短　论

用预报残差最小的逐步回归方法作黄河上游旱涝预测试验①

黄嘉佑
(北京大学地球物理系)
王云璋
(黄河水利委员会)

提　要

本文提出用预报残差最小的逐步回归方法作预报，较之传统的逐步回归具有计算简便、适应于小型计算机计算等优点。在作黄河上游旱涝预测试验中表明有较好的效果。

关键词　残差逐步回归　逐步回归　降水量预报

1 前　言

黄河流域上游(兰州以上地区)夏季(6—9 月)降水量的预报，对高原地区、大西北水资源开发以及黄河中、下游未来水情的预测具有十分重要的意义。目前在降水量长期预报中所使用的方法大多用统计方法，其中尤以逐步回归方法最为常用。但是这一方法在水文气象预报的应用中存在两个主要的缺点:

(1) 在传统的逐步回归中，因子的筛选是用一种所谓的双重检验准则实现的。因子每一步引入或剔除是以因子的方差贡献大小来判别。其贡献效果是使引入到回归方程中的所有因子的组合对预报量有较好的拟合。这一过程又称为最小二乘法，即预报量的观测值与回归方程的估计值—残差平方和最小。这一过程的缺点是它仅考虑预报量资料的拟合，对于用某些因子来描述预报量来说，这也许是表达较好的统计模型。但是，对水文气象来说，其主要的目的是预测，因此，用这一方法所建立的统计模型在未来的预报上并不是最好的。

(2) 在传统的逐步回归中，计算过程的开始往往从待选因子的相关阵出发，在计算机的计算过程中相关阵的元素将占据相当的单元。就水文气象中降水量预报而言，把前期的有关物理量作为因子的话，数量是可观的。以某一等压面半球范围内某气象要素为因子场的话，网格点格距选 10×10 的经纬度，由各网格点组成的因子群可达 500 个以上，也即它们将占据 $500\times500=250\,000$ 个机器单元，对于容量较少的小型计算机来说，计算的实现将是十分困难的。

①1989年6月26日收到，11月20日收到修改稿。

本文提出的预报残差最小的逐步回归方法主要是针对上述两个缺点的一种克服方案，该方法具有计算简便，适合于目前广大台站使用，试验表明它具有较好的预报效果。

2 方 法

本文提出的逐步回归方法以每步只引入一个因子为主要计算方案。该方案可以避免计算一开始就需要占据机器较多的单元。设要预报的降水量为 y，引入的因子为 x，则第 i 个样品的拟合残差可表示为

$$e_i = y_i - \hat{y}_i \tag{1}$$

其中$\hat{y}$为用因子作的回归估计[1]。用样本(样本容量为 n)中去掉第 i(i=1，2，…，n)个样品后的样本(即容量为 n−1 的样本)建立回归方程，作第 i 个样品的预报，此时该样品预报量观测值与预报估计值之差称为预报残差，此残差可用因子和预报量的实测值计算得到，即表示为[2]

$$e_{(i)} = e_i / (1 - h_{ii}) \tag{2}$$

其中h_{ii}为帽子矩阵中的元素(见文献[2])。易证残差平方和为

$$PRESS = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\left(1 - \frac{x_i^2}{n}\right)} \tag{3}$$

计算待选因子的残差平方和，取其最小者所对应的因子作为可选入的因子。下次选入的因子以上次残差为预报对象，如此逐步进行(参见文献[1])。选入因子个数可用相关系数绝对值大小控制，或人为控制。最后建立以常规的多元回归方法对选入因子建立回归方程。

3 实例分析

本文使用这一方法对黄河上游地区旱涝预测进行试验。取兰州以上流域 6—9 月降水量为预报量。选取 6 个因子，它们分别是 x_1：当年 1 月欧亚纬向环流指数；x_2：上年 10 月欧亚经向环流指数；x_3：当年 1 月西太平洋副高面积指数；x_4：当年 1 月西太平洋副高平均脊线位置；x_5：当年 1 月东亚槽位置(经度)；x_6：当年 1 月极涡位置(经度)。选取 1953—1982 年期间的资料作为依赖样本(即用该样本建立统计预报模型)。又取 1983—1988 年期间的资料作为独立样本(即使用该样本作为试报检查用)。检验预报效果的优劣用独立样本中的均方误来衡量。

表 1 给出了以预报残差平方和最小为原则的逐步挑选因子的情况。表中第二行列出每步所选入的因子，第三行给出了因子与每步的预报量的相关系数。从相关系数变化可见，随着因子的选入，其绝对值逐步减小，到了第 5 步，相关系数突然下降，下降幅度达 0.181，而且其绝对值达到小于 5%的显著水平。同时，每步的预报残差平方和值(PRESS)由上升转为下降，在第四步达到极小值。因此选入 4 个因子是合适的，并可建立相应的回归方程。用它在独立样本中进行试报，结果均方误为 58.6。与传统的逐步回归比较有所改进。表 2 给出了两种方法在选入 2—5 个因子的试报结果的比较情况。

由于每步所选入因子相同，它们在独立样本中试报 6 年所得均方误在选入 2 和 3 个因子时是一样的，但在选入 4 个因子时，由于两方法所根据挑选因子的原则毕竟不一样，因而出现不同的结果。

表 1　因子的逐步选择

Table 1　The factors selected with stepwise.

步　数	1	2	3	4	5
因　子	x_3	x_2	x_5	x_4	x_1
相关系数	0.526	−0.453	0.376	0.357	0.176
PRESS	22.7	31.2	33.6	32.3	35.5

表 2　两种方法均方误比较

Table 2　The comparison of the root-mean-square errors in two methods.

选入因子	2	3	4	5
残　回	52.5	61.2	58.6	57.2
逐　回	52.5	61.2	61.5	57.2

注：残回即预报残差逐步回归；逐回即传统的逐步回归。

从比较中可见，残回方法不仅计算简便，在因子选择上具有与逐回相同的效果，在某些情况下还能在预报中取得较好的效果。例如，取 1953—1987 年期间资料作依赖样本，用两方法对 1988 年作试报，结果残回选取的因子仍为 x_2，x_3，x_5 和 x_4，而逐回则选 x_2，x_3，x_6 及 x_1。残回预报 1988 年降水量距平为−12.39，而逐回则为 10.71，结果是残回预报正确(实况为−56.31)，而且残回所选的因子较为稳定。

以预报残差平方和最小为原则的逐步回归的预报效果随样本容量不同而不同。在长期预报中，一般希望预报模式要稳定少变，才能有较好的预报效果[3]。我们分别用 1983—1988 年期间每年作为一个独立样本，以它们以前的 15，20，25 及 30 年分别作为依赖样本。把各年的试报综合，再求其均方误。结果发现样本容量为 20 及 30 时，所得的均方误较小，分别为 25.6 及 27.7，而对应于容量为 15 及 25 时，分别为 38.2 及 27.9。实验表明以最接近依赖样本的一年作为独立样本有较好的预报效果。实验还表明样本容量为 20—30 时有较好的效果且预报方程也比较稳定，这与国内外某些试验结果是一致的[4—5]。表 3 给出了样本容量为 20 时，对 1983—1988 年逐年预报时预报方程中回归系数、预报量观测距平值(y_d)、预报距平值($\hat{y}_d$)及误差(e)的情况。预报的稳定性表现在挑选因子中，尽管在每次滑动预报的样本中在第三、四步的因子选入次序上略有出入外，全部选入方程的四个因子均是一致的，它们分别为 x_5，x_2，x_3 及 x_1。相应方程中的回归系数的值列在表 3 中(表 3 中 b_5，b_2，b_3 及 b_1 列)。系数的变化性可用 6 个试报值的标准差来反映(相应值列在表中最下一行)。从表中数字的变化可见，变化的幅度是很小的，其中最小的是 b_1，标准差仅为 0.03；最大的为 b_3。但从逐年预报的回归系数来看，仅 1983 年的 b_3 突出地异于其它年份，其余各年均相当稳定。另外，从各年预报误差来看，也是比较稳定的，除 1983 年误差较大外，其余各年均有较好的效果，6 年中除 1987 年距平符号报错外，其余各年均预报(符号)正确。用最后一个回归方程作 1989 年预报，预报为正距平，与实况符合。

表 3 样本容量为 20 时的预报

Table 3 The forecasting taking the sample size of 20.

年 份	b_5	b_2	b_3	b_1	y_d	$\hat{y}_d$	e
1983	2.91	−4.20	3.06	0.82	17.2	93.0	−75.8
1984	3.40	−4.25	1.42	0.86	13.8	3.8	10.0
1985	3.11	−4.39	1.53	0.85	3.1	35.0	−31.9
1986	3.14	−3.38	1.07	0.82	−37.1	−55.2	18.1
1987	3.13	−3.35	1.62	0.80	−61.1	3.1	−64.2
1988	2.78	−3.63	1.11	0.78	−51.9	−39.1	−12.8
平 均	3.08	−3.87	1.64	0.82			−26.1
标准差	0.20	0.43	0.67	0.03			35.1

参 考 文 献

〔1〕 黄嘉佑，气象统计预报讲义，北京大学地球物理系，1979年.

〔2〕 Montgomery, D. C. and E. A. Peck, Introduction to Linear Regression Analysis, John Wiley and Sons, 1982, P.504.

〔3〕 Bhalme, H. N., S. K. Jodhav, D. A. Mooley and BH. V. R. Murty, Forecasting of monsoon performance over India., *J. Climatology*, 1986, **6**, P. 347—354.

〔4〕 向元珍，长期天气预报中逐步回归应用的讨论，气象，1986年，第6期，16—18页.

〔5〕 Hastenrath, S., Prediction of Indian monsoon rainfall: Further exploration, *J. Climate*, 1988, **1**, P. 298—304.

THE STEPWISE REGRESSION METHOD WITH MINIMUM OF FORECAST ERROR: AN EXPERIMENT FOR DROUGHT AND FLOOD FORECASTING IN THE UPPER BASIN OF YELLOW RIVER

Huang Jiayou

(*Department of Geophysics, Beijing University*)

Wang Yunzhang

(*Irrigation Works Committee of Yellow River*)

Abstract

The new stepwise regression method with minimum of forecast error is proposed in this paper. It is more simple, convenient, and easier used in microcomputer than the traditional stepwise regression. The results of forecasting precipitation in the area of the upper basin of Yellow River show that it is more efficient than the traditional method.

Key words: Stepwise regression with predictable error; Stepwise regression; Forecasting for precipitation