

the replication principle

adam okulicz-kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Friday 6th September, 2019 17:15

outline

the idea

replication+stata=dofile

get code from others!

outline

the idea

replication+stata=dofile

get code from others!

bad excel

- ◇ never trust numbers that come from excel
 - no way to find out what happened, there's no code!
[*]<http://www.statisticalengineering.com/Weibull/excel.html>
[*]http://www.texasoft.com/excel/Should_You_Use_Excel_for_Statistics.pdf
[*]<http://andrewgelman.com/2013/04/17/excel-bashing/>
- ◇ tell a story about excel when I learned it hard way:
 - my first paper for ecological economics, done in excel
 - reviewers got back after 6mo, i had dozens of excel files
 - couldn't replicate my own results!

replication, replication

- ◇ replication=write computer code that will do *everything*
 - from raw data (eg FED, IMF) to results (eg regression)
- ◇ necessary for science
- ◇ otherwise we don't know what happened
- ◇ how was it calculated? is there a mistake? who knows?
- ◇ pol sci perspective
 - [*]<http://gking.harvard.edu/files/gking/files/replication.pdf>
- ◇ IT perspective
 - [*]<http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001745>

humans and mistakes

- ◇ a part of human nature is that we make mistakes
 - can't avoid it no matter what's your skills, experience, etc.
 - same pertains to academic research
- ◇ computers, on the other hand, never make mistakes
 - they just do whatever humans tell them to do
 - sometimes they execute our mistakes

implications for every day practice [will come back to it when we analyze data!]

- ◇ once you have coded everything, double/triple-check it
 - leave it aside and check again
 - show it to other people, post on your website
- ◇ the more times it is checked, the fewer mistakes
- ◇ cross-check end output with raw data—e.g. are there the same numbers for randomly chosen data points
- ◇ does it make sense?
- ◇ check with alternative data sources? do they tell the same story?
 - i always google tables and graphs of what i study

outline

the idea

replication+stata=dofile

get code from others!

dofile

- ◇ GUI and command window OK for playing around
- ◇ sometimes handy to use command window or GUI
- ◇ but in the end, everything must be in dofile
- ◇ can write in dofile and run from there: highlight+Ctrl-d
- ◇ dofile must do *everything*:
 - produce final output (usually descr and inferential stats)
 - from the very raw data (data someone gave you)
- ◇ so always first load raw data, manage, organize, manipulate
 - and only then produce some results

dofile

- ◇ just a text file (.do)
- ◇ click “new do-file editor” icon: new window pops up
- ◇ file-open...and open dofile for today
- ◇ it has all the code we will use today
- ◇ highlight code you want to run and press Ctrl-d
- ◇ can have many dofiles opened at the same time
- ◇ can copy-paste between dofile and:
 - command window, review window, and results window
- ◇ don't forget to save your dofile: file-save as

outline

the idea

replication+stata=dofile

get code from others!

examples: dofiles

◇ examples for intl, country level, comparative:

- <https://www.prio.org/JPR/Datasets/>
- <http://www.isanet.org/Publications/ISQ/Replication-Data>
- <https://huber.research.yale.edu/writings.html>

the easiest way to do research in 21st century

- ◇ start with code others wrote, and build on their work
- ◇ this is the fastest, most efficient way to do research
- ◇ any research very close to yours, just email author and ask her to share code with you
- ◇ even if it sas or spss etc—you'll be able to figure it out quickly what is going on there and then implement something similar in stata
- ◇ don't reinvent the wheel: almost as if you were to start research without reading literature and had to come up with all theories and ideas on your own!