Comprehensive Exam Practice Questions mgmt 737

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).

6. You've been put in charge of evaluating the impact of Connecticut's policy to forgive medical debt. The policy targeted people whose household income is up to $60,000.

   The state has tasked you with evaluating the impact of this policy on financial distress. You have been given a dataset with:

   - Y_i: financial distress of individual i
   - X_i: an individual's household income
   - D_i: the individual's initial amount of medical debt

   a. You have determined that you should estimate the effect of the policy on Y using a regression discontinuity design. What assumptions are necessary for this design to be consistent? How would you estimate the effect of the policy using this design?

   b. Describe two tests you could do to test the assumptions necessary for the regression discontinuity design to be consistent.

   c. How could you use this regression discontinuity design to estimate the effect of changes in medical debt on financial distress?

   d. You now realize that the policy actually had two cutoffs: $60k for a single person or $120k for a married couple. Describe, either formally or in words, how you would adjust your estimation procedure to account for this. Do you need anymore assumptions?

   e. The state wants to know what would happen if they changed the cutoff for the policy to $70k. What (well-defended and supported) answer would you give them?

   df. You have now discovered a serious issue with the data: the data on household income is bucketed in bins of $5000, so that everyone between 50,000 and 54,999 is reported as 50,000, everyone between 55,000 and 59,999 is reported as 55,000, etc. How would this affect your ability to estimate the effect of the policy using a regression discontinuity design? How could you address this issue?

7. We want to estimate a linear regression model of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{1}$$

where Y_i is the change in employment in the service sector in city i, X_i is an indicator for whether the city passed a law that increased the minimum wage, and ϵ_i is an error term. Assume you observe n cities.

   a. Assume that we could consistently estimate β_1 by using OLS. How would you calculate the standard errors for β_1 if you assumed that the errors, ϵ_i, were homoskedastic?

   b. Now, we will consider a model where the effect of X_i varies by city, but this heterogeneity is uncorrelated with whether the city passed a minimum wage law. Specifically, we assume that $\beta_i = \beta_1 + (\beta_i - \beta_1)$, $E(\beta_i - \beta_1 | X_i) = 0$, and

$$Y_i = \beta_0 + \beta_i X_i + \epsilon_i. \tag{2}$$

   How would you estimate β_1 in this case? How would your standard errors change? If they change, please specify how you would estimate them.

   c. Now, we realize that the cities don't pass minimum wage laws on their own – rather, the states passed the laws, making X_i correlated within states. We assume we observe S states and n_s cities in each state. Observations in our regression are still at the city level.

      i If we assume that there was no heterogeneity in β_1, how would we estimate standard errors for β_1? What assumptions are necessary?

      ii If we assume that there was heterogeneity in β_1 *and it is correlated within cluster*, with $\beta_i - \beta_1 = b_s + \tilde{\beta}_i$, how would we estimate standard errors for β_1?

8. You are interested in studying the impact of noncompete bans across states (A noncompete agreement is a contract between an employer and employee where the employee agrees not to work for a competitor for a specified period of time after leaving the company). You have a panel dataset with the following variables:

   - Y_{it}: the wages of workers in state i in year t
   - X_{it}: an indicator for whether the state has an abortion ban in year t

   You have n states, and $T = 3$ years of data. In the first year, no states have a ban. In the second year, $n_e = 20$ states have a ban put into place going forward, and in the third year, $n_l = 10$ more states have a ban, giving a total of 30 states with a ban, and 20 without.

   a. Describe how you would estimate the effect of the ban on the average wage in the second year. What assumptions are necessary for this estimate to be consistent?

   b. How would you expand your approach from part a to estimate the effect of the abortion ban on the average wage using all time periods? What assumptions are necessary for this estimate to be consistent?

   c. Are there any tests you could do to test the assumptions necessary for your estimates to be consistent?

   d. One important consideration when studying state-level data is that individuals can cross into other states to work. How would this affect your estimates of the effect of the bans? Do you have any ideas on how you could quantify this effect?

   e. Now, you get another year of data, and in this year, all states have a ban. What assumptions would you need to make to estimate the effect of the ban on the average wage?