不确定性数据的交互分析方法

周爱民

（郑州大学 图书馆，郑州 450002）

摘 要：本文提出利用三维非完备线性等级对数模型，分析不确定数据的方法，主张用迭代法求对数模型的数据解，进一步分析哪一个模型最适合，此模型的交互项就是各变量之间的交互项，交互关系的确立可为排除共线性提供依据。

关键词：三维；对数模型；不确定性数据；交互分析

中图分类号：O213　　文献标识码：A　　文章编号：1002-6487（2005）08-0017-03

有些数据有上下界，但不知其具体值；有些数据是定性数据。这些数据就是不确定性数据。

对于有上下界，但不知具体值的不确定性数据，我们可根据其上下界，将其分类。对于分类数据，我们就可利用多维对数线性模型对其进行分析。由于四维以上数据分析难度较大，一般可按数据的主要属性将其分成三维。由于数据的频数，在一些格子中的数可能是零。所以，我们可以利用三维非完备对数线性模型，对其进行分析。

1　三维非完备对数线性模型

有些三维非完备等级对数线性模型可能有解析解，求解本身较容易。但是，大多数此类模型解析解不存在；即使存在，识别哪些问题存在解析解也很困难。单纯求出解析解，我们无法确定出这个解析解是哪一个模型的解，而我们识别交互作用的依据是对数模型本身，因此单纯求出一组解析解对我们识别交互作用没有提供任何信息。数值迭代解与模型紧密相关，不同的模型给出不同的数值迭代解。哪一组数值迭代解给出最大似然估计最小值，那一组数值迭代解就是最优解。求出最优解，我们也就求出了最优模型，也就识别出了交互作用。若交互作用是在因变量与自变量之间，那么，这个自变量就是主因素。

三维观察值用 X_{ijk} 来表示（$i=1\cdots I$；$j=1\cdots J$；$k=1\cdots K$）。我们用 m_{ijk} 表示真值，$\hat{m}_{ijk}$ 表示真值的拟合值。三维非完备等级对数线性模型有以下 8 种形式：

模型 1

$$1nm_{jik}=u+u_{1(i)}+u_{2(j)}+u_{3(k)}+u_{12(ij)+}u_{13(ik)+}u_{23(jk)}$$

其迭代拟合式为 $m_{ijk}^{(0)}=\delta_{jik};\delta_{jik}$ 表示（i, j, k）格观察值 x_{ijk} 非零时，$\delta_{jik}=1$；当观察值 x_{ijk} 为零时，$\delta_{jik}=0$

$$\begin{cases} m_{ijk}^{(3r-2)}=\dfrac{m_{ijk}^{(3r-3)}X_{ij+}}{\sum\limits_{k=1}^{K} m_{ijk}} \\ m_{jik}^{(3r-1)}=\dfrac{m_{ijk}^{(3r-2)}X_{+ij}}{\sum\limits_{i=1}^{I} m_{ijk}^{(3r-2)}} \\ m_{jik}^{(3r)}=\dfrac{m_{ijk}^{(3r-1)}X_{i+k}}{\sum\limits_{j=1}^{J} m_{ijk}^{(3r-1)}} \end{cases}$$

由于缺少 u_{123} 项，这个模型的自由度为 $df=(I-1)(J-1)(K-1)-Z_e$，其中 Z_e 为表中零的个数。

模型 2

$$1nm_{jik}=u+u_1+u_2+u_3+u_{12}+u_{23}$$

其迭代拟合式为

$$\begin{cases} m_{ijk}^{(2r-1)}=\dfrac{m_{ijk}^{(2r-2)}X_{ij+}}{\sum\limits_{k=1}^{K} m_{ijk}^{(2r-2)}} \\ m_{jik}^{(2r)}=\dfrac{m_{ijk}^{(2r-1)}X_{+ik}}{\sum\limits_{i=1}^{I} m_{ijk}^{(2r-1)}} \end{cases}$$

其中 $m_{ijk}^{(0)}=\delta_{jik}$，由于缺少 $u_{123}+u_{13}$ 项，其自由度为 $df=(I-1)(J-1)(K-1)+(I-1)(K-1)-Z_e$。

模型 3

$$1nm_{ijk}=u+u_1+u_2+u_3+u_{13}+u_{23}$$

其迭代拟合式为

$$\begin{cases} m_{ijk}^{(2r-1)}=\dfrac{m_{ijk}^{(2r-2)}X_{i+k}}{\sum\limits_{j=1}^{J} m_{ijk}^{(2r-2)}} \\ m_{jik}^{(2r)}=\dfrac{m_{ijk}^{(2r-1)}X_{+jk}}{\sum\limits_{i=1}^{I} m_{ijk}^{(2r-1)}} \end{cases}$$

由于缺少 $u_{123}+u_{12}$ 项，其自由度为 $df=(I-1)(J-1)(K-1)+(I-1)(J-1)-Z_e$。

模型 4

$$1nm_{ijk}=u+u_1+u_2+u_3+u_{12}+u_{13}$$

其迭代拟合式为

$$\begin{cases} m_{ijk}^{(2r-1)}=\dfrac{m_{ijk}^{(2r-2)}X_{ij+}}{\sum\limits_{k=1}^{K} m_{ijk}^{(2r-2)}} \\ m_{jik}^{(2r)}=\dfrac{m_{ijk}^{(2r-1)}X_{j+k}}{\sum\limits_{k=1}^{J} m_{ijk}^{(2r-1)}} \end{cases}$$

由于缺少 $u_{123}+u_{23}$ 项，其自由度为 $df=(I-1)(J-1)(K-1)+$

$(J-1)(K-1)-Z_e$。

模型 5

$$\ln m_{ijk}=u+u_1+u_2+u_3+u_{12}$$

其迭代拟合式为

$$m_{ijk}^{(2r)}=\frac{m_{ijk}^{(2r-1)}X_{ij+}}{\sum_{k=1}^{K}m_{ijk}^{(2r-1)}}$$

$$df=(I-1)\ (J-1)\ (K-1)+(I-1)\ (K-1)+(J-1)\ (K-1)-Z_e$$

模型 6

$$\ln m_{ijk}=u+u_1+u_2+u_3+u_{13}$$

其迭代拟合式为

$$m_{ijk}^{(r)}=\frac{m_{ijk}^{(r-1)}X_{i+k}}{\sum_{j=1}^{J}m_{ijk}^{(r-1)}}$$

由于缺少 $u_{123}+u_{12}+u_{23}$ 项，其自由度为

$$df=(I-1)\ (J-1)\ (K-1)+(I-1)\ (J-1)+(J-1)\ (K-1)-Z_e。$$

模型 7

$$\ln m_{ijk}=u+u_1+u_2+u_3+u_{23}$$

其迭代拟合式为

$$m_{ijk}^{(r)}=\frac{m_{ijk}^{(r-1)}X_{+jk}}{\sum_{i=1}^{I}m_{ijk}^{(r-1)}}$$

由于缺少 $u_{123}+u_{12}+u_{13}$ 项，其自由度为

$$df=(I-1)\ (J-1)\ (K-1)+(I-1)\ (J-1)+(I-1)\ (K-1)-Z_e。$$

模型 8

$$\ln m_{ijk}=u+u_1+u_2+u_3$$

其迭代拟合式为

$$\begin{cases} m_{ijk}^{(3r-1)}=\dfrac{m_{ijk}^{(3r-2)}X_{i++}}{\sum_{j=1}^{J}\sum_{k=1}^{K}m_{ijk}^{(3r-2)}} \\ m_{ijk}^{(3r-2)}=\dfrac{m_{ijk}^{(3r-3)}X_{+j+}}{\sum_{i=1}^{I}\sum_{k=1}^{K}m_{ijk}^{(3r-3)}} \\ m_{ijk}^{(3r)}=\dfrac{m_{ijk}^{(3r-1)}X_{++k}}{\sum_{i=1}^{I}\sum_{j=1}^{J}m_{ijk}^{(3r-1)}} \end{cases}$$

自由度为

$$df=(I-1)\ (J-1)\ (K-1)+(I-1)\ (J-1)+(I-1)\ (K-1)+\ (J-1)\ (K-1)-Z_e。$$

2 模型优选

令 $$X^2=\sum_{i\ j\ k}\frac{(X_{ijk}-\hat{m}_{ijk})}{\hat{m}_{ijk}}$$

若 $X^2<X_\alpha^2(df)$ 则模型有统计意义，否则无统计意义。把 8 个模型中有统计意义的模型作为备选模型，在其中选最佳模型。

规定：对非零的 X_{ijk}，定义

$$G^2=2\sum_{i\ j\ k}\ln\frac{X_{ijk}}{\hat{m}_{ijk}}$$

对模型可求出 G^2。把一个模型含有另一模型，且比其只多一项的模型叫做相邻模型。

模型优选时，首先以第 8 模型为准佳模型，用相邻模型 i 与其比较。看不等式

$$G_8^2-G_i^2<X_\alpha^2(df_8-df_i)$$

是否成立。若成立，说明两模型相差的那一项不显著，可以为零。8 个模型中所有含这一项的模型都不是最佳模型，不用再考虑，在剩余的模型中继续比较。若不等式不成立，则说明两相邻模型相差的那一项不可忽略，必须保留；然后以这个相邻模型为准佳模型；继续在其相邻的模型中找显著项，显著项都找出来，就得到最佳模型。

3 用交互作用分析法分析书刊流失过程中的主要交互作用

科学的性质要求一个完整的研究必须理论联系实际，能够比较准确地刻划事物的现状，把握它的内在规律，解决一些相关的实际问题。

书刊流失一直是图书馆一个老话题，为了防止流失量过大，图书馆一般规定流失量不应超过千分之三。就这个问题《数理统计与管理》1994 年 1 期发表了《书刊合理流失量的确定及评价考核的统计方法》一文，通过数据分析，此文提出书刊流失与书库多少无关，与读者量大小相关。见表 1。

表 1

指标 \ 数量 \ 室名	文一	文二	文三	学理	原版	教文	教理	港台
88 藏书(册)	4820	4616	2804	2574	10057	15869	5529	9583
88 年实丢册数	24	13	19	7	4	0	0	0
3‰允许丢失数	14	14	8	8	30	48	17	4
馆员数	2	2	2	2	1	1	1	1
日均借阅人次	336	310	91	56	10	7	4	9
90 年藏书(册)	5671	5725	3522	2710	11113	17230	5903	1067
90 年实丢册数	12	8	35	12	0	0	0	1
30‰允许丢失量	17	17	11	8	33	52	18	3
馆员数	2	2	2	2	1	1	1	1
日均借阅人次	108	188	118	31	3	4	4	3

此文为书刊流失提供了一个实际的的原材料（表 1），这是难能可贵的。但在对书刊流失量 (y)，藏书量 (X_1)，读者日均人数 (X_2) 作多元分析时，单纯使用两两变量的简单相关系数，以此作为自变量筛选的依据，这可能有虚假性，可能不反映问题的真实情况。

我们把此数据按本文的方法重新分组。

流失量 i 分为三个等级

0–5（本） i=1

6–10（本） i=2

11 本以上 i=3

单馆员接待读者日均量 j 也分为三个等级

0-30(人)　　j=1
31-60(人)　　j=2
61 以上(人)　　j=3
藏书量 k 也分为三个等级
0-5 千　　k=1
5 千零 1-1 万　　k=2
1 万零 1 以上　　k=3

显然 i、j、k 都是不确定性数据。

我们把数据整理如表 2。

表 2

i	j=1			j=2			j=3		
	k=1	k=2	k=3	k=1	k=2	k=3	k=1	k=2	k=3
1	1	3	4	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1	0
3	1	0	0	2	1	0	2	0	0

模型 1　迭代结果与表 2 相同。

模型 2　迭代结果如表 3：

表 3

	j=1			j=2			j=3		
i	k=1	k=2	k=3	k=1	k=2	k=3	k=1	k=2	k=3
1	1.00013	2,9999	3.99999	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1	0
3	1	0	0	2	1	0	2	0	0

模型 3　迭代结果如表 4：

表 4

	j=1			j=2			j=3		
i	k=1	k=2	k=3	k=1	k=2	k=3	k=1	k=2	k=3
1	1	3	4	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1	0
3	1.0001	0	0	1.9999	1	0	1.9999	0	0

模型 4　迭代结果如表 5：

表 5

	j=1			j=2		j=3			
i	k=1	k=2	k=3	k=1	k=2	k=3	k=1	k=2	k=3
1	1	3	4	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1	0
3	1	0	0	1.9999999999999991	1.0000000000000069	0	2	0	0

模型 5　迭代结果如表 6：

表 6

	j=1			j=2			j=3		
i	k=1	k=2	k=3	k=1	k=2	k=3	k=1	k=2	k=3
1	2.6666666666666665	2.6666666666666665	2.6666666666666665	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1	0
3	1	0	0	1.5	1.5	0	2	0	0

模型 6　迭代结果如表 7：

表 7

	j=1			j=2			j=3		
i	k=1	k=2	k=3	k=1	k=2	k=3	k=1	k=2	k=3
1	1	3	4	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1	0
3	1.6666666666666667	0	0	1.6666666666666667	1	0	1.6666666666666667	0	0

模型 7　迭代结果如表 8：

模型 8　迭代结果如表 9：

表 8

i	j=1			j=2			j=3		
	k=1	k=2	k=3	k=1	k=2	k=3	k=1	k=2	k=3
1	1	3	4	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1	0
3	1	0	0	2	1	0	2	0	0

表 9

	j=1			j=2			j=3		
i	k=1	k=2	k=3	k=1	k=2	k=3	k=1	k=2	k=3
1	1.8572	2.14778	3.99847	0	0	0	0	0	0
2	0.75736	0	0	0	0	0	0	1.24233	0
3	1.23917	0	0	1.3911	1.6088			0	0

通过数据迭代我们发现只有模型 1 与模型 7很快达到原表值,没有任何误差,其余模型都有误差。这说明这两个模型拟合程度最好,但考虑到自由度,模型 7 是最佳模型。这说明藏书量与读者量有较大的相关性。反过来看原数据表一,我们可以发觉藏书量越大,读者越少,藏书量少的书库,读者相对较多。二者是负相关,这说明交互作用分析法的结论是正确的。

既然藏书量与读者量负相关,那么,说明藏书量大的书库,一定是可读性差的书库;藏书量小的书库一定是可读性强的书库。原文把可读程度不同的书库进行流失比较得出书库数量与书刊流失不相关的结论缺乏说服力。

若我们有可读性强的大中小三种藏书量书库的流失数据,有可读性中的大中小三种藏书量书库流失数据和可读性差的大中小三种藏书量书库的流失数据,那么我们就能分析出藏书量大小与书刊流失量有无关系。只有在各种藏书量处于平等的位置下来讨论藏书量与流失量有无关系,结论才会正确,才有说服力。

4　不确切数据交互分析的意义

交互分析的意义主要有以下三点:

(1)模型的交互项是两自变量的交互项,那么,这两个自变量共线,建模分析时必须排除共线的影响。

(2)模型的交互项是因变量与自变量的交互项,那么,这个自变量就是因变量的主要因素,这个方法就成了主因素分析方法的一种。

(3)确切数据是不确切数据的特例,所有确切数据都可当着不确切数据来处理。

参考文献:

[1]Bishop,Y.M.M.著.离散多元分析理论与实践[M].张尧庭译.北京:中国统计出版社,1998.

[2]孙姹姹,董云河.书刊合理流失量的确定及评价考核的统计分析方法[J].数理统计与管理,1994,(1).

(责任编辑/李友平)