



THE WORLD BANK



# 技术路线

## 讲座 4

### 工具变量

Christel Vermeersch  
北京, 中国, 2009

# 举例：志愿者工作培训项目

---

- 假设现在要评价一个志愿者工作培训项目
  - 任何一个失业的人都是符合培训条件的
  - 有些人报名参加(“培训组”)
  - 其他人不参加(“对照组”)
- 一些简单但不是很理想的评价方法:
  - 比较培训组在参加培训项目前后的情况
  - 比较培训组和对照组实施培训项目后的情况
  - 比较培训组和对照组实施培训项目前后的情况

# 志愿者工作培训项目

- 假设我们决定比较参与和未参与培训项目人员的结果，可以通过一个简单的模型来表述：

$$y = \alpha + \beta_1 D + \beta_2 x + \varepsilon$$

其中 D=1 如果此人参与了培训

D=0 如果此人没有参与培训

X= 可控制变量(外在的和可观测的变量)

为什么此公式效果不好？有两个问题

- 有些变量很重要，但我们把它们忽略了(因为各种原因)
- 决定参与培训是内在原因起作用



# 问题#1：被忽略的变量

---

即使在缜密的模型中，也可能出现以下遗漏

- “忽略”的特点：我们不知道它们的重要性
- 过于复杂而难以测量的特征

例如：

- 禀赋和动机不同
- 不同层次的信息
- 参与的机会成本不同
- 服务可及性的程度不同

完全“正确”的模型是

$$y = \alpha + r_1 D + r_2 x + r_3 M + \eta$$

我们使用的模型是：

$$y = \alpha + \beta_1 D + \beta_2 x + \varepsilon$$



## 问题#2: 决定参与的内生因素

---

- 参与是一个决策变量? → 内生的!
- (例如: 它取决于参与者自身)

$$y = \alpha + \beta_1 D + \beta_2 x + \varepsilon$$

$$D = \pi + \pi_2 M + \xi$$

$$\Rightarrow y = \alpha + \beta_1 (\pi + \pi_2 M + \xi) + \beta_2 x + \varepsilon$$

$$\Rightarrow y = \alpha + \beta_1 \pi + \beta_2 x + \beta_1 \pi_2 M + \beta_1 \xi + \varepsilon$$

- 因此: 在这两个案例中: 我们丢掉了M变量项, 这个M变量是我们需要用来估计正确模型的, 但我们不能很好地测量它。

□ 完全正确的模型:  $y = \alpha + \gamma_1 D + \gamma_2 x + \gamma_3 M + \eta$

□ 简化模型:  $y = \alpha + \beta_1 D + \beta_2 x + \varepsilon$

---

- 比如说我们用  $\beta_{1,OLS}$  来估测干预效果  $\gamma_1$
- 如果 M 与 D 有相关关系, 而我们并没有在简化模型中考虑 M, 那么 D 的参数估计值将混杂部分 M 的效果. 这样就会影响到 M 与 D 相关联的程度.
- 因此: 在 OLS 模型中, 我们有关干预效果项  $\gamma_1$  的变量  $\beta_{1,OLS}$  的估计值包含干预效果以外的其他变量(M)的效果.
- 这就意味着在估计值  $E(\beta_{1,OLS})$  和  $\gamma_1$  之间有一个差异
  - OLS 中估计的  $\beta_1$  预期值并不是真正的干预效果  $\gamma_1$ ,
  - $\beta_{1,OLS}$  是一个带有偏倚的干预效果  $\gamma_1$  的估计值.

□ 完整的修正模型:  $y = \alpha + \gamma_1 T + \gamma_2 x + \gamma_3 D + \eta$

□ 被简化的模型:  $y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$

---

- 这就意味着在  $E(\beta_{1,OLS})$  值和  $\gamma_1$  值 之间有一个差异
  - OLS 测算值  $\beta_1$  的预期值  $\gamma_1$  并未反映干预效果值
  - $\beta_{1,OLS}$  是偏倚的干预效果  $\gamma_1$  的测算值.

□ 为什么会这样?

■ 违反了 OLS 中一个基本的条件 BLUE:

□ 换句话说  $E(\beta_{1,OLS}) \neq \gamma_1$  (有偏差的估计值)

□ 甚至更糟糕的是.....  $plim(\beta_{1,OLS}) \neq \gamma_1$  (不一致的估计值)

# 我们怎样解决这个问题？

---

$$y = \alpha + \gamma_1 D + \gamma_2 x + \gamma_3 M + \eta$$

$$y = \alpha + \beta_1 D + \beta_2 x + \varepsilon$$

- 尝试着消除  $D$  和  $\varepsilon$  之间的相关性：
- 通过被忽略的变量  $M$  来分离  $D$  中与误差项  $\varepsilon$  并不相关的变量
- 我们可以运用工具变量( $IV$ ) 来达到目的

# 工具变量 IV 背后的基本原理

$$y = \alpha + \gamma_1 D + \gamma_2 x + \gamma_3 M + \eta$$

$$y = \alpha + \beta_1 D + \beta_2 x + \varepsilon$$

- 基本问题是  $\text{corr}(D, M) \neq 0$
- 找出一个可以满足以下两个条件的变量  $Z$ :
  1. 与  $D$  相关联:  $\text{corr}(Z, D) \neq 0$ 
    - $Z$  和  $D$  是相关联的, 或者  $Z$  可以用来预测部分  $D$
  2.  $Z$  与  $\varepsilon$  是并不关联的:  $\text{corr}(Z, \varepsilon) = 0$ 
    - $Z$  本身对  $y$  并无影响。它是通过影响  $D$  使  $y$  发生变化。  
 $Z$  对  $y$  的所有影响都是通过  $D$  来传导的。
- 在志愿者工作培训项目中有  $Z$  的例子?

# 二阶最小平方(2SLS)

- 还记得包含内生变量D的原始模型：

$$y = \alpha + \beta_1 D + \beta_2 x + \varepsilon$$

- 步骤一：做一次内生变量(D)与工具变量Z和其他外生变量的回归方程：

$$D = \delta_0 + \delta_1 x + \theta_1 Z + \tau$$

- 为每一个研究对象计算D的预测值( $\hat{D}$ )
- 因为Z和x并不与 $\varepsilon$ 相关联，也不与D的预测值  $\hat{D}$  相关
- 你将会为每一个潜在的内生回归值提供一个工具变量



# 二阶最小平方(2SLS)

- 步骤二：做一次 $y$ 与 $\hat{D}$ 变量和其它外生变量的回归：

$$y = \alpha + \beta_1 \hat{D} + \beta_2 x + \varepsilon$$

- ▶ 注意：在OLS第二步中的标准误差需要被纠正，因为 $\hat{D}$ 不是一个固定回归量
- ▶ 在实际情况中，运用STATA软件给予二阶最小平方命令，程序会立即报告出正确的标准误差。
- ▶ 直觉：通过运用 $Z$ 对 $D$ 的影响，我们消除了 $D$ 自身与 $\varepsilon$ 的相关性
- ▶ 这可以被显示为(在一定条件下)通过工具变量可以形成一个连续稳定的 $\gamma_1$ 估计值。(大样本理论)



# 工具变量的运用

---

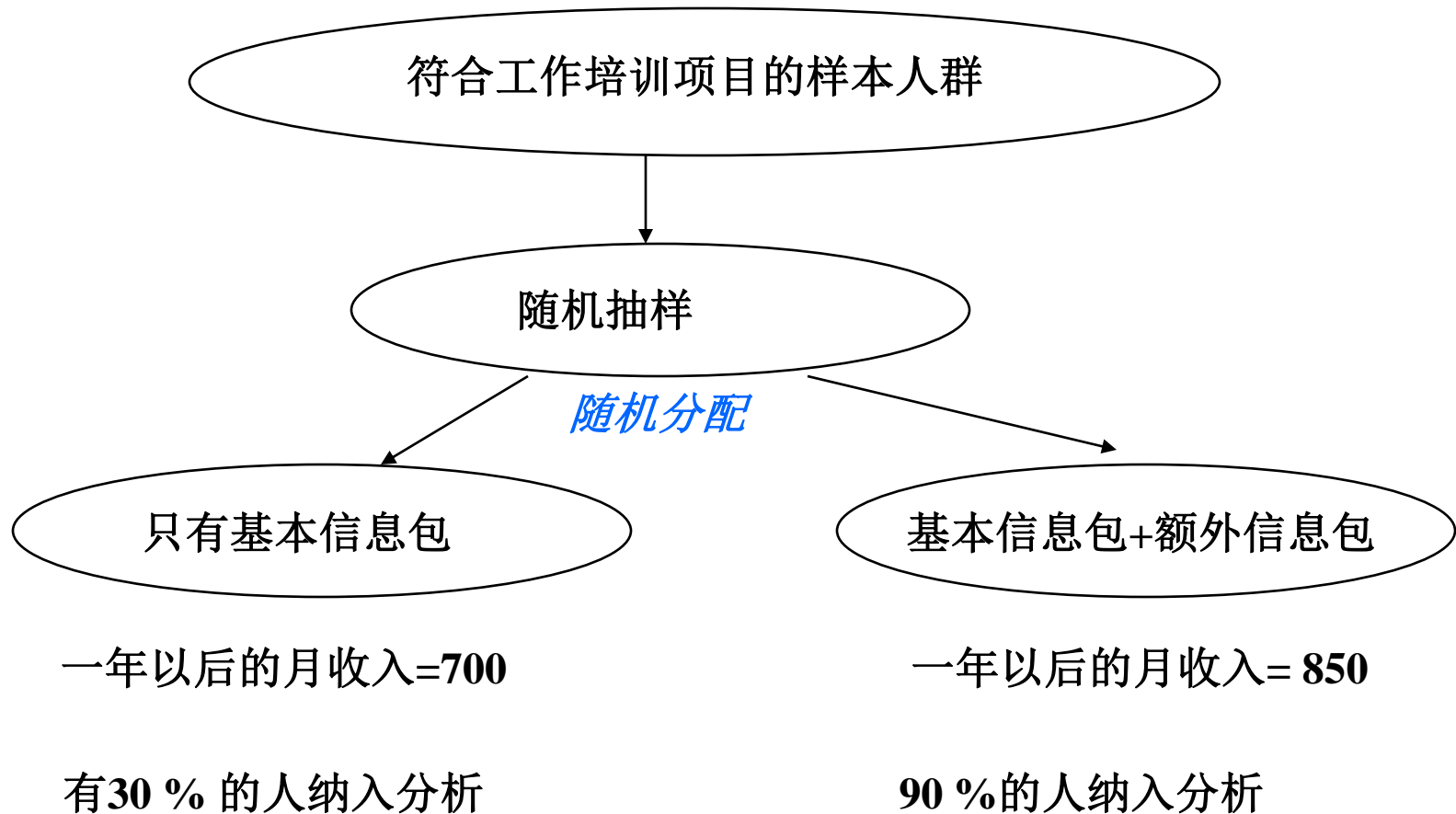
- 同时发生: X 和 Y 相互影响
  - ▣ 工具变量 X
- 被忽略的变量: X 包含了那些被忽略的变量的效果
  - ▣ 工具变量 X 包含一个与被忽略变量没有相关关联的变量
- 测量误差: X 并没有被精确地测量
  - ▣ 工具变量 X

# 我们从哪里寻找工具变量?

---

- 找出一个工具变量 ---- 很难!
- 利用数据建立一个
  - 当所有人都有资格参加干预项目时
  - 但有些人能获得比其他人更多的信息
    - 信息量大的人更有可能参与项目
  - 随机提供“额外信息”

# 事例 1: 志愿者工作培训项目



问题:在职培训项目的效果?

基本信息包

基本+额外信息包

一年后的月收入 = 700

一年后的月收入 = 850

30%的人纳入分析

90%人纳入分析

问题: 职业培训项目的影响是什么?

• “完全享有信息”组与“不完全享有信息组”之间的差异

.....

• 不同纳入率的校正

.....

• 实际: 效果 = .....

# 结合估算公式

---

## □ 步骤一:

- 以是否接受额外信息作为虚拟变量，做一个培训项目参与情况的回归方程(线性方程)
- 计算参与情况的预测值

## □ 步骤二:

- 以工资为变量，做参与情况预测值的回归

# 事例二： 尼泊尔学校的自治情况

---

- 目标是评估
  - A. 将学校的管理权交给社区
  - B. 学生成绩报告单
- 数据
  - 评估对象为1000所学校
  - 每个社区自主决定是否参与
  - 学生成绩报告单由非政府组织评定
  - 每个社区只有一所学校
- 任务: 设计能够评价的项目实施方案---提出评估方法的建议

# 尼泊尔学校的自治情况

		干预 B: 学生成绩报告单 由非政府组织进行干预		
		Yes	No	<i>Total</i>
干预A的工具变量: 非政府组织访问社区 并通知社区内的学校 由社区来管理的程序	Yes	300	300	<i>600</i>
	No	200	200	<i>400</i>
	<i>Total</i>	<i>500</i>	<i>500</i>	<i>1000</i>

# 提示....

---

## □ $corr(Z, \varepsilon) = 0$

- 如果  $corr(Z, \varepsilon) \neq 0$  “不合格的工具” ; 问题!
- ; 寻找一个好的工具变量是很难的!
- ; 既运用理论又利用常识去找寻一个工具变量!
- 我们可以通过设计来获得一个好的工具变量.

## □ $corr(Z, D) \neq 0$

- “无说服力的工具” :  $Z$  和  $D$  的相关性必须非常强.
- 如若不然, 大样本中同样存在偏倚