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(二)

葛毅[1]，胡良平[2]

1. 后勤指挥学院，北京 100858

2. 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

Ge Y, Hu LP. *J Chin Integr Med*. 2009; 7(12): 1188-1192.
Received November 9, 2009; accepted November 21, 2009; published online December 15, 2009.
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 www.jcimjournal.com.
Forward linking and reference linking via CrossRef.
DOI: 10.3736/jcim20091219

Open Access

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

Yi GE[1], Liang-ping HU[2]

1. Command Academy of Logistics, Beijing 100858, China

2. Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, Beijing 100850, China

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

在上一讲中主要介绍了采用加权等措施合并原因变量的方法处理高维列联表资料[1]，但如果列联表维数较多，且希望将各原因变量对结果变量所产生的影响都明确地显示出来时，加权 χ^2 及 CMH χ^2 分析就不能满足研究的需要了。如果想系统地评价各变量间的联系及变量间相互作用的大小，对数线性模型是达到前述分析目的比较好的分析方法。

对数线性模型最先由 Yule 提出，Bartlett 利用 Yule 1900 年定义的交叉乘积比对三维交互作用进行分析，然后结合方差分析的思想发展而来。其主要的思想为：把各分组变量(包括自变量和因变量)各水平组合下期望(或理论)频数的自然对数表示为各分组变量及其交互作用的线性函数，通过迭代计算求得模型中参数的估计值，进而运用方差分析的思想检验各主效应和交互作用的效应大小。值得注意的是，对数线性模型是在数据满足 possion 分布的情况下使用，而并不是方差分析时数据必须满足正态分布[2]。

对数线性模型主要考察各分类变量间的交互作用(关联性)，对主效应的分析相当于单变量分析。交互效应按其因素多少可分为两变量间交互效应和多变量间交互效应，它们依次又被称为一阶交互效应、二阶交互效应……，依此类推。

1 对数线性模型的计算原理

对数线性模型的主要思想是由方差分析的思想发展而来。先从 2×2 列联表说起，设有两个因素 A、B，各有两个水平，构成了一个 2×2 列联表，其中 A 的第 a 个水平与 B 的第 b 个水平交叉处的频数为 n_{ab}(a=1,2;b=1,2)，显然，n_{ab} 是随样本变化而变化的，不同样本的 n_{ab} 是不同的，且在抽样前无法确切地预测它将取什么值，n_{ab} 是随机变量，造成频数 n_{ab} 变异的原因正是因子的作用及随机误差。若记 n_{ab} 的总体均数为 μ_{ab}，则 μ_{ab} = 常数 × 因素 A 的主效应 × 因素 B 的主效应 × 因素 A 与因素 B 的交互作用的效应(与方差分析模型不同，各因素间是相乘的关系，所以取对数可以成为线性模型)，两边取自然对数得 $\ln(\mu_{ab})$ = ln(常数) + ln(因素 A 的主效应) + ln(因素 B 的主效应) + ln(因素 A 与因素 B 的交互作用的效应)，记 ln(常数)为 $\mu_{..}$，记 ln(因素 A 的主效

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

应)为 α_a,记 ln(因素 B 的主效应)为 β_b,ln(因素 A 与因素 B 的交互作用的主效应)为 $(\alpha\beta)_{ab}$,则 $\ln\mu_{ab}=\mu..+\alpha_a+\beta_b+(\alpha\beta)_{ab}$,这就是二维列联表的对数线性模型。其中 $\mu..$、α_a、β_b、$(\alpha\beta)_{ab}$ 为模型的参数,满足 $\alpha_1+\alpha_2=\beta_1+\beta_2=0$;$(\alpha\beta)_{11}+(\alpha\beta)_{12}=0$;$(\alpha\beta)_{12}+(\alpha\beta)_{22}=0$;$(\alpha\beta)_{11}+(\alpha\beta)_{21}=0$;$(\alpha\beta)_{21}+(\alpha\beta)_{22}=0$。

这 8 个参数 α_1、α_2、β_1、β_2、$(\alpha\beta)_{11}$、$(\alpha\beta)_{12}$、$(\alpha\beta)_{21}$、$(\alpha\beta)_{22}$ 中只有 4 个独立参数 μ、α_1(描述因素 A 的 1 水平作用)、β_1(描述因素 B 的 1 水平作用)和 $(\alpha\beta)_{11}$(描述因素 A 与 B 的交互作用),其他均可由这 4 个参数算得,这几个参数一般可由 SAS 分析结果的参数估计部分给出。$\alpha_2=-\alpha_1$,$\beta_2=-\beta_1$,$(\alpha\beta)_{12}=(\alpha\beta)_{21}=-(\alpha\beta)_{11}$,$(\alpha\beta)_{22}=(\alpha\beta)_{11}$,这种独立参数个数等于列联表格子数的对数线性模型被称为饱和对数线性模型,其中 $\ln\mu_{11}=\mu..+\alpha_1+\beta_1+(\alpha\beta)_{11}$;$\ln\mu_{12}=\mu..+\alpha_1+\beta_2+(\alpha\beta)_{12}$;$\ln\mu_{21}=\mu..+\alpha_2+\beta_1+(\alpha\beta)_{21}$;$\ln\mu_{22}=\mu..+\alpha_2+\beta_2+(\alpha\beta)_{22}$。

研究在 A 的不同水平下,B 各水平的作用是否相同,就是研究 $\frac{\mu_{11}/\mu_{12}}{\mu_{21}/\mu_{22}}=1$ 是否成立。其中模型中的各参数与样本总数密切相关,因此它是一个 possion 模型,而并不是一个多项式模型[2]。

$$\ln\frac{\mu_{11}/\mu_{12}}{\mu_{21}/\mu_{22}}=\ln\mu_{11}+\ln\mu_{22}-\ln\mu_{12}-\ln\mu_{21}=4(\alpha\beta)_{11};\ln\mu_{ab}=\mu..+\alpha_a+\beta_b+(\alpha\beta)_{ab}。$$

研究 $\frac{\mu_{11}/\mu_{12}}{\mu_{21}/\mu_{22}}=1$ 是否成立,在对数线性模型中,就是研究 A 与 B 的交互作用效应的对数,即 $(\alpha\beta)_{ab}$ 是否为零。针对此例的研究目的在对数线性模型中的原假设为 $H_0:(\alpha\beta)_{ab}=0$,$H_1:(\alpha\beta)_{ab}\neq0$;a=1,2;b=1,2。可采用似然比检验来进行假设检验,其检验统计量为:

$$\chi^2_{LR}=2\sum_a\sum_b n_{ab}ln\frac{n_{ab}}{\mu_{ab}}$$

当 H_0 为真时,该统计量渐近服从自由度为 1 的 χ^2 分布。设 $\hat{\mu}_{ab}$ 为理论频数,是在 H_0 成立时,对列联表各网格频数的总体均数 μ_{ab} 的估计值,对于二维列联表,当 $(\alpha\beta)_{ab}=0$ 成立时,$\hat{\mu}_{ab}=\frac{n_{a+}n_{b+}}{n}$。

此外,还应对各参数进行估计,并进行检验假设,其估计结果在 SAS 程序中会自动给出,具体的估计过程,此处就不详述了。如有兴趣,可参阅胡良平编著的《统计学三型理论在定量与定性资料分析中的应用》[3]或其他相关文献。高维列联表的对数线性模型与 2×2 列联表的模型类似,建模思想是完全一样的,只是参数的个数不同。

2 对数线性模型处理结果变量为二值变量的高维列联表资料

2.1 问题的提出 有一项关于冠心病的调查研究,共有 1 330 个病人,按照血压、血清胆固醇、冠心病 3 个变量进行了分类,其中血压有 4 个水平:"<127 mmHg、127～146 mmHg、147～166 mmHg、>167 mmHg",血清胆固醇有 4 个水平"<200 mg/100 mL、200～219 mg/100 mL、220～259 mg/100 mL、>260 mg/100 mL",冠心病是否发生有"是、否"2 个水平。具体数据见表 1[4]。试对此数据进行分析,探讨血压、血清胆固醇和冠心病之间的关系。

表 1 冠心病发生的调查数据

血压	血清胆固醇	冠心病发病与否	人数	
			是	否
<127 mmHg	<200 mg/100 mL		2	117
	200～219 mg/100 mL		3	121
	220～259 mg/100 mL		3	47
	>260 mg/100 mL		4	22
127～146 mmHg	<200 mg/100 mL		3	85
	200～219 mg/100 mL		2	98
	220～259 mg/100 mL		1	43
	>260 mg/100 mL		3	20
147～166 mmHg	<200 mg/100 mL		8	119
	200～219 mg/100 mL		11	209
	220～259 mg/100 mL		6	68
	>260 mg/100 mL		6	43
>167 mmHg	<200 mg/100 mL		7	67
	200～219 mg/100 mL		12	99
	220～259 mg/100 mL		11	16
	>260 mg/100 mL		11	33

2.2 分析及 SAS 程序 表 1 被称为结果变量为二值变量的三维列联表,要探讨血压、血清胆固醇、是否患冠心病之间的关系可以采用对数线性模型进行分析,其 SAS 程序[5-7]如下。

```
DATA ABC1;
DO a=1 TO 4;
DO b=1 TO 4;
DO c=1 TO 2;
INPUT f @@;
OUTPUT;
END; END; END; CARDS;
2   117
3   121
… . …… … ..
11  46
11  33
;
RUN;
```

```
PROC CATMOD;  WEIGHT f;
MODEL a*b*c=_RESPONSE_ /NOGLS
NOPARM NORESPONSE ML;
 LOGLIN a b c a*b a*c b*c a*b*c;   RUN;
 LOGLIN a b c a*b a*c b*c;         RUN;
 LOGLIN a b c a*b a*c;             RUN;
 LOGLIN a b c a*b b*c;             RUN.
 LOGLIN a b c a*c b*c;             RUN;
 LOGLIN a b c a*b;                 RUN;
 LOGLIN a b c a*c.                 RUN;
 LOGLIN a b c b*c;                 RUN;
 LOGLIN a b c;                     RUN;
```

a、b、c 分别代表血压、血清胆固醇、是否患冠心病。为了能深入探讨 3 个因素之间的关系，建立了只含独立因素、含一阶交互作用以及含二阶交互作用3类共 9 个模型，分别对这 9 个模型进行拟合。

在程序中，如果有零频数时，则无法计算，这时需在每个数上加上一个很小的正数，修改的方法是：INPUT w @@；f=w+1E-12；其他语句不变。在过程步中，若需输出对数线性模型中各参数的估计值和误差，只需去掉 MODEL 语句中的 NOPARM 选择项即可。ML 选项的作用是采用最大似然法进行参数估计；NOGLS 选项的作用是不采用加权最小二乘法进行参数估计；NORESPONSE 选项的作用是不输出反应矩阵；LOGLIN 的作用为定义线性模型的效应项，在本例中一共定义了 9 个含有不同效应项的模型。

2.3 结果及解释 上例 SAS 程序的运行结果如下。

Maximum likelihood analysis of variance

Source	*df*	Chi-square	*Pr* >ChiSq
a	3	30.29	<0.000 1
b	3	20.92	0.000 1
c	1	364.87	<0.000 1
a * b	9	5.62	0.777
a * c	3	24.2	<0.000 1
b * c	3	22.28	<0.000 1
a * b * c	9	4.56	0.870 6
Likelihood ratio	0	.	.

Maximum likelihood analysis of variance

Source	*df*	Chi-square	*Pr* >ChiSq
a	3	31.18	<0.000 1
b	3	35.57	<0.000 1
c	1	398.83	<0.000 1
a * b	9	19.92	0.018 4
a * c	3	25.25	<0.000 1
b * c	3	20.87	0.000 1
Likelihood ratio	9	4.77	0.853 5

以上依次是模型 1 和模型 2 的拟合结果。

模型 1 也是饱和模型的输出结果，其误差项的自由度为 0，无法进行拟合优度检验。事实上，因为是饱和模型，模型完全拟合了数据，也就不必检验了，但是由于 a * b * c 项中 $P=0.870\ 6>0.05$，与 0 的差别无统计学意义，因此最好能找到更简单一些的模型。从此模型拟合的结果看，以先去掉二级交互作用项 a * b * c 后再作决定为宜。第 2 至第 9 个对数线性模型可分为 4 组，每组依次比前一组中模型少一项，模型中所缺少的那一项就是假定该项为 0。若某个模型所对应的拟合优度检验（似然比检验）结果为 $P>0.05$，则可以认为此时的模型与资料的吻合度较好，在总体模型中所缺少的那些项的效应为 0。

模型 2 是假定 a * b * c 项为 0，这时最后一行的似然比检验结果表明，用此模型拟合该资料是相当满意的。因为对模型 2 的拟合优度检验（似然比检验）结果为 $P=0.853\ 5$，且进入该模型的所有项的效应不为 0（$P<0.05$）。

Maximum likelihood analysis of variance

Source	*df*	Chi-square	*Pr* >ChiSq
a	3	33.43	<0.000.1
b	3	253.28	<0.000 1
c	1	456.02	<0.000 1
a * b	9	24.92	0.003 1
a * c	3	30.2	<0.000 1
Likelihood ratio	12	24.06	0.02

这是模型 3 的拟合结果。它假定 a * b * c=0，b * c=0，此时，最后一行的似然比检验结果为 $P=0.02<0.05$，表明用此模型拟合是不合适的。因此假设 b * c=0 不成立。

Maximum likelihood analysis of variance

Source	*df*	Chi-square	*Pr* >ChiSq
a	3	53.73	<0.000 1
b	3	33.82	<0.000 1
c	1	494.6	<0.000 1
a * b	9	24.92	0.003 1
b * c	3	26.5	<0.000 1
Likelihood ratio	12	30.4	0.002 4

这是模型 4 的拟合结果。它假定 a * b * c=0，a * c=0，此时，最后一行的似然比检验结果为 $P=0.002\ 4<0.05$，表明此模型的拟合效果是不好的。提示假设 a * c=0 不成立。

Maximum likelihood analysis of variance

Source	*df*	Chi-square	*Pr* >ChiSq
a	3	31.15	<0.000 1
b	3	35.54	<0.000 1
c	1	397.1	<0.000 1
a * c	3	30.2	<0.000 1
b * c	3	26.5	<0.000 1
Likelihood ratio	18	24.4	0.142 3

这是模型 5 的拟合结果。它假定 a * b * c=0，a * b=0，此时，最后一行的似然比检验结果为 $P=0.142\ 3>0.05$，表明用此模型拟合该资料是可以

的，且假设 a * b=0 成立。

Maximum likelihood analysis of variance			
Source	*df*	Chi-square	*Pr* >ChiSq
a	3	53.72	<0.000 1
b	3	253.28	<0.000 1
c	1	579.36	<0.000 1
a * b	9	24.92	0.003 1
Likelihood ratio	15	54.51	<0.000 1

这是模型 6 的拟合结果。它假定 a * b * c=0，b * c=0，a * c=0，此时，最后一行的似然比检验结果为 P <0.000 1，表明用此模型拟合此资料是不合适的，且假设 b * c=0，a * c=0 不成立或不同时成立。

Maximum likelihood analysis of variance			
Source	*df*	Chi-square	*Pr* >ChiSq
a	3	31.14	<0.000 1
b	3	280.83	<0.000 1
c	1	456.04	<0.000 1
a * c	3	30.20	<0.000 1
Likelihood ratio	21	48.51	0.000 6

这是模型 7 的拟合结果。它假定 a * b * c=0，a * b=0，b * c=0，此时，最后一行的似然比检验结果为 P =0.000 6，表明用此模型拟合此资料是不合适的，且假设 b * c=0，a * c=0 不成立或不同时成立。

Maximum likelihood analysis of variance			
Source	*df*	Chi-square	*Pr* >ChiSq
a	3	80.17	<0.000 1
b	3	35.54	<0.000 1
c	1	494.60	<0.000 1
b * c	3	26.50	<0.000 1
Likelihood ratio	21	54.85	<0.000 1

这是模型 8 的拟合结果。它假定 a * b * c=0，a * b=0，a * c=0，此时，最后一行的似然比检验结果为 P <0.000 1，表明此模型是不合适的，且假设 a * b=0，a * c=0 不成立或不同时成立。

Maximum likelihood analysis of variance			
Source	*df*	Chi-square	*Pr* >ChiSq
a	3	80.17	<0.000 1
b	3	280.83	<0.000 1
c	1	579.36	<0.000 1
Likelihood ratio	24	78.96	<0.000 1

这是模型 9 的拟合结果。它假定 a * b * c=0，a * b=0，a * c=0，b * c=0，此时，最后一行的似然比检验结果为 P <0.000 1，表明用此模型拟合此资料是不合适的，且假设 a * b=0，b * c=0，a * c=0 不成立或不同时成立。

从以上的结果可以看出用模型 1、模型 2 和模型 5 进行拟合都是可以的，但是模型 2 与模型 5 相对简单，且对模型的似然比检验结果都为 P > 0.05。问题是模型 2 与模型 5 选哪一个比较合适是一个关键。模型 2 比模型 5 多了一个 a * b 项，但是其似然比检验的结果 P =0.853 5 又很理想。根据 Goodman 和 Fienberg 所提出的方法，即将两个模型的似然比 χ^2 值相减来评价在模型中增加参数后所引起的拟合优度的改变。对于此例，模型 5 与模型 2 的 χ^2 值相减后，结果为 24.4－4.77＝19.63，自由度相差为 18－9＝9，与自由度为 9 的 χ^2 值 16.92 相比较，差异有统计学意义（P <0.05）。说明在模型中增加了 a * b 项后，模型的拟合效果有了很大的改善，因此，最终应选择模型 2。其专业结论为：血压（变量 a）与血清胆固醇（变量 b）有关，血压（变量 a）与是否患冠心病（变量 c）有关，血清胆固醇（变量 b）与是否患冠心病（变量 c）有关。具体地说，血压偏高的人血清胆固醇的含量有偏高的倾向；血压和（或）血清胆固醇偏高的人，患冠心病的概率会有升高的趋势。

3 对数线性模型处理结果变量为多值名义变量的高维列联表资料

有一项关于老年人信仰的调查，主要目的是考察不同性别的老年人的信仰状况，具体数据见表 2[8]。

表 2 不同性别老年人的信仰状况

种族	性别	人数		
		信仰状况：	有信仰	不确定	没有信仰
白种人	女		371	49	74
	男		250	45	71
黑人	女		64	9	15
	男		25	5	13

在该研究中原因变量为“种族”和“性别”，其中“种族”和“性别”为二值名义变量；结果变量为“信仰状况”，有 3 个水平，分别为“有信仰”、“不确定”、“没有信仰”，属于多值名义变量。对以上数据的分析可以采用 CMH χ^2 检验，可以平衡掉“种族”这个分层因素。但是为了能够更深入地探讨这 3 个因素之间的关系，在此考虑用对数线性模型进行分析。其 SAS 程序[1-6] 如下。

程序	程序
DATA ABC2;	PROC CATMOD; WEIGHT f;
DO a=1 TO 2;	MODEL a*b*c=_RESPONSE_ /NOGLS
DO b=1 TO 2;	NOPARM NORESPONSE ML;
DO c=1 TO 3;	LOGLIN a b c a*b a*c b*c a*b*c, RUN,
INPUT f @@,	LOGLIN a b c a*b a*c b*c; RUN;
OUTPUT,	LOGLIN a b c a*b a*c; RUN,
END; END; END; CARDS,	LOGLIN a b c a*b b*c; RUN;
371 49 74	LOGLIN a b c a*c b*c, RUN,
250 45 71	LOGLIN a b c a*b; RUN;
64 9 15	LOGLIN a b c a*c; RUN,
25 5 13	LOGLIN a b c b*c, RUN;
;	LOGLIN a b c; RUN;

在程序中 a、b、c 分别代表“种族”、“性别”、“信仰状况”。在运行的结果中，考察所有可能的不饱和模型，依据拟合优度检验的结果，在模型成立的基础上，要求模型尽量简单。在进行综合比较后，选出最佳的模型为：

Maximum likelihood analysis of variance

Source	*df*	Chi-square	*Pr* >ChiSq
a	1	379.69	<0.000 1
b	1	10.88	0.001 0
c	2	501.40	<0.000 1
a * b	1	4.40	0.035 9
b * c	2	6.77	0.033 9
Likelihood ratio	4	2.85	0.583 6

假设 a * b * c=0 和 a * c=0 成立。由于 a * b 项的 $P<0.05$，说明种族（变量 a）与性别（变量 b）有关，两种种族的受试者其性别构成比不同；b * c 项的 $P<0.05$，说明性别（变量 b）与信仰状况（变量 c）有关。这说明不同性别的老年人在信仰问题上的态度是不同的，而种族与信仰状况则无太大关系。

4 讨 论

对数线性模型是分析高维列联表的一种非常有效的方法，它可以通过分析交互效应项来挖掘多个变量间的深层关系。在应用时，对数线性模型对样本含量要求比较严格，一般要求样本例数应为网格数的 4～5 倍，而且要有 80% 以上格子的理论频数大于 5。与之相比，logistic 回归模型似乎对样本含量要求不那么严。然而许多实际工作者在应用对数线性模型进行分析时，往往忽视了样本含量的要求，因此，为了使结果可靠，必须对样本含量有所要求，观察频数 $m_i \leqslant 5$ 的比例不能太多，对于 $m_i \leqslant 5$ 的比例较多的稀疏数据的情形，其参数估计值将无解释的价值，而且模型拟合优度检验统计量的渐近性也差，因此得出的结论也就缺乏科学性。

从理论上讲，对数线性模型和 logistic 回归模型密切相关，其分析的结果也基本上是一致的，但是 logistic 回归模型有明确的因变量，可用于考察各自变量对因变量的作用大小。因此，如何选择这两种方法，主要还是应根据分析目的，一般来讲，对数线性模型适合于进行探索性分析，而后者多用于流行病学等方面危险因素的筛选。

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 Agresti A. Categorical data analysis. New York: John Wiley & Sons, Inc. 2002: 318.

3 Hu LP. Application of triple-type theory of statistics in statistical expression and description. Beijing: People's Military Medical Press. 2008. Chinese.
胡良平. 统计学三型理论在统计表达与描述中的应用. 北京：人民军医出版社. 2008.

4 Everitt BS. Contingency table analysis. Beijing: Science Press. 1980: 95. Chinese. Translated by Liu YY, Zhou JL.
Everitt BS 著，刘韵源，周家丽译. 列联表分析. 北京：科学出版社. 1980: 95.

5 Hu LP. Modern statistics and application of SAS. Beijing: Press of Military Medical Sciences. 2000: 191-214. Chinese.
胡良平. 现代统计学与 SAS 应用. 北京：军事医学科学出版社. 2000: 191-214.

6 Hu LP. Applied course of statistical analysis by Windows SAS 6.12 & 8.0. Beijing: Press of Military Medical Sciences. 2001: 327-329. 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: 327-329.

7 SAS Institute Inc. SAS/Stat 9.1 user's guide. Cary, NC: SAS Institute Inc. 2004.

8 Agresti A. An introduction to categorical data analysis. New York: John Wiley & Sons, Inc. 2006: 114-120.