Spring 2023

In answering these questions, full marks are given for explanations, not just right answers. Good luck!

1 True/False/Uncertain

For the following problems, respond True False or Uncertain, and explain why. If you can refer to papers or results that is particularly valuable.

- If something is randomly assigned conditional on variables X, controlling for those variables in a linear regression is sufficient to estimate the average treatment effect.

- A difference-in-difference design with a single intervention time period suffers from negative weighting issues.

- It is still possible to partially identify a treatment effect in a regression discontinuity setting if there is bunching in the running variable.

- You should always cluster your standard errors in the most conservative way possible.

- The parallel trends test is a good test for research design validity in the difference-in-difference setting.

- If a treatment spills over and affects more than the treated units, it is not possible to estimate the average treatment effect of a treatment.

- It is not possible to identify anything about the compliers in an instrumental variables regression.

- The quantile regression estimation approach is robust to censoring in the outcome variable.

- In a duration model, it is fine to throw out observations that are right censored.

- A Bartik or shift-share research design rests on the validity of exogeneous shares.

- A randomized experiment doesn't need to control for any confounding factors.

- If you have an outcome variable y with zeros and positive values, you can estimate a percentage effect by using $\log(1 + y)$ as the outcome.

- As long as your first stage F statistic is larger than 10, you should have no weak instrument problems.

2 Application Questions

1. In Angrist, Imbens and Rubin (1996), they study as an application the effect of military service on civilian mortality. The relevant variables are:

    - Z_i: binary variable that person i received a low draft lottery number (such that they more likely to be drafted)
    - D_i: binary indicator that person i served in the military
    - Y_i: binary indicator that person i died between 1974 and 1983 given lottery

    (a) Write these variables in potential outcome notation. Describe, in words, what each potential outcome means. (Hint: there should be 4 total)

    (b) Under what assumptions would a regression of Y_i on D_i yield the causal effect of military service on mortality rates? Write out this estimand using potential outcomes.

(c) Define the exclusion restriction for draft lottery numbers in terms of potential outcomes.

(d) List an example violation of this restriction.

(e) Under what assumptions would a regression of Y_I on Z_i yield the causal effect of draft lottery numbers on mortality rates? Write out this estimand using potential outcomes.

(f) Finally, under what additional assumptions could you use an instrumental variables approach with the draft lottery numbers to identify the effect of military service on mortality? Be precise in defining your assumptions in terms of potential outcomes.

(g) Imagine that the exclusion restriction is violated. Under what settings would this not cause significant bias in the IV estimator?

2. Lee (2008) considers the impact of a Democrat winning on subsquent victory using a regression discontinuitiy design

   - Z_i: running variable – vote share margin of victory (RD at $Z_i = 0$)
   - D_i: winning election
   - Y_i: subseuquent victory in an election

   (a) Write out the estimand for the RD above

   (b) Describe a way in which this design could be violated

   (c) How would you estimate this effect? Describe the estimation procedure (not just what function you would use, but how it would be implemented. You do not need to be precise mathematically).

   (d) A graduate student colleague of yours suggests running a linear regresion on both sides of the regression, using the full dataset, and then taking the predicted value at the cutoff for each regression. What issues might you have with that? Feel free to draw a picture.

   (e) What issues arise with discrete running variables? How would you solve them?

3. Consider a random sample of individuals $i = 1, \ldots, n$, with treatment status D_i and outcome Y_i.

   (a) Write out the individual treatment effects estimands using the potential outcome notation.

   (b) Write the DAG for this effect

   (c) Now imagine that $i = 1, \ldots, n$ instead indexes pairs of roommates in a college dorm, with $Y_i = (Y_{i1}, Y_{i2})$. If we thought treatments from roommates had spillover effects, how would you write the potential outcome? Define the different potential estimands you could construct using this notation (Hint: there should be 4).

   (d) Write out the DAG for this setup.

   (e) Would a regression of outcomes on the number of people in a room who are treated ($X_i = D_{i1} + D_{i2}$) capture any of these effects? Explain.

4. Consider the effect of going to college D_i on earnings Y_i.

   (a) We are given a number of covariates, X_i, and told that conditional on X_i, strict ignorability holds for D_i and the outcome. Write out what that means in potential outcome notation.

   (b) Write down a DAG where this holds.

   (c) How would you implement a p-score procedure to estimate the average treamtent effect of D_i on Y_i, using the strict ignorability condition?

   (d) You're now given data on occupation for these individuals (W_i). We might expect that occupation causes changes in earnings, and the choice of occupation is causally shifted by the decision to go to college. Add this variable to the DAG.

(e) What would happen to our causal estimate if we now added W_i as a control to our estimation procedure?

5. We consider the roll-out of a COVID-19 lockdown across some states, but not others. For $t \geq 0$, $D_{it} = 1$ for states in the treatment group, and $D_{it} = 0$ for the control states. For $t < 0$, $D_{it} = 0$ for everyone. We're interested in the effect on economic activity Y_i, and will use difference-in-differences.

(a) Write out a simple parametric model that would let you identify the effect of this policy, while not necessarily assuming random assignment of D_i, given two time perods. What assumption is necessary?

(b) Write out the regression for how you would estimate a single treatment effect in the post-period for the treatment. (you may have done this in the last problem)

(c) What if D_{it} had been randomly assigned? What could you do instead?

(d) Describe a test you can do to test the validity of this model. What do you need? What are some potential issues with this test?

(e) Now, you get access to policies that have been implemented at different times (a staggered roll-out). Describe in words how having a staggered roll-out provides a more robust identification approach.

(f) What issues might arise if you ran the same regression as in part b) above with the staggered roll-outs? Describe in words, or with algebra, or with a graph (or all three).