the replication principle

adam okulicz-kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Tuesday 5th September, 2017 17:16

outline

bad excel

the idea

replication+stata=dofile

get code from others!

outline

bad excel

the idea

replication+stata=dofile

get code from others!

bad excel

- better teach nothing than excel
- take this from the class:
- even if you do not use statistical software:
- never trust numbers that come from excel
- in your future careers, do not trust people working with excel
- simply, it's very likely there are mistakes
- and worse, there is no way to find out what happened
- there's no code !

more elaboration

- Friends don't let Friends use Excel for Statistics
  [*]https://oit.utk.edu/research/documentation/Documents/ExcelStatProbs.pdf
- Excel's Checkered Statistical Past
  [*]http://www.statisticalengineering.com/Weibull/excel.html
- "Should you use Excel to teach statistics?"
  [*]http://www.texasoft.com/excel/Should_You_Use_Excel_for_Statistics.pdf
- See Andrew Gelman's blog. Funny.
  [*]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 file
  - couldn't replicate my own results!

outline

bad excel

the idea

replication+stata=dofile

get code from others!

replication, replication

- replication=write computer code that will do *everything*
- necessary for science
- otherwise we don't know what happened
- how was it calculated? is there a mistake? who knows?
- for elaboration see

[*]http://gking.harvard.edu/files/gking/files/replication.pdf

intuition

- replication=automation
- "have computer code that produces whatever is the end result from the raw data"
- the end result may be: a map (e.g. thematic map of voting), a graph (e.g. histogram of income), regression table, and so forth...
- the raw data is the data that somebody gave you
  - it is usually downloaded from some website, say IMF, UNDP, and so forth
- from the moment somebody gives you data, you are responsible for the rest

humans and mistakes

- a part of human nature is that we make mistakes
- you simply can't avoid it no mater what is your knowledge, skills, experience, etc.
- and the more complicated is your task, the more mistakes you're going to make
- same pertains to academic research
  - when we do academic research, we make mistakes along the way
  - the more complicated the research the more mistakes

computers and mistakes

- 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

- 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
- everything has been studied by others and it is good to cross-check

outline

bad excel

the idea

replication+stata=dofile

get code from others!

dofile

- we follow replication principle by writing dofiles
- 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
- and 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 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

bad excel

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`

the best way to do research in 21st century

- start with code others wrote, and build on their work
- this is the best, 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!