本章内容：

第四章 放宽基本假定的模型

- §4.1 多重共线性：回归元相关会怎么样?
- §4.2 异方差性：误差方差不是常数会怎么样?
- §4.3 自相关：误差项相关会怎么样?

本节内容：

§4.1 多重共线性：回归元相关会怎么样？

多重共线性的性质
——概念与内涵

- **多重共线性(multicollionearity)**
  - 含义：原意是指回归模型的解释变量之间存在“完全”或准确的线性关系。如：

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots\cdots \lambda_k X_k = 0$$

其中$\lambda_1, \lambda_2, \cdots, \lambda_k$为常数，但不同时为零。

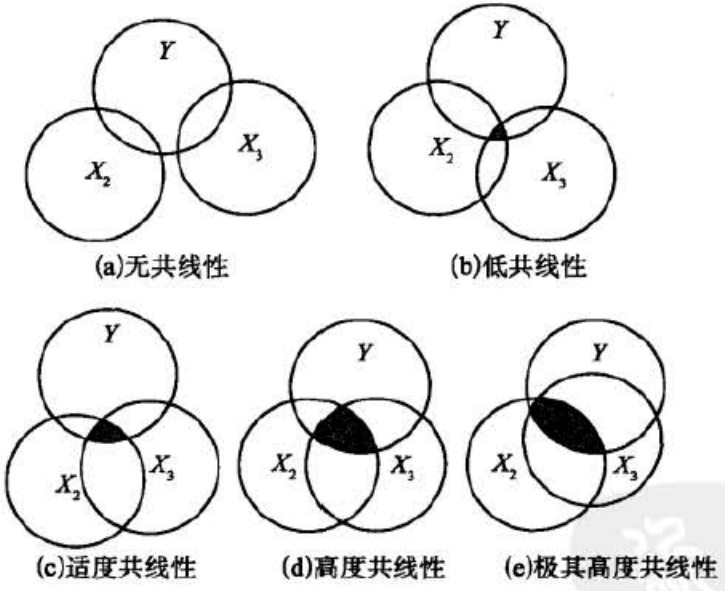
  - 现在多重共线性还用来泛指诸X变量之间有交互 相关，但又非完全相关：

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k + v_i = 0$$

其中v_i是随机误差项

多重共线性的性质
——一个例子

X_2	X_3	X_3^*
10	50	52
15	75	75
18	90	97
24	120	129
30	150	152



(a)无共线性　(b)低共线性

(c)适度共线性　(d)高度共线性　(e)极其高度共线性

多重共线性的性质
——概念与内涵

- **引起多重共线性的原因（Montgormery & Peck）：**
  - **数据采集所用的方法**：回归元限于一个范围
  - 模型或从中取样的总体受到约束：
    - 如做电力消费（Y）对收入（X2）和住房面积（X3）回归时，可能X2高的X3也大
  - **模型设定**：如在模型中加入多项式；数据范围小等
  - 一个过度决定的模型：回归元个数大于观测次数
    - 医学研究中可能只有少数病人
  - **相同的时间趋势**
    - 消费支出(Y)对收入、财富和人口的回归

完全多重共线性时的估计问题
——举例：三变量模型情况

- 在完全多重共线性的情况下，回归系数是不确定的，其标准误是无穷大。
  - 对于三变量回归：

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + e_i \quad \text{（式10.2.1）}$$

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad \text{（式7.4.7）}$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad \text{（式7.4.8）}$$

假定$X_{3i} = \lambda X_{2i}$，这里λ是一非零常数。代入（7.4.7）得：

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} = \frac{0}{0}$$

不完全多重共线性时的估计问题
——举例：三变量模型情况

- 假定不完全多重共线性的形式为：

$$x_{3i} = \lambda x_{2i} + v_i \quad \text{（式10.3.1）}$$

其中$\lambda \neq 0$，并且v_i是具有性质$\sum x_{2i} v_i = 0$的随机误差项

$$\hat{\beta}_2 = \frac{\sum (y_i x_{2i})(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum y_i x_{2i} + \sum y_i v_i)(\lambda \sum x_{2i}^2)}{\sum x_{2i}^2 (\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum x_{2i}^2)^2} \quad \text{（式10.3.2）}$$

- 如果v_i足够小，以至于接近于零，则（10.3.1）表示几乎完全共线性。这时我们又回到（ 10.2.2 ）的不定式情形。

多重共线性的理论后果

Micro-
numerosity

即使多重共线性是非常高的，如近似多重共线性，而OLS 估计量仍保持BLUE性质。

- 在近似多重共线性的情形下，OLS估计量仍然是无偏的。但是，无偏性只是一种多维样本（multisample）或重复抽样的性质。
- 共线性并不破坏最小方差性质。但是，这并不意味着，在任意给定的样本中，一个OLS估计量的方差一定是小的。仍然是BLUE
- 多重共线性**本质**上是一种**样本现象**。即使总体中X变量间不存在共线性，由于抽样方法或小样本问题也可能带来多重共线性问题。
  - 例如：

$$消费_i = \beta_1 + \beta_2 收入_i + \beta_3 财富_i + u_i$$

    - 一般地，收入高的人也更富裕，很难分开二者的各自影响。怎么办？

多重共线性的实际后果
——总览

- 更大的方差和协方差，估计精度大大下降。
  - 注：OLS估计仍是BLUE
- 置信区间变宽，系数的检验倾向于不显著！
  - 注：即更倾向接受原假设H_0，认为系数为零。
- 系数的t值倾向于统计上不显著，但R^2却会很高。

- OLS估计量及其标准误对数据的微小变化非常敏感。

多重共线性的实际后果
——更大的方差和协方差(1/2)

- 更大的方差和协方差

（式7.4.12）

$$\mathrm{var}(\hat{\beta}_2)=\frac{\sigma^2}{\sum x_{2i}^2(1-r_{23}^2)}$$

（式7.4.15）

$$\mathrm{var}(\hat{\beta}_3)=\frac{\sigma^2}{\sum x_{3i}^2(1-r_{23}^2)}$$

（式7.4.17）

$$\mathrm{cov}(\hat{\beta}_2,\hat{\beta}_3)=\frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum x_{2i}^2\sum x_{3i}^2}}$$

- 随着r_{23}增大，方差增大，协方差的绝对值也增大。增大的速度用方差膨胀因子(VIF, variance-inflating factor)衡量：

（式10.5.1）

$$VIF=\frac{1}{(1-r_{23}^2)}$$

多重共线性的实际后果
——更大的方差和协方差(2/2)

（式10.5.2）

$$\mathrm{var}(\hat{\beta}_2)=\frac{\sigma^2}{\sum x_{2i}^2}VIF$$

（式10.5.3）

$$\mathrm{var}(\hat{\beta}_3)=\frac{\sigma^2}{\sum x_{3i}^2}VIF$$

（式7.5.6）

$$\mathrm{var}(\hat{\beta}_j)=\frac{\sigma^2}{\sum x_j^2}\frac{1}{1-R_j^2}$$

（式10.5.4）

$$\mathrm{var}(\hat{\beta}_j)=\frac{\sigma^2}{\sum x_{ji}^2}VIF$$

R_j^2表示X_j对其余k - 2个回归元进行回归的R^2；
(注: 在k个变量的回归模型中有是k- 1个回归元。)

- 容许度(tolerance , TOL)：等于VIF 的倒数

（式10.5.5）

$$\mathrm{TOL}_j=\frac{1}{VIF_j}=1-R_j^2$$

当R_j = 1(即完全共线性)时，TOL=0;
当R_j =0 (即不存在共线性)时，TOL=1 。
由于VIF 和TOL 之间有密切关系，所以可以将它们互换使用。

多重共线性的实际后果
——更宽的置信区间

- 更宽的置信区间
  - 标准误增大，则有关总体参数的置信区间随之变大。
  - 在高度多重共线性的情形下，增加了接受错误假设的概率（第二类错误）

（Table 10-2）

表 10—2　增加共线性对 β_2 的 95%置信区间 "$\hat{\beta}_2 \pm 1.96se(\hat{\beta}_2)$" 的影响

r_{23} 值	β_2 的 95%置信区间
0.00	$\hat{\beta}_2 \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.50	$\hat{\beta}_2 \pm 1.96\sqrt{1.33}\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.95	$\hat{\beta}_2 \pm 1.96\sqrt{10.26}\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.995	$\hat{\beta}_2 \pm 1.96\sqrt{100}\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.999	$\hat{\beta}_2 \pm 1.96\sqrt{500}\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$

注：为方便起见，假定 σ^2 已知，因此可用正态分布，从而用 1.96 作为正态分布下的 95%置信因子。与各 r_{23} 值相对应的标准误取自表 10—1。

多重共线性的实际后果
——不显著”的t比率

- 不显著”的t比率：

在检验虚拟假设$H_0: \beta = 0$

$$t = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)}$$

- 将这一t值和临界t值相比较，即t检验高度共线性使估计的标准误增加很快，t值迅速变小。
- 因而，容易接受总体参数为零的虚拟假设

多重共线性的实际后果
——R^2值高而显著的t值少

- **R^2值高而显著的t值少：**
  - 在高度共线性情形中，有可能会发现一个或多个偏斜率系数基于t检验不是个别统计显著的，然而这时R^2却高达(比如说)0.9以上，从而根据F检验，可令人信服地拒绝H_0:

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

  - 但是，个别偏回归系数的t检验可能并不显著——这就是多重共线性的一个信号
  - 这里的真正问题在于估计量之间的协方差，而这些协方差是同回归元之间的相关性有关系的。

多重共线性的实际后果
——例子：OLS估计量及其标准误对数据的微小变化很敏感

（Table 10-3）

表 10—3　Y，X_2 和 X_3 的人为影响

Y	X_2	X_3
1	2	4
2	0	2
3	4	12
4	6	0
5	8	16

（式10.5.6）

$$\hat{Y}_i = 1.193\,9 + 0.446\,3X_{2i} + 0.003\,0X_{3i}$$
$$(0.773\,7)\quad (0.184\,8)\quad (0.085\,1)$$
$$t = (1.543\,1)\quad (2.415\,1)\quad (0.035\,8)$$
$$R^2 = 0.810\,1 \qquad r_{23} = 0.552\,3$$
$$\mathrm{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.008\,68 \qquad \mathrm{df} = 2$$

（Table 10-4）

表 10—4　Y，X_2 和 X_3 的人为数据

Y	X_2	X_3
1	2	4
2	0	2
3	4	0
4	6	12
5	8	16

（式10.5.7）

$$\hat{Y}_i = 1.210\,8 + 0.401\,4X_{2i} + 0.027\,0X_{3i}$$
$$(0.748\,0)\quad (0.272\,1)\quad (0.125\,2)$$
$$t = (1.618\,7)\quad (1.475\,2)\quad (0.215\,8)$$
$$R^2 = 0.814\,3 \qquad r_{23} = 0.828\,5$$
$$\mathrm{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.028\,2 \qquad \mathrm{df} = 2$$

多重共线性的实际后果
——讨论微数缺测性

- **讨论：**

  Goldberger认为微数缺测性（micronumerosity）是一个重要的问题

  - 微数缺测性：
    - 狭义是指样本大小n等于零的情形
    - 广义是指观测次数刚刚超过待估参数个数

一个假想的经济案例
——例子：消费支出(Y)与收入(X2)和财富(X3)的关系

（Table 10-5）

表 10—5　关于消费 Y，收入 X_2 和财富 X_3 的假想数据

Y	X_2	X_3
70	80	810
65	100	1 009
90	120	1 273
95	140	1 425
110	160	1 633
115	180	1 876
120	200	2 052
140	220	2 201
155	240	2 435
150	260	2 686

（式10.6.1）

$$Y_i = 24.774\,7 + 0.941\,5X_{2i} - 0.042\,4X_{3i}$$
$$(6.752\,5)\quad(0.822\,9)\quad(0.080\,7)$$
$$t = (3.669\,0)\quad(1.144\,2)\quad(-0.526\,1)$$
$$R^2 = 0.963\,5\quad \overline{R}^2 = 0.953\,1\quad df = 7$$

多重共线性诊断标准1：单个系数t值不显著，但多个系数的联合检验F值显著（高的R^2）。

（Table 10-6）

表 10—6　消费—收入—财富一例的 ANOVA 表

变异来源	SS	df	MSS
来自回归	8 565.554 1	2	4 282.777 0
来自残差	342.445 9	7	46.349 4

（式10.6.2）

$$F = \frac{4\,282.777\,0}{46.349\,4} = 92.401\,9$$

一个假想的经济案例
——例子：消费支出(Y)与收入(X2)和财富(X3)的关系

（Table 10-5）

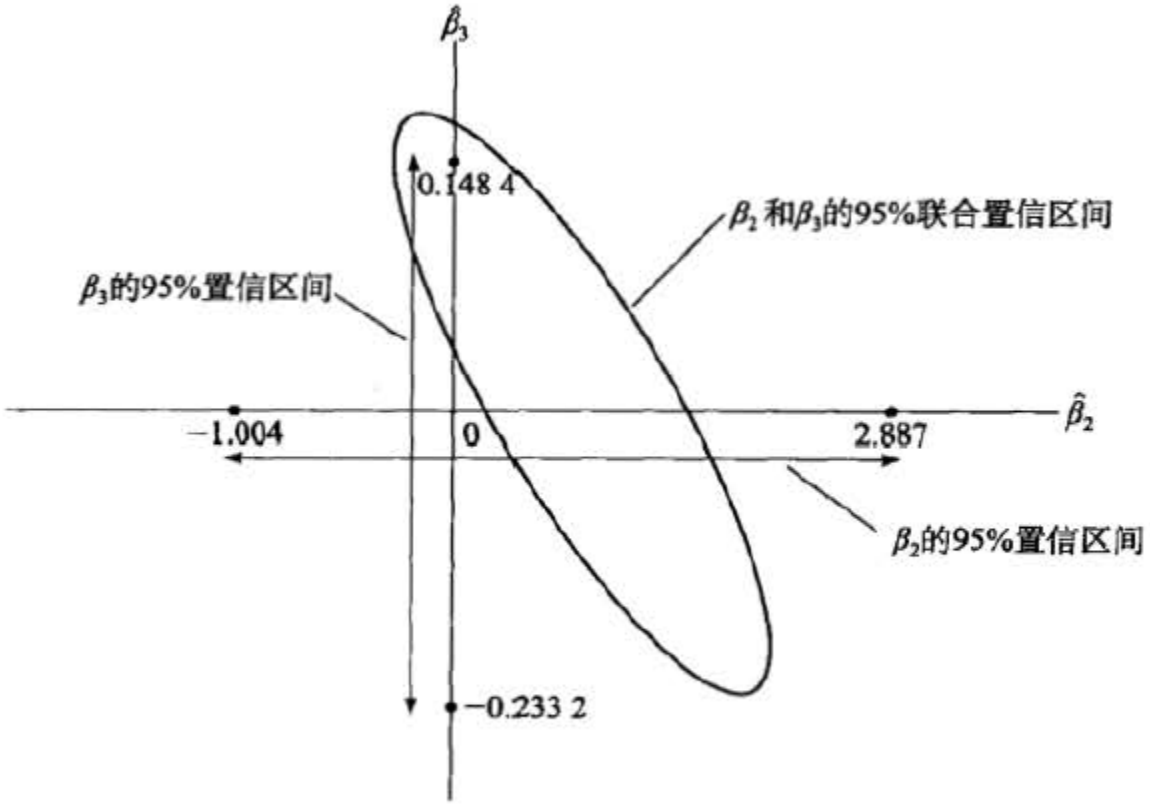


图 10—3　β_2 和 β_3 的个别置信区间与 β_2 和 β_3 的联合置信区间（椭圆形）

（式10.6.1）

（Table 10-6）

系数的T检验与F检验不一致：每个系数的置信区间均包含0（即无法拒绝H_0假设）；但是参数联合置信区间不包含0（即拒绝联合检验的H_0假设）

（式10.6.2）

一个假想的经济案例
——例子：消费支出(Y)与收入(X2)和财富(X3)的关系

- 如果我们做X3对X2的回归：

$$\hat{X}_{3i} = 7.545\,4 + 10.190\,9X_{2i} \quad \text{（式10.6.3）}$$
$$(29.475\,8) \quad (0.164\,3)$$
$$t = (0.256\,0) \quad (62.040\,5) \qquad R^2 = 0.997\,9$$

- 再做Y仅对X2的回归：

$$\hat{Y}_i = 24.454\,5 + 0.509\,1X_{2i} \quad \text{（式10.6.4）}$$
$$(6.413\,8) \quad (0.035\,7)$$
$$t = (3.812\,8) \quad (14.243\,2) \qquad R^2 = 0.962\,1$$

- 再做Y仅对X3的回归：

$$\hat{Y}_i = 24.411 + 0.049\,8X_{3i} \quad \text{（式10.6.5）}$$
$$(6.874) \quad (0.003\,7)$$
$$t = (3.551) \quad (13.29) \qquad R^2 = 0.956\,7$$

多重共线型诊断参考：Y分别对X2，X3的一元回归中，系数显著，但是Y同时对X2,X3的多元回归系数不显著。

一个实际例子：来自美国的案例(数据)

表 10—7　　1947—2000 年间美国的消费支出

年份	C	Yd	W	I
1947	976.4	1 035.2	5 166.815	−10.350 94
1948	998.1	1 090	5 280.757	−4.719 804
1949	1 025.3	1 095.6	5 607.351	1.044 063
1950	1 090.9	1 192.7	5 759.515	0.407 346
1951	1 107.1	1 227	6 086.056	−5.283 152
1952	1 142.4	1 266.8	6 243.864	−0.277 011
1953	1 197.2	1 327.5	6 355.613	0.561 137
1954	1 221.9	1 344	6 797.027	−0.138 476
1955	1 310.4	1 433.8	7 172.242	0.261 997
1997	5 423.9	5 854.5	32 664.07	3.12
1998	5 683.7	6 168.6	35 587.02	3.583 909
1999	5 968.4	6 320	39 591.26	3.245 271
2000	6 257.8	6 539.2	38 167.72	3.575 97

消费支出(C)

个人可支配收入(Yd)
财富(W)
利率(I)

一个实际例子：来自美国的案例(结果)

Dependent Variable: LOG (C)
Method: Least Squares
Sample: 1947-2000
Included observations: 54

	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.467711	0.042778	-10.93343	0.0000
LOG (YD)	0.804873	0.017498	45.99836	0.0000
LOG (WEALTH)	0.201270	0.017593	11.44060	0.0000
INTEREST	-0.002689	0.000762	-3.529265	0.0009

R-squared	0.999560	Mean dependent var.	7.826093
Adjusted R-squared	0.999533	S.D. dependent var.	0.552368
S.E. of regression	0.011934	Akaike info criterion	-5.947703
Sum squared resid.	0.007121	Schwarz criterion	-5.800371
Log likelihood	164.5880	Hannan-Quinn cariter.	-5.890883
F-statistic	37832.59	Durbin-Watson stat.	1.289219
Prob(F-statistic)	0.000000		

注：Log代表自然对数。

提问：如何解释回归分析结果？
思考：是否需要担心多重共线性的问题?

多重共线性的侦察
——一些经验规则

- 侦察多重共线性的经验规则：
  - 1. 观察主回归方程分析报告： R^2值高，F显著，但显著的t比率少（经典法则）
    - R^2值比较高，比如大于0. 8, F检验一般会拒绝所有偏回归系数同时为零，但是t检验却可能表明，没有或极少偏回归系数是统计上显著的

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i} + \hat{\beta}_6 X_{6i} + e_i$$

X_1 =价格指数; X_2 =GNP (以百万美元计); X_3 = 失业人数(以千人计) ; X_4=军队中的人数; X_5 =14 岁以上的非机构人口数; X_6 =年份:

一个引申的例子:朗利数据
——数据及模型

X_1 =价格指数; X_2 =GNP (以百万美元计); X_3 = 失业人数(以千人计) ; X_4=军队中的人数; X_5 =14 岁以上的非机构人口数; X_6 =年份:

	Y_EMPLOYEE	X1_CPI	X2_GNP	X3_JOBLESS__AMOUNT	X4_SODIER__AMOUN	X5_NONINSTITUTION...	X6_TIME
1947	60323	830	234289	2356	1590	107608	1
1948	61122	885	259426	2325	1456	108632	2
1949	60171	882	258054	3682	1616	109773	3
1950	61187	895	284599	3351	1650	110929	4
1951	63221	962	328975	2099	3099	112075	5
1952	63639	981	346999	1932	3594	113270	6
1953	64989	990	365385	1870	3547	115094	7
1954	63761	1000	363112	3578	3350	116219	8
1955	66019	1012	397469	2904	3048	117388	9
1956	67857	1046	419180	2822	2857	118734	10
1957	68169	1084	442769	2936	2798	120445	11
1958	66513	1108	444546	4681	2637	121950	12
1959	68655	1126	482704	3813	2552	123366	13
1960	69564	1142	502601	3931	2514	125368	14
1961	69331	1157	518173	4806	2572	127852	15
1962	70551	1169	554894	4007	2827	130081	16

一个引申的例子:朗利数据
——侦察：经济检验、t检验和F检验

Dependent Variable: Y_EMPLOYEE
Method: Least Squares
Date: 10/16/14 Time: 00:05
Sample: 1947 1962
Included observations: 16

存在多重共线性问题: 因为R^2 的值很高，但有几个变量不是统计显著的(X1、 X2 和X5)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	77270.12	22506.71	3.433204	0.0075
X1_CPI	1.506187	8.491493	0.177376	0.8631
X2_GNP	-0.035819	0.033491	-1.069516	0.3127
X3_JOBLESS__AMOUNT	-2.020230	0.488400	-4.136427	0.0025
X4_SODIER__AMOUN	-1.033227	0.214274	-4.821985	0.0009
X5_NONINSTITUTION_AMOUNT	-0.051104	0.226073	-0.226051	0.8262
X6_TIME	1829.151	455.4785	4.015890	0.0030

R-squared	0.995479	Mean dependent var	65317.00
Adjusted R-squared	0.992465	S.D. dependent var	3511.968
S.E. of regression	304.8541	Akaike info criterion	14.57718
Sum squared resid	836424.1	Schwarz criterion	14.91519
Log likelihood	-109.6174	Hannan-Quinn criter.	14.59449
F-statistic	330.2853	Durbin-Watson stat	2.559488
Prob(F-statistic)	0.000000		

多重共线性的侦察
——一些经验规则

- 侦察多重共线性的经验规则：
  - 2. 相关系数矩阵：
    - 参考准则：相关系数>0.9, 越多则越严重
    - 简单相关系数 vs 偏相关系数
      - 简单相关系数（零阶相关系数，参考标准）
      - 偏相关系数（高阶相关系数，精确标准）

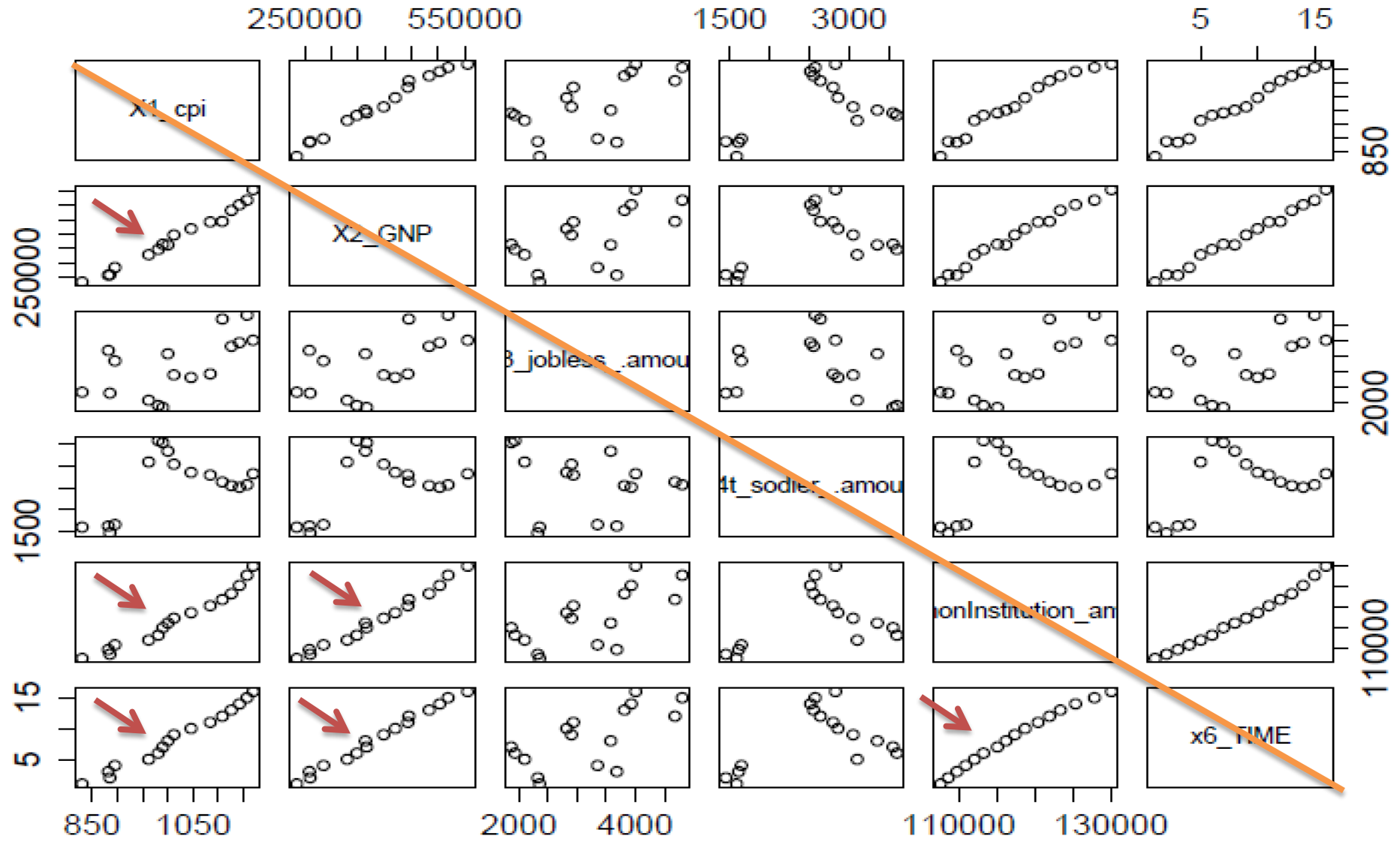
一个引申的例子:朗利数据
——侦察：相关系数矩阵

	X1_CPI	X2_GNP	X3_JOBLESS...	X4_SODIER...	X5_NONINSTITU...	X6_TIME
X1_CPI	1.000000	0.991589	0.620633	0.464744	0.979163	0.991149
X2_GNP	0.991589	1.000000	0.604261	0.446437	0.991090	0.995273
X3_JOBLESS...	0.620633	0.604261	1.000000	-0.177421	0.686552	0.668257
X4_SODIER_...	0.464744	0.446437	-0.177421	1.000000	0.364416	0.417245
X5_NONINSTI...	0.979163	0.991090	0.686552	0.364416	1.000000	0.993953
X6_TIME	0.991149	0.995273	0.668257	0.417245	0.993953	1.000000

多重共线性的侦察
——一些经验规则

- 侦察多重共线性的经验规则：
  - 3. 散点图矩阵
    - 判断规则：散点图表现为直线，这样的图越多则越严重

一个引申的例子:朗利数据
——侦察：散点图



多重共线性的侦察
——一些经验规则

（式10.7.3）

- **4. 辅助回归侦查法：**
  - 建立主回归：把Y对所有X进行回归，得到判定系数R^2

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i} + \hat{\beta}_6 X_{6i} + e_i$$

  - 建立辅助回归：把每一个X_i对其余X的回归，计算判定系数$\boldsymbol{R_i^2}$：

$$X_1 = \hat{\beta}_0 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i} + \hat{\beta}_6 X_{6i} + e_i$$
$$X_2 = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i} + \hat{\beta}_6 X_{6i} + e_i$$
$$X_3 = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i} + \hat{\beta}_6 X_{6i} + e_i$$
$$X_4 = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_5 X_{5i} + \hat{\beta}_6 X_{6i} + e_i$$
$$X_5 = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_6 X_{6i} + e_i$$
$$X_6 = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i} + e_i$$

克莱因的经验法则(Klein‘s rule of thumb): 仅当来自一个辅助回归的$\boldsymbol{R_i^2}$大于得自主回归中的总R^2值时，多重共线性才算是一个麻烦的问题。

一个引申的例子:朗利数据
——辅助回归1：X1_CPI to other

Dependent Variable: X1_CPI
Method: Least Squares
Date: 10/16/14 Time: 00:10
Sample: 1947 1962
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2044.583	533.3698	3.833331	0.0033
X2_GNP	0.002561	0.000948	2.700628	0.0223
X3_JOBLESS__AMOUNT	0.031922	0.015130	2.109831	0.0611
X4_SODIER__AMOUN	0.008802	0.007478	1.176973	0.2665
X5_NONINSTITUTION_AMOUNT	-0.017550	0.006331	-2.771998	0.0197
X6_TIME	-9.992189	16.66535	-0.599579	0.5621

R-squared	0.992622	Mean dependent var	1016.813
Adjusted R-squared	0.988933	S.D. dependent var	107.9155
S.E. of regression	11.35293	Akaike info criterion	7.976825
Sum squared resid	1288.890	Schwarz criterion	8.266546
Log likelihood	-57.81460	Hannan-Quinn criter.	7.991661
F-statistic	269.0649	Durbin-Watson stat	1.870344
Prob(F-statistic)	0.000000		

一个引申的例子:朗利数据
——辅助回归:2: X2_GNP to other

Dependent Variable: X2_GNP
Method: Least Squares
Date: 10/16/14 Time: 00:12
Sample: 1947 1962
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-480986.0	148413.8	-3.240845	0.0089
X1_CPI	164.6571	60.96993	2.700628	0.0223
X3_JOBLESS__AMOUNT	-13.78980	1.500185	-9.192068	0.0000
X4_SODIER__AMOUN	-2.998116	1.787322	-1.677435	0.1244
X5_NONINSTITUTION_AMOUNT	5.624360	1.180367	4.764926	0.0008
X6_TIME	10902.88	2570.756	4.241117	0.0017

R-squared	0.999441	Mean dependent var	387698.4
Adjusted R-squared	0.999161	S.D. dependent var	99394.94
S.E. of regression	2878.484	Akaike info criterion	19.04791
Sum squared resid	82856689	Schwarz criterion	19.33763
Log likelihood	-146.3833	Hannan-Quinn criter.	19.06275
F-statistic	3575.027	Durbin-Watson stat	1.665549
Prob(F-statistic)	0.000000		

一个引申的例子:朗利数据
——辅助回归3：X3_JOBLESS__AMOUNT

Dependent Variable: X3_JOBLESS__AMOUNT
Method: Least Squares
Date: 10/16/14 Time: 00:13
Sample: 1947 1962
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-28518.24	11446.89	-2.491354	0.0319
X1_CPI	9.649428	4.573555	2.109831	0.0611
X2_GNP	-0.064843	0.007054	-9.192068	0.0000
X4_SODIER__AMOUN	-0.271381	0.109011	-2.489492	0.0320
X5_NONINSTITUTION_AMOUNT	0.350986	0.095432	3.677879	0.0043
X6_TIME	768.5517	167.0507	4.600709	0.0010

R-squared	0.970255	Mean dependent var	3193.313
Adjusted R-squared	0.955382	S.D. dependent var	934.4642
S.E. of regression	197.3861	Akaike info criterion	13.68820
Sum squared resid	389612.8	Schwarz criterion	13.97792
Log likelihood	-103.5056	Hannan-Quinn criter.	13.70303
F-statistic	65.23778	Durbin-Watson stat	1.663054
Prob(F-statistic)	0.000000		

一个引申的例子:朗利数据
——辅助回归4：X4_SODIER__AMOUNT

Dependent Variable: X4_SODIER__AMOUN
Method: Least Squares
Date: 10/16/14 Time: 00:16
Sample: 1947 1962
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-11881.24	33002.42	-0.360011	0.7263
X1_CPI	13.82322	11.74472	1.176973	0.2665
X2_GNP	-0.073243	0.043664	-1.677435	0.1244
X3_JOBLESS__AMOUNT	-1.409910	0.566344	-2.489492	0.0320
X5_NONINSTITUTION_AMOUNT	0.199317	0.327633	0.608354	0.5565
X6_TIME	1167.779	561.6770	2.079094	0.0643

R-squared	0.721365	Mean dependent var	2606.688
Adjusted R-squared	0.582048	S.D. dependent var	695.9196
S.E. of regression	449.9064	Akaike info criterion	15.33595
Sum squared resid	2024158.	Schwarz criterion	15.62567
Log likelihood	-116.6876	Hannan-Quinn criter.	15.35079
F-statistic	5.177860	Durbin-Watson stat	1.369483
Prob(F-statistic)	0.013267		

一个引申的例子:朗利数据
——辅助回归5：X5_NONINSTITUTION_AMOUNT

Dependent Variable: X5_NONINSTITUTION_AMOUNT
Method: Least Squares
Date: 10/16/14 Time: 00:17
Sample: 1947 1962
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	95694.37	8682.033	11.02211	0.0000
X1_CPI	-24.75928	8.931925	-2.771998	0.0197
X2_GNP	0.123433	0.025905	4.764926	0.0008
X3_JOBLESS__AMOUNT	1.638107	0.445395	3.677879	0.0043
X4_SODIER__AMOUN	0.179055	0.294326	0.608354	0.5565
X6_TIME	-782.0409	587.1614	-1.331901	0.2125

R-squared	0.997495	Mean dependent var	117424.0
Adjusted R-squared	0.996242	S.D. dependent var	6956.102
S.E. of regression	426.4253	Akaike info criterion	15.22875
Sum squared resid	1818385.	Schwarz criterion	15.51847
Log likelihood	-115.8300	Hannan-Quinn criter.	15.24358
F-statistic	796.3020	Durbin-Watson stat	1.567875
Prob(F-statistic)	0.000000		

一个引申的例子:朗利数据
——侦察：辅助回归X6_TIME

Dependent Variable: X6_TIME
Method: Least Squares
Date: 10/16/14 Time: 00:18
Sample: 1947 1962
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.305049	15.40358	0.539164	0.6016
X1_CPI	-0.003473	0.005792	-0.599579	0.5621
X2_GNP	5.89E-05	1.39E-05	4.241117	0.0017
X3_JOBLESS__AMOUNT	0.000884	0.000192	4.600709	0.0010
X4_SODIER__AMOUN	0.000258	0.000124	2.079094	0.0643
X5_NONINSTITUTION_AMOUNT	-0.000193	0.000145	-1.331901	0.2125

R-squared	0.998682	Mean dependent var	8.500000
Adjusted R-squared	0.998024	S.D. dependent var	4.760952
S.E. of regression	0.211653	Akaike info criterion	0.012258
Sum squared resid	0.447969	Schwarz criterion	0.301979
Log likelihood	5.901938	Hannan-Quinn criter.	0.027094
F-statistic	1515.961	Durbin-Watson stat	1.297174
Prob(F-statistic)	0.000000		

一个引申的例子:朗利数据
——侦察:相关矩阵和辅助回归

X_1 =价格指数; X_2 =GNP (以百万美元计); X_3 = 失业人数(以千人计) ;
X_4=军队中的人数; X_5 =14 岁以上的非机构人口数; X_6 =年份:

表 10—10　　辅助回归的 R^2 值

因变量	R^2 值	容许度（TOL）=1－ R^2
X_1	0.992 6	0.007 4
X_2	0.999 4	0.000 6
X_3	0.970 2	0.029 8
X_4	0.721 3	0.278 7
X_5	0.997 0	0.003 0
X_6	0.998 6	0.001 4

$$R^2_{origin} = 0.995479$$

多重共线性的侦察
——一些经验规则

- **5. 方差膨胀因子(VIF)与容许度(TOL) ：**
  - **方差膨胀因子**(VIF)：

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \cdot \left(\frac{1}{1 - R_j^2} \right)$$

$$= \frac{\sigma^2}{\sum x_j^2} \cdot VIF_j$$

其中: $\hat{\beta}_j$是回归元 X_j 的偏回归系数，$R_j{}^2$是 X_j对其余（k-2）个回归元的辅助回归的 $\boldsymbol{R^2}$，VIF_j是方差膨胀因子

  - VIF_j越大，变量X_j与其余变量的共线性越严重。
  - **经验规则：** VIF大于10($R_j{}^2$大于0.90)，认为是高度共线性

  - **容许度**(TOL)：

$$TOL_j = (1 - R_j^2) = (1/VIF_j)$$

（式10.7.5）

  - **经验规则：** 无共线性，则 $\text{TOL}_j = 1$；
    完全共线性，则 $\text{TOL}_j = 0$

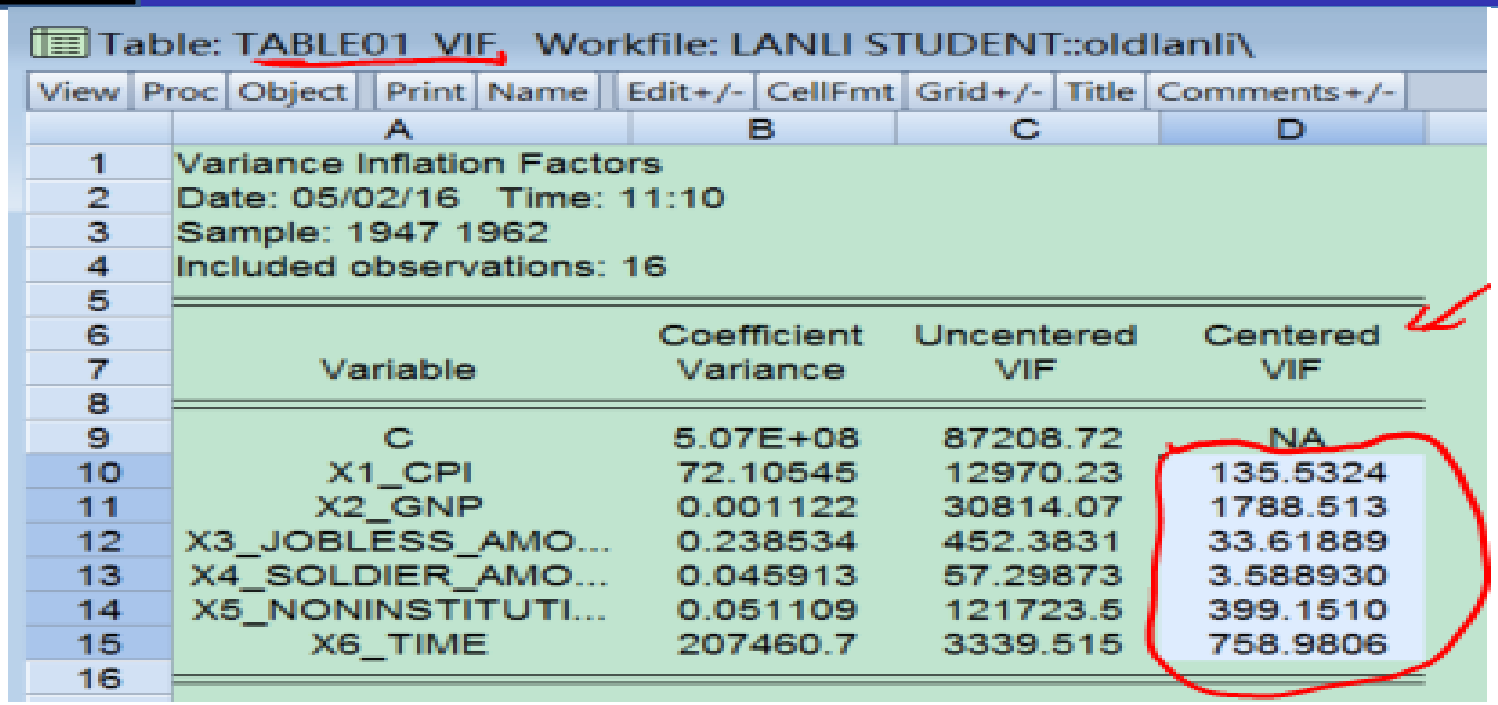


Table: TABLE01_VIF Workfile: LANLI STUDENT::oldlanli\

View | Proc | Object | Print | Name | Edit+/- | CellFmt | Grid+/- | Title | Comments+/-

	A	B	C	D
1	Variance Inflation Factors			
2	Date: 05/02/16 Time: 11:10			
3	Sample: 1947 1962			
4	Included observations: 16			
5				
6		Coefficient	Uncentered	Centered
7	Variable	Variance	VIF	VIF
8				
9	C	5.07E+08	87208.72	NA
10	X1_CPI	72.10545	12970.23	135.5324
11	X2_GNP	0.001122	30814.07	1788.513
12	X3_JOBLESS_AMO...	0.238534	452.3831	33.61889
13	X4_SOLDIER_AMO...	0.045913	57.29873	3.588930
14	X5_NONINSTITUTI...	0.051109	121723.5	399.1510
15	X6_TIME	207460.7	3339.515	758.9806
16				

	辅助回归(Eviews)	TOL公式	VIF公式	VIF(Eviews菜单)
变量	R^2_j	$TOL=1-R^2_j$	$VIF=1/(1-R^2_j)$	Centered VIF
X1	0.9926	0.0074	135.5324	135.5324
X2	0.9994	0.0006	1788.9088	1788.5135
X3	0.9703	0.0297	33.6191	33.6189
X4	0.7214	0.2786	3.5889	3.5889
X5	0.9971	0.0029	347.5843	399.1510
X6	0.9987	0.0013	758.7253	758.9806

多重共线性的侦察
——一些经验规则

- **6. 特征值、病态系数(*)**
  - a. 对所有X进行**主成分分析**，得到特征值(eigenvalues)
  - b. 计算**病态数k**:(condition number)

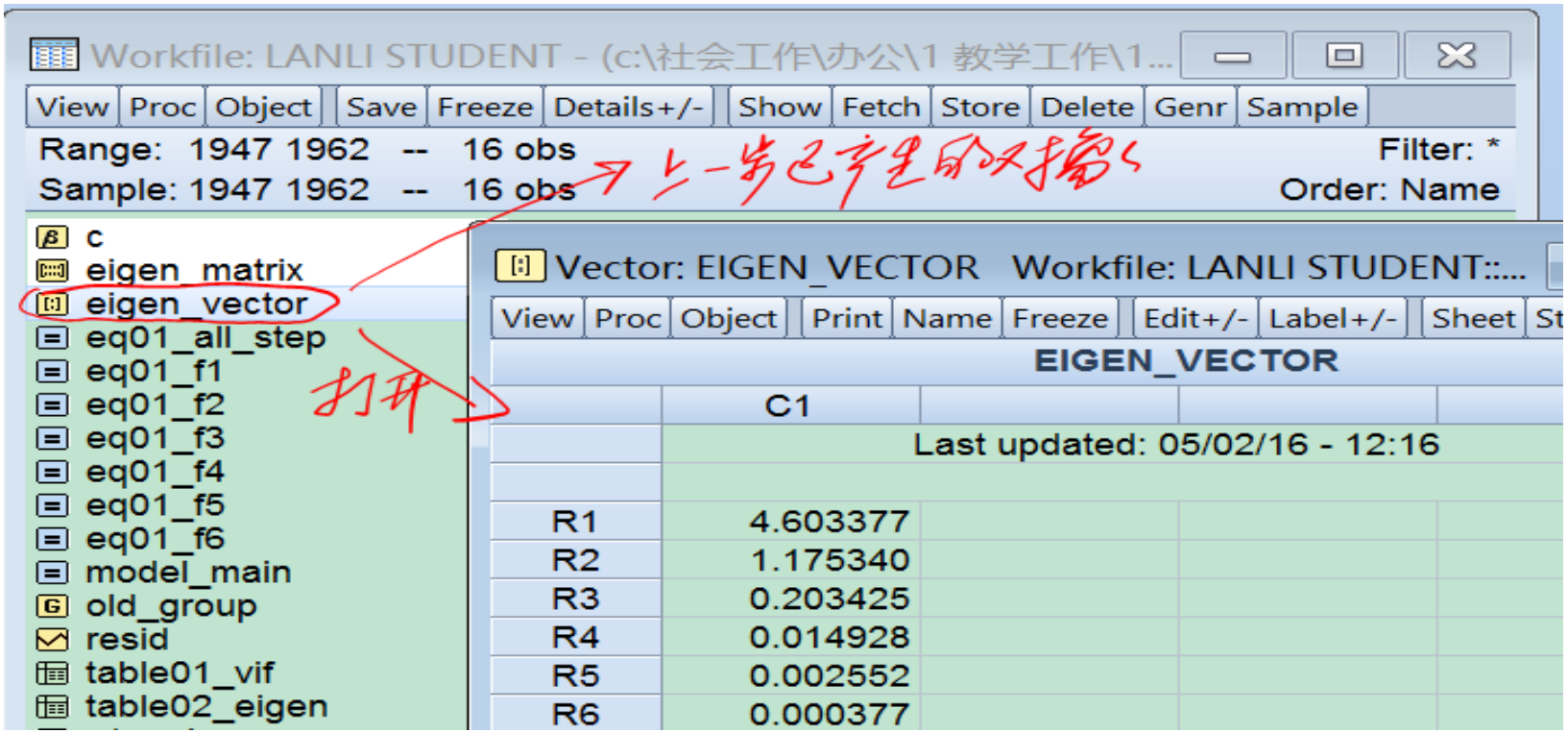
$$k=\frac{\text{最大特征值}}{\text{最小特征值}}$$

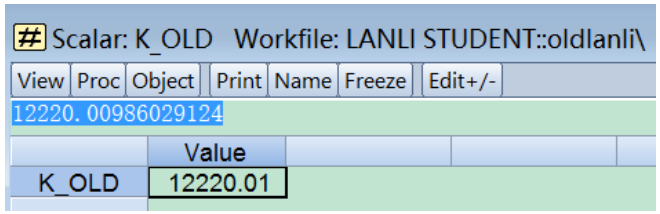
  - c. 计算**病态指数CI**(condition index,CI)

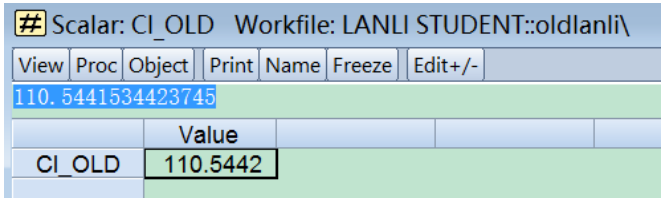
$$CI=\sqrt{k}$$

（式10.7.3）

- **经验准则：**k>1000，严重共线性；
  或CI>30，严重共线性


Workfile: LANLI STUDENT - (c:\社会工作\办公\1 教学工作\1...
View Proc Object Save Freeze Details+/- Show Fetch Store Delete Genr Sample
Range: 1947 1962 -- 16 obs
Sample: 1947 1962 -- 16 obs
Filter: *
Order: Name
c
eigen_matrix
eigen_vector
eq01_all_step
eq01_f1
eq01_f2
eq01_f3
eq01_f4
eq01_f5
eq01_f6
model_main
old_group
resid
table01_vif
table02_eigen
Vector: EIGEN_VECTOR Workfile: LANLI STUDENT::...
View Proc Object Print Name Freeze Edit+/- Label+/- Sheet
EIGEN_VECTOR
C1
Last updated: 05/02/16 - 12:16
R1 4.603377
R2 1.175340
R3 0.203425
R4 0.014928
R5 0.002552
R6 0.000377


Scalar: K_OLD Workfile: LANLI STUDENT::oldlanli\
View Proc Object Print Name Freeze Edit+/-
12220.00986029124
Value
K_OLD 12220.01


Scalar: CI_OLD Workfile: LANLI STUDENT::oldlanli\
View Proc Object Print Name Freeze Edit+/-
110.5441534423745
Value
CI_OLD 110.5442

补救措施：无为而治或经验程序
——措施1：利用先验信息

无为而治：多重共线性是上帝的意志，而不是OLS 或其他一般性统计方法的问题

多重共线性实质上是一个数据不足的问题（微数缺测性）

➢ 1. 利用先验信息（专家*）：

- 做消费Y对收入X2和X3的回归中

（式10.7.4）

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

- 如果事先得知 $\beta_3 = 0.10\beta_2$，则可以变为如下回归：

（式10.7.5）

$$Y_i = \beta_1 + \beta_2 X_{2i} + 0.10\beta_2 X_{3i} + u_i$$

$$= \beta_1 + \beta_2 X_i + u_i$$

- 怎样获得先验信息呢?它可以经验研究工作或者有关基础理论。例如，在柯布-道格拉斯生产函数(7.9. 1)中，如果人们预期规模报酬不变成立，则有$\beta_2+\beta_3 = 1$ 。如果劳动和资本之间存在共线性，这一变换就减轻或消除了共线性问题。

补救措施：无为而治或经验程序
——措施2：横截面与时间序列数据并用

- 2. 横截面与时间序列数据并用（不总是可行*）

$$\ln Y_t = \beta_1 + \beta_2 \ln P_t + \beta_3 \ln I_t + u_t$$

其中: Y—汽车需求，P—价格，I—收入，t—时间。
一般地，价格和收入之间存在共线性

- Tobin提出：
  - 利用横截面数据估计**收入弹性**，得到$\hat{\beta}_3$。(因为这些数据都产生于一定时点上，价格还不至于有多大变化,收入弹性的横截面估计将比较准确)
  - 再做如下回归，就能得到较准确的价格弹性

$$Y_t^* = \beta_1 + \beta_2 \ln P_t + u_t$$

其中，$Y^* = \ln Y - \hat{\beta}_3 \ln I$,即$Y^*$表示出去收入效应之后的$Y$值

补救措施：无为而治或经验程序
——措施3：剔除变量或设定误差

- 3. 剔除变量或设定误差（很常用！）

  - 面对严重的共线性，最简单的方法就是去掉某些变量。
  - 但剔除变量会导致设定误差。实际中需要权衡利弊。

- a. 简单删除法：根据经验和实际情况，酌情删除

- b. 逐步删除法：逐步回归法(Stepwise Least Square)

a.简单删除法：根据经验和实际情况，酌情删除

- **简单删除的依据：**
  - 改用真实GNP，不用名义GNP ：将名义GNP 除以价格指数X1 其次，
  - 留下X_5，去掉X_6：14 岁以上非机构人口数X_5随时间不断增长，它与时间变量X6高度相。
  - 去掉变量X_3：可能失业率是劳动力市场状况的一个更好的度量指标，但我们没有这方面的数据，而失业人数X_3也没有充分的理由包括进来。

Dependent Variable: Y_EMPLOYEE
Method: Least Squares
Date: 10/16/14 Time: 00:28
Sample: 1947 1962
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	65720.37	10624.81	6.185558	0.0000
X7_RGNP_NEW	97.36496	17.91552	5.434671	0.0002
X4_SODIER__AMOUN	-0.687966	0.322238	-2.134965	0.0541
X5_NONINSTITUTION_AMOUNT	-0.299537	0.141761	-2.112965	0.0562

R-squared	0.981404	Mean dependent var	65317.00
Adjusted R-squared	0.976755	S.D. dependent var	3511.968
S.E. of regression	535.4492	Akaike info criterion	15.61641
Sum squared resid	3440470.	Schwarz criterion	15.80955
Log likelihood	-120.9313	Hannan-Quinn criter.	15.62630
F-statistic	211.0972	Durbin-Watson stat	1.654069
Prob(F-statistic)	0.000000		

b.逐步删除法：逐步回归法(Stepwise Least Square)

- **逐步回归法的删除依据：**
  - 前向回归(forward stepwise);后向回归(backward stepwise)
  - p值<0.1(比较显著)；或p值<0.05(比较显著)；或p值<0.01(极其显著)
  - t值：2t法则

Equation: MODEL_STEP Workfile: LANLI STUDENT::ol...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y_EMPLOYEE
Method: Stepwise Regression
Date: 05/02/16 Time: 13:34
Sample: 1947 1962
Included observations: 16
No always included regressors
Number of search regressors: 7
Selection method: Stepwise backwards
Stopping criterion: p-value forwards/backwards = 0.05/0.05

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
X4_SOLDIER_AMOUNT	-1.014639	0.183734	-5.522333	0.0002
X3_JOBLESS_AMOU...	-2.088391	0.289970	-7.202082	0.0000
X2_GNP	-0.040190	0.016473	-2.439820	0.0328
C	74169.53	4251.585	17.44515	0.0000
X6_TIME	1887.410	382.7665	4.930969	0.0004

R-squared	0.995359	Mean dependent var	65317.00
Adjusted R-squared	0.993671	S.D. dependent var	3511.968
S.E. of regression	279.3955	Akaike info criterion	14.35344
Sum squared resid	858680.4	Schwarz criterion	14.59487
Log likelihood	-109.8275	Hannan-Quinn criter.	14.36580
F-statistic	589.7571	Durbin-Watson stat	2.603857
Prob(F-statistic)	0.000000		

Selection Summary

Removed X1_CPI
Removed X5_NONINSTITUTION_AMOUNT

*Note: p-values and subsequent tests do not account for stepwise selection.

补救措施：无为而治或经验程序
——措施4：变量代换(两种方式)

- 4. 变量代换法（**要有经济学基础！**）

v_t序列相关
减少了自由度

- 一阶差分法(first difference form)：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad （式10.8.1）$$

$$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} \quad （式10.8.2）$$

治疗比疾病更糟糕？

$$Y_t - Y_{t-1} = \beta_2 (X_{2t} - X_{2,t-1}) + \beta_3 (X_{3t} - X_{3,t-1}) + v_t \quad （式10.8.3）$$

即使X_2和X_3的水平值是高度相关的，其一阶差分回归可能不存在多重共线性

- 比率变换(ratio transformation)：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad （式10.8.4）$$

Y 为以真实价格表示的消费支出， X2 为GDP， X3为总人口。

  - 除以X3 得到以人均量为基础的模型：

$$\frac{Y_t}{X_{3t}} = \beta_1 + \beta_2 \frac{X_{2t}}{X_{3t}} + \beta_3 + \frac{u_t}{X_{3t}} \quad （式10.8.5）$$

治疗比疾病更糟糕？
异方差出现

  - 这样的变换可能会减少原有变量的共线性。（思考）

补救措施：无为而治或经验程序
——措施5：补充新数据

- 5. 补充新数据（有时候有用！）：

由于多重共线性是一个样本特性，故有可能在关于同样变量的另一样本中共线性没有第一个样本那么严重。

- 在三变量回归中，有：

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$

- 随着样本增加，$\sum x_{2i}^2$ 一般地说都会增加(为什么?)。
- 因此，对任何给定的r^2_{23}，$\hat{\beta}_2$的方差将减小，从而降低$\hat{\beta}_2$标准误，以使我们能更准确地估计$\hat{\beta}_2$。

10 次观测的消费Y 对收入X_2 和财富X_3的回归

$$\hat{Y}_i = 24.377 + 0.871\,6X_{2i} - 0.034\,9X_{3i} \quad \text{（式10.8.8）}$$

$$t = (3.875)\ (2.772\,6)\ (-1.159\,5) \qquad R^2 = 0.968\,2$$

40 次观测的消费Y 对收入X_2 和财富X_3的回归

$$\hat{Y}_i = 2.090\,7 + 0.729\,9X_{2i} + 0.060\,5X_{3i} \quad \text{（式10.8.9）}$$

$$t = (0.871\,3)\ (6.001\,4)\ (2.001\,4) \qquad R^2 = 0.967\,2$$

补救措施：无为而治或经验程序
——措施6/7：其他方法

- **6. 在多项式回归中降低共线性（*）：**
  - 多项式回归模型的一个特点是解释变量以不同的幂出现，从而容易导致多重共线性。
  - 处理办法：离差形式或正交多项式(orthogonal polynomials)方法

- **7. 拯救多重共线性的其他方法：**
  - 脊回归(ridge regression) 常被用来"解决"多重共线性问题。可惜这些技术都要利用矩阵代数才便于讨论。
  - 因子分析(factor analysis)
  - 主成分分析法(principal components)——Eviews
    - A.先根据主成分分析确定主成分个数(看累积解释百分比)
    - B.再用主成分得分（scoring）序列进行回归分析

主成分分析法(principal components) Eviews操作
——A.先根据主成分分析确定主成分个数

Principal Components Analysis
Date: 05/02/16 Time: 12:16
Sample: 1947 1962
Included observations: 16
Computed using: Ordinary correlations
Extracting 6 of 6 possible components

特征值

累积解释比例

Eigenvalues: (Sum = 6, Average = 1)

Number	Value	Difference	Proportion	Cumulative Value	Cumulative Proportion
1	4.603377	3.428037	0.7672	4.603377	0.7672
2	1.175340	0.971915	0.1959	5.778718	0.9631
3	0.203425	0.188497	0.0339	5.982143	0.9970
4	0.014928	0.012376	0.0025	5.997071	0.9995
5	0.002552	0.002175	0.0004	5.999623	0.9999
6	0.000377	---	0.0001	6.000000	1.0000

Eigenvectors (loadings):

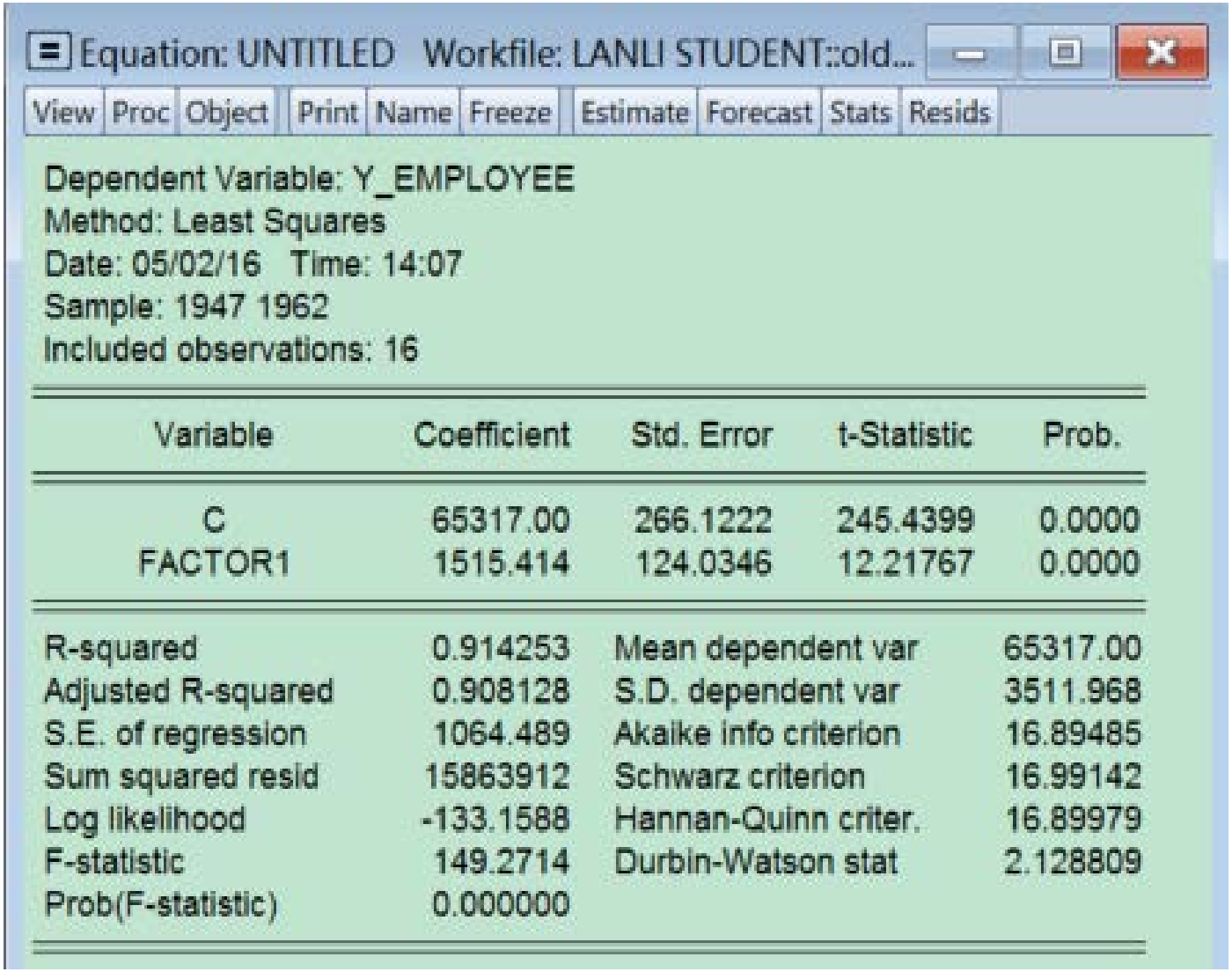
载荷

Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
X1_CPI	0.461835	0.057843	-0.149120	-0.792874	0.337938	-0.135187
X2_GNP	0.461504	0.053212	-0.277682	0.121621	-0.149573	0.818481
X3_JOBLESS_AM...	0.321317	-0.595514	0.728306	-0.007646	0.009232	0.107453
X4_SOLDIER_AM...	0.201510	0.798193	0.561608	0.077255	0.024252	0.017971
X5_NONINSTITU...	0.462279	-0.045544	-0.195985	0.589745	0.548578	-0.311571
X6_TIME	0.464940	0.000619	-0.128116	0.052287	-0.749543	-0.450409

Ordinary correlations: 相关系数

	X1_CPI	X2_GNP	X3_JOBLE...	X4_SOLDIE...	X5_NONIN...	X6_TIME
X1_CPI	1.000000					
X2_GNP	0.991589	1.000000				
X3_JOBLESS_AM...	0.620633	0.604261	1.000000			
X4_SOLDIER_AM...	0.464744	0.446437	-0.177421	1.000000		
X5_NONINSTITU...	0.979163	0.991090	0.686552	0.364416	1.000000	
X6_TIME	0.991149	0.995273	0.668257	0.417245	0.993953	1.000000

主成分分析法(principal components) Eviews操作
——B. 再用主成分得分序列进行回归分析



Equation: UNTITLED Workfile: LANLI STUDENT::old...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y_EMPLOYEE
Method: Least Squares
Date: 05/02/16 Time: 14:07
Sample: 1947 1962
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	65317.00	266.1222	245.4399	0.0000
FACTOR1	1515.414	124.0346	12.21767	0.0000

R-squared	0.914253	Mean dependent var	65317.00
Adjusted R-squared	0.908128	S.D. dependent var	3511.968
S.E. of regression	1064.489	Akaike info criterion	16.89485
Sum squared resid	15863912	Schwarz criterion	16.99142
Log likelihood	-133.1588	Hannan-Quinn criter.	16.89979
F-statistic	149.2714	Durbin-Watson stat	2.128809
Prob(F-statistic)	0.000000		

多重共线性一定是坏事吗?
——如果预测是唯一目的，就未必如此

- 如果回归分析的唯一目的是预测或预报，则多重共线性就不是一个严重的问题。因为R^2 值越高，预测越准。

本节内容：

§4.2 异方差性

- 4.2.1 异方差的性质
- 4.2.2 出现异方差性时的OLS估计
- 4.2.3 广义最小二乘法
- 4.2.4 出现异方差性时使用OLS的后果
- 4.2.5 异方差性的侦察
- 4.2.6 补救措施
- 4.2.7 总结性的例子
- 4.2.8 谨防对异方差性反应过度

异方差的概念与内涵
——回顾μ_i的方差协方差矩阵

- 一元线性回归模型：

（式11.1.1）

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

- 随机干扰项μ_i的方差协方差矩阵为：

（式11.1.2）

$$E(\mathbf{uu}') = \begin{bmatrix} E(u_1^2) & E(u_1u_2)\cdots E(u_1u_n) \\ E(u_2u_1) & E(u_2^2)\cdots E(u_2u_n) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ E(u_nu_1) & E(u_nu_2)\cdots E(u_n^2) \end{bmatrix}$$

异方差的概念与内涵
——CLRM假设：μ_i具有相同的方差

- 经典回归假设下残差μ_i具有相同的方差σ^2：

（式11.1.1）

$$E(\mathbf{uu}')=\begin{bmatrix} E(u_1^2) & E(u_1u_2)\cdots E(u_1u_n) \\ E(u_2u_1) & E(u_2^2)\cdots E(u_2u_n) \\ \cdots\cdots & \cdots\cdots \\ E(u_nu_1) & E(u_nu_2)\cdots E(u_n^2) \end{bmatrix}=\begin{bmatrix} \sigma^2 & 0 & 0\cdots 0 \\ 0 & \sigma^2 & 0\cdots 0 \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 0 & 0 & 0\cdots\sigma^2 \end{bmatrix}=\sigma^2\mathbf{I}$$

（式11.1.2）

- 同方差性(homoscedasticity)：即 μ_i 具有相同的方差，或者不随X变化而变化，用符号表示为：

$$E(u_i^2)=\sigma^2,\quad i=1,2,\cdots\cdots,n$$

异方差的概念与内涵
——违背CLRM假设：μ_i具有不相同的方差

- 违背经典回归假设下残差μ_i具有不相同的方差：

（式11.1.1）

（式11.1.2）

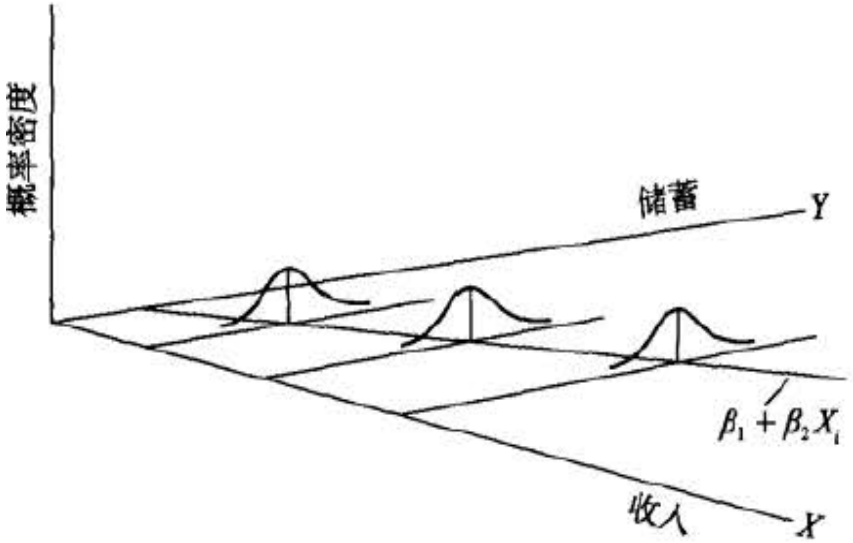
$$E(\mathbf{uu}')=\begin{bmatrix} E(u_1^2) & E(u_1u_2)\cdots E(u_1u_n) \\ E(u_2u_1) & E(u_2^2)\cdots E(u_2u_n) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ E(u_nu_1) & E(u_nu_2)\cdots E(u_n^2) \end{bmatrix}=\begin{bmatrix} \sigma_1^2 & 0 & 0\cdots 0 \\ 0 & \sigma_2^2 & 0\cdots 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ 0 & 0 & 0\cdots\sigma_n^2 \end{bmatrix}$$

- 异方差性（hetroscadasticity）：μ_i的条件方差（=Y_i的条件方差）随着X的变化而变化，用符号表示为：

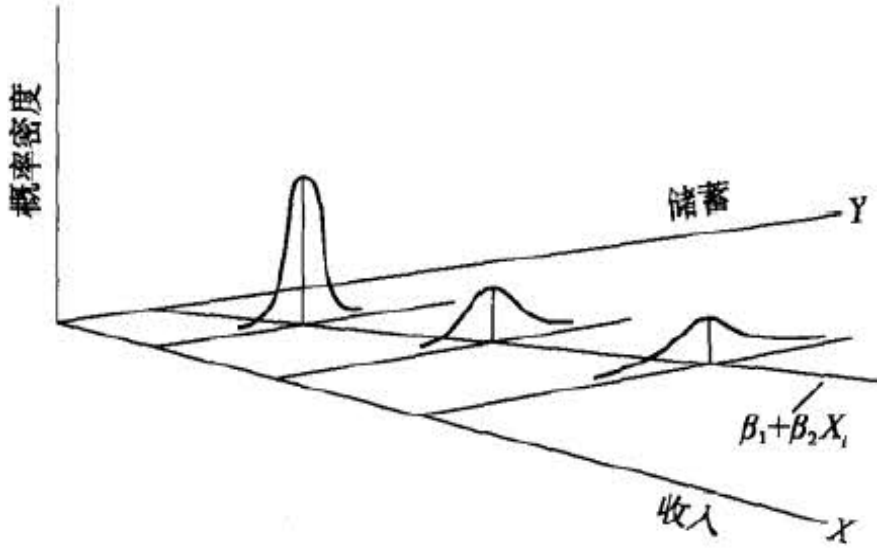
$$E(u_i^2)=\sigma_i^2,\quad i=1,2,\cdots\cdots,n$$

异方差的概念与内涵
——两种μ_i正态分布情形的比较

同方差



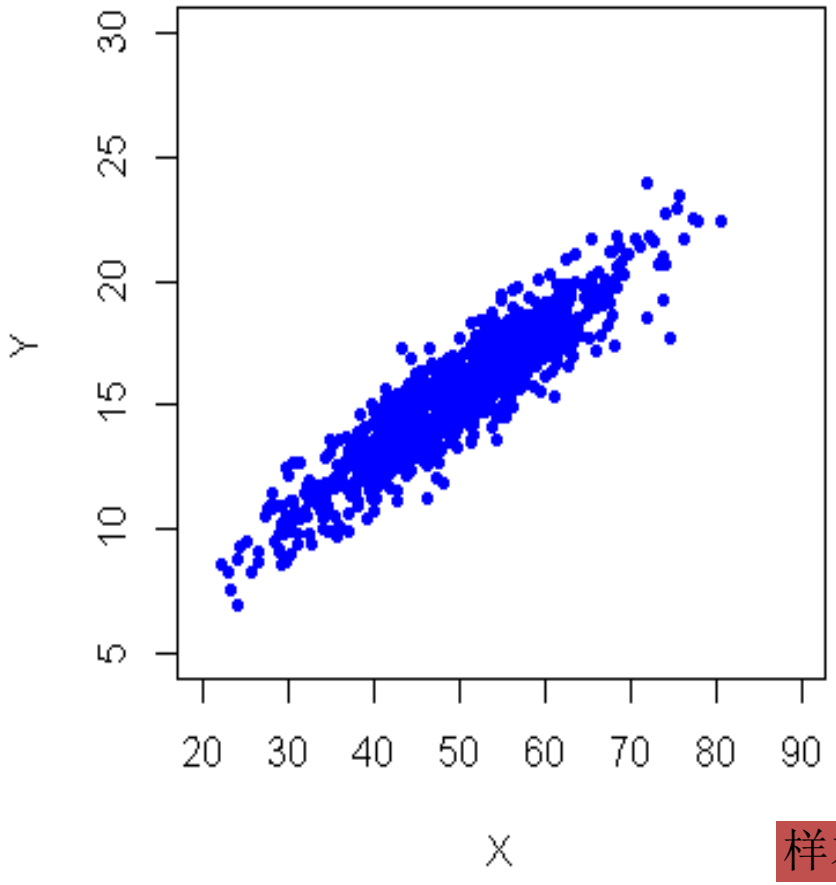
异方差



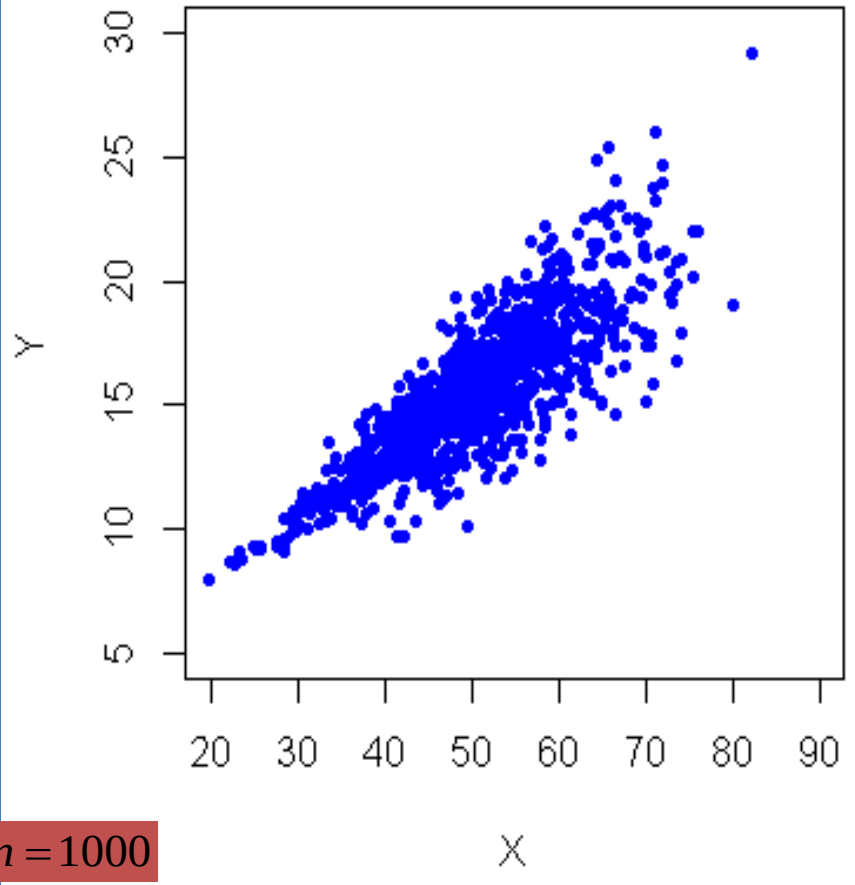
$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

异方差的概念与内涵
——数据直观比较

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$



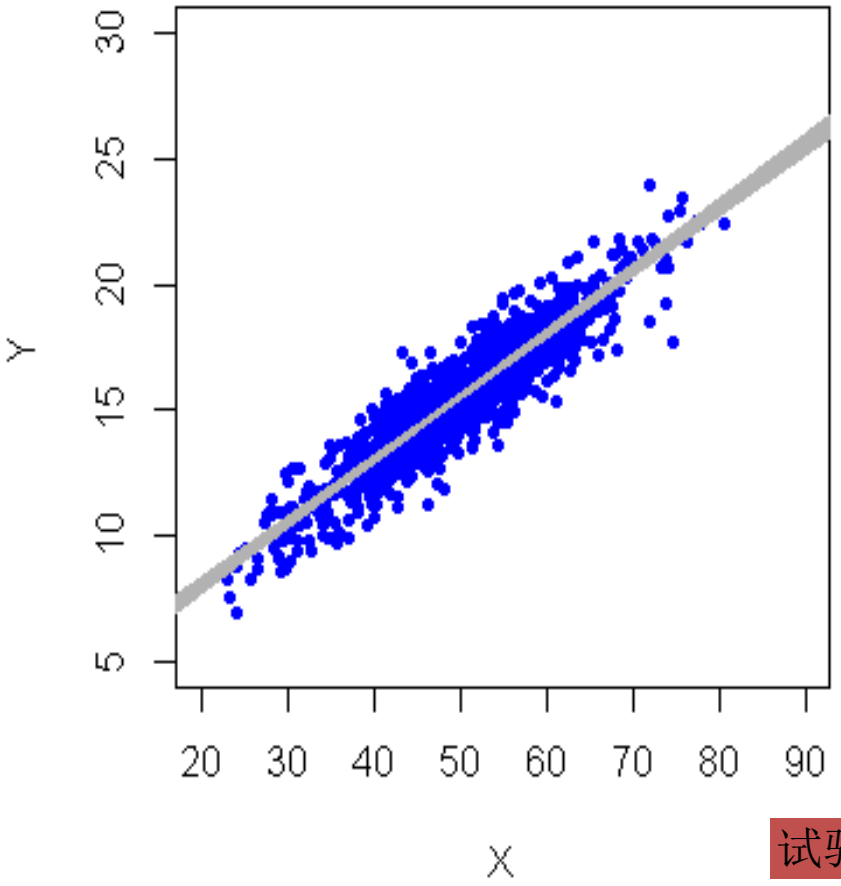
$$Y_i^* = \beta_1^* + \beta_2^* X_i + u_i^*$$



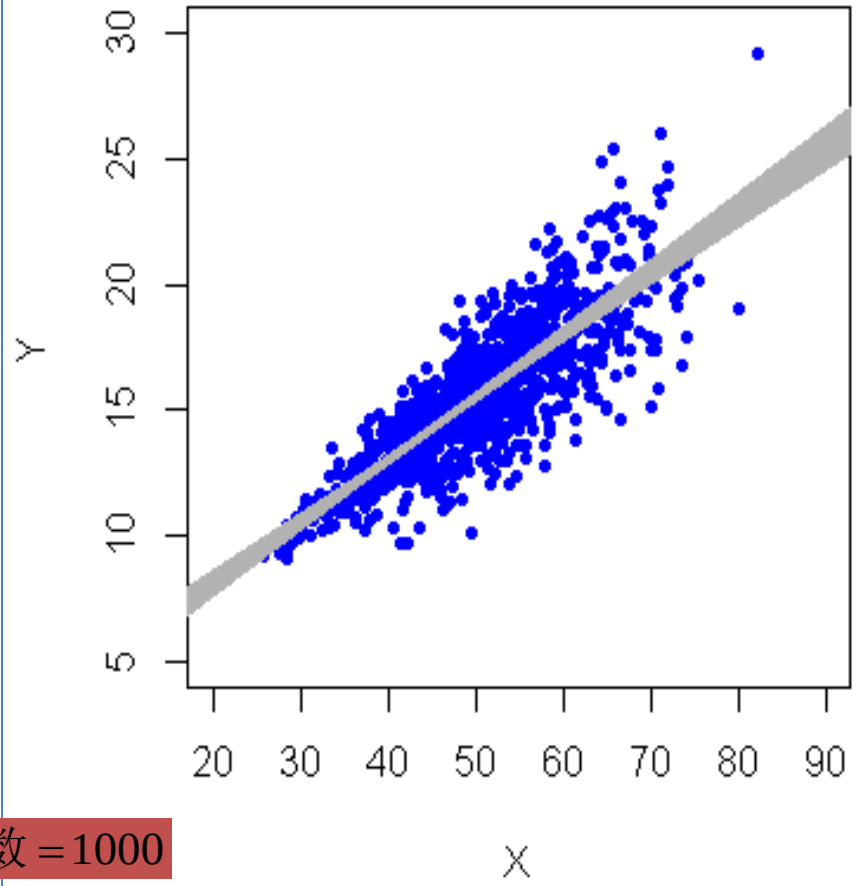
样本数$n = 1000$

异方差的概念与内涵
——OLS估计：直观比较1

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$



$$Y_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* X_i + e_i^*$$

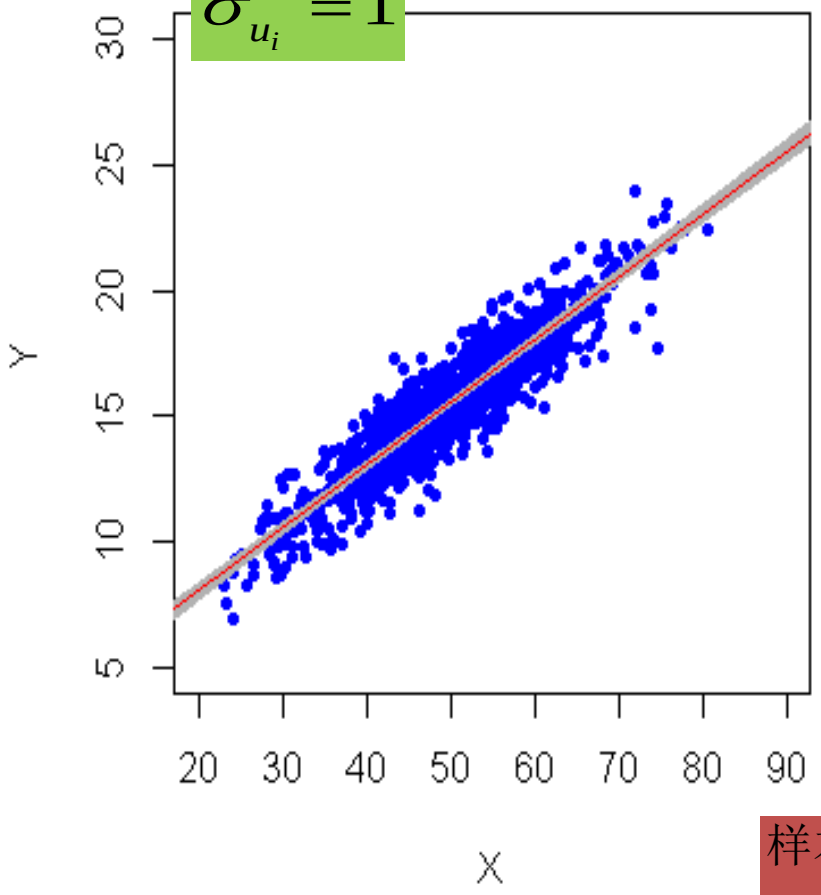


试验次数＝1000

异方差的概念与内涵
——OLS估计与真实值：直观比较2

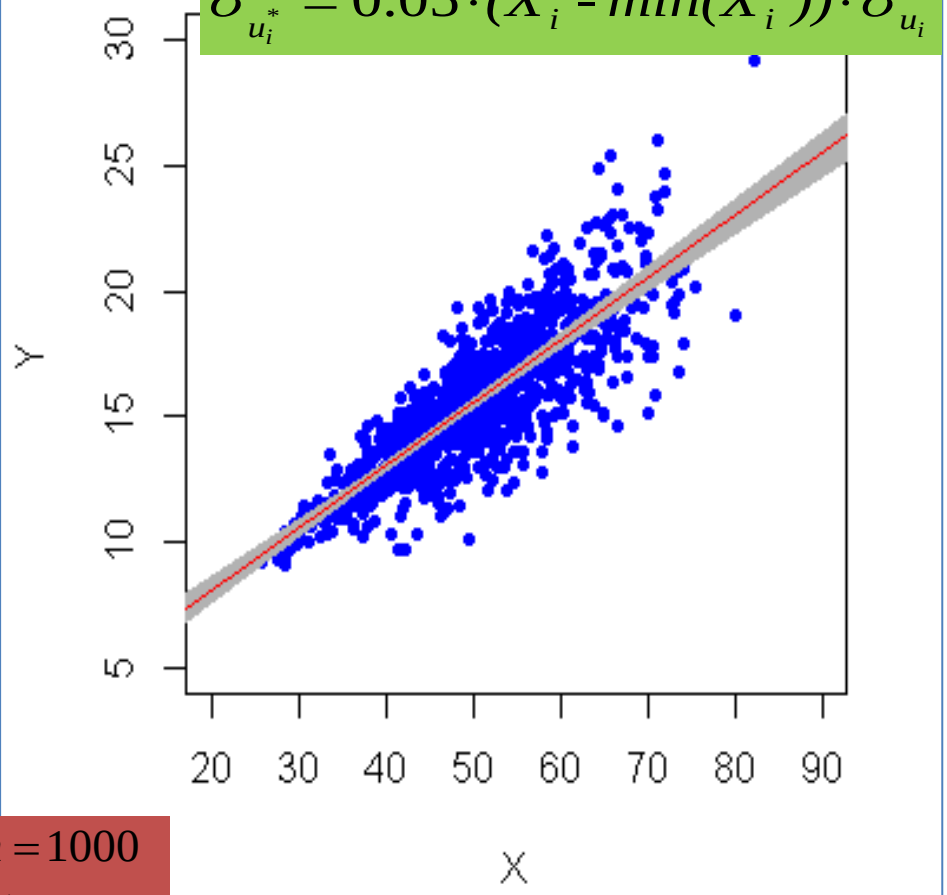
$$Y_i = 3 + 0.25X_i + u_i$$

$\sigma_{u_i} = 1$



$$Y_i^* = 3 + 0.25X_i + u_i^*$$

$\sigma_{u_i^*} = 0.05 \cdot (X_i - min(X_i)) \cdot \sigma_{u_i}$



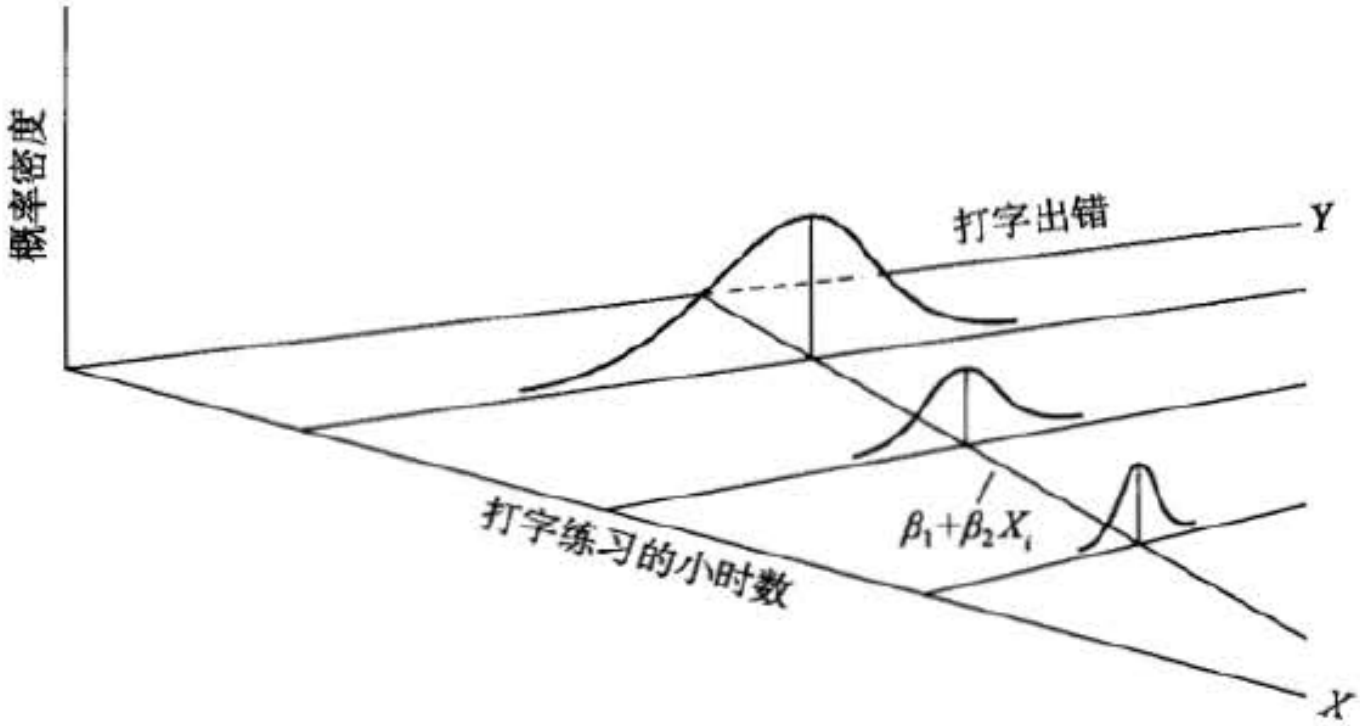
样本数$n = 1000$

试验次数 $= 1000$

异方差的性质
——异方差的来源

- **来源1：边错边改的学习模型（error-learning models)—即人们的行为误差随时间而减少。**
  - **实质：**人类行为固有模式特点
  - **解决：**无法改变，只能接受。

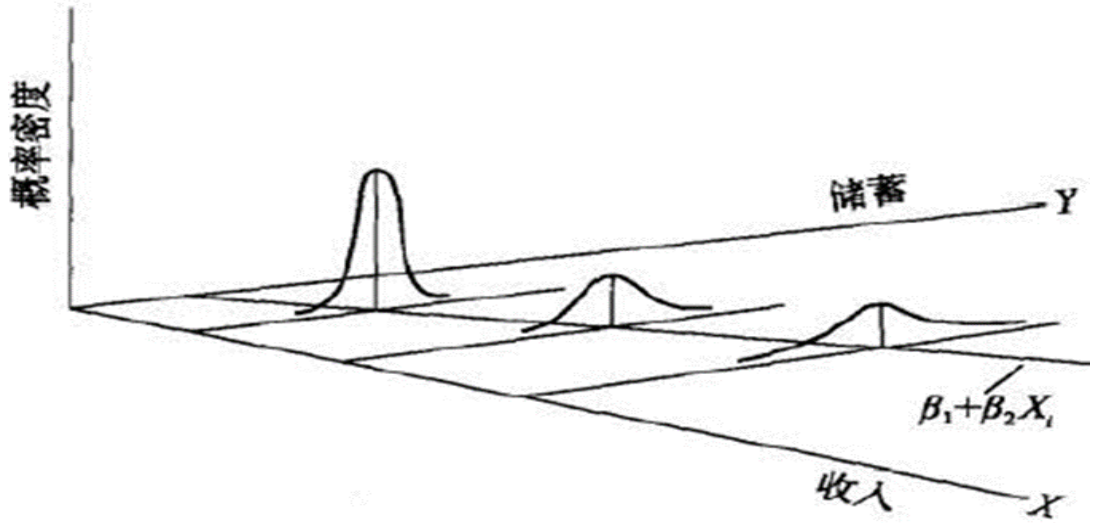
（Figure 11-3）



异方差的性质
——异方差的来源

- **来源2：随着收入增加，人们在支出和储蓄中有更大的灵活性（更大的方差）。**
  - **表现**：在做储蓄对收入的回归中，σ^2随着 收入增加而增大。
    - **经济含义**：富人比穷人具有更多**财务自由**，即富人可以选择较低或较高的储蓄率（σ^2更大）；而穷人只能选择较低储蓄率（σ^2更小）。（如何理解：“贫穷限制了我的想象力”）
  - **实质**：人类经济行为的固有特点。
  - **解决**：无法改变，只能接受。

（Figure 11-2）



异方差的性质
——异方差的来源

- **来源3：数据采集技术的差异**
  - **表现**：例如，有成熟的数据处理设备的银行，在为客户提供的月度或季度报表中，相对于没有这种设备的银行，会出现更少的差错。
  - **实质**：不同数据采集技术引发的差异。
  - **解决**：改进数据采集技术，σ^2可能减小

（Figure 11-2）

异方差的性质
——异方差的来源

- **来源4：异常值（outliers）导致的异方差性**
  - **表现**：样本中包含了不同于其他个体的特殊个体。
  - **实质**：抽样不当引发的异方差。
  - **解决**：剔除异常个体, 或者增加样本容量

（Figure 11-4）

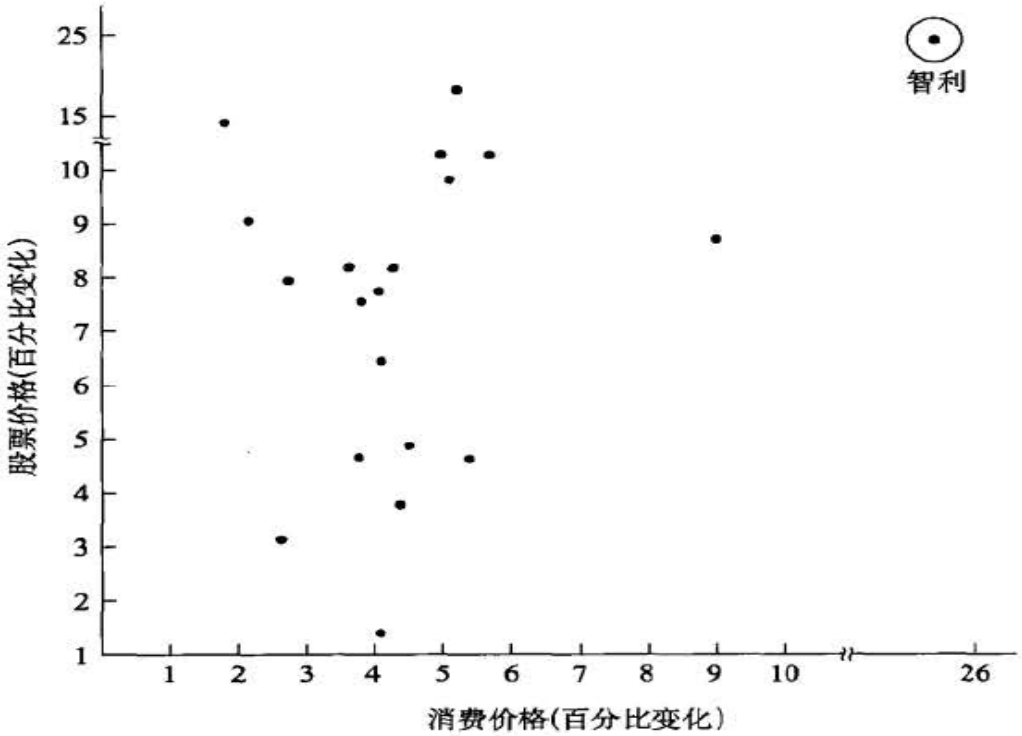


图 11—4　股票价格与消费价格的关系

异常观测是来自于与产生其余观测值的总体不同的另一个总体。

小样本(本例20个国家)，问题会更大。

异方差的性质
——异方差的来源

- **来源5：模型的设定偏误引发的异方差**
  - 情形1：忽略了重要的解释变量。
    - 表现：例如，做商品的需求量对价格的回归时，没有将互补品或替代品的价格包括进来，会引起异方差问题
    - 后果：样本方差大于总体的方差（高估）
    - 解决：将重要解释变量列入模型，然后进行估计。
  - 情形2：包含了无关解释变量
    - 表现：建模时，包含了不重要的变量。
    - 后果：样本方差小于总体的方差（低估）；自由度损失
    - 解决：识别并剔除无关的解释变量。

（Figure 11-5）

异方差的性质
——异方差的来源

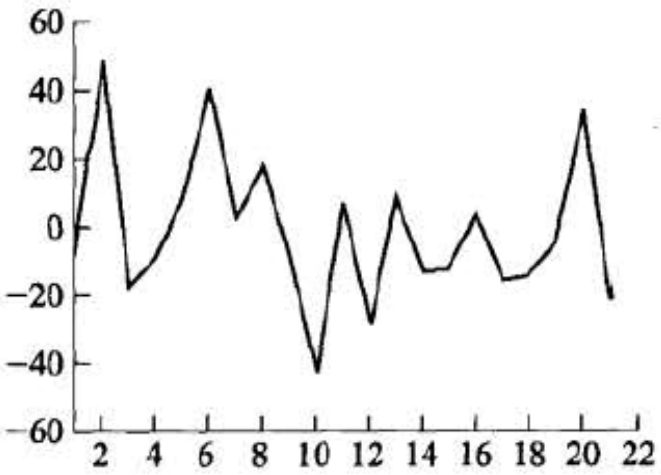
- **来源6：解释变量的分布偏态（skewness）**
  - **表现**：例如，收入和财富的分布不均；
  - **实质**：经济行为的本身固有的特点。
  - **解决**：数据变换，比如取对数
    - **经验法则：**
      - 分布偏态的原始数据，取对数后，分布会趋近于对称分布。
      - 对称分布的原始数据，取对数后，分布会偏离对称分布。

（Figure 11-5）

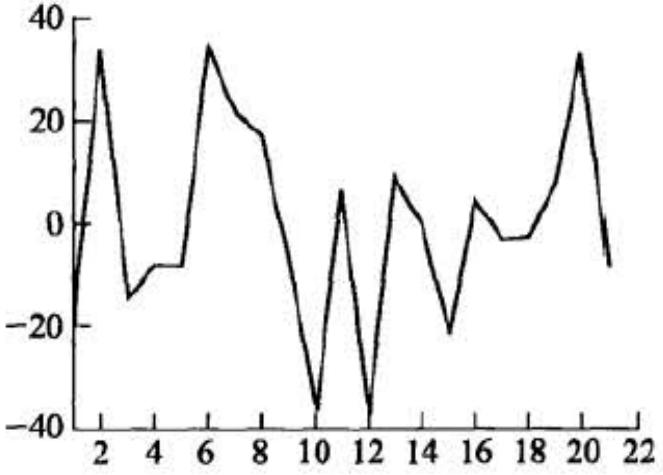
异方差的性质
——异方差的来源

- **来源7：错误的数据形式和函数形式导致的异方差。**
  - **表现：**
    - 不正确的数据形式：比率，差分，对数
      - 比如：对称分布取对数，导致更大的异方差。
    - 函数形式：线性、对数线性、U型、倒U型
  - **实质：**数据处理，或函数形式选择不当。
  - **解决：**谨慎选择数据和函数形式。

(Figure 11-5)



(a) 广告印象 Y 对广告支出 X 进行回归的残差



(b) 广告印象 Y 对 X 和 X^2 进行回归的残差

异方差的性质
——异方差的来源

- **来源8**：截面数据中个体的差异导致的异方差。（异方差的主要来源）

比如：表11-1表明，平均而言，大企业比小企业支付更多的工资。

（Table 11-1）

表 11—1　　1958 年按厂商职工人数划分的非耐用品制造行业的人均薪金　　（单位：美元）

行业	就业人数（平均职工人数）								
	1～4	5～9	10～19	20～49	50～99	100～249	250～499	500～999	1 000～2 499
食品干果	2 994	3 295	3 565	3 907	4 189	4 486	4 676	4 968	5 342
烟草产品	1 721	2 057	3 336	3 320	2 980	2 848	3 072	2 969	3 822
纺织品	3 600	3 657	3 674	3 437	3 340	3 334	3 225	3 163	3 168
器皿用具	3 494	3 787	3 533	3 215	3 030	2 834	2 750	2 967	3 453
纸张类	3 498	3 847	3 913	4 135	4 445	4 885	5 132	5 342	5 326
印刷与出版	3 611	4 206	4 695	5 083	5 301	5 269	5 182	5 395	5 552
化工产品	3 875	4 660	4 930	5 005	5 114	5 248	5 630	5 870	5 876
石油与煤炭	4 616	5 181	5 317	5 337	5 421	5 710	6 316	6 455	6 347
橡胶与塑料	3 538	3 984	4 014	4 287	4 221	4 539	4 721	4 905	5 481
皮革与皮革制品	3 016	3 196	3 149	3 317	3 414	3 254	3 177	3 346	4 067
平均薪金	3 396	3 787	4 013	4 104	4 146	4 241	4 388	4 538	4 843
薪金标准差	742.2	851.4	727.8	805.06	929.9	1 080.6	1 243.2	1 307.7	1 110.7
平均生产力	9 355	8 584	7 962	8 275	8 389	9 418	9 795	10 281	11 750

（Figure 11-6）

异方差的性质
——异方差的来源

行业类别		企业规模（员工人数）								
industry	id	1-4	5-9	10-19	20-49	50-99	100-249	250-499	500-999	1000-2499
Food …	1	2994	3295	3565	3907	4189	4486	4676	4968	5342
Tobacco Products	2	1721	2057	3336	3320	2980	2848	3072	2969	3822
Textile Mill Products	3	3600	3657	3674	3437	3340	3334	3225	3163	3168
Apparel …	4	3494	3787	3533	3215	3030	2834	2750	2967	3453
Paper …	5	3498	3847	3913	4135	4445	4885	5132	5342	5326
Printing …	6	3611	4206	4695	5083	5301	5269	5182	5395	5552
Chemicals …	7	3875	4660	4930	5005	5114	5248	5630	5870	5876
Petroleum …	8	4616	5181	5317	5337	5421	5710	6316	6455	6347
Rubber …	9	3538	3984	4014	4287	4221	4539	4721	4905	5481
Leather …	10	3016	3196	3149	3317	3414	3254	3177	3346	4067

企业规模（员工人数）	1-4	5-9	10-19	20-49	50-99	100-249	250-499	500-999	1000-2499
薪金标准差	743.7	851.4	727.8	805.06	929.9	1080.6	1243.2	1307.7	1112.5
平均薪金(Y)	3396	3787	4013	4104	4146	4241	4388	4538	4843
平均生产力(X)	9355	8584	7962	8275	8389	9418	9795	10281	11750

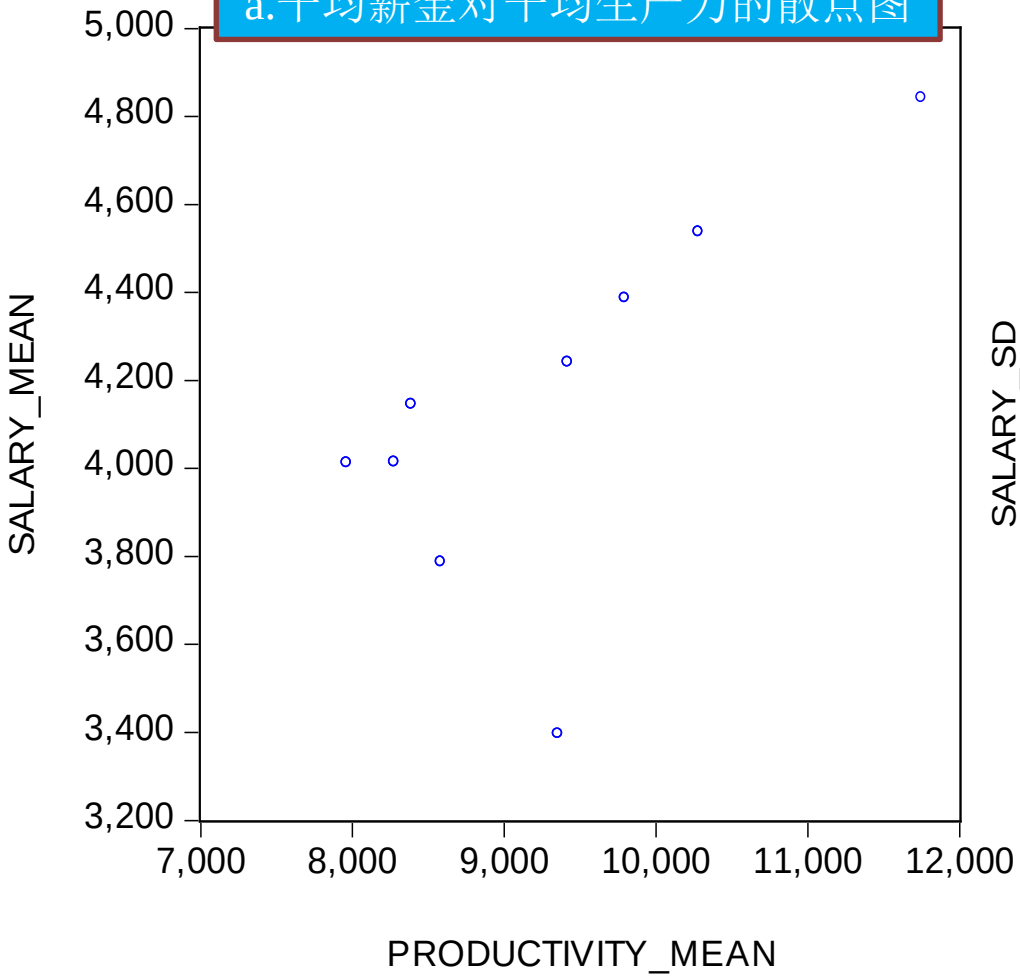
异方差的性质
——异方差的来源

企业规模 scale	薪金标准差 Salary sd	平均薪金Y Salary mean	平均生产力X Productivity mean
1-4	743.7	3396	9355
5-9	851.4	3787	8584
10-19	727.8	4013	7962
20-49	805.06	4104	8275
50-99	929.9	4146	8389
100-249	1080.6	4241	9418
250-499	1243.2	4388	9795
500-999	1307.7	4538	10281
1000-2499	1112.5	4843	11750

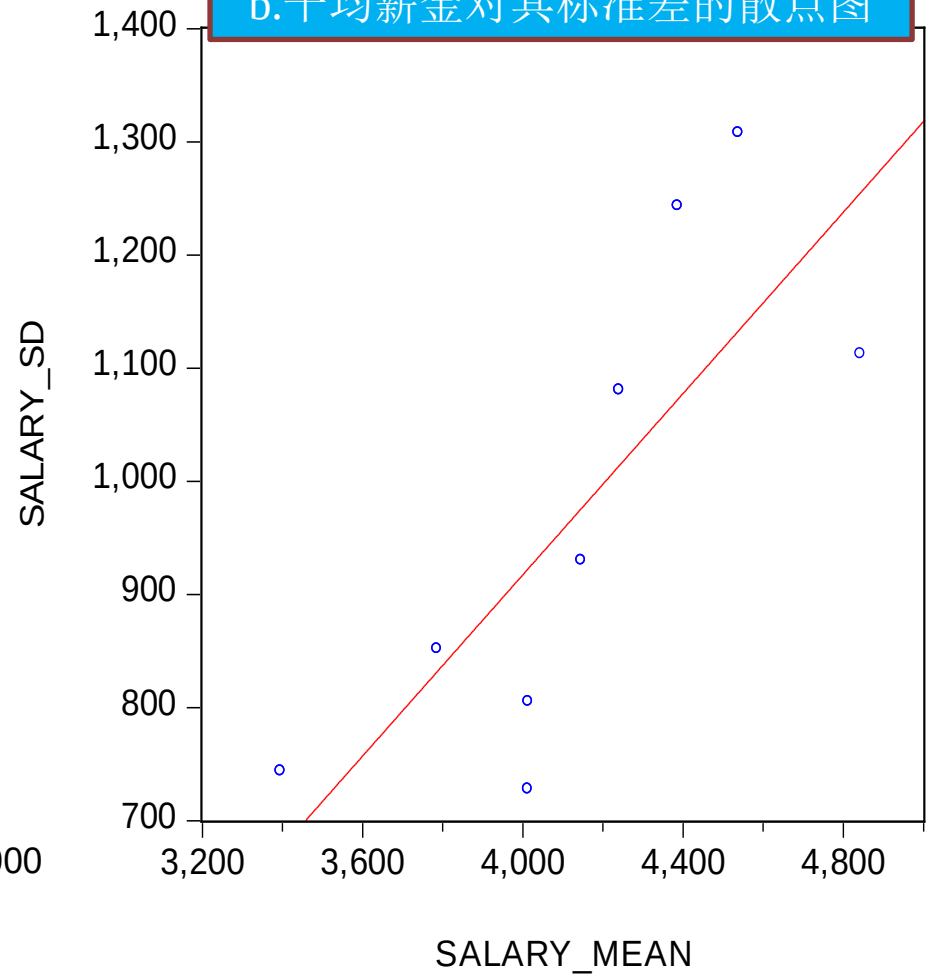
平均而言，大的厂家比小的厂家支付更多的工资。

异方差的性质
——异方差的来源

a.平均薪金对平均生产力的散点图



b.平均薪金对其标准差的散点图



$$Salary_i = \beta_1 + \beta_2 Productivity_i + u_i$$

Salary(Y)的方差并不是处处相等，而呈现明显增大，这有违CLRM假设。

出现异方差性时的OLS估计
——OLS估计量不再是BLUE

- 对于双变量模型(一元回归模型)：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

β_2的*OLS*估计量：

（式11.2.1）

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

异方差假定下，它的方差为：

（式11.2.2）

$$\text{var}(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2}$$

而同方差假定下，它的方差为：

（式11.2.3）

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$$

出现异方差性时的OLS估计
——OLS估计量不再是BLUE

- 在**经典模型**的各种假定下（包含同方差性假定），OLS估计量是BLUE（高斯-马尔可夫定理）。

- **异方差**情形下，异方差对OLS估计带来的后果：

  其他假定不变，同方差性假定不成立时，通常OLS估计量不再是BLUE。

  具体而言：OLS估计量仍然是**线性的**和**无偏的**（因这两条性质都与方差无关），但是，**不再是“最优的”或“有效的”**，即OLS估计量不再具有最小方差

- 那么，在出现异方差性时，什么才是BLUE呢?

$$\hat{\beta}_2, \hat{\beta}_3, \cdots, \hat{\beta}_n$$

（思考）

广义最小二乘法（GLS）
——认识GLS

表 11—1　1958 年按厂商职工人数划分的非耐用品制造行业的人均薪金　（单位：美元）

行业	就业人数（平均职工人数）								
	1～4	5～9	10～19	20～49	50～99	100～249	250～499	500～999	1 000～2 499
食品干果	2 994	3 295	3 565	3 907	4 189	4 486	4 676	4 968	5 342
烟草产品	1 721	2 057	3 336	3 320	2 980	2 848	3 072	2 969	3 822
纺织品	3 600	3 657	3 674	3 437	3 340	3 334	3 225	3 163	3 168
器皿用具	3 494	3 787	3 533	3 215	3 030	2 834	2 750	2 967	3 453
纸张类	3 498	3 847	3 913	4 135	4 445	4 885	5 132	5 342	5 326
印刷与出版	3 611	4 206	4 695	5 083	5 301	5 269	5 182	5 395	5 552
化工产品	3 875	4 660	4 930	5 005	5 114	5 248	5 630	5 870	5 876
石油与煤炭	4 616	5 181	5 317	5 337	5 421	5 710	6 316	6 455	6 347
橡胶与塑料	3 538	3 984	4 014	4 287	4 221	4 539	4 721	4 905	5 481
皮革与皮革制品	3 016	3 196	3 149	3 317	3 414	3 254	3 177	3 346	4 067
平均薪金	3 396	3 787	4 013	4 104	4 146	4 241	4 388	4 538	4 843
薪金标准差	742.2	851.4	727.8	805.06	929.9	1 080.6	1 243.2	1 307.7	1 110.7
平均生产力	9 355	8 584	7 962	8 275	8 389	9 418	9 795	10 281	11 750

异方差处理的思路:对来自变异较大的总体的观测值赋予较小的权重，而对来自较小变异的总体现测值赋予较大的权重。

- GLS比OLS更多地利用了样本数据所提供的信息!

广义最小二乘法（GLS）
——异方差情形下μ_i方差协方差矩阵

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$E(\mathbf{uu}') = \begin{bmatrix} E(u_1^2) & E(u_1u_2)\cdots E(u_1u_n) \\ E(u_2u_1) & E(u_2^2)\cdots E(u_2u_n) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ E(u_nu_1) & E(u_nu_2)\cdots E(u_n^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & 0 & 0\cdots 0 \\ 0 & \sigma_2^2 & 0\cdots 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ 0 & 0 & 0\cdots\sigma_n^2 \end{bmatrix} = \begin{bmatrix} w_1\sigma^2 & 0 & 0\cdots 0 \\ 0 & w_2\sigma^2 & 0\cdots 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ 0 & 0 & 0\cdots w_n\sigma^2 \end{bmatrix}$$

广义最小二乘法（GLS）
——认识GLS

- 对于一元回归模型：

（式11.3.1）
$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

（式11.3.2）
$$Y_i = \beta_1 X_{0i} + \beta_2 X_i + u_i$$

令X_{0i}=1

（式11.3.3）
$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{X_{0i}}{\sigma_i}\right) + \beta_2 \left(\frac{X_i}{\sigma_i}\right) + \left(\frac{u_i}{\sigma_i}\right)$$

两边除以已知的σ_i

（式11.3.4）
$$Y_i^* = \beta_1^* X_{0i}^* + \beta_2^* X_i^* + u_i^*$$

$E(\mu^*_i) = 0$

已知的σ_i

（式11.3.5）
$$\mathrm{var}(u_i^*) = E(u_i^*)^2 = E\left(\frac{u_i}{\sigma_i}\right)^2 = = \frac{1}{\sigma_i^2} E(u_i^2) = \frac{1}{\sigma_i^2}(\sigma_i^2) = 1$$

- 转换后模型的干扰项满足同方差性假定，再用OLS方法，就可以得到BLUE估计量
- 这就是GLS方法，得到的是GLS估计量！

广义最小二乘法（GLS）
——参数的GLS估计量：推导1

- 对于一元回归模型：

（式11.3.4）
PRF

$$Y_i^* = \beta_1^* X_{0i}^* + \beta_2^* X_i^* + u_i^*$$

（式11.3.6）
SRF

$$Y_i^* = \hat{\beta}_1^* X_{0i}^* + \hat{\beta}_2^* X_i^* + e_i^*$$

$$\frac{Y_i}{\sigma_i} = \hat{\beta}_1^* \left(\frac{X_{0i}}{\sigma_i}\right) + \hat{\beta}_2^* \left(\frac{X_i}{\sigma_i}\right) + \left(\frac{e_i}{\sigma_i}\right)$$

令X_{0i}=1

$$\sum e_i^{*2} = \sum (Y_i^* - \hat{\beta}_1^* X_{0i}^* - \hat{\beta}_2^* X_i^*)^2$$

（式11.3.7）

$$\sum \left(\frac{e_i}{\sigma_i}\right)^2 = \sum \left[\left(\frac{Y_i}{\sigma_i}\right) - \hat{\beta}_1^* \left(\frac{X_{0i}}{\sigma_i}\right) - \hat{\beta}_2^* \left(\frac{X_i}{\sigma_i}\right)\right]^2$$

广义最小二乘法（GLS）
——参数的GLS估计量：推导2

$$\sum e_i^{*2} = \sum (Y_i^* - \hat{\beta}_1^* X_{0i}^* - \hat{\beta}_2^* X_i^*)^2$$

$$\sum \left(\frac{e_i}{\sigma_i}\right)^2 = \sum \left[\left(\frac{Y_i}{\sigma_i}\right) - \hat{\beta}_1^* \left(\frac{X_{0i}}{\sigma_i}\right) - \hat{\beta}_2^* \left(\frac{X_i}{\sigma_i}\right)\right]^2$$

- 最小化并求偏导，得到 $\hat{\beta}_2^*$ 的GLS估计量和方差：

（式11.3.8）

$$\hat{\beta}_2^* = \frac{(\sum w_i)(\sum w_i X_i Y_i) - (\sum w_i X_i)(\sum w_i Y_i)}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2}$$

（式11.3.9）

$$\mathrm{var}(\hat{\beta}_2^*) = \frac{\sum \omega_i}{(\sum \omega_i)(\sum \omega_i X_i^2) - (\sum \omega_i X_i)^2}$$

其中 $\omega_i = \frac{1}{\sigma_i^2}$

广义最小二乘法（GLS）
——GLS与OLS的区别

- OLS的思想实质是最小化：

$$\sum e_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad \text{（式11.3.10）}$$

- GLS的思想实质是最小化：

$$\sum w_i e_i^2 = \sum w_i (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad \text{（式11.3.11）}$$

  - 因此，这里的GLS也被称为**加权最小二乘法**（weighted least squares，**WLS**）。
  - 其实，GLS是更一般的方法，包括工具变量法等。

广义最小二乘法（GLS）
——GLS与OLS的区别

- OLS的思想实质是最小化：

（式11.3.8）

（式11.3.9）

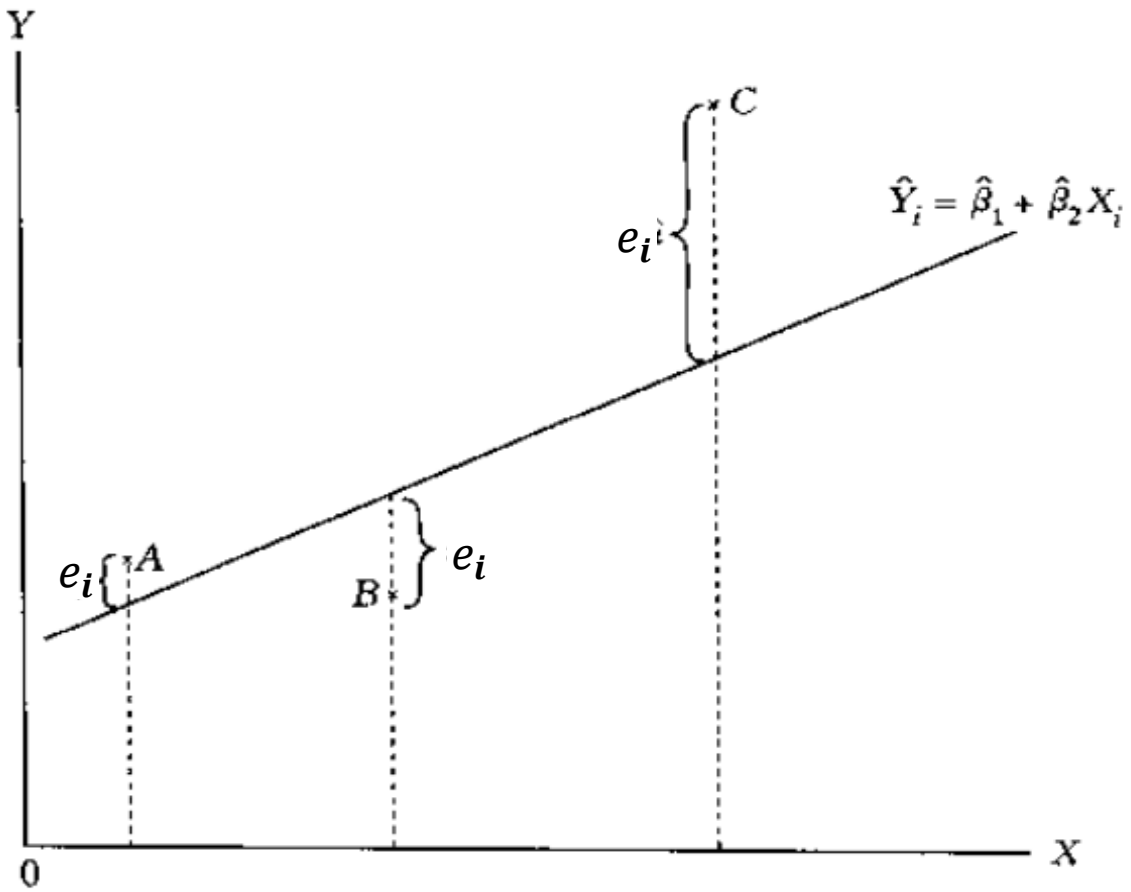


图 11.7　假想的散点图

出现异方差性时OLS的结果
——考虑异方差性的OLS估计

$\hat{\beta}_2$和$\hat{\beta}_2^*$都是(线性)无偏估计量，但$\hat{\beta}_2^*$才是有效估计量，即方差最小。如果我们继续使用OLS估计量$\hat{\beta}_2$，置信区间和假设检验会出现什么情况？下面分两种情况讨论：

➢ 考虑异方差性的OLS估计

- 如果我们仍然用$\hat{\beta}_2$，同时考虑异方差性而使用如下方差公式：（且假定σ_i^2已知）

（式11.2.2）

$$\mathrm{var}(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2}$$

考虑异方差的方差公式

- 结果是：
  - 利用$var(\hat{\beta}_2)$作出的置信区间将无谓地**增大**，t检验和F检验可能不准确。
  - 本来显著的系数可能变得统计上不显著了（t值过小）。
  - 因为：

$$\mathrm{var}(\beta_2^*) \le \mathrm{var}(\hat{\beta}_2)$$

证明略

出现异方差性时OLS的结果
——忽视异方差性的OLS估计

- 忽视异方差性的OLS估计
  - 在异方差性存在的情形下，我们不但使用了$\hat{\beta}_2$，而且使用如下同方差假设下的方差公式：

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \quad (式11.2.3)$$

假设同方差的方差公式

  - 结果是：
    - 上式给出的$\text{var}(\hat{\beta}_2)$是有偏的，可能低估或高估$\hat{\beta}_2$的真实的方差
    - 置信区间，t检验和F检验也将不准确
    - 因为：当异方差性出现时，$\hat{\sigma}^2$ 不再是σ^2的无偏估计量

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} \qquad E(\hat{\sigma}^2) \neq \sigma^2$$

如果我们忽视异方差性而执意使用惯常的检验程序，则无论我们得出什么结论或作出什么推断，都可能产生严重的误导。

异方差性的侦察
——非正式方法

- **1. 根据数据性质做判断（截面数据经常有）**
  - 在涉及不均匀、异质性（heterogeneous）单元（国家、省份、企业、家庭）的横截面数据中，异方差性可能是一种常规，而不是例外！
  - 例如，投资(Y)与销售量(X2)、利率(X3)等变量之间关系的横截面分析中， 如果样本同时包含小、中和大型厂家，一般都预期有异方差性。

- **2. 图解法（看残差e_i是否存在系统模式）**
  - 图解法A：残差e_i序列的描点图(dot plot)
  - 图解法B：残差e_i与解释变量X的XY散点图(scatter plot)

——图解法参考示例

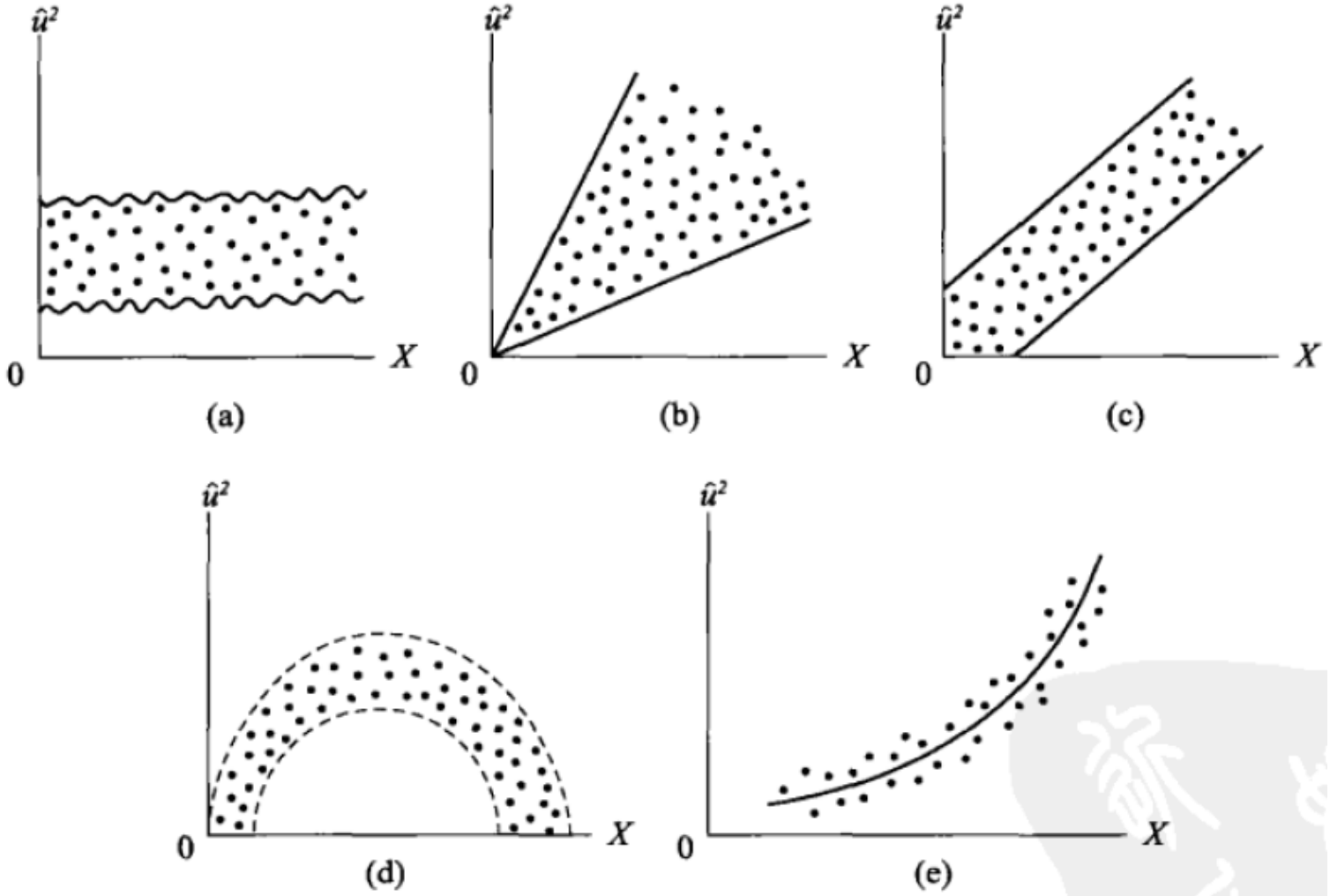


图 11—9　残差平方估计值相对于 X 的散点图

案例：平均薪资与平均生产力
——图解法：进行主回归

$$Salary_i = \hat{\beta}_1 + \hat{\beta}_2 Productivity_i + e_i$$

$$Salary_i = 1.992 + 0.233 Productivity_i + e_i$$

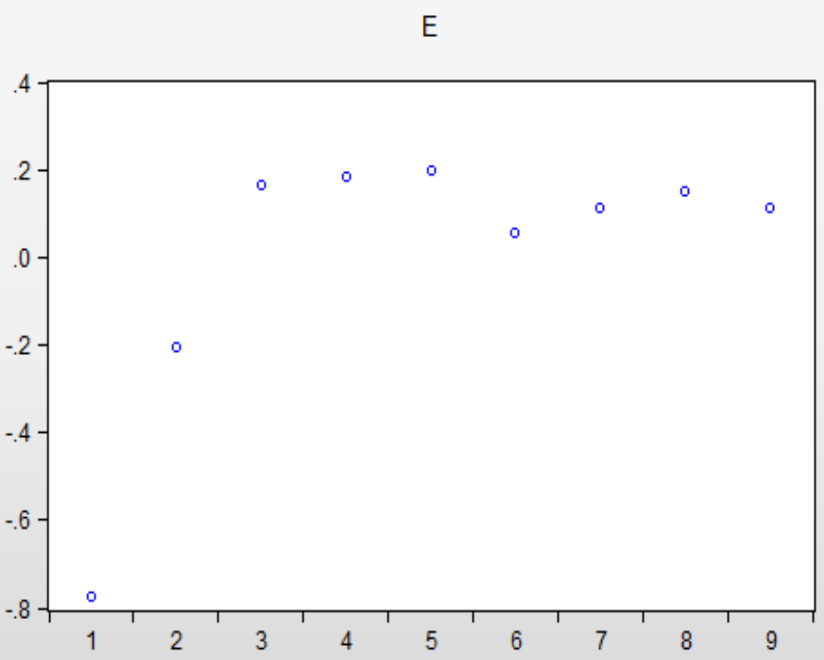
Dependent Variable: SALARY_MEAN
Method: Least Squares
Date: 10/19/14 Time: 13:53
Sample: 1 9
Included observations: 9

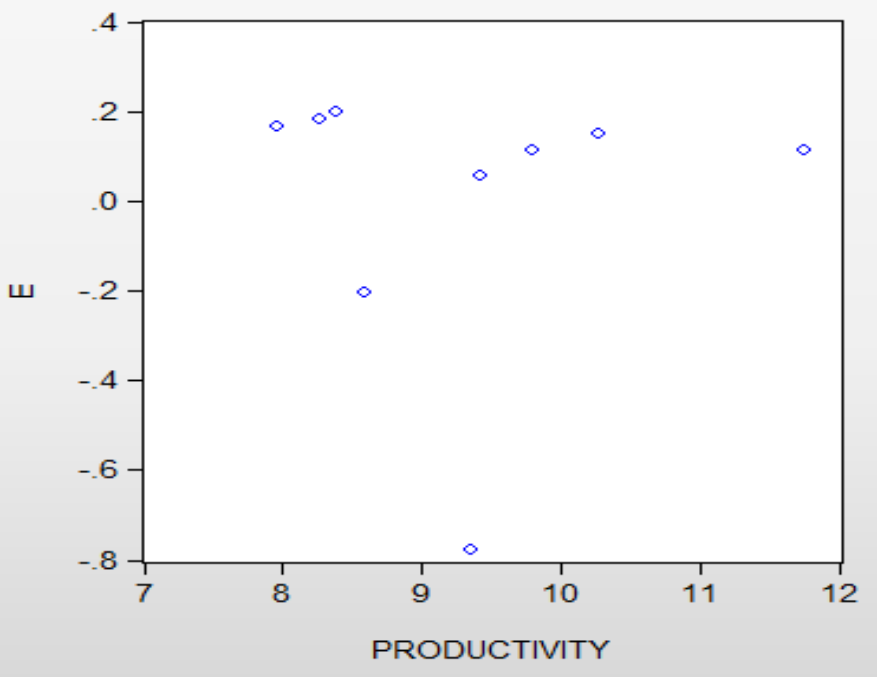
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.992062	0.936612	2.126880	0.0710
PRODUCTIVITY_MEAN	0.232999	0.099853	2.333428	0.0523

R-squared	0.437520	Mean dependent var	4.161778
Adjusted R-squared	0.357166	S.D. dependent var	0.420663
S.E. of regression	0.337274	Akaike info criterion	0.857290
Sum squared resid	0.796278	Schwarz criterion	0.901118
Log likelihood	-1.857805	Hannan-Quinn criter.	0.762710
F-statistic	5.444885	Durbin-Watson stat	0.616592
Prob(F-statistic)	0.052349		

	平均生产率X Productivity	残差 e_i	残差平方 e_i^2
1	9.355	-0.77577	0.60182
2	8.584	-0.20513	0.042078
3	7.962	0.165797	0.027489
4	8.275	0.183868	0.033808
5	8.389	0.199306	0.039723
6	9.418	0.05455	0.002976
7	9.795	0.113709	0.01293
8	10.281	0.150472	0.022642
9	11.75	0.113196	0.012813

案例：平均薪资与平均生产力
——图解法





图解法A： e_i**序列的描点图(dot plot)**

图解法B： e_i**序列与X序列的散点图(scatter plot)**

结论：描点图和散点图看出不系统模式，初步认为**不存在异方差性**

异方差性的侦察
——正式方法1： Park检验

- 1. Park检验（帕克检验）：对数化处理
  - 原理：若 σ_i^2 随 X_i 而变化(存在异方差)，则可形式化：

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{v_i}$$

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$

（式11.5.1）

$$\ln e_i^2 = \alpha + \beta \ln X_i + v_i$$

σ_i^2未知，用残差平方e_i^2代替

（式11.5.2）

  - 步骤：
    - 先做主回归，得到残差序列e_i和e_i^2
    - 再做辅助回归(park)，得到分析报告
  - 判断：看辅助回归分析报告
    - 如果**β的t检验**不显著，则认为同方差。
    - **如果β的t检验显著，则认为存在异方差；**

案例：平均薪金与平均生产力
——Park检验：过程与结论

Dependent Variable: SALARY_MEAN
Method: Least Squares
Date: 10/19/14 Time: 13:53
Sample: 1 9
Included observations: 9

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.992062	0.936612	2.126880	0.0710
PRODUCTIVITY_MEAN	0.232999	0.099853	2.333428	0.0523

R-squared	0.437520	Mean dependent var	4.161778
Adjusted R-squared	0.357166	S.D. dependent var	0.420663
S.E. of regression	0.337274	Akaike info criterion	0.857290
Sum squared resid	0.796278	Schwarz criterion	0.901118
Log likelihood	-1.857805	Hannan-Quinn criter.	0.762710
F-statistic	5.444885	Durbin-Watson stat	0.616592
Prob(F-statistic)	0.052349		

图1：主回归分析报告

Equation: EQ01_ORIGIN Workfile: CLASS DEMO45::Untitled\
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: Harvey

F-statistic	0.445907	Prob. F(1,7)	0.5257
Obs*R-squared	0.538976	Prob. Chi-Square(1)	0.4629
Scaled explained SS	0.194334	Prob. Chi-Square(1)	0.6593

Test Equation:
Dependent Variable: LRESID2 ← $\log(e_i^2)$
Method: Least Squares
Date: 05/05/16 Time: 20:07
Sample: 1 9
Included observations: 9

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.655653	9.346531	0.284132	0.7845
LOG(PRODUCTIVITY_MEAN)	-2.802022	4.196134	-0.667763	0.5257

不显著

R-squared	0.059886	Mean dependent var	-3.577070
Adjusted R-squared	-0.074416	S.D. dependent var	1.414819
S.E. of regression	1.466517	Akaike info criterion	3.796787
Sum squared resid	15.05470	Schwarz criterion	3.840615
Log likelihood	-15.08554	Hannan-Quinn criter.	3.702207
F-statistic	0.445907	Durbin-Watson stat	1.137101
Prob(F-statistic)	0.525681		

图2：Park辅助回归报告

Park检验初步结论：β_2的t检验不显著，认为主模型同方差

异方差性的侦察
——正式方法2：Glejser检验(格莱泽检验)*

- 2. Glejser检验(格莱泽检验)：
  - 原理：用e_i的**绝对值**对（被认为与σ^2密切相关的）X变量做如下多种形式的(Glejser)辅助回归：

$$|e_i| = \beta_1 + \beta_2 X_i + v_i$$
$$|e_i| = \beta_1 + \beta_2 \sqrt{X_i} + v_i$$
$$|e_i| = \beta_1 + \beta_2 \frac{1}{X_i} + v_i \quad \text{（式11.5.1）}$$

$$|e_i| = \beta_1 + \beta_2 \frac{1}{\sqrt{X_i}} + v_i$$
$$|e_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i$$
$$|e_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i \quad \text{（式11.5.2）}$$

  - 步骤：
    - 先做主回归，得到残差序列e_i和e_i^2
    - 再做辅助回归(park)，得到分析报告

v_i为随机干扰项

  - 判断：看辅助回归分析报告
    - 如果β_2的t检验不显著，则认为同方差。
    - **如果β_2的t检验显著，则认为存在异方差；**

案例：平均薪水与平均生产力
——Glejser检验：原始数据

Equation: EQ01_ORIGIN Workfile: CLASS DEMO45::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: Glejser

F-statistic	0.090875	Prob. F(1,7)	0.7718
Obs*R-squared	0.115342	Prob. Chi-Square(1)	0.7341
Scaled explained SS	0.114296	Prob. Chi-Square(1)	0.7353

Test Equation:
Dependent Variable: ARESID
Method: Least Squares
Date: 05/05/16 Time: 20:32
Sample: 1 9
Included observations: 9

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.407476	0.633189	0.643529	0.5404
PRODUCTIVITY_MEAN	-0.020350	0.067505	-0.301456	0.7718

R-squared	0.012816	Mean dependent var	0.217978
Adjusted R-squared	-0.128210	S.D. dependent var	0.214665
S.E. of regression	0.228012	Akaike info criterion	0.074290
Sum squared resid	0.363925	Schwarz criterion	0.118118
Log likelihood	1.665693	Hannan-Quinn criter.	-0.020290
F-statistic	0.090875	Durbin-Watson stat	1.015122
Prob(F-statistic)	0.771825		

图：格莱泽辅助回归分析报告

$$|e_i| = \hat{\beta}_1^* + \hat{\beta}_2^* Productivity\_mean_i + e_i^*$$

Glejser检验初步结论：β_2的t检验不显著，认为主模型是同方差

异方差性的侦察
——正式方法3：BPG检验

- 3. BPG检验(Breusch-Pagan-Godfrey)
  - 原理：构建如下模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \quad \text{（式11.5.12）}$$

$$\sigma_i^2 = f(\alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + \cdots + \alpha_m Z_{mi}) \quad \text{（式11.5.13）}$$

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + \cdots + \alpha_m Z_{mi} \quad \text{（式11.5.14）}$$

异方差性的侦察
——正式方法3：BPG检验（步骤）

- 3. BPG检验(Breusch-Pagan-Goldfrey)
  - 步骤：
    - 进行主回归，得到残差序列e_i=($e_1, e_2, \ldots, e_n$)
    - 计算$\hat{\sigma}^2 = \sum e_i^2/n$
    - 构造新序列$P_i = e_i^2/\hat{\sigma}^2$
    - 做P_i对各Z（部分或全部X变量）的辅助回归：

（式11.5.15）

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + \cdots + \alpha_m Z_{mi} + v_i$$

（式11.5.17）

    - 计算ESS(解释的平方和)，定义并有：

$$\chi^* = \frac{ESS}{2} \sim \chi^2(m-1)$$

  - 判断：
    - 若卡方检验不显著，即 $\chi^* < \chi^2$，则认为是同方差。
    - **若卡方检验显著，即 $\chi^* > \chi^2$，则认为存在异方差。**

案例2：平均薪水与平均生产力
——BPG检验：进行BPG辅助回归和χ^2检验

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	0.005998	Prob. F(1,7)	0.9404
Obs*R-squared	0.007705	Prob. Chi-Square(1)	0.9301
Scaled explained SS	0.009853	Prob. Chi-Square(1)	0.9209

结论：BPG的χ^2检验不显著，认为**不存在异方差性**

$$\Theta^* = \frac{ESS}{2} = 0.0098 < \chi^2_{0.05(1)} = 3.8414$$

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 10/19/14 Time: 14:35
Sample: 1 9
Included observations: 9

$$Resid_i^2 = 0.132 - 0.004 Productivity_i + e_i^*$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.132499	0.572573	0.231409	0.8236
PRODUCTIVITY_MEAN	-0.004728	0.061042	-0.077447	0.9404

R-squared	0.000856	Mean dependent var	0.088475
Adjusted R-squared	-0.141879	S.D. dependent var	0.192950
S.E. of regression	0.206184	Akaike info criterion	-0.126969
Sum squared resid	0.297582	Schwarz criterion	-0.083142
Log likelihood	2.571362	Hannan-Quinn criter.	-0.221549
F-statistic	0.005998	Durbin-Watson stat	1.072385
Prob(F-statistic)	0.940435		

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + \cdots + \alpha_m Z_{mi} + v_i$$

$$p_i = \frac{e_i^2}{\hat{\sigma}_i^2}; \quad \hat{\sigma}_i^2 = \frac{\sum e_i^2}{n}$$

$$Salary_i = 1.992 + 0.233 Productivity_i + e_i$$

异方差性的侦察
——正式方法4：怀特的一般异方差性检验（原理）

（式11.5.21）

- 4. 怀特(White) 异方差性检验：
  - 原理：构建如下模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

看 e_i^2 与 $X_{2i}, X_{3i}, X_{2i}^2, X_{3i}^2, X_{2i}X_{3i}$ 是否存在线性回归关系！

- 本质：怀特检验可能是(纯粹)**异方差性的一个检验**，或者是**设定错误的一个检验**，或者两者兼有。已经被证明，若怀特检验程序中没有出现交叉项， 则是对纯粹异方差性的检验；若出现交叉项，则既是对异方差性又是对设定偏误的检验。

异方差性的侦察
——正式方法4：怀特的一般异方差性检验（步骤）

（式11. 5. 15）

（式11. 5. 17）

- **检验步骤：**

步骤1.　估计（11.5.21),获得残差e_i

步骤2.　做辅助回归：

$$e_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i \qquad (11.5.22)$$

还可以引入回归元的更高次方。求出R^2

步骤3.　设置虚拟假设H_0：无异方差性。可证：

$$n \cdot R^2 \underset{asy}{\sim} \chi_{df}^2 \qquad (11.5.23)$$

其中df = 辅助回归中的回归元（不含常数项）的个数，这里$df = 5$

判断：　计算的χ^2值大于χ^2临界值，有异方差性；

$\chi^2 < \chi_\alpha^2(df)$，没有异方差性，即：

$$a_2 = a_3 = a_4 = a_5 = a_6 = 0$$

案例：平均薪水与平均生产力
——White检验：进行怀特辅助回归和χ^2检验

Heteroskedasticity Test: White

F-statistic	0.336607	Prob. F(2,6)	0.7269
Obs*R-squared	0.907947	Prob. Chi-Square(2)	0.6351
Scaled explained SS	1.161001	Prob. Chi-Square(2)	0.5596

结论：White的χ^2检验不显著，认为**不存在异方差性**

$nR^{2*} = 0.908 < \chi^2_{0.05(2)} = 3.8414$

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 10/19/14 Time: 15:04
Sample: 1 9
Included observations: 9

$$Resid_i^2 = -3.798 - 0.042 Productivity_i + 0.811 Productivity_i^2 + e_i^*$$

$$e_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + e_i^*$$

$$Salary_i = 1.992 + 0.233 Productivity_i + e_i$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.797516	4.845903	-0.783655	0.4630
PRODUCTIVITY_MEAN^2	-0.041705	0.051046	-0.817007	0.4451
PRODUCTIVITY_MEAN	0.811347	1.000815	0.810686	0.4485

R-squared	0.100883	Mean dependent var	0.088475
Adjusted R-squared	-0.198823	S.D. dependent var	0.192950
S.E. of regression	0.211262	Akaike info criterion	-0.010233
Sum squared resid	0.267790	Schwarz criterion	0.055509
Log likelihood	3.046047	Hannan-Quinn criter.	-0.152103
F-statistic	0.336607	Durbin-Watson stat	1.120893
Prob(F-statistic)	0.726857		

异方差性的侦察
——正式方法6：G-Q检验（原理）

- **G-Q检验(戈德菲尔德-匡特检验)：**
  - 原理：扩大差异，F检验

（式11.5.10）

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\sigma_i^2 = \sigma^2 X_i^2$$

σ^2为常数

该模型表明：着X_i值越大，σ_i^2也越大。如果情况正是如此，则模型中有异方差性是最为可能的。

异方差性的侦察
——正式方法6：G-Q检验（步骤）

- G-Q检验(戈德菲尔德-匡特检验)：
  - 步骤：
    - 对X排序，从小到大
    - 去掉中间大约c个数(预先确定的)
    - 分段回归:前一半样本(n-c)/2，后一半样本(n-c)/2
    - 得到残差平方和：RSS_1 和RSS_2　自由度=(n-c)/2-k
    - 构造F统计量(假定μ_i是正态分布)

$$\lambda=\frac{RSS_2/df}{RSS_1/df}\sim F((n-c-2k)/2,(n-c-2k)/2)$$

（式11.5.11）

    - 异方差判断依据：

计算的λ 值大于选定显著性水平的F 临界值，就拒绝同方差性假设，也就是说很可能出现了异方差性。

  - 不足：存在多个解释变量时，难以确定合适的排序依据。

案例:消费与收入的人为数据
——G-Q检验：人为的原始数据

表 11—3　为说明戈德菲尔德-匡特检验的假想消费 Y（美元）与收入 X（美元）数据

Y	X	按 X 值排序的数据 Y	按 X 值排序的数据 X
55	80	55	80
65	100	70	85
70	85	75	90
80	110	65	100
79	120	74	105
84	115	80	110
98	130	84	115
95	140	79	120
90	125	90	125
75	90	98	130
74	105	95	140
110	160	108	145
113	150	113	150
125	165	110	160
108	145	125	165
115	180	115	180
140	225	130	185
120	200	135	190
145	240	120	200
130	185	140	205
152	220	144	210
144	210	152	220
175	245	140	225
180	260	137	230
135	190	145	240
140	205	175	245
178	265	189	250
191	270	180	260
137	230	178	265
189	250	191	270

居中的 4 个观测值

n=30

案例:消费与收入的人为数据
——G-Q检验：过程与结论

对前13 个观测值做回归:

$$\hat{Y}_i = 3.409\,4 + 0.696\,8X_i$$

(8.704 9)　(0.074 4)　$r^2=0.888\,7$　$RSS_1=377.17$　df=11

对后13 个观测值做回归:

$$\hat{Y}_i = -28.027\,2 + 0.794\,1X_i$$

(30.642 1)　(0.131 9)　$r^2=0.768\,1$　$RSS_2=1\,536.8$　df=11

计算F比率值 λ:

$$\lambda = \frac{RSS_2/df}{RSS_1/df} = \frac{1\,536.8/11}{377.17/11}$$

$$\lambda = 4.07$$

$F_{0.05}(11,11)=2.82.$

拒绝H_0(同方差)，接受H_1(异方差)，认为误差方差中有异方差性!

补救措施：若σ_i^2已知
——加权最小二乘法(WLS)

（式11.6.1）

- 补救1：σ_i^2已知：加权最小二乘法(WLS)

$$\frac{Y_i}{\sigma_i}=\hat{\beta}_1^*\left(\frac{1}{\sigma_i}\right)+\hat{\beta}_2^*\left(\frac{X_i}{\sigma_i}\right)+\left(\frac{\hat{u}_i}{\sigma_i}\right)$$

  - 特别提示：通常总体的σ_i^2难以获知，故此补救措施只是理论探讨，实际中并不可行。

补救措施：加权最小二乘法(*WLS*)
——薪金与就业人数的例子

（Table 11-4）

表 11—4　　加权最小二乘回归的说明

酬金，Y	就业人数，X	σ_i	Y_i/σ_i	X_i/σ_i
3 396	1	742.2	4.566 4	0.001 3
3 787	2	851.4	4.448 0	0.002 3
4 013	3	727.8	5.513 9	0.004 1
4 104	4	805.06	5.097 8	0.005 0
4 146	5	929.9	4.458 5	0.005 4
4 241	6	1 080.6	3.924 7	0.005 5
4 387	7	1 241.2	3.528 8	0.005 6
4 538	8	1 307.7	3.470 2	0.006 1
4 843	9	1 110.7	4.353 2	0.008 1

注：在回归（11.6.2）中，因变量是（Y_i/σ_i）而自变量是（$1/\sigma_i$）和（X_i/σ_i）。

资料来源：Y_i 和 σ_i（薪金的标准差）的数据来自表 11—1。就业人数 1＝1～4 人职工组，2＝5～9 人职工组……其他数据也来自表 11—1。

$$\widehat{(Y_i/\sigma_i)} = 3\,406.639(1/\sigma_i) + 154.153(X_i/\sigma_i)$$
$$(80.983) \qquad (16.959)$$
$$t= \quad (42.066) \qquad (9.090) \qquad R^2=0.999\,3^{①}$$

WLS回归

$$\hat{Y}_i = 3\,417.833 + 148.767X_i$$
$$(81.136) \quad (14.418)$$
$$t= \quad (42.125) \quad (10.318) \qquad R^2=0.938\,3$$

OLS回归

补救措施：σ_i^2未知
——几种异方差情形：情形1（误差方差正比于X^2）

- 补救2：σ_i^2未知(情形1)（对应White检验辅助回归无交叉项的情形）

假定1.：　　误差方差正比于X_i^2

$$E(u_i^2)=\sigma^2 X_i^2 \qquad (11.6.5)$$

（式11.6.5）

用X_i去除原模型，得：

$$\frac{Y_i}{X_i}=\frac{\beta_1}{X_i}+\beta_2+\frac{u_i}{X_i}$$

$$=\beta_1\frac{1}{X_i}+\beta_2+v_i \qquad (11.6.6)$$

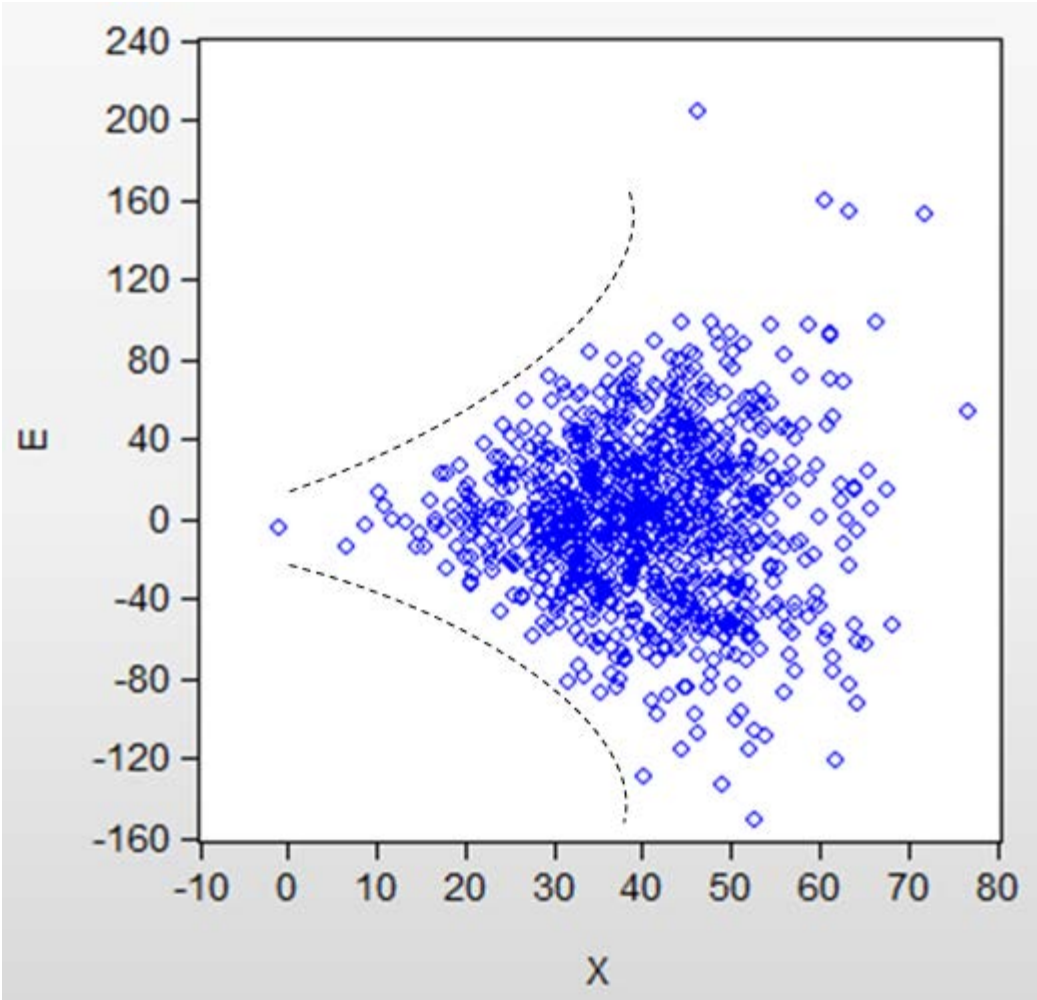
（式11.6.6）

$$E(v_i^2)=E\left(\frac{u_i}{X_i}\right)^2=\frac{1}{X_i^2}E(u_i^2)$$

$$=\sigma^2$$

补救措施：σ_i^2未知
——几种异方差情形：情形1（误差方差正比于X^2）

注：课本第391页，图11-10标识有误。



模拟结果(n=1000)：误差方差正比于X^2的XY散点图

补救措施：σ_i^2未知
——几种异方差情形：情形2（误差方差正比于X）

- 补救2：σ_i^2未知(情形2) （对应B-P-G检验辅助回归的情形）

假定2.: 误差方差正比于X_i:

$$E(u_i^2)=\sigma^2 X_i \qquad (11.6.7)$$

（式11.6.7）

平方根变换：

$$\frac{Y_i}{\sqrt{X_i}}=\frac{\beta_1}{\sqrt{X_i}}+\beta_2\sqrt{X_i}+\frac{u_i}{\sqrt{X_i}}$$

$$=\beta_1\frac{1}{\sqrt{X_i}}+\beta_2\sqrt{X_i}+v_i \qquad (11.6.8)$$

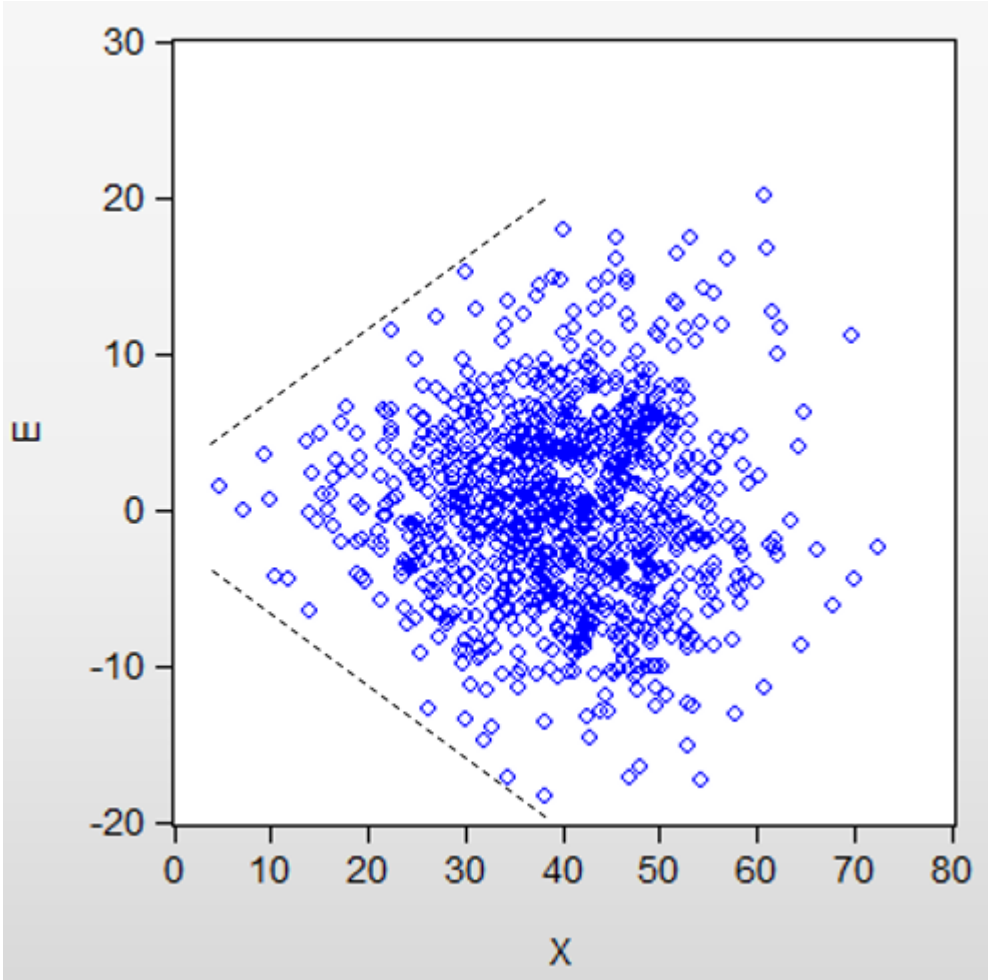
（式11.6.8）

其中$v_i=\frac{u_i}{\sqrt{X_i}}$，且$X_i>0$

可证：$E(v_i^2)=\sigma^2$,即为同方差性，可用*OLS*估计参数

补救措施：σ_i^2未知
——几种异方差情形：情形2（误差方差正比于X）

注：课本第392页，图11-11标识有误



模拟结果(n=1000)：误差方差正比于X的XY散点图

补救措施：σ_i^2未知
——几种异方差情形：情形3（误差正比于Y均值的平方）

- **补救2：σ_i^2未知(情形3)**

假定3.：　误差方差正比于Y均值的平方：

$$E(u_i^2) = \sigma^2[E(Y_i)]^2 \qquad (11.6.9)$$

现在　　$E(Y_i) = \beta_1 + \beta_2 X_i$

模型变换：

$$\frac{Y_i}{E(Y_i)} = \frac{\beta_1}{E(Y_i)} + \beta_2 \frac{X_i}{E(Y_i)} + \frac{u_i}{E(Y_i)}$$

$$= \beta_1\left(\frac{1}{E(Y_i)}\right) + \beta_2 \frac{X_i}{E(Y_i)} + v_i \qquad (11.6.10)$$

$E(Y_i)$不可知，利用$E(Y_i)$的一致性估计值$\hat{Y}_i$：

$$\frac{Y_i}{\hat{Y}_i} = \beta_1\left(\frac{1}{\hat{Y}_i}\right) + \beta_2\left(\frac{X_i}{\hat{Y}_i}\right) + v_i \qquad (11.6.11)$$

其中$v_i = \frac{u_i}{\hat{Y}_I}$。变换后的（11.6.11）一般具有良好的性质

补救措施：σ_i^2未知
——几种异方差情形：情形4（对数变换）

- 补救2：σ_i^2未知(情形4)

假定4.:　　线性回归模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

而如下的对数模型：

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \qquad (11.6.12)$$

通常能降低异方差性

- 对数变换降低异方差，必须满足下列条件：
  - 原始数据分布偏态；
  - 原始数据量纲差异较大；
  - 原始数据所有取值均大于0；
  - 原始数据分布偏态和量纲差异是异方差的主要来源。

补救措施：σ_i^2未知
——几个忠告

- **忠告1：**

  实际回归中，由于我们不能得到总体，从而也就无法获知σ_i^2，因此σ_i^2未知的情形，就是通常的状态。

- **忠告2：**

  由于σ_i^2未知，我们就不能采用σ_i^2进行加权，来降低异方差的影响，因此，上述4种情形的变换，实质上是通过猜测异方差的具体形式，并检验确认，然后进行加权变换，来达到降低异方差影响的效果。

- **忠告3：**

  加权变换只是有限程度地降低异方差的影响，并不能完全消除异方差的影响。

- **忠告4：**

  错误的加权变换会带来新的问题。在进行加权变换之前，务必反复确认：加权变换的方法与异方差形式相匹配。

总结性的例子
——美国18个产业群数据：R&D支出、销售额和利润

IND	SALES	RD	PROFITS
1	6375.3	62.5	185.1
2	11626.4	92.9	1569.5
3	14655.1	178.3	276.8
4	21869.2	258.4	2828.1
5	26408.3	494.7	225.9
6	32405.6	1083	3751.9
7	35107.7	1620.6	2884.1
8	40295.4	421.7	4645.7
9	70761.6	509.2	5036.4
10	80552.8	6620.1	13869.9
11	95294	3918.6	4487.8
12	101314.1	1595.3	10278.9
13	116141.3	6107.5	8787.3
14	122315.7	4454.1	16438.8
15	141649.9	3163.8	9761.4
16	175025.8	13210.7	19774.5
17	230614.5	1703.8	22626.6
18	293543	9528.2	18415.4

提问1：
是否存在多重共线性？

提问2：
是否存在异方差性？

总结性的例子
——美国产业群数据：多元回归模型（多重共线性诊断）

Dependent Variable: RD
Method: Least Squares
Date: 10/23/14 Time: 11:29
Sample: 1 18
Included observations: 18

$$RD_i = \hat{\beta}_1 + \hat{\beta}_2 SALES_i + \hat{\beta}_3 PROFITS_i + e_i$$

$$RD_i = -13.96 + 0.013 SALES_i + 0.24 PROFITS_i + e_i$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-13.95579	991.9935	-0.014068	0.9890
SALES	0.012559	0.017997	0.697818	0.4960
PROFITS	0.239844	0.198592	1.207726	0.2459

R-squared	0.524537	Mean dependent var	3056.856
Adjusted R-squared	0.461142	S.D. dependent var	3705.973
S.E. of regression	2720.441	Akaike info criterion	18.80599
Sum squared resid	1.11E+08	Schwarz criterion	18.95438
Log likelihood	-166.2539	Hannan-Quinn criter.	18.82645
F-statistic	8.274108	Durbin-Watson stat	3.173945
Prob(F-statistic)	0.003788		

提问1：
数据是否存在严重的多重共线性？（多重共线性是否引发了严重的问题？）

多重共线性诊断标准1：高的R平方，低的解释变量系数t值

总结性的例子
——美国产业群数据：多重共线性诊断与处理

Correlation

	SALES	PROFITS
SALES	1.000000	0.889829
PROFITS	0.889829	1.000000

多重共线性诊断标准2：解释变量之间高的相关系数。

Equation: EQ01 Workfile: 2.表11.5::1\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Variance Inflation Factors
Date: 04/10/18 Time: 17:55
Sample: 1 18
Included observations: 18

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
C	984051.1	2.393379	NA
SALES	0.000324	11.15236	4.802985
PROFITS	0.039439	11.10019	4.802985

多重共线性诊断标准3：VIF值大于10

问题：
是否存在严重的多重共线性？

多重共线性如何处理？

总结性的例子
——美国产业群数据：一元回归模型（异方差诊断）

Dependent Variable: RD
Method: Least Squares
Date: 04/10/18 Time: 22:21
Sample: 1 18
Included observations: 18

$$RD_i = \hat{\beta}_1 + \hat{\beta}_2 SALES_i + e_i$$

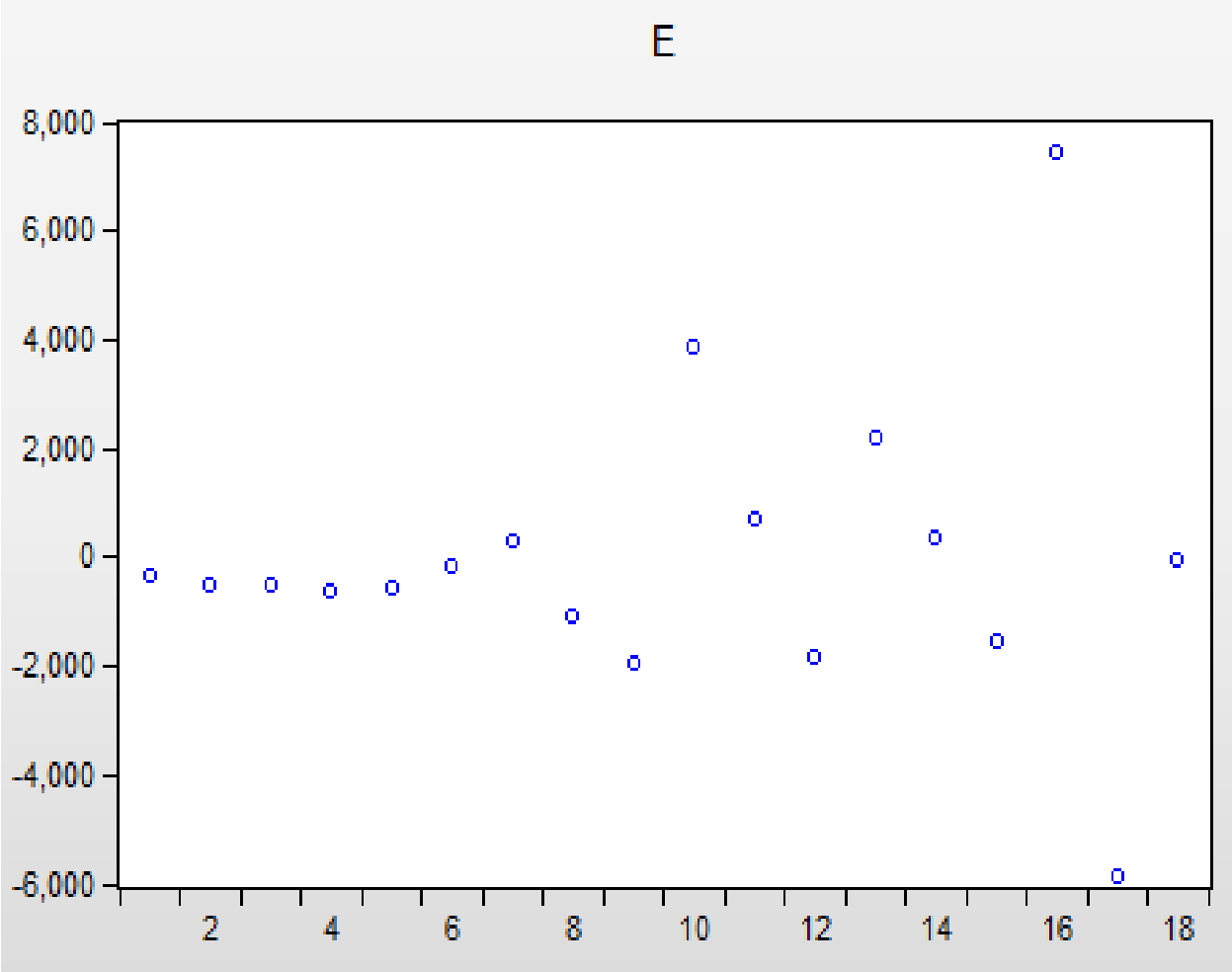
$$RD_i = 192.99 + 0.032 SALES_i + e_i$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	192.9931	990.9858	0.194749	0.8480
SALES	0.031900	0.008329	3.830033	0.0015

R-squared	0.478303	Mean dependent var	3056.856
Adjusted R-squared	0.445697	S.D. dependent var	3705.973
S.E. of regression	2759.153	Akaike info criterion	18.78767
Sum squared resid	1.22E+08	Schwarz criterion	18.88660
Log likelihood	-167.0891	Hannan-Quinn criter.	18.80132
F-statistic	14.66916	Durbin-Watson stat	3.015607
Prob(F-statistic)	0.001476		

提问：
是否存在严重的异方差性？（异方差是否引发了严重的问题？）

总结性的例子
——美国产业群数据：一元回归模型（异方差诊断）

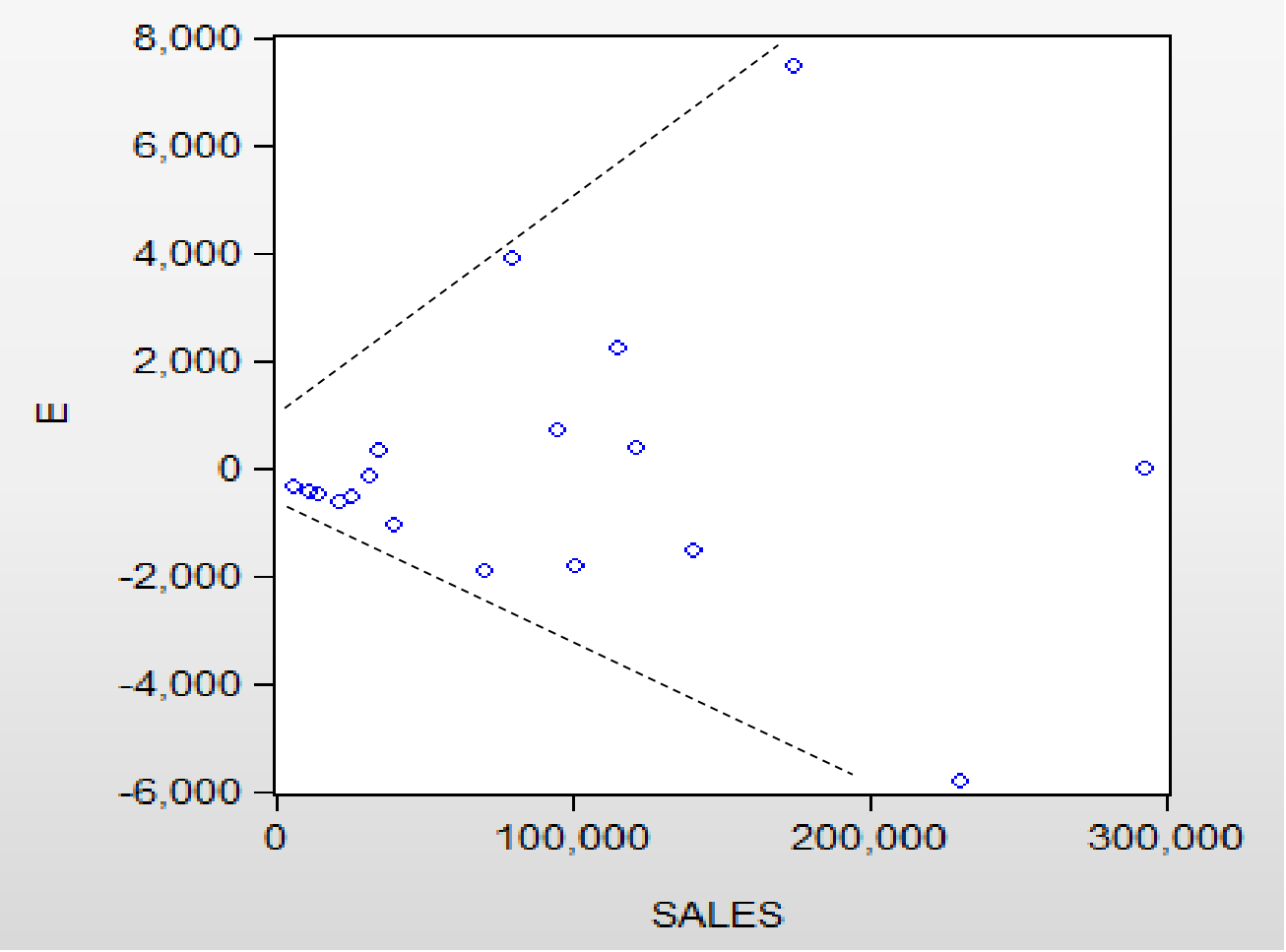


残差e_i序列的描点图(dot plot)

图解法A：

残差e_i序列没有位于某个固定的区间，呈现出发散的趋势，表明存在异方差。

总结性的例子
——美国产业群数据：一元回归模型（异方差诊断）



图解法B：

残差e_i随着解释变量sales取值的增加而呈现出发散趋势，表明存在异方差。

残差e_i与解释变量Sales的XY散点图(scatter plot)

总结性的例子
——美国产业群数据：一元回归模型（异方差诊断）

Heteroskedasticity Test: Harvey

F-statistic	1.351643	Prob. F(1,16)	0.2620
Obs*R-squared	1.402148	Prob. Chi-Square(1)	0.2364
Scaled explained SS	1.987316	Prob. Chi-Square(1)	0.1586

Test Equation:
Dependent Variable: LRESID2
Method: Least Squares
Date: 04/10/18 Time: 21:34
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.687749	6.635175	0.857212	0.4040
LOG(SALES)	0.701435	0.603332	1.162602	0.2620

R-squared	0.077897	Mean dependent var	13.36642
Adjusted R-squared	0.020266	S.D. dependent var	2.721342
S.E. of regression	2.693626	Akaike info criterion	4.924093
Sum squared resid	116.0899	Schwarz criterion	5.023023
Log likelihood	-42.31683	Hannan-Quinn criter.	4.937734
F-statistic	1.351643	Durbin-Watson stat	1.567339
Prob(F-statistic)	0.262039		

Park检验：
无法拒绝原假设（H0:同方差），故不存在异方差。

注：Park检验是Harvey检验的特殊形式。

总结性的例子
——美国产业群数据：一元回归模型（异方差诊断）

Heteroskedasticity Test: Glejser

F-statistic	4.380896	Prob. F(1,16)	0.0526
Obs*R-squared	3.869120	Prob. Chi-Square(1)	0.0492
Scaled explained SS	5.654785	Prob. Chi-Square(1)	0.0174

Test Equation:
Dependent Variable: ARESID
Method: Least Squares
Date: 04/10/18 Time: 21:41
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	578.5710	678.6950	0.852476	0.4065
SALES	0.011939	0.005704	2.093059	0.0526

R-squared	0.214951	Mean dependent var	1650.432
Adjusted R-squared	0.165886	S.D. dependent var	2069.046
S.E. of regression	1889.657	Akaike info criterion	18.03062
Sum squared resid	57132868	Schwarz criterion	18.12955
Log likelihood	-160.2756	Hannan-Quinn criter.	18.04426
F-statistic	4.380896	Durbin-Watson stat	1.743294
Prob(F-statistic)	0.052633		

Glejser检验：
无法拒绝原假设（H0:同方差），故不存在异方差。

注：显著性水平：5%

总结性的例子
——美国产业群数据：一元回归模型（异方差诊断）

Heteroskedasticity Test: White

F-statistic	2.458856	Prob. F(1,16)	0.1364
Obs*R-squared	2.397733	Prob. Chi-Square(1)	0.1215
Scaled explained SS	4.225069	Prob. Chi-Square(1)	0.0398

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/10/18 Time: 21:48
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3399398.	3959493.	0.858544	0.4033
SALES^2	0.000238	0.000152	1.568074	0.1364

R-squared	0.133207	Mean dependent var	6767046.
Adjusted R-squared	0.079033	S.D. dependent var	14706011
S.E. of regression	14112923	Akaike info criterion	35.86752
Sum squared resid	3.19E+15	Schwarz criterion	35.96645
Log likelihood	-320.8077	Hannan-Quinn criter.	35.88116
F-statistic	2.458856	Durbin-Watson stat	1.688933
Prob(F-statistic)	0.136426		

White检验（不含交叉项）：
无法拒绝原假设（H0:同方差），故不存在异方差。

注：显著性水平：5%

总结性的例子
——美国产业群数据：一元回归模型（异方差诊断）

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	4.564374	Prob. F(1,16)	0.0484
Obs*R-squared	3.995197	Prob. Chi-Square(1)	0.0456
Scaled explained SS	7.039975	Prob. Chi-Square(1)	0.0080

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/10/18 Time: 21:49
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-974469.1	4802343.	-0.202915	0.8418
SALES	86.23211	40.36253	2.136439	0.0484

R-squared	0.221955	Mean dependent var	6767046.
Adjusted R-squared	0.173328	S.D. dependent var	14706011
S.E. of regression	13370930	Akaike info criterion	35.75950
Sum squared resid	2.86E+15	Schwarz criterion	35.85843
Log likelihood	-319.8355	Hannan-Quinn criter.	35.77314
F-statistic	4.564374	Durbin-Watson stat	1.791548
Prob(F-statistic)	0.048439		

B-P-G检验：
拒绝原假设（H0:同方差），故存在异方差。

注：显著性水平：5%

总结性的例子
——美国产业群数据：一元回归模型（异方差校正结果）

Dependent Variable: RD
Method: Least Squares
Date: 04/10/18 Time: 21:51
Sample: 1 18
Included observations: 18
Weighting series: SALES
Weight type: Variance (average scaling)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-246.6769	381.1285	-0.647228	0.5267
SALES	0.036798	0.007114	5.172315	0.0001

Weighted Statistics

R-squared	0.625756	Mean dependent var	1583.280
Adjusted R-squared	0.602366	S.D. dependent var	1579.546
S.E. of regression	1297.542	Akaike info criterion	17.27877
Sum squared resid	26937844	Schwarz criterion	17.37770
Log likelihood	-153.5089	Hannan-Quinn criter.	17.29241
F-statistic	26.75284	Durbin-Watson stat	2.885313
Prob(F-statistic)	0.000093	Weighted mean dep.	929.6666

Unweighted Statistics

R-squared	0.467030	Mean dependent var	3056.856
Adjusted R-squared	0.433719	S.D. dependent var	3705.973
S.E. of regression	2788.805	Sum squared resid	1.24E+08
Durbin-Watson stat	2.961493		

异方差校正方法：
加权最小二乘法（WLS）

校正依据：
根据B-P-G检验的辅助回归结果，确认：误差方差正比于解释变量sales。

加权操作方法：
加权类型：Variance加权
权重序列：sales

注：显著性水平：5%

总结性的例子
——美国产业群数据：一元回归模型（结果对比）

Dependent Variable: RD
Method: Least Squares
Date: 04/10/18 Time: 22:21
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	192.9931	990.9858	0.194749	0.8480
SALES	0.031900	0.008329	3.830033	0.0015

R-squared	0.478303	Mean dependent var	3056.856
Adjusted R-squared	0.445697	S.D. dependent var	3705.973
S.E. of regression	2759.153	Akaike info criterion	18.78767
Sum squared resid	1.22E+08	Schwarz criterion	18.88660
Log likelihood	-167.0891	Hannan-Quinn criter.	18.80132
F-statistic	14.66916	Durbin-Watson stat	3.015607
Prob(F-statistic)	0.001476		

未进行异方差校正的回归结果

Dependent Variable: RD
Method: Least Squares
Date: 04/10/18 Time: 21:51
Sample: 1 18
Included observations: 18
Weighting series: SALES
Weight type: Variance (average scaling)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-246.6769	381.1285	-0.647228	0.5267
SALES	0.036798	0.007114	5.172315	0.0001

Weighted Statistics

R-squared	0.625756	Mean dependent var	1583.280
Adjusted R-squared	0.602366	S.D. dependent var	1579.546
S.E. of regression	1297.542	Akaike info criterion	17.27877
Sum squared resid	26937844	Schwarz criterion	17.37770
Log likelihood	-153.5089	Hannan-Quinn criter.	17.29241
F-statistic	26.75284	Durbin-Watson stat	2.885313
Prob(F-statistic)	0.000093	Weighted mean dep.	929.6666

Unweighted Statistics

R-squared	0.467030	Mean dependent var	3056.856
Adjusted R-squared	0.433719	S.D. dependent var	3705.973
S.E. of regression	2788.805	Sum squared resid	1.24E+08
Durbin-Watson stat	2.961493		

经过异方差校正的回归结果

总结：有关异方差问题的几个忠告

- **忠告1：** 谨防对异方差问题反应过度。

- **忠告2：**

  “一个好的模型，绝不会因异方差性的原因而被抛弃！”

  ——N.Gregory Mankiw(N.格利高里·曼昆)

- **忠告3：**

  “……只有在问题严重的时候，误差方差不相等的问题才值得去修正。”

  ——John Fox(约翰·福克斯)

- **忠告4：**

  尽管有异方差性的问题，但OLS 估计量仍是**线性无偏**和**渐近正态分布**(大样本)。（异方差只影响OLS估计系数的**方差最小**性质）

本节内容：

§4.3 自相关

自相关的性质
——含义及数学表达

- **自相关（autocorrelation）：**
  - 按时间（时间序列数据）或空间（横截面数据）排列的观测值序列的成员之间的相关。
    - 经典线性回归模型假定在干扰项之间不存在自相关：

$$E(u_i u_j) = 0, \quad i \neq j$$

（式8.1.4）

    - 出现自相关时，则为：

$$E(u_i u_j) \neq 0, \quad i \neq j$$

  - 自相关的分类：
    - 序列自相关（Serial Correlation）：时间序列数据
    - 空间自相关（Spatial Correlation）：横截面数据

（式12.1.1）

注：通常情况下，自相关通常是指时间序列数据的自相关。

自相关的性质
——含义及矩阵表达

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$E(\mathbf{uu}'|\mathbf{X}) = \begin{bmatrix} E(u_1^2) & E(u_1u_2)\cdots E(u_1u_n) \\ E(u_2u_1) & E(u_2^2)\cdots E(u_2u_n) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ E(u_nu_1) & E(u_nu_2)\cdots E(u_n^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \cdots\sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_2^2 & \cdots\sigma_{2n}^2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots\sigma_n^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & \sigma_{12}^2 & \cdots\sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma^2 & \cdots\sigma_{2n}^2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots\sigma^2 \end{bmatrix}$$

自相关的性质
——模式图

（Figure 12-1）

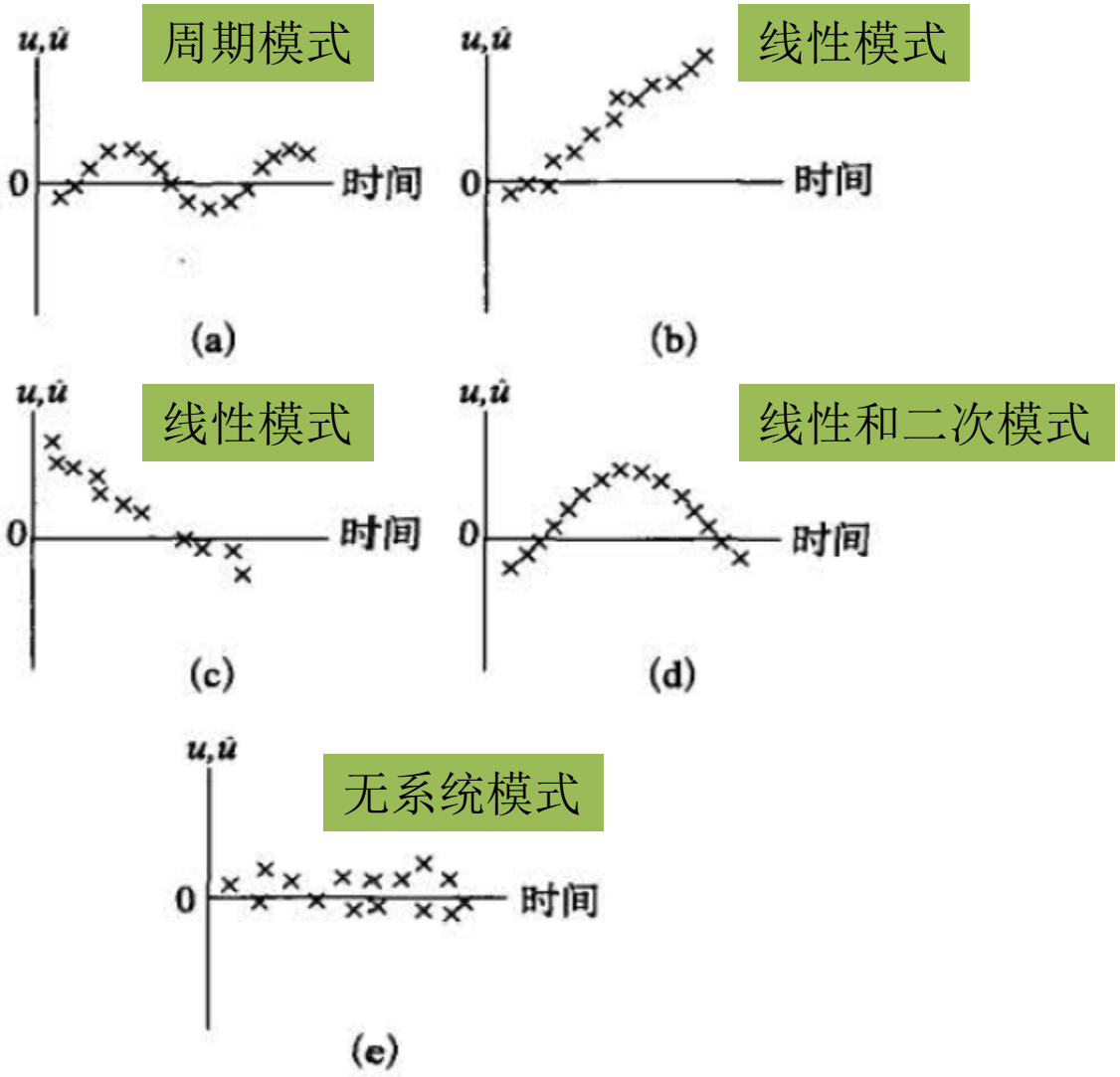


图 12—1　自相关模式与非自相关模式

自相关的性质
——产生自相关的原因

- **原因1：惯性（大多数经济时间序列都有）**
  - GNP、价格指数、生产、就业和失业等时间序列变量都呈现出商业周期。

- **原因2：设定偏误——遗漏重要解释变量**

正确的模型设定形式：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t \qquad (12.1.2)$$

其中：Y_i = 牛肉需求量, X_2 = 牛肉价格, X_3 = 消费者收入, X_4 = 猪肉价格, t = 时间。

实际建模时的模型设定形式（错误的模型设定）：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + v_t \qquad (12.1.3)$$

如果猪肉价格X_4对牛肉需求量Y_t影响明显，则此错误的模型设定中，残差v_t将会呈现出某种系统模式（即猪肉价格X_4的部分影响会体现在残差v_t中）。

自相关的性质
——产生自相关的原因

- **原因3：错误的函数形式**

在成本—产出研究中，正确模型的函数形式为：

$$\text{边际成本}_i = \beta_1 + \beta_2 \text{产出}_i + \beta_3 \text{产出}_i^2 + u_i \qquad (12.1.4)$$

实际建模时的函数形式：

$$\text{边际成本}_i = \alpha_1 + \alpha_2 \text{产出}_i + v_i \qquad (12.1.5)$$

那么，实际总体模型的随机扰动项为：

$$v_i = \beta_3 * \text{产出}_i^2 + u_i$$

最终，样本模型的直线和散点图关系如下：

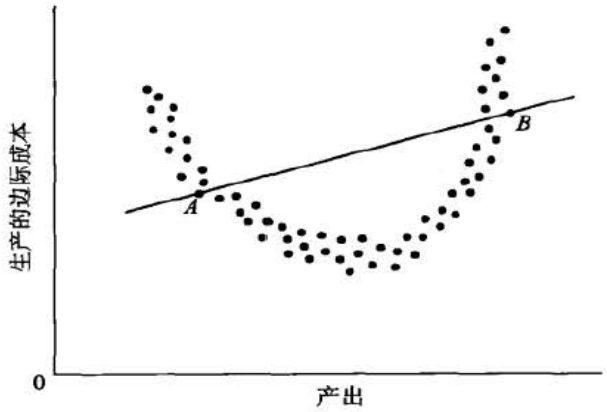


图 12—2　设定偏误：不正确的函数形式

自相关的性质
——产生自相关的原因

- **原因4：蛛网现象（Cobweb phenomenon）**
  - 供给对价格的反应要滞后一个时期

$$\text{供给}_t = \beta_1 + \beta_2 \text{价格}_{t-1} + u_t \qquad (12.1.6)$$

- **原因5：滞后效应**
  - 在消费支出对收入的时间序列回归中，当期消费还会受到前期消费水平的影响：

$$\text{消费}_t = \beta_1 + \beta_2 \text{收入}_t + \beta_3 \text{消费}_{t-1} + u_t \qquad (12.1.7)$$

  - 这种带有因变量的滞后值的回归也叫**自回归**（auto-regression）

自相关的性质
——产生自相关的原因

- **原因6：数据的“编造”**
  - 从月度数据计算得出季度数据，会减小波动，引进匀滑作用，使扰动项出现系统性模式
  - 数据的内插（interpolation）：人口普查10年一次
  - 数据的外推（extrapolation）
- **原因7：数据的变换**
  - 差分：

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1}$$

$$\Delta Y_t = \beta_2 \Delta X_t + \Delta u_t$$

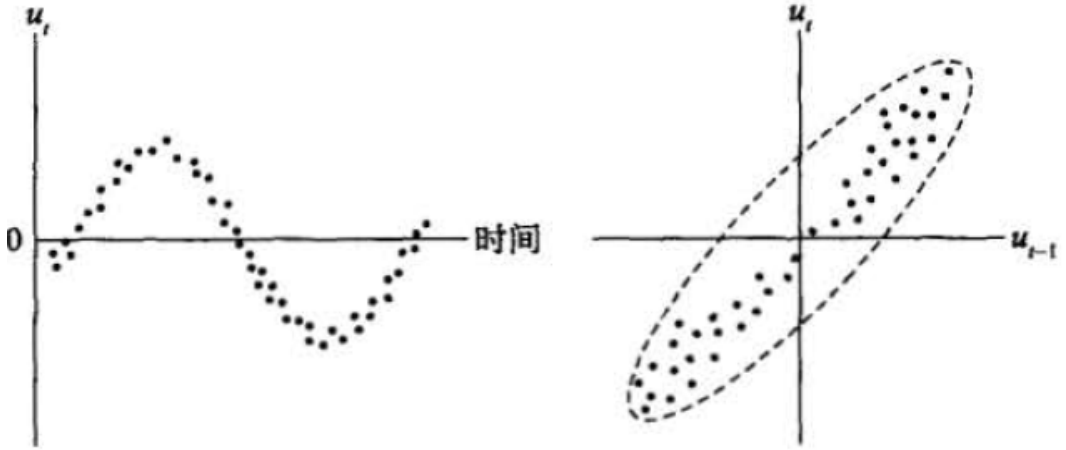
自相关的性质
——产生自相关的原因

- 原因8：时间序列非平稳性：

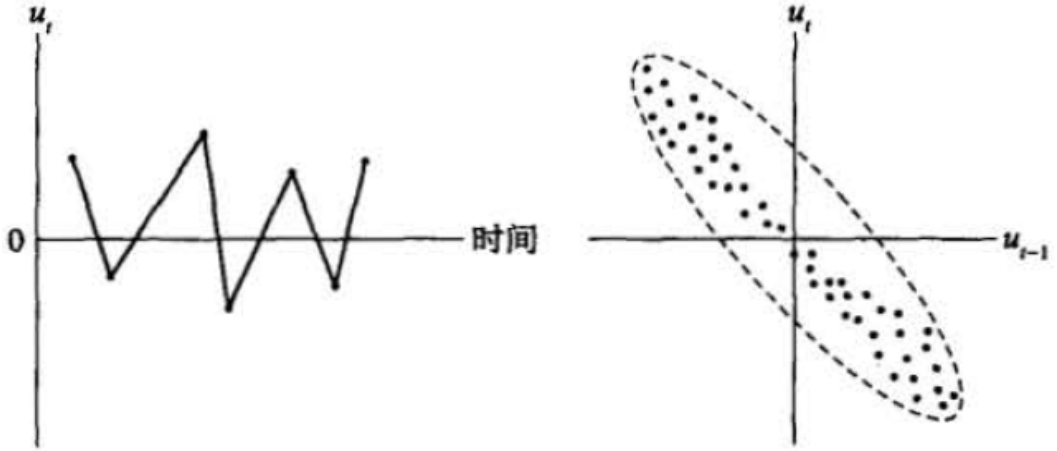
$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

  - Y 和X 很可能都是非平稳的，因此误差u也是非平稳的。此时， 误差项将表现出自相关.

自相关的性质
——自相关时μ_i的形态



(a)正自相关



(b)负自相关

自相关出现时的OLS估计量
——自相关的一种模型：AR(1)模型的形式

如果我们在干扰项中通过假定$E(\mu_t\mu_{t+s}) \neq 0, (s \neq 0)$引进自相关，但保留经典模型的所有其他假定，对OLS 估计量及其方差来说， 会出现什么情况呢?

- 对于双变量回归模型：

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

  - 可具体假定随机干扰项是如下产生的：

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad -1 < \rho < 1$$

马尔可夫一阶自回归模式:AR(1)

    - ρ被称为自协方差系数(coefficient of autocovariance)
    - ε_t是满足以下标准OLS 假定的随机干扰项:

$$E(\varepsilon_t) = 0$$

$$\text{var}(\varepsilon_t) = \sigma^2$$

$$\text{cov}(\varepsilon_t, \varepsilon_{t+s}) = 0, \quad s \neq 0$$

这里ε_t 也成为白噪声误差项(white noise error Term)

自相关出现时的OLS估计量
——自相关的一种模型：AR(1)模型的特征

- 给定AR(1)模式，可以证明：

（式12.2.3）方差：$$var(u_t) = E(u_t^2) = \frac{\sigma_\varepsilon^2}{1-\rho^2}$$

$-1 < \rho < 1$

（式12.2.4）协方差：$$cov(u_t, u_{t+s}) = E(u_t u_{t+s}) = \rho^s \frac{\sigma_\varepsilon^2}{1-\rho^2}$$

（式12.2.5）相关系数：$$cor(u_t, u_{t+s}) = \rho^s$$

$$cov(u_t, u_{t+s}) = cov(u_t, u_{t-s})$$
$$cor(u_t, u_{t+s}) = cor(u_t, u_{t-s})$$

- 若$\rho = 1$， 上述方差和协方差都没有定义。
- 若$|\rho| < 1$，μ_t的均值、方差和协方差都不随时间而变化（弱平稳），即AR (l)过程被称为是（弱）平稳的。此时，协方差的值将随着两个误差的时间间隔越远而越小。

问题：$|\rho| \geq 1$时会发生什么情况？

自相关出现时的OLS估计量
——自相关的一种模型：AR(1)模型的OLS估计量

- 对于双变量回归模型：

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

  - 在经典假设下，我们有：

$$\hat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2} \qquad \mathrm{var}(\hat{\beta}_2)^{\mathrm{OLS}}_{\mathrm{CLRM}} = \frac{\sigma^2}{\sum x_i^2}$$

  - 而在AR(1)假设下，我们可证明：

$$\mathrm{var}(\hat{\beta}_2)^{\mathrm{OLS}}_{\mathrm{AR1}} = \frac{\sigma^2}{\sum x_t^2} + \frac{2\sigma^2}{\sum x_t^2}\left[\rho\frac{\sum_{t=1}^{n-1} x_t x_{t+1}}{\sum_{t=1}^{n} x_t^2} + \rho^2\frac{\sum_{t=1}^{n-2} x_t x_{t+2}}{\sum_{t=1}^{n} x_t^2} + \cdots + \rho^{n-1}\frac{x_1 x_n}{\sum_{t=1}^{n} x_t^2}\right]$$

    - 若假定回归元X服从自相关系数为r的AR(1)过程，则：

$$var(\hat{\beta}_2)^{\mathrm{OLS}}_{\mathrm{AR1}} = \frac{\sigma^2}{\sum x_t^2}\left(\frac{1+r\rho}{1-r\rho}\right) = var(\hat{\beta}_2)^{\mathrm{OLS}}_{\mathrm{CLRM}}\left(\frac{1+r\rho}{1-r\rho}\right)$$

请验证，$r = 0.6, \rho = 0.8$时，两者的关系？！

自相关出现时的BLUE估计量*
——自相关的一种模型：AR(1)模型的GLS估计量

- 对于双变量回归模型，且假定AR(1)过程：

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad -1 < \rho < 1$$

  - 可以证明β_2的BLUE估计量和方差由下式给出：

$$\hat{\beta}_2 \Big|_{AR1}^{GLS} = \frac{\sum_{t=2}^{n}(x_t - \rho x_{t-1})(y_t - \rho y_{t-1})}{\sum_{t=2}^{n}(x_t - \rho x_{t-1})^2} + C$$

C、D表示校正因子，在实际中可以忽略

$$\text{var}\,\hat{\beta}_2 \Big|_{AR1}^{GLS} = \frac{\sigma^2}{\sum_{t=2}^{n}(x_t - \rho x_{t-1})^2} + D$$

GLS 中我们通过变量变换把额外的信息(异方差性或自相关性〉包括到估计程序中去，而在OLS 中我们并不直接考虑这种附加信息。

出现自相关时使用OLS 的后果
——“后果很严重，黎叔很生气！”

在自相关出现时，OLS 估计量仍是线性的、无偏的和一致性的，但不再是有效的(亦即最小方差)。

那么，如果我们继续使用OLS 估计量，我们平常的假设检验程序会遇到什么问题呢?主要有：

- 参数估计不再是有效估计量(亦即不再是方差最小)
- 参数的显著性检验失去意义
- 模型的预测失效

下面分两种情形来讨论：

- 考虑自相关时OLS估计的后果
- 忽视自相关时OLS估计的后果：

出现自相关时使用OLS 的后果
——“将错就错！”（并不聪明的“小聪明”）

- 考虑自相关时OLS估计的后果：问题较严重！
  - 假定AR(1)，我们考虑它，但仍使用OLS方法得到：

$$\text{var}(\hat{\beta}_2)^{\text{OLS}}_{\text{AR1}}$$

  - 该方差公式产生的系列影响：

$$\text{置信区间}_{\left(\text{var}(\hat{\beta}_2)^{\text{OLS}}_{\text{AR1}}\right)} > \text{置信区间}_{\left(\text{var}(\hat{\beta}_2)^{\text{GLS}}_{\text{AR1}}\right)}$$

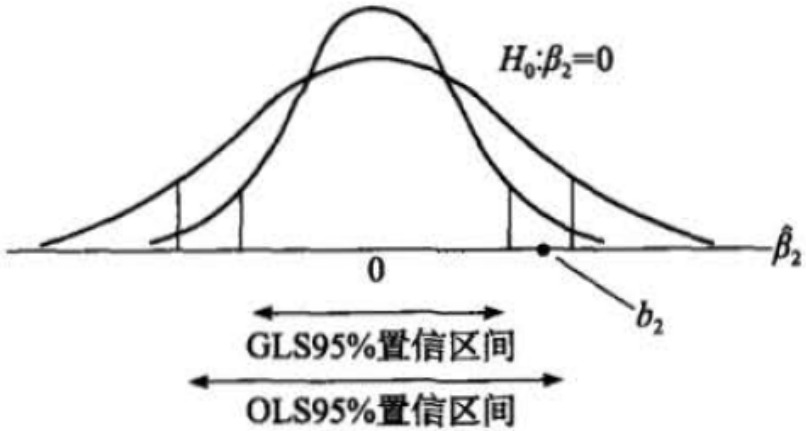


图 12—4 GLS 和 OLS 95%置信区间

启示：尽管OLS 估计量是无偏的和一致性的，但为了构造置信区间并检验假设，要用GLS 而不用OLS！

出现自相关时使用OLS 的后果
——“熟视无睹！”（跟掩耳盗铃无异也！）

- **忽视自相关时OLS估计的后果：问题会更严重！！**
  - OLS估计的残差方差$\hat{\sigma}^2$可能低估了真实方差σ^2；
  - OLS估计的R^2 可能高估了实际的R^2 ；
  - OLS估计的系数方差$\text{var}(\hat{\beta})_{CLRM}^{OLS}$可能低估了实际的系数方差$\text{var}(\hat{\beta})_{AR1}^{GLS}$ ；
  - OLS估计的t检验、F检验无效。

$$r=\frac{\sum_{t=1}^{n-1}x_t x_{t-1}}{\sum_{t=1}^{n}x_t^2}$$

CLRM假设下	自相关AR(1)假设下
$\hat{\sigma}^2=\frac{\sum e_i^2}{(n-2)},\ \mathrm{E}(\hat{\sigma}^2)=\sigma^2$	$\mathrm{E}(\hat{\sigma}^2)=\frac{\sigma^2\left\{n-\left[\frac{2}{(1-\rho)}-2\rho r\right]\right\}}{n-2}<\sigma^2$
$\text{var}(\hat{\beta}_2)_{\text{CLRM}}^{\text{OLS}}=\frac{\sigma^2}{\sum x_i^2}$	$\text{var}\,\hat{\beta}_2\Big\vert_{AR1}^{GLS}=\frac{\sigma^2}{\sum_{t=2}^{n}(x_t-\rho x_{t-1})^2}+D$

感受一下自相关的问题
——例子比较：先构建一个AR(1)

表 12—1　　正自相关误差项的一个假设例子

	ε_t	$u_t=0.7u_{t-1}+\varepsilon_t$
0	0	$u_0=5$（假设的）
1	0.464	$u_1=0.7\times5+0.464=3.964$
2	2.026 2	$u_2=0.7\times3.964+2.026\ 2=4.801\ 0$
3	2.455	$u_3=0.7\times4.801\ 0+2.455=5.815\ 7$
4	−0.323	$u_4=0.7\times5.815\ 7-0.323=3.748\ 0$
5	−0.068	$u_5=0.7\times3.748\ 0-0.068=2.555\ 6$
6	0.296	$u_6=0.7\times2.555\ 6+0.296=2.084\ 9$
7	−0.288	$u_7=0.7\times2.084\ 9-0.288=1.171\ 4$
8	1.298	$u_8=0.7\times1.171\ 4+1.298=2.118\ 0$
9	0.241	$u_9=0.7\times2.118\ 0+0.241=1.723\ 6$
10	−0.957	$u_{10}=0.7\times1.723\ 6-0.957=0.249\ 5$

均值为0 方差为1 的正态随机数

感受一下自相关的问题
——例子比较：查看AR(1)序列图

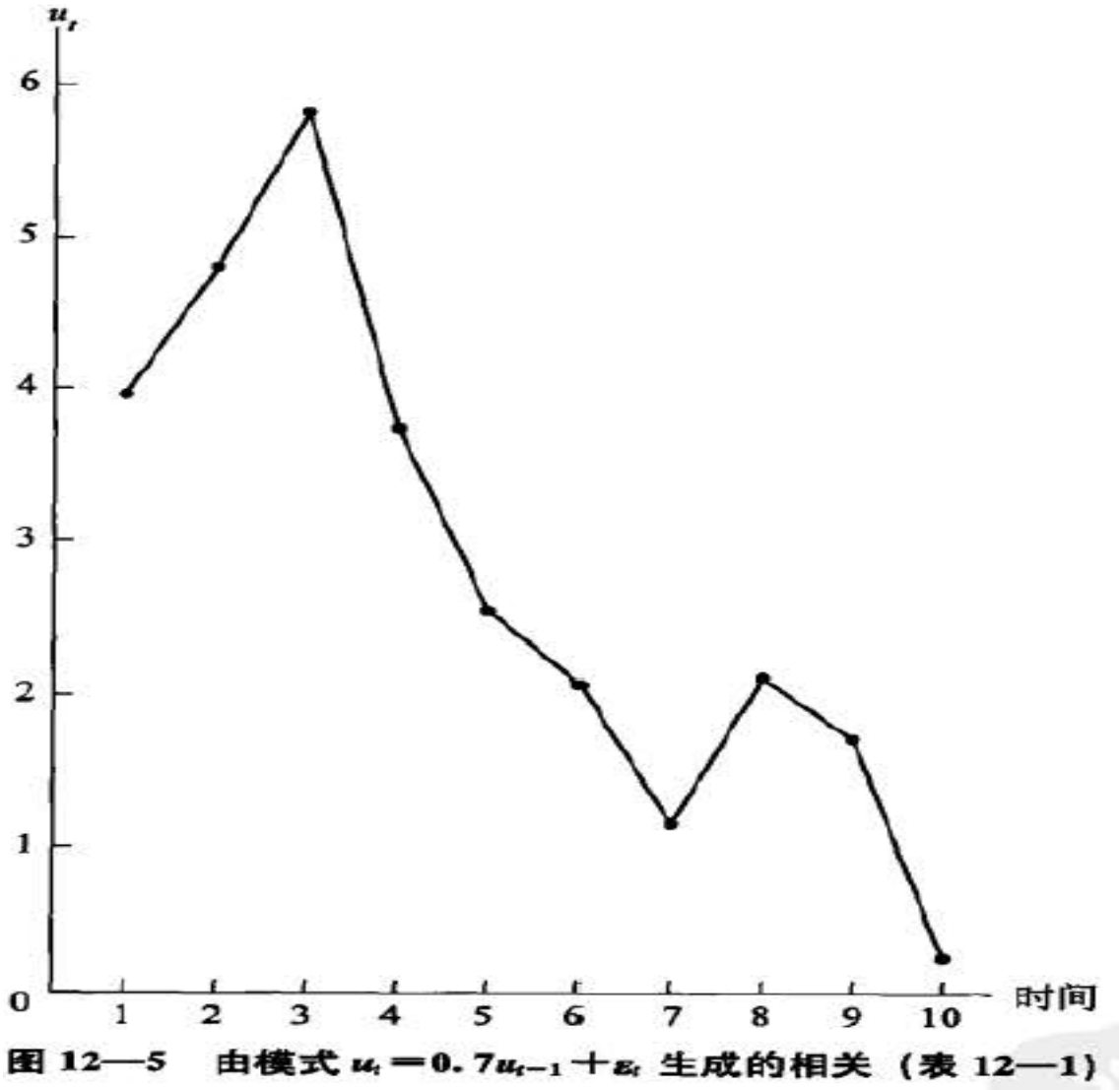


图 12—5　由模式 $u_t = 0.7u_{t-1} + \varepsilon_t$ 生成的相关（表 12—1）

感受一下自相关的问题
——例子比较：AR(1)模型下并进行OLS估计！

表 12—2　　Y 样本值的生成

X_t	u_t	$Y_t=1.0+0.8X_t+u_t$
1	3.964 0	$Y_1=1.0+0.8\times1+3.964\ 0=5.764\ 0$
2	4.801 0	$Y_2=1.0+0.8\times2+4.801\ 0=7.401\ 0$
3	5.815 7	$Y_3=1.0+0.8\times3+5.815\ 7=9.215\ 7$
4	3.748 0	$Y_4=1.0+0.8\times4+3.748\ 0=7.948\ 0$
5	2.555 6	$Y_5=1.0+0.8\times5+2.555\ 6=7.555\ 6$
6	2.084 9	$Y_6=1.0+0.8\times6+2.084\ 9=7.884\ 9$
7	1.171 4	$Y_7=1.0+0.8\times7+1.171\ 4=7.771\ 4$
8	2.118 0	$Y_8=1.0+0.8\times8+2.118\ 0=9.518\ 0$
9	1.723 6	$Y_9=1.0+0.8\times9+1.723\ 6=9.923\ 6$
10	0.249 5	$Y_{10}=1.0+0.8\times10+0.249\ 5=9.249\ 5$

注：u_t 数据来自表 12—1。

$$\hat{Y}_t = 6.545\ 2 + 0.305\ 1\ X_t$$
$$(0.615\ 3)\quad (0.099\ 2)$$
$$t = (10.636\ 6)\quad (3.076\ 3)$$
$$r^2 = 0.541\ 9 \qquad \hat{\sigma}^2 = 0.811\ 4$$

感受一下自相关的问题
——例子比较：AR(1)模型的OLS产生了偏差！

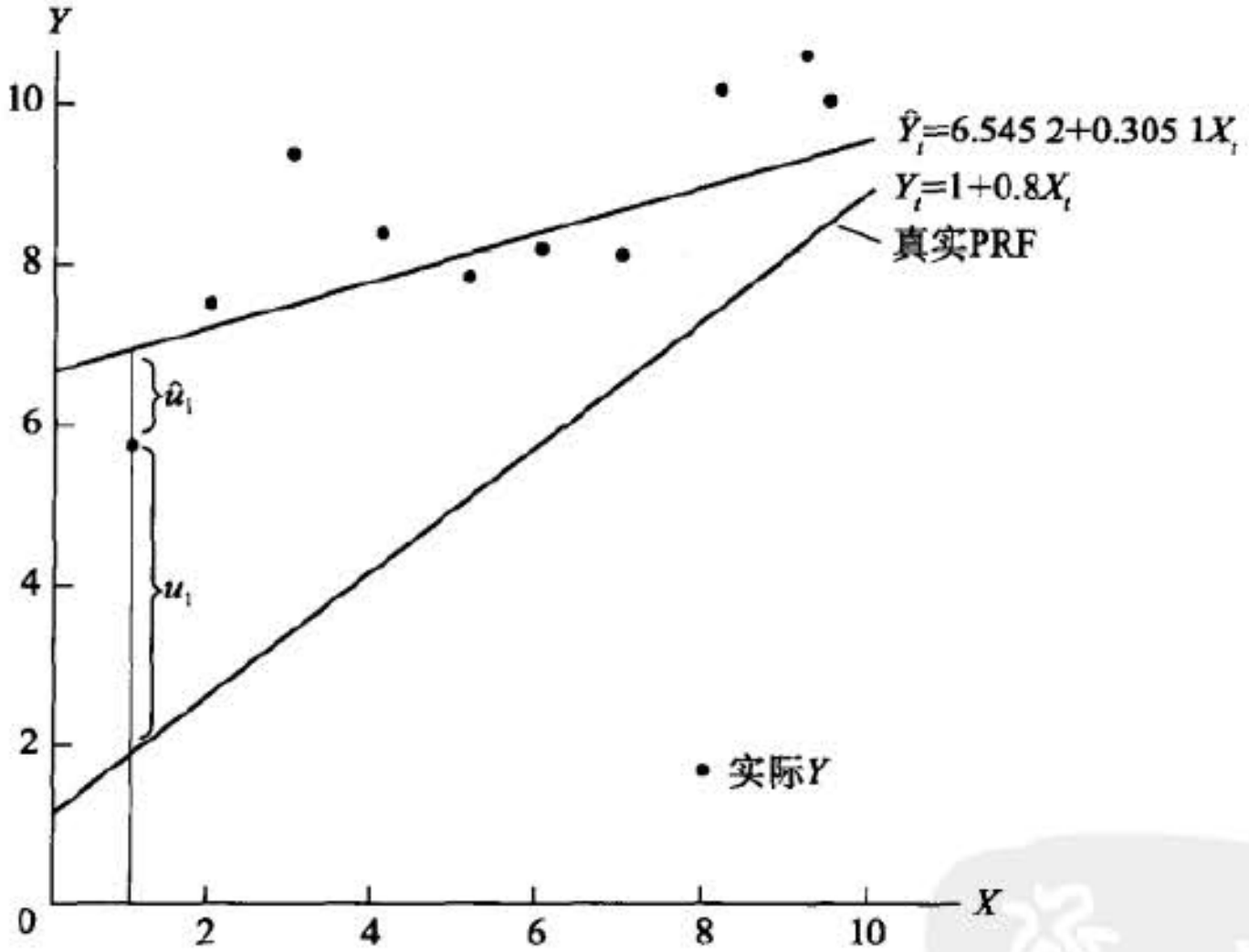


图 12—6　真实 PRF 与表 12—2 所给数据估计的回归线

感受一下自相关的问题
——例子比较：CLRM模型下进行OLS估计！

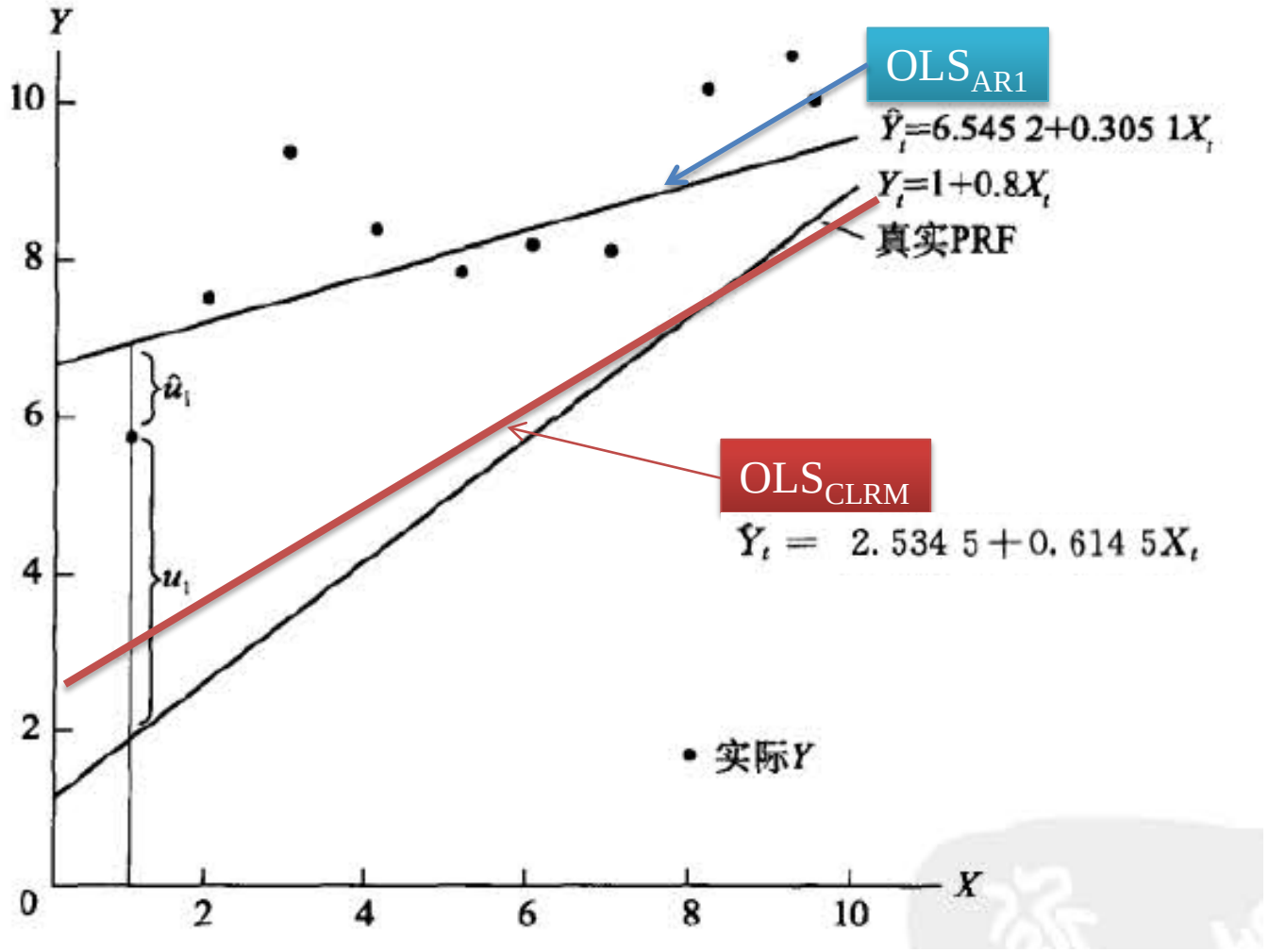
表 12—3　　零序列相关的 Y 样本值

X_t	$\varepsilon_t = u_t$	$Y_t = 1.0 + 0.8X_t + \varepsilon_t$
1	0.464	2.264
2	2.026	4.626
3	2.455	5.855
4	−0.323	3.877
5	−0.068	4.932
6	0.296	6.096
7	−0.288	6.312
8	1.298	8.698
9	0.241	8.441
10	−0.957	8.043

注：因为没有自相关，所以 u_t 和 ε_t 相同。ε_t 来自表 12—1。

$$\hat{Y}_t = 2.534\,5 + 0.614\,5X_t$$
$$(0.679\,6)\quad (0.108\,7)$$
$$t = (3.791\,0)\quad (5.654\,1)$$
$$r^2 = 0.799\,7 \qquad \hat{\sigma}^2 = 0.975\,2$$

感受一下自相关的问题
——例子比较：OLS_{CLRM}比OLS_{AR1}偏差要小



1960-2005年间美国商业部门工资与生产率之间的关系
——数据表

Year	Y	X	Year	Y	X	Year	Y	X
1960	60.8	48.9	1976	86.4	77.1	1992	100.0	100.0
1961	62.5	50.6	1977	87.6	78.5	1993	99.7	100.4
1962	64.6	52.9	1978	89.1	79.3	1994	99.0	101.3
1963	66.1	55.0	1979	89.3	79.3	1995	98.7	101.5
1964	67.7	56.8	1980	89.1	79.2	1996	99.4	104.5
1965	69.1	58.8	1981	89.3	80.8	1997	100.5	106.5
1966	71.7	61.2	1982	90.4	80.1	1998	105.2	109.5
1967	73.5	62.5	1983	90.3	83.0	1999	108.0	112.8
1968	76.2	64.7	1984	90.7	85.2	2000	112.0	116.1
1969	77.3	65.0	1985	92.0	87.1	2001	113.5	119.1
1970	78.8	66.3	1986	94.9	89.7	2002	115.7	124.0
1971	80.2	69.0	1987	95.2	90.1	2003	117.7	128.7
1972	82.6	71.2	1988	96.5	91.5	2004	119.0	132.7
1973	84.3	73.4	1989	95.0	92.4	2005	120.2	135.7
1974	83.3	72.3	1990	96.2	94.4			
1975	84.1	74.8	1991	97.4	95.9			

Y：人均真实工资指数 X：人均产出指数 T=46

1960-2005年间美国商业部门工资与生产率之间的关系
——散点图

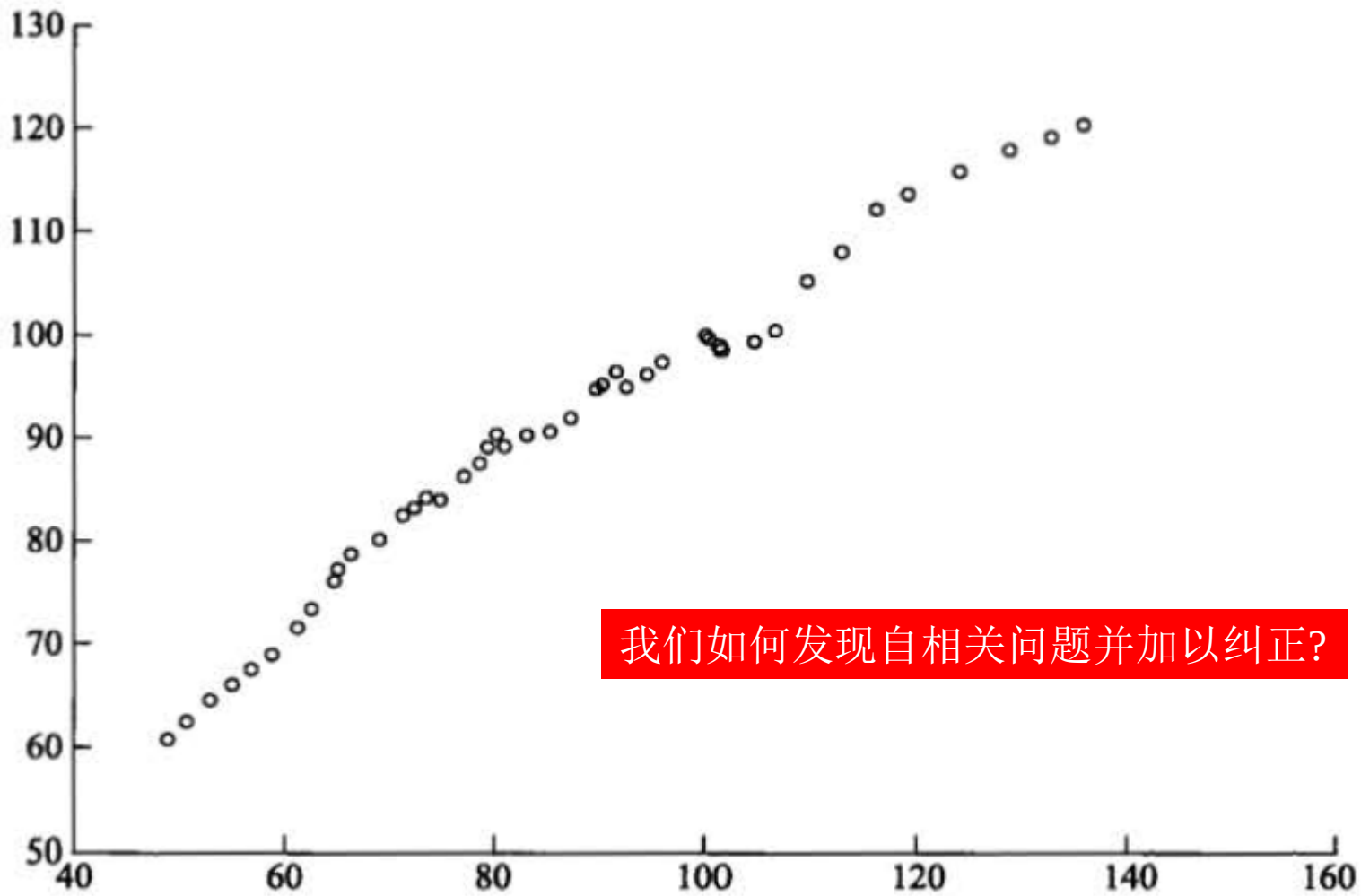


图 12—7　美国工资指数（Y）与生产率指数（X）：1960—2005 年

我们如何发现自相关问题并加以纠正?

1960-2005年间美国商业部门工资与生产率之间的关系
——两个回归模型

- 线性模型：

$$\hat{Y}_t = 32.741\,9 + 0.670\,4\,X_t$$
$$se = (1.394\,0)\ (0.015\,7)$$
$$t = (23.487\,4)(42.781\,3)$$
$$r^2 = 0.976\,5 \qquad d = 0.173\,9 \qquad \hat{\sigma} = 2.384\,5$$

- 对数线性模型：

$$\widehat{\ln Y_t} = 1.606\,7 + 0.652\,2\,\ln X_t$$
$$se = (0.054\,7)\quad (0.012\,4)$$
$$t = (29.368\,0)\ (52.799\,6)$$
$$r^2 = 0.984\,5 \qquad d = 0.217\,6 \qquad \hat{\sigma} = 0.022\,1$$

我们的数据是否受到自相关问题的困扰？

侦察自相关
——图解法：原理

- **图解法**
  - 自相关说明：
    - 经典模型的**无自相关假定**是对总体随机扰动项 μ_t 而言；
    - 对于样本的残差e_t而言，不一定成立；
    - 在实际估计中，通常用样本残差e_t替代总体随机扰动项μ_t，因此，有必要针对样本残差e_t进行自相关检验。
  - 两种图解方式：
    - 图解法1：e_t的时序图（ e_t对时间的散点图）
    - 图解法2：e_t和e_{t-1}的散点图

侦察自相关
——图解法：工资-生产率对数模型残差e_i

表 12—5　　残差的实际值、标准化值及滞后值

观测	S1	SDRES	S1(—1)	观测	S1	SDRES	S1(—1)
1960	—0.036 068	—1.639 433	NA	1983	0.014 416	0.655 291	0.038 719
1961	—0.030 780	—1.399 078	—0.036 068	1984	0.001 774	0.080 626	0.014 416
1962	—0.026 724	—1.214 729	—0.030 780	1985	0.001 620	0.073 640	0.001 774
1963	—0.029 160	—1.325 472	—0.026 724	1986	0.013 471	0.612 317	0.001 620
1964	—0.026 246	—1.193 017	—0.029 160	1987	0.013 725	0.623 875	0.013 471
1965	—0.028 348	—1.288 551	—0.026 246	1988	0.017 232	0.783 269	0.013 725
1966	—0.017 504	—0.795 647	—0.028 348	1989	—0.004 818	—0.219 005	0.017 232
1967	—0.006 419	—0.291 762	—0.017 504	1990	—0.006 232	—0.283 285	—0.004 818
1968	0.007 094	0.322 459	—0.006 419	1991	—0.004 118	—0.187 161	—0.006 232
1969	0.018 409	0.836 791	0.007 094	1992	—0.005 078	—0.230 822	—0.004 118
1970	0.024 713	1.123 311	0.018 409	1993	—0.010 686	—0.485 739	—0.005 078
1971	0.016 289	0.740 413	0.024 713	1994	—0.023 553	—1.070 573	—0.010 686
1972	0.025 305	1.150 208	0.016 289	1995	—0.027 874	—1.266 997	—0.023 553
1973	0.025 829	1.174 049	0.025 305	1996	—0.039 805	—1.809 304	—0.027 874
1974	0.023 744	1.079 278	0.025 829	1997	—0.041 164	—1.871 079	—0.039 805
1975	0.011 131	0.505 948	0.023 744	1998	—0.013 576	—0.617 112	—0.041 164
1976	0.018 359	0.834 515	0.011 131	1999	—0.006 674	—0.303 364	—0.013 576
1977	0.020 416	0.927 990	0.018 359	2000	0.010 887	0.494 846	—0.006 674
1978	0.030 781	1.399 135	0.020 416	2001	0.007 551	0.343 250	0.010 887
1979	0.033 023	1.501 051	0.030 781	2002	0.000 453	0.020 599	0.007 551
1980	0.031 604	1.436 543	0.033 023	2003	—0.006 673	—0.303 298	0.000 453
1981	0.020 801	0.945 516	0.031 604	2004	—0.015 650	—0.711 380	—0.006 673
1982	0.038 719	1.759 960	0.020 801	2005	—0.020 198	—0.918 070	—0.015 650

注：S1＝工资一生产率对数线性回归中得到的残差。
S1(—1)　一滞后一期的残差值。
SDRES＝标准化残差＝残差/估计值的标准误。

侦察自相关
——图解法1：工资-生产率对数模型的残差e_t序列图

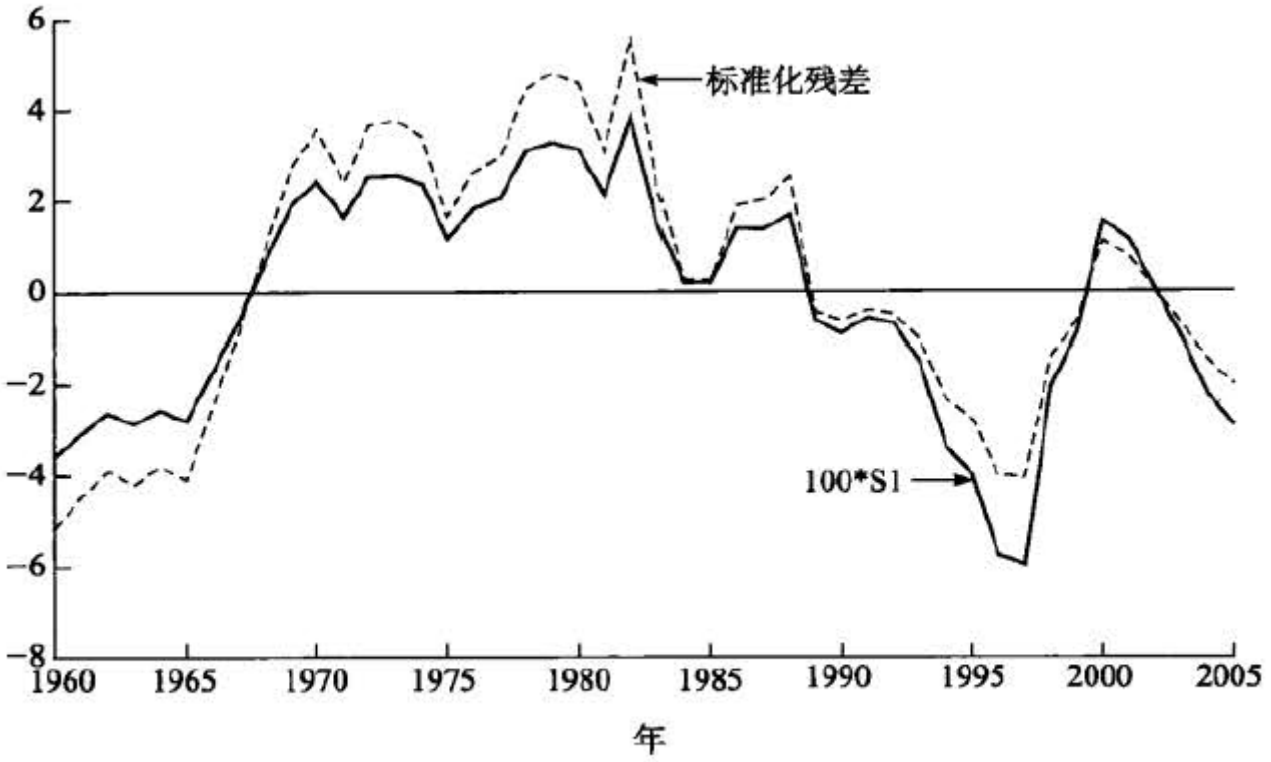


图 12—8　工资—生产率对数线性回归［模型（12.5.2)］中的残差（放大 100 倍）及标准化残差

提示：标准化残差服从均值为0，方差为1的标准正态分布；
而图中的标准化残差的方差明显大于1，这表明残差并非完全随机的，即可能存在某种相关性（比如AR(1)等）

侦察自相关

——图解法2：工资-生产率对数模型残差的$(e_t,\ e_{t-1})$散点图

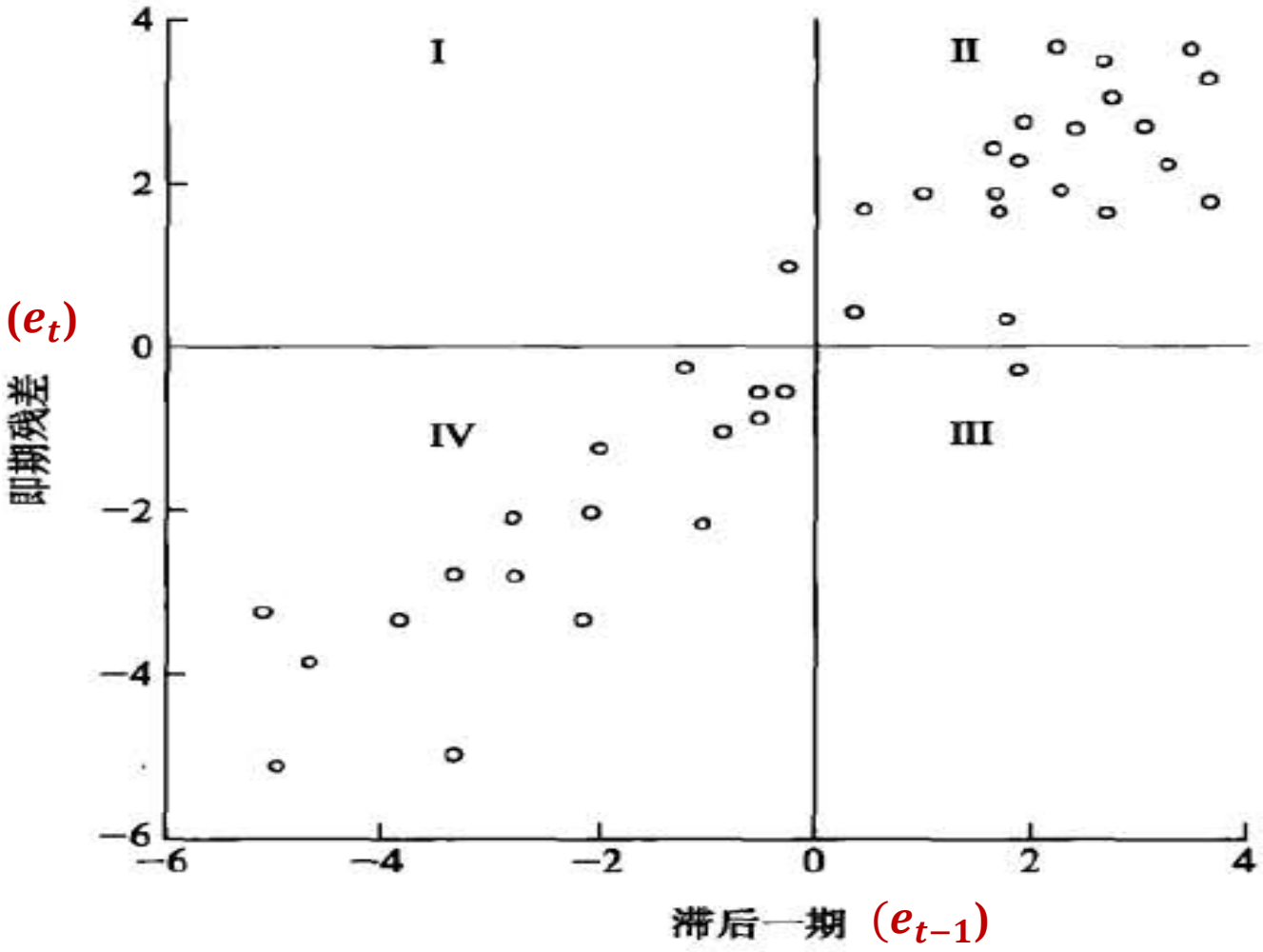


图 12—9　即期残差与滞后残差

侦察自相关
——Durbin-Watson d 检验：d统计量公式

（式12.6.5）

- **德宾-沃森d统计量(Durbin-Watson d statistic)：**

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

  - 它其实是用相继残差的差异的平方和与RSS之比
  - 由于取相继差异时损失一个观测值，d 统计量的分子只有n-1次观测值

侦察自相关
——Durbin-Watson d 检验：使用条件

- 使用d 统计量的条件：
  - 回归模型**含有截距项**，如果没有截距项，就必须重新做带有截距的回归
  - 诸解释变量**X是非随机的**，或者说，在反复抽样中是被固定的
  - 随机扰动项u_t是按AR(1)模式产生的：

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

（式12.6.6）

  - 被解释变量的滞后值(如Y_{t-1})**不能**当作解释变量。
    即如下模型的D-W检验无效：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt} + \gamma Y_{t-1} + u_t$$

  - **没有**数据缺损——如果数据缺失，d 统计量无法补偿

侦察自相关
——Durbin-Watson d 检验：判断准则

（式12.6.7）

$$d=\frac{\sum e_t^2+\sum e_{t-1}^2-2\sum e_t e_{t-1}}{\sum e_t^2}$$

$$\sum e_t^2 \approx \sum e_{t-1}^2$$

（式12.6.8）

$$\approx 2\left(1-\frac{\sum e_t e_{t-1}}{\sum e_t^2}\right)$$

样本一阶自相关系数 $\hat{\rho}=\frac{\sum e_t e_{t-1}}{\sum e_t^2}$

（式12.6.10）

$$=2(1-\hat{\rho})$$

已知：　$-1<\rho<1$　（12.2.1）

故有：　$0<d<4$

（式12.6.11）

注：课本第434页，公式12.6.11 表示有误。

具体而言，几个特殊取值的含义：

$\hat{\rho}=+1$，则$d=0$，残差中存在完全正序列相关

$\hat{\rho}=0$，则$d=2$，没有（一阶）序列相关

$\hat{\rho}=-1$，则$d=4$，残差中存在完全的负相关

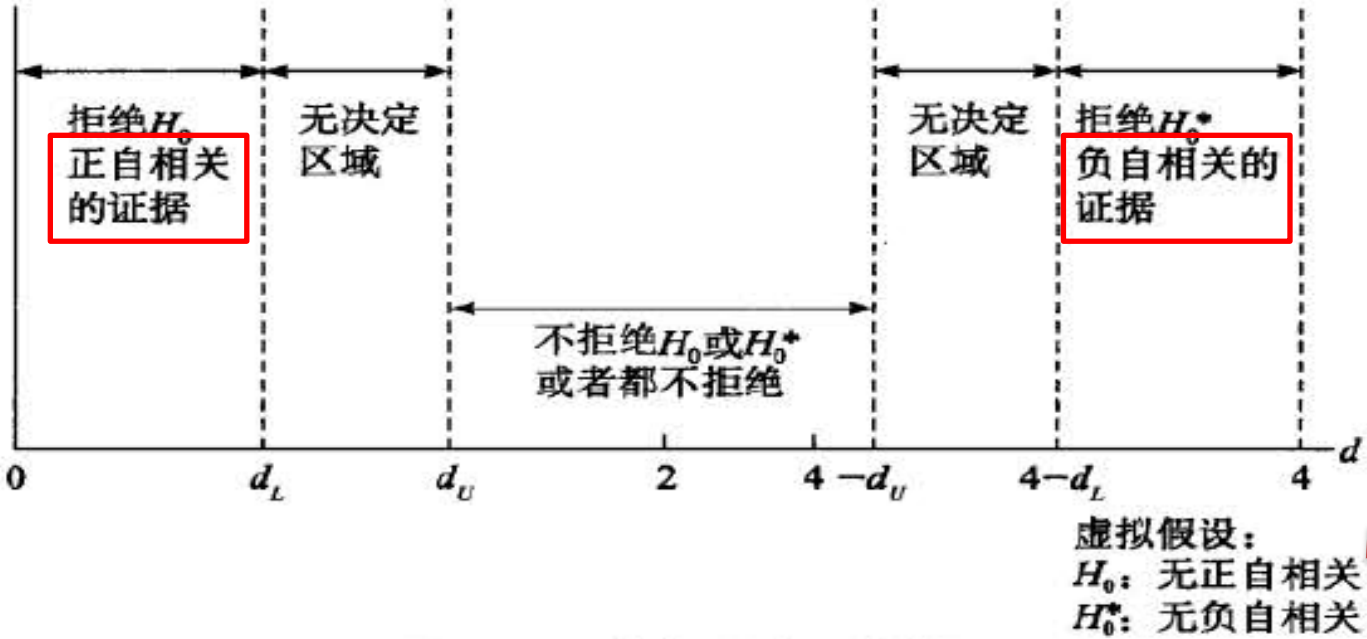


图 12—10　德宾-沃森 d 统计量

表 12—6　　德宾-沃森 d 检验：决策规则

虚拟假设	决策	如果
无正自相关	拒绝	$0<d<d_L$
无正自相关	无决定	$d_L \leqslant d \leqslant d_U$
无负自相关	拒绝	$4-d_L<d<4$
无负自相关	无决定	$4-d_U \leqslant d \leqslant 4-d_L$
无正或负自相关	不拒绝	$d_U<d<4-d_U$

侦察自相关
——Durbin-Watson d 检验：判断准则

- 步骤：
  - 做OLS 回归并获取残差。
  - 按方程(12.6.5)计算d(计算机程序大都给出d 值)
  - 对给定样本容量和解释变量个数下找出临界值d_L 和d_U
  - 按照表12-6 的决策规则行事。

$$\widehat{\ln Y_t} = 1.606\,7 + 0.652\,2\ln X_t$$
$$se = (0.054\,7) \quad (0.012\,4)$$
$$t = (29.368\,0) \quad (52.799\,6)$$
$$r^2 = 0.984\,5 \qquad d = 0.217\,6 \qquad \hat{\sigma} = 0.022\,1$$

N=46，1 个解释变量X，在5%水平上， $d_L = 1.475$ 和$d_U = 1.566$ 。

查表方法：课本第894页，附录：表D-5A $k' = 1$, n=45

侦察自相关
——拉格朗日乘数检验：基本原理

- **拉格朗日乘数检验(LM test)**，也称为布罗施-戈弗雷检验(Breusch-Goldfrey，BG test)
  - 对于双变量模型：$Y_t = \beta_1 + \beta_2 X_t + \mathrm{u}_t$
  - 假设u_t服从如下p阶自回归AR(p)模式：

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

  - BG检验的**原假设**：$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0$
- 优点：
  - 允许被解释变量滞后值(如Y_{t-1})作为解释变量；
  - 可以对随机扰动项u_t的高阶自相关，如AR(p)进行检验；

    当p=1时，BG检验也称Durbin's M test.
  - 允许u_t是白噪音ϵ_t的高阶移动平均，如MA(q)

    $$u_t = \epsilon_t + \lambda_1 \epsilon_{t-1} \epsilon_t + \lambda_2 \epsilon_{t-2} + \cdots + \lambda_q \epsilon_{t-q}$$

- 缺点：随机扰动项滞后期p不能先验设定。

  解决：赤池(Akaike)和施瓦茨(Schwarz)信息准则

（式12.6.6）

侦察自相关
——拉格朗日乘数检验(LM test)：步骤

- 检验步骤：
  - OLS估计回归方程，得到残差e_t
  - 作如下辅助回归，并得到R^2

$$e_t = \beta_1 + \beta_2 X_t + \varepsilon_t$$

  - 若样本容量很大，则有：

$$LM = (n-p)R^2 \sim \chi_p^2$$

如果：LM统计量大于临界值，就拒绝原假设，表明存在自相关。

自相关的补救措施
——自相关的结构已知：广义最小二乘法(GLS)

- **自相关的结构已知（ρ已知）**
  - 如果已知总体残差遵循一阶自回归方式：

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

  - 当自相关系数ρ为已知时，序列相关便可解决：

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1}$$

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2 X_t - \rho\beta_2 X_{t-1} + (u_t - \rho u_{t-1})$$

$$= \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t \quad \text{（式12.9.5）}$$

广义差分方程！

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t$$

由于ε_t满足所有OLS条件，变换后的模型就可通过OLS方法得到BLUE估计量。这实际上就是GLS！

$$Y_t^* = (Y_t - \rho Y_{t-1})$$
$$\beta_1^* = \beta_1(1-\rho)$$
$$X_t^* = (X_t - \rho X_{t-1})$$

自相关的补救措施
——自相关的结构未知：广义差分法(GLS)

➢ **自相关的结构未知（ρ未知）**

因为ρ落在-1到+1之间，下面的**广义差分方程**可以变换

$$(Y_t-\rho Y_{t-1})=\beta_1(1-\rho)+\beta_2(X_t-\rho X_{t-1})+\varepsilon_t$$

- 当ρ=+1时，广义差分方程便化为一阶差分方程：

$$\begin{aligned}Y_t-Y_{t-1}&=\beta_2(X_t-X_{t-1})+(u_t-u_{t-1})\\&=\beta_2(X_t-X_{t-1})+\varepsilon_t\\\Delta Y_t&=\beta_2\Delta X_t+\varepsilon_t\end{aligned}$$

- 当ρ=-1时，广义差分方程将变为：

$$Y_t+Y_{t-1}=2\beta_1+\beta_2(X_t+X_{t-1})+\varepsilon_t$$

$$\frac{Y_t+Y_{t-1}}{2}=\beta_1+\beta_2\frac{X_t+X_{t-1}}{2}+\frac{\varepsilon_t}{2}$$

- 这个模型叫做（2时期）移动平均回归

自相关的补救措施
——自相关的结构未知：基于DW d统计量法估计ρ

- **自相关的结构未知（ ρ未知）**
  - 利用Durbin-Watson d统计量估计ρ
    - 已经有下式：

$$d \approx 2(1-\hat{\rho})$$

$$\hat{\rho} \approx 1-\frac{d}{2}$$

    - 先从(12.9.13)估计出 ρ，即可按照(12.9.5)那样转换数据，然后进行平常的OLS估计

自相关的补救措施
——自相关的结构未知：基于残差中估计出来的ρ

- 自相关的结构未知（ ρ未知）
  - 基于残差中估计出来的ρ
    - 例如，假设AR(1)过程：

$$u_t = \rho_1 u_{t-1} + v_t$$

$$e_t = \hat{\rho}_1 e_{t-1} + \varepsilon_t$$

- 工资-生产率模型的估计结果如下：

$$e_t = 0.8678 e_{t-1}$$

$$t = (12.735\ 9) \qquad r^2 = 0.786\ 3$$

- 从而估计出来的$\hat{\rho} = 0.8678$

自相关的补救措施
——自相关的结构未知：基于迭代方法估计ρ

- **自相关的结构未知（ ρ未知）**
  - 基于迭代方法估计ρ
    - 科克伦–奥克特迭代法(Cochrane-Orcutt iterative procedure)
    - **科克伦–奥克特两步法(Cochrane-Orcutt two-step procedure)**
    - 德宾两步法(Durbin two-step procedure)
    - 希尔德雷思–卢扫描或搜寻程序(Hildreth-Lu scanning or search procedure) 等

自相关的补救措施
——自相关的结构未知：一些点评

- **自相关的结构未知（ ρ未知）**
  - 第一，由于在**大样本**情况下，即便存在自相关问题，OLS估计量仍是一致的，所以无论我们是从德宾-沃森d、从当期残差对前期残差的回归，还是从科克伦-奥克特迭代程序中估计ρ，都没有多大差别，因为这些方法也都是给出真实ρ的一致估计值。
  - 第二，上述方法基本上都是**两步法**。我们在第一步得到未知ρ的一个估计值，第二步用这个估计值变换变量去估计广义差分方程（实质上就是GLS）。但由于我们用的是$\hat{\rho}$而非真正的ρ，所以在文献中所有这些估计方法都被称为可行GLS (feasible GLS, FGLS) 或估计GLS (estimated GLS, EGLS)。
  - 第三，重要的是要指出，只要我们用FGLS 或EGLS 估计变换模型的参数，估计系数都不一定具有通常经典模型所具有的优良性质(比如BLUE)，特别是在小样本情况下。