Problem 5: Due April 4 at 10:00 AM

Suppose we wish to estimate a linear model of the effect of schooling on income,

$$income_i = \beta_0 + \beta_1 schooling_i + \beta_2 age_i + \epsilon_i, \tag{1}$$

with it being a cross-sectional model for simplicity (as we did in class). There exists a variable, $parents\_income_i$, that is unobserved to the econometrican but enters the data generating process for income:

$$income_i = \beta_0 + \beta_1 schooling_i + \beta_2 age_i + \beta_3 parents\_income_i + \epsilon_i \tag{2}$$

and it also enters the data generating process for schooling,

$$schooling_i = \gamma_0 + \gamma_1 age_i + \gamma_2 parents\_income_i + \gamma_3 peer\_effects_i + \varepsilon_i \tag{3}$$

1. Generate a 1000 observation dataset of all the variables using equations (2) and (3) and normal distributions for age_i, $parents\_income_i$, and $peer\_effects_i$, with distribution parameters and true model parameters of your choosing.

2. Estimate equation (1) with MLE and OLS. Bootstrap the standard errors with 200 samples. Is your estimate statistically significant? Does your estimate of β_1 correspond to the true value in your DGP?

3. Now instrument for the endogenous variable assuming the econometrician observes $peer\_effects_i$ but still does not observe $parents\_income_i$. Is your estimate of β_1 more accurate?