第四章　线性模型

1.变量之间的关系一般分为两类：

(1)完全确定的关系，也就是变量之间的关系可以用函数解析式表达出来；如$y = f(x)$

(2)非确定的关系，y与x的取值有关，但不能完全确定，也称为相关关系。

2.回归分析研究的主要内容：

(1)通过观察或实验数据的处理，找出变量间相关关系的定量数学表达式–经验公式，即进行参数估计，并确定经验回归方程的具体形式；

(2)检验所建立的经验回归方程是否合理；

(3)利用合理的回归方程对随机变量Y进行预测和控制；

例1：为研究商品价格与销售量之间的关系，现在收集了某商品在一个地区25个时间段内平均价格x(元) 和销售总额y(万元)，试研究y与x之间的关系。

x	35.3	29.7	30.8	58.8	61.4	71.3	74.4	76.7	70.7
y	10.98	11.13	12.51	8.40	9.27	8.73	6.36	8.50	7.82
x	46.4	28.9	28.1	39.1	46.8	48.5	59.3	70.0	70.0
y	8.24	12.19	11.88	9.57	10.94	9.58	10.09	8.11	6.83
x	72.1	58.1	44.6	33.4	28.6	57.7	74.5		
y	7.68	8.47	8.86	10.36	11.08	9.14	8.88		

第一节　一元线性回归模型

第一节　一元线性回归模型

一.基本概念

1.定义:设回归变量x与响应变量(因变量)y之间有这样的关系，$y = \beta_0 + \beta_1 x + \varepsilon$,其中$\beta_0, \beta_1$ 是未知参数，ε是随机项，且假定$E(\varepsilon) = 0, D(\varepsilon) = \sigma^2$,则称此模型为一元线性回归模型.
若$\varepsilon \sim N(0, \sigma^2)$，则称为一元正态线性回归模型.

2.设n对观测值$(x_i, y_i), i = 1, \cdots, n$，每一对$(x_i, y_i)$都有$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$,且$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2, i = 1, \cdots, n$，$Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$.
由此可以估计β_0, β_1，记为$\hat{\beta}_0, \hat{\beta}_1$，称其为回归系数. 记

$$\tilde{y}_i = \beta_0 + \beta_1 x_i.$$

第一节 一元线性回归模型

二.参数β_0, β_1的估计

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

表示总的偏离平方和

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0; \\ \frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0. \end{cases}$$

此方程组称为正规方程组
得方程组

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i; \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i, \end{cases}$$

第一节 一元线性回归模型

解得

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

记$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2, S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

则$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$分别称为$\beta_0, \beta_1$的最小二乘估计，这种方法称为最小二乘法, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 为经验回归直线.

第一节　一元线性回归模型

性质1： 对于一元线性回归模型有

(1) $E(\hat{\beta}_0) = \beta_0$，$E(\hat{\beta}_1) = \beta_1$,即$\hat{\beta}_0, \hat{\beta}_1$分别是$\beta_0, \beta_1$的无偏估计；

(2) $Var(\hat{\beta}_0) = (\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}})\sigma^2 = \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}\sigma^2$

$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{x}}{S_{xx}}\sigma^2$

性质2： $Cov(\hat{\beta}_1, \overline{y}) = 0$, 即$\hat{\beta}_1$与$\overline{y}$不相关.

性质3： $\hat{\beta}_0$，$\hat{\beta}_1$ 分别是β_0，β_1的最小方差线性无偏估计.

第一节　一元线性回归模型

三.参数σ^2的估计

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \cdots, n$为回归值，则称$\hat{\varepsilon}_i = y_i - \hat{y}_i$为第i个残差，$i = 1, \cdots, n$.

定义

$$Q_e = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

为残差平方和，它代表y_i与经验回归直线上点的纵坐标$\hat{y}_i$的离差平方和,反映了试验的随机误差.

性质4. $\hat{\sigma}^2 = \frac{Q_e}{n-2}$为$\sigma^2$的无偏估计.

第一节　一元线性回归模型

四.回归方程的显著性检验

$$H_0: \beta_1 = 0 \leftrightarrow H_1: \beta_1 \neq 0$$

当H_0成立时，说明y与x之间无线性相关关系；

当H_0不成立时，说明y与x之间线性相关关系显著。

第一节　一元线性回归模型

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2 \\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = Q_e + U \end{aligned}$$

其中$Q_e = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$,

$$U = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1 x_i-\hat{\beta}_0-\hat{\beta}_1\bar{x})^2 = \hat{\beta}_1{}^2 S_{xx},$$

S_{yy}为总的离差平方和，U为回归平方和，表示回归值$\hat{y}_i$的波动，Q_e为剩余(残差)平方和，反映了随机误差的存在而引起的因变量的波动.

第一节 一元线性回归模型

定理:若随机误差$\varepsilon \sim N(0, \sigma^2)$, 则

$$(1)\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}}), \quad (2)\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2),$$

$$(3)\frac{Q_e}{\sigma^2} \sim \chi^2(n-2), Q_e \text{与} \hat{\beta}_1 \text{相互独立},$$

$$(4)\text{在} H_0(\beta_1 = 0) \text{成立条件下}, \frac{U}{\sigma^2} \sim \chi^2(1),$$

$$F = \frac{U}{Q_e/(n-2)} = (n-2)\frac{U}{Q_e} \sim F(1, n-2),$$

$$(5)\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2), \text{其中} \hat{\sigma}^2 = \frac{Q_e}{n-2}$$

第一节　一元线性回归模型

设$\varepsilon \sim N(0, \sigma^2)$

1. F检验法

统计量$F = (n-2)\frac{U}{Q_e} \overset{H_0}{\sim} F(1, n-2)$, 对给定的显著性水平$\alpha$,拒绝域为$W = \{f > F_{1-\alpha}(1, n-2)\}$.

2. t检验法

$T = \frac{\hat{\beta}_1\sqrt{S_{xx}}}{\hat{\sigma}} \overset{H_0}{\sim} t(n-2)$ 对给定的显著性水平α,拒绝域为$W = \{|t| > t_{1-\frac{\alpha}{2}}(n-2)\}$

注:

(1)当落入拒绝域时，拒绝H_0,即认为y与x之间线性关系显著，或者说回归方程是有意义的;

(2)否则认为回归方程不合理，这种情况由多种原因引起。

第一节　一元线性回归模型

3. 相关系数检验法(判定系数法)

$$R=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}=\hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}},\quad R^2=\hat{\beta}_1{}^2\frac{S_{xx}}{S_{yy}}$$

R: 线性相关系数; R^2: 相关指数(判定系数), $|R|, R^2$越接近1, 说明线性相关程度越高. 对给定的显著性水平α, 拒绝域为$W=\{|R|>c\}$.

第一节　一元线性回归模型

注：$R^2 = U/S_{yy}$：回归平方和在总离差平方和中的比例，
T, F, R^2之间的关系：

$$Q_e = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{yy}(1 - \frac{S_{xy}^2}{S_{xx}S_{yy}}) = S_{yy}(1 - R^2),$$

$$U = S_{yy} - Q_e = R^2 S_{yy}, \quad \therefore F = (n-2)\frac{R^2}{1-R^2},$$

$T^2 = F = (n-2)\frac{R^2}{1-R^2}$，故三种检验在本质上是一致的，大部分软件都采用F检验。

第一节 一元线性回归模型

五.预测

1.点预测： 当给定x_0时，$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$就是$y_0$的点预测。
2.区间预测： 当给定x_0时，y_0的置信度为$1-\alpha$的置信区间，称为预测区间，即寻找y_1, y_2，使$P\{y_1 \le y_0 \le y_2\} = 1-\alpha$.
当$\varepsilon \sim N(0, \sigma^2)$，$y_0$的置信水平为$1-\alpha$的置信区间(即预测区间)为

$$[\hat{y}_0 - \delta(x_0),\ \hat{y}_0 + \delta(x_0)]$$

其中

$$\delta(x_0) = \hat{\sigma} t_{1-\frac{\alpha}{2}}(n-2)\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

第一节 一元线性回归模型

注：

(1)这个区间长度为$2\delta_0$，中心在$\hat{y}_0$，n, α固定，$Q_e \downarrow, S_{xx} \uparrow$，可提高预测精度.

(2)由样本观测值可以作两条曲线$y_1(x) = \hat{y}(x) - \delta(x), y_2(x) = \hat{y}(x) + \delta(x)$,这两条曲线把回归直线$\hat{y}(x) = \beta_0 + \beta_1 x$夹在中间，形成一条宽窄不等的区域，这个区域在$x = \bar{x}$处最窄.

(3)在利用回归方程进行预测时，样本容量不能太小，因为小样本也许不能真实反映变量之间的结构关系.

(4) n很大时，$t_{1-\frac{\alpha}{2}}(n-2) \approx u_{1-\frac{\alpha}{2}}$，若$\bar{x}$离$x_0$不太远，$y_0$的置信水平近似为$1-\alpha$的置信区间(即预测区间)为$[\hat{y}_0 - \hat{\sigma} u_{1-\frac{\alpha}{2}}, \hat{y}_0 + \hat{\sigma} u_{1-\frac{\alpha}{2}}]$

第一节　一元线性回归模型

3.控制：

控制问题是预测的反问题，若要y在某个范围$y_1 \le y \le y_2$，则变量x应控制在何处，即确定x_1, x_2 使

$$\begin{cases} \hat{y}(x_1) - \delta(x_1) \ge y_1 \\ \hat{y}(x_2) + \delta(x_2) \le y_2 \end{cases}$$

则当$x \in [\min\{x_1, x_2\}, \max\{x_1, x_2\}]$时，就以至少$1-\alpha$的概率保证x相对应的y落在区间$[y_1, y_2]$内。

第一节　一元线性回归模型

注： (1) β_1的置信水平为$1-\alpha$的置信区间

$$\because P\{-t_{1-\frac{\alpha}{2}}(n-2) \leq \frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}}\sqrt{S_{xx}} \leq t_{1-\frac{\alpha}{2}}(n-2)\} = 1-\alpha$$

$$\therefore [\hat{\beta}_1 - t_{1-\frac{\alpha}{2}}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{xx}}}]$$

(2) β_0的置信水平为$1-\alpha$的置信区间

$$\because P\{-t_{1-\frac{\alpha}{2}}(n-2) \leq \frac{\hat{\beta}_0-\beta_0}{\sigma\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}}} / \sqrt{\frac{Q_e}{\sigma^2}/(n-2)} \leq t_{1-\frac{\alpha}{2}}(n-2)\} = 1-\alpha$$

$$\therefore [\hat{\beta}_0 - t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}, \hat{\beta}_0 + t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}]$$

第一节 一元线性回归模型

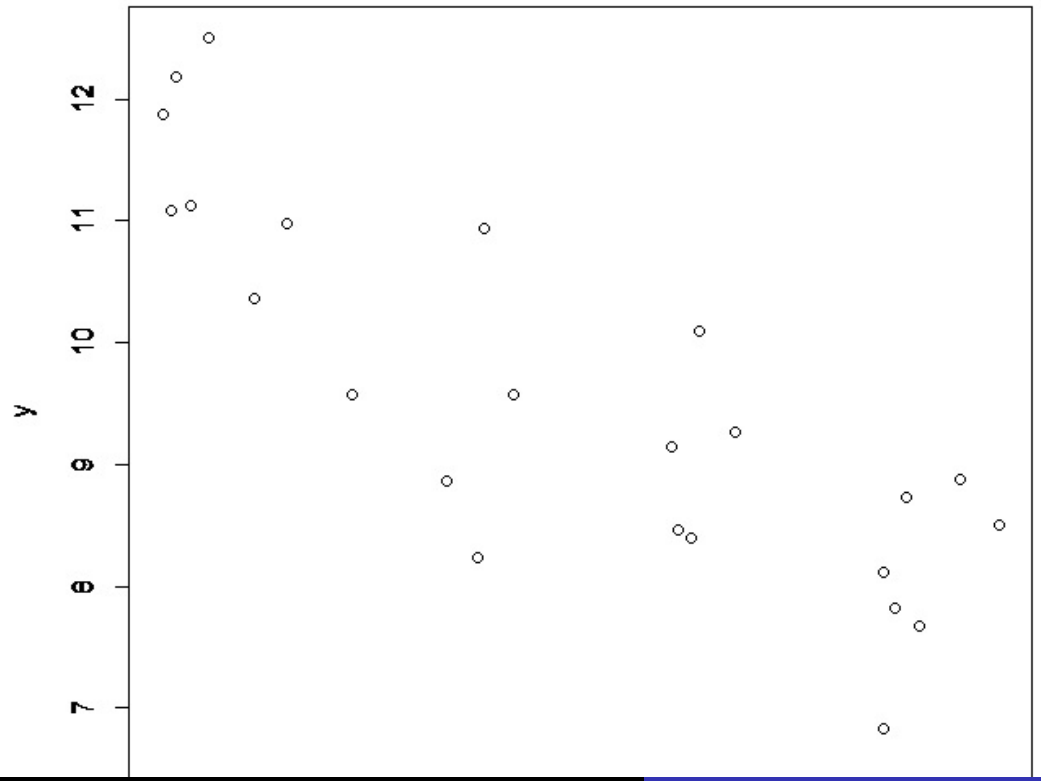
例2：为研究商品价格与销售量之间的关系，现在收集了某商品在一个地区25个时间段内平均价格x(元) 和销售总额y(万元)，试研究y与x之间的关系。

x	35.3	29.7	30.8	58.8	61.4	71.3	74.4	76.7	70.7
y	10.98	11.13	12.51	8.40	9.27	8.73	6.36	8.50	7.82
x	46.4	28.9	28.1	39.1	46.8	48.5	59.3	70.0	70.0
y	8.24	12.19	11.88	9.57	10.94	9.58	10.09	8.11	6.83
x	72.1	58.1	44.6	33.4	28.6	57.7	74.5		
y	7.68	8.47	8.86	10.36	11.08	9.14	8.88		

一元线性回归模型例1

解：

(1)画散点图



第一节　一元线性回归模型

(2)确定回归方程

由样本数据$\bar{x} = \frac{1}{25}\sum_{i=1}^{25} x_i = 52.60$, $\bar{y} = \frac{1}{25}\sum_{i=1}^{25} y_i = 9.424$

因而

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{25} x_i y_i - \frac{1}{25}(\sum_{i=1}^{25} x_i)(\sum_{i=1}^{25} y_i)}{\sum_{i=1}^{25} x_i^2 - 25\bar{x}^2} = -0.0798$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 13.623$$

所以$\hat{y} = 13.623 - 0.0798x$

第一节　一元线性回归模型

3)对回归方程进行显著性检验

利用F检验法：$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$

$$F = \frac{U}{Q_e/(n-2)} \overset{H_0}{\sim} F(1, n-2)$$

$S_{yy} = \sum_{i=1}^{25}(y_i - \bar{y})^2 = 63.82$，$U = \sum_{i=1}^{25}(\hat{y}_i - \bar{y})^2 = 45.59$,
$Q_e = \sum_{i=1}^{25}(y_i - \hat{y}_i)^2 = 18.82$

取$\alpha = 0.05$，则$F_{0.95}(1, 23) = 4.28$,而$f = 57.56 > 4.28$,所以拒绝H_0 即认为回归方程反映了该商品销售总额y与其价格x之间的相关关系。

(4)同样可计算$R^2 = \frac{U}{S_{yy}} = 0.714$

第一节 一元线性回归模型

(5)利用回归方程对销售额进行预测

当$x = x_0 = 28.6$元时，则$\hat{y}_0 = 11.34$万元

对$\alpha = 0.05$,查表$t_{1-\frac{\alpha}{2}}(23) = 2.069$,可得$\delta(x_0) = 1.95$

所以y_0的置信水平为95%的预测区间

为$[(\hat{y}_0 - \delta(x_0), \hat{y}_0 + \delta(x_0))] = [9.39, 13.29]$，即把价格定在28.6元时，销售总额落在区间[9.39,13.29] 的概率为95%。

(6)若使销售额控制在[10,12]之间，则由

$$\begin{cases} 10 = 13.623 - 0.0798x_1 \\ 12 = 13.623 - 0.0798x_2 \end{cases}$$

得到x的范围为[20.34,45.40].

第一节　一元线性回归模型

六.可化为一元线性回归的模型

1. 双曲线方程:$\frac{1}{y}=a+\frac{b}{x}$;
   令$y'=\frac{1}{y},x'=\frac{1}{x}$,则有$y'=a+bx'$。

2. 幂函数方程:$y=ax^b$;
   令$y'=\ln y,x'=\ln x,a'=\ln a$,则$y'=a'+bx'$.

3. 指数曲线方程:$y=a\mathrm{e}^{bx}$;
   令$y'=\ln y,a'=\ln a$,则$y'=a'+bx$.

第一节 一元线性回归模型

4. 指数曲线方程:$y = ae^{\frac{b}{x}}$;
   令$y' = \ln y, x' = \frac{1}{x}, a' = \ln a$,则$y' = a' + bx'$.

5. 对数曲线方程:$y = a + b\ln x$;
   令$x' = \ln x$,则$y = a + bx'$.

6. S型曲线方程:$y = \frac{1}{a+be^{-x}}$.
   令$y' = \frac{1}{y}, x' = e^{-x}$,则$y' = a + bx'$.

第二节　多元线性回归模型的参数估计

第二节　多元线性回归模型的参数估计

一.数学模型

假设随机变量y与k个变量$x_1, \cdots, x_k$之间存在下面的线性关系

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

其中ε是一个随机变量,满足$E\varepsilon = 0, D\varepsilon = \sigma^2(\sigma^2$为未知常数),称为随机误差, $\beta_0, \beta_1, \cdots, \beta_k$是未知参数.
设有n组独立的观测值$(y_i, x_{i1} \cdots x_{ik}), i = 1, \cdots, n$,则有

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1, \\ \quad \cdots\cdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n, \end{cases}$$

成立.

第二节 多元线性回归模型的参数估计

或写成矩阵形式为

$$Y = X\beta + \varepsilon,$$

其中

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

这里Y表示随机变量y的n次观测值组成的列向量,称为观测向量,X的元素是k个自变量$x_1, \cdots, x_k$在n次观测中的取值, β称为未知参数向量,ε称为随机误差向量,满足

$$E(\varepsilon) = 0, \quad Cov(\varepsilon, \varepsilon) = \Sigma.$$

第二节　多元线性回归模型的参数估计

称模型

$$Y = X\beta + \varepsilon, \quad E(\varepsilon) = 0, \quad Cov(\varepsilon, \varepsilon) = \Sigma$$

为n元线性回归模型.

若

$$Y = X\beta + \varepsilon, \varepsilon \sim N(\mathbf{0}, \sigma^2 I_n),$$

则称为多元正态线性回归模型.

第二节 多元线性回归模型的参数估计

二.参数β的估计

仍采用最小二乘法，求$y_i-(\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik})$的平方和。令$Q(\beta)=\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik})]^2$达到最小，若用矩阵表示，

$$Q(\beta)=(Y-X\beta)'(Y-X\beta)=Y'Y-2\beta'X'Y+\beta'X'X\beta$$

$$\frac{\partial Q}{\partial \beta}=-2X'Y+2X'X\beta=0$$

所以

$$X'X\beta=X'Y \tag{1}$$

若X是列满秩矩阵，$(X'X)^{-1}$存在，则$\hat{\beta}=(X'X)^{-1}X'Y$

第二节　多元线性回归模型的参数估计

正规方程组的解与β的最小二乘估计有以下关系：

定理：
(1)正规方程组的解必是β的最小二乘估计；
(2)β的最小二乘估计必是正规方程组的解.

因而$\hat{\beta}$即为β的最小二乘估计，$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)'$.

第二节 多元线性回归模型的参数估计

最小二乘估计的性质

性质1: $E(\hat{\beta}) = \beta$, $\hat{\beta}$是β的线性无偏估计.

性质2: $Cov(\hat{\beta}, \hat{\beta}) = \sigma^2(X'X)^{-1}$.

注: $\hat{\beta}$的各分量$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$一般并不独立.

$$S^{-1} := (X'X)^{-1} = \begin{pmatrix} c_{00} & c_{01} & \cdots & c_{0k} \\ c_{10} & c_{11} & \cdots & c_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{pmatrix}$$

第二节　多元线性回归模型的参数估计

$$Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}, \quad Var(\hat{\beta}_i) = \sigma^2 c_{ii}, i, j = 0, 1, ..., k.$$

选择使c_{ij}尽可能小的设计矩阵X. 如

$$c_{ij} = \delta_{ij} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \quad i, j = 0, 1, ..., k,$$

则称相应的X为正交的.

第二节 多元线性回归模型的参数估计

定义: 对任一$k+1$维向量$C=(c_0,c_1,...,c_k)'$, 若存在n维列向量L, 使$E(L'Y)=C'\beta$, 则称$C'\beta$为可估函数, 而可估函数$C'\beta$的最小方差线性无偏估计, 称为它的最好线性无偏估计(BLUE).

性质3(Guass-Markov定理): $C'\hat{\beta}$是$C'\beta$的最好线性无偏估计, 其中$\hat{\beta}$是β的最小二乘估计.

注: $C'\hat{\beta}$不一定是$C'\beta$的一切无偏估计中方差最小的.

性质4: 若$\varepsilon\sim N(0,\sigma^2 I_n)$, 则$C'\hat{\beta}$是$C'\beta$的UMVUE.

第二节　多元线性回归模型的参数估计

三.参数σ^2的估计

用残差向量$\hat{\varepsilon} = Y - X\hat{\beta}$来构造方差$\sigma^2$的估计.

性质5: (1) $E(\hat{\varepsilon}) = 0$, (2)$Cov(\hat{\varepsilon}, \hat{\varepsilon}) = \sigma^2 \left[I_n - X(X'X)^{-1}X'\right]$, (3)$Cov(\hat{\beta}, \hat{\varepsilon}) = 0$.

残差平方和$Q_e = \|\hat{\varepsilon}\|^2 = \hat{\varepsilon}'\hat{\varepsilon}$.

第二节　多元线性回归模型的参数估计

$$
\begin{aligned}
Q_e &= \|\hat{\varepsilon}\|^2 = \hat{\varepsilon}'\hat{\varepsilon} \\
&= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\
&= (AY)'(AY) = Y'AY \\
&= Y'Y - Y'X(X'X)^{-1}(X'X)(X'X)^{-1}X'Y \\
&= Y'Y - \hat{\beta}'X'X\hat{\beta} = Y'Y - \hat{Y}'\hat{Y}
\end{aligned}
$$

第二节 多元线性回归模型的参数估计

(1) 设n维随机向量Y满足$E(Y)=a, Cov(Y,Y)=\sigma^2 I_n$, A为n阶对称常数阵, 则

$$E(Y'AY)=a'Aa+\sigma^2 tr(A), \quad tr(A)=\sum_{i=1}^{n} a_{ii}.$$

(2) 设A,B s.t. AB, BA均为方阵, 则$tr(AB)=tr(BA)$, 特别有$tr(AA')=tr(A'A), tr(ab')=tr(b'a)=b'a$, a,b为相同维数向量.

(3) tr(A+B)=tr(A)+tr(B)

第二节　多元线性回归模型的参数估计

性质6: 记$\hat{\sigma}^2 = Q_e/(n-k-1)$, 称为残差方差, 则有$E(\hat{\sigma}^2) = \sigma^2$.

性质7: 若$\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I_n})$, 则
(1) $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$, $\hat{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{A})$ 且二者相互独立.

(2) $\hat{\beta}, Q_e$相互独立.　　(3) $\frac{Q_e}{\sigma^2} \sim \chi^2(n-k-1)$

(4) β的最小二乘估计$\hat{\beta}$也是β的极大似然估计, σ^2的极大似然估计为$\frac{Q_e}{n}$.

第二节　多元线性模型的参数估计

四.线性回归模型的中心化处理

设有n组独立的观测值$(y_i, x_{i1} \cdots x_{ik}), i = 1, \cdots, n$和$(\overline{x}_1, ..., \overline{x}_k; \overline{y})$

$$\begin{cases} y_1 - \overline{y} = \widetilde{\beta}_0 + \beta_1(x_{11} - \overline{x}_1) + \cdots + \beta_k(x_{1k} - \overline{x}_k) + \varepsilon_1, \\ \quad \cdots\cdots \\ y_n - \overline{y} = \widetilde{\beta}_0 + \beta_1(x_{n1} - \overline{x}_1) + \cdots + \beta_k(x_{nk} - \overline{x}_k) + \varepsilon_n, \end{cases}$$

其中$\widetilde{\beta}_0 := \beta_0 + \beta_1\overline{x}_1 + ... + \beta_k\overline{x}_k - \overline{y}$.

第二节 多元线性模型的参数估计

$$\widetilde{Y} = \begin{pmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix}, \ \widetilde{X} = \begin{pmatrix} x_{11} - \overline{x}_1 & \cdots & x_{1k} - \overline{x}_k \\ x_{21} - \overline{x}_1 & \cdots & x_{2k} - \overline{x}_k \\ \vdots & \vdots & \vdots \\ x_{n1} - \overline{x}_1 & \cdots & x_{nk} - \overline{x}_k \end{pmatrix}, \boldsymbol{\beta}_1 = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \widetilde{Y} = (\mathbf{1} \ \ \widetilde{X}) \begin{pmatrix} \widetilde{\beta}_0 \\ \boldsymbol{\beta}_1 \end{pmatrix} + \varepsilon,$$

$$\mathbf{1} = (1, ..., 1)', \ \ \mathbf{1}'\widetilde{X} = \mathbf{0}, \ \ \mathbf{1}'\widetilde{Y} = 0.$$

第二节　多元线性模型的参数估计

中心化线性回归模型的最小二乘估计

$$\begin{pmatrix} \hat{\tilde{\beta}}_0 \\ \hat{\beta}_1 \end{pmatrix} = \left[(\mathbf{1}\ \tilde{X})'(\mathbf{1}\ \tilde{X})\right]^{-1}(\mathbf{1}\ \tilde{X})'\tilde{Y} = \begin{pmatrix} \frac{1}{n} & \mathbf{0} \\ \mathbf{0} & (\tilde{X}'\tilde{X})^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{0} \\ \tilde{X}'\tilde{Y} \end{pmatrix}$$

$$\therefore \begin{cases} \hat{\tilde{\beta}}_0 &= 0 \\ \hat{\beta}_1 &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}. \end{cases}$$

所以可只考虑$\tilde{Y} = \tilde{X}\beta_1 + \varepsilon$.

第三节　多元线性回归模型的假设检验

一.回归方程的显著性检验

提出假设:

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

(1)若接受H_0, 则表明诸变量与y之间确实无线性相关关系;

(2)若拒绝H_0,则认为回归方程是有意义的，但是这个结论只说明至少有一个$\beta_i \neq 0$,也就是说在所选自变量中，至少有一部分对y来说是重要的，但不表示所有自变量都是重要的。

第三节 多元线性回归模型的假设检验

1.平方和分解： $\hat{Y} = X\hat{\beta}$为n个试验点处Y的回归值, 总的偏差平方和定义为

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = (Y - \mathbf{1}\bar{y})'(Y - \mathbf{1}\bar{y}) \\ &= \|Y - \mathbf{1}\bar{y}\|^2 = \|Y - \hat{Y} + \hat{Y} - \mathbf{1}\bar{y}\|^2 \\ &= \|Y - \hat{Y}\|^2 + \|\hat{Y} - \mathbf{1}\bar{y}\|^2 + 2(Y - \hat{Y})'(\hat{Y} - \mathbf{1}\bar{y}) \end{aligned}$$

$$\begin{aligned} (Y - \hat{Y})'(\hat{Y} - \mathbf{1}\bar{y}) &= (Y - \hat{Y})'\hat{Y} - (Y - \hat{Y})'\mathbf{1}\bar{y} \\ &= (AY)'(PY) - (Y - \hat{Y})'\mathbf{1}\bar{y} \\ &= Y'APY - (Y - \hat{Y})'\mathbf{1}\bar{y} = 0. \end{aligned}$$

$$S_{yy} = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \mathbf{1}\bar{y}\|^2 = Q_e + U.$$

$Q_e = \|Y - \hat{Y}\|^2$为剩余(残差)平方和, $U = \|\hat{Y} - \mathbf{1}\bar{y}\|^2$为回归平方和, 若$U \gg Q_e$, 则拒绝$H_0$.

第三节 多元线性回归模型的假设检验

2.构造统计量:

(1) F统计量

当$\varepsilon \sim N(0, \sigma^2 I_n)$, $Q_e/\sigma^2 \sim \chi^2(n-k-1)$, $S_{yy}/\sigma^2 \overset{H_0}{\sim} \chi^2(n-1)$.

$$S_{yy} = Q_e + U, \quad Q_e = Y'AY, \quad r(A) = n-k-1,$$

由Cochran分解定理可证明$U/\sigma^2 \overset{H_0}{\sim} \chi^2(k)$, Q_e与U相互独立, 则选取

$$F = \frac{U/k}{Q_e/(n-k-1)} \sim F(k, n-k-1)$$

对给定的显著性水平α,拒绝域$W = \{F > F_{1-\alpha}(k, n-k-1)\}$.

第三节 多元线性模型的假设检验

(2) $R^2 = \frac{U}{S_{yy}}$, 称R^2为全相关系数.
它刻画了全体自变量$x_1, \ldots, x_k$对于因变量y的线性相关程度.R^2越大,越接近于1,说明上述线性相关程度越显著,R^2可作为衡量回归方程总效果的一个数量指标.

注：$F = \frac{n-k-1}{k} \frac{R^2}{1-R^2}$, 所以F检验与$R^2$检验是等价的。

第三节 多元线性模型的假设检验

二.回归系数的显著性检验

检验x_i对y的影响是否显著，等价于检验回归系数$H_{0i}:\beta_i=0$
(1)若接受H_{0i}，则表明x_i对y的影响相对于整个模型来说比较小；
(2)若拒绝H_{0i}，则表明x_i对y确有一定的影响，称x_i为显著因子。

当$\varepsilon\sim N(0,\sigma^2 I_n)$, 若记$S=(c_{ij})=(X'X)^{-1}$,则由性质知$\hat{\beta}_i\sim N(\beta_i,\sigma^2 c_{ii})$, 且$\hat{\beta}_i$与$Q_e$相互独立，其中$c_{ii}$表示矩阵$S$的第i个对角元。

第三节 多元线性模型的假设检验

1. T检验：选取

$$T = \frac{\hat{\beta}_i}{\sqrt{c_{ii}}} \Big/ \sqrt{\frac{Q_e}{n-k-1}} \overset{H_{0i}}{\sim} t(n-k-1),$$

对给定的显著性水平α, 拒绝域

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

2. F检验：选取

$$F = \frac{\hat{\beta}_i^2}{c_{ii}} \Big/ \frac{Q_e}{n-k-1} \overset{H_{0i}}{\sim} F(1, n-k-1)$$

对给定的显著性水平α, 拒绝域

$$W = \{f > F_{1-\alpha}(1, n-k-1)\}$$

第三节 多元线性模型的假设检验

三.偏回归平方和

自变量对y的影响,是指从回归方程剔除了这个自变量后所造成的影响，称回归平方和的减少部分为y对这个自变量的偏回归平方和。

若在$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$中剔除自变量$x_i$,不能简单地抹去这一项而得到

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{i-1} x_{i-1} + \hat{\beta}_{i+1} x_{i+1} + \cdots + \hat{\beta}_k x_k,$$

应该重新估计回归系数,建立新的回归方程

$$\hat{y}^* = \hat{\beta}_0^* + \hat{\beta}_1^* x_1 + \cdots + \hat{\beta}_{i-1}^* x_{i-1} + \hat{\beta}_{i+1}^* x_{i+1} + \cdots + \hat{\beta}_k^* x_k.$$

第三节 多元线性模型的假设检验

一般地$\hat{\beta}_j^* \neq \hat{\beta}_j$, 下面给出中心化线性回归模型下偏回归平方和的表示.

回归平方和$U = \hat{\beta}_1^{\mathrm{T}} \widetilde{X}^{\mathrm{T}} \widetilde{Y} = \sum_{i=1}^{k} \hat{\beta}_i S_{i0}$, $S_{i0} = \sum_{l=1}^{n} (x_{li} - \bar{x}_i)(y_l - \bar{y})$ $i = 1, 2, ..., k$.

剔除自变量x_i后的回归平方和 $U_i' = \sum_{j \neq i} \hat{\beta}_j^* S_{j0}$,

y对x_i的偏回归平方和$U_i = U - U_i' = \hat{\beta}_i^2 / c_{ii}$.

第三节　多元线性模型的假设检验

注:

(1) 回归系数显著性检验的F统计量的分子即为偏回归平方和. 偏回归平方和越大, 此变量对y 的影响越显著.

(2) 得到回归方程后, 计算每个变量的偏回归平方和, 对偏回归平方和最小的变量, 如果相应的回归系数检验又不显著, 可将此变量剔除.

第三节 多元线性模型的假设检验

四.“最优”回归方程的选择

最优回归方程: 如回归方程中包含所有对y有显著影响的自变量, 不包含对y影响不显著的自变量, 同时在同类方程中残差平方和Q_e达到最小, 则称此回归方程为最优的.

(1) 全部比较法: 从所有可能的自变量组合的回归方程中选择最优者.
注: 总可找到最优方程; 但计算量大.

(2) 只出不进法: 从包含全部自变量的回归方程中逐个剔除不显著的自变量, 直到回归方程中所含自变量全部都是显著的为止. 首先考虑含所有自变量的回归方程, 剔除不显著自变量中偏回归平方和最小的, 再对其中的每个自变量进行显著性检验, 继续剔除, 直到所有自变量都显著.

第三节 多元线性模型的假设检验

四.“最优”回归方程的选择

注: 每剔除一次自变量就得重新计算回归系数, 考虑自变量不多时, 不显著自变量不多时, 可采用. 不显著自变量较多时, 计算量大.

(3) 只进不出法: 从一个自变量开始, 把显著的自变量逐个引入回归方程, 直到在余下的自变量中选出一个与已引入的自变量一起组成回归方程有最大偏回归平方和的自变量, 至经检验为不显著, 因而不被引入时为止.
注: 计算量少, 但不一定能得到最优方程. 由于自变量间的相关关系, 引入新的自变量后, 使原来引入的自变量成为不显著的.

第三节 多元线性模型的假设检验

四.“最优”回归方程的选择

(4) 逐步回归法: 综合方法(2)、(3), 将自变量按其对y的影响一个个引入, 同时每引入一个新的自变量, 即对原已引入的自变量逐个检验, 将不显著的剔除, 直到回归方程再也不能引入新的自变量, 同时也不能剔除任一自变量为止.

注: 有计算技巧, 计算量相对较小, 有较好的计算程序.

第四节　非线性回归模型

第四节　非线性回归模型

多项式回归

若随机变量y与自变量x之间的相关系数为

$$\begin{cases} y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon; \\ (\varepsilon \sim N(0, \sigma^2)). \end{cases}$$

称此模型为(正态)多项式回归模型。

只需令$x_i = x^i, i = 1, \cdots, k$，则可转化为多元线性回归模型

$$\begin{cases} Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon; \\ (\varepsilon \sim N(0, \sigma^2)). \end{cases}$$

第四节 非线性回归模型

例： 某种合金钢中的两种主要成分之和x与它的膨胀系数y之间有一定的数量关系，给出实验所得的13 组数据，求y与x的回归方程。

x	37	37.5	38	38.5	39	39.5	40	40.5	41	41.5	42	42.5	43
y	3.4	3	3	3.27	2.1	1.83	1.53	1.70	1.8	1.9	2.35	2.54	2.9

解： 先画散点图

设回归方程为$y = \beta_0 + \beta_1 x + \beta_2 x^2$

令$x_1 = x, x_2 = x^2$,则可确定回归方程为$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$,

经计算，$\hat{\beta}_0 = 257.063, \hat{\beta}_1 = -12.620, \hat{\beta}_2 = 0.156$

所以y对x的多项式回归方程为

$$\hat{y} = 257.063 - 12.620x + 0.156x^2$$

第五节　单因子试验方差分析

第五节　单因子试验方差分析

一.基本概念

1.指标–试验的结果(如产品的性能，质量，产量等)，用Y表示；

2.因子–试验中变化的因素(即影响指标的原因)，用A,B,C表示；

3.水平–因子在试验中所处的不同状态，如因子A有n个水平，用$A_1, A_2, \cdots, A_n$表示；

4.若试验中只有一个因素在变化，其他条件不变，则称为单因子试验，处理单因子试验的统计推断方法称为单因子方差分析。

第五节 单因子试验方差分析

例1：5个水稻产品比较试验，在成熟期随机抽取样本测定产量，每个品种取3个点，结果如下表

A	1	2	3	$\bar{Y}_{i.}$
A_1	41	39	40	40
A_2	33	37	35	35
A_3	38	35	35	36
A_4	37	39	38	38
A_5	31	34	34	33

指标：产量；因子：品种；水平：A_1, A_2, A_3, A_4, A_5

第五节　单因子试验方差分析

由表中数据我们可以分析：
(1)在同一水平$A_i(i=1,\cdots,5)$下，生产的条件虽然相同，但产量却有所不同，产生这种差异的原因是由于试验过程中随机因素的干扰及测量误差所致，称这类差异为随机误差或试验误差。说明试验结果是一个随机变量。
(2)5个不同的品种，从平均产量来看，它们是参差不齐的，其原因主要是由于品种的不同引起的差异（除了随机波动），称这类差异为系统误差。
(3)对同一品种进行3次重复试验的结果可看成是取自同一个总体的样本，表中的5组数据可以看成是来自5个不同总体的样本，记这些总体为$Y_1, Y_2, \cdots, Y_5$,每个试验结果记为$Y_{ij}, i=1,\cdots,5, j=1,2,3$

第五节　单因子试验方差分析

通常假定：1*.$Y_i \sim N(\mu_i, \sigma^2), i = 1, \cdots, 5,$(其它试验条件不变，因而认为所有试验的方差是相同的)。

2*.$Y_{i1}, Y_{i2}, \cdots, Y_{in_i}$是来自$Y_i$的样本，$n_i = 3, i = 1, \cdots, 5,$且$Y_1, Y_2, \cdots, Y_5$相互独立。

(4)设因子A有a个水平，每个水平A_i重复n_i次，若重复数n_i全相等，则称这类试验为等重复的单因子试验；反之，则称为不等重复的单因子试验。

(5)本例分析判断5个不同品种的产量之间的差异主要是由随机误差还是由于不同品种造成的问题，可归结为判定5个正态总体的均值是否相等的问题。若5个正态总体的均值相等，则认为产量之间的差异是由随机误差造成的；否则，认为产量之间的差异是由不同品种造成的。

第五节 单因子试验方差分析

二.数学模型

设因子A有a个不同水平$A_1, \cdots, A_a$,它们对应的总体$Y_1, \cdots, Y_a$相互独立，且$Y_i \sim N(\mu_i, \sigma^2), i = 1, \cdots, a$. 在水平$A_i$下进行$n_i$次独立观测，获得容量为$n_i$的一个样本$Y_{i1}, Y_{i2}, \cdots, Y_{in_i}, i = 1, \cdots, a$

水平	总体	样本
A_1	$Y_1 \sim N(\mu_1, \sigma^2)$	$Y_{11}, Y_{12}, \cdots, Y_{1n_1}$
A_2	$Y_2 \sim N(\mu_2, \sigma^2)$	$Y_{21}, Y_{22}, \cdots, Y_{2n_2}$
$\vdots$	$\vdots$	$\vdots$
A_a	$Y_a \sim N(\mu_a, \sigma^2)$	$Y_{a1}, Y_{a2}, \cdots, Y_{an_a}$

第五节 单因子试验方差分析

令$\varepsilon_{ij} = Y_{ij} - \mu_i$,则

$$\begin{cases} Y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, \cdots, a, j = 1, \cdots, n_i; \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{且}\varepsilon_{ij}\text{相互独立}. \end{cases}$$

为了找因子各水平对试验指标的影响，将μ_i分解。
令$\mu = \frac{1}{n}\sum_{i=1}^{a} n_i\mu_i$，$n = \sum_{i=1}^{a} n_i$，$\alpha_i = \mu_i - \mu, i = 1, \cdots, a$
其中μ为所有Y_{ij}的总的平均值，α_i为第i个水平对试验指标的效应，简称为水平A_i的效应，它反映了因子的第i个水平A_i对试验指标作用的大小。可以验证：$\sum_{i=1}^{a} n_i\alpha_i = \sum_{i=1}^{a} n_i(\mu_i - \mu) = 0$.
于是有

$$\begin{cases} Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, \cdots, a, j = 1, \cdots, n_i; \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{且}\varepsilon_{ij}\text{相互独立}; \\ \sum_{i=1}^{a} n_i\alpha_i = 0.\text{称为单因子方差分析模型}. \end{cases}$$

第五节 单因子试验方差分析

三.统计分析

判定因子A的a个水平下均值是否相等，归结为检验假设

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$$

或

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a$$

是否成立

1.显著性检验

记$\overline{Y} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}$,表示所有$Y_{ij}$的总平均值。

$\overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$,表示第$i$个水平下的样本均值。

考虑统计量$S_T = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$ 称为总偏差平方和，反映全部试验数据之间的差异(离散程度)。

第五节 单因子试验方差分析

将S_T分解:

$$\begin{aligned} S_T &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}[(Y_{ij}-\overline{Y}_{i\cdot})+(\overline{Y}_{i\cdot}-\overline{Y})]^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\cdot})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\overline{Y}_{i\cdot}-\overline{Y})^2 \\ &= S_E + S_A \end{aligned}$$

其中$S_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\cdot})^2$反映了在相同条件下各次试验的差异,称为误差平方和或组内平方和; $S_A = \sum_{i=1}^{a} n_i(\overline{Y}_{i\cdot}-\overline{Y})^2$反映了来自不同总体的样本之间的差异,也就是反映了因子各水平效应α_i的影响,称为组间平方和, S_A也与试验误差有关.

第五节　单因子试验方差分析

(1)若H_0成立，则$\alpha_i = 0, i = 1, \cdots, a$,模型变为

$$\begin{cases} Y_{ij} = \mu + \varepsilon_{ij}, i = 1, \cdots, a, j = 1, \cdots, n_i; \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{且}\varepsilon_{ij}\text{相互独立} \end{cases}$$

这时S_T表示仅由随机误差ε_{ij}所引起的偏差。

(2)若H_0不成立，则S_T中除包含由随机误差ε_{ij}所引起的偏差外，还应包含由α_i不全为0所引起的偏差。

若能把S_T中由ε_{ij}所引起的偏差和因子α_i不全为0所引起的偏差分开，并选取适当的统计量作为衡量它们之间差异的度量尺度，就可以检验假设H_0.

第五节 单因子试验方差分析

可以推出

$$\frac{S_A}{\sigma^2} \overset{H_0}{\sim} \chi^2(a-1), \frac{S_E}{\sigma^2} \sim \chi^2(n-a)$$

且S_A和S_E相互独立。
选取统计量

$$F = \frac{S_A/(a-1)}{S_E/(n-a)} \overset{H_0}{\sim} F(a-1, n-a)$$

拒绝域为

$$W = \{f > F_{1-\alpha}(a-1, n-a)\}$$

若拒绝H_0,则认为因子A的a个水平效应之间有显著性差异；否则认为因子A的a个水平效应之间没有显著性差异

第五节　单因子试验方差分析

方差来源	平方和S	自由度f	均方$\overline{S}$	F值	显著性
因子A	$S_A = \sum_{i=1}^{a} n_i(\overline{Y}_{i\cdot} - \overline{Y})^2$	$a-1$	$\overline{S}_A = \frac{S_A}{a-1}$	$F = \frac{\overline{S}_A}{\overline{S}_E}$	
误差	$S_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2$	$n-a$	$\overline{S}_E = \frac{S_E}{n-a}$		
总和	$S_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$	$n-1$			

第五节 单因子试验方差分析

例2： 研究5个品种产量之间是否有显著性差异($\alpha = 0.1$)

A	1	2	3	$\bar{y}_{i.}$
A_1	41	39	40	40
A_2	33	37	35	35
A_3	38	35	35	36
A_4	37	39	38	38
A_5	31	34	34	33

第五节 单因子试验方差分析

解： $H_0 : \mu_1 = \cdots = \mu_5, F = \frac{S_A/(5-1)}{S_E/(15-5)} \sim F(4,10)$

$S_A = \sum_{i=1}^{a} \frac{1}{n_i}(\sum_{j=1}^{n_i} Y_{ij})^2 - \frac{1}{n}(\sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij})^2 = 19962 - 19874.4 = 87.6$

$S_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij})^2 - \sum_{i=1}^{a} \frac{1}{n_i}(\sum_{j=1}^{n_i} Y_{ij})^2 = 19986 - 19962 = 24$

$S_T = S_A + S_E = 111.6$

方差来源	平方和S	自由度f	均方$\overline{S}$	F值	显著性
因子A	$S_A = 87.6$	4	21.9	9.13	
误差	$S_E = 24$	10	2.4		
总和	$S_T = 111.6$	14			

对于$\alpha = 0.1$,查表$F_{0.9}(4,10) = 2.641 < f$,所以拒绝$H_0$,说明5个品种有显著差异。

第五节 单因子试验方差分析

2.参数估计

a个水平效应之间有显著差异，也就是说$\mu_1, \cdots, \mu_a$不完全相同，还需对每一对μ_i, μ_j之间的差异程度作出估计，也就是对效应之差$\mu_i - \mu_j$进行区间估计。

(1)$E(\overline{Y}_{i\cdot}) = \mu + \alpha_i, \quad i = 1, 2, \cdots, a, \quad E(\overline{Y}) = \mu.$ 所以$\hat{\mu} = \bar{Y}, \hat{\alpha}_i = \overline{Y}_{i\cdot} - \overline{Y}$分别是$\mu$和$\alpha_i$的无偏估计.

(2) $Y_i \sim N(\mu_i, \sigma^2), Y_j \sim N(\mu_j, \sigma^2)$ $(i \neq j)$,求均值差$\mu_i - \mu_j = \alpha_i - \alpha_j$的区间估计.

第五节　单因子试验方差分析

$\overline{Y}_{i\cdot} \sim N(\mu_i, \frac{\sigma^2}{n_i}), i \neq j$时，$\overline{Y}_{i\cdot}$ 与$\overline{Y}_{j\cdot}$相互独立。
所以

$$\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot} \sim N(\mu_i - \mu_j, (\frac{1}{n_i} + \frac{1}{n_j})\sigma^2)$$

又因为$S_E/\sigma^2 \sim \chi^2(n-a)$,且$\hat{\sigma}^2 = \frac{S_E}{n-a}$

$$\frac{(\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}) - (\alpha_i - \alpha_j)}{\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\hat{\sigma}} \sim t(n-a)$$

于是均值差$\mu_i - \mu_j = \alpha_i - \alpha_j$的置信水平为$(1-\alpha)$置信区间为

$$[\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot} - \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\hat{\sigma} t_{1-\frac{\alpha}{2}}(n-a), \overline{Y}_{i\cdot} - \overline{Y}_{j\cdot} + \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\hat{\sigma} t_{1-\frac{\alpha}{2}}(n-a)]$$

第五节　单因子试验方差分析

注：

(1)若置信区间包含0，则以$(1-\alpha)$概率认为μ_i与μ_j没有显著差异；

(2)若置信区间上限小于0，则以$(1-\alpha)$概率认为$\mu_i < \mu_j$

(3)若置信区间下限大于0，则以$(1-\alpha)$概率认为$\mu_i > \mu_j$

第五节　单因子试验方差分析

3.当观测值过大或过小，可以经过线性变换使计算简单，令$Y'_{ij}=\frac{Y_{ij}-c}{b}, b\neq 0$, 用$Y'_{ij}$与$Y_{ij}$计算的F值是相同的。