4.1 工具变量与因果关系

孔彤阳

2019.10.17

工具变量IV Instrumental variables

- 满足条件独立性假设(CIA)—回归估计量具有因果效应的解释
- CIA不成立—未观测因素会造成选择偏差，内生性问题

  工具变量是解决内生性问题的一种经典方法

- 工具变量可用来估计线性联立方程的系数

  起源：Wright父子研究亚麻籽的供求联立方程组

一个例子

- 教育水平和收入：

$$Y_{si} = f_i(\mathrm{s})$$

$$f_i(\mathrm{s}) = \alpha + \rho s + \eta_i$$

引入控制变量 能力A_i

$$\eta_i = A_i'\gamma + \nu_i$$

假设s_i与η_i相关的唯一原因是A_i，所以

$$E[s_i \nu_i] = 0$$

如果A_i可观察到，则有

$$Y_i = \alpha + \rho s_i + A_i'\gamma + \nu_i$$

- A_i无法观察到，如何估计ρ？

工具变量法

引入工具变量Z_i

- 相关性：与解释变量s_i相关

$$cov(s_i, Z_i) \neq 0$$

- 排他性(exclusion restriction)：与决定被解释变量的其他因素无关

$$cov(\eta_i, Z_i) = 0$$

其中，$\eta_i = A_i'\gamma + \nu_i$

工具变量法

- 计算工具变量与结果之间的相关性

$$cov(Y_i, Z_i) = cov(\alpha, Z_i) + \rho cov(s_i, Z_i) + cov(\eta_i, Z_i)$$

因为

$$cov(\alpha, Z_i) = cov(\eta_i, Z_i) = 0$$

所以上式整理得

$$\rho = \frac{cov(Y_i, Z_i)}{cov(s_i, Z_i)} = \frac{cov(Y_i, Z_i)/Var(Z_i)}{cov(s_i, Z_i)/Var(Z_i)}$$

- 分子$cov(Y_i, Z_i)/Var(Z_i)$是Y_i对Z_i的总体回归系数，称之为“简约式”
- 分母$cov(s_i, Z_i)/Var(Z_i)$是s_i对Z_i的总体回归系数，称之为“第一阶段”

一个例子

Angrist和Krueger利用美国普查数据估计教育收益率

能力变量影响个体的教育选择和收入水平，但无法预测
无法利用匹配或回归方法得到教育的正确因果效应

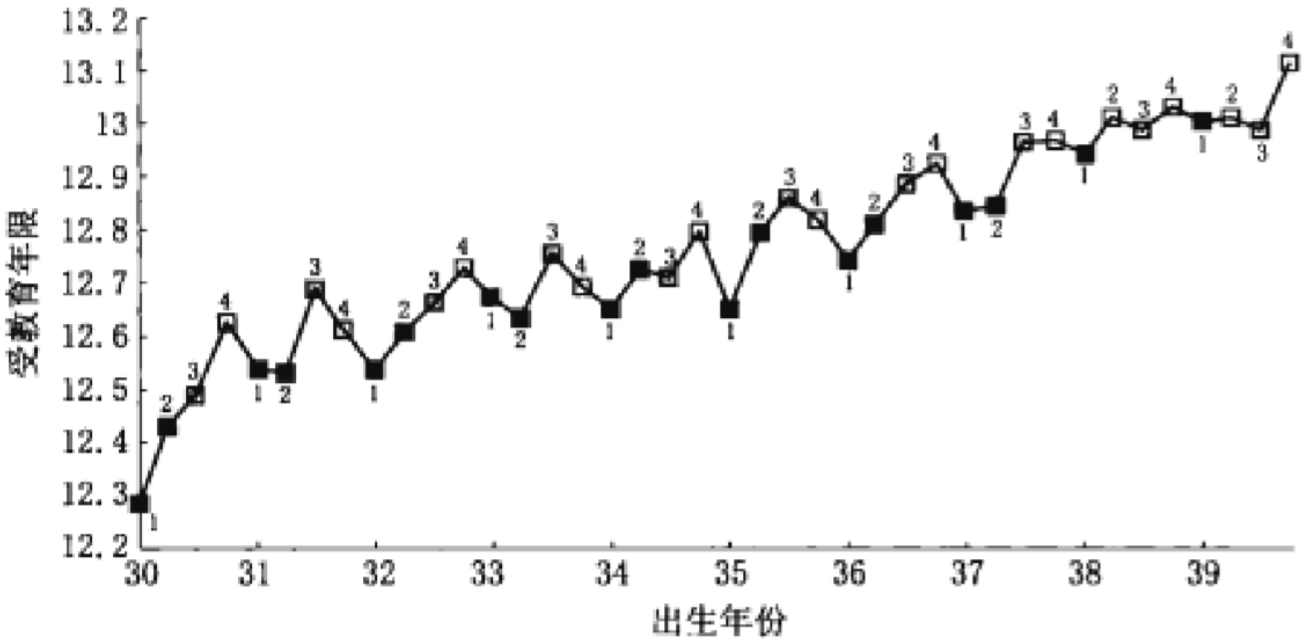
如何寻找工具变量?

- 与教育密切相关
- 与能力或其他未观测因素无关

一个例子

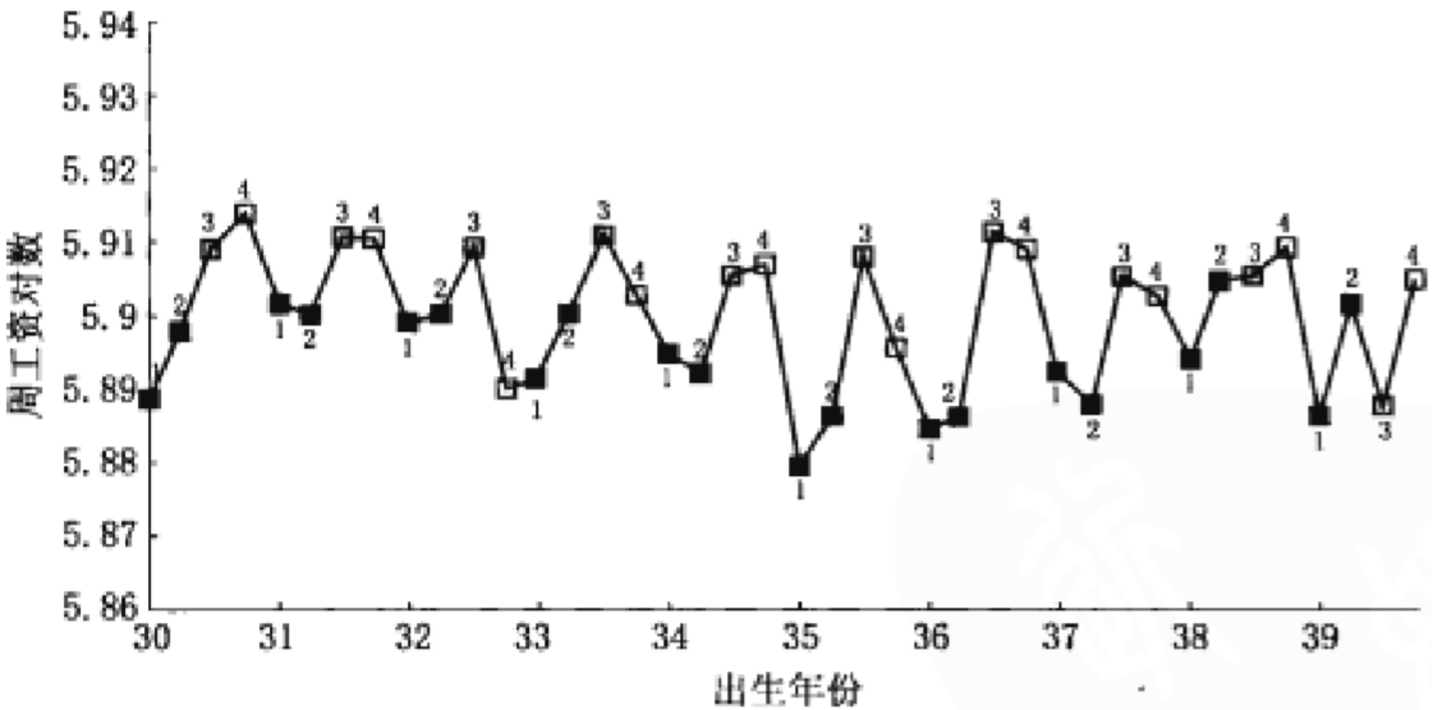
Angrist和Krueger选定出生季度作为工具变量

A. 按出生季度绘出的平均教育水平(第一阶段)



一个例子

B. 按出生季度绘出的平均周工资(简约式)



一个例子

- 第一阶段回归：

$$S_i = X_i'\pi_{10} + \pi_{11}Z_i + \xi_{1i}$$

- 简约式回归：

$$Y_i = X_i'\pi_{20} + \pi_{21}Z_i + \xi_{2i}$$

π_{11}—控制协变量X_i'后，工具变量Z_i对s_i的影响

π_{21}—控制协变量X_i'后，工具变量Z_i对Y_i的影响

$$\rho = \frac{\pi_{21}}{\pi_{11}} = \frac{cov(Y_i, \tilde{z}_i\ )}{cov(S_i, \tilde{z}_i)}$$

一个例子

- 存在协变量的因果模型：

$$Y_i = \alpha' X_i + \rho S_i + \eta_i$$

$\tilde{z}_i$与X_i不相关，$\tilde{z}_i$与η_i不相关

可证

$$cov(Y_i, \tilde{z}_i) = \rho cov(S_i, \tilde{z}_i)$$

4.1.1两阶段最小二乘回归

- 将$S_i = X_i'\pi_{10} + \pi_{11}Z_i + \xi_{1i}$代入因果模型得：

$$
\begin{aligned}
Y_i &= \alpha' X_i + \rho S_i + \eta_i \\
&= \alpha' X_i + \rho[X_i'\pi_{10} + \pi_{11}Z_i + \xi_{1i}] + \eta_i \\
&= X_i'[\alpha + \rho\pi_{10}] + \rho\pi_{11}Z_i + [\rho\xi_{1i} + \eta_i] \\
&= X_i'\pi_{20} + \pi_{21}Z_i + \xi_{2i}
\end{aligned}
$$

$\pi_{20} \equiv [\alpha + \rho\pi_{10}]$； $\pi_{21} \equiv \rho\pi_{11}$； $\xi_{2i} \equiv [\rho\xi_{1i} + \eta_i]$

$$Y_i = \alpha' X_i + \rho[X_i'\pi_{10} + \pi_{11}Z_i] + \xi_{2i}$$

- 可以直接用OLS回归得到ρ的一致估计

4.1.1两阶段最小二乘回归

- 第一步：估计第一阶段回归

$$\hat{s}_i = X_i'\hat{\pi}_{10} + \hat{\pi}_{11}Z_i$$

- 第二步：将第一阶段得到的拟合值代入结构式进行第二阶段回归

$$Y_i = \alpha'X_i + \rho\hat{s}_i + [\eta_i + \rho(s_i - \hat{s}_i)]$$

得到的估计值是ρ的一致估计

协变量和第一阶段拟合值既与η_i不相关，也与$(s_i - \hat{s}_i)$不相关

得到的标准误差是不正确的

4.1.1两阶段最小二乘回归

- 可以将两阶段回归估计量写为：

$$\hat{\rho} = \frac{Cov(Y_i,\ \hat{s}_i^*)}{Var(\hat{s}_i^*)}$$

其中$\hat{s}_i^*$是$\hat{s}_i$对X_i的回归残差

$$\begin{aligned}
Cov(s_i,\ \hat{s}_i^*) &= Cov(\hat{s}_i + (s_i - \hat{s}_i),\ \hat{s}_i^*) \\
&= Cov(\hat{s}_i,\ \hat{s}_i^*) + \underline{Cov(s_i - \hat{s}_i,\ \hat{s}_i^*)} \\
&= Cov(\hat{s}_i,\ \hat{s}_i^*) \\
&= Cov\left((\hat{s}_i - \hat{s}_i^*) + \hat{s}_i^*,\ \hat{s}_i^*\right) \\
&= \underline{Cov(\hat{s}_i - \hat{s}_i^*,\ \hat{s}_i^*)} + Var(\hat{s}_i^*) \\
&= Var(\hat{s}_i^*)
\end{aligned}$$

4.1.1两阶段最小二乘回归

- 将$Cov(s_i,\ \hat{s}_i^*) = Var(\hat{s}_i^*)$代入得

$$\hat{\rho} = \frac{Cov(Y_i,\ \hat{s}_i^*)}{Cov\left(s_i,\ \hat{s}_i^*\right)}$$

- 在只有单个内生变量和单个工具变量的模型中，2SLS估计值等于相应的间接最小二乘估计值，即工具变量的估计值

- 当出现多元工具变量时，2SLS和工具变量之间的关系还需进一步说明

  第一阶段回归变为：

$$s_i = X_i'\pi_{10} + \pi_{11}Z_{1i} + \pi_{12}Z_{2i} + \pi_{13}Z_{3i} + \xi_{1i}$$

一个例子

- 用出生季度作为工具变量进行2SLS回归求得的教育的经济回报见下表

	OLS		2SLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
受教育年数	0.071	0.067	0.102	0.13	0.104	0.108	0.087	0.057
	(0.000 4)	(0.000 4)	(0.024)	(0.020)	(0.026)	(0.020)	(0.016)	(0.029)
外生协变量								
年龄(按季度)								✓
年龄(按季度)的平方								✓
9 个出生年份虚拟变量		✓			✓	✓	✓	✓
50 个出生州的虚拟变量		✓			✓	✓	✓	✓
工具变量								
表示 QOB=1 的虚拟变量			✓	✓	✓	✓	✓	✓
表示 QOB=2 的虚拟变量				✓		✓	✓	✓
表示 QOB=3 的虚拟变量				✓		✓	✓	✓
出生季度虚拟变量和出生年份虚拟变量间的交互项(共 30 个工具变量)							✓	✓

工具变量和2SLS术语

- 被解释变量
- 内生变量
- 工具变量
- 外生协变量

4.1.2瓦尔德估计值

- 一个不含协变量的因果回归模型：

$$Y_i = \alpha + \rho s_i + \eta_i$$

- 假设工具变量Z_i为虚拟变量，$P(Z_i = 1) = p$，则

$$
\begin{aligned}
cov(Y_i, Z_i) &= E[Y_i Z_i] - E[Y_i]E[Z_i] \\
&= E\big[Z_i E[Y_i|Z_i]\big] - pE\big[E[Y_i|Z_i]\big] \\
&= pE[Y_i|Z_i = 1] - p(pE[Y_i|Z_i = 1] + \\
&\quad (1-p)E[Y_i|Z_i = 0]) \\
&= p(1-p)\{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]\}
\end{aligned}
$$

- 同理可得

$$cov(s_i, Z_i) = p(1-p)\{E[s_i|Z_i = 1] - E[s_i|Z_i = 0]\}$$

4.1.2瓦尔德估计值

- 瓦尔德估计值为：

$$\rho = \frac{cov(Y_i, Z_i)}{cov(s_i, Z_i)} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[s_i|Z_i = 1] - E[s_i|Z_i = 0]}$$

- 工具变量识别因果效应的核心：

被解释变量与工具变量之间的唯一联系是工具变量通过影响解释变量进而影响被解释变量

一个例子

- 用出生季度作为工具变量得到的教育回报的瓦尔德估计值和2SLS估计值

	(1) 在一年的第一个季度出生	(2) 在一年的第四个季度出生	(3) 差别 (标准误)
ln(周工资)	5.892	5.905	−0.013 5 (0.003 4)
受教育年限	12.688	12.839	−0.151 (0.016)
教育回报的瓦尔德估计值			0.089 (0.021)
教育回报的最小二乘估计值			0.07 (0.000 5)

一个例子

- 越战服役经历对退伍老兵收入的影响

  用参军资格作为工具变量

	收入		越战服役状态		
收入年	平均值 (1)	参军资格的影响 (2)	平均值 (3)	参军资格对服役概率的影响 (4)	服役对收入影响的瓦尔德估计值 (5)
1981	16 461	−435.8	0.267	0.159	−2 741
		(210.5)		(0.040)	(1.324)
1971	3 338	−325.9			−2 050
		(46.6)			(293)
1969	2 299	−2.0			
		(34.5)			

一个例子

- 家庭规模对母亲劳动力供给的影响
- 工具变量一：双胞胎；工具变量二：性别组成

			工具变量估计值			
			双胞胎		性别组成	
被解释变量	平均值 (1)	最小二乘法 (2)	第一阶段 (3)	瓦尔德估计值 (4)	第一阶段 (5)	瓦尔德估计值 (6)
就业率	0.528	−0.167	0.625	−0.083	0.067	−0.135
		(0.002)	(0.011)	(0.017)	(0.002)	(0.029)
工作周数	19.0	−8.05		−3.83		−6.23
		(0.09)		(0.76)		(1.29)
时间/周	16.7	−6.02		−3.39		−5.54
		(0.08)		(0.64)		(1.08)

4.1.3分组数据和两阶段最小二乘

- 分组数据(grouped data)将瓦尔德估计值和2SLS估计值联系在了一起：当用虚拟变量做工具变量时，2SLS估计值实际上等于对一系列分组数据的组内均值做广义最小二乘估计(GLS)

↓

- 广义最小二乘估计：瓦尔德估计值的线性组合
- 有些工具变量看上去是连续的，我们可以对其分组

一个例子

- R_i：用以确定参军资格的随机数
- $P[D_i = 1|R_i]$与R_i有关：

随机数R_i越小，获取参军资格的概率越大，志愿参军的可能性越大；随机数R_i越大，获取参军资格的概率越小，志愿参军的可能性越小

- 我们可以构造更多的组别进行比较，从而得到一组线性无关的瓦尔德估计值，每个瓦尔德估计值都一致地估计出了相同的因果效应

如何获得一个单一的估计值？

4.1.3分组数据和两阶段最小二乘

- 对用来构造瓦尔德估计值的组内均值拟合直线
- 我们使用二元常因果效应模型：

$$Y_i = \alpha + \rho D_i + \eta_i$$

其中，$\rho = Y_{1i} - Y_{0i}$是我们感兴趣的因果效应

因为$E[\eta_i|R_i] = 0$，$P[D_i = 1|R_i] = E[D_i|R_i]$，可得

$$E[Y_i|R_i] = \alpha + \rho P[D_i = 1|R_i]$$

我们可以用$E[Y_i|R_i]$和$P[D_i = 1|R_i]$的样本值来拟合直线

令$\bar{y}_i$：$E[Y_i|R_i = j]$，$\hat{p}_i$：$P[D_i = 1|R_i = j]$

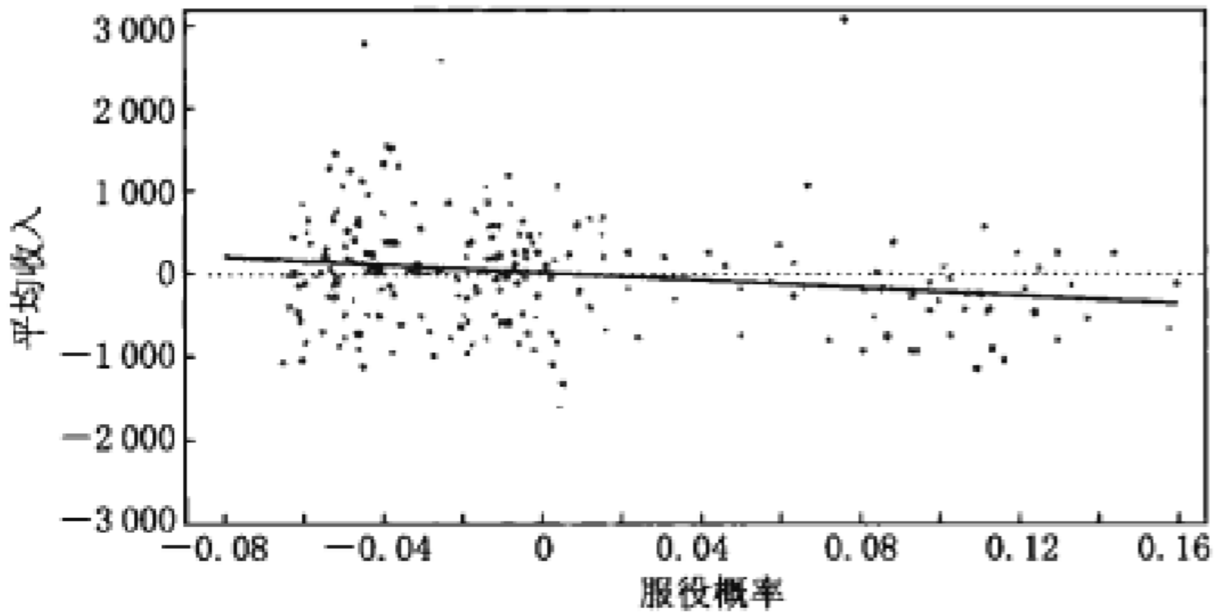
$$\bar{y}_i = \alpha + \rho \hat{p}_i + \bar{\eta}_i$$

4.1.3分组数据和两阶段最小二乘

- 分组方程往往是方差结构已知的异方差模型，可以使用加权最小二乘法计算参数，权重为$\bar{\eta}_i$的方差：
  假设分组内部的数据残差项是同方差的，记为${\sigma_\eta}^2$，那么$\bar{\eta}_i$的方差为$\frac{{\sigma_\eta}^2}{n_j}$，其中$n_j$是分组规模
- 运用广义最小二乘法从个分组数据中构造出的斜率是任何个线性独立的瓦尔德估计值的渐进有效线性组合
- 由于每个瓦尔德估计值都是工具变量估计值，所以广义最小二乘估计仍然是2SLS估计值
- 可以将任何使用一组虚拟变量做工具变量的2SLS估计值理解为相应瓦尔德估计值的线性组合

4.1.3分组数据和两阶段最小二乘

- 可视化工具变量(visual instrumental variables, VIV)



注：这是一个可视化工具变量的散点图，它以每五个数字为一个组，用1981—1984年的收入与参军的条件概率绘制散点图。样本包括了生于1950—1953年的白人男性。从这些点中得到的最小二回归估计值为−2 384，标准误为778。

4.2 2SLS的渐进推断

4.2.1两阶段最小二乘回归系数的渐进分布

- 令$V_i \equiv [X_i', \hat{s}_i]'$表示第二阶段回归中的回归元向量
- 2SLS估计值可以写为：

$$\hat{\Gamma}_{2SLS} \equiv \left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i Y_i$$

其中，$\Gamma \equiv [\alpha' \rho']$是相应的系数向量

$$\hat{\Gamma}_{2SLS} = \Gamma + [\textstyle\sum_i V_i V_i']^{-1} \sum_i V_i \, [\eta_i + \underline{\rho(s_i - \hat{s}_i)]}$$

$$= \Gamma + [\textstyle\sum_i V_i V_i']^{-1} \sum_i V_i \, \eta_i$$

- 斯拉茨基定理：将$[\sum_i V_i V_i']^{-1} \sum_i V_i \, \eta_i$中的样本拟合值替换为总体拟合值并不改变其极限分布 (用$[X_i' \pi_{10} + \pi_{11} Z_i]$代替$\hat{s}_i$)

4.2.1两阶段最小二乘回归系数的渐进分布

- $\hat{\Gamma}_{2SLS}$是渐进正态分布的，概率极限是Γ
- 统计量$[\sum_i V_i V_i']^{-1}[\sum_i V_i V_i' \eta_i{}^2][\sum_i V_i V_i']^{-1}$可一致地估计出协方差矩阵，与最小二乘估计的标准误类似，这个公式也是个三明治公式(sandwich formula)
- 给定协变量和工具变量，如果η_i是条件同方差的，那么对协方差矩阵的一致估计可简化为：

$$\left[\sum_i V_i V_i'\right]^{-1} \sigma_\eta^2$$

4.2.2过度识别与两阶段最小二乘的最小化元

- 过度识别：使用的工具变量个数大于内生变量的个数
- 记$Z_i \equiv \left[X_i' Z_{1i} \ldots Z_{Qi}\right]'$是由外生协变量和$Q$个工具变量构成的向量
- 记$W_i \equiv [X_i' s_i]'$是由外生协变量和单个内生变量组成的向量
- 因果模型(第二阶段)中的残差可以定义为：

$$\eta_i(\Gamma) \equiv Y_i - \Gamma' W_i = Y_i - [\alpha' X_i + \rho s_i]$$

4.2.2过度识别与两阶段最小二乘的最小化元

- 根据假设，$E[Z_i\eta_i(\Gamma)] = 0$，但这个方程不会严格成立，因为据条件的个数大于Γ中待估计参数个数。故其样本估计值为：

$$\frac{1}{N}\sum Z_i\eta_i(\Gamma) \equiv m_N(\Gamma)$$

- 2SLS实际上是一种广义矩估计值(GMM)，即求解使上述等式所表示的样本矩最接近0的参数向量
- 由中心极限定理，样本矩向量$\sqrt{N}m_N(\Gamma)$的渐进协方差矩阵为$E[Z_iZ_i'\eta_i(\Gamma)^2]$，我们将其称为$\Lambda$
- 我们将$m_N(\Gamma)$中的Γ替换为$\hat{g}$，表示Γ的估计值

4.2.2过度识别与两阶段最小二乘的最小化元

- $E[Z_i\eta_i(\Gamma)] = 0$的最优广义矩估计值应该可以最小化样本矩向量$m_N(\hat{g})$的二次型
- 在构造$m_N(\hat{g})$的二次型时，最优权重矩阵就是Λ^{-1}（实际中用Λ的一致估计来替代Λ）
- 我们要最小化的二次型为：

$$J_N(\hat{g}) \equiv N m_N(\hat{g})\Lambda^{-1} m_N(\hat{g})$$

- 当残差为条件同方差时，通过最小化$J_N(\hat{g})$得到的估计值正好是2SLS估计值；当条件同方差不成立时，通过最小化$J_N(\hat{g})$得到的估计值就是White得到的两阶段工具变量(2SLS的一般化)估计值
- $J_N(\hat{g})$：2SLS的最小化元

4.2.2过度识别与两阶段最小二乘的最小化元

- 因为η_i是条件同方差的，所以

$$\Lambda = E[Z_i Z_i' \eta_i(\Gamma)^2] = E[Z_i Z_i']\sigma_\eta^2$$

用它来替代Λ^{-1}并用y，Z和W表示相应向量和矩阵的样本值，于是需要最小化的二次型变为：

$$J_N(\hat{g}) = \frac{1}{N\sigma_\eta^2}(y - W\hat{g})' Z E[Z_i Z_i']^{-1} Z'(y - W\hat{g})$$

用样本的叉乘矩阵$\left[\frac{Z'Z}{N}\right]$代替$E[Z_i Z_i']$，可得：

$$J_N(\hat{g}) = \frac{1}{\sigma_\eta^2}(y - W\hat{g})' P_Z (y - W\hat{g})$$

其中，$P_Z = Z(Z'Z)^{-1}Z$。由此得到二次型的解：

$$\hat{g} = \hat{\Gamma}_{2SLS} = [W' P_Z W]^{-1} W' P_Z y$$

4.2.2过度识别与两阶段最小二乘的最小化元

- 通过求解2SLS的最小化元，我们也可得到检验过度识别问题的统计量
- 由于假设残差和工具变量不相关，故

$$J_N(\hat{g}) \sim \chi^2(Q-1)$$

- 通过比较2SLS最小化元的实际值和卡方分布的标准值，我们可以对原假设$H_0\colon E[Z_i\eta_i(\Gamma)]=0$做统计检验

4.2.2过度识别与两阶段最小二乘的最小化元

- 考虑一组互斥的虚拟变量做工具变量的情况
- 假设一组互斥的虚拟变量可取J个值，产生$N \times 1$维的拟合向量，其中每个工具变量都表示一个分组数据并对应于该组数据中拟合出的条件均值
- 易知拟合值向量中每个分组数据的拟合值都会出现n_j次，并有$\sum n_j = N$，工具变量的叉乘矩阵$[Z'Z]$是个$J \times J$对角阵，对角线元素为n_j，故

$$\hat{J}_N(\hat{g}) = \frac{1}{\sigma_\eta^2} \sum_j n_j \left(\bar{y}_i - \hat{g}'\bar{W}_j\right)^2$$

$\bar{W}_j$：矩阵W中处于第j行的样本均值

4.2.2过度识别与两阶段最小二乘的最小化元

- 无需同方差假设时有效的两阶段工具变量过程的最小化元：

$$\hat{J}_N(\hat{g}) = \sum_j \left(\frac{n_j}{\sigma_j^2}\right)\left(\bar{y}_i - \hat{g}'\overline{W}_j\right)^2$$

其中，σ_j^2是第j组中n_j的方差

- 2SLS估计值虽然不是有效的，但却还是一致的
- 过度识别统计量实际上度量了将$\bar{y}_i$和$\overline{W}_j$连接起来的直线的拟合优度
- 该统计量也是可视化工具变量回归散点图中对回归直线拟合优度所做的卡方检验，自由度等于工具变量和待估参数个数之差

4.2.2过度识别与两阶段最小二乘的最小化元

得到过度识别检验统计量的两种方法：

- 在同方差模型中，2SLS最小化元等于样本大小乘R^2，其中R^2来自用2SLS的残差关于工具变量所进行的回归

$$N\left[\frac{\hat{\eta}'P_z\hat{\eta}}{\hat{\eta}'\hat{\eta}}\right]$$

其中，$\hat{\eta}=Y-W\,\hat{\Gamma}_{2SLS}$，它是2SLS的残差向量

- “不同方法，同一结果”：

构造多个恰好识别的工具变量并用这些工具变量对相同的因果关系进行识别，然后在不同结果之间进行比较。如果我们把“所有可能的恰好识别估计值都相等”看作原假设，那么对该原假设进行的检验实际上就是在构造瓦尔德检验。

4.2.2过度识别与两阶段最小二乘的最小化元

实际中使用过度识别检验需要注意的问题：

- $\hat{J}_N(\hat{g})$度量了经过方差调整的拟合优度，过度识别统计量倾向于偏低
- 由于工具变量估计值往往不精确，即使单个估计量显示出足够的精确性，我们也不能因为某个工具变量估计值和另一个工具变量估计值相差不大而感到欣喜
- 当工具变量估计值相当精确时，过度识别检验拒绝原假设也并不意味着我们的识别策略错了，这可能是因果效应的异质性造成的