6.2　多元线性回归分析

一、多元线性回归模型

二、参数的估计

三、估计量的分布及性质

四、回归系数与回归方程的显著性检验

五、多元线性回归模型的预测

六、逐步回归

七、稳健回归

一、多元线性回归的数学模型

实际问题中的随机变量Y通常与多个普通变量$x_1, x_2, \cdots, x_m (m > 1)$有关.

对于自变量$x_1, x_2, \cdots, x_m$的一组确定值,Y具有一定的分布,若Y的数学期望存在,则它是$x_1, x_2, \cdots, x_m$的函数.

$$\mu_{Y|x_1, x_2, \cdots, x_m} = \mu(x_1, x_2, \cdots, x_m)$$

Y关于x的回归函数

若$\mu(x_1, x_2, \cdots, x_m)$是$x_1, x_2, \cdots, x_m$的线性函数, 即

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

$\beta_0, \beta_1, \cdots, \beta_m, \sigma^2$是与$x_1, \cdots, x_m$无关的未知参数.

称其为多元线性回归模型

其中$x_1, x_2, \cdots, x_m$称为回归变量,

$\beta_0, \beta_1, \cdots, \beta_m$称为回归系数,

因为 $EY = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$,则称

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

为Y关于$x_1, x_2, \cdots, x_m$的线性回归方程

如果$(x_{i1}, x_{i2}, \cdots, x_{im}, Y_i)(i=1,2,\cdots,n)$是$(x_1, x_2, \cdots, x_m, Y)$的$n$个观测值，

同时它们满足关系

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i,$$

$\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \cdots, n$, ε_i相互独立.

显然Y_i相互独立，且服从正态分布，即

$$Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}, \sigma^2), \quad i = 1, 2, \cdots, n$$

注：由于Y_i的期望不同，所以Y_i相互独立非同分布。

为了表述方便，引入矩阵

$$X=\begin{pmatrix}1 & x_{11} & x_{12} & \cdots & x_{1m}\\ 1 & x_{21} & x_{22} & \cdots & x_{2m}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm}\end{pmatrix}, Y=\begin{pmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{pmatrix}, \quad \beta=\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_m\end{pmatrix}, \quad \varepsilon=\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}.$$

则模型 $Y_i=\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}+\varepsilon_i$， $\varepsilon_i \sim N(0,\sigma^2)$，

$i=1,2,\cdots,n$， ε_i相互独立，可用矩阵表示为

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

同时 $EY=X\beta$， $\mathrm{Cov}(Y,Y)=\sigma^2 I_n$

一般假定$n>m$和矩阵X的秩等于$m+1$.

二、参数的估计

1. 参数向量β的最小二乘估计

$$\sum_{i=1}^{n}(Y_i-\sum_{j=0}^{m}\hat{\beta}_j x_{ij})^2=\min_{\beta}\sum_{i=1}^{n}(Y_i-\sum_{j=0}^{m}\beta_j x_{ij})^2$$

上式可以用矩阵表示为

$$\|Y-X\hat{\beta}\|^2=\min_{\beta}\|Y-X\beta\|^2$$

利用微分法求解$\hat{\beta}$,即

$$\sum_{i=1}^{n}(Y_i-\sum_{j=0}^{m}\hat{\beta}_j x_{ij})=0,\quad k=0$$ (对β_0求导的结果)

$$\sum_{i=1}^{n}(Y_i-\sum_{j=0}^{m}\hat{\beta}_j x_{ij})x_{ik}=0,\quad k=1,\cdots,m$$ (对β_j(j=1,⋯, m)求导的结果)

将$\sum_{i=1}^{n}(Y_i-\sum_{j=0}^{m}\hat{\beta}_j x_{ij})x_{ik}=0,\quad k=0,1,\cdots,m(k=0$时，$x_{i0}=1)$改写

$$\sum_{i=1}^{n}Y_i x_{ik}=\sum_{i=1}^{n}\sum_{j=0}^{m}\hat{\beta}_j x_{ij}x_{ik}=\sum_{j=0}^{m}(\sum_{i=1}^{n}x_{ij}x_{ik})\hat{\beta}_j,\quad k=0,1,\cdots,m$$

此式可以用矩阵改写为

$$X^T Y=(X^T X)\hat{\beta}$$

此方程称为正规方程。由于假设X的秩为$m+1$,所以$(X^T X)_{(m+1)\times(m+1)}$是正定矩阵，因而存在逆矩阵$(X^T X)^{-1}$,

于是 $\hat{\beta}=(X^T X)^{-1}X^T Y$

对于 $\hat{\beta}=(X^T X)^{-1}X^TY$

经常需要计算 $(X^T X)^{-1}$.

简单的2阶矩阵的逆矩阵的计算口诀:

主对调，次换号，除以行列式

推导：假设 $A=\begin{bmatrix} a & b \\ c & d \end{bmatrix}$，$a,b,c,d \in R$，且 A 可逆，那么根据知识储备 1.2 $A^*=\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

所以呢，$A^{-1}=\dfrac{A^*}{|A|}=\dfrac{\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}}{|A|}$

将$\hat{\beta}$代入回归方程，可得

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m$$

此方程也称为线性回归方程，此方程可以对Y预测.

2. 未知参数σ^2的估计

由6.1节可知，一元情形，σ^2的无偏估计为

$$\hat{\sigma}^{*2}=\frac{1}{n-2}\sum_{i=1}^{n}[Y_i-(\hat{\alpha}+\hat{\beta}x_i)]^2=\frac{1}{n-2}Q_e.$$

类似的可以得到多元情形时，σ^2的无偏估计为

$$\hat{\sigma}^{*2}=\frac{1}{n-\underline{(m+1)}}\sum_{i=1}^{n}[Y_i-\sum_{j=0}^{m}\hat{\beta}_j x_{ij}]^2$$

后面会证明其为无偏估计

未知参数的个数！！！

其矩阵形式为：

$$\hat{\beta}=(X^T X)^{-1} X^T Y$$

$$\hat{\sigma}^{*2}=\frac{1}{n-m-1}(Y-X\hat{\beta})^T(Y-X\hat{\beta})$$

$$=\frac{1}{n-m-1}(Y-X(X^T X)^{-1}X^T Y)^T(Y-X(X^T X)^{-1}X^T Y)$$

$$=\frac{1}{n-m-1}Y^T\boxed{(I_n-X(X^T X)^{-1}X^T)^T(I_n-X(X^T X)^{-1}X^T)}Y$$

$$=\frac{1}{n-m-1}Y^T[I_n-X(X^T X)^{-1}X^T-\cancel{X(X^T X)^{-1}X^T}$$

$$+X(X^T X)^{-1}\underbrace{X^T X(X^T X)^{-1}}_{I_{m+1}}X^T]Y$$

$$\cancel{X(X^T X)^{-1}X^T}$$

$$=\frac{1}{n-m-1}Y^T\boxed{[I_n-X(X^T X)^{-1}X^T]}Y$$

$$\hat{\sigma}^{*2}=\frac{1}{n-m-1}Y^T[I_n-X(X^TX)^{-1}X^T]Y$$

$$=\frac{1}{n-m-1}[Y^TY-Y^TX\underline{(X^TX)^{-1}X^TY}]$$

$$\hat{\beta}=(X^TX)^{-1}X^TY$$

$$=\frac{1}{n-m-1}[Y^TY-\hat{\beta}^TX^TY]$$

$$=\frac{1}{n-m-1}[Y^T-\hat{\beta}^TX^T]Y$$

无偏估计

例1 (p201例6. 5) 某种水泥在凝固时放出的热量Y与水泥中下列4种化学成份有关：

$(1)x_1: 3CaO \times Al_2O_3;$ $(2)x_2: 3CaO \cdot SiO_2$

$(3)x_3: 4CaO \cdot Al_2O_3 \cdot Fe_2O_3;$ $(4)x_4: 2CaO \cdot SiO_2$

通过实验得到下列数据：

序 号	x_1%	x_2%	x_3%	x_4%	Y_i
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.6
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

试求Y关于x_1，x_2，x_3，x_4的回归函数。

解 由于$\hat{\beta}=(X^TX)^{-1}X^TY$，代入相关数据，得

$$(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_4)^T=(62.45,1.55,0.51,0.10,-0.144)^T$$

将$\hat{\beta}$代入回归方程，

$$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_4x_4$$

可以得到回归方程为

$$\hat{Y}=62.45+1.55x_1+0.51x_2+0.10x_3-0.144x_4$$

三、估计量的分布及性质

1. $\hat{\beta}=(X^T X)^{-1}X^T Y$ 的分布

$\hat{\beta}$的每一个分量都是$Y_1,Y_2,\cdots,Y_n$的线性组合，因而由多元分布理论可知，随机向量$\hat{\beta}$服从$m+1$维正态分布，其期望向量为

$$E\hat{\beta}=(X^T X)^{-1}X^T EY=(X^T X)^{-1}X^T X\beta=\beta$$

因而,$\hat{\beta}$是β的无偏估计

$$\mathrm{cov}(\hat{\beta},\hat{\beta}) = (X^T X)^{-1} X^T \mathrm{cov}(Y,Y) X (X^T X)^{-1}$$

$$= (X^T X)^{-1} X^T (\sigma^2 I_n) X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1}$$

若令$C = \sigma^2 (X^T X)^{-1}$,则$\hat{\beta}$服从$m+1$维正态分布，其密度函数为

$$f(x) = (2\pi)^{-\frac{m+1}{2}} |C|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(X-\beta)^T C^{-1}(X-\beta)\},$$

其中 $x \in R^{m+1}$.

记为 $\hat{\beta} \sim N_{m+1}(\beta, \sigma^2 (X^T X)^{-1})$.

$\hat{\beta}=(X^TX)^{-1}X^TY$ 的性质

性质1 $\hat{\beta}$是Y的线性函数，服从$m+1$维正态分布，均值为β，协方差矩阵为$\sigma^2(X^TX)^{-1}$.

注： 若估计量为Y的线性函数，则称其为线性估计，由性质1知$\hat{\beta}$是β的线性无偏估计。若T是β的另一估计，且$\mathrm{Cov}(T,T)-\mathrm{Cov}(\hat{\beta},\hat{\beta})$为非负定矩阵，则称$\hat{\beta}$的方差不大于$T$的方差。

性质2 $\hat{\beta}$是β的最小方差线性无偏估计.

证 设T是β的任一线性无偏估计，则T必可表为

$$T = AY$$

并且$ET = E(AY) = AEY = AX\beta = \beta$.由$\beta$的任意性，则

$$AX = I_{m+1}$$

由于 $$\mathrm{Cov}(T,T) = A\mathrm{Cov}(Y,Y)A^T = \sigma^2(AA^T)$$

又考虑到

$$\boxed{AX = I_{m+1}}$$

$$[A-(X^TX)^{-1}X^T][A-(X^TX)^{-1}X^T]^T$$

$$=(AA^T)+(X^TX)^{-1}-(X^TX)^{-1}\underline{X^TA^T}-\underline{AX}(X^TX)^{-1}$$

$$=(AA^T)+(X^TX)^{-1}-(X^TX)^{-1}-(X^TX)^{-1}$$

$$=(AA^T)-(X^TX)^{-1}$$ ≥ 0 非负定

则由T及A的任意性,

可知, $\hat{\beta}$是β的最小方差线性无偏估计.

残差向量

称$\tilde{Y}=Y-X\hat{\beta}$为残差向量.

则有$\tilde{Y}=[I_n-X(X^TX)^{-1}X^T]Y$, 或有$\tilde{Y}_i=Y_i-\sum_{j=0}^{m}\hat{\beta}_j x_{ij}$.

性质3 $E\tilde{Y}=0$, $\mathrm{Cov}(\tilde{Y},\tilde{Y})=\sigma^2[I_n-X(X^TX)^{-1}X^T]$

证 $E\tilde{Y}=E(Y-X\hat{\beta})=X\beta-X\beta=0$

$$
\begin{aligned}
&\mathrm{Cov}(\tilde{Y},\tilde{Y})\\
&=[I_n-X(X^TX)^{-1}X^T]\mathrm{cov}(Y,Y)[I_n-X(X^TX)^{-1}X^T]^T\\
&=\sigma^2[I_n-X(X^TX)^{-1}X^T][I_n-X(X^TX)^{-1}X^T]^T\\
&=\sigma^2[I_n-X(X^TX)^{-1}X^T]
\end{aligned}
$$

$$\tilde{Y} = Y - X\hat{\beta} = [I_n - X(X^T X)^{-1} X^T]Y$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

性质4 $\tilde{Y}$与$\hat{\beta}$互不相关

证 $\mathrm{Cov}(\tilde{Y}, \hat{\beta})$

$$= [I_n - X(X^T X)^{-1} X^T]\mathrm{Cov}(Y, Y)[(X^T X)^{-1} X^T]^T$$

$$= \sigma^2 [I_n - X(X^T X)^{-1} X^T][(X^T X)^{-1} X^T]^T = 0$$

因而$\tilde{Y}$与$\hat{\beta}$互不相关.

后面定理6.2会进一步证明两者相互独立

由残差向量$\tilde{Y}$

可构造残差平方和.

$$\tilde{Y} = Y - X\hat{\beta},$$

$$= [I_n - X(X^T X)^{-1} X^T]Y,$$

设$Q = \tilde{Y}^T \tilde{Y} = \|\tilde{Y}\|^2$,称其为残差平方和，则

$$Q = Y^T [I_n - X(X^T X)^{-1} X^T]^T [I_n - X(X^T X)^{-1} X^T]Y$$

$$= Y^T [I_n - X(X^T X)^{-1} X^T]Y$$

回想$\hat{\sigma}^{*2} = \dfrac{1}{n-m-1} Y^T [I_n - X(X^T X)^{-1} X^T]Y$

$$\Longrightarrow \quad \hat{\sigma}^{*2} = \frac{1}{(n-m-1)} Q$$

验证$\hat{\sigma}^{*2}=\dfrac{1}{(n-m-1)}Q$是$\sigma^2$的无偏估计

由性质3已知

$$E\tilde{Y}=0,\ \mathrm{Cov}(\tilde{Y},\tilde{Y})=\sigma^2[I_n-X(X^TX)^{-1}X^T]$$

$$EQ=E\tilde{Y}^T\tilde{Y}=\sum_{i=1}^{n}E\tilde{Y}_i^2=\sum_{i=1}^{n}D\tilde{Y}_i$$

$$=\mathrm{tr}\{\mathrm{cov}(\tilde{Y},\tilde{Y})\}=\sigma^2\mathrm{tr}[I_n-X(X^TX)^{-1}X^T]$$

$$=\sigma^2[n-\mathrm{tr}I_{m+1}]=\sigma^2[n-m-1]$$

其中$\mathrm{tr}A=\sum_{i=1}^{n}a_{ii}$表示$n\times n$矩阵$A$的迹.

因此，$E\hat{\sigma}^{*2}=\dfrac{1}{(n-m-1)}EQ=\sigma^2$

无偏估计

$\hat{\sigma}^{*2} = \dfrac{1}{(n-m-1)}Q$的分布

定理6. 2 若$(x_{i1}, x_{i2}, \cdots, x_{im}, Y_i)(i=1,2,\cdots,n)$满足多元线性回归模型，则

(1)$\hat{\beta}$与$\tilde{Y}$相互独立，且服从正态分布；

(2)$\hat{\beta}$与$\hat{\sigma}^{*2}$相互独立；

(3) $(n-m-1)\hat{\sigma}^{*2}/\sigma^2 \sim \chi^2(n-m-1)$

n-(m+1)

m+1 未知参数个数

证 (1) 由于$(\hat{\beta},\tilde{Y})^T=\begin{pmatrix}\hat{\beta}\\ \tilde{Y}\end{pmatrix}=\begin{pmatrix}X(X^TX)^{-1}X^TY\\ [I_n-X(X^TX)^{-1}X^T]Y\end{pmatrix}$

$$=\begin{pmatrix}X(X^TX)^{-1}X^T\\ I_n-X(X^TX)^{-1}X^T\end{pmatrix}Y$$

为相互独立的且服从正态分布的

$Y_1,Y_2,\cdots,Y_n$的线性组合，因而由多元正态分布理论

可知，$(\hat{\beta},\tilde{Y})^T$服从多元正态分布，由性质3可知，$\hat{\beta}$与

$\tilde{Y}$不相关，因而二者独立.

(2) 由于$\hat{\sigma}^{*2}=\dfrac{1}{(n-m-1)}Q=\dfrac{\tilde{Y}^T\tilde{Y}}{(n-m-1)}$,

结合（1）可知，$\hat{\beta}$与$\hat{\sigma}^{*2}$相互独立。

(3) 设$B=X(X^TX)^{-1}X^T$,由于B是$n\times n$非负定矩阵，秩为

$m+1$,则存在n阶正交矩阵D使得

$$DBD^T=\begin{bmatrix}\lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_{\mathrm{m+1}} & & & \\ & & & \mathbf{0} & & \\ & & & & \ddots & \\ & & & & & \mathbf{0}\end{bmatrix}$$

其中$D^T D = I_n, \lambda_i > 0, i = 1, 2, \cdots, m+1$

同时

$$B^2 = BB^T = X(X^T X)^{-1} X^T [X(X^T X)^{-1} X^T]^T = B$$

因而 $$DB^2 D^T = DBD^T$$

即 $$\lambda_i = \lambda_i^2 \Rightarrow \lambda_i = 1,\ i = 1, 2, \cdots, m+1$$

则 $$DBD^T = \begin{pmatrix} I_{m+1} & 0 \\ 0 & 0 \end{pmatrix}$$

令 $Z = D(Y - X\beta) = (Z_1, Z_2, \cdots, Z_n)^T$，其中$D$为正交矩阵

$$EZ = D(EY - EX\beta) = 0$$

$$\mathrm{cov}(Z, Z) = D\,\mathrm{cov}(Y - X\beta, Y - X\beta)D^T$$

$$= D\,\mathrm{cov}(\varepsilon, \varepsilon)D^T = D\sigma^2 I_n D^T = \sigma^2 I_n$$

$$= \mathrm{cov}(\varepsilon, \varepsilon) \quad = \mathrm{cov}(Y - X\beta, Y - X\beta)$$

又因为Z为正态随机向量，上式可以表明

$Z_1, Z_2, \cdots, Z_n$相互独立，同服从于$N(0, \sigma^2)$分布.

又

$$Q=\tilde{Y}^T\tilde{Y}=(Y-X\hat{\beta})^T(Y-X\hat{\beta})$$

$$=Y^T(I_n-X(X^TX)^{-1}X^T)^T(I_n-X(X^TX)^{-1}X^T)Y$$

$$=Y^T(I_n-X(X^TX)^{-1}X^T)Y$$

可以容易地反推上式

$$=(Y-X\beta)^T(I_n-X(X^TX)^{-1}X^T)(Y-X\beta)$$

$$=Z^TD(I_n-X(X^TX)^{-1}X^T)D^TZ$$

$$=Z^TDD^TZ-Z^TDX(X^TX)^{-1}X^TD^TZ$$

$$=Z^TDD^TZ-Z^TDBD^TZ$$

$$=Z_1^2+\cdots+Z_n^2-(Z_1^2+\cdots+Z_{m+1}^2)$$

$$=Z_{m+2}^2+\cdots+Z_n^2$$

Q由$n-m-1$个独立的正态变量的平方和组成。

故$Q/\sigma^2 \sim \chi^2(n-m-1)$,即

$$\frac{(n-m-1)\hat{\sigma}^{*2}}{\sigma^2} \sim \chi^2(n-m-1)$$

得证。

又 $X\hat{\beta}-X\beta=X(X^TX)^{-1}X^TY-X\beta$

$$=X(X^TX)^{-1}X^T(Y-X\beta)=X(X^TX)^{-1}X^TD^TZ$$

则 $\|X\hat{\boldsymbol{\beta}}-X\boldsymbol{\beta}\|^2$

$$=(X\hat{\beta}-X\beta)^T(X\hat{\beta}-X\beta)$$

$$=Z^TDX(X^TX)^{-1}X^TX(X^TX)^{-1}X^TD^TZ$$

$$=Z^TDX(X^TX)^{-1}X^TD^TZ$$

$$=Z^TDB\ D^TZ=Z^T\begin{pmatrix} I_{m+1} & 0 \\ 0 & 0 \end{pmatrix}Z$$

$$=Z_1^2+Z_2^2+\cdots+Z_{m+1}^2$$

于是

$$\|Y - X\beta\|^2 = (Y - X\beta)^T (Y - X\beta) = Z^T D D^T Z$$

$$= Z^T Z = \sum_{i=1}^{n} Z_i^2 = Q + \|X\hat{\beta} - X\beta\|^2$$

推论1 $\dfrac{Q}{\sigma^2}$与$\dfrac{\|X\hat{\beta} - X\beta\|^2}{\sigma^2}$相互独立,

且$\dfrac{\|X\hat{\beta} - X\beta\|^2}{\sigma^2} \sim \chi^2(m+1)$.

四、回归系数及回归方程的显著性检验

1. 回归系数的显著性检验

假设检验$H_0:\beta_j=0 \leftrightarrow H_1:\beta_j \neq 0 (j=1,\cdots,m)$

构造检验统计量　　由于$\mathrm{cov}(\hat{\beta},\hat{\beta})=\sigma^2(X^TX)^{-1}$

设C_{jj}是矩阵$C=(X^TX)^{-1}$的主对角线上第$j+1$个元素

则　$D\hat{\beta}_j=C_{jj}\sigma^2$，因而

$$\frac{\hat{\beta}_j-\beta_j}{\sqrt{C_{jj}\sigma^2}} \sim N(0,1)$$

又因为

$\frac{Q}{\sigma^2} \sim \chi^2(n-m-1)$,且$Q$与$\hat{\beta}_j$相互独立,在假设成立时,

$$T_j = \frac{\hat{\beta}_j}{\sqrt{C_{jj}Q/(n-m-1)}} \sim t(n-m-1)$$

对于给定的显著性水平α，拒绝域为：

$$W = \{T_j \mid \quad |T_j| \geq t_{\frac{\alpha}{2}}(n-m-1)\}$$

当T_j属于拒绝域时，拒绝原假设，即系数β_j显著不为0.

2. 回归方程的显著性检验

假设检验

$$H_0: \beta_1 = \cdots = \beta_m = 0 \leftrightarrow H_1: \text{至少}\exists\text{一个}\beta_j \neq 0$$

构造检验统计量

$$\underline{Q_T} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2$$

总离差平方和

$$= \underline{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} + \underline{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} = Q + Q_B$$

平方和分解

残差平方和

回归平方和

混合项为

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})=\sum_{i=1}^{n}\tilde{Y}_i(\hat{Y}_i-\bar{Y})=\sum_{i=1}^{n}\tilde{Y}_i\hat{Y}_i-\sum_{i=1}^{n}\tilde{Y}_i\bar{Y}$$

由P206公式

当$k=1,2,...,m$时，有$\sum_{i=1}^{n}(Y_i-\sum_{j=0}^{m}\hat{\beta}_j x_{ij})x_{ik}=\sum_{i=1}^{n}\tilde{Y}_i x_{ik}=0.$

及 $Y_i=\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}+\varepsilon_i$, $\hat{Y}_i=\sum_{k=0}^{m}\hat{\beta}_k x_{ik}$

$\varepsilon_i \sim N(0,\sigma^2)$, $i=1,2,\cdots,n$, ε_i相互独立.

可知 $\sum_{i=1}^{n}\tilde{Y}_i\hat{Y}_i=\sum_{i=1}^{n}\tilde{Y}_i\sum_{k=0}^{m}\hat{\beta}_k x_{ik}=\sum_{k=0}^{m}\hat{\beta}_k\sum_{i=1}^{n}\tilde{Y}_i x_{ik}=0,$

$$\sum_{i=1}^{n} \tilde{Y}_i \bar{Y} = \bar{Y} \sum_{i=1}^{n} \tilde{Y}_i = \bar{Y} \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{m} \hat{\beta}_j x_{ij})$$

由P206公式　当$k = 0$时，有$\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{m} \hat{\beta}_j x_{ij}) = 0,$

$$so \quad \sum_{i=1}^{n} \tilde{Y}_i \bar{Y} = \bar{Y} \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{m} \hat{\beta}_j x_{ij}) = 0$$

所以

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n} (\tilde{Y}_i)(\hat{Y}_i - \bar{Y})$$

$$= \sum_{i=1}^{n} \tilde{Y}_i \hat{Y}_i - \sum_{i=1}^{n} \tilde{Y}_i \bar{Y} = 0$$

在H_0成立的条件下，可以证明

$$Q/\sigma^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2/\sigma^2 \sim \chi^2(n-m-1),$$

$$Q_B/\sigma^2=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2/\sigma^2 \sim \chi^2(m)$$

$$F=\frac{Q_B/(\sigma^2 m)}{Q/[\sigma^2(n-m-1)]}=\frac{Q_B(n-m-1)}{Qm} \sim F(m,n-m-1)$$

对于给定的显著性水平α，拒绝域为：

$$W=\{F \mid \quad F \geq F_\alpha(m,n-m-1)\}$$

当F属于拒绝域时，拒绝原假设，即所有系数β_j显著不全为0.

例2（续例1）（p216例6.6）

当$\alpha=0.05$时，试检验线性回归方程的显著性.

解 有给定的数据可以计算得到

$$Q_T=\sum_{i=1}^{13}(y_i-\hat{y}_i)^2=2715.763$$

$$Q_A=\sum_{i=1}^{13}(\hat{y}_i-\bar{y})^2=47.863$$

$$Q_B=Q_T-Q_A=2667.9$$

$$F=\frac{8Q_B}{4Q_A}=111.4795>F_{0.05}(4,8)=3.84$$

因此拒绝原假设，认为线性回归方程是显著的.

例3（续例1）（p212例6.7）

当$\alpha=0.05$以及$\alpha=0.1$时，试检验例6.5线性回归方程中回归系数是否显著为0.

解 有给定的数据可以计算得到

$$Q_A=\sum_{i=1}^{13}(y_i-\hat{y}_i)^2=47.863$$

$$\hat{\beta}_1=1.5511,\hat{\beta}_2=0.5101,\hat{\beta}_3=0.1019,\hat{\beta}_4=-0.1441$$

$$\hat{\sigma}^*=\sqrt{\frac{47.863}{13-4-1}}=2.446$$

$$t_1 = \frac{\hat{\beta}_1}{\hat{\sigma}^* \sqrt{C_{11}}} = 2.0817, \quad t_2 = \frac{\hat{\beta}_2}{\hat{\sigma}^* \sqrt{C_{11}}} = 0.7046,$$

$$t_3 = \frac{\hat{\beta}_3}{\hat{\sigma}^* \sqrt{C_{11}}} = 0.1350, \quad t_4 = \frac{\hat{\beta}_4}{\hat{\sigma}^* \sqrt{C_{11}}} = -0.2032$$

$$t_{0.025}(8) = 2.306, \qquad t_{0.05}(8) = 1.860$$

因此　　α=0. 05时，4个系数均显著为0.

　　　　α=0. 1时，只有$\hat{\beta}_1$是显著不为0,其他回归系数显著为0.

五、多元线性回归模型的预测

为了利用回归方程进行预测，在给出$x_1, x_2, \cdots, x_m$的一组观察值$x_{01}, x_{02}, \cdots, x_{0m}$时，

若记$x_0 = (1, x_{01}, x_{02}, \cdots, x_{0m})^{\mathrm{T}}$，可得

$$y_0 = x_0^{\mathrm{T}}\beta + \varepsilon_0, E(\varepsilon_0) = 0, \quad D(\varepsilon_0) = \sigma^2$$

以及y_0的预测值$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_m x_{0m} = x_0^{\mathrm{T}}\hat{\beta}$

$\hat{y}_0$具有如下性质：

(1)$\hat{y}_0$是y_0的无偏预测，即$E(\hat{y}_0)=E(y_0)$;

(2)在y_0的一切线性无偏预测中，$\hat{y}_0$的方差最小；

(3)如果$\varepsilon_0 \sim N(0,\sigma^2 I_n)$,则$\hat{y}_0 - y_0 \sim N(0,\sigma^2(1+x_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_0))$,且$\hat{y}_0 - y_0$与$\hat{\sigma}^{*2}$相互独立，其中$\hat{\sigma}^{*2}=Q/(n-m-1)$,$Q$为残差平方和。

(4)如果$\varepsilon_0 \sim N(0,\sigma^2 I_n)$,则

$$\frac{\hat{y}_0 - y_0}{\hat{\sigma}\sqrt{1+x_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_0}} \sim t(n-m-1)$$

(5)如果$\varepsilon_0 \sim N(0,\sigma^2 I_n)$,则$y_0$的置信度为$1-\alpha$的置信区间为

$$(\hat{y}_0 - t_{1-\alpha/2}(n-m-1)\hat{\sigma}\sqrt{1+x_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_0},$$

$$\hat{y}_0 + t_{1-\alpha/2}(n-m-1)\hat{\sigma}\sqrt{1+x_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_0})$$

例6.8 某商店将其连续18个月的库存占用资金情况、广告投入费用、员工薪酬以及销售额等方面的数据作了一个汇总，见表6.4(教材P213)。该商店的管理人员试图根据这些数据找到销售额与其他三个变量之间的关系，以便进行销售额预测并为未来的预算人员提供参考。试根据这些数据建立回归模型。如果未来某月库存资金额为150万元，广告投入预算为45万元，员工薪酬总额为27万元，试根据建立的回归模型预测该月的销售额。

解 建立y(销售额)关于x_1(库存资金额)、x_2(广告投入)和x_3(员工薪酬总额)的多元线性回归方程，运用参数估计公式，我们可以求出参数估计。经计算，参数估计为$\hat{\beta}_0 = 162.0632, \hat{\beta}_1 = 7.2739, \hat{\beta}_2 = 13.9575, \hat{\beta}_3 = -4.3996$.于时可以得到相应的回归方程

$$y = 162.0632 + 7.2739x_1 + 13.9575x_2 - 4.3996x_3$$

进一步对回归方程作显著性检验。计算数据为

方差来源	平方和	自由度	均方	F值	显著性
回归	3177186	3	1059062	105.0867	α=0.01
剩余	141091.8	14	10077.99		
总和	3318277	17			

查表得$F_{0.01}(3,14)=5.56$。由于F值$105.0867>F_{0.01}(3,14)=5.56$,这说明在$\alpha=0.01$的水平下，以上回归方程是显著的。

如果未来某月库存资金额为150万元，广告投入预算为45万元，员工薪酬总额为27万元，可以计算得出$y=1762.4465$(万元)，也即是说，这时利用回归模型预测该月的销售额为1762.4465万元。

Thank You!