体育经济分析：理论与应用

专题一：回归与因果2

周正卿

09 March 2022

引言

今天 + 0.5

- CEF $\longrightarrow$ LPF $\longrightarrow$ LPF^{ols} $\longrightarrow$ 因果推断
  - CEF、LPF和 LPF^{ols} 的系数解读
- Rubin 潜在结果框架
- 随机分配
- 在CEF-LPF框架下看 RCT+CIA
- 著名的STAR实例

- 基于实例的延伸
  - SUTVA
  - 因果关系估计偏差的类型与解决办法
  - 关于控制变量
  - 遗漏变量偏差
  - "坏"的控制
  - 内生性问题

CEF、LPF和 LPF^{ols} 的系数解读

CEF、LPF和 LPF^{ols} 的系数解读

- 虚拟变量：Black={1, black; 0, others}； female={1, female; 0, male}，{black, female}={1,0; 1,1; 0,1; 0,0}
- 线性CEF的参数

$$E[\log(wage) \mid Black, female] = -0.20Black - 0.24female + 0.10Black \times female + 3.06$$

线性CEF：黑人男性（相对于非黑人的男性）低20%，黑人女性（相对于非黑人女性）为(-20-24+10)=-34%。

- LPF的参数

$$\mathscr{P}[\log(wage) \mid Black, female] = -0.15Black - 0.23female + 3.06$$

LPF：黑人打工者（相对于非黑人)低15%，忽略了性别的作用(计算净影响)，控制变量保证条件相同。

CEF、LPF和 LPF^{ols} 的系数解读

Summary of functional forms with logarithms and interpretations

Model	Regressand	Regressor	Formulae	Interpretation
level - level	y	x	$\Delta y = b_1 \Delta x$	A percentage point change in x leads to a unit change in y
level - log	y	$\log(x)$	$\Delta y = (b_1/100)\%\Delta x$	A percentage point change in x leads to a unit change in y
log - level	$\log(y)$	x	$\%\Delta y = (100 b_1)\Delta x$	A percentage point change in x leads to a percentage change in y *(semi-elasticity model)*
log - log	$\log(y)$	$\log(x)$	$\%\Delta y = b_1 \%\Delta x$	A percentage change in x leads to a percentage change in y *(constant elsticity model)*

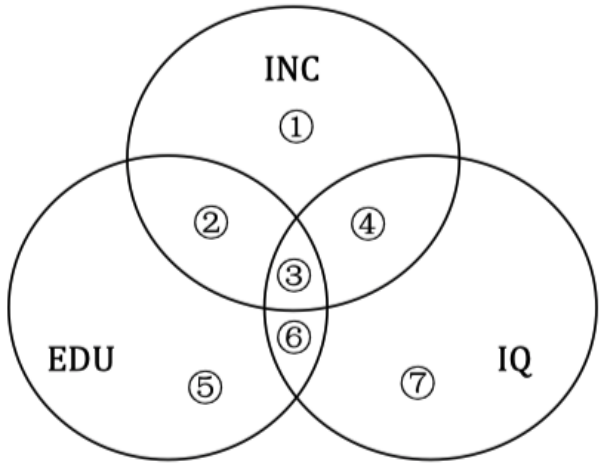
韦恩图理解LPF^{ols}

- 两两相交的部分指的是两个变量共同变化的部分，表明两个变量存在一定的线性相关关系

$$INC = \alpha + \beta_1 EDU + \beta_2 IQ + e$$

- $$E(e \mid EDU, IQ) = 0$$
$$Cov(EDU, IQ) \neq 0$$

- β_1 反映的只有(2)的信息, β_2 反映的只有(4)的信息, 而信息(3)（反映了共同的影响）被同时舍去
- 满足 $E(e \mid EDU, IQ) = 0$, 回归系数反映了EDU和 IQ 各自对INC的影响, 即因果关系



系数估计值: 两步求解

基于OLS得到任何 X_i (EDU)系数 $\hat{\beta}_i$ 都可分为两步求解

- 第一步: 将 X_i 作为因变量, 其他不包含 X_i 的解释变量
  作为自变量,OLS得到残差项 $\widetilde{X}_i = \hat{v}_i$,即(2)+(5):

  $$X_i = \hat{\gamma}_0 + \hat{\gamma}_1 X_1 + \cdots + \hat{\gamma}_{i-1} X_{i-1} + \hat{\gamma}_{i+1} X_{i+1} + \cdots + \hat{\gamma}_k X_k + \hat{v}_i$$

- 第二步: Y作因变量, $\widetilde{X}_i$ 作自变量, OLS得到 $\hat{\beta}_i$,即(2):

  $$Y = \hat{\alpha} + \hat{\beta}_i \widetilde{X}_i + \hat{\varepsilon}$$

  其中 $\hat{\beta}_i = \frac{Cov\left(Y, \widetilde{X}_i\right)}{Var(\widetilde{X}_i)}$

Rubin 潜在结果框架

关于潜在结果因果推断的版权

目前可以看到的文献，最早的是 Neyman 于 1923 年用波兰语写的博士论文，第一个在**试验设计**中提出了“潜在结果”（potential outcome）的概念。后来 Rubin 在**观察性研究**中重新（独立地）提出了这个概念，并进行了广泛的研究。 Rubin 早期的文章并没有引用 Neyman 的文章， Neyman 的文章也不为人所知。一直到 1990 年， Dabrowska 和 Speed 将 Neyman 的文章翻译成英文发表在 Statistical Science 上，大家才知道 Neyman 早期的重要贡献。今天的文献中，有人称 Neyman-Rubin Model，其实就是潜在结果模型。计量经济学家，如 Heckman 称，经济学中的 Roy Model 是潜在结果模型的更早提出者。在 Rubin 2004 年的 Fisher Lecture 中，他非常不满地批评计量经济学家，因为 Roy 最早的论文中，全文没有一个数学符号，确实没有明确的提出这个模型。详情请见， Rubin 的 Fisher Lecture，发表在 2005 年的 Journal of the American Statistical Association 上。研究 Causal Diagram 的学者，大多比较认可 Rubin 的贡献。但是 Rubin 却是 Causal Diagram 的坚定反对者，他认为 Causal Diagram 具有误导性，且没有他的模型清楚。他与 Heckman （诺贝尔经济学奖）， Pearl （图灵奖） 和 Robins 之间的激烈争论，成为了广为流传的趣闻。 -- by 丁鹏

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

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**

关心参数的命名

- **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和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

观测结果

观测结果:

- 个体根据它的接受干预状态而显现出来的对应的潜在结果
- 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, $Y_i = Y_i(1)$

“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: $(Y_i(0) \mid D_i = 0)$.

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

然而,上述答案不一定正确。

“朴素”估计量可能存在的三种偏差

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

“朴素”估计量可能存在的偏差

- 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. **接受干预与否**并非随机，即:是否接受干预与**潜在结果**是相关的
  2. 这三种产生偏差的原因都源于**接受干预与否**是个体自我选择的后果，称之为**选择偏差**

计算ATE实例: 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	2	3	1	5
2	7	3	4	1	7
3	4	1	3	1	4
4	3	2	1	0	2
5	8	3	5	0	3

- 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

计算ATE实例

- 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		处置情况	平均观测结果
如果处置	如果未处置		
$T1 = E[Y_i(1) \mid D_i = 1] = 5.3$	$T0 = E[Y_i(0) \mid D_i = 1] = 2$ (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)	$C0 = E[Y_i \mid D_i = 1] = E[Y_i(0) \mid D_i = 0] = 2.5$

计算ATE实例

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 -- 随机分配

随机分配

随机分配的两种理解

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

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

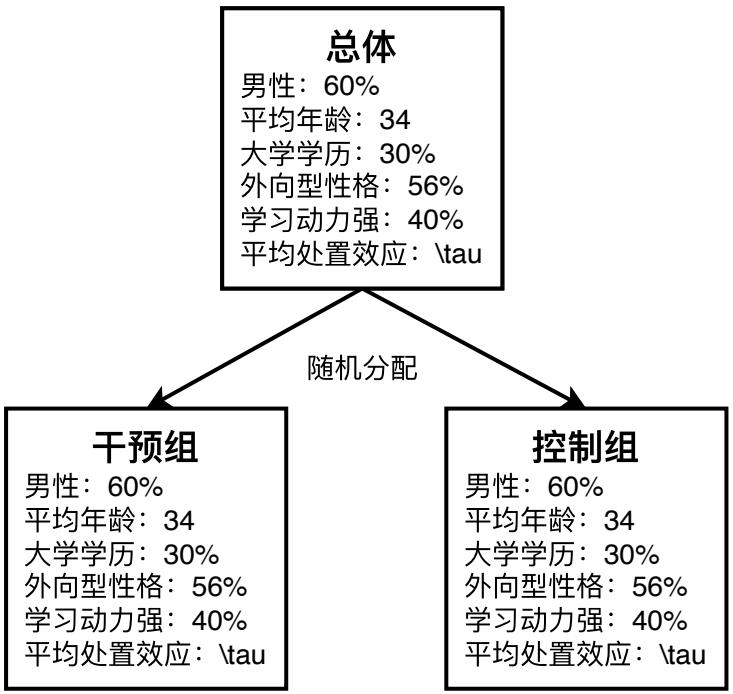
- 理解二：可观测特征、不可观测特征和处置效应完全独立于是否接受干预
  - 若潜在结果可以表示为可观测特征 X_i 、不可观测特征 e_i 和处置效应 τ_i 的函数

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

  - 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

随机分配的两种理解

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)

- 第一个独立性: 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)

- 第二个独立性: 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

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

随机分配小结

随机分配的假设下 → 潜在结果独立假设(强) → 2个均值独立假设(弱)

- $\{Y_i(0)\} \perp D_i \Rightarrow E[Y_i(0) \mid D_i = 0] = E[Y_i(0) \mid D_i = 1] \Longleftrightarrow C0 = T0$
- $\{Y_i(1)\} \perp D_i \Rightarrow E[Y_i(1) \mid D_i = 1] = E[Y_i(1) \mid D_i = 0] \Longleftrightarrow T1 = C1$

$\longrightarrow$

$$ATT \equiv T1 - T0 \ = ATU \equiv C1 - C0 \ = ATE \equiv \omega ATT + (1-\omega)ATU$$

- ATT、ATU和ATE都没有偏差
- 随机分配可以使得干预组和控制组的处置效应没有区别，均可以用观测到的T1 和 C0 去估计。
- 随机分配虽然潜在结果与是否干预独立，但是观测结果与是否干预相关的。因为 $E[Y_i(0) \mid D_i = 0] \neq E[Y_i(1) \mid D_i = 1]$
- 随机分配看起来完美解决了因果推断的根本难点， 但是现实中社会科学无法保证完全随机和对照处理，随机分配的思想指 导我们进一步研究。

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?

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 → 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 → 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)**

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

- 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）**

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:

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

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

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!

在之后介绍的STAR项目中,虽然是按照学校进行的控制组与处理组(across-school)分配，但在学校内(within-school)部是随机化分配学生的。

条件独立假设的扩展：多值干预变量

继续考虑： 教育程度对收入的例子

现在，我们将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。

条件独立假设的扩展：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”。

条件独立假设的扩展：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。

条件独立假设的扩展：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:

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

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

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

条件独立假设的扩展：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

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

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

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

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

$=$ **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

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

$=$ **给定** X_i **下**，高中是否毕业（为条件）的平均处置效应

条件独立假设的扩展：从条件到无条件

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?

条件独立假设的扩展：从条件到无条件

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)

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

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

$$= E[f_i(12) - f_i(11)] \quad \text{(迭代期望)}$$

LPF + CIA → 处置效应

现在设定具体函数形式，

假设总体的、线性的、处置效应为常数的模型：

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

之所以是总体模型，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 最终实现值 , 所以这里省略了 s 的下标 i 。该式还同时假设在 $f_i(s)$ 中唯一因人而异的部分是误差项 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 代入模型有：

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

其中 (A) 式中 ρ 是准确的处置效应, 而 (B) 式中 ρ 因为 s_i 的内生性原因(遗漏变量或选择偏差)不是真实的处置效应.

LPF + CIA → 处置效应

现在考虑给定一系列可观察的协变量 X_i, 且CIA成立。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 (因人而异的特点)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 回归的总体参数向量(意味着上式假设是可以通过最小二乘估计获得正确的参数估计)，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 → 处置效应

更进一步，由条件独立假设，我们有：

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

$$= E[f_i(s) \mid X_i] \quad \text{(根据CIA)}$$

$$= E[\alpha + \rho s_i + \eta_i \mid X_i] \quad \text{(代入B式)}$$

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

$$= \alpha + \rho s_i + X_i'\beta \quad \text{(最小二乘回归方程)}$$

回忆 *这里再次看到, 倘若 $f_i(s_i)$ 的条件期望函数CEF是线性的(倒数第二式), 意味着"正确†" 线性投影函数LPF就是CEF*

† 这里, "正确" 就是以正确的 X_i 为条件, 可以用CIA是实现RCT效果的意思。

LPF + CIA → 处置效应

所以线性因果(回归)方程为:

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

残差项 ν_i 与以下不相干:

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

系数 ρ 就是我们感兴趣的 s_i 对 Y_i 处置效应

这里需要再次强调的是我们做出的关键假设是: **可观察的特点 X_i 是导致 η_i 和 s_i 相关的唯一原因（想想C式代入B式）。** 这就是Barnow, Cain和Goldberger (1981) 针对回归讨论过的选择性偏误全部来自可观察变量的假设 (selection-on-observable assumption)。它已成为经济学中绝大多数经验研究的基础，换句话说为什么在实践中我们会在回归式中加入控制变量(协变量)，而且寻求"好"的控制变量的原因。 所有加入控制变量后的LPF的潜台词是: **基于我所控制了一些条件，获得的处置效应就如同实施了随机实验那样好。**

CIA 小结

综上

1. CIA赋予参数**因果解释** (消除了选择偏误).
2. 允许干预变量多值 —— 在没有使用LIE预期下得到是 **条件平均处置效应**.
3. CIA的挑战是: 你必须像"透视图"一样知道一系列的控制变量 (X_i) 才能使得干预分配像 **RCT一样**.

在CEF-LPF框架下理解 RCT+CIA

CEF-LPF框架下看处置效应

在观测研究中，我们仅能观测到一种状态下的潜在结果. $Y_i(0)$ 列和 $Y_i(1)$ 列实际上是观测不到的,只能看到 Y_i 和 D_i 两列，已知:

$$Y_i = Y_i(0) + [Y_i(1) + Y_i(0)] \times D_i$$

改写为:

$$\begin{aligned} Y_i &= \underbrace{E\left[Y_i(0)\right]}_{a} + \underbrace{[Y_i(1) - Y_i(0)] \times D_i}_{\tau} + \underbrace{Y_i(0) - E\left[Y_i(0)\right]}_{u_i} \\ &= a + \tau \times D_i + u_i \end{aligned}$$

- 系数: a : 所有个体的未干预潜在结果均值 $E[Y_i(0)]$
- τ : 处置效应 (假设所有个体处置效应相同)
- u_i : 第 i 个 个体未干预时的潜在结果和所有个体未处置时的平均潜在结果之差 ，且 $E[u_i] = E[Y_i(0) - E[Y_i(0)]] = E[Y_i(0)] - E[Y_i(0)] = 0$

CEF-LPF框架下看处置效应

将 Y_i 以 D_i 为条件，计算CEF：

$$E\left(Y_i \mid D_i\right)=a+\tau \times D_i+E\left(u_i \mid D_i\right)$$

- 该CEF是线性的，因此与其对应的LPF具有相同形式
- 经验研究，一般是从LPF出发，要获得处置效应，取决于 $E(u_i \mid D_i)$ 与 干预变量 D_i 的关系。

- 若果两者具有线性关系: $E(u_i \mid D_i)=\phi_0+\phi_1 D_i$
  - 代入可得 $E\left(Y_i \mid D_i\right)=a+\phi_0+\left(\tau+\phi_1\right) \times D_i$

- $u_i=Y_i(0)-E\left[Y_i(0)\right]$.若以服药为例, u_i 是个体 i 没有吃药的健康状况与所有个体没有吃药时平均健康的差异。如果 $u_i<0$ (表示所有吃药个体 i 的平均健康水平普遍较差），则意味着 $\phi_0<0, \phi_1<0$ （手动证明），通过LPF^{ols}得到的估计系数 $\tau+\phi_i$ 会低于处置效应真实值 τ

CEF-LPF框架下看处置效应

只有 $\phi_1 = 0 \iff E(u_i \mid D_i) = \phi_0 \iff u_i$ **均值独立于** D_i

$\Longrightarrow$ 建立在观测值 Y_i 和 D_i 基础上的 LPF^{ols} 是正确的估计

$\iff$ LPF^{ols}对应的LPF满足了 $E(e \mid X) = 0$ 的条件

$\Longrightarrow$ LPF^{ols} 满足了 CEF的条件

$\Longrightarrow$ 获得了因果性CEF所对应的LPF的正确处置效应估计

CEF-LPF框架下看处置效应

只有 $\phi_1 = 0 \iff E(u_i \mid D_i) = \phi_0 \iff u_i$ **均值独立于** D_i
$\Longrightarrow$ 建立在观测值 Y_i 和 D_i 基础上的 LPF^{ols} 是正确的估计
$\iff$ LPF^{ols}对应的LPF满足了 $E(e \mid X) = 0$ 的条件
$\Longrightarrow$ LPF^{ols} 满足了 CEF的条件
$\Longrightarrow$ 获得了因果性CEF所对应的LPF的正确处置效应估计
$\iff$ 本质是，**潜在结果均值独立于干预变量**

证明: $E(u_i \mid D_i) = \phi_0 \longrightarrow E\left(u_i \mid D_i = 1\right) = E\left(u_i \mid D_i = 0\right) = \phi_0$

将 $u_i = Y_i(0) - E\left[Y_i(0)\right]$ 代入公式可得：

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

$E\left[Y_i(0)\right]$ 是一个固定值, 与 D_i 无关, 因此上式等价于:

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

LPF^{ols} 与 "朴素"估计量

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 → 处置效应

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 → 处置效应

- 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）

控制变量正式定义

在CEF-LPF 框架下：

- 能够让干扰项 e_i 条件均值独立于干预变量 D_i 的变量，即给定了控制变量后，干扰项与干预变量不再相关
- 给定控制变量 X_i ，当 D_i 发生变化时， e_i 的均值不发生变化，因变量 Y_i 的均值变化就可以完全归因于 D_i 的变化，从而识别出处置效应

- 干扰项条件均值独立于干预变量，只是保证 D_i 的系数是 τ 的无偏估计 ，不能保证控制变量 X_i 的系数 β 的无偏估计
- 进一步将解释变量区分为：干预变量和控制变量

控制了X

$$Y_i = \alpha + \gamma D_i + \beta X_i + e_i$$

D_i 和 e_i 不相关

RCT实例：田纳西州班级师生比对学业成绩影响

实验的理论基础

- 为了振兴经济，提高教育质量，应对入学人数减少，在选民的强烈要求下，20世纪八十年代美国田纳西州进行了迄今最具权威性的小班化教育改革实验STAR计划。它耗资1200万，历时4年，实验学生达12,000名，被誉为美国历史上最伟大的教育实验。STAR计划及后续研究表明，小班化教育有利于学生学业成就的提高，尤其对处境不利学生更具优势，并且这种优势具有累积性
- 学校： 79所
- 学生随机分配到三类班级：普通班（22-25）；加强普通班（22-25，加辅导老师）；小班（13-17）
- 干预变量：班级类型
- 潜在结果：学习成绩

实验的理论基础

$$\begin{cases} Y_i(regular), D_i = regular \\ Y_i(regularaid), D_i = regularaid \\ Y_i(small), D_i = small \end{cases}$$

- 实验目的：研究小班和加强普通班相对于普通班对学习成绩的影响
  - 小班和加强普通班的平均处置效应，根据定义:

$$ATE(small) = E\left[Y_i(small)\right] - E\left[Y_i(regular)\right]$$
$$ATE(regularaid) = E\left[Y_i(regularaid)\right] - E\left[Y_i(regular)\right]$$

- 给定学校g, 学生的潜在成绩和班级类型是独立的, 即CIA条件独立假设 $\left\{Y_i\left(small\right), Y_i\left(regularaid\right), Y_i\left(regular\right) \perp D_i \mid school_i\right\}$ 意味着对于同一个学校, 不同班级的学生没有区别。

实验的理论 → 应用：确保随机分配

- step1: 检验是否随机分配: 确保可观测的特征在干预组和控制组是均匀的

```
Stata 命令：  table glclasstype,
c(mean female mean whiteasian mean age1985 mean glclasssize) cell ( 18 )
```

结果：除了小班平均人数少于加强普通班和普通班，其他均无差异

- step2: 为比较有无控制变量的差异，先对下模型回归

$$Y_{gi} = \alpha + \tau_1 small_{gi} + \tau_2 regularaid_{gi} + e_{gi}$$

score	Coeff.	Std.Err.
small	29.78802	2.8076
regularaid	11.93308	2.6862
_cons	1039.393	1.8361

glclasstype	Mean(score)
1	1069.181
2	1039.393 PS：普通班，常数项
3	1051.326

实验的理论 → 应用：消除偏误

- step3: 加入学校固定效应，对下模型回归

$$Y_{gi} = \alpha + \tau_1 small_{gi} + \tau_2 regularaid_{gi} + \beta school_g + e_{gi}$$

score	Coef.
small	29.005
regularaid	7.217
_Iglschid_123056	55.609
_Iglschid_128076	43.541
_Iglschid_128079	47.238

对这两项系数影响不太大

- 由于学校固定效应与班级大小有相关性，城市里班级会有更多非普通班，因此如果不控制学校固定效应，该项目就会进入LPF扰动项，造成有偏估计。只不过本例中相关性较小，加入学校固定效应对估计结果影响不大。

实验的理论 → 应用：提高估计精度

理论上，加入学校固定效应已经可以没有偏误。但若在干扰项中再“提炼”出与因变量学习成绩相关的控制变量，有助于降低干扰项的方差，提高估计精度。

- step4: step3 + 性别、年龄、种族

固定效应	Root MSE	Adj R—square
不加参数	90.512	0.0171
固定学校	79.17	0.2480
固定性别，年龄，人种	78.084	0.2685

- 加入这些控制变量后，step3的估计系数变化不大（标准误变小），说明新加入的控制变量与干预变量班级类型是不相关的；**这与step1中这些变量再不同班级类型中没有显著差异是一致的**

实验的理论 → 应用：考虑集群方差，增加结论稳健性

因为同方差假设不成立的话，一般情况会导致目标参数估计值的标准误差的**低估**，当低于临界值时，造成原本不显著的系数变得显著，导致对结论的过度自信。

⟶ 也就是说在考虑异方差、集群方差或者自相关情况下，得到系数依然显著，说明该结论是稳健的。

本例中，由于同一所学校的干扰项可能存在相关性，应该考虑集群方差：

step5：step4 + 集群方差

结果：**标准误差变大， t值下降，说明存在显著的集群方差**。但干预变量的估计系数仍然显著，可以提高结论的稳健性。

基于RCT实例的延伸

通过上述实例的分析，可以了解在RCT+LFP下必须基于干扰项条件均值独立于干预变量的假设，才能识别真实的处置效应。

此外，上述流程也是经验分析在应用过程中必须经历的。

本小节，再就几个问题进行延伸，作为本课程的第一部分完结。

SUTVA

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**和**一般均衡效应**。
    - 例如：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，从而提高大家的学习效率。相反，调皮多，可能产生负外部性，整个班级的学习效果都不好。再有较小，不会对项目外的劳动力市场产生影响。另外，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。

SUTVA

- 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.
- 可以证明：**即便我们放松了第2层假设，即允许 LPF^{ols} 中存在异质性的个体处置效应，当潜在结果均值独立假设成立时，也可以在LPF^{ols} 识别 平均处置效应ATE**。

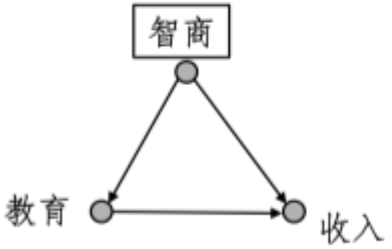
因果关系估计偏差: 类型与解决办法

因果关系估计偏差: 类型与解决办法

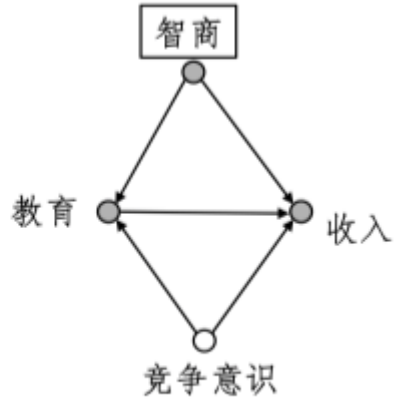
- 辛普森悖论 $\longrightarrow$ 相关关系未必一定反应因果关系
- 有向无环图
  - 因果路径
  - 混淆路径: A $\leftarrow$ B $\rightarrow$ C , B是A和C的混淆变量; 混淆变量会导致相关关系
  - 对撞路径: A $\rightarrow$ B $\leftarrow$ C , B是A和C的对撞变量; 对撞变量不会产生相关性
- 估计X与Y的因果关系的本质是找到二者间所有的因果路径，同时去 除二者间的非因果关系路径。

因果关系估计偏差: 混淆偏差

- 混淆偏差是指在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

因果关系估计偏差: 过度控制偏差

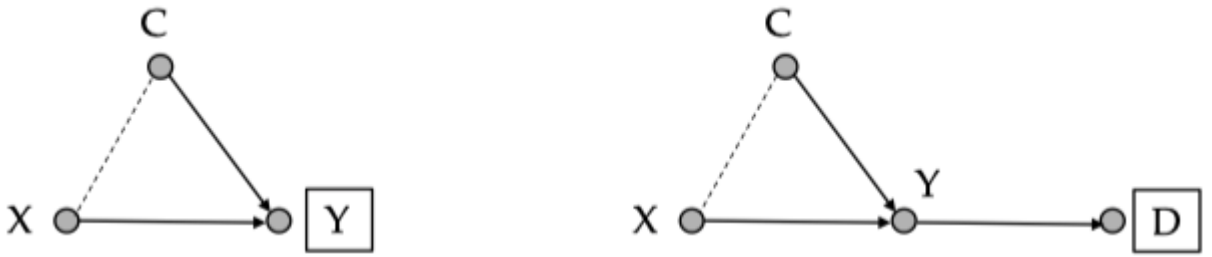
- 过度控制偏差是指控制了因果路径上的变量造成的偏差。
- 在研究中我们要避免控制受X影响并会影响Y的中介变量，否则会造成过度控制偏差。

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

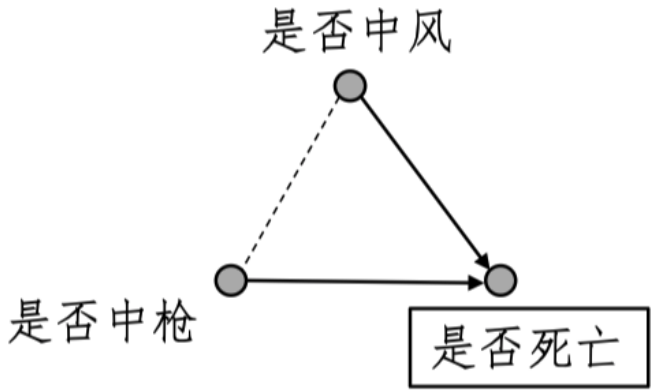
因果关系估计偏差: 对撞偏差

- 对撞偏差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结果变量），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：对撞偏差

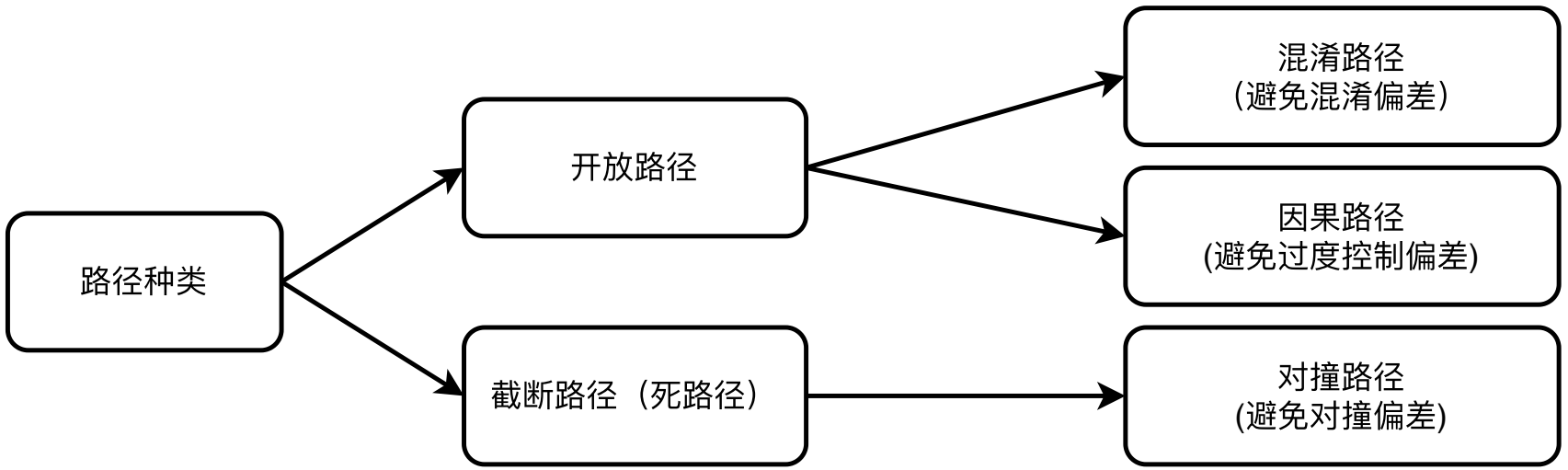
因果关系估计偏差: 对撞偏差



是否死亡	是否中风	是否中枪
否	否	否
是	否	是

因果关系估计偏差: 类型小结

- 由于因果关系通常无法被直接观测到，只能通过变量间的相关性去推测因果关系，因此从路径的角度上讲，分析因果关系的本质就是：截断混淆路径、避免过度控制以及避免对撞路径产生的衍生路径



因果关系估计偏差: 解决办法

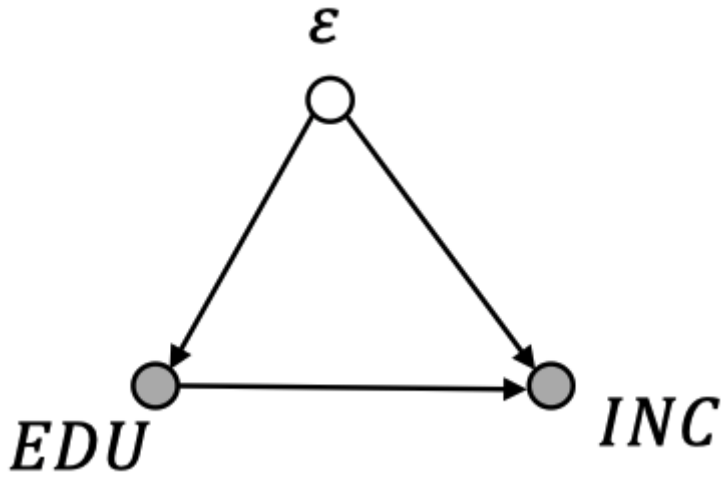
- 了解偏差来源 ⟺ 建立有向无环图 ⟺ 寻找对应解决办法 ⟺ 完全知晓数据产生过程
- 观察性研究很难做到

- 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 。

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



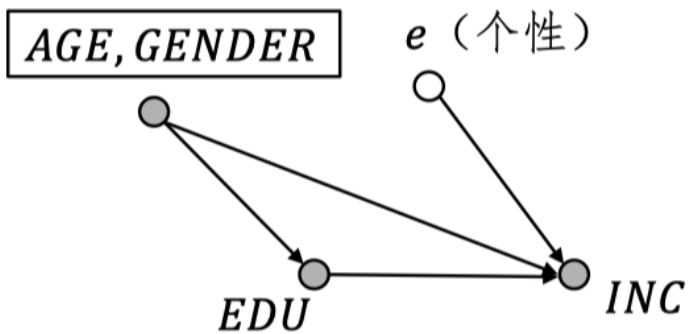
图：LPF等价于在DAC中忽略了年龄和性别

因果关系估计偏差: 解决办法

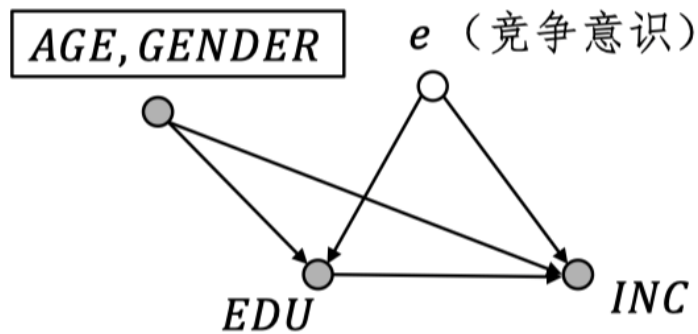
- 可以将 ε_{it} 中的可观测变量分离进行控制，得到 LPF[1]

  $INC_{it} = \alpha + \beta_1 EDU_{it} + \beta_2 AGE_{it} + \beta_3 GENDER_i + e_{it}$ ，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



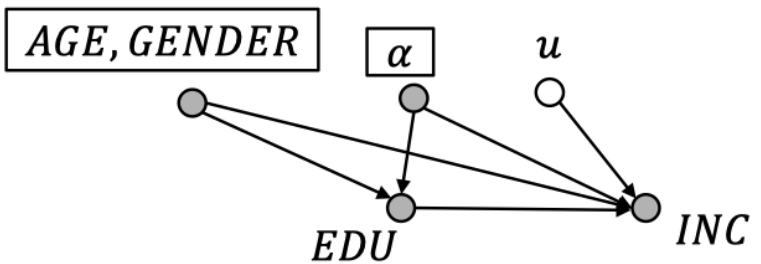
图：控制年龄和性别且e为混淆变量的情况

因果关系估计偏差: 解决办法

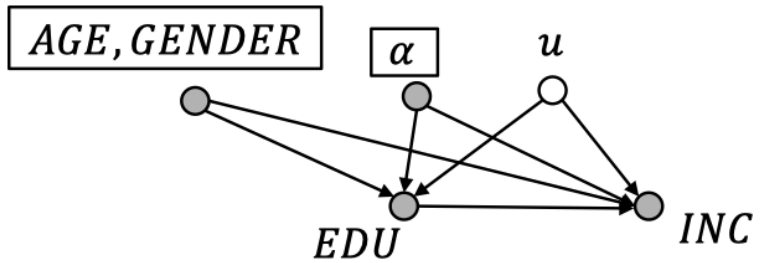
- 若真实关系是右图，仍无法识别因果关系
- 将 ε_{it} 进一步分解为：不可观测的非时变变量 α_i 和不可观测的时变变量 u_{it} ，即 $e_{it} = \alpha_i + u_{it}$

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

- 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。



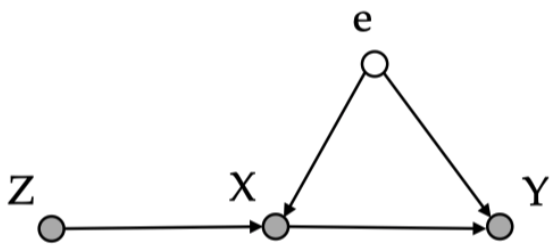
图：控制年龄、性别和α且u为无关变量的情况



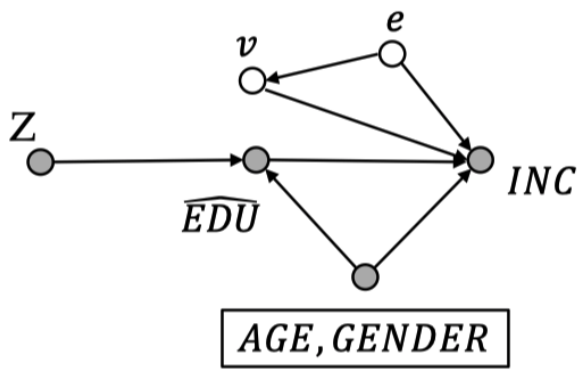
图：控制年龄、性别和α且u为混淆变量的情况

因果关系估计偏差: 解决办法

- 实际上，很多社会研究不太能轻易地在 ε_{it} 中区清晰地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 解释变量中不被 e_{it} 混淆的信息来估计解释和被解释变量的因果关系。
  - 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



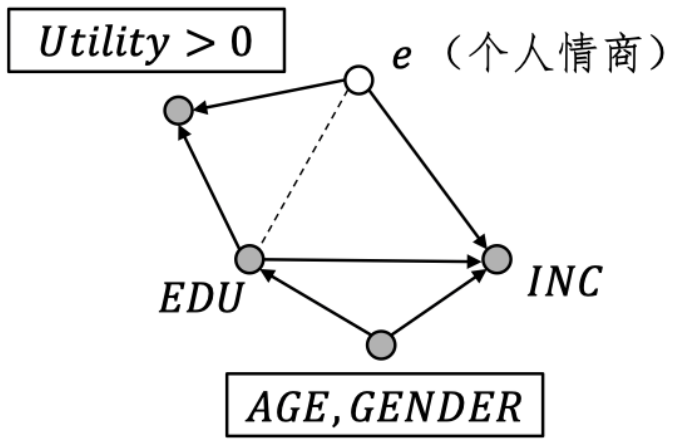
图：工具变量的相关性和外生性



图：引入工具变量的情况

因果关系估计偏差: 解决办法

- 若存在对撞偏差。实际中会情况是，对撞偏差不是因为解释变量和不可观测因素 e 在总体里存在相关性造成的，而是由于用来估计的样本不是从总体里 的随机抽取，导致样本里解释变量和不可观测因素 e 存在相关性。
    - 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：对撞偏差

因果关系估计偏差: 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

- 在STAR中，我们通过“简单回归 + RCT + 潜在结果条件均值独立于干预变量的假设 ”避免了可观测因素带来的混淆偏差

关于控制变量

关于控制变量

Q STAR估计过程显示增加控制变量可以提高回归估计值的可信度，逐渐的让回归的参数趋向于真实的因果效果，但是不是控制的变量越多越好?

A 当然不是。接下来就看下什么是"好的"控制与"坏的"控制?

但是之前，我们先引入"遗漏变量偏误"的概念。

关于控制变量：遗漏变量偏误

例子：教育回报率

我们有两个线性、总体模型:

$$Y_i = \alpha + \rho s_i + \eta_i \tag{1}$$
$$Y_i = \alpha + \rho s_i + X_i'\gamma + \nu_i \tag{2}$$

我们不能将 (1) 中的 $\hat{\rho}$ 解释为因果关系(考虑到可能存在的选择偏误)

对于模型 (2), 可以将 $\hat{\rho}$ 解释为因果关系 *考虑到假设* $Y_{si} \perp\!\!\!\perp s_i | X_i$ (CIA) 的成立.

换句话说, 条件独立假设告诉我们 **可观测的向量 X_i 必须能解释 s_i 与 η_i 间的全部相关性**.

关于控制变量：遗漏变量偏误公式

我们使用遗漏变量偏误公式(Omitted Variable Bias Formula)描述 **当回归包含不同的控制变量时，回归结果之间存在的关系。** 它提供了**长回归方程**和**短回归方程**估计系数之间的联系。

为更简明，假设控制变量可以简化为**家庭背景、智力和动机**所组成的控制变量集合。将这些变量所组成的向量记为 (A_i) 并将此变量简记为“能力”。在控制了能力后，对工资关于教育水平进行回归的方程就可以写成：

$$Y_i = \alpha + \rho s_i + A_i'\gamma + e_i \tag{1}$$

其中, α, ρ, γ 是总体回归系数, e_i 是回归残差。由定义可知 e_i 和所有的回归元都无关 (e_i 是控制了 A_i 后潜在收入水平的随机部分）。给定 A_i, 如果CIA成立, 那么系数 ρ 就是我们感兴趣的处置效应。但在实际中, 能力是很难度量的。 那么, 将“能力"排除在回归方程（1）之外的后果是什么呢?此时回归方程变为：

$$Y_i = \alpha + \beta s_i + v_i \tag{2}$$

关于控制变量：遗漏变量偏误公式

我们将方程 (2) 称为**短回归方程**, (1) 称为**长回归方程**。 将“能力"排除在外的短回归方程参数与长回归方程 (1)得到的参数之间存在的关系由下式给出: **遗漏变量偏误公式 (Omitted Variable Bias Formula)**

$$\hat{\beta}_{ols} = \frac{Cov(Y_i,\, s_i)}{Var(s_i)} = \rho + \gamma' \delta_{As}$$

其中, δ_{As} 是对 A_i 关于 s_i 回归得到的参数。

该公式表明：**短回归参数**等于**长回归参数**加上一个数，这个数等于**遗漏变量效应**乘以**遗漏变量对被包含变量的回归系数**。

因此，满足以下两个条件，那么长、短回归方程对教育回报率的估计将一样: (**a**) 受教育程度与能力大小无关 $(\delta_{As} = 0)$ *或者* (**b**) 在控制受教育程度后，能力大小与工资多少无关 $(\gamma = 0)$.

关于控制变量：例子(MHE)

表 3.2.1 教育回报率(MHE)

	1	2	3	4
教育程度	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
控制变量	None	Age Dum.	2 + Add'l	3 + AFQT

在这里，我们有四种具体控方式，关于工资对上学年限的回归（来自NLSY，美国青年纵向调查）。

关于控制变量：例子(MHE)

表 3.2.1 教育回报率(MHE)

	1	2	3	4
教育程度	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
控制变量	None	Age Dum.	2 + Add'l	3 + AFQT

列1 (没有控制变量) 意味着每额外获得1年教育，工资有13.2%的增长.

关于控制变量：例子(MHE)

表 3.2.1 教育回报率(MHE)

	1	2	3	4
教育程度	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
控制变量	None	Age Dum.	2 + Add'l	3 + AFQT

列2 (控制年龄) 意味着每额外获得1年教育，工资有13.1%的增长.

关于控制变量：例子(MHE)

表 3.2.1 教育回报率(MHE)

	1	2	3	4
教育程度	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
控制变量	None	Age Dum.	2 + Add'l	3 + AFQT

列3 (列2的控制变量再加上父母教育和自身人口学特征) 意味着每额外获得1年教育，工资有11.4%的增长.

关于控制变量：例子(MHE)

表 3.2.1 教育回报率(MHE)

	1	2	3	4
教育程度	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
控制变量	None	Age Dum.	2 + Add'l	3 + AFQT

列4 (列3列3 AFQT† 分数) 意味着每额外获得1年教育，工资有8.7%的增长.

† *AFQT* is *Armed Forces Qualification Test*，*武装部队资格测验*, 反映能力

关于控制变量：例子(MHE)

表 3.2.1 教育回报率(MHE)

	1	2	3	4
教育程度	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
控制变量	None	Age Dum.	2 + Add'l	3 + AFQT

随着我们加大控制力度，教育回报率估计值下降了4.5个百分点（系数下降34%），从**第1栏**到**第4栏**。

$$\frac{Cov(Y_i,\ s_i)}{Var(s_i)} = \rho + \gamma' \delta_{As}$$

如果我们常识性的认为**能力大小对工资多少具有积极(正的)的影响**，那么看起来我们也可以得到**能力大小对受教育程度同样具有积极(正的)选择**的结论。

关于控制变量：遗漏变量公式与条件独立假设

$$\frac{\mathrm{Cov}(\mathrm{Y}_i,\, \mathrm{s}_i)}{\mathrm{Var}(\mathrm{s}_i)} = \rho + \gamma' \delta_{As}$$

上式除了帮助我们思考和了解OVB的方向外，该公式还使我们认识到，为了能够将多元回归系数解释为因果关系，我们在条件独立假设上靠得很紧。

Q 什么时候条件独立假设是可信的?

A 两个潜在的答案

1. 随机实验
2. 分配方案符合具有任意截断/抽签的程序

利用回归分析进行处置效应参数识别时，应该如何选择控制变量呢？或者说什么样的变量才是好的控制变量，什么样的变量是坏的控制变量？

关于控制变量：坏的控制

好的控制变量是发生在干预变量[†]之前或取值不随干预变量变化的变量，即好的控制变量是潜在的混杂因素。

在估计大学教育收益率时，影响个体教育决策的变量是好的控制变量，比如考大学时的个体特征（不随教育决策变化的性别、民族等）、家庭背景（当时父母的收入教育等）、学习成绩（高中学习成绩）、地域等。但今天的家庭背景等信息可能不一定是好的控制变量。相应地，可能受到干预变量影响的变量或发生在干预变量之后的变量不是好的控制变量。

†在多元回归分析中，**解释变量(自)和控制变量(协)**的地位是不一样的。强调因果识别，关系的解释变量称为干预变量或干预变量。在回归分析中，主要关心干预变量的系数是否有处置效应的解释。对于控制变量，引人它们的目的是为了保证干预变量系数有处置效应的解释，控制变量本身的系数是否有处置效应解释并不重要.

关于控制变量：坏的控制

仍以教育收益率的估计为例，个人职业和就业行业就不是好的控制变量。

为什么?

因为个人职业及就业行业往往是教育完成之后个人选择的结果，也就是说，这些变量发生在教育之后，可能受到教育的影响。事实上，教育对职业选择有重要影响，职业和行业往往存在着密切的联系，这些变量就可能成为教育和收入的交汇变量，以交汇变量（包括交汇变量的子孙变量）为条件，将产生样本选择偏差.

例子

假设我们想知道**大学毕业对工资的影响**。

1. 只有两种类型的工作：蓝领和白领。
2. 白领工作的平均工资比蓝领工作高。
3. 大学毕业会增加获得白领工作的可能性。

Q 在考虑大学毕业对工资的影响时，应该控制职业类型吗？(职业会不会是一个遗漏的变量？)

A 不用.假设大学学位是随机分配的。当控制职业类型时，我们其实比较的是那些选择了蓝领工作的人中获得学位的与那些选择蓝领工作的人中没有获得学位的两个群体的工资差别。我们只对学位(干预变量)是随机分配进行假定，**并没有**对于工作类型(控制变量)是随机分配进行任何假设。

形式化推导1

更正式地，让

- W_i 表示 i 是否白领工作的虚拟变量
- Y_i 表示 i 的收入
- C_i 指 i 的**随机分配的**大学是否毕业状态

$$
\begin{aligned}
Y_i &= C_i Y_i(1) + (1 - C_i)\, Y_i(0) \\
W_i &= C_i W_i(1) + (1 - C_i)\, W_i(0)
\end{aligned}
$$

因为假设 C_i 是随机分配的，所以平均值的差异就是因果估计, *i.e.*

$$
\begin{aligned}
E[Y_i \mid C_i = 1] - E[Y_i \mid C_i = 0] &= E[Y_i(1) - Y_i(0)] \\
E[W_i \mid C_i = 1] - E[W_i \mid C_i = 0] &= E[W_i(1) - W_i(0)]
\end{aligned}
$$

形式化推导2

让我们看看当我们加入一些控制措施时会发生什么--*例如*,专注于大学毕业对白领工作的工资影响。

$$E[Y_i \mid W_i = 1, C_i = 1] - E[Y_i \mid W_i = 1, C_i = 0]$$

$$= E[Y_i(1) \mid W_i(1) = 1, C_i = 1] - E[Y_i(0) \mid W_i(0) = 1, C_i = 0]$$

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

$$\begin{aligned} = {} & E[Y_i(1) \mid W_i(1) = 1] - E[Y_i(0) \mid W_i(1) = 1] \\ & + E[Y_i(0) \mid W_i(1) = 1] - E[Y_i(0) \mid W_i(0) = 1] \end{aligned}$$

$$= \underbrace{E[Y_i(1) - Y_i(0) \mid W_i(1) = 1]}_{\text{白领工人的处置效应}} + \underbrace{E[Y_i(0) \mid W_i(1) = 1] - E[Y_i(0) \mid W_i(0) = 1]}_{Selection-bias}$$

形式化推导3

引入坏的控制，相当于额外地把选择偏误引入了进来。(之前没有控制就没有选择偏差)。具体来说，选择偏差项

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

前者是有大学文凭的白领工人的倘若没有毕业的收入(反事实);后者是没大学文凭的白领工人的没有大学毕业的收入；其实是描述了获得大学文凭如何改变白领工人阶层的构成。

注意 即便处置效应为零，选择偏误不一定为零.

关于控制变量：更棘手的例子

一个更棘手的例子工资差距（例如，女性-男性或黑人-白人）。

Q当考虑工资性别差距时，我们应该控制职业吗?

- 我们试图捕捉什么?
- 如果我们关注的是歧视问题，那么歧视似乎也可能影响到职业选择和雇用结果。
- 有些人用组间偏好差异来刻画职业的控制。

答案是什么呢?

关于控制变量：代理变量可能也是不好的控制

Angrist和Pischke提出了一个有趣的场景，它同时包含遗漏变量偏差和不良控制。

- 我们想估计教育回报率
- 能力是遗漏变量
- 我们有一个能力的代理变量--学校教育结束后的测试。

问题是:

1. 如果我们完全省略测试，就会有遗漏变量偏差。
2. 如果我们包括代理变量，就有了一个后方控制，也就是坏的控制。

通过一些数学/运气，我们可以用这些估计值来缩小真实的处置效应估计。

关于控制变量：例子(MHE)

回到教育回报率的例子，我们控制职业。

表 3.2.1 教育回报率(MHE)

	1	2	3	4	5
教育程度	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)
控制变量	None	Age Dum.	2 + Add'l	3 + AFQT	4 + Occupation

受教育程度可能影响职业选择；我们如何解释新的结果?

关于控制变量：例子(MHE)

回到教育回报率的例子，我们控制职业。

表 3.2.1 教育回报率(MHE)

	1	2	3	4	5
教育程度	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)
控制变量	None	Age Dum.	2 + Add'l	3 + AFQT	4 + Occupation

受教育程度可能影响职业选择；我们如何解释新的结果?

其实：我们很难解释是何种原因导致了这种下降。 教育水平的系数变小可能仅仅是选择偏误的一种表现。因此最好还是用不由教育水平决定的那些变量作为控制变量。

如何加入控制变量：再次重申(MHE)

- 对协变量的控制可以提高使多元回归估计值获得因果解释的可能性，但并非控制变量越多越好。有些控制变量是不合格的控制变量，将其加入回归固然可以改变回归系数，但实际上却不该将其加人。

- 由于我们总是可以将经验研究想象为一个实验，所以**不合格的控制变量**就是那些可以作为实验结果的变量。也就是说，不合格的控制变量本身可作为被解释变量。合格的控制变量是指当我们选定干预变量后，它的取值已经固定给出的那些变量。

时机很重要!

- 对坏的控制变量和代理性控制变量都适用的一个挑选准则是：**考虑控制变量被决定的时间**。一般来说在X产生之前就被决定的变量都是好的控制变量。但是有时必须面对控制变量被决定的时间不确定或未知的情况。在这种情况下，因果关系的准确考量需要我们做出哪个变量先被决定的假设，或者去说明没有任何个控制变量是由我们感兴趣的变量所影响的。

内生性问题

内生性问题

- 对于给定的线性回归模型

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e$$

如果干扰项和解释变量是相关的, 即

$$E\left(e \mid X_1, X_2, \cdots, X_k\right) \neq 0$$

那么可以说这个线性模型存在内生性问题。

- “内生性问题”指当干扰项和解释变量相关时，我们无法识别解释变量的因果关系系数的情形（常见于经济类文章）
- **简单概括：内生性问题是在CEF-LPF框架下对因果关系估计偏误的表述**

内生性问题

来源一: 遗漏变量

考虑模型: $INC = \alpha + \beta_1 EDU + \beta_2 IQ + e$

其中: $E(e \mid EDU, IQ) = 0, \mathrm{Cov}(EDU, IQ) \neq 0$

若遗漏了解释变量 IQ, 即使用 $INC = \alpha + \beta_1 EDU + v$ 进行回归, 则

$E(v \mid EDU) = E\left(\beta_2 IQ + e \mid EDU\right) = \beta_2 E(IQ \mid EDU) \neq 0$

其关系可以根据之前的OVB公式得到。

- 遗漏解释变量会造成内生性原因是遗漏变量和未遗漏 解释变量之间存在相关性

内生性问题

来源二: 测量误差

(1) 解释变量的测量误差

考虑模型:

$$Y^* = \beta_0 + \beta_1 X^* + e, \quad E(e \mid X^*) = 0$$

当解释变量 X^* 存在测量误差, 即 $X = X^* + u$, 同时

$$\mathrm{Cov}(u, X^*) = 0, \quad \mathrm{Cov}(u, Y^*) = 0, \quad E(u \mid X^*) = 0$$

此时模型变成

$$Y^* = \beta_0 + \beta_1(X - u) + e = \beta_0 + \beta_1 X + v$$
$$v = -\beta_1 u + e$$

内生性问题

来源二: 测量误差

$$
\begin{aligned}
&\operatorname{Cov}(X, v) \\
&= \operatorname{Cov}(X^* + u, -\beta_1 u + e) \\
&= \operatorname{Cov}(u, -\beta_1 u) \\
&= -\beta_1 \operatorname{Var}(u) \neq 0
\end{aligned}
$$

- 解释变量存在测量误差时, 会造成内生性问题
- 原因: 解释变量存在测量误差时会造成干扰项里面存在测量误差, 从而导致干扰项和观测的解释变量相关 (均包含了测量误差)

内生性问题

来源二: 测量误差

- (2) 被解释变量的测量误差 考虑模型:

$$Y^* = \beta_0 + \beta_1 X^* + e, \quad E(e \mid X^*) = 0$$

当解释变量 Y^* 存在测量误差, 即 $Y = Y^* + u$, 同时

$$\mathrm{Cov}(u, X^*) = 0, \quad \mathrm{Cov}(u, Y^*) = 0, \quad E(u \mid X^*) = 0$$

此时模型变成

$$Y = \beta_0 + \beta_1 X^* + e + u = \beta_0 + \beta_1 X^* + v$$
$$v = e + u$$

内生性问题

来源二: 测量误差

$$
\begin{aligned}
&\operatorname{Cov}(X^*, v) \\
&= \operatorname{Cov}(X^*, e + u) \\
&= 0
\end{aligned}
$$

- 当被解释变量存在测量误差时, 不会造成内生性问题
- 但是由于误差项 (噪音) 变大, 回归结果的显著性会有所降低(实践中有可能干预变量变的不显著，但显著的干预变量的参数估计是一致的)

内生性问题

来源三: 互为因果

- 若两个变量互为因果, 则任何一方都可以作为对方的解释变量, 此时任何一个单方面的回归都存在内生性问题。考虑如下情形

$$Y_1 = \beta_1 X_1 + \phi_1 Y_2 + e_1 \tag{1}$$
$$Y_2 = \beta_2 X_2 + \phi_2 Y_1 + e_2 \tag{2}$$

$$E\left(e_i \mid X_1, X_2\right) = 0; \quad i = 1, 2 \quad ; \mathrm{Cov}(e_1, e_2) = 0$$

将式（2）代入式（1）中,可以得到

$$Y_1 = \frac{\beta_1}{1-\phi_1\phi_2} X_1 + \frac{\beta_2\phi_1}{1-\phi_1\phi_2} X_2 + \frac{e_1}{1-\phi_1\phi_2} + \frac{e_2\phi_1}{1-\phi_1\phi_2} \tag{3}$$

内生性问题

来源三: 互为因果

由式（3）

$$
\begin{aligned}
&\operatorname{Cov}(Y_1, e_2) \\
&= \operatorname{Cov}\left(\frac{\beta_1}{1-\phi_1\phi_2} X_1 + \frac{\beta_2\phi_1}{1-\phi_1\phi_2} X_2 + \frac{e_1}{1-\phi_1\phi_2} + \frac{e_2\phi_1}{1-\phi_1\phi_2}, e_2 \right) \\
&= \operatorname{Cov}\left(\frac{e_2\phi_1}{1-\phi_1\phi_2}, e_2 \right) \\
&= \frac{\phi_1}{1-\phi_1\phi_2} \operatorname{Var}(e_2) \neq 0
\end{aligned}
$$

- 所以模型 (2) 存在内生性问题（对简化式2进行回归），模型 (1) 同理可证