

专题： 内生性与工具变量

什么是内生性问题？

- 定义：

内生性（Endogeneity）是指计量模型中的一个或多个解释变量（X）与随机扰动项（ μ ）**相关**的现象。

要点：违背经典线性回归模型假设2，即不满足“X独立于随机扰动项 μ ”的假定。

- 影响：

- 对模型BLUE性质的影响：OLS估计量**有偏**，且**不一致**
 - 最佳性（Best）：失去意义
 - 线性（Linearity）：不受影响
 - 无偏性（Unbiased）：消失（甚至，失去“一致性”）
- 假设检验失效：t检验、F检验结果不可靠。
- 回归结果的解释：错误的因果归因

内生性产生的三大原因

核心问题：为什么解释变量 (X) 与随机扰动项 (μ) 相关？

- **A. 遗漏变量偏差 (Omitted Variable Bias)**

遗漏的解释变量与模型其他解释变量相关，遗漏的变量影响被归入误差项，因此导致其他解释变量与随机扰动项相关。

例子：研究“教育对工资的影响”。如果遗漏了“个人能力”，而能力既影响学历又影响工资，那么学历的系数就会把能力的效果也算进去，从而被高估。

- **B. 双向因果 (Simultaneity / Reverse Causality)**

例子：“警察数量”与“犯罪率”的相互影响

- **C. 测量误差 (Measurement Error)**

当解释变量存在系统性误差时，误差项会与解释变量相关

如何解决内生性？

方法	核心思路
工具变量法 (IV)	寻找一个变量 Z （工具变量），它只通过影响 X 来间接影响 Y ，而与 Y 的误差项无关。常用 2SLS（两阶段最小二乘法） 估计。
固定效应模型 (FE)	针对面板数据，通过控制个体固定效应（如个人的性格、地区的文化），消除那些 不随时间变化 的遗漏变量。
双重差分法 (DID)	利用政策冲击（如法律颁布），构建“处理组”和“对照组”的对比，通过消除共性趋势来提取净因果效应。
断点回归 (RDD)	利用某个连续变量的临界点（如考试及格线）形成的自然实验进行推断。
滞后变量	在时间序列中，用 X 的过去值作为解释变量，以减少双向因果的影响。

工具变量法

- 什么是工具变量法？

思路：当自变量 X 变得“不干净”（与误差项相关）时，我们找一个“干净”的变量 Z （工具变量），利用 Z 去提取 X 中干净的部分，再用这部分纯净的信息去解释 Y 。

- 工具变量必须满足的两个条件：

- 相关性：工具变量 Z 必须与内生变量 X 显著相关，即 $Cov(Z_i, X_i) \neq 0$
- 外生性：工具变量 Z 必须与随机扰动项 μ 无关，即 $Cov(Z_i, u_i) = 0$

- 工具变量法的实现：两阶段最小二乘法（2SLS）

- 第一阶段：内生变量 X 作为被解释变量，工具变量 Z 作为解释变量，进行辅助回归，得到预测值 $\hat{X}$ ，辅助回归模型如下：

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

- 第二阶段：用预测值 $\hat{X}$ 替代原始 X ，进行主模型回归：

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \epsilon_i$$

上述主模型回归中，估计量 $\hat{\beta}_1$ 就是工具变量法中一致估计量，记为 $\hat{\beta}_{IV}$

工具变量法的相关检验：两个前置检验

- **弱工具变量检验 (Weak Instrument Test)**

- 目的：确保工具变量Z与内生变量X之间显著性相关，即 $Cov(Z_i, X_i) \neq 0$
- 方法：第一阶段回归的F统计量
- 判定标准：F统计量越大，工具变量Z与内生变量X越显著相关

- **过度识别检验 (Hansen J test)**

- 适用场景：工具变量个数L 大于 内生变量个数K
- 目的：确保所有工具变量均与随机扰动项 μ 无关，即 $Cov(Z_i, u_i) = 0$
- 假设
 - 原假设：所有工具变量都是外生的（与随机扰动项 μ 无关）
 - 备择假设：至少有一个工具变量是内生的（与随机扰动项 μ 相关）

- 统计量表达式：

$$J = n \cdot \bar{g}(\hat{\beta}_{GMM})' \hat{W} \bar{g}(\hat{\beta}_{GMM})$$

- 分布：J统计量服从自由度为L-K的卡方分布
- 判定标准：统计量对应的P值

工具变量法的相关检验：Hausman检验

Hausman检验

- 目的：判断解释变量X“内生性”和工具变量法的有效性

- 原理：

通过对比OLS估计量和IV估计量的差异，判断：解释变量X的内生性；工具变量法的有效性。

- 假设：

- H_0 : 解释变量外生（即OLS估计量与IV估计量差异不显著，OLS有效）
- H_1 : 解释变量内生（即OLS估计量与IV估计量差异显著，IV有效）

- 统计量表达式：

$$H = (\hat{\beta}_{OLS} - \hat{\beta}_{IV})' [Var(\hat{\beta}_{OLS}) - Var(\hat{\beta}_{IV})]^{-1} (\hat{\beta}_{OLS} - \hat{\beta}_{IV})$$

- 分布：Hausman检验统计量服从卡方分布
- 判断标准：统计量对应的P值