Backwards Design Assignment
Causal Inference

Nick Eubank

March 1, 2022

For the next step of your group data science project, you must document a plan using this backwards design template.

In completing this assignment, you should follow the template presented below. Note that this is an exercise that is fundamentally about what you do *before* you start working with data. You **should** look into what data is available, and you need to report specific datasets that will enable you to run the analyses you propose, but completing this assignment does not require the presentation of any data analysis.

Deadlines

- **Due**: Tuesday, March 15nd, Start of Class via Gradescope

I hope you're able to have fun with this exercise. It is rare in school that we get to invest in answering precisely the questions we find really exciting, and I hope you will see this as an opportunity to invest in learning about (and potentially helping address) a problem you care about personally. Moreover, as we've discussed before, this is a potential portfolio piece – one unique to your team – you can show future potential employers.

1 Topic:

What is your project about? What problem are you seeking to solve, or in which domain do you think you can contribute meaningfully?

2 Project Question

What specific question are you seeking to answer with this project? For this project, this must be a ***causal*** *question.*

3 Ideal Experiment

If you were a god, what experiment would you run to answer your question? Define both your treatment variable, and your outcome of interest.

4 Pick a Study Context

Where can you get data that (a) measures your outcome variable, and (b) includes variation in your treatment variable?

5 Project Design

Given the context you want to study (and data you can find), what design do you think would be feasible?

6 Model Results

One of the hardest parts of developing a good data science project is developing a question that is actually answerable. Perhaps the best way to figure out if your question is answerable is to see if you can imagine what an answer to your question would look like. Below, draw the graph, regression table, etc. that you would consider to be an answer to your question. Then draw it again, so you have a model result for if treatment has an effect, and a model result for if your treatment does not have an effect. (If the answer to your question is continuous, not discrete (like: what is the effect of health insurance on life expectancy), draw it for high values (high inequality) and low values (low inequality)).

Result if you hypothesis is true	**Result if you hypothesis is false**

7 Final Variables Required

Now that you've specified what an answer to your question looks like, what data do you need to generate that answer?

For each variable, define both the variable you need ***and*** *the population for which you need the variables to be defined.*

You don't have to be too specific ("I need variable 7 from the NHGIS 2019 census 1% sample release") – just define it in the most general terms that are still specific enough to meet your needs (e.g. I need income data for a nationally representative sample of US citizens from both before and after 2012).

8 Data Sources

Finally, given the variables you need for your analysis, what actual data sources do you think will have the data you need?

In specifying the datasets you need, if you list more than one ***also*** *indicate how you think you can relate these datasets (i.e. if you're gonna merge them, what variables do you think those datasets will provide that will allow you merge them? There's no use saying "I'll merge this political survey with medical records of who has received bad care" if the political survey doesn't provide identifying information you can use to link survey respondents to medical records, even if you have both the survey and medical records!)*