

**ECON 142**

**APPLIED EXERCISE #2**

*Due October 20, In Class*

Although the paper is technical, it will be helpful to read the Ashenfelter and Krueger (1994) paper in the reader as you do this problem set. This exercise examines violations of the assumptions underlying the Gauss-Markov theorem for the BLUE properties of the least squares estimator of the linear regression model. Again, feel free to work cooperatively and in groups. Each person must hand in his/her own problem set using his/her own words and interpretation of the results.

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 /user8/e142/data/pubtwins.dta` .
2. The original data set has 680 observations ordered by twin pairs. That is, there are 340 twin pairs in the data, and the data are sorted by these pairs. So, the first two observations are the first twin pair, the next two observations are the second twin pair, etc. Note that this is true of the original data (and won't be true if you have changed the ordering of the data and saved over the original data set).
3. For this problem set, 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 ( $\text{age}^2$ )
  - 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 between twins in their log-wages
  - 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

Question 1: Heteroskedasticity, Twins Correlation

- a. Based on the scatter plot of hourly wages on the y-axis and education on the x-axis, is there any evidence of homoskedasticity/heteroskedasticity in the wage regression model? What about with log-wages on the y-axis? (you need not attach the graphs)
- b. Suppose that there is heteroskedasticity in the residuals of the log-wage regression. Is the ordinary least squares (OLS) estimator of the returns to education unbiased and consistent? Efficient? Is the "conventional" estimator of the variance of the estimated return to education unbiased/consistent?

- c. Regress log-wages on education, age, age<sup>2</sup>, and the gender and racial indicators, using the “robust” subcommand in STATA to calculate the Eicker-White consistent standard errors [`reg lwage educ age age2 female white, robust`]. 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 from the same regression, but not including the “robust” subcommand. From this comparison, is there any evidence of heteroskedasticity?
- d. Using the “predict” STATA command [`predict (var. name), residual`], save the residuals from the log-wage regression. Now regress these residuals on the regressors (educ, age, age2, female, white). Explain why the R-squared and estimated coefficients from this regression are virtually zero.
- e. Now regress the squared values of the residuals from the two different regressions on education, age, age<sup>2</sup>, female, and white. From the R-squareds of these auxiliary 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, education<sup>2</sup>, age, age<sup>2</sup>, female, white, and the interactions education\*age, female\*age, female\*education, white\*age, and white\*education. Again, test for heteroskedasticity based on the R-squareds of the regressions.
- f. Explain how the assumption that the residuals from the log-wage regression are “pairwise” uncorrelated may be violated when using the twins data. What effect does this have on the properties of the OLS estimator of the returns to education and the estimated variance of the estimated return to education?
- g. Now 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, age<sup>2</sup>, female, and white using the “cluster” STATA subcommand to correct the estimated standard errors for correlation in the residuals between twins [`reg lwage educ age age2 female white, cluster(id)`]. Explain why the standard error on the estimated return to education is higher (and t-ratio lower) than when clustering is not corrected for.

### Question 2: Measurement Error and “Omitted Variables” Bias

- 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” 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, age2, white, female in terms of the “noise-to-total variance ratio”?
- b. Now run the following STATA command [`ivreg lwage age age2 female white (educ = educt_t)`]. This performs two-stage least squares estimation (a form of instrumental variables 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 a given 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.

- c. Suppose we are worried that there is an unmeasured factor that determines 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 this omitted variable is  $Z_i$  (with  $Z$  = vector of observations). Write out this "omitted variables bias" in terms of the linear relationships between education and  $Z$  and log-wages and  $Z$ .
- d. Now, suppose that these 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. Why didn't we include controls for age, gender, and race in the "first-differences" regression?
- e. 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 d. 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.
- f. 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 d. How does the unbiasedness of this estimate depend on the classical measurement error assumption? Will it be unbiased if the measurement error in an individual's self-report of his education is correlated with the measurement error in his twin's report of the individual's education?
- g. 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 using variation in twin pair differences in education and earnings to estimate the return to education reduces the omitted variables problem? Explain.