

# 用基于联系数的主因子分析预测法 预测流行性乙型脑炎

李秀央 陈坤 赵克勤

**【摘要】** 目的 探讨流行性乙型脑炎(乙脑)发生率与预测因子的关系,为预测提供一种简单、行之有效的新方法。方法 首先计算历史上预测因子与发病率的联系数,根据联系数值的大小依次排列出最主要因子、次主要因子、再次主要因子,剔除联系数值为最小的次要因子;再把新近出现的预测因子观测值与历史上同类因子值相比较,与之最接近的因子值为该次预测用因子值,并根据该因子值与当时发病率的同一度建立预测方程,代入新因子值,解此方程得到预测值;当预测用因子数为  $n'$  ( $n' \geq 2$ ) 时,取  $n'$  个预测值的平均作为该次预测值。结果 应用以上预测方法预测某地某年乙脑,预测值与实际发生率很接近,仅相差 0.0264/10 万,准确率为 97.94%。结论 可以应用基于联系数的主因子分析预测法预测乙脑。

**【关键词】** 乙型脑炎,流行性;预测;联系数;主因子

**A connection number-based principal factor analysis forecast method to forecast the encephalitis B epidemics** LI Xiu-yang\*, CHEN Kun, ZHAO Ke-qin. \*Department of Epidemiology & Health Statistics, Medical College, Zhejiang University, Hangzhou 310031, China

**【Abstract】 Objective** To detect the relations between incidence rate of the epidemical encephalitis B and related factors, to provide a simple, valid and practical new method for forecasting encephalitis B epidemics. **Methods** Connection number between the incidence rate of encephalitis B and the historical forecast factors was computed, before ranking the first, second and the third principal factor, to remove the factor with the smallest value in the light of the connection number before comparing the newest value of forecast factors with the same kind of history while the most nearly value becoming the forecasting factor value and to establish a forecasting equation according to the factor value and the consistent degree of the incidence rate of encephalitis B at that time. Finally, to put into the new factor value to get this forecast value under this equation. Assuming that there are  $n'$  ( $n' \geq 2$ ) forecast factors, this time forecast value can then be directly obtained from the average of these estimate values. **Results** Using above forecast method to forecast the incidence rate of encephalitis B at certain place and year, the predicting value is very much close to the actual incidence rate. Difference between the predicting value forecasted by the above-mentioned method and the actual incidence rate is only 0.0264/100 000 with an accurate rate of 97.94%. **Conclusion**

This principal factor analysis forecast method based on connection number in forecasting the incidence rate of encephalitis B prevention is acceptable.

**【Key words】** Encephalitis B; Forecast; Connection number; Principal factor

流行性乙型脑炎(乙脑)发生率的预测对其防治有重要意义,但这方面的文献不多,原因可能在于涉及到的预测因子较多,较难建立理想的数学预测模型。刘伟光等<sup>[1]</sup>采用模糊聚类预测法预测某地 1998 年乙脑发病率,取得较好的定性预测效果。我们根据集对分析理论和应用联系数的方法<sup>[2,3]</sup>,结合主成分分析思想,以文献[1]中给出的原始数据为实例,介绍用基于联系数的主因子分析预测法预测

乙脑发病率,不仅获得与文献[1]同样的定性预测结果,还进一步获得乙脑发生率的定量预测结果,与实际发生率有较高的符合程度。

## 基本原理

首先计算历史上预测因子与发病率的联系数,根据联系数值的大小依次排列出最主要因子、次主要因子、再次主要因子,剔除联系数值为最小的次要因子;再把新近出现的预测因子观测值与历史上同类因子值相比较,与之最接近的因子值为本次预测

用因子值,并根据该因子值与当时发病率的同一度建立预测方程,代入新因子值,解此方程得到本次预测值;当预测用因子数为  $n'$  ( $n' \geq 2$ ) 时,取  $n'$  个预测值的平均作为本次预测值。其具体步骤如下:

第一步:编秩。对表 1 中历史数据各年别的  $X_1, X_2, \dots, X_k$  和  $y$  按从小到大次序编上秩次 1、2、 $\dots$ 、 $k$ ,若有  $p$  个相同秩次的,则编以平均秩次。

第二步:写出  $X_k$  与  $y$  关于秩次对应情况的联系系数。分别观察各年别的  $X_k$  ( $k=1, 2, \dots, n$ ) 与  $y$  在秩次对上的同异反,秩次对是相同秩次的记为“同”,秩次对是(1)与(5)或(5)与(1)的记为“反”,其余情况的秩次对记为“异”;统计同、异、反对数,写出  $X_k$  与  $y$  关于秩次的同异反联系系数  $u_{(X_k, y)} = a + bi + cj$ ,其中  $a$  是相同的秩次对占总秩次数的比例, $b$  是相异的秩次对占总秩次数的比例, $c$  是相反的秩次对占总秩次数的比例;按联系系数定义, $j = -1$ ,  $i$  在  $[-1, 1]$  区间视不同情况取值。

第三步:找出  $X_k$  中的主要预测因子。具体是计算  $j = -1, i = 0$  ( $[-1, 1]$  区间的中间值)时  $u_{(X_k, y)}$  的值,称为联系数值,联系数值最大的  $X_k$  称为主要预测因子,联系数值次大的为次主要因子……,依次类推,当主要预测因子数  $n \geq 3$  时,按黄金分割法取前  $n' = 0.618n$  个因子作为主要预测因子参加预测;联系数值最小的因子为非主要因子,不参加预测。

第四步:确定预测用因子值。在被选用的  $X_k$  中,把被估计年  $m$  的  $X_k$  与已知年份的各年别的  $X_k$  作在被选用的  $X_k$  中,把对照比较,找出与之最接近的  $X_{kt}$  值,该  $X_{kt}$  即为被估计年乙脑发生率的预测用因子值,记为  $X_{kt}^*$ 。

第五步:计算预测值并做出预测。求解方程  $\frac{X_{kt}^*}{y_t} = \frac{X_{mk}}{y_m}$ ,求得  $y_m$ ;当  $X_{kt}^*$  有  $n'$  ( $n' \geq 2$ ) 个时,则以  $n'$  个  $y_m$  的平均值作为预测值。

### 实例分析

实例分析资料来自文献[1],预测用原始数据如表 1 所示。

按文献[1],乙脑可分为散发( $y \leq 1/10$  万)、小流行( $y = 1/10$  万 ~  $10/10$  万)和流行( $y > 10/10$  万)三类。

1. 主因子分析:从表 1 看出,  $X_k$  ( $k=1, 2, 3, 4$ ) 在各年别与  $y$  的秩次对应情况联系系数分别是

$$\begin{aligned} u_{(x_1, y)} &= \frac{1}{5} + \frac{2}{5}i + \frac{2}{5}j & u_{(x_2, y)} &= \frac{5}{5} + 0i + 0j \\ u_{(x_3, y)} &= \frac{3}{5} + \frac{2}{5}i + 0j & u_{(x_4, y)} &= \frac{5}{5} + 0i + 0j \end{aligned}$$

取  $i = 0, j = -1$ , 得各联系数值为  $u'_{(x_1, y)} = -\frac{1}{5}$ ,  $u'_{(x_2, y)} = 1$ ,  $u'_{(x_3, y)} = \frac{3}{5}$ ,  $u'_{(x_4, y)} = 1$ , 由此可见  $X_2, X_4$  这两个因子与  $y$  关于秩次的联系数值为最大,且都等于 1, 因此,这两个因子是预测用最主要因子。

$X_1$  这个因子与  $y$  关于秩次的联系数值为最小,所以不作为预测因子。

$X_3$  这个因子由于在 1996、1997 两年别上取值相同,在秩次上具有或然性,导致联系数值结果为  $\frac{3}{5}$ , 可看作次主要因子,这样连同  $X_1$  与  $X_4$  共有 3 个预测用主因子,按黄金分割法取  $3 \times 0.618 \approx 2$ , 综合前面分析,最后确定  $X_2$  与  $X_4$  为 1998 年乙脑预测用因子。

2. 主因子值:由于已知 1998 年的  $X_2 = 0.266$ ,  $X_4 = 0.289$ , 用观察法就可找出与之最接近的两个预测用因子值是  $X_{2(1990)}^* = 0.269$ ,  $X_{4(1991)}^* = 0.311$ 。

### 3. 预测结果:

表1 某地 1998 年乙脑发生率预测用原始数据

年份	当年 3 月份平均气温 ( $X_1, ^\circ\text{C}$ )	上年 8 月份发病率 ( $X_2, /10$ 万)	上年 9 月份发病率 ( $X_3, /10$ 万)	当年 7 月份发病率 ( $X_4, /10$ 万)	当年乙脑发病率 ( $y, /10$ 万)
1990	-0.3(3)	0.269(3)	0.358(3)	0.261(3)	1.22(3)
1991	-0.7(2)	0.349(4)	0.436(4)	0.311(4)	1.40(4)
1994	-0.8(1)	0.429(5)	0.513(5)	2.462(5)	12.48(5)
1996	0.8(5)	0.000(1)	0.085(1.5)	0.000(1)	0.17(1)
1997	0.5(4)	0.085(2)	0.085(1.5)	0.080(2)	0.58(2)
1998	-0.1	0.266	0.414	0.289	-

注:1998 年的  $y$  是当年乙脑发病率待预测;括号内数值是秩次

由预测方程  $\frac{X_{2(1990)}}{Y_{(1990)}} = \frac{X_{2(1998)}}{Y_{1(1998)}}$  即  $\frac{0.269}{1.22} = \frac{0.266}{Y_{1(1998)}}$  得

$$Y_{1(1998)} = 1.206$$

由预测方程  $\frac{X_{4(1990)}}{Y_{(1990)}} = \frac{X_{4(1998)}}{Y_{2(1998)}}$  即  $\frac{0.311}{1.40} = \frac{0.289}{Y_{2(1998)}}$  得

$$Y_{2(1998)} = 1.3012$$

$$\text{则 } Y_{(1998)} = \frac{1}{2} \sum_{i=1}^2 Y_{i(1998)} = \frac{1}{2} (1.206 + 1.3012) = 1.2536$$

即 1998 年该地乙脑的发病率预测值为 1.2536/10 万,该预测结果与实际发生率 1.28/10 万仅相差 0.0264/10 万<sup>[1]</sup>,相对误差为 2.0625%,准确率为 97.94%,与实际发生率符合较好。按发生程度的分类标准可知,该预测值位于小流行范围,所以本预测结果与文献[1]的预测结果完全相同,但比文献[1]更为优越的是,在于还给出了具体的发生率预测值。

### 讨 论

习惯上,对于多因子预测问题常采用多元回归方程处理,不仅计算量大,且要求有较大的样本含量,而本例只有 5 年的历史数据,所以不能使用多元回归方法处理;也不宜用灰色预测模型预测,因常用的灰色预测模型只适用于单一预测因子的预测,因此除采用模糊聚类预测法外<sup>[1]</sup>,目前对此类问题尚缺乏较为理想的预测方法。我们采用集对分析联系数与主成分分析相结合的思路,探索性地给出了一种基于联系数的预测方法,取得较好的预测结果,且方法简明,易为基层卫生工作者所掌握使用。

本文中给出的预测方法首先是利用了历史数据中所含有的秩次信息<sup>[4]</sup>和文献[5]中给出的基于集对分析的同异反数据结构分析思路,通过考察预测因子在历史上的数据与乙脑发病率  $y$  的系统结构关系,以联系数值的大小为依据和“黄金分割”原理,选出主要预测因子;其次又参考了“择近原则”<sup>[6]</sup>,按“择近原则”找出与当前主要预测用因子的值“最为接近”的同一类因子的历史数据值作为本次预测用因子值;此外还应用了“组合预测”的思想,即对有  $n'$  ( $n' \geq 2$ ) 个主要预测用因子值时,取  $n'$  个预测值的平均值作为预测结果,从而使本文所给出的预测方法具有较高的预测精度。

我们所用方法的合理性还在于正确地使用了联系数中的同一度这个概念,因为最后所采用的两个预测因子  $X_2 =$  上一年 8 月份乙脑发病率,  $X_4 =$  当年 7 月份乙脑发病率,本质上是该地乙脑在不同时间段上发病率的两个测度,这两个测度所含有的物理意义是完全同一的,所不同的仅仅是数量上的差异;相对而言,  $X_1 =$  当年 3 月份平均气温这个因子与乙脑发病率两者的物理意义是完全不同的,这或许是  $X_1$  这个因子在主因子分析过程中作为次要因子被剔除的一个本质上的原因;既然参加预测的因子与预测对象在本质上具有较强的同一性,利用一定的数学方法建立预测因子与预测对象的联系方程(本法中主要是线性方程),进而利用这种方程开展预测就有了客观依据,这是本文所给方法获得较好预测结果的一个重要原因。

本文所给的方法也有局限性,要求被分析的资料包括所有影响乙脑发病的重要因子,否则研究结果不够可靠。另外,预测结果不会超过历史上乙脑的最高发病率,但由于对乙脑一直在采取种种预防措施(如接种疫苗),新的发病率在非爆发年段一般不会超过历史最高发病率,因此本文给出的方法仍是一种可供基层卫生工作中采用的预测方法,而当有较多的历史数据可供筛选预测时,还可综合应用时间序列预测方法。

### 参 考 文 献

- 1 刘伟光,路秀萍,吴秋旺.用模糊聚类预测法预测流行性乙型脑炎.中国卫生统计,2001,18:163-164.
- 2 赵克勤,主编.集对分析及其初步应用.第1版.杭州:浙江科技出版社,2000.9-88.
- 3 覃杰,赵克勤.用联系数研究门诊人数增长与门诊医疗质量因子的关系.中国卫生统计,2003,20:137-139.
- 4 田凤调,主编.秩和比法的应用.第1版.北京:人民卫生出版社,2002.4-12.
- 5 周泽,赵毅毅,简贵儒,等.基于 SPA 的数据结构同异反分析在农业中的应用.农业系统科学与综合研究,2002,18:112-116.
- 6 高洁,盛昭瀚.集对分析聚类预测法及其应用.系统工程学报,2002,17:458-462.

(收稿日期:2003-11-26)

(本文编辑:张林东)