Christopher Gandrud

Reproducible Research with R and RStudio

Chapman & Hall/CRC Press

Expected Publication: August 2013

Stylistic Conventions **xiii**

and

Required R Packages **xv**

Additional Resources **xvii**
Chapter Examples xvii
Short Example Project xvii

List of Figures **xxii**

List of Tables **xxiii**

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? 7
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, 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 15
1.6.2 Reproduce this book 15
1.6.3 Contents overview 16

2 Getting Started with Reproducible Research **17**
2.1 The Big Picture: A workflow for reproducible research 17
2.1.1 Reproducible Theory 18
2.2 Practical tips for reproducible research 20
2.2.1 Document everything! 20
2.2.2 Everything is a (text) file 21
2.2.3 All files should be human readable 22
2.2.4 Explicitly tie your files together 24
2.2.5 Have a plan to organize, store, & make your files available 25

3 Getting Started with R, RStudio, and knitr **27**
3.1 Using R: the basics . . . 27
3.1.1 Objects . . . 28
3.1.2 Component Selection . . . 34
3.1.3 Subscripts . . . 36
3.1.4 Functions and commands . . . 37
3.1.5 Arguments . . . 38
3.1.6 The workspace & history . . . 39
3.1.7 Global R options . . . 41
3.1.8 Installing new packages and loading commands . . . 42
3.2 Using RStudio . . . 43
3.3 Using knitr: the basics . . . 44
3.3.1 What *knitr* does . . . 45
3.3.2 File extensions . . . 45
3.3.3 Code Chunks . . . 47
3.3.4 Global chunk options . . . 49
3.3.5 knitr package options . . . 51
3.3.6 Hooks . . . 52
3.3.7 *knitr* & RStudio . . . 52
3.3.8 *knitr* & R . . . 55

4 Getting Started with File Management **59**
4.1 File paths & naming conventions . . . 60
4.1.1 Root directories . . . 60
4.1.2 Subdirectories & parent directories . . . 60
4.1.3 Spaces in directory & file names . . . 61
4.1.4 Working directories . . . 61
4.2 Organizing your research project . . . 63
4.3 Setting directories as RStudio Projects . . . 64
4.4 R file manipulation commands . . . 64
4.5 Unix-like shell commands for file management . . . 67
4.6 File navigation in RStudio . . . 72

II Data Gathering and Storage **73**

5 Storing, Collaborating, Accessing Files, Versioning **75**
5.1 Saving data in reproducible formats . . . 76
5.2 Storing your files in the cloud: Dropbox . . . 77
5.2.1 Storage . . . 78
5.2.2 Accessing Data . . . 78
5.2.3 Collaboration . . . 79
5.2.4 Version control . . . 80
5.3 Storing your files in the cloud: GitHub . . . 80
5.3.1 Setting up GitHub: Basic . . . 83
5.3.2 Version Control with Git . . . 83

5.3.3 Remote Storage on GitHub . . . 91
5.3.4 Accessing on GitHub . . . 93
5.3.4.1 Collaboration with GitHub . . . 93
5.3.5 Summing up the GitHub workflow . . . 94
5.4 RStudio & GitHub . . . 94
5.4.1 Setting Up Git/GitHub with Projects . . . 95
5.4.2 Using Git in RStudio projects . . . 96

6 Gathering Data with R **99**
6.1 Organize your data gathering: makefiles . . . 99
6.1.1 R Make-like files . . . 100
6.1.2 GNU Make . . . 101
6.1.2.1 Example Makefile . . . 102
6.1.2.2 Makefiles and RStudio Projects . . . 106
6.1.2.3 Other information about Makefiles . . . 107
6.2 Importing locally stored data sets . . . 107
6.3 Importing data sets from the internet . . . 108
6.3.1 Data from non-secure (`http`) URLs . . . 108
6.3.2 Data from secure (`https`) URLs . . . 109
6.3.3 Compressed data stored online . . . 110
6.3.4 Data APIs & feeds . . . 112
6.4 Advanced Automatic Data Gathering: web scraping . . . 114

7 Preparing Data for Analysis **117**
7.1 Cleaning data for merging . . . 117
7.1.1 Get a handle on your data . . . 117
7.1.2 Reshaping Data . . . 119
7.1.3 Renaming variables . . . 122
7.1.4 Ordering data . . . 123
7.1.5 Subsetting data . . . 123
7.1.6 Recoding string/numeric variables . . . 125
7.1.7 Creating new variables from old . . . 127
7.1.8 Changing variables types . . . 130
7.2 Merging data sets . . . 131
7.2.1 Binding . . . 131
7.2.2 The merge command . . . 131
7.2.3 Duplicate values . . . 134
7.2.4 Duplicate columns . . . 135

III Analysis and Results **139**

8 Statistical Modelling and knitr **141**
8.1 Incorporating analyses into the markup . . . 142
8.1.1 Full code chunks . . . 142
8.1.2 Showing code & results inline . . . 144
8.1.2.1 LaTeX . . . 144

8.1.2.2 Markdown . . . 146
8.1.3 Dynamically including non-R code in code chunks . . 146
8.2 Dynamically including modular analysis files . . . 147
8.2.1 Source from a local file . . . 148
8.2.2 Source from a non-secure URL (`http`) . . . 149
8.2.3 Source from a secure URL (`https`) . . . 150
8.3 Reproducibly Random: `set.seed` . . . 151
8.4 Computationally intensive analyses . . . 152

9 Showing Results with Tables **155**
9.0.1 Basic *knitr* syntax for tables . . . 156
9.1 Table Basics . . . 156
9.1.1 Tables in LaTeX . . . 156
9.1.2 Tables in Markdown/HTML . . . 161
9.2 Creating tables from R objects . . . 165
9.2.1 *xtable* & *apsrtable* basics with supported class objects 165
9.2.1.1 *apsrtable* for LaTeX . . . 168
9.2.2 *xtable* with non-supported class objects . . . 171
9.2.3 Creating variable description documents with *xtable* . 173

10 Showing Results with Figures **177**
10.1 Including non-knitted graphics . . . 177
10.1.1 Including graphics in LaTeX . . . 178
10.1.2 Including graphics in Markdown/HTML . . . 180
10.2 Basic *knitr* figure options . . . 181
10.2.1 Chunk options . . . 181
10.2.2 Global options . . . 183
10.3 Knitting R's default graphics . . . 183
10.4 Including *ggplot2* graphics . . . 186
10.4.1 Showing Regression Results with Caterpillar Plots . . 190
10.5 JavaScript graphs with *googleVis* . . . 193

IV Presentation Documents **199**

11 Presenting with LaTeX **201**
11.1 The Basics . . . 201
11.1.1 Getting Started with LaTeX Editors . . . 201
11.1.2 Basic LaTeX command syntax . . . 202
11.1.3 The LaTeX preamble & body . . . 203
11.1.4 Headings . . . 206
11.1.5 Paragraphs & spacing . . . 206
11.1.6 Horizontal lines . . . 207
11.1.7 Text formatting . . . 207
11.1.8 Math . . . 208
11.1.9 Lists . . . 209
11.1.10 Footnotes . . . 210

11.1.11 Cross-references 210
11.2 Bibliographies with BibTeX 210
11.2.1 The *.bib* file 211
11.2.2 Including citations in a LaTeX document 212
11.2.3 Generating a BibTeX file of R packages 213
11.3 Presentations with LaTeX Beamer 216
11.3.1 Beamer basics 216
11.3.2 *knitr* with LaTeX slideshows 219

12 Large LaTeX Documents: Theses, Books, & Batch Reports 221
12.1 Planning large documents 221
12.2 Large documents with traditional LaTeX 222
12.2.1 Inputting/including children 223
12.2.2 Other common features of large documents 224
12.3 *knitr* and large documents 225
12.3.1 The parent document 225
12.3.2 Knitting child documents 226
12.4 Child documents in a different markup language 227
12.5 Creating batch reports 228

13 Presenting on the Web with Markdown 235
13.1 The Basics 235
13.1.1 Getting Started with Markdown Editors 236
13.1.2 Preamble and document structure 236
13.1.3 Headers 239
13.1.4 Horizontal Lines 239
13.1.5 Paragraphs and new lines 239
13.1.6 Italics and bold 240
13.1.7 Links 240
13.1.8 Special characters and font customization 240
13.1.9 Lists 240
13.1.10 Escape characters 241
13.1.11 Math with MathJax 241
13.2 Markdown with Pandoc and Custom CSS 242
13.2.1 Pandoc 242
13.2.2 CSS style files and Markdown 245
13.3 Presentations with `slidify` 247
13.4 Publishing Markdown Documents 254
13.4.1 Stand alone HTML files 254
13.4.2 Hosting webpages with Dropbox 254
13.4.3 GitHub Pages 255

14 Conclusion **257**
14.1 Citing reproducible research 257
14.2 Licensing your reproducible research 259
14.3 Sharing your code in packages 259
14.4 Project development: public or private? 260
14.5 Is it possible to completely future proof your research? . . . 261

Bibliography **263**