data

adam okulicz-kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Thursday 30th January, 2025 17:20

outline

replication

data basics

merge

tips

outline

replication

data basics

merge

tips

replication, replication

- replication=write computer code that will do *everything* from raw data (eg FED, IMF) to vis
- necessary for science– otherwise don't know what's up: how was it calculated? is there a mistake? who knows?
- `http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001745` [superb! read it!]

outline

replication

data basics

merge

tips

data basics

- dataset is a matrix
- cols are variables (var), rows are observations (obs; U/As), and vars are characteristics of obs
- eg 'edu, 'age', and 'inc are vars and persons are obs
  - each row is a separate person
- have data clean! eg only one top row for var names
  - (xls is typically a mess with unusable var names)

be careful and clear

- define key vars in as much detail as possible
    - eg “income”$-\ >$“median hh income in current USD”
- think about limitations, shortcomings
    - eg sampling error, missing data, etc
- try to triangulate: measure the concept with multiple vars

outline

replication

data basics

merge

tips

the power of merge

- merge as much as possible! great value!

◦ one of the most useful things you'll learn in class

◦ there's a ton of data and growing

- great value comes from simple fact of merging

◦ using just one data can only do so much

◦ by merging easily create dataset that nobody else has

◦ and produce insight nobody else has

- eg `https://www.amazon.com/gp/product/0063032376`

easy to merge; difficult to do it right

- the challenge is to check what happened after the merge
- **always investigate carefully non-merges**
- **make sure that *ALL* nonmerges are as expected**
- **even matches can be wrong**

○ use vis to investigate and be skeptical: does it makes sense?

- typically non-merges bc of diff coding, eg: “Poland≠ “Rep. of Poland”; “CAMDEN”≠”Camden”
- go back and fix it before merge:
- replace to “Poland” from “Rep. of Poland”
- often wasn’t supposed to merge

○ eg data A: 1995-2000, but B: 1990-1998

merging investigation

- tab _merge
- cross-tab _merge with geography and/or time
  - say year and state
- want to list relevant parts of df:
  - _merge and key/id vars: geo, time, etc
  - and sort on key vars
- it may take time to find out what happened
- be clear about nonmerges!
  - how many nonmerges and what you did about it
  - eg dropped, fixed, etc

what to merge on?

- geography! usually have some
- and can aggregate up, say groupby state
- time! say with weather (weather usually matters)
- occupation–there are occ codes eg `https://www.onetonline.org/find/descriptor/result/4.A.2.b.2`

outline

replication

data basics

merge

tips

data choice matters

- data management often takes 50-90% of time
- most of it is learning/figuring out data
- you'll spend 100+ hrs learning about specific datasets
- dont waste time! pick data that:
  - you're passionate about (eg sth you went to school to learn about, eg poverty, inequality, discrimination)
  - you'll use in other classes, possibly for thesis
  - advance your career after graduation, eg want to work for state–use data they produce or use a lot

make lots of comments in your code

- make comments in notebook in code cells, important!
  - eg explain to yourself what command does, what to look for
- and use plenty of text cells
- if you do not make comments, you will forget
- use handy keywords like “TODO”, “BUG”, “LATER”, “FIXME”
  - ctrl-f

datasets of the day

- climate/weather, down to county (easy access!)
  - https://wonder.cdc.gov/EnvironmentalClimateData.html
- religion!
  - https://www.thearda.com/data-archive?tab=1&fid=RCMSCY10
- state level policy https://www.statepolicyindex.com/data/