Christopher Gandrud

Reproducible Research with R and RStudio Second Edition

Contents

Preface **xiii**

Stylistic Conventions **xvii**

Required R Packages **xix**

Additional Resources **xxi**

List of Figures **xxv**

List of Tables **xxvii**

I Getting Started **1**

1 Introducing Reproducible Research **3**
1.1 What Is Reproducible Research? . . . 3
1.2 Why Should Research Be Reproducible? . . . 5
1.2.1 For science . . . 5
1.2.2 For you . . . 6
1.3 Who Should Read This Book? . . . 8
1.3.1 Academic researchers . . . 8
1.3.2 Students . . . 8
1.3.3 Instructors . . . 8
1.3.4 Editors . . . 9
1.3.5 Private sector researchers . . . 9
1.4 The Tools of Reproducible Research . . . 10
1.5 Why Use R, *knitr*/*rmarkdown*, and RStudio for Reproducible Research? . . . 11
1.5.1 Installing the main software . . . 13
1.6 Book Overview . . . 14
1.6.1 How to read this book . . . 16
1.6.2 Reproduce this book . . . 16
1.6.3 Contents overview . . . 17

2 Getting Started with Reproducible Research **19**
2.1 The Big Picture: A Workflow for Reproducible Research . . . 19
2.1.1 Reproducible theory . . . 20

2.2 Practical Tips for Reproducible Research 22
2.2.1 Document everything! 22
2.2.2 Everything is a (text) file 24
2.2.3 All files should be human readable 24
2.2.4 Explicitly tie your files together 26
2.2.5 Have a plan to organize, store, and make your files available 27

3 Getting Started with R, RStudio, and knitr/rmarkdown **29**
3.1 Using R: The Basics 29
3.1.1 Objects 30
3.1.2 Component selection 36
3.1.3 Subscripts 38
3.1.4 Functions and commands 39
3.1.5 Arguments 40
3.1.6 The workspace & history 42
3.1.7 Global R options 44
3.1.8 Installing new packages and loading functions 44
3.2 Using RStudio 45
3.3 Using *knitr* and *rmarkdown*: The Basics 47
3.3.1 What *knitr* does 48
3.3.2 What *rmarkdown* does 48
3.3.3 File extensions 50
3.3.4 Code chunks 50
3.3.5 Global chunk options 53
3.3.6 *knitr* package options 55
3.3.7 Hooks 55
3.3.8 *knitr*, *rmarkdown*, & RStudio 56
3.3.9 *knitr* & R 59
3.3.10 *rmarkdown* and R 61

4 Getting Started with File Management 65
4.1 File Paths & Naming Conventions 66
4.1.1 Root directories 66
4.1.2 Subdirectories & parent directories 66
4.1.3 Working directories 67
4.1.4 Absolute vs. relative paths 67
4.1.5 Spaces in directory & file names 68
4.2 Organizing Your Research Project 69
4.3 Setting Directories as RStudio Projects 70
4.4 R File Manipulation Commands 70
4.5 Unix-like Shell Commands for File Management 74
4.6 File Navigation in RStudio 78

II Data Gathering and Storage 81

5 Storing, Collaborating, Accessing Files, and Versioning 83
5.1 Saving Data in Reproducible Formats . . . 84
5.2 Storing Your Files in the Cloud: Dropbox . . . 85
5.2.1 Storage . . . 86
5.2.2 Accessing data . . . 86
5.2.3 Collaboration . . . 88
5.2.4 Version control . . . 88
5.3 Storing Your Files in the Cloud: GitHub . . . 89
5.3.1 Setting up GitHub: Basic . . . 91
5.3.2 Version control with Git . . . 92
5.3.3 Remote storage on GitHub . . . 100
5.3.4 Accessing on GitHub . . . 102
5.3.4.1 Collaboration with GitHub . . . 104
5.3.5 Summing up the GitHub workflow . . . 105
5.4 RStudio & GitHub . . . 105
5.4.1 Setting up Git/GitHub with Projects . . . 105
5.4.2 Using Git in RStudio Projects . . . 107

6 Gathering Data with R 109
6.1 Organize Your Data Gathering: Makefiles . . . 109
6.1.1 R Make-like files . . . 110
6.1.2 GNU Make . . . 111
6.1.2.1 Example makefile . . . 112
6.1.2.2 Makefiles and RStudio Projects . . . 116
6.1.2.3 Other information about makefiles . . . 116
6.2 Importing Locally Stored Data Sets . . . 117
6.3 Importing Data Sets from the Internet . . . 118
6.3.1 Data from non-secure (`http`) URLs . . . 118
6.3.2 Data from secure (`https`) URLs . . . 119
6.3.3 Compressed data stored online . . . 121
6.3.4 Data APIs & feeds . . . 123
6.4 Advanced Automatic Data Gathering: Web Scraping . . . 125

7 Preparing Data for Analysis 129
7.1 Cleaning Data for Merging . . . 129
7.1.1 Get a handle on your data . . . 129
7.1.2 Reshaping data . . . 131
7.1.3 Renaming variables . . . 134
7.1.4 Ordering data . . . 134
7.1.5 Subsetting data . . . 135
7.1.6 Recoding string/numeric variables . . . 137
7.1.7 Creating new variables from old . . . 139
7.1.8 Changing variable types . . . 142
7.2 Merging Data Sets . . . 143
7.2.1 Binding . . . 143

7.2.2 The merge command 143
7.2.3 Duplicate values 146
7.2.4 Duplicate columns 147

III Analysis and Results 151

8 Statistical Modeling and *knitr* 153
8.1 Incorporating Analyses into the Markup 154
8.1.1 Full code chunks 154
8.1.2 Showing code & results inline 156
8.1.2.1 LaTeX 156
8.1.2.2 Markdown 158
8.1.3 Dynamically including non-R code in code chunks . . 159
8.2 Dynamically Including Modular Analysis Files 159
8.2.1 Source from a local file 160
8.2.2 Source from a non-secure URL (`http`) 162
8.2.3 Source from a secure URL (`https`) 162
8.3 Reproducibly Random: `set.seed` 163
8.4 Computationally Intensive Analyses 164

9 Showing Results with Tables 167
9.1 Basic *knitr* Syntax for Tables 168
9.2 Table Basics 168
9.2.1 Tables in LaTeX 169
9.2.2 Tables in Markdown/HTML 173
9.3 Creating Tables from Supported Class R Objects 177
9.3.1 `kable` for Markdown and LaTeX 177
9.3.2 *xtable* for LaTeX and HTML 178
9.3.3 *texreg* for LaTeX and HTML 181
9.3.4 Fitting Large Tables in LaTeX 184
9.3.5 *xtable* with non-supported class objects 185
9.3.6 Creating variable description documents with *xtable* . 188

10 Showing Results with Figures 191
10.1 Including Non-knitted Graphics 191
10.1.1 Including graphics in LaTeX 192
10.1.2 Including graphics in Markdown/HTML 194
10.2 Basic *knitr*/*rmarkdown* Figure Options 195
10.2.1 Chunk options 195
10.2.2 Global options 196
10.3 Knitting R's Default Graphics 197
10.4 Including *ggplot2* Graphics 200
10.4.1 Showing regression results with caterpillar plots 204
10.5 JavaScript Graphs with *googleVis* 209
10.5.1 JavaScript Graphs with *htmlwidgets*-based packages . 212

IV Presentation Documents 213

11 Presenting with *knitr*/LaTeX 215
11.1 The Basics . . . 215
11.1.1 Getting started with LaTeX editors . . . 216
11.1.2 Basic LaTeX command syntax . . . 216
11.1.3 The LaTeX preamble & body . . . 217
11.1.4 Headings . . . 220
11.1.5 Paragraphs & spacing . . . 221
11.1.6 Horizontal lines . . . 221
11.1.7 Text formatting . . . 221
11.1.8 Math . . . 223
11.1.9 Lists . . . 224
11.1.10 Footnotes . . . 225
11.1.11 Cross-references . . . 225
11.2 Bibliographies with BibTeX . . . 225
11.2.1 The *.bib* file . . . 225
11.2.2 Including citations in LaTeX documents . . . 227
11.2.3 Generating a BibTeX file of R package citations . . . 227
11.3 Presentations with LaTeX Beamer . . . 230
11.3.1 Beamer basics . . . 231
11.3.2 *knitr* with LaTeX slideshows . . . 234

12 Large *knitr*/LaTeX Documents: Theses, Books, and Batch Reports 237
12.1 Planning Large Documents . . . 237
12.2 Large Documents with Traditional LaTeX . . . 238
12.2.1 Inputting/including children . . . 239
12.2.2 Other common features of large documents . . . 240
12.3 *knitr* and Large Documents . . . 241
12.3.1 The parent document . . . 241
12.3.2 Knitting child documents . . . 242
12.4 Child Documents in a Different Markup Language . . . 243
12.5 Creating Batch Reports . . . 244

13 Presenting on the Web and Other Formats with R Markdown 249
13.1 The Basics . . . 249
13.1.1 Getting started with Markdown editors . . . 250
13.1.2 Preamble and document structure . . . 250
13.1.3 Headings . . . 252
13.1.4 Horizontal lines . . . 253
13.1.5 Paragraphs and new lines . . . 253
13.1.6 Italics and bold . . . 254
13.1.7 Links . . . 254

13.1.8 Special characters and font customization 254
13.1.9 Lists 254
13.1.10 Escape characters 255
13.1.11 Math with MathJax 255
13.2 Further Customizability with *rmarkdown* 256
13.2.1 More on *rmarkdown* Headers 256
13.2.2 CSS style files and Markdown 260
13.3 Slideshows with Markdown, *rmarkdown*, and HTML 261
13.3.1 HTML Slideshows with *rmarkdown* 262
13.3.2 LaTeX Beamer Slideshows with *rmarkdown* 264
13.3.3 Slideshows with Markdown and RStudio's R Presentations 265
13.4 Publishing HTML Documents Created by R Markdown . . . 268
13.4.1 Standalone HTML files 268
13.4.2 Hosting webpages with Dropbox 268
13.4.3 GitHub Pages 269
13.4.4 Further information on R Markdown 270

14 Conclusion **271**
14.1 Citing Reproducible Research 271
14.2 Licensing Your Reproducible Research 273
14.3 Sharing Your Code in Packages 273
14.4 Project Development: Public or Private? 274
14.5 Is it Possible to Completely Future-Proof Your Research? . . 275

Bibliography **277**

Index **285**