Advanced Data Analysis with Python

Cecilia Graiff

November 5, 2025

ALMAnaCH, Inria Paris - École d'Affaires Publiques, SciencesPo
cecilia.graiff@sciencespo.fr

Resources used for this class

- Brady Neal - Causal Inference Course
- Towards Data Science - Causal Inference Tutorial
- Stanford Lab - Causal Inference Tutorial
- Matheus Facture - The Python Causality Handbook (Book and Notebooks)

Homework Correction

An introduction to causal inference

What is causal inference?

Definition (Causal Inference)

Causal inference entails statistical methods that analyze the response of an effect variable when one of its causes is changed. As such, it is used for the evaluation of the effects of a treatment or policy intervention.

Recall: Linear Regression

Model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y: dependent variable (e.g., household income)
- X: independent variable (e.g., years of education)
- β_0: intercept, β_1: slope, β_0 and β_1: coefficients or parameters
- ε: error term

Applied example: Predicting income from years of education. Each additional year increases expected income by β_1.

Insights of linear regression

Linear regression gives information about the relation between two variables, but does not imply **causation**.

Example

Research question: Does the amount of nurses taking care of a patient increase if the patient's health status is worse?

Can we model this question with linear regression?

Yes, we can:

$$nurses\_amount = \beta_0 + \beta_1 \times health\_status + \varepsilon$$

Yes, we can:

$$\mathit{nurses\_amount} = \beta_0 + \beta_1 \times \mathit{health\_status} + \varepsilon$$

Therefore:

To decrease patient mortality, hire less nurses.

Yes, we can:

$$\textit{nurses\_amount} = \beta_0 + \beta_1 \times \textit{health\_status} + \varepsilon$$

Therefore:

To decrease patient mortality, hire less nurses.

Correlation $\neq$ causation!

Correlation vs. Causation

Correlation

- Measures *association* between two variables.
- Does not imply causation.
- Examples:
  - Nurses per patient and mortality
  - Number of firefighters and fire damage

$\boxed{\text{Nurses per patient}} \xleftrightarrow{\text{Correlation}} \boxed{\text{Mortality}}$

Correlation vs. Causation

Correlation

- Measures *association* between two variables.
- Does not imply causation.
- Examples:
  - Nurses per patient and mortality
  - Number of firefighters and fire damage

Causation

- One variable *directly affects* another.
- Requires mechanism or intervention.
- Examples:
  - Smoking → Lung cancer
  - Job training → Higher employment

- Standard statistical analysis (e.g. linear regression): focused on **prediction**
- Causality: focused on **"what if?"** questions:
  - "What would happen if variable X was exposed to event Y?"
  - "What would happen if we reduced nurse hiring for the hospital?"

Direct Acyclic Graphs

Definition (Direct Acyclic Graphs)

A graph is a pair $G = (V, E)$, where V is a set whose elements are called vertices, and E is a set of unordered pairs v_1, v_2 of vertices, whose elements are called edges.

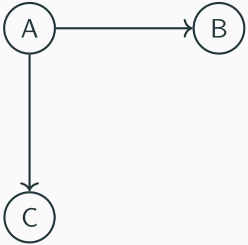
Source: Wikipedia

Direct Acyclic Graphs

Definition (Direct Acyclic Graphs)

A graph is a pair $G = (V, E)$, where V is a set whose elements are called vertices, and E is a set of unordered pairs v_1, v_2 of vertices, whose elements are called edges.

Source: Wikipedia



Basic Causal Relationship

X has a direct causal effect on Y.

Bias

- What makes causality and correlation differ is **bias**
- Possible bias causes:
  - Confounding
  - Omitted Variables
  - ... and many others

Confounding Factors

- A **confounder** is a variable that influences both the treatment and the outcome.

Confounding Factors

- A **confounder** is a variable that influences both the treatment and the outcome.

Example:

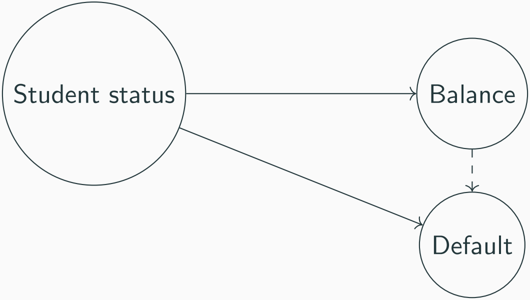
- Indipendent variable: Credit card balance
- Dipendent variable: Default of credit card
- Confounder: Student status

Confounding Factors

- A **confounder** is a variable that influences both the treatment and the outcome.

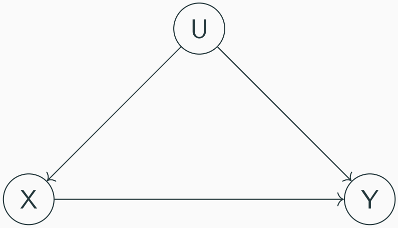
Example:

- Indipendent variable: Credit card balance
- Dipendent variable: Default of credit card
- Confounder: Student status

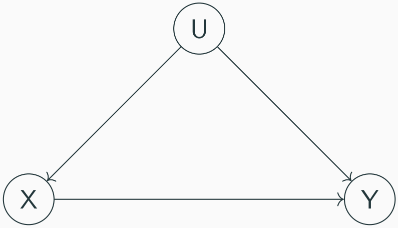


The example is taken from https://www.statlearning.com/

Confounding Example



Confounding Example



U is a confounder affecting both X and Y. Observing only X and Y may produce biased causal estimates.

Omitted Variable Bias (OVB)

- OVB occurs when a relevant variable (confounder) is left out of the analysis.

Omitted Variable Bias (OVB)

- OVB occurs when a relevant variable (confounder) is left out of the analysis.

Example:

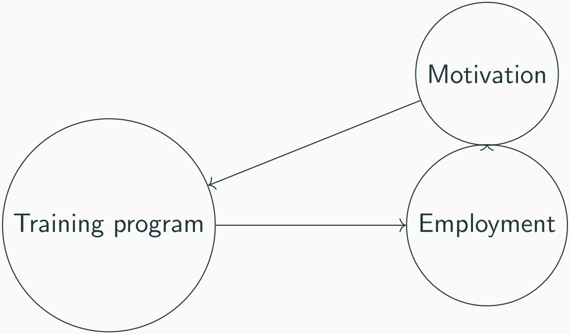
- Study: Effect of training program on employment
- Omitted variable: Student motivation

Omitted Variable Bias (OVB)

- OVB occurs when a relevant variable (confounder) is left out of the analysis.

Example:

- Study: Effect of training program on employment
- Omitted variable: Student motivation



The example is taken from https://matheusfacure.github.io/python-causality-handbook/10-Matching.html

Causal Inference Frameworks

How do we model **causal relationships?**

Causal Inference Frameworks

How do we model **causal relationships?**

- Rubin Model
- Fundamental Problem of Causal Inference
- Average Treatment Effect

The Rubin Model

Definition

The Rubin Model is a causal inference framework that defines what it means for a treatment or policy to **cause** an effect.

The Rubin Model

Definition

The Rubin Model is a causal inference framework that defines what it means for a treatment or policy to **cause** an effect.

For every individual i, there exist two **potential outcomes**:

The Rubin Model

Definition

The Rubin Model is a causal inference framework that defines what it means for a treatment or policy to **cause** an effect.

For every individual i, there exist two **potential outcomes**:

Scenario	Notation	Example (Education Policy)
If the person receives treatment	$Y_i(1)$	Earnings *if educated*
If the person does not receive treatment	$Y_i(0)$	Earnings *if not educated*

The Rubin Model

Definition

The Rubin Model is a causal inference framework that defines what it means for a treatment or policy to **cause** an effect.

For every individual i, there exist two **potential outcomes**:

Scenario	Notation	Example (Education Policy)
If the person receives treatment	$Y_i(1)$	Earnings *if educated*
If the person does not receive treatment	$Y_i(0)$	Earnings *if not educated*

Treatment effect: $Y_1 - Y_0$

Define the subject's **status** as follows:

- $D = 1$: subject received treatment
- $D = 0$: subject did not receive treatment

Define the subject's **status** as follows:

- $D = 1$: subject received treatment
- $D = 0$: subject did not receive treatment

Then we can define the **observed outcome** Y:

$$Y = (1 - D) * Y_0 + D * Y_1$$

The Fundamental Problem of Causal Inference

The Fundamental Problem of Causal Inference (Holland, 1986): For each individual, we can only observe one of the two potential outcomes — either $Y_i(1)$ or $Y_i(0)$, but never both.

The Fundamental Problem of Causal Inference

The Fundamental Problem of Causal Inference (Holland, 1986): For each individual, we can only observe one of the two potential outcomes — either $Y_i(1)$ or $Y_i(0)$, but never both.

However: We must estimate the missing counterfactual outcome. We use strategies such as **randomization** and **matching**.

Average Treatment Effect

On average, how much does the treatment change the outcome compared to not receiving the treatment?

Average Treatment Effect

On average, how much does the treatment change the outcome compared to not receiving the treatment?

$$ATE = \mathbb{E}[Y(1) - Y(0)]$$

Further notation

- **Covariates**: individual characteristics for each subject
  - Example: demographics (age, sex, ...), income, education level, ...
- **Propensity score**: probability of receiving treatment conditional on X
  - Defined as $p(X) = p(D = 1 \mid X)$

Overlap

Definition (Overlap)

The property of **overlap** is satisfied when **propensity score** is **bounded away from 0 and 1**:

$$\eta < e(x) < 1 - \eta \quad \text{for all } x.$$

Meaning:

- For all subject, some are treated and some are untreated
- No one has a propensity score extremely close to 0 (nobody is **almost sure** to be untreated)
- No one has a propensity score extremely close to 1 (nobody is **almost sure** to be treated)

Unconfoundedness

Definition

The property of **unconfoundedness** is respected when the probability of a subject being assigned to a group is **fixed** and **does not depend on its potential outcomes** (denoted as W_i):

$$Y_i(1), Y_i(0) \perp W_i \mid X_i$$

Setting

- **Randomized setting**:
    - Result of a RCT
    - **Unconfoundedness** and **overlap** satisfied

Setting

- **Randomized setting**:
  - Result of a RCT
  - **Unconfoundedness** and **overlap** satisfied
- **Observational setting**:
  - The above mentioned property of **unconfoundness** is **not valid** $\rightarrow$ the probability of the subject being assigned to a group is **influenced by (one of) its potential outcomes**

Setting

- **Randomized setting**:
  - Result of a RCT
  - **Unconfoundedness** and **overlap** satisfied
- **Observational setting**:
  - The above mentioned property of **unconfoundness** is **not valid** $\rightarrow$ the probability of the subject being assigned to a group is **influenced by (one of) its potential outcomes**

Example:

- At least 3 nurses are randomly assigned to a patient: **randomized setting**.
- Patients divided into groups depending on the amount of nurses assigned to them: **observational setting**.

Randomization

Definition (RCT)

A **Randomized Controlled Trial (RCT)** is an experiment method that **randomly** classifies subjects in a treatment group (outcome Y_0) and in a control group (outcome Y_1).

Randomized Controlled Trial

Reminder

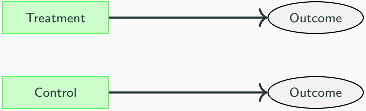
Correlation $\neq$ Causation

Therefore, $ATE = \mathbb{E}[Y(1) - Y(0)] \neq \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$

Randomized Controlled Trials vs Conditional Probability

Randomized Controlled Trials (RCT)

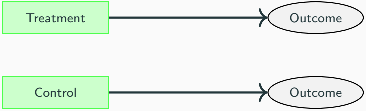
- Participants randomly assigned to treatment or control.
- Eliminates confounding.



Randomized Controlled Trials vs Conditional Probability

Randomized Controlled Trials (RCT)

- Participants randomly assigned to treatment or control.
- Eliminates confounding.



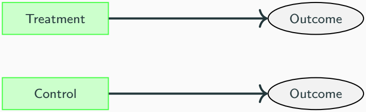
Conditional Probability

- Probability of outcome given a condition: $P(Y \mid X)$
- Does *not* guarantee causality.
- Sensitive to confounding variables.

Randomized Controlled Trials vs Conditional Probability

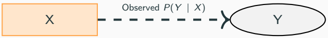
Randomized Controlled Trials (RCT)

- Participants randomly assigned to treatment or control.
- Eliminates confounding.



Conditional Probability

- Probability of outcome given a condition: $P(Y \mid X)$
- Does *not* guarantee causality.
- Sensitive to confounding variables.



RCT vs Conditional Probability: Example

RCT Example: Nurses per patient

- Randomly take 100 patients, from which:
  - 50 patients: assign more than 3 nurses (Treatment)
  - 50 patients: assign less than 3 nurses (Control)
- Measure deaths in both groups

RCT vs Conditional Probability: Example

RCT Example: Nurses per patient

- Randomly take 100 patients, from which:
  - 50 patients: assign more than 3 nurses (Treatment)
  - 50 patients: assign less than 3 nurses (Control)
- Measure deaths in both groups

Conditional Probability Example

- Group of 100 patients
- Compute $P(\text{Death} \mid \text{Nurses}>=3)$
- **Confounding factor**:
  - Gravity of illness

RCT vs Conditional Probability: Example

RCT Example: Nurses per patient

- Randomly take 100 patients, from which:
  - 50 patients: assign more than 3 nurses (Treatment)
  - 50 patients: assign less than 3 nurses (Control)
- Measure deaths in both groups

Conditional Probability Example

- Group of 100 patients
- Compute $P(\text{Death} \mid \text{Nurses}>=3)$
- **Confounding factor**:
  - Gravity of illness

Causal effect: RCT measures causal effect directly ($45/50 - 35/50 = 0.20$), while conditional probability may overestimate effect.

ATE Estimation

- Randomized setting (RCT):
  - Difference in-means estimator
- Observational setting:
  - Regression
  - Matching
  - Propensity score weighting

Difference in-means estimator

Definition

The difference-in-means estimator is the sample average of outcomes in treatment minus the sample average of outcomes in control.

$$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{T_i Y_i}{\bar{T}} - \frac{(1-T_i)Y_i}{1-\bar{T}}\right)$$

$$\text{with } \bar{T} = \frac{1}{N}\sum_{i=1}^{N} T_i$$

$$T_i = \begin{cases} 1 & \text{if individual } i \text{ is treated} \\ 0 & \text{if individual } i \text{ is in control} \end{cases}$$

Adjusted Linear Regression

- We want the causal effect of job training on employment.
- But many other variables (motivation, resources, experience, family background) affect both.
- Those variables are considered as **control**
- We compute the effect of **job training** by eliminating the effect of the other variables

Adjusted Linear Regression

- We want the causal effect of job training on employment.
- But many other variables (motivation, resources, experience, family background) affect both.
- Those variables are considered as **control**
- We compute the effect of **job training** by eliminating the effect of the other variables

$$\text{Employment} = \alpha + \kappa \, \text{job training} + \boldsymbol{\beta}' X + \varepsilon$$

- κ: effect of interest (job training of employment)
- $\boldsymbol{\beta}$: parameters for controls

Adjusted Linear Regression: Step 1

1. **Remove the predictable part of treatment**

- We model the **treatment** as a linear regression on the **control**

$$T_i = X_i \boldsymbol{\beta}_{aux} + v_i$$

$$v_i = T_i - X_i \beta_{aux}$$

- We estimate the linear relationship between job training and controls (motivation, resources, etc) to see how the controls influence job training
- We extract the part that is **not predictable** with the controls

Adjusted Linear Regression: Step 2

Residual-on-Residual Regression

$$\hat{\kappa} = \frac{\mathsf{Cov}(Y, \tilde{T})}{\mathsf{Var}(\tilde{T})}$$

- This is the coefficient of a regression of Y on the residualized T.
- It is the same as the coefficient on T from the full regression:

$$Y = \alpha + \kappa T + X\boldsymbol{\beta} + \varepsilon$$

Adjusted Linear Regression: Summary

- We first remove all parts of job training explained by background factors.
- Then we ask: does the remaining, "unexplained" part of job training still predict unemployment?
- That remaining slope (κ) measures the causal impact of job training, assuming:

$$(Y(1), Y(0)) \perp T \mid X$$

Adjusted Linear Regression: Summary

- We first remove all parts of job training explained by background factors.
- Then we ask: does the remaining, "unexplained" part of job training still predict unemployment?
- That remaining slope (κ) measures the causal impact of job training, assuming:

$$(Y(1), Y(0)) \perp T \mid X$$

κ is the **partial effect** of treatment T on outcome Y, controlling for X.

Matching

Definition

THe idea behind **matching** is that for each subject of the **treatment group**, you find a subject of the **control group** that has the same characteristics.

Matching

Definition

THe idea behind **matching** is that for each subject of the **treatment group**, you find a subject of the **control group** that has the same characteristics.

Person	Treated?	Age	Outcome
A	Yes	30	100
B	No	29	90
C	No	50	120
D	Yes	52	130

Table 1: Example of treated and control individuals for matching

Problem: Features X usually do not match completely.

Problem: Features X usually do not match completely.

- **Scale** the features
  - Variables such as age (max 3 digits) or income (potentially way more digits) would have a very different impact otherwise
- Define a **distance metric**
  - Usually **Euclidean distance**: $\mid X_i - X_j \mid$

Propensity Score

Definition

This method is based on the idea that it is not necessary to control **single counfounding variables**, it is sufficient to **control for a balancing score**:

$$e(X) = P(T \mid X)$$

In other words: **propensity score** allows matching individuals with similar characteristics (expressed by the variables X_i) without handling each variable singularly.