双评分准则逐步回归法

牛保山　　曹鸿兴

（河南商水县气象站，466100）　（中国气象科学研究院）

刘生长

（气象出版社）

提　要

阐明了以双评分准则(CSC)作逐步回归的基本原理，通过实例给出了计算方法与步骤。用此法所建模型能同时报好预报对象的数量和类别。

关键词： 统计天气预报　逐步回归　变量筛选

引　言

在气象、水文、地震等领域中，为了对各自的研究对象作出预报，面对众多的影响因素，常常采用逐步回归的预报方法。逐步回归分析方法是在改进全回归预报方法的基础上所产生的一种统计预报方法。它是从大量预报因子中按一定的衡量指标、根据预报因子对预报量的重要程度，逐次选入回归方程。在此过程中，某些次要因子可能始终未被引入方程，而先前被引入的因子，由于其后某些因子的引入而相形见绌被剔除，这样就可以从所有变量的所有组合中筛选出优化方程。然而应用逐步回归方法时，其预报结果又常因所选 F 检验临界值的不同而产生差异。因此，其结论的选择也就带上了一定的人为性。日本学者赤池(Akaike)在研究自回归模型定阶时提出了 AIC 准则方法，他还建议，从一组可供选择的模型中选择一个最佳模型时，应该选用 AIC 最小的模型。可是用这一准则建立自回归模型，或用它代替 F 检验进行逐步筛选，由于涉及到 Yule-Walker 递推和增广矩阵的扫描变换等问题，舍入量的干扰使计算精度大幅度降低，以致影响到预报精度的提高。我国气象科学工作者在文献[1,2]中提出了双评分准则(CSC)，用以对统计模型的变量进行筛选。由于这种方法计算简便，避免了舍入量的干扰；同时又以均生函数作为基函数，弱化了极值因素的影响。因此，所建预测模型比较稳健，预报精度也较高。实践证明，双评分准则(CSC)的提出，能较好地处理统计预报模型的选型问题，这对统计预报与气候预测来说，无疑是一个贡献。本文首先阐明双评分准则，然后说明用 CSC 作逐步筛选的步骤，最后给出计算实例。

1　基本原理

建立预报因子 $x_1, x_2, \cdots, x_m$ 与预报量 y 之间的回归

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \quad (1)$$

其中 $\beta_i (i=0,1,\cdots,m)$ 为回归系数。ε 为随机误差，假定为零均值白噪声。用最小二乘法求估计 $\hat{y}$，均方根误差为

$$\varepsilon = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{1/2} \quad (2)$$

式中 n 为样本量。

1.1　双评分准则

首先根据实际需要，将 n 个样品的预报

量 y 分成若干级别。如降水量预报中可分为偏多、正常和偏少3级，或分为涝、偏多、正常、偏少和旱5级。也可以根据变化趋势来分级，计算

$$u=\frac{1}{N-1}\sum_{t=2}^{N}|\Delta y_t|$$

式中

$$\Delta y_t=y_t-y_{t-1}\qquad t=1,2,\cdots,n$$

则预报对象可分为3级

y_A：当 $\Delta y_t>u$

y_B：当 $|\Delta y_t|\leqslant u$

y_C：当 $\Delta y_t<-u$

y_A、y_B、y_C 分别表示升、平、降3种趋势。筛选自变量的双评分准则就是要使回归模型的拟合误差，即模型的精评分，以及拟合趋势（或级别），即模型的粗评分，同时达到最优。

计算预报量的均值

$$\bar{y}=\frac{1}{n}\sum_{t=1}^{n}y_t$$

每次都以均值作预报，即气候预报。由

$$Q_y=\sum_{t=1}^{n}(y_t-\bar{y})^2 \qquad (3)$$

计算其总离差平方和。

设方程中引进了 l 个因子，用最小二乘法得的预报方程为：

$$\hat{y}_t=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_lx_l \qquad (4)$$

其残差平方和

$$Q_k=\sum_{t=1}^{n}(y_t-\hat{y}_t)^2 \qquad (5)$$

显然，Q_k 等价于均方根误差 ε，现以

$$S_1=(n-l)R^2=(n-l)(1-Q_k/Q_y)$$
$$=(n-l)R^2 \qquad (6)$$

作为双评分的第一项，即精评分。R^2 为复相关系数。式(6)中 Q_k/Q_y 的含意是，一个好的预报方法，其误差必须比气候预报小，即 $Q_k/Q_y<1$，乘$(n-l)$相当于一个惩罚因子，即要选取因子尽可能少的回归模型。

趋势或级别评分取

$$S_2=2I=2[\sum_{i=1}^{L}\sum_{j=1}^{L}n_{ij}\ln n_{ij}+n\ln n$$
$$-(\sum_{i=1}^{L}n_{i\cdot}\ln n_{i\cdot}+\sum_{j=1}^{L}n_{\cdot j}\ln n_{\cdot j})] \qquad (7)$$

式中 L 为预报分成的级别数，n_{ij} 为 i 类事件与 j 类事件的列联表中的个数，详见表1。其中

$$n_{\cdot j}=\sum_{i=1}^{L}n_{ij}\qquad n_{i\cdot}=\sum_{j=1}^{L}n_{ij}$$

$2I$ 称为最小判别信息统计量。于是双评分准则表示为

$$CSC_k=S_1+S_2=(n-l)R^2+2I \qquad (8)$$

表1 预报对象分级列联表

n_{ij}		预报 Ⅰ类	预报 Ⅱ类	预报 Ⅲ类	行总计 $n_{i\cdot}$
实况	Ⅰ类	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
实况	Ⅱ类	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
实况	Ⅲ类	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
列总计	$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

1.2 逐步回归中变量引进、剔除及终止的原则

变量引进、剔除中总的原则是使预报方程对历次的回报 CSC 值最大。

引进第一个变量，若

$$V_k=\max\{CSC_j\}\qquad j=1,2,\cdots,m \qquad (9)$$

则第 k 个变量即为第一个被引进的因子。然后，从余下的变量中引进第二个因子，这时把引进前后两个回归方程计算得到的 CSC 值进行比较，引进后增大的则引进，否则不引进。

引进3个因子后开始对已引进的因子逐个作剔除计算，以确定这些因子是否由于新因子的引进而相形见绌应予以剔除。

现设已引进了 l 个变量（因子），相应有 CSC_l，刚引进的变量记为 x_{l_0} 设被剔除的变量为 $x_j(j\neq l_0)$，建立没有 x_j 参加的 $l-1$ 个变量的回归，计算相应的 CSC

$$CSC_{l-1}(x_j^-)=[n-(l-1)]R^2+2I_{l-1} \qquad (10)$$

若

$$CSC_{l-1}(x_j^-)\geqslant CSC_l \qquad (11)$$

则剔除该变量 x_j，否则不剔除。然后考虑引

进新的变量。同理，设已引进了 l 个变量，从 $(m-l)$ 个待选变量中挑出一个，记为 x_j，建立 $l+1$ 个变量的回归，计算

$$CSC_{l+1}(x_j^+) = [n-(l+1)]R^2 + 2I_{l+1} \tag{12}$$

同理可得其他待选变量的 CSC_{l+1}，计算

$$V_k = \max_j CSC_{l+1}(x_j^+)$$

$$j = 1,2,\cdots,(m-l) \tag{13}$$

若

$$V_k > CSC_l \tag{14}$$

则引进第 k 个变量。

当引进和剔除都不能使方程的 CSC 值增大时，即停止筛选。此时所得预报方程为在已有备选的 m 个变量中的一个优化方程。

需要说明的是式(11)的原则与以方差为筛选标准的逐步回归不同，但它是合理的，这是由 CSC 的性质所决定的。由式(6)、(10)和(12)可见，精评分中的 l 起着惩罚因子的作用，即要求回归方程中的变量尽可能少，以体现节省原理。剔除一个变量后，根据回归理论，复相关系数就会减小，但若减小不多，就会被 $[n-(l-1)]$ 的增大而抵消，使得式(11)有可能成立。此外，根据式(7)知，CSC 中的粗评分项 S_2 并非是变量个数的单调函数，也就是当剔除一个变量后有可能使 $2I_{l-1}$ 和 $2I_l$ 增大。总之，剔除标准式(11)与引进标准式(14)只差一个等号，既使得回归方程的拟合达到更高水平，又体现使回归方程中变量尽可能少的节省原理。

2 计算实例

用 CSC 逐步回归法作河南省商水县1991年4月份降水量预报。选取上一年：7月平均最低气温 x_1，7月平均水汽压 x_2，5月 NW 风频率 x_3，4月 SW 风平均风速 x_4，6月 NNW 风平均风速 x_5，9月 NNW 风平均风速 x_6，7月份西太平洋副热带高压西伸脊点经度位置 x_7 等7个因子作为备选因子。以1960—1990年31年资料作样本。

根据气象上规定，将降水量小于历年平均值70%者定为偏少，大于历年平均值130%者定为偏多，其它定为正常。这样商水县降水量少于39.3mm 即为偏少，多于73.0mm 为偏多，其它情况为正常。

根据前述引入因子的原则，试引入第1个因子。将引入第1个因子的各回归模型的 CSC 值进行比较，$\max\{CSC_j^{(1)}\} = CSC_7^{(1)} = 18.2$，故引入 x_7。

试引入第2个因子。将引入第2个因子后的各回归模型的 CSC 值进行比较，经比较后 $\max\{CSC^{(2)}\} = CSC_3^{(2)} = 30.7$，故引入 x_3。

引入的第3个因子是 x_4。因为引入第3个因子后，各回归方程的 CSC 值经过比较，有

$$\max\{CSC^{(3)}\} = CSC_4^{(3)} = 37.2$$

为最大者。

在引进了3个因子后，就应考虑因子的剔除。剔除已引入的因子 x_7 后，有

$$CSC_{72}^{(3)} = 10.9 < CSC_4^{(3)} = 37.2,$$

剔除 x_3 后

$$CSC_{32}^{(3)} = 19.9 < 37.2$$

故不能剔除已引进的因子。

再继续引进因子。试引进第4个因子，由于

$$\max_j CSC_j^{(4)} = 22.9 < 37.2$$

故不能引入第4个因子。

引入3个因子后，既不能剔除因子，又不能再引进因子，故停止筛选。各步计算结果见表2。

通过筛选，在7个备选因子中选得 x_7，x_3 和 x_4，相应的回归方程为

$$\hat{y} = 1.16x_7 + 4.60x_3 - 1.50x_4 - 109.7$$

用此模型预报1991年4月商水县降水量为36.7mm，偏少。实况为37.4mm，偏少。预报完全正确。

该例若用一般的逐步回归分析，选取置信水平 $\alpha=0.05$，经筛选变量 x_3 和 x_7 进入，其回归方程为

$$\hat{y} = 4.67x_3 + 1.2x_7 - 120.6$$

预报1991年4月降水量为36.6mm。从数值结

果上看，两种方法基本一样，这佐证了 CSC 逐步回归的可行性。

我们还用双评分准则逐步回归法试作了1991年春季3—5月降水量预报，预报为138mm，属正常级，实况为131.2mm，正常级。另外还运用此法试作了1991年3月降水量预报，回归预报值为52.7mm，偏少级，实况降水量为92.9mm，偏少。预报正确。

表2 各步计算结果

步骤	方程所含变量				
	引入 $CSC^{(1)}$	引入 $CSC^{(2)}$	引入 $CSC^{(3)}$	引入 $CSC^{(4)}$	引入 $CSC^{(5)}$
x_1	$9.9, x_1$	$21.0, x_1, x_7$	$34.5, x_1, x_7, x_3$		$10.9, x_1, x_7, x_3, x_4$（x_1不能进入）
x_2	$14.0, x_2$	$24.4, x_2, x_7$	$32.4, x_2, x_7, x_3$		$10.9, x_2, x_7, x_3, x_4$（x_2不能进入）
x_3	$13.1, x_3$	$30.7, x_3, x_7$（x_3进入）		$19.9, x_7, x_4$（x_3不能剔除）	
x_4	$7.2, x_4$	$19.9, x_4, x_7$	$37.2, x_4, x_7, x_3$（x_4进入）		
x_5	$15.7, x_5$	$14.7, x_5, x_7$	$28.4, x_5, x_7, x_3$		$21.3, x_5, x_7, x_3, x_4$（x_5不能进入）
x_6	$8.3, x_6$	$21.3, x_6, x_7$	$35.8, x_6, x_7, x_3$		$22.9, x_6, x_7, x_3, x_4$（x_6不能进入）
x_7	$18.2, x_7$（x_7进入）			$10.9, x_3, x_4$（x_7不能剔除）	

3 结语

以方差贡献为衡量标准的一般的逐步回归，由于能通过不大的计算获得一个较优的回归，应用广泛；但理论上并不能证明通过该逐步筛选能得到最优的回归方程。当然，所有可能子集回归能获得最优回归，但计算量大。因此人们试图改进逐步筛选，本文就是这类尝试之一。

对方差贡献作 F-检验来进行逐步回归，事先必须给定置信水平 α，而给定 α 带有较大的人为性。70年代发展了 AIC 准则和 BIC 准则[3]，它的基本思想是在残差平方和与变量个数间进行权衡，当两个模型的残差平方和相同时，取变量少的模型[4]。我们提出的双评分准则，旨在使预报对象的数量拟合和类别拟合同时达到最优，因此用双评分准则作逐步筛选，就有可能使所建模型不但在定量上报好，而且在定性上也报好。这在实际问题中是十分重要的，因为譬如作下年的汛期降水预报，如果旱、涝趋势报反了，就会导致严重的预报失败。

总之，运用双评分准则进行逐步回归，不但免除确定置信水平的人为性，而且能兼顾数量和类别的预报。

参考文献

1 曹鸿兴等. 统计模型选择的双评分准则及其在气象、水文预报中的应用. 数理统计与应用概率，1989，4(1)：5—10.

2 魏凤英、曹鸿兴. 长期预测的数学模型及其应用. 北京：气象出版社，1990：29—36.

3 Akaike, H. . A new look at the statistical identification model. IEEE Trans. Auto. Control, 1974, 19: 716—723.

4 刘生长. 以最小信息准则(AIC)代替 F-检验的逐步回归分析方法. 新疆气象，1988，第5期.

A Method of Stepwise Regression with Couple Score Criterion

Niu Baoshan

(Shangshui County Meteorological Station, Henan Province, 466100)

Cao Hongxing (Chinese Academy of Meteorological Sciences)　Liu Shengchang (China Meteorological Press)

Abstract

Fundamentals of stepwise regression with the couple score criterion (CSC) is described, meanwhile, computational procedure and its details are exampled. A model built by this method is capable of predicting both quantity and category of a predictand.

Key Words: statistical weather forecasting　stepwise regression　variable screening