Homework 1
PP 414 Applied Regression Analysis
Instructors: Black and Delgado
Due: October 19th
Fall 2018

Question 1

Instructions: On Canvas, you will find two files. JTPA.csv is a data file for women in the classroom training component of the Job Training Partnership Act experiment. JTPAcodebook.txt is a text file with variable definitions. The outcome of interest is an employment indicator 18 months after random assignment. The expstatus variable indicates whether the individual is in the treatment group or control group. The variable treated indicates whether the individual received classroom training or not.

1. Describe the incidence of treatment by experimental status. Is there substantial crossover? Are the other observable variables balanced between treatment and control groups? How would you test this?

2. Estimate the Intent To Treat (ITT) parameter without covariates.

3. Estimate the impact of treatment on the compliers (the Bloom estimator) without covariates.

4. The data contain numerous predetermined covariates. Re-estimate the ITT using these covariates. Re-estimate the Bloom estimator using these covariates. What happens to the standard errors of these estimates?

5. How might the procedure in question 4 be abused?

Question 2

In this question you will create the dataset and simulate the decision to go to college. Assume an individual chooses to go to college if her earnings as a college graduate, Y_{1i}, are greater than her earnings as a high school graduate, Y_{0i}. Further assume the earnings are given by

$$Y_{0i} = \mu_0 + \varepsilon_{0i}$$

$$Y_{1i} = \mu_1 + \varepsilon_{1i}$$

where the error terms follow a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon_{0i} \\ \varepsilon_{1i} \end{pmatrix} \sim N\left(\mathbf{0}, \Sigma\right)$$

$$\Sigma = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}$$

1. Generate 1,000 observations and create their potential earnings as college and high school graduate. The parameter values are $\mu_0 = 40,000$, $\mu_1 = 50,000$, $\sigma_0 = 10,000$, $\sigma_1 = 20,000$, and the correlation between the error terms is 0. Given the potential earnings, create an indicator, D_i, that takes value 1 if the individual attends college and 0 otherwise. What percentage of individuals in your data decide to attend college?

2. Create the observed earnings for each individual, $Y_i = Y_{0i}(1 - D_i) + Y_{1i}D_i$. What is the "naive" estimator of the effects of obtaining a college degree on earnings?

3. Because we know the potential outcomes for each individual in the treated and untreated states, we can do better than the naive estimator. Estimate the ATE. Is ATE larger or lower than the naive estimator? Why?

4. Estimate ATT and ATN. In this hypothetical example, what would be the consequences of a policy that forces everyone go to college? Would everyone benefit from this policy?

5. Now assume the error terms are highly correlated with a correlation coefficient $\rho = 0.6$ (you might need to calculate the covariance). Are more people going to college in this new scenario? Calculate the naive estimator and ATE.

6. Now assume the error terms are negatively correlated with $\rho = -0.5$. Calculate the naive estimator and ATE. How are these estimates compared to your answer to the previous question? Why these differences?