第18卷 第2期 大 气 科 学 Vol. 18, No. 2
1994 年 3 月 SCIENTIA ATMOSPHERICA SINICA March 1994

多元门限回归模型的一种建模方法

严华生
(云南省气象台，昆明 650034)

曹 杰
(云南省红河哈尼族彝族自治州气象局，
云南蒙自 661100)

提 要

本文根据门限自回归模型的基本思想[1]，提出一种多元门限回归模型的建模方法。其特点是充分考虑了预报系统中某些特殊预报因子突变点对预报关系的改变作用。数值实例表明，该模型在模拟和预报精度上比一般线性逐步回归模型有一定程度的提高。

关键词：多元门限回归；突变；建模。

一、预报原理及数学模型

考虑m个预报因子x_1，x_2，…，x_m与预报对象y之间的预报关系，若其中有某个特殊因子x_i，当其值低于某个水平时，y与x_1，x_2，…，x_m是一种预报关系；而当x_i高于某个水平时，y与x_1，x_2，…，x_m又是另一种预报关系；则称x_i为门限变量，导致y与x_1，x_2，…，x_m预报关系发生改变的x_i值称为门限值。例如：统计长期天气预报实践中就遇到大气中出现异常强讯号造成常规预报关系改变的问题。因此在统计预报中很有必要考虑预报系统中的突变、跳跃等非线性情况，多元门限回归模型就是为解决这一问题而提出的。其基本思想来源于时间序列分析中的门限自回归模型[1]，把它运用到多元分析[2,3]中就得出多元门限回归模型。

我们把预报系统中预报对象y和预报因子集x之间的两段门限回归模型定义为

$$y=\begin{cases}a_1^{(1)}x_1+a_2^{(1)}x_2+\cdots+a_m^{(1)}x_m+a_0^{(1)}, & \text{当} \quad x_i<b_1\\ a_1^{(2)}x_1+a_2^{(2)}x_2+\cdots+a_m^{(2)}x_m+a_0^{(2)}, & \text{当} \quad x_i\geqslant b_1\end{cases} \tag{1}$$

式中x_i即为门限变量，b_1即为门限值。

多元门限回归模型的基本思想是：当给出预报因子资料后，首先根据门限变量和门限值决定在不同情况下使用不同预报关系的方程，以此解释各种类似于突变的现象。其实质是，把预报问题按状态空间的取值进行分类，用分段的线性回归模型来描述总体非线性问题。

1991 年 8 月 2 日收到，1992 年 3 月 10 日收到再改稿。

二、多元门限回归模型的建模步骤及预报方法

由于在时间序列分析中门限自回归模型的建模方法不适用于多元门限回归模型，因此本文提出一种多元门限回归模型的建模方法，其中包括门限变量和门限值的确定等。现简述如下：

1. 应用最优分割法确定门限变量和门限值

其理论根据为：当大气中某些强影响因子发生变化时，必然使预报关系发生改变、导致生成的预报对象有显著不同；统计学根据为当预报系统中某个因子大于或小于、等于某一临界值时，其所分成的两组子样本的组间方差有最大差异。据此就可寻找能使预报对象分组产生最大差异的预报因子作为门限变量和门限值。具体作法如下：

a) 设有预报对象y和m个预报因子$x=\{x_1, x_2, \cdots, x_m\}$，共$n$个样本。分别从$x$中依次取因子$x_i$，$i=1,2,\cdots,m$。将$x_i=\{x_{i1}, x_{i2}, \cdots, x_{in}\}$按其值从小到大顺序排列，对$y$样本值也随$x_i$的顺序相应重新排列，得新序列$y'$。

b) 对y'进行二分割，将分成的两段分别记为：$y'(1,k)$，$y'(k+1,n)$。于是得y'序列的总方差

$$V^2=\sum_{i=1}^{n}(y'_i-\bar{y}')^2.$$

所分成两段的组内方差和为

$$S^2=\sum_{i=1}^{k}[y'_i-\bar{y}'(1,k)]^2+\sum_{i=k+1}^{n}[y'_i-\bar{y}'(k+1,n)]^2,$$

组间方差

$$B^2=V^2-S^2.$$

c) 用$F=[B^2(n-2)]/S^2$来检验两组间是否存在显著性差异。对序列y'共进行$L=n-1$次二分割，计算$n-1$个F值，记为：$F_1(x_i)$，$F_2(x_i)$，$\cdots$，$F_L(x_i)$，选择$F_k(x_i)=\max\{F_1(x_i), F_2(x_i), \cdots, F_L(x_i)\}$，，所对应的分割点$k$即为第$i$个因子的最优分割点。

d) 对所有因子x_i，$i=1,2,\cdots,m$分别求出$F_k(x_1)$，$F_k(x_2)$，$\cdots$，$F_k(x_m)$，再选择$F_k^*(x_i^*)=\max\{F_k(x_1), F_k(x_2), \cdots, F_k(x_m)\}$，这样就确定了最优分割因子$x_i^*$和最优分割点$k^*$。最优分割因子即为门限变量，对应的最优分割点值$x_{ik}^*$，即为门限值。

e) 将x_i^*按大于和小于、等于x_{ik}^*分为两段，其余预报因子和预报对象也相应地随x_i^*分为两段。

2. 对子样本组作逐步回归

将已分成的两组样本分别作逐步回归，可得门限回归模型

$$\hat{y}=\begin{cases}a_1^{(1)}x_1+a_2^{(1)}x_2+\cdots+a_m^{(1)}x_m+a_0^{(1)}, & x_i^*<x_{ik}^*,\\ a_1^{(2)}x_1+a_2^{(2)}x_2+\cdots+a_m^{(2)}x_m+a_0^{(2)}, & x_i^*\geqslant x_{ik}^*.\end{cases}\tag{1}$$

当给出预报因子资料后，首先根据门限变量和门限值判定其属于哪一段，然后用该段回归模型代入预报因子实时资料即可作出预报。

三、应用实例

本文用云南河口1954—1990年5月降水资料作为预报对象；1953—1989年1—12月和1954—1990年1—3月太阳黑子相对数及北半球500hPa月平均极涡中心强度、乌拉尔山地区平均高度、印缅地区五点高度和、中国南海副高强度指数等共75个与河口降雨有关的资料作为预报因子。

按前述建模步骤，确定出河口5月降水多元门限回归长期预报模型中的门限变量为x_{51}，即头年6月乌拉尔山地区五点平均高度，门限值为67.3（略去百位数），其对应的F值为14.53，通过$\alpha=0.001$显著性检验。

河口5月降水量的门限回归长期预报模型为

$$\hat{y}=\begin{cases}-4.820x_{14}-7.742x_{61}+7.332x_{63}+4.731x_{67}-8.431x_{71}\\ \qquad +2.526x_{74}+3460.355, & x_{51}\leqslant 67.3\\ -22.221x_{14}+1.111x_{32}-15.876x_{57}+11746.941, & x_{51}>67.3\end{cases}\tag{2}$$

其中第一段$F=18.05$，第二段$F=21.49$均通过$\alpha=0.001$显著性检验；总拟合误差均方差仅为32.40。

按一般的线性逐步回归建立的模型为

$$\hat{y}=-9.563x_{14}+11.667x_{11}-6.845x_{72}+7802.506,\tag{3}$$

其中$F=8.88$，虽通过显著性检验，但拟合误差均方差较大，其值为90.0，几乎是(2)式的三倍。我们将(2)式与(3)式的历史回代结果列入表1中。其中y为实测值，$\hat{y}$为预报值，$\overline{V}=|(y-\hat{y})/y|$为相对误差，趋势评定中"√"为正确，"×"为错误。

从表1可看出：多元门限回归模型回代趋势准确率为30／37，平均相对误差为13.4%，一般线性逐步回归模型回代趋势准确率为18／37，平均相对误差为33.0%。

根据河口历年5月降雨量的实测值，(2)式的历史回代值及门限变量x_{51}的历年值绘得一点聚图（图1）。

比较图1中所绘两条(2)式的模拟直线，当门限变量$x_{51}<67.3$与$x_{51}\leqslant 67.3$时，这两条模拟直线的斜率明显不同；再比较(2)式和(3)式以及(2)式各段模型中的入选预报因子，除x_{14}是共同具有外，其余入选预报因子均不相同，且(2)式与(3)式、(2)式各段模型间，预报因子x_{14}对整个预报系统的影响也不相同。上述诸现象说明：一旦强影响因子即门限变量（例如：本文中的x_{51}）的异常强讯号即门限值（例如：本文中的$x_{51}<67.3$或$x_{51}\leqslant 67.3$）出现，便造成了预报系统中常规预报关系的改变，使某些原来对预报系统影响不大的因子变成了主要影响因子，而一些原来对预报对

象影响较大的因子则变成了次要因子等等。这样，最终使得整个预报系统发生了突变跳跃，即造成了预报关系的显著不同。

表 1　两种预报模型历史回代检验表

年	y	$\hat{y}$		V		趋势	评定	年	y	$\hat{y}$		V		趋势	评定
		(2)	(3)	(2)	(3)	(2)	(3)			(2)	(3)	(2)	(3)	(2)	(3)
1954	241	262	201	8.7%	16.6%	✓	✓	1973	218	212	140	2.8%	35.8%	✓	×
1955	255	223	180	12.5%	29.4%	✓	×	1974	123	118	179	4.1%	45.5%	✓	×
1956	287	335	296	16.7%	3.1%	✓	✓	1975	93	128	130	37.6%	39.8%	×	×
1957	114	108	115	5.3%	0.9%	✓	✓	1976	349	302	321	13.5%	8.0%	✓	✓
1958	102	150	113	47.1%	10.8%	×	✓	1977	170	190	226	11.8%	32.9%	✓	×
1959	587	546	290	7.0%	50.6%	✓	×	1978	332	245	330	26.2%	0.6%	×	✓
1960	145	136	160	6.2%	10.3%	✓	✓	1979	193	202	275	4.7%	42.5%	✓	×
1961	98	120	166	22.4%	69.3%	×	×	1980	134	129	49	3.7%	63.4%	✓	×
1962	197	196	274	0.5%	39.1%	✓	×	1981	209	216	234	3.3%	12.0%	✓	✓
1963	97	92	261	5.2%	169.1%	✓	×	1982	122	137	274	12.3%	124.6%	✓	×
1964	146	151	164	3.4%	12.3%	✓	✓	1983	99	165	181	66.7%	82.8%	×	×
1965	174	177	182	1.7%	4.6%	✓	✓	1984	253	242	292	4.3%	15.4%	✓	✓
1966	168	144	164	14.3%	2.4%	✓	✓	1985	477	479	289	0.4%	39.4%	✓	×
1967	110	119	142	8.2%	29.1%	✓	×	1986	251	281	217	12.0%	13.5%	✓	✓
1968	148	110	167	25.7%	12.8%	×	✓	1987	108	90	166	16.7%	53.7%	✓	×
1969	144	135	137	6.3%	4.9%	✓	✓	1988	212	175	207	17.5 $	2.4%	✓	✓
1970	242	242	176	0.0%	27.3%	✓	×	1989	244	323	346	32.4%	41.8%	×	×
1971	259	222	209	14.3%	19.3%	✓	✓	1990	498	512	411	2.8%	17.5%	✓	✓
1972	226	189	142	16.4%	37.2%	✓	×								

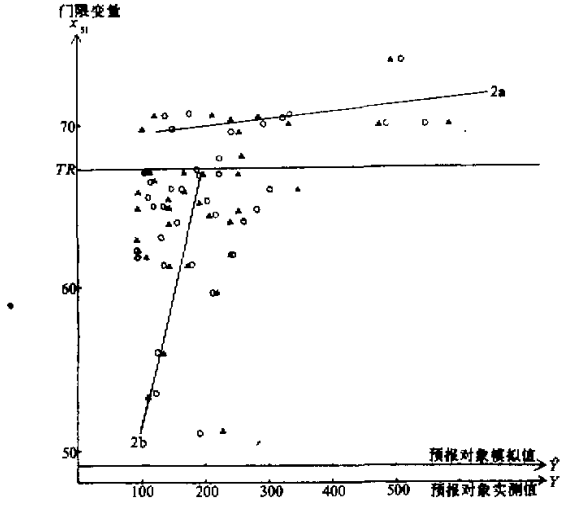

图 1　多元门限回归模拟图

"△"为实测值，"O"为 (2) 式历史回代值，TR为门限值.

2a 表示$x_{51} > 67.3$ 时所得模拟方程，2b 表示$x_{51} \leqslant 67.3$ 时所得模拟方程.

将此模式投入业务使用，于1991年4月作1991年河口5月雨量预报，得$\hat{y}$=157mm，趋势为偏少至特少。1991年5月雨量实况为147mm；若根据(3)式，则得$\hat{y}$=188mm。初次使用效果较好。可见，在长期预报中多元门限回归模型值得进一步探索试用。

四、工作展望

目前投入业务应用的统计预报模型多为线性模型，但自然界的现象是很复杂的，变量间的关系不可能仅用简单的线性关系就能描述，这或许是目前限制统计预报水平的一个因素，因此很有必要发展非线性统计预报。本文提出的多元门限回归模型就是非线性统计预报中很重要的一类分段线性化逼近方法。当样本数足够大时，还可把单门限变量两段多元门限回归模型拓广到多级门限(如(4)式)、多门限变量(如(5)式)、多级复合门限(如(6)式)的多元门限回归模型。

$$y=\begin{cases}a_1^{(1)}x_1+\cdots+a_m^{(1)}x_m+a_0^{(1)}, & x_i<b_2\\ a_1^{(2)}x_1+\cdots+a_m^{(2)}x_m+a_0^{(2)}, & b_2\leqslant x_i<b_3\\ a_1^{(l)}x_1+\cdots+a_m^{(l)}x_m+a_0^{(l)}, & b_l\leqslant x_i\end{cases}\tag{4}$$

$$y=\begin{cases}a_1^{(1)}x_1+\cdots+a_m^{(1)}x_m+a_0^{(1)}, & x_i\leqslant b_1,\ x_j\leqslant b_2\\ a_1^{(2)}x_1+\cdots+a_m^{(2)}x_m+a_0^{(2)}, & x_i\leqslant b_1,\ x_j>b_2\\ a_1^{(3)}x_1+\cdots+a_m^{(3)}x_m+a_0^{(3)}, & x_i>b_1,\ x_j\leqslant b_2\\ a_1^{(4)}x_1+\cdots+a_m^{(4)}x_m+a_0^{(4)}, & x_i>b_1,\ x_j>b_2\end{cases}\tag{5}$$

$$y=\begin{cases}a_1^{(1)}x_1+a_2^{(1)}x_2+\cdots+a_m^{(1)}x_m+a_0^{(1)}, & f_1(x)\\ \vdots\quad\vdots\quad\vdots\quad\vdots & \vdots\\ a_1^{(l)}x_1+a_2^{(l)}x_2+\cdots+a_m^{(l)}x_m+a_0^{(l)}, & f_l(x)\end{cases}\tag{6}$$

(6)式中$f_1(x),\cdots,f_l(x)$分别为各段模型的复合门限变量判别值。

(4)、(5)两式的建模方法可对前述两段门限变量多元门限回归建模思路进行拓广即得，(6)式可使用逐步判别来确定复合门限变量及门限值，再对每段分别用逐步回归建立方程。多元门限回归预报模型的主要特点是对处理预报问题的突变、跳跃等不连续或间断点有一定能力。

目前，对非线性统计预报模型的显著性检验是个还没有完全解决的科学问题。在本文中，仅只是对用门限变量把预报对象分段进行方差分析检验和分别对各段线性回归方程进行显著性检验，但对整个门限回归模型的显著性检验问题，从理论到方法都还有待进一步研究。

参 考 文 献

[1] 项静恬等，1986，动态数据处理，气象出版社。
[2] 张尧庭等，1983，多元分析引论，科学出版社。
[3] 严华生等，1991，多因变量及要素场统计预报，气象出版社。

Method of Building a Multivariate Threshold Regression Model

Yan Huasheng
(*Meteorological Bureau of Yunnan Province, Kunming* 650034)
Cao Jie
(*Honghe Hani and Yi Autonomous Prefecture Meteorological Bureau of Yunnan Province, Mengzi* 661100)

Abstract

In this paper, a method of building multivariate threshold regression model is given. The idea is mainly based on the fact that the relation between predictor and predictand will change when some special predictors change abruptly. Computational results have shown better effectiveness using multivariate threshold regression model than using general regression models.

Key words: multivariate threshold regression; catastrophe; model buile-up.

勘 误

本刊17卷6期的总目录中，由于编辑粗心大意，误将16卷1期的目录用作17卷1期的，特此勘误，并向广大读者致歉！

本刊编辑部