DOI:10.16016/j.1000-5404.1994.05.006

Logistic 回归模型异常值的判定原理探讨

易 东 许汝福 (第三军医大学预防医学系卫生统计学教研室) 630038
夏结来 (第四军医大学预防医学系卫生统计学教研室) 710032

提 要 笔者综合运用非线性回归诊断分析,建立了 Logistic 回归模型异常值判定的基本原理和方法。其基本思想是:(1)利用 Pearson 残差初步确定 Logistic 回归模型的可疑异常值;(2)利用 Score 检验统计量对可疑异常值进行假设检验。最后,作者将此方法应用于急性淋巴细胞性白血病的预后分析中,并对异常值的处理进行了初步探讨。

关键词 异常值;Logistic 模型;白血病;预后
中图法分类号 R195.1

Logistic 模型在病因分析中得到广泛应用。但是人们在对医学资料进行 Logistic 回归分析时往往忽略异常值的影响。本文综合运用统计学中新近发展起来的非线性回归诊断分析,提出实用而有效的 Logistic 回归模型的异常点判定方法。

1 异常点的诊断方法

在回归模型中所谓的异常点是反映对既定模型偏离很大的数据点,这里所说的"偏离"主要是针对模型的误差项而言。下面我们给出 logistic 回归模型及其基本假设:

设(y_i,x_i^T),$i=1,2\cdots,n$ 表示 n 组数据点,x_i^T 表示自变量组成的向量,未知参数为 $\beta=(\beta_1,\beta_2,\cdots\beta_p)^T$,假若 y_i 服从二项分布 $b(n_i,p_i)$,则得到 Logistic 回归模型,记为:

$$Logit(p_i)=Ln(\frac{p_i}{1-p_i})=x_i^T\beta,i=1,2\cdots,n \quad (1)$$

其中 n_i 表示第 i 组观察单位的个数,p_i 是出现某种预定结果(如生存、发病等)的概率。根据异常值的定义,下面我们将分两步给出其判别方法。

1.1 利用 Pearson 残差对模型进行初步诊断

此步的目的是根据拟合后残差的大小,初步断定哪些点可能是异常点,因此,首先要给出其残差表达式。Pearson 残差如下[1]:

$$rp_i=\frac{y_i-n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}} \quad (2)$$

其中 $\hat{p}_i$ 是 p_i 的估计值,由(3)式求得。

$$\hat{p}_i=\frac{e^{\beta_0-\Sigma\beta_ix_i}}{1+e^{\beta_0-\Sigma\beta_ix_i}} \quad (3)$$

和线性模型的普通残差类似,Pearson 残差也是 Logistic 回归模型的拟合优度统计量。如果 rp_i 较大,则说明拟合不太好,(y_i,x_i^T)就有可能成为异常点。下面将用假设检验,进一步确定其是否异常点。

1.2 利用 Score 检验统计量[2]进行假设检验

在利用 Pearson 残差得到可能异常点后,用 Score 统计量进行假设检验。Score 假设检验的基本依据是建立在数据删除模型 $\eta(i)=\chi(i)\beta$ 和均值漂移模型 $\eta=\chi\beta+\gamma C_i$ 的参数 β 之最大似然估计满足同样的似然方程。根据这一等价性定理,我们可建立假设检验如下:

$H_0:\gamma=0$,即第 i 组数据点为异常点

$H_1:\gamma\neq0$,即第 i 组数据点为非异常点

检验 H_0 的 Score 检验统计量为:

$$SC_i=\frac{rp_i^2}{1-h_{ii}} \quad (4)$$

当 Ho 成立时,SC_i 的渐近分布为 $\chi^2(1)$。h_{ii} 是投影矩阵 H 对角线上的第 i 个元素。其中

$$H=W^{\frac{1}{2}}X(X^TWX)^{-1}X^TW^{\frac{1}{2}} \quad (5)$$

$$W=diag(w_1,w_2,\cdots w_n) \quad (6)$$

$$w_i=n_i\hat{p}_i(1-\hat{p}_i) \quad (7)$$

2 应用

为了对急性淋巴细胞性白血病进行预后分析,这里统计了30名患者入院治疗时的外周血中白细胞数 x_1 (K/mm^3)、淋巴结浸润等级 x_2 (分为0,1,2,3四级)、出院后巩固治疗 x_3 (有巩固治疗为1,无巩固治疗为0),并随访取得病人的生存时间 t_i,对 t_i 进行重新赋值,若 $t_i\geq12$个月,则 $y_i=0$,若 $t_i<12$个月,则 $y_i=1$。其具体数据见表1第(1)~(6)列。

表1资料经拟合 Logistic 模型得:

$$Logit(\hat{p}_i)=Ln(\frac{\hat{p}_i}{1-\hat{p}_i})=0.5897-0.0060x_1+1.1106x_2-2.2440x_3$$

即 $$\hat{p}_i=\frac{e^{0.5897-0.0060x_1-1.1106x_2-2.2440x_3}}{1+e^{0.5897-0.0060x_1+1.1106x_2-2.2440x_3}}$$

此时 $\hat{p}_i$ 为患者一年内死亡的概率估计值。将各组数据 x_1、x_2、x_3 代入上式得 $\hat{p}_i$(见表1第(7)栏)。用 Pascal 语言将上述公式编制程序,在 AST 386机上运算得各组数据的 rp_i 值、SC_i 值及检验结果(见表2),第2、8两组数据SC检验 $P<0.05$,是两个异常点。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表1　30例急性淋巴细胞白血病患者资料

Tab 1　The data of thirty cases of acute lymphatic Leukemia

序号 (1)	x_{i1} (2)	x_{i2} (3)	x_{i3} (4)	y_i (5)	n_i (6)	$\hat{p}_i$ (7)
1	21	2	1	1	1	0.5649
2	6	0	1	1	1	0.1335
3	3	0	0	1	2	0.1335
4	3	2	1	1	1	0.5912
5	5	0	0	1	1	0.6364
6	6	0	0	1	1	0.6350
7	128	2	1	1	1	0.4059
8	4	0	1	1	1	0.1349
9	35	0	0	1	1	0.5938
10	62	0	0	0	1	0.5542
11	2	0	0	0	2	0.6405
12	11	0	1	0	1	0.1301
13	9	0	1	0	1	0.1314
14	22	0	1	0	1	0.1228
15	2	2	1	0	1	0.5926
16	2	0	1	0	3	0.1363
17	3	2	1	0	1	0.5912
18	4	0	1	0	2	0.1349
19	5	0	1	0	2	0.1342
20	245	2	1	0	1	0.2529
21	1	0	1	0	2	0.1370
22	32	0	1	0	1	0.1165
23	13	0	1	0	1	0.1287

表2　Logistic 回归的 Pearson 残差及 Score 检验

Tab 2　The Pearson error of Logistic model and the Scor test

序号	rP_i	w_i	h_{ii}	SC_i
1	0.878	0.2458	0.2076	0.972
2	2.548	0.1157	0.0630	6.926*
3	−0.410	0.4613	0.0004	0.168
4	0.832	0.2417	0.2399	0.910
5	0.756	0.2314	0.0005	0.572
6	0.758	0.2318	0.0007	0.575
7	1.210	0.2411	0.2669	1.997
8	2.532	0.1167	0.0636	6.848*
9	0.827	0.2412	0.0252	0.702
10	−1.115	0.2471	0.0810	1.353
11	−1.888	0.4605	0.0002	3.564
12	−0.387	0.1132	0.0617	0.159
13	−0.389	0.1142	0.0621	0.161
14	−0.374	0.1077	0.0606	0.149
15	−1.206	0.2414	0.2420	1.919
16	−0.688	0.3532	0.1930	0.587
17	−1.203	0.2417	0.2399	1.903
18	−0.558	0.2334	0.1272	0.357
19	−0.557	0.2324	0.1266	0.355
20	−0.582	0.1890	0.6858	1.077
21	−0.564	0.2365	0.1295	0.365
22	−0.363	0.1029	0.0613	0.140
23	−0.384	0.1122	0.0613	0.157

3　讨论

从表2中第2栏可看出，对30例急性淋巴细胞白血病患者的数据资料拟合 Logistic 回归模型时，其 rp_i 的绝对值也即是 Pearson 残差较大的有8组数据。但这只能初步判定有8个异常点。

在初步确定的8个可疑异常点的基础上，应对之进行 Score 假设检验。从表2中第5栏可看出第2、8两点是异常点（这里根据 SC_i 服从 $\chi^2_{(1)}$ 分布。其对应的 Pearson 残差也最大。

对于异常点的产生，其原因较复杂。如表1中30例急性淋巴细胞白血病患者数据资料经判定第2、8两点是异常点。这两个异常点的产生也许是由偶然误差所致，也许这种资料的本身就不能拟合 Logistic 回归模型，而应采用其它回归模型。

关于 Logistic 异常点的处理目前还不完善，其研究也非常活跃。但一般认为，若数据中有异常点，首先应根据专业知识、数据收集的情况，分析其产生原因后酌情处理。如来自测量或记录错误，应剔除或校正，否则处置就必须持慎重态度，考虑是否采用新的模型，而不能只是简单地删除就算完事。因为在许多场合，异常点的出现恰好是我们探测某些事先不清楚的或许更为重要因素的线索。

（本文承蒙尹全焕教授指导，特此致谢！）

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

参考文献

1 Wang PC. Residual plots for detecting nonlinerarity in generalized liner model. Technometrics, 1987, 29: 435

2 Cox PR. Thoeretical statistics. London: Chapmam and Hall, 1974

(收稿:1993-11-16;修回:1994-09-14)

The principled research to assess the outliers of Logistic regression model

Yi Dong, Xia Jielai, Xu Rufu (Department of Medical Statistics, Third Military Medical College, Chongqing, 630038)

Abstract On the basis of the non-linear regression diagnostic analysis, the authors put forward a diagnostic method to assess the outliers of the Logistic regression model. The basic principle is as follow: (1) To preliminarily assess the doubious outliers with Pearson error; (2) To take hypothesis test to the doubious outliers with Score statistic. Finally it was used in the prognosis for cases of acute lymphatic leukemia and it has been simply discussed that how to deal with the outliers.

Key words outlier; Logistic model; Leukemia; prognosis

TRH 抗失血性休克的肾上腺素能受体及阿片受体机理*

The adrenoceptor and opioid receptor mechanisms of thyrotropin releasing Hormone against hemorrhagic shock

刘良明 陈惠孙 胡德耀 (第三军医大学野战外科研究所) 630042

本研究以失血性休克为主要模型从两个方面进行了研究,第一是研究了促甲状腺素释放激素(TRH)抗休克的肾上腺素能受体机理,包括:(1)TRH 抗休克与哪些肾上腺素能受体有关;(2)它们是什么关系;(3)TRH 对休克动物肾上腺素能受体数量和亲和力有无影响;(4)TRH 对心肌跨膜离子转运有何影响。第二是研究 TRH 与阿片受体的关系,包括:(1)确证 TRH 与内阿片肽间的关系;(2)TRH 拮抗内阿片肽的作用是否直接作用在阿片受体上,TRH 与哪些阿片受体亚型有直接作用;(3)TRH 对休克动物脑阿片受体有无调节作用。研究结果显示,TRH 的抗休克作用主要由 β-肾上腺素能受体和 DA 受体介导,与 α-肾上腺素能受体关系不大;TRH 可增强休克大鼠心脏 β-受体和 DA 受体对肾上腺素能激动剂的反应性,增强其正性肌力作用,即 TRH 对休克大鼠心脏 β-受体和 DA 受体有一定的增敏作用;TRH 可使休克大鼠心脏 β-受体和 DA 受体发生增数调节,同时可提高 β 受体的亲和力;TRH 可增加心肌细胞慢钙内流,延长钙通道开放时间,增加肾上腺素能激动剂促钙内流作用;TRH 既可拮抗 β-内啡肽对正常动物的心血管抑制作用,又可拮抗 β-内啡肽对休克动物的心肌抑制作用;TRH 可以直接作用在阿片受体上拮抗阿片受体的作用,TRH 的这一作用主要与 δ-阿片受体有关,而与 μ-和 κ-受体关系不大;同时,TRH 可降低休克动物脑 δ-阿片受体数量。这些结果提示,TRH 增敏休克动物肾上腺素能 β 和 DA 受体,增强其对肾上腺素能受体激动剂的反应性可能是 TRH 抗休克的重要机理之一,TRH 增数调节休克动物肾上腺素能受体并提高其亲和力,延长心肌细胞钙通道开放时间,增加肾上腺素能受体激动剂促钙内流作用可能是 TRH 增敏调节肾上腺素能受体的重要途径。TRH 拮抗内阿片肽的心血管抑制作用可能是 TRH 抗休克的又一重要机理,TRH 直接作用在阿片受体上拮抗 δ-阿片受体,同时发生减数调节可能是 TRH 拮抗内阿片肽作用发挥抗休克作用的重要原因。

关键词 促甲状腺素释放激素;休克,出血性;肾上腺素能受体;阿片受体;阿片受体拮抗剂;膜片钳;钙内电流;兔;大鼠;豚鼠

中图法分类号 Q575.14;R605.971

* 国家自然科学基金资助项目,No. 39370663

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net