# Homework 1: Estimating the returns to education using the twins data

Shu Wang

April 18, 2019

1. Run the bivariate regression of log-wages on a constant and education and show the scatter plot. Now regress log-wage 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 age3 and age4 to the regression. Does this substantially improve the fit of the regression model?

2. 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 twins population? Now compare the mean characteristics of individuals with a college degree (educ=16) to individuals with just a high school degree (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?

3. 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-square, gender and race as controls. Does allowing for nonlinearities in the return to education improve the fit of regression model substantially?

4. Based on the scatter plot of hourly wages on the y-axis and education on the x-axis, is there any evidence on homoskedasticity/heteroskedacity in the wage regression model? What about with log-wages on the y-axis?

5. Regress log-wage on education, age, age2, and the gender and racial indicators, using the "robust" subcommand in STATA to calculate the Eicker-White consistent standard error. 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?

6. Using the "predict" STATA command [predict (var. name), residual], save the residual from both the wage and log-wage regression. Now regress the squared values of the residuals from

two sets of regressions on education, age, age2, female and white. From the R-squares of these regressions, test for heteroskedasticity in the two sets of residuals. Does one set of residual appear to be more heteroskedastic than the other? Now regress the squared residuals on education, education2, age, age2, female, white, and the interactions education*age, female*age, female*education, white*age, and white*education. Again, test for heteroskedasticity based on the R-squares of the regressions.

7. Explain how the assumption that the residuals from the log-wage regression are "pairwise" uncorrelated may be violated when using the twin 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, age2, 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) than when clustering is not corrected for.

8. 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 pait observations based on twin averages). Does this correct the "clustering" problem in the residuals? explain briefly.