Stata Recitation - Week 4 - Modifying Data I

McCourt School of Public Policy, Georgetown University

SWBAT

- Open and save non-system data sets
- Create new variables as functions of existing variables
- Use notes to document workflow
- Work in a do-file

Resource for Variable Creation Nagler, Jonathan. 1995. ÒCoding Style and Good Computing Practices.Ó PS: Political Science and Politics 28(3): 488-492.

Opening non-system data sets.

- Open Stata
- use auto.dta system data set
- save as a new data set using drop-down menus `Desktop\recitation4\myautodata.dta`
- Close stata
- Now you have a data set exactly as if it was distributed by TA online or from email.
- Open Stata
- Open data set
- You will see the `use` command in results window.

Basic variable creation command: Generate

- Make sure to start with an existing data set. Otherwise, there are no observations to assign a value to! "‘ clear sysuse auto generate x1 = 1
- list first 10 observations list if _n<=10

generate x2 = 2 list if _n<=10

generate y = x1 + x2 list if _n<=10

sum x1 x2 y

```
* What if we changed our mind about the value of `x1`?
* You can't generate it again:
```

gen x1=2 x1 already defined r(110);

```
* We could use the replace command:
```

replace x1 = -1

```
* Then we also have to replace y.
```

replace y = x1 + x2

```
* But this is a very bad idea.
* Requires a lot of extra typing, error prone, no clear record.
* The better approach is to re-create the variable correctly.
* You can do this easily, if you are working in a do-file.

### A note on drop (DANGER):
You can use the command `drop` to remove variables from your dataset. However, this should b
```

gen testdrop drop testdrop

```
### Exercise
* Create a do-file that creates x1, x2, and y.
* Do-file should include -use- commands
* Change the value of x1 and re-run the do-file.
* Use the do-file to input the remaining commands for this session.


## Saving Progress

- Now we have made some changes to the data set that we want to preserve.
- Your main product is the do-file, not the data set.
- We save both, but the do-file is the important part.
- You may want to turn on the `Save before do/run` option.
- Whenever you start working on that data again, start from the do-file, not the data set.
- Add a `save , replace` command to the end of the do-file.
- Give saved data set a new name



* Do file should have these commands.
```

use "...\auto-r4.dta", clear

generate x1 = 1 generate x2 = 2 generate y = x1 + x2

sum x1 x2 y save "...\auto-r4-modified.dta", replace

Re-run do file with `save` command.
Close out of Stata.
Reopen Stata, open do-file, run do-file.

Operators and Functions
- Add variable creation commands to do-file before the save command.
* We can also create new variables from existing variables
* A list of operators can be found here:

help operators generate weightforlength = weight/length

browse weight length weightforlength sum weight length weightforlength

generate mpgsq = mpg^2 browse mpg mpgsq sum mpg mpgsq

* You can also do more complex operations with functions

help functions

generate logmpg = log(mpg) browse mpg logmpg sum mpg logmpg

generate int_headroom = round(headroom)

* Or combine functions and operators

generate z1 = log(mpg* turn) + sqrt(abs(x1 * gear_ratio))

Commenting your work

- Some data transformations are self-explanatory, i.e. `logmpg = log(mpg)`
- Most are not.
- Use comments to document what you are doing,
- helpful for other people reading your code
- helpful for yourself when you re-read your code, and have no idea what you were doing.
- Go back and comment your do-file.

After Class Exercise

Answer the following questions using the example data set: `census.dta`

```
Use a do-file to produce the output that you used to arrive at your answers.

Use comments before and after the command to document the question you
are answering and the answer.

When your do-file is complete, create a log to save the results.

1. Create a new variable giving the total population age 17 and younger in each state. What
2. Create a new variable giving the proportion of residents age 17 and younger in each state
3. Create a new variable giving number of marriages as a proportion of total population.
   What state had the highest proportion of marriages?
   What was that proportion?
4. Create a new variable giving number of divorces as a proportion of total population.
   What state had the highest proportion of divorces?
   What was that proportion?
5. What is the average rural (not urban) population for all states?
6. Create a new variable giving the log of population for each state.
   Create histograms for population and log population.
   Title graphs appropriately.
   Export graphs and insert into a Word document.


### After Class Exercise Answer
```

clear sysuse census.dta

- 1. Create a new variable giving the total population age 17 and younger in each state.
- What is the average state under population of age 17 and younger?

generate poplt18 = poplt5 + pop5_17 summarize poplt18

- 1,272,229
- 2. Create a new variable giving the proportion of residents age 17 and younger in each state.
- Which state has the lowest proportion of population age 17 and younger?
- What was that proportion?

generate proportionlt18 = poplt18 / pop sum proportionlt18 list state proportionlt18 if proportionlt18<.25

- Florida 0.242

- 3. Create a new variable giving number of marriages as a proportion of total population.
- What state had the highest proportion of marriages?
- What was that proportion?

generate proportionmarriage = marriage / pop sum proportionmarriage list state proportionmarriage if proportionmarriage > 0.14

- Nevada 0.143
- 4. Create a new variable giving number of divorces as a proportion of total population.
- What state had the highest proportion of divorces?
- What was that proportion?

generate proportiondivorce = divorce / pop sum proportiondivorce list state proportiondivorce if proportiondivorce > 0.017

- Nevada 0.017
- 5. What is the average rural (non-urban) population for all states?

generate poprural = pop - popurban sum poprural

- 1,189,896
- 6. Create a new variable giving the log of population for each state.
- Create histograms for population and log population.
- Title graphs appropriately.
- Export graphs and insert into a Word document.

gen logpop = log(pop)

histogram pop , title(Distribution of State Population) graph export popdist.png, replace

histogram logpop , title(Distribution of State Log Population) graph export logpopdist.png, replace “‘