

第一章 因果关系入门

T_1 干预介入量 treatment intake ~ 现实(分组)

Y_1 观察结果变量 ~ 真理

$ATE = E[Y_1 - Y_0]$ 因果 平均处理效果

$ATT = E[Y_1 - Y_0 | T=1]$ (被干预者的)

$$\underbrace{E[Y|T=1] - E[Y|T=0]}_{\text{相关}} = \underbrace{E[Y_1 - Y_0 | T=1]}_{ATT} + \underbrace{E[Y_0 | T=1] - E[Y_0 | T=0]}_{BIAS \star}$$

当 $BIAS=0$, 相关 $\xrightarrow{\text{代表}} ATT \xrightarrow{\text{可交换}} E[Y_1 - Y_0] = \text{因果}$

02: 随机实验 Randomised Experiments 提供一种消除 BIAS 的方法

事实 $\xrightarrow{\text{干预}} \begin{matrix} \leftarrow \text{需要消除} \\ \text{潜在结果} \end{matrix} \xrightarrow{\text{干预}} \begin{matrix} \Rightarrow \text{需要研究} \\ \text{结果} \end{matrix}$

而要 $\xrightarrow{\quad} (Y_0, Y_1) \perp T$ 即: $E[Y_0 | T=0] = E[Y_0 | T=1] = E[Y_0]$ $BIAS=0$

RCT

03: 统计知识回顾

$$\text{估计值的标准误差: } SE = \frac{\sigma}{\sqrt{n}} \rightarrow \hat{\sigma}^2 = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

置信区间: 样本容量 \downarrow 标准误 \uparrow 置信区间宽度 \uparrow

假设检验: H_0 : T_1 生成的正态与 T_0 生成的正态相同
值

即使各个置信区间重叠, 差异仍可在统计上不等于。
 T_1/T_0 在 95% 下均值落入的区间 $\leftarrow T_1, T_0$ 的值之差。

P 值: $P(\text{data} | H_0)$ H_0 成立下, 发生观察到数据的概率更极端 (弃真错误)

置信度: 让估计值和总体参数在一定范围的误差之内 的概率 (置信度)

置信度 \uparrow 置信区间 \uparrow 准确性 \downarrow

显著性水平: 犯弃真错误的概率

当 P 值很小, \Rightarrow 当 H_0 , data 发生很小, 但它发生了 $\Rightarrow H_0$ 是错的

p 小过临界值 \leftrightarrow 临界值为显著性水平
样本 理性估计

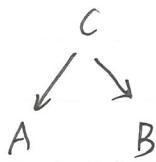
04：图因果模型 $A \rightarrow B$ 表示 A 影响 B，依赖性沿箭头传递

① $A \rightarrow B \rightarrow C$ (B 可阻断 $A \rightarrow C$)

当以 B 为条件 (B 已知) : $A \not\perp C$

$A \perp C | B$

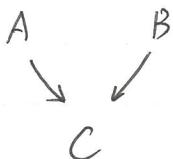
②



当 C 为条件: $A \not\perp B$

$A \perp B | C$

③



当以 C 为条件: $A \perp B$

$A \not\perp B | C$

混淆偏差:

替代混杂图表: 一些不可测的变量

个人总结

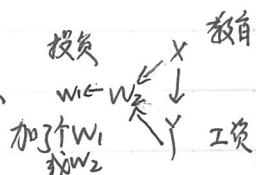
混淆偏差 :



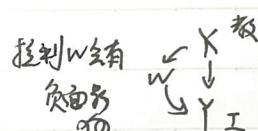
未控制开门的影响

控制偏差: 多控制

多控制共同结果,
冲淡因子打开关系因果

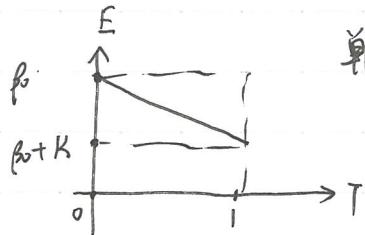


多控制
削弱因子之一会
成为因果



05: 线性回归的有效性

对个体 i : $Y_{ii} = \beta_0 + K$, 计算 K 正负



单回归: $E = \beta_0 + K \cdot T + u_i$

斜率, 为 ATE, 平均治疗效果

$$\downarrow \text{单变量回归 } \beta_1 = \frac{\text{Cov}(Y_i, T_i)}{\text{Var}(T_i)}$$

当 T 为随机分配的, 则 β_1 为 ATE ✓

多变量回归: 处理名影响 T_i 的多因素推到问题

$$y_i = \beta_0 + K T_i + \underbrace{\beta_1 X_{i1} + \dots + \beta_K X_{ki}}_{\text{辅助变量}} + u_i$$

$$K = \frac{\text{Cov}(Y_i, \tilde{T}_i)}{\text{Var}(\tilde{T}_i)}$$

\tilde{T}_i 为其他协变量在 $X_{i1} + \dots + X_{ki}$ 在 T_i 上回归的残差

←
先用 X_i 预测, 取残差 \tilde{T}_i : 其不可用 X_i 预测,

当回归过程中遗漏变量 (本应用多元而用了单元或漏了):

正解: $Wage_i = d + K Educi + \underbrace{A_i' \beta}_{\text{遗漏的项}} + u_i$

$$\frac{\text{Cov}(Wage_i, Educi)}{\text{Var}(Educi)} = K + \underbrace{\beta' \delta_{Ability}}_{\text{遗漏的项}} \quad A_i \text{ 为向量}$$

单元

δA 为 A 对 $Educ$ 回归的系数向量

↙ OVB Omitted Variable Bias

若模型中考虑了所有混杂变量, 则没有 OVB

RCT 切断联系; Regression

控制中介变量 KOKUYO

Dummy Regression

ob 分组和虚拟回归

异方差：因变量方差在各个特征变量的值域内不恒定

先分组：点数减少，同时分组降低了方差，但参数估计标准误差变大，t统计量也是
 ↗
 去了部分方差信息。

分组回归时要按组内样本数加权

虚拟变量回归

单变量： $wage_i = \beta_0 + \beta_1 T_i + e_i$
 (是否12年) \rightarrow 取0或1，叫结果差一个 β_1 $\begin{cases} T=0 \Rightarrow E = \text{intercept} \\ T=1 \Rightarrow E = \text{intercept} + \beta_1 \end{cases}$

T_i 虚拟，仅连接

$$wage_i = \beta_0 + \beta_1 T_i + \beta_2 IQ_i + \beta_3 IQ_i * T_i + e_i$$

↙

β_0 : $T=0$, $IQ=0$ 的期望

β_1 : $IQ=0$ (除 T 以外全为0)，每个 T_i 对 y_i 的效益

β_2 : $T_i=0$ (除 IQ 以外全为0)，每个 IQ 对 y_i 的效益

* β_3 : IQ 有对 $T_i=1$ 的收益，使 $T=1$ 的回归线 k 更大 ($\beta_3 > 0$)

对 T_i 的回归结果分析：计算每单元对总体的贡献 (按年限分组，对 i 组，通过 i 时 $T_i=1$)

对每一组，求出 $C(\text{educ})[T_i]$

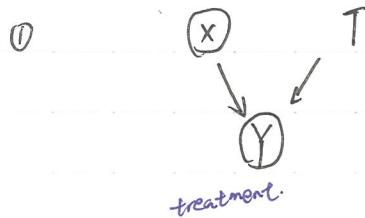
当考虑更多协变量，叫目标参数为对每个虚拟组影响的加权平均值；

$$E[Y|T=1, \text{Group}_i] - E[Y|T=0, \text{Group}_i] w_{\text{Group}_i}]$$

求 $C(\text{educ})[T_i]$ 和 $C(IQ-\text{bins})[T_i]$

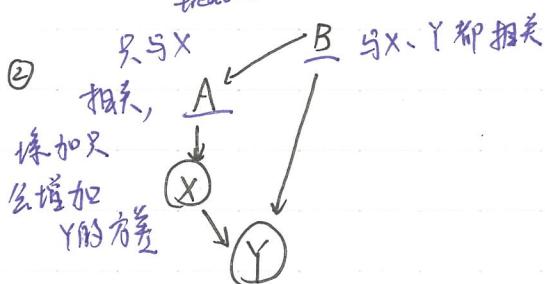
↑ 考察 T_i 的影响
 支: $\sim \text{Var}(T_i | \text{Group}_i)$ KOKUYO

07 控制混淆因素之外的考虑 (控制考虑的变量数)



在考虑 $X \rightarrow Y$ 时，加入控制 T 效果

s.t. X 对 Y 的影响更加明显，降低估计的方差



回归系数标准误差公式：

$$\hat{\beta}^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2$$

$$\text{Var}(\hat{\beta}_2) = \frac{\hat{\beta}^2}{\sum (x_i - \bar{x})^2}$$

\rightarrow 标准误与 X 方差成反比
样本差异越大，标准误差越小

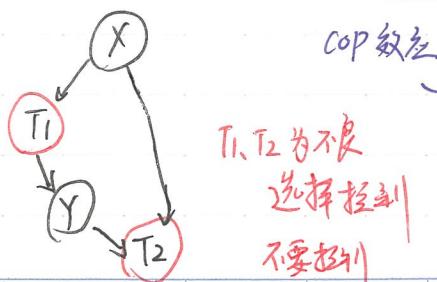
Confounding is the bias from failing to control for a common cause, 共同原因

Selection bias is when we control for a common effect or a variable in
between the path from cause to effect
控制共同结果 原因 \rightarrow 结果间的变量

Bad COP：对于是否治疗会影响效果是否的 treatment.

$$E[Y_i | T_i=1] - E[Y_i | T_i=0] \quad \text{相关 = 因果 (RCT中)}$$

$$= \underbrace{\{P(Y_i > 0 | T_i=1) - P(Y_i > 0 | T_i=0)\} * E[Y_i | Y_i > 0, T_i=1]}_{\text{Participation Effect: } y \text{ 为正的概率}} + \underbrace{\{E[Y_i | Y_i > 0, T_i=1] - E[Y_i | Y_i > 0, T_i=0]\} * P(Y_i > 0 | T_i=0)}_{\text{客户参与消费的可能性的增加}}$$



$$\rightarrow = E[Y_{i1} | Y_{i1} > 0] - E[Y_{i0} | Y_{i0} > 0]$$

$$= E[Y_{i1} - Y_{i0} | Y_{i1} > 0] + \underbrace{\{E[Y_{i0} | Y_{i0} > 0] - E[Y_{i0} | Y_{i0} > 0]\}}_{\text{causal effect}} \rightarrow \text{选择偏差}$$

对于参与者的因果效应
差异

\hookrightarrow 治疗参与与不治疗参与之差

08 工具变量 Instrumental Variables

当回归时：

$$Y_i = \beta_0 + K T_i + \beta W_i + u_i$$

T: 干预变量

当无 W_i 的数据：

$$\text{运行: } Y_i = \beta_0 + K T_i + v_i$$

W : 混淆因子

$$\rightarrow v_i = \beta W_i + u_i \quad \leftarrow \text{有偏} \quad \text{Cov}(T, v) \neq 0, K \text{ 的有偏估计}$$

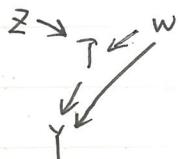
连接 T 与 V 的信息，不过

当加入仅与 T 相关（仅通过 T 与结果相关）的工具 Z： $\text{Cov}(Z, v) = 0$

对于 $Z \rightarrow Y$:

$$\begin{aligned} \text{Cov}(Z, Y) &= \text{cov}(Z, \beta_0 + K T_i + v_i) \\ &= K \text{Cov}(Z, T) + \text{cov}(Z, v) = K \text{Cov}(Z, T) \end{aligned}$$

$$K = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, T)} = \frac{\text{Cov}(Y_i, Z_i)/V(Z_i)}{\text{Cov}(T_i, Z_i)/V(Z_i)} \quad \begin{array}{l} \leftarrow Z \text{ 对 } Y \text{ 的影响 (仅能过 } T \text{)} : \text{ 简化形式系数} \\ \text{分子分母都是回归系数} \end{array}$$



$$K = \frac{\frac{\partial Y}{\partial Z}}{\frac{\partial T}{\partial Z}} = \frac{\frac{\partial Y}{\partial Z}}{\frac{\partial Y}{\partial T}} \cdot \frac{\partial T}{\partial Z} = \frac{\partial Y}{\partial T}$$

(0.1)

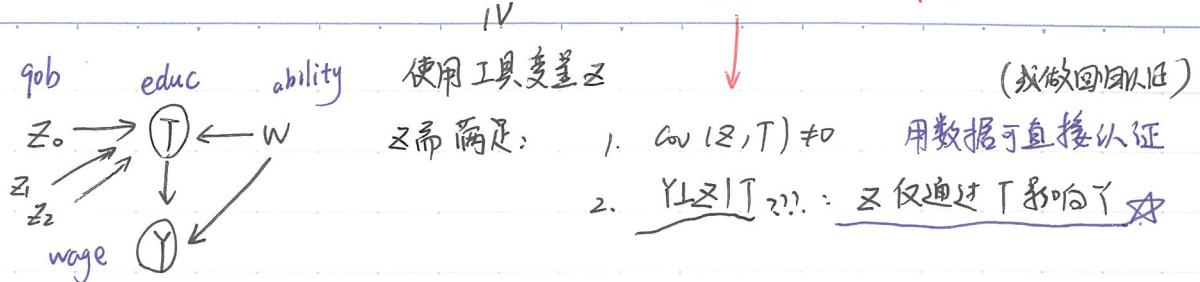
当工具变量为虚拟变量：

$$K = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[T|Z=1] - E[T|Z=0]} \quad \text{Wald Estimator}$$

△ 目前得 $Z \rightarrow Y$

后转化: $T \rightarrow Y$

不准确的 IV 修正



Step: 验证①: $T \sim Z_0$ \Rightarrow ①成立。

$Y \sim Z_0$ \Rightarrow ②有影响

{ 法1: $T \sim Z_0$ 与 $Y \sim Z_0$ 回归系数比 得简化的 $Y \sim T$

法2: 2SLS:

$$1. T \sim Z_0, Z_1, Z_2$$

$$2. Y \sim T, Z_1, Z_2$$

$$\sim (Z_0, Z_1, Z_2), Z_1, Z_2$$

编程时 $T \sim Z_0$, 因为会自动添加
2个元素

对 OLS: $Y \sim T + C(Z_0, Z_1, Z_2)$

结论: T 与 Z 相关性弱时, 估计值 between OLS & 2SLS 有异 (SE 会增加)

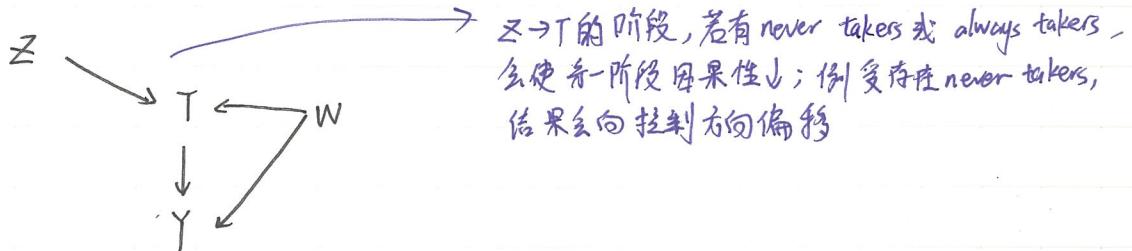
2SLS 是有偏的

{ 2SLS 偏向 OLS, 偏差方向(正负)一致
2SLS 一致, OLS 不一致

偏差会随工具变量个数而变, 当 IV 太多: 2SLS 会更像 OLS

09: 非服从性与局部平均效应

compliers	永远服从安排
never takers	永不从 T
always takers	永远是 T
defiers	反着来



Y_i 变为双参数: $Y_i(1,0)$ if $T_{i1} = 1, Z_i = 0$
 T_0 if $Z_i = 0$



- 工具变量假设 :
- $T_{i1}, T_{i2} \perp\!\!\!\perp Z_i$
 $Y_i(T_{i1}, 1), Y_i(T_{i1}, 0) \perp\!\!\!\perp Z_i \rightarrow Z_i \text{ 与潜在干预无关}$,
 不同工具变量组的人具有可比性
 - $Y_i(1, 0) = Y_i(1, 1) = Y_{i1}$
 $Y_i(0, 0) = Y_i(0, 1) = Y_{i0} \rightarrow \text{工具变量不影响潜在干预, 工具变量仅通过干预影响结果}$
 - $E[T_{i1} - T_{i0}] \neq 0$: 第一阶段存在, 工具变量确实会影响干预
 - $T_{i1} > T_{i0}$, 工具变量若被每个人执行, 干预水平会提高

Wald 估计:

$$\text{对 } ATE = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[T|Z=1] - E[T|Z=0]}$$

分子 $E[Y|Z=1] = E[Y_{i0} + T_{i1}(Y_{i1} - Y_{i0})|Z=1]$ 排除限制

$$= E[Y_{i0} + T_{i1}(Y_{i1} - Y_{i0})]$$
 由于独立性

同理 $E[Y|Z=0] = E[Y_{i0} + T_{i0}(Y_{i1} - Y_{i0})]$

分子: $E[(Y_{i1} - Y_{i0})(T_{i1} - T_{i0})]$ 0.25

$$= E[(Y_{i1} - Y_{i0})|T_{i1} > T_{i0}] P(T_{i1} > T_{i0})$$

分母: $E[T|Z=1] - E[T|Z=0] = E[T_{i1} - T_{i0}] = P(T_{i1} > T_{i0})$

n)

$$ATE = \frac{E[(Y_{i1} - Y_{i0})|T_{i1} > T_{i0}] P(T_{i1} > T_{i0})}{P(T_{i1} > T_{i0})} = E[(Y_{i1} - Y_{i0})|T_{i1} > T_{i0}]$$

这为 ATE 子群的条件,
即服从

IV 仅查找服从者的处理结果

$$\begin{cases} compliers: & T_{i1} > T_{i0} \\ never takers: & T_{i1} = T_{i0} = 0 \\ always takers: & T_{i1} = T_{i0} = 1 \\ defiers: & 对其干预不效 \end{cases}$$

10. 匹配 Matching

对于 ATE 的计算，由于混淆因素进行分层时，而对层加权计算总 ATE 加权：

- ① 按样本容量加权：ATE 接近样本容量多的组，但其方差可能小，因果不明显
非参数估计
- ② 按方差加权：更高 var 更高 weight：回归方法：线性、参数化、
高方差特征

①

子分类估计器 the subclassification estimator

当 X 未混淆，以 X 分层 (X 相同) 当 $(Y_0, Y_1) \perp | X$ ：

$$\text{ATE} = \int (E[Y|X, T=1] - E[Y|X, T=0]) dP(x)$$

以 X 对 K 个单元格 $\{x_1, x_2, \dots, x_K\}$

$$\hat{\text{ATE}} = \sum_{i=0}^K (\bar{Y}_{ki} - \bar{Y}_{k0}) * \frac{N_k}{N}$$

求和

②

KNN 方法

匹配估计器：matching estimator K nearest neighbour algorithm

利用样本间的距离定义相似度，在相似及接近的，但干预不同的个体间做对比

定义距离时需要归一化， $\frac{|x_i - x_j|}{\max}$

$$\hat{\text{ATE}} = \frac{1}{N} \sum_{i=0}^N \underbrace{(Z_{Ti} - 1)}_{(18)} \underbrace{(Y_i - Y_{jm(i)})}_{与 Y_i 最相似的另一个干预组的样本}$$

给药情况

Matching 的偏差

考虑 ATE 估计器 ATET:

$$\hat{ATET} = \frac{1}{N_1} \sum (Y_i - Y_{j(i)})$$

不收敛 $\sqrt{N_1}$ 增长快 —— 减小
 $\mu_0(x) = E[Y]x=x, T=0$

而中心极限的期望:

$$E[\sqrt{N_1} (\hat{ATET} - ATET)] = E[\sqrt{N_1} \underbrace{(\mu_0(X_i) - \mu_0(X_{j(i)}))}_{\downarrow}]$$

更新, 以收敛

X_i 的 y_j , 但 X_i 已处理 (反事实)

X_j 的 y_i , 但 X_j 未处理 (事实)

($x_i \approx x_j$, $y_{ji} \approx y_{ij}$)

$$\hat{ATET} = \frac{1}{N_1} \sum ((Y_i - Y_{j(i)}) - (\hat{\mu}_0(X_i) - \hat{\mu}_0(X_{j(i)})))$$

* 为估计, 可通过 LS 得到

影响 LS, 不影响 matching 估计器的核心

The curse of dimensionality 维度的诅咒

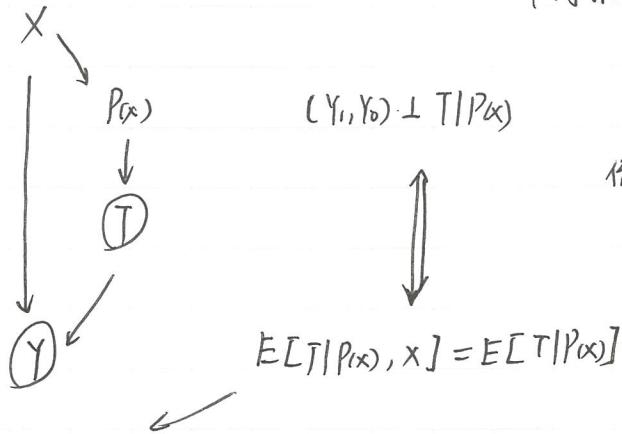
特征越多, 匹配距离越大

→ 单元格样本越少

11. 倾向得分

另一种处理混淆因素带来的偏差的方法：

(将混淆因素带来的倾向得分)



例：2个个体：一个干预组，一个对照组
但二者接受干预期相同（均为1）
则其为可比数据

左边： $E[T | P(x), X] = E[T | X] = P(x)$ (因为 $P(x)$ 只是 X 的函数)

右边： $E[T | P(x)] = E[E[T | P(x), X] | P(x)] = E[P(x) | P(x)] = P(x)$

IPFW

倾向加权：

$$E[Y|X, T=1] - E[Y|X, T=0] = E\left[\frac{Y}{P(x)}|X, T=1\right]P(x) - E\left[\frac{Y}{(1-P(x))}|X, T=0\right](1-P(x))$$

T=1

自我理解：

对 i 处理了 T，但每个 i 得到 T=1 的概率只有 $P(x)$ [由于 X 的原因]：

全：每个处理了 T 的 i，加权 $\frac{1}{P(x)}$ 以修正：| $P(x)$ 小，则其对 Y 的贡献小，模拟其先验分布

$T_i, m_i \rightarrow \text{加权 } \frac{1}{P(x)}$

对 T=1 没有处理： $\frac{1}{P(x)}, \frac{1}{(1-P(x))}$ 得： $Y_1: \rightarrow \text{全} \rightarrow T=1$ $Y_0: \rightarrow \text{全} \rightarrow T=0$ 而 $ATE = Y_1 - Y_0$

标准误差

$$IPFW: b_w^2 = \frac{\sum_{i=1}^n w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^n w_i}$$

但估计 $P(x)$ 也有误差

存在的问题：若不小心添加了非混淆因子的变量，会使倾向得分估计量方差巨大
可能：接受干预期的人比未接受干预期的人有更高的被干预概率

权重高于20会明显增加方差

倾向得分匹配 Propensity score Matching

(这东西太折磨)

将 $P(x)$ 替换 距离 D

- ① 提高接受干预可能的预测性 不会使因果估计更好，其会增加方差
- ② 若 干预和未干预的 倾向得分分布之间无良好重叠，则其会遇到问题

12. 双重稳健估计 (Doubly Robust Estimation) ← 二者结合

估计 $E[Y|T=1] - E[Y|T=0] | x$ | 线性回归
倾向得分加权

$$\hat{ATE} = \frac{1}{N} \sum \left(\frac{T_i(Y_i - \hat{\mu}_1(x_i))}{\hat{p}(x_i)} + \hat{\mu}_1(x_i) \right) - \frac{1}{N} \sum \left(\frac{(1-T_i)(Y_i - \hat{\mu}_0(x_i))}{1-\hat{p}(x_i)} + \hat{\mu}_0(x_i) \right)$$

$$\begin{aligned} & \text{第一部 } E[Y_1] \\ & = E[Y] \end{aligned}$$

$\hat{p}(x)$ 为倾向得分的估计 (逻辑回归等)

$\hat{\mu}_1(x)$ 为 $E[Y|x, T=1]$ 的估计

且而 $\hat{p}(x)$ 或 $\hat{\mu}(x)$ 其一正确

$\hat{\mu}_0(x)$ 为 $E[Y|x, T=0]$ 的估计

①

对 $\hat{E}[Y_1] = \frac{1}{N} \sum \left(\frac{T_i(Y_i - \hat{\mu}_1(x_i))}{\hat{p}(x_i)} + \hat{\mu}_1(x_i) \right)$ 当 $\hat{p}(x)$ 错误;

对 $T_i=0$,

$T_i=1, E[\hat{\mu}_1(x_i)] = Y_i$ ★
左边为 0; 仅用了 $\hat{\mu}_1(x_i)$

②:

$\hat{E}[Y_1] = \frac{1}{N} \sum \left(\frac{T_i Y_i}{\hat{p}(x_i)} - \left(\frac{T_i - \hat{p}(x_i)}{\hat{p}(x_i)} \right) \hat{\mu}_1(x_i) \right)$ 当 $\hat{\mu}_1(x_i)$ 错误

$E[T_i - \hat{p}(x_i)] = 0$ ★

理论正确, 跑数据正确,

使用单一模型时尽量用 DRE

14. 面板数据和固定效应 (避免混淆因素的不差)

面板数据：多个时间段内对同一人进行观察

→ 保持人这个变量固定

实现人对应的 IV：固定效应模型

引入，之前的 DID：D 处理。

$$\hat{Y}_0(1)/D=1 = \underbrace{Y_0(0)/D=1}_{\text{处理的城市处理前}} + \underbrace{(Y_0(1)/D=0 - Y_0(0)/D=0)}_{\text{未处理城市随时间的变化。}}$$

$$\hat{ATT} = \underbrace{Y_1(1)/D=1}_{\text{已知}} - \underbrace{\hat{Y}_0(1)/D=1}_{\text{估计}}$$

在 POA 市放广告牌的效果

平行趋势：治疗分配与潜在结果随时间的增长无关

$$(Y_d(t) - Y_d(t-1)) \perp D$$

→ 随时间变化的可变量。

固定效应：

$$y_{it} = \beta x_{it} + \gamma u_{it} + e_{it}$$

↑ 下 该一组不可见值，与 t 无关

个体 i 在 t 时的结果

13. 双重差分

Difference-in-Differences

评估宏观干预的影响

PID 估计量：D 表示 treatment T 表示时间，设 $Y_0(T)$ 是干预 D 在时间段 T 的潜在结果

① $\hat{ATE}_T = E[Y_1(D=1) - \underbrace{Y_0(D=1)}_{\text{反事实}}] \quad \text{同一时间内, 有干预-无干预}$

其中 $E[Y_0(D=1)] = E[Y_0(D=0)]$

用 $E[Y_0(D=1)]$ 估计 $E[Y_0(D=0)]$ 假设相等

该假设可能失效, 若干预无干预趋势, 二者不等

② $\hat{ATE}_T = E[Y_{1,0}] - E[Y_{0,0}] \quad \text{干预组与对照组}$

用 $E[Y_0(D=0)]$ 估计 $E[Y_{0,0}]$ 不可能失效,

$\hat{ATE}_T = (E[Y_{1,0}] - E[Y_{0,0}]) - (E[Y_{1,0}] - E[Y_{0,0}])$

(偏性校正有交叉)
(方差不知)

1 美 | 对 0

检查平行

14附:

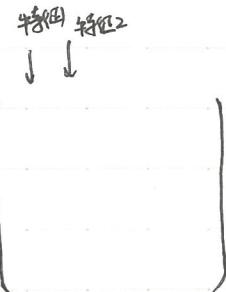
面板数据无用

面板数据的线性回归模型

特征 X_1 和另一组特征 X_2

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

- ① 逆向因果 reversed causality: 互为因果, 相互促进, 打架
 ② 不可测 混淆因素与时间相关



1. 以第二组特征回归 y : $\hat{y}^* = \hat{\beta}_1 X_2$
2. $\hat{X}_1 = \hat{\beta}_2 X_2$
3. 得到 $\tilde{X}_1 = X_1 - \hat{X}_1$ 和 $\tilde{y}_i = y_i - \hat{y}^*$
4. 将结果的残差回归到特征残差 $\tilde{y} = \hat{\beta}_1 \tilde{X}_1$

估计过程:

$$1. \hat{Y}_{it} = Y_{it} - \bar{Y}_i \quad \hat{X}_{it} = X_{it} - \bar{X}_i$$

2. 在 \hat{X}_{it} 上回归 \hat{Y}_{it}

$$U_{it} \text{ 消失}, (\bar{U}_{it} = U_{it})$$

时间效应

如通货膨胀

但可能也有随时间变的事物,

增加时间段维度,

每时段增加一个虚拟变量

 \Rightarrow

$$\begin{cases} Y_{it} = \beta X_{it} + \gamma U_{it} + e_{it} \\ \bar{Y}_i = \beta \bar{X}_{it} + \gamma \bar{U}_{it} + \bar{e}_{it} \end{cases}$$

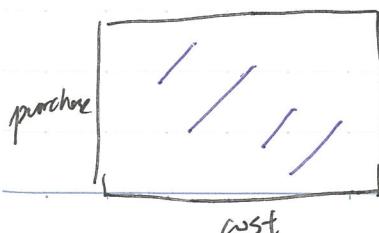
 \Rightarrow

$$(Y_{it} - \bar{Y}_i) = (\beta X_{it} - \beta \bar{X}_{it}) + (\gamma U_{it} - \gamma \bar{U}_{it}) + (e_{it} - \bar{e}_{it})$$

$$Y_{it} = \beta \hat{X}_{it} + \hat{e}_{it} \quad \text{消除所有未观察到的随时间不变的事物}$$

本质消除

按个体对数据进行分组并求得标准差总和, 若为0, 则其对任一个体, 该变量都不随时间变化



固定效应模型对每个城市拟合一条回归线

假设因果效应在所有个体中为常数

15 合成控制

(面板数据需要颗粒化的数据，若只有聚合数据，
或找不到相似对照，用多个单元组合为一个相似单元)

有 $J+1$ 个单元 : 单元 j 是受干预影响的单元

$2, 3, \dots, J+1$ 未处理 \rightarrow '供体池' donor pool

观察结果: $\begin{cases} Y_{jt}^I & \text{有干预的潜在结果} \\ Y_{jt}^N & \text{无干预 潜在结果} \end{cases}$ 共 J 个单元, 时间 t

干预
 T_0

当 $t > T_0$: $j=1$ 在 t 的影响 $Z_{1t} = Y_{1t}^I - \underbrace{Y_{1t}^N}_{\text{反事实}} \downarrow$

与 T 相关的函数

为了估计 Y_{jt}^N , 给定权重 $W = (w_2, \dots, w_{J+1})$

$$\hat{Y}_{jt}^N = \sum_{j=2}^{J+1} w_j Y_{jt} \quad \text{这可用回归弄}$$

找权重: OLS, 找最小化干预前期间供体池中单位的加权平均值与治疗单位之间的平方距离

合成控制作为线性回归的实现

① 用供体池构建与加州相似的虚拟州

$$\text{cigsale} \left[\begin{array}{cccc} \text{state}_1 & \text{state}_2 & \dots & \text{state}_{39} \\ \text{year1} \\ \text{year2} \\ \vdots \\ \text{years} \end{array} \right]$$

以加州为Y，X为其它州 $Y \sim X$

得 $\text{weight} = ([-0.436, \dots, -0.032])$

39个数，均为39个州的权重，加权后为虚拟州

以最小化治疗单位与供体池中单位之间的平方差

→ 该方法容易过拟合： 虚拟州在干预前与加州几乎重叠；干预后波动大且超过加州
 T大 N 大 (可用 Ridge 或 Lasso 回归)

② 一种避免过拟合的方法

之前过拟合的原因：外推（某-数据因为权重离谱而被外推到一个不可能存在的值）

凸组合；合成控制仍定义为：

$$Y_{it}^N = \sum_{j=2}^{J+1} w_j Y_{jt}$$

使用最小化的权重 $w = (w_2, \dots, w_{J+1})$

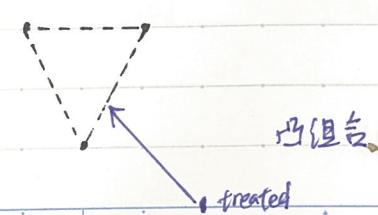
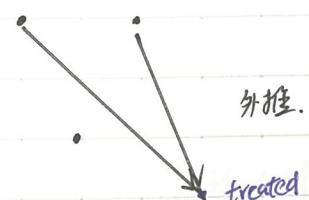
$$\|X_i - X_0 w\| = \left(\sum_{h=1}^k v_h (X_{hi} - \sum_{j=2}^{J+1} w_j X_{hj})^2 \right)^{\frac{1}{2}}$$

w_2, \dots, w_{J+1} 为正且其和为1。

每个变量的重要性；
以在最小化处理和合成控制之间的差异反映】

不同 v_h 会给出不同的权重

选择 v ： ①使每个变量都有均值为0和单位方差
②使更好的强调 Y 的变量有更高重要性



我们简单地为每个变量赋予相同的重要性：

代码：① 损失函数 ② scipy 二次规划优化，权重总和为 1 ③ 优化界限设置在 0~1
值得权重：

$[0, 0, 0, 0.00852, \dots]$ 和为 1，且释放，小 ($0 \sim 1$)

如何确定结果在统计上显著？：

Fisher 精确检验：

对所有 (39) 个州 都运行合成控制和效果估计

并计算



(这步应对未干预的州找无法找到任何显著的干预效果)

对加州：干预效果

其他州：安慰剂效应 (对照)

得到结果：

| 干预后效益高于干预前

| 干预前有些单化也无法很好地拟合

(删除某些州)



均方误差

设置干预前背景的拟值 $MSE = \frac{1}{N} \sum (Y_t - \hat{Y}_{t, \text{background}})^2$

检验显著：④ 得到所有州效应

⑤ P 值设为 $PV = \frac{1}{N} \sum I\{\hat{Z}_{\text{coeff}} > \hat{Z}_j\}$

(检验加州效应大于 0)

当你只有一个州) 的效应高于加州,

$P=0.0286$. 加州干预效果估计很

(有效) 极端,

16. 断点回归 Regression Discontinuity Design (RDD)

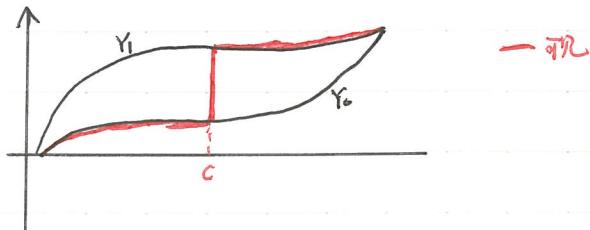
跳跃：受影响？

设干预变量 T ，潜在结果： T 为连续变量 R 的不连续函数，s.t.

$$D_i = I(R_i > c)$$

当 R 低于阈值 c 时处理为 0，否则为 1

$$\begin{cases} R > c \text{ 可观察 } Y_1 \\ R \leq c \text{ 可观察 } Y_0 \end{cases}$$



直观： $x: \text{age}$ · $y: \text{all/mva/suicide}$ \Rightarrow 可见 Y 与年龄的关系，断之

RDD 估计：若平滑，潜在结果的极限应该是相同的

$$\lim_{r \rightarrow c^+} E[Y_{ti} | R_i = r] = \lim_{r \rightarrow c^-} E[Y_{ti} | R_i = r]$$

若不平滑，间断处：

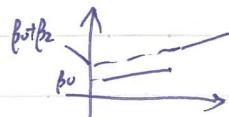
$$\begin{aligned} \lim_{r \rightarrow c^+} E[Y_{ti} | R_i = r] - \lim_{r \rightarrow c^-} E[Y_{ti} | R_i = r] &= \lim_{r \rightarrow c^+} E[Y_{ti} | R_i = r] - \lim_{r \rightarrow c^-} E[Y_{ti} | R_i = r] \\ &= E[Y_{ti} | R_i = r] - E[Y_{ti} | R_i = r] \\ &= E[Y_{ti} - Y_{ti} | R_i = r] \end{aligned}$$

局部平均干预结果 (LATE) \rightarrow

局部随机试验，只是接近 RCT

估计 干预效果：线性回归

$$y_i = \underbrace{\beta_0 + \beta_1 r_i}_{\text{低于阈值的回归}} + \underbrace{\beta_2 I(r_i > c)}_{\text{高于阈值的回归}} + \underbrace{\beta_3 I(r_i > c)r_i}_{\text{高于阈值的回归}}$$



摘要：

RDD 回归

$$y_i = \beta_0 + \beta_1 r_i + \beta_2 \mathbb{1}_{\{r_i > c\}} + \beta_3 \mathbb{1}_{\{r_i < c\}} r_i$$

$$\beta_0 = \lim_{r \rightarrow c^-} E[Y_i | R_i = r]$$

$$\beta_0 + \beta_2 : r \rightarrow c^+$$

$$\therefore \beta_2 = \lim_{r \rightarrow c^+} E[Y_i | R_i = r] - \lim_{r \rightarrow c^-} E[Y_i | R_i = r] = E[\text{ATE} | R = c]$$

标准误差，流汗是性

内核加权 Kernel Weighting

回归不连续性在很大程度上依赖于线性回归的外推特征

回归可能过于关注拟合其他数据点，

1. 给更接近阈值的点赋予更高的权重

使阈值处拟合不佳

triangular kernel 加权样本

$$K(R, c, h) = \mathbb{1}\{|R - c| \leq h\} * \left(1 - \frac{|R - c|}{h}\right)$$

离 C 远, weight ↓

wls

带权最小二乘

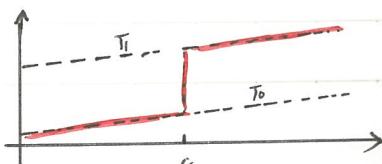
↓

带宽参数 h：是否接近阈值：只考虑 $c \pm h$ 内的数据

模糊 RDD： Fuzzy. RDD

(在断点处样本不分为 (0,1), 而是有概率)

值视为 IV, 若忽略 treatment 偏向。流分成,



$T_{1i} > T_{0i} \forall i$

假设潜在为单

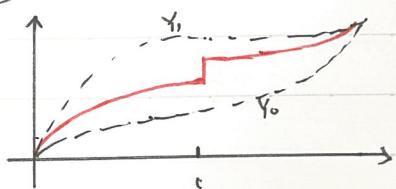
LATE 的 Wald Estimator *

$$\lim_{r \rightarrow c^+} E[Y_i | R_i = r] - \lim_{r \rightarrow c^-} E[Y_i | R_i = r]$$

$$\lim_{r \rightarrow c^+} E[T_i | R_i = r] - \lim_{r \rightarrow c^-} E[T_i | R_i = r]$$

$$= E[Y_{1i} - Y_{0i} | T_{1i} > T_{0i}, R_i = c]$$

| 只估计 RDD 位置
IV 位置



当P值大，说明不显著，则效应弱

麦克雷测试 McCrary Test 运行差程密度上的聚类现象 bunching (阈值周围是否有任何尖峰)

Part II: 应用

17. 预测模型 101 Predictive Models 101

ML: 估计 $E[Y|X]$ Y 在 X 下的期望



过复杂的模型会过拟合

交叉验证: 训练集 | 测试集

Unseen data

预测 和 治疗
predictions policies

估计器, 最大深度

R^2 函数: 用于评估预测连续变量的模型. $(-\infty, 1)$

说明结果中有多少方差由模型解释

对于平均值为负

当 depth ↑
在训练集 R ↑
测试 R ↓ \Rightarrow 过拟合

pandas: pd.qcut. 用模型预测的数据划分为分位数

模型波段

细粒度策略: 阈值 \rightarrow 连续

参与 or 不参与 要投入多少

18. 异质干预效应与个性化

与ML不同处：干预干预干预：

$$E[Y|X, T]$$

$\left\{ \begin{array}{l} X: \text{无法控制的外生特征} \\ T: \text{处理} \end{array} \right.$

(例：销售额)

(例：价格)

$$\text{优化: } \underset{T}{\operatorname{argmax}} E[Y|X, T]$$

从ATE到CATE：

平均干预效果 $E[Y_i - Y_0]$

有效的连续干预 $E[y^{(t)}]$

$y^{(t)}$ 为响应函数或结果的处理导数

条件平均干预效果 (CATE)

$E[Y_i - Y_0 | X]$ 或 $E[y^{(t)} | X]$

个性化干预

加入条件 X ：我们对每个 X 分开求 E ，并制定个性化方案

预测弹性：

$$\frac{\delta y_i}{\delta t_i} \quad \text{近似} \quad \frac{Y(t_i) - Y(t_i + \epsilon)}{t_i - (t_i + \epsilon)} \quad \text{但不可用 (因果推断)}$$

简单回归：

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 X_i + e_i \quad \frac{\delta y_i}{\delta t_i} = \beta_1 \quad \text{只能预测弹性}$$

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 X_i + \underline{\beta_3 t_i X_i} + e_i \quad \frac{\delta y_i}{\delta t_i} = \hat{\beta}_1 + \hat{\beta}_3 X_i \quad \checkmark$$

交互项 \star 特征向量

$$\frac{\delta y}{\delta t} = \frac{y(t+\epsilon) - y(t)}{t+\epsilon - t}$$

可用 β_1, β_3 写出 $y(t)$ 关系

类似：

$$\frac{\delta y}{\delta t} \approx \hat{y}(t+\epsilon) - \hat{y}(t)$$

19. 评估因果模型 (⑩ 强化：排序(分组))

使用非随机数据估计因果模型

对称 \downarrow 矩阵 \downarrow

② model1: $\text{sales}_i = \beta_0 + \beta_1 \text{price}_i + \beta_2 x_i + \beta_3 x_i \text{price}_i + e_i$ 强制 B2Z

① model2: $\text{sales}_i = G(x_i, \text{price}_i) + e_i$ 因果图示

③ model3: 输出随机数作为限制 (用来对冲) 随机校正
检查 R^2 (是否过拟合)

弹性模型萃: (来源于哪些单位对干预更敏感)

单变量线性回归估计弹性: $y_i = \beta_0 + \beta_1 t_i + e_i$

price 分段输出段内弹性 $\hat{\beta}_i = \frac{\sum_{i=1}^k (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^k (t_i - \bar{t})^2}$ \bar{t} 处理的均值.
 \bar{y} 输出的均值

累积弹性曲线: 最敏感的放在一组, 依次包含, 直到所有组 $\rightarrow ATE$

评估

$$\hat{y}(t)_k = \hat{\beta}_k = \frac{\sum_{i=1}^k (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^k (t_i - \bar{t})^2}$$

目标:

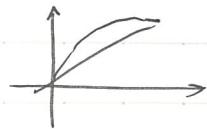
① $\hat{y}(t)_k > \hat{y}(t)_{k+1}$

② $\hat{y}(t)_k - \hat{y}(t)_{k+1}$ 最大



累积增加曲线

$$\hat{F}(t)k = \hat{\beta}_1 k * \frac{k}{N} = \frac{\sum_i^k (t_i - \bar{t})(y_i - \bar{y})}{\sum_i^k (t_i - \bar{t})^2} * \frac{k}{N}$$



置信区间: α :

$$S_{\beta_1} = \sqrt{\frac{\sum_i \hat{\epsilon}_i^2}{(n-2)\sum_i (t_i - \bar{t})^2}}$$

之后的 translation 为参考:

20. Plug-and-Play Estimators

之前: 当 $Y(0), Y(1) \perp X$ (干预对照组可交换) 可得 ATE.

现在: treatment 对谁效果更好? 更 sensitive

个人治疗效果

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

$$Y_i^{\text{obs}}(t) = \begin{cases} Y_i(1), & \text{if } t=1 \\ Y_i(0), & \text{if } t=0 \end{cases}$$

$$\text{ATE: } \tau = E[Y_i(1) - Y_i(0)] = E[\tau_i]$$

$$\text{CATE: } \tau(x) = E[Y_i(1) - Y_i(0)|x] = E[\tau_i|x]$$



$$y_i = \beta_0 + \beta_1 t_i + \beta_2 x_i + \beta_3 t_i x_i + e_i$$

$$\text{w.l.: } \hat{\tau}(x) = \hat{\beta}_1 + \hat{\beta}_3 x_i$$

→ 结果不关心 β_2 , 但若 x 与结果无关, 则 $\hat{\beta}_1, \hat{\beta}_3$ 有偏差

均方误差 MSE

$$E[(Y_i - \hat{Y}_i)^2] \leftarrow \text{结果的 MSE}$$

★ 目标: st. 治疗效果的 MSE

$$\underbrace{E[(\hat{\tau}(x)_i - \tau(x)_i)^2]}_{\min} = E[(Y_i(1) - Y_i(0) - \hat{\tau}(x)_i)^2] \text{ 最小化}$$

估计 $E[(Z(x)_i - \hat{z}(x))^2]$ 使其最小

中间变量: $Y_i^* = 2Y_{(1)} * T_i - 2Y_{(0)} * (1-T_i)$

① 当随机分配 (5.5) : $T \perp Y_{(1)}, Y_{(0)}$

$$E[T, Y_{(t)}] = E[T] * E[Y_{(t)}]$$

$$\text{Also: } Y_i * T_i = Y_{(1)i} * T_i \quad \text{and} \quad Y_i * (1-T_i) = Y_{(0)i} * (1-T_i)$$

从而可推

$$\begin{aligned} E[Y_i^* | X_i = x] &= E[2Y_{(1)i} * T_i - 2Y_{(0)i} * (1-T_i) | X_i = x] \\ &= 2E[Y_{(1)i} * T_i | X_i = x] - 2E[Y_{(0)i} * (1-T_i) | X_i = x] \\ &= 2E[Y_{(1)i} | X_i = x] * 0.5 - 2E[Y_{(0)i} | X_i = x] * 0.5 = Z(x)_i \end{aligned}$$

② 当非随机分配治疗, 引入 $e(x_i)$ 倾向性得分 / 当治疗以 p 进行, 用 p 替换 $e(x_i)$

$$Y_i^* = Y_i * \frac{T_i - e(x_i)}{e(x_i)(1-e(x_i))}$$

$$E[Y_i^* | X_i = x] = Z(x)_i$$

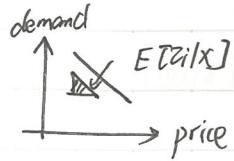
$$\begin{aligned} &\star = \frac{1}{e(x_i)} E[Y_{(1)i} | X_i = x] * E[T_i | X_i = x] - \\ &\frac{1}{1-e(x_i)} E[Y_{(0)i} | X_i = x] * E[(1-T_i) | X_i = x] \end{aligned}$$

从累积增益曲线可以看出: 转化后的目标 Y_i^* 是个对个体治疗效果非常敏感的估计
方差较大 (但在大数据下只是个小问题)

在连贯治疗中：

$$\bar{z}(x) = E[\partial Y_i(t) | X] = E[\tau_i | X]$$

↓



$$\hat{z}(x) = \text{Cov}(Y_i, \tau_i) / \text{Var}(\tau_i) = \frac{\sum (Y_i - \bar{Y})(\tau_i - \bar{\tau})}{\sum (\tau_i - \bar{\tau})^2}$$

对于单个单位：

$$Y_i^* = (Y_i - \bar{Y}) \cdot \frac{(\tau_i - \bar{\tau})}{\sigma_{\tau}^2} \quad (\text{随机治疗})$$

推导：

$$\because V_i = \frac{(\tau_i - \bar{\tau})}{\sigma_{\tau}^2} \quad E[V_i | X_i = x] = 0$$

有 $E[V_i | X_i = 0]$

$$E[T_i V_i | X_i = 0] = 0 \quad \text{条件} \quad E[\tau_i (\tau_i - \bar{\tau}) | X_i = x] = \underline{E[(\tau_i - \bar{\tau})^2 | X_i = x]} \\ E[T_i e_i | X_i = x] = E[\tau_i | X_i = x] E[e_i | X_i = x]. \quad \text{variance}$$

又 $Y_i = \alpha + \beta \tau_i + e_i | X_i = x$

$$E[Y_i^* | X_i = x] = E[(Y_i - \bar{Y}) V_i | X_i = x] \\ = E[(\alpha + \beta \tau_i + e_i - \bar{Y}) V_i | X_i = x] = \beta = z(x)$$

$$Y_i^* = (Y_i - \bar{Y}) \cdot \frac{(\tau_i - M(\tau_i))}{(\tau_i - M(\tau_i))^2} \quad (\text{连贯治疗}) M \text{ 为一个估计 } E[\tau_i | X_i = x] \text{ 改正} \\ \text{分子为校准，若只而比较排序，不需要分子}$$

非线性关系中：(目前无解决方式) $\text{maybe: } D_i = \frac{1}{P_i^{\alpha}}$

$$\log(D)_i = -\alpha * \log(P_i)$$

2). Meta Learners 元学习者 (一些ML方法)

仍使用非随机数据来训练模型，用随机数据验证，估计 CATE
 ↗ (去伪存真)

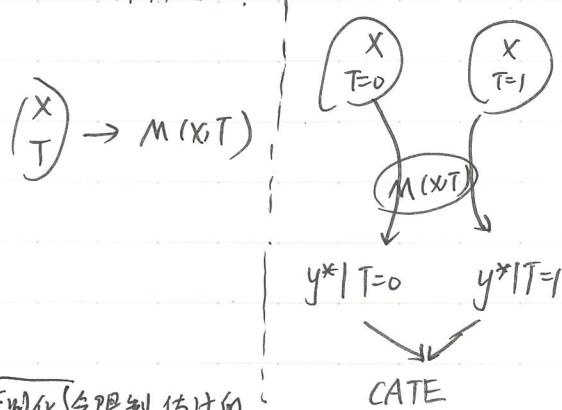
$$\begin{aligned} z_i(x) &= E[Y_{i(1)} - Y_{i(0)} | X] = E[\tau_i | X] \\ \text{or } E[\delta Y_i(t) | X] \end{aligned}$$

① S-Learner (Go-Horse Learner)

用ML model: MS 估计 $\mu(x) = E[Y | T, X]$

用累积增益函数可指是否过拟合：

(调整: max-depth)



缺点：倾向将治疗效果偏向。

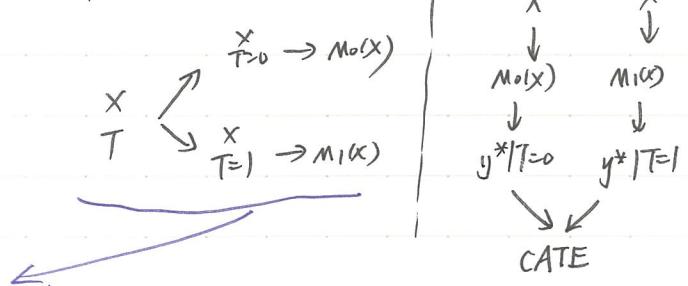
其 S-Learner 为正则化ML模型，正则化会限制估计的
 成当某些协变量影响很大，S-Learner 可能会抛弃处理变量 treatment effect

② T-Learner 通过分割解决完全抛弃处理的问题

$$\mu_0(x) = E[Y | T=0, X]$$

$$\mu_1(x) = E[Y | T=1, X]$$

$$\text{m: } \hat{\eta}(x)_i = \mu_1(x_i) - \mu_0(x_i)$$



将数据集缩小了： 例：T=0 数据很多，可避免过拟合

T=1 — 很少，模型简单，避免过拟合，使用正则化处理。

矛盾。

② X-Learner : 第二阶段，第一阶段与T-Learner近似，第二阶段作倾向性得分模型

$$\hat{M}_0(x) \approx E[Y|T=0, X]$$

} 假 T=0 or 1 时关系不同，替换

$$\hat{M}_1(x) \approx E[Y|T=1, X]$$

2.

$$\hat{\tau}(x, T=0) = \hat{M}_1(x, T=0) - Y_{T=0}$$

用 真值与缺失值的估计差异

$$\hat{\tau}(x, T=1) = Y_{T=1} - \hat{M}_0(x, T=1)$$

3.

$$\hat{M}_{20}(x) \approx E[\hat{\tau}(x) | T=0]$$

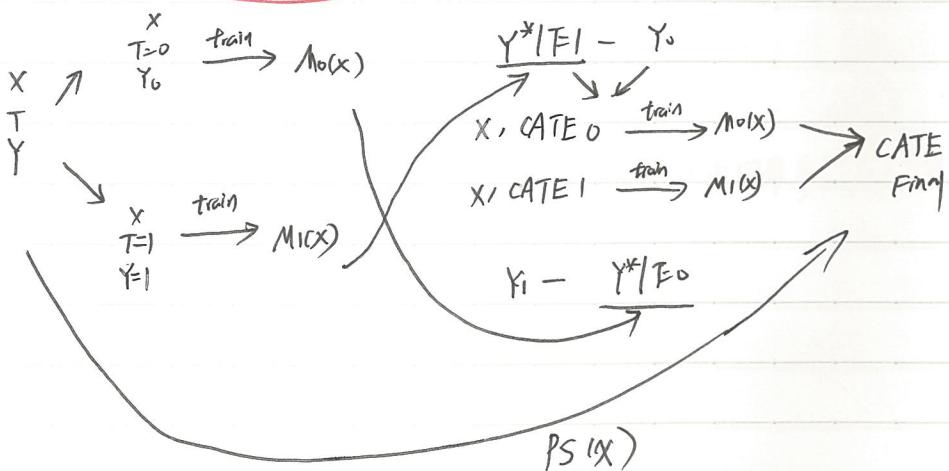
用 T=1 的数据估计出的模型，

$$\hat{M}_{21}(x) \approx E[\hat{\tau}(x) | T=1]$$

4.

$$\hat{\tau}(x) = \hat{M}_{20}(x)(\hat{e}(x)) + \hat{M}_{21}(x)(1-\hat{e}(x))$$

当 $\hat{e}(x)$ 越大，T=1 越多，更佳，
权重更大



22. Debiased / Orthogonal Machine Learning \star

去偏倚 / 正交机器学习

<< Double Machine Learning for Treatment and Causal
Parameters >> 2016

ML for Nuisance Parameters 滞扰参数的 ML

已有数据中有 bias, (因为 feature 间有相关) sales: price, temp, aust, weekday --
关系也可能是多次的

Frisch - Waugh - Lovell

有一个线性回归模型:

$$\hat{Y} = \hat{\beta}_1 \hat{X}_1 + \hat{\beta}_2 \hat{X}_2 \quad (\hat{X}_1, \hat{X}_2 \text{ 为特征矩阵, 每个特征一列, 每个观测量值一行, } \hat{\beta}_1, \hat{\beta}_2 \text{ 为行向量})$$

可通过以下步骤得到完全相同的 $\hat{\beta}$ 参数:

1. $\hat{y}^* = \hat{\gamma}_1 \hat{X}_2$

2. $\hat{x}_1 = \hat{\beta}_2 \hat{X}_2$

3. 得残差 $\tilde{x}_1 = x_1 - \hat{x}_1$ and $\tilde{y}_1 = y - \hat{y}^*$

4. 回归: $\tilde{y} = \hat{\beta}_1 \tilde{x}_1 \rightarrow \text{得 } \hat{\beta}_1$

FWL: $(Y - (Y \sim X)) \sim (T - (T \sim X))$ 实质是对以下假定中因果多效工的估计:

$$Y_i - E[Y_i | X_i] = Z \cdot (T_i - E[T_i | X_i]) + \varepsilon$$

FWL on steroids

FWL 理论适合 ML 方法：对 ATE

~~备注：~~

$$Y_i - \hat{M}_y(x_i) = z \cdot (T_i - \hat{M}_t(x_i)) + \varepsilon$$

↑ ↑

estimating $E[Y|X]$ estimating $E[T|X]$

denoising model

去除 Y 的方差

所有由 X 引起的方差都被
解释掉了

step:

1. Estimate the outcome Y with features X using a flexible ML regression model M_y
2. Estimate T X 所有偏差被模型 M_t 捕捉 debiased $\rightarrow M_t$
3. obtain the residuals $\tilde{Y} = Y - M_y(x)$ and $\tilde{T} = T - M_t(x)$ 移除 对治疗减损
4. regress the residuals of the outcome on the residuals of the treatment $\tilde{Y} = \alpha + z \tilde{T}$

二次拟合，一次回归

 z is the causal parameter ATE若过拟合：例 M_y 过拟合 $\rightarrow M_y$ 捕捉了不仅仅 $Y \sim X$ 间的关系（也许含了一些 $Y \sim T$ 的关系）则最后 z 值会偏向 0；（ M_y 捕捉了因果关系而不是将其留给了残差回归） M_t 过拟合 $\rightarrow M_t$ 解释了 T 中更多的方差 \rightarrow 处理残差的方差会更少若治疗方差过小，最终估计的方差偏大（当 T 为 X 的确定性函数，也会发生，
正向性被违反了）解决过拟合：数据分为 K 部分，对每部分，在其余 $K-1$ 个样本上估计 ML 模型，在 K 部分上做残差
最后结合所有 K 部分的预测，估计 $\tilde{Y} = \alpha + z \tilde{T}$

对 CATE

$$Y_i - M_y(x_i) = z(x_i) \cdot (T_i - M_t(x_i)) + \varepsilon_i$$

$$4. \quad \tilde{Y}_i = \alpha + \beta_1 \tilde{T}_i + \beta_2 x_i \tilde{T}_i + \varepsilon_i$$

$$5. \quad \hat{\mu}(1|sales_i, x_i) = M(\text{Price}=1, x_i) - M(\text{Price}=0, x_i)$$

 \Downarrow final model

R-learner

Non Parametric Double / Debiased ML (当 T 对 Y, B1 price 对 sales, ~~是因果关系~~)

对于连续型的关系: \rightarrow 预测因果 \rightarrow 预测治疗

$$\text{偏差误差项} \quad Y_i = \hat{M}_y(X_i) + \varepsilon_i(T_i - \hat{M}_t(X_i)) + \hat{\varepsilon}_i$$

$$\hat{\varepsilon}_i = (Y_i - \hat{M}_y(X_i)) - \varepsilon_i(T_i - \hat{M}_t(X_i)) \leftarrow \text{因果损失函数}$$

最小化这一损失的平方, 则可估计 ε_i

即 CATE

$$\hat{\Sigma}_n(\varepsilon_i) = \frac{1}{n} \sum_{i=1}^n [(Y_i - \hat{M}_y(X_i)) - \varepsilon_i(T_i - \hat{M}_t(X_i))]^2$$

R-loss: R-Learner 量化

化的东酉

等等:

$$\hat{\Sigma}_n(\varepsilon_i) = \frac{1}{n} \sum_{i=1}^n (\tilde{Y}_i - \varepsilon_i \tilde{T}_i)^2$$

$$= \frac{1}{n} \sum_{i=1}^n \tilde{T}_i^2 \underbrace{(\frac{\tilde{Y}_i}{\tilde{T}_i} - \varepsilon_i)^2}_{\text{最小化}}$$

括号内的内容为 predicting $\frac{Y_i}{T_i}$

即用一个线性模型去 predict $\frac{Y_i}{T_i}$ while using the weight $\frac{\tilde{T}_i^2}{\tilde{T}_i}$

非参数化: What is Non-Parametric About? 非参数化与非线性性

对于基本假设:

$$\tilde{Y}_i = \varepsilon_i \tilde{T}_i + \epsilon_i \quad \text{这是线性的}$$

Double/ML 会发现对非线性 CATE 的局部线性近似, (找到了某点的导数)

x_i

No.

Date

步骤： $Y: sales$ $T: 价格$ $X: 特征$.

\hat{Y} 由 5个LGBM 反复预测 $\sim X$.
 \hat{T} $\sim X$

$$\tilde{Y} = Y - \hat{Y}$$

$$\tilde{T} = T - \hat{T}$$

将 \tilde{Y}, X 打给 ML 来拟合 \tilde{Y} 行 (用 LGBM 例)
find N.

训练过程：

\hat{Y}' 用 5个LGBM 反复预测 $\sim X$

用 \hat{Y}' 替换 T_{new} 得 \hat{Y}'

↓ 我们假设的反事实数据，代入 (\hat{Y}', X) 得 \tilde{Y}' ，对每个 T_{new}
= $\frac{\text{该值为 } 0}{\text{该值为 } 1}$

用 \hat{Y}' 加上 \tilde{Y}' 得最终预测

→ 未完成

Non-Scientific Double / Debiased ML

非科学的双重/有偏见的ML

进行反事实预测：(ATE?)

从单个数据点构建整个结果曲线

$$\tilde{Y}_i = \mathbb{E}(X_i) \tilde{T}_i + e_i$$

修改

$\tilde{Y}_i = T(X_i, \tilde{T}_i) + e_i$ 一个奇怪的函数 用ML解决 (先预测 \tilde{T}_i , 再用 \tilde{T}_i, X 预测 \tilde{Y}_i)

先预测 \tilde{T}_i , 用 any ML model 得 $\hat{T}_i \sim T(X_i, \tilde{T}_i) + e_i$

LGBM

又有 $\hat{Price}_i = \hat{T}(X_i, \tilde{T}_i)$ 代入训练集计算 (处理过程与训练步骤相同)

利用了5倍交叉验证；训练集用5折

型的预测平均

一个预测值中，
避去

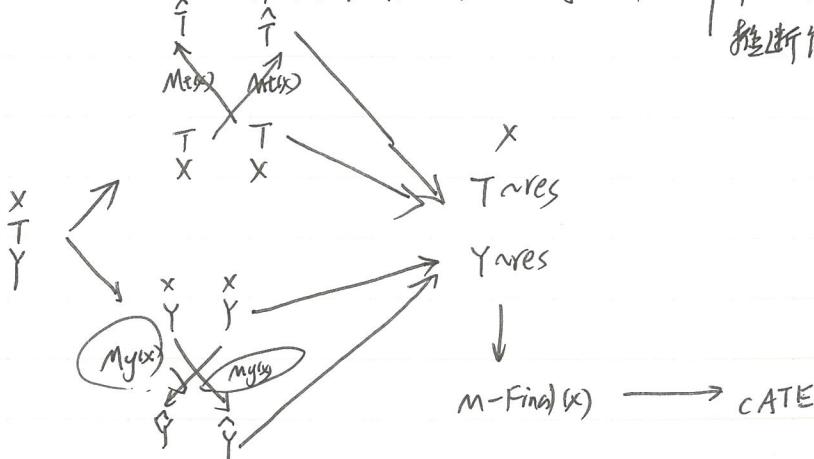
弊端：

① bias: 正则化使结果偏向于0

② 与选择的ML算法有关 (除了提升树 平滑预测效果不好)

推断能力不强：对以前从未有的

价格输出奇怪的
预测结果)



TEH 治疗效果异质性：对不同个体治疗效果不同

2) Effect Heterogeneity and Nonlinearity 效果异质性与非线性

例如：分析

$$Y_{latent} \sim N(-4.5 + 0.001 \text{income} + \text{nudge} + 0.01 \text{nudge} \cdot \text{age}, 1)$$

$$\text{nudge} \sim B(0, 5)$$

$$\text{age} \sim G(10, 4)$$

对所有样本效果相同

好的分类差异

$$\text{income} \sim G(20, 2)$$

设置

$$\text{conversion} = 1 \quad | \quad Y_{latent} > 0 \quad 50\%$$

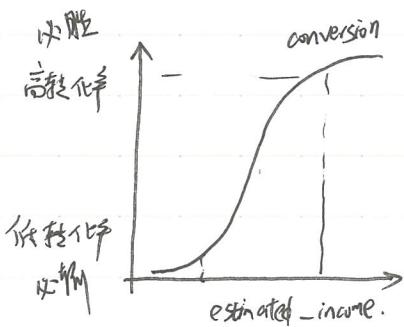
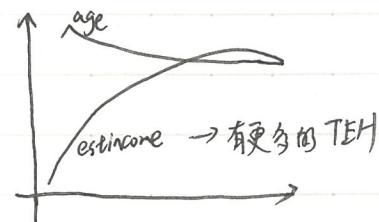
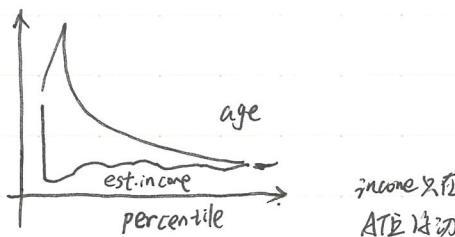
→ 真理 ATE 应为 1.4,

但 conversion = 0.394

因为 conversion 不是线性的

对于 effect on Latent outcome.

effect on conversion



收入对 conversion 有很强预测性：

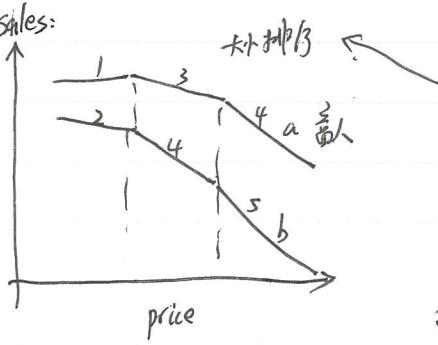
在两端转化率都极低，治疗效果小

当

转化率高：high baseline conversion 低治疗效果

转化率低：high baseline conversion 高治疗效果

第一种情况 Sales:



治疗效果与 a,b 与 price 都有关

可将 price 放元使 $y-x$ 线性；例 $-(\text{price}^4)$