Project Proposal

CS396-4 Causal Inference

February 4, 2022

Instructions

This assignment is due on Thursday, Feb 10 at 11:59pm CST. Assignments will be accepted up to 72 hours late, but with a 14.3% (1/7th) penalty per day late. If your assignment is less than 24 hours late, we'll grade it and you'll receive 85.7% of those points; if it's less than 48 hours late, you'll receive 71.4% of those points. If it's more than 72 hours late, you'll receive zero points. This assignment is worth 5% of your final grade. If you are working in a group and are concerned that some members of your group are not contributing equally, please email me or let me know in an anonymous course survey.

Please upload your (group's) proposal to Canvas as a single PDF. You can use this assignment TeX to fill in your answers. If you are working in a group, please only submit one PDF for the entire group. While this template is rather long because it includes both questions and partial example answers, you may delete these instructions from the pdf you submit. If you do, we expect your proposal to be *at the most* three pages long.

1 Group members

Please list your group members.

2 Problem Statement

Describe the problem you're considering, why it is important, and who else cares about it.
For example, you might say: "we want to understand how much smoking increases the risk of cardiovascular disease (CVD), which is important to public health experts because CVD is a leading cause of death."

3 Causal Questions or Hypotheses

Describe at least one causal question or hypothesis you would like to investigate.

(a) What are the treatment(s) and outcome(s)?

(b) Frame your question as a contrast of counterfactual random variables.

(c) If you haven't been able to decide which causal question(s) you want to ask, please list multiple possible treatments and outcomes and possible arguments for or against each.

For example, you might say:

(a) Our treatment A is smoking and our outcome Y is CVD.

(b) We are interested in the causal risk ratio $E[Y^{a=1}]/E[Y^{a=0}]$, or expected rate of CVD had everyone smoked divided by the expected rate of CVD had no one smoked.

(c) We might also be interested in using 'Cigarettes per day' as a treatment, which might provide a more fine-grained effect estimate, but also requires working with a continuous-valued treatment.

4 Dataset(s)

What dataset(s) do you plan to use?

For the following questions, answer (a) through (c) for *each dataset* you might use. You only need to answer (d) through (f) for *one dataset* – preferrably the largest dataset or the one you plan to work with first.

(a) Describe the dataset's background: how was it collected, what does it contain?

(b) What are the limitations of this dataset?

(c) Describe the format of the data. Can it be represented as a NxD matrix with N individuals and D features? If not, why?

(d) Load the data as a pandas dataframe (e.g. `pd.read_csv`) and provide a printout of at least ten rows.

(e) For at least six variables (columns) in your dataset:

   i. Describe that variable: what is it measuring? Is it a discrete or continuous variable? Does it have any missing values? What is its mean and standard deviation?

   ii. What are the possible[1] causal relationships between this variable and the other variables (in part e)?

(f) What is at least one variable that your dataset doesn't contain but might be a causal factor? How might such a variable complicate your causal question(s)?

For example, if you're working the Framingham dataset, your answers might *start* like:

(a) The Framingham Heart Study was collected starting in 1948, with an initial 5,209 subjects monitored over several years for clinical risk factors and cardiovascular outcomes, such as . . .

(b) The dataset has a few important limitations. First, it has been anonymized in such a way that makes it unsuitable for publication. Second, . . .

(c) Each row of the dataset indicates an observation for a given subject, but not all subjects have the same number of observations. There are a total of X observations across Y subjects. Each observation has Z variables.

(d) After loading the dataset into pandas, we see:

```
RANDID  SEX  TOTCHOL  AGE  SYSBP  DIABP  CURSMOKE  CIGPDAY    BMI
2448    1    195.0    39   106.0  70.0   0         0.0        26.97
2448    1    209.0    52   121.0  66.0   0         0.0        NaN
...
```

(e) Variables:

   i. `AGE` records the subject's age at the time of the observation. It is continuous, with a mean of X and standard deviation of Y. Age may be a cause of most other variables in the dataset, but cannot be caused by anything else.

[1] If you don't have the domain knowledge to answer this question, that's okay. Just focus on at least a few variables where the way the data was measured makes it clear. For example, someone's age cannot be caused by any other variables.

ii. `SYSBP` records the subject's Systolic Blood Pressure. It is continuous, with a mean of X and standard deviation of Y. We expect that SYSBP is caused by ...

(f) Socioeconomic status (SES) may be a relevant but unmeasured confounder. It likely affects all health outcomes such as X, and may influence our treatment variable Y. Trying to account for SES will make identifying our counterfactual $\mathbb{E}[Y^a]$ more difficult because ...

5 Expectations and Concerns

Write a few sentences about what you hope to learn during this project. Are there concepts from the class that you hope to explore with this particular dataset? Are there any challenges you expect to encounter while working on this project?