文章编号:1005－0523(2005)02－0144－03

通过交叉验证准则选择线性模型

闻 斌,江其保

(东南大学 数学系,江苏 南京 210096)

摘要:考虑建立在交叉验证准则基础上线性回归模型的选择问题.我们对原来的交叉验证准则进行改进,通过增加惩罚函数来解决交叉验证过程中模型过度拟合问题,从而提出一个新的模型选择准则.在一定的假设条件下,新准则确定的模型具有强相合性并且在样本容量充分大时能得到最小的真实模型.在本文中,我们将证明新准则确定的模型在一定条件具有强相合性,并给出一般条件下模型选择准则.

关 键 词:相合性;交叉验证;线性回归;模型选择

中图分类号:O212.4 **文献标识码**:A

1 引 言

考虑多元线性回归模型

$$Y_n = X_n\beta + e_n \tag{1.1}$$

其中:X_n 为 $n\times p$ 的矩阵,β 为 $p\times 1$ 的未知回归参数向量,e_n 为 $n\times 1$ 随机误差向量(e_n 的分量相互独立,但分布不必相同).$\{1,2,\ldots,p\}$ 的每个子集 μ 称为子模型,显然,多元线性回归有 2^p 个可能的子模型.如果对于任意 $i\notin\mu$,$\beta_i=0$;则称为子模型 μ 为真实模型.我们解决的问题是要找到最小的真实模型,其中最小的真实模型定义为它所有的真子模型都不是真实模型.

交叉验证是一种模型选择的方法,它能对模型进行预测.它的主要思想:假设有 n 个数据点,为了在模型集中选择一个适当的模型,首先,我们删掉一个数据点,用其余 $n-1$ 个数据点去拟合模型;然后用删掉的数据点检验模型,再对所有数据点都进行一次上面的操作,选择一个具有佳预测水平的模型.

在本文中,我们提出一个模型选择准则,通过在交叉验证中增加惩罚函数,把交叉验证准则与信息理论准则结合.正如我们所知的,交叉验证准则易于过度拟合模型,而相合信息理论准则在样本容量不大时,时常对模型拟合不足.新的模型选择准则把这两者的优势结合起来,从而能更好地拟合模型.

注:A 表示 $p\times p$ 的矩阵,a 为 $p\times 1$ 向量,$A(i)$ 表示 A 的前 i 列组成的矩阵,$A(-i)$ 表示去掉 A 的第 i 列组成的矩阵,$a(i)$ 表示 a 的前 i 个元素组成的向量,$a(-i)$ 表示去掉 a 的第 i 个元素组成的向量.

2 相合模型选择准则

考虑回归模型(1.1),定义:$X_n=(x_{1n},\ldots,x_{pn})=(x_1,\ldots,x_n)'$,$P_i$ 表示由 $x_{1n},\ldots,x_{in}$ 生成空间上的正交投影矩阵.下面给出得到我们主要结果所需的假设条件.

假设 1 存在常数 a_1 和 a_2,使得

$$0<a_1 n\leq\lambda_p(X_n'X_n)\leq\lambda_1(X_n'X_n)\leq a_2 n \tag{2.1}$$

收稿日期:2005－01－12

作者简介:闻 斌(1980－),男,江苏苏州人.

其中:$\lambda_i(X_n'X_n)$表示$(X_n'X_n)$的第 i 个特征值.

假设 2　存在常数 $\delta>0$,使得对每个 i, $1\leq i\leq p$

$$\sum_{j=1}^{n}(X_{in}^{j})^3=O[(x_{in}'x_{in})^{3/2}/(\log(x_{in}'x_{in}))^{1+\delta}] \tag{2.2}$$

其中:x_{in}^{j}表示 $x_{in}=(x_{in}^1,\ldots,x_{in}^n)$的第 i 个元素.

假设 3　$\max_{1\leq i\leq n}X_i'X_i=O(\log\log n)$.

假设 4　$e_n=(e_1,\ldots,e_n)'$的分量相互独立,期望为 0 并且满足矩条件:使得对每个 i, $1\leq i\leq n$, $0<\gamma^2\leq E(e_i^2)$, $E(|e_i|^3\leq\tau^3<\infty$. (2.3)

首先考虑 p 个连续的子模型集$\{M(1),\ldots,M(p)\}$, $M(k)$表示回归模型(1.1) 其中$(\beta=(\beta_1,\cdots,\beta_k\neq 0,\cdots,0)'$. 定义判别函数:$G_n^{cv}(k)=\sum_{j=1}^{n}(y_i-\hat{y}_{-i,k})^2+kC_n$,其中 $\hat{y}_{-i,k}=x_i'(k)\hat{\beta}_{-i}(k)$, $\hat{\beta}_{-i}(k)$表示在模型 $M(k)$ 中 $\beta(k)$基于数据点$\{(y_1,x_1(k)),\ldots,(y_{i-1},x_{i-1}(k)),(y_{i+1},x_{i+1}(k)),\ldots,(y_n,x_n(k))\}$ 的最小二乘估计; C_n 是关于 n 的函数,称为惩罚函数;满足条件:

$$\frac{C_n}{n}\to 0,\ \frac{C_n}{\log\log n}\to\infty. \tag{2.4}$$

我们提出的模型选择准则以判别函数 $G_n^{cv}(k)$ 为基础,选取的模型定义为 $M(\hat{k}_n)$,满足:

$$G_n^{cv}(\hat{k}_n)=\min_{1\leq k\leq p}G_n^{cv}(k). \tag{2.5}$$

下面的定理将证明按上面的方法选取的模型具有强相合性.

定理 2.1　假定对于 $n=1,2,\ldots$,假设 1—4 均成立, $M(k_0)$是最小的真实模型.如果对于充分大的 n, C_n 以概率为 1 满足(2.4),则按(2.5)选取的模型强相合于最小的真实模型;在样本容量充分大时能得到最小的真实模型.

为了证明定理 2.1,我们需要下面两个引理. $S(k)$表示模型 $M(k)$中的残差平方和.

引理 2.1　假定假设 **1**,2,4 成立,则

$(L1)$ $a_2 n\geq x_{in}'x_{in}\geq a_1 n$, $n\to\infty$, $1\leq i\leq p$;

$(L2)$ $a_2 n\geq x_{in}'(I-P_{i-1})x_{in}\geq a_1 n>0$, $1\leq i\leq p$;

$(L3)$ $x_{in}'e_n=O((\log\log n)^{1/2})$, $a\cdot s$, $1\leq i\leq p$

$(L4)$ $e_n'P_ie_n=O(\log\log n)$, $a\cdot s$, $1\leq i\leq p$;

证明见 Bai et al. 1999.

引理 2.2　假定假设 1 和 3 成立,则对每个 k, $1\leq k\leq p$, $\max_{1\leq i\leq n}w_{i,k}=O(\log\log n)$, $\max_{1\leq i\leq n}w_{i,-k}=O(\log\log n)$;其中 $w_{ik}=X_i'(k)(X_n'(k)X_n(k))^{-1}x_i(k)$, $w_{i,-k}=X_i'(-k)(X_n'(-k)X_n(-k))^{-1}x_i(-k)$.

证明　由假设 1 及 3, 引理 2.2 即得.

证明定理 2.1.令 $r_{ik}=y_i-x_i'\hat{\beta}(k)$.

$$\sum_{i=1}^{n}(y_i-\hat{y}_{-i,k})^2=\sum_{i=1}^{n}[(1-w_{ik})^{-1}r_{rk}]^2=\sum_{i=1}^{n}(1+2w_{ik}+O(W_{ik}^2))r_{ik}^2$$

$$=\sum_{i=1}^{n}r_{ik}^2+\sum_{i=1}^{n}(2w_{ik}+O(w_{ik}^2))r_{ik}^2\cdot a\cdot s \tag{2.6}$$

当 n 充分大时,由(2.6)和假设 3,可得

$$S(k)\leq\sum_{i=1}^{n}(y_i-\hat{y}_{-i,k})^2$$

$$\leq(1+3\max_{1\leq i\leq n}w_{ik})S(k),\ a\cdot s \tag{2.7}$$

当 $k<k_0$ 时,由(2.7),引理 2.1 和引理 2.2 对充分大的 n,都以概率为 1 成立.

$$G_n^{cv}(k)-G_n^{cv}(k_0)$$

$$\geq S(k)-(1+3\max_{1\leq i\leq n}w_{ik})S(k_0)+(k_0-k)C_n$$

$$\geq\beta^2(k_0)a_1 n+\beta(k_0)O((\log\log n)^{1/2})+O(\log\log n)-(k_0-k)C_n\quad a\cdot s.$$

由(2.4)条件:$n^{-1}C_n\to 0$, 从而对充分大的 n, $G_n^{cv}(k)-G_n^{cv}(k_0)>0$. $a\cdot s$. 这说明

$$\liminf\ \hat{k}_n k_n\cdot\ a\cdot s \tag{2.8}$$

当 $k>k_0$ 时,由(2.7),引理 2.1($L4$)和引理 2.2 对充分大的 n,都以概率为 1 成立.

$$G_n^{cv}(k)-G_n^{cv}(k_0)$$

$$\geq S(k)-(1+3\max_{1\leq i\leq n}w_{ik})S(k_0)+(k-k_0)$$

$$C_n\geq O(\log\log n)+C_n\cdot\ a\cdot s.$$

由(2.4)条件:$\frac{C_n}{\log\log n}\to\infty$. 从而对充分大的 n, $G_n^{cv}(k)-G_n^{cv}(k_0)>0$. $a\cdot s$. 这说明

$\limsup\ \hat{k}\leq k_0$. $a\cdot s$. 再由(2.8),可得 $\hat{k}_n\to k_0$. $a\cdot s$. 定理证毕.

3　一般情况下相合模型选择准则

考虑一般的情形, β 的每个分量可以取 0 或非 0.对于任何 β,重新组合的 β 分量及设计阵 X_n 的列向量,我们可以得到等价的回归模型:最小的真实模型在连续的子模型集$\{M(1),\ldots,M(p)\}$之中;然后再用上面介绍的模型选择准则,因为重新组合的条件下假设条件不会改变,估计的模型仍强相合于最小的真实模型;在重组模型选择中选取最小的 $\hat{k}$

即可.此方法的缺点是需要计算 2^p 个残差平方和,当 p 很大时,计算量很大.我们使用“删一”法(见 Rao and Wu,1989;Bai et al,1999)来选取最小的真实模型,此方法的优点是只需计算 $p+1$ 个残差平方和.

考虑模型 $\tilde{y}_n = X_n(-k)\beta(-k) + e_n$ (3.1)

记 $\hat{y}_{-i,-k} = x_i'(-k)\hat{\beta}_{-i}(-k)$,其中 $\hat{\beta}_{-i}(-k)$ 表示在模型(3.1)中 $\beta(-k)$ 基于数据点

$\{(y_1, x_1(-k)), \ldots, (y_{i-1}, x_{i-1}(-k)), (y_{i+1}, x_{i+1}(-k)), \ldots, (y_n, x_n(-k))\}$ 的最小二乘估计;

定义 对于 $i, 1 \leq i \leq p$;

$$G_n^{cv}(-k) = \sum_{i=1}^{n}(y_i - \hat{y}_{-i,-k})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_{-i,k})^2 - C_n \quad (3.2)$$

其中 C_n 满足条件(2.4).按如下准则选择模型:

对于 $i=1, \ldots, p$,如果 $G_n^{cv}(-k) \leq 0$,则 $\beta_i = 0$;如果 $G_n^{cv}(-k) > 0$,则 $\beta_i \neq 0$. (3.3)

定理 3.1 在定理 2.1 的条件下,按(3.3)选择的模型强相合于最小的真实模型;在样本容量充分大时能得到最小的真实模型.

证明与定理 2.1 的证明相似.(略)

4 结 论

本文针对利用交叉验证准则易于过度拟合线性模型,通过在交叉验证中增加惩罚函数来解决这一问题;在一定的假设条件下,新准则确定的模型强相合于最小的真实模型,并给出一般条件下模型选择准则.

参考文献:

[1] C·R·Rao. Linear model[M]. Springer-Verlag, 1995.

[2] Z·D·Bai, C·R·Rao, Y·Wu. Model selection with data-oriented penalty[J]. J·Statist·Plann·Inference, 1999, (77): 103—117.

[3] Akaike, H. Imformation theory and extension of the maximum likelihood principle[M]. Budapest, 1773, 267—281.

[4] Petrov. Sum of Independent Radom Variable. Springer-Verlag, 1975, 111—305.

[5] Zheng, X. Loh, W. Y. Consistent variable seletion in linaer models[J]. J·Amer·Statist·Assoc, 1995, (90): 151—156.

Linear Model Selection by Cross-validation

WEN Bin, JIANG Qi-bao

(Department of Mathematics, Southeast University, Nanjing 210096, China)

Abstract: We consider the problem of model selection in the classical regression model based on cross-validation with an add penalty term for penalizing overfitting. Under some weak considers, the new criterion is shown to be strongly consistent in the sense that with probability one, for all large n, the criterion chooses the smallest true model and we extends the criterion to the general case.

Key words: consistency; cross-validation; linear regression; model selection