Problem Set 4 Solutions

Causality and Review

EC 421: Introduction to Econometrics

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).

Answer: The treatment effects for individuals 1 through 6 are 9, 2, 6, 4, 0, -3.

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

Answer: The treatment effect is heterogeneous, as it varies across individuals.

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

Answer: The average causal effect of participation in the program is approximately 3.00.

1d. 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 ($\tau_i < 0$).

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

Answer:

```r
# 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 $\hat{\tau} \approx -3.667$.

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

Answer: No. Unless the treatment is exogenous (for example: randomized treatments), then we should not expect an unbiased estimate.

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*.

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.

1h. 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.

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

Answer: The selection bias in this sample is -4.

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

Answer: Time-series models attempt to model an outcome for a single unit (*e.g.*, individual) across time (*i.e.*, repeated observations). Cross-sectional models consider an outcome across individuals (typically for a specific moment in time).

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).

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

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

Answer:

- β_1: For each additional arrest in the previous period $(t-1)$, we expect β_1 additional arrests in period t (holding all else constant).
- β_2: For each additional officer on the street in time t, we expect β_2 additional arrests in period t (holding all else constant).
- β_3: For each additional officer on the street in time $t-1$, we expect β_3 additional arrests in period t (holding all else constant).
- β_4: For each additional unit of GDP in period t, we expect β_4 additional arrests in period t (holding all else constant).
- β_5: For each additional unit of GDP in period $t-1$, we expect β_5 additional arrests in period t (holding all else constant).

2f. How does heteroskedasticity affect OLS regression?

Answer:

1. Heteroskedasticity biases our standard-error estimates (which affects inference).
2. Heteroskedasticity reduces the efficiency of OLS

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

Answer: Autocorrelated disturbances have different effects depending upon whether the model includes a lagged outcome variable.

- **With a lagged outcome variable:** OLS is biased and inconsistent for the regression coefficients.
- **Without a lagged outcome variable:** OLS is consistent *if we have contemporaneous exogeneity*.

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

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

2i. 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.

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

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

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.

Answer: It suggests that our assumption of exogeneity is false: for exogeneity, we need the disturbances to independent of (uncorrelated with) the explanatory variables.

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.

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

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).

2m. 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.

2n. 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).

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

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

Answer: On average, we expecte a 1-percent increase in *health* to increase *happiness* by β_1 percent increase (all else equal).

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.

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

Answer: β_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). β_2 tells us the difference in homeownership rates between people of color and white individuals, holding everything else constant.