Guidelines for Final Projects
Modern Methods in Causal Inference

Summer 2021

1 Guidelines for Final Projects:

The final project for this course is your opportunity to apply what you have learned in the course to a real-world problem. Explicitly and thoughtfully apply each step of the causal roadmap to this problem. This does not mean you must have a perfect data analysis and write-up to turn in at the end of the semester. High quality analysis of real data takes time, and many of you will encounter challenges that we have not yet learned the tools to deal with. A good project, rather, is one in which you have thought hard about the issues at each step in the roadmap, have done your best given your training so far, and have clearly identified limitations of your work so far and next steps.

Requested exceptions to the below expectations should be discussed early with Iván.

- A description of the final project should be prepared. You will be required to give a 5 minute presentation to the group. The report and presentation should contain:
  - Name of each member of your group (groups of 3 max.)
  - Description of the problem and scientific question (Step 0. below), including a brief description of the scientific background
  - Description of the dataset that will be used to address the scientific question
  - An exploration of the data set you have available to answer your research question, where you are trying to identify possible issues such as positivity violations etc.
  - This report should be appropriate to show a non-statistician collaborator.
  - For this description (but this does not apply for the final project) it is not necessary to show code, but a Table 1, description of variables, and relevant plots will be helpful to ensure your data set has the necessary information to answer your causal question of interest.
- You will be required to give a group presentation on August 6th (details below). You can and are encouraged to incorporate feedback from the presentation into the final report.
- *Please check the announcement tab on canvas for deadlines for each of these assignments.*

The roadmap in the context of the project:

0. **Specify the Scientific Question (including the target population).** Give some background about why your question is interesting/important/relevant.

1. **Specify a Causal Model.** Use a SCM to represent your knowledge (and its limits) about the system you will study. If you have uncertainty in specifying your SCM, discuss that explicitly and explain why you made the decisions you did.

2. **Translate your question into a formal target causal parameter, defined using counterfactuals.** If you feel a more complex target parameter than those we have learned how to identify and estimate in the class would be of greater interest, explain why. But for this project choose a target you know how to identify and estimate.

3. **Specify your observed data and its link to the causal model.**

   - Describe your observed data and its link to the causal model you have specified. If you feel that in reality the link between your causal model and the observed data is more complex than we have learned in class (n i.i.d. copies of random variable O), explain why. But for this project, stick with the simple link we have learned in class.
   - Be sure to include a basic descriptive table of your data that provides information on the outcome, exposure, and covariate distributions. (i.e. a classic "Table 1" in the applied public health and medical literature.) Feel free to ask for guidance if you are not sure what this should look like.

4. **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 changes to your study design that would improve their plausibility?

5. **Commit to a Statistical Model and Estimand (target parameter of the observed data distribution).** State these explicitly.

6. **Estimate.**

   - Apply each of the three estimators we have learned in class: parametric G-computation (simple substitution estimator), inverse probability of treatment weighted estimator, AIPW, and TMLE. Use of the `tmle`, or other `R` packages is acceptable. Also report the unadjusted results for comparison.
   - Use Super Learner when implementing AIPW and TMLE. For comparison, you may also wish to use it when implementing your G-computation and IPTW estimators. A simple library is fine. (Writing wrappers to include your own parametric regressions as candidates is great). Include an assessment of the performance (cross validated risk estimate) of the algorithms in your library. It is helpful to include the simple mean as a benchmark. Also report an estimate of the cross-validated risk of the Super Learner and interpret.
   - Provide some formal assessment of the positivity assumption. Evaluate the distribution of your estimated propensity score $\hat{\mathsf{P}}(A_i = 1|W_i)$ for $i = 1, \ldots, n$, and corresponding non-stabilized weights (as well as of your stabilized weights if you use stabilized weights to fit the parameters of a MSM). Consider evaluating sensitivity to different truncation levels for $\hat{\mathsf{P}}(A = 1|W)$. Note that for TMLE, bounding or truncating $\hat{\mathsf{P}}(A = 1|W)$ away from 0 is recommended on the basis of both theory and finite sample performance; for IPTW it can help or hurt. Report how for how many observations was the estimated propensity score was truncated.
   - Present a detailed plan for statistical inference/variance estimation based on the non-parametric bootstrap, and implement it (understanding that time may be a limitation depending on your

SL library). Plot your bootstrap distribution and comment as appropriate. For TMLE (and IPTW), you can also report a influence curve based variance estimate for comparison, if you wish.

7. **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?

1.1 Guidelines for the grading of in-class presentations

"A"-level group presentations should:

- Have equal group participation
- Fit into time window: 15 minutes, with 5 min for discussion.
- Introduce the class to the data and question of interest
- Contain a concise description of your application of *each step* of the Roadmap to your problem
- Treat the real world problem you address seriously; think about each step of the roadmap in context
- Include results from analysis (or preliminary analysis) of your data using the estimators we have learned in class. Make 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, and discuss limitations and next steps

1.2 Guidelines for the grading of reports:

"A"-level reports should:

- Be clearly written and in full sentences. i.e. pretend that you are submitting an manuscript to a journal.
- 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/appendices/references
- Treat the real world problem you address seriously
- Apply each step of the causal roadmap appropriately and thoughtfully
- Use notation correctly
- Incorporate background/subject matter knowledge
- Make effective use of tables and figures
- Interpret results thoughtfully
- Accurately identify limitations
- Briefly consider future directions for the topic of interest