Problem Set 4

Statistics and Econometrics

Deadline: 11am, 25 November 2020

General Guideline

What we are looking for in the assignments is a demonstration that you can understand the econometrics and statistics questions and can solve them with R or conceptually. That means effective programming to get correct results is needed, but at the same time, clear explanations of economics/business concepts in well presented reports are equally important when assessing your work. In particular, you will be marked for successful (correct) programming (not the style of coding), good understanding of related concepts, and clear interpretations and explanations of results.

Please submit a pdf or html file converted from R markdown/notebook after you program in R.

Question 1

You need to use two data sets for this exercise, jtrain2.RData and jtrain3.RData. jtrain2 was obtained from the National Supported Work Demonstration job-training program conducted by the Manpower Demonstration Research Corporation in the mid 1970s in US. Training status was randomly assigned, so this is essentially experimental data. On the other hand, jtrain3 contains observational data, where individuals themselves largely determine whether they participate in job training. The data sets cover the same time period.

Source:

jtrain2: R.J. Lalonde (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76, 604-620.

jtrain3: R.H. Dehejia and S. Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94, 1053-1062

1. In the data set jtrain2, what fraction of the men received job training? What is the fraction in jtrain3? Why do you think there is such a big difference?
2. Using jtrain2, run a simple regression of $re78$ on $train$. What is the estimated effect of participating in job training on real earnings?
3. Now add as controls to the regression in part 2 the variables $re74$, $re75$, $educ$, age, $black$, and $hisp$. Does the estimated effect of job training on re78 change much? How come?
4. Do the regressions in parts 2 and 3 using the data in jtrain3, and report the results. What is the effect now of controlling for the extra factors, and why?
5. Define $avgre = (re74 + re75)/2$. Find the sample averages, standard deviations, and minimum and maximum values in the two data sets. Are these data sets representative of the same populations in 1978?
6. Almost 96% of men in the data set jtrain2 have $avgre$ less than $10,000. Using only these men, regress $re78$ on $train$, $re74$, $re75$, $educ$, age, $black$, and $hisp$ and report the result. Run the same regression for jtrain3, using only men with $avgre \leq 10$. For the subsample of low-income men, how do the estimated training effects compare across the experimental and nonexperimental data sets?
7. Now use each data set to run the simple regression $re78$ on $train$, but only for men who were unemployed in 1974 and 1975. How do the training estimates compare now?
8. Using your findings from the previous regressions, discuss the potential importance of having comparable populations underlying comparisons of experimental and nonexperimental estimates.

Question 2

Consider a regression model $y = \beta_0 + u$. Suppose we have a variable x, and we want to decide whether or not to include it as an independent variable. Suppose that the model including x, i.e., $y = \beta_0 + \beta_1 x + u$, has a lower AIC than the model with only the intercept β_0. Is it possible that x is statistically insignificant even at 10% level in the model $y = \beta_0 + \beta_1 x + u$? Explain.