可加风险模型下相依 I 型区间删失数据的一个 Copula 推断方法

赵慧¹, 崔琪², 孙建国^{3*}

1. 中南财经政法大学统计与数学学院, 武汉 430073;

2. 吉林大学数学学院应用统计研究中心, 长春 130012;

3. Department of Statistics, University of Missouri-Columbia, Missouri 65211, USA

E-mail: hzhao@mail.ccnu.edu.cn, qicui15@mails.jlu.edu.cn, sunj@missouri.edu

收稿日期: 2017-08-22; 接受日期: 2018-12-07; 网络出版日期: 2019-06-03; * 通信作者
国家自然科学基金 (批准号: 11471135) 资助项目

摘要 可加风险模型是生存分析中一类重要的回归模型, 许多学者对该模型进行过研究. 但是针对相依 I 型区间删失数据的研究却非常少, 且已有的研究都假设删失时间与寿命之间的关联系数已知. 显然, 该假设在实际中未必成立. 针对此问题, 本文放松这一假设, 提出一种新的基于 Copula 的方法对可加风险模型下相依 I 型区间删失数据进行回归分析, 给出参数部分估计量的渐近性质, 通过数值模拟检验所提方法在有限样本下的表现, 并进行实例分析.

关键词 可加风险模型 Copula 模型 I 型区间删失数据 相依删失

MSC (2010) 主题分类 62N01, 62N02, 62H20

1 引言

可加风险模型是生存数据分析中非常重要且常用的一类回归模型, 不少学者在各种背景下对该模型进行了研究 (参见文献 [1–6]). 特别地, 文献 [3, 4] 研究了 I 型区间删失数据的模型拟合. I 型区间删失数据也称为“现状数据”, 其中每个个体只被观测一次, 感兴趣的事件的精确失效时间无法得到, 而只能知道失效时间是小于等于观测时间还是大于观测时间, 也就是事件发生的时间被左删失或者被右删失. 这种数据广泛存在于横截面数据研究、人口调查和肿瘤实验等领域 [7]. 本文将探讨基于可加风险模型的相依 I 型区间删失数据的推断问题.

通常, 相依 I 型区间删失数据是指感兴趣的失效时间与观测时间 (也即删失时间) 是相关的. 例如, 在肿瘤实验中, 人们感兴趣的是肿瘤的产生时间, 但是一般只能在动物自然死亡, 或者被实验人员杀死的时候, 才能观测到肿瘤是否已经产生. 一般地, 如果这种肿瘤本身是致命的, 则肿瘤产生时间和死亡时间通常被视为是一致的; 如果肿瘤是非致命的, 则认为肿瘤产生时间与死亡时间是独立的. 但

英文引用格式: Zhao H, Cui Q, Sun J G. A Copula model approach for the additive hazards model with dependent current status data (in Chinese). Sci Sin Math, 2019, 49: 1261–1272, doi: 10.1360/SCM-2017-0561

是, 由于大多数肿瘤是介于致命和非致命之间的, 这表明肿瘤的发生时间和动物的死亡时间是相关的, 因此得到的观测数据是相依的 I 型区间删失数据.

目前已经有一些文献讨论过相依区间删失数据的统计分析问题, 这些方法大致可分为两类: 一类是基于脆弱模型的方法, 通过潜变量建立失效时间变量和删失时间变量的相关关系 (如文献 [8,9]). 由于在实际问题中, 潜变量不可观测, 需要关于它进行积分, 且积分往往没有显式表达式, 需要用到数值算法, 计算过程通常比较复杂且不稳定. 另一类是基于 Copula 建模方法, 采用 Copula 函数来度量变量之间的相关性, 通常比脆弱模型方法更能适用一般情形^[10]. 近年来, 一些学者讨论了相依 I 型区间删失数据的半参数回归分析, 例如, Ma 等^[11]、Zhao 等^[12]和 Xu 等^[13]分别在比例风险模型、可加风险模型和线性变换模型下对这种数据进行回归分析, 其中失效时间和删失时间的相关性都是采用 Copula 模型来刻画, 但他们的方法都需要假设 Copula 模型的类型和相关系数 α 事先已知. 但是, 通常情形下我们并不知道相关系数 α 的取值, 并且正如文献 [11] 中指出的, 当 Copula 模型的相关系数假定错误时, 得到的回归系数 β 的估计是有偏的. 本文将在可加风险模型下, 提出一个新的基于 Copula 模型的估计方法, 我们的方法不需假设 α 事先已知.

本文余下内容安排如下. 第 2 节给出本文将用到的一些记号、假设和模型. 特别是, 我们将假设感兴趣的失效时间服从可加风险模型, 失效时间与观测时间之间的关联通过 Copula 模型来刻画, 并给出相应的似然函数. 第 3 节将提出一种新的两步估计方法, 模型中未知的基准累积风险函数采用 I 样条来逼近, 同时给出估计量的渐近性质. 第 4 节通过数值模拟来评价本文所提出的估计方法在有限样本量下的表现. 实验结果表明我们的方法表现良好. 第 5 节给出一个关于小鼠肿瘤实验的实证分析. 第 6 节对本文的内容和方法进行总结.

2 模型假设

假设在某个失效时间研究中有 n 个独立观测个体, 每个个体仅被观测一次. 对于个体 i , 记 T_i 为我们感兴趣的失效时间, Z_i 为一个 p 维的协变量. 对于该个体, 假定存在两个潜在的观测时间或删失时间 C_i 和 ζ_i , 其中 C_i 与失效时间 T_i 相关, 而 ζ_i 与 T_i 独立. 在前面介绍的肿瘤实验例子中, 实验动物的自然死亡时间为 C_i , 被实验人员杀死或者研究结束的时间为 ζ_i . 记 $\tilde{C}_i = \min(C_i, \zeta_i)$, $\Delta_i = I(C_i \leq \zeta_i)$, $\delta_i = I(T_i \leq \tilde{C}_i)$, 则我们能得到的观测数据形如 $\{X_i = (\Delta_i, \delta_i, \tilde{C}_i, Z_i); i = 1, \dots, n\}$.

记 F_T 和 F_C 分别表示给定协变量时 T_i 和 C_i 的边缘分布函数, F 表示 T_i 和 C_i 的联合分布函数. 根据 Sklar 定理^[14], 存在定义在矩形 $I^2 = [0, 1] \times [0, 1]$ 上的 Copula 函数 $C_\alpha(u, v)$, 满足 $C_\alpha(u, 0) = C_\alpha(0, v) = 0$, $C_\alpha(u, 1) = u$, $C_\alpha(1, v) = v$, 且

$$F(t, c) = C_\alpha\{F_T(t), F_C(c)\}, \quad (2.1)$$

其中 α 称为关联系数, 刻画了 T 与 C 之间的关联性强弱. 如前所述, Copula 模型常被用来刻画多元寿命数据或相依寿命数据中相关变量或失效时间之间的关联性 (如文献 [11, 12, 15]). 本文假设 Copula 函数 C_α 的形式已知, 但参数 α 未知, 并在后文中对此进行说明.

为了描述协变量效应, 假设在给定 Z_i 的条件下, 感兴趣的失效时间 T_i 服从可加风险模型

$$\lambda^{(T)}(t | Z_i) = \lambda_1(t) + Z_i^T \beta, \quad (2.2)$$

同时假设删失时间 C_i 服从比例风险模型

$$\lambda^{(C)}(c | Z_i) = \lambda_2(c) \exp(Z_i^T \gamma), \quad (2.3)$$

其中 $\lambda_1(t)$ 和 $\lambda_2(c)$ 表示未知的基准风险函数, β 和 γ 表示 p 维回归系数向量. 在很多实际应用中, 人们主要对 β 的估计感兴趣.

记 f_C 为给定协变量时, C_i 的边际密度函数, 并且令 $\Lambda_T(t) = \int_0^t \lambda_1(s)ds$, $\Lambda_C(c) = \int_0^c \lambda_2(s)ds$, 则

$$F_T(t) = 1 - \exp\{-\Lambda_T(t) - Z_i^T \beta t\}, \quad F_C(c) = 1 - \exp\{-\Lambda_C(c)e^{Z_i^T \gamma}\},$$

并且

$$f_C(c) = \exp\{-\Lambda_C(c)e^{Z_i^T \gamma}\} \lambda_2(c) \exp(Z_i^T \gamma).$$

注意到, 在给定 C_i 和 Z_i 的条件下, T_i 的条件分布函数有如下形式:

$$P(T \leq t \mid C = c, Z_i) = \frac{\partial C_\alpha(u, v)}{\partial v} \bigg|_{u=F_T(t), v=F_C(c)} \triangleq m_\alpha\{F_T(t), F_C(c)\}.$$

进而, 记 $\theta = \{\beta^T, \alpha, \Lambda_T(\cdot)\}^T$, $\eta = \{\gamma^T, \Lambda_C(\cdot)\}^T$, 则参数 θ 与 γ 的似然函数具有如下的形式:

$$\begin{aligned} L(\theta, \eta) = & \prod_{i=1}^n \{([m_\alpha\{F_T(\tilde{C}_i), F_C(\tilde{C}_i)\}]^{\delta_i} [1 - m_\alpha\{F_T(\tilde{C}_i), F_C(\tilde{C}_i)\}]^{1-\delta_i} f_C(\tilde{C}_i))^{\Delta_i} \\ & \times ([F_T(\tilde{C}_i) - C_\alpha\{F_T(\tilde{C}_i), F_C(\tilde{C}_i)\}]^{\delta_i} [1 - F_T(\tilde{C}_i) - F_C(\tilde{C}_i) + C_\alpha\{F_T(\tilde{C}_i), F_C(\tilde{C}_i)\}]^{1-\delta_i})^{1-\Delta_i}\}. \end{aligned}$$

3 条件 Sieve 最大似然估计

本节讨论模型 (2.2) 和 (2.3) 的估计问题, 重点关注回归参数 β 的估计. 我们运用两步估计方法, 先估计参数 η , 然后估计 θ . 与文献 [12] 不同, 本文不仅给出回归系数 β 和 γ 的估计, 同时给出关联系数 α 的估计.

具体地, 在第一步中, 对于删失时间变量 C_i , 我们得到了完全数据或者右删失数据, 这样就可以用极大偏似然估计量和 Breslow 估计量^[16]来分别估计 γ 和 $\Lambda_C(\cdot)$ (参见文献 [17]). 记 $\hat{\gamma}$ 为参数 γ 的极大偏似然估计量, $\hat{\Lambda}_C$ 为 Λ_C 的 Breslow 估计量, $\hat{\eta} = (\hat{\gamma}, \hat{\Lambda}_C)$. 记 $N_i(t) = I(\tilde{C}_i \leq t, \Delta_i = 1)$, $Y_i(t) = I(\tilde{C}_i \geq t)$. Breslow 估计量具有如下的形式:

$$\hat{\Lambda}_C(c) = \sum_{i=1}^n \int_0^c \frac{dN_i(s)}{\sum_{i=1}^n Y_i(s) \exp(Z_i^T \hat{\gamma})}.$$

那么, 我们可以估计出 C_i 的边际分布函数为 $\hat{F}_C(c) = 1 - \exp\{-\hat{\Lambda}_C(c) \exp(Z_i^T \hat{\gamma})\}$.

在第二步中, 为了估计 θ , 对于给定的 $\hat{\eta} = (\hat{\gamma}, \hat{\Lambda}_C)$, 我们极大化条件似然函数

$$\begin{aligned} L(\theta \mid \hat{\eta}) = & \prod_{i=1}^n \{([m_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}]^{\delta_i} [1 - m_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}]^{1-\delta_i} \hat{f}_C(\tilde{C}_i))^{\Delta_i} \\ & \times ([F_T(\tilde{C}_i) - C_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}]^{\delta_i} [1 - F_T(\tilde{C}_i) - \hat{F}_C(\tilde{C}_i) + C_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}]^{1-\delta_i})^{1-\Delta_i}\}. \end{aligned}$$

另一方面, 可以看出由于 T_i 的基准累积风险函数 $\Lambda_T(\cdot)$ 是无穷维未知函数, 直接极大化上面的条件似然很困难. 根据文献 [11] 中的想法, 我们在极大化之前, 先使用 Sieve 方法, 也就是用单调的 I 样条函数来逼近 $\Lambda_T(\cdot)$ (参见文献 [18, 19]). Sieve 方法的一个主要优势是使用一些参数形式的光滑函数来逼近未知函数, 其中一类常用的光滑函数是文献 [19] 中讨论的 I 样条. 具体来讲就是对于一个有限闭区间 $[a, b]$, 记 $\mathcal{X} = \{x_i\}_{i=1}^{k_n+2m}$ 是一列分割 $[a, b]$ 的节点, 其中 $a = x_1 = \cdots = x_m < x_{m+1} < \cdots < x_{k_n+m} <$

$x_{k_n+m+1} = \cdots = x_{k_n+2m} = b$. 给定 $\mathcal{X}$, 一族阶数为 m 、内点数为 k_n 的 I 样条基函数 $\{I_i(t | m, x)\}_{i=1}^{m+k_n}$ 的定义为 $I_i(t | m, x) = \int_a^t M_i(u | m, x) du$, 其中函数 $\{M_i(t | m, x)\}_{i=1}^{m+k_n}$ 有如下的递归定义:

$$M_i(t | 1, x) = \begin{cases} \frac{1}{x_{i+1} - x_i}, & x_i \leq t < x_{i+1}, \\ 0, & \text{否则,} \end{cases}$$

并且, 对于 $m > 1$, 有

$$M_i(t | m, x) = \frac{m[(t - x_i)M_i(t | m-1, x) + (x_{i+m} - t)M_{i+1}(t | m-1, x)]}{(m-1)(x_{i+m} - x_i)},$$

可以看出, I 样条基函数是单调不减的, 任一连续函数都可以由这种基函数的线性组合来逼近. 特别地, 可以用它们的线性组合 $\Lambda_{T_n}(t) = \sum_{i=1}^{m+k_n} \xi_i I_i(t | m, x)$ (其中 $\xi_i \geq 0, i = 1, \dots, m+k_n$) 来逼近未知的基准累积风险函数 $\Lambda_T(t)$, 其中系数 $\xi_i \geq 0$ 保证了 $\Lambda_{T_n}(t)$ 单调递增.

记 M 是一个正的常数, $\{I_i(t)\}_{i=1}^{m+k_n}$ 是阶数为 m 、内点数为 k_n 的 I 样条基函数, 其中 $k_n = o(n^v)$, $0 < v < 0.5$, 定义 $\Theta_n = \{\theta_n = (\beta^T, \alpha, \Lambda_{T_n})^T\} = \mathcal{B} \otimes \mathcal{M}_n$, 其中 $\mathcal{B} = \{(\beta^T, \alpha)^T \in \mathbb{R}^{p+1}, \|\beta\| + \|\alpha\| \leq M\}$, $\|v\|$ 表示向量 v 的 Euclid 范数. $\mathcal{M}_n = \{\Lambda_{T_n} : \Lambda_{T_n}(t) = \sum_{i=1}^{m+k_n} \xi_i I_i(t), \xi_i \geq 0, i = 1, \dots, m+k_n, t \in [0, u_c]\}$, u_c 是所有观测时间点 $\{\tilde{C}_i : i = 1, \dots, n\}$ 的上界.

根据文献 [18, 引理 A₁] 可知, 可以把 Θ_n 当作原始参数空间 Θ 的 Sieve 空间. 这样, 就可以通过 I 样条基函数把之前的既包含有限维参数, 又包含非参数函数的半参数估计问题转化成只包含有限维参数的估计问题.

接下来用 Sieve 方法来估计参数 $\theta = (\beta^T, \alpha, \xi^T)^T$, 其中 $\xi^T = (\xi_1, \dots, \xi_{m+k_n})$. 记相应的 Sieve 极大似然估计量为 $(\hat{\beta}, \hat{\alpha}, \hat{\Lambda}_T(\cdot))$, 它使得条件对数似然函数 $l(\theta | \hat{\eta})$ 在 Θ_n 上取极大值, 其中 $l(\theta | \hat{\eta}) = \sum_{i=1}^n l^{(i)}(\theta | \hat{\eta})$, 这里

$$\begin{aligned} l^{(i)}(\theta | \hat{\eta}) &= \Delta_i \log \hat{f}_C(\tilde{C}_i) + (1 - \delta_i) \Delta_i \log[1 - m_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}] \\ &\quad + \delta_i \Delta_i \log[m_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}] + \delta_i(1 - \Delta_i) \log[F_T(\tilde{C}_i) - C_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}] \\ &\quad + (1 - \delta_i)(1 - \Delta_i) \log[1 - F_T(\tilde{C}_i) - \hat{F}_C(\tilde{C}_i) + C_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}]. \end{aligned}$$

注意到估计量 $\hat{\gamma}$ 和 $\hat{\Lambda}_C$ 的渐近性质已有研究结果, 具体来讲, 在给定了一定的正则条件后, 文献 [20, 定理 8.1–8.3] 表明 $\hat{\gamma}$ 和 $\hat{\Lambda}_C$ 都具有相合性, 且 $\sqrt{n}(\hat{\gamma} - \gamma_0)$ 和 $\sqrt{n}(\hat{\Lambda}_C(t) - \Lambda_{C0}(t))$ 分别收敛到一个多元正态分布和一个 Gauss 过程 (详见附录 A 中的说明), 其中 γ_0 和 Λ_{C0} 分别为 γ 和 Λ_C 的真值. 下面的定理给出我们最感兴趣的估计量 $\hat{\beta}$ 的渐近性质.

定理 3.1 假定附录 A 中所列的正则性条件 (C1)–(C4) 成立, 那么, 当 $n \rightarrow \infty$ 时, $\hat{\beta}$ 是相合的, 并且 $\sqrt{n}(\hat{\beta} - \beta_0)$ 收敛于均值为 0 的多元正态分布, 其中 β_0 表示 β 的真值.

我们将定理 3.1 的证明推迟到附录 A 中给出. 对于前面提出的估计方法, 我们需要指定 m 和 k_n . 通常样条的阶数 m 可根据真实的基准累积风险函数的光滑性来定. 例如, 一般情形下二次或三次样条就已经表现很好了. 另外一种常用的方法是采用 AIC (Akaike information criterion) 准则, 选择使模型的 AIC 值达到最小的 m 和 k_n . 对于内节点的选取, 通常的做法是在给定 m 和 k_n 后, 选取 $[0, u_c]$ 上所有观测时间点的等间隔分位点来做为内节点. 当然, 也可以尝试不同的内节点, 比较最终的结果来选择一种更合适的. 当确定了 m 和 k_n 以后, 计算 $\hat{\theta}$ 就比较容易了, 因为可以用一些现成的软件包, 例如, R 里面的函数 nlminb.

为了进行回归系数的统计推断, 我们还需要估计 $\hat{\beta}$ 的协方差矩阵. 一个自然的想法是推导一个相合的估计量, 但是从定理的证明中可以看到这样的估计量过于复杂, 并不实用. 这里运用 Bootstrap 自助法给出协方差的一个估计量. 令 B 是一个事先给定的正整数, 从原样本 $\{X_i = (\Delta_i, \delta_i, \tilde{C}_i, Z_i), i = 1, \dots, n\}$ 中有放回地抽取样本容量为 n 的自助样本 B 次, 新抽取的自助样本记为 $\{X_i^{(b)} = (\Delta_i^{(b)}, \delta_i^{(b)}, \tilde{C}_i^{(b)}, Z_i^{(b)}), i = 1, \dots, n\}, b = 1, \dots, B$. 对于第 b 个自助样本 $\{X_i^{(b)}\}$, 记 $\hat{\beta}^{(b)}$ 为基于 $\{X_i^{(b)}\}$ 得到的极大似然估计. 那么, 可以通过下式估计 $\hat{\beta}$ 的协方差矩阵:

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{1}{B-1} \sum_{b=1}^B \left(\hat{\beta}^{(b)} - \frac{1}{B} \sum_{b=1}^B \hat{\beta}^{(b)} \right)^2.$$

类似地, $\hat{\alpha}$ 也具有相合性并且渐近服从正态分布, 也可以运用自助法估计它的方差. 关于 B 的选择, 可以先尝试某个合理的整数, 然后再逐渐增大, 直至所得结果趋于稳定. 在大量的模拟实验中, 我们发现当 $B = 50$ 时计算结果已经比较稳定, 而且可以节省计算时间.

4 数值模拟研究

本节通过数值模拟来研究前面提出的条件 Sieve 极大似然估计方法在有限样本下的表现. 假设协变量 Z 服从成功概率为 0.5 的 Bernoulli 分布. 为了产生失效时间 T_i 和观测时间 C_i , 我们在模型 (2.2) 和 (2.3) 中取 $\lambda_1(t) = \lambda_2(c) = 1$. 首先从 $(0, 1)$ 区间上的均匀分布中产生两个独立的随机数 u_i 和 w_i , 然后求解方程

$$P(C \leq c_i | T = t_i, Z) = \frac{\partial C_\alpha(u, v)}{\partial u} \bigg|_{u=u_i, v=v_i} = w_i,$$

得到随机数 v_i . 令 $T_i = t_i, C_i = c_i$, 其中 t_i 和 c_i 分别表示方程 $F_T(t_i) = u_i$ 和 $F_C(c_i) = v_i$ 的解. 同时, 将与失效时间 T_i 独立的观测时间 ζ_i 取成常数.

为了产生模拟数据, 考虑了几种不同的 Copula 模型, 包括 Clayton 模型

$$C_\alpha(u, v) = (u^{-(\alpha-1)} + v^{-(\alpha-1)} - 1)^{-1/(\alpha-1)}, \quad \alpha > 1$$

和 Frank 模型

$$C_\alpha(u, v) = -\frac{1}{\alpha} \log \left\{ 1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{e^{-\alpha} - 1} \right\}, \quad \alpha \neq 0.$$

注意到, 对于不同的 Copula 模型, 关联系数 α 的取值范围不同, 并且它们的含义也不同. 因此, 在接下来的模拟研究中, 我们采用另一个相关系数, Kendall 相关系数 τ . 它也是一种全局关联系数, 具体定义为 $\tau = P\{(T_i - T_j)(C_i - C_j) > 0\} - P\{(T_i - T_j)(C_i - C_j) < 0\}$, 其中 (T_i, C_i) 和 (T_j, C_j) 是 (T, C) 的独立同分布的抽样. Kendall 相关系数 τ 取值总是在 -1 到 1 之间. 当两个随机变量独立时, Kendall 相关系数 τ 的取值为 0, 而且它是稳健的, 在单调变换下保持不变.

以下的模拟结果基于 1,000 次重复, 每次抽取的样本容量为 $n = 200$. 表 1 给出 Clayton 模型下回归系数 β_α 、 γ 和 Kendall 相关系数 τ 的模拟结果, 其中 β 和 γ 的真值设为 0、-0.5 或 0.5; τ 的真值为 0.3 或 0.5. 结果包括了估计的偏差 (Bias)、样本标准差 (the sample standard errors, SSE)、标准差估计的均值 (the estimated standard errors, SEE) 和 95% 经验覆盖率 (the coverage probability, CP). 对于 $\Lambda_T(\cdot)$ 的近似, 设定样条基函数的阶数为 2, 并选择 $\tilde{C}_i$ 的 0.2、0.4、0.6 和 0.8 分位数作为 4 个内

节点. 从表 1 中可以看到, 本文提出的两步估计方法得到的估计量是无偏的; SSE 与 SEE 接近, 说明基于 Bootstrap 方法得到的方差估计是合理的; 经验覆盖率接近 95%, 说明提出的估计量是近似正态分布的.

表 2 给出了 Frank 模型下的估计结果, 其他参数的设定与表 1 相同. 可以看出在 Frank 模型下, 我们提出的估计方法依然表现很好.

基于 Frank 模型和 Gumbel 模型考虑协变量为混合变量的情形, 即 $Z = (Z_1, Z_2)^T$, 其中 Z_1 服从成功概率为 0.5 的 Bernoulli 分布, Z_2 服从 $(0, 1)$ 上的均匀分布, 具体模拟结果见表 3. 该结果表明我们的方法对不同类型的协变量和不同 Copula 函数都表现良好, 这也说明了该方法具有一定稳健性. 此外, 为了评价 Sieve 方法中 I 样条的估计效果, 我们又考虑了 $\Lambda_T(t) = 1.2t^{1.2}$ 和 $\Lambda_C(c) = 1.2c^{1.2}$ 的情形, 表 4 中给出模拟结果, 该结果显示采用 I 样条是合理可靠的.

表 1 Clayton 模型下回归系数与关联参数的估计

参数	真值	Bias	SSE	SEE	CP	真值	Bias	SSE	SEE	CP
τ	0.3	0.017	0.218	0.212	0.947	0.3	0.010	0.213	0.205	0.930
β	0.0	-0.019	0.243	0.249	0.962	0.5	-0.010	0.283	0.292	0.958
γ	-0.5	-0.001	0.151	0.148	0.938	0.0	0.001	0.144	0.144	0.950
τ	0.3	0.040	0.229	0.222	0.936	0.3	0.019	0.224	0.220	0.966
β	0.0	0.028	0.279	0.285	0.955	0.5	0.008	0.288	0.296	0.967
γ	0.5	0.001	0.155	0.149	0.949	0.5	-0.001	0.148	0.149	0.948
τ	0.5	-0.051	0.164	0.163	0.944	0.5	-0.059	0.179	0.173	0.903
β	0.0	0.063	0.225	0.241	0.968	0.5	0.057	0.254	0.286	0.969
γ	-0.5	-0.002	0.151	0.148	0.945	0.0	0.006	0.143	0.144	0.954
τ	0.5	-0.055	0.189	0.188	0.935	0.5	-0.126	0.204	0.203	0.900
β	0.0	-0.064	0.240	0.264	0.968	0.5	-0.089	0.271	0.280	0.955
γ	0.5	0.002	0.143	0.149	0.958	0.5	0.008	0.148	0.149	0.951

表 2 Frank 模型下回归系数与关联参数的估计

参数	真值	Bias	SSE	SEE	CP	真值	Bias	SSE	SEE	CP
τ	0.3	0.058	0.329	0.304	0.946	0.3	0.049	0.297	0.275	0.927
β	0.0	-0.003	0.198	0.213	0.971	0.5	-0.042	0.291	0.309	0.952
γ	0.0	0.003	0.145	0.145	0.953	0.0	-0.006	0.149	0.144	0.945
τ	0.3	0.019	0.301	0.276	0.920	0.3	0.026	0.275	0.257	0.922
β	0.0	0.026	0.327	0.323	0.936	0.5	-0.022	0.318	0.312	0.959
γ	0.5	0.002	0.148	0.149	0.944	0.5	0.005	0.151	0.149	0.954
τ	0.5	-0.069	0.287	0.271	0.962	0.5	-0.034	0.241	0.226	0.925
β	0.0	-0.009	0.174	0.194	0.972	0.5	0.010	0.303	0.308	0.954
γ	0.0	-0.002	0.139	0.144	0.962	0.0	-0.010	0.142	0.144	0.952
τ	0.5	-0.033	0.249	0.233	0.926	0.5	-0.045	0.252	0.234	0.945
β	0.0	-0.029	0.304	0.303	0.946	0.5	-0.050	0.301	0.295	0.945
γ	0.5	0.002	0.149	0.148	0.954	0.5	0.006	0.152	0.149	0.946

表 3 协变量取二维混合变量时的参数估计

	参数	真值	Bias	SSE	SEE	CP	真值	Bias	SSE	SEE	CP
Frank 模型	τ	0.3	0.009	0.308	0.287	0.955	0.3	0.022	0.310	0.289	0.946
	β_1	0.3	-0.021	0.218	0.229	0.965	0.3	-0.023	0.298	0.302	0.954
	β_2	-0.3	-0.010	0.324	0.344	0.964	0.3	-0.047	0.437	0.479	0.964
	γ_1	0.3	0.003	0.145	0.147	0.953	0.3	-0.002	0.146	0.147	0.951
	γ_2	-0.3	-0.007	0.262	0.255	0.958	0.3	-0.004	0.250	0.256	0.959
	τ	0.5	-0.088	0.286	0.266	0.944	0.5	-0.078	0.273	0.268	0.956
	β_1	0.3	-0.038	0.205	0.212	0.962	0.3	-0.056	0.265	0.278	0.957
	β_2	-0.3	0.010	0.293	0.317	0.968	0.3	-0.058	0.410	0.441	0.967
	γ_1	0.3	0.004	0.153	0.147	0.941	0.3	0.005	0.149	0.147	0.956
	γ_2	-0.3	0.006	0.264	0.255	0.951	0.3	0.019	0.258	0.256	0.952
Gumbel 模型	τ	0.3	0.055	0.246	0.231	0.912	0.3	0.051	0.248	0.231	0.916
	β_1	0.3	-0.023	0.195	0.210	0.969	0.3	-0.027	0.276	0.285	0.954
	β_2	-0.3	-0.005	0.308	0.327	0.963	0.3	-0.048	0.426	0.454	0.968
	γ_1	0.3	0.001	0.151	0.147	0.944	0.3	0.010	0.153	0.147	0.939
	γ_2	-0.3	0.004	0.249	0.256	0.969	0.3	0.016	0.255	0.255	0.952
	τ	0.5	-0.021	0.232	0.212	0.910	0.5	-0.013	0.228	0.214	0.915
	β_1	0.3	-0.031	0.188	0.197	0.962	0.3	-0.049	0.255	0.268	0.954
	β_2	-0.3	0.015	0.298	0.298	0.953	0.3	-0.069	0.402	0.427	0.953
	γ_1	0.3	0.004	0.148	0.147	0.953	0.3	0.002	0.154	0.147	0.939
	γ_2	-0.3	0.003	0.257	0.256	0.956	0.3	-0.002	0.256	0.255	0.952

表 4 $\Lambda_T(t) = 1.2t^{1.2}$ 和 $\Lambda_C(c) = 1.2c^{1.2}$ 时, Frank 模型下参数的估计

参数	真值	Bias	SSE	SEE	CP	真值	Bias	SSE	SEE	CP
τ	0.3	0.016	0.325	0.300	0.953	0.3	0.032	0.291	0.272	0.935
β	0.0	-0.019	0.229	0.243	0.971	0.5	-0.033	0.311	0.331	0.957
γ	0.0	-0.002	0.147	0.143	0.941	0.0	0.002	0.146	0.143	0.949
τ	0.3	-0.012	0.276	0.260	0.952	0.3	-0.011	0.262	0.242	0.923
β	0.0	-0.011	0.337	0.337	0.955	0.5	-0.039	0.332	0.330	0.953
γ	0.5	0.005	0.150	0.147	0.948	0.5	0.006	0.151	0.147	0.948
τ	0.5	-0.086	0.276	0.266	0.950	0.5	-0.042	0.256	0.229	0.925
β	0.0	-0.007	0.205	0.224	0.972	0.5	0.023	0.321	0.329	0.961
γ	0.0	-0.005	0.146	0.143	0.946	0.0	0.009	0.141	0.143	0.940
τ	0.5	-0.058	0.249	0.234	0.953	0.5	-0.066	0.250	0.227	0.922
β	0.0	-0.069	0.335	0.339	0.946	0.5	-0.087	0.322	0.321	0.956
γ	0.5	-0.001	0.151	0.147	0.942	0.5	-0.001	0.143	0.147	0.957

如前所述, Zhao 等^[12] 讨论了相同的问题, 但是他们的方法假设了相关系数 α 或 Kendall 相关系数 τ 是已知的. 我们通过数值模拟, 比较了本文提出的两步估计方法和文献 [12] 提出的方法. 表 5 给出在相同的模拟数据下, τ 真值为 0.3 或 0.5 时两种方法的估计. 可以看出, 两种方法的估计效果都很

表 5 本文的方法与文献 [12] 的方法的比较

	参数	真值	本文的方法				文献 [12] 中的方法			
			Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
$\tau_0 = 0.3$	β	0.0	-0.003	0.198	0.213	0.971	-0.014	0.190	0.188	0.946
	γ	0.0	0.003	0.145	0.145	0.953	-0.002	0.146	0.143	0.945
	β	0.3	-0.010	0.246	0.262	0.963	0.017	0.233	0.222	0.951
	γ	0.3	-0.007	0.151	0.147	0.954	0.001	0.146	0.145	0.946
	β	0.5	-0.022	0.318	0.312	0.959	0.014	0.252	0.249	0.951
	γ	0.5	0.005	0.151	0.149	0.954	0.018	0.148	0.147	0.946
$\tau_0 = 0.5$	β	0.0	-0.009	0.174	0.194	0.972	-0.009	0.169	0.164	0.944
	γ	0.0	-0.002	0.139	0.144	0.962	0.000	0.144	0.140	0.947
	β	0.3	-0.033	0.234	0.242	0.964	-0.005	0.204	0.194	0.941
	γ	0.3	0.005	0.149	0.147	0.942	-0.004	0.146	0.143	0.947
	β	0.5	-0.050	0.301	0.295	0.945	0.018	0.220	0.219	0.939
	γ	0.5	0.006	0.152	0.149	0.946	0.006	0.146	0.145	0.939

好. 相对而言, 文献 [12] 提出的方法方差略小, 或者说更有效一点. 但是, 在大多数情形下, 两种估计方法的方差是比较接近的. 也就是说, 通常情形下, 我们提出的估计方法与文献 [12] 中的方法相比, 效率损失比较小. 但有一点必须明确, 如果参数 τ 的值不知道, 文献 [12] 中的算法就无法工作, 但本文的方法仍有效.

5 实际数据分析

本节将应用前面提出的模型和估计方法来分析一组小鼠肺部肿瘤的实验数据. 已有几位学者对这组数据进行过分析 (参见文献 [12, 21]). 本实验中, 144 只雄性小鼠被随机分为两组, 其中 96 只被分到常规环境 (conventional environment, CE) 中, 其他 48 只被分到无菌环境 (germ-free environment, GE) 中. 对每只小鼠, 得到的数据有死亡时间和在死亡时刻肺部是否出现肿瘤. 也就是说, 对于我们感兴趣的肿瘤发生时间, 我们仅能得到 I 型区间删失数据. 本实验的目的是比较两组之间肿瘤的发生风险率.

对于第 i 只小鼠, 记 T_i 为肿瘤发生时间, C_i 为小鼠死亡时间 (也即对肿瘤的观测时间). 如果第 i 只小鼠在无菌环境组, 那么记 $Z_i = 1$; 否则记 $Z_i = 0$. 首先考虑对于动物死亡时间的组别效应, 运用我们提出的估计方法的第一步, 可得 $\hat{\gamma} = -1.968$, 估计的标准差为 0.243, 相应的 p - 值接近于 0, 这表明无菌环境中小鼠的死亡率明显低于传统环境中小鼠的死亡率. 对于我们提出的估计方法的第二步, 我们考虑几种常见的 Copula 模型, 采用了几种不同的 m 和 k_n 的取值, 并运用 AIC 准则来选择其中最优的组合. 我们发现当 $m = 2$ 、 $k_n = 3$ 、取 Copula 函数为下面的 Gumbel 模型时:

$$C_\alpha(u, v) = \exp[-\{(-\log u)^\alpha + (-\log v)^\alpha\}^{1/\alpha}], \quad \alpha \geq 1,$$

AIC 取得最小值.

与模拟研究中的做法一样, 对于给定的 k_n , 选取 $[0, u_c]$ 区间上的等样本量分划, 选取相应的分位数作为内节点, 其中 u_c 是所有观测时间点的上界.

表 6 给出 Kendall 相关系数 τ 和回归系数 β 的估计, 包括在最优组合下的估计结果以及其他几

表 6 肿瘤实验分析结果

		参数	估计值	SEE	p - 值	AIC
Clayton 模型	$k_n = 3$	τ	-0.05560	0.16854	0.74146	839.3737
		β	0.00091	0.00062	0.14020	
	$k_n = 4$	τ	-0.07689	0.19037	0.68626	840.5355
		β	0.00094	0.00065	0.14727	
	$k_n = 5$	τ	-0.06651	0.19996	0.73939	842.6871
		β	0.00092	0.00069	0.18298	
Frank 模型	$k_n = 3$	τ	0.16343	0.16758	0.32943	839.0853
		β	0.00018	0.00062	0.77155	
	$k_n = 4$	τ	0.06929	0.16686	0.67795	840.6480
		β	0.00050	0.00062	0.41946	
	$k_n = 5$	τ	0.17184	0.16162	0.28767	842.3577
		β	0.00014	0.00061	0.81626	
Gumbel 模型	$k_n = 3$	τ	0.11691	0.09066	0.19722	838.4409
		β	0.00028	0.00047	0.54632	
	$k_n = 4$	τ	0.10949	0.09867	0.26714	839.7632
		β	0.00028	0.00045	0.53639	
	$k_n = 5$	τ	0.12813	0.11637	0.27088	841.6067
		β	0.00022	0.00047	0.62792	

种组合下的估计结果. 结果包括参数的估计、估计量的标准差和检验参数是否为 0 的 p - 值. 结果表明, 肿瘤的发生时间与动物的死亡时间似乎是不相关的, 这也与文献 [21] 中提出的可以认为肺部肿瘤非致命的观点一致. 同时, 结果也表明两种环境下, 小鼠肺部肿瘤的产生风险并没有显著的差异. 对比之下, 文献 [12] 提出, 如果认为肿瘤的发生时间与动物的死亡时间不显著相关, 那么存在一定的环境效应; 如果认为两个时间变量相关, 那就不存在环境效应.

6 结论

在前几节中, 我们在可加风险模型下, 针对相依 I 型区间删失数据, 提出了一种基于 Copula 模型的新方法, 该方法是一种两步估计算法, 其中采用了 I 样条函数来逼近模型的非参数部分, 也就是基准累积风险函数. 得到的估计量具有相合性和渐近正态分布. 与文献 [12] 中的方法相比, 本文提出的估计方法的主要优势在于我们不需要假设关联系数 α 已知. 一般而言, 如果没有额外的信息, 即使 Copula 函数的结构已知, 关联系数 α 也是不可估的. 本文用到的额外信息就是第一步中估计出来的边际分布函数 F_C , 在后面的估计过程中可以把这个边际分布函数当作已知的.

需要注意的是, 与其他文献中的做法类似, 这里也假设了基本 Copula 模型是已知的. 这是由于如果没有更强的假设和额外的信息, 估计 Copula 函数的结构是不太可能的 (参见文献 [11, 12, 15]). 实际应用中可以多尝试几种 Copula 函数, 选择计算结果好的来用. 另外, 本文假设感兴趣的失效时间服从可加风险模型, 也可以考虑将本文的两步估计算法推广到其他回归模型, 如线性转移模型.

本文所提方法仅限于 I 型区间删失数据, 这是一种特殊的区间删失数据. 针对这一局限, 将来的

一个研究方向是, 讨论更一般的区间删失数据, 也就是仅能观测到个体的失效时间属于某个区间 (参见文献 [7]). 显然, 对于更一般的区间删失数据, 也同样可能存在相依删失, 将本文的方法推广到这种情形也是很有意义的. 虽然我们仍可采用 Copula 模型来做, 但肯定会比本文的算法更复杂, 这主要是因为一般区间删失涉及两个观测时间点, 再加上失效时间, 我们就需考虑三个变量间的相关性, 就得采用三维 Copula 模型, 而不再是本文中的二维 Copula 模型.

致谢 感谢三位审稿人以及主编、副主编对本文提出的修改意见.

参考文献

- 1 Chen L, Sun J. A multiple imputation approach to the analysis of current status data with the additive hazards model. *Comm Statist Theory Methods*, 2009, 38: 1009–1018
- 2 Lin D Y, Ying Z. Semiparametric analysis of the additive risk model. *Biometrika*, 1994, 81: 61–71
- 3 Lin D Y, Oakes D, Ying Z. Additive hazards regression with current status data. *Biometrika*, 1998, 85: 289–298
- 4 Martinussen T, Scheike T H. Efficient estimation in additive hazards regression with current status data. *Biometrika*, 2002, 89: 649–658
- 5 Tong X W, Hu T, Sun J G. Efficient estimation for additive hazards regression with bivariate current status data. *Sci China Math*, 2012, 55: 763–774
- 6 Wang L, Sun J, Tong X. Regression analysis of case II interval-censored failure time data with the additive hazards model. *Statist Sinica*, 2010, 20: 1709–1723
- 7 Sun J. *The Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer, 2006
- 8 Zhang Z, Sun J, Sun L. Statistical analysis of current status data with informative observation times. *Stat Med*, 2005, 24: 1399–1407
- 9 Chen C M, Lu T F C, Chen M H, et al. Semiparametric transformation models for current status data with informative censoring. *Biom J*, 2012, 54: 641–656
- 10 Shih J H, Louis T A. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 1995, 51: 1384–1399
- 11 Ma L, Hu T, Sun J. Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika*, 2015, 102: 731–738
- 12 Zhao S, Hu T, Ma L, et al. Regression analysis of informative current status data with the additive hazards model. *Lifetime Data Anal*, 2015, 21: 241–258
- 13 Xu D, Zhao S, Hu T, et al. Regression analysis of informative current status data with the semiparametric linear transformation model. *J Appl Stat*, 2019, 46: 187–202
- 14 Nelsen R B. *An Introduction to Copulas*, 2nd ed. New York: Springer, 2006
- 15 Zheng M, Klein J P. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 1995, 82: 127–138
- 16 Breslow N E. Analysis of survival data under the proportional hazards model. *Internat Statist Rev*, 1975, 43: 45–58
- 17 Kalbfleisch J D, Prentice R L. *The Statistical Analysis of Failure Time Data*, 2nd ed. New York: John Wiley & Sons, 2002
- 18 Lu M, Zhang Y, Huang J. Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika*, 2007, 94: 705–718
- 19 Ramsay J O. Monotone regression splines in action. *Statist Sci*, 1988, 3: 425–441
- 20 Fleming T R, Harrington D P. *Counting Processes and Survival Analysis*. New York: John Wiley & Sons, 1991
- 21 Hoel D G, Walburg H E. Statistical analysis of survival experiments. *J Natl Cancer Inst*, 1972, 49: 361–372
- 22 Huang J. Efficient estimation for the proportional hazards model with interval censoring. *Ann Statist*, 1996, 24: 540–568
- 23 Huang J, Rossini A J. Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J Amer Statist Assoc*, 1997, 92: 960–967

附录 A 定理 3.1 的证明

首先说明 $\hat{\eta}$ 的渐近性质: 基于文献 [20, 定理 8.1–8.3] 可知, $\hat{\gamma}$ 和 $\hat{\Lambda}_C(\cdot)$ 都是相合的, 且 $\sqrt{n}(\hat{\gamma} - \gamma_0)$

和 $\sqrt{n}(\hat{\Lambda}_C(t) - \Lambda_{C0}(t))$ 分别弱收敛到一个多元正态分布和一个 Gauss 过程. 此外,

$$\hat{\gamma} - \gamma_0 = \frac{1}{n} \sum_{j=1}^n g_j + o_p(n^{-1/2}), \quad (\text{A.1})$$

$$\hat{\Lambda}_C(t) - \Lambda_{C0}(t) = \frac{1}{n} \sum_{j=1}^n b_j(t) + o_p(n^{-1/2}), \quad (\text{A.2})$$

其中 $\inf\{t : \Lambda_C(t) > 0\} < t < u_c$, 且上式中,

$$\begin{aligned} g_j &= \Sigma_{\gamma_0}^{-1} \int_0^{u_c} \{Z_j - e(\gamma_0, t)\} dM_j(t), \\ e(\gamma_0, t) &= \frac{\mathbb{E}\{I(t \leq \zeta_j) Z_j e^{\gamma_0^T Z_j}\}}{\mathbb{E}\{I(t \leq \zeta_j) e^{\gamma_0^T Z_j}\}}, \\ M_j(t) &= I(C_j \leq \zeta_j, C_j \leq t) - \int_0^t I(u \leq \zeta_j) e^{\gamma_0^T Z_j} d\Lambda_{C0}(u), \\ b_j(t) &= \int_0^t \frac{dM_j(u)}{\mathbb{E}\{I(t \leq \zeta_j) e^{\gamma_0^T Z_j}\}} - h^T(t) g_j, \\ h(t) &= \int_0^t e(\gamma_0, u) d\Lambda_{C0}(u). \end{aligned}$$

为了证明 $\hat{\beta}$ 的相合性和渐近正态性, 除了证明 $\hat{\eta}$ 的渐近性质所需的正则性条件外, 我们还需要如下正则性条件.

(C1) 对于删失时间 ζ_i , 成立 $P(\zeta_i \geq a_0) > 0$, 其中 a_0 表示最长的跟踪时间.

(C2) 协变量 Z 在实空间 $\mathbb{R}^p$ 上存在一个有界支撑.

(C3) (i) Copula 函数 $C(\cdot, \cdot)$ 存在有界的一阶偏导数, 且 $\partial C(u, v)/\partial u$ 和 $\partial C(u, v)/\partial v$ 是 Lipschitz 的. (ii) 假设对于任意开集 $E \in I^2$, 有 $\mu_C(E) > 0$, 其中 μ_C 表示给定 Z 时, 与 Copula 函数 $C(\cdot, \cdot)$ 相应的概率测度.

(C4) $\Lambda_T(\cdot)$ 的 κ 阶导数, 记为 $\Lambda_T^{(\kappa)}(\cdot)$, 是 Holder 连续的, 也就是说, 存在某个 $\varepsilon \in (0, 1]$, 使得 $|\Lambda_0^{(\kappa)}(t_1) - \Lambda_0^{(\kappa)}(t_2)| \leq M_0 |t_1 - t_2|^\varepsilon$, $\forall t_1, t_2 \in (0, u_c)$, 其中 M_0 是某个常数.

接下来给出 $\hat{\beta}$ 的相合性和渐近正态性的证明过程.

下面证明 $\hat{\theta}$ 的相合性. 令 $\theta_0 = (\beta_0^T, \alpha_0, \Lambda_{T0})^T$ 和 $\eta_0 = (\gamma_0^T, \Lambda_{C0})^T$ 分别为 θ 和 η 的真值. 对于任意的 $\epsilon > 0$, 由于 $\{|l(\hat{\theta} | \hat{\eta}) - l(\theta_0 | \eta_0)| \geq \epsilon\} \subset \{|l(\hat{\theta} | \hat{\eta}) - l(\hat{\theta} | \eta_0)| \geq \epsilon/2\} \cup \{|l(\hat{\theta} | \eta_0) - l(\theta_0 | \eta_0)| \geq \epsilon/2\}$, 我们有 $P\{|l(\hat{\theta} | \hat{\eta}) - l(\theta_0 | \eta_0)| \geq \epsilon\} \leq P\{|l(\hat{\theta} | \hat{\eta}) - l(\hat{\theta} | \eta_0)| \geq \frac{\epsilon}{2}\} + P\{|l(\hat{\theta} | \eta_0) - l(\theta_0 | \eta_0)| \geq \frac{\epsilon}{2}\}$.

一方面, 根据 Markov 不等式和连续映射定理, 给定 $\hat{\theta}$ 的当前值, 作为 η 的连续函数, 当 n 趋于无穷时, $l(\hat{\theta} | \hat{\eta})$ 依概率收敛于 $l(\hat{\theta} | \eta_0)$; 另一方面, 在正则条件 (C1)–(C4) 下, 对数似然函数 $l(\theta | \eta_0)$ 是凹函数, 于是, 极大值 $\hat{\theta}$ 存在, 并且是唯一的、相合的. 基于这两个事实, 当 $n \rightarrow \infty$ 时, $P\{|l(\hat{\theta} | \hat{\eta}) - l(\theta_0 | \eta_0)| \geq \epsilon\} \rightarrow 0$, 这表明 $\hat{\beta}$ 是相合的.

接下来证明 $\hat{\theta}$ 的渐近正态性. 首先由 $\dot{l}_\theta(\hat{\theta} | \hat{\eta})$ 在真值 θ_0 处的 Taylor 展开, 我们有

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\{n^{-1} \ddot{l}_\theta(\theta_0 | \hat{\eta})\}^{-1} n^{-1/2} \dot{l}_\theta(\theta_0 | \hat{\eta}) + o_p(1),$$

其中 $\ddot{l}_\theta(\theta_0 | \hat{\eta})$ 是 $l(\theta | \hat{\eta})$ 关于 θ 的二阶偏导在 $\theta = \theta_0$ 处的值, 且 $n^{-1} \ddot{l}_\theta(\theta_0 | \hat{\eta})$ 收敛到正定矩阵 $-\mathbb{E}\{\ddot{l}_\theta(\theta_0 | \hat{\eta})\}$. 因此, 为了证明 $\hat{\beta}_a$ 的渐近正态性, 只需要证明得分函数 $\dot{l}_\theta(\theta_0 | \hat{\eta})$ 可以写成 n 个独立

同分布的均值为 0 的随机变量之和, 加上一些可以忽略的无穷小量. 对此, 我们可以把 $\dot{l}_\theta(\theta_0 | \hat{\eta})$ 重新写成 $\dot{l}_\theta(\theta_0 | \hat{\eta}) = I + II$, 其中 $I = \dot{l}_\theta(\theta_0 | \eta_0)$, $II = \dot{l}_\theta(\theta_0 | \hat{\eta}) - \dot{l}_\theta(\theta_0 | \eta_0)$.

由文献 [22, 23] 中的论述可知, 第一项 I 可以写成 n 个独立同分布的均值为 0 的随机变量之和 $\sum_{i=1}^n \dot{l}_\theta^{(i)}(\theta_0 | \eta_0)$ 加上一些可以忽略不记的无穷小量. 接下来, 我们证明第二项 II 也可以写成这样的形式.

记 $\dot{l}_\theta^{(i)}(\theta | \gamma, \Lambda_C(\tilde{C}_i))$ 为 $l^{(i)}(\theta | \gamma, \Lambda_C(\tilde{C}_i))$ 关于 θ 的一阶偏导数. 定义

$$\ddot{l}_{\theta\gamma}^{(i)}(\theta | \gamma, \Lambda_C(\tilde{C}_i)) = \frac{\partial}{\partial \gamma} \dot{l}_\theta^{(i)}(\theta | \gamma, \Lambda_C(\tilde{C}_i)), \quad \ddot{l}_{\theta\Lambda_C}^{(i)}(\theta | \gamma, \Lambda_C(\tilde{C}_i)) = \frac{\partial}{\partial s} \dot{l}_\theta^{(i)}(\theta | \gamma, s) \Big|_{s=\Lambda_C(\tilde{C}_i)}.$$

那么, 通过多元 Taylor 展开, 可以得到

$$\begin{aligned} II &= \sum_{i=1}^n [\dot{l}_\theta^{(i)}(\theta_0 | \hat{\gamma}, \hat{\Lambda}_C(\tilde{C}_i)) - \dot{l}_\theta^{(i)}(\theta_0 | \gamma_0, \Lambda_{C0}(\tilde{C}_i))] \\ &= \sum_{i=1}^n [\ddot{l}_{\theta\gamma}^{(i)}(\theta_0 | \gamma_0, \Lambda_{C0}(\tilde{C}_i))(\hat{\gamma} - \gamma_0) + \ddot{l}_{\theta\Lambda_C}^{(i)}(\theta_0 | \gamma_0, \Lambda_{C0}(\tilde{C}_i))(\hat{\Lambda}_C(\tilde{C}_i) - \Lambda_{C0}(\tilde{C}_i))] + o_p(1). \end{aligned}$$

进一步, 由 (A.1) 和 (A.2), 有

$$\begin{aligned} II &= \sum_{i=1}^n \{\ddot{l}_{\theta\gamma}^{(i)}(\theta_0 | \gamma_0, \Lambda_{C0}(\tilde{C}_i))E(g_j) + \ddot{l}_{\theta\Lambda_C}^{(i)}(\theta_0 | \gamma_0, \Lambda_{C0}(\tilde{C}_i))E(b_j(\tilde{C}_i))\} + O_p(1) \\ &\triangleq \sum_{i=1}^n d_i(\theta_0, \eta_0) + O_p(1). \end{aligned}$$

于是, 得分函数 $\dot{l}_\theta(\theta_0 | \hat{\eta})$ 可以写成 n 个独立同分布的均值为 0 的随机变量之和加上一些可以忽略不记的误差项的形式, $\dot{l}_\theta(\theta_0 | \hat{\eta}) = \sum_{i=1}^n \{\dot{l}_\theta^{(i)}(\theta_0 | \eta_0) + d_i(\theta_0, \eta_0)\} + O_p(1)$. 因此, $\sqrt{n}(\hat{\theta} - \theta_0)$ 依分布收敛于一个零均值正态随机变量, 协方差矩阵 $\Sigma = \Gamma\phi\phi^T\Gamma$, 其中 $\Gamma = -\{n^{-1}E\{\ddot{l}_\theta(\theta_0 | \hat{\eta})\}\}^{-1}$, $\phi = E\{\dot{l}_\theta^{(1)}(\theta_0 | \eta_0) + d_1(\theta_0, \eta_0)\}$. 这表明 $\hat{\beta}$ 具有渐近正态性, 它的渐近协方差是 Σ 的前 $p \times p$ 顺序主子阵.

A Copula model approach for the additive hazards model with dependent current status data

Hui Zhao, Qi Cui & Jianguo Sun

Abstract The additive hazards model is one of the most commonly used regression models in survival data analysis and many authors have discussed its inference under various situations. However, there exists very little literature on the analysis of dependent current status data under the additive hazards model, and the existing methods assume that the association coefficient between the censoring time and the failure time is known. It is apparent that this may not be true in practice and corresponding to this, in this paper, we propose a new Copula model-based approach that does not need the assumption for analyzing the dependent current status data. The asymptotic properties of the resulting estimators are established, and a simulation study is conducted and suggests that the proposed method performs well for practical situations. Also an illustrative example is provided.

Keywords additive hazards model, Copula model, current status data, dependent censoring

MSC(2010) 62N01, 62N02, 62H20

doi: 10.1360/SCM-2017-0561