

逐步回归双重分析

李邦宪

(浙江金华市气象台)

提 要

本文运用逐步回归方法同时进行因子筛选和周期分析，建立预报方程。既考虑了前期因子的支配作用，又兼顾了预报对象本身的周期变化的影响。其效果明显优于周期分析和多元回归分析，预报效果较为稳定。

一、问题的提出

随着电子计算机的广泛应用，数理统计已成为长期预报的有力工具。统计预报方法目前主要有两大类，一类是利用前期预报因子和后期预报对象建立统计模式，通常称之为多元分析方法，如逐步回归、判别分析等。它假设气象要素的变化仅受各种前期因子所支配，其它影响可以忽略不计。另一类是通过对气象要素本身随时间演变规律的分析作出预报，通常称之为时间序列分析方法。它假设气象要素仅是由几个确定周期迭加而成的复杂波动，其变化取决于它自身的演变规律。

事实上，某种气象要素的变化不仅受前期因子（如大气环流、海温等）的支配，同时又有其自身的演变规律，只是在不同时间、不同条件下所起的作用不同而已。通常认为，气象要素序列的组成应包括趋势项、周期项和随机项，即

$$y(t) = f_1(x) + f_2(x) + \varepsilon$$

其中趋势项 $f_1(x)$ 体现了前期因子的支配作用，周期项 $f_2(x)$ 体现了要素本身的演变规律，随机项则是除此之外的因素所引起的随机误差。多元分析和时间序列分析方法都只考虑了问题的一个方面。李道松、潘仰耕等人针对两类统计方法所存在的缺陷，提出了一种回归分析与周期迭加相结合的预报方法。他们先用前期因子建立多元线性回归方程，然后对其残差序列用方差分析周期迭

加拟合，取得了一定的效果。但问题在于，用多元分析建立回归方程时，并没有把周期波从原序列中分离出来，若对回归方程拟合值进行周期分析，仍可找出显著周期。这样的方程拟合效果也许较好，但常可导致预报效果不稳定。而对残差进行周期分析，一是因其残差数值较小，实际上只起到对回归方程的订正作用；二是用残差序列分析的周期，无论在理论上和实践中都难以证明它就是原序列的主要周期。实际上前期因子和气象要素本身的演变规律同时支配着气象要素的变化，因此，本文用逐步回归方法同时进行因子筛选和周期分析，并建立预报方程，取得了较好的效果。

二、统计模型的建立

如前所述，气象要素的变化既受前期因子的支配，同时又存在自身的演变规律，因此，本文将统计模型取为：

$$y(t) = \sum_{i=1}^{k_1} a_i f_{1i}(t) + \sum_{i=1}^{k_2} b_i f_{2i}(t) + \varepsilon(t)$$

其中 $y(t)$ 为预报对象序列， $f_{1i}(t)$ 取为 $y(t)$ 的周期序列， $f_{2i}(t)$ 为前期因子资料， a_i 、 b_i 为权重系数， $\varepsilon(t)$ 为随机误差。本文将 $\varepsilon(t)$ 作为预报误差处理，即 $\varepsilon(t) = y(t) - \hat{y}(t)$ 。在实际工作中我们总是使 $|\varepsilon(t)|$ 尽可能小。

在上述模型中，关键是解决两个问题，一是 $y(t)$ 的周期如何提取，二是权重系数 a_i 、 b_i 如何确定。

魏凤英等人在1983年提出了逐步回归周期分析方法，统计模型为 $y(t) = \sum_{i=1}^k a_i f_i(t) + e(t)$ ，周期函数 $f_i(t)$ 取为原序列 $y(t)$ 的分组平均值序列，用逐步回归的方法确定权重系数 a_i ，从而提取隐含周期。而前期因子的筛选和 b_i 的估计也可用逐步回归来处理。根据这样的思路，我们自然就可以用逐步回归方法进行因子筛选和周期分析，并确定各自的权重系数 a_i 和 b_i ，从而建立起预报方程。我们将该方法称之为逐步回归双重分析。

三、计算方法

第一步：计算试验周期序列

将预报对象序列 $y(t)$ 依次按长度 l ($2 \leq l \leq m$) 进行分组：

$$\begin{aligned} &y(1), \dots, y(i), \dots, y(l) \\ &y(1+l), \dots, y(i+l), \dots, y(2l) \\ &\dots\dots\dots \\ &y[1+(n_0-1)], \dots, \\ &y[i+(n_0-1)], \dots, y(n) \end{aligned}$$

其中 n 为原序列长度， n_0 为满足 $i+(n_0-1)l \leq n$ 的最大整数， m 为不大于 $\frac{n}{2}$ 的最大整数。

对以上各组求平均，则得到一个长度为 l 的平均值序列，在此称为试验周期序列。按不同长度分组可得到 $m-1$ 个试验周期序列。将各序列按其周期外延，使其序列长度均为 n ，并将这 $m-1$ 个新序列视为因子 x_1, x_2, \dots, x_{m-1} 。

第二步：粗选前期因子

取一定的信度 α ，对大量的前期因子进行相关普查，选取相关系数 $r \geq r_\alpha$ 的 k 个前期因子，记为 $x_m, x_{m+1}, \dots, x_{m+k-1}$ ，并将原序列 $y(t)$ 视为因子 x_{m+k} 。

经过粗选，可大大减少逐步回归的计算量。如前期因子已经人工挑选，此步工作可省略。

第三步：逐步回归筛选因子和提取周

期

按照逐步回归计算步骤，在给定的 F 检验水平下，对 $x_1, x_2, \dots, x_{m+k-1}$ 进行变量的逐个引入或剔除，直到既无变量可剔除又无变量可引入为止，同时记下被选变量的序号 i 。然后计算被选入变量的回归系数，建立预报方程。显然，序号 $i < m$ 的变量就是原序列 $y(t)$ 所含的主要周期；序号 $i \geq m$ 的变量为前期因子。

四、实例计算分析

1. 资料选取

前期因子取 1956—1983 年各月北太平洋 37 个格点海温，500hPa 32 个环流特征量及 14 种本站要素资料，预报对象为金华市 5 月降水量。

2. 相关普查

为能较早地发布预报，1 月份资料取当年相关，其余各月资料取隔年相关， $r_\alpha = 0.4$ 。经上机普查，选取了 22 个前期因子。

3. 建立预报方程

由计算机自动分组计算各试验周期序列，与 22 个前期因子一起进行逐步回归双重分析计算。取 $F = 5$ 时，选入 6 个变量，方程复相关系数达 0.97。预报方程为：

$$\begin{aligned} y = &-345.4 + 0.777x_9 + 0.635x_{10} \\ &+ 0.597x_{11} - 6.90x_{25} \\ &+ 216.4x_{27} + 536.4x_{31} \end{aligned}$$

其中前三个变量分别为 9 年、10 年、11 年周期， x_{25} 为上年 2 月 D 指数， x_{27} 为上年 3 月亚欧地区经向环流指数， x_{31} 为上年 7 月亚欧地区经向环流指数。

五、与多元回归、周期分析的比较

取同样的前期因子资料用逐步回归计算，取 $F = 2.3$ 时，选入 4 个因子，复相关系数为 0.82，回归方程为：

$$\begin{aligned} y = &-13596.9 - 231.6x_1 + 700.5x_2 \\ &- 36.4x_3 + 14.15x_4 \end{aligned}$$

其中 x_1 为上年 7 月亚欧地区纬向环流

指数, x_2 为上年 7 月亚欧地区经向环流指数, x_3 为上年 4 月第 16 点海温, x_4 为 1 月上旬金华平均气压。

为使提取的周期更稳定, 取 1953—1983 年资料进行逐步回归周期分析。取 $F=8$ 时, 选入 5 年、7 年、10 年和 15 年 4 个周期, 复相关系数为 0.94, 预报方程为

$$y = -202.6 - 0.983x_5 + 0.803x_7$$

附表 逐步回归双重分析、逐步回归、逐步回归周期分析效果比较

年份	实况	双重分析	逐步归	周期分析	年份	实况	双重分析	逐步归	周期分析
1957	163	132	175	206	1972	153	144	197	92
1958	440	434	415	458	1973	516	472	401	500
1959	200	191	185	180	1974	218	219	228	244
1960	175	173	202	168	1975	215	192	216	209
1961	286	289	259	247	1976	120	90	85	148
1962	262	256	189	220	1977	370	388	293	426
1963	229	209	203	221	1978	96	138	196	123
1964	218	234	208	219	1979	174	176	147	154
1965	159	143	139	166	1980	155	211	150	168
1966	123	122	170	114	1981	137	126	171	148
1967	362	347	256	279	1982	82	116	114	160
1968	188	196	140	172	1983	197	237	310	171
1969	223	208	246	255	1984*	169	164	152	412
1970	232	194	283	254	1985*	125	220	256	196
1971	273	328	387	268	1986*	140	154	203	235

* 1984、1985、1986 三年为预报值。

六、结论

1. 逐步回归双重分析运用逐步回归技术同时进行因子筛选和隐含周期提取, 并给予不同的权重系数, 既考虑到前期因子的支配作用, 又兼顾了周期性变化影响, 有利于减少预报误差, 提高预报精度。

2. 从目前所做的工作来看, 逐步回归双

$$+ 1.126x_{10} + 0.95x_{15}$$

将以上两种预报方法的拟合和试报效果同列在附表。从表中可以看到, 逐步回归双重分析的拟合效果和试报效果都明显优于其它两种方法。按现行评分办法, 三年试报平均得分为: 逐步回归双重分析 90 分, 逐步回归 80 分, 逐步回归周期分析为 47 分。

重分析的预报效果明显优于周期分析和多元回归, 预报效果较为稳定。

3. 在给定的 F 检验水平下, 有时选入方程的变量全是前期因子, 这正是逐步回归双重分析的独到之处。当预报对象序列周期性变化不明显时, 它能自动舍弃周期项, 无需人为干预。

(参考文献略)