data formats

adam okulicz-kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Friday 14th September, 2018 12:44

outline

misc

data types

Stata

import/export

outline

today's class

- still going slow today; next week will be last slow class!
  - if too slow for you, do extra things!
  - check out recommended materials, start working on final project, etc
- today, still basic so that everybody has the basics covered
  - everybody is at different level
- if you are bored: help others, check extra materials, see help files and experiment with commands
- we will start little more advanced topics next week

outline

misc

data types

Stata

import/export

data basics

- dataset is a matrix
- columns are variables (var), rows are observations (obs)
- obs are also often referred to as U/A
- variables are characteristics of observations
- e.g., 'education', 'age', and 'income' are variables and persons are observations; each row is a separate person

paths

- a location of a file on hard drive
- e.g. C:\Documents and Settings\myfile.txt
- if there is a blank in path, as above, stata needs quotes "C:\ Documents and Settings\ myfile.txt"
- avoid blanks: computers understand blank as a character
- and avoid special characters: everything that is not a letter or a number, say $ % &
- special characters have special meaning for a computer

finding the path

- Windows: to find the path right-click the file$-$ $>$ properties
- Mac: ctrl-left-click the file $-$ $>$ get info

paths

- remember that you write code that should run on other computers
- and remember to cd first to desired directory, so you can say
- cd ????
- and then log using ps1.log
- as opposed to:

```
log using C:\Users\Documents\ASTATA\PS1.txt
```

- that won't run, because I do not have these dirs!
- and it is messy to repeat path for each reading/writing

putting data online

- usually the biggest issue was to put data online!
- eg for google sites i often get error:
  - "You need permission"
- so the file you've put up online was not made public
- maybe better try wordpress.com, or dropbox.com, or sth else
- make sure it works!
- say try it on apps.rutgers.edu or some other computer
- it is important it runs out of the box!
- i will be picky about it

data for today

- data we use is a subset of general social survey: `http://www.norc.org/gss+website/`
- probably the most comprehensive social science data for the US
- whatever you study you are likely to find it in gss
- we will look today at income, education and gender across US regions

data types

- there are dozens of data types/formats/files
- a basic distinction: binary files (.dta, .sas7bdat, .sav) v text files (.txt, .dat, .csv, .tab)
- you can open text file with text editor and you cannot open a binary file (there will be weird characters)
- in fact, can open text file with almost anything
  - (web browser, word processor, etc)
  - (best use text editor like notepad++ or just stata editor)
- but a binary file can be opened with specific software only
- `http://www.cs.umd.edu/class/sum2003/cmsc311/Notes/BitOp/asciiBin.html`
- `http://en.wikipedia.org/wiki/Binary_file`

what is a text or ascii file

- *not* ms word file (.doc)
- unformatted text (same font, no bold, italics, etc.)
- but you can have syntax highlighting – text editor will format *display* of the text (not the text) given some rules/keywords
- e.g. if you open dofile in stata editor it will apply colors etc
- but is just plain text that is displayed by editor based on some rules
- make des, sum, and other words that are stata commands blue
- make everything that follows "//" green etc

data files

- if you are unsure what format you have–open in text editor (can use Stata's dofile editor)
- if it opens, it is text, otherwise it is binary
- text format is great to archive your data for long periods of time (say over 10 years): text format will never change
- but binary format has a great advantage: saves extra info like labels, notes, and is faster

databases

- but wait, we have databases
- outside of academia, in the real world
- all data are in databases
- Oracle, MySQL, NoSQL, MsSQL, or even MsAccess
- we'll talk about those in SQL class in 2nd part of the semester
- sometimes you can use Stata and always Python to pull directly from databases

internet

- but wait, we have internet
- popular internet data types: html, xml, json
- we'll discuss internet data when we do Python
- in the 2nd half of the class
- Python is great at managing internet data

outline

misc

data types

Stata

import/export

make comments in your code

- for each class we will have dofile with Stata code
- make comments in the electronic code files – you will run electronic files not the printout
- if you do not make comments, you will forget...
- use very handy keywords like “LATER” and “FIXME”

commenting

- have preamble (notes, install packages, etc)
- *comment

```
/*comment
block */
```

stata command syntax and getting help

- <command> <variables> , <options>

  `sum var1 var2, detail`
- <variables> and <options> are optional
- command specific syntax is in help files,

  e.g. `help describe`
- `help` if you know command name, eg `help use`
  - esp options, examples, full pdf help

getting help using gui and google

- gui, eg to load/save, edit data, graphs, etc
- google: "stata"+"what you want to do"
  - eg "stata read excel"
- use google a lot! extremely useful!
- again, ucla website is the best:
  `https://stats.idre.ucla.edu/stata/`

tips

- if you did something wrong, load data again and start over
  - (replication: you have dofile and can always start over)
- page -up and -down to get previous/next command in command window
- don't memorize commands but reuse and share code
- learn (naturally) abbreviations, e.g. d for describe
  - (they are underlined in help files)

navigating

- you can navigate in stata: change, list/make/rm dirs and preview files

packages/user-written commands

- to get them either google or `findit`;
- say we want to load spss data e.g. `findit spss` and `help usespss`

outline

misc

data types

Stata

import/export

excel

- many people use it and you may need to import from there
- can save as csv and then insheet
- or just use gui to generate the code you need
- in some cases (as here) gui is useful to generate code
  - File-Import-Excel Spreadsheet
  - Worksheet: Cell Range: Import first row as variable names

fixed format, ascii (text file)

- .raw, .dat, .txt, etc
- in addition to data, a dictionary:
  - tells you which column is which var
- you can open in text editor to see yourself
- dofile

saving

```
//good
use data1.dta
...
save data2.dta

//bad
use data1.dta
...
outsheet data1.tab //loosing var/val labels,notes

//ugly
use data1.dta
...
save, replace //loosing  code in between
```