Workflow

Paul E. Johnson[1] [2]

[1]Department of Political Science

[2]Center for Research Methods and Data Analysis, University of Kansas

2018

KU

Outline

KU

Outline

Replication is a Priority

- In CRDMA, we need to replicate projects
  - Across time: a worker can exactly reproduce something
  - Across people: another person can exactly reproduce something
- The General Failure of Research Reproducibility is a high-priority topic at the current time.
  - Center for Open Science "Reproducibility crisis"
  - Many possible causes, today we are focused on poor data management and non-replicable calculations.

KU

Avoid non-replicable point-and-click methods

- "Friends don't let Friends use Microsoft Excel"
  - J.D. Cryer: http://www.amstat.org/sections/srms/proceedings/y2001/proceed/00470.pdf
  - "Excel's Checkered Statistical Past" http://www.statisticalengineering.com/Weibull/excel.html

TexaSoft's

Statistics Classroom

Vol. 8, No 1, 2010

Should you use Excel to teach Statistics?

The following articles discuss pitfalls of using Microsoft Excel® for teaching college-level or AP statistics. Links to web pages are included when possible, along with pertinent quotes from the article.

"Excel 2007, like its predecessors, fails a standard set of intermediate-level accuracy tests in three areas: statistical distributions, random number generation, and estimation…. Persons who wish to conduct statistical analyses should use some other package." McCullough & Heiser, 2008

"The journal Computational Statistics & Data Analysis recently published an article concluding that [Excel] is inadequate for substantive statistical analysis" (2000) Berkeley Lab Computing Newsletter http://www.lbl.gov/ICSD/CIS/compnews/2000/June/05_journal.html

"Teaching statistics is a big challenge, teaching statistics with Excel is an even bigger challenge." AUSTRIAN JOURNAL OF STATISTICS Volume 37 (2008), Number 2, 195–206
Example of accuracy problems in Excel: http://www.cmh.edu/stats/ask/accuracy.asp

"We find that the accuracy of various statistical functions in Excel 2007 range from unacceptably bad to acceptable but significantly inferior in comparison to alternative implementations." A.T. Yalta, 2008

"Microsoft Excel spreadsheets have become somewhat of a standard for data storage, at least for smaller data sets. This, along with the program often being packaged with new computers, naturally encourages its use for statistical analyses. This is unfortunate, since Excel is most decidedly **not** a statistical package." http://pages.stern.nyu.edu/~jsimonof/classes/1305/pdf/excelreg.pdf

What Excel is Not Good For: Analysis

- Paul Krugman, April 18, 2013, "The Excel Depression" http://www.nytimes.com/2013/04/19/opinion/krugman-the-excel-depression.html
- Mike Konczal, April 16, 2013, "Researchers Finally Replicated Reinhart-Rogoff, and There are Serious Problems", The Roosevelt Institute
- "Reinhart, Rogoff, and the Excel Error That Changed History"
- These concern
  - Reinhart, Carmen and Rogoff, Kenneth (2010) "Growth in a time of debt", NBER Working Paper Series #15639.

Outline

Write out everything

- The R (R Core Team, 2017) console is a place to 'doodle', find out what works.
- All calculations, from top-to-bottom, should be saved in a script file.
- Ideally, that would be able to repeat everything
  - data import
  - recode
  - analyze
  - create figures
  - and tables.

You need an editor that won't drive you crazy

- Mac, Linux, & Windows {Emacs, Rstudio}
- Windows only {Notepad++ with NPPTOR}
- Mac only {R.app}
- The R FAQ recommends Emacs, I've always respected that judgment
  - My Emacs Notes: "Emacs has no learning curve."
- Can't deny that beginners do better with RStudio☺
  - can't understand why non-beginners keep using it☹

Write Code in an R file

- The same things you would have typed in a terminal can be typed into a file
- An R-aware editor has a "send-this-line" or "send-this-region" shortcut key.
- Formatting: Editor should help you respect indentation and spaces. See my essay: Rstyle a vignette in rockchalk
- Use { } to separate nested sections of commands.
- # is the comment symbol. It is customary in the Emacs-using community to use two ## to represent a comment that is to be intended at the current level.
- Examples: `http://pj.freefaculty.org/R/WorkingExamples`

KU

Outline

Stages in (Almost All) Projects

1. Download "raw" data
2. Import data
3. Inspect data (descriptive and graphic)
4. Recode & Reorganize data
5. Analyze data
6. Generate Tables and Plots

- In a perfect world, these are all
  - "automated" and
  - more-or-less easily repeated.

Outline

Project Folders

Subdirectories in a CRMDA Project folder

data:	Data in "fresh" "pristine" "unaltered state".
R:	R files here
workingdata:	"Recoded", "Cleaned", "Subsetted"
output:	Graphs and tables
lit:	reading material
trash:	don't use the system-wide recycle bin

Relative Paths

- Suppose we are in R folder.
- Read data from "`../data`"
- Write/read working data from "`../workingdata`"
- Write output into "`../output`"

kutils:initProject will automate

```
library(kutils)
## Creates directories in current working
    directory
initProject()
```

- I'll test in the working directory "/tmp/practice"

```
Creating: data
Creating: workingdata
Creating: output
Creating: tmp
Creating: lit
Creating: writeup
Creating: R
Initialized empty shared Git repository in /tmp/practice/.git/
[master (root-commit) 5d2504b] Initialized project in /tmp/practice
 2 files changed, 5 insertions(+)  create mode 100644 ChangeLog
 create mode 100644 README.md
Please consider creating a remote repository to which this repo should
    be linked
[1] "/tmp/practice"
```

TRY THIS

Test this out by temporarily using the R temporary directory.

```
library(kutils)
owd <- getwd()
tdir <- tempdir()
setwd(tdir)
initProject()
list.files()
# go back to your original directory
setwd(owd)
```

Step 1. Download

- Could retrieve files in a web browser.
- R has functions for downloading files.
- My R code is in the R subdirectory. Use `dir.create()` to create a folder "../data"

```
if(!file.exists("../data"))
    dir.create("../data", showWarnings = TRUE)
```

  - Confession: the first thing a programmer will say is "don't type strings into code that way".

```
ddir <- "../data"
if(!file.exists(ddir)) dir.create(ddir,
    showWarnings = TRUE)
```

R's `download.file()` function can retrieve text files without trouble

Step 1. Download ...

```
fn <- "ortann.csv"
URL <-
    paste0("http://pj.freefaculty.org/guides/",
                   "stat/DataSets/OregonTemps/",
                      fn)
download.file(URL, destfile = file.path(ddir,
    "oregon.csv"))
list.files(ddir)
```

Step 2-5. Import, Analyze, Reconsider

- We will go through this in detail in the following sessions

Step 6. Output

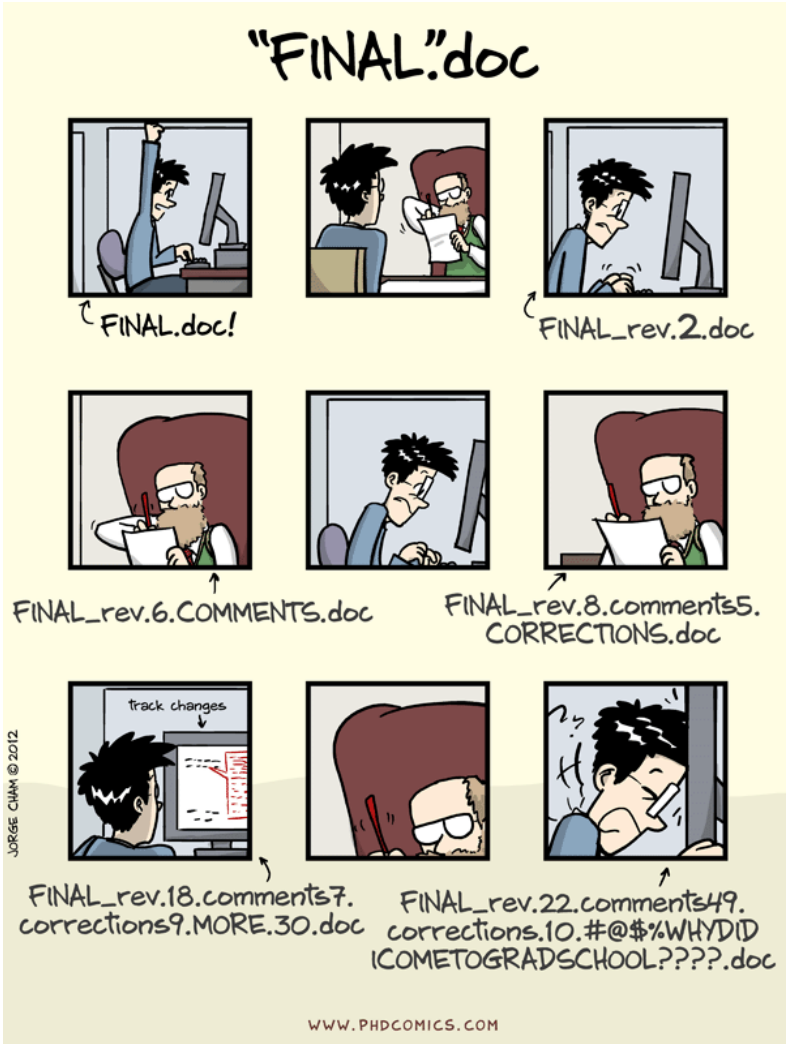
- Can export graphics to devices of many types (pdf, png, etc)
- Can export ready-for-production tables of many types
- Documents: Don't "embed"! Rather "link to" tables and graphs (so document does not need copy/paste editing to update figures and tables).
- Key idea: if need arises to correct, add, or remove data, possible to re-produce tables & graphs instantly

Outline

Version Management

- Use a Project Management framework.
  - Register and Track project files over time
  - Accumulate a comprehensive file history
- Avoid the "how did I get that number?" problem!
- Synchronize one computer's copy with a central server
- Integrate team of workers

Why Version Management?



- Git is one of the most popular programs for version management.
  - We wrote a long(ish) guide, "Git it Together" (http://crmda.ku.edu/guides).
- Remote server not necessary. Tracking can be strictly inside your computer
- Can use GitHub, Gitlab, BitBucket, etc
- not Dropbox

Reproducible Documents

bad: point and click, gaze at output, and hand-type a table into a paper

poor: point and click, copy/paste some output into a paper, re-arrange by hand

good: export tables and figures in format ready for inclusion in paper

better: Re-produce all calculations in process of producing paper.

Sweave

- For LaTeX documents, developed by Friedrich Leisch and the R Core for inclusion of R code in "literate documents"
- Here's the basic idea. Inside a LaTeX document, R code chunks are included

```
<<swexample1, fig=T, include=F, eval=F>>=
 x <- rnorm(100, mean = 39, sd = 22)
 hist(x, main = "Sweave code included")
@
```

- knitr, developed by Yihui Xie, similar to Sweave, offers different customizations
  - Xie, Yihui (2015) *Dynamic Documents with R and knitr*. 2nd edition. Chapman and Hall/CRC.
- Rmarkdown. A completely different framework that flows out of the "markdown" movement for preparing web pages

KU

I'm between good and better

- I Sweave most of my lectures
  - Eliminates the mismatch between code examples and what students actually get
- Here's an "I'm smarter than you" moment: http://pj.freefaculty.org/R/gloating/test2
- Sometimes I'll write two documents at once,
  1. An Sweave document that writes "bits" of graphs and tables into separate files.
  2. A Sweave document that incorporates results, makes a presentation
- Why separate those 2 steps?
  - Cobbling everything into one giant document sometimes causes hard-to-find bugs

KU

Project Management and Interacting with Clients

- In several CRMDA projects, we notice a dangerous tendency to develop R scripts with 100s of recodes and calculations that are impossible to proofread (without herculean effort).
- In a project for a US Government agency, we needed to
  - import 18 separate SPSS data files
  - deal with problem that each survey could change variable names and recode the scales!
- We developed a
  - Variable Key spreadsheet file so that the client could inspect the variable names and recodings
  - R code that could import the Variable Key and carry out the recoding

KU

Project Management and Interacting with Clients ...

	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	wave15	wave15v	wave16	wave16v	wave17	wave17v	wave18	wave18v	wave19	wave19v	wave20	wave20v	wave21	wave21v	
2	Wave 15_	Verbatim	Wave 16_	Verbatim	Wave 17_	Verbatim	Wave 18_	Verbatim	ANQAR_V	Verbatim	ANQAR W	Verbatim	Data_ANQ	Verbatim	
3	Q4.1	1|2|3	q1	1|2|3	q1	1|2|3	q1	1|2|3	q1	1|2|3	q1	1|2|3	q1	1|2|3	
4	Q4.2	1|2|3	q2	1|2|3	q2	1|2|3	q2	1|2|3	q2	1|2|3	q2	1|2|3	q2	1|2|3	
5	Q4.3	1|2|3|4	q3	1|2|3|4|5	q3	1|2|3|4	q3	1|2|3|4|5	q3	1|2|3|4|5	q3	1|2|3|4|5	q3	1|2|3|4|5	
6	Q4.5	1|2|3|4	q8	1|2|3|4|5	q10	1|2|3|4	q10	1|2|3|4|5	q10	1|2|3|4|5	q10	1|2|3|4|5	q10	1|2|3|4|5	
7	Q3.2	1|3|2	q25	1|3|2	q27	1|3|2	q27	1|3|2	q28	1|3|2	q26	1|3|2	q31	1|3|2	
8	Q4.16	6|5|4|3	q12	6|5|4|3|2	q15	6|5|4|3	q14	6|5|4|3|2|1	q14	6|5|4|3|2	q14	6|5|4|3|2	q15	6|5|4|3|2|1	
9	Q4.18	1|2|3|4	q14	1|2|3|4|5	q17	1|2|3|4	q16	1|2|3|4|5	q18	1|2|3|4|5	q17	1|2|3|4|5	q19	1|2|3|4|5	
10	Q4.21	6|5|4|3	q20	6|5|4|3|2	q22	6|5|4|3	q22	6|5|4|3|2|1	q23	6|5|4|3|2	q22	6|5|4|3|2	q27	6|5|4|3|2|1	
11	Q3.3	1|2|3|4	q18	1|2|3|4|5	q20	1|2|3|4	q20	1|2|3|4|5	q21	1|2|3|4|5	q20	1|2|3|4|5	q25	1|2|3|4|5	
12	Q4.19	1|2	q13	1|2	q16	1|2	q15	1|2	q16	1|2	q16	1|2	q18	1|2	
13	Q4.22	1|2	q17	1|2	q19	1|2	q19	1|2	q20	1|2	q19	1|2	q23	1|2	
14	Q2.4.5	1|2|3|4	q32e	1|2|3|4|5	q34e	1|2|3|4	q34e	1|2|3|4|5	q36e	1|2|3|4|5	q33e	1|2|3|4|5	q39e	1|2|3|4|5	
15	Q2.5.5	1|2|3|4	q30e	1|2|3|4|5	q32e	1|2|3|4	q32e	1|2|3|4|5	q34e	1|2|3|4|5	q31e	1|2|3|4|5	q37e	1|2|3|4|5	
16	Q2.4.1	1|2|3|4	q32a	1|2|3|4|5	q34a	1|2|3|4	q34a	1|2|3|4|5	q36a	1|2|3|4|5	NA	NA	q39a	1|2|3|4|5	
17	Q2.10.1	2|1	NA	NA	q40a	2|1	q41a	2|1	NA	NA	NA	NA	NA	NA	
18	Q2.10.2	2|1	NA	NA	q40b	2|1	q41b	2|1	NA	NA	NA	NA	NA	NA	
19	Q2.10.3	2|1	NA	NA	q40c	2|1	q41c	2|1	NA	NA	NA	NA	NA	NA	
20	Q2.9.1	1|2|3|4	NA	NA	q41	1|2|3|4	q42	1|2|3|4|5	NA	NA	q39b	1|2|3|4|5	i	1|2|3|4|5	
21	Q2.12.3	5|4|3|2	q48b	5|4|3|2|1	q53b	5|4|3|2	q54b	5|4|3|2|1	q50b	5|4|3|2|1	q50b	5|4|3|2|1	q52b	5|4|3|2|1	
22	Q2.5.1	1|2|3|4	q30a	1|2|3|4|5	q32a	1|2|3|4	q32a	1|2|3|4|5	q34a	1|2|3|4|5	q31a	1|2|3|4|5	q37a	1|2|3|4|5	
23	Q2.11.1	2|1	q37a	2|1	q43a	2|1	q44a	2|1	q41a	2|1	NA	NA	NA	NA	
24	Q2.11.2	2|1	q37b	2|1	q43b	2|1	q44b	2|1	q41b	2|1	NA	NA	NA	NA	
25	Q2.11.3	2|1	q37c	2|1	q43c	2|1	q44c	2|1	q41c	2|1	NA	NA	NA	NA	
26	Q2.9.2	1|2|3|4	q38	1|2|3|4|5	q44b	1|2|3|4	q45	1|2|3|4|5	NA	NA	q39c	1|2|3|4|5	q44d	1|2|3|4|5	
27	Q2.12.5	5|4|3|2	q48c	5|4|3|2|1	q53d	5|4|3|2	q54c	5|4|3|2|1	q50d	5|4|3|2|1	q50d	5|4|3|2|1	q52d	5|4|3|2|1	
28	NA	NA	q10	1|2|3|4	q13	1|2|3|4	q12	1|2|3|4	q12	1|2|3|4	q12	1|2|3|4	q12	1|2|3|4	
29	Q3.4	2|1	q70	2|1	q78	2|1	q77	2|1	q78	2|1	q76	2|1	q79	2|1	
30	NA	NA	NA	NA	q35a	1|2|3	NA	NA	NA	NA	NA	NA	NA	NA	
31	NA	NA	NA	NA	q35b	1|2|3	NA	NA	NA	NA	NA	NA	q40b	1|2|3	

- By separating the substance of variable names and recodes from the details of R code, we reduced the danger that typographical errors would slide through without notice.

Project Management and Interacting with Clients ...

- Also, we get improved communication with client, who was not fluent in R, but was fluent in English!
- To make the Variable Key framework work in day-to-day projects, we offer functions in `kutils` named `keyTemplate()`, `keyImport()`, and `keyApply()`, along with a host of support functions.

Outline

Research Life is 6 Steps

1. Download
2. Import
3. Review
4. Recode
5. Analyze
6. Plot & Table

References

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Session

```
sessionInfo()
```

```
R version 3.6.0 (2019-04-26)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 19.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/atlas/libblas.so.3.10.3
LAPACK: /usr/lib/x86_64-linux-gnu/atlas/liblapack.so.3.10.3

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
     LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8
     LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8
    LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
[1] compiler_3.6.0 tools_3.6.0
```