高维列联表资料的统计分析与 SAS 软件实现(五)

柳伟伟，胡良平
军事医学科学院生物医学统计学咨询中心，北京 100850

关键词：统计学；医学；数据分析，统计；定性资料；SAS 软件

Liu WW, Hu LP. *J Chin Integr Med*. 2010; 8(3): 287-291.
Received January 24, 2010; accepted February 25, 2010; published online March 15, 2010.
Indexed/abstracted in and full text link-out at PubMed. Journal title in PubMed: *Zhong Xi Yi Jie He Xue Bao*.
Free full text (HTML and PDF) is available at http://www.jcimjournal.com.
Forward linking and reference linking via CrossRef.
DOI: 10.3736/jcim20100315

Statistical analysis for data of multidimensional contingency table with SAS software package (Part five)

Wei-wei LIU, Liang-ping HU
Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, Beijing 100850, China

Keywords: statistics; medicine; data analysis, statistical; qualitative data; SAS software

针对高维列联表资料的分析，分别介绍了加权 χ^2 检验、CMH χ^2 检验、对数线性模型、结果变量为二值变量或多值有序变量高维列联表资料的 logistic 回归分析[1-4]。本文将讨论结果变量为多值名义变量的高维列联表资料的 logistic 回归分析，即采用扩展的多重 logistic 回归模型，也称为多项 logit 模型。在结果变量为多值名义变量的高维列联表资料中，结果变量是多值名义变量，原因变量可以是二值变量、多值名义变量或多值有序变量。在使用 logistic 回归模型进行分析时，原因变量也可以是连续变量，当然，此时资料就不再适合以列联表的形式呈现了，而需要以数据库(一行代表一个受试者的全部信息，而一列代表一个原因或结果)的形式呈现。此外，回归模型中还可以包括各个原因变量之间的交互项。

1 实例及数据结构

例 1 某研究小组研究髓过氧化物酶-463G/A 基因的多态性分型，探讨其与冠状动脉疾病(coronary artery disease, CAD)的关系。病例组均行冠状动脉造影证实患有 CAD，正常对照组冠状动脉造影示正常，均用聚合酶链反应-限制性片段长度多态性方法判断各组基因型，结果见表 1。试分析基因型频率与患病情况及性别之间的关系。

表 1 患病情况、性别与基因型频率的关系

分组	性别	例数		
		基因型: GG	GA	AA
对照组	男	46	38	9
	女	37	23	7
病例组	男	104	14	4
	女	68	15	2

此表中数据为分析需要而作了修改，不反映真实情况。

例 2 在某社区随机抽取了 100 名育龄妇女进行问卷调查，了解她们的年龄(A)、生育史(B)、收入(C)及采取的避孕方式(Y)，评价哪些因素会影响育龄妇女对避孕方式的选择。将调查对象按年龄分为小于 30 岁组(A=1)，30～35 岁组(A=2)，35 岁以上组(A=3)；按生育史分为未生育组(B=0)和已生育组(B=1)；按收入水平分为低收入组(C=1)和高收入组(C=2)；选择的避孕方式有宫内节育器

Correspondence: Prof. Liang-ping HU; Tel: 010-66931130; E-mail: lphu812@sina.com

(Y=1)、口服避孕药(Y=2)和避孕套(Y=3)3 种。相关资料见表 2。

表 2 育龄女性选择避孕方式的影响因素

编号	年龄(A)	生育史(B)	收入(C)	避孕方法(Y)	编号	年龄(A)	生育史(B)	收入(C)	避孕方法(Y)
1	2	1	2	2	51	2	1	1	1
2	2	1	1	1	52	3	0	2	3
3	1	1	2	2	53	2	1	2	2
…	…	…	…	…	…	…	…	…	…
49	1	0	2	2	99	3	1	1	1
50	2	1	2	2	100	1	0	2	3

2 SAS 软件实现与结果解释

2.1 SAS 程序及其说明 在 SAS 软件中,主要是通过 logistic 过程来实现 logistic 回归分析,其他可以进行 logistic 回归分析的过程还包括 catmod 过程、genmod 过程和 phreg 过程。这里需要特别指出的是,在 SAS 较早的版本中,logistic 过程无法实现多项 logit 模型,只能由 catmod 过程完成这一分析,在 SAS 9.1 及以后的版本中,logistic 过程已增加了这一功能,并且两个过程的输出结果是一致的[5]。本文将提供这两种实现途径的程序写法。

对例 1 中资料进行分析的 SAS 程序如下,程序名 prg1.sas:

```
data exam1;
  do a=1 to 2;
    do b=1 to 2;
      do c=1 to 3;
        input f@@;
        output;
      end;
    end;
  end;
cards;
104  14  4
68   15  2
46   38  9
37   23  7
;
run;
```

```
proc logistic data=exam1;
  model c=a b/link=glogit;
  weight f;
run;

proc catmod data=exam1;
  direct a b;
  model c=a b;
  weight f;
run;
```

首先建立数据集,本例使用循环方式读入表 1 中每个格子里的频数,变量 a、b、c、f 分别代表患病情况、性别、基因型和频数。其中 a 的取值 1、2 分别表示病例组和对照组,b 的取值 1、2 分别表示男性和女性,c 取 1、2、3 分别表示基因型 GG、GA 和 AA。

然后使用 logistic 过程拟合多项 logit 模型。在语句"model c=a b/link=glogit;"中,指定 c 为因变量,a 和 b 为自变量;选项 link=glogit 要求对多值的因变量拟合多项 logit 模型,在这里这个选项是必须的,因为在默认状态下,logistic 过程会将多值因变量作为多值有序变量来处理,拟合累积logistic 回归模型。语句"weight f;"指定 f 为频数变量,这是由于本例是以列联表的形式呈现并录入数据,如果是以原始数据库形式录入数据的话,则不需要写出该语句。

最后通过 catmod 过程拟合多项 logit 模型。为了与 logistic 过程的输出结果保持一致,此处使用了语句"direct a b;"将二值变量 a 和 b 指定为连续变量,然后按照处理连续变量的方式将它们纳入模型。

上述程序中尤其需要注意的是,在 logistic 过程中,如果不用 class 语句进行指定,SAS 系统会将 model 语句中列出的所有自变量作为连续变量处理,平时我们在进行分析时,经常是将二值变量和多值有序变量直接代入模型,并不专门使用 class 语句指定它们为分类变量。当自变量为多值名义变量时,则需要用 class 语句进行定义,此时用户无须专门再去为多值名义变量生成哑变量。而在 catmod 过程中,如果不用 direct 语句指定一个变量为连续变量,SAS 系统会将 model 语句中列出的所有自变量作为分类变量处理。这就是这两个过程在处理自变量方式上的区别,正因为如此,有的用户在分别使用这两个过程进行分析时,得到的结果会有所不同。

当然,在对该数据进行分析时,也可以在 logistic 过程中使用 class 语句定义 a 和 b 为分类变量;在 catmod 过程中去掉"direct a b;"语句,按照默认状态将它们作为分类变量处理。此时两个过程得到的结果仍然一致。

将二值变量 a 和 b 分别作为连续变量和分类变量处理,这两种情况下得到的回归方程会有所不同,但是对于自变量进行检验的结论是一致的。在对发生概率进行预测时,两者的计算结果也是相同的,只不过在将自变量代入回归方程时,其代入方式略有不同,关于这一点,在后面对输出结果进行解释时,会做进一步的说明。

在对例 2 中的资料进行分析时,其 SAS 程序的写法、输出结果及解释与例 1 大体相同,限于篇幅此处从略。由于例 2 中是以原始的数据库形式录入数据,所以将 logistic 过程和 catmod 过程中的语句"weight f;"去掉就可以了。当然,数据步中不需要写"do-end"语句,只需要有一个 input 语句就可以了。

2.2 SAS 输出结果的解释 由于程序 prg1.sas 中 logistic 过程和 catmod 过程的主要输出结果内容一致，本文仅列出 logistic 过程的部分结果如下：

Model information

Data set	Work.Exam1
Response variable	c
Number of response levels	3
Weight variable	f
Model	Generalized logit
Optimization technique	Fisher's scoring

Logits modeled use c=3 as the reference category.

这是关于模型信息的说明，其中需要指出的是本例中使用的模型是广义 logit 模型，也就是多项 logit 模型。另外，在建模时是以 c=3 为参照类，也就是以基因型 AA 为参照类估计方程中的参数值。Logistic 过程在拟合多项 logit 模型时，默认方式是将结果变量的最后一个类别作为参照类，这里 3 就是变量 c 取值顺序中的最后一位，这一点要注意与二值 logistic 模型以及累积 logistic 模型有所区别。如果想要改变参照类，比如以基因型 GG 为参照类别，只需要进行一定的设置即可[6]。

Testing global null hypothesis: BETA=0

Test	Chi-square	*df*	*Pr*>ChiSq
Likelihood ratio	42.059 7	4	<0.000 1
Score	41.671 2	4	<0.000 1
Wald	38.814 9	4	<0.000 1

这是对整个模型进行假设检验的结果，它的原假设是所有的回归系数都为 0，分别使用似然比检验、计分检验和 Wald 检验 3 种方法。检验结果中依次给出了 χ^2 值、自由度和 P 值，可以看出，3 种方法的 P 值都小于 0.000 1，可以认为该模型是成立的。

Type 3 analysis of effects

Effect	*df*	Wald chi-square	*Pr*>ChiSq
a	2	38.786 6	<0.000 1
b	2	0.022 0	0.989 1

这是对各个自变量所产生的效应进行分析的结果，这部分显示了各个自变量对结果变量的“整体”作用，该检验的原假设为原因变量对所有 logit 中的任何一个都没有作用。可以看到，患病情况 a 对于结果变量的影响有统计学意义，而性别 b 的作用没有统计学意义。

Analysis of maximum likelihood estimates

Parameter	c	*df*	Estimate	Standard error	Wald chi-square	*Pr*>ChiSq
Intercept	1	1	5.034 4	1.086 1	21.487 4	<0.000 1
Intercept	2	1	1.734 8	1.158 3	2.243 2	0.134 2
a	1	1	−1.709 5	0.497 0	11.830 9	0.000 6
a	2	1	−0.237 5	0.529 2	0.201 5	0.653 5
b	1	1	0.021 9	0.461 0	0.002 2	0.962 2
b	2	1	0.055 4	0.483 4	0.013 1	0.908 8

Odds ratio estimates

Effect	c	Point estimate	95% wald confidence limits	
a	1	0.181	0.068	0.479
a	2	0.789	0.280	2.225
b	1	1.022	0.414	2.523
b	2	1.057	0.410	2.726

这是参数估计的结果，作为因变量的基因型有 3 个水平，需要建立两个 logit 模型，每个 logit 模型对应一组参数，共有两组参数估计值。用 $P(GG)$、$P(GA)$和 $P(AA)$分别表示 3 种基因型出现的概率，模型的表达式可写为：

$$\ln\left[\frac{P(GG)}{P(AA)}\right]=5.034\ 4-1.709\ 5a+0.021\ 9b$$

$$\ln\left[\frac{P(GA)}{P(AA)}\right]=1.734\ 8-0.237\ 5a+0.055\ 4b$$

某受试对象归入 3 种基因型的概率分别为：

$$P(GG)=\frac{\exp(5.034\ 4-1.709\ 5a+0.021\ 9b)}{1+\exp(5.034\ 4-1.709\ 5a+0.021\ 9b)+\exp(1.738\ 4-0.237\ 5a+0.055\ 4b)}$$

$$P(GA)=\frac{\exp(1.738\ 4-0.237\ 5a+0.055\ 4b)}{1+\exp(5.034\ 4-1.709\ 5a+0.021\ 9b)+\exp(1.738\ 4-0.237\ 5a+0.055\ 4b)}$$

$$P(AA)=\frac{1}{1+\exp(5.034\ 4-1.709\ 5a+0.021\ 9b)+\exp(1.738\ 4-0.237\ 5a+0.055\ 4b)}$$

综上所述，基因型的分布与患病情况有关，而与性别因素没有关系。

利用上述方程预测基因型的分布概率时，直接将自变量的具体取值代入方程即可，以基因型 AA 为例，当受试对象属于病例组并且为女性时，其基因型为 AA 的概率为 0.028 5。

若将二值变量 a 和 b 作为分类变量处理，也就是在 logistic 过程中使用 class 语句定义 a 和 b 为分类变量或在 catmod 过程中去掉“direct a b;”语句时，可以得到如下的参数估计结果：

Analysis of maximum likelihood estimates

Parameter		c	*df*	Estimate	Standard error	Wald chi-square	*Pr*>ChiSq
Intercept		1	1	2.502 9	0.252 1	98.578 5	<0.000 1
Intercept		2	1	1.461 6	0.268 1	29.713 0	<0.000 1
a	1	1	1	0.854 8	0.248 5	11.830 9	0.000 6
a	1	2	1	0.118 8	0.264 6	0.201 5	0.653 5
b	1	1	1	−0.010 9	0.230 5	0.002 2	0.962 2
b	1	2	1	−0.027 7	0.241 7	0.013 1	0.908 8

模型的表达式可写为：

$$\ln\left[\frac{P(GG)}{P(AA)}\right]=2.502\ 9+\beta_{1i}^{i}+\beta_{i2},\ i,\ j=1,2$$

$\beta_{11}^{1}=0.854\ 8$, $\beta_{11}^{2}=-0.854\ 8$, $\beta_{12}^{1}=-0.010\ 9$, $\beta_{12}^{2}=0.010\ 9$

$$\ln\left[\frac{P(GA)}{P(AA)}\right]=1.461\ 6+\beta_{21}^{i}+\beta_{i2},\ i,\ j=1,2$$

$\beta_{21}^{1}=0.118\ 8$, $\beta_{21}^{2}=-0.118\ 8$, $\beta_{22}^{1}=-0.027\ 7$, $\beta_{22}^{2}=0.027\ 7$

其中 β_{11}^{1} 表示在第一个 logit 模型中，患病情况(变量 a)取值为 1 时的回归系数，其取值为 0.854 8；β_{11}^{2} 表示在第一个 logit 模型中，患病情况取值为 2 时的回归系数，其取值为 −0.854 8。符号 β_{11}^{1} 中上标位置的数字表示自变量的取值为 1；下标位置的第一个数字表示第一个 logit 模型，下标位置的第二个数字表示第一个自变量，也就是变量 a。

Logistic 过程或 catmod 过程在进行参数估计时，对于自变量为分类变量的情形，是以自变量的一个水平作为参照估计出多个参数，默认状态下是以自变量的最后一个取值为参照。本例中变量 a 和 b 都有两个水平，所以在每个 logit 中各有一个参数估计值，在输出结果中其变量名后紧跟着的数字是 1，说明这是其取值为 1 时的参数估计值。同时，SAS 系统对自变量各水平的参数施加“合计为 0”的限制条件，所以省略类别的参数值可以由其他类别很容易地计算出来，也就是当 a 和 b 的取值为 2 时，它们的参数估计值就是其取值为 1 时回归系数值的相反数[7]。以变量 a 在第一个 logit 中的回归系数为例，当它取值为 1 时，回归系数的值为 0.854 8；当它取值为 2 时，回归系数值正好是 0.854 8 的相反数，即 −0.854 8。使用这种情况下得到的方程预测发生概率时，就不能将自变量直接代入方程，而是应该根据自变量的不同取值选择其对应的回归系数值进行计算。仍以基因型 AA 为例，当受试对象属于病例组并且为女性，也就是变量 a 和 b 取值分别为 1 和 2 时，变量 a 在两个 logit 模型中的回归系数值分别为 0.854 8 和 0.118 8，变量 b 在两个 logit 模型中的回归系数值分别为 0.010 9 和 0.027 7，经计算可得其基因型为 AA 的概率为 0.028 5，这与前述我们得到的结果一致。

此外，在 logistic 过程或 catmod 过程中可以通过相应的选项或语句输出预测概率，当受试对象属于病例组并且为女性时，这两个过程输出的基因型为 AA 的预测概率都为 0.028 5。具体的程序和输出结果本文不再赘述，感兴趣的读者可自行尝试。

由于变量 b 的作用没有统计学意义，所以需要在模型中将其去掉，在程序中只要在语句“model c=a b/link=glogit;”或“model c=a b;”中删掉 b 就可以了，重新拟合的结果此处从略。

3 讨 论

在使用扩展的 logistic 回归模型对结果变量为多值名义变量的高维列联表资料进行分析时，需要建立多个 logit 模型，并且对每一个 logit 模型都要进行参数估计，因此，当结果变量有 K 个取值时，可以得到 $K-1$ 个回归方程，也就是有 $K-1$ 组参数估计值。在 SAS 程序中，对二值变量采用不同的处理方式，得到的方程会有所不同，使用者要注意其中的区别，当利用回归方程进行预测时，应该正确辨别方程的形式，根据具体形式选择计算的方式，不能盲目地将自变量按其取值直接代入方程。

REFERENCES

1 Ge Y, Hu LP. Statistical analysis for data of multidimensional contingency table with SAS software package (Part one). J Chin Integr Med. 2009; 7(11): 1086-1089. Chinese.
葛毅，胡良平. 高维列联表资料的统计分析与 SAS 软件实现(一). 中西医结合学报. 2009; 7(11): 1086-1089.

2 Ge Y, Hu LP. Statistical analysis for data of multidimensional contingency table with SAS software package (Part two). J Chin Integr Med. 2009; 7(12): 1188-1192. Chinese.
葛毅，胡良平. 高维列联表资料的统计分析与 SAS 软件实现(二). 中西医结合学报. 2009; 7(12): 1188-1192.

3 Ge Y, Hu LP. Statistical analysis for data of multidimensional contingency table with SAS software package

(Part three). J Chin Integr Med. 2010; 8(1): 90-94. Chinese.
葛毅, 胡良平. 高维列联表资料的统计分析与SAS 软件实现(三). 中西医结合学报. 2010; 8(1): 90-94.

4 Gao H, Hu LP, Guo J, Li CP. Statistical analysis for data of multidimensional contingency table with SAS software package (Part four). J Chin Integr Med. 2010; 8(2): 186-189. Chinese.
高辉, 胡良平, 郭晋, 李长平. 高维列联表资料的统计分析与SAS 软件实现(四). 中西医结合学报. 2010; 8(2): 186-189.

5 SAS Institute Inc. SAS/Stat 9.2 user's guide. Cary, NC: SAS Institute Inc. 2008: 1091-1228, 3253-3474.

6 Hu LP. Medical statistics: analysis of quantitative and qualitative data using the triple-type theory. Beijing: People's Military Medical Press. 2009: 376-393. Chinese.
胡良平. 医学统计学——运用三型理论分析定量与定性资料. 北京: 人民军医出版社. 2009: 376-393.

7 Wang JC, Guo ZG. Logistic regression models: methods and application. Beijing: Higher Education Press. 2001: 249-262. Chinese.
王济川, 郭志刚. Logistic 回归模型——方法与应用. 北京: 高等教育出版社. 2001: 249-262.

第五届全国中西医结合围手术期医学研讨会征文通知

为了促进现代外科理论和中西医结合围手术期管理研究的学术探讨和经验交流，推动本领域临床实践和研究的不断深入，中国中西医结合学会围手术期专业委员会定于 2010 年 5 月或 6 月在浙江省杭州市举办“第五届全国中西医结合围手术期医学研讨会”。届时将邀请美国、日本的国际知名外科专家及国内本领域的院士和专家作专题报告，介绍 21 世纪外科围手术期医学的新理论、新思路和新发展，以及中西医结合围手术期管理医学的最新进展，开展学术交流。现征文如下。

1 征文内容 (1)围手术期管理医学研究的新进展，包括围手术期管理医学研究的理论探讨，快速康复外科新理念研究的前沿动态，中医、中西医结合方法在术后快速康复中的应用展望；(2)快速康复外科与中医药和中西医结合医学研究，包括快速康复技术应用的中西医结合临床经验交流，围手术期营养支持研究，微创新技术、新成果研究，相关临床研究、基础研究等。

2 征文要求 (1)论文内容真实可靠，具备科学性、先进性、实用性。(2)全文 3 000 字左右，文稿需附 400 字左右中文摘要；(3)寄送全文、摘要打印稿各 1 份(A4 纸，加盖公章)，或用 E-mail 发送电子版；(4)来稿免收审稿费，但请务必注明作者姓名、工作单位、通讯地址、邮政编码和联系电话，是否同意参加大会交流；(5)论文截稿日期：2010 年 4 月 15 日。

3 联系方式 联系人：张勤(13588887282)，周济春(13777361870)；传真：0571-87077785；E-mail：zjc0305@live.cn；联系地址：浙江省杭州市邮电路 54 号浙江省中医院外科；邮政编码：310006。

4 其他事宜 采取大会演讲与讨论相结合的形式，参加会议者授予国家级继续教育学分 6 分。会议具体时间、地点和费用请见第二轮会议通知。

中国中西医结合学会

2010 年 1 月 8 日