具有连续辅助协变量的均值剩余寿命模型的统计推断

张虎, 杨青龙*

中南财经政法大学统计与数学学院, 武汉 430073

E-mail: zhh11497@sina.com, yangqinglong@zuel.edu.cn

收稿日期: 2019-10-25; 接受日期: 2020-03-03; 网络出版日期: 2020-10-23; * 通信作者

国家自然科学基金 (批准号: 11671131 和 11971324) 和中南财经政法大学中央高校基本科研业务费 (批准号: 2722020JCT030) 资助项目

摘要 在许多实际问题的研究中, 由于预算或者技术等因素限制, 重要协变量值只能对某一个核实集进行准确测量, 而对应重要协变量的连续辅助协变量则对研究队列中的所有个体均可观测. 利用辅助协变量的信息有助于提高统计研究的效率. 本文考虑如何利用辅助协变量的信息来提高比例均值剩余寿命模型的统计推断效率. 本文基于逆概率删失加权技巧构造一个估计方程来估计回归参数, 并证明所得估计是相合的和渐近正态的. 大量数值模拟验证了所提出的方法在有限样本下的有效性. 最后, 将所提出的方法应用于 Mayo 临床试验中原发性胆汁性肝硬化问题的研究中.

关键词 逆概率删失加权 估计方程 经验过程 核平滑 核实集

MSC (2010) 主题分类 62G05, 62N01

1 引言

为了研究协变量对生存时间的影响, 对危险率函数进行建模是生存分析中最基本和最有效的方式. 但在随机临床试验、人口学和精算学等领域的研究中, 剩余寿命是影响最终决策的一个关键指标. 生存时间 T 在时刻 t (≥ 0) 的均值剩余寿命函数 (mean residual life function, MRLF) 定义为 $m(t) = E(T - t | T > t)$. 它表示一个已经生存到 t 的个体的剩余寿命的均值, 是度量剩余寿命最直接的一个数量指标. 与危险率函数相比, 均值剩余寿命函数的统计解释更加直观, 能够提供更直接和有效的评价信息. 例如, 保险公司在对投保者目前的身体状况和存活至今的事实进行了解后, 对其剩余寿命的均值尤其感兴趣. 进一步地, 均值剩余寿命函数与生存函数具有一一对应关系, 因而可通过对均值剩余寿命函数建模达到对生存函数的了解. 因此, 均值剩余寿命模型的研究越来越得到人们的重视. Oakes 和 Dasu^[1] 于 1990 年最早提出了比例均值剩余寿命模型. Maguluri 和 Zhang^[2] 利用指数回归和比例风险率模型的思想给出了基于估计方程的两种不同的估计方法. Chen 和 Cheng^[3] 提出了拟偏

英文引用格式: Zhang H, Yang Q L. Statistical inference for proportional mean residual life model with continuous auxiliary covariate (in Chinese). Sci Sin Math, 2020, 50: 1631–1648, doi: 10.1360/SSM-2019-0263

得分估计方程方法并利用计数过程的理论建立了估计量的理论性质. Chen 等^[4]通过逆概率删失加权方法修正了 Maguluri 和 Zhang^[2]的估计方法来考虑数据删失. Chen 和 Cheng^[5]及 Chen^[6]研究了协变量对 MRLF 的加性影响. Sun 和 Zhang^[7]研究了一类转换均值剩余寿命模型, 并给出了独立删失和非独立删失情形下的推断方法, 此模型可包括比例均值剩余寿命模型和加法均值剩余寿命模型作为其特例. Sun 等^[8]研究了时间相依的变系数均值剩余寿命模型. Cai 等^[9]利用估计方程的思想提出了加乘均值剩余寿命模型.

前面提到的研究工作都假定数据是完全观测的, 但由于预算或技术等的限制, 有时只能观测全队列的一个子集的重要暴露变量, 这个子集称为核实集. 但一些辅助的协变量, 对全队列每个个体均可以观测到. 具有辅助协变量的数据在生物医学研究中经常出现 (参见文献 [10,11]). 对于这样的生存数据, 研究如何利用辅助协变量信息来提高统计研究效率是很有必要的. 在危险函数模型中, 已有很多统计方法来解决此类问题. 例如, Prentice^[12]在一定的严格假设条件下, 对 Cox 比例风险模型提出了诱导部分似然方法. Zhou 和 Pepe^[10]放宽了这些假设, 并在辅助协变量离散的情形下, 提出了估计部分似然方法. Zhou 和 Wang^[11]通过使用核平滑的方法将 Zhou 和 Pepe^[10]的估计方法拓展到具有连续性辅助协变量的问题. Liu 等^[13-15]和杨青龙等^[16]在边际危险率框架下发展了估计部分似然方法来处理具有辅助信息的多元失效时间数据. Kulich 和 Lin^[17]以及 Jiang 和 Zhou^[18]对加法危险模型提出了修正的伪得分估计方法. Chen 和 Wang^[19]在一些协变量随机缺失的情形下, 基于逆概率加权 and 扩充的逆概率加权方法提出了比例均值剩余寿命模型的统计推断方法.

据我们所知, 具有辅助协变量的均值剩余寿命模型相关工作寥寥无几. 本文将给出具有连续辅助协变量的比例均值剩余寿命模型的统计推断方法. 均值剩余寿命函数可通过核实集上所有观测的加权平均来得到经验估计. 因此, 本文基于逆概率删失加权技巧提出了一个估计方程, 并获得了估计的相合性和渐近正态性. 模拟研究表明提出的方法效果良好并比只使用核实集信息得到的估计更有效.

本文余下内容安排如下: 第 2 节引入模型, 提出在主协变量缺失但辅助协变量对所有个体均能观测的情况下的估计方程; 第 3 节建立估计的相合性和渐近正态性; 第 4 节通过大量的模拟来验证提出的方法的有限样本性质, 并将提出的方法应用于 Mayo 临床试验中原发性胆汁性肝硬化问题的研究. 所有的证明将在附录中给出.

2 模型和估计

假设有 n 个独立的个体, 记 $\tilde{T}_i$ 和 C_i 分别表示第 i 个个体的生存时间和相对应的删失时间. 于是观测时间 $T_i = \min(\tilde{T}_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$ 和 $Y_i(t) = I(T_i \geq t)$ 分别表示删失指示变量和风险过程, 其中 $I(\cdot)$ 表示示性函数. 令 $\mathbf{W}_i(t) = (\mathbf{Z}_i(t)^T, \mathbf{X}_i(t)^T)^T$ 表示一个 p 维协变量, $\mathbf{Z}(t)$ 是 p_1 维的主暴露变量, $\mathbf{X}(t)$ 是 p_2 维 ($p = p_1 + p_2$) 对全部个体均可观测的协变量. 假定第 i 个个体均值剩余寿命模型形式如下:

$$m(t | \mathbf{W}_i(t)) = m_0(t) \exp\{\beta_{10}^T \mathbf{Z}_i(t) + \beta_{20}^T \mathbf{X}_i(t)\}, \quad (2.1)$$

其中 $m_0(t)$ 是未知基准 MRLF, $\beta_0 = (\beta_{10}^T, \beta_{20}^T)^T$ 是感兴趣的待估参数.

记 V 为核实集, $\bar{V}$ 为非核实集, n_V 和 $n_{\bar{V}}$ 分别表示核实集 V 和非核实集 $\bar{V}$ 中包含个体数量, 核实集 V 中个体 $\mathbf{Z}(t)$ 和 $\mathbf{X}(t)$ 均可观测, 非核实集 $\bar{V}$ 中个体只有 $\mathbf{X}(t)$ 可观测. η_i 为第 i 个个体是否属于核实集 V 的示性变量. 令 $\rho = \lim_{n \rightarrow \infty} (n_V/n)$ 表示核实集比例, $\mathbf{A}(t)$ 是主暴露变量 $\mathbf{Z}(t)$ 的连续

辅助协变量, 对全队列均可观测. 假定给定 $\mathbf{Z}(t)$ 和 $\mathbf{X}(t)$ 的条件下, $\mathbf{A}(t)$ 对 MRLF 没有贡献, 即

$$m(t | \mathbf{Z}_i(t), \mathbf{X}_i(t), \mathbf{A}_i(t)) = m(t | \mathbf{Z}_i(t), \mathbf{X}_i(t)).$$

如果第 i 个个体属于 V , 则均值剩余寿命模型形式同 (2.1). 如果第 i 个个体属于 $\bar{V}$, 由 Prentice^[12] 的思想可得第 i 个个体的诱导 MRLF 为

$$\begin{aligned} m[t | \mathbf{X}_i(t), \mathbf{A}_i(t)] &= E[\tilde{T}_i - t | \tilde{T}_i > t, \mathbf{X}_i(t), \mathbf{A}_i(t)] \\ &= E[E[\tilde{T}_i - t | \tilde{T}_i > t, \mathbf{Z}_i(t), \mathbf{X}_i(t), \mathbf{A}_i(t)] | \tilde{T}_i > t, \mathbf{X}_i(t), \mathbf{A}_i(t)] \\ &= E[m_0(t) \exp\{\beta_{10}^T \mathbf{Z}_i(t) + \beta_{20}^T \mathbf{X}_i(t)\} | \tilde{T}_i > t, \mathbf{X}_i(t), \mathbf{A}_i(t)] \\ &= m_0(t) E[\exp\{\beta_{10}^T \mathbf{Z}_i(t)\} | \tilde{T}_i > t, \mathbf{X}_i(t), \mathbf{A}_i(t)] \exp\{\beta_{20}^T \mathbf{X}_i(t)\}. \end{aligned}$$

假定删失为非信息删失 (参见文献 [20]), 模型可重新写成

$$m[t | \mathbf{X}_i(t), \mathbf{A}_i(t)] = m_0(t) E[\exp\{\beta_{10}^T \mathbf{Z}_i(t)\} | Y_i(t) = 1, \mathbf{X}_i(t), \mathbf{A}_i(t)] \exp\{\beta_{20}^T \mathbf{X}_i(t)\}. \quad (2.2)$$

注意到 (2.2) 诱导出的模型仍然是比例均值剩余寿命模型, 相对均值剩余寿命函数为

$$\phi_i(\boldsymbol{\beta}_0, t) = E[\exp\{\beta_{10}^T \mathbf{Z}_i(t)\} | Y_i(t) = 1, \mathbf{X}_i(t), \mathbf{A}_i(t)] \exp\{\beta_{20}^T \mathbf{X}_i(t)\}.$$

基于模型 (2.1) 和 (2.2), 可将相对均值剩余寿命函数统一写成

$$r_i(\boldsymbol{\beta}_0, t) = \eta_i \psi_i(\boldsymbol{\beta}_0, t) + (1 - \eta_i) \phi_i(\boldsymbol{\beta}_0, t),$$

其中 $\psi_i(\boldsymbol{\beta}_0, t) = \exp\{\beta_{10}^T \mathbf{Z}_i(t) + \beta_{20}^T \mathbf{X}_i(t)\}$.

记 $G(t)$ 为 C 的生存函数, $\hat{G}(t)$ 是 $G(t)$ 基于 $\{T_i, 1 - \Delta_i; i = 1, \dots, n\}$ 的 Kaplan-Meier 估计. 定义

$$M_i(t) = \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} [T_i - t - m_0(t) r_i(\boldsymbol{\beta}_0, t)], \quad i = 1, \dots, n.$$

在模型 (2.1), $M_i(t)$ 为 0 均值的随机过程. 因此, 对于给定的 $\boldsymbol{\beta}$, $m_0(t)$ 的估计可通过求解下面方程得到:

$$\sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} [T_i - t - m_0(t) r_i(\boldsymbol{\beta}, t)] = 0, \quad 0 \leq t \leq \tau, \quad (2.3)$$

其中 τ 是研究结束的时间.

经过简单代数运算可导出基准 MRLF $m_0(t)$ 的一个估计

$$\hat{m}(\boldsymbol{\beta}, t) = \frac{\sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} [T_i - t]}{\sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} r_i(\boldsymbol{\beta}, t)}. \quad (2.4)$$

记 $\mathbf{f}^{(j)}(\boldsymbol{\beta}, t)$ 为 $f(\boldsymbol{\beta}, t)$ 关于 $\boldsymbol{\beta}$ 的第 j 阶导数, $f(\boldsymbol{\beta}, t)$ 为任一可导函数. 为了估计 $\boldsymbol{\beta}_0$, 使用广义估计方程方法和逆概率删失加权技巧, 可通过下面方程来估计 $\boldsymbol{\beta}_0$:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\tau \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} \frac{\mathbf{r}_i^{(1)}(\boldsymbol{\beta}, t)}{r_i(\boldsymbol{\beta}, t)} [T_i - t - r_i(\boldsymbol{\beta}, t) \hat{m}(\boldsymbol{\beta}, t)] dH(t) = 0, \quad (2.5)$$

其中 $H(t)$ 为 $[0, t]$ 上递增的权函数.

由于 (2.5) 无法直接计算, 需要估计 $r_i(\beta, t)$, 本质上要估计 $\phi_i(\beta, t)$. 记 $\mathbf{W}_i^*(t) = (\mathbf{A}_i(t)^T, \mathbf{X}_i(t)^T)^T$ 是 p 维向量. 若 $i \in \bar{V}$, 则 $\phi_i(\beta, t)$ 可通过 Nadaraya-Watson 估计得到 (参见文献 [21, 22])

$$\hat{\phi}_i(\beta, t) = \frac{\sum_{j=1}^n \eta_j Y_j(t) K_h(\mathbf{W}_j^*(t) - \mathbf{W}_i^*(t)) \exp\{\beta_1^T \mathbf{Z}_j(t)\}}{\sum_{j=1}^n \eta_j Y_j(t) K_h(\mathbf{W}_j^*(t) - \mathbf{W}_i^*(t))} \exp\{\beta_2^T \mathbf{X}_i(t)\},$$

其中 $K_h(\cdot) = K(\cdot/h)$, $K(\cdot)$ 是 p 元一致有界核函数, 具有有限支撑且满足 $\int K(\mathbf{u}) d\mathbf{u} = 1$ 和 $\int K^2(\mathbf{u}) d\mathbf{u} < \infty$, h 是窗宽. 对于 $\mathbf{u} = (u_1, \dots, u_p)^T$, $K(\mathbf{u}/h) = K(u_1/h, \dots, u_p/h)$. 从而可得到 $r_i(\beta, t)$ 的估计

$$\hat{r}_i(\beta, t) = \eta_i \psi_i(\beta, t) + (1 - \eta_i) \hat{\phi}_i(\beta, t).$$

用 $\hat{r}_i^{(l)}(\beta, t)$ ($l = 0, 1$) 在 (2.5) 中替换 $r_i^{(l)}(\beta, t)$, 可得新的估计方程

$$\hat{U}(\beta) = \sum_{i=1}^n \int_0^\tau \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} (T_i - t) \left[\frac{\hat{r}_i^{(1)}(\beta, t)}{\hat{r}_i(\beta, t)} - \frac{\hat{S}^{(1)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} \right] dH(t) = 0, \quad (2.6)$$

其中 $\hat{S}^{(l)}(\beta, t) = n^{-1} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} \hat{r}_i^{(l)}(\beta, t)$ ($l = 0, 1$). 记 $\hat{\beta}_p$ 是 $\hat{U}(\beta) = 0$ 的解, 可通过 Newton 迭代法得到. 记 $\hat{m}_p(t)$ 是 $m_0(t)$ 的相应估计, 在 (2.4) 中用 $\hat{r}_i(\hat{\beta}_p, t)$ 替换 $r_i(\beta, t)$ 即可得到.

3 渐近性质

为了给出 $\hat{\beta}_p$ 和 $\hat{m}_p(t)$ 的渐近性质, 引入以下记号: 对于向量 $\mathbf{a}$, 记 $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$, $\|\mathbf{a}\| = \sup_i |a_i|$; 对于矩阵 $\mathbf{M}$, 令 $\|\mathbf{M}\| = \sup_{i,j} |m_{ij}|$, m_{ij} 是 $\mathbf{M}$ 的第 (i, j) 个元素. 定义

$$\begin{aligned} \mathbf{S}^{(0)}(\beta, t) &= n^{-1} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{G(T_i)} r_i(\beta, t), \\ \mathbf{S}^{(1)}(\beta, t) &= n^{-1} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{G(T_i)} \mathbf{r}_i^{(1)}(\beta, t), \\ \mathbf{S}^{(2)}(\beta, t) &= n^{-1} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{G(T_i)} \mathbf{r}_i^{(2)}(\beta, t), \\ \mathbf{S}^{(3)}(\beta, t) &= n^{-1} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{G(T_i)} \left\{ \frac{\mathbf{r}_i^{(2)}(\beta, t)}{r_i(\beta, t)} \right\} r_i(\beta_0, t), \\ \mathbf{S}^{(4)}(\beta, t) &= n^{-1} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{G(T_i)} \left\{ \frac{\mathbf{r}_i^{(1)}(\beta, t)}{r_i(\beta, t)} \right\}^{\otimes 2} r_i(\beta_0, t). \end{aligned}$$

对于 $k = 1, \dots, 4$, 令 $\mathbf{s}^{(k)}(\beta, t) = E(\mathbf{S}^{(k)}(\beta, t))$. 为了保证估计的渐近性质, 给出如下条件:

- (C1) $H(t)$ 在 $t \in [0, \tau]$ 内几乎处处一致收敛到有界的非随机函数 $\tilde{H}(t)$;
- (C2) 函数 $G(t)$ 连续, $m_0(t)$ 在 $[0, \tau]$ 上连续可微;
- (C3) 对于 $i = 1, \dots, n$, $\mathbf{Z}_i(t)$ 和 $\mathbf{X}_i(t)$ 具有有界变差;
- (C4) $\Gamma(\beta_0)$ 是正定的, 其中

$$\Gamma(\beta_0) = \int_0^\tau \left[\frac{\mathbf{s}^{(4)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} - \left(\frac{\mathbf{s}^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right)^{\otimes 2} \right] s^{(0)}(\beta_0, t) m_0(t) d\tilde{H}(t);$$

(C5) 对于 $\mathbf{W}^*(t)$ 的每个取值及任意的 $t \in [0, \tau]$, $\Pr(Y(t) = 1 \mid \mathbf{W}^*(t) = \mathbf{w}) > 0$;

(C6) 核函数 $K(\cdot)$ 称为是 α_0 阶, 若 $\alpha_0 \equiv \inf\{|\boldsymbol{\alpha}| > p; \int_{\mathcal{R}^p} \mathbf{u}^\alpha K(\mathbf{u}) d\mathbf{u} \neq 0\}$, 其中 $\mathbf{u}^\alpha = u_1^{\alpha_1} \cdots u_p^{\alpha_p}$, $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_p$, $\mathbf{u} = (u_1, \dots, u_p)$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)$ 且 α_i 均为非负整数, 窗宽 $\mathbf{h}$ 满足 $\sqrt{n}\|\mathbf{h}\|^{\alpha_0} \rightarrow 0$ 和 $\frac{\log n}{\sqrt{n}\|\mathbf{h}\|^p} \rightarrow 0$;

(C7) 对于给定 t , 令 $H(\mathbf{u}, \mathbf{v}, \mathbf{s})$ 为 $(\eta Y(t), \mathbf{W}^*(t), \mathbf{X}(t))$ 的联合分布, 假定 $h(\mathbf{v}, \mathbf{s}) = \frac{\partial^2 H(1, \mathbf{v}, \mathbf{s})}{\partial \mathbf{v} \partial \mathbf{s}}$ 关于 $\mathbf{v}$ 的每个分量均有 α_0 阶偏导数.

条件 (C1)–(C4) 是文献 [3, 7] 中的正则条件. 条件 (C5) 是识别条件. 条件 (C6) 是通常的核光滑条件. 条件 (C7) 是一个便于渐近结果推导的技术假设.

下面定理给出了 $\hat{\beta}_p$ 的渐近性质, 定理证明将在附录中给出.

定理 3.1 若正则条件 (C1)–(C7) 成立, 则 $\hat{\beta}_p$ 是 β_0 的一个相合估计, 并且 $n^{1/2}(\hat{\beta}_p - \beta_0)$ 是渐近正态的, 其极限分布的均值为 0, 方差阵为 $\Sigma(\beta_0) = \Gamma^{-1}(\beta_0)[\Sigma_1(\beta_0) + \Sigma_2(\beta_0)]\Gamma^{-1}(\beta_0)$, 其中

$$\Sigma_1(\beta_0) = (1 - \rho)E(\boldsymbol{\mu}_i(\beta_0)\boldsymbol{\mu}_i(\beta_0)^T),$$

$$\Sigma_2(\beta_0) = \rho E(\boldsymbol{\nu}_i(\beta_0)\boldsymbol{\nu}_i(\beta_0)^T),$$

$$\boldsymbol{\mu}_i(\beta_0) = \int_0^\tau M_i(t) \left[\frac{\mathbf{r}_i^{(1)}(\beta_0, t)}{r_i(\beta_0, t)} - \frac{\mathbf{s}^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right] d\tilde{H}(t) + \int_0^\tau \frac{\mathbf{R}^*(\beta_0, t)}{\pi(t)} dM_i^c(t),$$

$$N_i^c(t) = I(T_i \leq t, \Delta_i = 0),$$

$$\Lambda^c(t) = -\log(G(t)),$$

$$M_i^c(t) = N_i^c(t) - \int_0^t Y_i(u) d\Lambda^c(u),$$

$$\mathbf{R}^*(\beta_0, t) = E \left\{ Y_i(t) \int_0^\tau M_i(u) \left[\frac{\mathbf{r}_i^{(1)}(\beta_0, u)}{r_i(\beta_0, u)} - \frac{\mathbf{s}^{(1)}(\beta_0, u)}{s^{(0)}(\beta_0, u)} \right] d\tilde{H}(u) \right\},$$

$$\pi(t) = \Pr(T > t),$$

$$\boldsymbol{\nu}_i(\beta_0) = \int_0^\tau M_i(t) \left[\frac{\mathbf{r}_i^{(1)}(\beta_0, t)}{r_i(\beta_0, t)} - \frac{\mathbf{s}^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right] d\tilde{H}(t) + \int_0^\tau \frac{\mathbf{R}^*(\beta_0, t)}{\pi(t)} dM_i^c(t) - \frac{1 - \rho}{\rho} \mathbf{Q}_i(\beta_0),$$

$$\mathbf{Q}_i(\beta_0) = \int_0^\tau \frac{\Delta_i Y_i(t)}{G(T_i)} \left(\frac{\boldsymbol{\phi}_i^{(1)}(\beta_0, t)}{\boldsymbol{\phi}_i(\beta_0, t)} - \frac{\mathbf{s}^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right) [\psi_i(\beta_0, t) - \phi_i(\beta_0, t)] m_0(t) d\tilde{H}(t).$$

$\hat{\beta}_p$ 的相合性可由 Foutz^[23] 的逆函数定理和现代经验过程的结果 (参见文献 [24–27]) 得到, 其渐近正态性可由估计函数的 Taylor 展式、经验过程理论和多元分布中心极限定理导出.

$\Sigma(\beta_0)$ 的相合估计可通过相应的样本估计值来替代:

$$\hat{\Sigma}(\hat{\beta}_p) = \hat{\Gamma}^{-1}(\hat{\beta}_p)[\hat{\Sigma}_1(\hat{\beta}_p) + \hat{\Sigma}_2(\hat{\beta}_p)]\hat{\Gamma}^{-1}(\hat{\beta}_p),$$

其中

$$\hat{\Gamma}(\beta) = \int_0^\tau \left[\frac{\hat{\mathbf{S}}^{(4)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} - \left(\frac{\hat{\mathbf{S}}^{(1)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} \right)^{\otimes 2} \right] \hat{S}^{(0)}(\beta, t) \hat{m}_p(t) dH(t),$$

$$\hat{\mathbf{S}}^{(4)}(\beta, t) = n^{-1} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} \frac{\{\hat{\mathbf{r}}_i^{(1)}(\beta, t)\}^{\otimes 2}}{\hat{r}_i(\beta, t)},$$

$$\hat{\Sigma}_1(\beta) = n^{-1} \sum_{i \in \bar{V}} \hat{\boldsymbol{\mu}}_i(\beta) \hat{\boldsymbol{\mu}}_i(\beta)^T,$$

$$\begin{aligned}
 \widehat{\Sigma}_2(\beta) &= n^{-1} \sum_{i \in V} \widehat{\nu}_i(\beta) \widehat{\nu}_i(\beta)^T, \\
 \widehat{\mu}_i(\beta) &= \int_0^\tau \widehat{M}_i(t) \left[\frac{\widehat{r}_i^{(1)}(\beta, t)}{\widehat{r}_i(\beta, t)} - \frac{\widehat{S}^{(1)}(\beta, t)}{\widehat{S}^{(0)}(\beta, t)} \right] dH(t) + \int_0^\tau \frac{\widehat{R}^*(\beta, t)}{\widehat{\pi}(t)} d\widehat{M}_i^c(t), \\
 \widehat{\pi}(t) &= n^{-1} \sum_{i=1}^n Y_i(t), \\
 \widehat{\Lambda}^c(t) &= n^{-1} \sum_{i=1}^n \int_0^t \frac{dN_i^c(u)}{\widehat{\pi}(u)}, \\
 \widehat{M}_i^c(t) &= N_i^c(t) - \int_0^t Y_i(u) d\widehat{\Lambda}^c(u), \\
 \widehat{R}^*(\beta, t) &= n^{-1} \sum_{i=1}^n Y_i(t) \int_0^\tau \widehat{M}_i(u) \left[\frac{\widehat{r}_i^{(1)}(\beta, u)}{\widehat{r}_i(\beta, u)} - \frac{\widehat{S}^{(1)}(\beta, u)}{\widehat{S}^{(0)}(\beta, u)} \right] dH(u), \\
 \widehat{\nu}_i(\beta) &= \int_0^t \widehat{M}_i(t) \left[\frac{\widehat{r}_i^{(1)}(\beta, t)}{\widehat{r}_i(\beta, t)} - \frac{\widehat{S}^{(1)}(\beta, t)}{\widehat{S}^{(0)}(\beta, t)} \right] dH(t) + \int_0^\tau \frac{\widehat{R}^*(\beta, t)}{\widehat{\pi}(t)} d\widehat{M}_i^c(t) - \frac{n_V}{n_V} \widehat{Q}_i(\beta), \\
 \widehat{M}_i(t) &= \frac{\Delta_i Y_i(t)}{\widehat{G}(T_i)} [T_i - t - \widehat{m}_p(t) \widehat{r}_i(\beta_p, t)], \\
 \widehat{Q}_i(\beta) &= \int_0^\tau \frac{\Delta_i Y_i(t)}{\widehat{G}(T_i)} \left(\frac{\widehat{\phi}_i^{(1)}(\beta, t)}{\widehat{\phi}_i(\beta, t)} - \frac{\widehat{S}^{(1)}(\beta, t)}{\widehat{S}^{(0)}(\beta, t)} \right) [\psi_i(\beta, t) - \widehat{\phi}_i(\beta, t)] \widehat{m}_p(t) dH(t).
 \end{aligned}$$

下面定理给出了基准均值剩余寿命函数估计的渐近性质.

定理 3.2 若正则条件 (C1)–(C7) 成立, 则 $\widehat{m}_p(t)$ 在 $t \in [0, \tau]$ 中几乎处处一致收敛到 $m_0(t)$. 进一步, $n^{1/2}(\widehat{m}_p(t) - m_0(t))$ 弱收敛到一个零均值的 Gauss 过程, 协方差函数为

$$\xi(s, t) = E(\Psi_1(s) \Psi_1(t)),$$

其中

$$\begin{aligned}
 \Psi_i(t) &= \eta_i \Phi_i(t) + (1 - \eta_i) \Upsilon_i(t), \\
 \Phi_i(t) &= \frac{M_i(t)}{s^{(0)}(\beta_0, t)} + \frac{\int_0^\tau \frac{\Theta(t, u)}{\pi(u)} dM_i^c(u)}{s^{(0)}(\beta_0, t)} - \zeta(\beta_0, t)^T \Gamma^{-1}(\beta_0) \nu_i(\beta_0) \\
 &\quad - \frac{1 - \rho}{\rho} m_0(t) \frac{\Delta_i Y_i(t)}{\widehat{G}(T_i)} \frac{[\psi_i(\beta_0, t) - \phi_i(\beta_0, t)]}{s^{(0)}(\beta_0, t)}, \\
 \Theta(t, u) &= E[Y_i(u) M_i(t)], \\
 \Upsilon_i(t) &= \frac{M_i(t)}{s^{(0)}(\beta_0, t)} + \frac{\int_0^\tau \frac{\Theta(t, u)}{\pi(u)} dM_i^c(u)}{s^{(0)}(\beta_0, t)} - \zeta(\beta_0, t)^T \Gamma^{-1}(\beta_0) \mu_i(\beta_0), \\
 \zeta(\beta_0, t) &= m_0(t) \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)}.
 \end{aligned}$$

附录中将说明 $n^{1/2}(\widehat{m}_p(t) - m_0(t))$ 与 $n^{-\frac{1}{2}} \sum_{i=1}^n \Psi_i(t)$ 渐近相等. 从而由多元分布中心极限定理和胎紧性的证明得到过程 $n^{1/2}(\widehat{m}_p(t) - m_0(t))$ 弱收敛到零均值的 Gauss 过程, 协方差函数为 $\xi(s, t) = E(\Psi_1(s) \Psi_1(t))$.

协方差函数 $\xi(s, t)$ 的相合估计可以通过相应的经验估计得到, 即

$$\hat{\xi}(s, t) = n^{-1} \sum_{i=1}^n \hat{\Psi}_i(s) \hat{\Psi}_i(t),$$

其中

$$\begin{aligned} \hat{\Psi}_i(t) &= \eta_i \hat{\Phi}_i(t) + (1 - \eta_i) \hat{\Upsilon}_i(t), \\ \hat{\Phi}_i(t) &= \frac{M_i(t)}{\hat{S}^{(0)}(\hat{\beta}_p, t)} + \frac{\int_0^t \frac{\hat{\Theta}(t, u)}{\hat{\pi}(u)} d\hat{M}_i^c(u)}{\hat{S}^{(0)}(\hat{\beta}_p, t)} - \hat{\zeta}(\hat{\beta}_p, t)^T \hat{\Gamma}^{-1}(\hat{\beta}_p) \hat{\nu}_i(\hat{\beta}_p) \\ &\quad - \frac{n_V}{n_V} \hat{m}_p(t) \frac{\Delta_i Y_i(t) [\psi_i(\hat{\beta}_p, t) - \hat{\phi}_i(\hat{\beta}_p, t)]}{\hat{G}(T_i) \hat{S}^{(0)}(\hat{\beta}_p, t)}, \\ \hat{\Theta}(t, u) &= n^{-1} \sum_{i=1}^n Y_i(u) \hat{M}_i(t), \\ \hat{\Upsilon}_i(t) &= \frac{\hat{M}_i(t)}{\hat{S}^{(0)}(\hat{\beta}_p, t)} + \frac{\int_0^t \frac{\hat{\Theta}(t, u)}{\hat{\pi}(u)} d\hat{M}_i^c(u)}{\hat{S}^{(0)}(\hat{\beta}_p, t)} - \hat{\zeta}(\hat{\beta}_p, t)^T \hat{\Gamma}^{-1}(\hat{\beta}_p) \hat{\mu}_i(\hat{\beta}_p), \\ \hat{\zeta}(\hat{\beta}_p, t) &= \hat{m}_p(t) \frac{\hat{S}^{(1)}(\hat{\beta}_p, t)}{\hat{S}^{(0)}(\hat{\beta}_p, t)}. \end{aligned}$$

4 模拟计算

本节通过数值模拟来评价第 2 节提出的估计方法在有限样本下的可行性; 将本文中得到的估计 $\hat{\beta}_p$ 与 3 个估计量进行比较: $\hat{\beta}_F$ 和 $\hat{\beta}_V$ 分别代表基于全队列完全观测数据的估计和仅利用核实集数据得到的估计 (参见文献 [7]), $\hat{\beta}_{CW}$ 是逆概率加权估计 (参见文献 [19]).

失效时间通过生存函数为

$$S(t | W) = \frac{m(0 | W)}{m(t | W)} \exp \left\{ - \int_0^t \frac{du}{m(u | W)} \right\}$$

的分布产生, 其中 $m(t | W) = m_0(t) \exp\{\beta_{10}Z + \beta_{20}X\}$. 基准均值剩余寿命函数设定为 $m_0(t) = -0.6t + 5$ 和 $m_0(t) = (-1/3t + 1.2)^2$. 真实回归系数 $\beta_0 = (\beta_{10}, \beta_{20})$ 取为 $(0.5, -0.5)$. 删失时间来自均匀分布 $U(0, C)$, 其中通过产生近似为 10% 和 30% 删失率来确定 C 的取值.

协变量 Z 从 $U(0, 1)$ 均匀分布产生. 协变量 X 服从 Bernoulli(0.5) 分布. 辅助协变量 A 定义为 $A = Z + e$, 其中 e 服从正态分布 $N(0, \sigma^2)$, σ^2 是控制 Z 和 A 关联强度的参数. 注意到 σ^2 增加, A 信息就比 Z 少. 取 σ^2 为 0.04 和 0.64 分别代表 A 和 Z 的相关性较强和中等. 核实集比例 ρ 分别取为 30% 和 50%. 使用窗宽 $h = 4\hat{\sigma}_A n^{-1/3}$ 的 Epanechnikov 核函数 [28], 其中 $\hat{\sigma}_A$ 是辅助变量 A 的样本标准差. 选取不同的两个权函数 $H(t)$:

- (1) $H_1(t) = \sum_{i=1}^n \Delta_i I(T_i \leq t)$, 此时 (2.6) 中的积分就是在所有非删失的失效时间上的加权和;
- (2) $H_2(t)$ 是在事先选定的时间点上的以 1 为跃度的阶梯函数.

样本量 n 分别取 200 和 600. 在每种组合下, 产生 10,000 个模拟数据集. 表 1-4 给出了两个不同

表 1 样本量 $n = 200$, $H(t) = H_1(t)$, $\beta = (0.5, -0.5)$ 和 $X \sim \text{Bernoulli}(0.5)$ 时的模拟结果

$m_0(t)$	C%	ρ	σ^2	Est.	$\beta_1 = 0.5$				$\beta_2 = -0.5$					
					Bias	SD	SE	CP	Bias	SD	SE	CP		
$-0.6t + 5$	10	0.3		$\hat{\beta}_F$	0.012	0.102	0.100	0.944	-0.010	0.069	0.069	0.949		
				$\hat{\beta}_V$	0.041	0.203	0.192	0.929	-0.036	0.135	0.131	0.936		
				$\hat{\beta}_{\text{CW}}$	0.001	0.249	0.248	0.942	-0.020	0.121	0.118	0.938		
		0.04		$\hat{\beta}_p$	0.023	0.127	0.125	0.946	-0.009	0.070	0.070	0.948		
				0.64	$\hat{\beta}_p$	-0.042	0.168	0.162	0.925	-0.007	0.071	0.071	0.944	
					$\hat{\beta}_V$	0.019	0.147	0.144	0.942	-0.021	0.101	0.099	0.941	
			$\hat{\beta}_{\text{CW}}$	0.003	0.190	0.186	0.935	-0.011	0.095	0.094	0.936			
			0.04	$\hat{\beta}_p$	0.021	0.117	0.115	0.946	-0.009	0.069	0.069	0.949		
			0.64	$\hat{\beta}_p$	-0.085	0.127	0.123	0.873	-0.007	0.071	0.071	0.949		
		30		$\hat{\beta}_F$	0.014	0.122	0.120	0.943	-0.013	0.083	0.082	0.944		
				0.3	$\hat{\beta}_V$	0.049	0.255	0.230	0.911	-0.047	0.170	0.154	0.912	
					$\hat{\beta}_{\text{CW}}$	-0.003	0.277	0.282	0.947	-0.026	0.136	0.133	0.935	
	0.04		$\hat{\beta}_p$	0.023	0.150	0.149	0.949	-0.012	0.084	0.083	0.945			
			0.64	$\hat{\beta}_p$	-0.047	0.200	0.193	0.924	-0.010	0.086	0.085	0.945		
				$\hat{\beta}_V$	0.026	0.181	0.172	0.934	-0.026	0.122	0.117	0.931		
	$(-1/3t + 1.2)^2$	10		$\hat{\beta}_{\text{CW}}$	-0.001	0.208	0.203	0.943	-0.012	0.105	0.103	0.940		
				0.04	$\hat{\beta}_p$	0.023	0.140	0.138	0.945	-0.012	0.083	0.082	0.944	
				0.64	$\hat{\beta}_p$	-0.089	0.151	0.145	0.884	-0.010	0.085	0.084	0.946	
			0.3		$\hat{\beta}_F$	0.027	0.133	0.129	0.932	-0.017	0.081	0.080	0.940	
					$\hat{\beta}_V$	0.069	0.244	0.224	0.907	-0.052	0.149	0.140	0.912	
					$\hat{\beta}_{\text{CW}}$	-0.030	0.210	0.198	0.931	0.000	0.098	0.104	0.943	
				0.04	$\hat{\beta}_p$	0.019	0.158	0.152	0.939	-0.015	0.081	0.080	0.942	
					0.64	$\hat{\beta}_p$	-0.103	0.183	0.173	0.884	-0.014	0.079	0.078	0.940
						$\hat{\beta}_V$	0.045	0.180	0.172	0.922	-0.035	0.111	0.107	0.920
0.5				$\hat{\beta}_{\text{CW}}$	-0.026	0.162	0.166	0.947	0.008	0.081	0.084	0.960		
				0.04	$\hat{\beta}_p$	-0.011	0.143	0.138	0.937	-0.015	0.081	0.080	0.940	
				0.64	$\hat{\beta}_p$	-0.068	0.153	0.147	0.908	-0.016	0.078	0.077	0.938	
		30		$\hat{\beta}_F$	0.027	0.162	0.159	0.954	-0.018	0.097	0.094	0.948		
				0.3	$\hat{\beta}_V$	0.068	0.304	0.259	0.881	-0.060	0.184	0.161	0.882	
					$\hat{\beta}_{\text{CW}}$	-0.035	0.225	0.232	0.941	0.006	0.116	0.117	0.941	
0.04			$\hat{\beta}_p$	0.017	0.189	0.181	0.933	-0.016	0.098	0.095	0.931			
			0.64	$\hat{\beta}_p$	-0.114	0.216	0.200	0.881	-0.016	0.094	0.092	0.929		
				$\hat{\beta}_V$	0.048	0.220	0.201	0.906	-0.038	0.133	0.125	0.907		
			$\hat{\beta}_{\text{CW}}$	-0.035	0.198	0.192	0.935	0.012	0.144	0.140	0.937			
			0.04	$\hat{\beta}_p$	0.013	0.177	0.169	0.932	-0.017	0.098	0.095	0.929		
			0.64	$\hat{\beta}_p$	-0.075	0.183	0.171	0.903	-0.018	0.093	0.091	0.930		

$\hat{\beta}_F$: 全队列数据估计; $\hat{\beta}_V$: 核实集估计; $\hat{\beta}_{CW}$: Chen 和 Wang^[19] 逆概率加权估计; $\hat{\beta}_p$: 本文提出的估计; C%: 删失率; ρ : 核实集比例; σ^2 : 辅助协变量与主暴露变量相关性强度.

表 2 样本量 $n = 600$, $H(t) = H_1(t)$, $\beta = (0.5, -0.5)$ 和 $X \sim \text{Bernoulli}(0.5)$ 时的模拟结果

$m_0(t)$	C%	ρ	σ^2	Est.	$\beta_1 = 0.5$				$\beta_2 = -0.5$			
					Bias	SD	SE	CP	Bias	SD	SE	CP
$-0.6t + 5$	10	0.3		$\hat{\beta}_F$	0.002	0.055	0.057	0.952	-0.005	0.039	0.039	0.943
				$\hat{\beta}_V$	0.011	0.112	0.106	0.938	-0.009	0.074	0.072	0.939
				$\hat{\beta}_{CW}$	0.002	0.137	0.137	0.952	-0.009	0.062	0.065	0.959
		0.64		$\hat{\beta}_p$	0.013	0.102	0.101	0.945	-0.001	0.041	0.041	0.947
				$\hat{\beta}_V$	0.003	0.078	0.081	0.959	-0.010	0.056	0.056	0.951
				$\hat{\beta}_{CW}$	-0.004	0.101	0.104	0.954	-0.007	0.051	0.053	0.961
	30	0.64		$\hat{\beta}_p$	0.007	0.075	0.079	0.953	-0.002	0.040	0.040	0.943
				$\hat{\beta}_F$	0.004	0.067	0.068	0.945	-0.005	0.046	0.047	0.950
				$\hat{\beta}_V$	0.014	0.137	0.127	0.928	-0.012	0.089	0.086	0.937
		0.3		$\hat{\beta}_{CW}$	0.002	0.150	0.145	0.941	-0.010	0.068	0.068	0.944
				$\hat{\beta}_p$	-0.036	0.114	0.112	0.929	-0.002	0.048	0.048	0.954
				$\hat{\beta}_V$	0.007	0.097	0.097	0.948	-0.010	0.067	0.066	0.952
	10	0.5		$\hat{\beta}_{CW}$	-0.004	0.111	0.112	0.949	-0.009	0.056	0.056	0.951
				$\hat{\beta}_p$	-0.007	0.093	0.092	0.946	-0.003	0.047	0.048	0.952
				$\hat{\beta}_F$	0.025	0.066	0.069	0.944	-0.015	0.044	0.043	0.953
		0.3		$\hat{\beta}_V$	0.033	0.138	0.128	0.921	-0.022	0.082	0.080	0.921
				$\hat{\beta}_{CW}$	-0.026	0.124	0.128	0.946	0.008	0.091	0.091	0.946
				$\hat{\beta}_p$	-0.015	0.116	0.113	0.941	-0.008	0.045	0.045	0.934
$(-1/3t + 1.2)^2$	30	0.64		$\hat{\beta}_V$	0.028	0.094	0.098	0.950	-0.020	0.063	0.062	0.924
				$\hat{\beta}_{CW}$	-0.030	0.106	0.107	0.938	0.009	0.080	0.081	0.932
				$\hat{\beta}_p$	0.006	0.089	0.093	0.948	-0.010	0.045	0.044	0.932
		0.3		$\hat{\beta}_F$	0.025	0.084	0.083	0.931	-0.015	0.051	0.051	0.933
				$\hat{\beta}_V$	0.032	0.168	0.151	0.908	-0.024	0.098	0.093	0.924
				$\hat{\beta}_{CW}$	-0.031	0.154	0.153	0.940	0.013	0.112	0.110	0.935
	10	0.64		$\hat{\beta}_p$	-0.023	0.141	0.133	0.921	-0.009	0.052	0.053	0.945
				$\hat{\beta}_V$	0.031	0.117	0.117	0.940	-0.021	0.076	0.073	0.930
				$\hat{\beta}_{CW}$	-0.033	0.123	0.127	0.944	0.011	0.090	0.092	0.937
		0.5		$\hat{\beta}_p$	0.012	0.111	0.111	0.949	-0.011	0.051	0.053	0.943

表格注解同表 1.

权函数的模拟结果. 在表中, Bias 是点估计偏差的样本平均值, SD 是 10,000 个点估计的样本标准差, SE 给出的是估计标准差的均值, CP 表示 95% 置信区间的覆盖率.

从表 1-4 可以看出, (i) 当 $\sigma^2 = 0.04$ 时, 提出的估计在所考虑的情形下均为近似无偏的. 当 $n = 200$ 和 $\sigma^2 = 0.64$ 时, 提出的估计 $\hat{\beta}_p$ 不够准确. 然而, 当样本量提高到 $n = 600$ 时, 偏差急剧降低. 我们的方法使用非参数平滑技巧来估计条件期望. 当辅助变量信息不足时, 需要较多的样本来得到条件期望的准确估计值. (ii) 提出的估计标准差 SE 是准确的, 基于正态近似的 95% 经验覆盖率是非常合理的. (iii) $\hat{\beta}_p$ 没有全队列数据估计 $\hat{\beta}_F$ 有效, 但是比核实集估计 $\hat{\beta}_V$ 和 $\hat{\beta}_{CW}$ 都更有效. 对于 β_1 , 即主协变量 Z 的影响, 当删失率 (C%) 降低, 核实集比例 (ρ) 提高或者辅助协变量提供更多信息 (即 σ^2 减少),

表 3 样本量 $n = 200$, $H(t) = H_2(t)$, $\beta = (0.5, -0.5)$ 和 $X \sim \text{Bernoulli}(0.5)$ 时的模拟结果

$m_0(t)$	C%	ρ	σ^2	Est.	$\beta_1 = 0.5$				$\beta_2 = -0.5$			
					Bias	SD	SE	CP	Bias	SD	SE	CP
$-0.6t + 5$	10	0.3		$\hat{\beta}_F$	0.012	0.104	0.102	0.945	-0.011	0.071	0.070	0.950
				$\hat{\beta}_V$	0.042	0.207	0.194	0.928	-0.037	0.138	0.133	0.936
				0.04 $\hat{\beta}_p$	0.023	0.130	0.127	0.946	-0.009	0.071	0.071	0.950
				0.64 $\hat{\beta}_p$	-0.042	0.171	0.165	0.924	-0.007	0.073	0.073	0.950
		0.5		$\hat{\beta}_V$	0.019	0.150	0.146	0.942	-0.022	0.103	0.101	0.943
				0.04 $\hat{\beta}_p$	0.021	0.119	0.117	0.945	-0.010	0.071	0.071	0.950
				0.64 $\hat{\beta}_p$	-0.086	0.129	0.126	0.876	-0.007	0.072	0.072	0.951
		30		$\hat{\beta}_F$	0.015	0.124	0.122	0.942	-0.014	0.084	0.083	0.943
				$\hat{\beta}_V$	0.049	0.257	0.232	0.911	-0.048	0.172	0.156	0.913
				0.04 $\hat{\beta}_p$	0.023	0.152	0.151	0.948	-0.012	0.085	0.084	0.944
				0.64 $\hat{\beta}_p$	-0.048	0.203	0.195	0.926	-0.010	0.087	0.086	0.943
	$(-1/3t + 1.2)^2$	0.3		$\hat{\beta}_V$	0.026	0.183	0.174	0.934	-0.027	0.124	0.119	0.931
				0.04 $\hat{\beta}_p$	0.023	0.141	0.139	0.945	-0.013	0.085	0.084	0.943
				0.64 $\hat{\beta}_p$	-0.090	0.153	0.147	0.882	-0.011	0.086	0.085	0.944
		0.5		$\hat{\beta}_F$	0.038	0.130	0.132	0.946	-0.023	0.080	0.078	0.934
				$\hat{\beta}_V$	0.070	0.251	0.229	0.906	-0.053	0.154	0.144	0.911
				0.04 $\hat{\beta}_p$	0.003	0.150	0.143	0.936	-0.019	0.081	0.079	0.936
				0.64 $\hat{\beta}_p$	-0.103	0.189	0.178	0.889	-0.015	0.082	0.081	0.940
		30		$\hat{\beta}_V$	0.046	0.186	0.176	0.922	-0.036	0.115	0.111	0.921
				0.04 $\hat{\beta}_p$	-0.001	0.140	0.134	0.939	-0.020	0.080	0.079	0.935
				0.64 $\hat{\beta}_p$	-0.069	0.158	0.151	0.907	-0.017	0.081	0.080	0.939
		0.5		$\hat{\beta}_F$	0.038	0.155	0.157	0.952	-0.025	0.094	0.091	0.947
				$\hat{\beta}_V$	0.069	0.310	0.262	0.880	-0.061	0.188	0.163	0.881
				0.04 $\hat{\beta}_p$	0.027	0.182	0.174	0.932	-0.022	0.095	0.092	0.931
				0.64 $\hat{\beta}_p$	-0.116	0.222	0.205	0.880	-0.017	0.097	0.094	0.931
$(-1/3t + 1.2)^2$	10	0.3		$\hat{\beta}_V$	0.049	0.225	0.204	0.904	-0.039	0.137	0.127	0.908
				0.04 $\hat{\beta}_p$	0.023	0.170	0.168	0.941	-0.022	0.095	0.092	0.947
				0.64 $\hat{\beta}_p$	-0.075	0.188	0.175	0.903	-0.018	0.096	0.093	0.930

表格注解同表 1.

给出的估计 $\hat{\beta}_p$ 的效率提升显著. 然而, 完全观测协变量 X 对 β_2 的影响, 在所有模拟环境中, $\hat{\beta}_p$ 和 $\hat{\beta}_F$ 几乎一样有效, 效率变化不大. 进一步可以看出, 在所有考虑的情形下, 两个不同的权函数 $H_1(t)$ 和 $H_2(t)$, 模拟结果相似.

表 5 给出在不同的样本容量、不同的删失率、核实集比例和权函数下, $\hat{\beta}_p$ 相对于核实集估计 $\hat{\beta}_V$ 和 $\hat{\beta}_{CW}$ 的相对效率. 相对效率 (RE) 通过对应估计的均方误差的比值来计算. 从表 5 可以看出, 对于固定的 n , 当 σ^2 或 ρ 减少, 或 C% 增加, $\hat{\beta}_p$ 相对 $\hat{\beta}_V$ 效率增加.

同样地, 对于固定的 n , 当 σ^2 或 ρ 减少, $\hat{\beta}_p$ 相对于 $\hat{\beta}_{CW}$ 的效率增加. 这些结果表明, 对于较低的核实集比例, 辅助协变量信息充足或者删失率较高时, 我们的估计比 Chen 和 Wang^[19] 的估计有效.

表 4 样本量 $n = 600$, $H(t) = H_2(t)$, $\beta = (0.5, -0.5)$ 和 $X \sim \text{Bernoulli}(0.5)$ 时的模拟结果

$m_0(t)$	C%	ρ	σ^2	Est.	$\beta_1 = 0.5$				$\beta_2 = -0.5$			
					Bias	SD	SE	CP	Bias	SD	SE	CP
$-0.6t + 5$	10	0.3		$\hat{\beta}_F$	0.002	0.055	0.057	0.952	-0.005	0.039	0.039	0.943
				$\hat{\beta}_V$	0.011	0.114	0.108	0.943	-0.009	0.076	0.074	0.944
				$\hat{\beta}_p$	0.013	0.103	0.102	0.945	-0.002	0.042	0.041	0.952
		0.5	0.64	$\hat{\beta}_V$	0.004	0.080	0.083	0.954	-0.009	0.057	0.057	0.945
				$\hat{\beta}_p$	0.007	0.078	0.081	0.951	-0.003	0.041	0.041	0.947
				$\hat{\beta}_F$	0.004	0.067	0.068	0.945	-0.005	0.046	0.047	0.950
	30	0.3		$\hat{\beta}_V$	0.014	0.139	0.128	0.929	-0.012	0.090	0.087	0.935
				$\hat{\beta}_p$	-0.036	0.117	0.113	0.952	-0.002	0.048	0.049	0.955
				$\hat{\beta}_V$	0.007	0.097	0.098	0.952	-0.011	0.068	0.067	0.953
		0.5	0.64	$\hat{\beta}_p$	-0.008	0.095	0.094	0.943	-0.003	0.047	0.049	0.952
				$\hat{\beta}_F$	0.025	0.066	0.069	0.944	-0.015	0.044	0.043	0.953
				$\hat{\beta}_V$	0.033	0.141	0.131	0.925	-0.023	0.085	0.082	0.915
$(-1/3t + 1.2)^2$	10	0.3		$\hat{\beta}_p$	-0.016	0.119	0.116	0.938	-0.008	0.047	0.046	0.936
				$\hat{\beta}_V$	0.029	0.100	0.102	0.941	-0.022	0.065	0.064	0.923
				$\hat{\beta}_p$	0.004	0.092	0.095	0.954	-0.010	0.046	0.046	0.934
		0.5	0.64	$\hat{\beta}_F$	0.025	0.084	0.083	0.931	-0.015	0.051	0.051	0.933
				$\hat{\beta}_V$	0.035	0.172	0.155	0.904	-0.024	0.102	0.096	0.921
				$\hat{\beta}_p$	-0.023	0.146	0.147	0.947	-0.009	0.054	0.055	0.943
	30	0.3		$\hat{\beta}_V$	0.031	0.120	0.120	0.941	-0.021	0.079	0.075	0.924
				$\hat{\beta}_p$	0.013	0.116	0.114	0.949	-0.011	0.053	0.054	0.944
		0.5	0.64									

表格注解同表 1.

表 5 当 $X \sim \text{Bernoulli}(0.5)$ 时, $\hat{\beta}_p$ 中 β_1 与核实集估计 $\hat{\beta}_V$ 和 $\hat{\beta}_{\text{CW}}$ 相比的相对效率

Beta	$m_0(t)$	n	C%	ρ	σ^2	$\hat{\beta}_V$		$\hat{\beta}_{\text{CW}}$
						$H_1(t)$	$H_2(t)$	
(0.5, -0.5)	$-0.6t + 5$	200	10	30	0.04	2.555	2.535	3.844
					0.64	1.460	1.465	2.197
					50	1.579	1.589	2.637
					0.64	1.340	1.352	2.238
					0.04	2.890	2.859	3.410
					0.64	1.626	1.603	1.918
		200	30	30	0.04	1.671	1.684	2.207
					0.64	1.437	1.431	1.897
					0.04	1.206	1.225	1.804
					0.64	1.082	1.052	1.814
		600	10	30	0.64	1.444	1.411	1.731
					0.64	1.088	1.043	1.425

								续表
Beta	$m_0(t)$	n	C%	ρ	σ^2	$\hat{\beta}_V$		$\hat{\beta}_{CW}$
						$H_1(t)$	$H_2(t)$	
(0.5, -0.5)	$(-1/3t + 1.2)^2$	200	10	30	0.04	2.385	2.800	1.767
					0.64	1.778	1.763	1.317
					50	1.584	1.765	1.283
				50	0.04	1.384	1.386	1.121
					0.64	2.587	2.901	1.417
					0.64	1.981	1.950	1.085
		200	30	30	0.04	1.545	1.752	1.251
					0.64	1.445	1.432	1.171
					50	1.415	1.404	1.413
				50	0.64	1.116	1.181	1.419
					0.64	1.420	1.387	1.193
					0.64	1.111	1.070	1.228

5 实例分析

用 Mayo Clinic 临床试验中关于原发性胆汁性肝硬化 (primary biliary cirrhosis, PBC) 的数据集来说明所提出的方法在实际中的应用, 此试验是 1974 年 1 月至 1984 年 5 月间进行的一项随机双盲试验. PBC 是一种慢性致死性胆汁淤积性肝病, 其组织学特征为非化脓性破坏小叶间胆管, 最终导致晚期肝纤维化、肝硬化和肝功能衰竭. PBC 的发病机制尚不清楚, 但免疫系统失调和遗传因素似乎都是导致疾病的原因. 通过血清中抗线粒体抗体的是否存在来诊断疾病. 受影响的患者往往有血液检查异常, 如血清肝生化指标无症状上升. 更详细的临床背景的描述、更深入的试验讨论及协变量记录参见文献 [29]. Fan 和 Wang^[30] 使用这些数据来阐明他们处理辅助协变量的方法.

在这项临床试验中, 共有 412 名 PBC 患者符合资格标准. 失效时间是登记到死亡之间的天数. PBC 感兴趣的风险因素变量有患者性别 (Sex, 男性取 0 和女性取 1)、患者年龄 (Age, 以年为单位)、血清总胆固醇 (Chol) 和疾病的病理阶段 (Stage). 在此数据集中, 大约有 31% 的胆固醇数值缺失. 丢弃这些个体的观测会导致效率损失. 另一方面, 血清胆红素 (Bili) 所有患者无缺失值. 根据临床文献, 将胆固醇和胆红素作对数变换. 胆固醇和血清胆红素均为连续变量值, 测量单位毫克/分升. 首先, 探究血清胆红素与胆固醇的关系. 给定 $\log(\text{Chol})$, $\log(\text{Bili})$ 的一个线性模型拟合可得斜率估计为 0.19055 ($p \leq 0.0001$, $R^2 = 0.2016$), 这表明了 $\log(\text{Bili})$ 和 $\log(\text{Chol})$ 的共线性关系. 因此, 这里血清胆红素可作为胆固醇的辅助变量. 根据前面章节的记号, 可知 $Z = \log(\text{Chol})$, $\mathbf{X} = (\text{Sex}, \text{Age}, \text{Stage})$, $A = \log(\text{Bili})$. 给定 $\log(\text{Bili})$, 为了检查 $\log(\text{Chol})$ 与 $(\text{Sex}, \text{Age}, \text{Stage})$ 是否条件独立, 将变量 $(\text{Sex}, \text{Age}, \text{Stage})$ 增加到 $\log(\text{Chol})$ 关于 $\log(\text{Bili})$ 的线性模型. 得到的结果表明是条件独立的. 因此, 令 $W^* = \log(\text{Bili})$.

为了分析这个数据集, 考虑模型 $m(t | Z_i, X_i) = m_0(t)r_i(\beta, t)$, 其中

$$r_i(\beta, t) = \begin{cases} e^{\beta_{10}Z_i + \beta_{20}^T \mathbf{X}_i}, & \text{当 Chol 可观测,} \\ \phi_i(\beta, t), & \text{当 Chol 缺失,} \end{cases}$$

这里 $\phi_i(\beta, t) = E[e^{\beta_{10}Z_i} | Y_i(t) = 1, W_i^*]e^{\beta_{20}^T \mathbf{X}_i}$. 对于这个数据, 选取 $H(t) = \sum_{i=1}^n \Delta_i I(T_i \leq t)$.

表 6 PBC 数据分析结果

估计方法		log(Chol)	Age	Sex	Stage
$\hat{\beta}_p$	Est.	-0.41709	-0.00002	-0.27517	-0.16128
	SE	0.15540	0.00002	0.17587	0.07803
	p 值	0.00728	0.29299	0.11769	0.03874
$\hat{\beta}_v$	Est.	-0.48752	-0.00004	-0.30345	-0.15173
	SE	0.16363	0.00002	0.17767	0.08686
	p 值	0.00289	0.13153	0.08765	0.08068
$\hat{\beta}_{CW}$	Est.	-0.33432	-0.00005	0.12896	-0.32738
	SE	0.16514	0.00001	0.12341	0.04534
	p 值	0.04292	0.00003	0.29604	<0.00001

表 6 给出了实际数据分析的结果. 分析结果表明, 在提出的方法中, 变量 $\log(\text{Chol})$ 和 Stage 在统计上是显著的. 在核实集方法中, 只有 $\log(\text{Chol})$ 是显著的. 在 Chen 和 Wang^[19] 的方法中, $\log(\text{Chol})$ 、Age 和 Stage 是统计显著的. 通过这 3 种方法的比较, 相对于核实集方法和 Chen 和 Wang^[19] 的方法, 我们提出的方法均有效率增加. 例如, $\log(\text{Chol})$ 对心力衰竭时间的影响, 我们提出的估计的标准误为 0.155, 核实集估计方法为 0.163, Chen 和 Wang^[19] 的方法为 0.165. 因此, 我们提出的估计是一个更精确的估计. 本例证实利用辅助信息, 所提出的方法提高了研究的统计功效, 如果只分析核实集, 这些功效可能会丢失.

6 总结

当主协变量只能对核实集观测, 且缺失机制可忽略, 而一些辅助协变量则对研究队列中的全部个体可以观测的情形下, 本文提出了比例均值剩余寿命模型的估计方法. 利用辅助变量的信息来提高研究的效率. 所提出方法的一个重要特点是关于主协变量与辅助变量之间相关性是非参数的. 得到的参数估计是相合和渐近正态的, 协方差矩阵可通过其相合估计来得到. 本文也给出基准均值剩余寿命函数的一个相合估计. 模拟结果表明所提出的方法优于只用核实集来分析的方法. 模拟研究发现, 当核实集比例提高时, 所提出方法可以获得更高的统计效率. 最后把所提出的方法应用于 Mayo 临床试验的原发性胆汁性肝硬化 (PBC) 的数据集的分析.

对该方法的局限性做出如下说明: 当辅助变量关于真实暴露变量的信息不足时 (例如, 模拟中 σ^2 比较大时), 提出的估计将趋于低估真实参数, 方差估计低估真实方差 (如 $n = 200$, $\sigma^2 = 0.64$, $\rho = 30\%$). 增加样本量或者提高核实集比例可以缓解这些问题 (如 $n = 600$, $\sigma^2 = 0.64$, $\rho = 50\%$). 因为本文的方法使用核平滑技巧, 需要更多的样本量才能获得更好的表现.

实际应用中, 我们会选择一类与感兴趣真实协变量相关并且比较容易测量的变量作为辅助变量. 多元辅助变量在实际中也会经常遇到 (参见文献 [18]), 本文所提出的方法也可以用来处理多元辅助变量的情形. 所提出的方法在 $\mathbf{W}^*(t)$ 低维的情形下非常有用. 众所周知, 当维数增加时, 核光滑的收敛速度会下降很快, 即为维数灾难. 当将所提方法用在高维的情形时, 这是一个必须面临的问题. 降维方法具有部分消除维数灾难的局限性 (参见文献 [31–34]). 这方面将来需要进一步研究.

本文利用比例均值剩余寿命函数对失效时间进行建模, 其他半参数模型, 如加性均值剩余寿命模型和转换均值剩余寿命模型, 也可以考虑. 另外, 在一些应用中, 删失可能依赖于剩余寿命, Sun 和

Zhang^[7] 考虑了删失依赖于协变量的情形, 但假定了给定协变量的条件下, 失效时间和删失时间独立, 我们将来的研究中也可以考虑将所提出的方法推广到删失依赖协变量的情形.

致谢 作者对审稿人的宝贵意见表示衷心感谢.

参考文献

- 1 Oakes D, Dasu T. A note on residual life. *Biometrika*, 1990, 77: 409–410
- 2 Maguluri G, Zhang C H. Estimation in the mean residual life regression model. *J R Stat Soc Ser B Stat Methodol*, 1994, 56: 477–489
- 3 Chen Y Q, Cheng S C. Semiparametric regression analysis of mean residual life with censored survival data. *Biometrika*, 2005, 92: 19–29
- 4 Chen Y Q, Jewell N P, Lei X, et al. Semiparametric estimation of proportional mean residual life model in presence of censoring. *Biometrics*, 2005, 61: 170–178
- 5 Chen Y Q, Cheng S C. Linear life expectancy regression with censored data. *Biometrika*, 2006, 93: 303–313
- 6 Chen Y Q. Additive expectancy regression. *J Amer Statist Assoc*, 2007, 102: 153–166
- 7 Sun L, Zhang Z. A class of transformed mean residual life models with censored survival data. *J Amer Statist Assoc*, 2009, 104: 803–815
- 8 Sun L, Song X, Zhang Z. Mean residual life models with time-dependent coefficients under right censoring. *Biometrika*, 2012, 99: 185–197
- 9 Cai J, He H, Song X, et al. An additive-multiplicative mean residual life model for right-censored data. *Biom J*, 2017, 59: 579–592
- 10 Zhou H, Pepe M S. Auxiliary covariate data in failure time regression. *Biometrika*, 1995, 82: 139–149
- 11 Zhou H, Wang C Y. Failure time regression with continuous covariates measured with error. *J R Stat Soc Ser B Stat Methodol*, 2000, 62: 657–665
- 12 Prentice R L. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 1982, 69: 331–342
- 13 Liu Y, Zhou H, Cai J. Estimated pseudopartial-likelihood method for correlated failure time data with auxiliary covariates. *Biometrics*, 2009, 65: 1184–1193
- 14 Liu Y, Wu Y, Zhou H. Multivariate failure times regression with a continuous auxiliary covariate. *J Multivariate Anal*, 2010, 101: 679–691
- 15 Liu Y, Yuan Z, Cai J, et al. Marginal hazard regression for correlated failure time data with auxiliary covariates. *Lifetime Data Anal*, 2012, 18: 116–138
- 16 杨青龙, 马学敏, 袁中尚. 带有辅助信息的多元失效时间数据的估计标准部分似然方法. *中国科学: 数学*, 2012, 42: 545–562
- 17 Kulich M, Lin D Y. Additive hazards regression with covariate measurement error. *J Amer Statist Assoc*, 2000, 95: 238–248
- 18 Jiang J, Zhou H. Additive hazard regression with auxiliary covariates. *Biometrika*, 2007, 94: 359–369
- 19 Chen X, Wang Q. Semiparametric proportional mean residual life model with covariates missing at random. *J Nonparametr Stat*, 2013, 25: 647–663
- 20 Fleming T R, Harrington D P. *Counting Processes and Survival Analysis*. New York: John Wiley & Sons, 1991
- 21 Nadaraya E A. On estimating regression. *Theory Probab Appl*, 1964, 10: 186–190
- 22 Watson G S. Smooth regression analysis. *Sankya Ser A*, 1964, 26: 359–372
- 23 Foutz R V. On the unique consistent solution to the likelihood equations. *J Amer Statist Assoc*, 1977, 72: 147–148
- 24 Pollard D. *Empirical Processes: Theories and Applications*. Hayward: Institute of Mathematical Statistics, 1990
- 25 Biliyas Y, Gu M, Ying Z. Towards a general asymptotic theory for Cox model with staggered entry. *Ann Statist*, 1997, 25: 662–682
- 26 van der Vaart A W, Wellner J A. *Empirical Processes with Applications to Statistics*. New York: John Wiley & Sons, 1996
- 27 van der Vaart A W. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 1998
- 28 Eubank R L. *Nonparametric Regression and Spline Smoothing*, 2nd ed. New York: Marcel Dekker, 1999
- 29 Dickson E R, Grambsch P M, Fleming T R, et al. Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, 1989, 10: 1–7
- 30 Fan Z Z, Wang X F. Marginal hazards model for multivariate failure time data with auxiliary covariates. *J Nonparametr Stat*, 2009, 21: 771–786

- 31 Chen C H, Li K C, Wang J L. Dimension reduction for censored regression data. *Ann Statist*, 1999, 27: 1–23
- 32 Xia Y, Zhang D, Xu J. Dimension reduction and semiparametric estimation of survival models. *J Amer Statist Assoc*, 2010, 105: 278–290
- 33 Nadkarni N V, Zhao Y, Kosorok M R. Inverse regression estimation for censored data. *J Amer Statist Assoc*, 2011, 106: 178–190
- 34 Lu W, Li L. Sufficient dimension reduction for censored regressions. *Biometrics*, 2011, 67: 513–523

附录 A 定理 3.1 和 3.2 的证明

引理 A.1 若正则条件 (C1)–(C7) 成立, 则存在 β_0 的一个邻域 $\mathcal{B}$, 使得

$$\sup_{t \in [0, \tau], \beta \in \mathcal{B}} \|\hat{\phi}_i^{(k)}(\beta, t) - \phi_i^{(k)}(\beta, t)\| \xrightarrow{\text{a.s.}} 0 \quad (k = 0, 1, 2),$$

$$\sup_{t \in [0, \tau], \beta \in \mathcal{B}} \|\hat{S}^{(k)}(\beta, t) - s^{(k)}(\beta, t)\| \xrightarrow{\text{a.s.}} 0 \quad (k = 0, \dots, 4).$$

进一步地, $s^{(k)}(\beta, t)$ ($k = 0, \dots, 4$) 是 $\beta \in \mathcal{B}$ 上的连续函数, 对于 $t \in [0, \tau]$, 一致地有 $s^{(k)}(\beta, t)$ ($k = 0, \dots, 4$) 在 $\mathcal{B} \times [0, \tau]$ 上有界, $s^{(0)}(\beta, t)$ 在 $\mathcal{B} \times [0, \tau]$ 上有界且大于 0.

引理 A.1 的证明可直接参见文献 [13] 中的证明, 本文略去相关证明.

定理 3.1 的证明 先证明 $\hat{\beta}_p$ 的相合性, 再证明 $\hat{\beta}_p$ 的渐近正态性.

根据 Foutz^[23] 的结果, 只要保证下面 4 条成立, 就能说明 $\hat{\beta}_p$ 是 β_0 的相合估计:

- (I) $n^{-1} \frac{\partial \hat{U}(\beta)}{\partial \beta}$ 在 β_0 的一个开邻域 $\mathcal{B}$ 内存在且连续;
- (II) $n^{-1} \frac{\partial \hat{U}(\beta_0)}{\partial \beta_0}$ 以概率趋近于 1 且为负定的;
- (III) $n^{-1} \frac{\partial \hat{U}(\beta)}{\partial \beta}$ 在 β_0 的一个开邻域内一致以概率收敛到一个已知函数, 记为 $-\Gamma(\beta)$;
- (IV) $n^{-1} \hat{U}(\beta_0)$ 以概率收敛到 0.

显然, 由于 $\hat{r}_i^{(l)}(\beta, t)$ ($l = 0, 1, 2$) 的连续性, (I) 成立. 为了说明 (II) 和 (III) 成立, 将 $n^{-1} \frac{\partial \hat{U}(\beta)}{\partial \beta}$ 分解成如下两部分:

$$\begin{aligned} n^{-1} \frac{\partial \hat{U}(\beta)}{\partial \beta} &= n^{-1} \sum_{i=1}^n \int_0^\tau \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} \left[\frac{\hat{r}_i^{(2)}(\beta, t)}{\hat{r}_i(\beta, t)} - \left(\frac{\hat{r}_i^{(1)}(\beta, t)}{\hat{r}_i(\beta, t)} \right)^{\otimes 2} - \frac{\hat{S}^{(2)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} \right. \\ &\quad \left. + \left(\frac{\hat{S}^{(1)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} \right)^{\otimes 2} \right] M_i(t) dH(t) \\ &\quad + n^{-1} \sum_{i=1}^n \int_0^\tau \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} \left[\frac{\hat{r}_i^{(2)}(\beta, t)}{\hat{r}_i(\beta, t)} - \left(\frac{\hat{r}_i^{(1)}(\beta, t)}{\hat{r}_i(\beta, t)} \right)^{\otimes 2} - \frac{\hat{S}^{(2)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} \right. \\ &\quad \left. + \left(\frac{\hat{S}^{(1)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} \right)^{\otimes 2} \right] r_i(\beta_0, t) m_0(t) dH(t) \\ &=: \Omega_{1n}(\beta) + \Omega_{2n}(\beta). \end{aligned}$$

由引理 A.1 和大数定律, 可以证明 $\Omega_{1n}(\beta)$ 以概率收敛到 0. 简单计算后, 第 2 项 $\Omega_{2n}(\beta)$ 可写成

$$\int_0^\tau \left[\hat{S}^{(3)}(\beta, t) - \hat{S}^{(4)}(\beta, t) - \frac{\hat{S}^{(2)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} S^{(0)}(\beta_0, t) + \left(\frac{\hat{S}^{(1)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} \right)^{\otimes 2} S^{(0)}(\beta_0, t) \right] m_0(t) dH(t).$$

由引理 A.1, $\Omega_{2n}(\beta)$ 以概率收敛到 $-\Gamma(\beta)$, 其形式为

$$\int_0^\tau \left[s^{(3)}(\beta, t) - s^{(4)}(\beta, t) - \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} s^{(0)}(\beta_0, t) + \left(\frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right)^{\otimes 2} s^{(0)}(\beta_0, t) \right] m_0(t) d\tilde{H}(t).$$

基于上述结果, (III) 成立.

当 $\beta = \beta_0$, $s^{(3)}(\beta_0, t) = s^{(2)}(\beta_0, t)$ 时,

$$-\Gamma(\beta_0) = \int_0^\tau \left[\frac{s^{(4)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} - \left(\frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right)^{\otimes 2} \right] s^{(0)}(\beta_0, t) m_0(t) d\tilde{H}(t).$$

从而, 条件 (II) 满足.

对于条件 (IV), 作类似分解, 将 $n^{-1}\hat{U}(\beta_0)$ 分成 2 项:

$$\begin{aligned} n^{-1} \int_0^\tau \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} r_i(\beta_0, t) \left[\frac{\hat{r}_i^{(1)}(\beta_0, t)}{\hat{r}_i(\beta_0, t)} - \frac{\hat{S}^{(1)}(\beta_0, t)}{\hat{S}^{(0)}(\beta_0, t)} \right] m_0(t) dH(t) \\ = n^{-1} \int_0^\tau \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} \frac{\hat{r}_i^{(1)}(\beta_0, t)}{\hat{r}_i(\beta_0, t)} r_i(\beta_0, t) m_0(t) dH(t) \\ - \int_0^\tau \frac{\hat{S}^{(1)}(\beta_0, t)}{\hat{S}^{(0)}(\beta_0, t)} n^{-1} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} r_i(\beta_0, t) m_0(t) dH(t) \\ =: L_{1n} - L_{2n}. \end{aligned}$$

由引理 A.1 和一致强大数定律^[24], 可得

$$L_{1n} \xrightarrow{P} \int_0^\tau s^{(1)}(\beta_0, t) m_0(t) d\tilde{H}(t), \quad L_{2n} \xrightarrow{P} \int_0^\tau s^{(1)}(\beta_0, t) m_0(t) d\tilde{H}(t).$$

从而, (IV) 成立.

因此, $\hat{\beta}_p$ 以概率收敛到 β_0 , 相合性得证.

由 $\hat{U}(\beta)$ 在 β_0 附近的 Taylor 展式, 可得

$$n^{-1/2}\hat{U}(\beta_0) = \left(-n^{-1} \frac{\partial \hat{U}(\beta^*)}{\partial \beta} \right) n^{1/2}(\hat{\beta}_p - \beta_0),$$

其中 β^* 介于 $\hat{\beta}_p$ 与 β_0 之间. $\hat{\beta}_p$ 的渐近正态性可从 (III) 和 $n^{-1/2}\hat{U}(\beta_0)$ 的渐近正态性导出.

为了得到 $n^{-1/2}\hat{U}(\beta_0)$ 的渐近正态性, 我们可将其表示成

$$\begin{aligned} n^{-1/2}\hat{U}(\beta_0) &= n^{-1/2} \int_0^\tau \sum_{i=1}^n M_i(t) \frac{G(T_i)}{\hat{G}(T_i)} \left[\frac{\hat{r}_i^{(1)}(\beta_0, t)}{\hat{r}_i(\beta_0, t)} - \frac{\hat{S}^{(1)}(\beta_0, t)}{\hat{S}^{(0)}(\beta_0, t)} \right] dH(t) \\ &+ n^{-1/2} \int_0^\tau \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} r_i(\beta_0, t) \left[\frac{\hat{r}_i^{(1)}(\beta_0, t)}{\hat{r}_i(\beta_0, t)} - \frac{\hat{S}^{(1)}(\beta_0, t)}{\hat{S}^{(0)}(\beta_0, t)} \right] m_0(t) dH(t). \end{aligned} \quad (A.1)$$

对 $\hat{r}^{(1)}\hat{G}^{-1}/\hat{r}$ 在 $r^{(1)}G^{-1}/r$ 使用一阶 Taylor 展式

$$\frac{x}{y} = \frac{x_0}{y_0} + \frac{x - x_0}{y_0} - \frac{(y - y_0)x_0}{y_0^2} + O((x - x_0)^2 + (y - y_0)^2),$$

(A.1) 的第 2 项可写成

$$-n^{-1/2} \sum_{i=1}^n \int_0^\tau \frac{\Delta_i Y_i(t)}{G(T_i)} \left[\frac{r_i^{(1)}(\beta_0, t)}{r_i(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right] [\hat{r}_i(\beta_0, t) - r_i(\beta_0, t)] m_0(t) dH(t) + o_p(1),$$

使用 Zhou 和 Wang^[11] 类似推导, 上式可证明以概率收敛到

$$n^{-1/2} \frac{n_V}{n_V} \sum_{i \in V} \int_0^\tau \frac{\Delta_i Y_i(t)}{G(T_i)} \left[\frac{r_i^{(1)}(\beta_0, t)}{r_i(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right] [\psi_i(\beta_0, t) - \phi_i(\beta_0, t)] m_0(t) d\tilde{H}(t).$$

从而, (A.1) 的第 2 项渐近等于 $-n^{-1/2} \frac{1-\rho}{\rho} \sum_{i \in V} \mathbf{Q}_i(\beta_0)$.

由引理 A.1, (A.1) 的第 1 项等于

$$n^{-1/2} \sum_{i=1}^n \left\{ \int_0^\tau M_i(t) \left[\frac{r_i^{(1)}(\beta_0, t)}{r_i(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right] d\tilde{H}(t) + \int_0^\tau \frac{\mathbf{R}^*(\beta_0, t)}{\pi(t)} dM_i^c(t) \right\} + o_p(1).$$

因此, 基于上面的结果,

$$\begin{aligned} n^{-1/2} \hat{\mathbf{U}}(\beta_0) &= n^{-1/2} \sum_{i=1}^n \left\{ \int_0^\tau M_i(t) \left[\frac{r_i^{(1)}(\beta_0, t)}{r_i(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right] d\tilde{H}(t) + \int_0^\tau \frac{\mathbf{R}^*(\beta_0, t)}{\pi(t)} dM_i^c(t) \right\} \\ &\quad - n^{-1/2} \frac{1-\rho}{\rho} \sum_{i \in V} \mathbf{Q}_i(\beta_0) + o_p(1) \\ &= n^{-1/2} \sum_{i \in \bar{V}} \left\{ \int_0^\tau M_i(t) \left[\frac{r_i^{(1)}(\beta_0, t)}{r_i(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right] d\tilde{H}(t) + \int_0^\tau \frac{\mathbf{R}^*(\beta_0, t)}{\pi(t)} dM_i^c(t) \right\} \\ &\quad + n^{-1/2} \sum_{i \in V} \left\{ \int_0^\tau M_i(t) \left[\frac{r_i^{(1)}(\beta_0, t)}{r_i(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right] d\tilde{H}(t) \right. \\ &\quad \left. + \int_0^\tau \frac{\mathbf{R}^*(\beta_0, t)}{\pi(t)} dM_i^c(t) - \frac{1-\rho}{\rho} \mathbf{Q}_i(\beta_0) \right\} + o_p(1). \end{aligned}$$

由多元分布的中心极限定理, $n_V^{-1/2} \sum_{i \in \bar{V}} \boldsymbol{\mu}_i(\beta_0)$ 和 $n_V^{-1/2} \sum_{i \in V} \boldsymbol{\nu}_i(\beta_0)$ 均以概率收敛到零均值的正态分布, 方差阵分别为 $E(\boldsymbol{\mu}_i(\beta_0) \boldsymbol{\mu}_i(\beta_0)^T)$ 和 $E(\boldsymbol{\nu}_i(\beta_0) \boldsymbol{\nu}_i(\beta_0)^T)$. 又因这两项是独立的, 于是得 $n^{-1/2} \hat{\mathbf{U}}(\beta_0)$ 是渐近正态的, 且均值为 0, 方差阵为 $\boldsymbol{\Sigma}_1(\beta_0) + \boldsymbol{\Sigma}_2(\beta_0)$. 再结合上面 Taylor 展式即可得 $\hat{\beta}_p$ 的渐近正态性结果. $\square$

定理 3.2 的证明 注意到

$$\begin{aligned} \sup_{t \in [0, \tau]} |\hat{m}_p(t) - m_0(t)| &\leq \sup_{t \in [0, \tau]} \left| \frac{n^{-1} \sum_{i=1}^n \frac{G(T_i)}{\hat{G}(T_i)} M_i(t)}{\hat{S}^{(0)}(\hat{\beta}_p, t)} \right| + \sup_{t \in [0, \tau]} \left| m_0(t) \frac{n^{-1} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} [r_i(\beta_0, t) - r_i(\hat{\beta}_p, t)]}{\hat{S}^{(0)}(\hat{\beta}_p, t)} \right| \\ &\quad + \sup_{t \in [0, \tau]} \left| m_0(t) \frac{n^{-1} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} [r_i(\hat{\beta}_p, t) - \hat{r}_i(\hat{\beta}_p, t)]}{\hat{S}^{(0)}(\hat{\beta}_p, t)} \right|, \end{aligned}$$

由引理 A.1, 上面不等式右端第 1 项几乎处处收敛到 0. 再由 $\hat{\beta}_p$ 的相合性和引理 A.1 可知, 第 2 和 3 项是渐近可忽略的. 因此, $\hat{m}_p(t)$ 关于 t 一致几乎处处收敛到 $m_0(t)$.

经简单分解, 可得

$$\begin{aligned} n^{1/2} (\hat{m}_p(t) - m_0(t)) &= \frac{n^{-1/2} \sum_{i=1}^n \frac{G(T_i)}{\hat{G}(T_i)} M_i(t)}{\hat{S}^{(0)}(\hat{\beta}_p, t)} + m_0(t) \frac{n^{-1/2} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} [r_i(\beta_0, t) - r_i(\hat{\beta}_p, t)]}{\hat{S}^{(0)}(\hat{\beta}_p, t)} \\ &\quad + m_0(t) \frac{n^{-1/2} \sum_{i=1}^n \frac{\Delta_i Y_i(t)}{\hat{G}(T_i)} [r_i(\hat{\beta}_p, t) - \hat{r}_i(\hat{\beta}_p, t)]}{\hat{S}^{(0)}(\hat{\beta}_p, t)}. \end{aligned} \quad (\text{A.2})$$

由引理 A.1, (A.2) 右侧第 1 项等于

$$\frac{n^{-1/2} \sum_{i=1}^n M_i(t)}{s^{(0)}(\beta_0, t)} + \frac{n^{-1/2} \sum_{i=1}^n \int_0^\tau \frac{\theta(t, u)}{\pi(u)} dM_i^c(u)}{s^{(0)}(\beta_0, t)} + o_p(1).$$

Taylor 展式表明 (A.2) 右侧第 2 项等于 $-m_0(t)(\frac{S^{(1)}(\beta^\dagger, t)}{\hat{S}^{(0)}(\hat{\beta}_p, t)})^T n^{1/2}(\hat{\beta}_p - \beta_0)$, 其中 $\beta^\dagger$ 在 $\hat{\beta}_p$ 与 β_0 之间的线段上. 再由定理 3.1 和引理 A.1, $\frac{S^{(1)}(\beta^\dagger, t)}{\hat{S}^{(0)}(\hat{\beta}_p, t)}$ 关于 t 一致几乎处处收敛到 $\zeta(\beta_0, t)$.

使用定理 3.1 同样的推导, (A.2) 右侧第 3 项可重新写成

$$-m_0(t)n^{-1/2} \frac{1-\rho}{\rho} \sum_{i \in V} \frac{\Delta_i Y_i(t)}{G(T_i)} \frac{[\psi_i(\beta_0, t) - \phi_i(\beta_0, t)]}{s^{(0)}(\beta_0, t)} + o_p(1).$$

因此, $n^{1/2}(\hat{m}_p(t) - m_0(t)) = n^{-1/2} \sum_{i=1}^n \Psi_i(t) + o_p(1)$ 的有限维分布收敛到一个零均值的正态随机变量. 若 $n^{1/2}(\hat{m}_p(t) - m_0(t))$ 胎紧性成立, 则可以得到 $n^{1/2}(\hat{m}_p(t) - m_0(t))$ 弱收敛到一个零均值的 Gauss 过程, 协方差函数为 $\xi(s, t) = E(\Psi_1(s)\Psi_1(t))$.

因此, 为了证明 $n^{1/2}(\hat{m}_p(t) - m_0(t))$ 的胎紧性, 只需证明

$$n^{-1/2} \sum_{i=1}^n M_i(t), \quad n^{-1/2} \sum_{i=1}^n \int_0^\tau \frac{\Theta(t, u)}{\pi(u)} dM_i^c(u), \quad n^{-1/2} \sum_{i \in V} \frac{\Delta_i Y_i(t)}{G(T_i)} [\psi_i(\beta_0, t) - \phi_i(\beta_0, t)]$$

的胎紧性. 这 3 项的胎紧性可分别从文献 [26, 第 215 页, 例 2.11.16]、[24, 定理 10.2] 和本文 $\hat{\beta}_p$ 的渐近正态性得到. $\square$

Statistical inference for proportional mean residual life model with continuous auxiliary covariate

Hu Zhang & Qinglong Yang

Abstract In many statistical studies involving failure data, the main covariate is only measured in a validation set due to financial and technical restriction, while some auxiliary information for the main covariate is collected for the full cohort. Making use of the auxiliary information will increase the efficiency of the study. In this paper, we consider statistical inference for the proportional mean residual life model when the primary covariate is measured only for a randomly chosen subcohort and some auxiliary variables are available for the whole study cohort. To further make use of the auxiliary information to improve the efficiency of the study, we propose an estimating equation based on the inverse probability of censoring weighting techniques. The resulting estimators are shown to be consistent and asymptotically normal. Extensive simulations are conducted to evaluate the finite sample performance of the proposed methods. We illustrate the proposed method with a real data set from the Mayo Clinic trial in primary biliary cirrhosis of the liver.

Keywords inverse probability of censoring weighting, estimating equation, empirical process, kernel smoothing, validation set

MSC(2010) 62G05, 62N01

doi: 10.1360/SSM-2019-0263