Christopher Gandrud

Reproducible Research with R and RStudio (Third Edition)

To my wife,

who is currently at the movies with our son, so that I can finish the third edition.

Contents

List of Tables

List of Figures

Preface

My motivation

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 Third Edition

To WRITE

- The book is created using *bookdown* (**?**), a format that builds on *rmarkdown* to compile books into many different formats.

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.

Acknowledgments

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

and there have been many other annonymous reviewers who have provided great feedback over the years.

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.

Additional Resources

Additional resources that supplement the examples in this book can be freely downloaded and experimented with. These resources include longer examples discussed in individual chapters and a complete short reproducible research project.

Chapter Examples

Longer examples discussed in individual chapters, including files to dynamically download data, code for creating figures, and markup files for creating presentation documents, can be accessed at: <`https://GitHub.com/christophergandrud/Rep-Res-Examples`}. Please see Chapter **??** for more information on downloading files from GitHub, where the examples are stored.

Short Example Project

To download a full (though very short) example of a reproducible research project created using the tools covered in this book go to: `https://GitHub.com/christophergandrud/Rep-Res-ExampleProject1`. Please follow the replication instructions in the main *README.md* file to fully replicate the project. It is probably a good idea to hold off looking at this complete example in detail until after you have become acquainted with the individual tools it uses. Become acquainted with the tools by reading through this book and working with the individual chapter examples.

The following two figures give you a sense of how the example's files are organized. Figure **??** shows how the files are organized in the file system. Figure **??** illustrates how the main files are dynamically tied together. In the *Data* directory we have files to gather raw data from the **?** on fertilizer consumption and from **?** on countries' levels of democracy. They are tied to the