

一个着眼于预测能力及稳定性的 因子普查方法

俞善贤

(浙江省气象科学研究所)

提 要

本文提出用系数 r_p 来衡量因子预测能力的方法。当 $r_p \leq 0$ 时,表明因子无预测能力;当 $r_p > 0$ 时,其值越大表明因子的预测能力越强。通过杭州市蔬菜产量预报因子普查表明,该方法不但能较好反映因子的预测能力,而且使选取的因子具有较好的稳定性。

一、引言

在气象统计预报中,使用的因子往往先是从几千个因子中进行普查,然后以单相关系数较高的因子(或区域平均值)作为建立预测模型的候选因子。对其中存在的一些问题,作者在文献[1]中进行了讨论。为此,本文提出一种着眼于因子预测能力和稳定性

的普查方法,并在预报对象与500hPa月平均场、北太平洋286个格点上的月平均海温场的普查中得以实现。

二、着眼于预测能力的系数 r_p

单相关系数事实上反映了 x 与 y 之间用一个线性函数来拟合的好坏,而不能完全反映预测的好坏。预测平方和(PRESS)可

以衡量预测能力的好坏(2), (3), 可用它作为准则选取预测能力较好的因子。

设有预报因子 X 和预报对象 Y 的 n 次观测值

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

考虑如下线性回归模型

$$Y = X\beta + e \quad (1)$$

式中

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

若把第 i 个试验点(样本)去掉, 余下的 $n-1$ 个试验点的模型记

$$Y(i) = X(i)\beta(i) + e(i) \quad (2)$$

式中 $Y(i)$ 、 $X(i)$ 和 $e(i)$ 分别是 Y 、 X 和 e 中删去第 i 行后得到的。由最小二乘估计可得出 $\hat{\beta}(i)$, 则在第 i 个试验点处的预测值为 $x'_i\hat{\beta}(i)$, 而预测偏差为 $f_i = y_i - x'_i\hat{\beta}(i)$ 。

对每一个试验点重复上述做法, 可分别得出 $f_1, f_2, \dots, f_n$, 其平方和

$$PRESS = \sum_{i=1}^n f_i^2 \quad (3)$$

称为预测平方和, 这个值的直观意义很明确, 它反映了对历史样本的预测能力。

在因子普查中, 为了便于比较, 定义系数

$$r_p = 1 - \left[PRESS / \sum_{i=1}^n (y_i - \bar{y})^2 \right] \quad (4)$$

r_p 取值有两种情况:

(1) 当 $r_p \leq 0$ 时, 则 $PRESS \geq \sum_{i=1}^n (y_i - \bar{y})^2$, 表示用该因子建立的线性回归方程, 对历史样本的预报, 总的来说, 还不如每次

都用 $\bar{y}$ 值来预测 $y_i (i=1, 2, \dots, n)$ 来得好, 因为 $(y_i - \bar{y})$ 可理解为用 $\bar{y}$ 来预报 y_i 的偏差, $\sum_{i=1}^n (y_i - \bar{y})^2$ 为总偏差平方和。这就说明了该因子无预测能力。

(2) 当 $r_p > 0$ 时, 则 $PRESS < \sum_{i=1}^n (y_i - \bar{y})^2$, 表示该因子有一定的预测能力, r_p 值越大表明 x_i 对 y_i 的预测能力越强。

为了计算方便, (3)式 $PRESS$ 值的计算可用下式(2)

$$PRESS = \sum_{i=1}^n \left(\frac{\hat{e}_i}{1 - h_{ii}} \right)^2 \quad (5)$$

其中 h_{ii} 为矩阵 $H = X(X'X)^{-1}X'$ 的第 i 个对角元。对于一元线性回归, h_{ii} 的计算式还可简化如下:

$$\text{由于 } X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \text{得 } (X'X) = \begin{pmatrix} n & \overline{nx} \\ \overline{nx} & \sum_{i=1}^n x_i^2 \end{pmatrix} \quad (6)$$

其 $(X'X)$ 的逆矩阵为

$$(X'X)^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{x}/s^2 & -\overline{x}/s^2 \\ -\overline{x}/s^2 & 1/s^2 \end{pmatrix} \quad (7)$$

其中 $\overline{X} = \sum_{i=1}^n x_i / n$

$$s^2 = \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n x_i^2 - n\overline{x}^2 \quad (8)$$

下面仅考虑矩阵 $H = X(X'X)^{-1}X'$ 的对角元部分

$$H = X(X'X)^{-1}X' = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \frac{1}{n} + \overline{x^2}/s^2 & -\overline{x}/s^2 \\ -\overline{x}/s^2 & 1/s^2 \end{pmatrix} \begin{pmatrix} 1 & 1 \cdots 1 \\ x_1 & x_2 \cdots x_n \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{n} + (x_1 - \overline{x})^2/s^2 & \cdots \cdots \\ \cdots \cdots & \frac{1}{n} + (x_2 - \overline{x})^2/s^2 & \cdots \cdots \\ \cdots \cdots & \cdots \cdots & \frac{1}{n} + (x_n - \overline{x})^2/s^2 \end{pmatrix}$$

得 $h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \quad (i=1, 2 \cdots n)$ (9)

这样用(5)式和(9)式, 计算就变得十分简单。

h_{ii} 有一定的几何意义, h_{ii} 表达式的第二项, 表示在自变量的空间中, 第*i*个试验点到样本中心 $\overline{X}$ 的马氏(Mahalanobis)距离。把PRESS值表示为残差 $\hat{e}_i$ ($i=1, 2, \cdots n$)的加权平方和, 如果 x_i 离 $\overline{X}$ 的距离越远, h_{ii} 就越大(接近于1), 权重 $W_i = \frac{1}{(1-h_{ii})^2}$ 也就越大, 因此PRESS比较强调远离 $\overline{X}$ 的点。

三、实际计算分析结果

在杭州市蔬菜产量与500hPa月平均场、北太平洋286个格点上月平均海温场的相关普查选取因子中加以应用。样本长度 $n=22$, 通过计算表明, 用 r 和 r_p 得出的结论有较大的差别。计算的部分结果在附表中作了对比。附表中的最大(小)相关系数是用“刀切法”[1]去掉某一个样本使相关系数绝对值达最大(小)。从表中可以看出, 当 r_p 值较小时, 相关系数的最大值与最小值之差就较大, 且最小值往往较小, 这说明波动较大, 缺乏稳定性。

一般容易误认为相关系数的最小值的置

信水平较高, 该因子预测能力就较强, 但在本例中却不完全如此。例如, 9月的蔬菜产量与因子 x_{15} 的最小相关系数达0.5111; 10月的蔬菜产量与因子 x_1 最小相关系数达-0.5045, 而它们的 r_p 值均为负值, 为此作如下点聚图1, 2。

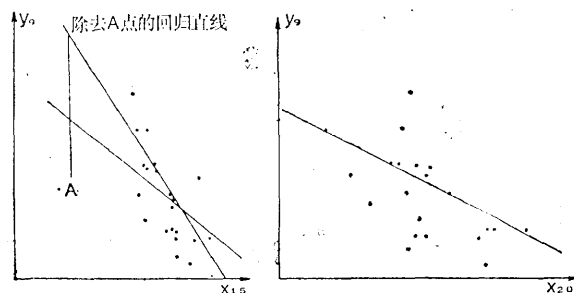


图1 9月蔬菜产量与 x_{15} , x_{20} 的点聚图

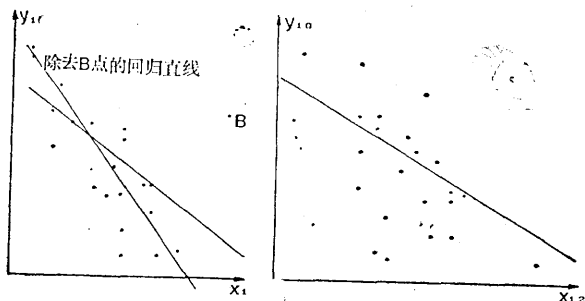


图2 10月蔬菜产量与 x_1 , x_{12} 的点聚图

从图1, 2中可以清楚地看出, 凡 r_p 值为

附表

相关系数 r 与 r_p 结果对比($r_{\alpha=0.01}=0.5168$)

预报对象	预测能力较差因子					预测能力较好因子				
	因子	r	r_{max}	r_{min}	r_p	因子	r	r_{max}	r_{min}	r_p
3月产量	x_{17}	-0.5522	-0.6844	-0.4870	0.0344	x_{56}	-0.5299	-0.6282	-0.4487	0.1679
	x_{48}	-0.5178	-0.6417	-0.4501	0.0102	x_{12}	0.5072	0.6139	0.4021	0.1291
	x_3	-0.5284	-0.6250	-0.3883	0.0749	x_5	-0.5651	-0.6195	-0.4918	0.1857
6月产量	x_{50}	-0.5852	0.7576	0.4072	-0.0085	x_{21}	-0.5915	-0.6633	-0.4784	0.1876
	x_{35}	-0.5584	-0.6640	-0.3509	0.0606	x_{32}	0.5738	0.6418	0.4692	0.1476
	x_{46}	0.5506	0.6469	0.3489	0.0496	x_{55}	0.5837	0.6784	0.4112	0.1276
	x_{23}	-0.5335	-0.6510	0.3687	0.0541	x_{57}	-0.5701	-0.6449	-0.4488	0.1264
9月产量	x_{15}	-0.5428	-0.6691	0.5111	-0.0531	x_{16}	-0.5139	-0.5723	-0.4312	0.1226
	x_{28}	0.5082	0.5829	0.4224	0.0773	x_{20}	-0.5092	-0.5651	-0.4224	0.1452
10月产量	x_1	-0.5833	-0.8222	-0.5045	-0.0094	x_{12}	-0.5773	-0.6354	-0.4926	0.2304
						x_6	0.5558	0.6207	0.4802	0.1968

负的点聚图中, 均有强影响点 (A)、(B) 存在, 尤其在图 2 中的因子 x_1 , 除 (B) 点外, 用其它点作线性回归的效果相当好, 但 (B) 点大大影响其效果, 这些强烈影响统计量的点称为强影响点。统计分析中我们希望每个数据, 既对参数估计或预测有一定的影响, 但影响又不太大, 如果影响过大, 那么包含这组数据的统计量和不包含这组数据的统计差异就很大, 考虑到数据获得具有一定的随机性, 于是包含有较大影响数据所导出的统计量就容易产生不稳定。对 x_{12} 而言, 虽然各点离回归直线的离散度大一点, 但各点分布较合理, 统计量较稳定, 且有一定的预报能力。

需要特别加以说明的是, 我们不能断然地说, 含有强影响点的数据作为因子是完全不可取的。原则上讲, 如果产生强影响点的原因来自数据收集的过失、或同研究问题的基本假设不相符合、或其它特殊原因, 则应去掉该点, 如果找不到产生的原因, 对这种情况的处理应十分小心, 往往这些数据中能提供其它数据不能提供的信息, 在建立数学模型时, 可采用一些特殊处理方法, 但在一个业务系统中, 却往往会产生很多麻烦, 因此如果不作一些特殊处理, 建议还是不要选

用该因子。

四、讨论

用 r_p 来普查因子, 我们认为有下列几方面的优点:

1. 计算量不大, 完全可以在低档微机中实现, 且便于在业务系统中统一处理。如在相邻格点上, 给一个 r_p 的临界值, 可以自动组合一个区域, 以区域平均值作为预报因子。

2. 同用相关系数来普查因子相比, 该方法在较大程度上弥补了用相关系数的不足, 用 r_p 可以排除选取因子含有强影响点。对因子在空间中的分布, r_p 的要求较高。

3. r_p 也反映了因子的稳定性, 比用刀切法考察相关系数稳定性的处理要方便的多, 刀切法虽可以找出强影响点, 但不能估计对预测的影响。

综合上述, 可以认为: 用 r_p 来普查因子是一种切实可行的做法, 有一定的使用价值, 应引起重视, 并不断地给以改进和发展。

参考文献

- (1) 俞善贤, 陈孝源, 气象数据回归分析中的若干问题及其对策, 气象学报, 第46卷, 第3期, 1988.
- (2) 陈希孺, 王松桂, 近代实用回归分析, 106—112, 广西人民出版社, 1984.
- (3) 俞善贤, 沈锦花, 基于预测平方和PRESS准则选取预报因子的向前、向后算法及其应用, 气象学报, 第47卷, 第3期, 1989.