文章编号:1000 - 582X(2003)01 - 0062 - 04

组合预测模型的回归分析方法*

谢 开 贵,周 家 启
(重庆大学 电气工程学院,重庆 400044)

摘　要:给出求解组合预测权系数的回归分析方法,文章首先给出了基于最小二乘和最小一乘准则的线性回归组合预测模型,然后应用最小二乘原理得到权系数最小二乘估计值。由于最小一乘准则下,目标函数不可微,传统的优化规划方法无法求解,故文中提出用基于最小二乘的逐步变权方法进行求解。同时,还给出了百分误差绝对值最小为目标的组合预测模型及权系数求解方法。通过实例分析,表明组合预测模型的预测精度很高,回归效果很显著。

关键词:组合预测模型;最小二乘准则;最小一乘准则;逐步变权法
中图分类号:G303;O241.5　　**文献标识码**:A

组合预测方法是一种全新的预测方法,其理论和方法逐步完善、应用范围不断拓广[1-6]。对同一预测问题而言,由于考虑的角度、方式和层次等不同,可为其提供不同的预测方法,将这些方法进行组合,可增大信息量,能更好地进行预测。组合预测将各种预测效果进行总体性综合考虑,比单个预测模型更系统、更全面,且 Bates 和 Granger 证明 2 种或 2 种以上无偏的单项预测可以组合出优于每个单项的预测结果[1],即能有效地提高预测精度。

1　组合预测权系数的特性分析

组合预测方法关键是确定组合权系数[2-6]:文献[2-3]从正定矩阵的角度出发,给出了权系数公式 $K = \frac{E^{-1}R}{R^{\mathrm{T}}E^{-1}R^{\mathrm{T}}}$ 和最优权系数为非负的一个充分条件;文献[4]从数学规划的角度出发,给出权系数之和为 1 且非负的非线性规划解法。但笔者认为:

1) 权系数之和不需一定为 1。设 y_t 为第 t 个实际观测值($t = 1,2,\cdots,n$)。f_{it} 为第 i 中方法的第 t 个预测值($i = 1,2,\cdots,k$;$t = 1,2,\cdots,n$)。w_i 为第 i 种方法的权重。如果由于预测方法不当,使得 $f_{it} < y_t$(当 $f_{it} > y_t$ 时同理),若 $\sum_{i=1}^{k} w_i = 1$,那么组合预测值 $\hat{y}_t = \sum_{i=1}^{k} f_{it}w_i < y_t$,即预测结果一致偏小,这将难以接受(事实上,$f_{it}$ 一致比 y_t 大或小在实际中是不多见的,但 $e_{it} = y_t - f_{it}$ 不服从正态分布确是可能的,这就将使组合预测的结果一致偏大或偏小成为可能)。如果 $\sum_{i=1}^{k} w_i$ 为大于 1 的某值,此时组合预测的误差可能服从正态分布,而且误差平方和、误差绝对值之和、百分误差绝对值之和都可能比较小。通常只有当各种方法的预测误差都服从正态分布时,才能严格地满足 $\sum_{i=1}^{k} w_i = 1$,但在实践中,这一点难于做到,所以没有必要限制 $\sum_{i=1}^{k} w_i = 1$。同时,若确定的权系数 $\sum_{i=1}^{k} w_i > 1$,这表明各种方法的预测值总的来讲偏小,这时可以修正一些预测方法,再进行组合预测。

2) 权系数可以有正有负。组合预测值中各种预测方法在其中所起的作用不同,有的大、有的小;同时有的相对于其它来讲与原数据列有正相关关系,也有的有负相关关系,所以权重有正有负更与实际相符,换句话说,这里权重只是一个系数。

当权系数 w_i 有正有负,且其和没有等于 1 的条件限制时,则可根据每个 w_i 以及 $\sum_{i=1}^{k} w_i$ 的值对某些预测方法进行取舍和修正,以反过来检验各种预测方法的

* 收稿日期:2002 - 09 - 10
基金项目:重庆大学青年骨干教师资助基金
作者简介:谢开贵(1972 -),男,四川眉山人,重庆大学副教授,博士,从事电力系统规划与可靠性、电力市场、人工智能研究。

好坏。

为此，笔者从回归分析的角度给出了以误差平方和最小、误差绝对值之和最小、百分误差绝对值之和最小为目标的权系数求解方法。

2　基于最小二乘的组合预测模型

e_{it}、f_{it}、y_t、w_i 意义同前，从而有：

$$y_t = \sum_{i=1}^{k}(w_i f_{it} + e_{it}) \quad t = 1,2,\cdots,n \tag{1}$$

设
$$Q_{ls} = \sum_{t=1}^{n}\left(y_t - \sum_{i=1}^{k} w_i f_{it}\right)^2 \tag{2}$$

式(2)中 Q_{ls} 表示预测的误差平方和。当以 Q_{ls} 为目标函数时，问题转化为：

$$\min Q_{ls} = \sum_{t=1}^{n}\left(y_t - \sum_{i=1}^{k} w_i f_{it}\right)^2 = \sum_{t=1}^{n} e_t^2 \tag{3}$$

其中，$e_t = y_t - \sum_{i=1}^{k} w_i f_{it}$，即时刻 t 的组合预测误差。

设，$X = (f_{ij})_{k\times n}$，$Y = [y_1, y_2, \cdots, y_n]^{\mathrm{T}}$

由最小二乘回归方法知：

权系数为：$\hat{W} = [\hat{w}_1, \hat{w}_2, \cdots, \hat{w}_k]^{\mathrm{T}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ (4)

3　基于最小一乘的组合预测模型

评价一种预测方法好坏的准则很多，主要有误差平方和最小、误差绝对值之和最小、百分误差绝对值之和最小。通常选用误差平方和最小准则的比较多，这主要是因为最小二乘估计方法成熟，估计形式简明，估计的参数具有许多优良性质[7]；但在一些应用，特别是在某些数量经济的问题中，误差不能认为有正态性，而是服从一种尾部占更大比重的分布，理论证明：在这些情况下，最小一乘估计的统计性能优于最小二乘估计；另外，最小一乘准则的稳健性比最小二乘准则的稳健性好，而且其受异常点的影响较小一点，所以将误差绝对值之和最小为目标也被广泛地应用[8]。

3.1　模型的建立

设
$$Q_{ls} = \sum_{t=1}^{n}\left|y_t - \sum_{i=1}^{k} w_i f_{it}\right| \tag{5}$$

式(5)中 Q_{ls} 表示误差绝对值之和。当以 Q_{ls} 为目标函数时，问题转化为：

$$\min Q_{ls} = \sum_{t=1}^{n}\left|y_t - \sum_{i=1}^{k} w_i f_{it}\right| = \sum_{t=1}^{n}|e_t| \tag{6}$$

由式(6)可以看出，由于目标函数为不可微函数，且不能用类似于式(4)的解析表达式，故其求解非常复杂。下面将给出利用加权最小二乘法求解该问题的算法。

3.2　组合预测权系数的求解

由式(3)可以得到对应的加权最小二乘法模型：

$$\min \quad Q_{wls} = \sum_{t=1}^{n} w_t e_t^2 \tag{7}$$

式(7)中，w_t 为时刻 t 的权值，e_t 意义同前，Q_{wls} 为加权误差平方和。

显然，当 $w_i = 1(i = 1,2,\cdots, n)$ 时，式(7)变为式(3)，此时模型为一般的最小二乘法的模型，即

$$\min \quad Q_{ls} = \sum_{t=1}^{n} e_t^2 \tag{8}$$

由式(8)可以看出，在最小二乘法中，每个误差的平方对 Q_{ls} 的影响是相等的，当 e_t 的绝对值较大时 e_t^2 亦大，即绝对值较大的误差在 Q_{ls} 中占了较大的比重。如果数据中有异常点存在，则它的误差在平方和 Q_{ls} 中将起着较大的影响，特别是数据不多时这种影响更大，因此，为了减小 Q_{ls}，就得“照顾”这种异常点，致使众多的“正常点”的误差绝对值加大，这是不甚合理的。正是由于这一原因，导致有时建模的成效不大，甚至失败。

特别地，在式(7)中取 $w_t = 1/|e_t|$，则式(7)变为

$$\min \quad Q_{ls} = \sum_{t=1}^{n}\frac{1}{|e_t|}e_t^2 = \sum_{t=1}^{n}|e_t| \tag{9}$$

式(9)中，Q_{ls} 意义同前，即误差绝对值之和。

式(9)即为最小一乘准则的组合预测模型。这种准则可以减小或避免上述最小二乘法的不足；但因绝对值函数关于自变量不是处处可导的，故在数学处理上有较大困难。

根据上述的讨论，可应用松驰算法求解，即在加权最小二乘的基础上采用“逐步变权”的方法，逐步迭代求出误差绝对值之和 Q_{ls} 为最小的组合预测模型。

应用这种方法的步骤如下：

step1：取定迭代精度 ε_0，令迭代代数 $M = 0$；

step2：加权最小乘法中取初值 $w_{0i} = 1$，即用最小二乘法求出回归系数和回归方程，进而计算误差向量 $e_0 = (e_{01}, e_{02}, \cdots, e_{0n})$(下标 0 表示第 0 代)；

step3：取 $w_{Mi} = \frac{1}{|e_{M-1\,i}|}$，应用式(7)求出回归系数和回归方程，进而求得误差 e_{Mi}；

step4：$M = M + 1$；

step5：计算相邻两代间的误差接近程度 $\Phi = \left|\frac{Q_{ls\,M+1} - Q_{ls\,M}}{Q_{ls\,M+1}}\right|$(这里 $Q_{ls\,M}$ 表示第 M 代的 Q_{ls})；

step6：若 $\Phi < \varepsilon_0$，则输出计算结果，停止计算；否则，转 step3；

计算中,当出现个别误差 $e_t = 0$ 时,则 $w_t = \frac{1}{|e_t|} = +\infty$。此时,可取 w_t 为一个较大数(如:w_t 取样本数),使运算继续下去。实际上,为保证算法的稳定性,w_t 不宜取得太大,如 w_t 太大使得第 t 项占绝对优势,算法出现震荡;如果 w_t 太小,算法收敛速度将会减慢。

上述方法将加权最小二乘法应用于最小一乘,算法收敛速度较快;同时,计算中便于得到首次运算的最小二乘法解和基于最小一乘准则的最终结果。

3.3 基于百分误差的组合预测模型

式(5)、式(6)体现的都是数值的绝对差的关系,未能体现误差与原始数据相对大小关系。如两数的绝对差相等,但由于基数不同,使得其百分误差相差甚远。下面以百分误差绝对值之和最小为目标建立如下模型:

$$\min Q_{1pa} = \sum_{t=1}^{n} \left| \frac{y_t - \sum_{i=1}^{k} w f_{it}}{y_t} \right| = \sum_{t=1}^{n} \left| 1 - \sum_{i=1}^{k} w_i \frac{f_{it}}{y_t} \right| = \sum_{t=1}^{n} |e'_t| \qquad (10)$$

比较式(6)和式(10)可以看出,此处只是将 e_t 修改为 e'_t 的形式,应用上一节的迭代方法能较容易地求解模型(10)。

4 实例分析

将文献[2,4]和笔者提出的方法分别用于"河南省化工行业人才预测研究"[4]。先用 3 种预测方法分别预测,得到专门人才的预测值及观测值如表 1[2,4]。

表 1 3 种预测方法的预测结果 人

年份	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
y_t	6 014.00	6 398.00	6 781.00	7 416.00	8 076.00	8 399.00	9 215.0	11 125.0	14 120.0	17 238.0	18 689.0	20 592.0	22 665.0
f_{1t}	5 865.28	6 033.17	6 435.57	7 072.49	7 943.94	9 049.90	10 390.4	11 965.4	13 744.9	15 819.0	18 097.5	20 610.6	23 358.2
f_{2t}	5 967.44	6 131.59	6 527.73	7 155.87	8 016.00	9 108.12	10 432.2	11 988.3	13 776.4	15 796.5	18 048.6	20 532.7	23 248.8
f_{3t}	4 921.74	5 641.31	6 488.58	7 463.55	8 566.22	9 796.60	11 154.7	12 640.5	14 254.0	15 995.2	17 864.1	19 860.7	21 985.0

说明:y_t 表示实际的化工人数;f_{1t}、f_{2t} 和 f_{3t} 分别表示 3 种不同方法对化工人才的预测结果。

几种方法确定的组合权系数及预测误差见表 2。

表 2 几种预测方法计算结果的比较

方法	w_1	w_2	w_3	误差平方和	误差绝对值和	平均百分误差 /%	备注
1	1.0	0	0	5 906 494.0	7 098.650	4.984 441	表 1 第 1 种方法
2	0	1.0	0	5 888 705.5	6 744.590	4.570 902	表 1 第 2 种方法
3	0	0	1.0	13 346 328.0	11 144.940	8.572 965	表 1 第 3 种方法
4	0.037 6	1.138 6	- 0.171 6	5 715 340.0	6 868.098	4.432 908 5	由文献[2]方法得到
5	0.525 8	0.475 3	0	5 867 100.0	6 867.795	4.651 895	由文献[3]方法得到
6	- 0.173 297	1.357 147	- 0.183 887 5	5 713 630.0	6 805.850	4.523 451 6	由本文式(3)得到
7	- 2.019 872	3.040 168	- 0.010 074	6 392 590.0	6 391.804	4.293 717	由本文式(6)得到
8	- 1.655 039	2.822 226	- 0.163 366	5 901 400.0	6 395.038	4.202 882	由本文式(10)得到

说明:方法 1 ~ 3 为单模型方法;方法 4 ~ 8 为组合预测方法。

从表 2 中可以看出:方法 6 的误差平方和是所有方法中最小的,方法 7、8 求得的误差绝对值之和、平均绝对百分误差分别为所有方法中最小的。本文提出的 3 种权系数确定方法,其算法易于理解,计算简便;且从 3 个不同的角度给出了 3 种方法,决策者可以根据实际选择组合预测对应的目标函数(如:$\sum_{t=1}^{n} e_t^2$、$\sum_{t=1}^{n} |e_t|$、$\sum_{t=1}^{n} \left|\frac{e_t}{y_t}\right|$),从而选择适当的方法进行组合预测。

同时,笔者还利用本文方法对重庆市(计及万涪黔)1995 ~ 2000 年人口进行了预测,各单模型及组合预测模型计算结果见表 3,预测误差的比较分析见表 4。

表 3 重庆市人口预测 万人

年份	1995	1996	1997	1998	1999	2000	2001	备注
实际人口	3 001.77	3 022.77	3 042.92	3 059.69	3 072.34	3 091.09		
方法 1	3 001.77	3 024.68	3 041.15	3 057.70	3 074.35	3 091.09	3 107.92	改进 GM(1,1)[9]
方法 2	3 001.77	3 024.65	3 041.11	3 057.67	3 074.33	3 091.09	3 107.95	改进 GM(1,1)[9]
方法 3	3 005.69	3 022.77	3 039.85	3 056.93	3 074.01	3 091.09	3 108.17	最小一乘法[8]
组合预测	3 002.17	3 024.59	3 041.93	3 058.07	3 074.25	3 089.84	3 104.83	由本文式(3)得到
组合预测	3 001.77	3 026.01	3 042.92	3 059.03	3 075.24	3 091.09	3 106.59	由本文式(6)得到
组合预测	3 001.77	3 026.01	3 042.92	3 059.03	3 075.24	3 091.09	3 106.59	由本文式(10)得到

说明:方法 1 表示直接应用文献[9]方法;方法 2 是将原数据进行对数变换后再应用文献[9]方法。

表 4　几种预测方法预测误差的比较

方法	w_1	w_2	w_3	误差平方和	误差绝对值和	平均百分误差 /%	备注
1	1.0	0	0	14.779 4	7.680 21	0.041 97	表 3 方法 1
2	0	1.0	0	14.859 4	7.703 83	0.042 10	表 3 方法 2
3	0	0	1.0	35.197 8	11.420 0	0.062 67	表 3 方法 3
4	64.721 38	- 64.134 78	0.412 992	12.289 30	7.992 2	0.043 61	由本文式(3)得到
5	45.25	- 44.25	- 0.0	19.301 34	6.795	0.037 164	由本文式(6)得到
6	45.25	- 44.25	- 0.0	19.301 34	6.795	0.037 164	由本文式(10)得到

从表 4 中可以看出，组合预测的结果比单模型方法好。事实上，当个别单模型方法预测效果较差时，预测效果将更加明显。

5　结　论

通过上述计算和分析可以看出，组合预测模型比单模型具有更高的预测精度，是一种有效的工程预测方法。

基于最小一乘准则和最小二乘准则，文中给出组合预测模型权系数的回归分析方法，并可根据预测工作者的需要选择误差平方和最小、误差绝对值之和最小以及绝对百分误差和为目标，以满足实际问题的需要。文中给出最小一乘模型求解的变权系数方法。该方法计算方便，编程容易，而且可以同时得到基于最小一乘和最小二乘两组解，减少了计算时间。

同时，文中用实例证实了方法得正确性和可行性。

参考文献：

[1] BATES J M, GRANGER C. The Combination of Forecast[J]. Operation Research Quarterly, 1969,20:451 - 468.

[2] 唐小我. 组合预测计算方法研究[J]. 预测,1991,10(4):35 - 40.

[3] 唐小我. 组合预测方法研究的若干新成果[J]. 预测,1992,11(5):39 - 46.

[4] 王明涛. 非线性规划在确定组合预测权系数中的应用[J]. 预测,1994,13(3):60 - 61.

[5] 周宗放,杨春德. 组合预测权系数的目标规划确定法[J]. 系统工程理论方法应用,1995,4(1):50 - 55.

[6] 王景,刘良栋,王作义. 组合预测方法的现状和发展[J]. 预测,1997,16(6):37 - 38.

[7] 王福林,张晋国. 二个模型构成的组合预测模型最优权估计的一种方法[J]. 预测,1995,14(6):54 - 55.

[8] 陈希孺. 最小一乘回归(上)[J]. 数理统计与管理,1989,(5):48 - 55.

[9] 谢开贵. 基于遗传算法的模型[J]. 系统工程学报,2000,15(2):168 - 172.

Combination Forecasting Models Using Regression Analysis Method

XIE Kai-gui, ZHOU Jia-qi

(College of Electrical Engineering, Chongqing University, Chongqing 400044, China)

Abstract: Using regression analysis method, the methods for solving the weights of combination forecasting model (CFM) are proposed. At first, the linear regress CFM are presented based on the least absolute criteria and least square criteria. Then the weights can be evaluated using the least square princinple. Because the objective function of CFM based on least absolute criteria is non-differential, the traditional programming methods can not solve it. So the least square method with the modified weights is proposed to solve this problem. At the same time, methods for solving CFM is given with the aim of minimizing sum of percentage error absolutes. From many cases, the results show that the forecasting precision of CFM is very high and the effect of regression is remarkable.

Key words: combination forecasting model; least square criteria; least absolute criteria; method of the modified weights step by step

(责任编辑　李胜春)