第16卷 第2期 大 气 科 学 Vol.16, No.2
1992年 3月 SCIENTIA ATMOSPHERICA SINICA Mar., 1992

基于 PRESS 准则选取预报因子的逐步算法

姚棣荣 俞善贤
（杭州大学地理系，310028） （浙江省气象科学研究所，杭州，310021）

提 要

本文提出了基于预测平方和(PRESS)准则选取预报因子的逐步算法．通过实例计算表明，这种算法不仅具有向前算法和向后算法的快速特点，而且同时具有相应的功能，因而它能更有效地寻找最优解，使算法更为完善，更具有普遍性和实用意义．

关键词： 预测平方和(PRESS)准则；预报因子；逐步算法．

一、引 言

在应用回归分析方法处理实际问题时，回归自变量的选择是一个十分重要的问题．目前被广泛采用的逐步回归分析就是一种选取回归自变量的有效方法，但是它存在一定的缺陷，因而近年来，选取回归自变量的各种准则相继出现[1,2]，其中由 Allen[3] 提出的预测平方和(PRESS)准则，引起了人们的重视和注意，这对于建立预测能力较强的回归模型无疑是很有意义的．在实际应用中，如何解决计算量过大，并且实现快速、有效的算法就显得更为重要．对此，俞善贤和沈锦花[4] 提出了向前、向后两个快速算法，使选取回归自变量的PRESS 准则能在实际中得以实现，通过实例计算表明，在 PRESS 准则下建立的回归方程的预测效果比采用逐步回归分析有较明显的提高；但是也发现，这二个快速算法不一定能得到最优解．本文在文献 [4] 的基础上，提出了一个基于 PRESS 准则选取回归自变量的逐步算法，它与逐步回归分析一样，应用了引进和剔除回归自变量的双重检验方法，期望能得到预测能力较强的最优模型．

二、PRESS 准则

设有 p 个自变量(预报因子) $X=(x_1, x_2, \cdots, x_p)'$ 和因变量(预报对象) Y 的 n 次观测值．考虑如下的线性回归模型

$$Y=X\beta+e \ , \tag{1}$$

式中 β, e 分别表示回归系数和残差的向量．

若去掉第 i 个以后试验点(样品)的模型记为

1989年3月28日收到，1990年2月28日收到再改稿

$$Y(i)=X(i)\beta+e(i)\,, \tag{2}$$

式中 $Y(i)$，$X(i)$和 $e(i)$ 分别是从 Y、X 和 e 中删去第 i 行后得到的．由最小二乘估计，可得到 $\hat{\beta}(i)$，则第 i 个试验点 $X_i=(x_{i1}, x_{i2}, \cdots, x_{ip})'$ 处的预测值为 $X_i'\hat{\beta}(i)$．而预测偏差为 $f_i=y_i-X_i'\beta(i)$，对每一个试验点重复上述做法，可得到 f_1，f_2，$\cdots$，f_n．其平方和

$$\mathrm{PRESS}=\sum_{i=1}^{n} f_i^2=\sum_{i=1}^{n}[y_i-X_i'\beta(i)]^2 \tag{3}$$

称为预测平方和．这是全模型的预测平方和，它度量了全模型预测能力的优劣．

对于选模型，

$$Y=X_l\beta_l+e\,, \tag{4}$$

上式表示在回归模型中只包含 l 个因子 $(l<p)$，其预测平方和值为

$$\mathrm{PRESS}(l)=\sum_{i=1}^{n}[y_i-X_{li}'\hat{\beta}_l(i)]^2\,. \tag{5}$$

于是，在预报因子 $x_1, x_2, \cdots, x_p$ 可能产生的一切子集中，以 PRESS 值达到最小的那个子集作为最优回归子集，这就是选取预报因子的 PRESS 准则．

为了方便，PRESS 值的计算可用如下公式[2,3]

$$\mathrm{PRESS}=\hat{e}'D^{-2}\hat{e}\,, \tag{6}$$

式中

$$\hat{e}=Y-X\beta=(I-X(X'X)^{-1}X')Y\,, \tag{7}$$

$$D=\mathrm{diag}(d_1, d_2, \cdots, d_n)\,, \tag{8}$$

其中 d_i 为方阵 $I-X(X'X)^{-1}X'$ 的第 i 个对角元，而 $H=X(X'X)^{-1}X'$ 称为帽子矩阵，它的第 i 个对角元记为 h_{ii}．

为了计算 p 个因子的所有可能子集的 PRESS 值，原算法需要对一个子集进行一次扫描运算，即设 A 为 n 阶方阵，现定义一个新方阵 $B=(b_{ij})_{n\times n}$，则以 a_{kk} 为枢轴的扫描运算为

$$b_{ij}=\begin{cases} 1/a_{kk}\,, & i=k,\ j=k \\ -a_{ik}/a_{kk}\,, & i\neq k,\ j=k \\ a_{kj}/a_{kk}\,, & i=k,\ j\neq k \\ a_{ij}-a_{ik}a_{kj}/a_{kk}\,, & i\neq k,\ j\neq k \end{cases} \tag{9}$$

这样共需 2^p-1 次扫描运算，当 p 很大时，计算量大得惊人而往往无法实现．但使用向前或向后的快速算法[4]，则至多只需施行 $p(p+1)/2$ 次扫描运算即可，大大减少了计算量．

三、逐步算法

为了更有效地寻找上述 PRESS 准则下的最优解，我们认为，采用向前算法和向后算法相结合的双重检验的逐步算法（简称为逐步算法），似乎更为完善．

设自变量 $x_i(i=1,2,\cdots,p)$和因变量 y 是经过中心化处理的变量，沿用前文的符号，但此时的设计矩阵 X 为$(n\times p)$阶矩阵，β 则为 p 维列向量，其余同前．现记 $A=X'X$，它是一个$(p\times p)$阶矩阵．

1．预报因子引入的计算

设已进行了 l 步，已引入了 l 个因子(即为x_{i_1}，x_{i_2}，$\cdots$，x_{i_l})此时的PRESS值为PRESS(l)，而且对 $X'X$ 作了 l 次扫描运算，并已计算了

$$\hat{e}_l=Y-X_l\hat{\beta}_l, \tag{10}$$

和矩阵

$$H_l=X_l(X_l'X_l)^{-1}X_l' \tag{11}$$

的对角元部分．

现在考虑第 i_k 个自变量 X_{i_k} 的引进，这里$i_k\in\{i_{l+1},i_{l+2},\cdots,i_p\}$，按向前算法[4]，则有

$$H_{l+1}=X_l(X_l'X)^{-1}X_l'+\frac{1}{a_{i_ki_k}}(X_lA_{i_k}-X_{i_k})(X_lA_{i_k}-X_{i_k})', \tag{12}$$

$$\hat{e}_{l+1}=\hat{e}_l-\frac{1}{a_{i_ki_k}}(X_lA_{i_k}-X_{i_k})(X_lA_{i_k}-X_{i_k})'Y, \tag{13}$$

式中

$$X_{i_k}=(X_{1i_k},X_{2i_k}\cdots X_{ni_k})',$$

$$A_{i_k}=\begin{pmatrix}a_{i_1i_k}\\a_{i_2i_k}\\\vdots\\a_{i_li_k}\end{pmatrix}, \tag{14}$$

而 $a_{i_ki_k}$ 为A矩阵的第 i_k 个对角元．

由(12)和(13)式可见，第$l+1$步上考虑 X_{i_k} 的引进时，只要在l步基础上，计算

$$\frac{1}{a_{i_ki_k}}(X_lA_{i_k}-X_{i_k})(X_lA_{i_k}-X_{i_k})'Y, \tag{15}$$

和矩阵

$$\frac{1}{a_{i_ki_k}}(X_lA_{i_k}-X_{i_k})(X_lA_{i_k}-X_{i_k}) \tag{16}$$

的对角元部分．这样，若引进 X_{i_k} 后的PRESS值为

$$\mathrm{PRESS}(i_k)=\hat{e}_{l+1}'D_{l+1}^{-2}\hat{e}_{l+1}. \tag{17}$$

利用(17)式，可以计算所有未选因子 X_{i_k}，$i_k\in\{i_{l+1},i_{l+2},\cdots,i_p\}$的PRESS值，从中找出最小的PRESS值，记为PRESS(i_k^*)，如果PRESS(i_k^*)<PRESS(l)，则引进 $X_{i_k}^*$，否则 $X_{i_k}^*$ 不能引进．

2．预报因子的剔除计算

按(10)和(11)式，现在考虑已选因子 X_{i_k} 的剔除问题，$i_k\in\{i_1,i_2,\cdots,i_l\}$．根据

向后算法[4]，则有

$$\boldsymbol{H}_{l-1}=\boldsymbol{H}_l-\frac{1}{h_{i_k i_k}}\boldsymbol{L}\boldsymbol{L}', \tag{18}$$

$$\hat{\boldsymbol{e}}_{l-1}=\hat{\boldsymbol{e}}_l+\frac{\hat{\beta}_{i_k}}{h_{i_k i_k}}\boldsymbol{L}, \tag{19}$$

式中

$$\boldsymbol{L}=\boldsymbol{X}_l(\boldsymbol{X}_l'\boldsymbol{X}_l)^{-1}\boldsymbol{R}, \tag{20}$$

这里 $\boldsymbol{R}$ 为第 i_k 行元素是1，其余元素是0的$(l\times 1)$阶矩阵，所以(20)式即为设计矩阵 $\boldsymbol{X}_l$ 乘 $(\boldsymbol{X}_l'\boldsymbol{X}_l)^{-1}$ 中的第 i_k 列，即 $\boldsymbol{L}$ 为$(n\times 1)$阶矩阵.

$$h_{i_k i_k}=\boldsymbol{R}'(\boldsymbol{X}_l'\boldsymbol{X}_l)^{-1}\boldsymbol{R}, \tag{21}$$

它等于$(\boldsymbol{X}_l'\boldsymbol{X}_l)^{-1}$ 的第 i_k 个对角元. $\hat{\beta}_{i_k}$ 为第 i_k 个因子的回归系数.

可见，在引入了 l 个因子的选模型中，考虑因子 $\boldsymbol{X}_{i_k}$ 的剔除问题，只要在 l 步基础上，计算

$$\frac{\hat{\beta}_{i_k}}{h_{i_k i_k}}\boldsymbol{L} \text{ 和矩阵 } \frac{1}{h_{i_k i_k}}\boldsymbol{L}\boldsymbol{L}'$$

的对角元. 这样，考虑剔除 $\boldsymbol{X}_{i_k}$ 后的PRESS值为

$$\mathrm{PRESS}(i_k)=\hat{\boldsymbol{e}}_{l-1}'\boldsymbol{D}_{l-1}^{-2}\hat{\boldsymbol{e}}_{l-1}. \tag{22}$$

对所有的已选因子 $\boldsymbol{X}_{i_k}$，$i_k\in\{i_1,i_2,\cdots,i_l\}$，按(22)式计算其PRESS值，从中选出最小的PRESS值，记为PRESS(i_k^*)，若 PRESS$(i_k^*)<$PRESS(l)，则剔除 $\boldsymbol{X}_{i_k^*}$，否则，$\boldsymbol{X}_{i_k^*}$ 不能剔除.

综合上述，逐步算法的步骤可以归纳如下：

(1) 计算$(\boldsymbol{X}'\boldsymbol{X})$ 矩阵，取初始的 PRESS 值和残差，即

$$\mathrm{PRESS}=\sum_{i=1}^{n}(y_i-\bar{y})^2, \tag{23}$$

$$\hat{e}_i=y_i-\bar{y},\ i=1,2,\cdots,n. \tag{24}$$

(2) 从未选因子中分别选入一个预报因子，由(12)、(13)和(17)式计算出相应的PRESS值.

(3) 找出选入一个因子后模型的PRESS值的最小者，若这个最小值大于第2步中当前模型的PRESS值，则该因子不能入选，转入第6步；否则，引进该因子，把这个最小的PRESS值作为当前模型的PRESS值，并作一次扫描运算，枢轴为当前选进因子的序号.

(4) 当入选因子数达到3个时，则要考虑因子的剔除问题. 从已选因子中分别剔除一个因子，由(18)、(19)和(22)式计算相应的PRESS值.

(5) 找出剔除一个因子后模型的PRESS值的最小者，若这个最小值大于当前模型的PRESS值，则该因子不能剔除，转入第2步；否则，剔除该因子，把这个最小的PRESS值作为当前模型的PRESS值，并作一次扫描运算，枢轴为当前剔除因子的序号，

转入第2步.

(6)逐步计算结束，把PRESS值最小的模型作为最终模型，并输出PRESS值、回归方程等结果.

四、实例计算和分析

我们选取二个实例进行了计算，为了便于比较，还同时用向前算法、向后算法以及逐步回归进行了计算.

实例1　嘉兴地区晚稻产量预报

样本大小 $n=24$，候选因子数 $p=8$. 四种计算法的计算结果见表1.

表1　四种算法的结果比较

项目	逐步、向后算法	向前算法	逐步回归	
			F=2.97	F=4.35
R	0.9724	0.9773	0.9773	0.9567
s_y	19.3478	18.1308	18.1308	22.7436
PRESS值	12521.8705	12904.9067	／	／
因子数	7	8	8	5

按逐步和向后算法的回归方程为

$$y_s = 603.5126 + 5.8122\,x_1 - 3.4295\,x_3 + 5.2664\,x_4 - 11.8164\,x_5 + 6.1660\,x_6 + 2.1561\,x_7 - 6.0390\,x_8 .$$

按向前算法和F=2.97(相当于显著性水平 $\alpha=0.10$)时逐步回归的回归方程为

$$y_R = 215.7942 + 5.4293\,x_1 + 4.1035\,x_2 - 2.5977\,x_3 + 4.6541\,x_4 - 9.9671\,x_5 + 6.0098\,x_6 + 2.1743\,x_7 - 6.9173\,x_8 .$$

$F=4.35$ ($\alpha=0.05$)时逐步回归的回归方程从略.

由表1可见，逐步算法与向后算法的结果是一致的，但它们与向前算法的结果却不同. 向前算法入选的因子比逐步算法(或向后算法)多了1个，而它与 $F=2.97$ 时逐步回归的结果是一致的，但在逐步回归中入选的因子数目却随F值的改变而改变，且 F 值的控制具有人为性，相比之下，采用PRESS准则来选取预报因子就更为客观；此外，向前算法的复相关系数 R 高于逐步算法(或向后算法)，残差均方差 s_y 前者小于后者，这意味着前者的回归模型合适程度比后者要好，注意到向前算法与 $F=2.97$ 时逐步回归的结果相一致的事实，上述的结论正体现了在逐步回归中以拟合的好坏来选取预报因子的实质，由此可以说明，从拟合的角度来看，逐步回归所得的模型可以认为是最优的. 但从预测的角度来看，逐步算法(或向后算法)的PRESS值却是小于向前算法的，这说明，由逐步算法和向后算法所得的模型是最优解，而逐步算法是向前、向后两种算法的结合，因此它比向前算法或者向后算法更为完善；同时还可以看出选取预报因子的PRESS准则与逐步回归的本质差别，就本例而言，从预测的角度来看，逐步回归所

得的模型就不能认为是最优的.

实例2 金华市6月份降水预报

样本大小 $n=28$，候选因子 $p=4$. 计算结果列入表2中.

表2 四种算法的结果比较

项 目	逐步、向前、向后算法	逐步回归 ($F=2.91$)
R	0.6652	0.6709
s_y	58.6001	58.1959
PRESS 值	104276.7420	/
因子数	2	2

在PRESS准则下的回归方程为

$$y_s=19.2945+3.3487x_2+0.6202x_3.$$

在 $F=2.91$ (相当于 $\alpha=0.10$)时逐步回归确定的方程为

$$y_R=947.4384-1.3878x_1+0.4619x_3.$$

上述方程对1985—1987年三年的试报结果见表3.

表3 三年试报结果(单位：mm)

年份	实况	逐步、向前、向后算法		逐步回归	
		$\hat{y}_s$	$\lvert\hat{e}_s\rvert$	$\hat{y}_R$	$\lvert\hat{e}_R\rvert$
1985	168	215	47	218	50
1986	177	203	26	235	58
1987	303	197	106	183	120
平均	/	/	60	/	76

在这一例子中，PRESS准则的三种算法的结果完全一致，入选的因子数目与 $F=2.91$ 时逐步回归相同(都是2个)，但它们各自的回归子集所包含的因子却不同. 从拟合的角度看，$F=2.91$ 时逐步回归的回归方程优于PRESS准则下的回归方程，但当 $F=4.23$ (相当于 $\alpha=0.05$)时逐步回归的回归方程就不再优于PRESS准则下的回归方程；从预测的角度看，PRESS值最小的应是由 x_2、x_3 组成的回归子集，三年的试报结果也证实了PRESS准则下的回归方程优于逐步回归所得的回归方程.

五、讨 论

通过上述的试验和比较，我们认为：

(1) 本文所提出的基于PRESS准则选取预报因子的逐步算法，与向前和向后两种算法一样，具有快速的优点，而且同时具有向前、向后算法的功能. 对 40×50 的样本，在IBM－PC/AT微机上计算时间一般不超过1分钟. 因而说明这一算法更具有普遍性和实用意义.

(2) 在用向前算法引入一个新的因子后，就考虑剔除多余因子(使模型的PRESS

值变大的因子）是很有必要的，向前算法有时比逐步算法多选几个因子，但 PRESS 值比逐步算法要大，这说明了逐步算法可以更有效地找出最优解，因而这一算法比向前算法和向后算法更为完善．

（3）本文所提出的逐步算法的计算过程与逐步回归有相似之处，但是它们在选取预报因子的着眼点以及在筛选因子的依据方面是完全不同的．逐步算法着眼于预测效果，在筛选因子时不作任何假设检验，是非参数化的；而逐步回归主要考虑拟合好坏，不能完全反映预测效果，且通过 F 检验来决定因子的取舍，方程中含有的因子数目随着 F 值的改变而改变，而 F 值的控制存在着主观性．因此，基于 PRESS 准则选取预报因子的逐步算法比逐步回归要客观．

（4）实例的计算表明，在适当的 F 值下，逐步回归所确定的回归方程的拟合程度一般优于 PRESS 准则下的回归方程，但也有例外；而就预测效果而言，PRESS 准则下的回归子集具有最小的 PRESS 值，均优于逐步回归所得的回归子集，实际的试报效果也证实了这一点，当然其肯定的程度，还需通过大量的实践来验证．另外，作者在计算中还得到过两者结果完全一致的实例．由此我们可以认为，既然基于 PRESS 准则的逐步算法所得的回归方程，比逐步回归所确定的回归方程具有更好的预测能力（不排斥两者相当的情形），所以就建立回归方程的目的为了用于预测这一意义来说，对于 PRESS 准则下的逐步算法应该引起重视，并不断给以改进和发展．

致谢：杭州大学地理系气象专业 88 届毕业生陈敏参加了部分工作，特致谢意．

参　考　文　献

[1] 陈希孺、王松桂，1984，近代实用回归分析，广西人民出版社．

[2] 陈希孺、王松桂，1987，近代回归分析，安徽教育出版社．

[3] Allen，D．M．，1971，Mean square error of prediction as a criterion for selecting variables，*Technometrics*，13，469－475．

[4] Yu Shanxian and Shen Jinhua，1988，Forward and backward algorithms for selecting predictor on the basis of the criterion from prediction sum of squares and their application，*Acta Meteorologica Sinica*，2，83－90．

The Stepwise Algorithm for Selecting Predictors on the Basis of the Criterion from Prediction Sum of Squares

Yao Dirong

(*Department of Geography，Hangzhou University，Hangzhou*，310028)

Yu Zhangxiao

(*Zhejiang Research Institute of Meteorological Science，Hangzhou*，310021)

Abstract

In this paper，the stepwise algorithm for selecting predictors on the basis of the criterion from prediction sum of squares (PRESS) are discussed．The calculation of examples shows that，this algorithm not only is faster than the forward and backward algorithms，but also remains the same function as both the forward and backward algorithms．It is more efficient to find an optimum solution．

Key words：The criterion from prediction sum of squares；Predictor；Stepwise algorithm．