Introduction to causal inference

Tools for Causality, Thematic Quarter on Causality

Elise Dumas

September 2023

EPFL

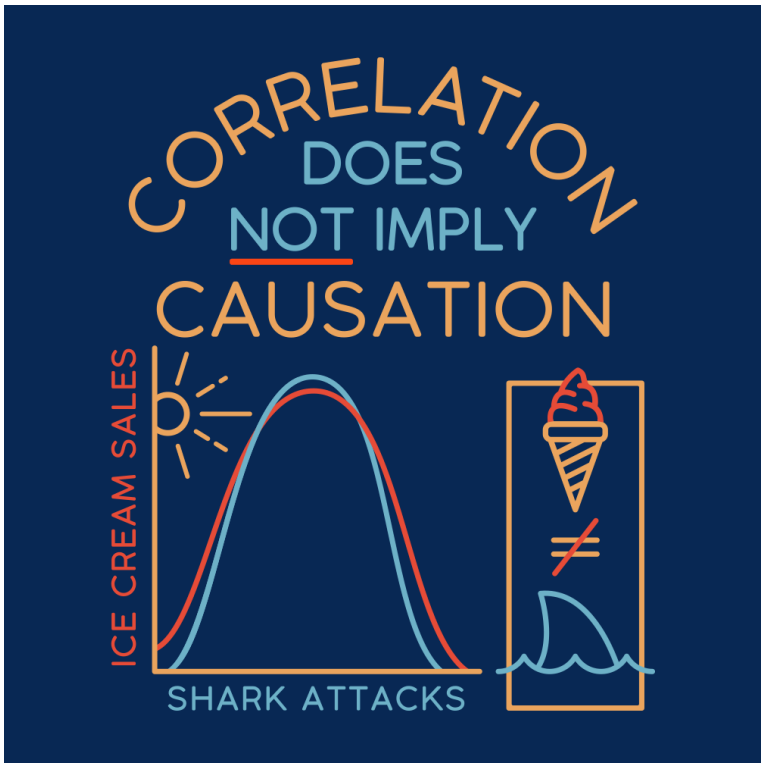
Contents

Contents

Objectives

- To introduce (review) basics of causal inference.
- Why asking causal questions ? And how to express them with mathematical language?
- How can we **identify** and **estimate** causal target parameters? Under which assumptions ?
- I've tried to encompass the names of libraries in R/Python that you can use (list non-exhaustive).

Introduction



Ok, but then, how can we define causation ? Can we derive causal relationships using statistical models and data ?

We are used to thinking causally

But sometimes we need help

- Does wearing masks prevent the spread of covid ?
- Does raising minimum wage increase labor demand ?
- Does immunotherapy reduce cancer cells growth?

Can we answer these causal questions with maths and data ?

Machine Learning are predictive tools, not causal tools (in general)

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Figure: **Recognition algorithms generalize poorly to new environments.** Top five labels and confidence produced by ClarifAI.com shown.

Interpolation is great, but we need to **extrapolate**.

And sometimes predicting is not enough, we need to intervene

Models predicting patients' survival are great, but what can we do with them ? They do not (in general) answer questions like: Why does this patient have a poorer prognosis ? What if I intervened and gave him/her/they this treatment ?

Inferential statistics : pipeline

- In statistics, we **estimate** theoretical quantities using data we have available.
- It is called **estimation**, and is composed of three steps :
  - Defining a **statistical estimand** : what theoretical value do we want to estimate?
  - Picking a **statistical estimator** : a function we apply on the data
  - Applying this function to the data we have available to get a **point estimate**
- Additionally : you can conduct some **statistical inference** around your point estimate : getting a CI, statistical hypothesis testing, *etc* (done thanks to theoretical properties of our estimator)...

A parallel with cooking recipe

estimand

Ingredients

- 150g unsalted butter, plus extra for greasing
- 150g plain chocolate, broken into pieces
- 150g plain flour
- ½ tsp baking powder
- ½ tsp bicarbonate of soda
- 200g light muscovado sugar
- 2 large eggs

Method

1. Heat the oven to 160C/140C fan/gas 3. Grease and base line a 1 litre heatproof glass pudding basin and a 450g loaf tin with baking parchment.

2. Put the butter and chocolate into a saucepan and melt over a low heat, stirring. When the chocolate has all melted remove from the heat.

estimator

estimate

Causal questions

- Define **units**, **treatment** (intervention), **outcome**
- What we want to know : does **treatment** causally impact the **outcome** in the **unit** population?
- What if we were to intervene on treatment status ? How would this affect the outcome ?

Standard statistics and probability cannot translate these questions. We need new variables (potential outcomes) and an additional layer in the inferential pipeline (identification).

Contents

A birthday example

- Does buying a **birthday cake** from a **cake shop** improve **the cake's look**?
- Versus home made cake.
- For my birthday, I asked for an Olaf cake.
- My family bought it in a cake shop ($T = 1$).
- I thought that the cake looked good ($Y = 1$).

A birthday example

- But to know if the cake was looking good because it was made in a cake shop rather than at home (in which case there is a causal effect),
- I would have needed to know the look it would have had had my family made it at home.

Potential outcomes

- We have a set of n units indexed by i
- Let T_i be the value of a treatment assigned to individual i.

Definition of potential outcomes

The **potential outcome** under treatment level t, denoted by $Y_i(t)$, is the value that the outcome would have taken were T_i set to t, possibly contrary to the fact.

- For binary T_i: $Y_i(1)$ is the value we would observe if unit i were *treated*.
- $Y_i(0)$ is the value we would observe if unit i were under *control*
- Potential outcomes are fixed attributes of the units.
- Sometimes you'll see potential outcomes written as:
  - Y_i^1, Y_i^0 or $Y_i^{d=1}$, $Y_i^{d=0}$
  - Y_{i0}, Y_{i1}
  - $Y_1(i)$, $Y_0(i)$

Potential Outcomes: history of the concept

Our chosen way to express counterfactual statements as probability quantities

- Started from Neyman (1923) and Fisher's (1935) work on understanding experiments
- Formalized by Donald Rubin in a series of famous papers (1st one in 1974)
- Potential Outcomes has evolved into an entire framework for causal inquiry.

An alternative way to express counterfactuals in probability is Judea Pearl's (2000) do operator.

The individual causal effect

Potential outcomes enable us to translate causal questions into the estimation of **a causal estimand**.

Individual Treatment Effect (ITE)

For each individual i, we can define the Individual Treatment Effect τ_i as the difference:

$$\tau_i = Y_i(1) - Y_i(0)$$

We can also consider other quantities than the difference, such as the ratio, or the percentage increase due to treatment. In any case, it is some contrast measure between two potential outcomes.

The Fundamental Problem of Causal Inference

- Can we do something to estimate the ITE ?
- It relies on both the value of $Y_i(1)$ and $Y_i(0)$
- But in practice, we get to observe, at best[1], one of the potential outcome.
- Even if my sister asks for another Frozen cake for her birthday, and asks it to be home-made, the difference in the cake's look could be attributable to many other things : Was the cake more difficult to make ? Was there a sugar shortage at the time of my sister's birthday?

The Fundamental Problem of Causal Inference (Holland 1986)

It is impossible to observe the value of $Y_i(1)$ and $Y_i(0)$ for the same unit. Therefore, it is impossible to observe the ITE.

[1] actually we need an assumption here, more on that later

The "Statistical" Solution to go around the fundamental problem of causal inference

It is very hard to learn the value of τ_i for individual units

Instead we can focus on the average treatment effect on the population:

The average treatment effect (ATE)

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$$

This is a causal estimand! Let's see if we can do something for it.

In an hypothetical world, we have :

	My birthday	My sister's	My father's	My mother's
	Cake shop	Home-made	Home-made	Cake shop
Y(1)				
Y(0)				

In reality, we observe [2] :

	My birthday	My sister's	My father's	My mother's
	Cake shop	Home-made	Home-made	Cake shop
Y(1)				
Y(0)				

Can we average the cake looks when coming from cake shop and compare it to the cake looks when home made? Yes, but under very stringent assumptions!

[2] We need an assumption here (SUTVA)

Contents

The SUTVA assumption (1/2)

How do we get from our observed data to potential outcomes? With an assumption.

The Stable Unit Treatment Value Assumption (SUTVA)

Observed outcome = Potential outcome of the observed treatment, i.e.,

$$Y_i(t) = Y_i \text{ if } T_i = t$$

For a binary treatment, sometimes write

$$Y_i = Y_i(1) \times T_i + Y_i(0) \times (1 - T_i)$$

We used it implicity several times already (see footnotes).

The SUTVA assumption (2/2)

SUTVA can be broken down into 2 main components:

- No interference: Manipulating another unit's treatment does not affect a unit's potential outcomes. (e.g. if I get a home-made cake birthday, it will not affect the potential look of my sister's birthday cake).
- Consistency : For each unit, there is no different form or version of each treatment level, which lead to different potential outcomes ($T_i = 1$ means the same thing for all i) (Same cake shop)

- SUTVA deals with the way we design the experiment.
- Be careful with "treatments" which are not interventions
- Paradigm of "no causation without manipulation" (Holland, 1986)

Positivity

Positivity

We assume that, for all units i and treatment levels, t:

$$Pr(T_i = t) > 0$$

i.e., each unit can be assigned every treatment level with some probability.

- Positivity is a "technical" assumption.
- This is needed to actually estimate quantities with our data. This comes from probability theory: we are not allowed to condition on an event with probability 0!
- It depends on how the treatment is assigned.
- Here : for each person's birthday, there is a non-zero probability of having both home-made cake and cake shop cake.

Ignorability

(Strong) ignorability

We assume that, for all treatment levels, t:

$$Y_i(t) \perp\!\!\!\perp T_i$$

The potential outcomes are independent of treatment assignment.

- Ignorability is sometimes referred to as **unconfoundedness** or **exchangeability**.
- Ignorability is a very strong assumption.
- It implies that

$$E[Y_i(1)] = E[Y_i(1)|T_i = 1] = E[Y_i(1)|T_i = 0]$$

- and that

$$E[Y_i(0)] = E[Y_i(0)|T_i = 1] = E[Y_i(0)|T_i = 0]$$

What does ignorability mean?

- Ignorability tells us that we can think of the treated group as a kind of "random sample" of $Y_i(1)$ and the control group a "random sample" of $Y_i(0)$.
- The average outcome in the treated group is representative of what we'd see on average if everyone got treated. Same for the controls.

Can it go wrong here ?

- Imagine the cake you sister ordered is much more complex to make.
- Cake complexity will increase both the probability of buying it in the cake shop,
- And the value of the potential outcome under control ($Y(0)$).
- ==> Violation of strong ignorability.

Identification proof using SUTVA, positivity, and ignorability

- Difference in means :

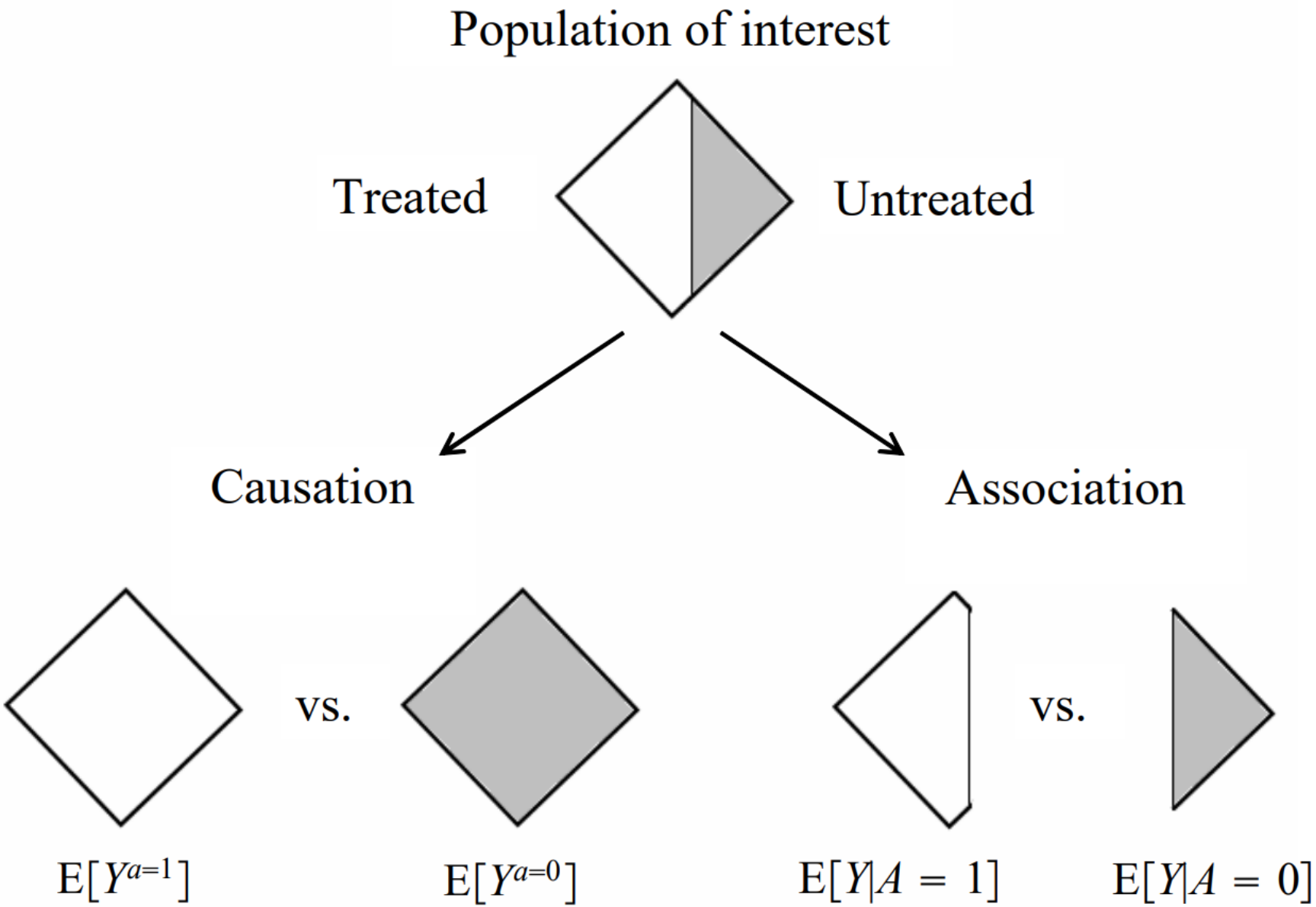
$$\mathbb{E}[Y_i|T_i = 1] - \mathbb{E}[Y_i|T_i = 0]$$

- Under **SUTVA, positivity, and ignorability**, ATE can be identified by the difference in means!
- First, the difference in means exists because of positivity (it enables conditioning)
- Then,

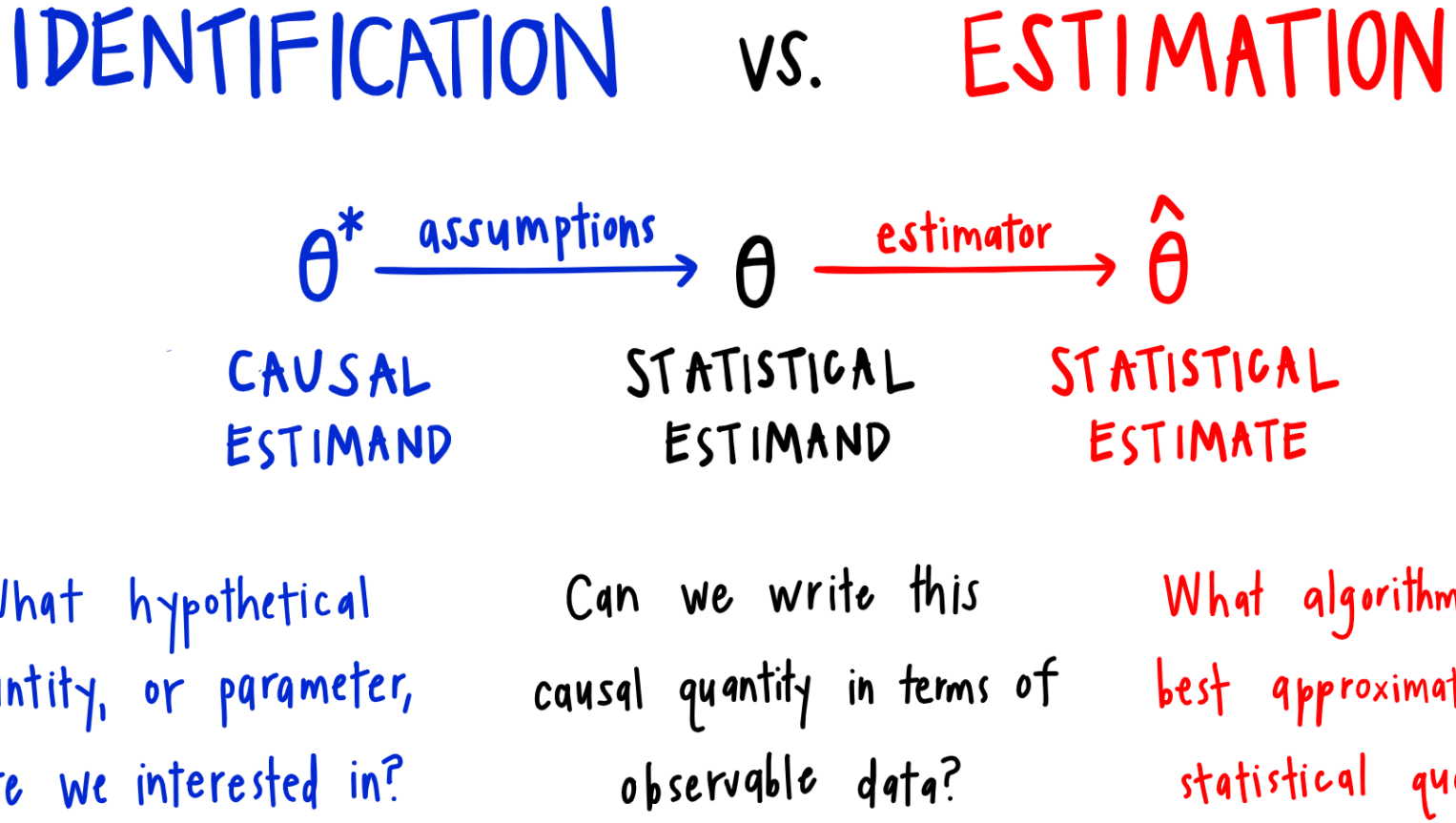
$$\begin{aligned}
\mathbb{E}[Y_i|T_i = 1] - \mathbb{E}[Y_i|T_i = 0] &= \mathbb{E}[Y_i(1)|T_i = 1] - \mathbb{E}[Y_i(0)|T_i = 0] && \text{(SUTVA)} \\
&= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] && \text{(Ignorability)} \\
&= ATE && \text{(Definition of ATE)}
\end{aligned}$$

At that point, we need to estimate $\mathbb{E}[Y_i|T_i = 1] - \mathbb{E}[Y_i|T_i = 0]$, which is a standard **statistical estimand**. We are back with classic inferential statistics (estimation)!

Association versus causation



Identification versus estimation [3]



[3] Source

Estimating the Average Treatment Effect

We need to estimate the quantities $E[Y_i|T_i = 1]$ and $E[Y_i|T_i = 0]$ with what we have in our sample. What's the most intuitive estimator? The sample means

within each subset.

$$\hat{\tau} = \underbrace{\frac{1}{n_t}\sum_{i=1}^{n} Y_i T_i}_{\text{treated sample mean}} - \underbrace{\frac{1}{n_c}\sum_{i=1}^{n} Y_i(1-T_i)}_{\text{control sample mean}}$$

- We can prove that this estimator is unbiased, consistent, and asymptotically normal.
- Several methods for statistical inference :
  - Neyman method using asymptotical normality
  - Fisher exact p-values, which relies only on assignment randomization.

How can we ensure positivity and ignorability ?

The easiest way to collect data that satisfies the assumptions for causal inference is to perform a randomized experiment.

(Randomized) Experiment

An experiment is a study in which the probability of treatment assignment $Pr(T_i = t)$ is directly under the control of a researcher.

We fix the distribution $Pr(T_i = t)$ so that it satisfies several causal assumptions:

1. **Positivity**: We make sure that each unit has a positive probability of receiving any treatment: $Pr(T_i = t) > 0$ for all t and i. Treatment is not **deterministic**.
2. **Ignorability**: We make sure that the probability of receiving treatment is independent of potential outcomes $T_i \perp\!\!\!\perp Y_i(1), Y_i(0)$

Fixing $Pr(T_i = t)$ does not garanty SUTVA!

Tools for randomized experiments

- Many things can be done with base functions (correlation, hypothetis testing)
- Also packages for more challenging settings :
  - Sample size calculations
  - Methods for non-compliance, arm switches.
  - Specific designs (sequential, two-stages, stratified)
  - Meta-analyses of randomized trials

In R, I refer you to the CRAN Task View: Clinical Trial Design, Monitoring, and Analysis In Python : experimental follow-up, overview of softare, analysis expan, ...

Contents

Why can't we always conduct a randomized experiment?

Randomized experiment is the **gold standard** : association is causation! But it is not alsways possible to conduct a randomized experiment :

- They can be unethical (e.g. effect of smoking on lung cancer)
- Need for a large sample size to assess treatment effect heterogeneity
  - Hence expensive (in terms of money, human needs, time, ...)
  - Difficult to recruit people
- Bias in population selection
  - The whole population may not be represented in the randomized experiment.
  - For instance, in medical studies, old patients, or patients with comorbidities are less represented.
- Clinical trials may lack follow-up (long-term effects?)
  - It is tedious and expensive to follow people for many years;
  - So that the treatment effect is often computed after at a pre-defined time point only (*e.g.* 1 year, 5 years)

Lots of observational data available

- There is a lot of data available other that randomized experiments.
- Which were not necessarily collected at first to answer the causal question we are trying to answer.

The problem : ignorability is very unlikely to hold. For example :

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

Can we still use this data to draw some causal conclusions? If yes, under which set of assumptions? And using which estimation methods?

There is still hope

- Our first set of assumptions is very unlikely to be satisfied
- But there exists **another set of assumptions** which enables the estimation of causal effects.
- SUTVA, **conditional** ignorability, **conditional** positivity.
- But : with slightly more complex estimation strategies.

Conditional Positivity

First we must assume that positivity holds conditionally on the covariates:

Definition

We assume that $0 < \Pr(T_i = 1|X_i = x) < 1$, for all x in the domain of X and treatment levels t.

Treatment is no deterministic in the covariates

- Knowing a unit's covariate values will never determine what treatment that unit gets with certainty
- Usually randomness comes from other factors that are not related to the outcome

Left: Non smoker and never treated

Right: Smokers and all treated

Conditional Ignorability

Second, and most important, we must assume that treatment is independent of the potential outcomes conditional on the covariates:

Definition

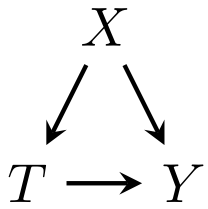
We assume that $Y_i(1), Y_i(0) \perp\!\!\!\perp T_i | X_i = x$ for all x and t.

- The covariates "tell the whole story" of the treatment assignment process.
- Within levels of X_i, treatment is assigned "as-if-random"
- There is no omitted factor that could induce confounding bias.
- Also called: No unmeasured confounding, selection on observables, no omitted variables, exogeneity, conditionally exchangeable, etc...

How to pick the good set of covariates ?

Contents

Directed Acyclic Graphs



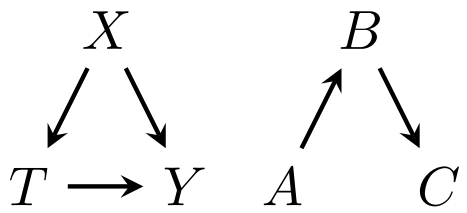
- **Nodes** represent random variables
- **Edges** between nodes denote the presence of **causal** effects (*i.e.* difference in potential outcomes). Here, $Y_i(t) \neq Y_i(t')$ for two different levels of T_i because there is an arrow from T to Y.
- **Lack of edges** between nodes denotes the **absence** of a **causal** relationship.

In practice

Before starting the analysis, draw the DAG with all the nodes and arrows you (or domain experts) think are relevant. If you do not know the DAG at all, you can try **causal discovery**.

Paths in DAGs

- A **path** between two nodes is a route that connects the two nodes following non-intersecting edges.
- A path between node A and B is such that : (i) it starts by A; (ii) it ends with B; (iii) it goes at most once through each node.
- The arrows in a path do not necessarily need to be in the good direction.



We distinguish :

- Causal paths : arrows are all in the same direction.
- From non-causal paths : arrows pointing in different directions

Open or blocked paths

A path can also be :

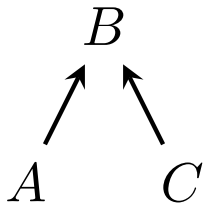
- **open** : "information" flows
- or **blocked** : "information" is blocked.

Our goal : block all non-causal paths from T to Y and let all causal paths from T to Y open.

- What we can use : we can "condition" on variables.
- When we are done (if it was possible), all the variables we have conditioned are the variables X such as conditional ignorability holds with respect to X.

Paths in DAGs

- To explain what blocked and open paths are, we need to introduce the notion of **collider**.
- A **collider** is a node of a path which "collides" two arrows.
- Path-specific!



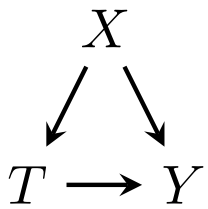
- A and C are both common causes of B.
- B is a *collider* here (two arrows "collide" into it)
- A and C are independent, but conditionally *dependent* given B.
- Conditioning on B or on a descendant of B induces a spurious correlation between A and C.

Example:

- A is result from dice 1, C is results from dice 2, B is sum of dice 1 and dice 2

Open and Blocked Paths

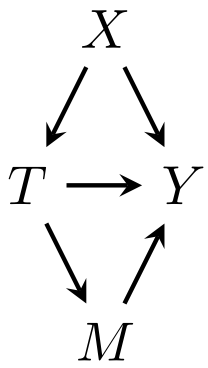
We can block or open paths between variables by conditioning on them:



A path is blocked (or d-separated) if

- The path contains a non-collider that has been conditioned on
- or the path contains a collider that has not been conditioned on (and has no descendant that has been conditioned on).

General rules



- Do not condition for variables on causal paths from treatment to outcome
- Condition on variables that block non-causal backdoor paths
  - Don't condition on colliders!

In the example above, to estimate the effect of T on Y, we should:

- Condition on X
- NOT condition on M because it is a descendant of T

Contents

What happens if you do not condition on the good variables?

- Your estimator will be biased!
- Two major types of bias may arise :
  - Confounding bias.
  - Selection bias.
- Also measurement bias

Confounding definition

- A backdoor path is a non-causal path from treatment to outcome, with an arrow pointing at treatment.

Confounding bias

Confounding bias arises when there are open **backdoor** paths between the treatment in the outcome.

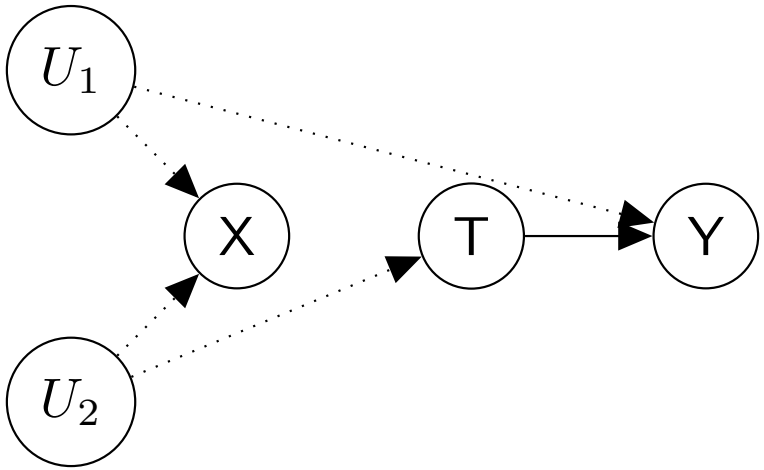
Confounder / confounding variable

A confounder is any variable that can be used to adjust for confounding bias.

A common mistake

Be careful ! Confounders are not all variables associated with both the treatment and the outcome!

- T: physical activity, Y: cervical cancer
- X: diagnostic test for pre-cervical cancer
- U_1: pre-cancer lesion, U_2: health-conscious personality (whether being in good health is of importance for you).



- No open backdoor path
- Conditioning on X (collider) will open the backdoor path and results in bias $\rightarrow X$ is not a confounder.
- But X is pre-treatment, associated with T and associated with Y.

Selection bias

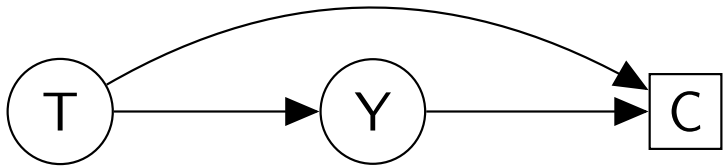
Selection bias

Bias resulting from conditioning on the common effect of two variables, one of which is either the treatment or associated with the treatment, and the other is either the outcome or associated with the outcome; thus opening a non-causal path in the graph.

- It comes from the procedure by which individuals are selected into the analysis (the design of the experiment).
- And has nothing to do with treatment assignment.
- In particular, it may happen even if treatment is randomized.

An example

- T: folic acid supplements given to pregnant women
- Y: the fetus's risk of developing a cardiac malformation
- C: fetus survived until birth
- Now let's say that the study was restricted to fetuses who survived until birth.



By conditioning on C, we opened a non-causal path from T to Y : $T \rightarrow C \leftarrow Y$. This path is not a backdoor path. But it is not part of the causal effect from T to Y neither.

Contents

Setting

- Observational dataset $(T_i, Y_i, X_i)_{i \in 1,\dots,n}$
- We assume
  - SUTVA
  - Conditional ignorability w.r.t X
  - Conditional positivity w.r.t X

  X may be multidimensional, and can include both continuous and categorical variables

There are many identification and estimation strategies here. Let's present briefly some of them.

An overview of available packages

In R

- Plenty of libraries in R
- Very nice and comprehensive CRAN task view for Causal Inference

In Python

- Only a few libraries until recently
- Several nice packages :
  - Multi-tools : causalinference, DoWhy, causalml
  - Instrumental variables : function IV2SLS from linearmodels.iv
  - Causal inference and machine learning : causalml, doubleML

Contents

(Post)-stratification (1/3)

- Conditional ignorability means that treatment assignment is as-if random within strata of the covariates.
- Within strata, we can identifiy the ATE by the difference-in-means !

$$E[Y(1) - Y(0)|X = x] = E[Y|T = 1, X = x] - E[Y|T = 0, X = x]$$

- The causal estimand $E[Y(1) - Y(0)|X = x]$ is denoted by Conditional Average Treatment Effect (CATE).
- And estimate it with the difference in sample means in treated and control within the strata

(Post)-stratification (2/3)

- And then we can aggregate the CATEs to estimate the ATE.

$$
\begin{aligned}
E[Y(1) - Y(0)] &= E[E[Y(1) - Y(0)|X]] && \text{(Law of total expectation)} \\
&= \sum_{x \in \mathcal{X}} (E[Y|T=1, X=x] - E[Y|T=0, X=x]) Pr(X=x)
\end{aligned}
$$

- Which can be easily estimated by :

$$\hat{\tau} = \sum_{x \in \mathcal{X}} \hat{\tau}(x) \times \frac{n_x}{n}$$

where $\hat{\tau}(x)$ is the estimator of the CATE (difference-in-means), n_x the number of units in strata x, and n the total number of units.

(Post)-stratification (3/3)

Pros

- Easy to understand and code, fast to run
- Non-parametric
- Statistical inference is easy (formula for variance, asymptotic normality)

Cons

- If too many strata, too few units in each strata, and high variance (curse of dimensionality)
- Infeasible with continuous covariates in X.

Contents

Intuition

- In a randomized experiment, covariate distribution are balanced across treatment groups.
- But in observational studies, covariates will be unbalanced (easy non-technical cakes will be more likely in the home-made cakes group).

Why not weighting units to balance the covariate distribution and mimmick a randomized experiment setting ?

$$\hat{\tau}_w = \frac{1}{n} \sum_{i=1}^{n} w_i (Y_i T_i - Y_i (1 - T_i))$$

Horowitz-Thompson estimator

In causal inference, the probability of receiving the treatment is commonly known as the propensity score.

Propensity Score

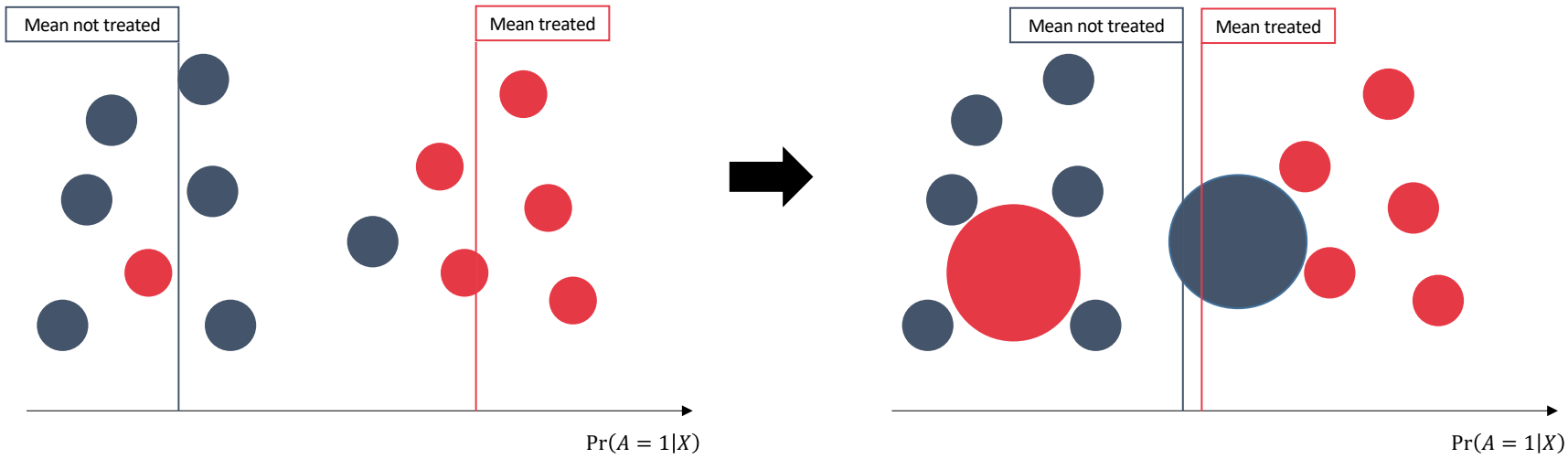
The propensity score is: $e(x) = \Pr(T = 1|X = x)$.

The most popular weighted estimator with propensity scores is the Horowitz-Thompson estimator.

- Also known as Inverse-probability-weighted estimator.
- Uses $w_i = \frac{1}{e(X_i)}$ as weights

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\frac{T_i Y_i}{e(X_i)}}_{\text{reweighted treated mean}} - \underbrace{\frac{(1-T_i)Y_i}{1-e(X_i)}}_{\text{reweighted control mean}}\right)$$

Horowitz-Thompson estimator



- In practice, we use a regularized version of the weights $w_i = \frac{Pr(T{=}t_i)}{e(X_i)}$
- Check for positivity assumption (and quality of adjustment on measured confounders) possible by **checking covariate balance after weighting**.
- Propensity scores need to be estimated (any model can fit).
- Need to encompass the variability due to the weight computation in computing variance. Easiest way : use **bootstrap**.

Weighting

Pros

- Works with high-dimensional covariates (including continuous)
- All units contribute to the final estimate
- Ajustment quality check feasible
- No use of an outcome model

Cons

- Can lead to high variance (large weights if positivity is nearly violated)
- Bootstrap to compute variance (long to run)
- Rely on the good specification of the model for the propensity scores.

Contents

Intuition

Recall the fundamental problem of causal inference: Among the treated units, we observe $Y_i(1)$ (it's Y_i), but we are missing $Y_i(0)$, and the opposite is true for the control units.

Idea: Could we find a way to impute ("fill in") the missing outcome for each unit?

Matching

For each unit, i we find the unit j with opposite treatment and most similar covariate values and use their outcome as the missing one for i.

- Since after this we have both potential outcomes for each unit, we can estimate ATE (and other causal estimands) very easily.

Matching : distance and algorithms

For each unit that we match to, i, we want to find some number (K) of units that are "close" in their covariates, but have the opposite treatment status. How do we define closeness?

- We rely on a distance metric between observations on their covariates X.
- Many choices of distance metric, which one you pick defines what you mean by "close" (euclidian on covariate values, propensity score-based, Mahalanobis, ...).
- There are two main ways to define a matched group:
  - Caliper Matching : we fix a value of γ and choose all units that have distance from i less than γ.
  - K Nearest Neighbor (KNN) matching : we fix a value of K and choose the K closest units to i.
- Many other more complex matching algorithm
  - Mixed version of caliper matching and KNN matching.
  - Choice of matched groups by optimization of a criteria : cardinality matching (Niknam and Z., 2022, JAMA)
  - Profile matching for a target population (Cohn and Z., 2022; Epidemiology)

Matching : several key remarks

- You need to choose if you let units matched to one observation be matched to another observation?
- Note that by construction, we will still have some residual imbalance between the treated and controls since the matches will not be perfect. In general, maching leads to **biased estimators** of the ATE.
- There exists procedures to remove bias, but otfen they rely on a parametric model.
- Matching can be considered as a weighting procedure, with discrete weights (number of times matched + 1).

Matching - pros and cons

Pros

- Matching is non-parametric: No need to specify a model of the data.
- Matching is interpretable: Estimates can be explained intuitively in terms of who is matched to whom.
- Matching allows to estimate many different causal quantities

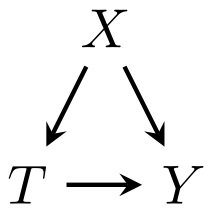
Cons

- Still a lot of parameters to set: caliper, number of neighbors, distance metric, with or without replacement
- Inexact matching is biased !

Contents

Regression-based methods : intuition

Remember that our goal was to "block" the non-causal open path from T to Y by adjusting on X.



- We've discussed weighting and matching which remove the association between X and T.
- But what if instead, we modeled the relationship between $Y_i(t)$ and X_i?

A common model : "multivariate" regression

Regression model to estimate the conditional expectation $E[Y|T, X]$. Let's assume a linear relationship :

$$E[Y_i|T_i, X_i] = \alpha + \beta T_i + X_i'\gamma$$

Can we interpret $\hat{\beta}$ as an estimate of the treatment effect?

Constant treatment effect

In order to interpret $\hat{\beta}$, the estimated coefficient on T_i as the ATE we need to assume that :

- The model is correctly specified.
- The treatment effect is constant across units: $E[Y_i(1) - Y_i(0)|X_i = x] = E[Y_i(1) - Y_i(0)|X_i = w]$ for all values of X.

- Then $\hat{\beta} \xrightarrow{p} \tau$.
- This is a strong assumption!
- What happens when it is violated?

Interaction terms

One way to deal with this problem is to add an interaction term to the regression model:

$$E[Y_i|T_i, X_i] = \alpha + \beta T_i + X_i'\gamma + T_i X_i'\lambda$$

- Now we allow different units to have different treatment effects
- Specifically, the covariates determine a unit's treatment effect
- The interaction coefficients, λ determine the impact of each covariate on the TE
- Can we interpret $\hat{\beta}$ as an estimate of the ATE in this model?
- No!!! $\hat{\beta}$ is an estimate of the Conditional Average Treatment Effect (CATE) when $X = 0$.
- The effect of the treatment on units such as all covariates are zero only!!
- In a linear model with interactions we generally have: $\tau = \beta + E[X_i]\lambda$

De-meaning covariates

A way to get $\tau = \beta$ in a linear model is to have $E[X_i] = 0$.

- We ensure this by de-meaning the covariates
- i.e., we subtract the sample average value of each covariate from each unit's value of that covariate:
- Let $\tilde{X}_{ij} = X_{ij} - \frac{1}{n}\sum_{k=1}^{n} X_{kj}$, be the de-meaned value of unit i's jth covariate
- Let $\tilde{X}_i$ be the vector of unit i's de-meaned covariates

De-meaning covariates

Then we estimate the regression:

$$E[Y_i|T_i, X_i] = \alpha + \beta T_i + \tilde{X}_i\gamma + T_i\tilde{X}_i\lambda$$

- $\beta = \tau$ in this regression
- Intuitively de-meaning ensures that $X_i = 0$ is the mean of the covariates
- Sometimes referred to as the “Lin" estimator (from a 2013 paper by Winston Lin). This is the preferred specification in Imbens and Rubin (2015) when doing a single regression to estimate a treatment effect.

Be careful with regressions

- Can we interpret the γ coefficients on X_{i1}, $X_{i2}, \ldots$ as causal effects?
- No! – we would need an entirely different set of assumptions!. Also, if T_i is a consequence of X_{i1} (as it would be if X_{i1} is a confounder), we have post-treatment bias.
- General rule: One regression, one causal effect.
- Don't try to interpret every model coefficient as causal!

In summary

If you use a standard multivariate model to infer Y with respect to T and X, be careful with interpretation!

- First, define clearly treatment and outcome.
- Then use causal DAG to know the set of covariates to adjust on.
- And discuss the other causal assumptions (conditional positivity, SUTVA).
- Fit your model.
- Can we interpret the coefficient of the treatment as a causal effect?
- Yes, if the treatment effect is constant for all units (strong assumption).
- If not, need to specify a model with the good interactions and if you de-mean your covariates, you can express the ATE with the regression coefficients (strong assumption, valid only if your model is well specified)
- **In summary, it works in some very specific cases with very strong assumptions**

Reminder: Conditional treatment effects

Our goal is to estimate the average treatment effect (ATE) τ:

$$\tau = E[Y_i(1)] - E[Y_i(0)]$$

- We assume that treatment is ignorable given X_i: $Y_i(t) \perp\!\!\!\perp T_i | X_i$.
- The CATE $\tau(x)$ can therefore be written in terms of observed Y_i (conditional difference-in-means statistical estimand):

$$\begin{aligned}\tau(x) &= E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] \\ &= E[Y_i|T_i = 1, X_i = x] - E[Y_i|T_i = 0, X_i = x]\end{aligned}$$

To get the ATE, we just aggregate up based on the density of X_i (stratification estimand, we just used the law of total probability.)

$$\tau = E[\tau(X)] = \sum_{i=1}^{n} \tau(x) \Pr(X_i = x)$$

Reminder: Stratification estimator

$$\tau = E[\tau(X)] = \sum_{i=1}^{n} \tau(x) \Pr(X_i = x)$$

With stratification estimator, we estimated :

- $\Pr(X_i = x)$ by $\frac{N_x}{N}$
- and $\tau(x)$ by conditional difference-in-means.

Limitation : If too many strata, not enough units in each strata, so not enough statistical power. A model could help us here (reduce the number of degrees of freedom)

Conditional treatment effects

- $\Pr(X_i = x)$ can be estimated $\frac{N_x}{N} = \sum_{i=1}^{n} \mathbb{1}(X_i = x)\frac{1}{n}$.
- So, $\tau = E[\tau(X)] = \sum_{i=1}^{n} \tau(x) \Pr(X_i = x)$ can be estimated by :

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \tau(X_i)$$

  We can think of this weighted average of conditional treatment effects as simply the average of the conditional treatment effects calculated at the covariate values for each unit $\tau(X_i)$
- We still need to estimate $\tau(X_i) = E[Y_i|T_i = 1, X_i] - E[Y_i|T_i = 0, X_i]$
- All we need to estimate each $\tau(X_i)$ are two conditional expectations: $E[Y_i|T_i = 1, X_i = x]$ and $E[Y_i|T_i = 0, X_i = x]$.
- Which we can estimate using outcome models (like linear regression).

Estimation

In S-learner, we use a single regression, regressing the outcome with respect to the outcome and the covariates. We denote

$$\mu(t,x) \triangleq E[Y|T=t, X=x]$$

We can then write our ATE estimator

$$\hat{\tau}_G = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}(1,X_i) - \hat{\mu}(0,X_i))$$

and our CATE estimator

$$\hat{\tau}_G(x) = \frac{1}{n_x}\sum_{i:X_i=x}(\hat{\mu}(1,x) - \hat{\mu}(0,x))$$

Implementation

- We estimate $\hat{\mu}$ by regression using all the data
- We use a Single regression (S-learner)
- $\hat{\tau}_G$ estimator is unbiased and consistent if $\hat{\mu}$ is unbiased and consistent (well specified if parametric!!)
- You can use any algorithm you want to fit the regression, however, complex machine learning algorithms require more data and will have bias due to regularization etc + overfitting problems (more on that later)
- Warning: Machine learning objective is to reduce the prediction error on Y, not to pick the right coefficient for the treatment
- In extreme cases, the algorithm could just ignore T as a variable, in particular if X is high-dimensional (e.g. think of the Lasso)

What about treatment heterogeneity?

- We fitted only one model to estimate $E[Y|T, X]$.
- If our model did not include interaction terms : it is as if we assumed constant treatment effect (no treatment heterogeneity).
- If there is treatment heterogeneity, and if we included the proper interaction terms in the model, then our estimate of an unbiased estimate of the ATE.
- Contrary to what we saw in the multivariate regression model.
- So it is already a bit more satisfying.
- The issue is that it is very hard in practice to know which interaction terms we need to use.
- T-learner can alleviate this issue.

Intuition

Can we do a bit better? Yes, by fitting two different models!

- Instead of fitting a single model to estimate $E[Y|T, X]$.
- We fit one model for the treated $E[Y|T = 1, X]$;
- And one model for the control $E[Y|T = 0, X]$.
- The treatment effect heterogeneity will be encompassed in the differences between the two models.
- If there is no treatment heterogeneity, this is pointless (the two models will be excactly the same).
- But in theory not harmless;
- In practice though it could be, since the models are fitted on less samples (model 1 fitted on the treated only; model 2 fitted on the control only).

Formalization

Again, we want to estimate the CATE $\tau(x)$, written in terms of observed Y_i

$$\begin{aligned}\tau(x) &= E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] \\ &= E[Y_i|T_i = 1, X_i = x] - E[Y_i|T_i = 0, X_i = x]\end{aligned}$$

To get the ATE, we just aggregate up based on the density of X_i

$$\tau = E[\tau(X)] = \sum_{i=1}^{n} \tau(x) \Pr(X_i = x)$$

We denote

$$\mu_t(x) \triangleq E[Y|T = t, X = x]$$

We can then write our ATE estimator

$$\hat{\tau}_G = \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$$

and our CATE estimator

$$\hat{\tau}_G(x) = \frac{1}{n_x} \sum_{i:X_i=x} (\hat{\mu}_1(x) - \hat{\mu}_0(x))$$

Practical implementation

The regression imputation estimator is:

$$\hat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i(1) - \hat{Y}_i(0)$$

If the two models are well specified, then $E[\hat{\tau}_{\text{reg}}] = \tau$ and $\hat{\tau}_{\text{reg}} \xrightarrow{p} \tau$, i.e. the regression imputation estimator is unbiased and consistent.

Variance: what about its variance? If the two models are well specified, then we also have:

$$\sqrt{n}(\hat{\tau}_{\text{reg}} - \tau) \xrightarrow{d} \mathcal{N}(0, Var[\hat{\tau}_{\text{reg}}])$$

- Which means that we can conduct approximate HP tests and construct CIs as we have been doing.
- But we need to estimate $Var[\hat{\tau}_{\text{reg}}]$ to use this!

Variance

How do we estimate the variance of the regression imputation estimator?

- In theory: it is possible to work out and implement a formula for the variance
- In practice: we can just use the bootstrap!

Double robustness

- Most estimators we saw so far (IPTW, S-learner, T-learner) are sensitive to model misspecification
- IPTW was sensitive to the model we used to estimate the propensity scores;
- S-learner and T-learner were sensitive to the model(s) we used to estimate the outcome;
- The idea of Doubly robust estimators is to combine them and create an estimator which is consistent if at least one of the models is well-specified
- Rationale: makes group more similar (like in IPTW) before modeling the outcome (like in S and T-learner).
- In practice, we combine regression estimate with an inverse propensity weighted estimate of the regression residuals

Augmented IPW: a doubly robust estimator

We define

- $\mu_{(t)}(x) := \mathbb{E}[Y_i(t) \,|\, X_i = x]$ the outcome models (like in S and T-learner);
- $e(x) = P(T_i = 1 | X_i = x)$ the propensity score (like in IPTW).
- the Augmented IPW (AIPW) estimator can be written :

Augmented IPW: a doubly robust estimator

$$\hat{\tau}_{AIPW} := \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + T_i \frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - T_i) \frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)} \right)$$

is consistent if **either** the $\hat{\mu}_{(w)}(\cdot)$ are consistent **or** $\hat{e}(\cdot)$ is consistent.

How is double robustness working

We can reorganize the AIPW estimator as

$$\hat{\tau}_{AIPW} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}\left(X_i\right) - \hat{\mu}_{(0)}\left(X_i\right)\right)}_{\text{a consistent treatment effect estimator}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\hat{e}\left(X_i\right)}\left(Y_i - \hat{\mu}_{(1)}\left(X_i\right)\right) - \frac{1-T_i}{1-\hat{e}\left(X_i\right)}\left(Y_i - \hat{\mu}_{(0)}\left(X_i\right)\right)\right)}_{\approx \text{ mean-zero noise}},$$

and in that case the propensity score is not really used in estimation

How is double robustness working

And we can also rearrange terms differently

$$\hat{\tau}_{AIPW} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\hat{e}\left(X_i\right)} - \frac{\left(1-T_i\right)Y_i}{1-\hat{e}\left(X_i\right)}\right)}_{\text{the IPW estimator}}$$

$$+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}\left(X_i\right)\left(1-\frac{T_i}{\hat{e}\left(X_i\right)}\right) - \hat{\mu}_{(0)}\left(X_i\right)\left(1-\frac{1-T_i}{1-\hat{e}\left(X_i\right)}\right)\right)}_{\approx \text{ mean-zero noise}},$$

and in that case the outcome models are not really used in estimation

Contents

Let's go a bit further

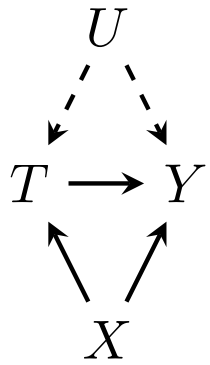
- Even with double robustness, we need at least one model to be **correctly specified**.
- For now, we talked only about simple models such as linear regressions.
- It is very unlikely that they represent the true data-generating process.
- Why not use more complex machine learning models?
- Several challenges arise here :
  - Machine learning models are used as predictive models in general. Because of regularization strategies (to avoid overfitting) the treatment may even not be encompassed in the model.
  - Few results on the properties of machine learning estimators : variance ? root-n consistency ? How to get confidence intervals ?
- Also, why not trying to learn the "best" models? (Maximum likelihood theories, Newton-Raphson algorithm)

(Roughly) Two major methods arise here : targeted minimum loss estimator (TMLE), and double Machine Learning (Double ML).

Contents

What can we do if we know/suspect some unmeasured confounding ?

- Conditional ignorability does not hold with respect to X.
- But we assume that there is an unmeasured variable U such that ignorability holds with respect to X and U



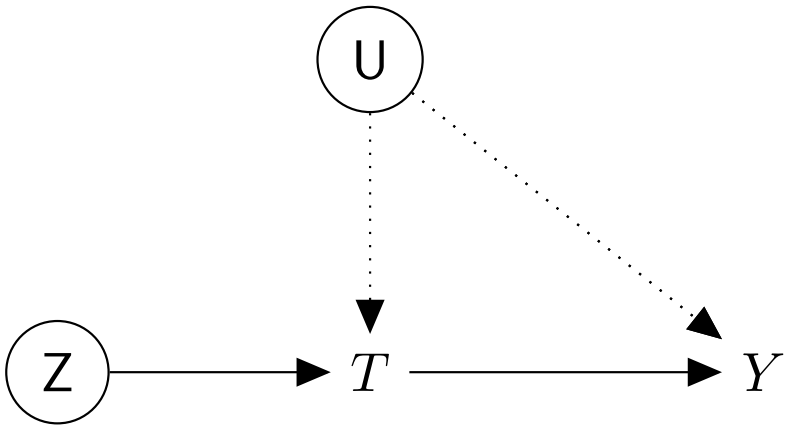
Two possibilities here :

- In certain settings, identification is still possible (but very specific cases) : instrumental variables, regression discontinuity designs, difference-in-differences.
- Otherwise : sensitivity analyses.

Contents

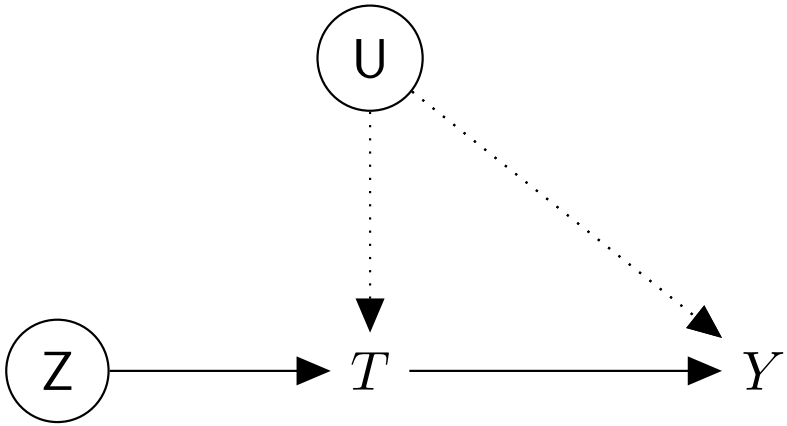
Instrumental variables (1/2)

Imagine that you have a covariate Z such that three assumptions hold:



- Relevance : Z has a causal effect on T.
- Exclusion restriction : The causal effect of Z on Y is fully mediated by T.
- Instrumental unconfoundedness : The relationship between Z and Y is unconfounded or confounded only by variables we measure and can adjust on.

Instrumental variables (2/2)



Then, if we assume further than either we are in a linear model, or that there is no defier (monotonicity assumption, $\forall i, T_i(Z=1) \geq T_i(Z=0)$), we can estimate the Local Average Treatment Effect (LATE):

LATE non parametric identification

$$E[Y(Z=1) - Y(Z=0)|T(1)=1, T(0)=0] = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[T|Z=1] - E[T|Z=0]}$$

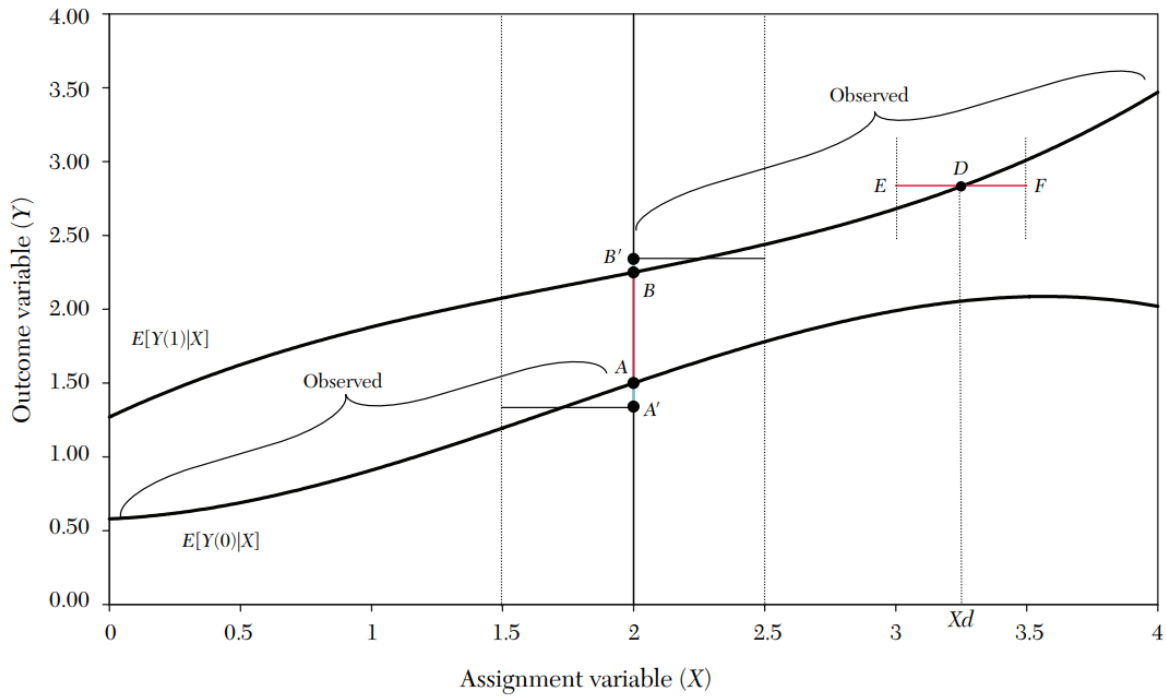
Contents

Regression discontinuity designs (1/2)

Treatment is sometimes assigned based on a **cut-off or threshold value** of a continuous running variable (merit aid scholarships awarded based on a test-score cut-off). Intuition : treatment is **quasi-randomized** at the cut-off.

Two assumptions here :

- Sharp RD : Treatment assignment is perfectly determined by the value of the forcing variable X_i and the threshold c.
- Continuity in potential outcomes : $E[Y_i(0)|X_i = x]$ and $E[Y_i(1)|X_i = x]$ are continuous at $x = c$ (at the discontinuity point).



Regression discontinuity designs (2/2)

Identification at the cut-off point

$$E[Y_i(1)|X_i = c] - E[Y_i(0)|X_i = c] = \lim_{x \to c+} E[Y_i|X_i = x] - \lim_{x \to c-} E[Y_i|X_i = x]$$

In reality, RDD set-up is equivalent to an instrumental variables design where the instrumental variable is the indicator for being above or below the cut-off.

Contents

Difference-in-differences (1/2)

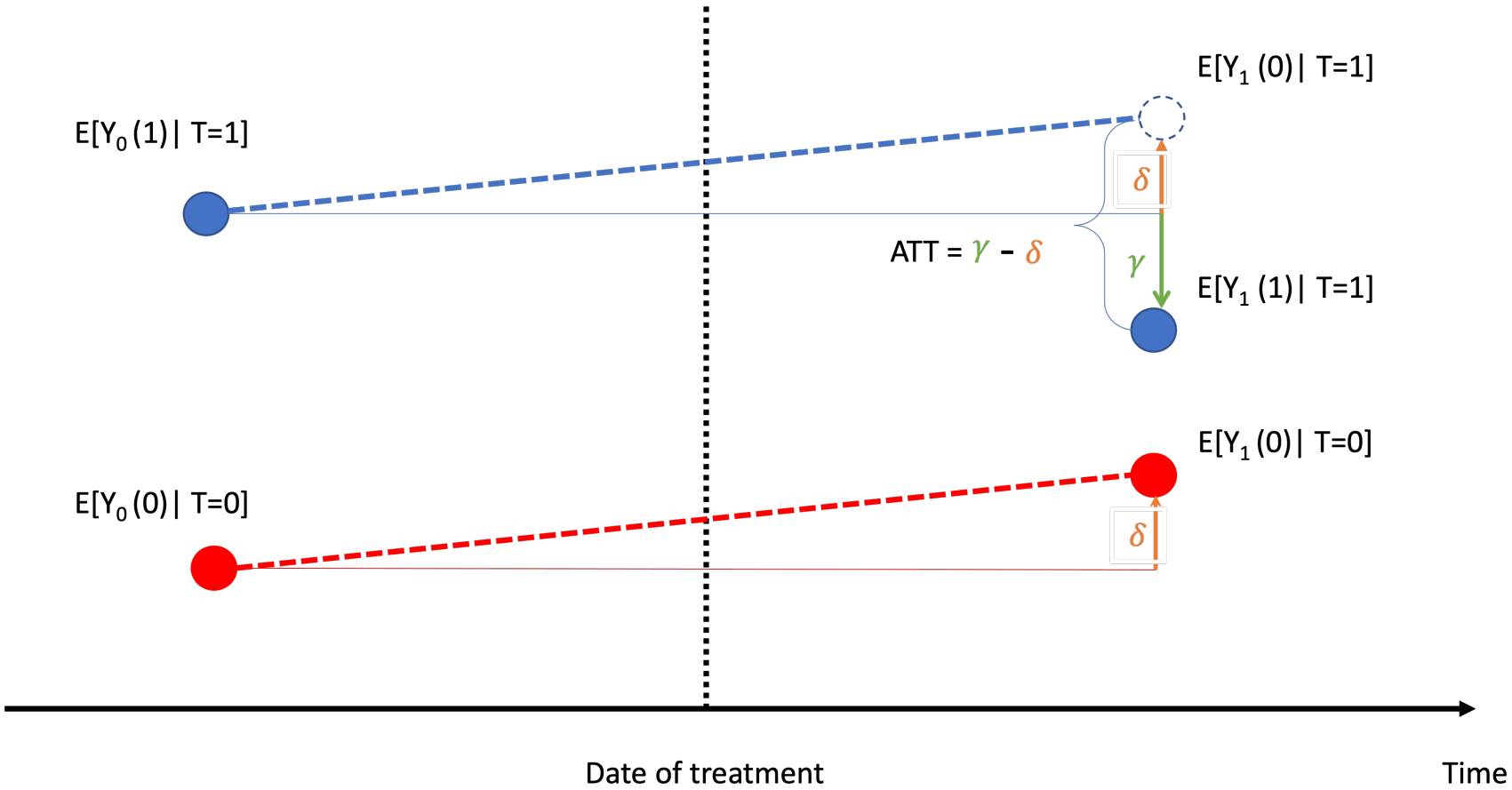
- Imagine now that we have two time points : outcome is measured both before and after the treatment is applied.
- And two groups : a control group and a treated group (treated only at the second time point).

Imagine that in 2020, France was cut into two. Half of France is put under lockdown (North). The other half : no lockdown (South). Can we infer the value of covid cases we would have observed in the treatment group if no lockdown ?

Yes, under strong assumptions

Difference-in-differences (2/2)

- Assumption 1 : SUTVA at both tine points.
- Assumption 2 : Parallel trends assumption. $(Y_1(0) - Y_0(0)) \perp\!\!\!\perp T$
- Assumption 3 : No pre-treatment effect assumption.
  $E[Y_0(1)|T = 1] = E[Y_0(0)|T = 1]$
- We can identify the Average Treatment Effect in the Treated (ATT).

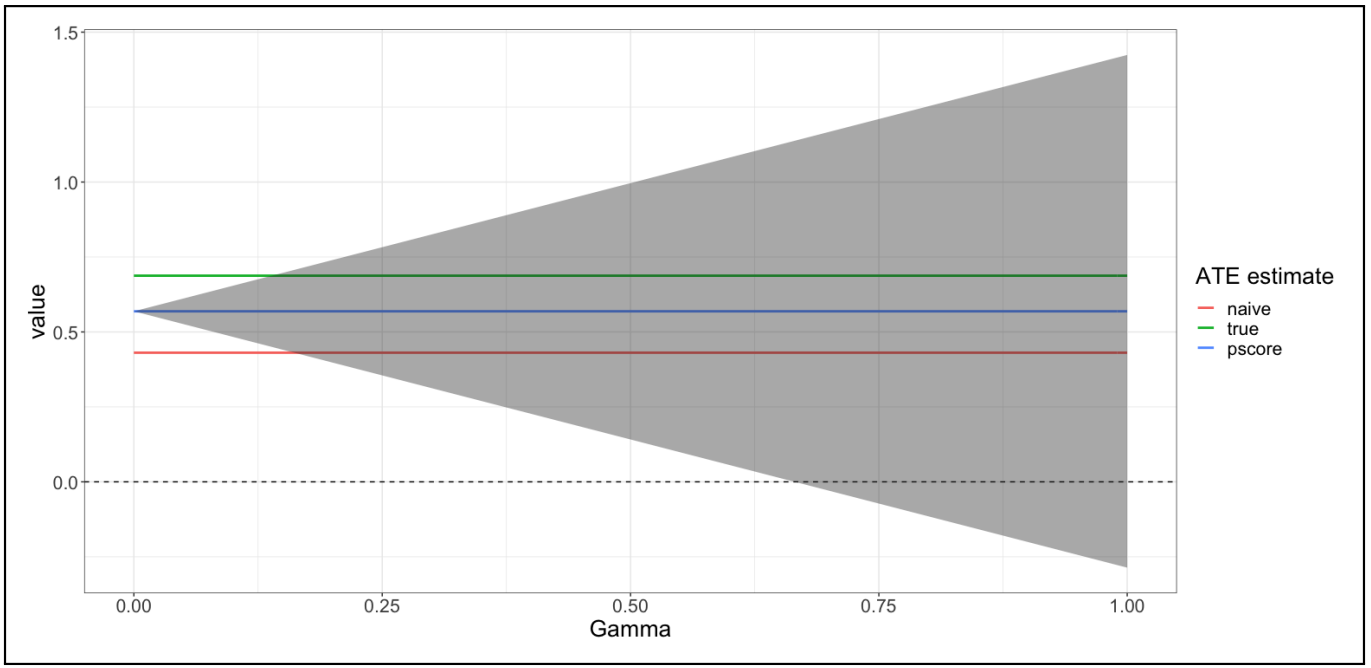


Contents

Sensitivity analyses

Very briefly here :

- Even if we assume unmeasured confounding, we can estimate the extent of bias due to the unmeasured confounder(s).
- By finding bounds for the ATE (*e.g.* Manski bounds).
- Or by estimating the association needed between U and T or Y to cancel the significance of the point estimate.



References and acknowledgement

Several good references

- Causal Inference : What if? (Hernán and Robins)
- The book of why (Pearl)
- Causal inference for statistics, social, and biomedical sciences (Imbens and Rubin)
- Targeted learning (van der Laan, Rose)
- Explanation in Causal Inference: Methods for Mediation and Interaction (VanderWeele)

Slides with some inspirations from

- Judith Abécassis
- Marco Morucci
- Mats Julius Stensrud
- Julie Josse

Any question?