2014.4.22

# 第十九讲. 因子变量



## 单因素方差分析(one-way anova)

假设第 $k$ 组 $y_{k1},\dots,y_{kn_k} \sim N(\mu_k,\sigma^2), k=1,\dots,\mathrm{K}$,各组独立,

$H0: \mu_1=\mu_2=\dots=\mu_{\mathrm{K}}$

线性模型表示:

$y_{ki}=\mu_k+\varepsilon_{ki}$, $\varepsilon_{ki}$ iid $\sim N(0,\sigma^2), i=1,2,\dots,n_k; k=1,\dots,\mathrm{K}$

$$\begin{array}{l} \text{第一组} \\ \text{第二组} \\ \text{第三组} \\ \\ \text{第}K\text{组} \end{array} \begin{pmatrix} \mathrm{Y}_1 \\ \mathrm{Y}_2 \\ \mathrm{Y}_3 \\ \vdots \\ \mathrm{Y}_K \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} & \dots & \mathbf{0}_{n_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_{n_K} & \mathbf{0}_{n_K} & \mathbf{0}_{n_K} & \dots & \mathbf{1}_{n_K} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_K \end{pmatrix} + \boldsymbol{\varepsilon}$$

模型: $\mathrm{Y}=\mathbf{x}_1\mu_1+\mathbf{x}_2\mu_2+\dots+\mathbf{x}_{\mathrm{K}}\mu_{\mathrm{K}}+\varepsilon$ (无截距项)



$\hat{Y}=(\bar{y}_{1\bullet},\dots\bar{y}_{1\bullet}, \dots, \bar{y}_{K\bullet},\dots\bar{y}_{K\bullet})$, 其中 $\bar{y}_{k\bullet}=(y_{k1}+\dots y_{kn_k})/n_k$ 组内平均

原假设下模型为: $Y=(\mathbf{x}_1+\mathbf{x}_2+\dots+\mathbf{x}_{\mathrm{K}})\mu_1+\varepsilon=\mathbf{1}\mu_1+\varepsilon$

$\hat{Y}_0=(\bar{y}_{\bullet\bullet},\dots\dots,\bar{y}_{\bullet\bullet})$, $\bar{y}_{\bullet\bullet}=\sum\sum y_{\mathrm{ij}}/n$, $n=n_1+..+n_K$ 为总平均

$$F=\frac{\|\hat{Y}-\hat{Y}_0\|^2/(K-1)}{\|Y-\hat{Y}\|^2/(n-K)}=\frac{\left(n_1(\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet})^2+\dots+n_K(\bar{y}_{K\bullet}-\bar{y}_{\bullet\bullet})^2\right)/(K-1)}{\sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{k\bullet})^2/(n-K)}$$

$$=\frac{SS_{\text{组间}}/(K-1)}{SS_{\text{组内}}/(n-K)}\boxed{\sim_{H_0} F_{K-1,n-K}}$$



事实上,通常的做法是直接进行方差(平方和)分解:

$$\mathrm{SS}_{\text{总}}=\sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{\bullet\bullet})^2=\sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{k\bullet}+\bar{y}_{k\bullet}-\bar{y}_{\bullet\bullet})^2$$

$$=\sum_{k=1}^{K}\sum_{j=1}^{n_k}(\bar{y}_{k\bullet}-\bar{y}_{\bullet\bullet})^2+\sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{k\bullet})^2 \triangleq SS_{\text{组间}}+SS_{\text{组内}}$$

$$F=\frac{SS_{\text{组间}}/(K-1)}{SS_{\text{组内}}/(n-K)}\sim F_{K-1,n-K} \quad \text{(原假设下)}$$

称为单因素方差分析(one-way anova)

方差分析表:

| 来源 | 因子/分组 | 残差 |
|---|---|---|
| 平方和 | $\mathrm{SS}_{\text{组间}}$ | $\mathrm{SS}_{\text{组内}}$ |
| 自由度 | $\mathrm{K}-1$ | $\mathrm{n}-k$ |

**AOV函数(Analysis of Variance)**

- **aov( y ~ group) # group** 代表了每个观察所属组别，因子变量。
- **lm (y ~ group) #**

例如，**Sleep1**数据

**TS**：哺乳动物睡眠时间

**D:** 动物所处环境危险等级，**5**个等级（水平）。检验睡眠时间是否与**D**有关，即**5**组数据是否有显著性差异：

```
> aov(TS~D, data=sleep1)
```

Terms:

| | D | Residuals |
|---|---|---|
| Sum of Squares | 457.2556 | 752.4122 |
| Deg. of Freedom | 4 | 53 |

← 方差分析表

# F= (457.2556/4)/(752.4122/ 53)=8.052



## 因子变量及其哑变量表示(dummy coding)

- 因子变量(属性变量、分类变量)代表的是分类而不是数值，为了数学/计算机上进行处理，需要将其编码(数值化)。

  比如因子变量 **color** 取值 **red, yellow, blue.** 你可以将**red, yellow, blue** 分别定义为**1,2,3,**但它们只是代表了类别而不具有数值含义，数学/计算机上仍不能处理。

- 上节课，我们以**K**个示性函数/**dummy variable**表示**K-**水平因子，其和为**1**，因此模型中没有截距项。这种表示不方便用于多个自变量的情形。

- 实际上为了方便, 如果有一个基准组(对照组)，我们可以用其余组的**K-1**个示性函数表示**K**个类(如果某个体的这**K-1**个示性变量都为**0**，那么它属于对照组). 比如性别因子变量取值为男、女，你可以将男性定义为**1**，女性定义为**0**，即使用男性的示性函数表示性别。



单因素方差分析模型：$y_{k1,\dots,}\ y_{kn_k}\ \text{iid} \sim N(\mu_k,\sigma^2),\ k=1,\dots,K$

$\Leftrightarrow y_{ki}=\mu_k+\varepsilon_{ki},\ \varepsilon_{ki}\sim N(0,\sigma^2),\ \ k=1,\dots,K;\ i=1,\dots,n_k$

重新参数化，令$\beta_2=\mu_2-\mu_1,\ \dots,\beta_K=\mu_K-\mu_1$，称为效应，则
$H_0:\mu_1=\mu_2=\dots=\mu_K \Leftrightarrow \beta_2=\dots=\beta_K=0$

$\Leftrightarrow y_{ki}=\mu_1+\beta_k+\varepsilon_{ki},\varepsilon_{ki}\sim N(0,\sigma^2),\ \beta_1=0,\ \ k=1,\dots,K;\ i=1,\dots,n_k$

$\Leftrightarrow$ 任何一个$y_i=\mu_1+x_{i2}\beta_2\dots+x_{iK}\beta_K+\varepsilon_i, i=1,\dots,n,$ 其中$x_{ik}=1_{(i属于第k组)}, k\geq 2$

$$\Leftrightarrow \begin{array}{l} 第一组 \\ 第二组 \\ 第三组 \\ \\ 第K组 \end{array} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_K \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} & \cdots & \mathbf{0}_{n_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_K} & \mathbf{0}_{n_K} & \mathbf{0}_{n_K} & \cdots & \mathbf{1}_{n_K} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{pmatrix} + \boldsymbol{\varepsilon}$$

$$\mathrm{Y} \ = \ \mathbf{1}\mu_1+\mathbf{x}_2\beta_2+\dots+\ \mathbf{x}_K\,\beta_K \ + \ \varepsilon$$



## 两因素方差分析(two-way anova)

两个因子变量,比如一个是处理，一个是区组/block.

取值分别为K，J类，在每个水平组合$(k,j)$下

的数据为:$y_{kj1},\dots,y_{kj\,n_{kj}}\ iid\sim N(\mu_{kj},\sigma^2).$

| | 1 | … | $J$ |
|---|---|---|---|
| 1 | $y_{11i}, i=1,..n_{11}$ | … | $y_{1Ji}, i=1,..n_{1J}$ |
| … | … | … | … |
| $K$ | $y_{K1i}, i=1,..n_{K1}$ | … | $y_{KJi}, i=1,..n_{KJ}$ |

假设可加模型：

$y_{kji}=\mu_{11}+\alpha_k+\beta_j+\varepsilon_{kji},\ \ k=1,\dots,K;\ j=1,\dots,J;\ i=1,2\dots,n_{jk}$，其中$\alpha_1=0,\beta_1=0$。

这种模型表示含义明显，易于解释，$\alpha_k$，$\beta_j$称为水平$k,j$的效应(effect).

但数学上不能处理

将所有$y$'s编号为1,2,....,$n$, 第$i$个响应$y_i$满足一般线性模型：

$$y_i=\mu_{11}+\sum_{k=2}^{K}\alpha_k x_{ik}+\sum_{j=2}^{J}\beta_j z_{ij}+\varepsilon_i$$

其中$x_{ik}=1_{(i属于第k个处理组)}$，$k\ge 2$; $z_{ij}=1_{(i属于第j个区组)}, j\ge 2$

$H0:\alpha_2=\dots=\alpha_K=0$

的F-检验可由一般F公式得到,或等价地由下页平方和分解得到。



平方和分解：

$$SS_{总}=\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{j=1}^{n_{kl}}(y_{klj}-\bar{y}_{\bullet\bullet})^2$$

$$=\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{l=1}^{n_{kj}}(\bar{y}_{k\bullet\bullet}-\bar{y}_{\bullet\bullet})^2+\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{l=1}^{n_{kj}}(\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet})^2+\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{l=1}^{n_{kj}}(y_{kjl}-\bar{y}_{k\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet})^2$$

$$=SS_{组间1}+SS_{组间2}+SS_{组内}$$

H0：$\alpha_2=\dots=\alpha_K=0$

$$F=\frac{SS_{组间1}/(K-1)}{SS_{组内}/(n-K-J+1)}\overset{H_0}{\sim}F_{K-1,n-K-J+1}$$

特别当所有$n_{kl}=1$时，$n=KJ, F=\dfrac{SS_{组间1}/(K-1)}{SS_{组内}/(K-1)(J-1)}\overset{H_0}{\sim}F_{K-1,(K-1)(J-1)}$



注1:

将K-水平因子表示为K-1个示性变量(dummy variable)可方便地处理

(1) 多个因子自变量情形：多因素方差分析模型

(2) 既有因子也有连续变量的情形：协方差分析 模型

注2：如果没有一个特定的对照组，即各组/水平地位平等，那么为了检验各组/水平均值相同，可考察各个均值与总平均的差：

$H_0:\mu_1=\mu_2=\dots=\mu_K$

令$\mu=(\mu_1+\dots+\mu_K)/K,\beta_k=\mu_k-\mu,k=1,2,\dots,K$,有约束：$\beta_1+\dots+\beta_K=0$

$H_0:\mu_1=\mu_2=\dots=\mu_K\Leftrightarrow\beta_1=\beta_2=\dots=\beta_K=0$

注3：我们假设了两因素方 差分析模型是可加的：

$$y_{kji}=\mu_{11}+\alpha_k+\beta_j+\varepsilon_{kji},\ \ k=1,\dots,K;\ j=1,\dots,J;\ i=1,2\dots,n_{jk},$$

即两个因素的效应是线 性可加的，否则可考虑 交互作用.



## 带交互作用的两因素方差分析

两个因子变量$x,z$，都是2个水平. $y_{ij1},\dots,y_{ijn_{ij}}\ iid\sim N(\mu_{ij},\sigma^2),\ \ i,j=0,1$

| | $z=0$ | $z=1$ |
|---|---|---|
| $x=0$ | $y_{00i},i=1,..n_{00}$ | $y_{01i},i=1,..n_{01}$ |
| $x=1$ | $y_{10i},i=1,..n_{10}$ | $y_{11i},i=1,..n_{11}$ |

| | | z: 0 | 1 |
|---|---|---|---|
| x | 0 | $\mu_{00}$ | $\mu_{01}$ |
| | 1 | $\mu_{10}$ | $\mu_{11}$ |

可加模型：

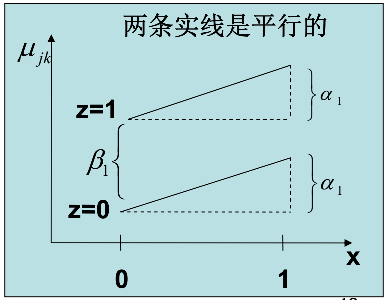
$\mu_{jk}=\mu+\alpha_j+\beta_k$，其中 $\alpha_0=0,\beta_0=0$

$\mu_{00}=\mu$

$\mu_{01}=\mu+\beta_1$

$\mu_{10}=\mu+\alpha_1$

$\mu_{11}=\mu+\alpha_1+\beta_1$

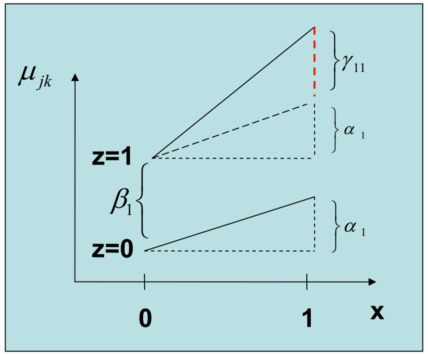


| | 0 | 1 |
|---|---|---|
| 0 | $\mu_{00}$ | $\mu_{01}$ |
| 1 | $\mu_{10}$ | $\mu_{11}$ |

交互作用模型：

$$\mu_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}$$

其中 $\alpha_0 = 0, \beta_0 = 0, \gamma_{0j} = \gamma_{i0} = 0$




$$\begin{aligned} \mu_{00} &= \mu \\ \mu_{01} &= \mu + \beta_1 \\ \mu_{10} &= \mu + \alpha_1 \\ \mu_{11} &= \mu + \alpha_1 + \beta_1 + \gamma_{11} \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} \mu &= \mu_{00}, \\ \beta_1 &= \mu_{01} - \mu_{00}, \\ \alpha_1 &= \mu_{10} - \mu_{00}, \\ \gamma_{11} &= \mu_{11} - \mu_{01} - \mu_{10} + \mu_{00} \end{aligned}$$



## 例： salary 数据

| 变量 | 描述 |
|---|---|
| *Sex* | 1 : 女， 0：男 |
| *Rank* | 职称 1: Assistant Prof, 2: Associate Prof, 3:Full Prof |
| *Year* | 拥有当前职称（rank）的时间（单位：年） |
| *Degree* | 最高学位。 1：博士, 0：硕士 |
| *YSdeg* | 工龄: 获得最高学位至今的时间（单位：年） |
| *Salary* | 年薪 ($) |

```
> Rank
 [1] 3 3 3 3 3 3 3 3 3 3 3 2 3 2 3 3 3 2 2 3 1 2 3 3 2 3 2 3 2 2 1 2 1 2
> is.factor(Rank)
[1] FALSE
```

因为3，2，1可能并不能代表三个职称上的差异（对工资来说），下面我们把它当作因子变量，分析工资与职称的关系：



### 单因素分析

将Rank当作因子时，

$$\text{Salary} = a + b \times \text{Rank2} + c \times \text{Rank3} + \varepsilon \qquad (*)$$

$b$表示Rank2与Rank1的Salary的平均差别，即副教授的效应

$c$表示Rank3与Rank1的Salary的平均差别，即正教授的效应

```
> lm(Salary~factor(Rank))
Coefficients:
 (Intercept)   factor(Rank)2    factor(Rank)3
      17769            5407            11890
```

故Rank的第二、三水平的效应 估计分别为 5407、11890

即副教授比助理教授多 5407$，正教授比助理教授多 11890$.

截距项 17769 为助理教授平均工资（ 副教授平均工资 ＝17769＋5407）



### 将Rank当作数值变量(取实数值1,2,3, 就像数据本身给出的那样)？

```
> lm(Salary~ Rank)
Coefficients:
(Intercept)       Rank
   11663          5953
```

$$\text{Salary} = a + b \times \text{Rank} + \varepsilon \qquad (**)$$

$\hat{b} = 5953$表示Rank增加一个单位Salary的平均增量

即副教授比助理教授，正教授比副教授都多5953.

模型 (**) 假设了Rank作为实数变化时Salary的变化率是常数，

换言之，假设了Rank1，Rank2，Rank3的工资是等间距的。

从模型 (*)的分析结果来看，模型(**) 是否合理呢？即3个效应0、5407、11890

是否可以认为是等间距的？基本上是！所以有理由采用有更高效率的模型(**)，

这也某种程度上说明了原数据为什么把三个职称分别赋值1,2,3而不是1,2,5

## 两因素分析

```
> lm(Salary~factor(Rank) +Sex)
Coefficients:
 (Intercept)    factor(Rank)2    factor(Rank)3        Sex
   18155            5145            11678            -870
```

截距项18155为男性助理教授平均工资;Rank的第二、三水平的效应估计分别为5145、11678; Sex 的第二水平（女性）效应为-870

即对于特定的性别，副教授比助理教授多5145$，正教授比助理教授多11678$.

对于给定的Rank，女性比男性平均少870$

女性助理教授平均工资 : $18155-870$

男性副教授平均工资 : $18155+5145$

女性副教授平均工资 : $18155+5145-870$

**该模型Sex与Rank2，Rank3是可加的/线性的，Rank的效应在男、女性别中是否可能不同？**



## 因子变量之间的交互作用

```
> lm(Salary~factor(Rank) * Sex)  # 考虑交互作用模型，即非可加模型

Coefficients:
 (Intercept)   factor(Rank)2   factor(Rank)3    Sex    factor(Rank)2:Sex  factor(Rank)3:Sex
   17920          5524           11953          -340        -1534              -728
```

男性副教授、正教授的效应分别是 **5524**、**11953**，

女性副教授、正教授的效应分别是 **5524-1534**、**11953-728**

比如女性正教授平均工资为**17920 -340 +** （**11953-728**）

交互作用效应是否显著（或可加模型是否合理）？

```
> anova( lm(Salary~factor(Rank) + Sex) , lm(Salary~factor(Rank) * Sex) )
Analysis of Variance Table

Model 1: Salary ~ factor(Rank) + Sex
Model 2: Salary ~ factor(Rank) * Sex
  Res.Df     RSS    Df   Sum of Sq       F       Pr(>F)
1    48 431871315
2    46 428769653  2    3101661      0.1664    0.8472
```



## 因子变量与连续变量的交互作用

```
> lm(Salary~factor(Rank) +Year)  # Year连续变量

Coefficients:
 (Intercept)     Rank2       Rank3       Year
   16203.3       4262.3      9454.5      375.7
```

```
交互作用：
> lm(Salary~factor(Rank) * Year,data=salary)
Coefficients:
    (Intercept)   Rank2    Rank3   Year  Rank2:Year   Rank3:Year
      16417       5354     8176    325     -130          151
```

$$\text{Salary} = a + b_2 \times \text{Rank2} + b_3 \times \text{Rank3} + c \times \text{Year}$$
$$+ d_2 \times \text{Rank2} \times \text{Year} + d_3 \times \text{Rank3} \times \text{Year} + \varepsilon$$

助理教授(Rank2、3 = 0) : $\text{Salary} = a + c \times \text{Year} + \varepsilon$

副教授(Rank2 = 1) : $\text{Salary} = a + b_2 + (c + d_2) \times \text{Year} + \varepsilon$

正教授(Rank3 = 1) : $\text{Salary} = a + b_3 + (c + d_3) \times \text{Year} + \varepsilon$