

数据生成过程与外生性

司继春

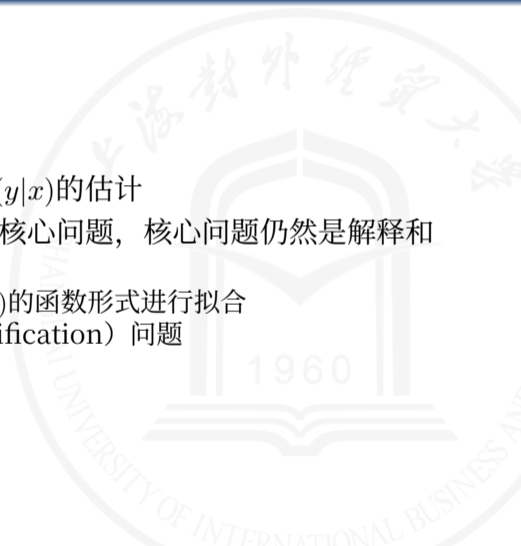
上海对外经贸大学

2026年4月



外生性

- 以上我们将线性回归理解为条件期望函数 $\mathbb{E}(y|x)$ 的估计
- 然而对于计量经济学而言，预测和拟合并非核心问题，核心问题仍然是解释和因果效应
 - 如果以拟合和预测为目的，通过设定 $\mathbb{E}(y|x)$ 的函数形式进行拟合
 - 如果以解释为目的，需要讨论识别 (identification) 问题



外生性

- 而对于模型

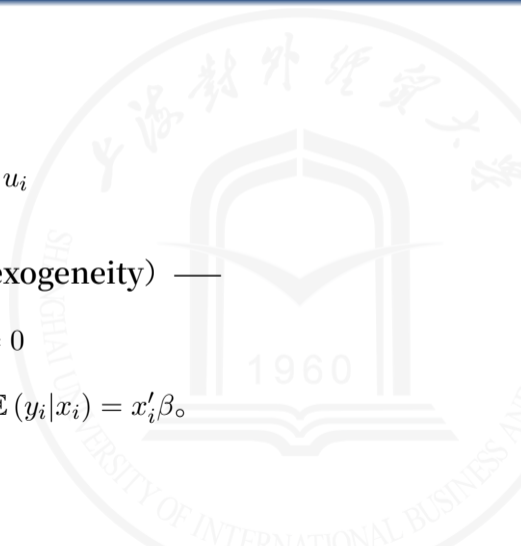
$$y_i = x_i' \beta + u_i$$

我们需要讨论何时 β 有因果效应的解释。

- 是的 β 具有因果解释的假设：**外生性假设 (exogeneity)** ——

$$\mathbb{E}(u_i | x_i) = 0$$

- 如果外生性成立，则 $\mathbb{E}(y_i - x_i' \beta | x_i) = 0 \Leftrightarrow \mathbb{E}(y_i | x_i) = x_i' \beta$ 。



预测v.s.解释

- 在预测中，先假设了 $\mathbb{E}(y_i|x_i) = x_i'\beta^0$ 从而得到了回归模型，进一步定义了 $e_i = y_i - \mathbb{E}(y_i|x_i) = y_i - x_i'\beta^0$ ，从而 $\mathbb{E}(e_i|x_i) = 0$ 。
 - x 是否必然为 y 的原因？
 - 不必然，比如通过购买行为判断性别

- 而在这里：先设定了模型

$$y_i = x_i'\beta + u_i$$

其中 x_i 为原因， y_i 为结果，而 u_i 为影响 y_i 的其他不可观测（unobservables）的原因。

- 此时，我们关注的是何时 β 具有因果效应的解释，或者 β 是否是我们希望得到的结构参数。

外生性

教育回报

在研究教育的回报问题时，我们通常使用所谓的Mincer方程进行估计：

$$\ln Y = \beta_0 + \beta_1 \cdot \text{edu} + \beta_2 \cdot \text{exp} + \beta_3 \cdot \text{exp}^2 + u$$

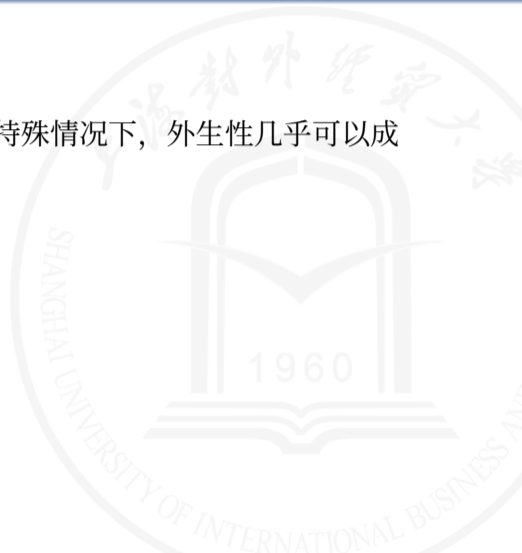
其中edu为教育年限，exp为经验。如果：

- 仅仅为了预测，我们不需要系数具有因果解释，可以使用OLS作为条件期望函数的逼近；
 - 此时获得的仅仅是相关性
 - 业界应用中较多
- 如果需要 β_1 具有因果解释，必须假设 $\mathbb{E}(u|\text{edu}, \text{exp}) = 0$
 - 一般不成立，例如 u 中包含着能力（ability），同时影响edu和 $\ln Y$
 - 关键：何时相关性具有因果性的解释。

外生性

一般情况下，很难argue外生性一定成立，但是特殊情况下，外生性几乎可以成立：

- 实验
- 自然实验
 - 改革开放（missing women）
 - 加入WTO
 -
- 根本要求：随机性！



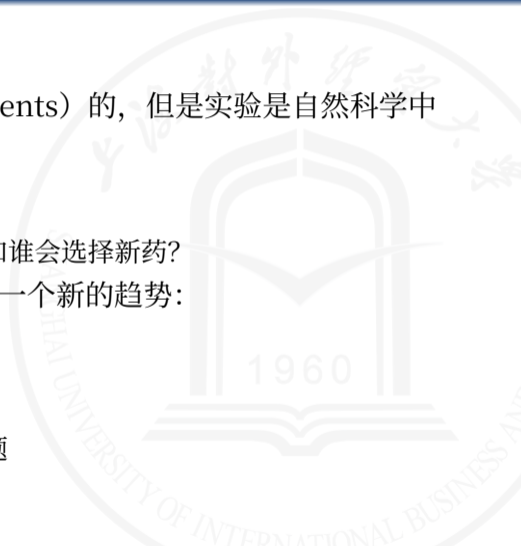
实验

传统上认为经济学是不可以进行实验（experiments）的，但是实验是自然科学中进行因果推断的最佳方法：

- 一个药物是否有效？
 - 最可靠的方法是进行实验
 - 如果不进行实验，就会有自选择问题，比如谁会选择新药？

随着经济学的发展，因果推断中引入实验成为了一个新的趋势：

- 越来越多的问题通过实验解决
- 但是实验并非万能的：
 - 很多问题不可以进行实验：伦理问题
 - 实验的结论可能不可外推：外部有效性问题



实验的准则

实验通过随机化（randomization），即随机的将实验对象分为实验组和对照组控制混淆因素：

- 理想的实验中，分组需要是完全随机的
- 这要求实验者能够完全控制分组（而非自己选择的）
- 随机分组保证了控制组和实验组平均而言其他特征都一样

实验不要求：

- 两个组别的样本量是一样的
 - 实际上样本量不一样完全没问题，特别是花费较高的实验
- 每个个体被选中的概率是一样的
 - 实际上只要被选中的概率已知（实验人员设定的）即可

自然实验

正是由于随机试验的这些限制，一些研究者转而使用所谓的自然实验（natural experiment）或者拟实验（quasi-experiment）。

- 与实验相同的是，自然实验同样强调核心解释变量的随机性
- 在自然实验中，核心解释变量的随机性并非由研究人员完全控制的随机分配得到的，而是从一些客观发生的、不随研究人员意志为转移的、自然发生的事件中得到的。
- 在此类方法中，研究人员通常强调研究设计（research design），控制混淆因素的主要方法来自于研究设计，而非结构建模或者其他统计方法。
- 自然实验的核心与随机实验一样，仍然是解释变量的随机性，或者外生性。

自然实验

真正随机的自然实验

Angrist等人(2002)使用了哥伦比亚1991年开始的PACES项目作为自然实验，研究了私立高中教育对于个人教育结果的影响。

- 现实中，是否上私立高中的决策受到家庭财富、家长的教育观念等多个因素的影响，不可观测的混淆因素导致外生性假设的违背。
- PACES项目为低收入家庭的符合条件的学生提供进入私立高中的凭证 (voucher)，这些凭证需要家庭进行申请，而一旦申请人数超过了供给，政府会通过抽签的方式分配。
- 作者使用了如下回归研究了凭证对于个人教育结果的影响：

$$y_{ic} = \alpha \cdot \text{voucher}_i + x_i' \beta + \delta_c + u_{ic}$$

Simpson悖论

- 辛普森悖论 (Simpson's paradox) 是指, 当我们对某个感兴趣的变量 y 进行比较时, 在每个组内比较的结果与忽略分组进行比较的结果可能是不相同的。
- 比如我们考虑如下的思想实验。我们已知男性的平均寿命比女性的平均寿命要短, 但是同时男性可能更愿意锻炼身体, 而锻炼身体对寿命有正向的促进作用。假设男性 ($\text{sex} = 1$) 有80%的人会从事锻炼, 而女性 ($\text{sex} = 0$) 只有30%;

Simpson悖论

- 假设不锻炼的女性平均寿命 (y) 为80岁，男性平均低10岁；而如果锻炼身体 ($exer = 1$)，平均可以延长3年的寿命。以上的数据生成过程可以总结如下：

$$y = 80 - 10 \cdot \text{sex} + 3 \cdot \text{exer} + e$$

其中 e 为影响寿命的其他与 sex 、 exer 无关的因素，假设 $\mathbb{E}(e|\text{sex}, \text{exer}) = 0$ ，那么有：

$$\mathbb{E}(y|\text{sex}, \text{exer}) = 80 - 10 \cdot \text{sex} + 3 \cdot \text{exer}$$

Simpson悖论

- 如果我们忽略性别这一因素，仅仅比较锻炼身体的人与不锻炼身体的人的寿命，即使用回归：

$$y = \beta_0 + \beta \cdot \text{exer} + u$$

由于 exer 为0/1变量，以上回归即直接对锻炼身体的一组与不锻炼身体的一组进行比较，从而：

$$\begin{aligned}\mathbb{E}(y|\text{exer}) &= \mathbb{E}(80 - 10 \cdot \text{sex} + 3 \cdot \text{exer} + e|\text{exer}) \\ &= 80 - 10 \cdot \mathbb{E}(\text{sex}|\text{exer}) + 3 \cdot \text{exer}\end{aligned}$$

其中 $\mathbb{E}(\text{sex}|\text{exer})$ 为给定（不）锻炼身体的群体中男性的比率。

Simpson悖论

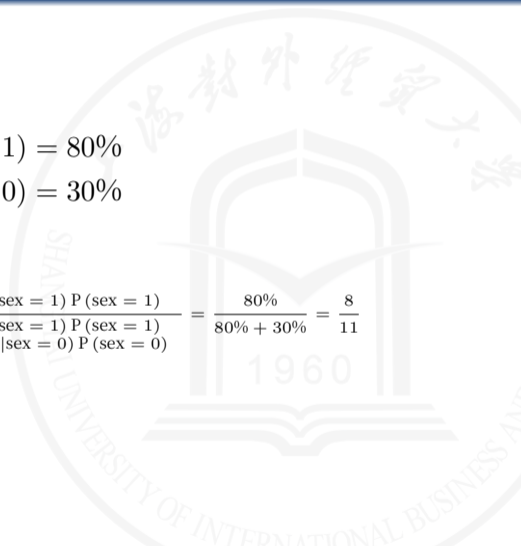
- 已知:

$$\begin{cases} P(\text{exer} = 1 | \text{sex} = 1) = 80\% \\ P(\text{exer} = 1 | \text{sex} = 0) = 30\% \end{cases}$$

根据贝叶斯公式:

$$\mathbb{E}(\text{sex} | \text{exer} = 1) = P(\text{sex} = 1 | \text{exer} = 1) = \frac{P(\text{exer} = 1 | \text{sex} = 1) P(\text{sex} = 1)}{P(\text{exer} = 1 | \text{sex} = 1) P(\text{sex} = 1) + P(\text{exer} = 1 | \text{sex} = 0) P(\text{sex} = 0)} = \frac{80\%}{80\% + 30\%} = \frac{8}{11}$$

- 同理 $\mathbb{E}(\text{sex} | \text{exer} = 0) = \frac{2}{9}$
- 从而 $\mathbb{E}(\text{sex} | \text{exer}) = \frac{2}{9} + \left(\frac{8}{11} - \frac{2}{9}\right) \cdot \text{exer}$



Simpson悖论

- 从而：

$$\begin{aligned}\mathbb{E}(y|\text{exer}) &= 80 - 10 \cdot \mathbb{E}(\text{sex}|\text{exer}) + 3 \cdot \text{exer} \\ &= 80 - 10 \cdot \left(\frac{2}{9} + \left(\frac{8}{11} - \frac{2}{9} \right) \cdot \text{exer} \right) + 3 \cdot \text{exer} \\ &= \frac{700}{9} - \frac{203}{99} \cdot \text{exer} \\ &\approx 77.78 - 2.05 \cdot \text{exer}\end{aligned}$$

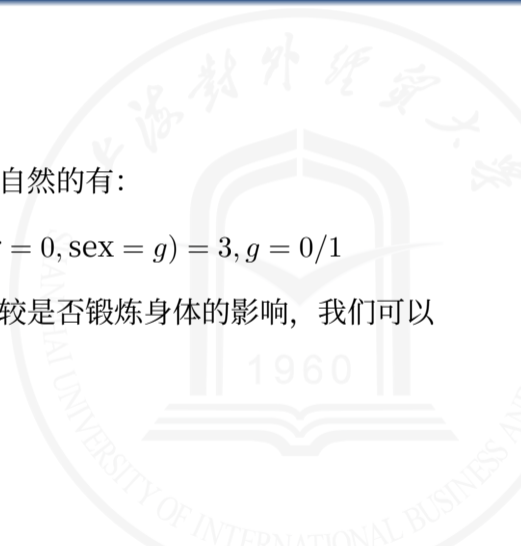
- 从上式可以看出，如果我们仅仅比较锻炼身体的组和不锻炼身体的组的寿命，那么我们会得到锻炼身体会减少2.05年寿命的结论，因而在这里，只包含锻炼与否的线性回归模型错误的估计了锻炼身体对于寿命的影响。

Simpson悖论

- 实际上，如果我们分性别进行比较，那么很自然的有：

$$\mathbb{E}(y|\text{exer} = 1, \text{sex} = g) - \mathbb{E}(y|\text{exer} = 0, \text{sex} = g) = 3, g = 0/1$$

因而如果分别在男性、女性两个组别内，比较是否锻炼身体的影响，我们可以得到正确答案。



Simpson悖论

