Medical Statistics 医学统计学

高维列联表资料的统计分析与 SAS 软件实现(四)

高辉[1]，胡良平[1]，郭晋[1]，李长平[2]
1. 军事医学科学院生物医学统计学咨询中心，北京 100850
2. 天津医科大学公共卫生学院，天津 300070

关键词：统计学；医学；数据分析，统计；定性资料；SAS 软件

Gao H, Hu LP, Guo J, Li CP. *J Chin Integr Med*. 2010; 8(2): 186-189.
Received December 24, 2009; accepted January 25, 2010; published online February 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/jcim20100215

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

Hui GAO[1], Liang-ping HU[1], Jin GUO[1], Chang-ping LI[2]
1. Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, Beijing 100850, China
2. School of Public Health, Tianjin Medical University, Tianjin 300070, China

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

生物医学研究，尤其是临床医学研究中经常需要考察 k 个($k \geqslant 2$)定性原因变量对多值有序结果变量的影响，这类资料被称为结果变量为多值有序变量的高维列联表资料。如某多中心临床试验中，原因变量为“不同就诊医院(A 医院、B 医院、C 医院)”和“药物种类(甲药、乙药)”，结果变量为“疗效(治愈、显效、好转、无效)”。本文结合医学研究实例，介绍如何正确分析此类资料及分析方法的 SAS 软件实现。

1 实际问题

例 1 有甲、乙两所医院，近 5 年共收治陈旧性踝关节骨折脱位患者 72 例，其中甲医院收治 37 例，乙医院收治 35 例。由于很多患者无法回忆起确切的骨折损伤日期，只能记起受伤大概所处的时间段，故将受伤后至手术之间的时间间隔划分为 3 段：21～40 d，41～90 d，91 d 以上。有关数据见表 1，请比较伤后手术间隔时间与疗效之间是否有关联？

表 1 72 例陈旧性踝关节骨折脱位患者损伤后手术时间间隔与疗效评定

就诊医院	伤后手术间隔时间(d)	例数			
	疗效：	优良	较好	一般	不佳
甲医院	21～40	12	4	3	0
	41～90	4	2	3	1
	≥91	2	1	3	2
乙医院	21～40	9	5	2	1
	41～90	3	4	2	1
	≥91	1	2	3	2

例 2 同例 1，但数据是以数据库形式记录，具体数据见表 2。其中，id 代表患者序号，a 代表“就诊医院”，a＝1、2 分别代表甲、乙医院；b 代表“伤后手术间隔时间”，b＝1、2、3 分别代表 21～40 d、41～90 d、≥91 d；y 代表“疗效”，y＝1、2、3、4 分别代表优良、较好、一般、不佳。

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

表 2 72 例陈旧性踝关节骨折脱位患者损伤后手术时间间隔与疗效评定

id	a	b	y
1	2	2	2
2	1	3	3
3	1	1	1
…	…	…	…
72	2	1	4

2 研究目的及分析方法

表 1 是以列联表形式呈现的，资料中包含两个原因变量："就诊医院"和"伤后手术间隔时间"。结果变量"疗效"是多值有序的，所以此资料是结果变量为多值有序变量的高维列联表资料。表 2 是以数据库形式呈现的，其性质同表 1，只是呈现形式不同而已。

分析结果变量为多值有序变量的高维列联表资料，若研究者考察资料的整体情况，一般可选择 CMH 校正的秩和检验或有序变量多重 logistic 回归分析。

2.1 *CMH 校正的秩和检验* 采用 CMH 统计分析方法分析高维列联表资料时，应注意结合资料的类型和研究目的选择合适的统计量。当行变量与列变量均为有序变量时，若只研究行变量不同水平之间在结果上差异是否有统计学意义，应采用行平均得分差异统计量(本质上是秩和检验统计量)；若希望研究行变量与列变量之间是否存在相关关系时，应采用非零相关统计量(本质上是秩相关分析的检验统计量)。当行变量为无序变量而列变量为有序变量时，采用行平均得分差异统计量即可。

2.2 *有序变量多重 logistic 回归分析* 若采用有序变量多重 logistic 回归分析，首先需要对平行线假设条件进行统计检验。如果这一假设条件被拒绝，便说明累积 logistic 回归模型不适合，需要采用其他模型来进行资料分析，如可引入二次项或交互项等复杂模型。

另外，在运用 SAS 软件进行有序变量多重 logistic 回归分析时，应注意原因变量的属性。如果原因变量是连续变量，一般不需要进行处理，可直接建立 logistic 回归模型[1]。但有时根据专业知识需对其进行分级以获得更有实际意义的结果时，连续变量就转换成了有序变量，此时可按影响结果变量由小到大的顺序赋值为 1、2……，并将它当作连续型变量处理或直接引入哑变量，建立 logistic 回归模型。比如在肺癌危险因素的病例对照研究中，研究者往往感兴趣的是年龄每增加 5 岁(根据专业知识和研究目的决定)，肺癌发病的危险性是基础状态时的多少倍；而年龄每增加 1 岁，肺癌发病的危险性是基础年龄时的多少倍往往没有多大实际意义。

如果原因变量是二值变量，一般可按 0、1 赋值。如果原因变量是多值名义变量，需引入哑变量，每个哑变量都是一个二值变量，所需哑变量的数目为多值名义变量的类别数减 1。

3 SAS 软件实现及结果解释

3.1 *SAS 程序及其用法* 对例 1，若采用 CMH 校正的秩和检验来处理，所用 SAS 程序如下，程序名为 exam1_1.sas。

```
data exam1;
  do a=1 to 2;
  do b=1 to 3;
  do y=1 to 4;
  input f @@;
  output;
  end; end; end;
  cards;
  12 4 3 0 4 2 3 1
   2 1 3 2 9 5 2 1
   3 4 2 1 1 2 3 2
  ;
run;
```

```
ods html;
proc freq;
  tables a*b*y/scores=rank
  cmh2;
  weight f;
  run;  ods html close;
```

data 步中 a 代表"就诊医院"，a=1 代表"甲医院"，a=2 代表"乙医院"；b 代表"伤后手术间隔时间"，b=1、2、3 分别代表 21～40 d、41～90 d、≥91 d；y 代表"疗效"，y=1、2、3、4 分别代表优良、较好、一般、不佳。

将 freq 过程步中 tables 语句后的 cmh2 选项换成 all，可在最后的结果中输出 3 个统计量，分别为非零相关、行平均得分差异和一般联系。若换成 cmh1，则在最后的结果中只输出第一个统计量，而 cmh2 本身则可输出前两个统计量。

如果结果变量的级别不是等间隔的，则可以选用修正的 Ridit 法计算，将 freq 过程步中 tables 语句后的"scores = rank"换成"scores = modridit"即可。

对例 1，若采用有序变量多重 logistic 回归分析来处理，所用 SAS 程序如下，程序名为 exam1_2.sas。

数据步同程序 exam1_1.sas，不再解释。第一个过程步是进行有序变量多重 logistic 回归分析，结果显示，因素 a 对结果变量的影响无统计学意义(见 3.2 部分有关结果)，故在 model 语句中删去因素 a，重新进行分析，见第二个过程步。Logistic 过程步中的 output 语句可将因素 a 和 b 组合而成的

各种条件下结果变量各种取值发生的概率输出到 sas 数据集中。后面的数据步和过程步用来对包含概率值的 sas 数据集进行操作，以使最后的计算结果便于读者阅读和理解。这几步需要有一定的 SAS 基础，感兴趣的读者可参阅有关 SAS 书籍[1]。

```
data exam1;
  do a=1 to 2;
  do b=1 to 3;
  do y=1 to 4;
  input f @@;
  output;
  end; end; end;
  cards;
  12 4 3 0 4 2 3 1
   2 1 3 2 9 5 2 1
   3 4 2 1 1 2 3 2
  ;
run;
ods html;
proc logistic data=exam1;
  weight f;
  model y=a b;
run;
proc logistic data=exam1;
  weight f;
  model y=b;
  output out=setb p=p;
run;
proc sql;
  create table setc as select
distinct
  b,_LEVEL_,p from setb;
run;
proc sort;
  by b;
run;
proc transpose
  data=setc(rename=
  (_LEVEL_=_name_))
  out=setd(drop=_name_);
  var p;
  by b;
run;
data sete;
  set setd;
  _4=1-_3;
  if b=1 then time="21-40";
  else if b=2 then time="41-90";
  else time=">90";
  drop _label_ b;
run;
proc print noobs label;
  label _1=P(=1);
  label _2=P(<=2);
  label _3=P(<=3);
  label _4=P(=4);
  id time;
run;
  ods html close;
```

表 2 是表 1 资料的数据库形式，二者本质相同。所以分析方法是完全一致的，只是在编写 SAS 程序时，需要对 data 步和过程步加以调整。data 步中的“do…;…;end;”需换成“input id a b y@@”，cards 语句后的数据换成表 2 中的数据即可。在采用 freq 过程或 logistic 过程对资料进行分析时，这两个过程中的“weight f;”语句删除即可，其他与程序 exam1_1.sas 和 exam1_2.sas 相同，不需要调整。

3.2 SAS 程序输出结果 程序 exam1_1.sas 的输出结果如下：

Summary statistics for b by y

Controlling for a Cochran-Mantel-Haenszel statistics (based on rank scores)

Statistics	Alternative hypothesis	*df*	Value	Prob
1	Nonzero correlation	1	11.420 5	0.000 7
2	Row mean scores differ	2	11.641 0	0.003 0

这是对“就诊医院”校正(本质上是分层计算再合并)后的“不同伤后手术间隔时间”疗效之间比较的 CMH 秩和检验的结果，查看“行平均得分差异(row mean scores differ)”可得：$Q_{CMH}=11.641\ 0$，$P=0.003\ 0<0.05$，所以差异有统计学意义，在统计上有理由认为“不同伤后手术间隔时间”对治疗陈旧性踝关节骨折脱位的疗效不相同。当然，由于“不同伤后手术间隔时间”因素也是个多值有序变量，所以本资料也可比较伤后手术间隔时间与疗效之间是否有相关关系。此时，需查看非零相关统计量(nonzero correlation)，其值为 11.420 5，$P=0.000\ 7<0.05$，可认为伤后手术间隔时间与疗效之间存在具有统计学意义的相关关系。

程序 exam1_2.sas 的输出结果如下：

Score test for the proportional odds assumption

Chi-square	*df*	*Pr* >ChiSq
2.1661	4	0.705 2

这是用“score test”做平行线假设的结果，用来考察资料是否满足累积 logistic 回归分析的前提条件。结果为 $P=0.705\ 2>0.05$，说明资料可以做有序变量多重 logistic 回归分析。

Testing global null hypothesis: beta=0

Test	Chi-square	*df*	*Pr* >ChiSq
Likelihood ratio	13.113 0	2	0.001 4
Score	12.317 5	2	0.002 1
Wald	12.302 0	2	0.002 1

这是对模型检验的信息，“score test”对应的 χ^2 值为 12.317 5，$P=0.002\ 1<0.05$，说明模型中考虑这些协变量对结果的影响是有统计学意义的。

Parameter	*df*	Estimate	Standard error	Wald chi-square	*Pr* >ChiSq
Intercept 1	1	1.717 5	0.855 4	4.0311	0.044 7
Intercept 2	1	2.915 4	0.903 9	10.402 9	0.001 3
Intercept 3	1	4.572 2	1.004 6	20.713 6	<0.000 1
a	1	−0.225 5	0.444 5	0.257 2	0.612 0
b	1	−1.010 2	0.290 9	12.062 5	0.000 5

这是模型中各参数的估计值，由于 a 的参数检验结果为 $P=0.612\ 0>0.05$，说明 a 没什么保留的价值，故在程序中删除 model 语句中的变量 a(见第二个过程步)，得结果如下：

Score test for the proportional odds assumption

Chi-square	*df*	*Pr* >ChiSq
0.301 0	2	0.860 3

Testing global null hypothesis: beta=0

Test	Chi-square	*df*	*Pr* >ChiSq
Likelihood ratio	12.856 1	1	0.000 3
Score	11.994 8	1	0.000 5
Wald	12.183 5	1	0.000 5

这部分的结果解释同上，此处从略。

Parameter	*df*	Estimate	Standard error	Wald chi-Square	*Pr* >ChiSq
Intercept 1	1	1.387 6	0.540 9	6.579 7	0.010 3
Intercept 2	1	2.583 5	0.605 9	18.180 2	<0.000 1
Intercept 3	1	4.242 9	0.739 6	32.910 5	<0.000 1
b	1	−1.015 5	0.290 9	12.183 5	0.000 5

这是仅含因素 b 的参数估计结果，有 3 个截距项(Intercept 1、Intercept 2、Intercept 3)，说明已拟合出 3 个 logistic 回归模型，分别用于计算“优良”的概率(P_1)、“优良和较好”的概率(P_2)、“一般以上”的概率(P_3)。其完整的 logistic 回归模型依次为：

$$P_1=\frac{e^{1.387\,6-1.015\,5*b}}{1+e^{1.387\,6-1.015\,5*b}}$$

$$P_2=\frac{e^{2.583\,5-1.015\,5*b}}{1+e^{2.583\,5-1.015\,5*b}}$$

$$P_3=\frac{e^{4.242\,9-1.015\,5*b}}{1+e^{4.242\,9-1.015\,5*b}}$$

Time (d)	*P* (=1)	*P* (≤2)	*P* (≤3)	*P* (=4)
21−40	0.591 97	0.827 51	0.961 85	0.038 15
41−90	0.344 49	0.634 74	0.901 32	0.098 68
≥91	0.159 92	0.386 30	0.767 90	0.232 10

这是程序输出的因素 b 各水平下结果变量各种取值发生的概率大小。*P* (=1)代表“优良”的概率，*P* (≤2)代表“优良＋较好”的概率，*P* (≤3)代表“一般以上”的概率，*P* (=4)代表“不佳”的概率。因为 a 因素两水平间(甲、乙医院)治疗陈旧性踝关节骨折脱位患者疗效的差异无统计学意义，所以上表只给出了不同伤后手术间隔时间下患者疗效的概率估计值。

由各列概率值可知：伤后手术间隔时间越长，疗效越差。

4 小 结

对于结果变量为有序变量的高维列联表资料或以数据库形式显示的此类型资料，可采用 CMH 校正的秩和检验或有序变量多重 logistic 回归分析来处理[2]。其中，CMH 校正的秩和检验是控制外层原因变量或多个原因变量的组合，研究最内层原因变量与结果变量之间的关系。有序变量多重 logistic 回归分析则可以分析多个原因变量对结果变量的影响，且可以计算出各种情形下结果变量各取值发生的概率，所得信息比 CMH 校正的秩和检验丰富，但在应用时须首先检查资料是否满足平行线假设条件。如不满足，则不适合进行有序变量多重 logistic 回归分析。

REFERENCES

1 Hu LP. Applied course of statistical analysis by Windows 6.12 & 8.0. Beijing: Press of Military Medical Sciences. 2001: 335-354. Chinese.
胡良平. Windows SAS 6.12 & 8.0 实用统计分析教程. 北京：军事医学科学出版社. 2001：335-354.

2 Hu LP. Medical statistics: analysis of quantitative and qualitative data using the triple-type theory. Beijing: People's Military Medical Press. 2009: 352-375. 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. 2009：352-375.