



# 计量经济学II (Econometrics II)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

[huhuaping01@hotmail.com](mailto:huhuaping01@hotmail.com)

2022-09-26

西北农林科技大学

# 模块04：联立方程模型 (SEM)

Chapter 17. 内生性问题与工具变量法

Chapter 18. 为什么要关心联立方程模型？

Chapter 19. 联立方程模型的识别问题

Chapter 20. 联立方程模型的估计方法

# 17. 内生性问题与工具变量法

17.1 内生自变量问题的定义和来源

17.2 内生变量法下的估计问题

17.3 工具变量及其选择

17.4 两阶段最小二乘法 (2SLS)

17.5 检验工具变量的有效性(Instrument validity)

17.6 检验自变量的内生性(regressor endogeneity)

## 17.1 内生自变量问题的定义和来源



# 知识回顾

对于总体回归模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (\text{PRM})$$

- 在CLRM假设下：**CLRM假设A2**—— $X$ 是固定的（给定的）或独立于误差项。也即自变量 $X$ 不是随机变量。此时，我们可以使用OLS方法，并得到**BLUE**。

$$\text{Cov}(X_i, u_i) = 0$$

$$E(X_i u_i) = 0$$

- 如果违背上述假设，也即自变量 $X$ 与随机干扰项相关。此时使用OLS估计将不再能得到**BLUE**，而应该采用**工具变量法（IV）**进行估计。

$$\text{Cov}(X_i, u_i) = 0$$

$$E(X_i u_i) = 0$$



# 好模型的标准与外生自变量

$$y = X\beta + u$$

**随机控制实验** (randomized controlled experiment)：理想情形下，自变量X的取值是随机分配变化的（**原因**），然后我们再来观测因变量Y的变化（**结果**）。

- 如果  $Y_i$  和  $X_i$  确实存在系统性的关系（线性关系），那么改变  $X_i$  则导致  $Y_i$  的相应变化。
- 除此之外的任何其他随机因素，都将放到随机干扰项  $u_i$  中，它对因变量  $Y_i$  的变动影响，应该是**独立于**  $X_i$  的影响作用的。



# 好模型的标准与外生自变量

外生自变量 (exogenous regressors) : 如果自变量  $X_i$  真的是如上所说的完美的随机取值 (randomly assigned) , 则称之为外生自变量。更准确地, 可以定义为:

严格外生性假设 (strictly exogeneity) :  $E(u_i | x_1, \dots, x_N) = E(u_i | \mathbf{x}) = 0$ 。

因为在随机控制实验, 给定样本  $i$  和样本  $j$ , 自变量的取值分别为  $X_i$  和  $X_j$ , 它们应该是相互独立的。因此可以把上述假设进一步简化为:

同期外生性假设 (contemporaneously exogeneity) :  $E(u_i | X_i) = 0, \text{ for } i = 1, \dots, N$ 。



# 大样本情况下OLS方法

在大样本情形下，上述严格外生性假设可以进一步转换为同期不相关假设：

- $E(u_i) = 0$ ，而且
- $\text{cov}(x_i, u_i) = 0$

因为我们可以证明（证明略），在大样本情况OLS方法下：

- $E(u_i | X_i) = 0 \Rightarrow E(u_i) = 0$
- $E(u_i | X_i) = 0 \Rightarrow \text{cov}(x_i, u_i) = 0$



# 内生自变量问题的定义

$$y = X\beta + u$$

在经典线性回归模型假设(CLRM)中，我们假设所有回归元  $X_i$  是给定的，且随机干扰项的条件期望为0 ( $E(u_i|X_i) = 0$ )。

- 回归元是**严格外生性**具有重要意义，因为理论上将表明模型的预测误差将是最小的（等于0）
- 实际上，我们的这一假设要求非常高，在**随机控制实验**中要求  $X_{ki}$  是**同期外生性**的  $E(u_i|X_i) = 0, \text{ for } i = 1, \dots, N$ 。

然而，现实中，回归元  $X_i$  可能是**随机的**；而且回归元与随机干扰项可以能是**相关的**。此时，我们称模型存在**内生自变量**（endogenous regressors）问题。正式地：

- 如果自变量与随机干扰项无关，则称之为**外生变量**（**exogenous**）
- 如果自变量与随机干扰项相关，则称之为**内生变量**（**enogenous**）。



# 内生自变量问题的几种情形

在应用计量经济学中，内生性通常以以下四种方式之一出现：

- 遗漏变量 (Omitted variables)
- 测量误差 (Measurement errors)
- 自相关问题 (Autocorrelation)
- 方程联立性问题 (Simultaneity)

西北农林科技大学  
NORTHWEST A&F UNIVERSITY

西北农林科技大学  
NORTHWEST A&F UNIVERSITY



# 内生自变量情形I：遗漏变量

假定假定工资水平的“真实模型”为：

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i \quad (\text{the assumed true model})$$

然而，因为个体的能力变量  $Abl$  往往无法直接观测得到，因此我们往往不能放入到模型中，并构建了一个有偏误的模型。

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i \quad (\text{the error specified model})$$

其中能力变量  $Abl$  被包含到新的随机干扰项  $v_i$  中，也即： $v_i = \beta_3 abl_i + u_i$

显然，我们认为偏误模型中，忽略了能力变量  $Abl$ ，而受教育年数变量  $Edu$  实际上又与之有相关关系。进而偏误模型中， $cov(Edu_i, v_i) \neq 0$ ，从而受教育年数变量  $Edu$  具有内生自变量问题。



# 内生自变量情形1：遗漏变量（演示1）

下面我们对遗漏变量情形做一个整体的直观演示：

假定工资水平的“真实模型”为：

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$





# 内生自变量情形1：遗漏变量（演示1）

下面我们对遗漏变量情形做一个整体的直观演示：

假定工资水平的“真实模型”为：

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

A同学构建遗漏变量的模型：

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$





# 内生自变量情形1：遗漏变量（演示1）

下面我们对遗漏变量情形做一个整体的直观演示：

假定工资水平的“真实模型”为：

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

A同学构建遗漏变量的模型：

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$



西北农林科技大学  
NORTHWEST A&F UNIVERSITY



# 内生自变量情形1：遗漏变量（演示1）

下面我们对遗漏变量情形做一个整体的直观演示：

假定工资水平的“真实模型”为：

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

A同学构建遗漏变量的模型：

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

  ~~$Abl_i$~~

遗漏  $\neq$  消失

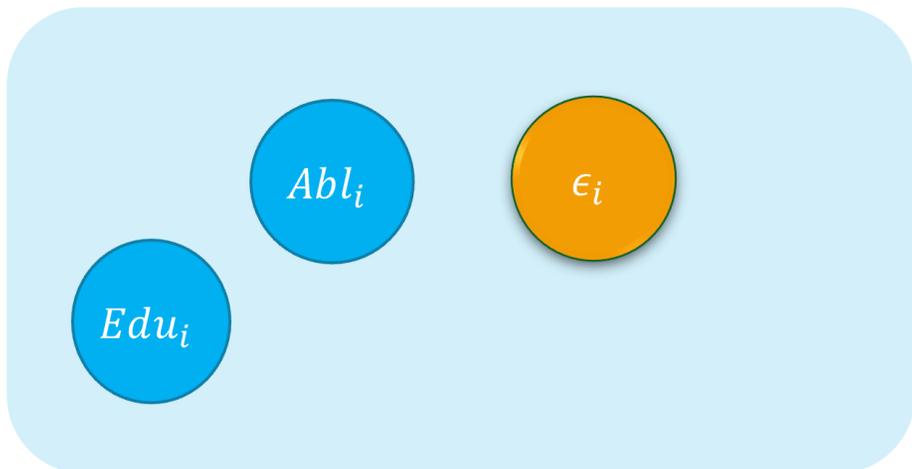


西北农林科技大学  
NORTHWEST A&F UNIVERSITY



## 内生自变量情形1：遗漏变量（演示2）

具体地，遗漏变量引发内生自变量问题的直观演示如下：



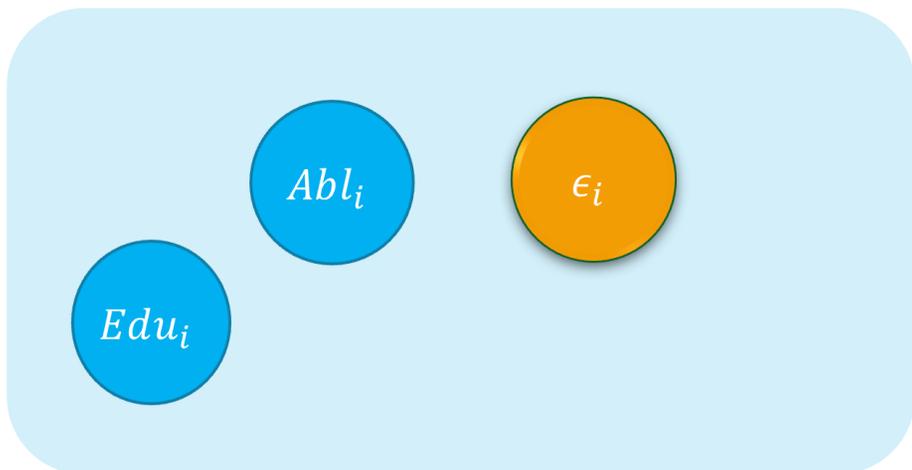
$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$



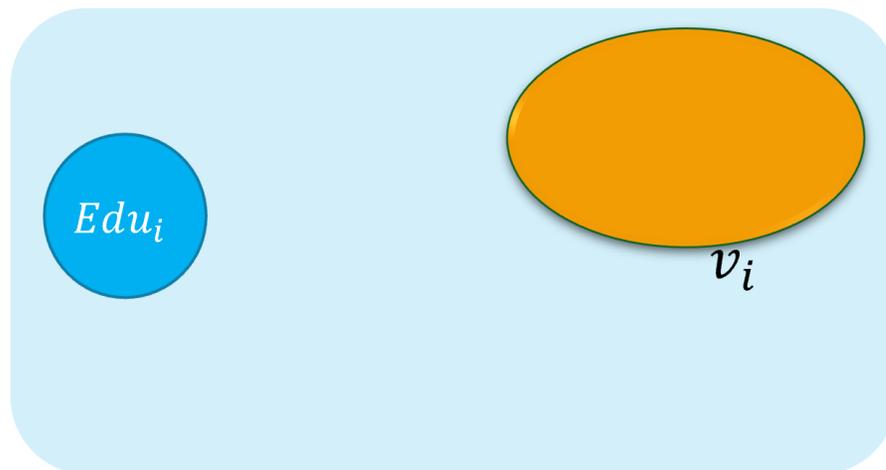


# 内生自变量情形1：遗漏变量（演示2）

具体地，遗漏变量引发内生自变量问题的直观演示如下：



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$



$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

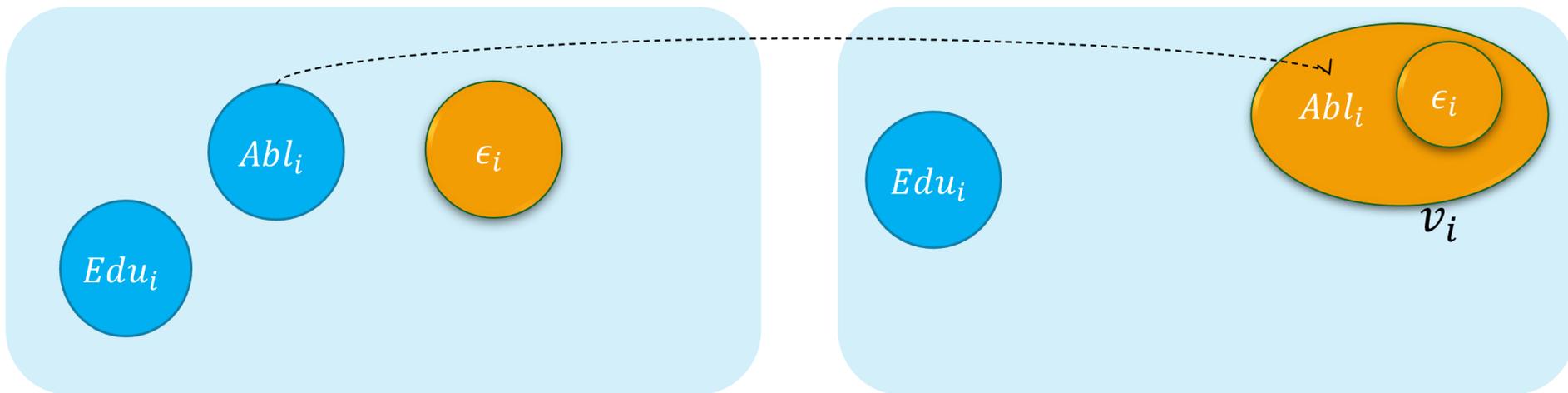
大学  
UNIVERSITY

西北  
NORTHWEST



# 内生自变量情形1：遗漏变量（演示2）

具体地，遗漏变量引发内生自变量问题的直观演示如下：



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$v_i = \beta_3 abl_i + \epsilon_i$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

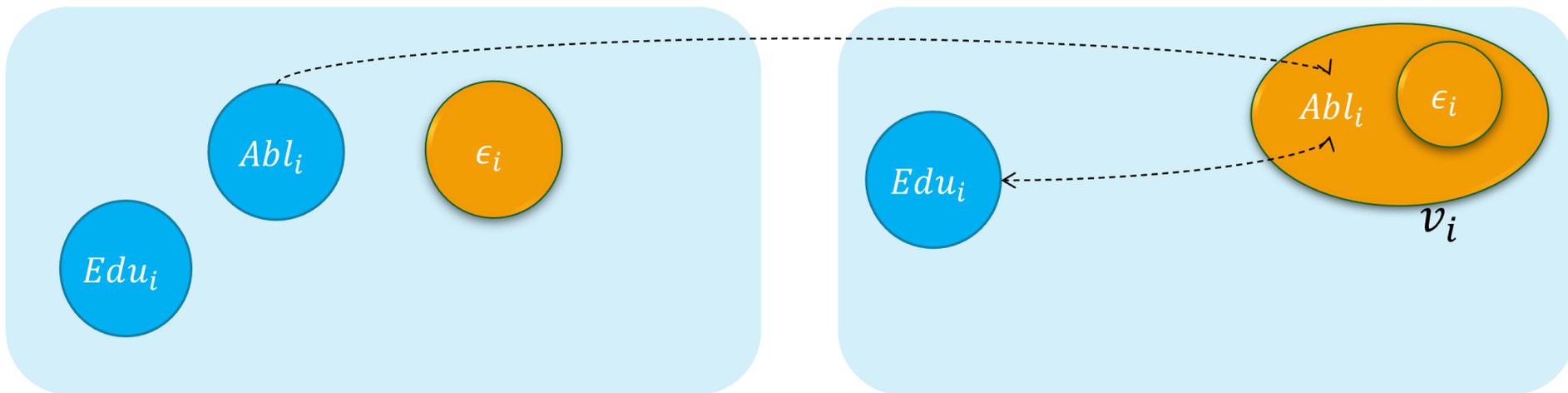
大学  
UNIVERSITY

西北  
NORTHWEST



# 内生自变量情形1：遗漏变量（演示2）

具体地，遗漏变量引发内生自变量问题的直观演示如下：



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$v_i = \beta_3 abl_i + \epsilon_i$$

$$Cov(Edu_i, Abl_i) \neq 0$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

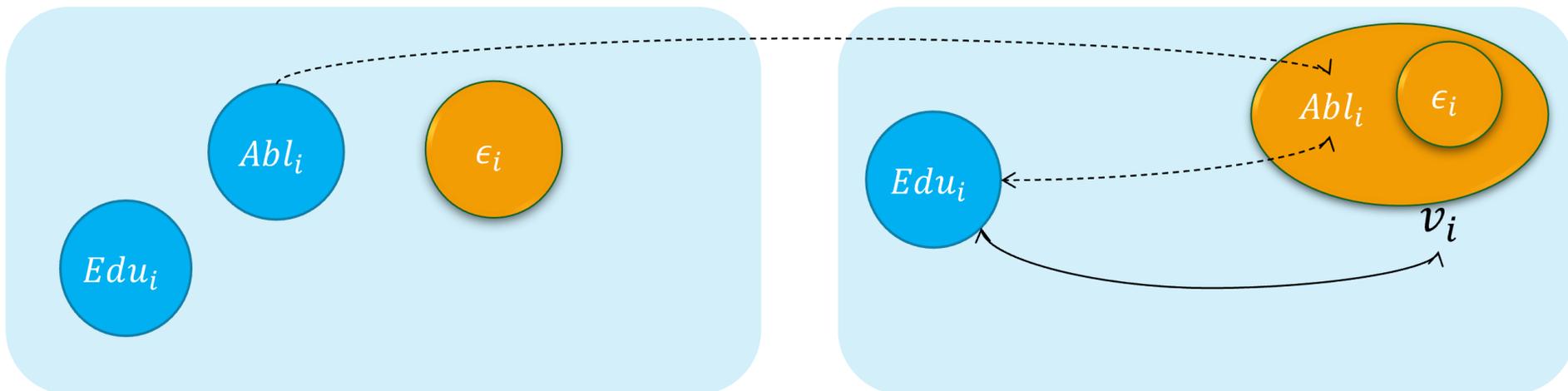
大学  
UNIVERSITY

西北  
NORTHWEST



# 内生自变量情形1：遗漏变量（演示2）

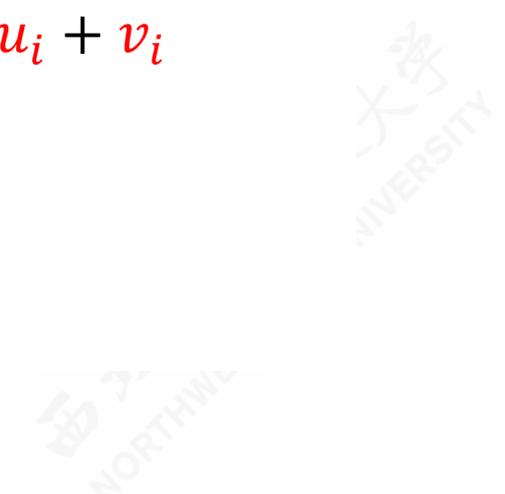
具体地，遗漏变量引发内生自变量问题的直观演示如下：



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

$$\left. \begin{aligned} v_i &= \beta_3 abl_i + \epsilon_i \\ Cov(Edu_i, Abl_i) &\neq 0 \end{aligned} \right\} \Rightarrow Cov(Edu_i, v_i) \neq 0$$





## 内生自变量情形2：测量误差

很多时候的模型中实际使用的某个自变量本身并不是准确观测的，而只是“近似物”，因此模型自变量中存在**测量误差**（measurement error）。



## 内生自变量情形2：测量误差

再次,假定工资决定的**真实模型** (real model) 是:

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + u_i \quad (\text{the assumed true model})$$

然而,因为个体的**能力变量**  $Abl$  往往无法直接观测得到,我们便会考虑使用**智商水平** 变量 ( $IQ_i$ ), 并构建如下有偏误的**代理变量模型** (proxy variable model) :

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i \quad (\text{the error specified model})$$

- 此时,智商水平  $IQ_i$  被认为是能力变量  $Abl_i$  的一个**代理变量** (proxy variable) 。
- 而实际上,能力水平变量  $Abl_i$  的内涵要远远大于智商水平变量  $IQ_i$ 。因此,受教育年数变量  $Edu$  会与随机干扰项中未纳入模型的  $Abl_i$  变量的测量误差部分存在相关关系。进而偏误模型中,  $cov(Edu_i, v_i) \neq 0$ , 从而受教育年数变量  $Edu$  具有内生自变量问题。



## 内生自变量情形2：测量误差（演示1）

下面我们对测量误差情形做一个整体的直观演示：

假定工资水平的“真实模型”为：

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$





## 内生自变量情形2：测量误差（演示1）

下面我们对测量误差情形做一个整体的直观演示：

假定工资水平的“真实模型”为：

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

B同学构建有测量误差的模型：

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

西北农林科技大学  
NORTHWEST A&F UNIVERSITY



## 内生自变量情形2：测量误差（演示1）

下面我们对测量误差情形做一个整体的直观演示：

假定工资水平的“真实模型”为：

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

B同学构建有测量误差的模型：

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$


$$Abl_i = \{IQ_i, Abl\_other_i\}$$

西北农林科技大学  
NORTHWEST A&F UNIVERSITY



## 内生自变量情形2：测量误差（演示1）

下面我们对测量误差情形做一个整体的直观演示：

假定工资水平的“真实模型”为：

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

B同学构建有测量误差的模型：

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

$$Abl_i = \{IQ_i, Abl\_other_i\}$$

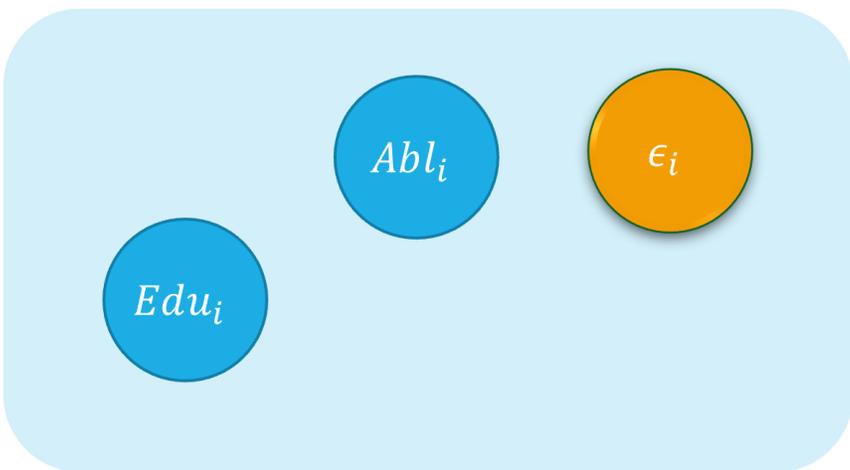
误差 ≠ 消失

西北农林科技大学  
NORTHWEST A&F UNIVERSITY



## 内生自变量情形2：测量误差（演示2）

具体地，测量误差引发内生自变量问题的直观演示如下：



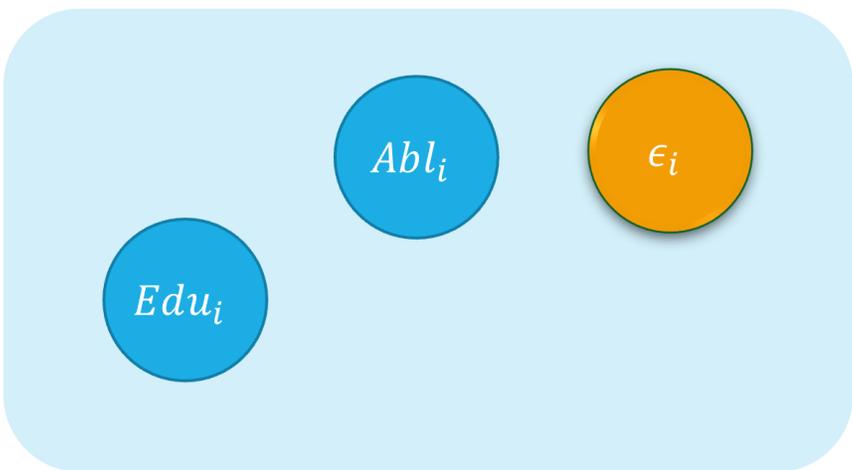
$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$



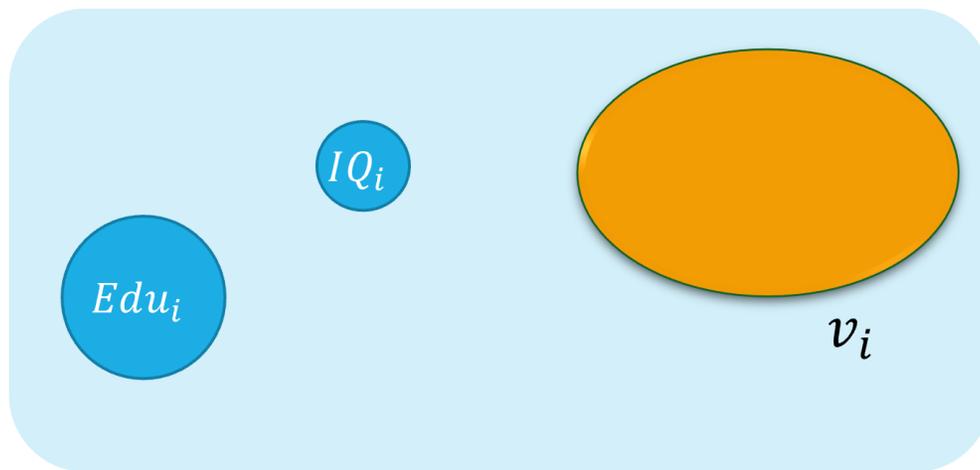


## 内生自变量情形2：测量误差（演示2）

具体地，测量误差引发内生自变量问题的直观演示如下：



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

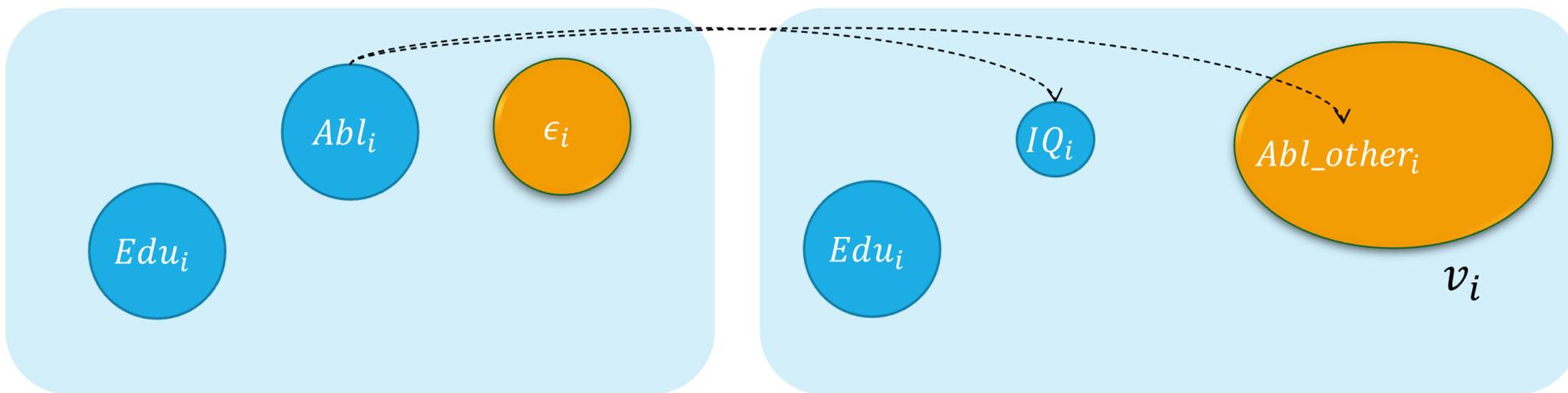


$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$



## 内生自变量情形2：测量误差（演示2）

具体地，测量误差引发内生自变量问题的直观演示如下：



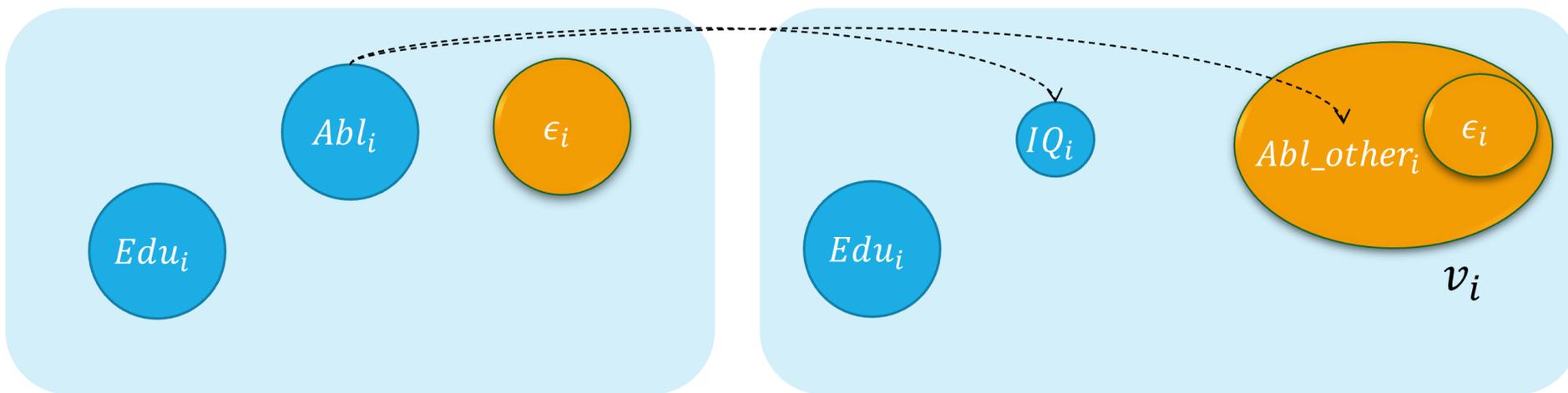
$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$



## 内生自变量情形2：测量误差（演示2）

具体地，测量误差引发内生自变量问题的直观演示如下：



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

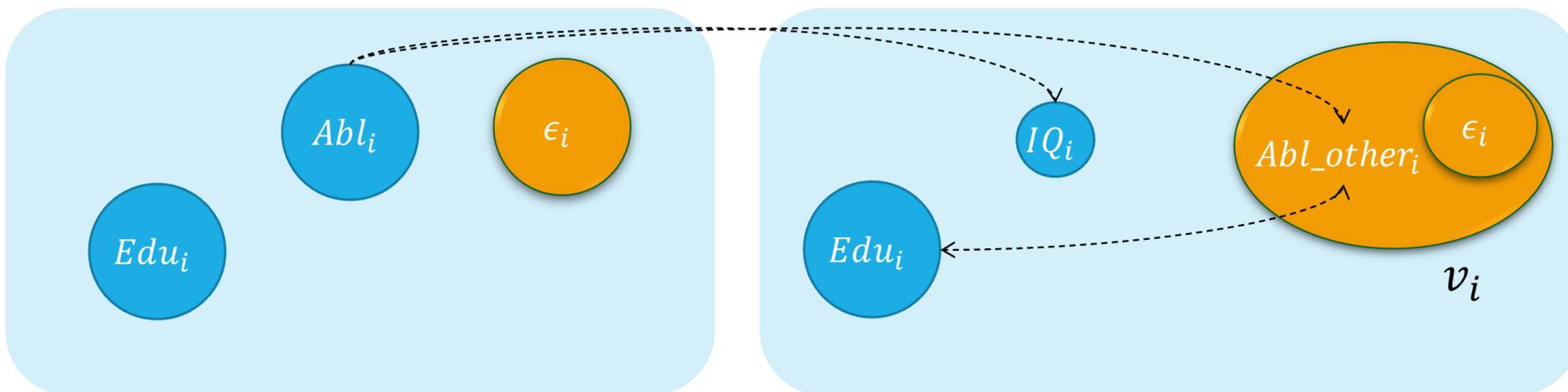
$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

$$v_i = \beta_3 Abl\_other_i + \epsilon_i$$



## 内生自变量情形2：测量误差（演示2）

具体地，测量误差引发内生自变量问题的直观演示如下：



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

$$v_i = \beta_3 Abl\_other_i + \epsilon_i$$

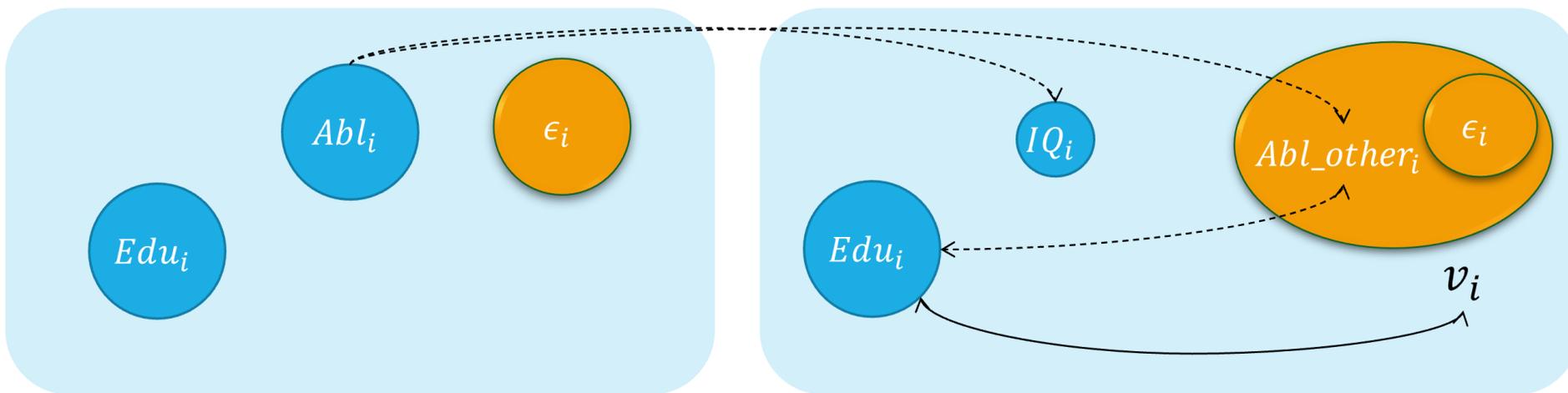
$$Cov(Edu_i, Abl\_other_i) \neq 0$$

西北  
NORTHWEST



# 内生自变量情形2：测量误差（演示2）

具体地，测量误差引发内生自变量问题的直观演示如下：



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$v_i = \beta_3 Abl\_other_i + \epsilon_i$$

$$Cov(Edu_i, Abl\_other_i) \neq 0$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

$$\Rightarrow Cov(Edu_i, v_i) \neq 0$$

西北农林科技大学  
NORTHWEST A&F UNIVERSITY



## 内生自变量情形3：序列自相关问题

自回归滞后变量模型：因变量的滞后变量（ $Y_{t-1}, \dots, Y_{t-p}, \dots$ ）作为回归元，出现在模型中。

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$

如果随机干扰项表现为一阶自相关AR(1)，也即：

$$u_t = \rho u_{t-1} + v_t$$

那么，显然  $cov(Y_{t-1}, u_{t-1}) \neq 0$ ，进而  $cov(Y_{t-1}, u_t) \neq 0$ 。因此，受教育年数变量  $Y_{t-1}$  具有内生自变量问题。



## 内生自变量情形3：序列自相关（演示1）

下面我们对序列自相关情形做一个整体的直观演示：

### 一阶自回归滞后AR(1)模型

西北农林科技大学  
NORTHWEST A&F UNIVERSITY



## 内生自变量情形3：序列自相关（演示1）

下面我们对序列自相关情形做一个整体的直观演示：

### 一阶自回归滞后AR(1)模型

$$\begin{cases} Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t & (\text{主模型}) \\ u_t = \rho u_{t-1} + \epsilon_t & (\text{辅助模型}) \end{cases}$$





# 内生自变量情形3：序列自相关（演示1）

下面我们对序列自相关情形做一个整体的直观演示：

## 一阶自回归滞后AR(1)模型

$$\begin{cases} Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t & \text{(主模型)} \\ u_t = \rho u_{t-1} + \epsilon_t & \text{(辅助模型)} \end{cases}$$



西北农林科技大学  
NORTHWEST A&F UNIVERSITY



# 内生自变量情形3：序列自相关（演示1）

下面我们对序列自相关情形做一个整体的直观演示：

## 一阶自回归滞后AR(1)模型

$$\begin{cases} Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t & (\text{主模型}) \\ u_t = \rho u_{t-1} + \epsilon_t & (\text{辅助模型}) \end{cases}$$

“隐身” ≠ 消失

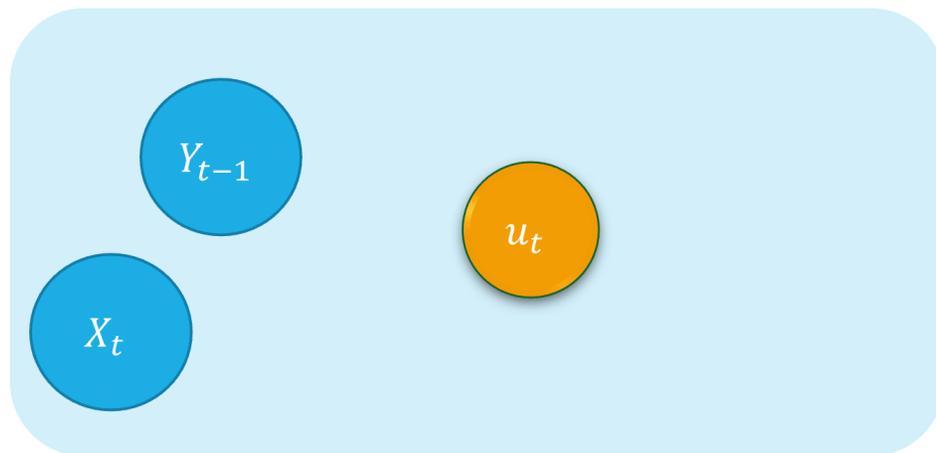


## 内生自变量情形3：序列自相关（演示2）

具体地，序列自相关引发内生自变量问题的直观演示如下：

(主模型)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$





## 内生自变量情形3：序列自相关（演示2）

具体地，序列自相关引发内生自变量问题的直观演示如下：

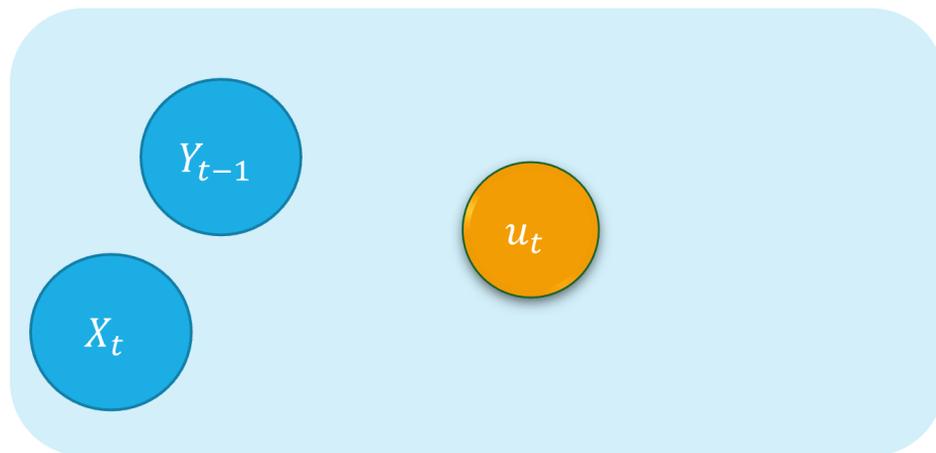
(主模型)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$



(衍生模型)

$$Y_{t-1} = \beta_1 + \beta_2 Y_{t-2} + \beta_3 X_{t-1} + u_{t-1}$$





# 内生自变量情形3：序列自相关（演示2）

具体地，序列自相关引发内生自变量问题的直观演示如下：

(主模型)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$

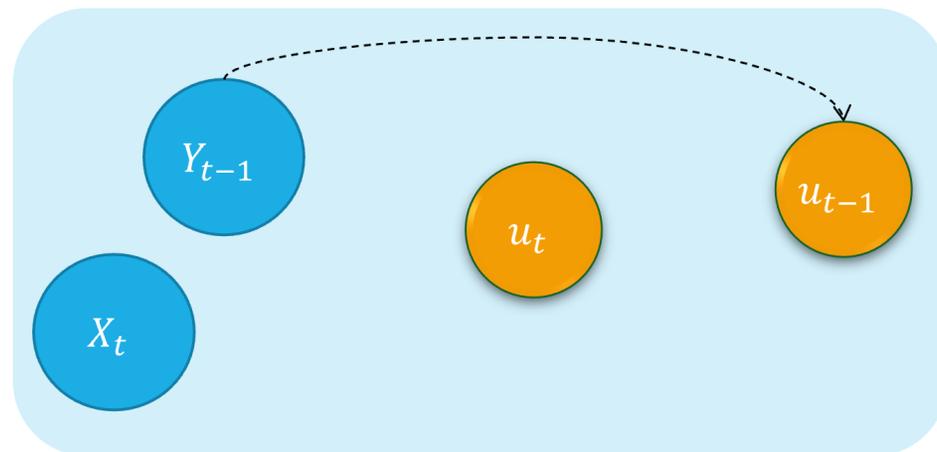


(衍生模型)

$$Y_{t-1} = \beta_1 + \beta_2 Y_{t-2} + \beta_3 X_{t-1} + u_{t-1}$$



$$\text{Cov}(Y_{t-1}, u_{t-1}) \neq 0$$





# 内生自变量情形3：序列自相关（演示2）

具体地，序列自相关引发内生自变量问题的直观演示如下：

(主模型)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$



(衍生模型)

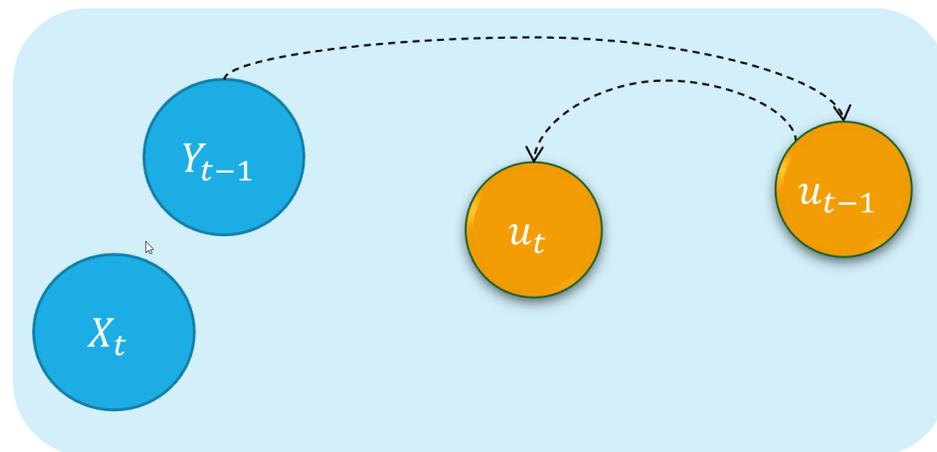
$$Y_{t-1} = \beta_1 + \beta_2 Y_{t-2} + \beta_3 X_{t-1} + u_{t-1}$$



$$\text{Cov}(Y_{t-1}, u_{t-1}) \neq 0$$

$$u_t = \rho u_{t-1} + \epsilon_t$$

(辅助模型)





# 内生自变量情形3：序列自相关（演示2）

具体地，序列自相关引发内生自变量问题的直观演示如下：

(主模型)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$



(衍生模型)

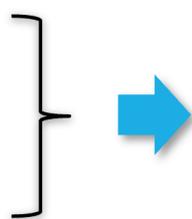
$$Y_{t-1} = \beta_1 + \beta_2 Y_{t-2} + \beta_3 X_{t-1} + u_{t-1}$$



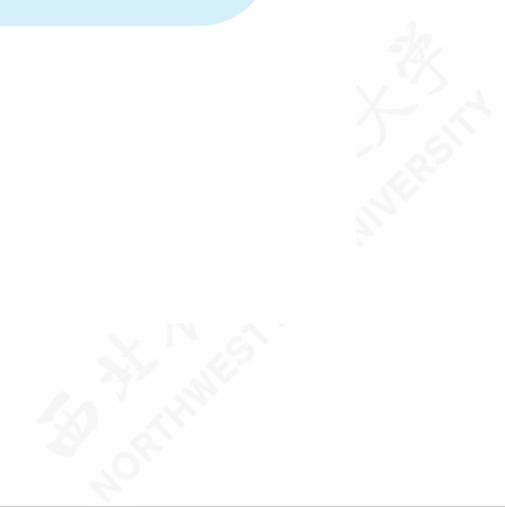
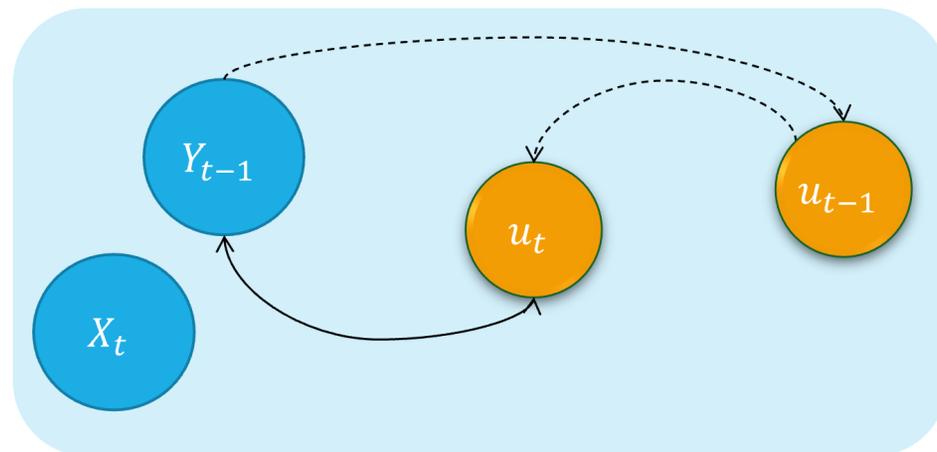
$$Cov(Y_{t-1}, u_{t-1}) \neq 0$$

$$u_t = \rho u_{t-1} + \epsilon_t$$

(辅助模型)



$$Cov(Y_{t-1}, u_t) \neq 0$$





## 内生自变量情形4：方程联立性

对于供需联立方程的结构化形式：

$$\begin{cases} \text{Demand: } Q_i = \alpha_1 + \alpha_2 P_i + u_{di} \\ \text{Supply: } Q_i = \beta_1 + \beta_2 P_i + u_{si} \end{cases}$$

众所周知，因为价格  $P_i$  变动会影响供给量和需求量  $Q_i$  的变动；反之亦然。两者之间存在相互反馈影响机制。

因此，可以证明  $cov(P_i, u_{di}) \neq 0$ ，而且  $cov(P_i, u_{si}) \neq 0$ ，从而产生内生性问题。



## 内生自变量情形4：方程联立性（演示1）

下面我们对方程联立性情形做一个整体的直观演示：

$$\left\{ \begin{array}{l} Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 I_i + u_{di} \quad (\text{需求 } \alpha_2 < 0, \alpha_3 > 0) \end{array} \right.$$





## 内生自变量情形4：方程联立性（演示1）

下面我们对方程联立性情形做一个整体的直观演示：

$$\begin{cases} Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 I_i + u_{di} & (\text{需求 } \alpha_2 < 0, \alpha_3 > 0) \\ Q_i = \beta_1 + \beta_2 P_i + u_{si} & (\text{供给 } \beta_2 > 0) \end{cases}$$

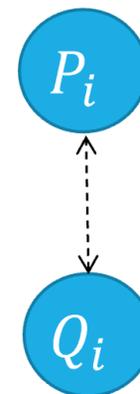




# 内生自变量情形4：方程联立性（演示1）

下面我们对方程联立性情形做一个整体的直观演示：

$$\begin{cases} Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 I_i + u_{di} & (\text{需求 } \alpha_2 < 0, \alpha_3 > 0) \\ Q_i = \beta_1 + \beta_2 P_i + u_{si} & (\text{供给 } \beta_2 > 0) \end{cases}$$

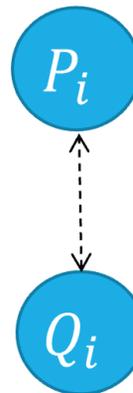




# 内生自变量情形4：方程联立性（演示1）

下面我们对方程联立性情形做一个整体的直观演示：

$$\begin{cases} Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 I_i + u_{di} & (\text{需求 } \alpha_2 < 0, \alpha_3 > 0) \\ Q_i = \beta_1 + \beta_2 P_i + u_{si} & (\text{供给 } \beta_2 > 0) \end{cases}$$



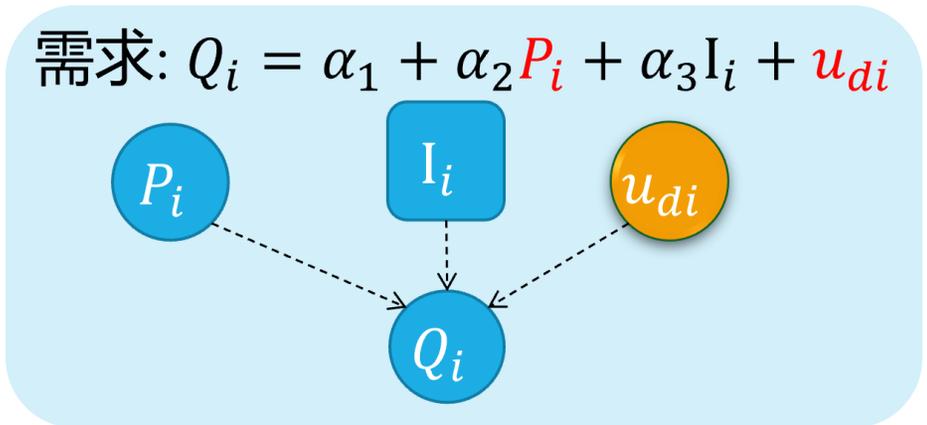
“复杂” ≠ 消失





# 内生自变量情形4：方程联立性（演示2）

具体地，方程联立性引发内生自变量问题的直观演示如下：



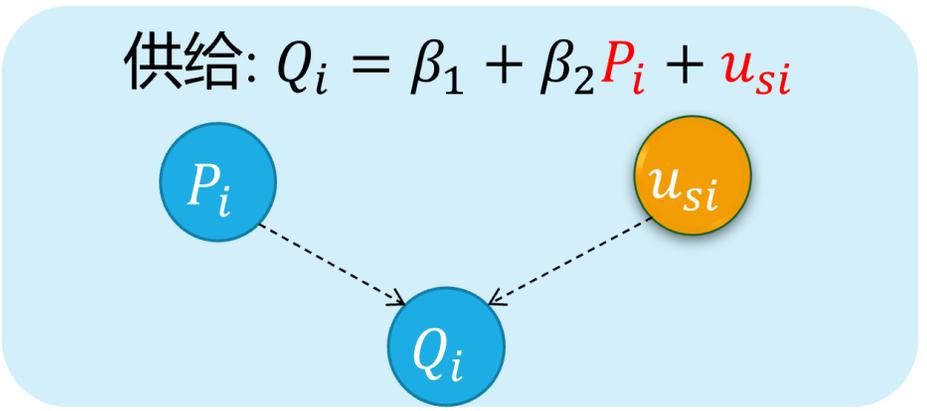
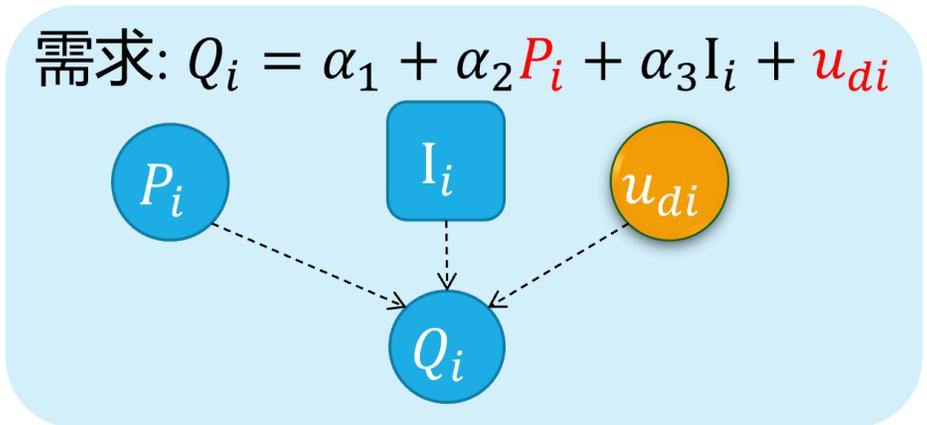
大学  
UNIVERSITY

西  
NORTH



# 内生自变量情形4：方程联立性（演示2）

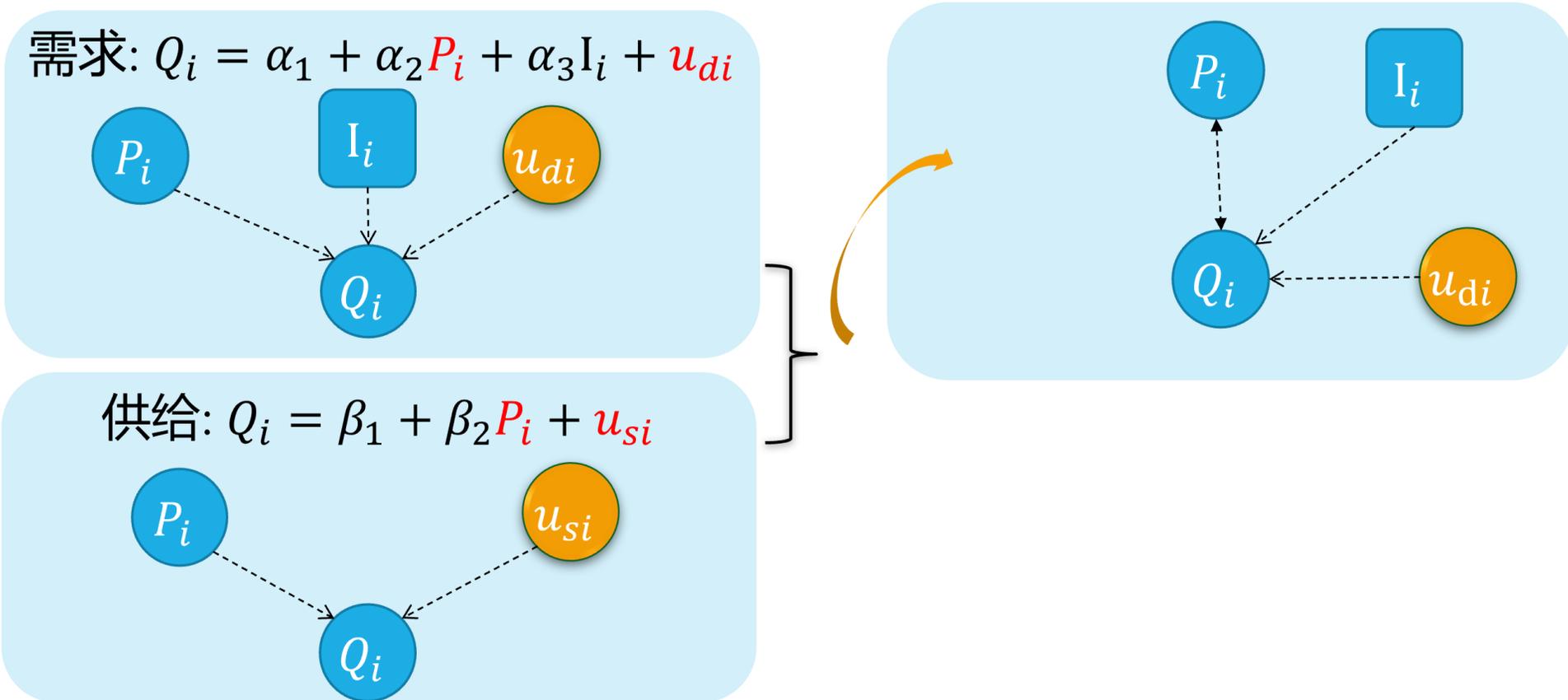
具体地，方程联立性引发内生自变量问题的直观演示如下：





# 内生自变量情形4：方程联立性（演示2）

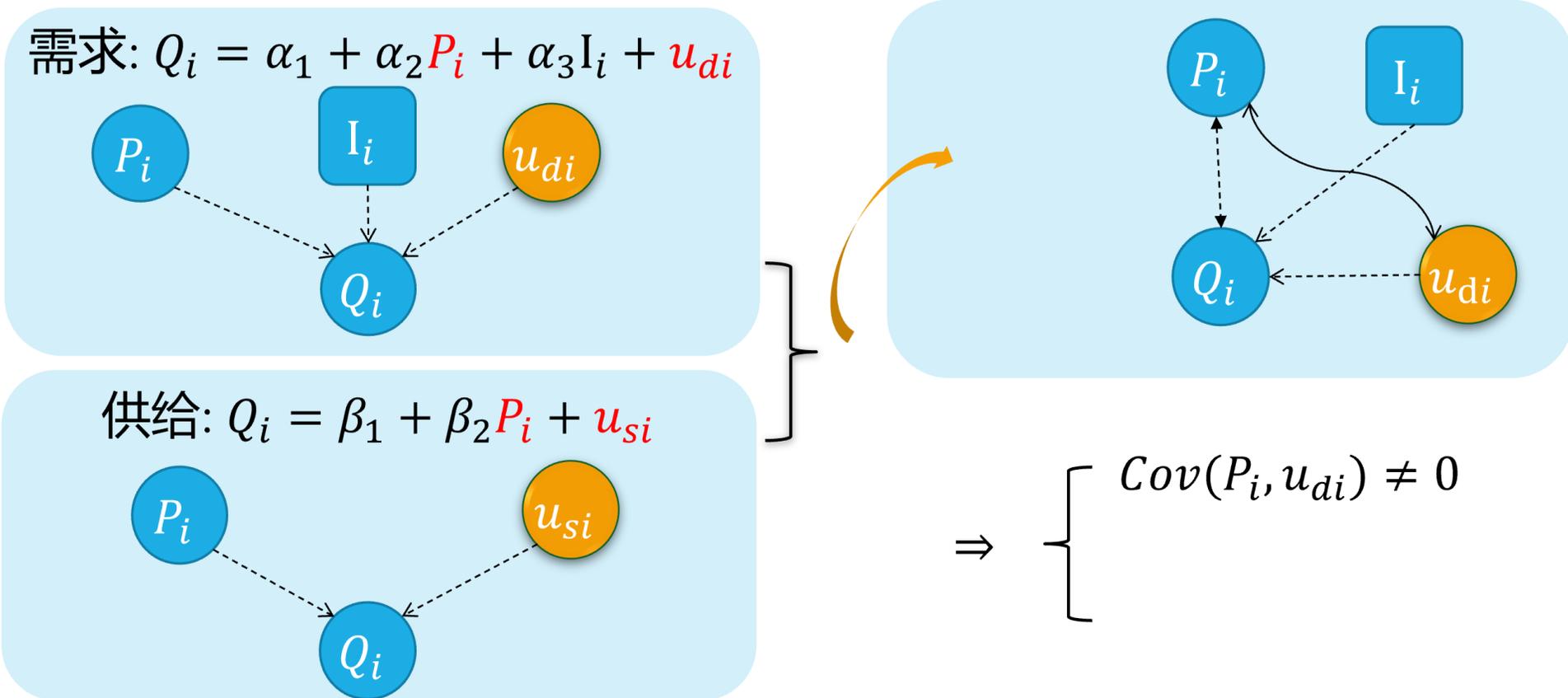
具体地，方程联立性引发内生自变量问题的直观演示如下：





# 内生自变量情形4：方程联立性（演示2）

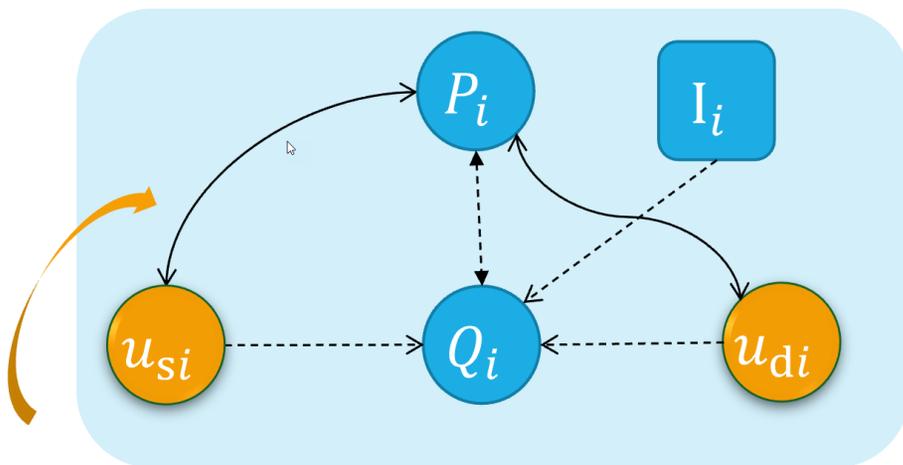
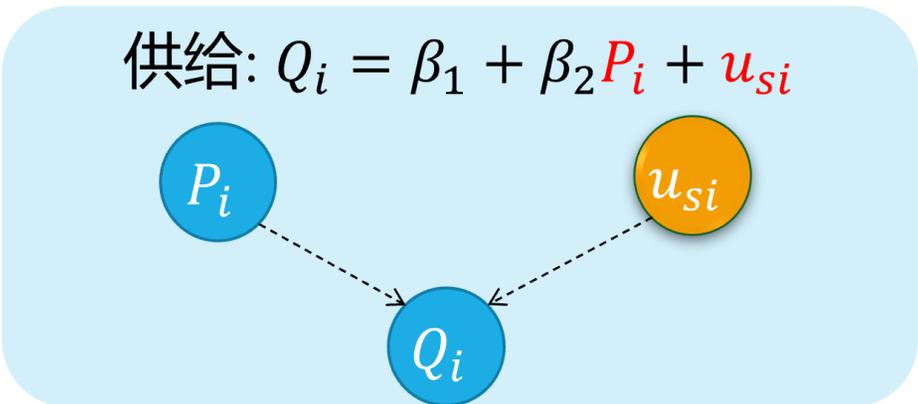
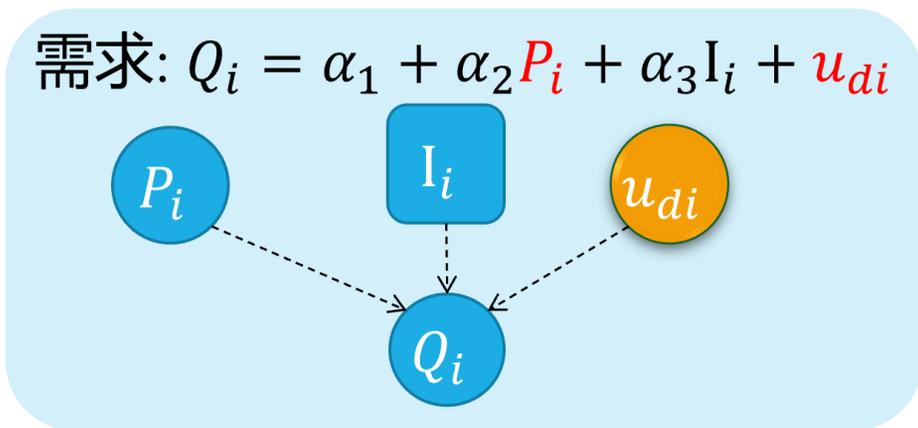
具体地，方程联立性引发内生自变量问题的直观演示如下：





# 内生自变量情形4：方程联立性（演示2）

具体地，方程联立性引发内生自变量问题的直观演示如下：



$$\Rightarrow \begin{cases} Cov(P_i, u_{di}) \neq 0 \\ Cov(P_i, u_{si}) \neq 0 \end{cases}$$

大学  
UNIVERSITY

西  
NORTH



# 学习成绩与逃课次数的例子

假设“真实模型”是：

$$score_i = \alpha_1 + \alpha_2 skipped_i + \alpha_3 abil_i + \alpha_4 mot_i + \alpha_5 income_i + u_i$$

一个遗漏了重要变量的“偏误模型”是：

$$score_i = \beta_1 + \beta_2 skipped_i + v_i$$

- 学习成绩受到逃课次数的影响，但是我们也很担心以上模型中  $skipped_i$  与  $v_i$  中的某些因素相关，例如越有能力  $abil_i$ 、越积极  $mot_i$  的学生，逃课也越少。
- 因为自变量  $skipped_i$  可能与随机干扰项  $v_i$  相关。此时，对于以上简单的回归，可能得不出可靠的估计。



# 学习成绩与逃课次数的例子

$$score_i = \beta_1 + \beta_2 skipped_i + v_i$$

逃课次数  $skipped_i$  的工具变量  $Z_i$  有哪些可供备选的呢？

- 宿舍跟上课地点的距离  $distance$ 。我们一般认为，它与逃课次数相关  $skipped_i$ ，但是它与  $v_i$  中的某些因素也会相关么？
- 如果收入水平  $income$  确实影响了学习成绩，但是模型却没有引入收入水平  $income$  变量，也就意味着  $v_i$  中包含了遗漏的重要变量——收入水平  $income$ 。此时，距离  $distance$  就会与收入水平  $income$  相关，进而与  $v_i$  相关。——因为收入少的学生，更倾向于在外租房（合租）；收入多的学生，更倾向于住校。



## 17.2 内生变量法下的估计问题



# 造成不一致性估计：遗漏变量情形

一般而言，由于A2不成立，相关重要变量的**遗漏**，会导致OLS方法的估计不一致。

假设真实的工资模型是**教育和能力**的函数：

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + u_i \quad (a)$$

然而个人能力往往是不能被观察到的。

因此能力被包含在误设模型的**随机干扰项**中：

$$Wage_i = \beta_1 + \beta_2 Edu_i + e_i \quad (b)$$

其中： $e_i = \beta_3 Abl_i + \epsilon_i$

问题是能力不仅影响工资，而且能力越强的人受教育的时间越长，这就导致了随机误差项和教育变量之间的正相关。 $Cov(Edu_i, e_i) > 0$ 。

牢记：

- 遗漏并不等于消失！"omission" does not means "disappear"！



# 造成不一致性估计：测量误差情形

如果真实模型为：

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (1)$$

我们希望观测自变量  $X$  对因变量  $Y$  的真实影响，但是很可能我们无法完全地观测得到自变量  $X$ ，从而退一步采用一个可以观测到的代理变量（如  $X^*$ ）。

$$X_i^* = X_i - v_i \quad (2)$$

其中：

- 随机变量  $v_i$  的期望为 0，方差为  $\sigma_v^2$
- $X_i, \epsilon_i$  与  $v_i$  是互为独立的（**pairwise independent**）。

从而，我们构造了一个包含测量误差的误设模型：

$$Y_i = \alpha_0 + \alpha_1 X_i^* + v_i \quad (3)$$



# 造成不一致性估计：测量误差情形

更一般地：

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{eq(1) assumed true model}$$

$$X_i^* = X_i - v_i \quad \text{eq(2) proxy variable}$$

$$X_i = X_i^* + v_i \quad \text{eq(3)}$$

$$Y_i = \beta_0 + \beta_1 X_i^* + u_i \quad \text{eq(4) error specified model}$$

把方程 (3) 带入方程 (1)，可以得到方程(5)：

$$Y_i = \beta_0 + \beta_1 X_i^* + \epsilon_i = \beta_0 + \beta_1 (X_i^* + v_i) + \epsilon_i = \beta_0 + \beta_1 X_i^* + (\epsilon_i + \beta_1 v_i) \quad \text{eq(5)}$$

这将表明**误设模型**中的随机误差项  $u_i = (\epsilon_i + \beta_1 v_i)$ ，从而导致  $\text{Cov}(X_i^*, u_i) \neq 0$ ，根据高斯马尔可夫定理，OLS方法将不能得到一致性估计量（具体见下一页）。



# 造成不一致性估计：测量误差情形

容易证明： $E(u_i) = E(\epsilon_i + \beta_1 v_i) = E(\epsilon_i) + \beta_1 E(v_i) = 0$

然而：

$$\begin{aligned}\text{Cov}(X_i^*, u_i) &= E[(X_i^* - E(X_i^*))(u_i - E(u_i))] \\ &= E(X_i^* u_i) \\ &= E[(X_i - v_i)(\epsilon_i + \beta_1 v_i)] \\ &= E[X_i \epsilon_i + \beta_1 X_i v_i - v_i \epsilon_i - \beta_1 v_i^2] \leftarrow \text{(pairwise independent)} \\ &= -E(\beta_1 v_i^2) \\ &= -\beta_1 \text{Var}(v_i) \\ &= -\beta_1 \sigma_{v_i}^2 \neq 0\end{aligned}$$

因此，方程4中的自变量  $X^*$  是内生自变量 (**endogenous**)，从而OLS的系数估计量  $\beta_1$  是不一致的。



# OLS估计: 违背CLRm假设A2

一般而言, 如果CLRm假设中的**A2**被违背, OLS估计量将会是有偏的 (biased estimator) :

我们已知, 真实参数  $\hat{\beta}$  的OLS估计量理论公式为:

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon \quad (6)$$

我们可以两边同时取期望:

$$\begin{aligned} E(\hat{\beta}) &= \beta + E\left((X'X)^{-1}X'\epsilon\right) \\ &= \beta + E\left(E\left((X'X)^{-1}X'\epsilon|X\right)\right) \\ &= \beta + E\left((X'X)^{-1}X'E(\epsilon|X)\right) \neq \beta \end{aligned}$$

如果CLRm假设中**A2**  $E(\epsilon|X) = 0$ 被违背, 也即意味着  $E(\epsilon|X) \neq 0$ , 从而OLS估计量是有偏的。



# OLS估计: 一致性估计量

那么, 在什么条件下我们才能得到一致估计量呢?

$$\begin{aligned} p \lim \hat{\beta} &= \beta + p \lim \left( (X'X)^{-1} X'\epsilon \right) = \beta + p \lim \left( \left( \frac{1}{n} X'X \right)^{-1} \frac{1}{n} X'\epsilon \right) \\ &= \beta + p \lim \left( \frac{1}{n} X'X \right)^{-1} \times p \lim \left( \frac{1}{n} X'\epsilon \right) \end{aligned}$$

通过弱大数定律(Weak Law of Large Numbers, WLLN):

$$\frac{1}{n} X'\epsilon = \frac{1}{n} \sum_{i=1}^n X_i \epsilon_i \xrightarrow{p} E(X_i \epsilon_i)$$

因此如果  $E(X_i \epsilon_i) = 0$ , 则  $\hat{\beta}$  是一致估计量。

需要注意的是:  $E(X_i \epsilon_i) = 0$  比CLRM假设中的A2  $E(\epsilon|X) = 0$  更容易满足。因此, 一些有偏的估计量, 在大样本情况下也可以是渐进一致的。



# 工资案例：误设模型

考虑如下的“误设模型”：

$$lwage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 expersq_i + v_i$$

如前所述，该“误设模型”的问题在于，随机误差项中包含不可观测的重要变量，例如个人能力水平（ $Abl_i$ ），它同时对工资水平因变量和受教育程度自变量产生影响。

换言之，自变量工资水平与随机干扰项相关，也即  $cov(educ_i, v_i) \neq 0$ ，因此它是内生自变量（endogenous regressor）。

注意：

- 在实践中，我们将使用受教育年数作为 **educ** 的代理变量，这本身也会带来前面提到的误差测量问题。



# 案例变量说明

研究者关注428名已婚女性时均工资  $wage$  与其受教育年数  $educ$  之间的关系，并考虑如下变量：

## 变量说明

vars	mark
lwage	时均工资
educ	受教育年数
exper	就业次数
fatheduc	父亲的受教育年数
motheduc	母亲的受教育年数
inlf	是否是劳动力
hours	工作时长

Showing 1 to 7 of 22 entries

Previous

1

2

3

4

Next



# 案例原始数据

数据集 (n=428)

id	lwage	educ	exper	expersq	fatheduc	motheduc
1	1.21	12	14	196	7	12
2	0.33	12	5	25	7	7
3	1.51	12	15	225	7	12
4	0.09	12	6	36	7	7
5	1.52	14	7	49	14	12
6	1.56	12	33	1089	7	14
7	2.12	16	11	121	7	14
8	2.06	12	35	1225	3	3

Showing 1 to 8 of 428 entries

Previous

1

2

3

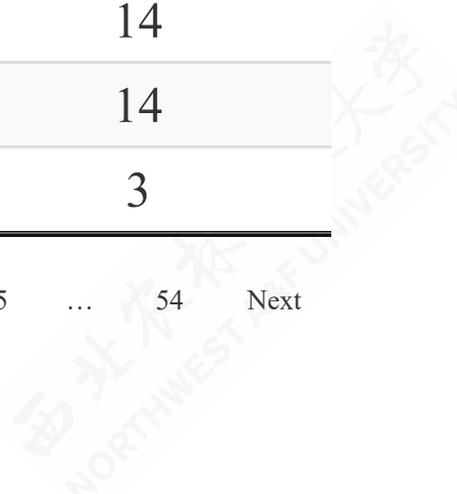
4

5

...

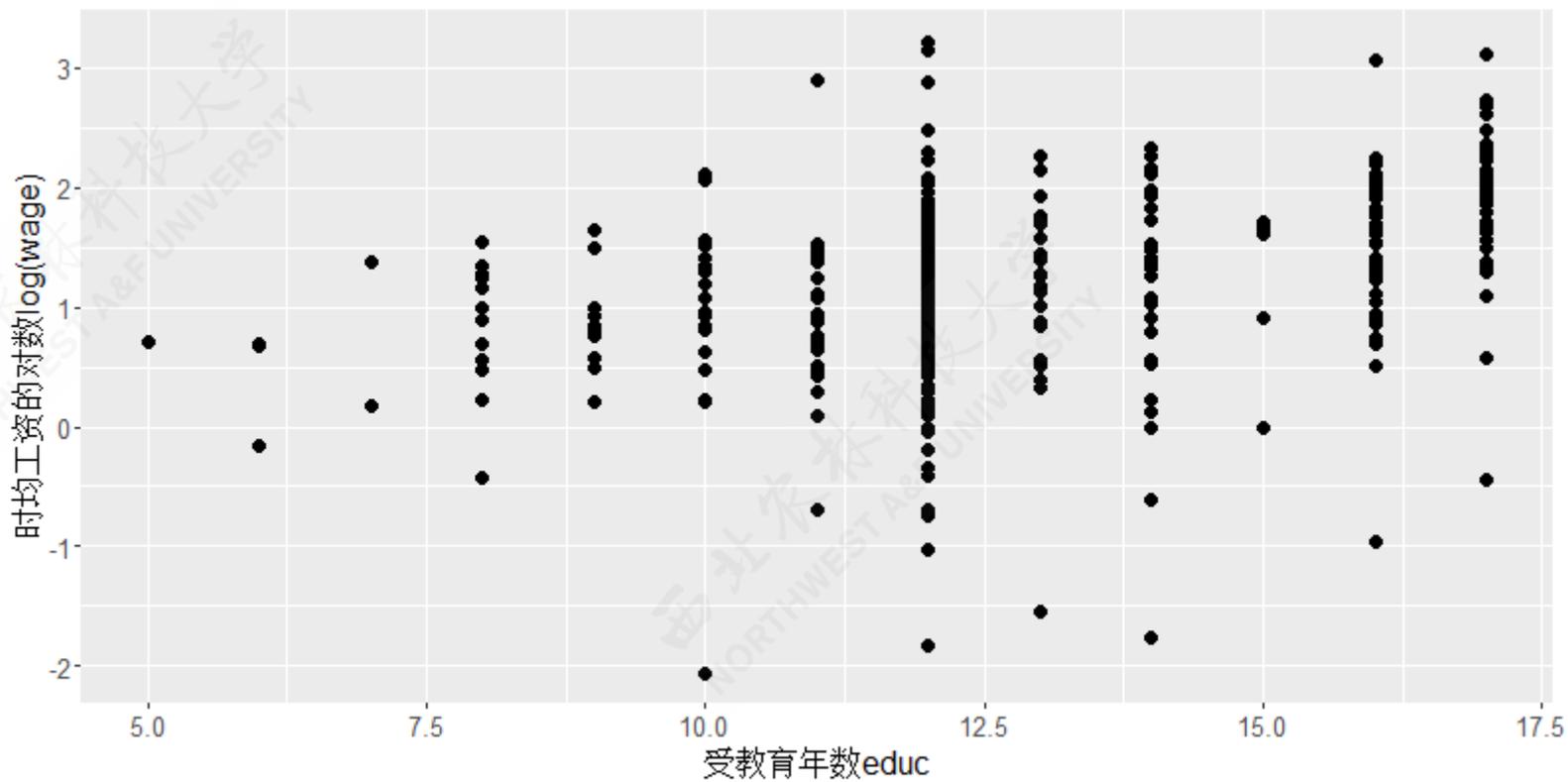
54

Next





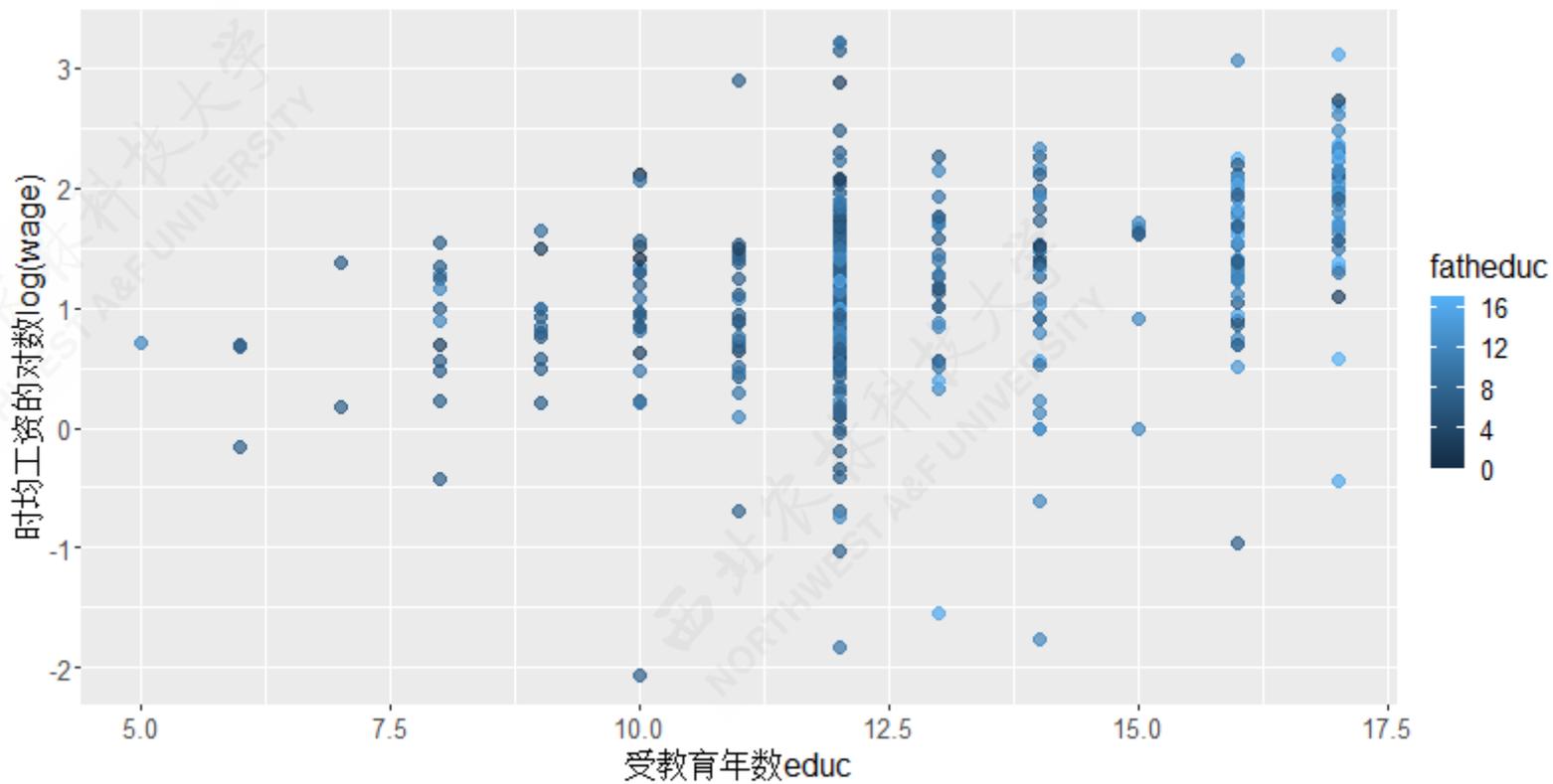
# 案例散点图1



受教育年数与时均工资的散点图



# 案例散点图2



考虑父亲受教育年数的散点图



# 案例误设模型的OLS回归

如果直接构建如下的“偏误模型”，并坚持采用OLS估计：

```
mod_origin <- formula(lwage ~ educ +exper+expersq)  
ols_origin <- lm(formula = mod_origin, data = mroz)
```

$$lwage = + \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 expersq + u_i$$

$$\widehat{lwage} = - 0.52 \quad + 0.11educ \quad + 0.04exper \quad - 0.00expersq$$

$$(t) \quad (-2.6282) \quad (7.5983) \quad (3.1549) \quad (-2.0628)$$

$$(se) \quad (0.1986) \quad (0.0141) \quad (0.0132) \quad (0.0004)$$

$$(fitness) R^2 = 0.1568; \bar{R}^2 = 0.1509$$

$$F^* = 26.29; p = 0.0000$$

## 17.3 工具变量及其选择



## 工具变量：缘由

至此，我们已经了解到如果模型出现一个或多个内生自变量，则参数  $\beta$  的 OLS 估计是有偏的。

OLS 方法的估计“问题”，来自于我们要求的 CLRM 假设中的  $E(X_i \epsilon_i) = 0$ ，这意味着我们相信样本数据满足：

$$X'e = 0$$

但是，实际上自变量与随机误差项存在相关关系，也即  $E(X_i \epsilon_i) \neq 0$ 。



# 工具变量：缘由

如果我们能够找到这样的一些解释变量（explanatory variables） $Z$ ，它们满足如下条件：

- 相关性（Relevance）： $Z$ 与 $X$ 相关
- 外生性（Exogeneity）： $Z$ 与随机干扰项 $\epsilon$ 不相关

我们称满足以上条件的变量 $Z$ 为工具变量（**Instrumental Variables, IV**）。



# 工具变量：估计量

正确使用工具变量后，参数估计量  $\hat{\beta}_{IV}$  可以表达为如下的正则表达式（normal equation）——更准确地是矩条件（moment condition）形式：

$$\mathbf{Z}'\hat{\epsilon} = \mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\beta}_{IV}) = 0$$

假定  $\mathbf{Z}'\mathbf{X}$  是非奇异方阵（non singular square matrix），则有：

$$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

上述关于  $\mathbf{Z}'\mathbf{X}$  是非奇异方阵的条件，直觉上是可以得到满足的，只要我们的工具变量<sup>[1]</sup>数不少于模型中的自变量数。

尽管如此，工具变量法下的参数估计量  $\hat{\beta}_{IV}$  在有限样本下仍然是有偏的，但是可以证明它是渐进一致的。

[1] 模型中的外生自变量，本质上也可以视作为工具变量。



# 工具变量：一致性

下面我们来证明  $\hat{\beta}_{IV}$  是渐进一致的。

$$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\beta + \epsilon) = \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\epsilon$$

$$\begin{aligned} p \lim \hat{\beta}_{IV} &= \beta + p \lim \left( (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\epsilon \right) \\ &= \beta + \left( p \lim \left( \frac{1}{n}\mathbf{Z}'\mathbf{X} \right) \right)^{-1} p \lim \left( \frac{1}{n}\mathbf{Z}'\epsilon \right) = \beta \end{aligned}$$

• 保证相关性条件(Relevance)

$$\begin{aligned} p \lim \left( \frac{1}{n}\mathbf{Z}'\mathbf{X} \right) &= p \lim \left( \frac{1}{n} \sum z_i X_i' \right) \\ &= E(Z_i X_i') \neq 0 \end{aligned}$$

• 保证内生性条件(Exogeneity)

$$\begin{aligned} p \lim \left( \frac{1}{n}\mathbf{Z}'\epsilon \right) &= p \lim \left( \frac{1}{n} \sum Z_i \epsilon_i \right) \\ &= E(Z_i \epsilon_i) = 0 \end{aligned}$$



# 工具变量：推断

下面我们来看一下随机干扰项方差  $\sigma^2$  的工具变量法估计情况。

$$\hat{\sigma}_{IV}^2 = \frac{\sum e_i^2}{n - k} = \frac{(y - X\hat{\beta}_{IV})' (y - X\hat{\beta}_{IV})}{n - k}$$

可以证明它是真实参数的无偏估计了（证明略）。

基于此，我们才可以进行后续各种假设检验。



# 工具变量的选择

然而，找到有效的工具量并非易事，它本身就是工具变量估计方法的一大现实困难。因为：

- 优良的工具变量需要同时满足相关性和外生性两个严苛的条件。
- 一个段子: If you can find a valid instrumental variable, you can get PhD from MIT.



# 工具变量的选择

我们可以证明IV估计量  $\hat{\beta}_{IV}$  的渐进方差 (asymptotic variance) 等于 (证明略) :

$$\text{Var}(\hat{\beta}_{IV}) = \sigma^2 (\mathbf{Z}'\mathbf{X})^{-1} (\mathbf{Z}'\mathbf{Z}) (\mathbf{X}'\mathbf{Z})^{-1}$$

其中:

- $\mathbf{X}'\mathbf{Z}$  是工具变量和自变量的协方差矩阵 (covariances matrix) 。
- 如果二者的相关程度较低, 则协方差矩阵  $\mathbf{X}'\mathbf{Z}$  的元素取值会接近于0, 因此逆矩阵  $(\mathbf{X}'\mathbf{Z})^{-1}$  元素取值会非常大。最后, 参数估计量的方差  $\text{Var}(\hat{\beta}_{IV})$  也会非常大, 也即估计精度会非常低。



# 工具变量的选择

对于误设模型（存在内生自变量问题）：

$$y = X\beta + v$$

一个**基本策略**是构造全体工具变量  $Z = (X_{ex}, X^*)$ ，其中：

- 工具变量  $X_{ex}$  是那些明确出现在模型中的、且被认定为**外生**的自变量。
- 其他工具变量  $X^*$  是那些没有明确出现在模型中、但是与模型密切相关的、通过某种努力找到的**外生**变量。



# 工具变量的选择

显然，如果模型中的自变量  $X$  被认定为都是外生的，那么  $X = Z$ ，因而高斯马尔可夫定律(Gauss-Markov theorem)是成立的。并且我们需要注意的是：

- IV估计量  $\hat{\beta}_{IV}$  并不会显示任何绝对估计效率的特征。
- 我们只能说它具有相对估计效率。换言之，我们只能通过不断选择更优的工具变量积合，从而使得在众多IV估计量中，能够找到相对更好的估计量。



# 多个工具变量可供选择的情形

下面考虑另一种情形，此时我们找到的工具变量数目要远多于内生自变量数目（后面我们会知道，这属于过度识别情形,over-identification）。

根据相对估计效率原则，我们将会从工具变量集中找到那些与自变量  $X$  高度相关的，从而使得IV估计量的方差最小化！



# 多个工具变量可供选择的情形

最好的办法就是：

- 我们首先使用OLS方法，把  $\mathbf{X}$  的每一列，都对全部工具变量  $\mathbf{Z}$  进行回归，从而得到拟合变量  $\hat{\mathbf{X}}$ ：

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{ZF}$$

- 然后，我们使用拟合得到的  $\hat{\mathbf{X}}$  作为新的自变量，再与因变量  $\mathbf{y}$  进行OLS回归，从而得到高斯马尔可夫一致性估计量(证明过程见下一页)：

$$\hat{\beta}_{IV} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} = (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

实际上，这就是我们经常听说的两阶段最小二乘法（2SLS）。



# 工具变量法解决方案：遗漏变量情形

假定如下的故事背景：

$$Wage_i = \beta_0 + \beta_1 Edu_i + \beta_2 Abl_i + \epsilon_i \quad (\text{ture model})$$

$$Wage_i = \beta_0 + \beta_1 Edu_i + v_i \quad (\text{error specification model})$$

其中， $v_i = \beta_2 Abl_i + \epsilon_i$ 。此时， $Edu$  是一个内生自变量。



# 工具变量法解决方案：遗漏变量情形

假设我们能够找到满足如下条件的工具变量  $Z$ :

首先:

- $Z$  不会直接影响因变量  $Wages$
- $Z$  与  $v$  不相关, 也即:

$$\text{Cov}(v, z) = 0$$

其次:

- $Z$  至少要与内生变量  $Edu$  相关 (relevance)

$$\text{Cov}(Z, Edu) \neq 0$$

这一条件是否满足, 可以利用如下简单OLS回归的  $\alpha_2$  显著性检验进行判断:

$$Edu_i = \alpha_1 + \alpha_2 Z_i + u_i$$



# 工具变量法解决方案：遗漏变量情形

一些经济学家建议使用家庭背景变量作为内生自变量  $Edu$  的工具变量：

- 例如，母亲受教育程度  $motherEdu$  与子代的受教育程度  $Edu$  相关。然而它跟子代的能力  $Abl$  可能存在一定相关关系。
- 又例如，家庭中兄弟姐妹数量  $Siblings$  与受教育程度  $Edu$  一般呈现负相关关系，而且它与能力  $Abl$  应该不相关



# 工具变量法解决方案：测量误差情形

下面我们看一下，IV方法如何处理测量误差导致的内生自变量问题。

$$\log(\text{Wage}_i) = \beta_0 + \beta_1 \text{Edu}_i + \beta_2 \text{Abl}_i + u_i \quad (\text{true model})$$

$$\log(\text{Wage}_i) = \beta_0 + \beta_1 \text{Edu}_i + \beta_2 \text{IQ}_i + u_i^* \quad (\text{error specification model})$$

此时，智商水平  $\text{IQ}_i$  可以考虑作为内生自变量受教育程度  $\text{Edu}$  的工具变量。但是要注意的是，工具变量  $\text{IQ}_i$  还是可能与随机干扰项  $u_i^*$  相关。



# 工具变量法下系数的估计过程

把上述“偏误模型”记为：

$$\begin{aligned} score_i &= \beta_1 + \beta_2 skipped_i + u_i \\ Y_i &= \beta_1 + \beta_2 X_i + u_i \end{aligned}$$

假设我们找到了理想的工具变量  $Z_i$ ，并构建如下的工具变量模型：

$$Y_i = \alpha_1 + \alpha_2 Z_i + v_i$$

$$cov(Z_i, Y_i) = \alpha_2 cov(Z_i, X_i) + cov(Z_i, u_i) \quad \leftarrow [cov(Z_i, u_i) = 0]$$

$$\alpha_2|_{IV}^{plim} = \frac{cov(Z_i, Y_i)}{cov(Z_i, X_i)}$$

$$= \frac{\sum z_i y_i}{\sum z_i x_i} \quad \leftarrow [if \quad X_i = Z_i]$$

$$= \frac{\sum x_i y_i}{\sum x_i^2} = \beta_2$$

这将意味着工具变量法IV会得到最小二乘法OLS下的估计结果。



# 工具变量法下系数的真实方差

对于“偏误模型”和工具变量模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (\text{PRM})$$

$$Y_i = \alpha_1 + \alpha_2 Z_i + v_i \quad (\text{IV})$$

如果如下三个条件成立：

$$\text{Cov}(Z_i, u_i) = 0$$

$$\text{Cov}(Z_i, X_i) \neq 0$$

$$E(v_i^2 | Z_i) \equiv \sigma^2 \equiv \text{var}(u_i)$$

可证明斜率系数  $\alpha_2$  渐近方差为：

$$\text{var}(\alpha_2) \simeq \frac{\sigma^2}{n\sigma_{X_i}^2 \rho_{(X_i, Z_i)}^2}$$

其中：

- $\sigma^2$  是  $v_i$  的总体方差，也即  $\text{var}(v_i) \equiv \sigma^2$ 。
- $\sigma_{X_i}^2$  是  $X_i$  的总体方差，也即  $\text{var}(X_i) \equiv \sigma_{X_i}^2$ 。
- $\rho_{(X_i, Z_i)}^2$  是  $X_i$  和  $Z_i$  的总体相关系数的平方，也即  $\rho_{(X_i, Z_i)}^2 \equiv \frac{[\text{cov}(X_i, Z_i)]^2}{\text{var}(X_i)\text{var}(Z_i)}$ ；



# 工具变量法下系数的样本方差

对于给定的样本数据，我们可以计算出

$$\text{var}(\alpha_2) \simeq \frac{\sigma^2}{n\sigma_{X_i}^2\rho_{(X_i,Z_i)}^2} \simeq \frac{\hat{\sigma}^2}{nS_{X_i}^2R_{(X_i,Z_i)}^2}$$

其中：

- $\sigma_{X_i}^2 \simeq S_{X_i}^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$ 。
- $\rho_{(X_i,Z_i)}^2 \simeq R^2$ ，其中  $R^2$  为通过做  $X_i$  对  $Z_i$  的回归来获得的判定系数。

$$X_i = \hat{\pi}_1 + \hat{\pi}_2 Z_i + \epsilon_i$$

- $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ ，是来自对工具变量回归的残差计算。

$$Y_i = \hat{\alpha}_1 + \hat{\alpha}_2 Z_i + e_i$$



# 已婚女性的教育回报案例

下面给出一个已婚女性的教育回报案例，对上述结论进行论证和分析。

西北农林科技大学  
NORTHWEST A&F UNIVERSITY

西北农林科技大学  
NORTHWEST A&F UNIVERSITY

西北农林科技大学  
NORTHWEST A&F UNIVERSITY



# 工具变量法回归 (IV): 手工分步计算

采用工具变量法的第一阶段回归:

$$educ = + \beta_1 + \beta_2 fatheduc + u_i$$

$$\widehat{educ} = + 10.24 \quad + 0.27 fatheduc$$

$$(t) \quad (37.0993) \quad (9.4255)$$

$$(se) \quad (0.2759) \quad (0.0286)$$

$$(fitness) R^2 = 0.1726; \bar{R}^2 = 0.1706$$

$$F^* = 88.84; p = 0.0000$$

采用工具变量法的第二阶段回归:

$$\widehat{lwage} = + 0.44 \quad + 0.06 educ. hat$$

$$(t) \quad (0.9443) \quad (1.6081)$$

$$(se) \quad (0.4671) \quad (0.0368)$$

$$(fitness) R^2 = 0.0060; \bar{R}^2 = 0.0037$$

$$F^* = 2.59; p = 0.1086$$



# 工具变量法回归 (IV) : R软件自动计算

采用R包AER的工具变量回归函数 `ivreg()`, 可以得到如下回归结果:

```
Call:
ivreg(formula = lwage ~ educ | fatheduc

Residuals:
    Min       1Q   Median       3Q      Max
-3.0870 -0.3393  0.0525  0.4042  2.0677

Coefficients:
              Estimate Std. Error t value
(Intercept)  0.44110    0.44610   0.989
educ         0.05917    0.03514   1.684
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.

Residual standard error: 0.6894 on 426
Multiple R-Squared: 0.09344, Adjuste
Wald test: 2.835 on 1 and 426 DF, p-va
```

工具变量回归模型:

$$\log(wage) = \lambda_1 + \lambda_2 educ | fatheduc + \epsilon_i$$

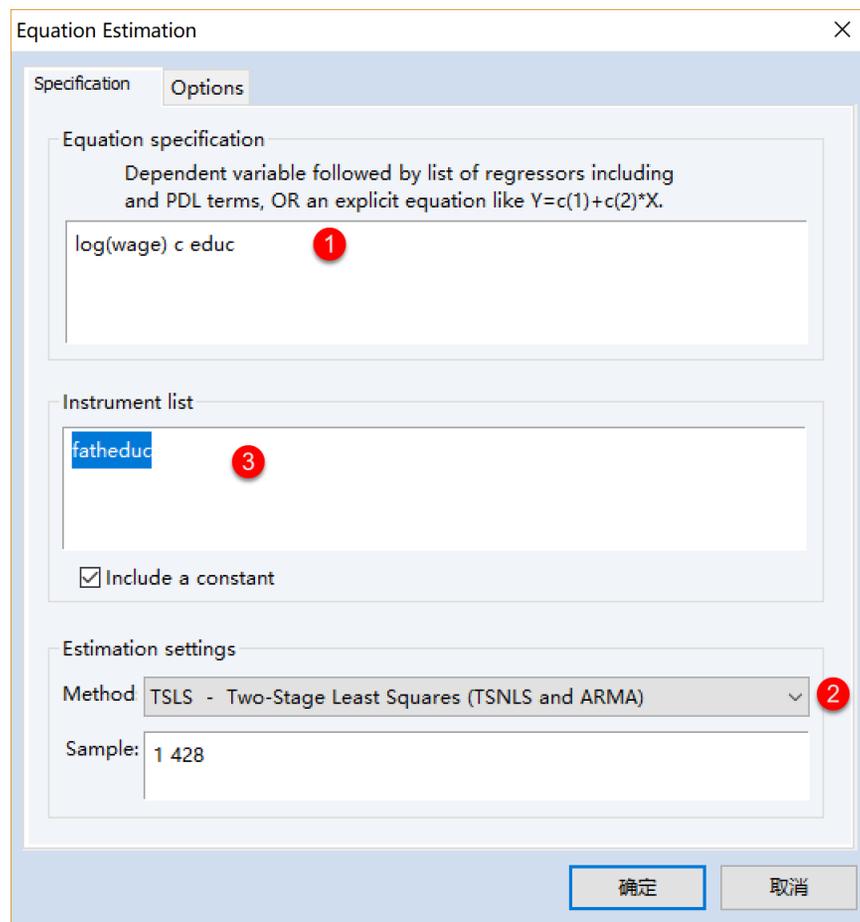
提问:

- 手工分步计算与软件自动计算有哪些不同?
- 判定系数和系数标准误差为什么会不同?



# 工具变量法回归 ( IV ) : EViews软件自动计算

EViews软件下工具变量法的实现:





# 工具变量法回归 ( IV ) : EViews软件自动计算

EViews软件下工具变量法的结果:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.441103	0.446102	0.988795	0.3233
EDUC	0.059173	0.035142	1.683850	0.0929

R-squared	0.093438	Mean dependent var	1.190173
Adjusted R-squared	0.091310	S.D. dependent var	0.723198
S.E. of regression	0.689390	Sum squared resid	202.4601
F-statistic	2.835351	Durbin-Watson stat	1.968194
Prob(F-statistic)	0.092943	Second-Stage SSR	221.9799
J-statistic	1.51E-41	Instrument rank	2

## 17.4 两阶段最小二乘法 (2SLS)



# 两阶段最小二乘法：基本过程

如果我们的工具变量数目多于内生自变量数目，则一致性估计量  $\hat{\beta}_{IV}$  可以通过两步法实现：

- **第1阶段**：对自变量矩阵  $\mathbf{X}$  的每一列都对全部工具变量  $\mathbf{Z}$  进行OLS回归。从而得到矩阵  $\mathbf{X}$  的拟合值矩阵  $\hat{\mathbf{X}}$ 。
- **第2阶段**：将因变量  $\mathbf{y}$  对拟合值矩阵  $\hat{\mathbf{X}}$  进行OLS回归。

以上两个步骤，一起被称为 **两阶段最小二乘法**（two-stage least squares, 2SLS/TSLS）。



## 工资案例：2SLS (无方差矫正)——阶段1 (模型设定)

首先，我们考虑使用母亲受教育情况  $mothereduc$  作为内生自变量  $educ$  的工具变量：

**2SLS的第1阶段：**内生自变量对全部工具变量进行OLS回归。

这一阶段中，我们将能够得到内生自变量的拟合变量  $\widehat{educ}$ ：

$$\widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\gamma}_3 expersq + \hat{\gamma}_4 mothereduc$$



# 工资案例：2SLS (无方差矫正)——阶段1 (回归结果)

以下是 2SLS 的第1阶段估计过程和结果 (R代码) :

```
mod_step1 <- formula(educ~exper + expersq + motheduc) # modle setting
ols_step1 <- lm(formula = mod_step1, data = mroz) # OLS estimation
```

$$\begin{array}{l} \widehat{educ} = + 9.78 \quad + 0.05exper - 0.00expersq + 0.27motheduc \\ (t) \quad (23.0605) \quad (1.1726) \quad (-1.0290) \quad (8.5992) \\ (se) \quad (0.4239) \quad (0.0417) \quad (0.0012) \quad (0.0311) \\ (fitness) R^2 = 0.1527; \bar{R}^2 = 0.1467 \\ F^* = 25.47; p = 0.0000 \end{array}$$

我们可以看到：*mothereduc*系数的样本t值大于2（2t法则），因此t检验显著（ $\alpha = 0.05$ 水平下），意味着工具变量和内生自变量之间存在明显的线性关系，而且是我们已经控制了其他变量的情况下。



# 工资案例：2SLS (无方差矫正)——阶段1 (拟合结果)

在2SLS的第1阶段过程中，我们很快可以获得内生自变量的OLS拟合值  $\widehat{educ}$ ，并把它列添加到数据集中：

```
mroz_add <- mroz %>% mutate(educHat = fitted(ols_step1)) # add fitted educ to data
```

id	lwage	educ	exper	expersq	fatheduc	motheduc	educHat
1	1.21	12	14	196	7	12	13.42
2	0.33	12	5	25	7	7	11.86
3	1.51	12	15	225	7	12	13.43
4	0.09	12	6	36	7	7	11.90
5	1.52	14	7	49	14	12	13.27
6	1.56	12	33	1089	7	14	13.74

Showing 1 to 6 of 428 entries

Previous 1 2 3 4 5 ... 72 Next



## 工资案例：2SLS (无方差矫正)——阶段2 (模型设定)

2SLS的第2阶段：使用母亲受教育情况 *mothereduc* 作为内生自变量 *educ* 的工具变量。

在第2阶段中，我们将因变量  $\log(wage)$  对前面得到的拟合值  $\widehat{educ}$  以及原来模型中的外生自变量继续进行OLS回归。

$$lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon}$$

```
mod_step2 <- formula(lwage~educHat + exper + expersq)
ols_step2 <- lm(formula = mod_step2, data = mroz_add)
```



## 工资案例：2SLS (无方差矫正)——阶段2 (回归结果)

通过利用新的数据集 `moroz_add`，2SLS的第2阶段回归结果如下：

```
fun_report_eq(lm.mod = mod_step2, lm.dt = mroz_add, lm.n = 4)
```

$$\begin{array}{l} \widehat{lwage} = + 0.20 \quad + 0.05educHat + 0.04exper - 0.00expersq \\ (t) \quad (0.4017) \quad (1.2613) \quad (3.1668) \quad (-2.1749) \\ (se) \quad (0.4933) \quad (0.0391) \quad (0.0142) \quad (0.0004) \\ (fitness) R^2 = 0.0456; \bar{R}^2 = 0.0388 \\ F^* = 6.75; p = 0.0002 \end{array}$$

但是请记住，用这种“step by step”的过程计算的标准误差是不正确的(为什么?)。

而正确的方法应该使用专用软件来求解工具变量模型。在R中，这样的函数是 `AER::ivreg()`。



# 广义工具变量回归模型：定义

我们将内生自变量模型表达为：

$$Y_i = \beta_0 + \sum_{j=1}^k \beta_j X_{ji} + \sum_{s=1}^r \beta_{k+s} W_{ri} + \epsilon_i$$

其中， $(X_{1i}, \dots, X_{ki})$  是内生自变量 (endogenous regressors)； $(W_{1i}, \dots, W_{ri})$  是外生自变量 (exogenous regressors)。而且假定我们还找到了  $m$  个工具变量 (instrumental variables)  $(Z_{1i}, \dots, Z_{mi})$ ，它们都满足工具相关性 (instrument relevance) 和工具外生性 (instrument exogeneity) 两大条件。

- 如果  $m = k$ ，则参数估计将是恰好识别的 (exactly identified)。
- 如果  $m > k$ ，则参数估计将是过度识别的 (over-identified)。
- When  $m < k$ ，则参数估计将是无法识别的 (underidentified)。
- 最后，只有  $m \geq k$  时，参数估计才是可识别的 (identified)。



# 广义工具变量回归模型：2SLS估计过程

## 两阶段最小二乘法(2SLS):

- **第1阶段:** 将自变量矩阵中的第1列  $X_{1i}$  都对常数1、所有工具变量  $(Z_{1i}, \dots, Z_{mi})$  以及所有外生自变量  $(W_{1i}, \dots, W_{ri})$  进行OLS估计, 并得到内生自变量的拟合值  $\hat{X}_{1i}$ 。对所有内生自变量都重复此步骤, 最后得到  $(\hat{X}_{1i}, \dots, \hat{X}_{ki})$ 。
- **第2阶段:** 将因变量  $Y_i$  对常数、所有拟合变量  $(\hat{X}_{1i}, \dots, \hat{X}_{ki})$ 、以及所有外生自变量  $(W_{1i}, \dots, W_{ri})$  继续进行OLS估计, 并得到参数估计值  $(\hat{\beta}_0^{IV}, \hat{\beta}_1^{IV}, \dots, \hat{\beta}_{k+r}^{IV})$

下面的几个例子中, 我们将使用**一次性的、整体性的**2SLS估计方案, 直接得到2SLS估计结果。也即:

- 对估计样本标准差进行某种**合理矫正**
- 一次性完成两个OLS估计步骤, 直接得到最后估计结果。
- 我们这里将使用R函数 `ARE::ivreg()` 来执行具体分析。



## 工资案例2SLS：仅使用母亲教育为IV（模型设定）

工资案例中，我们首先仅使用 *mothereduc* 作为内生自变量 *educ* 的工具变量。

$$\begin{cases} \widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\gamma}_3 expersq + \hat{\gamma}_4 mothereduc & \text{(stage 1)} \\ lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon} & \text{(stage 2)} \end{cases}$$



## 工资案例2SLS：仅使用母亲教育为IV（估计结果）

以下是使用R函数 `ARE::ivreg()` 进行2SLS估计的结果：

```
library("AER")
mod_iv_m <- formula(lwage ~ educ + exper + expersq | motheduc + exper + expersq)
lm_iv_m <- ivreg(formula = mod_iv_m, data = mroz)
summary(lm_iv_m)
```

Call:

```
ivreg(formula = mod_iv_m, data = mroz)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.10804	-0.32633	0.06024	0.36772	2.34351

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.1981861	0.4728772	0.419	0.67535
educ	0.0492630	0.0374360	1.316	0.18891
exper	0.0448558	0.0135768	3.304	0.00103 **
expersq	-0.0009221	0.0004064	-2.269	0.02377 *

我们可以看到 *educ* 前的t检验结果是不显著的。



# 工资案例2SLS：仅使用母亲教育为IV（整理结果）

2SLS回归结果整理（仅使用 $motheduc$ 作为工具变量）

term	estimate	std.error	statistic	p.value
(Intercept)	0.1982	0.4729	0.4191	0.6754
educ	0.0493	0.0374	1.3159	0.1889
exper	0.0449	0.0136	3.3039	0.0010
expersq	-0.0009	0.0004	-2.2690	0.0238



## 工资案例2SLS：仅使用父亲教育为IV（模型设定）

这里，我们再考虑仅使用  $fatheduc$  作为内生自变量  $educ$  的工具变量：

$$\begin{cases} \widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\gamma}_3 expersq + \hat{\gamma}_4 fatheduc & \text{(stage 1)} \\ lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon} & \text{(stage 2)} \end{cases}$$

同样，我们使用R软件进行2SLS估计。



## 工资案例2SLS：仅使用父亲教育为IV（估计结果）

通过运行如下的R代码，我们可以得到2SLS的估计结果：

```
mod_iv_f <- formula(lwage ~ educ + exper + expersq | fatheduc + exper + expersq)
lm_iv_f <- ivreg(formula = mod_iv_f, data = mroz)
summary(lm_iv_f)
```

Call:

```
ivreg(formula = mod_iv_f, data = mroz)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.09170	-0.32776	0.05006	0.37365	2.35346

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.0611169	0.4364461	-0.140	0.88870
educ	0.0702263	0.0344427	2.039	0.04208 *
exper	0.0436716	0.0134001	3.259	0.00121 **
expersq	-0.0008822	0.0004009	-2.200	0.02832 *

---

这里我们可以看到 *educ* 前的t检验结果是显著的（给定  $\alpha = 0.05$ ）。



# 工资案例2SLS：仅使用父亲教育为IV（整理结果）

2SLS回归结果整理（仅使用 *fatheduc* 作为工具变量）

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0611	0.4364	-0.1400	0.8887
educ	0.0702	0.0344	2.0389	0.0421
exper	0.0437	0.0134	3.2590	0.0012
expersq	-0.0009	0.0004	-2.2003	0.0283



## 工资案例2SLS：同时使用父亲和母亲教育为IV（模型设定）

当然，我们实际上也可以同时使用 *motheduc* 和 *fatheduc* 作为内生自变量 *educ* 的工具变量。

$$\begin{cases} \widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\beta}_3 expersq + \hat{\beta}_4 motheduc + \hat{\beta}_5 fatheduc & \text{(stage 1)} \\ lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon} & \text{(stage 2)} \end{cases}$$



## 工资案例2SLS：同时使用父亲和母亲教育为IV（估计结果）

类似地，通过运行如下的R代码，我们可以得到2SLS的估计结果：

```
mod_iv_mf <- formula(lwage ~ educ + exper + expersq | motheduc + fatheduc + exper  
lm_iv_mf <- ivreg(formula = mod_iv_mf, data = mroz)  
summary(lm_iv_mf)
```

Call:

```
ivreg(formula = mod_iv_mf, data = mroz)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.0986	-0.3196	0.0551	0.3689	2.3493

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.0481003	0.4003281	0.120	0.90442
educ	0.0613966	0.0314367	1.953	0.05147 .
exper	0.0441704	0.0134325	3.288	0.00109 **
expersq	-0.0008990	0.0004017	-2.238	0.02574 *

---

这里我们可以看到 *educ* 前的t检验结果是显著的（p值小于0.1）。



# 工资案例2SLS：同时使用父亲和母亲教育为IV（整理结果）

2SLS回归结果整理（使用 $motheduc$ 和 $fatheduc$ 作为工具变量）

term	estimate	std.error	statistic	p.value
(Intercept)	0.0481	0.4003	0.1202	0.9044
educ	0.0614	0.0314	1.9530	0.0515
exper	0.0442	0.0134	3.2883	0.0011
expersq	-0.0009	0.0004	-2.2380	0.0257



## 工资案例：多种估计方法下的估计结果对比

我们简单把前面的几类分析汇总一下。目前为止，我们实际上实施了总共**5**种参数估计，它们的估计方法或估计流程各有不同：

- a. 直接对误设模型（存在内生自变量问题）进行OLS估计。
- b. 一步一步的、“手动的”2SLS估计流程（仅使用 *motheduc* 作为工具变量），而且**没有**进行方差协方差矫正。
- c. 一次性的、“专门的”2SLS估计流程，并**进行**了方差协方差矫正。这里使用的是R软件里的专用函数 `ARE::ivreg()` 进行整体“打包式”估计。具体我们估计了3个模型：
  - 仅使用 *motheduc* 作为工具变量
  - 仅使用 *fatheduc* 作为工具变量
  - 同时使用 *motheduc* 和 *fatheduc* 作为工具变量

为了全面做出比较，我们把以上模型的估计结果展示在下一页幻灯片中。



# 工资案例：多种估计方法下的估计结果对比(图片形式)

lwage equation: OLS, 2SLS, and IV models compared

	Dependent variable: lwage				
	OLS (1)	explicit 2SLS (2)	IV mothereduc (3)	IV fathereduc (4)	IV mothereduc and fathereduc (5)
Constant	-0.5200*** (0.2000)	0.2000 (0.4900)	0.2000 (0.4700)	-0.0610 (0.4400)	0.0480 (0.4000)
educ	0.1100*** (0.0140)		0.0490 (0.0370)	0.0700** (0.0340)	0.0610* (0.0310)
educHat		0.0490 (0.0390)			
exper	0.0420*** (0.0130)	0.0450*** (0.0140)	0.0450*** (0.0140)	0.0440*** (0.0130)	0.0440*** (0.0130)
expersq	-0.0008** (0.0004)	-0.0009** (0.0004)	-0.0009** (0.0004)	-0.0009** (0.0004)	-0.0009** (0.0004)
Observations	428	428	428	428	428
R <sup>2</sup>	0.1600	0.0460	0.1200	0.1400	0.1400
Adjusted R <sup>2</sup>	0.1500	0.0390	0.1200	0.1400	0.1300
Residual Std. Error (df = 424)	0.6700	0.7100	0.6800	0.6700	0.6700
F Statistic (df = 3; 424)	26.0000***	6.8000***			





# 工资案例：多种估计方法下的估计结果对比(表格形式)

Dependent variable: lwage					
	OLS	explicit 2SLS	IV mothereduc	IV fathereduc	IV mothereduc and fathereduc
	(1)	(2)	(3)	(4)	(5)
Constant	-0.5220*** (0.1986)	0.1982 (0.4933)	0.1982 (0.4729)	-0.0611 (0.4364)	0.0481 (0.4003)
educ	0.1075*** (0.0141)		0.0493 (0.0374)	0.0702** (0.0344)	0.0614* (0.0314)
educHat		0.0493 (0.0391)			
exper	0.0416*** (0.0132)	0.0449*** (0.0142)	0.0449*** (0.0136)	0.0437*** (0.0134)	0.0442*** (0.0134)
	0.0000**	0.0000**	0.0000**	0.0000**	0.0000**



## 工资案例：多种估计方法下的估计结果对比

表格中的主要信息说明如下：

- 列 (1) ... (5) 分别表示前述5个模型的估计结果，其中 (1) 和 (2) 没有进行方差修正，而 (3)、(4)、(5) 则进行了方差修正。
- 需要注意的是，(3)、(4)、(5) 中的 *educ* 与 (2) 中的 *educHat* 是等价的。
- 括号里显示的是参数估计了的样本标准误差 (standard error of the estimator)。



# 工资案例：多种估计方法下的估计结果对比

5个模型估计结果比较的主要要点有：

- 首先，由表可知，教育在决定工资方面的重要性在模型(3)、(4)和(5)中相对要更小，系数分别为0.049,0.07,0.061。标准误差也随着估计模型(3)、(4)、(5)而减小。
- 其次，它还表明，明确的2SLS模型(2)和仅使用 *motheduc* 为工具变量的模型(3)产生相同的系数估计值，但**标准误差**不同。模型(2)中的2SLS的标准误差为0.039，比模型(3)估计值的标准误差0.037略大。
- 第三，当仅使用*motheduc*作为唯一工具变量时，模型(2)和模型(3)的教育系数的t检验不显著。
- 第四，我们这里可以充分感受和理解一下2SLS的“相对估计效率”！

## 17.5 检验工具变量的 有效性 (Instrument validity)



# 工具变量有效性：定义和内涵

考虑如下一般化的模型：

$$Y_i = \beta_0 + \sum_{j=1}^k \beta_j X_{ji} + \sum_{s=1}^r \beta_{k+s} W_{ri} + u_i$$

- $Y_i$  是因变量
- $\beta_0, \dots, \beta_{k+1}$  是  $1 + k + r$  个待估计回归系数
- $X_{1i}, \dots, X_{ki}$  是  $k$  个内生自变量
- $W_{1i}, \dots, W_{ri}$  是  $r$  个模型中外生自变量，它们都与  $u_i$  不相关
- $u_i$  是随机干扰项
- $Z_{1i}, \dots, Z_{mi}$  是  $m$  个工具变量。



# 工具变量有效性：定义和内涵

工具变量有效性(Instrument valid)意味着工具变量必须同时满足工具相关性(Instrument Relevance)和工具外生性(Instrument Exogeneity)两个条件：

$$E(Z_i X_i') \neq 0$$

$$E(Z_i u_i) = 0$$



# 检验工具相关性: 放松条件

实际研究中, 工具相关性也意味着, 如果存在  $k$  个内生自变量和  $m$  个工具变量  $Z$ , 只要  $m \geq k$ , 则一定可以得到如下的外生变量向量:

$$\left( \hat{X}_{1i}^*, \dots, \hat{X}_{ki}^*, W_{1i}, \dots, W_{ri}, 1 \right)$$

而且, 它也不应该是完全共线性(perfectly multicollinear)。

其中:

- $\hat{X}_{1i}^*, \dots, \hat{X}_{ki}^*$  是2SLS中第1阶段得到的  $k$  个内生自变量的OLS估计拟合值。
- 1 代表常数回归元, 对于有截距回归模型, 所有样本的常数回归元取值都等于1。

显然, 完全多重共线是比较少见的, 我们完全可以不用大费周章来仔细检验这种情形。

事实上, 我们真正需要注意的是被称为“弱工具性”(weak instruments)的问题。



## 检验工具相关性: 弱工具变量(问题)

**弱工具变量:** 如果我们找到的工具变量只能解释内生自变量变异的很少部分, 那么我们就称这样的工具变量为弱工具变量(**weak instruments**)。

正式地, 当  $\text{corr}(Z_i, X_i)$  接近于0时,  $z_i$  被称作为弱工具变量。

- 考虑简单回归的情形  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- 参数  $\beta_1$  的IV估计值为  $\hat{\beta}_1^{IV} = \frac{\sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})}$

Note that  $\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y}) \xrightarrow{p} \text{Cov}(Z_i, Y_i)$

and  $\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X}) \xrightarrow{p} \text{Cov}(Z_i, X_i)$ .

- 因此, 如果  $\text{Cov}(Z_i, X_i) \approx 0$ , 那么IV估计值  $\hat{\beta}_1^{IV}$  也将是无意义的。



## 检验工具相关性: 弱工具变量(案例)

下面的案例中, 我们考察吸烟 (smoking) 对出生婴儿体重 (birth weight) 的影响, 构建的模型如下:

$$\log(\text{bwght}) = \beta_0 + \beta_1 \text{packs} + \epsilon_i$$

其中  $\text{packs}$  是妈妈每天抽烟的盒数, 我们有理由认为这个变量是内生自变量 (为什么?). 另外, 假定我们使用香烟平均价格  $\text{cigprice}$  作为内生自变量  $\text{packs}$  的一个工具变量, 并假设它与随机干扰项  $\epsilon$  不相关。



## 检验工具相关性: 弱工具变量(案例)

然而, 妈妈抽烟盒数  $packs$  对香烟平均价格  $cigprice$  第1阶段OLS回归分析, 我们发现基本上二者并没有相关关系。

$$\widehat{packs} = 0.067 + 0.0003 \text{ cigprice}$$

(0.103) (0.0008)

在这种情况下, 如果我们执意使用  $cigprice$  作为工具变量, 并进行第2阶段的OLS回归, 我们会得到:

$$\log(\widehat{bwght}) = 4.45 + 2.99 \text{ packs}$$

(0.91) (8.70)

显然, 即便第2阶段的结果t检验是显著的, 但它已经完全没有的检验的意义和价值。因为  $cigprice$  表现为弱工具变量, 在第1阶段的回归就已经暴露出问题了。



## 检验工具相关性: 弱工具变量(策略)

如果手里拿到的是弱工具变量, 那么我们有两个实施策略:

- 忍爱放弃 **弱工具变量**, 再次开始寻找**强工具变量**:

虽然前者只是一个选项, 如果待估计参数仍然可识别时, 即便弱工具变量被舍弃, 参数估计还是可能的。但是后者可能就是极其困难的, 甚至可能需要我们重新设计整个研究。

- 坚持使用**弱工具变量**, 但要使用改进的2SLS方法。

这样的改进方法包括, 诸如**有限信息极大似然估计法**(limited information maximum likelihood estimation, **LIML**)。





## 弱工具变量检验( $F$ test): 1个内生自变量的情形

下面先简单考虑只有一个内生自变量的情况。如果在2SLS估计的第1阶段回归中所有工具变量的系数联合F检验不显著（接受  $H_0 : \alpha_1 = \alpha_2 = \dots = 0$ ）则该工具变量显然不具备工具相关性的要求。

我们可以使用以下经验法则：

- 进行2SLS估计的第1阶段回归

$$X_i = \hat{\alpha}_0 + \hat{\alpha}_1 Z_{1i} + \dots + \hat{\alpha}_m Z_{mi} + \hat{u}_i \quad (3)$$

- 通过计算F统计量，对如下联合假设进行检验： $H_0 : \hat{\alpha}_1 = \dots = \hat{\alpha}_m = 0$ 。
- 如果计算得到的样本统计量  $F^*$  比理论查表值小，则不能拒绝  $H_0$ ，表明这些工具变量都是弱工具变量。

这一经验法则在R中很容易实现。使用 `lm()` 函数运行第1阶段回归，然后通过 `car::linearHypothesis()` 函数计算得到统计量  $F^*$ 。



## 弱工具变量检验(Cragg-Donald test): 多个内生自变量的情形

然而，如果模型中存在多个内生自变量，前述的F检验就变得不可靠了。——即便我们确实也可以对每个内生自变量分别进行F检验。

此时，一个可行的检验方法是**Cragg-Donald test**，这一检验将依赖于计算如下的统计量：

$$F = \frac{N - G - B}{L} \frac{r_B^2}{1 - r_B^2}$$

- 其中： $G$  是外生自变量的个数； $B$  是内生自变量的个数； $L$  是工具变量的个数； $r_B$  是最小的canonical相关系数（lowest canonical correlation）。

**canonical相关系数**是对内生变量和外生变量之间相关性的度量，可以通过R函数 `cancor()` 来计算得到。后面我们会给出一个示例！



## 工资案例: 弱工具变量检验 ( $\mathcal{F}$ test): 模型设定

对于前面工资案例中的3个工具变量, 我们可以依次检验它们的工具相关性:

$$educ = \gamma_1 + \gamma_2 exper + \gamma_2 expersq + \theta_1 motheduc + v \quad (\text{relevance test 1})$$

$$educ = \gamma_1 + \gamma_2 exper + \gamma_2 expersq + \theta_2 fatheduc + v \quad (\text{relevance test 2})$$

$$educ = \gamma_1 + \gamma_2 exper + \gamma_2 expersq + \theta_1 motheduc + \theta_2 fatheduc + v \quad (\text{relevance test 3})$$



# 工资案例: 弱工具变量检验 ( $F$ test): 检验结果1

$$educ = \gamma_1 + \gamma_2 exper + \gamma_2 expersq + \theta_1 motheduc + v$$

```
library("car")  
linearHypothesis(ols_relevance1, c("motheduc=0"))
```

Linear hypothesis test

Hypothesis:  
motheduc = 0

Model 1: restricted model  
Model 2: educ ~ exper + expersq + motheduc

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	425	2219.2				
2	424	1889.7	1	329.56	73.946	< 2.2e-16 ***

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

- 受约束F检验(Restricted F test)的原假设为:  $H_0: \theta_1 = 0$ 。
- 以上结果表明样本统计量  $F^*$ 对应的概率p值小于0.01, 应显著拒绝  $H_0$ , 认为工具变量 *motheduc*满足工具相关性条件。



# 工资案例: 弱工具变量检验 ( $F$ test): 检验结果 I

需要注意的是: 受约束F检验(Restricted F test)是不同于经典F检验(classical F test)的。  
我们可以简单比较一下。

这是经典F检验(classical F test)结果:

$$\begin{aligned} \widehat{educ} &= + 9.78 && + 0.05exper - 0.00expersq + 0.27motheduc \\ (t) & (23.0605) && (1.1726) \quad (-1.0290) \quad (8.5992) \\ (se) & (0.4239) && (0.0417) \quad (0.0012) \quad (0.0311) \\ (fitness) R^2 &= 0.1527; \bar{R}^2 = 0.1467 \\ F^* &= 25.47; p = 0.0000 \end{aligned}$$





## 工资案例: 弱工具变量检验 ( $F$ test): 检验结果2

$$educ = \gamma_1 + \gamma_2 exper + \gamma_2 expersq + \theta_1 fatheduc + v \quad (\text{relevance test 2})$$

```
linearHypothesis(ols_relevance2, c("fatheduc=0"))
```

Linear hypothesis test

Hypothesis:

fatheduc = 0

Model 1: restricted model

Model 2:  $educ \sim exper + expersq + fatheduc$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	425	2219.2				
2	424	1838.7	1	380.5	87.741	< 2.2e-16 ***

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

- 受约束F检验(Restricted F test)的原假设为:  $H_0: \theta_1 = 0$ 。
- 以上结果表明样本统计量  $F^*$ 对应的概率p值小于0.01, 应显著拒绝  $H_0$ , 认为工具变量  $fatheduc$ 满足工具相关性条件。



## 工资案例: 弱工具变量检验 ( $F$ test): 检验结果3

$$educ = \gamma_1 + \gamma_2 exper + \gamma_2 expersq + \theta_1 motheduc + \theta_2 fatheduc + v \quad (\text{relevance test 3})$$

```
linearHypothesis(ols_relevance3, c("motheduc=0", "fatheduc=0"))
```

Linear hypothesis test

Hypothesis:

motheduc = 0

fatheduc = 0

Model 1: restricted model

Model 2: educ ~ exper + expersq + motheduc + fatheduc

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	425	2219.2				
2	423	1758.6	2	460.64	55.4	< 2.2e-16 ***

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

以上结果表明样本统计量  $F^*$  对应的概率  $p$  值小于 0.01, 应显著拒绝  $H_0$ , 认为工具变量  $motheduc$  和  $fatheduc$  之中起码有 1 个是满足工具相关性条件的。



## 工作时长案例: 弱工具变量检验 (Cragg-Donald test)

下面, 我们将构造含有2个内生自变量的模型, 并尝试使用Cragg-Donald test方法来检验我们的工具变量是否是弱工具变量。

假定如下误设工作时长回归模型 (包含2个内生自变量) :

$$hushrs = \beta_1 + \beta_2 mtr + \beta_3 educ + \beta_4 kidslt6 + \beta_5 nwifeinc + e$$

假定我们认为模型中有:

- 2个内生自变量: *educ* and *mtr*
- 2个外生自变量: *nwifeinc* and *kidslt6*
- 2个工具变量: *motheduc* and *fatheduc*.

- *hushrs* = 家庭中丈夫工作时长(1975年)
- *mtr* = 联邦政府对已婚女性征收的婚姻税
- *kidslt6* = 家庭中是否有年龄小于6岁的孩子 (虚拟变量)
- *nwifeinc* = 扣除妻子收入的家庭净收入



## 工作时长案例: 弱工具变量检验(Cragg-Donald test)

我们仍旧使用前面的数据集 `mroz`，且只用女性参加工作的样本( $inlf = 1$ )。

```
mroz1 <- wooldridge::mroz %>%
  filter(wage > 0, inlf == 1)
G<-2; L<-2; N<-nrow(mroz1)
x1 <- resid(lm(mtr ~ kidslt6 + nwifeinc, data = mroz1))
x2 <- resid(lm(educ ~ kidslt6 + nwifeinc, data = mroz1))
z1 <-resid(lm(motheduc ~ kidslt6 + nwifeinc, data = mroz1))
z2 <-resid(lm(fatheduc ~ kidslt6 + nwifeinc, data = mroz1))
X <- cbind(x1,x2)
Y <- cbind(z1,z2)
rB <- min(cancor(X, Y)$cor)
CraggDonaldF <- ((N-G-L)/L)/((1-rB^2)/rB^2)
```

运行上述R代码，结果显示 Cragg-Donald统计量  $F^* = 0.1008$ ，它远小于理论查表值4.58<sup>[1]</sup>。因此，我们无法拒绝  $H_0$ ，认为工具变量 `motheduc`和 `fatheduc`两个都是弱工具变量。

[1]理论查表值可以参阅《计量经济学原理》Hill, Griffiths and Lim(2011)的表10E.1



# 检验工具外生性: 主要的困难

工具变量外生性(Instrument Exogeneity) 意味着所有  $m$  个工具变量必须与随机干扰项不相关:

$$\text{Cov}(Z_{1i}, \epsilon_i) = 0; \quad \dots; \quad \text{Cov}(Z_{mi}, \epsilon_i) = 0.$$

- 在只有少数工具变量情形下, 我们会发现工具变量外生性的要求几乎无法被检验。(为什么?)
- 然而, 如果我们有比我们需要的更多工具变量, 那么我们可以有效地测试是否其中一些工具变量与随机干扰项无关。

因此, 下面我们将主要讨论过度识别 (over-identification) 的内生自变量问题模型。



# 检验工具外生性: 过度识别情形

我们已经知道, 过度识别 (over-identification) 情况下 ( $m > k$ ), 我们可以通过尝试组合不同的工具变量来进行IV法参数估计。显然, 理论上我们认为:

如果工具变量都是外生的, 组合不同工具变量, 那么得到的估计值应该是近似的。  
如果估计值非常不同, 则一些或所有工具变量可能不是外生的。

我们下面介绍的过度识别的受约束检验(overidentifying restrictions test) —— **J test**, 正是基于这一检验思想:

- **J test**原假设为工具变量是外生性的:

$$H_0 : E(Z_{hi}\epsilon_i) = 0, \text{ for all } h = 1, 2, \dots, m$$





# 检验工具外生性: Jtest检验流程

过度识别约束检验 (overidentifying restrictions test), 又被称为  $J$ -test 检验, 或者 Sargan test 检验。这种检验的原假设为工具变量都是外生性的。

过度识别约束检验的主要流程是:

- Step 1: 计算IV回归残差(IV regression residuals):

$$\hat{\epsilon}_i^{IV} = Y_i - \left( \hat{\beta}_0^{IV} + \sum_{j=1}^k \hat{\beta}_j^{IV} X_{ji} + \sum_{s=1}^r \hat{\beta}_{k+s}^{IV} W_{si} \right)$$

- Step 2: 运行辅助回归, 也即将IV回归残差对工具变量和外生自变量进行OLS回归估计。然后对该辅助回归进行如下的联合假设检验

$$H_0 : \alpha_1 = 0, \dots, \alpha_m = 0$$

$$\hat{\epsilon}_i^{IV} = \alpha_0 + \sum_{h=1}^m \alpha_h Z_{hi} + \sum_{s=1}^r \alpha_{m+s} W_{si} + v_i \quad (2)$$



# 检验工具外生性: Jtest检验流程

- **Step3:** 根据上述受约束联合F检验计算得到如下J统计量:  $J = mF^*$

其中  $F^*$  是前述  $m$  个受约束回归检验中的F统计量值。其约束条件为

$$H_0 : \alpha_1 = \dots = \alpha_m = 0 \text{ in eq(2)}$$

在原假设  $H_0$  下, 上面计算得到的J统计量统计量在大样本情况下服从卡方分布  $\chi^2(m - k)$ 。

$$J \sim \chi^2(m - k)$$

- 如果  $J$  小于卡方分布理论查表值, 则J检验不显著, 不能拒绝  $H_0$ , 意味着所有工具变量都是外生性的。
- 如果  $J$  大于卡方分布理论查表值, 则J检验显著, 拒绝  $H_0$ , 接受  $H_1$ , 意味着所有至少有1个工具变量不是外生性的。

下面的示例中, 我们将使用R软件的函数 `linearHypothesis()` 进行 J-test 检验。



## 工资案例: Jtest (主模型和辅助模型)

继续讨论工资案例，这里我们考虑同时使用 *motheduc* 和 *fatheduc* 作为内生自变量 *educ* 的工具变量。

初步可以判断，下述的IV模型是过度识别的，因此我们采用 **J-test** 来检验两个工具变量是不是全都是外生性的。

2SLS模型中，我们设定为：

$$\begin{cases} \widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\beta}_3 expersq + \hat{\beta}_4 motheduc + \hat{\beta}_5 fatheduc & \text{(stage 1)} \\ lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon} & \text{(stage 2)} \end{cases}$$

那么辅助回归，我们相应地设定为：

$$\hat{\epsilon}^{IV} = \hat{\alpha}_1 + \hat{\alpha}_2 exper + \hat{\alpha}_3 expersq + \hat{\alpha}_4 motheduc + \hat{\alpha}_5 fatheduc + v \quad \text{(auxiliary model)}$$



# 工资案例: Jtest (得到工具变量估计的残差)

```
mroz_resid <- mroz %>%  
  mutate(resid_iv_mf = residuals(lm_iv_mf)) # obtain residual of IV regression
```

id	lwage	educ	exper	expersq	fatheduc	motheduc	resid_iv_mf
1	1.21	12	14	196	7	12	-0.0169
2	0.33	12	5	25	7	7	-0.6547
3	1.51	12	15	225	7	12	0.2690
4	0.09	12	6	36	7	7	-0.9254
5	1.52	14	7	49	14	12	0.3515
6	1.56	12	33	1089	7	14	0.2930

Showing 1 to 6 of 428 entries

Previous 1 2 3 4 5 ... 72 Next

这里展示了在进行执行2SLS的第1阶段回归后，我们将**IV回归残差**添加到原来的数据集中，从而得到新的数据集。



## 工资案例: Jtest (运行辅助回归)

下一步, 我们运行前面设定的**辅助回归**, 并得到如下结果 (事实上这里不能得到外生性的任何结论):

```
mod_jttest <- formula(resid_iv_mf ~ exper +expersq +motheduc +fatheduc)
lm_jttest <- lm(formula = mod_jttest, data = mroz_resid)
summary(lm_jttest)
```

Call:

```
lm(formula = mod_jttest, data = mroz_resid)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.1012	-0.3124	0.0478	0.3602	2.3441

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.096e-02	1.413e-01	0.078	0.938
exper	-1.833e-05	1.333e-02	-0.001	0.999
expersq	7.341e-07	3.985e-04	0.002	0.999
motheduc	-6.607e-03	1.189e-02	-0.556	0.579
fatheduc	5.782e-03	1.118e-02	0.517	0.605



## 工资案例: Jtest (受约束 $F$ 检验结果)

实际上, 关键的步骤是我们对**辅助回归**进行如下的**受约束联合 $F$ 检验**, 并得到 $F$ 统计量值  $F^* = 0.19$

```
restricted_ftest <- linearHypothesis(lm_jtest, c("motheduc = 0", "fatheduc = 0"),  
restricted_ftest
```

Linear hypothesis test

Hypothesis:

motheduc = 0

fatheduc = 0

Model 1: restricted model

Model 2: resid\_iv\_mf ~ exper + expersq + motheduc + fatheduc

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	425	193.02				
2	423	192.85	2	0.1705	0.187	0.8295

请注意代码块中的 `c("motheduc = 0", "fatheduc = 0")`; 同时要注意**受约束 $F$ 检验**不同于**经典 $F$ 检验**



# 工资案例: Jtest (计算J统计量)

根据受约束联合F检验可计算出卡方统计量，并得到检验结论。

```
(jtest <- linearHypothesis(lm_jtest, c("motheduc = 0", "fatheduc = 0"), test = "Ch
```

Linear hypothesis test

Hypothesis:

motheduc = 0

fatheduc = 0

Model 1: restricted model

Model 2: resid\_iv\_mf ~ exper + expersq + motheduc + fatheduc

	Res.Df	RSS	Df	Sum of Sq	Chisq	Pr(>Chisq)
1	425	193.02				
2	423	192.85	2	0.1705	0.374	0.8294

最后得到的卡方统计量值为  $\chi^{2*} = 0.37$ 。需要注意的是，R软件中 `linearHypothesis()` 报告的概率  $p$  值是不正确的，因为卡方统计量的自由度错误地设定成了2，而根据我们的理论公式，实际自由度应该是  $(m - k) = 1$ 。所以，还需要对自由度进行调整。



## 工资案例: Jtest (调整自由度)

因为R软件中 `linearHypothesis()` 默认卡方自由度是  $m$ 。现在，我们需要设定正确的卡方检验自由度为  $m - k$ :

```
# compute correct p-value for J-statistic  
pchisq(jtest[2, 5], df = 1, lower.tail = FALSE)  
pchisq
```

```
[1] 0.5408401
```

R软件中，我们可以直接使用 `pchisq()` 函数，计算卡方统计量  $\chi^{2*} = 0.37$  对应的概率  $p$  值，并做出假设检验的判断。（当然，我们也可以通过卡方分布的理论查表值做出假设检验判断）

因为计算得到的卡方概率值  $p = 0.5408$ ，比0.1还要大。因此，我们不能拒绝原假设，从而认为所有工具变量 `motheduc` 和 `fatheduc` 都是外生性的!

## 17.6 检验自变量的 内生性 (regressor endogeneity)



# 检验自变量的内生性：内涵和思路

由于OLS通常比IV方法更有效(回想一下，如果高斯-马尔科夫假设成立，则OLS估计为BLUE)。

这也就以为着，如果我们并不想得到一致性估计量时，我们实际上并不需要使用IV方法。

当然，如果我们确实想要得到一致性估计量，面对内生自变量问题模型，我们还需要检验内生自变量是不是真的是内生性的，也即：

$$H_0 : \text{Cov}(X, \epsilon) = 0 \text{ vs. } H_1 : \text{Cov}(X, \epsilon) \neq 0$$



# 检验自变量的内生性：内涵和思路

**Hausman test**将会告诉我们：如果**不能拒绝**原假设  $H_0$ ，我们直接使用OLS方法估计就很有效；如果**显著拒绝**原假设  $H_0$ ，那么使用IV法才能得到参数的一致性估计量。

下面给出的是**Hausman test**检验的基本思想和逻辑：

- 如果自变量  $X$  确实是**外生性的**，那么我们采用OLS方法和采用IV方法，两者的参数估计结果应该是**一样的**。
- 如果自变量  $X$  确实存在**内生性**，那么我们采用OLS方法和采用IV方法，两者的参数估计结果应该是**不一样的**。



## 检验自变量的内生性: Hausman检验

**Hausman test**检验的关键，就是比较OLS方法和IV方法下参数估计值之间的**差异性**。

- 如果两种估计方法的差异是**微小**，我们可以推测OLS和IV是**一致的**，也即模型中自变量都是**外生的**。我们可以直接使用OLS方法。
- 如果两种估计方法的差异**很大**，意味着OLS和IV估计量是**不一致的**。在这种情况下，模型可能存在内生自变量问题，那么我们应该使用IV法。



## 检验自变量的内生性: Hausman检验

下面给出的是Hausman test的具体检验形式:

$$\hat{H} = n \left[ \hat{\beta}_{IV} - \hat{\beta}_{OLS} \right]' \left[ \text{Var} \left( \hat{\beta}_{IV} - \hat{\beta}_{OLS} \right) \right]^{-1} \left[ \hat{\beta}_{IV} - \hat{\beta}_{OLS} \right] \xrightarrow{d} \chi^2(k)$$

- 如果样本统计量  $\hat{H}$  比卡方理论查表值 **小**, 则Hausman test **不显著**, 不能显著拒绝  $H_0$ , 从而认为所有自变量应该不是内生性的。
- 如果样本统计量  $\hat{H}$  比卡方理论查表值 **大**, 则Hausman test是**显著的**, 显著拒绝  $H_0$ , 接受  $H_1$ , 从而认为至少有部分自变量是内生性的。



## 工资案例: Hausman test (工具变量模型设定)

再次使用工资案例进行说明。我们继续同时使用 *motheduc* 和 *fatheduc* 作为内生自变量 *educ* 的工具变量。并做出如下的2SLS模型设定：

$$\begin{cases} \widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\beta}_3 expersq + \hat{\beta}_4 motheduc + \hat{\beta}_5 fatheduc & \text{(stage 1)} \\ lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon} & \text{(stage 2)} \end{cases}$$

在R软件中，我们可以使用IV模型诊断工具来进行**Hausman test**。其中只需要设定函数 `summary(lm_iv_mf, diagnostics = TRUE)` 中的参数 `diagnostics = TRUE`，我们就能得到**Hausman test**结论。



## 工具变量: Hausman test (模型诊断)

```
summary(lm_iv_mf, diagnostics = TRUE)
```

Call:

```
ivreg(formula = mod_iv_mf, data = mroz)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.0986	-0.3196	0.0551	0.3689	2.3493

Coefficients:

Estimate	Std. Error	t value	Pr(> t )
----------	------------	---------	----------

- **(Wu-)Hausman test**用于内生性检验, 拒绝原假设, 认为自变量 *Educ*是内生性的。
- **Weak instruments test**用于弱工具变量检验, 拒绝原假设, 认为至少有1个工具变量不是弱工具变量。
- **Sargan overidentifying restrictions**用于检验外生性。结果发现不能拒绝原假设, 意味着工具变量随机干扰项是不相关的 (外生性的)。

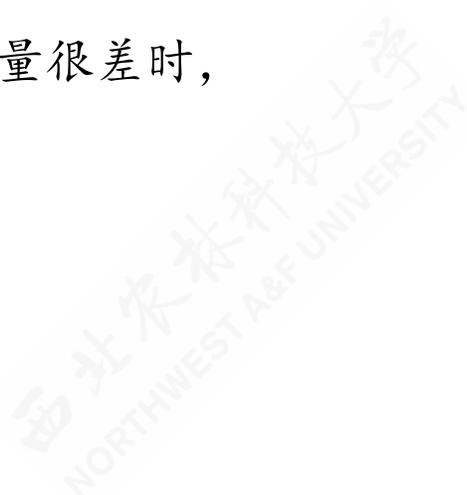


## 小结

- 一个**工具变量**必须有两个属性：
  1. 必须与随机干扰项不相关（**工具外生性**）；
  2. 必须与内生解释变量部分相关（**工具相关性**）。

找到具有这两个属性的变量通常很有挑战性。

- 虽然我们永远不能测试**所有的**工具变量是否是外生的，但我们至少可以测试它们中的一些否是外生的。
- 当工具变量有效时，我们可以进一步检验解释变量是否为**内生性的**。
- **两阶段最小二乘（2SLS）**方法在社会科学中经常使用。但是当工具变量很差时，2SLS可能比OLS方法更糟糕。





## 练习案例: Card (1995)

In Card (1995) education is assumed to be endogenous due to omitted **ability** or **measurement error**. The standard wage function

$$\ln(wage_i) = \beta_0 + \beta_1 Educ_i + \sum_{m=1}^M \gamma_m W_{mi} + \varepsilon_i$$

is estimated by **Two Stage Least Squares** using a **binary instrument**, which takes value 1 if there is an **accredited 4-year public college in the neighborhood** (in the "local labour market"), 0 otherwise.





## 练习案例: Card (1995)

The dataset is available online at [http://davidcard.berkeley.edu/data\\_sets.html](http://davidcard.berkeley.edu/data_sets.html) and consists of 3010 observations from the National Longitudinal Survey of Young Men.

- **Education** is measured by the years of completed schooling and varies between 2 and 18 years.



## 练习案例2: Angrist and Krueger (1991)

The data is available online at

<http://economics.mit.edu/faculty/angrist/data1/data/angkru1991> and consists of observations from 1980 Census documented in Census of Population and Housing, 1980: Public Use Microdata Samples.

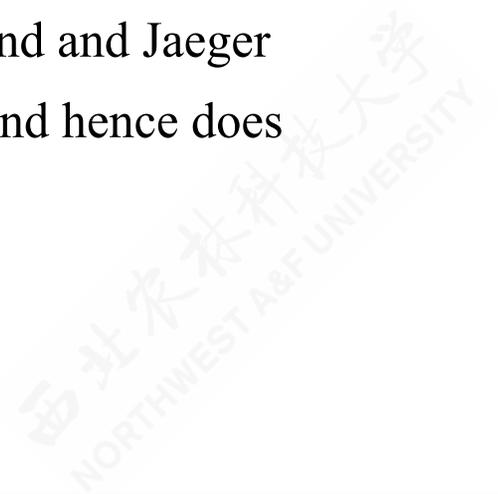


## 练习案例2: Angrist and Krueger (1991)

- They observe that individuals born earlier in the year (first two quarters) have less schooling than those born later in the year.

It is a consequence of **the compulsory schooling laws**, as individuals born in the first quarters of the year reach *the minimum school leaving age* at the lower grade and might legally leave school with less education.

- The main criticism of Angrist and Krueger (1991) analysis, pointed out by Bound, Jaeger and Baker (1995) is that the quarter of birth is a **weak instrument**.
- A second criticism of Angrist and Krueger (1991) results, discussed by Bound and Jaeger (1996) is that quarter of birth might be **correlated** with unobserved ability and hence does **not** satisfy the **instrumental exogeneity condition**.



# 本章結束

