Final Project Guidelines

Introduction to Causal Inference

The final project is your opportunity to apply what you have learned in the course to a real-world problem. Focus on a point treatment problem (a single intervention node), as that has been the focus of the course. Explicitly and thoughtfully apply each step of the causal roadmap to this problem. This does not mean that you must have a perfect data analysis and write-up. High quality analysis of real data takes time, and many of you will encounter challenges, going beyond the tools of this course. A good project, rather, is one in which you have thought hard about teach step in the roadmap, have done your best given your training thus far, and have clearly identified limitations of your work so far and next steps.

The roadmap in the context of the project

1. **Specify the Scientific Question.** Give some background about why your question is interesting, important and/or relevant.

2. **Specify a Structural Causal Model.** Use a SCM to represent your knowledge (and its limits) about the system you will study.

3. **Translate your question into a formal target causal parameter, defined using counterfactuals.** If you feel a more complex target parameter would be of greater interest, explain why. For this project, however, choose a target you know how to identify and estimate.

4. **Specify your observed data and its link to the causal model.** Describe the data you are working with and its link to the causal model you have specified.

5. **Identify.** Is your target causal parameter identified under your initial causal model? If not, under what additional assumptions would it be identified? How plausible are these for your particular problem? Are there additional data or modifications to your study design that would improve their plausibility?

6. **Commit to a Statistical Model and Estimand.** State these explicitly.

7. **Estimate.** For estimation, apply each of the three estimators: simple or non-targeted substitution estimator, inverse probability of treatment weighted estimator and TMLE. Use of the `tmle` package is acceptable. Use Super Learner when implementing TMLE and IPTW. Provide some justification for your choice of library. Also present a detailed plan for statistical inference/variance estimation based on the non-parametric bootstrap. Implementation of a variance estimate is not strictly required. (It may be quite time consuming if you are bootstrapping SuperLearner on a large data set.) However, if you do not implement the bootstrap, please provide a step-by-step plan for its implementation in R, as well as a description of how you will use the results to test your hypothesis and generate confidence intervals.

8. **Interpret results.** What is the statistical interpretation of your analyses? Discuss differences (or lack thereof) in the estimates provided by the different estimators. What is the causal interpretation of your results and how plausible is it? What are key limitations of your analysis? How might these results (if at all) inform policy, understanding and/or the design of future studies?

There are 3 main deliverables:

1. **Group membership and brief project description** - submit by email
2. **Group presentation** - in class
3. **Group report** - submit by email

Requested exceptions to the above should be discussed with the Professor as soon as possible.

Guidelines for the project description:

The content will not be graded. However, this is your opportunity to make sure your group is on the right track. Please include a very brief description of

- The question in words
- The target causal parameter
- The data, including key variables available
- Any anticipated challenges and how you will address them

Guidelines for the grading of presentations

"A"-level group presentations will

- Have equal group participation; everyone should speak!
- Introduce the class to the data and question of interest.
- Contain a concise description of the application of *each step* of the roadmap to your problem.
- Treat the real-world problem you address seriously.
- If complete, include results from analysis (or preliminary analysis), making use of tables and figures as appropriate.
- Alternatively, provide a detailed analysis plan, including specifying the adjustment set, proposed Super Learner library and possible concerns regarding data support.
- Interpret your results; discuss limitations and next steps
- Fit into the allotted time frame, while leaving room for questions

Guidelines for the grading of reports

"A"-level reports will

- Be clearly written
- Be a group collaboration with explicit statement of the contributions of each author
- Be no more than 12 pages single spaced, not including tables, figures, or appendices.
- Carefully and thoughtfully, apply each step of the causal roadmap.
- Treat the real-world problem seriously.
- Make effective use of tables and figures.
- Accurately identify limitations and briefly consider future directions.
- If you did not implement a variance estimator, provide a step-by step description of how you would in an appendix.
- Use notation correctly.