

# 【干货分享】实证论文中必须认真解决的内生性问题，到底有哪几种处理方案？最全的总结（附示例）



## 一、控制代理变量

**代理变量：**用来代替观测数据中难以得到或无法测量的信息。

举个例子：例如研究“是否上大学对收入的影响”，要控制住上大学这个条件之外所有对收入可能造成影响的条件，比如“能力”，家庭条件等有很多难以测量的信息被遗漏

好的代理变量满足以下**三个假设**，由于比较难理解，穿插例子的形式进行理解：

比如说我们想看是否上大学对收入的影响，

即原本的关系满足：




我们找到了衡量“能力”的代理变量 IQ，满足



原本的关系变成：



衡量 IQ 是一个好的代理变量需要满足以下三个假设：

- 假设 1，代理变量与所缺失的混杂因素相关。即 IQ 与 Ability 相关， $\gamma_1$  不为 0；
- 假设 2，如果将该代理变量纳入方程内生性问题，则不存在。即 Ability 放入 Income 式子中，方程原本的内生性不存在，即新产生的  
 不与 college 或 iq 相关；
- 假设 3，无法被代理变量所解释的那部分缺失变量与其他自变量无相关。即 e 不与 college 或 iq 相关。

#### 4. 代理变量的不足

- 代理变量可以大概率减少该变量所在的内生性问题，但是无法完全替代我们研究中所忽略的那个变量。
- 



## 二、固定效应模型

基于分析**面板数据**（对同一样本进行重复观测，比如说家庭追踪调查）

### 公式解释固定效应

通过对同一样本进行重复观测，得到简单的线性回归：

$y_{it}$ ，其中残差项  $u_{it}$  可以分成两部分  $\alpha_i$  和  $u_{it}$ ，

$y_{it}$ ，其中  $\alpha_i$  指那些影响  $y$  却不随时间变化的不可观测变量，有时被称作固定效应，指代不可观测的异质性； $u_{it}$  指随时间变化的，但不影响外的不可观测变量。满足独立同分布。

### 3. 一阶差分模型（FD）

当面板数据的时间为两个阶段的时候就是一阶差分模型，当时间大于两个阶段的时候就是固定效应模型，一阶差分模型以第一时间段为基期水平，进行相减，从而抵消固定效应，而固定效应模型以所有期的平均水平为基准点，每一期减去平均水平消去固定效应。

- 两年数据模型分别表示为

$$T=2: y_{i2} = \beta_0 + \delta + \beta_1 x_{i2} + \alpha_i + u_{i2}$$

$$T=1: y_{i1} = \beta_0 + \beta_1 x_{i1} + \alpha_i + u_{i1}$$

- $(T=2)-(T=1)$

$$\Rightarrow (y_{i2} - y_{i1}) = \delta + \beta_1(x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

$$\text{Or } \Delta y_{i1} = \beta_1 \Delta x_i + \Delta u_i$$

一阶差分估计

(First difference, 简称FD)

需注意：

- $\alpha_i$  消掉了
- 要获得一致无偏估计，需假设  $\Delta x_i$  与  $\Delta u_i$  不能相关
- $\Delta x_i$  要有变异。如果  $x_i$  不随时间变化，那么一阶差分后  $\Delta x_i$  就被抵消了

知乎 @二三两两

3. 注意，如果我们使用固定效应模型去分析  $x$  对  $y$  的影响时，那么  $x$  需要是随时间变化的变量，因为非时变变量会被固定效应消去，无法估计。

## 4. 拓展 -- 随机效应模型 (RE)

如果我们非要去研究非时变变量对于  $y$  的影响，那么可以使用随机效应模型。随机效应模型既可以去估计非时变量的影响，也可以估计时变变量的影响。但是需要满足一定条件，

$\mu_i$ ， $t=1,2, \dots, T$  中， $\mu_i$  与  $\epsilon_{it}$  不相关，即  $\text{Cov}(\mu_i, \epsilon_{it}) = 0$ ， $t=1,2, \dots, T$ ，其中  $\mu_i$  在固定效应模型中是指那些影响  $y$  却不随时间变化的不可观测变量，在随机效应模型中， $\epsilon_{it}$  满足独立同分布，所以才能研究非时变变量对于  $y$  的影响。

### 4.1 Stata 命令实现 RE 模型

```
xtset id year // 设定panel variable和time variable
xtreg y x, fe // 固定效应模型
est sto fe // 把估计结果暂时存储进内存，命名为fe

xtreg y x, re // 随机效应模型
est sto re // 把估计结果暂时存储进内存，命名为re

Hausman fe re // Hausman检验, 显著选FE，不显著选RE
```



## 三、工具变量 (IV)

通过构建工具变量，来检测不可观测的因素的影响。

### 工具变量的来源

#### 自然现象

- 霍克斯比：Y = 地区教育质量；X = 地区学校数量；Z = 地区河流数量（通过河流划分学区）
- 阿西莫格鲁：Y = 国家人均收入；X = 制度；Z = 殖民地时代死亡率
- 安谷瑞斯特：Y = 母亲就业；X = 孩子数；Z = 老大老二的性别组合

#### 时空距离（自然历史实验）

- 安古瑞斯特：Y = 收入；X = 教育年；Z = 出生的季度
- 卡德：Y = 收入；X = 教育年；Z = 家距离大学远近

- 钱楠筠：Y = 男女性别比；X = 家庭收入男女性别比；Z = 茶叶加工
- 陈云松：Y = 幸福感；X = 是否信教；Z = 解放前宗教场所
- 陈云松：Y = 政治信任；X = 城市餐饮, 解放前宗教场所；Z = 餐饮, 参与社群

### 3. 公式理解：



，如果 ，可以考虑使用工具变量方法。

### 4. 引入工具变量 $z$ ， $z$ 需要满足两个关键假设：

- 相关性： $z$  与  $x$  相关
- 外生性： $z$  与  $u$  不相关， $z$  只能通过  $x$  影响  $y$



### 5. 工具变量求解两步走

- 1) 先用工具变量  $z$  作自变量， $x$  做因变量对  $z$  进行回归，目的在于把  $x$  分为两个部分，与  $z$  无关（内生部分，即与误差相关。），与  $z$  有关（外生部分）
- 2) 与  $z$  相关部分作为 ，用  $y$  对  回归

### 工具变量一个经典例子

- 例：Angrist (1990)

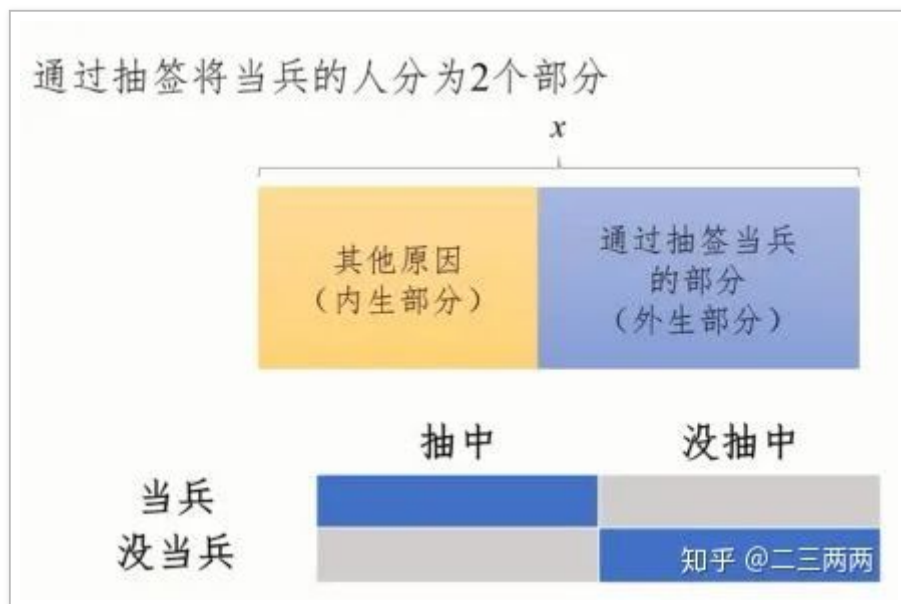
#### 当兵经历对之后工作收入的影响？

- 当兵经历具有内生性
- 工具变量：抽签
  - 背景：越南战争期间，美国用抽签的方式决定谁去当兵，抽签的标准是出生日期。
- 工具变量有效性：
  - 抽签影响是否当兵，意味着  $Cov(x, z) \neq 0$
  - 抽签是随机的，意味着  $Cov(u, z) = 0$

知乎 @二三两两

## 6. 对变量的潜在问题

- 1) 工具变量估计的是局部平均处理效应（LATE），就以上述例子为例，抽签的影响是部分的，只能研究外生部分。因 LATE，因果效应难以推广；



- 2) 工具变量很难找且容易被攻击

## 7. Stata 命令实现工具变量

最常用：

- 线性模型：ivregress
- 非线性模型：ivprobit, ivtobit

最紧凑：

- 既可线性也可非线性：cmp

相关检验

- 过分识别检验（Overidentification test）：overid
  - 内生性检验（Endogeneity of X in LS model）：ivendog
  - 异方差检验：ivhettest
- 知乎 @二三两两



## 四、赫克曼方法

### 1. 赫克曼模型

如果样本不是随机被选择的，那么如果某些样本的缺失可能会造成偏差，这种**样本选择问题**的解决方案可以采用赫克曼模型分析。比如想要拟合一个收入模型，那么只有工作的人才有收入，没有工作的人将不会被纳入分析，此时样本有偏可能导致结果有偏。

## 2. 解决方法：

2.1 增加一个选择方程，对应的结果模型也是具有选择的，以“收入”为例，增加的模型如下：

• 增加了一个选择方程

$$work^* = \gamma Z + v$$

选择模型

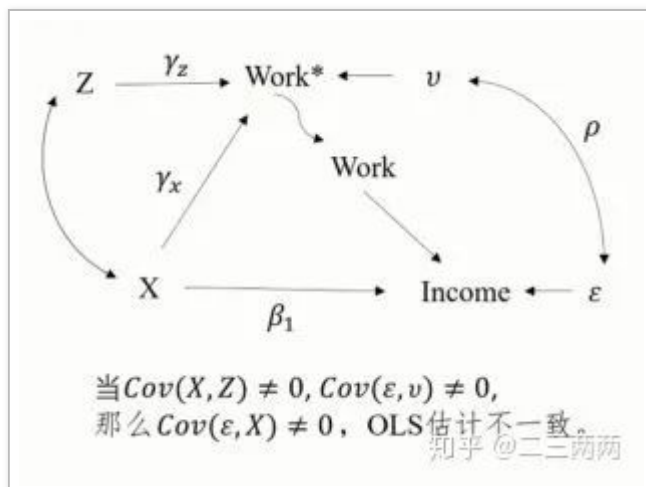
$$work = \begin{cases} 1 & \text{if } \gamma Z + v \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

结果模型

$$\begin{cases} Income = \beta_0 + \beta_1 X + \varepsilon & \text{if } work = 1 \\ Income \text{ 观测不到} & \text{if } work = 0 \end{cases}$$

### 2.2 增加选择方程时考虑两点：

- 1) 样本的选择性来自哪里？（明确研究的目标群体是什么？实际分析的目标群体又是什么？）
- 2) 如何基于选择性样本，获得无偏估计？（即上图中的 Z 如何选定）
  - Heckman 模型假定
    - 【外生性】，控制了 X 之后，误差项  $\varepsilon$  和  $v$  都满足 iid
    - 【单调性】，加入的选择模型要么增长，要么下降，是一个连续模型





逆米尔斯率 (Inverse Mills Ratio, IMR)  $\longrightarrow \lambda(Z\gamma) = \frac{\phi(Z\gamma)}{\Phi(Z\gamma)}$

← 标准正态概率密度函数  
← 标准正态概率累积函数

## 2.3 赫克曼模型校正

- 如何进行校正？

结果模型

$$Income = \beta_0 + \beta_1 X + \varepsilon$$

选择模型

$$work = \gamma Z + v$$

- 两步估计法 (two-step estimation)

第一步：对所有的n个个案，估计一个probit模型 (work decision)，从而得到 $\gamma$ ，并对每个个案计算IMR， $\hat{\lambda} = \lambda(Z, \hat{\gamma})$ ；

第二步：基于选择性样本（即，work=1的样本），用Income对X和 $\hat{\lambda}$ 进行回归。

通过两步法得到校正了选择性误差之后的系数。

- 模型不足：

- Z的选择需要符合 **exclusion restriction** (即，该变量影响selection, not the outcome)

知乎 @二三两两

不足：z 的选择非常难，跟工具变量一样

## 3. Stata 命令实现赫克曼模型

- Stata命令：heckman (Y是连续变量), heckprobit (Y是二分变量)

比如，

heckman lnincome educ age, select(work=married children educ age) twostep

知乎 @二三两两



## 五、倍差法 (DID)

基于实验的设计，结果是否有效，取决于实验设计，有些人用来研究政策的影响。

满足“**共同趋势假设**”的话，倍差法实现会变得简单，“共同趋势假设”是指不进行干预，处理组的变化情况与控制组相同



- 两个组：
  - 处理组(Control)
  - 控制组(Treatment)
- 时间：
  - 施加处理前(Pre)
  - 施加处理后(Post)

	Pre	Post
Control	$\bar{y}_{C,Pre}$	$\bar{y}_{C,Post}$
Treatment	$\bar{y}_{T,Pre}$	$\bar{y}_{T,Post}$

$$DD = (\bar{y}_{T,Post} - \bar{y}_{T,Pre}) - (\bar{y}_{C,Post} - \bar{y}_{C,Pre})$$

知乎 @二三两两

- A=1是处理组，0是控制组**
- 基于潜在结果框架：
 
$$E(Y|A=1) - E(Y|A=0)$$

$$= \underbrace{E(Y^1|A=1) - E(Y^0|A=1)}_{ATT} + \underbrace{E(Y^0|A=1) - E(Y^0|A=0)}_{\text{样本选择性偏差}}$$
  - 如果我们有接受处理前，处理组和控制的信息，即 $Y^{pre}$ 。
  - 假设 **处理组假如没有接受处理的情况**

$$E(Y^0|A=1) - E(Y^{pre}|A=1) = E(Y^0|A=0) - E(Y^{pre}|A=0)$$

共同趋势假设 (parallel trends assumption)
  - 该假设是倍差法的重要假设。 **处理组接受处理以前的情况**

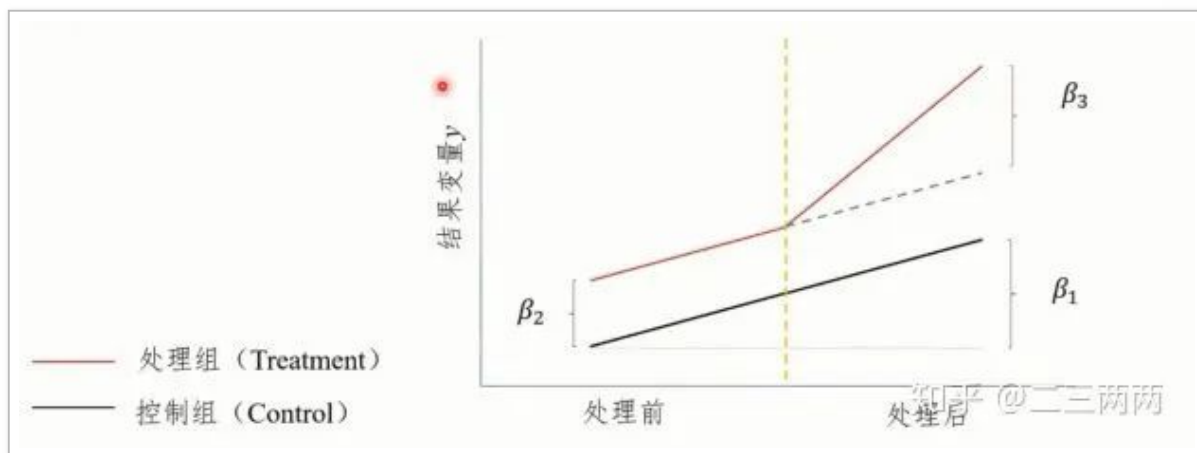
知乎 @二三两两

## 2. 在回归中表达 DID

time 和 treatment 都是 0, 1 变量

$$y = \beta_0 + \beta_1 \cdot Time + \beta_2 \cdot Treatment + \beta_3 \cdot Treatment \cdot Time + \varepsilon$$





## 六、断点回归 (RDD)

断点回归可以用来考察政策实施的影响，有以下三个**特点**：

- 估计二分变量  $D$  对  $Y$  的因果效应；
- 处理变量分配不随机；
- 常常是否接受处理取决于一个确定的规则；
  - 变量  $X$  取值决定是否接触处理， $X$  被称为分配变量 (Running variable or assignment variable)

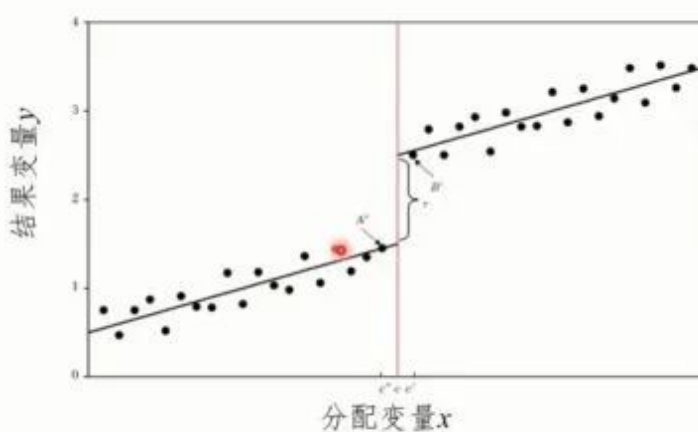
2. 例子：“颁发国家奖学金是否有助于提高大学生的学业成绩？”，定 650 分为基准线，大等于 650 分意味着能提高，低于 650 则不能，650 是临界点，在该点存在一个跳跃

### RD设定

- $Y$ 轴是大学成绩
- $x$ 轴是高考分数
- 在临界点650分会会有一个跳跃，跳跃的大小就是获得奖学金的效应。

### 注意：

- 简单起见，我们设定分配变量 $x$ 与结果变量 $y$ 的关系为线性，但是实际分析中，两者的关系可以更加复杂。



### 3. 精确断点回归

## 精确断点回归 (sharp regression discontinuity)

- 精确断点，是指在临界点 $X = X_0$ 处， $D$ 的取值从0跳至1。
  - 当 $X \geq X_0$ 时， $D = 1$
  - 当 $X < X_0$ 时， $D = 0$
- 此时， $D$ 的取值完全由 $X$ 决定。控制了 $X$ 之后 $D$ 是常数，因此必然满足 $(Y_1, Y_0) \perp D | X$

知乎 @二三两两

### 3.1 断点回归的多种形式：

第一种情况：线性关系

$$Y = \beta_0 + \beta_1 X + \beta_2 1(X > X_0) + \varepsilon$$

第二种情况：多项式

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 1(X > X_0) + \varepsilon$$

第三种情况：临界点两边变化趋势不同

$$X' = X - X_0$$

$$Y = \beta_0 + \beta_1 X' + \beta_2 1(X > X_0) X' + \beta_3 1(X > X_0) + \varepsilon$$

只要临界点两边 $f(X)$ 的形态设定正确，可以得到treatment的有效估计

知乎 @二三两两

## 4. 模糊断点回归

### 4.1 精确断点回归设计存在问题

一旦规则确定分配结果也确定在现实情况下很难满足，实际分配处理不一定严格按照规则执行。这意味着临界点不再是一个清晰断点，而是模糊的，更多反映的是接受处理的概率，这样的断点称为模糊断点回归。比如说，成绩高于 650 分，获得奖学金的概率更高，而不是一定获得奖学金。

### 4.2 构造变量 Z

- 模糊断点回归设计并不是断点的位置模糊，只是在断点处，D的值不是全部从0跳至1

如何估计？

- 可通过构造变量Z
  - 当 $X \geq X_0$ 时， $Z=1$
  - 当 $X < X_0$ 时， $Z=0$
- Z很大程度上决定随机分组情况，在控制X以后，Z并不直接影响Y
- Z就相当于工具变量

知乎 @二三两两

如何实现？

- 情况1

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k + \delta D + \varepsilon$$

- 当只有D是内生变量，可以Z作为其工具变量通过两阶段最小二乘法求解。

- 情况2

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k + \delta D + \delta_1 DX + \delta_2 DX^2 + \dots + \delta_k DX^k + \varepsilon$$

- 当D、DX、DX<sup>2</sup>...DX<sup>k</sup>都是内生变量，可以Z、ZX、ZX<sup>2</sup>、...ZX<sup>k</sup>作为工具变量通过两阶段最小二乘法求解。

知乎 @二三两两

## 5. RDD 关键：寻找跳跃

### 5.1 跳跃需要符合两个条件：

- 1) 让自变量和因变量同时跳起来的连续变量 x（自变量和因变量跳的幅度越大，断点回归设计越有效）；
- 2) 同时其他影响因变量的协变量在断点处不能有跳跃。

### 5.2 断点回归的不足

- 满足要求的 X 不好找；
- 基于临界点前后的样本进行估计，若样本量小，随机波动会很大；
- 断点回归法估计的是局部平均处理效应（LATE）。

## 七. 解决内生性问题的方法组合

### 例子 1：固定效应 + 赫克曼

$$Y_{it} = \beta_{0t} + \beta_1 S_{it} + \beta_2 X_{it} + \beta_3 \hat{\lambda}_{it} + \alpha_{it} + \varepsilon_{it} \quad (4)$$

$$\lambda_{it} = \gamma_{0t} + \gamma_1 Z_{it} + \gamma_2 X_{it} + \mu_{it} \quad (5)$$

$$Y_{i(t-1)} = \beta_{0(t-1)} + \beta_1 S_{i(t-1)} + \beta_2 X_{i(t-1)} + \beta_3 \hat{\lambda}_{i(t-1)} + \alpha_{i(t-1)} + \varepsilon_{i(t-1)} \quad (6)$$

$$\lambda_{i(t-1)} = \gamma_{0(t-1)} + \gamma_1 Z_{i(t-1)} + \gamma_2 X_{i(t-1)} + \mu_{i(t-1)} \quad (7)$$

Then we subtract Eq. (6) from Eq. (4) and obtain:

$$\begin{aligned} Y_{it} - Y_{i(t-1)} &= (\beta_{0t} - \beta_{0(t-1)}) + \beta_1 (S_{it} - S_{i(t-1)}) + \beta_2 (X_{it} - X_{i(t-1)}) \\ &\quad + \beta_3 (\hat{\lambda}_{it} - \hat{\lambda}_{i(t-1)}) + (\alpha_{it} - \alpha_{i(t-1)}) + (\varepsilon_{it} - \varepsilon_{i(t-1)}) \end{aligned} \quad (8)$$

Since we assume that individual unobserved variables are time-invariant, that is  $\alpha_{it} - \alpha_{i(t-1)} = 0$ , Eq. (8) ends up as:

$$\Delta Y_i = \beta_0 + \beta_1 \Delta S_i + \beta_2 \Delta X_i + \beta_3 \Delta \hat{\lambda}_i + \Delta \varepsilon_i \quad (9)$$

where “ $\Delta$ ” denotes the change from  $t$  to  $(t-1)$ . Eq. (9) is thus a first-difference Heckit model (Heckit-FD), predicting changes of outcomes as a function of changes of the independent variables.

收入      声望      个人因素

$$Y_{it} = \beta_{0t} + \beta_1 S_{it} + \beta_2 X_{it} + \beta_3 \hat{\lambda}_{it} + \alpha_{it} + \varepsilon_{it}$$

知乎 @二三两两



## 例子 2：工具变量 + 赫克曼

OLS 模型方程可以写作：

$$W_{ig} = \beta_0 + \beta_1 S_g + \beta_2 X_{ig} + \beta_3 V_g + \epsilon \quad (1)$$

其中  $W_{ig}$  代表农民工在城市的工资收入,  $i$  表示第  $i$  个农民工,  $g$  表示第  $g$  个村庄,  $S_g$  表示村庄的外出打工人数,  $X_{ig}$  表示个人层面的控制变量,  $V_g$  是村庄层面的控制变量,  $\epsilon$  则是非观测因素的联合效应, 也即误差项。因此,  $S_g$  是主解释变量, 而  $\beta_1$  代表了社会网效应。注意, 获得  $\beta_1$  的无偏估计量的前提是  $\text{Cov}[S_g, \epsilon] = 0$ 。但是这个假设可能有很大的问题。

在 OLS 模型的基础上, 用赫克曼方法来解决样本选择问题, 因此 Heckit 模型可以写成由(2)和(3)组成的方程组：

$$W_{ig} = \beta_0 + \beta_1 S_g + \beta_2 X_{ig} + \beta_3 V_g + \beta_4 \hat{P}_{ig} + \epsilon \quad (2)$$

$$\text{是否打工 } P_{ig} = \gamma_0 + \gamma_1 F_{ig} + \gamma_2 S_g + \gamma_3 X_{ig} + \gamma_4 V_g + \mu \quad (3)$$

赫克曼一定要找到排除限定

其中  $\hat{P}_{ig}$  即反向 Mills 比率,  $F_{ig}$  是作为排除限定的家庭劳动力人数, 不会在方程(2)出现。方程(2)的所有解释变量, 是方程(3)解释变量的严格子集。 $\mu$  是方程(3)中的误差项。

最后, 基于以上, 结合了赫克曼方法和工具变量方法的 IV-Heckit 模型由方程(4)、(5)、(6)组成： $s$  的变化一下给  $w$  就有因果关系

$$W_{ig} = \beta_0 + \beta_1 \hat{S}_g + \beta_2 X_{ig} + \beta_3 V_g + \beta_4 \hat{P}_{ig} + \epsilon \quad (4)$$

$$\text{米尔斯率 } P_{ig} = \gamma_0 + \gamma_1 F_{ig} + \gamma_2 N_g + \gamma_3 X_{ig} + \gamma_4 V_g + \mu \quad (5)$$

$$\text{社会资本 } S_g = \alpha_0 + \alpha_1 N_g + \alpha_2 X_{ig} + \alpha_3 V_g + \alpha_4 \hat{P}_{ig} + \eta \quad (6)$$

先估算5再6后4

其中  $N_g$  代表村庄所遭受的自然灾害的强度,  $\eta$  是误差项。注意, 此时方程(5)不应包括内生解释变量  $S_g$ , 而纳入了外生工具变量  $N_g$

知乎 @二三两两

来源及版权：本文来源微信公众号刘西川阅读写作课，原文作者知乎用户“二三两两”版权归作者所有！

每天更新内容，包括：空间经济、城市经济、土地经济、交通经济、城市规划、经济地理、区域经济等学科领域

敬请关注，微信号：quyujingji

微信公众号二维码：



## “区域经济”公众号

### 联系方式

微信：qq312462147

E-mail: 312462147@qq.com

全文完

本文由 简悦 SimpRead 转码，用以提升阅读体验，原文地址