非线性回归方程的探索

姜子俊　　乔志敏

（内蒙古自治区气象学校）　　（内蒙农牧业环保站）

提　要

本文使用常见的15类函数，采取粗选与精选两步，选择非线性预报因子。根据预报对象与因子时空尺度相匹配的原则，确定待选因子，建立非线性回归预报方程。试验表明，预报因子出现的形态高达390类，且非线性方程的效果优于线性方程。

一、引　言

现实世界，包括观测数据以及我们所研究的大气现象，是一个具有无限的现实联系并具有无限转化可能性的客观系统。因此，在讨论预测问题时，仅用线性因子显然不能全面的表述预报对象。本文借助计算机，探讨非线性预报因子和回归方程，期望能使预报方程和预测效果更接近客观实际。

二、非线性因子的选择

1．一个惊人的数字

预报模型：$Y=f(X)+e$，当X中至少有一个为非线性时，称为非线性模型。

为选到适宜的预报因子，我们使用较常见的15类函数，即$f(X)$中的变量分别为x、x^2、x^3、x^4、$1/x$、$1/x^2$、$\sqrt{x}$、$1/\sqrt{x}$、e^{-x}、$\ln x$，以及指数函数（$Y=ae^{bx}$），幂函数（$Y=ax^b$），双曲函数（$1/Y=a+b/x$），负指数函数（$Y=ae^{b/x}$），S型函数（$Y=1/(a+be^{-x})$）。前10种函数（以下称为一类函数，后5类函数，以下称为二类函数。）是显函数，系数已确定。其中第一种是线性函数，其余9种函数，我们可以通过对自变量作适当的函数变换，使变换所得的变量与依变量Y之间成线性关系，就可很方便的计算相关系数。它们的相关系数计算公式为：$r=S_{xy}/S_xS_y$或经变换后的公式$r=S_{x'y}/S_{x'}S_y$。其中S_{xy}为x、y的协方差，S_x为x的标准差，S_y为y的标准差；$S_{x'y}$为$f(X)$中的各类变量与Y的协方差，$S_{x'}$为$f(X)$中的各类变量的标准差。

在用计算机筛选因子时，还将一类函数线性组合后，用逐步回归法挑选组合因子并计算复相关系数，即

$$R=\sqrt{U/S_{yy}}$$

$$=\sqrt{1-\Sigma(Y_i-\hat{Y}_i)^2/\Sigma(Y_i-\overline{Y})^2}。$$

本试验中，计算出的一类函数的R，共组合因子数曾分别出现2、3、4个三种组合情况。从理论上讲，就有$C_{10}^2=45$，$C_{10}^3=120$，$C_{10}^4=210$，即有可能组合形式375种。也就是说，所选预报因子的可能形态的数量是惊人的。然而，却尚未包括所有可能的组合形式（当然，理论上所有的组合形式不可能都出现）。这反映了预报因子性态的复杂性，并间接体现预报对象性态的可能现实。

2．计算相关指数

对于二类函数，由于带有待定参数a、b，是隐函数。其相关系数不能直接计算。为了解非线性相关程度，我们通过一元非线性函数的线性化处理，即对自变量和依变量均作适当的函数变换，使变换所得的变量之间成线性关系，并用最小二乘法确定系数，就得到线性回归方程。再根据变换公式，确定回归曲线中相应的参数a、b[1]，最后构成曲线回归方程，再以曲线回归方程的估计值（$\hat{\hat{Y}}$），计算相关指数[2]。相关指数的计算公式为：

$$R^* = \sqrt{1 - \Sigma(Y_i - \hat{\hat{Y}})^2 / \Sigma(Y_i - \overline{Y})^2}。$$

R^*只取正值。为确定二类函数的参数a、b，对依变量Y均需经过函数变换，故它们的残差平方和$Q = \Sigma(y_i - \hat{\hat{y}}_i)^2$和$R^*$不能在求解正规方程时得到[1]，而只能用上面提到的公式计算R^*。

当选择二类函数中某一种函数型作为因子后，在加入方程运算时，其作为因子的表现形式就改变为一元曲线回归方程的类型。例如，双曲函数（$1/y = a + b/x$）的回归方程应为原函数的倒数，即$\hat{\hat{Y}} = \dfrac{1}{a + b/x}$。也就是说，应以$\hat{\hat{y}}$，即$\dfrac{1}{a + b/x}$作为新因子的数值进入方程参与运算。

由（1）（2）两步可知，若某预报因子经计算r、R后出现的函数类型总数为m，R^*的类型为5，则描述相关密切程度的总类别数为$G = m + 5$。为叙述方便，以R'（包含所有的$|r|$、R与R^*）表示相关密切程度的量。比较全部R'，若h属于G类内任一类别，有

$$R'_h = \max_{1 \le g \le G} R'_g$$

$$(g = 1, 2, \cdots, m, m+1, \cdots, G) \quad (1)$$

则表明该预报因子属于h类。

3. 粗选与精选结合的筛选因子方法

第一步粗选。用一类函数粗选因子。

根据童绍颜[3]的工作，我们以内蒙古中部地区的黑白灾害为预报对象。样本资料为1951—1976年共26年。由1、5、8三个月的500hPa月平均高度网格点资料计算相关场。以$|r|$或$R \geqslant 0.40$为标准进行普查，初选的网格点数共159个。

第二步精选。用一、二两类函数判定待选因子的类别及具体形式。

从相关场中分析初步选入的网格点分布，各月都有几个集中区域。根据筛选预报因子应注意预报对象与因子的时空尺度相匹配的原则，我们选集中区≥3个网格点的连续区，取其平均值作为一个精选因子。这样共精选16个因子（具体地理位置略），作为方程的待选因子。它们与预报对象关系的密切程度由一、二两类函数计算，以（1）式判定每个因子的类别。

4. 确定线性待选因子

为与非线性待选因子作比较，我们以相同的样本年代，用第3节的方法与标准，粗选出与预报对象线性关系密切的格点136个。也有类似的集中区，共得到精选因子13个（具体地理位置略）。应指出的是，线性与非线性相关程度≥0.4的网格点，多数重合，但也略有差别。从以上结果可看出，选出的非线性网格点数比线性的多23个；集中区（精选因子）多3个。

三、建立回归预报方程

回归模型：$Y = X\beta + e$ (2)

当X均为线性因子时，为线性回归方程；当X中至少有一个为非线性因子时，组成非线性回归方程。求回归系数均用最小二乘法，即$Q = \Sigma(y_i - \hat{y}_i)^2 \Rightarrow \min$原则。由于随着因子的增加，残差$Q$减小，为便于对回归效果比较，方程中的因子均取3个。今以第二部分之结果，叙述如下。

线性逐步回归方程：

$$\hat{Y} = 3.694 - 2.005x_2 - 1.036x_5 + 0.999x_9 \quad (3)$$

与线性因子同区域组成的非线性回归方程：$\hat{Y} = 4.76967 - 2.475511x_2^2 - 1.192362x_5^2 - 1.625626e^{-x_9}$（4）

非线性逐步回归方程：

$$\hat{Y} = 3.185 - 3.098x_2^2 - 1.204x_6^2 + 1.215x_{11} \quad (5)$$

为比较3种方程的优劣，将有关结果列于表1。由表1可看出，非线性方程优于线性方程，非线性逐步回归效果最佳。

表1　　3 种 回 归 方 程 效 果 之 比 较

方　法	R	S	U	F	E	SE
线性逐步回归	0.8137	0.4860	10.1831	14.3726	0.3523	0.934
非线性回归	0.8455	0.4464	10.9947	18.3905	0.3347	0.898
非线性逐步回归	0.8569	0.4310	11.2926	20.2660	0.3064	0.877

注：R：复相关系数，S：标准差，U：回归平方和，F：F检验值，E：平均绝对误差，SE：试报5年平均绝对误差。

四、二类函数的非线性方程实例

根据文献[3]，我们还建立了内蒙古西部地区的黑白灾预报方程，样本资料1951—1976年，共26年。以1、8两月的500hPa月平均高度网格点资料计算相关场。所得结果与前述的实例特点相似，同样说明了非线性筛选因子的优越性（有关数据略）。现仅列出回归方程。由于线性逐步回归方程只选定2个因子，为便于比较，各类方程的因子均取2个。

线性逐步回归方程：

$$\hat{Y} = 1.080887 + 2.737084x_2 + 1.290754x_4 \quad (6)$$

与线性因子同区域组成的非线性回归方程：

$$\hat{Y} = -0.834477 + 0.80796(1.634243e^{1.174191x_2}) + 0.5279757(1.679043e^{0.9248981x_4}) \quad (7)$$

非线性逐步回归方程：

$$\hat{Y} = 0.078 + 1.9x_8^3 + 0.901(1.634243e^{1.174191x_2}) \quad (8)$$

上述（7）、（8）式中，凡括弧中的内容，均为新因子的具体形式。本例中，（7）式中有2项，（8）式中有一项，均为指数函数（$Y = ae^{bx}$）。其出现的具体形式，实际上就是单因子曲线回归方程$\hat{\hat{Y}}$（即单因子的曲线回归方程之预报估计值）。

为比较3种方程之优劣，将有关结果列于表2。由表2可看出，非线性方程优于线性方程，非线性逐步回归效果最佳。

表2　　3 种 回 归 方 程 效 果 之 比 较

方　法	R	S	U	F	E	SE
线性逐步回归	0.8455	0.639769	23.59562	28.82405	0.4772	0.7820
非线性回归	0.8544	0.6225697	24.09498	31.08284	0.4725	0.7784
非线性逐步回归	0.8828	0.5626509	25.72837	40.63538	0.4193	0.7674

注：各符号代表的意义同表1。

五、讨论

（1）以15类常见的函数挑选预报因子，探讨建立非线性回归，经比较，效果较好。这说明从筛选非线性预报因子出发，经粗选与精选两步工作；又考虑到预报因子与预报

对象时空尺度的匹配；预报因子共有百种以上的可能性态。这是否可说明更逼近实际大气的变化状态。

（2）由表1、表2均可看出，虽然非线性逐步回归最佳，但从试报结果分析，$SE>E$。这表明，简单的选择方程（如方程只选3个因子，无论从理论角度还是经验之谈，均不为最佳数），难以解决预报效果的稳定性问题；不能克服拟合好而预报效果差的通病，这有待做进一步的探讨。

（3）为探索本方法与最优化方法〔4〕之优劣，我们将两种方法作一比较。其结果见表3和表4。表3中的差值，是由最优化方法之值减去本方法之值而得。“+”值表示最优化方法较本方法的相关程度高。由表3可知，总体上讲，最优化方法所选因子的相关程度较本方法高，最大差值为+0.0154。由于最大差值约为1.5%，这可能对回归效果无明显影响。凡差值为“-”，其函数形式均为组合因子（其中x_{12}，由表3之注③知，文献〔4〕计算的R值小于作者计算的值，故不便比较），这可反映本方法的部分优点。

在对文献〔4〕的方法与本方法的方程效果比较之前，需说明一点，即由于计算机编程中小数取位不同等原因而造成的舍入误差，故我们用〔4〕的资料算得之方程的回归

表3　最优化方法所选相关系数与本方法所选相关系数绝对值的比较

因子	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{12}
最优化	0.6840	0.6459	0.5972	0.4917	0.6505	0.4762	0.5077	0.6341	0.4073	0.5398	0.4974
本方法	0.6758	0.6459	0.5843	0.4895	0.6475	0.4608	0.4949	0.6734	0.4594	0.5291	0.4979
函数形式	x^4	$\ln x$	$1/x^2$	$1/\sqrt{x}$	e^{-x}	$1/x^2$	$1/x^2$	$x^4+\frac{1}{x^2}$	$\frac{1}{x}+\frac{1}{x^2}+\frac{1}{\sqrt{x}}$	$1/x^2$	x^2
差值	+0.0082	0.0	+0.0129	+0.0022	+0.0030	+0.0154	+0.0128	−0.0393	−0.0516	+0.0107	-0.0005

注：①x_{11}从缺，是因为有两个数值印刷不清，②x_9，冯文计算的a=−11.7489在22个样本中，最小值为第11个样本60.3，而$60.8^{-11.7489}=1.2955\times10^{-21}\doteq0$，则其余样本更近于零值，估计有误。③本文计算了11个因子的线性相关系数，除x_{12}外，均与〔4〕同。x_{12}的R值，〔4〕为0.4845，本文所得为0.4851。

表4　3个方程回归效果之比较

方法	Q	R	F	S	E
线性	44176.67	0.8915	16.44991	50.97674	36.2810
最优化	42444.81	0.8960	17.29452	49.96752	30.9150
本方法	40001.69	0.9023	18.61036	18.50815	30.2032

注：Q：残差平方和，其余同表1。

系数等量值与〔4〕略有差异。为便于比较，将本文所用程序计算出的回归方程列出。3种方法的回归方程如下。

线性回归方程〔4〕：

$$\hat{Y}=341.0641-3.066882x_1+28.77729x_5-2.791786x_{10}-0.5555066x_{12}$$

非线性回归方程〔4〕：

$$\hat{Y}=65.21443-4.021605E-10x_1^{5.778}-4736.639x_5^{-3.2631}+1.305906E+0.9x_{10}^{4.0314}-7.729911E-0.3x_{12}^{846}*$$

本方法之非线性回归方程：

$$\hat{Y}=30.16087-1.518778E-0.6x_1^4-2696.993e^{-x_5}+501190.8/x_{10}^2-3.582155E-0.3x_{12}^2$$

由表4可知，本方法的效果略优于最优化方法〔4〕之结果，参与方程之函数均属一类，计算方便，表达简单。

参考文献

〔1〕 屠其璞等，气象应用概率统计学，271—273，279。气象出版社，1984年

〔2〕 么枕生，气候统计学基础，213—216，240—242，科学出版社，1984年。

〔3〕 竜绍颜，内蒙古地区黑白灾标准及其气候特点的初步分析，内蒙古气象，1981年第3期。

〔4〕 冯耀煌、杨旭，最优化方法在天气预报中的应用，气象，第13卷第8期，1987年。

Probe of nonlinear regression equation

Jiang Zijun
(Inner Mongolia Meteorological School)
Qiao Zhimin
(Inner Mongolia Agriculture Environmental Protective Monitoring station)

Abstract

In this paper, fifteen kinds of ordinary functions were applied to select the nonlinear predictors. Coarse and fine selections were accepted. A nonlinear forecasting equation has been built. The results showed that the predictors had 390 kinds of formation and the effects of the nonlinear forecasting equation in this method were better than those of the linear equations.