讲 座

灰色多序列数据残差辨识

鞍钢卫生防疫站 魏振宇 杨彦华 刘桐树

随着在卫生事业管理和疾病控制中引入系统工程理论和方法，寻找科学精确的数学模型愈来愈重要。准确的发病预测，可以量化对疾病的预防、控制、管理决策，也是系统分析与规化决策的基础。

灰色系统预测是近年来形成的新的理论，已经在社会经济、农业、生命系统中得到广泛的应用。在医学卫生系统中，存在着许多具有灰色信息特性的因素、现象，并具有随机性。不同于概率论的是，处理这些灰色量是通过关联度，通过数据处理来分析和对待随机量，也就是通过数据到数据的“映射”来处理随机量和发现规律。采用灰色系统方法建模，对原始数据要求不苛刻，这是其它建模方法所不能比拟的优点。

灰色残差信息是灰色系统预测的基本概念；基本机理是灰色信息的开发和利用；数据残差辨识是其基本方法之一。

本文旨在介绍与灰色多序列数据残差辨识有关的概念、计算方法，以及通过实例说明该方法在医学上的应用。

灰色残差信息

灰色建模的主要方法是最小二乘法。但最小二乘法的解仅是近似解，存在着残差。比如有矩阵关系

$$m=A\delta$$

$$m=\begin{bmatrix}1\\0\\1\end{bmatrix}\quad A=\begin{bmatrix}1&0\\2&-1\\-1&2\end{bmatrix}\quad \delta=\begin{bmatrix}\delta_1\\\delta_2\end{bmatrix}$$

按最小二乘法得

$$\hat{\delta}=[A^TA]^{-1}A^Tm=\begin{bmatrix}0.57\\0.857\end{bmatrix}$$

若$\hat{\delta}$是方程$m=A\delta$的精确解，将$\hat{\delta}$代回原式，两边应相等，但结果却不相等

$$A\hat{\delta}=[0.57\quad 0.285\quad 1.14]$$

其残差Δ为

$$\Delta=m-A\delta=[0.43\quad -0.285\quad -0.14]^T$$

因此可以说灰色残差信息是具有灰色量特征的“边角余料”信息。所谓的残差辨识是将残差中有用的信息挖掘出来再利用，以提高信息的利用率和模型的预测精度。

数学方法概述和计算步骤

1.给定原始数据列

$$\{X^{(0)}_{k(i)}\}\qquad i=1,\cdots,N\quad k=1,\cdots,h$$

2.构造GM(0，h)模型，其数学方程式为

$$X^{(0)}_{1(t)}=\sum_{i=1}^{n-1}biX^{(0)}_{i+1}(t)+b_o$$

系数向量为

$$\hat{a}=[b_1,\cdots,b_h,b_o]^T$$

用最小二乘法求解

$$\hat{a}=[B^TB]^{-1}B^TY_N$$

B为数据矩阵，Y_N为N列的数据向量。

3.数据残差辨识

$$A_{N-1}=0$$

$$B_{N-1}=\begin{bmatrix}X^{(0)}_1(N-1)\cdots X^{(0)}_1(N-k)\\\vdots\qquad\qquad\vdots\\X^{(0)}_h(N-1)\cdots X^{(0)}_h(N-k)\end{bmatrix}\qquad(1)$$

$$Y_N=[X^{(0)}_1(N),\cdots,X^{(0)}_h(N)]^T\qquad(2)$$

令δ_1为待辨识的参数列

$$\delta_1=[\delta_{11},\delta_{12},\cdots,\delta_{1k}]^T$$

则δ_1与数据矩阵$X_{N-1}(B)$、Y_N之间有如下关系

$$Y_N=X_{N-1}(B)\delta_1\qquad X_{N-B}(B)\triangleq B_{N-1}$$

按最小二乘法求解有

$$\hat{\delta}_1=[X_{N-1}(B)^TX_{N-1}(B)]^{-1}X_{N-1}(B)^TY\qquad(3)$$

根据最小二乘解的残差概念，定义N−1级残差$\hat{\Delta}_{N-1}$为

$$\hat{\Delta}_{N-1}=Y_N-X_{N-1}(B)\,\hat{\delta}_1\qquad(4)$$

令下式成立

$$\hat{\Delta}_{N-1}=X_{N-2}(B)\delta_2$$

$$A_{N-2}=0$$

$$B_{N-2}=\begin{Bmatrix} X_1^{(0)}(N-2)\cdots X_1^{(0)}(N-k-1) \\ \vdots \qquad\qquad \vdots \\ X_h^{(0)}(N-2)\cdots X_h^{(0)}(N-k-1) \end{Bmatrix} \quad (5)$$

$$\delta_2=[\delta_{21},\ \delta_{22},\ \cdots,\ \delta_{2k}]^T$$

则N－2级中的残差按下式寻找

$$\hat{\delta}_2=[X_{N-2}(B)^TX_{N-2}(B)]^{-1}X_{N-2}(B)^T \hat{\Delta}_{N-1} \quad (6)$$

以此类推便得到如下公式

$$\hat{\Delta}_{N-2}=\hat{\Delta}_{N-1}-X_{N-2}(B)\hat{\delta}_2 \quad (7)$$

$$\vdots$$

$$\hat{\delta}_{N-k}=[X_k(B)^TX_k(B)]^{-1}X_k(B)^T \hat{\Delta}_{k+1} \quad (8)$$

$$\hat{\Delta}_{N-k}=\hat{\Delta}_{N-k+1}-X_k(B)\hat{\delta}_{N-k} \quad (9)$$

综合（1）～（9）式有

$$Y_N=X_{N-1}(B)\hat{\delta}_1+X_{N-2}(B)\hat{\delta}_2+\cdots+X_k(B)\delta_{N-k}+\Delta_{N-k}$$

将此式外推，便得到数据残差辨识的预测模型

$$\hat{Y}_{N+1}=X_N(B)\hat{\delta}_1+X_{N-1}(B)\hat{\delta}_2+\cdots+X_{k+1}(B)\hat{\delta}_{N-k}+\hat{\Delta}_{N-k} \quad (10)$$

4.模型精度检验，一般多采用后验差检验方法

（1）求原始数据和残差的平均值

$$\bar{X}=\frac{1}{N}\sum_{t=1}^{N}X_t \quad (11)$$

$$\bar{\varepsilon}=\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t \quad (12)$$

（2）求原始数据方差和残差方差

$$S_1^2=\frac{1}{N}\sum_{t=1}^{N}(X_t-\bar{X})^2 \quad (13)$$

$$S_2^2=\frac{1}{n}\sum_{t=1}^{n}(\varepsilon_t-\bar{\varepsilon})^2 \quad (14)$$

（3）计算后验差比值C和小误差概率P，并根据检验表对精度进行估计

$$C=\frac{\sqrt{\frac{S_2^2}{n-1}}}{\sqrt{\frac{S_1^2}{N-1}}} \quad (15)$$

$$P=\{|\varepsilon_t-\bar{\varepsilon}|<0.6745S_1\} \quad (16)$$

医学实例计算

表1是鞍钢职工1984～1988年心脏病、脑血管病和癌症的死亡率。现根据1984～1987年的数据建立多序列数据残差辨识模型，并预测1988年上述三病的死亡率。

表1 心脏病、脑血管病、癌症死亡率(/10万)数据

	1984 (1)	1985 (2)	1986 (3)	1987 (4)	1988 (5)
心脏病	74.8	78.3	82.0	86.0	88.6
	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
脑血管病	112.2	125.6	127.4	102.4	128.3
	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
癌症	178.4	180.7	176.2	176.4	172.9
	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}

根据公式（1）～（9）计算有关数据。

构造数据阵

$$X_4(B)=\begin{Bmatrix} X_{14} & X_{13} \\ X_{24} & X_{23} \\ X_{34} & X_{33} \end{Bmatrix}=\begin{Bmatrix} 86.0 & 82.0 \\ 102.4 & 127.4 \\ 176.4 & 176.2 \end{Bmatrix}$$

$$X_3(B)=\begin{Bmatrix} X_{13} & X_{12} \\ X_{23} & X_{22} \\ X_{33} & X_{32} \end{Bmatrix}=\begin{Bmatrix} 82.0 & 78.3 \\ 127.4 & 125.6 \\ 176.2 & 180.7 \end{Bmatrix}$$

$$X_2(B)=\begin{Bmatrix} X_{12} & X_{11} \\ X_{22} & X_{21} \\ X_{32} & X_{31} \end{Bmatrix}=\begin{Bmatrix} 78.3 & 74.8 \\ 125.6 & 112.2 \\ 180.7 & 178.4 \end{Bmatrix}$$

$$Y_4=\begin{Bmatrix} 86.0 \\ 102.4 \\ 176.4 \end{Bmatrix}$$

$$\hat{\delta}_3=[X_3(B)^TX_3(B)]^{-1}X_3(B)^TY_4$$

$$=\left(\begin{bmatrix} 82.0 & 127.4 & 176.2 \\ 78.3 & 125.6 & 180.7 \end{bmatrix}\begin{bmatrix} 82.0 & 78.3 \\ 127.4 & 125.6 \\ 176.2 & 180.7 \end{bmatrix}\right)^{-1}$$

$$\times\begin{pmatrix}82.0 & 127.4 & 176.2\\ & & \\ 78.3 & 125.6 & 180.7\end{pmatrix}\times\begin{pmatrix}86.0\\102.4\\176.4\end{pmatrix}$$

$$=\begin{pmatrix}-0.3254\\ \\ 1.2588\end{pmatrix}$$

$$\hat{\Delta}_3=Y_4-X_3(B)\,\hat{\delta}_3$$

$$=\begin{pmatrix}86.0\\102.4\\176.4\end{pmatrix}-\begin{pmatrix}82.0 & 78.3\\127.4 & 125.6\\176.2 & 180.7\end{pmatrix}\times\begin{pmatrix}-0.3254\\ \\ 1.2588\end{pmatrix}$$

$$=\begin{pmatrix}14.0\\-14.2\\6.3\end{pmatrix}$$

$$\hat{\delta}_2=[X_2(B)^TX_2(B)]^{-1}X_2(B)^T\hat{\Delta}_3$$

$$=\left(\begin{pmatrix}78.3 & 125.6 & 180.7\\ & & \\ 74.8 & 112.2 & 178.4\end{pmatrix}\begin{pmatrix}78.3 & 74.8\\125.6 & 112.2\\180.7 & 178.4\end{pmatrix}\right)^{-1}$$

$$\begin{pmatrix}78.3 & 125.6 & 180.7\\ & & \\ 74.8 & 112.2 & 178.4\end{pmatrix}\begin{pmatrix}14.0\\-14.2\\6.3\end{pmatrix}$$

$$=\begin{pmatrix}-1.4471\\ \\ 1.5198\end{pmatrix}$$

$$\hat{\Delta}_2=\hat{\Delta}_3-X_2(B)\,\hat{\delta}_2$$

$$=\begin{pmatrix}14.0\\-14.2\\6.3\end{pmatrix}-\begin{pmatrix}78.3 & 74.8\\125.6 & 112.2\\180.7 & 178.4\end{pmatrix}\begin{pmatrix}-1.4471\\ \\ 1.5198\end{pmatrix}$$

$$=\begin{pmatrix}3.6\\-3.0\\-3.8\end{pmatrix}$$

根据公式（10）预测1988年三病死亡率，结果见表2。

$$Y_5=X_4(B)\,\hat{\delta}_3+X_3(B)\hat{\delta}_2+\hat{\Delta}_2$$

$$=\begin{pmatrix}86.0 & 82.0\\102.4 & 127.4\\176.4 & 176.2\end{pmatrix}\begin{pmatrix}-0.3254\\ \\ 1.2588\end{pmatrix}+$$

$$\begin{pmatrix}82.0 & 78.3\\127.4 & 125.6\\176.2 & 180.7\end{pmatrix}\begin{pmatrix}-1.4471\\ \\ 1.5198\end{pmatrix}+\begin{pmatrix}13.6\\-3.6\\-3.8\end{pmatrix}$$

$$=\begin{pmatrix}89.1\\130.6\\180.2\end{pmatrix}$$

表2 1988年三病死亡率观测值与预测值比较（/10万）

	观测值	预测值	差值(ε)	误差(%)
心脏病	88.6	89.1	−0.5	0.6
脑血管病	128.3	130.6	−2.3	1.8
癌　症	172.9	180.9	−8	4.6

计算结果表明，预测值与实际观测值十分接近。由于样本序列数不够，不能做后验差检验，但预测误差仅分别为0.6%、1.8%和4.6%，足以证明模型精度较高。这样的预测误差，对于决策很可能不会产生偏差，是能够被接受的。

讨　论

数据残差辨识预测在其它领域已取得满意的成果，在医学卫生方面的应用还是个偿试，结果已证明具有较高的实用价值。从理论上说，该模型是将时间序列转化为微分方程，建立抽象系统发展变化的动态模型，能够描述医学及卫生管理科学等系统内部的物理和化学过程的本质特征，并且适用广泛。该模型具有如下特点：（1）计算简便；（2）适合摆动幅度不大的非平稳随机过程；（3）精度较高；（4）残差信息利用率的提高有利于系统的量化分析；（5）主要的缺点是预测期短。

参考文献

1.邓聚龙.灰色系统.北京：国防工业出版社，1985;60～70.
2.肖明耀.误差理论与应用.北京：计量出版社，1985；233～239.

（1989年2月1日收稿，同年5月28日修回）