Econ 335: Labor Economics

Stata Exercise, due Monday, December 8th

In this exercise, you'll use CPS data to reproduce a few simple findings relating to earnings, and then to design and implement your own analyses. Completed projects should include a do file, a dictionary file, and a write-up in Word including tables and graphs.

1. Find the data

Go to the Current Population Survey home page at http://www.census.gov/cps/
Navigate to the DataWeb FTP page, http://thedataweb.rm.census.gov/ftp/cps_ftp.html
Skip to the March Supplement Section, and find the 2014 installment at the top.
Documentation: ftp://ftp2.census.gov/programs-surveys/cps/techdocs/cpsmar14.pdf
Data: http://thedataweb.rm.census.gov/pub/cps/march/asec2014_pubuse.zip

2. Write a dictionary file

You can create a dictionary file using most simple text editing programs. You can save it e.g. as "dictionary.dct". The content should have the following structure:

```
dictionary using "C:\[path to data file]\asec2014_pubuse.dat"     {
   _column(1)       type       %1f
   _column(19)      age        %2f
   _column(21)      marital    %1f
   _column(24)      sex        %1f
   _column(25)      educ       %2f
[and more variables…]
}
```

Each line in your dictionary file refers to one variable from the data set that you'd like to work with. The first entry on each line indicates the column that the variable begins on. The second entry indicates the name you'd like to give to the variable. (You can give them whatever names you like.) The third entry indicates how many digits the variable has. You can use the documentation to find out the starting column and length of each variable – this starts on page 68 of the pdf.

3. Import into Stata

Use the "infile" command in Stata, as follows:

```
infile using "C:\[path to dictionary file]\dictionary.dct"
```

Now you should be able to see the data you imported using the Data Editor (Browse) feature.

4. Drop household and family records

The first column of the data is 1 if the line is a household record, 2 if it is a family record, and 3 if it is a person record. Let's just use the person records here. I called this first variable "type" in my dictionary (as in the example above), so I would focus on person records using the command

```
drop if type ~= 3
```

This should leave you with 139,415 observations.

5. Generate new variables from existing variables

For example, you can make an "age squared" variable using the command

```
gen AGE2 = age^2
```

You can create an indicator variable (or "dummy variable") for "male" using the command

```
gen MALE = 1 if sex == 1
replace MALE = 0 if sex == 2
```

You can use the education variable to create indicator variables for different education levels using commands like

```
gen CHILD = 0
replace CHILD = 1 if educ == 0
gen SOMECOL = 0
replace SOMECOL = 1 if educ == 40
```

You can lump different ages together using the cut function

```
egen agecat = cut(age), at (25, 30, 35, 40, 45, 50, 55, 60, 65, 70)
```

6. Run regressions

Regress earnings (I've been using the variable starting on column 588, but you can also experiment with other definitions) on age, age squared, and indicators for male, high school graduate (as with the other education indicators, this should defined so as to exclude those who have reached higher levels of education), some college, college degree, graduate degree, white, black, Hispanic, Asian, military experience, born in America, and union. For example:

```
regress totalearn age AGE2 MALE HSGRAD SOMECOL COLLDEGREE GRADDEGREE
WHITE BLACK HISPANIC ASIAN MILITARY BORNINUS UNION if CHILD == 0 &
NOHS == 0
```

Present your results as tables, and discuss each of the coefficients; are they intuitive? Try a few variations on these and see what you get. If you regress earnings on just one of these variables (for example, the indicator for a college degree), how does the coefficient change? Explain.

7. Create graphs

Create a series of age-earnings profiles, i.e. functions with age on the horizontal axis and earnings on the vertical axis. On the first graph, draw age-earnings profiles for each of the following levels of education: some high school, high school degree, some college, associates degree, bachelors degree, masters degree, doctorate, and professional degree. You can repeat this exercise separately for males and for females, as well as for the combined population, if you like. Second, make a graph of the age-earnings profiles of males in general and females in general. Third, make a graph of the age-earnings profiles of different race categories covered in the survey, e.g. white, black, Asian, and Hispanic.

For some of these, the graph will be cleaner if you divide age into categories rather than reporting a separate data point for each precise value of age. (Two possible exceptions are the general male and female age-earnings profiles, because each age value seems to have a decent number of observations.)

There are plenty of ways to graph these in Stata, and you're more than welcome to do that, but I personally preferred to import the information to Excel in table form, and graph it there.

Here is an example of how I graph the profile for males with a doctorate. First, I define a new variable giving the earnings of males with doctorates.

```
gen earn_male_doctorate = totalearn if MALE == 1 & DOCTORATE == 1
```

Second, I create a table of average earnings in this category, according to age group.

```
mean earn_male_doctorate, over(agecat)
```

Third, I highlight the newly created table, use the Copy Table option (regular copy won't work), bring it into Excel, and graph it.

8. Create your own analyses

Now you have a very rich data set at your fingertips, plus a few ideas about what Stata can do with it and how. The next step is to play around with the data on your own, and create new findings. Look at the documentation to learn what other variables you might use – there are quite a lot of possibilities! You can run regressions, create graphs, make comparisons, perform hypothesis tests, or do anything else you like. The goal is to find answers in the data to questions that you or others might be curious about. My default expectation is that the amount of work you produce here is roughly similar to the amount of work in the structured exercise above, but I have no specific quota or limit in mind. By default I would expect you to stick mostly with the March 2014 CPS, but if you feel a strong desire to bring in other years or even other data sets for the purpose of comparison, I suppose that's alright. So the main directive here is to explore and follow your interests.