0577-6619/2008/66(4)-0547-54 *Acta Meteorologica Sinica* 气象学报

复共线性关系对逐步回归预报方程的影响研究*

金 龙 黄小燕 史旭明
JIN Long HUANG Xiaoyan SHI Xuming

广西气象减灾研究所，南宁，530022
Guangxi Research Institute of Meteorological Disasters Mitigation, *Nanning* 530022, *China*
2007-07-31 收稿，2007-12-29 改回.

Jin Long, Huang Xiaoyan, Shi Xuming. 2008. A study on impact of multicollinearity on stepwise regression prediction equation. ***Acta Meteorologica Sinica*****, 66(4):547－554**

Abstract The accuracy of traditional stepwise regression meteorological prediction equation (SRMPE) is limited by the existence of multicollinearity among predictors of the equation, this paper introduces conditional number into the prediction modeling to minimize it in the traditional SRMPE. In the prediction modeling of novel SRMPE, the conditional number is used to determine the predictor set which has the lowest multicollinearity among the possible sets from a number of preliminary screening-out predictors (independent variables), and is then used to construct the novel SRMPE. The novel prediction modeling based on condition number is exampled with typhoon track prediction, which is a well known nodus in meteorological disaster prediction. 12 typhoons track latitude/longitude stepwise regression prediction equations have been built employing both the traditional and novel prediction modeling methods, respectively, but using a large number of identical samples. And the comparison and analysis results indicate that under the condition of same predictors (independent variable) and predictands (dependent variables), despite the fitting accuracy of typhoon tracks of the novel prediction model to the historical modeling samples is slightly lower than that of the traditional model, the prediction accuracy to the independent samples is obviously improved, with an averaged prediction error of the novel model for July, August, and September being 153.9 km, 75.3 km smaller than that of the tradition model (a reduction of 33%), due to the effectively minimizing of multicollinearity by the computation and analysis of condition number in modeling. It is further shown that when $F=1.0$, 2.0 and 3.0, the prediction errors of the traditional stepwise regression prediction equations are also obviously larger than those of the novel model. Furthermore, the extremely large/unreasonable errors occurred at the individual points of typhoon tracks in the independent sample prediction experiments of the traditional prediction model due to the impact of the multicollinearity in its predictor set.
Key words Multicollinearity, Meteorological prediction, Stepwise regression

摘 要 针对气象预报中常用的逐步回归预报建模方法，由于没有直接考虑筛选出的预报因子之间可能存在复共线性关系会影响气象预报方程的预报性能问题，提出了在初选的大量气象预报因子(自变量)中，采用条件数计算分析方法，选择复共线性关系小的预报因子组合建立预报模型的方法。以重要气象灾害的预报难点——台风预报为例，用大样本分别建立了12个台风移动经度、纬度的条件数预报方程和逐步回归预报方程。对比分析结果表明，由于条件数计算分析有效控制了预报因子间的复共线性关系，因此，在相同的预报因子(自变量)和预报对象(因变量)条件下，分月建立的条件数台风移动路径预报方程，虽然历史建模样本的拟合精度略低于逐步回归预报方程，但是对独立样本的预报精度明显提高，其中7、8和9月条件数预报方程的预报误差平均为153.9 km，而相应的逐步回归预报误差平均为229.2 km，两者相差75.3 km。进一步研究发现，在F值分别取1.0、2.0和3.0的情况下，建立的台风移动路径的逐步回归预报方程，其预报误差也明显大于条件数预报方程。另外，由于预报因子组合的复共线性的影响，逐步回归方程还出现了在个别点预报误差极大的不合理情况。
关键词 复共线性，气象预报，逐步回归
中图法分类号 P457.8

---

* 资助课题：国家自然科学基金项目(40675023)和国家科技部社会公益性研究专项(2004CB418306)。
作者简介：金龙，主要从事非线性人工智能气象预报技术研究。E-mail: jinlong01@163.com

1 引 言

多元回归分析是众多学科领域最常用的统计预报建模方法,尤其是在大气学科中,很多省和地市气象台站制作气象预报时,最为常用的预报方法是采用逐步回归预报建模方法(周家斌等, 1997;谢炯光等, 2003;姚愚等, 2004;高洁等, 2005),这主要是因为逐步回归预报方法的理论基础十分成熟,通用性好,没有可调参数,客观方便。同时,在气象预报问题中,对一般的降水或气温等各种气象要素(因变量)的预报都能较为方便地找到很多与之有很好相关关系的预报因子(自变量)。而如何从这些众多相关预报因子中筛选出最好的预报因子组合对预报量作未来状况的预测,虽然有各种不同的方法,但是相对而言逐步回归方法方便、快捷和有效是较为公认的。然而,在很多实际气象回归预报方程的建立过程中,一般都是直接采用逐步回归方法筛选预报因子建立预报方程(金龙等, 2003;刘还珠, 2004;陈豫英, 2006;刘锦銮, 2006),很少考虑由逐步回归方法筛选获得的预报因子组合是否具有最好的预报能力,以及会对回归参数估计产生什么影响。为此,本文试图通过大样本的台风移动路径气象预报问题,采用条件数值计算方法,分析讨论一般的逐步回归气象预报建模方法,由于筛选出的预报因子如果存在明显的复共线性关系,会导致预报量(因变量)的预报误差显著增大,而严重影响实际气象预报的准确性问题。

2 回归模型的变量选择和参数估计

在气象预报中,最为常用的线性回归模型为

$$Y = a_0 + X\beta + \varepsilon \tag{1}$$

其中 $X=(X_1\cdots X_p)$ 为 $n\times p$ 的预报因子矩阵,Y 为 $p\times 1$ 的预报量向量,β 为 $p\times 1$ 的回归系数向量,ε 为 $n\times 1$ 的误差向量,a_0 为常数项。一般如果方程(1)中预报因子(自变量)矩阵由某种形式确定后,则采用一般的最小二乘估计(LS)方法就能方便地确定出回归系数 β 的参数估计值 $\hat{\beta}$。然而,实际的大量回归分析问题,都会面临如何从众多预报因子矩阵中选择最合适的一组预报因子子集来建立预报能力最强的回归方程问题。在实际气象预报工作中,采用逐步回归筛选预报因子时,由于逐步回归预报方法本身并不直接提供甄别选出的预报因子组的自变量间是否存在明显的复共线性关系,因此讨论这种复共线性关系可能对回归方程预报能力的影响显得十分重要。一般逐步回归方法主要是通过计算分析预报因子(自变量)的相关矩阵,再根据一定的显著性检验标准,从全部自变量中选择一个使剩余方差下降最多,即方差贡献最大的自变量引进方程。由于逐步回归方法是根据显著性检验标准逐个引进变量(施能, 1995),因此,在引进后面的自变量时,可能会出现其中的变量方差贡献又不显著的情况,需要进一步计算每一个引入方程的变量的方差贡献,并将其中方差贡献最小的变量作显著性检验,确定是否剔除该变量。因此,可以看到逐步回归方法在选择因子时,主要是以预报因子的方差贡献作为标准。根据一定的显著性检验标准进行逐步的因子筛选后,便可获得最终的回归预报方程。对此,可以进一步分析预报因子组的变量之间复共线性对最小二乘估计的影响。

对于式(1)的线性回归模型的回归系数用最小二乘估计得到

$$\hat{\beta}_{\mathrm{LS}} = (X'X)^{-1}X'Y \tag{2}$$

设 $\hat{\beta}_{\mathrm{LS}}$ 为 β 的无偏估计,为了评价参数估计的好坏,可以采用以下公式

$$E\,\|\,\hat{\beta}_{\mathrm{LS}}-\beta\,\|^2 = \sigma^2\,\mathrm{tr}(X'X)^{-1} \tag{3}$$

其中 $\hat{\beta}_{\mathrm{LS}}=(X'X)^{-1}X'Y$,如果设 $X'X$ 的顺序特征根为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$,由于 $X'X$ 可逆,因此 $(X'X)^{-1}$ 的特征根可表示为:$\lambda_1^{-1},\lambda_2^{-1}\cdots\lambda_p^{-1}$,由此,式(3)可以写成

$$E\,\|\,\hat{\beta}_{\mathrm{LS}}-\beta\,\|^2 = \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i} \tag{4}$$

可以看到,如果 λ_k 很小,接近于零,则 $E\,\|\,\hat{\beta}_{LS}-\beta\,\|^2$ 可能就很大,这时候 $\hat{\beta}_{\mathrm{LS}}$ 的估计就会较差。假定 $X'X$ 的一个数值很小的特征根 λ_k 对应的标准正交化特征向量 C_k,由 λ_k 很小,接近于零,则有

$$X'XC_k = \lambda_k C_k \approx 0 \tag{5}$$

式(5)两边同乘 C'_k,得到

$$C'_kX'XC_k = C'_k\lambda_kC_k = \lambda_k \approx 0 \tag{6}$$

从而有

$$XC_k \approx 0 \tag{7}$$

因为预报因子阵 $X=(X_1\cdots X_p)$,$C_k=(C_{k_1},C_{k_2}\cdots C_{k_p})'$,因此有

$$C_{k_1}X_1+C_{k_2}X_2+\cdots+C_{k_p}X_P\approx 0 \qquad (8)$$

说明 X 的列向量 X_1,$X_2\cdots X_p$ 之间有近似的线性关系,而这种列向量间的多重共线性关系会导致用最小二乘方法作参数估计的性质变坏。

3　复共线性对回归模型的影响

由于在回归模型的自变量组合中,如果存在复共线性关系,会导致最小二乘参数估计性质变坏,这是否影响预报方程的预报能力,是本文想要进一步深入分析的。在上一节中看到,一般气象预报最常用的逐步回归预报建模方法,主要是考虑自变量因子的方差贡献,而并不直接考虑自变量组合的复共线性关系问题。为此可以通过直接采用逐步回归分析建立预报模型和采用条件数计算分析、诊断选择复共线性关系小的自变量因子组合建立预报模型后,来比较两种预报模型的性能差异。对此需要观察,如果自变量组合中存在复共线性关系,建立的回归模型会出现什么情况。对于一般的回归方程式(1)可以写成如下形式

$$\begin{cases} Y_i=\alpha_i+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i \\ E(\varepsilon_i)=0,\ D(\varepsilon_i)=\sigma^2 \\ b_{ij}=Cov(\varepsilon_i,\varepsilon_j)=0 \quad (i\neq j) \end{cases} \qquad (9)$$

如果假设该回归模型的设计矩阵(x_{ij})的自变量组合存在复共线性关系,并且再不妨假定该自变量组合中,第 1 至 $p-1$ 个自变量间为线性无关,只是将自变量 X_p 加入后,该自变量组合变成线性相关,由式(8)应该有

$$C_{k_1}X_1+C_{k_2}X_2+\cdots+C_{k_p}X_p=0$$

则一定有 $C_{k_p}\neq 0$,可以得到

$$X_p=\frac{C_{k_1}}{C_{k_p}}X_1+\cdots+\frac{C_{k_{p-1}}}{C_{k_p}}X_{p-1} \qquad (10)$$

将式(10)代入式(9)可以得到

$$Y_i=\alpha_i+(\beta_1+\frac{C_{k_1}}{C_{k_p}})x_{i1}+\cdots+(\beta_{p-1}+\frac{C_{k_{p-1}}}{C_{k_p}})x_{ip-1}+\varepsilon_i \qquad (11)$$

式中 $i=1,2\cdots n$,由此可以看到,在前 $p-1$ 个变量中,再加入第 p 个变量 X_p 是没有必要的,因为前 $p-1$个变量已包含了全部的信息。所以,在建立回归模型时,进行自变量组合的复共线性诊断,并剔除没有预报信息的预报因子是非常重要的。

4　计算实例分析

由前面两节的分析可以看到,逐步回归分析方法与复共线性分析方法是两种不同的分析方法,由此得到的回归预报模型效果如何,本节将通过实例来进行计算分析。目前,关于自变量组合的复共线性诊断分析,已有不少有效方法(Rice, 1966;陈希孺等, 1982;Walker, 1989),如特征分析法,方差扩大因子以及条件数方法等。本文将采用较为方便的条件数方法对重要气象灾害的台风路径预报问题进行研究,并进一步与逐步回归预报模型进行对比分析。实际的回归模型 p 个自变量组合设计矩阵$X'X$的条件数可以由下面的简单公式计算得出

$$k=\lambda_1/\lambda_p \qquad (12)$$

其中 k 即为条件数,$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$ 为设计矩阵由大到小顺序排列的 p 个特征根。

首先根据式(12)计算台风路径预报模型的自变量组合的条件数,据此选择出 k 值小的自变量组合来建立回归预报模型。台风作为最重要的自然灾害之一,其未来移动路径的预报,是防御这种气象灾害影响的重要手段和方法,并且这也是气象灾害预报的重点和难点(Davidson,et al, 1993;吕纯濂等, 1996)。这主要是因为台风的移动路径变化不仅与其自身的强度、能量积累补充等因素有关,还与台风外围环境流场的非线性变化密切相关,十分复杂。因此,本文以中国气象局出版的南中国海海域 1960—2003 年 44 a 台风资料为基础,将在南中国海海域生成或进入的具有 48 h 以上生命史的台风个例作为预报研究对象。预报因子主要考虑两类,第一类为表征台风自身变化的气候持续因子,即包括台风前 12 和 24 小时的经度位置变化、纬度位置变化和强度变化等。第二类是根据美国 NCEP/NCAR 全球再分析资料的物理量场计算得出的代表台风周围环境流场的预报因子。并且,以 1960—1989 年 30 a 的台风个例作为建模样本,1990—2003 年 14 a 的台风个例作为独立样本进行预报试验。并且规定进入南中国海海域第一时刻起,每间隔 12 h 的台风位置作为一个统计样本,这样所得到的各个月台风移动预报建模样本和独立预报样本的个数列于表 1。

表1 各月建模样本和独立预报样本容量

Table 1 Sizes of the modeling and independence samples of typhoon track prediction experiments for July, August and September

	7月	8月	9月
建模样本(1960—1989年)	330	462	490
独立预报样本(1990—2003年)	154	151	163

对台风移动路径的预报,实际是要预报未来24小时台风的经度和纬度,因此对于7、8和9月每个月的预报均要建立两个预报方程,3个月共需要建立6个预报方程。表2给出了各个月台风移动经度和纬度为预报对象的初选获得的各气候持续预报因子和由美国NCEP/NCAR全球再分析资料得到的前24小时大气层各层位势高度、水汽通量、等压面比湿、风场、涡度等大气物理量预报因子。在表2中根据气候持续法原理(谢玲娟, 1989;薛宗元等, 1995)初选的预报因子是以达到或超过0.01置信度水平作为预报因子的选择标准。而NCEP/NCAR数值预报产品物理量预报因子,同样以0.01置信度水平作为初选预报因子的标准。因此,最终对各月台风移动经度和纬度预报进行初选获得的预报因子个数列于表2。

表2 气候持续因子和物理量因子个数统计

Table 2 Numbers of CLIPER and numerical weather prediction production predictors for July, August and September

	7月	8月	9月
经度预报因子	46	57	40
纬度预报因子	100	62	75

从上述预报对象初选获得的预报因子统计得到,每个预报量(共6个),不仅有从40个到100个数量众多的预报因子,并且这些预报因子是从台风自身变化和周围环境流场变化两个方面计算分析得出的,其中每个预报对象都有两个以上的相关系数在0.7至0.9(样本长度在330—490)的高相关预报因子。

以下,将利用各预报对象初选得到的预报因子组,分别采用条件数计算方法,选择预报因子子集建立回归预报方程和采用逐步回归分析方法筛选预报因子建立预报模型进行对比分析。

在采用条件数计算方法选择最终的预报因子组合建立台风移动经度和纬度的最终预报方程的步骤为:

(1) 从某预报对象的全部 m 个初选预报因子中,先选择第1列和第2列两个预报因子 x_1、x_2,令 $x=(x_1,x_2)$,计算 $x'x$ 的特征根,再根据条件数分析的原则考察它们之间的复共线性严重程度,如果所计算得到的 k 值较大,表明复共线性明显,则只取其中一个预报因子,如果 k 值较小,则将两个预报因子全部保留。

(2) 再从其余 $m-2$ 个预报因子中,选择一个预报因子 x_3,构成新的预报因子子集 $x=(x_1,x_2,x_3)$,计算该预报因子组合矩阵的特征根和 k 值,同样考察预报因子之间是否存在复共线性关系,如果存在,则剔除 x_3,否则将预报因子 x_3 保留。

(3) 按上述步骤,将全部 m 个预报因子复共线性关系较小的预报因子保留,并以此作为最终选定的预报因子组合,构建新的设计矩阵。再利用最小二乘方法作出回归参数的估计,建立预报方程。

根据上述条件数计算选择预报因子的方法,设定 k 值不大于500作为最终选择入选预报因子的界限,在7月南海台风移动经度预报方程中共选入16个预报因子;而7月的纬度预报方程共选入17个预报因子。再用回归分析方法,建立了南海台风7月移动经度和纬度的两个预报方程(13)和(14)。

$$\begin{aligned} Y_{24(7)} = {} & 11.2356 + 0.0059x_{32}^{*} + 0.026x_7 - \\ & 0.0865x_{12}^{*} + 0.0094x_{33}^{*} - 0.0479x_8 - \\ & 0.0037x_{31}^{*} + 0.0256x_2 + 94.3871x_{28}^{*} + \\ & 94.3923x_{23}^{*} - 0.0013x_{34}^{*} - 0.0067x_{14}^{*} - \\ & 0.0575x_{29}^{*} - 94.3573x_{25}^{*} + 2.9484x_1 - \\ & 2.0349x_{11} + 94.4689x_{26}^{*} \\ & (\sigma = 1.5301,\ R = 0.9318) \end{aligned} \tag{13}$$

$$\begin{aligned} X_{24(7)} = {} & 3.3409 + 0.0086x_{33}^{*} - 19.2891x_5 + \\ & 0.0019x_8 + 0.0059x_{60}^{*} - 0.0047x_5^{*} - \\ & 0.0035x_{58}^{*} + 0.0039x_{61}^{*} + 0.0249x_4 + \\ & 0.0045x_{56}^{*} - 0.0139x_{24} + 0.0076x_{28} + \\ & 0.0941x_{13}^{*} - 0.0076x_{25} - 0.0025x_{57}^{*} + \\ & 178.8577x_1 - 177.0504x_{18} - 0.9617x_{33} \\ & (\sigma = 0.9839,\ R = 0.9359) \end{aligned} \tag{14}$$

式中 σ 表示剩余标准差,R 表示复相关系数,加 $*$ 表示物理量预报因子,其余因子为气候持续预报因子。

并且根据上述同样的方法,再分别建立了8月

和9月南海台风移动经度和移动纬度的24小时预报方程(预报方程略)。

利用建立的6个条件数预报方程分别对1990—2003年14 a的南海台风个例进行独立样本的预报试验。并且在实际的独立样本预报中,采用了逐次的预报试验方法,即在根据1960—1989年7月330个台风移动经度建模样本建立方程(13),利用该预报方程作出第331个独立样本预报后,再进行第332个独立样本预报时,又将第331个台风经度实况值加入前330个建模样本中,以此类推,直到用483个样本建立预报方程对最后第484个样本做出预报。

根据这样的预报步骤,表3给出了由6个条件数预报方程分别预报得出的南海台风7月(154次),8月(151次),和9月(163次)移动经度、纬度逐次预报的平均绝对误差和平均相对误差。在表3中给出的预报距离(km)平均误差是根据预报的经度、纬度值与台风实际的经度、纬度值的差值,按以下的计算公式换算得出

$$\Delta S=\frac{1}{n}\sum_{i=1}^{n}\sqrt{(y_i-\hat{y}_i)^2+(x_i-\hat{x}_i)^2}\times 110 \qquad (15)$$

其中,110(km)为一个经度、纬度的大致平均距离,$\hat{y}_i$ 为预测的台风经度,$\hat{x}_i$ 为预测的台风纬度,y_i 和 x_i 为实测的台风经度、纬度。由表3可以看到,采用条件数计算方法选择预报因子建立的7月台风路径预报方程对154次逐次预报的平均误差距离为162.1 km;采用同样的条件数方法对8月(151次)、9月(163次)逐次预报的平均误差距离分别为145.9和153.8 km。

表3　条件数分析方法各月独立样本预报结果

Table 3　Prediction results of the stepwise regression meteorological prediction equation based on condition number analysis for the independence samples for July, August and September

	7月		8月		9月	
	绝对误差(度)	相对误差	绝对误差(度)	相对误差	绝对误差(度)	相对误差
经度	1.1996	1.0711%	0.9132	0.8205%	1.2091	1.0598%
纬度	0.8564	4.1033%	0.9616	4.6876%	0.7017	3.5493%
距离	162.1 km		145.9 km		153.8 km	

为了分析复共线性关系对回归预报方程性能的影响,进一步以表2相同的初选预报因子和预报量,直接采用逐步回归方法,进行预报因子筛选,并建立逐步回归方程。并且为了进行对比分析,在建立7、8和9月台风移动经度和纬度6个逐步回归预报方程时,首先通过调整 F 值,将各个逐步回归预报方程入选的预报因子个数与条件数计算分析得出的预报方程的预报因子个数相当。例如,条件数计算分析筛选确定的7月纬度预报方程的预报因子个数为17个。因此,在建立逐步回归预报方程时,对于7月纬度预报方程,F 值取2.5时逐步回归方程也选择了17个预报因子建立预报方程(是从全部100个预报因子中自动筛选出17个预报因子);而8月纬度预报方程由条件数计算分析选择了7个预报因子,所以逐步回归计算取 F 值为4.0,则也自动筛选出7个预报因子建立预报方程(两种方法均是以同样的62个初选预报因子为基础),其余方程类似。表4给出了根据各个预报量的全部初选预报因子得出的预报方程、拟合平均误差、复相关系数等方程指标(表4中也同时给出了条件数计算分析选择的预报因子组建的预报方程各项指标)。由表4可以得出以下一些主要差异:在条件数选择的预报因子方程和逐步回归筛选预报因子的对应预报方程(共6对12个方程)中,(1)条件数方程的复相关系数均小于逐步回归方程,剩余标准差均为前者大于后者。(2)条件数方程对历史样本的拟合平均相对误差和拟合平均绝对误差均大于相应的逐步回归方程;相反条件数方程的独立样本预报平均相对误差、平均绝对误差无一例外均小于相应的逐步回归方程。特别是8和9月台风移动纬度预报误差,逐步回归方程几乎是条件数方程预报误差的2和1倍;实际上由图1可以看到8和9月逐步回归方程的纬度预报出现了异乎寻常的跳越式极端预报误差,这实际是方程参数估计造成的"病态"反映。由此不难看出虽然是基于相同的初选预报因子,但是如果不考虑最终选入方程的预报因子组的复共线性问题,直接采

表 4 条件数方程和逐步回归方程预报性能的比较

Table 4 Prediction performances of stepwise regression prediction equations with and without condition number analysis for July, August and September

			复相关系数	剩余标准差	预报因子数	拟合平均相对误差(%)	预报平均相对误差(%)	拟合平均绝对误差	预报平均绝对误差	预报样本
月	经度	条件数	0.9318	1.5301	16	1.0028	1.0711	1.1199	1.1996	154
		逐步回归	0.9364	1.4819	17	0.9603	1.1434	1.0720	1.2808	154
	纬度	条件数	0.9359	0.9839	17	3.7896	4.1033	0.7513	0.8564	154
		逐步回归	0.9479	0.8904	17	3.3499	4.5353	0.6643	0.9411	154
月	经度	条件数	0.9559	1.4778	16	1.0228	0.8205	1.1495	0.9132	151
		逐步回归	0.9633	1.3531	17	0.9252	1.1298	1.0388	1.2713	151
	纬度	条件数	0.9030	1.0281	7	4.1269	4.6876	0.8026	0.9616	151
		逐步回归	0.9068	1.0088	7	4.0925	12.9447	0.7141	2.5850	151
月	经度	条件数	0.9600	1.2982	5	0.8676	1.0598	0.9742	1.2091	163
		逐步回归	0.9626	1.2566	5	0.8470	1.1957	0.9508	1.3661	163
	纬度	条件数	0.9445	0.9294	19	3.8866	3.5493	0.6930	0.7017	163
		逐步回归	0.9507	0.8770	18	3.6994	6.1762	0.6652	1.2050	163

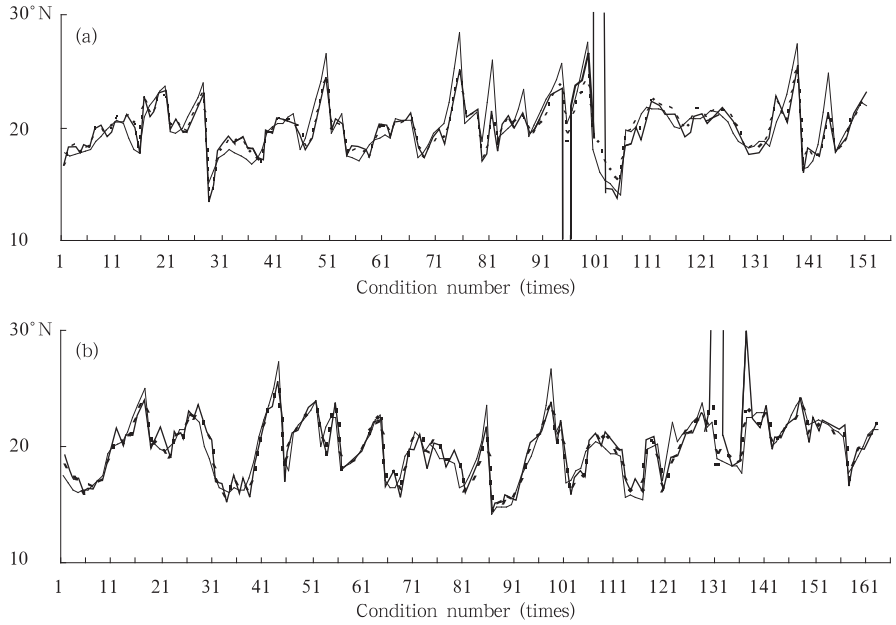

图 1 8(a)和 9 月(b)条件数和逐步回归方法台风移动纬度预报

(粗实线为实况值;细实线为条件数预报值;虚线为逐步回归预报值)

Fig. 1 Latitude prediction of August (a) and September (b) stepwise regression prediction equations with (dash line) and without condition number analysis (thin solid line) (thick solid line: observed value)

用逐步回归方法筛选预报因子建立预报方程,不仅会使得实际的预报误差明显偏大,甚至出现难以忍受的极端预报误差,这显然是不可取的。表 5 还进一步给出了分别由条件数方法和逐步回归方法建立的 7 月经度和纬度预报方程所最终选入的预报因子情况。虽然两种方法依据的初选预报因子是相同的,但是最终选入方程的因子多数是不同的,而正是这种不同导致了上述两种预报方程产生了显著的差异。

上述对比分析,主要是基于预报量相同,初选的预报因子相同,以及最终选入的预报因子个数基本相同的条件进行的相同独立样本预报的预报试验。为了进一步作客观对比分析,再根据这相同的 6 个预报量初选得出预报因子群,调整 F 值(F 值分别取 1.0、2.0 和 3.0)建立相应的逐步回归预报方程,

表 5　7 月台风经度、纬度条件数方法和逐步回归方法入选因子

Table 5　Predictors of July stepwise regression prediction equations with and without condition number analysis

	方法	入选因子
经度	条件数法	$x_1\ x_2\ x_7\ x_8\ \ x_{11}\ x_{12}^*\ x_{14}^*\ x_{23}^*\ x_{25}^*\ x_{26}^*\ x_{28}^*\ x_{29}^*\ x_{31}^*\ x_{32}^*\ x_{33}^*\ x_{34}^*$
	逐步回归	$x_1\ x_3\ x_8\ x_{10}\ x_{11}\ x_2^*\ x_5^*\ x_7^*\ x_8^*\ x_9^*\ x_{12}^*\ x_{13}^*\ x_{19}^*\ x_{20}^*\ x_{24}^*\ x_{26}^*\ x_{29}^*$
纬度	条件数法	$x_1\ x_4\ x_5\ x_8\ \ x_{18}\ x_{24}\ x_{25}\ x_{28}\ x_{33}\ x_5^*\ x_{13}^*\ x_{33}^*\ x_{56}^*\ x_{57}^*\ x_{58}^*\ x_{60}^*\ x_{61}^*$
	逐步回归	$x_1\ x_4\ x_6\ x_{27}\ x_{30}\ x_{33}\ x_3^*\ x_{13}^*\ x_{21}^*\ x_{22}^*\ x_{27}^*\ x_{35}^*\ x_{41}^*\ x_{48}^*\ x_{52}^*\ x_{59}^*\ x_{61}^*$

比较不同的 F 值条件下,这 18 个逐步回归预报方程与条件数预报方程的性能差异(表 6)。由表 6 可以清楚地看到,两种方法的预报方程预报性能差异与表 4 结论完全一样,即逐步回归方程对历史样本的拟合效果、复相关系数、剩余标准差均好于条件数方程,但是条件数方程的独立样本的预报效果却明显好于逐步回归方程,并且逐步回归方程中出现了极端的预报误差情况。因此可以看到在实际的气象预报工作中,在建立回归预报方程时,考察预报因子组合的复共线性关系十分重要,也是改进和提高气象预报模型预报能力的重要措施。其结论无论从理论分析和实践预报试验中都是十分清楚的。

表 6　条件数方程与不同预报因子组合回归方程预报性能的比较

Table 6　Prediction performances of condition number models and stepwise regression prediction equations (F=1.0, 2.0, 3.0) for July, August and September

				复相关系数	剩余标准差	预报因子数	拟合平均相对误差(%)	预报平均相对误差(%)	拟合平均绝对误差	预报平均绝对误差	预报样本
月	经度	条件数		0.9318	1.5301	16	1.0028	1.0711	1.1199	1.1996	154
		逐步回归	F=1.0	0.9356	1.4818	13	0.9763	1.1362	1.0898	1.2724	154
			F=2.0	0.9348	1.4855	11	0.9758	1.1230	1.0895	1.2565	154
			F=3.0	0.9338	1.4918	9	0.9805	1.1181	1.0951	1.2522	154
	纬度	条件数		0.9359	0.9839	17	3.7896	4.1033	0.7513	0.8564	154
		逐步回归	F=1.0	0.9525	0.8709	31	3.2262	4.7880	0.6392	1.0425	154
			F=2.0	0.9498	0.8815	22	3.3031	4.3596	0.6548	0.9519	154
			F=3.0	0.9456	0.9035	13	3.4115	4.1306	0.6771	0.9056	154
月	经度	条件数		0.9559	1.4778	16	1.0228	0.8205	1.1495	0.9132	151
		逐步回归	F=1.0	0.9637	1.3511	21	0.9166	1.4983	1.0288	1.6983	151
			F=2.0	0.9632	1.3531	17	0.9252	1.1158	1.0388	1.2538	151
			F=3.0	0.9623	1.3659	14	0.9464	1.0022	1.0622	1.1227	151
	纬度	条件数		0.9030	1.0281	7	4.1269	4.6876	0.8026	0.9616	151
		逐步回归	F=1.0	0.9147	0.9811	20	3.9272	10.6195	0.7607	2.2059	151
			F=2.0	0.9140	0.9815	17	3.9261	13.1939	0.7598	2.6523	151
			F=3.0	0.9097	0.9982	11	4.0218	12.1729	0.7785	2.5273	151
月	经度	条件数		0.9600	1.2982	5	0.8676	1.0598	0.9742	1.2091	163
		逐步回归	F=1.0	0.9646	1.2335	13	0.8292	1.1226	0.9313	1.2842	163
			F=2.0	0.9645	1.2342	12	0.8275	1.1241	0.9293	1.2858	163
			F=3.0	0.9639	1.2401	9	0.8341	1.1782	0.9363	1.3453	163
	纬度	条件数		0.9445	0.9294	19	3.8866	3.5493	0.6930	0.7017	163
		逐步回归	F=1.0	0.9516	0.8758	25	3.6605	7.1398	0.6576	1.4210	163
			F=2.0	0.9499	0.8806	15	3.7126	6.1750	0.6672	1.2049	163
			F=3.0	0.9493	0.8837	13	3.7370	6.2166	0.6718	1.2104	163

5　结　论

由于各种气象灾害的预报比较容易获得与预报对象相关密切的大量前期相关大气环流、气象要素的预报因子,但是如何从大量的预报因子中选出更好的预报因子组合,以建立预报能力更强的预报方程是提高气象预报准确性的重要研究课题。本文针对目前气象预报中大量采用逐步回归预报方法建立灾害性天气预报方程,可能存在预报因子间的复共线性关系影响预报方程的预报精度,提出采用条件

数计算分析方法,对气象预报因子进行预报因子的复共线性关系诊断分析,选择复共线性小的预报因子组合,建立预报方程以获得更好的预报效果,这对改进和提高气象客观预报方法的预报精度有重要意义。从理论分析和多个台风移动路径的预报方程的对比分析结果表明,用传统的逐步回归分析方法选择预报因子建立气象预报方程,确实可能由于预报因子间存在复共线性关系,导致预报方程预报准确率不高和个别点出现有悖常理的预报误差极大的情况。这提示我们在建立气象统计预报方程时,面对大量的初选预报因子,诊断分析预报因子间的复共线性关系是十分重要的。当然造成复共线性关系存在可能有很多不同的原因,如何进一步诊断分析造成不同预报因子(自变量)组合的复共线性原因,有待进一步深入探讨。

References

Chen Yuying, Chen Xiaoguang, Ma Jinren, et al. 2006. A study on subtle MOS forecasting method of wind. Scientia Meteor Sinica (in Chinese), 26(2):210-215

Chen Xiru, Wang Songgui. 1982. Modern regression analysis (in Chinese). Beijing:Science Press, 217-226

Davidson N E, Wadsley J, Puri K, et al. 1993. Implementation of the JMA typhoon bogus in the BMRC tropical prediction system. J Meteor Soc Japan, 71:437-467

Gao Jie, Liu Duanci, Jin Yingyan. 2005. A regression estimation of event probabilities for forecasting scatter of radiation fog at Xian yang Airport. Meteor Mon(in Chinese), 31(4):81-84

Rice J L. 1966. A theory of condition number. SIAM J Numer Anal, 3:287-310

Jin Long, Lin Xi, Jin Jian, et al. 2003. A numerical prediction product interpretation and application based on a modular fuzzy neural network. Acta Meteor Sinica(in Chinese), 61(1):78-83

Liu Huanzhu, Zhao Shengrong, Lu Zhishan, et al. 2004. Objective element forecasts at NMC: a MOS system. J Appl Meteor Sci (in Chinese), 15(2):181-191

Liu Jinluan, He Jian, Chen Xinguang. 2006. Agricultural weather forecast technical research of Guangdong province. Meteor Mon(in Chinese), 32(2):116-120

Lu Chunlian, Chen Shunhua, Zhu Yongti. 1996. Research and apply multiple dynamic interdependent models to predict typhoon track, intensity and wind-speed, simultaneously. Acta Meteor Sinica(in Chinese), 54(6):737-744

Shi Neng. 1995. The Multianalysis Method of Meteorological Research and Prediction(in Chinese). Beijing:China Meteorological Press, 66-90

Walker E. 1989. Detection of collinearity-influential observations. Comm Stat Theory Meth, 5:1675-1690

Xue Zongyuan, Li Zuofeng. 1995. The statistical dynamic operational model (SD-90) and test results of Western North Pacific Tropical Cyclone path. Res Appli Atmos Sci(in Chinese), 5:59

Xie Lingjuan. 1989. A climate-continuity model for the forecasts of the South Sea cyclone pathes. Marine Forecasts(in Chinese), 6(2): 20-30

Xie Jiongguang, Zeng Cong, Ji Zhongping. 2003. Statistical forecast of meteorology for the last 30 years in China. Meteor Sci Tech (in Chinese), 31(2):67-79

Yao Yu, Yan Huasheng, Cheng Jiangang. 2004. Relationship between subtropical high indexes at the main raining period with the rainfall of 160 sampling stations in China. J Tropical Meteor(in Chinese), 20(6):657-661

Zhou Jiabin, Huang Jiayou. 1997. Advances in statistical meteorology in China in recent years. Acta Meteor Sinica(in Chinese), 55(3):297-304

附中文参考文献

陈豫英,陈晓光,马金仁等. 2006. 风的精细化 MOS 预报方法研究. 气象科学, 26(2): 210-215

陈希孺,王松桂. 1982. 近代回归分析原理方法及应用. 北京:科学出版社, 217-226

高洁,刘端次,勒英燕. 2005. 用事件概率回归方法预报咸阳机场辐散雾消散. 气象, 31(4):81-84

金龙,林熙,金健等. 2003. 模块化模糊神经网络的数值预报产品释用预报研究. 气象学报, 61(1):78-83

刘还珠,赵声蓉,路志善等. 2004. 国家气象中心气象要素的客观预报:MOS 系统. 应用气象学报, 15(2):181-191

刘锦銮,何键,陈新光. 2006. 广东农用天气预报技术研究. 气象, 32(2):116-120

吕纯濂,陈舜华,朱永□. 1996. 多维动态关联模型在台风路径、强度和风速同时预报中的应用研究. 气象学报, 54(6):737-744

施能. 1995. 气象科研与预报中的多元分析方法. 北京:气象出版社, 66-90

薛宗元,李佐凤. 1995. 西北太平洋热带气旋路径的统计动力预报方案(SD-90)及其业务试验结果. 大气科学研究与应用, 5:59

谢玲娟. 1989. 南海台风路径预报的气候、持续性模式. 海洋预报, 6(2): 20-30

谢炯光,曾琮,纪忠萍. 2003. 中国近 30 年来气象统计预报进展. 气象科技, 31(2): 67-79

姚愚,严华生,程建刚. 2004. 主汛期(6-8 月)副高各指数与中国 160 站降雨的关系. 热带气象学报, 20(6):657-661

周家斌,黄嘉佑. 1997. 近年来中国气象统计学的新进展. 气象学报, 55(3): 297- 304