Causal Inference

🪀 Paper: Counterfactual Fairness

🧴 Lab review

Neyman–Rubin causal model. We focus on the framework of potential outcomes.

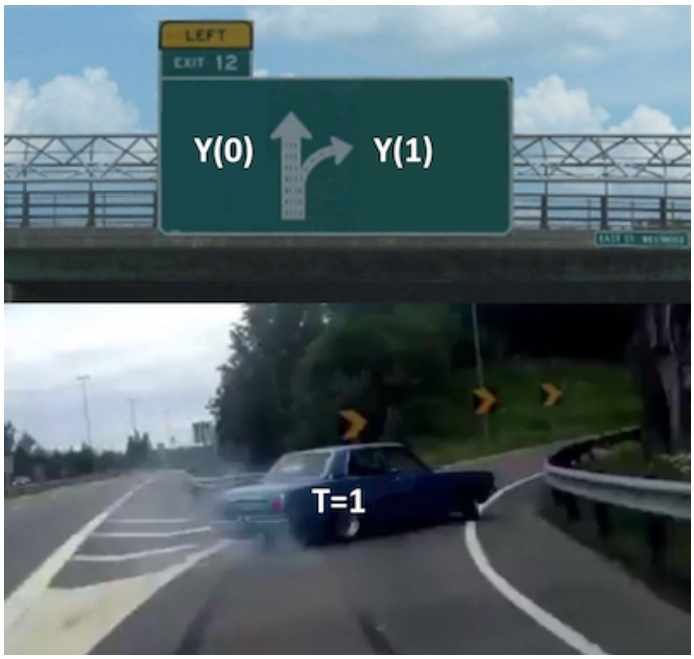
ML: good at prediction, bad at inference the counterfact

association is not causation: Bias is what makes association different from causation.

The covariates influence the treament! Example of tablets used in the school.

fundamental problem of causal inference: we can not observe all potential outcomes (with and without treatment).

The act of setting the treatment to 0 or 1 materializes one of the potential outcomes and makes it impossible for us to ever know the other one.



The potential output

$$Y_i(0), Y_i(1)$$

does not depend on the treatment. Whether or not receive treatment, the potential outcome remains the same, for one individual i.

If treatment, we would observe Y_i(1); if not, we observe Y_i(0): **Consistency assumption**

average treatment effect (ATE)

$$ATE = E[Y_1 - Y_0]$$

$$\underbrace{E[Y|T=1] - E[Y|T=0]}_{ATE \text{ (under consistency assumption)}} = \underbrace{E[Y_1 - Y_0|T=1]}_{ATT} + \underbrace{\{E[Y_0|T=1] - E[Y_0|T=0]\}}_{BIAS}$$

ATE is what we are interested in.

Task of causal inference: finding clever ways of **removing bias and making the treated and the untreated comparable** so that all the difference we see is only the average treatment effect.

A first tool we have to make the bias vanish: **Randomised Controlled Trials (RCT).**

It randomly assigns individuals in a population to a treatment or to a control group, which makes the **potential outcomes** to be **independent** of the treatment

随机分配 treatment，potential outcome 当然也就独立于 treatment

BUT: the **outcomes is still dependent** of the treatment (of course, it should be).

$$(Y_0, Y_1) \perp\!\!\!\perp T$$

Thus, in expectation, the **potential outcomes** are the same in the treatment or the control group. The only thing generating a difference between the outcome in the treated and in the control group.

$$E[Y_0|T=0] = E[Y_0|T=1] = E[Y_0]$$
$$E[Y|T=1] - E[Y|T=0] = E[Y_1 - Y_0] = ATE$$

a simple difference in means between treatment and control is thus the treatment effect.

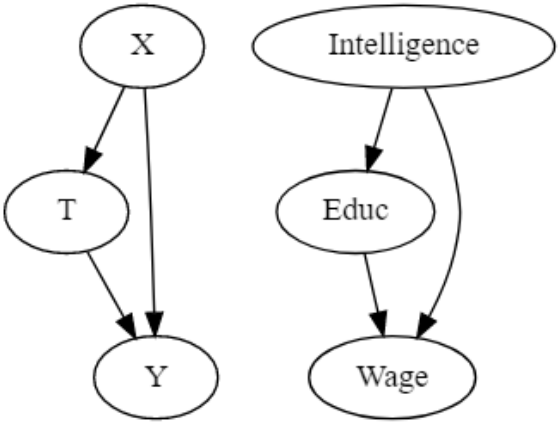
BUT, RCT tend to be either very expensive or just plain unethical. Sometimes, we simply can't control the assignment mechanism. 抽烟对孕妇和婴儿的影响，不能随机地要求一部分孕妇去吸烟；研究最低工资对失业率的影响，不能随机地要求一部分国家有最低工资，另一部分国家没有

RCT —> Observational study (draws inferences from a sample to a population where the independent variable is **not under the control of the researcher** because of ethical concerns or logistical constraints)

Conditional Independence:

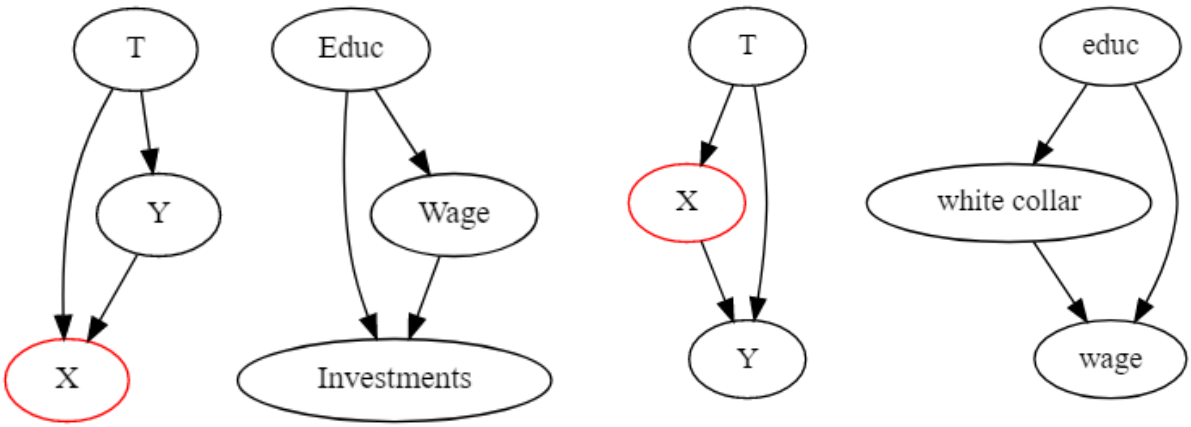
$$(Y_0, Y_1) \perp T|X$$

The first significant cause of bias is confounding. It happens when the treatment and the outcome share a common cause. To eliminate the confounders, we can condition on them.



Sometimes, we can't control the confounder because it is unmeasurable (eg. Intelligence). However, we have other measured variables that can act as a proxy for the confounder. Those variables are not in the backdoor path, but controlling for them will help lower the bias (but it won't eliminate it). Those variables are sometimes referred to as **surrogate confounders**.

The second significant source of bias is what we will call selection bias. It may due to conditioning on a common effect, or due to excessive controlling of mediator variables.



Propensity score:

you don’t have to condition on the entirety of X to achieve independence of the potential outcomes on the treatment. It is sufficient to condition on this single variable, which is the propensity score:

$$e(x) = P(T_i = 1|X_i = x)$$

$$(Y_1, Y_0) \perp T|e(x)$$

It is like a kind of dimensionality reduction on the feature space.

If someone has a low probability of treatment, that individual looks like the untreated. However, that same individual was treated. This must be interesting. We have a treated that looks like the untreated, so we will give that entity a high weight. This creates a population with the same size as the original, but where everyone is treated.

positivity assumption of causal inference:

$$\eta < e(x) < 1 - \eta, \text{ for some } \eta > 0$$

$$\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|X] - \mathbb{E}[Y_i(0)|X_i]] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[T_i|X_i] \cdot \mathbb{E}[Y_i(1)|X]}{e(X_i)} - \frac{\mathbb{E}[1 - T_i|X_i]_i(0)|X_i]}{1 - e(X_i)}\right] \text{ def. of } e(x) \\
&= \mathbb{E}\left[\frac{\mathbb{E}[T_i \cdot Y_i(1)|X_i]}{e(X_i)} - \frac{\mathbb{E}[(1 - T_i) \cdot Y_i(0)|X_i]}{1 - e(X_i)}\right] \text{ uncounfoundness} \\
&= \mathbb{E}\left[\frac{T_i \cdot Y_i}{e(X_i)} - \frac{(1 - T_i) \cdot Y_i}{1 - e(X_i)}\right]
\end{aligned}$$

To **estimate Propensity score**, one common way of doing so is using logistic regression, but other machine learning methods, like gradient boosting, can be used as well.

By producing weights 1/e(X), it creates the population where everyone is treated and by providing the weights 1/(1−e(X)), it creates the population where everyone is untreated.

the estimation of the propensity scores helps to remove confounders such that a better balance between treated and control units is achieved.

Inverse Propensity Weighting (IPW) estimator of ATE:

$$\hat{\tau}_{IPW} = \frac{1}{n} \sum_i [\frac{T_i Y_i}{e(X_i)} - \frac{(1 - T_i)Y_i}{1 - e(X_i)}]$$

Plug-in estimator: estimate Propensity score and inject into IPW estimator.

Note: **The variance of the IPW estimator is relatively large. Its variance depends of the inverse of the propensity score**. Consequently, **when overlap is not large or when some probabilities to be treated or non treated are close to 1, the estimator is unstable**. Normalized IPW can be an alternative to stabilize the estimator.

Linear regression to estimate ATE:

We assume:

$$Y_i(t) = c(t) + X_i\beta(t) + \epsilon_i(t) \text{ where } \mathbb{E}(\epsilon_i(t)|X_i) = 0, Var(\epsilon_i(t)|X_i) = \sigma^2$$

We note:

$$\hat{\mu}_t(X_i) = \mathbb{E}[Y_i(t)|X_i] = c(t) + X_i\beta(t)$$

And we estimate ATE by linear regression:

$$\hat{\tau}_{OLS} = \frac{1}{n}\sum_i(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_1)) = \frac{1}{n}\sum_i[(\hat{c}_1 + X_i\hat{\beta}_1) - (\hat{c}_0 + X_i\hat{\beta}_0)]$$

> Notice: When the models are **mispecified** (identifiability assumptions are not met **as we do not include all the confounders**), there is a **large bias for OLS and IPW estimators.**

Doubly Robust Estimation

It is a way of combining propensity score and linear regression to estimate ATE. It only requires one of the models to be correct.

Thus, 2 regressions to be done: propensity score over (subset of) covariates; output over (subset of) covariates.

$$\hat{\tau} = \frac{1}{N}\sum\left(\frac{T_i(Y_i - \hat{\mu_1}(X_i))}{\hat{e}(X_i)} + \hat{\mu_1}(X_i)\right) - \frac{1}{N}\sum\left(\frac{(1-T_i)(Y_i - \hat{\mu_0}(X_i))}{1-\hat{e}(X_i)} + \hat{\mu_0}(X_i)\right)$$

- If linear regression is correct:

Assume that $\hat{\mu_1}(x)$ is correct. If the propensity score model is wrong, we wouldn't need to worry. Because if $\hat{\mu_1}(x)$ is correct, then $E[T_i(Y_i - \hat{\mu_1}(X_i))] = 0$. That is because the multiplication by T_i selects only the treated and the residual of $\hat{\mu_1}$ on the treated have, by definition, mean zero. This causes the whole thing to reduce to $\hat{\mu_1}(X_i)$, which is correctly estimated $E[Y_1]$ by assumption. So, you see, that by being correct, $\hat{\mu_1}(X_i)$ wipes out the relevance of the propensity score model. We can apply the same reasoning to understand the estimator of $E[Y_0]$.

- If propensity score is correct: (the notation is slightly different)

Now, assume that the propensity score $\hat{P}(X_i)$ is correctly specified. In this case, $E[T_i - \hat{P}(X_i)] = 0$, which wipes out the part dependent on $\hat{\mu_1}(X_i)$. This makes the doubly robust estimator reduce to the propensity score weighting estimator $\frac{T_i Y_i}{\hat{P}(X_i)}$, which is correct by assumption. So, even if the $\hat{\mu_1}(X_i)$ is wrong, the estimator will still be correct, provided that the propensity score is correctly specified.

> More often, what ends up happening is that neither the propensity score nor the outcome model are 100% correct. They are both wrong, but in different ways. When this happens, it is not exactly settled if it's better to use a single model or doubly robust estimation.

Heterogeneous Treatment Effects and Personalization

The treatment might be beneficial to treat one unit but not another. we want to estimate the Conditional Average Treatment Effect (CATE):

$$\tau(x) = E[Y_1 - Y_0 | X = x]$$

Estimation of CATE:

- Modified ML models. Example: **Causal Forest**
- Model-free approches: **Meta-learners**