Stentor的天空

Stentor的技术随笔

多元线性回归

能用office07发布简直是太好了，这下子省了很多事。

1、多元线性回归模型

假定被解释变量Y与多个解释变量$X_1,X_2,\cdots,X_k$之间具有线性关系，是解释变量的多元线性函数，称为多元线性回归模型。即

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\mu \quad (1.1)$$

其中Y为被解释变量，$X_j(j=1,2,\cdots,k)$为k个解释变量，$\beta_j(j=0,1,2,\cdots,k)$为$k+1$个未知参数，$\mu$为随机误差项。

被解释变量Y的期望值与解释变量$X_1,X_2,\cdots,X_k$的线性方程为：

$$E(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k \quad (1.2)$$

称为多元总体线性回归方程，简称总体回归方程。

对于n组观测值$Y_i,X_{1i},X_{2i},\cdots,X_{ki}(i=1,2,\cdots,n)$，其方程组形式为：

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+\mu_i,(i=1,2,\cdots,n) \quad (1.3)$$

即

$$\begin{cases}Y_1=\beta_0+\beta_1X_{11}+\beta_2X_{21}+\cdots+\beta_kX_{k1}+\mu_1\\Y_2=\beta_0+\beta_1X_{12}+\beta_2X_{22}+\cdots+\beta_kX_{k2}+\mu_2\\\cdots\cdots\\Y_n=\beta_0+\beta_1X_{1n}+\beta_2X_{2n}+\cdots+\beta_kX_{kn}+\mu_n\end{cases}$$

其矩阵形式为

$$\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix}=\begin{bmatrix}1&X_{11}&X_{21}&\cdots&X_{k1}\\1&X_{12}&X_{22}&\cdots&X_{k2}\\\vdots&\vdots&\vdots&\vdots&\vdots\\1&X_{1n}&X_{2n}&\cdots&X_{kn}\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\\\vdots\\\beta_k\end{bmatrix}+\begin{bmatrix}\mu_1\\\mu_2\\\vdots\\\mu_n\end{bmatrix}$$

即

$$\mathbf{Y}=\mathbf{X\beta}+\boldsymbol{\mu} \quad (1.4)$$

其中

$\mathbf{Y}_{n\times1}=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix}$为被解释变量的观测值向量；$\mathbf{X}_{n\times(k+1)}=\begin{bmatrix}1&X_{11}&X_{21}&\cdots&X_{k1}\\1&X_{12}&X_{22}&\cdots&X_{k2}\\\vdots&\vdots&\vdots&\vdots&\vdots\\1&X_{1n}&X_{2n}&\cdots&X_{kn}\end{bmatrix}$为解释变量的观测值矩阵；$\boldsymbol{\beta}_{(k+1)\times1}=\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\\\vdots\\\beta_k\end{bmatrix}$为总体回归参数向量；$\boldsymbol{\mu}_{n\times1}=\begin{bmatrix}\mu_1\\\mu_2\\\vdots\\\mu_n\end{bmatrix}$为随机误差项向量。

总体回归方程表示为：

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \quad (1.5)$$

多元线性回归模型包含多个解释变量，多个解释变量同时对被解释变量Y发生作用，若要考察其中一个解释变量对Y的影响就必须假设其它解释变量保持不变来进行分析。因此多元线性回归模型中的回归系数为偏回归系数，即反映了当模型中的其它变量不变时，其中一个解释变量对因变量Y的均值的影响。

由于参数$\beta_0, \beta_1, \beta_2, \cdots, \beta_k$都是未知的,可以利用样本观测值$(X_{1i}, X_{2i}, \cdots, X_{ki}; Y_i)$对它们进行估计。若计算得到的参数估计值为$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$，用参数估计值替代总体回归函数的未知参数$\beta_0, \beta_1, \beta_2, \cdots, \beta_k$，则得多元线性样本回归方程：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki} \quad (1.6)$$

其中$\hat{\beta}_j (j = 0,1,2,\cdots,k)$为参数估计值，$\hat{Y}_i (i = 1,2,\cdots,n)$为$Y_i$的样本回归值或样本拟合值、样本估计值。

其矩阵表达形式为:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \quad (1.7)$$

其中$\hat{\mathbf{Y}}_{n\times 1} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$为被解释变量样本观测值向量$Y$的$n\times 1$阶拟合值列向量；$\mathbf{X}_{n\times(k+1)} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}$为解释变量$X$的$n\times(k+1)$阶样本观测矩阵；$\hat{\beta}_{(k+1)\times 1} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$为未知参数向量$\beta$的$(k+1)\times 1$阶估计值列向量。

样本回归方程得到的被解释变量估计值$\hat{Y}_i$与实际观测值Y_i之间的偏差称为残差e_i。

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_{2i} + \cdots + \hat{\beta}_{ki} X_{ki}) \quad (1.8)$$

2、多元线性回归模型的假定

与一元线性回归模型相同，多元线性回归模型利用普通最小二乘法(OLS)对参数进行估计时，有如下假定：

假定1 零均值假定：$E(\mu_i) = 0, i = 1,2,\cdots,n$，即

$$E(\boldsymbol{\mu}) = E\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} E(\mu_1) \\ E(\mu_2) \\ \vdots \\ E(\mu_n) \end{bmatrix} = 0 \quad (2.1)$$

假定2 同方差假定(μ的方差为同一常数)：

$$Var(\mu_i) = E(\mu_i^2) = \sigma^2, (i = 1,2,\cdots,n) \quad (2.2)$$

假定3 无自相关性：

$$Cov(\mu_i, \mu_j) = E(\mu_i \mu_j) = 0, (i \neq j, i, j = 1,2,\cdots,n)$$

Algorithm for Data Mining(9)
Data Structure and Algorithm
Database System(SQL Server,Oracle)(2)
Dynamic Language (1)
Numerical Computation(1)
Office Software(VBA etc)(4)
Software Design(Language,Disign，Test etc)(2)
Statistic Analysis(SAS,R etc) (34)

随笔档案

2011年1月(1)
2010年5月(1)
2009年11月(2)
2009年9月(1)
2009年5月(2)
2009年4月(2)
2009年3月(3)
2009年2月(3)
2009年1月(1)
2008年12月(2)
2008年11月(3)
2008年10月(5)
2008年9月(4)
2008年8月(24)

Statistic

Agri521'Blog
我的英语技术博客

最新评论

1. Re:逻辑回归模型
能发一份word吗?
2602196069@qq.com
--一厘米的阳光

2. Re:逻辑回归模型
能给一份word吗?
953022837@qqq.com
--katty_fat

3. Re:多元线性回归
学神!
--知好奇行

4. Re:多元线性回归
质量很高，很多没看懂。。。
--AutumnLight

5. Re:多元线性回归
数学太好了！！我看的一头雾水。。。
--Lansing999

阅读排行榜

1. 多元线性回归(17073)
2. 逻辑回归模型(16897)
3. 关联规则(apriori algorithm)(1)(15010)
4. Excel中如何作特大值数据的条形图(11882)
5. 在R中使用支持向量机(SVM)(2.Kernlab包)(11207)

评论排行榜

1. 逻辑回归模型(22)
2. 关联规则(apriori algorithm)(1)(8)
3. 多元线性回归(7)

$$E(\boldsymbol{\mu}\boldsymbol{\mu}') = E\left[\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{bmatrix}(\mu_1,\mu_2,\cdots,\mu_n)\right] = E\begin{bmatrix}\mu_1^2 & \mu_1\mu_2 & \cdots & \mu_1\mu_n\\ \mu_2\mu_1 & \mu_2^2 & \cdots & \mu_2\mu_n\\ \vdots & \vdots & \vdots & \vdots\\ \mu_n\mu_1 & \mu_n\mu_2 & \cdots & \mu_n^2\end{bmatrix}$$

$$= \begin{bmatrix}E(\mu_1^2) & E(\mu_1\mu_2) & \cdots & E(\mu_1\mu_n)\\ E(\mu_2\mu_1) & E(\mu_2^2) & \cdots & E(\mu_2\mu_n)\\ \vdots & \vdots & \vdots & \vdots\\ E(\mu_n\mu_1) & E(\mu_n\mu_2) & \cdots & E(\mu_n^2)\end{bmatrix}$$

$$= \begin{bmatrix}\sigma_\mu^2 & 0 & \cdots & 0\\ 0 & \sigma_\mu^2 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & \sigma_\mu^2\end{bmatrix} = \sigma_u^2\mathbf{I}_n \quad (2.3)$$

假定4 随机误差项μ与解释变量X不相关(这个假定自动成立):

$$Cov(X_{ji},\mu_i) = 0, (j = 1,2,\cdots,k, i = 1,2,\cdots,n) \quad (2.4)$$

假定5 随机误差项μ服从均值为零，方差为σ^2的正态分布:

$$\mu_i \sim N(0,\sigma_\mu^2\mathbf{I}_n) \quad (2.5)$$

假定6 解释变量之间不存在多重共线性:

$$rank(\mathbf{X}) = k+1 \le n$$

即各解释变量的样本观测值之间线性无关，解释变量的样本观测值矩阵$\mathbf{X}$的秩为参数个数k+1，从而保证参数$\beta_0,\beta_1,\beta_2,\cdots,\beta_k$的估计值唯一。

3、多元线性回归模型的参数估计

3.1回归参数的最小二乘估计

对于含有k个解释变量的多元线性回归模型

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i \quad (i = 1,2,\cdots, n)$$

设$\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$分别作为参数$\beta_0,\beta_1,\cdots,\beta_k$的估计量，得样本回归方程为:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$$

观测值Y_i与回归值$\hat{Y}_i$的残差e_i为:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_{2i} + \cdots + \hat{\beta}_{ki} X_{ki})$$

由最小二乘法可知$\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$应使全部观测值$Y_i$与回归值$\hat{Y}_i$的残差$e_i$的平方和最小，即使

$$Q(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_k) = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

$$= \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki})^2 \quad (3.1)$$

取得最小值。根据多元函数的极值原理，Q分别对$\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$求一阶偏导，并令其等于零，即

$$\frac{\partial Q}{\partial \hat{\beta}_j} = 0, (j = 1,2,\cdots,k) \quad (3.2)$$

即

4. Excel中如何作特大值数据的条形图(6)
5. 遗传编程算法(4)

推荐排行榜

1. 逻辑回归模型(4)
2. 多元线性回归(4)
3. 通过示例学SAS（1)--从外部文件读入数据(3)
4. 在R中使用支持向量机(SVM)（1）(2)
5. 遗传编程算法(2)

$$\begin{cases}\dfrac{\partial Q}{\partial \hat{\beta}_0}=2\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_{1i}-\hat{\beta}_2X_{2i}-\cdots-\hat{\beta}_kX_{ki})(-1)=0\\ \dfrac{\partial Q}{\partial \hat{\beta}_1}=2\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_{1i}-\hat{\beta}_2X_{2i}-\cdots-\hat{\beta}_kX_{ki})(-X_{1i})=0\\ \cdots\cdots\\ \dfrac{\partial Q}{\partial \hat{\beta}_k}=\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_{1i}-\hat{\beta}_2X_{2i}-\cdots-\hat{\beta}_kX_{ki})(-X_{ki})=0\end{cases}$$

化简得下列方程组

$$\begin{cases}n\hat{\beta}_0+\hat{\beta}_1\sum X_{1i}+\hat{\beta}_2\sum X_{2i}+\cdots+\hat{\beta}_k\sum X_{ki}=\sum Y_i\\ \hat{\beta}_0\sum X_{1i}+\hat{\beta}_1\sum X_{1i}^2+\hat{\beta}_2\sum X_{2i}X_{1i}+\cdots+\hat{\beta}_k\sum X_{ki}X_{1i}=\sum X_{1i}Y_i\\ \cdots\cdots\\ \hat{\beta}_0\sum X_{ki}+\hat{\beta}_1\sum X_{1i}X_{ki}+\hat{\beta}_2\sum X_{2i}X_{ki}+\cdots+\hat{\beta}_k\sum X_{ki}^2=\sum X_{ki}Y_i\end{cases} \quad (3.3)$$

上述$(k+1)$个方程称为正规方程，其矩阵形式为

$$\begin{bmatrix}n & \sum X_{1i} & \sum X_{2i} & \cdots & \sum X_{ki}\\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{2i}X_{1i} & \cdots & \sum X_{ki}X_{1i}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \sum X_{ki} & \sum X_{1i}X_{ki} & \sum X_{2i}X_{ki} & \cdots & \sum X_{ki}^2\end{bmatrix}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \vdots\\ \hat{\beta}_k\end{bmatrix}=\begin{bmatrix}\sum Y_i\\ \sum X_{1i}Y_i\\ \vdots\\ \sum X_{ki}Y_i\end{bmatrix} \quad (3.4)$$

因为

$$\begin{bmatrix}n & \sum X_{1i} & \sum X_{2i} & \cdots & \sum X_{ki}\\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{2i}X_{1i} & \cdots & \sum X_{ki}X_{1i}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \sum X_{ki} & \sum X_{1i}X_{ki} & \sum X_{2i}X_{ki} & \cdots & \sum X_{ki}^2\end{bmatrix}$$

$$=\begin{bmatrix}1 & 1 & \cdots & 1\\ X_{11} & X_{12} & \cdots & X_{1n}\\ X_{21} & X_{22} & \cdots & X_{2n}\\ \vdots & \vdots & \vdots & \vdots\\ X_{k1} & X_{k2} & \cdots & X_{kn}\end{bmatrix}\begin{bmatrix}1 & X_{11} & X_{21} & \cdots & X_{k1}\\ 1 & X_{12} & X_{22} & \cdots & X_{k2}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn}\end{bmatrix}=\mathbf{X'X}$$

$$\begin{bmatrix}\sum Y_i\\ \sum X_{1i}Y_i\\ \vdots\\ \sum X_{ki}Y_i\end{bmatrix}=\begin{bmatrix}1 & 1 & \cdots & 1\\ X_{11} & X_{12} & \cdots & X_{1n}\\ X_{21} & X_{22} & \cdots & X_{2n}\\ \vdots & \vdots & \vdots & \vdots\\ X_{k1} & X_{k2} & \cdots & X_{kn}\end{bmatrix}\begin{bmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{bmatrix}=\mathbf{X'Y}$$

设 $\hat{\boldsymbol{\beta}}=\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \vdots\\ \hat{\beta}_k\end{bmatrix}$为估计值向量

样本回归模型$\mathbf{Y}=\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}$两边同乘样本观测值矩阵$\mathbf{X}$的转置矩阵$\mathbf{X'}$，则有

$$\mathbf{X'Y}=\mathbf{X'X}\hat{\boldsymbol{\beta}}+\mathbf{X'e}$$

得正规方程组：

$$\mathbf{X'Y}=\mathbf{X'X}\hat{\boldsymbol{\beta}} \quad (3.5)$$

由假定(6)，$R(\mathbf{X})=k+1$，$\mathbf{X'X}$为$(k+1)$阶方阵，所以$\mathbf{X'X}$满秩，$\mathbf{X'X}$的逆矩阵$(\mathbf{X'X})^{-1}$存在。因而

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (3.6)$$

则为向量$\boldsymbol{\beta}$的OLS估计量。

以二元线性回归模型为例，导出二元线性回归模型的OLS估计量的表达式。由(1.3)式得二元线性回归模型为

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \mu_i$$

为了计算的方便，先将模型中心化。

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ji}, x_{ji} = X_{ji} - \bar{X}_j, (j = 1,2)$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, y_i = Y_i - \bar{Y}$$

$$L_{pq} = \sum x_{pi}x_{qi}, (p,q = 1,2)$$

$$L_{jY} = \sum x_{ji}y_i, (j = 1,2)$$

$$L_{YY} = \sum y_i^2$$

设$\alpha_0 = \beta_0 + \beta_1\bar{X}_1 + \beta_2\bar{X}_2$，则二元回归模型改写为中心化模型。

$$Y_i = \alpha_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \mu_i \quad (3.7)$$

记

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \alpha_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & 0 & 0 \\ 0 & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ 0 & \sum x_{2i}x_{1i} & \sum x_{2i}^2 \end{bmatrix}, \mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum Y_i \\ \sum x_{1i}Y_i \\ \sum x_{2i}Y_i \end{bmatrix} \quad (3.8)$$

将$L_{pq} = \sum x_{pi}x_{qi}, (p,q = 1,2)$代入得

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & 0 & 0 \\ 0 & L_{11} & L_{12} \\ 0 & L_{21} & L_{22} \end{bmatrix} \quad (3.9)$$

因为

$$\sum_{i=1}^{n} x_{ji}Y_i = \sum_{i=1}^{n} x_{ji}(y_i + \bar{Y}) = \sum_{i=1}^{n} x_{ji}y_i + \bar{Y}\sum_{i=1}^{n} x_{ji}$$

$$= \sum_{i=1}^{n} x_{ji}y_i = L_{jY}, (j = 1,2) \quad (3.10)$$

则

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum Y_i \\ L_{1Y} \\ L_{2Y} \end{bmatrix}$$

由(3.6)式得

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \mathbf{L}^{-1} \end{bmatrix} \begin{bmatrix} \sum Y_i \\ L_{1Y} \\ L_{2Y} \end{bmatrix} \quad (3.11)$$

其中

$$\mathbf{L}^{-1}=\begin{bmatrix}L_{11} & L_{12}\\ L_{12} & L_{22}\end{bmatrix}^{-1}=\frac{1}{L_{11}L_{22}-L_{12}L_{21}}\begin{bmatrix}L_{22} & -L_{12}\\ -L_{12} & L_{11}\end{bmatrix}$$

由(3.11)式可知

$$\hat{\alpha}_0=\overline{Y}$$

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}=\mathbf{L}^{-1}\begin{bmatrix}L_{1Y}\\ L_{2Y}\end{bmatrix}=\frac{1}{L_{11}L_{22}-L_{12}^2}\begin{bmatrix}L_{22} & -L_{12}\\ -L_{12} & L_{11}\end{bmatrix}\begin{bmatrix}L_{1Y}\\ L_{2Y}\end{bmatrix}$$

得

$$\hat{\beta}_1=\frac{L_{1Y}L_{22}-L_{2Y}L_{12}}{L_{11}L_{22}-L_{12}^2}\quad(3.12)$$

$$\hat{\beta}_2=\frac{L_{2Y}L_{11}-L_{1Y}L_{12}}{L_{11}L_{22}-L_{12}^2}\quad(3.13)$$

$$\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X}_1-\hat{\beta}_2\overline{X}_2\quad(3.14)$$

3.2随机误差项μ的方差σ_μ^2的估计量

样本回归方程得到的被解释变量估计值$\hat{Y}_i$与实际观测值Y_i之间的偏差称为残差e_i

$$e_i=Y_i-\hat{Y}_i=Y_i-(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_{2i}X_{2i}+\cdots+\hat{\beta}_{ki}X_{ki})$$

则

$$\mathbf{e}=\mathrm{Y}-\hat{\mathrm{Y}}=\mathrm{Y}-\mathrm{X}\hat{\boldsymbol{\beta}}=(\mathrm{X}\boldsymbol{\beta}+\boldsymbol{\mu})-\mathrm{X}[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}]$$

$$=(\mathrm{X}\boldsymbol{\beta}+\boldsymbol{\mu})-\mathrm{X}[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'(\mathrm{X}\boldsymbol{\beta}+\boldsymbol{\mu})]$$

$$=\mathrm{X}\boldsymbol{\beta}+\boldsymbol{\mu}-\mathrm{X}[\boldsymbol{\beta}+(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\mu}]$$

$$=\boldsymbol{\mu}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\mu}$$

$$=[\mathrm{I}_n-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']\boldsymbol{\mu}$$

设$\mathbf{P}=\mathrm{I}_n-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$，可以得出$\mathbf{P}$是$n$阶对称幂等矩阵，$\mathbf{P}=\mathbf{P}'$，$\mathbf{P}^2=\mathbf{P}$。于是

$$\mathbf{e}=\mathbf{P}\boldsymbol{\mu}$$

而残差的平方和为

$$\sum e_i^2=\mathbf{e}'\mathbf{e}=(\mathbf{P}\boldsymbol{\mu})'(\mathbf{P}\boldsymbol{\mu})=\boldsymbol{\mu}'\mathbf{P}'\mathbf{P}\boldsymbol{\mu}=\boldsymbol{\mu}'\mathbf{P}\boldsymbol{\mu}$$

$$=\boldsymbol{\mu}'[\mathrm{I}_n-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']\boldsymbol{\mu}$$

$$E(\mathbf{e}'\mathbf{e})=E\{\boldsymbol{\mu}'[\mathrm{I}_n-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']\boldsymbol{\mu}\}$$

$$=\sigma_\mu^2 tr[\mathrm{I}_n-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']$$

$$=\sigma_\mu^2[tr\mathrm{I}_n-tr\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']$$

$$=\sigma_\mu^2[n-(k+1)]$$

其中"tr"表示矩阵的迹，即矩阵主对角线元素的和。于是

$$\sigma_\mu^2=\frac{E(\mathbf{e}'\mathbf{e})}{n-(k+1)}=E\left(\frac{\mathbf{e}'\mathbf{e}}{n-(k+1)}\right)$$

随机误差项μ的方差σ_μ^2的无偏估计量，记作S_e^2，即$E(S_e^2)=\sigma_\mu^2$，$S_e^2=\hat{\sigma}_\mu^2$，S_e为残差的标准差(或回归标准差)。

因此

$$S_e^2 = \frac{\sum e_i^2}{n-k-1} = \frac{\mathbf{e}'\mathbf{e}}{n-k-1} \quad (3.15)$$

其中

$$\sum e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$= \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

$$= \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \quad (3.16)$$

例如,对于二元线性回归模型($k = 2$)

$$S_e^2 = \frac{\mathbf{e}'\mathbf{e}}{n-3} = \frac{\sum e_i^2}{n-3} \quad (3.17)$$

$$\sum e_i^2 = \mathbf{e}'\mathbf{e} = L_{YY} - \hat{\beta}_1 L_{1Y} - \hat{\beta}_2 L_{2Y}$$

$$= \sum Y_i^2 - \hat{\beta}_1 \sum X_{1i} Y_i - \hat{\beta}_2 \sum X_{2i} Y_i \quad (3.18)$$

3.3、估计参数的统计性质

1、线性性

指最小二乘估计量$\hat{\boldsymbol{\beta}}$是被解释变量的观测值$Y_1, Y_2, \cdots, Y_k$的线性函数。

由于

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

设$\mathbf{P} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$，则矩阵$\mathbf{P}$为一非随机的$(k+1)\times n$阶常数矩阵。所以

$$\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y} \quad (3.19)$$

显然最小二乘估计量$\hat{\boldsymbol{\beta}}$是被解释变量的观测值$Y_1, Y_2, \cdots, Y_k$的线性函数。

2、无偏性

将$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu}$代入(3-16)式得

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu}$$

$$= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu} \quad (3.20)$$

则

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu}]$$

$$= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\mu})$$

$$= \boldsymbol{\beta}$$

所以$\hat{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的无偏估计量。

3.最小方差性

设$\mathbf{P}$为$n\times p$阶数值矩阵，$\mathbf{X}$为$p\times n$阶随机矩阵(随机变量为元素的矩阵)，$\mathbf{Q}$为$n\times n$阶数值矩阵，则

$$E(\mathbf{PXQ}) = \mathbf{P}(E(\mathbf{X}))\mathbf{Q}$$

下面推导$\hat{\boldsymbol{\beta}}$的方差、协方差矩阵。

定义：$Var(\hat{\boldsymbol{\beta}}) = E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right]$

$$= E\left[\begin{bmatrix}\hat{\beta}_0-\beta_0\\ \hat{\beta}_1-\beta_1\\ \vdots\\ \hat{\beta}_k-\beta_k\end{bmatrix}\left(\hat{\beta}_0-\beta_0,\hat{\beta}_1-\beta_1,\cdots,\hat{\beta}_k-\beta_k\right)\right]$$

$$= \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0,\hat{\beta}_1) & \cdots & Cov(\hat{\beta}_0,\hat{\beta}_k)\\ Cov(\hat{\beta}_1,\hat{\beta}_0) & Var(\hat{\beta}_1) & \cdots & Cov(\hat{\beta}_1,\hat{\beta}_k)\\ \vdots & \vdots & \vdots & \vdots\\ Cov(\hat{\beta}_k,\hat{\beta}_0) & Cov(\hat{\beta}_k,\hat{\beta}_1) & \cdots & Var(\hat{\beta}_k)\end{bmatrix}$$

由(3.20)式得

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu}$$

$$\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)' = \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu}\right]' = \boldsymbol{\mu}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

所以

$$Var(\hat{\boldsymbol{\beta}}) = E\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\right]$$

$$= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu}\boldsymbol{\mu}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\mu}\boldsymbol{\mu}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma_\mu^2\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \sigma_\mu^2(\mathbf{X}'\mathbf{X})^{-1} \quad (3.21)$$

这个矩阵主对角线上的元素表示$\hat{\boldsymbol{\beta}}$的方差，非主对角线上的元素表示$\hat{\boldsymbol{\beta}}$的协方差。例如$Var(\hat{\beta}_i)$是位于$\sigma_\mu^2(\mathbf{X}'\mathbf{X})^{-1}$的第$i$行与第$i$列交叉处的元素(主对角线上的元素)；$Cov(\hat{\beta}_i,\hat{\beta}_j)$是位于$\sigma_\mu^2(\mathbf{X}'\mathbf{X})^{-1}$的第$i$行与第$j$列交叉处的元素(非主对角线上的元素)

在应用上，我们关心的$\hat{\boldsymbol{\beta}}$的方差，而忽略协方差，因此把(3.21)式记作

$$Var(\hat{\boldsymbol{\beta}}) = \sigma_\mu^2(\mathbf{X}'\mathbf{X})_{ii}^{-1} \quad (3.22)$$

记$\mathbf{S}^{-1} = (\mathbf{X}'\mathbf{X})^{-1} = (C_{ij}),(i,j=0,1,2,\cdots,k)$，则$Var(\hat{\beta}_i) = \sigma_\mu^2 C_{ii}$，所以$\hat{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的最小方差线性无偏估计。这说明，在(1.1)式系数的无偏估计量中，OLS估计量的方差比用其它估计方法所得的无偏估计量的方差都要小，这正是OLS的优越性所在。

用S_e^2代替σ_μ^2则得$\hat{\beta}_i$的标准估计量的估计值，乃称为标准差。

$$S(\hat{\beta}_i) = \sqrt{C_{ii}S_e^2} \quad (3.23)$$

其中

$$S_e^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$$

对于二元回归模型($k=2$)，求估计量$\hat{\beta}_1,\hat{\beta}_2$的方差，由(3.22)式得

$$Var(\hat{\boldsymbol{\beta}}) = \sigma_\mu^2(\mathbf{X}'\mathbf{X})_{ii}^{-1} = \sigma_\mu^2\begin{bmatrix}\frac{1}{n} & 0\\ 0 & \mathbf{L}^{-1}\end{bmatrix}_{ii}$$

其中

$$\mathbf{L} = \begin{bmatrix}L_{11} & L_{12}\\ L_{12} & L_{22}\end{bmatrix}$$

于是

$$Var\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}=\sigma_\mu^2 L_{ii}^{-1}\frac{\sigma_\mu^2}{L_{11}L_{22}-L_{12}^2}\begin{bmatrix}L_{22} & -L_{12}\\ -L_{12} & L_{11}\end{bmatrix}_{ii}$$

所以

$$Var(\hat{\beta}_1)=\sigma^2(\hat{\beta}_1)=\frac{L_{22}}{L_{11}L_{22}-L_{12}^2}\sigma_\mu^2 \quad (3.24)$$

$$Var(\hat{\beta}_2)=\sigma^2(\hat{\beta}_2)=\frac{L_{11}}{L_{11}L_{22}-L_{12}^2}\sigma_\mu^2 \quad (3.25)$$

$$S(\hat{\beta}_1)=\sqrt{\frac{L_{22}}{L_{11}L_{22}-L_{12}^2}S_e^2} \quad (3.26)$$

$$S(\hat{\beta}_2)=\sqrt{\frac{L_{11}}{L_{11}L_{22}-L_{12}^2}S_e^2} \quad (3.27)$$

其中

$$S_e^2=\frac{\mathbf{e}'\mathbf{e}}{n-3}$$

4. 显著性检验

4.1 拟合优度检验

4.1.1总离差平方和分解

设具有k个解释变量的回归模型为

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+\mu_i$$

其回归方程为

$$\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\cdots+\hat{\beta}_kX_{ki}$$

离差分解：

$$Y_i-\overline{Y}=(Y_i-\hat{Y}_i)+(\hat{Y}_i-\overline{Y})$$

总离差平方和分解式为：

$$\sum(Y_i-\overline{Y})^2=\sum(\hat{Y}_i-\overline{Y})^2+\sum(Y_i-\hat{Y})^2 \quad (4.1)$$

即

$TSS=ESS+RSS$ (4.2)总离差平方和分解为回归平方和与残差平方和两部分。$\sum(Y_i-\overline{Y})^2$体现了观测值$y_1,y_2,\cdots y_n$总波动大小，称为总偏差平方和，记作TSS. $\sum(\hat{Y}_i-\overline{Y})^2$体现了n个估计值$\hat{y}_1,\hat{y}_2,\cdots\hat{y}_n$的波动大小，它是由于Y与自变量$x_1,x_2,\cdots x_k$的变化而引起，被称作为回归平方和，记为ESS（Explained Sum of Squares）或U；$\sum(Y_i-\hat{Y})^2$称为残差平方和，记为RSS（Residual Sum of Squares）或Q.

4.1.2样本决定系数

对于多元回归方程，其样本决定系数为复决定系数或多重决定系数。

$R_{YX}^2,(i=1,2,\cdots,k)$，简记为$R^2$。

$$R^2=\frac{ESS}{TSS} \quad (4.3)$$

根据式(4.2)

$$R^2=1-\frac{RSS}{TSS} \quad (4.4)$$

因为

$$TSS = \sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - n\overline{Y}^2$$

由(3.16)式知

$$RSS = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$$

所以

$$ESS = TSS - RSS = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\overline{Y}^2$$

$$R^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\overline{Y}^2}{\mathbf{Y}'\mathbf{Y} - n\overline{Y}^2} \quad (4.5)$$

R^2作为检验回归方程与样本值拟合优度的指标：$R^2(0 \le R^2 \le 1)$越大，表示回归方程与样本拟合的越好；反之，回归方程与样本值拟合较差。

具体的，当$k = 2$时,求样本决定系数

$$R^2 = \frac{\sum (\hat{Y}_i - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2} = \frac{\sum y_i^2 - \sum e_i^2}{\sum y_i^2}$$

由(3.8)式，得$\sum e_i^2 = L_{YY} - \hat{\beta}_1 L_{1Y} - \hat{\beta}_2 L_{2Y}$，因此有

$$R^2 = \frac{\hat{\beta}_1 L_{1Y} + \hat{\beta}_2 L_{2Y}}{L_{YY}} \quad (4.6)$$

4.1.3调整后的样本决定系数

在使用R^2时，容易发现R^2的大小与模型中的解释变量的数目有关。如果模型中增加一个新解释变量，总离差TSS不会改变，但总离差中由解释变量解释的部分，即回归平方和ESS将会增加，这就是说R^2与模型中解释变量个数有关。但通过增加模型中解释变量的数目而使R^2增大是错误的，显然这样R^2来检验被回归方程与样本值拟合优度是不合适的，需要对R^2进行调整，使它不但能说明已被解释离差与总离差的关系，而且又能说明自由度的数目。

以$\overline{R}^2$表示调整样本决定系数，

$$\overline{R}^2 = 1 - \frac{S_e^2}{S_y^2} \quad (4.7)$$

其中

$$S_e^2 = \frac{\sum e_i^2}{n-k-1}, S_y^2 = \frac{\sum (Y_i - \overline{Y})^2}{n-1}$$

这里$n-k-1$是残差平方和的自由度，$n-1$是总离差平方和的自由度。

由(4.7)式得

$$\overline{R}^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \overline{Y})^2} \times \frac{n-1}{n-k-1} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

其中,n是样本观测值的个数,k是解释变量的个数。从式中可以看出，当增加一个解释变量时，由前面分析可知R^2会增加，引起$(1-R^2)$减少，而$\frac{n-1}{n-k-1}$增加，因而$\overline{R}^2$不会增加。这样用$\overline{R}^2$判定回归方程拟合优度，就消除了R^2对解释变量个数的依赖。

R^2或$\overline{R}^2$只能说明在给定的样本条件下回归方程与样本观测值拟合优度，并不能做出对总体模型的推测，因此不能单凭R^2或$\overline{R}^2$来选择模型，必须对回归方程和模型中各参数的估计量做显著性检验。

4.2方程显著性检验

由离差平方和分解(4.2)式可知，总离差平方和TSS的自由度为$n-1$，回归平方和ESS是由k个解释变量$X_1,X_2,\cdots,X_k$对Y的线性影响决定的。因此它的自由度为k。所以，残差平方和的自由度由总离差平方和的自由度减去回归平方和的自由度，即为$n-k-1$。

检验回归方程是否显著，

第一步，作出假设

$$H_0:\beta_1=\beta_2=\cdots=\beta_k=0$$

备择假设H_1：b_1、b_2、...、b_k不同时为0

第二步，在H_0成立的条件下，计算统计量F

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

第三步，查表临界值

对于假设H_0，根据样本观测值计算统计量F给定显著水平α，查第一个自由度为k，第二个自由度为$n-k-1$的F分布表得临界值$F_\alpha(k,n-k-1)$。当$F\geq F_\alpha(k,n-k-1)$时，拒绝H_0，则认为回归方程显著成立；当$F<F_\alpha(k,n-k-1)$时，接受H_0，则认为回归方程无显著意义。

4.3 参数显著性检验

回归方程显著成立，并不意味着每个解释变量$X_1,X_2,\cdots,X_k$对被解释变量Y的影响都是重要的。如果某个解释变量对被解释变量Y的影响不重要，即可从回归模型中把它剔除掉，重新建立回归方程，以利于对经济问题的分析和对Y进行更准确的预测。为此需要对每个变量进行考查，如果某个解释变量X对被解释变量Y的作用不显著，那么它在多元线性回归模型中，其前面的系数可取值为零。因此必须对β_i是否为零进行显著性检验。

由(3.23)式

$$S(\hat{\beta}_i)=\hat{\sigma}(\hat{\beta}_i)=\sqrt{C_{ii}S_e^2}\quad(4.8)$$

其中

$$S_e^2=\frac{\mathbf{e}'\mathbf{e}}{n-k-1}$$

C_{ii}为L^{-1}的第i个对角元素，而$L=\tilde{X}'\tilde{X}$，$\tilde{X}$是中心化的数据阵。

对回归系数$\hat{\beta}_i$进行显著性t检验，步骤如下：

(1)提出原假设$H_0:\beta_i=0$；备择假设$H_1:\beta_i\neq 0$。

(2)构造统计量$t=\frac{\hat{\beta}_i-\beta_i}{S(\hat{\beta}_i)}$，当$\beta_i=0$成立时,统计量$t=\frac{\hat{\beta}_i}{S(\hat{\beta}_i)}\sim t(n-k-1)$。这里$S(\hat{\beta}_i)$是$\hat{\beta}_i$的标准差，$k$为解释变量个数，计算由式(4.8)给出。

(3)给定显著性水平α，查自由度为$n-k-1$的t分布表，得临界值$t_{\frac{\alpha}{2}}(n-k-1)$。

(4)若$|t|\geq t_{\frac{\alpha}{2}}(n-k-1)$，则拒绝$H_0:\beta_i=0$，接受$H_1:\beta_i\neq 0$，即认为$\beta_i$显著不为零。若$|t|<t_{\frac{\alpha}{2}}(n-k-1)$，则接受$H_0:\beta_i=0$，即认为$\beta_i$显著为零。

5.回归变量的选择与逐步回归

5.1变量选择问题

在实际问题中，影响因变量Y的因素（自变量）很多，人们希望从中挑选出影响显著的自变量来建立回归关系式，这就涉及到自变量选择的问题。

在回归方程中若漏掉对Y影响显著的自变量，那么建立的回归式用于预测时将会产生较大的偏差。但回归式若包含的变量太多，且其中有些对Y影响不大，显然这样的回归式不仅使用不方便，而且反而会影响预测的精度。因

而选择合适的变量用于建立一个"最优"的回归方程是十分重要的问题。

选择"最优"子集的变量筛选法包括逐步回归法(Stepwise),向前引入法（Forward）和向后剔除法(Backwad)。

向前引入法是从回归方程仅包括常数项开始，把自变量逐个引入回归方程。具体地说，先在m个自变量中选择一个与因变量线性关系最密切的变量，记为 x_{i_1}，然后在剩余的m-1个自变量中，再选一个 x_{i_2}，使得 $\{x_{i_1}, x_{i_2}\}$ 联合起来二元回归效果最好，第三步在剩下的m-2个自变量中选择一个变量 x_{i_3}，使得 $\{x_{i_1}, x_{i_2}, x_{i_3}\}$ 联合起来回归效果最好，...如此下去，直至得到"最优"回归方程为止。

向前引入法中的终止条件为，给定显著性水平 α，当某一个对将被引入变量的回归系数作显著性检查时，若p-value $\geq \alpha$，则引入变量的过程结束，所得方程即为"最优"回归方程。

向前引入法有一个明显的缺点，就是由于各自变量可能存在着相互关系，因此后续变量的选入可能会使前面已选入的自变量变得不重要。这样最后得到的"最优"回归方程可包含一些对Y影响不大的自变量。

向后剔除法与向前引入法正好相反，首先将全部m个自变量引入回归方程，然后逐个剔除对因变量Y作用不显著的自变量。具体地说，从回归式m个自变量中选择一个对Y贡献最小的自变量，比如 x_{j_1}，将它从回归方程中剔除；然后重新计算Y与剩下的m-1个自变量回归方程，再剔除一个贡献最小的自变量，比如 x_{j_2},依次下去，直到得到"最优"回归方程为止。向后剔除法中终止条件与向前引入法类似。

向后剔除法的缺点在于，前面剔除的变量有可能因以后变量的剔除，变为相对重要的变量，这样最后得到的"最优"回归方程中有可能漏掉相对重要的变量。

逐步回归法是上述两个方法的综合。向前引入中被选入的变量，将一直保留在方程中。向后剔除法中被剔除的变量，将一直排除在外。这两种方程在某些情况下会得到不合理的结果。于是，可以考虑到，被选入的的变量，当它的作用在新变量引入后变得微不足道时，可以将它删除；被剔除的变量，当它的作用在新变量引入情况下变得重要时，也可将它重新选入回归方程。这样一种以向前引入法为主，变量可进可出的筛选变量方法，称为逐步回归法。

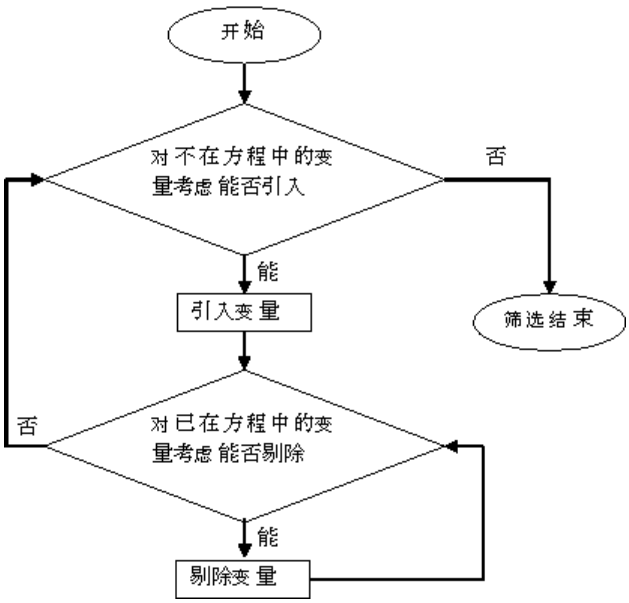
5.2逐步回归分析

5.2.1基本思想

逐个引入自变量。每次引入对Y影响最显著的自变量，并对方程中的老变量逐个进行检验，把变为不显著的变量逐个从方程中剔除掉，最终得到的方程中既不漏掉对Y影响显著的变量，又不包含对Y影响不显著的变量。

5.2.2筛选的步骤

首先给出引入变量的显著性水平 α_{in}和剔除变量的显著性水平 α_{out}，然后按下图筛选变量。



5.2.3逐步筛选法的基本步骤

逐步筛选变量的过程主要包括两个基本步骤：一是从回归方程中考虑剔除不显著变量的步骤；二是从不在方程中的变量考虑引入新变量的步骤。

（1）考虑可否引入新变量的基本步骤。假设已入选r个变量，不在方程中的变量记为 $x_{j_1},\cdots,x_{j_{m-r}}$.

1. 计算不在方程中的变量 x_{j_k} 的偏回归平方和 P_{j_k}：

$P_{j_k}=Q(i_1,\cdots,i_r)-Q(i_1,\cdots,i_r,j_k)$，Q表示括号中这些变量的回归模型的残差平方和。并设

$P_{j_0}=\max(P_{j_1},\cdots,P_{j_{m-r}})$，即不在方程中的变量 x_{j_0} 是对Y影响最大的变量。

1. 检验变量 x_{j_0} 对Y的影响是否显著。对变量 x_{j_0} 作回归系数的显著性检验，即检验 $H_0:\beta_{j_0}=0$，检验统计量为

$F_{j_0}=\dfrac{P_{j_0}}{Q(i_1,\cdots,i_r,j_0)/(n-r-2)}$ 及 $p=P\{F\geq F_{j_0}\}$，其中F～F(1,n-r-1)．

若p< α_{in},则引入 x_{j_0},并转入考虑可否剔除变量的步骤。若 $p\geq\alpha_{in}$,则逐步筛选变量的过程结束。

（2）考虑可否剔除变量的基本步骤。假设已引入回归方程的变量为 $x_{i_1},x_{i_2},\cdots,x_{i_r}$.

1. 计算已在方程中的变量 x_{i_k} 的偏回归平方和 P_{i_k}。

$P_{i_k}=Q(i_1,\cdots,i_{k-1},i_{k+1},\cdots,i_r)-Q(i_1,\cdots,i_r)$

$=U(i_1,\cdots,i_r)-U(i_1,\cdots,i_{k-1},i_{k+1},\cdots,i_r)$ 其中Q表示括号中这些变量的回归模型的残差平方和，U表示其回归平方和。设

$P_{i_0}=\min(P_{i_1},\cdots,P_{i_r})$，即相应的变量 x_{i_0} 是方程中对Y影响最小的变量。

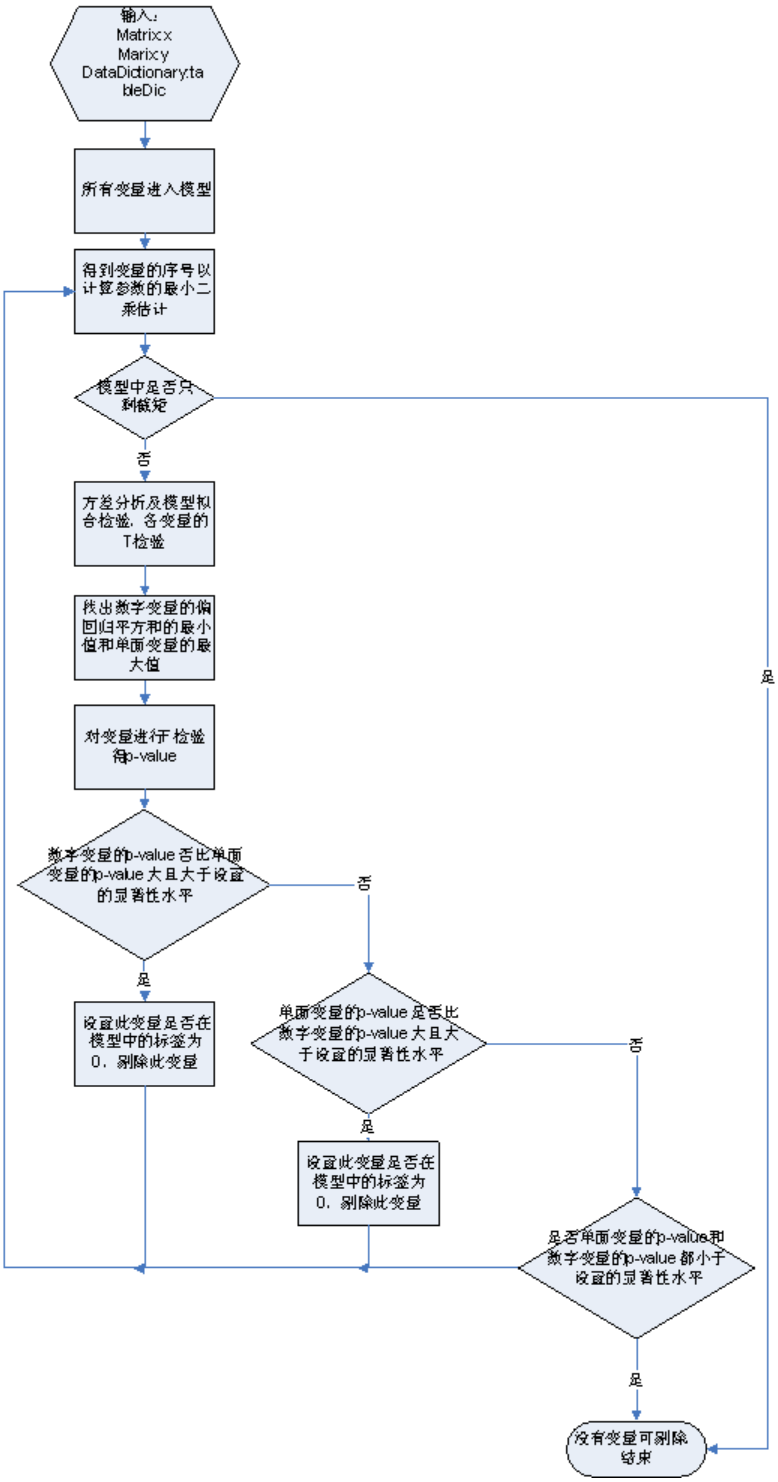
1. 检验 x_{i_0} 对Y的影响是否显著。对变量 x_{i_0} 进行回归系数的显著性检验，即检验 $H_0:\beta_{i_0}=0$，检验统计量为

$F_{i_0}=\dfrac{P_{i_0}}{Q(i_1,\cdots,i_r)/(n-r-1)}$ 及 $p=P\{F\geq F_{i_0}\}$,其中F～F（1,n-r-1）。

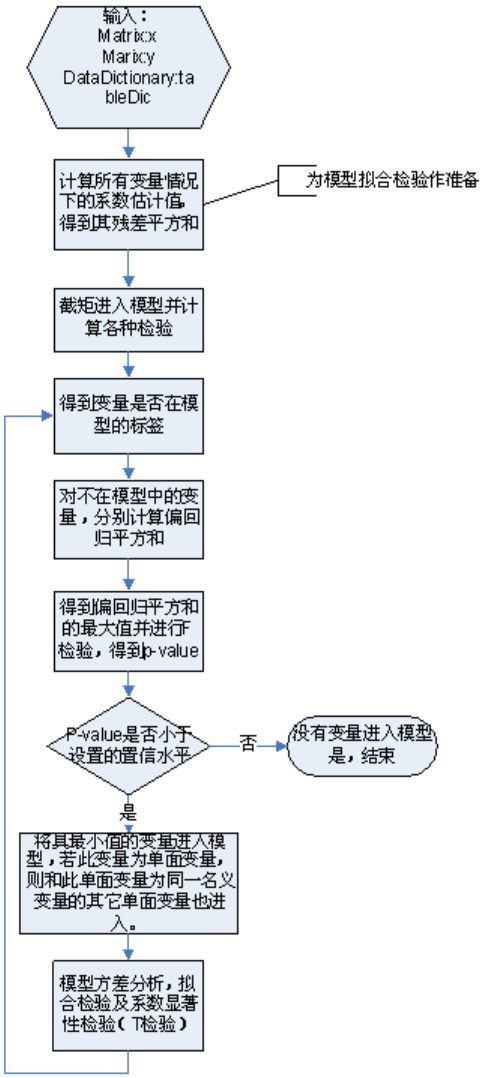
若p大于等于 α_{out}，则剔除 x_{i_0}，重新建立Y与其余r-1个变量的回归方程，然后再检验方程中最不重要的变量可否删除，直到方程中没有变量可删除后，转入考虑能否引入新变量的步骤。

5.3流程图

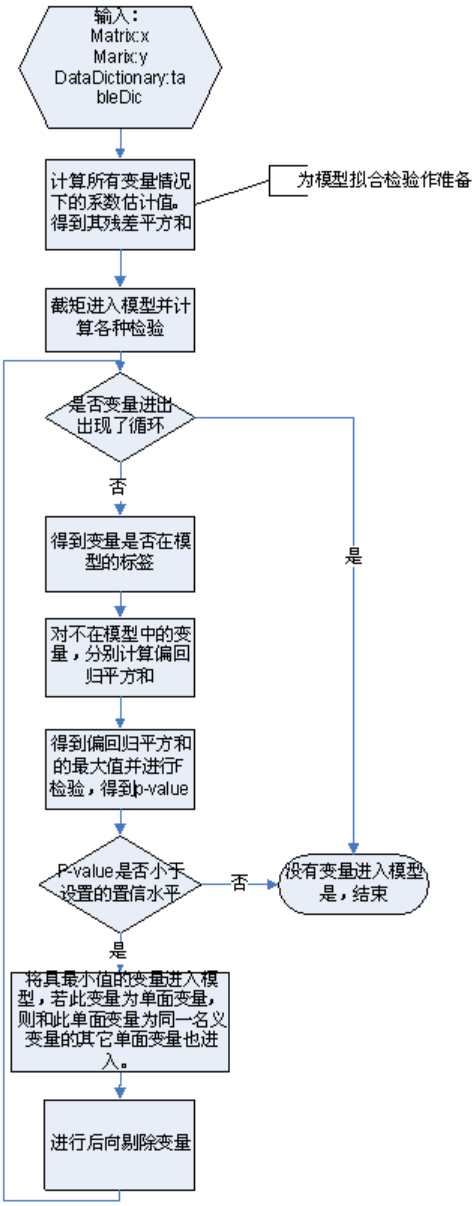
（1）后向选择



(2) 前向引入（Forward）



(3)逐步回归(Stepwise)



分类: Algorithm for Data Mining

zgw21cn
关注 - 1
粉丝 - 68
+加关注

4 0

« 上一篇： 逻辑回归Stepwise的R代码
» 下一篇： Sql server 行列互换制作交叉表格报表

posted on 2008-12-24 13:14 zgw21cn 阅读(17073) 评论(7) 编辑 收藏

评论

#1楼 2009-01-07 09:43 aliu

请问怎么把多元化俄回归模型转化为中心化的模型?

在你上面的ols回归只是以二元回归为例，把二元回归模型改写成了中心化模型。

支持(0) 反对(0)

#2楼 2009-01-07 17:49 stentor

@aliu
对观测数据进行中心化处理只是为了方便下面得出估计值B的具体的表达式而已。若需要对多元化的数据进行中心化处理，只需将各变量减去其平均值即可。

支持(0) 反对(0)

#3楼 2009-04-08 16:30 lz

帅哥，有没有word版本，传一个给我，参考下，谢谢了
邮箱：leezhiwei@126.com

支持(0) 反对(0)

#4楼 2015-11-22 20:03 丁丁的大森林

有没有word哦。。。

支持(0) 反对(0)

#5楼 2016-03-01 22:39 Lansing999

数学太好了！！我看的一头雾水。。。

支持(0) 反对(0)

#6楼 2016-10-21 14:38 AutumnLight

质量很高，很多没看懂。。。

支持(0) 反对(0)

#7楼 2016-10-29 15:19 知好奇行

学神!

支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册， 访问 网站首页。