第四章 违背基本假设的情况

4.1 异方差性产生的背景和原因
4.2 一元加权最小二乘估计
4.3 多元加权最小二乘估计
4.4 自相关性问题及其处理
4.5 异常值与强影响点
4.6 本章小结与评注

第四章 违背基本假设的情况

Gauss-Markov条件

$$\begin{cases} E(\varepsilon_i)=0, \quad i=1,2,\cdots,\ n \\ cov(\varepsilon_i,\varepsilon_j)=\begin{cases} \sigma^2\ , & i=j \\ 0\ , & i\neq\ j \end{cases} \end{cases} \qquad (i\ ,j=1,2,\cdots,n)$$

4.1 异方差性产生的背景和原因

一、异方差产生的原因

例4.1 居民收入与消费水平有着密切的关系。用x_i表示第i户的收入量，y_i表示第i户的消费额，一个简单的消费模型为：

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i \quad , \quad i=1,2,\ldots,n$$

低收入的家庭购买差异性比较小，

高收入的家庭购买行为差异就很大。

导致消费模型的随机项ε_i具有不同的方差。

4.1 异方差性产生的背景和原因

二、异方差性带来的问题

当存在异方差时，普通最小二乘估计存在以下问题:

(1)参数估计值虽是无偏的,但不是最小方差线性无偏估计;

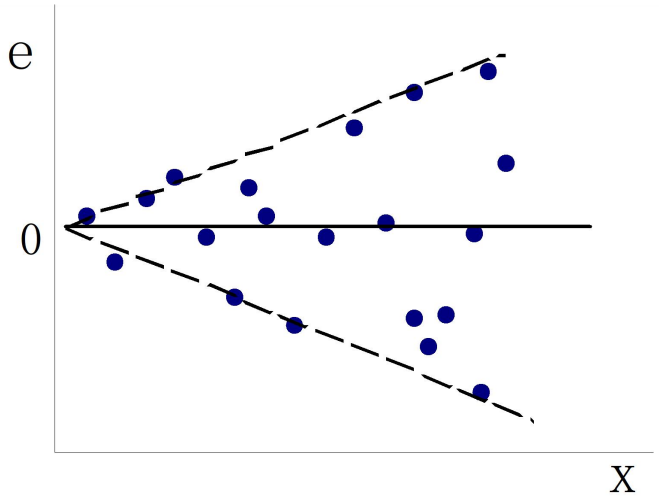
(2)参数的显著性检验失效;

(3)回归方程的应用效果极不理想。

4.2 一元加权最小二乘估计

一、异方差性的检验

（一）残差图分析法



(b)

图2.5（b)
存在异方差

4.2 一元加权最小二乘估计

一、异方差性的检验

（二）等级相关系数法

等级相关系数检验法又称斯皮尔曼(Spearman)检验，是一种应用较广泛的方法。这种检验方法既可用于大样本，也可用于小样本。进行等级相关系数检验通常有三个步骤。

第一步，作y关于x的普通最小二乘回归，求出ε_i的估计值，即e_i的值。

4.2 一元加权最小二乘估计

（二）等级相关系数法

第二步,取e_i的绝对值,分别把x_i和$|e_i|$按递增（或递减）的次序分成等级,按下式计算出等级相关系数:

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2$$

其中,n为样本容量,d_i为对应于x_i和$|e_i|$的等级的差数。

4.2 一元加权最小二乘估计

（二）等级相关系数法

第三步,做等级相关系数的显著性检验。在$n>8$的情况下,用下式对样本等级相关系数r_s进行t检验。检验统计量为:

$$t=\frac{\sqrt{n-2}\ r_s}{\sqrt{1-r_s^2}}$$

如果$t\le t_{\alpha/2}(n-2)$可认为异方差性问题不存在，

如果$t>t_{\alpha/2}(n-2)$,说明x_i与$|e_i|$之间存在系统关系,异方差性问题存在。

4.2 一元加权最小二乘估计

例4.3 设某地区的居民收入与储蓄额的历史统计数据如表4.1。

(1)用普通最小二乘法建立储蓄y与居民收入x的回归方程,并画出残差散点图;

(2)诊断该问题是否存在异方差;

序号	储蓄 y（万元）	居民收入 x（万元）
1	264	8777
2	105	9210
3	90	9954
…	…	…
31	2300	38200

4.2 一元加权最小二乘估计

序号	储蓄y	居民收入x	x_i等级	残差e_i	$\|e_i\|$	$\|e_i\|$等级	d_i	d_i^2
1	264	8777	1	169.0	169.0	16	-15	225
2	105	9210	2	-26.6	26.6	3	-1	1
3	90	9954	3	-104.6	104.6	7	-4	16
4	131	10508	4	-110.5	110.5	8	-4	16
5	122	10979	5	-159.4	159.4	15	-10	100
6	107	11912	6	-253.4	253.4	23	-17	289
7	406	12747	7	-25.1	25.1	2	5	25
8	503	13499	8	8.2	8.2	1	7	49
9	431	14269	9	-129.0	129.0	9	0	0
10	588	15522	10	-78.0	78.0	4	6	36
11	898	16730	11	129.7	129.7	10	1	1
12	950	17663	12	102.7	102.7	6	6	36
13	779	18575	13	-145.5	145.5	14	-1	1
14	819	19635	14	-195.3	195.3	19	-5	25
15	1222	21163	15	78.4	78.4	5	10	100
…	…	…	…	…	…	…	…	…
31	2300	38200	31	-286.1	286.1	24	7	49

4.2 一元加权最小二乘估计

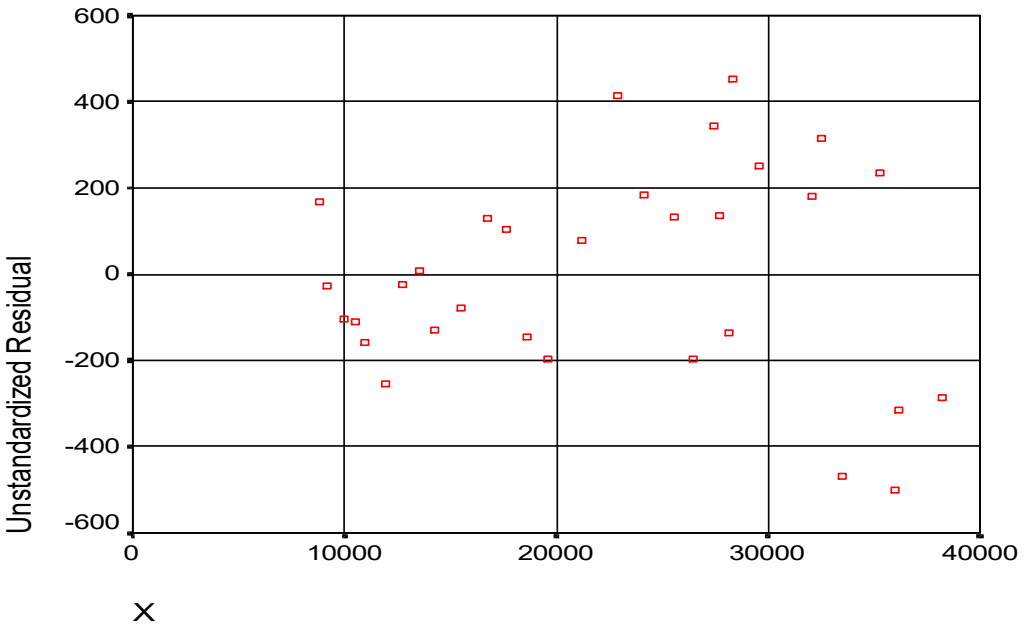


图4.1 残差图

4.2 一元加权最小二乘估计

用SPSS计算等级相关系数。

Correlations

			ABSE	X
Spearman's rho	ABSE	Correlation Coefficient	1.000	.686
		Sig. (2-tailed)	.	.000
		N	31	31
	X	Correlation Coefficient	.686	1.000
		Sig. (2-tailed)	.000	.
		N	31	31

4.2 一元加权最小二乘估计

（2）计算等级相关系数。

$$r_s = 1 - \frac{6}{31(31^2 - 1)} \times 1558 = 0.6859$$

$$t = \frac{\sqrt{31-2}\ \times 0.6859}{\sqrt{1-0.6859^2}} = 5.076$$

4.2 一元加权最小二乘估计

Spearman等级相关系数可以反映非线性相关的情况，Pearson简单相关系数不能反映非线性相关的情况。例如x与y的取值如下，

序号	1	2	3	4	5	6	7	8	9	10
x	1	2	3	4	5	6	7	8	9	10
y	1	4	9	16	25	36	49	64	81	100

$y_i = x_i^2$ 具有完全的曲线相关。

容易计算出y与x的简单相关系数r=0.9746，
而y与x的等级相关系数r_s=1

4.2 一元加权最小二乘估计

二、一元加权最小二乘估计

消除异方差性的方法通常有：

- 加权最小二乘法,
- Box-Cox变换法,
- 方差稳定性变换法

加权最小二乘法(Weighted Least Square,简记为WLS)是一种最常用的消除异方差性的方法。

4.2 一元加权最小二乘估计

二、一元加权最小二乘估计

一元线性回归普通最小二乘法的残差平方和为：

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

一元线性回归的加权最小二乘的离差平方和为：

$$Q_w(\beta_0,\beta_1)=\sum_{i=1}^{n}w_i(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}w_i(y_i-\beta_0-\beta_1x_i)^2$$

4.2 一元加权最小二乘估计

加权最小二乘估计为：

$$\begin{cases} \hat{\beta}_{0w} = \bar{y}_w - \hat{\beta}_{1w}\bar{x}_w \\ \hat{\beta}_{1w} = \dfrac{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2} \end{cases}$$

其中，$\bar{x}_w = \dfrac{1}{\Sigma w_i}\sum w_i x_i$ 是自变量的加权平均；

$\bar{y}_w = \dfrac{1}{\Sigma w_i}\sum w_i y_i$ 是因变量的加权平均。

4.2 一元加权最小二乘估计

观测值的权数应该是观测值误差项方差的倒数,即

$$w_i = \frac{1}{\sigma_i^2}$$

在实际问题中,误差项的方差是未知的,常与自变量x的幂函数x^m成比例，其中m是待定的未知参数。此时权函数为

$$w_i = \frac{1}{x_i^m}$$

4.2 一元加权最小二乘估计

三、寻找最优权函数

利用SPSS软件可以确定幂指数m的最优取值。

依次点选Analyze-Regression-Weight Estimation进入估计权函数对话框，默认的幂指数m的取值为m=-2.0，-1.5，-1.0，-0.5，0，0.5，1.0，1.5，2.0。

先将因变量y与自变量x选入各自的变量框，再把x选入Weight变量框，幂指数（Power）取默认值，计算结果如下（格式略有变动）：

4.2 一元加权最小二乘估计

Log-likelihood Function = -224.258830	POWER value = -2.000
Log-likelihood Function = -221.515008	POWER value = -1.500
Log-likelihood Function = -218.832193	POWER value = -1.000
Log-likelihood Function = -216.252339	POWER value = -.500
Log-likelihood Function = -213.856272	POWER value = .000
Log-likelihood Function = -211.773375	POWER value = .500
Log-likelihood Function = -210.185972	POWER value = 1.000
Log-likelihood Function = -209.316127	POWER value = 1.500
Log-likelihood Function = -209.379714	POWER value = 2.000

The Value of POWER Maximizing Log-likelihood Function = 1.500

Log-likelihood Function = -209.316127

4.2 一元加权最小二乘估计

Multiple R	.96744
R Square	.93595
Adjusted R Square	.93374
Standard Error	.12532

Analysis of Variance:

	DF	Sum of Squares	Mean Square	F	Sig
Regression	1	6.6548981	6.6548981	423.741	0.000
Residuals	29	.4554477	.0157051		

Variables in the Equation

Variable	B	SE B	Beta	T	Sig T
X	.08793	.004272	.967443	20.585	.0000
(Constant)	-719.12	78.316		-9.182	.0000

4.2 一元加权最小二乘估计

幂指数m的最优取值为m=1.5。

加权最小二乘的r^2=0.9360，F值=423.741；

普通最小二乘的r^2=0.912，F值=300.732。

说明加权最小二乘估计的效果好于普通最小二乘的效果。

4.2 一元加权最小二乘估计

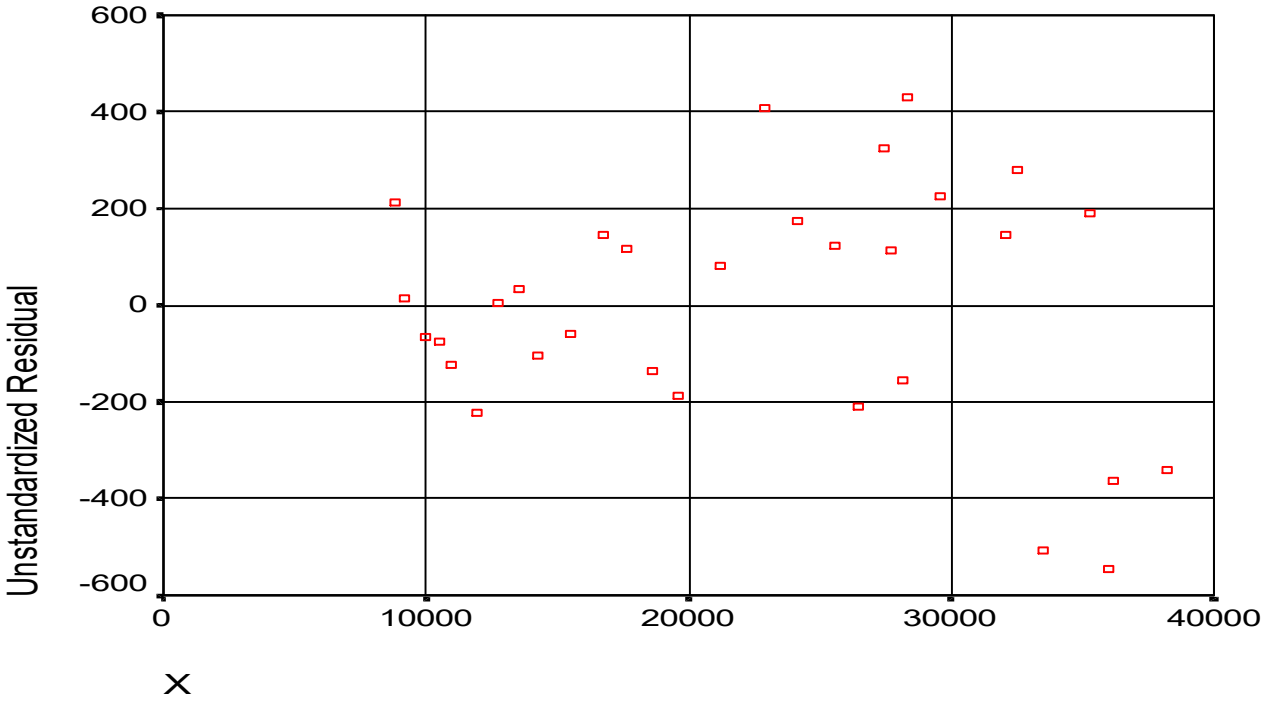


图4.2 加权最小二乘残差图残差图

4.2 一元加权最小二乘估计

	序号	y_i	x_i	w_i	e_i	e_{iw}
小方差组	1	264	8777	1.2161E-06	169	211
	2	105	9210	1.1314E-06	-27	14
	3	90	9954	1.0069E-06	-105	-66
	4	131	10508	9.2837E-07	-111	-74
	5	122	10979	8.6927E-07	-159	-124
	6	107	11912	7.6917E-07	-253	-221
	7	406	12747	6.9485E-07	-25	4
	8	503	13499	6.3760E-07	8	35
	9	431	14269	5.8669E-07	-129	-105
	10	588	15522	5.1710E-07	-78	-58

4.2 一元加权最小二乘估计

	序号	y_i	x_i	w_i	e_i	e_{iw}
中等方差组	11	898	16730	4.6212E-07	130	146
	12	950	17663	4.2599E-07	103	116
	13	779	18575	3.9501E-07	-146	-135
	14	819	19635	3.6346E-07	-195	-188
	15	1222	21163	3.2481E-07	78	80
	16	1702	22880	2.8895E-07	413	409
	17	1578	24127	2.6684E-07	183	176
	18	1654	25604	2.4408E-07	134	122
	19	1400	26500	2.3181E-07	-195	-211
	20	1829	27670	2.1726E-07	134	115
	21	2200	28300	2.1005E-07	452	431

4.2 一元加权最小二乘估计

	序号	y_i	x_i	w_i	e_i	e_{iw}
大方差组	22	2017	27430	2.2012E-07	343	324
	23	2105	29560	1.9676E-07	250	225
	24	1600	28150	2.1173E-07	-135	-156
	25	2250	32100	1.7388E-07	180	147
	26	2420	32500	1.7068E-07	317	281
	27	2570	35250	1.5110E-07	234	190
	28	1720	33500	1.6309E-07	-468	-507
	29	1900	36000	1.4640E-07	-500	-546
	30	2100	36200	1.4519E-07	-317	-364
	31	2300	38200	1.3394E-07	-286	-340

4.3 多元加权最小二乘

当误差项ε_i存在异方差时，加权离差平方和为

$$Q_w = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

记

$$\mathbf{W} = \begin{pmatrix} w_1 & & & \\ & w_2 & & \\ & & \cdots & \\ & & & w_n \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{W}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

加权最小二乘估计
WLS的矩阵表达

4.3 多元加权最小二乘估计

通常取权函数W为某个自变量x_j（j=1, 2, ..., p）的幂函数，即，$W = x_j^m$

在$x_1, x_2, ..., x_p$这p个自变量中取哪一个？

这只需计算每个自变量x_j与普通残差的等级相关系数，选取等级相关系数最大的自变量构造权函数。

4.3 多元加权最小二乘估计

例4.4 续例3.2，研究北京市各经济开发区经济发展与招商投资的关系。

因变量y为各开发区的销售收入（百万元），

x_1为截至1998年底各开发区累计招商数目，

x_2为招商企业注册资本（百万元）。

计算出普通残差的绝对值abse=$|e_i|$与x_1、x_2的等级相关系数，$r_{e1}=0.443$，$r_{e2}=0.721$，因而选取x_2构造权函数。

4.3 多元加权最小二乘估计

Correlations

			ABSE	X1	X2
Spearman's rho	ABSE	Correlation Coefficient	1.000	.443	.721
		Sig. (2-tailed)	.	.098	.002
		N	15	15	15
	X1	Correlation Coefficient	.443	1.000	.432
		Sig. (2-tailed)	.098	.	.108
		N	15	15	15
	X2	Correlation Coefficient	.721	.432	1.000
		Sig. (2-tailed)	.002	.108	.
		N	15	15	15

4.3 多元加权最小二乘估计

仿照例4.3，用Weight Estimate估计幂指数m，得m的最优值为m=2。

由于m=2是在默认范围[-2，2]的边界，因而应该扩大范围重新计算。取m从1到5，步长仍为0.5，得m的最优值为m=2.5

4.3 多元加权最小二乘估计

Multiple R	.92163
R Square	.84941
Adjusted R Square	.82431
Standard Error	.03238

	DF	Sum of Squares	Mean Square	F	Sig
Regression	2	.07096521	.03548261	33.84	0.000
Residuals	12	.01258145	.00104845		

Variable	B	SE B	Beta	T	Sig T
X1	1.696439	.404370	.587146	4.195	.0012
X2	.470312	.149306	.440853	3.150	.0084
(Constant)	-266.9621	106.742		-2.501	.0279

4.3 多元加权最小二乘估计

加权最小二乘的R^2=0.84941，F值=33.84；

普通最小二乘的R^2=0.842，F值=31.96。

加权最小二乘估计的拟合效果略好于普通最小二乘。

加权最小二乘的回归方程为：

$$\hat{y}=-266.96+1.696x_1+0.4703x_2$$

普通最小二乘的回归方程为：

$$\hat{y}=-327.039+2.036x_1+0.468x_2$$

4.3 多元加权最小二乘估计

方差稳定变换

(1)如果σ_i^2与$E(y_i)$存在一定的比例关系,使用$y' = \sqrt{y}$;

(2)如果σ_i与$E(y_i)$存在一定的比例关系,使用$y' = \log(y)$;

(3)如果$\sqrt{\sigma_i}$与$E(y_i)$存在一定的比例关系,使用$y' = \frac{1}{y}$

4.3 多元加权最小二乘估计

Box-Cox变换

$$Y^{(\lambda)}=\begin{cases}\dfrac{Y^{\lambda}-1}{\lambda}, & \lambda\neq 0\\ \ln Y, & \lambda=0\end{cases}$$

§4.4 自相关性问题及其处理

如果一个回归模型的随机误差项

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) \neq 0$$

则称随机误差项之间存在着自相关现象。

这里的自相关现象不是指两个或两个以上的变量之间的相关,而指的是一个变量前后期数值之间存在的相关关系。

§4.4 自相关性问题及其处理

一、自相关性产生的背景和原因

1.遗漏关键变量时会产生序列的自相关性。

2.经济变量的滞后性会给序列带来自相关性。

3.采用错误的回归函数形式也可能引起自相关性。

4.蛛网现象(Cobweb phenomenon)可能带来序列的自相关性。

5.因对数据加工整理而导致误差项之间产生自相关性。

§4.4 自相关性问题及其处理

二、自相关性带来的问题

1.参数的估计值不再具有最小方差线性无偏性。

2.均方误差MSE可能严重低估误差项的方差。

3.容易导致对t值评价过高,常用的F检验和t检验失效。如果忽视这一点,可能导致得出回归参数统计检验为显著,但实际上并不显著的严重错误结论。

4.当存在序列相关时,仍然是β的无偏估计量,但在任一特定的样本中,可能严重歪曲β的真实情况,即最小二乘估计量对抽样波动变得非常敏感。

5.如果不加处理地运用普通最小二乘法估计模型参数,用此模型进行预测和结构分析将会带来较大的方差甚至错误的解释。

§4.4 自相关性问题及其处理

三、自相关性的诊断

（一）图示检验法

1. 绘制（e_t，e_{t-1}）的散点图。

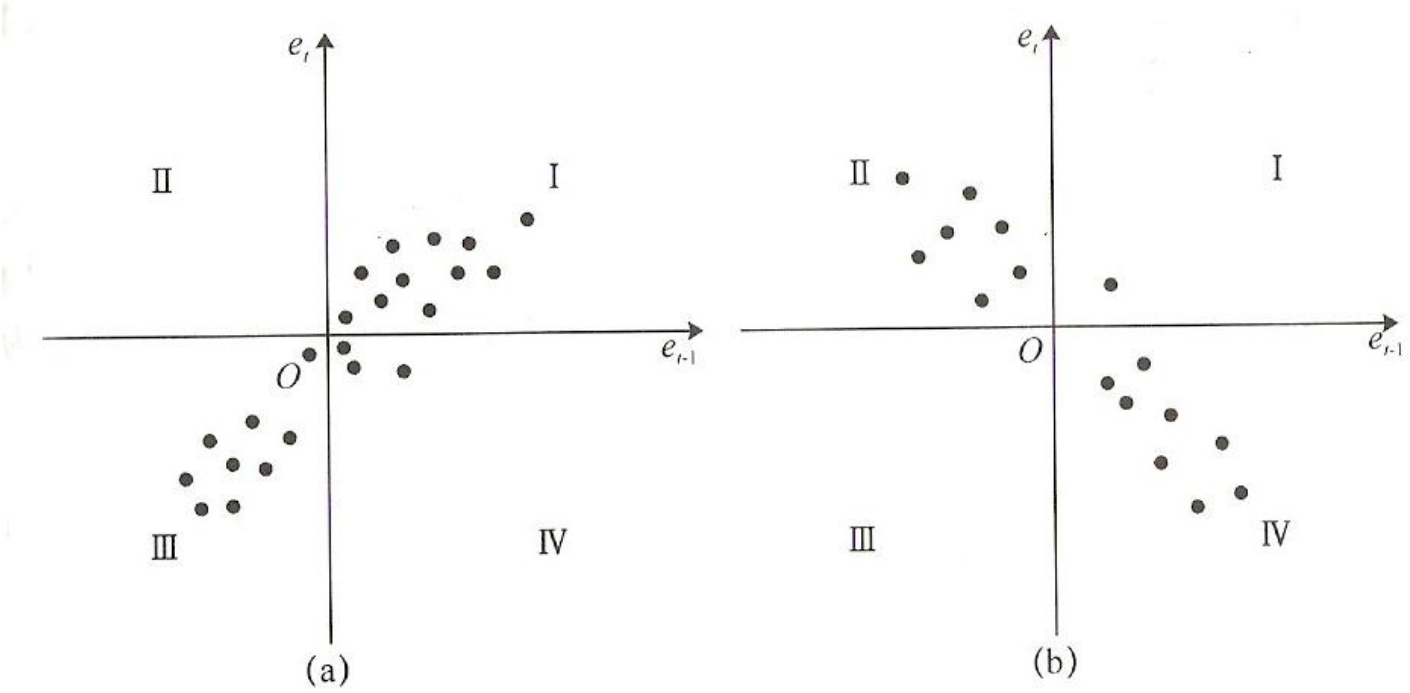


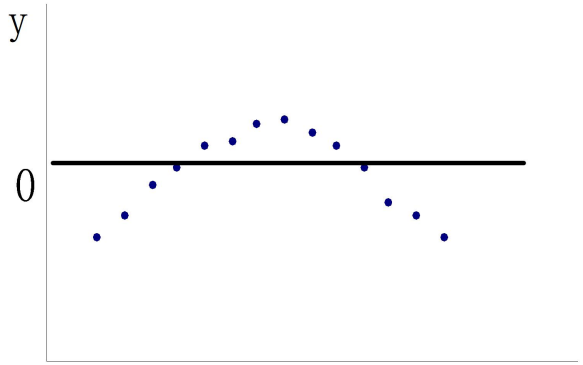
图 4.3

§4.4 自相关性问题及其处理

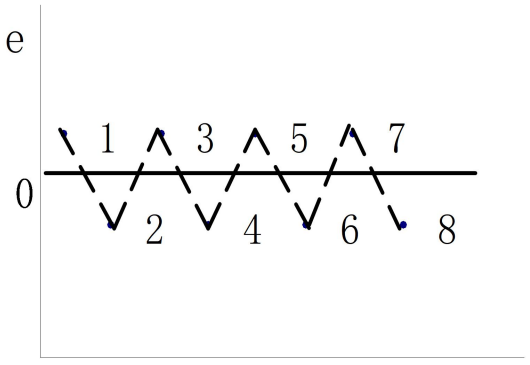
三、自相关性的诊断

（一）图示检验法

2.按照时间顺序绘制回归残差项e_t的图形。



(c)



(d)

§4.4 自相关性问题及其处理

三、自相关性的诊断

（二）自相关系数法

误差序列$\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n$的自相关系数定义为

$$\rho=\frac{\sum_{t=2}^{n}\varepsilon_t\varepsilon_{t-1}}{\sqrt{\sum_{t=2}^{n}\varepsilon_t^2}\sqrt{\sum_{t=2}^{n}\varepsilon_{t-1}^2}}$$

自相关系数的估计值为

$$\hat{\rho}=\frac{\sum_{t=2}^{n}e_te_{t-1}}{\sqrt{\sum_{t=2}^{n}e_t^2}\sqrt{\sum_{t=2}^{n}e_{t-1}^2}}$$

§4.4 自相关性问题及其处理

三、自相关性的诊断

（三）D.W检验

D.W检验是J.Durbin和G.S.Watson于1951年提出的一种适用于小样本的一种检验方法。

D.W检验验只能用于检验随机扰动项具有一阶自回归形式的序列相关问题。

这种检验方法是建立计量经济学模型中最常用的方法,一般的计算机软件都可自动产生出D.W值。

§4.4 自相关性问题及其处理

（三）D.W检验

随机扰动项的一阶自回归形式为:

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

其中u_t是不相关序列。

为了检验序列的相关性,构造的假设是

$$H_0: \rho = 0$$

§4.4 自相关性问题及其处理

（三）D.W检验

定义D.W统计量为:

$$D.W=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=2}^{n}e_i^2}$$

$$D.W=\frac{\sum_{t=2}^{n}e_t^2+\sum_{t=2}^{n}e_{t-1}^2-2\sum_{t=2}^{n}e_te_{t-1}}{\sum_{t=2}^{n}e_t^2}\approx 2\left(1-\frac{\sum_{t=2}^{n}e_te_{t-1}}{\sum_{t=2}^{n}e_i^2}\right)$$

§4.4 自相关性问题及其处理

（三）D.W检验

$$\hat{\rho}=\frac{\sum_{t=2}^{n}e_te_{t-1}}{\sqrt{\sum_{t=2}^{n}e_t^2}\sqrt{\sum_{t=2}^{n}e_{t-1}^2}}\approx\frac{\sum_{t=2}^{n}e_te_{t-1}}{\sum_{t=2}^{n}e_t^2}$$

得 $D.W\approx 2(1-\hat{\rho})$

D.W的取值范围为:0≤D.W≤4

§4.4 自相关性问题及其处理

（三）D.W检验

因而D.W值与 $\hat{\rho}$ 的对应关系为

$\hat{\rho}$	D.W	误差项的自相关性
-1	4	完全负自相关
（-1，0）	（2，4）	负自相关
0	2	无自相关
（0，1）	（0，2）	正自相关
1	0	完全正自相关

§4.4 自相关性问题及其处理

（三）D.W检验

根据样本容量n和解释变量的数目k(这里包括常数项),查D.W分布表,得临界值d_L和d_U,然后依下列准则考察计算得到的D_W值,以决定模型的自相关状态:

$0 \leqslant D.W \leqslant d_L$,	误差项 ε_1, ε_2, ..., ε_n间存在正相关;
$d_L < D.W \leqslant d_U$,	不能判定是否有自相关;
$d_U < D.W < 4-d_U$,	误差项 ε_1, ε_2, ..., ε_n间无自相关;
$4-d_U \leqslant D.W < 4-d_L$,	不能判定是否有自相关;
$4-d_L \leqslant D.W \leqslant 4$,	误差项 ε_1, ε_2, ..., ε_n间存在负相关。

§4.4 自相关性问题及其处理

（三）D.W检验

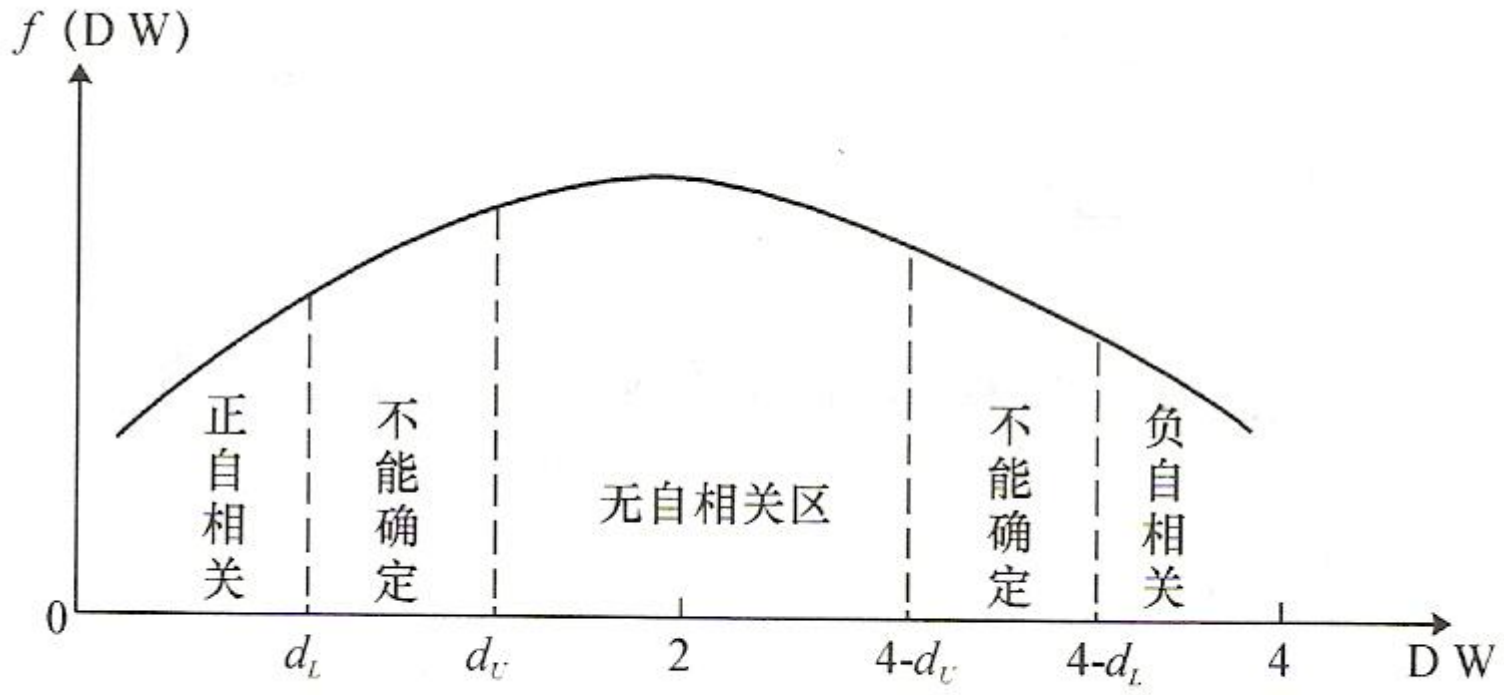


图 4.5

§4.4 自相关性问题及其处理

表 4　DW 检验上下界表

n 是观察值的数目；k 是解释变量的数目，包括常数项　　5%的上下界

n	$k=2$		$k=3$		$k=4$		$k=5$		$k=6$	
	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83

§4.4 自相关性问题及其处理

（三）D.W检验

D.W检验尽管有着广泛的应用,但也有明显的缺点和局限性。

1. D.W检验有一个不能确定的区域,一旦D.W值落在这个区域,就无法判断。这时,只有增大样本容量或选取其他方法。

2. D.W统计量的上、下界表要求n＞15,这是因为样本如果再小,利用残差就很难对自相关的存在性作出比较正确的诊断。

3. D.W检验不适应随机项具有高阶序列相关的检验。

§4.4 自相关性问题及其处理

四、自相关问题的处理方法

（一）迭代法

以一元线性回归模型为例, 设一元线性回归模型的误差项存在一阶自相关

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

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

$$\begin{cases} E(u_t) = 0, \ t = 1, 2, \cdots, \ n \\ \operatorname{cov}(u_t, u_s) = \begin{cases} \sigma^2, & t = s \\ 0, & t \neq s \end{cases} \end{cases} \qquad (t, s = 1, 2, \cdots, n)$$

§4.4 自相关性问题及其处理

（一）迭代法

根据回归模型 $y_t=\beta_0+\beta_1 x_t+\varepsilon_t$

有 $y_{t-1}=\beta_0+\beta_1 x_{t-1}+\varepsilon_{t-1}$

则有 $(y_t-\rho y_{t-1})=(\beta_0-\rho\beta_0)+\beta_1(x_t-\rho x_{t-1})+(\varepsilon_t-\rho\varepsilon_{t-1})$

令 $y'_t = y_t - \rho y_{t-1}$ $\quad \beta'_0 = \beta_0(1-\rho)$

$x'_t = x_t - \rho x_{t-1}$ $\quad \beta'_1 = \beta_1$

得 $y'_t = \beta'_0 + \beta'_1 x'_t + u_t$

§4.4 自相关性问题及其处理

四、自相关问题的处理方法

（一）迭代法

其中自相关系数ρ用公式 $\hat{\rho} \approx 1 - \frac{1}{2} D.W$ 估计。

用变换因变量与变换自变量作普通最小二乘回归。

如果误差项确实是一阶自相关，通过以上变换，回归模型已经消除自相关。

§4.4 自相关性问题及其处理

（一）迭代法

实际问题中，有时误差项并不是简单的一阶自相关，而是更复杂的自相关形式，（4.24）式的误差项u_t可能仍然存在自相关，这就需要进一步对（4.24）式的误差项u_t做D.W检验，以判断u_t是否存在自相关，如果检验表明误差项u_t不存在自相关，迭代法到此结束。如果检验表明误差项u_t存在自相关，那末对回归模型（4.24）式重复用迭代法，这个过程可能要重复几次，直至最终消除误差项自相关。这种迭代消除自相关的过程正是迭代法名称的由来。

§4.4 自相关性问题及其处理

（二）差分法

一阶差分法通常适用于原模型存在较高程度的一阶自相关的情况。

在迭代法（4.24）式中，当 $\rho=1$ 时，得

$$(y_t-y_{t-1})=\beta_1(x_t-x_{t-1})+(\varepsilon_t-\varepsilon_{t-1})$$

以 $\Delta y_t=y_t-y_{t-1}$，$\Delta x_t=x_t-x_{t-1}$ 代之，得

$$\Delta y_t=\beta_1\Delta x_t+u_t$$

是不带有常数项的回归方程

$$\hat{\beta}_1=\frac{\sum_{t=2}^{n}\Delta y_t\Delta x_t}{\sum_{t=2}^{n}\Delta x_t^2}$$

§4.4 自相关性问题及其处理

（二）差分法

一阶差分法的应用条件是自相关系数ρ=1，在实际应用中，ρ接近1时我们就采用差分法而不用迭代法，这有两个原因。

第一，迭代法需要用样本估计自相关系数ρ，对ρ的估计误差会影响迭代法的使用效率；

第二，差分法比迭代法简单,人们在建立时序数据的回归模型时,更习惯于用差分法。

§4.4 自相关性问题及其处理

（三）科克伦—奥克特（Cochrane-Orcutt）迭代

方法（一）中的迭代法近似取 $\hat{\rho} \approx 1 - \frac{1}{2}\mathrm{DW}$ 可以使用其他迭代法给出的更精确的估计，最常用的是科克伦—奥克特迭代法。

以一元线性回归为例，方法（一）的迭代是1步迭代，根据1步迭代计算出的 ρ 和回归系数，由（4.18）式的回归方程重新计算残差，得到新的残差序列后就可以计算出新的DW值，新的 ρ 和回归系数，如果新的 $\hat{\rho}$ 与前一次迭代的相差很小，低于给定的界限，就停止迭代，否则继续下一步迭代。

§4.4 自相关性问题及其处理

（三）科克伦—奥克特（Cochrane-Orcutt）迭代

有一点需要说明的是，迭代的起始步骤认为是从第0步开始的，就是用（4.18）式做普通最小二乘回归，相当于认为$\rho=0$。这样方法（一）中的迭代实际上包括第0步和第1步共两步迭代过程，也称为科克伦—奥克特两步法。通常情况下，科克伦—奥克特多步迭代与两步迭代相差不大。

§4.4 自相关性问题及其处理

（四）普莱斯—温斯登（Prais-Winsten）迭代法

采用迭代法用（4.23）式计算迭代值时不能计算第1期的迭代值，因此样本量从n减少到n-1。对大样本量时这无足轻重，但是当样本量较小时每一个样本值都是宝贵的。为此可以使用普莱斯—温斯登变换，

对t=1，令，$y_1' = \sqrt{1-\rho^2}\, y_1$　　$x_1' = \sqrt{1-\rho^2}\, x_1$

经过普莱斯—温斯登变换的迭代法就称为普莱斯—温斯登迭代法。

§4.4 自相关性问题及其处理

五、 自相关实例分析

【例4.5】 续例2.2

年份	人均国民收入（元）	人均消费金额（元）	年份	人均国民收入（元）	人均消费金额（元）
1980	460	234.75	1990	1634	797.08
1981	489	259.26	1991	1879	890.66
1982	525	280.58	1992	2287	1063.39
1983	580	305.97	1993	2939	1323.22
1984	692	347.15	1994	3923	1736.32
1985	853	433.53	1995	4854	2224.59
1986	956	481.36	1996	5576	2627.06
1987	1104	545.40	1997	6053	2819.36
1988	1355	687.51	1998	6392	2958.18
1989	1512	756.27			

§4.4 自相关性问题及其处理

年份	序号	x_t	y_t	e_t	x'_t	y'_t	e'_t
1980	1	460	234.75	-12.11			
1981	2	489	259.26	-.81	229.56	126.86	5.92
1982	3	525	280.58	4.13	249.20	134.36	4.46
1983	4	580	305.97	4.47	283.90	147.72	2.00
1984	5	692	347.15	-5.33	364.88	174.59	-8.08
1985	6	853	433.53	7.75	462.71	237.74	10.45
1986	7	956	481.36	8.69	474.91	236.85	4.00
1987	8	1104	545.40	5.35	564.82	273.91	.04
1988	9	1355	687.51	33.18	732.34	379.90	29.62
1989	10	1512	756.27	30.47	747.78	368.52	11.19
1990	11	1634	797.08	15.73	781.23	370.54	-2.05
1991	12	1879	890.66	-2.22	957.42	441.11	-11.85
1992	13	2287	1063.39	-15.24	1227.24	561.05	-14.98
1993	14	2939	1323.22	-52.24	1649.13	723.47	-45.02
1994	15	3923	1736.32	-87.12	2265.40	990.02	-59.58
1995	16	4854	2224.59	-22.70	2641.43	1245.31	24.18
1996	17	5576	2627.06	51.07	2838.34	1372.39	61.43
1997	18	6053	2819.36	26.21	2908.14	1337.70	-5.09
1998	19	6392	2958.18	10.70	2978.11	1368.07	-6.64

§4.4 自相关性问题及其处理

4. 方法比较

自回归方法	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_0' = (1-\hat{\rho})\hat{\beta}_0$	$\hat{\beta}_1 = \hat{\beta}_1'$	DW	$\hat{\sigma}_u$
迭代法	0.564	37.202	16.220	0.456	1.372	26.96
差分法	——	——	0	0.465	1.596	29.34
精确最大似然	0.544	33.532	15.291	0.457	——	27.055
科克伦—奥克特	0.563	37.214	16.263	0.456	1.381	27.840
普莱斯—温斯登	0.570	33.110	14.237	0.457	1.385	27.039

§4.4 自相关性问题及其处理

对回归模型　$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$

做变换　$y'_t = y_t - \rho y_{t-1}, \quad x'_t = x_t - \rho x_{t-1}$

得　$y'_t = \beta'_0 + \beta'_1 x'_t + u_t$

其中　$\beta'_0 = \beta_0(1-\rho), \quad \beta'_1 = \beta_1$

问题：为什么变换后的回归模型参数估计性质好

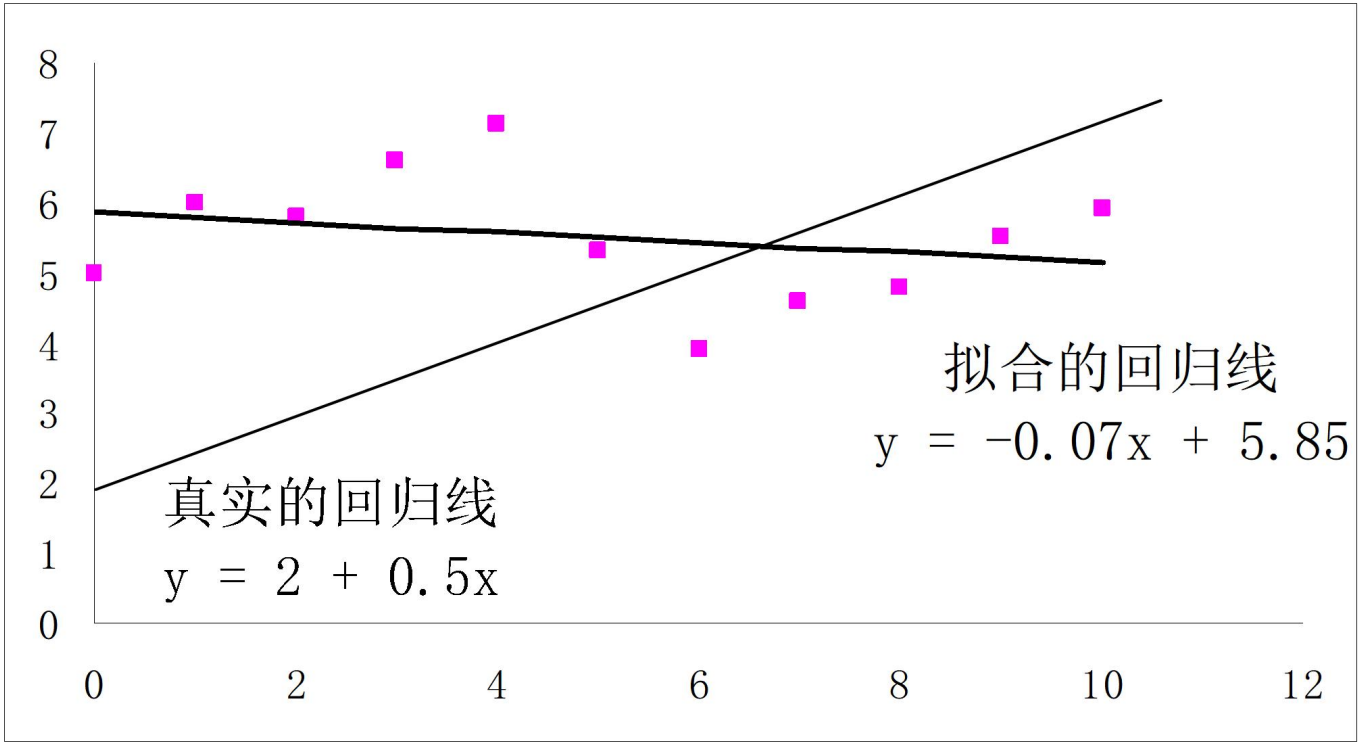
§4.4 自相关性问题及其处理

自相关的危害

t	u_t	$\varepsilon_t=\varepsilon_{t-1}+u_t$	$y=2+0.5t+\varepsilon_t$
0		3	5
1	0.5	3.5	6
2	-0.7	2.8	5.8
3	0.3	3.1	6.6
4	0	3.1	7.1
5	-2.3	0.8	5.3
6	-1.9	-1.1	3.9
7	0.2	-0.9	4.6
8	-0.3	-1.2	4.8
9	0.2	-1	5.5
10	-0.1	-1.1	5.9

§4.4 自相关性问题及其处理

自相关的危害



§4.4 自相关性问题及其处理

5 预测

以迭代法为例说明回归预测值 $\hat{y}_t$ 和残差 e'_t 的计算方法。

$$\hat{y}'_t = 16.22 + 0.456\, x'_t$$

将 $y'_t = y_t - 0.564 y_{t-1}$，$x'_t = x_t - 0.564 x_{t-1}$ 代入，还原为原始方程

$$\hat{y}_t = 37.20 + 0.564 y_{t-1} + 0.456 (x_t - 0.564 x_{t-1})$$

$$= 37.20 + 0.564 y_{t-1} + 0.456 x_t - 0.257 x_{t-1}$$

§4.4 自相关性问题及其处理

5 预测

其一般性的公式为

$$\hat{y}_t = \hat{\beta}_0' + \hat{\rho} y_{t-1} + \hat{\beta}_1'(x_t - \hat{\rho} x_{t-1})$$

注意：在自相关回归中，回归预测值 $\hat{y}_t$ 不是用 $\hat{\beta}_0 + \hat{\beta}_1 x_t$ 计算

SPSS软件提供的3种方法可以直接保存回归预测值 $\hat{y}_t$ 和残差 e_t'

§4.4 自相关性问题及其处理

另外一种计算 $\hat{y}_t$ 的想法是对 $\hat{\beta}_0+\hat{\beta}_1 x_t$ 做修正。在误差项没有自相关时，我们实际上就是直接用估计值 $\hat{\beta}_0+\hat{\beta}_1 x_t$ 作为回归预测值 $\hat{y}_t$。现在误差项存在自相关 $\varepsilon_t=\rho\varepsilon_{t-1}+u_t$，需要从残差 e_t 中提取出有用的信息对估计值 $\hat{\beta}_0+\hat{\beta}_1 x_t$ 做修正，其中 $e_t=y_t-(\hat{\beta}_0+\hat{\beta}_1 x_t)$ 是误差项 ε_t 的估计值。计算过程如下：

t=1 时，取 $\hat{y}_1=\hat{\beta}_0+\hat{\beta}_1 x_1$，$e_1=y_1-(\hat{\beta}_0+\hat{\beta}_1 x_1)$

$t\geqslant 2$ 时，取 $\hat{y}_t=\hat{\beta}_0+\hat{\beta}_1 x_t+\hat{\rho}e_{t-1}$，　$e_t=y_t-(\hat{\beta}_0+\hat{\beta}_1 x_t)$

注意：e_t 是 ε_t 的估计值，　$e_t'=e_t-\hat{\rho}e_{t-1}$ 是 u_t 的估计值

§4.4 自相关性问题及其处理

例如，取x_{20}=6600，则

$$\hat{y}_t = \hat{\beta}_0' + \hat{\rho} y_{t-1} + \hat{\beta}_1'(x_t - \hat{\rho} x_{t-1})$$

$$\hat{y}_{20} = 16.22 + 0.564 \times 2958.15 + 0.456(6600 - 0.564 \times 6392) = 3050.31$$

第二种方法

$t \geqslant 2$ 时，取 $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\rho} e_{t-1}$，　　$e_t = y_t - (\hat{\beta}_0 + \hat{\beta}_1 x_t)$

$$e_{19} = 2958.18 - (37.202 + 0.456 \times 6392) = 6.226$$

$$\hat{y}_{20} = 37.202 + 0.456 \times 6600 + 0.564 \times 6.226 = 3050.31$$

§4.5 异常值与强影响值

异常值分为两种情况：

一种是关于因变量y异常；

另一种是关于自变量x异常。

§4.5 异常值与强影响值

一、关于因变量y的异常值

在残差分析中，认为超过$\pm 3\hat{\sigma}$的残差为异常值。

标准化残差 $$ZRE_i = \frac{e_i}{\hat{\sigma}}$$

学生化残差 $$SRE_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

其中h_{ii}是帽子矩阵$\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$的主对角线元素。

§4.5 异常值与强影响值

当数据中存在关于 y 的异常观察值时，异常值把回归线拉向自己，使异常值本身的残差减少，而其余观察值的残差增大，这时回归标准差 $\hat{\sigma}$ 也会增大，因而用“3σ”准则不能正确分辨出异常值。解决这个问题的方法是改用删除残差。

§4.5 异常值与强影响值

删除残差的构造思想是：

在计算第 i 个观察值的残差时，用删除掉这第 i 个观察值的其余 n-1 个观察值拟合回归方程，计算出第 i 个观察值的删除拟合值 $\hat{y}_{(i)}$，这个删除拟合值与第 i 个值无关，不受第 i 个值是否为异常值的影响，第 i 个观察值的删除残差为：

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

可以证明：

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}$$

§4.5 异常值与强影响值

进一步可以给出第 i 个观察值的删除学生化残差，记为 $SRE_{(i)}$

$$SRE_{(i)} = SRE_i \left(\frac{n-p-1}{n-p-2} - \frac{SRE_i^2}{n-p-2} \right)^{-\frac{1}{2}}$$

用 SPSS 软件可以直接计算出删除学生化残差 $SRE_{(i)}$ 的数值，$|SRE_{(i)}|>3$ 的观测值即判定为异常值。

§4.5 异常值与强影响值

二、关于自变量x的异常值

在 $D(e_i)=(1-h_{ii})\sigma^2$ 中，h_{ii} 是帽子矩阵中主对角线的第 i 个元素，它是调节 e_i 方差大小的杠杆，因而称 h_{ii} 为第 i 个观察值的杠杆值。类似于一元线性回归，多元线性回归的杠杆值 h_{ii} 也是表示自变量的第 i 次观测值与自变量平均值之间距离的远近。较大的杠杆值的残差偏小，这是因为大杠杆值的观测点远离样本中心，能够把回归方程拉向自己，因而把杠杆值大的样本点称为强影响点。

§4.5 异常值与强影响值

二、关于自变量x的异常值

根据（3.22）式，$tr(H)=\sum_{i=1}^{n} h_{ii} = p+1$，则杠杆值 h_{ii} 的平均值为

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{p+1}{n}$$

一个杠杆值 h_{ii} 的如果大于 2 倍或 3 倍的 $\bar{h}$ 就认为是大的

§4.5 异常值与强影响值

二、关于自变量x的异常值

SPSS 软件计算出的是中心化杠杆值 ch_{ii}，也就是自变量中心化后生成的帽子矩阵的主对角线元素，由参考文献[2]可知，

$$ch_{ii}=h_{ii}-1/n$$

因此，$\sum_{i=1}^{n} ch_{ii}$ =p，中心化杠杆值 ch_{ii} 的平均值是

$$\overline{ch}=\frac{1}{n}\sum_{i=1}^{n} ch_{ii}=\frac{p}{n}$$

§4.5 异常值与强影响值

二、关于自变量x的异常值

虽然强影响点并不总是y的异常值点，不能单纯根据杠杆值h_{ii}的大小判断强影响点是否异常，但是我们对强影响点应该有足够的重视。为此引入库克距离，用来判断强影响点是否为y的异常值点。库克距离的计算公式为：

$$D_i = \frac{e_i^2}{(p+1)\hat{\sigma}^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

§4.5 异常值与强影响值

二、关于自变量x的异常值

库克距离反应了杠杆值 h_{ii} 与残差 e_i 大小的一个综合效应。

对于库克距离，判断其大小的方法比较复杂，一个粗略的标准是

当$D_i<0.5$时，认为不是异常值点，

当$D_i>1$时，　认为是异常值点。

§4.5 异常值与强影响值

三、异常值实例分析

以例3.2的北京开发区的数据为例，做异常值的诊断分析。分别计算普通残差e_i，学生化残差SRE_i，删除残差$e_{(i)}$，删除学生化残差$SRE_{(i)}$，杠杆值ch_{ii}，库克距离D_i，见表4.10

§4.5 异常值与强影响值

序号	x_1	x_2	y	e_i	SRE_i	$e_{(i)}$	$SRE_{(i)}$	ch_{ii}	D_i
1	25	3547.79	553.96	-832	-2.340	-1490	-3.038	0.375	1.445
2	20	896.34	208.55	75	0.167	84	0.160	0.043	0.001
3	6	750.32	3.10	-34	-0.075	-38	-0.072	0.054	0.000
4	1001	2087.05	2815.40	127	0.376	253	0.363	0.432	0.047
5	525	1639.31	1052.12	-458	-1.034	-529	-1.037	0.068	0.055
6	825	3357.70	3427.00	502	1.305	768	1.348	0.280	0.302
7	120	808.47	442.82	147	0.326	164	0.313	0.036	0.004
8	28	520.27	70.12	96	0.218	112	0.209	0.070	0.003
9	7	671.13	122.24	121	0.271	138	0.261	0.060	0.004
10	532	2863.32	1400.00	-697	-1.606	-837	-1.735	0.100	0.172
11	75	1160.00	464.00	95	0.209	104	0.201	0.021	0.001
12	40	862.75	7.50	-151	-0.336	-169	-0.323	0.040	0.005
13	187	672.99	224.18	-145	-0.324	-164	-0.312	0.052	0.005
14	122	901.76	538.94	195	0.431	216	0.416	0.029	0.007
15	74	3546.18	2442.79	958	2.613	1613	3.810	0.339	1.555

§4.5 异常值与强影响值

绝对值最大的学生化残差为SRE_{15}=2.613，小于3。

绝对值最大的删除学生化残差为$SRE_{(15)}$=3.810，因而根据学生化残差诊断认为第15个数据为异常值。其中心化杠杆值ch_{ii}=0.339位于第3大，库克距离 D_i=1.555位于第一大。由于

$$\overline{ch}=\frac{p}{n}=\frac{2}{15}=0.13333$$

第 15 个数据 h_{ii}=0.339>2$\bar{h}$，因而从杠杆值看第 15 个数据是自变量的异常值，同时库克距离 D_{15}=1.555>1，这样第 15 个数据为异常值的原因是由自变量异常与因变量异常两个原因共同引起的。

§4.5 异常值与强影响值

异常值原因	异常值消除方法
1. 数据登记误差，存在抄写或录入的错误	重新核实数据
2. 数据测量误差	重新测量数据
3. 数据随机误差	删除或重新观测异常值数据
4. 缺少重要自变量	增加必要的自变量
5. 缺少观测数据	增加观测数据，适当扩大自变量取值范围
6. 存在异方差	采用加权线性回归
7. 模型选用错误，线性模型不适用	改用非线性回归模型

§4.5 异常值与强影响值

对本例的数据，通过核实认为不存在登记误差和测量误差。

删除第 15 组数据，用其余 14 组数据拟合回归方程，发现第 6 组数据的删除学生化残差增加为 $SRE_{(6)}=4.418$，仍然存在异常值现象，因而认为异常值的原因不是由于数据的随机误差。

实际上，在本章第三节中已经诊断出本例数据存在异方差，应该采用加权最小二乘回归。权数为 $W_i=x_2^{-2.5}$。用 SPSS 软件计算出加权最小二乘回归的有关变量值如下表所示：

§4.5 异常值与强影响值

序号	x_1	x_2	y	e_i	SRE_i	$e_{(i)}$	$SRE_{(i)}$	ch_{ii}	D_i
1	25	3547.79	553.96	-890	-1.149	-1165	-1.1658	0.2341	0.1360
2	20	896.34	208.55	20	0.135	23	0.1293	0.0604	0.0009
3	6	750.32	3.10	-93	-0.795	-110	-0.7824	0.0501	0.0385
4	1001	2087.05	2815.40	403	1.175	716	1.1963	0.4294	0.3581
5	525	1639.31	1052.12	-343	-1.135	-429	-1.1498	0.1864	0.1081
6	825	3357.70	3427.00	715	0.937	841	0.9320	0.1471	0.0515
7	120	808.47	442.82	126	0.949	139	0.9448	0.0093	0.0318
8	28	520.27	70.12	45	0.717	74	0.7015	0.1339	0.1115
9	7	671.13	122.24	62	0.617	76	0.6008	0.0463	0.0287
10	532	2863.32	1400.00	-582	-0.926	-677	-0.9199	0.1366	0.0466
11	75	1160.00	464.00	58	0.281	65	0.2702	0.0748	0.0033
12	40	862.75	7.50	-199	-1.391	-223	-1.4544	0.0324	0.0765
13	187	672.99	224.18	-143	-1.611	-224	-1.7424	0.2272	0.4951
14	122	901.76	538.94	175	1.137	189	1.1528	0.0112	0.0360
15	74	3546.18	2442.79	916	1.173	1179	1.1939	0.2209	0.1317

§4.5 异常值与强影响值

采用加权最小二乘回归后，删除学生化残差$SRE_{(i)}$的绝对值最大者为$|SRE_{(13)}|=1.7424$，库克距离都在0.5至1.0之间，说明数据没有异常值。这个例子也说明了用加权最小二乘法处理异方差性问题的有效性。