第8章　协方差分析模型

吴密霞

北京工业大学统计与数据科学系

E-mail:　*wumixia@bjut.edu.cn*

教材

- 吴密霞, 王松桂. 2024. 线性模型引论（第2版）, 科学出版社.

本章目录

- 协方差分析的基本概念
- 参数估计
- 检验问题
- 案例

协方差分析的基本概念

- 协方差分析对误差项方差的影响

  如在研究三种促销策略对某种苏打饼干销量的影响试验中, 我们选择15个商店, 每个商店被随机指派采用其中一种促销策略, 每种促销策略指派5家商店. 三种促销策略分别为

  - 策略1: 常规货架空间, 顾客可在店内免费品尝;
  - 策略2: 在常规位置增加货架空间;
  - 策略3: 除常规货架空间外,在过道两端设置特殊陈列架.

  为了可比性, 要求各商店在促销期间该苏打饼干的售价和广告等相关条件相同.

但促销前各商店该苏打饼干的销量x_{ij}无法控制相等

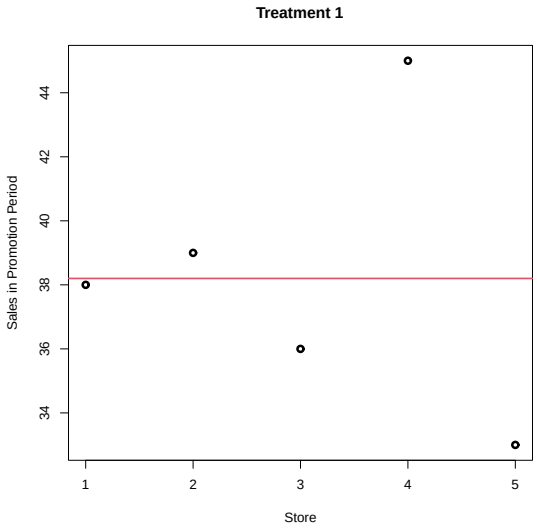
协方差分析的基本概念

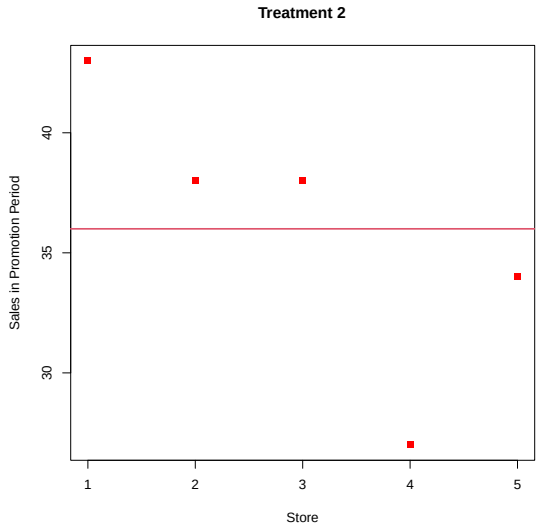
各商店的该苏打饼干的促销前和促销期间的销量

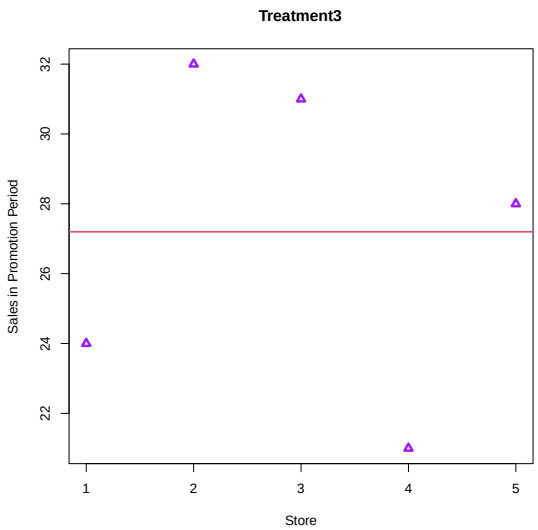
Table: 苏打饼干销量

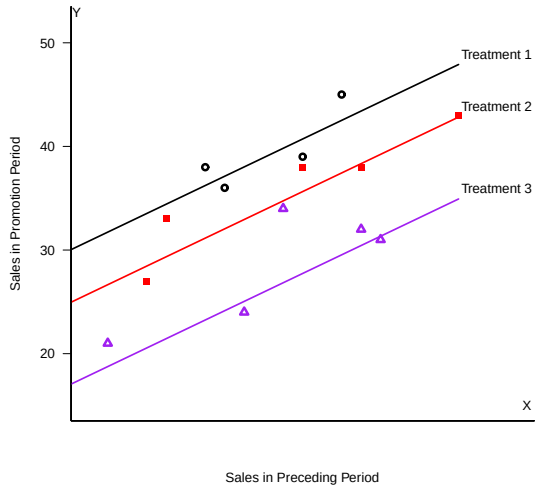
促销	商店									
策略	1		2		3		4		5	
i	y_{i1}	x_{i1}	y_{i2}	x_{i2}	y_{i3}	x_{i3}	y_{i4}	x_{i4}	y_{i5}	x_{i5}
1	38	21	39	26	36	22	45	28	33	19
2	43	34	38	26	38	29	27	18	34	25
3	24	23	32	29	31	30	21	16	28	29

协方差分析的基本概念


Treatment 1
Sales in Promotion Period
Store
34
36
38
40
42
44
1
2
3
4
5


Treatment 2
Sales in Promotion Period
Store
30
35
40
1
2
3
4
5


Treatment3
Sales in Promotion Period
Store
22
24
26
28
30
32
1
2
3
4
5


Y
X
Sales in Promotion Period
Sales in Preceding Period
Treatment 1
Treatment 2
Treatment 3
20
30
40
50

伴随变量的选择

- 伴随变量的选择

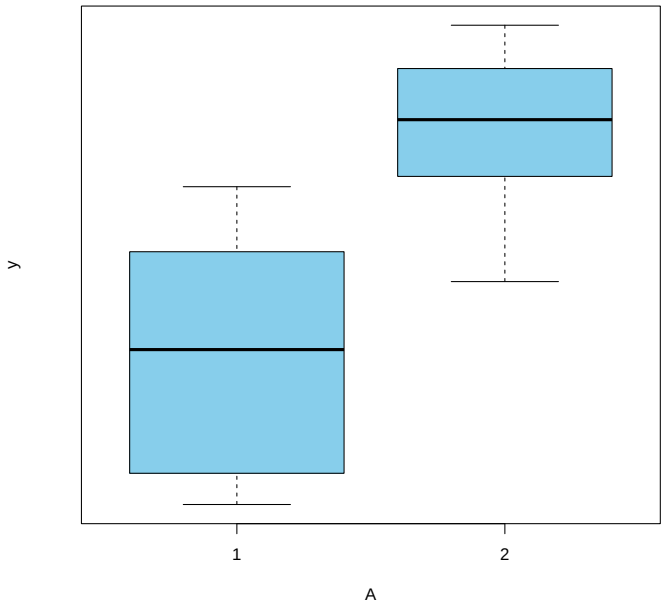
1) 伴随变量通常包括研究对象前期的状态变量, 如被研究对象的年龄、社会经济地位、事先的态度等. 当零售店为研究对象时, 伴随变量可能是试验前的销售额或员工人数.
2) 伴随变量不受不同处理或方案的影响. 为了清楚地解释研究结果, 伴随变量应该是在研究前已被观测或者虽然是在研究期间进行观察，但该变量应该不受试验处理或方案的任何影响.

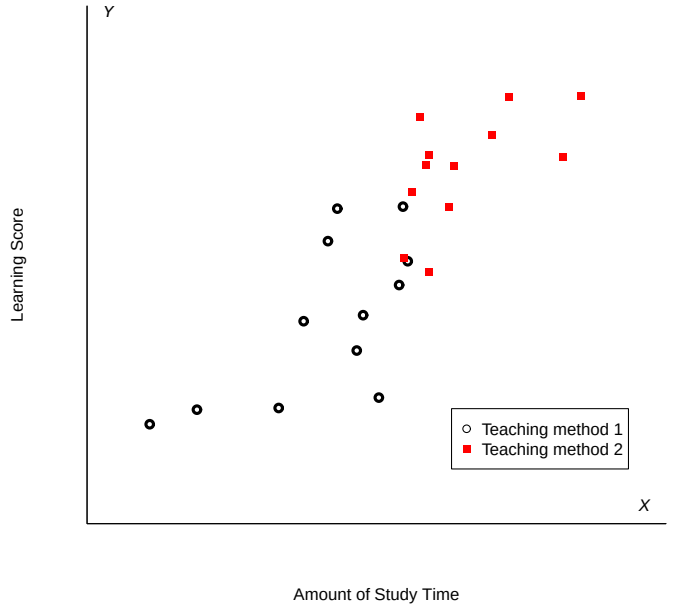
例8.1.1中该苏打饼干促销前各商店的销售量就不受后面研究中三种促销方案的影响.

伴随变量的选择

- 错选伴随变量的影响

  例如一家公司正在为工程师开办一所培训学校, 主要培训内容是会计和预算原理. 为了对比两种教学方法的培训效果, 将学员们随机分配到两种方法中的一种, 课程结束时, 每位学员都会得到一个反映学习程度的分数. 从各组随机抽取12名学员的成绩.





伴随变量的选择

- 两组教学方法下学员的成绩有显著差异.
- 方差分析的结果证实第二种培训方法更优.
- 将学习时间作为伴随变量引入模型后, 协方差分析结果显示培训方法对培训成绩几乎没有影响

方差分析和协方差分析的分析结果为什么矛盾?

因为学员学习时间和因变量（学员成绩）都受到了培训方法的影响, 可以看出分配到第二种培训方法组的学员普遍比分配给第一组的学习时间长和成绩高.

伴随变量不受处理方法的影响，否则可能会导致严重误导性结论.

协方差分析模型

- 考虑一般的协方差分析模型

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{e}, \quad \boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

这里, $\boldsymbol{y}$为$n \times 1$观测向量,

$\mathbf{X}\boldsymbol{\beta}$为模型的方差分析部分, $\mathbf{Z}\boldsymbol{\gamma}$为模型的回归部分,

$\mathbf{X} = (x_{ij})$为$n \times p$已知矩阵, 其元素x_{ij}皆为0或1

$\mathbf{Z} = (z_{ij})$为$n \times q$已知矩阵, 其元素(z_{ij})可以取任意实数值.

$\boldsymbol{\beta}$为因子效应向量, $\boldsymbol{\gamma}$为$q \times 1$的回归系数向量.

假定$\mathbf{Z}$是列满秩, 并且$\mathbf{Z}$的列与$\mathbf{X}$的列线性无关, 即

$$\mathcal{M}(\mathbf{X}) \cap \mathcal{M}(\mathbf{Z}) = \{\mathbf{0}\}, \quad \mathrm{rk}(\mathbf{Z}) = q.$$

参数估计

- 应用分块LS估计（6.2节），得γ 和可估函数$c'\beta$的BLU估计:

$$\gamma^* = (\mathbf{Z}'\mathbf{N}_\mathbf{X}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{N}_\mathbf{X}\boldsymbol{y},$$

$$\boldsymbol{c}'\boldsymbol{\beta}^* = \boldsymbol{c}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}\gamma^* = \boldsymbol{c}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}'\mathbf{X}_\mathbf{Z}\gamma^*,$$

其中$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\boldsymbol{y}$, $\mathbf{X}_\mathbf{Z} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}$, $\mathbf{N}_\mathbf{X} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$.

$$\mathrm{Cov}(\gamma^*) = \sigma^2(\mathbf{Z}'\mathbf{N}_\mathbf{X}\mathbf{Z})^{-1},$$

$$\mathrm{Var}(\boldsymbol{c}'\boldsymbol{\beta}^*) = \sigma^2[\boldsymbol{c}'(\mathbf{X}'\mathbf{X})^-\boldsymbol{c} + \boldsymbol{c}'\mathbf{X}_\mathbf{Z}(\mathbf{Z}'\mathbf{N}_\mathbf{X}\mathbf{Z})^{-1}\mathbf{X}_\mathbf{Z}\boldsymbol{c}].$$

注 关于方差分析部分, 常采用虚拟变量法,将设计阵转成可逆矩阵.

假设检验

- 假设检验 $H_{01}: \mathbf{H}\boldsymbol{\beta} = \mathbf{0}$.

  F统计量为

$$F_1 = \frac{(\mathrm{SS}^*_{\mathbf{H}_e} - \mathrm{SS}^*_e)/m}{\mathrm{SS}^*_e/(n-r-q)},$$

  其中, $\mathrm{SS}^*_{\mathbf{H}_e}$ 和SS^*_e 分别在H_{01}下模型的误差平方和与原模型的误差平方和,

$$\mathrm{SS}^*_{\mathbf{H}_e} = (\boldsymbol{y} - \mathbf{X}'\boldsymbol{\beta}^*_H - \mathbf{Z}\boldsymbol{\gamma}^*_H)'(\boldsymbol{y} - \mathbf{X}'\boldsymbol{\beta}^*_H - \mathbf{Z}\boldsymbol{\gamma}^*_H)$$

$$\mathrm{SS}^*_e = (\boldsymbol{y} - \mathbf{X}'\boldsymbol{\beta}^* - \mathbf{Z}\boldsymbol{\gamma}^*)'(\boldsymbol{y} - \mathbf{X}'\boldsymbol{\beta}^* - \mathbf{Z}\boldsymbol{\gamma}^*)$$

  当$\mathbf{H}\boldsymbol{\beta} = \mathbf{0}$为真时, $F_1 \sim F_{m,\, n-r-q}$, 这里$r = \mathrm{rk}(\mathbf{X})$, $m = \mathrm{rk}(\mathbf{H})$.

假设检验

- 假设检验$H_{02}: \mathbf{H}\boldsymbol{\beta} = \mathbf{0}$

  H_{02}成立, 模型就变为纯方差分析模型, 记SS_e为相应误差平方和.

  F 统计量

$$F_2 = \frac{(\mathrm{SS}_e - \mathrm{SS}_e^*)/q}{\mathrm{SS}_e^*/(n-r-q)}. \tag{1.1}$$

  当$\gamma = 0$ 成立时, $F_2 \sim F_{q,n-r-q}$. 故H_{02}的拒绝域为

$$F_2 > F_{q,n-r-q}(\alpha).$$

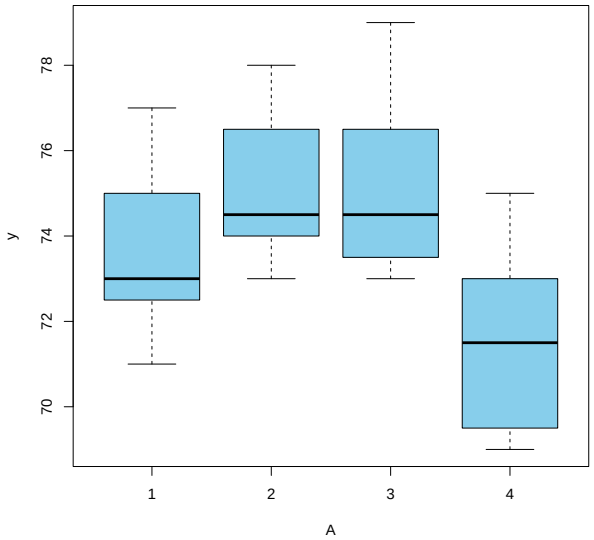
  如果经检验, 假设$\gamma = 0$被接受, 则可以认为协变量的影响不存在, 我们只要研究纯方差分析模型就够了.

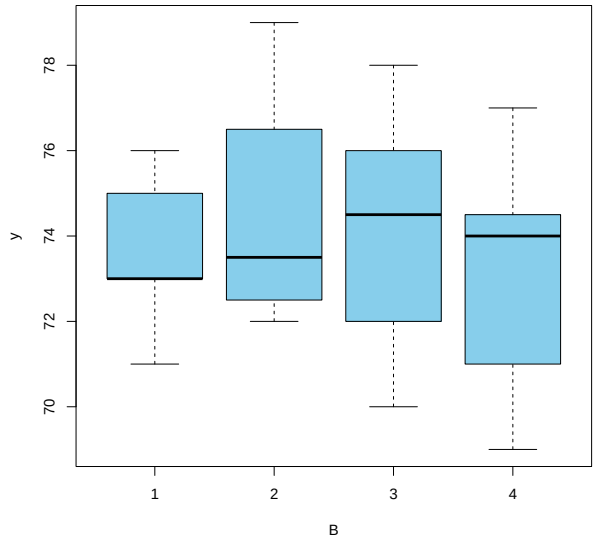
例8.4.2

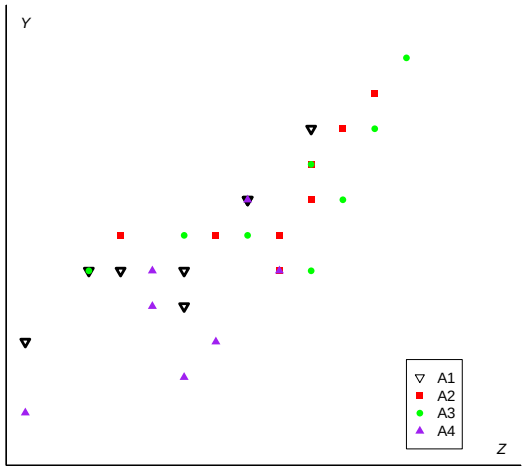
在化学纤维生产中影响化纤弹性(Y) 的因素有收缩率A和总拉伸倍数B. 对A, B各取四个水平进行试验, 各个试验重复1次. 由于试验中电流周波(Z)不能完全控制, 把它作为协变量, 试验数据如表

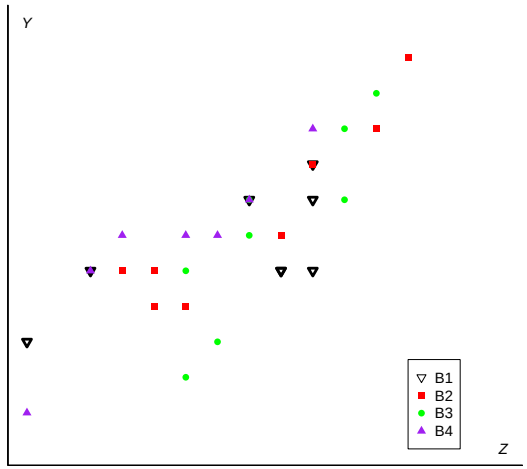
变量			伸缩率 A_1		A_2		A_3		A_4	
总拉伸倍数	B_1	Z	49.0	49.2	49.8	49.9	49.9	49.9	49.7	49.8
		y	71	73	73	75	76	73	75	73
	B_2	Z	49.5	49.3	49.9	49.8	50.2	50.1	49.4	49.4
		y	72	73	76	74	79	77	73	72
	B_3	Z	49.7	49.5	50.1	50.0	49.7	50.0	49.5	49.6
		y	75	73	78	77	74	75	70	71
	B_4	Z	49.9	49.7	49.6	49.3	49.5	49.2	49.0	48.9
		y	77	75	74	74	74	73	69	69

案例分析


y
78
76
74
72
70
1
2
3
4
A


y
78
76
74
72
70
1
2
3
4
B


Y
Z
A1
A2
A3
A4


Y
Z
B1
B2
B3
B4

假设检验

图显示:

- 因子A 的不同水平下y的分布有明显的差异,尤其是第四水平下, y的取值普遍比较低, 但因子B的4水平下y的Box图重叠部分较多, 中位数较靠近;
- (Z, Y)散点图都呈现一定线性趋势, 且在因子不同水平下散点的分布大致没有太大差异.

  因此, 初步可认为收缩率因子A和电流周波Z对化纤弹性y有影响, 后者的影响是线性的, 而总拉伸倍数因子B的影响可能不大.

案例分析

需对如下假设作严格检验.

- 收缩率(因子A)的显著性检验

  $H_{01}:\alpha_1=\alpha_2=\alpha_3=\alpha_4=0 \longleftrightarrow H_{11}:\alpha_1,\alpha_2,\alpha_3,\alpha_4$不全为零;

- 总拉伸倍数(因子B)的显著性检验

  $H_{02}:\beta_1=\beta_2=\beta_3=\beta_4=0 \longleftrightarrow H_{12}:\beta_1,\beta_2,\beta_3,\beta_4$不全为零;

- 协变量回归系数的显著性检验

  $H_{03}:\gamma=0 \longleftrightarrow H_{13}:\gamma\neq 0$

案例分析

1. 协变量回归系数的显著性检验

分别计算原模型和假设$H_{03}:\gamma=0$下模型的残差平方和,

$$\mathrm{SS}_e^*=37.399,\quad \mathrm{SS}_e^*(A,B)=101.031,$$

其自由度分别为24, 25. 于是可得协变量的回归平方和

$$\mathrm{SS}_\gamma^*=\mathrm{SS}_e^*(A,B)-\mathrm{SS}_e^*=101.031-37.399=63.632,$$

自由度为$25-24=1$. 检验统计量

$$F_3=\frac{\mathrm{SS}_\gamma^*}{\mathrm{SS}_e^*/24}=\frac{63.632}{37.399/24}=40.83>F_{1,24}(0.05)=4.26,$$

因此, 在显著性水平$\alpha=0.05$下, 拒绝$H_{03}:\gamma=0$, 认为电流周波Z化纤弹性有显著影响.

2. 收缩率(因子A)的显著性检验

计算$H_{01}:\alpha_1=\alpha_2=\alpha_3=\alpha_4=0$ 下, 模型修正的残差平方和:

$$\mathrm{SS}^*_{eH_{01}}=50.434.$$

因子A的修正平方和为

$$\mathrm{SS}^*_{eH_{01}}-\mathrm{SS}^*_e=50.434-37.399=13.035.$$

于是H_{01}的检验统计量为

$$F_1=\frac{(\mathrm{SS}^*_{eH_{01}}-\mathrm{SS}^*_e)/3}{\mathrm{SS}^*_e/24}=2.788<F_{3,24}(0.05)=3.01,$$

由此初步判定收缩率(因子A)对纤维弹性Y的影响不显著.

3 总拉伸倍数(因子B)的显著性检验

在假设$H_{02}:\beta_1=\beta_2=\beta_3=\beta_4=0$ 得模型修正的残差平方和

$$\mathrm{SS}^*_{eH_{02}}=43.277,$$

故因子A的修正平方和为

$$\mathrm{SS}^*_{eH_{02}}-\mathrm{SS}^*_e=43.277-37.399=5.878.$$

由于H_{02}的检验统计量

$$F_2=\frac{(\mathrm{SS}^*_{eH_{02}}-\mathrm{SS}^*_e)/3}{\mathrm{SS}^*_e/24}=\frac{5.878/3}{37.399/24}=1.257<F_{3,24}(0.05)=3.01,$$

因此, 接受H_{02}, 认为总拉伸倍数B对纤维弹性Y影响不显著.

案例分析

因B的影响不显著, 故采用单因子协方差模型：

$$y_{ijk} = \mu + \alpha_i + \gamma z_{ijk} + e_{ijk}, \quad i = 1, \cdots, 4, \ j = 1, \cdots, 4, \ k = 1, 2,$$

```
yreg2<-lm(y A+Z, data=data)
Anova(yreg2,type=2)
Anova Table (Type II tests)
Response: y
```

	Sum Sq	Df	F value	Pr(>F)
A	15.315	3	3.185	0.03974 *
Z	66.348	1	41.394	6.802e-07 ***
Residuals	43.277	27		

同时置信区间

Table: 两两效应差的置信区间和同时置信区间

两两效应差	点估计	置信区间	Scheffé区间	Bonferroni区间
$\alpha_1 - \alpha_2$	2.360	[0.957, 3.763]	[0.322, 4.398]	[0.413, 4.307]
$\alpha_1 - \alpha_3$	2.424	[1.014, 3.834]	[0.377, 4.471]	[0.468, 4.380]
$\alpha_1 - \alpha_4$	4.001	[2.698, 5.304]	[2.109, 5.893]	[2.193, 5.809]
$\alpha_2 - \alpha_3$	0.064	[-1.235, 1.363]	[-1.476, 1.604]	[-1.738, 1.866]
$\alpha_2 - \alpha_4$	1.642	[0.198, 3.086]	[-1.822, 1.950]	[-0.362, 3.646]
$\alpha_3 - \alpha_4$	1.578	[0.123, 3.033]	[-0.535, 3.691]	[-0.440, 3.596]

- 只有$\alpha_2 - \alpha_3$的置信系数为95%的置信区间包含了零点, 因此, 可以认为因子A的第二水平α_2和第三水平α_3间的差异不显著.

- 基于以上置信区间和检验,可以给因子A的第四水平排序如下:

$$\alpha_1 > \alpha_2 = \alpha_3 > \alpha_4.$$

- 除了关于α_1和α_2的序外, 这个排序与图描述统计分析结果大致相同, 在该图表现出的是$\alpha_1 < \alpha_2$. 这点不同恰好说明了描述性统计分析的局限性, 协方差分析可以将数据潜在信息挖掘出来.
- 对于无交互效应的情形，协方差分析表的结果可调用R软件的程序包“car"中的函数Anova()中的type=2选项实现, 当模型存在交互效应时，程序类似于例7.3.2.