PLSC 598: Causal Inference Midterm

Spring 2023

March 1, 2023

You have from 2pm to 5pm today, March 1, to complete this midterm. This exam is open book (including whatever materials you want to use: internet, class notes, GPT3, prayer), but *no consultation or group work*.

Upload a written document, the output of your code, and the code itself, to Canvas. Don't worry about formatting everything prettily (although I won't mind if you do), just make sure it's all clearly labeled in terms of what question you're answering where.

The exam will be curved; the points next to each question are only a measure of their relative importance.

1 True or False: Explain Your Answer (5 Points Each)

1. True or false: we always prefer to use a unbiased estimator to a biased estimator.

2. True or false: a balance test of all measured covariates can tell us whether a regression is causally identified or not.

3. True or false: Matching and weighting are not causal identification strategies.

4. True or false: Bootstrapping is not a causal identification strategy.

5. True or false: Clustering is not a causal identification strategy.

6. True or false: The "sharp null" is a more severe test than the standard null hypothesis test; when we reject the standard null hypothesis, we would always reject the sharp null as well.

2 Short Answer (5 Points Each)

1. What is the fundamental problem of causal inference?

2. What is meant by the "design-focused" approach to causal inference?

3. What are some solutions to the problem of a lack of common covariate support across "treatment" and "control" groups (scare quotes to indicate this is not a randomized experiment)?

4. If the number of clusters is large, kind of problems will we encounter if we fail to adjust for cluster assignment in our estimation? Think in terms of bias and consistency.

3 Literature (30 Points Total)

Munger et al (2022) uses a panel of British survey respondents to estimate the causal effect of consuming political tweets sent by media or politician Twitter accounts on citizens' knowledge of factual political questions. There's other stuff in the paper, feel free to skim or consult, but the key for this question is what's written here.

The outcome of interest, Y_i, is whether the survey respondent correctly answered the question "Over the past five years, has the number of immigrants to the United Kingdom from other EU countries been: Less than 100,000 per year, **Between 100,000 and 300,000 per year,** Between 300,000 and 500,000 per year, More than 500,000 per year?" (Correct answer in **bold**).

This outcome is measured at times t_0 (July 31, 2014) and t_1 (June 17, 2015). The Treatment D_i is the number of tweets about immigration sent by media accounts that person followed during this time period. The covariates X_i are survey-based measures of various aspects of the respondent, including age, gender, education, income, vote choice in the previous election, and a series of questions about offline media diet.

This specific test is found in Figure 5, the fourth coefficient estimate ("Facts: Immigration").

The "Results" section of the paper, page 117 and the first paragraph of page 118, discuss our identification strategy.

Summarize this identification strategy using terms from the course. You can use a DAG, potential outcomes notation, or just English words, but try to be precise. What is our target of inference? What is our research design? (Don't worry about estimation strategy here).

What are some potential problems with this strategy? Don't worry about your substantive knowledge of this case (UK politics, or Twitter), just think about how we've discussed causal identification in the course.

What would an ideal experiment look like in this case?

4 Data Analysis (20 Points Total)

Code to read in the data file is in the attached midterm.R. This is a simplified version of the dataset with

- fact_immigrants_w3_correct_dk0 is the (binary) outcome variable measured at time t_1
- fact_immigrants_w2_correct_dk0 is the same question measured at time t_0
- log_tweets_immigration is the continuous, logged treatment variable

I've dropped all the subjects who did not respond to the entire survey or for whom treatment was not measured; for this portion of the exam, ignore these issues and assume that this is a random sample of the target population of UK citizens.

1. Transform the treatment variable into a binary variable for whether the respondent was exposed to more or less than the median number of tweets in the sample. Conduct some kind of balance test (graphical, regression-based, whatever you think makes sense) of the covariates across these new binary treatment and control groups. Where are there issues of common support?
2. In the sense developed in Aronow and Samii (2016), calculate the Regression Weights for each observation in the full regression defined in the code on line 9. What % of the total weight is contributed by the top 10% of the observations?
3. (Extra credit: do this last, if you have time). Conduct randomization inference on this sample, randomizing the empirical distribution of the treatment variable across the observations and recording the coefficient of the treatment each time. Can you reject the sharp null hypothesis of zero treatment effect?