第八章　方差分析与回归分析

§8.1　方差分析

1. 单因子方差分析

(1) 问题与数据

设某因子有 r 个水平,记为 $A_1,A_2,\cdots,A_r$,在每一水平下各做 m 次独立重复试验,若记第 i 个水平下第 j 次重复的试验结果为 y_{ij},所有试验的结果可列表如下:

因子水平	试验数据				和	平　均
A_1	y_{11}	y_{12}	$\cdots$	y_{1m}	T_1	$\bar{y}_1$
A_2	y_{21}	y_{22}	$\cdots$	y_{2m}	T_2	$\bar{y}_2$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
A_r	y_{r1}	y_{r2}	$\cdots$	y_{rm}	T_r	$\bar{y}_r$
					T	$\bar{y}$

对这个试验要研究的问题是:r 个水平 $A_1,A_2,\cdots,A_r$ 间有无显著差异.

(2) 基本假定

A1:第 i 个水平下的数据 $y_{i1},y_{i2},\cdots,y_{im}$ 是来自正态总体 $N(\mu_i,\sigma_i^2)$ 的一个样本,$i=1,2,\cdots,r$;

A2:r 个方差相同,即 $\sigma_1^2=\sigma_2^2=\cdots=\sigma_r^2=\sigma^2$;

A3:诸数据 y_{ij} 都相互独立.

在这三个基本假定下,要检验的假设是

$$H_0:\mu_1=\mu_2=\cdots=\mu_r \qquad \text{vs} \qquad H_1:\mu_1,\mu_2,\cdots,\mu_r \text{ 不全相等}.$$

方差分析就是在方差相等的条件下,对若干个正态均值是否相等的假设检验.

(3) 平方和分解式

$$S_T=S_A+S_e,\quad f_T=f_A+f_e,$$

若记 $\bar{y}_{i\cdot}=\frac{1}{m}\sum_{j=1}^{m}y_{ij},\bar{y}=\frac{1}{rm}\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}=\frac{1}{r}\sum_{i=1}^{r}\bar{y}_{i\cdot}$.上述诸平方和分别为

- $S_T=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2$ 称为总平方和,其自由度 $f_T=n-1$.

- $S_A=m\sum_{i=1}^{r}(\bar{y}_{i\cdot}-\bar{y})^2$ 称为组间平方和或因子 A 的平方和,其自由度 $f_A=r-1$.

- $S_e=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2$ 称为组内平方和或误差平方和,其自由度 $f_e=n-r$.

注:数据 y_{ij} 的平移 $y'=y_{ij}-a$ 不会改变其平方和的值.用此性质可简化计算.

(4) 方差分析表

来源	平方和	自由度	均方	F 比
因子	$S_A=\frac{1}{m}\sum_{i=1}^{r}T_i^2-\frac{T^2}{rm}$	$f_A=r-1$	$MS_A=S_A/f_A$	$F=MS_A/MS_e$
误差	$S_e=S_T-S_A$	$f_e=r(m-1)$	$MS_e=S_e/f_e$	
总和	$S_T=\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}^2-\frac{T^2}{rm}$	$f_T=rm-1$		

(5) 判断

在 H_0 成立下,$F=MS_A/MS_e\sim F(f_A,f_e)$,对给定的显著性水平 $\alpha(0<\alpha<1)$,其拒绝域为 $W=\{F\geqslant F_{1-\alpha}(f_A,f_e)\}$,其中 $F_{1-\alpha}(f_A,f_e)$ 可从附表 5 中查得.

- 若 $F\geqslant F_{1-\alpha}(f_A,f_e)$,则认为因子 A 显著,即诸正态均值间有显著差异.

- 若 $F<F_{1-\alpha}(f_A,f_e)$,则说明因子 A 不显著,即接受原假设 H_0.

给出检验的 p 值是更常用的,若以 X 记服从 $F(f_A,f_E)$ 的随机变量,F 为统计量,$F=MS_A/MS_e$ 的观测值为 F_0,则 $p=P(X\geqslant F_0)$,可用软件计算,如在 Matlab 中使用如下命令:$p=1-\text{fcdf}(F_0,f_A,f_E)$.

2. 数据结构式及其参数估计

(1) 数据结构式

$$y_{ij}=\mu+a_i+\varepsilon_{ij},\qquad i=1,2,\cdots,r,j=1,2,\cdots,m,$$

其中 μ 为总均值，a_i 为第 i 个水平的效应，且 $\sum_{i=1}^{r} a_i = 0$，ε_{ij} 为试验误差，所有 ε_{ij} 可作为来自 $N(0,\sigma^2)$ 的一个样本，在上述数据结构式下，$y_{ij} \sim N(\mu + a_i, \sigma^2)$.要检验的假设可改写为

$$H_0: a_1 = a_2 = \cdots = a_r = 0 \quad \text{vs} \quad H_1: a_1, a_2, \cdots, a_r \text{ 不全为 } 0.$$

（2）点估计

- 总均值 μ 的估计 $\hat{\mu} = \bar{y}$.
- 水平均值 μ_i 的估计 $\hat{\mu}_i = \bar{y}_{i\cdot}$，$i = 1, 2, \cdots, r$.
- 主效应 a_i 的估计 $\hat{a}_i = \bar{y}_{i\cdot} - \bar{y}$，$i = 1, 2, \cdots, r$.
- 误差方差 σ^2 的估计 $\hat{\sigma}^2 = MS_e = S_e / f_e$.

（3）$1-\alpha$ 置信区间（$0<\alpha<1$）

- μ_i 的 $1-\alpha$ 置信区间为 $[\bar{y}_{i\cdot} \pm \hat{\sigma} t_{1-\alpha/2}(f_e) / \sqrt{m}]$.

3. 单因子试验的统计分析可得如下三个结果

- 因子 A 是否显著.
- 试验误差方差 σ^2 的估计.
- 诸水平均值 μ_i 的点估计与区间估计（此项在因子 A 不显著时无需进行）.

4. 重复数不等情形下的方差分析

（1）数据略有不同

设因子 A 有 r 个水平 $A_1, A_2, \cdots, A_r$，并且第 i 个水平 A_i 下重复进行 m_i 次试验，获得如下数据：

水平	重复数	数　据	和	平　均
A_1	m_1	$y_{11}, y_{12}, \cdots, y_{1m_1}$	T_1	$\bar{y}_{1\cdot}$
A_2	m_2	$y_{21}, y_{22}, \cdots, y_{2m_2}$	T_2	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	m_r	$y_{r1}, y_{r2}, \cdots, y_{rm_r}$	T_r	$\bar{y}_{r\cdot}$
合计	n	T	T	$\bar{y}$

（2）基本假定、平方和分解、方差分析及判断准则都和前面一样，只是因子 A 的平方和 S_A 的计算公式略有不同：记 $n = \sum_{i=1}^{r} m_i$，则

$$S_A=\frac{T_1^2}{m_1}+\frac{T_2^2}{m_2}+\cdots+\frac{T_r^2}{m_r}-\frac{T^2}{n}.$$

（3）数据结构式及其参数估计基本同前，但要注意以下两点：

- 总均值 $\mu=\frac{1}{n}\sum_{i=1}^{r}m_i\mu_i$.
- 主效应的约束条件为 $\sum_{i=1}^{r}m_ia_i=0$.

§8.2 多 重 比 较

1. 问题 在单因子方差分析中，当因子 A 显著时，就要继续研究如下问题：在多个水平均值中同时比较任意两个水平间有无明显差异的问题，这个问题的检验法则称**多重比较**.

若因子 A 有 r 个水平，则同时检验 $r(r-1)/2$ 个假设

$$H_0^{ij}:\mu_i=\mu_j,\quad 1\leqslant i<j\leqslant r,$$

其拒绝域 $W=\bigcup_{1\leqslant i<j\leqslant r}\{|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|\geqslant c_{ij}\}$. 对给定的显著性水平 $\alpha(0<\alpha<1)$，诸临界值 c_{ij} 由 $P(W)=\alpha$ 决定.

2. 重复数相等场合的 T 法 在各水平试验次数相同时，其诸临界值 c_{ij} 也相同，具体为

$$c=q_{1-\alpha}(r,f_e)\hat{\sigma}/\sqrt{m},$$

其中 $\hat{\sigma}=\sqrt{MS_e}$，$q_{1-\alpha}(r,f_e)$ 是分布 $q(r,f_e)$ 的 $1-\alpha$ 分位数，可从附表 8 中查得.

3. 重复数不等场合的 S 法 在各水平试验次数不同时，其诸临界值 c_{ij} 也不同，具体为

$$c_{ij}=\sqrt{(r-1)F_{1-\alpha}(r-1,f_e)\left(\frac{1}{m_i}+\frac{1}{m_j}\right)\hat{\sigma}^2},$$

其中 $F_{1-\alpha}(r-1,f_e)$ 是相应的 F 分布的 $1-\alpha$ 分位数.

§ 8.3　方差齐性检验

1. 问题　方差齐性即诸方差相等,是方差分析的基本假定之一,方差齐性检验就是检验这个假定是否成立.该检验问题的一对假设为

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 \quad \text{vs} \quad H_1: \text{诸 } \sigma_i^2 \text{ 不全相等}.$$

2. 哈特利检验(在重复数相等场合使用)　在重复次数均为 m 时,采用哈特利检验,检验统计量是

$$H = \frac{\max\{s_1^2, s_2^2, \cdots, s_r^2\}}{\min\{s_1^2, s_2^2, \cdots, s_r^2\}},$$

其中 s_i^2 是第 i 个水平 A_i 下重复试验数据的样本方差.拒绝域为

$$W = \{H > H_{1-\alpha}(r, f)\},$$

其中 α 为显著性水平,$f = m-1$,$H_{1-\alpha}(r, f)$ 是统计量 H 的分布的 $1-\alpha$ 分位数,它的值可查附表 10.

3. 巴特利特检验(可在重复数不等场合使用,但样本量不得低于 5)　在重复次数不等且每个水平下试验次数均不低于 5 时,可采用巴特利特检验,检验统计量为

$$B = \frac{1}{C}\left[f_e \ln MS_e - \sum_{i=1}^{r} f_i \ln s_i^2\right],$$

其中诸 s_i^2 同上,$f_i = m_i - 1$ 为 s_i^2 的自由度,$MS_e = \frac{1}{f_e}\sum_{i=1}^{r} f_i s_i^2$,$f_e = \sum_{i=1}^{r} f_i$ 为误差方差的自由度.$C = 1 + \frac{1}{3(r-1)}\left(\sum_{i=1}^{r} \frac{1}{f_i} - \frac{1}{f_e}\right)$,拒绝域为

$$W = \{B > \chi_{1-\alpha}^2(r-1)\},$$

其中 $\chi_{1-\alpha}^2(r-1)$ 是自由度为 $r-1$ 的 χ^2 分布的 $1-\alpha$ 分位数.

4. 修正的巴特利特检验(在样本量相等或不等,样本量较小或较大场合均可使用)　一般场合,可采用修正的巴特利特检验,检验统计量是

$$B' = \frac{f_2 BC}{f_1(A - BC)},$$

其中 B 与 C 同前，$f_1=r-1$，$f_2=\dfrac{r+1}{(C-1)^2}$，$A=\dfrac{f_2}{2-C+2/f_2}$，拒绝域为

$$W=\{B'>F_{1-\alpha}(f_1,f_2)\},$$

其中 $F_{1-\alpha}(f_1,f_2)$ 是相应 F 分布的 $1-\alpha$ 分位数，f_2 的值可能不是整数，这时可通过对 F 分布的分位数表施行线性内插法获得.

§8.4　一元线性回归

1. 问题　考察两个变量 x 与 y 之间是否存在线性相关关系，其中 x 是一般（可控）变量，y 是随机变量，其线性相关关系可表示如下（可用散点图显示）：

$$y=\beta_0+\beta_1 x+\varepsilon,$$

其中 β_0 为截距，β_1 为斜率，ε 为随机误差，常假设 $\varepsilon\sim N(0,\sigma^2)$. 这里 β_0，β_1，σ^2 是三个待估参数. 上式表明，y 与 x 之间有线性关系，但受到随机误差的干扰.

2. 数据　对 x 与 y 通过试验或观察可得 n 对数据（注：数据是成对的，不允许错位）. 在 y 与 x 之间存在线性关系的假设下，有如下统计模型：

$$\begin{cases} y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\ i=1,2,\cdots,n, \\ \text{各 } \varepsilon_i \text{ 独立同分布，其分布为 } N(0,\sigma^2). \end{cases}$$

利用成对数据可获得 β_0 与 β_1 的估计，设估计分别为 $\hat\beta_0$ 与 $\hat\beta_1$，则称 $\hat y=\hat\beta_0+\hat\beta_1 x$ 为回归方程，其图形称为回归直线.

3. 参数估计　用最小二乘法可得 β_0 与 β_1 的无偏估计

$$\begin{cases} \hat\beta_1=l_{xy}/l_{xx}, \\ \hat\beta_0=\bar y-\hat\beta_1\bar x, \end{cases}$$

其中 $\bar x=\dfrac{1}{n}\sum x_i$，$\bar y=\dfrac{1}{n}\sum y_i$（此处 $\sum$ 表示 $\sum\limits_{i=1}^{n}$，下同），

$$l_{xy}=\sum(x_i-\bar x)(y_i-\bar y)=\sum x_i y_i-n\,\bar x\cdot\bar y=\sum x_i y_i-\frac{1}{n}\sum x_i\sum y_i,$$

$$l_{xx}=\sum(x_i-\bar{x})^2=\sum x_i^2-n\bar{x}^2=\sum x_i^2-\frac{1}{n}\left(\sum x_i\right)^2,$$

$$l_{yy}=\sum(y_i-\bar{y})^2=\sum y_i^2-n\bar{y}^2=\sum y_i^2-\frac{1}{n}\left(\sum y_i\right)^2.$$

4. 回归方程的显著性检验　回归方程的显著性检验就是要对如下一对假设作出判断：

$$H_0: \beta_1=0 \quad \text{vs} \quad H_1: \beta_1\neq 0.$$

对此可采用如下两种等价的检验方法.

（1）F **检验**

如下的平方和分解式是非常重要的，它在许多统计领域得到应用：

$$S_T=S_R+S_e,\ f_T=f_R+f_e,$$

其中

$S_T=\sum(y_i-\bar{y})^2=\sum y_i^2-\frac{1}{n}(\sum y_i)^2=l_{yy}$是总平方和，其自由度$f_T=n-1$，

$S_R=\sum(\hat{y}_i-\bar{y})^2=\sum(\hat{\beta}_0+\hat{\beta}_1x_i-\bar{y})^2=\hat{\beta}_1l_{xy}=\hat{\beta}_1^2l_{xx}$是回归平方和，其自由度$f_R=1$，

$S_e=\sum(y_i-\hat{y}_i)^2=\sum(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$是残差平方和，其自由度$f_e=n-2$.

而$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$是在$x=x_i$的回归值（拟合值），它与实测值$y_i$通常是不相等的.

在原假设H_0成立的条件下，检验统计量$F=\frac{S_R}{S_e/(n-2)}\sim F(1,n-2)$，拒绝域为

$$W=\{F\geqslant F_{1-\alpha}(1,n-2)\}.$$

上述检验过程一般用如下方差分析表列出：

来源	平方和	自由度	均方	F比
回归	S_R	$f_R=1$	$MS_R=S_R$	$F=\frac{MS_R}{MS_e}$
残差	S_e	$f_e=n-2$	$MS_e=\frac{S_e}{n-2}$	
总计	S_T	$f_T=n-1$		

（2）t 检验

检验统计量为 $t=\dfrac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{l_{xx}}}$，其中 $\hat{\sigma}=\sqrt{S_e/(n-2)}$，在原假设成立下，$t\sim t(n-2)$，因此拒绝域为

$$W=\{|t|>t_{1-\alpha/2}(n-2)\}.$$

注意到 $t^2=F$，因此 t 检验与 F 检验是等价的，选其中之一使用即可.

5. 相关系数及其检验

（1）相关系数

对容量为 n 的二维样本 $\{(x_i,y_i),i=1,2,\cdots,n\}$ 的线性相关程度可用如下(样本)相关系数度量

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}.$$

- $r=\pm1$，n 个点完全在一条直线上，此时两者之间可能是确定性关系.
- $r>0$，当 x 增加时，y 有线性增加趋势，此时称正相关.
- $r<0$，当 x 增加时，y 反而有线性减少趋势，此时称负相关.
- $r=0$，n 个点可能毫无规律，也可能呈某种曲线趋势，此时称不(线性)相关.

（2）相关系数的检验

记 ρ 为二维总体的相关系数，于是可建立如下假设：

$$H_0:\rho=0 \quad \text{vs} \quad H_1:\rho\neq0.$$

对此，采用检验统计量 $r=\dfrac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$，拒绝域为

$$W=\{|r|>r_{1-\alpha}(n-2)\},$$

其中 $r_{1-\alpha}(n-1)$ 是 $|r|$ 分布的 $1-\alpha$ 分位数，可查附表 9.

（3）检验统计量 r 与 F 统计量之间关系

$$r^2=\frac{F}{F+(n-2)},$$

这表明 $|r|$ 是 F 的严格增函数，所以相关系数检验与前面的 F 检验也是等价的.

6. 估计与预测——回归方程的应用

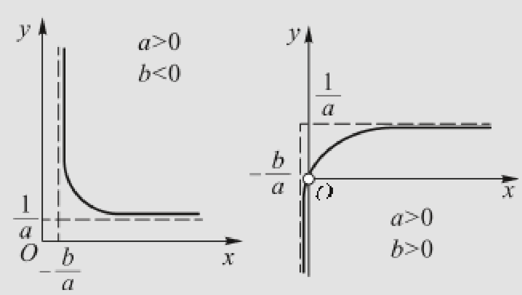
- 当 $x=x_0$ 时,$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_0$ 是 $E(y_0)=\beta_0+\beta_1x_0$ 的点估计.
- 当 $x=x_0$ 时,$E(y_0)=\beta_0+\beta_1x_0$ 的置信水平为 $1-\alpha$ 的置信区间是 $[\hat{y}_0-\delta_0,\hat{y}_0+\delta_0]$,其中 $\delta_0=t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{l_{xx}}}$, $\hat{\sigma}=\sqrt{MS_e}$.
- 当 $x=x_0$ 时,$y_0=\beta_0+\beta_1x_0+\varepsilon_0$ 的 $1-\alpha$ 预测区间是 $[\hat{y}_0-\delta,\hat{y}_0+\delta]$,其中 $\delta=\delta(x_0)=t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{l_{xx}}}$.

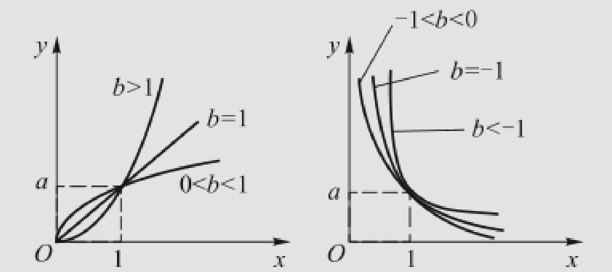
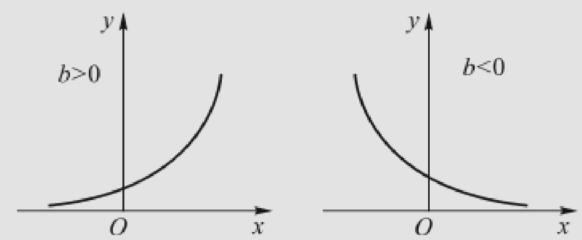
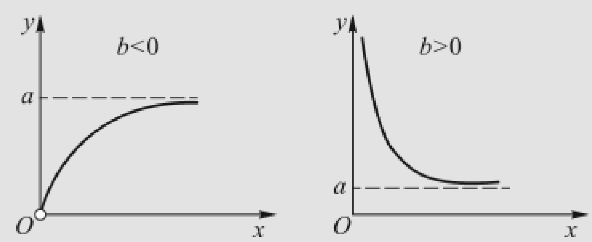
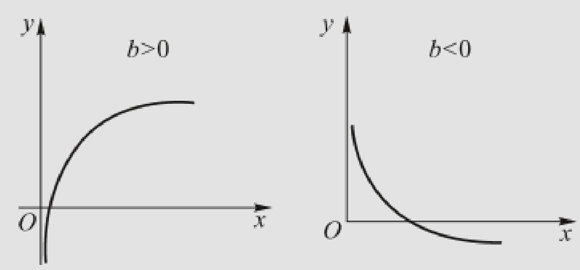
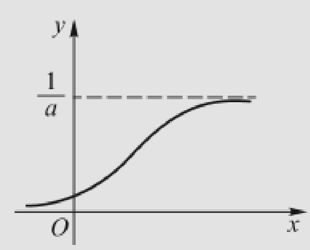
注:$E(y_0)$是未知参数,而 y_0 是随机变量.对 $E(y_0)$谈论的是置信区间,对 y_0 谈论的是预测区间,两者是不同的,显然,预测区间要比置信区间宽很多.

要提高预测区间(置信区间也一样)的精度,即要使 δ(或 δ_0)较小,这要求:(1) 增大样本量 n;(2) 增大 l_{xx},即要求 $x_1,x_2,\cdots,x_n$ 较为分散;(3) 使 x_0 靠近 $\bar{x}$.

§8.5 一元非线性回归

1. 非线性函数形式 根据二维样本的散点图确定可能的非线性函数形式,部分常见的非线性函数及其图形如下表.注意:若有两个或两个以上非线性函数可用,可对它们分别拟合非线性回归并根据后面提及的一些准则进行选择.

函数名称	函数表达式	图 像	线性化变换
双曲线函数	$\dfrac{1}{y}=a+\dfrac{b}{x}$		$v=\dfrac{1}{y}$ $u=\dfrac{1}{x}$

函数名称	函数表达式	图像	线性化变换
幂函数	$y=ax^b$		$v=\ln y$ $u=\ln x$
指数函数	$y=ae^{bx}$		$v=\ln y$ $u=x$
	$y=ae^{b/x}$		$v=\ln y$ $u=\dfrac{1}{x}$
对数函数	$y=a+b\ln x$		$v=y$ $u=\ln x$
S 型曲线	$y=\dfrac{1}{a+be^{-x}}$		$v=\dfrac{1}{y}$ $u=e^{-x}$

2. 参数估计　通过适当变换，把非线性函数转化为线性函数形式，然后对未知参数寻求最小二乘估计.譬如由$\dfrac{1}{y}=a+\dfrac{b}{x}$可转化为 $v=a+bu$ 只要令$\dfrac{1}{y}=v,\dfrac{1}{x}=u$ 即可.

3. 评价标准 常用的曲线回归方程好坏的评价标准有两个:

(1) 决定系数 $R^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$,愈大愈好;

(2) 剩余标准差 $s=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}$愈小愈好.

这两个评价标准是一致的,只是从两个侧面作出评价.