DOI:10.16016/j.1000-5404.2003.18.034
第25卷第18期 第三军医大学学报 Vol. 25, No. 18
1670 2003年9月 ACTA ACADEMIAE MEDICINAE MILITARIS TERTIAE Sep. 2003

技术方法

文章编号:1000-5404(2003)18-1670-02

Breslow 算法非容许误差校正

Rectification of non-tolerance error in Breslow's algorithm

梁正东[1],王开发[1],杨 旭[2] ([1]第三军医大学基础医学部数学教研室,重庆 400038;[2]重庆电力教育培训中心,重庆 400053)

提 要:目的 对多元生存分析Breslow算法的基础生存率$S_o(t)$的估计会在实用中产生与概率意义不相符的非容许误差进行校正。方法 根据Cox比例危险度概率模型原理,利用参数极大似然估计法对$S_o(t)$做推导分析。结果 导出了$S_o(t)$的无偏计算公式。结论 纠正了原估$S_o(t)$计公式的非容许误差,并通过不同算法对比的算例,说明新估计公式的合理性。

在多元生存分析中,Cox算法虽然稳定可靠,但由于该法需计算各个重复死亡点上的不同组合[1],对于大样本多重复数据面临着计算技术上的困难,产生所谓"组合爆炸"。Breslow算法避免了计算组合,是对Cox算法的一种改进。相对说来计算较简单,节约机时,微机存贮空间也能满足。无重复时,两种方法对β的估计一致;重复点不多且重复数目不大时,两者差别也不明显。从实用上看,Breslow算法是有效的,仍是目前普遍采用的一种方法。正如文献[2]指出的那样,计算基础生存率$S_0(t)$,"一般采用乘积极限法或Breslow法估计"。

然而,在文献[3]中Breslow算法给出的基础生存率$S_0(t)$的计算公式[见下述(1)式],会在实用中发生括号内第二项值于个别点大于1,使得$S_0(t)$值为负的情形,显然在概率意义下不合理,是非容许的误差。

$$S_0(t_i)=\prod_{t_l\leqslant t_i}[1-m_l/\sum_{j\in R(t_l)}(z_j\beta)] \qquad (1)$$

由于文献[2]中省略了推导过程,很难直观判断(1)式的来历根据。事实上,下面将推导出(1)式只是下述(2)式的一个粗糙近似,是造成实用中非容许误差的原因。因此笔者建议计算$S_o(t)$不用(1)式,而改用(2)式进行,即可避免上述误差的产生。

$$S_0(t_i)=\prod_{t_l\leqslant t_i}exp[-m_l/\sum_{j\in R(t_l)}exp(z_j\beta] \qquad (2)$$

1 推证

在生存分析中,生存率函数定义为

$$S(t,z)=P(T\geqslant t,z)=1-F(t,z) \qquad (3)$$

它表示具有协变量z(p维向量)的个体生存期T大于等于t的概率,$F(t,z)$为死亡率函数。

危险率函数$h(t,z)$的定义及其与$s(t,z)$的关系为

$$h(t,z)=\lim_{\triangle t\to 0}\frac{P(t\leqslant T\leqslant t+\triangle t|\ T\geqslant t,z)}{\triangle t}$$

$$=\lim_{\triangle t\to 0}\frac{[F(t+\triangle t,z)-F(t,z)]/\triangle t}{s(t,z)}$$

$$=\frac{dF(t,z)}{s(t,z)dt}=-\frac{ds(t,z)}{s(t,z)dt} \qquad (4)$$

它表示个体生存期T已达到t的条件下而在t时刻的瞬时死亡率。对(4)式从0到t积分,并化简得

$$S(t,z)=exp[-\int_o^t h(u,z)du] \qquad (5)$$

实际采样时,时间都是离散的。设采样数据记为(T_i, w_i, z_i), i=1, 2, …, n, T_i为第i个个体的观测时间,w_i为标识变量,$w_i=1$为第i个个体死亡,$w_i=0$则为截尾(失访或终止随访),z_i为其p个协变量构成的向量。将$w_i=1$者所对应的T_i按由小到大排序,记为$t_1<t_2<\cdots<t_k$,其中k表示不同死亡时间点的个数,显然$k\leqslant n$。记数出各死亡时刻t_i处的在危险个体数r_i、死亡个体数m_i、区间内的截尾数目l_i。用$R(t_i)$表示t_i时刻的r_i个在危险个体构成的集合。

Breslow法是将Cox法的比例危险度模型中的基础危险函数$h_0(t)$,假设为各死亡点之间的区间内是连续的,且不同区间为一不同常数。即在(t_{i-1}, t_i)内令$h_{0i}(t)=\lambda_i$(λ_i为常数)。这时的比例危险度模型形式为。

$$h_i(t,z)=\lambda_i\cdot exp(z\beta) \qquad (6)$$

$exp(z\beta)$与时间无关,而$(z\beta)$是协变量的线性式,β为待估的参数向量。

将(6)式代入(5)式,并从t_{i-1}到t积分(特别地$t_0=0$),得到第i个区间内生存函数为

$$s_i(t,z)=exp[-\lambda_i(t-t_{i-1})\cdot exp(z\beta)]$$

当$z=0$时,即为该区间内的基础生存函数。形为下述(8)式,是区间上的一个指数函数。

$$s_{0i}(t)=exp[-\lambda_i(t-t_{i-1})] \qquad (8)$$

于是由(6)式及(7)式可写出似然函数为

$$L(\lambda_i,\beta)=\prod_{i=1}^{k}\left\{\prod_{j=1}^{m_i}[\lambda_i exp(z_j\beta)]\prod_{j\in R(t_i)}exp[-\lambda_i(t_i-t_{i-1})\cdot exp(z_j\beta)]\right\}$$

作者简介:梁正东(1947-),男,四川省阆中市人,副教授,主要从事生物统计学方面的研究,发表论文14篇。电话:(023)68752211

收稿日期:2002-05-04;修回日期:2003-01-09

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

$$= \prod_{i=1}^{k}\left\{ exp(s_i\beta)\lambda_i^{m_i} \prod_{j\in R(t_i)} exp[-\lambda_i(t_i - t_{i-1})\cdot exp(z_j\beta)]\right\} \quad (9)$$

第一个等式花括号内第一个乘积号表示在 t_i 时刻 m_i 个死亡个体的死亡概率, 第二个乘积号表示 r_i 个在危险个体的存活概率。其中 s_i 为 m_i 个死亡个体的 z 值之和。对数似然函数为

$$logL(\lambda_i, \beta) = \sum_{i=1}^{k}\left\{ s_i\beta + m_i log\lambda_i - \lambda_i(t_i - t_{i-1}) \sum_{j\in R(t_i)} exp(z_j\beta)\right\} \quad (10)$$

令 $\frac{\partial logL}{\partial \lambda_i} = \frac{m_i}{\lambda_i} - (t_i - t_{i-1}) \sum_{j\in R(t_i)} exp(z_j\beta) = 0$, 得到

$$\lambda_i = m_i / [(t_i - t_{i-1}) \sum_{j\in R(t_i)} exp(z_j\beta)] \quad (11)$$

将(11)式代入(10)式, 得

$$\log L(\beta) = \sum_{i=1}^{k}\left\{ s_i\beta - m_i \log \sum_{j\in R(t_i)} \exp(z_j\beta) + [m_i \log m_i - m_i \log(t_i - t_{i-1}) - m_i]\right\}$$

由于方括号内各项与β 无关, 故可写成

$$logL(\beta) = \sum_{i=1}^{k}[s_i\beta - m_i log \sum_{j\in R(t_i)} exp(z_j\beta)] \quad (12)$$

由此式并运用 Newton-Raphson 迭代解法便可估计出β 来。(11)及(12)式就是文献[2]中所给出的结果。将(11)式代入(6)式, 再将(6)式代入(5)式, 取 z = 0 并从 0 到 t_i 积分, 就得到个体的基础生存率估计公式为:

$$S_0(t_i) = exp[-\sum_{l=1}^{i} \int_{t_{l-1}}^{t_l} \lambda_l du]$$

$$= exp[-\sum_{l=1}^{i} \frac{m_l(t_l - t_{l-1})}{(t_l - t_{l-1}) \sum_{j\in R(t_l)} exp(z_j\beta)}]$$

$$= \prod_{t_l \leqslant t_i} exp[-m_l / \sum_{j\in R(t_l)} exp(z_j\beta)] \quad (13)$$

这就是(13)式即[(2)式]作为替代式的根据。注意到(13)式的因子是一个指数式, (1)式与之相比, 可见只是取了(13)式的幂级数展开式的前两项。由于取项太少, (1)式是一个粗糙的近似式, 因此在实用中于个别死亡点会产生负值的情况。

2　算例

表 1 资料摘自文献[1]。表中所列的实验数据, 是为研究抗癌药物 6-MP 对治疗白血病的效果而观测到的患者的缓解时间 t(单位为周)。其中"*"号表示截尾(censored), 并设立了一个对照组。这里涉及的协变量 z 只一个, 且规定 z = 1 为对照, z = 2 为用 6-MP。样本容量 n = 42。

表 1　白血病患者的缓解时间 t(单位: 周)

样本 1(6-MP)	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
样本 2(对照)	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

表 2 所列是用几种不同算法计算出的基础缓解率 $s_0(t)$ [$s_0(t) = P(T \geqslant t)$, 即缓解时间 T 大于等于 t 的概率]的估计结果。

3　讨论

表 2 中第 2 列的 Breslow 算法, 在 4 个时间点上出现了负值(见 *号), 而第 3 列的改进算法完全避免了这种情况的产生。第 3 列与第 5 列的非参数法的 Kaplan-Meier 估计结果比较, 曲线同样呈单调下降且具有一致性。第 3 列与第 4 列稳定性好的 Cox 算法结果相比较, 两者也很贴近。可见改进算法是适宜可行的。Cox 模型经过众多领域长期广泛应用结果表明, 它是多元生存分析中理论成熟、方法可靠的模型。为充分发挥该模型提取信息的效力, 需进一步着力研究在大样本多重复数据情况下的计算机算法技术。

表 2　$s_0(t)$ 的四种算法结果比较

缓解时间点 t (1)	Breslow 法 (2)	改进法 (3)	Cox 法 (4)	Kaplan-Meier 法 (5)
1	0.64	0.70	0.69	0.95
2	0.40	0.48	0.47	0.90
3	0.32	0.39	0.38	0.88
4	0.17	0.25	0.24	0.83
5	9.2×10^{-2}	0.16	0.15	0.78
6	1.7×10^{-2}	0.07	0.07	0.71
7	1.2×10^{-2}	5.2×10^{-2}	5.3×10^{-2}	0.68
8	-1.9×10^{-3}*	1.6×10^{-2}	1.9×10^{-2}	0.59
10	-1.2×10^{-3}*	1.1×10^{-2}	1.3×10^{-2}	0.56
11	-2.0×10^{-4}*	4.8×10^{-3}	6.2×10^{-3}	0.51
12	8.5×10^{-6}	1.7×10^{-3}	2.5×10^{-3}	0.45
13	2.7×10^{-6}	8.5×10^{-4}	1.3×10^{-3}	0.42
15	8.2×10^{-7}	4.2×10^{-4}	6.8×10^{-4}	0.39
16	1.4×10^{-7}	1.8×10^{-4}	3.2×10^{-4}	0.36
17	1.9×10^{-8}	7.8×10^{-5}	1.4×10^{-4}	0.34
22	-2.9×10^{-8}*	6.1×10^{-6}	2.7×10^{-5}	0.26
23	8.4×10^{-8}	1.2×10^{-7}	3.1×10^{-6}	0.18

关键词: 生存分析; 基础生存率; 校正误差
中图法分类号: R311　　文献标识码: B

参考文献:

[1] Cox D R. Regression models and life-table(with discussion)[J]. J Roy Statist Soc B, 1972, 34(2): 187-220.
[2] 张丕德. 预后因子存在共线性时 Cox 模型的拟合方法[J]. 中国卫生统计, 2000, 17(6): 357-358.
[3] Breslow N. Covariance analysis of censored survival data[J]. Biometrics, 1974, 30: 89-99.

(编辑　冯崇英)

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