体育经济分析: 原理与应用

单元4: 体育中的相关或因果

周正卿

26 April 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 + 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。
  - 资源是有限的，而干预措施是要花钱的。

例子：球队胜率与上座率

假设联盟经理想知道哪些因素对球队上座率最重要。

- 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
- 通常我们认为**期望**球队获胜越多，观众就越多。用上座率ATT作为因变量，获胜率WPCT作自变量。之所以ATT是因变量，因为它取决于胜率。WPCT作自变量，是我们可以**操纵的变量**，往往不依赖于影响ATT中的其他变量： $ATT = f(WPCT)$
- 根据常识，我们可以**理论假设** $\frac{\partial ATT}{\partial WPCT} > 0$
- **检验**上述观点的最直接方式是**散点图**。但这样无法指导当胜率从0.5提高到0.51时，上座率会增加多少。提高胜率通常意味着要花钱买球员。
- 为了样本数据结合，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 $\hat{\tau}$ ，然而我们感兴趣的是 τ , 这两者什么关系?
  - 举个例子：进校门要人脸识别，我们通过闸机的时候，它用我们照片去跟照片库里的我们去匹配。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 τ ，这个过程就是 **识别（identification）**
  - 这个例子中，我们比较好命，至少有个照片库，但是现实中 τ 往往是不知道的或者说它存在但是看不到，我们只能通过**理论建构**
- 现在我们有 $\hat{\tau}$ ，它只少要存在两个方面的意义，统计学（现实）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，需要看**效应值**与**经济学解释**
- 若是 τ 具有**因果性**，上述过程是**因果推断**

例子：球队胜率与上座率

$$\text{ATT} = -2.4536 + 65.23 * \text{WPCT}$$

- 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%的上座率波动)，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针对性。
  - 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 τ 是**偏误**的
- 我们将可回归的线性表达式改为 $ATT_i = \boldsymbol{X}\beta + \tau \times WPCT_i + \varepsilon_i$
  - 市场规模与联赛差异进入了 $\boldsymbol{X}\beta$

例子：球队胜率与上座率

ATT=-.421+1.174POP-2.014 LEAGUE + 59.31*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

遗漏变量偏误与控制变量

增加控制变量可以提高回归估计值的可信度，可以让系数趋向于因果，但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：

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

- 其中, α, τ, γ 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 τ 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：

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

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, δ_{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	0.131	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† 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

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

- 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。[†]
- **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[†]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人。**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

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

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

- 那么可以说这个线性模型存在内生性问题。
  - 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: $E(e \mid EDU, IQ) = 0, \mathrm{Cov}(EDU, IQ) \neq 0$

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

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

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 β_1 。这时, 对于 β_1 的 OLS 估计为

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

- 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 σ_x^2/σ_v^2 **信噪比**有关

来源二: 测量误差

(2) 因变量存在测量误差

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

当因变量 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

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

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

- 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。

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

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

- 将式（2）代入式（1）中,可以得到

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

来源三: 互为因果

- 由式（3）

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

- 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

如何去找“真相”

- 大部分时候我们存着立场、偏见、固念和坐标。可是我们还能不能用别人的观点帮助自己了解这个世界?
- 看见为了相信 to see is to believe
- 真实世界，绝大多少只是看到我们相信。but you see what you believe
- 而且你只相信你希望的得到的真相 and you believe what you hope to be true
- p → q == -q → -p
- p → q != q → p
- （A & B & C & D）→ 显著 == 不显著 → （-A ^ -B ^ -C ^ -D）
- 怎么办？→ 去猜
- 直接猜 P(A) 很困难，所以去猜 P(A|B)=P(A&B)/P(B)
- 这时候就需要看案例的背景资料和文献，去寻找更多信息，来发现潜在B的可能性，去减少犯错的可能性

条件分布 (Conditional Distribution)

- 通常关心两个变量的关系
- 观察**条件期望**是最直接、简单办法
- 一般假设我们最感兴趣的 Y 与 X 是随机变项（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、

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

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

条件密度函数

- 离散形式：

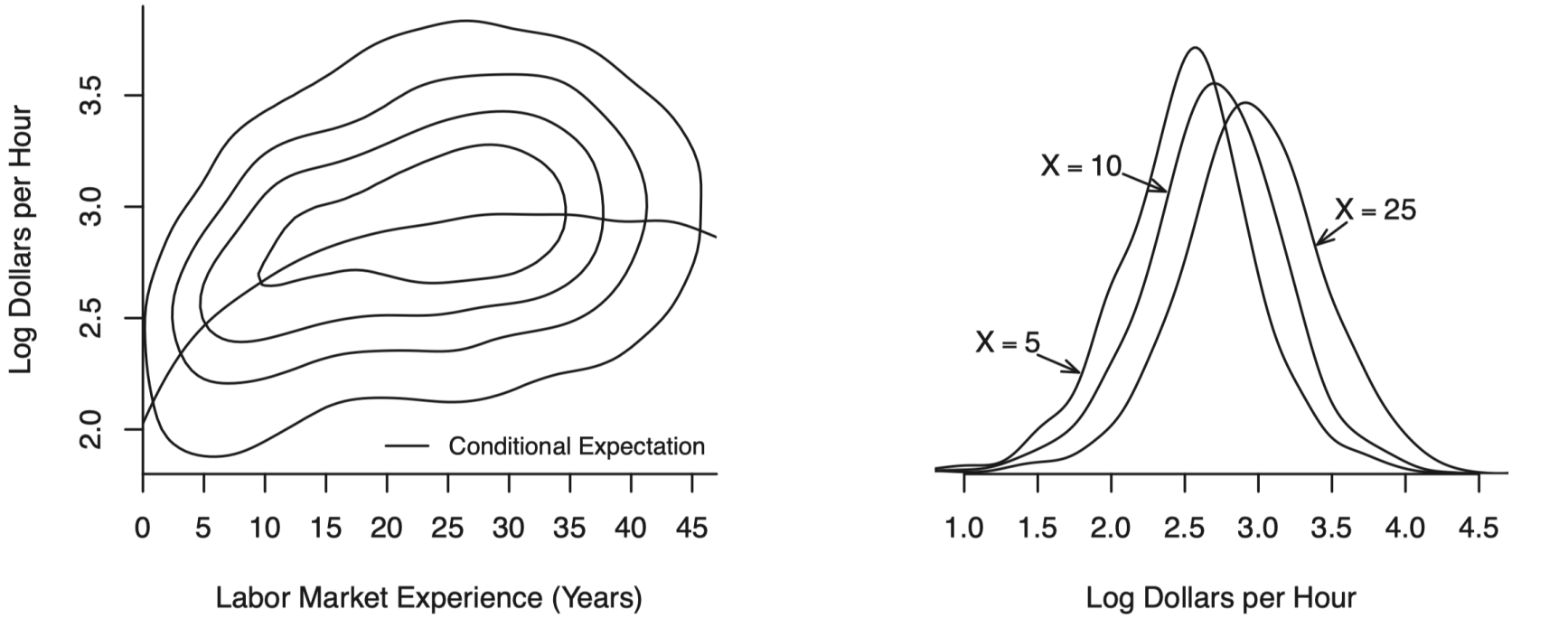
$$P(y|x) = \frac{P(y,x)}{P(x)}$$

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”

条件密度函数

为什么用联合概率分布函数和联合密度函数来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

条件期望函数的误差项

- 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

- CEFE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 $h(x)$, $E(h(X) \cdot e) = 0$ → 通常利用该性质进行线性变换

CEF与建模的关系

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。

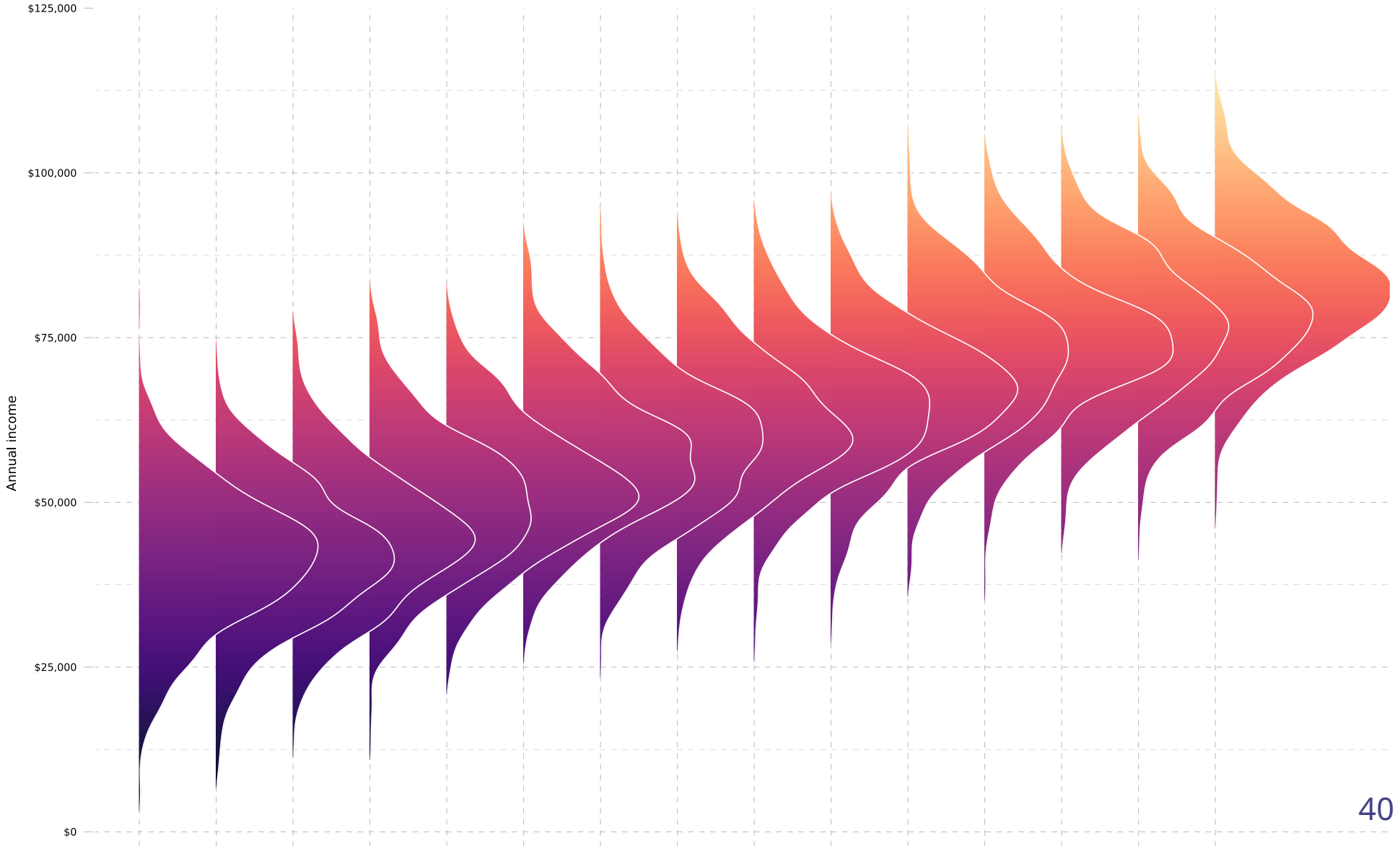
如截距模型，线性模型，Logit模型等。

基于样本的CEF

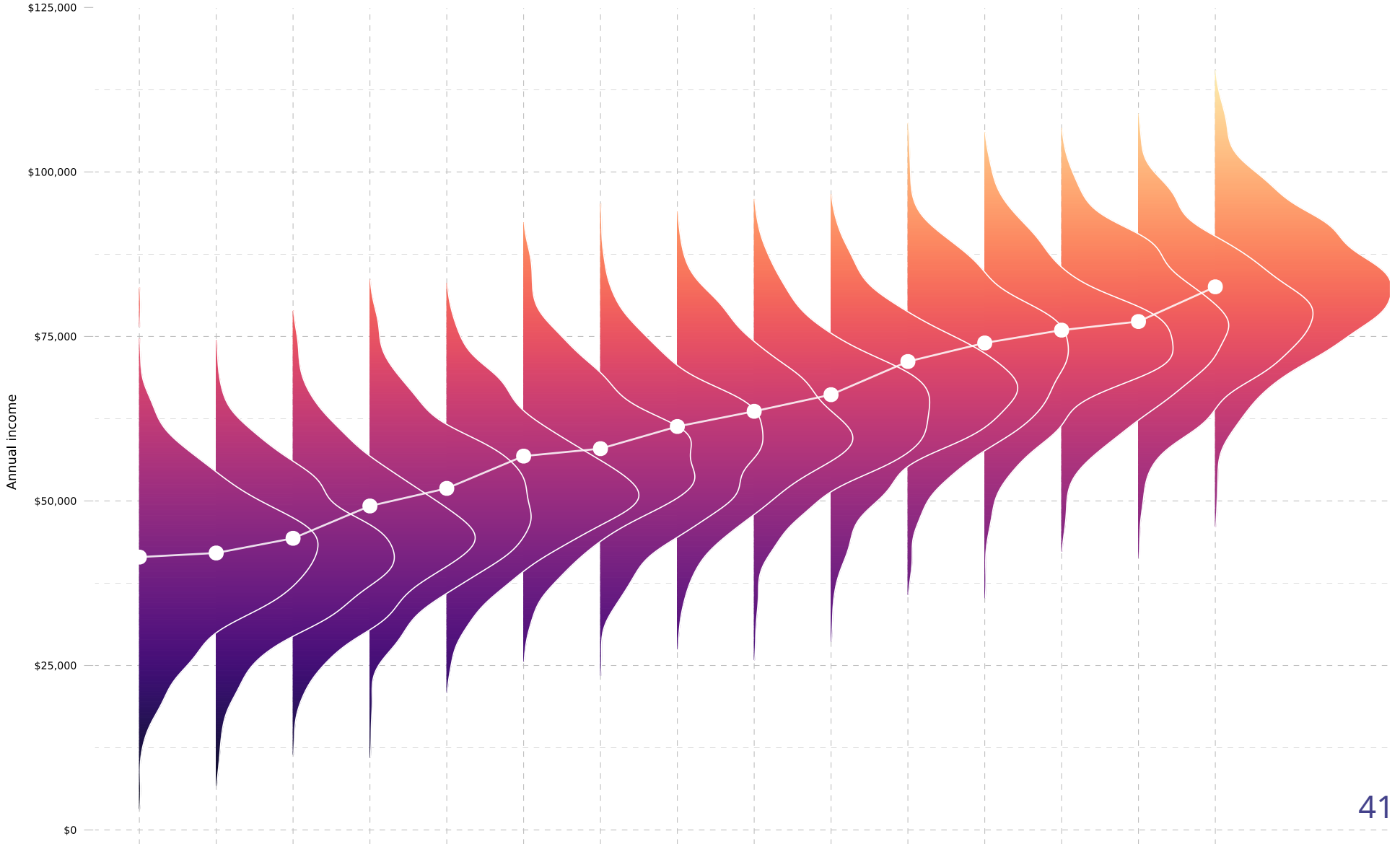
- 期望本身是总体概念（价值观）
- 实际中，我们是基于样本信息推断总体信息，例如用样本均值推断总体期望
- 将CEF写作基于样本的CEF: $E[Y_i \mid X_i]$

从图形上看CEF...

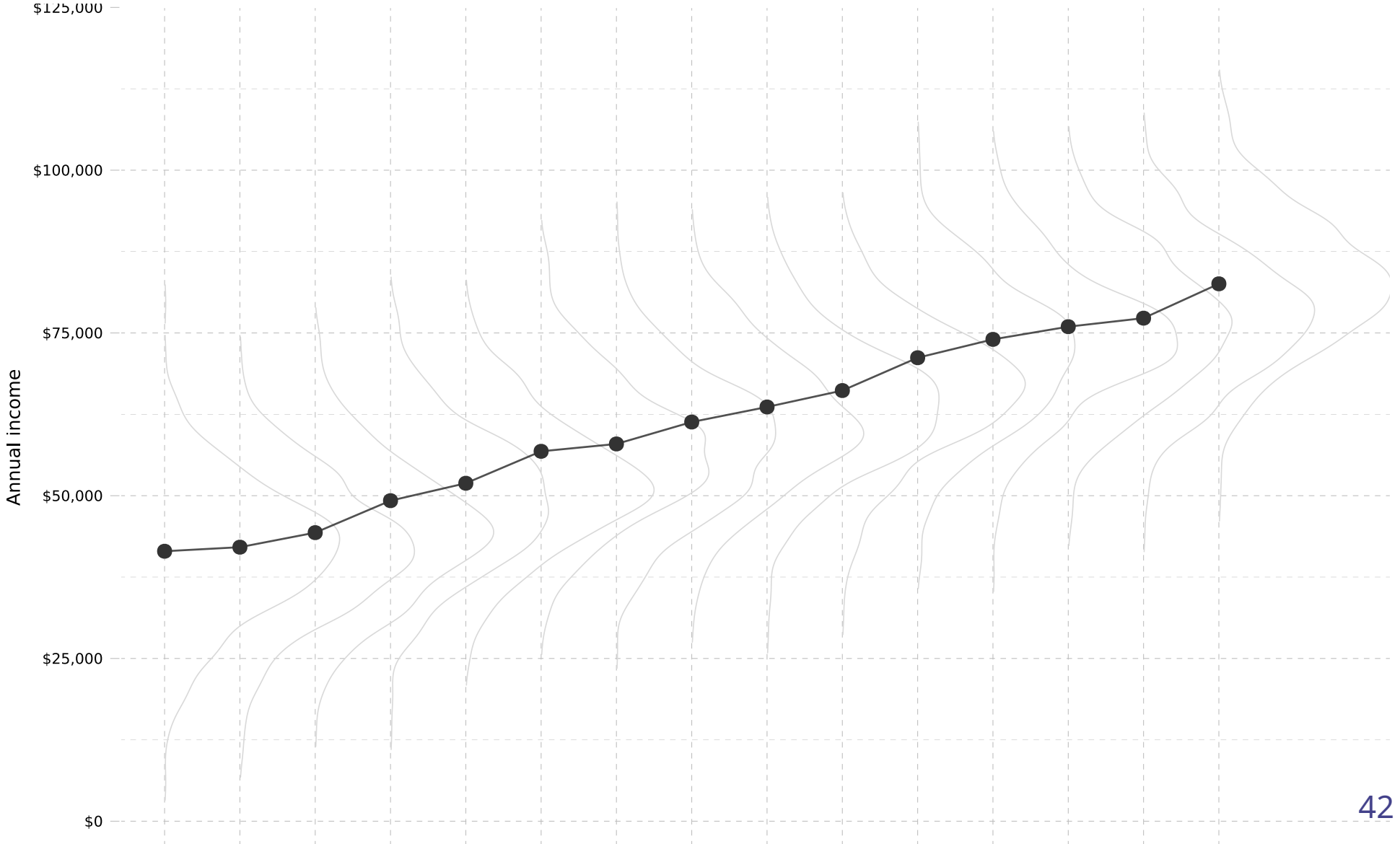
条件分布 Y_i, 对于8, ..., 22不同教育年限的 $X_i = x$.



条件期望函数 $E[Y_i \mid X_i]$ 其实是这些条件分布的均值



若只关注条件期望函数 $E[Y_i \mid X_i]$...



实际数据（MHE）

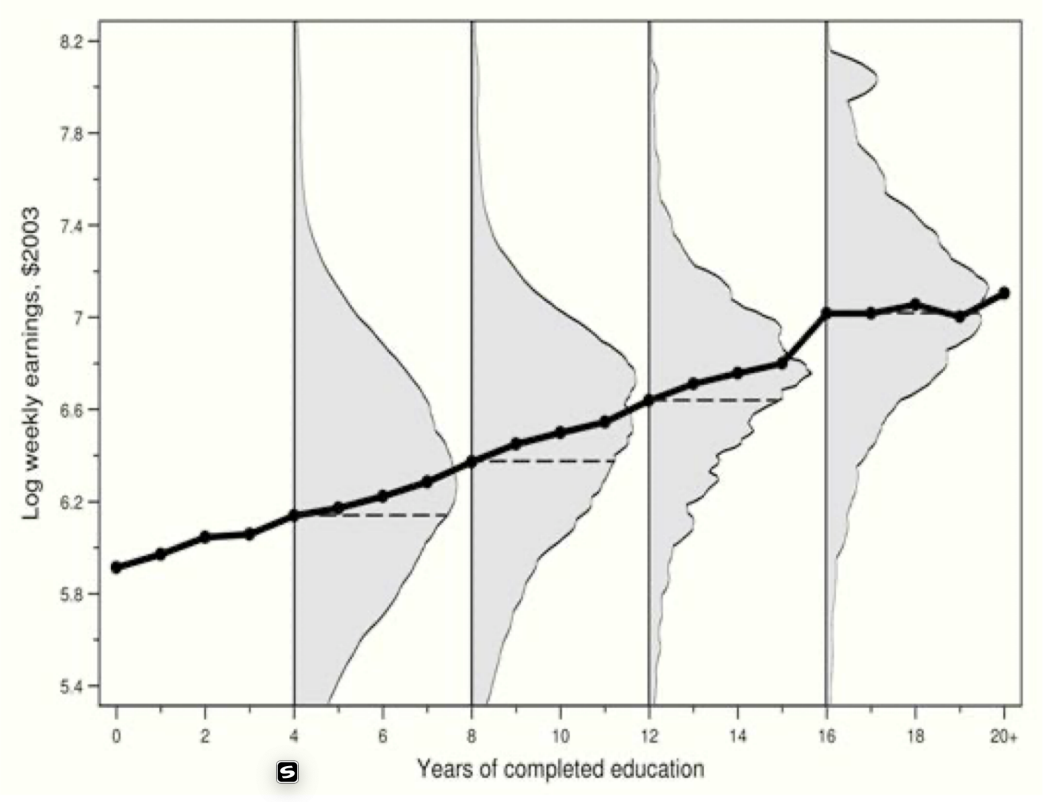


Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49. The data are from the 1980 IPUMS5 percent sample.

手动操作：研究问题 ⟶ CEF

CEF基于数据出发，对于理解变量间至关重要，但经验研究的第一问题是如何将理论结果转换为实证模型，$E\left[\text{工资}_i \mid \text{运动技能}_i\right]$：

- 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 $E(Y \mid X = x)$

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

$$Y_i = E[Y_i \mid X_i] + e_i$$

1. 被 X_i 解释的部分(*i.e.*, CEF $E[Y_i \mid X_i]$)
2. 具有特殊性质的干扰项†

    i. e_i 均值独立于 X_i, *i.e.*, $E[e_i \mid X_i] = 0$
    ii. e_i 与 X_i的任何函数不相干

† 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

$$Var(Y_i) = E[Var(Y_i \mid X_i)] + Var(E[Y_i \mid X_i])$$

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**）

$$E[Y_i \mid X_i] = \underset{m(X_i)}{argmin} \; E\left[(Y_i - m(X_i))^2\right]$$

CEF是给定 X_i 能够预测 Y_i 最好预测方式.

注意 m 可以是任意形式函数（包含非线性），但我们偏好形式简单、解释力强的模型 → **线性投影函数（LPF）**

为什么不是线性CEF，而是LPF呢? 线性CEF也是常见的线性回归模型(linear regression model)。其中一个原因是 $m(x1, x2)$ 完整线性CEF为

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

为什么是LPF而不是CEF?

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**

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

为什么是LPF而不是CEF?

- LPF是MSE最小的线性函数:

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

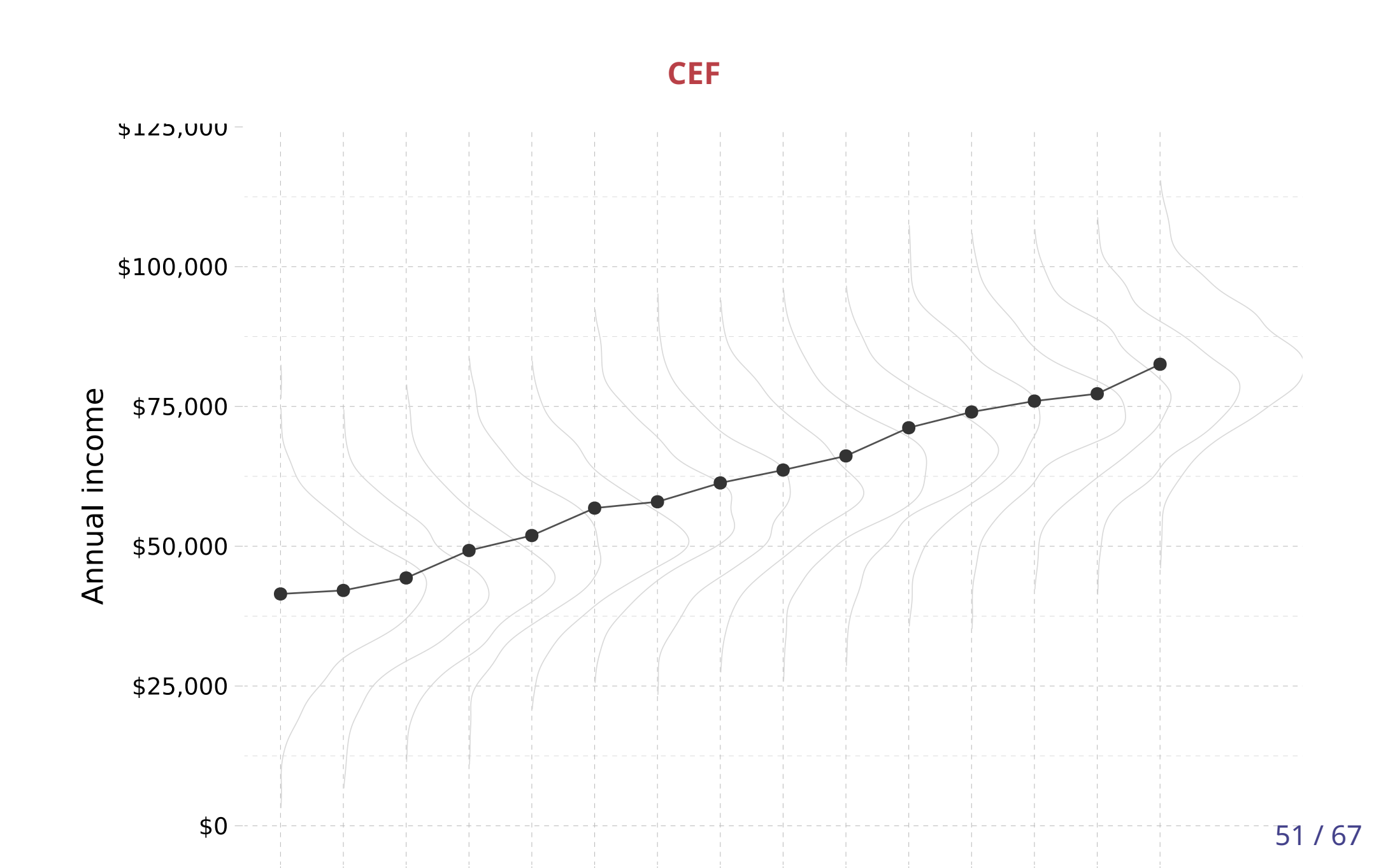
- 依据一阶条件: $E\left[X_i\left(Y_i - X_i'b\right)\right] = 0$得到 b 的最优解 $\beta = E\left[X_iX_i'\right]^{-1}E[X_iY_i]$
- $X_i'\beta$ 是 Y_i 在 X_i 上的最优线性投影（best linear projection, BLP）, 向量 β 是线性投影系数（linear projection coefficient）
- 根据一阶条件重新构建 $E\left[X_i\left(Y_i - X_i'\beta\right)\right] = 0$, 也就是说 Y 的线性投影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

为什么是LPF而不是CEF?

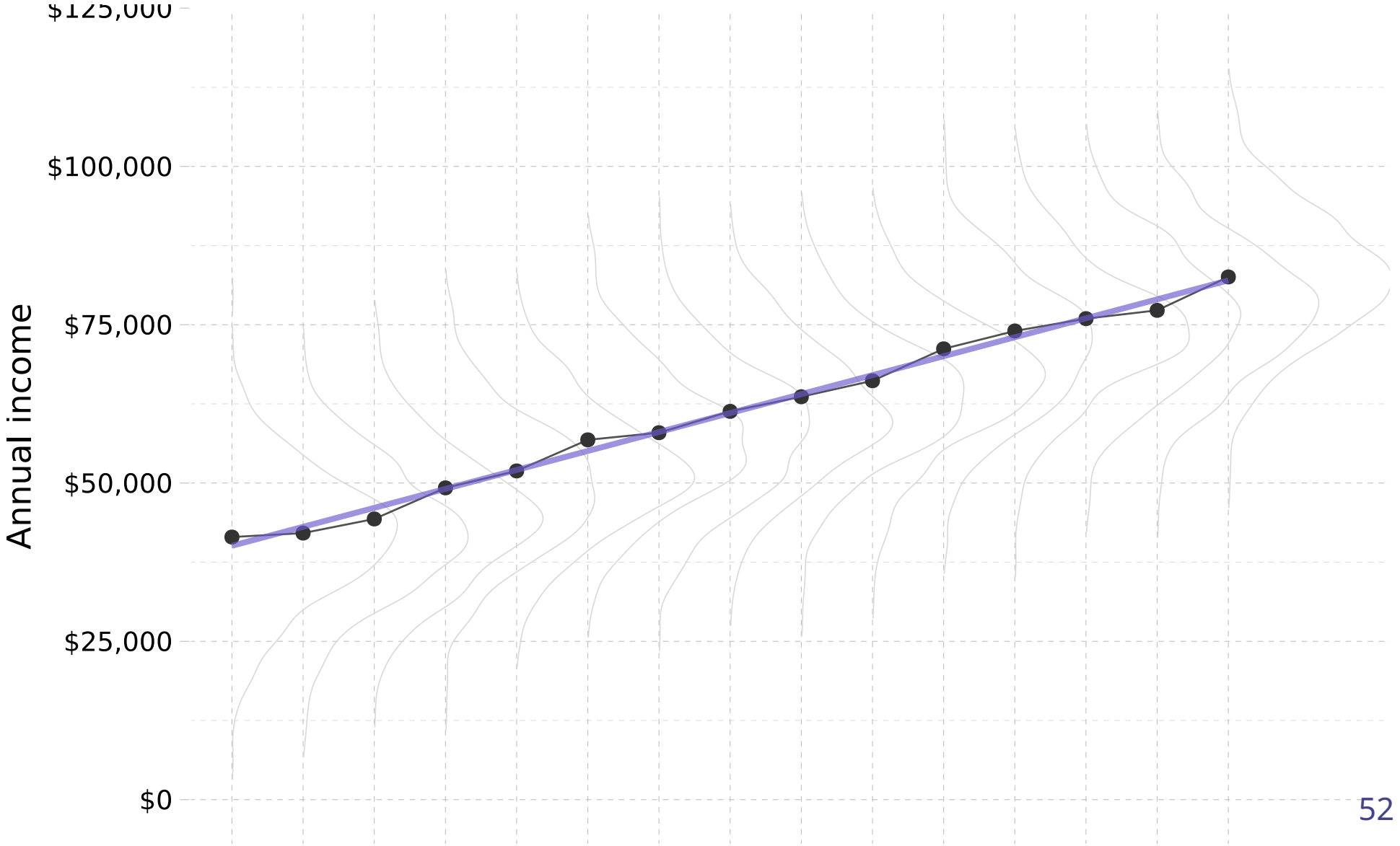
- LPF是MMSE，CEF也是MMSE。继续使用最小化MSE**准则**：

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

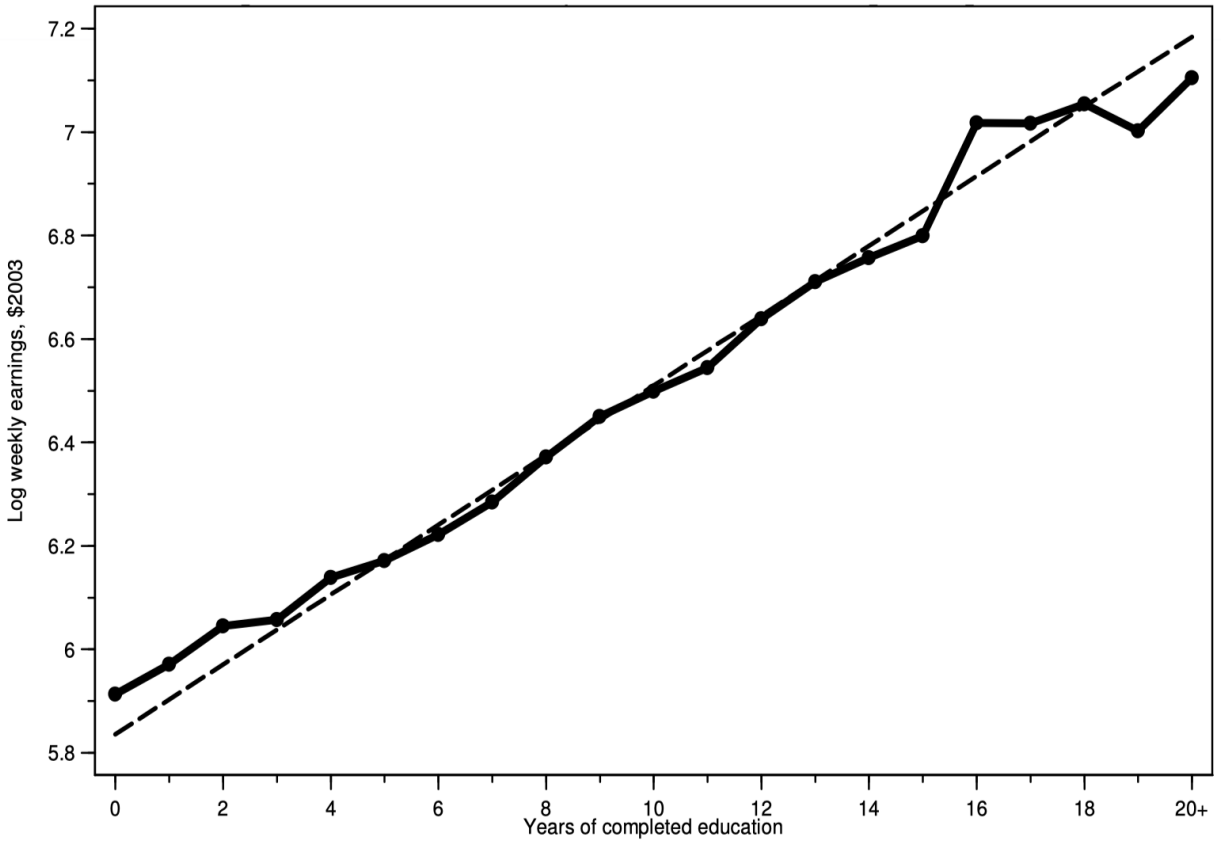
- 回归与条件期望函数定理（Regression-CEF Theorem）
- 结论：
  - **LPF同样是CEF的MMSE和BLP**
  - 通常而言，CEF不一定是线性的；但CEF若是线性的, 那么LPF就是CEF


CEF
$125,000
$100,000
$75,000
$50,000
$25,000
$0
Annual income

LPF去估计CEF



实际数据



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

因果关系基于CEF而不是LPF

若CEF是相关关系的 $\longrightarrow$ LPF是相关的

若CEF是因果关系 $\longrightarrow$ LPF是因果的

问题是：怎样获得一个因果的 CEF？ (客观)

$\longrightarrow$ 必须依赖于理论认知(主观)

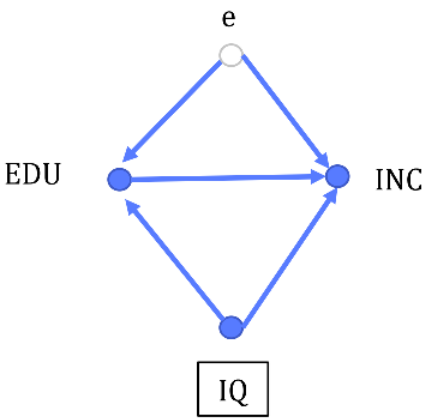
因果关系基于CEF而不是LPF

- **实际上**直接的做法是使用 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**通过假定 **干扰项条件均值独立于自变量**，即 $E(e \mid X) = E(e) = c$ ，才保证LPF的估计系数是有效的

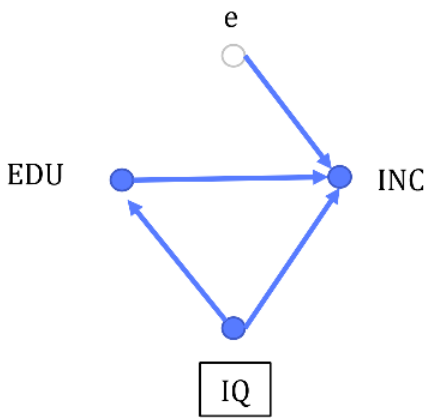
干扰项条件均值独立于自变量

假设 LPF：

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



干扰项与解释变量相关



干扰项与解释变量不相关

- **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 → INC 直接因果路径
- EDU ← IQ → INC 混淆路径1
- EDU ← e → 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

干扰项条件均值独立于自变量

- $E(e \mid EDU, IQ) = E(e) = c$ 意味着：给定我们班{EDU = 研究生, IQ = 高} 干扰项的平均值一样(c)，假如性格（不可观测）进入 e，干扰项条件均值独立于自变量成立就表示符合上面两个条件的同学的性格是一样的。
  - 这样，若EDU与IQ其中一个变化，另一不变，收入的条件平均值变化就可以归因于其中那个变化了的条件，获得“净”的因果效应。
- 因此，LPF中的系数 β_1 和 β_2 就是 EDU 和 IQ 对 INC 的因果效应。
- 将 $E(e \mid EDU, IQ) = c$ 的常数c 并入 LPF常数项，就获得了与CEFE性质一样的 $E(e \mid EDU, IQ) = 0$ 重要假设。 只不过该假设是从 LPF 出发建立的，也就是常在计量教材见到的线性回归方程的**干扰项条件均值为0**的假设。

干扰项条件均值独立于自变量的理解

- 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:

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

- 将线性CEF对EDU求偏导：**偏回归系数**

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

β_1 表示在IQ固定不变，INC 的期望值（均值）随 EDU 如何变化。

因果关系基于CEF而不是LPF

若 $E(e \mid EDU, IQ) \neq 0$, 干扰项的存在如何影响因果关系的估计?

假设只观测到了INC和EDU, 将 LPF 写作：

$$INC = \alpha + \beta_1 EDU + \varepsilon, \quad \varepsilon = \beta_2 IQ + e$$

此时:

$$E(\varepsilon \mid EDU) = E(\beta_2 IQ + e \mid EDU) = \beta_2 E(IQ \mid EDU) \neq 0$$

此时的LPF为:

$$INC = \alpha + \beta_1 EDU + \varepsilon, \quad E(\varepsilon \mid EDU) \neq 0$$

因果关系基于CEF而不是LPF

- 假如此时我们错误地把 LPF 当成了正确的、具有因果关系的线性CEF,
- 实际上相当于理解为 我们建构的模型 LPF：
  $INC = \gamma_0 + \gamma_1 EDU + u, \quad E(u \mid EDU = 0)$ *的对应的"具有因果关系"（其实是相关关系）CEF*： $E(INC \mid EDU) = \gamma_0 + \gamma_1 EDU$
- 求偏导：

$$\frac{dE(INC \mid EDU)}{dEDU} = \gamma_1$$

- γ_1 反映了INC 的期望值随EDU如何变化, 但并没有控制IQ不变。

因果关系基于CEF而不是LPF

计算 γ_1 和 β_1 的关系:

$$
\begin{aligned}
&E(INC \mid EDU) \\
&= E\left(\alpha + \beta_1 EDU + \beta_2 IQ + e \mid EDU\right) \\
&= \alpha + \beta_1 EDU + \beta_2 E(IQ \mid EDU)
\end{aligned}
$$

- 对EDU求导:

$$
\frac{dE(INC \mid EDU)}{dEDU} = \beta_1 + \beta_2 \frac{dE(IQ \mid EDU)}{dEDU}
$$

即

$$
\gamma_1 = \beta_1 + \beta_2 \frac{dE(IQ \mid EDU)}{dEDU}
$$

反映相关关系的CEF

- 假设受教育程度于智商之间存在线性相关关系:

$$E(IQ \mid EDU) = \phi_0 + \phi_1 EDU$$

即

$$\frac{dE(IQ \mid EDU)}{dEDU} = \phi_1$$

则:

$$\gamma_1 = \beta_1 + \beta_2 \phi_1$$

为什么需要推断?

- **之前** 重点关注了CEF与LPF，都代表**总体意义**。刻画总体我们只能通过样本, 通过统计推断的方式进行。
- 将LPF 设定为: $Y = X'\boldsymbol{\beta} + e, \quad E(e \mid X) = 0$
- 展开: $Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + e$

$$E(Y \mid X) = X'\boldsymbol{\beta}$$

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

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

为什么需要推断?

- 由一阶条件可得:

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

- 此条件同等与:

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

- 由此可见, LS的本质是求解系数 $\hat{\boldsymbol{\beta}}_{ols}$ ， 使得自变量 X 与残差 ε 不相关

$$\hat{\boldsymbol{\beta}}_{ols}=E\left[\boldsymbol{X}\boldsymbol{X}'\right]^{-1}E[\boldsymbol{X}Y]$$

为什么需要推断?

- 将LPF代入上式:

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

  其中**由于假设**： $E(e \mid X) = 0$

- 故 $E[\boldsymbol{X}e] = E_{\boldsymbol{X}}[E(\boldsymbol{X}e \mid \boldsymbol{X})] = E_{\boldsymbol{X}}[\boldsymbol{X}E(e \mid \boldsymbol{X})] = \mathbf{0}$
- 故 $\hat{\boldsymbol{\beta}}_{ols} = \boldsymbol{\beta}$ 以上讨论说明, 最小二乘法估计系数 $\hat{\boldsymbol{\beta}}_{ols}$ 就是LPF系数，同样也是线性条件期望函数 $E(Y \mid X) = \boldsymbol{X}'\boldsymbol{\beta}$ 的系数 $\boldsymbol{\beta}$

为什么需要推断?

- **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，通常**靠经验和理论判断**;
- **残差项** ε 是最小二乘法计算出来的， 总是与X正交。

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

e

- 最小二乘只是估计方法.super[.pink[†]]，常见的估计方法还有矩方法、最大似然估计等
- 总体 $\hat{\boldsymbol{\beta}}_{ols}=E[\boldsymbol{X}\boldsymbol{X}']^{-1}E[\boldsymbol{X}Y]$
- 样本 $\hat{\boldsymbol{\beta}}_{ols}=\left(\sum_i X_iX_i'\right)^{-1}\left(\sum_i X_iY_i\right)$

.footnote[.pink[†] 矩方法(method-of-moments) 。根据大数定律和中心极限定理使用样本矩 $\frac{1}{n}\sum_i X_iX_i'$ 估计总体矩 $E\left[X_iX_i'\right]$. 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.]

推断的大样本性质

1. **异方差影响会显著性**

> ...如果异方差变化很大，比如使标准误上升了30%或者低于传统意义上的标准误[†](通常会大于)，可能标志着在稳健性计算中存在有限样本偏误。

2. 现代微观实证建立在大样本假定下，已经避免了强的经典假设(正态分布、自变量非随机、线性 CEF、同方差) .

† 标准差(SD)是样本数据的方差的平方根，衡量样本数据的离散程度; 标准误(SE)是样本均值的估计值(是随机变量)的标准差，衡量样本均值的离散程度。实际中用样本均值来推断总体均值，那么样本均值的离散程度（标准误）越大，抽样误差就越大。所以用标准误（SE）来衡量抽样误差的大小，所以统计软件中报告是SE。