文章编号:1000-5862(2016)04-0354-04

右删失数据下的限制平均寿命问题

邓文丽,欧阳菲
(江西师范大学数学与信息科学学院,江西 南昌 330022)

摘要:在右删失数据下构造了"新"的随机变量,在删失变量分布已知的条件下,该变量的样本均值是限制平均寿命的方差有限的无偏估计;在删失变量分布未知的条件下,利用该变量得到的样本均值是限制平均寿命的相合估计. 数值模拟说明所提出方法的有效性和可行性.

关键词:限制平均寿命;乘积极限估计;相合估计
中图分类号:O 212.7 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2016.04.05

0 引言

在生存分析和可靠性研究中,除了关心试验对象的平均生存时间外,限制平均寿命(restricted mean lifetime)也常常是关注的焦点,如在不同治疗方案对肿瘤病人的治疗效果的比较中,治疗效果通常会在有限时间内显现,治疗效果的比较通常也是基于有限时间内进行. 在小儿科的肝脏移植问题中,接受过肝脏移植的小孩大多数在 10 年内还需要进行下一次肝脏移植,如果用 T 表示首次肝脏移植后的存活时间,对 $E\{\min(T,10)\}$ 进行研究,应该会比研究 $E(T)$ 更有意义.

许多学者在 Cox 比例风险模型的背景下提出了关于比较限制平均寿命的方法. T. Karrison[1] 利用协变量对 2 个组的限制平均寿命进行比较,假定协变量对失效时间的影响符合比例风险模型,各个组有不同的基准风险,基准风险函数是分段的指数分布,不同组中协变量对风险函数的影响都是乘法效应. D. M. Zucker[2] 也采用了分层 Cox 模型,各个组的基准风险函数不相同,但是没有作具体假定. Chen Peiyun 等[3] 提出调整协变量后两组之间的限制平均寿命差异的点估计. Zhang Min 等[4] 基于信息删失数据在比例风险模型中对 2 个组的限制平均寿命进行了比较. Zhang Min 等[5] 推导出一种半参数双稳健的限制平均寿命差异的估计. 在对现有工作状态持续时间的研究中,Pan Qing 等[6] 建立了晋升和退休的半参数竞争风险模型,对现有工作任期的限制平均时间进行了估计. P. Royston[7] 通过限制平均寿命的差异来估计新的医疗方案所产生的效应. P. Royston 等[8] 基于限制平均寿命提出了一种联合 Cox 检验和置换检验的新检验方法.

本文拟在非参数模型下,对右删失数据的限制平均寿命进行估计. 在限制平均寿命问题中,采用无偏转换方法构造出的"新"的随机变量,不仅数学期望等于限制平均寿命,其方差也是显然有界的. 这种无偏转换的做法最早出现在文献[9]中,主要用于右删失数据下回归分析问题的处理. 文献[10-12]将无偏转换方法用于区间删失数据的回归问题的处理. 考虑到所构造估计量的方差有限问题,无偏转换方法在限制平均寿命的估计中更能发挥作用. 右删失数据下寿命随机变量的生存函数的乘积极限估计是 E. L. Kaplan 等[13] 提出来的,其在生存分析中的地位相当于经验分布函数在经典数理统计中的地位,其大样本性质已经得到了充分证明,但是它尾部的估计效果不是很理想. 在限制平均寿命的估计中,这个缺陷却不会显现出来. 基于以上考虑,本文拟利用无偏转换思想和生存函数的乘积极限估计,对非参数模型下的限制平均寿命进行估计.

1 基本模型

用 T 表示失效时间变量,C 表示删失时间变量. 对 n 个独立试验个体,T_i 为第 i 个个体的失效时间,C_i 为第 i 个个体的删失时间,T_i 与 C_i 是相互独立的. 试验中得到的观测值为 (Z_i,δ_i),其中 $Z_i=\min(T_i,$

收稿日期:2016-03-22
基金项目:国家自然科学青年基金(71001046)资助项目.
作者简介:邓文丽(1974-),女,江西南昌人,副教授,博士,主要从事数理统计研究.

$C_i)$,$\delta_i = I(T_i \leqslant C_i)$,$i = 1,2,\cdots,n$. 根据这组样本对限制平均寿命 $E[\min(T,L)]$ 进行估计,其中 L 为大于 0 的常数.

记 $T_i^{(L)} = \min(T_i,L)$,$C_i^{(L)} = \min(C_i,L)$,$Z_i^{(L)} = \min(Z_i,L)$,$\delta_i^{(L)} = I(T_i^{(L)} \leqslant C_i^{(L)})$,$i = 1,2,\cdots,n$. 且 $T_i^{(L)}, C_i^{(L)} \in (0,L]$,$Z_i^{(L)} = \min(T_i,C_i,L) = \min(T_i^{(L)}, C_i^{(L)})$,$\delta_i^{(L)}$ 可以通过 δ_i 得到. 当 $\delta_i = 1$ 时,则 $T_i \leqslant C_i$,显然 $T_i^{(L)} \leqslant C_i^{(L)}$,$\delta_i^{(L)} = 1$;当 $\delta_i = 0$ 时,$T_i > C_i$,需要分情况讨论,若 $Z_i = \min(T_i,C_i) \geqslant L$,则 $T_i^{(L)} = C_i^{(L)} = L$,$\delta_i^{(L)} = 1$;若 $Z_i < L$,则 $T_i^{(L)} > C_i^{(L)}$,$\delta_i^{(L)} = 0$.

2　限制平均寿命的估计

基于一组“新”的右删失数据 $(Z_i^{(L)},\delta_i^{(L)})(i = 1,2,\cdots,n)$,对 $E[T^{(L)}]$ 进行估计,$T^{(L)} = \min(T,L)$. $T_i^{(L)}$ 和 $C_i^{(L)}$ 的分布分别记为 $F(\cdot)$ 和 $G(\cdot)(i = 1,2,\cdots,n)$. 若 $G(\cdot)$ 已知,则构造

$$T_i(G) = \delta_i^{(L)}\varphi_1(Z_i^{(L)},G(Z_i^{(L)})) + (1-\delta_i^{(L)})\varphi_2(Z_i^{(L)},G(Z_i^{(L)})), \quad (1)$$

其中 φ_1,φ_2 是连续可微函数,它们与 $F(\cdot)$ 无关,且满足方程

$$[1-G(t)]\varphi_1(t,G(t)) + \int_0^t \varphi_2(x, G(x))\,\mathrm{d}G(x) = t, t \in [0,L], \quad (2)$$

从而

$$ET_i(G) = E[\delta_i^{(L)}\varphi_1(Z_i^{(L)}) + (1-\delta_i^{(L)})\varphi_2(Z_i^{(L)})] = \iint\limits_{t\leqslant c\leqslant L} \varphi_1(t,G(t))\,\mathrm{d}F(t)\,\mathrm{d}G(c) + \iint\limits_{c<t\leqslant L} \varphi_2(c, G(c))\,\mathrm{d}F(t)\,\mathrm{d}G(c) = \int_0^L \left\{[1-G(t)]\varphi_1(t, G(t)) + \int_0^t \varphi_2(c,G(c))\,\mathrm{d}G(c)\right\}\mathrm{d}F(t) = ET_i^{(L)}.$$

取 $\varphi_1(t,G(t)) = t/[1-G(t)]$,$\varphi_2(t,G(t)) = 0, t \in [0,L]$,显然 φ_1,φ_2 是连续可微函数,且满足(2)式,$T_i(G) = \delta_i^{(L)}Z_i^{(L)}/[1-G(Z_i^{(L)})]$,以及

$$\mathrm{Var}(T_i(G)) = \int_0^L t^2/[1-G(t)]^2\mathrm{d}F(t) - [E(T(G))]^2 = \mathrm{Var}(T^{(L)}) + \int_0^L (1/[1-G(t)]^2 - 1)\,t^2\mathrm{d}F(t) \geqslant \mathrm{Var}(T^{(L)}).$$

由于 $Z_i^{(L)}$ 有界,φ_1,φ_2 连续,所以 $\mathrm{Var}(T_i(G))$ 有界. 因此可以构造限制平均寿命 $E[T^{(L)}]$ 的无偏估计 $\overline{T}(G) = \frac{1}{n}\sum_{i=1}^n T_i(G)$. 记 $H(G) = \{\overline{T}(G) \mid T_i(G)(i = 1,2,\cdots,n)$ 满足(1)式,φ_1,φ_2 连续可微且满足(2)式$\}$,显然集合 $H(G)$ 中所有元素都是限制平均寿命 $ET^{(L)}$ 的无偏估计.

在实际问题中,$G(\cdot)$ 通常是未知的,可以用 Kaplan-Meier 估计得到 $\hat{G}(\cdot)$. 因而在(1)式中可以用 $\hat{G}(\cdot)$ 代替 $G(\cdot)$,得到

$$T_i(\hat{G}) = \delta_i^{(L)}\varphi_1(Z_i^{(L)},\hat{G}(Z_i^{(L)})) + (1-\delta_i^{(L)})\varphi_2(Z_i^{(L)},\hat{G}(Z_i^{(L)})), \quad (3)$$

所以限制平均寿命的估计 $\overline{T}(\hat{G}) = \frac{1}{n}\sum_{i=1}^n T_i(\hat{G})$. 从而得到在 G 未知的情形下限制平均寿命的估计类 $H(\hat{G}) = \{\overline{T}(\hat{G}) \mid T_i(\hat{G})(i = 1,2,\cdots,n)$ 满足(3)式,φ_1,φ_2 连续可微且满足(2)式$\}$.

定理 1　若 $H(\hat{G})$ 非空,其任意元素都是限制平均生存时间的强相合估计,则收敛速度为 $\sqrt{(\log\log n)/n}$.

证　$H(\hat{G})$ 非空显然. 任给 $\overline{T}(\hat{G}) \in H(\hat{G})$ 均有对应的 $\overline{T}(G) \in H(G)$,

$$T_i(\hat{G}) = \delta_i^{(L)}\varphi_1(Z_i^{(L)},\hat{G}(Z_i^{(L)})) + (1-\delta_i^{(L)})\varphi_2(Z_i^{(L)},\hat{G}(Z_i^{(L)})),$$

$$T_i(G) = \delta_i^{(L)}\varphi_1(Z_i^{(L)},G(Z_i^{(L)})) + (1-\delta_i^{(L)})\varphi_2(Z_i^{(L)},G(Z_i^{(L)})),$$

φ_1,φ_2 是连续可微函数,且满足方程

$$[1-G(t)]\varphi_1(t,G(t)) + \int_0^t \varphi_2(x, G(x))\,\mathrm{d}G(x) = t, t \in [0,L].$$

因为

$$|\overline{T}(\hat{G}) - E(T^{(L)})| = |\overline{T}(\hat{G}) - \overline{T}(G))| + |\overline{T}(G) - E(T^{(L)})|,$$

其中 $\lim\limits_{n\to+\infty}|\overline{T}(G) - E(T^{(L)})| = 0$ a. s.,所以只需证 $\lim\limits_{n\to+\infty}|\overline{T}(G) - \overline{T}(\hat{G})| = 0$ a. s..

由 φ_1,φ_2 连续知,$\forall\varepsilon > 0$,$\exists\varepsilon_1$,当 $|\hat{G}(t) - G(t)| < \varepsilon_1$ 时,

$$|\varphi_1(t,\hat{G}(t)) - \varphi_1(t,G(t))| < \varepsilon,$$

$$|\varphi_2(t,\hat{G}(t)) - \varphi_2(t,G(t))| < \varepsilon.$$

由 A. V. Peterson[14] 的关于乘积极限估计的强相合性结论知,除去 1 个零测集,$\forall\varepsilon_1 > 0$,$\exists N$,当 $n > N$ 时,有 $\sup\limits_{t\in[0,L]}|\hat{G}(t) - G(t)| < \varepsilon_1$.

综上所述,除去 1 个零测集,$\forall\varepsilon > 0$,$\exists N$,当 $n > N$ 时,

$$|\bar{T}(\hat{G}) - \bar{T}(G)| \leqslant \frac{1}{n}\sum_{i=1}^{n} |T_i(\hat{G}) - T_i(G)| < \varepsilon,$$

因为 $\sup_{t\in[0,J]} |\hat{G}(t) - G(t)| = O(\sqrt{(\log\log n)/n})$ a. s. ,φ_1,φ_2 是连续可微函数,故

$$|\bar{T}(\hat{G}) - E(T^{(L)})| = O(\sqrt{(\log\log n)/n}) \text{ a. s. .}$$

3 数值模拟

在模拟计算中失效时间 T 采用威布尔分布随机生成,这是因为威布尔分布的危险率不是常数,它与指数分布相比有较广泛的应用,其可以用于调查深槽轮滚珠轴承的疲劳寿命、描写电子管的失效等. 威布尔分布函数为 $F(t) = 1 - e^{-(\lambda t)^{\gamma}}$,其中 γ 是分布曲线的形状参数,λ 是尺度参数. 选取 $\lambda = 0.5$,$\gamma = 2$,$L = 0.8$,得到限制平均寿命为 0.402. 删失时间 C 的分布选取为在 $(0,A)$ 上的均匀分布,调整 A 的大小可以改变删失的比例. 模拟计算次数均为 1 000 次.

$H(\hat{G})$ 给出的是 1 个估计类,不同的 (φ_1,φ_2) 对应着不同的估计. 表 1 列出了在不同 (φ_1,φ_2) 下得到的估计效果,其中样本数 $n = 200$,删失时间 C 服从均匀分布 $U(0,1)$,删失比率为 0.434. 直接利用生存函数的乘积极限估计也可以求出限制平均寿命的估计. 结果表明: 在无偏转换方法中选择合适的 (φ_1,φ_2),可以得到比直接采用乘积极限估计方法更优良的估计量.

表 1 在不同 (φ_1,φ_2) 下 $E[T^{(L)}]$ 估计效果的比较

估计方法	估计量	偏度	标准差
乘积极限估计方法	0.432 0	0.030 0	0.018 4
$\varphi_1(t,G(t)) = t/[1-G(t)]$,$\varphi_2(t,G(t)) = 0$	0.432 0	0.030 0	0.018 4
$\varphi_1(t,G(t)) = t/[1-G(t)] - \int_0^t s\mathrm{d}G(s)$, $\varphi_2(t,G(t)) = t[1-G(t)] - \int_0^t s\mathrm{d}G(s)$	0.431 8	0.029 8	0.018 4
$\varphi_1(t,G(t)) = \varphi_2(t,G(t)) = \frac{t}{1-G(t)} - \int_0^t \frac{s\mathrm{d}G(s)}{(1-G(s))^2}$	0.384 9	0.017 1	0.016 8

表 2 列出了本文方法在不同样本下的估计效果,可以看出随着样本数的增大,估计量的标准差显著减小,这符合相合估计的特点.

表 2 在不同样本下 $E[T^{(L)}]$ 估计效果的比较

样本数 n	估计量	偏度	标准差
100	0.356 4	0.045 6	0.055 6
200	0.360 1	0.041 9	0.042 4
500	0.364 2	0.037 8	0.031 6
1 000	0.368 6	0.033 4	0.024 9

注: $\varphi_1(t,G(t)) = t/[1-G(t)]$,$\varphi_2(t,G(t)) = 0$,$C$ 服从 $U(0,0.8)$,删失比率为 0.540.

表 3 列出了不同删失比率对估计的影响. 选取删失时间 C 服从分布分别为 $U(0,1.2)$,$U(0,1.0)$,$U(0,0.6)$,对应的删失比率分别为 0.360,0.468,0.684,从模拟计算的结果可以看出,随着删失比率的增大估计的偏度和标准差都增大,这就是删失造成“信息损失”所带来的后果.

表 3 在不同删失比率下 $E[T^{(L)}]$ 估计效果的比较

删失比率	估计量	偏度	标准差
0	0.431 9	0.029 9	0.014 5
0.360	0.432 5	0.030 5	0.017 8
0.468	0.431 5	0.029 5	0.019 1
0.684	0.254 5	0.147 5	0.049 7

4 讨论

生存函数的乘积极限估计有很好的大样本性质,但是它的尾部效果并不好. 无偏转换方法构造的寿命数据的均值估计,通常都具有强相合性和渐近正态性,在删失变量分布已知的条件下,收敛速度甚至能达到 $(n^{-1}\log\log n)^{-1/2}$,但是当寿命变量的分布尾部比删失变量的分布尾部厚很多时,尾部过度的信息删失会使估计量的效果很差,甚至会出现估计量方差为无穷大的情况. 在限制平均寿命的估计问题中,因为不用考虑在尾部问题上的局限性,这 2 种方法都能发挥很好的作用.

利用无偏转换方法进行限制平均寿命的估计,得到的是 1 个估计类. 模拟结果显示这些估计在偏度和标准差上还是有明显差别的. 在今后的研究中,可以考虑在这个估计类中寻找某种意义下最优的估计量.

5 参考文献

[1] Karrison T. Restricted mean life with adjustment for cova-

riates [J]. Journal of the American Statistical Association, 1987,82(400):1169-1176.

[2] Zucker D M. Restricted mean life with covariates: Modification and extension of a useful survival analysis method [J]. Journal of the American Statistical Association, 1998,93(442):702-709.

[3] Chen Peiyun, Tsiatis A A. Causel inference on the difference of the restricted mean lifetime between two groups [J]. Biometrics,2001,57(4):1030-1038.

[4] Zhang Min, Schaubel D E. Estimating differences in restricted mean lifetime using observational data subject to dependent censoring [J]. Biometrics,2011,67(3):740-749.

[5] Zhang Min, Schaubel D E. Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies [J]. Biometrics, 2012, 68(4): 999-1009.

[6] Pan Qing, Gastwirth J L. Estimating restricted mean job tenures in semi-competing risk data compensating victims of discrimination [J]. Annals of Applied Statistics,2013, 7(3):1474-1496.

[7] Royston P. Estimating the treatment effect in a clinical trial using difference in restricted mean survival time [J]. Stata Journal,2015,15(4):1098-1117.

[8] Royston P, Parmar M K B. Augmenting the log rank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated [J]. B M C Medical Research Methodology,1981,16(1):1-13.

[9] Koul H, Susarla V, Van R J. Regression analysis with randomly right-censored data [J]. Ann Statist,1981,9(6): 1276-1288.

[10] Zheng Zukang. A class of estimators of the mean survival time from interval censored data with application to linear regression [J]. Appl Math J Chinese Univ,2008,23(4): 377-390.

[11] Deng Wenli, Tian Yong, Lü Qiuping. Parametric estimator of linear model with interval-censored data [J]. Communications in Statistics-Simulation and Computation,2012,41 (10):1794-1804.

[12] Deng Wenli, Zheng Zukang, Zhang Riquan. Nonparametric regression with interval-censored data [J]. Acta Mathematica Sinica: English Series,2014,30(8):1422-1434.

[13] Kaplan E L, Meier P. Non-parametric estimation from incomplete observations [J]. J Am Stat Assoc, 1958, 53 (282):457-481.

[14] Peterson A V. Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions [J]. Journal of the American Statistical Association,1977,72(360):854-858.

The Estimation of the Restricted Mean Lifetime with Right Censored Data

DENG Wenli, OUYANG Fei

(College of Mathematics and Informatics, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: A group of "new" random variables were constructed with right censored data. If the distribution of the censoring variable is known, the sample means are the unbiased estimations of the restricted mean lifetime. If the distribution of the censoring variable is unknown, it can be proved that the sample means are consistent estimations of the restricted mean lifetime. The effectiveness and feasibility of the estimators can also be seen in the simulation studies.

Key words: restricted mean lifetime; product limit estimator; consistent estimator

(责任编辑: 曾剑锋)