Reproducible research with RStudio and `knitr`

Ben Bolker

McMaster University, Mathematics & Statistics and Biology

3 February 2016

Outline

Reproducible research

- Science requires reproducibility
- Computational methods: harder to reproduce than math/analytic methods
  (but easier than non-computational science)
- Maintain integrity
- Disseminate knowledge
- Maintain personal sanity

Literate programming (**?**)

- ancestor of RR
- similar tools (WEB/weave/tangle), but different scope
- targets code as a document with interwoven discussion
- some notes on the LP-RR ecosystem

TeX/LaTeX

- ?/?
- mathematical (and general-purpose) typesetting system
- *pro*: beautiful, widely used, cross-platform, customizable, stable
- *con*: old-fashioned, arcane
- troubleshooting: TeX Stack Exchange

R

- Gentleman and Ihaka, 1990s
- statistical programming language/data analysis environment
- *pro*: powerful, widely used (3000+ packages), cross-platform, customizable
- *con*: relatively slow; organic/inconsistent

Sweave (**?**)

- literate programming tool, allowing LaTeX chunks in R
- highlighted code chunks (`echo=TRUE`)
- automatically generated figures, optionally in a figure environment
- *pro*: super-convenient, once you get used to it
- *con*: one more software layer;
  less suitable for *big* projects/code

knitr (?)

- updated version of `Sweave`
- just plain better

RStudio (Allaire et al.)

- full-featured front-end for R
- one-button front end for `knitr` (“Compile PDF”)
- *pro*: beginner-friendly; cross-platform;
  zoomable graphics, code highlighting, tab completion,
  environment listing, etc.
- *con*: R-centric; restriction to built-in editor;
  one more software layer

Outline

Getting started

- bookmark the knitr web page, especially the options page
- switch RStudio to compile documents with `knitr` (`Tools/Global options/Sweave/Weave Rnw files using ...`)
- make sure LaTeX is installed/working and the `rmarkdown` package is installed (Packages menu or `install.packages(c("rmarkdown"))`; also install `tikzDevice` package
- build this document, or use (File/New File/R Sweave) to generate an empty template (need to add *something* to it); RStudio recognizes `.Rnw` extension
- code chunks start with `<<>>=` and end with `@`

Troubleshooting

- use `knitr::knit` or `rmarkdown::render` from the console
- R code failing? Run it interactively in the console, or `purl()` to pull the code into a separate file
- in the console:
  `knit2pdf("myfile.Rnw")` = pushing the button
- step by step: `knit("myfile.Rnw")` +
  externally `pdflatex myfile`
- **always** name your code chunks!
- pressing the button compiles PDF in a clean environment
- MikTeX may hang when LaTeXneeds to download a new package
  (e.g. first time using TikZ)

Code options

Set per chunk, e.g. `<<mychunk,echo=TRUE,eval=FALSE>>=`
or globally via `opts_chunk$set(...)`

- `eval`: evaluate?
- `echo`: show code?
- `warning`/`message`/`error`: show/stop? (`knitr::knit` does *not* stop on errors by default, but `rmarkdown::render` does)
- `results`: `"markup"` is default, alternatives `"hide"` or `"asis"`
- `tidy`: reformat code?
- `cache`: cache results?

More code issues

- if you're using beamer, need to use `\begin{frame}[fragile]` to show code (i.e., `echo=TRUE`)
- code in chunks must be complete/syntactically correct: no fragments allowed;
  can't (e.g.) separate parts of a `for` loop, even if `eval=FALSE`
- in-line expressions via `\Sexpr{}` (don't forget to round numeric values)

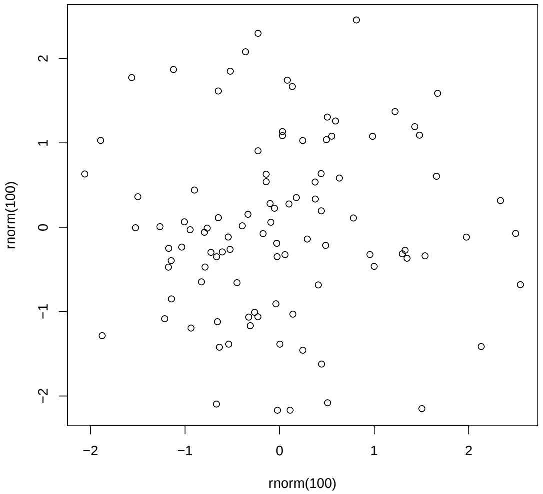
Code example (using `fragile` option)

```r
library(nlme)
## comments get formatted nicely too
fm1 <- lme(distance ~ age, data = Orthodont)
```

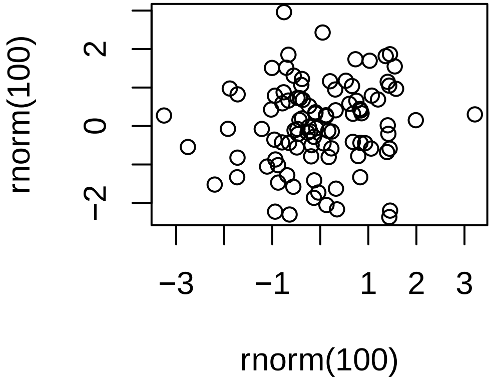
Graphics basics

- Graphics are automatically run (stored in `figures` directory)
- `fig.width`, `fig.height` control the size/aspect ratio of the *plot window* (in inches!)
- `out.width` controls the size of the printed plot (in LaTeX units, e.g. `"0.7\\textwidth"`) (note double backslashes)
- `dev` controls device: default is `"pdf"`, may want `"png"` for huge figures or `"tikz"` for LaTeX fonts and symbols (small figures only!)
- `fig.cap` generates a figure caption and puts the plot in a `figure` environment (need math mode where appropriate, and double backslashes!); use `fig.scap` if you have a super-long caption; can use `fig.pos` to force figure position

Graphics example: basic

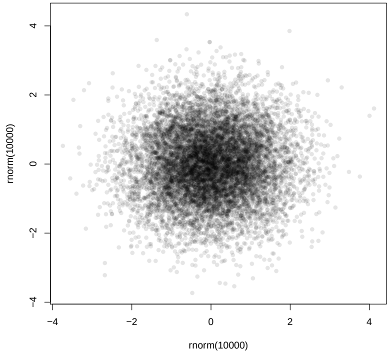


Graphics example: `fig.width=3,fig.height=3`



Graphics example: `dev="png"`

```
plot(rnorm(1e4),rnorm(1e4),
     pch=16,
     col=adjustcolor("black",alpha=0.1))
```



Graphics example: `dev="tikz"`

```r
plot(rnorm(100),rnorm(100),
     xlab="${\\cal R}_0$",ylab="$\\sqrt{\\xi^\\alpha}$")
```

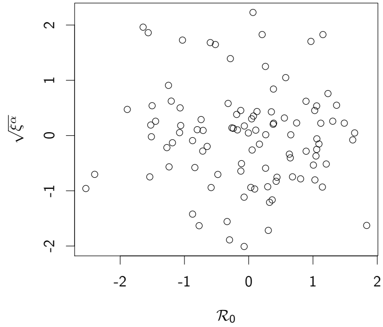


Table basics

- any old tables: `knitr::kable`, `Hmisc::latex`, `xtable`
- regression output: `stargazer`, `rockchalk`
- tables for markdown/HTML output: `pander`

knitr::kable (results="asis")

```r
library("knitr")
df <- data.frame(A = c(1.00123, 33.1, 6),
                 B = c(111111, 3333333, 3123.233))
kable(df)
```

A	B
1.00123	111111.000
33.10000	3333333.000
6.00000	3123.233

xtable (results="asis")

```r
library("xtable")
df <- data.frame(A = c(1.00123, 33.1, 6),
                 B = c(111111, 3333333, 3123.233))
xtable(df)
```

	A	B
1	1.00	111111.00
2	33.10	3333333.00
3	6.00	3123.23

stargazer (results="asis")

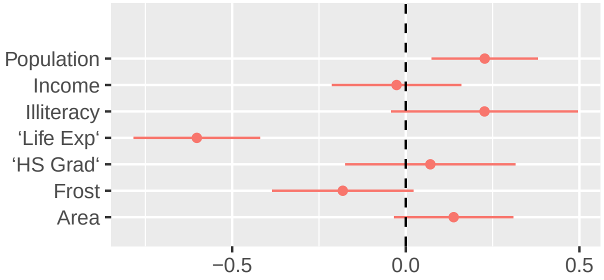
```
library("stargazer")
m2 <- lm(Murder~Illiteracy+Income+Population,
         data=as.data.frame(state.x77))
stargazer(m2,float=FALSE)
```

	Dependent variable:
	Murder
Illiteracy	4.111^{***}
	(0.671)
Income	0.0001
	(0.001)
Population	0.0002^{**}
	(0.0001)
Constant	1.340
	(3.369)
Observations	50
R^2	0.567
Adjusted R^2	0.539
Residual Std. Error	2.507 (df = 46)
F Statistic	20.072^{***} (df = 3; 46)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

eschew tables? (?)

dotwhisker and broom packages

```
library("dotwhisker")
m2 <- lm(Murder~.,
         data=as.data.frame(scale(state.x77)))
dwplot(m2)+geom_vline(xintercept=0,lty=2)
```



Outline

Advanced tricks

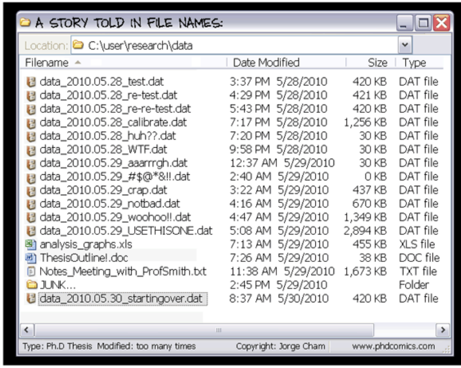
- other programming languages (e.g. Python)
- other markup languages (e.g. markdown)
- other output formats (e.g. docx, HTML)
- other ways of documenting/disseminating results: commented R code (`spin()`); R packages/vignettes; `roxygen2` package
- large/batch jobs: caching gets tricky, use Makefiles instead?
- figure tricks: 3D (`rgl`) plots, animation ...

Workflow tips

- batch vs interactive processing
- DRY (don't repeat yourself) – functions should be in one place, re-usable/re-used
- organic process:
  - experiments in console window
  - rough code in main script
  - code $\rightarrow$ functions in main script
  - functions $\rightarrow$ separate file
  - functions $\rightarrow$ package
  - batch runs

Version control and collaboration

- Dropbox
- Github, Bitbucket
- Overleaf



A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified		Size	Type
data_2010.05.28_test.dat	3:37 PM	5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM	5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM	5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM	5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM	5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM	5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM	5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM	5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM	5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM	5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM	5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM	5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM	5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM	5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM	5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM	5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM	5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

PhD Comics

Further resources

- knitr web page, including the demos and showcase . . .
- StackOverflow
- my examples on Rpubs
- reproducible research task view
- knitr book on Amazon

References