·基础研究·

主成分改进的 Logistic 回归方法探讨及其应用

郑 伟[1], 高 歌[1], 魏以璧[2]
(1.苏州大学放射医学与公共卫生学院 流行病与卫生统计学教研室, 江苏苏州 215123;
2.徐州市中区医院 信息科, 江苏徐州 221009)

摘要:**目的** 探讨使用主成分分析处理 Logistic 回归中共线性问题的方法及其在医学科研中的应用。**方法** 采用多重线性回归中的共线性诊断方法诊断 Logistic 回归模型的共线性, 使用主成分分析处理 Logistic 回归中的共线性问题, 全部计算采用 SAS 软件。**结果** 在 216 例高血压脑出血患者的预后影响因素分析中, 使用主成分改进的 Logistic 回归, 各估计系数的标准误均有所减小, 提示模型结构较为稳定, 其结果的可靠性更高。**结论** 使用主成分改进的 Logistic 回归进行多重共线性的诊断和处理是有效及可行的。

关键词:多重共线性; Logistic 回归; 高血压; 脑出血; SAS
中图分类号: R195.4 **文献标识码**: A **文章编号**: 1673-0399(2008)04-0517-04

The Modified Logistic Regression Model and Its Application

ZHENG Wei[1], GAO Ge[1], WEI Yi-bi[2]
(1.Dept of Epidemiology and Health Statistics, Radiation Medicine and Public Health School of Suzhou University, Jiangsu Suzhou 215123,China; 2.Dept of Information Section,Xuzhou Center Hospital Jiangsu Xuzhou 221009, China)

Abstract:Objective To explore the method that principal component analysis was used to solve the multicollinearity in the logistic regression analysis and its application in the medical reseach. Methods The data with multivariable multicollinearity were diagnosed by the method which was used to diagnose multivariable multicollinearity in the multiple linear regression model, treated using principal component analysis.All calculation was completed by SAS. Results Using the logistic regression was improved by principal component analysis to analyze the prognosis of 216 hypertensive intracerebral hemorrhage patients and all estimate standard error had decreased . It meant that the construction of the model was steady and its result was more reliable. Conclusion The new method is effective and feasible for diagnosis and treatment of multivariable multicollinearity in the logistic regression model analysis.

Key words: multicollinearity; Logistic regression; hypertension; cerebral hemorrhage; SAS

Logistic 回归是现今进行病因分析、生存分析常用的多元统计方法。但在 Logistic 回归中的变量筛选及参数估计, 都要求各自变量之间相互独立, 而在很多研究中各自变量间并不独立, 而是相互之间存在一定程度的线性依存关系, 被称为多重共线性(multico linearity), 这种多重共线性关系常会增大估计参数的均方误差和标准误, 有的甚至使回归系数的方向相反, 导致方程极不稳定, 从而引起 Logistic 回归模型拟合上的矛盾及不合理[1]。本研究采用主成分改进的 Logistic 回归分析, 先采用主成分分析产生若干主成分, 它们必定会将相关性较强的变量综合在同一个主成分中, 而不同的主成分又是互相独

基金项目:国家自然科学基金资助项目(Q30571620) 收稿日期:2008-03-17
作者简介:郑 伟(1984-),女,江苏徐州人,在读医学硕士,研究方向为敏感问题的抽样调查。通讯作者:高 歌

立的。只要多保留几个主成分，原变量的信息不致过多损失。然后，以这些主成分为自变量进行 Logistic 回归就不会再出现共线性的困扰。如果原有 p 个自变量 $X_1, X_2, \cdots, Xp$，那么，采用全部 p 个主成分所作回归完全等价于直接对原变量的回归；采用一部分主成分所作回归虽不完全等价于对原变量的回归，但往往能摆脱某些虚假信息，而出现较合理的结 果[2]。

1 原理与方法

1.1 多重共线性的诊断

由于 SAS 软件的 Logistic 过程不提供共线性诊断，本研究采用 SAS 软件中的多重线性回归的共线性诊断方法来进行诊断。利用 SAS 软件的 REG 过程步的 collion 选项，计算条件指数。由于只关心自变量之间的关系，因此条件指数的估计与应变量的函数形式无关[3]。

1.2 主成分改进的 Logistic 回归分析的原理与方法[4]

首先将原设计矩阵的各列解释变量观察值矩阵标准化，标准化后的矩阵用 X 表示，然后进行主成分变换，选择使得前 r 个特征根之和在 p 个特征根总和中所占比例>80%，根据选定的 r 将矩阵 X'X 的特征向量构成的正交阵 ϕ 剖分为 $\phi=\phi_r+\phi_*$，$Z_j=X\phi_r$ 为前 r 个主成分的得分值。具体公式如下：

$$Z_j=\sum_i l_{ij}X_i \qquad j=1,2,\cdots,r \qquad (1)$$

由于各主成分 $Z_1, Z_2, \cdots, Z_r$ 间的相关系数为 0，则可以 $Z_1, Z_2, \cdots, Z_r$ 为自变量，用通常的 Logistic 回归模型估计方法，得到 $Z_1, Z_2, \cdots, Zr$ 的回归系数估计 $a_1, a_2, \cdots, a_r$ 拟合的模型为：

$$P(Y=1 \mid Z)=\frac{e^{a_0}+\sum_{i=1}^{r}a_iZ_i}{1+e^{a_0}+\sum_{i=1}^{r}a_iZ_i} \qquad (2)$$

将(1)式代入(2)中，得到原变量回归系数 β 的估计 b_i 拟合的模型为：

$$P(Y=1 \mid X)=\frac{e^{b_0}+\sum_{i=1}^{p}b_iX_i}{1+e^{b_0}+\sum_{i=1}^{p}b_iX_i} \qquad (3)$$

其中 $b_i=l_i^Ta=l_{i1}a_1+l_{i2}a_2+\cdots+l_{ir}a_r$，$i=1,2,\cdots,p$ (4)

获得了原变量回归系数 β 的估计值后，还须对之进行 Wald 检验。由 SAS 统计软件可得 $a_1, a_2, \cdots, a_r$ 的协方差矩阵：

$$\sum=\begin{bmatrix} var(a_1) & cov(a_1,a_2) & \cdots & cov(a_1,a_r) \\ & & & \\ cov(a_r,a_1) & cov(a_r,a_2) & \cdots & var(a_r) \end{bmatrix}$$

$$=\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1r} \\ & & & \\ \sigma_{r1} & \sigma_{r2} & \cdots & \sigma_{rr} \end{bmatrix} \qquad (5)$$

由极大似然估计的性质，$a_r=(a_1,a_2,\cdots,a_r)$ 渐近服从正态分布 $N[a_i, cov(a_i)]$。因为正态分布具有线性变换的不变性，由 $b_i=l_i^Ta=l_{i1}a_1+l_{i2}a_2+\cdots+l_{ir}a_r$，故 b_i 渐近服从正态分布。只要得到 b_i 的方差，就可以求得 $\beta=0$ 的 Wald 检验统计量。而 bi 的方差可由(4)，(5)式获得：

$$var(b_i)=\sum_j\sum_k l_{ij}l_{jk}\sigma_{jk} \qquad (6)$$

因此，$\beta=0$ 的 Wald 检验统计量为：

$$u_i=b_i/\sqrt{var(b_i)} \qquad (7)$$

P 值运用 SAS 中 probnorm 函数，P=1- probnorm [abs(u)]计算得出。

Logistic 回归系数和流行病学中常用的反映各危险因素对疾病作用大小的优势比(比数比)有直接的联系，并把单因素两水平下的优势比的定义扩展到多因素任意取值下都可以定义。当只有一个自变量 X_j 由 C_0 改变为 C_1 而固定其他自变量取值时，有

$$OR_j=\exp[\beta_j(C_1-C_0)] \qquad (8)$$

以上计算由 SAS 统计软件计算完成。

2 应用实例

2.1 资料来源

2.1.1 调查时间 收集徐州市中心医院 2006 年 1 月 1 日至 2006 年 12 月 30 日间 216 例高血压脑出血住院患者资料，采用 Epi Data 软件进行资料的录入。

2.1.2 调查内容 (1)预后影响因素：X_1(性别：男=1、女=0)，X_2(年龄：岁)，X_3(血糖：mmol/L)，X_4(白细胞计数：$\times10^9$ 个/L)，X_5(体温：摄氏度)，X_6(收缩压：mmHg)，X_7(舒张压：mmHg)，X_8(平均动脉压：mmHg)，X_9(Glassgow 昏迷指数)，X_{10}(出血量：ml)，X_{11}(是否伴中线移位：是=1，否=0)，X_{12}(是否伴脑梗

死: 是=1,否=0), X_{13}(是否伴脑积水: 是=1,否=0), X_{14}(是否破入脑室: 是=1,否=0), X_{15}(是否伴脑疝: 是=1,否=0), X_{16}(治疗方式: 保守治疗=1,外科治疗=0), X_{17}(是否继发感染: 是=1,否=0), X_{18}(是否伴消化道出血: 是=1、否=0), X_{19}(是否伴心功能衰竭: 是=1,否=0), X_{20}(是否伴心律失常: 是=1,否=0), X_{12}(是否伴电解质失常: 是=1,否=0), X_{22}(是否有糖尿病病史: 是=1,否=0), X_{23}(是否有脑卒中史: 是=1,否=0), X_{24}(是否有肾脏疾病史: 是=1,否=0), X_{25}(是否有心脏疾病史: 是=1,否=0)。其中,白细胞计数,体温,收缩压,舒张压, 平均动脉压为患者起病后 1~5 d 的最高值。(2)观察结局变量: Y(死亡: Y=0,好转: Y=1)。

2.2 研究结果

2.2.1 共线性诊断结果 共线性诊断结果显示条件指数 (condition index) 最大为 228.261 86, > 100, 表示有强相关。进行因素间的相关分析结果显示: X_5(体温)与 X_4(白细胞计数), X_5(体温)与 X_9(Glassgow昏迷指数), X_6(收缩压)与 X_7(舒张压), X_6(收缩压)与 X_8(平均动脉压), X_7(舒张压)与 X_8(平均动脉压), X_3(血糖)与 X_{22}(是否患糖尿病)间相关系数做假设检验,均 P<0.0001,说明这些变量间存在相关性。

2.2.2 多因素 Logistic 回归分析 用逐步法筛选变量,进行多因素 Logistic 回归分析,结果见表 1。由有 1 可见, X_3(血糖)、X_5(体温)、X_9(Glassgow昏迷指数)、X_{14}(是否破入脑室)、X_{15}(是否伴脑疝)、X_{16}(治疗方式)、X_{18}(是否伴消化道出血)对高血压脑卒中患者预后的影响有统计学意义。

表 1 多因素 Logistic 回归分析结果

变量	b_j	S_{b_j}	u	P	OR
X_3	0.1868	0.0662	7.9675	0.0048	1.205
X_5	0.7400	0.2604	8.0747	0.0045	2.096
X_9	-0.2013	0.0553	13.2491	0.0003	0.818
X_{14}	1.7060	0.4650	13.4587	0.0002	5.507
X_{15}	2.3559	0.8654	7.4117	0.0065	10.548
X_{16}	1.6337	0.5603	8.4999	0.0036	5.123
X_{18}	1.3373	0.5671	5.5614	0.0184	3.809

2.2.3 主成分改进的 Logistic 回归分析 运用 SAS 软件中的 PRINCOMP 过程步, 对 25 个变量进行主成分分析,选择前 15 个主成分(累积贡献率达到 81.83%)进入 Logistic 回归模型。然后将原 25 个变量回代入估计的模型中,得到原始变量的回归系数估计值。通过 Wald 检验对所得的原变量系数估计值进行显著性检验。运用 SAS 软件中的 IML 过程步,按照公式(6), (7), (8)计算出 b_j, U, P, OR。结果见表 2。由表 2 可见, X_1(性别)、X_2(年龄)、X_3(血糖)、X_4(白细胞计数)、X_5(体温)、X_6(收缩压)、X_7(舒张压)、X_8(平均动脉压)、X_9(Glassgow昏迷指数)、X_{11}(是否伴中线移位)、X_{12}(是否伴脑梗塞)、X_{13}(是否伴脑积水)、X_{14}(是否破入脑室)、X_{15}(是否伴脑疝)、X_{16}(治疗方式)、X_{17}(是否继发感染: 是=1、否=0), X_{18}(是否伴消化道出血), X_{19}(是否伴心功能衰竭), X_{20}(是否伴心律失常), X_{21}(是否伴电解质失常), X_{23}(是否有脑卒中史), X_{25}(是否有心脏疾病史)对高血压脑出血患者预后的影响有统计学意义。各估计系数的标准误均有所减小,提示模型结构较为稳定,其结果的可靠性更高。

3 讨论

目前临床和流行病学研究越来越多的是多个因素对某个指标的影响, 因此多重线性回归法和 Logistic 回归分析得到广泛的应用。但有些人忽视了变量之间的相关性, 或是认为多重线性回归模型中需要处理共线性问题, Logistic 回归不需要处理共线性,把尽可能多的变量放入模型中得到阳性结果。

如果忽视了变量之间的相关性, 就会导致参数估计值的标准误变得很大; 回归方程不稳定,增加或减少某几个观察值,估计值可能会发生很大的变化; t 检验不准确, 误将应保留在模型中的重要变量舍弃,甚至导致估计值正负号改变,导致结果与客观实际不符合。

现在人们常用的消除共线性的方法就是应用逐步回归法筛选变量,剔除一些变量。在多重线性回归中,消除共线性的方法很多,例如,在 SAS 软件中 REG 过程步的 PCOMIT 选项就可以针对每个数值 m, REG 过程剔除最小的 m个主成分后进行非完全主成分分析,其原理与本研究中的一致,但在 Logistic 回归分析中消除共线性的方法就很少。本研究采用的主成分改进 Logistic 回归分析方法, 吸取了主成分分析的优点, 从众多原始指标之间的相互关系入手,寻找少数综合指标以概括原始指标信息,这些综合指标即保留了原始指标的主要信息, 又互不相关,消除了因素间的共线性,取得了较好的结果。

参考文献:

[1] 刘韵源, 主编. 状态风险分析及其在生物医学中的应用——定常协变量问题 [M]. 北京: 北京科学出版社,

表 2 主成分改进的 Logistic 回归分析结果

变量	b_j	S_{b_j}	u	P	OR
X_1	0.1237197	0.0087135	9.6609	0.00000	1.13170
X_2	0.2303586	0.0030739	5.0990	0.00000	1.25905
X_3	0.1879743	0.0018544	6.8970	0.00000	1.20680
X_4	0.3313857	0.0068413	32.7800	0.00000	1.39290
X_5	0.4121836	0.0006879	30.2228	0.00000	1.51011
X_6	0.2207328	0.0009278	5.5958	0.00000	1.24699
X_7	0.0634789	0.0009280	4.1676	0.000015	1.06554
X_8	0.1388789	0.0026839	4.7440	0.000001	1.14898
X_9	- 0.82214	0.0050502	- 11.0768	0.000000	0.43949
X_{11}	0.1075684	0.0024864	2.9436	0.001622	1.11357
X_{12}	0.3480523	0.0040796	5.8049	0.000000	1.41631
X_{13}	0.1989104	0.0030716	4.4061	0.000005	1.22007
X_{14}	0.2369157	0.0017937	8.9867	0.000000	1.26733
X_{15}	0.390472	0.0014913	17.8151	0.000000	1.47768
X_{16}	0.2832678	0.0015456	12.4702	0.000000	1.32746
X_{17}	0.1446714	0.0045251	2.1753	0.014803	1.15566
X_{18}	0.2946157	0.0016442	12.1913	0.000000	1.34261
X_{19}	0.3882448	0.0014272	18.5088	0.000000	1.47439
X_{20}	0.0972353	0.0025267	2.6007	0.004652	1.10212
X_{21}	0.2401643	0.0004303	37.9733	0.000000	1.27146
X_{23}	0.3666159	0.0018745	13.3073	0.000000	1.44284
X_{25}	0.0835794	0.0015108	3.7642	0.000084	1.08717

1990: 97.

[2] 方积乾.医学统计学与电脑实验[M].第 2 版.上海:上海科学技术出版社, 2001:441.

[3] 王 骏.Logistic 回归诊断及 SAS 实现[J].数理医药学杂志, 2005, 18(1): 36.

[4] 裘炯良. 主成分改进的 Logistic 回归模型方法在流行病学分析中的应用[J].中国热带医学,2005, 5(2): 207- 209.

·短篇论著·

皂擦浴和贝复剂治疗烧伤后残余创面 36 例分析

展红波

(靖江市人民医院 烧伤科, 江苏靖江 214500)

关键词: 皂擦浴; 贝复剂; 烧伤

中图分类号: R644 文献标识码: B 文章编号: 1673- 0399(2008) 04- 0520- 01

近年的研究结果提示,局部氧治疗、负压引流以及组织工程全层皮肤作为活性敷料能有效修复烧伤后期残余创面[1]。但由于其对器材的特殊要求,该法在基层医院的使用受到了一定的限制。本院自 2000 年 6 月至 2007 年 8 月,应用香皂擦浴和贝复剂(重组牛成纤维细胞生长因子- 2)治疗烧伤后期残余创面,取得良好疗效,现报道如下。

1 临床资料

1.1 一般资料

本组共 36 例,本院治疗病例 15 例,外院 21 例,创面均为烧伤后 40 d 以上未愈残余创面, 分别为深Ⅱ度烧伤自行愈合后重新开放的创面、Ⅲ度烧伤创面及其愈合后破溃创面、皮片间隙因感染扩大的创面、小面积深度烧伤经久不愈的创面。创面单处直径 0.5~2.0 cm,最大直径 4.0 cm。致伤原

(下转第 558 页)

收稿日期: 2008- 03- 11 作者简介: 展红波(1971-),男,江苏靖江人,主治医师,医学学士,主要从事烧伤科工作。