



Chizuru7

A Junior in XMU

Earth, Solar System

文章

6

分类

0

标签

1



9个月前发表 9个月前更新 JUNWEI LIN 19分钟读完(大约2778个字)

Causal Effect Estimation

Causal effect is different from **associative relationship**. In reality, we are usually interested in how an action matters. For example, it is obvious that a high advertisement payment appeals to consumers, which is **associative relationship**. However, how to allocate your advertisement money, like focusing on young people, will get the best effect? It requires the **causal effect**.

Shallow men believe in luck or in circumstance. Strong men believe in cause and effect.

—Ralph Waldo Emerson

Prerequisites

Structural Causal Model

Like statistical learning, we'd better use a concise "language" to tell the causal effect. And here is the **structural causal model** (SCM). A SCM can always be drawn within a **Directed Acyclic Graph** (DAG). One reason is that the relationship doesn't have feedback effect, with which SCM and DAG are consistent.



Fig1: A Simple SCM

In Fig1, it shows a simple causal relationship that X affects Y , and X is the cause. The causal effect can be written as,

$$Y \leftarrow f_X(X)$$

In general, X is the treatment binary variable to be 0 or 1. Therefore, the causal effect is,

$$\tau = Y_{X=1} - Y_{X=0} \quad (1)$$

Counterfactual

最新文章



2021-06-18
Casual Effect Estimation



2021-05-07
Ridge and Lasso Regression



2021-05-04
Challenge: Ordered Logit



2021-03-31
Report - Regression Splines



归档

六月 2021

1

五月 2021

2

三月 2021

3

标签

微观计量 6

订阅更新

链接

Hexo

hexo.io

Bulma

bulma.io

Observed Data from the Real world



No data from the Counterfactual world



Fig2: Counterfactual

It leads a severe problem that we cannot estimate the treatment effect (causal effect) for one specific individual because the equation 1 requires us to analyze some variations in different worlds.

Potential Outcomes Model

Potential Outcomes Model combines the treatment, potential outcomes, observable outcome and treatment effect into one equation, which is,

$$Y_i = X_i Y_i^1 + (1 - X_i) Y_i^0 \quad (2)$$

, where X_i is the treatment. Y_i^1 denotes the outcome after treatment while Y_i^0 doesn't. Therefore, Y_i is the final outcome observed. That is, if a treatment is taken, we will get $Y_i = Y_i^1$.

Confounding Variables

Some variables may affect treatment and outcomes simultaneously. One DAG is,

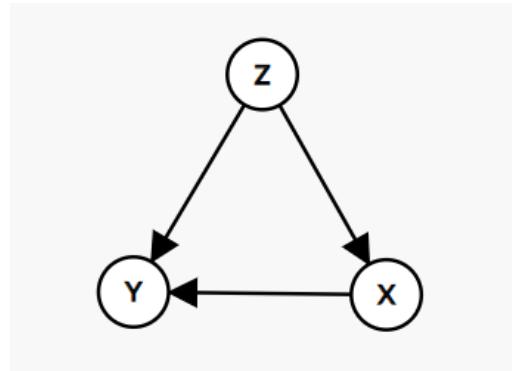


Fig3: Confounding Variables

We call Z blocks every path $X \rightarrow Y$. So variable Z makes variable X and Y **dependent**. X denotes "going to university", Y is salary and Z is gender. Thus, we cannot observe the true treatment effect ($X \rightarrow Y$) by only observing X, Y . This kind of variables called **confounding variables**.

Average Treatment Effect

- No Confounding Variables here.

We'd like to change our mind to estimate the **average treatment effect (ATE)**.

$$\begin{aligned} \tau = ATE &= E(Y_i^1 - Y_i^0) \\ &= E(Y_i^1 - Y_i^0 | X_i = 1) * p(X_i = 1) + E(Y_i^1 - Y_i^0 | X_i = 0) * p(X_i = 0) \end{aligned} \quad (3)$$

Suppose the sample here is created by the God, and individuals are homogeneous.

	Y_i^1	Y_i^0	Treatment
1	1	0	0
2	1	0	0
3	1	1	0
4	0	0	0
5	1	1	1
6	1	0	1
7	1	0	1
8	0	0	1

The ATE is,

$$\begin{aligned}\tau = ATE &= E(Y_i^1 - Y_i^0 | X_i = 1) * p(X_i = 1) + E(Y_i^1 - Y_i^0 | X_i = 0) * p(X_i = 0) \\ &= \frac{(1 - 0) + (1 - 0) + (1 - 1) + (0 - 0)}{4} * \frac{1}{2} + \frac{(1 - 1) + (1 - 0) + (1 - 0) + (0 - 0)}{4} \\ &= \frac{1}{2}\end{aligned}$$

For this sample, if we only treat four individuals, the *ATE* will not change because we know the outcomes before and after treatment.

ATU & ATT & SDO

- **ATU** is the shorthand of **average treatment effect of the untreated**.

$$ATU = E(Y_i^1 - Y_i^0 | X_i = 0)$$

- **ATT** is the shorthand of **average treatment effect of the treated**.

$$ATT = E(Y_i^1 - Y_i^0 | X_i = 1)$$

- **SDO** is the shorthand of **simple difference in mean outcomes**.

$$SDO = E(Y_i^1 | X_i = 1) - E(Y_i^0 | X_i = 0)$$

In the sample, we can easily calculate,

$$\begin{aligned}ATU &= \frac{3 - 1}{4} = \frac{1}{2} \\ ATT &= \frac{3 - 1}{4} = \frac{1}{2} \\ SDO &= \frac{3}{4} - \frac{1}{4} = \frac{1}{2}\end{aligned}$$

Since *SDO* is **not affected** by counterfactual to be accurately computed, we wonder whether it can be used to estimate *ATE*. More importantly, we can do a simple transposition for *ATE* based on equation 3, (let $p = p(X_i = 1)$)

$$\begin{aligned}ATE &= ATT * p + ATU * (1 - p) \\ &= ATT + (ATU - ATT) * (1 - p) \\ &= SDO + E(Y_i^0 | X_i = 0) - E(Y_i^0 | X_i = 1) + (1 - p)(ATU - ATT) \\ \Rightarrow SDO &= ATE + \underbrace{E(Y_i^0 | X_i = 1) - E(Y_i^0 | X_i = 0)}_{\text{Selection Bias}} + \underbrace{(1 - p)(ATT - ATU)}_{\text{Heterogeneous Bias}}\end{aligned} \quad (5)$$

So how the two terms influence the estimation?

Selection Bias

If we intentionally assign our treatments to individual 1,3,5,7. The sample shows below,

	Y_i^1	Y_i^0	Treatment
1	1	0	1
2	1	0	0
3	1	1	1
4	0	0	0
5	1	1	1
6	1	0	0
7	1	0	1
8	0	0	0

If we calculate SDO ,

$$SDO = \frac{1+1+1+1}{4} - \frac{0+0+0+0}{4} = 1$$

Simply regrading it as ATE , then a huge effect reflects if a treatment is implemented. However, the true ATE is $\frac{1}{2}$. The selection bias is,

$$\text{Selection Bias} = \frac{0+0+1+1}{4} - \frac{0+0+0+0}{4} = \frac{1}{2}$$

What exactly contributes to the **Selection Bias**? The answer is Y_i^0 . There is a probability that treated ones and untreated ones make different choices if no treatment assigned. Say, **in order to ensure my estimated ATE is not distorted due to sampling bias, I must ensure treatment assignment strategy does not yield a significant difference in the potential outcome given no treatment (Y_i^0) of treated and untreated individuals.**

Confounders are generally the main cause of selection bias.

Heterogeneous Treatment Effect Bias

If there are two economists, the God surprisingly finds that the treatment has no effect on them, and gives the sample below,

	Y_i^1	Y_i^0	Treatment
1	1	0	1
2	1	0	0
3	1	1	0
4✓	0	0	1
5	1	1	1
6	1	0	0
7	1	0	0
8✓	0	0	1

, where individuals 4, 8 are economists. Calculating the SDO ,

$$SDO = \frac{1+1+0+0}{4} - \frac{0+1+0+0}{4} = \frac{1}{4}$$

While with the selection bias is,

$$\text{Selection Bias} = \frac{1+0+0+0}{4} - \frac{0+1+0+0}{4} = 0$$

The heterogeneous treatment effect bias is,

$$\begin{aligned} HTE &= (1 - \frac{1}{2}) * (ATT - ATU) \\ &= \frac{1}{2} * (\frac{1}{4} - \frac{3}{4}) = -\frac{1}{4} \end{aligned}$$

What exactly causes **Heterogeneous Treatment Effect (HTE)**? The answer is Y_i^1 . When there is no sampling bias, Y_i^1 is the main factor that influences ATT and ATU because the group of the treated and the untreated may not be the same. In this case, economists, who are not affected by the treatment, are all in the treated. In order to ensure my estimated ATE is not distorted due to sampling bias, I must ensure that my treatment assignment strategy does not yield a significant difference in the potential outcome given treatment (Y_i^1) of treated and untreated individuals.

Randomized Controlled Trials

How to eliminate those bias? Applying **Randomized Controlled Trials** is a good idea. Then, treatments are **independent** with potential outcomes Y_i^1 and Y_i^0 . We will have,

$$\begin{aligned} E(Y_i^1|X_i=1) &= E(Y_i^1|X_i=0) \\ E(Y_i^0|X_i=1) &= E(Y_i^0|X_i=0) \\ \tau = ATE &= E(Y_i^1 - Y_i^0) = E(Y_i^1|X_i=0/1) = ATT = ATU \end{aligned}$$

ATE under Confoundedness

If there is observable confounding variables, how to estimate **ATE**? There are two criteria to figure out the **ATE**.

Back-door Criterion

Backdoor Criterion is simply controlling confounding variables **constant** (or nearly equal to the constant). By doing so, we will reach a magic place where treatment and outcome are **independent**. Remember the Fig3, the variable Z is ‘invisible’ in this case. Of course, we call it as **conditional independent**. An efficient method is [matching](#).

As we do above, we estimate **conditional ATE (CATE)**,

$$CATE = E(Y_i^1 - Y_i^0|Z) \quad (6)$$

Then the **ATE** is,

$$ATE = \sum CATE_{Z_i} p(Z_i)$$

If there are many confounding variables, controlling all confounding variables are difficult, which is known as [curse of dimensionality](#). Under this circumstance, we can apply [propensity score matching](#), which compresses variables into one score and all we need to do is to control the score.

Front-door Criterion

Another efficient way is **Front-door Criterion**. Based on Fig3, we add M between X and Y ,

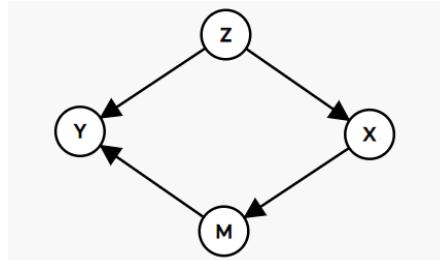


Fig4: Front-door Criteria

M is a mediator, then the causal effect $X \rightarrow Y$ is the combination of $X \rightarrow M$ and $M \rightarrow Y$. Here, we can easily control X to control M . As a result, the influence of Z is eliminated.

$$ATE_{X \rightarrow Y} = ATE_{X \rightarrow M} * ATE_{M \rightarrow Y}$$

Front-door criteria is a very useful way because I can even have no idea about Z !

Causal Regression

Based on **Potential Outcomes Model**, we make a transposition on equation 2,

$$\begin{aligned} Y_i &= X_i Y_i^1 + (1 - X_i) Y_i^0 \\ &= X_i (Y_i^1 - Y_i^0) + Y_i^0 \end{aligned} \quad (7)$$

It reminds a simple regression that contains a binary variable. And the coefficient is the *ATE*.

Assumptions

Unconfoundedness

$$X_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | Z_i$$

This assumption tells that if there are unobserved variables, those variables are not confounders. This assumption aims at eliminating **selection bias**.

Overlap

$$\forall x \in \text{supp}(X), \quad 0 < P(W = 1 | X = x) < 1$$

It states that every point of the covariate space we can find both the treated and control individuals.

Disjunctive Cause Criterion

Disjunctive Cause Criterion tells a methodology to select variables to control observed confounding variables. Actually, the criterion is to control variables that satisfy backdoor criterion if there are any. An example is,

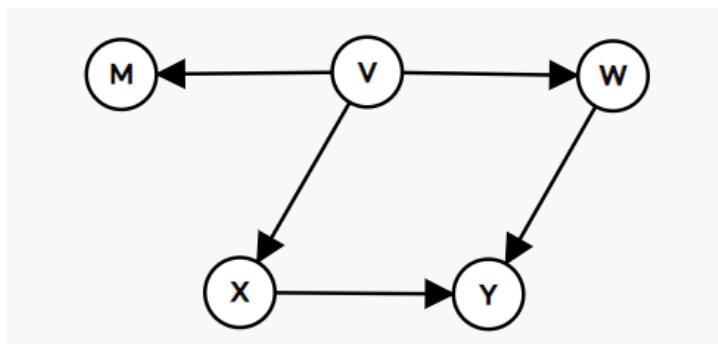


Fig5: Disjunctive Example

And based on the criterion, we should choose V and W because they satisfy the backdoor criterion for $X \rightarrow Y$ treatment.

Regression for Causal Inference

Therefore, we can add some control variables into our model based on equation 7,

$$Y_i = Y_i^0 + \tau X_i + \beta Z_i$$

And,

$$E(y|do(x = a)) = \int \mathbb{E}[y|x = a, z]p(z)dz$$

$$\tau = \begin{cases} \frac{d\mathbb{E}[y|do(x)]}{dx}, & \text{if continuous} \\ \mathbb{E}[y|do(x = 1)] - \mathbb{E}[y|do(x = 0)], & \text{if binary} \end{cases} \quad (8)$$

Simulation

Homogeneous Treatment Effect

Simulation 1

First, we simulate under Fig3, where an observed confounding variable affects treatment and result simultaneously.

```
x ← N(3, 4)
Treatment ← Bernoulli(1 - exp(-x))
y = 1 + 3 * Treatment + 0.2x + N(0, 0.25)
```

» Unfold to See the Code



The result is shown below,

	Coefficient	P-value
Intercept	0.9849	0.00
Treatment	3.0233	0.00
x	0.2001	0.00

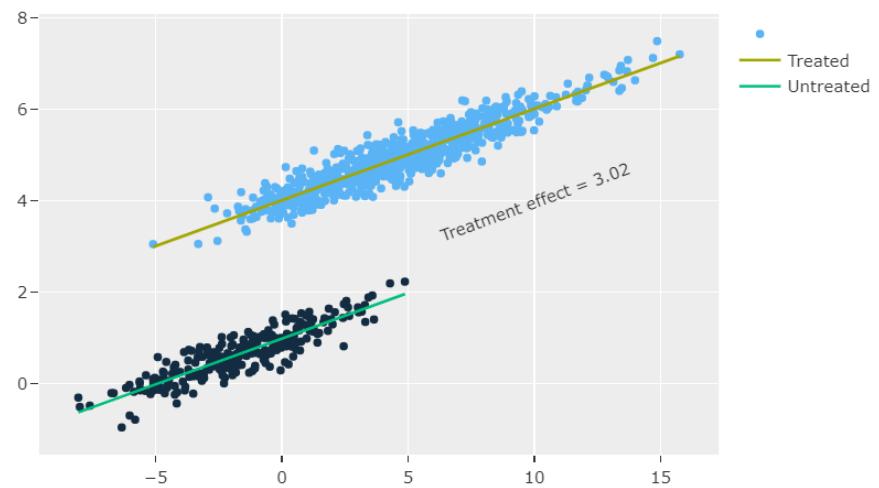


Fig:6 Simulation 1

Heterogeneous Treatment Effect

Simulation 2

```
x ← N(3, 4)
Treatment ← Bernoulli(1 - exp(-x))
y = 1 + α * Treatment + 0.2x + N(0, 0.25), α ~ N(3, 1)
```

[» Unfold to See the Code](#)

The result is,

	Coefficient	P-value
Intercept	1.0116	0.00
Treatment	3.0125	0.00
x	0.2020	0.00

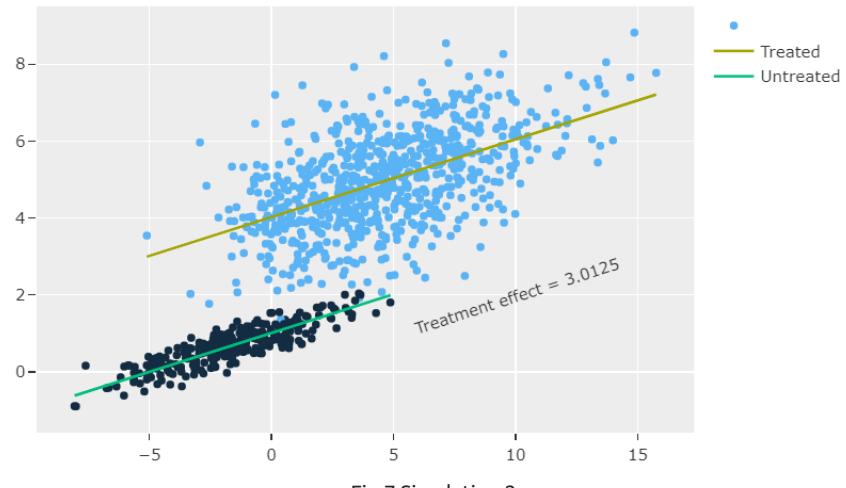


Fig:7 Simulation 2

Here, we can see that if heterogeneous treatment effect is independent with x , we can easily get ATE from the mean of the distribution of the coefficient term ($\mathcal{N}(3, 1)$), via equation 8, $\mathbb{E}(\alpha) = 3$.

Simulation 3

$$\begin{aligned} x &\leftarrow \mathcal{N}(3, 4) \\ \text{Treatment} &\leftarrow \text{Bernoulli}(1 - \exp(-x)) \\ y &= 1 + 0.5 * \text{Treatment} + 0.2x + 0.2x * \text{Treatment} + \mathcal{N}(0, 0.25) \end{aligned}$$

[» Unfold to See the Code](#)

The result is,

	Coefficient	P-value
Intercept	0.9813	0.00
Treatment	0.5251	0.00
x	0.1979	0.00
x*treatment	0.2025	0.00

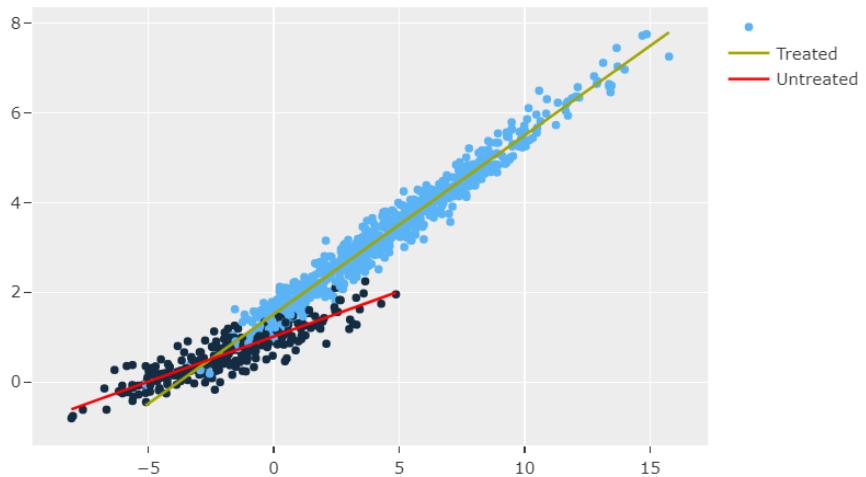


Fig:7 Simulation 3

By the equation 8,

$$\tau = ATE = \mathbb{E}(0.2x + 0.5) = 1.1$$

OtherMethods

Inverse-propensity score weighting

Rosenbaum and Rubin (1983) have shown that whenever unconfoundedness holds, it is sufficient to control for the **propensity score**. The propensity score serves a single-dimensional variable that summarizes how observables affect the treatment probability. Let,

$$e(z) = P(X_i = 1 | Z_i = z)$$

∴ Unconfoundedness, $Y_i \perp X_i | e(Z)$

That is, a comparison of two people with the same propensity score, one of whom received the treatment and one who did not, should in principle adjust for confounding variables. Propensity score weighting (PSW) provides our starting point as a method to reduce the effects of confounding in observational studies. The basic idea is to weight the observations to obtain similar baseline characteristics. The following results can be shown to hold,

$$E(Y_i(1)) = \mathbb{E}\left(\frac{Y_i X_i}{e(Z_i)}\right), \quad E(Y_i(0)) = \mathbb{E}\left(\frac{Y_i(1 - X_i)}{1 - e(Z_i)}\right)$$

Thus,

$$\tau = ATE = \mathbb{E}\left(\frac{Y_i X_i}{e(Z_i)} - \frac{Y_i(1 - X_i)}{1 - e(Z_i)}\right)$$

Weighted OLS

Using weights on OLS,

$$w = \frac{X}{e(Z)} + \frac{1 - X}{1 - e(Z)}$$

For observations, we apply weights to make the treated and the untreated as in the similar baseline characteristics.

Causal Tree

Another important way to estimate **heterogeneous treatment effect** is the **Causal Tree**. It is an adjusted version of decision tree. Simply speaking, decision tree is efficient at capture **nonlinear relationships**. Thus, after classification, those 'similar' samples are in the same leaf. Say, we are using decision tree to **control** the

confounders (satisfying the backdoor criterion). In each leaf, we will get the *CATE* by equation 6.

Causal Tree introduces a new splitting method and minimization criterion called **Honest**. Honest applies the thought of CV, and it creates \mathcal{S}^{est} to avoid overfitting in training.

- **Honest Target,**

$$\text{MSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) = \frac{1}{\#(\mathcal{S}^{te})} \sum_{i \in \mathcal{S}^{te}} \{(\tau_i - \hat{\tau}(X_i; \mathcal{S}^{est}, \Pi))^2 - \tau_i^2\}$$

$$EMSE_\tau(\Pi) = \mathbb{E}_{\mathcal{S}^{te}, \mathcal{S}^{est}} \text{MSE}_\tau$$

- **Honest Split,**

$$EMSE_\tau = -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi) + \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}} \right) \sum_{\ell \in \Pi} \left(\frac{S_{treated}^2}{p} + \frac{S_{controlled}^2}{1-p} \right)$$

, where \mathcal{S}^{te} denotes the test set, Π denotes the splits. This splitting method ensures the **balance** in leaves.

References

List of Works

- Jiaming Mao, [Data Analysis - Causality Part \(Part II\)](#)
- Ken Acquah, [Causal Flows](#)
- Susan Athey, [ATE Tutorial](#)
- Guido Imbens & Susan Athey, Recursive partitioning for heterogeneous causal effects.

Picture Reference

Pixiv by Alcxome

Casual Effect Estimation
<http://example.com/2021/06/18/Causal-Regression/>

作者	发布于	更新于	许可协议
Junwei Lin	2021-06-18	2021-06-23	



微观计量

Ridge and Lasso Regression ➤