Guidelines for Final Projects: PH 252D Spring 2018

The final project for this course 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 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 you 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 with Maya and your GSIs early.

The roadmap in the context of the project:

1. **Specify the Scientific Question.** Give some background about why your question is interesting/important/relevant.
2. **Specify a Casual 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.
3. **Translate your question into a formal target casual 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.
4. **Specify your observed data and its link to the casual model.**
    - Describe the data you are working with and its link to the casual model you have specified. If you feel that in reality the link between your casual model and the observed data is more complex than we have learned in class (n i.i.d. copies of random variable $O \subseteq X$), 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.

5. **Identify.** Is your target casual 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/ changes to your study design that would improve their plausibility?

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

7. **Estimate.**

   - Apply each of the three estimators we have learned in class (simple or non-targeted substitution estimator-a.k.a G-computation estimator), Inverse probability of treatment weighted estimator, and TMLE) to estimate your target parameter. Use of the `tmle`, `ltmle`, or other `R` packages is acceptable. Also report unadjusted results for comparison.

   - Use Super Learner when implementing TMLE. For comparison, you may wish to use it when implementing your G-computation and IPTW estimators also. 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) 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 SL and interpret.

   - Provide some formally assessment of the positivity assumption. Evaluate the distribution of your estimated propensity score $g_n(A = 1|W_i), i = 1, ..., n$, and corresponding non-stabilized weights (as well as of your stabilized weights if you use stabilized weights to fit an MSM). Consider evaluating sensitivity to different truncation levels for g_n. Note that for TMLE, bounding or truncating g_n 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 $g_n(A|W_i)$ 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.

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 preliminary project description:** due March 21.
2. **Group presentation:** in class April 25, April 30, May 2. Due in electronic form (sent to Lina) by 1pm April 24th.
3. **Group report:** due via bcourses by May 7th.

Guidelines for preliminary project description:

Content is not graded. However, this is your opportunity to make sure your group is on the right track. If you are delayed in getting access to data, you may have up to a week extension on turning this in. In it please include a very brief description of:

- Group membership
- Target population/inclusion criteria of study
- Your causal question in words
- Your target causal parameter
- A brief description of the data you will work with, including key variables available. Ideally, this will include a basic assessment of feasiblity, including sample size, an assessment of the marginal distribution of your exposure or treatment variable (e.g. for a binary A, make sure you have a reasonable number of 0s and 1s), and an assessment of the marginal distribution of your outcome (e.g. if you have a binary outcome, make sure there is some reasonable degree of variability. You will have very limited ability to adjust if you have an outcome that is very rare or occurs for almost everyone).
- Any anticipated challenges and how you will address them.

Guidelines for the grading of presentations

"A"-level group presentations should:

- Have equal group participation (everyone should speak!)
- Be a group collaboration with explicit statement of the contributions of each author
- Fit into a 12 minute time window with 5 minutes for questions
- Introduce the class to the data and question of interest
- Contain a concise description of application of *each step* of the roadmap above 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 set using the estimators we have learned in class, making use of tables and figures as appropriate.
- Interpret your results, and discuss limitations and next steps

Guidelines for the grading of reports

"A"-level reports should:

- 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.
- 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