CS&SS 321 - Data Science and Statistics for Social Sciences

Module III - Introduction to causal inference and linear models

Ramses Llobet

Module III

- This module introduces and reviews the topic of causation in science.
  - *randomization*.
  - *applied causal inference*.
  - *causal modeling* (module IV).
- It also introduces the **linear regression model** and the method of **least squares** (LS).

The statistics war of the late XXth century

The statistics war of the XXIth century

- Causal inferences requires a model outside of the statistical model.

Causes in, causes out

- Why do experiments work? When do they work?
- What if treatment is imperfect assigned?
- Should you *control* for anything? Everything?

Answers depend upon **causal assumptions** ($\rightarrow$).

- **Definition**: an assumption is a **premise** or **supposition** that is accepted without direct evidence, often forming the basis for reasoning or an argument.

Causes in, causes out

- Causal assumptions requires **causal knowledge** of social systems. For example:
  - Where X represents **rain** and Y represents **puddles**.
  - What **causal assumption** ($\rightarrow$) you find more reasonable?

$X \leftarrow Y$

$X \rightarrow Y$

Causal design

- **Step 1**: sketch a (scientific) casual model: $X \rightarrow Y$.
  - *Causes in*: assumptions reflect **background knowledge** (*literature review*).
- **Step 2**: use the model to design **data collection** and **statistical procedures**.
- **Step 3**: use statistical analyses to **hypothesis test** and report results.
  - *Causes out*: assumptions have implications about the **causal mechanism**.

Causal design: intervention

- In **experimental designs** with complete control over settings, interventions involve assigning **treatments** to test causal assumptions.
  - *Example*: Pouring a bucket of water on the floor creates a puddle; does rain follow?
- We formalize this is via the **potential outcomes** framework.

Causation in science

Treatment indicator: $T_i \in \{0, 1\}$, where i refers respondents.

- **(1) example:**
  - $T_i = 0$ indicates no membership in a union.
  - $T_i = 1$ indicates membership in a union.
- **(2) example:**
  - $T_i = 0$ indicates no daughters.
  - $T_i = 1$ indicates having daughters.

Outcome: Y_i

- **(1) example:** redistribution attitudes (*gincdif*).
- **(2) example:** pro-feminist attitudes (*progressive.vote*).

Causation in science

- Consider the treatments' (T) **causal mechanisms** ($\rightarrow$) that drives the **outcome** (Y).
  - **Why** does labor **union membership** increase the sense of solidarity?
  - **Why** does having a **daughter** increase pro-feminist attitudes?

Potential outcomes $Y_i(0)$, $Y_i(1)$, where:

- **(1) example:**
  - $Y_i(0)$ represents redistribution attitudes *without* membership.
  - $Y_i(1)$ represents redistribution attitudes *with* membership.
- **(2) example:**
  - $Y_i(0)$ represents pro-feminist attitudes *without* daughters.
  - $Y_i(1)$ represents pro-feminist attitudes *with* daughters.

Causation in science

The **fundamental problem of causality**, we cannot observe two outcomes at the same time:

$$\text{individual treatment effect} = Y_{\text{Ramses}}(1) - Y_{\text{Ramses}}(0) \tag{1}$$

Instead, we **estimate** effects by taking the differences in means between **treatment**, $\bar{Y}(1)$, and **control**, $\bar{Y}(0)$, groups.

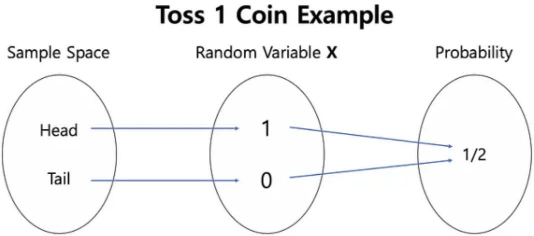
$$\text{average treatment effect} = \bar{Y}(1) - \bar{Y}(0) \tag{2}$$

However, we can identify **ATE** if, and only if, the treatment D has been **randomly assigned** to each respondent `i`. Formally,

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \tag{3}$$

Causation in science

- Think about random assignment as flipping a coin.
  - In **expectation** (as $n \to \infty$), a fair coin has a probability of 0.5 to show tails (0) or heads (1).
  - By definition, a random event has a probability of 0.5.



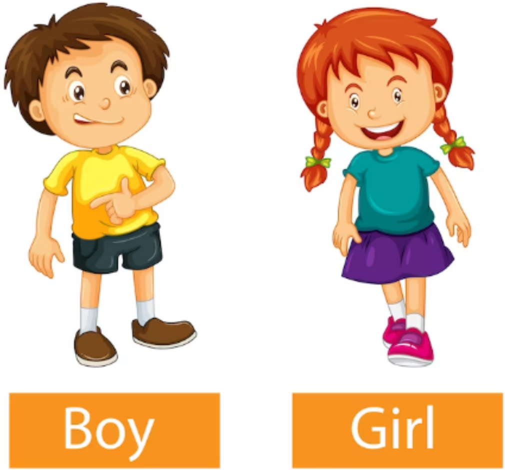
- **What if**, in expectation, a coin has a probability of 0.7 ?

Causation in science

- Is labor union membership a random occurrence?

Causation in science

- Is having a girl (instead of a boy) a random occurrence?

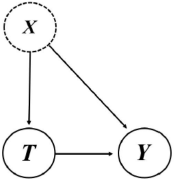


Causation in science

- **Selection bias**: Self-selection and unbalanced factors introduce bias in our statistical estimations.
  - *Self-selection*: Left-wing individuals are more likely to become labor union activists.
  - *Unbalanced factors*: Labor union members may systematically differ from non-union members in terms of factors such as occupation and income.

Causation in science

- In observational studies, unconditional treatment effects are unlikely due to the influence of **confounding** factors, both **observed** and **unobserved**.



- However, sometimes we can assume **conditional random effects**.

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i. \tag{4}$$

Causation in science

- Let's work a short coding example.
- Open the file `unions_sweden.Rmd`, we will do only the **first** section.
- We will finish the remaining section next week.

From previous model: Data Generating Process

- A **Data Generating Process** (DGP) refers to the hypothetical or real mechanism that generates a dataset.
  - It is a conceptual model that describes **how** the observed data is generated or produced.
- **Distributions** represent **systematic behavior** (aka, DGP).
- When looking at a distributions:
  - think in terms of a **DGP**, and
  - **how** the data was generated.

From previous model: Data Generating Process

- Two very useful pieces of information from a DGP are its **mean** and **standard deviation**.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad ; \quad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where

- $\bar{X}$ represents the **sample mean**.
- n is the number of **observations** in the sample.
- X_i represents **values** from a variable in the sample.
- S represents the **sample standard deviation**.

Standard devitation and variance

- The **standard deviation** and **variance** are both measures of the spread of a distribution.
  - To estimate the variance (S^2), we simply take the **square** of the standard deviation (S).

$$S^2 = \left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^2 \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- S^2 is the **sample** variance.
- Q: Why choose the standard deviation over the variance to report **summary statistics**?