ECON526: Quantitative Economics with Data Science Applications

Introduction to Causality

Jesse Perla

jesse.perla@ubc.ca

University of British Columbia

Phil Solimine

philip.solimine@ubc.ca

University of British Columbia

Table of contents

Overview

Summary

- Introduction and motivation for causal inference and randomization
- We will introduce the concepts of treatment effects, potential outcomes, and the fundamental problem of causal inference
- Material includes much adapted from Causal Inference for the Brave and True: Introduction to Causality
- Using the following packages and definitions

```
import pandas as pd
import numpy as np
from scipy.special import expit
import seaborn as sns
from matplotlib import pyplot as plt
from matplotlib import style
```

Introduction

Prediction and Inference

- Machine learning is often criticized as being only about “prediction” and sometimes “inference”

  → This isn’t quite true, but it provides a good starting point to ask what prediction really means

- “Inference” is used in different ways within ML and datascience

  → Sometimes the “point estimate” of some $\hat{f}(X)$ approximation even if we think $y = f(X) + \epsilon$ is the true model

  → Other times means the entire distribution of y given X (e.g., Bayesian inference) or some approximation around the mean with normal covariance (confidence intervals)

- But prefer thinking in probabilities. If there was some true $f(\cdot)$ function,

  → Take some X_1 and X_2 and want to find the distribution $\mathbb{P}(f(X_1, X_2) \mid X_1, X_2)$

Forecasts and Prediction

- The key becomes the distribution itself and what you can and can't condition on. e.g. permissable X_2 values
- From this perspective, prediction is just an unconditional evaluation of the probability distribution, maybe the mean, a sample from it, or with confidence intervals - and not really special

  → The question is whether you have the right joint distribution!
- Forecasts typically just condition on the past observations, but could condition on future events

  → i.e., how might GDP grow if a tax cut is passed in 3 years

Counterfactuals: “What If?”

- Most interesting problems in economics are about counterfactuals in one way or another
  - → What would have happened to the economy if the government had not intervened?
  - → What would have been her income if she had not gone to college, or if she wasn’t subjected to gender bias?
- By definition these are not observable. If we had the data we wouldn’t need to ponder “What if?”. How? One way or another....

YOU HAVE TO MAKE $HIT UP

The Role of Theory

- There is no data interpretation without some theory - even if it is sometimes implicit
- The role of both data and theory is then to help constrain the set of possible counterfactual
- So any criticisms of ML as “merely prediction” are basically a statement on whether the theory makes sense

  → i.e., if you fit $y = f(X) + \epsilon$ on data to find a $\hat{f}(X)$ function, then theory tells you if you made the right assumptions (e.g., that the X data is representative and wouldn’t change for your counterfactual of interest, etc)
- Some models (e.g., random assignment) have easier to swallow assumptions than others.

Approaches

- Always remember: you need assumptions in one form or another because the counterfactuals are inherently not factual
- Broadly there are three approaches to conducting counterfactuals. They are not mutually exclusive
  1. Structural models: i.e. emphasize theory + data to put structure on the joint distribution of $\mathbb{P}(X_1, X_2)$
  2. Causal inference using matching, instrumental variables, etc. which use theoretical assumptions on independence to adjust for bias and missing latents
  3. Randomized Experiments/Treatment Effects where you can get good data which truly randomizes some sort of “treatment”.

Why do People Love Randomized Experiments?

- Because the assumptions are often easy to believe if you trust your random assignment

  → It often requires fewer assumptions beyond random assignment - for better or worse

- However:

  → They are not always possible, and even when they are, they are not always ethical

  → And even when possible and ethical, the inherent difficulty in randomization means it has limited scope and generalizability. i.e., you can learn an effect in one circumstance, but how common are those exact circumstances?

Potential Outcomes Framework

Treatments

- A coherent approach, which fits will with randomized trials, is to emphasize "treatment". This means conditioning on binaries. Language/tools best thought of in terms of pharmaceutical trails
  - → Call the value $T_i \in \{0, 1\}$ as the treatment
  - → Let $Y_i(T_i)$ be the observed outcome
  - → Let $Y_i(0)$ be the outcome if $T_i = 0$
  - → Let $Y_i(1)$ be the outcome if $T_i = 1$
- The key: **you never get to see both**. One is always counterfactual

Potential Outcomes

- Many economic questions posed as: what would have happened if T_i was different for person i? (or country i, etc)
- A "structural" model might be able to help answer that question, but might require a lot of assumptions on the underlying structure of i
- Alternatively, maybe we can make fewer (or different) assumptions and ask:
  - → **Average Treatment Effect**: $\mathbb{E}[Y_i(1) - Y_i(0)]$
  - → **Average Treatment Effect on the Treated**: $\mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1]$
- Note here that we are taking expectations over the distribution of i. Hides lots of probability.

Potential Outcomes Framework

- The potential outcomes framework is a way to formalize causal inference
- It involves defining potential outcomes Y_{0i} and Y_{1i} for each unit under different treatment conditions
- The treatment variable T_i is a binary variable that indicates whether unit i receives the treatment ($T_i = 1$) or not ($T_i = 0$)
- The treatment effect on a unit of type i is the difference between the potential outcomes under different treatment conditions: $\tau_i = Y_{1i} - Y_{0i}$

Treatment Effects

- We are generally intersted in treatment effects of the form $\tau_i = Y_{1i} - Y_{0i}$. However, we cannot observe *both* potential outcomes for a given unit. Instead, we can estimate

  → The **average treatment effect (ATE)**, which is the average of the treatment effects across all units: $\tau = E[Y_{1i} - Y_{0i}]$

  → The **average treatment effect on the treated (ATT)**, which is the average of the treatment effects for units that receive the treatment: $\tau_T = E[Y_{1i} - Y_{0i}|T_i = 1]$

- In randomized experiments, we can estimate the ATE and ATT using the difference in means between the treatment and control groups

  → Why is randomization important? To find out, let's look at a stylized example

A Stylized Example

Let's say that we are interested in calculating the effect of providing a tablet on the academic performance of students, measured by their test scores.

In some strange universe, imagine that we are actually able to observe the outcome for each unit under both treatment conditions. Suppose the data look like this:

```
import pandas as pd
import numpy as np

df = pd.DataFrame(dict(
    i= [1,2,3,4],
    Y0=[500,600,800,700],
    Y1=[450,600,600,750],
    T= [0,0,1,1],
    test_score= [500,600,600,750]
))
df.head()
```

	i	Y0	Y1	T	test_score
0	1	500	450	0	500
1	2	600	600	0	600
2	3	800	600	1	600
3	4	700	750	1	750

A Stylized Example

```
1 df.head()
```

	i	Y0	Y1	T	test_score
0	1	500	450	0	500
1	2	600	600	0	600
2	3	800	600	1	600
3	4	700	750	1	750

Then we can simply calculate the average treatment effect directly:

```
1 df["TE"] = df.Y1 - df.Y0
2 df.TE.mean()
```

```
-50.0
```

And the ATT would be the mean of the last column for units that receive the treatment:

```
1 treatment_rows = df["T"] == 1
```

```
-75.0
```

Bias

A More Realistic Example

In practice, we can't really ever observe both potential outcomes for each unit. Instead, we can only observe the outcome for each unit under the treatment condition that they actually receive.

Suppose that instead of what we had before, the data look like this:

```
df_real = pd.DataFrame(dict(
    i= [1,2,3,4],
    Y0=[500,600,np.nan,np.nan],
    Y1=[np.nan,np.nan,600,750],
    T= [0,0,1,1],
    test_score= [500,600,600,750]
))
df_real.head()
```

	i	Y0	Y1	T	test_score
0	1	500.0	NaN	0	500
1	2	600.0	NaN	0	600
2	3	NaN	600.0	1	600
3	4	NaN	750.0	1	750

A More Realistic Example

```
1 df_real.head()
```

	i	Y0	Y1	T	test_score
0	1	500.0	NaN	0	500
1	2	600.0	NaN	0	600
2	3	NaN	600.0	1	600
3	4	NaN	750.0	1	750

Let's try to estimate the ATE again. We don't observe the counterfactual outcomes, but we can try to estimate the ATE using the difference in means between the treatment and control groups:

```
1 group_averages = df_real.groupby("T").test_score.mea
2 group_averages[1] - group_averages[0]
```

```
125.0
```

This is **not** the correct answer! Why not?

Bias

- In this case, the $\bar{Y}_0$ for the treated is different than the $\bar{Y}_0$ for the control group, (which is what we are using in the calculation)
- In the example of schools adopting a tablet program, the $\bar{Y}_0$ for the treated is higher than the $\bar{Y}_0$ for the control group
  - → This is because the schools that adopt the tablet program would have had higher test scores even if they had not adopted the program
  - → However, once they adopt the program, their test scores drop
- Blindly taking the difference in means between the treatment and control groups will give us a biased estimate of the ATE
  - → In other words, the expected value of our estimator is not equal to the true value of the parameter we are trying to estimate
- **Bias** is measured as the difference between the **expected value of our**

Bias

- Using potential outcome notation, we can write the conditional mean calculation (from the previous example) as:

$$E[Y \mid T=1] - E[Y \mid T=0] = E[Y_1 \mid T=1] - E[Y_0 \mid T=0]$$

- Then adding and subtracting $E[Y_0 \mid T=1]$, and rearranging gives:

$$E[Y \mid T=1] - E[Y \mid T=0] = \underbrace{E[Y_1 - Y_0 \mid T=1]}_{ATT} + \underbrace{E[Y_0 \mid T=1] - E[Y_0 \mid T=0]}_{BIAS}$$

Randomized Experiments

Randomization

- Earlier, we said that randomization is important for estimating treatment effects. Why is that?
- Randomization is the easiest way to ensure that the treatment and control groups are similar on average

  → If the treatment and control groups have the same counterfactual average outcome, then the bias term is zero
- To see why, note that if the treatment and control groups are similar on average, then $E[Y_0 \mid T = 1] = E[Y_0 \mid T = 0] = E[Y_0]$, the population average of the potential outcome under the control condition

  → The bias term is $E[Y_0 \mid T = 1] - E[Y_0 \mid T = 0] = E[Y_0] - E[Y_0] = 0$

  → Therefore, the difference in means between the treatment and control groups is equal to the ATT

Randomized Experiments

- Experimental randomization is the gold standard for estimating treatment effects
- In a randomized experiment, the treatment assignment is independent of the potential outcomes

  → This means that the treatment assignment is independent of the outcome under both treatment conditions

  → In other words, Y_0 and Y_1 are independent of T
- These randomized experiments are also called **randomized controlled trials (RCTs)**

Randomized Experiments - The Difficulties

- RCTs are the gold standard for estimating treatment effects because they ensure that the treatment and control groups are similar on average
- However, RCTs are not always ethical or feasible. For example, it would be unethical to
  - → Randomly assign people to smoke cigarettes
  - → Randomly assign people to be exposed to a deadly virus
- Throughout (recent) history, there have been many examples of unethical experiments
  - → The Tuskegee experiment (1932-1972)
  - → The Stanford prison experiment (1971)
  - → The Milgram experiment (1961-1962)
- These experiments have led to the development of ethical guidelines for

Estimating Treatment Effects Without Randomization

- While many RCTs, especially to answer economic questions, are not unethical, they are often not feasible
  - → For example, it would be difficult to randomly assign people to be born in different countries
  - → It could also be very expensive to randomly assign people to receive a house or to supplement their income
- Due to the ethical and practical constraints, we often have to rely on observational data to estimate treatment effects
  - → In this case, we have to make assumptions about the data-generating process to estimate treatment effects
  - → These assumptions are called **identification assumptions**

Estimating Treatment Effects Without Randomization

- Many identification assumptions are based on the idea of **selection on observables**
  - → This means that the treatment assignment is independent of the potential outcomes *conditional* on the observed covariates
  - → In other words, Y_0 and Y_1 are independent of T conditional on X
- This allows us to do things like **match** units in the treatment and control groups that are similar on average
  - → We can then estimate the treatment effect by comparing the outcomes of the matched units, assuming that the matched units are similar on average
- Another option is to use **instrumental variables** to estimate the treatment effect
- Alternatively, a **structural model** can be used to estimate the treatment effect

Credits

This lecture draws heavily from Causal Inference for the Brave and True: Introduction to Causality by Matheus Facure