INTRODUCTION TO CAUSAL INFERENCE

PROF. SEBASTIÁN VALLEJO VERA

WHY CAUSALITY?

- Most of the questions of interest for us (i.e., social scientists) are causal.
  - There are some questions that are not causal and are interesting nonetheless (e.g., normative questions).
- Social scientists strive to establish causal relationships, that is, evaluate how changes in a variable of interest (X) affects the values in another variable (Y).

Education → Income

Partisanship → Vote Choice

Small Loans → Poverty

- We are not interested in a mere statistical association between two variable through a correlation analysis.
- We want to show—within previously agreed margins of uncertainty—that changes in one variable *cau*se change in another.
  - And then use our comparative advantage as social scientist to tell convincing stories about society and power.

WHAT IS CAUSAL INFERENCE?

WHAT IS CAUSAL INFERENCE?

- The ”causal” component refers to establishing cause-to-effect relationships of the type X→Y (where X is the cause—independent/explanatory variable—and Y the effect—dependent/outcome variable);

WHAT IS CAUSAL INFERENCE?

- The ”causal” component refers to establishing cause-to-effect relationships of the type X→Y (where X is the cause—independent/explanatory variable—and Y the effect—dependent/outcome variable);
- The ”inference” component refers to forming judgments about the reliability of statistical relationships in a population on the basis of random sampling.

WHAT IS CAUSAL INFERENCE?

- Thus, causal inference refers to the process of drawing a conclusion about the response of the effect variable (Y) when the values of the causal variable (X) is changed in a population on the basis of a random sample from that population.

EXAMPLES OF CAUSALITY

- Obvious examples of causality $|E[Y^{a=1}] - E[Y^{a=0}]| > 0$:
  - Moving the light switch causes the lights to turn on.
  - Pulling the thingy of the chair causes the chair to lower.

- Less obvious examples of causality $|E[Y^{a=1}] - E[Y^{a=0}]| > 0$:
  - Getting a college degree causes higher income.
  - Tariffs cause a reduction in imports.
  - Other?

EXAMPLES OF CORRELATIONS

- Examples of correlations $|E[Y|A=1] - E[Y|A=0]| > 0$ that are not causal $E[Y^{a=1}] - E[Y^{a=0}] = 0$:
  - More people drown when the ice cream sales increase.
  - Colds tend to go away after you take Vitamin C.
  - When Liberals govern, the economy worsens.
  - Other?

IMPORTANT NOTE:

- That $|E[Y^{a=1}] - E[Y^{a=0}]| > 0$ does not mean that A is the only variable that causes Y.
- Also, it does not mean that Y will always change when A=1.
  - When know the switch turns on the lights, but it won't if there is no light bulb.
- However, we would still say that the switch turns on the lights. What's important is that the probability (stochastic and non-deterministic) of Y changing increases/decreases when A=1 (but not necessarily that this will always happens).

THE PATH TO CAUSALITY IS A DIFFICULT ONE

(AS ARE THE PATHS OF LOVE)

- Establishing causal relationships is not easy.
- It requires the comparison between two states of the world:
  - A world where $[Y^{a=1}]$ and a world where $[Y^{a=0}]$.
  - Unless you have a time-machine or a quantum-string something or other, those worlds do not exist (or our current technology does not allow us to observe them).
- For example, the same voter cannot be both a partisan and a nonpartisan at the same time to evaluate how partisanship affects vote choice.

- This constitutes the fundamental problem of causal inference: the counterfactual problem.

ESTIMATING CAUSAL EFFECTS

RANDOMIZED TRAILS

- Among the numerous research designs at hand, the one that circumvents best the fundamental problem of causal inference is the randomized experiment, also called randomized trial.
- In a randomized trial, the researcher randomly assigns specific values of the causal variable to groups of people. Generally, members of one group receive an experimental manipulation like a small loan—labeled treatment group—and members of the other group do not receive it (no loan)—control group.
- Random assignment of the experimental manipulation makes other things equal hold on average across the groups.

RANDOMIZED TRAILS

- Ok, but why does this work?

- One way to think about this is in the same way that we think about sampling:
  - The distribution of the means of infinite samples from a population will have a normal distribution (i.e., central limit theorem). With a big enough N, samples are increasingly more likely to be similar.
  - Now, take two of these random samples. If the risk of receiving the treatment is the same for both sample (i.e., if it does not matter which groups receives the treatment; if being treated is independent from the outcome), then the only difference between these two groups is the treatment itself. It's as if these groups were… exchangeable.

RANDOMIZED TRAILS

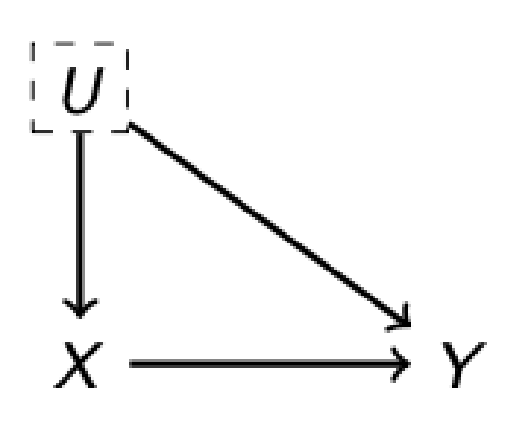
- Let's see this in action (go to code for 3 Intro to Causal Inference.R / Vignette 2.1)

OBSERVATIONAL DATA

- The problem is that experimental data are not always readily available, and some social phenomena simply cannot be studied through randomized trials.
  - We cannot randomly assign *democracy* to a country to estimate its effect on *development*.
- Frequently, researchers must rely on nonexperimental data, also called observational data or retrospective data.
- When working with observational data, researchers have a hard time holding other relevant things equal. In other words, finding a perfect counterfactual to a specific state of the world (to which we could compare to estimate the causal effect of interest).
  - Why is Canada under a Liberal Government in 1995 a bad counterfactual to Canada under a Conservative Government in 1985?

OBSERVATIONAL DATA

- When estimating causal effects using observational data, researchers face one important problem: **omitted variables bias.**
  - A situation where one or many variables that affect both the dependent and the independent variable of interest are left uncontrolled (e.g., leaving motivation out when explaining how education affects income).



- Let's see this in action (go to code for 3 Intro to Causal Inference.R / Vignette 2.2)

OBSERVATIONAL DATA

- Unlike our simulated example, we can never be sure that such relevant variable(s) have been left out and their omission might cause us to mistakenly draw (biased) inferences.
- These unmeasured and/or unaccounted attributes are also called confounders or lurking variables or unobserved heterogeneity.

OBSERVATIONAL DATA

- In randomized trials, random assignment implies that the observed and unobserved factors that affect the dependent/outcome variable are equally likely to be present in the control and treatment groups.
- Luckily, social scientists (mainly econometricians) have developed statistical tools to isolate (as best as possible) causal effects using observational data.
- In this course, we will explore the three following tools (time permitting):
  - Regression analysis
  - Regression discontinuity design (RDD) (like in the code!)
  - Difference-in-Differences (DiD)
- There are other techniques like instrumental variables and matching (which you can explore on your own in most textbooks in the Syllabus).

THE STRUCTURE OF THE DATA

- Observational data come under different structures and each data structure requires specific treatment to evaluate causal relationships.
- In this course, we focus mostly on one data structure: cross-sectional data. Cross-sectional data consist of a sample of individuals, households, cities, states/provinces, countries, or any variety of other units, taken at a given point in time (Prof. Lebo will show you how things that vary in time can cause inference problems).

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

LAB EXERCISES

- From 'wooldridge':
- C5. The data 'fertil2' were collected on women living in the Republic of Botswana in 1988. The variable children refers to the number of living children. The variable electric is a binary indicator equal to one if the woman's home has electricity, and zero if not.
  - (i) Find the smallest and largest values of children in the sample. What is the average of children?
  - (ii) What percentage of women have electricity in the home?
  - (iii) Compute the average of children for those without electricity and do the same for those with electricity. Comment on what you find.
  - (iv) From part (iii), can you infer that having electricity "causes" women to have fewer children? Explain.