Motivation of using OLS

- As it turns out, an extremely useful property of OLS is that, even if the relationship between the x variables and the y variable is nonlinear, OLS yields the best linear predictor (BLP) of y given x.
- The sense of "best" is that the OLS prediction is the linear prediction that minimizes the mean squared error.
- This is a significant (though often overlooked) advantage of OLS over nonlinear estimation methods.

- Consider the following estimation problem: choose f so that it minimizes some function of the forecast error, $y - f(\boldsymbol{x})$.
- If the criterion is to minimize the **mean squared error**,

$$\mathrm{E}\big[(y - f(\boldsymbol{x}))^2\big],$$

then it turns out, the function f that minimizes the mean squared error is simply the **conditional expectation function** (CEF), $\mathrm{E}[y|\boldsymbol{x}]$.

$$\begin{aligned}(y - f(\boldsymbol{x}))^2 &= \big((y - \mathrm{E}[y|\boldsymbol{x}]) + (\mathrm{E}[y|\boldsymbol{x}] - f(\boldsymbol{x}))\big)^2 \\ &= (y - \mathrm{E}[y|\boldsymbol{x}])^2 + 2(y - \mathrm{E}[y|\boldsymbol{x}])(\mathrm{E}[y|\boldsymbol{x}] - f(\boldsymbol{x})) + (\mathrm{E}[y|\boldsymbol{x}] - f(\boldsymbol{x}))^2\end{aligned}$$

Take expectations, it could easily be seen

$$\begin{aligned}\mathrm{MSE} \equiv \mathrm{E}\big[(y - f(\boldsymbol{x}))^2\big] &= \mathrm{E}\big[(y - \mathrm{E}[y|\boldsymbol{x}])^2\big] + \mathrm{E}\big[(\mathrm{E}[y|\boldsymbol{x}] - f(\boldsymbol{x}))^2\big] \\ &\geq \mathrm{E}\big[(y - \mathrm{E}[y|\boldsymbol{x}])^2\big]\end{aligned}$$

- Note: $\mathrm{E}[y|\boldsymbol{x}]$ could be highly nonlinear.

- Therefore, the conditional expectation $\mathrm{E}[y|\boldsymbol{x}]$ is the best predictor of y (in the sense that it minimizes the mean squared error).
- What is the best linear predictor of y based on $\boldsymbol{x}$?

$$\arg\min_{\tilde{\boldsymbol{\beta}}} \mathrm{E}\big[(y - \boldsymbol{x}'\tilde{\boldsymbol{\beta}})^2\big]$$

- The first order condition gives

$$\mathrm{E}[\boldsymbol{x}(y - \boldsymbol{x}'\tilde{\boldsymbol{\beta}})] = \mathbf{0},$$

so the solution for $\boldsymbol{b}$ is

$$\boldsymbol{\beta}^* \equiv \mathrm{E}[\boldsymbol{x}\boldsymbol{x}']^{-1}\mathrm{E}[\boldsymbol{x}y]$$

- $\boldsymbol{\beta}^*$ is defined as the **least squares projection coefficients**, while $\boldsymbol{x}'\boldsymbol{\beta}^*$ is defined as the **least squares** (or **linear**) **projection** of y on $\boldsymbol{x}$.

CEF and linear projection

1. Suppose the CEF is linear. Then the least squares projection is the CEF.
   - Proof: Suppose $\mathrm{E}[y|\boldsymbol{x}] = \boldsymbol{x}'\tilde{\boldsymbol{\beta}}$ for some $\tilde{\boldsymbol{\beta}}$. Note that $\mathrm{E}[\boldsymbol{x}(y - \mathrm{E}[y|\boldsymbol{x}])] = \mathbf{0}$, substitute $\mathrm{E}[y|\boldsymbol{x}] = \boldsymbol{x}'\tilde{\boldsymbol{\beta}}$ to find that $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* = \mathrm{E}[\boldsymbol{x}\boldsymbol{x}']^{-1}\mathrm{E}[\boldsymbol{x}y]$.
   - Therefore, if it happens that CEF is linear, then the linear projection is the best predictor.
2. The best linear approximation to $\mathrm{E}[y|\boldsymbol{x}]$ is $\boldsymbol{x}'\boldsymbol{\beta}^*$.
   - Proof: This is the same as saying that $\boldsymbol{\beta}^*$ minimizes

$$\arg\min_{\tilde{\boldsymbol{\beta}}} \mathrm{E}\big[(\mathrm{E}[y|\boldsymbol{x}] - \boldsymbol{x}'\tilde{\boldsymbol{\beta}})^2\big].$$

- To see why, write

$$
\begin{aligned}
(y-\boldsymbol{x}'\tilde{\boldsymbol{\beta}})^2 &= \big((y-\mathrm{E}[y|\boldsymbol{x}])+(\mathrm{E}[y|\boldsymbol{x}]-\boldsymbol{x}'\tilde{\boldsymbol{\beta}})\big)^2 \\
&= (y-\mathrm{E}[y|\boldsymbol{x}])^2+(\mathrm{E}[y|\boldsymbol{x}]-\boldsymbol{x}'\tilde{\boldsymbol{\beta}})^2 \\
&\quad + 2(y-\mathrm{E}[y|\boldsymbol{x}])(\mathrm{E}[y|\boldsymbol{x}]-\boldsymbol{x}'\tilde{\boldsymbol{\beta}})
\end{aligned}
$$

- The first term does not involve $\tilde{\boldsymbol{\beta}}$, and the last term has expectation zero. The CEF-approximation problem, therefore has the same solution as the best-linear-predictor problem, $\arg\min_{\tilde{\boldsymbol{\beta}}} \mathrm{E}\big[(y-\boldsymbol{x}'\tilde{\boldsymbol{\beta}})^2\big]$.

- Therefore, in the prediction context, linear projection is the best linear predictor while CEF is the best unrestricted predictor.
- On the other hand, even if CEF is nonlinear, linear regression provides the best linear approximation to it.

第 7 章　异方差与 GLS

7.1　异方差的后果

“异方差”(heteroskedasticity)是违背球型扰动项假设的一种情形，即$\text{Var}(\varepsilon_i \mid \boldsymbol{X})$依赖于$i$，不是常数。

在异方差的情况下：

(1) OLS 估计量依然无偏、一致且渐近正态，因为在证明这些性质时并未用到“同方差”的假定。

(2) OLS 估计量方差$\text{Var}(\boldsymbol{b}\,|\,\boldsymbol{X})$的表达式不再是$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$，因为$\text{Var}(\boldsymbol{\varepsilon}\,|\,\boldsymbol{X}) \neq \sigma^2\boldsymbol{I}$。因此，通常的$t$检验、$F$检验也失效了。

(3) 高斯–马尔可夫定理不再成立，OLS 不再是 BLUE。在异方差的情况下，GLS 才是 BLUE。

为何 OLS 不再是 BLUE？

假设$\text{Var}(\varepsilon_i\,|\,\boldsymbol{X})$是某解释变量$x_i$的增函数，参见图 7.1。

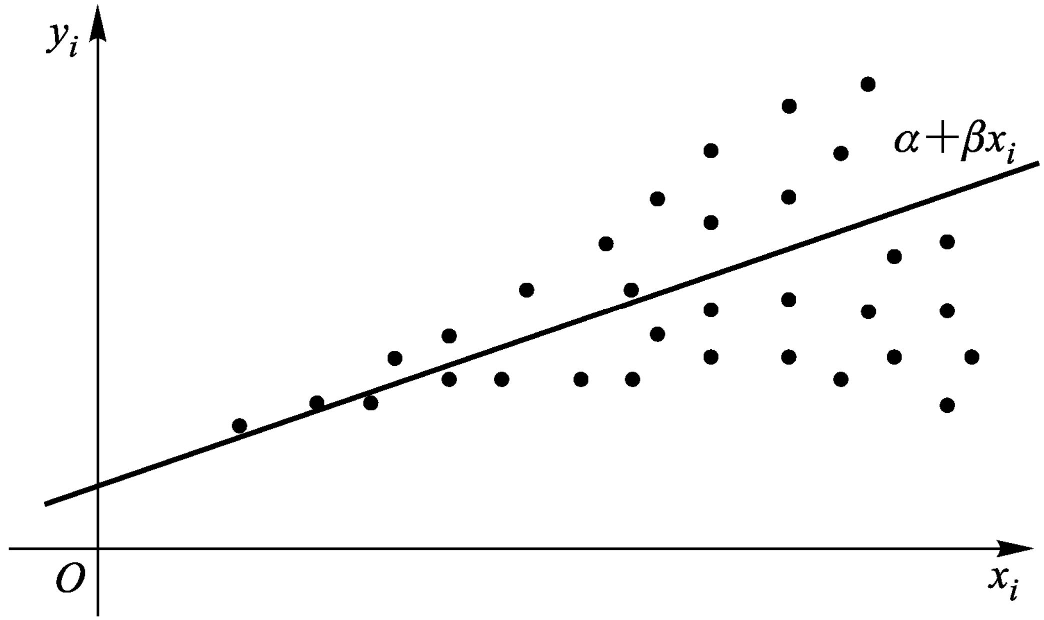


图 7.1　异方差的一种情形

GLS 及其特例“加权最小二乘法”(Weighted Least Square，简记 WLS)，通过对不同数据包含信息量的不同进行加权以提高效率。

7.2 异方差的例子

(1) 消费函数：

$$C_i = \alpha + \beta Y_i + \varepsilon_i$$

其中，C 为消费，Y 为收入。富人的消费计划较有弹性，而穷人的消费多为必需品。富人的消费支出难测量，包含较多测量误差。

(2) 企业的投资、销售收入与利润：大型企业的商业活动以亿元计，而小型企业以万元计，扰动项规模不同。

(3) 组间异方差：样本包含两组(类)数据，第一组为自我雇佣者(企业主、个体户)的收入，第二组为打工族的收入，前者的收入波动比后者大。

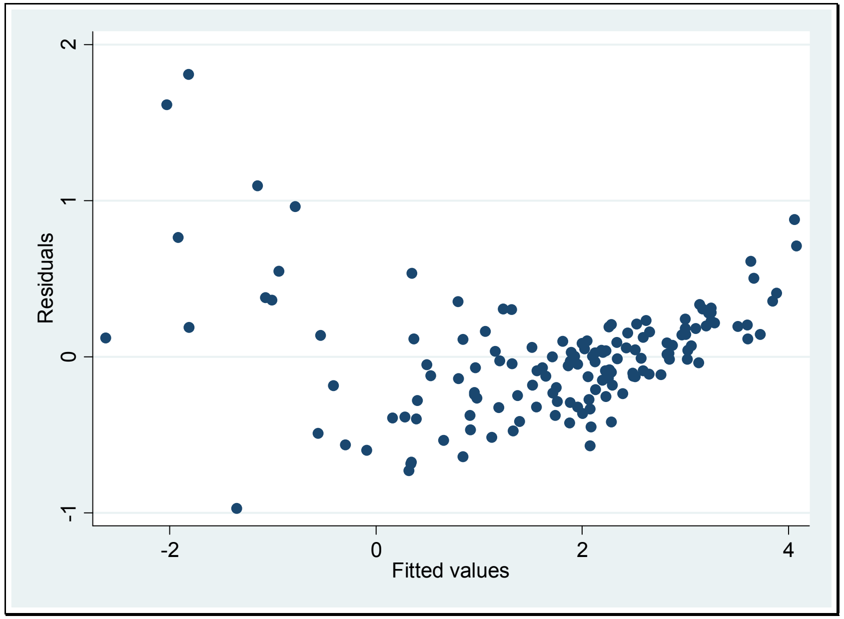
(4) 组平均数：如果数据为组平均数，则大组平均数的方差要比小组平均数的方差小。比如，全国各省的人均 GDP，人口多的省份其方差较小，方差与人口数成反比。

(5) 时间序列数据中也可能出现条件异方差，比如第 22 章的 ARCH 模型。

7.3 异方差的检验

1. 看残差图(residual plot)

看"残差e_i与拟合值$\hat{y}_i$的散点图"(residual-versus- fitted plot)，或"残差e_i与某个解释变量x_{ik}的散点图"(residual-versus-predictor plot)。



2．怀特检验(White test)

在条件同方差下，稳健标准误还原为普通标准误，二者的差别

可用来度量条件异方差。怀特检验正是基于这一思想。

在同方差的原假设$H_0: \mathrm{E}(\varepsilon_i^2 \mid \boldsymbol{X}) = \sigma^2$下，稳健协方差矩阵与普通协方差矩阵之差收敛到一个零矩阵：

$$\hat{\mathbf{S}} - s^2 \boldsymbol{S}_{XX} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' - s^2 \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \frac{1}{n}\sum_{i=1}^{n} (e_i^2 - s^2) \boldsymbol{x}_i \boldsymbol{x}_i' \xrightarrow{p} \mathbf{0}_{K\times K}$$

实际操作上，进行辅助回归：

$$e_i^2 \xrightarrow{\text{OLS}} \text{常数} + \boldsymbol{\psi}_i' \boldsymbol{\gamma}$$

并检验$\boldsymbol{\psi}_i$中所有变量的系数$\boldsymbol{\gamma}$均为 0。

如果R^2很低，意味着e_i^2无法由解释变量及其平方项与交叉项来解释，故倾向于接受同方差的原假设。可以证明：

$$nR^2 \xrightarrow{d} \chi^2(m)$$

如果nR^2很大(超过临界值)，则拒绝原假设H_0。

在大样本中，nR^2与检验整个方程显著性的F统计量渐近等价。

对于回归方程 $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$，检验原假设$H_0 : \beta_2 = \cdots = \beta_K = 0$，则$F$统计量：

$$F=\frac{R^2/(K-1)}{(1-R^2)/(n-K)}\sim F(K-1,n-K)$$

在大样本下，F分布与χ^2分布等价(第 5 章附录，见下)，即

$$(K-1)F=\frac{(n-K)R^2}{(1-R^2)}\xrightarrow{d}\chi^2(K-1)$$

在原假设成立的情况下，当$n\to\infty$时，$(n-K)\to n$，$(1-R^2)\to 1$，故$(K-1)F\to nR^2$，故F检验与nR^2检验在大样本下等价。

怀特检验的优点是，可以检验任何形式的异方差；缺点是，如果H_0被拒绝，并不提供有关异方差具体形式的信息。

A5.3 F 分布与 χ^2 分布在大样本下是等价的

命题 假设 $F \sim F(m, n-K)$ 分布，则当 $n \to \infty$ 时，$mF \xrightarrow{d} \chi^2(m)$。

证明：因为 $F \sim F(m, n-K)$，故可设 $F = \dfrac{\chi^2(m)/m}{\chi^2(n-K)/(n-K)}$。

根据 χ^2 分布的性质，$\mathrm{E}\left[\chi^2(n-K)\right] = n-K$，而 $\mathrm{Var}\left[\chi^2(n-K)\right] = 2(n-K)$。

故 F 统计量分母的期望值为：

$$\mathrm{E}\left[\chi^2(n-K)/(n-K)\right] = 1$$

F 统计量分母的方差为：

$$\mathrm{Var}\left[\chi^2(n-K)\big/(n-K)\right]=\frac{2(n-K)}{(n-K)^2}=\frac{2}{n-K}\to 0 \quad (当 n\to\infty 时)$$

故分母依均方收敛于 1。

因此，分母依概率收敛于 1，即 $\chi^2(n-K)\big/(n-K)\xrightarrow{p}1$。

所以，$F\xrightarrow{d}\chi^2(m)\big/m$。

故 $mF\xrightarrow{d}\chi^2(m)$。

3．BP 检验(Breusch and Pagan, 1979)

假设回归模型为$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$，检验以下原假设：

$$H_0:\ \mathrm{E}(\varepsilon_i^2 \mid x_2, \cdots, x_K) = \sigma^2$$

如果H_0不成立，则条件方差$\mathrm{E}(\varepsilon_i^2 \mid x_2, \cdots, x_K)$是$(x_2, \cdots, x_K)$的函数，称为“条件方差函数”(conditional variance function)。

BP 检验假设此条件方差函数为线性函数：

$$\varepsilon_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + u_i$$

原假设简化为

$$H_0 : \delta_2 = \cdots = \delta_K = 0$$

由于ε_i不可观测，故使用e_i^2进行辅助回归：

$$e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i$$

使用nR^2统计量：

$$nR^2 \xrightarrow{d} \chi^2(K-1)$$

BP 检验与怀特检验的区别在于，后者还包括平方项与交叉项。

BP 检验的优点在于其建设性，可帮助确认异方差的具体形式。

7.4　异方差的处理

1．使用“OLS + 稳健标准误”

一种处理方法是，仍进行 OLS 回归，但使用稳健标准误。

这是最简单，也是目前通用的方法。只要样本容量较大，即使在异方差的情况下，若使用稳健标准误，则所有参数估计、假设检验均可照常进行。

但还可能存在比 OLS 更有效的方法，比如 GLS。

2．广义最小二乘法(GLS)

假设$\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{V}(\boldsymbol{X}) \neq \sigma^2 \boldsymbol{I}_n$，其中$\boldsymbol{V}(\boldsymbol{X})$为对称正定矩阵且已知，可能依赖于$\boldsymbol{X}$。

GLS 的基本思想是，通过变量转换，使得转换后的模型满足球型扰动项的假定。

命题 对于对称正定矩阵$\boldsymbol{V}_{n\times n}$，存在非退化矩阵$\boldsymbol{C}_{n\times n}$，使得$\boldsymbol{V}^{-1} = \boldsymbol{C}'\boldsymbol{C}$。

在一维情况下，“$\boldsymbol{V}$ 正定”即要求$\boldsymbol{V}$为正数，故$\dfrac{1}{\boldsymbol{V}}$也是正数，可分解为$\dfrac{1}{\sqrt{\boldsymbol{V}}}\cdot\dfrac{1}{\sqrt{\boldsymbol{V}}}$；如果$\boldsymbol{V}$为负数，则无法进行此分解。

矩阵$\boldsymbol{C}$不唯一，但不影响 GLS 的最终结果。

将原回归模型$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$两边同时左乘矩阵$\boldsymbol{C}$：

$$\boldsymbol{Cy}=\boldsymbol{CX\beta}+\boldsymbol{C\varepsilon}$$

定义变量转换：

$$\tilde{\boldsymbol{y}}\equiv\boldsymbol{Cy},\ \tilde{\boldsymbol{X}}\equiv\boldsymbol{CX},\ \tilde{\boldsymbol{\varepsilon}}\equiv\boldsymbol{C\varepsilon}$$

可将模型写为

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$$

变换后的模型仍满足严格外生性：

$$\mathrm{E}(\tilde{\boldsymbol{\varepsilon}} \mid \tilde{\boldsymbol{X}}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon} \mid \boldsymbol{C}\boldsymbol{X}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \boldsymbol{C}\,\mathrm{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \mathbf{0}$$

球型扰动项的假定也得到满足：

$$\begin{aligned}\mathrm{Var}(\tilde{\boldsymbol{\varepsilon}} \mid \tilde{\boldsymbol{X}}) &= \mathrm{E}(\tilde{\boldsymbol{\varepsilon}}\,\tilde{\boldsymbol{\varepsilon}}' \mid \boldsymbol{X}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}'\boldsymbol{C}' \mid \boldsymbol{X}) = \boldsymbol{C}\,\mathrm{E}(\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}' \mid \boldsymbol{X})\boldsymbol{C}' = \sigma^2\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}' \\ &= \sigma^2\boldsymbol{C}(\boldsymbol{V}^{-1})^{-1}\boldsymbol{C}' = \sigma^2\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}' = \sigma^2\boldsymbol{C}\boldsymbol{C}^{-1}(\boldsymbol{C}')^{-1}\boldsymbol{C}' = \sigma^2\boldsymbol{I}_n\end{aligned}$$

故高斯-马尔可夫定理成立。

对变换后的模型使用 OLS 即得到 GLS 估计量：

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{GLS}} &= (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} = \left[(\boldsymbol{C}\boldsymbol{X})'(\boldsymbol{C}\boldsymbol{X})\right]^{-1}(\boldsymbol{C}\boldsymbol{X})'\boldsymbol{C}\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}\end{aligned}$$

虽然$\boldsymbol{C}$不唯一，但$\hat{\boldsymbol{\beta}}_{\text{GLS}}$唯一，因为$\hat{\boldsymbol{\beta}}_{\text{GLS}}$不依赖于$\boldsymbol{C}$。

由于高斯-马尔可夫定理成立，故$\hat{\boldsymbol{\beta}}_{\text{GLS}}$是 BLUE，比 OLS 更有效。但前提是必须知道协方差矩阵$\boldsymbol{V}$。

3．加权最小二乘法(WLS)

假设仅存在异方差，无自相关，$\boldsymbol{V}(\boldsymbol{X})$为对角矩阵。

方差小的数据提供的信息量大。WLS 根据信息量大小进行加权。

假定$\mathrm{E}(\varepsilon_i^2 \mid \boldsymbol{x}_i) = \mathrm{Var}(\varepsilon_i \mid \boldsymbol{x}_i) = \sigma^2 v_i(\boldsymbol{X})$，即

$$\boldsymbol{V} = \begin{pmatrix} v_1 & & & 0 \\ & v_2 & & \\ & & \ddots & \\ 0 & & & v_n \end{pmatrix}, \quad \boldsymbol{V}^{-1} = \begin{pmatrix} 1/v_1 & & & 0 \\ & 1/v_2 & & \\ & & \ddots & \\ 0 & & & 1/v_n \end{pmatrix}$$

由于$\boldsymbol{V}^{-1}=\boldsymbol{C}'\boldsymbol{C}$，可知

$$\boldsymbol{C}=\boldsymbol{C}'=\begin{pmatrix} 1/\sqrt{v_1} & & & 0 \\ & 1/\sqrt{v_2} & & \\ & & \ddots & \\ 0 & & & 1/\sqrt{v_n} \end{pmatrix}$$

$$\tilde{\boldsymbol{y}}\equiv\boldsymbol{C}\boldsymbol{y}=\begin{pmatrix} 1/\sqrt{v_1} & & & 0 \\ & 1/\sqrt{v_2} & & \\ & & \ddots & \\ 0 & & & 1/\sqrt{v_n} \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}=\begin{pmatrix} y_1/\sqrt{v_1} \\ y_2/\sqrt{v_2} \\ \vdots \\ y_n/\sqrt{v_n} \end{pmatrix}$$

$$\tilde{\boldsymbol{X}} \equiv \boldsymbol{C}\boldsymbol{X} = \begin{pmatrix} 1/\sqrt{v_1} & & & 0 \\ & 1/\sqrt{v_2} & & \\ & & \ddots & \\ 0 & & & 1/\sqrt{v_n} \end{pmatrix} \begin{pmatrix} x_{11} & \dots & x_{1K} \\ x_{21} & \dots & x_{2K} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nK} \end{pmatrix}$$

$$= \begin{pmatrix} x_{11}/\sqrt{v_1} & \dots & x_{1K}/\sqrt{v_1} \\ x_{21}/\sqrt{v_2} & \dots & x_{2K}/\sqrt{v_2} \\ \vdots & & \vdots \\ x_{n1}/\sqrt{v_n} & \dots & x_{nK}/\sqrt{v_n} \end{pmatrix}$$

故权重为$1/\sqrt{v_i}$(标准差的倒数)。对于第 i 个观测值，回归方程为

$$\frac{y_i}{\sqrt{v_i}} = \beta_1 \frac{x_{i1}}{\sqrt{v_i}} + \beta_2 \frac{x_{i2}}{\sqrt{v_i}} + \cdots + \beta_K \frac{x_{iK}}{\sqrt{v_i}} + \frac{\varepsilon_i}{\sqrt{v_i}}$$

新扰动项为$\varepsilon_i/\sqrt{v_i}$，可将 WLS 视为最小化“加权的残差平方和”：

$$\min_{\tilde{\boldsymbol{\beta}}} \ \mathrm{SSR} = \sum_{i=1}^{n} \left(e_i / \sqrt{v_i} \right)^2 = \sum_{i=1}^{n} \frac{e_i^2}{v_i}$$

从这个角度来看，权重为$1/v_i$，Stata 也是这样约定的。

4．可行广义最小二乘法(Feasible GLS，简记 FGLS)

必须先用样本数据估计$\boldsymbol{V}(\boldsymbol{X})$，然后才能使用 GLS，称为 FGLS 或“可行加权最小二乘法”(Feasible WLS，简记 FWLS)，即

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$$

其中，$\hat{\boldsymbol{V}}$是$\boldsymbol{V}$的一致估计。

$\boldsymbol{V}(\boldsymbol{X})$包含很多参数。实践中，常考虑只有异方差，或只有一阶自相关的情形(参见第 8 章)。

以 FWLS 为例。在作 BP 检验时，通过辅助回归

$e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i$就可获得$\sigma_i^2$的估计值$\hat{\sigma}_i^2$。

为保证$\hat{\sigma}_i^2$为正，假设辅助回归为指数函数的形式：

$$e_i^2 = \sigma^2 \exp(\delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK}) v_i$$

其中，v_i为乘积形式的扰动项。取对数后可得

$$\ln e_i^2 = (\ln \sigma^2 + \delta_1) + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + \ln v_i$$

得到对$\ln e_i^2$的预测值，记为$\ln \hat{\sigma}_i^2$，进而得到拟合值$\hat{\sigma}_i^2 = e^{\ln \hat{\sigma}_i^2}$，然后以$1/\hat{\sigma}_i^2$为权重进行 WLS 估计。

5．究竟使用“OLS + 稳健标准误”还是 FWLS

理论上，GLS 是 BLUE，但 FGLS 既非线性估计，也不是无偏估计，无资格参加 BLUE 的评选。

根据方程(7.22)，$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$，而$\hat{\boldsymbol{V}}$是数据$(\boldsymbol{y},\boldsymbol{X})$的非线性函数，故$\hat{\boldsymbol{\beta}}_{\text{FGLS}}$是$\boldsymbol{y}$的非线性函数，一般来说是有偏的。

FWLS 的优点主要体现在大样本理论中。如果$\hat{\boldsymbol{V}}$是$\boldsymbol{V}$的一致估计，则 FWLS 一致，且在大样本下比 OLS 更有效。

FWLS 的缺点是必须估计条件方差函数$\text{Var}(\varepsilon_i \mid \boldsymbol{x}_i)$，而通常不知道条件方差函数的具体形式。如果该函数的形式设定不正确，则

根据 FWLS 计算的标准误可能失效，导致不正确的统计推断。

使用“OLS + 稳健标准误”的好处是，对回归系数及标准误的估计都一致，不需要知道条件方差函数的形式。Stata 操作十分简单，只要在命令 `reg` 之后加选择项“`robust`”即可。

总之，“OLS + 稳健标准误”更为稳健，而 FWLS 更有效。必须在稳健性与有效性之间做选择。

由于“病情”通常难以诊断，故特效药也可能失效或起反作用。如果对 $\boldsymbol{V}$ 的估计不准，则 FGLS 的性能可能还不如 OLS。Stock and Watson (2011)推荐，大多数情况下应使用“OLS + 稳健标准误”。

第 8 章　自　相　关

8.1　自相关的后果

如果存在$i \neq j$，使得$\mathrm{E}(\varepsilon_i \varepsilon_j \mid \boldsymbol{X}) \neq 0$，即$\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X})$的非主对角线元素不全为 0，则存在“自相关”(autocorrelation)或“序列相关”(serial correlation)。

在有自相关的情况下：

(1) OLS 估计量依然无偏且一致，因为在证明这些性质时，并未用到“无自相关”的假定；

(2) OLS 估计量依然服从渐近正态分布；

(3) OLS 估计量方差$\mathrm{Var}(\boldsymbol{b}\,|\,\boldsymbol{X})$的表达式不再是$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$，因为$\mathrm{Var}(\boldsymbol{\varepsilon}\,|\,\boldsymbol{X}) \neq \sigma^2\boldsymbol{I}$，通常的 t 检验、F 检验也失效了；

(4) 高斯-马尔可夫定理不再成立，OLS 不再是 BLUE。

假设扰动项存在正自相关，即$\mathrm{E}(\varepsilon_i\varepsilon_j\,|\,\boldsymbol{X}) > 0$，参见图 8.1。

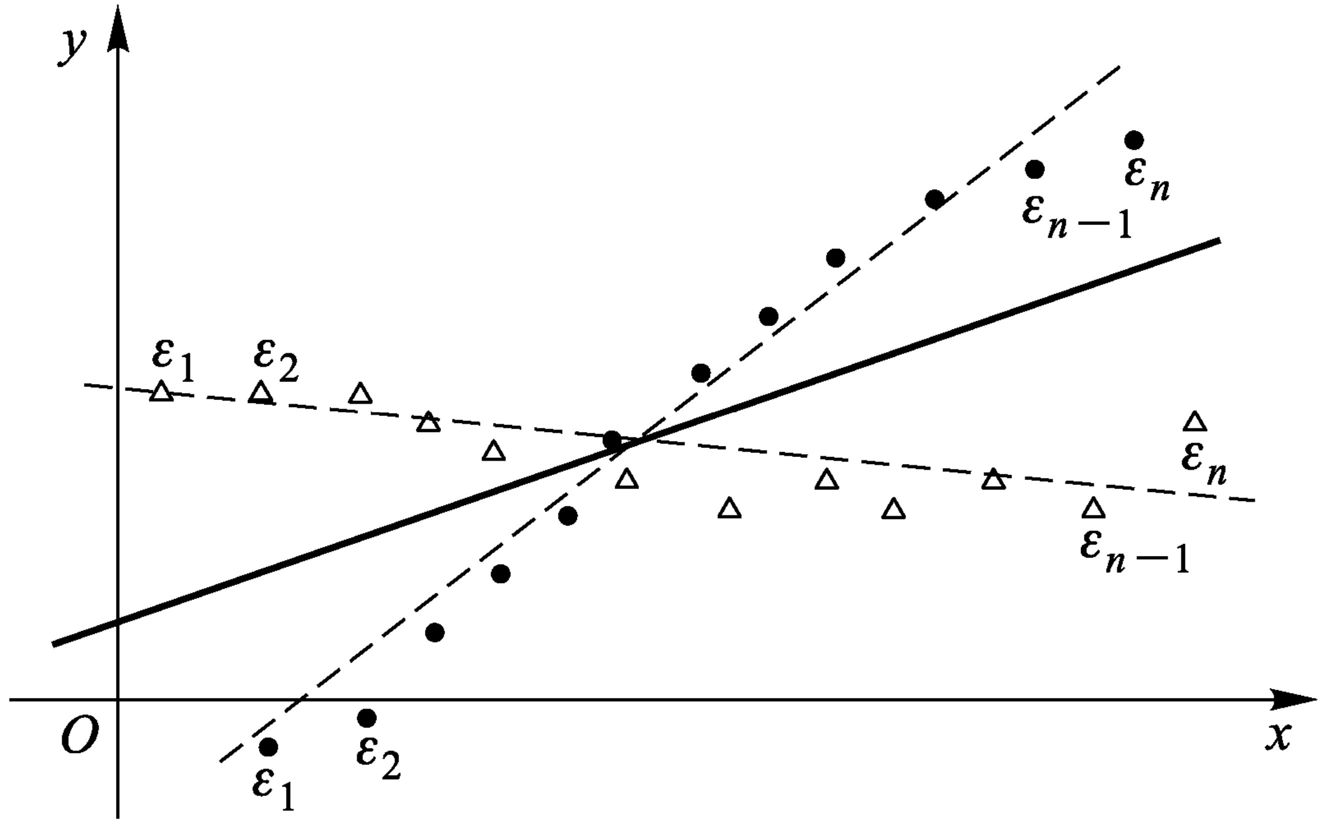

y
ε_1
ε_2
ε_n
ε_{n-1}
ε_n
ε_{n-1}
O
ε_1
ε_2
x

图 8.1　自相关的后果

8.2 自相关的例子

(1) 时间序列的自相关：经济活动具有持久性，比如，相邻两年的 GDP 增长率、通货膨胀率；意外事件或新政策的效应需逐步释放；最优资本存量需若干年投资才能达到(滞后的调整过程)。

(2) 截面数据的自相关：相邻单位间可能存在“溢出效应”(spillover effect or neighborhood effect)，称为“空间自相关”(spatial autocorrelation)。比如，相邻省份、国家间的经济活动相互影响；相邻地区的农产量受类似天气影响；同一社区内房屋价格相关。

(3) 对数据的人为处理：数据中包含移动平均数(moving average)、内插值或季节调整时。

(4) 设定误差(misspecification)：模型设定中遗漏了某个自相关的解释变量，被纳入到扰动项中。

8.3　自相关的检验

1．画图

可将e_t与e_{t-1}画成散点图。

也可画残差的“自相关图”(correlogram)，显示各阶样本自相关系数(命令 ac)或偏自相关系数(命令 pac)。此法虽直观，不严格。

2．BG 检验

对于$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_K x_{tK} + \varepsilon_t$，假设存在一阶自相关，即$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$，其中$u_t$为白噪声，并检验$H_0 : \rho = 0$。

由于可能存在高阶自相关，考虑p阶自回归：

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \cdots + \rho_p \varepsilon_{t-p} + u_t$$

检验$H_0 : \rho_1 = \cdots = \rho_p = 0$。由于$\{\varepsilon_t\}$不可观测，故用$\{e_t\}$替代，并引入所有解释变量，考虑辅助回归：

$$e_t \xrightarrow{\text{OLS}} x_{t1}, \cdots, x_{tK}, e_{t-1}, \cdots, e_{t-p} \quad (t = p+1, \cdots, n)$$

由于使用e_{t-p}，损失 p 个样本值，故样本容量仅为$(n-p)$。

使用nR^2形式的 LM 统计量：

$$(n-p)R^2 \xrightarrow{d} \chi^2(p)$$

如果$(n-p)R^2$超过$\chi^2(p)$的临界值，拒绝“无自相关”的原假设。此检验被称为“Breusch-Godfrey 检验”，简称 BG 检验。

Davidson and MacKinnon(1993)建议：

把残差向量$\boldsymbol{e}$中因滞后而缺失的项，用其期望值$\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$来代替；

保持样本容量仍为 n，使用统计量：

$$nR^2 \xrightarrow{d} \chi^2(p)$$

Davidson-MacKinnon 方法为 Stata 的默认设置。

3．Box–Pierce Q 检验

残差的各阶样本自相关系数：

$$\hat{\rho}_j \equiv \frac{\sum_{t=j+1}^{n} e_t e_{t-j}}{\sum_{t=1}^{n} e_t^2} \quad (j = 1, 2, \cdots, p)$$

如果$H_0 : \rho_1 = \cdots = \rho_p = 0$成立，则$\hat{\rho}_j \xrightarrow{p} 0$，$\sqrt{n}\hat{\rho}_j \xrightarrow{d}$正态分布，$j = 1, 2, \cdots, p$。

残差的各阶样本自相关系数平方和的 n 倍，就是“Box–Pierce Q 统计量”：

$$Q_{\mathrm{BP}} \equiv n \sum_{j=1}^{p} \hat{\rho}_j^2 \xrightarrow{d} \chi^2(p)$$

经改进的“Ljung−Box Q 统计量”：

$$Q_{\mathrm{LB}} \equiv n(n+2)\sum_{j=1}^{p}\frac{\hat{\rho}_j^2}{n-j} \xrightarrow{d} \chi^2(p)$$

这两种 Q 统计量在大样本下等价，但 Ljung−Box Q 统计量的小样本性质更好，为 Stata 所采用。

如何确定自相关阶数 p？如果 p 太小，可能忽略高阶自相关的存在；如果 p 较大，则 Q 统计量的小样本分布可能与 $\chi^2(p)$ 相差较远。

Stata 默认的 p 值为 $p = \min\{\text{floor}(n/2) - 2,\ 40\}$，其中 $\text{floor}(n/2)$ 为不超过 $n/2$ 的最大整数。

4．DW 检验

“DW 检验” (Durbin and Watson, 1950)较早出现，已不常用。只能检验一阶自相关，且要求解释变量满足严格外生性。

DW 检验的统计量为

$$\mathrm{DW} \equiv d \equiv \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{\sum_{t=2}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1} + \sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=1}^{n} e_t^2}$$

$$\approx 2 - 2\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \equiv 2(1 - \hat{\rho}_1)$$

其中，$\hat{\rho}_1$为残差的一阶自相关系数。

当$d = 2$时，$\hat{\rho}_1 \approx 0$，无一阶自相关；

当$d = 0$时，$\hat{\rho}_1 \approx 1$，一阶正自相关；

当$d = 4$时，$\hat{\rho}_1 \approx -1$，一阶负自相关。

DW 统计量依赖于数据矩阵$\boldsymbol{X}$，无法制表，须使用上限分布d_U与下限分布d_L($d_L < d < d_U$)来判断。得到d_U与d_L的临界值后，仍存在无结论区域。

DW 统计量本质就是残差的一阶自相关系数，不能指望它提供太多的信息。

8.4 自相关的处理

1．使用“OLS + 异方差自相关稳健的标准误”

仍用 OLS 来估计回归系数，但使用“异方差自相关稳健的标准误”(Heteroskedasticity and Autocorrelation Consistent Standard Error，简记 HAC)。

此法称为“Newey−West 估计法”(Newey and West, 1987)，只改变标准误的估计值，不改变回归系数的估计值。

为什么第 5 章的“异方差稳健标准误”不适用于自相关的情形？问题出在假定 5.5，即 $\boldsymbol{g}_i \equiv \boldsymbol{x}_i \varepsilon_i$ 为鞅差分序列的假定。

命题 如果回归模型含有截距项，则假定 5.5 意味着扰动项ε_i无自相关。

证明：根据假定 5.5，$\boldsymbol{g}_i$为鞅差分序列，故

$$\mathrm{E}(\boldsymbol{g}_i \mid \boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1) = \mathrm{E}(\boldsymbol{x}_i \varepsilon_i \mid \boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1) = 0$$

因为模型含有截距项，故向量$\boldsymbol{g}_i \equiv \boldsymbol{x}_i \varepsilon_i$的第一个元素为$\varepsilon_i$。因此，$\mathrm{E}(\varepsilon_i \mid \boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1) = 0$。由于$\{\varepsilon_{i-1}, \cdots, \varepsilon_1\} \subset \{\boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1\}$(前者是后者的子集，故前者的信息完全包含于后者之中)，根据迭代期望定律可得

$$\begin{aligned}
\mathrm{E}(\varepsilon_i \mid \varepsilon_{i-1}, \cdots, \varepsilon_1) &= \mathrm{E}_{\boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1}\left[\mathrm{E}(\varepsilon_i \mid \varepsilon_{i-1}, \cdots, \varepsilon_1) \middle| \boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1\right] \\
&= \mathrm{E}_{\boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1}\underbrace{\left[\mathrm{E}(\varepsilon_i \mid \boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1)\right]}_{=0} = 0
\end{aligned}$$

因此，ε_i均值独立于$(\varepsilon_{i-1}, \cdots, \varepsilon_1)$，故扰动项$\varepsilon_i$无自相关。

根据第 5 章，异方差稳健的协方差矩阵$\boldsymbol{S}_{XX}^{-1}\hat{\boldsymbol{S}}\,\boldsymbol{S}_{XX}^{-1}$为夹心估计量，其中$\boldsymbol{S}_{XX} \equiv \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$，$\hat{\boldsymbol{S}} \equiv \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'$。

异方差自相关稳健的协方差矩阵也是夹心估计量，其形式为$\boldsymbol{S}_{XX}^{-1}\hat{\boldsymbol{Q}}\boldsymbol{S}_{XX}^{-1}$，其中

$$\hat{\boldsymbol{Q}} = \hat{\boldsymbol{S}} + \frac{1}{n}\sum_{j=1}^{p}\sum_{t=j+1}^{n}\left(1-\frac{j}{p+1}\right)e_t e_{t-j}(\boldsymbol{x}_t \boldsymbol{x}_{t-j}' + \boldsymbol{x}_{t-j}\boldsymbol{x}_t')$$

p 为自相关的阶数，也称“截断参数”(truncation parameter)。建议令 $p = n^{1/4}$ 或 $p = 0.75 n^{1/3}$，再取整数。

考虑一元回归情形，

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n$$

OLS 估计量为，

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\left[\beta_1 (x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})\right]}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

其中，由于$\overline{y}=\beta_0+\beta_1\overline{x}+\overline{\varepsilon}$，故$y_i-\overline{y}=\beta_1(x_i-\overline{x})+\varepsilon_i-\overline{\varepsilon}$。因此，

$$\hat{\beta}_1-\beta_1=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(\varepsilon_i-\overline{\varepsilon})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

其中，$\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(\varepsilon_i-\overline{\varepsilon})=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\varepsilon_i-\frac{1}{n}\overline{\varepsilon}\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})}_{=0}$。记$v_i\equiv(x_i-\overline{x})\varepsilon_i$，在大样本中，$\hat{\beta}_1-\beta_1\approx\frac{\frac{1}{n}\sum_{i=1}^{n}v_i}{\sigma_x^2}$，其中$\sigma_x^2$为$x_i$的方差。

故在大样本中，

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right)}{\left(\sigma_x^2\right)^2}$$

考虑$n = 2$的最简单情形，则上式分子为，

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right) = \mathrm{Var}\left[\frac{1}{2}(v_1 + v_2)\right] = \frac{1}{4}\left[\mathrm{Var}(v_1) + \mathrm{Var}(v_2) + 2\,\mathrm{Cov}(v_1, v_2)\right]$$

$$= \frac{1}{2}\sigma_v^2 + \frac{1}{2}\rho_1\sigma_v^2 = \frac{1}{2}\sigma_v^2(1 + \rho_1) \equiv \frac{1}{2}\sigma_v^2 f_2$$

其中，$\sigma_v^2 \equiv \mathrm{Var}(v_i)$，$\rho_1 \equiv \mathrm{corr}(v_1, v_2)$为一阶自相关系数，而

$f_2 \equiv (1+\rho_1)$是修正系数。

如不存在自相关，$\rho_1=0$，$f_2=1$，则$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right)=\frac{1}{2}\sigma_v^2$，得到通常的方差公式。

如存在自相关，$\rho_1 \neq 0$，方差公式有所不同。

考虑样本容量为n的一般情况，则$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right)=\frac{1}{n}\sigma_v^2 f_n$，其中$f_n \equiv 1+2\sum_{j=1}^{n-1}\left(\frac{n-j}{n}\right)\rho_j$为对应于样本容量为$n$的修正系数，而$\rho_j$为$j$阶自相关系数；因此，

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_v^2}{n\left(\sigma_x^2\right)^2} \cdot f_n$$

上式是普通方差公式的f_n倍。f_n包含未知的自相关系数ρ_j，需对其进行估计，比如

$$\hat{f}_n \equiv 1 + 2\sum_{j=1}^{n-1}\left(\frac{n-j}{n}\right)\hat{\rho}_j$$

其中，$\hat{\rho}_j$为j阶样本自相关系数。但待估计参数$(\rho_1, \cdots, \rho_{n-1})$太多，且随样本容量$n$增长，导致此估计量不一致。

反之，仅考虑前几阶自相关系数(比如，只考虑ρ_1)的估计量也不一致，因为忽略了高阶自相关。

正确的做法是，包括足够多阶数的自相关系数，并让此阶数p随着样本容量的增长而增长。

一般建议取$p = n^{1/4}$或$p = 0.75n^{1/3}$，作为截断参数。

实践中，建议使用不同的截断参数，考察 HAC 标准误是否对于截断参数的取值敏感。

2．使用“OLS + 聚类稳健的标准误”

如果样本观测值可以分为不同的“聚类”(clusters)，在同一聚类里的观测值互相相关，而不同聚类之间的观测值不相关，这种样本称为“聚类样本”(cluster sample)。

【例】在 Nerlove(1963)对美国电力企业的研究中，同一个州的电力企业可能受到相同州政策的影响而自相关，但不同州之间的电力企业可能不相关。此时，“州”(state)被称为“聚类变量”(cluster variable)。

【例】如果以全班同学为样本，则聚类变量可能是宿舍或专业。

如果将观测值按聚类的归属顺序排列，则扰动项的协方差矩阵为“块对角”(block diagonal)。

仍可用 OLS 来估计系数，但需使用“聚类稳健的标准误”(cluster robust standard errors)。

假设样本容量为 N，包括 M 个聚类，其中第 j 个聚类包含 M_j 个个体。记第 j 个聚类个体 i 的解释变量为 $\boldsymbol{x}_{ij}$，残差为 e_{ij}，然后定义 $\boldsymbol{u}_j \equiv \sum_{i=1}^{M_j} e_{ij}\boldsymbol{x}_{ij}$，则聚类稳健的协方差矩阵可以写为

$$\frac{N-1}{N-K}\frac{M}{M-1}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{j=1}^{M}\boldsymbol{u}_j'\boldsymbol{u}_j\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

其中，$\dfrac{N-1}{N-K}\dfrac{M}{M-1}$为对自由度的调整。

聚类稳健的标准误也是夹心估计量。在推导过程中并未假定同方差，故也是异方差稳健的。

使用聚类稳健标准误的前提是，聚类中的个体数M_j较少，而聚类数很多$(M \to \infty)$；则聚类稳健标准误是真实标准误的一致估计。

处理面板数据时，常使用聚类稳健的标准误。

3．使用可行广义最小二乘法(FGLS)

首先估计$\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X})$。为减少待估参数，假设扰动项为 AR(1)：

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t,\quad |\rho| < 1,\quad u_t \text{为白噪声}$$

记扰动项的j阶协方差$\rho_j \equiv \mathrm{Cov}(\varepsilon_t, \varepsilon_{t-j} \mid \boldsymbol{X})$，则

$$\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \begin{pmatrix} \rho_0 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{n-2} \\ \vdots & \vdots & & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & \rho_0 \end{pmatrix}$$

容易证明，$\rho_0=\sigma^2=\mathrm{Var}(\varepsilon_t)=\dfrac{\sigma_u^2}{1-\rho^2}$，其中$\sigma_u^2\equiv\mathrm{Var}(u_t)$。

$\rho_1=\rho\sigma^2$，故$\dfrac{\rho_1}{\rho_0}=\dfrac{\rho\sigma^2}{\sigma^2}=\rho$为一阶自相关系数；$\rho_2=\rho^2\sigma^2$，…，$\rho_{n-1}=\rho^{n-1}\sigma^2$，故

$$\mathrm{Var}(\boldsymbol{\varepsilon}\,|\,\boldsymbol{X})=\sigma^2\begin{pmatrix}1 & \rho & \dots & \rho^{n-1}\\ \rho & 1 & \dots & \rho^{n-2}\\ \vdots & \vdots & & \vdots\\ \rho^{n-1} & \rho^{n-2} & \dots & 1\end{pmatrix}\equiv\sigma^2\boldsymbol{V}$$

只要估计唯一的参数ρ，就可使用 FGLS。Stata 默认的估计方

法为使用 OLS 对残差进行辅助回归，$e_t = \hat{\rho} e_{t-1} + error_t$。也可通过残差一阶自相关系数$\hat{\rho} = \dfrac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$，或$\hat{\rho} = 1 - \dfrac{\mathrm{DW}}{2}$来估计$\rho$。

并将 $\boldsymbol{V}$ 的逆矩阵分解为$\boldsymbol{V}^{-1} = \boldsymbol{C}'\boldsymbol{C}$。可以证明

$$\boldsymbol{C} = \frac{1}{\sqrt{1-\rho^2}} \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \dots & 0 & 0 \\ -\rho & 1 & \dots & 0 & 0 \\ 0 & -\rho & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & -\rho & 1 \end{pmatrix}$$

以$\sqrt{1-\rho^2}\boldsymbol{C}$左乘原模型，并定义$\tilde{\boldsymbol{y}} \equiv \sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{y}$，$\tilde{\boldsymbol{X}} \equiv \sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{X}$，$\tilde{\boldsymbol{\varepsilon}} \equiv \sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{\varepsilon}$，则变换后的扰动项$\tilde{\boldsymbol{\varepsilon}}$满足球型扰动项的假设，故高斯-马尔可夫定理成立(此变换是 GLS 的特例)：

$$\tilde{\boldsymbol{y}} = \sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{y} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \dots & 0 & 0 \\ -\rho & 1 & \dots & 0 & 0 \\ 0 & -\rho & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & -\rho & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sqrt{1-\rho^2}\, y_1 \\ y_2 - \rho y_1 \\ \vdots \\ y_n - \rho y_{n-1} \end{pmatrix}$$

$$\tilde{\boldsymbol{X}} = \sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{X} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \dots & 0 & 0 \\ -\rho & 1 & \dots & 0 & 0 \\ 0 & -\rho & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & -\rho & 1 \end{pmatrix} \begin{pmatrix} x_{11} & \dots & x_{1K} \\ x_{21} & \dots & x_{2K} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nK} \end{pmatrix}$$

$$= \begin{pmatrix} \sqrt{1-\rho^2}x_{11} & \dots & \sqrt{1-\rho^2}x_{1K} \\ x_{21}-\rho x_{11} & \dots & x_{2K}-\rho x_{1K} \\ \vdots & & \vdots \\ x_{n1}-\rho x_{n-1,1} & \dots & x_{nK}-\rho x_{n-1,K} \end{pmatrix}$$

$$\tilde{\boldsymbol{\varepsilon}} = \sqrt{1-\rho^2}\,\boldsymbol{C}\boldsymbol{\varepsilon} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \dots & 0 & 0 \\ -\rho & 1 & \dots & 0 & 0 \\ 0 & -\rho & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & -\rho & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \sqrt{1-\rho^2}\,\varepsilon_1 \\ \varepsilon_2 - \rho\varepsilon_1 \\ \vdots \\ \varepsilon_n - \rho\varepsilon_{n-1} \end{pmatrix}$$

写出每个观测值(个体)的回归方程:

$$\sqrt{1-\rho^2}\,y_1 = \sqrt{1-\rho^2}\,\beta_1 + \sqrt{1-\rho^2}\,\beta_2 x_{12} + \cdots + \sqrt{1-\rho^2}\,\beta_K x_{1K} + \tilde{\varepsilon}_1$$

$$y_2 - \rho y_1 = (1-\rho)\beta_1 + \beta_2(x_{22} - \rho x_{12}) + \cdots + \beta_K(x_{2K} - \rho x_{1K}) + \tilde{\varepsilon}_2$$

$$\cdots\cdots\cdots\cdots$$

$$y_n - \rho y_{n-1} = (1-\rho)\beta_1 + \beta_2(x_{n2} - \rho x_{n-1,\,2}) + \cdots + \beta_K(x_{nK} - \rho x_{n-1,\,K}) + \tilde{\varepsilon}_n$$

第一个方程的形式与其他方程不同。用 OLS 估计变换后的模型，即为“Prais−Winsten 估计法”(简记 PW)。

为计算方便，将第一个方程删去，称为“Cochrane−Orcutt 估计法”(简记 CO)。该法有更简洁的推导过程。原模型为，

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + \varepsilon_t$$

将上式滞后一期，然后方程两边同时乘以ρ得

$$\rho y_{t-1} = \rho\beta_1 + \rho\beta_2 x_{t-1,\,2} + \cdots + \rho\beta_K x_{t-1,\,K} + \rho\varepsilon_{t-1}$$

将两方程相减可得：

$$y_t - \rho y_{t-1} = (1-\rho)\beta_1 + \beta_2(x_{t2} - \rho x_{t-1,\,2}) + \cdots + \beta_K(x_{tK} - \rho x_{t-1,\,K}) + \varepsilon_t - \rho\varepsilon_{t-1}$$

新扰动项$\varepsilon_t - \rho\varepsilon_{t-1} = u_t$满足球型扰动项的古典假定。

此法也称“准差分法”(quasi differences)。

在操作中，常使用迭代法，首先用 OLS 估计原模型，作辅助回归得到$\hat{\rho}^{(1)}$(对ρ的第一轮估计)，再用$\hat{\rho}^{(1)}$进行 FGLS 估计，使用新的残差估计$\hat{\rho}^{(2)}$(对ρ的第二轮估计)，再用$\hat{\rho}^{(2)}$进行 FGLS 估计，……，直至收敛。

使用 FGLS 处理自相关，如果对自相关系数的估计较准确，且满足严格外生性的假定，则 FGLS 比 OLS 更有效率。

如果不满足严格外生性，而仅满足前定解释变量的假定，则

FGLS 可能不一致，尽管 OLS 依然一致。

使用准差分法时，变换后的新扰动项为$(\varepsilon_t - \rho\varepsilon_{t-1})$，而新解释变量为$(x_{tk} - \rho x_{t-1,k})$，二者可能存在相关性，导致不一致估计。

总之，FGLS 不如 OLS 稳健。

4．修改模型设定

自相关的深层原因可能是模型设定有误，比如，遗漏了自相关的解释变量；或将动态模型(解释变量中包含被解释变量的滞后值)误设为静态模型，而后者也可视为遗漏了解释变量。

假设真实模型为

$$y_t = \rho y_{t-1} + \boldsymbol{x}_t' \boldsymbol{\beta} + \varepsilon_t$$

由于y_t是y_{t-1}的函数，故$\{y_t\}$存在自相关。假设这个模型被错误地估计成

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta} + \underbrace{\left[\rho y_{t-1} + \varepsilon_t\right]}_{=v_t}$$

ρy_{t-1}被纳入到扰动项v_t中，导致扰动项$\{v_t\}$自相关。

时间序列的自相关，有时可通过引入被解释变量的滞后值来消除。由于模型设定误差而导致的自相关，最好改进模型。