ECON 20: Lecture #3
Stata Introduction (Answers)

1 Getting started

Your first step is to download the dataset `cpsmar08_10pct.dta` using the link on canvas. Create a new, empty folder on your computer called `stata_intro` and move the file you just downloaded into that folder. Take note of this new folder's "location" on your computer.

Once you have done this, you should launch STATA. The first thing we need to get comfortable with is the idea of a *working directory*. This is the location (i.e., folder) on your computer where STATA will work from. You should tell STATA to change the working directory to your new `stata_intro` folder using the `cd` command:

```
/// EXAMPLE ON MAC ///
cd "~/Dropbox/Econ20/stata_intro/"

/// EXAMPLE ON PC ///
cd "C:\Users\steve\Dropbox\Econ20\stata_intro\"
```

You will need to replace the directory locations after the `cd` to the folder names corresponding to the location of the folder on your own computer.

2 Loading the data

Now that we are working from the correct directory, we can load our dataset into STATA by typing the following into the command line:

```
use cpsmar08_10pct.dta, clear
```

Note: You can always open a dataset by double clicking on the file on your computer. On most systems, this will automatically change the working directory to the location of that file on your computer. You can also open a dataset by going to `File->Open` in STATA, but

I recommend against this practice. You need to get used to the idea of a working directory and doing *everything* by issuing commands in the command line.

Now is a good time to familiarize ourselves with the STATA environment. This will be discussed using a screenshare during class. The two most important panels in the STATA environment are the *Results* and *Command* panels. The command panel is where you type commands in for STATA for perform. The results panel displays the output corresponding to a given command.

3 Exploring the data

With a dataset loaded into STATA, you can view the dataset with the `browse` command. This opens up a spreadsheet-like view of the dataset. Some questions that we might have (or be able to answer) after using the `browse` command:

1. What kind of data are these?
2. Why are the words `male` and `female` colored blue? (*Hint:* Look at the codebook)
3. What do you think it means that some of the values of `wage_hr` seem to be a period?

Answer (1): These are cross-sectional data, one row (or "observation") per person
Answer (2): The blue color means that STATA is displaying what's called a *value label*. By examining the codebook or clicking on an instance of *male* or *female* when browing the data, we can see that the underlying values in the dataset are male = 1 and female = 2. Stata has been "programmed" with a value label to display *Male* when gender= 1 and *Female* when gender = 2.
Answer (3): The periods indicate *missing* values.

Some other useful ways to get some basic facts about the dataset are the `describe` and `summarize` commands. The `describe` command (or `desc` for short)) lists all the variables in the dataset, with information on storage type as well as value and variable labels. The `summarize` (or `sum` for short) provides basic summary statistics for all the variables in the dataset. Try entering the `desc` and `sum` commands into the command line and making sense of the output.

4 Basic operations

Creating variables

You can create new variables using the `generate` (or `gen` for short) command. Start by generate a new variable for the natural log of the wage:

```
gen lnwage = ln(wage_hr)
```

You can see the mean and standard deviation of your new variable (or any variable) in the dataset by issuing the `sum` command followed by the variable name. So if we issue the command `sum lnnwage`, we get the output:

```
. summ lnwage

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      lnwage |      1661    2.849478    .6470944   .2766323   5.441418
```

Regressions

You can estimate a regression using the `reg` command. The syntax for the regression command is `reg Y X`, where Y is the left-hand side variable and X is the right-hand side variable. Note that for multiple regression, we would use `reg Y X1 X2 X3`.

Start by estimating a regression of log wages on years of education using the command `reg lnwage yrsed`. This gives us the output:

```
. reg lnwage yrsed

      Source |       SS       df       MS              Number of obs =    1661
-------------+------------------------------           F(  1,  1659) =  568.58
       Model |  177.419565     1  177.419565           Prob > F      =  0.0000
    Residual |  517.674217  1659  .312039914           R-squared     =  0.2552
-------------+------------------------------           Adj R-squared =  0.2548
       Total |  695.093782  1660  .418731194           Root MSE      =  .55861

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
       yrsed |   .1177832   .0049396    23.84   0.000     .1080948    .1274717
       _cons |   1.214163   .0699375    17.36   0.000     1.076988    1.351338
------------------------------------------------------------------------------
```

Notice the key elements. The intercept ($\hat{\beta}_0$) is 1.214163. It has standard error .0699375. The slope ($\hat{\beta}_1$) is .1177832, with standard error .0049396.

A few questions for you:

1. Is the slope statistically significant?

   Answer: Yes, highly significant. You can see this from the very high t-statistic ($t = 23.84$) and correspondingly very low p-value ($p < 0.001$). We will learn more about hypothesis testing soon.

2. How do you interpret the slope?

   Answer: Generally, we would interpret the slope as a one unit increase in x is associated with a 0.1178 unit change in y. Taking into account the units here, and paying attention to the fact that $y = \log(wage)$, our interpretation should be that a one additional year of education is associated with 11.77% higher wages.

3. What is the R^2? Briefly interpret this number.

   Answer: $R^2 = 0.2552$. We interpret this as saying that about 25% of the variation in wages is explaned by education. This means other things besides education explain a lot of the variation in wages – about 75%.

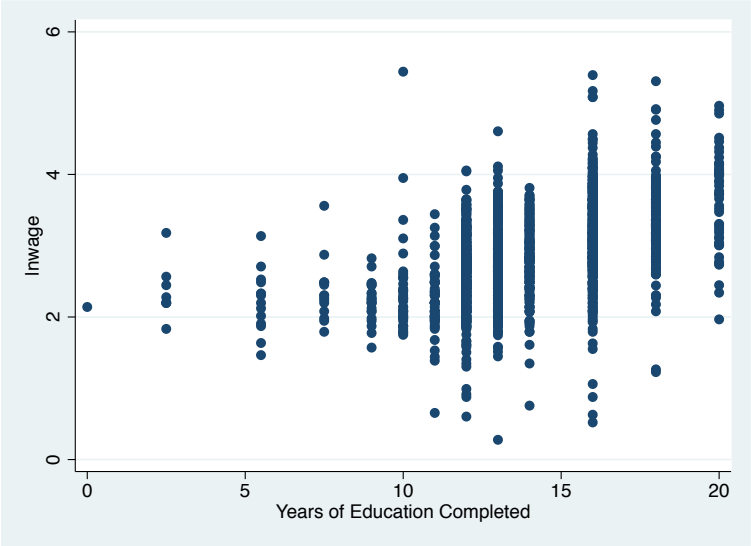
4. What is the MSE?

   Answer: MSE = 0.312. We can get this by squaring the root MSE which is reported in the regresion output above, $0.55861^2 = 0.312$. The root MSE is the standard deviation of the residuals and the MSE is the variance of the residuals.

Plotting relationships

When examining the relationship between two variables (such as education and wages), it is often a good idea to "visualize" the relationship by plotting the data. There are many benefits to doing so. For example, a plot will help us answer the questions: (a) Is the relationship between education and wages linear? (b) Is the regression estimate from above driven by outliers in the data?

The simplest way is `scatter lnwage educ`, which will plot each data point i, with observation i's value of education on the x-axis and observation i's value of lnwage on the y-axis. The graph looks like this:

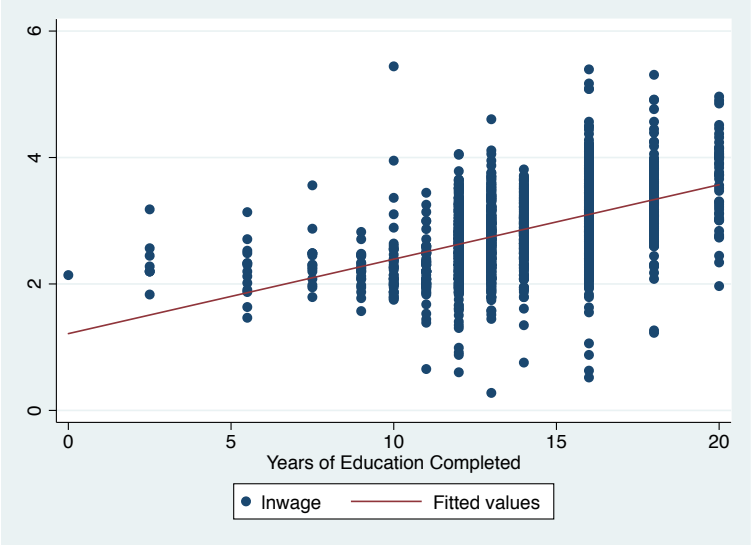


After creating a plot in STATA, you can save the plot as picture file on your computer using `graph export wagescatter1.png, replace`, which will store the graph in your working directory as `.png` image file that could then be inserted into, say, a word document.

We can also add the regression line from above to our scatter plot using the command:

```
graph twoway (scatter lnwage yrsed) (lfit lnwage yrsed)
```

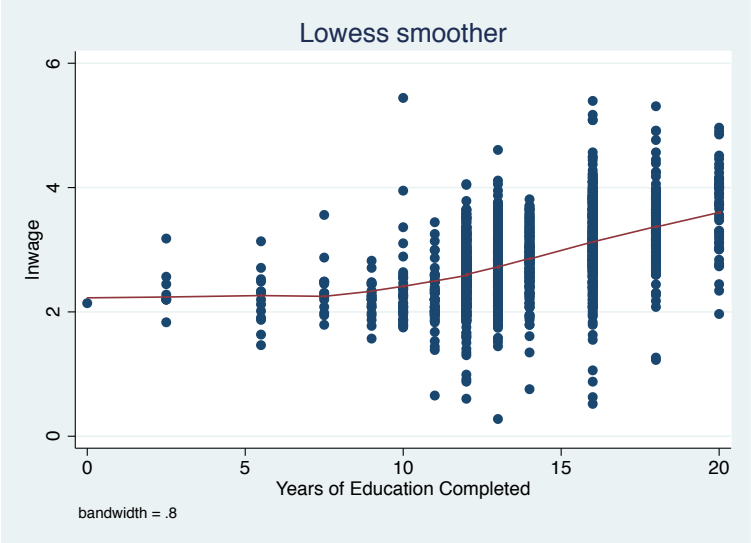
The two expressions in parentheses correspond to two different plots that will be shown on the same graph. The first part is our same scatter plot from above. The second part (`lfit`) will add the linear fitted line (same as our regression from above) to the graph:



A useful alternative to the above graph is to fit a general, nonlinear relationship between the two variables (this is called a *local linear regression*, you do not need to understand what this is right now). To get this more general fit, we can use:

```
lowess lnwage yrsed
```

This command generates the following graph:



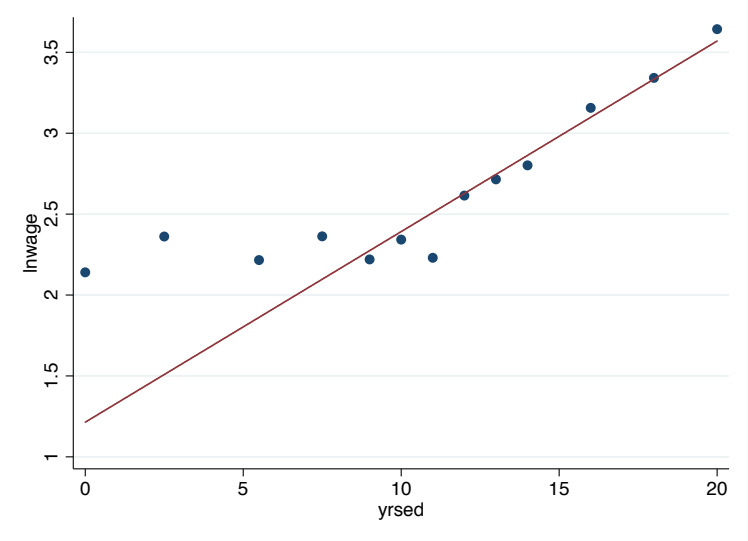
Once we have the nonlinear fit, does the relationship between years of education and ln(wages) look linear?

Answer: No, it appears to be convex. The wage gain associated with a one-year increase in education is lower at low levels of education than at higher levels of education.

Another very useful graphing command in STATA is called a binned scatter. The `binscatter` command chops the x variable up into bins (i.e., 5-6 years of education) and then plots the average value of y in that bin. In mathematical terms, we can think of `binscatter` as an empirical approximation of the conditional expectation function $E[y|x]$. The `binscatter` command is not natively installed in STATA, we first need to install it. To create a binned scatter plot, use the commands:

```
ssc install binscatter
binscatter lnwage yrsed
```

The first line installs the `binscatter` program. Once you have done this once, you never need to do it again. The command creates the following plot:



5 Stata programming

Entering calls into the command line one-by-one and examing output in the results panel is called using STATA *interactively*. This is a useful way to open STATA for the first time and get a sense of a few basic commands, but is a bad practice more generally. Going forward, we will instead use *programs* to analyze data. A STATA program, called a ".do file," is really just a list of commands that will be passed to STATA line-by-line all at once. To create a new do-file, you can click the labeled `do-file editor` at the top of your STATA window.

There is a template do-file on canvas that you should download. Your do-file should always begin with the following three lines:

```
clear all
cap log close
set more off
```

The next step in every do-file is to initiate a "log" file. A log file prints all the output of your do-file into a text file that you can read after your program has finished running. A good practice is to name your log file the same thing as the your do-file, so that you can always keep track of which log file corresponds to which STATA program.

```
log using "LOGFILENAME.log", replace
```

After initiating your log file, you enter STATA commands line-by-line into your do-file. A simple do-file that only performs the `describe` and `summarize` commands would look like this:

```
clear all
cap log close
set more off

log using "LOGFILENAME.log", replace

describe
summarize

cap log close
```

The last line, `cap log close`, tells STATA to stop writing in the log-file. To execute a do-file, type `do DOFILENAME` into the commmand line.

Assigment for you: Download the template do-file and save it in the folder you created in the first step above. Rename the do-file as `stata_intro.do`. Modify the do-file so that it creates a log file called `stata_intro.log`. Then modify the do-file so that it executes the following commands:

1. Create the new variable `lnwage`.
2. Estimate the regression of `lnwage` on `yrsed`.
3. Construct a simple scatter plot fo `lnwage` against `yrsed`.
4. Store that plot as `wagescatter.png`.

Run the do-file and make sure it works. Examine the log-file and make sure it looks how you expect it to look. Make sure there is a new image file called `wagescatter.png` stored on your computer.

6 Additional exercises

Predicted values and residuals

Immediately after any regression command (`reg`), you can construct new variables containing either the predicted values, $\hat{y}$, or residuals, $\hat{u}$.

1. Add the command `predict yhat` to your program after `reg lnwage yrsed`. This constructs the predicted values. Take note that, in this case, `predict yhat` is equivalent in this case to using `gen yhat = 1.214163 + .1177*yrsed`. You can verify that these are equivalent directly in STATA if you want to get creative.

2. Construct the residuals by adding the command `predict uhat, resid`. Take note again that, in this specific case, this is equivalent to using the command
   `gen uhat = lnwage  (1.214163 + -.1177* yrsed)`.

3. Now regress `uhat` on `yrsed`. Is there a statistically significant relationship between the residuals and years of education. Does this speak to whether ther relationship estimated via `reg lnwage yrsed` is causal (in light of assumpion SLR.4)? Why or why not?

   Answer: I find little evidence of a relationship: the coefficient is tiny and statistically insignificant. In principle, it should be exactly zero (and it would be if it were not for the fact that STATAs calculations are not infinitely precise!) The reason for this is that residuals from OLS regression are DEFINED to be uncorrelated with the error. In essence, the line is fit to the data by imposing that any residual variation be uncorrelated with the independent (X) variable in this case, education. (If this were not the case, the line would not appear to go through the points in the right place.) This means that assumption SLR.4, which says the unobservable errors are uncorrelated with the X variable, cannot be tested by examining the residuals.

4. Explain in words (no symbols) what needs to be true in order for the estimated slope to provide a causal estimate?

   Answer: Heres a sufficient answer: In words, we need that third factors – besides education – which affect wages to be uncorrelated with education. Heres a longer

explanation: This assumption is necessary because our regression compares the earnings of people with more education to those with less. In order to attribute all of the observed difference in earnings to the education differences, we need to assume that the people with more education do not systematically have other unmeasured attributes which also raise their earnings. Third factors and what I call here unmeasured attributes are what in our statistical model we call the "error term."

5. Now estimate the reverse regression, that is regress `yrsed` on `lnwage`. Is the slope just the reciprocal of the "forward" regresion taht you already ran?

   Answer: Not even close, $1/0.1177 = 8.5 \neq 2.17$. Notice that there is no reason to expect them to be reciprocals given our expressions for the bivariate regression coeffients using covariances and variances.

6. How does the R^2 compare to the forward regression?

   Answer: They are the same (0.2552) and this is not an accident. In a bivariate regression, the R^2 is just the squared correlation between X and Y. That correlation does not depend on which variable is labeled X and which is labeled Y, the formula is symmetric: $corr(x, y) = cov(x, y)^2/(var(x)var(y))$.