DOI: 10.13656/j.cnki.gxkx.1997.04.008
广西科学 Guangxi Sciences 1997, 4 (4): 267～ 270

一种提高灰色预测模型的方法
A Method of Improving the Grey Forecasting Model

伍艳春
Wu Yanchun

(桂林工学院基础部 桂林市建干路 12号 541004)
(Guilin Institute of Technology, 12 Jianganlu, Guilin, Guangxi, 541004)

摘要 根据文献 [1] 提出的科技文献增长和老化的灰色模型,适当改变原始数据,可以使模型精度提高.
关键词 增长 老化 灰色模型 精度

Abstract Basing on the grey model on the increasing and ageing of scientific literature, which is advanced in literature [1]. The author, in this essay, appropriately changes the firsthand data, so that the accuracy of the model can be improved.

Key words increasing, ageing, grey model, accuracy
中图法分类号 O 221.3; O 213.9

1 文献增长灰色模型简介

设 $F=\{F(1),F(2),\cdots,F(n)\}$ 为逐年文献量数列,令 $x^{(0)}_{(k)}=1/F(k)$, $k=1,2,\cdots,n$.
得原始数列 $x^{(0)}=\{x^{(0)}_{(1)},x^{(0)}_{(2)},\cdots,x^{(0)}_{(n)}\}$

令 $x^{(1)}_{(k)}=\sum_{i=1}^{k}x^{(0)}_{(i)}$,得累加生成列

$x^{(1)}=\{x^{(1)}_{(1)},x^{(1)}_{(2)},\cdots,x^{(1)}_{(n)}\}$

则 $x^{(1)}$ 服从的规律之白化形式为:

$$\frac{dx^{(1)}}{dt}=ax^{(1)}_{(t)}+bt+c$$

上式离散化

$$x^{(0)}_{(k)}=a\frac{1}{2}[x^{(1)}_{(k-1)}+x^{(1)}_{(k)}]+(k-\frac{1}{2})b+c$$

$$k=2,3,4,\cdots,n \tag{1}$$

将 $x^{(0)},x^{(1)}$ 代入式(1),用最小二乘法确定 a、b、c:

$$(c\quad b\quad a)^T=(B^TB)^{-1}B^Ty \tag{2}$$

其中

$$B=\begin{pmatrix}1 & \frac{3}{2} & \frac{1}{2}(x^{(1)}_{(1)}+x^{(1)}_{(2)})\\ 1 & \frac{5}{2} & \frac{1}{2}(x^{(1)}_{(2)}+x^{(1)}_{(3)})\\ \cdots & \cdots & \cdots\\ 1 & n-\frac{1}{2} & \frac{1}{2}(x^{(1)}_{(n-1)}+x^{(1)}_{(n)})\end{pmatrix}$$

$$y=\begin{pmatrix}x^{(0)}_{(2)}\\ x^{(0)}_{(3)}\\ \cdots\\ x^{(0)}_{(n)}\end{pmatrix}$$

于是原始数列的模型值和预测值为:

$$\hat{x}^{(0)}_{(k)}=-\frac{b}{a}+(x^{(0)}_{(1)}+\frac{b}{a}+\frac{b+ac}{a^2})(1-e^{-a})e^{a(k-1)}\qquad k=2,3,4,\cdots \tag{3}$$

逐年文献量模型值 $\hat{F}(k)=1/x^{(0)}_{(k)}$.

2 $x^{(0)}_{(1)}$ 对文献增长灰色模型的影响

我们要证明原始数列模型值式(3)与 $x^{(0)}_{(1)}$ 无关.亦即

今有原始数据列 $x^{(0)}=\{x^{(0)}_{(1)},\cdots,x^{(0)}_{(n)}\}$ 和原始数据列 $Z^{(0)}=\{Z^{(0)}_{(1)},Z^{(0)}_{(2)},\cdots,Z^{(0)}_{(n)}\}$, $Z^{(0)}$ 与 $x^{(0)}$ 的区别仅在于第一个数据:

$$Z^{(0)}_{(1)}=x^{(0)}_{(1)}+r,\quad (r\text{是任一实数})$$

$$Z^{(0)}_{(k)}=x^{(0)}_{(k)},\quad k=2,3,\cdots,n \tag{4}$$

由 $x^{(0)}$ 如前建模得原始数据模型值为式(3).
由 $Z^{(0)}$ 如前建模得原始数据模型值为:

$$\hat{Z}^{(0)}_{(k)}=-\frac{U}{T}+(Z^{(0)}_{(1)}+\frac{U}{T}+\frac{U+TV}{T^2})(1-e^{-T})e^{T(k-1)},\qquad k=2,3,\cdots \tag{5}$$

我们要证明 $\hat{Z}^{(0)}_{(k)}=\hat{x}^{(0)}_{(k)}$, $k=2,3,\cdots$

由 $Z^{(0)}$ 作累加生成列:

$Z^{(1)}=\{Z^{(1)}_{(1)},Z^{(1)}_{(2)},\cdots,Z^{(1)}_{(n)}\}$

1997-04-21收稿.

其中 $Z^{(1)}_{(k)} = \sum_{i=1}^{k} Z^{(0)}_{(i)}, k = 1, 2, \cdots, n.$

显然有 $Z^{(1)}_{(k)} = x^{(1)}_{(k)} + r$ (6)

式(5)中系数 T, U, V由下式给出(仿(2)):

$$(V\quad U\quad T)^T = (B_1^T B_1)^{-1} B_1^T Z \tag{7}$$

这里

$$B_1 = \begin{pmatrix} 1 & \frac{3}{2} & \frac{1}{2}(Z^{(1)}_{(1)} + Z^{(1)}_{(2)}) \\ 1 & \frac{5}{2} & \frac{1}{2}(Z^{(1)}_{(2)} + Z^{(1)}_{(3)}) \\ \cdots & \cdots & \cdots \\ 1 & n - \frac{1}{2} & \frac{1}{2}(Z^{(1)}_{(n-1)} + Z^{(1)}_{(n)}) \end{pmatrix}$$

$$Z = \begin{pmatrix} Z^{(0)}_{(2)} \\ Z^{(0)}_{(3)} \\ \cdots \\ Z^{(0)}_{(n)} \end{pmatrix}$$

由关系式(4)与(6)以及线性代数知识[4]可知:

$B_1 = BQ, \quad Z = y$

其中

$$Q = \begin{pmatrix} 1 & 0 & r \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

所以 $B_1^T B_1 = (BQ)^T (BQ) = Q^T B^T B Q,$

$(B_1^T B_1)^{-1} = Q^{-1} (B^T B)^{-1} (Q^T)^{-1},$

$$\begin{aligned} &(V\quad U\quad T)^T \\ &= (B_1^T B_1)^{-1} B_1^T Z \\ &= [Q^{-1}(B^T B)^{-1}(Q^T)^{-1}](BQ)^T y \\ &= Q^{-1}(B^T B)^{-1}(Q^T)^{-1} Q^T B^T y \\ &= Q^{-1}[(B^T B)^{-1} B^T y] \\ &= Q^{-1}(c\quad b\quad a)^T. \end{aligned}$$

由于

$$Q^{-1} = \begin{pmatrix} 1 & 0 & -r \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

所以 $(V\quad U\quad T)^T = (c - ar\quad b\quad a)^T$

即 $T = a, \quad U = b, \quad V = c - ar.$ (8)

将(8)式代入(5)式得:

$$\begin{aligned} \hat{Z}^{(0)}_{(k)} &= -\frac{b}{a} + \left(x^{(0)}_{(1)} + r + \frac{b}{a} + \frac{b + a(c - ar)}{a^2}\right) \cdot (1 - e^{-a}) e^{a(k-1)} \\ &= -\frac{b}{a} + \left(x^{(0)}_{(1)} + \frac{b}{a} + \frac{b + ac}{a^2}\right)(1 - e^{-a}) e^{a(k-1)} \\ &= \hat{x}^{(0)}_{(k)} \end{aligned}$$

亦即(3)式与(5)式是相同的.

现在我们看到 $x^{(0)}$ 与 $Z^{(0)}$ 的第一个数据虽然不同,建立的模型却是相同的.据此,为了充分利用原始数据,可以如下处理已有的原始数据:

设已统计到逐年文献量数据

$$F(2), F(3), \cdots, F(n) \tag{9}$$

$F(2)$ 前面的文献量是一个灰数 $\otimes$[5],任取一个实数 $F(1)$ 作为 $\otimes$ 的白化值,以数据

$$F(1), F(2), \cdots, F(n) \tag{10}$$

建立文献增长模型,如前所述,$F(1)$ 是否为 $F(2)$ 前面的文献量对模型值不产生影响.但用(10)式建立模型比用(9)式建立模型多用一个原始数据,理论上说,更具有可靠性.

3 算例

文献[2]给出了河南省高校 1986年~1990年文献累积量数据列: 1269, 1307, 1373, 1432, 1485(万册).文献[2]根据文献[1]提供的方法建立的数学模型为:

$$x(k) = -0.52363 + 1257.43937 e^{0.04169(K-1)} \tag{11}$$

这样文献累积量模型值为(仅到出 1987年~1990年的):

1310, 1366, 1424, 1485(万册)

河南省高校文献累积量在 1985年的值文献[2]没有给出,它是一个灰数 $\otimes$.第二段指出,把它加在统计到的数据前面,由于其实际大小不影响模型值,故取值为 1000,即可以认为 1985年~1990年的文献累积量为:

1000, 1269, 1307, 1373, 1432, 1485.

用这 6个数据按文献[1]的方法建模,有:

$$x(k) = 360.232 + 855.913 e^{0.055(k-1)} \tag{12}$$

这样得到 1986年~1990年文献累积量模型值为:

1265, 1316, 1370, 1427, 1487.

用文献[5]给出的后验差法检验模型精度,模型(11)的后验差比值 $c_1 = 0.071$,模型(12)的后验差比值 $c_2 = 0.066$, $c_1 > c_2$,所以模型(12)比(11)精度高.(后验差比值的计算在下一段介绍)

4 文献老化模型提高精度的一种方法

文献[1]的文献老化模型是根据巴尔顿—凯普勒老化方程

$$y = 1 - \left(\frac{a}{e^{0.1(x-1)}} + \frac{b}{e^{0.2(x-1)}}\right)$$

提出来的.

设有原始数据列

$$x^{(0)} = \{x^{(0)}_{(1)}, x^{(0)}_{(2)}, \cdots, x^{(0)}_{(n)}\} \tag{13}$$

累加生成数据列为 $x^{(1)} = \{x^{(1)}_{(1)}, x^{(1)}_{(2)}, \cdots, x^{(1)}_{(n)}\}$,则 $x^{(1)}_{(t)}$ 满足的白化方程是:

$$\frac{dx^{(1)}}{dt} + 0.1x^{(1)} = u + ae^{-0.2(t-1)},$$

离散化得:

$$x^{(0)}_{(k)} + \frac{0.1}{2}\left[x^{(1)}_{(k)} + x^{(1)}_{(k-1)}\right] = u + a\frac{1}{2}\left[e^{-0.2(k-1)} + e^{-0.2(k-2)}\right], \quad k = 2, 3, \cdots, n. \tag{14}$$

由(14)式用最小二乘法确定 u 和 a:

$$\begin{pmatrix} u \\ a \end{pmatrix} = (B^T B)^{-1} B^T y \tag{15}$$

其中

$$B = \begin{pmatrix} 1 & \frac{1}{2}(e^{-0.2} + 1) \\ 1 & \frac{1}{2}(e^{-0.4} + e^{-0.2}) \\ \cdots & \\ 1 & \frac{1}{2}(e^{-0.2(n-1)} + e^{-0.2(n-2)}) \end{pmatrix}$$

$$y = \begin{pmatrix} x^{(0)}_{(2)} + \frac{0.1}{2}(x^{(1)}_{(2)} + x^{(1)}_{(1)}) \\ x^{(0)}_{(3)} + \frac{0.1}{2}(x^{(1)}_{(3)} + x^{(1)}_{(2)}) \\ \cdots \\ x^{(0)}_{(n)} + \frac{0.1}{2}(x^{(1)}_{(n)} + x^{(1)}_{(n-1)}) \end{pmatrix}$$

把 u 和 a 代入白化方程,得到原始数据模型值和预测值:

$$\hat{x}^{(0)}_{(k)} = \frac{(1 - 10u + 10ae^{0.2})(1 - e^{0.1})e^{-0.1}}{e^{0.1(k-1)}} + \frac{-10a(1 - e^{0.2})}{e^{0.2(k-1)}} \quad k = 2, 3, \cdots, n. \tag{16}$$

今由数据(13)构造如下数据列

$$Z^{(0)} = \{Z^{(0)}_{(1)}, Z^{(0)}_{(2)}, \cdots, Z^{(0)}_{(n)}\} \tag{17}$$

其中 $Z^{(0)}_{(1)} = x^{(0)}_{(1)} + r$, $Z^{(0)}_{(k)} = x^{(0)}_{(k)}$, $k = 2, 3, \cdots, n$.我们称数据列(17)为次生成数据列.

仿建立模型(16)的方法,由次生数据建模得到

$$\hat{Z}^{(0)}_{(k)} = \frac{(1 - 10\_ + 10\mathrm{T}e^{0.2})(1 - e^{0.1})e^{0.1}}{e^{0.1(k-1)}} + \frac{-10\mathrm{T}(1 - e^{0.2})}{e^{0.2(k-1)}} \quad k = 2, 3, \cdots, n. \tag{18}$$

其中系数 _ ,T由下式确定:

$$\begin{pmatrix} \_ \\ \mathrm{T} \end{pmatrix} = (B_1^T B_1)^{-1} B_1^T Z \tag{19}$$

而

$$B_1 = \begin{pmatrix} 1 & \frac{1}{2}(e^{-0.2} + 1) \\ 1 & \frac{1}{2}(e^{-0.4} + e^{-0.2}) \\ \cdots & \\ 1 & \frac{1}{2}(e^{-0.2(n-1)} + e^{-0.2(n-2)}) \end{pmatrix}$$

$$Z = \begin{pmatrix} Z^{(0)}_{(2)} + \frac{0.1}{2}(Z^{(1)}_{(2)} + Z^{(1)}_{(1)}) \\ Z^{(0)}_{(3)} + \frac{0.1}{2}(Z^{(1)}_{(3)} + Z^{(1)}_{(2)}) \\ \cdots \\ Z^{(0)}_{(n)} + \frac{0.1}{2}(Z^{(1)}_{(n)} + Z^{(1)}_{(n-1)}) \end{pmatrix}$$

其中 $Z^{(1)}_{(k)}$, $k = 1, 2, 3, \cdots, n$.是次生数据的累加生成数据.显然有:

$$Z^{(1)}_{(k)} = x^{(1)}_{(k)} + r \tag{20}$$

$$\begin{pmatrix} \_ \\ \mathrm{T} \end{pmatrix} = (B_1^T B_1)^{-1} B_1^T Z \tag{21}$$

显然 $B_1 = B$,由式(20)可知

$$Z = y + R \tag{22}$$

$$R = (0.1r, 0.1r, \cdots, 0.1r)^T$$

将(22)式代入(21)式得到(还要考虑到关系式(15)):

$$\begin{pmatrix} \_ \\ \mathrm{T} \end{pmatrix} = (B^T B^{-1}) B^T (y + R) = \begin{pmatrix} u \\ a \end{pmatrix} + (B^T B)^{-1} B^T R \tag{23}$$

下面求 $(B^T B)^{-1} B^T R$.

在式(14)中,将其左端改为 $0.1r$,视其为关于 u, a 的方程组,这时:

$$u = 0.1r, \quad a = 0$$

是它的唯一解,由线性代数广义逆矩阵理论[4]得到:

$$(B^T B)^{-1} B^T R = (0.1r, 0)^T \tag{24}$$

将(24)式代入(23)式,得到

$$\begin{pmatrix} \_ \\ \mathrm{T} \end{pmatrix} = \begin{pmatrix} u \\ a \end{pmatrix} + \begin{pmatrix} 0.1r \\ 0 \end{pmatrix} = \begin{pmatrix} u + 0.1r \\ a \end{pmatrix}$$

即 $\_ = u + 0.1r$, $\mathrm{T} = a$.

代入式(18),整理后将 $\hat{Z}^{(0)}_{(k)}$ 与式(16)所确定的 $\hat{x}^{(0)}_{(k)}$ 比

较,有

$$\hat{Z}^{(0)}_{(k)} = \hat{x}^{(0)}_{(k)} + \frac{-r(1-e^{0.1})}{e^{0.1k}},$$

$$k = 2, 3, \cdots, n \tag{25}$$

上式说明,对于文献老化模型,第一个原始数据对模型值是有影响的.或者说,第一个原始数据的改变可引起模型精度的变化.下面我们将讨论怎样选择 r 的值,使得模型精度最高.

我们用后验差比值 c 来衡量精度.由文献[5]

$$c = \frac{s_2}{s_1} \tag{26}$$

由于对于第一个原始数据,我们给它一个改变量 r,所以检验精度中不用第一个原始数据.故(26)式中的

$$s_1^2 = \frac{1}{n-1}\sum_{k=2}^{n}(Z^{(0)}_{(k)} - \bar{Z}^{(0)})^2,$$

$$s_2^2 = \frac{1}{n-1}\sum_{k=2}^{n}(X_k - \bar{X})^2.$$

而 $\bar{Z}^{(0)} = \frac{1}{n-1}\sum_{k=2}^{n} Z^{(0)}_{(k)}$; $X_k = Z^{(0)}_{(k)} - \hat{Z}^{(0)}_{(k)}$,

$k = 1, 2, 3, \cdots, n$; $\bar{X} = \frac{1}{n-1}\sum_{k=2}^{n} X_k$.

显然,无论 r 取什么值,s_1 是一个常量.所以只须考虑 r 对 s_2 的影响.

因为 $Z^{(0)}_{(k)} = x^{(0)}_{(k)}$, $k = 2, 3, \cdots, n$.

由式(25)可知:

$$X_k = Z^{(0)}_{(k)} - \hat{Z}^{(0)}_{(k)} = x^{(0)}_{(k)} - \hat{x}^{(0)}_{(k)} + \frac{r(1-e^{0.1})}{e^{0.1k}}$$

$$= X_k^{(0)} + \frac{r(1-e^{0.1})}{e^{0.1k}}, k = 2, 3, \cdots, n \tag{27}$$

其中 $X^{(0)}_{(k)} = x^{(0)}_{(k)} - \hat{x}^{(0)}_{(k)}$, $k = 2, 3, \cdots, n$.

所以

$$\bar{X} = \frac{1}{n-1}\sum_{k=2}^{n} X_k = \frac{1}{n-1}\sum_{k=2}^{n} X_k^{(0)} + \frac{1}{n-1}\sum_{k=2}^{n}\frac{r(1-e^{0.1})}{e^{0.1k}}$$

$$= \bar{X}^{(0)} + \frac{1}{n-1} r(1-e^{0.1}) e^{-0.2}\frac{1-e^{-0.1(n-1)}}{1-e^{0.1}}. \tag{28}$$

上式用了等比数列求和公式,其中

$$\bar{X}^{(0)} = \frac{1}{n-1}\sum_{k=2}^{n} X_k^{(0)}.$$

式(28)还可化简为:

$$\bar{X} = \bar{X}^{(0)} - \frac{r}{n-1}\frac{1-e^{-0.1(n-1)}}{e^{0.1}}. \tag{29}$$

这样

$$S_2^2 = \frac{1}{n-1}\sum_{k=2}^{n}(X_k - \bar{X})^2$$

$$= \frac{1}{n-1}\sum_{k=2}^{n}\left[(X_k^{(0)} - \bar{X}^{(0)}) - r\left(\frac{1-e^{0.1}}{e^{0.1k}} - \frac{1-e^{-0.1(n-1)}}{(n-1)e^{0.1}}\right)\right]^2$$

$$= Lr^2 + Mr + N \tag{30}$$

其中

$$L = \frac{1}{n-1}\sum_{k=2}^{n}\left(\frac{1-e^{0.1}}{e^{0.1k}} - \frac{1-e^{-0.1(n-1)}}{(n-1)e^{0.1}}\right)^2 \tag{31}$$

$$M = -\frac{2}{n-1}\sum_{k=2}^{n}(X_k^{(0)} - \bar{X}^{(0)})\left(\frac{1-e^{0.1}}{e^{0.1k}} - \frac{1-e^{-0.1(n-1)}}{(n-1)e^{0.1}}\right) \tag{32}$$

$$N = \frac{1}{n-1}\sum_{k=2}^{n}(X_k^{(0)} - \bar{X}^{(0)})^2 \tag{33}$$

式(30)中,$L > 0$,所以这个关于 r 的二次函数当

$$r = -\frac{M}{2L} \tag{34}$$

时 S_2^2 取最小值,由式(26)知,这时 c 取最小值,即模型的精度最高.

致谢

本文在撰写过程中,得到了李正吾教授的热情指导和帮助,在此谨致以诚挚的谢意.

参考文献

1 李正吾.科学文献增长和老化的灰色预测模型(GM),情报学报,1990,(5):342~352.

2 白广思.科学文献灰色预测模型的分析及修正.情报学报,1993,(2):140~144.

3 李 华,马文光.应用灰色模型进行情报预测存在的问题.情报学报,1993,(1).

4 Strang G. 线性代数及其应用.侯自新等译.天津:南开大学出版社,1990,149~157.

5 邓聚龙.灰色控制系统.武汉:华中工学院出版社,1985,318~324.

(责任编辑:邓大玉)