

introduction

Adam Okulicz-Kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Thursday 13th September, 2018 13:49

outline

misc

why data management?

Stata v other software

let's fire up stata

[*] bonus-data sources [skip, can look at home]

outline

misc

why data management?

Stata v other software

let's fire up stata

[*] bonus—data sources [skip, can look at home]

about myself...

- `theaok.github.io`
- and more importantly `http://scholar.google.com/citations?hl=en&user=pz9RYloAAAAJ&view_op=list_works&sortby=pubdate`
- and i mostly use: wvs, gss, brfss, and other surveys
- if you use same data then i can share my code with you!
-
- except grades and ps submission on Sakai, everything at:
`theaok.github.io/dirStu`

yourself? (see if others overlap: can collaborate!)

- ◇ research interests and data?
- ◇ software? eg SAS, SPSS, Stata, Python, R
- ◇ specific expectations for this class?
-
- [i'll email you some data based on your interets; bother me for more!]

weekly labs

- small class so will just flip it often, esp in 2nd part

mission statement/core values

- replication [replication.pdf]
- coding/programming; no pointing-clicking
- collaboration, sharing, open-source
- be lazy! (copy from others incl my code)
- it is a directed study, so you must have the topic and data or figure it out asap!

outline

misc

why data management?

Stata v other software

let's fire up stata

[*] bonus—data sources [skip, can look at home]

data management is fundamental

- in order to analyze data you need to manage it first
- GIGO (Garbage In Garbage Out)
 - if data management fails, data analysis fails

preparing data vs analyzing data

- it takes more time to prepare data than to analyze it
- it may take years to prepare a dataset, and hours to analyze it
- it may be hundreds of lines of code to prepare dataset and several lines to analyze it
- start early !

it's all about productivity

- productivity = output/time \approx \$
- reuse your code, copy from others, google
- minimize time spent on coding
- do other things: eg read journal articles, books
- be strategic: do spend time on coding if this will cut more time later than you invest now doing it

outline

misc

why data management?

Stata v other software

let's fire up stata

[*] bonus—data sources [skip, can look at home]

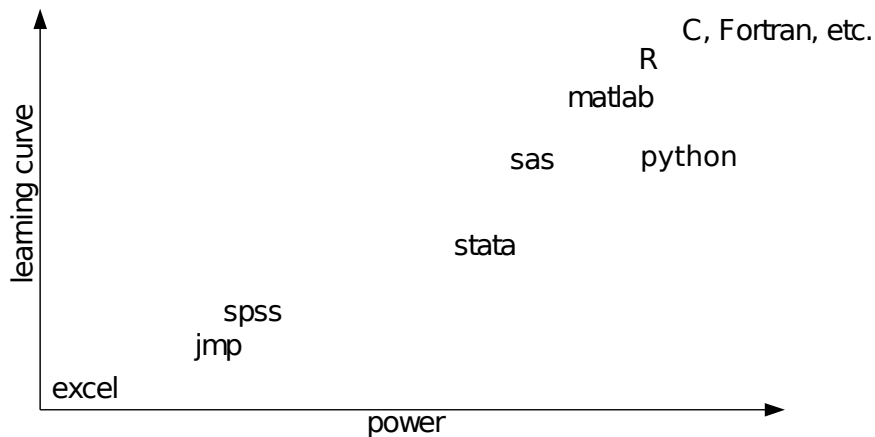
a critical decision!

- it takes months to get productive with software
- it takes years to master software
- huge time investment
- in soc sci dat man the choice is: Sas, Stata, R, Python
- there's more (Lisrel, HLM, etc) but the above are major
- excel and spss are junk that no one should use

which one?

- Stata: powerful, no need to learn any other software; sufficient for vast majority of projects: data analysis, data management and graphics
- R: most powerful statistical software (and coming up!)
- Stata: user friendly, fast, very concise code
- R: user unfriendly, slow; weird code!
- Stata, R: great user community: listserv, websites, etc.
- Sas: a dinosaur (still, often industry standard), very verbose
- R: free, Stata: around \$200; sas over \$1k

which one?



which stata

stata editions	# observations	# variables
small(student version) [useless!]	1,000	99
Intercooled (standard version)	based on ram in your computer	2,047
se [for large datasets]		32,767
mp (multi-processor) [fastest]		32,767

- just use Stata-IC (Intercooled), perpetual license: \approx \$200
- <https://www.stata.com/order/new/edu/gradplans/campus-gradplan>

outline

misc

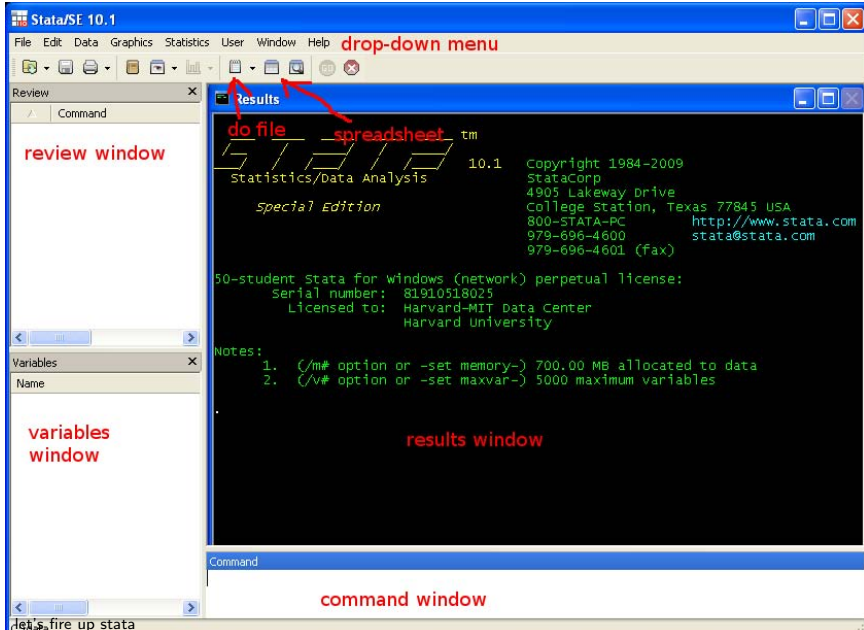
why data management?

Stata v other software

let's fire up stata

[*] bonus—data sources [skip, can look at home]

stata interface; and do intro.do



looking

- exploring your data is critical!
- we will learn many commands to understand data
- most basic: `d` `sum` `tab`
- remember this:
 - to manage data well, you need it to understand it well
 - it takes a lot of time to understand it well, and hence
 - either manage data you are already familiar with
 - or data that you are ready to invest a lot of time into knowing
- again, the bottomline in this class is to manage data that is of great interest to you

outline

misc

why data management?

Stata v other software

let's fire up stata

[*] bonus-data sources [skip, can look at home]

data.gov

- `http://www.data.gov/`

data sources

- ◇ <http://www.worldvaluessurvey.org/>
- ◇ <http://www.norc.uchicago.edu/GSS+Website/>
- ◇ <http://www.icpsr.umich.edu/icpsrweb/ICPSR/>
- ◇ <http://www.thearda.com/>
- ◇ <http://ksghome.harvard.edu/~pnorris/Data/Data.htm>

more data sources

- ◇ <http://www.measureofamerica.org/>
- ◇ <http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/0,,contentMDK:20388241~menuPK:665266~pagePK:64165401~piPK:64165026~theSitePK:469382,00.html>
- ◇ <http://usa.ipums.org/usa/>
- ◇ <https://international.ipums.org/international/>

“non-traditional” data

- ◇ `http://dvn.iq.harvard.edu/dvn/dv/patent`
- ◇ `http://www.trustlet.org/wiki/Trust_network_datasets`

happiness data

- ◇ <http://www.bmj.com/content/337/bmj.a2338.full>
- ◇ http://apps.facebook.com/usa_gnh/
- ◇ <http://www.facebook.com/notes/facebook-data-team/relationships-and-happiness/304457453858>
- ◇ <http://www.springerlink.com/content/757723154j4w726k/fulltext.pdf>
- ◇ <http://www.wefeelfine.org/>

facebook data

- ◇ `http://apps.facebook.com/usa_gnh/`
- ◇ `http://www.facebook.com/notes/facebook-data-team/relationships-and-happiness/304457453858`
- ◇ `http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919`
- ◇ `http://cyber.law.harvard.edu/node/4682`
- ◇ `http://www.thefacebookproject.com/resource/datasets.html`

more data

- <http://www.stateoftheusa.org/blog.php>
- <http://www.stateoftheusa.org/content/health-measures-for-the-develo.php>
- <http://www.stateoftheusa.org/content/fbi-report-violent-crime-down.php>
- <http://www.stateoftheusa.org/content/economy-seen-as-prompting-cohabitation.php>
- <http://stateoftheusa.org/content/measuring-economic-well-being.php>
- <http://www.stateoftheusa.org/content/report-hispanics-outlive-other-american.php>

LaTeX

- ◇ we may have a separate class, but some links follow:
- ◇ <http://people.hmdc.harvard.edu/~akozaryn/myweb/latex/>
- ◇ <http://www.ats.ucla.edu/stat/stata/latex/default.htm>
- ◇ i will post my do-files that output into LaTeX