非线性稳健回归在天气预报中的应用

冯耀煌

（沈阳区域气象中心研究所）

提　要

统计预报中常用的多元回归模型，由于预报因子的非线性及模型数据的非正态性等情况，影响预报效果，针对这些问题，本文提出非线性稳健回归模型，使预报的准确率得到提高。

一、前言

线性回归模型是统计预报中常用的一种数学模型，但以下问题影响预报效果。首先，假设线性回归模型自变量（预报因子）是线性的，而实际上它们与预报量的关系都是非线性的；其次是一些无法预料的错误和误差，使某些观测值严重偏离了样本中其它观测值，使样本数据受到“污染”；第三则是经典统计方法大多以正态分布假设及最小二乘法为基础，而实际上有的统计量不服从正态分布。针对这些问题，本文提出非线性稳健回归模型，并以实例说明在预报中的应用及试验效果。

二、线性稳健回归

随机变量y与m个自变量（预报因子）x_1，$x_2\cdots x_m$关系的线性回归模型为：

$$y=\beta x'+\varepsilon$$

其中$x'=(1,x_1,x_2\cdots x_m)$，$\beta=\begin{bmatrix}\beta_0\\ \beta_1\\ \cdots\\ \beta_m\end{bmatrix}$

经典的最小二乘估计就是选取β，使残差平方和：

$$Q=\sum_{i=1}^{N}e_i^2=\sum_{i=1}^{N}(y_i-\hat{\beta}x'_i)^2\to\min$$

其中$x'_i=(1,x_{1i},x_{2i}\cdots x_{mi})$。

这种方法虽然计算简单，估计的精度也比较高，但所得到的方程不够稳定，易受前述因素的影响。

稳健回归的基本思想是用某种函数$P(e_i)$去代替e^2_i，再求

$$\sum_{i=1}^{N}P(e_i)\to\min$$

的解。函数$P(e_i)$的值随e_i的增长速度必须小于e^2_i的增长速度，因而上式的解对e_i的影响不如最小二乘估计的解敏感。函数$P(e_i)$的形式可有多种，在预报问题中，人们常常关心的是预报的拟合误差精度，一般要求平均绝对误差越小越好[1]，即要求：

$$ERM=\frac{1}{N}\sum_{i=1}^{N}|e_i|\to\min$$

但上式等价于

$$\sum_{i=1}^{N}|e_i|\to\min$$

这时，目标函数为

$$f(\hat{\beta})=\sum_{i=1}^{N}|e_i|=\sum_{i=1}^{N}|y_i-x'_i\hat{\beta}|\to\min$$

这样求出的β估计值为最小绝对偏差L_1估计[2]。所得到的回归方程便是稳健回归方程。

最小绝对偏差L_1估计没有最小二乘估计求解方法那样简单，不能用间接的解析法

求出，而是要用多变量直接寻优方法——单纯形加速法（具体方法可看参考文献[3]）。

三、非线性稳健回归

以上是假设因变量y与m个自变量x_1、x_2、……x_m的关系是线性关系，但一般实际情况是非线性的，现假设预报因子x_i除了线性类型外，还可以有x_i^a、e^{ax_i}、$\ln x_i$三种非线性类型[4]，x_i究竟属于哪种类型，则要计算y与x_i、x_i^a、e^{ax_i}、$\ln x_i$这四种类型的相关系数的绝对值。公式如下：

$$|R_k| = \left| \frac{\sum_{i=1}^{N}(x'_{ij}-\overline{x'})(y_j-\overline{y})}{\sqrt{\sum_{i=1}^{N}(x'_{ij}-\overline{x'})^2 \sum_{j=1}^{N}(y_j-\overline{y})^2}} \right| \quad (k=1,2,3,4)$$

$k=1$时，$x'_{ij}=x_{ij}$，$\overline{x'}=\frac{1}{N}\sum_{j=1}^{N}x_{ij}$，

$k=2$时，$x'_{ij}=x^a_{ij}$，$\overline{x'}=\frac{1}{N}\sum_{j=1}^{N}x^a_{ij}$；

$k=3$时，$x'_{ij}=e^{ax_{ij}}$，$\overline{x'}=\frac{1}{N}\sum_{j=1}^{N}e^{ax_{ij}}$；

$k=4$时，$x'_{ij}=\ln x_{ij}$，$\overline{x'}=\frac{1}{N}\sum_{j=1}^{N}\ln x_{ij}$。

第1、4两种类型，可直接计算得$|R_1|$，$|R_4|$，但第2、3两种类型，还含有一个参数a，这就要用单变量的直接寻优方法——外推内插法[3]，求出使目标函数$f(a)=|R_2|$最大的a值及$|R_2|$，同样求出$|R_3|$，然后比较$|R_1|$、$|R_2|$、$|R_3|$、$|R_4|$，则x_i属于相关系数绝对值最大的那种类型。

当每个预报因子$x_i(i=1,2,\cdots m)$类型确定以后，再对每个因子作变量变换，如x_1属于第2种类型，则令$W_1=x_1^a$（a为已知），其余类推。这样得到变换后的m个新自变量$W_1,W_2,\cdots\cdots W_m$，因W_i一般都是非线性类型，所以y与m个新自变量的关系为非线性回归模型：

$$y=W'\beta+\varepsilon$$

其中$W'=(1,W_1,W_2,\cdots W_m)$，$\beta=\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_m\end{bmatrix}$

这时，对回归系数β进行估计，如果用最小二乘估计，则令目标函数

$$Q(\hat{\beta})=\sum_{i=1}^{N}(y_i-W'_i\hat{\beta})^2\rightarrow\min$$

其中$W'_i=(1,W_{1i},W_{2i},\cdots W_{mi})$

求得的$\hat{\beta}$便是一般非线性回归系数。如果用最小绝对偏差L_1估计，则令目标函数

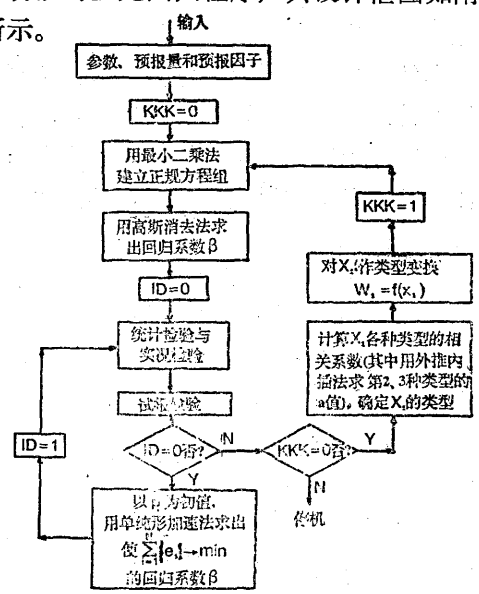
$$f(\hat{\beta})=\sum_{i=1}^{N}|y_i-W'_i\hat{\beta}|\rightarrow\min$$

这样求得的$\hat{\beta}$便是非线性稳健回归系数。

以上两种估计的求解方法是和线性的求解方法完全一样的。

四、程序设计

为了做各种方法的对比，使一次上机计算就能得到各种方法的计算结果，特设计了多功能的多元回归程序，其设计框图如附图所示。



附图　多功能多元回归程序框图

图中最小二乘法和高斯消去法都已为大家所熟悉，外推内插法是非线性单变量直接寻优方法，单纯形加速法是非线性多变量直

接寻优方法，这后两种方法在参考文献[3]中有详细的论述和程序框图。

多功能多元回归程序是用FORTRAN语言编写的，它在IBM微机上通过使用，给计算带来了很大的方便，只要把参数、预报量和预报因子输入，很快就能得到一般线性回归方程、线性稳健回归方程及一般非线性回归方程和非线性稳健回归方程，同时还得到各种方程的统计检验、实况检验和试报检验的结果。在这基础上，只要互相比较一下，就能选出效果最好的最优预报方程。

五、应用举例

应用上面程序进行长期降水预报、年平均流量预报和气象产量预报等试验，都取得类似的效果。现以长期降水预报为例，预报量y为辽宁省3—5月降水距平百分率，选出5个前一年11—12月500hPa月平均高度场球函数展开的振幅和位相作为预报因子，其中x_1为前一年11月振幅C_7^3，x_2为前一年11月振幅C_{10}^4，x_3为前一年12月振幅C_8^2，x_4为前一年11月位相Q_7^3，x_5为前一年12月位相Q_3^1。资料样品容量为33年，用独立样品试报2年（数据从略）。我们利用这些资料，做了以下两种试验：

（一）使用全部样本点，建立方程

把参数、预报量和预报因子原始数据全部输入机器，便可得到以下4个预报方程：

1. 一般线性回归方程

$$\hat{y} = -48.6091 + 0.047x_1 - 0.1756x_2 + 0.1597x_3 + 0.2409x_4 + 0.0918x_5 \quad (1a)$$

2. 线性稳健回归方程

$$\hat{y} = -44.8357 + 0.0117x_1 - 0.1518x_2 + 0.1609x_3 + 0.1988x_4 + 0.1019x_5 \quad (2a)$$

3. 一般非线性回归方程

$$\hat{y} = 24.4644 + 50.8266\left(\frac{x_1 - 29}{419}\right)^{6.2826} + 5.5514\left(\frac{x_2 - 23}{145}\right)^{-0.1988} - 63.7677e^{-1.5291\left(\frac{x_3 - 9}{357}\right)} - 28.1054e^{-0.9844\left(\frac{x_4}{119}\right)} + 19.7003\left(\frac{x_5 - 3}{356}\right)^{1.5651} \quad (3a)$$

4. 非线性稳健回归方程

$$\hat{y} = 19.9408 + 47.6239\left(\frac{x_1 - 29}{419}\right)^{6.2826} + 5.2632\left(\frac{x_2 - 23}{145}\right)^{-0.1988} - 65.4645e^{-1.5291\left(\frac{x_3 - 9}{357}\right)} - 28.868e^{-0.9844\left(\frac{x_4}{119}\right)} + 30.6623\left(\frac{x_5 - 3}{356}\right)^{1.5651} \quad (4a)$$

现把这4个方程的实况检验、统计检验和试报检验结果列于表1

表 1　　四个方程检验结果

项目 \ 方程	1a	2a	3a	4a
Q	9410.808	10110.78	5980.085	6513.353
R	0.7407	0.7177	0.8445	0.8292
F	6.5639	5.7357	13.4275	11.8861
σ	18.6695	19.3513	14.8824	15.5318
E	13.5005	13.0054	10.2788	9.3320
$\gamma_{拟}$	81.8	87.9	84.8	90.9
$\gamma_{试}$	50.0	50.0	100.0	100.0

注：Q为残差平方和；R为复相关系数；F为F检验值；σ为标准差；E为平均绝对误差；$\gamma_{拟}$和$\gamma_{试}$分别为拟合趋势和试报趋势正确百分率(%)。

（二）使用去掉“污染点”的资料，建立方程

我们对四个方程的残差分析发现：第19个样本点的残差都比较大，且这个点预报的趋势都相反，它很可能已受“污染”，因

此把这个样本点舍去，用32个样本点资料建立方程，这时得到的另外4个预报方程如下：

1． 一般线性回归方程

$$\hat{y} = -56.439 + 0.0496x_1 - 0.1476x_2 + 0.1798x_3 + 0.2393x_4 + 0.116x_5 \quad (1b)$$

2． 线性稳健回归方程

$$\hat{y} = -51.3839 + 0.0386x_1 - 0.2024x_2 + 0.1715x_3 + 0.2874x_4 + 0.1082x_5 \quad (2b)$$

3． 一般非线性回归方程

$$\hat{y} = 31.7006 + 47.0633\left(\frac{x_1 - 29}{419}\right)^{0.3512} + 2.5003\left(\frac{x_2 - 23}{145}\right)^{-0.2256} - 73.4745e^{-1.6607\left(\frac{x_3 - 9}{357}\right)} - 28.6391e^{-0.9844\left(\frac{x_4}{119}\right)} + 32.2201\left(\frac{x_5 - 3}{356}\right)^{1.8873} \quad (3b)$$

4． 非线性稳健回归方程

$$\hat{y} = 25.4029 + 48.4715\left(\frac{x_1 - 29}{419}\right)^{0.3512} + 4.0072\left(\frac{x_2 - 23}{145}\right)^{-0.2256} - 72.4576e^{-1.6607\left(\frac{x_3 - 9}{357}\right)} - 26.6236e^{-0.9844\left(\frac{x_4}{119}\right)} + 29.6339\left(\frac{x_5 - 3}{356}\right)^{1.8873} \quad (4b)$$

现把这4个方程的实况检验、统计检验和试报检验结果列于表2。

效果比较：

表1和表2中8个方程都通过显著性检验（$R > R_{0.01} = 0.57$，$F > F_{0.01} = 3.79$），因我们建立方程目的在于预报，而试报样品才2个，不能说明问题，只可做参考，因此，看方程好坏就看回代时平均绝对误差和拟合的趋势正确百分率。现先做线性和非线性、全部点和去掉“污染”点、一般和稳健三个方面的比较（每种都有4个方程），比较结果见表3。

表 2 去掉“污染”点后4个方程检验结果

项目 \ 方程	1b	2b	3b	4b
Q	7274.576	7473.235	4055.134	4541.472
R	0.8027	0.7966	0.8954	0.8829
F	9.4215	9.0328	21.0298	18.2209
σ	16.7270	16.9538	12.4887	13.2164
E	12.5111	12.1112	8.9695	8.3674
Y拟	81.3	84.4	84.4	90.6
Y试	50.0	50.0	100.0	100.0

注：说明同表1

表 3 各种方程效果比较

回归方程	方程序号	平均绝对误差和	拟合趋势正确百分率
线性	(1a)(2a)(1b)(2b)	51.1282	83.8%
非线性	(3a)(4a)(3b)(4b)	36.9477	87.7%
全部点	(1a)(2a)(3a)(4a)	46.1167	86.4%
去掉污染点	(1b)(2b)(3b)(4b)	41.9592	85.2%
一般	(1a)(3a)(1b)(3b)	45.2599	83.1%
稳健	(2a)(4a)(2b)(4b)	42.8160	88.5%

从表3比较结果看出：平均绝对误差和，非线性比线性减少27.7%，去掉“污染”点比全部点减少11.2%，稳健比一般减少5.4%；拟合趋势正确百分率，稳健比一般提高5.4%，非线性比线性提高3.9%，而去掉“污染”点没有提高，反而下降1.2%。从此看出，要提高预报准确率主要矛盾还是线性与非线性的矛盾，而稳健回归比一般回归有更高的拟合趋势正确百分率。

现再从单个方程比较，从各种检验结果综合来看，这8个方程中去掉“污染”点的非线性稳健回归方程(4b)为最好，其平均绝对误差最小，比一般线性回归减少33.1%，其拟合趋势准确率为90.6%，比一般线性回归提高9.3%，根据两年试报，平均绝对误

差比一般回归减少40.7%，试报趋势完全正确，按省评分办法均为100分。

六、结语

1. 通过上面实例计算结果说明，非线性稳健回归模型是较好的多元统计预报模型，它比一般线性回归模型的预报准确率有较大的提高，拟合和试报效果都较好，用不同的数据试验，也取得同样的效果，只是提高程度不同而已。

2. 对于回归模型的数据是否存在受"污染"的数据和是否符合正态分布假设的问题，开始是不知道的，通过各种回归模型的残差分析就可发现，如发现残差较大的点就可能是"污染"点，可以把它剔除，这样建立的方程会有更好的效果。同时，我们通过多功能多元回归程序，把各种类型的结果都计算出来，最后经过比较，选出最优预报方程，这样效果较好。

3. 现在用最小绝对偏差L_1估计作为稳健估计，这种稳健估计的优点是对偏离统计假设的情况远远没有经典统计的最小二乘估计方法那么敏感；对假设的微小偏差并不影响方法的性能，较大的偏差也不致于导致荒谬的效果。同时最小绝对偏差对预报的要求来说也是合理的，至于是否还有更好的稳健估计，有待今后进一步研究。

参考文献

[1] 方开泰等编，实用回归分析，科学出版社，1988年10月。

[2] 俞善贤等，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年8月。

[3] 范鸣玉等编，最优化技术基础，清华大学出版社，1982年。

[4] 冯耀煌、杨旭，论最优预报因子与最优预报方程，气象学报，第47卷第1期，1988年2月。

An application of the nonlinear robust regression to the weather forecastings

Feng Yaohuang

(Institute of Meteorology, Shenyang Center)

Abstract

The multivariate regression model is often used in the statistical forecasts. But the forecast accuracy is influenced by both non-linearity of predictors and the non-normality of the date. The model of nonlinear robust regression is introduced in this paper, in order to improve the forecast accuracy greatly.