PLSC 598: Causal Inference Final Exam

Spring 2023

May 2, 2023

You have from noon today, March 2 until 11:59am tomorrow, May 3 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*.

First, read the full exam and see if you have any clarification questions. If you do, send me an email; I will respond to these with an email to everyone (so make sure that everyone has the same information) at 3pm on May 2.

Upload a written document, the output of your code, and the code itself, to Canvas. Don't worry about formatting everything prettily, 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 (4 Points Each)

1. The choice of a loss function in the context of supervised machine learning should be made through cross-validation.

2. It doesn't matter whether an instrument variable *causes* the treatment, only whether it *predicts* it in the data.

3. The goal of minimizing a loss function in the context of supervised machine learning is best achieved by evaluating the bias-variance tradeoff.

4. The regression discontinuity design violates the mutual support condition.

5. Our store of social scientific knowledge is steadily accumulating as we continue to publish academic papers. Not today, and perhaps not tomorrow, but we will eventually complete social science and usher in an era of perfect and permanent harmony.

2 Short Answer (5 Points Each)

1. How should we decide what the appropriate bandwidth for a regression discontinuity design is?

2. When is the ITT the estimand of interest? Provide an example.

3. Under what conditions is the "no defiers" assumption about principal strata (used to justify certain IV estimators) plausible / implausible? Provide an example.

4. Under what circumstances are we interested only in prediction and not causality?

5. In the context of causal mediation, when are we interested in the Controlled Direct Effect versus the Natural Direct Effect? Provide an example.

6. A colleague shows you a graph of the potential outcomes of their dataset

in order to justify their parallel trends assumption to motivate a difference-in-difference design. Is this sufficient evidence to accept this assumption?

7. What is the "manipulability" conception of causality, and how does it relate to the Neyman-Rubin "potential outcomes" framework?

8. When is clustering sufficient to handle the problem of spillovers? When is it not?

9. What is the exclusion restriction? Provide an example in a hypothetical instrumental variables design.

10. In the context of external validity, what is the "unconfounded location" assumption? Provide an example in which this assumption is plausible.

3 Literature (20 Points Total)

One of the key skills in the application of causal inference is the ability to read a paper and explain the research design. I've included the paper "Taking Part without Blending In: Legalization Policies and the Integration of Immigrants" by Professor Stephanie Zonszein.

Zonszein uses a Regression Discontinuity Design to study the process by which immigrants become acculturated to and (potentially) assimilate into the larger culture of the host country. Please read the paper, skimming the theoretical parts which are no relevant for this assignment (another important skill!) and focusing on the methodological details.

1. What are the outcomes of interest? How are they measured?
2. What is the "discontinuity" in the central RDD design?
3. What causal identification assumptions are we required to make in order to accept this research design?
4. Are these assumptions plausible? (Yes: this paper got her a job at Berkeley.) What substantive evidence might convince us that these assumptions are *not* plausible?
5. Look at the analyses presented in Appendices D and E. What do they tell us about the plausibility of the research design?
6. Figure 2 presents the main results of the paper. What is the bandwidth used in this analysis? How is this justified?

4 Data Analysis (20 Points Total)

Conduct a simulation in R to demonstrate the problems generated by – and the solutions to – the problems of non-compliance and non-response. This question is also designed to test your ability to translate the words used in this course into data.

It is important that your code be well-commented so that I can track your thinking about this operationalization.

First, generate a dataset with 20,000 data points. For each of them, generate a covariate X that is evenly distributed between -1 and 1 (that is, with overall mean 0).

Then give each unit a treatment *assignment*; the treatment is binary and randomly assigned.

Now assign each unit potential outcomes equal to their covariate value plus 1 times the value of their *realized* value of the treatment. (Note that we have not yet defined the *realized* treatment value, only their *assigned* value.) The true value of the ATE is therefore homogeneous and equal to the unit-level treatment effect.

Now we are going to assign each unit to one of the four principal stratum: complier, defier, never-taker and always-taker. Do this three times, according to the instructions below. Each time, calculate the empirical ATE and the empirical ITT.

1. Each unit is randomly assigned to one of the four principal strata with probability .25.

2. Each unit is randomly assigned to one of the four principal strata with the following probability distribution: .4 always-taker, .2 never-taker, .2 defier, .2 complier.

3. Units with covariate value below 0 are randomly assigned to one of the four principal strata with the following probability distribution: .2 always-taker, .2 never-taker, 0 defier, .6 complier. Units with covariate value 0 and above are randomly assigned to one of the four principal strata with the following probability distribution: .2 always-taker, .4 never-taker, .2 defier, .2 complier.

Now we are going to complicate the problem with the addition of non-compliance. Replicate the analysis of the ATE and ITT for each of the three steps above for the following two situations:

1. Defiers have a 50% random chance of non-response; each other type of unit has a 20% random chance of non-response.

2. Units with a realized outcome of less than .5 have a 50% chance of non-response; other units always respond.