Problem Set 4

Causality and Review

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Thursday, 11 March 2021

DUE Upload your answers on Canvas *before* midnight on Thursday, 11 March 2021.

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 you are using RMarkdown, you can turn in one file, but it must be an HTML or PDF that includes your responses and R code (not just the RMD file).

If we ask you to create a figure or run a regression, then the figure or the regression results should be in the document that you submit (not just the code—we want the actual figure or regression output with coefficients, standard errors, *etc.*).

OBJECTIVE This problem set has three purposes: (1) reinforce the topics of causality, IV, and statistical inference; (2) build your R toolset; (3) review course material.

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 and IV

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	2	4
2	0	3	5
3	0	1	3
4	1	9	5
5	1	0	0
6	1	6	4

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

1b. Within the control group: Is the treatment effect heterogeneous or homogeneous? Briefly explain your answer.

1c. Across the treatment group and control group (for both groups, jointly): Is the treatment effect heterogeneous or homogeneous? Briefly explain your answer.

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

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

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

1g. Calculate the selection bias in this setting.

1h. Why does the difference in groups' means in **1f** differ from the true average treatment effect **1d**?

1i. Define and explain selection bias.

1j. How does randomly assigning individuals into treatment or control help avoid selection bias?

1k. Give an example of when randomization can still suffer from selection bias.

1l. What are the two requirements of a valid instrument? Explain each requirement.

1m. Suppose your boss wants you to estimate the effect of whether counties have COVID-related shutdowns on the counties' infection rates (infections per 10,000), *i.e.*,

$$(\text{Infections rate})_i = \beta_1 + \beta_1 (\text{Has shutdown})_i + u_i$$

Should you be concerned with endogeneity in this regression? Explain your answer.

1n. Now your boss suggests using whether the county's (state's) governor is a Democrat as an instrument. In other words: The proposed instrumental variable is an indicator for whether the governor is a Democract for the state that contains county i.

Is this a valid instrument? Explain using both of the requirements for a valid instrument.

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. Write out an ADL(1,1) model where the outcome variable is the **log** number of arrests and the explanatory variables are (**a**) the **logged** number of police officers (*e.g.*, $\log(\text{Police}_t)$) and (**b**) the **logged** GDP (*e.g.*, $log(\text{GDP}_t)$) (in addition to the appropriate lags of the outcome and explanatory variables).

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

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

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

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

2h. 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.

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

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

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

2k. 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.

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