第六章 多重共线性、异方差及自相关

§6.1 多重共线性

- 一、多重共线性的概念
- 二、多重共线性的后果
- 三、多重共线性的诊断
- 四、克服多重共线性的方法
- 五、案例

一、多重共线性的概念

对于模型 $$Y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

其基本假设之一是解释变量是互相独立的。

如果某两个或多个解释变量之间出现了相关性，则称为多重共线性(**Multicollinearity**)。

含义：解释变量的样本向量近似线性相关。

多重共线性来源：

（1）解释变量x受到同一个因素的影响；

例如：政治事件对很多变量都产生影响，这些变量同时上升或同时下降。

（2）解释变量X自己的当期和滞后期；

（3）错误设定。

二、多重共线性的后果

1、完全共线性下参数估计量不存在

$$Y = X\beta + \varepsilon$$

的OLS估计量为：

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

完全共线性指的是解释变量中某个变量是其他变量的线性组合，即$c_1X_1+c_2X_2+\ldots+c_kX_k=0$
其中c_i不全为0，$i=1,\ldots k$

如果存在完全共线性，则$(X'X)^{-1}$不存在，无法得到参数的估计量。

例：对离差形式的二元回归模型

$$y = \beta_1 x_1 + \beta_2 x_2 + \mu$$

如果两个解释变量完全相关，如$x_2 = \lambda x_1$，则

$$y = (\beta_1 + \lambda\beta_2) x_1 + \mu$$

这时，只能确定综合参数$\beta_1 + \lambda\beta_2$的估计值：

$$\widehat{\beta_1 + \lambda\beta_2} = \sum x_{1i} y_i \Big/ \sum x_{1i}^2$$

一个方程确定两个未知数，有无穷多个解。

2、近似共线性下OLS估计量非有效

近似共线性下，可以得到OLS参数估计量，但参数估计量**方差**的表达式为 $Cov(\hat{\boldsymbol{\beta}})=\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

由于$(X'X)^{-1}\approx 0$，引起主对角线元素 $(X'X)^{-1}$ 较大，使参数估计值的方差增大，**OLS参数估计量非有效。**

近似共线性指的是解释变量中某个变量不完全是其他解释变量的线性组合，还差个扰动项。即$c_1X_1+c_2X_2+\ldots+c_kX_k+v_i=0$，其中$c_i$不全为0，i=1,…k

3、参数估计量经济含义不合理

如果模型中两个解释变量具有线性相关性，例如 $X_2 = \lambda X_1$，

这时，X_1和X_2前的参数β_1、β_2并不反映各自与被解释变量之间的结构关系，而是反映它们对被解释变量的共同影响。

β_1、β_2已经失去了应有的经济含义，于是经常表现出似乎反常的现象：例如β_1本来应该是正的，结果恰是负的。

注：除非是完全共线性，多重共线性并不意味着任何基本假设的违背。

三、多重共线性的诊断

多重共线性表现为解释变量之间具有相关关系，所以用于多重共线性的检验方法主要是统计方法：如判定系数检验法、逐步回归检验、方差膨胀因子（VIF）法等。

多重共线性诊断的任务是：

（1）检验多重共线性是否存在；

（2）估计多重共线性的范围，即判断哪些变量之间存在共线性。

1、检验多重共线性是否存在

(1) 对两个解释变量的模型，采用简单相关系数法

求出X_1与X_2的简单相关系数r，若|r|接近1，则说明两变量存在较强的多重共线性。

(2) 对多个解释变量的模型，采用综合统计检验法

若 在OLS法下：R^2与F值较大，但t检验值较小，说明各解释变量对Y的联合线性作用显著，但各解释变量间存在共线性而使得它们对Y的独立作用不能分辨，故t检验不显著。

2、判明存在多重共线性的范围

如果存在多重共线性，需进一步确定究竟由哪些变量引起。

(1) 判定系数检验法

使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

$$X_{ji}=\alpha_1 X_{1i}+\alpha_2 X_{2i}+\ldots\alpha_k X_{ki}$$

的判定系数较大，说明 X_j 与其他X间存在共线性。

具体可进一步对上述回归方程作F检验：

构造如下F统计量

$$F_j = \frac{R_j^2 / k-1}{(1-R_j^2)/(n-k)} \sim F(k-1, n-k)$$

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可决系数，

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α，计算F值，并与相应的临界值比较，来判定是否存在相关性。

另一等价的检验是:

在模型中排除某一个解释变量 X_j，估计模型；

如果拟合优度与包含X_j时十分接近，则说明X_j与其它解释变量之间存在共线性。

(2)逐步回归法

以Y为被解释变量，逐个引入解释变量，构成回归模型，进行模型估计。

根据拟合优度的变化决定新引入的变量是否独立。

如果拟合优度变化显著，则说明新引入的变量是一个独立解释变量；

如果拟合优度变化很不显著，则说明新引入的变量与其它变量之间存在共线性关系。

(3) 方差膨胀因子
（VIF: Variance Inflation Factor ）

VIF指标： $VIF(b_K)=\dfrac{1}{1-R_K^2}$　　$K=1,\ldots,k$

VIF范围：[+1,+∞)

X_k与其余变量回归所得的可决系数

判断：若VIF≥5，则认为多重共线性强，不可接受。

(4) 条件数（Condition Indix）

$$条件数=\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

解释变量的相关矩阵的最大特征值与最小特征值相比

调用数据库 neiyun.dta 讲解。

条件数大于30，认为多重共线存在。

四、克服多重共线性的方法

如果模型被检验证明存在多重共线性，则需要发展新的方法估计模型，最常用的方法有三类。

1、第一类方法：排除引起共线性的变量

找出引起多重共线性的解释变量，将它排除出去。

以**逐步回归法**得到最广泛的应用。

- **注意：**

这时，剩余解释变量参数的经济含义和数值都发生了变化。

2、第二类方法：差分法

时间序列数据、线性模型：将原模型变换为差分模型:

$$\Delta Y_i=\beta_1 \Delta X_{1i}+\beta_2 \Delta X_{2i}+\ldots+\beta_k \Delta X_{ki}+\Delta \mu_i$$

可以有效地消除原模型中的多重共线性。

一般讲，增量之间的线性关系远比总量之间的线性关系弱得多。

3、第三类方法：减小参数估计量的方差

多重共线性的主要后果是参数估计量具有较大的方差，所以

采取适当方法减小参数估计量的方差，虽然没有消除模型中的多重共线性，但确能消除多重共线性造成的后果。

例如：

增加样本容量可使参数估计量的方差减小。

六、案例——中国粮食生产函数

根据理论和经验分析，影响粮食生产（Y）的主要因素有：

农业化肥施用量（X_1）；粮食播种面积(X_2)

成灾面积(X_3)；农业机械总动力(X_4)；

农业劳动力(X_5)

已知中国粮食生产的相关数据，建立中国粮食生产函数：

$$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5+\mu$$

调用数据库E:\博士计量课程软件应用\multi

表 4.3.3 中国粮食生产与相关投入资料

年份	粮食产量 Y (万吨)	农业化肥施用量 X_1（万公斤）	粮食播种面积 X_2（千公顷）	受灾面积 X_3（公顷）	农业机械总动力 X_4（万千瓦）	农业劳动力 X_5（万人）
1983	38728	1659.8	114047	16209.3	18022	31645.1
1984	40731	1739.8	112884	15264.0	19497	31685.0
1985	37911	1775.8	108845	22705.3	20913	30351.5
1986	39151	1930.6	110933	23656.0	22950	30467.0
1987	40208	1999.3	111268	20392.7	24836	30870.0
1988	39408	2141.5	110123	23944.7	26575	31455.7
1989	40755	2357.1	112205	24448.7	28067	32440.5
1990	44624	2590.3	113466	17819.3	28708	33330.4
1991	43529	2806.1	112314	27814.0	29389	34186.3
1992	44264	2930.2	110560	25894.7	30308	34037.0
1993	45649	3151.9	110509	23133.0	31817	33258.2
1994	44510	3317.9	109544	31383.0	33802	32690.3
1995	46662	3593.7	110060	22267.0	36118	32334.5
1996	50454	3827.9	112548	21233.0	38547	32260.4
1997	49417	3980.7	112912	30309.0	42016	32434.9
1998	51230	4083.7	113787	25181.0	45208	32626.4
1999	50839	4124.3	113161	26731.0	48996	32911.8
2000	46218	4146.4	108463	34374.0	52574	32797.5

1、用OLS法估计上述模型：

$$\hat{Y} = -12816.44 + 6.213X_1 + 0.421X_2 - 0.166X_3 - 0.098X_4 - 0.028X_5$$

(-0.91)　(8.39)　(3.32)　(-2.81)　(-1.45)　(-0.14)

$R^2=0.9828$　$\bar{R}^2=0.9756$　$F=137.11$　$DW=1.81$

R^2接近于1；

给定α=5%，得F临界值 $F_{0.05}(5,12)=3.11$

F=137.11 > 3.11，

故认上述粮食生产的总体线性关系显著成立。

但X_4、X_5的参数未通过t检验，且符号不正确，

故**解释变量间可能存在多重共线性**。

2、检验简单相关系数

列出X_1，X_2，X_3，X_4，X_5的相关系数矩阵：

	X1	X2	X3	X4	X5
X1	1.00	0.01	0.64	0.96	0.55
X2	0.01	1.00	-0.45	-0.04	0.18
X3	0.64	-0.45	1.00	0.69	0.36
X4	0.96	-0.04	0.69	1.00	0.45
X5	0.55	0.18	0.36	0.45	1.00

- **发现：** X_1与X_4间存在高度相关性。

3、找出最简单的回归形式

分别作Y与X_1，X_3，X_2，X_4，X_5间的回归：

$$\hat{Y} = 30867.64 + 4.576X_1$$

(25.58)　　(11.49)

$R^2=0.8919$　$F=132.1$

$$\hat{Y} = -33821.18 + 0.699X_2$$

(-0.49)　　(1.14)

$R^2=0.075$　$F=1.30$

$$\hat{Y} = 37712.86 + 0.3549X_3$$

(1.74)　　(7.25)

$R^2=0.1596$　$F=3.04$

$$\hat{Y} = 31919.0 + 0.380X_4$$

(17.45)　　(6.68)

$R^2=0.7527$　$F=48.7$

$$\hat{Y} = -28259.19 + 2.240X_5$$

(-1.04)　　(2.66)

$R^2=0.3064$　$F=7.07$

- 可见，应选第1个式子为初始的回归模型。

4、逐步回归

将其他解释变量分别导入上述初始回归模型，寻找最佳回归方程。

	C	X1	X2	X3	X4	X5	$\bar{R}^2$	DW
Y=f(X1)	30868	4.23					0.8852	1.56
t 值	25.58	11.49						
Y=f(X1,X2)	-43871	4.65	0.67				0.9558	2.01
t 值	-3.02	18.47	5.16					
Y=f(X1,X2,X3)	-11978	5.26	0.41	-0.19			0.9752	1.53
t 值	0.85	19.6	3.35	-3.57				
Y=f(X1,X2,X3,X4)	-13056	6.17	0.42	-0.17	-0.09		0.9775	1.80
t 值	-0.97	9.61	3.57	-3.09	-1.55			
Y=f(X1,X3,X4,X5)	-12690	5.22	0.40	-0.20		0.07	0.9798	1.55
t 值	-0.87	17.85	3.02	-3.47		0.37		

5、结论

回归方程以Y=f（X_1，X_2，X_3）为最优：

$$Y = -11978 + 5.26X_1 + 0.41X_2 - 0.19X_3$$

§5.2 异方差（Heteroscedasticity）

1、同方差假定及异方差定义

模型的假定条件(1) 给出Var(ε) 是一个对角矩阵，

$$\mathrm{Var}(\varepsilon) = \sigma^2 I = \sigma^2 \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix}$$

且 ε 的方差协方差矩阵主对角线上的元素都是常数且相等，即每一误差项的方差都是有限的相同值（同方差假定）；且非主对角线上的元素为零（无自相关假定），

当这个假定不成立时，

Var(ε) 不再是一个纯量对角矩阵。

$$\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\sigma}^2\boldsymbol{\Omega} = \boldsymbol{\sigma}^2\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1T} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2T} \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{T1} & \sigma_{T2} & \cdots & \sigma_{TT} \end{pmatrix} \neq \sigma^2 \boldsymbol{I}.$$

当误差向量 ε 的方差协方差矩阵主对角线上的元素不相等时，称该随机误差系列存在异方差，即误差向量 ε 中的元素取自不同的分布总体。非主对角线上的元素表示误差项之间的协方差值。比如 Ω 中的 σ_{ij}，（$i \neq j$）表示与第i组和第j组观测值相对应的ε_i与ε_j的协方差。若 Ω 非主对角线上的部分或全部元素都不为零，误差项就是自相关的（后面讲自相关）。

2. 异方差的表现

异方差通常有三种表现形式，
（1）递增型（2）递减型（3）条件自回归型。
递增型异方差见图5.21和5.22。图5.23为递减型异方差。
图5.24为条件自回归型异方差(复杂性异方差)。

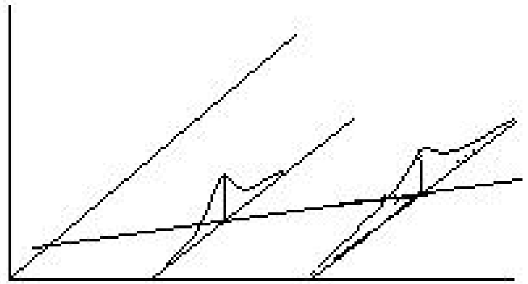


图5.21 递增型异方差情形

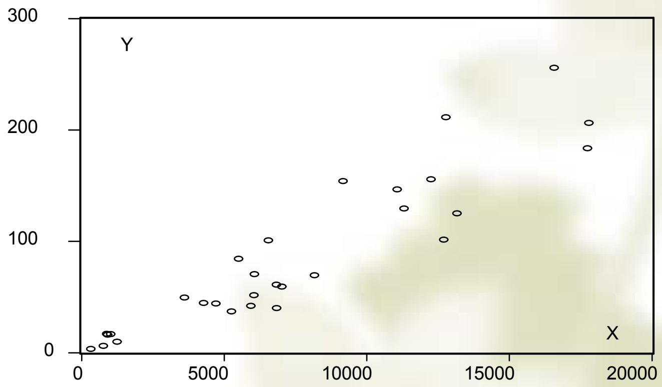


图5.22 递增型异方差

随着解释变量值的增大，被解释变量取值的差异性增大

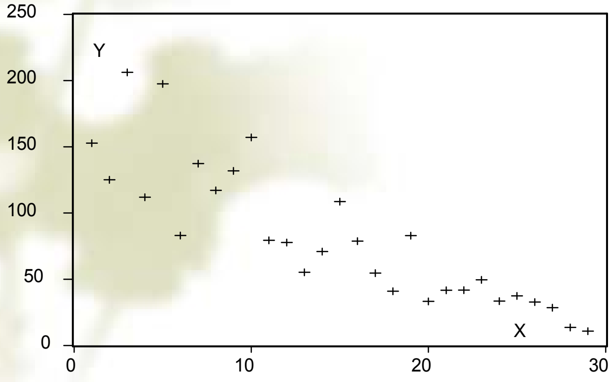


图5.23　递减型异方差

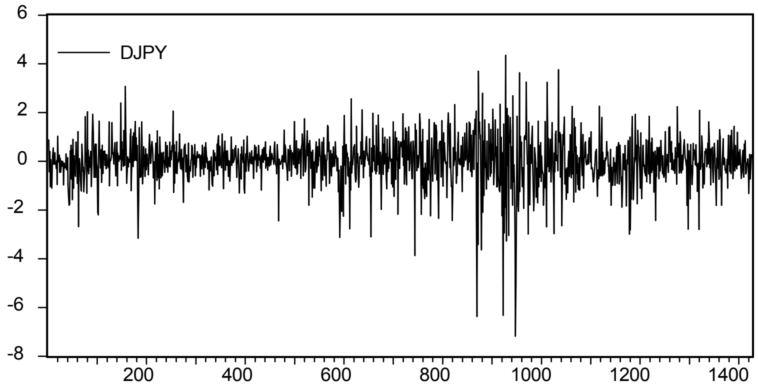


图5.6　复杂型异方差

注： 时间序列数据和截面数据中都有可能存在异方差。
经济时间序列中的异方差常为递增型异方差。
金融时间序列中的异方差常表现为自回归条件异方差。
无论是时间序列数据还是截面数据。递增型异方差的来源主要是因为随着解释变量值的增大，被解释变量取值的差异性增大。

3. 异方差的后果

下面以简单线性回归模型为例讨论异方差对参数估计的影响。

对模型 $y_t = \beta_0 + \beta_1 x_t + u_t$

（1）．当Var(u_t) = σ_t^2为异方差时（σ_t^2是一个随时间或序数变化的量），回归参数估计量仍具有无偏性和一致性，但是回归参数估计量不再具有有效性。

以 $\hat{\beta}_1$为例 $E(\hat{\beta}_1) = E\left(\frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sum (x_t - \bar{x})^2}\right)$

$$= E\left(\frac{\sum (x_t - \bar{x})[\beta_1 (x_t - \bar{x}) + u_t]}{\sum (x_t - \bar{x})^2}\right)$$

$$= \beta_1 + \frac{\sum (x_t - \bar{x}) E(u_t)}{\sum (x_t - \bar{x})^2} = \beta_1$$

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{E}(\hat{\beta}_1 - \beta_1)^2 = \mathrm{E}\left(\frac{\sum (x_t - \bar{x}) u_t}{\sum (x_t - \bar{x})^2}\right)^2$$

$$= \mathrm{E}\left(\frac{(\sum (x_t - \bar{x}) u_t)^2}{(\sum (x_t - \bar{x})^2)^2}\right) = \frac{\sum (x_t - \bar{x})^2 \mathrm{E}(u_t)^2}{(\sum (x_t - \bar{x})^2)^2}$$

$$= \frac{\sum (x_t - \bar{x})^2 \sigma_t{}^2}{(\sum (x_t - \bar{x})^2)^2} \neq \frac{\sigma^2}{\sum (x_t - \bar{x})^2}$$

上式不等号左侧项分子中的$\sigma_t{}^2$不是一个常量，不能从累加式中提出，所以不等号右侧项不等于不等号左侧项。而不等号右侧项是同方差条件下β_1的最小二乘估计量的方差。因此异方差条件下的失去有效性。

（2）参数估计量的方差估计是真实方差的有偏估计

$$E(\mathrm{var}(\hat{\beta}_1)) \neq \mathrm{var}(\beta_1)$$

（3）t 检验失效

4. 异方差的诊断

((1))经济变量规模差别很大时容易出现异方差。如个人收入与支出关系，投入与产出关系。

(2) 利用散点图做初步判断。

(3) White检验　检验假设：H_0: $\sigma_i^2 = \sigma^2$ 即，同方差

White检验的具体步骤如下：

第一步：$Y \underset{OLS}{\text{———}} X \longrightarrow e_1, \ldots, e_n$

第二步：$e_i^2 \xrightarrow{\text{OLS}}$ 原变量、原变量平方、交叉项

$\longrightarrow R^2$

第三步：$nR^2 \sim \chi^2(k-1)$

第四步：$nR^2 > \chi^2(k-1)$ 拒绝H_0，存在异方差。

（4）自回归条件异方差（*ARCH*）检验

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σ_t^2 看作是x_t的函数，而是把σ_t^2 看作误差滞后项u_{t-1}^2，u_{t-2}^2，… 的函数。*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：

$$\hat{u}_t^2 = \alpha_0 + \alpha_1 \hat{u}_{t-1}^2 + \dots + \alpha_n \hat{u}_{t-n}^2$$

检验假设：H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$

LM 统计量定义为：ARCH=$nR^2 \sim \chi^2(k-1)$

5. 克服异方差的方法

（1）采用GLS估计

若对于不全为零的实数$x_1,x_2,\ldots x_n$总有$f=f(x_1,x_2,\ldots x_n)>0$,则f称为正定二次型。

设模型为： $Y = X\beta + u$

其中$E(u) = 0$，$Var(u) = E(uu') = \sigma^2\Omega$， Ω已知

因为Ω是一个T阶正定矩阵，所以必存在一个非退化$T\times T$阶矩阵M使下式成立。$M\Omega M' = I_{T\times T}$ ➡ $M'M = \Omega^{-1}$ ➡

$MY = MX\beta + Mu$ 取$Y^* = MY$, $X^* = MX$, $u^* = Mu$

➡ $Y^* = X^*\beta + u^*$ ➡

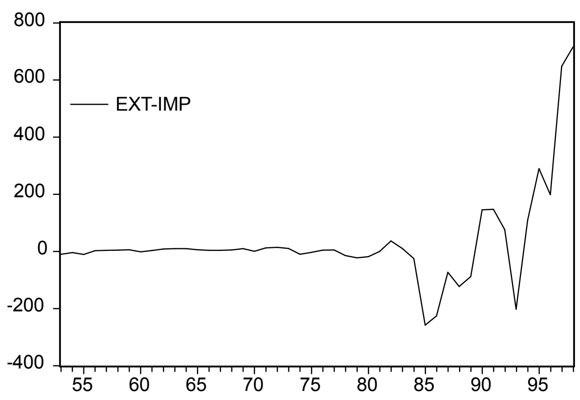
满秩矩阵

$$\mathrm{Var}(u^*) = \mathrm{E}(u^* u^{*\prime}) = \mathrm{E}(M u u' M')$$
$$= M \sigma^2 \Omega M' = \sigma^2 M \Omega M' = \sigma^2 I$$

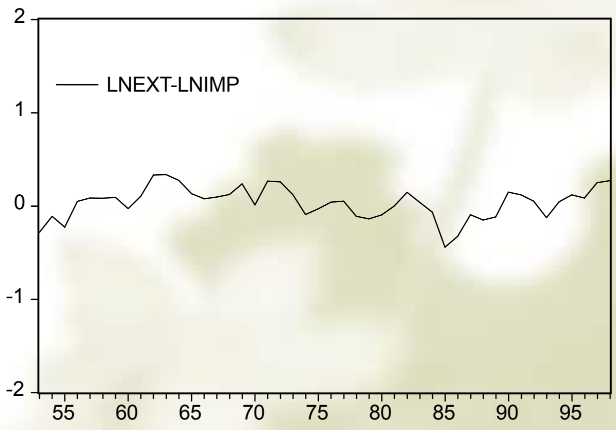
$$\hat{\beta}_{(GLS)} = (X^{*\prime} X^*)^{-1} X^{*\prime} Y^* = (X'M'MX)^{-1} X'M'MY$$
$$= (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}Y$$

这种方法成为广义最小二乘法（**GLS**）

(2) 通过对数据取对数消除异方差。



中国进出口贸易额差（**1953-1998**）



对数的中国进出口贸易额之差

§ 5.3 自相关 （ Autocorrelation ）

1. 非自相关的假定及自相关定义

对于模型

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i \qquad i=1,2,\cdots,n$$

随机误差项互不相关的基本假设表现为：

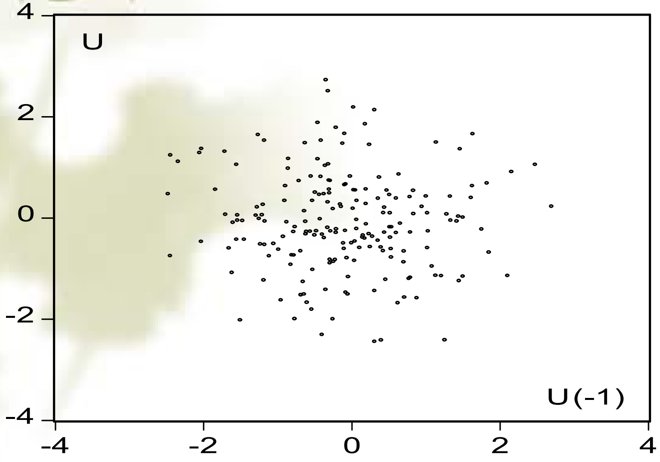
$$Cov(\mu_i, \mu_j) = 0 \qquad i \neq j, \quad i,j=1,2,\cdots,n$$

即误差项 u_t 的取值在时间上是相互无关的。
称误差项 u_t 非自相关。

注：自相关又称序列相关。

如果 $\text{Cov}(u_i, u_j) \neq 0, \ (i \neq j)$

则称误差项 u_t 存在自相关。



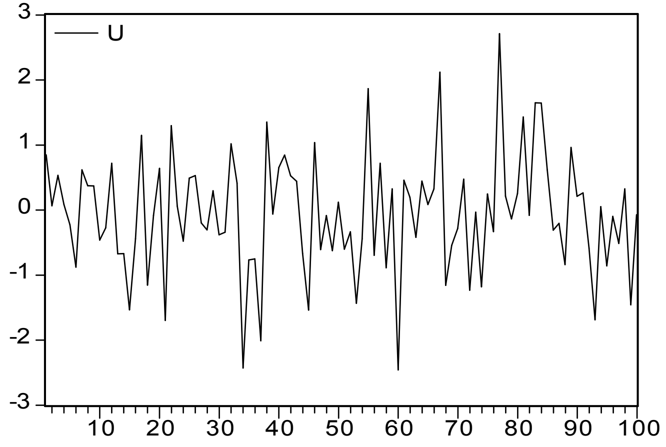
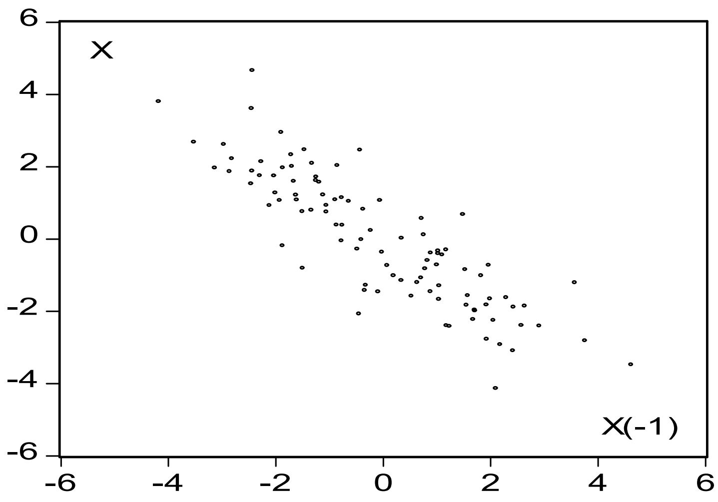


图1 非自相关的散点图

图2 非自相关的序列图



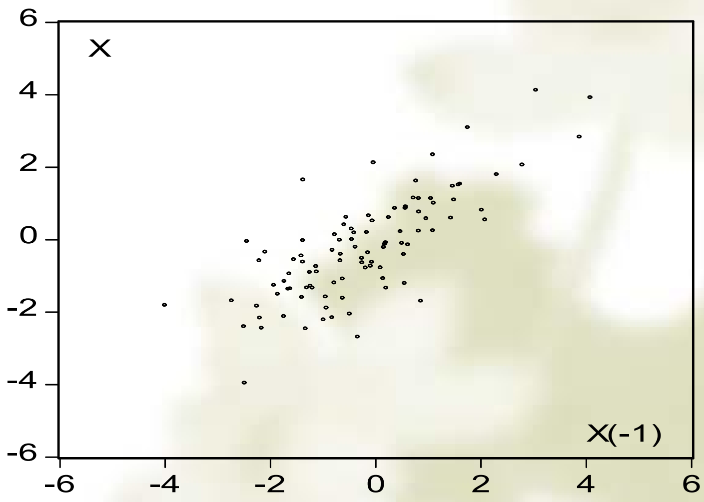


图3 负自相关的散点图

图4 正自相关的散点图

2. 自相关产生的原因

（1）惯性

大多数经济时间数据都有一个明显的特点，就是它的惯性。

GDP、价格指数、生产、就业与失业等时间序列都呈周期性，如周期中的复苏阶段，大多数经济序列均呈上升势，序列在每一时刻的值都高于前一时刻的值，似乎有一种内在的动力驱使这一势头继续下去，直至某些情况（如利率或课税的升高）出现才把它拖慢下来。

（2）设定偏误1：模型中未含应包括的变量

例如：如果对牛肉需求的正确模型应为

$$Y_t=\beta_0+\beta_1X_{1t}+\beta_2X_{2t}+\beta_3X_{3t}+\mu_t$$

其中：Y=牛肉需求量，X_1=牛肉价格，

X_2=消费者收入，X_3=猪肉价格

如果模型设定为：

$$Y_t= \beta_0+\beta_1X_{1t}+\beta_2X_{2t}+v_t$$

则该式中，$v_t= \beta_3X_{3t}+\mu_{t}$,

于是在猪肉价格影响牛肉消费量的情况下，这种模型设定的偏误往往导致随机项中有一个重要的系统性影响因素，使其呈序列相关性。

(3) 设定偏误2：不正确的函数形式

例如：如果真实的边际成本回归模型应为：

$$Y_t= \beta_0+\beta_1X_t+\beta_2X_t^2+\mu_t$$

其中：Y=边际成本，X=产出，

但建模时设立了如下模型：$Y_t= \beta_0+\beta_1X_t+v_t$

因此，由于$v_t= \beta_2X_t^2+\mu_t$，包含了产出的平方对随机项的系统性影响，随机项也呈现序列相关性。

(4) 蛛网现象

例如：农产品供给对价格的反映本身存在一个滞后期：

$$供给_t= \beta_0+\beta_1价格_{t-1}+\mu_t$$

意味着，农民由于在年度t的过量生产（使该期价格下降）很可能导致在年度t+1时削减产量，因此不能期望随机干扰项是随机的，往往产生一种蛛网模式。

3. 自相关性的后果

（1）参数估计量非有效

虽然回归系数 $\hat{\beta}$ 仍具有无偏性。

$$E(\hat{\beta}) = E\left[(X'X)^{-1}X'Y\right]$$
$$= E\left[(X'X)^{-1}X'(X\beta + \mu)\right]$$
$$= \beta + (X'X)^{-1}X'E(\mu) = \beta$$

但是 $\hat{\beta}$ 丧失有效性。

$$\operatorname{var}(\hat{\beta}) = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)'$$
$$= E\left[(X'X)^{-1}X'\mu\mu'X(X'X)^{-1}\right]$$
$$= (X'X)^{-1}X'E(\mu\mu')X(X'X)^{-1}$$
$$= (X'X)^{-1}X'\Omega X(X'X)^{-1} \neq \sigma^2(X'X)^{-1}$$

$\hat{\beta}_1$ 的方差比 μ_t 非自相关时大，失去有效性。

（2）变量的显著性检验失去意义

用OLS法估计时仍然用 $\sigma^2(X'X)^{-1}$ 估计，

所以会低估 $\hat{\beta}$ 的方差。

等于过高估计统计量t的值，从而把不重要的解释变量保留在模型里，使显著性检验失去意义。

(3) Var($\hat{\beta}_1$) 和s_u^2都变大，都不具有最小方差性。所以用依据普通最小二乘法得到的回归方程去预测，预测是无效的。

关于名词白噪声序列：零期望、同方差、无自相关序列。

4. 自相关检验

基本思路：

首先，采用 OLS 法估计模型，以求得随机误差项的“近似估计量”，用 $\tilde{e}_i$ 表示：

$$\tilde{e}_i = Y_i - (\hat{Y}_i)_{0ls}$$

然后，通过分析这些“近似估计量”之间的相关性，以达到判断随机误差项是否具有自相关性的目的。

（1）Durbin-Watson 检验法

D-W检验是J.Durbin）和G. S. Watson于1951年提出的一种检验序列自相关的方法，该方法的假定条件是：

（1）解释变量 X非随机；

（2）随机误差项μ_i为一阶自回归形式：

$$\mu_i=\rho\mu_{i-1}+\varepsilon_i$$

（3）回归模型中不应含有滞后因变量作为解释变量，即不应出现下列形式：

$$Y_i=\beta_0+\beta_1X_{1i}+\ldots\beta_kX_{ki}+\gamma Y_{i-1}+\mu_i$$

（4）回归含有截距项；

D.W.统计量

Durbin 和 Watson 假设：

$H_0: \rho = 0$， 即 μ_i 不存在一阶自回归；

$H_1: \rho \neq 0$， 即 μ_i 存在一阶自回归

并构如下造统计量：

$$D.W. = \frac{\sum_{i=2}^{n} (\tilde{e}_i - \tilde{e}_{i-1})^2}{\sum_{i=1}^{n} \tilde{e}_i^2} \qquad (2.5.5)$$

该统计量的分布与出现在给定样本中的X值有复杂的关系，因此其精确的分布很难得到。

DW检验步骤如下。给出假设：

H_0：ρ = 0 （μ_t 不存在自相关）

H_1：ρ ≠ 0 （μ_t 存在一阶自相关）

用残差值 e_t 计算统计量DW。

$$\mathbf{DW} = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^{\,2}} = \frac{\sum_{t=2}^{T} e_t^{\,2} + \sum_{t=2}^{T} e_{t-1}^{\,2} - 2\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^{\,2}}$$

因为 $\sum_{t=2}^{T} e_t^{\,2} \approx \sum_{t=2}^{T} e_{t-1}^{\,2} \approx \sum_{t=1}^{T} e_t^{\,2}$

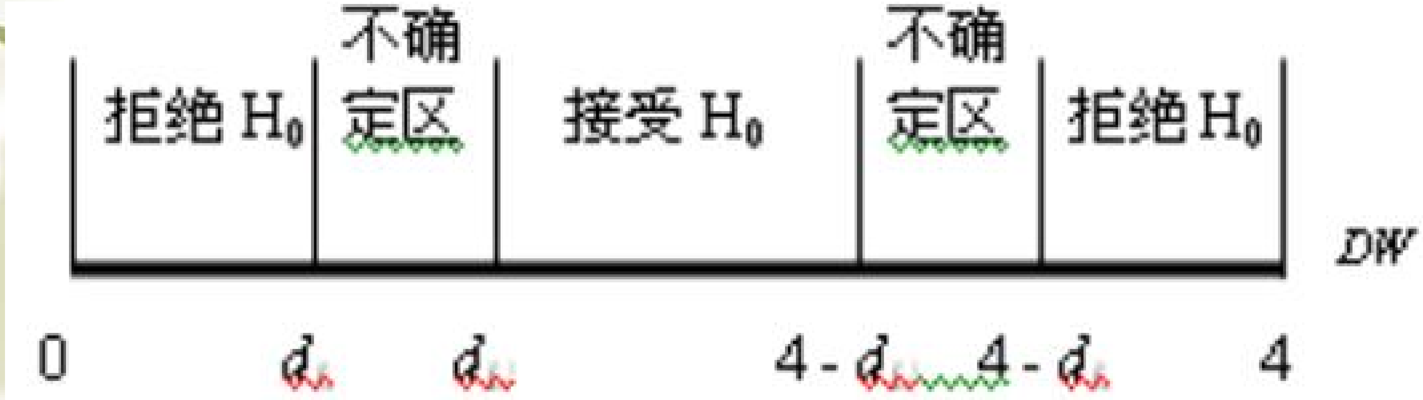
$$\mathbf{DW} = \frac{2\sum_{t=2}^{T} e_{t-1}^{\,2} - 2\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=2}^{T} e_{t-1}^{\,2}} = 2\left(1 - \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=2}^{T} e_{t-1}^{\,2}}\right) = 2(1-\hat{\rho})$$

因为 ρ 的取值范围是［-1，1］，
所以DW统计量的取值范围是［0，4］。

ρ 与DW值的对应关系见下表。

ρ	DW	μ_t 的表现
$\rho=0$	$DW=2$	μ_t非自相关
$\rho=1$	$DW=0$	μ_t完全正自相关
$\rho=-1$	$DW=4$	μ_t完全负自相关

实际中$DW=0, 2, 4$ 的情形是很少见的。当DW取值在（0, 2），（2, 4）之间时，怎样判别误差项 μ_t是否存在自相关呢？
推导统计量DW的精确抽样分布是困难的，因为DW是依据残差e_t 计算的，而e_t的值又与x_t的形式有关。DW检验与其它统计检验不同，它没有唯一的临界值用来制定判别规则。然而Durbin-Watson根据样本容量和被估参数个数，在给定的显著性水平下，给出了检验用的上、下两个临界值d_U和d_L 。判别规则如下：



若 $0<D.W.<d_l$ 则存在正自相关

$d_l<D.W.<d_u$ 不能确定

$d_u<D.W.<4-d_u$ 无自相关

$4-d_u<D.W.<4-d_l$ 不能确定

$4-d_l<D.W.<4$ 存在负自相关

可以看出，当D.W.值在2左右时，模型不存在一阶自相关。

当*DW*值落在“不确定”区域时，有两种处理方法。
①加大样本容量或重新选取样本，重作*DW*检验。
②选用其它检验方法。

注意：①因为***DW***统计量是以解释变量非随机为条件得出的，所以当有滞后的内生变量作解释变量时，***DW***检验无效。

②不适用于联立方程模型中各方程的序列自相关检验。

③***DW***统计量不适用于对高阶自相关的检验。

但在实际计量经济学问题中，一阶自相关是出现最多的一类序列相关；另外，经验表明，如果不存在一阶自相关，一般也不存在高阶序列相关。

（2）图示法

图示法就是依据残差e_t 对时间 t的序列图作出判断。

由于残差e_t是对误差项u_t 的估计，所以尽管误差项u_t 观测不到，但可以通过e_t的变化判断u_t 是否存在自相关。

图示法的具体步骤是，

（1）用给定的样本估计回归模型，

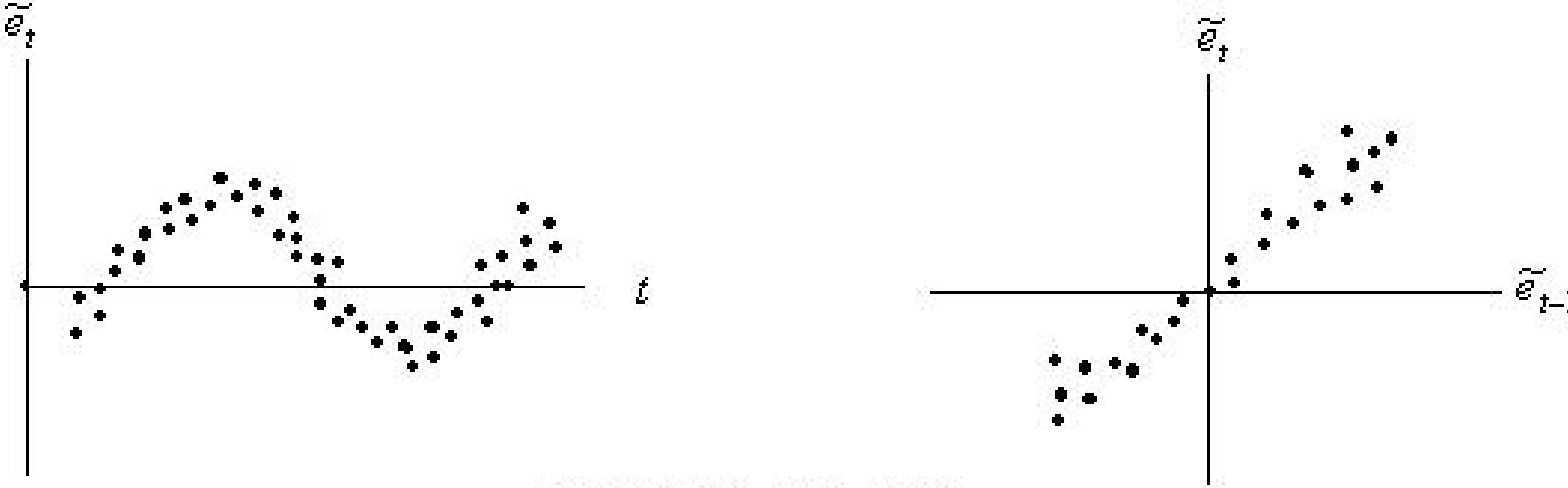
计算残差e_t ，（t = 1，2，… T），绘制残差图；

（2）分析残差图。

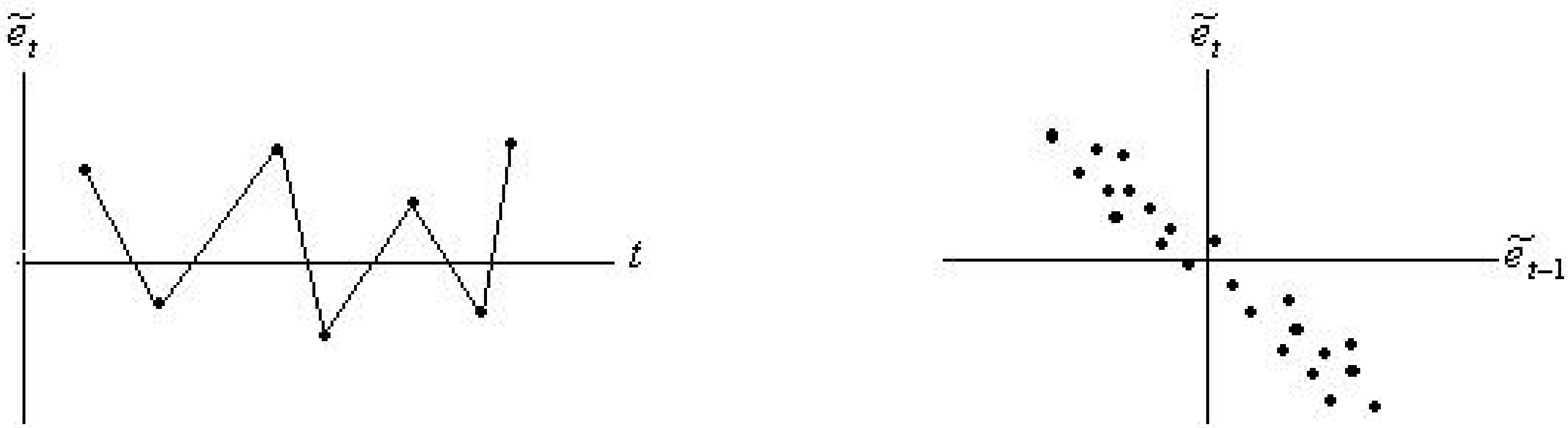
Stata实现语句：

```
reg y x
predict r,residuals
gen rlag＝r[_n-1]
Scatter r rlag
```

调用水果数据库



正序列相关（正自相关）



负序列相关（负自相关）

(3) 回归检验法

优点是，（1）适合于任何形式的自相关检验，
（2）若结论是存在自相关，则同时能提供出自相关的具体形式与参数的估计值。

缺点是，计算量大。

回归检验法的步骤如下：

①用给定样本估计模型并计算残差e_t。

②对残差序列e_t , (t = 1 ,2 ,… , T) 用OLS进行不同形式的回归拟合。如

$$e_t = \rho e_{t-1} + v_t$$

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

$$e_t = \rho e_{t-1}^2 + v_t$$

$$e_t = \rho \sqrt{e_{t-1}} + v_t$$

…

(3) 对上述各种拟合形式进行显著性检验，从而确定误差项u_t存在哪一种形式的自相关。

5. 克服自相关

如果模型的误差项存在自相关，首先应分析产生自相关的原因。如果自相关是由于错误地设定模型的数学形式所致，那么就应当修改模型的数学形式。

如果自相关是由于模型中省略了重要解释变量造成的，那么解决办法就是找出略去的解释变量，把它做为重要解释变量列入模型。

只有当以上两种引起自相关的原因都消除后，才能认为误差项u_t“真正”存在自相关。在这种情况下，解决办法是变换原回归模型，使变换后的随机误差项消除自相关，进而利用普通最小二乘法估计回归参数。这种变换方法称作广义最小二乘法（GLS）。

广义最小二乘法

对于模型

$$Y=X\beta+N \qquad \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

如果存在序列相关，同时存在异方差，即有

$$E(\mathrm{N})=0$$

$$Cov(\mathrm{NN}')=E(\mathrm{NN}')=\sigma^2\Omega$$

$$\Omega=\begin{bmatrix} w_1 & w_{12} & \cdots & w_{1n} \\ w_{21} & w_2 & \cdots & w_{2n} \\ & & \ddots & \\ w_{n1} & w_{n2} & \cdots & w_n \end{bmatrix}$$

设 $\Omega = DD'$

用$\mathbf{D}^{-1}$左乘上式两边，得到一个新的模型：

$$\mathbf{D}^{-1}\mathbf{Y} = \mathbf{D}^{-1}\mathbf{X}\beta + \mathbf{D}^{-1}\mathbf{N}$$

即 $$\mathbf{Y}^* = \mathbf{X}^*\beta + \mathbf{N}^* \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

该模型具有同方差性和随机误差项互相独立性。

可以证明：

$$E(\mathrm{N}^*\mathrm{N}^{*\prime}) = E(\mathbf{D}^{-1}\mathrm{N}\mathrm{N}'\mathbf{D}^{-1\prime})$$

$$= \mathbf{D}^{-1}E(\mathrm{N}\mathrm{N}')\mathbf{D}^{-1\prime}$$

$$= \mathbf{D}^{-1}\sigma^2\mathbf{\Omega}\mathbf{D}^{-1\prime}$$

$$= \mathbf{D}^{-1}\sigma^2\mathbf{D}\mathbf{D}'\mathbf{D}^{-1\prime}$$

$$= \sigma^2\mathbf{I}$$

于是，可以用OLS法估计模型(2)，得

$$\hat{\mathrm{B}} = (\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^{*}$$

$$= (\mathbf{X}'\mathbf{D}^{-1\prime}\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1\prime}\mathbf{D}^{-1}\mathbf{Y}$$

$$= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{Y}$$

这就是原模型(1)的广义最小二乘估计量(GLS estimators)，是无偏的、有效的估计量。

如何得到矩阵Ω?

仍然是对原模型(1)首先采用普通最小二乘法，
得到随机误差项的近似估计量，以此构成矩阵的估计量Ω，即

$$\hat{\Omega} = \begin{bmatrix} \tilde{e}_1^2 & \tilde{e}_1\tilde{e}_2 & \cdots & \tilde{e}_1\tilde{e}_n \\ \tilde{e}_2\tilde{e}_1 & \tilde{e}_2^2 & \cdots & \tilde{e}_2\tilde{e}_n \\ & & \ddots & \\ \tilde{e}_n\tilde{e}_1 & \tilde{e}_n\tilde{e}_2 & \cdots & \tilde{e}_n^2 \end{bmatrix}$$

差分法

差分法是一类克服自相关性的有效的方法，被广泛地采用。

差分法是将原模型变换为差分模型，分为

一阶差分法(first-difference method)和

广义差分法(generalized difference method)。

下面以一元线性模型为例说明。

（1）一阶差分法

将原模型 $Y_i = \beta_0 + \beta_1 X_i + \mu_i \qquad i = 1,\dots n$

变化为 $\Delta Y_i = \beta_1 \Delta X_i + (\mu_i - \mu_{i-1})$

其中 $\Delta Y_i = Y_i - Y_{i-1}$

$\Delta X_i = X_i - X_{i-1}$

如果原模型存在完全一阶正自相关，即在

$\mu_i=\rho\mu_{i-1}+\varepsilon_i$

中，$\rho=1$。 模型可变换为：

$\Delta Y_i= \beta_1\Delta X_i+\varepsilon_i$

由于ε_i不存在序列相关，该差分模型满足应用**OLS**法的基本假设；

因此，用**OLS**法估计可得到原模型参数的无偏的、有效的估计量。

即使对于非完全一阶正相关的情况，只要存在一定程度的一阶正相关，差分模型就可以有效地加以克服。

（2）广义差分法

如果原模型存在 $\mu_i = \rho_1 \mu_{i-1} + \rho_2 \mu_{i-2} + \dots + \rho_l \mu_{i-l} + \varepsilon_i$

可以将原模型变为

$$
\begin{aligned}
& Y_i - \rho_1 Y_{i-1} - \dots - \rho_l Y_{i-l} \\
& = \beta_0 (1 - \rho_1 - \dots - \rho_l) + \beta_1 (X_i - \rho_1 X_{i-1} - \dots \rho_l X_{i-l}) + \varepsilon_i \\
& i = 1 + l, 2 + l, \dots n
\end{aligned}
$$

此模型称为广义差分模型，该模型不存在序列相关问题。采用**OLS**法估计可以得到原模型参数的无偏、有效的估计量。

广义差分法可以克服所有类型的序列相关带来的问题，一阶差分法是它的一个特例。

随机误差项相关系数ρ的估计

应用广义差分法，必须已知不同样本点之间随机误差项的相关系数ρ_1，ρ_2，…，ρ_l。实际上，人们并不知道它们的具体数值，所以必须首先对它们进行估计。

常用的方法有：

（1）科克伦-奥科特（Cochrane-Orcutt）迭代法。

（2）杜宾（durbin）两步法

（1）科克伦-奥科特迭代法。

首先，采用OLS法估计原模型

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

得到的随机误差项的“近似估计值”，并以之作为观测值采用OLS法估计下式

$$\mu_i=\rho_1\mu_{i-1}+\rho_2\mu_{i-2}+\ldots\rho_L\mu_{i-L}+\varepsilon_i$$

得到 $\hat{\rho}_1,\hat{\rho}_2,\cdots,\hat{\rho}_l$，作为随机误差项的相关系数 $\rho_1,\rho_2,\cdots,\rho_l$ 的第一次估计值。

其次，将上述$\hat{\rho}_1,\hat{\rho}_2,\cdots,\hat{\rho}_l$代入广义差分模型

$$Y_i-\rho_1 Y_{i-1}-\cdots-\rho_l Y_{i-l}=\beta_0(1-\rho_1-\cdots-\rho_l)+\beta_1(X_i-\rho_1 X_{i-1}-\cdots-\rho_l X_{i-l})+\varepsilon_i$$

$$i=1+l,2+l,\cdots,n$$

并对之进行 OLS 估计，得到$\hat{\hat{\beta}}_0$、$\hat{\hat{\beta}}_1$。

再次，将$\hat{\hat{\beta}}_0$、$\hat{\hat{\beta}}_1$代回原模型，计算出原模型随机误差项的新的“近拟估计值”，并以之作为模型

$$\mu_i=\rho_1\mu_{i-1}+\rho_2\mu_{i-2}+\cdots+\rho_l\mu_{i-l}+\varepsilon_i$$

的样本观测值，采用 OLS 法估计该方程，得到$\hat{\hat{\rho}}_1,\hat{\hat{\rho}}_2,\cdots,\hat{\hat{\rho}}_l$，作为相关系数$\rho_1,\rho_2,\cdots,\rho_l$的**第二次估计值**。

类似地，可进行第三次、第四次迭代。

关于迭代的次数，可根据具体的问题来定。

一般是事先给出一个精度，当相邻两次ρ_1，ρ_2，…，ρ_L的估计值之差小于这一精度时，迭代终止。

实践中，有时只要迭代两次，就可得到较满意的结果。两次迭代过程也被称为科克伦-奥科特两步法。

（2）杜宾（durbin）两步法

该方法仍是先估计ρ_1，ρ_2，…，ρ_L，再对差分模型进行估计

第一步，变换差分模型为下列形式：

$$Y_i = \rho_1 Y_{i-1} + \ldots + \rho_l Y_{i-l} + \beta_0 (1 - \rho_1 - \ldots - \rho_l) + \beta_1 (X_i - \rho_1 X_{i-1} - \cdots \rho_l X_{i-l}) + \varepsilon_i$$

$$i = 1 + l, 2 + l, \ldots n$$

采用 OLS 法估计该方程，得各$Y_j (j = i-1, i-2, \cdots i-l)$前的系数$\rho_1, \rho_2, \cdots, \rho_l$的估计值$\hat{\rho}_1, \hat{\rho}_2, \cdots, \hat{\rho}_l$。

第二步，将估计的$\hat{\rho}_1,\hat{\rho}_2,\cdots,\hat{\rho}_l$代入差分模型

$$Y_i-\rho_1 Y_{i-1}-\cdots-\rho_l Y_{i-l}=\beta_0(1-\rho_1-\cdots-\rho_l)+\beta_1(X_i-\rho_1 X_{i-1}-\cdots-\rho_l X_{i-l})+\varepsilon_i$$

$$i=1+l,2+l,\cdots,n$$

采用 OLS 法估计，得到参数$\beta_0(1-\hat{\rho}_1-\cdots-\hat{\rho}_l),\beta_1$的估计量，记为$\hat{\beta}_0^*$，$\hat{\beta}_1^*$。

于是：

$$\hat{\beta}_0=\hat{\beta}_0^*\big/(1-\hat{\rho}_1-\cdots-\hat{\rho}_l)\,,\qquad \hat{\beta}_1=\hat{\beta}_1^*$$

五、案例：地区商品出口模型

某地区商品出口总值与国内生产总值的关系研究。

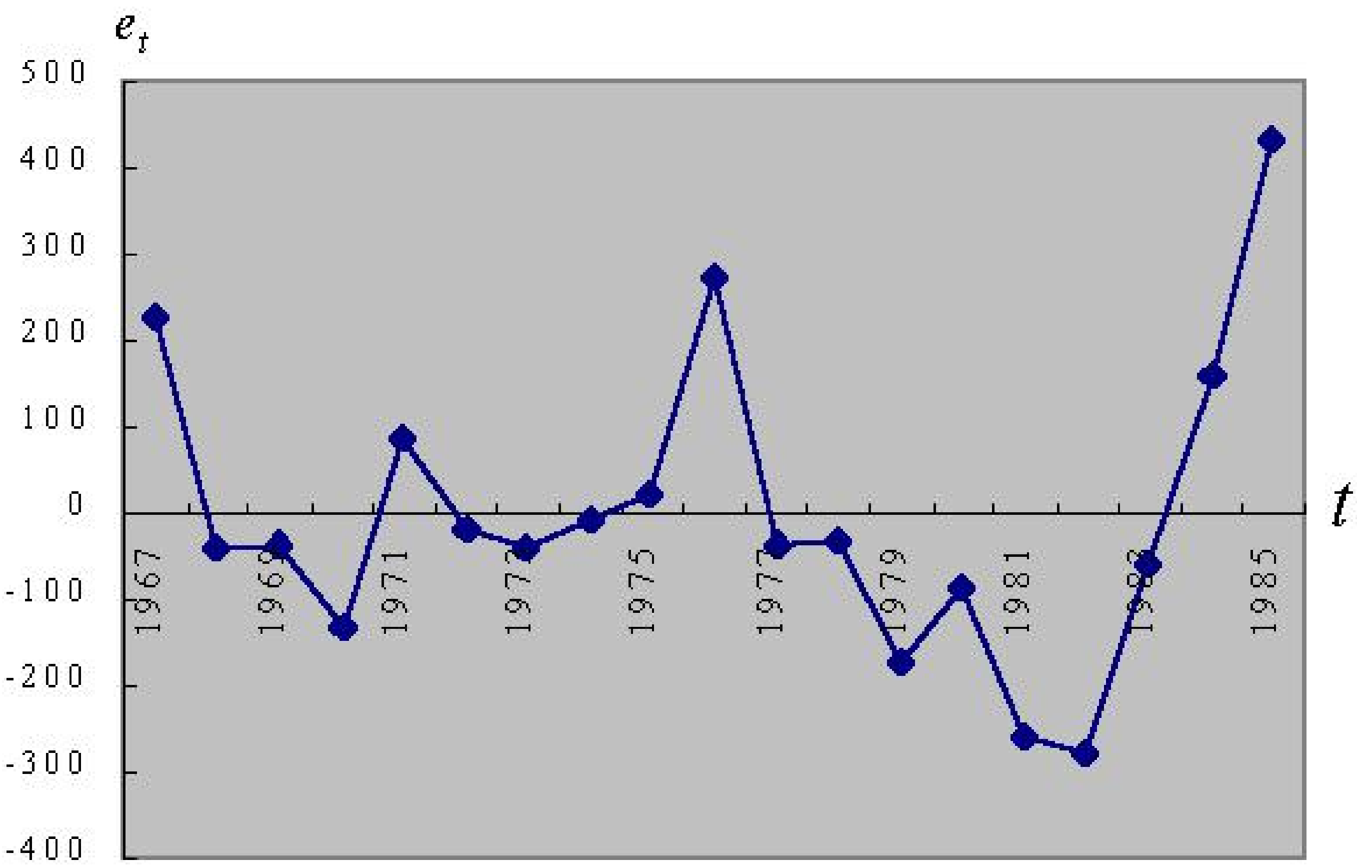
单位：万元

年份	出口 Y	国内生产总值 X	年份	出口 Y	国内生产总值 X
1967	4010	22418	1977	5628	29091
1968	3711	22308	1978	5736	29450
1969	4004	23319	1979	5946	30705
1970	4151	24180	1980	6501	32372
1971	4569	24893	1981	6549	33152
1972	4582	25310	1982	6705	33764
1973	4697	25799	1983	7104	34411
1974	4753	25886	1984	7609	35429
1975	5062	26868	1985	8100	36200
1976	5669	28134			

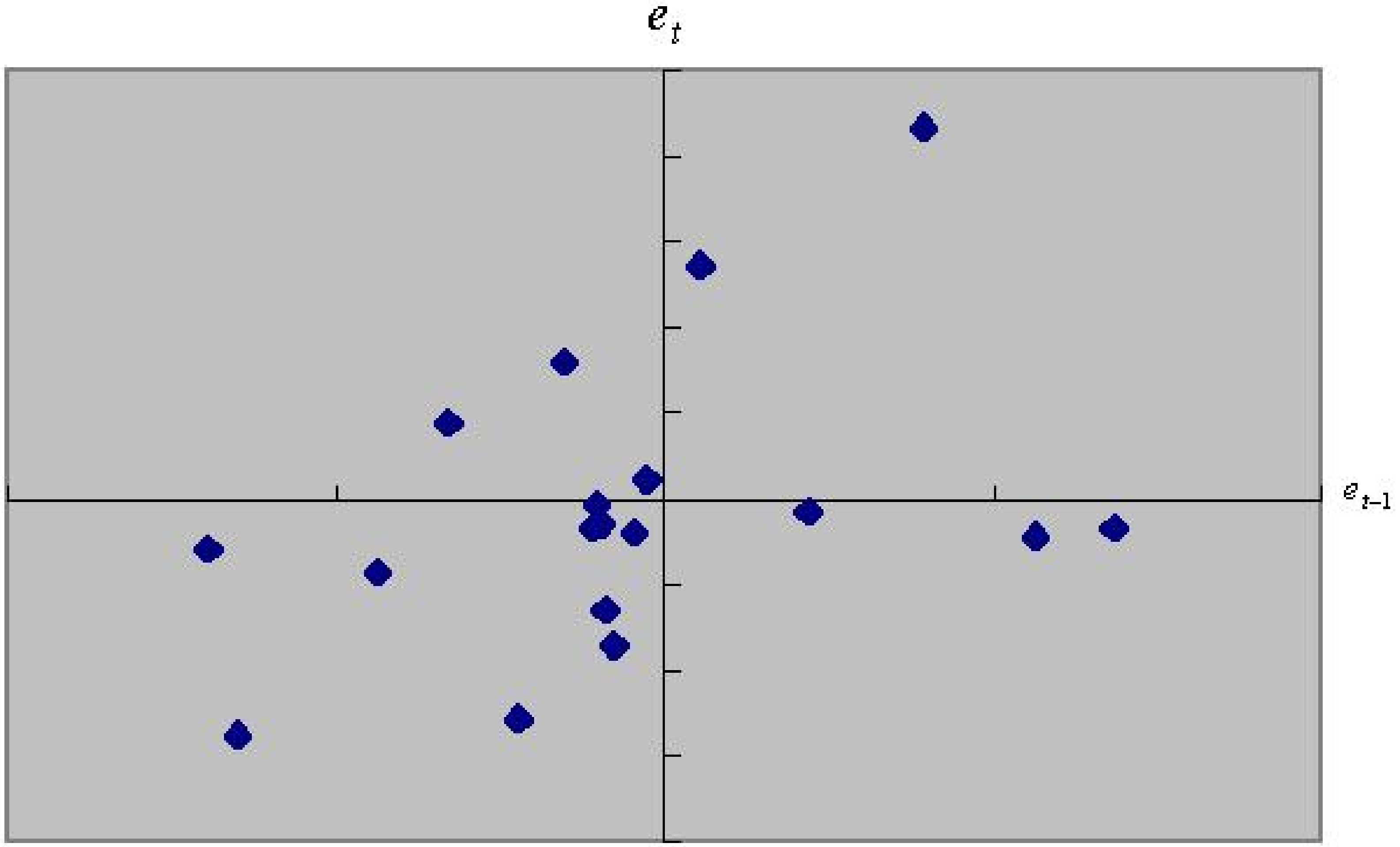
调用数据库E:\博士计量课程软件应用\export-gdp

1、检验

（1）图示法检验



e_t
500
400
300
200
100
0
-100
-200
-300
-400
t
1967
1969
1971
1973
1975
1977
1979
1981
1983
1985


e_t
e_{t-1}

（2）D.W.检验：

回归结果：

$$\hat{Y}_t = -2531.83 + 0.28X_t$$

(-9.34)　　(30.11)

r^2=0.9816,　　R^2=0.9805　　D.W.=0.9505

在5%在显著性水平下，n=19，k=2(包含常数项)，查表得d_L=1.18，d_U=1.40，

由于DW=0.9505<d_L，故存在正自相关。

2、自相关的处理

（1）一阶差分法

$$\Delta\hat{Y}_t = 0.3185\Delta X_t + \varepsilon_t$$

$$(6.8098)$$

R^2=0.4747,　D.W.=1.8623

由于DW>du=1.39(注：样本容量为18个)，已不存在自相关。

（2）广义差分法：

①采用杜宾两步法估计ρ

1）估计模型

$$Y_t = \beta_0^* + \rho Y_{t-1} + \beta_1^* X_t + \beta_2^* X_{t-1} + \varepsilon_{t-1}$$

得：

$$\hat{Y}_t = -1334.79 + 0.5939 Y_{t-1} + 0.3348 X_t - 0.2109 X_{t-1}$$

（-1.86）　(2.01)　(3.41)　(-1.53)

$r^2=0.9862$,　$R^2=0.9832$,　D.W.=1.6282

2)将 $\hat{\rho}$=0.5939 代入差分模型

$$Y_t - 0.5939Y_{t-1} = \beta_0^* + \beta_1^*(X_t - 0.5939X_{t-1}) + \varepsilon_t$$

OLS 法估计得：

$$Y_t - 0.5939Y_{t-1} = -1351.01 + 0.3083(X_t - 0.5939X_{t-1}) + \varepsilon_t$$

（-5.53）　(15.58)

r^2=0.9382,　R^2=0.9343,　D.W.=1.6570

由于DW>=1.39(注：样本容量为19-1=18个)，已不存在自相关。于是原模型估计式为：

$$\hat{Y}_t = -3326.79 + 0.3083X_t$$

调用数据库程序：multi-hetro-autoco