

Foundations of causality | R

Agenda

Statistical Modeling & Causal Inference – Oswald | Ramirez-Ruiz

- Revisiting lecture
 - Causality
 - Potential Outcomes Framework
 - NATE and biases
- Getting started with R
 - Data-wrangling with `dplyr`

Causal Inference

The *reasoning* process of

- **ruling out** non-causal explanations of the observed association
- pointing out the **assumptions** necessary to rule out such sources

plus

- providing **evidence** to support or refute these assumptions

Potential Outcomes Framework

Statistical Modeling & Causal Inference – Oswald | Ramirez-Ruiz

Key concept: Every individual has a potential outcome (Y_i) both under treatment and under control (no treatment).

The fundamental problem of causal inference: we can only ever observe one of these states.

So, we cannot observe the individual treatment effect (ITE), nor directly observe the average treatment effect (ATE).

POF Notation

NATE

$$= E[y_{1i}|d_i = 1] - E[y_{0i}|d_i = 0]$$

“The expected outcome when treated, for those in the treatment group”

$E[y_{0,i}] / E[y_{0,i}]$ “expected outcomes”

$y_{0,i} / y_{1,i}$ “potential outcomes”

POF Logic

$$\text{ATE} = E[\underline{y_{1i} - y_{0i}}] = E[y_{1i}] - E[y_{0i}]$$

$$\text{ATT} = E[y_{1i}|d_i = 1] - \underline{E[y_{0i}|d_i = 1]}$$

$$\text{ATC} = \underline{E[y_{1i}|d_i = 0]} - E[y_{0i}|d_i = 0]$$

$$\text{NATE} = E[y_{1i}|d_i = 1] - E[y_{0i}|d_i = 0]$$

Unattainable: we cannot observe
counterfactuals.

Example

- Contact hypothesis (Allport, 1954)

Inter-group contact (X)  Prejudice (Y)

Each individual i in a student sample is exposed ($d_i = 1$) to the cause, or not exposed ($d_i = 0$) (here: contact with member of different ethnic group).

$y_{0,i}$ = non-exposure

$y_{1,i}$ = exposure

ATE, ATT, ATC

If we could observe counterfactuals...

...we could know:

Student (<i>i</i>)	Prejudice		δ_i	Contact
	y_{0i}	y_{1i}		
1	6	5	-1	0
2	4	2	-2	1
3	4	4	0	0
4	6	7	1	0
5	3	1	-2	1
6	2	2	0	1
7	8	7	-1	0
8	4	5	1	0

→

$$ATE = E[\delta_i] = \frac{-1 + (-2) + 0 + 1 + (-2) + 0 + (-1) + 1}{8} = -0.5 \quad (5)$$

$$ATT = \frac{-2 + (-2) + 0}{3} = -1.333$$

$$ATC = \frac{-1 + 0 + 1 + (-1) + 1}{5} = 0$$

NATE

We can only observe half of the potential outcomes we need to get to the ATE...

...so we can only calculate a naïve average treatment effect.

Student (<i>i</i>)	Prejudice		Contact
	y_{0i}	y_{1i} δ_i	
1	6		0
2		2	1
3	4		0
4	6		0
5		1	1
6		2	1
7	8		0
8	4		0

Information we *do* have

$$\begin{aligned} NATE &= E[y_{1i}|d_i = 1] - E[y_{0i}|d_i = 0] \\ &= \frac{2 + 1 + 2}{3} - \frac{6 + 4 + 6 + 8 + 4}{5} \\ &= 1.666 - 5.6 \\ &= -3.933 \end{aligned}$$

NATE and biases

Student (<i>i</i>)	Prejudice		Contact
	Y_{0i}	Y_{1i} δ_i	
1	6		0
2		2	1
3	4		0
4	6		0
5		1	1
6		2	1
7	8		0
8	4		0

Information we *do* have

The treated and untreated groups may differ in more ways than just being treated or not and, therefore, have different potential outcomes.

$$NATE = ATE + \underbrace{E[Y_0|D=1] - E[Y_0|D=0]}_{\text{selection bias}} + \underbrace{(1-p)(ATT - ATU)}_{\text{HTE bias}}$$

baseline bias

heterogeneous /
differential treatment
effect bias

NATE and biases

$$NATE = ATE + \underbrace{E[Y_0|D = 1] - E[Y_0|D = 0]}_{\text{selection bias}} + \underbrace{(1 - p)(ATT - ATU)}_{\text{HTE bias}}$$

Baseline bias: difference in average outcome without treatment for the treatment and control groups.

Differential treatment effect bias: the difference in the average treatment effect between the treatment and control groups, weighted by the proportion of the population in the control group.

Tackling biases

Randomization: randomly assigning subjects to $D=0$ or $D=1$.

- The **probability** of being assigned to treatment is the same for all subjects.
- Being assigned to treatment does **not depend of any characteristic** of the subjects.
- The treatment and control groups have (on average) the same potential outcomes
- **Key point:** *when using random assignment* (and the SUTVA holds), then $ATE = NATE$

$$ATE = E[Y_{1i}] - E[Y_{0i}] \longrightarrow ATE = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

Picking up the lecture discussion again...

Statistical Modeling & Causal Inference – Oswald | Ramirez-Ruiz

You are part of the newly established EU Policy Impact Evaluation Unit.

- **Your mission** is to evaluate a brand new policy that allocates funds to EU regions to combat climate change by fostering green energy, industry, housing, etc.
- To qualify for the funding regions have to be above 125% of the EU average of CO2 emissions per capita.
- **You are given full control** in the pilot phase (i.e., you alone can decide how funds are allocated).
- **What design do you propose** to evaluate the impact of the policy on CO2 emission reduction at the regions level?



Coding with `dplyr`

We'll often use the pipe operator (`%>%`) to string together commands, and rely on the `dplyr` “verbs”. For example:

`select`: subset columns

`filter`: subset rows

`arrange`: reorder rows

`mutate`: add columns to existing data

`summarize`: summarize values in the dataset

`group_by`: defines groups within dataset

Further Resources

R basics: <https://tinyurl.com/vkebh2f>

RMarkdown: The definitive guide <https://tinyurl.com/y4tyfqmg>

Dplyr: <https://tinyurl.com/vyrv596>

Dplyr video tutorial: <https://www.youtube.com/watch?v=jWjqLW-u3hc>

Summary of lab materials – [Lab homepage](#)

For any coding issues – [Stackoverflow](#)

Hertie's Data Science Lab – [Research Consulting](#)