

UGOD 5020

Quantitative Social

Science

We 9:00 - 11:50

208 E2

# 基本无害的计量经济学

## 变量因果关系

## 设计实验找因果关系

很多情况下，我们做不到像上一节随机实验那样通过把 $x$ 随机分配来识别因果效应，即使如此，大部分情况下我们依然可以判断在统计学意义上 $x$ 与 $y$ 是否相关，为此引入条件期望函数

(conditional expectation function, CEF)。

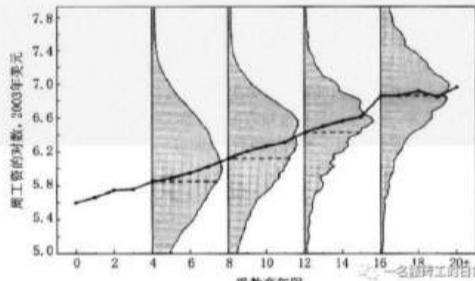
给定一个被解释变量 $Y_i$ ，和一组解释变量 $X_i$  我们希望判断这一组 $X_i$  在多大程度上能够解释 $Y_i$  的变化。比如给定个体受教育年限，如何预测其未来收入。首先定义CEF如下：

$$E(Y_i|X_i = x) = \int t f_y(t|X_i = x) dt$$

如果 $Y_i$ 是离散的，那么就是

$$E(Y_i|X_i = x) = \sum_t P(Y_i = t|X_i = x) dt$$

修改：式子中应是 $tP$ ，少写了个 $t$ 。需要注意的是这里有总体和样本的区别，期望的概念是总体的均值，而一般我们的数据只是总体中的一部分样本，所以需要使用样本来推测总体，这一步需要用到大数定律。将在后面具体讨论。



上图反映了教育和收入的关系，在每个受教育年限上画出了这个群体的收入分布，中间的连线则为 $E(Y_i|X_i = x)$ ，在这个图中清晰的看到了如何用CEF对收入做预测，即中间的连线。在进一步讨论实际回归操作之前，需要明白几个CEF的性质：

### 1. Law of iterated expectation:

$$E(Y_i) = E_x E_y(Y_i|X_i)$$

### 2. 由此性质可以得到 CEF分解性质：

$$Y_i = E(Y_i|X_i) + \epsilon_i$$

根据构造， $\epsilon_i$  关于 $X_i$  均值独立。即 $E[\epsilon_i|X_i] = 0$  且任何函数 $h(\cdot)$  均满足 $E[h(X_i)\epsilon_i] = 0$

### 3. CEF是在所有 $X_i$ 函数中最小化残差平方和的函数：

$$E(Y_i|X_i) = \arg \min_{m(X_i)} E[Y_i - m(X_i)]^2$$

而这个函数和我们一般见到的线性回归函数有什么联系呢？

## 识别策略

## 统计推断

首先对我们熟知的线性回归做一个定义，对总体要拟合一个线性回归方程，其系数满足以下最小化残差平方和的条件

$$\beta = \arg \min_b E[(Y_i - X'_i b)^2]$$

求导可以解得

$$E[X_i(Y_i - X'_i b)] = 0 \\ \beta = E[X_i X'_i]^{-1} E[X_i Y_i]$$

在只有一个变量和常数项的情况下，系数又可以写成

$$\beta = \frac{Cov(Y_i, x_1)}{V(x_1)}$$

在有多个变量的情况下，系数可以写成

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$$

其中 $\tilde{x}_{ki}$  表示第 $k$ 个变量对于其他所有变量回归后的残差。(Frisch-Waugh-Lovell theorem)

之前得到CEF是使得在所有关于 $X$ 的函数中残差平方和最小的函数，该线性回归和CEF的关系是怎样的呢？有以下三个定理，略过证明：

1. 如果CEF是线性的，那么上面这个用总体回归出来的方程就是这个CEF函数。
2. 该线性回归结果在对 $Y_i$  拟合的所有线性函数中是最优的，即最优线性估计量(best linear estimator)，这里的最优指的是使得残差平方和最小。 $\min E[(Y_i - X'_i b)^2]$
3. 该线性回归也是对CEF的最优线性拟合，即使得CEF的残差平方和最小化：  
 $\min E[(E[Y_i|X_i] - X'_i b)^2]$

以上定理说明了CEF与线性回归的关系，CEF是对于 $Y$ 的拟合所有可能函数中最优的，而线性回归是对于 $Y$ 的拟合所有线性函数中最优的。并且由定理3线性回归也是对于CEF的拟合中的最优线性估计，这个结论意味着在做线性回归时其实并不需要所有样本，只需要每个 $x$ 下的均值和权重即可，即对CEF的拟合结果和对所有样本的拟合结果一样。在下图可以清楚看出如果只使用每个年龄的收入均值(CEF)做线性回归，那么这个线性回归和之前使用全样本的回归结果是一样的。这个方法在微观数据不可得而分组数据(CEF) 可以得到时非常有用。如Angrist(1998) 使用分组数据研究自愿参军对于收入的影响。

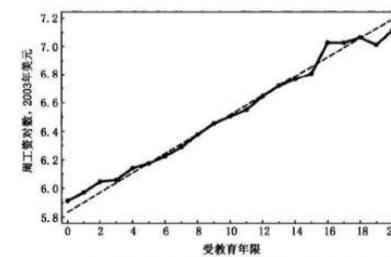


图 3.2 回归线贯穿了给定受教育年限下周平均工资的条件期望函数  
(点=条件期望函数; 折线=回归线) 一名领竣工的日常

CEF和线性回归

然而需要注意的是，使用总体估计和只是用分组数据估计的标准差会有所差别，所以在做统计推断时需要小心。

## 选择性偏差

### 2.1 选择性偏差

举一个例子来说明实验对于发现因果关系的重要性。“在医院接受医疗服务能使人更健康吗？”对于大部分人来讲，这个答案应该是显而易见。那么我们拿出一份医院调研数据来看：

| 分组      | 样本    | 平均健康水平 |
|---------|-------|--------|
| 过去一年住过院 | 7774  | 2.79   |
| 一年内没住过院 | 90049 | 2.07   |

其中健康水平从1（最健康）到5（最糟糕）。这是否说明住院导致人更不健康了呢？明显不是这样，因为住院的人之所以去住院，就是因为本身健康水平就比其余的人低，那么接受完治疗也不一定能使其恢复到和其他人一样的水平。这就是选择性偏差的一个反映，用数学的方法进行刻画如下，这个刻画方式将会贯穿整本书的计量分析。定义一个二值变量 $D_i = \{0, 1\}$  为处理变量，等于0表示个体*i*没接受治疗，这里就表示没住院。定义一个结果变量 $Y_i$ ，这里表示个体*i*的健康水平。那么住院导致的因果效应是什么呢？即如果一个个体*i*接受了住院，那么相比于他没有接受住院时候的健康水平提升多少。这里要对可能产生的结果做出定义，即

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

这里 $Y_{0i}$  ( $Y_{1i}$ ) 分别表示假如个体*i*未住院(住院)的健康水平。然而同一个个体我们只能看到他在一种情况下的结果，即 $Y_{0i}$ 或者 $Y_{1i}$ ，而我们想要得到的是同一个人在两种情况下的差别。如果我们直接将观测到的两类人的结果相减，则会得到如下：

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) = [E(Y_{1i}|D_i = 1) - E(Y_{0i}|$$

第一个等号为定义，即我们用观测到的住院的人的健康水平 $E(Y_{1i}|D_i = 1)$ 减去没住院的人的健康水平 $E(Y_{0i}|D_i = 0)$ 。第二个等号为同时加减 $E(Y_{0i}|D_i = 1)$ 。第一个方括号即为我们想知道的效果，对于实际上住院的人，住院后的效果 $E(Y_{1i}|D_i = 1)$ 比如果他们不住院大的效果 $E(Y_{0i}|D_i = 1)$ 高多少。但是这个算式中还包括第二个方括号，表示实际住院的人如果不住院的话，他们的健康水平和一直没住院的人 $E(Y_{0i}|D_i = 0)$ 的差别。这第二项就是选择性偏差。由于住院的人本身身体条件比不住院的差，这个选择性偏差是负数，也就导致了我们用观测到的数据相减是负数，即使住院效果是正的。

选择性偏差  
身体完点才住院

解决办法：随机实验 一便 $D_i$ 和 $Y_i$ 不相关

→ 内生性 (加入控制变量) 一下 $Y_i = \alpha + \rho D_i + \beta X_i + \eta_i$

## OLS 的渐进分布

使用总体可以求得的线性回归的系数为

$$\beta = E[X_i X'_i]^{-1} E[X_i Y_i]$$

然而我们需要的是使用总体的一部分样本来对该值进行估计，这就带来了OLS估计值：

$$\hat{\beta} = [\Sigma_i X_i X'_i]^{-1} [\Sigma_i X_i Y_i]$$

这里用到的是使用样本的均值来估计总体的均值 $E$ 。而这个估计值 $\hat{\beta}$  多大程度上接近总体的真实值 $\beta$ ，取决于对于样本的假设，最极端的当然是如果样本就是总体，那么均值就是期望值。当样本小于总体时，这个估计值分布就会取决于对样本的假设。此时就可以用大数定律来确定估计值的分布。最重要的有以下几个，对于推导 $\hat{\beta}$  的渐进分布至关重要，简要说明，任何一本书都有推导细节！

1. THE LAW OF LARGE NUMBERS: 样本矩依概率收敛到总体矩。

2. THE CENTRAL LIMIT THEOREM: 样本矩以一定方式服从正态分布。

3. SLUTSKY'S THEOREM: 两变量加减乘积等运算后的值也依概率收敛到相对应每个变量的总体矩的运算值，如其中一个依概率收敛到一个分布，那么运算后值收敛到这个分布与另一个总体矩的运算值。

4. THE CONTINUOUS MAPPING

THEOREM: 任何函数 $h(x_n)$  收敛到对应

$h(x^*)$ ，其中 $x^*$  为 $x_n$ 的收敛值。

5. THE DELTA METHOD: 对于一个渐进正态分布变量 $x_n$ ，其任何可微函数 $h(x_n)$  也是渐进正态分布。

有些工具后，可以推出OLS的估计值 $\hat{\beta}$  的渐进分布为正态分布。其均值和方差分别为

$$E[\hat{\beta}] = \beta + E[\Sigma_i X_i X'_i]^{-1} [\Sigma_i X_i \epsilon_i]$$
$$var(\hat{\beta}) = E[X_i X'_i]^{-1} E[X_i X'_i \epsilon_i^2] E[X_i X'_i]^{-1}$$

t统计量使用的标准差即为这个式子相对应角线元素的平方根。这个方差中的 $\epsilon_i^2$  是使用样本残差估计的，右下标的*i*说明了异方差的性质，这就是robust error, 或White standard error。当CEF是非线性函数时，OLS下的 $\epsilon_i^2$  则必定是异方差：

$$E[(Y_i - X'_i \beta)^2 | X_i] = E[(Y_i - CEF + CE$$

第二个等号中由CEF性质误差项与X不相关，得到交乘项为0所以略去。CEF的非线性保证了第二项 $E[(CEF - X'_i \beta)^2 | X_i]$  不会为一个常数，所以OLS的误差项将会异方差。

在大部分计量经济学教科书中，有许多的经典的假设，在那些假设下OLS估计量会满足无偏，一致等。不过大样本下使用大数定律得到的渐进分布略去一些假设下依然可以成立，并且大部分实证研究都是基于大样本理论下，所以研究渐进分布就显得有必要。





## Review

### • CEF 条件期望函数 Conditional Expectation Function

### • CIA 条件独立假设 Conditional Independence Assumption

bias /

均值的差可看作因果关系  $ATE = E[Y|X] - E[Y_0|X]$

Average Treatment Effect

关心的变量

$Y_i \sim LS_i | X_i$

因变量

控制分组

treatment is as good as random

CATE: Conditional Average Treatment Effect

OVB

## Omitted – Variable Bias and Controls

e.g.  $Y_i = \alpha + \rho S_i + \eta_i$  ① CIA does not hold → selection bias

& correlation between A and Y (+)

$Y_i = \alpha + \rho S_i + X_i r + \nu_i$  ② if CIA holds → can be interpreted causally

↓ OV → OV会使估计系数偏大

$\xrightarrow{S \xrightarrow{+} A \xrightarrow{+} Y}$  what happened if we can't get data? Bias

1. will yields:  $\beta = \frac{\text{Cov}(Y_i, S_i)}{\text{Var}(S_i)} = \rho + \delta_{AS}$  → coefficients from regression of the elements of  $A_i$  on  $S_i$ .

$$A_i = \alpha + \delta_{Si} + e_i$$

$\delta$  corr between  $S_i$  and  $A$  (+)

selection bias:  $S, Y$  → If we think ability positively affects wages,

then it looks like <sup>we</sup> also have positive selection into schooling

(know  $OVB > 0$ , ability → wages, ability → schooling.)

causality: OVB does not require either of the models to be causal.

helps us understand the CIA

Bad controls are variables that are also affected by treatment

e.g.  $\left\{ \begin{array}{l} Y_i = C_i Y_i(1) + (1-C_i) Y_i(0) \\ W_i = C_i W_i(1) + (1-C_i) W_i(0) \end{array} \right.$   $C$  is randomly assigned

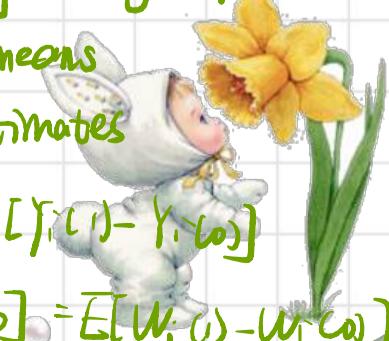
differences in means

$\left. \right\} E[Y_i | C_i=1] - E[Y_i | C_i=0] = E[Y_i(1)] - E[Y_i(0)]$

↪ Dummy

$$E[W_i | C_i=1] - E[W_i | C_i=0] = E[W_i(1)] - E[W_i(0)]$$

$$E[W_i | C_i=1] - E[W_i | C_i=0] = E[W_i(1) - W_i(0)]$$



when control for occupation (condition on  $W_i = 1$ )

$$E[Y_i | C_i=1, W_i=1] - E[Y_i | C_i=0, W_i=1] = E[Y_i(1) | C_i=1, W_i=1] - E[Y_i(0) | C_i=0, W_i=1].$$

$$= E[Y_i(1) | W_i(1)=1] - E[Y_i(0) | W_i(0)=1] \quad (\because C_i \text{ is randomized})$$

$$= E[Y_i(1) | W_i(1)=1] - \underbrace{E[Y_i(0) | W_i(1)=1]}_{\substack{\text{describes how college} \\ \text{graduation changes the} \\ \text{composition of the pool of} \\ \text{white-class workers}}} + \underbrace{E[Y_i(0) | W_i(1)=1]}_{\substack{\text{blue-collar or white-collar}}} - E[Y_i(0) | W_i(0)=1]$$

$$= \underbrace{E[Y_i(1) - Y_i(0) | W_i(1)=1]}_{\text{conditional treatment effect}} + \underbrace{E[Y_i(0) | W_i(1)=1] - E[Y_i(0) | W_i(0)=1]}_{\text{selection bias}}$$

(occupation choice)

- By introducing a bad control, we introduced selection bias into a setting that did not have selection bias without controls

If we're concerned about discrimination, it seems likely that discrimination also affects occupational choice and hiring outcomes.

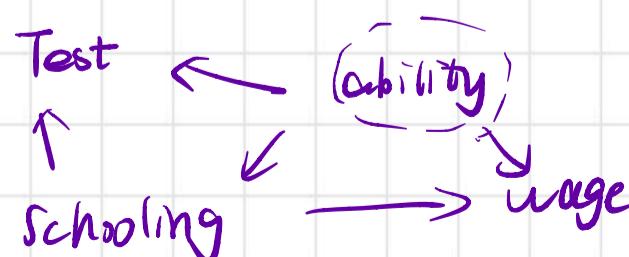
- Proxy variables intersects omitted-variable bias and bad controls

① Suppose we want to estimate the returns to education.

→ Ability is omitted

② We have a proxy for ability — a test taken after

schooling finishes





- If we omit the test altogether, we've got omitted-variable bias.
- If we include our proxy, we've got a back control  
→ selection bias
- If we control the test, → we need to adjust the selection bias

### 3.2.2 遗漏变量 (omitted variable)

由之前得到，教育与收入的回归方程如下：

$$Y_i = \alpha + \rho s_i + X'_i \gamma + v_i$$

在保证该模型中残差项  $v_i$  与  $X$  与  $s$  均不相关时，所得到的系数  $\rho$  就是因果效应。

如果我们不加入这个控制变量  $X$ ，由之前可知会产生选择性偏差，这时候回归方程如下

$$Y_i = \alpha + \rho s_i + \eta_i$$

$\eta_i = X'_i \gamma + v_i$ ，如果对这个方程做回归，回归系数为

$$\frac{\text{cov}(Y_i, S_i)}{V(S_i)} = \rho + \gamma' \delta_{xs}$$

这个推导直接将  $Y$  带入即可。其中  $\delta_{xs}$  是  $x$  对  $s$  的回归系数。这里可以看出，如果我们不将表示能力水平的变量  $x$  加入回归，那么回归结果会高估，因为能力水平和教育水平是正相关  $\delta_{xs} > 0$ ，而且能力水平和收入也正相关  $\gamma > 0$ 。这和之前讨论的选择性偏差的方向是一致的。

需要注意的是，这里可以看出如果遗漏变量  $x$  和  $s$  不相关， $\delta_{xs} = 0$ ，那么回归就不会有问题，这就是有条件独立假设时的情况。同样如果有更多的遗漏变量，那么这个回归偏差还需要看那些遗漏变量与  $s$  和  $y$  的相关性而定。

### 3.2.3 坏控制变量 (bad control)

虽然遗漏变量会产生问题，但是也不是加越多控制变量就越好的。以下两种情况的控制变量并不一定会改善选择性偏差的问题。

#### 1. 控制变量本身由政策变量决定。

在教育与收入的例子中，职业选择同样会决定收入高低，那么假如职业选择是否能估计的更准确呢？这里的问题在于，职业选择一定程度上也会由教育水平决定，假设只有两个职业，白领和非白领，而受过大学教育更容易选择白领工作，这样一来如果我们加入白领作为控制变量，那么对比的则是在所有白领人员中，上过大学的和没上过大学的。而由于上大学本身会使得成为白领的概率变大，在已经从事白领工作的人中，是否上过大学就不再是随机的了。因此在加入控制变量时，一般不要加入在政策变量之后会受政策变量影响的变量，而需要加入在之前发生的影响政策变量的变量 (predetermined)。

#### 2. 使用代理变量作为控制变量

同样的例子，我们担心本身能力会成为遗漏变量，因此加  $X$  来控制相同能力水平，然而我们衡量能力的指标可能并不能反映天生的能力水平，而可能会使用诸如成绩或智力测试的指标，而这些指标本身会随着教育水平的增加而增加，也就是说在这个回归中

$$\begin{aligned} Y_i &= \alpha + \rho s_i + \gamma a_i + v_i \\ X_i &= \pi_0 + \pi_1 s_i + \pi_2 a_i + e_i \end{aligned}$$

这里的  $a_i$  才是我们真正关心的遗漏变量能力水平，只不过我们无法观测到这个值， $X$  则是能够观测到的衡量指标，称  $X$  为代理变量。然而如果使用  $X$  作为  $a$  来回归，回归方程变为：

$$Y_i = (\alpha - \gamma \frac{\pi_0}{\pi_2}) + (\rho - \gamma \frac{\pi_1}{\pi_2}) s_i + \frac{\gamma}{\pi_2} X_i - \gamma \frac{\pi_2}{\pi_2} e_i$$

可以看到此时回归系数将会不等于  $\rho$ ，并且会低估，只有当  $\pi_1 = 0$ ，此时  $s$  和  $X$  不相关时，回归系数才是我们想要的。



# Matching



- Treatment is as good as random conditional on a known set of covariates
- Fundamental problem of causal inference: we cannot simultaneously both  $Y_i(1)$  and  $Y_i(0)$

We match untreated observations to treated observations using  $X_i$ , i.e., calculate a  $\hat{Y}_i(0)$  for each  $Y_i(1)$ , based upon "matched" untreated individuals

## 3.3.1 回归与匹配

除了用回归来获得因果效应，也可以直接使用最直观的匹配来获得因果效应，由Day7：

$$E[Y_i|X_i, C = 1] - E[Y_i|X_i, C = 0] = E[Y_i]$$

这个式子表示对于每个X水平的人，其受到政策影响的因果效应为  $\delta_X$ ，这个值可以直接由等式左边代入相应观测值数据得到。这一步就要求对于每个X的水平都需要有一部分人接受政策一部分人没有接受，如果对于一个X水平下所有人都接受或没接受，那么就无法估计。

由此，根据law of iterated expectation我们通过加权平均不同X的人的因果效应来获得政策的平均效应：Treatment on Treated TOT：

$$\delta_{TOT} = E[Y_{1i} - Y_{0i}|D_i = 1] = E[E[Y_{1i} -$$

由此，政策对那些接受政策的人影响TOT就是  $\delta_X$  的条件期望：

$$\delta_{TOT} = \sum_x \delta_X P(X_i = x | D_i = 1)$$

根据law of iterated expectation，政策对所有人的平均期望average treatment effect ATE为：

$$\delta_{ATE} = E[Y_{1i} - Y_{0i}] = E[E[Y_{1i} - Y_{0i}|X_i]]$$

由此，政策对那些所有人的平均影响ATE就是  $\delta_X$  的无条件期望：

$$\delta_{ATE} = \sum_x \delta_X P(X_i = x)$$

这就是使用匹配得到的因果效应估计值，那么回归结果的  $\rho$  和匹配估计值有什么区别呢？

根据数学推导简而言之，回归结果的值也是  $\delta_X$  一个加权平均值，只不过权重不是匹配中的概率  $P$ ，而是每个X下对应政策变量的方差：

$$\rho = \frac{E[\sigma_D^2(X_i)\delta_X]}{E[\sigma_D^2(X_i)]}$$

其中  $\sigma_D^2$  为政策变量D的条件方差

$$\sigma_D^2 \equiv E[(D - E[D_i|X_i])^2 | X_i]$$

除非X与D不相关，即此时

$$\rho = \frac{E[\sigma_D^2\delta_X]}{E[\sigma_D^2]} = E[\delta_X]$$

，否则匹配得到的ATE和回归得到的系数不相等。

当样本中存在在某一X水平下，所有个体均接受或没有接受政策处理时，这个样本在两种估计下都不会有任何权重。对于匹配， $\delta_X$  将会无法识别，对于回归， $\sigma_D^2$  为0。





e.g.  $\hat{ITE}_2 = \hat{Y}_2(1) - \hat{Y}_2(0) = Y_2(1) - Y_4(0) = 1 - 0 = 1$   $ATE = E(\hat{ITE})$

Simple illustration

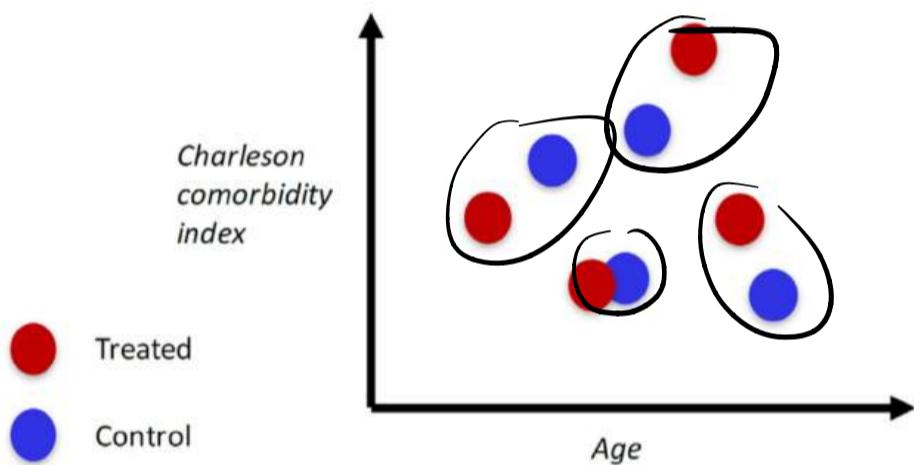
| i | X  | T | $Y(1)$ | $Y(0)$ | $ITE = Y(1) - Y(0)$ |
|---|----|---|--------|--------|---------------------|
| 1 | 20 | 0 |        | 0      | ?                   |
| 2 | 31 | 1 | 1      |        | ?                   |
| 3 | 43 | 1 | 0      |        | ?                   |
| 4 | 30 | 0 |        | 0      | ?                   |
| 5 | 20 | 1 | 1      |        | ?                   |

Problem: If  $T_i=1$ , then observe only

$\hat{Y}_2(1)$ , need estimates for  $\hat{Y}_4(0)$

Idea: find some  $j$ , who is similar to  $i$  &  $T_j=0$ , then  $\hat{Y}_2(0) = \hat{Y}_j(0)$ .

## NN (Nearest neighbor covariate matching)



- choosing the single closest control observation using the metric (distance) between  $x_i$  and  $x_j$ .

$$\begin{aligned} j(i) &= \underset{j: T_i \neq T_j}{\operatorname{argmin}} \text{distance}(x_i, x_j) \\ &= (x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \dots \end{aligned}$$

## Conditional average treatment effect

$j_i$  = the nearest neighbor of  $i$

The estimated individual treatment effect is then

$$\hat{ITE}_i = \hat{Y}_{i(1)} - \hat{Y}_{i(0)} = \hat{Y}_{i(1)} - \hat{Y}_{j(i)(0)}$$

$$\hat{ITE}_i = \hat{Y}_{i(1)} - \hat{Y}_{i(0)} = \hat{Y}_{j(i)(1)} - \hat{Y}_{j(i)(0)}$$

Produces causal estimates if CIA is valid and we have sufficient overlap

Suffers from arbitrary choices of units





## NN matching with Mahalanobis distance

- Nearest-neighbor matching chooses the single closest control observation using the Mahalanobis distance between  $X_i$  and  $X_j$

$$d(X_i, X_j) = (X_i - X_j)' \Sigma_X^{-1} (X_i - X_j)$$

- Does not suffer from arbitrary choices of units

↑  
↑  
↑  
↑  
↑

- We want to construct a counterfactual for each individual (say, with  $T_i=1$ )
- The counterfactual for  $i$  should only use individuals that match  $X_i$
- Let  $w_i(j)$  be the weights on the controls, our estimate:  $\hat{Y}_i(w) = \sum_{j: T_j=0} w_i(j) Y_j(w)$
- Then, our estimated individual treatment effect (for the treated) is

$$\hat{ITE}_i = \hat{Y}_{i(1)} - \hat{Y}_{i(0)} = Y_{i(1)} - \sum_{j: T_j=0} w_i(j) Y_j(w)$$

Weighted average of all individuals in the control group

Matching estimator for the ATE:

$$\hat{ATE}_M = \frac{1}{N_T} \sum_{i: T_i=1} [Y_{i(1)} - \sum_{j: T_j=0} w_i(j) Y_j(w)].$$

## More neighbors

### More neighbors

- Kernel matching gives positive weight to all control observations within some bandwidth  $h$ , with higher weight for closer matches determined by some kernel function  $K(\cdot)$

$$w_i(j) = \frac{K\left(\frac{|x_i - x_j|}{h}\right)}{\sum_{j: T_j=0} K\left(\frac{|x_i - x_j|}{h}\right)}$$

normalization

$w_i(j) = 0 \text{ if } |x_i - x_j| > h$



① Interpretable especially in small-sample regime

② Nonparametric

③ Heavily reliant on the underlying metric

④ Outliers / far-away observations that are hard to match, → need bias correction

$$\hat{Y}_i(w) = \sum_{j: T_j=0} w_i(j) Y_j(w)$$

Weighted average of all individuals in the control group

So all we need is those weights and we're done: Special cases:

Mean difference: NN estimator

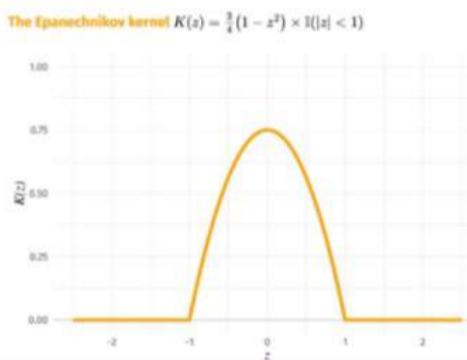
(unmatched observed difference)

$$w_i(j) = \begin{cases} \frac{1}{N_c} & \text{if } j = \text{NN} \\ 0 & \text{otherwise} \end{cases}$$

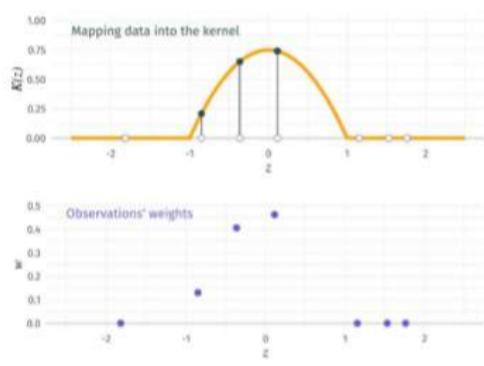




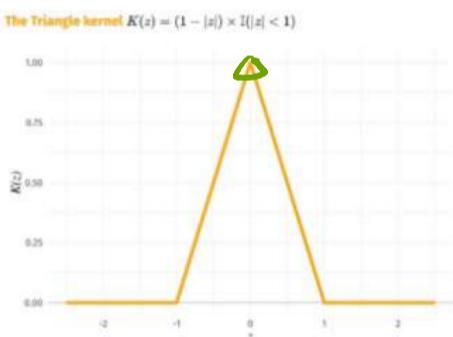
## Epanechnikov kernel



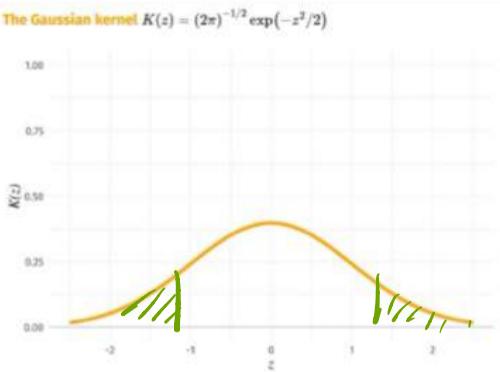
## Epanechnikov kernel



## Triangle kernel



## Gaussian kernel



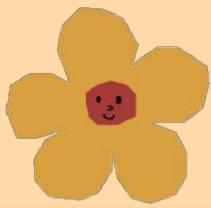
## Adding neighbors

- As we add more neighbors — for example by increasing our bandwidth — we potentially increase the efficiency of our estimator
- We need to be careful not to add too many controls for each treated i
- CIA requires that we're actually conditioning on the observables — it does not allow us to take a simple average across all control observations.

## The curse of dimensionality

- It turns out kernel- and bandwidth-selection are not our biggest enemies
- As the dimension of  $x_i$  expands (matching on more variables), it becomes harder and harder to find a nice, close control for each treated unit.
- We need a way to shrink the dimensionality of  $x_i$





## 工具变量

加权回归 weight regression

有限被解释变量 limited dependent variable, LDA | 假设没变  
若 OLS 满足

