《生物试验设计》
第十章　协方差分析

王超

广东药科大学

Email: wangchao@gdpu.edu.cn

2023-10-04

第十章　协方差分析

方差与协方差分析

- 在方差分析中，一些分析变量本身是一个受到另一个或多个自变量影响的因变量
- 对这些自变量难以进行有效控制，但又要消除其对因变量的影响，以提高试验结果的可靠程度
- 例如：不同饲料对动物增重的影响
  - 试验动物的初始体重不同
  - 可能会增大处理间差异，也可能会增大处理内差异

协方差分析

- 协方差分析 ♣
  - 初始体重对增重的影响可以通过回归分析来度量
  - 矫正初始体重的影响后再进行方差分析
  - 在协方差分析中，y 为因变量，x 为协变量
- 把回归分析与方差分析结合起来的分析方法。将离均差的乘积和与平方和同时按照变异来源进行分解，利用协变量来降低试验误差，矫正处理平均数，实现统计控制，分析不同变异来源的相关关系，及对缺失数据进行估计
- 协方差分析也是用于分析多组均数之间的差异有无显著意义，只是多考虑了一个协变量因素

- 要提高试验的精确度和灵敏度，必须严格控制试验条件的均匀性
- 但是某些情况下，很难使试验控制达到预期要求
- 如果试验条件 x 与因变量 y 之间存在直线回归关系
  - 可以通过 x 来矫正 y
  - 使 y 的比较能够在相同试验条件下进行
  - 得到正确结论
- 用统计方法来矫正因自变量的不同而对因变量所产生的影响，使试验误差减小，对试验处理效应的估计更加准确

- 协方差

$$COV_{xy} = \frac{\sum (x - \mu_x)(y - \mu_y)}{N}$$

- 总体相关系数可以用总体方差和协方差表示

$$\rho = \frac{COV}{\sigma_x \sigma_y}$$

- 根据以上，可以得到不同来源协方差组分的估计值
- 并进一步估算两个变量之间各个变异来源的相关系数

- 方差分析中估计缺失数据是建立在误差平方和最小的基础上
  - 缺点：处理平方和发生偏倚
- 协方差分析中估计缺失数据
  - 保证误差平方和最小
  - 得到无偏的处理平方和

- 设有 k 组双变量资料，每组样本皆有 n 对 (x, y) 观测值，数据模式为

$$
\begin{array}{cccccccc}
1: & x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\
 & y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1n} \\
2: & x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\
 & y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2n} \\
\vdots & & & & & & \\
i: & x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} \\
 & y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{in} \\
\vdots & & & & & & \\
k: & x_{k1} & x_{k2} & \cdots & x_{kj} & \cdots & x_{kn} \\
 & y_{k1} & y_{k2} & \cdots & y_{kj} & \cdots & y_{kn}
\end{array}
$$

- x 和 y 的总乘积和（SP_T）可以分解为组间乘积和（SP_t）与组内乘积和（SP_e）两部分

$$SP_T = SP_t + SP_e$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x})(y_{ij}-\bar{y}) = n\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_{i:}-\bar{x})(\bar{y}_{i:}-\bar{y}) + \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i:})(y_{ij}-\bar{y}_{i:})$$

- 对应的自由度

$$df_T = df_t + df_e$$

$$nk - 1 = (k-1) + k(n-1)$$

- 协方差分析的数学模型为

$$y_{ij} = \mu_y + \alpha_i + \beta(x_{ij} - \mu_x) + \epsilon_{ij}$$

  - μ_x 和 μ_y 为平均值，α_i 为第 i 个处理效应，β 为总体回归系数
- 若令 $y'_{ij} = y_{ij} - \alpha_i$，此时协方差分析就是 y'_{ij} 与 x_{ij} 的线性回归分析

$$y'_{ij} = \mu_y + \beta(x_{ij} - \mu_x) + \epsilon_{ij}$$

- 若令 $y''_{ij} = y_{ij} - \beta(x_{ij} - \mu_x)$，此时协方差分析就是在消除了 x_{ij} 不一致对 y_{ij} 影响之后，对 y''_{ij} 的方差分析

$$y''_{ij} = \mu_y + \alpha_i + \epsilon_{ij}$$

- x 是固定的变量（各个处理的效应是固定的常量）
- ϵ_{ij} 是独立的，跟处理效应无关，服从正态分布
- 各个处理的 (x, y) 总体都是线性的，具有共同的回归系数 β
  - 各处理总体的回归是一条平行的直线
  - 对样本来说，各个处理之间回归系数的差异不显著（回归系数的齐性/同质性检验）

- 3 种饲料 (A_1, A_2, A_3) 饲喂试验中猪的始重 x 与增重 y 资料

A_1 :	x :	18	16	11	14	14	13	17	17
	y :	85	89	65	80	78	83	91	85
A_2 :	x :	17	18	18	19	21	21	16	22
	y :	95	100	94	98	104	97	90	106
A_3 :	x :	18	23	23	20	24	25	25	26
	y :	91	89	98	82	100	98	102	108

- 猪的始重在各组相差较大
- 始重大的猪增重快，始重小的猪增重慢
- 直接使用方差分析的方法忽略了始重不同的影响，不能反映 3 种饲料的真实效应
- 需要用协方差分析的方法，矫正始重对增重的影响

R DEMO

```
pig_weight <- data.frame(x = c(18, 16, 11, 14, 14, 13, 17, 17, 1
                         y = c(85, 89, 65, 80, 78, 83, 91, 85, 9
                         group = as.factor(rep(c("A1", "A2", "A3
                         )
summary(aov(y ~ group, data = pig_weight))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## group         2   1216   608.0   11.34 0.000458 ***
## Residuals    21   1126    53.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

R DEMO

```
model1 <- aov(y ~ x, data = pig_weight)
model2 <- aov(y ~ group + x, data = pig_weight)
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ group + x
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     22 865.68
## 2     20 395.35  2    470.33 11.897 0.0003946 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

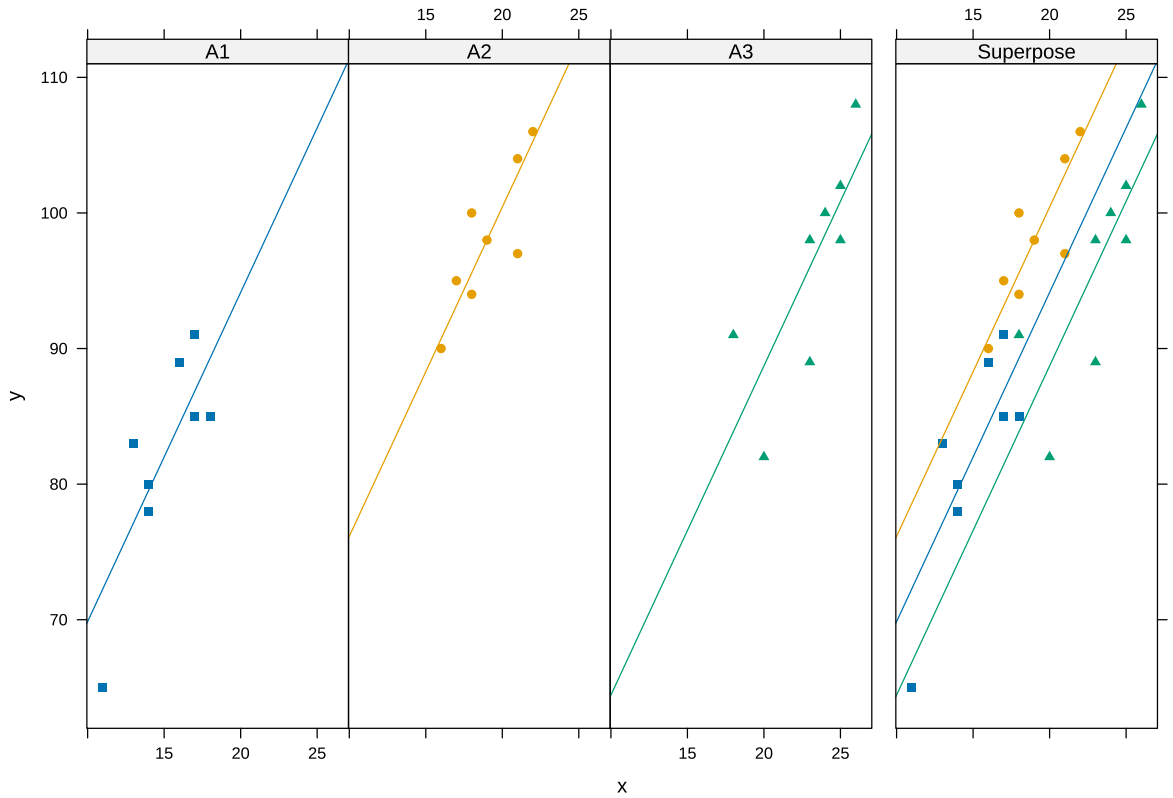
- 从组内项变异中找出 x(始重) 与 y(增重) 之间是否存在真实的线性回归关系
  - 计算处理内（组内）的回归系数，并对线性回归关系进行显著性检验
- 假设 $H_0 : \beta = 0, H_A : \beta \neq 0$
  - 如果接受 $H_0 : \beta = 0$，则二者之间回归关系不显著
  - 说明 y(增重) 不受 x(始重) 的影响
  - 不用考虑始重，直接对增重进行方差分析
- 计算误差项（处理内或组内）的回归系数、回归平方和与自由度

$$\begin{cases} b_e = \frac{SP_e}{SP_{e_x}} \\ U_e = \frac{SP_e^2}{SP_{e_x}} \\ df_e = 1 \end{cases}$$

- 组内回归关系显著，需要用组内回归系数对 y 进行矫正
- 矫正 y 计算出的各项平方和，实际上就是去除了 x 影响的部分
- 矫正 y 的各项平方和等于其相应变异项的离回归平方和及自由度

R DEMO

```
> library(HH)
> ancovaplot(y ~ group + x, data= pig_weight)
```



- y 对 x 有显著的直线回归关系，确实受到影响

R DEMO

```
> model <- ancova(y ~ x + group , data= pig_weight)
> summary(model)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## x             1 1476.3  1476.3   74.69 3.47e-08 ***
## group         2  470.3   235.2   11.90 0.000395 ***
## Residuals    20  395.3    19.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 通过矫正后消除始重的影响后，各饲料组间矫正增重达到显著水平，需要进行多重比较