# DATA EXERCISE #1

Due by 11:00pm ET on Friday, April 16

*Instructions:* Turn in your answers to all questions as a single PDF file. Copy and paste your entire STATA log file at the very end of your answer document. If a particular question is only asking you to, say, generate a new variable, you can skip writing an answer to that question. If a question is asking about the output of a regression, output of a formal test, or graphical output, you should copy the relevant STATA output as part of your answer to that question. *Remember to complete the Pre-Lab before beginning the assignment*

## PART I. Understanding alternative right-hand side variables

1. The template program creates a new dummy variable called *momhs* which indicates whether an individual's mother completed high school. Compute the mean of *momhs* and interpret that mean in words.

2. Regress *lnwage* on *momhs.* Interpret the coefficient on *momhs.*

3. There is another way we could have computed the coefficient on *momhs* in the regression above. What is it? Explain briefly in words and then perform the steps in STATA in you need to perform this calculation.

4. Regress *lnwage* on *educ* and *momhs*. Interpret the coefficient on *momhs*.

5. Generate a new variable for the *interaction* between an individual's education and whether the individual's mother completed high school using the code *gen educ_x_momhs = momhs * educ*. Regress *lnwage* on *educ, momhs,* and *educ_x_momhs*. Interpret the coefficient on *educ_x_momhs* in the above regression. In 1-2 sentences, explain the "story" this regression is telling us.

6. In labor economics, researchers often want to "control for" work experience when estimating regressions with wages as the left-hand side variable. However, in many datasets, actual work experience is not measured. Instead, we use a formula for *potential experience = age – years of education – 6*. Use this formula to generate a new variable called *potexp.*

7. Graph the relationship between *lnwage* (which should be on the vertical axis) and *potexp*. Does the relationship look linear or nonlinear? *(hint:* you may find the *lowess or binscatter* commands we discussed last week useful).

8. Construct a new variable called *potexp2* which equals *potexp* squared. Regress *lnwage* on *educ, potexp*, and *potexp2*. What can we learn by examining the coefficient and standard error on *potexp* and *potexp2* in this regression? Your answer should be in words and 2-3 sentences long.

**PART II. Controlling for regional variation when estimating the returns to schooling**

1. The variable *region14* indicates and individual's region of residence at age 14 and takes four possible values: *1=Northeast; 2=Midwest; 3=South; 4=West.* Construct dummy variables for each region. When you are trying to construct dummy variables for multiple categories like this, a useful shortcut in is the "tab, gen" command, with syntax:

   **`tab existing variable with categories, gen(prefix for new variables)`**

   So here, we can use the command *tab region14, gen(reg)* which will create four new variables *regi* for *i*=1,…,4 and *regi* is a dummy variable *region14=i*.

2. One drawback of the above command is that the names are not very informative. For interpreting regressions, it might be useful to rename them using the codes above: *rename reg1 northeast; rename reg2 midwest; rename reg3 south; rename reg4 west.*

3. Regress *lnwage* on *midwest, south, west*. Interpret the coefficients on the three region dummy variables.

4. Regress *lnwage* on *educ.* Interpret the coefficient on *educ.*

5. Regress *lnwage* on *educ* and the *midwest, south,* and *west* dummy variables. Interpret the coefficient on *educ.*

6. Compare the coefficient on *educ* in the regressions from part 4 and part 5. If we are interested in the effect of education on wages, were the region dummies *omitted* from the regression in part 4. State yes or no and briefly explain why.

7. Interpret the difference in the coefficient on *educ* in the "short" and "long" regressions in part 4 and part 5. What specifically can we infer from how the coefficient on *educ* changes once we include the region dummies in the regression? Your answer should be in words, tell a story, and be 2-3 sentences long.

8. Regress *lnwage* on *midwest, south, and west.* Store the residuals from this regression as a new variable called *rwage* (hint: *predict rwage, resid)*.

9. Regress *educ* on *midwest, south,* and *west.* Store the residuals from this regression as a new variable called *reduc*.

10. Regress *rwage* on *reduc*. Compare the coefficient on *reduc* in this regression to the coefficient on *educ* in part 4 above. Are you surprised by your answer? Why or why not?

**PART III. Non-pecuniary returns to schooling**

1. Labor economists often argue that education impacts many facets of individuals' lives beyond labor market earnings. In this question, we briefly examine the relationship between education and marital status. As a first step, explore the value labels of the variable *marstatus* (*hint:* browse the dataset) and create a new dummy variable for whether the individual is married. Call this new variable *married.*

2. Regress *married* on *educ*. Interpret the coefficient on *educ*.

3. We know that we are looking at a sample of relatively young (age 24-34) individuals. Hence, we may expect that age is highly correlated with marital status in this dataset. Regress *married* on *educ* and *age*. Interpret the coefficient on *age* in this regression.

4. What can you learn by comparing the coefficient on *educ* in the regressions in part 2 and part 3 above?

5. Suppose I told you that the "true" causal effect of education on the likelihood of marriage is positive but relatively small. Explain, as precisely as you can (3-4 sentences), two possible reasons that we may be finding a negative coefficient on *educ* in our regressions?

**PART IV. Sheepskin Effects**

1. Another theory in labor economics is that the returns to education we observe in datasets such as the CPS or the NLS represent so-called "sheepskin effects." Specifically, employers reward individuals who complete degrees with higher wages, but do not necessarily pay higher wages for one additional year of education. In this question, we consider the labor market return to completing college. Your first step is to construct a new dummy variable called *college* for whether an individual graduated from college, which would be measured in our dataset as having completed at least 16 years of education.

2. If we are interested in the causal effect of college completion on wages, we might be interested in a "balance" test. Briefly describe the purpose of a balance test.

3. Regress *college* on all the following variables: *age, married, iq, motheduc, fatheduc, momdad14, midwest, south,* and *west.*

4. Test the joint significance of the individual characteristic variables (we are going to cover hypothesis testing more on Thursday) in the regression from part 3 using the command: *test age married iq.* Briefly explain in words how to interpret this test.

5. Test the joint significance of the family background variables (*motheduc, fatheduc, momdad14)* in the regression from part 3. Briefly interpret the test.

6. Test the joint significance of the region dummies (*midwest, south, west)* in the regression from part 3. Briefly interpret the test.

7. Briefly comment on how your answers to parts 4-6 relate to your answer to part 2.

8. Suppose you estimate the effect of college completion on wages by regressing *lnwage* on *college* plus all the other variables included in the regression in part 3. What assumption is needed for the coefficient on *college* to provide a causal estimate? Explain in words and, ideally, explain how your answer relates to a "selection on observables" argument.

---

A few more tips on STATA:

**In programs, don't break commands across lines** (except see below).

STATA treats each line of the program as a separate command, and if you try to wrap a command across multiple lines of the program it will get confused – it'll try to run the second part of your command as a separate command. You do sometimes have very long commands that really need to be split across multiple lines (e.g., graph commands where you specify a lot of options about how the graph looks.) STATA has a way to deal with this:

**To break up a long command across lines, put a "///" at the end of the line to tell STATA that the command continues**. (Note: there needs to be a space before the ///.)

You can try an inane example to see that this works, like

**summ** ///
**wage** ///
**yrsed** ///
**female**

**Help**

Help is a very useful command in STATA. Typing "help *command*" gives detailed information about the proper syntax for *command*.

**More commands**

There are some nice "cheat sheets" at http://geocenter.github.io/StataTraining/portfolio/01_resource/. See also the reference section of the library for full STATA manuals.

**A FINAL WORD: EXITING STATA**

When you want to exit STATA, all you need to do is type

**exit, clear**

On the command line. If you try to exit by hitting "Quit" in the file menu, STATA will probably ask you if you want to save changes to your data. The answer is probably "No." The reason is that if you overwrite the original data with the modified data (i.e. where you added lnwage etc.) your STATA program may not run anymore -- it will crash when it comes to the line where you create lnwage because you will already have lnwage in the dataset. If you want to save your data, save it under a different name from the original.

Also, **don't forget to save your program!!!** STATA will prompt you for that, too. In this case, you probably DO want to save changes.