Applied Exercise #2

– Due Tuesday 03/11/07 –

This exercise examines violations of the assumptions underlying the Gauss-Markov theorem for the properties of the least squares estimator of the linear regression model. It will be helpful to read the paper by Ashenfelter and Krueger 1994 in the reader as you do this problem set. Feel free to work in groups. Each student must hand in their own solution to the problem set using own words to explain and interpret the results.

Data: The data is an extract form a survey conducted at the Twins Day Festival in Twinsburg, Ohio (see Ashenfelter and Krueger 1994). For sets of twins information was collected on earnings, education, age race, gender, etc.

Access the data on bspace site for Econ C142: problem set 2/ pubtwins.dta (STATA 9 format). The observational unit of the data is the individual. The 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 tow 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 any more if you have changed the ordering of the data and saved over the original data set).

The key variables for this exercise are:
- hrwage – self reported hourly wage of the individual (in dollars)
- lwage – natural log of hourly wage
- age and age2 – the age of a person and its square
- female – an indicator variable equal 1 if the person is female, zero otherwise
- white – indicator variable equal to 1 if the person is white, zero otherwise
- educ – educational attainment of the individual
- educt_t – the other twin’s report of the individual’s education
- first – indicator equal to one if the twin is first born, equal to . otherwise
- dlwage – difference between twins in their log-wages
Question 1: Heteroskedasticity and Twin 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? How does the scatter and the impression of homoskedasticity/heteroskedasticity change if you plot log-wages on the y-axis?

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 returns to education unbiased / consistent?

c) Regress log-wages on education, age, age2, female, white, using the “robust” option in STATA to calculate corrected 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 standard errors, which you get from the same regression without the robust command? From this comparison, is there any evidence of heteroskedasticity?

d) Regress (1) wage and (2) log-wage on education, age, age2, female, white as regressors. Use the “predict” STATA command [predict var_name, residual] and save the residuals from the two regressions. Now regress each of these residuals on the regressors (education, age, age2, female, white). Explain why the R-squared and estimated coefficients of these regressions are virtually zero.

e) Now regress the squared values of the residuals from the two different regressions on education, age, age2, female, white. Does one set of residuals appear to be more heteroskedastic than the other? Next, regress the squared residuals on education, education², age, age2, female, white, and the interactions education*age, female*age, female*education, white*age, white*education. Use a test based on the R-squareds of these regressions to test for heteroskedasticity in the sample.

f) Explain how the assumption that the residuals from the log-wage regression are “pair-wise” 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) 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, age2, female, white, using the “cluster” option in STATA to correct the estimated standard errors for correlation in the residuals between twins [regress lwage educ age age2 female white, cluster(id)]. Explain why the standard error on the estimated return to education is higher (and t-statistic 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 and uncorrelated with the true level of schooling. What is the formula for the bias in the estimated return to education from the regression of log-wages on educ, age, age2, female, white in terms of their “noise-to-total variance” ratio?

b) Run the following STATA command [ivreg lwage age age2 female white (educ=educt_t)] This performs the 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 educational attainment as an instrument for the individual’s self-reported education (for a given individual, both the individual and his/her twin were asked about the individual’s educational 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, …) 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 A. Write out the “omitted variables” bias in terms of the linear relationships between education and A and log-wages and A.
d) 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 \[\text{regress dlwage deduc if first==1, noconstant}\]. 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 the twin’s self reported education (deduc) on the x-axis using only 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 on 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 \[\text{ivreg dlwage (deduc=deduct) if first==1, noconstant}\]. This two-stage least squares regression uses deduct as an instrumental variable 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 her/his education is correlated with the measurement error in her/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.