·基础理论与方法·

精确 logistic 回归及其 SAS 应用程序

刘启军 曾庆 周燕荣

【摘要】 目的 精确 logistic 回归在最大似然法估计结果不可靠或者不存在情况下，作为传统 logistic 回归的一种补充方法可适用于小样本资料、资料结构不平衡及高度分层资料。方法 对 37 例乳腺癌患者术后预后资料采用 SAS 进行精确 logistic 回归分析。结果 发现组织学分级、淋巴结转移两因素对乳腺癌的术后预后具有统计学意义。结论 在样本量很小，最大似然法估计无效的情况下，精确 logistic 回归是一种极好的分析手段。

【关键词】 logistic 回归，精确；SAS；小样本

Exact logistic regression and its performance to SAS system *LIU Qi-jun, ZENG Qing, ZHOU Yan-rong. Department of Health Statistics, Chongqing University of Medical Sciences, Chongqing* 400016, China

【**Abstract**】 **Objective** To explore the feasibility of exact logistic regression, used as a complemental method for the maximum liklihood estimation, and to analyse with data small sample, unbalanced structure and highly stratal nature under the situations of questionable results or inexistence of the maximum likelihood estimation. **Methods** Data from 37 postoperative breast cancaer cases were analyzed in 1997 by exact logistic regression under SAS system. **Results** Data caculated by SAS software showed that Quasi-complete separation of data points was detected but the results of maximum likelihood estimation did not exist, SAS outputs conflicted the results of the last maximum likelihood iteration (likelihood Chi-square and score Chi-square have similar P, less than 0.05, but the Wald chi-square had a larger P, more than 0.05). Under conditional exact parameter estimation, it appeared that: (1) the joint effect of conditional score statistics was 21.12 with $P = 0.000\ 3$; (2) for individual parameters, the effect conditional score statistics of histological classification (grades) was 5.80 with $P = 0.020\ 8$; axillary node metastasis (diversion) was 5.74 with $P = 0.019\ 5$; tumor size (size) was 0.79, with $P = 0.647\ 6$. The effects of tumor histological classification and axillary node metastasis were statistically significant on breast cancaer tumour. **Conclusion** Exact logistic regression seemed to be a very useful method in analyzing data from small sample when the maximum likelihood estimation was either with no effect or did not exist.

【**Key words**】 Logistic regression, exact; SAS; Small sample

非条件 logistic 回归模型参数估计的传统方法是最大似然法，反应变量为二分类及样本量较大是该方法的基本条件。在流行病学研究中经常遇到小样本、资料结构不平衡及高度分层（含层内重复测量）的情况，最大似然法估计结果不可靠甚至不存在。Cox，Snell[1] 于 1970 年提出了精确 logistic 回归以处理此类资料，但是由于手工计算的繁杂，该方法没有得到广泛应用。随着精确条件分布（exact conditional distribution）方面研究的深入[2,3]，尤其是递归运算和多元变换运算（multivariate shift algorithm）的提出[4,5] 使精确估计趋于简便高效，更由于计算机技术的发展促进了精确 logistic 回归的推广应用。

作者单位：400016 重庆医科大学卫生统计学教研室

基本原理

一、基本思想

精确 logistic 回归推断是研究在回归参数相应的充分统计量（sufficient statistics）的精确条件分布（或称排列分布）的基础上建立条件似然函数[5]，并进一步求出充分统计量的分布函数而得。

二、模型

对每一个观测建立非条件 logistic 模型为[5]

$$\log it\pi_j = \ln\left[\frac{\pi_j}{1-\pi_j}\right] = \gamma + x_j'\beta \qquad (1)$$

其中 $j = 1, 2, \cdots, n$ 表示观测数，π_j 表示 $Y_j = 1$ 时的概率。式(1)表示对每一个观测建立一个概率函数。在二项分布概率的基础上得到非条件似然函数[5]，即一系列的观测结果出现的概率为

$$\Pr(Y_1 = y_1, Y_2 = y_2, \Lambda Y_n = y_n) =$$

$$\frac{\exp[\sum_{y=1}^{n} y_j(x_j'\beta+\gamma)]}{\prod_{j=1}^{n}[1+\exp(x_j'\beta+\gamma)]} \quad (2)$$

我们只对模型中的参数 β 感兴趣，在模型剩余参数(nuisance parameter)的充分统计量的实测值 m ($m=\sum_{j=1}^{n} y_j$)的条件下得到其条件似然函数为

$$\Pr(Y_1=y_1, Y_2=y_2, \Lambda\ Y_n=y_n \mid m) = \frac{\exp(\sum_{y=1}^{n} y_j x_j'\beta)}{\sum_R[\exp(\sum_{j=1}^{n} y_j x_j'\beta)]} \quad (3)$$

其中分母中包含了集合 R 的全部元素的各种组合，即对所有 n 维 y 向量的求和

$$R=\{(y_1, y_2, \Lambda\ y_n): \sum_{j=1}^{n} y_j = m\}$$

对于此条件似然函数的参数估计有两种方法：其一为常用于匹配资料分析的最大化条件似然函数的渐近方法，其二为本文介绍的精确法。对于后者，从式(3)可得参数 β 的相应$P\times 1$ 为充分统计量向量的实测值

$$t=\sum_{j=1}^{n} y_j x_j$$

T 分布的非条件似然函数为

$$\Pr(T_1=t_1, T_2=t_2, \Lambda\ T_P=t_P) = \frac{c(t)\,e^{t'\beta}}{\sum_u c(u)\,e^{u'\beta}} \cdots \text{K K} \quad (4)$$

其中 $c(t)=|S(t)|$，$S(t)=\{(y_1, y_2, \cdots, y_n)$ $\sum_{j=1}^{n} y_j = m, \sum_{j=1}^{n} y_j x_{ij} = t_i, i=1,2,\cdots,p\}$。$|S|$表示集合 S 的元素的各种不同组合数；分母是所有使 $c(u)\geqslant 1$的 u 向量空间的求和。参数 β 的精确推断就需要先计算出系数，如 $c(t)$。

三、精确条件推断

1.单个参数的统计推断：

(1)条件似然函数：引入单个参数的充分统计量(T_p)的条件似然函数[5]，给定 t_1、t_2、…、t_{p-1}的情况下 t_p 仅依赖于β_p，以 $f(t_p \mid \beta_p)$表示 $\Pr(T_p=t_p \mid T_1=t_1, \cdots, T_{p-1}=t_{p-1})$，由式(4)

$$f(t_p \mid \beta_p) = \frac{c(t_1, t_2, \Lambda, t_p)\, e^{\beta_p t_p}}{\sum_u c(t_1, t_2, \Lambda, t_{p-1}, u)\, e^{\beta_p u}} \quad (5)$$

分母部分是所有使 $c(t_1、t_2、\cdots、t_{p-1}, u)\geqslant 1$ 的 u 值下的求和。由于式(6)未包含 $p-1$ 个剩余参数(β_1, $\beta_2, \cdots, \beta_{p-1}$)，所以应用该式估计感兴趣的参数 β_p 十分方便。

(2)假设检验：令无效假设 H_0：$\beta_p=0$。要求得 H_0 是否成立的双侧精确概率，可在一判别区域(临界区域) E 内对式(5)求和，即得

$$P=\sum_{v\in E} f(v \mid \beta_p=0) \quad (6)$$

对式(6)有两种检验方法，即条件概率(conditional probability)检验和条件分(conditional score)检验，不同的方法采用不同的判别区域 E。

①适用于条件概率检验的 E[5]，即

$$E_{cp}=\{v: f(v \mid \beta_p=0) \leqslant f(t_p \mid \beta_p=0)\}$$

②适用于条件分检验的 E[5]，即

$$E_{cs}=\{v: (v-\mu_p)^2\sigma_p^{-2} \geqslant (t_p-\mu_p)^2\sigma_p^{-2}\}$$

其中 μ_p、σ_p^{-2} 是 T_p 在 $\beta_p=0$ 条件下的均数和方差阵。

上面两种检验均是在 t_1、t_2、…、t_{p-1}固定，v 可在 T_p 的所有范围内任意取值的情况下，求得系数 $c(t_1, t_2, \cdots, t_{p-1}, v)$，从而计算得所需的精确概率。

(3)参数估计：

①区间估计：为了求得 α 水平的参数 β_p 的可信区间(β_-, β_+)，定义

$$F_1(t_p \mid \beta)=\sum_{v\geqslant t_p} f(v \mid \beta) \cdots \text{和} \cdots F_2(t_p \mid \beta)=\sum_{v\leqslant t_p} f(v \mid \beta)$$

设 t_{min}、t_{max}为式(5)中 t_p 的最小和最大可能值，有

下限 β_- 为：$F_1(t_p \mid \beta_-)=\alpha/2$ $(t_{min}<t_p\leqslant t_{max})$

$\beta_-=-\infty$ $(t_p=t_{min})$

上限 β_+ 为：$F_2(t_p \mid \beta_+)=\alpha/2$ $(t_{min}\leqslant t_p<t_{max})$

$\beta_+=\infty$ $(t_p=t_{max})$

由上可得参数 β_p 的 $100(1-\alpha)\%$ CI。

②点估计：有两种方法估计，即条件最大似然估计(β_{cmle})和中值无偏估计(β_{ume})[6]。

β_{cmle}通过选择 β 的不同值以最大化 $f(t_p \mid \beta)$，求得。当 t_p 取极值时 β_{cmle}不确定，此时应选用 β_{ume}

$$\beta_{ume}=(\beta_-+\beta_+)/2$$

与条件最大似然估计不同的是，中值无偏估计总可以得到明确的点估计值。

2.多个参数的统计推断：对于多个参数的假设检验和参数估计类似于单个参数。

目前可实现精确 logistic 回归计算的软件有 SAS 8.0及以上的版本[6]和 LogXact[7]。由于 SAS 是目前应用比较普遍的统计软件，故本文采用 SAS 计算。

实例分析

采用重庆医科大学附属第一医院肿瘤中心 1997 年 37 例女性乳腺癌患者术后 5 年生存状态及其影响因素的资料。变量赋值情况：size(肿瘤大小)：< 2 cm 为 0，2 ~ 5 cm为 1，> 5 cm为 2；grades(肿瘤组织学分级)：Ⅰ级为 0，Ⅱ级为 1，Ⅲ级为 2；diver(淋巴结转移)：无转移为 0；1 ~ 3 个为 1；4 ~ 9 个为 2；> 10 个为 3；y5scun(生存状态)满 5 年赋值为 0，否则为 1。进行 logistic 回归分析，SAS 程序为：

```
data abc;
```

```
input size grades diver y5scun@@;
cards;
……
;;;;
proc logistic;
model y5scun = size grades diver;
exact size grades diver
      / joint estimate = both outdist = dist;
proc print data = dist;
run;
```

以上程序同时进行似然法和精确法的计算。其中 exact 是 SAS 8.0及以上版本新增的功能，完成精确法计算，其位置必须在 model 后，其后的变量表表示要进行精确推论的变量；joint（联合检验）选项要求对所有变量的偏回归系数是否同时等于 0 作精确检验；estimate 选项要求作区间估计和点估计，等于 both 时要求产生参数估计和比数比结果，等于 odds 时要求比数比结果，等于 parm 时要求参数估计；outdist 选项产生包含变量表的充分统计量的所有可能的精确条件分布的数据集，当上面指定 joint 后，该数据集将包含充分统计量的各种可能排列；同时 exact 声明还可指定 onesided（要求单侧可信区间和单侧的单个参数和比数比的 P 值）、jointonly（强制只产生 joint 的结果，不产生单个变量的假设检验结果）及检验水准 alpha 选项。

经 SAS 计算，给出最大似然估计不存在的提示信息，表明最大似然估计的结果不可靠。最后一次最大似然迭代的结果如表 1。

表1 $\beta=0$ 是否成立的假设检验结果（$\alpha=0.05$）

统计量	χ^2 值	ν	P 值
似然比	23.533 8	3	< 0.000 1
Score	15.304 9	3	0.001 6
Wald	0.959 5	3	0.811 1

从表 1 可知上述 3 个统计量的显著性检验结果不统一，无从判断该模型的有效性。

我们注意到该资料在 3 个变量上均线性可分，即无淋巴结转移、肿瘤≤2 cm、组织学分级为Ⅰ的病例均至少可生存 5 年，这是导致最大似然估计失败的原因。

精确 logistic 回归估计各因素相应的偏回归系数的充分统计量向量 T 的实测值为 $(t_0, t_1, t_2, t_3) = (30, 56, 24, 40)$。

同时产生相应的充分统计量的精确条件分布表，通过该表可直接计算单侧 P 值。联合检验结果表明三因素的效应具有统计学意义。肿瘤组织学分级及淋巴结转移对乳腺癌术后预后的影响按照检验水准（α）具有统计学意义[8]，肿瘤大小不具有统计学意义（表 2～4）。

表2 条件精确检验结果（$\alpha=0.05$）

因 素	检验	统计量	Exact P 值	Mid P 值
joint（联合检验）	条件分	14.891 3	0.000 3	0.000 3
	条件概率	2.914E-6	0.000 4	0.000 4
size（肿瘤大小）	条件分	0.788 2	0.647 6	0.504 8
	条件概率	0.285 7	0.647 6	0.504 8
grades（组织学分级）	条件分	5.795 9	0.020 8	0.013 4
	条件概率	0.014 9	0.020 8	0.013 4
diver（淋巴结转移）	条件分	5.737 3	0.019 5	0.014 9
	条件概率	0.009 12	0.019 5	0.014 9

注：Mid P 值为对离散分布作调整后的 P 值

表3 精确参数估计（$\alpha=0.05$）

因 素	点估计	95% *CI*	P 值
size（肿瘤大小）	−0.985 5	−4.993 7～1.517 9	0.723 8
grades（组织学分级）	−2.020 6*	−∞～−0.167 4	0.029 8
diver（淋巴结转移）	−1.712 4*	−∞～−0.206 6	0.018 2

* 中位无偏估计

表4 精确 *OR* 值估计（$\alpha=0.05$）

因 素	*OR* 估计值	95% *CI*	P 值
size（肿瘤大小）	0.373	0.007～4.563	0.723 8
grades（组织学分级）	0.133*	0.000～0.846	0.029 8
diver（淋巴结转移）	0.180*	0.000～0.813	0.018 2

* 同表 3

讨 论

1.国外学者将精确 logistic 回归大量应用于小样本量的药物剂量-反应分析，作为处理小样本（或因分层而致各层例数相对较少）、最大似然法结果不可靠或者不存在情况的补充方法。近年来精确 logistic 回归方法广泛应用于病因筛选研究中，相同总体的大样本资料的似然法估计和小样本资料的精确 logistic 回归估计具有相同的功效[5]。

2.由于分层和不分层资料回归参数的充分统计量的分布函数具有相同的形式，且 SAS 计算程序一样，故本文仅讨论了不分层模型的参数估计。同单因素分析中 R × C 表的 Fisher 精确法相比，精确 logistic 回归不但能进行单因素分析，还可进行多因素的联合分析[5]。

3.当非条件渐近估计和条件精确法结果冲突时有学者建议在小样本、似然法 $P<0.10$ 时精确法结果更恰当[6]。

4.不足之处在于在样本量较大、因素个数较多

的情况下精确 logistic 估计需要耗费大量的计算机内存,可能出现耗时过多甚至出现计算机死机的情况,因此在条件满足似然法时,建议采用最大似然法。

5.研究进展:目前研究主要集中在稀有事件分析中,当最大似然法与精确法结果近似时的结果选择 鞍点(saddlepoint)研究和马尔可夫链 Monte Carlo 抽样研究。

对于满足 logistic 回归建模的离散型资料分析时建议先采用最大似然法估计回归参数,在最大似然法估计结果不可靠的情况下,再采用精确 logistic 回归估计。本文旨在将精确 logistic 回归方法介绍给国内致力于病因学研究的同行,以期多一种病因学研究的方法。相信随着精确 logistic 回归的推广应用,它将在病因学研究中发挥及其重要的作用。

参 考 文 献

1 Cox DR, Snell EJ. Analysis of binary data. 2nd ed. Chapman and Hall. London, 1989.
2 Tritchler D. An algorithm for exact logistic regression. JASA, 1984, 79: 709-711.
3 Gail MH, Lubin JH, Rubenstein LV. Liklihood caculations for matched case-control studies and survival studies with tied death times. Biometrics, 1981, 68: 703-707.
4 Hirji KF, Mehta CR, Patel NR. Computing distributions for exact logistic regression. JASA, 1987, 82: 1110-1117.
5 Mehta CR, Patel NR. Exact logistic regression: theory and examples. Statistics Medicine, 1995, 14: 2143-2160.
6 Stokes ME, Davis. Categorical data analysis using the SAS system, cary, NC: SAS institute, 2000.
7 LogXact. Software for exact logistic regression, cytel software corporation, Cambridge, MA, 1992.
8 余乃登,贾美琳,卢文娜,等.多元分析在乳腺癌预后预测中的应用.贵阳医学院学报,1995,20:309-310.

(收稿日期:2002-09-09)
(本文编辑:张林东)

·疾病控制·

河南省方城县独树镇人间布鲁氏菌病流行情况调查

郭恩朝 金玉玲 安玉凤 李金刚 陈曦

1.材料与方法:布鲁氏菌试管凝集抗原和布氏菌素为中国疾病预防控制中心传染病预防控制所生产(批号 2000.1 和 2002.1)。调查采用局部普查和线索调查两种方法。①普查:已经确诊 4 例布鲁氏菌病(布病)患者的前庄村进行全民普查,首先对 1 岁以上人群做皮变试验,在前臂内侧皮内注射布氏菌素0.1 ml,48 h观察结果,2.5 cm×2.5 cm为阳性,然后对皮变试验阳性者采集静脉血2 ml,离心取血清,进行试管凝集试验。②线索调查:对该疫区所属的马库庄、金银店、中信庄、小街等行政村由调查人员入户访问,填写调查表,根据接触史、既往史、临床症状确定疑似患者,并采集疑似患者静脉血2 ml。③试管凝集试验:凝似患者血清1:100^{++}以上者即确诊为布病,对1:100^{+}和1:50^{++}者定为疑似患者,经访问体检,症状典型者可确诊为布病患者。

2.结果:在对前庄村普查中,共进行皮变试验 851 人,皮变阳性 56 例,阳性率为6.58%。对该 5 个村的调查结果见表 1。经对确诊的布病患者访问、体检发现,大多数患者临床表现轻微,症状不典型,体征不明显,主要表现为长期低热,乏力、纳差、身痛或关节疼痛等,但随着病程延长,症状逐渐加重。

表1 方城县独树镇布病疫区患病率

村 别	人数	采血份数	1:100^{++}以上例数	患病率(/10 万)
前庄村	1 650	56	33	2 000.00
马库庄	1 430	48	20	1 398.60
中信庄	1 245	43	2	160.64
金银店	2 215	11	1	45.14
小 街	1 395	20	0	0.00
合 计	7 935	178	56	705.57

3.讨论:调查结果显示,前庄村人群感染率为6.58%,且发病年龄差异有显著性。发病年龄集中在 10 ~ 59 岁之间,以男性青壮年为主,这与青壮年男性从事放牧、宰杀及接生羊羔机会较多有关,也符合我国人间布病流行特点。患者病程多集中在 1 ~ 6 年,表明经过 1987 年普查普治后,布病疫情虽得到暂时控制,但并未终止,近年来逐年上升,现已显现出流行态势。畜间传染源的存在和蔓延是该病流行的根本原因。由于该地母羊流产现象普遍存在,且畜牧部门近年来没有对畜间布病采取任何防治措施,导致畜间布病流行,而人间免疫措施不力是该病流行的重要因素。因防治经费不足导致近 10 年来该疫区未采取任何免疫措施,人群易感性高,一旦细菌侵入极易发病。

作者单位:473200 河南省方城县卫生防疫站(郭恩朝、金玉玲、安玉凤、陈曦);方城县独树镇卫生院(李金刚)

(收稿日期:2002-10-28)
(本文编辑:尹廉)