Lab 6: Measurement error – how bad is it?

due: 5:00pm on Friday, April 19

Many economists are interested in the relationship between education and wage. In the ideal world, we would base our measures of educational attainment on official records (like transcripts). In the *real* world, however, we often have to rely on self-reported measures. This problem will investigate one type of problem with self-reported measures, and a potential solution.

Data

This problem set makes use of data collected by Ashenfelter and Krueger on identical twins. important variables we'll use are `educ`: self-reported education, `educ_t`: twin's report of educational attainment, and `lwage`: logarithm of earned wages.

Problems

1. Draw a causal graph representing the relationship between education (`TrueEduc`), reported education (`educ`), and log wages (`lwages`). Denote the observed variables with boxes, and the unobserved variables with circles. including error terms for each observed variable.
2. Assume that reported income is the sum of education and an error term that is independent of all other variables, and normally distributed. If we assume the *only* relevant causal variables are the ones we've discussed, will a regression of wages on reported education yield an unbiased estimate of the effect of education on wages? If so, explain why; if not, what will be the direction of the bias?
3. Create a new .do file which will document your STATA commands for the remainder of the assignment. Using comments (*), give your .do file a title, list your name(s), the date the file was created, and give a brief description of what the file is meant to do.
4. Run the regression of wages (`lwages`) on self-reported education (`educ`), and interpret the coefficient on educ, $\widetilde{\beta}$, temporarily ignoring any measurement error problems.
5. Beyond estimating the *direction* of the bias, we can (in some circumstances) estimate the *magnitude* of the bias – doing so depends on our ability to model the error associated with the reported variable of interest. Write down the equation for the *true* relationship between education and wage as a function of $\widetilde{\beta}$, Var(`TrueEduc`), and Var(e). (Hint: see slide 9 from measurement error lecture).

One strategy for estimating the error of a report of education is to compare the report to other estimates. The `pubtwins` dataset includes a twin's report of a subject's education, `educ_t`[1]. Let's assume that the twin's report is the sum of the true level of education and an independent error term drawn from the same distribution as the error associated with `educ`,

6. Create a new variable, `Diffeduc`, which reports the difference between an individual's self-reported education, and the report of their twin.
7. If `educ` = `TrueEduc` + e_1 and `educ_t` = `TrueEduc` + e_2, what is `Diffeduc`?
8. Using the fact that $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$ for uncorrelated variables, find Var(`Diffeduc`), and Var(`TrueEduc`) in terms of Var(e), and Var(`educ`).
9. Estimate Var(`Diffeduc`) using STATA. What is your estimate for Var(e)?
10. Estimate Var(`educ`) in the data. What is your estimate for Var(`TrueEduc`)?
11. Using your estimates of Var(e) and Var(`TrueEduc`), return to the equation you wrote in problem (5) to estimate the (unbiased) relationship between education and wage.

[1] That is, my own report of my educational attainment is given by `educ`, while my twin brother's report of my educational attainment is given by `educ_t`.