

# 预报对象和因子间的相关性探讨

王 双 一

(总参气象局气象室)

在气象统计预报中,时常涉及到因子和预报对象之间的相关性问题。一般说来,因子与预报对象之间的相关性越好,则因子对预报对象的指示性也越好。然而,由于因子和预报对象之间还有一个相关稳定性的问题,所以,只有在因子和预报对象之间的相关性好且稳定的情况下,方可以较有把握地说,该因子对预报对象的预报指示性好。在实际作统计预报时的因子选取过程中,总希望被选入的因子与预报对象之间的相关性很好,这并非总是能够如愿的。本文讨论如何通过变换预报对象,使新预报对象与原因子有较好相关性,以及与新预报对象有较好相关性的各因子对于新预报对象的相互配合问题。

一、预报对象与因子间的线性相关的表示——线性相关系数  $r_{xy}$

设预报对象  $Y$  和因子  $X$  的  $n$  次试验值为:

$$Y = (y_1, y_2, \dots, y_n)^T \quad X = (x_1, x_2, \dots, x_n)^T$$

$$\text{记 } \bar{X} = \frac{1}{n} \sum_{t=1}^n x_t, \quad \bar{Y} = \frac{1}{n} \sum_{t=1}^n y_t,$$

$$S_{xx} = \sum_{t=1}^n (x_t - \bar{X})^2, \quad S_{xy} = \sum_{t=1}^n (y_t - \bar{Y})(x_t - \bar{X}),$$

$$S_{yy} = \sum_{t=1}^n (y_t - \bar{Y})^2$$

式中  $S_{xx}, S_{xy}, S_{yy}$  分别为  $X$  与  $X, X$  与  $Y, Y$  与  $Y$  之间的协方差。

$X$  和  $Y$  之间的线性相关系数  $r_{xy}$  为:

$$r_{xy} = S_{xy} / \sqrt{S_{xx} S_{yy}}$$

二、新预报对象与原因子之间的线性相关系数  $r_{xy'}$

记新预报对象为  $Y' = Y + F$ , 其中  $F$  的  $n$  次试验值为  $F = (f_1, f_2, \dots, f_n)^T$ 。

$f$  为一种函数关系,它表示了单个因子的一种变换关系或多因子之间的一种组合关系。由  $f$  所构成的函数的变量中,不准许含有与预报对象  $Y$  同期出现的因子,这样新构成的预报对象对老预报对象  $y$  仍旧是可预报的。否则,新预报对象对老预报对象将失去预报意义。下面将  $f$  看成是一个由  $f$  函数得到的因子。

记  $S_{xy'}$ ,  $S_{y'y'}$  分别为  $X$  和  $Y', Y'$  和  $Y'$  的协方差,  $r_{xy'}$  为  $X$  和新预报对象  $Y'$  之间的线性相关系数。

$$S_{xy'} = \sum_{t=1}^n (x_t - \bar{X})(y'_t - \bar{Y}')$$

$$\text{其中 } y'_t = y_t + f_t, \quad \bar{Y}' = \sum_{t=1}^n (y_t + f_t)$$

$$= \bar{Y} + \bar{f}, \quad \bar{f} = \sum_{t=1}^n f_t$$

$$\text{代入得 } S_{xy'} = \sum_{t=1}^n (x_t - \bar{X})(y_t + f_t - \bar{Y} - \bar{f})$$

$$= \sum_{t=1}^n (x_t - \bar{X})(y_t - \bar{Y})$$

$$+ \sum_{t=1}^n (x_t - \bar{X})(f_t - \bar{f})$$

$$= S_{xy} + S_{xf}$$

其中  $S_{xf} = \sum_{t=1}^n (x_t - \bar{X})(f_t - \bar{f})$ , 为  $X$  和  $F$  之间的协方差。

$$S_{y'y'} = \sum_{t=1}^n (y'_t - \bar{Y}')^2 = \sum_{t=1}^n (y_t - \bar{Y} + f_t - \bar{f})^2$$

$$= \sum_{t=1}^n (y_t - \bar{Y})^2 + 2 \sum_{t=1}^n (y_t - \bar{Y})(f_t - \bar{f})$$

$$+ \sum_{t=1}^n (f_t - \bar{f})^2$$

y 与 F 之间的协方差为

$$S_{yf} = \sum_{t=1}^n (y_t - \bar{Y})(f_t - \bar{f}),$$

F 与 F 之间的协方差为  $S_{ff} = \sum_{t=1}^n (f_t - \bar{f})^2$ ,

$$\begin{aligned} \text{则 } S_{y'y'} &= S_{yy} + 2S_{yf} + S_{ff} \\ r_{xy'} &= S_{xy'} / \sqrt{S_{xx}S_{y'y'}} \\ r_{xy'}^2 &= S_{xy'}^2 / (S_{xx}S_{y'y'}) \\ &= (S_{xy} + S_{xf})^2 / [S_{xx}(S_{yy} \\ &\quad + 2S_{yf} + S_{ff})] \\ &= (S_{xy}^2 + 2S_{xy}S_{xf} + S_{xf}^2) / \\ &\quad (S_{xx}S_{yy} + 2S_{xx}S_{yf} + S_{ff}S_{xx}) \end{aligned} \quad (1)$$

X 与 F 的相关系数  $r_{xf} = S_{xf} / \sqrt{S_{xx}S_{ff}}$

Y 与 F 的相关系数  $r_{yf} = S_{yf} / \sqrt{S_{yy}S_{ff}}$

将  $r_{xy}$ ,  $r_{xf}$ ,  $r_{yf}$  代入 (1) 式得:

$$\begin{aligned} r_{xy'}^2 &= (r_{xy}^2 S_{xx}S_{yy} + 2r_{xy}r_{xf}S_{xx}\sqrt{S_{yy}S_{ff}} \\ &\quad + r_{xf}^2 S_{xx}S_{ff}) / (S_{xx}S_{yy} + 2r_{yf}S_{xx} \\ &\quad \cdot \sqrt{S_{yy}S_{ff}} + S_{xx}S_{ff}) \end{aligned}$$

$$\begin{aligned} \text{令 } a &= S_{xx}S_{yy}, \quad b = S_{xx}\sqrt{S_{yy}S_{ff}}, \\ C &= S_{yy}S_{ff} \end{aligned}$$

$$\text{则: } r_{xy'}^2 = r_{xy}^2 \cdot (a + 2\frac{r_{xf}}{r_{xy}}b + \frac{r_{xf}^2}{r_{xy}^2}c) / (a + 2br_{yf} + c) \dots \dots (2)$$

显然,  $a, b, c > 0$ ,

$$-1 \leq r_{xy}, r_{xf}, r_{yf}, r_{xy'} \leq 1$$

### 三、 $r_{xy}$ 和 $r_{xy}$ 的比较

由 (2) 可知:

1. 当  $r_{xy}$  和  $r_{xf}$  同号, 且  $|r_{xy}| \leq |r_{xf}|$  时, 有  $r_{xf}/r_{xy} \geq 1$ , 则必有  $r_{xf}/r_{xy} \geq r_{yf}$ 。

所以此时也必有  $|r_{xy'}| \geq |r_x|$ 。而只有当  $r_{xy} = r_{xf}$  和  $r_{yf} = 1$ , 等号才成立。要有  $r_{yf} = 1$ , 一般是不可能的。这说明, 只要 X 与 Y 和 X 与 F 之间同时具有正相关或负相关, 且 X 与 F 的线性相关优于 X 与 Y 的线性相关时, 就必有 X 与 Y' 的线性相关优于 X 与 Y 的线性相关。

2. 当  $r_{xf}$  与  $r_{xy}$  异号, 且  $|r_{xy}| \ll |r_{xf}|$  时, 有  $r_{xf}/r_{xy} >> |r_{xf}/r_{xy}|$ 。

此时  $|r_{xy'}| \geq |r_{xy}|$  的可能性是很大的。 $r_{yf}$  的负值越大, 越有利于  $|r_{xy'}| \geq |r_{xy}|$ 。

这说明当 X 与 Y 和 X 与 F 之间分别为正、负相关, 或负、正相关, 且 X 与 F 之间的相关性远优于 X 与 Y 之间的相关性时, 则 X 与 Y' 之间的线性相关优于 X 与 Y 之间的线性相关的可能性很大, 且当 Y 与 F 的负相关性越好, 则越有利于  $|r_{xy'}| \geq |r_{xy}|$ 。

### 四、F 因子存在的可能性

从第三节的讨论中可以看出, 要使  $|r_{xy'}| \geq |r_{xy}|$ , 一般都需要有  $|r_{xy}| \leq |r_{xf}|$ 。这就要求去选取一个因子 F。使 F 与因子之间的相关性优于 Y 与因子之间的相关性。下面讨论 F 存在的可能性。

因子间的同期相关, 一般优于因子间的前期相关。在统计预报中, 预报对象与因子间的相关均为前期相关。当取构造的 F 因子与因子间为同期相关时, 就很有希望达到  $|r_{xy}| \leq |r_{xf}|$  这一要求。下面给出一个取  $f_t = -x_{1t}$  时的例子 ( $f_t$  为 F 的第 t 次试验值,  $x_{1t}$  为  $x_1$  的第 t 次试验值,  $x_1$  为北京 14 点地面风速)。显然, 此时的 F 与因子间是同期相关的。

表 1 给出预报对象 Y 为 24 小时以后的北京 14<sup>h</sup> 地面风速与各预报因子的相关系数及新预报对象与因子的相关系数。从中可见有  $|r_{xiy}| < |r_{xif}|$  和  $|r_{xiy'}| > |r_{xiy}|$ 。在这个例子中, 并没有去考虑一个恰当的 F, 但作为一个例子, 也说明了得到一个使  $|r_{xf}| \geq |r_{xy}|$  的 F 是不难的。

表 1

$x_t$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$r_{xiy}$	0.123	0.028	0.054	-0.026	-0.026
$r_{xif}$	-1.000	-0.240	-0.204	0.201	-0.121
$r_{xiy'}$	-0.660	-0.160	-0.113	-0.111	-0.111

表中预报对象 Y 为 24 小时以后北京 14<sup>h</sup> 地面风速,  $x_1$  为北京 14<sup>h</sup> 地面风速,  $x_2$  为 02 时北京地面风速,  $x_3$  为 08<sup>h</sup> 850hPa 北京风向,  $x_4$  为 08<sup>h</sup> 700hPa 北京温度,  $x_5$  为 08<sup>h</sup> 700hPa 北京风向。 $r_{xiy}$ ,  $r_{xif}$ ,  $r_{xiy'}$  分别为  $x_i$  与 Y、F 及新预报对象 Y' ( $Y' = y + F$ ) 之间的相关系数。

### 五、因子和预报对象间的线性相关信度

若 X 和 Y 间线性相关系数  $r_{xy}$  可达到信度  $\alpha$ , X 和 Y' 间线性相关系数  $r_{xy'}$  可达到信度  $\alpha_1$ 。

用  $F_a$  和  $F_{a_1}$  分别表示  $r_{xy}$  和  $r_{xy'}$  的  $F$ -统计量,

$$\text{则 } F_a = [r_{xy}^2 / (1 - r_{xy}^2)] \times (n - 2)$$

$$F_{a_1} = [r_{xy'}^2 / (1 - r_{xy'}^2)] \times (n - 2)$$

$F_a$  和  $F_{a_1}$  的自由度均为  $(1, n - 2)$

当  $|r_{xy}| \leq |r_{xy'}|$  时, 有  $F_a \leq F_{a_1}$ , 故  $\alpha_1 \leq \alpha$

这说明, 随着因子和预报对象间的线性相关性增强, 因子与预报对象间的线性相关的信度也随之提高了。

## 六、对新预报对象而言的各因子之间的相互配合

在某个预报对象的产生过程中, 往往受到许多物理因素的同时作用。而这许多物理因素在对预报对象形成的作用过程中, 又包含着各物理因素之间的相互作用过程。这种物理因素对预报对象产生的作用和各物理因素之间的相互作用, 从统计学角度来看, 就表示为因子和预报对象之间的相关性及各因子之间的相关性。

在统计预报的实践中, 人们总希望统计模式包含的因子与预报对象有较好的相关性, 而各因子之间的相关较差, 也就是各因

子之间的配合要好。但这往往不能如愿, 因为在因子的选取过程中, 往往是与预报对象相关性比较好的各因子之间, 也有较好的相关性, 即这些因子之间的配合可能是不好的。那么在构成新预报对象之后, 与新预报对象相关性较好的那些因子之间的配合又如何呢?

表 2 是表 1 中因子  $x_i$  与  $x_j$  之间的相关系数  $r_{x_i x_j}$  ( $i, j = 1-5$ ), 可以看出各因子之间的线性相关性并不显得都好或都差。最大的相关系数为  $r_{x_3 x_5} = 0.574$ , 最小的为  $r_{x_2 x_5} = 0.090$ 。这表明这些因子之间对于新预报对象具有一定的配合性。

表 2 因子之间的相关系数  $r_{x_i x_j}$

$r_{x_i x_j}$ \ $x_j$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$x_1$	1.000				
$x_2$	0.240	1.000			
$x_3$	0.204	0.178	1.000		
$x_4$	-0.201	-0.223	-0.396	1.000	
$x_5$	0.121	0.090	0.574	-0.405	1.000