长度偏差右删失数据下均值剩余寿命的复合估计

马秋霞^{1,2}, 潘生¹, 周勇^{3*}

1. 上海财经大学统计管理学院, 上海 200433;

2. 上海立信会计金融学院统计与数学学院, 上海 201209;

3. 华东师范大学经管学部统计交叉科学研究院, 上海 200241

E-mail: maqiuxia2012@163.com, pansheng201601@163.com, yzhou@amss.ac.cn

收稿日期: 2018-04-24; 接受日期: 2019-02-25; 网络出版日期: 2019-04-25; * 通信作者

摘要 长度偏差右删失数据是一类复杂的数据, 观察到的数据分布与总体分布有所改变且其删失是有信息删失, 通常的统计分析方法并不能直接应用到长度偏差数据中. 本文将在长度偏差右删失数据下研究均值剩余寿命函数, 提出其非参数估计方法, 在估计中通过加入长度偏差右删失数据辅助信息, 即截断变量和进入试验后的剩余存活时间同分布的辅助信息来提高估计的效率. 虽然极大似然方法是有效估计, 但是其构造复杂且计算需要迭代来实现, 计算量大. 为此, 本文考虑通过简单的加入辅助信息的方法来构造估计量, 并给出估计量的相合性及渐近正态性. 本文提出的加入辅助信息估计方法与以往类似方法相比具有较简单的显式表达式, 计算方便.

关键词 长度偏差 右删失 均值剩余寿命模型 非参数估计 辅助信息 置信区

MSC (2010) 主题分类 62N02

1 引言

在医学统计研究中, 生存分析是一个非常重要的工具, 目前已有许多统计方法和模型可以用来拟合生存曲线数据, 如 Kaplan-Meier 估计^[1]和 Cox 比例风险模型^[2]等, 但是风险函数和生存函数并不是一个直接刻画剩余寿命很好的选择. 那些模型和方法更多地关注个体存活概率或瞬时死亡的概率, 对于非统计专业人士来说很难简单地理解其内容. 注意到平均剩余寿命是一个可以直接刻画剩余寿命的方法. 剩余寿命, 定义为一个个体在 t 时刻存活时, 还能继续生存的剩余时间, 即 $T - t | T > t$, 其中 T 为生存时间. 它具有更多的信息量和更大的灵活性, 因为它是时变的, 有助于研究一些随时间变化的疾病. 在医学临床试验中, 剩余寿命通常是患者们最感兴趣的量, 尤其那些患慢性病或不可治愈疾病 (如癌症) 的患者, 想知道某种治疗是否可以延长其寿命. 而医生也需要知道该病到底是如何发展的, 所采取的治疗方案是否有效, 以及是否用新的治疗方法就能延长患者的寿命. 或者患者和医生都想知道某种疾病是否值得进行治疗, 那么这些都需要对患者的剩余寿命进行研究. 此外, 剩余寿命比

英文引用格式: Ma Q X, Pan S, Zhou Y. Composite estimator of mean residual lifetime with length-biased and right-censored data (in Chinese). Sci Sin Math, 2019, 49: 781–798, doi: 10.1360/N012018-00114

风险函数更易于解释主要体现剩余寿命均值意义, 例如, 我们可以这样解释: 一种新的治疗方法可以有效延长患者的剩余寿命超过 10 年, 而不是解释为减少患者 10% 死亡的危险. 但是, 许多经典的统计方法在研究剩余寿命问题中是无效的, 因为它们大多依赖于疾病起源的时间数据, 如 Kaplan-Meier 估计和 Cox 比例风险模型.

在关于生存时间的观测性研究中, 为了减少采集样本的成本和时间, 人们通常采用有效的流行队列研究. 这种抽样设计只是从具备某种条件的个体 (例如, 得了某种疾病的人) 中抽取样本, 可以根据其他的记录得到初始事件发生的时间, 然后进行跟踪观察, 这时可能只需要短时间内就能观测到失效事件. 但是在筛选样本之前已经发生死亡的个体并不能被观测到, 所以得到的数据是左截断的, 而且样本的长度越长 (生存时间越长), 被观测到的概率也就越大, 实际观测到的生存时间其实是比真正总体中的生存时间要长的, 这样我们得到的观测样本是有偏样本. 另外, 在对所选的样本进行跟踪观察的过程中, 有一些个体可能因为某些原因退出试验或者到试验截止并未发生死亡或感兴趣的事件, 这使得观测到的数据是右删失的. 正是由于这种特殊的抽样方式, 使得得到的样本具有特殊的数据结构. 特别地, 当事件的初始发生过程是一个平稳 Poisson 过程, 也就是截断变量服从均匀分布时, 我们称之为平稳性假设^[3], 这种数据为长度偏差数据 (见图 1). 由于观测的失效时间和删失时间拥有共同的因子, 即截断变量, 因此, 它们是相关的, 从而这类数据的删失是有信息删失. 其次在有偏抽样下, 对无偏总体假设的模型结构通常会发生改变, 即观测的样本不再服从原来的模型结构. 如果不考虑平稳性假设, 长度偏差数据就是左截断右删失数据. 如果直接忽略这种观测数据的偏差进行分析, 则会造成误导性的结果. 因此, 对于传统生存数据的分析方法都不再适用的情形, 有必要提出新的方法来分析长度偏差数据, 本文将研究长度偏差数据下均值剩余寿命函数的非参数统计推断.

剩余寿命越来越受到人们的关注, 目前主要有两种测量的方式: 均值剩余寿命函数和剩余寿命中位数函数, 分别由文献 [4, 5] 提出. 这两种函数都已有很多方法研究过, 涉及参数、非参数和半参数方法, 以及包不包含自变量等^[6, 7]. 在删失数据下, 文献 [8] 提出了一类灵活的回归模型来估计均值剩余寿命, 包括文献 [6] 等研究的模型. 以上这些研究大多数都是基于完全观察数据或是右删失数据下对均值剩余寿命进行统计建模或分析. 然而对于长度偏差数据删失数据下的研究还不多见, 最近, 一个重要的工作是文献 [9] 在长度偏差数据下研究了均值剩余寿命半参数模型, 他们的方法主要基于估计方程的方法. 事实上, 长度偏差数据在平稳性假设下具有更多的信息, 在没有删失情形下, 可以发现长度偏差数据中的截断变量 A 和剩余寿命 V (见图 1) 具有相同的分布. 如果能充分利用这些信息, 则可以改进对均值剩余寿命函数的估计效率^[10]. 因此, 本文通过构造不同的分布函数估计方法, 充分利用长度偏差数据的这个特征, 给出均值剩余寿命函数简单且相对有效的估计.

本文余下内容组织如下: 第 2 节介绍长度偏差和右删失数据的结构, 并且介绍均值剩余寿命函数; 第 3 节介绍均值剩余寿命模型的估计方法及渐近性质; 第 4 节给出模拟研究, 比较三种估计值, 可以

图 1 长度偏差右删失数据

看出, 我们给出的估计值形式更简单, 有很好的显式表达式, 在效率相当的情形下, 我们的计算速度更快; 第 5 节给出实际数据分析, 对 Channing House 数据集分别用三种方法进行分析; 第 6 节对本文进行总结, 并给出结论; 附录 A 对本文给出的定理进行详细证明.

2 定义和模型

2.1 长度偏差右删失数据

假设 $\tilde{T}$ 是从发病到事件终止的时间, 如果终止事件是死亡, 那么 $\tilde{T}$ 指病后存活时间, 这是我们感兴趣的无偏生存时间; $\tilde{O}$ 指发病的时间. 令 ξ 表示试验开始时间, 则 $\tilde{A} = \xi - \tilde{O}$ 表示从发病到试验开始的时间段. 假定 $\tilde{T}$ 与 $\tilde{O}$ 独立, 试验开始时间 ξ 独立于 $(\tilde{O}, \tilde{T})$, 则 $\tilde{A}$ 与 $\tilde{T}$ 也独立. 此外, 假定在观测时间段内, 该疾病的发病率保持不变 (该假定对慢性病来说是合理的), 从而 $\tilde{A}$ 在给定的区间内服从均匀分布, 此情形也被文献 [11] 称为“平稳性假设”, 是产生长度偏差数据的重要假设.

通常在流行队列样本中, 只有在检验时刻仍然存活的患病个体, 即满足 $\tilde{T} \geq \tilde{A}$ 的个体, 才可被招收进入试验. 此情形下, $\tilde{T}$ 被 $\tilde{A}$ 左截断. 令观测到的变量 (O, A, T) 分别表示该样本发病时间、左截断时间和病后存活时间变量, 则 (A, T) 的联合分布与 $(\tilde{A}, \tilde{T}) | \tilde{T} \geq \tilde{A}$ 的条件联合分布相同. 记 $A = \xi - O$, $V = T - A$, 则流行队列样本中 A 、 V 和 T 的生存分布函数分别由 $S_A(t)$ 、 $S_V(t)$ 和 $S_T(t)$ 表示, 边际密度函数分别由 $f_A(t)$ 、 $f_V(t)$ 和 $f_T(t)$ 表示.

令 f_T 表示长度偏差变量 T 的密度函数, f 表示无偏数据 $\tilde{T}$ 的密度函数, 它们关系如下^[12]:

$$f_T(t) = \frac{tf(t)}{\mu}, \quad \mu = \int_0^\infty uf(u)du, \quad \mu < \infty. \quad (2.1)$$

长度偏差删失数据的统计分析存在着两个挑战, 主要表现为, 由于观测的失效时间和删失时间拥有共同的因子, 从而删失是有信息删失, 见图 1, 生存时间 $T = A + V$ 和删失时间 $A + C$ 有共同因子 A . 其次在有偏抽样下, 与无偏总体假设的模型结构相比, 现实观察样本所服从的模型发生了改变, 即观测的样本不再服从原来的模型结构, 参见 (2.1), 总体分布 $f(t)$ 改变为观察样本的分布 $f_T(t)$. 如果直接忽略这种观测数据的偏差进行分析会导致错误的结果. 另外, 截断变量与无删失的剩余寿命具有相同分布^[13], 尽管它们并不一定独立, 但在统计推断中提供了更多的辅助信息. 因此, 我们提出新的方法来分析长度偏差数据下均值剩余寿命函数.

若 $f(t)$ 和 $S(t)$ 分别代表 $\tilde{T}$ 的边际密度函数和生存函数, 文献 [13] 给出长度偏差样本 (A, T) 有如下的联合密度函数:

$$f_{A,T}(a, t) = \frac{f(t)}{\mu} I(t \geq a > 0), \quad (2.2)$$

其中 $\mu = \int_0^\infty uf(u)du$. 因此, 在长度偏差抽样下, T 、 A 和 V 的边际密度函数很容易被获得:

$$f_T(t) = \frac{tf(t)}{\mu} I(t > 0), \quad (2.3)$$

$$f_A(t) = f_V(t) = \frac{S(t)}{\mu} I(t > 0). \quad (2.4)$$

长度偏差数据虽然存在结构上的复杂性, 但它能提供更多的辅助信息. 从 (2.4) 可见, 在长度偏差抽样下, A 和 V 有相同的边际分布函数, 在我们的估计方法中可利用该同分布的信息提高估计效

率. 此外, 在实际观测中, 可能会因为试验中途终止或个体提前离开试验而造成个体生存时间数据的右删失. 令 C 代表从试验开始到删失发生的时间, 并假定 C 与 (A, V) 独立. 然而, 由于删失变量 $A + C$ 和生存时间变量 $T = A + V$ 含有共同的左截断变量 A , 这里的删失是有信息删失, 因此, 很多经典的统计方法在有信息删失的情形下已失效, 需要发展新的方法来对此问题进行研究. 另外, 令 $Y = \min(T, A + C)$, $\tilde{V} = \min(V, C)$. 因此, 对于 n 个独立随机样本, 观测的数据集具有独立同分布样本 $\{(O_i, A_i, Y_i, \Delta_i), i = 1, \dots, n\}$, 其中 $Y_i = \min(T_i, A_i + C_i) = \min(A_i + V_i, A_i + C_i) = A_i + \tilde{V}_i$, $\Delta_i = I(V_i \leq C_i)$.

$\tilde{T}$ 表示无偏生存时间, 那么 t 时刻均值剩余寿命函数定义为

$$m(t) = E(\tilde{T} - t | T > t) = \frac{\int_t^\infty S(u) du}{S(t)} I(S(t) > 0). \quad (2.5)$$

本文在长度偏差右删失数据下研究均值剩余寿命非参数估计. 在平稳条件下充分利用长度偏差右删失数据的辅助信息, 以便于提高估计的准确性和有效性. 由于技术原因, 我们仅考虑在 $(0, \tau)$ 上的估计, 其中 $0 < \tau < \sup\{t : S(t)G(t) > 0\}$.

由均值剩余寿命函数的定义可知, 均值剩余寿命函数与生存函数 $S(t)$ 是一一对应的. 因此, 寻找生存函数 $S(t)$ 合适估计, 是对均值剩余寿命 $m(t)$ 进行估计的一个重要步骤. 在长度偏差右删失数据下, 结合 $S(t)$ 已有的几个估计方法, 在此基础上, 给出一个更好且简单的估计, 并对这些估计进行比较. 一个最为普通的方法是在长度偏差删失数据下给出 $S(t)$ 的非参数极大似然估计, 从而由极大似然估计的不变性可获得 $m(t)$ 的极大似然估计.

3 估计方法和渐近性质

3.1 非参数极大似然估计

在流行队列样本下, 观察到的长度偏差右删失数据可表示为 $\{(O_i, A_i, Y_i, \Delta_i), i = 1, \dots, n\}$, $\tilde{V}_i = Y_i - A_i = \min(V_i, C_i)$. 根据 (2.1)–(2.4), 全似然函数可推出如下形式:

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^n \frac{f(Y_i)^{\Delta_i} S(Y_i)^{1-\Delta_i}}{\mu} \\ &= \left\{ \prod_{i=1}^n \frac{f(Y_i)^{\Delta_i} S(Y_i)^{1-\Delta_i}}{S(A_i)} \right\} \left\{ \prod_{i=1}^n \frac{S(A_i)}{\mu} \right\} = \mathcal{L}_C \times \mathcal{L}_M. \end{aligned} \quad (3.1)$$

首先, 文献 [14] 通过最大化 (3.1) 中的似然函数 $\mathcal{L}$, 给出了 $S(t)$ 的一个极大似然估计, 即

$$\hat{S}_n^{\text{mle}}(t) = 1 - \frac{\int_0^t s^{-1} d\hat{G}_n(s)}{\int_0^\tau s^{-1} d\hat{G}_n(s)}, \quad (3.2)$$

其中 $\hat{G}_n(\cdot)$ 为极大化如下函数的非参数极大似然估计 [14]:

$$\prod_{i=1}^n dG(Y_i)^{\Delta_i} \left\{ \int_{s \geq Y_i} s^{-1} dG(s) \right\}^{1-\Delta_i}, \quad (3.3)$$

其中 $G(t) = \int_0^t f_T(u) du$. $\hat{G}_n(\cdot)$ 没有显式表达式, 但是可用文献 [15] 提出的一个 EM (expectation-maximization) 算法用于计算.

令

$$\begin{aligned} G^*(t) &= P(A + R \leq t \mid \delta = 1), \\ F^*(t) &= P(A + C \leq t \mid \delta = 0), \\ p &= P(\delta = 1) = P(R \leq C), \end{aligned}$$

令 f^* 和 g^* 分别表示 $F^*(\cdot)$ 和 $G^*(\cdot)$ 的密度函数, 即 $f^*(t) = (d/dt)F^*(t)$, $g^*(t) = (d/dt)G^*(t)$. 定义线性算子 $\mathcal{F}$:

$$\mathcal{F}(u)(t) = p \int_0^t \frac{g^*(x)}{g(x)} du(x) + (1-p) \int_0^t y \left(\int_{y \leq z} \frac{u(z)}{z^2} dz \right) d \left\{ \left(\frac{h(t)}{h(y)} - 1 \right) \frac{f^*(y)}{h(y)} \right\}, \quad (3.4)$$

其中 $h(t) = \int_t^\tau z^{-1} dG(z)$. 进一步, 定义 U 过程如下:

$$U(\cdot) = \mathcal{F}^{-1}(V)(\cdot). \quad (3.5)$$

引理 1 假定附录 A 中假设 1 成立, 定义

$$\mathcal{L}_y(x) = \frac{I_{[y, \infty)}(x) - S(y)}{x},$$

则在 $[0, \tau]$ 上, 有

$$\sqrt{n} \{ \hat{S}_n^{\text{mle}}(t) - S(t) \} \xrightarrow{\mathcal{W}} \mu \int_0^\tau \mathcal{L}_t(x) dU(x),$$

其中 $U(\cdot)$ 在 (3.5) 中定义. $\xrightarrow{\mathcal{W}}$ 表示弱收敛, 上式中右边过程的协方差函数为

$$r_S(y, z) = \mu^2 \int_0^\tau \int_0^\tau \psi(s, t) d\mathcal{L}_y(t) d\mathcal{L}_z(s),$$

其中

$$\psi(s, t) = \text{cov}(U(s), U(t)).$$

引理 1 的证明参见文献 [16, 定理 1] 的证明.

从上面的方法及其渐近结果可以看出, 极大似然估计形式相当复杂, 需要迭代算法, 计算量较大. 同时极大似然估计 $\hat{S}_n^{\text{mle}}(t)$ 的渐近方差是一个极其复杂的数学结构, 因此对其进行估计也是复杂的, 甚至其计算量相当大, 特别是其渐近方差没有简单形式可以进行直接估计. 下面基于极大似然估计, 可以获得均值剩余寿命函数的估计, 并推导其渐近性质. 记

$$\hat{m}_n^{\text{mle}}(t) = \frac{\int_t^\infty \hat{S}_n^{\text{mle}}(u) du}{\hat{S}_n^{\text{mle}}(t)} I(\hat{S}_n^{\text{mle}}(t) > 0). \quad (3.6)$$

定理 1 定义

$$\widetilde{\mathcal{L}}_t(x) = \frac{(m(t) + t)I(x > t)}{x} + I(x < t).$$

假定附录 A 中假设 1 成立, 则 $\sqrt{n}(\hat{m}_n^{\text{mle}}(t) - m(t))$ 在 $[0, \tau]$ 上弱收敛到一个 Guass 过程,

$$G^{\text{mle}}(t) = \frac{\mu}{S(t)} \int_0^\infty \widetilde{\mathcal{L}}_t(x) dx.$$

它具有零期望, 且协方差函数 $\bar{\Gamma}(s, t)$ 为

$$\bar{\Gamma}(s, t) = \frac{\mu}{S^2(t)} \int_0^\infty \int_0^\infty r_S(u, v) d\tilde{\mathcal{L}}_s(u) \tilde{\mathcal{L}}_t(v).$$

从上面的渐近理论可以看出, 此渐近分布卷入了复杂算子, 除了 Bootstrap 方法, 现实很难直接获得其渐近方差的估计.

3.2 基于修改乘积限估计均值剩余寿命估计方法

在把长度偏差数据仅看成为左截断右删失数据下, 即没有给定左截断时间 A_i 分布函数的情形下, 文献 [17] 证明了用过极大化 $\mathcal{L}_C$ 得到的极大似然估计是有效的. 文献 [18] 提出了一个非参数乘积极限估计 (TJW (Tsai-Jewell-Wang) 估计), 形式如下:

$$\hat{S}_n^{\text{TJW}}(t) = \prod_{u \in [0, t]} \left\{ 1 - \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n \tilde{R}_i(u)} \right\}, \quad (3.7)$$

其中 $N_i(t) = \Delta_i I(Y_i \leq t)$ 和 $\tilde{R}_i(t) = I(Y_i \geq t \geq A_i)$. 在左截断右删失数据下, 文献 [19] 等基于此估计 $\hat{S}_n^{\text{TJW}}(t)$ 给出均值剩余寿命的估计, 并推导出其渐近性质.

然而, 在长度偏差样本下, TJW 估计不是十分有效的, 因为它没有考虑 $\mathcal{L}_M$ 中 A 所包含的信息. 为了解决这个问题, 可以把长度偏差右删失数据中的辅助信息加入, 即 A 和 V 具有相同的分布, 参见 (2.4). 文献 [13] 提出了一个估计 $S_A(t)$ 的一个方法: 合并 A 和 V 量来构造新的生存分析函数估计. 首先利用数据 $(A_i, \Delta_i \equiv 1)$ 和 $(\tilde{V}_i, \Delta_i)$, 然后通过修改 TJW 估计 (3.7), 进而得到估计值 $\hat{S}_A(t)$ 如下:

$$\hat{S}_A(t) = \prod_{u \in [0, t]} \left\{ 1 - \frac{d\tilde{Q}(u)}{\tilde{P}(u)} \right\}, \quad (3.8)$$

其中 $\tilde{Q}(t) = n^{-1} \sum_{i=1}^n \{I(A_i \leq t) + \Delta_i I(\tilde{V}_i \leq t)\}$, $\tilde{P}(t) = n^{-1} \sum_{i=1}^n \{I(A_i \geq t) + I(\tilde{V}_i \geq t)\}$. $\hat{S}_A(t)$ 是 $S_A(t)$ 的相合估计, 它不同于仅仅利用 A 的边际生存函数得到的估计.

因此, 文献 [13] 给出 $S(t)$ 的估计时, 考虑了 A 的辅助信息且是无删失的, 且 A 和 V 具有相同的分布. 在修改 TJW 乘积估计 (3.7) 时, 考虑到

$$\frac{1}{n} \sum_{i=1}^n \tilde{R}_i(t) = \frac{1}{n} \sum_{i=1}^n I(Y_i \geq t \geq A_i) = \frac{1}{n} \sum_{i=1}^n I(Y_i \geq t) - \frac{1}{n} \sum_{i=1}^n I(A_i > t),$$

从而可得 $S(t)$ 的估计值 $\hat{S}_n^{\text{HQ}}(t)$ 如下:

$$\hat{S}_n^{\text{HQ}}(t) = \prod_{u \in [0, t]} \left\{ 1 - \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n I(Y_i \geq u) - n\hat{S}_A(u)} \right\}. \quad (3.9)$$

根据文献 [13] 可知, $\sqrt{n}(\hat{S}_n^{\text{HQ}}(t) - S(t))$ 弱收敛到一个 Guass 过程, 其均值为 0, 协方差阵为

$$\sigma_S(t_1, t_2) = S(t_1)S(t_2)E\{\psi_1(t_1)\psi_1(t_2)\},$$

其中

$$\psi_1(t) = \int_0^t K(u)^{-2} \{I(A_i \leq u \leq Y_i)\} dW(u) - \frac{\Delta_i I(Y_i \leq t)}{K(Y_i)}$$

$$+ \int_0^t K(u)^{-2} \{I(A_i \geq u) - S_A(u) - S_A \phi_i(u)\} dW(u),$$

且 $W(t) = P(Y \leq t, \Delta = 1)$, $S_A(t)$ 是生存函数 A 的边际分布函数.

基于 $\hat{S}_n^{\text{HQ}}(t)$, 可以获得均值剩余寿命函数的一个估计

$$\hat{m}_n^{\text{HQ}}(t) = \frac{\int_t^\infty \hat{S}_n^{\text{HQ}}(u) du}{\hat{S}_n^{\text{HQ}}(t)} I(\hat{S}_n^{\text{HQ}}(t) > 0). \quad (3.10)$$

记 $M(t) = \int_t^\tau S(u) du$. 对均值剩余寿命函数估计 $\hat{m}_n^{\text{HQ}}(t)$, 可以证明如下的结论.

定理 2 假定附录 A 中假设 1 成立, 则 $\sqrt{n}(\hat{m}_n^{\text{HQ}}(t) - m(t))$ 在 $[0, \tau]$ 上弱收敛到 Guass 过程

$$G^{\text{HQ}}(t) = \frac{1}{S^2(t)} \left[S(t) \int_t^\tau Z^{\text{HQ}}(u) du - Z^{\text{HQ}}(t) \int_t^\tau S(v) dv \right],$$

且具有零期望和协方差函数 $\Gamma^{\text{HQ}}(s, t)$, 其中 $Z^{\text{HQ}}(t)$ 是 $\sqrt{n}(\hat{S}_n^{\text{HQ}}(t) - S(t))$ 的渐近 Guass 过程, $\Gamma^{\text{HQ}}(s, t)$ 表示如下:

$$\begin{aligned} \Gamma^{\text{HQ}}(s, t) = & \frac{1}{S^2(s)S^2(t)} \left[S(s)S(t) \int_s^\tau \int_t^\tau S(u)S(v) E\{\psi_1(u)\psi_1(v)\} dudv \right. \\ & + S(s)S(t) E\{\psi_1(s)\psi_1(t)\} \int_s^\tau S(u) du \int_t^\tau S(u) du \\ & - S(s) \left(\int_t^\tau S(u) du \right) \left(\int_s^\tau S(t)S(u) E\{\psi_1(t)\psi_1(u)\} du \right) \\ & \left. - S(t) \int_s^\tau S(u) du \left(\int_t^\tau S(s)S(u) E\{\psi_1(s)\psi_1(u)\} du \right) \right]. \end{aligned}$$

Lancaster^[13] 证明了生存分布 $\hat{S}_n^{\text{HQ}}(u)$ 估计是相合估计, 而且通过模拟计算发现此估计与极大似然估计的效率相近, 同时计算比极大似然估计要简单且计算容易很多. 但是, 他的估计仍有不足, 在估计 (3.9) 中, 其分母是 plug-in 估计量, 可能会使分母产生负数, 从而导致此估计在某个点上发生的概率大于 1, 造成生存分布的错误估计, 因此也会导致均值剩余寿命函数错误估计. 为此, 我们考虑一个可以方便加入长度偏差数据辅助信息, 且形式简单易于计算, 并同著名的 Kaplan-Meier 估计有相似的形式.

3.3 加入辅助信息的均值剩余寿命估计

从 (2.4) 可以看出, 当数据没有删失时, 随机变量 A 和 $V = T - A$ 具有相同的边际密度函数. 也就是 A 和 V 有相同的边际分布. 这是长度偏差数据特有的额外信息, 一个简单有效利用此额外信息的方法, 就是采用文献 [20] 提出的组合似然方法, 这是因为给定 A 下 V 的条件密度函数与给定 V 下 A 的分布是相同的, 也称具有可交换性. 然而, 数据存在删失时, A 和 $\tilde{V} = \min(T - A, C)$ 的联合分布因为右删失的出现不再可交换. 在给定观测的生存时间未删失时, 给定 A 下 $\tilde{V}$ 的分布和给定 $\tilde{V}$ 下 A 的分布是相同的. 因此可利用这个性质通过用复合条件似然函数的方法来获得生存分布的估计. 为了便于理解, 我们给出长度偏差右删失数据下非参数似然估计. 设 $m = \sum_{i=1}^n \delta_i$ 表示未被删失变量的个数. 不妨设前 m 个观察是未被删失的数据, 后面的 $n - m$ 个观察是删失数据, 故组合条件似然函数是

$$\prod_{i=1}^n \left\{ \frac{f(Y_i)}{S(A_i)} \frac{f(Y_i | \mathbf{X}_i)}{S(\tilde{V}_i)} \right\} \times \prod_{i=m+1}^n \frac{S(Y_i)}{S(A_i)} = \prod_{i=1}^m \frac{\{\lambda(Y_i)\}^{2\delta_i} \exp\{-(1 + \Delta_i)\lambda(Y_i)\}}{\exp\{-(\Lambda(A_i) + \delta_i \Lambda(\tilde{V}_i))\}}. \quad (3.11)$$

对于半参数模型, 文献 [21] 在长度偏差数据下扩展了文献 [20] 的组合似然方法到 Cox 模型. 极大化 (3.11) 可以获得如下累积风险函数的估计:

$$\hat{\Lambda}_0(t) = \sum_{i=1}^n \frac{2\delta_i I(Y_i \leq t)}{\sum_{j=1}^n [I(A_j \leq Y_i \leq Y_j) + \delta_j I(\tilde{V}_j \leq Y_i \leq Y_j)]} = \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n R_i(u)}, \quad (3.12)$$

其中

$$R_i(t) = 2^{-1} \{I(A_i \leq t \leq Y_i) + \Delta_i I(\tilde{V}_i \leq t \leq Y_i)\}.$$

由此估计可获得生存分布估计记为 $\hat{S}_n^{\text{new}}$:

$$\hat{S}_n^{\text{new}}(t) = \prod_{u \in [0, t]} \{1 - d\hat{\Lambda}(u)\} = \prod_{u \in [0, t]} \left\{1 - \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n R_i(u)}\right\}, \quad (3.13)$$

很明显, 这样得到的估计形式比 $\hat{S}_n^{\text{HQ}}(t)$ 估计和极大似然估计的形式简单, 且容易计算, 比 TJW 估计更有效, 形式与著名的 Kapla-Meier 估计相同. 在附录 A 里, 可见 $n^{1/2}\{\hat{S}_n^{\text{new}}(t) - S(t)\}$ 在 $0 < t < \tau$ 时收敛到一个零期望的 Guass 过程, 其方差 - 协方差函数如下:

$$\tilde{\sigma}_S(t_1, t_2) = S(t_1)S(t_2)E\{\phi_1(t_1)\phi_1(t_2)\},$$

其中

$$\phi_i(t) = \frac{1}{2} \int_0^t K(u)^{-2} \{I(A_i \leq u \leq Y_i) + \Delta_i I(\tilde{V}_i \leq u \leq Y_i)\} dW(u) - \frac{\Delta_i I(Y_i \leq t)}{K(Y_i)}, \quad (3.14)$$

这里

$$K(t) = \frac{1}{2} E\{I(A \leq t \leq Y) + \Delta I(\tilde{V} \leq t \leq Y)\}, \quad W(t) = E\{\Delta I(Y \leq t)\}.$$

基于生存分布估计 $\hat{S}_n^{\text{new}}(t)$ 可以获得均值剩余寿命函数的估计

$$\hat{m}_n^{\text{new}}(t) = \frac{\int_t^\infty \hat{S}_n^{\text{new}}(u) du}{\hat{S}_n^{\text{new}}(t)} I(\hat{S}_n^{\text{new}}(t) > 0).$$

记 H 为

$$\left\{ \frac{1}{2} S(t) \left[\int_0^t K(u)^{-2} \{I(a \leq u \leq y) + I(v \leq c) I(a \wedge c \leq u \leq y)\} dW(u) - \frac{I(a \wedge c \leq u \leq y) I(y \leq t)}{K(y)} \right] : t \in [0, \tau] \right\}$$

的封套函数 (参见文献 [22] 的 “envelope function” 定义).

定理 3 在假设 1 (见附录 A) 下, 若 H 平方可积, 则 $\sqrt{n}(\hat{m}_n^{\text{new}}(t) - m(t))$ 弱收敛到一个 Guass 过程,

$$G^{\text{new}}(t) = \frac{1}{S^2(t)} \left[S(t) \int_t^\tau \tilde{Z}(u) du - \tilde{Z}(t) \int_t^\tau S(v) dv \right]$$

具有零期望, 协方差函数为 $\Gamma^{\text{new}}(s, t)$, 其中 $\tilde{Z}(t)$ 是 $\sqrt{n}(\hat{S}_n^{\text{new}}(t) - S(t))$ 的弱收敛极限, $\Gamma^{\text{new}}(s, t)$ 具有如下形式:

$$\Gamma^{\text{new}}(s, t) = \frac{1}{S^2(s)S^2(t)} \left[S(s)S(t) \int_s^\tau \int_t^\tau S(u)S(v) E\{\phi_1(u)\phi_1(v)\} dudv \right]$$

$$\begin{aligned}
& + S(s)S(t)E\{\phi_1(s)\phi_1(t)\} \int_s^\tau S(u)du \int_t^\tau S(u)du \\
& - S(s) \left(\int_t^\tau S(u)du \right) \left(\int_s^\tau S(t)S(u)E\{\phi_1(t)\phi_1(u)\}du \right) \\
& - S(t) \int_s^\tau S(u)du \left(\int_t^\tau S(s)S(u)E\{\phi_1(s)\phi_1(u)\}du \right),
\end{aligned}$$

这里 $\phi_i(t)$ 的定义见 (3.14).

对于此方差一个相合且方便的估计是使用插入法, 即对 $\Gamma^{\text{new}}(s, t)$ 中未知参数代入其相合估计, 即

$$\begin{aligned}
\hat{\Gamma}^{\text{new}}(s, t) = & \frac{1}{(\hat{S}_n^{\text{new}})^2(s)(\hat{S}_n^{\text{new}})^2(t)} \left[\hat{S}_n^{\text{new}}(s)\hat{S}_n^{\text{new}}(t) \int_s^\tau \int_t^\tau \hat{S}_n^{\text{new}}(u)\hat{S}_n^{\text{new}}(v)\hat{E}\{\phi_1(u)\phi_1(v)\}dudv \right. \\
& + \hat{S}_n^{\text{new}}(s)\hat{S}_n^{\text{new}}(t)\hat{E}\{\psi_1(s)\psi_1(t)\} \int_s^\tau \hat{S}_n^{\text{new}}(u)du \int_t^\tau \hat{S}_n^{\text{new}}(u)du \\
& - \hat{S}_n^{\text{new}}(s) \left(\int_t^\tau \hat{S}_n^{\text{new}}(u)du \right) \left(\int_s^\tau \hat{S}_n^{\text{new}}(t)\hat{S}_n^{\text{new}}(u)\hat{E}\{\phi_1(t)\phi_1(u)\}du \right) \\
& \left. - \hat{S}_n^{\text{new}}(t) \int_s^\tau \hat{S}_n^{\text{new}}(u)du \left(\int_t^\tau \hat{S}_n^{\text{new}}(s)\hat{S}_n^{\text{new}}(u)\hat{E}\{\phi_1(s)\phi_1(u)\}du \right) \right],
\end{aligned}$$

其中

$$\hat{E}\{\phi_1(s)\phi_1(u)\} = \frac{1}{n} \sum_{i=1}^n \hat{\phi}_i(s)\hat{\phi}_i(u)$$

且

$$\begin{aligned}
\hat{\phi}_i(t) &= \frac{1}{2n} \sum_{j=1}^n \frac{\Delta_j \{I(A_i \leq Y_j \leq Y_i) + \Delta_i I(\tilde{V}_i \leq Y_j \leq Y_i)\}}{\hat{K}^2(Y_j)} - \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i I(Y_i \leq t)}{\hat{K}(Y_i)}, \\
\hat{K}(t) &= \frac{1}{2n} \sum_{i=1}^n \{I(A_i \leq t \leq Y_i) + \Delta_i I(\tilde{V}_i \leq t \leq Y_i)\}.
\end{aligned}$$

$\hat{E}\{\psi_1(s)\psi_1(t)\} = \frac{1}{n} \sum_{i=1}^n \hat{\psi}_i(s)\hat{\psi}_i(t)$ 为 $E\{\psi_1(s)\psi_1(t)\}$ 的经验版本. 由于 $\hat{S}_n^{\text{new}}(t)$ 的相合性, 可以获得渐近协方差的相合估计, 即 $\hat{\Gamma}^{\text{new}}(s, t)$ 依概率收敛于 $\Gamma^{\text{new}}(s, t)$.

4 模拟研究

本节通过模拟来计算我们的估计量有限样本性质, 同时与本文中提到的其他两个估计量进行比较. 按照长度偏差数据的结构, 左截断变量为均匀分布, 我们按如下过程生成长度偏差数据: 选择 A 为在 $[0, 100]$ 上的均匀分布, $\tilde{T}$ 为参数为 $(2, 2)$ 的 Weibull 分布. 继而真实均值剩余寿命函数为 $m(t) = e^{(-0.5t)^2}$. 重复生成 A 和 $\tilde{T}$ 至 $\tilde{T} + A$ 大于 100, 直到产生 50 个长度偏差数据后停止. 删失时间 C 按在区间 $[0, 5.5]$ 上的均匀分布来生成. 从而获得观察数据 (A_i, Y_i, δ_i) , $i = 1, 2, \dots, n$, 其中 $Y_i = \min(T_i, A_i + C_i)$, $\delta_i = I(T_i \leq A_i + C_i)$.

在模拟中, 我们重复 500 次, 每次生成了 50 个数据. 我们选择了在区间 $[0, 1.8]$ 间隔为 0.2 的 10 个点来计算剩余均值寿命估计量, 用 Bootstrap 方法来计算标准差 (记为 SE), Bootstrap 次数选择 50. 样本标准差记为 SD. 图 2 显示了三种方法基于 Bootstrap 估计方差的逐点的置信区间.

表 1 中的 “NEW” 代表本文提出的估计量, “HQ” 代表文献 [13] 提出的估计量, “MLE” 代表极大似然估计. 从表 1 可以看到, 当 t 取不同值时, 利用 E-M 算法得到的极大似然量的方差比我们给出

的估计值得到的方差要略小, 符合人们的预期, 极大似然估计是最有效估计. 我们提出的估计的 SE 比 HQ 的 SE 要小. 但是极大似然估计是没有解析式的. 虽然文献 [13] 的估计量有解析式, 但是我们的估计量在形式上更简洁, 且从表 2 可以看出其计算时间远远比极大似然估计和 HQ 都要短. 同时从表 1 也可以看出极大似然估计的覆盖率是比较差的, 而其余两种方法基本表现很接近. 图 3 为三种估计量 500 次生成数据的估计值减去真实寿命均方误差根的盒状图, 可以看出极大似然估计量的盒装图最窄, 其他两种方法则相差不大. 从盒状图中可看到中位数均在一个水平上, 这也表明所提出的方法靠近极大似然估计.

图 2 三种方法的置信区间 (点划线代表本文给出的估计, 点线代表极大似然估计, 虚线代表文献 [13] 的估计量)

表 1 模拟数据下三种方法的表现

t	估计量	SD	SE	Bias	CP
0.4	NEW	0.1409	0.1404	0.0180	0.9370
	HQ	0.1388	0.1446	0.0138	0.9340
	MLE	0.1372	0.1274	0.0170	0.9040
0.6	NEW	0.1409	0.1404	0.0180	0.9370
	HQ	0.1388	0.1446	0.0138	0.9340
	MLE	0.1372	0.1274	0.0170	0.9040
0.8	NEW	0.1347	0.1319	0.0153	0.9470
	HQ	0.1334	0.1371	0.0109	0.9530
	MLE	0.1317	0.1194	0.0136	0.9280
$t = 1.0$	NEW	0.1269	0.1261	0.0159	0.9250
	HQ	0.1258	0.1328	0.0113	0.9340
	MLE	0.1238	0.1134	0.0139	0.9010

表 2 三种方法用 5,000 个样本在 10 个点处计算剩余寿命所消耗的时间

	NEW	HQ	MLE
计算时间 (秒)	84.27	235.66	195.99

图 3 三种方法在不同时间点处的表现盒装图

5 实际数据分析

本节将应用本文中的估计量到一个实际数据集来分析: Channing House 数据集. Channing House 是一个位于加利福尼亚 Palo Alto 的退休中心. 我们的目标是分析出 Channing House 的医疗程序对剩余寿命的影响. 这个数据集包括 97 名男性、365 名女性共 462 个个体. 数据集收集时间为 1965 年到 1975 年. 数据收集方法为招募居民住到 Channing House 直到死亡, 并支付包括除了物理治疗外的任何医疗需要. 那些在 1965 年前死亡的不能被招募, 因此可视为左截断. 进一步的研究表明, 如果选

表 3 Channing House 数据集下三种方法的表现

t	估计量	剩余寿命	SE
900	NEW	134.18056	18.74198
	HQ	152.92438	17.78285
	MLE	137.87304	11.35684
950	NEW	104.29870	25.50352
	HQ	119.59626	18.88793
	MLE	101.37126	10.85439
1,000	NEW	77.88734	30.25895
	HQ	89.83317	21.36786
	MLE	71.75796	12.67684
1,050	NEW	69.89393	56.42448
	HQ	74.81847	26.20001
	MLE	54.49691	14.02142

择一个年龄大于 786 个月的数据集作为子样本 (共 448 人, 包括 94 名男性和 354 名女性), 那么子样本是一个长度偏差的数据^[23]. 在这个数据集中, 观测到的时间是从出生到招募, 而剩余寿命是从招募到死亡. 在研究结束时, 大量的参与者仍然活着, 所以, 它们是右删失. 删失率约为 61.9%. 我们计算了在时间点 900、950、1,000 和 1,050 处三个不同估计量所给出的剩余寿命值 (见表 3), 可见在退休中心时间越久, 剩余寿命的估计值也越低, 与实际情形相符.

6 结论

本节在长度偏差右删失数据下, 首先基于均值剩余寿命的定义构建估计方程, 提出不含协变量的非参数估计方法, 并在估计中考虑了截断变量和进入试验后的剩余存活时间同分布的辅助信息, 来提高估计的效率. 最后得到的估计方法与基于极大似然估计 $\hat{S}^{\text{mle}}(t)$ 、文献 [13] 给出的估计 $\hat{S}_n^{\text{HQ}}(t)$ 进行比较, 很明显, 我们的估计方法比基于 HQ 估计方法更有效, 同时与极大似然估计相比效率损失也不大, 且我们提出估计有较简单的显式表达式, 容易计算并方便推出所提出估计的大样本性. 进一步还可以研究半参数下的剩余寿命均值模型、剩余寿命分位数半参数模型等, 虽然有一些工作研究过这两个模型^[7, 10], 但是在复杂的长度偏差右删失数据下仍然有许多待解决的问题.

参考文献

- 1 Kaplan E L, Meier P. Nonparametric estimation from incomplete observations. J Amer Statist Assoc, 1958, 53: 457–481
- 2 Cox D R. Regression models and life-tables. J R Stat Soc Ser B Stat Methodol, 1972, 34: 187–220
- 3 De Uña-álvarez J. Nonparametric estimation under length-biased sampling and type I censoring: A moment based approach. Ann Inst Statist Math, 2004, 56: 667–681
- 4 Chiang C L. A stochastic study of the lifetime table and its applications: I. Probability distributions of the biometric functions. Biometrics, 1960, 16: 618–635
- 5 Schmittlein D C, Morrison D G. The median residual lifetime: A characterization theorem and an application. Oper Res, 1981, 29: 392–399

- 6 Sun L, Song X, Zhang Z. Mean residual life models with time-dependent coefficients under right censoring. *Biometrika*, 2012, 99: 185–197
- 7 Wang Y X, Liu P, Zhou Y. Quantile residual lifetime for left-truncated and right-censored data. *Sci China Math*, 2015, 58: 1217–1234
- 8 Yang G, Zhou Y. Semiparametric varying-coefficient study of mean residual life models. *J Multivariate Anal*, 2014, 128: 226–238
- 9 Bai F F, Huang J, Zhou Y. Semiparametric inference of mean residual life model with right censored and length-biased data. *Statist Sinica*, 2016, 26: 1129–1158
- 10 Liu Y T, Zhang S C, Zhou Y. Semiparametric quantile-difference estimation for length-biased and right-censored data. *Sci China Math*, 2019, doi: 10.1007/s11425-017-9250-0
- 11 Wang M C, Jewell N P, Tsai W Y. Asymptotic properties of the product limit estimate under random truncation. *Ann Statist*, 1986, 14: 1597–1605
- 12 Lancaster T. *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press, 1990
- 13 Huang C Y, Qin J. Nonparametric estimation for length-biased and right-censored data. *Biometrika*, 2011, 98: 177–186
- 14 Asgharian M, Wolfson D B. Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. *Ann Statist*, 2005, 33: 2109–2131
- 15 Vardi Y. Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*, 1989, 76: 751–761
- 16 Asgharian M, M'Lan C E, Wolfson D B. Length-biased sampling with right censoring. *J Amer Statist Assoc*, 2002, 97: 201–209
- 17 Wang M C. Nonparametric estimation from cross-sectional survival data. *J Amer Statist Assoc*, 1991, 86: 130–143
- 18 Tsai W Y, Jewell N P, Wang M C. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 1987, 74: 883–886
- 19 Zhao M, Jiang H M, Liu X. A note on estimation of the mean residual life function with left-truncated and right-censored data. *Statist Probab Lett*, 2013, 83: 2332–2336
- 20 Arnold B C, Strauss D. Bivariate distributions with exponential conditionals. *J Amer Statist Assoc*, 1988, 83: 522–527
- 21 Huang C, Qin J. Composite partial likelihood estimation under length-biased sampling, with application to a prevalent cohort study of dementia. *J Amer Statist Assoc*, 2012, 107: 946–957
- 22 van der Vaart Wellner J A. *Weak Convergence and Empirical Processes*. New York: Springer, 1996
- 23 Chen X, Zhou Y. Quantile regression for right-censored and length-biased data. *Acta Math Appl Sin Engl Ser*, 2012, 28: 433–462
- 24 van der Vaart Aad W. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000

附录 A

下面给出前面主要定理的证明. 考虑如下假设总是成立的:

假设 1 $F(\cdot)$ 和 $G(\cdot)$ 是绝对连续分布函数, 且有有限支撑 $[0, \tau]$.

文献 [13, 14] 中的渐近结果都需要在假设 1 下才成立. 且可以看出, 在假设 1 下有限距条件总是满足的.

为了证明本文的结果, 首先考虑如下一个重要结论. 假设 $\hat{S}_n(t)$ 是生存函数的一个估计量, 令 $\hat{m}_n(t) = \frac{\int_t^\infty \hat{S}_n(u) du}{\hat{S}_n(t)} I(\hat{S}_n(t) > 0)$. 假设 $\tau < \infty$, $S(\tau) > 0$. 利用函数 Delta 方法, 我们能够推导出下面结论. 令 $\mathbb{D}[0, \tau]$ 表示一个在区间 $[0, \tau]$ 上具有右连续和左极限点的函数空间, 一致度量为其上的度量.

命题 1 若 $n^{1/2}(\hat{S}_n(t) - S(t))$ 在 $\mathbb{D}[0, \tau]$ 上弱收敛到一个 Guass 过程 $Z(t)$, 则 $\sqrt{n}(\hat{m}_n(t) - m(t))$ 弱收敛到一个 Guass 过程

$$G(t) = \frac{1}{S^2(t)} \left[S(t) \int_t^\tau Z(u) du - Z(t) \int_t^\tau S(v) dv \right].$$

它具有零期望, 且协方差函数 $\Gamma(s, t)$ 如下:

$$\begin{aligned}\Gamma(s, t) = & \frac{1}{S^2(s)S^2(t)} \left[S(s)S(t) \int_s^\tau \int_t^\tau E[Z(u)Z(v)]dudv \right. \\ & + E[Z(s)Z(t)] \int_s^\tau S(u)du \int_t^\tau S(u)du - S(s) \left(\int_t^\tau S(u)du \right) E \left(Z(t) \int_s^\tau Z(u)du \right) \\ & \left. - S(t) \int_s^\tau S(u)du E \left(Z(s) \int_t^\tau Z(u)du \right) \right].\end{aligned}$$

命题 1 的证明可参见文献 [19, 定理 2] 的证明. 为了使本文的主要结果证明完整, 在此我们给出此命题的证明梗概.

命题 1 的证明 令 $\mathbb{D}[0, \tau]$ 是一个具有右连续和左极限点的函数空间, 定义如下映射:

$$\phi: \mathbb{D} \mapsto \mathbb{D}, \quad \phi(\theta)(t) = \frac{\int_t^\tau \theta(u)du}{\theta(t)}, \quad \theta \in \mathbb{D}.$$

由文献 [24, 引理 20.10] 可知, 映射 $\theta(t) \rightarrow \int_t^\tau \theta(u)du$ 是 Hadamard 可微的. 同时我们可以观察到, 对任意收敛序列 $b_n \rightarrow 0$ 和任意的 $t \in [0, \tau]$, 若 $\sup_{t \in [0, \tau]} |h_n(t) - h(t)| \rightarrow 0$, 则有

$$\begin{aligned}& \lim_{n \rightarrow \infty} \sup_{t \in [0, \tau]} \left| \frac{\phi(\theta(t) + b_n h_n(t)) - \phi(\theta(t))}{b_n} - \left[\frac{\int_t^\infty h(u)du}{\theta(t)} - \frac{h(t) \int_t^\infty \theta(u)du}{\theta^2(t)} \right] \right| \\ &= \lim_{n \rightarrow \infty} \sup_{t \in [0, \tau]} \left| \frac{\theta(t) \int_t^\infty h_n(u)du}{\theta(t)(\theta(t) + b_n h_n(t))} - \frac{h_n(t) \int_t^\infty \theta(u)du}{\theta(t)(\theta(t) + b_n h_n(t))} \right. \\ & \quad \left. - \left[\frac{\int_t^\infty h(u)du}{\theta(t)} - \frac{h(t) \int_t^\infty \theta(u)du}{\theta^2(t)} \right] \right| \\ &= 0.\end{aligned}\tag{A.1}$$

由文献 [24, 第 296 页的定义] 可知, $\frac{\int_t^\infty h(u)du}{\theta(t)} - \frac{h(t) \int_t^\infty \theta(u)du}{\theta^2(t)}$ 为 ϕ 在 θ 处的 Hadamard 导数, 由函数 Delta 定理 (参见文献 [24, 定理 20.8]), 取 $\theta(t)$ 为生存函数, 函数 h 为 Guass 过程 Z , 则 $\sqrt{n}(\hat{m}_n(t) - m(t))$ 弱收敛到

$$B(t) = \frac{1}{S^2(t)} \left[S(t) \int_t^\tau Z(u)du - Z(t) \int_t^\tau S(v)dv \right].$$

由于 Guass 过程的连续线性变换是正态分布 (参见文献 [24, 例 20.11]), 而两个正态分布随机变量的线性组合仍是正态分布, 任意有限个正态分布随机变量联合分布是联合正态分布, 因此, $B(t)$ 是 Guass 过程. $\square$

引理 2 在假设 1 下, 当 $n \rightarrow \infty$ 时, 对于 $0 < t < \tau$, 随机过程 $n^{1/2}\{\hat{S}^{\text{HQ}}(t) - S(t)\}$ 满足

$$n^{1/2}\{\hat{S}^{\text{HQ}}(t) - S(t)\} = n^{-1/2} \sum_{i=1}^n S(t)\psi_i(t) + o_p(1)$$

一致成立, 且收敛于一个零期望的 Guass 过程, 其方差 - 协方差函数为

$$\sigma_S(t_1, t_2) = S(t_1)S(t_2)E\{\psi_1(t_1)\psi_1(t_2)\},$$

其中

$$\psi_i(t) = \int_0^t K(u)^{-2} \{I(A_i \leq u \leq Y_i)\} dW(u) - \frac{\Delta_i I(Y_i \leq t)}{K(Y_i)}$$

$$+ \int_0^t K(u)^{-2} \{I(A_i \geq t) - S_A(u) - S_A \phi_i(u)\} dW(u),$$

$W(t) = P(Y \leq t, \Delta = 1)$, $S_A(t)$ 是 A 的边际生存函数.

这个引理可以参见文献 [13, 定理 1] 的证明.

定理 2 的证明 由命题 1 和引理 3 知, $\sqrt{n}(\hat{m}_n^{\text{HQ}}(t) - m(t))$ 弱收敛到一个 Gauss 过程. 由 Fubini 定理知, 其协方差为

$$\begin{aligned} \Gamma^{\text{HQ}}(s, t) = & \frac{1}{S^2(s)S^2(t)} \left[S(s)S(t) \int_s^\tau \int_t^\tau S(u)S(v)E\{\psi_1(u)\psi_1(v)\}dudv \right. \\ & + S(s)S(t)E\{\psi_1(s)\psi_1(t)\} \int_s^\tau S(u)du \int_t^\tau S(u)du \\ & - S(s) \left(\int_t^\tau S(u)du \right) \left(\int_s^\tau S(t)S(u)E\{\psi_1(t)\psi_1(u)\}du \right) \\ & \left. - S(t) \int_s^\tau S(u)du \left(\int_t^\tau S(s)S(u)E\{\psi_1(s)\psi_1(u)\}du \right) \right]. \end{aligned}$$

记 H 为

$$\left\{ \frac{1}{2} S(t) \left[\int_0^t K(u)^{-2} \{I(a \leq u \leq y) + I(v \leq c)I(a \wedge c \leq u \leq y)\} dW(u) \right. \right. \\ \left. \left. - \frac{I(a \wedge c \leq u \leq y)I(y \leq t)}{K(y)} \right] : t \in [0, \tau] \right\}$$

的封套函数. □

引理 3 在假设 1 下, 若 H 平方可积, 则当 $n \rightarrow \infty$ 时, 随机过程 $n^{1/2}\{\hat{S}(t) - S(t)\}$ 在 $0 < t < \tau$ 时收敛到一个零期望的 Gauss 过程, 其方差 - 协方差函数如下:

$$\tilde{\sigma}_S(t_1, t_2) = S(t_1)S(t_2)E\{\phi_1(t_1)\phi_1(t_2)\},$$

其中

$$\phi_1(t) = \frac{1}{2} \int_0^t K(u)^{-2} \{I(A_i \leq u \leq Y_i) + \Delta_i I(\tilde{V}_i \leq u \leq Y_i)\} dW(u) - \frac{\Delta_i I(Y_i \leq t)}{K(Y_i)}.$$

证明 此引理证明类似文献 [13, 定理 1] 的证明. 对于长度偏差数据而言, (A, V) 和 (A, T) 的联合分布如下:

$$f_{A,V}(a, v) = \frac{1}{\mu} f(a+v)I(a > 0, v > 0), \quad (\text{A.2})$$

其中 $\mu = \int_0^\infty tf(t)dt$. 我们有

$$\begin{aligned} E[I(A \leq t \leq Y)] &= E[I(A \leq t, A+V \geq t, A+C \geq t)] \\ &= \int \int_{a \leq t, a+v \geq t} f_{A,V}(a, v) G(t-a) da dv \\ &= \frac{1}{\mu} \int_0^t G(t-a) S(t) da \\ &= \frac{1}{\mu} S(t) w(t), \end{aligned} \quad (\text{A.3})$$

其中

$$w(t) = \left[\int_0^t G(u) du \right].$$

另一方面,

$$\begin{aligned} E[\Delta I(\tilde{V} \leq t \leq Y)] &= E[I(V \leq t \leq A + V, V \leq C)] \\ &= \int \int_{0 \leq v \leq t, a+v \geq t} f_{A,V}(a, v) G(v) da dv \\ &= \frac{1}{\mu} S(t) w(t). \end{aligned} \quad (\text{A.4})$$

定义函数

$$\begin{aligned} \tilde{K}_n(t) &= \frac{1}{n} \sum_{i=1}^n R_i(t) = \frac{1}{2n} \sum_{i=1}^n \{I(A_i \leq t \leq Y_i) + \Delta_i I(\tilde{V}_i \leq t \leq Y_i)\}, \\ \tilde{K}(t) &= \frac{1}{n} \sum_{i=1}^n \{I(A_i \leq t \leq Y_i)\}, \\ \tilde{W}_n(t) &= \frac{1}{n} \sum_{i=1}^n N_i(t) = \frac{1}{n} \sum_{i=1}^n \Delta_i I(Y_i \leq t), \\ K(t) &= \frac{1}{2} E\{I(A \leq t \leq Y) + \Delta I(\tilde{V} \leq t \leq Y)\}, \\ W(t) &= E\{\Delta I(Y \leq t)\}. \end{aligned}$$

因此, 由 (A.3) 和 (A.4) 有

$$K(t) = E\tilde{K}_n(t) = \mu^{-1} S(t) w(t), \quad (\text{A.5})$$

$$dW(t) = E[dN(t)] = \mu^{-1} f(t) \int_0^t G(u) du dt = \mu^{-1} f(t) w(t) dt. \quad (\text{A.6})$$

T^0 的累积风险函数可推导如下:

$$\Lambda(t) = \int_0^t \frac{f(u)}{S(u)} du = \int_0^t \frac{\mu^{-1} f(u) w(u) du}{\mu^{-1} S(u) w(u)} = \int_0^t \frac{dW(u)}{K(u)}. \quad (\text{A.7})$$

我们可通过下式估计 $\Lambda(t)$:

$$\hat{\Lambda}(t) = \int_0^t \frac{d\tilde{W}_n(u)}{\tilde{K}_n(u)}.$$

由于 $n^{1/2}\{\tilde{W}_n(t) - W(t)\}$ 和 $n^{1/2}\{\tilde{K}_n(t) - K(t)\}$ 是由独立同分布样本构成的经验过程, 且其组成部分可分解为区间示性函数的加减, 因此为 Donsker 族. 由 $n^{1/2}\{\tilde{W}_n(t) - W(t), \tilde{K}_n(t) - K(t)\}$ 弱收敛性, 应用函数 Delta 方法, 可以推导出下面的渐近表达式:

$$\begin{aligned} n^{1/2}\{\hat{\Lambda}(t) - \Lambda(t)\} &= \int_0^t \frac{n^{1/2} d\tilde{W}_n(u)}{K(u)} - \int_0^t \frac{n^{1/2} \tilde{K}(u) dW(u)}{K(u)^2} \\ &\quad - \int_0^t \frac{1}{2\sqrt{n}} \sum_{i=1}^n \Delta_i I\{Y_i \geq u \geq \tilde{V}_i\} \frac{dW(u)}{K(u)^2} + o_P(1) \end{aligned}$$

$$= -n^{-1/2} \sum_{i=1}^n \phi_i(t) + o_P(1),$$

其中

$$\phi_i(t) = \frac{1}{2} \int_0^t K(u)^{-2} \{I(A_i \leq u \leq Y_i) + \Delta_i I(\tilde{V}_i \leq u \leq Y_i)\} dW(u) - \frac{\Delta_i I(Y_i \leq t)}{K(Y_i)}.$$

可以证明 $E[\phi_1(t)] = 0$.

而且, 我们能够确立 $n^{1/2}\{\hat{S}_n^{\text{new}}(t) - S(t)\}$ 的弱收敛性. 由定义

$$\hat{S}_n^{\text{new}}(t) = \prod_{u \in [0, t]} \{1 - d\hat{\Lambda}(u)\}$$

和文献 [24, 引理 20.14], 利用函数 Delta 方法, 可得渐近表达式

$$n^{1/2}\{\hat{S}_n^{\text{new}}(t) - S(t)\} = n^{-1/2} \sum_{i=1}^n S(t)\phi_i(t) + o_P(1) \quad (\text{A.8})$$

在 $[0, \tau]$ 上一致成立. 下面证明 $n^{-1/2} \sum_{i=1}^n S(t)\phi_i(t)$ 在 $\mathbb{D}[0, \tau]$ 上的弱收敛性, 首先证明

$$\left\{ \frac{1}{2} S(t) \left[\int_0^t K(u)^{-2} \{I(a \leq u \leq y) + I(v \leq c) I(a \wedge c \leq u \leq y)\} dW(u) - \frac{I(a \wedge c \leq u \leq y) I(y \leq t)}{K(y)} \right] : t \in [0, \tau] \right\} \quad (\text{A.9})$$

是 P-Donsker 族, 其中 P 是 A 、 C 、 Y 和 V 的联合分布. 可以看到

$$S(t) \int_0^t K(u)^{-2} \{I(a \leq u \leq y)\} dW(u) = S(t) \left[I(a \leq y) I(y \leq t) \int_a^y K(u)^{-2} dW(u) + I(a \leq t) I(y > t) \int_a^t K(u)^{-2} dW(u) \right]$$

能分解成示性函数和有界均匀单调函数的 $+$ 、 $-$ 和 $\times$ 运算, 由文献 [22, 例 2.10.7] 和 H 平方可积可知, (A.9) 是 P-Donsker 族. 继而 $n^{1/2}\{\hat{S}_n^{\text{new}}(t) - S(t)\}$ 满足等度连续性, 满足渐近紧的条件, 而由中心极限定理可知, 其任意有限维也依分布收敛, 因此在 $\mathbb{D}[0, \tau]$ 上弱收敛 (参见文献 [22, 第 1.5 小节]). 引理 3 证明完成. $\square$

定理 3 的证明 由命题 1 和引理 3 知, $\sqrt{n}(\hat{m}_n^{\text{new}}(t) - m(t))$ 弱收敛到一个 Guass 过程. 由 Fubini 定理知, 其协方差为

$$\begin{aligned} \Gamma^{\text{new}}(s, t) = & \frac{1}{S^2(s)S^2(t)} \left[S(s)S(t) \int_s^\tau \int_t^\tau S(u)S(v) E\{\phi_1(u)\phi_1(v)\} dudv \right. \\ & + S(s)S(t) E\{\psi_1(s)\psi_1(t)\} \int_s^\tau S(u)du \int_t^\tau S(u)du \\ & - S(s) \left(\int_t^\tau S(u)du \right) \left(\int_s^\tau S(t)S(u) E\{\phi_1(t)\phi_1(u)\} du \right) \\ & \left. - S(t) \int_s^\tau S(u)du \left(\int_t^\tau S(s)S(u) E\{\phi_1(s)\phi_1(u)\} du \right) \right]. \end{aligned}$$

证毕. $\square$

Composite estimator of mean residual lifetime with length-biased and right-censored data

Qiuxia Ma, Sheng Pan & Yong Zhou

Abstract Length-biased data are very common in prevalent cases by cross-sectional sampling when studying time-to-event data, the distribution of length bias data changes from the population distribution and the censoring is informative censoring. But it is difficult to estimate lifetime exactly from such biased observational data with censoring. Bias and the informative censoring in the length-biased data make many classical statistical methods invalid. In this paper, we study a mean residual lifetime model when sample is subject to length-biased and right-censored data. Nonparametric estimation methods are proposed based on estimating equations. To improve the efficiency of the estimator, we take account into the informative censoring and auxiliary information of length-bias data. Although the maximum likelihood estimation is efficient, its computation has so many iterations. Thus we propose a simple way to add the auxiliary information, where the new estimator needs less calculation times. At the same time, asymptotic properties of nonparametric estimators for the mean residual lifetime are established. In simulation studies, it can be seen that our estimator is more effective and simpler.

Keywords length-biased, right-censored, mean residual lifetime model, nonparametric estimation method, confidence intervals

MSC(2010) 62N02

doi: 10.1360/N012018-00114