体育经济分析: 原理与应用

单元4: 体育与计量经济学

周正卿

27 February 2023

大纲

大纲

- Level 1
  - 一个例子
- Level 2
  - 基本概念
- Level 3
  - 具体实战

潜在结果框架帮助理解“真相”

个体处置效应

- Y_i: 对个体的 i **观察结果**, 每个个体都有2个**潜在结果**
- D_i: 二元 **干预状态**

1. $Y_i(1)$ 若 $D_i = 1$
   表示： i 干预后的结果

1. $Y_i(0)$ 若 $D_i = 0$
   表示： i 没有被干预的结果

两者之差就是 **个体处置效应**,

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

- 个体处置效应存在异质性

因果推断的根本难点在于反事实无法观测

问题是 无法直接计算: $\tau_i = Y_i(1) - Y_i(0)$

- 数据上只能同时观察每个个体的 (Y_i, D_i)
- 永远无法**同时**观 $Y_i(0)$ 和 $Y_i(1)$, 必须借助**反事实（conterfactual）**概念

$\longrightarrow$ **两个潜在结果只能观测其一**，这就是Holland(1986)提出的**因果推断的根本难点**

系数的重新命名

- **个体处置效应：** $\tau_i = Y_i(1) - Y_i(0)$
  - 关键点: **因人而异**
  - 由于潜在结果根本矛盾而永远无法获得
- 作为替代转向**总体平均处置效应 (Average Treatment Effect)**： 用于描述处置效应的平均效果
  - $ATE = E[Y_i(1) - Y_i(0)]$，ATE只是这些异质性干预的平均值。
- 干预组平均处置效应(最关注的效应，是干预行为的直接后果):
  - $ATT = E[Y_i(1) - Y_i(0)|D_i = 1]$
- 控制组平均处置效应:
  - $ATU = E[Y_i(1) - Y_i(0)|D_i = 0]$
- 协变量条件平均处置效应:
  - $ATE(x) = E[Y_i(1) - Y_i(0)|D_i = 1, X_i = x]$

ATE与ATT、ATU的关系

- 总体平均处置效应 (ATE)

$$
\begin{aligned}
ATE &= E\left[Y_i(1) - Y_i(0)\right] \\
&= E\left[Y_i(1)\right] - E\left[Y_i(0)\right] \\
&= \omega \times ATT + (1-\omega) \times ATU
\end{aligned}
$$

- ATE是ATT和ATU的加权平均

观察结果

- 个体根据是否接受了干预而表现出来的潜在结果
- 可表示为潜在结果和干预状态的函数 $Y_i = Y_i(0) + [Y_i(1) - Y_i(0)] \times D_i$
- $D_i = 0$ 表示个体 i 没有接受干预, $Y_i = Y_i(0)$
- $D_i = 1$ 表示接受了干预, $Y_i = Y_i(1)$

所谓的“朴素”估计量

问题 既然 ATE、ATT和ATU均无法获得

简单方案:

直接比较 干预组 ($Y_i(1) \mid D_i = 1$) 和 控制组 均值, 即: ($Y_i(0) \mid D_i = 0$).

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$$

3种“朴素”估计偏误形式

$$
\begin{aligned}
& E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \\
& \quad = \underbrace{E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 1]}_{ATT \text{ 😀}} + \underbrace{E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]}_{ATT\text{估计偏差 😞}} \\
& \quad = \underbrace{E[Y_i(1) \mid D_i = 0] - E[Y_i(0) \mid D_i = 0]}_{ATU \text{ 😀}} + \underbrace{E[Y_i(1) \mid D_i = 1] - E[Y_i(1) \mid D_i = 0]}_{ATU\text{估计偏差 😞}} \\
& \quad = \underbrace{\omega \times (E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 1]) + (1-\omega) \times (E[Y_i(1) \mid D_i = 0] - E[Y_i(0) \mid D_i = 0])}_{ATE \text{ 😀}} \\
& \quad + \underbrace{\omega \times (E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]) + (1-\omega) \times (E[Y_i(1) \mid D_i = 1] - E[Y_i(1) \mid D_i = 0])}_{ATE\text{估计偏差 😀}}
\end{aligned}
$$

选择偏误

- ATE估计偏差 $= \omega \times$ ATT估计偏差 $+ (1-\omega)$ ATU估计偏差
  - 造成ATE 估计偏差的原因包含造成 ATT 和 ATU 估计偏差的原因
- 造成“朴素”估计量估计处置效应产生偏差的原因：
  1. **是否接受干预**不是随机的
  2. 原因都源于**接受干预与否**是个体自我选择的后果，称之为**选择偏误**（selection bias）

例子：吃药 → 健康

个体i	潜在结果		处置效应	处置状态	观测结果
	如果处置	如果未处置			
i	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$	D_i	Y_i
1	5	<u>2</u>	3	1	5
2	7	<u>3</u>	4	1	7
3	4	<u>1</u>	3	1	4
4	<u>3</u>	2	1	0	2
5	<u>8</u>	3	5	0	3

- “上帝”视角
- 阴影部分为观测结果，有下划线的部分为无法观测到的反事实结果

例子： 吃药 → 健康

- 干预组： $T1 = E[Y_i(1) \mid D_i = 1]$;　$T0 = E[Y_i(0) \mid D_i = 1]$ (反事实)
- 控制组： $C0 = E[Y_i(0) \mid D_i = 0]$;　$C1 = E[Y_i(1) \mid D_i = 0]$ (反事实)

平均潜在结果：如果处置	平均潜在结果：如果未处置	处置情况	平均观测结果
$T1 = E[Y_i(1) \mid D_i = 1] = 5.3$	$T0 = E[Y_i(0) \mid D_i = 1] = 2$ (反事实结果)	$D_i = 1$ (处置组)	$T1 = E[Y_i \mid D_i = 1] = E[Y_i(1) \mid D_i = 1] = 5.3$
$C1 = E[Y_i(1) \mid D_i = 0] = 5.5$ (反事实结果)	$C0 = E[Y_i(0) \mid D_i = 0] = 2.5$	$D_i = 0$ (控制组)	$C0 = E[Y_i \mid D_i = 1] = E[Y_i(0) \mid D_i = 0] = 2.5$

例子： 吃药 → 健康

若知道所有个体的潜在结果, 就可以得到准确的平均处置效应

- ATT (接受干预的个体的平均处置效应)$= T1 - T0 = 3.3$
- ATU (未接受干预的个体的平均处置效应)$= C1 - C0 = 3$
- ATE (总体平均处置效应)$= \omega \times ATT + (1 - \omega) \times ATU = 3.18$

但在实际情况中, 无法观测到反事实结果。

- “朴素” 估计量 $= T1 - C0 = 2.8$
- ATT 估计误差 $= T0 - C0 = -0.5$
- ATU 估计误差 $= T1 - C1 = -0.2$
- ATE 估计误差 $= \omega \times (T0 - C0) + (1 - \omega) \times (T1 - C1) = -0.38$
- 三组有不同程度的偏差

问题：既然由于反事实的根本问题存在，通常使用"朴素"估计量又会存在估计偏差，那么如何通过观测数据识别处置效应?

回答： 通过实验设计 -- 例如，随机分配

随机实验从理论到实战

- 理解一：潜在结果独立性假设（independence assumption）

$$\{Y_i(1), Y_i(0)\} \perp D_i$$

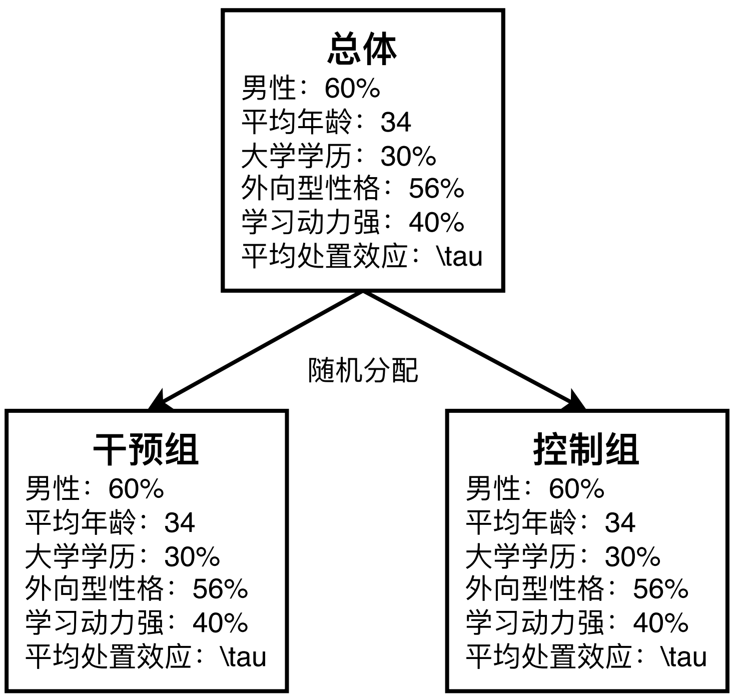
- 理解二：**可观测特征、不可观测特征和处置效应**完全独立于是否接受干预，也就是说那些干扰因素在随机分配后都要被控制
  - 若潜在结果可以表示为可观测特征 X_i 、不可观测特征 e_i 和处置效应 τ_i 的函数

$$\begin{aligned}
&Y_i(0) = a + bX_i + e_i, D_i = 0 \\
&Y_i(1) = a + \tau_i + bX_i + e_i, D_i = 1 \\
&(X_i, e_i, \tau_i) \perp D_i
\end{aligned}$$

  - 通俗理解: 将总体随机分为干预组和控制组, 个体的特征在总体、干预组、控制组均一致

随机实验从理论到实战

研究问题是：班级人数对学生成绩的影响？



- 总体随机抽取各1000人
- 可观测特征：性别、年龄、教育程度
- 不可观测特征：个性、学习动力
- 处置效应：在两组分布没有差异

潜在结果独立假设包含的两个“独立”(1)

- 独立性维度1: 未接受干预(的个体)的潜在结果独立于干预变量

$$\{Y_i(0)\} \perp D_i$$

  - 意味着, 它的均值也和 D_i 不相关

$$E[Y_i(0) \mid D_i = 0] = E[Y_i(0) \mid D_i = 1]$$

  - 化简为: $E\left[Y_i(0) \mid D_i\right] = E\left[Y_i(0)\right]$
  - 该条件就意味着， $T0 = C0$

- 通俗理解: 可以用控制组的观测结果 $C0$ 来衡量不可观测的反事实结果 $T0$, 此时干预组的平均处置效应ATT无偏

$$T1 - C0 = \underbrace{(T1 - T0)}_{\text{ATT}} + \underbrace{(T0 - C0)}_{\text{ATT的偏差}=0} = ATT$$

潜在结果独立假设包含的两个“独立”(2)

- 独立性维度2: 接受干预(的个体)的潜在结果独立于干预变量

$$\{Y_i(1)\} \perp D_i$$

  - 意味着, 它的均值也和 D_i 不相关

$$E[Y_i(1) \mid D_i = 1] = E[Y_i(1) \mid D_i = 0]$$

  - 同理: $E\left[Y_i(1) \mid D_i\right] = E\left[Y_i(1)\right]$
  - 该条件就意味着， $C1 = T1$
- 通俗理解: 可以用干预组的观测结果 $T1$ 来衡量不可观测的反事实结果 $C1$, 此时控制组的平均处置效应ATU无偏

$$T1 - C0 = \underbrace{(C1 - C0)}_{\text{ATU}} + \underbrace{(T1 - C1)}_{\text{ATU的偏差}=0} = ATT$$

从随机实验到回归

由于RCT实验昂贵且以人为实验对象会受伦理审查委员会的保护。那么当不是随机分配时候，能够使用"朴素"估计量呢?

回答： 可以。只要潜在结果的差异是由是否接受干预和可观测的个体特征造成时，就可以通过控制可观测的个体特征来消除选择偏差。

控制可观测特征可以消除选择偏差

- 药物效果实验
  - 服药个体普遍年龄偏大，年龄大的个体普遍的潜在健康状况差
  - 对干预组和控制组的年龄进行分类，控制年龄以消除不同年龄段潜在健康状况的差异。同一个年龄段，干预组和控制组可以看成随机分配，满足前一节的独立性假设

潜在结果		处置情况	观测结果
如果处置	如果未处置		
T1(30) $=\mathbb{E}[Y_i(1) \mid D_i=1, X_i=30]$	T0(30) $=\mathbb{E}[Y_i(0) \mid D_i=1, X_i=30]$	D = 1	T1(30) $=\mathbb{E}[Y_i(1) \mid D_i=1, X_i=30]$
C1(30) $=\mathbb{E}[Y_i(1) \mid D_i=0, X_i=30]$	C0(30) $=\mathbb{E}[Y_i(0) \mid D_i=0, X_i=30]$	D = 0	C0(30) $=\mathbb{E}[Y_i(1) \mid D_i=0, X_i=30]$

- $ATT(30) = ATU(30) = ATE(30) = T1(30) - C0(30)$
- $ATT(40) = ATU(40) = ATE(40) = T1(40) - C0(40)$
- $ATT = P(30|D=1) \times ATT(30) + P(40|D=1) \times ATT(40)$

控制可观测特征在CMI假设下消除选择偏差

对于给定的可观测特征条件 $X_i = x$ 的干预组和控制组

$$ATT(x) = T1(x) - C0(x)$$
$$ATT = \sum_x P(x \mid D = 1) \times ATT(x)$$

- 则有, $ATE = E_x[ATE(X)] = \sum_x P(x) \times ATE(x)$

该假设称为: **条件均值独立假设(CMI)**

$E\left[Y_i(0) \mid D_i = 1, X_i = x\right] = E\left[Y_i(0) \mid D_i = 0, X_i = x\right] = E\left[Y_i(0) = x\right]$
$E\left[Y_i(1) \mid D_i = 1, X_i = x\right] = E\left[Y_i(1) \mid D_i = 0, X_i = x\right] = E\left[Y_i(0) = x\right]$

- 满足CMI最直接的方式是条件随机分配, 如给定30岁群体, 从中随机抽取 一些人服药、一些人不服药
- CMI只能估计该条件下ATE, 更强的假设是**条件独立假设（CIA）**

想要更多：需要条件独立假设 CIA

定义:

- 在 X_i 的条件下,潜在结果 $(Y_i(0), Y_i(1))$ 与干预变量 D_i 独立(选择偏误消失), 数学形式为:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i | X_i$$

$$\begin{aligned} \text{选择偏误} &= E[Y_i(0) \mid X_i, D_i = 1] - E[Y_i(0) \mid X_i, D_i = 0] \\ &= E[Y_i(0) \mid X_i] - E[Y_i(0) \mid X_i] \\ &= 0 \end{aligned}$$

想要更多：需要条件独立假设 CIA

CIA意思是：在控制某些协变量 X_i 后, 干预措施的分配就像 ***随机分配一样***.

将之前的"朴素"估计量写为在控制 X_i 的条件下

$$
\begin{aligned}
& E[Y_i \mid X_i, D_i = 1] - E[Y_i \mid X_i, D_i = 0] \\
&= E[Y_i(1) \mid X_i] - E[Y_i(0) \mid X_i] \\
&= E[Y_i(1) - Y_i(0) \mid X_i]
\end{aligned}
$$

RCT + CIA 才完整!

多次干预呢？CIA扩展-多值干预变量

继续考虑： 教育程度对收入的例子

现在，将CIA拓展到干预变量取多值的情况，比如受教育年数教育 (s_i) 取值为整数 t $\in \{0, 1, \ldots, T\}$。由于受教育水平和收入之间的因果关系可能因人而异，所以我们用个体的收入函数：

$$Y_{si} \equiv f_i(s)$$

$Y_i(1)$ 为个体 i 接受教育(是否干预)后的潜在结果, Y_{si} 代表个体 i 接受 s 年教育后会获得的潜在收入，函数 $f_i(s)$ 告诉我们：每个人任意的受教育水平 s 下个体 i 可能的收入，是依据教育与收入的理论建立的。

换句话说, $f_i(s)$ 回答了“如果……，就会……”这样的一个因果性问题。

模型建构具有一般性，适用于不同理论：在人力资本和收入之间关系的理论模型中， i 教育回报率的函数形式是不同的，可能由个体行为某个特点决定，也可能被市场力量决定，或二者兼而有之。

多次干预呢？CIA扩展-多值干预变量

以上将 CIA 扩展到干预变量的多值情形(multi-value).

CIA表示在给定控制变量集合 X_i的条件下，潜在结果 Y_{si} 和 s_i 是相互独立的，在更一般的条件下，CIA变为：

$$Y_{si} \perp\!\!\!\perp s_i \mid X_i \quad \text{对于所有 } s$$

在RCT中，由于 s_i 是在给定 X_i 下随机分配的，所以CIA自然成立。在使用观察数据进行的研究中，CIA意味着给定 X_i 下 s_i "就像被随机分配的那样好"。

多次干预呢？CIA扩展-多值干预变量

给定 X_i ，多接受一年教育带来的平均处置效应就是 $E[f_i(s)-f_i(s-1)\mid X_i]$ ，多接受四年教育带来的平均处置效应就是 $E[f_i(s)-f_i(s-4)\mid X_i]$ 。

数据只能告诉我们 $Y_i=f_i(s_i)$ ，也就是当 $s=s_i$ 时的 $f_i(s_i)$ 。

在CIA"护身符"下，给定 X_i ，不同教育水平下平均收入的差异就可解释为教育的处置效应。因此多接受1年教育的处置效应可以写为:

$$E[Y_i\mid X_i, s_i=s]-E[Y_i\mid X_i, s_i=s-1]=E[f_i(s)-f_i(s-1)\mid X_i]$$

对任何的 s 都成立。下面证明。

多次干预呢？CIA扩展-多值干预变量

在CIA下，给定 X_i, Y_{si} (潜在结果) 和 s_i (理解为用药的剂量)是独立的:

$$
\begin{aligned}
&E[Y_i \mid X_i, s_i = s] - E[Y_i \mid X_i, s_i = s-1] \\
&= E[f_i(s_i) \mid X_i, s_i = s] - E[f_i(s_i) \mid X_i, s_i = s-1] \\
&= E[f_i(s) \mid X_i, s_i = s] - E[f_i(s-1) \mid X_i, s_i = s-1] \\
&= E[Y_{si} \mid X_i, s_i = s] - E\left[Y_{(s-1)i} \mid X_i, s_i = s-1\right]
\end{aligned}
$$

$$CIA: f_i(s) \perp\!\!\!\perp s_i \mid X_i$$

$$
\begin{aligned}
&= E[Y_{si} \mid X_i] - E\left[Y_{(s-1)i} \mid X_i\right] \\
&= E\left[Y_{si} - Y_{(s-1)i} \mid X_i\right] \\
&= E[f_i(s) - f_i(s-1) \mid X_i]
\end{aligned}
$$

CIA下, 不同教育水平下的平均收入的差异可能解释为教育的处置效果

多次干预呢？CIA扩展-多值干预变量

例子 可以比较教育水平为11年和12年的个体间平均收入的差别，以此来了解高中毕业带来的平均处置效应

$E[Y_i \mid X_i, s_i = 12] - E[Y_i \mid X_i, s_i = 11]$

$= E[f_i(12) \mid X_i, s_i = 12] - E[f_i(11) \mid X_i, s_i = 11]$

$= E[f_i(12) \mid X_i, s_i = 12] - E[f_i(11) \mid X_i, s_i = 12]$ (CIA)

$= E[f_i(12) - f_i(11) \mid X_i, s_i = 12]$

$=$ **给定 X_i 下，已高中毕业学生**因高中毕业带来的平均处置效应

$= E[f_i(12) - f_i(11) \mid X_i]$ (再次CIA)

$=$ **给定 X_i 下**，高中是否毕业（为条件）的平均处置效应

多次干预呢？CIA扩展-多值干预变量（从多条件到无条件）

到目前为止，对 X_i 可取的每一个值都构造了一个处置效果 $ATE_{X_i=x}$ 。这样做的结果是协变量 X_i 取多少值就可能会存在多少处置效果。

对上面的例子而言，如果CIA假设满足，我们可以计算任意条件(组合)下的教育年限为12和11的人的平均收入的差来得到该条件下的处置效应。例如 X_i 包含的变量为（Sex, Age）。那么，Sex=1表示女性，Age的取值范围从20-60。在上面的条件下，一个因果关系可以表示为：

- $E[f_i(12)-f_i(11)|\text{ Sex}=1,\text{ Age}=20\text{至}30]$ 表示年龄段为20~30岁的女性，高中毕业比高中肄业的平均教育回报水平。
- $E[f_i(12)-f_i(11)|\text{ Sex}=0,\text{ Age}=65\text{岁以上}]$ 表示65岁以上的男性，高中毕业比高中肄业的平均教育回报水平。
- 能不能用相对综合的指标概括一系列处置效应?

多次干预呢？CIA扩展-多值干预变量（从多条件到无条件）

Q 那么**无条件**的高中毕业相对于高中肄业的平均处置效应是什么？

A 我们可以利用迭代期望定理对不同的因果效果进行综合。首先, 回忆下刚证明的...

$$E[Y_i \mid X_i, s_i = 12] - E[Y_i \mid X_i, s_i = 11] = E[f_i(12) - f_i(11) \mid X_i]$$

现在取两边的期望值并应用迭代期望法则(LIE)

$$E_X\Big(E[Y_i \mid X_i, s_i = 12] - E[Y_i \mid X_i, s_i = 11] \Big)$$

$$= E_X\Big(E[f_i(12) - f_i(11) \mid X_i] \Big)$$

$$= E[f_i(12) - f_i(11)]$$ (迭代期望)

LPF + CIA → 因果效应

现在我们将LPF与随机实验的研究设计整合在一起。假设我们能够根据理论提炼出，总体的、线性的、**同质因果效应**模型：

$$f_i(s) = \alpha + \tau s + \eta_i \tag{A}$$

- 总体模型是因为 (A) 式告诉我们的是个体 i 在 s 的任意值下能够赚得的收入(这里是潜在收入)， 而不是依据 s_i 观测值，所以省略了 s 的下标 i 。
- 该式假设在 $f_i(s)$ 中唯一因人而异的部分是干扰项 η_i , 其均值为 0 , 用以捕捉决定潜在收入水平 $f_i(s)$ 的其他不可观测因素。将观察到的 s_i 和观察值 Y_i 代入模型，就得到了**样本回归模型**:

$$Y_i = \alpha + \tau s_i + \eta_i \tag{B}$$

- 其中 (A) 式中 τ 是**真实的处置效应**，而 (B) 式中 τ 通常由于 s_i 存在的内生性问题(遗漏变量、测量偏误和反向因果)不是真实的处置效应。在文章的**研究设计**部分要说明如何才能得到真实的处置效应。

LPF + CIA → 因果效应

现在就可以加入多个**可观察**的协变量 X_i， 它们CIA成立，意图排除**干扰因素**。我们将潜在收入水平 $f_i(s)$ 的随机项表达为可观察变量 X_i (因人而异)和**残差项** v_i 的线性函数：

$$\eta_i = X_i'\beta + \nu_i \tag{C}$$

其中 β 是 η_i 对 X_i 回归的总体系数向量(意味着上式假设是可以通过最小二乘估计获得正确的系数估计)，所以有：

1. $E[\eta_i \mid X_i] = X_i'\beta$
2. 残差项 v_i 与 X_i 不相关

LPF + CIA → 因果效应

进一步，由CIA我们可以：

$E[f_i(s) \mid X_i, s_i]$

$= E[f_i(s) \mid X_i]$ (根据CIA)

$= E[\alpha + \tau s_i + \eta_i \mid X_i]$ (代入B式)

$= \alpha + \tau s_i + E[\eta_i \mid X_i]$

$= \alpha + \tau s_i + X_i'\beta$ (最小二乘回归方程)

回忆 这里再次使用到，若 $f_i(s)$ 的CEF是线性的(如倒数第2行)，则意味着"正确[†]" LPF就是CEF。

† 这里"正确"是指若控制了 X_i 就像RCT一样。

LPF + CIA → 因果效应

所以我可以把模型设置如下：

$$Y_i = \alpha + \tau s_i + X_i'\beta + \nu_i$$

通过限制扰动项 ν_i 的性质:

1. s_i (根据 CIA)
2. X_i (根据定义 β 是 η 对 X_i 回归的总体系数向量)

   就是 来得到我们最感兴趣的因果效应 τ

LPF^{ols} 与 "朴素"估计量的关系

回忆："朴素"估计量是干预组与控制组的观测结果均值之差 $E[Y_i|D_i=1]-E[Y_i|D_i=0]$

当干预变量为二值时，可以证明回归系数 $\hat{\tau}_{OLS}$ 等于处理组与控制组样本均值之差(by Mixtape)。在样本视角下：

$$\hat{\tau}_{OLS}=\frac{1}{N_T}\sum_{i=1}^{n}(y_i \mid d_i=1)-\frac{1}{N_C}\sum_{i=1}^{n}(y_i \mid d_i=0)=\bar{Y}_T-\bar{Y}_C$$

在大样本下:

$$\hat{\tau}_{OLS}=\bar{Y}_T-\bar{Y}_C \xrightarrow{p} E\left[Y_i \mid D_i=1\right]-E\left[Y_i \mid D_i=0\right]=\tau_{OLS}$$

综上, $\hat{\tau}_{OLS}=\bar{Y}_T-\bar{Y}_C \xrightarrow{p} \tau_{OLS}=$ “朴素” 估计量

"朴素"估计量 = ATE + 选择偏误 +异质性干预偏误 (by Mixtape) , 基于SUTVA第三项为零

LPF^{ols}+控制变量+假设 → 因果效应

现在我们已经知道在： $E\left[Y_i(0) \mid D_i=1\right] \neq E\left[Y_i(0) \mid D_i=0\right]$ 时, 无法识别处置效应。

假如造成差异的原因: 个体未干预时的潜在结果 $Y_i(0)$ 是可观测特征和不可观测特征的线性函数

$$Y_i(0)=\alpha+\beta X_i+e_i$$

代入方程: $Y_i=\underbrace{E\left[Y_i(0)\right]}_{a}+\underbrace{\left[Y_i(1)-Y_i(0)\right] \times D_i}_{\tau}+\underbrace{Y_i(0)-E\left[Y_i(0)\right]}_{u_i}$

得: $Y_i=\alpha+\tau D_i+\beta X_i+e_i$ （观测结果、干预状态、可观测特征、不可观测特征的关系）

将 Y_i 对 D_i、X_i 归回: $E\left(Y_i \mid D_i, X_i\right)=\alpha+\ D_i+\beta X_i+E\left[e_i \mid D_i, X_i\right]$

LPF^{ols}+控制变量+假设 → 因果效应

- 与CIA思路一样，若要使得条件期望函数的 D_i 的系数等于 τ ,需要 以观测结果、干预状态、可观测特征为基础的LPF的干扰项 e_i 的条件均值独立于干预变量：

  $E[e_i \mid D_i, X_i] = E[e_i \mid X_i]$

- 可证明：**(建立在LPF^{ols}基础上的)干扰项条件均值独立于干预变量**和 **平均未干预潜在结果条件独立**（ $E\left[Y_i(0) \mid D_i = 1\right] = E\left[Y_i(0) \mid D_i = 0\right]$ ）是等价的

- 这个条件使得LPF^{ols}可以通过加入控制变量X 来达到估计处置变量D 的真实因果效应系数 τ 的目的

- **CMI** 与 **CIA** 是直接建立在 **潜在结果**上的；**干扰项条件均值独立于干预变量** 和 **平均潜在结果条件独立** 是建立在**平均潜在结果**基础上（是在CEF-LPF 框架下能够识别处置效应的关键条件）

条件期望的性质（自学）

条件期望值函数的性质

- **性质1** (期望迭代法则,law of iterated expectation)

$$E[E[Y \mid X]] = E[Y]$$

$E[Y|X]$ 的期望值是 $[Y]$ 的无条件期望值。

例如：

$$\begin{aligned}
&\mathbb{E}\left[\log(wage) \mid gender = man\right]\mathbb{P}\left[gender = man\right] \\
&+\mathbb{E}\left[\log(wage) \mid gender = woman\right]\mathbb{P}\left[gender = woman\right] \\
&=\mathbb{E}\left[\log(wage)\right].
\end{aligned}$$

Or numerically,

$$3.05 \times 0.57 + 2.81 \times 0.43 = 2.95.$$

- **性质1**推论

$$E\left[E[Y|X_1,X_2]|X_1\right]=E[Y|X_1]$$

  - 内部期望值以X1和X2同时为条件,外部期望值只以X1为条件。迭代后的期望值可以得到简单的答案E[Y|X1],即只以X1为条件的期望值。《E》表述为"较小的信息集获胜" → 以小谋大

    例：

$$\begin{aligned}&\mathbb{E}\left[\log(wage)\mid gender=man,\ race=white\right]\mathbb{P}\left[race=white\mid gender=man\right]\\&+\mathbb{E}\left[\log(wage)\mid gender=man,\ race=Black\right]\mathbb{P}\left[race=Black\mid gender=man\right]\\&+\mathbb{E}\left[\log(wage)\mid gender=man,\ race=other\right]\mathbb{P}\left[race=other\mid gender=man\right]\\&=\mathbb{E}\left[\log(wage)\mid gender=man\right]\end{aligned}$$

    or numerically

$$3.07\times 0.84+2.86\times 0.08+3.03\times 0.08=3.05.$$

- **性质2** (线性)

$$E[a(X)Y + b(X)|X] = a(X)E[Y|X] + b(X)$$

对于函数 $a(\cdot)$ and $b(\cdot)$.

- **性质3**（独立意味着均值独立）

若 X 与 Y 独立, 则 $E[Y|X] = E[Y]$

- **性质3**的证明 (以离散变量为例):

$$
\begin{aligned}
E[Y|X] &= \sum_{i=1}^{N} y_i P(Y = y_i|X) \\
&= \sum_{i=1}^{N} y_i \frac{P(Y = y_i, X)}{P(X)} \\
&= \sum_{i=1}^{N} y_i \frac{P(Y = y_i) \times P(X)}{P(X)} \quad = E[Y].
\end{aligned}
$$

用到了 $P(Y = y, X = x) = P(X = x)P(Y = y)$.

- **性质4** （均值独立意味着不相干）
  若 $E[Y|X] = E[Y]$, 则 $Cov(X, Y) = 0$.

  - $E[Y|X] = E[Y]$ is 均值独立(**mean independence**)
  - 记住: 均值独立意味着不相干，反过来不一定成立.

- **性质5** （条件期望值是最小均值平方误差）
  假设对于任意函数 g 有 $E[Y^2] < \infty$ 并 $E[g(X)] < \infty$, 那么

$$E[(Y - \mu(X))^2] \leq E[(Y - g(X))^2]$$

  其中 $\mu(X) = E[Y|X]$
  解读:

  - 假设使用某种函数形式 g 和数据 X 来解释 Y
  - 那么 g 的最小均方误（ **the mean squared error** ）就是条件期望。

- **性质5**的证明:

$$
\begin{aligned}
E[(Y-g(X))^2] &= E[\{(Y-\mu(X))+(\mu(X)-g(X))\}^2] \\
&= E\left[(Y-\mu(X))^2\right] + E\left[(\mu(X)-g(X))^2\right] \\
&+ 2E\left[(Y-\mu(X))\left(\mu(X)-g(X)\right)\right].
\end{aligned}
$$

使用期望迭代法则

$$
\begin{aligned}
E\left[(Y-\mu(X))\left(\mu(X)-g(X)\right)\right] &= E\left\{E\left[(Y-\mu(X))\left(\mu(X)-g(X)\right)|X\right]\right\} \\
&= E\left\{(\mu(X)-g(X))\left(E[Y|X]-\mu(X)\right)\right\} \\
&= 0
\end{aligned}
$$

所以,

$$
E[(Y-g(X))^2] = E\left[(Y-\mu(X))^2\right] + E\left[(\mu(X)-g(X))^2\right]
$$

上式取最小值，当且仅当 $g(X)=\mu(X)$.

- **概率迭代法则**

$$P(Y) = \sum_{i=1}^{N} P(Y|x_i)P(x_i)$$

X是离散随机变量

- **方差加法法则**

$$Var(Y) = E[V(Y|X)] + V[E(Y|X)]$$

SUTVA

在之前的例子中，都是假设个体处置效应是相同的，即 $\tau_i = \tau$。这里正式提出 **SUTVA** 假设。

- **稳定个体干预值假设（The Stable Unit Treatment Value Assumption， SUTVA）**：简单说，每个个体的潜在结果不依赖于其他个体的干预状态。有两层含义：1. 不同个体的潜在结果之间不会有交互影响。2. 干预水平对所有个体都是相同的。
- 第1个含义：它排除了**外部性**或**均衡效应**。
    - 例：研究班级规模对个体学习效果的影响，同学之间往往存在外部性，如果班级里好学生多，相互讨论、相互促进，产生正外部性，从而提高了整体学习效率。
    - 例：如果劳动力培训项目规模很大，改变整改市场技能结构，使得技能劳动力供给很多，则接受培训的个体干预效果就不显著。

- 第2层含义：处置效应对所有个体相同。
    - 例如：教育对个人收入影响。要求纳入的教育程度要求教育质量相同.
- 社会科学，实际中对第1项更为关注.

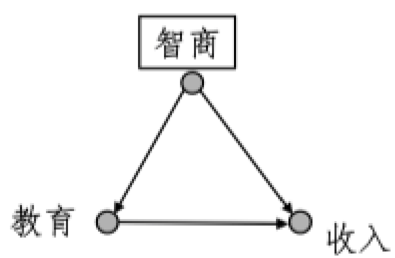
偏误类型与解决办法

有向无环图表达偏误类型

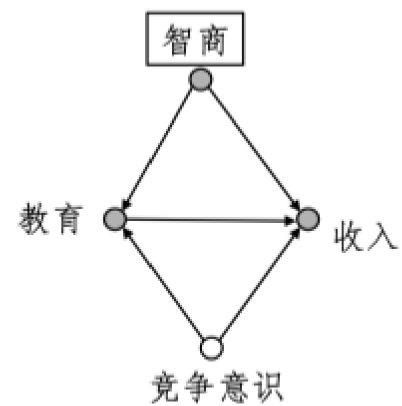
- 因果路径
- 混淆路径: A ← B → C , B是A和C的混淆变量; 混淆变量会导致相关关系
- 对撞路径: A → B ← C , B是A和C的对撞变量; 对撞变量不会产生相关性
- 估计X与Y的因果关系的本质是找到二者间所有的因果路径，同时去 除二者间的非因果关系路径。

混淆偏误（好的控制）

- 混淆偏误是指在X和Y之间存在未截断的混淆路径，造成X和Y的相关性不仅包含因果关系，还包含非因果关系。
- 截断混淆路径是通过给定混淆变量（conditional on confounding variable）为条件，从而排除混淆变量的干扰。给定混淆变量可以简单的理解为固定混淆变量的值。在关系图中，我们加个方框表示这个变量是给定的。
- 当混淆变量给定时，X和Y的相关性就与混淆变量无关，二者相关性就是因果关系。



图：截断混淆路径



图：存在未截断的混淆路径

过度控制偏误

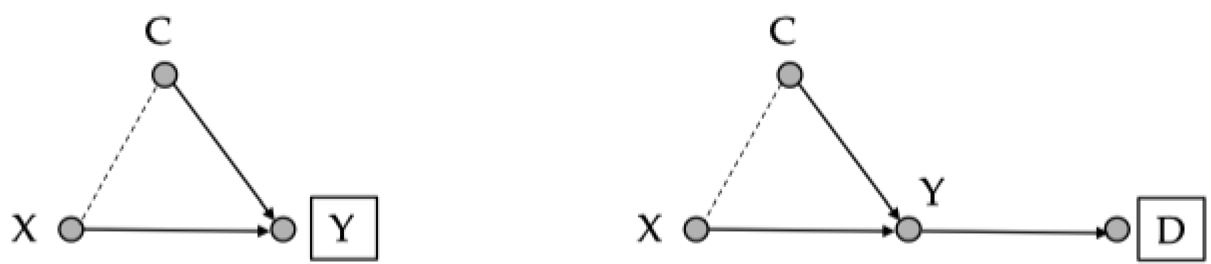
- 过度控制偏误是指控制了因果路径上的变量造成的偏误
- 在研究中我们要避免控制受X影响并会影响Y的中介变量，否则会造成过度控制偏差。

图：过度控制偏差

- 控制了生活规律，就截断了一条因果路径，估计得到的只是锻炼对健康的直接因果关系，相当于低估真实值

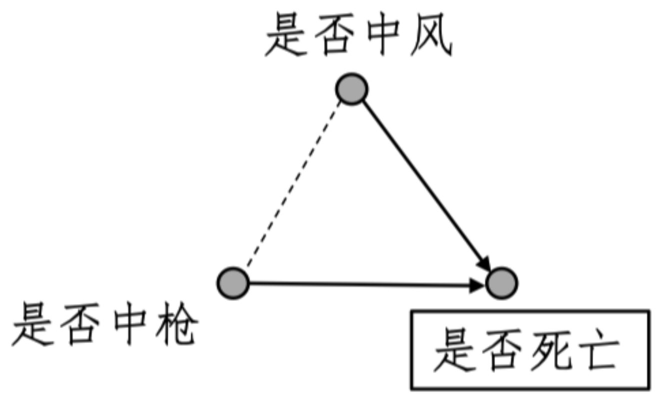
对撞偏误

- 对撞偏误可以理解为当给定两个变量的共同结果（对撞变量） 时（或者对撞变量的延伸因变量），两个变量间会产生一个衍生路径。 衍生路径会造成两个原本不相关的变量变为相关，或造成两个原本相关 的变量的相关性发生改变。



图：对撞偏误

对撞偏误



是否死亡	是否中风	是否中枪
否	否	否
是	否	是

解决办法

- 文献寻找偏误潜在来源 $\Longleftrightarrow$ 建立有向无环图 $\Longleftrightarrow$ 寻找对应解决办法 $\Longleftrightarrow$ 深挖**数据生成过程**

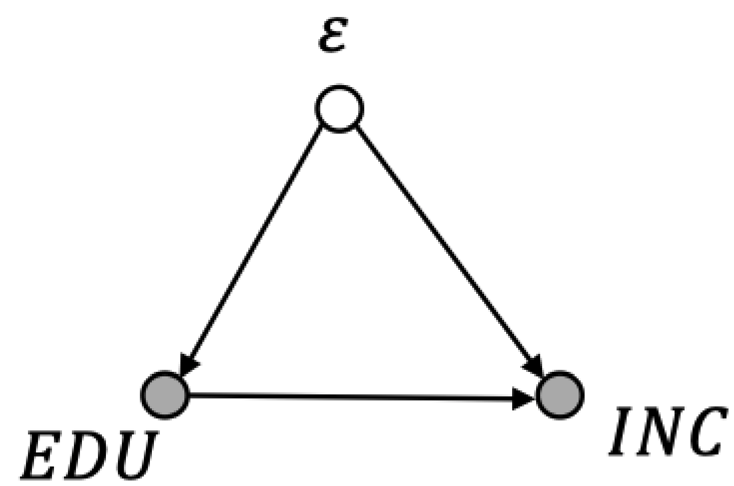
- **回忆：** 教育程度对收入

- 假设最初只是通过建立如下LPF^{0}来估计处置效应

$$INC_{it} = \alpha + \beta_1 EDU_{it} + \varepsilon_{it}$$

- 当可观测变量（性别、年龄）和不可观测变量，同时进入扰动项 ε_{it}，导致 ε_{it} , 所以无法识别因果影响系数 β 。

$$\varepsilon_{it} = \beta_2 AGE_{it} + \beta_3 GENDER_i + e_{it}$$



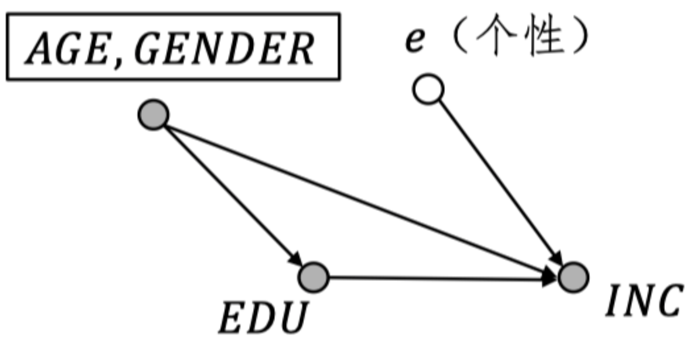
图：LPF等价于在DAC中忽略了年龄和性别

例子：解决办法

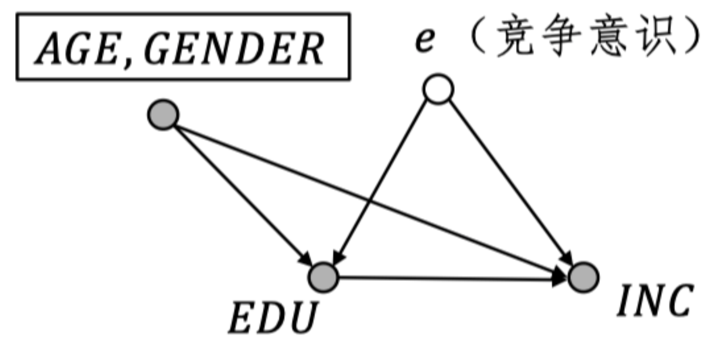
- 可以将 ε_{it} 中的可观测变量分离进行控制，得到 LPF[1]

  $INC_{it} = \alpha + \beta_1 EDU_{it} + \beta_2 AGE_{it} + \beta_3 GENDER_i + e_{it}$ ，其中 e_{it} 为不可观测变量，如（个性、竞争意识）．

- AGE_{it} 和 EDU_{it} 为时变变量； $GENDER_i$ 为非时变变量（虚拟变量、类别变量）



图：控制年龄和性别且e为无关变量的情况



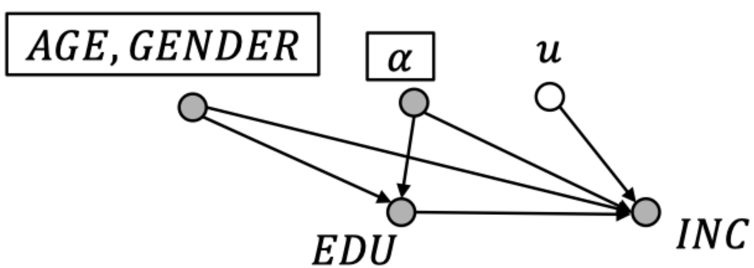
图：控制年龄和性别且e为混淆变量的情况

例子：解决办法

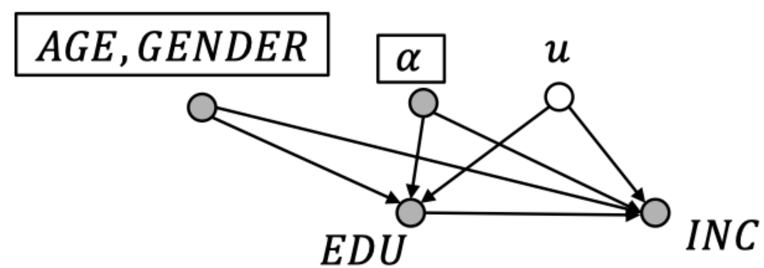
- 倘若真实关系是右图，仍无法识别因果关系
- 将 ε_{it} 进一步分解为：不可观测的非时变变量 α_i 和不可观测的时变变量 u_{it} ，即 $e_{it} = \alpha_i + u_{it}$

$$INC_{it} = \alpha + \beta_1 EDU_{it} + \beta_2 AGE_{it} + \beta_3 GENDER_i + \alpha_i + u_{it}$$

- 如果混淆路径是 α_i 造成的，我们希望控制 α_i 截断混淆路径。即采用面板数据分析法可以达到控制不可观测的非时变变量。



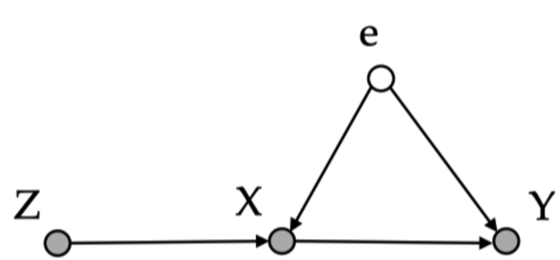
图：控制年龄、性别和α且u为无关变量的情况



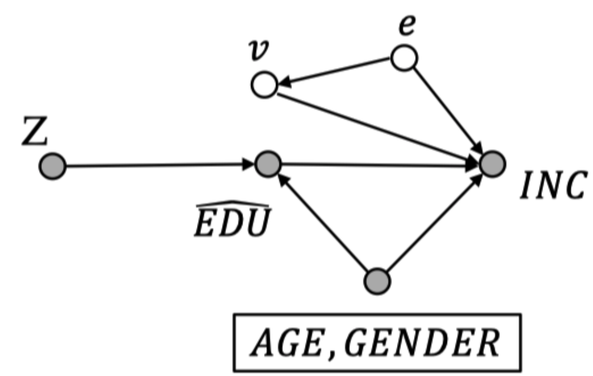
图：控制年龄、性别和α且u为混淆变量的情况

例子：解决办法

- 实际上，大部分社会研究不太能轻易地在 ε_{it} 中分离出不可观测和可观测变量
- 引入入具变量 Z_i 分解出 EDU_{it} 变化中与 e_{it} 无关的部分，即 $EDU_{it} = \widehat{EDU}_{it} + v_{it}$，其中 $\widehat{EDU}_{it}$ 是 EDU_{it} 与 e_{it} 无关的部分。通过工具变量分解出 自变量中不被 e_{it} 混淆的信息来估计解释和因变量的因果关系。
    - 工具变量要符合两个条件：外生性和相关性
    - 这意味着 Z 对 Y 的作用 $=$ Z 对 X 的作用 $\times$ X 对 Y 的作用



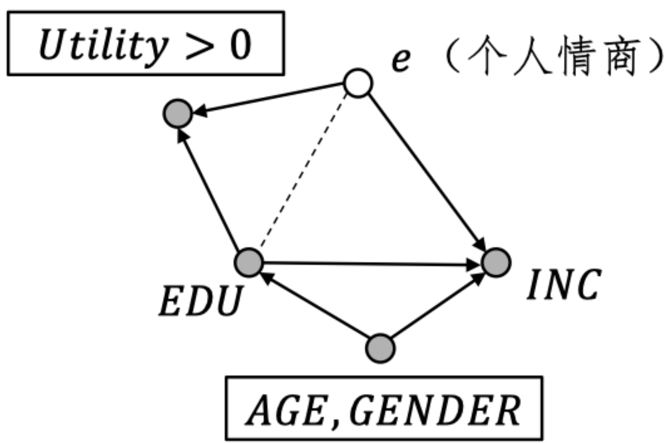
图：工具变量的相关性和外生性



图：引入工具变量的情况

解决办法

- 倘若样本从总体随机抽取的，会导致样本里自变量和不可观测因素 e 存在相关性。下图刻画该情形。
- 由于样本中只包括了参加工作的个体，是否参加工作则有效用变量 Utility 表示。



图：对撞偏误

解决办法小结

方法	解决的因果关系中的偏差
简单回归、匹配法	可观测因素造成的混淆偏差
面板数据分析法	可观测因素+不随时间变化的不可观测因素造成的混淆偏差
工具变量法、双重差分法、断点回归法	可观测因素+不可观测因素造成的混淆偏差
样本自选择模型	包含不可观测因素造成的对撞偏差