the replication principle

adam okulicz-kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Tuesday 16$^{\text{th}}$ January, 2018   15:15

__outline__

bad excel

the idea

replication+stata=dofile

get code from others!

<u>outline</u>

bad excel

the idea

replication+stata=dofile

get code from others!

bad excel

$\diamond$ better teach nothing than excel

$\diamond$ take this from the class:

$\diamond$ even if you do not use statistical software:

$\diamond$ never trust numbers that come from excel

$\diamond$ in your future careers, do not trust people working with excel

$\diamond$ simply, it's very likely there are mistakes

$\cdot$ and worse, there is no way to find out what happened

$\cdot$ 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*

· 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

$\diamond$ a part of human nature is that we make mistakes

$\diamond$ can't avoid it no mater what is your knowledge, skills, experience, etc.

$\diamond$ same pertains to academic research

$\diamond$ computers, on the other hand, never make mistakes

$\diamond$ they just do whatever humans tell them to do

$\diamond$ 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

$\diamond$ we follow replication principle by writing dofiles

$\diamond$ GUI and command window OK for playing around

$\diamond$ sometimes handy to use command window or GUI

$\diamond$ but in the end, everything must be in dofile

$\diamond$ and can write in dofile and run from there:
  highlight+Ctrl-d

$\diamond$ dofile must do *everything*:

$\cdot$ produce final output (usually descr and inferential stats)

$\cdot$ from the very raw data (data someone gave you)

$\diamond$ so always first load raw data, manage, organize,
  manipulate

$\cdot$ 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

<u>outline</u>

bad excel

the idea

replication+stata=dofile

get code from others!

examples: dofiles

$\diamond$ 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 best way to do research in 21st century

$\diamond$ start with code others wrote, and build on their work

$\diamond$ this is the best, most efficient way to do research

$\diamond$ any research very close to yours, just email author and ask her to share code with you

$\diamond$ 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

$\diamond$ 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!