Stata Exercises Version 5

10/31/19
Chelsea Swete
Econ 280 Fall 2019

Exercise 1

1. Set up econ 280 folder (and relevant subfolders) in \\acsnfs4.ucsd.edu\unix-home\econgrad\yourusername
    - If you are using your personal computer, map the drive (need to have VPN on if not on campus)
2. Open Stata
    a. If you are on a lab computer, just open Stata 15
    b. If you are at home and have Stata on your computer, just open Stata
    c. If you are at home and on campus and don't have Stata, remote into the the ACMS Stata server
    d. If you are at home and not on campus and don't have Stata, use VPN and then the ACMS Stata server

Exercise 2

1. Open and save a do file
2. Create your own header
3. Start a log file
4. Use the built in "auto" dataset
5. Create a frequency table of rep78 including missing as a category
6. Find the mean of mpg
7. Create a variable called high_mpg which is:
    a. 1 if mpg is above mean
    b. 0 if below mean
    c. missing if mpg missing
8. Show mean and range of mpg by high_mpg (check if ranges correct)
9. Create make_only (the text before the first space of the make variable) WITHOUT deleting the make variable
10. Create make_only_num as a numeric version of make_only with the labels of the old strings
11. Create a frequency table of make_only_num
12. Regress price on high_mpg and series of dummy variables based on make_only

13. End log file
14. Clear data
15. Save do file

Commands that you may want to use:
encode
generate
help fvvarlist
log
regress
replace
sort
split
summarize
sysuse
tabulate

Functions you may want to use:
missing()
strpos()
substr()

Don't forget to look at the help files if these commands are new to you. Some of the examples will also use the auto data.

Exercise 3

Source: **IPUMS-USA, University of Minnesota, www.ipums.org**

Download some variables from the 2014 ACS using IPUMS:

https://usa.ipums.org/usa/

year
sample
serial
cbserial
hhwt
statefip
metro
puma
gq
pernum
perwt
nchild
nchlt5
relate
related
sex
age
race
raced
hispan
hispand
educ
educd
empstat
empstatd
labforce
occ
uhrswork
inctot
incwage

Save in your data folder (or subfolder for raw data) as acs_raw.dta

Exercise 4

1. Go to https://usa.ipums.org/usa/
2. Look up the documentation for the variables that are in your dataset
    a. Priority: incwage, sex, educ, race
    b. More if you have time
3. Write down how you would recode/clean the variables to make them more useful for analysis
    a. Consider how missing values are coded
    b. Consider if you may want to code indicators or categorical variables

Exercise 5

Note: if things are running slowly while you're coding, take a sample and run it on that (but don't forget to run on the whole thing at the end)

Don't drop the original variables, you'll check them against the new ones next

1. Open Stata do-file
2. Set up program, including:
    a. Header
    b. References to your personal data folder
    c. Log file
    d. Settings
3. Set directory
4. Use raw data from exercise 3
5. Create race and ethnicity variables as follows:
    a. white: race = 1 and hispan = 0
    b. black: race = 2
    c. asian: race is 4, 5, or 6
    d. race_other: race is 3, 7, 8, or 9
    e. hispanic: hispan is 1, 2, 3, or 4
6. Create a female indicator using the sex variable (look up in codebook how to recode)
7. Create lnwage as the log of wage if wage is not 9999998 or 9999999
8. Download tabmiss and use it on lnwage
9. Create (and label the values) a categorical variable for education:

value	label	How to create

1	Less than HS	Educ between 0 and 5
2	HS	Educ is 6
3	Some College	Educ is between 7 and 9
4	College	Educ is 10 or educd is 100 or 101
5	Graduate School	Educd is 114, 115, or 116
.		If educ and educd are missing

10. Create a (very approximate) experience variable (exp) and experience squared (exp2)
    a. First, create an estimate of how old they were when they started working, assuming
        i. If they didn't complete high school, they started working at 16
        ii. High school grads started at 18
        iii. Some college started at 20
        iv. College started at 22
        v. Master's degree (educd = 114) started at 24
        vi. Professional/Doctor degree (educd = 115 or 116) started at 28
    b. exp = age - age_started_work
    c. If experience is negative, replace as 0
    d. Create experience squared
11. Create a full time indicator where uhrswork is greater than or equal to 35 (and not missing)
12. Preserve your data
13. Import excel file with occupation recodes using "import excel"
14. Rename occ_acs as occ
15. Restore your data
16. Merge in occupation recode
17. Save data as acs_clean
18. Close log and clear data

Exercise 6

Choose 4 variables that you have just created in exercise 5 and prove that you created them correctly by comparing them in some way with the source variable and showing that their ranges/averages are reasonable.

Exercise 7

Calculate the male-female wage gap in this data.

1. Create a new program, header, set up, log, etc
2. Set up Output folder
3. Use clean data from exercise 5 (that was checked in exercise 6)
4. Keep people aged 25 - 64
5. "Human Capital Specification" regress lnwage on:
    a. Female
    b. race/ethnicity categories (leave out white)
    c. education categories (leave out completed high school)
    d. Experience
    e. experience squared
        - Use perwt as a pweight
6. Store estimate
7. "Full/Occ Specification" regress lnwage on:
    a. a-e in step 3
    b. Recoded occupation categories
        - Use perwt as a pweight
8. Store estimate
9. "Fulltime Specification" For fulltime workers, repeat step 6
10. Store estimate
11. Using outreg or outreg2 or estout, export the three estimates as one table (in latex, excel, or word).
    a. Make sure that your table reports standard errors
12. Create a model with interactions (with factor variable notation): add female*black, female*hispanic, female*asian, and female*race_other
13. Export as a table

Exercise 8

Graph similar to:
https://www.stata.com/support/faqs/graphics/gph/graphdocs/bar-chart-with-multiple-bars-graphed-over-another-variable/index.html

1. Create a new program, header, set up, log, output location, etc

2. Make a bar graph similar to the link above: within the categories of black and white, graph male and female average income
    a. Sample: full time workers aged 25-64
    b. Use incwage (make sure you deal with the missings like in exercise 5)
    c. Using collapse, get the average of incwage for each female/black/white combination
    d. Discard information about when white and black are both 0
    e. Reshape so that you have incwage_female and incwage_male (hint: you will need to rename incwage0 as male and incwage1 as female)
3. Formatting the graph:
    a. Y-axis should read "Average Earned Income" in large font
    b. Title should be "Income by Race and Gender for fulltime workers"
    c. Make the light blue background white
    d. The number labels on the y-axis should be horizontal
    e. Use value labels to manipulate the x-axis categories until they read "White" and "Black".
4. Export the graph as a png

Exercise 9

1. Bring in clean data
2. Recode both inctot and incwage so that 9999999 and 9999998 become missing
    a. Create a global that contains inctot and incwage
    b. Loop over the global variable to replace with missing
3. Loop over inctot and incwage again to run regressions of inctot/incwage on female

Exercise 10

1. Flag households where somebody in the household is 17.
    a. Hint: you can use egen and a true/false expression
    b. Note: serial identifies households
2. Using your cleaned data, create an indicator for whether the individuals are married WITHOUT using the marst variable, only using "related". Individuals are married if:
    a. Related = 201 (Q: How are you related to the head of household? Ans: Spouse)
    b. OR {Related = 101 (indicates head of household) AND someone else in the household responded 201 to related}

3. Create an indicator for whether the individuals are part of a cohabiting relationship using related. Individuals are in a cohabiting relationship if:
    a. Related = 1114 (Q: How are you related to the head of household? Ans: Spouse)
    b. OR {Related = 101 (indicates head of household) AND someone else in the household responded 1114 to related}

Exercise 11

1. Create a nested loop that will display month-year combinations from January 2000 to December 2003:
   Jan2000
   Feb2000
   Mar2000
   …
   Nov2003
   Dec2003