



### 第三节 序列相关、异方差和 多重共线的处理技术

## ■ 回归模型的基本假定

1. 零期望值假定  $E(\varepsilon_i) = 0$

$$\Rightarrow E(\varepsilon) = E \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} E\varepsilon_1 \\ E\varepsilon_2 \\ \vdots \\ E\varepsilon_n \end{pmatrix} = \mathbf{0}$$

2. 同方差假定  $D(\varepsilon_i) = \sigma^2$  

3. 无自相关假定  $Cov(\varepsilon_i, \varepsilon_j) = E\varepsilon_i\varepsilon_j = 0$  

4. 随机误差项与解释变量 不相关:  $Cov(x_{ij}, \varepsilon_j) = 0$

5.  $\varepsilon_i \sim N(0, \sigma^2)$


6. 解释变量之间不存在多重共线性


各解释变量的观测值之间线性无关

$$\text{rank}(X) = P + 1 \quad \text{rank}(X'X) = P + 1$$



■ 序列相关 

■ 异方差 

■ 多重共线 

# 一、序列相关

- 定义:  $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i - E\varepsilon_i)(\varepsilon_j - E\varepsilon_j) = E\varepsilon_i\varepsilon_j \neq 0$

- 产生原因

- 惯性: 大多数经济时间序列数据都具有惯性或延续性

国民生产总值、就业率、货币供给、价格指数、消费和投资呈周期波动

- 偏误: 模型函数形式的设定误差

- 蛛网现象  $Q_t = b_0 + b_1P_{t-1} + e_t$

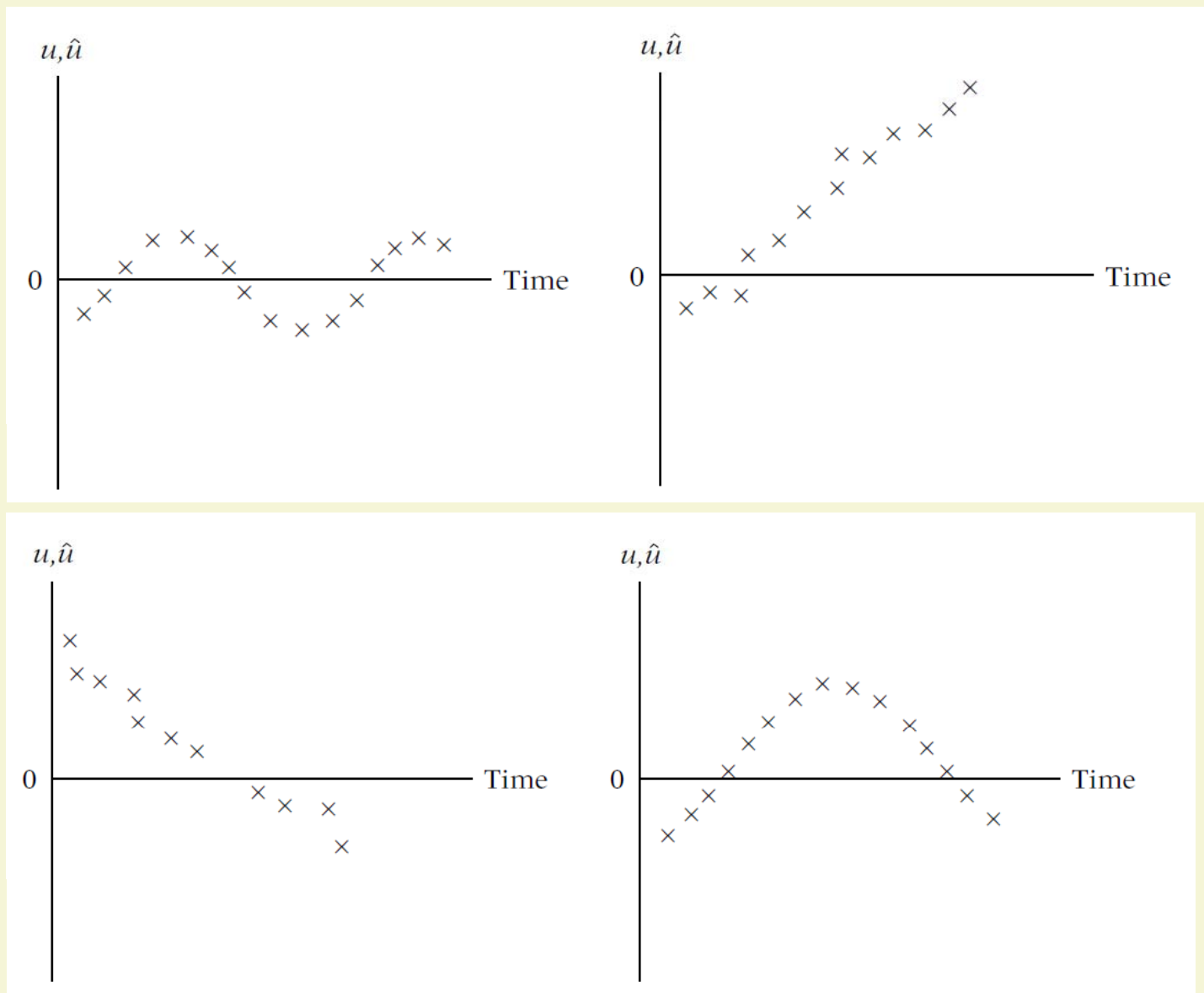
- 其他原因

真正的模型： $y_t = b_0 + b_1x_{1t} + b_2x_{2t} + b_3x_{3t} + e_t$

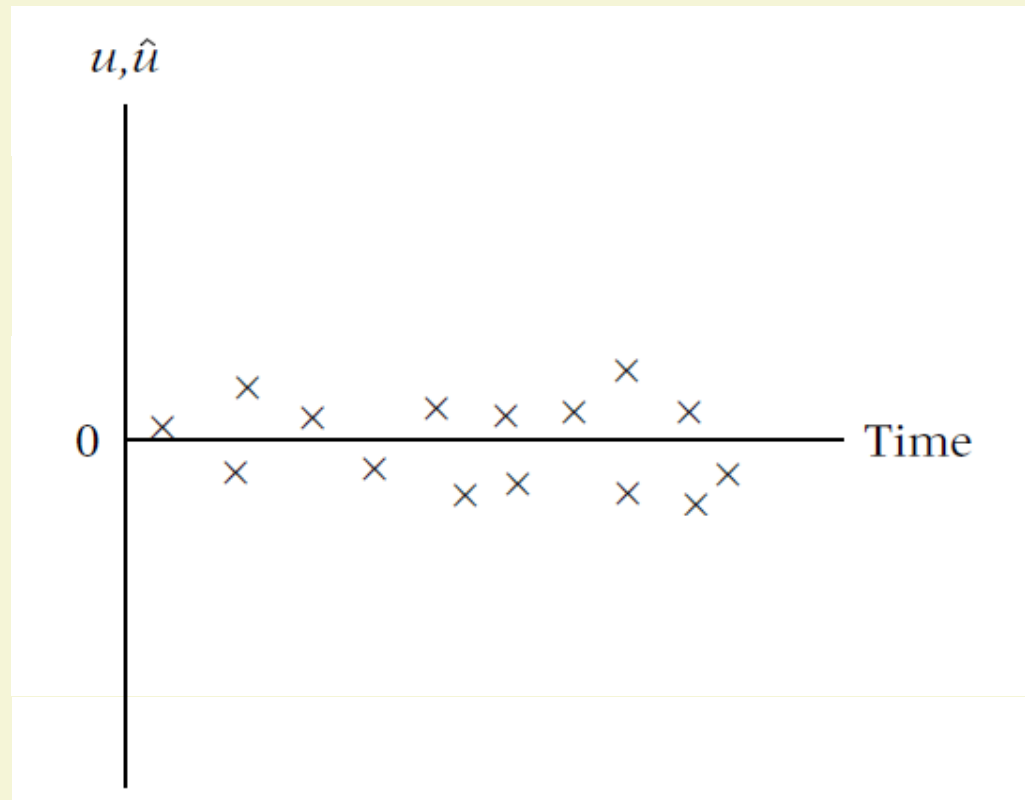
建立的模型： $y_t = b_0 + b_1x_{1t} + b_2x_{2t} + V_t$

那么： $V_t = b_3x_{3t} + e_t$

$$\begin{aligned} EV_tV_{t-1} &= E(b_3x_{3t} + e_t)(b_3x_{3t-1} + e_{t-1}) \\ &= b_3^2Ex_{3t}x_{3t-1} + b_3Ee_tx_{3t-1} + b_3Ex_{3t}e_{t-1} + Ee_te_{t-1} \\ &= b_3^2Ex_{3t}x_{3t-1} \end{aligned}$$



存在序列相关



不存在序列相关



■ 出现形式:

1. 一阶自相关  $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$

其中:  $E(v_t) = 0$ ;  $E(v_{t1}, v_{t2}) = 0$ ;  $E\varepsilon_t v_{t+1} = 0$

2. 高阶自相关  $\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \cdots + \rho_p\varepsilon_{t-p} + v_t$

## ■ 后果

- 参数估计值仍是无偏的，但不再具有最小方差，不再是有效估计量。
- 随机误差项的方差一般会低估

如果随机误差项存在一阶自相关性，可以证明

$$E(\sum e_t^2) < \sigma^2(n-2), \text{ 而 } \hat{\sigma}^2 = \frac{\sum e_t^2}{n-2}$$

- 模型的统计检验功能减小，F、t检验不可靠
- 区间估计和预测区间的精度降低

例如：  $t$  检验

$$\text{剩余标准差} \quad S_y = \sqrt{\frac{Q}{n-p-1}} = \sqrt{\frac{Y'Y - \hat{B}'X'Y}{n-p-1}} = \hat{\sigma}$$
$$t_i = \frac{\hat{b}_i}{S_y \cdot \sqrt{C_{ii}}}$$

其中：  $C_{ii}$  为  $(X'X)^{-1}$  的第  $i+1$  个对角元。

对于给定的显著性水平  $\alpha$ ，查  $t$  分布表 得  $t_\alpha(n-p-1)$ ，

若  $|t_i| \geq t_\alpha(n-p-1)$ ，则  $x_i$  的作用显著，必须保留  $x_i$  在回归方程中。

## ■ 检验方法

**1.D-W检验：**检验一阶自相关性

此方法假定：（1）解释变量x为非随机的；

（2） $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$ ， $v_t$  为满足古典假定的误差项

（3）线性回归模型中不应含有滞后内生变量作为解释变量，  
即不应出现下列形式  $y_t = b_0 + b_1x_t + b_2y_{t-1} + \dots$

（4）截距项不为0

（5）统计数据比较完整，无缺失项

记  $e_t = y_t - \hat{y}_t$

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

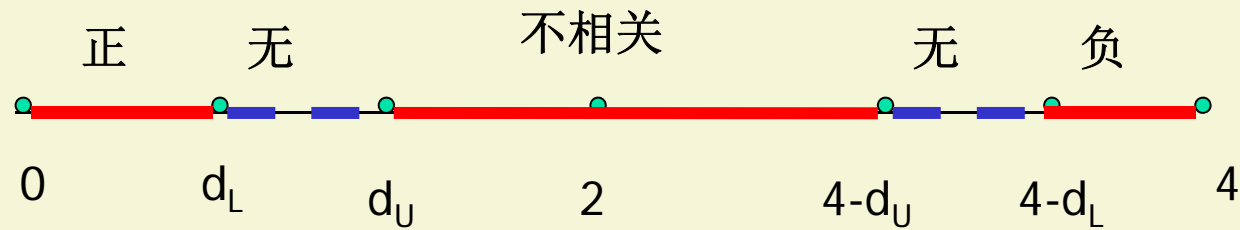
若令  $\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}$ , 则  $d \approx 2(1 - \hat{\rho})$

$\hat{\rho} = 1, \quad d = 0 \quad e_t$  完全正相关

$\hat{\rho} = 0, \quad d = 2 \quad e_t$  不存在一阶自相关

$\hat{\rho} = -1, \quad d = 4 \quad e_t$  完全负相关  $0 \leq d \leq 4$

对于给定的显著性水平, 查D-W表, 得 $d$ 的上下限 $d_U$ 、 $d_L$



- 对于无结论区，一般可以通过调整样本容量，减小无结论区

2.拉格朗日乘数检验：检验高阶自相关性  
设自相关形式为

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \cdots + \rho_p \varepsilon_{t-p} + v_t$$

(1) 利用OLS估计模型，得到残差序列 $e_t$ 。

(2) 利用  $e_t$ 、 $e_{t-1}$ 、 $\cdots$ 、 $e_{t-p}$ 进行回归

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \cdots + \rho_p e_{t-p} + v_t$$

得到可决系数 $R^2$ 。

(3) 对于给定的显著性水平 $\alpha$ ，计算 $nR^2$ 。

若 $nR^2 > \chi^2_\alpha(p)$ ，则存在相关性。

## ■ 消除方法

### 1. 差分法

$$y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + \cdots + b_p x_{pt} + \varepsilon_t$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t$$

$$\begin{aligned} y_t - \rho y_{t-1} &= b_0 + b_1 x_{1t} + b_2 x_{2t} + \cdots + b_p x_{pt} + \rho \varepsilon_{t-1} + v_t \\ &\quad - \rho(b_0 + b_1 x_{1,t-1} + b_2 x_{2,t-1} + \cdots + b_p x_{p,t-1} + \varepsilon_{t-1}) \end{aligned}$$

$$\text{令 } y_t' = y_t - \rho y_{t-1} \quad x_{it}' = x_{it} - \rho x_{i,t-1}$$

$$\text{则有 } y_t' = b_0(1 - \rho) + b_1 x_{1t}' + b_2 x_{2t}' + \cdots + b_p x_{pt}' + v_t$$

$$\text{Cov}(v_i, v_j) = 0 \quad i \neq j$$

$$\rho \text{ 的取法: } \hat{\rho} = 1 - \frac{d}{2}$$



## ■ 例：书101页第2题

2. 某地区的年消费  $C$  与可支配收入有如下表的历史数据：

单位：万元

年 份	$C$	$Y$	年 份	$C$	$Y$
1990	11 378	11 617	1996	20 074	21 512
1991	13 012	13 297	1997	21 439	23 124
1992	15 263	15 790	1998	22 833	24 724
1993	16 873	18 017	1999	24 205	26 175
1994	17 764	19 214	2000	25 307	27 219
1995	18 857	20 198			

应用普通最小二乘法，建立了以下线性模型

$$\hat{C} = 8526 + 0.65Y \quad r^2 = 0.953$$

试回答下述问题

- (1) 求残差平方和并检验自相关。
- (2) 若存在一阶自相关，试估计出自相关系数  $\rho$  的值。
- (3) 若存在一阶自相关，应如何消除呢？

$$\hat{C} = 8526 + 0.65Y \quad r^2 = 0.953$$

$$\rho = 1 - \frac{d}{2} = 0.8762 \quad C = 521.3821 + 0.9085Y$$

例、书101页第2题。

样本序号	年份	C	Y	$\hat{C}$	$e_i = C - \hat{C}$	C'	Y'	$\hat{C}'$	$e_i = C - \hat{C}$
1	1990	11378	11617	16077	-4699			11075	302
2	1991	13012	13297	17169	-4157	3043	3118	12602	410
3	1992	15263	15790	18789	-3526	3862	4139	14867	396
4	1993	16873	18017	20237	-3364	3500	4182	16890	-16
5	1994	17764	19214	21015	-3251	2980	3428	17977	-213
6	1995	18857	20198	21655	-2798	3292	3363	18871	-14
7	1996	20074	21512	22509	-2435	3752	3815	20065	9
8	1997	21439	23124	23557	-2118	3675	4275	21530	-90
9	1998	22833	24724	24597	-1764	4048	4463	22983	-150
10	1999	24205	26175	25540	-1335	4199	4512	24301	-96
11	2000	25307	27219	26218	-911	4099	4284	25250	57
12	2001	27020	28915	27321	-301	4846	5066	26791	229

解：(1)  $Q = 9.7737 \times 10^7$   $\sum_{t=2}^{12} e_t e_{t-1} = 8.5637 \times 10^7$   $d = 2(1 - \frac{8.5637 \times 10^7}{9.7737 \times 10^7}) = 0.2476$

$\alpha = 0.01$   $n = 15$ 时  $d_L = 0.81$   $d_U = 1.07$   $d < d_L$   $\therefore$  存在正自相关

表 VI—1

1%杜宾—瓦特森检验统计量的下界和上界

解 释 变 量 个 数							
n	1		2		3		4
	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$
15	.81	1.07	.70	1.25	.59	1.46	.49
16	.84	1.09	.74	1.25	.63	1.44	.53
17	.87	1.10	.77	1.25	.67	1.43	.57
18	.90	1.12	.80	1.26	.71	1.42	.61
19	.93	1.13	.83	1.26	.74	1.41	.65
20	.95	1.15	.86	1.27	.77	1.41	.68
21	.97	1.16	.89	1.27	.80	1.41	.72
22	1.00	1.17	.91	1.28	.83	1.40	.75

(2) 若存在一阶自相关, 自相关系数  $\rho = 1 - \frac{d}{2} = 0.8762$

(3) 消除方法: 用广义差分法, 构造  $y_t', c_t'$  如上表, 其中

$$y_t' = y_t - \rho y_{t-1} \quad c_t' = c_t - \rho c_{t-1}$$

用  $y_t'$ 、 $c_t'$  线性回归, 得到

$$b_0' = 64.5471 \quad b_1' = 0.9085$$

$$c_t' = 64.5471 + 0.9085 Y_t'$$

$$\text{已知 } b_0' = b_0(1 - \rho) \Rightarrow b_0 = 521.3821$$

$$b_1' = b_1 \Rightarrow b_1' = b_1 = 0.9085$$

所以原回归方程应为:

$$C = 521.3821 + 0.9085 Y$$

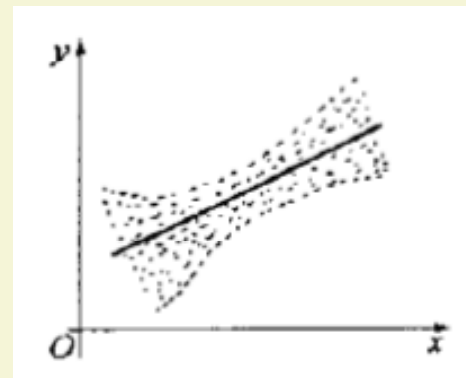
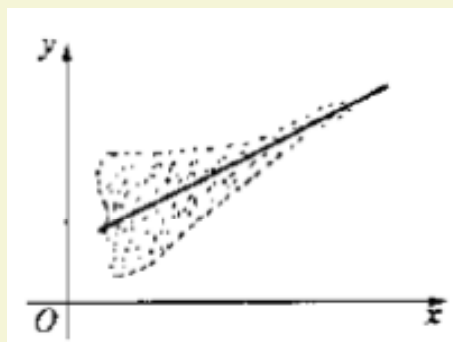
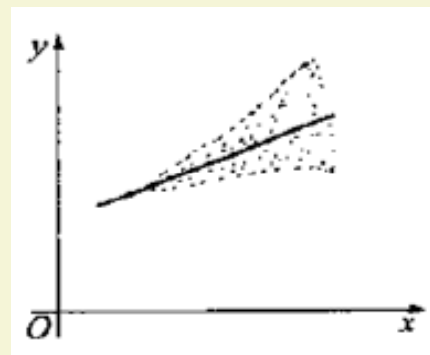
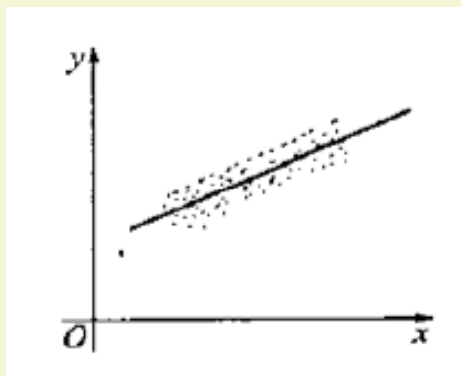
$$\text{相应的 } Q = 5.5898 \times 10^5$$



## 二、异方差性

- 定义:  $D(\varepsilon_i) \neq D(\varepsilon_j)$

异方差的直观表现



## ■ 产生原因

- 模型中遗漏了某些解释变量
- 模型函数形式的设定误差
- 样本数据的测量误差
- 随机因素的影响

## ■ 后果

- 回归系数的估计是无偏的,但不再具有最小方差,不再有效估计。
- 建立在 $t$ 分布和 $F$ 分布上的置信区间和假设检验不可靠。用 $t$ 、 $F$ 检验可能得到错误结论。

## ■ 考虑一元线性回归

$$\hat{b} = \frac{\sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i} = \sum k_i y_i \quad \text{其中: } k_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

在同方差假设下:

$$D(\hat{b}) = D(\sum k_i y_i) = \sum k_i^2 D(y_i) = \sigma^2 \sum k_i^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

其中 $\sigma^2$ 的无偏估计为  $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$ 。

若存在异方差, 令  $\varepsilon_t \sim N(0, \lambda_t \sigma^2)$ , 则

$$\begin{aligned} D(\hat{b}^*) &= D(\sum k_i y_i) = \sum k_i^2 D(y_i) = \sigma^2 \sum k_i^2 \lambda_i = \sigma^2 \sum k_i^2 \frac{\sum k_i^2 \lambda_i}{\sum k_i^2} \\ &= D(\hat{b}) \frac{\sum k_i^2 \lambda_i}{\sum k_i^2} \end{aligned}$$

## ■ 检验方法

### 1. White 检验

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \varepsilon_i$$

(1) 对上式进行 *OLS* 回归，求残差  $e_i$ 。

(2) 做辅助回归式

$$e_i^2 = a_0 + a_1 x_{1i} + a_2 x_{2i} + a_3 x_{1i}^2 + a_4 x_{2i}^2 + a_5 x_{1i} x_{2i} + v_i$$

求辅助回归式的可决系数  $R^2$ 。

若  $nR^2 \leq \chi_\alpha^2(5)$ ，则  $\varepsilon_i$  同方差；若  $nR^2 > \chi_\alpha^2(5)$ ，则  $\varepsilon_i$  异方差。

2. 图示法；

3. 戈里瑟检验；



(2) 直观判断法。首先不考虑它是否存在异方差性,就使用普通最小二乘法建立起回归模型,然后计算出残差  $e_i = y_i - \hat{y}_i$ , 绘制  $(\hat{y}_i, e_i^2)$  的平面点聚图。例如,图 7.8 表示不存在异方差性; 图 7.9 表示随  $\hat{y}$  值的增大其方差有递增趋势; 图 7.10 表示  $\hat{y}_i$  与  $e_i^2$  有线性关系; 图 7.11 与图 7.12 表示  $e_i^2$  与  $\hat{y}_i$  有二次曲线关系。

图 7.8 至图 7.12 是考察  $(\hat{y}_i, e_i^2)$  的点聚图。除此之外,还可以作  $(x_i, e_i^2)$  或  $(x_{1i}, e_i^2), (x_{2i}, e_i^2), \dots, (x_{ki}, e_i^2)$  的点聚图, 通过点聚图进行直观判断。

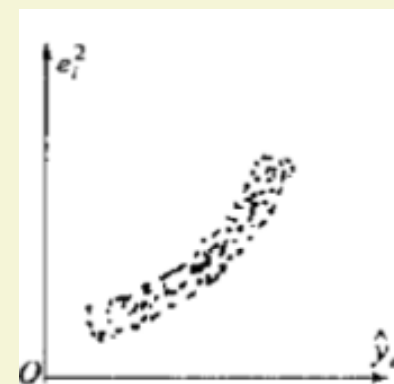


图7.12

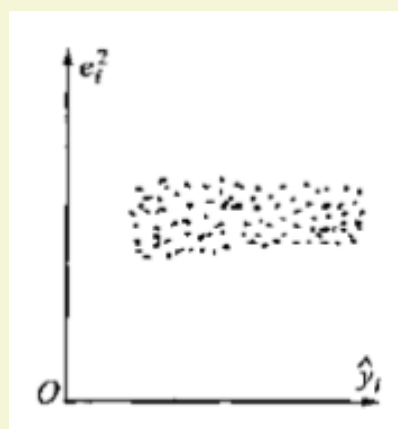


图7.8

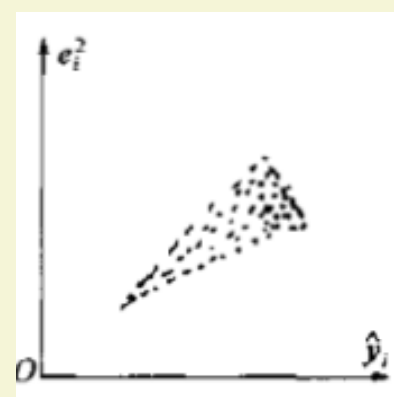


图7.9

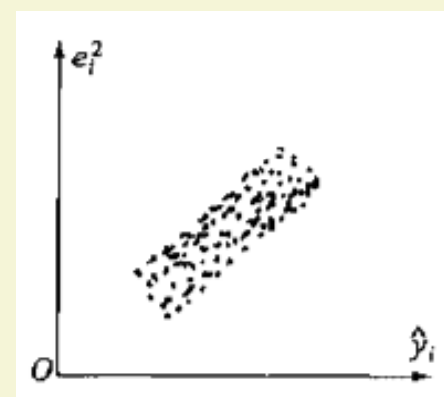


图7.10

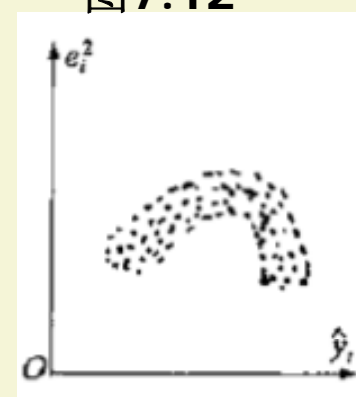


图7.11

(4) 戈里瑟 (Glejser) 检验。此法的应用步骤如下:

①将因变量  $y$  对自变量依普通最小二乘法求出回归方程, 并求出残差

$$e_i = y_i - \hat{y}_i$$

②将  $|e_i|$  对自变量  $x_i$  进行回归, 回归的模式可以参考点聚图决定。例如

$$|e_i| = \alpha_0 + \alpha_1 x_i^2$$

$$|e_i| = \alpha_0 + \alpha_1 \frac{1}{x_i}$$

$$|e_i| = \alpha_0 + \alpha_1 \sqrt{x_i} \quad \text{要求 } x_i > 0$$

等等。

③根据可决系数的大小, 估计量的标准差, 选择最优拟合的回归模式。

④对  $\alpha_0$ ,  $\alpha_1$  进行显著性检验, 若显著异于 0, 则接受异方差性, 如  $\alpha_0$  与 0 无显著差异, 但  $\alpha_1$  与 0 有显著差异, 则认为存在纯异方差性, 若  $\alpha_0$ ,  $\alpha_1$  均与 0 无显著差异, 则认为不存在异方差性。

## ■ 消除方法

### 1. 模型变换法

原回归模型  $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_p x_{pi} + \varepsilon_i$

若  $D(\varepsilon_i) = \sigma^2 f(x_{1i}, x_{2i}, \cdots, x_{pi})$ , 将原回归模型两端除以  $\sqrt{f(x_{1i}, x_{2i}, \cdots, x_{pi})}$ ,

$$\text{得 } \frac{y_i}{\sqrt{f(\dots)}} = \frac{b_0}{\sqrt{f(\dots)}} + b_1 \frac{x_{1i}}{\sqrt{f(\dots)}} + \cdots + b_p \frac{x_{pi}}{\sqrt{f(\dots)}} + \frac{\varepsilon_i}{\sqrt{f(\dots)}}$$

$$\text{此时 } D\left(\frac{\varepsilon_i}{\sqrt{f(\dots)}}\right) = \sigma^2$$

- 例：假设回归模型为  $y_i = a + bx_i + \varepsilon_i$ ，其中  $\text{Var}(\varepsilon_i) = \sigma^2 x_i^2$ ，则使用模型变换法估计模型时，应将模型变换为

$$\frac{y_i}{x_i} = \frac{a}{x_i} + b + \frac{\varepsilon_i}{x_i}$$

## 2. 加权最小二乘法

## 3. 模型的对数变换

若  $y_i = a + bx_i + \varepsilon_i$

令  $y_i' = \ln y_i, x_i' = \ln x_i$ , 建立  $y_i'$  和  $x_i'$  的模型

$$y_i' = a' + b'x_i' + \varepsilon_i'$$

- 例：已知1998年我国30个地区城镇居民平均每人全年家庭可支配收入 $x$ 与交通和通讯支出 $y$ 的数据如下，试预测随着收入的增加，人们对交通和通讯的需求。

城市	山西	宁夏	吉林	...	北京	上海	浙江
$x$	4098.73	4112.41	4206.64	...	8471.98	8773.1	7836.76
$y$	137.11	231.51	172.65	...	369.54	384.49	388.79

- 解: (1) 利用OLS建立x和y的一元线性回归方程

$$\hat{y}_t = -56.91798 + 0.058075x_t$$

$$t = (1.572049) \quad (8.962009)$$

$$R^2 = 0.741501 \quad S.E. = 50.48324$$

(2) 异方差检验: 戈里瑟检验

$x$	4098.73	4112.41	4206.64	...	8471.98	8773.1	7836.76
$y$	137.11	231.51	172.65	...	369.54	384.49	388.79
$\hat{y}$	181.12	181.91	187.38	...	435.09	452.58	398.20
$ \hat{e} $	44.01	49.6	14.73	...	65.55	68.09	9.41
$x^2$	...	...	...	...	...	...	...
$\sqrt{x}$	64.02	64.13	64.86	...	92.04	93.66	88.53
$\frac{1}{x}$	0.00024	...	...	...	...	...	...

分别建立回归模型

$$\left| \hat{e}_t \right| = -11.34475 + 1.38 \times 10^{-6} x_t^2$$

$$t = (-1.00849) \quad (4.763249) \quad R^2 = 0.447606$$

$$\left| \hat{e}_t \right| = -156.4946 + 2.579356 \sqrt{x_t}$$

$$t = (-3.569) \quad (4.324) \quad R^2 = 0.4004$$

$$\left| \hat{e}_t \right| = 133.7844 - 521981.0 \frac{1}{x_t}$$

$$t = (4.836) \quad (-3.773) \quad R^2 = 0.3371$$

$$\because \alpha = 0.05 \quad t_\alpha(28) = 2.05$$

$\therefore$  存在异方差

### (3) 异方差消除: 模型的对数变换

$x$	4098.73	4112.41	4206.64	...	8471.98	8773.1	7836.76
$y$	137.11	231.51	172.65	...	369.54	384.49	388.79
$\ln x$	8.3184	8.3218	8.3444	...	9.0445	9.079	8.9666
$\ln y$	4.9208	5.4446	5.1513	...	5.9123	5.9519	5.9630

得到模型:

$$\ln \hat{y}_t = -4.48084 + 1.164771 \ln x_t$$

$$t = (-4.5444) \quad (10.12156)$$

$$R^2 = 0.785352$$





### 三、多重共线性

#### ■ 定义:

对于变量  $x_1 x_2 \cdots x_p$ , 如果存在不全为零的数  $\lambda_1 \lambda_2 \cdots \lambda_p$ , 使得

$$\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_p x_p = 0 \quad \text{完全多重共线性}$$

$$\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_p x_p + u = 0 \quad u \text{ 为随机误差项} \quad \text{不完全多重共线性}$$

$$\text{记 } X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{pmatrix}, \text{ 则当存在多重共线性时, 有}$$

$$\text{Rank}(X) < p + 1$$

## ■ 产生原因:书109页

### (1) 经济变量之间的共同趋势

**时间序列样本：**经济繁荣时期，各基本经济变量（收入、消费、投资、价格）都趋于增长；衰退时期，又同时趋于下降。

**横截面数据：**生产函数中，资本投入与劳动力投入往往出现高度相关情况，大企业二者都大，小企业都小。

### (2) 滞后变量的引入

例如， $\text{消费} = f(\text{当期收入}, \text{前期收入})$   
显然，两期收入间有较强的线性相关性。

### (3) 在建模过程中解释变量选择不当

## ■ 后果

- 参数估计的精度降低，某些回归系数的标准偏差很大，不能正确反映自变量与因变量之间的关联程度。

$$S_{bi} = \sqrt{c_{ii}} \cdot S_y$$

- 回归系数的估计值对样本的敏感性增强。
- 回归系数可能出现与事理意义不符的符号。
- 可能将有用的变量排除掉。

## ■ 检验方法

1. 计算自变量之间的相关系数，对变量  $x_i$ 、 $x_j$

$$r_{ij} = \frac{\sum_{t=1}^n x_{it} x_{jt}}{\sqrt{\sum_{t=1}^n x_{it}^2} \sqrt{\sum_{t=1}^n x_{jt}^2}}$$

$r_{ij}=1$   $x_i$ 、 $x_j$  完全相关，完全多重共线；

$r_{ij}=0$   $x_i$ 、 $x_j$  完全不相关；

$r_{ij}^2 > R^2$  (可决系数) 时，共线性严重，应予以消除。

2. 计算不包含某个变量的复可决系数 $r_j^2$

若回归方程为  $y = f(x_1, x_2, \dots, x_m)$

分别构造不含某个变量  $x_j$  的  $m$  个回归方程

$$y_j = f_j(x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_m)$$

对每个方程估计可决系数 $r_1^2, r_2^2, \dots, r_m^2$ .

$$r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$r_j^2$ 越大，则  $x_j$  与其他解释变量发生共线性严重。

等价的检验方法：

模型中每一个解释变量分别以其余解释变量为解释变量进行回归，并计算相应的拟合优度。

如果某一种回归

$$X_{ji} = \alpha_1 X_{1i} + \alpha_2 X_{2i} + \dots + \alpha_L X_{Li}$$

的判定系数较大，说明 $X_j$ 与其他 $X$ 间存在共线性。

构造如下F统计量

$$F_j = \frac{R_{j\cdot}^2 / (k - 2)}{(1 - R_{j\cdot}^2) / (n - k + 1)} \sim F(k - 2, n - k + 1)$$

式中： $R_{j\cdot}^2$ 为第j个解释变量对其他解释变量的回归方程的决定系数，

若存在较强的共线性，则 $R_{j\cdot}^2$ 较大且接近于1，这时（ $1 - R_{j\cdot}^2$ ）较小，从而 $F_j$ 的值较大。

因此，给定显著性水平 $\alpha$ ，计算F值，并与相应的临界值比较，来判定是否存在相关性。

## ■ 消除方法

### 1. 去掉不必要的解释变量：

回归系数最小； $t$ 检验最小；系数符号与经济意义不符。

### 2. 合并变量：差分法

### 3. 增加样本量，增加数据量

### 4. 采用逐步回归法，减少多重共线性的影响。



例、某省工业产值 $x_1$ ，农业产值 $x_2$ ，固定资产投资 $x_3$ 和运输业产值  $y$  数据如下，试建立回归模型。

序号	年份	工业产值 $x_1$	农业产值 $x_2$	， 固定资产投资 $x_3$	运输业产值 $y$
1	1970	57.82	27.05	14.54	3.09
2	1971	58.05	28.89	16.83	3.40
3	1972	59.15	33.02	12.26	3.88
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
16	1985	126.50	94.01	43.86	10.03
17	1986	138.89	103.23	48.90	10.83
18	1987	160.56	119.33	60.98	12.90

解：建立 $y$  与 $x_1$ ,  $x_2$ ,  $x_3$ 的回归模型

$$\hat{y} = -1.003590 + 0.05528x_1 - 0.00398x_2 + 0.0907x_3$$

$$t_{b1} = 2.9574 \quad t_{b2} = -0.2865 \quad t_{b3} = 3.4918$$

$$R^2 = 0.9943 \quad S = 0.3354 \quad F = 405.58$$

计算相关系数矩阵：

	$x_1$	$x_2$	$x_3$	$y$
$x_1$	1.0	0.9972	0.9763	0.9892
$x_2$		1.0	0.9523	0.9648
$x_3$			1.0	0.9875
$y$				1.0

$$0.9972 =$$

$$\frac{57.82 \times 27.05 + 58.05 \times 28.89 + \cdots + 160.56 \times 119.33}{\sqrt{57.82^2 + 58.05^2 \cdots + 160.56^2} \cdot \sqrt{27.05^2 + 28.89^2 \cdots + 119.33^2}}$$

其中  $0.9972 \times 0.9972 = 0.9944 > R^2$ ，存在多重共线性。

剔除  $x_2$ ，重新建立回归方程

$$\hat{y} = -0.8989 + 0.0513x_1 + 0.0912x_3$$

$$t_{b1} = 4.201 \quad t_{b3} = 3.6323$$

$$R^2 = 0.9971 \quad S = 0.3250 \quad F = 647.99$$

可见剔除  $x_2$ ，提高了  $t$  检验值， $x$  与  $y$  的相关性增加了，减小了误差，预测精度提高了。

