高级计量经济学II课堂笔记*

刘旭浚† 赵凡青‡

2024 年 6 月：第2版
2020 年 3 月：第1版

*笔记主要内容由赵凡青于2020年9月份的第一版整理完成，第2版内容由刘旭浚进行的补充与调整（作者排序按照拼音顺序）。在第二版笔记的整理过程中，同样参考了石邦汝学长的笔记，特此感谢。本笔记主要为上海财经大学经济学院李聪老师所开设的高级计量经济学II的课堂内容笔记，仅供学习参考使用。

†上海财经大学金融学院2023级博士研究生；
‡上海财经大学金融学院2019级博士研究生。

写在前面

特殊时期为了便于自己复习，顺便学习下LaTeX的用法，特此整理高计II笔记。有错误还请及时告知我。

赵凡青，2020年3月

偶然发现的很棒的计量笔记，顺便为了整理自己学习计量经济学以来的笔记和思考，决定在此基础上对赵凡青学长的笔记进行一定程度上的更新与整理。本次更新纯属个人行为，如有与之前笔记的龃龉之处,，或有所不足，或有所纰漏，也欢迎与我讨论。

刘旭浚，2024年6月

目录

第一章　计量I回顾

1.1　最小二乘

1.1.1　最小二乘估计

对于经典的线性回归模型

$$y = \mathrm{X}\beta + u \tag{1.1}$$

其中，y和u为$n \times 1$维的向量，$\mathrm{X} = (x_1, x_2, \ldots, x_k)$为$n \times k$维的矩阵、$x_i$为$n \times 1$的向量，$\beta$为$k \times 1$维的向量。[1]在古典假设成立的情况下，参数估计量为BLUE（Best Linear Unbiased Estimator）.

为求解所估计的$\hat{\beta}$，根据OLS定义，最小化残差平方和：

$$\min_{\beta} \underbrace{(y - \mathrm{X}\beta)'}_{1\times n} \underbrace{(y - \mathrm{X}\beta)}_{n\times 1}$$

为求解β，有

$$\begin{aligned} \text{FOC}: \frac{\partial}{\partial \beta}(y'y - \beta'\mathrm{X}'y - y'\mathrm{X}\beta + \beta'\mathrm{X}'\mathrm{X}\beta) \\ = -\mathrm{X}'y - \mathrm{X}'y + (\mathrm{X}'\mathrm{X} + \mathrm{X}'\mathrm{X})\beta \\ = -2\mathrm{X}'y + 2\mathrm{X}'\mathrm{X}\beta = 0 \\ \Longrightarrow \hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y \\ = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'u + \beta \end{aligned} \tag{1.2}$$

由于$\hat{\beta}$为u的线性组合，在古典假设中，我们假定$u \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_n)$，因此为求得估计量$\hat{\beta}$的分布，我们仅需求得$\hat{\beta}$的均值与方差即可。

[1]通常，我们设定x_1为单位向量为保证模型包括了常数项。

$\hat{\beta}$的均值.

$$\begin{aligned} \mathrm{E}[\hat{\beta}] &= \mathrm{E}[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y] \\ &= \mathrm{E}[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'(\mathrm{X}\beta+u)] \\ &= \beta+\mathrm{E}[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'u] \\ (\because \mathrm{E}[\mathrm{X}'u]=0) &= \beta \end{aligned} \tag{1.3}$$

由于$\mathrm{E}[\hat{\beta}]=\beta$，故OLS估计$\hat{\beta}$为无偏估计量。

$\hat{\beta}$的方差.

$$\begin{aligned} \mathrm{V}ar(\hat{\beta}) &= \mathrm{E}[(\hat{\beta}-\beta)'(\hat{\beta}-\beta)] \\ &= \mathrm{E}[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'uu'\mathrm{X}(\mathrm{X}'\mathrm{X})] \\ &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{E}[uu']\mathrm{X}(\mathrm{X}'\mathrm{X}) \\ &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\sigma^2\mathrm{I}\mathrm{X}(\mathrm{X}'\mathrm{X}) \\ &= \sigma^2(\mathrm{X}'\mathrm{X})^{-1} \end{aligned} \tag{1.4}$$

当σ^2已知时，$\hat{\beta}\sim\mathrm{N}(\beta,\sigma^2(\mathrm{X}'\mathrm{X})^{-1})$。但实际中，我们通常并不知道随机扰动项方差的真值（true value），因此，我们需要估计σ^2。

$$\hat{\sigma^2}=\frac{\hat{u}'\hat{u}}{n-k} \tag{1.5}$$

其中

$$\begin{aligned} \hat{u} &= \hat{y}-\mathrm{X}\hat{\beta} \\ &= \underbrace{[\mathrm{I}_n-\overbrace{\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'}^{\text{P：投影矩阵（Projection matrix）}}]}_{\text{M：消灭矩阵（Annihilator matrix）}}y \end{aligned} \tag{1.6}$$

对于矩阵P和矩阵M而言，他们具有如下性质：（1）$\mathrm{PX}=\mathrm{X},\mathrm{P}u=0,\mathrm{MX}=0$；（2）P和M都是对称矩阵和

幂等矩阵。[2] 回到Equation 1.5，我们需要求得$\mathrm{E}[\hat{u}'\hat{u}]$来得到无偏估计量$\hat{\sigma}^2$，故有

$$\begin{aligned}
\mathrm{E}[\hat{u}'\hat{u}] &= u'\mathrm{M}u \quad (\because \hat{u} = \mathrm{M}y = \mathrm{M}(\mathrm{X}\beta + u) = \mathrm{M}u) \\
&= tr(\mathrm{E}[u'\mathrm{M}u]) \\
&= \mathrm{E}[tr(u'\mathrm{M}u)] \text{（线性算子可交换运算顺序）} \\
&= \mathrm{E}[tr(\mathrm{M}uu')] \mathrm{tr(AB)=tr(BA)} \\
&= tr(\mathrm{E}[\mathrm{M}uu']) \\
&= tr(\mathrm{E}[\mathrm{M}\mathrm{E}[uu'|\mathrm{X}]]) \\
&= \sigma^2\mathrm{E}[tr(\mathrm{M})] \\
&= \sigma^2\mathrm{E}[tr(\mathrm{I}) - tr(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}')] \text{（tr(A+B)=tr(A)+tr(B)）} \\
&= \sigma^2\mathrm{E}[tr(\mathrm{I}) - tr((\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X})] = \sigma^2(n-k) \qquad (1.7)
\end{aligned}$$

故，可得σ^2的无偏估计如下：

$$\mathrm{E}[\hat{\sigma}^2] = \mathrm{E}[\frac{\hat{u}'\hat{u}}{n-k}] = \sigma^2 \qquad (1.8)$$

所以估计量$\hat{\beta}$的分布为：

$$\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2(\mathrm{X}'\mathrm{X})^{-1}) \qquad (1.9)$$

将向量$\hat{\beta}$中的元素标准化后，有

$$z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2(\mathrm{X}'\mathrm{X})^{-1}_{jj}}} \sim \mathrm{N}(0,1) \qquad (1.10)$$

1.1.2 参数检验

*t*检验：单变量

但通常情况下，σ^2是未知的，故我们需要使用σ^2的无偏估计来进行代换。

$$\begin{aligned}
\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2(\mathrm{X}'\mathrm{X})^{-1}_{jj}}} &= \frac{(\hat{\beta}_j - \beta_j)/\sqrt{\sigma^2(\mathrm{X}'\mathrm{X})^{-1}_{jj}}}{\sqrt{[(n-k)\hat{\sigma}^2/\sigma^2]/(n-k)}} \\
&= \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{(n-k)}/(n-k)}} \qquad (1.11)
\end{aligned}$$

[2] 矩阵M是将向量变换到与x向量所张成平面垂直的平面上，所以MX = 0；矩阵P是将向量变换到与x所张成平面平行的向量（该平面的投影projection）,所以PX = X；P和M是一组正交分解，所以$\mathrm{P} + \mathrm{M} = \mathrm{I}_n$，$\mathrm{P}y + \mathrm{M}y = y$向量加法（平行四边形法则），P称为projection matrix，M称为orthogonal projection matrix，M和P都是symetric idempotent matrix（对称、幂等矩阵）。

其中，分母中的$(n-k)s^2/\sigma^2 \sim \chi^2_{(n-k)}$，由Theorem 1.1.1可知，

$$(n-k)\hat{\sigma}^2/\sigma^2 = \frac{\hat{u}'\hat{u}}{\sigma^2} = \frac{u'\mathrm{M}u}{\sigma^2} = (\frac{u}{\sigma})'\mathrm{M}(\frac{u}{\sigma}) \sim \chi^2(tr(\mathrm{M})) = \chi^2(n-k)$$

其中$u/\sigma \sim \mathrm{N}(0, 1_n)$.

Theorem 1.1.1. 标准正态向量中幂等二次型的分布

if $\underset{m\times 1}{z} \sim \mathrm{N}(0, \mathrm{I}_m)$, *matrix A is a* $m \times m$ *symetric idempotent matrix, then*

$$z'\mathrm{A}z \sim \chi^2(df)$$

其中，χ^2分布的自由度$df = rank(\mathrm{A}) = tr(\mathrm{A})$（幂等矩阵的秩等于它的迹）。

但为了确定Equation 1.11的分布，我们还需证明分子分母随机变量的独立性。在证明其独立性之前，我们先介绍线性函数与二次型独立性的定理。

Theorem 1.1.2. 线性函数与二次型的独立性

一个标准正态向量x的线性函数$\mathrm{L}x$与对称幂等二次型$x'\mathrm{A}x$独立，如果$\mathrm{LA} = 0$。

证明.首先，因为A为幂等矩阵，故$x'\mathrm{A}x = x'\mathrm{A}'\mathrm{A}x = (\mathrm{A}x)'\mathrm{A}x$。由于$x \sim \mathrm{N}(0, \mathrm{I}_n)$，故$\mathrm{L}x$与$x'\mathrm{A}x$均服从正态分布。对于正态分布而言，若$Cov(\mathrm{L}x, (\mathrm{A}x)') = 0$，则说明二者独立。

$$\begin{aligned} Cov(\mathrm{L}x, (\mathrm{A}x)') &= \mathrm{E}[\mathrm{L}xx'\mathrm{A}'] - \mathrm{E}[\mathrm{L}x]\mathrm{E}[\mathrm{A}x] \\ &= \mathrm{L}\mathrm{I}_n\mathrm{A}' = \mathrm{LA} \end{aligned}$$

故，当$\mathrm{LA} = 0$时，$Cov(\mathrm{L}x, (\mathrm{A}x)') = 0$，有$\mathrm{L}x$与$x'\mathrm{A}x$独立（这个证明稍微有点奇怪）。

Equation 1.11的分子分母分别为

$$\begin{aligned} &\text{分子：} && \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2(\mathrm{X}'\mathrm{X})^{-1}_{jj}}} = \frac{(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'}{\sqrt{(\mathrm{X}'\mathrm{X})^{-1}_{jj}}} \cdot \frac{u}{\sigma} \\ &\text{分母：} && \frac{(n-k)\hat{\sigma}^2}{\sigma^2} = (\frac{u}{\sigma})'\mathrm{M}(\frac{u}{\sigma}) \end{aligned}$$

又因为

$$\frac{(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'}{\sqrt{(\mathrm{X}'\mathrm{X})^{-1}_{jj}}}\mathrm{M} = \frac{(\mathrm{X}'\mathrm{X})^{-1}(\mathrm{XM})'}{\sqrt{(\mathrm{X}'\mathrm{X})^{-1}_{jj}}} = 0$$

故，Equation 1.11的分子分母独立，有Equation 1.11中的

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 (\mathrm{X}'\mathrm{X})^{-1}_{jj}}} \sim t(n-k)$$

***F*检验：多变量联合**

已知方差：Wald test. 为检验所估计参数（$\hat{\beta}$）是否满足J个线性约束是否，本小节将介绍F检验。首先，假设

$$\begin{cases} \mathrm{H}_0 : \mathrm{R}\hat{\beta} = q \\ \mathrm{H}_1 : \mathrm{R}\hat{\beta} \neq q \end{cases} \Longrightarrow \begin{array}{l} \mathrm{H}_0 : d = \mathrm{R}\hat{\beta} - q = 0 \\ \mathrm{H}_1 : d = \mathrm{R}\hat{\beta} - q \neq 0 \end{array}$$

由于$\hat{\beta}$服从正态分布，d为$\hat{\beta}$的线性组合、服从于正态分布。故，对于d的分布，我们仅需要知道其均值与方差。

- **均值**：$\mathrm{E}(d) = \mathrm{R}\mathrm{E}(\hat{\beta}) - q \overset{\mathrm{H}_0}{=} 0$
- **方差**：$\mathrm{V}ar(d) = \mathrm{R}\mathrm{V}ar(\hat{\beta})\mathrm{R}' = \sigma^2 \mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}'$

故有，

$$d = \mathrm{R}\hat{\beta} - q \sim \mathrm{N}(0, \Sigma) \tag{1.12}$$

其中，$\Sigma^2 = \sigma^2 \mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}'$。故，在方差$\sigma^2$已知的情况下，可直接使用标准化$d$（服从于标准正态分布）来进行假设检验。如前所述，在通常的实证中，我们并不知道随机扰动项的真实方差σ^2，因而，我们将使用Wald检验：

$$w = d'\Sigma^{-1}d = (\mathrm{R}\hat{\beta} - q)'[\sigma^2 \mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}']^{-1}(\mathrm{R}\hat{\beta} - q)$$

其中，J为约束条件的个数。根据Theorem 1.1.3，可知$w \sim \chi^2(\mathrm{J})$。

Theorem 1.1.3. *If a $n \times 1$ vector $x \sim \mathrm{N}(\mu, \Sigma)$, then*

$$\begin{aligned} \Sigma^{-1/2}(x-\mu) &\sim \mathrm{N}(0, \mathrm{I}_n) \\ (x-\mu)'\Sigma^{-1}(x-\mu) &\sim \chi^2(n) \end{aligned}$$

未知方差：F test. 类似于对$\hat{\beta}$的检验，由于Σ中σ^2未知，因此，我们将使用$s^2 = \hat{\sigma}^2$来替代σ^2。

$$\begin{aligned} \mathrm{F} &= (\mathrm{R}\hat{\beta} - q)'[\hat{\sigma}^2 \mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}']^{-1}(\mathrm{R}\hat{\beta} - q)/\mathrm{J} \\ &= \frac{(\mathrm{R}\hat{\beta} - q)'[\sigma^2 \mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}']^{-1}(\mathrm{R}\hat{\beta} - q)/\mathrm{J}}{\frac{s^2(n-k)}{\sigma^2}/(n-k)} \sim \mathrm{F}(\mathrm{J}, n-k) \end{aligned} \tag{1.13}$$

$\mathrm{F}=\frac{\chi_1^2/n_1}{\chi_2^2/n_2}$要求分子分母中的随机变量独立(与$t$统计量一样)。而对于正态随机变量而言，随机变量不相关与独立等价，故，仅需证明Equation 1.13中分子分母的随机变量不相关、即可证明其服从于F分布。为了证明不相关，考虑分子

$$\frac{\mathrm{R}\hat{\beta}-q}{\sigma}\overset{\mathrm{H}_0}{=}\frac{\mathrm{R}(\hat{\beta}-\beta)}{\sigma}=\frac{\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'u}{\sigma}=\frac{\mathrm{A}u}{\sigma}$$

而分母，

$$\frac{s^2(n-k)}{\sigma^2}=\frac{\frac{\hat{u}'\hat{u}}{n-k}(n-k)}{\sigma^2}=\frac{\hat{u}'\hat{u}}{\sigma^2}=\frac{u'\mathrm{M}'\mathrm{M}u}{\sigma^2}=(\frac{\mathrm{M}u}{\sigma})'(\frac{\mathrm{M}u}{\sigma})$$

因此只需要考虑$\frac{\mathrm{A}u}{\sigma}$和$\frac{\mathrm{M}u}{\sigma}$的相关性，有

$$Cov(\frac{\mathrm{A}u}{\sigma},\frac{\mathrm{M}u}{\sigma})=\mathrm{E}(\frac{\mathrm{A}u}{\sigma})(\frac{\mathrm{M}u}{\sigma})'=\mathrm{E}(\frac{1}{\sigma^2}\mathrm{A}uu'\mathrm{M}')=\mathrm{E}(\mathrm{AM})=0$$

其中，$\mathrm{AM}=\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{M}=0$（$\mathrm{MX}=0$）。所以相关性为0，分子分母独立，Equation 1.13服从$\mathrm{F}(\mathrm{J},n-k)$。

上述假设检验的思想在于检验实际的估计是否满足所施加的约束，但另一方面，施加了约束的模型的拟合优度通常会比没有约束的模型更低，因而，从拟合有度损失的角度，我们同样可以推导出F检验的等价形式。

等价形式. F检验可等价表示为（$\mathrm{H}_0:\ \mathrm{R}\beta=q$），

$$\mathrm{F}=\frac{(u^{*\prime}u^*-u'u)/\mathrm{J}}{u'u/(n-k)}=\frac{(\mathrm{SSR}^*-\mathrm{SSR})/\mathrm{J}}{\mathrm{SSR}/(n-k)}$$

其中，带有（不带有）上标*表示参数由带有（不带有）约束的模型估计得到。

下面我们将推导上述F统计量的等价形式。首先，对于带有约束的模型而言，在最小化残差平方和的过程变为，

$$\begin{aligned}\min_{\beta}\quad & (y-\mathrm{X}\beta)'(y-\mathrm{X}\beta)\\ s.t.\quad & \mathrm{R}\beta-q=0\end{aligned}$$

有拉格朗日方程ℓ为，

$$\ell(\beta,\lambda)=(y-\mathrm{X}\beta)'(y-\mathrm{X}\beta)+2\underset{1\times\mathrm{J}}{\lambda'}(\mathrm{R}\beta-q)$$

故，一阶最优条件（FOC）为，

$$\frac{\partial\ell}{\partial\beta}=\begin{pmatrix}\frac{\partial\ell}{\partial\beta_1}\\ \frac{\partial\ell}{\partial\beta_2}\\ \vdots\\ \frac{\partial\ell}{\partial\beta_k}\end{pmatrix}\quad\text{or}\quad\frac{\partial\ell}{\partial\beta'}=(\frac{\partial\ell}{\partial\beta_1},\frac{\partial\ell}{\partial\beta_2},\dots,\frac{\partial\ell}{\partial\beta_k})$$

保证$\frac{\partial \ell}{\partial \beta}$的维度和微分前一致，即与$\beta$（$k\times 1$）或者$\beta'$（$1\times k$）保持一致。对于带有约束的模型而言，

$$\begin{cases} \dfrac{\partial \ell}{\partial \beta}|_{\beta=\hat{\beta}^*} = -2\mathrm{X}'(y-\mathrm{X}\hat{\beta}^*)+2\mathrm{R}'\hat{\lambda}^* = 0 & (a) \\ \dfrac{\partial \ell}{\partial \lambda}|_{\lambda=\hat{\lambda}^*} = 2(\mathrm{R}\hat{\beta}-q) = 0 & (b) \end{cases} \tag{1.14}$$

进一步地，有

$$\begin{aligned} (a)\ in\ \text{Equation 1.14} &\Longrightarrow \mathrm{R}'\hat{\lambda}^* = \mathrm{X}'y-\mathrm{X}'\mathrm{X}\hat{\beta}^* \\ \text{上式两边同时左乘}\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1} &\Longrightarrow \mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}'\hat{\lambda}^* \\ &= \mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y-\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}\hat{\beta}^* \\ (y=\mathrm{X}\hat{\beta}+e) &= \mathrm{R}\hat{\beta}+\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'u-\mathrm{R}\hat{\beta}^* \\ &= \mathrm{R}(\hat{\beta}-\hat{\beta}^*) = \mathrm{R}\hat{\beta}-q \end{aligned}$$

解得$\hat{\lambda}^*$如下

$$\begin{aligned} \hat{\lambda}^* &= [\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}']^{-1}(\mathrm{R}\hat{\beta}-q) \\ \mathrm{R}'\hat{\lambda}^* &= \mathrm{R}'[\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}']^{-1}(\mathrm{R}\hat{\beta}-q) \\ \mathrm{R}'[\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}']^{-1}(\mathrm{R}\hat{\beta}-q) &= \mathrm{X}'y-\mathrm{X}'\mathrm{X}\hat{\beta}^* \\ &= \mathrm{X}'(\mathrm{X}\hat{\beta}+u)-\mathrm{X}'\mathrm{X}\hat{\beta}^* \\ &= \mathrm{X}'\mathrm{X}(\hat{\beta}-\hat{\beta}^*) \\ \Longrightarrow \hat{\beta}-\hat{\beta}^* &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}'[\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}']^{-1}(\mathrm{R}\hat{\beta}-q) \end{aligned}$$

又$\mathrm{SSR}^* = u^{*\prime}u^*$和$\mathrm{SSR} = u'u$，根据$\hat{\beta}-\hat{\beta}^*$，可得

$$\begin{aligned} u^* &= y-\mathrm{X}\hat{\beta}^* = u+\mathrm{X}(\hat{\beta}-\hat{\beta}^*) \\ u^{*\prime}u^* &= u'u+(\hat{\beta}-\hat{\beta}^*)'\mathrm{X}'\mathrm{X}(\hat{\beta}-\hat{\beta}^*) \\ (\text{代入}\hat{\beta}-\hat{\beta}^*) &= u'u+(\mathrm{R}\hat{\beta}-q)'[\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}']^{-1}(\mathrm{R}\hat{\beta}-q) \end{aligned}$$

所以，

$$\begin{aligned} & \frac{(\mathrm{SSR}^*-\mathrm{SSR})/\mathrm{J}}{\mathrm{SSR}/(n-k)} \\ = & \frac{(\mathrm{R}\hat{\beta}-q)'[\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}']^{-1}(\mathrm{R}\hat{\beta}-q)/\mathrm{J}}{\hat{u}'\hat{u}/(n-\mathrm{K})} \\ = & \frac{(\mathrm{R}\hat{\beta}-q)'[\mathrm{R}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{R}']^{-1}(\mathrm{R}\hat{\beta}-q)/\mathrm{J}}{s^2} \end{aligned} \tag{1.15}$$

化简后Equation 1.13和Equation 1.15一致。

1.2 古典假设的放松

1.2.1 随机解释变量

在之前的所有讨论中，包括了$u \sim \mathrm{N}(0, \sigma^2)$、X非随机等假设，但由于数据的获取的方式往往具有随机性(比如调差问卷)，所以这些假设过于实际问题过于严格。因此，当数据是随机的时候，需要用到以下工具。

Theorem 1.2.1. *Law of Large Number (LLN)*

Let z_i be an i.i.d and $\mathrm{M} \times p$ matrix of observations, with $\mathrm{E}(z_i) = \mu$, assume $\mathrm{E}|z_i|^2$ is finite, then

$$\frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow{p} \mathrm{E}(z_i) = \mu$$

Theorem 1.2.2. *Central Limit Theory (CLT)*

Let z_i be an i.i.d and $\mathrm{M} \times 1$ vector of observations, with $\mathrm{E}(z_i) = \mu$ and $\mathrm{V}ar(z_i) = \Omega$, then

$$\mathrm{Z}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(z_i - \mu) \xrightarrow{d} \mathrm{N}(0, \Omega)$$

其中，p和d分别代表依概率与依分布收敛。[3][4]LLN说明样本的均值依概率收敛到总体的均值，而CLT说明样本均值的抽样分布依分布收敛到正态分布。

Definition 1.2.1. *Converge in Mean Square Error（or say, Converge in Mean Square, $\xrightarrow{\mathrm{L}^2}$)*

A sequence of random variable x_n converges to a constant θ in mean square error (MSE), that is $x_n \xrightarrow{\mathrm{MSE}} \theta$, if

$$\lim_{n\to\infty} \mathrm{E}(x_n - \theta)^2 = 0$$

[3]依概率收敛的定义：$\forall \varepsilon > 0$, $\lim_{n\to\infty} \mathrm{P}rob.(|x_n - x| > \varepsilon) = 0$，则称$x_n$依概率收敛于$x$，记作$x_n \xrightarrow{p} x$。

[4]依分布收敛的定义：若在$\mathrm{F}(x)$的所有连续点上，均有$\lim_{n\to\infty} |\mathrm{F}_n(x_n) - \mathrm{F}(x)| = 0$，则称$x_n$依分布收敛于$x$，记作$x_n \xrightarrow{d} x$。

对于MSE收敛而言，其同时要求随机变量的方差为零与随机变量无偏，

$$\begin{aligned} & \mathrm{E}(x_n-\theta)^2 \\ = \ & \mathrm{E}(x_n-\mathrm{E}(x_n)+\mathrm{E}(x_n)-\theta)^2 \\ = \ & \mathrm{E}[(x_n-\mathrm{E}x_n)^2+(\mathrm{E}x_n-\theta)^2+2(x_n-\mathrm{E}x_n)(\mathrm{E}x_n-\theta)] \\ = \ & \mathrm{E}(x_n-\mathrm{E}x_n)^2+(\mathrm{E}x_n-\theta)^2 \\ = \ & \mathrm{V}ar(x_n)+\mathrm{B}ias(x_n)^2 \end{aligned}$$

故，极大收敛之间的关系大致如下：$\overset{\text{MSE}}{\rightarrow} \Longrightarrow \overset{p}{\rightarrow}(\text{consistency}) \Longrightarrow \overset{d}{\rightarrow}$。

在引入了渐近分布相关的性质后，我们开始推导在u不一定服从正态分布时，$\hat{\beta}$的渐近分布（渐近正态）。首先，从Equation 1.1中，我们可以得到所估参数为

$$\begin{aligned} \hat{\beta}^{\text{OLS}}-\beta &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'u \\ \sqrt{n}(\hat{\beta}^{\text{OLS}}-\beta) &= (\frac{1}{n}\mathrm{X}'\mathrm{X})^{-1}\frac{1}{\sqrt{n}}\mathrm{X}'u \\ &= (\frac{1}{n}\sum_i^n x_i x_i')^{-1}\frac{1}{\sqrt{n}}\sum_i^n x_i u_i \\ \text{LLN}: &\quad (\frac{1}{n}\sum_i^n x_i x_i')^{-1} \overset{p}{\rightarrow} \mathrm{E}[x_i x_i']^{-1} \\ \text{CLT}: &\quad \frac{1}{\sqrt{n}}\sum_i^n x_i u_i \overset{d}{\sim} \mathrm{N}(0,\sigma^2\mathrm{E}[x_i x_i']) \\ \sqrt{n}(\hat{\beta}_{\text{OLS}}-\beta) &\overset{d}{\sim} [\mathrm{E}(x_i x_i')]^{-1}\mathrm{N}(0,\sigma^2[\mathrm{E}(x_i x_i')]) = \mathrm{N}(0,\sigma^2[\mathrm{E}(x_i x_i')]^{-1}) \\ 又\sigma^2[\mathrm{E}(x_i x_i')]^{-1} &= \sigma^2[\frac{1}{n}\sum_i^n x_i x_i']^{-1} \\ &= n\sigma^2(\mathrm{X}'\mathrm{X})^{-1} \\ \hat{\beta}^{\text{OLS}} &\overset{d}{\sim} \mathrm{N}(\beta,\sigma^2(\mathrm{E}[x_i' x_i])^{-1}) \end{aligned}$$

对于某一个系数$\hat{\beta}_j$而言，有

$$\frac{\hat{\beta}_j-\beta_j}{\sqrt{\sigma^2(\mathrm{X}'\mathrm{X})^{-1}_{jj}}} \overset{d}{\sim} \mathrm{N}(0,1)$$

分子分母趋于0的速度相同，所以商不为0，根据LLT和CLT，当$n\to\infty$，上式服从标准正态分布，而不再是小样本下的t分布。

1.2.2 球形扰动项

在经典假设下，随机扰动项的方差结构为

$$\mathrm{V}ar(u) = \mathrm{E}[u'u] = \sigma^2 \mathrm{I}_n = \begin{pmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{pmatrix}_{n\times n}$$

放宽假设，如果$\sigma_1^2 \neq \sigma_2^2 \neq \ldots \neq \sigma_n^2 \neq \sigma^2$且对角线以外的元素不为零，会导致$\mathrm{V}ar(\hat{\beta}^{\mathrm{OLS}}) = \hat{\sigma}^2(\mathrm{X}'\mathrm{X})^{-1}$不再成立，即$\hat{\beta}^{\mathrm{OLS}}$不再有效，但仍满足无偏性$\mathrm{E}[\hat{\beta}^{\mathrm{OLS}}] = \beta$。

(1) **Heteroskadasticity (HET)** 需要估计n个参数

$$\begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \sigma_i^2 & \vdots \\ 0 & \dots & \sigma_n^2 \end{pmatrix}_{n\times n}$$

(2) **Serial correlation (AUTO)** 由某一个参数数量较少的表达式估计协方差矩阵

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 & \dots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \dots & \sigma_{2n}^2 \\ \vdots & \sigma_{ij}^2 & \ddots & \vdots \\ \sigma_{n1}^2 & \dots & \dots & \sigma_{nn}^2 \end{pmatrix}_{n\times n}$$

在不满足经典假设的情况下$\hat{\beta}^{\mathrm{OLS}}$不是最优的估计量(not best),所以可以使用GLS（Generalized Least Square）。对$y = \mathrm{X}\beta + u$两边同时左乘一个矩阵R，得到$\mathrm{R}y = \mathrm{RX}\beta + \mathrm{R}u$，使得新的随机误差项$\mathrm{R}u$满足古典假设中的球形扰动项假设，但使用GLS的前提是估计$\mathrm{V}ar(u)$的结构。由于$\mathrm{V}ar(u)$一共有$n \times n$个参数，因此需要增加其他假设，减少参数的个数。

一个自然的想法是：协方差矩阵Σ是少量参数θ的函数，即，我们可以将方差σ_i^2考虑为与x_i的某些特征有关，通常我们乘这些特征为z_i（可能为x_i的函数）。

习惯上，我们设定$\sigma_i^2 = \sigma^2 \exp(z_i'\alpha)$，$\alpha$是$k \times 1$向量$k \ll n$，或者记为$\sigma_i^2 = \sigma_i^2(\theta), \theta(\sigma^2, \alpha)$。

假设$y_i \sim \mathrm{N}(x_i'\beta, \sigma_i^2(\theta, z_i))$，由极大似然估计MLE

$$\begin{aligned} \ln \mathscr{L}(\beta, \theta | x, y, z) &= \sum_i^n f(y_i | x_i, z_i, \beta_i, \theta) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^n \ln \sigma_i^2(\theta) - \frac{1}{2}\sum_{i=1}^n \frac{(y_i - x_i'\beta)^2}{\sigma_i^2(\theta)} \end{aligned}$$

如果$\sigma_i^2(\theta) = \sigma^2$，则MLE等价于OLS。现在我们有MLE等价于以$\frac{1}{\sigma_i^2(\theta)}$为权重的GLS。

组间异方差（Groupwise HET）． 实际中，我们通常会遇到的异方差问题即组间异方差，该类异方差假设不同组别之间随机扰动项的方差结构不同，但组内是同方差的。对于*i.i.d*的数据而言（组别出现的顺序可变）：

$$\begin{pmatrix} \sigma_{\mathrm{B}}^2 & & & \\ & \sigma_{\mathrm{S}}^2 & & \\ & & \sigma_{\mathrm{B}}^2 & \\ & & & \ddots \end{pmatrix} \text{ or } \begin{pmatrix} \sigma_{\mathrm{B}}^2 & & & \\ & \sigma_{\mathrm{B}}^2 & & \\ & & \sigma_{\mathrm{S}}^2 & \\ & & & \ddots \end{pmatrix}$$

若方差σ^2已知，则有β的估计为

$$\hat{\beta}^{\mathrm{GLS}} = [\sum_{g=1}^{\mathrm{G}} \frac{1}{\hat{\sigma}_g^2} \mathrm{X}_g' \mathrm{X}_g]^{-1} [\sum_{g=1}^{\mathrm{G}} (\frac{1}{\hat{\sigma}_g^2} \mathrm{X}_g' y_g)]$$

$$\mathrm{X} = \begin{pmatrix} \mathrm{X}_1 \\ \vdots \\ \mathrm{X}_{\mathrm{G}} \end{pmatrix}, \hat{\sigma}_g^2 = \frac{u_g' u_g}{n_g}$$

在$\sigma_i^2 = \sigma^2 \exp(z_i'\alpha)$中，Z是分组dummy Matrix（$x_i$属于哪一组，则Z的哪一列就等于1，否则等于0）

$$\mathrm{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ & \vdots & & \end{pmatrix}_{\mathrm{G} \times \mathrm{G}}$$

假设R是变换矩阵（Transformation Matrix）

$$\begin{aligned} \mathrm{R}y &= \mathrm{RX}\beta + \mathrm{R}u \\ \Longrightarrow y^* &= \mathrm{X}^*\beta + u^* \\ \Longrightarrow \hat{\beta}^{\mathrm{WLS}} &= (\mathrm{X}^{*\prime}\mathrm{X}^*)^{-1}\mathrm{X}^{*\prime}y^* \\ &= (\mathrm{X}'\mathrm{R}'\mathrm{RX})^{-1}\mathrm{X}'\mathrm{R}'\mathrm{R}y, \quad \mathrm{R}'\mathrm{R} = \frac{1}{\hat{\sigma}_g^2} \end{aligned}$$

异方差稳健标准误（Heteroskadasticity Robust Standard Error）． 对于经典的OLS估计而言，有

$$\begin{aligned} \hat{\beta}^{\mathrm{OLS}} &= \beta + (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'u \\ \mathrm{V}ar(\hat{\beta}^{\mathrm{OLS}}) &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{E}uu'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1} \\ &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\Sigma'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1} \end{aligned}$$

由于异方差问题，$\Sigma \neq \sigma^2 \mathrm{I}_n$，所以$\mathrm{V}ar(\hat{\beta}^{\mathrm{OLS}}) \neq \sigma^2(\mathrm{X}'\mathrm{X})^{-1}$。为得到$\mathrm{V}ar(\hat{\beta}^{\mathrm{OLS}})$，我们需要首先估计$\Sigma$的结构。但$\Sigma$仍旧有$n$个待估参数（对角线上的元素），所以直接估计$(\mathrm{X}'\Sigma\mathrm{X})_{k\times k}$Lfi一般有$k \ll n$。

$$\begin{aligned} \mathrm{X}'\Sigma\mathrm{X} &= \sum_i \sum_j \sigma_{ij} x_i x_j' \\ (\because \sigma_{ij} = 0) &= \sum_i \sigma_{ii} x_i x_i' \end{aligned}$$

又，

$$\mathrm{E}(u_i^2 x_i x_i') = \mathrm{E}[\mathrm{E}(u_i^2|x_i) x_i x_i'] = \mathrm{E}[\sigma_i^2 x_i x_i']$$

当$n \to \infty$，有

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i{}^2 x_i x_i' \to \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 x_i x_i'$$

故，$\hat{\beta}^{\mathrm{OLS}}$方差的估计为

$$\widehat{\mathrm{V}ar}(\hat{\beta}^{\mathrm{OLS}}) = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' \begin{pmatrix} \hat{u_1^2} & & \\ & \ddots & \\ & & \hat{u_n^2} \end{pmatrix} \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}$$

该文件标准误就是在Stata等软件中常用的，虽然看上去只是把Σ矩阵进行了替换,但由于$\hat{u}^2 \neq \hat{\sigma}^2$，所以本质上是替换了$\mathrm{X}'\Sigma\mathrm{X}$矩阵。

自相关. 关于自相关，考虑随机扰动项服从AR(1)，即

$$\begin{aligned} u_t &= \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1 \\ \Longrightarrow \mathrm{V}ar(u_t) &= \sigma_u^2 = \rho^2 \sigma_u^2 + \sigma_\varepsilon^2 \\ &= \frac{\sigma_\varepsilon^2}{1-\rho^2} \end{aligned}$$

其中，ε_t为白噪声（white noise）。

由于u是一个AR(1)，故有，

$$\mathrm{E}(u_t u_{t-1}) = \rho\sigma_u^2, \quad \mathrm{E}(u_t u_{t-2}) = \rho^2\sigma_u^2, \quad \dots$$

$$\mathrm{E}(uu') = \frac{\sigma_\varepsilon^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \dots \\ \rho & 1 & \rho & \vdots \\ \rho^2 & \rho & \ddots & \vdots \\ \vdots & \dots & \dots & 1 \end{pmatrix}$$

故，已知随机扰动项协方差矩阵Σ的结构，就可以找到变换矩阵用GLS估计，令变换矩阵

$$\mathrm{R} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \dots & 0 \\ -\rho & 1 & \ddots & \vdots \\ 0 & -\rho & 1 & \vdots \\ \ddots & \dots & -\rho & 1 \end{pmatrix}, \text{ 则} u^* = \mathrm{R}u = \begin{pmatrix} \sqrt{1-\rho^2}u_1 \\ u_2 - \rho u_1 \\ u_3 - \rho u_2 \\ \vdots \\ u_n - \rho u_{n-1} \end{pmatrix}$$

其中，估计ρ时，首先得到$\hat{u}$序列，然后用$\hat{u}$来得到$\hat{\rho}$。

因此方程$y = \mathrm{X}\beta + u$转化为$\mathrm{R}y = \mathrm{R}x + \mathrm{R}u$, $\mathrm{V}ar(u)$是一个对角线以外元素不为0的方阵，而$\mathrm{V}ar(\mathrm{R}u) = \mathrm{R}\mathrm{V}ar(u)\mathrm{R}'$是一个对角阵，且对角线元素相等（满足球形扰动项假设）。

对于自相关问题，回到$\mathrm{V}ar(\hat{\beta})$的结构上，我们知道$\mathrm{V}ar(\hat{\beta}) = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\Sigma'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}$，其中

$$\Sigma = \hat{u}\hat{u}' = \begin{pmatrix} \hat{u_1}^2 & \hat{u_1}\hat{u_2} & \dots & \hat{u_1}\hat{u_n} \\ \hat{u_2}\hat{u_1} & \hat{u_2}^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \dots & \dots & \dots & \hat{u_n}^2 \end{pmatrix}$$

在异方差问题中，我们直接将估计出的$diag = (\hat{u_1^2}, \hat{u_2^2}, \dots, \hat{u_n^2})$代入为$\hat{\Sigma}$，即可获得修正后的标准方差估计，但对于自相关问题而言，由于$(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\hat{u}\hat{u}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1} = 0$（因为$\mathrm{X}'\hat{u} = 0$）,所以不能简单地认为$\hat{\Sigma} = \hat{u}\hat{u}'$。

Newey-West Adjusted SE. Newey和West（1987）通过为协方差赋权的方式来对$\hat{u}\hat{u}'$进行调整，即为越靠近对角线的元素赋予更高的权重。

$$\begin{aligned} \text{Unadjusted: } \mathrm{X}'\hat{\Sigma}\mathrm{X} &= \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \hat{u_t^2} x_t x_t' + \frac{1}{\mathrm{T}}\sum_{l=1}^{\mathrm{L}}\sum_{t=l+1}^{\mathrm{T}} \hat{u_t^2}\hat{u_{t-l}^2}(x_t x_{t-l}' + x_{t-l}x_t') \\ \text{NW-adjusted: } \mathrm{X}'\hat{\Sigma}'\mathrm{X} &= \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \hat{u_t^2} x_t x_t' + \frac{1}{\mathrm{T}}\sum_{l=1}^{\mathrm{L}}\sum_{t=l+1}^{\mathrm{T}} w_l \hat{u_t^2}\hat{u_{t-l}^2}(x_t x_{t-l}' + x_{t-l}x_t') \end{aligned}$$

权重为

$$w_l = 1 - \frac{l}{\mathrm{L}+1}$$

其中，T为样本在时序上的维度，L为研究中所考虑的自相关阶数，通常取$T^{\frac{1}{4}}$。当L = 0时，Newey-West调整的标准误实际为 Huber–White标准误（HET consistent）。

1.2.3 内生变量

在之前的假设中，大样本放松了随机扰动项服从正态分布的假设，而关于异方差的讨论放松了球形扰动项的假设，本小节将进一步讨论如何放松u和x不相关的假设。首先，u和X不相关意味着：

$$\mathrm{E}[u|\mathrm{X}] = 0$$

当X和u相关时，此时$\hat{\beta} - \beta = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'u$是有偏的：

$$\begin{aligned}
\hat{\beta} - \beta &= (\frac{\mathrm{X}'\mathrm{X}}{n})^{-1}(\frac{\mathrm{X}'u}{n}) \\
&= (\frac{1}{n}\sum x_i x_i')^{-1}(\frac{1}{n}x_i u_i) \\
\mathrm{E}[\hat{\beta} - \beta] &= (\mathrm{E}[x_i x_i'])^{-1}(\mathrm{E}x_i u_i) \neq 0 \ when\ n \to \infty \\
&= 0 \ if\ \mathrm{E}([x_i u_i]) = \mathrm{E}([x_i \underbrace{\mathrm{E}[u_i|x_i]}_{=0}]) = 0
\end{aligned}$$

因此，当$\mathrm{E}[u|\mathrm{X}] \neq 0$时，但$\mathrm{E}[u_i|\mathrm{X}_i] = 0$，$\hat{\beta}$虽然不是无偏的但是是一致的，即$\hat{\beta} - \beta \xrightarrow{p} 0$。

Unbiasness vs Consistency

- 考虑$\mathrm{E}(y_t) = \mu$，估计量$\hat{\mu_1} = \frac{1}{n+1}\sum_{t=1}^{n} y_t$，$\mathrm{E}\hat{\mu_1} = \frac{n}{n+1}\mu \neq 0$，但是$p\lim_{n\to\infty}\frac{n}{n+1}\frac{1}{n}\sum_{t=1}^{n} y_t = \mu$。
  - 有偏但是一致。
- 考虑$\hat{\mu_2} = 0.01y_1 + \frac{0.99}{n-1}\sum_{t=2}^{n} y_t$，$\mathrm{E}\hat{\mu_2} = \mu$，但是$p\lim_{n\to\infty}\hat{\mu_2} = 0.01y_1 + 0.99\mu \neq \mu$。
  - 是无偏但是不一致。

一致性要求 $\forall i, (\mathrm{E}[x_i u_i]) \to 0\ when\ n \to \infty$，但是当存在内生变量时，$\mathrm{E}[xu]$就不是零向量，因而会影响所有$\beta_j$的估计。

1.3 内生性问题

内生性问题出现的原因主要有三种：测量误差，互为因果，遗漏变量。接下来内容将逐个介绍三种问题。

1.3.1 测量误差（Measurement Error）

首先，在实际的数据收集过程中，由于各种原因，所收集的数据将会收到一系列的干扰，从而使得所收集的数据包含了某些噪声信息。对于真实模型：

$$y_t^0 = \beta_1 + \beta_2 x_t^0 + u_t^0 \tag{1.16}$$

其中，由于测量误差，实际数据（x_t与y_t）为

$$x_t = x_t^0 + v_{1t} \qquad y_t = y_t^0 + v_{2t}$$

故，实际模型为

$$\begin{aligned} y_t - v_{2t} &= \beta_1 + \beta_2(x_t - v_{1t}) + u_t^0 \\ \Longrightarrow y_t &= \beta_1 + \beta_2 x_t + \underbrace{[u_t^0 + v_{2t} - \beta_2 v_{1t}]}_{=u_t} \end{aligned}$$

此时，$\mathrm{V}ar(u_t) \geq \sigma^2(\mathrm{X}'\mathrm{X})^{-1}$，使得$\hat{\mathrm{V}ar}(\hat{\beta}^{\mathrm{OLS}})$变大，使得对$\hat{\beta}$的统计推断并不准确，即估计并不有效、存在方差更小的估计方法。由于测量误差，有x_t与u_t相关：

$$\mathrm{C}ov(x_t, u_t) = \mathrm{E}(x_t u_t) = \mathrm{E}[(x_t^0 + v_{1t})(u_t^0 + v_{2t} - \beta_2 v_{1t})] = -\beta \mathrm{V}ar(v_{1t})$$

举例而言，考虑收入和消费的关系：

$$\Uparrow\downarrow y_t = \beta_1 + \beta_2 x_t \Uparrow + u_t \downarrow \tag{1.17}$$

x_t增加会导致y_t增加$\Uparrow$，但由于内生性，所以x_t和u_t的负相关性导致u_t对y_t产生反方向的影响，使得低估了β_2（但不会使得符号相反）。

1.3.2 互为因果（Simultaneous Causality）

互为因果为最常见的内生性问题，一般直接的方法是使用工具变量来解决互为因果的问题，本文先介绍互为因果的问题，然后在介绍基于工具变量的回归方法。考虑经典的供求均衡问题，对于均衡时的成交数量而言，有

$$\begin{cases} q_t = \gamma_d p_t + x_t^d \beta_d + u_t^d \\ q_t = \gamma_s p_t + x_t^s \beta_s + u_t^s \end{cases} \tag{1.18}$$

但同时，对于均衡时的成交价格p_t, q_t与u_t^d, u_t^s而言，有

$$\begin{pmatrix} q_t \\ p_t \end{pmatrix} = \begin{pmatrix} 1 & -\gamma_d \\ 1 & -\gamma_s \end{pmatrix}^{-1} + \left[\begin{pmatrix} x_t^d & \beta_d \\ x_t^s & \beta_s \end{pmatrix} \begin{pmatrix} u_t^d \\ u_t^s \end{pmatrix} \right] \tag{1.19}$$

Equation 1.18和Equation 1.19表明价格与数量互相联系，并不能够直接形成所谓的因果推断，因此以下内容将引入工具变量来尝试解决内生性问题。

1.4 工具变量

IV. 工具变量是指与内生变量相关，但与被解释变量无关的变量（满足相关性与无关性两个条件），后续内容会介绍如何检验工具变量的相关性与无关性假设。因而，一个直观的解决内生性变量的方法为，将内生变量中与被解释变量无关的部分提取出来，使得该部分与残差项无关：

$$\begin{aligned} y &= \mathrm{X}_1\gamma_1 + \mathrm{X}_2\gamma_2 + u \\ &\Rightarrow \mathrm{X}_1\gamma_1 + \mathrm{X}_{\mathrm{IV}}\gamma_2 + u \end{aligned}$$

其中X_1是外生变量，X_2是内生变量，而$\rho(\mathrm{X}_{\mathrm{IV}}, u) = 0$。

2SLS. 实现提取内生变量中与被解释变量无关部分，并将其作为新的解释变量的方法叫做两阶段最小二乘（2-stage Least Square）。

$$\begin{aligned} \text{First Stage: } \hat{\mathrm{X}}_2 &= (\mathrm{X}_1, \mathrm{X}_{\mathrm{IV}}) \begin{pmatrix} \hat{\Pi}_1 \\ \hat{\Pi}_2 \end{pmatrix} = \mathrm{W} \begin{pmatrix} \hat{\Pi}_1 \\ \hat{\Pi}_2 \end{pmatrix} \\ \text{Second Stage: } y &= \mathrm{X}_1\gamma_1 + \hat{\mathrm{X}}_2\gamma_2 + u \end{aligned}$$

其中要求$\hat{\Pi}_2$显著（sigificant），通常来讲，$\mathrm{F} \leq 10$代表此时的工具变量为弱工具变量（weak IV）。又因为有

$$\begin{aligned} \hat{\mathrm{X}}_2 &= \mathrm{P}_{\mathrm{W}}\mathrm{X}_2 \\ \mathrm{X}_1 &= \hat{\mathrm{X}}_1 = \mathrm{P}_w\mathrm{X}_1 \\ \Longrightarrow \hat{\mathrm{X}} &= (\hat{\mathrm{X}}_1, \hat{\mathrm{X}}_2) = \mathrm{P}_w\mathrm{X} \end{aligned}$$

所以，第二阶段的估计为：

$$\begin{aligned} y &= \mathrm{X}_1\gamma_1 + \hat{\mathrm{X}}_2\gamma_2 + u \\ &= (\hat{\mathrm{X}}_1, \hat{\mathrm{X}}_2)\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} + u \\ &= \hat{\mathrm{X}}\delta + u \end{aligned}$$

故，最小二乘估计量为

$$\begin{aligned} \hat{\delta}^{\mathrm{2SLS}} &= (\hat{\mathrm{X}}'\hat{\mathrm{X}})^{-1}\hat{\mathrm{X}}'y \\ &= (\mathrm{X}'\mathrm{P}_{\mathrm{W}}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\mathrm{X}'\mathrm{P}_{\mathrm{W}}'y \\ &= (\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\mathrm{X}'\mathrm{P}_{\mathrm{W}}y \end{aligned}$$

工具变量法的检验.

1. 判断是否存在内生性，是否需要工具变量

2. 工具变量与内生变量是否有足够强的关系，F-test，F>10 in first stage

3.X_{IV}不能直接影响y，即不能直接影响u（若X_{IV}影响y，且未出现在$(\mathrm{X}_1, \mathrm{X}_2)$里，则必定与$u$相关）。**Overidentifying Restrictions Test (Sargan Test).** 用来判断工具变量是否是外生的。由于检验存在局限性，所以当直觉和检验出现矛盾的时候一般还是更依赖于直觉。对于工具变量的外生性而言，检验假设为

$$\begin{aligned} &\mathrm{H}_0 : \mathrm{E}[\mathrm{W}'u] = 0 \quad \mathrm{W} = (\mathrm{X}_1, \mathrm{X}_{\mathrm{IV}}) \\ &\mathrm{H}_1 : \mathrm{E}[\mathrm{W}'u] \neq 0 \end{aligned}$$

考虑回归模型

$$\begin{aligned} y &= \mathrm{X}_1\beta_1 + \mathrm{X}_2\beta_2 + u \\ \mathrm{W} &= (\mathrm{X}_1, \mathrm{X}_{\mathrm{IV}}) \end{aligned}$$

因此，我们仅需要估计出$\hat{u}$，然后计算$\mathrm{E}[\mathrm{W}'\hat{u}]$，即可检验原假设，利用2SLS估计：

$$y = \mathrm{X}_1\hat{\beta}_1^{\mathrm{2SLS}} + \mathrm{X}_2\hat{\beta}_2^{\mathrm{2SLS}} + \hat{u}$$

将$\hat{u}$回归到W上，即$\hat{u} = \mathrm{W}\hat{b} + e$，可以计算出$\mathrm{R}^2$，当$\mathrm{R}^2$越高时，说明W和$\hat{u}$之间的相关性越强。但是由

于R^2的分布不能直接查表判断，所以要构造包含R^2的标准分布，考虑LR检验，计算nR^2：

$$\text{LR test: } nR^2 = n\frac{SSE}{SST} = \frac{\hat{b}'W'W\hat{b}}{\frac{1}{n}\hat{u}'\hat{u}} \tag{1.20}$$

根据之前的笔记$\hat{\sigma^2} = \frac{\hat{u}'\hat{u}}{n-k}$, $y = X\beta + u$, $My = Mu = \hat{u}$, $Py = \hat{y}$, $y = \hat{y} + \hat{u}$，所以$W\hat{b} = P_W\hat{u}$，Equation 1.20可以改写为：

$$\frac{\hat{b}'W'W\hat{b}}{\frac{1}{n}\hat{u}'\hat{u}} = \frac{\hat{u}'P_W'P_W\hat{u}}{\frac{1}{n}\hat{u}'\hat{u}} \sim \chi^2(q)$$

接下来将计算χ^2分布的自由度q，首先，

$$\begin{aligned}\hat{u} &= y - X\hat{\beta}_{2SLS} \\ &= [I - X(X'P_WX)^{-1}X'P_W](X\beta + u) \\ &= [I - X(X'P_WX)^{-1}X'P_W]u\end{aligned}$$

其中，$W_{n\times l} = (X_1, X_{IV})$，$X_{n\times k} = (X_1, X_2)$，$\hat{\beta}^{2SLS} = (X'P_WX)^{-1}X'P_Wy$，故有

$$\begin{aligned}\hat{u}'P_W\hat{u} &= u'[I - P_WX(X'P_WX)^{-1}X']P_W[I - X(X'P_WX)^{-1}X'P_W]u \\ &= u'[P_W - P_WX(X'P_WX)^{-1}P_W]u \sim \chi^2(q)\end{aligned}$$

而q取决于矩阵$[P_W - P_WX(X'P_WX)^{-1}P_W]$的迹（trace），

$$\begin{aligned}tr(P_W - P_WX(X'P_WX)^{-1}P_W) &= tr(P_W) - tr(P_WX(X'P_WX)^{-1}X'P_W) \\ &= tr(W(W'W)^{-1}W') - tr(P_WX(X'P_WX)^{-1}X'P_W) \\ &= tr(W'W)^{-1}W'W) - tr((X'P_WX)^{-1}X'P_WP_WX) \\ &= tr(I_l) - tr(I_k) = l - k\end{aligned}$$

$$\hat{u}'P_W\hat{u} \sim \chi^2(l-k) \tag{1.21}$$

其中l是W含有变量的个数，k是X含有变量的个数。χ^2分布存在要求自由度$l-k$大于0，也即工具变量的个数大于内生变量的个数，所以又叫Overidentifying Restriction Test，必须要在过度识别的情况下才能检验（如果恰好识别则χ^2的自由度等于0，则不能检验）。

Hausman Test. Hausman Test旨在比较两个估计量的准度与效率的问题，尝试找到两个不同的估计量，一个估计量在H_0和H_1下都是一致的，另一个在H_1下不一致，但在H_0下是一致且有效的。

考虑假设：

H_0：OLS is consistent, 2SLS is consistent ,but *not* efficient

H_1：OLS is *not* consistent, 2SLS is consistent

因此想法将$\hat{\beta}^{\text{OLS}} - \hat{\beta}^{\text{2SLS}}$变成一个标准的分布，记$\hat{\beta}^{\text{IV}} = \hat{\beta}^{\text{2SLS}}$ 已知

$$\hat{\beta}^{\text{OLS}} = (\text{X}'\text{X})^{-1}\text{X}'y$$
$$\hat{\beta}^{\text{IV}} = (\text{X}'\text{P}_\text{W}\text{X})^{-1}\text{X}'\text{P}_\text{W}y$$

$$\begin{aligned}
\hat{\beta}^{\text{IV}} - \hat{\beta}^{\text{OLS}} &= (\text{X}'\text{P}_\text{W}\text{X})^{-1}\text{X}'\text{P}_\text{W}y - (\text{X}'\text{X})^{-1}\text{X}'y \\
&= (\text{X}'\text{P}_\text{W}\text{X})^{-1}\text{X}'\text{P}_\text{W}y - (\text{X}'\text{P}_\text{W}\text{X})^{-1}\text{X}'\text{P}_\text{W}\text{X}(\text{X}'\text{X})^{-1}\text{X}'y \\
&= (\text{X}'\text{P}_\text{W}\text{X})^{-1}\text{X}'\text{P}_\text{W}[\text{I}_n - \text{X}(\text{X}'\text{X})^{-1}\text{X}']y \\
&= (\text{X}'\text{P}_\text{W}\text{X})^{-1}\text{X}'\text{P}_\text{W}\text{M}_\text{X}y \\
&= (\text{X}'\text{P}_\text{W}\text{X})^{-1}\text{X}'\text{P}_\text{W}\text{M}_\text{X}u
\end{aligned} \tag{1.22}$$

将$\hat{\beta}^{\text{IV}} - \hat{\beta}^{\text{OLS}}$转换成含$u$的表达式后，可根据$u \sim \text{N}$，来找$\chi^2$的自由度。

$$\begin{aligned}
\text{test} &= (\hat{\beta}^{\text{IV}} - \hat{\beta}^{\text{OLS}})'[\text{V}ar(\hat{\beta}^{\text{IV}} - \hat{\beta}^{\text{OLS}})]^{-1}(\hat{\beta}^{\text{IV}} - \hat{\beta}^{\text{OLS}}) \\
&= \chi^2(?)
\end{aligned}$$

由于$\text{V}ar(\hat{\beta}^{\text{IV}} - \hat{\beta}^{\text{OLS}})$不可逆（见下），所以$\chi^2$分布的自由度并不是$k$，

$$\begin{aligned}
\hat{\beta}^{\text{IV}} - \hat{\beta}^{\text{OLS}} &= (\text{X}'\text{P}_\text{W}\text{X})^{-1}\begin{pmatrix}\text{X}_1' \\ \text{X}_2'\end{pmatrix}\text{P}_\text{W}\text{M}_\text{X}u \\
&= (\text{X}'\text{P}_\text{W}\text{X})^{-1}\begin{pmatrix}\text{X}_1'\text{P}_\text{W}\text{M}_\text{X}u \\ \text{X}_2'\text{P}_\text{W}\text{M}_\text{X}u\end{pmatrix} \qquad (\text{X}_1'\text{P}_\text{W} = \text{X}_1, \text{X}_1\text{M}_\text{X} = 0) \\
&= (\text{X}'\text{P}_\text{W}\text{X})^{-1}\begin{pmatrix}0 \\ \text{X}_2'\text{P}_\text{W}\text{M}_\text{X}u\end{pmatrix}
\end{aligned} \tag{1.23}$$

$$
\begin{aligned}
\mathrm{V}ar(\hat{\beta}^{\mathrm{IV}}-\hat{\beta}^{\mathrm{OLS}}) &= \mathrm{E}\left[(\hat{\beta}^{\mathrm{IV}}-\hat{\beta}^{\mathrm{OLS}})(\hat{\beta}^{\mathrm{IV}}-\hat{\beta}^{\mathrm{OLS}})'\right] \\
&= \mathrm{E}\left[(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\begin{pmatrix}0\\ \mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}u\end{pmatrix}(0,\ u'\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2)(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\right] \\
&= \mathrm{E}\left[(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\begin{pmatrix}0 & 0\\ 0 & \mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}uu'\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2\end{pmatrix}(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\right]
\end{aligned}
\tag{1.24}
$$

如果$\mathrm{E}(uu')=\sigma^2\mathrm{I}$，不失一般性，将中间矩阵的左上角部分记为A，得到

$$
\mathrm{V}ar(\hat{\beta}_{\mathrm{IV}}-\hat{\beta}_{\mathrm{OLS}}) = \mathrm{E}\left[(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\begin{pmatrix}\mathrm{A} & 0\\ 0 & \mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}\sigma^2\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2\end{pmatrix}(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\right] \tag{1.25}
$$

$$
\mathrm{V}ar(\hat{\beta}_{\mathrm{IV}}-\hat{\beta}_{\mathrm{OLS}})^{-1} = \mathrm{E}\left[(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})\begin{pmatrix}\mathrm{A} & 0\\ 0 & \mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}\sigma^2\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2\end{pmatrix}^{-1}(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})\right] \tag{1.26}
$$

故，test统计量为：

$$
\begin{aligned}
\mathrm{test} &= (\hat{\beta}^{\mathrm{IV}}-\hat{\beta}^{\mathrm{OLS}})'[\mathrm{V}ar(\hat{\beta}^{\mathrm{IV}}-\hat{\beta}^{\mathrm{OLS}})]^{-1}(\hat{\beta}^{\mathrm{IV}}-\hat{\beta}^{\mathrm{OLS}}) \\
&= (0'\,u'\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2)(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\mathrm{E}\left[(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})\begin{pmatrix}\mathrm{A}^{-1} & 0\\ 0 & (\mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}\sigma^2\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2)^{-1}\end{pmatrix}(\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})\right] \\
&\quad (\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\begin{pmatrix}0\\ \mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}u\end{pmatrix} \\
&= (0'\,u'\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2)\begin{pmatrix}\mathrm{A}^{-1} & 0\\ 0 & (\mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}\sigma^2\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2)^{-1}\end{pmatrix}\begin{pmatrix}0\\ \mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}u\end{pmatrix} \\
&= u'\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2(\mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}\sigma^2\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2)^{-1}\mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}u \\
&= \frac{1}{\sigma^2}u'\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2(\mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2)^{-1}\mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}u
\end{aligned}
$$

所以A的取值对最后test统计量的表达形式没有本质上的影响。

$$
tr(\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2(\mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2)^{-1}\mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}) = tr(\mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2)^{-1}\mathrm{X}_2'\mathrm{P}_{\mathrm{W}}\mathrm{M}_{\mathrm{X}}\mathrm{M}_{\mathrm{X}}\mathrm{P}_{\mathrm{W}}\mathrm{X}_2) = k_2 \tag{1.27}
$$

因此，在Hausman Test中test$\sim\chi^2(k_2)$，其中k_2为Equation 1.27中的取值。

第二章　广义矩估计

2.1　广义矩估计

矩估计（Method of Moment，MM）或者广义矩估计（Generalized Method of Moment，GMM）是一种类似LS，2SLS，MLE的估计方法。但不同的是，广义矩估计的核心思想是找到总体矩条件，再找样本矩条件(大数定理满足的情况下，样本矩条件等价于总体矩条件)，根据样本矩条件寻找使得矩条件违背程度最小的参数作为广义矩估计的估计参数，类似于k_1个方程解k_2个未知数，在n个观测中寻找使得方程在最大程度上满足的参数。矩估计通常有以下三步：

- **总体矩条件**：Poplution moment condition/restriction (POPMC)

$$\text{E(functions of r.v. and parameter)} = 0$$

- **样本矩条件**：Sample moment condition/restriction (SMC)

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)=0 \Longrightarrow \bar{x}-\mu=0 \Rightarrow \hat{\mu}=\bar{x}$$

- **解矩条件**：估计矩条件中的参数（通常有几个参数就需要几个方程）。

2.1.1　常见方法的广义矩估计等价形式

以OLS、2SLS和MLE举例而言，

Example 1: OLS. 对于经典模型$y=\mathrm{X}\beta+u$而言，有

$$\begin{aligned}
\text{总体矩条件：}&\quad \mathrm{E}\mathrm{X}'u=0\\
\text{样本矩条件：}&\quad \mathrm{E}x_iu_i=0\\
\text{求解方程：}&\quad \mathrm{X}'(y-\mathrm{X}\beta)=0\\
\Longrightarrow&\quad \hat{\beta}^{\mathrm{MM}}=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y=\hat{\beta}^{\mathrm{OLS}}
\end{aligned}$$

Example 2: 2SLS. 如果$\mathrm{E}(\mathrm{X}u) \neq 0$（出现内生性问题），则模型为

$$y = \mathrm{X}_1\beta + \mathrm{X}_2\beta + u$$

此时有$\mathrm{EX}'u \neq 0$，但于工具变量而言，$\mathrm{EW}'u = 0$，所以有

$$\begin{aligned}
\text{总体矩条件：} \quad & \mathrm{W}'u = 0 \\
\text{样本矩条件：} \quad & \mathrm{E}w_i u_i = 0 \\
\text{求解方程：} \quad & \mathrm{W}'(y - \mathrm{X}\beta) = 0 \\
\Longrightarrow \quad & \hat{\beta}^{\mathrm{MM}} = (\mathrm{W}'\mathrm{X})^{-1}\mathrm{W}'y = \hat{\beta}^{\mathrm{OLS}}
\end{aligned}$$

对于$\hat{\beta}_{\mathrm{2SLS}}$而言，

$$\begin{aligned}
\hat{\beta}^{\mathrm{2SLS}} &= (\mathrm{X}'\mathrm{P}_{\mathrm{W}}\mathrm{X})^{-1}\mathrm{X}'\mathrm{P}_{\mathrm{W}}y \\
&= (\mathrm{X}'\mathrm{W}(\mathrm{W}'\mathrm{W})^{-1}\mathrm{W}'\mathrm{X})^{-1}\mathrm{X}\mathrm{W}(\mathrm{W}'\mathrm{W})^{-1}\mathrm{W}'y
\end{aligned}$$

由于$\mathrm{X}'\mathrm{W}, \mathrm{W}\mathrm{W}, \mathrm{W}'\mathrm{X}$都是方阵，所以

$$\begin{aligned}
(\mathrm{X}'\mathrm{W}(\mathrm{W}'\mathrm{W})^{-1}\mathrm{W}'\mathrm{X})^{-1} &= (\mathrm{W}'\mathrm{X})^{-1}(\mathrm{W}'\mathrm{W})(\mathrm{X}'\mathrm{W})^{-1} \\
\Rightarrow \hat{\beta}^{\mathrm{2SLS}} &= (\mathrm{W}'\mathrm{X})^{-1}(\mathrm{W}'\mathrm{W})(\mathrm{X}'\mathrm{W})^{-1}\mathrm{X}\mathrm{W}'(\mathrm{W}'\mathrm{W})^{-1}\mathrm{W}'y \\
&= (\mathrm{W}'\mathrm{X})^{-1}\mathrm{W}'y
\end{aligned}$$

故，综上可得

$$\begin{aligned}
\mathrm{X}'u = 0: \quad & \hat{\beta}_{\mathrm{MM}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y = \hat{\beta}_{\mathrm{OLS}} \\
\mathrm{W}'u = 0: \quad & \hat{\beta}_{\mathrm{MM}} = (\mathrm{W}'\mathrm{X})^{-1}\mathrm{W}'y = \hat{\beta}_{\mathrm{2SLS}}
\end{aligned}$$

Example 3: MLE. 对极大似然估计而言，首先写出其似然函数。

$$\theta_0 = \underset{\theta \in \text{Ⓗ}}{\arg\max}\, \mathrm{E}\ln f(x, y|\theta)$$

$$\begin{aligned}
\text{总体矩条件：} \quad & \mathrm{E}\left(\frac{\partial \ln f(x,y|\theta_0)}{\partial \theta}\right) = 0 \\
\text{样本矩条件：} \quad & \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ln f(x_i,y_i|\hat{\theta}^{\mathrm{MLE}})}{\partial \theta} = 0 \\
\text{求解方程：} \quad & \theta_0^{\mathrm{MLE}}
\end{aligned}$$

2.1.2 广义矩估计

具体估计

本节将具体介绍GMM方法下估计的推导。

首先，对于一组总体矩条件（L个）$g(z,\theta)$

$$g(z,\theta)=\begin{pmatrix} g_1(z,\theta) \\ g_2(z,\theta) \\ \vdots \\ g_{\mathrm{L}}(z,\theta) \end{pmatrix}_{\mathrm{L}\times 1}$$

其样本矩条件的表达形式为：

$$\hat{g}_n(\theta)=\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} g_1(z_i,\theta) \\ \frac{1}{n}\sum_{i=1}^{n} g_2(z_i,\theta) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} g_{\mathrm{L}}(z_i,\theta) \end{pmatrix}_{\mathrm{L}\times 1} = \frac{1}{n}\sum_{i=1}^{n} g(z_i,\theta)$$

其中，k为未知参数的个数（# unknown parameters），L为矩条件的个数（# unknown independent restrictions)，如果$\mathrm{L}>k$（$\mathrm{L}=k$），则我们称该估计为GMM（MM）。即$\mathrm{X}'_{k\times n}u_{n\times 1}=0_{k\times 1}$。对于$\mathrm{L}>k$，实际并不能找到满足所有矩条件都为0的参数，即$\hat{g}_n(\theta)\neq 0$。

定义GMM的目标函数为

$$\mathrm{Q}_n^{\mathrm{W}}(\theta)=\underset{1\times \mathrm{L}}{\hat{g}_n(\theta)'}\quad \underset{\mathrm{L}\times \mathrm{L}}{\mathrm{W}}\quad \underset{\mathrm{L}\times 1}{\hat{g}_n(\theta)}$$

其中W是权重矩阵，其表示在不同矩条件的权重，任意的正定矩阵都能作为权重矩阵（最简单的有$\mathrm{W}=\mathrm{I}_n$。

$$\hat{\theta}^{\mathrm{GMM}}=\arg\min \mathrm{Q}_n^{\mathrm{W}}(\theta)\xrightarrow{p}\theta_0(as\ n\to\infty).$$

故有总体矩条件为$\mathrm{E}(r.v.,para)=\hat{\theta}^{\mathrm{GMM}}-\theta_0=0$，证明其成立则需要找到$\hat{\theta}^{\mathrm{GMM}}$的分布（如$\hat{\beta}_{\mathrm{OLS}}\sim \mathrm{N}(\beta,\sigma^2(\mathrm{X}'\mathrm{X})^{-1})$）。

Theorem 2.1.1. ***Asymptotic normality of GMM Estimator***

Under appropriate conditions, we have

$$\sqrt{n}(\hat{\theta}^{\mathrm{GMM}}-\theta_0)\xrightarrow{d}\mathrm{N}(0,(\mathrm{G}'\mathrm{W}\mathrm{G})^{-1}\mathrm{G}'\mathrm{W}\Omega_0\mathrm{W}\mathrm{G}(\mathrm{G}'\mathrm{W}\mathrm{G})^{-1})$$

其中$\mathrm{G}(\theta)=\mathrm{E}(\nabla_\theta g(z,\theta)),\mathrm{G}=\mathrm{G}(\theta_0)_{\mathrm{L}\times k},\Omega_0=\mathrm{E}[\underset{\mathrm{L}\times 1}{g(z,\theta_0)}\underset{1\times\mathrm{L}}{g(z,\theta_0)'}]$ *and*,

$$\underset{\mathrm{L}\times 1}{\nabla_\theta g(z,\theta)}=\frac{\partial g(z,\theta)}{\partial\theta'}=\begin{pmatrix}\frac{\partial g_1}{\partial\theta_1}&\cdots&\frac{\partial g_1}{\partial\theta_k}\\ \vdots&\ddots&\vdots\\ \frac{\partial g_{\mathrm{L}}}{\partial\theta_1}&\cdots&\frac{\partial g_{\mathrm{L}}}{\partial\theta_k}\end{pmatrix}_{\mathrm{L}\times k}$$

$$\nabla_{\theta'}g=\frac{\partial g'}{\partial\theta}=\left(\frac{\partial g}{\partial\theta}\right)'_{k\times\mathrm{L}}$$

证明.首先，找到目标函数的一阶条件：

$$\begin{aligned}\mathrm{FOC}:\frac{\partial\mathrm{Q}_n^{\mathrm{W}}(\hat{\theta})}{\partial\theta}&=\underset{k\times\mathrm{L}}{\frac{\partial\hat{g}'_{1\times\mathrm{L}}}{\partial\theta_{k\times 1}}}\mathrm{W}\hat{g}+\left(\underset{\substack{1\times\mathrm{L}\\ \mathrm{L}\times 1}}{\hat{g}'\mathrm{W}\frac{\partial\hat{g}_{\mathrm{L}\times 1}}{\partial\theta'_{1\times k}}}\right)'\\&=\underset{k\times\mathrm{L}}{2\nabla_{\theta'}\hat{g}_n(\hat{\theta})}\underset{\mathrm{L}\times\mathrm{L}}{\mathrm{W}}\underset{\mathrm{L}\times 1}{\hat{g}_n(\hat{\theta})}=0\end{aligned}$$

由于该一阶条件并不能显式求解$\hat{\theta}$，因此考虑Taylor Expansion

$$\hat{g}_n(\hat{\theta})=\hat{g}_n(\theta_0)+\nabla_\theta\hat{g}_n(\bar{\theta})(\hat{\theta}-\theta_0)$$

将$\hat{g}_n(\hat{\theta})$代入上式有，

$$\begin{aligned}\frac{\partial\hat{\mathrm{Q}}(\hat{\theta})}{\partial\theta}&=2\nabla_{\theta'}\hat{g}_n\mathrm{W}[\hat{g}_n(\theta_0)+\nabla_\theta\hat{g}_n(\bar{\theta})(\hat{\theta}-\theta_0)]=0\\ \sqrt{n}(\hat{\theta}-\theta_0)&=-[\nabla_{\theta'}\hat{g}_n\mathrm{W}\nabla_\theta\hat{g}_n(\bar{\theta})]^{-1}\nabla_{\theta'}\hat{g}_n\mathrm{W}\sqrt{n}\hat{g}_n(\theta_0)\\ \nabla_{\theta'}\hat{g_n}(\hat{\theta})&=\nabla_{\theta'}\frac{1}{n}\sum_{i=1}^{n}g(z_i,\theta)\to\mathrm{E}\left[\nabla_{\theta'}g(z_i,\hat{\theta})\right]=\mathrm{G}'(\hat{\theta})\to\mathrm{G}'(\theta_0)=\mathrm{G}'\end{aligned}$$

故，此时有各项方差为：

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta_0)&=(\mathrm{G}'\mathrm{W}\mathrm{G})^{-1}\mathrm{G}'\mathrm{W}\sqrt{n}\hat{g}_n(\theta_0)\\ \sqrt{n}\hat{g}_n(\theta_0)&=\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}g(z_i,\theta_0)\sim\mathrm{N}(0,\mathrm{E}gg')=\mathrm{N}(0,\Omega_0)\\ \mathrm{V}ar(\sqrt{n}(\hat{\theta}-\theta_0))&=(\mathrm{G}'\mathrm{W}\mathrm{G})^{-1}\mathrm{G}'\mathrm{W}\Omega_0\mathrm{W}\mathrm{G}(\mathrm{G}'\mathrm{W}\mathrm{G})^{-1}\end{aligned}$$

如果$\hat{\theta}_{\text{GMM}} \not\to \theta_0$，则不能使用Taylor Expansion。在最后，我们从直觉上说明为何$\hat{\theta}_{\text{GMM}} \to \theta_0$。根据Uniform Weak Law of Large Number（见**??**），如果$\hat{\text{Q}}_n(\theta) \to \text{Q}_0(\theta)$，则有$\hat{\theta}^{\text{GMM}} = \arg\min \hat{\text{Q}}^n(\theta) \to \theta_0 = \arg\min \text{Q}_0(\theta)$。而根据LLN，有$\hat{\text{Q}}_n(\theta) \to \text{Q}_0(\theta)$（如下所示）：

$$\begin{aligned}\hat{\text{Q}}_n(\theta) &= \hat{g}_n(\theta)'\text{W}\hat{g}_n(\theta) = \left[\frac{1}{n}\sum_{i=1}^{n} g(z_i,\theta)\right]' \text{W} \left[\frac{1}{n}\sum_{i=1}^{n} g(z_i,\theta)\right] \\ \text{Q}_0(\theta) &= \text{E}\left[g(z,\theta)'\right] \text{WE}[g(z,\theta)]\end{aligned}$$

Theorem 2.1.2. ***Consistency of GMM Estimator***

If there is a function $\text{Q}_0(\theta)$ *such that*

i. $\text{Q}_0(\theta)$ *is uniquely minimizaed at* θ_0*;*

ii. Ⓗ *is a compact set of* $\mathscr{R}^k$*;*

iii. $\text{Q}_0(\theta)$ *is continuous in* $\theta \in \text{Ⓗ}$*;*

iv. $\hat{\text{Q}}_n(\theta)$ *uniformly converges to* $\text{Q}_0(\theta)$ *in prob.*

Then we have $\hat{\theta} \xrightarrow{p} \theta$.

最优权重

在Theorem 2.1.1中，若令$\text{W} = \Omega_0^{-1}$，有

$$(\text{G}'\text{WG})^{-1}\text{G}'\text{W}\Omega_0\text{WG}(\text{G}'\text{WG})^{-1} = (\text{G}'\Omega_0^{-1}\text{G})^{-1}$$

此时，$\text{W} = \Omega_0^{-1}$被称为最优权重矩阵（optimal weighting matrix），$(\text{G}'\Omega_0^{-1}\text{G})^{-1}$被称为最优方差（optimal variance），$\hat{\theta}^{\text{GMM}}$被称为有效GMM估计（efficient GMM estimator）。但由于$\text{W} = \Omega_0^{-1}$是未知的（infeasible），所以需要对Ω_0进行估计（$\hat{\Omega}_0$）。

1. Two Step Feasible Efficient GMM Estimation

Step 1: 估计最优权重矩阵（estimate Ω_0 by $\hat{\Omega}$）

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(z_i,\hat{\theta})g(z_i,\hat{\theta})'$$

$$\tilde{\theta} = \arg\min_{\theta\in\text{Ⓗ}}[\hat{g}_n'](\theta)\hat{g}_n(\theta)] = \arg\min_{\theta\in\text{H}} \text{Q}_n^{\text{I}}(\theta)$$

因为无论取什么样的W，$\tilde{\theta}$都是一致的，所以一般取$\text{W} = \text{I}$（先采用任意W执行GM得到$\hat{\Omega}(\hat{\theta})$）。

Step 2: 执行GMM得到最小方差（$\tilde{\theta}^*_{\text{GMM}} = \arg\min \hat{g}_n'(\theta)[\hat{\Omega}(\hat{\theta})]^{-1}\hat{g}_n(\theta)$）

$$\text{V}ar(\sqrt{n}(\tilde{\theta}^*_{\text{GMM}} - \theta_0)) = (\hat{\text{G}}'(\tilde{\theta}^*_{\text{GMM}})\hat{\Omega}^{-1}(\tilde{\theta}^*_{\text{GMM}})\hat{\text{G}}(\tilde{\theta}^*_{\text{GMM}}))^{-1}$$

2. Continuous Updating Method (Hansen, Heaton, & Yaron, 1996)

将$\hat{\Omega}(\theta)$作为θ的函数，直接代入GMM式子中进行统一的最小化损失函数。

$$\hat{\theta} = \arg\min \text{Q}_n(\theta) = \arg\min \hat{g}_n(\theta)'[\hat{\Omega}(\theta)]^{-1}\hat{g}_n(\theta)$$

GMM与之前的方法

GMM本质上包含了所有以前学过的估计方法

0) $\text{L} < k$: Parameters can not be identified (if we do not apply machine learning).

1) $\text{L} = k$：L is # moment conditions and k is # of parameters.

$$\text{F.O.C}: \frac{\partial \hat{\text{Q}}(\theta)}{\partial \theta} = -2\nabla_{\theta'}\hat{g}_n(\hat{\theta})\text{W}\hat{g}_n(\theta) = 0$$

在$\text{L} = k$的情况下，$\nabla_{\theta'}\hat{g}_n(\hat{\theta})$和W都是满秩矩阵，都存在逆矩阵，可直接左乘其逆矩阵进行化简，得到$\hat{g}_n(\theta) = 0$，其与MM一致。

$$\begin{aligned}
\text{V}ar(\sqrt{n}\hat{\theta}^{\text{GMM}}) &= (\text{G}'\text{WG})^{-1}\text{G}'\text{W}\Omega_0\text{WG}(\text{G}'\text{WG})^{-1} \\
&\quad \text{（由于G是方阵，所以}(\text{G}'\text{WG})^{-1}\text{可以直接展开）} \\
&= \text{G}^{-1}\text{W}^{-1}\text{G}'^{-1}\text{G}'\text{W}\Omega_0\text{WGG}^{-1}\text{W}^{-1}\text{G}'^{-1} \\
&= \text{G}^{-1}\Omega_0\text{G}'^{-1} = (\text{G}'\Omega^{-1}\text{G})^{-1}
\end{aligned}$$

该过程说明当$\text{L} = k$时，权重矩阵对最后估计量的方差没有影响。

Example 1: OLS. $\text{L} = k,\ y = \text{X}\beta + u$

使用GMM，样本矩条件为$\frac{1}{n}\text{X}'u = 0$，即$\hat{g}_n(\beta) = \frac{1}{n}\text{X}'(y - \text{X}\beta) = 0$，故有

$$\begin{aligned}
\hat{\beta}^{\text{OLS}} = \hat{\beta}^{\text{GMM}} &= (\text{X}'\text{X})^{-1}\text{X}'y \\
&= \beta + (\text{X}'\text{X})^{-1}\text{X}'u
\end{aligned}$$

根据之前的论述，under homo $\text{V}ar(\hat{\beta}_{\text{OLS}}) = \sigma^2(\text{X}'\text{X})^{-1}$，因此从GMM角度来计算估计的方差。分别计

算G与Ω，有

$$\begin{aligned}\mathrm{G} &= \mathrm{E}[\nabla_\beta g(x_i,\beta_0)] = \mathrm{E}[\nabla_\beta \frac{1}{n}\sum_{i=1}^{n} g(x_i,\beta_0)] \\ &= \mathrm{E}[\nabla_\beta \hat{g}_n(\beta_0)] = -\frac{1}{n}\mathrm{E}(\mathrm{X}'\mathrm{X})\end{aligned}$$

又有，

$$\begin{aligned}\Omega &= \mathrm{E}(g(x_i,\beta_0)g(x_i,\beta_0)') \\ &= \frac{1}{n}\mathrm{E}[\sum_{i=1}^{n} g(x_i,\beta_0)\sum_{j=1}^{n} g(x_j,\beta_0)'] \\ &\text{（由于independence，故}\mathrm{E}(gg')=0 \text{ when, } i\neq j\text{）} \\ &= n\mathrm{E}[\frac{1}{n}\sum_{i=1}^{n} g(x_i,\beta_0)\frac{1}{n}\sum_{j=1}^{n} g(x_j,\beta_0)'] \\ &= n\mathrm{E}[\hat{g}_n(\beta_0)\hat{g}_n(\beta_0)'] \\ &= \frac{1}{n}\mathrm{E}[\mathrm{X}'uu'\mathrm{X}]\end{aligned}$$

$$\begin{aligned}Var(\sqrt{n}\hat{\beta}^{\mathrm{GMM}}) &= (\mathrm{G}'\Omega^{-1}\mathrm{G})^{-1} \\ &= [\mathrm{E}(\frac{\mathrm{X}'\mathrm{X}}{n})]^{-1}\mathrm{E}[\frac{\mathrm{X}'uu'\mathrm{X}}{n}][\mathrm{E}(\frac{\mathrm{X}'\mathrm{X}}{n})]^{-1} \\ &= \sigma^2(\frac{\mathrm{X}'\mathrm{X}}{n})^{-1}\end{aligned}$$

有$\mathrm{Var}(\hat{\beta}^{\mathrm{GMM}})=\mathrm{Var}(\hat{\beta}^{\mathrm{OLS}})$.

Example 2: IV Estimation. $\mathrm{L}=k$, $\mathrm{W}=(\mathrm{X},\mathrm{X}_{\mathrm{IV}})$, weighting matrix is Z.

由前可知，由内生变量时，总体矩条件为$\mathrm{E}[\mathrm{W}'u]=0$

$$\begin{aligned}\hat{\mathrm{Q}}_n &= u'\mathrm{W}\mathrm{Z}^{-1}\mathrm{W}'u \\ &\quad \text{（严格来说，当代入样本矩条件时，这里漏了俩}\frac{1}{n}\text{）} \\ &= (y-\mathrm{X}\beta)'\mathrm{W}\mathrm{Z}'\mathrm{W}'(y-\mathrm{X}\beta) \\ &= y'\mathrm{W}\mathrm{Z}^{-1}\mathrm{W}'y-2\beta'\mathrm{X}'\mathrm{W}\mathrm{Z}^{-1}\mathrm{W}'y+\beta'\mathrm{X}'\mathrm{W}\mathrm{Z}^{-1}\mathrm{W}\mathrm{X}\beta\end{aligned}$$

求解$\hat{\mathrm{Q}}_n$的FOC，有

$$\begin{aligned}\frac{\partial\hat{\mathrm{Q}}_n}{\partial\beta} &= -2\mathrm{X}'\mathrm{W}\mathrm{Z}^{-1}\mathrm{W}'y+2\mathrm{X}'\mathrm{W}\mathrm{Z}^{-1}\mathrm{W}'\mathrm{X}\hat{\beta}=0 \\ \Rightarrow \hat{\beta}^{\mathrm{GMM}} &= (\mathrm{W}'\mathrm{X})^{-1}\mathrm{W}'y=\hat{\beta}^{\mathrm{IV}}\end{aligned}$$

对比$Var(\hat{\beta}^{GMM})$和$Var(\hat{\beta}^{IV})$，有

$$\begin{aligned}
Var(\sqrt{n}\hat{\beta}^{GMM}) &= (G'\Omega^{-1}G')^{-1} \\
&= ((-\frac{1}{n}EW'X)'(\frac{1}{n}EW'uu'W)^{-1}(-\frac{1}{n}EW'X))^{-1} \\
&= \sigma^2(\frac{X'P_WX}{n})^{-1} \\
\sqrt{n}(\hat{\beta}^{IV}-\beta) &= (\frac{W'X}{n})^{-1}\frac{1}{\sqrt{n}}W'u \\
where \quad & (\frac{W'X}{n})^{-1} \to (EW_iX_i')^{-1}, Var\frac{1}{\sqrt{n}}W'u \to \sigma^2EW_iW_i' \\
\Rightarrow \sqrt{n}(\hat{\beta}^{IV}-\beta) &\sim N\left(0, \sigma^2\left(\frac{X'P_WX}{n}\right)^{-1}\right)
\end{aligned}$$

Example 3: IV Estimation. $L > k$, $W = (X, X_{IV})$, weighting matrix is Z.

此时的矩条件为$E[W'u] = 0$，目标函数为

$$\begin{aligned}
\hat{Q}_n &= u'W[Var(W'u)]^{-1}W'u \\
\Omega_0 &= E[gg'] = Var(g(z_i, \theta_0)) \\
&= Var(\frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta)) \\
&= \frac{1}{n}Var(g(z_i, \theta))
\end{aligned}$$

此时的矩条件$g(z_i, \theta) = W'u$，有$Var(W'u) = \frac{1}{\sigma^2}W'W$，故有

$$\begin{aligned}
\hat{Q}_n &= u'W[Var(W'u)]^{-1}Wu' \\
&= \frac{1}{\sigma^2}u'W(W'W)^{-1}W'u \\
&= \frac{1}{\sigma^2}u'P_Wu \\
&= \frac{1}{\sigma^2}(y-X\beta)'P_W(y-X\beta)
\end{aligned}$$

求解一阶条件FOC，

$$\begin{aligned}
\frac{\partial \hat{Q}_n}{\partial \beta} &= -2X'P_Wy + 2X'P_WX\hat{\beta} = 0 \\
\Longrightarrow \hat{\beta}^{GMM} &= (X'P_WX)^{-1}X'P_Wy (= \hat{\beta}^{2SLS}) \\
Var(\sqrt{n}\hat{\beta}^{GMM}) &= \sigma^2(\frac{X'P_WX}{n})^{-1}
\end{aligned}$$

2.2 基于广义矩估计的检验

一般在极大似然估计中会用到三大检验（Wald、LM和LR检验），本节将介绍如何从GMM的视角来进行着三大检验。首先根据一般教科书的惯例，从MLE的角度介绍三种检验的区别。考虑如下非线性约束：

$$\begin{aligned} \mathrm{H}_0 &: r(\theta_0) = 0 \\ \mathrm{H}_1 &: r(\theta_0) \neq 0 \end{aligned}$$

其中，$r : \mathscr{R}^k \to \mathscr{R}^q$。

如Figure 2.1所示，三大检验的主要思想如下：

- Wald Test: 先MLE最大化，得到参数的估计值，代入到$r(\theta)$，如果约束成立则统计量$\mathrm{W} = 0$（只需要求解无约束的最大化问题）；
- LM Test: 求解有约束的最大化问题，得到$\hat{\theta}_{\mathrm{R}}$，比较有约束的Score和0的差异（无约束的Score为0，所以只需要求解有约束的最大化问题）；
- LR Test: 分别求解有约束和无约束的最大化问题，比较似然函数的大小。

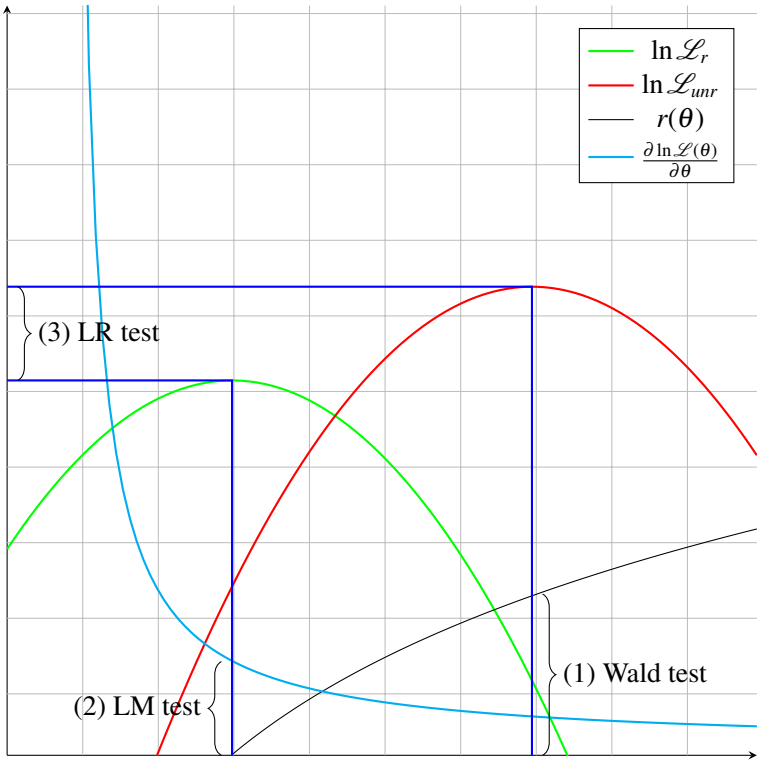

Figure 2.1 三种广义矩估计检验示意图

Wald Test. 找到受约束的分布进行检验（Delta Method）。

$$\begin{aligned}\hat{\theta}_{unr} &= \underset{\theta\in\text{Ⓗ}}{\arg\min}\, \hat{g}_n(\theta)'\hat{\Omega}^{-1}\hat{g}_n(\theta) \\ \sqrt{n}(\hat{\theta}_{unr}-\theta_0) &\sim \mathrm{N}(0,\mathrm{V}_0)\end{aligned}$$

其中，V_0 是用GMM计算的方差。根据Delta Method in Theorem 2.2.1，可得$\sqrt{n}(r(\hat{\theta}_{unr})-r(\theta_0))$的分布为

$$\sqrt{n}(r(\hat{\theta}_{unr})-r(\theta_0)) \sim \mathrm{N}(0,\mathrm{R}_0\mathrm{V}_0\mathrm{R}_0')$$

其中 $\mathrm{R}_0=\mathrm{R}(\theta_0)$, $\mathrm{R}(\theta)=\left(\frac{\partial r(\theta)}{\partial\theta'}\right)_{q\times k}$。故有，

$$\text{Wald test} = nr(\hat{\theta}_{unr})\left[\hat{\mathrm{R}}\hat{\mathrm{V}}\hat{\mathrm{R}}'\right]^{-1} r(\hat{\theta}_{unr}) \sim \chi^2(q)$$

其中$\mathrm{R}=\mathrm{R}(\hat{\theta}_{unr})$，$\hat{\mathrm{V}}=(\hat{\mathrm{G}}'\hat{\Omega}^{-1}\hat{\mathrm{G}})^{-1}$，$\hat{\mathrm{G}}=\frac{1}{n}\sum_{i=1}^n \nabla_\theta g(z_i,\hat{\theta}_{unr})$，$\hat{\Omega}=\frac{1}{n}\sum_{i=1}^n g(z_i,\hat{\theta}_{unr})g(z_i,\hat{\theta}_{unr})'$。

Theorem 2.2.1. ***Delta Method***

If $\hat{\beta}\sim\mathrm{N}(\beta,\mathrm{V}ar(\beta))$ *and* $f(\hat{\beta})=f(\beta)+f^{(1)}(\beta)(\hat{\beta}-\beta)$, *then*

$$f(\hat{\beta})\sim\mathrm{N}\left(f(\beta),f^{(1)}(\beta)\mathrm{V}ar(\hat{\beta})f^{(1)}(\beta)'\right)$$

LM test. 拉格朗日乘子检验：找到一阶条件的分布进行检验。

如Figure 2.1所示，LM主要检验的是受约束的目标函数一节条件是否为零，即

$$\frac{\partial\hat{\mathrm{Q}}_n(\hat{\theta}_r)}{\partial\theta}=2\nabla_{\theta'}\hat{g}_n(\hat{\theta}_r)\hat{\Omega}^{-1}\hat{g}_n(\hat{\theta}_r)=0$$

构造LM统计量，

$$\begin{aligned}\mathrm{LM} &= \underset{1\times k}{\left(\frac{\partial\hat{\mathrm{Q}}(\hat{\theta}_r)}{\partial\theta}\right)'}\left[\underset{k\times k}{\mathrm{V}ar\left(\frac{\partial\hat{\mathrm{Q}}(\hat{\theta}_r)}{\partial\theta}\right)}\right]^{-1}\left(\frac{\partial\hat{\mathrm{Q}}(\hat{\theta}_r)}{\partial\theta}\right) \\ \frac{\partial\hat{\mathrm{Q}}_n(\hat{\theta}_r)}{\partial\theta} &= 2\mathrm{G}(\hat{\theta}_r)'\hat{\Omega}^{-1}\hat{g}_n(\hat{\theta}_r) \\ \mathrm{V}ar\left(\frac{\partial\hat{\mathrm{Q}}_n(\hat{\theta}_r)}{\partial\theta}\right) &= \frac{4}{n}\mathrm{G}'(\hat{\theta}_r)\hat{\Omega}^{-1}\mathrm{V}ar(\sqrt{n}\hat{g}_n(\hat{\theta}_r))\hat{\Omega}^{-1}\mathrm{G}(\hat{\theta}_r) \\ &= \frac{4}{n}\underset{k\times\mathrm{L}}{\mathrm{G}'(\hat{\theta}_{res})}\quad\underset{\mathrm{L}\times\mathrm{L}}{\hat{\Omega}^{-1}}\quad\underset{\mathrm{L}\times k}{\mathrm{G}(\hat{\theta}_{res})}\end{aligned}$$

故有，

$$\mathrm{LM} = n[\hat{g}_n(\hat{\theta}_{res})'\underset{1\times k}{\hat{\Omega}^{-1}}\mathrm{G}(\hat{\theta}_{res})][\mathrm{G}'(\hat{\theta})_{res}\underset{k\times k}{\hat{\Omega}^{-1}}\mathrm{G}(\hat{\theta}_{res})]^{-1}[\mathrm{G}'(\hat{\theta}_{res})\underset{k\times 1}{\hat{\Omega}^{-1}}\hat{g}_n(\hat{\theta}_{res})] \sim \chi^2(q),\ q < k$$

LR test. 似然比检验：找到目标函数在受限与非受限情况之差的分布进行检验。

$$\hat{\mathrm{Q}}_n(\theta) = \hat{g}_n(\theta)'\Omega^{-1}\hat{g}_n(\theta)$$
$$\Omega = \mathrm{E}gg' = \mathrm{V}ar(\sqrt{n}\hat{g}_n) = n\mathrm{V}ar(\hat{g}_n) \Rightarrow \Omega^{-1} = \frac{1}{n}\mathrm{V}ar^{-1} \to 0$$
$$\therefore objective function is \quad n\hat{\mathrm{Q}}(\theta)$$

无约束的GMM目标函数：

$$n\hat{\mathrm{Q}}(\theta_{unr}) = (\sqrt{n}\hat{g}_n(\hat{\theta}_{unr}))'[\mathrm{V}ar\sqrt{n}\hat{g}_n(\hat{\theta}_{unr})]^{-1}(\sqrt{n}\hat{g}_n(\hat{\theta}_{unr})) \sim \chi^2(l-k)$$

有约束的GMM目标函数：

$$n\hat{\mathrm{Q}}_n(\theta_r) = (\sqrt{n}\hat{g}_n(\hat{\theta}_r))'[\mathrm{V}ar\sqrt{n}\hat{g}_n(\hat{\theta}_r)]^{-1}(\sqrt{n}\hat{g}_n(\hat{\theta}_r)) \sim \chi^2(l-(k-q))$$

有约束与无约束的GMM目标函数之差：

$$\mathrm{LR} = n\hat{\mathrm{Q}}_n(\hat{\theta}_{\mathrm{R}}) - n\hat{\mathrm{Q}}_n(\hat{\theta}) \sim \chi^2(q)$$

Note: Wald/LM/LR tests are equivalent if consistent estimation of Ω_0 is used. (Newey & West, 1987)

Test of Moment Restrictions. If the number of momentum restrictions is greater than that of the parameters (L > K), we might carefully select some of conditions based on the overidentification test. We begin by partitioning the moment restrictions into a set of k reliable moment conditions that identifies θ_0 $\mathrm{E}(g_l(z,\theta_0)) = 0$ for $l = 1,2,\ldots,k$ and a set of remaining questionable moment restrictions that comprise the H_0

$$\mathrm{H}_0 \text{æ} \mathrm{E}(g_l(z,\theta_0)) = 0 \quad l = k+1, k+2, \ldots, \mathrm{L}$$
$$\mathrm{H}_1 : \mathrm{E}(g_l(z,\theta_0)) \neq 0 \quad for some\ l = k+1, k+2, \ldots, \mathrm{L}$$

The logics behind this is that after selecting k from L moment conditions to estimate the parameters, we could test the rest of moment conditions. Hence test of moment conditions requires $\mathrm{L} > k$.

增广矩条件（Augumented Moment Condition）：将检验矩条件是否冗余转化为检验参数。

$$g^a(z,\theta,\varphi)=[\underbrace{g_1(z,\theta),\ \ldots,\ g_k(z,\theta)}_{reliable\quad \mathrm{E}(\cdot)=0},\ g_{k+1(z,\theta)}-\varphi_1,\ \ldots,\ g_{\mathrm{L}}(z,\theta)-\varphi_{\mathrm{L}-k}]'$$

故，此时的原假设与备择假设为

$$\mathrm{H}_0:\varphi_j=0,\ \forall j=1,2,\ldots,\mathrm{L}-k$$
$$\mathrm{H}_1:\varphi_j\neq 0,\ \exists j=1,2,\ldots,\mathrm{L}-k$$

此时可以使用前述的三大检验以检验“冗余”的矩条件是否被现有参数满足。以似然比检验为例（LR test)：

$$\begin{aligned}\mathrm{LR}&=n[\hat{\mathrm{Q}}^a(\hat{\theta}_r,\overset{\varphi_j=0}{0})-\hat{\mathrm{Q}}^a(\hat{\theta}_{unr},\hat{\varphi}_{unr})]\\&=n[\hat{\mathrm{Q}}_n(\hat{\theta}_{unr})-0]\\&=n\hat{\mathrm{Q}}_n(\hat{\theta})\sim\chi^2(\mathrm{L}-k)\end{aligned}$$

第三章　M估计

3.1　Estimation

An estimation of $\hat{\theta}$ is a M-estimator if there is an objective function $\hat{\mathrm{Q}}(w_i, \theta)$, where $w_i = \{y_i, x_i\}$ such that

$$\hat{\theta} = \arg\max / \min \hat{\mathrm{Q}}(w_i, \theta),\ \theta \in \text{Ⓗ}$$

This obejective function is more general than what we have learned above. M-estimation seems to be the most general form of one class of extremum estimators, the other is Minimum Distance Estimators, in which the objective function is a measure of a *distance*. For the name, "M" stands for a maximum or minimum estimators (Huber, 1967).

The examples of M-estimation we have learned:

- OLS: $\hat{\beta}^{\mathrm{OLS}} = \arg\min(y - \mathrm{X}\beta)(y - \mathrm{X}\beta)'$
- MLE: $\hat{\beta}^{\mathrm{MLE}} = \arg\max \ln \mathrm{L} = \arg\max(\sum_{i=1}^{\mathrm{T}} \ln f(x_i, y_i, \beta))$
- GMM: $\hat{\theta}^{\mathrm{GMM}} = \arg\max \hat{g}_n'(\theta)\mathrm{W}\hat{g}_n$

本章主要介绍的是非线性模型的参数估计（Nonlinear Regression)。具体而言，此处介绍的依旧是最小二乘框架下的非线性模型估计，应当为Nonlinear Least Square Regression（NLLS）。

$$\begin{aligned} \text{Parametric model} \quad &: \quad y_i = m(x_i, \theta) + u_i \qquad m()\text{已知，估计}\theta \\ \text{Non-parametric model} \quad &: \quad y_i = m(x) + u_i \qquad m()\text{未知，估计}m \end{aligned}$$

以logistic regression为例，我们已知$m(\mathrm{X}, \theta) = \frac{\exp(\mathrm{X}\theta)}{1+\exp(\mathrm{X}\theta)}$，

$$\begin{aligned} y &= m(\mathrm{X}, \theta) + u\theta \\ \mathrm{E}(y|\mathrm{X}) &= \mathrm{E}[m(\mathrm{X}, \theta)] + \mathrm{E}[u|\mathrm{X}] \\ \textit{assume}\ \mathrm{E}[u|\mathrm{X}] = 0 \Rightarrow \mathrm{E}(y|\mathrm{X}) &= m(\mathrm{X}, \theta) \end{aligned}$$

The Existence of Estimator

Assumption 3.1.1. *For some* $\theta_0 \in \Theta$, $\mathrm{E}[y|\mathrm{X}] = m(\mathrm{X},\theta)$ *when minimizing* $\mathrm{E}[y-m(x,\theta)]^2$.

根据Assumption 3.1.1，在求解目标函数过程中，我们有：

$$
\begin{aligned}
\mathrm{E}(y-m(x,\theta))^2 &= \mathrm{E}[y-\mathrm{E}(y|\mathrm{X})+\mathrm{E}(y|\mathrm{X})-m(\mathrm{X},\theta]^2 \\
&= \mathrm{E}[(y-m(\mathrm{X},\theta_0))+(m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta))]^2 \\
&= \mathrm{E}[y-m(\mathrm{X},\theta_0)]^2+2\mathrm{E}[(y-m(\mathrm{X},\theta_0))(m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta))]+\mathrm{E}[m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta)]^2 \\
&= \mathrm{E}[y-m(\mathrm{X},\theta_0)]^2+2\mathrm{E}\left[\mathrm{E}[(y-m(\mathrm{X},\theta_0))(m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta))|\mathrm{X}]\right]+\mathrm{E}[m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta)]^2 \\
&= \mathrm{E}[y-m(\mathrm{X},\theta_0)]^2+2\mathrm{E}\left[\mathrm{E}[y-m(\mathrm{X},\theta_0)|\mathrm{X}](m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta))\right]+\mathrm{E}[m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta)]^2 \\
&= \mathrm{E}[y-m(\mathrm{X},\theta_0)]^2+2\mathrm{E}\left[\underset{=0)}{(\mathrm{E}(y|\mathrm{X})-m(\mathrm{X},\theta_0))}(m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta))\right]+\mathrm{E}[m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta)]^2 \\
&= \underset{\text{与}\theta\text{无关}}{\mathrm{E}[y-m(\mathrm{X},\theta_0)]^2}+\underset{\text{取决于}\theta\text{的取值}}{\mathrm{E}[m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta)]^2}
\end{aligned}
\tag{3.1}
$$

在Equation 3.1中，我们讨论θ的取值对估计的影响：

1) if $\theta=\theta_0$ then $\theta=\theta_0=\arg\min_{\theta\in\Theta}\in \mathrm{E}[y-m(\mathrm{X},\theta_0)]^2$

2) if $\theta\neq\theta_0$ then $\mathrm{E}[m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta)]^2\geq 0$:

   - $\mathrm{E}[m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta)]^2=0$: θ_0 *cannot* be uniquely identified;
   - $\mathrm{E}[m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta)]^2>0$: θ_0 *can* be uniquely identified.

The Uniqueness of Estimator

Assumption 3.1.2. *If* $\mathrm{E}[m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta)]^2>0$, $\theta\neq\theta_0\ \forall\theta\in\Theta$.

以线性模型为例（假设$m(\mathrm{X},\theta)=\mathrm{X}\theta$），有

$$
\begin{aligned}
\mathrm{E}[m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta)]^2 &= \mathrm{E}[(\mathrm{X}\theta_0-\mathrm{X}\theta)'(\mathrm{X}\theta_0-\mathrm{X}\theta)] \\
&= \mathrm{E}[(\theta_0-\theta)'\mathrm{X}'\mathrm{X}(\theta_0-\theta)]>0
\end{aligned}
$$

因此要求X′X是正定矩阵（positive definite），即X矩阵列满秩（无完全多重共线性），$rank(\mathrm{E}[\mathrm{X}'\mathrm{X}])=k$。

一个不满足NLS 2的例子，假设：

$$
\begin{aligned}
\text{真实模型} &: m(x,\theta_0)=\theta_{10}+\theta_{20}x_2 \\
\text{待估计的模型} &: m(\mathrm{X},\theta)=\theta_1+\theta_2x_2+\theta_3x_3^{\theta_4}
\end{aligned}
$$

可得$\hat{\theta}_1 = \theta_{10}, \hat{\theta}_2 = \theta_{20}, \hat{\theta}_3 = 0, \hat{\theta}_4 =$ any value。所以违背了Assumption 3.1.2。

广义M估计. 在引入相关假设之后，我们首先表示广义M估计的目标函数为（不局限于Least Square）：

$$\min_{\theta \in \text{Ⓗ}} \mathrm{E}[q(w, \theta)] \tag{3.2}$$

The identification in Assumption 3.1.2 requires:

$$\mathrm{E}[q(w, \theta_0)] < \mathrm{E}[q(w, \theta)], \ \forall \theta \in \text{Ⓗ} \ and \ \theta \neq \theta_0 \tag{3.3}$$

用算术平均值代替期望

$$\hat{\theta} = \min_{\theta \in \text{Ⓗ}} \frac{1}{n} \sum_{i=1}^{n} q(w_i, \theta) \tag{3.4}$$

此时，我们的问题转变为：在什么条件下，估计量满足一致性条件，即$\hat{\theta} \xrightarrow{p} \theta_0$。与GMM中的思想类似（在一定的条件下），我们能发现，如果目标函数是一致的，则其对应的估计值也是一样的。

Theorem 3.1.1. ***Uniform Weak Law of Large Numbers***

If

1. *Data w_i is i.i.d;*
2. *$\theta \in$ Ⓗ, Ⓗ is a compact set;*
3. *$\forall w_i$, $q(w)$ is continuous on Ⓗ;*
4. *$|q(w_i, \theta)| \leq b(w_i)$, $\forall \theta \in$ Ⓗ, and $\mathrm{E}[b(w_i)] < \infty$(有界).*

Then we have

$$\frac{1}{n} \sum_{i=1}^{n} q(w_i, \theta) \xrightarrow{p} \mathrm{E}[q(w, \theta)]$$

Theorem 3.1.2. ***Consistency of M-estimator***

Theorem 3.1.1 and Assumption 3.1.2 hold, then we have

$$\hat{\theta} \xrightarrow{p} \theta_0$$

Proof. See Newey and Mcfadden(1994).

考虑更一般的情形，如果$\hat{\theta} \xrightarrow{p} \theta_0$ as $n \to \infty$，那么$\frac{1}{n} \sum_{i=1}^{n} r(w_i, \hat{\theta}) \xrightarrow{?} \mathrm{E}(r(w, \theta_0))$

Lemma 3.1.1. *Suppose that $\hat{\theta} \to \theta_0$ and assume any functions $r(w_i, \theta)$ satisfies the same assumption as in Theorem 3.1.2 (which means $r(w, \theta)$ is continuous and bounded), then*

$$\frac{1}{n}\sum_{i=1}^{n} r(w_i, \hat{\theta}) \xrightarrow{p} \mathrm{E}\left[r(w, \theta_0)\right]$$

回到Equation 3.4，此时我们要解决的问题是如何找到$\hat{\theta}$的分布（$\hat{\theta} \sim ?$）

$$\begin{aligned} &\text{objective function:} && \min_{\theta \in \Theta} \tfrac{1}{n}\textstyle\sum_{i=1}^{n} q(w_i, \theta) \\ &\text{FOC:} && \textstyle\sum_{i=1}^{n} \frac{\partial q(w_i, \hat{\theta})}{\partial \theta} \overset{\text{Taylor Exp}}{=} \sum_{i=1}^{n} \frac{\partial q(w_i, \theta_0)}{\partial \theta} + \sum_{i=1}^{n} \frac{\partial^2 q(w_i, \bar{\theta})}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0) = 0 \end{aligned}$$

Let $\mathrm{H}_i = \mathrm{H}(w_i, \bar{\theta}) = \frac{\partial^2 q(w_i, \bar{\theta})}{\partial \theta \partial \theta'}$ be the *Hessian matrix* of the objective function, $\mathrm{S}(w_i, \theta_0) = \frac{\partial q(w_i, \theta_0)}{\partial \theta}$ be the *Score* of the objective function.

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 q(w_i, \bar{\theta})}{\partial \theta \partial \theta'} \sqrt{n}(\hat{\theta} - \theta_0) &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial q(w_i, \theta_0)}{\partial \theta} \\ \frac{1}{n}\sum_{i=1}^{n} \mathrm{H}(w_i, \bar{\theta}) \sqrt{n}(\hat{\theta} - \theta_0) &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{S}(w_i, \theta_0) \\ \Longrightarrow \sqrt{n}(\hat{\theta} - \theta_0) &= (\frac{1}{n}\sum_{i=1}^{n} \mathrm{H}_i)^{-1} (-\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{S}(w_i, \theta_0)) \end{aligned}$$

根据Lemma 3.1.1，可得

$$\begin{aligned} \frac{1}{n}\sum_{i}^{n} \mathrm{H}_i = \frac{1}{n}\sum_{i=1}^{n} \mathrm{H}(w, \bar{\theta}) &\xrightarrow{p} \mathrm{E}[\mathrm{H}(w, \theta_0)] \overset{def}{=} \mathrm{A}_0 \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} [-\mathrm{S}(w_i, \theta_0)] &\sim \mathrm{N}(0, \underbrace{\mathrm{E}\mathrm{S}_i\mathrm{S}_i'}_{=\mathrm{B}_0}) \\ \Longrightarrow \sqrt{n}(\hat{\theta} - \theta_0) &\sim \mathrm{N}(0, \mathrm{A}_0^{-1}\mathrm{B}_0\mathrm{A}_0^{-1}) \\ \sqrt{n}(\hat{\theta} - \theta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} [-\mathrm{A}_0^{-1}\mathrm{S}_i(\theta_0)] + o_p(1) \\ &\overset{def}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} r_i(\theta_0) + o_p(1) \end{aligned} \tag{3.5}$$

which is the influnce function representation of $\hat{\theta}$, where $e(w_i, \theta_0)$ is called the influence function. [1]

$\sqrt{n}(\hat{\theta} - \theta_0) \sim \mathrm{N}(0, \mathrm{A}_0^{-1}\mathrm{B}_0\mathrm{A}_0^{-1})$ 要求

- $\mathrm{E}(\mathrm{S}_i) = 0$: M-estimator的目标函数为$\mathrm{E}[q(w_i, \theta)]$，其FOC为$\frac{\partial \mathrm{E}[q(w_i, \theta)]}{\partial \theta} = 0$，即$\mathrm{E}\frac{\partial [q(w_i, \theta)]}{\partial \theta} = 0$;

[1] For the notation here, we define $o(1)$ as a series converaging to 0, and $o_p(1)$ a series of random variable converaging to 0. If $\frac{a_n}{b_n} = o(1)$, we could note it as $a_n = o(b_n)$. If we use capital letter O, it means the series are bounded.

- A_0可逆：在最小化问题中，如果Assumption 3.1.2成立，则有Hessian矩阵正定， A_0必然可逆。

$$\begin{aligned} q(w,\theta_0) &= \frac{1}{2}(y-m(x_i,\theta_0))^2 \\ S_i &= -\frac{\partial m_i}{\partial \theta}(y_i-m_i) = -\nabla_{\theta_0} m_i'(y_i-m_i) \\ E[S_i] &= E[E(S_i|X_i)],\ E(S_i|X_i)=0 \end{aligned}$$

举例说明能否得到M-estimator的情况：

Example 1: OLS.

$$m(X,\theta) = X\theta, \Longrightarrow A_0 = E(X'X)$$

Example 2: NLLS.

$$\begin{aligned} m(X,\theta) &= \theta_1 + \theta_2 X_2 + \theta_3 X_3^{\theta_4}, where\ \theta_3 = 0,\ \theta_4 = \text{any value} \\ H(w,\theta) &= \nabla_\theta m(X,\theta)'\nabla_\theta m(X,\theta) - \nabla_\theta^2 m(X,\theta)(y-m(X,\theta)) \\ \Longrightarrow A_0 &= E[H(w,\theta_0)] = E[\nabla_\theta m(X,\theta)'\nabla_\theta m(X,\theta)] \\ &= E\left[\begin{pmatrix} 1 \\ x_2 \\ x_3^{\theta_4} \\ \underset{=0}{\theta_3 x_3^{\theta_4} ln(x_3)} \end{pmatrix} \left(1,\ x_2,\ x_3^{\theta_4},\ \underset{=0}{\theta_3 x_3^{\theta_4} ln(x_3)}\right)\right] \\ &= E\left[\begin{pmatrix} & & & 0 \\ & \ddots & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\right] \end{aligned}$$

由于A_0的行列式$|A_0| = 0$，故A_0不可逆，所以此时无法使用M估计。

3.2 Two Step M-Estimation

Sometimes, nonlinear models depend not only on our parameter of interest θ, but nuisance parameters or unobserved variables in some way. It is common to estimate θ using a "two-step" procedure:

$$\begin{aligned} 1^{st}\text{-stage} &: y_2 = g(w_i,\gamma) + e \Longrightarrow \text{We estimate } \hat{\gamma}, \text{ say } \gamma \\ 2^{nd}\text{-stage} &: y = m(w_i,\theta,\hat{\gamma}) + u \Longrightarrow \text{We estimate } \hat{\theta}, \text{ given } \hat{\gamma}. \end{aligned}$$

Finnally, the objective function for our last M-estimation:

$$\min_{\theta} \sum_{i=1}^{n} q(w_i, \theta, \hat{\gamma}) \ (= u'u)$$

We will test the properties of two-step M-estimation through Weighted Nonlinear Least Square (WNLS), that include *consistency* by a uniform weak LLN and asymptotic normality by CLT (Pagan, 1984, 1986, in which it is calld *generated regressors*).

Example: WNLS. 该模型满足如下设定：

$$\begin{aligned} y_i &= m(x_i, \theta) + u_i, \ \mathrm{E}(u_i^2 | x_i) = h(x_i, \gamma_0) \\ \frac{y_i}{\sqrt{h(x_i, \gamma_0)}} &= \frac{m(x_i, \theta)}{\sqrt{h(x_i, \gamma_0)}} + \frac{u_i}{\sqrt{h(x_i, \gamma_0)}} \end{aligned}$$

其中，γ_0为权重函数中的参数。最后，目标函数为：

$$\begin{aligned} &\min_{\theta \in \text{Ⓗ}} \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - m(x_i, \theta))^2}{h(x_i, \gamma_0)} \\ 1^{st}\text{-stage for } \gamma\text{: } &\hat{u}_i^2 = h(x_i, \gamma) + \varepsilon_i \overset{\text{M–E}stimation}{\Longrightarrow} \hat{\gamma} \\ &\min_{\theta \in \text{Ⓗ}} \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - m(x_i, \theta))^2}{h(x_i, \hat{\gamma})} \end{aligned}$$

The Existence of Estimator

Assumption 3.2.1. *For some* $\theta_0 \in \text{Ⓗ}$, $\mathrm{E}[\frac{u}{\sqrt{h(x_i, \gamma_0)}} | \mathrm{X}] = 0$, *that is,*

$$\mathrm{E}\left[\frac{y_i}{\sqrt{h(x_i, \gamma_0)}}\right] = \mathrm{E}\left[\frac{m(x_i, \theta)}{\sqrt{h(x_i, \gamma_0)}}\right] + \mathrm{E}\left[\frac{u_i}{\sqrt{h(x_i, \gamma_0)}}\right]$$

which is the same as Assumption 3.1.1.

The Uniqueness of Estimator

Assumption 3.2.2. $\forall \theta \in \textcircled{H}$ *and* $\theta \neq \theta_0$,

$$\mathrm{E}\left[\frac{(m(\mathrm{X},\theta_0)-m(\mathrm{X},\theta))^2}{h(\mathrm{X},\gamma^*)}\right] > 0$$

Or say, in a more general way,

$$\mathrm{E}[q(\mathrm{W},\theta_0,\gamma^*)] < \mathrm{E}[q(\mathrm{W},\theta,\gamma^*)]$$

Where γ^* *is the value* γ *approaches to as* $n \to \infty$, *which*

- 在模型设定正确的情况下有，$\mathrm{V}ar(y|\mathrm{X}) = h(\mathrm{X},\gamma_0), \hat{\gamma} \xrightarrow{p} \gamma_0$
- 在模型设定错误的情况下有，$\mathrm{V}ar(y|\mathrm{X}) \neq h(\mathrm{X},\gamma_0), \hat{\gamma} \xrightarrow{p} \gamma^*$

Like what we have in Theorem 3.1.2, we finnally get the consistency of estimations:

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} q(w_i,\theta,\gamma^*) \xrightarrow{p} \mathrm{E}[q(w_i,\theta,\gamma^*)] \\ \text{Identification Condition} \end{cases} \Longrightarrow \text{Consistency: } \hat{\theta} \xrightarrow{p} \theta_0 \tag{3.6}$$

Lemma 3.2.1. *Like Lemma 3.1.1, Suppose that* $\hat{\theta} \to \theta_0$ *and assume any functions* $q(w_i,\theta)$ *satisfies the same assumption as in Theorem 3.1.2 (which means* $q(w,\theta)$ *is continuous and bounded), then we have*

$$\frac{1}{n}\sum_{i=1}^{n} q(w_i,\hat{\theta},\hat{\gamma}) \to \mathrm{E}[q(w_i,\theta,\gamma^*)]$$

Based on the theorem and lemma, and like Equation 3.5, we have

$$\sqrt{n}(\hat{\theta}-\theta_0) = \mathrm{A}_0^{-1}(-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{S}_i(\theta_0,\hat{\gamma})) + o_p(1)$$

For $\hat{\gamma}$, we take the Taylor expansion at γ^* with risiduals (Liu: I don't remeber the exactly name of this form).

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{S}_i(\theta_0,\hat{\gamma}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{S}_i(\theta_0,\gamma^*) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \mathrm{S}_i(\theta_0,\gamma^*)}{\partial \gamma^*}(\hat{\gamma}-\gamma^*) + o_p(1) \tag{3.7}$$

Because we take $\hat{\gamma}$ from the 1^{st}-stage M-estimation, $\sqrt{n}(\hat{\gamma}-\gamma^*) \overset{p}{\sim} \mathrm{N}(0,?) = \mathrm{O}_p(1)$.

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \mathrm{S}_i(\theta_0,\gamma^*)}{\partial \gamma^*}(\hat{\gamma}-\gamma^*) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathrm{S}_i(\theta_0,\gamma^*)}{\partial \gamma^*}\sqrt{n}(\hat{\gamma}-\gamma^*) \to \mathrm{E}\left[\mathrm{S}_i(\theta_0,\gamma^*)\gamma^*\right]\mathrm{O}_p(1)$$

令$F_0 \overset{def}{=} E\left[\nabla_\gamma S(W,\theta_0,\gamma^*)\right]$，Equation 3.7则化简为：

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\theta_0,\hat{\gamma}) \to \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\theta_0,\gamma^*) + \underbrace{E\left[\nabla_\gamma S(W,\theta_0,\gamma^*)\right]}_{=F_0}\sqrt{n}(\hat{\gamma}-\gamma^*) + o_p(1) \tag{3.8}$$

此时，F_0的取值决定了$\sqrt{n}(\hat{\theta}-\theta_0)$的分布。

- if $F_0 = 0$.

此时，第一阶段对$\hat{\gamma}$的估计精度不影响结果，可得

$$\sqrt{n}(\hat{\theta}-\theta_0) \sim N(0, A_0^{-1}B_0A_0^{-1}) \tag{3.9}$$

其中，$A_0 = E\left[H(w,\theta_0,\gamma^*)\right]$, $B_0 = E\left[S(w,\theta_0,\gamma^*)S(w,\theta_0,\gamma^*)'\right]$。

- if $F_0 \neq 0$.

对于Probit，Tobit模型而言，$F_0 \neq 0$。此时我们需考虑由于$\hat{\gamma}$对$\hat{\theta}$方差的影响，类似于Equation 3.5，Influence function representation of $\hat{\gamma}$:

$$\sqrt{n}(\hat{\gamma}-\gamma^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} r_i(\gamma^*) + o_p(1)$$

. For $\sqrt{n}(\hat{\theta}-\theta)$, plugging the above in, we have

$$\begin{aligned}
\sqrt{n}(\hat{\theta}-\theta_0) &= -A_0^{-1}[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\theta_0,\gamma^*) + F_0\sqrt{n}(\hat{\gamma}-\gamma^*)] + o_p(1) \\
&= -A_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[S_i(\theta_0,\gamma^*) + F_0 r(\gamma^*)] + o_p(1) \\
&= A_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[-g_i(\theta_0,\gamma^*)] + o_p(1)
\end{aligned}$$

其中$E[-g_i(\theta_0,\gamma^*) = 0]$，本质上它是两个阶段M估计中各自FOC条件的和，此处合写为一个$g(\cdot)$。有$D_0 \overset{def}{=} Var(-g_i(\theta_0,\gamma^*) = E[g_i(\theta_0,\gamma^*)g_i(\theta_0,\gamma^*)']$。故，最后我们可得

$$\sqrt{n}(\hat{\theta}-\theta_0) \sim N(0, A_0^{-1}D_0A_0^{-1})$$

其中，

$$D_0 = E[S_i(\theta_0,\gamma^*)S_i(\theta_0,\gamma^*)'] + F_0E[r_i(\gamma*)r_i(\gamma^*)']F_0' = B_0 + \text{Positive Definte} > B_0$$

而$A_0 = E[H(W,\theta_0)]$, $B_0 = E[S(W,\theta_0,\gamma^*)S(W,\theta_0,\gamma^*)']$。

但由于上述方差都是渐近方差，因而需要用数据进行估计（样本方差代替总体方差）。

Method 1.

$$\begin{aligned} \mathrm{A} &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{H}(w_i,\hat{\theta}) \to \mathrm{A}_0 \\ \mathrm{B} &= \frac{1}{n}\sum_{i=1}^{n}[\mathrm{S}(w_i,\hat{\theta})\mathrm{S}(w_i,\hat{\theta})'] \to \mathrm{B}_0 \end{aligned}$$

但在实际情况中，计算二阶导对算力要求较高（其复杂度为$o(n^2)$），因而通常会用其他方式来求近似。

Method 2. 在模型设定正确的情况下，$\mathrm{A}(\mathrm{X},\theta_0)=\mathrm{E}[\mathrm{H}(\mathrm{W},\theta_0)|\mathrm{X}]$，对于非线性模型而言，$y=m(\mathrm{X},\theta_0)+u$，此时有，

$$\begin{aligned} \mathrm{H}(\mathrm{W},\theta_0) &= \underset{k\times 1}{\nabla_\theta m(\mathrm{X},\theta_0)'}\underset{1\times k}{\nabla_\theta m(\mathrm{X},\theta_0)} - \nabla_\theta^2 m(\mathrm{X},\theta_0) \quad \underset{\text{Assumption 3.2.1}\Rightarrow \mathrm{E}(u|\mathrm{X})=0}{[y-m(\mathrm{X},\theta_0)]} \\ \mathrm{A}_0 &= \mathrm{E}[\mathrm{H}(\mathrm{X},\theta_0)|\mathrm{X}] \\ &= \mathrm{E}\left[\nabla_\theta m(\mathrm{X},\theta_0)'\nabla_\theta m(\mathrm{X},\theta_0)\right] \\ \hat{\mathrm{A}} &= \frac{1}{n}\sum_{i=1}^{n}\underset{\hat{\mathrm{A}}_i}{\mathrm{A}(\mathrm{X}_i,\hat{\theta})} \to \mathrm{A}_0 \\ \hat{\mathrm{B}}_0 &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{S}(w_i,\hat{\theta})\mathrm{S}(w_i,\hat{\theta})' \to \mathrm{B}_0 \\ \sqrt{n}(\hat{\theta}-\theta_0) &\to \mathrm{N}(0,\hat{\mathrm{A}}_0^{-1}\hat{\mathrm{B}}_0\hat{\mathrm{A}}_0^{-1}) \end{aligned}$$

最后，我们可得在模型设定正确与否情况下$\hat{\theta}$的方差：

$$\hat{Var}(\hat{\theta}) = \begin{cases} (\sum_{i=1}^{n}\hat{\mathrm{H}}_i)^{-1}(\sum_{i=1}^{n}\hat{\mathrm{S}}_i\hat{\mathrm{S}}_i')(\hat{\mathrm{H}}_i)^{-1} & \text{Fully Robust Estimator} \\ (\sum_{i=1}^{n}\hat{\mathrm{A}}_i)^{-1}(\sum_{i=1}^{n}\hat{\mathrm{S}}_i\hat{\mathrm{S}}_i')(\hat{\mathrm{A}}_i)^{-1} & \text{Semi Robust Estimator} \end{cases} \tag{3.10}$$

对于非线性模型而言，可得

$$\begin{aligned} \hat{\mathrm{A}}(\mathrm{X}_i,\hat{\theta}) &= \underset{k\times 1}{\nabla_\theta \hat{m}_i'}\underset{1\times k}{\nabla_\theta \hat{m}_i} \\ \hat{\mathrm{S}}_i &= -\nabla_\theta \hat{m}_i(y_i-\hat{m}_i) = -\nabla_\theta \hat{m}_i\hat{u}_i \\ \hat{Var}(\hat{\theta}) &= (\sum_{i=1}^{n}\nabla_\theta \hat{m}_i'\hat{m}_i)^{-1}(\sum_{i=1}^{n}\hat{u}_i\nabla_\theta \hat{m}_i'\hat{m}_i)(\sum_{i=1}^{n}\nabla_\theta \hat{m}_i'\hat{m}_i)^{-1} \end{aligned} \tag{3.11}$$

在Stata等统计软件中，robust 一般指semi rubust estimator，即 Heteroskadasticity Robust Variance Estimation,并不是模型正确设定与否的robustness（fully robust）。

3.3 Homoscedasticity

回到M-estimation估计量方差的具体形式，我们发现其仍然受到u方差结构的影响，因而在一些情况下，为简化讨论，我们为M-estimation和Two step M-estimation分别引入了同方差假设来计算其估计量具体的方差形式。

3.3.1 M-Estimation

Homoscedasticity

Assumption 3.3.1. *For $y = m(\mathrm{X}, \theta_0) + u$, we assume that*

$$\mathrm{V}ar(y|\mathrm{X}) = \mathrm{V}ar(u|\mathrm{X}) = \sigma_0^2$$

基于Equation 3.9，我们可得

$$\begin{aligned}
\mathrm{B}_0 &= \mathrm{E}\left[\mathrm{S}(w,\hat{\theta})\mathrm{S}(w,\hat{\theta})'\right] \\
&= \mathrm{E}[\mathrm{E}[u^2|\mathrm{X}]\nabla_\theta m(\mathrm{X},\theta_0)'\nabla_\theta m(\mathrm{X},\theta_0)] \\
&= \sigma_0^2\mathrm{E}[\nabla_\theta m(\mathrm{X},\theta_0)'\nabla_\theta m(\mathrm{X},\theta_0)] \\
&= \sigma_0^2\mathrm{E}[\mathrm{H}(w,\theta_0)] \\
&= \sigma_0^2\mathrm{A}_0 \text{ Generalized Information Matrix Equality (GIME)} \\
\mathrm{V}ar(\sqrt{n}\hat{\theta}) &= \mathrm{A}_0^{-1}\mathrm{B}_0\mathrm{A}_0^{-1} = \sigma_0^2\mathrm{A}_0^{-1} \\
\mathrm{V}ar(\hat{\theta}) &= \sigma_0^2\mathrm{A}_0^{-1} \\
\hat{\mathrm{V}ar}(\hat{\theta}) &= \hat{\sigma}^2(\sum_{i=1}^n \hat{\mathrm{H}}_i)^{-1} \text{ or } \hat{\sigma}^2(\sum_{i=1}^n \hat{\mathrm{A}}_i)^{-1}
\end{aligned}$$

其中，$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^n \hat{u}_i^2$

Under Assumption 3.1.1,Assumption 3.1.2,Assumption 3.3.1, we FINNALLY have

$$\hat{\mathrm{V}ar}(\hat{\theta}) = \hat{\sigma}^2(\sum_{i=1}^n \nabla_\theta m(x_i,\hat{\theta})'\nabla_\theta m(x_i,\hat{\theta}))^{-1}$$

Example. If $y_i = exp(x_i\theta) + u_i$, we can get

$$\hat{\mathrm{V}ar}(\hat{\theta}) = \hat{\sigma}^2(\sum_{i=1}^n exp(2x_i\hat{\theta})x_i'x_i)^{-1}$$

3.3.2 Two Step M-Estimation

Recall Equation 3.8, final form of estimator variance is decided by the discussion on the value of F_0. Here we follow the discussion to see the final version of estimation variance.

- if $F_0 = 0$.

If $F_0 = E[\nabla_\gamma S(W, \theta_0, \gamma^*)] = 0$

$$\hat{Var}(\hat{\theta}) = \begin{cases} (\sum_{i=1}^n \hat{H}_i)^{-1} (\sum_{i=1}^n \hat{S}_i \hat{S}_i') (\sum_{i=1}^n \hat{H}_i)^{-1} \\ (\sum_{i=1}^n \hat{A}_i)^{-1} (\sum_{i=1}^n \hat{S}_i \hat{S}_i') (\sum_{i=1}^n \hat{A}_i)^{-1} \end{cases} \tag{3.12}$$

where $\hat{S}_i, \hat{H}_i, \hat{A}_i$ depend on $\hat{\gamma}_i, \hat{\theta}_i$.

Taking $y = m(x, \theta_0) + u$ as example, under Assumption 3.2.1 and Assumption 3.2.2, we have

$$\begin{aligned} E[q(W, \theta, \gamma^*)] &= E\left[\frac{1}{2}\frac{(y - m(X, \theta))^2}{h(X, \gamma^*)}\right] \\ S(W, \theta_0, \gamma^*) &= \frac{\partial q(W, \theta_0, \gamma^*)}{\partial \theta} \\ &= -\frac{\nabla_\theta m(X, \theta_0)'(y - m(X, \theta_0))}{h(X, \gamma^*)} = -\frac{\nabla_\theta m' u}{h} \\ H(W, \theta_0, \gamma^*) &= \frac{\partial^2 q}{\partial \theta \partial \theta'} = \nabla_\theta m' \nabla_\theta^2 m (y - m)/h \text{(Liu: seemingly wrong form.)} \\ E[\nabla_\gamma S(W, \theta_0, \gamma^*)] &= 0 \\ \hat{Var}(\hat{\theta}) &= (\sum_{i=1}^n \nabla_\theta m_i' \nabla_\theta m_i / h_i)^{-1} (\sum_{i=1}^n \nabla_\theta \hat{m}_i' \hat{u}_i^2 \nabla_\theta \hat{m}_i / \hat{h}_i) (\sum_{i=1}^n \nabla_\theta m_i' \nabla_\theta m_i / h_i)^{-1} \end{aligned}$$

- if $F_0 \neq 0$.

与Assumption 3.3.1，我们对模型随机扰动项的方差结构如下假设以简化分析：

Homoscedasticity

Assumption 3.3.2. *For $y = m(X, \theta_0) + u$, we assume that*

$$Var(y|X) = \sigma_0^2 h(X, \gamma_0)$$

基于Assumption 3.3.2和Equation 3.10，可得

$$
\begin{aligned}
\mathrm{B}_0 &= \sigma_0^2 \mathrm{E}\left[\frac{\nabla_\theta m(x,\theta_0)'\nabla_\theta m(x,\theta_0)}{h(x,\gamma_0)}\right] = \sigma_0^2 \mathrm{A}_0 \\
\mathrm{A}_0 &= \mathrm{E}\left[\frac{\nabla_\theta m(x,\theta_0)'\nabla_\theta m(x,\theta_0)}{h(x,\gamma_0)}\right] \\
\hat{Var}(\hat{\theta}) &= \hat{\sigma}^2 \left(\sum_{i=1}^{n} \frac{\nabla_\theta \hat{m}(x_i,\theta_0)'\nabla_\theta \hat{m}(x_i,\theta_0)}{h(x_i,\gamma_0)}\right)^{-1} \\
\hat{\sigma}^2 &= \frac{1}{n-k}\sum_{i=1}^{n}\left(\frac{\hat{u}_i^2}{\sqrt{\hat{h}(x_i,\gamma_0)}}\right)^2
\end{aligned}
$$

由于$\mathrm{F}_0 \neq 0$，有$\mathrm{E}[\nabla_\gamma \mathrm{S}(\mathrm{W},\theta_0,\gamma^*)] \neq 0$，故，

$$
\hat{Var}(\hat{\theta}) = \begin{cases} (\sum_{i=1}^{n}\hat{\mathrm{H}}_i)^{-1}(\sum_{i=1}^{n}\hat{g}_i\hat{g}_i')(\sum_{i=1}^{n}\hat{\mathrm{H}}_i)^{-1} \\ (\sum_{i=1}^{n}\hat{\mathrm{A}}_i)^{-1}(\sum_{i=1}^{n}\hat{g}_i\hat{g}_i')(\sum_{i=1}^{n}\hat{\mathrm{A}}_i)^{-1} \end{cases}
$$

其中

$$
\begin{aligned}
\hat{g}_i &= \hat{\mathrm{S}}_i + \hat{\mathrm{F}}_i + \hat{r}_i \\
\hat{\mathrm{F}}_i &= \frac{1}{n}\sum_{i=1}^{n}\nabla_\gamma \mathrm{S}_i(\hat{\theta},\hat{\gamma})
\end{aligned}
$$

3.4 Numerical Optimization

由于涉及到大规模的矩阵求导与求逆，解析解（close form）在最优化过程中通常并不存在，本节将介绍常用的数值解解法。

3.4.1 Newton-Raphson Method

如Figure 3.1所示，直观而言，Newton-Raphson Method的核心思想是从初始值$x^{(0)}$开始、沿着梯度的方向，逼近梯度为0的点（FOC=0），但同样的，该方法在不同的初始值设定下容易收敛到局部解，并且在二阶导不存在的情况下可能无法工作（无法判断是否为最大或最小值）。 回到我们学习的内容，

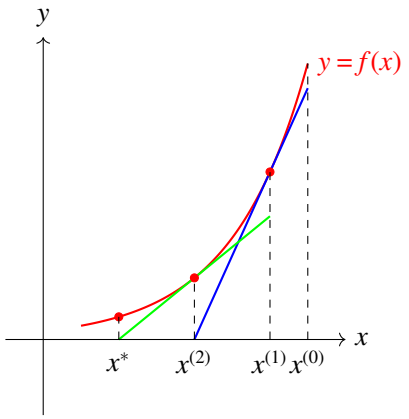

Figure 3.1 Newton-Raphson Method Plot

以FOC = 0为切入口，可以得到递推公式如下：

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{S}(\mathrm{W}_i, \hat{\theta}) &= 0 \\
\sum_{i=1}^{n} \mathrm{S}_i(\theta^{\{g+1\}}) &= \sum_{i=1}^{n} \mathrm{S}_i(\theta^{\{g\}}) + [\sum_{i=1}^{n} \mathrm{H}_i(\theta^{\{g\}})](\theta^{\{g+1\}} - \theta^{\{g\}}) + r^{\{g\}} \\
\text{Let } \sum_{i=1}^{n} \mathrm{S}_i(\theta^{\{g+1\}}) &= 0,\ r^{\{g\}} = 0 \\
\theta^{\{g+1\}} &= \theta^{\{g\}} - [\sum_{i=1}^{n} \mathrm{H}_i(\theta^{\{g\}})]^{-1} [\sum_{i=1}^{n} \mathrm{S}_i(\theta^{\{g\}})] \quad \text{iterative method}
\end{aligned}$$

Stopping rule. $\theta^{\{g+1\}}$ is very close to $\theta^{\{g\}}$

1. $|\theta_j^{\{g+1\}} - \theta_j^{\{g\}}|$, for $j = 1, 2, \ldots, k$ be smaller than some small constant (precision).
2. $\frac{|\theta_j^{\{g+1\}} - \theta_j^{\{g\}}|}{\theta_j^{\{g\}}|}$largest percentage change is parameter values be smaller than some small constant.
3. quadratic form

$$[\sum_{i=1}^{n} \mathrm{S}_i(\theta^{\{g\}})]' [\sum_{i=1}^{n} \mathrm{H}_i(\theta^{\{g\}})] [\sum_{i=1}^{n} \mathrm{S}_i(\theta^{\{g\}})] \tag{3.13}$$

Drawbacks.

- H_i requires second derivative;
- At particular value of $\hat{\theta}$, H_i might not be positive definite.

3.4.2 Berndt, Hall, Hall, and Hausman Method

To overcome the issue with Hessian matrics in Newton-Raphson method, the BHHH algorithm is a numerical optimization algorithm similar to the Newton–Raphson algorithm, but it replaces the observed

negative Hessian matrix with the outer product of the gradient. This approximation is based on the information matrix equality and therefore only valid while maximizing a likelihood function.

$$\theta^{\{g+1\}} = \theta^{\{g\}} - \left[\sum_{i=1}^{n} S_i(\theta^{\{g\}}) S_i(\theta^{\{g\}})'\right]^{-1} \left[\sum_{i=1}^{n} S_i(\theta^{\{g\}})\right] \tag{3.14}$$

相比于Newton-Raphson Method，BHHH规避了对目标函数的二次求导，同时保证了搜寻目标值时不会找错方向。

NL Model $B_0 = \sigma_0^2 A_0$ GIME(Generalized Information Matrix Equality)

$$\sum_{i=1}^{n} S_i(\theta^{\{g\}}) S_i(\theta^{\{g\}})' = \sigma_0^2 \sum_{i=1}^{n} H_i(\theta^{\{g\}}) \tag{3.15}$$

Stopping rule. $\theta^{\{g+1\}}$ is very close to $\theta^{\{g\}}$

1. $|\theta_j^{\{g+1\}} - \theta_j^{\{g\}}|$, for $j = 1, 2, \ldots, k$ be smaller than some small constant (precision).
2. $\frac{|\theta_j^{\{g+1\}} - \theta_j^{\{g\}}|}{\theta_j^{\{g\}}|}$largest percentage change is parameter values be smaller than some small constant.
3.

$$[\sum_{i=1}^{n} S_i(\theta^{\{g\}})]'[\sum_{i=1}^{n} S_i(\theta^{\{g\}}) S_i(\theta^{\{g\}})']^{-1}[\sum_{i=1}^{n} S_i(\theta^{\{g\}})] = nR^2 \sim \chi^2(k) \tag{3.16}$$

The blue part in Equation 3.14 is the coefficients from regression of vector 1 on $S_i(\theta^{\{g\}})'$. Hence, the blue R^2 above is uncentered R^2. In summary, the test in Equation 3.16 is the same as $nR^2 = \frac{SSE}{SST} * n$.

3.4.3 Gauss-Newton Method

For generalized Gauss-Newton method, we could have

$$\theta^{\{g+1\}} = \theta^{\{g\}} - r[\sum_{i=1} n A_i(\theta^{\{g\}})]^{-1}[\sum_{i=1}^{n} S_i(\theta^{\{g\}})]$$

其中$A(X_i, \theta_0) = E[H(W_i, \theta_0)|X_i]$。

下面以具体的形式来考虑该算法。考虑$y = m(X_i, \theta_0) + u_i$

$$\theta^{\{g+1\}} = \theta^{\{g\}} - \left[\sum_{i=1}^{n} \nabla_\theta m(X_i, \theta^{\{g\}})' \sum_{i=1}^{n} \nabla_\theta m(X_i, \theta^{\{g\}})]^{-1}[\sum_{i=1}^{n} \nabla_\theta m(X_i, \theta^{\{g\}})' u_i^{\{g\}}\right]$$

Like the regression of 1 on $S_i(\theta^{\{g\}})'$，here we reg $u_i^{\{g\}}$ on $\nabla_\theta m(X_i, \theta^{\{g\}})$ to get the blue and test $nR^2 \sim \chi^2(k)$.

具体而言，考虑$y = m(\mathrm{X}_i, \theta_0) + u_i$，Taylor Expansion would yield

$$\begin{aligned} m(\mathrm{X}, \theta^{\{2\}}) &\approx m(\mathrm{X}, \theta^{\{1\}}) + \nabla_\theta m(\mathrm{X}, \theta^{\{1\}})(\theta^{\{2\}} - \theta^{\{1\}}) \\ y - m(\mathrm{X}, \theta^{\{1\}}) &\approx \nabla_\theta m(\mathrm{X}, \theta^{\{1\}})(\theta^{\{2\}} - \theta^{\{1\}}) + y - m(\mathrm{X}, \theta^{\{2\}}) \end{aligned}$$

The second line above is the core regression we care. For the stopping value,

$$\begin{gathered} \mathrm{H}_0 : b = \theta^{\{2\}} - \theta^{\{1\}} = 0 \\ \text{if } b \neq 0, \theta^{\{2\}} = b + \theta^{\{1\}} \\ \vdots \\ \text{until } \theta^{\{i+1\}} - \theta^{\{i\}} = 0 \end{gathered}$$

更具体而言，考虑$y_i = \beta_1 x_{1i} + \beta_2 x_{2i}^{\beta_3} + u_i$，此时有

$$\begin{gathered} \nabla_\beta m(x_i, \beta) = (x_1, x_2^{\beta_3}, \beta_2 x_2^{\beta_2} \, ln x_2) \\ \text{initial value: } (\beta_1, \beta_2, \beta_3) = (1, 1, 1) \end{gathered}$$

reg $y - x_1 - x_2$ on x_1, x_2, $x_2 \, ln x_2$

$$b = \begin{pmatrix} \beta_1^{\{2\}} - \beta_1^{\{1\}} \\ \beta_2^{\{2\}} - \beta_2^{\{1\}} \\ \beta_3^{\{2\}} - \beta_3^{\{1\}} \end{pmatrix} \stackrel{?}{=} 0 \tag{3.17}$$

检验$\mathrm{H}_0 : b = 0$是否成立，若不成立则将$(\beta_1^{\{n\}}, \beta_2^{\{n\}}, \beta_3^{\{n\}})$代入原方程，递推进行下一次迭代、然后继续检验，直到不能拒绝原假设。

在此问题中，最后的估计结果如何检验单个系数，比如$\mathrm{H}_0 : \beta_3 = 1$。检验方法包括：

- 1. 用M-estimation 估计 β_3：已知$\hat{\beta} \sim \mathrm{N}(\beta_0, \sigma_0^2 \mathrm{A}_0^{-1}/n)$

$$\hat{Var}(\hat{\beta}) = \hat{\sigma}^2 (\sum_{i=1}^{n} \nabla_\beta m(x_i, \hat{\beta})' \nabla_\beta m(x_i, \hat{\beta}))^{-1}, \ \hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{u}_i^2$$

$$t = \frac{\hat{\beta}_1^{\mathrm{M}} - 1}{\mathrm{SE}(\hat{\beta}_3)}$$

- 2. 用t, F, Wald, LM, LR等方法检验。

除此之外，还有另一种检验的方法$\mathrm{H}_0 : \beta_3 = 1$。We first impose this restriction into the model, and run

$y = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + u$ to get the esitmated residuals.

$$\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, 1)$$

$$\tilde{u} = y - \tilde{\beta}_1 x_1 - \tilde{\beta}_2 x_2$$

$$\nabla_\beta m(x, \beta) = (x_1, x_2, \tilde{\beta}_2 x_2 \, ln \, x_2)$$

Second, we could reg $\tilde{u}$ on $\nabla_\beta m(x, \beta)$ which means a regression like $\tilde{u} = \theta_1 x_1 + \theta_2 x_2 + \theta_3 \tilde{\beta}_2 x_2 \, ln \, x_2$.

If $H_0 : \beta_3 = 1$ holds, the initial value for β_3 is 1 then the iteration would stop. Hence we could observe $\theta_3 = 0$. We transfer the $H_0 : \beta_3 = 1$ into $\theta_3 = 0$.

核心思想：将$\tilde{\beta}$作为某一次迭代过程的得到的值，则该回归方程的系数表示两次迭代的差，所以$\alpha_3 = 0$表示$\beta_3^{\{i+1\}} = \beta_3^{\{i\}}$。

3.5 Quantile Regression

本质上分位数回归依然属于M-estimation的内容，逻辑上其应该放在GMM之后和M-estimation之前来着，但不知道为啥会放在M-estimation之后来讲这部分内容，为延续老师讲课的脉络，本节将介绍分位数回归的内容。

Motivation. From Mosteller and Tukey (1977):

> *What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.*

Let y_i denote a random draw from a population. $q(\tau)$ is a τ-th quantitle , where $0 < \tau < 1$, if $P(y_i \leq q(\tau)) = \tau$ and $P(y_i \geq q(\tau)) = 1 - \tau$. Let $Q_\tau(y_i)$: τ-th quantile of y_i, then we try to model

$$Q_\tau(y_i | x_i) = \beta_0(\tau) + x_i \beta_1(\tau)$$

What we do here is to project the $y's$ $\tau - quantile$ on x. The key is to find an appropriate objective function. To see why does the *objective function* work here, we consider the OLS and Least Absolute Deviation (LAD) to illustrate how to model quantile by objective function.

OLS. For OLS, $\min \sum_{i=1}^n (y_i - q)^2$, where FOC $= \sum_{i=1}^n 2(y_i - q)(-1) = 0$. We can get $q = \frac{1}{n} \sum_{i=1}^n y_i = Ey_i$ *as* $n \to \infty$.

LAD. For LAD, $\min_q \sum_{i=1}^n |y_i - q|$, where $|y_i - q| = 1_{\{(y_i \geq q)\}}(y_i - q) + 1_{\{(y_i < q)\}}(q - y_i)$.

$$\begin{aligned}
\text{FOC} &= \sum_{i=1}^{n}(-1)\left[1_{\{(y_i \geq q)\}} + 1_{\{(y_i < q)\}}\right] \\
&= \sum_{i=1}^{n}(-1)[1 - 1_{\{(y_i < q)\}}] + \sum_{i=1}^{n} 1_{\{(y_i < q)\}} \\
&= -n + 2\sum_{i=1}^{n} 1_{\{(y_i < q)\}} = 0 \\
\Longrightarrow \frac{1}{2} &= \frac{1}{n}\sum_{i=1}^{n} 1_{\{(y_i < q)\}} \\
\Longrightarrow \mathrm{E}\left[1_{\{(y_i < q)\}}\right] &= \frac{1}{2} \\
\Longrightarrow \mathrm{P}(y_i < q) &= \frac{1}{2}
\end{aligned}$$

Finnally, from LAD we can get $q = \text{Median}(y_i)$. Hence, the discussion above implies that abosulate value function as objective function helps to treat quantile of y as the center in minimization.

除对于其他quantile，我们可以通过给绝对值> 0和< 0分配不同的权重而实现对不同分位数的回归。

$$\text{Asymetric error loss: } \mathrm{L}(e) = \begin{cases} \mathrm{E}[(1-\tau)|e|] & e < 0 \\ \mathrm{E}[\tau|e|] & e \geq 0 \end{cases} \tag{3.18}$$

其中τ为y的分位数。此时，目标函数为

$$\min_{q \in \mathrm{R}} \mathrm{L}|y - q| = \min \mathrm{E}\left\{\left[\tau 1_{\{y-q \geq 0\}} + (1-\tau)1_{\{(y-q<0)\}}\right]|y_i - q|\right\}$$

为表达简洁，不妨定义对勾函数（check function）如下：

$$\rho_\tau(x) = \left[\tau 1_{\{x \geq 0\}} + (1-\tau)1_{\{x<0\}}\right]|x|$$

故，分位数回归最后的目标函数变为：

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - x_i\beta)$$

$$\text{in M-estimation } \hat{\theta}(\tau) = (\hat{\alpha}(\tau), \hat{\beta}(\tau)') \to \theta_0(\tau) = (\alpha_0(\tau), \beta_0(\tau)')$$

题外话：**LAD or OLS.** As implied by Figure 3.2,

- OLS is *sensitive* to changes in outlier
- LAD is *insensitive* to changes in outlier

Besides robustness to the outlier, quantile regressions could provide more information about the distributions. Hence, to figure out which assumptions guarantee these for LAD is important, especially, under which assumptions LAD and OLS are equivalent.

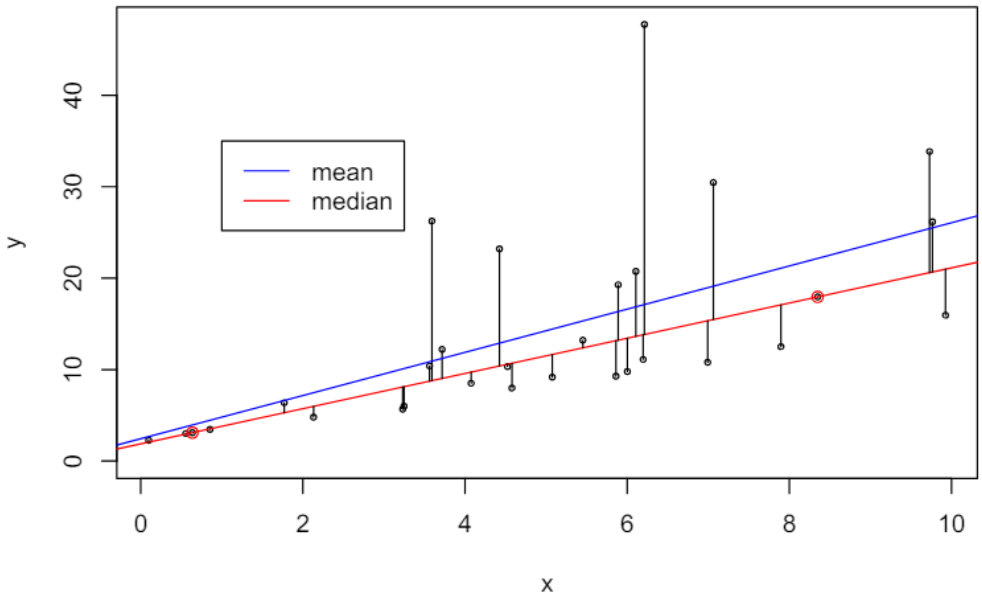

Figure 3.2 Mean and Median Regression

Consider the basic form of regression:

$$y = \alpha_0 + x\beta + u$$

Case 1: Distribution of u is symmetric about zero. Let D denote the distribution, we could get $\mathrm{D}(y|x) = \alpha_0 + x\beta + \mathrm{D}(u|x)$, where $\mathrm{E}(u|x) = \mathrm{M}ed(u|x) = 0$. Hence, $\mathrm{E}(y|x) = \alpha_0 + x\beta_0 = \mathrm{M}ed(y|x)$

Case 2: u and x are independent. That is, $\mathrm{D}(u_i|x_i) = \mathrm{D}(u_i)$ and $\mathrm{E}[u_i] = 0$. $\mathrm{E}(y_i|x_i) = \alpha_0 + x_i\beta_0$. $\mathrm{M}ed(y_i|x_i) = \alpha_0 + x_i\beta_0 + \mathrm{M}ed(u_i|x_i)$. Let $\mathrm{M}ed(u_i) = \eta_0$, hence $\mathrm{M}ed(y_i|x_i) = (\alpha_0 + \eta_0) + x_i\beta_0$. In this case, only are the slopes from OLS and LAD the same.

由于在M-estimation in LAD中通常并不会假设分布的情况，所以Case 1和2基本都不被满足，但有一种不满足假设，却仍可以用LAD代替OLS的方法。考虑收入y，通常是一个右偏的分布，因此取$\ln y$，有模型为$\ln y_i = \alpha_0 + x_i\beta_0 + u_i$。下面我们将分别考察两个Case下OLS与LAD是否等价。

Case 1 and OLS.

$$\begin{aligned} \mathrm{E}[\ln y_i|x_i] &= \alpha_0 + x_i\beta_0 \\ e^{\ln y_i} &= e^{(\alpha_0+x_i\beta_0+u_i)} = y_i \\ \mathrm{E}[y_i|x_i] &= e^{(\alpha_0+x_i\beta_0)}\mathrm{E}[e^{u_i}|x_i] \end{aligned}$$

Case 1 and LAD.

$$\begin{aligned} \mathrm{M}ed(\ln y_i|x_i) &= \alpha_0 + x_i\beta_0 \\ e^{\ln y_i} &= e^{(\alpha_0+x_i\beta_0+u_i)} = y_i \\ \mathrm{M}ed[y_i|x_i] &= e^{(\alpha_0+x_i\beta_0)}\mathrm{M}ed[e^{u_i}|x_i] \\ &= e^{(\alpha_0+x_i\beta_0)}e^{\mathrm{M}ed[u_i|x_i]} \ (e^x\text{单调}) \\ &= e^{(\alpha_0+x_i\beta_0)} \neq \mathrm{E}[y_i|x_i] \end{aligned}$$

Case 2 and OLS.

$$\mathrm{E}[\ln y_i|x_i] = \alpha_0 + x_i\beta_0 + \cancel{\mathrm{E}[u_i|x_i]}$$

$$\mathrm{E}(y_i|x_i) = e^{\alpha_0+x_i\beta_0}\mathrm{E}(e^{u_i}|\cancel{x_i})$$

不需要计算$\mathrm{E}(e^{u_i}|x_i)$，只需要计算$\mathrm{E}(e^{u_i})(\mathrm{E}(e^{\hat{u}_i}))$ *Case 2 and LAD.*

$\mathrm{M}ed(\ln y_i|x_i) = \alpha_0 + x_i\beta_0 + \mathrm{M}ed(u_i|x_i) = (\alpha_0 + \eta_0) + x_i\beta_0)$

$\alpha_0 + \eta_0$是一起估计出来的。

$$u_i = \mathrm{M}ed(u_i) + \tilde{u}_i = \eta_0 + \tilde{u}_i$$

$$u_i - \mathrm{M}ed(u_i) = \tilde{u}_i \quad \mathrm{M}ed(\tilde{u}_i) = 0$$

$$e^{u_i} = e^{\eta_0}e^{\tilde{u}_i}$$

where $\tilde{u}_i$ is the error term in $\ln y_i = \alpha_0 + \eta_0 + x_i\beta_0 + \tilde{u}_i$ by using LAD

$$\mathrm{E}e^{u_i} = e^{\eta_0}\mathrm{E}e^{\tilde{u}_i}$$

$$e^{\alpha_0+x_i\beta_0}\mathrm{E}(e^{u_i}) = e^{\alpha_0+x_i\beta_0}e^{\eta_0}\mathrm{E}e^{\tilde{u}_i}$$

虽然无法完全规避计算均值的影响，但$\mathrm{E}(y_i|x_i) = e^{\alpha_0+x_i\beta_0}e^{\eta_0}\mathrm{E}e^{\tilde{u}_i}$可以用LAD计算得到。

上述讨论表明试图take log 并完全用LAD的方法，规避E(.)是做不到的

OLS的优点是可以使用Law of Iterated Expectation $\mathrm{E}(x_i y_i) = \mathrm{E}[x_i \mathrm{E}(y_i|x_i)]$ 但是 $\mathrm{M}ed(x_i y_i) = \mathrm{M}ed[x_i \mathrm{M}ed(y_i|x_i)]$,其次M*ed*不能进行线性计算。

考虑$y_i = a_i + x_i b_i$ a_i, b_i are random and independent of x_i

$$\begin{aligned}\mathrm{E}(y_i|x_i) &= \mathrm{E}(a_i|x_i) + x_i \mathrm{E}(b_i|x_i) \\ &= \alpha_0 + x_i \beta_0 \quad \text{OLS average partial effect}\end{aligned}$$

$$\begin{aligned}\mathrm{M}ed(y_i|x_i) &= \mathrm{M}ed(a_i|x_i) + x_i \mathrm{M}ed(b_i|x_i) \\ &= \mathrm{M}ed(a_i) + x_i \mathrm{M}ed(b_i)\end{aligned}$$

y_i	a_i	x_i	b_i
3.1	2.1	1	1
4	2	1	2
2.1	0	1	1.1

L*eftside* $= 3.1 \neq$ R*ightside* $= 4$

题外话结束. 回到分位数回归的参数估计和分布推导，采用M估计的思路，我们有

$$y_i = x_i \theta_0 + u_i, \quad \mathrm{Q}_\tau(u_i|x_i) = 0$$

$$q(w_i, \theta) = \tau 1_{\{y_i - x_i\theta \geq 0\}}(y - x_i\theta) - (1-\tau)1_{\{(y_i - x_i\theta < 0\}}(y_i - x_i\theta)$$

在求解FOC过程中，由于绝对值函数的顶点不可导，之前求解的过程中，在尖点的导数是错误的。但是$0 = y_i - x_i\theta_0 = u_i$, $\mathrm{P}(u_i = 0) = 0$的点出现的概率是0，因此$\hat{\theta} \to \theta_0$ as $n \to \infty$。为解决该问题，考虑对概率加权平均的结果（期望）求导以规避绝对值函数顶点不可导的问题。若直接求导，则有

$$\mathrm{S}_i(\theta) = -x_i' \{\tau 1_{\{(y_i - x_i\theta \geq 0)\}} - (1-\tau)1_{\{y_i - x_i\theta < 0\}}\}$$

$$\mathrm{H}(x_i, \theta) = \frac{\partial \mathrm{S}_i}{\partial \theta'} = 0$$

此时，由于部分点不可导，有A_i不可逆，在之前的求解过程中包括$\frac{\partial \mathrm{E}q}{\partial \theta} = \mathrm{E}\frac{\partial q}{\partial \theta}$当$q$连续时，$\mathrm{E}, \partial$可交换，但在这里有S不连续，所以不能交换。此时，考虑利用迭代期望律、直接计算$\mathrm{E}[\mathrm{S}_i(\theta)]$。首先，计

算$\mathrm{E}[\mathrm{S}_i(\theta)|x_i]$

$$\begin{aligned}
\mathrm{E}[\mathrm{S}_i(\theta)|x_i] &= -x_i'\{\tau \mathrm{P}(y_i - x_i\theta \geq 0|x_i) - (1-\tau)\mathrm{P}(y_i - x_i\theta < 0|x_i)\} \\
&= -x_i'\{\tau \mathrm{P}(u_i \geq x_i(\theta-\theta_0)|x_i) - (1-\tau)\mathrm{P}(u_i < x_i(\theta-\theta_0)|x_i)\} \\
&= -x_i'\{\tau[1-\mathrm{F}_u(x_i(\theta-\theta_0)|x_i)] - (1-\tau)\mathrm{F}_u(x_i(\theta-\theta_0)|x_i)]\} \\
&= -x_i'\{\tau - \mathrm{F}_u(x_i(\theta-\theta_0)|x_i)\}
\end{aligned}$$

其中F为累积分布函数。因为F_u连续，所以有$\mathrm{E}\frac{\partial \mathrm{E}(\cdot|x)}{\partial\theta'} = \frac{\partial \mathrm{E}[\mathrm{E}(\cdot|x)]}{\partial\theta'} = \frac{\partial \mathrm{E}(\cdot)}{\partial\theta'}$

$$\begin{aligned}
\frac{\partial \mathrm{E}[\mathrm{S}_i(\theta|x_i)]}{\partial\theta'} &= f_u(x_i(\theta-\theta_0)|x_i)x_i'x_i \\
\mathrm{A}_0 &= \mathrm{A}(\theta_0) = \mathrm{E}[f_u(0|x_i)x_i'x_i] \text{ (蓝色部分的估计见后)} \\
\sqrt{n}(\hat{\theta}-\theta_0) &= \mathrm{A}_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{S}_i(\theta_0) + o_p(1) \sim \mathrm{N}(0, \mathrm{A}_0^{-1}\mathrm{B}_0\mathrm{A}_0^{-1})
\end{aligned}$$

其中f为概率密度函数。而$\sqrt{n}(\hat{\theta}-\theta_0)$方差中的$\mathrm{B}_0$计算如下：

$$\begin{aligned}
\mathrm{B}_0 &= \mathrm{E}[\mathrm{S}_i(\theta_0)\mathrm{S}_i(\theta_0)'] \\
&= \mathrm{E}\left[x_i'x'(\tau^2 1_{\{y \geq x_i\theta_0\}} + (1-\tau)^2 1_{\{(y_i < x_i\theta_0\}}\right] \\
&= \mathrm{E}\left[x_i'x_i(\tau^2(1-\tau) + (1-\tau)^2\tau)\right] \\
&= \tau(1-\tau)\mathrm{E}\left[x_i'x_i\right] \\
\hat{\mathrm{B}}_0 &= \tau(1-\tau)\frac{1}{n}\sum_{i=1}^{n}x_i'x'
\end{aligned}$$

而验证一阶条件成立，$\mathrm{E}[\mathrm{S}_i(\theta_0)|x_i] = -x_i[\tau - \mathrm{F}_u(0|x_i)]$，因为$\mathrm{F}_u = \mathrm{P}(u_i < x_i(\theta-\theta_0)|x_i), \mathrm{F}_u(0) = \mathrm{P}(y_i - x_i\theta_0 < 0|x_i)$，$x_i\theta_0$ is the τ percentile of y，所以$\mathrm{P}(y_i < x_i\theta_0|x_i) = \tau$,即满足$\mathrm{E}[\mathrm{S}_i(\theta_0)] = \mathrm{E}[\mathrm{E}[\mathrm{S}_i(\theta_0)|x_i]] = -x_i[\tau - \tau] = 0$。

最后，我们仅需估计出f_u，即可得到最后的分布形式，根据导数定义，有

$$\begin{aligned}
f_u(0|x_i) &\approx \frac{[\mathrm{F}_u(h|x_i) - \mathrm{F}_u(-h|x_i)]}{[h-(-h)]} \\
&= \mathrm{P}(-h \leq u_i \leq h|x)_i)/2h \\
&= \mathrm{P}(|u_i| \leq h|x_i)/2h \\
&= \mathrm{E}[1_{\{|u_i| \leq h\}}|x_i]/2h
\end{aligned}$$

So, recall the expression of A_0

$$\begin{aligned} A_0 &= E\left[f_u(0|x_i)x_i'x_i\right] \\ &= E\left[E\left[1_{\{|u_i|\le h\}}|x_i\right]x_i'x_i\right]/2h \\ &= \frac{1}{2h}E\left[1_{\{|u_i|\le h\}}x_i'x_i\right] \\ \hat{A}_0 &= \frac{1}{2nh}\sum_{i=1}^{n}1_{\{|\hat{u}_i|\le h\}}x_i'x_i \end{aligned}$$

where, h here is called bandwith or smoothing parameter and $\hat{u}_i(\tau)$ is estimated from M-estimation. 利用非参估计，无需假设$pdf = f(x)$。

第四章　面板数据

4.1　时序数据

时间序列的最终目标是通过建模来拟合真实数据，通常的做法是先介绍一些数据的基本特征事实（例如金融中收益率序列的波动率聚集等），然后尝试用所构建的模型来拟合或解释这些事实。一般的时间序列建模过程为：

Step 1 假定一个probability model来表示时间序列数据

Step 2 估计模型的参数

Step 3 时间序列关注模型的R^2（截面数据通常不需要做预测，所以只关注变量之间的因果关系，不要求高R^2）

Step 4 用模型解释数据，帮助我们加深对数据的理解

Step 5 预测

4.1.1　基本概念

在正式开始内容之前，我们首先定义一系列基本术语：

Definition 4.1.1. *Strictly Stationary (Distribution Stationary)*

A stochastic process is strickly stationary if the distribution of $(y_{t_1},\ldots,y_{t_k})$ *is the same as that of* $(y_{t_{1+h}},\ldots,y_{t_{k+h}})$, $\forall(t_1,\ldots,t_k)$ *and* $k,h=1,2,3,\ldots$.

Definition 4.1.2. *Weekly Stationary (Covariance Stationary)*

A stochastic process ($\{x_t\}$) *is strickly stationary if*

$$\begin{cases} \mathrm{E}(x_t)=\mu \\ \mathrm{V}ar(x_t)=\gamma(0)<\infty \\ \mathrm{C}ov(x_{t+h},x_t)=\gamma(h)<\infty,\forall h=\pm1\pm2,\ldots \end{cases}$$

Compared to strictly stationary, here we just require the first- and second-order moment of these distributions are the same (if the second-moment exists).

Definition 4.1.3. *White Noise*

x_t *is white noise if i)* $\mathrm{E}x_t = 0$*; ii)* $\mathrm{E}x_t^2 = \sigma^2$*; iii)* $\mathrm{E}x_t x_s = 0 \quad \forall s \neq t$.

We will introduce some common time series to enhance our understanding on these new concepts.

Example 1: Tread Stationary. For a series $y_t = \alpha + \beta t + z_t$, $z_t \sim w.n.(0, \sigma^2)$. Because $\mathrm{E}[y_t] = \alpha + \beta t$, it's not stationary.

Example 2: Random Walk. For a series $y_t = y_{t-1} + z_t$, $z_t \sim w.n.(0, \sigma^2)$ then $y_t = y_{t-2} + z_{t-1} + z_t = z_1 + z_2 + \cdots + z_t$ (assume $y_0 = 0$).

Because $\mathrm{E}[y_t] = 0$ and $\mathrm{V}ar(y_t) = t\sigma^2$, y_t is not stationary.

Example 3: Random Walk with drift For a series $y_t = \mu + y_{t-1} + z_t = z_1 + z_2 + \cdots + z_t + t\mu$

Because $\mathrm{E}y_t = t\mu$ and $\mathrm{V}ar(y_t) = t\sigma^2$, y_t is not stationary. Besides, we will introudce some common tools, would help to judge the characteristics of time series.

Definition 4.1.4. *ACVF (Auto CoVariance Function)*

$\gamma(h) = \mathrm{C}ov(y_{t+h}, y_t)$, $\forall h = 0, \pm 1, \pm 2, \ldots$

$\gamma(0) = \mathrm{V}ar(y_t)$

Definition 4.1.5. *ACF (Auto Correlation Function)*

$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \mathrm{C}orr(y_{t+h}, y_t)$ $\rho(0) = 1$

$\gamma(0) = \mathrm{V}ar(y_t)$

Definition 4.1.6. *PACF (Partial Auto Correlation Function)*

$$\rho^*(h) = \mathrm{C}orr\left(y_t - \mathrm{E}\left[y_t | y_{t-1}, \ldots, y_{t-h+1}\right], y_{t-h} - \mathrm{E}\left[y_{t-h} | y_{t-1}, \ldots, y_{t-h+1}\right]\right)$$

Intuitively, PACF focuses on the partial relations, which means the relations condition on the other determinants. The focus of PACF is consistent with the coefficients from OLS, the representive of partial effect. Hence, we usually estimate the PACF from OLS.

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_{h-1} y_{t-h+1} + \rho^*(h) y_{t-h} + u_t$$

We quickly prove it here. First, illustrate the procedure from a bivariate regressions:

$$\begin{aligned} y &= x_1\beta_1 + x_2\beta_2 + u \\ \mathrm{M}_{x_1} y &= \underbrace{\mathrm{M}_{x_1} x_1}_{\mathrm{M}_{x_1} x_1 = 0}\beta_1 + \mathrm{M}_{x_1} x_2\beta_2 + \underbrace{\mathrm{M}_{x_1} u}_{=0} \\ \Longrightarrow \mathrm{M}_{x_1} y &= \mathrm{M}_{x_1} x_2\beta_2 + \mathrm{M}_{x_1} u \end{aligned}$$

Back to the OLS regressions on the lagged term,

$$\begin{aligned} \underbrace{y_t - \mathrm{E}[y|y_{t-1},\ldots,y_{t-h+1}]}_{\mathrm{M}y_t=(\mathrm{I}-\mathrm{P})y_t} &= (y_{t-h} - \mathrm{E}(y_{t-h}|y_{t-1},\ldots,y_{t-h+1}))\beta \\ \beta &= [(y_{t-h} - \mathrm{E}(y_{t-h}|\cdot)'(y_{t-h} - \mathrm{E}(y_{t-h}|\cdot)]^{-1}(y_{t-h} - \mathrm{E}(y_{t-h}|\cdot)'(y_t - \mathrm{E}(y_t|\cdot)) \\ &= \mathrm{V}ar^{-1}\mathrm{C}ov_{t-h} = \rho^*(h) \end{aligned}$$

因此求$\rho^*(h)$，只需要做OLS回归，如下：

$$\begin{aligned} (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{h-1}, \rho^*(h)) &= \Gamma_h^{-1}\gamma_h \\ \Gamma_h = (\gamma(i-j))_{i,j=1}^h, \gamma_h &= (\gamma(1), \gamma(2), \ldots, \gamma(h))' \\ &= \begin{pmatrix} \gamma(0) & \gamma(1) & \ldots & \gamma(h-1) \\ \gamma(1) & \gamma(0) & & \vdots \\ \vdots & & \ddots & \\ \gamma(h-1) & \gamma(h-2) & \ldots & \gamma(0) \end{pmatrix} \end{aligned}$$

For any stationary process, $|\rho(h)| \to 0$ as $h \to \infty$. Based on the characteristics in approaching to zero, we category these processes into:

1. short term memory: $\rho(h) = 0$ *if* $h > q$, where q is a finite integer;

2. medium term memory: $|\rho(h)| = \mathrm{O}(|\xi|^h)$, where $|\xi| < 1$.

3. long term memory: Not covered.

4.1.2 经典模型

（Liu: 一点点题外话，这一节内容的讲述可能会有点点乱，限于本人水平，实在很难理解老师这部分内容在按照什么逻辑讲述，通常来讲，会先从最简单MA、AR模型开始介绍，然后深入到ARMA和ARIMA模型。）

One model in time series: ARMA(p,q).

$$x_t = \overbrace{\phi_1 x_{t-1} + \cdots + \phi_p x_{t-p}}^{\text{AR}(p)} + \underbrace{z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q}}_{\text{MA}(q)} \tag{4.1}$$

where AR(p) is a p-order auto regressive process and MA(q) is a q-order moving average process.

ARMA平稳性要求AR序列特征方程的特征根>1，落在单位圆外，MA的特征根不影响ARMA的平稳性，其特征根>1表示MA可逆，转化为AR（蓝色部分会在后面逐个补充说明）。

首先，用以判断AR(p)序列是否平稳的特征函数（characteristic function）如下

$$\text{特征函数：} \phi(z) = 1 - \phi_1(z) - \cdots - \phi_p z^p$$
$$\text{特征方程：} 0 = 1 - \phi_1(z) - \cdots - \phi_p z^p$$
$$\text{特征根：} z*, \text{ the solution from特征方程}$$

如果$|z*| > 1$则该AR(p)序列平稳（特征根在单位圆之外，为什么和单位圆作比较是源于求解特征根是有一项$\frac{1}{1-z}$，这是AR序列的性质）。ARMA模型的平稳性由AR部分决定，可逆性由MA部分决定（可逆是指MA序列可以转化为AR(∞)模型）。因此只要AR部分平稳则对应的ARMA模型平稳。

其次，MA(q)的特征方程如下

$$0 = \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

如果$|z*| > 1$则该MA(q)序列可逆（特征根在单位圆之外）。

在讨论ARMA过程是short term还是medium term之前，我们先以最简单的例子来分别看ARMA最简单的组成部分的性质。

(1) AR(1): $x_t = \phi x_{t-1} + z_t$, where $\mathrm{E}[x_t] = 0$, $|\phi| < 1$ and $z_t \sim \mathrm{WN}(0, \sigma^2)$.

$$(\times x_t)\ \mathrm{E}[x_t x_t] = \mathrm{E}[\phi x_{t-1} x_t] + \mathrm{E}[z_t x_t]$$
$$\Longrightarrow \gamma(0) = \phi\gamma(1) + \sigma^2$$

where $\mathrm{E}[z_t x_t] = \phi\mathrm{E}[x_{t-1} z_t] + \mathrm{E}[z_t^2] = \sigma^2$.

$$
\begin{aligned}
(\times x_{t-1})\ \mathrm{E}[x_{t-1}x_t] &= \mathrm{E}[\phi x_{t-1}x_{t-1}] + \mathrm{E}[z_t x_{t-1}] \\
&\Longrightarrow \gamma(1) = \phi\gamma(0) \\
&\vdots \\
\gamma(h) &= \phi\gamma(h-1) = \phi^h\gamma(0) \\
\rho(h) &= \frac{\gamma(h)}{\gamma(0)} = \phi^h
\end{aligned}
$$

Hence, AR is medium-term memory.

(2) MA(1) $x_t = z_t + \theta z_{t-1},\ z_t \sim \mathrm{WN}(0, \sigma^2)$

$$
\begin{aligned}
(\times x_t)\ \mathrm{E}[x_t x_t] &= \mathrm{E}[x_t z_t] + \mathrm{E}[\theta x_t z_{t-1}] \\
\Longrightarrow \gamma(0) = \sigma^2 + \theta^2\sigma^2 (\times x_{t-1})\ \mathrm{E}[x_{t-1}x_t] &= \mathrm{E}[x_{t-1}z_t] + \mathrm{E}[\theta x_{t-1}z_{t-1}] \\
&\Longrightarrow \gamma(1) = \theta\sigma^2 \\
(\times x_{t-2})\ \mathrm{E}[x_{t-2}x_t] &= \mathrm{E}[x_{t-2}z_t] + \mathrm{E}[\theta x_{t-2}z_{t-1}] \\
&\Longrightarrow \gamma(2) = 0 \\
&\vdots \\
\gamma(h) &= 0
\end{aligned}
$$

Hence, MA is a short memory process.

$$
\rho(h) = \begin{cases} \dfrac{\theta}{1+\theta^2}, & \text{if } h = 1 \\ 0 & \text{,if } h > 1 \end{cases}
$$

(3) ARMA(1,1): $x_t = \phi x_{t-1} + z_t + \theta z_{t-1},\ z_t \sim \mathrm{WN}(0, \sigma^2)$. 根据同样在等式两边乘以滞后项的方式，可得

$$
\rho(h) = \begin{cases} \dfrac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 + \theta^2 + 2\theta\phi} & \text{if } h = 1 \\ \phi^{h-1}\dfrac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 + \theta^2 + 2\theta\phi} & \text{if } h > 1 \end{cases}
$$

推广到更一般的时序模型：AR(p), MA(q), ARMA(p, q).

- AR(p): ACF: $\rho(h) \to 0$ as $h \to \infty$, PACF: $\rho^*(h) = 0$ if $h > p$（ACF拖尾，PACF截尾）
- MA(q): ACF: $\rho(h) = 0$ if $h > p$, PACF: $\rho^*(h) \to 0$ as $h \to \infty$ （ACF截尾，PACF拖尾）

考察MA(q)的PACF时，可把MA转换为AR。以MA(1)为例，其可以表达为AR(∞), *if* $|\theta| < 1$。

$$
\begin{aligned}
x_t &= z_t - \theta z_{t-1} \\
x_{t-1} &= z_{t-1} - \theta z_{t-2} \\
&\vdots \\
\text{AR}(\infty): \ x_t &= -\theta x_{t-1} - \theta^2 x_{t-2} - \cdots - \theta^n x_{t-n}
\end{aligned}
$$

if we find AR(p) model fit data well when p is very large, we can add MA part to fit the data, vice versa.

4.1.3 参数估计

Box-Jenkins Method. Before going further, we'd like to know the basic procedure to model the time series. A popular way is called *Box-Jenkins Method*.

- Transform the data to a stationary series (by difference, detrending) and decide the (p,q) in ARMA(p,q);
- Estimate the model;
- Test whether the estimated residulas follow a WN process;
- Forecast.

Estimation. Here, we focus on the estimation of ARMA(p,q). We list what we have learned before to see which methods should we apply here.

- OLS: only work fors AR(p), not for ARMA$(1,1)$ (the residuals are correlated with y_t)
- Method of Moments: works for all but with much complexity.
- MLE: works for all. With the form of MA, z_t and y_t follow the normal distribution so we could use to contruct the likelihood function.
- NLS: Taking MA(1) as example, we can do as follows:

$$
\begin{aligned}
x_t &= \mu + z_t - \alpha_1 z_{t-1} \\
x_1 &= \mu - \alpha_1 z_0 + z_1 \\
x_2 &= \mu - \alpha_1 z_1 + z_2 \\
&= (\mu + \alpha_1 \mu) - \alpha_1 x_1 - \alpha_1^2 z_0 + z_2 \\
x_t &= (\sum_{s=0}^{t-1} \alpha_1^s)\mu - \sum_{s=1}^{t-1} \alpha_1^s x_{t-s} - \alpha^t z_0 + z_t
\end{aligned}
$$

Distribution. So far, we know the key for hypothesis test is the distribution of estimators rather than estimation procedure. We introduce some theorem to help us with the estimator distribution.

Theorem 4.1.1. ***Normality***

if $\{y_t\}_{t=1}^n$ *is an AR(p) process with* $z_t \sim iid\text{WN}(0,\sigma^2)$, $\hat{\Phi}_p$ *is the estimation of* Φ_p, *then*

$$\sqrt{n}(\hat{\Phi}_p - \Phi_p) \sim \text{N}(0, \sigma^2\Gamma_p^{-1})$$

where $\Phi_p = (\phi_1,\ \phi_2,\ \ldots, \phi_p)'$ *and* Γ_p, *the covariance matrix, is equal to* $[\gamma(i-j)]_{i,j=1}^p$.

Lemma 4.1.1. ***Normality***

If $\{y_t\}_{t=1}^n$ *is an AR(p) process with* $z_t \sim iid\text{WN}(0,\sigma^2)$ *and for* $h > p$, $\Phi_h = (\hat{\phi}_1,\ \hat{\phi}_2,\ \ldots, \hat{\phi}_h)' = \Gamma_h^{-1}\gamma_h$, *then*

$$\sqrt{n}(\hat{\Phi}_h - \Phi_h) \overset{d}{\sim} \text{N}(0, \sigma^2\Gamma_h^{-1})$$

In particular, for $h > p$, $\sqrt{n}\hat{\phi}_h \sim \text{N}(0,1)$.

Example. AR(1): $y_t = \rho y_{t-1} + u_t,\ u_t \sim \text{WN}(0,\sigma^2)$ with $\text{H}_0:\ \rho = 0$

$$\hat{\rho} = \frac{\sum_t y_{t-1}y_t}{\sum_t y_{t-1}^2} = \rho + \frac{\sum_t y_{t-1}u_t}{\sum_t y_{t-1}^2}$$
$$\sqrt{n}(\hat{\rho} - \rho) = \frac{\frac{1}{\sqrt{n}}\sum_t y_{t-1}u_t}{\frac{1}{n}\sum_t y_{t-1}^2}$$

一般情况下对分子使用CLT，对分母使用LLN，可以得到$\sqrt{n}(\hat{\rho}-\rho)$的分布，但现在由于分子和分母都不是*iid*的，不能直接得到依概率收敛到总体分布。因此考虑对分子使用鞅差中心极限定理（见后），对分母通过证明MSE收敛来证依概率收敛。

分母. MSE收敛的关键是$\text{Bias}^2 \to 0$ and $\text{Variance} \to 0$ as $n \to \infty$。

$$\begin{aligned}
\text{Bias} &= \text{E}[\frac{1}{n}\sum_t y_t^2] - \sigma_y^2 \\
&= \frac{1}{n}\sum_t \text{E}[y_t^2] - \sigma_y^2 \\
&= \frac{1}{n}n\sigma_y^2 - \sigma_y^2 = 0
\end{aligned}$$

故有$\text{B}ias = 0, \text{B}ias^2 = 0.$

Assume that $|y_t|^4 = \mathrm{C} < \infty$ (four order moments are finity).

$$\begin{aligned} \mathrm{V}ar(\frac{1}{n}\sum_t y_{t-1}^2) &= \frac{1}{n^2}\left[\sum_t \mathrm{V}ar(y_{t-1}^2) + 2\sum_t\sum_{s>t}\mathrm{C}ov(y_{t-1}^2, y_{s-1}^2)\right] \\ &= \frac{1}{n^2}[\mathrm{O}(n) + ?] \end{aligned} \tag{4.2}$$

为判断?的量级，我们首先关注$\mathrm{C}ov(y_{t-1}^2, y_{s-1}^2)$的结构，以AR(1)为例，有$y_s^2 = \rho^2 y_{s-1}^2 + u_s^2 + 2\rho y_{s-1}u_s$，可得$\mathrm{C}ov(y_s^2, y_{s-1}^2) = \mathrm{C}ov(\rho^2 y_{s-1}^2, y_{s-1}^2) + 0 = \rho^2 \mathrm{V}ar(y_{s-1}^2)$，推广可得

$$\mathrm{C}ov(y_s^2, y_{s-t}^2) = \rho^{2t}\mathrm{C}$$

将其代入原式，有

$$\begin{aligned} \mathrm{V}ar(\frac{1}{n}\sum_t y_{t-1}^2) &= \frac{1}{n^2}[n\mathrm{C} + 2\sum_t\sum_{s>t}\mathrm{C}\rho^{2(s-t)}] \\ &= \frac{1}{n^2}\left\{ n\mathrm{C} + \underset{=\mathrm{O}(n)}{\frac{\mathrm{C}\rho^2\left[(n-1) - \frac{\rho^2(1-\rho^{2(n-1)})}{1-\rho^2}\right]}{1-\rho^2}} \right\} \\ &= \mathrm{O}(\frac{1}{n}) + \mathrm{O}(\frac{1}{n}) = \mathrm{O}(\frac{1}{n}) \to 0 \end{aligned}$$

故有，$\mathrm{B}ias^2 \to 0$, $\mathrm{V}ar \to 0$ as $n \to \infty$，所以分母收敛到$\frac{1}{n}\sum_t y_{t-1}^2 \to \sigma_y^2$

分子. 在深入细节之前，我们先介绍一些关于鞅差序列的基本概念。

Definition 4.1.7. *Martingale*

A series $\{x_t\}$ *is a martingale if* $\mathrm{E}[x_t|x_{t-1}, x_{t-2}, \dots] = \mathrm{E}[x_{t-1}]$ *(for example the random walk). It means that* x_{t-1} *is the best guass for* x_t *based on current information set* I_{t-1}.

Definition 4.1.8. *Martingale Difference*

A series $\{x_t\}$ *is a martingale if* $\mathrm{E}[x_t|x_{t-1}, x_{t-2}, \dots] = 0$.

Theorem 4.1.2. ***Martingale Difference Central Limit Theory (MDCLT)***

如果x_t是一个鞅差分序列，并有$\mathrm{E}[x_t x_t'] = \Sigma$，其中$\Sigma$为一个有限正定矩阵。定义$\bar{z}_\mathrm{T} = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} z_t$，则有

$$\sqrt{\mathrm{T}}\bar{z}_\mathrm{T} \overset{d}{\sim} \mathrm{N}(0, \Sigma)$$

此时，考虑分子，分子$y_{t-1}u_t$是Martigale Difference Process，即有$\mathrm{E}[y_{t-1}u_t|\mathrm{I}_{t-1}] = y_{t-1}\mathrm{E}[u_t|\mathrm{I}_{t-1}] = 0$，$\mathrm{I}_{t-1}$为$t-1$期及之前的信息集。

此时有，分子 $\frac{1}{\sqrt{n}}\sum y_{t-1}u_t \sim \mathrm{N}(0,(\mathrm{E}y_{t-1})^2(\mathrm{E}u_t)^2) = \mathrm{N}(0,\sigma_y^2\sigma_u^2)$（其中有$y_{t-1}$和$u_t$独立，有$\sigma_y^2 = \frac{\sigma_u^2}{1-\rho^2}$）。所以 $\sqrt{n}(\hat{\rho}-\rho) \to \mathrm{N}(0,\frac{\sigma_u^2}{\sigma_y^2}) = \mathrm{N}(0,1-\rho^2)$。当$h>p$时，有$\rho=0$ $\sqrt{n}\hat{\rho} \sim \mathrm{N}(0,1)$（$\rho=0$）。

根据上述讨论，我们可以尝试区分AR(p)和MA(q)过程。

Theorem 4.1.3. ***AR Process Test***

If the true data follows the AR(p) process, then we can distinguish it from a MA process by testing whether $\rho^*(h)=0$ *for h>p*

$$\sqrt{n}(\hat{\rho}_h^* - \rho_h^*) \sim \mathrm{N}(0,1)$$

Theorem 4.1.4. ***MA Process Test***

If the true data follows the MA(q) process, then we can distinguish it from a AR process by testing whether $\rho(h)=0$ *for h>q. Formally, for a series* $\{y_t\}_{t=1}^n$, $y_t = \sum_{j=\infty}^{\infty}\phi_j z_{t-j}$, *where* $z_t \sim iid\mathrm{WN}(0,\sigma^2)$. *And we assume that* $\sum_{-\infty}^{\infty}|\phi_j| < \infty$, $\mathrm{E}[z_t]^4 < \infty$. *For* $h \in \{1,2,\dots\}$, *we have*

$$\sqrt{n}(\hat{\rho}_h - \rho_h) \sim \mathrm{N}(0,\mathrm{W})$$

where

$$\hat{\rho}_h = (\hat{\rho}(1),\dots,\hat{\rho}(h))'$$

$$\rho_h = (\rho(1),\dots,\rho(h))'$$

$$\mathrm{W}_{ij} = \sum_{k=-\infty}^{\infty}\{\rho(k+i)\rho(k+j)+\rho(k-j)\rho(k+j)+2\rho(i)\rho(j)\rho(k)^2-2\rho(i)\rho(k)\rho(k+j)+2\rho(j)\rho(k)\rho(k+i)\}$$

To be more precise, if $\{y_t\}_{t=1}^n$ *is a MA process, when* $h>q$, *we have*

$$\sqrt{n}\hat{\rho}(h) \sim \mathrm{N}(0,\mathrm{V}),\ \textit{where}\ \mathrm{V} = 1+2\sum_{s=1}^{q}\rho^2(s)$$

序列相关性检验. 首先先介绍对单个相关系数的检验。对于AR(1)模型：$y_t = \rho y_{t-1}+u_t$，如何检验ρ是否等于0。假设 $\mathrm{H}_0:\ \rho(h)=0,\ \mathrm{H}_1:\ \rho(h)\neq 0$ for each $h>0$。根据ρ的定义

$$\hat{\rho}(1) = \frac{\gamma(1)}{\gamma(0)} = \frac{\frac{1}{\sqrt{n}}\sum y_t y_{t-1} \to \mathrm{MDCLT}}{\frac{1}{n}\sum y_t^2 \to \mathrm{LLN}}$$

所以有$\sqrt{n}\hat{\rho}(h) \sim \mathrm{N}(0,1)$

上述讨论检验的是1个h，解下来将介绍如何同时检验多个h（Box & Pierce, 1970）。

$H_0 : \rho(h) = 0, h = 1,2,\ldots,p$; $H_1 : \rho(h) \neq 0,$ for some h. 因为$\sqrt{n}\hat{\rho}(h) \sim N(0,1)$，可以构造

$$Q = n\sum_{h=1}^{p}\hat{\rho}(h)^2 \sim \chi^2(p)$$

残差相关性检验. 上述内容讨论的是数据本身序列相关性的检验，下面讨论模型残差项的序列相关性检验：Test error serial correlation。如果残差没有序列相关性，则说明模型完备，但是否可以用刚才的方法检验error的序列相关性呢？分布会有所不同吗？

这取决于残差项本身的分布。

- 可用的情况

Example 1. 模型为$y_t = x_t'\beta + u_t$，且x严格外生，即$E[u_t|x_1,x_2,\ldots,x_t] = 0$。

$$\begin{aligned}
\gamma(1) &= E[u_{t-1}u_t] \\
\hat{u}_t &= y_t - x_t'\beta = u_t - x_t'(\hat{\beta}-\beta) \\
\hat{u}_{t-1} &= y_{t-1} - x_{t-1}'\beta = u_{t-1} - x_{t-1}'(\hat{\beta}-\beta) \\
\hat{\gamma}(1) &= \frac{1}{n}\sum_t \hat{u}_{t-1}\hat{u}_t \\
&= \frac{1}{n}\sum[u_tu_{t-1} - u_{t-1}x_t'(\hat{\beta}-\beta) - u_tx_{t-1}'(\hat{\beta}-\beta) + (\hat{\beta}-\beta)x_{t-1}'x_t(\hat{\beta}-\beta)] \\
&\overset{def}{=} A_1 - A_2 - A_3 + A_4
\end{aligned}$$

For the concept, if $A = O_p(\frac{1}{n})$, $B = O_p(\frac{1}{n^2})$, then A is leading term, B is s.o.(smaller order) term. 对于$a_n = O_p(b_n)$, 如果有$E|a_n| = O(\frac{1}{n})$ 则$a_n = O_p(\frac{1}{n})$,如果有$E(a_n)^2 = O(\frac{1}{n^2})$ 则 $a_n = O_p(\frac{1}{n})$

$$\begin{aligned}
A_1 &= \frac{1}{n}\sum_t u_tu_{t-1} \\
E[A_1] &= E[u_tu_{t-1}] = 0 \\
Var(\frac{1}{n}\sum_t u_tu_{t-1}) &= \frac{1}{n^2}Var(\sum_t u_tu_{t-1}) \\
&= \frac{1}{n}Var(u_tu_{t-1}) = \frac{1}{n}\sigma_u^2\sigma_u^2 \\
&= \frac{1}{n}\sigma^4 = O(\frac{1}{n})
\end{aligned}$$

For the second component A_2,

$$\begin{aligned}
A_2 &= [\frac{1}{n}\sum_t u_{t-1}x_t'](\hat{\beta}-\beta) = \underset{1\times k}{B_2}C_2 \\
\sqrt{n}(\hat{\beta}-\beta) &\sim N(0,\frac{\cdot}{n}) = O_p(\frac{1}{\sqrt{n}}) \\
Var(B_2') &= \frac{1}{n^2}[nVar(x_t u_{t-1}) + 2\sum_t \sum_{s>t} Cov(x_t u_{t-1}, x_s u_{s-1})] \\
&= \frac{1}{n}E[x_t u_{t-1}^2 x_t'] = \frac{1}{n}\underset{\sigma_u^2}{E[u_{t-1}^2]}\underset{finite}{E[x_t x_{t-1}]} \\
B_2 &= O_p(\frac{1}{\sqrt{n}}) \\
\Longrightarrow A_2 &= A_3 = B_2C_2 = O_p(\frac{1}{\sqrt{n}})O_p(\frac{1}{\sqrt{n}}) = O_p(\frac{1}{n})
\end{aligned}$$

At the end, let's focus on A_4,

$$\begin{aligned}
A_4 &= \frac{1}{n}\sum_t (\hat{\beta}-\beta)x_{t-1}x_t'(\hat{\beta}-\beta) \\
&= (\hat{\beta}-\beta)\frac{1}{n}\sum_t x_{t-1}x_t'(\hat{\beta}-\beta) \\
&= O_p(\frac{1}{\sqrt{n}})(?)O_p(\frac{1}{\sqrt{n}}) \\
Var(\frac{1}{n}\sum x_{t-1}x_t') &= E[\underset{\xrightarrow{p} O(\frac{1}{n})}{\frac{1}{n}\sum x_{t-1}x_t'}]^2 - (\underset{O_p(1)}{E[x_{t-1}x_t']})^2 = O_p(1) \\
\Rightarrow ? &= O_p(1) \\
A_4 &= O_p(\frac{1}{\sqrt{n}})O_p(1)O_p(\frac{1}{\sqrt{n}}) = O_p(\frac{1}{n})
\end{aligned}$$

So A_1 is leading term.

$$\begin{aligned}
\hat{\gamma}(1) &= A_1 + O_p(\frac{1}{n}) \\
\sqrt{n}\hat{\gamma}(1) &= \sqrt{n}A_1 + O_p(\frac{1}{\sqrt{n}}) \sim N(0,\sigma_u^4) \\
\hat{\gamma}(0) &= \frac{1}{n}\sum \hat{u}_t^2 = \frac{1}{n}\underset{\to Eu_t^2}{\sum_t u_t^2} - \frac{2}{n}\underset{O_p(1)}{\sum_t u_t x_t'(\hat{\beta}-\beta)} + (\hat{\beta}-\beta)\underset{O_p(1)}{\frac{1}{n}\sum_t x_t' x_t}(\hat{\beta}-\beta) \\
\text{where } \hat{u}_t^2 &= u_t^2 - 2u_t x_t'(\hat{\beta}-\beta) + x_t'(\hat{\beta}-\beta)(\hat{\beta}-\beta)' x_t \\
\hat{\gamma}(0) &\to E[u_t^2] = \sigma_u^2 \\
\text{Hence, } \sqrt{n}\hat{\rho}(1) &= \sqrt{n}\frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} \sim N(0,1)
\end{aligned}$$

- 不能使用的情况

Example 2. 模型AR(1)为$y_t = \phi y_{t-1} + u_t$, $u_t \sim \text{WN}(0, \sigma^2)$. $\text{H}_0 : \gamma(1) = \text{E}(u_{t-1}u_t) = 0$.

$$\begin{aligned}
\hat{\gamma}(1) &= \frac{1}{n}\sum_t \hat{u}_t\hat{u}_{t-1} \\
&= \dots \\
&= \frac{1}{n}\sum_t [u_t u_{t-1} - u_{t-1}y_{t-1}(\hat{\phi}-\phi) - u_t u_{t-2}(\hat{\phi}-\phi) + y_{t-1}y_{t-2}(\hat{\phi}-\phi)^2] \\
&= \text{A}_1 - \text{A}_2 - \text{A}_3 + \text{A}_4
\end{aligned}$$

类似的，有$\text{A}_1 = \text{O}_p(\frac{1}{\sqrt{n}})$, $\text{A}_3 = \text{O}_p(\frac{1}{n})$, $\text{A}_4 = \text{O}_p(\frac{1}{n})$，唯一不一样的是$\text{A}_2$

$$\begin{aligned}
\text{A}_2 &= [\frac{1}{n}\sum_t y_{t-1}u_{t-1}](\hat{\phi}-\phi) \\
\text{Var}([\frac{1}{n}\sum_t y_{t-1}u_{t-1}]) &= \text{E}\left[\frac{1}{n}\sum_t y_{t-1}u_{t-1}\right]^2 - (\underbrace{\text{E}[y_{t-1}u_{t-1}]}_{=\text{O}(1)})^2 \to 0 \\
\underbrace{\text{A}_2}_{=\text{O}(1)} &= \text{O}_p(1)\text{O}_p(\frac{1}{\sqrt{n}}) = \text{O}_p(\frac{1}{\sqrt{n}})
\end{aligned}$$

Back to the distribution of $\hat{\phi}-\phi$:

$$\begin{aligned}
\text{A}_2 &= [\frac{1}{n}\sum_t y_{t-1}u_{t-1}](\hat{\phi}-\phi) \\
&= \left[\text{E}[y_{t-1}u_{t-1}] + o_p(1)\right]\left[\frac{1}{n}\sum y_{t-1}^2\right]^{-1}\left[\frac{1}{n}\sum y_{t-1}u_t\right] \\
&\approx \sigma_u^2(\frac{1-\phi^2}{\sigma_u^2})[\frac{1}{n}\sum y_{t-1}u_t] \\
&= (1-\phi^2)[\frac{1}{n}\sum y_{t-1}u_t].
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
\sqrt{n}\hat{\gamma}(1) &= \sqrt{n}(\text{A}_1 - \text{A}_2) + \text{O}_p(1) \\
&\approx \sqrt{n}\left(\frac{1}{n}\sum u_t u_{t-1} - (1-\phi^2)\frac{1}{n}\sum y_{t-1}u_t\right) \\
&= \frac{1}{\sqrt{n}}\sum_t u_t[u_{t-1} - (1-\phi^2)y_{t-1}]
\end{aligned}$$

根据MDCLT，$\text{E}\left[u_t[u_{t-1} - (1-\phi^2)y_{t-1}]\right] = 0$，可得（证明见Equation 4.3）：

$$\text{Var}(u_t[u_{t-1} - (1-\phi^2)y_{t-1}]) = \sigma_u^2\phi^2$$

所以有，

$$\begin{aligned}\sqrt{n}\hat{\gamma}(1) &\sim \mathrm{N}(0,\sigma_u^4\phi^2)\\ \hat{\gamma}(0) &\to \mathrm{E}[u_t^2]=\sigma_u^2\\ \sqrt{n}\hat{\rho}(1) &= \sqrt{n}\frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} \overset{\mathrm{H}_0}{\sim} \mathrm{N}(0,\phi^2)\end{aligned}$$

其中

$$\begin{aligned}\mathrm{E}\left[u_t[u_{t-1}-(1-\phi^2)y_{t-1}]\right] &= \sigma_u^2\mathrm{E}\left[u_{t-1}^2+(1-\phi^2)^2y_{t-1}^2-2u_{t-1}(1-\phi^2)y_{t-1}\right]\\ &= \sigma_u^2\left[\sigma_u^2+(1-\phi^2)^2(\frac{1-\phi^2}{\sigma_u^2})^{-1}-2(1-\phi^2)\sigma_u^2\right]\\ &= \sigma_u^4\phi^2\end{aligned} \tag{4.3}$$

4.1.4 滞后回归

Auto Regressive Distributional Lag Model. In ARDL, the regressors may include lagged values of the dependent variable and current and lagged values of one or more explanatory variables. This model allows us to determine what the effects are of a change in a policy variable. A general ARDL model is as follows:

$$\underbrace{y_t=\mu+\gamma_1y_{t-1}+\cdots+\gamma_py_{t-p}}_{autoregressive}+\underbrace{\beta_0x_t+\beta_1x_{t-1}+\cdots+\beta_rx_{t-r}}_{distribution\,lag}+\varepsilon_t \tag{4.4}$$

$$\Delta \text{ in } x\begin{cases}\text{shocks in short run}\begin{cases}\mathrm{S}hort\mathrm{R}un & \mathrm{Y}es\\ \mathrm{L}ong\mathrm{R}un & \mathrm{N}o\end{cases}\\ \text{permenant change}\begin{cases}\mathrm{S}hort\mathrm{R}un & \mathrm{Y}es\\ \mathrm{L}ong\mathrm{R}un & \mathrm{Y}es\end{cases}\end{cases}$$

Table 4.1展示了在t^*时，x增加一单位对y的影响。 Suppose an equilibrium with (x^*,y^*), we have

表 4.1 Effect of Change in x on change in y

Δx	Time	Δy
$\Delta x=1$	t^*	β_0
	t^*+1	$\beta_0+\beta_1+\gamma_1\beta_0$
	t^*+2	$\beta_0+\beta_1+\beta_2+\gamma_2(\beta_0+\beta_1+\gamma_1\beta_0)$

$$y^* = \mu + \gamma_1 y^* + \cdots + \gamma_p y^* + \beta_0 x^* + \beta_x^* + \cdots + \beta_r x^* + \varepsilon$$
$$\Longrightarrow y^* = \frac{\hat{\mu}^2}{1-\sum_{i=1}^{p}\hat{\gamma}_i} + \frac{\sum_{i=0}^{r}\hat{\beta}_j}{1-\sum_{i=1}^{p}\hat{\gamma}_i} x^*$$

For simplicity, we define some tools based on the lag operator.

Definition 4.1.9. *Lag Operator*

A simple lag operator is defined as the function to let $\mathrm{L}y_t = y_{t-1}$*. Meanwhile, it statisfies (1)* $\mathrm{L}^2 y_t = \mathrm{L}(\mathrm{L}y_t) = y_{t-2}$*;(2)* $(1-\mathrm{L})y_t = y_t - y_{t-1} = \Delta y_t$*.*

Definition 4.1.10. *For the simplicity in analysing* $\mathrm{ARDL}(r,p)$*, we define two functions:*

$$\mathrm{B}(\mathrm{L}) = \beta_0 + \beta_1 \mathrm{L} + \beta_2 \mathrm{L}^2 + \cdots + \beta_r \mathrm{L}^r$$
$$\mathrm{C}(\mathrm{L}) = 1 - \gamma_1 \mathrm{L} - \cdots - \gamma_p \mathrm{L}^p$$

Hence, the $\mathrm{ADL}(r,p)$ *model can be written as* $\mathrm{C}(\mathrm{L})y_t = \mu + \mathrm{B}(\mathrm{L})x_t + \varepsilon_t$*.*

Example: Partial Adjustment Model. The main model is as follows:

$$\begin{aligned} y_t^* &= \alpha + \beta x_t + \delta w_t + \varepsilon_t \\ y_t - y_{t-1} &= (1-\lambda)(y_t^* - y_{t-1}) \end{aligned}$$

Plugging the change in y, we could have

$$\begin{aligned} y_t &= \alpha(1-\lambda) + \beta(1-\lambda)x_t + \delta(1-\lambda)w_t + \lambda y_{t-1} + (1-\lambda)\varepsilon_t \\ y_t &= \alpha' + \beta' x_t + \delta' w_t + \lambda y_{t-1} + \varepsilon_t' \\ \mathrm{C}(\mathrm{L})y_t &= \alpha' + \beta' x_t + \delta' w_t + \varepsilon_t', \quad \text{where } \mathrm{C}(\mathrm{L}) = 1 - \lambda \mathrm{L} \\ \frac{1}{\mathrm{C}(\mathrm{L})} = \frac{1}{1-\lambda \mathrm{L}} &= 1 + \lambda \mathrm{L} + (\lambda \mathrm{L})^2 + (\lambda \mathrm{L})^3 + \ldots \quad |\lambda| < 1 \end{aligned}$$

$$y_t = \underbrace{[\alpha' + \lambda\alpha' + \lambda^2\alpha' + \ldots]}_{\text{Permenant Change}} + \underbrace{\beta'[x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \ldots]}_{\text{Shocks}} + \delta'[w_t + \lambda w_{t-1} + \ldots] + [\varepsilon_t' + \lambda\varepsilon_{t-1}' + \ldots] \tag{4.5}$$

Example: Common Factor. The main model is as follows:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

Organizing the conditions, we could have

$$\begin{aligned}
(1-\rho \mathrm{L})\varepsilon_t = u_t \quad &\Rightarrow \quad y_t = \beta_0 + \beta_1 x_t + \frac{u_t}{1-\rho \mathrm{L}} \\
y_t - \rho y_{t-1} \quad &= \quad (\beta_0 - \rho\beta_0) + \beta_1 x_t - \beta_1 \rho x_{t-1} + u_t \\
y_t \quad &= \quad \beta_0 + \beta_1 x_t - \beta_1 \rho x_{t-1} + \rho y_{t-1} + u_t \\
\text{(ARDL) } y_t \quad &= \quad \gamma_0' + \gamma_1 x_t + \gamma_2 x_{t-1} + \gamma_3 y_{t-1} + u_t
\end{aligned}$$

Common factor restriction means $\gamma_2 = -\gamma_1\gamma_3$ then $y_t = \gamma_0 + \gamma_1 x_t + \gamma_2 x_{t-1} + \gamma_3 y_{t-1} + u_t$ can be converted to $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ with AR(1) error. Note that β_1 is long run effect and γ_2 is short run effect.

In Particular, for an ARDL, we can test the serial correlation in some ways.

$$y_t = lag\ y_t's + \beta x_t + lag\ x_t's + \underset{serial\ corr}{error}$$

- BG test;
- Durbin's *h* test;
- Add extra lag term in ARDL to test the coefficients significance of these extra lagged terms;

Example: Vector Autoregression.

$$y_t = \mu + \underset{e.g.4\times4}{\Gamma_1}\ y_{t-1} + \Gamma_2 y_{t-2} + \cdots + \Gamma_p y_{t-p} + \varepsilon_t, \varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$$

有$y_t = (y_{1t}, y_{2t}, y_{3t}, y_{4t})'$， $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t}, \varepsilon_{3t}, \varepsilon_{4t})'$。

(1) $\mathrm{E}\varepsilon_t = 0$, (2) $\mathrm{E}\varepsilon_t\varepsilon_s' = 0$, $\forall t \neq s$, (3)However, $\mathrm{E}\varepsilon_t\varepsilon_t' = \Omega_{4\times4}$中非对角线元素可以不为0。[1]

对于冲击的影响在期数较短时（一期）具有很好的解释，Γ_1对应位置的系数即代表边际影响。但如果期数较高时，由于冲击的影响会随时间累积，此时Γ_1中系数并不仅仅代表某次冲击对下一期产出的影响，而代表的是累积效应，其难以翻译。我们采用以下形式来解释该问题：

$$\begin{aligned}
y_t \quad &= \quad \hat{\mu} + \hat{\Gamma}_1 y_{t-1} + \hat{\Gamma}_2 y_{t-2} + e_t \\
y_{t-1} \quad &= \quad \hat{\mu} + \hat{\Gamma}_1 y_{t-2} + \hat{\Gamma}_2 y_{t-3} + e_t \\
y_t \quad &= \quad (\hat{\mu} + \hat{\Gamma}_1\hat{\mu}) + (e_t + \hat{\Gamma}_1 e_{t-1}) + (\hat{\Gamma}_1^2 + \hat{\Gamma}_2) y_{t-2} + \hat{\Gamma}_1\hat{\Gamma}_2 y_{t-3} \\
&\vdots
\end{aligned}$$

[1]此回归中，分开执行的OLS（separate OLS）和GLS是等价的，见Theorem 4.1.5。

Theorem 4.1.5. ***Equivalence of FGLS and OLS***

If $x_{i1} = x_{i2} = \cdots = x_{iG}$ for all i, that is, if the same regressors show up in each equation (for all observations) then OLS equation by equation and FGLS are identical.

4.1.5 非平稳序列

以上讨论基本都是基于平稳的时间序列进行的，本节讲介绍一些典型的非平稳序列和其估计出来的分布。主要集中于两类：

- Trend Stationary: $y_t = \alpha + \beta t + u_t$
- Random Walk: $y_t = y_{t-1} + u_t$

Trend. 首先，trend series可以写作

$$\begin{aligned} y_t &= \alpha + \beta t + u_t \\ \Rightarrow \text{OLS } y &= \mathrm{X}(\alpha,\ \beta)' + u \end{aligned}$$

where

$$\mathrm{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & \mathrm{T} \end{pmatrix}$$

可得系数的估计为：

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'u$$

根据之前所学，可得

$$\mathrm{E}\left[\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}\right] = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},\ \mathrm{V}ar\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}$$

此时我们则关心的是估计值以怎样的速度趋近于真值，即，方差中的$(\mathrm{X}'\mathrm{X})^{-1}$部分以什么速度趋近于0。

首先求出$(X'X)^{-1}$的表达式：

$$X'X = \begin{pmatrix} n & \frac{n(n+1)}{2} \\ \frac{n(n+1)}{2} & \frac{1}{6}n(n+1)(2n+1) \end{pmatrix} \Rightarrow (X'X)^{-1} = \frac{1}{n}\begin{pmatrix} \frac{2(2n+1)}{n-1} & -\frac{6}{n-1} \\ -\frac{6}{n-1} & \frac{12}{n^2-1} \end{pmatrix} \to 0 \; as \; n \to \infty$$

Hence,

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} - \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \frac{2(2n+1)}{n-1} & -\frac{6}{n-1} \\ -\frac{6}{n-1} & \frac{12}{n^2-1} \end{pmatrix} \begin{pmatrix} \sum u_t \\ \sum t u_t \end{pmatrix} = \frac{1}{n}\begin{pmatrix} O(1) & O(\frac{1}{n}) \\ O(\frac{1}{n}) & O(\frac{1}{n^2}) \end{pmatrix} \begin{pmatrix} O(n^{1/2}) \\ O(n^{3/2}) \end{pmatrix} = \begin{pmatrix} O(n^{-1/2}) \\ O(n^{-3/2}) \end{pmatrix}$$

where

$$\begin{aligned} Var(\sum u_t) &= n\sigma_u^2 = O(n) \\ Var(\sum t u_t) &= \sum t^2 \sigma_u^2 = \frac{\sigma_u^2}{6} n(n+1)(2n+1) \end{aligned}$$

To get a standard normal distribution by applying the "CLT", we scale the estimation by its speed to zero.

$$\begin{aligned} \begin{pmatrix} \sqrt{n}(\hat{\alpha}-\alpha) \\ n^{\frac{3}{2}}(\hat{\beta}-\beta) \end{pmatrix} &= \begin{pmatrix} \sqrt{n} & 0 \\ 0 & n^{\frac{3}{2}} \end{pmatrix} (X'X)^{-1} \begin{pmatrix} \sqrt{n} & 0 \\ 0 & n^{\frac{3}{2}} \end{pmatrix} \underbrace{\begin{pmatrix} \frac{\sum u_t}{\sqrt{n}} \\ \frac{\sum t u_t}{n^{\frac{3}{2}}} \end{pmatrix}}_{=A} \\ &= \begin{pmatrix} \frac{2(2n+1)}{n-1} & -\frac{6n}{n-1} \\ -\frac{6n}{n-1} & \frac{12n^2}{n^2-1} \end{pmatrix} A \to \underbrace{\begin{pmatrix} 4 & -6 \\ -6 & 12 \end{pmatrix}}_{B} A \end{aligned}$$

where $A = (\frac{\sum u_t}{\sqrt{n}}, \frac{\sum t u_t}{n^{\frac{3}{2}}})'$ with $E[A] = 0$ and

$$Var(A) = E[AA'] = E\begin{pmatrix} \frac{\sum u_t}{\sqrt{n}}\frac{\sum u_t}{\sqrt{n}} & \frac{\sum u_t}{\sqrt{n}}\frac{\sum t u_t}{n^{\frac{3}{2}}} \\ \frac{\sum u_t}{\sqrt{n}}\frac{\sum t u_t}{n^{\frac{3}{2}}} & \frac{\sum u_t}{\sqrt{n}}\frac{\sum t u_t}{n^{\frac{3}{2}}}\frac{\sum u_t}{\sqrt{n}}\frac{\sum t u_t}{n^{\frac{3}{2}}} \end{pmatrix} = \begin{pmatrix} \sigma^2 & \frac{1}{n^2}\sigma^2\frac{n(n+1)}{2} \\ \frac{1}{n^2}\sigma^2\frac{n(n+1)}{2} & \frac{\sigma^2 n(n+1)(2n+1)}{6n^3} \end{pmatrix} \to \sigma^2\begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix}$$

So we have the distribution of A

$$A \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2\begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix})$$

Applying *Lindberg Feller CLT*, the distribution

$$\begin{pmatrix} \sqrt{n}(\hat{\alpha}-\alpha) \\ n^{\frac{3}{2}}(\hat{\beta}-\beta) \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, B\underbrace{AA'}_{=B^{-1}}B) = N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2\begin{pmatrix} 4 & -6 \\ -6 & 12 \end{pmatrix})$$

Random Walk.

$$y_t = \rho y_{t-1} + u_t$$

If $|\rho| < 1$, this process is stationary while $|\rho| = 1$, this implies a random walk. When If $|\rho| < 1$, we could have $\sqrt{n}(\hat{\rho} - \rho) \sim \mathrm{N}(0, 1-\rho^2)$.

Browian Motion. In this subsection, we will define an important process in finance, called *Browian Motion*.[2]

Definition 4.1.11. $\mathrm{W}(r)$ *is a standard Brownian motion if for* $r \in [0,1]$ $\mathrm{W}(r)$ *satisfies the following:*

1. $\mathrm{W}(0) = 0$
2. $\forall r > s, \mathrm{W}(r) - \mathrm{W}(s) \sim \mathrm{N}(0, r-s)$ *and* $[\mathrm{W}(r) - \mathrm{W}(s)] \perp \mathrm{W}(s)$
3. $\mathrm{W}(r)$ *has a continuous sample path (but not differentiable).*

Based on the definitions, we could know $w(s) \sim \mathrm{N}(0, r)$,*for* $s > 0$.

Example. Assume that we observe data $y_1, \ldots, y_n$. Let $[nr]$ denote the integer part of nr. Then as r changes from 0 to 1, $[nr]$ changes from 0 to n discretely. Let $y_n(r) = \sum_{t=1}^{[nr]} u_t$, where u_t is iid process with $\mathrm{E}[u_t] = 0$ and $\mathrm{V}ar(u_t) = \sigma^2$. Then we have,

$$y_n(r) = \begin{cases} 0, & 0 \le r < \dfrac{1}{n} \\ u_1, & \dfrac{1}{n} \le r < \dfrac{2}{n} \\ u_1 + u_2, & \dfrac{2}{n} \le r < \dfrac{3}{n} \\ \vdots & \vdots \\ u_1 + u_2 + \cdots + u_n, & r = 1 \end{cases} \tag{4.6}$$

For any fixed $r \in [0,1]$, we can show $\frac{1}{\sqrt{n}} y_n(r) \sim \mathrm{N}(0, r\sigma^2)$

$$\frac{1}{\sqrt{n}} y_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} u_t = \frac{\sqrt{[nr]}}{\sqrt{n}} \left(\frac{1}{\sqrt{[nr]}} \sum_{t=1}^{[nr]} u_t \right) \to \sqrt{r}\mathrm{N}(0, \sigma^2) = \mathrm{N}(0, r\sigma^2) = \sigma \mathrm{W}(r)$$

[2]这一小节的内容可能是正常学金融随机分析两三周的内容，我理解的引入内容按逻辑来讲应该包括：对称随机游走、按比例缩小型对称随机游走及其极限概率分布、布朗运动、简单过程的随机积分的构造、简单过程的伊藤积分的性质（引入二次变差）、一般过程的伊藤积分、伊藤-德布林公式（又叫伊藤引理）。感兴趣的同学可以移步Shreve(2004)'s Stochastic Calculus for Finance II Continuous Time Models。北大的李东风老师有一个从数学角度讲解金融随机过程的lecture，见https://www.math.pku.edu.cn/teachers/lidf/course/stochproc/stochprocnotes/html/_book/index.html。

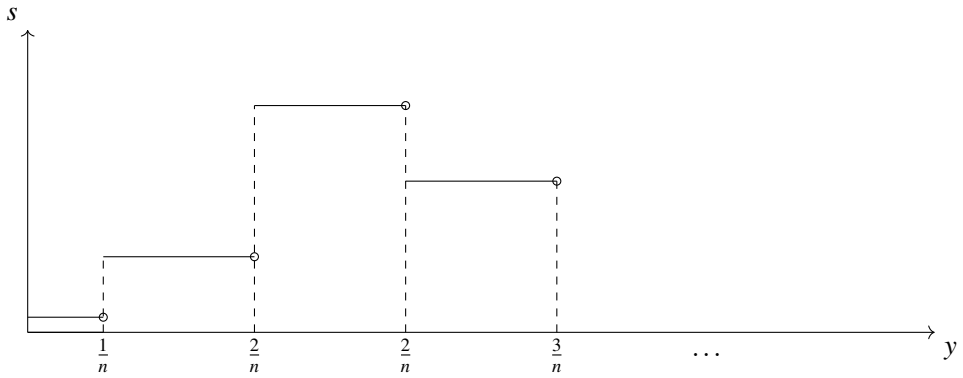

Figure 4.1 Realization of $y_n(t)$

下面我们将证明该过程的增量独立

$$\mathrm{C}_n(r+\Delta r)-\mathrm{C}_n(r) \perp \mathrm{C}_n(r)$$
$$\sum_{t=[nr]+1}^{[n(r+\Delta r)]} u_t \perp \sum_{t=1}^{[n(r)]} u_t$$
$$\text{if } r=1,\ \frac{1}{\sqrt{n}}\frac{y_n(1)}{\sigma} \to \mathrm{W}(1)$$

随机积分. 在介绍随机积分之前，我们先回顾之前数分学过的黎曼积分的形式（取微小变元区间的左端点代表该段的平均值，极限情况中微小变元的区间无限趋近于0，故在这个框架下取哪都一样）：

$$\int_0^1 g(x)dx = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} g(\xi_i)\Delta x_i = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} g(x_{i-1})$$

但对于随机变量积分而言，由于其区间右端点并不能取到（并且不连续），因此将采用*Itô*积分的方式对该过程进行积分，可得

$$\int_0^1 g(w(s))dw(s) = \lim_{n\to\infty}\sum_{i=1}^{n} g(w(s_i))(w(s_{i+1})-w(s_i))$$
$$g(w(s_{i+1}))\times \quad \text{对应}\frac{1}{n}$$

$w(s_i)$不能取$w(s_{i+1})$。一些基本的离散运算和积分运算包括：(1) $\frac{1}{n}\sim dr$; (2) $\sum_i^n \sim \int_0^1$; (3) $\frac{1}{\sqrt{n}}y_{t-1} = \frac{1}{\sqrt{n}}\sum_{s=1}^{[nr]} u_s \sim \sigma_u w_u(r)$; (4) $\frac{u_t}{\sqrt{n}} = \frac{\mathrm{Y}_t-\mathrm{Y}_{t-1}}{\sqrt{n}} = d\sigma_u w_u(r)$。基于此，可得

$$\begin{aligned} n(\hat{\rho}-1) &= \frac{n\sum y_{t-1}u_t}{\sum y_{t-1}^2} = \frac{\sum \frac{y_{t-1}}{\sqrt{n}}\frac{u_t}{\sqrt{n}}}{\frac{1}{n}\sum(\frac{y_{t-1}}{\sqrt{n}})^2} \\ &\to \frac{\int_0^1 \sigma_u^2 w_u(r)dw_u(r)}{\int_0^1 \sigma_u^2 w_u^2(r)dr} \end{aligned}$$

在单位根检验中，该统计量并不服从于t分布，其值是通过蒙特卡洛模拟得到的。

协整检验. 最后，以协整检验为例来考虑上述的应用。有y_t和x_t是I(1)过程，$y_t = x_t\beta + u$，其中

$$\left.\begin{array}{l} x_t = x_{t-1} + v_t \\ y_t = y_{t-1} + \varepsilon_t \end{array}\right\} v_t, \varepsilon_t \ iid \text{ random walk}$$

有

$$\begin{aligned} \hat{\beta} - \beta &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'u \\ \to n(\hat{\beta} - \beta) &= (\frac{1}{n^2}\sum x_t^2)^{-1}\frac{1}{n}\sum x_t u_t \\ &= (\frac{1}{n^2}\sum x_t^2)^{-1}(\frac{1}{n}\sum x_{t-1}u_t + \frac{1}{n}\sum v_t u_t) \end{aligned}$$

其中有

$$\begin{aligned} \frac{1}{n}\sum(\frac{x_t}{\sqrt{n}})^2 &\to \int_0^1 \sigma_v^2 w_r^2(r)dr \\ \sum\frac{x_{t-1}}{\sqrt{n}}\frac{u_t}{\sqrt{n}} &\to \int_0^1 \sigma_v w_v(r)d\sigma_u w_u(r) \\ \frac{1}{n}\sum v_t u_t &\to \mathrm{E}[v_t u_t] \overset{assume}{=} 0 \end{aligned}$$

化简得到

$$n(\hat{\beta} - \beta) \sim \left[\int_0^1 \sigma_v^2 w_r^2(r)dr\right]^{-1}\left[\int_0^1 \sigma_v w_v(r)\sigma_u dw_u(r)\right]$$

如果不选择用左端点计算，而选择用不能达到的右端点做积分会导致结果有问题，以下面为例

$$\begin{aligned} y_t u_t &= y_{t-1}u_t + u_t^2 \\ \frac{1}{n}\sum y_t u_t &= \frac{1}{n}\sum y_{t-1}u_t + \frac{1}{n}\sum u_t^2 \\ \text{右端点 } \frac{1}{n}\sum y_t u_t &\to \int_0^1 \sigma_u^2 w_u(r)dw_u(r) \\ \text{左端点 } \frac{1}{n}\sum y_{t-1}u_t &\to \int_0^1 \sigma_u^2 w_u(r)dw_u(r) \\ \frac{1}{n}\sum u_t^2 &\to \sigma^2 \\ \Longrightarrow \sigma^2 &= 0 \quad \text{与条件矛盾} \end{aligned}$$

4.2 面板数据

对于OLS回归而言，如果模型中包含部分不随时间变化、只随个体变化而变化的部分，诸如可观测的性别、种族、居住地，不可观测的基因、偏好等，此时模型应像如下设定

$$y_{it} = x_{it}'\beta + c_i + u_{it}$$

- Pooled OLS: c_i是一个常数项。If c_i only contains a constant term then OLS provides consisitent and efficient estimation for β.

  Fixed Effect Model: c_i is unobserved and varied, but correlated with X_{it}. We can re-write c_i as $c_i'\alpha$ to estimate their effects.

  3) Random Effect Model: Because c_i is uncorrelated with X_{it}, we can take it as a new error term $v_{it} = c_i + u_{it}$.

接下来我们将逐个介绍Fixed Effect和Random Effect模型。

4.2.1 Fixed Effect Model

Least Square Dummy Variable. For LSDV, we define observations at individual level like (the notation for i would be kind of mixed).

$$\underset{\mathrm{T}\times 1}{y_i} = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{i\mathrm{T}} \end{pmatrix}, \underset{\mathrm{T}\times 1}{i} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Stacked the vector for each individual, we can get

$$\underset{n\mathrm{T}\times 1}{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}} = \underset{n\mathrm{T}\times 1}{\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}} \beta + \underset{n\mathrm{T}\times n}{\underset{d_1 \ldots d_n}{\begin{pmatrix} i & & 0 \\ & \ddots & \\ 0 & & i \end{pmatrix}}} \underset{n\times 1}{\begin{pmatrix} \mathrm{C}_1 \\ \vdots \\ \mathrm{C}_n \end{pmatrix}} + \underset{n\mathrm{T}\times 1}{\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}}$$

$$\Longrightarrow \mathrm{FE}: \ y = \mathrm{X}\beta + \mathrm{D}\alpha + u$$

在估计β之前，首先将虚拟变量去掉（左乘一个消灭矩阵）。

$$\begin{aligned}
\mathrm{M_D}y &= \mathrm{M_D X}\beta + \underbrace{\mathrm{M_D DC}}_{=0} + u \\
\Longrightarrow \hat{\beta} &= (\mathrm{X'M_D X})^{-1}(\mathrm{X'M_D}y) \\
where\ \mathrm{M_D} &= \underset{n\mathrm{T}\times n\mathrm{T}}{\mathrm{I}} - \mathrm{D}(\mathrm{D'D})^{-1}\mathrm{D'} \\
&= \mathrm{I} - \begin{pmatrix} \frac{1}{\mathrm{T}}ii' & & 0 \\ & \vdots & \\ 0 & & \frac{1}{\mathrm{T}}ii' \end{pmatrix} = \begin{pmatrix} \mathrm{M}^0_{\mathrm{T}\times\mathrm{T}} & & 0 \\ & \ddots & \\ 0 & & \mathrm{M}^0 \end{pmatrix} \\
with\ \mathrm{M}^0 &= \mathrm{I} - \frac{1}{\mathrm{T}}ii'
\end{aligned}$$

但本质上$\mathrm{M_D}y_i = y_i - \mathrm{I}\bar{y}_i$、$\mathrm{M_D}x_i = x_i - \mathrm{I}\bar{x}$，我们所进行的回归其实是将变量去掉了组内均值后再做回归。其中，

$$\hat{\mathrm{C}} = (\mathrm{D'D})^{-1}\mathrm{D'}(y - \mathrm{X}\hat{\beta})$$

where

$$\begin{aligned}
\hat{\mathrm{C}}_i &= \frac{1}{\mathrm{T}}\sum_{t=1}\mathrm{T}y_{it} - \frac{1}{\mathrm{T}}\sum_{t=1}\mathrm{T}x'_{it}\hat{\beta} \\
&= \bar{y}_i - \bar{x}'_i\hat{\beta}
\end{aligned}$$

残差项方差的估计为

$$\hat{\sigma}^2 = \frac{(y - \mathrm{M_D X}\hat{\beta})'(y - \mathrm{M_D X}\hat{\beta})}{n\mathrm{T} - n - k}$$

最后，我们可以用*F-test*来检验不同个体的截距项是否相同（个体固定效应是否存在），即，H_0：$c_1 = c_2 = \cdots = c_n = 0$。或者包含截距项，构造替代的$\mathrm{D'}$以检验$\mathrm{H}_0$：$c_2 = c_3 = \cdots = c_n = 0$。

$$\mathrm{D'} = \begin{pmatrix} i & & 0 \\ 1 & \ddots & \\ 1 & & i \end{pmatrix}$$

上面的方法叫做one-way fixed effect，类似的也可以控制时间固定效应。一般而言，单向固定效应有两种估计方法：

- 组内估计量：分别减去两个组内的均值（等价于LSDV）；
- 组间估计量：用整组的平均值做估计。

两种方法都能得到一致有效的估计量，但其对u_t施加了不同的假设$\mathrm{E}[u_t|\mathrm{I}]$，不同方法的条件I是不同的。除此之外，我们常用的还包括two-way fixed effect，即同时控制两种固定效应。

4.2.2 Random Effect

如果C_i在回归中并不以截距项的形式出现，而出现在误差项中，我们则需要用到随机效应模型：

$$y_{it} = x_{it}'\beta + c_i + u_{it} = x_{it}'\beta + v_{it}$$

由于误差项的协方差矩阵不是对角阵（球形扰动项假设不成立），所以要用(F)GLS估计。首先，做一系列假设以方便推导：

$$\mathrm{E}(c_i|\mathrm{X}) = 0$$

$$\mathrm{E}(u_{it}|\mathrm{X}) = 0$$

$$\mathrm{E}(u_{it}^2|\mathrm{X}) = \sigma_u^2$$

$$\mathrm{E}(c_i^2|\mathrm{X}) = \sigma_c$$

$$\mathrm{E}(u_{it}c_j|\mathrm{X}) = 0$$

$$\mathrm{E}(u_{it}u_{js}|\mathrm{X}) = 0, i \neq j \text{ or } s \neq t$$

$$\mathrm{E}(c_i c_j|\mathrm{X}) = 0 \text{if } i \neq j$$

对合成的误差项$v_{it} = \mathrm{C}_i + u_{it}$，误差成分模型（ECM）如下：

$$\begin{aligned}
\mathrm{E}[v_{it}^2|\mathrm{X}] &= \sigma_c^2 + \sigma_u^2 \\
\mathrm{E}[v_{it}v_{is}|\mathrm{X}] &= \sigma_c^2, \forall t \neq s \\
\mathrm{E}[v_{it}v_{js}|\mathrm{X}] &= \sigma_c^2, \forall t \neq s \text{ and } i \neq j
\end{aligned}$$

对于同一个个体不同时期的扰动项之间的自相关系数ρ不随时间距离$t-s$而改变的模型，我们称之为等相关模型或可交换扰动项模型。如果ρ越大，则复合扰动项中个体效应的部分(C_i)越重要。回到复合扰动项，此时有

$$v_i = \begin{pmatrix} v_{i1} \\ \vdots \\ v_{i\mathrm{T}} \end{pmatrix}, \Sigma = \mathrm{E}v_i v_i' = \begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \ddots & & \\ & & \ddots & \\ \sigma_c^2 & \ddots & \ddots & \sigma_c^2 + \sigma_u^2 \end{pmatrix} = \sigma_u^2 \mathrm{I}_\mathrm{T} + \sigma_c^2 i_\mathrm{T} i_\mathrm{T}'$$

对于所有个体的观测而言，有[3]

$$\Omega = \mathrm{E}[\mathrm{VV}'] = \mathrm{I}_n \otimes \Sigma = \begin{pmatrix} \Sigma & & 0 \\ & \ddots & \\ 0 & & \Sigma \end{pmatrix}$$

故，最后的估计结果为：

$$\begin{aligned} \hat{\beta}^{\mathrm{GLS}} &= (\mathrm{X}'\Omega^{-1}\mathrm{X})^{-1}\mathrm{X}'\Omega^{-1}y \\ &= (\sum_{i=1}^{n} x_i'\Sigma^{-1}x_i)^{-1}(\sum_{i=1}^{n} x_i\Sigma^{-1}y_i) \end{aligned}$$

即对原数据做$\Omega^{-\frac{1}{2}}$变化，由于$\Omega^{-\frac{1}{2}} = [\mathrm{I}_n \otimes \Sigma]^{-\frac{1}{2}}$，故我们仅需知道$\Sigma$即可，

$$\begin{aligned} \Sigma^{-\frac{1}{2}} &= \frac{1}{\sigma_u}[\mathrm{I} - \frac{\theta}{\mathrm{T}} ii'] \\ \theta &= 1 - \frac{\sigma_u}{\sqrt{\sigma_u^2 + \mathrm{T}\sigma_c^2}} \\ \Sigma^{-\frac{1}{2}} y_i &= \frac{1}{\sigma_u} \begin{pmatrix} y_{i1} - \theta \bar{y}_{i\cdot} \\ \vdots \\ y_{i\mathrm{T}} - \theta \bar{y}_{i\cdot} \end{pmatrix} \end{aligned}$$

其中$y_{i\mathrm{T}} - \theta \bar{y}_{i\cdot}$被成为*quasi-demean*，它和前述的固定模型思路一致，只是在demean时给了均值一个权重。若$\theta = 1$，则$\sigma_c = 0$，模型就是固定效应模型。

$\sigma_u^2\sigma_c^2$的估计. 但实际情况中，我们通常并不知道σ_u^2和σ_c^2，因而，我们还需要对σ_u^2和σ_c^2进行估计。首先我们先利用$y_{it} = x_{it}'\hat{\beta} + \hat{v}_{it}$对$\sigma_u^2 + \sigma_c^2$进行估计。

$$\widehat{\sigma_u^2 + \sigma_c^2} = \frac{1}{n\mathrm{T} - k - 1} \sum_{i=1}^{n} \sum_{t=1}^{\mathrm{T}} \hat{\mathrm{v}}_{it}^2$$

[3]For the definition of Kronecker product, we illustrate it as

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad \mathrm{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$\mathrm{A} \otimes \mathrm{B} = \begin{pmatrix} a_{11}\mathrm{B} & a_{12}\mathrm{B} \\ a_{21}\mathrm{B} & a_{22}\mathrm{B} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{pmatrix}$$

利用FE模型中所估计参数的一致性，且其扰动项为$u_{it}-\bar{u}_{i\cdot}$，所以可以用FE的残差来估计σ_u^2，即

$$\begin{aligned}\bar{y}_{i\cdot} &= \bar{x}_{i\cdot}'\beta + c_i + \bar{u}_{i\cdot}\\ y_{it}-\bar{y}_{i\cdot} &= (x_{it}'-\bar{x}_{i\cdot}')\beta + u_{it}-\bar{u}_{i\cdot}\\ \ddot{y}_{it} &= \ddot{x}_{it}\beta + \ddot{u}_{it}\end{aligned}$$

故有

$$\begin{aligned}\mathrm{E}[\ddot{u}_{it}^2] &= \mathrm{E}\left[u_{it}^2 - 2u_{it}\bar{u}_{i\cdot} + \bar{u}_{i\cdot}^2\right]\\ &= \mathrm{E}\left[u_{it}^2 - 2u_{it}\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}u_{i\cdot} + \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}u_{i\cdot}^2\right]\\ &= \frac{\mathrm{T}-1}{\mathrm{T}}\sigma_u^2\\ \sum_{i=1}^{n}\sum_{t=1}^{\mathrm{T}}\mathrm{E}[\ddot{u}_{it}^2] &= n(\mathrm{T}-1)\sigma_u^2\\ \sigma_u^2 &= \mathrm{E}\left[\sum_{i=1}^{n}\sum_{t=1}^{\mathrm{T}}\frac{\ddot{u}_{it}^2}{n(\mathrm{T}-1)}\right]\end{aligned}$$

修正自由度k后得到σ_u^2的无偏估计为

$$\begin{aligned}\hat{\sigma}_u^2 &= \frac{1}{n(\mathrm{T}-1)-k}\sum_{i=1}^{n}\sum_{t=1}^{\mathrm{T}}\ddot{u}_{it}^2\\ \hat{\beta}^{\mathrm{RE-FGLS}} &= (\mathrm{X}'\hat{\Omega}^{-1}\mathrm{X})^{-1}\mathrm{X}'\hat{\Omega}^{-1}y\\ &= (\sum_{i=1}^{n}x_i'\hat{\Sigma}^{-1}x_i)^{-1}(\sum_{i=1}^{n}x_i\hat{\Sigma}^{-1}y_i)\end{aligned}$$

总结而言，c_i和x_{it}有关时用FE，无关时用RE和FE都可，但是RE更有效（因为RE只是quasi-demean，损失的自由度比FE模型更少）。

对于Nonlinear Panel c_i不容易被消除

总结. 概括而言，几种模型的假设为

- RE1: $\mathrm{E}(u_{it}|x_i,c_i)=0, \mathrm{E}(c_i|x_i)=\mathrm{E}(c_i)=0$; FE1: $\mathrm{E}(u_{it}|x_i,c_i)=0$.
- RE2: $rank(\mathrm{E}(x_i'\Omega^{-1}x_i))=k$; FE2: $rank(\mathrm{E}(\ddot{x}_i'\ddot{x}_i))=k$.
- RE3: $\mathrm{E}(u_iu_i'|x_i,c_i)=\sigma_u^2\mathrm{I}_\mathrm{T}, \mathrm{E}(c_i^2|x_i)=\sigma_c^2$; FE3: $\mathrm{E}(u_iu_i'|x_i,c_i)=\sigma_u^2\mathrm{I}_\mathrm{T}$.

$\mathrm{E}(u_{it}|x_i)\neq 0$即严格外生假设，所有时期都无关，因为估计的时候用了demean处理,所以要求所有数据都无关。当不同假设被违反之后，需要用不同的方式来解决（逐个展开）。

RE3 violation. We could solve it either by *robust estimator* or *GLS*. If RE3 fails,

$$\sqrt{n}(\hat{\beta}^{\mathrm{RE}}-\beta)=(\frac{1}{n}\sum_{i=1}^{n}x_i'\Sigma^{-1}x_i)^{-1}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i'\Sigma^{-1}v_i)$$

where $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i'\Sigma^{-1}v_i\sim \mathrm{N}(0,\mathrm{E}[x_i'\Sigma^{-1}\mathrm{E}(v_iv_i'|x_i)\Sigma^{-1}x_i])$. Hence, the robust variance matrix estimation is

$$\hat{\mathrm{V}ar}(\sqrt{n}\hat{\beta}^{\mathrm{RE}})=(\frac{1}{n}\sum_{i=1}^{n}x_i'\Sigma^{-1}x_i)^{-1}(\frac{1}{n}\sum_{i=1}^{n}x_i'\Sigma^{-1}\hat{v}_i\hat{v}_i'\Sigma^{-1}x_i)(\frac{1}{n}\sum_{i=1}^{n}x_i'\Sigma^{-1}x_i)^{-1}$$

where $\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}\hat{v}_i\hat{v}_i'$. Here both RE and OLS estimations are consistent.

For the Gerernal Feasible Generalized LS (GFGLS),

$$\hat{\beta}^{\mathrm{GFGLS}} = (\sum_{i=1}^{n}x_i'\tilde{\Sigma}^{-1}x_i)^{-1}(\sum_{i=1}^{n})^{-1}(\sum_{i=1}^{n}x_i'\tilde{\Sigma}y_i) \tag{4.7}$$

$$\hat{\mathrm{V}ar}(\sqrt{n}\hat{\beta}^{\mathrm{GFGLS}}) = (\sum_{i=1}^{n}x_i'\tilde{\Sigma}^{-1}x_i)^{-1} \tag{4.8}$$

where $\tilde{=}\frac{1}{n}\sum_{i=1}^{n}\tilde{v}_i\tilde{v}_i'$, $\tilde{v}_i$ is from OLS.

FE3 Violation. If FE3 is violated, we can fix it by *robust estimator* or *GLS*. For the robust estimations, we know

$$\sqrt{n}(\beta^{\mathrm{FE}}-\beta)=(\frac{1}{n}\sum_{i=1}^{n}\ddot{x}_i'\ddot{x}_i)^{-1}(\frac{1}{n}\sum_{i=1}^{n}\ddot{x}_i'\hat{\ddot{u}}_i)$$

Hence, the robust estimator is as follows:

$$\hat{\mathrm{V}ar}(\sqrt{n}\hat{\beta}^{\mathrm{FE}})=(\frac{1}{n}\sum_{i=1}^{n}\ddot{x}_i'\ddot{x}_i)^{-1}(\frac{1}{n}\sum_{i=1}^{n}\ddot{x}_i'\hat{\ddot{u}}_i\hat{\ddot{u}}_i'\ddot{x}_i)(\frac{1}{n}\sum_{i=1}^{n}\ddot{x}_i'\ddot{x}_i)^{-1}$$

FE3本质上要求u_{it}的方差结构位对角阵，所以我们在FEGLS中，假设$\mathrm{E}(u_iu_i'|x_i,\mathrm{C}_i)=\Lambda$，$\Lambda$为任意形式$\mathrm{T}\times\mathrm{T}$的正定矩阵。$\ddot{u}_i=\mathrm{Q}_{\mathrm{T}}u_i$, Q_{T} is demean Transformation.

$$\begin{aligned}
\mathrm{E}(\ddot{u}_i\ddot{u}_i'|\ddot{x}_i) &= \mathrm{E}(\ddot{u}_i\ddot{u}_i')=\mathrm{Q}_{\mathrm{T}}\mathrm{E}(u_iu_i')\mathrm{Q}_{\mathrm{T}}=\mathrm{Q}_{\mathrm{T}}\Lambda\mathrm{Q}_{\mathrm{T}}\overset{def}{=}\Sigma\\
\hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\ddot{u}}_i\hat{\ddot{u}}_i'\\
\hat{\Sigma}^{-\frac{1}{2}}\ddot{y}_i &= \hat{\Sigma}^{-\frac{1}{2}}\sum_{i=1}^{n}\ddot{x}_i\beta+\hat{\Sigma}^{-\frac{1}{2}}\ddot{u}_i\\
\Longrightarrow\hat{\beta}^{\mathrm{FEGLS}} &= (\sum_{i=1}^{n}\ddot{x}_i'\hat{\Sigma}\ddot{x}_i)^{-1}(\frac{1}{n}\sum_{i=1}^{n}\ddot{x}_i'\hat{\Sigma}\ddot{y}_i)\\
\mathrm{V}ar(\sqrt{n}\hat{\beta}^{\mathrm{FEGLS}}) &= (\sum_{i=1}^{n}\ddot{x}_i'\hat{\Sigma}\ddot{x}_i)^{-1}
\end{aligned}$$

由于对$\ddot{u}_i$做了demean处理，$\hat{\Sigma}$是$\mathrm{T}-1\times\mathrm{T}-1$矩阵，去掉了某一个时间的值（不然不是满秩无法求逆）。

RE1 Violation. 我们可以用REIV的方法来客服内生性，定义工具变量$(z_i)_{\mathrm{T}\times\mathrm{L}}$ $(\mathrm{L}>k)$。原RE表达式为

$$(\mathrm{I}_n\otimes\Sigma^{-\frac{1}{2}})\quad \Omega^{-\frac{1}{2}}y=\Omega^{-\frac{1}{2}}\mathrm{X}\beta+\Omega^{-\frac{1}{2}}v$$
$$\Omega^{-\frac{1}{2}}y=\Omega^{-\frac{1}{2}}\mathrm{Z}\beta+\Omega^{-\frac{1}{2}}v\quad (2\mathrm{SLS})$$

现在采用两阶段最小二乘来估计带权重的X。First stage,

$$\begin{aligned}\Omega^{-\frac{1}{2}}\mathrm{X} &= \Omega^{-\frac{1}{2}}\mathrm{Z}\delta+e\\ \Longrightarrow \widehat{\Omega^{-\frac{1}{2}}\mathrm{X}} &= \Omega^{-\frac{1}{2}}\mathrm{Z}(\mathrm{Z}'\Omega^{-1}\mathrm{Z})^{-1}\mathrm{Z}'\Omega^{-1}\mathrm{X}.\end{aligned}\tag{4.9}$$

Second stage,

$$\begin{aligned}\Omega^{-\frac{1}{2}}y &= \widehat{\Omega^{-\frac{1}{2}}\mathrm{X}}\beta+error\\ \hat{\beta}^{\mathrm{REIV}} &= (\mathrm{X}\Omega^{-1}\mathrm{Z}(\mathrm{Z}'\Omega^{-1}\mathrm{Z})^{-1}\mathrm{Z}'\Omega^{-1}\mathrm{X})^{-1}\mathrm{X}'\Omega^{-1}\mathrm{Z}(\mathrm{Z}'\Omega^{-1}\mathrm{Z})^{-1}\mathrm{Z}'\Omega^{-1}y\\ \sqrt{n}(\hat{\beta}^{\mathrm{REIV}}-\beta) &\sim \mathrm{N}\left(0,\left[\frac{\mathrm{X}'\Omega^{-1}\mathrm{Z}}{n}(\frac{\mathrm{Z}'\Omega^{-1}\mathrm{Z}}{n})^{-1}\frac{\mathrm{Z}'\Omega\mathrm{X}}{n}\right]^{-1}\right)\end{aligned}$$

但如何确认该模型中是否还存在内生性问题呢，我们将引入关于内生性的检验（Test of endogeneity）：Control function approach。其思想为，如果把v_{it}放入回归中，如果其系数显著则说明仍然存在内生性问题，如果不显著就外生。所以我们应该先设定总体的模型为

$$y_{it1}=z_{it1}\delta+y_{it2}\alpha_1+y_{it3}\gamma_1+v_{it1}$$

假设$\mathrm{H}_0:\mathrm{E}(y_{is3}|v_{it1})=0,\ \forall s=1,\ldots,\mathrm{T}$，从模型$y_{it3}=z_{it}\Pi_3+v_{it3}\Rightarrow\hat{v}_{it3}$，其中$z_{it}=(\overset{z_{it1}}{exo},\overset{z_{it2}}{iv})$。最后将其代入到原来的表达式中，

$$y_{it1}=z_{it1}\delta+y_{it2}\alpha_1+y_{it3}\gamma_1+\hat{v}_{it3}\rho_1+error$$

若$\rho_1=0$，则y_{it3}外生。

FEIV Violation. 对于模型$y_{it}=x_{it}'\beta+c_i+u_{it}$，我们假设随机扰动项严格外生：$\mathrm{E}(u_{it}|x_i)=0,\ \mathrm{E}(u_{it}|z_i)=0$。

First stage

$$\ddot{x}_i=\ddot{z}_i\delta+error\Rightarrow\hat{\sigma}=(\sum_{i=1}^{n}\ddot{z}_i'\ddot{z}_i)^{-1}\sum_{i=1}^{n}\ddot{z}_i'\ddot{x}_i$$

Second stage, plugging $\ddot{\hat{x}}_i = \ddot{z}_i\hat{\sigma}$, we have

$$\begin{aligned}\ddot{y}_i &= \ddot{\hat{x}}_i\beta + error \\ \hat{\beta}^{\text{FEIV}} &= (\sum_{i=1}^{n}\ddot{\hat{x}}_i'\ddot{\hat{x}}_i)^{-1}\sum_{i=1}^{n}\ddot{\hat{x}}_i'\ddot{y}_i \\ \sqrt{n}(\hat{\beta}^{\text{FEIV}} - \beta) &\sim \text{N}(0, \sigma^2[(\text{E}\ddot{x}_i'\ddot{z}_i)(\text{E}\ddot{z}_i'\ddot{z}_i)^{-1}(\text{E}\ddot{z}_i'x_i)])\end{aligned}$$

Test of endogeneity

$$y_{it3} = z_{it}\Pi_3 + v_{it3} \Rightarrow \hat{v}_{it3} \quad z_{it} \tag{4.10}$$

$$\text{FEIV}(z_{it}, y_{it3}, \hat{v}_{it3}) \tag{4.11}$$

First Difference. 最后，补充一种解决c_i与x_{it}有关的方法，叫做一阶差分（First Difference, FD）。同样的，对于一阶差分而言，我们有三个重要的假设以保证估计量的性质：

- FD1: $\text{E}(u_{it}|x_i, c_i) = 0$.
- FD2: $rank(\text{E}(\Delta x_i'\Delta x_i)) = k$.
- FE3: $\text{E}(e_ie_i'|x_i, c_i) = \sigma^2\text{I}_\text{T}$.

其中，$e_{i,t} \overset{def}{=} u_{i,t} - u_{i,t-1}$。$e_{it}$序列无关则意味着$u_{i,t}$是随机游走，具有真实的序列相关性。假设模型如下

$$\begin{aligned}y_{it} &= x_{it}'\beta + c_i + u_{it}\ t = 1, \ldots, \text{T} \\ \Delta y_{it} &= \Delta x_{it}'\beta + \Delta u_{it}\ t = 2, \ldots, \text{T}\end{aligned}$$

一阶差分估计量$\hat{\beta}^{\text{FD}}$是上式回归的混合OLS估计量，其在FD1下是一致的。FD模型的严格外生条件由$\text{E}[\Delta u_{it}|\Delta x_{it}] = 0$保证，其仅要求$\text{E}[u_{it}|x_{i,t-1}, x_{i,t}, x_{i,t+1}] = 0$。回到上述模型，我们可得

$$\begin{aligned}\widehat{\beta}^{\text{FD}} &= \hat{\sigma}_e^2(\Delta x'\Delta x)^{-1} \\ \text{where } \hat{\sigma}_e^2 &= \frac{\sum_{i=1}^{n}\sum_{t=2}^{\text{T}}\hat{e}_{it}^2}{n\text{T} - n - k} \\ \text{and } \hat{e}_{it}^2 &= \Delta y_{it} - \Delta x_{it}\hat{\beta}^{\text{FD}}\end{aligned}$$

FD3 Violation. Similarily, we could utilize the robust estimator as follows:

$$\widehat{\text{V}ar(\hat{\beta}^{\text{FD}})} = (\Delta x'\Delta x)^{-1}(\sum_{i=1}^{n}\Delta x'\hat{e}_{it}\hat{e}_{it}'\Delta x)(\Delta x'\Delta x)^{-1}$$

FD1 Violation. We still need the instrument variable here. However, how do we apply an IV-approach here? If we have access to a great set of IV satisfying FEIV1, we can replicate the exercises in FEIV1. Actually, FEIV1 is to much for FDIV1, which means that we don't need a full IV series matched to x_{it} ($\mathrm{E}[Z_{it}'u_{it}]=0$). FDIV1's requirements on IV are matched to the difference series ($\mathrm{E}[w_{it}'\Delta u_{it}]=0$).

Hence, we define IV as $w_i = diag(w_{i2}, w_{i3}, \ldots, w_{i\mathrm{T}})_{(\mathrm{T}-1)\times \mathrm{L}}$ and use w_{it} as IV. $\mathrm{E}(w_{it}'\Delta u_{it})=0$:

$$\mathrm{E}\left[\begin{pmatrix} w_{i2}' & & \\ & \ddots & \\ & & w_{i\mathrm{T}}' \end{pmatrix}\begin{pmatrix} \Delta u_{i2} \\ \vdots \\ \Delta u_{i\mathrm{T}} \end{pmatrix}\right]=0$$

我们可以对每个时间t都找一个工具变量来预测内生变量，且不需要严格外生假定。接下来，假设FDIV2：$rand(\mathrm{E}[w_i'w_i])=\mathrm{L}$ and $rand(\mathrm{E}[w_i'\Delta x_i])=k$.

System 2SLS. 首先，我们先做$\mathrm{T}-1$个第一阶段回归，得到$\widehat{\Delta x_i}$：

$$\left.\begin{aligned} \widehat{\Delta x_{it}}' &\leftarrow \Delta x_{it}' \quad \overset{\mathrm{IV}}{\leftarrow} w_{it}, \quad i=1,\ldots,n \\ \widehat{\Delta x_{it-1}}' &\leftarrow \Delta x_{it-1}' \quad \overset{\mathrm{IV}}{\leftarrow} w_{it-1}, \quad i=1,\ldots,n \\ &\vdots \end{aligned}\right\} \text{T-1 seperate OLS}$$

然后再把预测值代入原模型做Second stage regression，得到最后的估计量。这一流程叫做System 2SLS。

如前所属，这里的工具变量并不需要满足严格外生性，仅满足序列外生性（Sequential Exogeneity）即可（$\mathrm{E}[w_{it}'\Delta u_{it}]=0$）。但问题的关键往往并不是工具变量法的估计，而是如何寻找到一个合适的工具变量w_{it}。

第五章　Nonlinear Model

本章回到了简单的截面回归中来考虑一些典型的非线性回归。

5.1　离散选择模型

离散选择模型主要分为二元选择模型（binary choice models）和多元选择模型（multiple choice models），其中多元选择又分为有序（ordered）和无序的（unordered）。

5.1.1　Binary Choice Model

Linear Probability Model.　对于线性概率模型（LPM）而言，其特点是被解释变量$y=0$ or 1两个取值，其模型为

$$\begin{aligned} y_i &= x_i'\beta+\varepsilon_i \text{ with} \mathrm{E}(\varepsilon_i|x_i)=0 \\ \mathrm{E}[y_i|x_i] &= x_i'\beta = 1\times \mathrm{P}(y_i=1|x_i)+\cancel{0\times \mathrm{P}(y_i=0|x_i)} \\ \mathrm{V}ar(\varepsilon_i|x_i) &= \mathrm{E}(\varepsilon_i^2|x_i)=(1-x_i'\beta)^2x_i\beta+(-x_i\beta)^2(1-x_i\beta) \\ &= (1-x_i'\beta)x_i'\beta \end{aligned}$$

该模型潜在的问题是，所拟合出的概率可能会小于0或大于1，同时$\mathrm{V}ar(\varepsilon_i|x_i)$一定存在异方差问题。

Random Utility Model.　此处我们将y_i建模为基于一定标准做决策的行为，若一件事情的效用高于另一件事情，例如，如果认为在某人的决策过程中，若租房的效用（U_i^{rent}）高于买房的效用（U_i^{buy}），则我们认为$y_i=1$。故，模型演变为对效用的回归：

$$\begin{aligned} \mathrm{U}_i^{rent} &= x_i'\beta_r+\varepsilon_{ir} \\ \mathrm{U}_i^{buy} &= x_i'\beta_b+\varepsilon_{ib} \\ \mathrm{U}_i^{rent}-\mathrm{U}_i^{buy}=y_i^* &= x_i'\beta+\varepsilon_i \end{aligned}$$

y_i^* is latent variable and if $\mathrm{U}_i^{rent} - \mathrm{U}_i^{buy} > 0$ $y_i^* = 1$ otherwise $y_i^* = 0$.

$$\begin{aligned} \mathrm{P}(y_i = 1|x_i) &= \mathrm{P}(y_i^* > 0|x_i) \\ &= \mathrm{P}(x_i'\beta + \varepsilon_i > 0|x_i) \\ &= \mathrm{P}(\varepsilon_i > -x_i'\beta|x_i) \\ \text{如果}\varepsilon_i\text{是对称分布} &= \underset{\text{CDF}}{\mathrm{F}}(x_i'\beta) \end{aligned}$$

此时，ε的分布是未知的，我们仅用抽象表达式代替，而模型中的参数仅包括待估计的β，整个模型是半参数模型（Semi-parametrix）。直接而言，我们可以让数据自己产生分布，只假设β, 对ε非参假设。但如果对ε的分布进行不同的假设我们可以得到不同的模型，例如

- Probit model: $\varepsilon_i \sim \mathrm{N}(0,1)$, $\mathrm{P}(y_i = 1|x_i) = \Phi(x_i'\beta)$;
- Logit model: $\mathrm{P}(y_i = 1|x_i) = \Lambda(x_i'\beta) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$, $\mathrm{P}(y_i = 0|x_i) = 1 - \Lambda(x_i'\beta)$.

通常来说，我们用极大似然法来做参数估计：

$$\begin{aligned} \mathrm{P}(y_i|x_i,\beta) &= \left[\mathrm{F}(x_i'\beta)^{y_i}[1-\mathrm{F}(x_i'\beta)^{1-y_i}]\right] \\ \ln\mathscr{L}(\beta|x',y') &= \sum_{i=1}^{n}\left[y_i ln\mathrm{F}(x_i'\beta) + (1-y_i)ln(1-\mathrm{F}(x_i'\beta))\right] \\ \frac{\partial\mathscr{L}(\beta|x',y')}{\partial\beta} &= \sum_{i=1}^{n}\left[\frac{y_i f_i}{\mathrm{F}(x_i'\beta)} - \frac{(1-y_i)f_i}{1-\mathrm{F}(x_i'\beta)}\right]x_i \end{aligned}$$

Probit Model with Continous Endogenous Variable. 由于两阶段最小二乘法（2SLS）只适用于线性模型，在Random Utility Model的框架下并不适用，但可以在前述的LPM的框架下使用2SLS。本节我们以Probit模型为例，介绍非线性模型中使用**工具变量的两步法**。

首先定义$y_1 = 1_{\{y_1^*>0\}}$，z_1为外生变量，y_2为内生变量（z_2为对应的工具变量），且$u \sim \mathrm{N}(0,1)$，有模型如下：

$$\begin{aligned} y_1^* &= z_1\delta_1 + \alpha_1 y_2 + u_1 \\ \text{First step: } y_2 &= z_1\delta_{21} + z_2\delta_{22} + v_2 = z\delta_2 + v_2 \Rightarrow \hat{v}_2 \\ \text{Second step: } y_1^* &= z_1\delta_1 + \alpha_1 y_2 + \theta_1\hat{v}_2 + e_1 \\ \text{where (assume) } v_2 &\sim \mathrm{N}(0,\tau_2^2) \\ \text{then } \theta_1 &= \frac{\mathrm{C}ov(u_1,v_2)}{\mathrm{V}ar(v_2)} \overset{def}{=} \frac{\eta_1}{\tau_2^2} \end{aligned} \tag{5.1}$$

此时e_1尽管服从正态分布，但并不是标准正态分布。Recall $e_1 = u_1 - \theta v_2$, then

$$\begin{aligned}
\mathrm{E}[e_1] &= \mathrm{E}[u_1 - \theta_1 v_2] = 0 \ (\mathrm{E}[u_1] = \mathrm{E}[v_2] = 0) \\
(1 =)\mathrm{V}ar(u_1) &= \theta_1^2 \mathrm{V}ar(v_2) + \mathrm{V}ar(e_1) \\
\Longrightarrow \mathrm{V}ar(e_1) &= 1 - \frac{\eta_1^2}{\tau_2^2} \overset{def}{=} 1 - \rho_1^2, \text{ where } \rho_1 = \frac{cov(u_1, v_2)}{\sqrt{var(u_1)var(v_2)}} = \frac{\eta_1}{\tau_2} \\
e_1 &\sim \mathrm{N}(0, 1 - \rho_1^2)
\end{aligned}$$

而对于Equation 5.1而言，需要要求残差项$e_1 \sim \mathrm{N}(0,1)$，才满足Probit模型的定义，故此时我们有

$$\mathrm{P}(y_1 = 1 | z, y_2, \hat{v}_2) = \Phi\left(\frac{z_1\delta_1 + \alpha_1 y_2 + \theta_1 \hat{y}_2}{\sqrt{(1 - \hat{\rho}_1^2)}}\right)$$

where $\rho_1 = \theta_1 \tau_2$.总结而言，工具变量两步法旨在将内生变量与残差项相关的部分从残差项中提取出来，进而使用新的残差项进项回归。

另一方面，我们也可以直接使用最大似然估计来克服内生性的问题（不再进行分步回归）。

$$\begin{aligned}
f(y_1, y_2 | z) &= f(y_1 | y_2, z) f(y_2 | z) \text{ continuous case} \\
&= \mathrm{P}(y_1 | y_2, z) f(y_2 | z) \text{ discrete case} \\
\mathrm{P}(y_1 = 1 | y_2, z) &= \Phi\left(\frac{z_1\delta_1 + \alpha_1 y_2 + \frac{\rho_1}{\tau_2}(y_2 - z\delta_2)}{\sqrt{1 - \rho_1^2}}\right) = \Phi(w)
\end{aligned}$$

由前可知，y_2是均值为$z\delta$、方差为V_2的正态分布，故有

$$\begin{aligned}
f(y_2 | z) &= \frac{1}{\sqrt{2\pi}\tau} e^{\frac{(y_2 - z\delta_2)^2}{2\tau_2^2}} \\
\Rightarrow f(y_1, y_2 | z) &= \Phi(w)^{y_1} (1 - \Phi(w))^{1 - y_1} \frac{1}{\tau_2} \phi\left(\frac{y_2 - z\sigma_2}{\tau_2}\right)
\end{aligned}$$

Probit Model with Discrete Endogenous Variable. 当内生变量是离散时，工具变量两步法会由于无法估计出$\hat{v}_2$而失效，此时我们只能使用极大似然法来处理。同之前一样，MLE最核心的思想是

$$f(y_1, y_2 | z) = f(y_1 | y_2, z) f(y_2 | z)$$

对于此时的模型为：$y_1 = 1_{\{z_1\delta_1 + \alpha_1 y_2 + u_1 > 0\}}$，$y_2 = 1_{\{z\delta_2 + v_2 > 0\}}$，故极大似然的表达式化简为

$$\mathrm{P}(y_1 = 1, y_2 = 1) = \mathrm{P}(y_1 = 1 | y_2 = 1, z) \mathrm{P}(y_2 = 1 | z)$$

可得

$$\begin{aligned}
\mathrm{P}(y_1=1,y_2=1) &= \mathrm{P}(y_1=1|y_2=1,z)\underbrace{\mathrm{P}(y_2=1|z)}_{\text{From Probit}} \\
\mathrm{P}(y_1=1|y_2=1,z) &= \mathrm{E}\left[\mathrm{P}(y_1=1|v_2,z)|y_2=1,z\right] \text{ 只是}v_2\text{的函数} \\
(y_2=1_{\{v_2>-z\delta_2\}} &\Rightarrow \mathrm{E}[y_2]=1=\int_{-z\delta_2}^{\infty}\Phi(\cdot)\mathrm{d}v_2) \\
(\text{Integral of } v_2) &= \int_{-z\delta_2}^{\infty}\mathrm{P}(y_1=1|v_2,z)f(v_2|y_2=1,z)\mathrm{d}v_2 \\
&= \int_{-z\delta_2}^{\infty}\Phi\left(\frac{z_1\delta_1+\alpha_1 y_2+\rho_1 v_2}{\sqrt{1-\rho_1^2}}\right)\frac{\phi(v_2)}{\mathrm{P}(v_2>-z\delta_2)}\mathrm{d}v_2 \\
&= \frac{1}{\Phi(z\delta_2)}\int_{-z\delta_2}^{\infty}\Phi\left(\frac{z_1\delta_1+\alpha_1 y_2+\rho_1 v_2}{\sqrt{1-\rho_1^2}}\right)\phi(v_2)\mathrm{d}v_2
\end{aligned} \tag{5.2}$$

类似地，可得

$$\begin{aligned}
\mathrm{P}(y_i=0|y_2=1,z) &= 1-\mathrm{P}(y_1=1|y_2=1,z) \\
&= \frac{1}{1-\Phi(z\delta_2)}\int_{-\infty}^{-z\delta_2}\Phi\left(\frac{z_1\delta_1+\alpha_1 y_2+\rho_1 v_2}{\sqrt{1-\rho_1^2}}\right)\phi(v_2)\mathrm{d}v_2
\end{aligned}$$

Something interesting is when we plug the integral form in the MLE:

$$\begin{aligned}
\mathrm{P}(y_1=1,y_2=1|z) &= \mathrm{P}(y_1=1|y_2=1,z)\mathrm{P}(y_2=1|z) \\
&= \int_{-z\delta_2}^{\infty}\int_{-z_1\delta_1-\alpha_1 y_2}^{\infty}\frac{1}{2\pi\sqrt{1-\rho_1^2}}e^{-\frac{1}{2(1-\rho_1^2)}[u_1^2-2\rho_1 u_1 u_2+v_2^2]}
\end{aligned} \tag{5.3}$$

该形式为二元正态分布的分布函数（bivariate normal density）。在Stata中, IV probit命令适用于连续变量，离散内生变量应该用biprobit解决内生性问题。

Bivariate Probit Model. [1] 对二元Probit模型（两个y），我们通常设定其为

$$\begin{aligned}
y_1 &= 1[x_1\beta_1+e_1>0] \\
y_2 &= 1[x_2\beta_2+e_2>0] \\
\begin{pmatrix}e_1\\e_2\end{pmatrix} &\sim \mathrm{N}(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix})
\end{aligned}$$

- $\rho=0$：如果误差项之间不相关，那么我们仍可以用分开的Probit模型来处理该任务；

[1] 二元Probit模型的情况已经很接近目前利用机器学习解决的诸多任务了，比如常见的垃圾邮件分类等等，但只是这里的所使用的模型结构尚不能处理自变量维度很大的情况，但其设定出来的二元因变量情形已经比较有意思了。

- $\rho \neq 0$：此时我们应当执行的bivariate Probit（Equation 5.3），否则所得估计量尽管是一致的但并不有效。

5.2 Truncated Data

截断数据（Truncated Data）是在没有其他影响的时候，由于受到部分外生因素影响而我们并不能观测到部分缺失的x，导致其经验分布与总体分布不一致。假设此时，当$x < a$时，我们并不能观察到y的数据，因此本节的目的就是尝试用$x > a$的观测来推断全样本的结果（在假设了x本来分布的情况下）。

$$f(x|x>a) = \frac{f(x)}{\mathrm{P}(x>a)} \Leftarrow \text{条件密度} = \frac{\text{联合密度}}{\text{边际密度}} \tag{5.4}$$

where we assume that $x \sim \mathrm{N}(\mu, \sigma^2)$. So we have $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ and $\mathrm{P}(x>a) = 1-\Phi(\frac{a-\mu}{\sigma})$.

$$\begin{aligned}
f(x|x>a) &= \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{1-\Phi(\frac{a-\mu}{\sigma})} \\
\mathrm{E}[x|x>a] &= \int_a^\infty x f(x|x>a)\mathrm{d}x \\
&= \int_a^\infty x \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{1-\Phi(\frac{a-\mu}{\sigma})}\mathrm{d}x \\
(def) \quad & (\frac{x-\mu}{\sigma} \overset{def}{=} v,\ \frac{a-\mu}{\sigma} \overset{def}{=} \alpha) \\
&= \int_\alpha^\infty (\sigma v+\mu)\frac{\phi(v)}{1-\Phi(\alpha)}\mathrm{d}v \\
(-\phi'(x) = x\phi(x)) \quad &= \frac{\sigma}{1-\Phi(\alpha)}\int_\alpha^\infty (-\phi'(v))dv + \frac{\mu}{1-\Phi(\alpha)}\int_\alpha^\infty \phi(v)\mathrm{d}v \\
&= \frac{\sigma}{1-\Phi(\alpha)}\left[-\phi(v)|_\alpha^\infty\right] + \frac{\mu}{1-\Phi(\alpha)}\Phi(v)|_\alpha^\infty \\
&= \underbrace{\mu}_{\text{总体均值}} + \underbrace{\sigma\frac{\phi(\frac{a-\mu}{\sigma})}{1-\Phi(\alpha)}}_{\text{已观测数据调整项}}
\end{aligned}$$

现在，代入我们已知数据进行回归，有$y_i = x_i'\beta + \varepsilon_i$, $\varepsilon_i \sim \mathrm{N}(0,\sigma^2)$，$y_i|x_i \sim \mathrm{N}(x_i'\beta, \sigma^2)$。

$$\begin{aligned}
\mathrm{E}[y_i|y_i>a] &= x_i'\beta + \sigma\frac{\phi(\frac{a-x_i'\beta}{\sigma})}{1-\Phi(\frac{a-x_i'\beta}{\sigma})} \\
\text{Likelihood in MLE: } \mathscr{L}(\beta,\sigma;x,y) &= \prod_{i=1}^n \frac{\frac{1}{\sigma}\phi(\frac{y_i-x_i'\beta}{\sigma})}{1-\Phi(\frac{a-x_i'\beta}{\sigma})}
\end{aligned}$$

第六章　Non Paramtric Model

非参数模型基本上放弃了有关函数形式和分布的一切固有假定，极其有限地限定了数据的结构。其较少的设定通常并不能提供十分精确的推断（除了在样本充分大的情况下），但使用非参数方法所得到的信息往往极为稳健。

6.1　Kernel Function

下面将用一个例子来介绍非参数估计的基本知识点。估计一枚硬币是否均匀。记n此抛硬币的结果为$(x_1, x_2, \cdots, x_n)$。记$\mathrm{P(H)} = \mathrm{P}(x_i \leq x) = \mathrm{F}(x)$，有

$$\hat{\mathrm{P}}(\mathrm{H}) = \frac{\#\ of\ heads}{n} = \frac{1}{n}\{\ of\ x_i \leq\} = \mathrm{F}_n(x)$$

直观而言，若某处附近的小区间内点越多，则代表该处的密度越高（核函数，kernel function）。理论上而言，$f(x)$真值满足

$$\begin{aligned} \text{In theory}, f(x) &= \frac{\mathrm{dF}(x)}{\mathrm{d}x} = \lim_{h\to 0}\frac{\mathrm{F}(x+h)-\mathrm{F}(x-h)}{2h} = \lim_{h\to 0}\frac{\mathrm{P}(x-h \leq \mathrm{X} \leq x+h)}{2h} \\ \text{Empirically}, \hat{f}(x) &= \lim_{h\to 0}\frac{\{\#\text{ of } x_i'\text{s falling in the interval [x-h,x+h]}\}}{2hn} \end{aligned}$$

为衡量点是否落在区间内，我们定义核函数的表达形式如下：

$$\text{Def } k(\frac{x_i - x}{h}) = k(z) = \begin{cases} \frac{1}{2} & if|z| < 1 \\ 0 & \text{otherwise} \end{cases}$$

所以有

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{h} k(\frac{x_i - x}{h})$$

但基于邻域定义的核函数会在很大程度上受到带宽h的影响，从而导致错误地计数落在该邻域的数据

点。为改进这一缺点，另一种思想是为离中心越远的点赋予更小的权重以规避选择h的问题，即令$k(v)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}v^2}, -\infty<v<\infty$。

A 2^{nd} order kernel function is defined if it satisfies

$$\begin{cases} \int k(v)dv = 1 \quad k(v): \text{奇函数} \\ \int vk(v)dv = 0 \quad vk(v): \text{奇函数} \\ \int v^2k(v)dv = k_2 > 0 (< \infty) \end{cases}$$

A 4^{th} order kernel function is defined if it satisfies

$$\begin{cases} \int k(v)dv = 1 \\ \int vk(v)dv = 0 \\ \int v^2k(v)dv = 0 \\ \int v^3k(v)dv = 0 \\ \int v^4k(v)dv = k_4 \neq 0 \end{cases}$$

回到基于邻域定义的核函数，我们要证明$\hat{f}(x)$是$f(x)$的一致估计，通常并不直接证明$\hat{f}(x) \xrightarrow{p} f(x)$，而是证明$\hat{\text{MSE}}(x) \xrightarrow{p} f(x)$（因为前面的不好证）。

$$\begin{aligned}
\hat{f}(x) &= \frac{1}{nh}\sum_{i=1}^{h} k(\frac{x_i - x}{h}) \\
\text{E}\hat{f}(x) &= \frac{1}{nh}\sum_{i=1}^{n} \text{E}\left[k(\frac{x_i - x}{h})\right] = \frac{1}{h}\text{E}\left[k(\frac{x_i - x}{h})\right] \\
\text{E}\left[k(\frac{x_i - x}{h})\right] &= \int f(x_i)k(\frac{x_i - x}{h})dx_i \quad (\frac{x_i - x}{h} = v, x_i = x + hv) \\
&= \int f(x+hv)k(v)hdv \\
(\text{T}aylor\ \text{E}xpansion) \quad &= \int [f(x) + f'(x)hv + \frac{1}{2}f''(x)h^2x^2 + o(h^2)]k(v)hdv \\
&= hf(x) + \frac{h^3}{2}f''(x)\int v^2k(v)dv + o(h^3) \quad (\int k(v)\text{d}v = 1, \int vk(v)\text{d}v = 0) \\
\text{E}[\hat{f}(x)] &= f(x) + \frac{h^2}{2}f''(x)k_2 + o(h^2) \quad (\int v^2k(v)dv = k_2) \\
\text{Bias: E}[\hat{f}(x)] - f(x) &= \frac{h^2}{2}f''(x)k_2 + o(h^2)
\end{aligned}$$

又有

$$\begin{aligned}
\mathrm{V}ar(\hat{f}(x)) &= \frac{1}{n^2h^2}n\mathrm{V}ar\left(k(\frac{x_i-x}{h})\right) \\
&= \frac{1}{nh^2}\left\{\mathrm{E}\left[k^2\left(\frac{x_i-x}{h}\right)\right]-\left[\mathrm{E}\left[k\left(\frac{x_i-x}{h}\right)\right]\right]^2\right\}
\end{aligned}$$

其中，有

$$\begin{aligned}
\mathrm{E}\left[k^2\left(\frac{x_i-x}{h}\right)\right] &= \int f(x_i)k^2(\frac{x_i-x}{h})dx \\
&= \int f(x+hv)k^2(v)hdv \\
&= \int [f(x)+f'(x)hv+\frac{f''(x)}{2}h^2v^2+o(h^2)]k^2(v)hdv \\
&= hf(x)\int k^2(v)dv+o(h)
\end{aligned}$$

所以有，

$$\begin{aligned}
var(\hat{f}(x)) &= \frac{1}{nh^2}\left\{hf(x)\int k^2(v)dv+o(h)-\left[f(x)+\frac{h^2}{2}f''(x)k_2+o(h^2)\right]^2\right\} \\
&= \frac{1}{nh}f(x)\int k^2(v)dv+o(\frac{1}{nh}) \\
&= \frac{1}{nh}f\mathrm{K}+o(\frac{1}{nh}),\quad \mathrm{K}=\int k^2(v)dv
\end{aligned}$$

回到MSE，有

$$\begin{aligned}
\mathrm{MSE}\hat{f}(x) &= \mathrm{Bias}^2\hat{f}(x)+\mathrm{Var}\hat{f}(x) \\
&= \frac{h^4}{4}(k_2f'')^2+\frac{f\mathrm{K}}{nh}+o(h^4+\frac{1}{nh})
\end{aligned}$$

上面展示了MSE其实是带宽的一个函数，带宽的选取是在bias和variance之间的权衡。我们可以用一阶条件来确定最优带宽：

$$\begin{aligned}
\frac{\partial \mathrm{MSE}\hat{f}(x)}{\partial h}=0 \quad &\Rightarrow \quad h'_{opt}=\left[\frac{\mathrm{K}f(x)}{(k_2f''(x))^2}\right]^{\frac{1}{5}}n^{-\frac{1}{5}} \quad \text{a function of } x_i \\
\text{One for all: } \int \mathrm{MSE}\hat{f}(x)\mathrm{d}x &= \frac{h^4}{4}k_2^2\int f''(x)^2\mathrm{d}x+\frac{\mathrm{K}}{nh}\int f(x)\mathrm{d}x+\mathrm{O}(h^4+\frac{1}{nh}) \\
\text{FOC for h: } \frac{\partial\int \mathrm{MSE}\hat{f}(x)\mathrm{d}x}{\partial h} &= h^3k_2^2\int f''(x)^2\mathrm{d}x-\frac{\mathrm{K}}{nh^2}=0 \\
\Longrightarrow h_{opt} &= \left[\frac{\mathrm{K}}{(k_2f''(x))^2}\right]^{\frac{1}{5}}n^{-\frac{1}{5}}=\mathrm{C}n^{-\frac{1}{5}}
\end{aligned}$$

对于带宽的确定，通常有几种方法（此处就不在展开了）：

- Rule of thumb: $h_{opt} \approx 1.06\sigma n^{-\frac{1}{5}}$, assume $v \sim \mathrm{N}(0,1)$, $x \sim \mathrm{N}(\mu, \sigma^2)$;
- Plug-in method;
- Likelihood Cross-Validation;
- Least Square Cross-Validation.

高维. 对于高维概率密度的核估计，我们可以用联合的核函数：

$$
\begin{aligned}
\hat{f}(x_1,x_2,\ldots,x_q) &= \frac{1}{nh_1,h_2,\ldots,h_q}\sum_{i=1}^{n} k(\frac{x_{i1}-x_1}{h_1}\frac{x_{i2}-x_2}{h_2}\ldots\frac{x_{iq}-x_q}{h_q}) \\
\mathrm{E}[\hat{f}] &= \frac{1}{h_1\ldots h_q}\mathrm{E}\left[k(\frac{x_{i1}-x_1}{h_1})\ldots k(\frac{x_{iq}-x_q}{h_q})\right] \\
\text{where } \mathrm{E}\left[k(\frac{x_{i1}-x_1}{h_1})\ldots k(\frac{x_{iq}-x_q}{h_q})\right] &= \int \frac{x_{i1}-x_1}{h_1}\ldots\frac{x_{iq}-x_q}{h_q} f(x_{i1},\ldots,x_{iq})dx_{i1},\ldots,dx_{iq} \\
\text{变量代换：} &= \int k(v_1)\ldots k(v_q)f(x_1+h_1v_1,\ldots,x_q+h_qv_q)h_1\ldots h_q dv_1\ldots dv_q \\
\text{多维泰勒展开：} &= \int k(v_1)\ldots k(v_q)\{f(x_1,\ldots,x_q)+\sum_{s=1}^{q} h_s v_s f_s + \frac{1}{2}\sum_s\sum_t h_s h_t v_s v_t f_{st} + (s.o.)\}h_1\ldots h_q dv_1\ldots dv_q \\
&= h_1\ldots h_q f(x_1,\ldots,x_q) + (h_1\ldots h_q)\frac{\sum_s h_s^2 f_{ss}}{2}\int v_s^2 k(vs)ds + (s.o.) \\
\mathrm{E}[\hat{f}] &= f + \underbrace{\frac{\sum_s h_s^2 f_{ss}}{2}\mathrm{K}_2 + (s.o.)}_{\text{Bias}} \\
\mathrm{V}ar(\hat{f}) &= \frac{1}{nh_1^2\ldots h_q^2}\{\mathrm{E}k^2(\frac{x_{i1}-x_1}{h_1})\ldots k^2(\frac{x_{iq}-x_q}{h_q})[\mathrm{E}k(\frac{x_{i1}-x_1}{h_1})\ldots k(\frac{x_{iq}-x_q}{h_q})]\} \\
&= \frac{1}{nh_1\ldots h_q} f(x_1,\ldots,x_q)(\int k^2(v_i)dv)^q + (s.o.) \\
\mathrm{MSE}\hat{f} &= \mathrm{E}[\hat{f}]^2 + \mathrm{V}ar(\hat{f})
\end{aligned}
$$

核函数阶数. 在实际过程中，使用2^{nd}、4^{th}、6^{th}、... 阶核函数是有区别的，通常我们假设2阶核函数。

6.2 Kernel Regression.

介绍完基本的预备知识，我们在本节正式进入非参数估计的估计部分。下面将以单变量回归y =

$g(x)+\varepsilon$为例说明：

$$
\begin{aligned}
g(x) &= \mathrm{E}(y|x)=\int yf(y|x)dy=\int y\frac{f(y,x)}{f(x)}dy \\
\hat{g}(x) &= \frac{\int y\hat{f}(y,x)dy}{\hat{f}(x)}=\frac{\hat{m}(x)}{\hat{f}(x)} \\
\text{where } \hat{f}(x) &= \frac{1}{nh}\sum_{i=1}^{n}k(\frac{x_i-x}{h}) \\
\hat{m}(x) &= \int y\frac{1}{nh_xh_y}\sum_{i=1}^{n}k(\frac{x_i-x}{h_x})k(\frac{y_i-y}{h_y})dy \\
&= \frac{1}{nh_xh_y}\sum_{i=1}^{n}k(\frac{x_i-x}{h_x})\int yk(\frac{y_i-y}{h_y})dy \\
(y_i=y+h_yv)\int yk(\frac{y_i-y}{h_y})dy &= \int_{+\infty}^{-\infty}(y_i-h_yv)k(v)(-h_y)dv \\
&= \int_{-\infty}^{+\infty}(y_i-h_yv)k(v)(h_y)dv \\
&= h_y\int_{-\infty}^{+\infty}(y_i-h_yv)k(v)dv=h_yy_i \\
\text{hence } \hat{m}(x) &= \frac{1}{nh_x}\sum_{i=1}^{n}y_ik(\frac{x_i-x}{h_x}) \\
\hat{g}(x) &= \frac{\frac{1}{nh}\sum_{i=1}^{n}y_ik(\frac{x_i-x}{h_x})}{\frac{1}{nh}\sum_{i=1}^{n}k(\frac{x_i-x}{h})}=\sum_{i=1}^{n}y_i\frac{k(\frac{x_i-x}{h_x})}{\sum_{i=1}^{n}k(\frac{x_i-x}{h})}=\sum_{i=1}^{n}y_iw_i
\end{aligned}
$$

该权重通过x_i在空间中的分布（经验分布，非假定的分布）对y进行赋值，从而输出y_i的期望。最后，我们在MSE框架下检验该估计的效率。计算MSE有

$$
\begin{aligned}
\hat{g}(x) &= \frac{\frac{1}{nh}\sum_{i=1}^{n}y_ik(\frac{x_i-x}{h_x})}{\frac{1}{nh}\sum_{i=1}^{n}k(\frac{x_i-x}{h})}=\sum_{i=1}^{n}y_i\frac{k(\frac{x_i-x}{h_x})}{\sum_{i=1}^{n}k(\frac{x_i-x}{h})}=\frac{\hat{m}(x)}{\hat{f}(x)} \\
&= \frac{\mathrm{E}\hat{m}+\hat{m}-\mathrm{E}\hat{m}}{\mathrm{E}\hat{f}+\hat{f}-\mathrm{E}\hat{f}}=\frac{\mathrm{E}\hat{m}+\hat{m}-\mathrm{E}\hat{m}}{\mathrm{E}\hat{f}[1+\frac{\hat{f}-\mathrm{E}\hat{f}}{\mathrm{E}\hat{f}}]} \\
&= \frac{\mathrm{E}\hat{m}+\hat{m}-\mathrm{E}\hat{m}}{\mathrm{E}\hat{f}}\left[1-\frac{\hat{f}-\mathrm{E}\hat{f}}{\mathrm{E}\hat{f}}+(\frac{\hat{f}-\mathrm{E}\hat{f}}{\mathrm{E}\hat{f}})^2-(\frac{\hat{f}-\mathrm{E}\hat{f}}{\mathrm{E}\hat{f}})^3+\dots\right] \\
&= \frac{\mathrm{E}\hat{m}}{\mathrm{E}\hat{f}}+\frac{\hat{m}-\mathrm{E}\hat{m}}{\mathrm{E}\hat{f}}-\frac{\mathrm{E}\hat{m}(\hat{f}-\mathrm{E}\hat{f})}{(\mathrm{E}\hat{f})^2}-\frac{(\hat{m}-\mathrm{E}\hat{m})(\hat{f}-\mathrm{E}\hat{f})}{(\mathrm{E}\hat{f})^2}+\frac{\mathrm{E}\hat{m}(\hat{f}-\mathrm{E}\hat{f})^2}{\mathrm{E}(\hat{f})^3}+\dots
\end{aligned}
$$

观察上式，可知要求得$\hat{g}(x)$，我们应该求得$\hat{m}$的均值和方差，还有和$\hat{f}$的协方差。

均值.

$$
\begin{aligned}
\mathrm{E}[\hat{m}] &= \mathrm{E}\{\mathrm{E}(\hat{m}(x)|x_i)\} = \mathrm{E}\{\mathrm{E}(\frac{1}{nh}\sum y_i k(\frac{x_i-x}{h})|x_i)\} \\
&= \mathrm{E}\{\frac{1}{nh}\sum \mathrm{E}(y_i|x_i)k(\frac{x_i-x}{h})\} \\
&= \mathrm{E}\{\frac{1}{nh}\sum g(x_i)k(\frac{x_i-x}{h})\} = \frac{1}{h}\mathrm{E}g(x_i)k(\frac{x_i-x}{h}) \\
&= \frac{1}{h}\int g(x_i)k(\frac{x_i-x}{h})f(x_i)dx_i \\
&= \frac{1}{\cancel{h}}\int g(x+hv)k(v)f(x+hv)\cancel{h}dv \\
\text{Taylor Expansion} &= \int [g+g'hv+\frac{g''}{2}h^2v^2+o(h^2)]k(v)[f+f'hv+\frac{f''}{2}h^2v^2+o(h^2)]dv \\
&= gf+\frac{h^2}{2}[2g'f'+gf''+g''f]k_2+o(h^2)
\end{aligned}
$$

方差. 我们先分解一下方差的结构：law of total variance/variance decomposion/conditional variance/law of iterated variance, 也可以用回归来理解。

$$
\mathrm{V}ar(\hat{m}) = \mathrm{E}\{\underset{\mathrm{A}}{\mathrm{V}ar(\hat{m}|x_i)}\} + \underset{\mathrm{B}}{\mathrm{V}ar\{\mathrm{E}(\hat{m}|x_i)\}}
$$

$$
\begin{aligned}
\mathrm{A} &= \mathrm{E}\{\mathrm{V}ar[\frac{1}{nh}\sum y_i k(\frac{x_i-x}{h})|x_i]\} \\
&= \mathrm{E}\{\mathrm{V}ar[\frac{1}{nh}\sum y_i k(\frac{x_i-x}{h})|x_i]\} \\
&= \mathrm{E}\{\frac{1}{n^2h^2}n\underset{=\sigma^2}{\mathrm{V}ar(y_i|x_i)}k^2(\frac{x_i-x}{h})+\underset{=0}{\mathrm{C}ov}\},\ (assume\ \mathrm{V}ar(y_i|x_i)=\mathrm{V}ar(\varepsilon)=\sigma^2) \\
&= \frac{\sigma^2}{nh^2}\mathrm{E}k^2(\frac{x_i-x}{h}) = \frac{\sigma^2}{nh}f\int k^2(v)dv \\
&= \frac{\sigma^2}{nh}f\mathrm{K}+o(\frac{1}{nh}) \\
\mathrm{B} &= \mathrm{V}ar\{\mathrm{E}[\frac{1}{nh}\sum y_i k(\frac{x_i-x}{h})|x_i]\} \\
&= \mathrm{V}ar\{\frac{1}{nh}\sum g(x_i)k(\frac{x_i-x}{h})\} \\
&= \frac{1}{\cancel{n}h^2}\cancel{n}\mathrm{V}ar(g(x_i)k(\frac{x_i-x}{h})) \\
&= \frac{1}{nh^2}\{\mathrm{E}[g^2k^2(\frac{x_i-x}{h})]-[\underset{o_p(1)}{\mathrm{E}gk(\frac{x_i-x}{h})}]^2\} \\
&= \frac{1}{nh^2}g^2(f+o(\frac{1}{nh}))\int k^2(v)dv \\
&= \frac{1}{nh}g^2f\mathrm{K}+o(\frac{1}{nh})
\end{aligned}
$$

Hence that

$$\mathrm{V}ar(\hat{m}(x)) = \frac{1}{nh}[\sigma^2 f + g^2 f]\mathrm{K} + o(\frac{1}{nh})$$

协方差.

$$\begin{aligned}
\mathrm{C}ov(\hat{m},\hat{f}) &= \mathrm{E}\{(\hat{m}-\mathrm{E}\hat{m})(\hat{f}-\mathrm{E}\hat{f})\} = \mathrm{E}\{\mathrm{E}(\hat{m}|x_i)(\hat{f}-\mathrm{E}\hat{f})\} \\
&= \mathrm{E}\{\frac{1}{nh}\sum g(x_i)\mathrm{K}(\frac{x_i-x}{h})[\frac{1}{nh}\sum \mathrm{K}(\frac{x_i-x}{h}) - \frac{1}{nh}\sum \mathrm{E}[\mathrm{K}(\frac{x_i-x}{h})]]\} \\
&= \mathrm{E}\{\frac{1}{n^2h^2}\sum g(x_i)\mathrm{K}(\frac{x_i-x}{h})\sum[\mathrm{K}(\frac{x_i-x}{h}) - \mathrm{E}[\mathrm{K}(\frac{x_i-x}{h})]]\} \\
&= \frac{1}{nh^2}\{\mathrm{E}g(x_i)\mathrm{K}^2(\frac{x_i-x}{h}) - \mathrm{E}g(x_i)\mathrm{K}(\frac{x_i-x}{h})\mathrm{E}\mathrm{K}(\frac{x_i-x}{h})\} \\
&= \frac{1}{nh}gf\mathrm{K} + o(\frac{1}{nh})
\end{aligned}$$

MSE. 回到MSE，对$\hat{g}(x)$，有

$$\begin{aligned}
\mathrm{E}[\hat{g}(x)] &= \frac{\mathrm{E}\hat{m}}{\mathrm{E}\hat{f}} - \frac{\mathrm{C}ov(\hat{m},\hat{f})}{(\mathrm{E}f)^2} + \frac{\mathrm{E}[\hat{m}]\mathrm{V}ar(\hat{f})}{(\mathrm{E}\hat{f})^3} \\
&\approx \frac{\{gh + \frac{h^2}{2}[g'' + gf'' + 2g'f']\int v^2 k(v)dv\}}{f + \frac{h^2 f''}{2}\int v^2 k(v)dv} = \frac{\{gh + \frac{h^2}{2}[g'' + gf'' + 2g'f']k_2\}}{f[1 + \frac{h^2 f''}{2f}k_2]} \\
&= \frac{\{gh + \frac{h^2}{2}[g'' + gf'' + 2g'f']k_2\}}{f}[1 - (\frac{h^2 f''}{2f}k_2) + ()^2 - ()^3 + \ldots] \\
&= g + \frac{h^2}{2f}[g''f + gf'' + 2g'f']k^2 - \frac{h^2 gf''}{2f}\mathrm{K}_2 + o(h^2)
\end{aligned}$$

$$
\begin{aligned}
\mathrm{V}ar(\hat{g}) &= \mathrm{E}[\hat{g}-\mathrm{E}\hat{g}]^2 \\
\hat{g}-\mathrm{E}\hat{g} &\approx \frac{\hat{m}-\mathrm{E}\hat{m}}{\mathrm{E}\hat{f}} - \frac{\mathrm{E}\hat{m}(\hat{f}-\mathrm{E}\hat{f})}{(\mathrm{E}\hat{f})^2} \\
\mathrm{V}ar(\hat{g}) &= \frac{\mathrm{V}ar(\hat{m})}{(\mathrm{E}\hat{f})^2} + \frac{(\mathrm{E}\hat{m})^2\mathrm{V}ar\hat{f}}{(\mathrm{E}\hat{f})^4} - \frac{2\mathrm{E}\hat{m}\mathrm{C}ov(\hat{m},\hat{f})}{(\mathrm{E}\hat{f})^3} \overset{def}{=} \mathrm{A}+\mathrm{B}-\mathrm{C} \\
\mathrm{A} &\approx \frac{\frac{1}{nh}(\sigma^2 f + g^2 f)\int k^2 dv}{f^2} = \frac{\sigma^2+g^2}{nhf}\int k^2 dv \\
\mathrm{B} &\approx \frac{(gh)^2\frac{f}{nh}\int k^2 dv}{f^4} = \frac{g^2\int k^2 dv}{nhf} \\
\mathrm{C} &\approx \frac{2gf\frac{1}{nh}gf\int k^2 dv}{f^3} = \frac{2g^2\int k^2 dv}{nhf} \\
\mathrm{V}ar(\hat{g}) &= \mathrm{A}+\mathrm{B}-\mathrm{C} \\
&= \frac{1}{nhf}[\sigma^2+g^2+g^2-2g^2]\int k^2(v)dv + o(\frac{1}{nh}) \\
&= \frac{\sigma^2}{nhf}\int k^2(v)dv + o(\frac{1}{nh}) \\
\mathrm{MSE}\hat{g} &= \frac{h^4}{4f^2(x)}[g''f+2g'f']^2\mathrm{K}_2^2 + \frac{\sigma^2\mathrm{K}}{nhf} + o(h^4+\frac{1}{nh})
\end{aligned}
$$

可得各变量的收敛速度如下：

$$
h^4 \propto \frac{1}{nh} \Rightarrow h \propto n^{-\frac{1}{5}}
$$

$$
\mathrm{MSE}\hat{g} \propto n^{-\frac{4}{5}}
$$

$$
\hat{f} = f + o_p(n^{-\frac{2}{5}})
$$

$$
\hat{g} = g + o_p(n^{-\frac{2}{5}})
$$

总结可得，带宽h的增大会增加bias，但会减少variance，但样本量n的增加会降低variance。同时，真实方程的弯曲程度g''越大，bias也会越大。由于$f(x)$出现分母，部分空间中x密度较低同样会导致MSE较大。