Problem Set 4

Causality and Review

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Friday, 05 June 2020

OPTIONAL This problem set is optional. If you submit this problem set, then we will replace your lowest previous assignment grade with this problem set's grade. If you do not submit anything, then your grade will be unaffected.

- If you are worried about your grade, then you should do this assignment.
- If you want a nice way to review for the final, then you should do this assignment.

DUE If you choose to do this assignment: upload your answers on Canvas *before* midnight on Friday, 05 June 2020.

IMPORTANT You must submit **two files**:

1. your typed responses/answers to the question (in a Word file or something similar)
2. the R script you used to generate your answers. Each student must turn in her/his own answers.

If use RMarkdown, you can turn in one file, but it must be an HTML or PDF with your responses *and* R code.

OBJECTIVE This problem set has three purposes: (1) reinforce the topics of time series and statistical inference; (2) build your R toolset; (3) start building your intuition about causality within econometrics/regression.

INTEGRITY If you are suspected of cheating, then you will receive a zero. We may report you to the dean. Everything you turn in must be in your own words.

1. Causality

Imagine that we are interested in analyzing a government program. We consider individuals as *treated* if they participated in the program (and untreated if they did not). Following the notation of the Rubin causal model, imagine that we observe the following sample (which would be impossible observe in real life):

Table: Imaginary dataset

i	Trt.	y_1	y_0
1	0	17	8
2	0	7	5
3	0	10	4
4	1	5	1
5	1	0	0
6	1	1	4

1a. Calculate and report the treatment effect **for each individual** (*i.e.*, τ_i).

1b. Is the treatment effect heterogeneous or homogeneous? Briefly explain your answer.

1c. Calculate and interpret the **average treatment effect** for the sample.

1d. What does it mean if $\tau_i < 0$ for one individual and $\tau_j > 0$ for another individual?

1e. **Estimate the average treatment effect** by comparing the **mean of the treatment group** to the **mean of the control group**. Report your estimate.

1f. Should we expect our estimator in **1e** to provide unbiased estimates? **Explain.**

1g. Why would it be impossible to actually observe all of the data in the table (in real life)? Specifically: Which parts of the dataset would we not observe in real life? Think about *the fundametal problem of causal inference*.

1h. Define and explain selection bias.

1i. Calculate (and report) the selection bias in this sample.

2. General Review

These questions cover concepts that we discussed throughout the course.

2a. Define "standard error".

2b. What is the difference between u_i and e_i?

2c. How do time-series models and cross-sectional models differ?

2d. Write out an ADL(1,1) model where the outcome variable is the number of arrests and the explanatory variables are (**a**) the number of police officers (*e.g.*, Police_t) and (**b**) the GDP (*e.g.*, GDP_t) (in addition to the appropriate lags of the outcome and explanatory variables).

2e. Interpret each of the coefficients in **2d**.

2f. How does heteroskedasticity affect OLS regression?

2g. How do autocorrelated disturbances affect OLS regression? *Hint:* Distinguish between models with lagged outcome variables and models without lagged outcome variables.

2h. What does it mean for a variable to violate variance stationarity?

2i. Why do we care if our standard errors are biased?

2j. What does it mean for a relationship to be *spurious*?

2k. Imagine that you can actually observe the disturbances (u_i). Suppose you care about how the variable x_i affects the outcome y_i. Imagine you observe that the mean of u_i is 3 for individuals where $x_i < 5$ and the mean of u_i is 20 for individuals where $x > 5$. Does this suggest our assumption of exogeneity is true or false? Explain.

2l. Using the following model of test scores, suppose we run a regression that **omits ability**. Will the OLS estimate for β_1 be biased upward, biased downward, or unbiased? Explain your answer.

$$(\text{Test score})_i = \beta_0 + \beta_1 (\text{Hours studied})_i + \beta_2 \text{Ability}_i + u_i$$

2m. How do dynamic models relax the strong assumptions of a static model?

2n. What is measurement error and how does it affect OLS regression?

2o. Interpret β_1 below. All variables are continuous, numeric variables.

$$\log(\text{Happiness}_i) = \beta_0 + \beta_1 \log(\text{Health}_i) + \beta_2 \log(\text{Wealth}_i) + u_i$$

2p. Interpret β_1 and β_2 below. All variables are binary indicator variables, *e.g.*, the outcome variable is an indicator for whether the individual owns her/his home.

$$\text{Homeowner}_i = \beta_0 + \beta_1 \text{Female}_i + \beta_2 (\text{Non-white race})_i + u_i$$