Putting Reproducibility into Practice: Workflows and Case Studies

Putting Reproducibility into Practice: Workflows and Case Studies

Rutgers Workshop on Reproducibility in Experimental and Computational Science Research

Ryan Womack

Data Librarian, Rutgers University, rwomack@rutgers.edu

October 10, 2016

Libraries

- I approach reproducibility as a *librarian*
- Libraries are about *enabling learning*, and building on other's knowledge work
- Libraries take the long view with respect to keeping and archiving information
- But we are constantly introducing new users to that information
- So issues like *usability* and *availability* are paramount

Data Librarian

- What is a *data librarian*?
- We acquire, locate, and help others use and reuse data
- So reproducibility concerns are a natural part of my daily work
- Librarians cheer the larger impacts on science, but others are more qualified to analyze them
- As librarians, we care about this from the other end
- We see the frustration when users cannot do what they want with data that is inaccessible or incomprehensible

Exemplars from the Social Sciences

- ICPSR has been in operation for over 50 years, with well-established archiving practices and data documentation via codebooks and metadata
- IPUMS is reformatting and making data compatible across many decades and different projects, to enable international comparisons of microdata
- Coming from the world in the social sciences where long-term is, if not always routine, at least well-established
- Disciplinary separation of practices is diminishing when similar computational techniques can be applied to physical sciences or digital humanities

Outline

- We will illustrate some practices in a few contexts
  - an individual researcher
  - a team or research group
  - ongoing, large-scale collaboration

The Data In Itself

Some basic practices:

- Keep raw data pristine and separate from any working data
- Document your variables and data collection
  - anything you yourself would forget when revisiting the project 3 years later in response to a query
  - that will be the same thing other users need too!
- Don't work in Excel [if you can] or other manual editing environment
  - you should write down all your steps if you are doing this
  - better to use code or an environment that will at least record your steps

DOI

DOI, the Digital Object identifier, is the great success story

- makes it easy to have a permanent reference and good citation practice
- usually associated with quality data repositories
- encapsulates a lot of good stuff
- moral: defined standards and centralized tools make adoption and use easy
- Treat your local data as if you were pulling it from a DOI, and you will be baking in reproducibility

Code/Data/Document Integration

We will discuss examples in R, but other programming environments support this as well (Mathematica notebooks, Python/Jupyter)

- originally implemented in LaTeX + Sweave
- can embed R code and run it as the document is generated
- Code that is tangled in with text can be extracted, and formatted documents can be woven from the literate program.
- always ensures that the latest data and results are actually incorporated
- helps to document and explain code in context (literate programming)
- PDF, document, and HTML formats are easy to obtain

knitr and Markdown

- Markdown + knitr has become a popular, lightweight replacement
- Simple syntax and implementation
- Integrated into RStudio
- Publish documents with one click at RPubs
- Can fall back on LaTeX/Sweave for more complex document formatting

A short Rmd file

— title: ”A short literate programming exercise” author: ”Ryan Womack” date: ”October 10, 2016” output: pdf_document —
“‘{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE) “‘
Read in the data
Let’s read in the data with the following commands:
“‘{r load} library(readxl)
download.file(”http://ryanwomack.com/data/PharmaDemo.xls”, ”mydata.xls”)
mydata<-read_excel(”mydata.xls”)
names(mydata)
attach(mydata)
“‘
Describe the Data
Then we will get some summary statistics on the Age and Weight variables:
“‘{r summary} summary(Age)
summary(Weight) “‘
Now plot the data:
“‘{r plot, echo=FALSE} library(ggplot2)
ggplot(mydata, aes(Weight, Age))+ geom_point() “‘
Regression
“‘{r regression} summary(lm(Age˜Weight))
ggplot(mydata, aes(Weight, Age))+ geom_point()+ stat_smooth() “‘
All done!

RStudio environment

RStudio

File Edit Code View Plots Session Build Debug Tools Help

```
30
31 Then we will get some summary statistics on the Age and Weight variables:
32
33 ```{r summary}
34 summary(Age)
35
36 summary(Weight)
37 ```
38
39 Now plot the data:
40
41 ```{r plot, echo=FALSE}
42 library(ggplot2)
43
44 ggplot(mydata, aes(Weight, Age))+ geom_point()
45 ```
46
47 ## Regression
48
49 ```{r regression}
50 summary(lm(Age~Weight))
51
52 ggplot(mydata, aes(Weight, Age))+ geom_point()+ stat_smooth()
53 ```
54
55 All done!
```

```
    Min      1Q  Median      3Q     Max
-24.626  -5.943   1.047   6.411  19.980

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 83.00577    2.89586  28.664  < 2e-16 ***
Weight      -0.18489    0.03108  -5.949 1.21e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.056 on 198 degrees of freedom
Multiple R-squared:  0.1516,    Adjusted R-squared:  0.1473
F-statistic: 35.39 on 1 and 198 DF,  p-value: 1.205e-08

> getwd()
[1] "/home/ryan/R"
> setwd("/home/ryan/womack/documents/ryan/work/2016/")
>
```

A short literate programming exercise

Ryan Womack

October 10, 2016

Read in the data

Let's read in the data with the following commands:

```
library(readxl)

download.file("http://ryanwomack.com/data/PharmaDemo.xls", "mydata.xls")

mydata<-read_excel("mydata.xls")

names(mydata)
```

```
##  [1] "Age"               "Gender"            "Weight"
##  [4] "IV_APAP"           "Epidural"          "Opi_N_T"
##  [7] "Average_Pain_Score" "Tot_Opi"          "Tramadol"
## [10] "TOT_LOS_H"         "Painkiller"
```

```
attach(mydata)
```

Describe the Data

Then we will get some summary statistics on the Age and Weight variables:

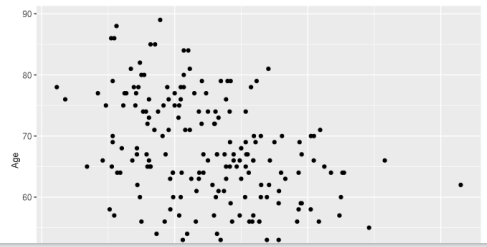
```
summary(Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    43.0    60.0    66.0    66.2    74.0    89.0
```

```
summary(Weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44.55   74.09   90.46   90.87  103.70  166.00
```

Now plot the data:



The computing environment

- Open source is an important enabler of reproducibility
- Anyone can grab copies of the software to execute
- And can get older versions if necessary for compatibility
- You can also record information about your computing environment (session.Info() in R)
- The checkpoint package automates this process in R

Other issues

- Dont save output. Where did it come from? This should be done in the code.
- Clean, well-formatted data (tidyr) and code (formatR) are a plus
- If using readme files approach, document everything

Other platforms

- Jupyter grew out of iPython
  - now over 40 languages supported
- Mathematica Notebooks, cloud support
- Cloud services making sharing much easier
- Becoming an expectation

Collaboration

- The same forces (cloud computing, shared platforms, standards) are making collaboration easier than ever
- Github, Bitbucket, and others enable easy collaboration on programming
  - with significant side benefits for reproducibility due to availability of code
- The Open Science Framework provides a more data-specific approach
- A key feature is that the same platform is used for private work and then public sharing
- Psychology reproducibility study uses OSF.
  - See the Science article for a start

XKCD

- http://xkcd.com/1742/

Enabling Collaboration

- Collaborative projects must/should agree on:
  - data practices and sharing platform
  - coding practices and sharing platform
- This is made easier by already existing platforms and practices

Yale ISPS

The Yale Institute for Social and Policy Studies is an example of a research group that enables reproducibility.

- Data and papers archived together onsite
- Handles (not DOIs) for data
- Code and documentation archived
- Code review for correct execution
- Good example of providing explanatory metadata for studies
- Possible because the Institute requires compliance as a condition of grants

Big Science

Multi-year, multi-institutional projects that may continue beyond original PIs *require* reproducibility.

- Many people will be coming on and off of the project over time
- Many unanticipated uses are anticipated ("known unknowns" or something like that)
- Collaboration and continuity must be consciously planned for
- Decisions should be made with more consideration for future use than current convenience
- But disciplinary expertise is building in these areas
- PDB, of course
- Also, https://www.teamsciencetoolkit.cancer.gov/

Big Science - Infrastructure

- In big science, the main node(s) are enablers of future reuse
- They provide basic infrastructure (OOI)
- But also provide a clearinghouse for other projects that link to and build on the central node
- One major future goal is to have more generic, all-purpose collaborative infrastructure
- Rutgers is developing a Virtual Data Collaboratory for this purpose
- Open infrastructure like Zenodo, Dataverse, OpenICSPR and the Open Science Data Cloud have been developed

Big Science - Practices

- Standards such as DOI and ORCID enable the broader community to coalesce around good practice
- Data repositories are developing standards
- Re3data.org is a directory of repositories
- The Data Seal of Approval is awarded to repositories using sound data practices
- ISO 16363 (Trusted Digital Repositories) is a more stringent standard
- One important step is to plan for what happens when the project winds down (expectedly or unexpectedly)
- What is the equivalent standard for computing and reproducibility?

Big Science - Genomic Data Sharing

- Massive investments, massive amounts of data
- Many repositories too (NIH GDS)
- Existing repositories are useful and aggregate many software tools
- But researchers want even larger pooled databases, especially for human genome
- Technical issues are complex, but the rights and permissions involved are equally complex
- How can data be federated for maximum discoverability?

Open Data -> Reproducible Research

- Increasing openness is a long-term trend
  - Internet, Government, Data, Software, Cloud Computing
- Pressure for Reproducible Research can only increase as these trends intensify
- Good news is...
  - Benefits are clear for society and for knowledge creation!
  - Tools to enable this are getting easier all the time!
  - Eventually it will be a standard we will take for granted!