§7.2　多元线性回归

前面我们讨论了只包含一个自变量的一元线性回归，在现实世界中一个因变量往往受到多个因素的影响，也即存在着多个自变量，例如：商品的销售额要受到下列因素的影响：商品的价格、广告宣传费、个人可支配收入等因素.

本节将讨论包含多个自变量的回归问题，重点介绍多元线性回归模型及其基本假设，回归模型未知参数的估计及其性质，回归方程及回归系数的显著性检验等. 利用回归方程进行预测，多元线性回归比一元线性回归更加复杂，计算量也大大增加.

一　多元线性回归模型的一般形式

设因变量y是可以观测的随机变量, 自变量$x_1, x_2, \cdots, x_p$是可以精确测量或可控制的一般变量，因变量与自变量之间的相关关系表示为

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

其中未知参数$\beta_1, \beta_2, \cdots, \beta_p$称为回归系数，$\beta_0$称为常数项，$\varepsilon$是随机误差，一般假设

$$E(\varepsilon) = 0, \quad D(\varepsilon) = \sigma^2,$$

在对未知参数进行区间估计或假设检验时，还需要假设误差服从正态分布，即$\varepsilon \sim N(0, \sigma^2)$. 因而

$$y \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2).$$

对于一个实际问题，如果我们获得了n组独立观测数据$(y_i;x_{i1},x_{i2},\cdots,x_{ip})$，$i=1,2,\cdots,n$，则线性回归模型表示为

$$\begin{cases} y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}+\varepsilon_i, \quad i=1,2,\cdots,n, \\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n \text{相互独立, 且} \varepsilon_i \sim N(0,\sigma^2) \end{cases}$$

写成矩阵形式为

$$y=X\beta+\varepsilon$$

其中

$$y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

$$\boldsymbol{\beta}=\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_p\end{pmatrix}\qquad \boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}\qquad \boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2 I_n).$$

矩阵X是$n\times(p+1)$矩阵，称为回归设计矩阵或资料矩阵. 为了参数估计的需要, 我们假定$rank(X)=p+1$，即矩阵X的$(p+1)$个列向量线性无关. 这要求$n\geq p+1$. 即要求样本的个数不少于未知参数的个数.

注：$E(y)=\beta_0+\beta_1x_1+\cdots+\beta_px_p$, $\dfrac{\partial E(y)}{\partial x_i}=\beta_i$,

由此可得到回归系数β_i的意义.

二　多元线性回归模型的参数估计

一般采用最小二乘方法估计未知参数$\beta_0,\beta_1,\cdots,\beta_p$.

令 $$Q(\beta_0,\beta_1,\cdots,\beta_p)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_px_{ip})^2,$$

$\hat\beta_0,\hat\beta_1,\cdots,\hat\beta_p$应满足$Q(\hat\beta_0,\hat\beta_1,\cdots,\hat\beta_p)=\min Q(\beta_0,\beta_1,\cdots,\beta_p)$,

令 $$\begin{cases}\dfrac{\partial Q}{\partial\hat\beta_0}=-2\sum_{i=1}^{n}(y_i-\hat\beta_0-\hat\beta_1x_{i1}-\cdots-\hat\beta_px_{ip})=0\\ \dfrac{\partial Q}{\partial\hat\beta_1}=-2\sum_{i=1}^{n}(y_i-\hat\beta_0-\hat\beta_1x_{i1}-\cdots-\hat\beta_px_{ip})x_{i1}=0\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ \dfrac{\partial Q}{\partial\hat\beta_p}=-2\sum_{i=1}^{n}(y_i-\hat\beta_0-\hat\beta_1x_{i1}-\cdots-\hat\beta_px_{ip})x_{ip}=0\end{cases}$$

这组方程称为正规方程组，经整理得

$$\begin{cases} n\hat{\beta}_0+\sum_{i=1}^{n}x_{i1}\hat{\beta}_1+\cdots+\sum_{i=1}^{n}x_{ip}\hat{\beta}_p=\sum_{i=1}^{n}y_i \\ \sum_{i=1}^{n}x_{i1}\hat{\beta}_0+\sum_{i=1}^{n}x_{i1}^2\hat{\beta}_1+\cdots+\sum_{i=1}^{n}x_{i1}x_{ip}\hat{\beta}_p=\sum_{i=1}^{n}x_{i1}y_i \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \sum_{i=1}^{n}x_{ip}\hat{\beta}_0+\sum_{i=1}^{n}x_{ip}x_{i1}\hat{\beta}_1+\cdots+\sum_{i=1}^{n}x_{ip}^2\hat{\beta}_p=\sum_{i=1}^{n}x_{ip}y_i \end{cases}$$

上述正规方程组写成矩阵形式为

$$(X'X)\hat{\beta}=X'y$$

由于$rank(X)=p+1$，上述方程组有解

$$\hat{\beta}=(X'X)^{-1}X'y$$

这就是未知参数向量β的最小二乘估计.

$\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\beta}$称为回归向量，$\tilde{\boldsymbol{y}}=\boldsymbol{y}-\hat{\boldsymbol{y}}$称为残差向量.

定理5.4　对于多元线性回归模型，$\boldsymbol{y}=\boldsymbol{X}\beta+\varepsilon$,

若$\boldsymbol{E}(\varepsilon)=\mathbf{0}$，$\boldsymbol{D}(\varepsilon)=\sigma^2\boldsymbol{I}_n$．则 $\hat{\beta}$ 有下列性质

(1) $\boldsymbol{E}(\hat{\beta})=\beta$,　$\boldsymbol{D}(\hat{\beta})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$;

(2) 设$\tilde{\beta}$是β的任意线性无偏估计, 则$\boldsymbol{D}(\tilde{\beta})\geq\boldsymbol{D}(\hat{\beta})$;

(3) $\mathrm{Cov}(\hat{\beta},\tilde{\boldsymbol{y}})=\mathbf{0}$.

进一步, 若设 $\varepsilon\sim N(\mathbf{0},\sigma^2\boldsymbol{I}_n)$.　　则

(1)* $\hat{\beta}\sim N(\beta,\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$,

$\tilde{\boldsymbol{y}}\sim N(\mathbf{0},\sigma^2\left[\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right])$;

(2)* $\hat{\beta}$与$\tilde{\boldsymbol{y}}$相互独立;

(3)* $\hat{\beta}$是β的极大似然估计.

定理5.4表明：

（1）$\hat{\beta}$是β的线性无偏估计；

（2）$\hat{\beta}$的估计精度(方差)不但和随机误差有关，也与回归设计矩阵有关，这要求数据不能太集中；

（3）β的线性无偏估计中，$\hat{\beta}$的协方差阵最小；据此可得到(Gauss-Markve定理).

β的任一线性函数$c'\beta$的最小方差线性无偏估计为$c'\hat{\beta}$. 其中c是任一$p+1$维的列向量.

（4）残差向量$\tilde{y}$与$\hat{\beta}$的不相关.

证明：$\hat{\beta}=(X'X)^{-1}X'y$，$E(y)=X\beta$，$D(y)=\sigma^2 I_n$.

(1) $E(\hat{\beta})=E\left[(X'X)^{-1}X'y\right]=\left[(X'X)^{-1}X'\right]E(y)$

$$=\left[(X'X)^{-1}X'\right]X\beta=\beta.$$

$$D(\hat{\beta})=D\left[(X'X)^{-1}X'y\right]=\left[(X'X)^{-1}X'\right]D(y)\left[X(X'X)^{-1}\right]$$

$$=\left[(X'X)^{-1}X'\right]\sigma^2 I_n\left[X(X'X)^{-1}\right]$$

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

(2) 设$\tilde{\beta}=Ay$. 则 $D(\tilde{\beta})=\sigma^2(AA')$，由于

$$E(\tilde{\beta})=AE(y)=AX\beta=\beta$$

所以 $AX=I_{p+1}$，因此

$$0\le\left[A-(X'X)^{-1}X'\right]\cdot\left[A-(X'X)^{-1}X'\right]'$$

$$=\left[A-(X'X)^{-1}X'\right]\cdot\left[A'-X(X'X)^{-1}\right]=AA'-(X'X)^{-1}$$

从而 $AA'\geq(X'X)^{-1}$， 故 $D(\tilde{\beta})\geq D(\hat{\beta})$;

（3） $\tilde{y}=y-\hat{y}=y-X\hat{\beta}=y-X(X'X)^{-1}X'y$

$$=\left[I_n-X(X'X)^{-1}X'\right]y$$

矩阵$P=X(X'X)^{-1}X'$，$Q=I_n-X(X'X)^{-1}X'$都是对称幂等矩阵, 因而它们的特征值都是 1 或 0.

矩阵P有$p+1$个特征值为1, 有$n-p-1$个特征值为0.

矩阵Q有$p+1$个特征值为0, 有$n-p-1$个特征值为1.

$$\mathrm{Cov}(\hat{\beta},\tilde{y})=\mathrm{Cov}((X'X)^{-1}X'y,Qy)$$

$$=(X'X)^{-1}X'\mathrm{Cov}(y,y)Q^T$$

$$=(X'X)^{-1}X'(\sigma^2 I_n)\left[I_n-X(X'X)^{-1}X'\right]=0.$$

若再设 $\varepsilon \sim N(0,\sigma^2 I_n)$. 则

$$y \sim N(X\beta,\sigma^2 I_n).$$

$\hat{\beta}$与$\tilde{y}$是$y_1,\cdots,y_n$的线性组合，故它也服从正态分布.

由于 $E(\tilde{y}) = E\left\{\left[I_n - X(X'X)^{-1}X'\right]y\right\}$

$$= \left[I_n - X(X'X)^{-1}X'\right]X\beta = 0$$

$$D(\tilde{y}) = D(Qy) = QD(y)Q^T = Q(\sigma^2 I_n)Q = \sigma^2 Q$$

$$= \sigma^2\left[I - X(X'X)^{-1}X'\right]$$

故 $\hat{\beta} \sim N(\beta,\sigma^2(X'X)^{-1})$,

$$\tilde{y} \sim N(0,\sigma^2\left[I - X(X'X)^{-1}X'\right]);$$

由于$\hat{\beta}$与$\tilde{y}$不相关，因而它们相互独立.

似然函数为

$$L(\beta,\sigma^2)=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^{2n}}(y-X\beta)^T(y-X\beta)\right\}$$

$$\ln L(\beta,\sigma^2)=-\frac{n}{2}\ln 2\pi\sigma^2-\frac{1}{2\sigma^{2n}}(y-X\beta)^T(y-X\beta)$$

β的极大似然估计满足$\dfrac{\partial \ln L(\beta,\sigma^2)}{\partial\beta}=0$，整理得

$$\frac{\partial\left[(y-X\beta)^T(y-X\beta)\right]}{\partial\beta}=0.$$

此即为正规方程组, 因此$\hat{\beta}$是β的极大似然估计.

三　回归方程的显著性检验

在实际问题中，事先并不能断定随机变量 y 与自变量$x_1, x_2, \cdots, x_p$之间是否确有线性关系，在求线性回归方程之前，线性回归模型只是一种假设. 这种假设常常基于某种定性分析和经验判断. 因此求得线性回归方程后, 需要对 y 与$x_1, x_2, \cdots, x_p$的线性关系进行显著性检验. 如果没有线性关系, 则回归系数为零, 反之则不全为零. 即相当于进行下列假设检验:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \Leftrightarrow H_1 : \beta_i \text{不全为零}.$$

如果接受H_0，则表明随机变量 y与自变量$x_1, x_2, \cdots, x_p$之间的关系不能由线性回归模型来表示，如果拒绝H_0，则表明二者之间的关系可以由线性模型表示，但还需要对每个回归系数进行检验，以消除对因变量 y 影响不显著的自变量.

和一元线性回归分析一样，为建立对H_0的检验统计量，我们利用总偏差平方和分解式.

总偏差平方和

$$S_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}\left[(y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})\right]^2$$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})$$

由于$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\hat{\beta}_0)+(\hat{\beta}_0-\overline{y})\sum_{i=1}^{n}(y_i-\hat{y}_i)$$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{\beta}_1x_{i1}+\cdots+\hat{\beta}_px_{ip})+(\hat{\beta}_0-\overline{y})\sum_{i=1}^{n}(y_i-\hat{y}_i)$$

$$=\hat{\beta}_1\sum_{i=1}^{n}(y_i-\hat{y}_i)x_{i1}+\cdots+\hat{\beta}_p\sum_{i=1}^{n}(y_i-\hat{y}_i)x_{ip}+(\hat{\beta}_0-\overline{y})\sum_{i=1}^{n}(y_i-\hat{y}_i)$$

$=0$ (利用正规方程)

因而 $$S_T=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2=S_e+S_R$$

上式即为平方和分解式.

其中$S_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$，称为回归平方和，它表示$H_0$可能不真时，$E(y)$随 $x=(x_1,x_2,\cdots,x_p)$的变化而变化，从而在每一个$(x_{i1},x_{i2},\cdots,x_{ip})$处的回归值$\hat{y}_i$不同，$S_R$描述了$\hat{y}_i$的波动大小. $S_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$称为残差平方和，它表示随机误差等因素引起的$y_i$的波动大小.

下面的定理给出了S_R与S_e的概率分布.

定理5.5 设 $y\sim N(X\beta,\sigma^2 I_n)$，有

(1) $S_e/\sigma^2\sim\chi^2(n-p-1)$;

(2) 若H_0成立，则有$S_R/\sigma^2\sim\chi^2(p)$;

(3) S_R与S_e、$\bar{y}$独立（或$\hat{\beta}$与S_e、$\bar{y}$独立）.

证明： 由定理5.4知

$$\tilde{y} \sim N(0, \sigma^2 \left[I - X(X'X)^{-1} X' \right])$$

矩阵$Q = I_n - X(X'X)^{-1} X'$是对称幂等矩阵, 它有$p+1$个特征值为0, 有$n-p-1$个特征值为1. 因而存在正交矩阵H，使得

$$HQH^T = \begin{pmatrix} I_{n-p-1} & o \\ o & o_{p+1} \end{pmatrix} \stackrel{\Delta}{=} I^*$$

令$z = H\tilde{y}$，则$z \sim N(0, \sigma^2 I^*)$，因此

$$\frac{S_e}{\sigma^2} = \frac{\tilde{y}^T \tilde{y}}{\sigma^2} = \frac{z^T z}{\sigma^2} = \frac{z_1^2}{\sigma^2} + \frac{z_2^2}{\sigma^2} + \cdots + \frac{z_{n-p-1}^2}{\sigma^2}$$

$$\sim \chi^2(n-p-1);$$

若H_0成立，则 $\hat{\beta} \sim N(0, \sigma^2 (X'X)^{-1})$,

$$X\hat{\beta} \sim N(0, \sigma^2 X(X'X)^{-1} X'),$$

根据矩阵$X(X'X)^{-1}X'$也是幂等矩阵，此时亦有

$$HPH^T = H(I-Q)H^T = \begin{pmatrix} o_{n-p-1} & o \\ o & I_{p+1} \end{pmatrix} \stackrel{\Delta}{=} I^{**}$$

同上类似可证 $S_R + n\bar{y}^2 / \sigma^2 \sim \chi^2(p+1)$

再由定理5.4知:S_R与S_e、$\bar{y}$独立(或$\hat{\beta}$与S_e、$\bar{y}$独立).

$$S_R \sim \chi^2(p)$$ 证毕.

注：若H_0成立, 则 $y \sim N(0, \sigma^2 I_n)$,

由 $\sum_{i=1}^{n} y_i^2 = S_R + S_e + n\bar{y}^2$，亦可由柯赫伦定理证明.

根据定理5.5，假设H_0的检验统计量可取为

$$F=\frac{S_R/p}{S_e/(n-p-1)}$$

当H_0成立时，$F\sim F(p,n-p-1)$，给定显著性水平α，H_0的拒绝域为

$$W=\{F>F_\alpha(p,n-p-1)\}$$

整个检验可以列成方差分析表.

来源	平方和	自由度	均方和	F比	临界值	显著性
回归	S_R	p	S_R/p	$F=\frac{S_R/p}{S_e/n-p-1}$	F_α	
残差	S_e	n-p-1	S_e/n-p-1			
总和	S_T	n-1				

三 回归系数的显著性检验

如果回归方程经过假设检验后，拒绝H_0，则认为随机变量 y 与自变量$x_1,x_2,\cdots,x_p$之间线性关系显著，这表明$\beta_1,\beta_2,\cdots,\beta_p$不全为零，但不排除某些$\beta_j=0$. 因此需要对每个系数$\beta_j$是否为0进行检验. 即进行如下的检验：

$$H_0:\beta_j=0 \Leftrightarrow H_1:\beta_j\neq 0$$

$$\hat{\beta}\sim N(\beta,\sigma^2(X'X)^{-1})\Rightarrow \hat{\beta}_j\sim N(\beta_j,c_{jj}\sigma^2)$$

$$\left.\begin{array}{c}\Rightarrow(\hat{\beta}_j-\beta_j)\big/\sigma\sqrt{c_{jj}}\sim N(0,1)\\ S_e\big/\sigma^2\sim\chi^2(n-p-1)\\ \hat{\beta},S_e\text{独立}\end{array}\right\}\Rightarrow t=\frac{(\hat{\beta}_j-\beta_j)\big/\sqrt{c_{jj}}}{\sqrt{S_e\big/(n-p-1)}}$$

$$\sim t(n-p-1)$$

或 $$\Rightarrow F=\frac{(\hat{\beta}_j-\beta_j)^2\big/c_{jj}}{S_e\big/(n-p-1)}\sim F(1,n-p-1)$$

检验检验统计量为

$$t=\frac{\hat{\beta}_j\big/\sqrt{c_{jj}}}{\sqrt{S_e\big/(n-p-1)}}\overset{H_0\text{成立}}{\sim}t(n-p-1)$$

或 $$F=\frac{\hat{\beta}_j^2\big/c_{jj}}{S_e\big/(n-p-1)}\overset{H_0\text{成立}}{\sim}F(1,n-p-1)$$

H_0的拒绝域为

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

或 $$W=\left\{F>F_{\alpha}(1,n-p-1)\right\}$$

当有多个自变量对因变量 y 无显著性影响时，由于$\hat{\beta}$ 的各分量间的相关性，不能一次去掉所有不显著的自变量，原则上每次只剔除一个变量，先剔除其中F值或$|t|$值最小的变量，然后再对求得的新的回归方程进行检验，有不显著的再剔除，直到保留的变量都对 y 有显著影响为止.

三　回归系数的区间估计

由于
$$t = \frac{(\hat{\beta}_j - \beta_j) \big/ \sqrt{c_{jj}}}{\sqrt{S_e \big/ (n-p-1)}} \sim t(n-p-1)$$

将上述的t作为枢轴变量，即可得β_j的置信区间.

四　估计与预测

若回归方程经过检验是显著的，此时可以用来做估计和预测.

- 当$x = x_0 = (x_{01}, x_{02}, \cdots, x_{0p})$时，类似一元线性回归可将$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_0 x_{ip}$作为$E(y_0)$的点估计，亦可求其区间估计(略).
- 当$x = x_0$时, 可给出y_0的概率为$1-\alpha$的预测区间.

注:

$$\hat{y}_0 - E(y_0) \sim N(0, \sigma^2 x_0 (X'X)^{-1} x_0'),$$

$$y_0 - \hat{y}_0 \sim N(0, \sigma^2 \left[1 + x_0 (X'X)^{-1} x_0'\right]),$$

$$S_e/\sigma^2 \sim \chi^2(n-p-1).$$

根据S_e分别与$\hat{y}_0 - E(y_0)$、$y_0 - \hat{y}_0$相互独立，构造t分布即可(略).

五　实例分析