

# 体育经济分析: 原理与应用

## 单元4: 从相关到因果的体育经济

周正卿

08 July 2023

# 大纲

# 大纲

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

# 什么是计量经济学

- Econometrics
- 计量经济学只是统计学的经济学版本吗?
  - 经济学理论 + 统计学技巧 = 计量经济学
  - 使用统计学的技巧（工具包），并将其应用于与经济学中的问题和现象有关的问题。
- 相关性足够有趣，为什么更偏好因果?
  - 预知某项干预/政策对感兴趣变量的影响结果 → 预判问题
  - 资源是有限的，而干预措施是要花钱的 → 更为精确

## 例子：球队胜率与上座率

假设联盟总裁想知道哪些因素能提升比赛上座率最重要

- 一个方法是用平均观众人数反映上座率现实情况，但是它无法反映变量间的关系
- 理论或者常识告诉我们球队获胜越多，观众就越多。把上座率ATT作为因变量，因为它是我们的目标变量，理论上我们认为它取决于胜率。而WPCT作自变量，是我们可以操纵变量。
- 将理论简化为模型  $ATT = f(WPCT)$
- 可以假设  $\frac{\partial ATT}{\partial WPCT} > 0$  以与理论描述一致
- 经过样本，检验上述观点的最直接方式是散点图。但这样无法精确计算胜率从0.5提高到0.51时，上座率会增加多少。因为提高胜率通常意味着要花钱买球员
- 为了与样本结合，最简单的方法是将理论模型改写为可回归的线性表达式 $ATT_i = \beta + \tau \times WPCT_i + \varepsilon_i$ 
  - $i$  代表每支球队

## 例子：球队胜率与上座率

- 接下来要选择一种**回归方法**，这里采取最广泛使用的一般最小二乘OLS对线性表达式进行**拟合**
- 通过限定一些**回归假设**，我们就可以得到**估计值**  $\hat{\tau}$ ，然而我们感兴趣的是**真实值**  $\tau$ ，前者是否真的代表后者? → identification
  - 举个例子：进校门人脸识别，闸机中有图片库和模型，用抽样照片去跟库里的匹配。两者匹配好肯定是根据某些算法进行的，当误差缩小到一定范围之后，闸门就会开放。现场照片就是  $\hat{\tau}$ ，而照片库里的我们就是  $\tau$ ，这个**过程**就是**识别**
  - 上面的例子中，至少有个库。但理论中的参数  $\tau$  往往不知道(或存在但观察不到)，只能经过**理论建构、稳健性检验、排除竞争性理论**等方式**推断**(猜) → 识别过程无标准
- 得到的估计值  $\hat{\tau}$  至少存在两方面意义，统计学(理论)意义和经济学(现实)意义
  - 统计学意义 (statistical significance) ，需要**统计推断**
  - 经济学意义 (economic significance, 要看**效应值**与**经济学解释**)
- 若  $\tau$  具有**因果性**，上述过程是**因果推断**

## 例子：球队胜率与上座率

$$ATT = -2.4536 + 65.23 * WPCT$$

- 告诉我们WPCT每增加一个单位(+0.001), 就会增加65.23个球迷到场。当把平均值  $WPCT=0.500$ 代入等式, 得到 30.163 或30163 球迷, 这是当年的平均上座率
- $R^2 = 0.311$  反映**拟合度**(代表胜率能够解释31.1%的上座率变化), t统计值代表显著性( $H_0 = 0, H_1 \neq 0$ )
- 但是其余的 68.9% 呢? 说明还有其他很多因素会影响, 基于以上结论的决策并没有足够说服力
  - 市场规模是一个**干扰因素**。纽约有8500万人, 辛辛那提只有30万人 → **控制**
  - 联赛差异 (分区) 是另一个干扰因素。AL和NL是存在差异的 → **控制**
  - 最关键的, 因为这两个因素很有可能会影响球队胜率, 若是忽略它们会导致**遗漏变量偏差**, 产生所谓的**内生性**问题, 导致我们用  $\hat{\tau}$  代表  $\tau$  是**偏误的**
- 我们将可回归的线性表达式改为  $ATT_i = \mathbf{X}\beta + \tau \times WPCT_i + \varepsilon_i$ 
  - 市场规模与联赛差异进入了  $\mathbf{X}\beta$

## 例子：球队胜率与上座率

$$ATT = -.421 + 1.174 \text{POP} - 2.014 \text{LEAGUE} + 59.31 * \text{WPCT}$$

R-squared = 0.4245

Adj R-squared = 0.358

Dependent variable is Attendance (in thousands)

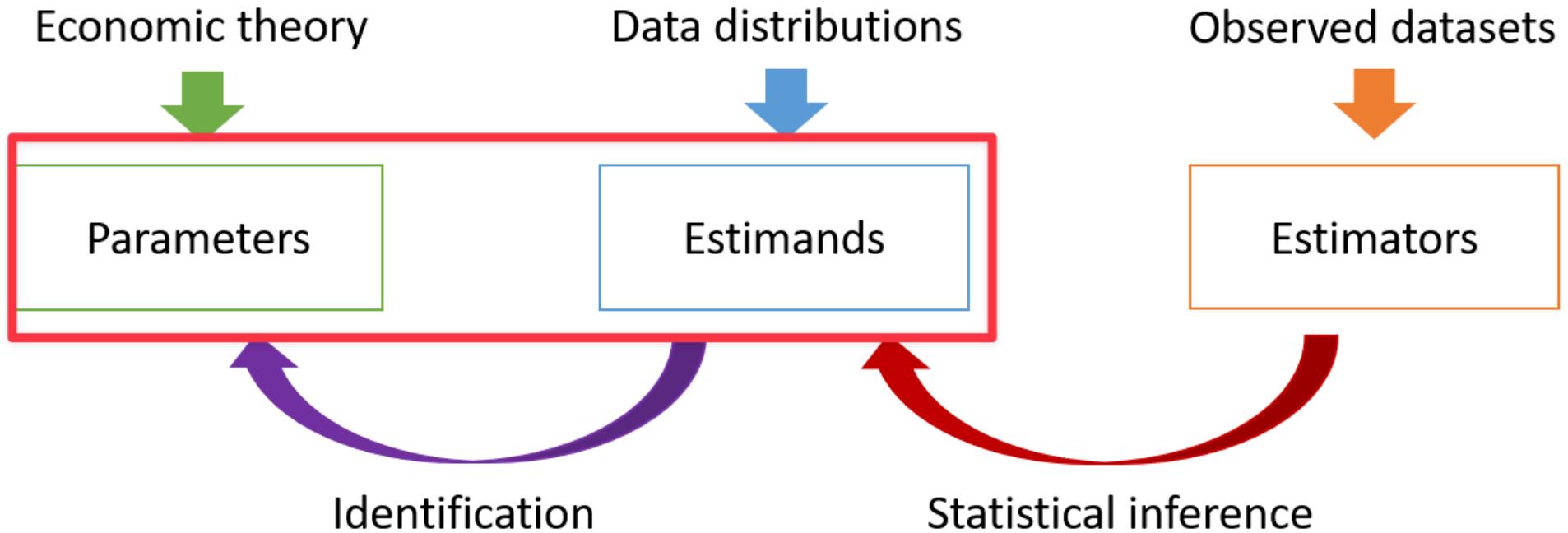
Variable	Coefficient	Std.error	t-statistic	p-value
POP (in millions)	1.174	.5602	2.10	.046
WPCT	59.310	17.939	3.31	.003
LEAGUE	-2.014	2.283	-0.88	.386
Constant	-.421	8.83	-.05	.962

- 练习：系数如何解释？具有如何意义？Dummy(AL=1) vs Continuous
- 讨论：59.31是否具有因果性？
- 59.310; 17.939; 3.31; 1.96; 0.003; 0.005

# 如何去找“真相”

按照既有的逻辑(理论、科学论证或者推断), 减少犯错的可能性

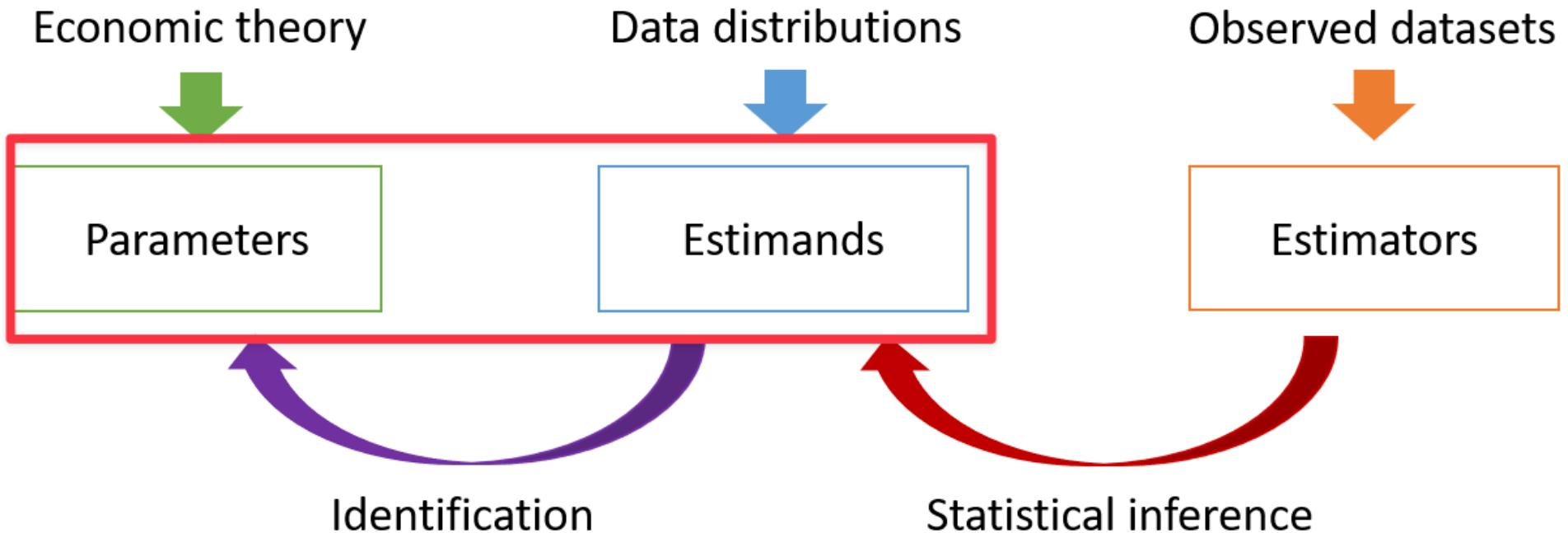
## 理论、总体分布与样本



- 样本(estmators) → 事实(=estmands?) → 因果模型(=parameter?) → Knowledge
- $2.2 \rightarrow 2.21 \rightarrow \mathbf{2.21} \rightarrow 2.21!!!$

# 从事实到因果模型：避免遗漏变量偏误 与控制变量滥用

## 理论、总体分布与样本



增加控制变量确实可以提高回归模型，让系数趋向于“因果”，但也不控制变量越多越好。为了说明这个道理，引入**遗漏变量偏误**的概念

例子：教育回报率

## 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。
- 将所有控制变量简化为**家庭背景、智力和动机**所组成的控制变量集合，标为  $(A_i)$  并记为“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：

$$Y_i = \alpha + \tau s_i + A'_i \gamma + e_i \quad (1)$$

- 其中， $\alpha, \tau, \gamma$  是总体回归系数（没有跟样本结合）， $e_i$  是回归残差。给定  $A_i$ ，如果CIA成立（ $e_i$  与其他回归元不再相关），那么系数  $\tau$  具有因果性。但在实际中，能力是很难度量的。那么，假如回归方程（1）遗漏了“能力”，此时回归方程变为：

$$Y_i = \alpha + \beta s_i + v_i \quad (2)$$

## 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**。那么**遗漏变量偏误公式 (Omitted Variable Bias Formula)**为：

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

其中,  $\delta_{As}$  是对  $A_i$  关于  $s_i$  回归得到的系数。

该公式表明：**短回归系数**等于**长回归系数**加上**bias**, 等于**遗漏变量效应**乘以**遗漏变量对自变量的回归系数**。

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 ( $\gamma = 0$ )

.

## 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，我们有四种增加控制变量的方式，关于工资对上学年限的回归（来自NLSY，美国青年纵向调查）

## 例子(MHE)

表 3.2.1 教育回报率(MHE)

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

第1列 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。

## 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			

第2列 控制年龄，意味着每额外获得1年教育，工资有13.1%的增长.

## 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

第3列，第2列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。

## 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

- 第4列(第3列又控制 AFQT<sup>†</sup> 分数)意味着每额外获得1年教育，工资有8.7%的增长。

<sup>†</sup> AFQT is Armed Forces Qualification Test, 武装部队资格测验, 反映能力

## 例子(MHE)

表 3.2.1 教育回报率(MHE)

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

- 可以看到，随着控制变量的增加，从第1列到第4列教育回报率估计值下降了4.5个百分点（系数下降34%）。

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

- 讨论为什么？

## 遗漏变量偏误公式

- 关注最想要的 OVB公式的使用~~并不要求~~每一个回归模型都能正确识别因果关系。该公式比较了~~短模型~~中的回归系数和~~长模型~~中同一变量的回归系数。<sup>†</sup>
- 条件独立假设**的重要性

$$\frac{Cov(Y_i, x_i)}{Var(x_i)} = \tau + \gamma' \delta_{Omitted-x_i}$$

- 可信的条件独立假设? → 随机分配

## 坏的控制变量

- 好的控制变量是发生在干预变量<sup>†</sup>之前或取值不受自变量影响
- 仍以教育收益率为例，个人职业和就业行业就不是好的控制变量。
- 为什么？

## 例子(MHE)

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

- 第5列，再控制职业。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

- 有些控制变量是不合格的，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入。**不合格的控制变量**会有严重的问题（比如某些药会导致高血压）。
- **时间原则**是普遍被接受的，也就是**考虑控制变量被决定的时间**。一般来说在自变量被记录之前就决定的变量大部分是好控制。
- 但是某些情况下，要考虑到**人的预期**。比如重大赛事前会超前部署

# 从事实到因果模型：避免内生性问题

- 给定一个多元线性回归模型

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

- 如果干扰项和自变量是相关的, 即

$$E(e | X_1, X_2, \dots, X_k) \neq 0$$

- 那么可以说这个线性模型存在内生性问题。
  - 无法识别自变量的因果关系系数的情形 (常见于经济类文章)
  - 若无法将内生性控制在可信服的水平下, 那么回归结果基本是无效的

## 来源一: 遗漏变量

考慮模型:  $INC = \alpha + \beta_1 EDU + \beta_2 IQ + e$

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

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

$$E(v | EDU) = E(\beta_2 IQ + e | EDU) = \beta_2 E(IQ | EDU) \neq 0$$

## 来源二: 测量误差

### (1) 自变量存在测量误差

考虑模型:

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

- 满足扰动项  $u_i$  与自变量均值独立, 且与自变量X的测量误差也独立, 且测量误差的均值为零会怎样。考虑含有测量误差的自变量观测值

$$x_{1i}^{obs} = x_{1i} + v_i$$

- 那么我们估计的方程实际上是

$$y_i = \beta_0 + \beta_1 x_{1i}^{obs} + e_i$$

其中  $e_i = (u_i - \beta_1 v_i)$

## 来源二: 测量误差

- 虽然干扰项  $e$  中的  $u$  和  $v$  是相互独立的, 但是里面含了系数  $\beta_1$ 。这时, 对于  $\beta_1$  的 OLS 估计为

$$\begin{aligned} plim \hat{\beta}_1^{OLS} &= \beta_1 + plim \frac{\sum_i \tilde{x}_{1i}^{obs} e_i}{\sum_i (\tilde{x}_{1i}^{obs})^2} \\ &= \beta_1 + \frac{-\beta_1 \sigma_v^2}{\sigma_{\tilde{x}_1}^2 + \sigma_v^2} \\ &= \beta_1 \left( \frac{\sigma_{x_1}^2}{\sigma_{x_1}^2 + \sigma_v^2} \right) \end{aligned}$$

- 自变量存在测量误差时, **会有内生性问题**
- 自变量存在测量误差时, **系数估计值在绝对值上都会减小** (经济意义不足), 该偏误叫**衰减偏误**(attenuation bias), 但**不会改变系数估计值符号**
- 偏离程度和  $\sigma_x^2 / \sigma_v^2$  **信噪比**有关

## 来源二: 测量误差

### (2) 因变量存在测量误差

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

当因变量  $Y^*$  存在测量误差, 即  $Y = Y^* + u$ , 同时

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

此时模型变成

$$\begin{aligned} Y &= \beta_0 + \beta_1 X^* + e + u = \beta_0 + \beta_1 X^* + v \\ v &= e + u \end{aligned}$$

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

- 当因变量存在测量误差时, 不会造成内生性问题
- 干扰项(噪音)变大, 导致回归结果显著性下降(现实中必须要排除掉的不显著原因, 但系数估计是一致的)

### 来源三: 互为因果

若因变量与自变量互为因果关系，即任何一方都可以作对方的自变量。

$$Y_1 = \beta_1 X_1 + \phi_1 Y_2 + e_1 \quad (1)$$

$$Y_2 = \beta_2 X_2 + \phi_2 Y_1 + e_2 \quad (2)$$

$$E(e_i | X_1, X_2) = 0; \quad i = 1, 2 \quad ; \text{Cov}(e_1, e_2) = 0$$

- 将式 (2) 代入式 (1) 中, 可以得到

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

## 来源三: 互为因果

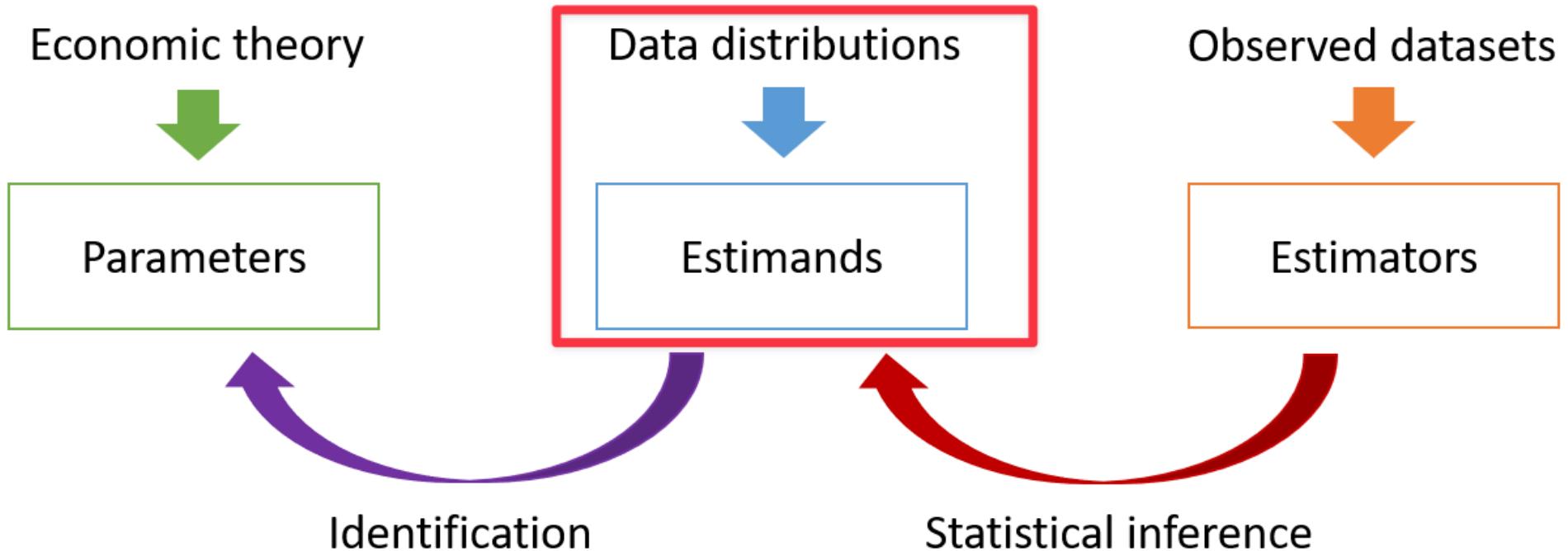
- 由式 (3)

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

- 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

事实是建立在总体意义上的模型

## 理论、总体分布与样本



- 模型代表什么?
  - 总体意义上的抽象关系

## 条件分布 (Conditional Distribution)

- 刻画变量间关系
  - 观察**条件期望**是最直接、简单办法
- 最感兴趣的  $Y$  与  $X$  是随机变量
  - $Y$  是因变量 (因变量 | 因变量 | 被解释变量) ;  $X$  是自变量 (自变量 | 干预变量 | 解释变量) .
  - 是随机变量就会有概率分布, 而最常见的是**正态分布**

## 例子：想知道工资与性别的关系

- 工资对数的条件均值可以写成如下形式：

$$E[\log(wage) \mid gender = man] = 3.05$$

$$E[\log(wage) \mid gender = woman] = 2.81$$

若是我们还好奇在种族与工资的关系，还可以增加新的条件、

$$E[\log(wage) \mid gender = man, race = white] = 3.07$$

$$E[\log(wage) \mid gender = woman, race = black] = 2.73$$

## 通过条件密度函数获得条件期望值

- 离散形式：

$$P(y|x) = \frac{P(y,x)}{P(x)}$$

其中  $P(x) = \sum_{i=1}^N P(y_i, x)$ ， 条件密度相当于联合密度  $f(y, x)$  在保持 $x$ 不变情况下的随机化“切片”

- 概率迭代法则

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

- 方差加法法则

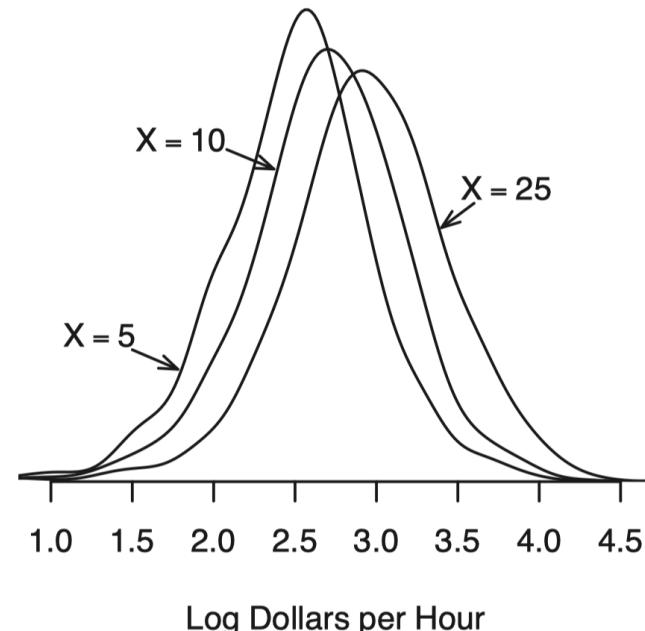
$$Var(Y) = E[V(Y|X)] + V[E(Y|X)]$$

## 通过条件密度函数获得条件期望值

为什么用联合概率分布函数和联合密度函数也可以捕捉两个变量的关系?



(a) Joint Density of Log Wage and Experience



(b) Conditional Density of Log Wage given Experience

Figure 2.4: Log Wage and Experience

事实是良好且能够被认知的

## 条件期望函数及其误差项的优良性质

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

$$E[E[Y | X]] = E[Y]$$

$E[Y|X]$  的期望值是  $[Y]$  的无条件期望值。

例如：

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

Or numerically,

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

- **性质1推论**

$$E[E[Y|X_1, X_2]|X_1] = E[Y|X_1]$$

- 内部期望值以 $X_1$ 和 $X_2$ 同时为条件,外部期望值只以 $X_1$ 为条件。迭代后的期望值可以得到简单的答案 $E[Y|X_1]$ ,即只以 $X_1$ 为条件的期望值。《E》表述为"较小的信息集获胜" → 以小谋大

例:

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

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):

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

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:**

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

使用期望迭代法则

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

所以,

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

上式取最小值, 当且仅当  $g(X) = \mu(X)$ .

## 条件期望函数及其误差项的优良性质

- Conditional Expectation Function Error (CEFE)

$$e = Y - E(Y|X) = Y - m(x)$$

- $X$  是随机变量,  $E(Y|X)$  也是随机变量
- $e$  是误差项, 也是随机变量, 具有概率分布

- CEEF 优良性质

1.  $E(e|X) = 0$
2.  $E(e) = 0$

3. 对于随机变量  $X$  任意函数形式  $h(x)$ ,  $E(h(X) \cdot e) = 0 \rightarrow$  通常利用该性质进行线性变换

## CEF与可回归的模型间的关系

step1: 定义条件期望函数  $m(x) = E(Y|X)$

step2: 定义条件期望函数的误差项  $e = Y - m(x)$

推导出：

$$Y = m(x) + e$$

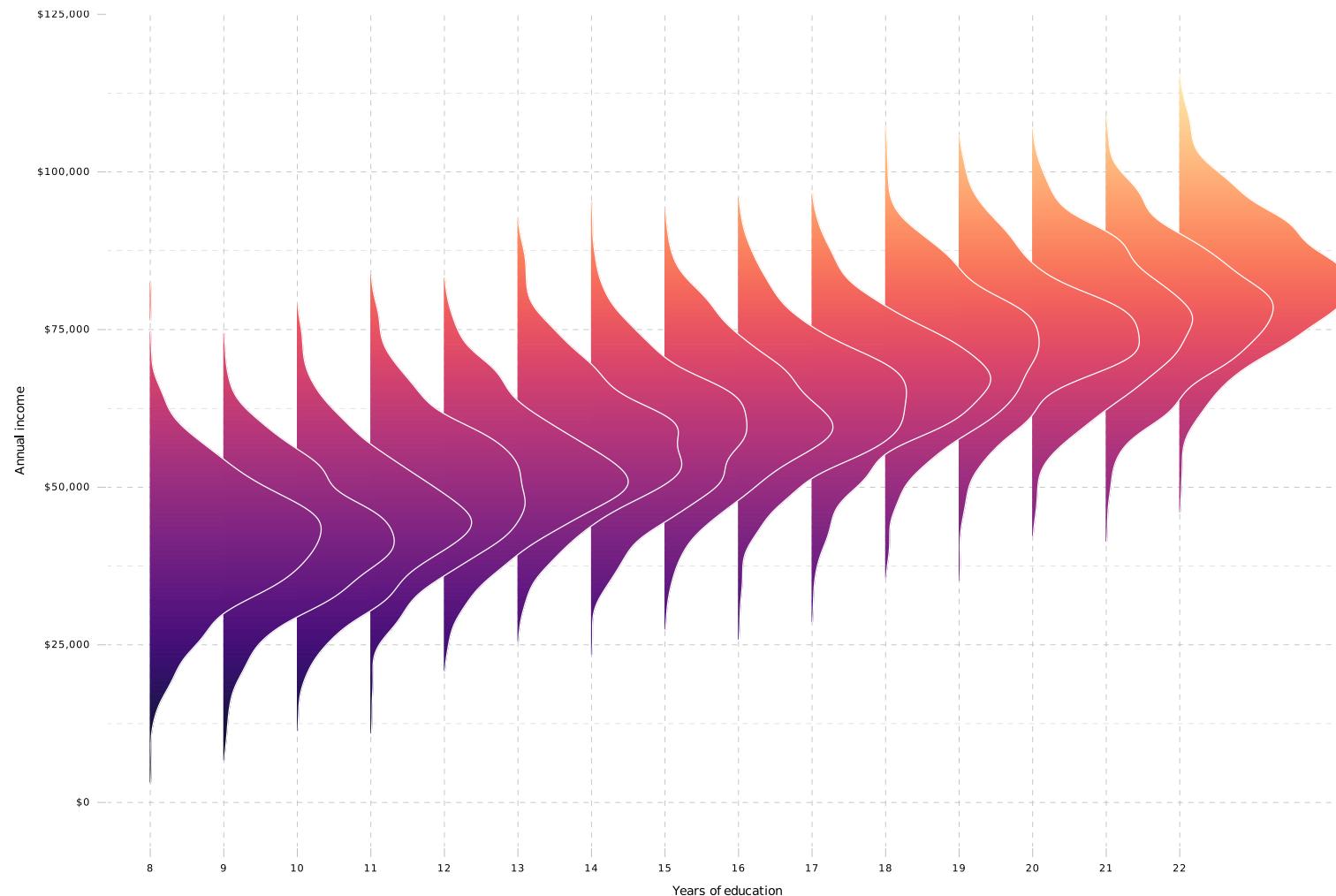
因此模型类别由  $m(x)$  形式决定：如截距模型，线性模型，Logit模型等

## CEF是从样本到总体的桥梁

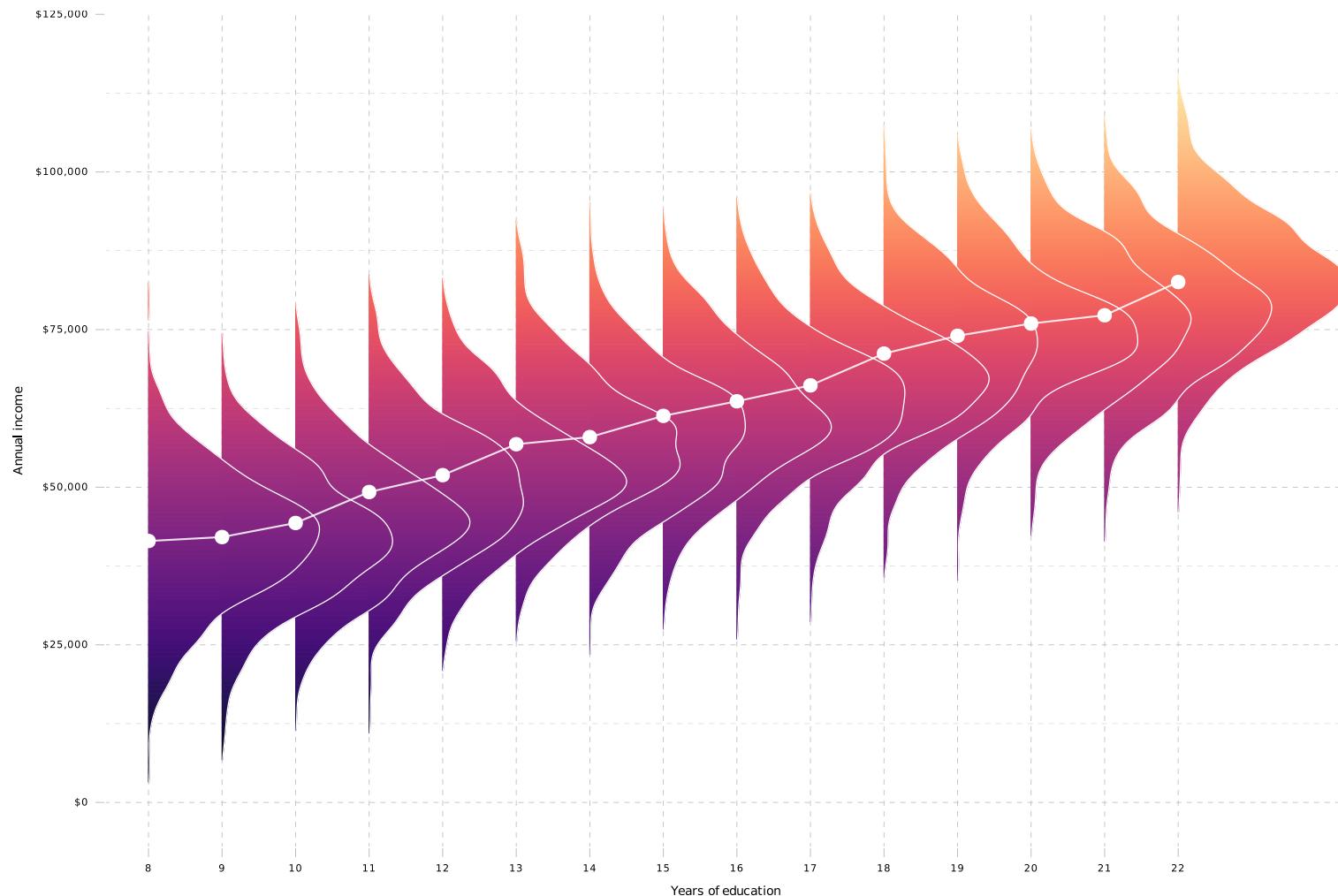
- 期望本身是总体概念（价值观）
- 实际中，我们是基于样本信息推断总体信息，例如用样本均值推断总体期望
- 将CEF写作基于样本的CEF:  $E[Y_i | X_i]$

从图形上看CEF...

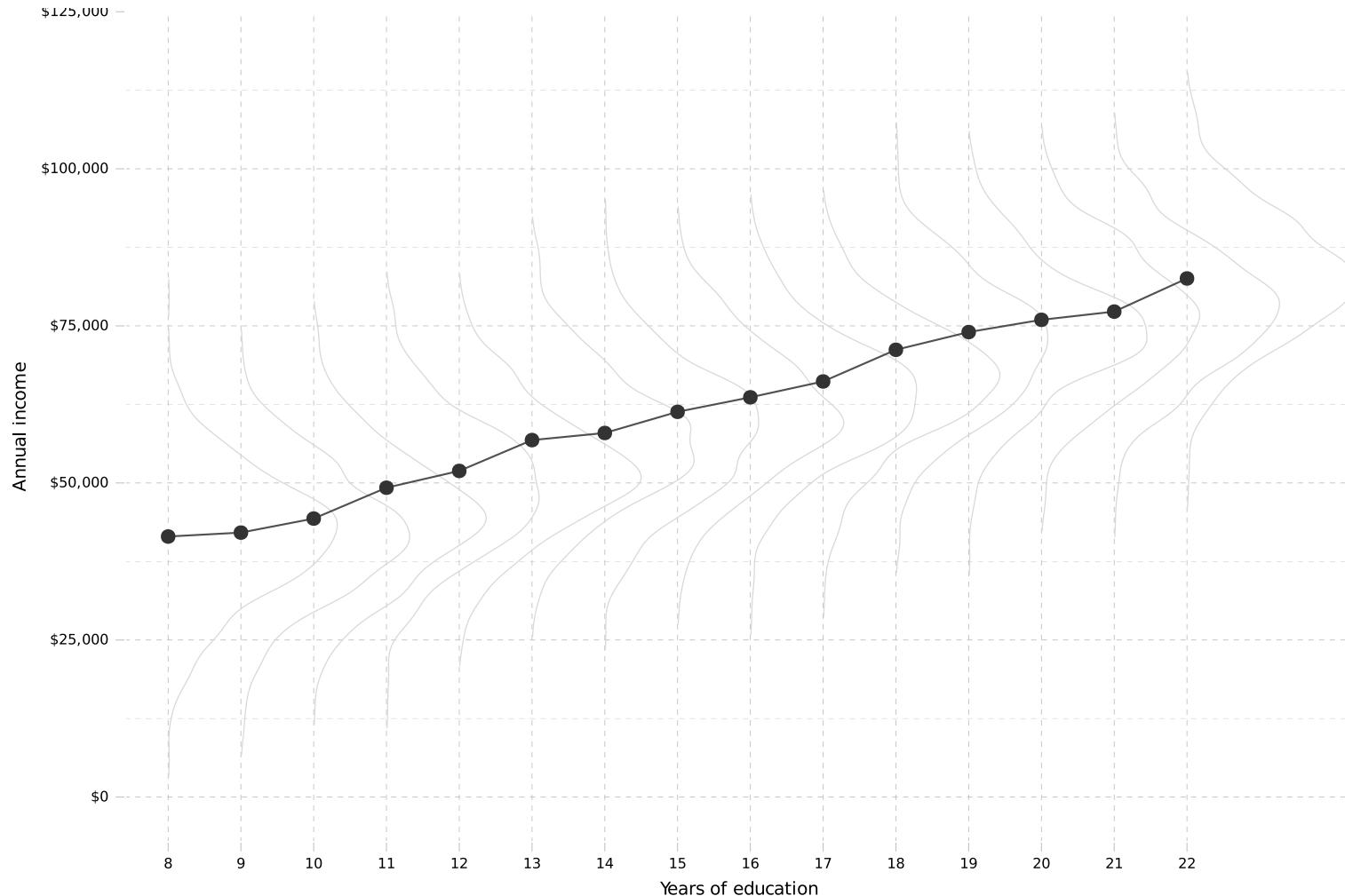
条件分布  $Y_i$ , 对于8, ..., 22不同教育年限的  $X_i = x$ .



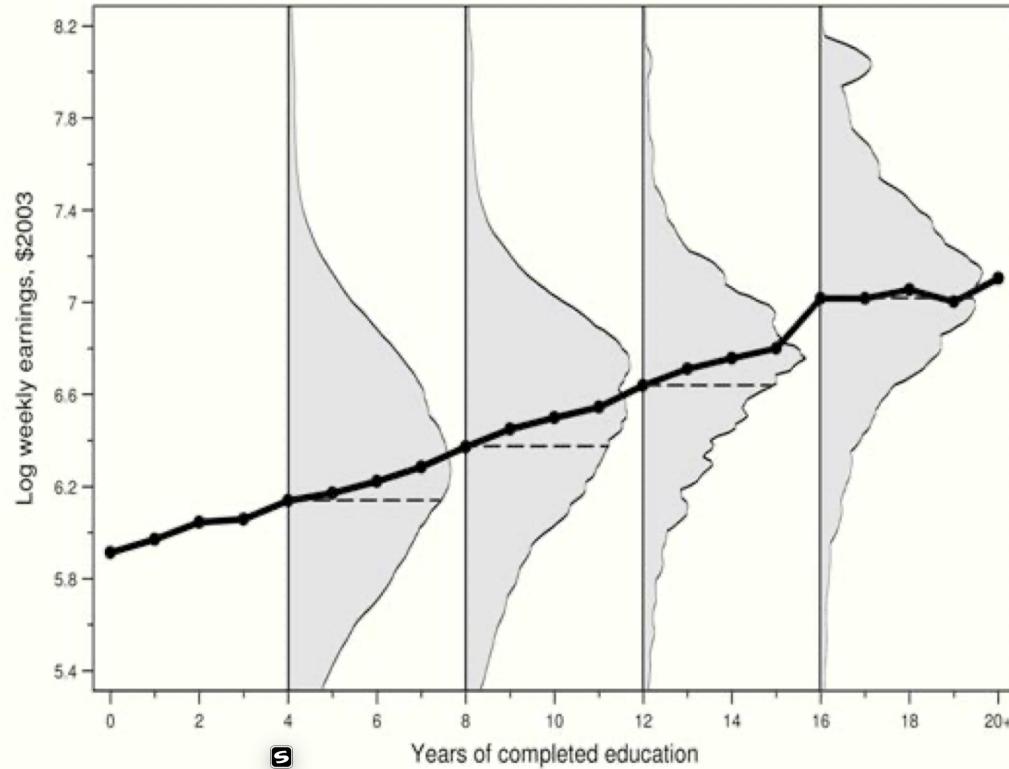
条件期望函数  $E[Y_i | X_i]$  其实是这些条件分布的均值



若只关注条件期望函数  $E[Y_i | X_i]$ ...



# 实际数据 (MHE)



© Princeton University Press. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49. The data are from the 1980 IPUMS5 percent sample.

## CEF的性质1

分解结构清楚: CEF将观测的因变量分解成两部分

$$Y_i = E[Y_i | X_i] + e_i$$

1. 被  $X_i$  解释的部分(*i.e.*, CEF  $E[Y_i | X_i]$ )
2. 具有特殊性质的干扰项<sup>†</sup>
  - i.  $e_i$  均值独立于  $X_i$ , *i.e.*,  $E[e_i | X_i] = 0$
  - ii.  $e_i$  与  $X_i$  的任何函数不相干

<sup>†</sup> 回忆之前的例子

## CEF的性质2

### ANOVA 定理:

无条件方差与条件方差的关系：可将因变量  $Y_i$  方差分解为两部分

$$Var(Y_i) = E[Var(Y_i | X_i)] + Var(E[Y_i | X_i])$$

1. 组内方差(的均值)(within group variance)。每个"等级"内Y的分布的方差的期望值(均值)。
2. 组间方差(across group variance)。条件期望值在"等级"间的分布的方差

解释为：因变量的变动 = CEF的方差(CEF可以解释) + 干扰项的方差(CEF无法解释)

## CEF的性质3

良好预测:  $m(X_i)$  为  $X_i$  任意形式函数, CEF是最小均方误差 (**性质5**)

$$E[Y_i | X_i] = \underset{m(X_i)}{\operatorname{argmin}} E[(Y_i - m(X_i))^2]$$

CEF是给定  $X_i$  能够预测  $Y_i$  最好预测方式.

$m$  可以是任意形式函数 (包含非线性) , 但更偏好**线性投影函数 (LPF)** (也叫总体回归模型)

## 练习：手算条件期望 → 从数据出发

CEF基于数据出发，对于理解变量间至关重要

研究问题  $E[\text{工资}_i \mid \text{运动技能}_i]$  :

- step 1: 选取  $Y$  与  $X$  (从研究问题出发)
- step 2: 在总体中重复抽样，获得样本
- step 3: 对  $X$  "切片"，获得  $Y \mid X = x$  的 条件密度和条件分布
- step 4: 制作联合密度表格  $P(Y = y, X = x)$
- step 5: 计算边缘密度  $P(X = x)$
- step 6: 制作条件密度表格  $P(Y \mid X = x) = \frac{P(Y=y, X=x)}{P(X=x)}$
- step 7: 计算条件期望  $E(Y \mid X = x)$

## 经验研究为什么从LPF而不是CEF开始?

- 简洁且经济含义明确：线性CEF也是常见的线性回归模型(linear regression model)。其中一个原因是  $m(x_1, x_2)$  完整线性CEF为

$$m(x_1, x_2) = x_1\beta_1 + x_2\beta_2 + x_1^2\beta_3 + x_2^2\beta_4 + x_1x_2\beta_5 + \beta_6$$

CEF是具有好的预测性质，那么什么是“好”的准则？

定义损失函数(**loss function**)，表达为常用的二次型形式：

$$L(Y, g(x)) = (Y - g(x))^2$$

其中  $L(\cdot)$  是r.v.，取期望得**均值平方误差 (mean squared error, MSE)**，简称**均方误**

$$R(Y, g(x)) = E[L(Y, g(x))] = E[(Y - g(x))^2]$$

## 经验研究为什么从LPF而不是CEF开始?

- LPF是MSE最小的线性函数:

$$\beta = \underset{b}{\operatorname{argmin}} E \left[ (Y_i - X'_i b)^2 \right]$$

- 依据一阶条件:  $E[X_i(Y_i - X'_i b)] = 0$  得到  $b$  的最优解  $\beta = E[X_i X'_i]^{-1} E[X_i Y_i]$
- $X'_i \beta$  是  $Y_i$  在  $X_i$  上的最优线性投影 (best linear projection, BLP) , 向量  $\beta$  是线性投影系数 (linear projection coefficient)
- 根据一阶条件重新构建  $E[X_i(Y_i - X'_i \beta)] = 0$ , 也就是说  $Y$  的线性投影函数误差(linear projection function error, LPFE)  $e_i = Y_i - X'_i \beta$  与  $X_i$  不相关, 也就是说LPF具有  $E(X_i e_i) = 0$  (矩阵形式为  $E[Xe] = 0$ ) 的性质.
- **思考:** 与CEFE的性质比较

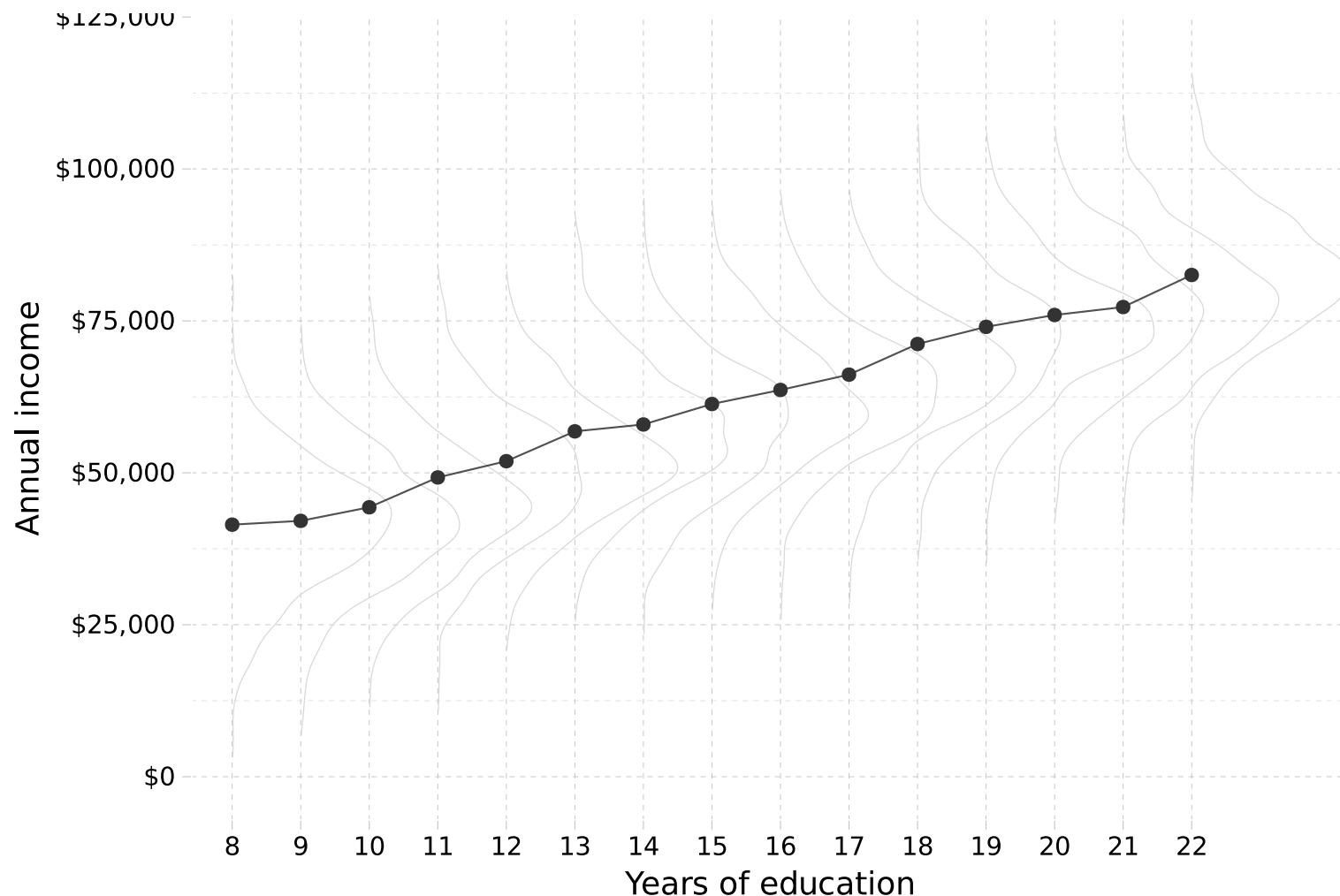
## 经验研究为什么从LPF而不是CEF开始?

- LPF是MMSE, CEF也是MMSE。继续使用最小化MSE准则:

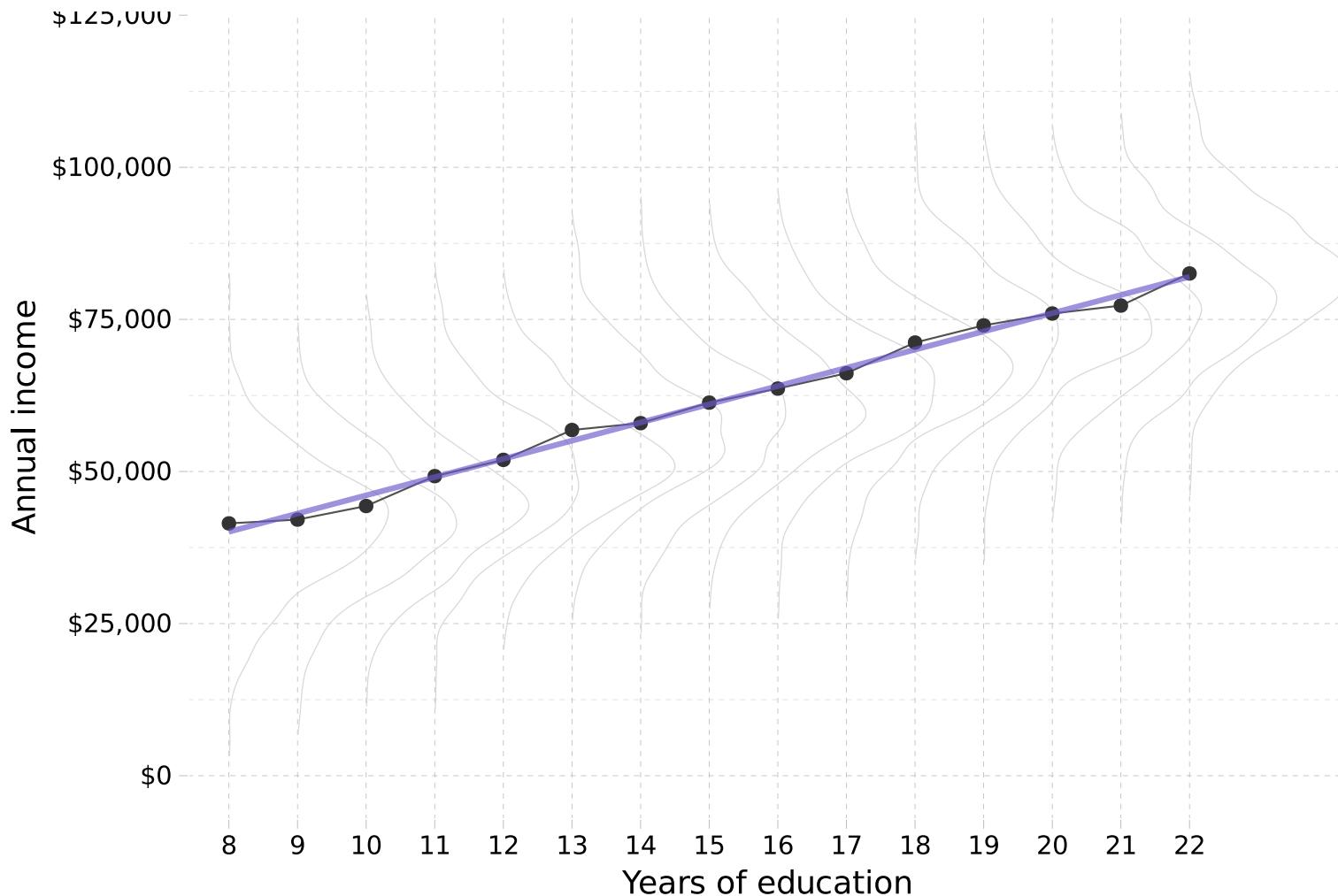
$$\beta = \underset{b}{\operatorname{argmin}} E \left[ (m(X_i) - X_i' b)^2 \right]$$

- 回归与条件期望函数定理 ( Regression-CEF Theorem)
- 结论:
  - LPF同样是CEF的MMSE和BLP
  - 通常而言, CEF不一定是线性的; 但CEF若是线性的, 那么LPF就是CEF

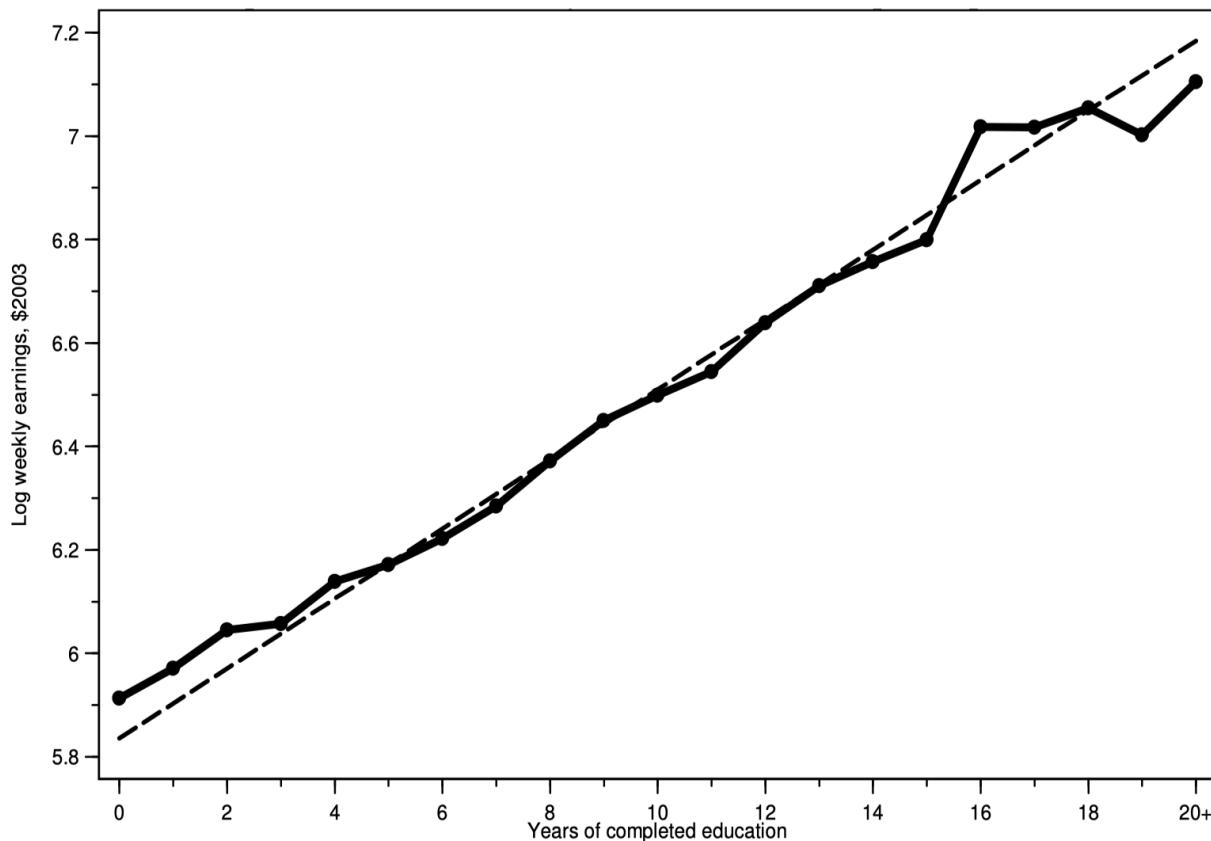
## CEF



## LPF去估计CEF



## 实际数据



© Princeton University Press. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Sample is limited to white men, age 40-49. Data is from Census IPUMS 1980, 5% sample.

Figure 3.1.2: Regression threads the CEF of average weekly wages given schooling

## 推断因果是基于CEF而不是LPF

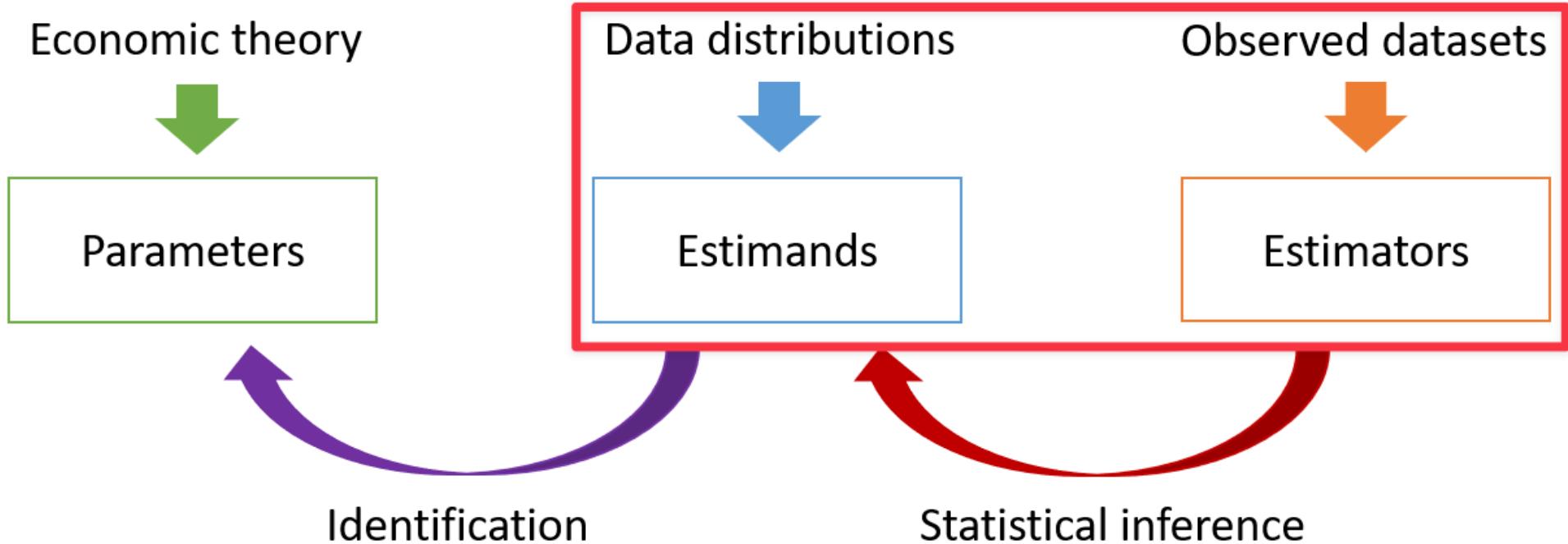
- 若CEF是相关关系  $\rightarrow$  LPF是相关的
- 若CEF是因果关系  $\rightarrow$  LPF是因果的
- 问题是：怎样获得一个因果的 CEF？(客观)  
 $\rightarrow$  必须依赖于理论认知(主观)

## 推断因果是基于CEF而不是LPF，但经验研究更多从LPF出发

- 实际上的做法是使用 LPF 进行建模(to see is to believe)
- 由于只有 线性CEF = LPF，即使使用了模型设定正确的LPF，一部分信息先天地进入到了干扰项  $e$ ，所以必须假定 干扰项条件均值独立于自变量，即  $E(e | X) = E(e) = c$ ，才保证LPF的估计系数距离线性CEF的真实值不远
- 即便统计推断可靠，但线性CEF出现概率仍旧很小，这就凸显了识别过程的重要性

# 从样本到事实：原材料的重要性

## 理论、总体分布与样本



## 为什么需要统计推断?

- 之前 重点从事实到模型, 关注了CEF与LPF, 它都是**总体意义的**
- 现在回到**样本**: 看着样本, 想着总体, 通过统计推断的方式进行
- 将LPF设定为:  $Y = X'\beta + e, \quad E(e | X) = 0$
- 展开:  $Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + e$

$$E(Y | X) = X'\beta$$

- 利用**最小二乘法**求解系数  $\hat{\beta}_{ols}$ , 就是最小化  $Y$  与线性投影预测值  $\hat{Y} = X'b$  的残差  $\varepsilon = Y - \hat{Y}$  的MSE

$$\hat{\beta}_{ols} = \operatorname{argmin}_b E \left[ (Y - X'b)^2 \right]$$

## 为什么需要统计推断?

- 由一阶条件可得:

$$E \left[ \mathbf{X} \left( Y - \mathbf{X}' \hat{\boldsymbol{\beta}} \right) \right] = 0$$

- 此条件同等与:

$$E \left[ \mathbf{X} \left( Y - \mathbf{X}' \hat{\boldsymbol{\beta}} \right) \right] = E[\mathbf{X}\varepsilon] = 0$$

- 为什么不是e?
- 由此可见, 最小二乘的本质就是通过样本求解系数  $\hat{\boldsymbol{\beta}}_{ols}$

$$\hat{\boldsymbol{\beta}}_{ols} = E[\mathbf{X}\mathbf{X}']^{-1} E[\mathbf{XY}]$$

## 为什么需要统计推断?

- 将LPF代入上式:

$$\begin{aligned}\hat{\beta}_{ols} &= E[\mathbf{X}\mathbf{X}']^{-1}E[\mathbf{XY}] = E[\mathbf{X}\mathbf{X}']^{-1}E[\mathbf{X}(\mathbf{X}'\boldsymbol{\beta} + e)] \\ &= \boldsymbol{\beta} + E[\mathbf{X}\mathbf{X}']^{-1}E[\mathbf{X}e]\end{aligned}$$

其中**由于假设**:  $E(e | X) = 0$

- 故  $E[\mathbf{X}e] = E_X[E(\mathbf{X}e | \mathbf{X})] = E_X[\mathbf{X}E(e | \mathbf{X})] = \mathbf{0}$
- 故  $\hat{\beta}_{ols} = \boldsymbol{\beta}$  以上讨论说明:  
最小二乘法估计系数  $\hat{\beta}_{ols}$  就是**LPF**系数, 同样也是**线性CEF**  $E(Y | X) = \mathbf{X}'\boldsymbol{\beta}$  的系数  $\boldsymbol{\beta}$
- **样本到事实的前提**
  - 假设 线性CEF  $\rightarrow$  允许LPF代表事实(CEF)
  - 假设  $E(e | X) = 0 \rightarrow$  对LPF使用OLS估计值可以得到总体真实值

## 为什么需要统计推断?

- 干扰项  $e$  包含了除  $X$  外的其他影响  $Y$  的因素, 与  $X$  是否相关无法检验  
→ 只能通过理论和经验判断
- 残差项  $\varepsilon$  是估计方法计算出来的, 总会与  $X$  正交

$$E \left[ \mathbf{X} \left( Y - \mathbf{X}' \hat{\boldsymbol{\beta}} \right) \right] = E[\mathbf{X}\varepsilon] = 0$$

- 最小二乘法只是估计方法的一种<sup>†</sup>, 常见的估计方法还有矩方法、最大似然估计等
- 总体  $\hat{\boldsymbol{\beta}}_{ols} = E[\mathbf{X}\mathbf{X}']^{-1} E[\mathbf{XY}]$
- 样本  $\hat{\boldsymbol{\beta}}_{ols} = (\sum_i X_i X'_i)^{-1} (\sum_i X_i Y_i)$

<sup>†</sup> 矩方法(method-of-moments)。根据大数定律和中心极限定理使用样本矩  $\frac{1}{n} \sum_i X_i X'_i$  估计总体矩  $E[X_i X'_i]$ 。还可以使用其他估计方法, e.g.  $Y_i$  给定  $X_i$  去最小化  $Y_i$  的MSE.

## 统计推断依赖的大样本性质

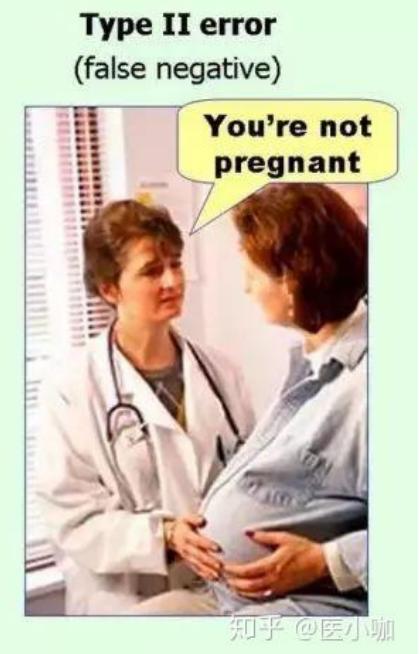
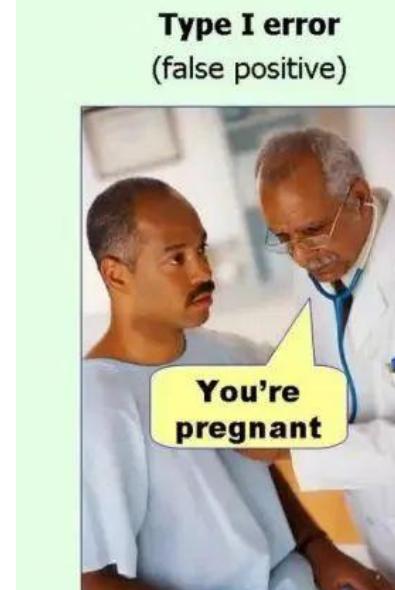
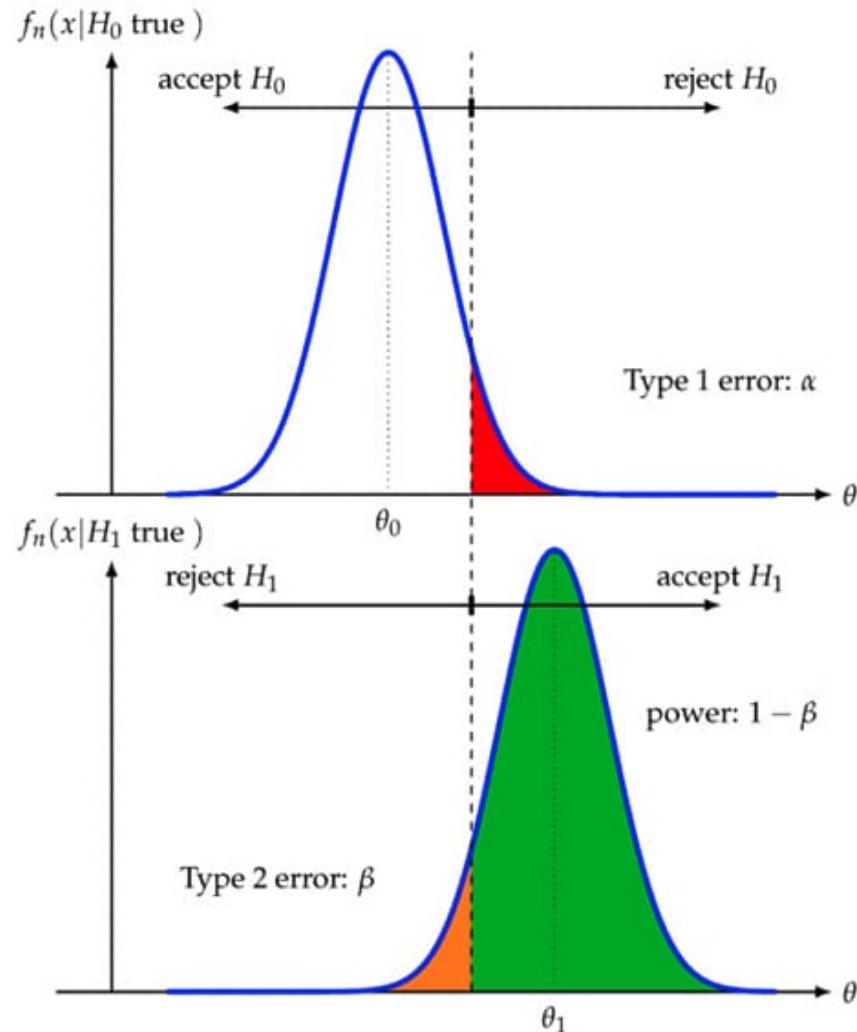
- 总体估计值  $\hat{\beta}$  是随机变量，因此具有分布(均值和方差)
- 样本量：在  $n > 200$ ，样本估计值  $\hat{\beta}$  是总体估计值(真实值)的  $\beta$  的一致估计( $\text{plim } \hat{\beta} = \beta$ )
  - 现代微观实证建立在大样本假定下，避免了传统的强假设(正态分布、自变量非随机、线性 CEF、同方差)
- 关注异方差 ← 社科研究的常态
  - 影响显著性。如果异方差问题严重，使标准误上升30% (极少情况会减少)
  - 修复异方差。使用异方差一致性标准误差或稳健标准误差，STATA中在回归后加上 `vce(robust)`，或增加聚类
  - 模型设定错误也会产生异方差。若CEF是非线性，而使用LPF产生异方差

$$E[(Y_i - X'_i \beta)^2 | X_i] = E\left[\left(\{Y_i - E[Y_i | X_i]\} + \{E[Y_i | X_i] - X'_i \beta\}\right)^2 | X_i\right] = Var(Y_i | X_i) + (E[Y_i | X_i] - X'_i \beta)^2$$

即使  $Var(Y_i | X_i)$  是常数，第二项导致异方差

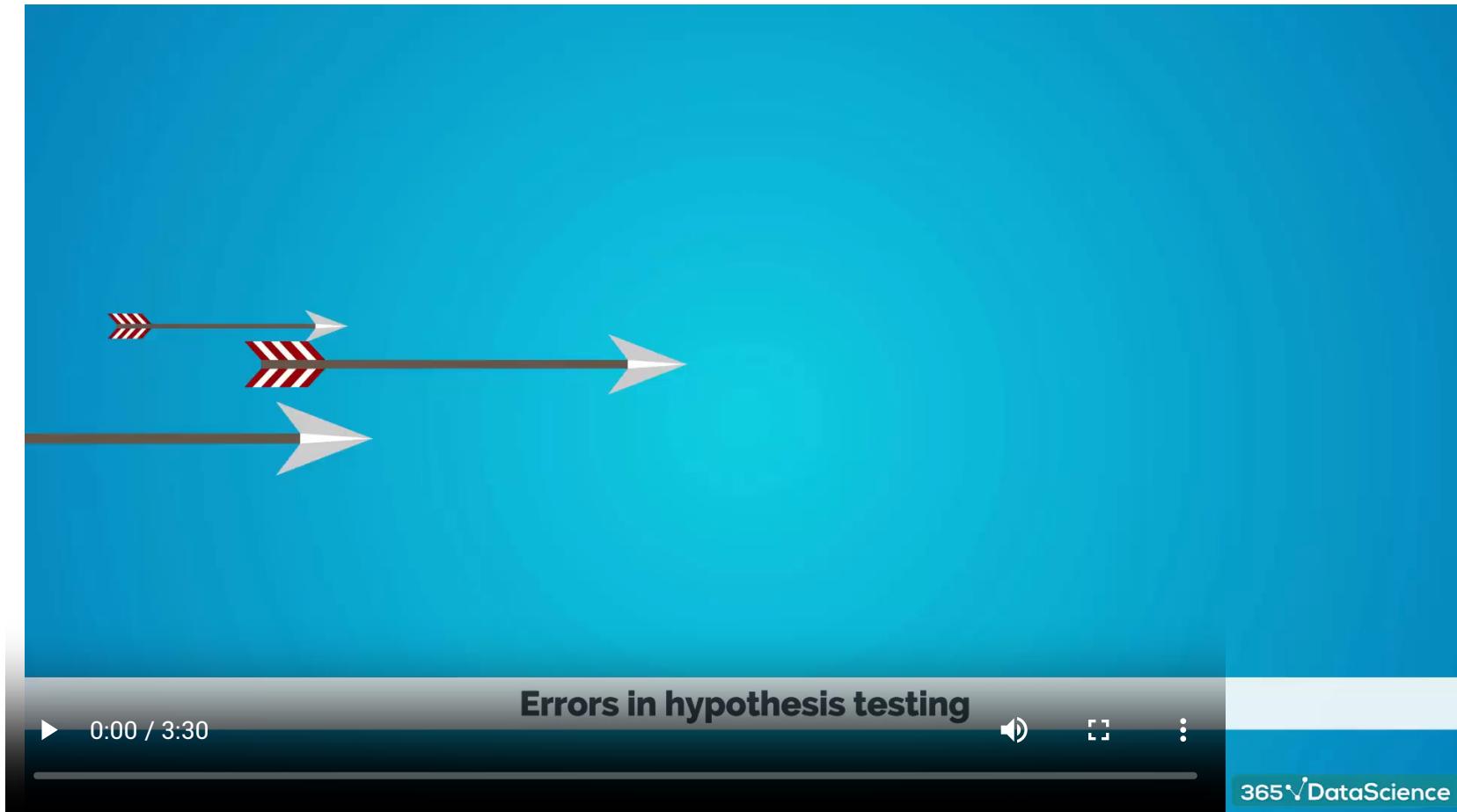
## 统计推断依赖的大样本性质

- 标准差(SD)是总体估计值(是随机变量)分布(未知但存在)的标准差；标准误(SE，报告)是样本估计值的均值的分布
- 使用**样本估计值(也是随机变量)**的分布均值来推断总体估计值  
→ 样本估计值的宽，导致抽样的误差就越大
  - 增加的统计的power, >80%
  - 统计软件中会报告SE



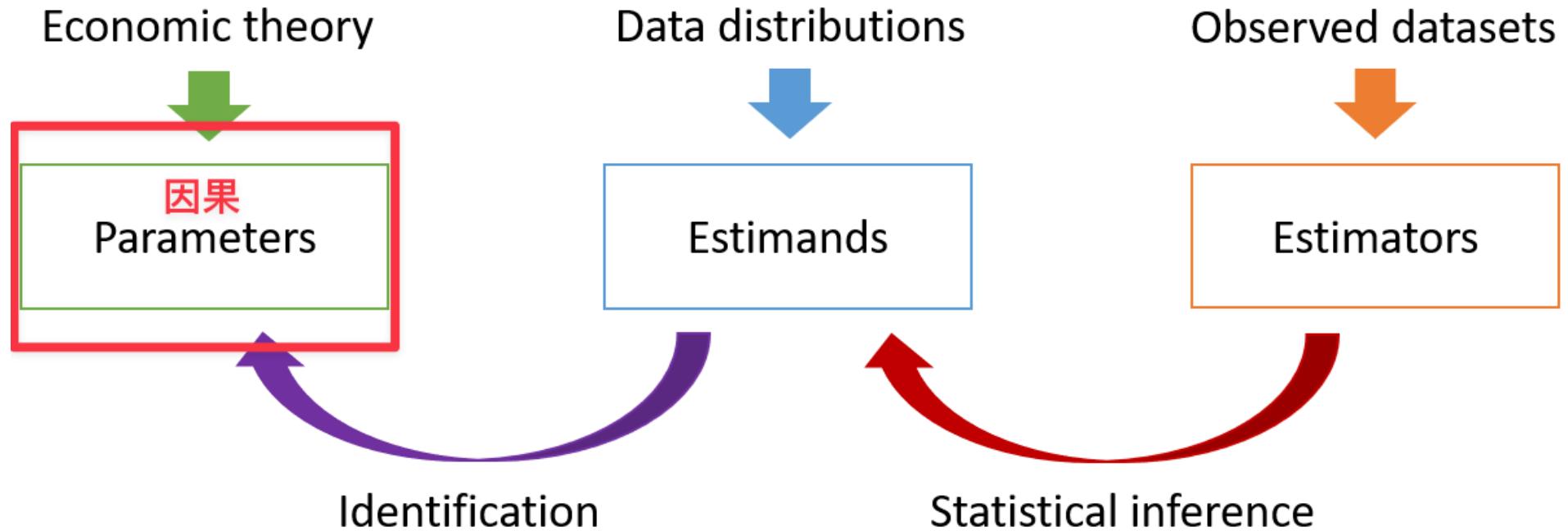
Type I error (significance level, P-value)、statistical power(sensitivity)、expected effect size、sample size

## Type I/II error



# 潜在结果框架下的因果模型

## 理论、总体分布与样本



## 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 (counterfactual)** 概念

→ 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)

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

- 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

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

## 选择偏误 selection bias

- 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. 非随机因素导致接受干预
  2. 若这个非随机因素是**个体的自我选择**, 其造成的估计偏误就是**选择偏误** (selection bias)
  3. 目标, 使得选择偏误  $\rightarrow 0$

## 例子：吃药对健康的影响

个体 <i>i</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>i</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	8	3	5	0	3

- “上帝”视角
- 阴影部分为可观测到的结果，而下划线部分为无法观测到的**反事实结果**

## 例子：吃药对健康的影响

- 干预组:  $T1 = E[Y_i(1) | D_i = 1]$ ;  $T0 = E[Y_i(0) | D_i = 1]$  (反事实)
- 控制组:  $C0 = E[Y_i(0) | D_i = 0]$ ;  $C1 = E[Y_i(1) | D_i = 0]$  (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		
$T1 = E[Y_i(1)   D_i = 1]$ = 5.3	$T0 = E[Y_i(0)   D_i = 1]$ = 2 (反事实结果)	$D_i = 1$ (处置组)	$T1 = E[Y_i   D_i = 1]$ = $E[Y_i(1)   D_i = 1]$ = 5.3
$C1 = E[Y_i(1)   D_i = 0]$ = 5.5 (反事实结果)	$C0 = E[Y_i(0)   D_i = 0]$ = 2.5	$D_i = 0$ (控制组)	$C0 = E[Y_i   D_i = 1]$ = $E[Y_i(0)   D_i = 0]$ = 2.5

## 例子：吃药对健康的影响

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) =  $\omega \times ATT + (1 - \omega) \times ATU = 3.18$

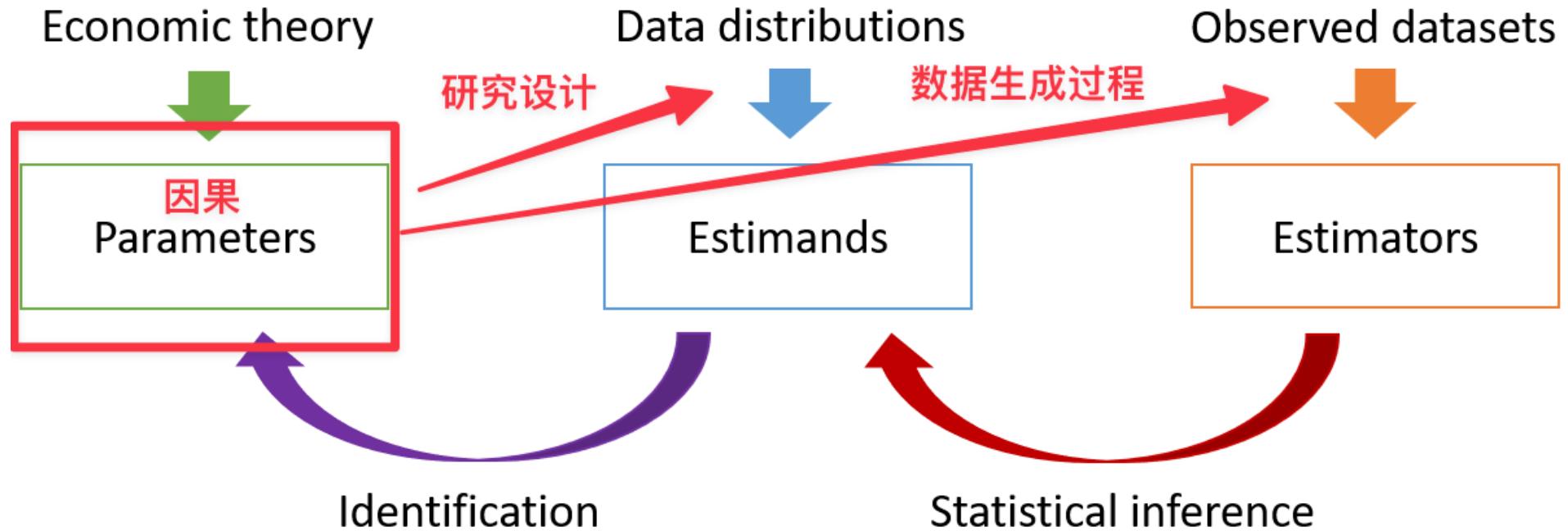
但在实际情况中，无法观测到反事实结果

- 存在偏误的“朴素”估计量 =  $T1 - C0 = 2.8$
- 存在偏误的ATT =  $T0 - C0 = -0.5$
- 存在偏误的ATU =  $T1 - C1 = -0.2$
- 存在偏误的ATE =  $\omega \times (T0 - C0) + (1 - \omega) \times (T1 - C1) = -0.38$
- 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"估计量又会存在选择偏误，那么如何通过观测数据识别处置效应？

**回答：**通过研究设计

## 理论、总体分布与样本



## 研究设计：随机实验

- 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  $\tau_i$  的函数

$$Y_i(0) = a + bX_i + e_i, D_i = 0$$

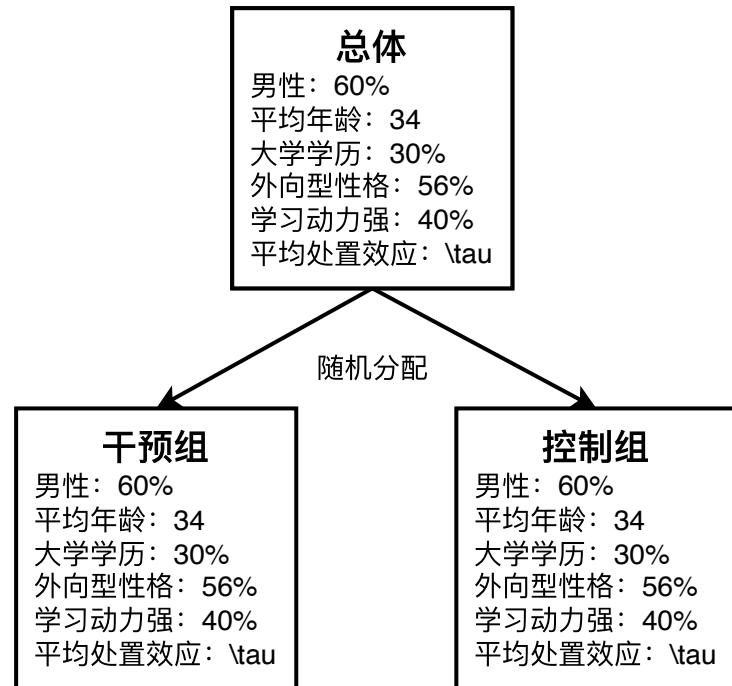
$$Y_i(1) = a + \tau_i + bX_i + e_i, D_i = 1$$

$$(X_i, e_i, \tau_i) \perp D_i$$

- 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

## 研究设计：随机实验

- 



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)

- 独立性的第1个维度: 未受干预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:  $E[Y_i(0) \mid D_i] = E[Y_i(0)]$
- 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)

- 独立性的第2个维度: 接受干预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

$$E[Y_i(1) | D_i = 1] = E[Y_i(1) | D_i = 0]$$

- 同理:  $E [Y_i(1) | D_i] = E [Y_i(1)]$
- 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ATT无偏

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

## 研究设计：类似RCT的回归

- RCT实验昂贵
- 以人为实验对象会受伦理审查委员会严格审查
- 那么当不是RCT时，是否也可以使用"朴素"估计量呢？
- **回答：**可以，但需要施加**额外假设**。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

## 研究设计：CMI假设下，控制可观测特征 + 回归 → 消除选择偏误

- 药物与健康的例子
  - 服药个体普遍年龄偏大，且年龄大的个体普遍的潜在健康状况差 → 年龄因素与健康 Y负相关
  - 对干预组和控制组的年龄进行分类，相同年龄段来比较用药前后的健康状况的差异（同年龄段内，干预组和控制组可以看成随机分配，满足潜在结果独立性假设）

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

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

## 研究设计：CMI假设下，控制可观测特征 + 回归 → 消除选择偏误

- 给定可观测特征条件  $X_i = x$  的干预组和控制组

$$ATT(x) = T1(x) - C0(x)$$

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

- 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)**

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

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

- 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)**

## 若我们关心总体：从CMI到条件独立假设 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:

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

- 选择偏误  $= E[Y_i(0) | X_i, D_i = 1] - E[Y_i(0) | X_i, D_i = 0]$   
 $= E[Y_i(0) | X_i] - E[Y_i(0) | X_i]$   
 $= 0$

## 若我们关心总体：从CMI到条件独立假设 CIA

- CIA的意思是：控制协变量  $X_i$  后，干预措施就像 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

$$\begin{aligned} & E[\textcolor{blue}{Y}_i \mid X_i, \textcolor{red}{D}_i = 1] - E[\textcolor{blue}{Y}_i \mid X_i, \textcolor{red}{D}_i = 0] \\ &= E[\textcolor{blue}{Y}_i(1) \mid X_i] - E[\textcolor{blue}{Y}_i(0) \mid X_i] \\ &= E[\textcolor{blue}{Y}_i(1) - \textcolor{blue}{Y}_i(0) \mid X_i] \end{aligned}$$

## 若我们关心剂量多少而不是是否干预：扩展的CIA

### 继续考虑：教育回报率的例子

- 现在，将CIA拓展到**多值**干预变量的情况，如**接受教育年限** ( $s_i$ ) 取值为整数  $t \in \{0, 1, \dots, T\}$ 。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  $i$  是否接受教育的潜在收入  $\rightarrow Y_{si}$  是个体  $i$  接受  $s$  年教育后的潜在收入，函数  $f_i(s)$  告诉我们：即使个体  $i$  接受  $s$  的潜在收入是因人而异的（符合理论） $\rightarrow f_i(s)$  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，因为两个人即使接受相同的教育年限，但潜在的收入也可能是不同的

## 若我们关心剂量多少而不是是否干预：扩展的CIA

- 将 CIA 扩展到多值干预变量
- 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：

$$Y_{si} \perp\!\!\!\perp s_i \mid X_i \text{ 对于 } s \text{ 的每个取值}$$

- 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  $s = s_i$  取定每个人接受的教育年限时的  $f_i(s_i)$
- 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  $s$  的取值都成立 → 该假设可能**很强**，因为多接受小学1年和大学1年很可能不同

## 若我们关心剂量多少而不是是否干预：扩展的CIA

在CIA下，给定  $X_i$ , 潜在结果  $\mathbf{Y}_{si}$  和每个人的干预剂量多少  $s_i$  是独立的：

$$\begin{aligned} & E[\mathbf{Y}_i \mid X_i, s_i = s] - E[\mathbf{Y}_i \mid X_i, s_i = s - 1] \\ &= E[f_i(s_i) \mid X_i, s_i = s] - E[f_i(s_i) \mid X_i, s_i = s - 1] \\ &= E[f_i(s) \mid X_i, s_i = s] - E[f_i(s - 1) \mid X_i, s_i = s - 1] \\ &= E[\mathbf{Y}_{si} \mid X_i, s_i = s] - E[\mathbf{Y}_{(s-1)i} \mid X_i, s_i = s - 1] \end{aligned}$$

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

- CIA下  $\rightarrow$  控制条件后，相差1年教育的人的收入均值的差异就可以解释为 **多接受1年教育的平均处置效果**

## 若我们关心剂量多少而不是是否干预：扩展的CIA

**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

$$\begin{aligned} & E[Y_i | X_i, s_i = 12] - E[Y_i | X_i, s_i = 11] \\ &= E[f_i(12) | X_i, s_i = 12] - E[f_i(11) | X_i, s_i = 11] \\ &= E[f_i(12) | X_i, s_i = 12] - E[f_i(11) | X_i, s_i = 12] \quad (\text{CIA}) \\ &= E[f_i(12) - f_i(11) | X_i, s_i = 12] \\ &= \text{给定 } X_i \text{ 下，已高中毕业学生因高中毕业带来的平均处置效应} \\ &= E[f_i(12) - f_i(11) | X_i] \quad (\text{再次CIA}) \\ &= \text{给定 } X_i \text{ 下，高中是否毕业（为条件）的平均处置效应} \end{aligned}$$

## 若我们关心剂量多少而不是是否干预：扩展的CIA（从多条件到无条件）

- 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  $X_i$  有多少个条件取值就可能会存在多少处置效果
- 就刚才例子，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：
- $E[f_i(\textcolor{red}{12}) - f_i(\textcolor{red}{11}) | \text{Sex} = 1, \text{Age} = 20\text{至}30]$  表示20至30岁的女性，高中毕业比高中肄业的平均教育回报水平。
- $E[f_i(\textcolor{red}{12}) - f_i(\textcolor{red}{11}) | \text{Sex} = 0, \text{Age} = 65\text{岁以上}]$  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
- 能否获得高中毕业比高中肄业的平均教育回报水平呢？

## 若我们关心剂量多少而不是是否干预：扩展的CIA（从多条件到无条件）

问题 那么无条件高中毕业相对于高中肄业的平均处置效应是什么？

回答 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...

$$E[\textcolor{blue}{Y}_i \mid X_i, s_i = 12] - E[\textcolor{blue}{Y}_i \mid X_i, s_i = 11] = E[f_i(12) - f_i(11) \mid X_i]$$

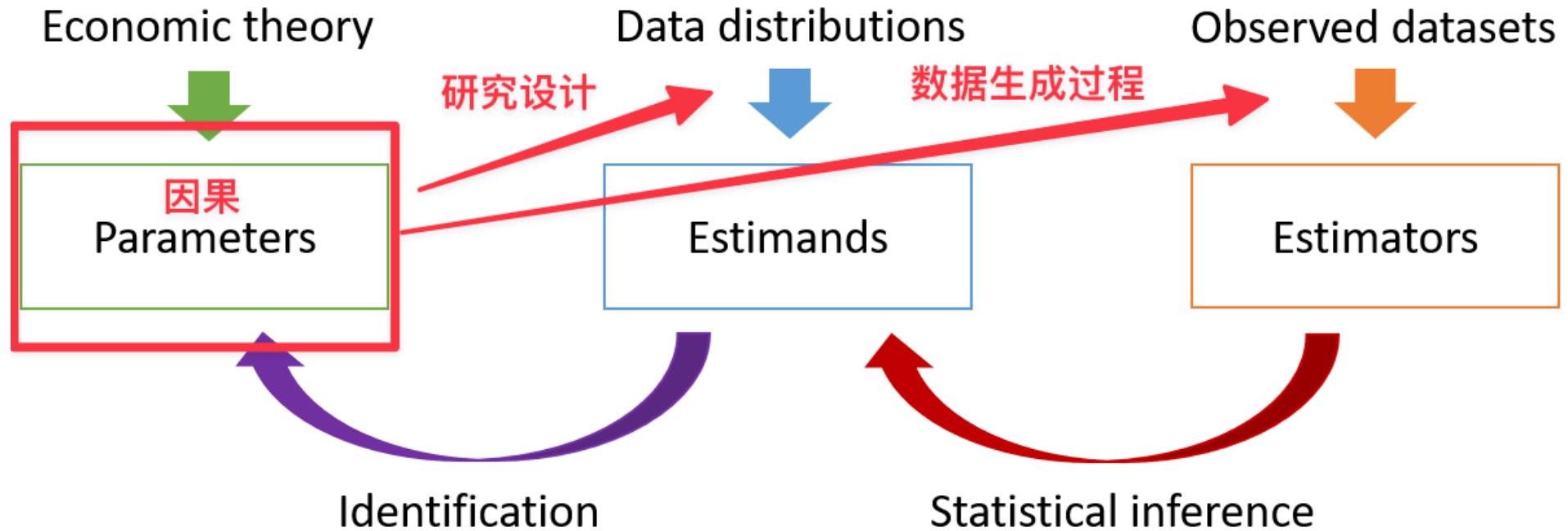
现在取两边的期望值并应用迭代期望法则(LIE)

$$E_X \left( E[\textcolor{blue}{Y}_i \mid X_i, s_i = 12] - E[\textcolor{blue}{Y}_i \mid X_i, s_i = 11] \right)$$

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

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

## 理论、总体分布与样本



## 基于RCT理念的LPF得到因果效应还需要CIA

- RCT的研究设计 → LPF (总体) → 能够得到因果效应么? → CIA
- 别忽略**线性CEF**的假设!
- **LPF模型设置的假设**: 线性、**同质的**的LPF可以刻画**线性CEF**:

$$f_i(\textcolor{red}{s}) = \alpha + \tau s + \eta_i \quad (\text{A})$$

- 为什么 (A) 式是LPF? → 个体  $i$  在  $s$  的任意取值下潜在收入, 而不是依据  $s_i$  观测值, 所以省略了  $s$  的下标  $i$
- (A) 假设在  $f_i(s)$  中唯一因人而异的部分是干扰项  $\eta_i$  的无条件均值为 0 (回想CEF) 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  $\textcolor{blue}{Y}_i$  代入模型, 就得到了**可回归的模型**:

$$\textcolor{blue}{Y}_i = \alpha + \tau \textcolor{red}{s}_i + \eta_i \quad (\text{B})$$

- (A) 式中  $\tau$  是**真实处置效应**, 而 (B) 式  $\tau$  的**样本估计值**  $\hat{\tau}$  通常会因为  $s_i$  的**样本选择问题**产生偏误
- 文章的**研究设计**中说明如何能得到真实处置效应(内生性问题必须说明)

## 基于RCT理念的LPF得到因果效应还需要CIA

- CIA下，意味着加入多个**可观察**的协变量  $X_i$ ，能够排除潜在的**干扰因素**
- 将潜在收入  $f_i(s)$  的干扰项结构化为**可观察因素**  $X_i$  (因人而异)和**残差项**  $v_i$  的线性函数：

$$\eta_i = X'_i \beta + v_i \quad (\text{C})$$

- $\beta$  是  $\eta_i$  对  $X_i$  回归的**总体系数向量**(意味着可以通过最小二乘估计获得正确的系数估计)，所以有：
  1.  $E[\eta_i | X_i] = X'_i \beta$
  2. 残差项  $v_i$  与  $X_i$  不相关

## 基于RCT理念的LPF得到因果效应还需要CIA

根据CIA可以得到：

$$E[f_i(\textcolor{red}{s}) \mid X_i, \textcolor{red}{s}_i]$$

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

$$= E[\alpha + \tau \textcolor{red}{s}_i + \eta_i \mid X_i] \quad (\text{代入B式})$$

$$= \alpha + \tau \textcolor{red}{s}_i + E[\eta_i \mid X_i]$$

$$= \alpha + \tau \textcolor{red}{s}_i + X'_i \beta \quad (\text{最小二乘回归方程})$$

- **回忆** 这里再次使用到，若  $f_i(\textcolor{red}{s})$  所代表的CEF是线性的(倒数第2行)，则意味着"**正确地模型设定为LPF<sup>†</sup>**" 可以正确代表线性CEF

<sup>†</sup> 这里"正确"是指若控制了  $X_i$  就像RCT一样。

## 基于RCT理念的LPF得到因果效应还需要CIA

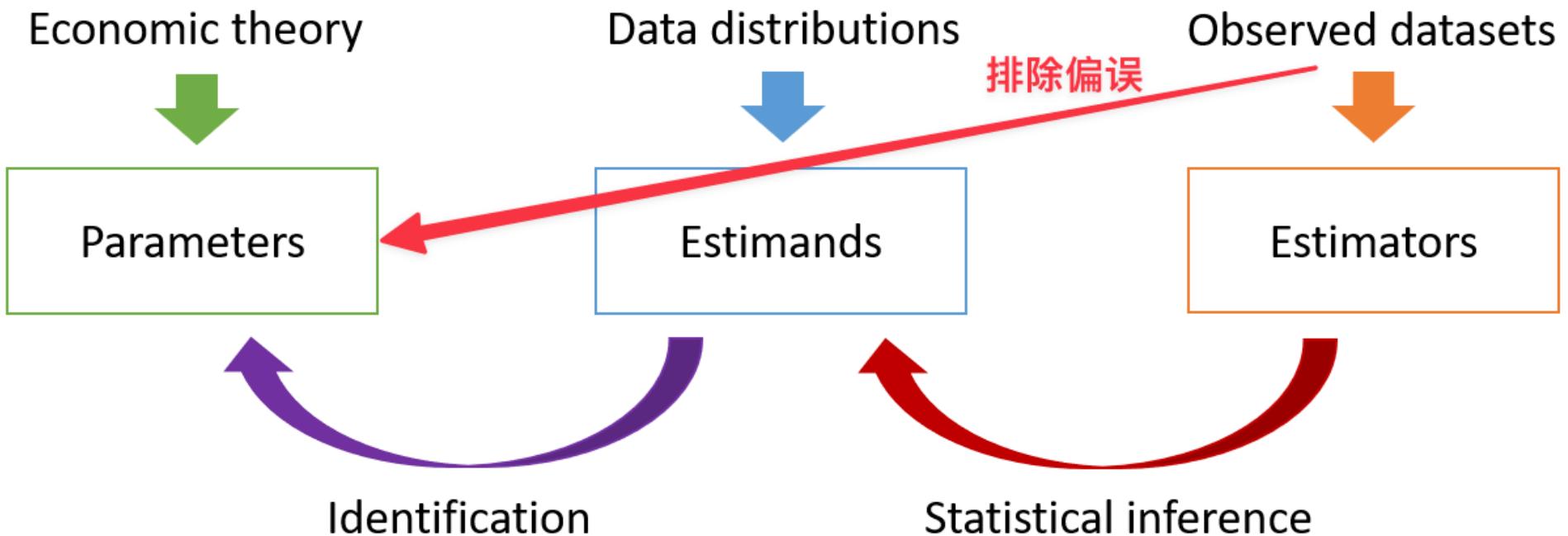
所以把可回归的模型形式设定如下：

$$Y_i = \alpha + \tau s_i + X'_i \beta + \nu_i$$

- 通过假设，限制扰动项  $\nu_i$  来估计真正的因果效应  $\tau$ 
  1.  $s_i$  (根据 CIA)
  2.  $X_i$  (根据定义  $\beta$  是  $\eta$  对  $X_i$  回归的总体系数向量)

从样本出发寻找“真相”

## 理论、总体分布与样本



## 理解"朴素"估计值与LPF<sup>ols</sup>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:

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

- 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

↑ 基于SUTVA假设可以使第三项为0

## "朴素"估计值 + 控制变量 + CIA $\rightarrow$ 因果效应

现在我们已经知道在:  $E[Y_i(0) | D_i = 1] \neq E[Y_i(0) | D_i = 0]$  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:  $Y_i = \underbrace{E[Y_i(0)]}_a + \underbrace{[Y_i(1) - Y_i(0)] \times D_i}_\tau + \underbrace{Y_i(0) - E[Y_i(0)]}_{u_i}$

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:  $E(Y_i | D_i, X_i) = \alpha + D_i + \beta X_i + E[e_i | D_i, X_i]$

"朴素"估计值+控制变量+假设  $\rightarrow$  LPF<sup>ols</sup>  $\rightarrow$  有因果理论支撑的线性CEF

- 与CIA思路一样，若要使得条件期望函数的  $D_i$  的系数等于  $\tau$ ，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：

$$E[e_i | D_i, X_i] = E[e_i | X_i]$$

- 可证明：**(建立在LPF<sup>ols</sup>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** ( $E[Y_i(0) | D_i = 1] = E[Y_i(0) | D_i = 0]$ ) 是等价的
- 这个条件使得LPF<sup>ols</sup>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  $\tau$  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）

## SUTVA

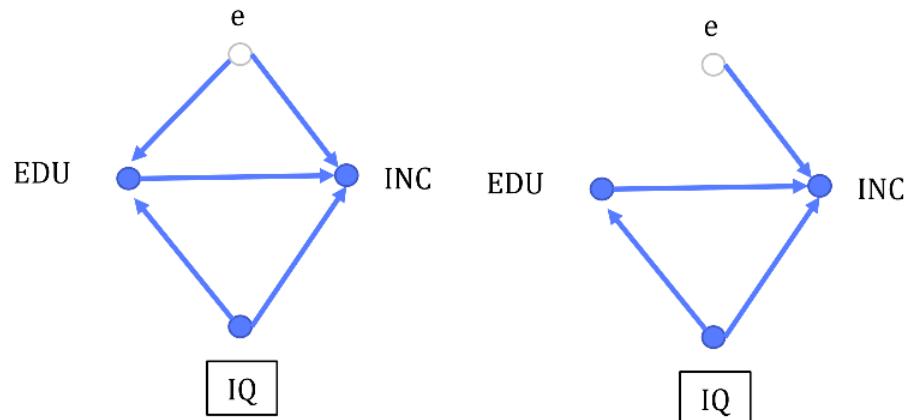
- 在之前的例子中，都是假设个体处置效应是相同的，即  $\tau_i = \tau$
- **稳定个体干预值假设（The Stable Unit Treatment Value Assumption, SUTVA）：**  
    简单说，每个个体的潜在结果不依赖于其他个体的干预状态。有两层含义：1. 不同个体的潜在结果之间不会有交互影响。2. 干预水平对所有个体都是相同的
- 第1个含义：它排除了**外部性或整体均衡效应**
  - 例：研究班级规模对个体学习效果的影响，同学之间往往存在外部性，如果班级里好学生多，相互讨论、相互促进，产生正外部性，从而提高了整体学习效率
  - 例：如果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

从总体理解，从样本着手：偏误类型与解决办法

# 有向无环图

- 干扰项**条件均值独立于**(CMI)自变量
- 教育回报率**
- LPF设定为：

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



- 有向无环图**(directed acyclic graph)
- 实心-可观测；空心-不可观测；单向箭头-因果关系；无法递归
- $EDU \rightarrow INC$  直接因果路径
- $EDU \leftarrow IQ \rightarrow INC$  混淆路径1
- $EDU \leftarrow e \rightarrow INC$  混淆路径2
- 右图：干扰项条件均值独立于自变量
- 左图：即便控制了IQ无效

## 偏回归系数

- 对于LPF:  $INC = \alpha + \beta_1 EDU + \beta_2 IQ + e$
- 两边取条件期望, 得到对应的线性CEF:

$$\begin{aligned} E(INC | EDU, IQ) \\ = \alpha + \beta_1 EDU + \beta_2 IQ + E(e | EDU, IQ) \\ = \alpha + \beta_1 EDU + \beta_2 IQ \end{aligned}$$

- 将线性CEF对EDU求偏导: **偏回归系数**

$$\frac{dE(INC | EDU, IQ)}{dEDU} = \beta_1$$

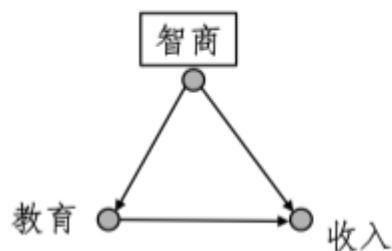
$\beta_1$  表示在IQ固定不变, INC 的期望值 (均值) 随 EDU 如何变化

## 有向无环图表达偏误类型

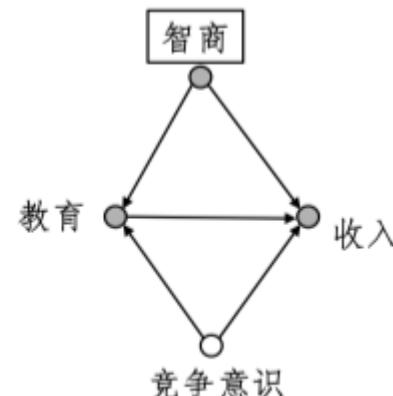
- 因果路径
- 混淆路径:  $A \leftarrow B \rightarrow C$ , B是A和C的混淆变量; 混淆变量会产生相关关系  $\rightarrow$  必须控制
- 对撞路径:  $A \rightarrow B \leftarrow C$ , B是A和C的对撞变量; 对撞变量不会产生相关性, 但若错误地控制了就会产生相关关系
- 估计X与Y的因果关系的本质是找到二者间所有的因果路径, 同时排除二者间所有非因果关系路径

## 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

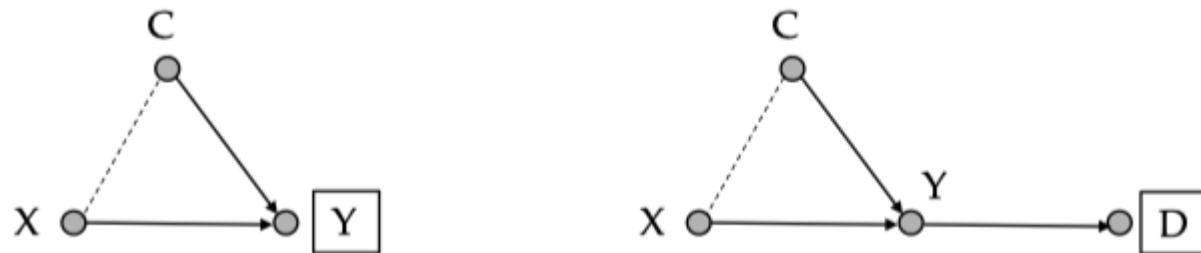


图：过度控制偏差

- 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

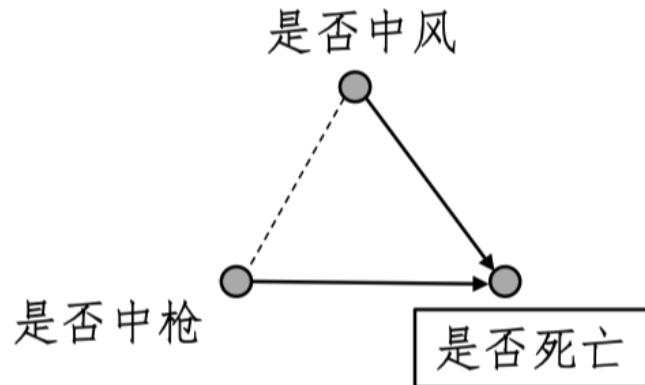
## 对撞偏误

- 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

## 偏误的解决办法

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

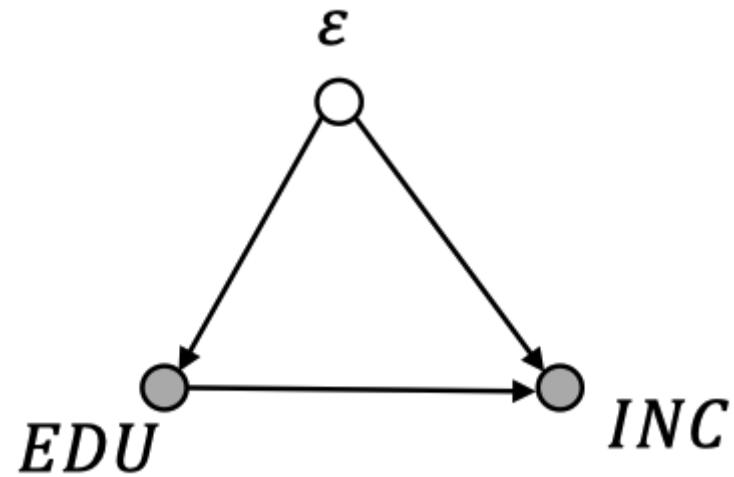
### 教育回报率的例子

- 假设最初只是通过建立如下LPF<sup>0</sup>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  $\varepsilon_{it}$ ，导致  $\varepsilon_{it}$ ，所以无法识别因果影响系数  $\beta$ 。

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



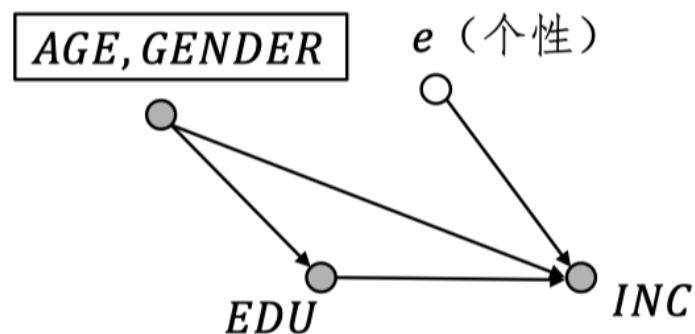
图：LPF等价于在DAC中忽略了年龄和性别

## 例子：解决办法

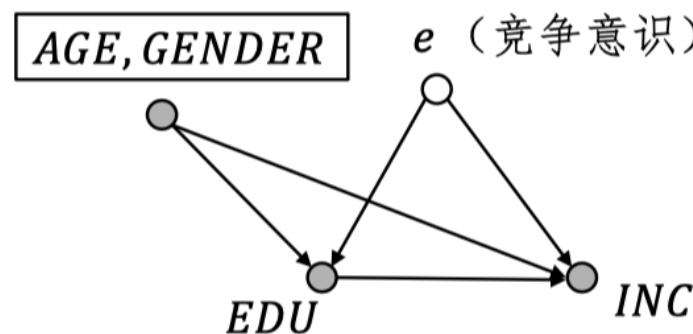
- 可以将  $\varepsilon_{it}$  中的可观测变量分离进行控制，得到 LPF<sup>1</sup>

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



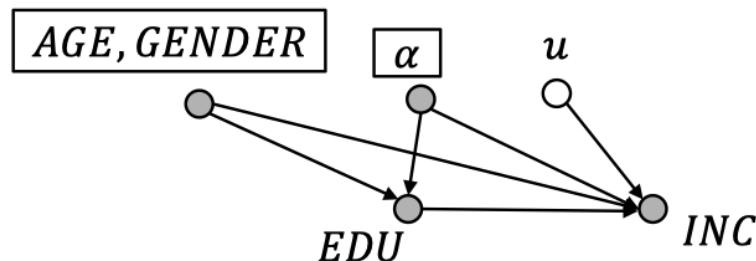
图：控制年龄和性别且  $e$  为混淆变量的情况

## 例子：解决办法

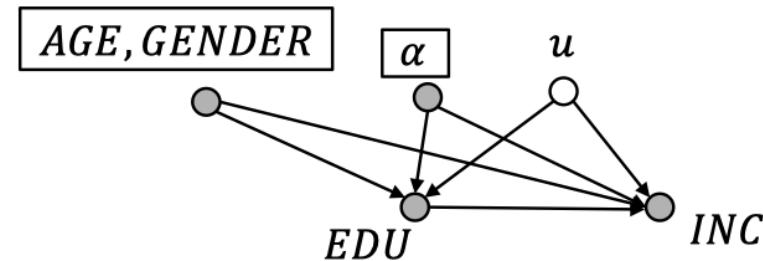
- 倘若真实关系是右图，仍无法识别因果关系
- 将  $\varepsilon_{it}$  进一步分解为：不可观测的非时变变量  $\alpha_i$  和不可观测的时变变量  $u_{it}$ ，即  
$$e_{it} = \alpha_i + u_{it}$$

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

- 如果混淆路径是  $\alpha_i$  造成的，我们希望控制  $\alpha_i$  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。



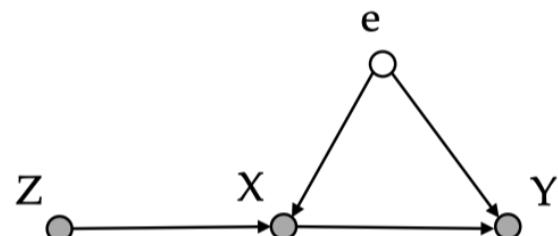
图：控制年龄、性别和 $\alpha$   
且 $u$ 为无关变量的情况



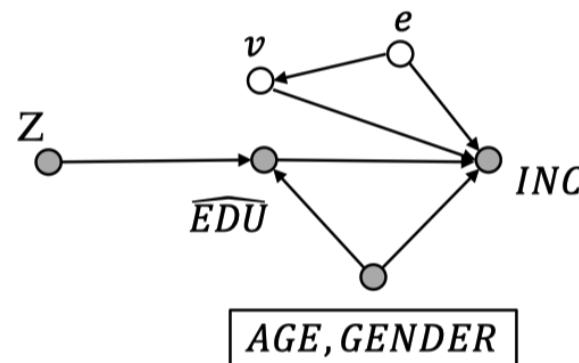
图：控制年龄、性别和 $\alpha$   
且 $u$ 为混淆变量的情况

## 例子：解决办法

- 实际上，大部分社会研究不太能轻易地在  $\varepsilon_{it}$  中分离出不可观测和可观测变量
- 引入工具变量  $Z_i$  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



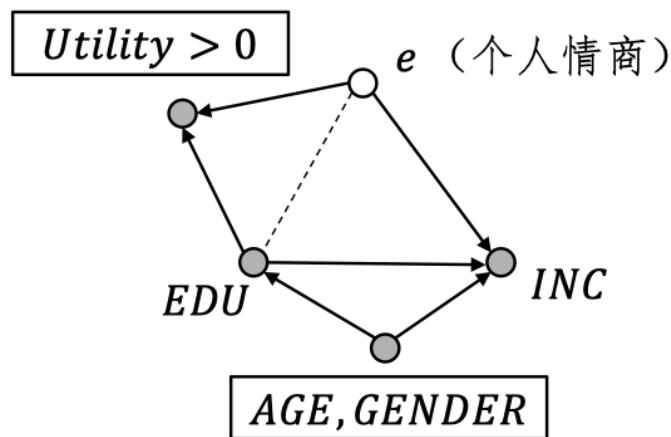
图：工具变量的相关性和外生性



图：引入工具变量的情况

## 例子：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