

预报因子残差相关选择法

姚益平

(浙江省金华市气象台)

提 要

本文从残差角度考虑预报因子，利用残差相关分析组织待选因子集。为在众多因子中客观选择配合较好的预报因子提供了一种有效方法。

一、前言

在建立回归方程的过程中，一个极其重要的问题，就是如何在为数众多的因素中选择变量，以建立最优回归方程。逐步回归是目前较常用的方法之一。但对于大样本资料，由于受计算速度和内存的限制，在微机上即使进行逐步回归计算，仍有不便和困难。因此，在建立MOS预报和作物产量预报等统计预报模式过程中，常采用因子分批相关普查，以相关系数高低为原则，初步筛选因子，然后利用逐步回归，在待选因子中进一步挑选因子，建立预报方程。由于用相关普查初步筛选的因子，其因子之间的相关关系往往较密切，而且有大量的“落选”因子未能参加逐步回归，由此建立的预报方程往往不是最优回归方程。

本文提出“预报因子残差相关选择法”，目的是弥补上述不足。经实际使用，效果较理想。

二、思路与方法

“残差相关选择法”的基本思路是：在通过相关普查初步筛选因子的基础上，选择对预报对象贡献最大的少量因子，初步建立方

程，以其残差代替预报对象，重新进行相关普查（称残差相关分析），把对残差贡献大的因子加入待选因子集，建立最后回归方程。具体过程如下：

记预报对象为 y ；设有 m 个因子，写成集合形式：

$$X = \{x_i | i = 1, 2, \dots, m\}$$

分别计算 y 与每个因子的相关系数，确定临界相关系数 r_0 ，由相关系数大于 r_0 的因子组成待选因子集 X_0 。利用 X_0 进行逐步回归，建立方程：

$$\hat{y}_0 = b_{00} + b_{01}x_{01} + b_{02}x_{02} + \dots + b_{0m}x_{0m} \quad (1)$$

令 $\Delta y_1 = \hat{y}_0 - y$ ，计算 Δy_1 与 X 的相关系数（称残差相关系数），重新确定 r_0 ，得残差相关系数大于 r_0 的因子集 X_1 。合并 X_0 与 X_1 ，得到新的待选因子集 $X_0 \cup X_1$ 。因子的原始相关系数较大，其对应的残差相关系数一般较小，即 X_0 与 X_1 相交为空集。因此，不必计算 X_0 集因子的残差相关系数，在残差相关分析时，可将 X_0 集因子从 X 中剔除。

利用 $X_0 \cup X_1$ 进行逐步回归，重新建立预报方程：

$$\hat{y}_1 = b_{10} + b_{11}x_{11} + b_{12}x_{12} + \dots +$$

$$b_{1m1}x_{1m1} \quad (2)$$

在建立(1)式的逐步回归计算过程中，通过提高 F 临界值，尽量减少入选因子数(m_0 一般控制在1或2)。在重新选择预报因子时，可另行确定适当的 F 临界值。由于 X_1 中的因子与残差(Δy)的相关系数最高，它们与方程(1)中的因子的配合往往优于 X_0 中的其它因子而入选方程，即方程(2)中含有 X_1 中的因子，不属于 X_0 ，从而在一定程度上防止因初选因子不当而失去更佳的因子组合，得到更优的回归方程。

根据需，要可连续进行多次残差相关分析。进行第 k 次分析时，令残差 $\Delta y_k = \hat{y}_{k-1} - y$ ，计算 Δy_k 与 X 中除待选因子外的其余因子的残差相关系数，从中选得残差相关系数较高的因子集 X_k 。利用待选因子集 $X_0 \cup X_1 \cup X_2 \dots \cup X_k$ 进行逐步回归，最后建立预报方程：

表 1 待选因子(X_0)

时段(旬/月)	要素		蒸发量		相对湿度		日照时数			降水量		
			下/4—下/5		上/5—上/6		下/9—中/10		下/12—上/1		上/5—下/5	
	i	1	2	3	4	5	6	7	8			
原始相关系数		0.8663	0.7802	-0.6688	-0.6806	-0.5751	0.5903	0.7107		-0.6379		
残差相关系数		-0.0002	-0.1404	-0.0218	0.0001	0.1532	-0.1552	0.0388		0.0043		

注：表中8—12月为上年时间，表2同。

分批求出 Δy_0 与其余997个因子的残差相关系数， X_1 由残差相关系数最大的4个因子组成(见表2)。 X_0 与 X_1 合并($X_0 \cup X_1$)共有12个因子。对 $X_0 \cup X_1$ 进行逐步回归，估计方程中取3—4个变量， $F_{0.01}(3, 21) = 4.87$ 作为 F 检验水平，最后建立预报方程：

$$y_1 = 16.0052 + 0.5337x_{01} - 0.4528x_{04} + 0.1293x_{14} \quad (5)$$

X_1 中残差相关系数最高的 x_{14} 入选方程。

若按传统方法，仅用 X_0 集因子进行逐步回归， x_{01} 和 x_{04} 入选后，将 F 临界值降低

$$\hat{y}_k = b_{k0} + b_{k1}x_{k1} + b_{k2}x_{k2} + \dots + b_{kmk}x_{mk} \quad (3)$$

三、应用实例

预报对象(y)为金华早稻气象产量(趋势产量用步长为8年的调和权重法拟合)，资料年代为1963—1987年；供普查的因子为前一年8月至当年7月间逐旬平均气温、平均相对湿度、降水量、日照时数、蒸发量，以及通过逐2至6旬滑动相加所产生的膨化因子^[1]，共计因子1005个，即 $m=1005$ 。

分批计算 y 与各因子的原始相关系数，临界相关系数 r_0 取值0.55，剔除同一要素时段重叠的部分因子，初步筛选因子8个(见表1)。利用该8个因子(即 X_0)进行逐步回归， F 临界值取15，得：

$$\hat{y}_0 = -0.9645 + 0.5643x_{01} - 0.3864x_{04} \quad (4)$$

表 2 新增加的待选因子(X_1)

时段(旬/月)	要素		蒸发量		相对湿度		日照时数			
			中/3—下/3		中/3—下/3		中/11—下/11		中/3—下/3	
	i	1	2	3	4	5	6	7	8	
原始相关系数		0.1119	-0.1305	0.1801	0.2107					
残差相关系数		-0.6042	0.5960	-0.4451	-0.6476					

至1.1才有新的因子入选，所建立的方程为：

$$\hat{y} = -21.2055 + 0.4989x_{01} + 0.0864x_{04} - 0.3374x_{14} \quad (6)$$

显然， x_{14} 与 x_{01} 和 x_{04} 的配合明显优于 X_0 中的其它几个因子。从生物学意义上讲，

表 3 试报效果

年份	趋势产量	\hat{y} (气象产量)			亩产(kg)				误差(%)		
		(4)式	(5)式	(6)式	(4)式	(5)式	(6)式	实况	(4)式	(5)式	(6)式
1988	329.0	-24.4	-28.3	-22.7	304.6	300.7	306.3	292	+4.3	+3.0	+4.9
1989	330.6	-47.9	-44.5	-46.1	282.7	286.1	284.5	288	-1.8	-0.7	-1.2

x_{14} 的入选，与育秧期出现严重烂秧烂种气象条件而将影响产量有关。

分别用(4)、(5)、(6)式对1988和1989年作试报，结果见表3，方程(5)的试报效果最好。可见，利用“残差相关选择法”得到了更佳的因子组合。

四、结论

1. “预报因子残差相关选择法”针对因子数多、计算量大而无法同时参加逐步回归计算问题而提出，它从残差角度考虑因子的

优化组合。根据需要，可适当提高残差相关分析的次数(k)，以增加待选因子数，为选择更优化的因子组合提供机会。

2. 本方法简单、计算量相对较小，并且可将众多因子分批单独计算。因此，为利用内存较小的微机在众多因子中客观选择配合较好的预报因子提供了一种较有效的途径。可与其它方法结合使用。

参 考 文 献

- [1] 汤志成等，江苏省双季早稻产量预报的累加型模式，江苏农业学报，1986年第2期。