Problem Set 3

Statistics and Econometrics

Deadline: 11am, 18 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

Use the data in kielmc.RData, only for the year 1981, to answer the following questions. The data are for houses that sold during 1981 in North Andover, Massachusetts; 1981 was the year construction began on a local garbage incinerator.

1. To study the effects of the incinerator location on housing price, consider the simple regression model

$$\log(price) = \beta_0 + \beta_1 \log(dist) + u,$$

   where *price* is housing price in dollars and *dist* is distance from the house to the incinerator measured in feet. Interpreting this equation causally, what sign do you expect for β_1 if the presence of the incinerator depresses housing prices? Estimate this equation and interpret the results.
2. To the simple regression model in part 1, add the variables $\log(intst)$, $\log(area)$, $\log(land)$, *rooms*, *baths*, and *age*, where *intst* is distance from house to interstate (i.e., a major system of highways running between US states) entrance ramp measured in feet, *area* is square footage of the house, *land* is the lot size in square feet, *rooms* is total number of rooms, *baths* is number of bathrooms, and *age* is age of the house in years. Now, what do you conclude about the effects of the incinerator? Explain why parts 1 and 2 give conflicting results.
3. Add $[\log(intst)]^2$ to the model from part 2. Now what happens? What do you conclude about the importance of functional form?
4. Is the square of $\log(dist)$ significant when you add it to the model from part 3?

Question 2

Use the data in gpa2.RData for this exercise. Consider the equation

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u,$$

where *colgpa* is cumulative college grade point average; *hsize* is size of high school graduating class, in hundreds; *hsperc* is academic percentile in graduating class; *sat* is combined SAT score; $female$ is a binary gender variable; and *athlete* is a binary variable, which is one for student athletes.

1. Estimate the equation and report the results. What is the estimated GPA differential between athletes and nonathletes? Is it statistically significant?
2. Drop *sat* from the model and reestimate the equation. Now, what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part 1.

3. In the model, allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no difference between women athletes and women nonathletes.
4. Does the effect of *sat* on *colgpa* differ by gender? Justify your answer.

Question 3

Consider the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

Our goal is to understand the causal impact of x_1 on y. However, there is an issue of multicollinearity in the model, meaning that x_1 and x_2 are highly correlated (say, correlation between the two is greater than 0.95), and thus we cannot correctly estimate the partial impact of x_1 on y due to inflated standard errors. Discuss how we can estimate the causal impact of x_1 on y in this situation.