第五章　异方差

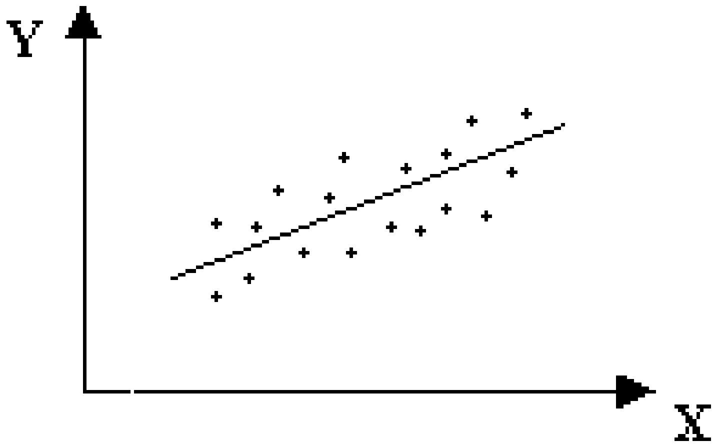
(Heteroscedasticity)

第一节　异方差性问题

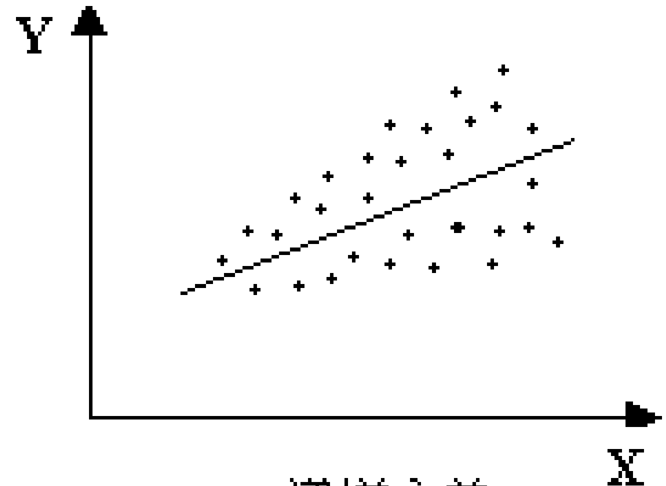
一、异方差性问题(Heteroscedasticity)

在经典线性回归模型（**CLRM**）中，我们假定随即干扰项具有同方差性，即： $\mathbf{Var(u_i|X_i)=E[u_i-E(u_i)|X_i]^2 = E(u_i^2|X_i]^2 = \sigma^2}$

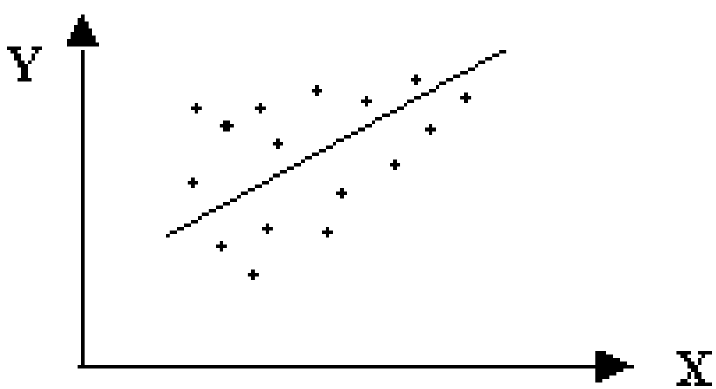
这实际上是假定了被解释变量 $\mathbf{Y_i}$ 的值围绕其期望值的分散程度相同。实际上，对应于解释变量的不同取值，方差可能不同，即本假定不成立。



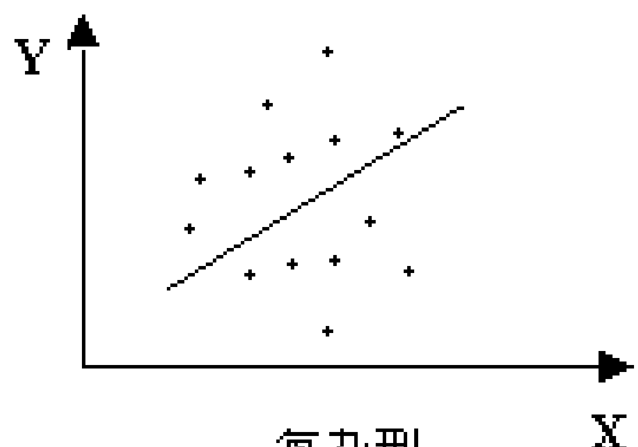
同方差



递增方差



递减方差



复杂型

如果保持随机项的协方差为0，则 $\mathbf{Y} = \beta_0 + \beta_1 X + u$ 的协方差矩阵为：

$$E(UU^T) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \cdots & & & \cdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

或者说，$Var(u_i) = \sigma_i^2 \neq \text{常数}, Cov(u_i, u_j) = 0, (i \neq j)$ 。

在这种情况下，称随机项$\mathbf{u_i}$具有异方差性。

二、异方差的原因：

1、省略了重要的解释变量引起异方差。

2、模型形式设定不当,引起异方差。

3、统计资料误差引起异方差。

(时间序列数据中，观测技术的缺陷引起的观测值的误差。)

三、异方差的后果

基于CLRM假定的OLS估计参数结果将受到影响。

1、考虑异方差性的OLS估计

如果假定$Var(u_i)=E(u_i^2)=\sigma_i^2 \neq$常数，保留其它的CLRM假定，以一元回归模型为例，普通OLS估计为：

$$Y_i=\beta_0+\beta_1 X_i+u_i$$

$$\hat{\beta}_1=\frac{\sum x_i y_i}{\sum x_i^2}=\frac{\sum x_i(Y_i-\bar{Y})}{\sum x_i^2}=\frac{\sum x_i(\beta_0+\beta_1 X_i+u_i)}{\sum x_i^2}$$

$$=\beta_1+\sum c_i u_i$$

(同方差假定下，$Var(\hat{\beta}_1)=\frac{\sigma^2}{\sum x_i^2}$)

当u_i存在异方差时，β的估计量为：

$$\tilde{\beta}_1=\beta_1+\sum c_i u_i,\quad E\tilde{\beta}_1=\beta_1,\quad Var（\tilde{\beta}_1）=\sum c_i^2\delta_i^2$$

不妨假设δ_i^2随X_i^2而变化，即$\delta_i^2=\delta^2 X_i^2$

$$Var(\tilde{\beta}_1)=\sum c_i^2 X_i^2\delta^2=\delta^2\sum\left(\frac{x_i}{\sum x_i^2}\right)^2 X_i^2$$

$$=\frac{\delta^2}{\sum x_i^2}\times\frac{\sum x_i^2 X_i^2}{\sum x_i^2}=Var(\hat{\beta}_1)\times\frac{\sum x_i^2 X_i^2}{\sum x_i^2}$$

$$\geq Var(\hat{\beta}_1)$$

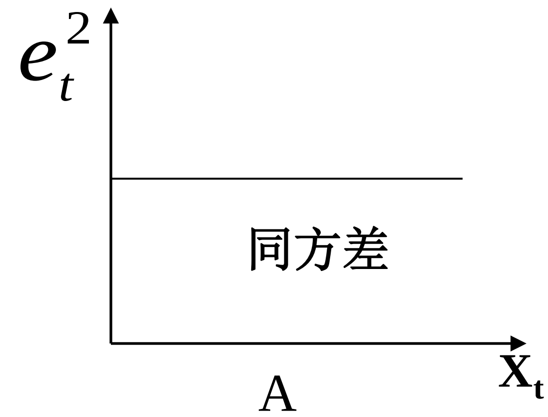
因此，估计量是线性无偏的，但不是最优估计量（具有最小方差性）。

2、参数的显著性检验失去意义

t、F检验都是在同方差下推出的，如果出现异方差，t、F检验将失去意义。

3、预测精度降低（预测结果不可信）

第二节　异方差性检验

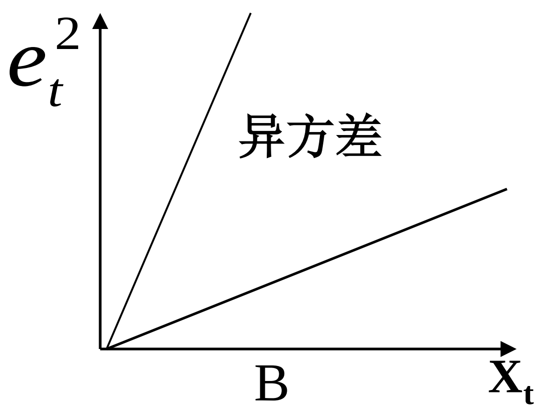


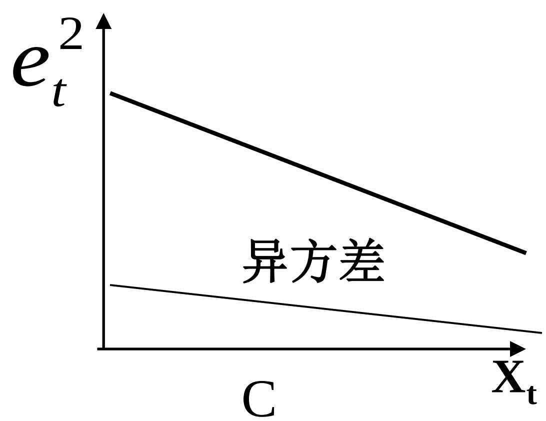
一、图示法

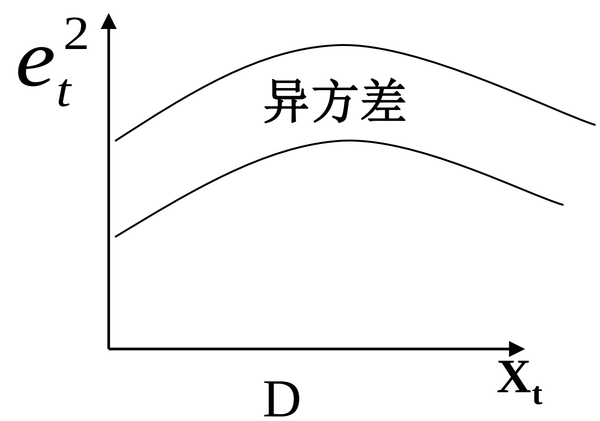
1、作回归；

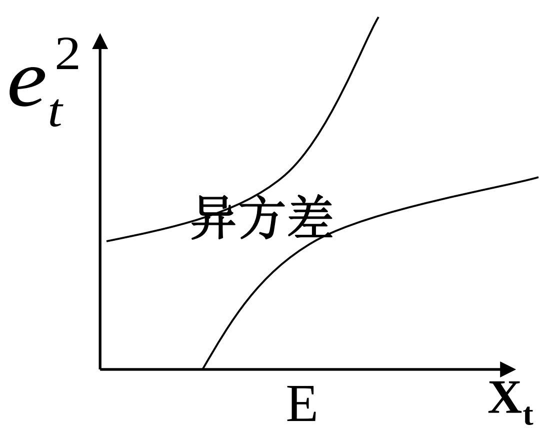
2、计算 $e_t = Y_t - \hat{Y}_t$

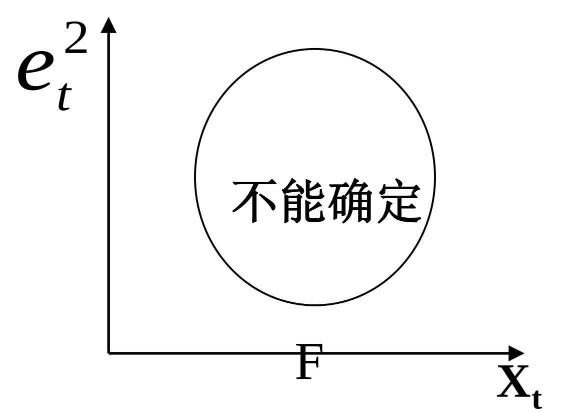
3、作散点图（X_t，e_t^2）











二、斯皮尔曼（Spearman）等级相关系数检验（小样本）

通过随机项的方差与解释变量的等级相关系数的显著性检验，判断是否存在异方差性。步骤：

1、作OLS估计，得到e_i。

2、把$|e_i|$和X_i按升序或降序赋予等级值（1,2,⋯，n）。

3、计算斯皮尔曼等级相关系数：

$$r = 1 - 6\left[\frac{\sum d_i^2}{n(n^2-1)}\right],$$ 其中d_i为第i组观测值的$|e_i|$和X_i的分类等级差。

4、可以证明：$t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2),$

若显著（超过临界值），则说明存在异方差性。否则，不存在异方差性。

可以证明：$r \sim N\left(0, \frac{1}{n-1}\right)$

则：$U = \frac{r-0}{\sqrt{\frac{1}{n-1}}} = r\sqrt{n-1} \sim N(0,1)$

若显著（超过临界值），则说明存在异方差性。否则，说明不存在异方差性。

这一检验的依据，其实就是检查随着解释变量的变化,方差是否随之变化(意味着等级差异随之变动)。

例，设某种商品1982—1991年的销售量Y（万斤）与价格X（元）的统计资料如下表，试用Spearman等级相关系数法检验模型是否存在异方差性。

年份	Y_i	X_i	X等级	$\hat{Y}_i$	$\lvert e_i \rvert$	$\lvert e_i \rvert$等级	d_i	d_i^2
1982	1.1	5.1	9	0.4994	0.6006	9	0	0
1983	1.3	3.4	7	1.8534	0.5534	8	-1	1
1984	1.3	3.6	8	1.6941	0.3941	6	2	4
1985	1.6	3.1	6	2.0924	0.4924	7	-1	1
1986	2.1	2.7	4	2.4110	0.3110	5	-1	1
1987	2.6	2.8	5	2.3313	0.2687	4	1	1
1988	2.4	2.6	3	2.4906	0.0906	1	2	4
1989	2.8	2.4	2	2.6499	0.1501	2	0	0
1990	3.1	2.1	1	2.8889	0.2111	3	-2	4
1991	3.5	2.1	1	2.8889	0.6111	10	-9	81

解：根据表中的数据，利用普通最小二乘得：

$$\hat{Y}_t = 4.5615 - 0.7965 X_t$$

(8.59)　(-4.66)　　$R^2 = 0.83$

将$\hat{Y}_t$、e_t、d_t计算于表中，且$\sum d_t^2 = 97$

*Spearman*等级相关系数：

$$r = 1 - \frac{6 \times 97}{10 \times (10^2 - 1)} = 0.4121$$

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.4121\sqrt{8}}{\sqrt{1-0.4121^2}} = 1.2793$$

$$U^* = r\sqrt{n-1} = 3r = 1.2363$$

结论:不存在异方差性。

三、戈里瑟（Glejser）检验(只能检验有异方差)

假定σ_i^2与某一解释变量X_{ik}有关。

可以对以下函数形势作回归：

$$|e_i|\, or\ e_i^2 = \beta_0 + \beta_1 f(X_{ik}) + v$$

$$|e_i|\, or\ e_i^2 = \beta_0 + \beta_1 X_{ik} + v_i$$

$$|e_i|\, or\ e_i^2 = \beta_0 + \beta_1 \sqrt{X_{ik}} + v_i$$

$$|e_i|\, or\ e_i^2 = \beta_0 + \beta_1 \frac{1}{X_{ik}} + v_i$$

……

进行回归，对 β和回归方程作显著性检验。

若显著，则存在异方差。

如果回归结果表明异方差与多个变量有关，可以引入多个变量进行回归，并进行检验。

戈里瑟（Glejser）检验的优点在于，在检验异方差的同时，可以得到异方差形式的信息（与解释变量的关系），据此修正回归模型，以得到最优线性无偏估计。

四、帕克（Pack）检验(只能检验有异方差)

假定σ_i^2与某一解释变量X_k有关：

$$\sigma_i^2=\sigma^2 X_k^{\beta} e^{v_i}, \text{或} \ln(\sigma_i^2)=\ln(\sigma^2)+\beta\ln(X_k)+v_i$$

由于σ_i^2未知，以同方差假定下OLS估计得到的e_i^2代替：

$$\ln(e_i^2)=\alpha+\beta\ln(X_k)+v_i$$

进行回归，对β作显著性检验。

若显著，则存在异方差。

同时能确定影响随机项的解释变量。

五、戈德菲尔德—夸特（**Goldfied-Quandt**）检验（大样本）

G-Q检验适用于大样本、随机项的方差与某个解释变量存在正相关的情况。检验的前提条件是：随机项服从正态分布；无序列相关。步骤：

1、把样本按解释变量X_i观测值大小顺序排列。

2、略去居中的c个样本，把样本分为两个子样本。（略去的样本数c以总样本数的1/4为宜）

3、分别对两个子样本进行OLS回归，并分别计算出RSS：

$$RSS_1=\sum e_{i1}^2,\quad RSS_2=\sum e_{i2}^2$$

4、计算统计量：$F=\dfrac{RSS_2\Big/\left(\frac{n-c}{2}-p-1\right)}{RSS_1\Big/\left(\frac{n-c}{2}-p-1\right)}=\dfrac{RSS_2}{RSS_1}$

$$\sim F\left(\frac{n-c}{2}-p-1,\frac{n-c}{2}-p-1\right)$$

若显著（超过临界值F_α），则说明存在异方差性。

若 F ≤ F_α，则为同方差；如果 F≥ F_α 则存在异方差，F值越大（超过临界值），说明存在异方差性的可能性就越大。

6、怀特(White)检验（截面数据）

White检验不需要对观测值排序，也不依赖于随机误差项服从正态分布，它是通过一个辅助回归式构造 χ2 统计量进行异方差检验。White检验的具体步骤如下。

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + u_t$$

①首先对上式进行OLS回归，求残差e_t^2

②做如下辅助回归式

$$e_t^2 = \alpha_0 + \alpha_1 \mathrm{x}_{t1} + \alpha_2 \mathrm{x}_{t2} + \alpha_3 \mathrm{x}_{t1}^2 + \alpha_4 \mathrm{x}_{t2}^2 + \alpha_5 \mathrm{x}_{t1} \mathrm{x}_{t2} + \mathrm{v}_t$$

拟合优度为 R^2 检验统计量为: $nR^2 \sim \chi^2(m)$

m为上式中解释变量个数，这里m=5。

若$nR^2 < \chi^2_\alpha(m)(or\ p > \alpha)$，接受H0 （$u_t$ 具有同方差）

若$nR^2 \geqslant \chi^2_\alpha(m)(or\ p \leqslant \alpha)$，拒绝H0 （$u_t$ 具有异方差）

7、自回归条件异方差（***ARCH***）检验（时间序列数据）

异方差的另一种检验方法称作自回归条件异方差（*ARCH*）检验。这种检验方法不是把原回归模型的随机误差项看作是***xt***的函数，而是把它 看作误差滞后项的函数 。

$$e_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \alpha_2 e_{t-2}^2 + \ldots + \alpha_p e_{t-k}^2$$

ARCH是误差项二阶矩的自回归过程。恩格尔（**Engle 1982**）针对***ARCH***过程提出***LM***检验法。辅助回归式定义为

LM统计量定义为:

$$ARCH = nR^2 \sim \chi^2(k)$$

若 $nR^2 < \chi_\alpha^2(k)(or\ p > \alpha)$，接受H0 （$u_t$ 具有同方差）

若 $nR^2 \geqslant \chi_\alpha^2(k)(or\ p \leqslant \alpha)$，拒绝H0 （$u_t$ 具有异方差）

第三节　异方差模型的处理

思想：变异方差为同方差，或尽量减少方差变异的程度。

一、模型变换法（适用于异方差已知的情况）

如果随机项的方差σ_i^2已知，则：

设原模型为：$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + u_i$,

$$\sigma_i^2 = \sigma^2 f(X_{i1}, X_{i2}, \cdots X_{im})$$

以$\sqrt{f(X_{i1}, X_{i2}, \cdots X_{im})}$除以原模型两边，

可得到满足**CLRM**假定的新模型：

$$\frac{Y_i}{\sqrt{f(X_{i1,}X_{i2},\cdots X_{im})}}=\frac{\beta_0}{\sqrt{f(X_{i1,}X_{i2},\cdots X_{im})}_0}+\beta_1\frac{X_{i1}}{\sqrt{f(X_{i1,}X_{i2},\cdots X_{im})}}+\cdots$$

$$+\beta_p\frac{X_{ip}}{\sqrt{f(X_{i1,}X_{i2},\cdots X_{im})}}+\frac{u_i}{\sqrt{f(X_{i1,}X_{i2},\cdots X_{im})}}$$

$$\frac{Y_i}{\sqrt{f(X_{i1,}X_{i2},\cdots X_{im})}}=\frac{\beta_0}{\sqrt{f(X_{i1,}X_{i2},\cdots X_{im})}}+\beta_1\frac{X_{i1}}{\sqrt{f(X_{i1,}X_{i2},\cdots X_{im})}}+\cdots$$

$$+\beta_p\frac{X_{ip}}{\sqrt{f(X_{i1,}X_{i2},\cdots X_{im})}}+u_i'$$

$$Var(u_i') = Var\left(\frac{u_i}{\sqrt{f(X_{i1}, X_{i2}, \cdots X_{im})}}\right) = \frac{Var(u_i)}{f(X_{i1}, X_{i2}, \cdots X_{im})}$$

$$= \frac{f(X_{i1}, X_{i2}, \cdots X_{im})\delta^2}{f(X_{i1}, X_{i2}, \cdots X_{im})} = \delta^2$$

如果知道$f(X_{i1}, X_{i2}, \cdots X_{im})$，即可进行估计。

因此，关键的问题是找出异方差的具体形式。

特别，以一元线性回归为例，若$f(X_i)=X_i^2$

则变换后的模型为：

$$\frac{Y_i}{X_i}=\frac{\beta_0}{X_i}+\beta_1+\frac{u_i}{X_i}$$

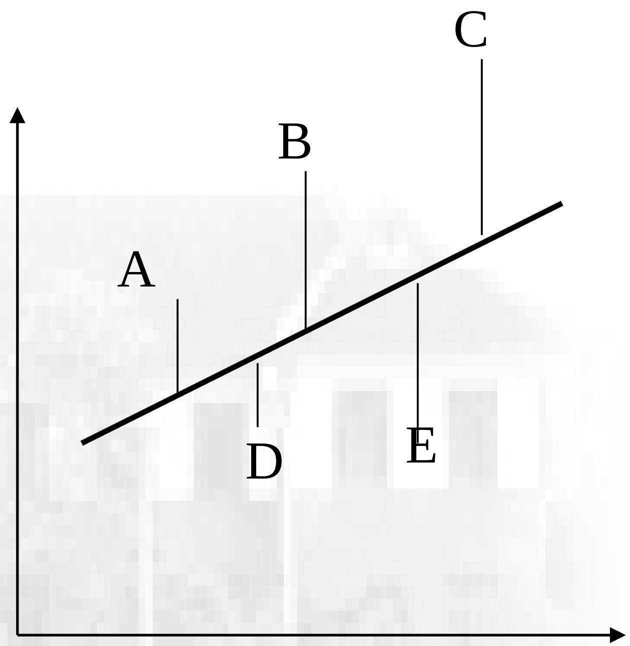
若$f(X_i)=X_i$

则变换后的模型为：

$$\frac{Y_i}{\sqrt{X_i}}=\frac{\beta_0}{\sqrt{X_i}}+\beta_1\sqrt{X_i}+\frac{u_i}{\sqrt{X_i}}$$

二、加权最小二乘法（**WLS**）

对每个点，如果残差较大，则给予较小的权重；如果残差较小，则给予较大的权重。



以一元为例，设$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + u_i \\ Var(u_i) = \delta^2 f(X_i) \end{cases}$

由$Q(\hat{\beta}_0^*, \hat{\beta}_1^*) = \sum w_i e_i^2 = \sum w_i (Y_i - \hat{Y}_i)^2 = \sum w_i (Y_i - \hat{\beta}_0^* - \hat{\beta}_1^* X_i)^2$

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{\beta}_0^*} = 0 \\ \dfrac{\partial Q}{\partial \hat{\beta}_1^*} = 0 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_1^* = \dfrac{\sum w_i x_i^* y_i^*}{\sum w_i x_i^{*2}} \\ \hat{\beta}_0^* = \bar{Y}^* - \hat{\beta}_1^* \bar{X}^* \end{cases}$$

其中, $\bar{X}^* = \dfrac{\sum w_i X_i}{\sum w_i}, \bar{Y}^* = \dfrac{\sum w_i Y_i}{\sum w_i}, x_i^* = X_i - \bar{X}^*, y_i^* = Y_i - \bar{Y}^*$

当$w_1 = w_2 = \cdots = w_n = w$时，则

$$WLS \to OLS$$

实际应用中，常取$w_i = \dfrac{1}{e_i^2}\ or\ \dfrac{1}{0.001+e_i^2}\ or\ \dfrac{1}{X_i^2}\ or\ \dfrac{1}{X_i}$

加权最小二乘估计（WLS）与通过模型变换法得到的估计量是一致的。

第四节　案例分析