高级计量经济学 I

Based on lectures by Yu Jihai(GSM)

Wang Bencheng

Department of Applied Economics, GSM

2024 年 1 月 24 日

Linear Regression Model

- 多元线性回归的矩阵形式、OLS 估计量
- OLS 的小样本性质、大样本渐进性质以及假设检验
- GLS 与 WLS 方法，异方差与时间序列问题，稳健标准误 [1]

Maximum Likelihood Method

- MLE 的基本性质、MLE 估计量的渐进正态性质
- MLE 三大检验：*Wald Test*、*LM Test*、*LR Test*
- *Discrete Choice Model* 及其 MLE 估计量、检验 [2]

IV and General Mement Method

- IV 的基本性质、矩估计与 2SLS 估计量
- GMM 估计量及其检验 [3]

Panel Data

- 固定效应模型与随机效应模型
- 动态面板模型 [4]

Spatial Economics

- 空间计量模型 [5]

目录

1 *Linear Regression Model*

1. OLS 统计量及其 BLUE 性质；

2. OLS 小样本正态假定下的假设检验；

3. OLS 大样本渐进正态性质；

4. GLS 方法与 WLS 方法：异方差与自回归问题；

本部分内容为最经典最基础最完备的计量经济理论，采用矩阵语言重写中级计量的多元回归内容，在实践应用中几乎没有任何参考性，但是其基本范式是一脉相承的，即讨论估计量的无偏性、一致性以及渐进正态分布，进而构造假设检验。

1.1 多元线性回归的矩阵处理

给定多元线性回归模型：

$$Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \varepsilon_{n\times 1}$$

相应的对应于 n 个方程 k 个参数，当 $k \geq n$ 时方程有解，如果方程个数多于未知数个数，则无解。注意到，自变量 X 可以包括截距项 $x = 1$，包括主回归变量和控制变量等，因而 X 是广义的回归项；转化为矩阵形式表示为：

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

一般情况下对于该模型做如下假定：

1. 条件独立性假定 CIA 或均值独立假定：$E(\varepsilon|X) = 0 \to E(\varepsilon X) = 0$；
2. 同方差假定：给定协方差矩阵[1]$E = E(\varepsilon\varepsilon') = E\begin{bmatrix} \varepsilon_1^2 & \dots & \varepsilon_1\varepsilon_n \\ \dots & \dots & \dots \\ \varepsilon_n\varepsilon_1 & \dots & \varepsilon_n^2 \end{bmatrix}$，同方差假定意味着 $E(\varepsilon\varepsilon') = \sigma^2 I_n$,$I_n$ 表示单位对角阵，同方差且不存在相关性；在此基础上可以做两点拓展：一是异方差假定，即 $E(\varepsilon_i\varepsilon_i') = \sigma_i I_n$，即主对角线元素不同；二是自相关问题，即 $E(\varepsilon_i\varepsilon_j) \neq 0$，即除主对角线元素以外还有非零元素；当存在自相关或者异方差问题时，需要考察 OLS 的估计量性质以及 GLS 估计量的性质；
3. X 是满秩矩阵，这意味着 X 不存在多重共线性问题；如果存在多重共线性，可以利用方差膨胀因子 VIF 进行判定，其中 $VIF = \frac{1}{1-R_j^2}$，R_j^2 表示除了 x_j 以外的其他变量 x_{-j} 回归到 x_j 得到的 R^2；

1.1.1 OLS 估计量

最小化误差平方项得到

$$\begin{aligned} min \sum e_i^2 &= min(Y - X\beta)'(Y - X\beta) \\ FOC.\ \frac{\partial(Y - X\beta)'(Y - X\beta)}{\partial\beta} &= -2X'(Y - X\beta) = 0 \\ \to \hat{\beta} &= (X'X)^{-1}(X'Y) \end{aligned}$$

[1]一般给定向量 X 为列向量，对于向量有 $X'X$ 表示加总 $\sum x_j^2$；XX' 表示矩阵；基本的线性代数矩阵运算需要掌握。

对于 OLS 模型，相关系数定义为 $R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$，其中 SSE 表示可由 X 解释的部分，SSR 表示由残差解释的部分。考虑到总是可通过加入更多的变量来提高模型的预测能力，因此引入调整后 $ad.R^2 = 1 - \frac{SSR/(n-k)}{SST/(n-1)}$，其中 k 为变量个数，此时即使加入更多的无关控制变量，导致 $n-k$ 增加但是模型解释力 SSR 不变，调整后 R^2 反倒降低。如果加入相关变量，调整后 R^2 的变化方向并不明确。[2]

1.1.2 投影矩阵与残差矩阵

定义投影矩阵 $P = X(X'X)^{-1}X'$，表示高维空间中任意曲线向 X 平面投影后的长度，PY 表示 Y 中与 X 相关的部分（可被 X 解释的部分）；投影矩阵性质如下：

1. $PX = X$；
2. $P \cdot P = P, P' = P$，投影矩阵为幂等阵且为对称矩阵；
3. $tr(P) = k$，证明 $tr(X(X'X)^{-1}X') = tr(X'X(X'X)^{-1}) = tr(I_k) = k$；

定义残差矩阵 $M = I - P = I - X(X'X)^{-1}X'$，表示高维空间中任意曲线向 X 平面投影中的垂线高度，MY 表示 Y 中与 X 不相关的部分（不可被 X 解释的部分），垂直项表明在 Y 中剔除了与 X 有关的信息；残差矩阵的性质：

1. $MX = 0$;
2. $M \cdot M = M, M' = M$，残差矩阵为幂等阵且为对称矩阵；
3. $tr(M) = n - k$；

重新使用投影矩阵和残差矩阵解释回归模型：[3]

$$Y = \hat{y} + \hat{\varepsilon} = PY + MY = X\hat{\beta} + e$$

其中 PY 表示 X 可以解释的部分，MY 表示 x 不可以解释的部分，三者构成了空间中的直角三角形（勾股定理）。在上述转化中，可以注意到如下关系：[4]

1. $X'e = 0$（相互垂直）；$MX = 0, MY = e, PX = X, PM = 0$;
2. $e'e = y'My = \varepsilon' M\varepsilon, MY = M\varepsilon$

1.1.3 *Partial Regression FWL*

给定多元回归

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon = X\beta + \varepsilon, X = (X_1, X_2), X' = \begin{bmatrix} X_1' \\ X_2' \end{bmatrix}$$

回归系数可以表示为

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}(X_1'M_2Y)$$
$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1}(X_2'M_1Y)$$

中级计量中给出 FWL 定理用于理解偏回归系数：以 $\hat{\beta}_1$ 为例，FWL 给出了三步走，首先将 X_1 回归到 X_2 上得到残差 M_2x_1；其次将 X_1 回归到 Y 上得到残差 M_2Y；最后将残差 M_2Y 回归到 M_2x_1 上得到 OLS 估计量，即 $\hat{\beta}_1 = (X_1'M_2'M_2X_1)^{-1}(X_1'M_2'M_2Y) = (X_1'M_2X_1)^{-1}(X_1'M_2Y)$。具体的含义是：为了得到干净的 X_1 的偏效应，必须要将 X_2 的影响剔除干净，避免 X_2 和 X_1, Y 相关带来的偏误，因此首先将 X_1, Y 分别回归到 X_2 得到残差，此时的残差中不包含任何其他信息，残差进行回归可以得到偏效应估计。

---

[2] 注意在带有截距项的模型中存在 $SST = SSR + SSE$，否则 $SST = SSR + SSE + 2\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$，存在误差。

[3] 为什么作上述变换呢？OLS 中最小化模型误差，为了是误差项最小，就需要保证是垂直投影，因此相应的拆解投影矩阵和误差矩阵，此时实现了 OLS 的几何解释。

[4] e 是回归中的残差项，ε 是模型中的随机误差项。

采用矩阵形式可以给定为：

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = (X'X)^{-1}(X'Y) = \begin{bmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{bmatrix}^{-1}\begin{bmatrix}X_1'Y\\ X_2'Y\end{bmatrix}$$

1.2 *OLS* 的 *BLUE* 性质

一般意义上，给定估计量 $\hat{\beta} = CY = CX\beta + C\varepsilon$，如果 BLUE 需满足如下条件（*W.Green*）：

1. 无偏性：$E(CY) = \beta, CX = I_k$，此时 $\hat{\beta}$ 是线性无偏估计量；

2. 有效性：$var(\hat{\beta}) = \sigma^2 CC'$，如果估计量是最佳线性无偏估计量 BLUE，则需满足对于任意的 C，均有 $var(\hat{\beta}) \leq var(CC')$

考察估计量 $\hat{\beta}$ 的无偏性：

$$\hat{\beta} = (X'X)^{-1}(X'Y) = \beta + (X'X)^{-1}(X'\varepsilon)$$
$$E(\hat{\beta}) = \beta + (X'X)^{-1}E(X'\varepsilon) = \beta$$

其次考察估计量 $\hat{\beta}$ 的方差：

$$var(\hat{\beta}) = (X'X)^{-1}X'var(\varepsilon)X(X'X)^{-1} = \sigma(X'X)^{-1}$$

现在证明 OLS 估计量的方差是最小的：对于 $(X'X)^{-1} \leq CC'$，等价于 $(X'X) \geq (CC')^{-1}$，也即 $(X'X - (CC')^{-1})$ 是半正定矩阵（PSD），给定无偏性条件 $CX = I$，得到 $X'X - X'C'(CC')^{-1}CX$ 是半正定矩阵，即 $x'[I - C'(CC')^{-1}C']X = X'M_C X$，$M_C$ 表示在 C 方向上投影的残差矩阵，根据几何性质可知矩阵半正定。这意味着，OLS 估计量是所有无偏估计量中方差最小的，因而是 $BLUE$。

进一步考察随机项方差的估计量

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum e_i^2 = \frac{e'e}{n-k}$$

下面证明估计量的无偏性质：

$$E(e'e) = E(y'My) = E(\varepsilon'M\varepsilon) = E(\sum_i \sum_j m_{ij}\varepsilon_i\varepsilon_j)$$

假定不存在自相关性 $cov(\varepsilon_i, \varepsilon_j) = 0$，上式进一步简化为

$$E(e'e) = E(\sum_i m_{ii}\varepsilon_i^2) = \sigma^2 E(\sum M_{ii}) = \sigma tr(M) = \sigma^2(n-k)$$
$$tr(M) = tr(I_n - X(X'X)^{-1}X') = n - tr(P) = n - k$$
$$\hat{\sigma}^2 = \frac{1}{n-k}e'e = \frac{1}{n-k}\sum e_i^2$$

相应的

$$E(\hat{\sigma}^2) = \frac{1}{n-k}E(e'e) = \sigma^2$$

1.3 *OLS* 统计推断

给定估计量和分布情况，可以从样本推断总体的性质，即假设检验。任何的估计都基于两个步骤：一是构造无偏且有效的估计量；二是估计量基础上的假设检验。小样本情况下一般需要给定正态分布假设，大样本基于大数定理和中心极限定理给出了渐进正态分布。本节利用小样本的正态分布假设进行检验。

1.3.1 正态分布假设

假定随机项服从正态分布：$\varepsilon \sim N(0,\sigma^2), E(\varepsilon)=0, E(\varepsilon'\varepsilon)=\sigma^2$，相应的可以得到 OLS 估计量的分布情况

$$\hat{\beta}=\beta+(X'X)^{-1}(X'\varepsilon)\sim N(\beta,\sigma^2(X'X)^{-1})$$

其中 $var((X'X)^{-1}(X'\varepsilon))=(X'X)^{-1}X'var(\varepsilon)X(X'X)^{-1}=\sigma^2(X'X)^{-1}$（三明治）。相应的

$$\hat{\sigma}^2=\frac{1}{n-k}\varepsilon' M\varepsilon=\frac{\sigma^2}{n-k}(\frac{\varepsilon}{\sigma}M)'(\frac{\varepsilon}{\sigma}M)\sim\frac{\sigma^2}{n-k}\chi^2_{n-k}$$

其中 $\frac{\varepsilon}{\sigma}\sim N(0,1)$,$M$ 控制检验自由度 $(n-k)$。

1.3.2 单个约束的 t 检验

给定单个方程系数检验：H_0:$\beta_{=}0$，相应的构造 t 检验

$$t=\frac{\hat{\beta}_k}{\sqrt{[\hat{\sigma}^2(X'X)^{-1}]_{kk}}}\sim t_{n-k}$$

注意到估计中采用的是估计值 $\hat{\sigma}^2$,如果有真实的 σ^2,可以直接进行 z 检验。一般的有 $\hat{\beta}_k\sim N(\beta_k,\sigma^2(X'X)^{-1}_{kk})$;对于上述问题转化为

$$t=\frac{(\hat{\beta}_k-\beta_k)/\sigma^2(X'X)^{-1}_{kk}}{\sqrt{\hat{\sigma}^2/\sigma^2}}$$

分子 $(\hat{\beta}_k-\beta_k)/\sigma^2(X'X)^{-1}_{kk}\sim N(0,1)$，分母 $\hat{\sigma}^2/\sigma^2\sim\frac{1}{n-k}\chi^2_{n-k}$，这是因为 $\hat{\sigma}^2\sim\frac{\sigma^2}{n-k}\chi^2_{n-k}$，分子为标准正态分布，分母为卡方分布开方，服从 t 分布。[5]

1.3.3 多个约束的 t 检验

给定单一条件检验：H_0:$r'\beta_{=}c$，其中 c 为常数，r 表示 $1\times k$ 的向量，例如条件 $\beta_1+\beta_2=1,\beta_1\neq 0$ 等，注意到这里只有一个约束，$\beta_1=\beta_2=\beta_3=0$ 的多重约束并不适用该方法。首先确定检验条件的分布情况：

$$r'\beta-c\sim N(r\beta-c,\sigma^2r'(X'X)^{-1}r)$$

构造 t 检验

$$\begin{aligned}t&=\frac{r'\beta-c}{\sqrt{[\hat{\sigma}^2r'(X'X)^{-1}r]}}\\&=\frac{[r'\beta-c]/[\sigma^2r'(X'X)^{-1}r]}{\sqrt{\hat{\sigma}^2/\sigma^2}}\\&\sim t_{n-k}\end{aligned}$$

1.3.4 多个约束的联合 F 检验

给定多重约束的联合检验：H_0:$R\beta=q$，其中 J 表示限制条件的个数。q 为 $J\times 1$ 的向量，R 为 $J\times k$ 矩阵，例如 $\beta_1=\beta_2=\beta_3=0$ 同时成立的联合检验，此时可以表述为如下形式（假定三个回归系数）

$$R\beta=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\\\beta_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}$$

[5]实际上，严格的证明服从 t 分布还需证明：分子与分母独立，参考（$W.Green$）给出的引理。

为什么不使用逐次的 t 检验？考虑如下两个问题：一是逐次进行 t 检验会导致第一类错误的概率更高，即更容易原假设，将估计结果高估为显著；二是如果存在多重共线性，逐次检验会导致单个检验的 *power* 不大，检验效力受限。此时采用联合 F 检验是更合理的选择。

给定约束条件的分布情况

$$R\hat{\beta}-q\sim N(R\beta-q,\sigma^2R(X'X)^{-1}R')$$

构造 F 检验

$$F=(R\hat{\beta}-q)'[\hat{\sigma}^2R(X'X)^{-1}R']^{-1}(R\hat{\beta}-q)/J\sim F_{J,n-k}$$

下面给出逐步证明：首先构造 *Wald* 检验，假定知道总体 σ^2

$$wald=(R\hat{\beta}-q)'[\sigma^2R(X'X)^{-1}R']^{-1}(R\hat{\beta}-q)=(\frac{R\hat{\beta}-q}{\sqrt{\sigma^2R(X'X)^{-1}R'}})'(\frac{R\hat{\beta}-q}{\sqrt{\sigma^2R(X'X)^{-1}R'}})\sim\chi_J^2$$

相应的 F 检验可以转化为

$$F=\frac{wald/J}{\hat{\sigma}^2/\sigma^2}\sim\frac{\chi_J^2/J}{\frac{1}{n-k}\chi_{n-k}^2}\sim F_{J,n-k}$$

其中分子分母相互独立。在大样本情况下，$\hat{\sigma}^2\to_p\sigma^2$，$F$ 检验收敛为 χ_J^2/J 检验。

1.3.5 F 检验与 R^2 关系

给定线性模型 $y=X\beta+\varepsilon$,$rank(X)=K$，同时 $\varepsilon\sim N(0,\sigma^2I_n)$. 构造假设检验 H_0:$R\beta=q$，下面讨论 F 检验与 R^2 关系：

已知 $R^2=1-SSE/SST$，其中

$$SST=Y'M_1Y=y'Qy/(n-K)$$
$$SSE=\hat{\sigma}^2(n-K)=e'e/(n-K)$$

因此转化为

$$\begin{aligned}R^2=&\frac{1}{y'Qy}[y'Qy-e'e]=\frac{1}{y'Qy}[y'Qy-y'My]\\=&\frac{1}{y'Qy}[y'QX_2(X_2'QX_2)^{-1}X_2'Qy]\end{aligned}$$

F 检验进一步表示为

$$\begin{aligned}F=&\frac{\hat{\beta}_2'(X_2'QX_2)^{-1}\hat{\beta}_2}{\hat{\sigma}^2(K-1)}\\=&\frac{(n-K)R^2}{(K-1)(1-R^2)}\end{aligned}$$

1.3.6 线性模型检验

给定线性模型

$$Y=X_1\underbrace{\beta_1}_{k_1\times1}+X_2\underbrace{\beta_2}_{k_2\times1}+\varepsilon=X\beta+\varepsilon,X=(X_1,X_2)$$

case 1：原假设为 H_0:$\beta_1 = 0$(k_1 个约束的联合检验)，构造 F 检验

$$\begin{aligned} F &= (R\hat{\beta} - q)'[\hat{\sigma}^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - q)/J \\ &= \hat{\beta}_1'[(I_{k_1}, 0)\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1}\begin{bmatrix} I_{k_1} \\ 0 \end{bmatrix}]^{-1}\hat{\beta}_1/k_1\hat{\sigma}^2 \\ &= \frac{\hat{\beta}_1'(X_1'M_2X_1)\hat{\beta}_1}{k_1\hat{\sigma}^2} \end{aligned}$$

其中 $[(I_{k_1}, 0)\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1}\begin{bmatrix} I_{k_1} \\ 0 \end{bmatrix}]$ 只需要考虑矩阵的左上角元素的逆，矩阵左上角元素的逆求解得到 $(X_1'M_2X_1)^{-1}$，代入估计中可以消除逆符号。

给定偏回归的系数，得到 $\hat{\beta}_1 = (X_1'M_2X_1)^{-1}(X_1'M_2Y)$,$\hat{\sigma}^2 = \frac{y'My}{n-k}$，代入检验中得到

$$\begin{aligned} F &= \frac{\hat{\beta}_1'(X_1'M_2X_1)\hat{\beta}_1}{k_1\hat{\sigma}^2} \\ &= \frac{n-k}{k_1}\frac{y'P_{1|2}y}{y'My} \sim F_{k_1, n-k_1-k_2} \end{aligned}$$

其中 $P_{1|2} = M_2X_1(X_1'M_2X_1)^{-1}X_1'M_2'$ 表示 X_1 对 X_2 回归得到的投影矩阵。

case 2：原假设为 H_0:$\beta_2 = 0$(k_2 个约束的联合检验)，构造 F 检验思路与上述基本类似，构造检验得到

$$\begin{aligned} F &= (R\hat{\beta} - q)'[\hat{\sigma}^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - q)/J \\ &= \hat{\beta}_2'[0, (I_{k_2})\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ I_{k_2} \end{bmatrix}]^{-1}\hat{\beta}_1/k_2\hat{\sigma}^2 \\ &= \frac{\hat{\beta}_2'(X_2'M_1X_2)\hat{\beta}_2}{k_2\hat{\sigma}^2} \\ &= \frac{y'M_1X_2(X_2'M_1X_2)X_2'M_1Y}{k_2\frac{y'My}{n-k}} \\ &= \frac{n-k}{k_2}\frac{y'P_{2|1}y}{y'My} \sim F_{k_2, n-k_1-k_2} \end{aligned}$$

其中 $P_{1|2} = M_1X_2(X_2'M_1X_2)^{-1}X_2'M_1'$ 表示 X_2 对 X_1 回归得到的投影矩阵。

1.4 大样本渐进性质

1.4.1 OLS 大样本渐进正态性质

考察 OLS 估计量的大样本渐进性质：

$$\begin{aligned} \hat{\beta} &= \beta + (X'X)^{-1}(X'\varepsilon) \\ &= \beta + (X'X/n)^{-1}(X'\varepsilon/n) \\ &\to_p \beta \end{aligned}$$

其中 $(X'X/n) \to_p Q = X'X, X'\varepsilon/n \to E(X'\varepsilon) = 0$。进一步考察渐进正态性质

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &= (X'X/n)^{-1}(X'\varepsilon/\sqrt{n}) \\ X'\varepsilon/\sqrt{n} &\sim N(0, \lim\frac{\sigma^2}{n}(X'X)) \\ \sqrt{n}(\hat{\beta} - \beta) &\sim N(0, \sigma^2(X'X)^{-1}) \\ \hat{\beta} &\sim N(\beta, \sigma^2(X'X)^{-1}/n) \end{aligned}$$

1.4.2 大样本假设检验

在大样本情况下进行统计推断，首先考察 t 检验

$$t = \frac{\hat{\beta}_k - \beta_k}{\sqrt{[\hat{\sigma}^2(X'X)^{-1}]_{kk}}}$$

大样本环境中有

$$\begin{aligned}\hat{\sigma}^2 = e'e/(n-k) &= \frac{n}{n-k}\frac{1}{n}(\varepsilon'\varepsilon - \varepsilon' M\varepsilon)\\ \frac{1}{n}\varepsilon'\varepsilon \to_p \sigma^2, &\frac{1}{n}\varepsilon' M\varepsilon \to_p 0\\ \to\hat{\sigma}^2 &\to_p \sigma^2\end{aligned}$$

此时 t 检验退化为 z 检验，得到

$$t = \frac{\hat{\beta}_k - \beta_k}{\sqrt{[\hat{\sigma}^2(X'X)^{-1}]_{kk}}} \sim_d z$$

进一步的考察联合 F 检验

$$F \sim \frac{\chi^2_J/J}{\chi^2_{n-k}/(n-k)}$$

考虑到 $\chi^2_{n-k}/(n-k) \to E(\chi^2_{n-k})/(n-k) = 1$，因此 F 检验退化为

$$F \sim_d \chi^2_J/J$$

1.5 *GLS Method*

给定线性模型，协方差矩阵表示为 $\Omega = E(\varepsilon\varepsilon') = E\begin{bmatrix}\varepsilon_1^2 & \dots & \varepsilon_1\varepsilon_n\\ \dots & \dots & \dots\\ \varepsilon_n\varepsilon_1 & \dots & \varepsilon_n^2\end{bmatrix}$，同方差假定意味着 $E(\varepsilon\varepsilon') = \sigma^2 I_n$,$I_n$ 表示单位对角阵，同方差且不存在相关性；在此基础上可以做两点拓展：一是异方差假定，即 $E(\varepsilon_i\varepsilon_i') = \sigma_i I_n$，即主对角线元素不同；二是自相关问题，即 $E(\varepsilon_i\varepsilon_j) \neq 0$，即除主对角线元素以外还有非零元素；当存在自相关或者异方差问题时，需要考察 OLS 的估计量性质以及 GLS 估计量的性质；

1.5.1 异方差的 OLS 估计

当存在异方差问题 $E(\varepsilon_i'\varepsilon_i) \neq E(\varepsilon_j'\varepsilon_j)$，此时考察 OLS 估计量的性质

$$\hat{\beta} = \beta + (X'X)^{-1}(X'\varepsilon)$$

如果满足 $E(\varepsilon|X) = 0$，OLS 估计量是无偏的；在证明大样本渐进性质时，考察

$$\begin{aligned}var((X'X)^{-1}(X'\varepsilon)) &= (X'X)^{-1}(X'\Omega X)(X'X)^{-1}\\ &= \frac{1}{n}(X'X/n)^{-1}(X'\Omega X/n)(X'X/n)^{-1}\end{aligned}$$

其中 $(X'\Omega X/n) = \frac{1}{n}\sum\sum x_i x_j'\sigma_{ij}$，如果 σ_{ij} 下降的足够快，那么 OLS 估计是一致的，但不是有效的。σ_{ij} 下降的足够快意味着距离较远的样本之间相关性很弱，这一般是满足的，OLS 估计量一般而言是一致的。

1.5.2 异方差的 GLS 估计

给定线性模型

$$Y = X\beta + \varepsilon$$

考虑到协方差矩阵结构为 Ω，做如下处理

$$\Omega^{-1/2}Y = \Omega^{-1/2}X\beta + \Omega^{-1/2}\varepsilon$$

此时有 $var(\Omega^{-1/2}\varepsilon) = \Omega^{-1/2}\Omega\Omega^{-1/2} = I_n$，此时不存在异方差问题，OLS 估计量是一致的，由此可以得到 GLS 估计量

$$\begin{aligned}\hat{\beta}_{GLS} &= [(\Omega^{-1/2}X)'(\Omega^{-1/2}X)]^{-1}[(\Omega^{-1/2}X)'(\Omega^{-1/2}Y)] \\ &= (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y)\end{aligned}$$

可以证明，GLS 估计量是异方差情形下的 BLUE（最优线性无偏估计量）：

$$\begin{aligned}E(\hat{\beta}) &= \beta + E[(X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}\varepsilon)] = \beta \\ var(\hat{\beta}) &= (X'\Omega^{-1}X)^{-1}\end{aligned}$$

OLS 和 GLS 在两种情况下等价：一是当 $\Omega = I_n$ 时两者等价，自然地不存在异方差两者等价，GLS 退化为 OLS；二是如果存在 Γ 满足 $\Omega X = X\Gamma$, 则 OLS 和 GLS 等价，证明如下：

$$X'\Omega^{-1} = (\Gamma^{-1})'X'$$

代入 *GLS* 估计量得到

$$\hat{\beta}_{GLS} = (\Gamma^{-1})'X'X)^{-1}(\Gamma^{-1})'X'Y) = (X'X)^{-1}(X'Y) = \hat{\beta}_{OLS}$$

进一步的，尽管 GLS 给出了更有效的估计方法，但是如何在样本中应用 GLS?

- *Infeasible GLS*: 已知协方差矩阵结构 Ω, 直接利用 GLS 进行估计 $\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y)$；但是一般情况下并不知道，因此该方法实际上并不可行；
- *Feasible GLS*：采用样本估计的协方差矩阵替代进行估计 $\hat{\beta}_{GLS} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}Y)$；

1.5.3 WLS 估计量

给定一种特殊情形，$\sigma_{ii} \neq \sigma_{jj}, i \neq j$ 且 $\sigma_{ij} = 0, i \neq j$，这表明协方差矩阵仅仅在主对角线上存在元素，但是主对角线上元素不相等，是一种简单的异方差结构

$$\Omega = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \dots & \sigma_i^2 & \dots \\ 0 & \dots & \sigma_n^2 \end{bmatrix}$$

此时 GLS 可以转化为 WLS（加权最小二乘法）。给定大样本下 OLS 估计量

$$var(\hat{\beta}) = \frac{1}{n}(\frac{\sum x_i x_i'}{n})^{-1}(\frac{\sum \sigma_i^2 x_i x_i'}{n})(\frac{\sum x_i x_i'}{n})^{-1}$$

注意到不能直接使用 e_i^2 估计 σ_i^2，无法保证一一对应。但是在大样本环境下，可以进行整体估计，即 $\hat{\Omega} = \frac{1}{n}\sum e_i^2 x_i x_i' \to \frac{1}{n}\sum \sigma_i^2 x_i x_i'$，此时 GLS 估计转化为

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y) = (\sum \frac{1}{\sigma_i^2} X_i X_i')^{-1}(\sum \frac{1}{\sigma_i^2} X_i Y_i')^{-1}$$

其中 $\frac{1}{\sigma_i^2}$ 作为权重对于样本进行加权处理。矩阵运算

$$X'\Omega^{-1}X = (X_1, \ldots, X_n)\begin{bmatrix} 1/\sigma_1^2 & \ldots & 0 \\ \ldots & 1/\sigma_i^2 & \ldots \\ 0 & \ldots & 1/\sigma_n^2 \end{bmatrix}\begin{bmatrix} X_1' \\ \ldots \\ X_n' \end{bmatrix} = \sum \frac{1}{\sigma_i^2} X_i X_i'$$

现在来理解 WLS 中权重到底体现了什么含义。对于线性模型

$$Y = X\beta + \varepsilon$$

$X\beta$ 可以视为信号（*signal*），ε 可以视为噪音（*noise*），如果 σ_i^2 上升表明模型中的噪音增加，信噪比降低，这个时候为了尽可能的利用多的信息，需要认为调低高噪音样本的权重，因此使用 $1/\sigma_i^2$ 进行加权以保证估计量的有效性。

当拿到数据，首先需要判断是够存在异方差问题，如果不存在异方差，使用 OLS 是合理的；如果存在异方差，则使用 WLS（GLS）或者在 OLS 情况下使用稳健标准误进行处理。可以说，WLS 是一种特殊的 GLS 方法，本身来自于特殊的协方差结构，从形式上可以进一步解释为权重。除了在方差结构中使用 WLS，使用样本变量（例如人口、规模等）作为权重的方法非常常见。也可以利用 WLS 进行估计。

异方差问题的检验与处理：如果存在异方差问题，假定异方差结构为 $\sigma_i^2 = h(z_i'\gamma) = exp(z_i'\gamma), ln(\sigma_i^2) = \alpha_0 + \alpha_1 y + \alpha_2 y^2 + v_i$，可以使用如下方法进行处理：

- 利用 $Y = X\beta + \varepsilon$ 得到回归残差 e_i；
- 将回归残差 e_i 回归到 y 上：$ln(e_i^2) = \alpha_0 + \alpha_1 y + \alpha_2 y^2 + v_i$，得到回归系数；
- 构造关于回归系数的假设检验：$\alpha_1 = \alpha_2 = 0$，如果不能拒绝原假设则不存在异方差问题，否则存在异方差问题；

如果已知存在异方差问题，那么可以利用 *step 1* 中得到的回归残差 e_i^2 作为 $\hat{\sigma}_i^2$ 的估计量，从而使用 *WLS* 进行估计。

1.5.4 自相关问题

以时间序列为例讨论自相关问题下 OLS 估计的问题以及解决方法。给定一般化的时间序列模型：

$$\begin{aligned} y_t &= x_t\beta + \varepsilon_t \\ \varepsilon_t &= \underbrace{\rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \cdots + \rho_p\varepsilon_{t-p}}_{AR(p)} \\ &\quad + \underbrace{r_1 u_{t-1} + r_2 u_{t-2} + \cdots + r_q u_{t-q}}_{MA(q)} \end{aligned}$$

对于 AR 过程而言，$|\rho| < 1$ 时收敛，$|\rho| > 1$ 时发散，$|\rho| = 1$ 时为随机游走（*random walk*）过程。简单起见，考察 $AR(1)$ 过程：

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, |\rho| < 1, u \sim N(0, \sigma_u^2)$$

首先确定 $AR(1)$ 过程的协方差结构：

$$\begin{aligned} cov(\varepsilon_1, \varepsilon_2) &= \rho var(\varepsilon_1) \\ \varepsilon_t &= \rho\varepsilon_{t-1} + u_t \\ &= \rho(\rho\varepsilon_{t-2} + u_{t-1}) + u_t \\ &= \rho^{m+1}\varepsilon_{t-(m+1)} + \rho^m u_{t-m} + \cdots + u_t \end{aligned}$$

其中 $\rho^{m+1}\varepsilon_{t-(m+1)} \to_p 0$，从而可以确定

$$E(\varepsilon_t^2) = \sigma_u^2(1+\rho^2+\rho^4+\cdots+\rho^{2m}) = \frac{\sigma_u^2}{1-\rho^2}$$
$$E(\varepsilon_t\varepsilon_{t-1}) = \rho E(\varepsilon_{t-1}^2) + Eu_t\varepsilon_{t-1} = \frac{\rho}{1-\rho^2}\sigma_u^2$$

其中假定 $Eu_t\varepsilon_{t-1}=0$，即 u_t 仅与 ε_t 有关而与 ε_{t-1} 无关。进一步的可以确定任意协方差

$$E(\varepsilon_t\varepsilon_s) = \frac{\rho^{t-s}}{1-\rho^2}\sigma_u^2, t>s$$

考虑到协方差矩阵的对称性，因此只需要求解 $t>s$ 单侧情形即可。相应的 $AR(1)$ 过程的协方差结构表示为

$$\Omega = \frac{\sigma_u^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{T-1} & \rho^{T-2} & \dots & \rho & 1 \end{bmatrix}$$

协方差结构表明存在自相关问题（此处尚未处理异方差问题，主对角线元素保持一致）。对于该问题，使用 GLS 估计是合理的，下面给出两种 GLS 的处理方法。

Quasi-Difference Method：基本思路是考虑一阶差分，对模型进行一阶差分之后可以得到同方差结构，从而 OLS 估计量是有效的：

$$y_t - \rho y_{t-1} = (x_t - \rho x_{t-1})\beta + (\varepsilon_t - \rho\varepsilon_{t-1})$$
$$y_t^* = x_t^*\beta + u_t$$

考虑到 u_t 是同方差的随机项，可以得到 OLS 估计量 $\hat{\beta} = (x_t^{*'}x_t^*)^{-1}(x_t^{*'}y_t^*)$。但是该方法存在问题：对于第一期的数据无法进行差分因而难以估计第一期参数，缺少第 1 期信息。

Prais-Winsten Method：根据 GLS 的基本思路，将方差结构中的 $\frac{1}{1-\rho^2}$ 进行处理

$$\sqrt{1-\rho^2}y_t = \sqrt{1-\rho^2}x_t\beta + \sqrt{1-\rho^2}\varepsilon_t$$

此时 $var(\sqrt{1-\rho^2}\varepsilon_t) = \sigma_u^2$。对于矩阵形式可以做如下处理：

$$pY = pX\beta + p\varepsilon$$
$$p = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & \dots & 0 \\ -\rho & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$
$$var(p\varepsilon) = p\Omega p' = \sigma^2 I_n$$

可以得到 GLS 估计量 $\hat{\beta} = (X'p'pX)^{-1}(X'p'pY)$。该方法有效利用了所有期的信息，因此可以得到全部参数的一致估计。

$AR(2)$ 过程的 *GLS* 估计思路：给定 $AR(2)$ 过程

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + u_t, |\rho|<1, u \sim N(0,\sigma_u^2)$$

首先确定 $AR(2)$ 过程的协方差结构，之后利用上述两种 GLS 方法的任意一种可以确定 GLS 估计量。

如果想要在序列相关的情况下使用 OLS 估计，如何得到稳健标准误呢？给定估计量方差

$$var(\hat{\beta}) = (X'X)^{-1}X'\Omega X(X'X)^{-1} = (X'X)^{-1}(\sum \sigma_{ij}x_ix_j')(X'X)^{-1}$$

这里不能使用 e_{ij} 单个估计 σ_{ij}，而是采用 e_ie_j 代替估计 σ_{ij}。此时的含义是伴随着 ij 距离的增加相应的权重随之下降，如果相差太远则人为去除。

给出较为正式的证明：*Newey and West(1987,1991)* 分别给出了如下关系

$$\begin{aligned}\frac{1}{n}X'\Omega X &= \frac{1}{n}(\sum \sigma_{ij}x_ix_j') \\ &= \frac{1}{n}(\sum e_i^2x_ix_j')\tfrac{1}{n}\sum_{l}^{L}\sum_{i=L+1}^{n}(1-\frac{l}{L+1})e_ie_{i-l}(x_ix_{i-l}'+x_{i-l}x_i')\end{aligned}$$

当 $i-j>L$ 时则 $\sigma_{ij}\to 0$，表明距离过远则人为去掉，这是因为距离较远的情况下一般相关性较弱是成立的。其中 L 是具有相关性的范围，权重 $1-\frac{l}{L+1}$ 表示距离越近权重越大。一般而言 $L=O(T^{1/4})$, 例如样本量为 $T=10000$, 相应的相关性范围为 10 的整数倍。

1.5.5 序列相关的检验与处理：DW 检验

对于时间序列模型的应用首先要判定是否存在序列相关，构造 DW 检验：原假设为 H_0：$\rho=0$，不存在序列相关，统计量构造为

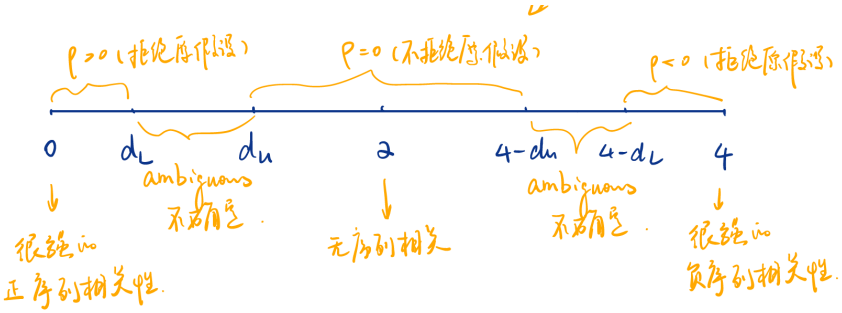
$$\begin{aligned}\mathcal{DW} &= \frac{\sum^T(e_t-e_{t-1})^2}{\sum^T e_{t-1}^2} \\ &= [\sum^T e_t^2+\sum^T e_{t-1}^2-2\sum^T e_te_{t-1}]/\sum^T e_{t-1}^2 \\ &= 2-2\frac{\sum^T e_te_{t-1}}{\sum^T e_{t-1}^2} \\ &\to_p 2(1-\rho)\in[0,4]\end{aligned}$$

其中 $\sum^T e_t^2\neq\sum^T e_{t-1}^2$，但是在 $T\to\infty$ 时近似相等。相应的，原假设可以转化为如下关系：

1. 如果 $\rho=1$，则 $\mathcal{DW}=0$, 表示完全的正序列相关；
2. 如果 $\rho=0$，则 $\mathcal{DW}=2$，不存在序列相关；
3. 如果 $\rho=-1$，则 $\mathcal{DW}=4$，表示存在完全的负序列相关；

给定如下判定法则：

- 如果 $\rho\in(d_u,4-d_u)$，相应的 $\mathcal{DW}$ 在 2 附近，不能拒绝原假设，表明不存在显著的序列相关；
- 如果 $\rho\in(0,d_l)$ 或者 $\rho\in(4-d_l,4)$, 相应的 $\mathcal{DW}$ 远离 2，表明存在显著的序列相关，如果 $\mathcal{DW}$ 靠近 4 表明存在负序列相关，如果 $\mathcal{DW}$ 靠近 0 表明存在正序列相关；
- 如果 $\rho\in(d_l,d_u)$ 或者 $\rho\in(4-d_u,4-d_l)$，则不能确定是否存在序列相关；



1.6 似不相关回归 *SUR*

Seemingly Unrelated Regressions 顾名思义就是对两个或多个表面上看起来没有关系的方程进行联合估计（连享会）。SUR 的基本思想是同时估计多个回归方程，例如男性和女性的消费函数，如果男性和女性两者消费不存在相关性，则直接使用 OLS 单个估计即可；但是一般而言男女性消费存在显著的相关性，因而单个估计会受到影响，估计不再是有效的（标准误偏高）；此时使用联合估计方法是比较合理的。区别于 SEM 联立方程模型中 Y 出现在等式右侧，SUR 中 Y 只出现在等式左侧。一旦考虑到来联合估计，就是为了避免方程组之间相关性导致的估计偏误问题。

给定一般形式的 SUR 问题：

$$
\begin{aligned}
Y_1 &= X_1\beta + u_1 \\
&\cdots \\
Y_m &= X_m\beta + u_m
\end{aligned}
$$

给定上述 m 个方程组，例如 Y_1 是对于水的需求，Y_m 是对于汽油的需求，X 表示自变量（不同方程间可以相同可以不同），考虑到消费者对于两类商品的需求存在相关性，因此分别估计 OLS 存在偏误，采用 SUR 更加有效，实际上 SUR 也更多的应用于居民们对各种食物的需求体系。

给定协方差结构

$$
E(u_i u_i') = \sigma_{ii} I_n, E(u_i u_j') = \sigma_{ij} I_n, E(u_i | X_I) = 0
$$

矩阵形式表示为

$$
\begin{aligned}
Y &= X\beta + u \\
\begin{bmatrix} Y_1 \\ \cdots \\ Y_m \end{bmatrix} &= \begin{bmatrix} X_1 & & \\ & \cdots & \\ & & X_m \end{bmatrix} \begin{bmatrix} \beta_1 \\ \cdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} u_1 \\ \cdots \\ u_m \end{bmatrix} \\
\Omega = E(uu') &= \begin{bmatrix} \sigma_{11} I_n & \cdots & \sigma_{1m} I_n \\ \cdots & \cdots & \cdots \\ \sigma_{m1} I_n & \cdots & \sigma_{mm} I_n \end{bmatrix} = \Sigma_m \otimes I_n
\end{aligned}
$$

其中 $\otimes$ 表示 *Kronecker* 积。构造 GLS 估计量

$$
\begin{aligned}
\hat{\beta}_{GLS} &= (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y) \\
&= [\begin{bmatrix} X_1' & & \\ & \cdots & \\ & & X_m' \end{bmatrix} (\Sigma_m^{-1} \otimes I_n) \begin{bmatrix} X_1 & & \\ & \cdots & \\ & & X_m \end{bmatrix}]^{-1} [\begin{bmatrix} X_1' & & \\ & \cdots & \\ & & X_m' \end{bmatrix} (\Sigma_m^{-1} \otimes I_n) \begin{bmatrix} Y_1 \\ \cdots \\ Y_m \end{bmatrix}]
\end{aligned}
$$

如果不存在相关性，则该问题简化为

$$
\Sigma_m = \begin{bmatrix} \sigma_{11} & \cdots & 0 \\ \cdots & \sigma_{ii} & \cdots \\ 0 & \cdots & \sigma_{mm} \end{bmatrix}
$$

代入 GLS 估计量得到

$$
\hat{\beta}_{GLS} = \begin{bmatrix} (X_1'X_1)^{-1}(X_1'Y_1) \\ \cdots \\ (X_m'X_m)^{-1}(X_m'Y_m) \end{bmatrix} = \hat{\beta}_{OLS}
$$

即如果不存在相关性问题，GLS 和 OLS 具有等价性（当然异方差可以进一步简化为 WLS）。

SUR 中可行 FGLS 给定为：首先，利用 OLS 估计每一个方程并得到残差，进一步计算得到协方差结构 $\hat{\sigma}_{ij} = \frac{e_i' e_j}{n}$。其次，代入可行 GLS 估计

$$\begin{aligned}\hat{\beta}_{GLS} &= (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}Y) \\ &= (X'(\hat{\Sigma}_m^{-1} \otimes I_n)X)^{-1}(X'(\hat{\Sigma}_m^{-1} \otimes I_n)Y)\end{aligned}$$

考虑在 SUR 系统中 OLS 与 GLS 的等价性，存在两种情形：一是方程间不存在相关性可以退化为 OLS 处理；二是 SUR 中自变量全部一致，从而可以使用 OLS 进行估计。严格的证明使用如下定理：如果给定矩阵 Γ 满足 $\Omega\Gamma = X\Gamma$，则 *OLS* 与 *GLS* 估计量是等价的。

对于第 *1* 种不存在相关性的情形，容易得到

$$\begin{bmatrix}\sigma_{11}I_n & \dots & 0 \\ \dots & \sigma_{ii} & \dots \\ 0 & \dots & \sigma_{mm}I_n\end{bmatrix}\begin{bmatrix}X_1 & & \\ & \dots & \\ & & X_m\end{bmatrix} = \begin{bmatrix}X_1 & & \\ & \dots & \\ & & X_m\end{bmatrix}\Gamma$$

可以得到

$$\Gamma = \begin{bmatrix}\sigma_{11}I_n & \dots & 0 \\ \dots & \sigma_{ii} & \dots \\ 0 & \dots & \sigma_{mm}I_n\end{bmatrix} = \Omega$$

因此 *OLS* 和 *GLS* 是等价的。

对于情形 *2*, 全部自变量一致的情形

$$\begin{bmatrix}X_1 & & \\ & \dots & \\ & & X_m\end{bmatrix} = \begin{bmatrix}X_c & & \\ & \dots & \\ & & X_c\end{bmatrix} = I_m \otimes X_c$$

从而可以得到

$$\begin{aligned}(\Sigma_m \otimes I_n)(I_m \otimes X_c) = \Sigma_m \otimes X_c = (I_m \otimes X_c)\Gamma \\ \Gamma = (\Sigma_m \otimes I_k)\end{aligned}$$

其中 k 表示自变量的维度，此时 *OLS* 和 *GLS* 是等价的

2 Maximum Likelihood Method

1). MLE 估计量的基本性质，识别与估计，Information Matrix

2). MLE 的大样本渐进性质

3). MLE 的三大统计检验：Wald 检验、LM 检验以及 LRT 检验

4). 线性模型的 MLE（给定参数分布，直接利用 MLE 方法估计参数）

5). Newton-Raphson 最速下降法（数值解法）、NR 方法的 MATLAB 编程

6). 高级计量结合陈强的《高级计量经济学与 Stata 应用》code

2.1 *MLE* 基本性质

给定参数的分布，可以采用 *MLE* 方法估计任意形式回归模型的参数（包括线性和非线性），前提是只要给定参数的分布假设；因而在结构估计中只要施加了参数分布假设，就可以采用 *MLE* 方法估计相应的参数；

***MLE* 的思路**：给定参数分布，一个自然的想法就是利用可以关测到的数据去尽量贴合分布状况，因而可以最优化参数从而使得可观测数据的分布与参数分布尽可能相似，也就是最大化概率乘积以寻找最优的拟合参数，这就是按分布拟合的 *MLE* 方法。

MLE* 的基本性质与处理方法 *Transformation

$$f(y_1,\ldots,y_n|\theta)=\Pi f(y_i|\theta)=L(\theta|Y)$$
$$lnL(\theta|Y)=\sum f(y_i|\theta)$$

其中，参数 θ 表示参数集合，例如正态分布中为 $\theta=\{\beta,\sigma^2\}$。在这定义中，前者给定的是给定参数 θ 的情况下，实际观测到的数据分布概率 $f(y|\theta)$；在转化为似然函数的时候，需要求解的是根据可观测的数据分布，最优化参数，因而需要求解的函数就是给定现实可观测的数据分布，寻找最优的参数 θ 来尽可能好的拟合，这就将给定参数分布下真实数据分布转换为给定真实分布下参数的优化问题，从而可以使用最大化来估计。

$$max_\theta\ L(\theta|Y)\rightarrow max_\theta \ln[L(\theta|Y)]$$

离散形式和连续形式的 *MLE* 方法

- 连续形式：概率密度函数 PDF $\Pi f(x_i)$ 乘积；离散形式：概率函数乘积 $\Pi prob(x_i\leq X)$
- 连续形式和离散形式的混合 $\Pi_1 f(x_i)\Pi_2 prob(x_i\leq X)$，离散的时候用概率，连续的时候用概率密度函数

Quasi-MLE：当不知道参数分布的时候，利用正态分布估计和识别模型中的参数（在大样本中考虑到渐进正态性质是可行的）

Identification：在何种情况下说 *MLE* 能够识别出 θ_0（真值）? 给定识别条件，即 $Eln(\theta)<Eln(\theta_0),\theta\neq\theta_0$，即最大化问题存在解

- 更具体的可以展开为一阶条件：$\partial L(\hat{\theta})/\partial\theta=0, E(\frac{\partial L(\theta_0)}{\partial\theta})=0$，这是 *MLE* 能够识别出 θ_0 的必要条件（可能存在多个解因而不是充分条件）
- 如果 *FOC* 可以唯一的确定参数 θ，那么实现了 *Identification*，如果对应多个参数，则无法判断到底哪一个是正确的估计量，因而无法识别（识别本身的含义就是唯一确定参数）

Information Identify：似然函数关于参数的二阶导矩阵，类似于 *Hessian Matrix*

$$E(\frac{\partial L(\theta_0)}{\partial \theta} \cdot \frac{\partial L(\theta_0)}{\partial \theta'}) = -E(\frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'}) \tag{2.1}$$

其中 θ_0 表示参数的真实值，前者表示最大似然函数的信息矩阵（似然函数一阶导的方差）表达式，该式的含义是可以用似然函数的二阶导来替换 *Information Matrix* 的求解。基本含义是：似然函数一阶导的方差等于二阶导期望的负数。

Information Identify（二阶条件）保证了 $\hat{\theta}$ 可以识别 θ_0，一阶条件 FOC 保证了必要性，两者保证了 *MLE* 可以识别唯一的参数 θ_0。

2.2 *MLE* 大样本渐进性质

一致性

$\frac{1}{n}L(\hat{\theta}_{MLE}) - E(\frac{1}{n}L(\theta_0)) \to_p 0$，根据 *Identification* 条件可以判定 θ_0 最大化期望函数，FOC 保证收敛到 0，进一步的该式表明 $\hat{\theta}_{MLE} \to_p \theta_0$.

渐进正态性质

一是确定 *MLE* 估计量的渐进正态性质；二是比较 *MLE* 估计量与 OLS 估计量的大样本渐进性质（判断估计量的优劣）。给定一阶条件

$$\begin{aligned}
0 = \frac{\partial L(\hat{\theta})}{\partial \theta} &= \underbrace{\frac{\partial L(\theta_0)}{\partial \theta} + \frac{\partial^2 L(\bar{\theta})}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0)}_{Taylor\ Extension} \\
\hat{\theta} - \theta_0 &= (-\frac{\partial^2 L(\bar{\theta})}{\partial \theta \partial \theta'})^{-1} \frac{\partial L(\theta_0)}{\partial \theta} \\
\sqrt{n}(\theta - \theta_0) &= (-\frac{1}{n}\frac{\partial^2 L(\bar{\theta})}{\partial \theta \partial \theta'})^{-1} (\frac{1}{\sqrt{n}}\frac{\partial L(\theta_0)}{\partial \theta})
\end{aligned}$$

其中等式 (2.5) 表示 FOC 在真实值 θ_0 附近一阶展开，$\bar{\theta}$ 表示 $(\theta_0, \hat{\theta})$ 的拉格朗日中值。等式 (2.6) 给出了渐进分布的基本形式，同时这个也是在 NR 方法中的基本搜索方程。等式 (2.7) 进一步给出 *MLE* 的大样本渐进正态分布形式：

$$\begin{aligned}
(\frac{1}{n}\frac{\partial^2 L(\bar{\theta})}{\partial \theta \partial \theta'}) &\to_p E(\frac{1}{n}\frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'}) \\
(\frac{1}{\sqrt{n}}\frac{\partial L(\theta_0)}{\partial \theta}) &\to_d N(0, \frac{1}{n}E(\frac{\partial L(\theta_0)}{\partial \theta} \cdot \frac{\partial L(\theta_0)}{\partial \theta'})) \\
\hat{\theta} - \theta_0 &\to_d (E(\frac{1}{n}\frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'}))^{-1} N(0, \frac{1}{n}E(\frac{\partial L(\theta_0)}{\partial \theta} \cdot \frac{\partial L(\theta_0)}{\partial \theta'}))
\end{aligned}$$

By Information Identify

$$\hat{\theta} - \theta_0 \to_d N(0, -(E(\lim \frac{1}{n}\frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'}))^{-1}) \tag{2.2}$$

据此给出 *MLE* 的大样本渐进正态分布，其中 $E(\hat{\theta}_{MLE}) = \theta_0, var(\hat{\theta} - \theta_0) = [-\frac{1}{n}E\frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'}]^{-1}$.

Cramer-Rao Lower Bound (参数估计下界问题)

给定 $\hat{\theta} = argmin\, var(\hat{\theta} - \theta_0)$ 可以确定方差最小(最有效率)的估计量,在 *MLE* 估计中,有 $var(\hat{\theta}_{MLE} - \theta_0) \leq var(\hat{\theta}_{others} - \theta_0)$。如果可以证明某个无偏估计量的方差等于 CRLB，那么它就是 BUE。除此以外，CRLB 再不能给更多东西了CRLB 数学推导。

在 *MLE* 的估计中，CRLB 等价于 $I^{-1}(\theta)$，同时 *MLE* 参数估计量 $var(\hat{\theta})$ 等于该下界，可以证明 *MLE* 估计是最有效的。

2.3 线性模型的 *MLE* 估计量

给定线性模型 $y = x\beta+\varepsilon$, $\varepsilon \sim N(0,\sigma^2 I)$,可以确定被解释变量的条件分布为 $y|\varepsilon \sim N(x\beta,\sigma^2 I)$, $f(y_i|\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-x_i\beta)^2}{2\sigma^2}}$, 通过上述转换可以确定如下最大似然函数:

$$
\begin{aligned}
L(\beta,\sigma^2) &= \Pi f(y_i|\varepsilon_i) \\
lnL &= -\frac{n}{2}ln2\pi - \frac{n}{2}ln\sigma^2 - \frac{1}{\sigma^2}(y-x\beta)(y-x\beta)' \\
\partial L/\partial\beta &= 0 \\
\partial L/\partial\sigma^2 &= 0 \\
\hat{\beta}_{MLE} &= (X'X)^{-1}(X'Y) = \hat{\beta}_{ols} \\
\hat{\sigma}^2_{MLE} &= \frac{e'e}{n} \neq \hat{\sigma}^2_{ols} = \frac{e'e}{n-k}
\end{aligned}
$$

如上利用 *MLE* 得到了线性模型的估计量，$\hat{\beta}$ 的估计量是无偏且一致的，但是 $\hat{\sigma}^2$ 是一致但有偏的(OLS 估计是无偏且一致的)。注意到，$\hat{\sigma}^2_{MLE} < \hat{\sigma}_{ols}$，采用 *MLE* 方法低估了数据的方差，导致更容易出现显著性（过度拒绝原假设的问题)。

方差估计的有偏性：$E(\hat{\sigma}^2_{MLE}) = \frac{n-k}{n}E(\varepsilon'\varepsilon) = \frac{n-k}{n}\sigma^2$ ，小样本中是有偏的。

进一步的，需要确定 *Information Matrix*，即明确估计量方差的相关性问题；信息矩阵包含参数的方差和协方差关系，注意到信息矩阵中 $L(\theta_0)$ 代入的是真实值:

$$
\begin{aligned}
E(-\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'}) &= \begin{bmatrix} \frac{\partial^2 L}{\partial\beta^2} & \frac{\partial^2 L}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 L}{\partial\beta\partial\sigma^2} & \frac{\partial^2 L}{\partial(\sigma^2)^2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma_0^2}X'X & E\frac{1}{\sigma_0^4}X'\varepsilon \\ E\frac{1}{\sigma_0^4}X'\varepsilon & -\frac{n}{2\sigma_0^4} + \frac{1}{\sigma_0^6}E(\varepsilon'\varepsilon) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma_0^2}X'X & 0 \\ 0 & \frac{n}{2\sigma_0^4} \end{bmatrix} \\
[E(-\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'})]^{-1} &= \begin{bmatrix} \sigma_0^2(X'X)^{-1} & 0 \\ 0 & 2\sigma_0^4/n \end{bmatrix}
\end{aligned}
$$

等式 1 根据 *Information Matrix* 的基本定义进行展开（采用了 *Information Identify*)；等式 3 协方差项全部等于 0，因为 $E(X'\varepsilon) = 0$，等式 4 有 $E(\varepsilon'\varepsilon) = n\sigma_0^2$。进一步写出参数的分布情况，即

$$
\sqrt{n}\begin{bmatrix} \hat{\beta}_{MLE} - \beta_0 \\ \hat{\sigma}^2_{MLE} - \sigma_0^2 \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma_0^2 \lim[\frac{1}{n}(X'X)]^{-1} & 0 \\ 0 & 2\sigma_0^4 \end{bmatrix})
$$

一旦确定了信息矩阵，就可以明确 *MLE* 估计量的联合分布情况。

2.4 无约束检验 *Wald Test*

给定检验：$H_0:\ h(\theta) = 0, h: R^h \to R^q$，如何根据 *MLE* 估计量构造假设检验?

首先明确，*unconstrained model* 是指没有任何约束条件（没有 H_0 ）的原模型，*constrained model* 是指原假设成立情况下的约束模型；区分这两者并构造三种检验，其原因在于有的时候，约束条件很复杂，约束优化问题很难求解，而直接求解原问题很容易求解，那么使用原问题估计量构造 Wald test 就很方便；如果说原问题很难优化，但是将约束条件代入后极大地简化了 model，很多参数都被消除，更容易求解估计量，那么使用约束模型构造 LM 检验就很方便。其基本思路就是求解约束（无约束）优化的 *MLE* 估计量，构造假设检验量进行检验。关键一点在于构造检验量并明确分布。

何种情况下使用 *Wald Test*：当约束条件较为复杂（特别是包含非线性的约束，例如 $\beta_1\beta_2+e^{\beta_3}=0$），直接求解原来的约束优化问题比较简单，这个时候确定无约束参数 $\hat{\theta}$，进一步构造 $h(\hat{\theta})$ 的分布以及检验量，来判定是否足够接近于零（检验约束条件是否真的成立）。

基本思路：在无约束的情况下求解最优化问题得到 *MLE* 的估计 $\hat{\theta}$，检验在这种情况下 $h(\hat{\theta})$ 是否足够接近 0，如果显著区别于 0 则表明原假设被拒绝，约束条件不等于 0。

给定无约束情况下的 *MLE* 估计量 $\hat{\theta}$，构造 Wald 检验（q 个约束条件）：

$$h'(\hat{\theta})[\frac{\partial h(\hat{\theta})}{\partial \theta'}(-E\frac{\partial^2 L(\hat{\theta})}{\partial\theta\partial\theta'})^{-1}\frac{\partial h'(\hat{\theta})}{\partial\theta}]^{-1}h(\hat{\theta})\sim\chi_q^2$$

下面讨论具体的检验量构造过程，已经知道了 *MLE* 的渐进分布，现在只需要明确的是 $h(\theta)$ 的分布，根据 Delta Method 或者 Taylor 一阶展开，可以得到

$$h(\hat{\theta})\sim N(h(\theta_0),[\frac{\partial h(\hat{\theta})}{\partial\theta'}var(\hat{\theta})\frac{\partial h'(\hat{\theta})}{\partial\theta}])\sim N(0,[\frac{\partial h(\hat{\theta})}{\partial\theta'}(-E\frac{\partial^2 L(\hat{\theta})}{\partial\theta\partial\theta'})^{-1}\frac{\partial h'(\hat{\theta})}{\partial\theta}])$$

如上便得到了相应的约束条件的分布情况，下面构造 Wald Test, 在 OLS 中的 F 检验量或者 Wald Test 构造方法是：$\frac{(R\beta-q)'var(\hat{\beta})^{-1}(R\beta-q)}{J}$，同样的在这里构造 Wald Test (经典的三明治形式)，同时使用估计值 $\hat{\theta}$ 来表示真实值，代入即可得到 *Wald Test*。

2.5 约束检验 *LM Test*

何种情况下使用 *LM Test*：原始函数形式较为复杂，例如 $y_i=\beta_0+\beta_1x_1+\beta_2x^2+\beta_3x^3+\varepsilon$，此时如果给定原假设 $\beta_2=\beta_3=0$，将其作为已知条件代入方程可以简化得到 $y=\beta_0+\beta_1x$；LM 的基本思路就是在给定 H0 可以极大简化函数形式的时候，利用一阶条件构造估计量，考察在 H0 成立的条件下一阶条件是不是得到满足。

基本思路：在约束情况下求解最优化问题得到 *MLE* 的估计 $\bar{\theta}$，检验这种情况下 $\frac{\partial lnL(\bar{\theta})}{\partial\theta}$ 是否足够接近 0，如果足够接近于 0 则接受原假设，否则则拒绝原假设。

给定约束条件下 *MLE* 估计量 $\bar{\theta}$，构造 LM 检验（q 个约束条件）：

$$\frac{\partial lnL(\bar{\theta})}{\partial\theta'}[var(\frac{\partial lnL(\bar{\theta})}{\partial\theta})]^{-1}\frac{\partial lnL'(\bar{\theta})}{\partial\theta}$$
$$\frac{\partial lnL(\bar{\theta})}{\partial\theta'}[-E(\frac{\partial^2 lnL(\bar{\theta})}{\partial\theta\partial\theta'})]^{-1}\frac{\partial lnL'(\bar{\theta})}{\partial\theta}\sim\chi_q^2$$

在这个过程中，首先需要确定无约束情况下的 *MLE* 估计量。其次，在原始的 *MLE* 估计中，需要得到一阶和二阶条件，并用于构造 LM 检验，最终信息矩阵和一阶条件中采用无约束的 *MLE* 估计量。

2.6 混合检验 *LR Test*

何种情况下使用 *LR Test*：对于约束优化和无约束优化都比较简单的形式，例如 $y=\beta_1x_1+\beta_2x_2$，原假设为 $\beta_2=0$，此时可以得到无约束优化估计量 $\hat{\theta}$ 和约束优化估计量 $\bar{\theta}$，从而考察两种情况下似然函数是不是足够接近，如果两种情况下的似然函数足够接近，则接受原假设，否则拒绝。

给定无约束优化估计量 $\hat{\theta}$ 和约束优化估计量 $\bar{\theta}$，构造 LR 检验（q 个约束条件）：

$$lnL(\hat{\theta})-lnL(\bar{\theta})$$

如果说两个估计量基本一致，那么可以预期 $lnL(\hat{\theta})-lnL(\bar{\theta})=0$。下面需要确定估计量的分布函数：在真值 θ_0 附近二阶展开得到

$$Q(\theta)=Q(\theta_0)+\frac{\partial Q(\theta_0)}{\partial\theta}(\theta-\theta_0)+\frac{1}{2}\frac{\partial^2 Q(\theta_0)}{\partial\theta\partial\theta'}(\theta-\theta_0)^2$$

定义函数 $Q(\theta) = lnL(\hat{\theta}) - lnL(\bar{\theta})$, 相应的可以得到估计量:

$$2(lnL(\hat{\theta}) - lnL(\bar{\theta})) \sim \chi_q^2$$

2.7 线性模型的三大检验

以线性模型为例, 给出 *MLE* 三大检验的具体形式: 给定线性模型 $y = X_1\beta_1 + X_2\beta_2 + \varepsilon, \varepsilon \sim N(0, \sigma^2)$, 其中 X_1, X_2 均为满秩矩阵, $rank(X_1) = k_1, rank(X_2) = k_2$。原假设 H_0: $\beta_2 = 0$。

2.7.1 Wald Test

无约束情况下, $y \sim N(X_1\beta_1 + X_2\beta_2, \sigma^2)$, 构造似然函数得到:

$$L(\beta_1, \beta_2, \sigma^2) = \Pi^n f(y_i|\beta, \sigma^2) = \Pi^n \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y_i - X_{1i}\beta_1 - X_{2i}\beta_2)^2}{2\sigma^2})$$
$$lnL(\beta_1, \beta_2, \sigma^2) = -\frac{n}{2}ln2\pi - -\frac{n}{2}ln\sigma^2 - \frac{1}{2\sigma^2}(y - X_1\beta_1 - X_2\beta_2)'(y - X_1\beta_1 - X_2\beta_2)$$

最优化似然函数的一阶条件表示为:

$$\frac{\partial L}{\partial \beta_1} = \frac{1}{\sigma^2}X_1'(y - X_1\beta_1 - X_2\beta_2) = 0$$
$$\frac{\partial L}{\partial \beta_2} = \frac{1}{\sigma^2}X_2'(y - X_1\beta_1 - X_2\beta_2) = 0$$
$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\varepsilon'\varepsilon = 0$$

根据一阶条件可以得到:

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}(X_1'M_2y)$$
$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1}(X_2'M_1y)$$
$$\hat{\sigma}^2 = \frac{e'e}{n} = \frac{y'My}{n}$$

定义 $\theta = (\beta, \sigma^2)$。进一步的, 确定 *Information Matrix*, 根据一阶条件可以得到

$$\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'} = \begin{bmatrix} -\frac{1}{\sigma_0^2}X'X & -\frac{1}{\sigma_0^2}X'(Y - X\beta_0) \\ -\frac{1}{\sigma_0^2}X'(Y - X\beta_0) & \frac{n}{2\sigma_0^4} - \frac{1}{\sigma_0^6}\varepsilon'\varepsilon \end{bmatrix}$$
$$= \begin{bmatrix} -\frac{1}{\sigma_0^2}X'X & 0 \\ 0 & \frac{n}{2\sigma_0^4} - \frac{1}{\sigma_0^6}\varepsilon'\varepsilon \end{bmatrix}$$
$$-E(\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'}) = \begin{bmatrix} \frac{1}{\sigma_0^2}X'X & 0 \\ 0 & \frac{n}{2\sigma_0^4} \end{bmatrix}$$

其中，β_0, σ_0^2 表示参数真值，$X = (X_1, X_2), X'X = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}$。$(X'X)^{-1} = \begin{bmatrix} * & * \\ * & (X_2'M_1X_2)^{-1} \end{bmatrix}$。

构造 *Wald* 统计量：

$$\begin{aligned}
\mathcal{WT} &= \hat{\beta}_2'[(0,1,0)[-E(\frac{\partial^2 L(\hat{\theta})}{\partial\theta\partial\theta'})]^{-1}(0,1,0)']^{-1}\hat{\beta}_2 \\
&\to \hat{\beta}_2'[[0,1,0]\begin{bmatrix} \frac{1}{\hat{\sigma}^2}X'X & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}]^{-1}\hat{\beta}_2 \\
&\to \frac{\hat{\beta}_2'(X_2'M_1X_2)\hat{\beta}_2}{\hat{\sigma}^2} \\
&\to n\frac{[(X_2'M_1X_2)^{-1}(X_2'M_1y)]'(X_2'M_1X_2)[(X_2'M_1X_2)^{-1}(X_2'M_1y)]]}{y'My} \\
&\to \frac{y'P_{x_{2|1}}y}{y'My/n} \sim \chi_{k_2}^2
\end{aligned}$$

其中，$P_{x_{2|1}} = (M_1X_2)[X_2'M_1X_2']^{-1}(M_1X_2)'$，表示 x_2 对 x_1 回归的投影矩阵。

2.7.2 LM Test

约束情况下，$y \sim N(X_1\beta_1, \sigma^2)$, 最优化似然函数的一阶条件表示为：

$$\begin{aligned}
\frac{\partial L}{\partial \beta_1} &= \frac{1}{\sigma^2}X_1'(y - X_1\beta_1 - X_2\beta_2) = 0 \\
\frac{\partial L}{\partial \beta_2} &= \frac{1}{\sigma^2}X_2'(y - X_1\beta_1 - X_2\beta_2) = 0 \\
\frac{\partial L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\varepsilon'\varepsilon = 0
\end{aligned}$$

根据一阶条件可以得到：

$$\begin{aligned}
\bar{\beta}_1 &= (X_1'X_1)^{-1}(X_1'y) \\
\bar{\beta}_2 &= 0 \\
\bar{\sigma}^2 &= \frac{e'e}{n} = \frac{y'M_1y}{n}
\end{aligned}$$

定义 $\theta = (\beta, \sigma^2)$。相应的将 *MLE* 估计量代入一阶条件得到 $\frac{\partial lnL(\bar{\theta})}{\partial\theta}$，显然在存在约束情况下一阶条件不能全部得到满足：

$$\begin{aligned}
\frac{\partial L(\bar{\theta})}{\partial \beta_1} &= \frac{1}{\sigma^2}X_1'e = 0 \\
\frac{\partial L(\bar{\theta})}{\partial \beta_2} &= \frac{1}{\sigma^2}X_2'M_1Y \\
\frac{\partial L(\bar{\theta})}{\partial \sigma^2} &= -\frac{n}{2\bar{\sigma}^2} + \frac{1}{2\sigma^4}e'e = 0
\end{aligned}$$

进一步的，确定 *Information Matrix*, 根据一阶条件可以得到

$$\begin{aligned}
\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'} &= \begin{bmatrix} -\frac{1}{\sigma_0^2}X'X & -\frac{1}{\sigma_0^2}X'(Y - X\beta_0) \\ -\frac{1}{\sigma_0^2}X'(Y - X\beta_0) & \frac{n}{2\sigma_0^4} - \frac{1}{\sigma_0^6}\varepsilon'\varepsilon \end{bmatrix} \\
&= \begin{bmatrix} -\frac{1}{\sigma_0^2}X'X & 0 \\ 0 & \frac{n}{2\sigma_0^4} - \frac{1}{\sigma_0^6}\varepsilon'\varepsilon \end{bmatrix} \\
-E(\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'}) &= \begin{bmatrix} \frac{1}{\sigma_0^2}X'X & 0 \\ 0 & \frac{n}{2\sigma_0^4} \end{bmatrix}
\end{aligned}$$

构造 LM 检验量，首先 $\frac{\partial lnL(\bar{\theta})}{\partial \theta}=(0,\frac{1}{\bar{\sigma}^2}X_2'M_1Y,0)$，$LM$ 统计量构造如下：

$$\begin{aligned}\mathcal{LM}&=(0,\frac{1}{\bar{\sigma}^2}X_2'M_1Y,0)'\begin{bmatrix}\frac{1}{\bar{\sigma}^2}X'X & 0\\ 0 & \frac{n}{2\bar{\sigma}^4}\end{bmatrix}^{-1}\begin{bmatrix}0\\ \frac{1}{\bar{\sigma}^2}X_2'M_1Y\\ 0\end{bmatrix}\\&=\frac{1}{\bar{\sigma}^2}Y'M_1X_2(X_2'M_1X_2)^{-1}X_2'M_1Y\\&=\frac{n}{YM_1Y}Y'P_{X2|1}Y\sim\chi^2_{k_2}\end{aligned}$$

其中，$P_{x_{2|1}}=(M_1X_2)[X_2'M_1X_2']^{-1}(M_1X_2)'$，表示 x_2 对 x_1 回归的投影矩阵。

2.7.3 *LR Test*

无约束条件下有

$$lnL(\hat{\theta})=-\frac{n}{2}ln2\pi--\frac{n}{2}ln\hat{\sigma}^2-\frac{1}{2\hat{\sigma}^2}(y-X_1\hat{\beta}_1-X_2\hat{\beta}_2)'(y-X_1\hat{\beta}_1-X_2\hat{\beta}_2)$$

其中 $\hat{\sigma}^2=\frac{y'My}{n}$。

对于约束条件下有

$$lnL(\bar{\theta})=-\frac{n}{2}ln2\pi--\frac{n}{2}ln\bar{\sigma}^2-\frac{1}{2\bar{\sigma}^2}(y-X_1\bar{\beta}_1-X_2\bar{\beta}_2)'(y-X_1\bar{\beta}_1-X_2\bar{\beta}_2)$$

其中 $\bar{\sigma}^2=\frac{y'M_1y}{n}$。

构造 LRT 检验量：

$$\begin{aligned}\mathcal{LRT}&=2(lnL(\hat{\theta})-lnL(\bar{\theta}))=nln(\frac{\bar{\sigma}^2}{\hat{\sigma}^2})\\&=nln(\frac{y'M_1y}{y'My})\sim\chi^2_{k_2}\end{aligned}$$

2.8 *Newton-Rapson Method*

如果给定的最大似然函数非常复杂，无法直接通过简单的最优化求解显式解，此时需要利用数值迭代算法进行求解，此处提供最简单的 *Newton-Rapson Method*。该方法的基本思路是按照梯度方向进行搜索，导数中沿梯度方向下降最快，因而沿梯度方向搜索是最快的路径。函数 $Q(\theta)$ 在初始值附近 θ_0 二阶线性展开可以得到：

$$Q(\theta)=Q(\theta_1)+\frac{\partial Q(\theta_1)}{\partial\theta}(\theta-\theta_1)+\frac{1}{2}\frac{\partial^2Q(\theta_1)}{\partial\theta\partial\theta'}(\theta-\theta_1)^2$$

等式左右侧对 θ 求导得到：

$$\begin{aligned}\frac{\partial Q(\theta)}{\partial\theta}&=\frac{\partial Q(\theta_1)}{\partial\theta}+\frac{\partial^2Q(\theta_1)}{\partial\theta\partial\theta'}(\theta-\theta_1)\\\theta&=\theta_1+[-\frac{\partial^2Q(\theta_1)}{\partial\theta\partial\theta'}]^{-1}\frac{\partial Q(\theta_1)}{\partial\theta}\end{aligned}$$

上式的基本含义表示为：给定初始值 θ_1，不断的利用等式 (2.67) 进行迭代，其中搜索方向为一阶导 $\frac{\partial Q(\theta_1)}{\partial\theta}$，调整因子为二阶导的倒数 $[-\frac{\partial^2Q(\theta_1)}{\partial\theta\partial\theta'}]^{-1}$。这就保证，如果初始值小于最优值，那么梯度向量大于 0，沿着最优值的方向进行迭代。

Newton-Rapson Method 的迭代思路如下：

1. 给定 j 步初始值 $\hat{\theta}_j$；
2. 根据迭代关系计算 $\hat{\theta}_{j+1}=\hat{\theta}_j+[-\frac{\partial^2Q(\hat{\theta}_j)}{\partial\theta\partial\theta'}]^{-1}\frac{\partial Q(\hat{\theta}_j)}{\partial\theta}$，得到 $j+1$ 步估计值 θ_{j+1}；
3. 如果 $|\theta_{j+1}-\theta_j|\le\varepsilon$，停止搜索；其中 ε 表示任意规定的最小收敛距离。

DI

3 *Discrete Choice Model*

Outline: 本部分所有模型估计采用 *MLE*（给定参数分布–*CDF* 或者 *PDF*–采用 *MLE* 可以得到参数估计值）。首先要搞清楚离散选择模型中存在什么问题，利用 *OLS* 估计有什么问题，如何转化为 *MLE* 估计？*OLS* 存在偏误的解释，在统计中可以解释为存在其他项干扰了估计量，在因果推断中解释为数据的内生性缺失（*Sample Selection* 问题）。

1). 离散选择模型 *Probit* 和 *Logit* 模型：对应的问题，需要处理的具体场景，模型设定，*MLE* 估计方法

2). *Truncation Data* 和 *Censored Data Model*：数据存在的问题、*OLS* 估计存在的偏误、模型设定、*MLE* 估计方法

3). *Sample Selection*

3.1 *LDV* 受限因变量

举个简单的例子，考试成绩有的人考了 100 分，有些人是只有 100 分的能力，但是有的人是有 200 分的水平，而只有 100 分的总成绩。首先因变量的意思是可观测的数据无法反映真实情况，因而被解释变量是 *limited*。受限因变量是无处不在的，但是在这里只关注 *dummy* 的情况，即转化为概率模型。

考虑线性回归模型：$I_i = x_i\beta + \varepsilon_i$，$I_i$ 表示工作与否的 0-1 变量，那么在这里就可以处理为 $P(I_i = 1) = x_i\beta$，转化为线性概率模型，估计的是参加工作（取值为 1）的概率，一个自然而然的问题就是，使用 OLS 估计如上线性概率模型，是否是合适的呢?

(1). 采用线性概率模型估计的一个自然问题就是存在 $x_i\beta$ 大于 1 或者小于 0 的情形，这不符合实际的概率情况，简单的处理是进行 censor（缩尾取值）

(2). 考察方差 $var(\varepsilon_i|x_i) = p(1-p) = x_i\beta(1-x_i\beta)$，存在异方差问题（使用 GLS 或者异方差稳健标准误处理）

(3). **关键问题**：*partial effect* 依赖于 x_i，即 $\partial P/\partial x = f(x_i\beta)\beta$，这种情况下偏效应对于不同人群具有不同的结果

综上，采用 OLS 估计线性概率模型不是一个好的选择。自然地，有必要将模型推广到非线性模型，给定参数的分布假设，采用 MLE 的方法进行估计，这个是更为合理的。

3.2 *Discrete Choice Model*

离散选择模型的基本设定如下：

- 给定 $E(y_i = 1) = F(x_i\beta)$（CDF 分布函数），相应的 $E(y_i = 0) = 1 - F(x_i\beta)$
- $0 < Prob(y_i = 1) < 1$，$var(y_i) = F(1-F) > 0$，$\partial Prob(y_i = 1)/\partial x = f(x_i\beta)\beta$ 偏效应同时包含参数与观测值
- 给定如下数据，$u_0 = v_0 + e_0$，$u_1 = v_1 + e_1$，前者表示不参加工作的效用，后者表示参加工作的效用，v 表示可观测变量，e 表示不可观测变量，做如下转换：$u_1 - u_0 = (v_1 - v_0) + (e_1 - e_0) \to y^* = x\beta + u$，该式表示真实的数据生成过程

给定真实数据生成过程 $y^* = X'\beta + u$，可观测的数据（从真实数据到可观测数据的处理过程）

$$y = \begin{cases} 1 & y^* > 0 \\ 0 & y^* \leq 0 \end{cases}$$

统计含义在于知道真实的数据生成过程以及参数 u 的分布情况，如果有全部的数据，那么就可以利用 OLS 估计；但是现在观察到的数据是二元离散数据，该二元离散数据是真实数据生成过程的一种处理结果，那么直接估计 $y = X'\beta + \varepsilon$ 是有偏误的吗?

给定 $F(X'\beta)$，偏效应等价于 $\frac{\partial F(X'\beta)}{\partial_k} = f(X'\beta)\beta_k$，这表明偏效应不等于固定的常数，而是依赖于 X 及其分布，采用 OLS 估计无法准确的识别 β，这是离散选择模型和后续的 *truncation data* 和 *censored data* 使用 OLS 中存在的问题（其实给定了参数分布，使用 MLE 是大概率的选择!）

一般情况下，针对二元离散选择模型，**采用 Probit 和 Logit 模型**，他们关于参数的分布假定存在差异：

- *Probit*：假定参数服从标准正态分布，给定 CDF $F(u) = \int \phi(u)du, \phi(w) = \frac{1}{\sqrt{2\pi}} exp(-w^2/2)$
- *Logit*：假定参数服从 logistics 分布，给定 CDF $F(u) = \frac{e^u}{1+e^u}$，$var(u) = \frac{\pi^3}{3}$

在这里，的基本思路是：给定了参数的 CDF 分布，那么就可以根据 CDF 估计参数值，例如给定 $u \sim N(0, \sigma^2)$，那么相应的 DGP 服从 $y^* \sim N(X'\beta, \sigma^2)$，既然给定了分布就可以利用 MLE 估计，后面处理这类问题模型和问题数据的时候，都可以使用 MLE 方法。当然，关于参数不同的分布假设得到的估计结果是不同的，关于偏效应系数估计也存在差异，即 $\beta_{probit} \neq \beta_{logit}$。

3.3 Logit 和 Probit 模型

确定最大似然函数

$$\begin{aligned} L(\beta) &= \Pi_{y_i=1} Prob(y_i = 1) \Pi_{y_i=0} Prob(y_i = 0) = \Pi_{y_i=1} F(X'\beta) \Pi_{y_i=0} (1 - F(X'\beta)) \\ lnL(\beta) &= \begin{cases} \sum lnF(X'\beta) & y_i = 1 \\ \sum ln(1 - F(X'\beta)) & y_i = 0 \end{cases} \\ &= \sum [y_i lnF(X'\beta) + (1 - y_i) ln(1 - F(X'\beta))] \\ F.O.C.\ \frac{\partial lnL}{\partial \beta} &= \sum [(\frac{y_i}{F(X'\beta)} - \frac{1 - y_i}{1 - F(X'\beta)}) f(X_i\beta) X_i] = 0 \end{aligned}$$

其中 F 表示 CDF，而 f 表示 PDF。

基本的思路：

1. 确定最大似然函数（对数形式）：在这里实际上就是取值为 1 和取值为 0 的离散形式的分布函数乘积；$L(\beta) = \Pi_{y_i=1} F(X'\beta) \Pi_{y_i=0} (1 - F(X'\beta))$，可以看出在离散数据的处理中，CDF 函数是至关重要的，决定了如何估计参数的估计值（MLE）；两类非线性概率模型的唯一差异在于对于参数的假设；
2. MLE 求解思路：FOC 一阶条件确定参数的取值，借助 FOC 可以确定 $\hat{\beta}_{mle}$，并进行假设检验；
3. Newton-Raphson 迭代计算方法：当无法直接得到参数的显式解的时候，可以采用梯度下降算法寻找数值解，迭代过程如下

$$\hat{\beta}_{j+1} = \hat{\beta}_j - (\frac{ln^2 L(\hat{\beta})}{\partial\beta\partial\beta'})^{-1} \frac{\partial L(\hat{\beta})}{\partial\beta}$$

非线性概率模型的拟合度问题

参考 R^2 的处理，在非线性概率模型中引入 pseudo R^2 的概念，定义为：$pR^2 = \frac{lnL - lnL_0}{lnL_{max} - lnL_0} = 1 - \frac{lnL}{lnL_0}$，其中 lnL 表示最大似然函数的实际对应值，lnL_0 表示不放入任何解释变量的对数函数值（对应的 y_i 估

计值就是 y_i 的均值），lnL_{max} 表示最大解释力度（取 1），其含义就是说的模型相比于不加入任何解释变量的模型的解释力度（整体的解释力度是 1 减去不加入任何解释变量模型的解释力度）。

该比例简单的告诉了模型的解释力度（都是相对于不加入任何解释变量–也就是均值作为估计量模型–的解释力度）。

3.4 *Truncation Data* 截断数据

首先区分截断数据（*truncated data*）和缩尾数据（*censor data*）：简单的理解是截断数据完全不包含某类数值以外的数据（y, x 都是），缩尾数据可以观测到 x 但是对于 y 的观测是被处理的。

截断数据处理的基本思路：截断数据的一个基本问题就是超过某个数值或者小于某个数值的数据缺失，这部分数据缺失很有可能是因为内生性缺失，例如高收入群体不愿意报告自己的收入从而完全无法观测这部分数据；那么给定参数的分布，实际上可以对估计方法做一些调整来处理这类问题，其基本的思路是说现在观测到的最大似然函数适用于观测到的样本区间，例如 $y_i \le c_i$，因此需要对于这个项进行调整，除以该区间在整个的参数分布区间上的概率 $Prob(\varepsilon \le c_i)$，该调整就是将观测到的截断数据最大似然函数推广到整个分布上面，解决高收入群体收入无法观测的问题。

对于不存在截断数据的分布，有 $h(y_i|x_i\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y_i - x_i'\beta)^2}{2\sigma^2})$，该式可以进一步化简为标准正态下的情形（标准正态转换），即

$$h(y_i|x_i\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}) = \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}} exp(-\frac{(\frac{y_i - x_i'\beta}{\sigma})^2}{2}) = \frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma})$$

不存在截断情况下的最大似然估计表示为：$L = \Pi h(y_i|x_i'\beta)$。

对于存在截断的数据，采用如下调整形式–MLE 估计方法：首先给定模型设定，$y_i = X'\beta + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$，只能观测到 $y_i \le c_i$ 的数据（高位截断），下面开始构造最大似然函数，首先确定可观测数据的 PDF，即

$$f(y_i|x_i\beta, c_i) = \frac{h(y_i|x_i\beta, c_i)}{Prob(y_i \le c_i)} = \frac{h(y_i|x_i\beta, c_i)}{\Phi(y_i|x_i\beta, c_i)} = \frac{\frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma})}{\Phi(\frac{c_i - x_i'\beta}{\sigma})}$$

其中，h 是 *original data* 的概率密度函数，Φ 和 ϕ 分别是标准正态分布的 CDF 和 PDF（MLE 只需要知道参数分布就可以构造统计量）。注意到这里是简单的概率处理，分母的含义是截断数据在全样本分布中的比例，即 $prob(y_i \le c_i) = prob(\varepsilon_i \le c_i - x_i'\beta) = \Phi(\frac{c_i - x_i'\beta}{\sigma})$；同理，如果是 $c_1 \le y_i \le c_2$，可以得到 $\Phi(\frac{c_2 - x_i'\beta}{\sigma}) - \Phi(\frac{c_1 - x_i'\beta}{\sigma})$。

一旦确定了概率分布密度，就可以构造最大似然函数：$lnL(\beta, \sigma) = \sum f(y_i|x_i\beta, c_i)$，根据一阶条件可以确定参数估计量。FOC 存在显式解的时候求解显式解；没有显式解的时候根据 *Newton-Raphson* 方法求解数值解。

三种形式的调整项：

1. $y_i < c_i$：$prob(y_i \le c_i) = prob(\varepsilon_i \le c_i - x_i'\beta) = \Phi(\frac{c_i - x_i'\beta}{\sigma})$
2. $y_i > c_i$：$prob(y_i > c_i) = prob(\varepsilon_i > c_i - x_i'\beta) = 1 - \Phi(\frac{c_i - x_i'\beta}{\sigma})$
3. $c_1 \le y_i \le c_i$：$prob(c_1 \le y_i \le c_2) = \Phi(\frac{c_2 - x_i'\beta}{\sigma}) - \Phi(\frac{c_1 - x_i'\beta}{\sigma})$

3.5 *Censored Data* 缩尾数据

给定观测数据

$$y = \begin{cases} y_i^* & y^* > 0 \\ 0 & y^* \le 0 \end{cases}$$

其中，$y_i^* \sim N(x_i'\beta, \sigma^2)$ 。

这是一个显然的缩尾数据，体现在如果真实数据生成过程是负数，那么相应的实际数据处理为 0，例如国际贸易中所有的负值处理为 0；这意味着观测到的数据并非都是真实的数据，尽管可以观测到数据信息，但是信息是有限的，特别是存在如上二元缩尾处理。这种情况下，也需要假定参数的分布情况并根据参数分布确定 MLE 估计量。

$$
\begin{aligned}
prob(y_i = 0) &= prob(y_i^* < 0) = prob(x_i\beta + \varepsilon_i < 0) = \Phi(-\frac{x_i'\beta}{\sigma}) \\
prob(y_i > 0) &= prob(y_i^* > 0) = \frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma}) \\
f(y_i) &= [\frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma})]^{1(y_i>0)}[\Phi(-\frac{x_i'\beta}{\sigma})]^{1(y_i=0)} \\
lnL(\beta,\sigma) &= \sum_{y_i>0} ln[\frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma})] + \sum_{y_i=0} ln[\Phi(-\frac{x_i'\beta}{\sigma})]
\end{aligned}
$$

根据 MLE 可以确定估计量 $\hat{\theta} = (\hat{\beta}, \hat{\sigma})$。其中 $y_i = 0$ 时为离散情况（概率函数），$y_i > 0$ 时为连续情况（PDF）。同理，可以确定任意缩尾门槛值 c_i 的极大似然函数。

3.6 *Truncation Data* 和 *Censored Data* 估计量

基本的处理逻辑：截断数据是数据处理过程中存在部分缺失值，所以需要使用截断部分在总分布中的比例进行调整，其余部分保持不变；缩尾数据的问题在于部分数据被 Group 成某个数值，所以在求解最大似然函数的过程中需要确定某个具体数值处的概率分布，从而叠加生成最大似然函数（类似于 Mixture 概率分布函数），利用 *Index Function* 进行调节。

3.6.1 OLS 估计下的偏误

1. 截断数据存在的问题就是部分数据观测不到导致的潜在内生性问题，造成了估计偏误
2. 缩尾数据存在的问题更大（在 MLE 估计量的处理上更为明显），几乎无法识别出待估计参数

3.6.2 模型设定

要研究的问题是，如果真实的数据生成过程是 $y_i = x_i'\beta + \varepsilon_i$，那么是否可以用 OLS 估计 $y_i = x_i'\beta + u_i$ 估计得到真实的参数 β?

给定可观测数据

$$
y = \begin{cases} y_i^* & y^* > 0 \\ 0 & y^* \le 0 \end{cases}
$$

如果观测到如下形式，表明存在缩尾问题（Tobit I model：被解释变量 y_i 是被选择的，但是其选择和本身是否大于 0 有关）

$$
\begin{cases} (y_i^*, x_i) & y^* > 0 \\ (0, x_i) & y^* \le 0 \end{cases}
$$

如果观测到如下形式，表明存在截断问题，即只能观测到 (y_i^*, x_i) 当且仅当 $y_i^* > 0$（低位截断）。

首先分析截断数据的 OLS 估计存在的问题（基本的思路就是概率转化：将可观测数据转化为真实的数据，并根据分布假设确定函数形式）

$$
E(y_i) = E(y_i^*|y_i^* > 0) = x_i'\beta + E(\varepsilon_i|\varepsilon_i > -x_i\beta) = x_i'\beta + \sigma\frac{\phi(\frac{x_i'\beta}{\sigma})}{\Phi(\frac{x_i'\beta}{\sigma})} \neq x_i'\beta
$$

其中 $\frac{\phi(\frac{x_i'\beta}{\sigma})}{\Phi(\frac{x_i'\beta}{\sigma})}$ 为 *Inverse Mills Ratio*（*IMR* 来自于条件期望的求解公式，参考 Green 教材），因此在截断数据中使用 OLS 估计存在很大的偏误，这个偏误取决于分布的相关参数。当然，如果对于被截断的数据（$y_i^* \leq 0$）不感兴趣，可以直接使用截断的数据进行回归。

其次分析缩尾数据的 OLS 估计存在的问题

$$\begin{aligned} E(y_i) &= prob(y_i = 0)0 + prob(y_i > 0)E(y_i|y_i > 0) = prob(\varepsilon_i > -x_i'\beta)E(y_i^*|\varepsilon_i > -x_i'\beta) \\ &= \Phi(\frac{x_i'\beta}{\sigma})x_i'\beta + \sigma\phi(\frac{x_i'\beta}{\sigma}) \end{aligned}$$

其中，概率 $prob(\varepsilon_i > -x_i'\beta) = \Phi(\frac{x_i'\beta}{\sigma})$，后者期望则可以根据截断数据的表达式代入得到，在缩尾数据的期望中可以明显的看到不等于实际生成过程，甚至很难直接去识别 β, σ。

3.6.3 两种数据估计量处理

1. 给定截断数据的最大似然函数：$L_1 = \Pi_{y_i>0}[Prob(y_i > 0)]^{-1} f(y_i)$ （其中分母表示在截断数据中的调整项）

2. 给定缩尾数据的最大似然函数：

$$\begin{aligned} L_2 &= \Pi_{y_i=0} Prob(y_i = 0)\Pi_{y_i>0} f(y_i) \\ &= \underbrace{\Pi_{y_i=0} Prob(y_i = 0)\Pi_{y_i=1} Prob(y_i > 0)}_{Probit} \underbrace{\Pi_{y_i=1}[Prob(y_i > 0)]^{-1} f(y_i)}_{Truncated} \end{aligned}$$

前两者的乘积是 *Probit* 模型估计量，后两者的乘积是 L_1（截断数据估计量）

3.6.4 *Heckman* 两阶段估计

如上模型为提供了利用截断数据估计 OLS 的调整方法，即 *Heckman* 两阶段：

Step 1. 利用全样本估计 *Probit* 模型，根据 *MLE* 得到 $(\frac{\hat{\beta}}{\sigma})$ 的估计量；

Step 2. 估计每个样本的 $\hat{\lambda}_i = \frac{\phi(\frac{x_i'\beta}{\sigma})}{\Phi(\frac{x_i'\beta}{\sigma})}$ 作为调整项，利用截断数据估计方程：$y_i = x_i'\beta + \sigma\lambda_i + \eta_i$，可以估计得到 $(\hat{\beta}, \hat{\sigma})$；

基本思想是在全样本中估计受选择样本的决定方程，将其作为调整项（*IMR*）放入存在选择偏误的回归方程中修正，从而得到 OLS 的无偏一致估计。

3.7 *Sample Selection: Tobit II Model*

Truncation Data 和 *Censored Data* 中的问题是无法观测到所有可以观测的数据，适应于第一类 *Tobit* 模型。现在讨论**样本选择**问题，最常见的例子就是已婚女性的工资决定问题，数据中可以发现只有部分已婚女性进入劳动力市场，其余部分已婚女性并未进入劳动力市场从而无法观测其劳动收入。问题在于，已婚女性是否进入劳动市场是内生选择的过程，受到众多变量的影响，导致可观测数据中劳动收入的缺失数据并不是随机的，而是存在选择性，接下来考察这种选择性带来的影响以及对应的措施：*Tobit II* 模型或者 *Heckman* 两步走矫正法。

3.7.1 样本选择的模型设定

给定 y_i 不可直接观测，例如所有已婚女性的公司水平，实际的数据生成满足

$$y_i^* = x_i\beta + \varepsilon_{1i}$$

给定样本选择方程：

$$d_i = 1(z_i\gamma + \varepsilon_{0i} > 0)$$

该式表示已婚女性是否进入劳动市场的决策，当进入劳动力市场的效用高于不进入的效用时已婚女性选择进入劳动力市场。其中 z_i 表示可观测变量，$\varepsilon_1, \varepsilon_0$ 表示不可观测变量。z_i 和 x_i 都表示了可观测变量，在样本选择模型中，一般要求 ***exclusion condition*** 满足 z_i 和 x_i 不完全一致，即决定女性进入市场的因素和女性收入的决定因素不完全相同，否则在估计中可能会导致完全共线性问题（不过考虑到 *IMR* 是非线性函数，因此影响不大）。另外从样本的角度来看，假定全样本为 1000 个，女性进入市场的样本为 800 个，那么要求利用的 z_i 是全样本数据 1000 个，从而估计 d_i 的概率模型；在工资决定过程中，y_i, x_i 使用的是可观测到的 800 个样本数据。

相应的可观测的样本数据 y_i 表示为：

$$y_i = \begin{cases} y_i^* & d_i = 1 \\ 0 & d_i = 0 \end{cases}$$

该处和截断数据的差别在于：数据内生于样本选择而导致无法观测，主导因素是样本选择方程。对于残差项，假定

$$(\varepsilon_0, \varepsilon_1)' \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix})$$

其中 $cov(\varepsilon_0\varepsilon_i) = \rho$ 表明回归方程和样本选择的不可观测变量存在相关性。

3.7.2 样本选择下 OLS 估计的偏误

只有在已婚女性进入劳动市场之后才能观测到劳动收入，给定条件期望函数

$$\begin{aligned} E(y_i|d_i = 1) &= x_i'\beta + E(\varepsilon_{1i}|\varepsilon_{0i} > -z_i\gamma) \\ &= x_i'\beta + \rho\frac{\phi(z_i\gamma)}{\Phi(z_i\gamma)} == x_i'\beta + \rho\lambda_i \\ &\neq x_i'\beta \end{aligned}$$

这表明使用 OLS 估计存在样本选择的方程是存在偏误的，偏误的方向性取决于 ρ，即工资决定和就业决策中不可观测变量之间的相关性方向。接下来，给定了误差项的概率分布函数，可以利用 *MLE* 进行估计（比较复杂）；同时可以利用 *Heckman* 两阶段修正工资决定方程后利用 OLS 进行估计。下面介绍 *Heckman* 两阶段的步骤和基本思路。

3.7.3 *Heckman* 两阶段

Heckman 两阶段步骤如下：

step 1. 在全样本数据（1000）中利用 *Probit* 或者 *Logit* 估计选择方程：$d_i = 1(z_i\gamma + \varepsilon_{0i})$，得到估计值 $\hat{d}_i$，代入 *IMR* 中得到 $\hat{\lambda} = \frac{\phi(\hat{d}_i)}{\Phi(\hat{d}_i)}$；

step 2. 在选择性样本（800）中利用 OLS 估计如下方程：$y_i = x_i\beta + \rho\hat{\lambda}_i + \eta_i$，得到 x_i 一致估计；

其中 *IMR* 作为控制变量加入原来的工资决定方程中，矫正存在的样本选择偏误，关心的是核心解释变量的系数；如果估计结果表明 *IMR* 显著，则说明存在明显的样本选择问题，两阶段估计系数和 OLS 系数相差较为明显；如果 *IMR* 不显著，则说明样本选择问题并不明显。

接下来，需要思考 *Heckman* 两阶段究竟意味着什么？*The solution here is therefore to predict the likelihood of participation in the labor force at first stage using a probit model and the exclusion restriction*

(the same criteria for valid instruments apply here), calculate the predicted inverse Mills ratio for each observation, and in second stage, estimate the wage offer using the as a predictor in the model (Wooldridge 2009). If the coefficient on $\hat{\lambda}$is statistically equal to zero, there is no evidence of sample selection (endogeneity), and OLS results are consistent and can be presented. If the coefficient on $\hat{\lambda}$is statistically significantly different from zero, you will need to report the coefficients from the corrected model.

对于存在选择性偏差的样本，既然直接利用 OLS 估计存在偏误，那么就需要一种方法来修正模型中存在的偏误。如何修正呢？首先第一步需要考察到底是什么样的个体更容易进入劳动力市场，即估计个体决策方程；当一旦估计得到了个体进入市场的选择方程，就可以进一步估计逆米尔斯比率，它反映了对于样本选择性偏误的修正，或者说在全样本中每个样本所占比例作为修正（类似 *truncated data* 修正项）。既然存在这种分布上的选择性，接下来控制这种选择性，此时估计 OLS 就不再是有偏的了。如果 IMR 显著说明进入市场的个体确实存在显著差异，这就表明存在样本选择问题。

4 工具变量 *IV*

4.1 *Endogeneity*

IV 的出现是为了解决内生性问题，一般意义上，内生性来自于三种方式：

1. 遗漏变量偏误（*OVB*）：偏误公式 $\hat{\beta} = \beta + \frac{cov(x,z)}{var(x)}$，其中 z 为遗漏变量；如果遗漏变量和内生变量正相关表明估计量存在高估；
2. 联立方程问题 (*SEM*): 联立方程导致因素相互决定的均衡系统（供求系统）；
3. 测量误差问题：给定测量误差描述框架

$$\begin{aligned} y_i - u_{yi} &= (x_i - u_{xi})\beta + \varepsilon_i \\ y_i &= x_i\beta + (\varepsilon_i + u_{yi} - u_{xi}\beta) \end{aligned}$$

考虑到 y_i 的测量误差 u_{yi} 与 x_i 无关，因此被解释变量的测量误差不会导致估计偏误；x_i 中的测量误差 u_{xi} 会导致估计偏误，即趋零偏误问题 $\hat{\beta} < \beta$；

4.2 *SEM* 中 *IV* 的识别问题

讨论 IV 问题需要首先明确 IV 的识别条件是什么，即满足什么条件下 IV 能够识别结构参数。一般形式上给出 IV 估计下的结构模型：

$$\begin{aligned} \underbrace{y_j}_{n\times 1} &= \underbrace{Y_j}_{n\times m_j}\underbrace{\gamma_j}_{m\times 1} + \underbrace{x_j}_{n\times G_j}\underbrace{\beta_j}_{G_j\times 1} + \underbrace{\varepsilon_j}_{n\times 1} \\ &= z_j\delta_j + \varepsilon_j \end{aligned}$$

其中 y_j 表示被解释变量，Y_j 表示内生变量，x_j 表示外生变量（IV），定义 $z_j = (Y_j, X_j)$。其中 m_j 表示内生变量的个数，G_j 表示外生变量 IV 的个数。现在需要明确是满足什么情况时，IV 能够识别模型参数。

为了更清楚的定义模型参数，定义 *structural* 模型：

$$y_t' \underbrace{\Gamma}_{m\times m} + x_t' \underbrace{B}_{G\times m} = \varepsilon_t'$$

$$\begin{bmatrix} y_{1t} \\ \cdots \\ y_{mt} \end{bmatrix}' \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \cdots & \cdots & \cdots \\ \gamma_{m1} & \cdots & \gamma_{mm} \end{bmatrix} + \begin{bmatrix} x_{1t} \\ \cdots \\ x_{Gt} \end{bmatrix}' \begin{bmatrix} \beta_{11} & \cdots & \beta_{1m} \\ \cdots & \cdots & \cdots \\ \beta_{G1} & \cdots & \beta_{Gm} \end{bmatrix} = \begin{bmatrix} \varepsilon_{1t} \\ \cdots \\ \varepsilon_{mt} \end{bmatrix}'$$

给定如上 SEM 结构模型，的任务是利用数据识别出参数 Γ 和 B，如果能够利用 reduce form 识别出上述参数，则说该模型是可识别的（*identification*）。

定义如下 *reduce form* 模型：结构模型左右侧同时乘以 Γ^{-1} 得到

$$y_t' = x_t\Pi + v_t'$$

其中 $\Pi = -B\Gamma^{-1}$，约简式估计中唯一能够无偏识别的是参数 Π。现在的问题转化为，约简估计 Π 能否得到唯一的 Γ 和 B? 注意到，令 $\bar{\Gamma} = \Gamma F, \bar{B} = BF, \bar{\varepsilon} = \varepsilon F$，相应的 $\Pi = -\bar{B}\bar{\Gamma}^{-1} = -B\Gamma^{-1}$，这表明不能直接通过 *reduce form* 估计得到唯一的 Γ 和 B 估计值，因此在该形式的 *reduce form* 估计中，参数 Γ 和 B 是不可识别的。

4.2.1 第 I 种识别策略

对于第 j 个方程，从 (4.6) 中可以得到：

$$(y_{1j}\gamma_{1j} + \cdots + y_{mj}\gamma_{mj}) + (x_{1j}\beta_{1j} + \cdots + x_{Gj}\beta_{Gj}) = \varepsilon_{jt}$$

转化为一般形式的方程得到

$$y_j = Y_j'\gamma_j + (Y_j^*)'\gamma_j^* + X_j'\beta_j + (X_j^*)'\beta_j^* + \varepsilon_j$$

其中 y_j 表示被解释变量，Y_j, Y_j^* 分别表示在第 j 个方程中出现的内生变量和未出现的内生变量，对应的个数分别为 m_j, m_j^*；X_j, X_j^* 表示在第 j 个方程中出现的外生变量和未出现的外生变量，对应的个数分别为 G_j, G_j^*。根据上述方程，可以得到必然有 $\gamma_j^* = \beta_j^* = 0$。那么现在的任务就是如何利用 *reduce form* 识别参数 γ_j 和 β_j。

等价问题表示为：给定结构模型

$$\begin{aligned} Y'\Gamma_j + X'B_j &= \varepsilon_j \\ (y_j, Y_j', (Y_j^*)')(1, -\gamma_j', 0')' + (X_j', (X_j^*)')(-\beta_j', 0') &= \varepsilon_j \end{aligned}$$

将该结构模型转化为 *reduce form* 形式的问题，得到：

$$\begin{aligned} (y_j, Y_j', (Y_j^*)') &= (X_j', (X_j^*)')\Pi + (v_j, V_j', (V_j^*)') \\ \Pi &= \begin{bmatrix} \pi_j & \Pi_j & \bar{\Pi}_j \\ \pi_j^* & \Pi_j^* & \bar{\Pi}_j^* \end{bmatrix} \end{aligned}$$

考虑到 $\Pi = -B\Gamma^{-1}$，等价转化为 $\Pi\Gamma_j + B_j = 0$，等价于：

$$\begin{bmatrix} \pi_j & \Pi_j & \bar{\Pi}_j \\ \pi_j^* & \Pi_j^* & \bar{\Pi}_j^* \end{bmatrix} \begin{bmatrix} 1 \\ -\gamma_j \\ 0 \end{bmatrix} = \begin{bmatrix} \beta_j \\ 0 \end{bmatrix}$$

求解该等式可以得到

$$\underbrace{\Pi_j^*}_{G_j^* \times m_j} \gamma_j = \underbrace{\pi_j^*}_{G_j^* \times 1}$$

对于该线性方程组，需要保证存在唯一解 γ_j，从而保证了 IV 可以识别出结构模型中的参数。相应的，转化为如下判别条件：

1. 矩条件（必要条件）：$G_j^* > m_j$，含义是作为 IV 的外生变量的个数需要大于内生变量个数；在 *SEM* 中的含义表示为在该系统中未出现但是在其他系统中出现的外生变量个数需要大于该系统中内生变量的个数，例如供给方程需要需求侧外生变量作为识别来源；[6]

2. 秩条件（充分条件）：$rank(\Pi_j^*) = m_j$，仅仅满足矩条件并不能保证能够完全识别参数，根据线性代数有解的条件必然是系数矩阵是满秩矩阵。

[6]对于一般的一个回归方程只需要按照矩条件判定，即外生变量个数和内生变量个数相等为恰好识别，外生变量更多表示过度识别。

4.2.2 第 II 种识别策略

定义矩阵

$$\begin{bmatrix}\Gamma\\ B\end{bmatrix}=\begin{bmatrix}1 & A_1\\ -\gamma_j & A_2\\ 0 & A_3\\ -\beta_j & A_4\\ 0 & A_5\end{bmatrix}$$

其中第一列为第 j 个方程的参数，A_1-A_5 表示不在第 j 个方程中对应的参数量，可以从经济学理论中给出。该系统能够有助于更好的理解识别条件。如果想要唯一的识别参数，那么就不能存在 $(\Gamma F, BF)$ 使得上式同样成立，定义 (f^0, f^1) 表示 F 第 j 列元素，可以得到：

$$\begin{bmatrix}\Gamma\\ B\end{bmatrix}\begin{bmatrix}f^0\\ f^1\end{bmatrix}=\begin{bmatrix}f^0+A_1f^1\\ -\gamma_j f^0+a_2f^1\\ A_3f^1\\ -\beta_j f^0+A_4f^1\\ A_5f^1\end{bmatrix}=\begin{bmatrix}1\\ \hat{-}\gamma_j\\ 0\\ \hat{-}\beta_j\\ 0\end{bmatrix}$$

这里的核心思想在于不允许存在其他的 $(\Gamma F, BF)$ 使得上式成立，否则就无法唯一的识别参数，因此为了保证上式不成立，等价转化为 $\begin{bmatrix}A_3\\ A_5\end{bmatrix}$ 是满秩矩阵，这样 $\begin{bmatrix}A_3\\ A_5\end{bmatrix}f^0=0$ 不存在解，上式也就无法成立。

进一步的，$\begin{bmatrix}A_3\\ A_5\end{bmatrix}$ 是 $(m_j^*+G_j^*)\times(m-1)$ 的矩阵，所以可以相应的得到判别条件：

1. 矩条件（必要条件）：$m_j^*+G_j^*\geq 1+m_j+m_j^*-1\rightarrow G_j^*\geq m_j$，矩条件和第一种识别策略一致；
2. 秩条件（充分条件）：$rank(\begin{bmatrix}A_3\\ A_5\end{bmatrix})=m-1$，即矩阵式列满秩的；

上述两种识别策略关于矩条件是一致的，即要求外生变量个数应该大于内生变量个数才能识别参数。关于秩条件，需要从模型判定，如果能够确定 Π_j^* 的形式，利用第 I 种识别策略判别充分条件更有效；如果可以根据经济学理论得到 $\begin{bmatrix}A_3\\ A_5\end{bmatrix}$ 的形式并判定秩，那么第 II 种识别策略更有效。总而言之，在一般的单一 IV 回归中判定个数条件即可，但是在 SEM 中 IV 的识别条件需要进行严格的界定，否则会出现识别不足（*incomplete*）的问题。

4.3 *IV* 与 *2SLS* 估计量

Notes：IV 和 2SLS 的核心差异在于当 IV 的数量大于内生变量时，采用 2SLS 方法更加便捷。

4.3.1 *IV* 估计量

给定回归模型[7]

$$y_j=Y_j\gamma_j+X_j\beta_j+\varepsilon_j=z_j\delta_j+\varepsilon_i$$

[7]这里全部使用了 j 个方程的形式，是为了从 SEM 的角度建立 IV 模型；在第 j 个方程中每个变量依旧是矩阵；这样做的好处是可以在 3SLS 以及 SEM 中更好的进行衔接。

其中 Y_j 表示内生变量，x_j 表示外生变量，$z_j = (Y_j, X_j)$ 表示回归方程的自变量，其中包含内生变量和非内生变量；给定上述情形，OLS 估计量是有偏的，这是因为

$$\hat{\delta}_j = (Z'Z)^{-1}(Z'y) = \delta_j + (Z'Z)^{-1}(Z'\varepsilon)$$

由于 $cov(Y_j, \varepsilon_j) \neq 0$，因此 OLS 估计量是不一致的。相应的，使用 IV 方法进行估计。给定 W_j 作为 Z_j 的 IV，作如下假定：

1. $\frac{W_j'Z_j}{n}$ 是非奇异矩阵（满秩矩阵），表明 IV 和内生变量存在相关性；
2. $\frac{W_j'\varepsilon_j}{n} \to 0$，表明 IV 和随机性无关，满足外生性约束；
3. $\frac{W_j'W_j}{n}$ 是正定矩阵，表明 IV 必须是不同的，排除存在完全共线性的可能；

根据矩条件可以得到 IV 估计量：Iv 识别依赖于核心的矩条件 $W_j'\varepsilon_j = 0$

$$W_j'\varepsilon_j = W_j'(y_j - Z_j'\delta_j) = 0 \to \hat{\delta}_{IV} = (W_j'Z_j)^{-1}(W_j'y_j)$$

进一步的考察 IV 估计量的基本性质：首先考察一致性

$$\begin{aligned}\hat{\delta}_{IV} &= (W_j'Z_j)^{-1}(W_j'y_j) \\ &= \delta_j + (W_j'Z_j)^{-1}(W_j'\varepsilon_j) \\ &= \delta_j + (W_j'Z_j/n)^{-1}(W_j'\varepsilon_j/n) \\ &\to_p \delta\end{aligned}$$

其中 $W_j'\varepsilon_j/n \to_p 0$，因此 IV 估计量是一致的。进一步考察 IV 估计量的渐进正态分布性质：

$$\begin{aligned}\hat{\delta}_{IV} &= \delta_j + (W_j'Z_j)^{-1}(W_j'\varepsilon_j) \\ \sqrt{n}(\hat{\delta}_{IV} - \delta_j) &\to_d \mathcal{N}(0, \sigma^2 \lim_{n\to\infty} \frac{1}{n}(W_j'Z_j)^{-1}W_j'W_j(Z_j'W_j)^{-1})\end{aligned}$$

Proof:

$$\begin{aligned}\sqrt{n}(\hat{\delta}_{IV} - \delta_j) &= (W_j'Z_j/n)^{-1}(W_j'\varepsilon_j/\sqrt{n}) \\ W_j'Z_j/n &\to_p W_j'Z_j \\ W_j'\varepsilon_j/\sqrt{n} &\to_d \mathcal{N}(0, \frac{\sigma^2}{n}W_j'W_j) \\ (W_j'Z_j/n)^{-1}(W_j'\varepsilon_j/\sqrt{n}) &\to_d (W_j'Z_j)^{-1}\mathcal{N}(0, \frac{\sigma^2}{n}W_j'W_j) \\ &\to_d \mathcal{N}(0, \sigma^2 \lim_{n\to\infty} \frac{1}{n}(W_j'Z_j)^{-1}W_j'W_j(Z_j'W_j)^{-1})\end{aligned}$$

4.3.2 *2SLS* 估计量

2SLS(两阶段最小二乘法) 是一种特殊的 IV，适用于当 IV 个数大于内生变量个数的情况，首先构造一个估计量来汇总多个 IV 中的信息，以此为基础进行估计，两阶段步骤如下：给定工具变量 W_j

1. $Z_j = \alpha W_j + v_j$，内生变量对所有的 IV 进行回归，得到估计量 $\hat{Z}_j = P_{W_j}Z_j = W_j(W_j'W_j)^{-1}W_j'Z_j$；
2. $y_j = \delta_j\hat{Z}_j + \varepsilon_j$，被解释变量对一阶段估计量回归得到估计系数 $\hat{\delta}_{2SLS} = (\hat{Z}_j'\hat{Z}_j)^{-1}(\hat{Z}_j'y_j)$；

两阶段估计量表示为

$$\hat{\delta}_{2SLS} = (\hat{Z}_j'\hat{Z}_j)^{-1}(\hat{Z}_j'y_j) = (Z_j'P_{W_j}Z_j)^{-1}(Z_j'P_{W_j}y_j)$$

进一步考察 2SLS 估计量的基本性质：首先考察一致性

$$\begin{aligned}\hat{\delta}_{2SLS} &= (Z_j' P_{W_j} Z_j)^{-1}(Z_j' P_{W_j} y_j) \\ &= \delta + (Z_j' P_{W_j} Z_j)^{-1}(Z_j' P_{W_j} \varepsilon_j) \\ &= \delta + (Z_j' P_{W_j} Z_j/n)^{-1}(Z_j' P_{W_j} \varepsilon_j/n) \\ &\to_p \delta\end{aligned}$$

其中 $W_j'\varepsilon_j/n \to_p 0; (Z_j' P_{W_j}\varepsilon_j/n) = Z_j'(W_j(W_j'W_j)^{-1}W_j')\varepsilon_j/n \to_p 0$。进一步考察 2SLS 的渐进正态分布性质：

$$\sqrt{n}(\hat{\delta}_{2SLS} - \delta) \to_d \mathcal{N}(0, \sigma^2 \lim_{n\to\infty} \frac{1}{n}(Z_j' P_{W_j} Z_j)^{-1})$$

Proof:

$$\begin{aligned}\sqrt{n}(\hat{\delta}_{2SLS} - \delta_j) &= (Z_j' P_{W_j} Z_j/n)^{-1}(Z_j' P_{W_j}\varepsilon_j/\sqrt{n}) \\ Z_j' P_{W_j} Z_j/n &\to_p Z_j' P_{W_j} Z_j \\ Z_j' P_{W_j}\varepsilon_j/\sqrt{n} &\to_d \mathcal{N}(0, \frac{\sigma^2}{n} Z_j' P_{W_j} Z_j) \\ \sqrt{n}(\hat{\delta}_{2SLS} - \delta_j) &\to_d \mathcal{N}(0, \frac{\sigma^2}{n}(Z_j' P_{W_j} Z_j)^{-1})\end{aligned}$$

4.3.3 IV 与 2SLS 估计量

问题 1：弱工具变量问题。如果工具变量和内生变量之间相关性很弱（一阶段 F 检验小于 10 或者不显著），说明是弱 IV，这种情况下使用 IV 进行统计推断存在明显的偏误；这是因为给定 IV 统计量 $(W_j'Z_j)^{-1}(W_j'y_j)$，如果 IV 是弱工具变量，则 $(W_j'Z_j)^{-1}$ 非常大从而使得估计量存在偏误，弱 IV 时的偏误甚至会高于 OLS 估计的偏误，因此存在弱 IV 必须要严格对待，例如使用弱工具变量稳健估计（*Isaiah Andrews*，*ARE*）。对于弱工具变量的判断标准一般以一阶段联合 F 检验为主，需要一阶段 F 值大于 10（经验法则），一般而言至少需要的大于 15；*Lee(2022 AER)* 给出了一种基于 F-t 的检验表，给出了每种 F 检验对应的 t 统计量临界值。

问题 2：IV 和 2SLS 估计量的关系：如果 IV 数量和内生变量个数一致，IV 估计量和 2SLS 估计量哪种更有效率？两者是等价的，证明如下：如果 IV 个数和内生变量个数一致，则表明 Z, W 同阶，从而可以求解逆矩阵

$$\begin{aligned}\hat{\delta}_{2SLS} &= (Z'W(W'W)^{-1}W'Z)^{-1}(Z'W(W'W)^{-1}W'Y) \\ &= (W'Z)^{-1}W'W(Z'W)^{-1}Z'W(W'W)^{-1}W'Y \\ &= (W'Z)^{-1}W'Y = \hat{\delta}_{IV}\end{aligned}$$

因而恰好识别的情况下没有必要使用 2SLS 方法。2SLS 有用的情形是 IV 个数多于内生变量的个数从而需要进行有效的汇总信息。

问题 3：对于任意 IV 个数多于内生变量个数的情形，可以证明 2SLS 估计量是最有效的，该命题表示为数学形式：对于工具变量 W, 现在给定矩阵 A, 使得 WA 表示 IV 的线性组合，则 $W(W'W)^{-1}W'Z$ 是最优估计量。

Proof: 给定 IV 的渐进方差表示为 $(A'WZ)^{-1}A'H'HA(Z'HA)^{-1}$（将 WA 作为 W_j 代入得到）; *2SLS* 的渐进方差表示为 $(Z'W(W'W)^{-1}W'Z)^{-1}$, 需证 $(Z'W(W'W)^{-1}W'Z)^{-1} \le (A'WZ)^{-1}A'H'HA(Z'HA)^{-1}$; 仅需证明 $Z'W(W'W)^{-1}W'Z \ge (Z'HA)(A'H'HA)^{-1}(A'WZ)$, 后者左右侧同乘以 P_W 得到 $Z'P_W P_{WA} P_W Z$, 仅需比较 $Z'P_W P_{WA} P_W Z$ 和 $Z'P_W Z$ 的大小关系。

问题 4：*Many IV Issue(Becker, 1994)*。问题表述为如果存在数量众多的 IV（例如动态面板中滞后项都可以作为 IV），那么 IV 估计量是不是一致的？给定特殊情形，如果 IV 数量大于样本数量 n，则可

以证明 $W(W'W)^{-1}W'$ 收敛到 I，此时 IV 估计量退化为 OLS 估计量，因此估计是存在偏误的。一般的，估计偏误程度表示为 $\frac{k}{n}$，其中 k 表示 IV 个数，n 表示样本量，IV 数量越大会导致估计的偏误越严重。

4.4 *3SLS*

对于 SEM 系统，如何有效的同时估计每一个回归方程? 一个自然段的想法是，如果每个方程间不存在相关性（随机项不相关），则可以使用 2SLS 方法逐个估计方程。但是如果 SEM 的方程间存在相关性（随机项相关），逐个估计便存在偏误（容易证明使用其他方程中的变量作为 IV 同样存在内生性问题），因此转而使用 3SLS 方法同时估计 SEM。

三阶段最小二乘法是联立方程模型的一种完全信息估计方法，所谓“完全信息估计法”指利用所有可用的信息，同时估计模型中的所有方程。本法基本思想是：应用两阶段最小二乘法的估计误差构造模型随机扰动项协方差矩阵的统计量，从而对整个模型进行广义最小二乘估计。主要步骤为：(1) 模型系统要求可识别，抽去所有的定义方程式 (即恒等式)；(2) 对模型简化式作最小二乘估计；(3) 以上述估计量为工具变量对模型结构式进行最小二乘估计 (即两阶段最小二乘估计)，并计算估计误差；(4) 以两阶段估计误差构造扰动项方差的统计量，进行广义最小二乘估计。3SLS 估计结果在一定条件下比 2SLS 具有更好的渐近有效性。

对于第 j 个方程有

$$y_j = z_j\delta_j + \varepsilon_i \; E(\varepsilon_i\varepsilon_j') = \sigma_{ij}I_n$$

该方程表明 SEM 不同方程存在相关性（随机项相关）；SEM 表示为

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix} = \begin{bmatrix} Z_1 & & & \\ & Z_2 & & \\ \dots & \dots & \dots & \\ & & & Z_m \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \dots \\ \delta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_i \\ \varepsilon_2 \\ \dots \\ \varepsilon_m \end{bmatrix}$$

矩阵形式表示为

$$Y = Z\delta + \varepsilon$$

其中方差协方差矩阵表示为

$$E(\varepsilon\varepsilon') = \begin{bmatrix} \sigma_{11}I_n & \dots & \sigma_{1m}I_n \\ \dots & \dots & \dots \\ \sigma_{m1}I_n & \dots & \sigma_{mm}I_n \end{bmatrix} = \Sigma_m \otimes I_n$$

其中 $\otimes$ 表示矩阵的*Kronecker* 积。从协方差矩阵中可以清晰的看出 SEM 内部方程间存在相关性，单独估计某个方程存在估计偏误。

给定 3SLS 的估计步骤：

- 第 j 方程的内生变量 z_j 回归到 X 上面，其中 X 表示 SEM 中所有的 x 变量集合，包括第 j 个方程中的变量和非第 j 个方程中的变量，得到估计值 $\hat{z}_j$；
- 得到两阶段估计量 $\hat{\delta} = (\hat{z}_j'\hat{z}_j)^{-1}(\hat{z}_j'y_j)$，同时可以得到不同方程间的协方差估计 $\hat{\sigma}_{ij} = \frac{e_i'e_j}{n}$，从而可以得到协方差矩阵的估计 $\hat{\Sigma}_m$；
- 3SLS 估计量表示为 $\hat{\delta}_{3SLS} = [\hat{Z}_j'(\hat{\Sigma}_m^{-1} \otimes I_n)\hat{Z}_j]^{-1}[\hat{Z}_j'(\hat{\Sigma}_m^{-1} \otimes I_n)Y]$；

4.5 *Hausman Test*

Hausman Test 针对的基本问题是：存在估计量 $\hat{\theta}$ 在 H_0 成立时是一致且有效的估计量，但是在 H_0 不成立时存在偏误；存在估计量 $\bar{\theta}$ 不论 H_0 成立与否都是一致的，但是在 H_0 成立时不如 $\hat{\theta}$ 有效；简言之，存在两类估计量，有效估计量但不稳健、稳健估计量但不有效，如何在两类估计两种进行选择?

以 IV 的内生性检验为例，H_0 表示为 $cov(z,\varepsilon)=0$，即不存在内生性问题；OLS 估计量在 H_0 成立时是有效且一致的，但是在 H_0 失效时存在偏误；IV 估计量不论 H_0 是否成立都有效，但是 H0 时不如 OLS 有效。如果 *Hausman Test* 检验选择了 IV 估计量则表明存在内生性问题，如果选择 OLS 估计量则表明内生性问题并不严重，从而通过对估计量的选择实现内生性的检验。但是注意，*Hausman Test* 中的内生性检验是一种统计性检验，不能提供给关于因果的内生性信息，因此因果的内生性绝对不能依赖于 *Hausman Test*（江艇，2023）。

给定两个估计量 θ_1,θ_2，其中估计量 θ_1 在 H_0[8]成立的情况下更有效但是在 H_0 不成立时不一致；估计量 θ_2 在 H_0 和 H_1 成立的情况下都是一致估计，但是在 H_0 成立的情况下并非有效。如何构造检验判断选择哪一个估计量?

Hausman 检验构造为:

$$(\theta_1-\theta_2)'V^{-1}(\theta_1-\theta_2)\sim\chi^2(k)$$

其中 $V=var(\theta_1-\theta_2)$,V^{-1} 可以表示为广义逆矩阵（允许其中部分元素为 0），自由度 k 表示内生变量个数。如果 H_0 成立，则有 $V=var(\theta_1-\theta_2)=V(\theta_2)-V(\theta_1)$[9]

$$(\theta_1-\theta_2)'[V(\theta_1)-V(\theta_2)]^{-1}(\theta_2-\theta_1)\sim\chi^2(k)$$

对于线性模型，构造 *Hausman* 检验，其中 OLS 估计量表示为 $\beta_{OLS}=(X'X)^{-1}(X'Y)$，IV 估计量表示为 $\beta_{2SLS}=(X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'Y)$，给定 H_0：$cov(x,\varepsilon)=0$（不存在内生性因而 OLS 是有效且一致的）成立，构造检验

$$\mathcal{HS}=\frac{1}{\sigma^2}(\beta_{2SLS}-\beta_{OLS})'[(X'Z(Z'Z)^{-1}Z'X)^{-1}-(X'X)^{-1}]^{-1}(\beta_{2SLS}-\beta_{OLS})\sim\chi^2(k)$$

其中 k 表示 X 中内生变量个数。如果估计量 $\mathcal{HS}$ 大于临界值，则倾向于拒绝原假设，从而存在内生性问题；反之则不存在内生性问题。

4.6 *LATE*

本部分是在 *Angrist and Imbens(1994ECMA)*[10]基础上阐释 IV 估计量的因果含义，建立局部平均处理效应（LATE）的框架。

如果工具变量 Z 为虚拟变量，估计量表示为

$$\beta=\frac{E(y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)}$$

其中 Z 表示工具变量，D 表示内生变量，工具变量满足 $E(\varepsilon|Z)=0$。上式可以进一步理解为组间均值差异，其中识别变动性的来源是工具变量 Z 的变动性。相应的可以利用样本数据进行估计得到 *Wald* 估计量

$$\hat{\beta}_{wald}=\frac{E_N(y_i|Z_i=1)-E_N(Y_i|Z_i=0)}{E_N(D_i|Z_I=1)-E_N(D_i|Z_i=0)}$$

[8]原假设与备择假设的选择是灵活的，取决于具体的研究情景，其中原假设保证估计量 1 是有效的一致估计量，例如在 IV 估计中原假设表示为 $cov(x,\varepsilon)=0$，即不存在内生性问题，此时 *Hausman* 检验可以视为内生性检验，但只是一种估计量本身性质的统计检验，不能代替识别。

[9]其中 θ_1 在原假设成立下更有效，因而 $var(\theta_1)<var(\theta_2)$.

[10]*Identification and Estimation of Local Average Treatment Effects, Econometrica, March 1994.*

Angrist and Imbens(1994ECMA) 将全部人群划分为四类：*never taker* $(D_i^1 = 1 = D_i^0 = 0)$ 、*defiers*$(D_i^1 = 0, D_i^0 = 1)$、*always taker* $(D_i^1 = D_i^0 = 1)$ 、*compliedr* $(D_i^1 = 1, D_i^0 = 0)$ 。简言之，如果处理发生，那么 *compliedr* 会进行选择 A，如果处理不发生，*compliedr* 会选择 B，因此 *compliedr* 的选择是与处理本身高度相关的，当然 *defiers* 也与政策高度相关，因此需要依赖假定进行排除。如果个体选择是内生的，但是处理是外生的（例如抽签随机），那么可以利用 IV 来识别处理的效应，或者 IV 的 LATE。

下面给出完整的LATE 定理：

$$y_i = \alpha_0 + D_i\beta_i + \varepsilon_i$$
$$D_i = \pi_0 + \pi_{1i}Z_i + \eta_i$$

其中 D_i 表示是否接受处理，例如是否参军服兵役等，参数 π_{1i} 表征了 $z_i = 1$ 对于参军服兵役的影响。**给定如下识别假设**[11]

1. 独立性：接受处理的潜在结果与 Z 无关，$E(Y_i(D_{1i}, 1), Y_i(D_{0i}, 0), D_{1i}, D_{0i}|Z) = 0$；或者说建立在潜在因果的框架下，实际的随机抽签不影响接受处理的潜在结果；
2. 排他性：$E(y_i.z_i|D_i) = 0$，z_i 只通过影响 D 影响 y;
3. 一阶段条件：$E(D_{1i} - D_{0i}) \neq 0$，即 $\pi_{1i} \neq 0$，不存在弱 IV 问题；
4. 单调性：$D_{1i} - D_{0i} \geq 0$ 或者 $D_{1i} - D_{0i} \leq 0$，这排除了 *defiers* 的存在，因此可以识别 *complier* 的平均处理效应；单调性的成立保证了因果效应是建立在对于异质性个体 *complier* 的基础上的。

IV 估计量可以表示为 *complier*$(D_{1i} - D_{0i} > 0)$ 的平均处理效应：

$$\begin{aligned}\hat{\beta}_{wald} &= \frac{E_N(y_i|Z_i = 1) - E_N(Y_i|Z_i = 0)}{E_N(D_i|Z_I = 1) - E_N(D_i|Z_i = 0)} \\ &= E(Y_{1i} - Y_{0i}|D_{1i} > D_{0i}) \\ &= E(\beta_i|\pi_{1i} > 0)\end{aligned}$$

Angrist (1990)[12]使用越战时期的抽签号码作为服兵役 (*dummy*) 的工具变量，用以估算服兵役对晚年收入的影响。一方面，抽签号码是随机的满足外生性，同时抽签中签与参军具有高度的正相关性（不完全一致，部分中签的会因为身体或者教育原因等无法参军）；另一方面，抽签是为了征兵设计的，因此只通过影响征兵服兵役进而影响个体收入，具有排他性。给定 IV 识别假设，利用抽签号码作为 IV 识别的因果效应实际上是 *complier* 的平均处理效应[13]。

[11]因果推断的关键不在于统计方法而在于识别假设是否得到满足，只有在满足识别假设的情况下传统的统计模型才具有更强大的因果解释力，因此识别假设是因果模型的关键，理解因果性首先需要理解识别假设，寻找满足识别假设的环境进行研究设计。

[12]Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*. June, 80:3, pp. 313–36.

[13]事实上，精确的异质性 IV 分析框架需要明确的界定四类个体并在此基础上进行识别，但是目前 IV 应用研究大部分对于该框架缺乏有效的应用，因而解释力度欠缺；中村惠美（2021RES）利用火山爆发作为迁移决策的工具变量，很好的界定了四类个体并讨论了平均处理效应。

5 *GMM*

5.1 *GMM* 估计量

给定矩条件即期望条件:

$$Ef(x,\beta_0)=0$$

其中 f 是 $r\times 1$ 矩阵（方程个数或者工具变量个数），β_0 是 $q\times 1$ 矩阵（未知参数个数或内生变量个数）。如果 $r<q$ 表示识别不足（*under identified*）；$r>q$ 表示过度识别（*over identified*）；$r=q$ 表示恰好识别（*just identified*）。例如 $r=1<q=3$ 意味着 $y=\alpha_0+\alpha_1x_1+\alpha_2x_2+\varepsilon$，一个方程识别三个参数意味着识别不足。如果说恰好识别，可以直接利用矩条件识别，方程个数等于未知量个数从而可以确定估计量；但是如果说方程个数多于未知数个数，线性方程组就会出现无解的情况，需要利用广义矩方法进行估计，其基本思路是尽可能降低样本矩以满足矩条件。下面给出 *GMM* 的完备形式。

矩条件给定为

$$g_n(x,\beta)=g_n(\beta)=\frac{1}{n}\sum f(x,\beta)$$

例如在 OLS 估计中矩条件可以表示为

$$E(x\varepsilon)=0\to\frac{1}{n}\sum x_i'\varepsilon_i=\frac{1}{n}\sum x_i'(y_i-x_i\beta)=0$$

上述矩条件可以确定参数 β。IV 的矩条件可以进一步表示为

$$E(z\varepsilon)=0\to\frac{1}{n}\sum z_i'\varepsilon_i=\frac{1}{n}\sum z_i'(y_i-x_i\beta)=0$$

如果 IV 个数大于内生变量个数，则表明方程个数多于内生变量个数，出现过度识别问题，使用线性方程组的方法无法识别模型参数。恰好识别情形下可以使用矩方法 $g_n=0$（线性方程组有解）进行估计，过度识别则需要利用 *GMM* 进行估计。这是因为对于矩条件 $g_n(\beta)=0$，其中 g_n 是 $r\times q$ 矩阵，β 是 $q\times 1$ 矩阵，如果 $r>q$ 则出现过度识别，无法唯一确定参数估计量。

过度识别情况下矩条件 $g_n=0$ 难以成立，*GMM* 的基本思路就是尽可能将矩条件估计值尽量小

$$min_\beta\ g_n'(\beta)A_ng_n(\beta)$$

其中 A 是 $r\times r$ 矩阵，表示正定的权重矩阵（*weighting matrix*），正定要求每个估计方程都是有用的（*information*）。

5.2 *GMM* 估计量性质

给定一阶条件得到

$$\frac{\partial g(\hat\beta)}{\partial\beta}A_ng(\hat\beta)=0$$

其中根据 *Taylor* 展开得到 $g(\hat\beta)=g(\beta_0)+\frac{\partial g(\bar\beta)}{\partial\beta}(\hat\beta-\beta_0)$，其中 β_0 表示参数真值，$\bar\beta\in(\hat\beta,\beta_0)$，定义 $G(\beta)=\frac{\partial g(\beta)}{\partial\beta}$，简化得到

$$\hat\beta-\beta_0=-(\hat G'A_n\bar G)^{-1}(\hat G'A_ng_n(\beta))$$

其中 $\hat G=\frac{\partial g(\beta)}{\partial\beta}|_{\beta=\hat\beta}$。

给定 $g_n(\beta)$ 的渐进正态性质

$$g_n=\frac{1}{n}\sum g(x_i,\beta_0)\to_p Ef$$

$$\sqrt{n}g_n(\beta_0)\to_d N(0,\lim S_n)$$

其中 $g_n = \frac{1}{n}\sum g(x_i,\beta_0), \lim S_n = var(\frac{1}{\sqrt{n}}\sum g(x_i,\beta_0))$，$S_n$ 表示矩条件对应的方差矩阵。

$$\sqrt{n}(\hat{\beta}-\beta_0) \to_d N(0, \lim(G_n' A_n G_n)^{-1} G_n' A_n S_n A_n G_n (G_n' A_n G_n)^{-1})$$

其中 A, G 为对称矩阵。

有效 *GMM* 的权重给定为 $A_n = S_n^{-1}$，渐进方差最小，有效 *GMM* 估计量渐进分布简化为

$$\sqrt{n}(\hat{\beta}_{GMM}-\beta_0) \to_d N(0, \lim(G_n' S_n^{-1} G_n)^{-1})$$

进一步需要证明有效 *GMM* 的渐进方差最小，仅需证明

$$\begin{aligned} G_n' S_n^{-1} G_n &> [(G_n' A_n G_n)^{-1} G_n' A_n S_n A_n G_n (G_n' A_n G_n)^{-1}]^{-1} \\ &= [(G_n' S^{-1/2} S^{1/2} A_n G_n)^{-1} G_n' A_n S^{1/2} S^{1/2} A_n G_n (G_n' A_n S^{-1/2} S^{1/2} G_n)^{-1}]^{-1} \end{aligned}$$

5.2.1 两步 *GMM*

给定 *GMM* 的估计量形式和渐进正态分布性质，但是只有样本数据而没有真实值，借鉴 GLS 的基本思路：利用任意可行 *GMM* 估计方差矩阵，将样本估计所得的方差矩阵代入 *GMM* 估计量，因此可以进行 *two step* 得到 *GMM* 估计量。

Step 1：根据样本数据得到方差结构：

$$min_\beta \ g_n'(\beta) g_n(\beta), \ A_n = I \to \hat{S}_n$$

得到样本估计的方差矩阵 $\hat{S}_n$.

Step 2：将估计的方差矩阵代入最优化问题

$$min_\beta \ g_n'(\beta) \hat{S}_n g_n(\beta), \ A_n = \hat{S}_n^{-1}$$

从而最终得到 *GMM* 估计量 $\hat{\beta}_{GMM}$。

5.3 过度识别检验

过度识别约束检验是对工具变量的外生性进行检验。在恰足识别情况下，无法对工具变量的外生性进行直接检验，但是在过度识别情况下，就可以检验多余的工具变量是否与干扰项 ε 不相关。过度识别检验（连享会）

5.3.1 *Sargent J* 检验

Sargent J 检验：原假设为 H_0：$g_n(\beta) = 0$，例如在 IV 中表示为所有的工具变量都是外生的，构造 *Sargent J* 估计量

$$\begin{aligned} J_n &= n g_n'(\hat{\beta}) S_n^{-1} g_n(\hat{\beta}) \to_d \chi^2_{r-q} \\ &= \sqrt{n} g_n'(\hat{\beta}) S_n^{-1} \underbrace{\sqrt{n} g_n(\hat{\beta})}_{N(0,S)} \end{aligned}$$

其中 $(r-q)$ 表示工具变量和内生变量个数的差值；如果 J_n 足够小，则不能拒绝 IV 均为外生的原假设。

以 IV 检验为例给出 *Sargent J* 检验。给定线性模型：

$$y = \beta_1 x_1 + \cdots + \beta_{k-r} x_{K-r} + \beta_{K-r+1} x_{K-r+1} + \cdots + \beta_K x_K + \varepsilon$$

其中 $K-r$ 个外生变量和 r 个内生变量；假定存在 m 个工具变量 z，其中 $m>r$，过度识别原假设给定为：H_0:$cov(z_1,\varepsilon)=\cdots=cov(z_m,\varepsilon)$。考虑到扰动项无法观测，只能通过 2SLS 估计量的残差项考察工具变量与扰动项的相关性：

$$e_{2SLS}=\gamma_1 x_x+\cdots+\gamma_{K-r}x_{K-r}+\delta_1 z_1+\cdots+\delta_m z_m+v$$

原假设相应的可以转换为 H_0：$\delta_1=\cdots=\delta_m=0$，记辅助回归系数为 R^2, 构造 *Sargent J* 检验：

$$nR^2\rightarrow_d\chi^2_{m-r}$$

其中自由度表示为工具变量个数减去内生变量个数。

5.4 *OLS* 与 *2SLS* 下的 *GMM* 估计量

对于 OLS 估计，矩条件给定为

$$g_n=\frac{1}{n}X'\varepsilon, S=(\frac{\sigma^2}{n}X'X)$$

进一步的构造 *GMM* 估计量

$$min_\beta\ \varepsilon'X(X'X)^{-1}X'\varepsilon=(Y-X\beta)'X(X'X)^{-1}X'(Y-X\beta)$$

相应的根据一阶条件可以得到

$$\hat{\beta}_{GMM}=(X'X)^{-1}(X'Y)=\hat{\beta}_{OLS}$$

对于 IV 估计，矩条件给定为

$$g_n=\frac{1}{n}Z'\varepsilon, S=(\frac{\sigma^2}{n}Z'Z)$$

进一步的构造 *GMM* 估计量

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

相应的根据一阶条件可以得到

$$\hat{\beta}_{GMM}=[X'Z(Z'Z)^{-1}Z'X]^{-1}[X'Z(Z'Z)^{-1}Z'Y]=\hat{\beta}_{2SLS}$$

上述两种情形说明，OLS 和 2SLS 都是 *GMM* 的一种特殊形式。

6 *Panel Data*

给定面板模型（*TWFE* 双向固定效应模型）

$$y_{it} = x'_{it}\beta + \alpha_i + \lambda_t + \varepsilon_{it}$$

控制个体固定效应 α_i 和时间固定效应 λ_y，如果个体固定效应和时间固定效应与随机项无关（不存在固定效应导致的内生性问题），则可以直接使用随机效应模型。如果是短面板（样本 N 远远大于时间跨度 T），不控制时间固定效应是可以的，但是对于长面板（时间跨度 T 很长）或者 *VAR model*（长时间序列），不控制时间固定效应存在 *incidental parameters problem*。一般情况下面板数据需要控制双向固定效应，考虑短面板，模型可以处理为（简单起见，后文仅考虑控制个体固定效应的面板数据模型）

$$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}$$

随机效应模型的优势在于：（1）待估计参数给定；（2）随机效应模式在 *RE*（不存在固定效应内生性）情况下是有效率的；（3）可以控制不随时间变化或个体变化的因素，例如 $z'_i\gamma$。其缺点在于其假设条件是否得到满足？如果存在固定效应下的内生性，随机效应模型估计是不一致的。

固定效应模型的缺点在于：（1）待估计参数随着样本增加而增加（固定效应）；（2）在 RE 情况下不如随机效应估计量有效；（3）不可以控制不随时间或个体变化的因素；其优点在于 RE 假定不成立的情况下，FE 估计量是一致的。

***Hausman* 检验**：FE 估计量和 RE 估计量的权衡取舍。给定 H_0:$cov(\alpha_i, \varepsilon_{it}) = 0$（不存在内生性问题），构造 *Hausman* 检验。但是注意，不能将内生性的判定依赖于统计上的 *Hausman* 检验[14]。

6.1 *Fixed Effect Model*

给定固定效应模型：

$$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}$$

其识别假设是 $E(x_{it}, \varepsilon_{it}|\alpha_i) = 0, cov(\alpha_i, \varepsilon_{it}) \neq 0$，即存在个体层面不随时间变化因素导致的内生性问题，控制个体固定效应之后满足条件独立性，下面不同的视角来理解固定效应统计量究竟是什么含义。

LSDV* 估计 *(Least Square Dummy Variable)：将 OLS 估计应用于面板固定效应模型

$$min_{\alpha_0,\alpha_1,\beta} \sum_i \sum_t (y_{it} - x'_{it}\beta - \alpha_i)^2$$

根据一阶条件得到

$$\begin{aligned}
\hat{\beta}_{LSDV} &= [\sum_i \sum_t (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)']^{-1}[\sum_i \sum_t (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)] \\
&= [\sum_i \sum_t x_{it}(x_{it} - \bar{x}_i)']^{-1}[\sum_i \sum_t x_{it}(y_{it} - \bar{y}_i)] \\
&= [\sum_i \sum_t \hat{x}_{it}\hat{x}'_{it}]^{-1}[\sum_i \sum_t \hat{x}_{it}\hat{y}_{it}] \\
\hat{\alpha}_i^{LSDV} &= \bar{y}_i - \bar{x}'_i\hat{\beta} = \frac{1}{T}\sum_t y_{it} - \frac{1}{T}\sum_t x_{it}\hat{\beta}_{LSDV}
\end{aligned}$$

[14]对于因果推断中的固定效应和随机效应模型的理解：使用面板数据的原因就是为了控制个体层面的固定效应从而消除不可观测的个体因素的干扰，正是因为个体维度的因素造成的内生性所以需要将截面数据拓展为面板维度，但是随机效应模型的基本要求就是不存在个体固定效应造成的内生性问题，这与截面数据回归没有本质区别，因而从因果的角度看使用面板数据就等价于存在个体层面的内生性问题。尽管 *Hausman* 检验可以从统计的角度给出是否内生性的测度，但是这是一种完全统计的方法并且是基于统计量效率与一致性的 *tradeoff*，天然的不具备因果解释力，因此内生性解释绝对不依赖于 *Hausman* 检验。

其中 $\hat{x}_{it} = x_{it} - \bar{x}_i$ 表示组内去均值，其中 $\bar{x}_i = \frac{1}{T}\sum x_{it}$ 表示同组内整个时间维度上的均值，因此在这里的处理是组内去均值处理（*within equation*）。

可以证明：$\hat{\beta}_{LSDV}$ 是一致估计量，其含义在于利用组间去均值处理可以消除 α_i 从而利用 OLS 进行估计：

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

注意组内去均值的处理实际上在处理中少利用一期信息，这和一阶差分是一致的，因此实际上 *LSDV* 中利用的是 $T-1$ 期信息。一阶差分表示为

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta v_{it}$$

其中随机项 Δv_{it} 可能存在序列相关的问题，因而使用 *LSDV* 方法更为有效。

利用矩阵语言重新表述组内去均值处理：定义 $J_T = I_T - \frac{1}{T}l_T'l_T$，其中 $l_T = (1, 1, \ldots, 1)'$，J_T 表示针对同组内时间维度上的的组内去均值处理[15]，即 $J_T y_i = y_i - \frac{1}{T}\sum_t y_{it}$，具体的可以表示为

$$J_T y_i = \begin{bmatrix} y_{i1} - \frac{1}{T}\sum_t y_{it} \\ y_{i2} - \frac{1}{T}\sum_t y_{it} \\ \cdots \\ y_{iT} - \frac{1}{T}\sum_t y_{it} \end{bmatrix}$$

面板模型可以表示为

$$\begin{aligned} y_i &= X_i\beta + \alpha_i l_T + V_i \\ \rightarrow J_T y_i &= J_T X_i \beta + J_T V_i \\ \rightarrow \hat{\beta} &= (\sum_i x_i' J_T x_i)^{-1}(\sum_i x_i' J_T y_i) \end{aligned}$$

其中 y_i 是 $T \times 1$ 的列向量，X_i 表示 $T \times k$ 的矩阵，将其定义为 ***within equation***（组内去均值处理），其基本含义在于：个体固定效应可以通过一阶差分（*FD*）的形式消除，也可以通过组内去均值的方式消除，本质上是差分用于消除固定效应从而将回归转化为常见的 OLS 形式进行估计。可以证明，组内去均值的估计量 $\hat{\beta}_w = \hat{\beta}_{LDSV}$：

$$\begin{aligned} \hat{\beta} &= (\sum_i x_i' J_T x_i)^{-1}(\sum_i x_i' J_T y_i) \\ &= \beta + (\sum_i x_i' J_T x_i)^{-1}(\sum_i x_i' J_T \varepsilon_i) \\ E(\hat{\beta}) = \beta, var(\hat{\beta}) &= \sigma^2(\sum_i x_i' J_T x_i)^{-1} \end{aligned}$$

OLS* 估计**：除了一阶差分、组内去均值，还可以利用 ***between equation（组间去均值处理）方法消除固定效应，此时将面板数据转化为横截面数据进行使用

$$\bar{y}_i = \bar{x}_i'\beta + \alpha_i + \bar{v}_i = \bar{x}_i'\beta + \alpha + (\bar{\varepsilon}_i + \alpha_i - \alpha)$$

其中 $var(\bar{\varepsilon}_i + \alpha_i - \alpha) = \sigma_\alpha^2 + \frac{1}{T}\sigma_v^2$，OLS 估计量表示为[16]

$$\hat{\beta}_b = [\sum_i (\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})']^{-1}[\sum_i (\bar{x}_i - \bar{\bar{x}})(\bar{y}_i - \bar{\bar{y}})]$$

$$var(\hat{\beta}_b) = (\sigma_\alpha^2 + \frac{1}{T}\sigma_v^2)[\sum_i (\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})']^{-1}$$

其中 $\bar{x}_i = \frac{1}{T}\sum_t x_{it}, \bar{\bar{x}} = \frac{1}{N}\frac{1}{T}\sum_i\sum_t x_{it}$。

[15] J_T 以及 $\frac{1}{T}l_T l_T$ 分别表示常数项对应的残差矩阵与投影矩阵，两者互相垂直；上述形式是组内去均值的常规处理方法，其中 $\frac{1}{T}l_T l_T$ 用于表示同组内时间维度上的均值。

[16] $\hat{\beta}_w$ 表示组内去均值（*within*）处理的估计量；$\hat{\beta}_b$ 表示组间去均值（*between*）处理的估计量。

6.2 *Random Effect Model*

给定随机效应模型，假定 $cov(\alpha_i, x_i) = 0$（不存在内生性问题）：

$$y_i = X_i'\beta + \alpha_i l_T + V_i = X_i'\beta + \alpha l_T + \underbrace{(V_i + (\alpha_i - \alpha)l_T)}_{u_i}$$

其中 u_i 与 x_i 相互独立，但是对于相同的个体具有时间相关性 $cov(u_{is}, u_{it}) \neq 0$，存在异方差问题。构造 GLS 估计量

$$\hat{\beta} = (\sum_i x_i'\Omega x_i)^{-1}(\sum_i x_i'\Omega y_i)$$

进一步可以得到方差项：

$$var(u_i) = var(V_i + \alpha_i l_T) = \sigma_v^2 I_T + \sigma_\alpha^2 l_T l_T' = V$$

注意到，方差项 V 相对于 J_T 而言更加有效，这是随机效应模型比固定效应更有效的原因所在。相应的 (α 表示常数项对应系数或截距项，β 表示非常数项系数)

$$\begin{bmatrix}\alpha\\ \beta\end{bmatrix}_{GLS} = [\sum_i (l_T, x_i')' V^{-1}(l_T, x_i')]^{-1}[\sum_i (l_T, x_i')' V^{-1}(l_T, y_i')]$$

考察方差项 V^{-1} 得到

$$\begin{aligned} V^{-1} &= [\sigma_v^2 I_T + \sigma_\alpha^2 l_T l_T']^{-1} \\ &= [\sigma_v^2 J_T + (T\sigma_\alpha^2 + \sigma_v^2)\frac{1}{T} l_T l_T']^{-1} \\ &= \frac{1}{\sigma_v^2} J_T + \frac{1}{\sigma_1^2}\frac{1}{T} l_T l_T' \\ &= \frac{1}{\sigma_v^2}[J_T + \phi^2 \frac{1}{T} l_T l_T'] \end{aligned}$$

其中 $\sigma_1^2 = T\sigma_\alpha^2 + \sigma_v^2, \phi = \frac{\sigma_1}{\sigma_v}$，$J_t$ 与 $\frac{1}{T} l_T l_T'$ 相互垂直[17]。GLS 估计量转化为

$$\begin{aligned} \begin{bmatrix}\alpha\\ \beta\end{bmatrix}_{GLS} &= [\sum_i (l_T, x_i')'[J_T + \phi^2 \frac{1}{T} l_T l_T'](l_T, x_i')]^{-1}[\sum_i (l_T, x_i')'[J_T + \phi^2 \frac{1}{T} l_T l_T'] y_i] \\ &= \begin{bmatrix} nT\phi^2 & T\phi^2 \sum_i \bar{x}_i \\ * & \sum_i x_i J_T x_i' + T\phi^2 \sum_i \bar{x}_i \bar{x}_i' \end{bmatrix}^{-1} \begin{bmatrix} T\phi^2 \sum_i \bar{y}_i \\ \sum_i x_i J_T y_i + T\phi^2 \sum_i \bar{x}_i \bar{y}_i \end{bmatrix} \end{aligned}$$

进一步的简化得到

$$\begin{aligned} \hat{\beta}_{GLS} &= [\frac{1}{T}\sum_i x_i J_T x_i' + \phi^2 \sum_i (\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})']^{-1}[\frac{1}{T}\sum_i x_i J_T y_i + \phi^2 \sum_i (\bar{x}_i - \bar{\bar{x}})(\bar{y}_i - \bar{\bar{y}})] \\ &= [W_{xx} + \phi^2 B_{xx}]^{-1}[W_{xy} + \phi^2 B_{xy}] \\ &= W_1 \beta_{LSDV} + W_2 \beta_{OLS} \\ \hat{\alpha}_{GLS} &= \bar{y} - \hat{\beta}_{GLS}\bar{\bar{x}} \end{aligned}$$

其中 $W_1 = [W_{xx} + \phi^2 B_{xx}]^{-1} W_{xx}, W_2 = [W_{xx} + \phi^2 B_{xx}]^{-1}\phi^2 B_{xx}$，$W_{xx} = \frac{1}{T}\sum_i x_i J_T x_i'$，$B_{xx} = \sum_i (\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})'$。进一步的，LSDV 估计量（组内去均值）可以表述为 $\hat{\beta}_{LSDV} = W_{xx}^{-1} W_{xy}$，OLS 估计量（组间去均值）可以表示为 $\hat{\beta}_{OLS} = B_{xx}^{-1} B_{xy}$。同时

$$\phi^2 = \frac{\sigma_v^2}{T\sigma_\alpha^2 + \sigma_v^2}$$

[17]第 3 步利用相互垂直向量的逆矩阵形式 $(aM + bP)^{-1} = \frac{1}{a}M + \frac{1}{b}P$，其中 M, P 分别为残差和投影矩阵。

如果 $\phi^2=0$ 转化为 *LSDV* 估计量，如果 $\phi=1$ 转化为 *OLS* 估计量。因此 *GLS* 估计量是 *LSDV* 估计量与 *OLS* 估计量的加权均值估计。当 $T\to\infty$ 的情况下，使用 *LSDV* 就可以，不需要使用 *GLS* 估计。

GLS 估计量方差表示为

$$var(\hat{\beta}_{GLS})=\sigma_v^2[\sum_i x_i' J_T x_i+T\phi^2\sum_i(\bar{x}_i-\bar{\bar{x}})(\bar{x}_i-\bar{\bar{x}})']^{-1}$$

可以证明，*GLS* 估计量相对于 *LSDV* 和 *OLS* 估计量是更有效的估计量。

可行 *GLS* 估计量：在 *GLS* 估计量中并不知道 ϕ，因此需要根据样本数据进行估计。下面给出可行 *GLS* 的估计构造，存在三种估计形式：

1. *Wallase and Hnssian(1969)*：使用 *OLS* 估计残差估计 σ_v^2 和 σ_α^2；
2. *Amemiya (1971)*：使用 *LSDV* 估计残差 σ_v^2 和 σ_α^2；
3. *Swany(1972)*：*LSDV* 与 *Between equation* 估计 σ_v^2 和 σ_α^2；

下面给出第三种形式的估计量构造：首先给定组内去均值的模型，得到 σ_v^2 的估计：

$$y_{it}-\bar{y}_i=(x_{it}-\bar{x}_i)\beta+(v_{it}-\bar{v}_i)\to\sigma_v^2=\frac{1}{n(T-1)}\sum_t\sum_i(v_{it}-\bar{v}_i)^2$$

其次给定组间去均值模型，得到 $\sigma_\alpha^2+\frac{1}{T}\sigma_v^2$ 的整体估计：

$$\bar{y}_i=\bar{x}_i\beta+\alpha+(\alpha_i-\alpha)+\bar{v}_i\to E(\alpha_i-\alpha+\bar{v}_i)^2=\sigma_\alpha^2+\frac{1}{T}\sigma_v^2,\frac{T}{n}\sum\hat{u}_i^2=T\hat{\sigma}_\alpha^2+\hat{\sigma}_v^2$$

Proof: 首先对于组内去均值处理

$$\begin{aligned}E(v_{it}-\bar{v}_i)^2&=E(v_{it}-\frac{1}{T}(v_{i1}+\cdots+v_{iT}))^2\\&=E(v_{it})^2+\frac{1}{T^2}(v_{i1}+\cdots+v_{iT})^2-2E(v_{it}\frac{1}{T}(v_{i1}+\cdots+v_{iT}))\\&=\sigma_v^2+\frac{1}{T^2}\sigma_v^2-\frac{2}{T}\sigma_v^2\\&=\frac{T-1}{T}\sigma^2\\\to\hat{\sigma}_v^2&=\frac{\sum_i\sum_t(\hat{v}_{it}-\bar{v}_i)}{n(T-1)}\end{aligned}$$

同理，对于组间去均值模型可以得到

$$E((\alpha_i-\bar{\alpha})+(\bar{v}_i-\bar{\bar{v}}))^2=E(\alpha_i-\bar{\alpha})^2+E(\bar{v}_i-\bar{\bar{v}})^2=\sigma_\alpha^2(\frac{n-1}{n})+\sigma_v^2(\frac{n-1}{n})$$

从而可以得到期望

$$E(\bar{u}_i-\bar{\bar{u}})^2=\frac{1}{n-1}\sum_i(\bar{u}_i-\bar{\bar{u}})^2\to\sigma_\alpha^2+\frac{1}{T}\sigma_v^2$$

实际上去均值期望等价于：

$$E(v_{it}-\bar{v}_i)^2=(1-\frac{1}{T})\sigma_v^2,\bar{v}_i=\frac{1}{T}\sum_t v_{it}$$

$$E(v_{it}-\bar{v}_t)^2=(1-\frac{1}{n})\sigma_v^2,\bar{v}_t=\frac{1}{n}\sum_i v_{it}$$

最后得到 ϕ 的估计：

$$\hat{\phi}^2=\frac{\hat{\sigma}_v^2}{T\hat{\sigma}_\alpha^2+\hat{\sigma}_v^2}$$

6.3 *Panel Data Test*

6.3.1 *Hausman* 检验

为了选择两种固定效应和随机效应估计量，构造 *Hausman* 检验。其中，随机效应估计量在内生性不成立的时候是有效且一致的，当存在内生性时估计量不一致；固定效应估计量总是一致的，但是不存在内生性时不如随机效应有效。因此，对于固定效应和随机效应估计量选择的检验可以间接转化为对于模型内生性的检验。原假设为 H_0:$cov(\varepsilon_{it},\alpha_i)=0$，表示不存在个体层面不随时间变化因素导致的内生性问题，构造 *Hausman* 统计量：

$$(\hat{\beta}_{RE}-\hat{\beta}_{FE})'[var(\hat{\beta}_{FE})-var(\hat{\beta}_{RE})]^{-1}(\hat{\beta}_{RE}-\hat{\beta}_{FE})\sim\chi^2$$

其中 β_{FE} 表示固定效应估计量，β_{RE} 表示随机效应估计量，在原假设成立的情况下，随机效应估计量是有效的，因而方差更小。

6.3.2 F 检验

给定控制个体固定效应的面板数据模型

$$y_{it}=\alpha_i+x_{it}\beta_i+v_{it}$$

给定如下原假设，现在来考察该面板模型中系数或者固定效应估计量（截距项）是否随个体而改变：

1. H_1：$\beta_1\ldots=\beta_n$，允许截距项 α_i 存在异质性，检验 β_i 是否存在个体异质性；
2. H_2：$\alpha_1=\cdots=\alpha_n$，允许系数 β_i 存在异质性，检验个体而已的截距项 α_i 是否具有异质性，注意到如果系数 β_i 存在个体异质性，那么个体层面的截距项 α_i 同质是没有意义的，因而该假设检验没有存在的必要性；
3. H_3：$\alpha_1=\cdots=\alpha_n$,$\beta_1=\cdots=\beta_n$，检验截距项和系数项是否均为同质性；
4. H_4：$\alpha_1=\cdots=\alpha_n$ *given* $\beta_1=\cdots=\beta_n$，该假设给定系数项 β_i 具有同质性，检验个体层面的截距项 α_i 是否具有同质性；区别于假设检验 H_3，如果 H_3 被拒绝，实际上并不清楚究竟是因为系数具有异质性还是截距项具有异质性，或者两者均存在，无法分离三种效应，因此额外增加检验 H_4 用来区分不同情形；

针对上述三种可行检验 H_1,H_3,H_4，依次进行考察。对于上述联合检验的形式，一般使用 F 检验进行处理。在截面数据中，对于原假设 H_0：$R_{k\times q}\beta=r$，假定 H_0 成立时（限制模型）残差估计量表示为 SSR_r，H_0 不成立时（非限制模型）残差估计量表示为 SSR_u，一般而言限制模型的残差更大，这是因为在模型中施加更多假定从而导致模型解释力有所下降，因而 $SSR_r>SSR_u$，此时构造 F 检验：

$$F=\frac{(SSR_r-SSR_u)/q}{SSR_u/df}\sim F_{q,df}$$

其中 q 表示约束条件个数，df 表示非限制模型中的自由度。相应的，将截面数据的检验方式应用于面板数据检验中。

首先，对于非限制模型（*unrestricted model*），可知截距项 α_i 和系数 β_i 均具有个体层面异质性，矩阵形式表示为

$$\begin{aligned}Y&=\begin{bmatrix}\alpha_1 l_T\\ \ldots\\ \alpha_n l_T\end{bmatrix}+\begin{bmatrix}x_1\beta_1\\ \ldots\\ x_n\beta_n\end{bmatrix}+\begin{bmatrix}v_1\\ \ldots\\ v_n\end{bmatrix}=\begin{bmatrix}l_T & & \\ & \ldots & \\ & & l_T\end{bmatrix}\begin{bmatrix}\alpha_1\\ \ldots\\ \alpha_n\end{bmatrix}+\begin{bmatrix}x_1 & & \\ & \ldots & \\ & & \beta_n\end{bmatrix}+\begin{bmatrix}v_1\\ \ldots\\ v_n\end{bmatrix}\\ &=\begin{bmatrix}l_T x_1 & & \\ & \ldots & \\ & & l_T x_n\end{bmatrix}\begin{bmatrix}(\alpha_1,\beta_1)'\\ \ldots\\ (\alpha_n,\beta_n)'\end{bmatrix}+\begin{bmatrix}v_1\\ \ldots\\ v_n\end{bmatrix}=Z_1\delta_n+u\end{aligned}$$

给定残差矩阵 $M_1 = I - Z_1(Z_1'Z_1)'Z_1'$，可知该模型 $SSR = S_1 = e'e = Y'M_1Y$。δ_n 表示截距项和系数在内的所有参数矩阵。

其次，对于最大化限制模型（*most restricted model*），即不存在个体层面异质的截距项 α_i 以及系数 β_i，矩阵形式表示为

$$Y = \alpha \begin{bmatrix} l_T \\ \dots \\ l_T \end{bmatrix} + \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} \beta + \begin{bmatrix} v_1 \\ \dots \\ v_n \end{bmatrix} = (l_{nT} X) \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + u = Z_3 \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + u$$

给定残差矩阵 $M_3 = I - Z_3(Z_3'Z_3)'Z_3'$，可知该模型 $SSR = S_3 = e'e = Y'M_3Y$。

最后，对于中间限制模型（*intermediate model*），即假定只存在个体层面异质性的截距项 α_i 而系数具有同质性 β，矩阵形式表示为

$$Y = \begin{bmatrix} \alpha_1 l_T \\ \dots \\ \alpha_n l_T \end{bmatrix} + \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} \beta + u = \begin{bmatrix} l_T & & & x_1 \\ & \dots & & \dots \\ & & l_T & x_n \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \dots \\ \alpha_n \\ \beta \end{bmatrix} + u = Z_2 \begin{bmatrix} \alpha_1 \\ \dots \\ \alpha_n \\ \beta \end{bmatrix} + u$$

给定残差矩阵 $M_2 = I - Z_2(Z_2'Z_2)'Z_2$，可知该模型 $SSR = S_2 = e'e = Y'M_2Y$。现在可以分情况对上述三种情形进行检验。在正式的进行检验之前，先做一些准备工作：首先，需要明确分子是谁减谁，即根据原假设成立于不成立确定限制模型与非限制模型的残差项，并确定两者大小（总是限制模型减非限制模型）；其次，需要确定分子和分母的自由度。分子考察的是限制模型的约束条件，分母考察的是非限制模型的自由度。

对于 H_3，假定不存在个体层面异质性，受限制模型为最大化限制模型（如果 H_3 成立），非限制模型为模型 *1*（如果 H_3 不成立 α_i, β_i 即均为异质性），相应的可以构造检验：

$$F_{H_3} = \frac{(S_3 - S_1)/[(n-1)(k+1)]}{S_1/[nT - n(k+1)]} \sim F_{(n-1)(k+1), nT-n(k+1)}$$

其中约束条件是 $\alpha_1 = \cdots = \alpha_n$，共 $n-1$ 个限制；$\beta_1 = \cdots = \beta_n$ 共有 $(n-1)k$ 个限制（β 包含 k 个自变量），因此限制个数为 $(n-1)(k+1)$。对于分母的自由度，给定无限制模型，自由度表示为 $df = nT - n(1+k)$，其中 $(1+k)$ 表示 α 和 β。

对于 H_1，假定只存在个体层面异质性的截距项，受限制模型为中间限制模型（如果 H_1 成立），非限制模型为模型 *1*（如果 H_1 不成立 α_i, β_i 即均为异质性），相应的可以构造检验：

$$F_{H_1} = \frac{(S_2 - S_1)/[(n-1)k]}{S_1/[nT - n(k+1)]} \sim F_{(n-1)k, nT-n(k+1)}$$

其中约束条件是 $\beta_1 = \cdots = \beta_n$ 共有 $(n-1)k$ 个限制（β 包含 k 个自变量），因此限制个数为 $(n-1)k$。

对于 H_4，假定不存在个体层面异质性，受限制模型为模型最大化限制模型（如果 H_4 成立），非限制模型为模型中间限制模型（如果 H_4 不成立 α_i 具有异质性），这个检验的构造就是为了区分 H_3 被拒绝时的几种情形，在这里如果 H_4 不成立则表明存在 α_i 层面的异质性，而 β_i 的异质性没有探讨（可以根据 H_4 对称的给出）：

$$F_{H_4} = \frac{(S_3 - S_2)/[(n-1)]}{S_2/[nT - (k+n)]} \sim F_{(n-1), nT-(k+n)}$$

其中约束条件是 $\alpha_1 = \cdots = \alpha_n$，共 $n-1$ 个限制；对于分母的自由度，给定无限制模型，自由度表示为 $df = nT - (n+k)$，包括 n 个 α 以及 *1* 个相同的 β（包含 k 个自变量）。

6.4 *Dynamic Panel Model*

6.4.1 *LSDV*

本部分给出动态面板模型，并进一步证明为何 *LSDV* 估计量存在问题，在此基础上下一节给出动态面板的 *GMM* 估计量。给出如下面板模型：

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + v_{it}, i \in [1,n], t \in [1,T]$$

可以给出 *LSDV* 估计量（组内去均值处理）：

$$\begin{aligned} y_{it} - \bar{y}_i &= \gamma[y_{i,t-1} - \bar{y}_{i,-1}] + v_{it} - \bar{v}_i \\ \hat{\gamma}_{LSDV} &= [\sum_i \sum_t (y_{i,t-1} - \bar{y}_{i,-1})^2]^{-1} [\sum_i \sum_t (y_{i,t-1} - \bar{y}_{i,-1})(y_{it} - \bar{y}_i)] \\ &= \gamma + [\sum_i \sum_t (y_{i,t-1} - \bar{y}_{i,-1})^2]^{-1} [\sum_i \sum_t (y_{i,t-1} - \bar{y}_{i,-1})(v_{it} - \bar{v}_i)] \end{aligned}$$

其中 $\bar{y}_i = \frac{1}{T}(y_{i1} + \cdots + y_{it}) . \bar{y}_{i,-1} = \frac{1}{T}(y_{i1} + \cdots + y_{i,t-1})$。注意到，如果 $E(\bar{y}_{i,-1}\bar{v}_i) = \frac{1}{nT}\sum_i \sum_t (y_{i,t-1} - \bar{y}_{i,-1})(v_{it} - \bar{v}_i) = 0$，此时 *LSDV* 估计量是一致的，不存在估计问题。但是，利用迭代关系处理得到

$$y_{it} = v_{it} + \gamma v_{i,t-1} + \cdots + \gamma^{t-1} v_{i1} + \gamma^t y_{i0} + \alpha_i(1 + \gamma + \gamma^2 + \cdots + \gamma^{t-1})$$

该式可以展开为

$$\begin{aligned} y_{i0} &= y_{i0} \\ y_{i1} &= v_{i0} + \gamma y_{i0} + \alpha_i \\ y_{i2} &= v_{i2} + \gamma v_{i1} + \gamma^2 y_{i0} + \alpha_i(1 + \gamma) \end{aligned}$$

该模型的含义是说利用滞后期的被解释变量 y_{it} 回归到 y_{it} 上面。取均值处理得到

$$\frac{1}{T}(y_{i1} + y_{i2} + \cdots + y_{i,T-1}) = \frac{1}{T} \begin{bmatrix} (1 + \gamma + \cdots + \gamma^{t-1}) v_{i1} \\ +(1 + \gamma + \cdots + \gamma^{t-2} v_{i2}) \\ \cdots \\ +v_{i,t-1} \end{bmatrix} + f(y_{i0}, \alpha)$$

期望表示为

$$\begin{aligned} E(\bar{y}_{i,-1}\bar{v}_i) &= \frac{\sigma^2}{T} \begin{bmatrix} (1 + \gamma + \cdots + \gamma^{t-1}) \\ +(1 + \gamma + \cdots + \gamma^{t-2} \\ \cdots \\ +1 \end{bmatrix} \\ &= \frac{\sigma^2}{T}(\frac{1 - \gamma^{T-1}}{1 - \gamma} + \frac{1 - \gamma^{T-2}}{1 - \gamma} + \cdots + \frac{1 - \gamma}{1 - \gamma}) \\ &= \frac{\sigma^2}{T(1 - \gamma)}[(T - 1) - \gamma(1 + \gamma + \cdots + \gamma^{T-2})] \\ &= \frac{\sigma^2}{T(1 - \gamma)}[1 - \frac{1}{T} - \frac{\gamma(1 - \gamma^{T-1})}{T(1 - \gamma)}] = \sigma^2 h_T \sim o(\frac{1}{T}) \end{aligned}$$

可以看出，如果说 T 很小，期望不等于 0，此时存在估计偏误，偏离误差的幅度为 $\frac{1}{T}$（*order of* $\frac{1}{T}$），只有当 $T \to \infty$ 的时候误差才会消除。如果说 T 足够大，$\hat{\gamma}$ 是一致的但是存在渐进偏误。因此，*LSDV* 估计量在 T 不是很大的时候存在估计偏误，不是一个足够好的偏误。既然 *LSDV* 估计量存在偏误，那么可以给出偏差修正后的 *LSDV* 估计量：

$$\hat{\gamma}_{BC} = \hat{\gamma}_{LSDV} + [\sum_i \sum_t (y_{i,t-1} - \bar{y}_{i,-1})^2]^{-1} \frac{1}{T} \frac{1}{1 - \hat{\gamma}}$$

重新来理解一下这里的逻辑:

$$E(\hat{\gamma}_{LSDV}) = \gamma + 0(\frac{1}{T})$$
$$E(\sum_t (y_{i,t-1} - \bar{y}_{i,-1})(v_{it} - \bar{v}_i)) = E(\sum_t y_{i,t-1}v_{it} - T\bar{y}_{i,-1}\bar{v}_i) = -E(T\bar{y}_{i,-1}\bar{v}_i) = o(1)$$

相应的渐进性质表示为

$$\sqrt{nT}(\hat{\gamma}_{LSDV} - \gamma + [\sum_i \sum_t (y_{i,t-1} - \bar{y}_{i,-1})^2]^{-1} T\bar{y}_{i,-1}\bar{v}_i) = \sqrt{nT}[\sum_i \sum_t (y_{i,t-1} - \bar{y}_{i,-1})^2]^{-1} \sum_i \sum_t y_{i,t-1}v_{it}$$
$$\to N(0, \phi)$$

6.4.2 *IV-GMM* 估计量

考虑一阶差分处理

$$\Delta y_{it} = \gamma \Delta y_{i,t-1} + \Delta v_{it}$$

其中 $\Delta y_{it} = y_{it} - y_{i,t-1}$，注意到 $cov(\Delta y_{i,t-1}, \Delta v_{it}) \neq 0$，这是因为 $y_{i,t-1}$ 和 $v_{i,t-1}$ 存在相关性导致。这种情况下使用 OLS 估计存在偏误，注意到 $E(y_{i,t-s}\Delta v_{it}) = 0$ 对于任意的 $s \geq 2$，也即只要滞后期超过 $T-1$ 相关性便不存在，因此这种滞后项可以用作 IV 进行估计。

给定 $(T-1) \times m$ 矩阵

$$Z_i = \begin{bmatrix} y_0 & & & & & \\ & y_0 & y_1 & & & \\ & & \cdots & & & \\ & & & y_0 & \cdots & y_{T-2} \end{bmatrix}$$

其中第一行表示 Δv_2 的 IV 是 y_0，相应的第二行表示 Δv_3 的 IV 是 y_0, y_1，最后一行是 Δv_{iT} 的 IV 是 $y_0, \ldots, y_{i,T-2}$。其含义就是必须利用更早之前的被解释变量作为对应的 IV 进行估计以保证不存在内生性。其中 $m = 1 + 2 + \cdots + (T-1) = \frac{T(T-1)}{2}$，表示可以用于模型估计的 IV 的数量，表示矩阵的列数，矩阵行数表示为 $T-1$。给定残差矩阵 $\Delta u_i = (\Delta v_{i2}, \ldots, \Delta v_{iT})'$，矩条件可以给定为

$$g_i = z_i' \Delta u_i \to g = Z'u$$
$$Z = \begin{bmatrix} z_1 \\ \ldots \\ z_n \end{bmatrix}_{n(T-1) \times m} \qquad u = \begin{bmatrix} u_1 \\ \ldots \\ u_n \end{bmatrix}_{n(T-1) \times 1}$$

为了更好的理解，把矩条件拆开表示为

$$g_i = \begin{bmatrix} y_{i0}\Delta v_{i2} \\ y_{i0}\Delta v_{i3} \\ y_{i1}\Delta v_{i3} \\ \cdots \\ y_{i,T-2}\Delta v_{iT} \end{bmatrix} = 0$$

即所有的滞后期工具变量与随机项满足的矩条件，一共有 m 个。使用 GMM 进行求解：最小化矩条件可以得到

$$min\ (\Delta y - \gamma \Delta y_{-1})' Z A_n Z' (\Delta y - \gamma \Delta y_{-1})$$
$$\to \hat{\gamma} = (\Delta y_{-1}' Z A_n Z' \Delta y_{-1})^{-1} (\Delta y_{-1}' Z A_n Z' \Delta y)$$

其中 $A_n = (\frac{1}{n}\sum_i z_i'\hat{\Delta}u_i\hat{\Delta}u_i'z_i')^{-1}$，或者采用如下两步走方法确定：

$$A_n = (\frac{1}{n}\sum_i z_i'Hz_i)^{-1}$$

$$H = \begin{bmatrix} 2 & -1 & 0 & \\ -1 & 2 & -1 & \\ & & \cdots & \\ & & & 2 \end{bmatrix}$$

注意到 $E(\Delta u_i \Delta u_i') = E\begin{bmatrix} \Delta v_{i2} \\ \dots \\ \Delta v_{iT} \end{bmatrix}(\Delta v_{i,2}, \dots, \Delta v_{i,T})$ 该矩阵表示方差-协方差矩阵，而在该模型中距离超过 2 个的协方差关系为 0（不具有相关性）。其中，$E(\Delta v_{i2}, \Delta v_{i2}) = var(v_{i2}) + var(v_{i1}) = 2$，因而对角线的方差表示为 2，临近左右侧的协方差表示为 1，其余位置均为 0。

7 *Spatial Model*

7.1 空间自回归 *SAR*

Spatial Autoregressive Model（空间自回归）给定为

$$\begin{aligned} y_i &= \lambda(w_{i1}y_1 + \cdots + w_{in}y_n) + x_i'\beta + v_i \\ &= \lambda\sum_{i\neq j} w_{ij}y_j + x_i'\beta + v_i \end{aligned}$$

这表示周围的被解释变量 y_j 会影响 y_i 的数据生成过程，例如空间溢出效应、空间竞争效应、网络效应与蔓延效应等。其中 w_{ij} 表示自回归权重，可以解释为连接矩阵、地理距离、经济距离等因素的影响程度；既可以人为的给定 w_{ij} 的设定，例如 $w_{ij} = \frac{1}{d_{ij}}$，其中 d_{ijj} 表示两者间的距离；也可以利用非参数方式 $w_{ij} = h(d_{ij}), h(\cdot)$ 表示未知函数，利用非参估计方法确定权重 w_{ij} 设定。在空间计量模型中，权重 w_{ij} 的设定是十分关键的，能够来自于经济理论是最自然的设定形式。简单起见，本节假定 w_{ij} 是给定的，不去考虑权重的设定细节，将注意力放在空间计量模型的参数估计上面。在空间自回归模型中最为重要的是参数 λ，相比之下 β 并不是很重要，如果可以确定参数 λ 的取值，因此可以将问题转化为 OLS 进行估计，所以自回归参数 λ 是估计的关键所在。

转化为矩阵形式表示为

$$\begin{aligned} Y_n &= \lambda W_n Y_n + X_n\beta + V_n \\ \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix} &= \lambda \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ & \cdots & \\ w_{n1} & \cdots & w_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix} + \begin{bmatrix} x_1 \\ \cdots \\ x_n \end{bmatrix} \beta + \begin{bmatrix} v_1 \\ \cdots \\ v_n \end{bmatrix} \end{aligned}$$

接下来，要考察 (λ, β) 的估计问题，采用不同的方式进行识别与估计，并讨论其等价性。

给定一般的线性回归 $y_n = x_n\beta + v_n$，随机项服从 $v_n \sim N(0, \sigma^2)$，可以得到

$$f(y_n) = (\frac{1}{\sqrt{2\pi\sigma^2}})exp[-\frac{(y_n - x_n\beta)'(y_n - x_n\beta)}{2\sigma^2}]$$

相应的对于空间自回归模型，如果假定 $Z_n = (W_nY_n, X_n), \delta = (\lambda, \beta')'$，因此模型转化为

$$Y_n = Z_n\delta + V_n \to \delta_{OLS} = (Z'Z)^{-1}(Z'Y) = \beta + (Z'Z)^{-1}(Z'v_n)$$

其中 $cov(Z_nv_n) = cov(Y_n, v_n) \neq 0$，存在内生性问题，直接使用 OLS 估计存在偏误，这说明对于空间自回归使用 OLS 估计是有偏和不一致的，所以需要考虑 OLS 之外的估计方法。

7.2 *MLE* 估计量

假定 X 服从分布 $f_X(x)$，同时 $Y = g(X)$，其中 g 是单调函数，定义

$$X = \{x : f_X(x) > 0\}, Y = \{y : y = g(x), x \in X\}$$

例如 $Y_n = X_n\beta + V_n = g(X_n)$，相应的 $V_n = Y_n - X_n\beta = g^{-1}(y_n)$。对应的，在空间自回归中可以得到

$$V_n = Y_n - \lambda W_nY_n - X_n\beta = (I_n - \lambda W_n)Y_n - X_n\beta = S_n(\lambda)Y_n - X_n\beta,\ S_n = (I_n - \lambda W_n)$$

相应的 Y 的分布给定为

$$f_Y(y) = f_X(g^{-1}(y))[\frac{d}{dy}g^{-1}(y)],\ y \in Y;\ f_Y(y) = 0,\ otherwise$$

对数似然函数给定为

$$lnL_n(\lambda,\beta,\sigma^2)=-\frac{n}{2}ln(2\pi)-\frac{n}{2}ln\sigma^2+ln|S_n(\lambda)|-\frac{1}{2\sigma^2}[S_n(\lambda)Y_n-X_n\beta]'[S_n(\lambda)Y_n-X_n\beta]$$

其中 $S_n(\lambda)=(I_n-\lambda W_n),|S_n|$ 表示雅可比项（行列式）。一阶条件表示为

$$\begin{aligned}\frac{\partial lnL}{\partial\beta}&=\frac{1}{\sigma^2}X_n'(Y_n-\lambda W_nY_n-X_n\beta)=0\to E(X_n'v_n)=0\\\frac{\partial lnL}{\partial\lambda}&=-tr(G_n)+\frac{1}{\sigma^2}(W_nY_n)'(Y_n-\lambda W_nY_n-X_n\beta)=0\\\frac{\partial lnL}{\partial\sigma^2}&=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^2}e_n'e_n=0\end{aligned}$$

其中 $e_n=Y_n-\lambda W_nY_n-X_n\beta$。给定求导法则 $\frac{\partial lnDet(F)}{\partial x}=tr(\frac{\partial F}{\partial x}F^{-1})$，相应的有

$$\frac{\partial ln|S_n|}{\partial\lambda}=\frac{\partial|I_n-\lambda W_n|}{\partial\lambda}=-tr(G_n),\ G_n=W_nS_n^{-1}=W_n(I_n-\lambda W_n)$$

可以考察一阶条件的含义：(1) 根据 $E(X_n'V_n)=0$ 可以得到；(2) 根据空间自回归模型可以得到

$$S_nY_n=X_n\beta+V_n\to Y_n=S_n^{-1}(X_n\beta+V_n)\to W_nY_n=G_n(X_n\beta+V_n)$$

可以很清楚的看到 $E(Y_nv_n)\neq0$，存在内生性问题，这表明使用 OLS 估计方法存在偏误；对于 (2) 式代入得到 $E[-tr(G_n)+\frac{1}{\sigma^2}(W_nY_n)'V_n]=0$ 成立。根据 $W_nY_n=W_nS_n^{-1}X_n\beta+W_nS_n^{-1}V_n$，进一步该式可以转化为

$$\begin{aligned}\frac{\partial lnL}{\partial\lambda}&=-tr(G_n)+\frac{1}{\sigma^2}(W_nY_n)'(Y_n-\lambda W_nY_n-X_n\beta)\\&=\frac{1}{\sigma^2}(G_nX_n\beta)'V_n+\frac{1}{\sigma^2}(G_nV_n)'V_n-tr(G_n)\\&=\frac{1}{\sigma^2}(G_nV_n)'V_n-tr(G_n)\end{aligned}$$

其中根据矩条件 $E(X_nv_n)=0$ 可以得到 $(G_nX_n\beta)'V_n=0$。

尽管可以给出直接的一阶形式，但是由于雅可比项的存在导致无法求解显式解，因此利用数值解法更为方便。注意到，在空间自回归的估计中 MLE 的一阶条件是 *Baseline*，有助于进行计量理论的分析，特别是后续进行三种估计量的对比分析时需要利用。

7.3 *2SLS* 估计量

给定空间自回归模型，相应的约简式估计转化为

$$W_nY_n=W_nS_n^{-1}[X_n\beta+v_n]=G_nX_n\beta+G_nv_n$$

其中 $E(V_n'G_nV_n)=\sigma^2tr(G_n),G_n=W_nS_n^{-1}=W_n(I_n-\lambda W_n)^{-1}$，这表明 (G_nX_n,G_nv_n) 回归存在相关性，利用 OLS 估计存在偏误，转而利用 IV 进行估计。

Kelejian and Prucha(1998) 建议使用

$$H=(X_n,W_nX_n,W_n^2X_n,\dots)$$

作为 IV；这是因为近似展开得到

$$G_n=W_n(I_n-\lambda W_n)^{-1}=W_n(I_n+\lambda W_n+\lambda^2W_n^2+\dots)=W_n+\lambda W_n^2+\lambda^2W_n^3+\dots$$

当 n 非常大的时候，*Lee*（*2003*）指出最优 IV 矩阵采用如下形式更有效：

$$H_n^*=(W_nS_n^{-1}X_n\beta_0,X_n)$$

定义 $Z_n = (W_nY_n, X_n), \delta = (\lambda, \beta')'$，自回归模型转化为

$$Y_n = Z_n\delta + v_n$$

给定 (W_nY_n, X_n) 的工具变量 H_n[18]，2SLS 估计量给定为

$$\delta_{2SLS} = (Z_n'H_n(H_n'H_n)^{-1}H_n'Z_n)^{-1}(Z_n'H_n(H_n'H_n)^{-1}H_n'Y_n)$$

但是，2SLS 估计量不能用于 *pure SAR*($\beta = 0$) 模型参数估计。

7.4 *GMM:Linear and Quadratic Moments*

给定空间自回归模型：

$$W_nY_n = W_nS_n^{-1}[X_n\beta + v_n] = G_nX_n\beta + G_nv_n$$

定义 (W_nY_n, X_n) 的工具变量矩阵 Q_n，随机项 $v_n(\theta) = S_n(\lambda)Y_n - X_n\beta$，线性矩条件给定为 $Q_n'v_n(\theta)$。现在考虑工具变量矩阵 $n \times n$ 的矩阵 P_n，其中 $tr(P) = 0$，相应的线性矩条件表示为 $P_nv_n(\theta)$，需要满足如下条件：

1. 与随机项无关：$E((P_nv_n)'v_n) = E(v_n'P_nv_n) = \sigma_0^2tr(P) = 0$[19]；
2. 与 W_nY_n 相关：$E((P_nv_n)'W_nS_n^{-1}v_n) = \sigma_0^2tr(P_n'W_nS_n^{-1}) \neq 0$;

其中 P 取值包括（保证迹为 0）：

$$P_1 = W_n - \frac{tr(W_n)}{n}I_n, P_2 = W_n^2 - \frac{tr(W_n^2)}{n}I_n, \dots$$

矩条件相应的表示为 $E(Q_n'v_n) = 0, E(v_n'P_nv_n) = 0$，GMM 给定为

$$g_n(\theta) = \begin{bmatrix} v_n'(\theta)P_{1n}v_n(\theta) \\ \cdots \\ v_n'(\theta)P_{mn}v_n(\theta) \\ Q_n'v_n(\theta) \end{bmatrix}$$

最优 GMM 给定为最小化如下表达式：

$$g_n'(\theta)[var(g_n)]^{-1}g_n(\theta)$$

可以得到有效 GMM 估计对应的工具变量

$$Q^* = [G_nX_n\beta_0, X_n], P_n^* = G_n - \frac{tr(G_n)}{n}I_n$$

其中 Q^* 对应于 (λ, β)，P^* 对应于 β。注意到，Q^* 与 2SLS 最优 IV 相对应，可知 2SLS 是 GMM 的一种特殊形式，即有效 GMM 情况下 GMM 退化为 2SLS 形式。

进一步考察 MLE 估计量与 GMM 估计量的关系：MLE 估计中的一阶条件如何得到最优的 P^*, Q^*？重新考察对于待估计参数 λ 的一阶条件

$$\begin{aligned} \frac{\partial lnL}{\partial \lambda} &= \frac{1}{\sigma^2}[V_n'G_nV_n - \sigma^2tr(G_n)] \\ &= \frac{1}{\sigma^2}V_n'(G_n - \frac{tr(G_n)}{n}I_n)V_n \end{aligned}$$

其中 $\sigma^2tr(G_n) = E(V_n'G_nV_n) = \sigma^2\frac{tr(G_n)}{n}$，据此可以得到 $P^* = G_n - \frac{tr(G_n)}{n}I_n$。当随机分布不是正态分布时，GMM 相比于 MLE 估计更为有效；当随机分布是正态分布时，有效 GMM 与正态假设下的 MLE 估计是一致的。一般而言，GMM 计算更加方便，往往效率更高。MLE 更多的是提供一种理论 *benchmark* 用于思考。

[18]模型中假定 X_n 是外生变量，不需要构造单独的 IV 处理。

[19]$E(\varepsilon'A\varepsilon) = \sigma^2tr(A)$, 其中 $cov(A\varepsilon) = 0, var(\varepsilon) = \sigma^2$

8 *Reference*

1. *Yu Jihai, Lecture Notes On Graduate Level Economics Theory I, GSM, 2023.*
2. *William H. Green, Econometric Analysis, Eight Edition, Prentice Hall, 2018.*
3. *Jeffrey M. Wooldridge, Econometric Analysis of Cross Section and Panel Data, MIT Press.*
4. *Bruce E. Hansen, Econometrics, 2021.*
5. *MacKinnon, James G., and Russell Davidson, Foundations of Econometrics, 2021.*