Stata Recitation - Week 10 - End of Semester Project

McCourt School of Public Policy, Georgetown University

Students Should Know

- Develop a do-file to transform initial data set to analysis data set
- Document results using comments
- Use relative path names to organize a project

RESOURCES

[http://www.haverford.edu/TIER]

Warm Up problems using the `nslw88.dta` example dataset:

```
sysuse nlsw88.dta, clear
```

1. Regress hourly wage on total workforce experience. As your workforce experience increases, what is the predicted effect on your wage?
2. Predict the wage for a worker with 14 years of experience.
3. Three people in the dataset have exactly 14.25 years of experience. List their actual wages.
4. Regress the effect of hourly wage on total workforce experience, being in a union, living in a city, and being a college graduate. As your workforce experience increases, controlling for these factors, what is the predicted effect on your wage?

Answer to the warm up problems

```
*1
regress wage ttl_exp

* Effect of 1 unit (year) increase in experience is a 0.3314 increase in hourly wage

*2
display 3.61+0.3314*14
*or:
predict wagep
list wagep if ttl_exp == 14
```

```
*3
list wage if ttl_exp == 14.25

*4
regress wage ttl_exp union c_city collgrad
* Effect of a 1 unit (year) increase in experience is a 0.295 increase in wage, controlling
```

End of Semester Project

At the end of this semester, students will be given a simulation project. It will be larger and more complex than any exercises we've done so far.

This recitation will simulate a large data project. We will complete a small and simple project, but organize the project as if it were a large and complex project.

Set-up

First create your main project folder, titled `recitation10`, on the desktop.

Open Stata, open the `auto.dta` data set

Save a copy of the `auto.dta` data set, named `auto_raw.dta` in the new `recitation10` folder

So far, we've generally been working with full path names.

This is for small projects, or if you're always working on the same computer.

If you have a large project, where you're saving files, graphs, etc, the full path name becomes cumbersome

If you have to change computers, you'll have to change the path name throughout.

Instead, for a large project, set the ***working directory***.

Working Directory

Stata has a single working directory where it is looking at any given time. It is displayed at the bottom of the window or use pwd.

Anything that you save using typed commands will save to this directory. If you try to open anything with a typed command, it will only look in this directory.

See what is currently in your working directory: ls / dir

Working directory is unrelated to any saving or opening that you do using the drop-down menus in Stata or in the do-file editor.

The only way to change it is with the cd command or the Change Working Directory menu item.

Use the Change Working Directory drop-down to change to the recitation folder. Copy the cd command to the do-file as the first line

```
*change working directory
*cd "C:\Users\gppilab\Desktop\Recitation 10"

sysuse auto, clear

save auto_raw.dta, replace

clear
use auto_raw.dta
```

COMMANDS FOR DATA MANIPULATION

```
save auto_final.dta, replace
```

- As you progress through the project, you will develop the do-file to create
- your final analysis data set. You generally do not have to save intermediate
- data sets. Instead, just fill in the commands to create your final data set.
- If you make a mistake, just fix it in the do-file and re-run it.

Suppose you received the following data description. Fill in your do-file with commands and comments to replicate the analysis that follows:

This study looks at the relationship between vehicle mileage, weight, length, and price. We also examine whether this relationship changes according to the vehicle's repair record.

Five records were dropped from our data, as they were missing data on their repair record. Data on mileage for Datsun vehicles is known to be inaccurate, so all Datsun automobiles were dropped from the sample. Thus, our final analysis sample contained 65 records.

We had to correct two problems in price reporting. First, prices for domestic automobiles were reported in 1978 nominal dollars, while prices for foreign automobiles were reported in 1977 nominal dollars. To correct for this discrepency, we inflated the prices of foreign automobiles by the 1977 inflation rate of 6.8%.

We also know that some very high and very low prices are mistakes in the data. The maximum automobile price in 1978 was $15,000 and the minimum price was 4,000. For prices outside this range, we recoded the value of price to missing, or ".". We did not want to loose these observations completely, so we did not drop them from the data set. This correction resulted in 7 missing values for the price variables. The mean of the corrected price variable is 6364.73.

In addition to the basic variables, mileage, weight, and length, we analysed specifications with variable transformations including the natural logs of mileage, weight, and length, and the ratio of weight to length.

When analyzing records with different repair records, we divided the sample into two groups. The first group had a low repair record, with a value of 1, 2, or 3. The second group had a high repair record, with a value of 4 or 5. The distribution of the final sample was roughly 60% low, 40% high.

We also wanted to compare cars by categories of price. We created a new categorical variable, price_cat, to divide cars into 4 categories according to the following conditions on price: 1 Less than 4,000 2 Greater or equal to 4,000 and less than 5,000 3 Greater or equal to 5,000 and less than 10,000 4 Greater or equal to 10,000

- transformation by sub-group
- continuous to categorical with multiple categories

```
clear
*use auto_raw.dta
sysuse auto



*** Sample Restriction ***




* Drop 5 records with missing data from rep78
codebook rep78
drop if rep78==.
* Drop Datsuns
tab make
drop if word(make,1)=="Datsun"
* Verify resulting number of observations:
count
* 65 observations


*** Correct price data ***

replace price = . if price < 4000
replace price = . if price > 15000
```

```
codebook price
* Verify number of missing: 7.
* Verify mean of corrected price variable: 6,364.73.


*** Variable Creation ***
* First examine variables
codebook mpg weight length price
sum mpg weight length price, de

* Create Log variables

gen lnmpg   =ln(mpg)
gen lnweight=ln(weight)
gen lnlength=ln(length)

* Label log variables
lab var lnmpg "Log MPG"
lab var lnweight "Log Weight"
lab var lnlength "Log Length"

* Create and label ratio variable
gen weightperinch = weight/length
lab var weightperinch "Weight Per Inch"



* Create indicator for high rep78 (4 or 5)
gen highrep = 0
replace highrep = 1 if rep78 >= 4
replace highrep = . if rep78 ==.

*OR in one line with recode
recode rep78 (1 2 3 = 0 "Low Rep") (4 5 = 1 "High Rep"), gen(highrep2)

* verify variable creation
tab rep78 highrep, missing
tab rep78 if highrep ==1

* verify 60-40 distribution
tab highrep, m


* Create Price categorical
gen price_cat=.
replace price_cat=1 if price < 4000
replace price_cat=2 if price >=4000  & price < 5000
replace price_cat=3 if price >=5000  & price < 10000
replace price_cat=4 if price >=10000 & price < .

* Verify variable creation with sum of the different categories.
sum price if price_cat== 1
sum price if price_cat== 2
sum price if price_cat== 3
sum price if price_cat== 4

* Report distribution of categories:
tab1 price_cat

*Save our modifed version of the data
save auto_final.dta, replace
```