# 应用数理统计

# Ch5 回归分析

## ------5.5 逐步回归

2014年6月25日

# 多元线性回归中的复共线性问题

## 一、在线性回归方程中，项数是否越多越好？

例如，对于同一批观测值$(x_{i1}, x_{i2}, y_i), i=1,2,...,n$，分别建立了下列两种形式的回归方程：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad (1)$$

$$y = \beta_0^* + \beta_1^* x_1 + \varepsilon \qquad (2)$$

问哪一个方程更好？

从残差平方和越小越好的角度来看，回归方程中的项数越多越好

对于上面例子中的回归方程（1）来说，问题相当于要求$\beta_0$，$\beta_1$，$\beta_2$的估计使得

$$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \qquad \text{达到最小}$$

对于上面例子中的回归方程（2）来说，问题相当于要求$\beta_0^*$，$\beta_1^*$的估计使得

$$Q^* = \sum_{i=1}^{n} (y_i - \beta_0^* - \beta_1^* x_{i1})^2 \qquad \text{达到最小}$$

如果我们已经求得$\beta_0^*=\beta_0^*$，$\beta_1^*=\beta_1^*$，使得$Q^*$达到最小值

只要令$\beta_0=\beta_0^*$，$\beta_1=\beta_1^*$，$\beta_2=0$，就可以使得

$$Q=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2})^2=\sum_{i=1}^{n}(y_i-\beta_0^*-\beta_1^*x_{i1})^2=Q_{\min}^*$$

$Q^*$能取到的最小值，$Q$必定也能取到，而

$Q$能取到的最小值，$Q^*$却不一定能取到

回归函数中自变元个数越多，回归的残差平方和**SSe**越小

## （2）但是,回归方程中的项数也并非越多越好

定义 对回归方程$y=\beta_0+\beta_1x_1+\cdots+\beta_m+\varepsilon$进行$n$次观测，

得到观测值（$x_{i1},x_{i2},\cdots,x_{im},y_i$） $i=1,2,...,n$

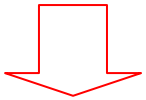
如果有一组不全为0的常数$\alpha_0,\alpha_1,\cdots,\alpha_m$，使得

$$\alpha_0\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}+\alpha_1\begin{bmatrix}x_{11}\\ \vdots\\ x_{n1}\end{bmatrix}+\cdots+\alpha_m\begin{bmatrix}x_{1m}\\ \vdots\\ x_{nm}\end{bmatrix}=\begin{bmatrix}0\\ \vdots\\ 0\end{bmatrix}$$

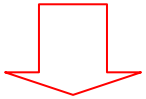
则称自变量$x_1,x_2,\cdots,x_m$之间存在复共线性（*multicollinearity*，也称为多重共线性）.

即: $X=\begin{bmatrix}1 & x_{11} & \cdots & x_{1m}\\ \vdots & \vdots & & \vdots\\ 1 & x_{n1} & \cdots & x_{nm}\end{bmatrix}$ $X$的秩$r(X)<X$的列数$m+1$

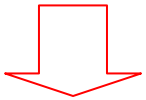
自变量$x_1, x_2, \cdots, x_m$存在复共线性



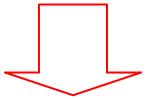
$X$的秩$r(X) < X$的列数$m+1$



$r(X^T X) < m+1$



$X^T X$不可逆



$\beta=(X^T X)^{-1} X^T Y$　　不能求出

## 二、在什么情况下，会发生复共线性？

**(1) 当观测次数n小于线性回归方程的项数m+1，就一定会产生负复线性**

矩阵$X=\begin{bmatrix}1 & x_{11} & \cdots & x_{1m}\\ \vdots & \vdots & & \vdots\\ 1 & x_{n1} & \cdots & x_{nm}\end{bmatrix}$ 有$n$行，$m+1$列，因为矩阵的秩总是小于它的行数，$n<m+1$，则显然后$r(X)<m+1$，这时一定会发生复共线性

这个问题怎样解决？

（1）增加试验次数n；（2）减少方程中的项数m+1

## (2) 有时即使观测次数n不小于回归方程的项数m+1，也会产生复共线性

### 例 （国际数学建模竞赛1993年A题）加速餐厅堆肥的生成

一家自助餐厅，每天把顾客吃剩下的食物搅拌成浆状，混入厨房里废弃的碎绿叶菜和少量撕碎的报纸，再加入真菌和细菌。混合原谅在真菌和细菌的消化作用下生成堆肥。

设$x_1, x_2, x_3$是三种堆肥原料的百分比含量，$y$ 是生成堆肥所需要的时间。题目中给出了$x_1, x_2, x_3$和$y$的一批观测数据，要求寻找生成堆肥所需时间与原料百分比含量之间的关系。

这显然是一个回归分析问题。如果我们认为$y$ 与$x_1, x_2, x_3$ 之间，只是简单的线性关系：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

用这样的方程进行回归分析求$\beta_1, \beta_2, \beta_3$的估计，肯定会发生问题。

因为$x_1, x_2, x_3$都是百分比含量，堆肥原料就是由着三种成分组成的，所以，对任何一组观测值来说，这3种成分的百分比含量加起来必定等于1，例如有

$$60\%+30\%+10\%=100\%=1,$$

……

$$50\%+35\%+15\%=100\%=1,$$

这时，显然可以找到一组不全为0的常数

$$\alpha_0=-1,\ \alpha_1=1,\ \alpha_2=1,\ \alpha_3=1$$

使得

$$\alpha_0\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}+\alpha_1\begin{bmatrix}x_{11}\\ \vdots\\ x_{n1}\end{bmatrix}+\alpha_2\begin{bmatrix}x_{12}\\ \vdots\\ x_{n2}\end{bmatrix}+\alpha_3\begin{bmatrix}x_{13}\\ \vdots\\ x_{n3}\end{bmatrix}$$

$$=-1\times\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}+1\times\begin{bmatrix}60\%\\ \vdots\\ 50\%\end{bmatrix}+1\times\begin{bmatrix}30\%\\ \vdots\\ 15\%\end{bmatrix}+1\times\begin{bmatrix}10\%\\ \vdots\\ 35\%\end{bmatrix}=\begin{bmatrix}0\\ \vdots\\ 0\end{bmatrix}$$

可见，矩阵X中的各列线性相关，用这样的回归方程作回归分析，必定会发生复共线性。这个问题的发生，与观测次数的多少无关，观测次数再多，也还是会出现复共线性。

# 2 逐步回归——克服复共线性的一种方法

## 一、逐步回归的基本思想

从一个只含常数项的回归方程出发，通过逐步引入和删除一些项的方法，选取一部分对回归贡献最大的项进入回归方程，使残差平方和尽可能小，而又不发生复共线性的现象

## 二、怎样衡量线性回归方程中各项贡献的大小

回归函数中自变元个数越多，回归的残差平方和**SSe**越小

设在一个线性回归方程中，除了常数项以外，有$m$个非常数项，残差平方和为$SS_e$,设删除第$j$项后的残差平方和为$SS_j$　　$SS_j \geq SS_e$

$SS_j - SS_e$越大，说明第$j$项的贡献越大

用统计理论可以证明，如果第j项对回归方程实际上没有任何贡献

$$F_j = \frac{SS_j - SS_e}{SS_e/(n-m-1)} \sim F(1, n-m-1)$$

## 三、逐步回归的具体步骤

事先给定两个非负常数

$F_{in}$——引入水平界限，

$F_{out}$——删除水平界限。

从一个只含常数项的线性回归方程$y=\beta_0+\varepsilon$出发。首先，在所有未引入回归方程的项中，找出一个$F_j$最大的项，如果它的$F_j>F_{in}$，就引入这一项。然后，在所有已经引入回归方程的项中，找出一个$F_j$最小的项，如果它的$F_j\le F_{out}$，就删除这一项,……。就这样一步一步做下去，引入，删除，引入，删除,……，直到方程内所有项都满足$F_j>F_{out}$，方程外所有项都满足$F_j\le F_{out}$为止。

为了避免出现“死循环”，事先给定的常数$F_{in}$和$F_{out}$必须满足$F_{in}\ge F_{out}$。

为什么？因为，如果$F_{in}<F_{out}$，就有可能出现某一项的$F_j$值正好有$F_{in}<F_j\le F_{out}$，从$F_j>F_{in}$来看，应该引入这一项，但是引入后，从$F_j\le F_{out}$来看，又应该删除这一项，这样一会儿引入，一会儿删除，一会儿引入，一会儿删除,……，就会陷入无休无止的循环反复中，永远也得不到结果。所以，要避免这样的情况，就要规定$F_{in}\ge F_{out}\ge 0$。

## 四、容许值和容许值水平界限

逐步回归的目的，是要避免出现复共线性，但是，仅仅依靠上面的步骤，还不足以保证不出现复共线性。

在回归分析的计算过程中，关键的一步，是要计算矩阵 $X^TX$ 的逆矩阵。如果存在复共线性。就会出现矩阵 $X^TX$ 不可逆或近似不可逆的现象。当 $X^TX$ 不可逆或近似不可逆时，按公式计算逆阵，就会遇到分母为0或分母近似为0的情况。这时，或者计算会溢出，或者会产生很大的计算误差，使计算结果非常不可靠。

为了避免这种情况，我们事先给定一个值，称为**容许值水平界限（Tolerance Level）**，记为Tol，通常取 $Tol = 10^{-2} \sim 10^{-7}$ ，在逐步回归过程中，每当我们要引入一项，都要看一下求你矩阵时用到的分母的绝对值的最小值，这个值称为**容许值（Tolerance）**，如果容许值小于事先给定的容许值水平界限，即使其它条件满足，我们也不引入这一项。这样，就可以完全避免出现复共线性。

## 五、逐步回归计算实例

例 1932年，H. Woods，H. H. Steinour和H. R. Starke为了研究波特兰水泥的成分与水泥固话时放出的热量之间的关系，收集了13个水泥样品的数据，进行回归分析。

回归方程为 $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\varepsilon$

其中，自变量是4种成分在水泥总重量中所占的百分比：

$x_1$ ——$3CaO\cdot Al_2O_3$在水泥总重量中所占的百分比；

$x_2$ ——$3CaO\cdot SiO_2$在水泥总重量中所占的百分比；

$x_3$ ——$4CaO\cdot Al_2O_3\cdot Fe_2O_3$在水泥总重量中所占的百分比；

$x_4$ ——$2CaO\cdot SiO_2$在水泥总重量中所占的百分比；

因变量是

$y$ ——单位质量的水泥固化时放出的热量（单位：卡/克）。

由于水泥主要是由这4种成分构成的，所以，这4种成分的百分比含量加起来近似等于100%，即有 $x_1+x_2+x_3+x_4\approx 100\%$

在这个回归方程中，存在着复共线性。为了避免复共线性可能会带来的不良结果，考虑采用逐步回归。事先给定：引入水平界限Fin=4.0，删除水平界限Fout=3.9，容许值水平界限Tol=0.00001.

初始状态：回归方程中没有变量，只有常数项

| 方程内的项 | $\hat{\beta}_j$ | $F_j$ | 方程外的项 | 容许值 | $F_j$ |
|---|---|---|---|---|---|
| 常数项 | 95.423 | | | | |
| | | | x1 | 1.000 | 12.60 |
| | | | x2 | 1.000 | 21.96 |
| | | | x3 | 1.000 | 4.40 |
| | | | x4 | 1.000 | 22.80 |

第一步：在方程外的项中，$x_4$的$F_j$最大，而且$F_4=22.80>4.0=F_{in}$，它的容许值为1.0000>0.0001=Tol，引入$x_4$.

| 方程内的项 | $\hat{\beta}_j$ | $F_j$ | 方程外的项 | 容许值 | $F_j$ |
|---|---|---|---|---|---|
| 常数项 | 117.57 | | | | |
| | | | x1 | 0.940 | 108.22 |
| | | | x2 | 0.053 | 0.17 |
| | | | x3 | 0.999 | 40.29 |
| x4 | -0.7382 | 22.80 | | | |

第二步：在方程内的项中，$x_4$的$F_j$最小，但是$F_4=22.80>3.9=F_{out}$，不删除。在方程外的项中，$x_1$的$F_j$最大，而且$F_4=108.22>4.0=F_{in}$，它的容许值为0.940>0.0001=Tol，引入$x_1$.

| 方程内的项 | $\hat{\beta}_j$ | $F_j$ | 方程外的项 | 容许值 | $F_j$ |
|---|---|---|---|---|---|
| 常数项 | 103.10 | | | | |
| x1 | 1.440 | 108.22 | | | |
| | | | x2 | 0.053 | 5.03 |
| | | | x3 | 0.289 | 4.24 |
| x4 | -0.6140 | 159.30 | | | |

第三步：在方程内的项中，$x_1$的$F_j$最小，但是$F_1=108.22>3.9=F_{out}$，不删除。在方程外的项中，$x_2$的$F_j$最大，而且$F_2=5.03>4.0=F_{in}$，它的容许值为0.053>0.0001=Tol，引入$x_2$.

| 方程内的项 | $\hat{\beta}_j$ | $F_j$ | 方程外的项 | 容许值 | $F_j$ |
|---|---|---|---|---|---|
| 常数项 | 71.648 | | | | |
| x1 | 1.452 | 154.01 | | | |
| x2 | 0.4161 | 5.03 | | | |
| | | | x3 | 0.021 | 0.02 |
| x4 | -0.2365 | 1.86 | | | |

第四步：在方程内的项中，$x_4$的$F_j$最小，而且$F_4=1.86<3.9=F_{out}$，删除x4。
在方程外的项中，$x_3$的$F_j$最大，但是$F_3=0.02<4.0=F_{in}$，不引入。

| 方程内的项 | $\hat{\beta}_j$ | $F_j$ | 方程外的项 | 容许值 | $F_j$ |
|---|---|---|---|---|---|
| 常数项 | 52.577 | | | | |
| x1 | 1.4683 | 146.52 | | | |
| x2 | 0.66225 | 208.58 | | | |
| | | | x3 | 0.318 | 1.83 |
| | | | x4 | -0.053 | 1.86 |

第五步：在方程内的项中，$x_1$的$F_j$最小，但是$F_1=146.52>3.9=F_{out}$，不删除。

在方程外的项中，$x_4$的$F_j$最大，但是$F_4=0.02<4.0=F_{in}$，不引入。

这时，既没有可删除的项，也没有可引入的项，逐步回归结束。

得到回归方程:

$$\hat{y}=52.277+1.4683x_1+0.66225x_2$$

它的残差平方和为 $SS_e=57.90$

估计的标准差为 $\hat{\sigma}=2.406$

多重相关系数为 $r=0.9893$