第 7 章　异方差

7.1　异方差的后果

条件异方差(conditional heteroskedasticity)，简称异方差(heteroskedasticity)，是违背球型扰动项假设的一种情形，即条件方差$\mathrm{Var}(\varepsilon_i \mid \mathbf{X})$依赖于$i$，而不是常数$\sigma^2$。

在存在异方差的情况下：

(1) OLS 估计量依然是无偏、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$\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X})$的表达式不再是$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$，因为$\mathrm{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X}) \neq \sigma^2\mathbf{I}$。因此，使用普通标准误的$t$检验、$F$检验失效。

(3) 高斯-马尔可夫定理不再成立，OLS 不再是 BLUE(最佳线性无偏估计)。

在存在异方差的情况下，本章将要介绍的“加权最小二乘法”才是 BLUE。

考虑一元回归$y_i = \alpha + \beta x_i + \varepsilon_i$，并假设$\mathrm{Var}(\varepsilon_i \mid \mathbf{X})$是解释变量$x_i$的增函数，即$x_i$越大则$\mathrm{Var}(\varepsilon_i \mid \mathbf{X})$越大，参见图 7.1。

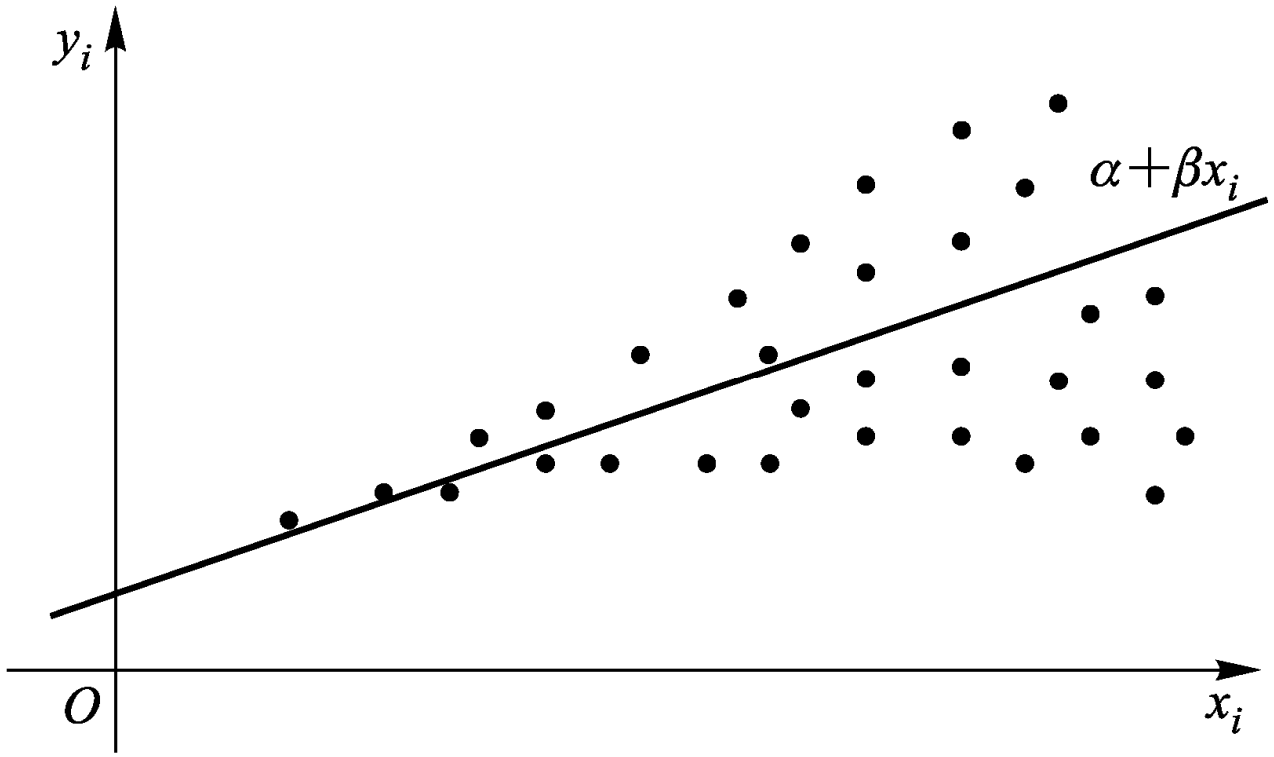

图 7.1　异方差示意图

OLS 回归线在x_i较小时可以较精确地估计，而在x_i较大时则难以准确估计。

方差较大的数据包含的信息量较小，但 OLS 却对所有的数据等量齐观地进行处理。

异方差的存在使得 OLS 的效率降低。

加权最小二乘法(Weighted Least Square，简记 WLS)正是通过对不同数据所包含信息量的不同进行相应的处理以提高估计效率的。比如，给予信息量大的数据更大的权重。

计量经济学所指的“异方差”都是“条件异方差”，而非“无条件异方差”。

大样本 OLS 理论要求样本数据为平稳过程，而平稳过程的方差为不变。

这是否意味着，大样本 OLS 理论已经假设了同方差？

这里关键要区分**无条件方差**(unconditional variance)与**条件方差**(conditional variance)。

以一元回归模型$y_i = \alpha + \beta x_i + \varepsilon_i$为例，假设$\{x_i, y_i\}$为平稳过程。

则$\varepsilon_i = y_i - \alpha - \beta x_i$也是平稳过程，故其无条件方差$\mathrm{Var}(\varepsilon_i) = \sigma^2$为常数，不随 i 而变。

所有个体的条件方差函数$\mathrm{Var}(\varepsilon_i \mid x_1,\cdots,x_n)$在函数形式上也完全相同；比如，$\mathrm{Var}(\varepsilon_i \mid x_1,\cdots,x_n)=x_i^2$。

但此条件方差函数的具体取值却依赖于x_i，故仍可存在条件异方差。

比如，$\mathrm{Var}(\varepsilon_1 \mid x_1,\cdots,x_n)=x_1^2$，$\mathrm{Var}(\varepsilon_2 \mid x_1,\cdots,x_n)=x_2^2$，以此类推。

7.2 异方差的例子

(1) 考虑以下消费函数：

$$c_i=\alpha+\beta y_i+\varepsilon_i \qquad (7.1)$$

其中，c_i为消费，y_i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，很少变动。

富人的消费支出可能更难测量，故包含较多测量误差。

$\mathrm{Var}(\varepsilon_i \mid y_i)$可能随$y_i$的上升而变大。

(2) 企业的投资、销售收入与利润：大型企业的商业活动可能动辄以亿元计，而小型企业则以万元计；因此，扰动项的规模也不相同。

若将大、中、小型企业放在一起回归，则可能存在异方差。

(3) 组间异方差：如果样本包含两组(类)数据，则可能存在组内同方差，但组间异方差的情形。

比如，第一组为自我雇佣者(企业主、个体户)的收入，而第二组为打工族的收入；则自我雇佣者的收入波动可能比打工族更大。

(4) 组平均数：如果数据本身就是组平均数，则大组平均数的方差通常要比小组平均数的方差小。

比如，考虑全国各省的人均 GDP，每个省一个数据。显然，人口较多的省份其方差较小，方差与人口数成反比。

7.3 异方差的检验

1．画残差图(residual plot)

由于残差可视为扰动项的实现值，故可通过残差的波动来大致考察是否存在异方差。

可以看残差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)，但不严格。

2．BP 检验(Breusch and Pagan, 1979)

判断异方差的严格方法仍须通过统计检验。假设回归模型为

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \quad (7.2)$$

记$\mathbf{x}_i = (1 \;\; x_{i2} \;\; \cdots \;\; x_{iK})$。

假设样本数据为 iid，则$\mathrm{Var}(\varepsilon_i \mid \mathbf{X}) = \mathrm{Var}(\varepsilon_i \mid \mathbf{x}_i)$。“条件同方差”的原假设为

$$H_0 : \; \mathrm{Var}(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2 \qquad (7.3)$$

由于$\mathrm{Var}(\varepsilon_i \mid \mathbf{x}_i) = \mathrm{E}(\varepsilon_i^2 \mid \mathbf{x}_i) - [\underbrace{\mathrm{E}(\varepsilon_i \mid \mathbf{x}_i)}_{=0}]^2 = \mathrm{E}(\varepsilon_i^2 \mid \mathbf{x}_i)$，故可将原假设写为

$$H_0:\ \mathrm{E}(\varepsilon_i^2 \mid \mathbf{x}_i) = \sigma^2 \qquad (7.4)$$

如果H_0不成立，则条件方差$\mathrm{E}(\varepsilon_i^2 \mid \mathbf{x}_i)$是$\mathbf{x}_i$的函数，称为**条件方差函数**(conditional variance function)。

假设此条件方差函数为线性函数：

$$\varepsilon_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + u_i \qquad (7.5)$$

故原假设可简化为

$$H_0 : \delta_2 = \cdots = \delta_K = 0 \qquad (7.6)$$

在方程(7.5)中，由于扰动项ε_i不可观测，故使用残差平方e_i^2替代之，进行以下**辅助回归**(auxiliary regression)：

$$e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i \qquad (7.7)$$

记此辅助回归的拟合优度为R^2。

R^2越高，则此辅助回归方程越显著，越可以拒绝$H_0 : \delta_2 = \cdots = \delta_K = 0$。

Breusch and Pagan(1979)使用*LM*统计量，进行*LM*检验(Lagrange Multiplier Test, 参见第 11 章)：

$$LM = nR^2 \xrightarrow{d} \chi^2(K-1) \qquad (7.8)$$

若 LM 统计量大于 $\chi^2(K-1)$ 的临界值，则拒绝同方差的原假设。

为什么 LM 统计量是 nR^2 呢？

在大样本中，nR^2 与检验整个回归方程显著性的 F 统计量是渐近等价的。

首先，对于辅助回归(7.7)，检验原假设“$H_0 : \delta_2 = \cdots = \delta_K = 0$”的 F 统计量为(参见第 5 章)

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F(K-1, n-K) \qquad (7.9)$$

其次，在大样本情况下，F 分布与 χ^2 分布是等价的(参见第 6 章)，即 $(K-1)F = \frac{(n-K)R^2}{(1-R^2)} \xrightarrow{d} \chi^2(K-1)$。

在 $H_0 : \delta_2 = \cdots = \delta_K = 0$ 成立的情况下，辅助回归方程(7.7)仅对常数项回归，故当 $n \to \infty$ 时，$R^2 \xrightarrow{p} 0$，而 $(1-R^2) \xrightarrow{p} 1$。因此，

$$(K-1)F = \frac{(n-K)R^2}{1-R^2} \xrightarrow{p} (n-K)R^2 \qquad (7.10)$$

在大样本下，$(n-K)R^2$与nR^2并无差别，故LM检验与F检验渐近等价。

如果认为异方差主要依赖于被解释变量的拟合值，可将辅助回归(7.7)改为

$$e_i^2 = \delta_1 + \delta_2 \hat{y}_i + error_i \qquad (7.11)$$

其中，$\hat{y}_i$为回归方程(7.2)的拟合值；然后检验$H_0 : \delta_2 = 0$(可使用F或LM统计量)。

Breusch and Pagan(1979)的最初检验假设扰动项ε_i服从正态分布。

Koenker (1981)将此假定减弱为独立同分布(iid)，在实际中较多采用。

3．怀特检验(White,1980)

BP 检验假设条件方差函数为线性函数，只是对条件方差函数的一阶近似，可能忽略了高次项。

怀特检验(White, 1980)在 BP 检验的辅助回归(7.7)中加入所有的二次项(含平方项与交叉项)。

不失一般性，考虑以下二元回归：

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad (7.12)$$

其中，除常数项外，只有两个解释变量x_{i2}与x_{i3}，故二次项包括x_{i2}^2，x_{i3}^2与$x_{i2}x_{i3}$。

怀特检验的辅助回归为

$$e_i^2 = \delta_1 + \delta_2 x_{i2} + \delta_3 x_{i3} + \delta_4 x_{i2}^2 + \delta_5 x_{i3}^2 + \delta_6 x_{i2} x_{i3} + error_i \qquad (7.13)$$

其中，e_i^2为回归方程(7.12)的残差平方。

对原假设“$H_0 : \delta_2 = \cdots = \delta_6 = 0$”进行 F 检验或 LM 检验。

怀特检验的优点：可检验任何形式的异方差；因为根据泰勒展开式(Taylor expansion)，二次函数可以很好地逼近任何光滑函数。

怀特检验的缺点：若解释变量较多，解释变量的二次项(含交叉项)将更多，在辅助回归中将损失较多有效样本容量以及自由度。

7.4 异方差的处理

1. 使用“OLS + 稳健标准误”

如果发现存在异方差，一种处理方法是，仍然进行 OLS 回归(OLS 仍然无偏、一致且渐近正态)，但使用在异方差情况下也成立的稳健标准误。

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，只要使用稳健标准误，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 更有效率的方法，比如 WLS。

2．加权最小二乘法(WLS)

由于方差较小的观测值包含的信息量较大，故对于异方差的另一处理方法是，给予方差较小的观测值较大的权重，然后进行加权最小二乘法估计。

对于存在异方差的数据，WLS 的基本思想是，通过变量转换，使得变换后的模型满足球形扰动项的假定(变为同方差)，然后进行 OLS 估计，即为最有效率的 BLUE。

考虑线性回归模型：

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \qquad (7.14)$$

假定$\mathrm{Var}(\varepsilon_i \mid \mathbf{x}_i) \equiv \sigma_i^2 = \sigma^2 v_i$，而且个体 i 的异方差因子$\{v_i\}_{i=1}^n$为已知。

在上式两边同时乘以权重$1/\sqrt{v_i}$(个体 i 的标准差倒数)可得

$$\frac{y_i}{\sqrt{v_i}} = \beta_1 \frac{1}{\sqrt{v_i}} + \beta_2 \frac{x_{i2}}{\sqrt{v_i}} + \cdots + \beta_K \frac{x_{iK}}{\sqrt{v_i}} + \frac{\varepsilon_i}{\sqrt{v_i}} \qquad (7.15)$$

新扰动项$\varepsilon_i/\sqrt{v_i}$不再存在异方差，因为

$$\mathrm{Var}\left(\varepsilon_i/\sqrt{v_i}\right) = \frac{1}{v_i}\mathrm{Var}(\varepsilon_i) = \frac{\sigma^2 v_i}{v_i} = \sigma^2 \qquad (7.16)$$

对方程(7.15)进行 OLS 回归，即为 WLS。

加权之后的回归方程满足球形扰动项的假定，故是 BLUE。

根据方程(7.15)，可将 WLS 定义为最小化“加权的残差平方和”：

$$\min \sum_{i=1}^{n}\left(e_i/\sqrt{v_i}\right)^2=\sum_{i=1}^{n}\frac{e_i^2}{v_i} \qquad (7.17)$$

WLS 的权重为$1/v_i$ (即方差的倒数)，在 Stata 中也是这样约定的。

加权最小二乘法的R^2通常没有太大意义。

它衡量的是变换之后的解释变量$\left(x_{ik}/\sqrt{v_i}\right)$对变换之后的被解释变量$\left(y_i/\sqrt{v_i}\right)$的解释力，而我们一般对此没有太大兴趣。

3．可行加权最小二乘法(FWLS)

使用 WLS 虽可得到 BLUE 估计，但必须知道每位个体的方差，即$\left\{\sigma_i^2\right\}_{i=1}^n$。

在实践中，我们通常不知道$\left\{\sigma_i^2\right\}_{i=1}^n$，故 WLS 事实上“不可行”(infeasible)。

解决方法是先用样本数据估计$\left\{\sigma_i^2\right\}_{i=1}^n$，然后再使用 WLS，称为**可行加权最小二乘法**(Feasible WLS，简记 FWLS)。

在作 BP 检验时，进行如下辅助回归：

$$e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i \qquad (7.18)$$

其中，e_i^2为原方程(7.14)的残差平方。通过此辅助回归的拟合值，即可获得σ_i^2的估计值：

$$\hat{\sigma}_i^2 = \hat{\delta}_1 + \hat{\delta}_2 x_{i2} + \cdots + \hat{\delta}_K x_{iK} \qquad (7.19)$$

上式中可能出现“$\hat{\sigma}_i^2 < 0$”的情形，而方差不能为负。

为保证$\hat{\sigma}_i^2$始终为正，一般假设条件方差函数为对数形式：

$$\ln e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i \qquad (7.20)$$

对此方程进行 OLS 回归，可得$\ln e_i^2$的预测值，记为$\ln \hat{\sigma}_i^2$。

得到拟合值$\hat{\sigma}_i^2 = \exp(\ln \hat{\sigma}_i^2)$(一定为正数)，然后以$1/\hat{\sigma}_i^2$为权重对原方程(7.14)进行 WLS 估计，记此估计量为$\hat{\boldsymbol{\beta}}_{\mathrm{FWLS}}$。

4．究竟使用“OLS + 稳健标准误”还是 FWLS

在理论上，WLS 是 BLUE。

但实践中使用的 FWLS 并非线性估计，因为权重$1/\hat{\sigma}_i^2$也是$\mathbf{y}$的函数。由于$\hat{\boldsymbol{\beta}}_{\mathrm{FWLS}}$是$\mathbf{y}$的非线性函数，一般来说是有偏的。

$\hat{\boldsymbol{\beta}}_{\mathrm{FWLS}}$甚至无资格参加 BLUE 的评选。

FWLS 的优点主要体现在大样本理论中。

如果$\hat{\sigma}_i^2$是σ_i^2的一致估计，则 FWLS 一致，且在大样本下比 OLS 更有效率。

FWLS 的缺点是必须估计条件方差函数$\hat{\sigma}_i^2(\mathbf{x}_i)$，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，它对回归系数及标准误的估计都是一致的，并不需要知道条件方差函数的形式。

在 Stata 中的操作也十分简单，只要在命令 `reg` 之后加选择项“`robust`”即可。

“OLS ＋ 稳健标准误”更为稳健(即适用于更一般的情形)，而FWLS 更有效率。

必须在稳健性与有效性之间做选择。

前者相当于“万金油”(指谁都适用)，而后者相当于“特效药”。

由于“病情”通常难以诊断(无法判断条件异方差的具体形式)，故特效药也可能失效，甚至起反作用。

如果对σ_i^2估计不准确，则 FWLS 即使在大样本下也不是 BLUE，其估计效率可能还不如 OLS。

Stock and Watson (2012)推荐，在大多数情况下应使用“OLS + 稳健标准误”。Wooldridge(2009)指出，如果确实存在严重的异方差，则可通过使用 FWLS 来提高估计效率。

若对于条件异方差函数的具体形式没有把握，可在进行 WLS 回归时仍使用异方差稳健的标准误，以保证 FWLS 标准误的有效性。

如果被解释变量取值为正，有时将被解释变量取对数，可以缓解异方差问题(参见习题)。

7.5 处理异方差的 Stata 命令及实例

以数据集 nerlove.dta 为例(参见第 6 章)，演示如何在 Stata 中处理异方差。

此数据集包括以下变量：*tc* (总成本)，*q* (总产量)，*pl* (工资率)，*pk* (资本的使用成本) 与 *pf* (燃料价格)，以及相应的对数值 *lntc*，*lnq*，*lnpl*，*lnpk* 与 *lnpf*。

1. 画残差图

完成回归后，可使用以下命令得到残差图：

```
rvfplot                (residual-versus-fitted plot)
rvpplot varname  (residual-versus-predictor plot)
```

首先，打开数据集 nerlove.dta，并以 OLS 估计对数形式的成本函数：

```
. use nerlove.dta,clear
. reg lntc lnq lnpl lnpk lnpf
```

Source	SS	df	MS	Number of obs	=	145
				F(4, 140)	=	437.90
Model	269.524728	4	67.3811819	Prob > F	=	0.0000
Residual	21.5420958	140	.153872113	R-squared	=	0.9260
				Adj R-squared	=	0.9239
Total	291.066823	144	2.02129738	Root MSE	=	.39227

lntc	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lnq	.7209135	.0174337	41.35	0.000	.6864462	.7553808
lnpl	.4559645	.299802	1.52	0.131	-.1367602	1.048689
lnpk	-.2151476	.3398295	-0.63	0.528	-.8870089	.4567136
lnpf	.4258137	.1003218	4.24	0.000	.2274721	.6241554
_cons	-3.566513	1.779383	-2.00	0.047	-7.084448	-.0485779

为初步考察异方差，画残差与拟合值的散点图，结果参见图 7.2。

```
. rvfplot
```

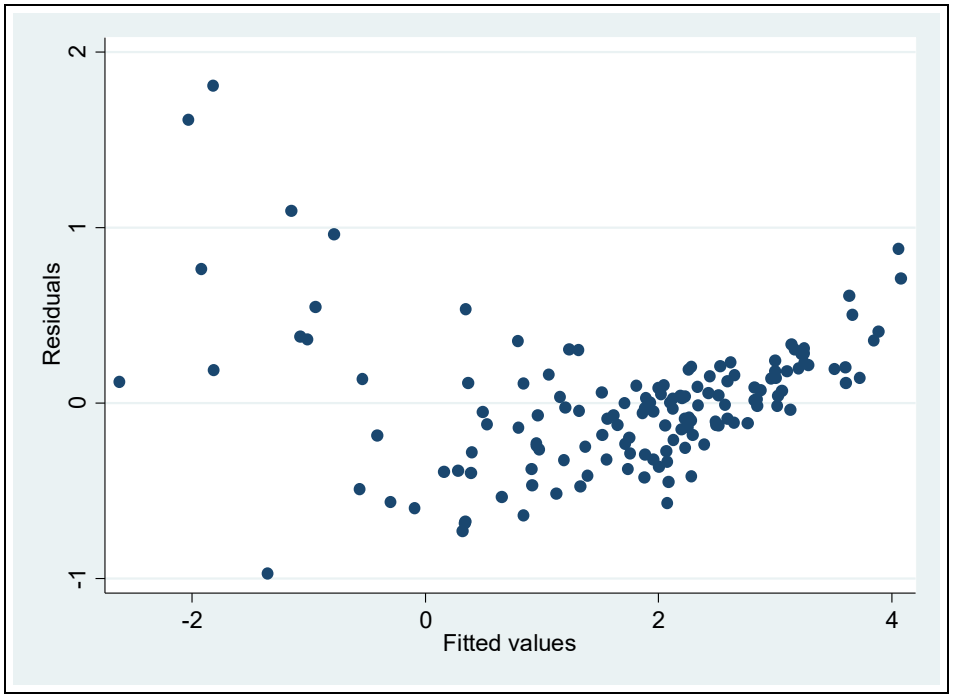

图 7.2　残差与拟合值的散点图

当总成本对数(*lntc* 的拟合值)较小时，扰动项的方差较大。

进一步考察残差与解释变量 *lnq* 的散点图，结果参见图 7.3。

```
. rvpplot lnq
```

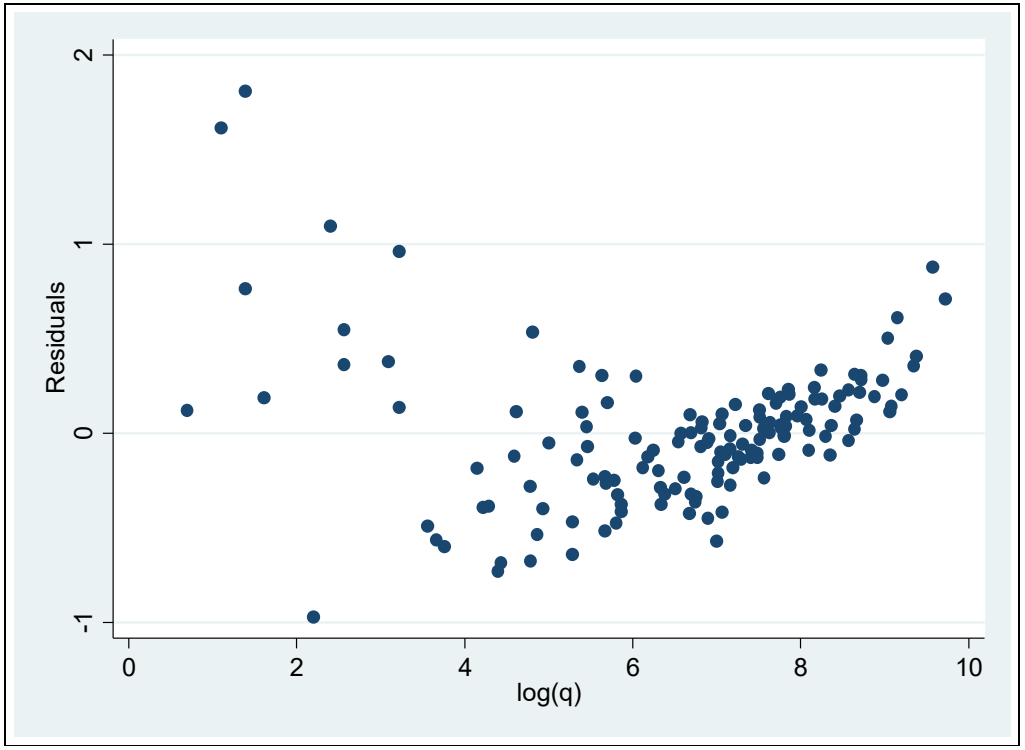

图 7.3　残差与解释变量 *lnq* 的散点图

当总产量对数(*lnq*)越小时，扰动项的方差越大。

2．BP 检验

在 Stata 中完成回归后，可使用以下命令进行 BP 检验：

```
estat hettest,iid rhs
```

“estat”指 post-estimation statistics(估计后统计量)，即在完成估计后所计算的后续统计量；

“hettest”表示 heteroskedasticity test。

选择项“iid”表示仅假定数据为 iid，而无须正态假定。

选择项“rhs”表示，使用方程右边的全部解释变量进行辅助回归，默认使用拟合值$\hat{y}$进行辅助回归。

如果想指定使用某些解释变量进行辅助回归，可使用如下命令：

```
estat hettest [varlist],iid
```

“[varlist]”为指定的变量清单；而“[]”表示其中的内容可出现在命令中，也可不出现。

回到 Nerlove(1963)的例子：

```
. quietly reg lntc lnq lnpl lnpk lnpf
```

其中，前缀(prefix)“quietly”表示执行此命令，但不在 Stata 的结果窗口显示运行结果。

首先，使用拟合值$\hat{y}$进行 BP 检验。

```
. estat hettest, iid
```

```
Breusch–Pagan/Cook–Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variable: Fitted values of lntc

H0: Constant variance

    chi2(1) =  29.13
Prob > chi2 = 0.0000
```

其次，使用所有解释变量进行 BP 检验。

```
. estat hettest, iid rhs
```

```
Breusch–Pagan/Cook–Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: All independent variables

H0: Constant variance

    chi2(4) =  36.16
Prob > chi2 = 0.0000
```

最后，使用变量 *lnq* 进行 BP 检验。

```
. estat hettest lnq,iid
```

```
Breusch–Pagan/Cook–Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variable: lnq

H0: Constant variance

    chi2(1) =  32.10
Prob > chi2 = 0.0000
```

以上各种形式 BP 检验的 p 值都等于 0.0000，故强烈拒绝同方差的原假设，认为存在异方差。

3．怀特检验

在 Stata 完成回归后，可使用如下命令进行怀特检验：

```
estat imtest,white
```

其中，“`imtest`”指 information matrix test(信息矩阵检验)。

继续以 Nerlove(1963)为例：

```
. estat imtest,white
White's test
H0: Homoskedasticity
Ha: Unrestricted heteroskedasticity

   chi2(14) =  73.88
Prob > chi2 = 0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	73.88	14	0.0000
Skewness	22.79	4	0.0001
Kurtosis	2.62	1	0.1055
Total	99.29	19	0.0000

p值(Prob>chi2)等于 0.0000，故强烈拒绝同方差的原假设，认为存在异方差。

4. WLS

在得到扰动项方差的估计值$\left\{\hat{\sigma}_i^2\right\}_{i=1}^n$后，可作为权重进行 WLS 估计。

假设已把$\left\{\hat{\sigma}_i^2\right\}_{i=1}^n$存储在变量 `var` 上，则可通过如下 Stata 命令来实现 WLS：

```
reg y x1 x2 x3 [aw=1/var]
```

其中，“`aw`”表示 analytical weight，为扰动项方差的倒数。

继续以 Nerlove(1963)为例。

首先计算残差，并记为 e1：

```
. quietly reg lntc lnq lnpl lnpk lnpf
. predict e1,residual
```

其次，生成残差的平方，并记为 e2：

```
. gen e2=e1^2
```

将残差平方取对数，

```
. gen lne2=log(e2)
```

假设$\ln\hat{\sigma}_i^2$为变量 *lnq* 的线性函数，进行以下辅助回归：

```
. reg lne2 lnq
```

Source	SS	df	MS	Number of obs	=	145
				F(1, 143)	=	21.54
Model	105.722127	1	105.722127	Prob > F	=	0.0000
Residual	701.999749	143	4.90908916	R-squared	=	0.1309
				Adj R-squared	=	0.1248
Total	807.721876	144	5.60917969	Root MSE	=	2.2156

lne2	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lnq	-.4479545	.0965276	-4.64	0.000	-.6387597	-.2571492
_cons	-.7452062	.6591018	-1.13	0.260	-2.048048	.5576351

尽管变量 *lnq* 在 1%水平上显著，但R^2仅为 0.1309，且常数项不显著(p值为 0.26)。

下面去掉常数项，重新进行辅助回归。

```
. reg lne2 lnq,noc
```

Source	SS	df	MS	Number of obs	=	145
				F(1, 144)	=	419.95
Model	2065.53636	1	2065.53636	Prob > F	=	0.0000
Residual	708.275258	144	4.91857818	R-squared	=	0.7447
				Adj R-squared	=	0.7429
Total	2773.81162	145	19.1297353	Root MSE	=	2.2178

lne2	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lnq	-.5527533	.0269733	-20.49	0.000	-.6060681	-.4994384

R^2上升为 0.7447(尽管无常数项的R^2与有常数项的R^2不具有可比性)，残差平方的变动与 *lnq* 高度相关。

计算以上辅助回归的拟合值，并记为 lne2f：

```
. predict lne2f
(option xb assumed; fitted values)
```

去掉对数后，即得到方差的估计值，并记为 e2f：

```
. gen e2f=exp(lne2f)
```

最后，使用方差估计值的倒数作为权重，进行 WLS 回归：

```
. reg lntc lnq lnpl lnpk lnpf [aw=1/e2f]
```

Source	SS	df	MS	Number of obs	=	145
				F(4, 140)	=	895.03
Model	173.069988	4	43.2674971	Prob > F	=	0.0000
Residual	6.76790874	140	.048342205	R-squared	=	0.9624
				Adj R-squared	=	0.9613
Total	179.837897	144	1.24887428	Root MSE	=	.21987

lntc	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lnq	.8759035	.0153841	56.94	0.000	.8454883	.9063187
lnpl	.5603879	.1734141	3.23	0.002	.2175389	.9032369
lnpk	-.0929807	.1960402	-0.47	0.636	-.4805627	.2946014
lnpf	.4672438	.0616476	7.58	0.000	.3453632	.5891243
_cons	-5.522088	.9928472	-5.56	0.000	-7.485	-3.559176

WLS 回归的结果显示，*lnpk* 的系数估计值由“-0.22”(OLS 估计值)改进为“-0.09”(其理论值应为正数)。

使用 OLS 时，变量 *lnpl* 的 *p* 值为 0.13，在 10%的水平上也不显著；而使用 WLS 后，该变量的 *p* 值变为 0.002，在 1%的水平上显著不为 0。

由于 Nerlove(1963)数据存在明显的异方差，使用 WLS 后提高了估计效率。

如果担心对条件方差函数的设定不准确，导致加权变换后的新扰动项仍有一定的异方差，可使用稳健标准误进行 WLS 估计：

```
. reg lntc lnq lnpl lnpk lnpf [aw=1/e2f],r
```

```
Linear regression                               Number of obs     =        145
                                                F(4, 140)         =     534.50
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9624
                                                Root MSE          =     .21987

------------------------------------------------------------------------------
             |               Robust
        lntc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnq |   .8759035   .020787    42.14   0.000     .8348064    .9170006
        lnpl |   .5603879   .2090099    2.68   0.008      .147164    .9736118
        lnpk |  -.0929807   .3016444   -0.31   0.758    -.6893478    .5033864
        lnpf |   .4672438   .0439915   10.62   0.000     .3802702    .5542173
       _cons |  -5.522088   1.671596   -3.30   0.001    -8.826924   -2.217252
------------------------------------------------------------------------------
```

无论是否使用稳健标准误，WLS 的回归系数都相同。

在此例中，多数解释变量(*lnq*, *lnpl*, *lnpk*)的稳健标准误大于普通标准误；但变量 *lnpf* 的稳健标准误反而小于普通标准误。

7.6 Stata 命令的批处理

在进行计量分析时，有时需要使用一系列命令对数据集进行处理。

可把所有命令放入一个 Stata “do 文件” (即以 “do” 为扩展名的程序文件)，进行批处理。

在 Stata 中，点击菜单 “窗口” (Window)→ “do 文件编辑器” (Do-file Editor)→ “新 do 文件编辑器” (New Do-file Editor)，或直接点击 New Do-file Editor 快捷键(参见图 7.4)

即可打开一个 “do 文件编辑器” (Do-file Editor)，在其中写入需要执行的命令。

图 7.4　Stata 的 New Do-file Editor 快捷键

以上文的加权最小二乘法为例。

假设数据文件 `nerlove.dta` 在当前路径，则可在“do 文件编辑器”中输入如下命令：

```
* WLS for Nerlove(1963)
capture log close
log using wls_nerlove.smcl,replace
set more off
use nerlove.dta, clear
reg lntc lnq lnpl lnpk lnpf
predict e1,res
gen e2=e1^2
gen lne2=log(e2)
reg lne2 lnq,noc
predict lne2f
gen e2f=exp(lne2f)
* Weighted least square regression
reg lntc lnq lnpl lnpk lnpf [aw=1/e2f]
reg lntc lnq lnpl lnpk lnpf [aw=1/e2f],r
log close
```

```
exit
```

“*”表示不执行其后的命令，常用来作为注释。

“capture log close”表示如有已打开的日志文件，先将其关闭(如有打开的日志文件，无法定义新的日志文件)。

命令“log using wls_nerlove.smcl,replace”表示在当前路径创建名为“wls_nerlove.smcl”的日志文件(选择项 replace 表示可覆盖此文件的原有内容)，并将 Stata 运行结果记录于此日志文件。

命令“set more off”使得 Stata 输出结果可自动连贯显示，而无须点击“more”翻屏。

输入以上命令后，点击 Do-file Editor 的“执行(do)”(即 Execute (do))快捷键即可运行此程序，参见图 7.5。

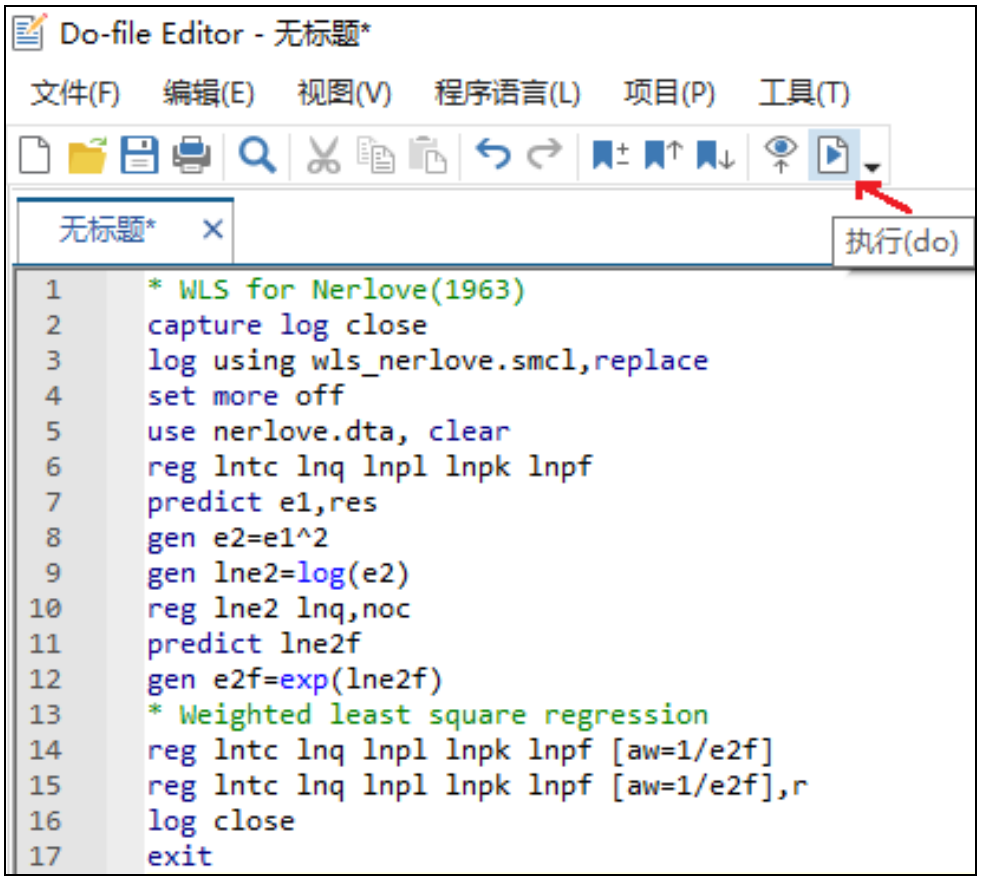

图 7.5　Do-file Editor 的 Execute (do)快捷键

如果要存储此程序文件，可点击 Do-file Editor 窗口的菜单“文件”(File)→“保存”(Save)或“另存为”(Save As)；比如，将此程序文件存为“wls_nerlove.do”，参见图 7.6。

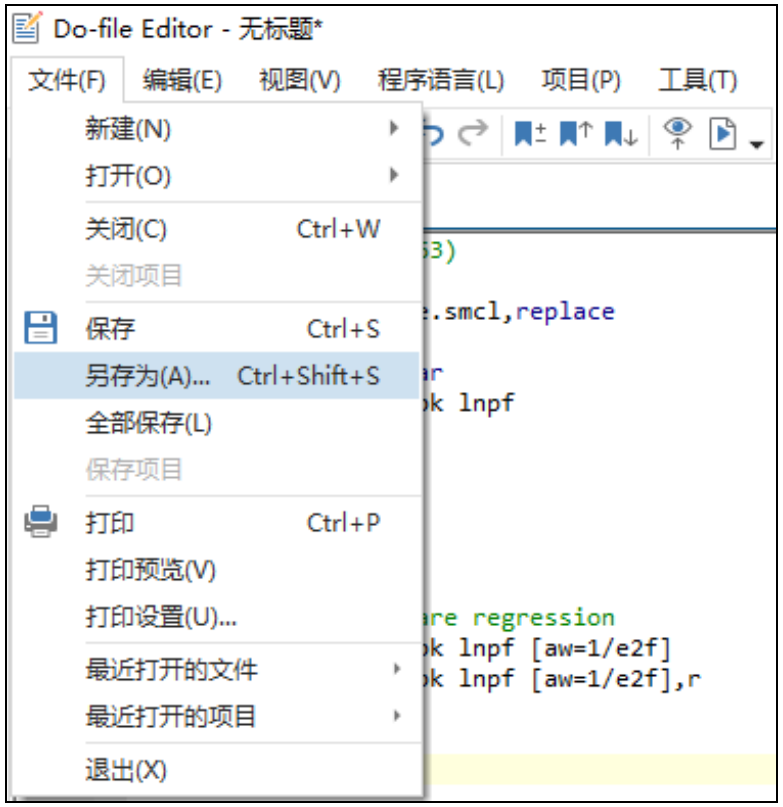

图 7.6　存储 do 文件

存储 Stata 的 do 文件后，可在 Stata 中点击菜单“文件”(File)→“”(Do)，寻找“wls_nerlove.do”文件，然后执行此文件，参见图 7.7。

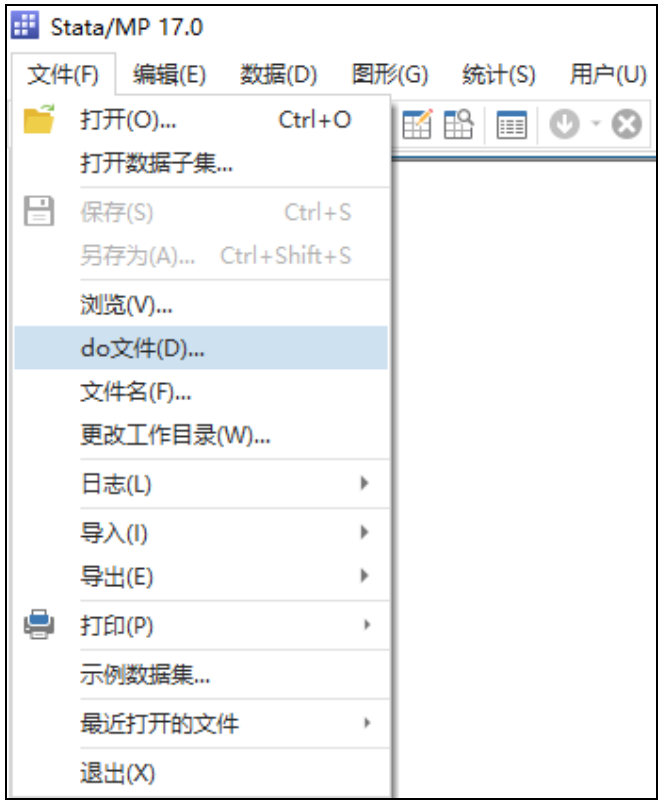

图 7.7　执行 do 文件

如果要编辑此文件，可以用鼠标右键点击“wls_nerlove.do”的图标，然后选择用 Stata 或“记事本”(Notepad)打开，编辑后直接存盘即可，参见图 7.8。

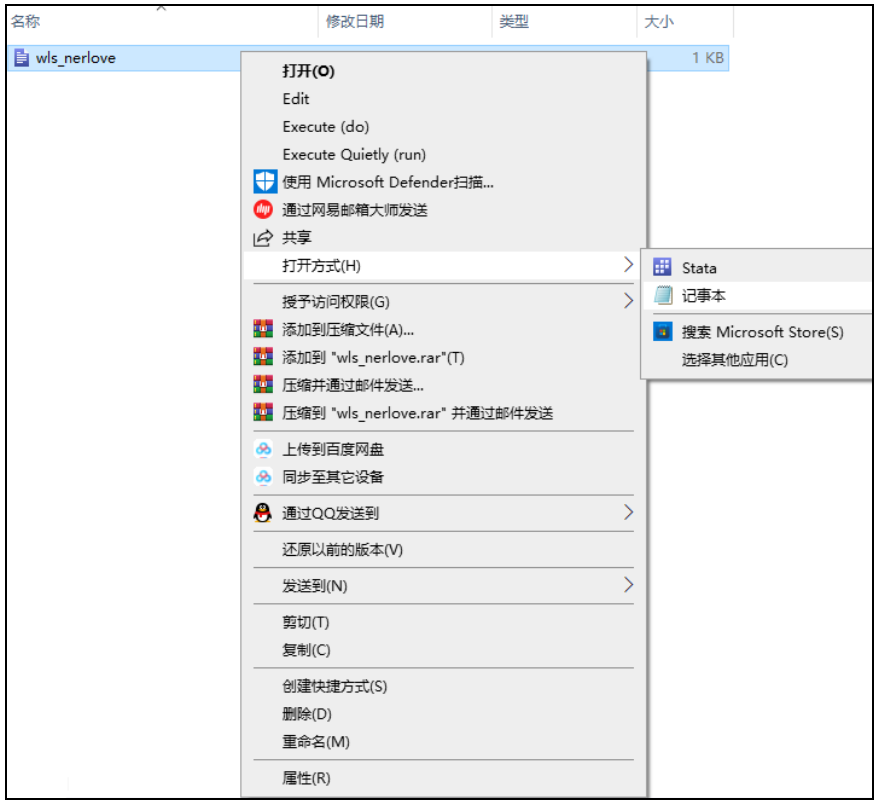

图 7.8　编辑 do 文件