*Christopher Gandrud*

# *Reproducible Research with R and RStudio*

***Chapman & Hall/CRC Press***

***Expected Publication: January 2014***

**I Getting Started** **1**

**1 Introducing Reproducible Research** **3**

*and*

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 . . . 9
1.5 Why use R, knitr, and RStudio for reproducible research? . 10
1.5.1 Installing the Main Software . . . 12
1.6 Book overview . . . 13
1.6.1 How to read this book . . . 14
1.6.2 Reproduce this book . . . 15
1.6.3 Contents overview . . . 15

**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 . . . 26

**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 . . . 35
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 . . . 41

3.2 Using RStudio . . . 42
3.3 Using knitr: the basics . . . 44
3.3.1 What *knitr* does . . . 44
3.3.2 File extensions . . . 45
3.3.3 Code Chunks . . . 45
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 57**
4.1 File paths & naming conventions . . . 58
4.1.1 Root directories . . . 58
4.1.2 Subdirectories & parent directories . . . 58
4.1.3 Spaces in directory & file names . . . 59
4.1.4 Working directories . . . 59
4.2 Organizing your research project . . . 59
4.3 Setting directories as RStudio Projects . . . 61
4.4 R file manipulation commands . . . 62
4.5 Unix-like shell commands for file management . . . 65
4.6 File navigation in RStudio . . . 69

**II Data Gathering and Storage 71**

**5 Storing, Collaborating, Accessing Files, Versioning 73**
5.1 Saving data in reproducible formats . . . 74
5.2 Storing your files in the cloud . . . 75
5.2.1 Dropbox . . . 76
5.2.2 Storage . . . 76
5.2.2.1 Accessing Data . . . 76
5.2.3 Collaboration . . . 77
5.2.3.1 Version control . . . 78
5.2.4 GitHub . . . 78
5.2.4.1 Setting up GitHub: Basic . . . 80
5.2.4.2 Version Control with Git . . . 81
5.2.4.3 Remote Storage on GitHub . . . 87
5.2.4.4 Accessing on GitHub . . . 89
5.2.4.5 Collaboration with GitHub . . . 89
5.2.5 Summing up the GitHub workflow . . . 90
5.3 RStudio & GitHub . . . 90
5.3.1 Set Up . . . 90
5.3.2 Using Git in RStudio projects . . . 92

**6 Gathering Data with R** **95**
6.1 Organize your data gathering: make files . . . . . 95
6.1.1 R Make-like files . . . . . 96
6.1.2 GNU Make . . . . . 97
6.1.2.1 Example Makefile . . . . . 98
6.1.2.2 Makefiles and RStudio Projects . . . . . 102
6.1.2.3 Other information about Makefiles . . . . . 102
6.2 Importing locally stored data sets . . . . . 103
6.2.1 Importing a single locally stored file . . . . . 103
6.3 Importing data sets from the internet . . . . . 105
6.3.1 Data from non-secure (`http`) URLs . . . . . 105
6.3.2 Data from secure (`https`) URLs . . . . . 105
6.3.3 Compressed data stored online . . . . . 107
6.3.4 Data APIs & feeds . . . . . 108
6.4 Advanced Automatic Data Gathering: web scraping . . . . . 111

**7 Preparing Data for Analysis** **113**
7.1 Cleaning data for merging . . . . . 113
7.1.1 Get a handle on your data . . . . . 113
7.1.2 Reshaping Data . . . . . 114
7.1.3 Renaming variables . . . . . 117
7.1.4 Ordering data . . . . . 118
7.1.5 Subsetting data . . . . . 119
7.1.6 Recoding string/numeric variables . . . . . 121
7.1.7 Creating new variables from old . . . . . 122
7.1.8 Changing variables types . . . . . 124
7.2 Merging data sets . . . . . 125
7.2.1 Binding . . . . . 125
7.2.2 The merge command . . . . . 125
7.2.3 Duplicate values . . . . . 127
7.2.4 Duplicate columns . . . . . 128

**III Analysis and Results** **131**

**8 Statistical Modelling and knitr** **133**
8.1 Incorporating analyses into the markup . . . . . 134
8.1.1 Full code chunks . . . . . 134
8.1.2 Showing code & results inline . . . . . 136
8.1.2.1 LaTeX . . . . . 136
8.1.2.2 Markdown . . . . . 138
8.1.3 Dynamically including non-R code in code chunks . . 138
8.2 Dynamically including modular analysis files . . . . . 139
8.2.1 Source from a local file . . . . . 140
8.2.2 Source from a non-secure URL (`http`) . . . . . 141
8.2.3 Source from a secure URL (`https`) . . . . . 142

8.3 Computationally Intensive Analyses . . . 143

**9 Showing Results with Tables** **145**
9.0.1 Basic *knitr* syntax for tables . . . 146
9.1 Table Basics . . . 146
9.1.1 Tables in LaTeX . . . 146
9.1.2 Tables in Markdown/HTML . . . 151
9.2 Creating tables from R objects . . . 155
9.2.1 *xtable* & *apsrtable* basics with supported class objects 155
9.2.1.1 *xtable* for LaTeX . . . 156
9.2.1.2 *xtable* for Markdown/HTML . . . 157
9.2.1.3 *apsrtable* for LaTeX . . . 158
9.2.2 *xtable* with non-supported class objects . . . 161
9.2.3 Creating variable description documents with *xtable* . 164

**10 Showing Results with Figures** **167**
10.1 Including graphics non-knitted graphics . . . 167
10.1.1 Including graphics in LaTeX . . . 168
10.1.2 Including graphics in Markdown/HTML . . . 170
10.2 Basic knitr figure options . . . 171
10.2.1 Chunk options . . . 171
10.2.2 Global options . . . 172
10.3 Knitting R's default graphics . . . 173
10.4 Including ggplot2 graphs . . . 176
10.4.1 Showing Regression Results with Caterpillar Plots . . 180
10.5 JavaScript graphs with *googleVis* . . . 184

**IV Presentation Documents** **189**

**11 Presenting with LaTeX** **191**
11.1 The Basics . . . 191
11.1.1 Getting Started with LaTeX Editors . . . 191
11.1.2 Basic command syntax . . . 192
11.1.3 The preamble & the body . . . 192
11.1.4 Headings . . . 196
11.1.5 Paragraphs & spacing . . . 196
11.1.6 Horizontal lines . . . 196
11.1.7 Text formatting & math . . . 197
11.1.8 Lists . . . 199
11.1.9 Footnotes . . . 200
11.1.10 Cross-References . . . 200
11.2 Bibliographies with BibTeX . . . 200
11.2.1 The *.bib* file . . . 200
11.2.2 Including citations in a LaTeX document . . . 202
11.2.3 Generating a BibTeX file of R packages . . . 202
11.3 Presentations with LaTeX Beamer . . . 205

11.3.1 Beamer basics . . . . . 206
11.3.2 knitr with LaTeX slideshows . . . . . 209

**12 Large LaTeX Documents: Theses, Books, & Batch Reports 211**
12.1 Planning large documents . . . . . 212
12.1.1 Planning theses and books . . . . . 212
12.2 Large documents with traditional LaTeX . . . . . 213
12.2.1 Other common features of large documents . . . . . 214
12.3 *knitr* and large documents . . . . . 215
12.3.1 The parent document . . . . . 215
12.3.2 Knitting child documents . . . . . 216
12.4 Child documents in a different markup language . . . . . 217
12.5 Creating Batch Reports . . . . . 218

**13 Presenting on the Web and More with Markdown 223**
13.1 The Basics . . . . . 223
13.1.1 Getting Started with Markdown Editors . . . . . 224
13.1.2 Preamble and document structure . . . . . 224
13.1.3 Headers . . . . . 227
13.1.4 Horizontal Lines . . . . . 227
13.1.5 Paragraphs and new lines . . . . . 227
13.1.6 Italics and bold . . . . . 228
13.1.7 Links . . . . . 228
13.1.8 Special characters and font customization . . . . . 228
13.1.9 Lists . . . . . 228
13.1.10 Escape characters . . . . . 229
13.1.11 Math with MathJax . . . . . 229
13.2 Markdown with Pandoc and Custom CSS . . . . . 230
13.2.1 Pandoc . . . . . 230
13.2.2 CSS style files and Markdown . . . . . 233
13.3 Presentations with `Slidify` . . . . . 236
13.4 Publishing Markdown Documents . . . . . 241
13.4.1 Stand alone HTML file . . . . . 241
13.4.2 Hosting webpages with Dropbox . . . . . 241
13.5 GitHub Pages . . . . . 242

**14 Conclusion 245**
14.1 Citing reproducible research . . . . . 245
14.2 Licensing your reproducible research . . . . . 246
14.3 Is it possible to completely future proof your research? . . . 247