "devtools", "dplyr",
                   "ggplot2", "googleVis",
                   "knitr", "MCMCpack",
                   "repmis", "RCurl",
                   "rmarkdown", "texreg",
                   "tidyr", "WDI",
                   "xtable", "Zelig"))
```

Once you enter this code, you may be asked to select a CRAN "mirror" to download the packages from.[1] Simply select the mirror closest to you.

In Chapter 9 we use the *Zelig* package (Owen et al., 2013) to create a simple Bayesian normal linear regression. For this to work properly you will need to install an additional package called *ZeligBayesian* (Owen, 2011). To do this, type the following code into your R console:

```
install.packages("ZeligBayesian",
                 repos = "http://r.iq.harvard.edu/",
                 type  = "source")
```

[1] CRAN stands for the Comprehensive R Archive Network.

Special issues for Windows and Linux Users

If you are using Windows, you will also need to install *Rtools* (Ripley and Murdoch, 2012). You can download *Rtools* from: `http://cran.r-project.org/bin/windows/Rtools/`. Please use the recommended installation to ensure that your system PATH is set up correctly. Otherwise your computer will not know where the tools are.

On Linux you will need to install the *RCurl* (Temple Lang, 2015) and *XML* (Temple Lang, 2013) packages separately. Use your Terminal to install these packages with the following code:

```
sudo apt-get update

sudo apt-get install libcurl4-gnutls-dev
sudo apt-get install libxml2-dev
sudo apt-get install r-cran-xml
sudo apt-get install r-cran-rjava
```