# 工具变量

慧航

上海对外经贸大学

2024年12月

# 工具变量

不管是遗漏变量、反响因果、度量误差还是自选择问题，都会导致核心解释变量与误差项的相关性，即在回归方程：

$$y_i = \gamma \times w_i + \tilde{x}_i'\delta + u_i$$

中$w_i$与误差项$u_i$之间相关。比如：

- 如果$w_i$为国家的GDP，$y_i$为一个国家内战爆发的次数
- 如果$w_i$为孩子的数量，$y_i$为孩子的教育经费（quantity-quality tradeoff）

解决方案：找到外生的影响$w_i$但是不会直接影响$y_i$的变量，即工具变量（**instrument variable**）$z_i$，比如：

- 内战例子中，可以使用天气作为国家GDP的工具变量
- 生育问题中，计划生育政策作为孩子数量的工具变量

# 工具变量的识别

如果我们要估计的结构方程（structural equation）为：

$$y_i = \alpha + \gamma w_i + u_i$$

然而$\mathbb{C}(w_i, u_i) \neq 0$，但是我们可以找到一个$z_i$，使得$\mathbb{C}(z_i, u_i) = 0$，且：

$$w_i = \eta + \phi z_i + v_i$$

我们把以上方程带入结构式，得到：

$$\begin{aligned} y_i &= \alpha + \gamma(\eta + \phi z_i + v_i) + u_i \\ &= (\alpha + \gamma\eta) + \gamma\phi z_i + u_i + \gamma v_i \end{aligned}$$

我们称上式为简约式（reduced form）。

# 工具变量的识别

现在两个式子：

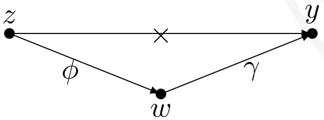
$$w_i = \eta + \phi z_i + v_i$$

以及

$$\begin{aligned} y_i &= (\alpha + \gamma\eta) + \gamma\phi z_i + u_i + \gamma v_i \\ &\triangleq \alpha^* + \gamma^* z_i + e_i \end{aligned}$$

都不存在内生性问题，因而我们可以一致估计$\eta, \phi, \alpha^*, \gamma^*$。进而，可以得到Wald估计量：

$$\hat{\gamma} = \frac{\hat{\gamma}^*}{\hat{\phi}} \xrightarrow{p} \frac{\gamma^*}{\phi} = \frac{\gamma\phi}{\phi} = \gamma$$



# 工具变量的识别假设

识别假设:

- 工具变量$z_i$与$u_i$不相关
  - 暗含$z_i$不直接影响$y_i$，即$z_i$不是$y_i$的直接原因，否则$z_i$会进入到$u_i$中
  - 此外，$z_i$也与$u_i$中的遗漏变量、度量误差等因素均不相关
- 分母$\phi = \frac{\mathbb{C}(z_i, w_i)}{\mathbb{V}(z_i)} \neq 0$（工具变量与内生变量高度相关）。
  - 不相关：欠识别（underidentification）
  - 相关性弱：弱工具（weak instrument）

# 工具变量

- 工具变量可以不止一个
- 如果估计方程：

$$y_i = \gamma w_i + \tilde{x}_i'\delta + u_i$$

其中$w_i$为关心的内生变量，即$\mathbb{C}\left(w_i, u_i\right) \neq 0$，但是存在（可以为多个）工具变量$\tilde{z}$，使得：

  - $\mathbb{E}\left(u_i|\tilde{z}\right) = 0$
  - 且$\mathbb{E}\left(w_i|\tilde{x}_i, \tilde{z}\right) \neq \mathbb{E}\left(w_i|\tilde{x}_i\right)$

- 那么$w_i$的简约式可以写为：

$$w_i = \tilde{z}_i'\phi + \tilde{x}_i'\xi + v_i$$

可以通过工具变量达到对$\gamma$的识别。

# 工具变量的一般形式

- 工具变量可以不止一个
- 如果估计方程：

$$y_i = \gamma w_i + \tilde{x}_i'\delta + u_i \triangleq x_i'\beta + u_i$$

其中$w_i$为关心的内生变量，即$\mathbb{C}(w_i, u_i) \neq 0$，但是存在（可以为多个）工具变量$\tilde{z}_i$，使得：
  - $\mathbb{E}(u_i|\tilde{x}_i, \tilde{z}) = 0$
  - 且$\mathbb{E}(w_i|\tilde{x}_i, \tilde{z}) \neq \mathbb{E}(w_i|\tilde{x}_i)$
- 那么$w_i$的简约式可以写为：

$$w_i = \tilde{z}_i'\phi + \tilde{x}_i'\xi + v_i \triangleq z_i'\psi + v_i$$

可以通过工具变量达到对$\gamma$的识别。

# 两阶段最小二乘

估计：两阶段最小二乘法（two-stage least squares, 2SLS）

1. 第一阶段回归（fisrt stage），对$w_i$的简约式进行回归：

$$w_i = \tilde{z}'\phi + \tilde{x}_i'\xi + v_i$$
$$\hat{w}_i = \tilde{z}'\hat{\phi} + \tilde{x}_i'\hat{\xi}$$

从而$\hat{w}_i$与$u_i$无关

2. 第二阶段回归，用$\hat{w}_i$代替$w_i$：

$$y_i = \hat{\gamma}\hat{w}_i + \tilde{x}_i'\hat{\xi} + \hat{\epsilon}_i$$

注意：不要手动算两阶段最小二乘，差别：在算标准误的时候用$\hat{\epsilon}_i$还是$\hat{u}_i = y_i - \hat{\gamma}w_i + \tilde{x}_i'\hat{\xi}$?

# 2SLS示例

## 降水与内战

经济条件（如经济增长）可能会影响一个国家爆发内战的可能性，然而经济条件与内战之间的关系可能存在着互为因果：经济差的时候更容易发生内战，但同时内战也会对经济增长带来负面影响。为了估计经济条件对内战的影响，Miguel，Satyanath和Sergenti （2004）使用降水作为经济增长的工具变量，研究了经济条件与内战爆发之间的关系。作者使用了撒哈拉以南的非洲国家作为样本，由于这些国家很多都是农业国家，因而其GDP很容易受到降水的影响。其估计的结构方程为

$$conflict_{it} = \beta_0 + \gamma \cdot growth_{it} + X_{it}'\delta + \alpha_i + \tau_t + \delta_i \cdot t + u_{it}$$

其中$conflict_{it}$为国家$i$在$t$年是否发生内战，$growth_{it}$为经济增长，$X_{it}$为其他控制变量，$\alpha_i$为国家固定效应，$\tau_t$为年份固定效应，在文章的设定中，作者加入了个体的时间趋势，即允许每个国家具有不同的时间固定效应系数。（CivilConflict.do）

# 多个内生变量

- 内生变量也可以不止一个
- 如果估计方程：

$$y_i = w_i'\gamma + \tilde{x}_i'\delta + u_i \triangleq x_i'\beta + u_i$$

此时需要工具变量$\tilde{z}_i$的个数大于等于内生变量$w_i$的个数（order condition），同样$\mathbb{E}\left(u_i|z_i\right) = 0$且$\mathbb{E}\left(w_i|\tilde{x}_i, \tilde{z}_i\right) \neq \mathbb{E}\left(w_i|\tilde{x}_i\right)$

- $w_i$的简约式：

$$\begin{aligned} w_{1i} &= \tilde{z}_i'\phi_1 + \tilde{x}_i'\xi_1 + v_{1i} \\ &\vdots \\ w_{Gi} &= \tilde{z}_i'\phi_G + \tilde{x}_i'\xi_G + v_{Gi} \end{aligned} \tag{1}$$

注意：**每个内生变量的第一阶段方程使用的外生变量是一样的！**

## 多个内生变量：Order Condition

- 如果工具变量个数小于内生变量个数，则不能识别

### Order Condition

如果结构方程为

$$y_i = \alpha + \gamma_1 w_{1i} + \gamma_2 w_{2i} + u_i$$

第一阶段方程：

$$\begin{aligned} w_{1i} &= \eta_1 + \phi_1 \tilde{z}_i + v_{1i} \\ w_{2i} &= \eta_2 + \phi_2 \tilde{z}_i + v_{2i} \end{aligned}$$

带入得到简约式：

$$\begin{aligned} y_i &= (\alpha + \gamma_1 \eta_1 + \gamma_2 \eta_2) + (\gamma_1 \phi_1 + \gamma_2 \phi_2)\, \tilde{z}_i + u_i + \gamma_1 v_{1i} + \gamma_2 v_{2i} \\ &= \alpha^* + \gamma^* \tilde{z}_i + e_i \end{aligned}$$

# 多个内生变量：Order Condition

## Order Condition

仿照前叙的识别方法：

$$\frac{\gamma^*}{\phi_1} = \gamma_1 + \gamma_2 \frac{\phi_2}{\phi_1}$$
$$\frac{\gamma^*}{\phi_2} = \gamma_2 + \gamma_1 \frac{\phi_1}{\phi_2}$$

联立以上两个方程，其中$\gamma^*, \phi_1, \phi_2$可以被一致估计，现在需要解出$\gamma_1, \gamma_2$。然而将上面两个式子稍微整理，都可以被整理为

$$\gamma_1\phi_1 + \gamma_2\phi_2 = \gamma^*$$

从而两个未知数只有一个方程，方程有无穷多组解，不可识别。

# 多个内生变量：Rank Condition

- 然而即使order contition满足，$\gamma_k$也不一定能被识别。

## Rank Condition

如果结构方程为

$$y_i = \alpha + \gamma_1 w_{1i} + \gamma_2 w_{2i} + u_i$$

第一阶段方程：

$$\begin{aligned} w_{1i} &= \eta_1 + \phi_{11}\tilde{z}_{1i} + \phi_{12}\tilde{z}_{2i} + v_{1i} \\ w_{2i} &= \eta_2 + \phi_{21}\tilde{z}_{1i} + \phi_{22}\tilde{z}_{2i} + v_{2i} \end{aligned}$$

带入得到简约式：

$$y_i = \alpha^* + \gamma_1^*\tilde{z}_{1i} + \gamma_2^*\tilde{z}_{2i} + \gamma_3^*\tilde{z}_{3i} + e_i$$

其中

$$\begin{cases} \gamma_1^* = \gamma_1\phi_{11} + \gamma_2\phi_{21} \\ \gamma_2^* = \gamma_1\phi_{12} + \gamma_2\phi_{22} \end{cases}$$

# 多个内生变量：Rank Condition

## Rank Condition

如果记

$$\Phi = \begin{bmatrix} \phi_{11} & \phi_{21} \\ \phi_{12} & \phi_{22} \end{bmatrix}$$

上式变为

$$\Phi\gamma = \begin{bmatrix} \phi_{11} & \phi_{21} \\ \phi_{12} & \phi_{22} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} \gamma_1^* \\ \gamma_2^* \end{bmatrix} = \gamma^*$$

从而$\gamma$能被识别的假设是

$$\text{rank}\left(\Phi\right) = 2$$

# 工具变量的一般设定

- 记

$$z_i = \begin{bmatrix} \tilde{z}_i \\ \tilde{x}_i \end{bmatrix}_{L\times 1}, x_i = \begin{bmatrix} w_i \\ \tilde{x}_i \end{bmatrix}_{K\times 1}, \beta = \begin{bmatrix} \gamma \\ \delta \end{bmatrix}_{K\times 1}$$

- 此外，记$\dim(w_i) = G, \dim(\tilde{x}_i) = M, \dim(\tilde{z}_i) = H$，从而$H + M = L, G + M = K$。
- Order condition 要求$L \geq K$，即$H \geq G$，否则无法识别。
- Rank condition：对于第一阶段回归式(1)，

$$\operatorname{rank}(\Phi) = \operatorname{rank}\left(\begin{bmatrix} \phi_1' \\ \vdots \\ \phi_G' \end{bmatrix}\right) = G$$

  - order condition是rank condition成立的必要条件

# 情形一：$H = G, \mathrm{rank}\left(\Phi\right) = G$

- 使用矩估计法：

$$\mathbb{E}\left(z_i u_i\right) = \mathbb{E}\left(z_i\left(y_i - x_i'\beta\right)\right) = \mathbb{E}\left(\left(y_i - w_i'\gamma - \tilde{x}_i'\delta\right)\begin{bmatrix}\tilde{z}_i \\ \tilde{x}_i\end{bmatrix}\right) = 0$$

从而

$$\mathbb{E}\left(z_i x_i'\right)\beta = \mathbb{E}\left(z_i y_i\right)$$

即

$$\beta = \left[\mathbb{E}\left(z_i x_i'\right)\right]^{-1}\mathbb{E}\left(z_i y_i\right)$$

从而

$$\hat{\beta} = \left[\sum_{i=1}^{N}\left(z_i x_i'\right)\right]^{-1}\sum_{i=1}^{N}\left(z_i y_i\right)$$

- 注意如果$z_i = x_i$，从而$\tilde{z}_i = w_i$，以上估计即OLS估计量：OLS估计量即所有的解释变量作为自己的工具变量。

# 情形二：$H > G, \text{rank}(\Phi) = G$

- 前面已经知道，如果$\text{rank}(\Phi) < G$，那么$\gamma$不可被识别，此时称欠识别（underidentification）
- 然而如果$\text{rank}(\Phi) = G$，可以区分两种情况：
  - $H = G$，此时工具变量个数等于内生变量个数，称为恰好识别（just identified）
  - $H > G$，此时工具变量个数多余内生变量个数，此时是否可以识别？此时称为过度识别（overidentified）

# 情形二：$H > G, \operatorname{rank}(\Phi) = G$

## Rank Condition

如果结构方程为

$$y_i = \alpha + \gamma_1 w_{1i} + \gamma_2 w_{2i} + u_i$$

第一阶段方程：

$$\begin{aligned} w_{1i} &= \eta_1 + \phi_{11}\tilde{z}_{1i} + \phi_{12}\tilde{z}_{2i} + \phi_{13}\tilde{z}_{3i} + v_{1i} \\ w_{2i} &= \eta_2 + \phi_{21}\tilde{z}_{1i} + \phi_{22}\tilde{z}_{2i} + \phi_{23}\tilde{z}_{3i} + v_{2i} \end{aligned}$$

带入得到简约式，类似得到

$$\begin{cases} \gamma_1^* = \gamma_1\phi_{11} + \gamma_2\phi_{21} \\ \gamma_2^* = \gamma_1\phi_{12} + \gamma_2\phi_{22} \\ \gamma_3^* = \gamma_1\phi_{13} + \gamma_2\phi_{23} \end{cases} \Rightarrow \begin{bmatrix} \phi_{11} & \phi_{21} \\ \phi_{12} & \phi_{22} \\ \phi_{13} & \phi_{23} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} \gamma_1^* \\ \gamma_2^* \\ \gamma_3^* \end{bmatrix} \tag{2}$$

# 情形二： $H > G, \operatorname{rank}(\Phi) = G$

## Rank Condition

- 注意以上有两个未知数，三个方程，即使

$$\operatorname{rank}(\Phi) = \operatorname{rank}\left(\begin{bmatrix} \phi_{11} & \phi_{21} \\ \phi_{12} & \phi_{22} \\ \phi_{13} & \phi_{23} \end{bmatrix}\right) = 2$$

  以上方程也可能无解。
- 然而注意到，如果任取其中两个方程，都可以得到恰好识别。
- 从总体的层面，只要$\mathbb{E}(u_i \tilde{z}_{1i}) = \mathbb{E}(u_i \tilde{z}_{2i}) = \mathbb{E}(u_i \tilde{z}_{3i}) = 0$，方程组(2)应该是成立、有解的，此时也是可以被识别的。
  - 取任意两个方程得到解，带入第三个方程也应该是成立的。

# 估计：2SLS

估计：只要rank $(\Phi) = G$，不管$H > G$或者$H = G$，都可以使用2SLS

1. 第一阶段回归，对$w_i$的简约式进行回归：

$$\hat{w}_{1i} = \tilde{z}_i'\hat{\phi}_1 + \tilde{x}_i'\hat{\xi}_1$$
$$\vdots$$
$$\hat{w}_{Gi} = \tilde{z}_i'\hat{\phi}_G + \tilde{x}_i'\hat{\xi}_G$$

记$\hat{w}_i = \begin{bmatrix} \hat{w}_{1i} & \cdots & \hat{w}_{Gi} \end{bmatrix}'$

2. 第二阶段回归，用$\hat{w}_i$代替$w_i$：

$$y_i = \hat{w}_i'\hat{\gamma} + \tilde{x}_i'\hat{\xi} + \hat{\epsilon}_i$$

同样：不要手动算两阶段最小二乘

# 2SLS示例

## 降水与内战

在内战的例子中，作者还加入了滞后的经济增长率作为额外的内生变量：

$$conflict_{it} = \beta_0 + \gamma_0 \cdot growth_{it} + \gamma_1 \cdot growth_{i,t-1} + X_{it}'\delta + \alpha_i + \tau_t + \delta_i \cdot t + u_{it}$$

此时，有两个内生变量，需要为两个内生变量至少找两个工具变量，一个自然的选择是使用降水的滞后项作为经济增长的滞后项的工具变量。此外，原文中还引入了其他降水的数据来源，共4个工具变量：

# 2SLS示例

## 降水与内战

```
// 内生变量：增长率GDP 、增长率之后；工具变量：降水增长率gdp_gGDPgdp_g_l 、降水增长率滞后GPCP_g GPCP_g_l // 第一阶段
reg gdp_g GPCP_g GPCP_g_l i.ccode i.year i.ccode#c.year
reg gdp_g_l GPCP_g GPCP_g_l i.ccode i.year i.ccode#c.year
// 两阶段最小二乘，三种命令
ivregress 2sls any_prio (gdp_g gdp_g_l = GPCP_g GPCP_g_l ) i.ccode i.year i.ccode#c.year, cl(ccode)
ivreghdfe any_prio (gdp_g gdp_g_l = GPCP_g GPCP_g_l ), absorb(i.ccode i.year i.ccode#c.year) cl(ccode)
ivreg2  any_prio (gdp_g gdp_g_l = GPCP_g GPCP_g_l ) i.ccode i.year i.ccode#c.year, cl(ccode)
// 两阶段最小二乘，多余的四个()工具变量
ivreghdfe any_prio (gdp_g gdp_g_l = GPCP_g GPCP_g_l NCEP_g NCEP_g_l), absorb(i.ccode i.year i.ccode#c.year) cl(ccode)
```

# GMM估计

- 两阶段最小二乘即在特定矩条件下的广义矩估计（GMM）
- 矩条件为：$\mathbb{E}(z_i u_i) = 0, z_i = \begin{bmatrix} \tilde{z}' & \tilde{x}_i' \end{bmatrix}', x_i = \begin{bmatrix} w_i & \tilde{x}_i' \end{bmatrix}', \beta = \begin{bmatrix} \gamma & \delta' \end{bmatrix}'$，且$\dim(z_i) = L, \dim(x_i) = K$，则GMM目标函数：

$$\min \left[ \sum_i z_i \left( y_i - x_i'\beta \right) \right]' \mathcal{W} \left[ \sum_i z_i \left( y_i - x_i'\beta \right) \right]$$

- 如果取$\mathcal{W} = Z'Z$，其中

$$Z = \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix}_{N \times L}$$

则得到了两阶段最小二乘（2SLS）。
  - 在同方差假定下，以上的权重矩阵为最优权重矩阵。
  - 当同方差假定不满足时，以上的权重矩阵不是最优权重矩阵。可以使用twostep、igmm等计算最优权重矩阵。

# 过度识别检验

- 在$L > K$即$H > G$的情形下（不能等于），我们可以检验矩条件（工具变量）是否有效。
- 直觉：
  - 考虑方程组(2)，如果$\mathbb{E}\left(u_i\tilde{z}_{1i}\right) = \mathbb{E}\left(u_i\tilde{z}_{2i}\right) = \mathbb{E}\left(u_i\tilde{z}_{3i}\right) = 0$，方程组(2)应该都是成立的，从而任意两个方程得到的结果带入到第3个方程，等号同样成立。
  - 但是，只要有一个不为0，那么方程组(2)应该是不成立的，从而任意两个方程得到的结果带入到第3个方程，等号不成立。

# 过度识别检验

- 可以证明，当使用最优权重矩阵

$$\mathcal{W}^* = \left[\sum_i \left(u_i^2 z_i z_i'\right)\right]^{-1}$$

时，GMM目标函数渐进服从$\chi^2$分布，自由度为$L-K$：

$$\left[\sum_i z_i \left(y_i - x_i'\beta\right)\right]' \mathcal{W}^* \left[\sum_i z_i \left(y_i - x_i'\beta\right)\right] \overset{a}{\sim} \chi^2 \left(L-K\right)$$

  - Hansen's J-statistics
  - 显著好还是不显著好？
- Sargan test：首先得到残差，再使用残差对所有外生变量做回归，使用F检验对除常数项外的所有外省变量系数均为0的原假设做检验。

# 控制函数法

控制函数法：如果估计方程：

$$y_i = w_i'\gamma + \tilde{x}_i'\delta + u_i$$

以及第一阶段：

$$w_i = \tilde{z}_i'\phi + \tilde{x}_i'\xi + v_i$$

在简约式中，$w_i$被分解为两部分：

1. $\tilde{z}_i'\phi + \tilde{x}_i'\xi$是与$u_i$不相关的部分
2. 如果$w_i$与$u_i$相关，则相关性必在$v_i$里面。

# 控制函数法

因而，如果假定：

$$u_i = \rho v_i + \epsilon_i$$

那么：

$$\begin{aligned}\mathbb{E}\left(y_i|\tilde{x}_i, \tilde{z}_i, v_i\right) &= \alpha\mathbb{E}\left(w_i|\tilde{x}_i, \tilde{z}_i, v_i\right) + \tilde{x}_i'\delta + \mathbb{E}\left(u|\tilde{x}_i, \tilde{z}_i, v_i\right) \\ &= \alpha w_i + \tilde{x}_i'\delta + \rho v_i + \epsilon_i\end{aligned}$$

待估计方程变为：

$$y_i = w_i'\gamma + \tilde{x}_i'\delta + \rho v_i + \epsilon_i$$

因而如果观察到$v_i$，就可以直接控制「内生性」$v_i$从而达到识别。

# 控制函数法

估计步骤：

1. 估计第一阶段$w_i = \tilde{z}_i'\phi + \tilde{x}_i'\xi + v_i$得到$\hat{v}_i$
2. 将$\hat{v}_i$带入主方程，使用回归：

$$y_i = w_i'\gamma + \tilde{x}_i'\delta + \rho\hat{v}_i + \epsilon_i$$

   进行估计。

检验内生性：$H_0 : \rho = 0$（或者Hausman检验）。

# 控制函数法

- 或者，可以使用极大似然估计，即假设$(u_i, v_i)$的联合（正态）分布，进而：

$$f\left(y_i, w_i | \tilde{x}_i, \tilde{z}_i\right) = f\left(y_i | w_i, \tilde{x}_i, \tilde{z}_i\right) \cdot f\left(w_i | \tilde{x}_i, \tilde{z}_i\right)$$

  如果假定$(u_i, v_i)$为联合正态分布，即得到有限信息极大似然估计（LIML）。
- 其他估计方法：$k$−class估计量

# 控制函数法

虽然在最简单的情况下，控制函数法与2SLS等价，但是一般情况下，两者不等价：

- $y = \alpha w + \gamma w^2 + x'\beta + u$
  - 2SLS: 使用$z, z^2$做工具?
  - 控制函数法：$y = \alpha w + \gamma w^2 + x'\beta + \rho v + e$
- $y = \alpha w + \gamma x_k w + x'\beta + u$
  - 2SLS: 使用$z, x_k z$做工具?
  - 控制函数法：$y = \alpha w + \gamma x_k w + x'\beta + \rho v + e$
- $y = \alpha d + x'\beta + u$，$d = 1\{z'\gamma + x'\delta + v \geq 0\}$
  - $\mathbb{E}(y|x, z, v) = \alpha d + x'\beta + \mathbb{E}(u|v \geq -z'\gamma - x'\delta)$
  - 假定$(u_i, v_i)$联合正态，$\mathbb{E}(u|v \geq -z'\gamma - x'\delta)$为Inverse-Mills ratio

# 工具变量

工具变量的两个假定:

1. 与误差项不相关——Hansen's test
2. 与内生变量高度相关

如果第二项假定不满足?

# 欠识别检验

- 对于第一阶段：

$$w_i = \tilde{z}_i'\phi + \tilde{x}_i'\xi + v_i$$

- 我们要求工具变量至少要和内生变量相关，如果不相关：**欠识别（underidentification）**
- 检验：
  - 原假设：$H_0 : \phi = 0$ 使用$F$检验——第一阶段$F$值
  - 多个内生变量时，还需要rank condition：

$$\text{rank}\left(\Phi\right) = G$$

  - Stata：Kleibergen and Paap rk Wald F

# 弱工具

以上得知，对于要估计的方程：

$$y_i = \alpha + \gamma w_i + u_i$$

以及第一阶段：

$$w_i = \eta + \phi z_i + v_i$$

Wald估计可以写为：

$$\hat{\beta}_{2SLS} = \frac{\mathbb{C}\left(y_i, z_i\right)}{\mathbb{C}\left(w_i, z_i\right)}$$

如果分母趋向于0，IV估计的结果会非常不稳定，并偏向OLS估计量——弱工具。

# 工具变量

- Angrist and Krueger (1991)研究教育回报时，使用了出生季节与州、年份的乘积作为上学时间的工具变量
- Bound, Jaeger and Baker (1996)对此研究做出了批评，发现随机生成一组出生季节仍然能得到相似的结论。
- 工具变量太多导致估计结果偏向OLS的估计结果

# 工具变量

- 对于少量弱工具：
  - 诊断弱工具
  - 使用LIML，并调整置信区间。
- 对于Many instruments：
  - 2SLS表现非常差
  - 使用LIML，或者REQML(Chamberlain and Imbense, 2004)
  - 使用Lasso对第一阶段工具进行挑选

# 弱工具诊断

弱工具诊断:

- 第一阶段$F$值，一般在一个内生变量一个工具变量的情况下，为了使得2SLS估计量偏差不大于10%，$F$要大于10
- Cragg-Donald 统计量（$F - value$推广，当只有一个内生变量时就是$F - value$）
- Stock and Yogo (2002)提出了针对Cragg-Donald 统计量的临界值

置信区间调整:

- Anderson and Rubin (1949)
- Lee等人（2022）：根据第一阶段$F$值进行调整
- Valid $t$-ratio inference, Lee et al.(2022)

# $tF$标准误

- 诊断与置信区间调整方法的结合
- Stock and Yogo (2005)的诊断方法可以理解为:

$$P\left\{t^2 > c^*, F > F^*\right\} \le \alpha$$

其中$F$为第一阶段的$F$值。以上步骤即首先判断第一阶段$F$值是否大于某个临界值，进而使用传统的$t$检验

- Lee等人（2022）：

$$P\left\{t^2 > c_\alpha\left(F\right)\right\} \le \alpha$$

即根据第一阶段$F$值调整$t$统计量的临界值

  - 缺点：目前只支持一个内生变量、一个工具变量

# $tF$标准误

*Panel A. Selected values of tF critical values,* $\sqrt{c_{0.05}(F)}$, *and tF standard error adjustments,* $\sqrt{c_{0.05}(F)}/1.96$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $F$ | 4.000 | 4.008 | 4.015 | 4.023 | 4.031 | 4.040 | 4.049 | 4.059 | 4.068 | 4.079 |
| $\sqrt{c_{0.05}(F)}$ | 18.656 | 18.236 | 17.826 | 17.425 | 17.033 | 16.649 | 16.275 | 15.909 | 15.551 | 15.201 |
| $\sqrt{c_{0.05}(F)}/1.96$ | 9.519 | 9.305 | 9.095 | 8.891 | 8.691 | 8.495 | 8.304 | 8.117 | 7.934 | 7.756 |
| | 4.090 | 4.101 | 4.113 | 4.125 | 4.138 | 4.151 | 4.166 | 4.180 | 4.196 | 4.212 |
| | 14.859 | 14.524 | 14.197 | 13.878 | 13.566 | 13.260 | 12.962 | 12.670 | 12.385 | 12.107 |
| | 7.581 | 7.411 | 7.244 | 7.081 | 6.922 | 6.766 | 6.614 | 6.465 | 6.319 | 6.177 |
| | 4.229 | 4.247 | 4.265 | 4.285 | 4.305 | 4.326 | 4.349 | 4.372 | 4.396 | 4.422 |
| | 11.834 | 11.568 | 11.308 | 11.053 | 10.804 | 10.561 | 10.324 | 10.091 | 9.864 | 9.642 |
| | 6.038 | 5.902 | 5.770 | 5.640 | 5.513 | 5.389 | 5.268 | 5.149 | 5.033 | 4.920 |
| | 4.449 | 4.477 | 4.507 | 4.538 | 4.570 | 4.604 | 4.640 | 4.678 | 4.717 | 4.759 |
| | 9.425 | 9.213 | 9.006 | 8.803 | 8.605 | 8.412 | 8.222 | 8.037 | 7.856 | 7.680 |
| | 4.809 | 4.701 | 4.595 | 4.492 | 4.391 | 4.292 | 4.195 | 4.101 | 4.009 | 3.919 |
| | 4.803 | 4.849 | 4.897 | 4.948 | 5.002 | 5.059 | 5.119 | 5.182 | 5.248 | 5.319 |
| | 7.507 | 7.338 | 7.173 | 7.011 | 6.854 | 6.699 | 6.549 | 6.401 | 6.257 | 6.117 |
| | 3.830 | 3.744 | 3.660 | 3.578 | 3.497 | 3.418 | 3.341 | 3.266 | 3.193 | 3.121 |
| | 5.393 | 5.472 | 5.556 | 5.644 | 5.738 | 5.838 | 5.944 | 6.056 | 6.176 | 6.304 |
| | 5.979 | 5.844 | 5.713 | 5.584 | 5.459 | 5.336 | 5.216 | 5.098 | 4.984 | 4.872 |
| | 3.051 | 2.982 | 2.915 | 2.849 | 2.785 | 2.723 | 2.661 | 2.602 | 2.543 | 2.486 |
| | 6.440 | 6.585 | 6.741 | 6.907 | 7.085 | 7.276 | 7.482 | 7.702 | 7.940 | 8.196 |
| | 4.762 | 4.655 | 4.550 | 4.448 | 4.348 | 4.250 | 4.154 | 4.061 | 3.969 | 3.880 |
| | 2.430 | 2.375 | 2.322 | 2.270 | 2.218 | 2.169 | 2.120 | 2.072 | 2.025 | 1.980 |
| | 8.473 | 8.773 | 9.098 | 9.451 | 9.835 | 10.253 | 10.711 | 11.214 | 11.766 | 12.374 |
| | 3.793 | 3.707 | 3.624 | 3.542 | 3.463 | 3.385 | 3.309 | 3.234 | 3.161 | 3.090 |
| | 1.935 | 1.892 | 1.849 | 1.808 | 1.767 | 1.727 | 1.688 | 1.650 | 1.613 | 1.577 |
| | 13.048 | 13.796 | 14.631 | 15.566 | 16.618 | 17.810 | 19.167 | 20.721 | 22.516 | 24.605 |
| | 3.021 | 2.953 | 2.886 | 2.821 | 2.758 | 2.696 | 2.635 | 2.576 | 2.518 | 2.461 |
| | 1.542 | 1.507 | 1.473 | 1.440 | 1.407 | 1.376 | 1.345 | 1.315 | 1.285 | 1.256 |
| | 27.058 | 29.967 | 33.457 | 37.699 | 42.930 | 49.495 | 57.902 | 68.930 | 83.823 | 104.67 |
| | 2.406 | 2.352 | 2.299 | 2.247 | 2.197 | 2.147 | 2.099 | 2.052 | 2.006 | 1.96 |
| | 1.228 | 1.200 | 1.173 | 1.147 | 1.121 | 1.096 | 1.071 | 1.047 | 1.024 | 1.00 |

# $tF$标准误

- 或者可以使用软件包：
  - net install tf, force from(http://www.princeton.edu/~davidlee/wp/)
  - 需要首先安装ivreg2、ranktest
  - 用法与ivreg2一致