Project Update

CS396 Causal Inference

April 22, 2024

Instructions

This assignment is due on Friday, May 17 at 11:59pm CDT. It will accepted up to 24 hours late with a 20% grade penalty.

Please upload a single PDF for your group to Canvas. You are encouraged to use the TeX file for this assignment to create and format your PDF. Your project update should no more than five pages total, not including references.

As always, your work must be your own. It's fine to use published packages or preprocessing code, but don't claim credit for work that you didn't do. If you use information from other sources, you must cite those.

Rubric (10 points total)

- (1 point) The update is at most five pages long, not including references
- (1 point) List your group members
- (1 point) Describe major changes to your project since your proposal
- (1 point) Include a causal graph and describe the variables
- (1 point) Describe and interpret your causal estimand
- (1 point) Provide and interpret a point estimate of your causal effect
- (1 point) Produce a synthetic data distribution that matches your assumptions
- (1 point) Estimate and interpret the causal effect from synthetic data
- (1 point) Propose at least one next step for the final report
- (1 point) Ask at least one question you want feedback on

1 Group members

Please list your group members.

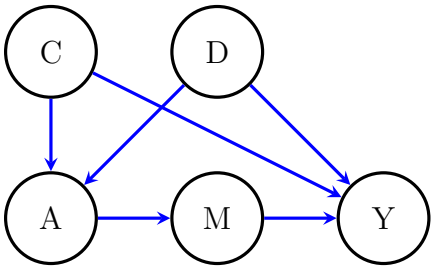
2 Big changes

What's changed in what you're focusing on since you wrote your proposal? This is your opportunity to let me know about any revisions to the project you've made since writing your proposal. If you need help deciding on possible revisions, please reach out to me before turning this in.

3 Causal graph

Draw a causal graph for your problem.

You can use any reasonable method to produce an image of your graph as long as it is easy to read. You could, for example, use the ID algorithm GUI at this website or the `tikz` latex package, like below. Your graph should use as many variables from your dataset as you can, but you aren't required to use them all. If you have a single treatment, a single outcome, and many confounders, you can just include a single confounder node and then indicate that it represents a vector of variables. For each node in your graph, please say what column(s) of your dataset it corresponds to, if it isn't obvious.



4 Counterfactual function

For your treatment A and outcome Y, write out $E[Y^a]$ as a function of the observed data. For example, in the frontdoor example where we had an unobserved U, we wrote:

$$E[Y^a] = \sum_m p(M = m \mid A = a) \sum_{a'} E[Y \mid A = a', M = m] p(A = a')$$

You can find use ID algorithm GUI at this website to find such a function, but please use LaTeX to reformat it rather than taking a screenshot.

4.1 Assumptions

What are the assumptions required for your function above to be an unbiased estimator of the causal effect? For example, are you assuming consistency? Conditional exchangeability? What else? For the conditional exchangeability statements, write them in the form of $X \perp Y \mid Z$.

5 Estimation and Interpretation

Choose a causal effect to estimate, such as the risk difference or risk ratio. Describe how you will estimate that (i.e., what models will you fit?). Will you use the backdoor or frontdoor estimator, or something else? If so, say so. If you have a more complicated graph and will need a more complicated estimator, describe why neither the backdoor nor frontdoor estimator will work. Rather than trying to develop your own estimator for a more complicated setting, you can use a package like Y_0, Ananke, or DoWhy.

5.1 Point estimate

Use your estimator to get a single estimate for a causal effect your project focuses on. If you proposed using several outcomes and treatments in your proposal, you can just focus on one for now. Explain what approach you took to estimating the causal effect and share the numerical estimate of that effect.

5.2 Interpretation

What does your estimate of the causal effect mean for your problem? Tie this back to the reasons why you were interested in this dataset in the first place. If you were able to make policy decisions based on your analysis of this data, what would these results tell you?

For example, if your treatment were pet ownership, your outcome were cardiovascular health, and you chose to estimate the risk ratio, you might say something like "our results indicate that if we intervened to require someone to own a pet, their risk of cardiovascular disease decreases by 10%."

6 Synthetic data

For the real-world dataset you consider, we can't check to see whether your assumptions are valid or your effect estimate is close to the real answer. To convince yourself that the methodology you're using is implemented correctly and that *if* your assumptions are correct *then* your estimator should be unbiased, you'll apply your methods to a synthetic dataset.

6.1 Synthetic generation

First, construct a synthetic dataset. This should be as close in format to your real-world dataset as possible. For example, if your real-world outcome is continuous, your synthetic outcome should be too. If your real-world dataset has many confounding variables, be sure to at least include two in your synthetic data.

Describe your data-generating process. You can do so either in words or as a code snippet that shows how you sample the data. You can refer to the `observed` function in HW1 as a guide.

Generate a dataset with at least 1000 rows that you will use in the next part.

6.2 Synthetic estimation

For the dataset you generated, what is the true causal effect of the treatment on the outcome? In the HW1 example, that was 0.5. Explain how you know what it is in your setting.

Apply the estimator you used from §5 to estimate this causal effect. Do you get (approximately) the right answer? The answer to that question should be yes; try to debug your methodology or increase your synthetic data sample size if not.

7 Next steps for the project

The work above covers the most fundamental aspect of a causal analysis. For your final project, you might consider trying a different estimator or choosing a different treatment or outcome to look at other interpretations of your data.

In addition to any such new analyses, you need to consider **at least one** new methodological considerations that we have or will cover in the second half of the course. These are listed in Appendix A.

Pick (at least) one of these challenges and discuss how it would apply to your project thus far. You don't to have implemented anything by the time you turn in this assignment, but you should describe what you plan to do.

8 Ask for feedback

List at least one part of your project thus far on which you'd specifically like feedback from me. For example: are you stuck on trying to use an existing package or aren't sure how to preprocess your data to apply one of the estimators you wrote? Do you have some initial results but aren't sure how to move forward on interpreting them?

References

Please cite any sources you referenced, including links to your datasets and any packages you used.

A Possible directions for next steps

Here are a few possible directions you might consider; you need to pick at least one and discuss in some detail how you hope to do so. **You do not need to include the text for all these possible directions in your update writeup.**

Unobserved confounding Think about possible variables that could be important to your problem but are not in the dataset you're analyzing. Do such variables introduce confounding between your treatment and outcome? If so, return to Section 4 and propose a new function that would allow you to handle this confounding, and return to Section 5 to see how it changes your estimates of the causal effect.

Causal discovery What if the causal graph you drew in Section 3 is wrong? Use a causal discovery algorithm to try to learn the causal graph from your dataset, and supplement the algorithm with any domain knowledge you're confident in. The TETRAD library is commonly used for causal discovery, but can be a bit difficult to use. I've forked the `causal-learn` library here (a python port of TETRAD) to which I've added a "fast conditional independence test (FCIT)" which can more easily handle a mix of discrete and continuous data.

If you decide to focus on causal discovery, you might want to explore how well different methods (e.g. constraint-based versus score-based) work on your particular data. You can also discuss how does changing your causal graph change how you'd choose your function in Section 4 and your estimation in Section 5.

Measurement error What if one or more variables you're using in your causal graph are actually just noisy proxies for the variables you wish you had? For example, if you're conditioning on BMI as a covariate, you might consider trying to formulate that as a mismeasured proxy for obesity. To apply these methods, you might need to consider whether you can find data to estimate the error rate $p(C^* \mid C)$, e.g., $p(\text{BMI} \mid \text{Obesity})$.

If there's a machine learning classifier which predicts a variable that would be helpful for your analysis, you could treat its outputs as a noisy proxy and use its classification error rate as the $p(C^* \mid C)$.

Double Machine Learning What if the relationship between A and Y is confounded by a nonlinear relationship involving a high-dimensional confounder? Just training a linear regression $E[Y \mid A, C]$ won't work in general, but so-called double machine learning provides a way to get unbiased estimates of the causal effect. While implementing these methods is outside the scope of this class, you can use existing implementations from the DoWhy or similar packages.

Missing data We haven't covered this yet, but will do so soon. If you want additional information on this topic to decide whether it's something you're interested in, you can see the Files folder on Canvas which has slides from 2022.

If you have missing values in your data, what additional assumptions do you have to make in order to identify the causal effect? If you only consider rows that have complete data, you're making an "MCAR" assumption that might not be very believable. Depending on the extent of the missingness, you can approach this by thinking hard about the underlying graphical model and figuring out an approach specific to your dataset. On the other hand, you can use a general

approach like MICE to impute the missing values from the observed data. If you decide to focus on missing data, you'll want to think about what additional assumptions your chosen method makes, and explore how your estimates in Section 5 change.

Selection bias We haven't covered this yet, but will do so soon. If you want additional information on this topic to decide whether it's something you're interested in, you can see the Files folder on Canvas which has slides from 2022.

Selection bias affects all datasets in some way, but may affect certain analyses more than others. If you want to focus on selection bias, think about the data-generating process – are there certain kinds of rows that are more likely to end up in your dataset than "in the wild?" If so, what does that selection 'mechanism' depend on?

We talked about two ways to approach selection bias: either try to find external data on $p(S = 1 \mid X)$ to counteract $P(X \mid S = 1)$, or target a causal effect such as the odds ratio that has built-in symmetry between the treatment and the outcome. Either way, you will have to tweak your function in Section 4 and your estimator in Section 5.