Data Exercise #2
Due by 11:00pm ET on Friday, May 7

Instructions: Turn in your answers to all questions as a single PDF file. You should copy and paste your entire STATA log file at the very end of your answer document. If a particular question is asking for a regression or figure, include the corresponding output in your answer to that particular question. There is no pre-lab for this data exercise.

Summary: You will use data to replicate some of the results from Ashenfelter and Krueger (1994), "Estimates of the Economic Returns to Schooling from a New Sample of Twins." *American Economic Review* 84(5): pp. 1157-73. If you are interested, you could read this study; a summary of the relevant issues appears in chapter 6 of the Angrist and Pischke textbook. Ashenfelter and Kreuger, two Princeton economists, descended on Twinsburg, Ohio's annual Twin Festival. They collected data on the education levels and earnings of a large sample of twins to see if they could use a panel of twin families to get rid of bias in the OLS estimates of the returns to schooling.

Data Source: pubtwins.dta
To do these exercises, you will use the Stata dataset pubtwins.dta. You can download the dataset here or download it from the files/data section on Canvas. You will have to create additional variables from pubtwins.dta to conduct the analysis.

Data notes:

- The data set contains 680 observations on individuals ordered by twin pairs. There are therefore 340 twin pairs, with the first two observations representing the first twin pair, the next two observations representing the second twin pair, etc.
- Key variables:

  hrwage = the self-reported hourly wage of the individual (in dollars)
  lwage = the natural log (ln) of the hourly wage
  age and age2 = the age of an individual and its square (age2)
  female = an indicator variable equal to one if the person is female, zero otherwise
  white = an indicator variable equal to one if the person is white, zero otherwise
  educ = the educational attainment of the individual
  educ_t = the other twin's report of the individual's education
  first = an indicator equal to one if the twin was the first-born (equal to 0. otherwise)
  dlwage = the difference in the log wages of twins
  deduc = the difference in twins' education based on their self-reports
  deduct = the difference in twins' education based on each twin's report of the other twin's education

Note: As always, you should do this problem set by writing a program (a *.do file in Stata). Here is a template program (also available on Canvas in the files/data folder). In your solution packet, include a well-annotated log file program, as well as relevant STATA output. And for this problem set, **I will also require you also to turn in an "outreg2" table** (worth 0.5 points of your problem set). Outreg2 is a STATA package for producing publication-quality regression tables in STATA

that can be exported to excel and then copied into a word document. Here is a handout on using outreg2 in STATA. You can also find this handout on Canvas in the files/project/handout section.

To use outreg2 in STATA, you must first install it. Type:
`ssc install outreg2`
on STATA's command line to get the command to work (you only ever have to do this once). I have put the first two outreg2 commands into the template program to help get you started. I recommend reading the handout above and "tinkering" with the command to figure out how it works.

1. Cross-section regressions

a. Run the bivariate regression of log wages on a constant and education and show the scatter plot. The outreg command could be something like:

   outreg2 using table1, excel ctitle(Ques 1a,ln wg-OLS) replace

   Notice that "ctitle" adds a column header to the table1 file that is sent to your folder, with the comma indicating the title will be split across two rows. Note that on macs, you will have to open this file from Excel rather than double clicking.

b. Regress log wages on a constant, education, age, age2, and white and female. Compare the estimated return to education from the multivariate regression to the one from the bivariate regression in (a). **Are the returns now different?**

   This time, the outreg2 command should be something like:

   outreg2 using table1, excel ctitle(Ques 1b,ln wg-OLS) append

   where now we "append" a column instead of replacing.

c. Regress education on the other controls in (b). Add a column to your outreg table (table1) that makes clear that the dependent variable is now "educ" instead of ln(wage).

   Is education significantly related to these other controls?

   Here's a suggestion: after your regression and before your outreg command, add an F-test command "test age age2 white female". Then you can add your F-stat to the outreg table in the bottom row with the option…

   outreg2 …. using table1, excel … append addstat(F-stat,r(F),p-value,r(p))

   where r(F) and r(p) are the stored values of the F-stat and p-value after a test command (on some computers this will only work if you put them in the `' the weird STATA quotes (used also for loops, i.e., `r(F)').

   Can you think of variables that we have not controlled for that may be related to both

educational attainment and earnings? What does this imply about how we should interpret the ordinary least squares estimate of the relationship between log wages and education? (Hint: use the omitted variables bias formula; if this were an exam question, your answer would be judged on your logic, not whether you are "right")

Make sure all of these regressions include the robust standard error correction for heteroskedasticity!

2. Measurement Error in the Cross Section

Note: This question presumes your understanding of CE #2, answers to which will be posted on Tuesday night.

a. Suppose that a twin's self-report of education is an imperfect measure of the twin's actual educational attainment due to misreporting. In addition, suppose that this measurement error is "classical" (or, loosely speaking, "just noise" uncorrelated with anything.) In this case you have a second (independent) report of years of education: each twin was asked about the education of the other twin. This is recorded in the "educt_t" variable. (Careful: there is also a variable called "educ_t which is the other twin's actual education. You want the one with two "t"s)**. Regress educt_t on educ and add this to the table. What would be the coefficient estimate if there were no measurement error? What does this slope coefficient instead represent?**

b. If you have two independent reports, you can also correct estimates for measurement error using instrumental variables regressions. So now run STATA commands estimating by instrumental variables the effect of education on wages, using "educt_t" as an instrument for "educ," i.e., using the other twin's report of the individual's education level as an instrument for the individual's self-reported education. Do this with and without controls. **How do these estimates compare to the ones you got in 1(a) and 1(b). Are the results roughly consistent with your estimates of the attenuation factor (the lambda from CE #2 is sometimes known as the attenuation factor)?**[1]

Note: the remainder of this Problem Set is based on material that we will cover in the second half of class on Tuesday and in class on Thursday.

[1] Note: the two would only be exactly the same if you instead used educ as an instrument for educt_t, since educt_t and educ may have different amounts of measurement error.

3. Panel techniques, measurement error in differences, and omitted variables bias

a. Now suppose that there is a common unobserved factor between pairs of twins related to their wages and potentially also to their educational attainment. **Write down a model which describes this possibility.**

b. Now suppose all omitted factors are held constant when comparing identical twins. Run the regression of the difference in log wages between twins on the difference in educational attainment using the STATA command [reg dlwage deduc if first==1, noconstant] (and, as always, add it to your table with appropriate headers). **How does this estimate of the return to education compare to the one based on the regression of lwage on educ, age, age2, female, white in, say, part 1(b)? Explain what this might imply about the omitted variables bias.**

c. Now estimate with fixed effects by "absorbing" family with the "areg" command. **How does this compare to your answer in 4(b)? Why?**

   For this question, it will be helpful to add a row to the bottom of the table indicating that you are now controlling for family fixed effects. For example, can do this in outreg with the "addtext" option, such as:

   outreg2 using table1, … addtext(Family Effects?, Yes)

d. For another possible answer to 4(b) run the STATA command ivreg dlwage (deduc = deduct) if first==1. This two-stage least squares regression uses the twin difference in education as reported by the other twin as an instrument for deduc. **How does this compare to 4(b)? Why do you think this is? Now what do you think of the omitted variables bias in the conventional OLS estimate of the returns to education (i.e., the estimate from regressing lwage on educ, age, age2, female, white)?**

e. Notice that in this part of the question I never asked you to include any of the other controls in the regression. For example, I did not ask you to include "age" or "daded" (father's education" or any other family background controls. **Why is this? Do you think that comparing twin pair differences across families reduces omitted variables problem?**

4. Standard Errors

There are multiple (two!) observations on people in each family in these data. You might think that such people have common unobservables which make their error terms correlated. As such, standard errors imposing that they are independent (the OLS standard errors) will be biased. The variable "family" identifies the family number. Returning to part 1, run the cross-sectional regression of log wages on education, age, age2, female, and white applying a standard error calculation that accounts for error correlation between twins. Compare the

standard error and t-ratio on education in this regression with the standard error and t-ratio on education in the regression without the clustered standard error. **Explain why the standard errors on the estimated return to education are higher (and t-ratio lower) than when clustering is not corrected for.**