工具变量与广义矩估计

司继春

上海对外经贸大学统计与信息学院

1 内生性问题

在上一节中，我们在外生性（exogeneity）假设 $\mathbb{E}(u_i|x_i)=0$ 下得到了线性回归的最小二乘估计。然而在现实中，出于种种原因，外生性假设并不容易满足，即误差项 u_i 与解释变量 x_i 之间存了某种相关性，即 $\mathrm{Cov}(x_i u_i)\neq 0$，导致最小二乘估计量不再一致，我们称这种情况为**内生性**（**endogeneity**）问题。因而在使用线性回归时，外生性假设是最重要的假设，如果外生性不满足，那么会导致我们以解释为目的的线性回归最终得到错误的结论。现实中，许多原因都可能导致内生性问题，比如遗漏变量、度量误差、反向因果、样本选择、自选择等等，下面我们就讨论几个常见的可能导致内生性的原因。

1.1 遗漏变量

遗漏变量是非常常见的导致内生性问题的原因。如果我们关心 x_i 对 y_i 的影响，并使用如下回归方程进行建模：

$$y_i = x_i'\beta + u_i$$

那么我们必须要保证 x_i 中包含了所有影响 y_i 同时又与 x_i 潜在可能相关的因素，如果某一个变量 q_i 即对 y_i 有影响，同时与 x_i 相关，然而我们在回归方程中并没有包含 q_i，那么就会导致内生性问题，进而导致最小二乘结果失效。

一个典型的例子是教育的回报问题。如果我们关心教育 edu_i 对收入 $income_i$ 的因果效应，那么我们可以设定如下回归模型：

$$income_i = \gamma \cdot edu_i + x_i'\beta + u_i$$

其中 x_i 为控制变量。然而实际上，可能有其他不可观测的因素，比如能力 $(ability_i)$ 即影响了个人的收入，又影响了个人对于教育程度 edu_i 的决策，真实的数据生成过程可能为：

$$income_i = \gamma \cdot edu_i + x_i'\beta + \gamma \cdot ability_i + v_i$$

因而误差项 $u_i = \gamma \cdot ability_i + v_i$。如果 $\mathrm{Cov}\left(ability_i, edu_i\right) \neq 0$，且 $\gamma \neq 0$，那么遗漏的变量 $ability_i$ 就会导致线性回归的结果不一致。

更一般的，如果真实的数据生成过程为：

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \gamma q_i + v_i$$

而变量 q_i 是观测不到的，并且假设 q_i 与 x_i 之间存在着相关性：

$$q_i = \delta_0 + \delta_1 x_{1i} + \cdots + \delta_K x_{Ki} + e_i$$

那么将以上方程带入结构式，得到：

$$y_i = \left(\beta_0 + \gamma\delta_0\right) + \left(\beta_1 + \gamma\delta_1\right) x_{1i} + \cdots + \left(\beta_K + \gamma\delta_K\right) x_{Ki} + \gamma e_i + v_i$$

如果令 $u_i = \gamma e_i + v_i$，我们有 $\mathbb{E}\left(u_i|x_i\right) = 0$，因而那么如果我们忽略了变量 q_i，那么实际得到的回归系数为 $\beta_k^* = \beta_k + \gamma\delta_k$，存在着偏误。

在以上教育的例子中，q_i 为 $ability_i$，如果我们认为具有更高能力的个人更容易上大学，即 $\delta_{edu} > 0$，且能力 $ability_i$ 对收入有正向影响，即 $\gamma > 0$，那么我们使用最小二乘法得到的 edu_i 对 $income_i$ 的影响就被高估了。

1.2 度量误差

在线性回归中，我们假设我们所观察到的所有变量都是准确的，然而现实中，我们所观察到的 x_i 可能会出于种种原因出现度量误差（measurement error）。在度量误差存在的情况下，也会导致内生性问题。

为了简单起见，我们考虑一个一元线性回归，假设数据的真实生成过程为：

$$y_i = \beta_0 + \beta x_i^* + v_i$$

其中 x_i^* 为真实值。然而现实中，我们可能观察不到 x_i^*，只能观察到有误差的 $x_i = x_i^* + e_i$。如果我们直接用 y_i 对 x_i 做回归，即：

$$y_i = \beta_0^* + \beta^* x_i + u_i$$

那么 $u_i = v_i - \beta e_i$。如果假设 $\mathbb{E}\left(e_i|x_i^*\right) = 0$，那么

$$\mathrm{Cov}\left(x_i, e_i\right) = \mathrm{Cov}\left(x_i^* + e_i, e_i\right) = \mathrm{Var}\left(e_i\right)$$

因而 $\mathrm{Cov}\left(x_i, u_i\right) = \mathrm{Cov}\left(x_i, v_i - \beta e_i\right) = -\beta \mathrm{Var}\left(e_i\right) \neq 0$，因而导致了内生性问题。

如果我们直接使用带有度量误差的 x_i 进行回归，那么我们将得到：

$$
\begin{aligned}
\text{plim}\hat{\beta} &= \frac{\text{Cov}\left(x_i, y_i\right)}{\text{Var}\left(x_i\right)} \\
&= \frac{\text{Cov}\left(x_i^* + e_i, \beta_0 + \beta x_i^* + v_i\right)}{\text{Var}\left(x_i^*\right) + \text{Var}\left(e_i\right)} \\
&= \beta \cdot \frac{\text{Var}\left(x_i^*\right)}{\text{Var}\left(x_i^*\right) + \text{Var}\left(e_i\right)}
\end{aligned}
$$

得到的 $\left|\hat{\beta}\right| < |\beta|$，即估计的系数的绝对值总是小于真实值的绝对值，存在着向中性偏误（attenuation bias）。

1.3 反向因果

在线性回归中，我们希望使用 x_i 解释 y_i，希望得到 x_i 对 y_i 的因果效应。然而经济变量中，很多时候存在着互为因果的情况，即不仅仅 x_i 对 y_i 有因果效应，同时反过来，y_i 对 x_i 也有因果效应。这种现象在经济学中非常普遍，比如金融制度对经济增长有影响，反过来经济增长也会导致金融的发展。由于反向因果的存在，使得我们很难区分哪些是单纯的制度对经济增长的影响，哪些是经济增长对制度的影响。

如果我们考虑两个相互影响的变量 y_{1i} 和 y_{2i}，其结构方程为：

$$y_{1i} = \alpha y_{2i} + x_i'\delta + u_i$$

$$y_{2i} = \gamma y_{1i} + w_i'\beta + v_i$$

即 y_2 对 y_1 有因果效应，同时 y_1 对 y_2 也有因果效应。如果我们联立以上方程，可以解得：

$$y_{1i} = \frac{\alpha w_i'\beta + x_i'\delta + u_i + \alpha v_i}{1 - \alpha\gamma}$$

$$y_{2i} = \frac{\gamma x_i'\delta + w_i'\beta + v_i + \gamma u_i}{1 - \alpha\gamma}$$

注意到在上式中，有 $\text{Cov}\left(y_{2i}, u_i\right) = \frac{\gamma}{1-\alpha\gamma}\text{Var}\left(u_i\right) \neq 0$，同理 $\text{Cov}\left(y_{1i}, v_i\right) \neq 0$，因而如果我们直接做 y_{1i} 对 y_{2i} 的回归，或者相反，都会导致内生性问题。

2 工具变量

当外生性假设不满足时，或者模型中存在内生性问题时，会导致我们线性回归的估计量不一致，得到错误的结果。一般而言，内生性问题是非常普遍而且非常难以解决的。尽管如此，在一些特殊情况下，我们还是可以通过一些计量方法得到因果效应，这其中最为常用的是**工具变量**（**instrumental variables**）方法。

简单而言，工具变量法即找到对 y_i 没有直接影响，但是与内生变量 x_i 高度相关的变量 z_i，通过 z_i 的外生变动得到 x_i 对 y_i 的因果效应。

为了方便说明，我们首先考虑一个最简单的例子。如果我们希望估计某农产品的需求曲线，假设 y_i 为成交量，x_i 为农产品的价格。假设农产品的需求曲线为：

$$y_i^d = \beta_0 + \beta x_i + u_i$$

供给曲线为：

$$y_i^s = \delta_0 + \delta x_i + v_i$$

均衡的成交量和价格应该使得供给需求相等，即：

$$\beta_0 + \beta x_i + u_i = \delta_0 + \delta x_i + v_i$$

解得均衡的价格为：

$$x_i = \frac{\delta_0 - \beta_0 + v_i - u_i}{\beta - \delta}$$

注意到 $\mathrm{Cov}\left(x_i, u_i\right) \neq 0$，因而如果我们使用 y_i 对 x_i 做回归，并不能得到 β 的一致估计。

现在假设该农产品的供给受到天气的影响，同时天气并不影响该农产品的需求。假设天气变量为 z_i，我们修改以上的供给曲线为：

$$y_i^s = \delta_0 + \delta x_i + \delta_1 z_i + v_i$$

那么均衡价格为：

$$x_i = \frac{\delta_1 z_i + \delta_0 - \beta_0 + v_i - u_i}{\beta - \delta} \triangleq \gamma_0 + \gamma z_i + \epsilon_i$$

其中 $\gamma_0 = \frac{\delta_0 - \beta_0}{\beta - \delta}$，$\gamma = \frac{\delta_1}{\beta - \delta}$，$\epsilon_i = \frac{v_i - u_i}{\beta - \delta}$。如此我们得到了价格 x_i 随着天气 z_i 变动的一个相关关系，即由于天气的外生变化导致价格变化的关系。由于 z_i 外生的影响产品的供给，因而我们可以假设 $\mathbb{E}\left(u_i | z_i\right) = \mathbb{E}\left(v_i | z_i\right) = 0$，那么 $\mathbb{E}\left(\epsilon_i | z_i\right) = 0$，因而我们可以直接使用最小二乘回归得到 γ_0 和 γ 的估计 $\hat{\gamma}_0$ 和 $\hat{\gamma}$。我们称 z_i 为工具变量，即与误差项不相关，但是与我们的内生变量 x_i 高度相关。

现在将以上 x_i 与 z_i 关系带入到需求曲线中，得到：

$$\begin{aligned} y_i^d &= \beta_0 + \beta x_i + u_i \\ &= \beta_0 + \beta \gamma_0 + \beta \gamma z_i + \beta \epsilon_i + u_i \\ &\triangleq \eta_0 + \eta z_i + e_i \end{aligned}$$

其中 $\eta_0 = \beta_0 + \beta \gamma_0$，$\eta = \beta \gamma$，$e_i = \beta \epsilon_i + u_i$。如此我们得到了因为天气的外生变

动导致的成交量的变化。注意由于天气的变动只对该农产品的供给有影响，而对需求没有直接影响，因而天气变动对需求的影响只通过价格来影响。以上因变量对

注意由于 $\mathbb{E}(u_i|z_i)=\mathbb{E}(v_i|z_i)=0$，因而 $\mathbb{E}(e_i|z_i)=0$，因而我们仍然可以使用最小二乘法得到 η_0 和 η 的估计，$\hat{\eta}_0$ 和 $\hat{\eta}$。而由于我们希望得到的结构参数 $\beta=\frac{\eta}{\gamma}$，而我们已经得到了 η 和 γ 的估计值，因而我们可以得到 β 的估计值：

$$\hat{\beta}=\frac{\hat{\eta}}{\hat{\gamma}}$$

如此，我们就得到了该农产品需求参数 β 的识别。

一般地，如果我们关心结构方程：

$$y_i=\beta_0+\beta x_i+u_i$$

的识别，其中 $\mathrm{Cov}(x_i,u_i)\neq 0$，即 x_i 为内生变量。如果我们可以找到一个变量 z_i，z_i 不会直接影响 y_i，$\mathbb{E}(u_i|z_i)=0$，同时 z_i 与 x_i 高度相关，那么我们称 z_i 为内生变量 x_i 的工具变量。其中内生变量 x_i 与 z_i 之间的相关性为：

$$x_i=\gamma_0+\gamma z_i+\epsilon_i$$

z_i 与 x_i 高度相关意味着 $\gamma\neq 0$。注意上述方程仅仅代表了 z_i 与 x_i 之间的相关性，不是结构方程，因而 $\mathbb{E}(\epsilon i|z_i)=0$。将上式带入结构方程，得到：

$$y_i=\beta_0+\beta\gamma_0+\beta\gamma z_i+u_i+\beta\epsilon_i\overset{\Delta}{=}\eta_0+\eta z_i+e_i$$

我们称以上方程为简约式（reduced-form）。我们可以使用最小二乘回归分别得到 γ 和 η 的估计，由于 $\beta=\frac{\eta}{\gamma}$，因而可以得到 β 的估计：

$$\hat{\beta}=\frac{\hat{\eta}}{\hat{\gamma}}$$

以上估计量我们称之为 Wald 估计量。

以上我们讨论了一个内生变量、一个工具变量的情形。实际上，我们可以将上述进行推广。如果我们关心结构方程：

$$y_i=w_i'\gamma+z_{1i}'\delta+u_i=x_i'\beta+u_i$$

其中 w_i 为 $G\times 1$ 维的内生变量，即 $\mathbb{E}(u_i|w_i)\neq 0$，而 z_{1i} 为外生的对 y_i 有影响的变量，$\mathbb{E}(u_i|z_{1i})=0$。记 $x_i=(w_i',z_{1i}')'$ 为 $K\times 1$ 维的结构方程的解释变量，β 为结构参数。

另外，假设存在着 G 个工具变量 z_{2i}，满足 $\mathbb{E}(u_i|z_{2i})=0$。记 $z_i=(z_{1i}',z_{2i}')'$ 为所有所有的外生变量，那么我们有 $\mathbb{E}(u_i|z_i)=0$，因而有 $\mathbb{E}(z_iu_i)=0$，因而我

们可以使用矩估计对 β 进行估计。由于 $u_i = y_i - x_i'\beta$，因而 $\mathbb{E}\left(z_i\left(y_i - x_i'\beta\right)\right) = 0$，如果 $\mathbb{E}\left(z_i x_i'\right)$ 可逆，那么：

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

因而我们可以使用样本矩：

$$\frac{1}{N}\sum_{i=1}^{N} z_i x_i' = Z'X \text{和} \frac{1}{N}\sum_{i=1}^{N} z_i y_i = Z'Y$$

替代总体矩，得到：

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

其中：

$$Z = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_N' \end{bmatrix}$$

这里需要注意的是，以上矩估计要求 $\mathbb{E}\left(z_i x_i'\right)$ 可逆，实际上要求每个内生变量 w_i 都要找到与之高度相关的工具变量。如果该条件不满足，那么矩阵 $\mathbb{E}\left(z_i x_i'\right)$ 不可逆，结构参数 β 是无法被识别的。

以上讨论了当有 G 个内生变量，且刚好有 G 个工具变量的情形。然而实际上，对于 G 个内生变量，我们可以使用 $L_1 > G$ 个工具变量。记 $L = L_1 + K - G$，即所有外生变量的个数。由于 $L_1 > G$，因而 $L > K$，意味着当我们有超过 G 个工具变量时，以上的矩估计中我们有 L 个矩条件或者方程，K 个参数，方程无解。此时，我们必须对矩估计进行推广，即所谓的广义矩估计。

3 广义矩估计

在此之前我们介绍了矩估计的思想，即使用样本矩代替总体矩进行估计。在矩估计中，如果我们对参数 θ 感兴趣，θ 为 $K \times 1$ 维向量，那么只要我们能够找到足够多的矩条件 $m_i\left(x_i, \theta\right)$，使得：

$$\begin{cases} \mathbb{E}\left[m_1\left(x_i, \theta\right)\right] = 0 \\ \vdots \\ \mathbb{E}\left[m_K\left(x_i, \theta\right)\right] = 0 \end{cases}$$

具有唯一解，那么我们就可以使用其样本的等价形式：

$$\begin{cases} \sum_{i=1}^{N} m_1\left(x_i, \hat{\theta}\right) = 0 \\ \vdots \\ \sum_{i=1}^{N} m_K\left(x_i, \hat{\theta}\right) = 0 \end{cases}$$

对未知参数 θ 进行估计。

以上对于 K 个参数 θ，我们使用了 K 个矩条件，而且必须使用 K 个矩条件。如果矩条件个数少于 K 个，那么方程个数少于参数个数，意味着我们得不到唯一解；如果矩条件个数多于 K 个，那么方程个数大于参数个数，意味着方程很有可能无解。

然而在应用中，我们经常有多于 K 个矩条件可以使用，更多的矩条件为参数 θ 的估计带来了更多的信息，因而有可能提高对 θ 估计的精度，那么我们是否可以使用多于 K 个矩条件呢?

例如，在正态分布参数估计的例子中，如果 $x_i \sim N\left(\mu, \sigma^2\right)$，我们前面使用了前两阶矩：

$$\begin{cases} \mathbb{E}\left(x_i\right) - \mu = 0 \\ \mathbb{E}\left(x_i^2\right) - \mu^2 - \sigma^2 = 0 \end{cases}$$

进行估计，在此情况下，我们有两个未知数两个方程，可以解得 μ 和 σ^2。然而我们是不是也可以使用其三阶矩：$\mathbb{E}\left(x_i^3\right) - \mu^3 - 3\mu\sigma^2 = 0$ 进行估计呢? 如此我们得到了三个矩条件：

$$\begin{cases} \mathbb{E}\left(x_i\right) - \mu = 0 \\ \mathbb{E}\left(x_i^2\right) - \mu^2 - \sigma^2 = 0 \\ \mathbb{E}\left(x_i^3\right) - \mu^3 - 3\mu\sigma^2 = 0 \end{cases}$$

其样本等价形式为：

$$\begin{cases} m_1\left(x, \theta\right) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i\right) - \mu = 0 \\ m_2\left(x, \theta\right) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i^2\right) - \mu^2 - \sigma^2 = 0 \\ m_3\left(x, \theta\right) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i^3\right) - \mu^3 - 3\mu\sigma^2 = 0 \end{cases}$$

然而在此情况下，三个方程两个未知数导致方程无解。

一个简单的想法是，既然无法保证每个矩条件都等于 0，那么我们就尽量的让三个矩条件都尽量靠近 0。一个解决方法是，可以直接解最小化三个矩条件的平方和：

$$\hat{\theta} = \arg\min_{\theta}\left[m_1^2\left(x, \theta\right) + m_2^2\left(x, \theta\right) + m_3^2\left(x, \theta\right)\right]$$

如此，尽管我们不能保证每个矩条件都等于 0，但是我们可以保证每个矩条件都

足够贴近于 0。

以上想法可以继续推广，如果记：

$$m(x,\theta)=\begin{bmatrix} m_1(x,\theta) \\ m_2(x,\theta) \\ m_3(x,\theta) \end{bmatrix}$$

那么上述最小化问题可以写为

$$\min_{\theta} m(x,\theta)' m(x,\theta)$$

更进一步，我们可以使用任意一个正定矩阵 W，解最小化问题：

$$\min_{\theta} m(x,\theta)' W m(x,\theta)$$

也可以保证每个矩条件都足够贴近于 0。以上就是**广义矩估计**（**Generalized method of moments, GMM**）的思想。

一般的，如果对于 $K\times 1$ 维参数 θ，我们有 L 个矩条件：$\mathbb{E}[m(x_i,\theta)]=0$，其中 $m(x_i,\theta)$ 为 $L\times 1$ 的向量函数，其中 $L\geq K$，那么广义矩估计即解如下问题：

$$\hat{\theta}=\arg\min_{\theta}\left[\sum_{i=1}^{N} m(x_i,\theta)\right]' \hat{W}\left[\sum_{i=1}^{N} m(x_i,\theta)\right]$$

其中 $\hat{W}$ 为 $L\times L$ 的实对称正定矩阵，我们称其为加权矩阵（weighting matrix）。

实际上，可以证明，在一定的条件下，只要满足：

1. $m(x_i,\theta)$ 为 θ 的连续函数
2. $\hat{W}\xrightarrow{p} W_0$
3. $\mathbb{E}[m(x_i,\theta)]=0$ 有唯一解 θ_0，即真值

那么广义矩估计一定是真值的一致估计，即 $\hat{\theta}\xrightarrow{p}\theta_0$。

此外，我们还可以得到广义矩估计的大样本分布。在一定条件下，如果满足：

1. $m(x_i,\theta)$ 为 θ 的连续可微函数
2. $m(x_i,\theta)$ 的每个分量都有有限的二阶矩
3. 记 $G_0=\mathbb{E}\left[\frac{\partial m(x_i,\theta_0)}{\partial\theta}\right]$，$\operatorname{rank}(G_0)=K$

那么广义矩估计量 $\hat{\theta}$ 的大样本分布为：

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)\overset{a}{\sim} N\left(0, A_0^{-1}B_0A_0^{-1}\right)$$

其中 $A_0 = G_0' W_0 G_0$，$B_0 = G_0' W_0 \Lambda_0 W_0 G_0$，$\Lambda_0 = \mathbb{E}\left[m(x_i,\theta_0)\, m(x_i,\theta_0)'\right]$。

以上我们给出了广义矩估计的大样本性质。然而注意到，虽然目前我们对加权矩阵 $\hat{W}$ 的要求仅仅为 $L \times L$ 的实对称正定矩阵，但是给定不同的 $\hat{W}$，所得到的广义矩估计量的方差是不同的。可以证明，当加权矩阵 $W = \Lambda_0^{-1}$ 时，广义矩估计量可以达到最小的方差，我们称此加权矩阵为最优加权矩阵（optimal weighting matrix）。当使用了最优加权矩阵时，广义矩估计量的大样本方差为 $A_0^{-1} = (G_0' W_0 G_0)^{-1}$。

在实践中，由于最优加权矩阵 $\Lambda_0^{-1} = \left(\mathbb{E}\left[m(x_i,\theta_0)\, m(x_i,\theta_0)'\right]\right)^{-1}$ 的估计依赖于 θ，因而在得到 θ 的估计之前，我们无法计算最优加权矩阵。实际上，只要我们给定任意的实对称正定矩阵 W，都可以得到一致估计。因而实践中，可以先使用任意的加权矩阵（比如单位阵）带入广义矩估计的目标函数中进行计算，得到一个估计 $\hat{\theta}^0$，然后使用该估计，计算权重矩阵：

$$\hat{W} = \hat{\Lambda}^{-1} = \left(\frac{1}{N}\sum_{i=1}^{N}\left[m\left(x_i,\hat{\theta}^0\right) m\left(x_i,\hat{\theta}^0\right)'\right]\right)^{-1}$$

进而使用该加权矩阵继续带入目标函数中，得到新的估计 $\hat{\theta}^1$。以上过程可以不断迭代，直至收敛。

以上介绍了广义矩估计的一般思路和结论。广义矩估计使得我们只要得到矩条件，可以很方便的将其带入到该框架中，得到一些列的推断结果。然而实际中，有时尽管我们可以得到很多矩条件，但是我们并不能保证所有矩条件都是正确无误的。比如在以上正态分布的例子中，如果 x_i 的确服从正态分布，那么三个矩条件必然都成立。但是如果我们的假设错误，x_i 不服从正态分布，那么三个矩条件就是错的。那么我们有没有办法检验矩条件是不是成立呢？当矩条件个数 $L > K$ 的时候，我们可以一定程度上回答这一问题。

可以证明，如果我们在计算广义矩估计时使用了最优加权矩阵，那么其目标函数渐进服从 χ^2 分布：

$$\frac{1}{N}\left[\sum_{i=1}^{N} m(x_i,\theta)\right]' \hat{\Lambda}^{-1}\left[\sum_{i=1}^{N} m(x_i,\theta)\right] \overset{a}{\sim} \chi^2(L-K)$$

因而在原假设：

$$H_0 : \mathbb{E}\left[m(x_i,\theta)\right] = 0$$

的条件下，当样本充分大时，以上目标函数应该足够贴近于 0。如果目标函数值过大，大于 χ^2 分布的临界值点，那么我们就有理由拒绝原假设，认为矩条件不成立。以上检验称为 Hansen 检验。

4 两阶段最小二乘

4.1 2SLS 的估计

在有了广义矩估计这一工具之后，我们就可以使用广义矩估计来解决工具变量个数大于内生变量个数时矩估计无解的问题了。

类似以上的设定，如果我们关心结构方程：

$$y_i = w_i'\gamma + z_{1i}'\delta + u_i = x_i'\beta + u_i$$

其中 w_i 为 $G \times 1$ 维的内生变量，即 $\mathbb{E}(u_i|w_i) \neq 0$，而 z_{1i} 为外生的对 y_i 有影响的变量，$\mathbb{E}(u_i|z_{1i}) = 0$。记 $x_i = (w_i', z_{1i}')'$ 为 $K \times 1$ 维的结构方程的解释变量，β 为结构参数。

另外，假设存在着 $L_1 > G$ 个工具变量 z_{2i}，满足 $\mathbb{E}(u_i|z_{2i}) = 0$。记 $z_i = (z_{1i}', z_{2i}')'$ 为 $L \times 1$ 维向量，包含了所有所有的外生变量，那么我们有 $\mathbb{E}(u_i|z_i) = 0$，因而有 $\mathbb{E}(z_i u_i) = 0$，以上即我们的矩条件。

使用广义矩估计的思路，我们可以通过最小化：

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

获得该问题的估计。注意到，我们的矩条件为 $\mathbb{E}(z_i u_i) = 0$，因而最优加权矩阵 $W_0 = \mathbb{E}\left(u_i^2 z_i z_i'\right)$，在同方差的假定下，$\mathbb{E}\left(u_i^2 z_i z_i'\right) = \sigma^2 \mathbb{E}(z_i z_i')$。由于 σ^2 为常数，因而 $\mathbb{E}(z_i z_i')$ 即为最优加权矩阵，我们可以使用 $\sum_{i=1}^{N} z_i z_i' = Z'Z$ 对以上最优加权矩阵进行估计。

注意到由于 $\sum_{i=1}^{N} z_i y_i = Z'Y$，$\sum_{i=1}^{N} z_i x_i' = Z'X$，带入最优加权矩阵，以上目标函数即：

$$\min_{\beta} [Z'Y - Z'X\beta]' (Z'Z)^{-1} [Z'Y - Z'X\beta]$$

对以上目标函数求一阶条件，得到：

$$X'Z (Z'Z)^{-1} Z'X\hat{\beta} = X'Z (Z'Z)^{-1} Z'Y$$

从而：

$$\hat{\beta} = \left(X'Z (Z'Z)^{-1} Z'X\right)^{-1} X'Z (Z'Z)^{-1} Z'Y$$

以上就是工具变量的广义矩估计。

注意到，令 $P_Z = Z (Z'Z)^{-1} Z'$，那么 P_Z 实际为幂等矩阵，以上估计量可以写为：

$$\hat{\beta} = \left[(P_Z X)' (P_Z X)\right]^{-1} \left[(P_Z X)' Y\right]$$

其中 $P_Z X$ 即使用 X 对 Z 回归得到的预测值。记 $\hat{X}=P_Z X$，以上的估计量即

$$\hat{\beta}=\left(\hat{X}'\hat{X}\right)^{-1}\hat{X}'Y$$

因而上述广义矩估计等价于在得到 $\hat{X}$ 之后，使用 Y 对 $\hat{X}$ 做回归。以上步骤可以整理为:

1. 首先使用 X 对 Z 做回归，得到 $\hat{X}$。由于 $x_i=(w_i',z_{1i}')'$，$z_i=(z_{1i}',z_{2i}')'$，因而以上回归实际上只需要用内生变量 w_i 对所有外生变量 z_i（包括 z_{1i}）做回归，得到 $\hat{w}_i$。由于 z_{1i} 包含在 z_i 中，因而 $\hat{z}_{1i}=z_{1i}$。记 $\hat{x}_i=(\hat{w}_i',z_{1i}')'$。此为第一阶段回归。

2. 使用 Y 对 $\hat{X}$ 做回归，即使用 y_i 对 $\hat{x}_i$ 做回归。此为第二阶段回归。

由于以上广义矩估计量等价于以上两阶段回归，因而这一估计量也被成为**两阶段最小二乘**（**two-stage least squares**，**2SLS**）。

4.2 2SLS 的推断

根据广义矩估计的一致性结论，我们知道以上两阶段最小二乘估计量是一致估计量。实际上，其一致性也可以通过如下步骤获得:

$$\begin{aligned}
\hat{\beta}&=\left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1}X'Z\left(Z'Z\right)^{-1}Z'Y\\
&=\beta+\left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1}X'Z\left(Z'Z\right)^{-1}Z'U\\
&=\beta+\left[\frac{\sum_{i=1}^N x_iz_i'}{N}\left(\frac{\sum_{i=1}^N z_iz_i'}{N}\right)^{-1}\frac{\sum_{i=1}^N z_ix_i'}{N}\right]^{-1}\\
&\cdot\left[\frac{\sum_{i=1}^N x_iz_i'x_i}{N}\left(\frac{\sum_{i=1}^N z_iz_i'}{N}\right)^{-1}\frac{\sum_{i=1}^N z_iu_i}{N}\right]
\end{aligned}$$

对不同部分使用大数定律，同时由于 $\frac{1}{N}\sum_{i=1}^N z_iu_i\xrightarrow{p}\mathbb{E}\left(z_iu_i\right)=0$，因而以上估计是一致估计。

类似的，我们还可以建立起渐进正态性，在同方差假定下，其渐进分布为

$$\sqrt{N}\left(\hat{\beta}-\beta\right)\overset{a}{\sim}N\left(0,\sigma^2\left\{\mathbb{E}\left(x_iz_i'\right)\left[\mathbb{E}\left(z_iz_i'\right)\right]^{-1}\mathbb{E}\left(z_ix_i'\right)\right\}\right)$$

我们可以使用 $\widehat{\mathrm{Var}\left(\hat{\beta}\right)}=\hat{\sigma}^2\left(\hat{X}'\hat{X}\right)^{-1}$ 对以上渐进方差进行估计，其中 $\hat{\sigma}^2=\frac{1}{N-K}\sum_{i=1}^N\hat{u}_i^2$。注意这里

$$\hat{u}_i=y_i-x_i'\hat{\beta}\neq y_i-\hat{x}_i'\hat{\beta}$$

因而如果计算出 $\hat{x}$，再使用 $\hat{x}$ 计算第二阶段回归，尽管其系数的估计是等价的，但是其方差的估计是不正确的。

在异方差的情况下，也可以使用异方差稳健的方差估计量，即

$$\widehat{\mathrm{Var}\left(\hat{\beta}\right)}=\left(\hat{X}'\hat{X}\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2\hat{x}_i\hat{x}_i'\right)^{-1}\left(\hat{X}'\hat{X}\right)^{-1}$$

4.3 2SLS 中的检验

在得到工具变量估计之后，我们可以对变量的内生性进行检验。我们知道，如果所有变量都是外生的，那么在同方差的假定下，最小二乘估计是一致的且最有效的估计，而工具变量同样也是一致估计；而如果内生性的确存在，此时最小二乘估计失效，而工具变量仍然是一致估计。在原假设 $H_0:\mathrm{Cov}\left(w_i,u_i\right)=0$ 以及备择假设 $H_1:\mathrm{Cov}\left(w_i,u_i\right)\neq 0$ 的假设下，有如下关系：

	H_0	H_1
$\hat{\beta}_{OLS}$	一致、有效	不一致
$\hat{\beta}_{2SLS}$	一致	一致

在这种情况下，可以证明，两个估计量只差的方差等于两个估计量方差之差：

$$\mathrm{Var}\left(\hat{\beta}_{OLS}-\hat{\beta}_{2SLS}\right)=\mathrm{Var}\left(\hat{\beta}_{2SLS}\right)-\mathrm{Var}\left(\hat{\beta}_{OLS}\right)$$

因而可以使用：

$$\left(\hat{\beta}_{OLS}-\hat{\beta}_{2SLS}\right)'\left[\mathrm{Var}\left(\hat{\beta}_{2SLS}\right)-\mathrm{Var}\left(\hat{\beta}_{OLS}\right)\right]^{-1}\left(\hat{\beta}_{OLS}-\hat{\beta}_{2SLS}\right)\overset{a}{\sim}\chi^2\left(q\right)$$

检验 OLS 估计量与 2SLS 估计量之间是否有差异。其中 χ^2 分布的自由度 $q=\mathrm{rank}\left(\mathrm{Var}\left(\hat{\beta}_{2SLS}\right)-\mathrm{Var}\left(\hat{\beta}_{OLS}\right)\right)$。如果差异不显著，那么可以认为并不存在内生性问题。以上检验成为 Hausman 检验。

在存在多于一个工具变量的情况下，我们还可以检验工具变量的有效性，或者过度识别检验（overidentifying）。在原假设 $H_0:\mathbb{E}\left(u_i|z_i\right)=0$ 的条件下，如果工具变量的个数大于内生变量的个数，就可以检验这些工具变量是否外生。实际上，检验工具的有效性即检验矩条件的是否成立，因而广义矩估计中的 Hansen 检验可以直接用来检验工具变量的有效性。除此之外，我们还可以使用 Sargan 检验。

Sargan 检验的原理是，如果我们得到了工具变量的估计，那么残差 $\hat{u}_i=y_i-x_i'\hat{\beta}_{2SLS}$ 应该与所有的外生变量 z_i 无关，因而我们可以使用残差对所有的外生变量 z_i 做回归，并联合检验所有的系数全都等于 0。可以证明，在原假设条件下，以上回归的 R^2 满足 $NR^2\overset{a}{\sim}\chi^2\left(L-K\right)$。如果发现有系数不等于 0，即拒绝了以上原假设，那么可以认为有不满足外生性的工具变量存在。实际上，即使通过了上述检验，也不能保证所有的工具变量都是有效的，但是通不过以上检验则说明工具变量很大可能性存在着问题。

5 工具变量的其他估计方法

除了两阶段最小二乘意外，实际上还有其他的工具变量的估计方法可以使用，在这其中，**控制函数法**（control function）以及**有限信息极大似然**（**limited information maximum likelihood**）方法是最经常使用的方法。

控制函数法的思想是，对于结构方程：

$$y_i = w_i'\gamma + z_{1i}'\delta + u_i = x_i'\beta + u_i \tag{1}$$

内生性的存在是由于 w_i 和 u_i 之间存在着某种程度的相关性，如果我们可以在控制变量中把这些相关性予以控制，那么就可以得到结构参数的一致估计了。将内生变量 w_i 在所有的外生变量上进行投影，得到：

$$w_i = \Gamma z_i + v_i \tag{2}$$

其中 Γ 为 $G \times L$ 的参数向量，当只有一个内生变量，即 $G = 1$ 时，上式等价于：

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

实际上，以上内生变量 w_i 在所有的外生变量上的投影就是 2SLS 中的第一阶段回归。

注意到，由于 z_i 为外生变量，如果 w_i 与 u_i 相关，那么所有的相关性都应该被包含在 v_i 中，而 Γz_i 是 w_i 中外生的部分。因而，我们可以在结构方程中通过控制 v_i，消除 w_i 的内生性。

特别的，如果我们假设 $\mathbb{E}\left(\left(u_i, v_i\right)|z_i\right) = 0$，且：

$$u_i = v_i'\eta + \epsilon_i \tag{3}$$

其中 $\mathbb{E}\left(\epsilon_i | v_i, z_i\right) = 0$，从而我们有：

$$\mathbb{E}\left(\epsilon_i | w_i, z_{1i}, v_i\right) = \mathbb{E}\left(z_i, v_i\right) = 0$$

因而我们可以使用回归：

$$y_i = w_i'\gamma + z_{1i}'\delta + v_i'\eta + \epsilon_i$$

得到结构参数的一致估计。

然而，现实中，v_i 是不可观测的，因而我们可以使用第一阶段回归的残差，即：

$$\hat{v}_i = w_i - \hat{\Gamma} z_i \tag{4}$$

替代 v_i 进行估计，即估计如下回归方程：

$$y_i = w_i'\gamma + z_{1i}'\delta + \hat{v}_i'\eta + \epsilon_i \tag{5}$$

注意到，如果我们根据分步回归的结论，以上回归等价于：

1. 首先使用 w_i 对 $\hat{v}_i$ 做回归得到残差，然而由于 $\hat{v}_i$ 是式 (4) 中的残差，因而得到的残差就是 Γz_i；
2. 使用 z_{1i} 对对 $\hat{v}_i$ 做残差，然而由于在式 (2) 中，z_i 包含了 z_{1i}，因而这一步得到的残差就是 z_{1i}；
3. 使用 y_i 对以上的两个残差做回归，即使用 y_i 对 Γz_i 以及 z_{1i} 做回归，得到回归系数。

实际上如果观察以上第 (3) 步可以看到，Γz_i 实际上就是 2SLS 中第一阶段的拟合值，因而实际上以上步骤表明控制函数法与 2SLS 是等价的。虽然在线性模型中，2SLS 与控制函数法是等价的，但是在一般的非线性模型，比如 Probit、Logit 回归中，一般来说控制函数法可以得到一致的估计，而 2SLS 步骤是不可用的。

在得到了以上的估计以后，在式 (5) 中，我们可以通过检验：$H_0:\eta=0$ 进行内生性检验：如果没有内生性，那么意味着 u_i 和 w_i 没有相关性，而 w_i 与 u_i 的相关性都体现在了 v_i 上，因而我们可以通过检验 v_i 的系数检验内生性是否存在。

或者，我们可以使用极大似然的方法同时估计以上的结构方程 (1) 和第一阶段方程 (2)。如果我们假设 (u_i, v_i) 服从一个联合正态分布，那么根据联合正态分布的性质，我们一定可以将其写成方程 (3) 的形式，且 ϵ_i 和 v_i 是相互独立的。如此，我们有：

$$f\left(y_i|w_i,z_i\right)=\frac{1}{\sqrt{2\pi}\sigma_\epsilon}\exp\left\{-\frac{\left(y_i-w_i'\gamma-z_{1i}'\delta-\left(w_i-\Gamma z_i\right)'\eta\right)^2}{2\sigma_\epsilon^2}\right\}$$

同时：

$$f\left(w_i|z_i\right)=\frac{1}{\sqrt{2\pi}\sigma_v}\exp\left\{-\frac{\left(w_i-\Gamma z_i\right)^2}{2\sigma_v^2}\right\}$$

从而：

$$f\left(y_i,w_i|z_i\right)=f\left(y_i|w_i,z_i\right)\cdot f\left(w_i|z_i\right)$$

将以上条件密度函数取对数，并根据样本加总，得到对数似然函数并最大化即可得到有限信息条件极大似然估计。