兼顾类别的回归模型及其应用*

魏凤英 曹鸿兴
(中国气象科学研究院,北京 100081)

提 要

在本文中设计了一种用双因变量和多自变量构成的回归模型. 因变量由预报量本身及其类别组成,所建模型可同时对预报量和类别作出预报. 天气预报实例表明,本模型可在许多预测问题中使用.

关键词: 多元回归;判别分析;双评分准则;天气预报.

一、引 言

在许多预测问题中,例如市场预测、天气预报、产量预测等等,人们主要关心未来的趋势变化,市场价格是上涨、平稳还是下降? 汛期是旱还是涝? 在趋势预报正确的前提下,我们才进一步要求预测的数值更接近观测值. 为此,我们设计了一种既可以预报类别又可以预报数量的回归模型. 这一模型由双因变量(预报量)和多自变量构成. 文中还叙述了筛选变量的双评分准则(CSC). 最后,给出了在天气预报中应用的实例.

二、模 型

考虑双预报量 Y_i, $i=1,2$, Y_1 是观测值, Y_2 是划分的类别; 预报因子 $x_i, i=1,2,\cdots,p$ 和系数 b_i, $i=1,2$. 那么,资料矩阵为

$$\underset{n\times 2}{Y}=\begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \\ \cdots\cdots & \\ Y_{n1} & Y_{n2} \end{bmatrix},\quad \underset{n\times(p+1)}{X}=\begin{bmatrix} 1 & X_{11} & X_{12}\cdots X_{1p} \\ 1 & X_{21} & X_{22}\cdots X_{2p} \\ & \cdots\cdots & \\ 1 & X_{n1} & X_{n2}\cdots X_{np} \end{bmatrix},\quad \underset{(p+1)\times 2}{B}=\begin{bmatrix} b_{01} & b_{02} \\ b_{11} & b_{12} \\ \cdots\cdots & \\ b_{p1} & b_{p2} \end{bmatrix}, \tag{1}$$

其中 n 是样本量, p 是预报因子个数. 回归模型为

$$Y=XB+\varepsilon. \tag{2}$$

(2)式的最小二乘估计为

$$\hat{B}=(X^TX)^{-1}X^TY, \tag{3}$$

其中 T 表示矩阵转置, -1 表示矩阵求逆.

于是,我们得到预报数量和类别的两个方程

1993 年 3 月 24 日收到, 5 月 31 日收到修改稿.
* 气象基金资助项目.

$$\begin{bmatrix}\hat{Y}_1\\ \hat{Y}_2\end{bmatrix}=\begin{bmatrix}b_{01}\\ b_{02}\end{bmatrix}1+\begin{bmatrix}b_{11}\\ b_{12}\end{bmatrix}X_1+\cdots+\begin{bmatrix}b_{P1}\\ b_{P2}\end{bmatrix}X_P \tag{4}$$

三、类别划分

把预报量按照预报的要求划分为 n 类，例如汛期降水量根据出现的概率分为 5 类，即

$$\begin{array}{ll}1\text{类：}P_1<15\% & \text{旱}\\ 2\text{类：}15\%\leqslant P_2\leqslant 35\% & \text{偏旱}\\ 3\text{类：}35\%<P_3<65\% & \text{正常}\\ 4\text{类：}65\%\leqslant P_4\leqslant 85\% & \text{偏涝}\\ 5\text{类：}P_5>85\% & \text{涝}\end{array} \tag{5}$$

其中 $P_i=n_i/n, i=1,2,\cdots 5$，$P_i$ 和 n_i 分别为累积概率和频数。

四、判别重心

确定矩阵 Y 中 Y_2 的一个简单方法是对第 g 类的每个样本假定 $Y_{i2}=Z_g, g=1,2,\cdots,G; i=1,2,\cdots,n$，$Z_g$ 称为第 g 类的重心或判别参考点，G 是类别数。

设

$$Z_g=\bar{Y}_{g1},\qquad g=1,2,\cdots,G \tag{6}$$

$$\bar{Y}_{g1}=\frac{1}{n_g}\sum_{i=1}^{n_g}Y_{i1},$$

其中 n_g 是第 g 类的样本个数[1-2]。

按照预报的具体要求，确定划分类别的阈值。例如对 3 类预报来讲，确定两个阈值 Y_u 和 Y_l，于是 Y_{i2} 可以简单地取为

$$Y_{i2}=\begin{cases}1 & \text{当 } Y_{i1}>Y_u \text{ 时},\\ 2 & \text{当 } Y_l\leqslant Y_{i1}\leqslant Y_u \text{时},\\ 3 & \text{当 } Y_{i1}<Y_l \text{ 时}.\end{cases}\qquad i=1,2,\cdots,n \tag{7}$$

五、筛选因子的双评分准则

用 S_1 表示数量评分，S_2 表示趋势评分，双评分准则[3]定义为

$$\mathrm{CSC}=S_1+S_2, \tag{8}$$

式中 S_1 与 S_2 的定义如下确定。令

$$Q_k=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \tag{9}$$

为回归模型的残差平方和（RSS）。其中 k 为模型的独立参数个数，Y_i 是预报量，$\hat{Y}_i$ 为其估计值。显然，Q_k 等价于均方根误差（RMSE）

$$\varepsilon = \left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right]^{1/2}.$$

令总离差平方和（TSS）为预报评分，即

$$Q_Y = \sum_{i=1}^{n}(Y_i - \bar{Y})^2. \tag{10}$$

TSS 实际上是气候学预报（每次都用均值作为预报值）的评分.

设

$$S_1 = \frac{Q_k}{Q_Y}, \tag{11}$$

或等价地取

$$S_1 = 1 - \frac{Q_k}{Q_Y}. \tag{12}$$

S_1 表示一个好的模型其预报效果应优于气候学预报，即 $Q_k < Q_Y$.

趋势评分我们取最小判别信息统计量 $2I$[4]

$$S_2 = 2I = 2\left[\sum_{i=1}^{G}\sum_{j=1}^{G} n_{ij}\ln n_{ij} + n\ln n - \left(\sum_{i=1}^{G} n_{i.}\ln n_{i.} + \sum_{j=1}^{G} n_{.j}\ln n_{.j}\right)\right], \tag{13}$$

式中 G 为观测和预报趋势类别数，n_{ij} 为 i 类观测事件与 j 类预报事件 $G \times G$ 列联表中的个数，其中

$$n_{i.} = \sum_{j=1}^{G} n_{ij},\quad n_{.j} = \sum_{i=1}^{G} n_{ij}.$$

在线性模型中我们取 S_1 为

$$S_1 = (n-k)\left(1 - \frac{Q_k}{Q_Y}\right), \tag{14}$$

这里 k 为模型中因子个数. 逐步回归筛选因子的多个应用实例表明[5]，CSC 是一种行之有效的变量选择准则. 但是，目前尚未从理论上证明它的分布，也没有严格的数学推导证明 S_1 和 S_2 完全独立. 我们不妨将它作为一种经验准则使用.

六、计 算 步 骤

为简化计算，下面我们采用了一种简单的计算过程.

（a）用(6)式或(7)式确定 Y_2，构造出预报量矩阵 Y.

（b）计算每个预报因子 $x_l, l = 1,2,\cdots,P$ 的 CSC，这里(14)式中 $k = 1$. 建立一元回归

$$Y = XB + \varepsilon,$$

其中 $X = (1, x_l), B = (b_0, b_l)^T$.

（c）按照 CSC 值从大到小的原则，逐个引进预报因子 x_l，当 CSC 达到最大时，就确定了方程人选的因子.

与普通回归所不同的是，根据上述过程得到的预报因子可以同时预报出预报量的数

量和类别．另外，如果预报的数量和类别是一致的，那么就会增强我们对这一预报的信心．反之，如果数量和类别不一致，则提示我们要谨慎使用这一预报，并寻求更多的资料和方法来验证这一预报．这就是我们提出建立兼顾类别的回归模型的目的所在．

七、计 算 实 例

例 1　预报量 Y_1 取江苏常熟 1957—1978 年 4 月的降水量，按预报经验取定 $Y_*=40\text{mm}$，$Y_L=15\text{mm}$．Y_2 用(7)式来确定．预报因子为：

X_1：前一年 5 月日照时数，X_2：前一年 9 月日照时数，

X_3：前一年 8 月平均气温，X_4：前一年 12 月平均气温，

X_5：前一年 12 月降水量，X_6：前一年 9 月平均气温，

X_7：前一年 10 月平均气温．

这样我们构造出

$$Y^T=\begin{bmatrix}28.60 & 8.60 & 16.80 & 62.70\cdots\cdots & 14.00\\ 2 & 1 & 2 & 3 & \cdots\cdots 1\end{bmatrix},$$

用每个预报因子分别建立一个回归模型，例如对 X_1，回归系数求出是

$$b_{01}=47.41,\quad b_{11}=-0.09169,$$
$$b_{02}=2.455,\quad b_{12}=-0.002405,$$
$$S_1=0.597.$$

按照(4)式我们可以得到引入 X_1 的 $\hat{Y}_{i2}, i=1,2,\cdots,n$．表 1 为引进预报因子 X_1 的列联表．

表 1　引进预报因子 x_1 的列联表

实况 \ 预报	1 类	2 类	3 类	总计 $n_{i.}$
1 类	2	1	1	4
2 类	3	5	5	13
3 类	1	3	1	5
总计 $n_{.j}$	6	9	7	$n=22$

用(13)式计算出 $2I=1.98$．因此，$\mathrm{CSC}(x_1)=2.58$．同理，得到 $\mathrm{CSC}(x_2)=12.80$，$\mathrm{CSC}(x_3)=10.61$，$\mathrm{CSC}(x_4)=5.57$，$\mathrm{CSC}(x_5)=10.74$，$\mathrm{CSC}(x_6)=6.93$，$\mathrm{CSC}(x_7)=0.721$．

预报因子依 CSC 值从大到小进入方程．表 2 给出引进因子 CSC 的变化．

表 2　引进因子 CSC 的变化

预报因子	X_1	X_5	X_3	X_6	X_4	X_2	X_7
CSC	12.80	16.64	24.45	24.47	23.94	43.07	40.99

从表中看出，当方程引入 x_2, x_5, x_3, x_6, x_4 和 x_1 后，CSC 达到最大，CSC = 43.07 > $x^2_{10,0.01}$ = 23.21，停止筛选。利用(3)式的最小二乘估计，我们得到

$$\hat{Y}_1 = -158.8 + 0.3233x_2 + 3.182x_5 + 9.497x_3 - 2.701x_6 - 0.4963x_4 - 0.3350x_1,$$

$$\hat{Y}_2 = -4.204 + 0.01369x_2 + 0.08593x_5 + 0.3393x_3 - 0.05496x_6 - 0.04557x_4 - 0.01229x_1,$$

方程 Y_1 的均方根误差 (RMSE) 是 9.39mm，类别预报 Y_2 与实况一致的有 17 年，准确率为 0.77。用上述方程作出 1979 和 1980 年的预报。其中 $\hat{Y}_1(1979) = 27.37$mm，实况 $Y_1(1979) = 16.2$mm；$\hat{Y}_1(1980) = 3.69$mm，实况 $Y_1(1980) = 12.9$mm。$\hat{Y}_2(1979) = 1.84 \approx 2, Y_2(1979) = 2; \hat{Y}_2(1980) = 0.80 \approx 1, Y_2(1980) = 1$。预报类别与实况吻合。

从表 2 中我们看到，当 X_2, X_5, X_3 和 X_6 引入方程后，CSC = 24.47 达极大。这一事实提示我们：在预报因子不是很多时，最好的方法是考查所有因子的组合，筛选最优回归子集。

例 2 预报量为长江流域 1952—1987 年 5—9 月降水指数，给定 $y_u = 10, y_l = -10$，Y_2 用(7)式划分。根据计算相关系数选取了冬季和春季北太平洋海温、副热带高压面积指数及这一地区气温等级等 10 个预报因子。表 3 为引入因子后的 CSC 值。当 CSC = 18.5 时达到最大，回归模型共包括 6 个预报因子。

表 3 引入因子后的 CSC 值

预报因子	X_4	X_6	X_1	X_9	X_{10}	X_3
CSC	7.79	12.5	12.1	12.1	16.8	18.5
S_1	0.91	2.95	3.16	3.16	6.62	6.40
S_2	6.88	9.59	8.91	8.91	10.2	12.1

Y_1 的 RMSE 是 13.56，Y_2 的准确率为 23/36 = 0.64%。

八、小 结

本文设计了一个兼顾数量和类别的回归模型，它是一种包括回归和判别分析的混合模型。因此，它有较广泛的应用领域。实例表明，本文提出的模型和筛选变量的过程是适用的。当然，仍有许多需要进一步改进和完善之处。譬如，我们需要研究确定更合适重心的方案；设计从所有可能子集回归中根据变量选择标准，建立最优回归模型的方案。另外，就三类判别分析而言，(1)式中的 Y_2 应由两列构成。

参 考 文 献

[1] 曹鸿兴，1978，最小方差准则的判别分析，大气科学，Vol. 2, No. 2., 169—173.

[2] Cao Hongxing and Yang ziqiang, 1983, Multi-dichotomy Variant of Stepwise Disriminant and *Its Application to Long-range Weather Forecasting*. Preprints of Eight Conference on Probability and Stalistics in Atmospheric Sciences, NOV. 16—18, Hot spring, Arkonsas.

[3] 赖风英、曹鸿兴，1990，长期预测的数学模型及其应用，气象出版社，29—47.

[4] Kullback S., 1968, Information Theory and Statistics, Dover Publications, Inc., New York, 113—119.
[5] 牛保山、曹鸿兴、刘生长，1993，双评分准则逐步回归法，气象，Vol. 19, No. 8, 18—21.

Regression Model Adjoined by Category Prediction and Its Application

Wei Fengying　Cao Hongxing
(Chinese Academy of Meteorological Sciences, Beijing 100081)

Abstract

A special regression model consisting of bi-predictand and multipredictor is proposed in this paper. The bi-predictand is composed of predictand and categories specified. The model can predict simultaneously both the quantity and the category of the predictand. Examples in weather forecast indicate the model is feasible for some prediction studies.

Key words: Multivariate regression; Discriminant analysis; Couple score criterion; Weather forecast.