

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

03 February 2024

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

- 计量经济学是经济学的一个分支，专注于使用数学和统计方法来测试假设、验证理论和估计模型。
- 它将理论经济学、数学和统计学的技术结合起来，旨在提供定量分析经济现象的实证证据。
- 计量经济学的核心目的是识别并量化经济变量之间的关系，以此来预测未来趋势、支持政策制定和决策过程。

计量经济学的研究框架

采用计量经济学的研究框架通常遵循以下步骤：

- **问题定义与假设提出**: (Problem Definition and Hypothesis Formulation):首先明确研究问题，并基于经济理论提出可验证的假设。这一步骤是研究的基础，它确定了研究的方向和焦点。**Topic → Problem → Question**
- **模型构建(Model Specification)**: 根据假设构建经济模型，模型通常以数学方程式的形式表示，描述变量之间的关系。模型的选择和构建依赖于理论知识和实际经验。
- **数据收集(Data Collection)**: 根据模型的需要，收集相应的数据。数据可以是时间序列、横截面数据或面板数据。数据的质量和适用性对后续分析至关重要。
- **估计与测试 (Estimation and Testing)** : 使用统计方法对模型进行估计和假设检验。常用的估计方法包括普通最小二乘法 (OLS) 、最大似然估计 (MLE) 和广义矩估计 (GMM) 等。估计结果用于检验理论假设的有效性。

计量经济学的研究框架

- **模型验证与诊断（Model Validation and Diagnostic Testing）**：通过各种统计检验和诊断方法验证模型的准确性和稳健性。这包括残差分析、异方差性测试、自相关检验等。
- **政策分析与预测（Policy Analysis and Forecasting）**：基于估计和测试结果，进行政策效果分析和经济预测。这一步骤将研究成果应用于实际经济问题和政策制定中。
- **结果解释与报告（Interpretation and Reporting）**：最后，清晰地解释研究结果，并将其撰写成报告或论文。这一步骤要求研究者以易于理解的方式传达复杂的统计和经济分析结果。

应用：球队胜率与上座率

在体育经济学中，球队上座率受哪些因素影响是一个常见的问题。通过经济计量模型进行量化分析，来回答该问题：

第一步：理论或常识指向研究假设

在文献综述之后，需要指出，球队的获胜比例（WPCT）可能是影响上座率（ATT）的一个重要因素（是符合理论或常识的），而且理论会指向一个基本假设：球队表现越好，吸引的观众越多，从而上座率更高

第二步：模型构建与可验性

模型 $ATT = f(WPCT)$ 是对这种关系的抽象表示，而进一步假设 $\frac{\partial ATT}{\partial WPCT} > 0$ 表明，随着胜率的提高，预期上座率也会增加。

应用：球队胜率与上座率

第三步：总体线性模型表达式与样本链接

将理论模型转化为总体线性表达式 $ATT = \beta + \tau \times WPCT + e$ 是为了便于使用统计方法进行估计。这里， β 是截距项， τ 是你感兴趣的斜率参数（即胜率对上座率的影响）， e 是误差项，代表其他未观测因素的影响。

应用：球队胜率与上座率

第四步：参数估计与方法的限制条件

OLS是估计线性回归模型参数的常用方法。通过最小化误差项的平方和来找到最佳拟合线，从而估计模型中的 τ 。但使用OLS方法在许多假设（如误差项的同方差性和独立性）被满足的情况下，可以提供有效且一致的参数估计。

- **关于估计值 $\hat{\tau}$ 是否为真实值 τ ？** 估计值 $\hat{\tau}$ 是基于样本数据对真实参数 τ 的最佳猜测。在理想情况下，如果模型正确指定且所有OLS的假设都被满足， $\hat{\tau}$ 会随着样本大小的增加趋向于真实值 τ 。但在实际应用中，可能因为**模型设定错误、样本选择偏差、遗漏变量**等问题，导致 $\hat{\tau}$ 与 τ 存在偏差。因此， $\hat{\tau}$ 是对 τ 的一个估计，它的准确性和可信度取决于模型设定的合理性和样本数据的质量。

应用：球队胜率与上座率

第四步：参数估计与方法的限制条件

- 上述问题演变成计量经济学的核心任务，即"**识别**" (**identification**)
 - 涉及确定模型中的相关关系是否可以从数据中可靠地推断出来
 - 通过与已知的标准或模型进行比较，从而确定未知参数的过程

因果效应的识别过程

- 大部分问题期盼因果结论，因此识别过程的本质就演变为能够准确区分和确定特定参数或因果效应的能力。这个过程要求研究者通过合理的**研究设计**策略来区分因果关系，而不仅仅是相关性。
- 在社会科学论文中，确立识别策略意味着**明确地指出如何使用数据和模型来估计因果效应，这通常涉及顺应研究传统、排除竞争性理论、实施自然实验等具体问题

应用：球队胜率与上座率

第四步：参数估计与方法的限制条件

推断是包含在识别过程中必要环节

- 在统计学传统中，**推断** (inference) 是指基于样本数据对总体参数进行估计和假设检验的过程。这相当于使用现场照片 ($\hat{\tau}$) 去推测数据库中的真实照片 (τ)。
- 经济学的推断不仅仅关注于统计学的显著性，还要考虑估计值的经济学意义，即它们在现实世界中的应用和重要性。

区分统计学意义与经济学意义

- **统计学意义**关注的是估计结果是否不太可能是随机波动所导致的，通常通过p值、置信区间等统计测试来评估。
- **经济学意义**则更加关注估计结果的实际影响大小和方向，以及这些结果在经济理论和政策制定中的应用。

应用：球队胜率与上座率

第四步：参数估计与方法的限制条件

因果推断

- 当我们说估计值 $\hat{\tau}$ 具有**因果性**时，指的是延续科学中的**识别**学术传统去**推断**真实值 τ 。隐含的前提是：真实值 τ 已经具有因果特质。因果推断的目标是使用观察到的数据来估计未被观察到的因果效应。
- 实现因果推断需要满足一系列假设，包括但不限于无遗漏变量偏差、稳定的单元处理值假设（SUTVA）以及潜在结果的独立性假设。

总之，识别和推断是计量经济学中最重要的概念，它们使研究者能够从数据中提取有意义的因果关系。通过理论建构、稳健性检验、排除竞争性理论等步骤，研究者可以确立他们的研究不仅在统计上显著，而且在经济上具有意义，从而为经济政策和理论提供有力的支持。

例子：球队胜率与上座率

第五步：结果解读与应对质疑

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

- 估计结果能够识别胜率 (WPCT) 对上座率 (ATT) 有显著的正面影响。但是模型的解释力度 ($R^2 = 0.311$) 表明还有大部分变异未被模型捕捉，这引出了模型可能存在的问题和进一步改进的方向

质疑1：模型的解释力度不够

尽管胜率是上座率的一个重要因素，但显然还有其他因素影响着上座率。这导致了两个主要问题：

- 研究问题是否足够重要？即使胜率是一个显著的因素，但如果其他未考虑的因素对上座率有更大的影响，那么仅研究胜率可能不足以全面理解上座率的变化。
- 模型是否过度简化？在实际应用中，很少有单一因素能够完全解释一个经济现象。因此，模型可能需要包含更多的解释变量来提高解释力度。

例子：球队胜率与上座率

第五步：结果解读与应对质疑

质疑 2:用估计值代替真实值是偏误的

内生性问题通常出现在模型中的解释变量与误差项相关时。例如，市场规模和联赛差异可能同时影响胜率和上座率，如果不将这些因素纳入模型，可能会导致估计偏误。

- 控制变量的引入：通过将市场规模和联赛差异作为控制变量纳入模型（ $\mathbf{X}\beta$ ），可以帮助减少遗漏变量偏差，提高模型估计的准确性。
 - 将总体线性表达式修正为 $ATT = \mathbf{X}\beta + \tau \times WPCT + e$
 - 市场规模与联赛差异进入了 $\mathbf{X}\beta$
 - 最关键的：这两个因素同时影响球队胜率和上座率。若忽略它们会导致**遗漏变量偏差**，产生所谓的**内生性**问题 →

内生性解决方法：除了加入控制变量外，还可以使用工具变量等方法解决内生性问题，前提是能找到合适的工具变量。

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

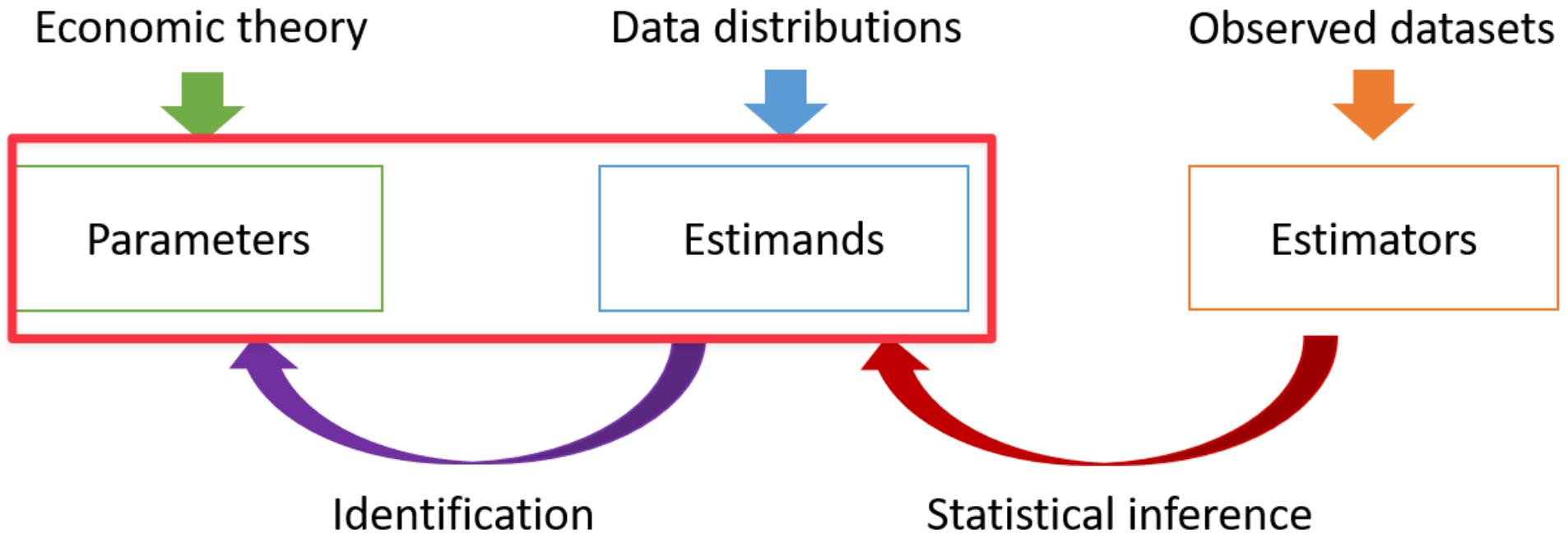
Adj R-squared = 0.358

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

- 那么59.31就能代表真实值了么？
- 练习：系数如何解释？具有如何意义？Dummy(AL=1) vs Continuous

经验研究的研究逻辑



- 观察样本(estmators) → 客观事实(=estmands) → 因果模型(=parameter) → 理论知识(knowledge)
- 在本次课程中，为建立样本数据与因果推断的逻辑体系，将重点介绍两个最重要的考
 - 平衡遗漏变量与过度控制的问题 + 避免内生性问题的质疑

从客观事实到因果模型

- 遗漏变量偏误（Omitted Variable Bias, OVB）的讨论指出，如果重要的解释变量没有被包括在模型中，可能会导致对其他变量效应的估计产生偏误。这是因为遗漏的变量可能与模型中的变量相关，并且直接影响到结果变量。
- 过度控制问题涉及到将那些受到干预变量影响的变量作为控制变量加入模型，这可能会干扰真实的因果关系，导致对感兴趣效应的错误估计。

如何平衡

- 研究者应当根据理论知识以及对研究领域的深入理解，仔细选择控制变量。理想的控制变量应当是那些影响结果变量、与干预变量相关，但不是由干预变量决定的变量。
- 遵循时间原则，优先考虑那些在干预变量发生前就已确定的变量作为控制变量。这有助于减少因逆向因果或干预变量引起的变化而产生的偏误。

从客观事实到因果模型

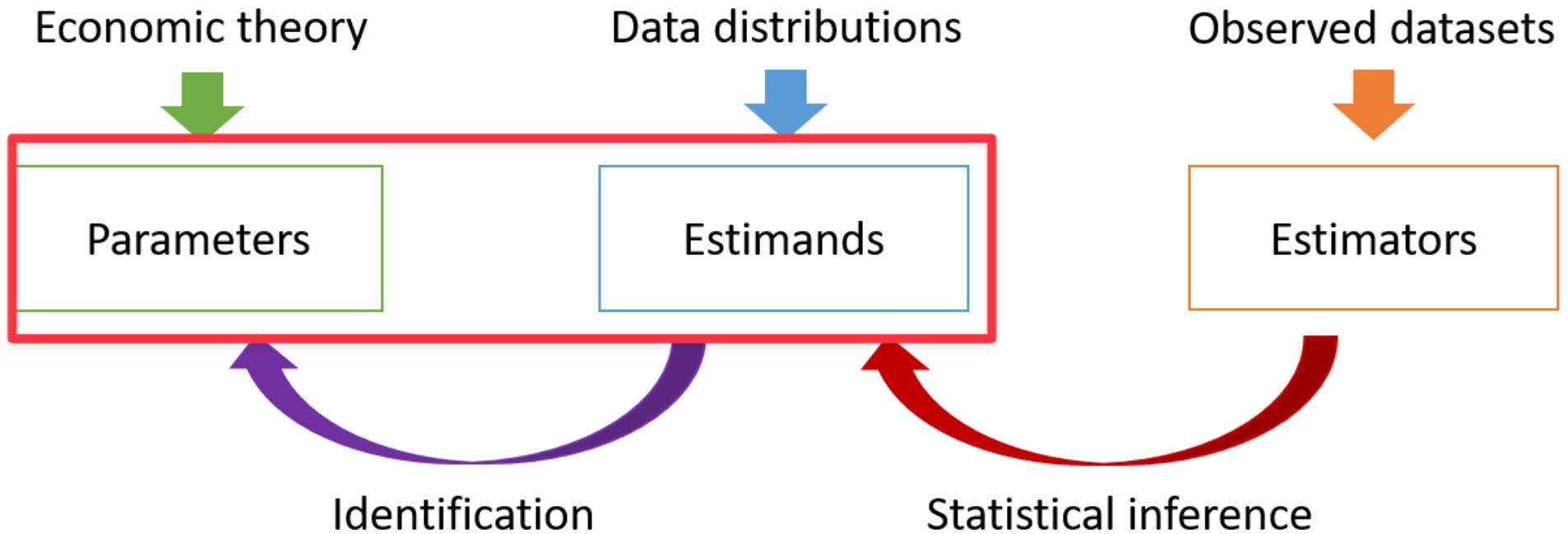
实施建议

- **理论指导**: 研究设计和控制变量的选择应该基于坚实的理论基础，确保模型能够反映研究问题的本质。
- **稳健性检验**: 通过稳健性检验，如使用不同的模型规格和控制变量组合，检查估计结果的稳定性。
- **考虑人的预期**: 在某些研究情景下，考虑人的预期和行为对未来的预测也非常重要，这可能要求引入动态模型或使用前瞻性数据。

通过教育回报率的例子具体阐释了如何实施这一过程

从客观事实到因果模型！
→ 平衡遗漏变量与过度控制

经验研究的研究逻辑



- 增加控制变量确实可以提高回归模型解释力，但过度控制会产生失去模型的意义
→ 引入遗漏变量偏误公式工具

应用：教育回报率

- 遗漏变量偏误公式提供了一个强有力的工具，用于分析当模型遗漏了重要解释变量时，估计系数偏误的方向和大小。
- 长回归方程和短回归方程。**长回归方程 (1) 考虑了能力 A_i 对收入 INC_i 的影响，而短回归方程 (2) 则没有。这种简化假设可能导致对教育 EDU_i 的回报率 β 的估计出现偏误。

$$INC_i = \alpha + \tau EDU_i + A'_i \gamma + e_i \quad (1)$$

$$INC_i = \alpha + \beta EDU_i + v_i \quad (2)$$

- 注意， α, τ, γ 是总体意义上的， e_i 表示扰动项。暂且假设系数 τ 具有因果性，是我们最感兴趣的

- 遗漏变量偏误公式 (Omitted Variable Bias Formula) 为：

$$\hat{\beta}_{ols} = \frac{Cov(INC_i, EDU_i)}{Var(EDU_i)} = \tau + \gamma' \delta_{A_{EDU}}$$

- 其中, $\delta_{A_{EDU}}$ 是对 A_i 关于 EDU_i 回归得到的系数

解释和含义

在教育回报率的例子中，考虑能力等潜在的混杂变量对于避免偏误和错误的因果推断至关重要，要正确识别感兴趣的参数，以下条件必须满足其一：

1. 如果受教育程度与能力大小无关 ($\delta_{A_{EDU}} = 0$)，这意味着能力不会影响个体选择受教育的程度。
2. 如果在控制受教育程度后，能力大小与工资多少无关 ($\gamma = 0$)，则意味着能力对于收入的影响可以被忽略。

实际应用中的挑战

实际上，满足以上两个条件是非常困难的，因为教育程度通常与个体能力强相关，而个体的能力又显著影响其收入。

- 尽量把能力这个看不到的变量给控制住 → 找到代理变量

例子(MHE)

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

MHE给出了4种增加控制变量的方式，关于工资(Y)对上学年限(X)的回归（来自NLSY, 美国青年纵向调查）

the regression of Y on X , regress Y on X

例子(MHE)

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)

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)

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)

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。

[†] 武装部队资格测验 (AFQT) 是美国军队招募的基本资格测验。它是由美国国防部于1950年开发的一个筛查测试，用于评估一个人是否符合入伍资格。

例子(MHE)

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%）。

$$\frac{Cov(INC_i, EDU_i)}{Var(EDU_i)} = \tau + \gamma' \delta_{A_{EDU}}$$

- 讨论为什么？

遗漏变量偏误公式

- OVB公式**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 → 降低了误判的可能性
- τ 是否因果还应该有其他的假设：**条件独立假设** → 更理想的方式是RCT

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

加入坏的控制变量 → 过度控制问题

- 在加入控制变量时，选择哪些变量进行控制至关重要。理想的控制变量应该是那些与处理变量和潜在结果都相关，但不是由干预变量引起的变量。
- 不良控制变量，如个人职业和就业行业，在教育研究中可能不适合作为控制变量，因为这些变量可能是教育的结果而非原因。控制这些变量可能引入过度控制问题，从而掩盖教育对收入的真实影响。
- 控制变量的选择应基于理论和先前的研究，确保这些变量不是处理变量引起的。此外，也需要考虑到变量之间可能存在的动态关系和因果路径

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？

教育水平的系数变小可能仅仅是(过度控制导致的)选择偏误表现

如何选择控制变量

1. **理论指导**: 控制变量的选择应该由相关理论和以往研究来指导。理论框架可以帮助识别可能影响研究结果的关键变量。
2. **先前研究**: 查阅相关文献，了解在相似研究背景下哪些变量被控制，可以提供有价值的参考。
3. **与干预和结果相关**: 选择与干预变量（例如，受教育程度）和潜在结果（例如，收入）都相关，但不是由处理变量直接引起的变量。
4. **避免结果变量**: 避免使用可能是干预变量结果的变量作为控制变量，因为这可能引入过度控制问题。
5. **数据可得性**: 在实际研究中，控制变量的选择还受限于数据的可得性。选择可靠和准确测量的变量。

避免过度控制问题

1. **识别中介变量**: 中介变量是介于干预变量和结果变量之间的变量，干预变量通过中介变量影响结果变量。在分析中控制中介变量可能会屏蔽干预变量的部分或全部效应。
2. **考虑变量的时间顺序**: 理想的控制变量应该是在干预变量发生之前就已经确定的变量。这有助于保证控制变量不是由干预变量引起的。
3. **动态关系和因果路径**: 在选择控制变量时，考虑变量之间的动态关系和可能的因果路径。理解变量之间是如何相互作用的，可以帮助避免不当的控制变量选择。
4. **稳健性检验**: 通过进行一系列稳健性检验，如在模型中逐步加入或移除控制变量，可以评估控制变量选择对估计结果的影响。

从客观事实到因果模型II

→ 避免内生性问题

- 内生性问题通常源于三个主要因素：遗漏变量、测量误差和互为因果。
- 如果干扰项 e 和干预变量 X_i 是相关的，那么我们称这个模型存在内生性问题。

内生性问题的影响

- **严重问题**：如果没有清楚指出解释变量与因变量之间的因果关系，可能会被学术期刊拒稿。
- **轻微问题**：如果无法在可信的水平下控制内生性问题，可能无法发表在顶级期刊。

测量误差对教育回报率估计的影响

(1) 教育存在测量误差

如果我们在测量教育水平时存在误差，比如受访者报告的受教育年数不准确，比如可能忘记了准确的学习年限，或者在提供信息时夸大或缩小了数字：

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

这里， y_i 是个体 i 的收入， x_{1i} 是受教育年数， u_i 是模型中的随机干扰项。

但是，如果我们实际观察到的教育年数 x_{1i}^{obs} 包含了一些误差 v_i ：

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

那么我们实际估计的模型将变为：

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

这里的 $e_i = (u_i - \beta_1 v_i)$ 是新的干扰项，它包含了原始干扰项 u_i 和由于测量误差 v_i 引入的额外误差。

测量误差对教育回报率估计的影响

测量误差导致的主要问题是，我们试图估计的教育对收入影响的系数 β_1 会受到影响：

$$\text{plim} \hat{\beta}_1^{OLS} = \beta_1 \left(\frac{\sigma_{x_1}^2}{\sigma_{x_1}^2 + \sigma_v^2} \right)$$

这意味着，由于存在测量误差，我们估计的教育对收入影响的系数 $\hat{\beta}_1^{OLS}$ 将会比真实的影响 β_1 小。这种现象被称为衰减偏误 (attenuation bias)，它使得我们低估了教育对收入的真实影响。

在该例子，对教育的测量无论是系统性夸大还是低报，**估计系数都会是真实值的下限**。

测量误差对教育回报率估计的影响

(2) 收入存在测量误差 如果接受调查的个人选择瞒报、乱说收入，这意味这测量误差与收入多少和教育程度都没有关系 → 收入与教育估计上是没有相关性的

- 尽管结果变量的测量误差不会直接引起内生性问题，但使得相关性在统计学意义不成立了，这意味可能更难以在统计上发现变量之间的真实关系。

Identification of Endogenous Social Effects: The Reflection Problem

CHARLES F. MANSKI
University of Wisconsin-Madison

First version received December 1991; final version accepted December 1992 (Eds.)

A variety of terms in common use connote endogenous social effects, wherein the propensity of an individual to behave in some way varies with the prevalence of that behaviour in some reference group containing the individual. These effects may, depending on the context, be called "social norms", "peer influences", "neighbourhood effects", "conformity", "imitation", "contagion", "epidemics", "bandwagons", "herd behaviour", "social interactions", or "interdependent preferences".

Mainstream economics has always been fundamentally concerned with a particular endogenous effect: how an individual's demand for a product varies with price, which is partly determined by aggregate demand in the relevant market. Economists have also

Why do such different perspectives persist? Why do the social sciences seem unable to converge to common conclusions about the channels through which society affects the individual? I believe that a large part of the answer is the difficulty of the identification problem. Empirical analysis of behaviour often cannot distinguish among competing hypotheses about the nature of social effects.

- Charles F. Manski 在 1993年的论文中，深入探讨了内生社会效应的识别问题，90年代初期社会科学对个体行为和群体行为相互作用的兴趣日增
- 曼斯基的工作是在这一时期对社会互动效应进行量化和识别方法论的重要贡献之一

互为因果

论文中提出的模型可以表述为：

$$Y = \underbrace{\beta E(Y | g)}_{\text{endogeneous}} + \underbrace{\gamma_1 E(X | g)}_{\text{exogenous}} + \gamma_2 X + \underbrace{\gamma_3 g}_{\text{contextual}} + u$$

这里， Y 是个体的行为或结果变量， X 是解释变量（可以是个体的特征或行为）， g 代表个体所在的群体（如班级、俱乐部等）， $E(Y|g)$ 和 $E(X|g)$ 分别表示在群体 g 内个体行为 Y 和解释变量 X 的平均水平。

- 模型的参数 β 、 γ_1 、 γ_2 和 γ_3 代表不同的效应，其中 β 是内生效应参数，表示个体行为如何受到群体行为平均水平的影响； γ_1 和 γ_2 是解释变量的外生效应参数； γ_3 是群体特征的影响； u 是误差项。
- "同伴"（Peers）的定义往往不是很明确，但通常指的是如班级、俱乐部等具有共同特征或共同目标的群体。群体成员可能在年龄、兴趣、地理位置或社会经济地位等方面具有相似性。在分析同伴效应时，研究人员关注的是这些群体内部成员间的相互作用及其对个体行为、态度或成就的影响。
- 在教育回报率中，同伴效应指学生在班级环境中相互影响的过程，对个体行为的影响

互为因果

曼斯基关注的一个关键问题是群体中的同伴 (peers) 效应，并指出在实证研究中通常难以明确定义什么是同伴群体。尽管同伴群体通常是指如班级、俱乐部这样的集体，但这些群体的边界和构成在现实世界中可能并不清晰。

进一步，曼斯基展开了模型的简约形式 (Reduced Form, RF)：

$$Y = \gamma_1 / (1 - \beta) E(X | g) + (\gamma_2 / (1 - \beta)) X + (\gamma_3 / (1 - \beta)) g + \tilde{u}$$

该方程形式揭示的是：在考虑内生的社会效应时，个体行为 Y 如何受到个体特征 X 、同伴效应 $E(X|g)$ 和群体特征 g 的综合影响。通过这种形式，曼斯基试图提供一种方法来估计内生社会效应的大小，同时指出在存在内生性时进行准确估计的困难。

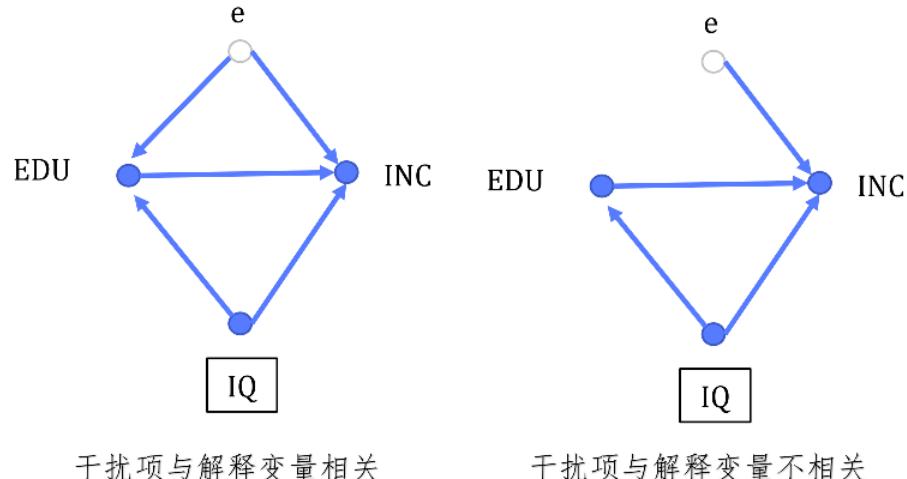
曼斯基的这篇论文为理解和分析社会互动效应提供了一个重要的理论框架，同时也指出了在实际应用这些模型时面临的挑战，特别是如何准确识别和量化这些效应。这项工作对后来的研究产生了深远的影响，促进了对社会互动和个体决策之间复杂关系更深入的理解和分析。

小结: 偏误类型与解决办法

有向无环图

- 教育回报率
- 根据理论将总体模型设定为：

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



- **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干预变量
- 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:

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

- 对EDU求偏导: **偏回归系数**

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

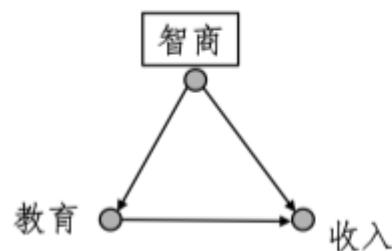
β_1 表示在IQ固定不变, INC 的期望值 (均值) 随 EDU 如何变化

有向无环图表达偏误类型

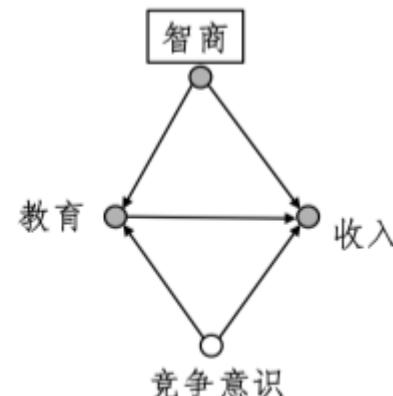
- 因果路径
- 混淆路径: $A \leftarrow B \rightarrow C$, B是A和C的混淆变量; 混淆变量会产生相关关系 \rightarrow 必须控制
- 对撞路径: $A \rightarrow B \leftarrow C$, B是A和C的对撞变量; 对撞变量不会产生相关性, 但若错误地控制了就会产生相关关系
- 估计X与Y的因果关系的本质是找到二者间所有的因果路径, 同时排除二者间所有非因果关系路径

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

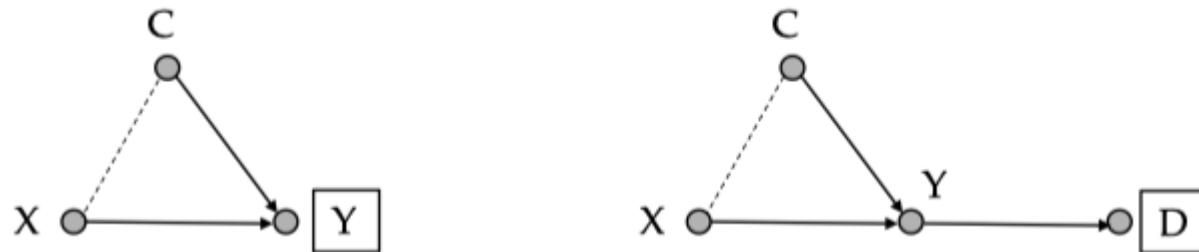


图：过度控制偏差

- 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

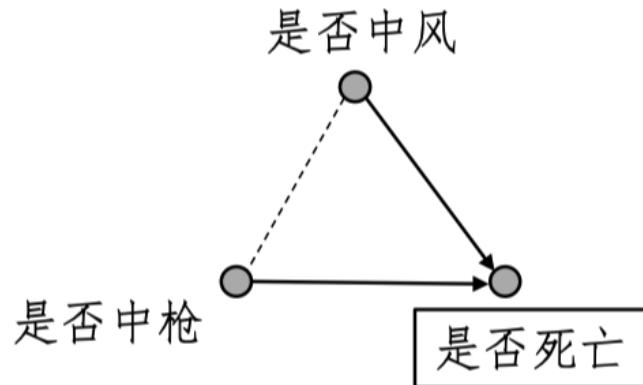
对撞偏误

- 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

在面对社会科学研究中的偏误时，我们可以采取一系列步骤来识别和解决这些问题。这些步骤包括：

1. 深入了解数据生成过程
2. 寻找偏误的潜在来源
3. 建立有向无环图 (DAG)
4. 寻找对应的解决办法

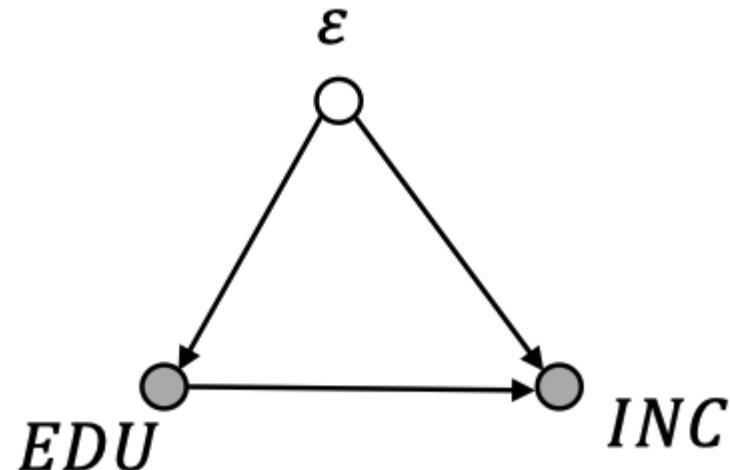
教育回报率的例子

考虑一个简单的案例，我们想要估计教育对收入的影响：

- 假设我们最初通过下面线性总体模型 (LPF) 来估计教育的效应：

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

- 当可观测变量（如性别、年龄）和不可观测变量同时进入误差项 ε_{it} 时，导致我们无法准确识别因果影响系数 β ：



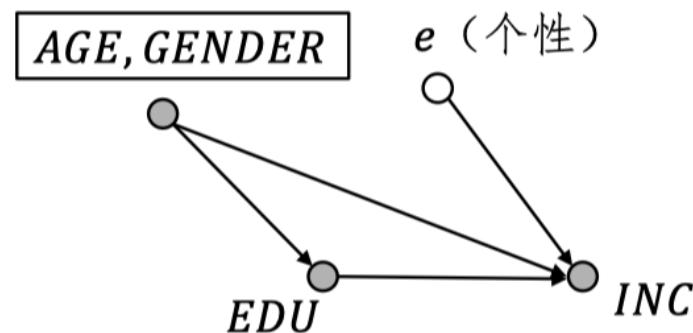
解决办法

为了解决这个问题，我们可以将误差项 ε_{it} 中的可观测变量分离出来进行控制：

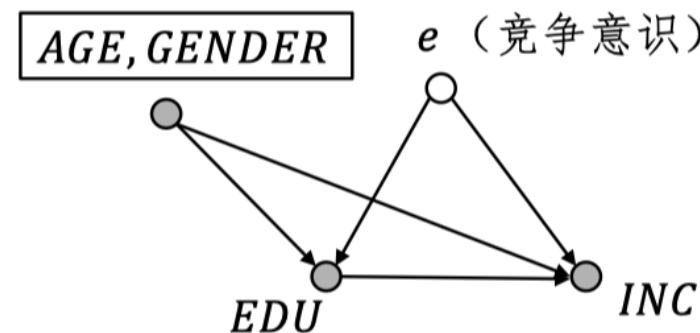
- 通过将年龄和性别作为控制变量纳入模型，我们可以得到改进的模型：

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

- 这样， e_{it} 就只包含不可观测的变量（如个性和竞争意识），从而帮助我们更准确地估计教育的真实效应。
- 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

在面对复杂的因果关系时，特别是当结果变量与干预变量可能存在互为因果的关系时，我们需要采取更精细化的方法来识别和估计因果效应。

进一步分解误差项

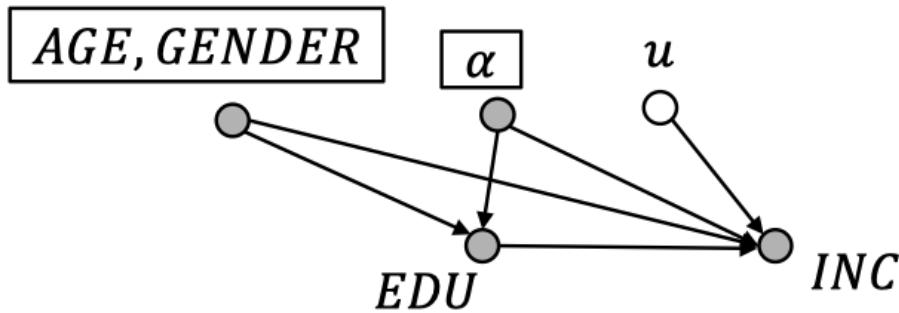
一个有效的策略是将误差项 ε_{it} 进一步分解为不可观测的非时变变量 α_i 和不可观测的时变变量 u_{it} ，即：

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

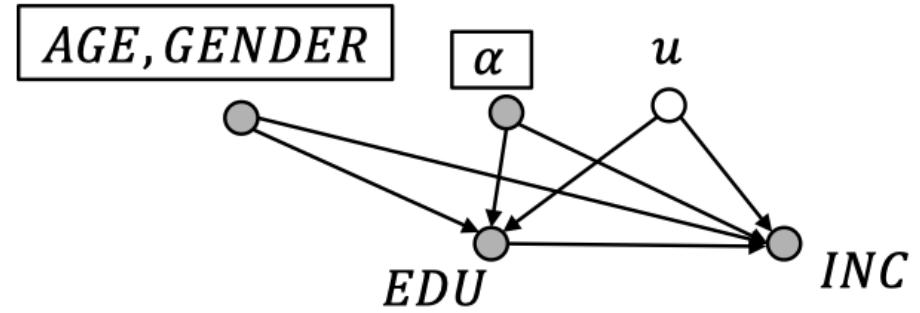
这样做的目的是尝试隔离和控制那些可能导致偏误的不可观测因素。

控制不可观测的非时变变量

- 如果我们认为混淆路径主要是由 α_i 造成的，那么我们可以通过控制 α_i 来截断这些混淆路径。
- 面板数据分析法提供了一种有效的方式来达成这一目标，通过考虑数据中的时间维度来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

在社会科学研究中，尤其是在处理可能存在内生性问题的情形时，分离可观测和不可观测变量对于确保因果推断的准确性至关重要。然而，实际上，很难直接在误差项 ε_{it} 中区分这两类变量。这就需要我们采取一些特别的方法来解决这个问题，其中引入工具变量 Z_i 是一种有效的策略。

引入工具变量

工具变量 Z_i 的引入旨在帮助分解干预变量\$EDU\{it\}\$中与误差项\$e\{it\}\$无关的部分，即：

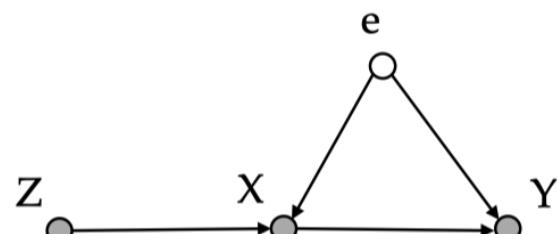
$$EDU_{it} = \widehat{EDU}_{it} + v_{it}$$

这里的 \widehat{EDU}_{it} 表示 EDU_{it} 中与 e_{it} 无关的部分，而 v_{it} 则是剩余的部分，可能与 e_{it} 相关。通过这种分解，我们能够更清晰地识别出 EDU_{it} 对结果变量 Y 的影响，而不被 e_{it} 的混淆所影响。

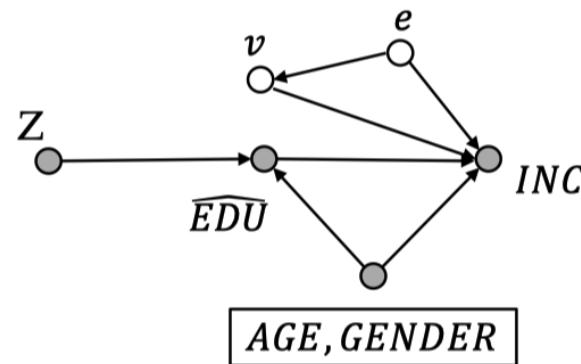
引入工具变量

工具变量需要满足两个主要条件：

1. **外生性**：工具变量 Z_i 必须与误差项 e_{it} 无关，这意味着它不受模型中未观察到的变量的影响。
2. **相关性**：工具变量 Z_i 必须与干预变量 EDU_{it} 有统计上的相关性，但这种相关性不通过误差项 e_{it} 产生。



图：工具变量的相关性和外生性



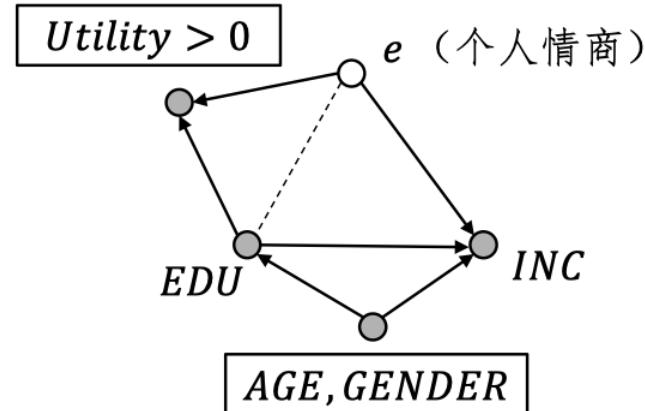
图：引入工具变量的情况

选择偏误 (Selection Bias)

指在选择研究样本的过程中由于非随机选择导致的研究结果系统性偏差。选择偏误可以以多种形式出现，主要包括以下几种：

1. **自我选择偏误 (Self-Selection Bias)** : 发生在研究参与者能够自主决定是否参加研究的情况下。参与者的小数是主动的，基于他们自己的特征、偏好或行为。这导致了参与研究的样本可能在某些关键特征上与那些选择不参与的个体有显著差异
2. **样本选择偏误 (Sample-Selection Bias)** : 发生在从总体中选择样本的过程中由于非随机选择而引起的偏差。主要特点是选择过程受到某种机制的影响，导致最终的样本不能准确反映整个总体的特性：研究设计的限制、数据收集方法等问题导致某些观测值更容易被包括或排除在样本之外
 - 假设想评估某种培训或教育对个体收入的影响。然而，如果我们的样本仅仅包括了选择参加工作的个体，那么我们的样本并不是从总体随机抽取的。这种情况下，选择参加工作的决定（由"Utility"变量表示）可能与个体的教育水平和其他未观察到的因素（如个人能力、动机等）有关，从而导致样本中的干预变量（如教育水平）和不可观测因素 e 之间存在相关性。这种情况下，就是因样本选择产生了**对撞偏误 (collider bias)**

解决方法



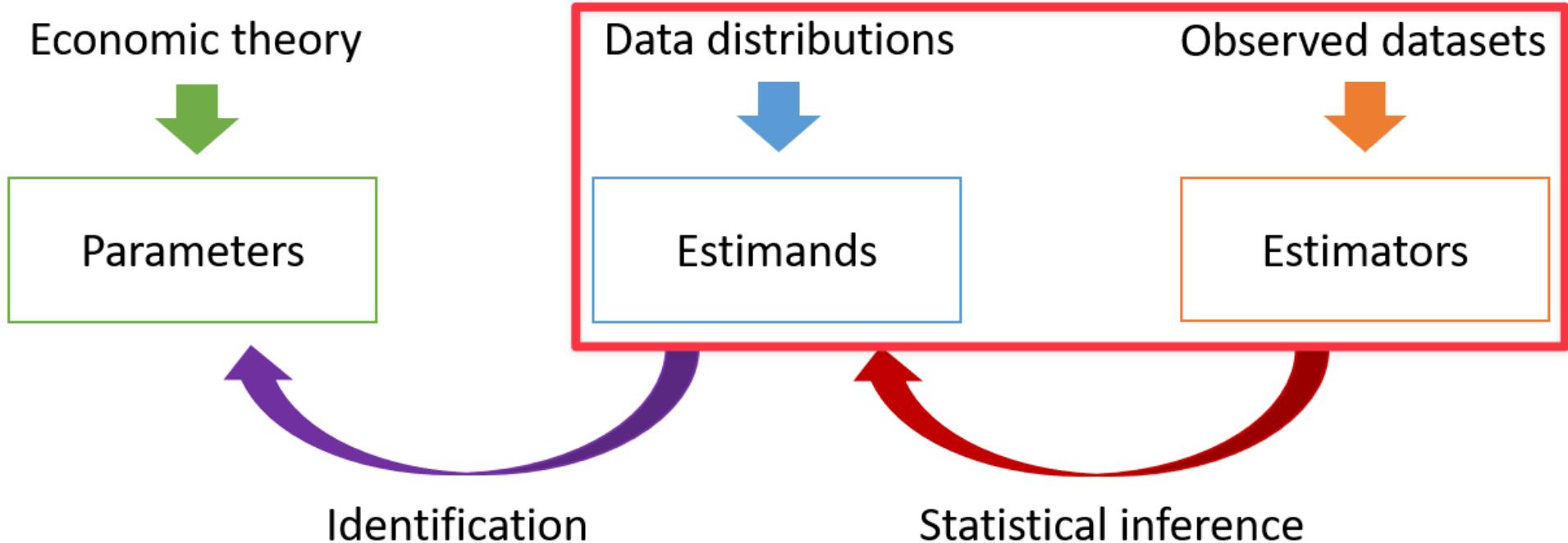
- 尽管两者都会影响研究结果的代表性和准确性，但解决这两种偏误的策略可能不同。
- **自我选择偏误**，可能需要采用如 **使用倾向得分匹配**通过估计个体选择参加工作的概率（倾向得分），然后在选择参加工作和未选择参加工作的个体之间进行匹配，以创建一个在观察到的协变量上平衡的样本
- **样本选择偏误**，需要设计更加周密的抽样方案或统计方法来调整选择过程中的偏差：
 - **采用Heckman两步法**。第一步使用一个选择方程来建模个体选择参加工作的决策，然后在第二步的收入方程中加入从第一步得到的逆米尔斯比率（Inverse Mills Ratio）作为控制变量，以纠正选择偏差

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

从观测样本到客观事实
→ 避免 Garbage in, Garbage out

理论、总体分布与样本



假设检验和统计推断

是统计学中两个基本且互相关联的概念。它们共同构成了科学的研究中推断总体参数的基础

假设检验

- 假设检验是一种统计方法，用于判断样本数据是否支持对总体参数的某个假设。通常涉及两个假设：零假设（ H_0 ）和备择假设（ H_1 ）
- 零假设通常表示为没有效应或没有差异，而备择假设表示有效应或有差异。

步骤：

- 设定零假设和备择假设。
- 选择合适的检验统计量。
- 确定显著性水平（通常是0.05）。
- 根据样本数据计算检验统计量和p值。
- 根据p值决定是否拒绝零假设。

假设检验和统计推断

统计推断

在进行经济学或社会科学研究时，我们经常面临从样本数据中推断总体特征的需求。统计推断提供了一种方法，让我们能够基于样本信息来估计和推断总体参数（如平均值、比例或回归系数等），并评估这些估计的可靠性。

- 统计推断的目标是基于样本数据对总体进行可靠的推断
- 点估计：提供一个单一的数值作为总体参数的估计
- 区间估计：提供一个区间，预计总体参数会落在这个区间内。最常见的是置信区间

假设检验和统计推断

统计推断

线性投影模型 (LPF)

在引入统计推断的背景下，我们通常从一个确定的模型开始，如线性投影模型 (LPF)，它假设一个变量（比如个体收入 Y ）可以通过一组其他变量 (X) 的线性组合来预测，同时考虑到一个误差项(e)，这个误差项代表了除 X 之外的其他所有影响 Y 的因素

$$Y = X'\beta + e, \quad E(e | X) = 0$$

目标是使用最小二乘法 (OLS) 来估计模型中的系数 β ，这些系数告诉我们变量 X 如何线性影响变量 Y

统计推断的关键

为了进行有效的统计推断，我们需要理解样本估计值与总体真值之间的关系。

- 大样本理论告诉我们，随着样本量的增加，通过样本计算得到的系数估计值会越来越接近总体真实的系数值。
 - **样本量**: 在 $n > 200$, 样本估计值 $\hat{\beta}$ 是总体估计值(真实值)的 β 的一致估计($\text{plim } \hat{\beta} = \beta$)
- 标准误差给出了估计值可能的变异范围
 - 较小的标准误差意味着我们的估计值更加接近总体真实值

功效分析

除了经济领域注重识别过程外，通常研究者会进行功效分析来确定在**预期效应大小**和所选**显著性水平**下，达到足够统计功效所需的**最小样本量**。这有助于确保研究设计可以有效地检测到感兴趣的效应，同时控制研究的成本和可行性。

统计功效

以教育回报为例：为确保研究设计有足够的统计功效，即：假如教育真的对收入有显著影响，我们能够检测到这一点。反之，我们的研究设计没有足够的统计功效，那么我们可能会错误地得出教育对收入无显著影响的结论

- 这就好比我们试图听到远处的轻声细语，但周围的噪音太大，我们的“听力”（即统计功效）不足以分辨出那些微弱的声音

样本大小

- 样本大小起到了决定性作用。
- 如果我们只调查了很少数的毕业生，那么得出的结论可能会因为样本太小而不具代表性
 - 试图通过观察一个小村庄的天气来推断整个国家的气候一样不可靠。增加样本大小，相当于增加了研究“听力”，让我们更有可能准确捕捉到教育对收入的影响

显著性水平

- 显著性水平决定了我们愿意接受多大风险，错误地拒绝零假设（事实是没有效应，但拒绝了）
 - 设定显著性水平，就像是设置一个“听力测试”的通过标准，决定了什么样的结果我们认为是显著的
 - 如果我们设置的显著性水平过于严格（例如 $\alpha=0.01$ ），那么我们可能错过一些真正的效应，因为我们要求的证据过于强硬
 - 相反，如果显著性水平过于宽松（例如 $\alpha=0.10$ ），我们可能会错误地认为教育对收入有显著影响，尽管实际上这种影响可能只是随机变化的结果。

教育回报率的情景化例子

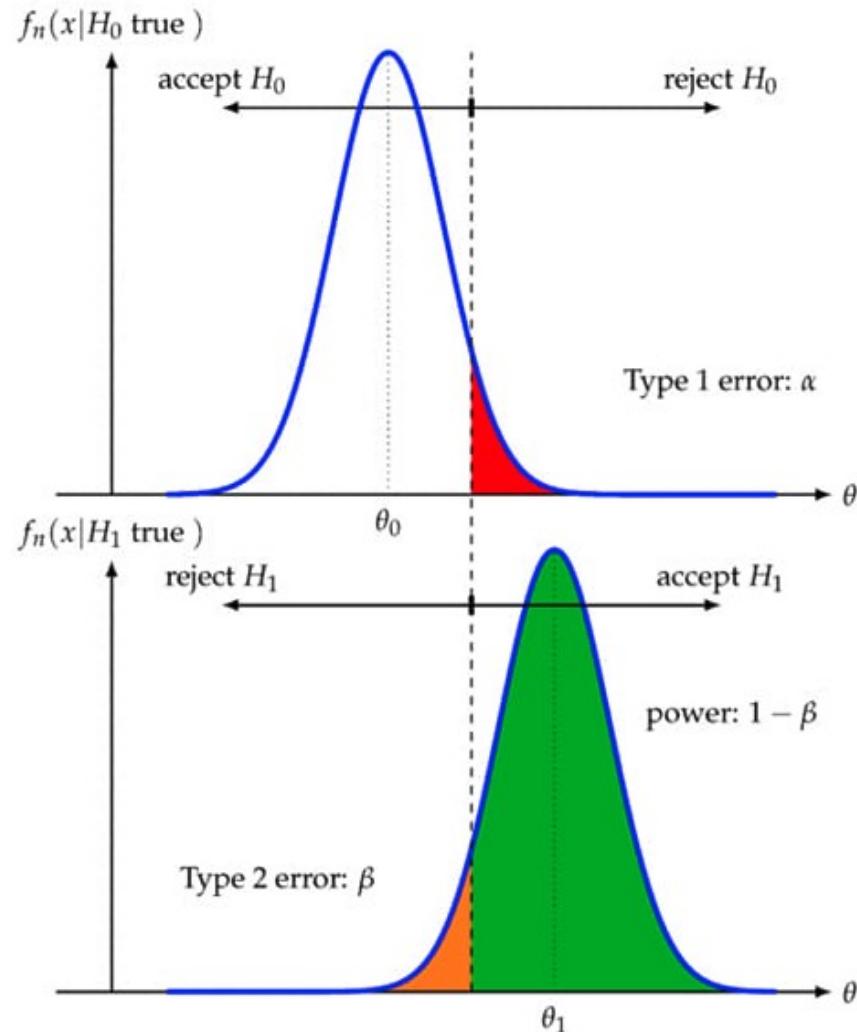
- 想象我们正在研究一个大型的数据集，包括成千上万的个人，旨在分析高等教育对个人收入的影响。
- 我们决定设置显著性水平为0.05，意味着我们愿意接受5%的错误拒绝真实无效果的零假设的风险
- 通过计算得知，为了确保80%的统计功效能够检测到教育对收入的最小显著影响，我们需要至少1000名参与者的数据
- 通过这种设计，如果我们发现高等教育与收入增加之间有显著的相关性，就可以相对自信地说，这种关系不太可能是偶然发现的。
 - 这就好比我们通过大量数据和合理的“听力测试”标准，准确地听到了教育对收入提升这一细微却重要的“声音”

第I型错误 (False Positive) → 设定显著性水平 α

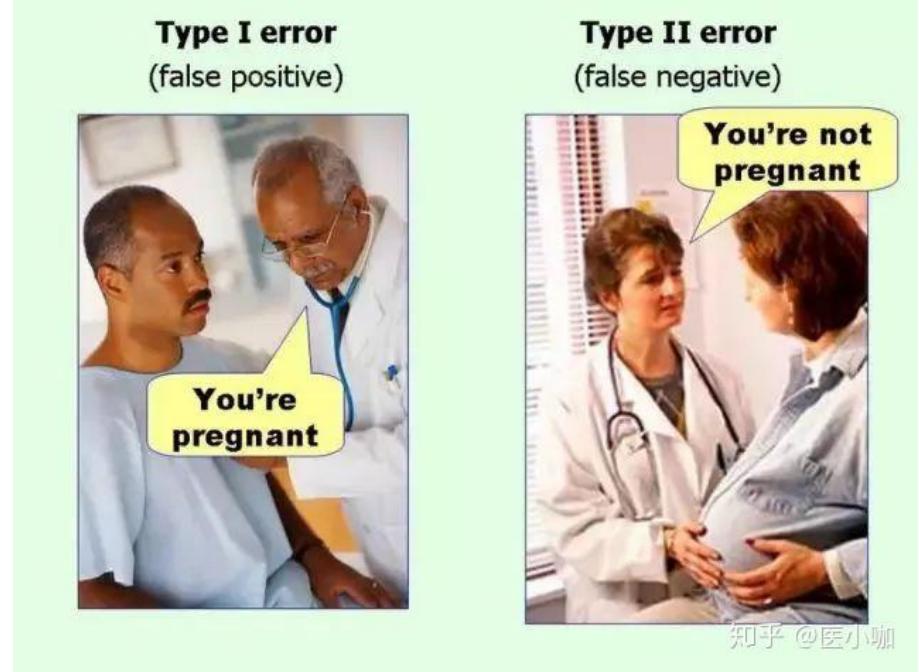
- 第I型错误，或称为假阳性，发生在错误地拒绝了一个实际为真的零假设。假设正在研究是否更高的教育水平会导致更高的收入。零假设 (H_0) 是教育水平不影响收入，即教育回报率为零。如果实际上教育并不影响收入，但的研究错误地得出“教育水平提高了收入”的结论，那么就犯了第一型错误。
- 想象这就像是误判一位无辜的人有罪。设置的显著性水平（比如5%或0.05）就是愿意接受犯第一型错误的最大概率。

第II型错误 (False Negative) → 统计功效 $1 - \beta$

- 第II型错误，或称为假阴性，发生在错误地没有拒绝一个实际为假的零假设。还是以教育和收入的研究为例，如果实际上更高的教育水平确实能显著提高收入，但的研究未能检测到这种影响，那么就犯了第二型错误。
- 这就像是误判一位有罪的人无辜。统计功效衡量正确发现实际效应（如果存在的话）的能力。功效越高，越有可能正确地识别出教育对收入有显著正面影响（如果真的有的话）。功效受多种因素影响，包括样本大小、效应大小、显著性水平和数据的变异性。提高样本量、增加效应大小或接受更高的第一型错误风险都可以增加统计功效。

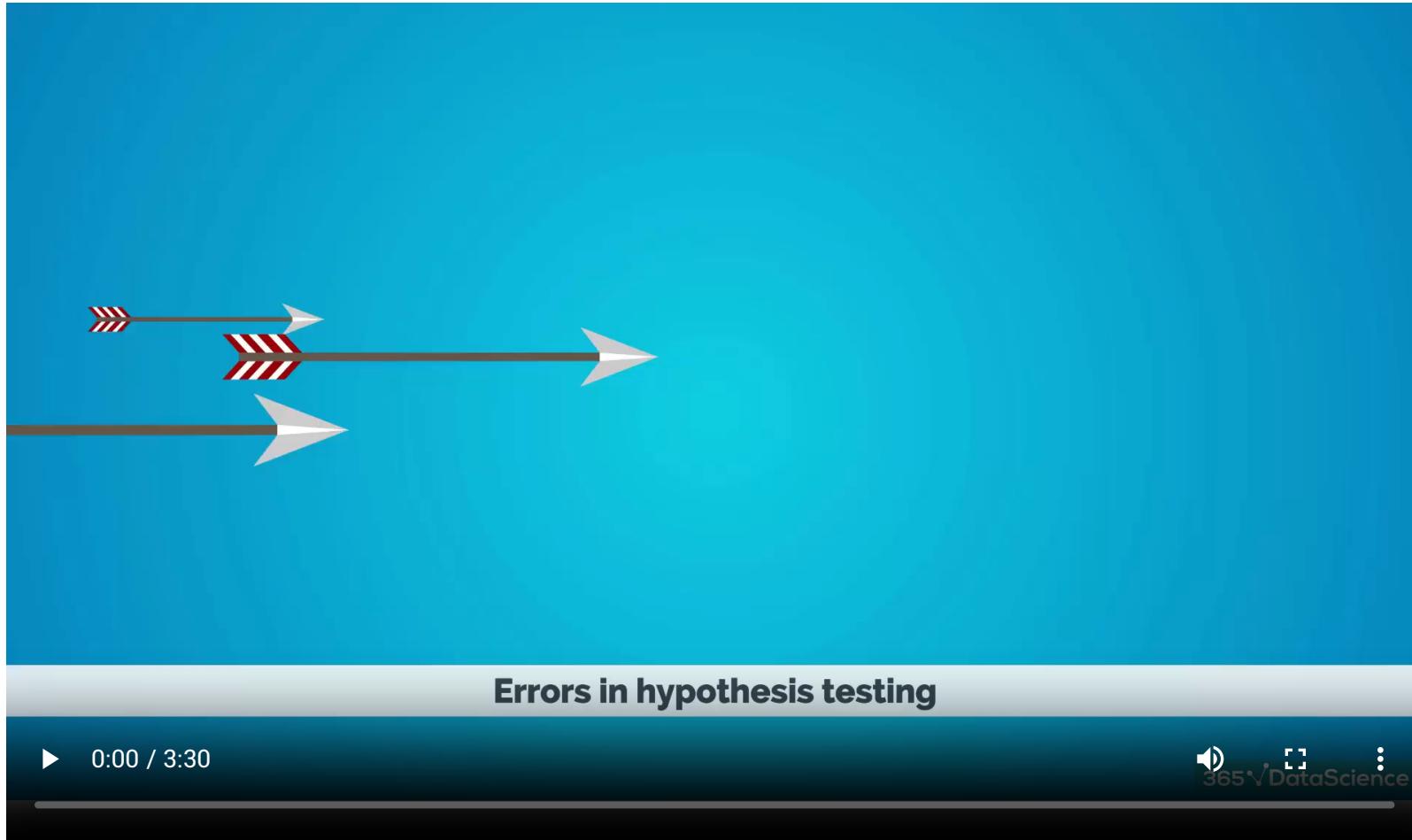


Type I error (significance level, P-value)、statistical power(sensitivity)、expected effect size、sample size



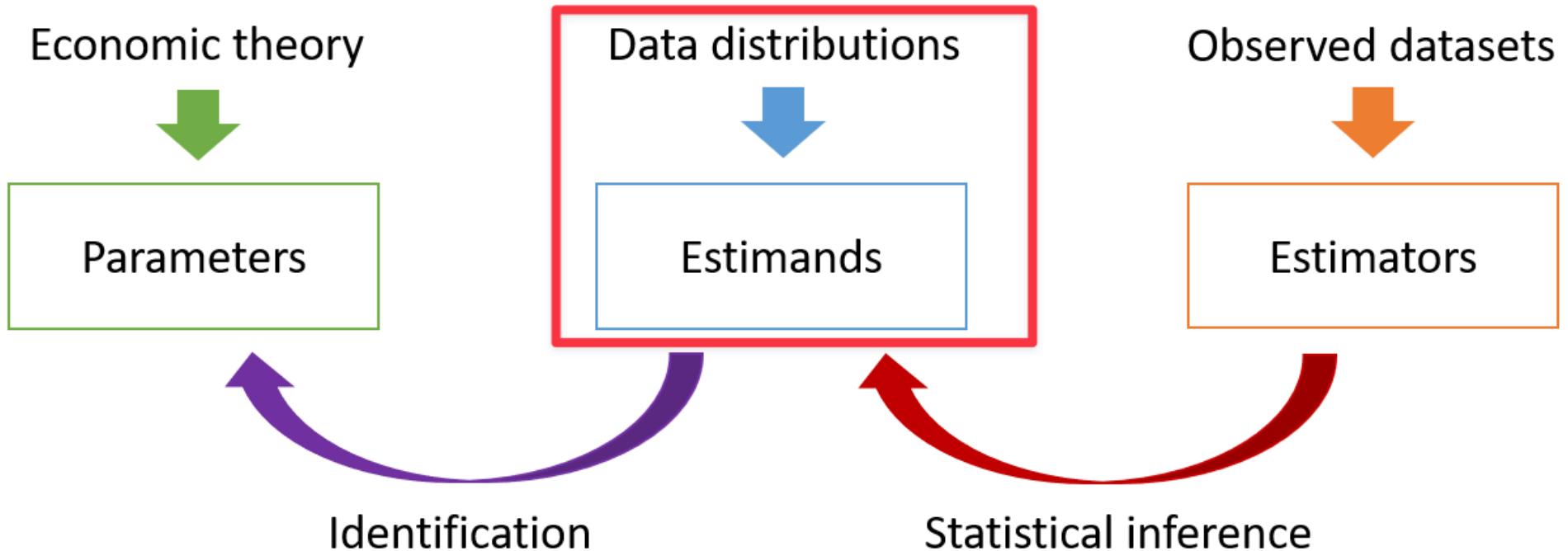
- **统计功效与样本大小**: 在其他条件不变的情况下，样本大小增加会增加统计功效，使研究更有可能检测到实际存在的效应。
- **统计功效与显著性水平**: 在样本大小固定的情况下，降低显著性水平（设置更严格的标准）会降低统计功效。换句话说，更严格的判断标准使我们更难拒绝正确的零假设，除非数据提供了更强的证据。
- **样本大小与显著性水平**: 在设计研究时，研究者通常需要在保持适当统计功效的同时选择合适的样本大小和显著性水平。这通常涉及权衡，因为更大的样本需要更多的资源。

Type I/II error



经验研究中的客观事实是什么?
→ 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 是结果变量 (结果变量 | 结果变量 | 被解释变量) ; X 是干预变量 (干预变量 | 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

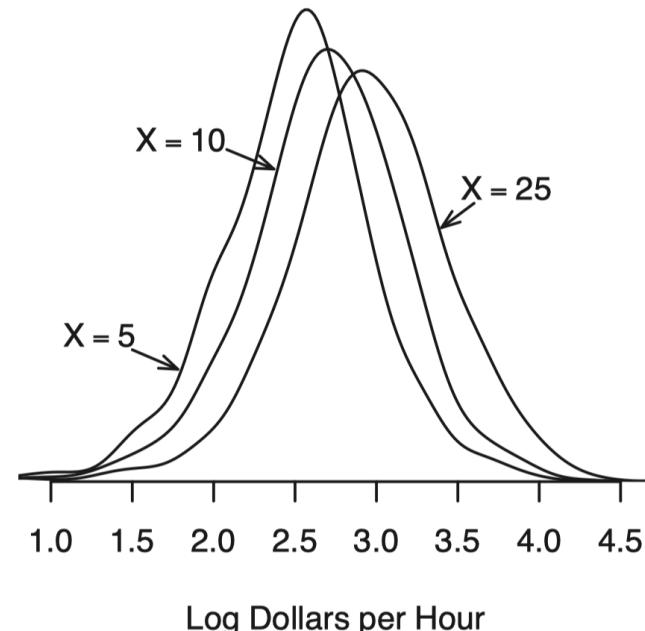
$$Var(Y) = E[V(Y|X)] + V[E(Y|X)]$$

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

特性良好且能被认知的客观事实
→ 经验研究是有边界的

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与总体模型间的关系

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：

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

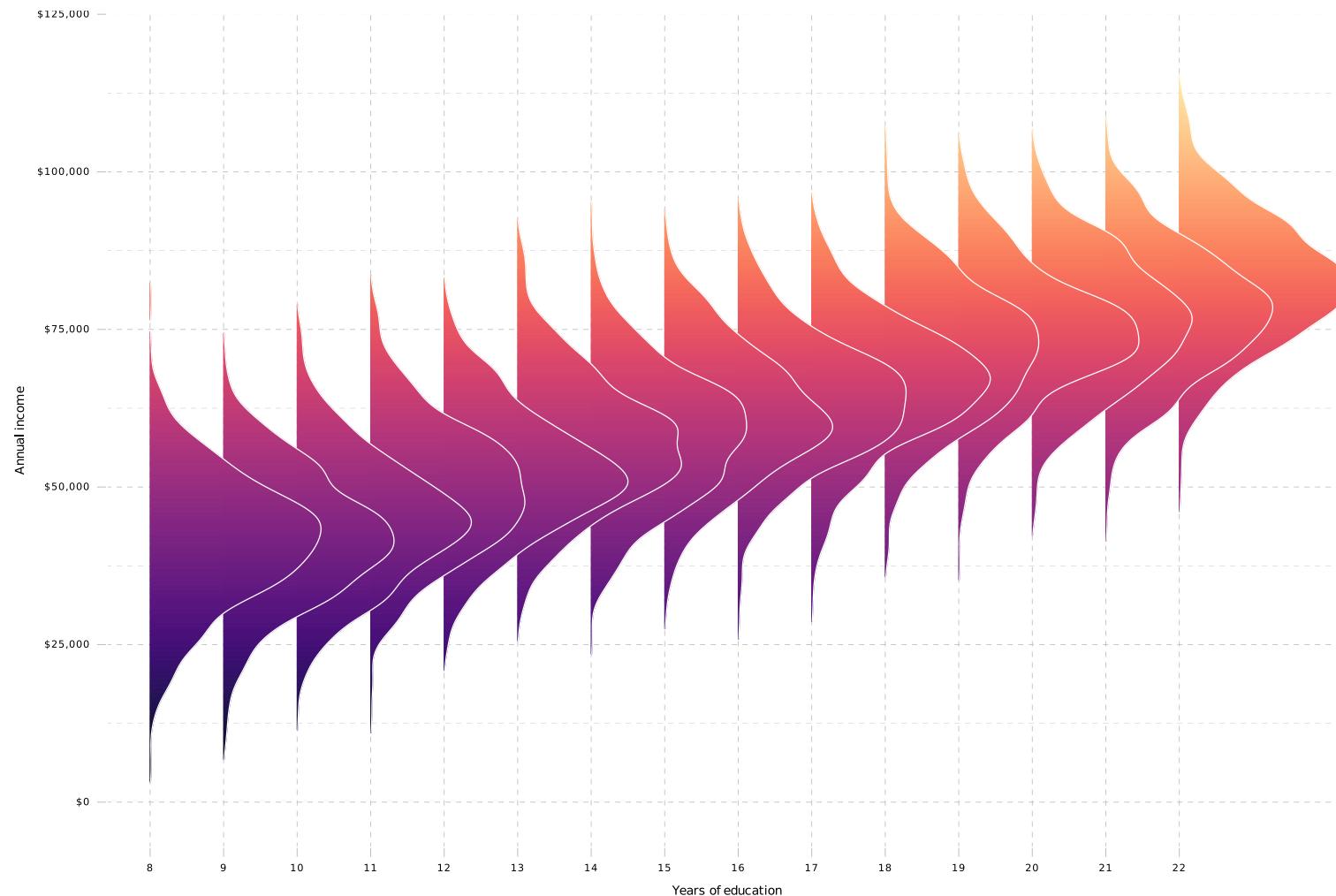
因此模型类别由 $m(x)$ 形式决定：如截距模型，线性模型，Logit模型等

CEF是从样本到总体的桥梁

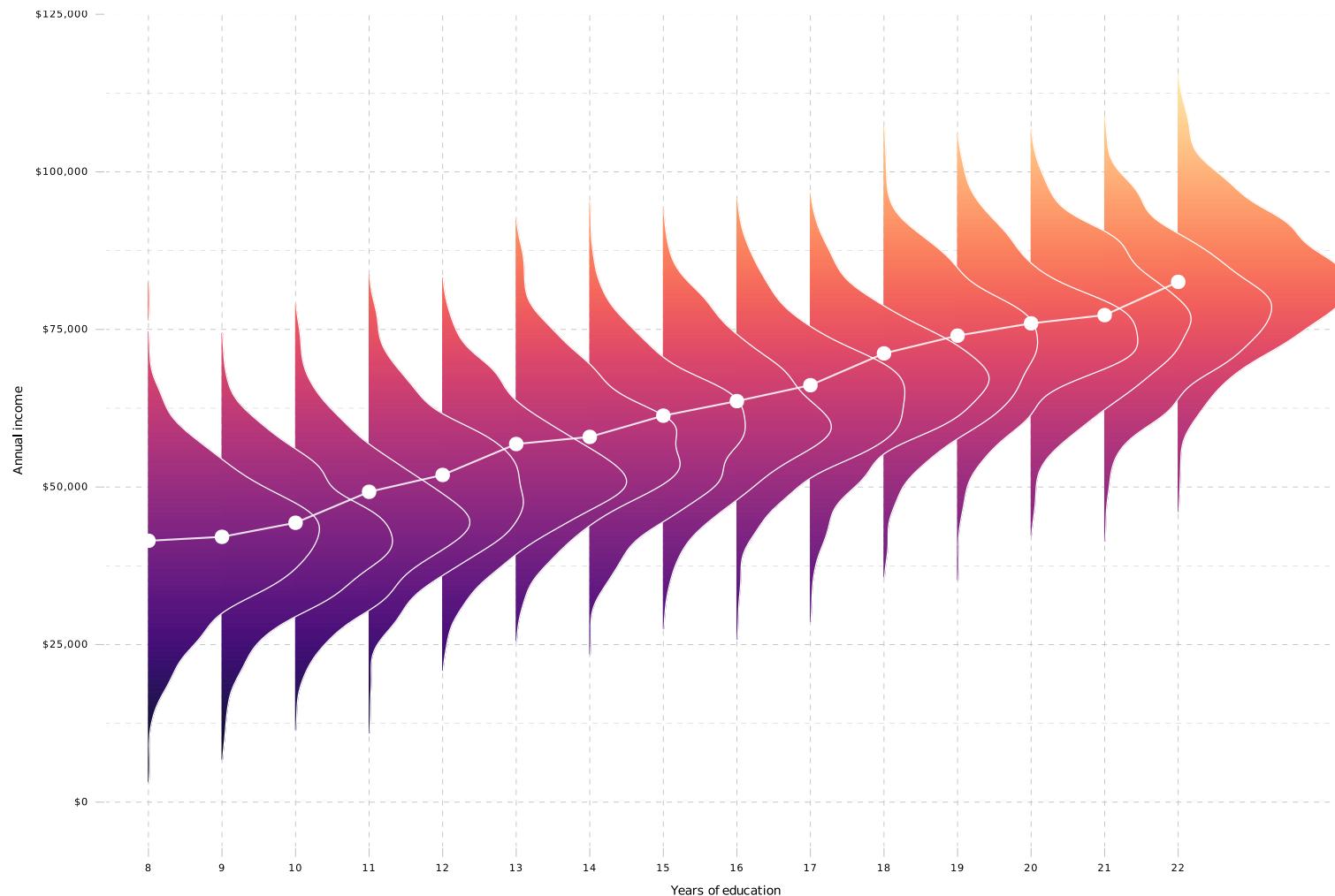
- 期望本身是总体概念（价值观）
- 实际中，我们是基于样本信息推断总体信息，例如用样本均值推断总体期望
- 将CEF写作基于样本的CEF: $E[Y_i | X_i]$

从图形上看CEF...

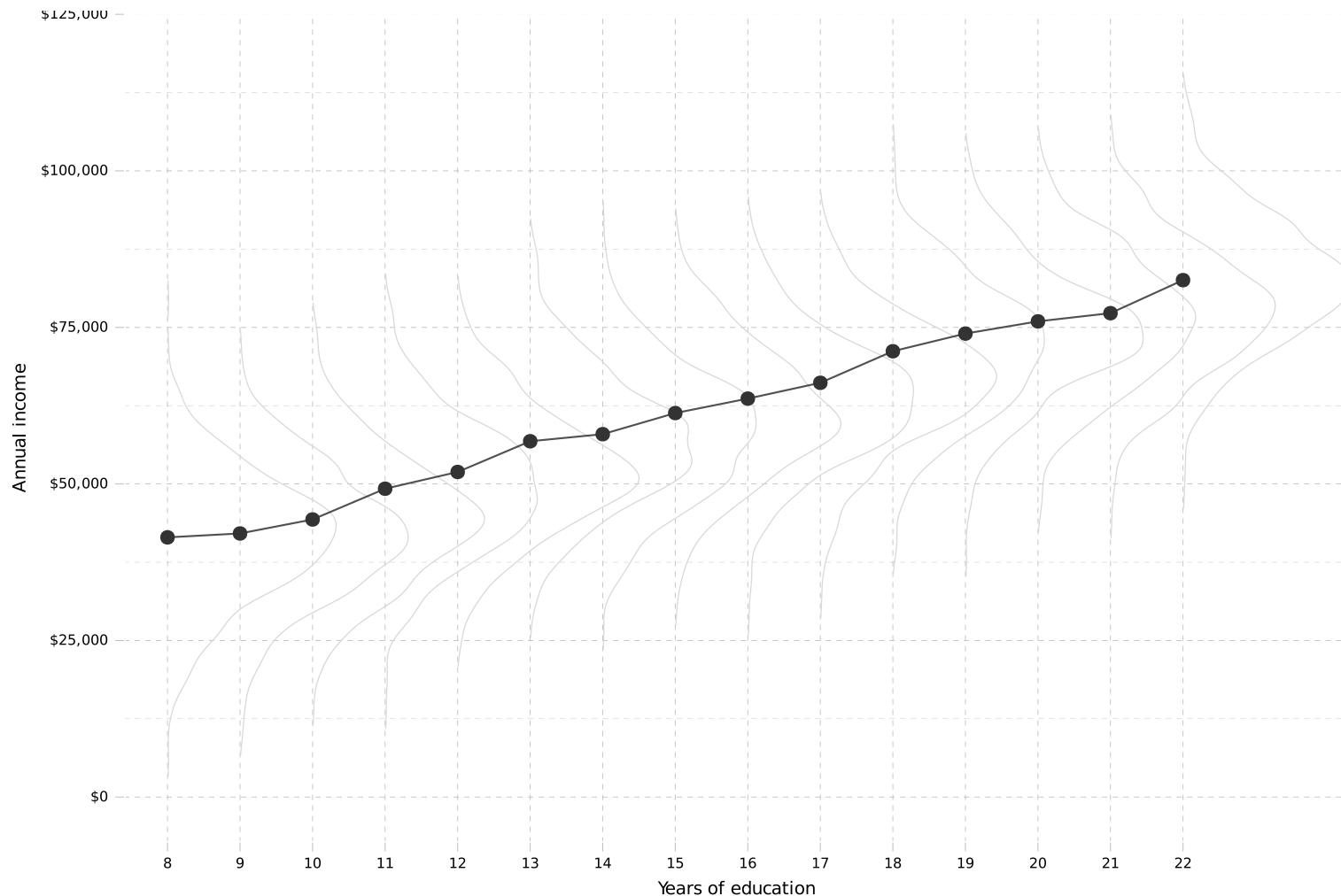
条件分布 Y_i , 对于8, ..., 22不同教育年限的 $X_i = x$.



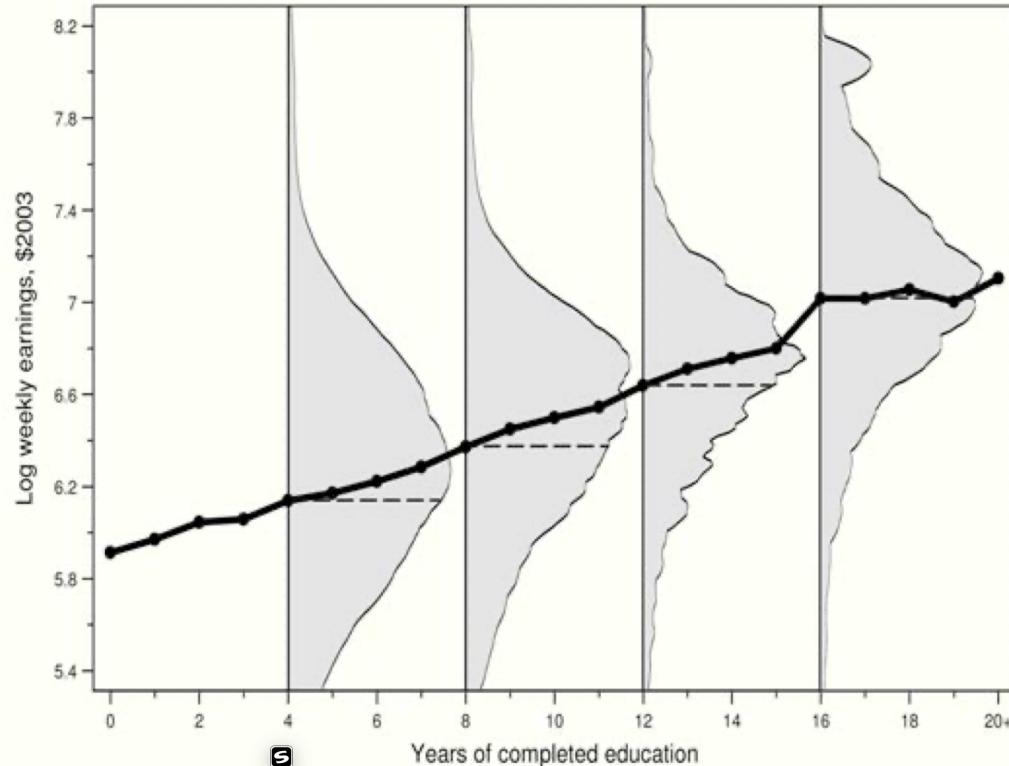
条件期望函数 $E[Y_i | X_i]$ 其实是这些条件分布的均值



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将观测的结果变量分解成两部分

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

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[†]
 - 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

[†] 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：可将结果变量 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：结果变量的变动 = 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 $E(Y \mid X = x)$

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 $g(X) = \mu(X)$.

经验研究为什么从LPF而不是CEF开始?

- **Meaningful** !! → 实证模型的建立基于的理论模型
- 线性CEF不是也有经济意义么? 未必。
→ 一个原因是: 总体模型 $m(x_1, x_2)$ 等价线性CEF形式为
$$m(x_1, x_2) = x_1\beta_1 + x_2\beta_2 + x_1^2\beta_3 + x_2^2\beta_4 + x_1x_2\beta_5 + \beta_6$$
- 转向LPF(linear projection function)
$$m^{LPF}(x_1, x_2) = x_1\beta_1 + x_2\beta_2 + \beta_3$$

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 β 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?

- 补充一个知识点: CEF还有一个非常好的性质 → 预测
- 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**

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

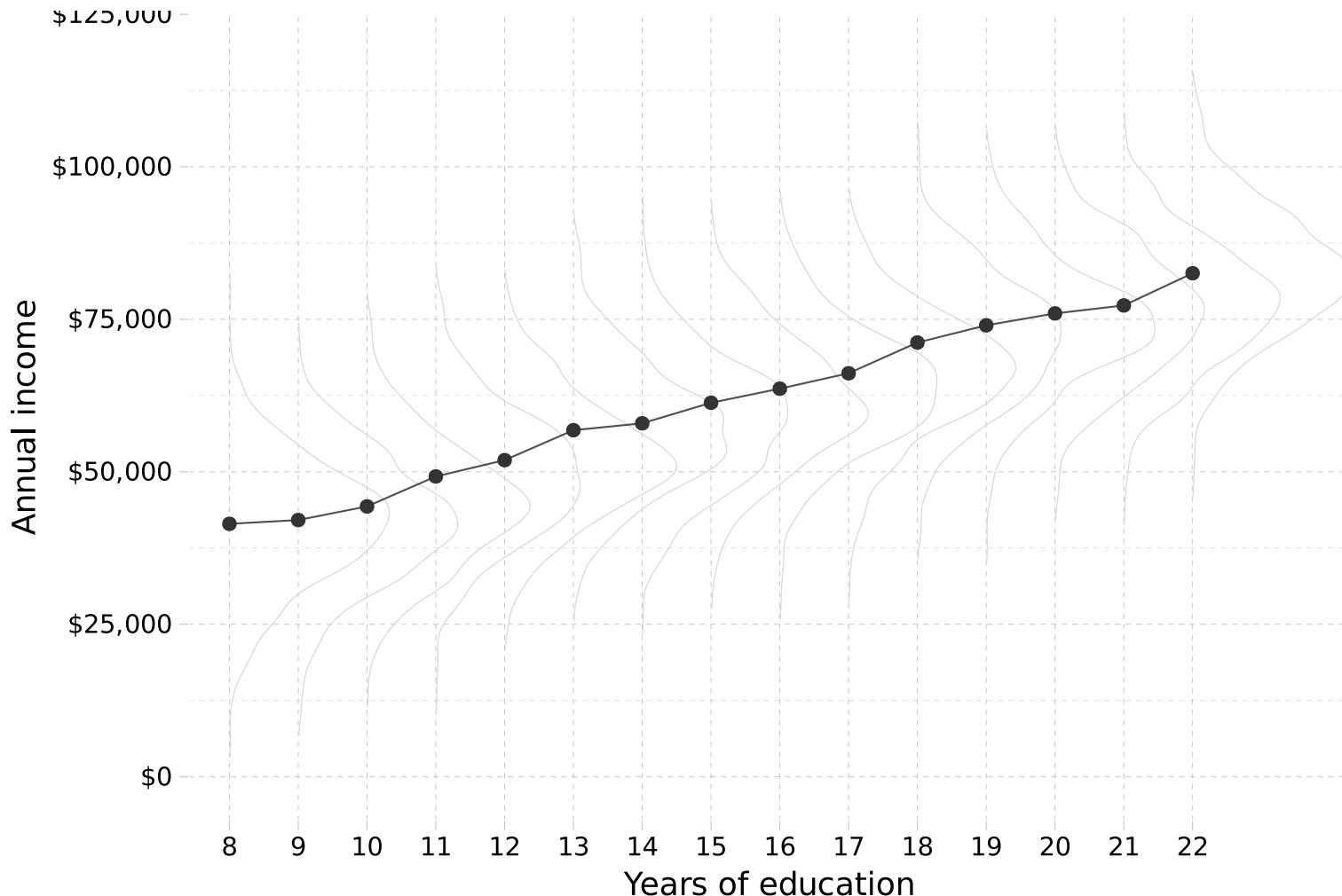
经验研究为什么从LPF而不是CEF开始?

- CEF是MMSE → LPF 也是MMSE。继续使用最小化MSE准则：

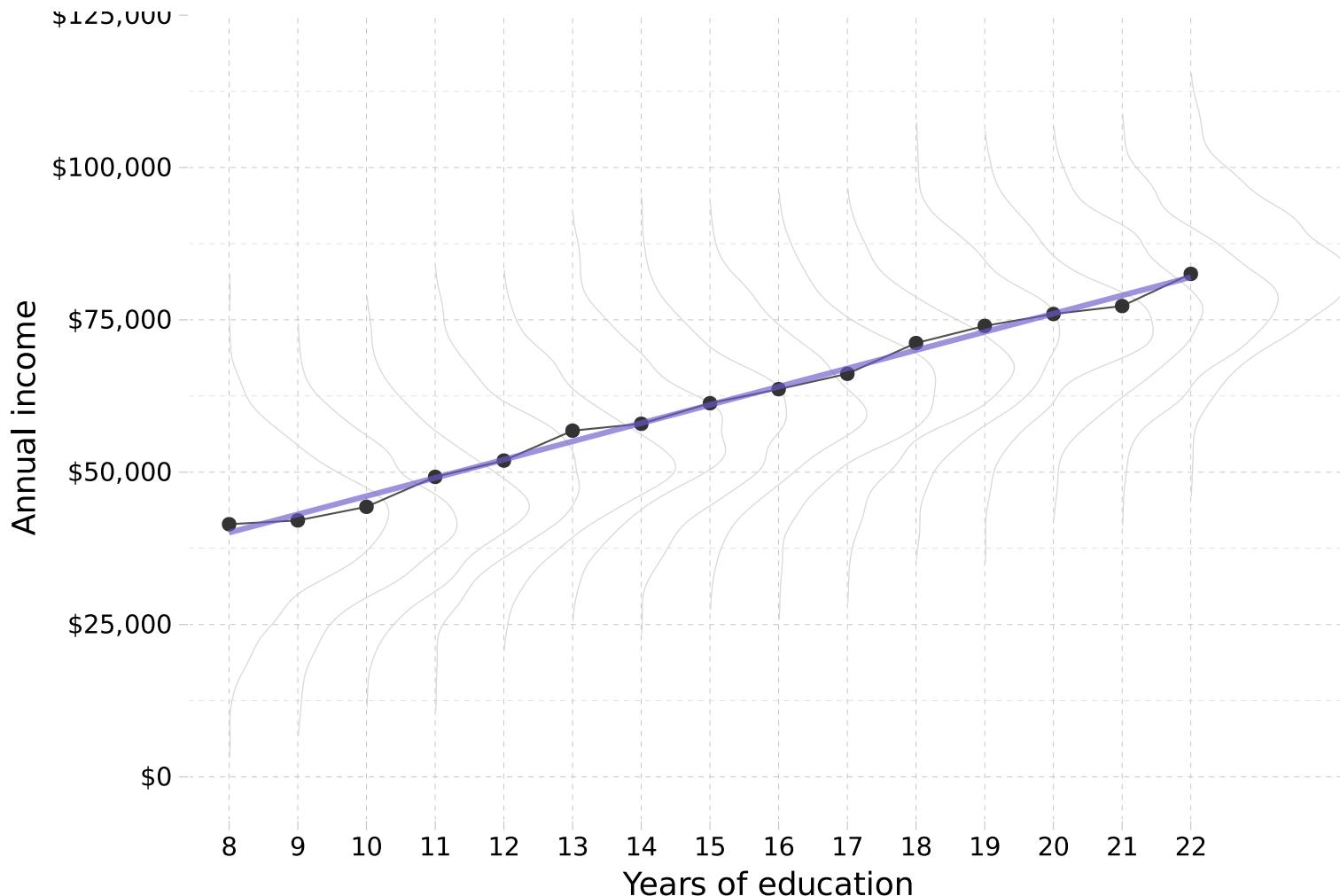
$$\beta = \underset{b}{\operatorname{argmin}} E \left[(m(X_i) - X_i' b)^2 \right]$$

- 回归与条件期望函数定理 (Regression-CEF Theorem)
- 结论：
 - LPF是CEF的MMSE(最小均方误)和BLP(最优线性预测)
 - 通常而言，CEF不一定是线性的(才需要多项式表达)
但CEF若是线性的，那么LPF就是CEF

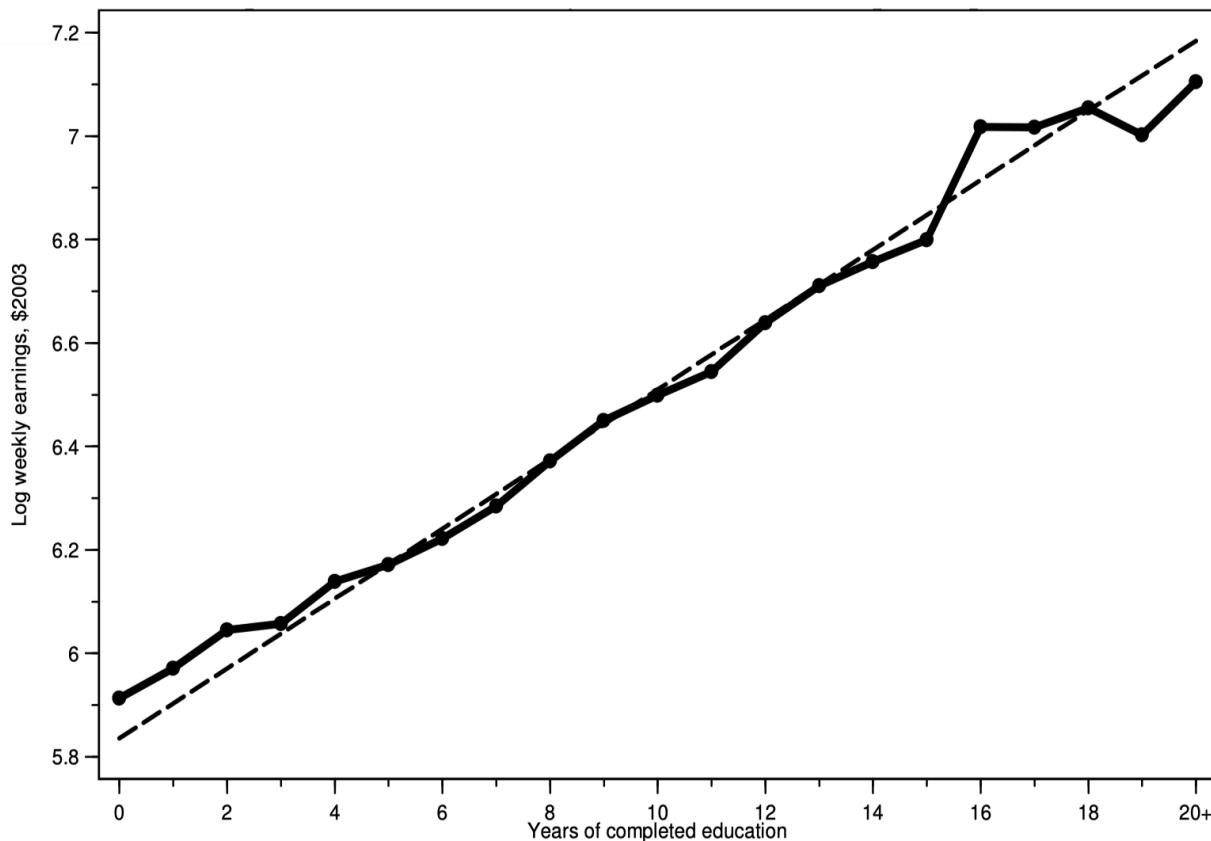
CEF



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)

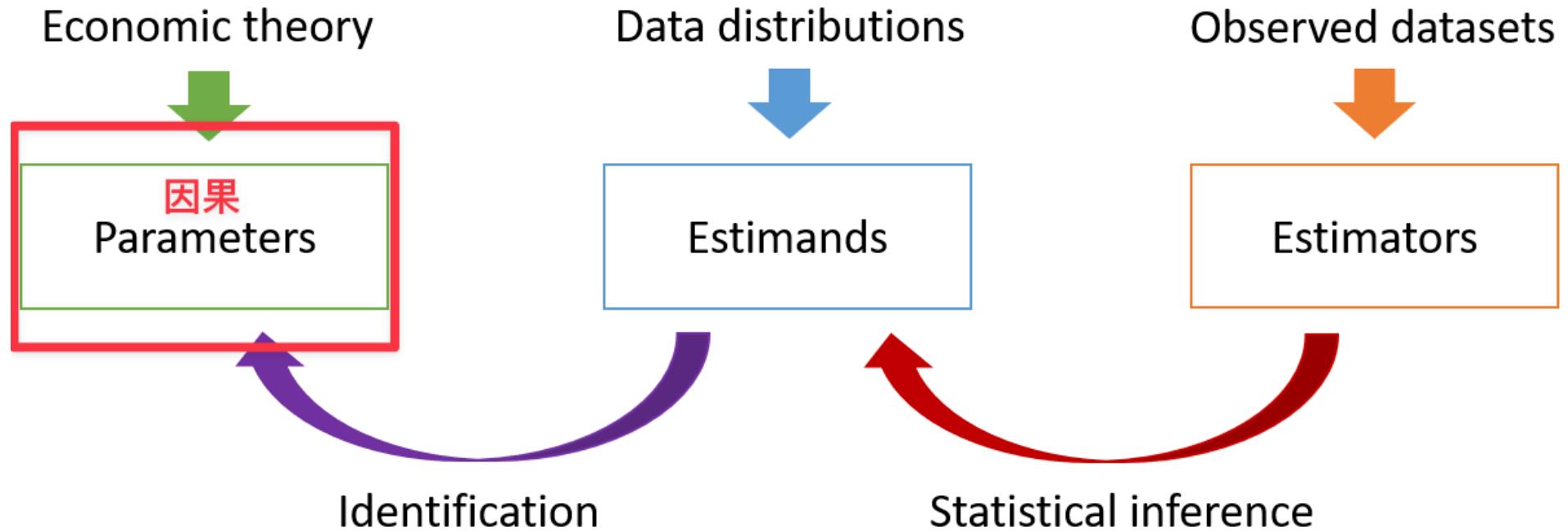
推断因果是基于CEF而不是LPF，但经验研究更多从LPF出发

- 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(高阶项)也会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干预变量，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，总体是线性CEF的形式仍然是概率事件，这就凸显了识别过程的重要性

因果模型

→ 建立在潜在结果框架之下

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

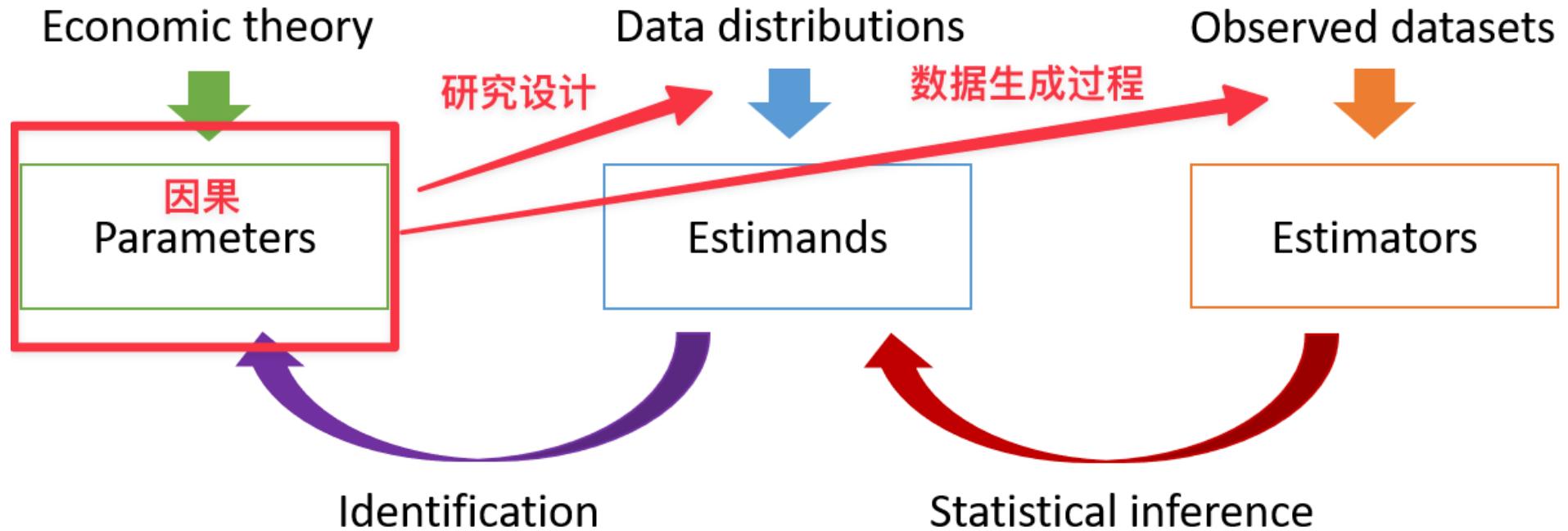
但在实际情况中，无法观测到反事实结果

- 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 τ_i 的函数

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

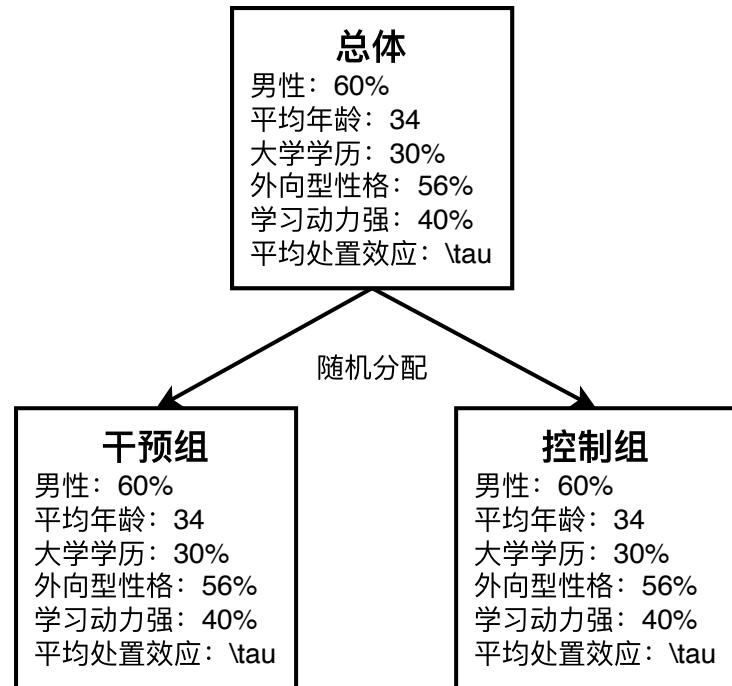
$$Y_i(1) = a + \tau_i + bX_i + e_i, D_i = 1$$

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

- 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

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

- 同理: $E[Y_i(1) \mid D_i] = E[Y_i(1)]$
- 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)**

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

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

- 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)**

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 Y_{si} 和每个人的干预剂量多少 s_i 是独立的：

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

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

- 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...

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

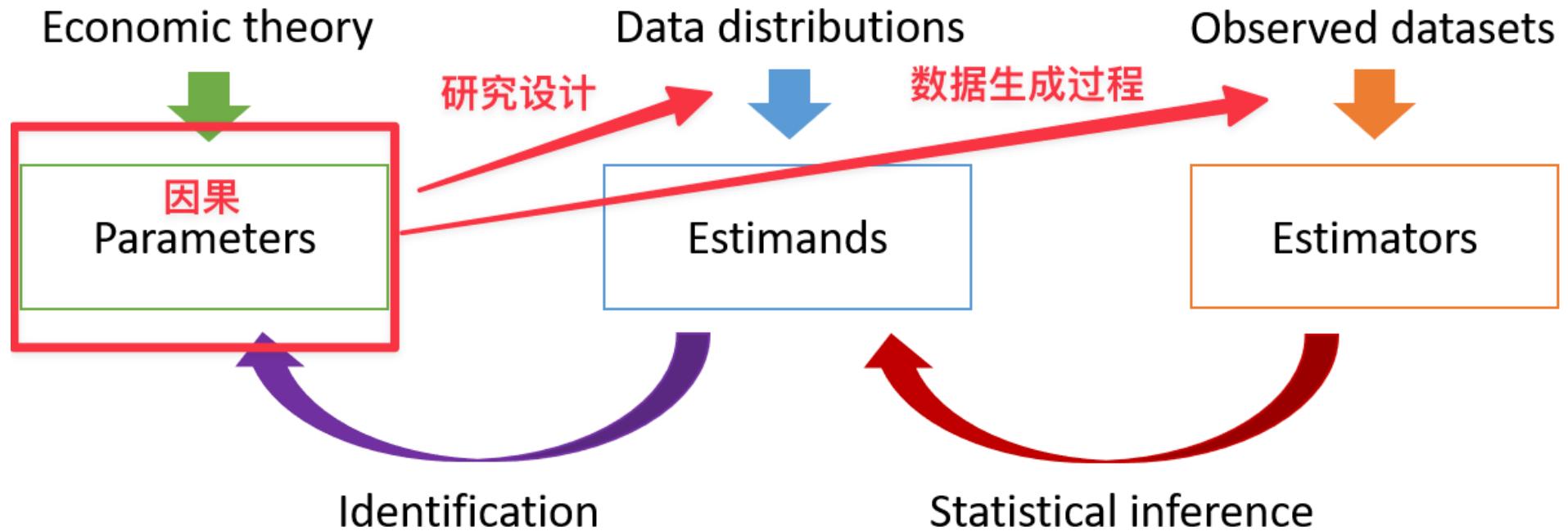
现在取两边的期望值并应用迭代期望法则(LIE)

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

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

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

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 η_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 τ 是**真实处置效应**, 而 (B) 式 τ 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：

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

- β 是 η_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[†]**" 可以正确代表线性CEF

[†] 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：

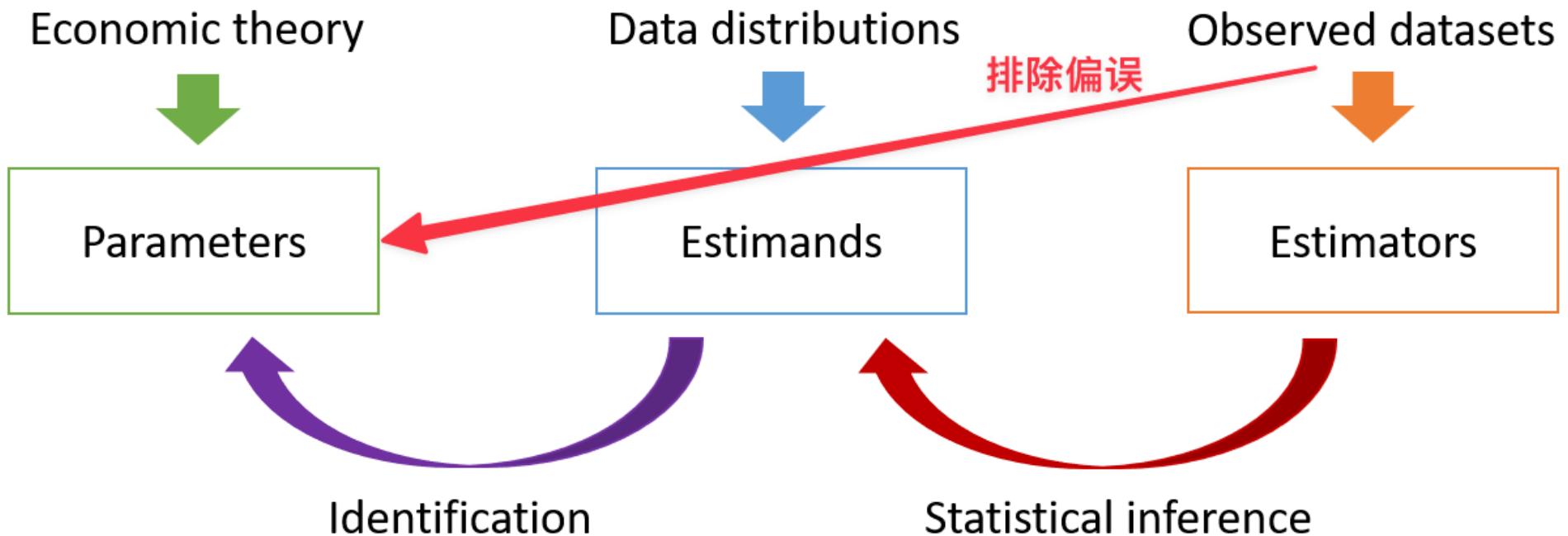
$$Y_i = \alpha + \tau s_i + X'_i \beta + \nu_i$$

- 通过假设，限制扰动项 ν_i 来估计**真正的因果效应** τ
1. s_i (根据 CIA)
 2. X_i (根据定义 β 是 η 对 X_i 回归的总体系数向量)

从观测样本到因果模型

→ 必须排除的偏误

理论、总体分布与样本



理解"朴素"估计值与LPF^{ols}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^{ols} \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 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：

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

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

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