高级计量经济学

Lecture 8: Nonlinear Regression and IV Estimator

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

办公室 粤海校区汇文楼1510
E-mail huangjp@szu.edu.cn
Website https://huangjp.com

非线性回归模型和
MM估计量

线性模型与非线性模型

Linear and Nonlinear Models

线性回归模型可以表达为

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \mid \boldsymbol{X} \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

或

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \mid \boldsymbol{X} \sim \mathrm{IID}(0, \sigma^2), \quad t = 1, \ldots, n$$

定义信息集（information set）Ω_t 为所有可能对 y_t 产生影响的变量的集合。一般情况下我们要求 Ω_t 中的变量为外生或者前定变量。

基于信息集 Ω_t，我们可以将线性模型拓展为下面的非线性模型

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \mid \Omega_t \sim \mathrm{IID}(0, \sigma^2), \quad t = 1, \ldots, n$$

因此 $E[u_t \mid \Omega_t] = 0$

这里 $x_t(\boldsymbol{\beta})$ 是参数向量 $\boldsymbol{\beta}$ 的非线性函数。非线性模型也可以写成

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

注：$x_t(\boldsymbol{\beta})$ 也可以写成 $f(\boldsymbol{x}_t, \boldsymbol{\beta})$。为了使符号和线性模型相对应，这里采用 $x_t(\boldsymbol{\beta})$ 的写法。

矩条件

Moment Conditions

在线性模型下，如果假设前定性 $E[u_t \mid \boldsymbol{X}_t] = 0$，我们可以推出 $E[\boldsymbol{X}_t^\top u_t] = \mathbf{0}$ （Lecture 5），而这恰好是求 MM 估计量的条件（Lecture 2）。此条件对应的样本条件是

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}$$

左式称为 moment conditions

在非线性模型下，我们可以从信息集 Ω_t 选择 k 个变量并写成 $1 \times k$ 向量 $\boldsymbol{W}_t$。此时可得 $E[u_t \mid \boldsymbol{W}_t] = 0$，因此 $E[\boldsymbol{W}_t^\top u_t] = \mathbf{0}$，对应的矩条件为

$$\boldsymbol{W}^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big) = \mathbf{0}, \qquad \boldsymbol{W} \text{ 的第 } t \text{ 行是 } \boldsymbol{W}_t$$

非线性模型的矩条件中包含 k 个关于系数的非线性方程，其解（若存在）就是 $\boldsymbol{\beta}$ 的 MM 估计量，可记作 $\hat{\boldsymbol{\beta}}_{\text{MM}}$。

关于 MM 估计量

- MM 估计量是 $\boldsymbol{W}^{\top}\big(\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta})\big)=\mathbf{0}$ 的解。因为条件是非线性的，我们无法获得解的一般表达式。

- 矩条件基于 $\boldsymbol{W}_t$ 的前定性 $E[u_t \mid \boldsymbol{W}_t]=0$ 而非外生性，因此 MM 估计量可能有偏。

- MM 估计量的性质会随 $\boldsymbol{W}$ 中变量的选择而变化。$\boldsymbol{W}$ 的选择一般不会影响 MM 估计量的一致性，但会影响它的渐进协方差矩阵。因此我们需要找到使 MM 估计量满足渐进有效性的 $\boldsymbol{W}$。

回归系数的渐进识别*

Asymptotic Identification

识别（Identification）指在确定样本和估计方法后，可以求出唯一的参数值的情况。如果一种估计方法在 $n \to \infty$ 时可以求出唯一的参数值，我们称其为渐进可识别。

在线性模型的 OLS 估计中，如果 $\boldsymbol{X}$ 列满秩，则 $\boldsymbol{\beta}$ 是可识别的（标准方程有唯一解）。在非线性模型的 MM 估计中，如果 $\boldsymbol{W}^{\top}\big(\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta})\big)=\mathbf{0}$ 存在唯一解，则 $\boldsymbol{\beta}$ 是可识别的。

MM 估计量的渐进可识别性取决于 $\frac{1}{n}\boldsymbol{W}^{\top}\big(\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta})\big)$ 的概率极限是否可以求出唯一的参数值。令

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^{\top}\big(\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta})\big)=\boldsymbol{\alpha}(\boldsymbol{\beta}),\qquad \boldsymbol{\alpha}(\boldsymbol{\beta})\text{ 为确定性函数。}$$

已知 $\frac{1}{n}\boldsymbol{W}^{\top}\big(\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta}_0)\big)=\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{W}_t^{\top}u_t$，$E[\boldsymbol{W}_t^{\top}u_t]=\mathbf{0}$，根据 LLN 可得

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^{\top}\big(\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta}_0)\big)=\mathbf{0}$$

因此，如果 $\boldsymbol{\beta}_0$ 是 $\boldsymbol{\alpha}(\boldsymbol{\beta})=\mathbf{0}$ 的唯一解，MM 估计量是渐进可识别的。

需要注意的是，可识别性无法推出渐进可识别性，反之亦然（pp.217-218）。

MM 估计量的一致性*

Consistency of MM Estimators

假设 MM 估计量是渐进可识别的，则

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^\top\left(\boldsymbol{y}-\boldsymbol{x}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\right)=\boldsymbol{\alpha}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})=\mathbf{0}$$

存在唯一解。如果我们假设 $\operatorname*{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}=\boldsymbol{\beta}_\infty$（非随机），则根据上式可得 $\boldsymbol{\alpha}(\boldsymbol{\beta}_\infty)=\mathbf{0}$，但是渐进可识别性的定义告诉我们只有 $\boldsymbol{\beta}_0$ 满足 $\boldsymbol{\alpha}(\boldsymbol{\beta})=\mathbf{0}$，因此

$$\operatorname*{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}=\boldsymbol{\beta}_0$$

即 MM 估计量满足一致性。

MM 估计量的渐进可识别性取决于如何选取 $\boldsymbol{W}$ 里的变量。

MM 估计量的其他渐进性质*

Other Asymptotic Properties of MM Estimators

非线性模型的 MM 估计量还具有以下渐进性质：

- 渐进正态性：$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{\beta}_0)$ 服从均值为 $\mathbf{0}$ 的渐进正态分布（pp.220-222）
- 渐进有效性：$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{\beta}_0)$ 的渐进协方差矩阵是

$$\sigma_0^2 \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n}\boldsymbol{X}_0^\top \boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{X}_0\right)^{-1} = \sigma_0^2 \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n}\boldsymbol{X}_0^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}_0\right)^{-1}$$

其中 $\boldsymbol{X}_0 = \boldsymbol{X}(\boldsymbol{\beta}_0) = \left[\partial x_t(\boldsymbol{\beta}) \middle/ \partial \beta_i\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$。

当 $\boldsymbol{W} = \boldsymbol{X}_0$ 时，$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{\beta}_0)$ 在 MM 估计量当中最有效。（pp.222-223）

需要注意的是，$\boldsymbol{X}_0$ 是未知量 $\boldsymbol{\beta}_0$ 的函数，因此这样定义的有效估计量不可行（infeasible）。

线性模型与 MM 估计量

线性模型 $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ 是非线形模型的一个特例，因此非线形模型的 MM 估计量也适用于线性模型。

在线性模型下，MM 估计量是

$$\boldsymbol{W}^\top\left(\boldsymbol{y} - \boldsymbol{X\beta}\right) = \boldsymbol{0}$$

的解。根据定义 $\boldsymbol{X}(\boldsymbol{\beta}) = \left[\partial x_t(\boldsymbol{\beta}) \big/ \partial\beta_i\right] = \boldsymbol{X}$，因此 $\boldsymbol{X}_0 = \boldsymbol{X}$，最有效的 MM 估计量满足

$$\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{MM}}\right) = \boldsymbol{0} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}}_{\mathrm{MM}} = \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$$

非线性最小二乘法
Nonlinear Least Squares

因为 $\boldsymbol{X}_0 = \boldsymbol{X}(\boldsymbol{\beta}_0) = \left[\partial x_t(\boldsymbol{\beta}) \big/ \partial \beta_i\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ 是未知参数 $\boldsymbol{\beta}_0$ 的函数，我们无法通过它求得最有效的 MM 估计量。

假设估计量 $\hat{\boldsymbol{\beta}}$ 是下面 MM 条件的解

$$\boldsymbol{X}^\top(\boldsymbol{\beta})\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big) = \mathbf{0}$$

同时我们也可以考虑非线形最小二乘法

$$\min_{\boldsymbol{\beta}} \big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top \big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)$$

上面的 MM 条件等价于非线性最小二乘法的一阶条件。因此，我们将 $\hat{\boldsymbol{\beta}}$ 称为非线性最小二乘估计量（NLS），并记为 $\hat{\boldsymbol{\beta}}_{\text{NLS}}$。

当 $n \to \infty$ 时，$\hat{\boldsymbol{\beta}}_{\text{NLS}}$ 收敛于有效 MM 估计量（$\boldsymbol{W} = \boldsymbol{X}_0$）。

广义最小二乘法

线性模型与广义最小二乘法

OLS 和 NLS 估计量都需要误差项满足同方差性。下面我们放松这个假设，考虑线性模型

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \quad E[\boldsymbol{u}\boldsymbol{u}^\top] = \boldsymbol{\Omega}$$

一般情况下 $\boldsymbol{\Omega} \neq \sigma^2 \boldsymbol{I}$，因此不满足 Gauss-Markov 定理的条件。

为了获得有效估计量，我们可以将原模型变换为满足条件的新模型，并将新模型的 OLS 估计量作为 $\boldsymbol{\beta}$ 的估计量。这种方法被称为广义最小二乘法（generalized least squares，GLS）。

广义最小二乘估计量

GLS Estimator

已知 $\boldsymbol{\Omega}$ 是正定矩阵且可逆，因此存在 $n \times n$ 非奇异矩阵 $\boldsymbol{\Psi}$ 满足

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$$

将回归方程 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ 从左侧乘以 $\boldsymbol{\Psi}^{\top}$ 可得 $\boldsymbol{\Psi}^{\top}\boldsymbol{y} = \boldsymbol{\Psi}^{\top}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Psi}^{\top}\boldsymbol{u}$。

GLS 估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 是变换后模型的 OLS 估计量，即

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\boldsymbol{X}^{\top}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\boldsymbol{y} = (\boldsymbol{X}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{y}$$

变换后的误差项的协方差矩阵是

$$E[\boldsymbol{\Psi}^{\top}\boldsymbol{u}\boldsymbol{u}^{\top}\boldsymbol{\Psi}] = \boldsymbol{\Psi}^{\top}E[\boldsymbol{u}\boldsymbol{u}^{\top}]\boldsymbol{\Psi} = \boldsymbol{\Psi}^{\top}(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top})^{-1}\boldsymbol{\Psi} = \boldsymbol{I}$$

由此可得到 GLS 估计量的协方差矩阵

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = 1 \cdot (\boldsymbol{X}^{\top}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\boldsymbol{X})^{-1} = (\boldsymbol{X}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}$$

广义最小二乘法

变换后模型 $\boldsymbol{\Psi}^\top \boldsymbol{y} = \boldsymbol{\Psi}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Psi}^\top \boldsymbol{u}$ 的 OLS 估计量是 SSR 最小化问题的解，该问题可以写成

$$
\begin{aligned}
& \min_{\boldsymbol{\beta}} \left(\boldsymbol{\Psi}^\top \boldsymbol{y} - \boldsymbol{\Psi}^\top \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{\Psi}^\top \boldsymbol{y} - \boldsymbol{\Psi}^\top \boldsymbol{X}\boldsymbol{\beta}\right) \\
= & \min_{\boldsymbol{\beta}} \left(\boldsymbol{\Psi}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)^\top \left(\boldsymbol{\Psi}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \\
= & \min_{\boldsymbol{\beta}} \ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Psi}\boldsymbol{\Psi}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
= & \min_{\boldsymbol{\beta}} \ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})
\end{aligned}
$$

我们将目标函数 $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ 称为 GLS 准则函数（GLS criterion function）。

GLS 估计量是 MM 估计量

最小化问题 $\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ 的一阶条件是

$$\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}$$

因此，$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 也可以看作满足矩条件 $\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = \mathbf{0}$ 的 MM 估计量。一般情况下，矩条件 $\boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}$ 的解可以写成

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} = (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{y}$$

$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 是 $\boldsymbol{W} = \boldsymbol{\Omega}^{-1} \boldsymbol{X}$ 时的 $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$。

GLS 估计量的统计学性质*

GLS 估计量的统计学性质类似于 MM 估计量的统计学性质。

- 当 $\boldsymbol{X}$ 和 $\boldsymbol{W}$ 外生时，即 $E[\boldsymbol{u} \mid \boldsymbol{X}, \boldsymbol{W}] = \mathbf{0}$ 时，$\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ 非偏。
- 当 $\boldsymbol{W}$ 满足前定性，即 $E[u_t \mid W_t] = 0$ 时，$\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ 满足一致性。
- $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ 的协方差矩阵是

$$
\begin{aligned}
\operatorname{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}) &= E\left[(\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} - \boldsymbol{\beta}_0)^\top\right] \\
&= E\left[(\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{u}\boldsymbol{u} \boldsymbol{W}^\top (\boldsymbol{X}^\top \boldsymbol{W})^{-1}\right] \\
&= (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{\Omega} \boldsymbol{W}^\top (\boldsymbol{X}^\top \boldsymbol{W})^{-1}
\end{aligned}
$$

已知，$\operatorname{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = (\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X})^{-1}$，可以通过证明两者之差是半正定矩阵证明 GLS 估计量的有效性。

OLS 估计量也是 MM 估计量，因此 GLS 估计量至少和 OLS 估计量同样有效，且在大多数情况下比 OLS 估计量更有效。

GLS 估计量的计算*

虽然 GLS 估计量有很好的性质，但是并不容易计算。

$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{y}$，即使 $\boldsymbol{\Omega}$ 已知，我们需要求它的逆矩阵。当 n 很大时，$\boldsymbol{\Omega}^{-1}$ 的计算需要占用大量的计算和储存资源（例如 $n = 10{,}000$ 时，储存 $\boldsymbol{\Omega}$ 和 $\boldsymbol{\Omega}^{-1}$ 需要 1600Mb 空间）。

如果 $\boldsymbol{\Psi}$ 已知，且我们可以将数据变换为 $\boldsymbol{\Psi}^\top\boldsymbol{y}$ 和 $\boldsymbol{\Psi}^\top\boldsymbol{X}$ 而不需要储存 $\boldsymbol{\Psi}$，就可以通过 OLS 估计 $\hat{\boldsymbol{\beta}}_{\text{GLS}}$。

OLS 估计不需要用到误差项的方差，因此我们可以利用这一特性。如果 $\boldsymbol{\Omega} = \sigma^2\boldsymbol{\Delta}$，且 $\boldsymbol{\Delta}$ 为已知，则可以用 $\boldsymbol{\Delta}$ 替代 $\boldsymbol{\Omega}$ 求 $\hat{\boldsymbol{\beta}}_{\text{GLS}}$。具体做法是找到满足 $\boldsymbol{\Delta}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$ 的 $\boldsymbol{\Psi}$，并对原模型进行变换，再用变换后模型进行 OLS 估计，所获得的估计量就是原模型的 GLS 估计量：

$$(\boldsymbol{X}^\top\boldsymbol{\Delta}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Delta}^{-1}\boldsymbol{y} = (\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{\text{GLS}}$$

这时 GLS 估计量的协方差矩阵可以写成 $\sigma^2(\boldsymbol{X}^\top\boldsymbol{\Delta}^{-1}\boldsymbol{X})^{-1}$。如果 σ^2 未知，我们依然要对其进行估计，即 $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = s^2(\boldsymbol{X}^\top\boldsymbol{\Delta}^{-1}\boldsymbol{X})^{-1}$，$s^2$ 是变换后模型误差项的估计量。

加权最小二乘估计量
Weighted Least Square Estimator

如果误差项是异方差但是无自相关，那么就可以很容易地计算 GLS 估计量。这时的 $\boldsymbol{\Omega}$ 是对角矩阵，我们将其第 t 对角要素写为 ω_t^2，即
$\boldsymbol{\Omega} = \mathrm{diag}(\omega_1^2, \ldots, \omega_n^2)$，则可得

$$\boldsymbol{\Omega}^{-1} = \mathrm{diag}(\omega_1^{-2}, \ldots, \omega_n^{-2}),\ \boldsymbol{\Psi} = \mathrm{diag}(\omega_1^{-1}, \ldots, \omega_n^{-1})$$

变换后的模型可以写成

$$\omega_t^{-1} y_t = \omega_t^{-1} \boldsymbol{X}_t \boldsymbol{\beta} + \omega_t^{-1} u_t, \qquad t = 1, \ldots, n$$

通过对这个模型进行 OLS 估计得到的估计量被称为加权最小二乘估计量（WLS 估计量），因为这可以看作是用权重 ω_t^{-1} 给样本中第 t 观测值进行加权。

因为变换后模型的误差项 $\omega_t^{-1} u_t$ 满足同方差性，WLS 估计可以作为应对异方差性的一种方法。

如何决定 ω_t *

WLS 估计的难点是如何决定 ω_t。这可以分为以下几种情况：

- 通过理论或者对样本数据的检验，我们相信 $E[u_t^2]$ 和某个可观测变量 z_t^2 成正比。此时可以用 z_t^{-1} 作为权重。
- 样本变量是针对不同大小的集合获得的统计数据。
  - 如果变量是集合中的均值，例如不同城市的人均可支配收入，则 u_t 的方差和集合的要素数 N_t 呈反比。此时可用 $N_t^{1/2}$ 作为权重。
  - 如果变量是总和而不是均值，则 u_t 的方差和 N_t 成正比，此时需用 $N_t^{-1/2}$ 作为权重。

工具变量估计

内生性

Endogeneity

至今为止我们讨论过的估计方法（OLS，MM，GLS）都需要假设解释变量（或者信息集）是外生的或者前定的。

例如在 MM 估计中，我们需要从信息集 Ω_t 中选取变量 $\boldsymbol{W}_t$，以保证 $E[u_t \mid \boldsymbol{W}_t] = 0$。

在实践中很难保证所有解释变量都和误差项不相关。如果某个解释变量和误差项相关，它就是内生变量。内生性可以导致 OLS 估计量有偏且不一致。

内生性可以分为以下几种：

- Errors in variables (measurement error)
- Simultaneity（联立方程、或称双向因果）
- Omitted variables（遗漏变量）

工具变量

Instrumental Variables

考虑下面的线性回归模型

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \quad E[\boldsymbol{u}\boldsymbol{u}^\top] = \sigma^2 \boldsymbol{I}$$

且 $\boldsymbol{X}$ 中至少有一个内生变量。

假设针对任意观测值 t，我们都能找到信息集 Ω_t 使其满足 $E[u_t \mid \Omega_t] = 0$，并且能定义 $n \times k$ 矩阵 $\boldsymbol{W}$ 使其第 t 行 $\boldsymbol{W}_t$ 的要素都包含在 Ω_t 中。

这样定义的 $\boldsymbol{W}$ 中的变量被称为工具变量（instrumental variables, or instruments）。

工具变量应该是外生的或者前定的，且包含 $\boldsymbol{X}$ 中所有外生或前定变量。

IV 估计量

Instrumental Variables Estimator

工具变量 $\boldsymbol{W}$ 满足矩条件

$$\boldsymbol{W}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})=\boldsymbol{0}$$

此等式的解 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 称为 IV 估计量，即

$$\hat{\boldsymbol{\beta}}_{\text{IV}}=(\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{W}^\top\boldsymbol{y}$$

如果忽略模型上的假设，IV 估计量
和 MM 估计量的表达式相同

从 MM 估计量的性质可知，在前定性和可识别性条件下，$\hat{\boldsymbol{\beta}}_{\text{IV}}$ 满足一致性和渐进正态性。基于样本 $\boldsymbol{X}$ 的可识别性是 $\boldsymbol{W}^\top\boldsymbol{X}$ 可逆，而渐进可识别性条件是

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{X}\ \text{是非奇异确定矩阵}$$

事实上，$\hat{\boldsymbol{\beta}}_{\text{IV}}$ 的一致性不需要前定性条件 $E[u_t \mid \boldsymbol{W}_t]=0$，而只需要

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{u}=\boldsymbol{0}$$

$$E[u_t \mid \boldsymbol{W}_t]=0 \ \underset{\not\Leftarrow}{\Rightarrow}\ \operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{u}=\boldsymbol{0}$$

1. 当 $\boldsymbol{X}$ 和 $\boldsymbol{W}$ 都只包含一个变量时，可识别性条件意味着
   $\mathrm{Cov}[w_t, x_t]\neq 0$
2. 前定性条件可以推出
   $\mathrm{Cov}[\boldsymbol{W}_t, u_t]=0$

这个条件被称为工具变量的渐进不相关（asymptotic uncorrelated）条件。

IV 估计量的有效性*

当真实参数值是 $\boldsymbol{\beta}_0$ 和 σ_0^2 时，

$$\operatorname{Var}\left[\operatorname*{plim}_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}}-\boldsymbol{\beta}_0)\right]=\sigma_0^2\operatorname*{plim}_{n\to\infty}\left(\tfrac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}$$

因此，IV 估计量的渐进有效性取决于如何选择 $\boldsymbol{W}$ 中的变量。我们称 IV 估计量满足渐进有效性的工具变量为最优工具变量（optimal instruments）。

理论上，我们可以定义矩阵 $\bar{\boldsymbol{X}}$，使其第 t 行为 $\bar{\boldsymbol{X}}_t=E[\boldsymbol{X}_t\mid\Omega_t]$，且满足

$$\boldsymbol{X}=\bar{\boldsymbol{X}}+\boldsymbol{V},\qquad E[\boldsymbol{V}_t\mid\Omega_t]=0$$

可以将其理解为生成 $\boldsymbol{X}$ 的 DGP

由此假设可以证明 $\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}=\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{X}}^{\top}\boldsymbol{P}_{\boldsymbol{W}}\bar{\boldsymbol{X}}$。当 $\boldsymbol{W}=\bar{\boldsymbol{X}}$ 时，右侧的概率极限等于 $\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{X}}^{\top}\bar{\boldsymbol{X}}$，在所有可选择的 $\boldsymbol{W}$ 中最有效，因此 $\bar{\boldsymbol{X}}$ 是最优工具变量（详见 p.318）。

令 $\boldsymbol{Z}$ 为包含 $\boldsymbol{X}$ 中外生或前定变量的子矩阵，则 $\bar{\boldsymbol{Z}}=\boldsymbol{Z}$，因此 $\boldsymbol{Z}$ 也是 $\bar{\boldsymbol{X}}$ 的子矩阵。这就解释了为什么 $\boldsymbol{W}$ 应该包含 $\boldsymbol{X}$ 中的所有外生或前定变量。

和 MM 估计量的渐进有效性类似，我们无法观测 $\bar{\boldsymbol{X}}$，而只能想办法找到它的一致估计量。

IV 估计中的识别
Identification in IV Estimation

至此，我们假设了工具变量矩阵 $\boldsymbol{W}$ 是 $n \times k$ 矩阵，所以工具变量的个数等于 $\boldsymbol{\beta}$ 中参数的个数。

在实践中，有时我们可以从信息集中找出 ℓ 个工具变量，从而构建 $n \times \ell$ 矩阵 $\boldsymbol{W}$。根据矩条件 $\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}$，可知其中共包含 ℓ 个等式，因此：

- 当 $\ell > k$ 时，我们称模型为过度识别（overidentified），此时矩条件的个数大于参数的个数，满足条件的估计量往往不存在；
- 当 $\ell = k$ 时，我们称模型为恰好识别（just/exactly identified），此时矩条件的个数等于参数的个数，因此存在唯一解；
- 当 $\ell < k$ 时，我们称模型为识别不足（underidentified），此时矩条件不存在唯一解。

当 $\ell > k$ 时，最有效的 IV 估计量称为广义 IV 估计量（generalized IV estimator，or GIVE）。$\ell = k$ 时的 IV 估计量可称为简单 IV 估计量（simple IV estimator）。

广义 IV 估计量*

Generalized IV Estimator

当 $\ell > k$ 时，我们可以从 ℓ 个工具变量中选取 k 种线性结合，从而构筑 k 个矩条件。这可以通过定义 $\ell \times k$ 矩阵 $\boldsymbol{J}$，从而使 $\boldsymbol{WJ}$ 为 $n \times k$ 矩阵，并建立矩条件 $\boldsymbol{J}^\top \boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{0}$ 来完成。

在选择矩阵 $\boldsymbol{J}$ 时，应当使其满足下列条件：

1. $\mathrm{rank}(\boldsymbol{WJ}) = k$，这是为了保证可识别性
2. $\boldsymbol{J}$ 至少应该是渐进确定的（asymptotic deterministic）
3. $\boldsymbol{J}$ 应当使 IV 估计量满足渐进有效性

因此，矩条件 $\boldsymbol{J}^\top \boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{0}$ 之解 $(\boldsymbol{J}^\top \boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{J}^\top \boldsymbol{W}^\top \boldsymbol{y}$ 代表一种估计量的集合，而是其中最有效的是 GIVE。

广义 IV 估计量*

Generalized IV Estimator

从简单 IV 估计量的性质可知，当 $\boldsymbol{X} = \bar{\boldsymbol{X}} + \boldsymbol{V}$，$E[\boldsymbol{V}_t \mid \Omega_t] = 0$时，用 $\boldsymbol{WJ}$ 替代 $\boldsymbol{W}$ 获得的渐进协方差矩阵是

$$\sigma_0^2 \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n}\bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{WJ}} \bar{\boldsymbol{X}}\right)^{-1}$$

简单 IV 估计量中的最优工具变量是 $\boldsymbol{W} = \bar{\boldsymbol{X}}$。

根据定义，$\bar{\boldsymbol{X}}_t \in \Omega_t$，因此 $\bar{\boldsymbol{X}}_t$ 是 $\boldsymbol{W}_t$ 中变量的确定函数，但不一定是线性函数。一般情况下不存在满足 $\bar{\boldsymbol{X}} = \boldsymbol{WJ}$ 的矩阵 $\boldsymbol{J}$。

$\boldsymbol{WJ}$ 是 $\boldsymbol{W}$ 的列空间 $\mathcal{S}(\boldsymbol{W})$ 中的点集，如果无法在 $\mathcal{S}(\boldsymbol{W})$ 中寻找最优解，作为次优解，我们可以选择 $\bar{\boldsymbol{X}}$ 在 $\mathcal{S}(\boldsymbol{W})$ 上的正交投影。即令

$$\boldsymbol{WJ} = \boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}} = \boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \bar{\boldsymbol{X}}$$

此时 $\boldsymbol{J} = (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \bar{\boldsymbol{X}}$。

可以证明当 $\boldsymbol{WJ} = \boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}}$ 时，IV 估计量满足渐进有效性（比较对象是所有可能的 $\boldsymbol{WJ}$ ，详见 p.320）。

广义 IV 估计量*

Generalized IV Estimator

和前面一样，$\bar{\boldsymbol{X}}$ 未知，因此无法直接计算 $\boldsymbol{P_W}\bar{\boldsymbol{X}}$。但是从 $\boldsymbol{J} = (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\bar{\boldsymbol{X}}$ 可得，

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\boldsymbol{J} &= \operatorname*{plim}_{n\to\infty}(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W})^{-1}(\frac{1}{n}\boldsymbol{W}^\top\bar{\boldsymbol{X}}) \\ &= \operatorname*{plim}_{n\to\infty}(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W})^{-1}(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{X})\end{aligned}$$

$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^\top\bar{\boldsymbol{X}} = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{X}$ （p.318）

因此，我们可以用 $\boldsymbol{P_W}\boldsymbol{X}$ 替代 $\boldsymbol{P_W}\bar{\boldsymbol{X}}$ 而不改变估计量的渐进性质。

当选择 $\boldsymbol{W}\boldsymbol{J} = \boldsymbol{P_W}\boldsymbol{X}$ 时，矩条件变为

$$\boldsymbol{X}^\top\boldsymbol{P_W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}$$

GIV 估计量为 $\hat{\boldsymbol{\beta}}_{\mathrm{GIV}} = (\boldsymbol{X}^\top\boldsymbol{P_W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{P_W}\boldsymbol{y}$。 （在 $\ell = k$ 时，$\hat{\boldsymbol{\beta}}_{\mathrm{GIV}} = \hat{\boldsymbol{\beta}}_{\mathrm{IV}}$）

GIV 估计量也可以作为最优化问题 $\min_{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{P_W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ 的解导出。

两阶段最小二乘估计

Two Stage Least Squares Estimation

GIV 估计量可以写成

$$\hat{\boldsymbol{\beta}}_{\mathrm{GIV}} = (\boldsymbol{X}^\top \boldsymbol{P_W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P_W} \boldsymbol{y} = (\boldsymbol{X}^\top \boldsymbol{P_W}^\top \boldsymbol{P_W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P_W}^\top \boldsymbol{y}$$

从最后一项可以看出，$\hat{\boldsymbol{\beta}}_{\mathrm{GIV}}$ 是回归模型

$$\boldsymbol{y} = \boldsymbol{P_W} \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{v}$$

的 OLS 估计量。其中的解释变量 $\boldsymbol{P_W X}$ 是用 $\boldsymbol{W}$ 回归每一个 $\boldsymbol{x}_i$ 所得的预测值所组成的矩阵。

因此，GIV 估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{GIV}}$ 可以通过下面的两阶段最小二乘回归（2SLS）获得：

1. 第一阶段（first stage）：对 $\boldsymbol{x}_i = \boldsymbol{W\beta} + \boldsymbol{w}$ 进行 OLS 估计，并计算 $\hat{\boldsymbol{x}}_i$；
2. 第二阶段（second stage）：令 $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_k]$，并对 $\boldsymbol{y} = \hat{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{v}$ 进行 OLS 估计。

在第二阶段回归中获得的 OLS 估计量就是原模型的 GIV 估计量。

在计算能力缺乏的时代，2SLS 不失为一种计算 GIV 估计量的好方法。但是通过 2SLS 无法求出原模型中 σ^2 的一致估计量。在实际应用中不需要特意使用 2SLS，因为现在的计量软件都可以直接进行 IV 估计，并正确计算回归标准误差。2SLS 的优点是可以帮助我们理解 IV 估计量的一些性质。

IV 估计量的小样本性质

即使 IV (GIV) 估计量能满足一致性、渐进有效性等大样本性质，在有限样本下，它几乎永远是有偏的。

导致 IV 估计量有限样本偏差的原因包括：

- 工具变量的数量 ℓ 过多，使第一阶段回归的拟合效果非常好（R^2 接近于 1），导致 $\hat{\boldsymbol{X}}$ 的取值非常接近 $\boldsymbol{X}$。此时第二阶段回归的结果就非常接近于原模型的 OLS 估计。这种情况下 IV 估计和 OLS 估计的偏差相似。
- 第一阶段中存在解释能力很低的模型（R^2 很小或 F 统计值不显著），称之为存在弱工具变量（weak instruments）。此时 IV 估计量的有限样本分布和其渐进分布可能差别很大，导致有限样本偏差。

因此，选择工具变量时应使其满足：

1. 工具变量与误差项不相关（外生性、前定性、或渐近不相关性）；
2. 工具变量与内生变量相关（第一阶段回归存在 OLS 解）。

通常我们把这两项总结为：工具变量只通过 $\boldsymbol{X}$ 对 $\boldsymbol{y}$ 产生影响。

MM 估计量总结

	矩条件	估计量	最小化目标函数
NLS	$\boldsymbol{X}^\top(\boldsymbol{\beta})\big(\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta})\big)=\mathbf{0}$	矩条件的唯一解	$\big(\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta})\big)^\top\big(\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta})\big)$
GLS	$\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})=\mathbf{0}$	$(\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{y}$	$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$
IV (GIV)	$\boldsymbol{J}^\top\boldsymbol{W}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})=\mathbf{0}$ $\boldsymbol{J}=(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\bar{\boldsymbol{X}}$	$(\boldsymbol{X}^\top\boldsymbol{P}_W\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{P}_W\boldsymbol{y}$	$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{P}_W(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$

以上估计量都是广义矩方法（generalized method of moments, GMM）的特例。
关于 GMM 可参考书中第 9 章。

课外阅读

- Angrist, J. D. and Kruger, A. B. (1991).
  Does Compulsory School Attendance Affect Schooling and Earnings?
  The Quarterly Journal of Economics, 106:4, 979-1014.
  https://www.jstor.org/stable/2937954

- Bound, J., Jaeger, D. A., and Baker, R. M. (1995).
  Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak.
  Journal of the American Statistical Association, 90:430, 443-450.
  https://www.jstor.org/stable/2291055

- Angrist, J. D., Imbens, G. W., and Kruger, A. B. (1999).
  Jackknife Instrumental Variables Estimation.
  Journal of Applied Econometrics, 14:1, 57-67.
  https://www.jstor.org/stable/223249