

Christopher Gandrud

Reproducible Research with R and RStudio

Contents

Preface	vii
Stylistic Conventions	xv
Required R Packages	xvii
Additional Resources	xix
Chapter Examples	xix
Short Example Project	xix
List of Figures	xxiv
List of Tables	xxv
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, <i>knitr</i> , 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	16
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	22
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 avail- able	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 <i>knitr</i> : the Basics	44
3.3.1	What <i>knitr</i> does	45
3.3.2	File extensions	45
3.3.3	Code chunks	47
3.3.4	Global chunk options	49
3.3.5	<i>knitr</i> package options	51
3.3.6	Hooks	52
3.3.7	<i>knitr</i> & RStudio	52
3.3.8	<i>knitr</i> & 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	80
5.2.4	Version control	80
5.3	Storing Your Files in the Cloud: GitHub	81
5.3.1	Setting up GitHub: basic	83
5.3.2	Version control with Git	84
5.3.3	Remote storage on GitHub	91
5.3.4	Accessing on GitHub	94
5.3.4.1	Collaboration with GitHub	96
5.3.5	Summing up the GitHub workflow	96
5.4	RStudio & GitHub	97
5.4.1	Setting up Git/GitHub with Projects	98
5.4.2	Using Git in RStudio Projects	99
6	Gathering Data with R	101
6.1	Organize Your Data Gathering: Makefiles	101
6.1.1	R Make-like files	102
6.1.2	GNU Make	103
6.1.2.1	Example makefile	104
6.1.2.2	Makefiles and RStudio Projects	108
6.1.2.3	Other information about makefiles	109
6.2	Importing Locally Stored Data Sets	109
6.3	Importing Data Sets from the Internet	110
6.3.1	Data from non-secure (http) URLs	111
6.3.2	Data from secure (https) URLs	111
6.3.3	Compressed data stored online	114
6.3.4	Data APIs & feeds	115
6.4	Advanced Automatic Data Gathering: Web Scraping	117
7	Preparing Data for Analysis	121
7.1	Cleaning Data for Merging	121
7.1.1	Get a handle on your data	121
7.1.2	Reshaping data	123
7.1.3	Renaming variables	126
7.1.4	Ordering data	126
7.1.5	Subsetting data	127
7.1.6	Recoding string/numeric variables	129
7.1.7	Creating new variables from old	131
7.1.8	Changing variable types	134
7.2	Merging Data Sets	135
7.2.1	Binding	135

7.2.2	The merge command	135
7.2.3	Duplicate values	138
7.2.4	Duplicate columns	139
III	Analysis and Results	143
8	Statistical Modelling and knitr	145
8.1	Incorporating Analyses into the Markup	146
8.1.1	Full code chunks	146
8.1.2	Showing code & results inline	148
8.1.2.1	LaTeX	148
8.1.2.2	Markdown	150
8.1.3	Dynamically including non-R code in code chunks . .	150
8.2	Dynamically Including Modular Analysis Files	152
8.2.1	Source from a local file	152
8.2.2	Source from a non-secure URL (http)	153
8.2.3	Source from a secure URL (https)	154
8.3	Reproducibly Random: set.seed	155
8.4	Computationally Intensive Analyses	157
9	Showing Results with Tables	159
9.1	Basic <i>knitr</i> Syntax for Tables	160
9.2	Table Basics	160
9.2.1	Tables in LaTeX	161
9.2.2	Tables in Markdown/HTML	165
9.3	Creating Tables from R Objects	169
9.3.1	<i>xtable</i> & <i>apsrtable</i> basics with supported class objects	169
9.3.1.1	<i>apsrtable</i> for LaTeX	172
9.3.2	<i>xtable</i> with non-supported class objects	175
9.3.3	Creating variable description documents with <i>xtable</i> .	177
10	Showing Results with Figures	181
10.1	Including Non-knitted Graphics	181
10.1.1	Including graphics in LaTeX	182
10.1.2	Including graphics in Markdown/HTML	184
10.2	Basic <i>knitr</i> Figure Options	185
10.2.1	Chunk options	185
10.2.2	Global options	187
10.3	Knitting R's Default Graphics	187
10.4	Including <i>ggplot2</i> Graphics	192
10.4.1	Showing regression results with caterpillar plots . . .	195
10.5	JavaScript Graphs with <i>googleVis</i>	198
IV	Presentation Documents	203

11 Presenting with LaTeX	205
11.1 The Basics	205
11.1.1 Getting started with LaTeX editors	205
11.1.2 Basic LaTeX command syntax	206
11.1.3 The LaTeX preamble & body	207
11.1.4 Headings	210
11.1.5 Paragraphs & spacing	210
11.1.6 Horizontal lines	211
11.1.7 Text formatting	211
11.1.8 Math	212
11.1.9 Lists	213
11.1.10 Footnotes	214
11.1.11 Cross-references	214
11.2 Bibliographies with BibTeX	215
11.2.1 The <i>.bib</i> file	215
11.2.2 Including citations in LaTeX documents	216
11.2.3 Generating a BibTeX file of R package citations	217
11.3 Presentations with LaTeX Beamer	220
11.3.1 Beamer basics	220
11.3.2 <i>knitr</i> with LaTeX slideshows	223
12 Large LaTeX Documents: Theses, Books, & Batch Reports	225
12.1 Planning Large Documents	225
12.2 Large Documents with Traditional LaTeX	226
12.2.1 Inputting/including children	227
12.2.2 Other common features of large documents	228
12.3 <i>knitr</i> and Large Documents	229
12.3.1 The parent document	229
12.3.2 Knitting child documents	230
12.4 Child Documents in a Different Markup Language	231
12.5 Creating Batch Reports	232
13 Presenting on the Web with Markdown	239
13.1 The Basics	239
13.1.1 Getting started with Markdown editors	240
13.1.2 Preamble and document structure	240
13.1.3 Headers	243
13.1.4 Horizontal lines	243
13.1.5 Paragraphs and new lines	243
13.1.6 Italics and bold	244
13.1.7 Links	244
13.1.8 Special characters and font customization	244
13.1.9 Lists	244
13.1.10 Escape characters	245
13.1.11 Math with MathJax	245

13.2	Markdown with Pandoc and Custom CSS	246
13.2.1	Pandoc	246
13.2.2	CSS style files and Markdown	250
13.2.3	<i>knitr</i> 's pandoc command	251
13.3	Slideshows with Markdown, <i>knitr</i> , and HTML	254
13.3.1	Slideshows with Markdown, <i>knitr</i> , and RStudio's R Pre- sentations	254
13.3.2	Slideshows with Markdown, <i>knitr</i> , and slidify	256
13.4	Publishing Markdown Documents	262
13.4.1	Stand alone HTML files	262
13.4.2	Hosting webpages with Dropbox	263
13.4.3	GitHub Pages	263
14	Conclusion	265
14.1	Citing Reproducible Research	265
14.2	Licensing Your Reproducible Research	267
14.3	Sharing Your Code in Packages	267
14.4	Project Development: Public or Private?	268
14.5	Is it Possible to Completely Future Proof Your Research? . .	269
	Bibliography	271
	Index	279