# Problem Set 4

## Causality and Review

**EC 421:** Introduction to Econometrics

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

# 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.*, $\tau_i$).

**Answer:** The treatment effects for individuals 1 through 6 are -2, -2, -2, 4, 0, 2.

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

**Answer:** The treatment effect is homogeneous **within the control group**: each control individual has a treatment effect of -2.

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

**Answer:** The treatment effect is heterogeneouos **in the sample**: The treatment effect varies across (some) individuals.

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

**Answer:** The average causal effect of participation in the program is 0.00.

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

**Answer:** The program positively affected some people $\left(\tau_j > 0\right)$ and negatively affected other people $(\tau_i < 0)$.

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

**Answer:**

```
# The groups' means
t1 = rubin_df %>% filter(trt == 1) %$% y1 %>% mean()
t0 = rubin_df %>% filter(trt == 0) %$% y0 %>% mean()
# Our estimate of the ATE
ate_est = t1 - t0
```

Our estimate for the average treatment effect is $\hat{\tau} \approx 1$.

**1g.** Calculate the selection bias in this setting.

**Answer:** The selection bias is the difference in the average **untreated** outcome $(y_{0i})$ for the treatment and control groups. Here it is -1.

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

**Answer:** If an individual is treated, then we do not get to observe $y_0$, and if the individual is untreated, then we do not get to observe $y_1$.

**1i.** Define and explain selection bias.

**Answer:** The selection bias is the difference between the average untreated outcome for the treated and untreated groups. It tells us how much the treated and untreated observations differ **in their untreated outcomes**. In other words: It tells us to what extent the untreated individuals provide a good counterfactual for the treated individuals.

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

**Answer:** By randomizing the assignment of treatment, we help the control group and treatment group to have similar distributions (since they are both random samples from the same distribution). Another way to think about it: When we randomizee treatment, we break the relationship between treatment and omitted variables.

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

**Answer:** If our individuals do not comply with their randomized treatment assignment, then we can still have selection bias.

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

**Answer:** An valid instrumental variable must be **relevant** and **exogenous**. Relevance requires that the instrument is actually correlated/predictive/affective with the endogeneous variable of interest. Exogeneity requires that the instrument is uncorrelated with the disturbance.

**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.

**Answer:** We should probably be concerned with exogeneity: there are likely omitted variables in the disturbance that both affect shutdown status and infection rate. We might also be concerned that *has shutdown* is affected by *infection rate* (reverse causality).

**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.

**Answer:** The instrument is probably not valid. It is probably relevant—it seems likely that counties will shutdown for COVID more often in states that have Democrats as governoors. But the proposed instrument is probably not exogenous: there are likely other variables in the disturbance that correlate with more Democrat states (*e.g.*, population density, mask policies, mask wearing, *etc.*).

# 2. General Review

These questions cover concepts that we discussed throughout the course.

**2a.** Define "standard error".

**Answer:** The standard error tells us about an estimator's variability (which tells us about the uncertainty underlying its estimates). More formally, the standard error is the standard deviation of an estimator's distribution.

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

**Answer:** $u_i$ gives the unobservable population disturbance, whereas $e_i$ is the sample-regression-based residual.

**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).

**Answer:**
$\log(\text{Arrests}_t) = \beta_0 + \beta_1 \log(\text{Arrests}_{t-1}) + \beta_2 \log(\text{Police}_t) + \beta_3 \log(\text{Police}_{t-1}) + \beta_4 \log(\text{GDP}_t) + \beta_5 \log(\text{GDP}_{t-1}) + u_t$

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

**Answer:**

- $\beta_1$: For a 1-percent increase in arrests in the previous period ($t-1$), we expect a $\beta_1$ percent increase in arrests in period $t$ (holding all else constant).
- $\beta_2$: For a 1-percent increase in the number of police officers on the street in time $t$, we expect a $\beta_2$ percent increase in arrests in period $t$ (holding all else constant).
- $\beta_3$: For a 1-percent increase in the number of police officers on the street in time $t-1$, we expect a $\beta_3$ percent increase in arrests in period $t$ (holding all else constant).
- $\beta_4$: For a 1-percent increase in GDP in period $t$, we expect $\beta_4$ percent increase in arrests in period $t$ (holding all else constant).
- $\beta_5$: For a 1-percent increase in GDP in period $t-1$, we expect $\beta_5$ percent increase in arrests in period $t$ (holding all else constant).

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

**Answer:** A variable is variance stationary if its variance is constant throughout time.

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

**Answer:** We care abbout biased standard errors because standard errors tell us about the uncertainty underlying our estimates. If our standard errors are biased, then our test statistics, confidence intervals, and hypothesis tests are all wrong. Thus, we are unable to learn about the precision or uncertainty of our point estimates.

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

**Answer:** Spurious relationships appear to be real (or significant) but are, in fact, false.

**2h.** Using the following model of test scores, suppose we run a regression that **omits ability**. Will the OLS estimate for $\beta_1$ be biased upward, biased downward, or unbiased? Explain your answer.

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

**Answer:** Suppose the covariance between ability and hours studied is negative and the effect of ability on test scores is positive. Then our estimate for the effect of studying on test scores will be biased **downward** (we will underestimate the true effect).

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

**Answer:** Dynamic models allow effects to occur across time periods, rather than the rigid assumption of static models that says effects only happen in one period.

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

**Answer:** Measurement error means we have a mis-measured (mis-recorded) variable. We often think of these issue as observing the actual variable "plus noise." Measurement error in the explanatory variable attenuates our estimates (biasing them toward zero).

**2k.** Interpret $\beta_1$ and $\beta_2$ below. All variables are binary indicator variables, *e.g.*, the outcome variable is an indicator for whether the individual owns her/his home.

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

**Answer:** $\beta_1$ tells us the difference in homeownership rates between female and male individuals, holding everything else constant (how much more likely female individuals are to own homes, relative to non-females, holding race constant). $\beta_2$ tells us the difference in homeownership rates between people of color and white individuals, holding everything else constant.