

**ECON 244**

**APPLIED EXERCISE #1**

This exercise examines regression analysis of the returns to education using twins data. Feel free to work cooperatively and in groups. However, each person must hand in his/her own problem set using his/her own words and interpretation of the results. **Summarize the results/answers to each question concisely and not by referencing a morass of STATA output.**

We will analyze the following STATA data set:

Data Source: pubtwins.dta

This data extract is from a survey of twins conducted at the Twins Day Festival in Twinsburg, Ohio (see Ashenfelter and Krueger 1994). The observational unit of the data is the individual. For sets of twins, information was collected on the earnings, education, age, race, gender, etc. of individuals.

Data Notes:

1. Link to data via command: `ln -s /class/e244/e244/data/pubtwins.dta pubtwins.dta`
2. The data set has 680 observations ordered by twin pairs. That is, there are 340 twin pairs in the data, and the data is sorted by these pairs. So, the first two observations are the first twin pair, the next two observations are the second twin pair, etc.
3. The key variables are:
  - `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 a person and its square ( $age^2$ )
  - `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
  - `educt_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 . 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
- a. Run the bivariate regression of log-wages on a constant and education and show the scatter plot. Now regress log-wages on a constant, education, age, age-squared, and the gender and racial indicators. Briefly interpret the "economic meaning" of each slope coefficient. What do the coefficients on age and age-squared imply about the life-cycle profile of earnings? Would including just a linear term for age lead to a more appropriate regression model? Explain. Now add  $age^3$  and  $age^4$  to the regression. Does this substantially improve the fit of the regression model?
- b. Compare the estimated return to education to the one from the bivariate regression model. Are they different? What might this imply about how education is distributed across the population? Now compare the mean characteristics of individuals with a college degree (`educ=16`) to individuals with

just a high school degree ( $\text{educ}=12$ ). 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 least squares estimate of the relation between log-wages and education?

- c. Now create dummy variables for each of the eleven levels of schooling (8-18). Regress both wages and log-wages on just the dummy variables. Is the effect of education on wages linear in education? How about its effect on log-wages? Focusing on log-wages, describe where the “nonlinearities” are, if any. Now run the dummy variable regression for log-wages including age, age-squared, gender, and race as controls. Does allowing for nonlinearities in the return to education improve the fit of the regression model substantially?
- d. Based on the scatter plot of hourly wages on the y-axis and education on the x-axis, is there any evidence on homoskedasticity/heteroskedasticity in the wage regression model? What about with log-wages on the y-axis?
- e. Regress log-wages on education, age,  $\text{age}^2$ , and the gender and racial indicators, using the “robust” subcommand in STATA to calculate the Eicker-White consistent standard errors. Explain briefly how these estimates of the standard errors are corrected for heteroskedasticity. How do they compare to the “uncorrected” (conventional) least squares estimates of the standard errors. Is there any evidence of heteroskedasticity?
- f. Using the “predict” STATA command [`predict (var. name), residual`], save the residuals from both the wage and log-wage regressions. Now regress the squared values of the residuals from the two sets of regressions on education, age,  $\text{age}^2$ , female, and white. From the R-squareds of these regressions, test for heteroskedasticity in the two sets of residuals. Does one set of residuals appear to be more heteroskedastic than the other? Now regress the squared residuals on education,  $\text{education}^2$ , age,  $\text{age}^2$ , female, white, and the interactions  $\text{education*age}$ ,  $\text{female*age}$ ,  $\text{female*education}$ ,  $\text{white*age}$ , and  $\text{white*education}$ . Again, test for heteroskedasticity based on the R-squareds of the regressions.
- g. Explain how the assumption that the residuals from the log-wage regression are “pairwise” uncorrelated may be violated when using the twins data. Use the following STATA commands to create a variable that separately identifies each twin pair in the data set (Note: the data must be in its original order for this to work):

```
. gen id=_n
. replace id=id/2
. replace id=round(id,1)
```

Run the regression of log-wages on education, age,  $\text{age}^2$ , female, and white using the “cluster” STATA subcommand to correct the estimated standard errors for correlation in the residuals between twins. Explain why the standard errors on the estimated return to education are higher (and t-ratio lower) than when clustering is not corrected for.
- h. Now run the regression of the average of the log-wages of each twin pair on each twin pair’s average education (i.e., you now have 340 twin pair observations based on twin averages). Does this correct the “clustering” problem in the residuals? Explain briefly.
- i. 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” in the sense that it is independently and identically distributed. What is the formula for the bias in the estimated return to education from the regression of log-wages on educ, age,  $\text{age}^2$ , white, female in terms of the “noise-to-total variance ratio”?

- j. Now run the following STATA command `[ivreg lwage age age2 female white (educ = educt_t)]`. This performs two-stage least squares estimation of the return to education using the other twin's report of the individual's education level as an instrument for the individual's self-reported education (for each individual, both the individual and his twin were asked about the individual's education level). Explain why the estimated return to education from this procedure is greater than the estimated return from standard OLS. Calculate the reliability ratio of the education data under the assumption that the measurement errors are classically distributed.
- k. Suppose there are unmeasured factors that are associated with both an individual's educational attainment and an individual's earnings (e.g., innate ability, family background, school quality). Explain how this could lead to "omitted variables" bias in the least squares estimate of the return to education. Suppose that the omitted variable is  $A_i$ . Write out the "omitted variables bias" in terms of the linear relationships between education and  $A$  and log-wages and  $A$ .
- l. Now suppose that 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]`. 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`? Explain what this might imply about the omitted variables bias.
- m. Graph a scatter plot with the difference in twins' log-wages (`dlwage`) on the y-axis and the difference in twins' self-reported education (`deduc`) on the x-axis only using the first-born twin's observations (`first=1`). Where are most of the observations clustered with respect to the x-axis of `deduc`? What could this imply about the importance of measurement error in this variable? How might this be another explanation for the result you found in part (l) that the estimated return to education is lower when running the "first-differences" regression? Explain how first-differencing the data may exacerbate the measurement error problem.
- n. Now run the STATA command `[ivreg dlwage (deduc = deduct) if first==1, noconstant]`. This two-stage least squares regression uses `deduct` as an instrument for `deduc`. Explain why the estimated return to education is now larger than the one in part (l). How does the unbiasedness of this estimate depend on the classical measurement error assumption? Will it be unbiased if the measurement errors between an individual's self-report of education and his twin's report of the individual's education are correlated? Describe a solution to this problem.
- o. Suppose that the classical measurement error assumption holds, what might one conclude about the size 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`)? Do you think that comparing twin pair differences across families reduces the omitted variables problem? Explain.