武汉大学金融系2023秋季学期

金融学/金融工程专业博士生方法论专题课

第二讲：外生变动与工具变量回归

授课人：刘 岩

2023年10月9日

本讲内容

1. 回归变量的内生性问题与工具变量回归理论
2. 工具变量回归相关检验
3. 工具变量回归示例
4. 工具变量回归的注意事项
5. 公司金融领域内生性偏误讨论

回归变量的内生性问题与工具变量回归理论

回归模型的内生性问题

- 回归模型：$Y_i = \alpha + \beta D_i + \boldsymbol{X}_i^{\top}\boldsymbol{\phi} + e_i$
  - 处置变量D_i的系数是需要识别的对象
- 识别条件：干扰项e_i的条件均值独立于处置变量D_i，即
$$\mathbb{E}(e_i|D_i, \boldsymbol{X}_i) = 0$$
- 内生性：$\mathbb{E}(e_i|D_i, \boldsymbol{X}_i) \neq 0 \Leftrightarrow \mathbb{E}(e_i D_i) = \mathrm{cov}(D_i, e_i) \neq 0$，即模型残差项中包含处置变量$D_i$的信息
  - 来源：倒向因果，遗漏变量，测量误差
- 两种方法来满足
  - “清理”干扰项：添加控制变量
  - “清理”解释变量：将处置变量中与干扰项不相关的部分分离出来⇒寻找工具变量

工具变量估计法的理解

- 假设结果方程为

$$Y_i = \alpha + \beta_1 D_i + e_i$$

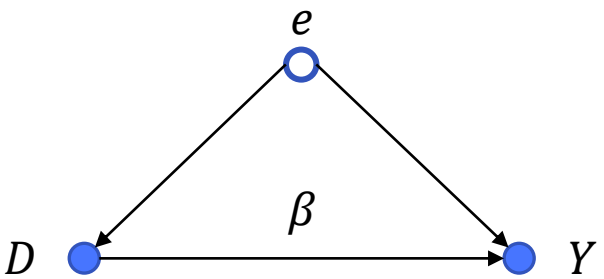
- 其中，D_i和干扰项e_i是相关的，即$\mathrm{cov}(D_i, e_i) \neq 0$
- OLS回归的系数$\hat{\beta}_1^{OLS}$为$\hat{\beta}_1^{OLS} = \frac{\widehat{\mathrm{cov}}(Y_i, D_i)}{\widehat{\mathrm{var}}(D_i)}$
- $\hat{\beta}_1^{OLS}$的样本概率极限值为

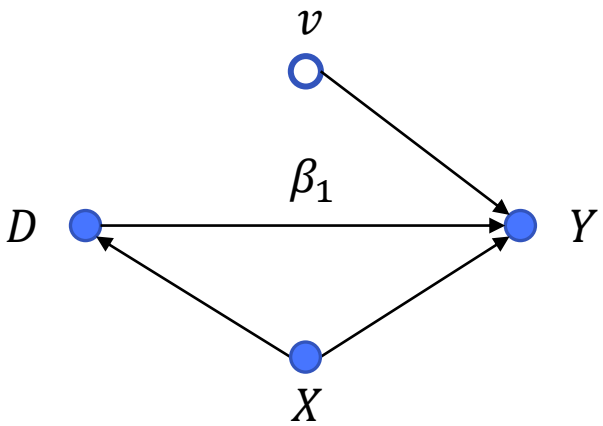
$$\mathrm{plim}\hat{\beta}_1^{OLS} = \mathrm{plim}\frac{\widehat{\mathrm{cov}}(Y_i, D_i)}{\widehat{\mathrm{var}}(D_i)} = \mathrm{plim}\frac{\widehat{\mathrm{cov}}(\alpha + \beta_1 D_i + e_i, D_i)}{\widehat{\mathrm{var}}(D_i)}$$

$$= \beta_1 + \mathrm{plim}\frac{\widehat{\mathrm{cov}}(e_i, D_i)}{\widehat{\mathrm{var}}(D_i)} = \beta_1 + \frac{\mathrm{cov}(e_i, D_i)}{\mathrm{var}(D_i)}$$

工具变量估计法的理解

- 因$\mathrm{cov}(D_i, e_i) \neq 0$，所以$\hat{\beta}_1^{OLS}$不以概率收敛为$\beta_1$
- D到Y的因果影响路径有两条：$D \to Y$（因果路径）和$D \leftarrow e \to Y$（混淆路径）



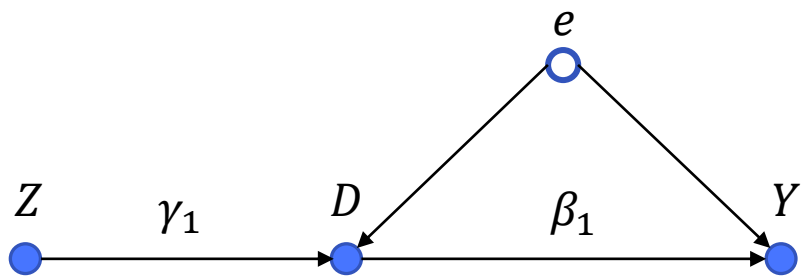


- 增加控制变量：通过观测得到X，将方程分解为

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + v_i$$

工具变量估计法的理解

- 使用工具变量：工具变量Z“清理”掉内生变量中与干扰项相关的变化（“坏”的变化），再用与干扰项不相关的变化（“好”的变化）去估计对Y的作用



  - 外生性：Z本身是“干净”的，Z和e_i不相关，即 $\text{cov}(Z_i, e_i) = 0$
  - 相关性：Z能够清理内生变量D，Z和D必须相关，即 $\text{cov}(Z_i, D_i) \neq 0$

估计方法——间接最小二乘法

- 假设内生变量D和工具变量Z的回归关系为

$$D_i = \gamma_0 + \gamma_1 Z_i + u_i$$

- 工具变量Z和被解释变量Y的关系为

$$Y_i = \alpha + \beta_1 D_i + e_i = \alpha + \beta_1(\gamma_0 + \gamma_1 Z_i + u_i) + e_i$$
$$= \alpha + \beta_1\gamma_0 + \beta_1\gamma_1 Z_i + \beta_1 u_i + e_i = \pi_0 + \pi_1 Z_i + \xi_i$$

- 通过回归得到$\pi_1 = \mathrm{cov}(Y_i, Z_i)/\mathrm{var}(Z_i)$，又$\pi_1 = \beta_1\gamma_1$，可得

$$\beta_1^{ILS} = \frac{\mathrm{cov}(Y_i, Z_i)/\mathrm{var}(Z_i)}{\mathrm{cov}(D_i, Z_i)/\mathrm{var}(Z_i)} = \frac{\mathrm{cov}(Y_i, Z_i)}{\mathrm{cov}(D_i, Z_i)}$$

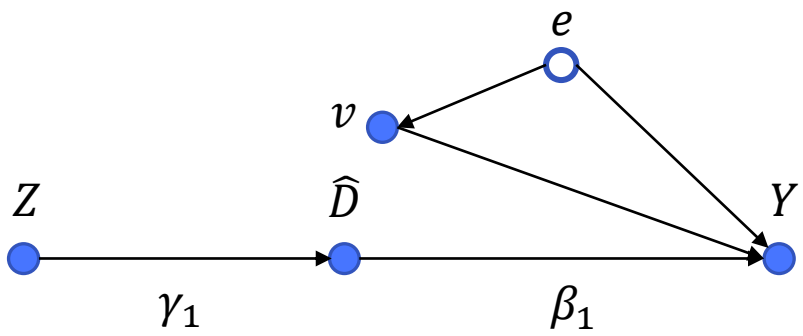
估计方法——两阶段最小二乘法

- 两阶段最小二乘法：通过直接分解出D_i中与干扰项e_i不相关的变化部分$\widehat{D}_i$来进行估计
- 第一阶段：通过工具变量Z_i将D_i分解为两个不相关变量

$$D_i = \widehat{D}_i + v_i = \gamma_0 + \gamma_1 Z_i + v_i$$

- 第二阶段：用D中“好的部分”$\widehat{D}$估计D对Y的影响

$$Y_i = \alpha + \beta_1 D_i + e_i = \alpha + \beta_1 \left(\widehat{D}_i + v_i\right) + e_i = \alpha + \beta_1 \widehat{D}_i + \delta_i$$



估计方法——两阶段最小二乘法

- 此时$\widehat{D}_i$与v_i和e_i都不相关，因此$\mathrm{cov}\left(\widehat{D}_i, \delta_i\right) = 0$，对其进行回归，得到正确的$\beta_1$

$$\beta_1^{2SLS} = \frac{\mathrm{cov}(Y_i, \widehat{D}_i)}{\mathrm{var}(\widehat{D}_i)} = \frac{\mathrm{cov}(Y_i, \gamma_0 + \gamma_1 Z_i)}{\mathrm{var}(\gamma_0 + \gamma_1 Z_i)} = \frac{\gamma_1 \mathrm{cov}(Y_i, Z_i)}{\gamma_1^2 \mathrm{var}(Z)} = \frac{\mathrm{cov}(Y_i, Z_i)}{\gamma_1 \mathrm{var}(Z_i)}$$

$$= \frac{\mathrm{cov}(Y_i, Z_i)}{\frac{\mathrm{cov}(D_i, Z_i)}{\mathrm{var}(Z_i)} \mathrm{var}(Z_i)} = \frac{\mathrm{cov}(Y_i, Z_i)}{\mathrm{cov}(D_i, Z_i)}$$

- 与间接最小二乘法得到的系数相同，方法处理方式稍有不同，本质上是一样的

工具变量数量问题

- 工具变量数量＜内生变量数量
  “识别不足”：两种方法都无法估计出内生变量的系数
- 工具变量数量=内生变量数量
  “刚好识别”：模型中内生变量系数可得到唯一估计值
- 工具变量数量＞内生变量数量
  - 分别使用不同的工具变量，会得到不同的β_1估计值
  - 使用两个工具变量的线性组合，使用2SLS可以得到一个多工具变量的“最佳组合”，来最佳拟合D，及通过回归方程

$$D_i = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + v_i$$

得到$\widehat{D}_i = \hat{\gamma}_0 + \hat{\gamma}_1 Z_{1i} + \hat{\gamma}_2 Z_{2i}$

工具变量数量问题

- 此时两阶段最小二乘法用两个工具变量的线性组合最大限度地分离出D_i中的外生部分$\widehat{D}_i$，再使用“最佳”外生部分$\widehat{D}_i$，将Y_i对$\widehat{D}_i$回归

$$Y_i = \alpha + \beta_1 \widehat{D}_i + \delta_i$$

- 因此更常使用两阶段最小二乘法

两阶段最小二乘法

- 假设结果模型是

$$Y_i = \alpha + \beta_1 D_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + e_i$$

- 工具变量Z_{1i}满足两个条件
  - 外生性：Z_{1i}与干扰项e_i不相关，$\mathrm{cov}(Z_{1i}, e_i) = 0$，“干净”地分离$D_{1i}$
    - 英文文献亦称其为exclusion restriction，中文翻译常见“排他性”条件
  - 相关性：控制外生变量与D_{1i}的相关性后，Z_{1i}与D_{1i}仍存在相关性

$$D_i = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 X_{2i} + \cdots + \gamma_k X_{ki} + v_i$$

其中，工具变量的系数$\gamma_1 \neq 0$

两阶段最小二乘法——模型估计

- 第一阶段：将内生变量对工具变量和所有外生变量进行回归

$$D_{1i} = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 X_{2i} + \cdots + \gamma_k X_{ki} + v_i$$

用所得到的系数计算内生变量的预测值

$$\widehat{D}_i = \hat{\gamma}_0 + \hat{\gamma}_1 Z_{1i} + \hat{\gamma}_2 X_{2i} + \cdots + \hat{\gamma}_k Z_{ki}$$

- 第二阶段：用预测值$\widehat{D}_i$替代结果模型中的内生变量D_{1i}，进行回归

$$Y_i = \alpha + \beta_1 \widehat{D}_i + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \delta_i$$

多个内生变量和多个工具变量

- 假设要估计的模型是

$$Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + e_i$$

  - D_{1i}的有一个工具变量Z_{1i}，D_{2i}的有两个工具变量Z_{2i}和Z_{3i}

- 第一阶段：将每个内生变量单独对所有工具变量和所有其他控制变量进行回归

$$D_{1i} = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \gamma_3 Z_{3i} + \gamma_4 X_{3i} + \cdots + \gamma_{k+1} X_{ki} + v_{1i}$$
$$D_{2i} = \theta_0 + \theta_1 Z_{1i} + \theta_2 Z_{2i} + \theta_3 Z_{3i} + \theta_4 X_{3i} + \cdots + \theta_{k+1} X_{ki} + v_{2i}$$

- 第二阶段：用内生变量预测值替代模型中的内生变量并进行回归

$$Y_i = \alpha + \beta_1 \widehat{D}_{1i} + \beta_2 \widehat{D}_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \delta_i$$

- 得到D_1和D_2系数的2SLS估计$\hat{\beta}_1^{2SLS}$和$\hat{\beta}_2^{2SLS}$是β_1和β_2的一致估计量

工具变量估计法的局限性——大样本

- 偏差性

$$\text{plim}\hat{\beta}_1^{2SLS} = \text{plim}\frac{\widehat{\text{cov}}(Y_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} = \beta_1 + \text{plim}\frac{\widehat{\text{cov}}(Z_i, e_i)}{\widehat{\text{cov}}(D_i, Z_i)} = \beta_1 + \frac{\text{cov}(Z_i, e_i)}{\text{cov}(D_i, Z_i)}$$

  - 当工具变量完全外生时，即$\text{cov}(Z_i, e_i) = 0$，2SLS大样本偏差项$= 0$，$\text{plim}\hat{\beta}_1^{2SLS} = \beta_1$，$\hat{\beta}_1^{2SLS}$是一致估计量
  - 当$\text{cov}(Z_i, e_i) \neq 0$时，即使分子很小，如果分母$\text{cov}(D_i, Z_i)$很小，偏差会被放得很大
  - 与内生变量相关性很小的工具变量被称为弱工具变量

工具变量估计法的局限性——大样本

- 比较$\hat{\beta}_1^{2SLS}$和$\hat{\beta}_1^{OLS}$在大样本下的偏差

$$\frac{\text{plim}\hat{\beta}_1^{2SLS} - \beta_1}{\text{plim}\hat{\beta}_1^{OLS} - \beta_1} = \frac{\text{cov}(Z_i, e_i)}{\text{cov}(D_i, Z_i)} \frac{\text{var}(D_i)}{\text{cov}(D_i, e_i)}$$

  - 如果工具变量不完全外生，且工具变量很弱，那么上式的比率可能远大于1，即使是在大样本中，两阶段工具变量估计量$\hat{\beta}_1^{2SLS}$也可能比OLS估计量的偏差还大。

工具变量估计法的局限性——大样本

- 有效性

  工具变量两阶段估计量$\hat{\beta}_1^{2SLS}$是渐近正态分布的

$$\hat{\beta}_1^{2SLS} \xrightarrow{\mathrm{d}} N\left(\beta_1, \mathrm{Avar}\left(\hat{\beta}_1^{2SLS}\right)\right)$$

  在同方差情况下，其渐进方差为

$$\mathrm{Avar}\left(\hat{\beta}_1^{2SLS}\right) = \frac{\sigma_e^2}{N\sigma_D^2\rho_{DZ}^2}$$

  不考虑内生性使用OLS估计模型，得到系数渐进方差

$$\mathrm{Avar}\left(\hat{\beta}_1^{OLS}\right) = \frac{\sigma_e^2}{N\sigma_D^2}$$

工具变量估计法的局限性——大样本

- 对比$\hat{\beta}_1^{2SLS}$和$\hat{\beta}_1^{OLS}$的方差，得到

$$\frac{\text{Avar}\left(\hat{\beta}_1^{2SLS}\right)}{\text{Avar}\left(\hat{\beta}_1^{OLS}\right)} = \frac{1}{\rho_{DZ}^2}$$

  - ρ_{DZ}小于1，故工具变量估计值的方差总是大于OLS估计值的方差，即$\text{Avar}\left(\hat{\beta}_1^{2SLS}\right) > \text{Avar}\left(\hat{\beta}_1^{OLS}\right)$因为工具变量估计值只使用了$D_i$中与工具变量相关的一部分信息，而OLS使用了全部信息
  - 在弱工具变量情况下，ρ_{DZ}很小，能分解出的内生变量“好”的信息很少，造成$\text{Avar}\left(\hat{\beta}_1^{2SLS}\right)$很大，此时通常的显著性检验并不可靠

工具变量估计法的局限性——有限样本

- 偏差性

  有限样本中，模型的$\hat{\beta}_1^{2SLS}$的偏差为

$$\hat{\beta}_1^{2SLS}\text{偏差} = \mathbb{E}\left(\hat{\beta}_1^{2SLS}\right) - \beta_1 = \frac{K\rho}{N}\left(\frac{1}{R^2} - 1\right)$$

  - 其中，K为工具变量数量，N为样本数量，ρ为内生变量与干扰项的相关系数及R^2为工具变量与内生变量的相关系数
  - 则弱工具变量会造成较大偏差，尤其样本数量较小的时候

工具变量估计法的局限性——有限样本

- 有限样本中，$\hat{\beta}_1^{2SLS}$与$\hat{\beta}_1^{OLS}$偏差的比率

$$\frac{\mathbb{E}\left(\hat{\beta}_1^{2SLS}\right)-\beta_1}{\mathbb{E}\left(\hat{\beta}_1^{OLS}\right)-\beta_1}=\frac{K}{NR^2}$$

  - 即使工具变量处理了内生性问题，如果工具变量太弱，在有限样本里的偏差有可能比OLS的还糟糕

- 有效性
  - 在有限样本里，$\hat{\beta}_1^{2SLS}$的分布未知，因此通常使用的t检验在小样本里并不适用

工具变量回归相关检验

工具变量运用的检验

- 是否需要使用工具变量：内生性检验
- 工具变量是否满足相关性：弱工具变量检验
- 工具变量是否是外生的：过度识别检验

工具变量运用的检验

- 是否需要使用工具变量：内生性检验
  - Durbin-Wu-Hausman χ^2_J 检验
  - 回归形式的Wu-Hausman F检验

工具变量运用的检验

Durbin-Wu-Hausman χ_J^2 检验

- 如果可能的内生变量 D_i 有一个合理的工具变量 Z_i：
  - 构造检验统计量如下：

    $$H = \left(\hat{\beta}_1^{2SLS} - \hat{\beta}_1^{OLS}\right)' \left[\text{Avar}\left(\hat{\beta}_1^{2SLS}\right) - \text{Avar}\left(\hat{\beta}_1^{OLS}\right)\right]^{-1} \left(\hat{\beta}_1^{2SLS} - \hat{\beta}_1^{OLS}\right) \sim \chi_J^2$$

  - 原假设H_0：D_i是外生的
  - 在原假设下，OLS估计量和工具变量估计得到的参数估计是一致的；备择假设下，OLS估计量是有偏的，所以二者差别$\hat{\beta}_1^{2SLS} - \hat{\beta}_1^{OLS}$应该较大，$H$值远远不等于零，故$H$值较大时，拒绝原假设，$D_i$是内生的

工具变量运用的检验

回归形式的Wu-Hausman F检验

- 考虑简单模型：

$$Y_i = \alpha + \beta_1 D_i + e_i$$
$$D_i = \gamma_0 + \gamma_1 Z_i + u_i$$
$$\mathrm{cov}(D_i, e_i) = \mathrm{cov}(u_i, e_i)$$

  - 原理：检验D_i的外生性就是检验$\mathrm{cov}(u_i, e_i)$是否为零
  - 操作：把干扰项的线性关系表示为$e_i = \rho u_i + \tau_i$，代入原估计方程$Y_i = \alpha + \beta_1 D_i + \rho u_i + \tau_i$，使用$\hat{u}_i = D_i - \hat{\gamma}_0 - \hat{\gamma}_1 Z_i$代入，估计检验原假设$H_0: \rho = 0$，如果$\rho$显著异于0，则拒绝$D_i$是外生的假设

工具变量运用的检验

工具变量是否满足相关性：弱工具变量检验

- 弱工具变量在有限样本和大样本下都对估计方差有很大负面影响，要避免使用弱工具变量
- 只有一个内生变量：观察2SLS中第一阶段关于所有工具变量系数同时为0的F检验，如果F值非常低，存在弱工具变量问题
  - 临界值有随机模拟计算得到专门列表
- 存在多个内生变量：Stock and Yogo (2005) 提供了一个检验方法，计算一个被称作Minimum eigenvalue的统计量，如果该统计量高于Stock and Yogo (2005) 给出的对应关键值，则不存在弱工具变量的问题

工具变量运用的检验

工具变量是否是外生的：过度识别检验

- 本质而言，我们不可能检验这个条件，因为干扰项无法被观测，但可以一定程度上对外生条件进行检验
- 恰当识别：无法检验
- 要得到一致的估计量$\hat{\beta}_1$的前提假设是工具变量外生：$\text{cov}(Z_i, \hat{e}_1) = \text{cov}\left(Z_i, Y_i - \hat{\beta}_1 D_i\right) = \text{cov}(Z_i, Y_i) - \hat{\beta}_1 \text{cov}(Z_i, D_i) = \text{cov}(Z_i, Y_i) - \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, D_i)} \text{cov}(Z_i, D_i) = 0$
  - 通过$\text{cov}(Z_i, \hat{e}_1) = 0$检验工具变量外生性没有意义

工具变量运用的检验

工具变量是否是外生的：过度识别检验

- 过度识别：假设存在两个工具变量，先假设第一个变量Z_1是外生的，并且只用Z_1作为工具变量得到系数$\hat{\beta}_1^{Z_1}$，那么残差$\hat{e}_1^{Z_1} = Y_i - \hat{\beta}_1^{Z_1} D_i$是一致的；接着使用残差对第二个工具变量检验$\mathrm{cov}\left(\hat{e}_1^{Z_1}, Z_2\right) = 0$即可检验第二个工具变量的外生性
- 必须先假设一个工具变量的外生性再去检验另一个的外生性，对结果的解释也必须谨慎：Z_2不通过外生性检验有可能是因为对Z_1的外生性假设不正确，因此，两个中至少有一个是内生的；Z_2通过外生性检验也可能是因为对Z_1的外生性假设不正确，因此，通过检验时也不能得出所有工具变量都是外生的结论

工具变量运用的检验

工具变量是否是外生的：过度识别检验

- 实际运用中的操作：假设检验
- 原假设H_0:所有工具变量都是外生的
  - 检验统计量：使用所有工具变量用2SLS进行回归得到残差项；将残差项作为被解释变量，所有工具变量作为解释变量OLS回归，得到R^2；检验统计量$NR^2 \sim \chi_q^2$，如果NR^2大于相关χ_q^2临界值，拒绝H_0，如果NR^2小于相关χ_q^2临界值，不拒绝H_0
- χ_q^2的取值：Sargan和Basmann的χ_q^2统计值；Wooldridge的稳健得分过度识别检验；Hansen的J统计值

工具变量的使用步骤

- 清晰地定义研究问题，描述经济机制，设置基本模型，对基本模型进行OLS回归，理解并描述模型可能存在的内生性问题的原因（反向因果，测量误差，遗漏变量等）
- 根据经济机制和理论基础选择有效的工具变量，并解释工具变量的外生性和相关性
  - 相关性一般比较容易解释
  - 最具挑战性、最关键之处：使用描述性语言和经济原理说明工具变量的外生性
- 使用工具变量估计法对模型进行估计，同时进行必要的统计检验，并谨慎的对结果进行解释
  - 检验变量的内生性：Hausman检验
  - 检验工具变量的相关性：报告第一阶段的F统计量，检验是否存在弱工具变量问题
  - 检验工具变量的外生性：过度识别检验

工具变量的使用步骤

- 将工具变量估计结果和OLS结果进行比较，理解为何结果存在差异
- 第三步中的三个检验只是辅助性检验，它们得到的结论是有限的，不能够代替前两步中对处置变量可能内生的原因和工具变量有效性的讨论

工具变量回归示例

工具变量运用举例

- Acemoglu, D., S. Johnson, and J. A. Robinson. 2001. The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review* 91:1369–1401.
- 研究问题：好的社会制度对经济发展是否有促进作用
- 经济机制：良好的社会制度意味着更好的产权保护和更少的扭曲资源配置的政策，会促进资产和人力资源的投入，并更有效率的产出

工具变量运用举例

- 基本模型

$$\log \text{ppp}\, GDP = \alpha + \beta_1 avexpr + \beta_2 lat\_abst + e$$

- $\log ppp\ GDP$：按购买力平价计算出的GDP取对数
- $avexpr$：1985~1995年企业免受政府盘剥的指数平均值，该值越大证明制度越好（国家制度）
- lat_abst：首都的纬度（地理位置）

工具变量运用举例

- 数据展示

```
. use "C:\Users\lenovo.LAPTOP-A3K6LB3T\Desktop\工具变量\maketable3.dta"

. des

Contains data from C:\Users\lenovo.LAPTOP-A3K6LB3T\Desktop\工具变量\maketable3.dta
  obs:           376
 vars:            11                          18 Jan 2010 22:27
                                              (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
lat_abst        float   %9.0g                 Abs(latitude of capital)/90
euro1900        float   %9.0g                 European settlers 1900, AJR
excolony        float   %9.0g                 =1 if was colony FLOPS definiti
avexpr          float   %9.0g                 average protection against expropriation risk
logpgp95        float   %9.0g                 log PPP GDP pc in 1995, World Bank
cons1           float   %9.0g                 cons on exec in 1st year indep
indtime         float   %9.0g                 years independent: 1995 minus firstyr
democ00a        float   %9.0g                 democracy in 1900
cons00a         float   %9.0g                 constraint on executive in 1900
extmort4        float   %9.0g                 corrected mort.
logem4          float   %9.0g                 log settler mortality
-------------------------------------------------------------------------------
Sorted by:
```

工具变量运用举例

- OLS回归结果

```
. regress logpgp95 avexpr lat_abst,robust

Linear regression                               Number of obs     =        111
                                                F(2, 108)         =     183.95
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6225
                                                Root MSE          =      .7108
```

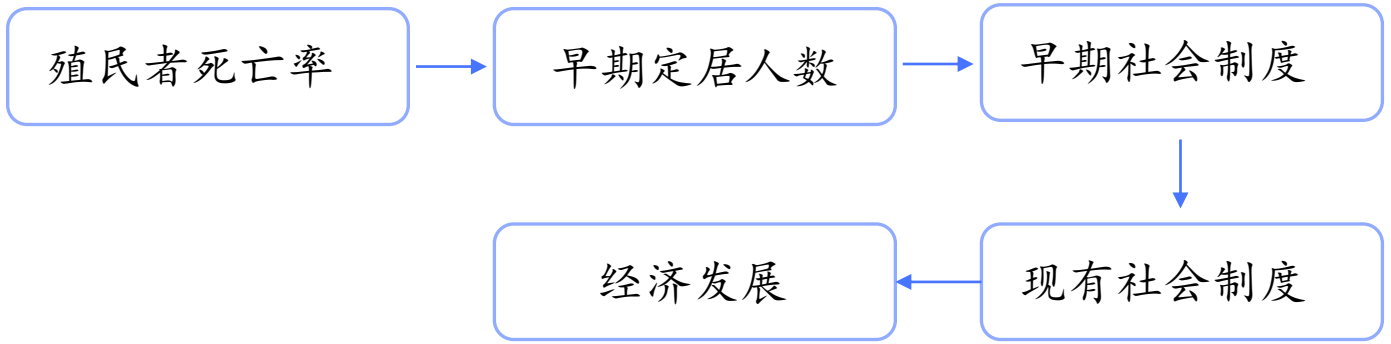
logpgp95	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
avexpr	.4634816	.0521728	8.88	0.000	.360066	.5668972
lat_abst	.8721613	.4993736	1.75	0.084	-.1176837	1.862006
_cons	4.872922	.2807513	17.36	0.000	4.316424	5.42942

工具变量运用举例

- OLS可能存在的问题
  - 发展较好的经济体更有可能建立良好的制度（反向因果）
  - 文化差异等其他变量同时影响国家经济状况和制度（混淆路径）
  - 社会制度不易准确衡量，测量可能存在较大偏差（测量误差）
  - 存在内生性问题：$\mathrm{cov}(avexpr, e) \neq 0$

工具变量运用举例

- 寻找有效的工具变量
  - 相关性：国家现在的社会制度一定程度上是过去制度的延续，早期社会制度又与欧洲殖民者的殖民政策有关



  - 外生性：殖民者死亡率的外生性最强

工具变量运用举例

- 使用工具变量法对模型进行估计

```
. ivregress 2sls logpgp lat_abst (avexpr = logem4),first

First-stage regressions
_____________________

                                      Number of obs     =         70
                                      F(   2,     67)   =      19.53
                                      Prob > F          =     0.0000
                                      R-squared         =     0.3682
                                      Adj R-squared     =     0.3494
                                      Root MSE          =     1.2523

------------------------------------------------------------------------------
      avexpr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lat_abst |   3.125466   1.203964     2.60   0.012     .7223438    5.528588
     logem4 |  -.4537058   .1304823    -3.48   0.001    -.7141496   -.1932619
       _cons |   8.094943   .7590112    10.67   0.000      6.57995    9.609936
------------------------------------------------------------------------------
```

工具变量运用举例

- 使用工具变量法对模型进行估计

```
Instrumental variables (2SLS) regression          Number of obs   =         70
                                                  Wald chi2(2)    =      39.18
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.0670
                                                  Root MSE        =     1.0159
------------------------------------------------------------------------------
    logpgp95 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avexpr |    1.029084   .2332977     4.41   0.000      .5718288    1.486339
    lat_abst |   -1.784366   1.494275    -1.19   0.232     -4.713092    1.144359
       _cons |     1.65175   1.322986     1.25   0.212     -.9412546    4.244754
------------------------------------------------------------------------------
Instrumented:  avexpr
Instruments:   lat_abst logem4
```

工具变量运用举例

- 对工具变量进行检验
  - Hausman检验解释变量是否外生

```
. estat endogenous

  Tests of endogeneity
  Ho: variables are exogenous

  Durbin (score) chi2(1)          =  16.4466  (p = 0.0001)
  Wu-Hausman F(1,66)              =  20.2691  (p = 0.0000)

.
```

工具变量运用举例

- 对工具变量进行检验
  - 检验工具变量是否为弱工具变量

```
. estat firststage

First-stage regression summary statistics
```

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	F(1,67)	Prob > F
avexpr	0.3682	0.3494	0.1529	12.0905	0.0009

```
Minimum eigenvalue statistic = 12.0905

Critical Values                       # of endogenous regressors:    1
Ho: Instruments are weak              # of excluded instruments:     1
```

	5%	10%	20%	30%
2SLS relative bias		(not available)		

	10%	15%	20%	25%
2SLS Size of nominal 5% Wald test	16.38	8.96	6.66	5.53
LIML Size of nominal 5% Wald test	16.38	8.96	6.66	5.53

工具变量运用举例

- 对工具变量进行检验
  - 过度识别情况下，检验工具变量是否是外生的
  - 检验工具变量有效性的另一个条件是外生性，此时需要过度识别，采用殖民者死亡率和定居人数同时作为工具变量进行估计

```
. ivregress 2sls logpgp lat_abst (avexpr = logem4 euro1900),first

First-stage regressions
```

```
                                  Number of obs     =         69
                                  F(    3,     65)  =      15.07
                                  Prob > F          =     0.0000
                                  R-squared         =     0.4103
                                  Adj R-squared     =     0.3830
                                  Root MSE          =     1.2271
```

avexpr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lat_abst	1.845078	1.355143	1.36	0.178	-.8613306	4.551486
logem4	-.3197937	.1453475	-2.20	0.031	-.6100726	-.0295148
euro1900	.0160626	.0077541	2.07	0.042	.0005765	.0315487
_cons	7.458333	.8117933	9.19	0.000	5.83707	9.079597

工具变量运用举例

- 对工具变量进行检验
  - 过度识别情况下，检验工具变量是否是外生的

```
Instrumental variables (2SLS) regression          Number of obs   =         69
                                                  Wald chi2(2)    =      47.23
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.1599
                                                  Root MSE        =     .96035
```

logpgp95	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
avexpr	.9735225	.1886416	5.16	0.000	.6037918	1.343253
lat_abst	-1.629427	1.291443	-1.26	0.207	-4.160609	.9017557
_cons	1.975545	1.073346	1.84	0.066	-.1281749	4.079265

```
Instrumented:  avexpr
Instruments:   lat_abst logem4 euro1900
```

工具变量运用举例

- 对工具变量进行检验
  - 过度识别检验

```
. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  .104916  (p = 0.7460)
  Basmann chi2(1)        =  .098985  (p = 0.7531)
```

工具变量运用举例

- 解释工具变量的经济显著性，并将估计结果和OLS结果进行对比，理解结果为何有差异
  - 经济显著性：制度对经济发展水平有较大实质性影响
  - 和OLS结果相比：2SLS的回归系数比OLS的回归系数高；同时，我们知道，前面提到的测量误差会导致向下偏差，逆向关系和缺少变量会导致向上偏差；因此，测量误差可能是造成系数差别的主要原因。

工具变量回归的注意事项

工具变量运用常见问题

- 用计量软件估计工具变量模型，不要自己手动进行两步回归
- 第一阶段回归的解释变量应包含所有的外生变量
- 避免使用组均值作为工具变量
- 避免使用内生变量的滞后项作为工具变量
- 模型含有二次项的工具变量的用法
- 模型存在交叉项时工具变量的用法
- 工具变量估计结果只是局部平均处置效应
- 工具变量越多越好吗？
- 工具变量是解决内生性的万灵药吗？

工具变量运用常见问题

- 用计量软件估计工具变量模型，不要自己手动进行两步回

$$\mathrm{Avar}\left(\hat{\beta}^{2SLS}\right) = \hat{\sigma}_e^2\left(\widehat{\boldsymbol{D}}'\widehat{\boldsymbol{D}}\right)^{-1}$$

$$\hat{\sigma}_e^2 = \frac{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}{N}, \qquad \hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{D}'\hat{\beta}^{2SLS}$$

- 注意此时计算残差时用到的时内生变量的原值而不是预测值，手动计算容易出$\hat{\boldsymbol{e}} = \boldsymbol{Y} - \widehat{\boldsymbol{D}}'\hat{\beta}^{2SLS}$的错误

工具变量运用常见问题

- 第一阶段回归的解释变量应包含所有的外生变量
  - 工具变量使用的一个常见的错误是，只使用工具变量作为第一阶段回归的解释变量
  - 正确的做法是将工具变量和所有的外生的解释变量都包含在第一步回归中（如果解释变量包含个体固定效应和时间固定效应，那么也应该加入第一步回归）
  - 不这么做的后果：第二步的参数估计不具备一致性

工具变量运用常见问题

- 避免使用组均值作为工具变量
  - 使用组内均值的理由：组内个体特征与组内均值有关（相关性）；组内其他个体的平均或加总特征不直接影响个体（外生性）
  - 实际上，组内均值作为工具变量，不满足外生性的要求

工具变量运用常见问题

- 避免使用内生变量的滞后项作为工具变量
  - 使用内生变量的滞后项作为工具变量的要求

    $$Y_{i,t} = \alpha + \beta X_{i,t} + e_{i,t}, \qquad \mathrm{cov}(X_{i,t}, X_{i,t-1}) \neq 0, \qquad \mathrm{cov}(X_{i,t-1}, e_{i,t}) = 0$$

  - 以上一方面要求内生变量是序列相关的，另一方面要求干扰项不存在序列相关，存在矛盾
  - 更详细讨论见Roberts and Whited (2013, Handbook chapter, sec. 3.6)

工具变量运用常见问题

- 模型含有二次项的工具变量的用法
  - 考虑模型$Y = \alpha + \beta_1 X + \beta_2 X^2 + e$，$X$有工具变量$Z$
  - 正确做法：分别使用Z和Z^2作为工具变量，使用计量软件估计模型

工具变量运用常见问题

- 模型存在交叉项时工具变量的用法
  - 考虑模型$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e$，$X_1$有工具变量$Z_1$，$X_2$外生
  - 正确做法：Z_1作为X_1的工具变量，$Z_1 X_2$作为$X_1 X_2$的工具变量

工具变量运用常见问题

- 工具变量估计结果只是局部平均处置效应
  - OLS参数估计值使用了解释变量所有的信息去估计被解释变量的变化，因此OLS估计得到的系数是平均处置效应：解释变量变动一个单位，引起的被解释变量均值的变化
  - 工具变量估计系数是使用了与工具变量相关的信息去估计被解释变量的变化，得到的系数描述的是对于那些其特征X会收到工具变量Z影响的个体，X变化一个单位引起的他们的Y均值的变化，即局部平均处置效应，局部指的是一部分个体
  - 工具变量的估计结果是局部平均处置效应也意味着选取不同的工具变量会得到不同的估计结果

工具变量运用常见问题

- 工具变量越多越好吗？
  - 找到更多的好的工具变量可以提高有效性，使用差的工具变量会放大估计偏差
  - 使用多工具变量权衡有效性和偏差性
  - 避免添加弱工具变量和同质工具变量
  - 比较不同工具变量的估计结果来决定是否添加某个工具变量

工具变量运用常见问题

- 工具变量是解决内生性的万灵药吗？
  - 好的工具变量可以有效地解决内生性问题
  - 外生性无法通过统计检验确认
  - 实际运用中外生性和相关性很难兼得：一个较为明确的外生工具变量通常相关性很低，一个相关性较高的工具变量又不满足外生性

公司金融领域内生性偏误讨论

公司金融领域工具变量回归的讨论

- Jiang, W. 2017. Have Instrumental Variables Brought Us Closer to the Truth. *Review of Corporate Finance Studies* 6:127–140.
  - 姜纬，哥伦比亚大学商学院教授
- 255篇2003-2014年金融Top 3 IV论文的梳理：80%的情况下，IV估计系数比OLS大，且平均而言大9倍
- 三类OLS内生性估计偏误
  - 证实性内生性：affirmative endogeneity，OLS会高估作用效果，论文占比67.1%
  - 纠正性内生性：corrective endogeneity，OLS会低估作用效果，论文占比18%
  - 待定内生性

偏误类型及分布

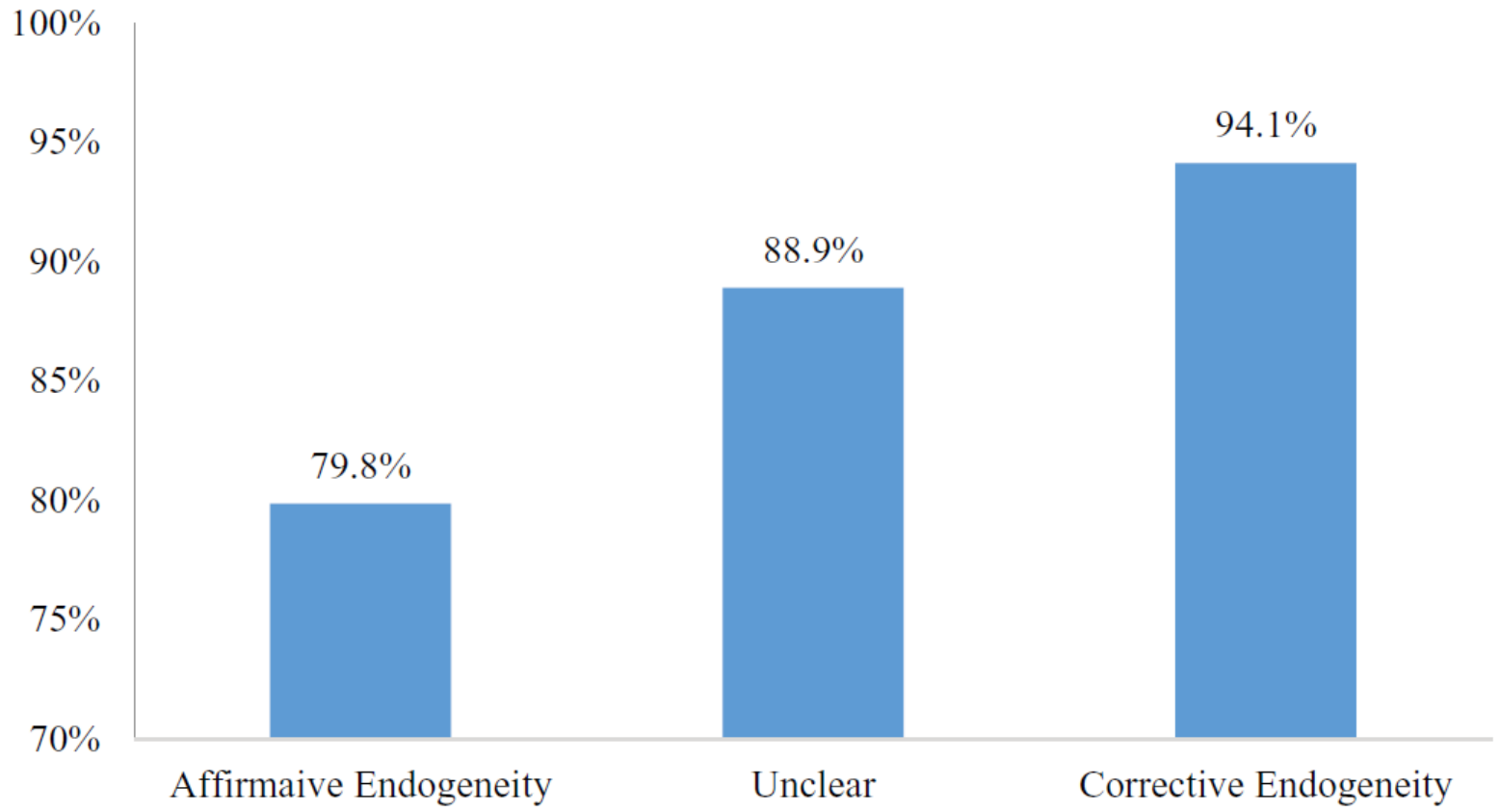


Figure 2
Percentage of papers with $|\beta^{IV}| > |\beta^{OLS}|$.
This chart shows the percentage of papers where the magnitude of IV estimates exceeds that of the OLS estimates, separately for the three categories of endogeneity, based on a priori information and economic reasoning.

偏误大小及分布

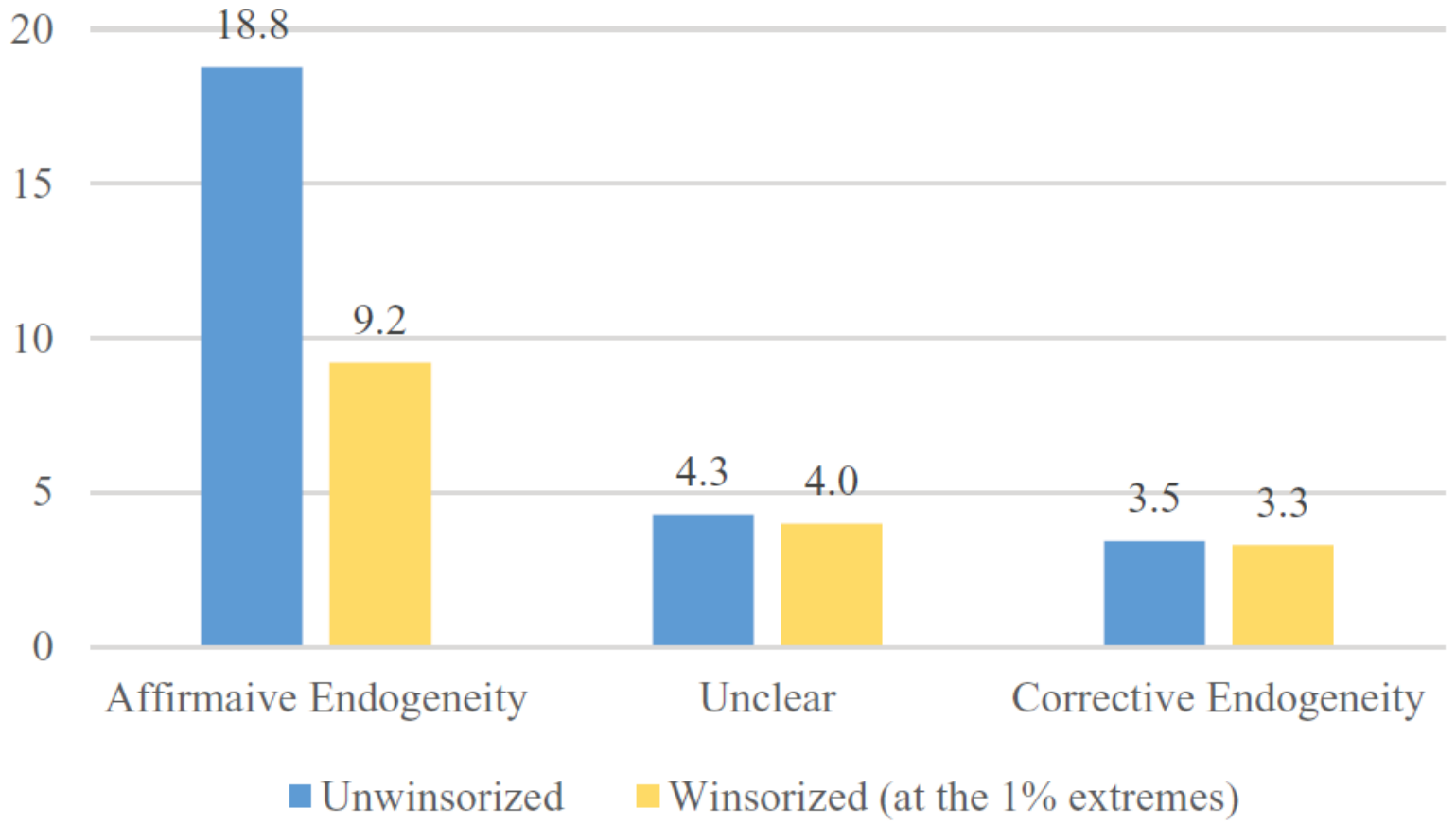


Figure 3
Ratio of $|\beta^{IV}/\beta^{OLS}|$.
This figure shows the average ratio of IV and OLS estimates across all papers by each endogeneity category.

偏误来源

- 工具变量回归本质是在估计局部处理效应(local average treatment effect, LATE)，而OLS估计可能反映全局平均效应
  - IV回归主要捕捉工具变量发生变化及其引起的内生变量的局部变化，产生的效应
- 弱工具变量：1阶段$X_i = \gamma Z_i + u_i$，2阶段结构方程$Y_i = \beta X_i + \epsilon_i = \beta X_i + \iota Z_i + \eta_i$

$$\hat{\beta}^{IV} = \frac{\mathrm{cov}(Y_i, Z_i)}{\mathrm{cov}(X_i, Z_i)} = \frac{\mathrm{cov}(\beta X_i + \iota Z_i + \eta_i, Z_i)}{\mathrm{cov}(X_i, Z_i)} = \frac{\beta \mathrm{cov}(X_i, Z_i) + \iota \mathrm{var}(Z_i)}{\mathrm{cov}(X_i, Z_i)} = \beta + \frac{\iota}{\gamma}$$

  - 弱工具变量意味着γ接近0，则给定IV固有偏差ι，$\hat{\beta}^{IV}$偏误与γ成反比
- 发表选择偏误：弱工具变量下，$\hat{\beta}^{IV}$必须足够大，才能使得IV估计结果满足显著性要求

IV回归诊断建议

- 明确讨论$\hat{\beta}^{IV}$的先验偏误，亦即$\hat{\beta}^{OLS}$的先验偏误：从理论出发，内生性问题会造成系数估计往哪个方向偏离，并结合估计结果进行讨论
- 在1阶段回归汇报IV变量的部分(partial)R方，即除去IV后，1阶段回归R方减少多少，如果IV在1阶段中对内生变量解释力贡献很小，则需要警惕弱工具变量偏误，无论是否通过“弱工具变量”检验
- 详细讨论IV回归系数的经济/现实合理性
  - 参考Mitton (2022b)：计算解释变量1个标准差变动所带来的被解释变量的变化，占被解释变量标准差的百分比