应用概率统计 第 38 卷
第 2 期 2022 年 4 月

Chinese Journal of Applied Probability and Statistics
Apr., 2022, Vol. 38, No. 2, pp. 195-218
doi: 10.3969/j.issn.1001-4268.2022.02.003

基于比例风险模型中协变量调整方法的研究 *

荣国才[1] 王亚楠[2,3] 韦程东[3] 邓立凤[1⋆]

([1]山东科技大学数学与系统科学学院, 青岛, 266590; [2]确山县第一高级中学, 驻马店, 463200;
[3]南宁师范大学数学与统计学院, 南宁, 530299)

摘 要: 在实际数据中, 尤其是医学数据, 其协变量受到某些因素的污染或干扰, 而真实的协变量无法观测. 本文所讨论的是在比例风险模型中如何对受干扰的协变量进行调整的问题. 目前所存在的协变量调整方法不能直接用于生存数据, 为了解决这个问题, 我们运用核函数来构造干扰因子的分布函数, 对受干扰的协变量进行平滑得到真实协变量的估计值, 再代入到模型中得到参数的回归估计值, 并完成了估计值满足相合性和渐近正态性证明. 又提出运用极小极大算法 (minorization-maximization algorithm, MM) 得到参数估计值, 第一个 M 是通过指数函数和负对数函数的凸性来构造一个黑塞矩阵为对角矩阵的替代函数; 第二个 M 是对替代函数求最大值. 最后通过数值模拟和真实数据研究来说明我们所提出方法的可行性.

关键词: 协变量调整; 比例风险模型; 渐近性质; 极小极大算法; 自助法

中图分类号: O212.1; O212.4; O212.7

英文引用格式: RONG G C, WANG Y N, WEI C D, et al. Research on covariate adjustment method based on proportional hazards model [J]. Chinese J Appl Probab Statist, 2022, 38(2): 195–218. (in Chinese)

§1. 引 言

生存数据存在于很多领域, 例如金融、公共卫生、流行病学、医学等. 在观测时间范围之内, 当被观测的个体发生了所关心的事件时, 就定义个体发生了死亡. 当个体在发生死亡之前退出或被移除, 就定义个体发生了右删失. 对于这种带有右删失的生存数据, 我们采用应用较为广泛的比例风险模型 [1] 进行研究. 比例风险模型一经提出, 许多学者对其进行深入研究, 并在多方面进行推广和延伸. 如 Cao 等 [2] 提出了广义病例队列设计下比例风险模型中对未知参数估计的推理方法, 并建立了在给定预算下获得最大功率的最优样本容量分配模型; Ding 等 [3] 在比例风险模型中针对参数带约束条件的估计问题提出了一种新的 MM 算法, 先将高维相乘的对数似然函数转换成一维相加的代理函数, 再将参数本身所带有的约束条件通过中位数加到算法中, 得到参数的最优值; Ji 等 [4] 在 Box-Cox 变换模型中

*山东科技大学人才引进科研启动基金项目 (批准号: 2019RCJJ021) 和广西高等教育本科教学改革工程项目 (批准号: 2014JGB415) 资助.

⋆通讯作者, E-mail: laiji1234@163.com.

本文 2020 年 1 月 15 日收到, 2020 年 11 月 8 日收到修改稿.

提出了两步方法针对变系数进行了估计. Liu 等[5] 在可加可乘模型中运用 B 样条插值法对非参数累积风险函数进行了估计, 并提出了同时筛选极大似然估计方法用于估计回归参数; Hamad 和 Kachouie[6] 提出了将 Kaplan Meier 的生存函数的非参数估计方法与在比例风险模型中部分似然法的逻辑函数的参数估计方法结合起来以估计比例风险模型中的基底风险函数.

在许多实际数据中, 由于数据本身的属性, 所收集到数据可能被污染, 即协变量受到了非随机性因子的干扰, 而真实的协变量无法观测到. 例如, 在肾脏疾病饮食改良研究[7,8] 中, 只能观测到受到体表面积干扰的肾小球过滤率和血清肌酐的数据, 而其真实数据无法观测到. 若直接运用这些受干扰的协变量进行分析, 可能得到有偏的参数估计, 从而导致错误的统计推断. 对于完整数据受污染问题, Şentürk 和 Müller[9] 在线性回归模型中提出了参数估计值调整方法, 它的思路是首先对可观测到的受干扰的协变量和响应变量进行回归分析, 再对得到的参数估计值除以干扰因子函数, 将干扰因素去掉, 从而得到真正的参数估计值. 但对于更多的一般模型, 找到这种直接有用的干扰函数是非常困难的. Cui 等[10] 在非线性回归模型中提出协变量调整方法, 它的思路是先对受干扰的协变量和响应变量进行调整, 得到两者的估计值再进行回归分析从而得到真正的参数估计值的分布函数. Li 等[11] 将协变量调整方法[10] 推广到部分线性回归模型中; Ma 和 Luan[12] 将参数调整方法[9] 运用到时间序列模型中; Li 和 Lu[13] 在高维线性回归模型中, 首先运用协变量调整方法对受干扰的协变量进行调整, 再运用 Lasso 进行变量筛选; Lu 等[14] 在变系数模型中运用非参数方法对受干扰的变量进行调整, 再运用局部最小二乘法得到参数估计.

本论文所讨论的数据为受污染的生存数据, 数据中包含观测时间、事件是否发生的示性变量、受干扰的协变量和干扰因子. 这里可以把观测时间和时间是否发生的示性变量看作是带有右删失的响应变量. 这种类型的数据与其他非生存数据相似, 因此我们采用协变量调整方法[10], 先运用核函数来构造干扰因子的分布函数, 再对受干扰的协变量除以分布函数, 将干扰因子去掉得到调整后的协变量估计值, 再带入比例风险模型中进行回归分析得到参数估计值. 由于协变量受到干扰, 所以在参数估计过程中会出现黑塞矩阵不可逆的问题. 而极小极大 (Minorization-Maximization) 算法[15] 正可以解决这个问题. 它的基本思路是通过不等式的放缩找到原似然函数的替代函数, 再对替代函数求最优解. 当替代函数取得最优解时, 原似然函数也得到最优解. 本文是在比例风险模型中求解参数的极大似然估计值. 为了解决运算过程中黑塞矩阵不可逆的问题, 我们通过指数函数 e^x 和负对数函数 $-\ln x$ 的凸性来得到替代函数, 并且其对应的黑塞矩阵为对角矩阵, 再通过对此替代函数求最大值得到了参数的估计值.

下面是对本文的章节安排, 第二章介绍比例风险模型中协变量调整方法; 第三章介绍参数估计的渐近性质; 第四章介绍了比例风险模型中的 MM 算法、CV 准则和 Bootstrap 方法; 第五章给出了几种模拟研究, 比较了协变量不受干扰、受干扰和受干扰调整后的估计值; 第六章针对心力衰竭患者的数据进行了真实数据分析; 第七章介绍了本文研究的主

要结论.

§2. 协变量调整方法

1) 模型的介绍

设有 n 个观测样本, $\widetilde{T}_i$ 表示第 i 个样本的死亡时间, C_i 表示第 i 个样本的右删失时间, T_i 表示第 i 个样本的观测时间, $T_i = \min\{\widetilde{T}_i, C_i\}$; $\Delta_i = I(\widetilde{T}_i \leqslant C_i)$ 表示第 i 个样本是否发生了右删失; $Y_i(t) = I(T_i \geqslant t)$ 表示第 i 个样本是否历险; $N_i(t) = \Delta_i I(T_i \leqslant t)$ 表示第 i 个样本的计数过程; 其中 $I(\cdot)$ 为示性函数.

Cox[1] 提出了著名的比例风险模型, 如下所示

$$x(t \mid Z) = x_0(t) \exp\{\beta^{\mathsf{T}} Z\},$$

其中, $x_0(t)$ 是基底风险函数, $Z = (Z_1, Z_2, \cdots, Z_p)^{\mathsf{T}}$ 为 p 维协变量, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^{\mathsf{T}}$ 为 p 维回归参数. 采用部分似然函数[16] 来估计 β.

$$L(\beta) = \prod_{i=1}^{n} \Big\{ \exp\{\beta^{\mathsf{T}} Z_i\} \Big/ \Big[\sum_{l=1}^{n} Y_l(T_i) \exp\{\beta^{\mathsf{T}} Z_l\} \Big] \Big\}^{\Delta_i}.$$

其对应的对数部分似然函数为

$$l(\beta) = \sum_{i=1}^{n} \Delta_i \Big\{ \beta^{\mathsf{T}} Z_i - \ln \Big[\sum_{l=1}^{n} Y_l(T_i) \exp\{\beta^{\mathsf{T}} Z_l\} \Big] \Big\}, \tag{1}$$

对应的得分方程为

$$F(\beta) = \sum_{i=1}^{n} \Delta_i \Big[Z_i - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \Big],$$

其中

$$S^{(k)}(\beta, t) = \frac{1}{n} \sum_{l=1}^{n} Y_l(t) Z_l(t)^{\otimes k} \exp\{\beta^{\mathsf{T}} Z_l(t)\}, \qquad k = 0, 1, 2,$$

其中 $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, $a^{\otimes 2} = aa^{\mathsf{T}}$.

$$E(\beta, t) = \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}, \qquad V(\beta, t) = \frac{S^{(2)}(\beta, t)}{S^{(0)}(\beta, t)} - E(\beta, t)^{\otimes 2}.$$

其极大似然估计值 $\widehat{\beta}$ 可由得分方程 $F(\beta) = 0$ 求解得到.

2) 协变量调整

前面所提到的 Z 为未受干扰的协变量, 本小节研究的问题是协变量受到干扰的情况, 这也是本文的重点所在. 令 $\widetilde{Z}$ 表示受干扰的协变量, U 表示干扰因子. 我们假定 $\widetilde{Z}_{ij}$ 与 Z_{ij} 的关系式可写为

$$\widetilde{Z}_{ij} = \phi_j(U_i) Z_{ij}, \tag{2}$$

其中函数 $\phi_j(U)$, $j=1,2,\cdots,p$ 是未知的, 协变量 $\widetilde{Z}$ 和干扰因子 U 是可观测的, 而未受到干扰的真正的协变量 Z 是无法观测到的. 此时观测数据结构可表示为 $(T_i,\Delta_i,\widetilde{Z}_i,U_i)$, $i=1,2,\cdots,n$. 因此得分方程 (1) 式的协变量为 $\widetilde{Z}$, 其所对应的得分方程为

$$\widetilde{F}(\beta)=\sum_{i=1}^{n}\Delta_i\Big[\widetilde{Z}_i-\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\Big], \tag{3}$$

如果直接利用 (3) 式求解参数的估计必然会得到错误的结果, 这是因为协变量受到了干扰, 而真正的协变量却无法观测到, 在第 5 章的模拟实验中给出了具体的比较结果. 为了避免这个错误, 受 Cui 等[10] 思路的启发, 我们采用核平滑方法对协变量进行调整. 首先给出文中用到的基本假设

(a) $\mathsf{E}[\phi_j(U)]=1$;

(b) (Z_j,U) 之间是相互独立.

由这两个基本假设可得

$$\begin{aligned}\mathsf{E}[\mathsf{E}(\widetilde{Z}_{ij}\,|\,U_i)]&=\mathsf{E}[\mathsf{E}(\phi_j(U_i)Z_{ij}\,|\,U_i)]=\mathsf{E}[\phi_j(U_i)\mathsf{E}(Z_{ij}\,|\,U_i)]\\&=\mathsf{E}[\phi_j(U_i)\mathsf{E}(Z_{ij})]=\mathsf{E}[\phi_j(U_i)]Z_{ij}=Z_{ij}.\end{aligned} \tag{4}$$

现在我们的目标是基于观测数据 $(T,\Delta,\widetilde{Z},U)$ 来估计未知参数 β, 由前面的假设可知

$$\phi_j(U)=\frac{\mathsf{E}(\widetilde{Z}_j\,|\,U)}{\mathsf{E}(Z_j)},\qquad 1\leqslant j\leqslant p.$$

但是 U 的分布函数是未知的, 所以 (2) 式是无法求出的. 为了解决这个问题, 我们通过核方法来构造干扰函数 $\phi_j(U)$, 如下所示

$$\widehat{\phi}_j(u)=\frac{(nh)^{-1}\sum\limits_{i=1}^{n}K((u-U_i)/h)\widetilde{Z}_{ij}}{(nh)^{-1}\sum\limits_{i=1}^{n}K((u-U_i)/h)}\times\frac{1}{\overline{\widetilde{Z}_j}}\triangleq\frac{\widehat{g}_j(u)}{\widehat{p}(u)}\times\frac{1}{\overline{\widetilde{Z}_j}}, \tag{5}$$

其中 $\overline{\widetilde{Z}_j}=n^{-1}\sum\limits_{i=1}^{n}\widetilde{Z}_{ij}$, h 是一个带宽, $K(\cdot)$ 是一个核函数. 由 (2) 式和 (5) 式可得

$$\widehat{Z}_{ij}=\frac{\widetilde{Z}_{ij}}{\widehat{\phi}_j(U_i)},\qquad \widehat{Z}_i=(\widehat{Z}_{i1},\widehat{Z}_{i2},\cdots,\widehat{Z}_{ip})^{\mathsf{T}}. \tag{6}$$

此时 $\widehat{Z}$ 表示经过调整后的协变量, 此时得分方程应改写为

$$\widehat{F}(\beta)=\sum_{i=1}^{n}\Delta_i\Big[\widehat{Z}_i-\frac{\widehat{S}^{(1)}(\beta,t)}{\widehat{S}^{(0)}(\beta,t)}\Big], \tag{7}$$

其极大似然估计值 $\widehat{\beta}_I$ 可由得分方程得出

$$\widehat{F}(\beta)=0. \tag{8}$$

§3. 渐近性质

为了证明渐近性, 我们首先引入一些符号说明. β_0 为 β 的真值. τ 为生存研究停止的时间. $p(U)$ 为干扰因子 U 的概率密度函数, 定义

$$g_{\widetilde{Z}}(U)=\mathsf{E}(\widetilde{Z}\,|\,U)p(U),\qquad g_{\widetilde{Z}_j}(U)=\mathsf{E}(\widetilde{Z}_j\,|\,U)p(U).$$

$$g_{\widetilde{Z}}(U)=\frac{(nh)^{-1}\sum\limits_{i=1}^{n}K((u-U_i)/h)\widetilde{Z}_i}{(nh)^{-1}\sum\limits_{i=1}^{n}K((u-U_i)/h)},\qquad g_{\widetilde{Z}_j}(U)=\frac{(nh)^{-1}\sum\limits_{i=1}^{n}K((u-U_i)/h)\widetilde{Z}_{ij}}{(nh)^{-1}\sum\limits_{i=1}^{n}K((u-U_i)/h)}.$$

Andersen 和 Gill[17] 讨论了如何将比例风险模型扩展到在多变量计数过程的情况中, 并对删失情况和时间相依协变量进行了回归分析, 推导出估计量的渐近性质.

$$\widehat{S}^{(k)}(\beta,t)=\frac{1}{n}\sum_{l=1}^{n}Y_l(t)\widehat{Z}_l(t)^{\otimes k}\exp\{\beta^{\mathsf{T}}\widehat{Z}_l(t)\},\qquad k=0,1,2,$$

其中 $a^{\otimes 0}=1$, $a^{\otimes 1}=a$, $a^{\otimes 2}=aa^{\mathsf{T}}$.

下面再给出贯穿全文的假设条件:

(A_1) 核函数 $K(\cdot)$ 的支撑区间为 $[-1,1]$, 关于 0 对称, 且满足

$$\int_{-1}^{1}K(u)\,\mathrm{d}u=1,\qquad \int_{-1}^{1}u^iK(u)\,\mathrm{d}u=0,\quad i=1,2,3.$$

(A_2) 当 $n\to\infty$ 时, $\exists\,\delta>0$, $\eta>2$, 都有 $h\to 0$, $n^{1-2\eta^{-1}-2\delta}h\to 0$ 成立.

(A_3) 设存在 (A_2) 中的 η 和区间 $[a,b]$ 成立, $\mathsf{E}|Z|^{\eta}<\infty$ 成立, 且

$$\sup_{u\in[a,b]}\int|z|^{\eta}f(u,z)\,\mathrm{d}x<\infty,$$

其中, f 表示 (U,Z) 的联合密度函数.

(A_4) 对所有 $g_{\widetilde{Z}_j}(u)=\phi_j(u)p(u)$, 其中 $\phi_j(u)$ 与 $p(u)$ 是与 0 较远的正常数, 且这三个函数是可微的, 其导数满足以下条件: 存在某一领域 Δ 和常数 $c>0$, 对任意的 $\delta\in\Delta$, 有

$$\big|g^{(3)}_{\widetilde{Z}_j}(u+\delta)-g^{(3)}_{\widetilde{Z}_j}(u)\big|\leqslant c|\delta|,\quad 1\leqslant j\leqslant p,\qquad |p^{(3)}(u+\delta)-p^{(3)}(u)|\leqslant c|\delta|.$$

(B_1) $\int_0^{\tau}x_0(t)\,\mathrm{d}t<\infty$.

(B_2) 存在某一 $\delta>0$, 有

$$n^{-1/2}\sup_{l;t\in[0,\tau]}|Z_l|Y_l(t)I\{\beta_0^{\mathsf{T}}Z_l(t)>-\delta|Z_l(t)|\}\xrightarrow{\mathsf{P}}0.$$

(B_3) 存在定义在 $\mathscr{B}\times[0,\tau]$ 上的三个函数 $s^{(0)}(\beta,t)$, $s^{(1)}(\beta,t)$ 和 $s^{(2)}(\beta,t)$, 满足以下三个条件

(i) $\sup\limits_{t\in[0,\tau];\beta\in\mathscr{B}}\|S^{(k)}(\beta,t)-s^{(k)}(\beta,t)\|\xrightarrow{\mathsf{P}}0$, $k=0,1,2$;

(ii) 对 $\beta\in\mathscr{B}$, $t\in[0,\tau]$, 由 $\beta\to s^{(k)}(\beta,t)$ 为连续的, $k=0,1,2$; $s^{(1)}(\beta,t)=(\partial/\partial\beta)\cdot s^{(0)}(\beta,t)$, $s^{(2)}(\beta,t)=(\partial^2/\partial\beta^2)s^{(0)}(\beta,t)$. $s^{(0)}(\beta,t)$ 是有界的且远离零;

(iii) 矩阵

$$\Sigma(\beta)=\int_0^\tau \nu(\beta,t)s^{(0)}(\beta,t)x_0(t)\,\mathrm{d}t,$$

在 β_0 处是正定的, 其中

$$e(\beta,t)=\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)},\qquad \nu(\beta,t)=\frac{s^{(2)}(\beta,t)}{s^{(0)}(\beta,t)}-e(\beta,t)^{\otimes 2}.$$

运用以上条件, 我们可以得到下面的结论, 具体的证明过程在附录中.

定理 1 由假设条件 (A_1-A_4), (B_1-B_3), 对受干扰的协变量调整后所得 $\widehat{\beta}_I$ 具有相合性, 即

$$\widehat{\beta}_I\xrightarrow{\mathsf{P}}\beta_0,\qquad n\to\infty.$$

定理 2 由假设条件 (B_1-B_3) 和引理 5 (见附录), $\widehat{\beta}_I$ 具有渐近正态性, 即

$$\sqrt{n}\,(\widehat{\beta}_I-\beta_0)\xrightarrow{\mathrm{D}}\mathrm{N}(0,[\widehat{\Sigma}(\beta_0)]^{-1}).$$

§4. 算法实现

在第 3 章中, 我们给出了协变量调整后所得参数估计值的渐近性质并完成了证明. 但是对 (8) 式的优化运算是非常复杂的, 尤其是当协变量受到干扰时, 若运用经典的 Newton-Rapson 算法迭代, 可能会出现黑塞矩阵不可逆的问题. 为了方便快捷地计算出参数估计值, 并便于说明所提出的协变量调整方法的有效性, 我们需要找到一个可行算法. 首先给出 (7) 式 $\widehat{F}(\beta)$ 的工作函数

$$\widehat{l}(\beta)=\sum_{i=1}^n\Delta_i\Big\{\beta^T\widehat{Z}_i-\ln\Big[\sum_{l=1}^n Y_l(T_i)\exp\{\beta^{\mathsf{T}}\widehat{Z}_l\}\Big]\Big\}. \tag{9}$$

则求解得分方程 (8) 式与 $\widehat{l}(\beta)$ 的优化问题等价.

Ding 等[3] 在比例风险模型中针对参数带约束条件的优化问题提出了新的 MM 算法. 结合这个思路, 我们将工作函数求解最大值问题转化为对带有对角黑塞矩阵的替代函数 $Q(\beta\,|\,\beta^{(m)})$ 求最大值问题.

1) MM 算法

MM 算法的最关键的问题是构造替代函数, 将复杂的目标函数转化为较简单的替代函数. 本文依照 Ding 等[3] 构造替代函数的思路, 首先根据指数函数 e^x 的凸性, 对任意的正数 $\{\alpha_k\}_{k=1}^p$ 满足 $\sum\limits_{k=1}^{p}\alpha_k=1$, 有

$$\mathrm{e}^{\sum\limits_{k=1}^{p}\alpha_k x_k}\leqslant\sum_{k=1}^{p}\alpha_k\mathrm{e}^{x_k}. \tag{10}$$

令 $\widehat{Z}_l=\big(\widehat{Z}_{l1},\widehat{Z}_{l2},\cdots,\widehat{Z}_{lp}\big)^{\mathsf{T}}$ 表示第 l 个样本观测值, $\beta^{(m)}=\big(\beta_1^{(m)},\beta_2^{(m)},\cdots,\beta_p^{(m)}\big)^{\mathsf{T}}$ 表示 (9) 式中参数估计值 $\widehat{\beta}$ 的第 m 次逼近值. 对 (9) 式中的指数项, 有

$$\begin{aligned}\exp\big\{\beta^{\mathsf{T}}\widehat{Z}_l\big\}&=\exp\big\{\widehat{Z}_l^{\mathsf{T}}(\beta-\beta^{(m)})\big\}\cdot\exp\big\{\widehat{Z}_l^{\mathsf{T}}\beta^{(m)}\big\}\\&=\exp\Big\{\sum_{k=1}^{p}\lambda_{lk}\big[\lambda_{lk}^{-1}\widehat{Z}_{lk}(\beta_k-\beta_k^{(m)})\big]\Big\}\cdot\exp\big\{\widehat{Z}_l^{\mathsf{T}}\beta^{(m)}\big\}\\&\leqslant\Big\{\sum_{k=1}^{p}\lambda_{lk}\exp\big\{\lambda_{lk}^{-1}\widehat{Z}_{lk}(\beta_k-\beta_k^{(m)})\big\}\Big\}\cdot\exp\big\{\widehat{Z}_l^{\mathsf{T}}\beta^{(m)}\big\}\\&=\sum_{k=1}^{p}\lambda_{lk}\exp\big\{\lambda_{lk}^{-1}\widehat{Z}_{lk}(\beta_k-\beta_k^{(m)})+\widehat{Z}_l^{\mathsf{T}}\beta^{(m)}\big\}.\end{aligned}$$

其中 $\lambda_{lk}=|\widehat{Z}_{lk}|\big/\big[\sum\limits_{k'=1}^{p}|\widehat{Z}'_{lk}(T_i)|\big]$. 若 $\widehat{Z}_{lk}=0$, 则 $\lambda_{lk}^{-1}\widehat{=}\,0$. 再由不等式 $-\ln x\geqslant 1-\ln y-x/y$, 令

$$x=\sum_{l=1}^{n}\omega_l Y_l(T_i)\exp\{\beta^{\mathsf{T}}\widehat{Z}_l\},\qquad y=\sum_{l=1}^{n}\omega_l Y_l(T_i)\exp\{\widehat{Z}_l^{\mathsf{T}}\beta^{(m)}\},$$

可得

$$\begin{aligned}&-\ln\Big[\sum_{l=1}^{n}\omega_l Y_l(T_i)\exp\big\{\beta^{\mathsf{T}}\widehat{Z}_l\big\}\Big]\\&\geqslant 1-\ln\Big[\sum_{l=1}^{n}\omega_l Y_l(T_i)\exp\big\{\widehat{Z}_l\beta^{(m)}\big\}\Big]-\frac{\sum\limits_{l=1}^{n}\omega_l Y_l(T_i)\exp\big\{\beta^{\mathsf{T}}\widehat{Z}_l\big\}}{\sum\limits_{l=1}^{n}\omega_l Y_l(T_i)\exp\big\{\widehat{Z}_l\beta^{(m)}\big\}}\\&\geqslant 1-\ln\Big[\sum_{l=1}^{n}\omega_l Y_l(T_i)\exp\big\{\widehat{Z}_l\beta^{(m)}\big\}\Big]-\frac{\sum\limits_{l=1}^{n}\sum\limits_{k=1}^{p}\omega_l\lambda_{lk}\exp\big\{\lambda_{lk}^{-1}\widehat{Z}_{lk}(\beta_k-\beta_k^{(m)})+\widehat{Z}_l^{\mathsf{T}}\beta^{(m)}\big\}}{\sum\limits_{l=1}^{n}\omega_l Y_l(T_i)\exp\big\{\widehat{Z}_l^{\mathsf{T}}\beta^{(m)}\big\}}.\end{aligned}$$

根据 (10) 式, 得到 $\widehat{l}(\beta)$ 的替代函数 $Q(\beta\,|\,\beta^{(m)})$.

$$Q(\beta\,|\,\beta^{(m)})=c_0+\sum_{i=1}^{n}\Delta_i\left[\beta^{\mathsf{T}}\widehat{Z}_i-\frac{\sum\limits_{l=1}^{n}\sum\limits_{k=1}^{p}Y_l(T_i)\lambda_{lk}g_{lk}(\beta_k\,|\,\beta^{(m)})}{\sum\limits_{l=1}^{n}Y_l(T_i)\exp\big\{\beta^{(m)\mathsf{T}}\widehat{Z}_l\big\}}\right], \tag{11}$$

其中 $c_0=\sum\limits_{i=1}^{n}\Delta_i\big\{1-\ln\big[\sum\limits_{l=1}^{n}Y_l(T_i)\exp\{\beta^{(m)\mathsf{T}}\widehat{Z}_l\}\big]\big\}$, $\lambda_{lk}=|\widehat{Z}_{lk}|\big/\big(\sum\limits_{k'=1}^{p}|\widehat{Z}_{lk'}|\big)$, $k=1,2,\cdots,p$, $g_{lk}(\beta_k\,|\,\beta^{(m)})=\exp\{\lambda_{lk}^{-1}\widehat{Z}_{lk}(\beta_k-\beta_k^{(m)})+\beta^{(m)\mathsf{T}}\widehat{Z}_l\}$.

由 (11) 式的构造过程可知 $\widehat{l}(\beta)\geqslant Q(\beta\,|\,\beta^{(m)})$, 当且仅当 $\beta=\beta^{(m)}$ 时, $\widehat{l}(\beta)=Q(\beta\,|\,\beta^{(m)})$. 再由 $Q(\beta^{(m+1)}\,|\,\beta^{(m)})\geqslant Q(\beta^{(m)}\,|\,\beta^{(m)})$ 可得

$$\begin{aligned}\widehat{l}(\beta^{(m+1)})&=\widehat{l}(\beta^{(m+1)})-Q(\beta^{(m+1)}\,|\,\beta^{(m)})+Q(\beta^{(m+1)}\,|\,\beta^{(m)})\\&\geqslant\widehat{l}(\beta^{(m)})-Q(\beta^{(m)}\,|\,\beta^{(m)})+Q(\beta^{(m)}\,|\,\beta^{(m)})\\&=\widehat{l}(\beta^{(m)}).\end{aligned}$$

由此可以得出 $\widehat{l}(\beta)$ 为严格单调递增函数, 因此对于协变量调整后的部分似然函数 $\widehat{l}(\beta)$ 的优化问题可转化为替代函数 $Q(\beta\,|\,\beta^{(m)})$ 的优化问题.

$$\beta^{(m+1)}=\arg\max_{\beta}Q(\beta\,|\,\beta^{(m)}),$$

且当 $m\to\infty$ 时, 保证 $\beta^{(m)}$ 收敛于 $\widehat{\beta}_I$.

并由 (11) 式可求得当 $\beta=\beta^{(m)}$ 时, $Q(\beta\,|\,\beta^{(m)})$ 的得分向量为

$$\frac{\partial Q(\beta^{(m)}\,|\,\beta^{(m)})}{\partial\beta}=\sum_{l=1}^{n}\Delta_i\left[\widehat{Z}_i-\frac{\sum\limits_{l=1}^{n}Y_l(T_i)\widehat{Z}_l\exp\big\{\beta^{(m)\mathsf{T}}\widehat{Z}_l\big\}}{\sum\limits_{l=1}^{n}Y_l(T_i)\exp\big\{\beta^{(m)\mathsf{T}}\widehat{Z}_l\big\}}\right],\tag{12}$$

以及负的黑塞矩阵为

$$-\frac{\partial^2Q(\beta^{(m)}\,|\,\beta^{(m)})}{\partial\beta\partial\beta^{\mathsf{T}}}=\mathrm{diag}\Big(-\frac{\partial^2Q(\beta^{(m)}\,|\,\beta^{(m)})}{\partial\beta_1^2},\cdots,-\frac{\partial^2Q(\beta^{(m)}\,|\,\beta^{(m)})}{\partial\beta_p^2}\Big),\tag{13}$$

其中

$$-\frac{\partial^2Q(\beta^{(m)}\,|\,\beta^{(m)})}{\partial\beta_k^2}=\sum_{l=1}^{n}\Delta_i\frac{\sum\limits_{l=1}^{n}Y_l(T_i)(\widehat{Z}_{lk}^2/\lambda_{lk})\exp\big\{\beta^{(m)\mathsf{T}}\widehat{Z}_l\big\}}{\sum\limits_{l=1}^{n}Y_l(T_i)\exp\big\{\beta^{(m)\mathsf{T}}\widehat{Z}_l\big\}}.$$

由 (13) 式可知 $Q(\beta^{(m)}\,|\,\beta^{(m)})$ 的黑塞矩阵为对角矩阵必可逆. 因此可以采用 Newton-Raphson 迭代方法得到参数的估计值.

将前面所提到的 MM 算法总结如下:

第一步: 利用指数函数和负对数函数的凸性构造出 $\widehat{l}(\beta)$ 的替代函数 $Q(\beta^{(m)}\,|\,\beta^{(m)})$;

第二步: 求出 $Q(\beta^{(m)}\,|\,\beta^{(m)})$ 的得分向量 (12) 式和负黑塞矩阵 (13) 式, 并运用下面的公式得到 $\beta^{(m+1)}$,

$$\beta^{(m+1)}=\beta^{(m)}+\Big[-\frac{\partial^2Q(\beta^{(m)}\,|\,\beta^{(m)})}{\partial\beta\partial\beta^{\mathsf{T}}}\Big]^{-1}\frac{\partial^2Q(\beta^{(m)}\,|\,\beta^{(m)})}{\partial\beta\partial\beta^{\mathsf{T}}}.$$

2) CV 准则和 Bootstrap 方法

CV 准则

在对受干扰的协变量进行调整时, 采用了非参数核平滑法, 其中光滑参数带宽 h 选用 CV 准则得到最优带宽. 主要思想是在一部分已知的数据上建立模型, 然后利用建立的模型预测剩余的数据, 并计算平均预测误差, 当平均预测误差达到最小时所得的模型是最好的. 其中 CV 选择函数为

$$\mathrm{CV}(h)=\sum_{j=1}^{n}\Big[\frac{1}{nh}\sum_{i=1}^{n}K\Big(\frac{u_j-u_i}{h}\Big)\Big]^2(u_j-u_{j-1})-\frac{2}{n}\sum_{j=1}^{n}\Big[\frac{2}{nh}\sum_{j\neq i}K\Big(\frac{u_j-u_{-i}}{h}\Big)\Big],$$

其中 u_{-i} 表示 $u_1,u_2,\cdots,u_n$ 中去掉第 i 个元素 u_{-i}, 最优步长 h 的函数定义为

$$\widehat{h}=\arg\min_{h}\mathrm{CV}(h).$$

Bootstrap 方法

对 $\widehat{\beta}$ 的标准误差计算, 我们采用非参数 Bootstrap 方法, 其主要思想是通过对观测样本的放回式抽样得到经验分布函数, 具体的步骤为:

第一步: 设经过调整的观测样本数据为 $X_1,X_2,\cdots,X_n$, 其中 $X_i=(T_i,\Delta_i,\widehat{Z}_i)$, 运用 Bootstrap 抽样方法从中放回式抽取样本 $\{X_1^*,X_2^*,\cdots,X_n^*\}$, 其中 $X_i^*=(T_i,\Delta_i^*,\widehat{Z}_i^*)$.

第二步: 由 $\{X_1^*,X_2^*,\cdots,X_n^*\}$ 来计算 $\widehat{\beta}^*$.

第三步: 重复第二步 B 次, 得到 B 个参数估计值 $\{\beta^*(1),\beta^*(2),\cdots,\beta^*(B)\}$, 则参数估计值 $\widehat{\beta}$ 的第 k 个元素的标准误差计算公式为

$$\widehat{\mathrm{Se}}(\widehat{\beta}_k)=\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\Big[\widehat{\beta}_k^*(b)-\frac{1}{B}\sum_{b=1}^{B}\widehat{\beta}_k^*(b)\Big]^2},\qquad k=1,2,\cdots,p.$$

§5. 模拟计算

正如前面多提到的, 在分析过程中若忽视协变量受干扰的情况可能得到错误的结论[9,10]. 我们构造了几种模拟来比较协变量是否受干扰以及受干扰后调整得到的参数估计值的情况, 其中 $\widehat{\beta}_O$ 表示协变量未受干扰时参数估计值, $\widehat{\beta}_P$ 表示协变量受干扰时参数估计值, $\widehat{\beta}_I$ 表示协变量受干扰经调整后的参数估计值. 我们构建比例风险模型, 在给定协变量 (Z_1,Z_2) 条件下, 死亡时间 $\widetilde{T}$ 的风险函数为

$$x(t\,|\,Z)=x_0(t)\exp\{\beta_1Z_1+\beta_2Z_1\}.$$

进一步, 我们运用两种算法分别估计回归参数 β_1,β_2, 算法 I: Newton-Raphson 算法; 算法 II: MM 算法.

1) Newton-Raphson 算法

运用 Newton-Raphson 算法时, 需计算对数似然函数的黑塞矩阵. 但当样本量较小时, 尤其当样本量受到干扰时, 黑塞矩阵可能出现不可逆的情况. 为避免这种情况的出现, 这里所取的样本量较大.

情况 I: 设协变量 $Z_1 \sim \mathrm{N}(2, 1.44)$, $Z_2 \sim \mathrm{U}(7/3 - \sqrt{19}/2,\ 7/3 + \sqrt{19}/2)$; 干扰因子 $U \sim \mathrm{U}(4 - \sqrt{7},\ 4 + \sqrt{7})$. 设参数 (β_1, β_2) 真值为: $(-0.5, 0.693)$. 基底风险函数 $x_0(t)$ 分别取为 1, $2t$, $3t^2$, 则死亡时间 $\widetilde{T}$ 的边际分布函数分别为失败率为 $\exp\{\beta_1 Z_1 + \beta_2 Z_2\}$ 的指数分布、形状参数为 2 与尺度参数为 $[\exp\{\beta_1 Z_1 + \beta_2 Z_2\}]^{-1/2}$ 的威布尔分布以及形状参数为 3 与尺度参数为 $[\exp\{\beta_1 Z_1 + \beta_2 Z_2\}]^{-1/3}$ 的威布尔分布. 删失时间 C 是由 $\mathrm{U}(0, c)$ 上产生的随机数, 其中 c 的选定要依据删失率为 $\rho = 30\%, 50\%, 60\%, 80\%$ 来取值. 这里取样本量为 500.

对每一种设置, 我们都比较了三种回归参数的估计值 $\widehat{\beta}_O$, $\widehat{\beta}_P$, $\widehat{\beta}_I$, 并通过 1 000 次独立产生模拟数据计算得到样本偏差 (BIAS)、标准均方误差 (SMSE)、估计的样本标准差 (SD)、以及运用 Bootstrap 方法中提到的 500 次非参数自助法得到的估计标准差的均值 (SE)、95% 的置信区间的收敛概率 (CP), 这里只讨论协变量 X 干扰因子的分布函数是否受到干扰, $\phi(u) = (u + 10)^2/194.9160$, 其核函数为高阶核函数 $K(t) = (15/32)(3 - 7t^2)$ $\cdot(1 - t^2)$ $(|t| \leqslant 1)$, 运用 CV 准则选择最优步长, 计算精度设置为 $\varepsilon = 0.0001$, 模拟结果汇总表 1 中.

情况II: 在第一种情况中, 协变量 Z_1, Z_2 都满足连续分布, 这里假设 $Z_1 \sim \mathrm{B}(1, 0.5)$, $Z_2 \sim \mathrm{U}(7/3 - \sqrt{19}/2,\ 7/3 + \sqrt{19}/2)$; 干扰因子 $U \sim \mathrm{U}(4 - \sqrt{7},\ 4 + \sqrt{7})$. 参数真值为 $(-0.5, 0.5)$, $(-0.5, 0.693)$, $(-0.25, 0.5)$, 基底风险函数 $x_0(t)$ 分别取为 1, $2t$, 即死亡时间 $\widetilde{T}$ 分别满足失败率为 $\exp\{\beta_1 Z_1 + \beta_2 Z_2\}$ 的指数分布和形状参数为 2 与尺度参数为 $[\exp\{\beta_1 Z_1$ $+ \beta_2 Z_2\}]^{-1/2}$ 的威布尔分布. 这里设定删失时间 C 与协变量相关, 满足 $E(1)I(Z_1 = 0)$ $+ E(1/3)I(Z_1 = 1)$. 样本容量为 300 和 500, 非参数自助法设置为 100 次, 结果汇总于表 2.

在所有的设定下, 我们可以看出协变量未受干扰和受干扰调整后所得的参数估计值 $\widehat{\beta}_O$ 和 $\widehat{\beta}_I$ 都是无偏的. 其 SMSE、SD 和 SE 三者比较接近, CP 值都在比较合理范围之内. 当协变量 Z_2 受干扰时, 参数 β_2 的估计值为有偏的, SMSE 与 SD 的差距较大, 且 CP 值较低. 这也说明若忽视协变量受干扰的情况会造成估计错误. 在经过非参数调整后, 参数 β_2 的估计值结果都较为合理. 例如, 在表 1 中, 当 $x_0(t) = 1$, $\rho = 30\%$ 时, β_2 在协变量 Z_2 受干扰和受干扰后调整的偏差分别为 -0.1768 和 -0.0125, SMSE 分别为 0.1818 和 0.0542, CP 值分别为 0.022 和 0.935. 因为模拟过程中只设定了对协变量 Z_2 进行干扰, 所以参数 β_1 的估计值为无偏的, 且 CP 值也在合理范围之内, 这些结果也说明我们提出的协变量调整方法是非常有效的. 在表 2 中, 由于删失时间 C 与协变量相关, 所以删失率各不相同. 并且在相同条件下, 随着样本量的增加, SMSE、SD 和 SE 的值均在减小.

表 1　$N=500$, $\beta=(-0.5,0.693)$, $Z_1\sim \mathrm{N}(2,1.44)$, $Z_2\sim \mathrm{U}(7/3-\sqrt{19}/2,7/3+\sqrt{19}/2)$, $u\sim \mathrm{U}(4-\sqrt{7},4+\sqrt{7})$

ρ	Type	$x_0(t)=1$									
		BIAS	SMSE	SD	SE	CP	BIAS	SMSE	SD	SE	CP
30%	$\widehat{\beta}_O$	-0.0029	0.0424	0.0423	0.0429	0.948	0.0035	0.0512	0.0511	0.0510	0.957
	$\widehat{\beta}_P$	0.0304	0.0521	0.0423	0.0426	0.890	-0.1774	0.1826	0.0431	0.0420	0.015
	$\widehat{\beta}_I$	0.0023	0.0424	0.0424	0.0429	0.945	-0.0138	0.0551	0.0533	0.0509	0.927
50%	$\widehat{\beta}_O$	-0.0020	0.0494	0.0493	0.0496	0.953	0.0030	0.0598	0.0597	0.0594	0.942
	$\widehat{\beta}_P$	0.0271	0.0569	0.0501	0.0496	0.902	-0.1862	0.1918	0.0459	0.0471	0.029
	$\widehat{\beta}_I$	0.0023	0.0497	0.0497	0.0497	0.953	-0.0146	0.0634	0.0617	0.0588	0.916
60%	$\widehat{\beta}_O$	-0.0044	0.0550	0.0548	0.0551	0.944	0.0006	0.0678	0.0679	0.0668	0.957
	$\widehat{\beta}_P$	0.0206	0.0591	0.0554	0.0551	0.933	-0.1936	0.2005	0.0525	0.0514	0.049
	$\widehat{\beta}_I$	0.0018	0.0541	0.0541	0.0553	0.952	-0.0120	0.0672	0.0661	0.0661	0.944
80%	$\widehat{\beta}_O$	-0.0053	0.0738	0.0736	0.0768	0.966	0.0064	0.0918	0.0916	0.0969	0.960
	$\widehat{\beta}_P$	0.0088	0.0749	0.0744	0.0766	0.964	-0.1995	0.2106	0.0676	0.0696	0.192
	$\widehat{\beta}_I$	-0.0034	0.0733	0.0733	0.0769	0.964	-0.0134	0.0918	0.0908	0.0940	0.958
ρ	Type	$x_0(t)=2t$									
30%	$\widehat{\beta}_O$	-0.0013	0.0429	0.0429	0.0434	0.956	0.0017	0.0521	0.0521	0.0519	0.950
	$\widehat{\beta}_P$	0.0327	0.0540	0.0430	0.0431	0.873	-0.1785	0.1838	0.0437	0.0430	0.031
	$\widehat{\beta}_I$	0.0036	0.0429	0.0428	0.0434	0.955	-0.0148	0.0571	0.0552	0.0518	0.927
50%	$\widehat{\beta}_O$	-0.0027	0.0501	0.0501	0.0504	0.955	0.0011	0.0604	0.0604	0.0603	0.951
	$\widehat{\beta}_P$	0.0270	0.0571	0.0503	0.0502	0.912	-0.1851	0.1913	0.0484	0.0484	0.050
	$\widehat{\beta}_I$	0.0013	0.0503	0.0504	0.0505	0.952	-0.0159	0.0628	0.0608	0.0599	0.936
60%	$\widehat{\beta}_O$	-0.0061	0.0557	0.0554	0.0565	0.960	0.0063	0.0689	0.0687	0.0677	0.948
	$\widehat{\beta}_P$	0.0214	0.0588	0.0548	0.0564	0.937	-0.1848	0.1928	0.0548	0.0531	0.085
	$\widehat{\beta}_I$	-0.0015	0.0563	0.0563	0.0566	0.956	-0.0127	0.0717	0.0706	0.0668	0.926
80%	$\widehat{\beta}_O$	-0.0099	0.0753	0.0746	0.0779	0.954	0.0113	0.0995	0.0989	0.0969	0.939
	$\widehat{\beta}_P$	0.0087	0.0762	0.0757	0.0776	0.948	-0.1926	0.2062	0.0736	0.0709	0.242
	$\widehat{\beta}_I$	-0.0070	0.0748	0.0745	0.0779	0.956	-0.0099	0.0982	0.0977	0.0944	0.931
ρ	Type	$x_0(t)=3t^2$									
30%	$\widehat{\beta}_O$	-0.0038	0.0448	0.0446	0.0439	0.944	0.0032	0.0528	0.0528	0.0525	0.940
	$\widehat{\beta}_P$	0.0312	0.0547	0.0450	0.0436	0.879	-0.1762	0.1815	0.0435	0.0439	0.027
	$\widehat{\beta}_I$	0.0012	0.0449	0.0449	0.0439	0.947	-0.0141	0.0573	0.0556	0.0524	0.921
50%	$\widehat{\beta}_O$	-0.0032	0.0514	0.0513	0.0513	0.950	0.0053	0.0601	0.0599	0.0613	0.964
	$\widehat{\beta}_P$	0.0276	0.0580	0.0511	0.0511	0.916	-0.1771	0.1837	0.0488	0.0499	0.072
	$\widehat{\beta}_I$	0.0014	0.0517	0.0517	0.0513	0.952	-0.0120	0.0631	0.0620	0.0609	0.938
60%	$\widehat{\beta}_O$	-0.0048	0.0559	0.0557	0.0565	0.951	0.0046	0.0658	0.0657	0.0681	0.951
	$\widehat{\beta}_P$	0.0246	0.0615	0.0564	0.0563	0.925	-0.1836	0.1912	0.0534	0.0543	0.091
	$\widehat{\beta}_I$	0.0000	0.0560	0.0560	0.0566	0.953	-0.0138	0.0682	0.0669	0.0675	0.938
80%	$\widehat{\beta}_O$	-0.0051	0.0789	0.0788	0.0792	0.947	0.0023	0.0933	0.0933	0.0968	0.961
	$\widehat{\beta}_P$	0.0131	0.0803	0.0792	0.0792	0.945	-0.1926	0.2053	0.0711	0.0725	0.255
	$\widehat{\beta}_I$	-0.0022	0.0786	0.0786	0.0792	0.951	-0.0172	0.0956	0.0941	0.0947	0.942

注: $\widehat{\beta}_O$ 表示协变量未受干扰的参数估计. $\widehat{\beta}_P$ 表示协变量受干扰的参数估计. $\widehat{\beta}_I$ 表示受干扰协变量调整后的参数估计. BIAS 表示样本偏差. SMSE 表示标准均方误差. SD 表示估计的样本标准差. SE 表示估计标准差的均值. CP 表示 95% 置信区间的收敛概率. 算法: Newton-Raphson 算法.

表 2 $Z_1 \sim \mathrm{B}(1, 0.5)$, $Z_2 \sim \mathrm{U}(7/3 - \sqrt{19}/2, 7/3 + \sqrt{19}/2)$, $u \sim \mathrm{U}(4 - \sqrt{7}, 4 + \sqrt{7})$, $C \sim E(1)I(Z_1 = 0) + E(1/3)I(Z_1 = 1)$

n	β	ρ	Type	$x_0(t) = 1$									
				BIAS	SMSE	SD	SE	CP	BIAS	SMSE	SD	SE	CP
300	$(-0.5, 0.5)$	0.2090	$\widehat{\beta}_O$	-0.0017	0.1375	0.1375	0.1373	0.947	0.0022	0.0602	0.0602	0.0591	0.954
			$\widehat{\beta}_P$	0.0204	0.1403	0.1389	0.1376	0.947	-0.1137	0.1245	0.0508	0.0504	0.379
			$\widehat{\beta}_I$	0.0019	0.1380	0.1380	0.1374	0.943	-0.0091	0.0608	0.0602	0.0590	0.944
	$(-0.5, 0.693)$	0.1611	$\widehat{\beta}_O$	-0.0025	0.1333	0.1334	0.1329	0.951	0.0035	0.0636	0.0635	0.0624	0.949
			$\widehat{\beta}_P$	0.0313	0.1400	0.1365	0.1337	0.930	-0.1764	0.1842	0.0528	0.0519	0.105
			$\widehat{\beta}_I$	0.0045	0.1353	0.1353	0.1332	0.951	-0.0183	0.0693	0.0669	0.0621	0.922
	$(-0.25, 0.50)$	0.1940	$\widehat{\beta}_O$	0.0007	0.1335	0.1335	0.1339	0.951	0.0027	0.0599	0.0599	0.0584	0.949
			$\widehat{\beta}_P$	0.0131	0.1337	0.1331	0.1343	0.948	-0.1122	0.1237	0.0523	0.0500	0.376
			$\widehat{\beta}_I$	0.0024	0.1341	0.1341	0.1339	0.949	-0.0082	0.0620	0.0615	0.0583	0.937
500	$(-0.5, 0.5)$	0.2094	$\widehat{\beta}_O$	-0.0062	0.1055	0.1054	0.1055	0.945	0.0019	0.0452	0.0452	0.0451	0.946
			$\widehat{\beta}_P$	0.0120	0.1072	0.1066	0.1055	0.944	-0.1134	0.1197	0.0384	0.0386	0.175
			$\widehat{\beta}_I$	-0.0030	0.1056	0.1056	0.1056	0.951	-0.0073	0.0468	0.0462	0.0452	0.935
	$(-0.5, 0.693)$	0.1617	$\widehat{\beta}_O$	-0.0070	0.1023	0.1021	0.1020	0.942	0.0022	0.0479	0.0479	0.0479	0.948
			$\widehat{\beta}_P$	0.0270	0.1063	0.1028	0.1026	0.940	-0.1758	0.1807	0.0418	0.0399	0.016
			$\widehat{\beta}_I$	-0.0012	0.1027	0.1028	0.1022	0.945	-0.0141	0.0524	0.0505	0.0479	0.929
	$(-0.25, 0.50)$	0.1941	$\widehat{\beta}_O$	-0.0017	0.1022	0.1023	0.1026	0.946	0.0046	0.0462	0.0460	0.0451	0.940
			$\widehat{\beta}_P$	0.0097	0.1035	0.1031	0.1028	0.954	-0.1113	0.1179	0.0388	0.0387	0.190
			$\hat{\beta}_I$	0.0000	0.1033	0.1033	0.1028	0.949	-0.0041	0.0466	0.0464	0.0452	0.935
n	β	ρ	Type	$x_0(t) = 2t$									
300	$(-0.5, 0.5)$	0.2837	$\widehat{\beta}_O$	0.0043	0.1443	0.1443	0.1467	0.959	0.0033	0.0624	0.0624	0.0626	0.945
			$\widehat{\beta}_P$	0.0239	0.1480	0.1462	0.1468	0.952	-0.1111	0.1235	0.0540	0.0536	0.469
			$\widehat{\beta}_I$	0.0089	0.1448	0.1446	0.1470	0.952	-0.0088	0.0639	0.0633	0.0626	0.938
	$(-0.5, 0.693)$	0.2434	$\widehat{\beta}_O$	0.0051	0.1397	0.1396	0.1415	0.956	0.0056	0.0653	0.0651	0.0667	0.951
			$\widehat{\beta}_P$	0.0414	0.1491	0.1433	0.1418	0.926	-0.1749	0.1831	0.0542	0.0555	0.131
			$\widehat{\beta}_I$	0.0112	0.1410	0.1406	0.1420	0.950	-0.0177	0.0686	0.0663	0.0669	0.927
	$(-0.25, 0.50)$	0.2741	$\widehat{\beta}_O$	0.0047	0.1404	0.1404	0.1424	0.958	0.0044	0.0619	0.0618	0.0626	0.943
			$\widehat{\beta}_P$	0.0167	0.1409	0.1400	0.1423	0.955	-0.1085	0.1215	0.0546	0.0535	0.454
			$\widehat{\beta}_I$	0.0068	0.1401	0.1400	0.1426	0.961	-0.0066	0.0629	0.0626	0.0625	0.941
500	$(-0.5, 0.5)$	0.2840	$\widehat{\beta}_O$	-0.0033	0.1116	0.1116	0.1120	0.948	0.0023	0.0491	0.0491	0.0479	0.937
			$\widehat{\beta}_P$	0.0179	0.1137	0.1123	0.112	0.940	-0.1138	0.1209	0.0410	0.041	0.223
			$\widehat{\beta}_I$	-0.0001	0.1118	0.1119	0.1122	0.952	-0.0065	0.0500	0.0496	0.0479	0.920
	$(-0.5, 0.693)$	0.2435	$\widehat{\beta}_O$	-0.0023	0.1078	0.1078	0.1075	0.951	0.0032	0.0522	0.0521	0.0506	0.938
			$\widehat{\beta}_P$	0.0333	0.1123	0.1073	0.1081	0.932	-0.1777	0.1828	0.0428	0.0420	0.020
			$\widehat{\beta}_I$	0.0023	0.1086	0.1086	0.1079	0.948	-0.0141	0.0558	0.0540	0.0507	0.917
	$(-0.25, 0.50)$	0.2739	$\widehat{\beta}_O$	-0.0062	0.1111	0.1110	0.1090	0.945	0.0001	0.0487	0.0487	0.0477	0.942
			$\widehat{\beta}_P$	0.0060	0.1097	0.1096	0.1090	0.950	-0.1146	0.1223	0.0429	0.0408	0.212
			$\widehat{\beta}_I$	-0.0044	0.1116	0.1116	0.1091	0.946	-0.0090	0.0500	0.0492	0.0476	0.936

注: $\widehat{\beta}_O$ 表示协变量未受干扰的参数估计. $\widehat{\beta}_P$ 表示协变量受干扰的参数估计. $\widehat{\beta}_I$ 表示受干扰协变量调整后的参数估计. BIAS 表示样本偏差. SMSE 表示标准均方误差. SD 表示估计的样本标准差. SE 表示估计标准差的均值. CP 表示 95% 置信区间的收敛概率. 算法: Newton-Raphson 算法.

2) MM 算法

在上一小节中运用 Newton-Rapshon 算法进行模拟时, 样本量必须高于 300, 否则会出现黑塞矩阵不可逆的现象. 为了避免这个问题我们运用 MM 算法, 对替代函数 $Q(\beta\,|\,\beta^{(m)})$ 求极大似然估计, 其黑塞矩阵为对角矩阵, 即使在样本量较小时也不会出现不可逆的情况. 在文献 [3] 的模拟结果说明运用 MM 算法所得的估计结果与运用 Newton-Rapshon 算法很接近, 所以可以运用 MM 算法来替代 Newton-Rapshon 算法进行模拟运算. 仍假设 $Z_1 \sim \mathrm{B}(1, 0.5)$, $Z_2 \sim \mathrm{U}(7/3 - \sqrt{19}/2, 7/3 + \sqrt{19}/2)$; 干扰因子 $U \sim \mathrm{U}(4 - \sqrt{7}, 4 + \sqrt{7})$. 基底风险函数 $x_0(t)$ 分别取为 $1, 2t, 3t^2$, 参数 β 的真值为 $(-0.25, 0.5)$, 样本量 N 设定为 120. 结果汇总于表 3.

从表 3 中的结果可以看出, 虽然样本量减少了, 并且协变量 Z_1 为离散分布, 但是运用 MM 算法依然可以得到参数估计值, 未出现黑塞矩阵不可能的情况. 由于只对协变量 Z_2 进行了干扰, 参数 β_1 的估计值都是无偏的, 并且 SMSE, SD 和 SE 三者值比较接近. 当协变量 Z_2 受到干扰时, 参数 β_2 的估计仍然是有偏的, 例如, 死亡时间 $\widetilde{T}$ 满足指数分布, 删失率为 60% 时, 参数 β_2 的估计偏差为 -0.1122, SE 的值为 0.1108, CP 值为 0.807. 说明此估计值不满足渐近性. 经调整之后的其他结果也比较合理, 说明我们所提出的算法是切实可行的.

针对另一个参数 β 的真值 $(-0.5, 0.5)$, 在这里我们又增加了参数 bootstrap 方法, 其主要思想是用每次循环中的参数估计值作为参数的真值, 代入模型中产生一组观测样本数据为 $\{Y_1^*, Y_2^*, \cdots, Y_n^*\}$, $Y_i^* = (T_i^*, \delta_i^*, Z_i^*)$. 由 $\{Y_1^*, Y_2^*, \cdots, Y_n^*\}$ 计算得到 β^*, 由此参数估计值产生 B 组观测样本数据得到 B 个参数估计值 $\{\beta_{(1)}^*, \beta_{(2)}^*, \cdots, \beta_{(B)}^*\}$. 由于模拟的过程是对协变量 Z_2 进行干扰和调整的, 所以我们以 β_2^* 的大小排序后取 95% 置信区间的置信下限和置信上限. 循环这个过程 1000 次, 分别得到 $\widehat{\beta}_O$、$\widehat{\beta}_p$ 和 $\widehat{\beta}_I$ 三种情况下 95% 置信区间的置信下限 (LCL) 和置信上限 (UCL) 的均值. 结果汇总于表 4.

在表 4 中, 当协变量 Z_2 受到干扰时, β_2 的真值落在 95% 置信区间之外. 例如, 死亡时间 $\widetilde{T}$ 满足尺度参数为 3 的威布尔分布, 删失率为 30% 时, $\widehat{\beta}_P$ 的 95% 置信下限和置信上限分别为 0.2080 和 0.4257, 说明 $\widehat{\beta}_P$ 与真值 0.5 之间的偏差很大, 为错误的结论. 所以参数 bootstrap 方法也可以作为判别估计好坏的一种方法.

§6. 真实数据分析

该数据集是关于心力衰竭患者的, 来自 “Machine Learning Repository” 网站 (http://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records). 原始数据是由巴基斯坦费萨拉巴德大学 (政府学院) Tanvir Ahamd 等提供, 后由 Devide chiceo (加拿大多伦多科伦比尔研究所) 整理得到现在的数据集.

表 3 $N=120$, $\beta=(-0.25,0.5)$, $Z_1\sim \mathrm{B}(1,0.5)$, $Z_2\sim \mathrm{U}(7/3-\sqrt{19}/2,7/3+\sqrt{19}/2)$, $u\sim \mathrm{U}(4-\sqrt{7},4+\sqrt{7})$

ρ	Type	$x_0(t)=1$									
		BIAS	SMSE	SD	SE	CP	BIAS	SMSE	SD	SE	CP
30%	$\widehat{\beta}_O$	-0.0030	0.0845	0.0845	0.0851	0.942	0.0125	0.1009	0.1002	0.1021	0.951
	$\widehat{\beta}_P$	0.0078	0.0843	0.0840	0.0850	0.941	-0.1047	0.1340	0.0837	0.0873	0.756
	$\widehat{\beta}_I$	-0.0018	0.086	0.0861	0.0851	0.9439	-0.0035	0.103	0.1030	0.1025	0.9520
50%	$\widehat{\beta}_O$	-0.0045	0.0986	0.0985	0.1012	0.955	0.0134	0.1178	0.1171	0.1211	0.952
	$\widehat{\beta}_P$	0.0038	0.0980	0.098	0.1011	0.955	-0.1109	0.1487	0.099	0.0998	0.777
	$\widehat{\beta}_I$	-0.0025	0.0995	0.0996	0.1013	0.953	-0.0099	0.1180	0.1176	0.1197	0.941
60%	$\widehat{\beta}_O$	-0.0029	0.1099	0.1099	0.1131	0.956	0.0168	0.1329	0.1319	0.1363	0.956
	$\widehat{\beta}_P$	0.0035	0.1091	0.1091	0.1133	0.961	-0.1122	0.1556	0.1079	0.1108	0.807
	$\widehat{\beta}_I$	-0.0019	0.1100	0.1101	0.1134	0.953	-0.0049	0.1311	0.1311	0.1340	0.954
80%	$\widehat{\beta}_O$	-0.0078	0.1626	0.1625	0.1636	0.958	0.0197	0.2075	0.2067	0.1973	0.939
	$\widehat{\beta}_P$	0.0000	0.1607	0.1608	0.1642	0.955	-0.1122	0.1862	0.1487	0.1578	0.881
	$\widehat{\beta}_I$	-0.0067	0.1625	0.1625	0.1638	0.956	-0.0105	0.1940	0.1938	0.1920	0.952
ρ	Type	$x_0(t)=2t$									
30%	$\widehat{\beta}_O$	-0.0095	0.0876	0.0871	0.0875	0.948	0.0090	0.0991	0.0988	0.1043	0.959
	$\widehat{\beta}_P$	-0.0008	0.0863	0.0863	0.0876	0.948	-0.1050	0.1347	0.0843	0.0895	0.758
	$\widehat{\beta}_I$	-0.0053	0.0879	0.0878	0.0875	0.946	-0.0134	0.1017	0.1009	0.1043	0.950
50%	$\widehat{\beta}_O$	-0.0113	0.1024	0.1018	0.1050	0.953	0.0114	0.1184	0.1179	0.1232	0.960
	$\widehat{\beta}_P$	-0.0027	0.1034	0.1034	0.1050	0.952	-0.1080	0.1478	0.1009	0.1035	0.806
	$\widehat{\beta}_I$	-0.0096	0.1029	0.1025	0.1051	0.950	-0.0081	0.1215	0.1213	0.1228	0.953
60%	$\widehat{\beta}_O$	-0.0099	0.1133	0.1129	0.1180	0.952	0.0100	0.1323	0.1320	0.1391	0.962
	$\widehat{\beta}_P$	-0.0031	0.1109	0.1109	0.1180	0.958	-0.1091	0.1535	0.1080	0.1153	0.818
	$\widehat{\beta}_I$	-0.0073	0.1136	0.1134	0.1183	0.954	-0.0114	0.1323	0.1318	0.1376	0.957
80%	$\widehat{\beta}_O$	-0.0134	0.1593	0.1588	0.1686	0.958	0.0202	0.2062	0.2053	0.1998	0.945
	$\widehat{\beta}_P$	-0.0083	0.1619	0.1618	0.1739	0.960	-0.1209	0.1947	0.1527	0.1642	0.884
	$\widehat{\beta}_I$	-0.0082	0.1617	0.1616	0.1739	0.964	-0.0107	0.1877	0.1875	0.1997	0.954
ρ	Type	$x_0(t)=3t^2$									
30%	$\widehat{\beta}_O$	-0.0096	0.0885	0.0880	0.0883	0.953	0.0099	0.1021	0.1017	0.1062	0.954
	$\widehat{\beta}_P$	-0.0007	0.0887	0.0887	0.0881	0.944	-0.1022	0.1338	0.0864	0.0910	0.783
	$\widehat{\beta}_I$	-0.0070	0.0887	0.0885	0.0883	0.948	-0.0117	0.1040	0.1034	0.1061	0.945
50%	$\widehat{\beta}_O$	-0.0110	0.1037	0.1031	0.1071	0.949	0.0100	0.1207	0.1204	0.1264	0.958
	$\widehat{\beta}_P$	-0.0015	0.1006	0.1007	0.1068	0.949	-0.1081	0.1471	0.0999	0.1065	0.807
	$\widehat{\beta}_I$	-0.0060	0.1045	0.1044	0.1052	0.953	-0.0043	0.1205	0.1205	0.1270	0.952
60%	$\widehat{\beta}_O$	-0.0113	0.1177	0.1172	0.1214	0.957	0.0104	0.1355	0.1352	0.1425	0.955
	$\widehat{\beta}_P$	-0.0031	0.1175	0.1175	0.1212	0.963	-0.1087	0.1583	0.1152	0.1196	0.829
	$\widehat{\beta}_I$	-0.0086	0.1182	0.1179	0.1215	0.950	-0.0111	0.1360	0.1356	0.1414	0.952
80%	$\widehat{\beta}_O$	-0.0116	0.1650	0.1647	0.1791	0.963	0.0140	0.1941	0.1937	0.2058	0.961
	$\widehat{\beta}_P$	-0.0105	0.1640	0.1638	0.1802	0.960	-0.1194	0.1961	0.1556	0.1702	0.900
	$\widehat{\beta}_I$	-0.0091	0.1646	0.1644	0.1795	0.964	-0.0105	0.1881	0.1879	0.2046	0.962

注: $\widehat{\beta}_O$ 表示协变量未受干扰的参数估计. $\widehat{\beta}_P$ 表示协变量受干扰的参数估计. $\widehat{\beta}_I$ 表示受干扰协变量调整后的参数估计. BIAS 表示样本偏差. SMSE 表示标准均方误差. SD 表示估计的样本标准差. SE 表示估计标准差的均值. CP 表示 95% 置信区间的收敛概率. 算法: MM 算法.

表 4　$N=200$, $\beta=(-0.5,0.5)$, $Z_1\sim \mathrm{B}(1,0.5)$, $Z_2\sim \mathrm{U}(7/3-\sqrt{19}/2,7/3+\sqrt{19}/2)$, $u\sim \mathrm{U}(4-\sqrt{7},4+\sqrt{7})$

ρ	Type	$x_0(t)=1$									
		BIAS	SMSE	SD	LCL	UCL	BIAS	SMSE	SD	LCL	UCL
30%	$\widehat{\beta}_O$	-0.0047	0.0701	0.0070	-0.4898	-0.5355	0.0061	0.0732	0.0730	0.3873	0.6351
	$\widehat{\beta}_P$	0.0135	0.0710	0.0698	-0.4694	-0.5053	-0.1109	0.1276	0.0632	0.2006	0.4157
	$\widehat{\beta}_I$	-0.0005	0.0696	0.0696	-0.4850	-0.5299	-0.0082	0.0755	0.0751	0.3566	0.6076
50%	$\widehat{\beta}_O$	-0.0063	0.0815	0.0813	-0.4978	-0.5402	0.0064	0.0877	0.0875	0.3666	0.6595
	$\widehat{\beta}_P$	0.0078	0.0817	0.0814	-0.4819	-0.5117	-0.1131	0.1358	0.0753	0.1748	0.4313
	$\widehat{\beta}_I$	-0.0028	0.0819	0.0819	-0.4883	-0.5349	-0.0081	0.0908	0.0905	0.3352	0.6298
60%	$\widehat{\beta}_O$	-0.0086	0.0907	0.0904	-0.5022	-0.5451	0.0066	0.0977	0.0975	0.3492	0.6806
	$\widehat{\beta}_P$	0.0036	0.0916	0.0916	-0.4961	-0.5107	-0.1169	0.1436	0.0835	0.1529	0.4439
	$\widehat{\beta}_I$	-0.0064	0.0908	0.0906	0.4987	-0.5408	-0.0085	0.0994	0.0991	0.3172	0.6498
80%	$\widehat{\beta}_O$	-0.0146	0.1255	0.1247	-0.5167	-0.5538	0.0072	0.1423	0.1422	0.2839	0.7627
	$\widehat{\beta}_P$	-0.0076	0.1261	0.1259	-0.5117	-0.5449	-0.1258	0.1676	0.1108	0.0706	0.5007
	$\widehat{\beta}_I$	-0.0123	0.1252	0.1246	-0.5128	-0.5492	-0.0113	0.1395	0.1391	0.2471	0.7229
ρ	Type	$x_0(t)=2t$									
30%	$\widehat{\beta}_O$	-0.0049	0.0697	0.0696	-0.4868	-0.5404	0.0050	0.0762	0.0761	0.3844	0.6359
	$\widehat{\beta}_P$	0.0142	0.0681	0.0667	-0.4684	-0.5001	-0.1040	0.1235	0.0667	0.2078	0.4246
	$\widehat{\beta}_I$	-0.0015	0.0676	0.0676	-0.4816	-0.5284	-0.0046	0.0793	0.0792	0.3594	0.6147
50%	$\widehat{\beta}_O$	-0.0066	0.0804	0.0802	-0.4972	-0.5464	0.0066	0.0906	0.0904	0.3659	0.6605
	$\widehat{\beta}_P$	0.0081	0.0802	0.0798	-0.4803	-0.5131	-0.1081	0.1337	0.0787	0.1823	0.4383
	$\widehat{\beta}_I$	-0.0032	0.0808	0.0808	-0.4904	-0.5376	-0.0074	0.0937	0.0934	0.3366	0.6326
60%	$\widehat{\beta}_O$	-0.0071	0.0880	0.0877	-0.4934	-0.5467	0.0061	0.1013	0.1012	0.3500	0.6787
	$\widehat{\beta}_P$	0.0043	0.0922	0.0921	-0.4879	-0.5180	-0.1077	0.1379	0.0861	0.1656	0.4523
	$\widehat{\beta}_I$	-0.0061	0.0916	0.0914	-0.4942	-0.5379	-0.0040	0.1055	0.1055	0.3236	0.6546
80%	$\widehat{\beta}_O$	-0.0059	0.1253	0.1253	-0.5101	-0.5501	0.0085	0.1457	0.1455	0.2854	0.7610
	$\widehat{\beta}_P$	-0.0029	0.1279	0.1280	-0.5027	-0.5417	-0.1192	0.1670	0.1170	0.0791	0.5065
	$\widehat{\beta}_I$	-0.0075	0.1271	0.1270	-0.5079	-0.5480	-0.0099	0.1467	0.1464	0.2493	0.7272
ρ	Type	$x_0(t)=3t^2$									
30%	$\widehat{\beta}_O$	-0.0053	0.0709	0.0708	-0.4870	-0.5359	0.0051	0.0767	0.0766	0.3843	0.6368
	$\widehat{\beta}_P$	0.0140	0.0701	0.0687	-0.4648	-0.5017	-0.1032	0.1241	0.0691	0.2080	0.4257
	$\widehat{\beta}_I$	-0.0018	0.0677	0.0678	-0.4837	-0.5295	-0.0034	0.0818	0.0818	0.3594	0.6167
50%	$\widehat{\beta}_O$	-0.0071	0.0829	0.0826	-0.4927	-0.5451	0.0079	0.0915	0.0912	0.3694	0.6594
	$\widehat{\beta}_P$	0.0085	0.0829	0.0825	-0.4766	-0.5114	-0.1053	0.1315	0.0788	0.1870	0.4382
	$\widehat{\beta}_I$	-0.0045	0.0823	0.0822	-0.4850	-0.5307	-0.0051	0.0966	0.0966	0.3405	0.6330
60%	$\widehat{\beta}_O$	-0.0071	0.0903	0.0901	-0.4961	-0.5463	0.0085	0.1016	0.1013	0.3594	0.6728
	$\widehat{\beta}_P$	0.0064	0.0894	0.0892	-0.4818	-0.5138	-0.1083	0.1394	0.0878	0.1722	0.4463
	$\widehat{\beta}_I$	-0.0050	0.0906	0.0905	-0.4935	0.5344	-0.0034	0.1045	0.1045	0.3310	0.6471
80%	$\widehat{\beta}_O$	-0.0083	0.1322	0.1320	-0.5045	-0.5549	0.0107	0.1497	0.1494	0.3204	0.7220
	$\widehat{\beta}_P$	-0.0060	0.1306	0.1305	-0.5054	-0.5373	-0.1130	0.1646	0.1198	0.1223	0.4796
	$\widehat{\beta}_I$	-0.0113	0.1297	0.1293	-0.5075	-0.5464	-0.0039	0.1461	0.1461	0.2901	0.6935

注: $\widehat{\beta}_O$ 表示协变量未受干扰的参数估计. $\widehat{\beta}_P$ 表示协变量受干扰的参数估计. $\widehat{\beta}_I$ 表示受干扰协变量调整后的参数估计. BIAS 表示样本偏差. SMSE 表示标准均方误差. SD 表示估计的样本标准差. LCL 表示 95% 置信区间的置信下限. UCL 表示 95% 置信区间的置信上限. 算法: MM 算法.

该数据集包含了随访期间收集的 299 例心力衰竭患者的病例, 每个患者档案均具有 13 个临床特征, 包括: 年龄、是否贫血、肌酐磷酸激酶、是否患有糖尿病、射血分数、是否患有高血压、血小板、血清肌酐、血清钠、性别、是否抽烟、观测时间和是否发生死亡. 我们比较关心的是观测时间和是否发生死亡与其他临床特征的关系, 这是一组带有右删失的生存数据, 与我们文章中所研究的数据相吻合. 在剩余的 11 个临床特征中, 我们分别对其进行相关性分析, 其中相关性显著的是射血分数和血清钠、是否贫血和肌酐磷酸激酶, 结合医学方面的知识, 可知射血分数会对血清钠造成干扰, 是否贫血会对肌酐磷酸激酶造成干扰. 我们主要的研究目的是在对血清钠和肌酐磷酸激酶分别运用射血分数和是否贫血进行平滑之后, 评估这些临床特征与心力衰竭之间的关联.

本数据中有 96 例死亡, 删失率接近 67.9%. 由于数据中协变量的量纲不一致, 我们对每一列协变量进行标准化后, 令年龄为 Z_1、肌酐磷酸激酶为 Z_2、是否患有糖尿病为 Z_3、是否患有高血压为 Z_4、血小板为 Z_5、血清肌酐为 Z_6、血清钠为 Z_7、性别为 Z_8、是否抽烟为 Z_9, 射血分数为干扰因子 U_1、是否贫血为干扰因子 U_2.

我们先将协变量 $Z_1, Z_2, \cdots, Z_9$ 代入比例风险模型中,

$$x(t \mid Z) = x_0(t) \exp\Big\{ \sum_{i=1}^{9} \beta_i Z_i \Big\}, \tag{14}$$

得到带有干扰因素的参数估计值为 (2.4106, 1.3012, 0.1553, 0.4941, −0.2904, 2.4139, −2.0584, −0.0165, 0.1192)$^{\intercal}$. 再将射血分数 U_1 作为协变量 $\widetilde{Z}_7$ 的干扰因子, 是否贫血 U_2 作为协变量 $\widetilde{Z}_2$ 的干扰因子, 运用本文中所提出的方法对 $\widetilde{Z}_2$ 和 $\widetilde{Z}_7$ 进行核平滑得到估计量 $\widehat{Z}_2$ 和 $\widehat{Z}_7$. 将调整后的协变量代入比例风险模型 (14) 式中得到参数估计值为 (2.5675, 5.0545, 0.1448, 0.2514, −0.5456, 2.1519, 11.4276, −0.0111, 0.0642)$^{\intercal}$, 这里的核函数和带宽的选择方法与第五章模拟研究中所用的方法是一致的.

协变量经过调整之后, 肌酐磷酸激酶的参数估计值从 1.3012 调整到 5.0545, 血清钠的参数估计值从 −2.0584 调整为 11.4276, 其他协变量的参数估计值经过调整之后变化较小, 符合医学常识, 从而说明我们的方法对真实数据的分析是有效的.

§7. 结　　论

本文针对生存数据受干扰的问题, 在比例风险模型中提出了协变量调整方法, 运用核函数构造干扰因子分布函数对协变量进行调整, 并给出参数估计量的相合性和渐近正态性的证明. 因为协变量受到干扰, 会出现估计量计算困难问题, 我们运用了 MM 算法通过构造替代函数解决了黑塞矩阵不可逆问题. 模拟研究表明, 在有限样本下协变量调整方法具有很好的性能.

本文讨论的模型为比例风险模型, 所提出的方法可以应用到其他生存模型中, 例如加速失效时间模型[18–20], 可乘可加模型[21], 加速风险模型[22] 和一类半参数变换模型[23]. 我们所研究的数据为带有右删失的生存数据, 为了节约成本并提高信息利用率, 可采用 Case-Cohort, 广义 Case-Cohort 依赖响应变量等抽样方法对参数进行统计推断. 本文运用核函数对受干扰的协变量进行平滑, 也可采用 B 样条插值法和多项式差值法对协变量进行调整.

本文只是对生存数据中协变量受到干扰的问题进行讨论分析, 还可以考虑观测时间有误差的问题. 由于实验过程中的长期性和复杂性, 在对病人进行观测时, 时间的记录会出现测量误差. 我们认为这里的误差是满足正态分布的, 并且与真实的观测时间是相加的关系, 而非本文所考虑的协变量与干扰因子之间的相乘关系, 可运用半参数 copula 方法[24] 进行调整. 同时考虑带测量误差的响应变量和带干扰的协变量的生存问题也是我们以后的研究方向之一.

附录: 渐近性质的证明

引理 3　由假设条件 (A_1–A_4), 则下面相合性成立

$$\widehat{Z}_i - Z_i = O_p(h + [nh/\ln(1/h)]^{-1/2}).$$

证明:　由文献 [25], 我们可以得到非参数估计 $\widehat{g}_{\widetilde{Z}}(u)$ 的一致收敛性.

$$\sup_{u\in[a,b]} |\widehat{g}_{\widetilde{Z}}(u) - g_{\widetilde{Z}}(u)| = O_p(h + [nh/\ln(1/h)]^{-1/2}). \tag{15}$$

首先证明 $\widehat{Z}_{ij} - Z_{ij}$, 由 (2)、(5) 和 (6) 式, 可得

$$\begin{aligned}
\widehat{Z}_{ij} - Z_{ij} &= \frac{\widetilde{Z}_{ij}}{\widehat{\phi}_j(U_i)} - Z_{ij} = \frac{Z_{ij}\phi_j(U_i)}{\widehat{\phi}_j(U_i)} - Z_{ij} = Z_{ij}\Big[\frac{\phi_j(U_i)}{\widehat{\phi}_j(U_i)} - 1\Big] \\
&= Z_{ij}\Big[\frac{g_{\widetilde{Z}_j}(U_i)}{\mathsf{E}Z_j}\frac{\overline{\overline{Z}}_j}{\widehat{g}_{\widetilde{Z}_j}(U_i)} - 1\Big] \\
&= Z_{ij}\Big[\frac{g_{\widetilde{Z}_j}(U_i)}{\widehat{g}_{\widetilde{Z}_j}(U_i)} - 1 + \frac{g_{\widetilde{Z}_j}(U_i)}{\widehat{g}_{\widetilde{Z}_j}(U_i)}\frac{\overline{\overline{Z}}_j - \mathsf{E}Z_j}{\mathsf{E}Z_j}\Big],
\end{aligned}$$

对于上式 $g_{\widetilde{Z}_j}(U_i)/\widehat{g}_{\widetilde{Z}_j}(U_i)$, 由 (15) 式可得

$$\frac{g_{\widetilde{Z}_j}(U_i)}{\widehat{g}_{\widetilde{Z}_j}(U_i)} = 1 - \frac{\widehat{g}_{\widetilde{Z}_j}(U_i) - g_{\widetilde{Z}_j}(U_i)}{\widehat{g}_{\widetilde{Z}_j}(U_i)} = O_p(h + [nh/\ln(1/h)]^{-1/2}),$$

再对 $(\overline{\widetilde{Z}}_j - \mathsf{E}Z_j)/\mathsf{E}Z_j$ 运用大数定律, 可得

$$\begin{aligned}\widehat{Z}_{ij} - Z_{ij} &= O_p\Big(Z_{ij}\cdot\frac{\widehat{g}_{\widetilde{Z}}(U_i) - g_{\widetilde{Z}_j}(U_i)}{\widehat{g}_{\widetilde{Z}_j}(U_i)}\Big) + O_p\Big(Z_{ij}\cdot\frac{\widehat{g}_{\widetilde{Z}_j}(U_i) - g_{\widetilde{Z}_j}(U_i)}{\widehat{g}_{\widetilde{Z}_j}(U_i)}\cdot n^{-1/2}\Big)\\ &= O_p(h + [nh/\ln(1/h)]^{-1/2}).\end{aligned}$$

引理得证. □

引理 4 定义

$$\begin{aligned}A(\beta,t) &= \sum_{i=1}^n\int_0^t \beta^{\mathsf{T}}Z_i(S)\mathrm{d}N_i(S) - \sum_{i=1}^n\int_0^t \ln\Big[\sum_{l=1}^n Y_l(S)\mathrm{e}^{\beta^{\mathsf{T}}Z_l(S)}\Big]\mathrm{d}N_i(S),\\ \widehat{A}(\beta,t) &= \sum_{i=1}^n\int_0^t \beta^{\mathsf{T}}\widehat{Z}_i(S)\mathrm{d}N_i(S) - \sum_{i=1}^n\int_0^t \ln\Big[\sum_{l=1}^n Y_l(S)\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(S)}\Big]\mathrm{d}N_i(S),\end{aligned}$$

则由 $(\mathrm{A}_1-\mathrm{A}_4)$ 可得 $\widehat{A}(\beta,t)\xrightarrow{\mathsf{P}}A(\beta,t)$.

证明: 因为

$$\begin{aligned}\widehat{A}(\beta,t) - A(\beta,t) &= \sum_{i=1}^n\int_0^t \beta^{\mathsf{T}}[\widehat{Z}_i(S) - Z_i(S)]\mathrm{d}N_i(S) - \sum_{i=1}^n\int_0^t \ln\Bigg[\frac{\sum\limits_{l=1}^n Y_l(S)\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(S)}}{\sum\limits_{l=1}^n Y_l(S)\mathrm{e}^{\beta^{\mathsf{T}}Z_l(S)}}\Bigg]\mathrm{d}N_i(S)\\ &= \sum_{i=1}^n\int_0^t \Bigg\{\beta^{\mathsf{T}}[\widehat{Z}_i(S) - Z_i(S)] - \ln\Bigg[\frac{\sum\limits_{l=1}^n Y_l(S)\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(S)}}{\sum\limits_{l=1}^n Y_l(S)\mathrm{e}^{\beta^{\mathsf{T}}Z_l(S)}}\Bigg]\Bigg\}\mathrm{d}N_i(S),\end{aligned}$$

由引理 3 可知

$$\sum_{i=1}^n \beta^{\mathsf{T}}[\widehat{Z}_i(S) - Z_i(S)]\xrightarrow{\mathsf{P}}0.$$

下证

$$\sum_{i=1}^n \ln\Bigg[\frac{\sum\limits_{l=1}^n Y_l(S)\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(S)}}{\sum\limits_{l=1}^n Y_l(S)\mathrm{e}^{\beta^{\mathsf{T}}Z_l(S)}}\Bigg]\longrightarrow 0,$$

只需证

$$\sum_{l=1}^n Y_l\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(S)} - \sum_{i=1}^n Y_l\mathrm{e}^{\beta^{\mathsf{T}}Z_l(S)}\longrightarrow 0,$$

即 $\sum\limits_{l=1}^n Y_l(S)(\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(S)} - \mathrm{e}^{\beta^{\mathsf{T}}Z_l(S)})\longrightarrow 0$.

因为 e^x 为可微函数, 且一阶导函数在 Z_l 处连续, 则由引理 3 可得

$$\sum_{l=1}^n Y_l(\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(S)} - \mathrm{e}^{\beta^{\mathsf{T}}Z_l(S)})\xrightarrow{\mathsf{P}}0,$$

则 $\widehat{A}(\beta,t)$ 渐进收敛于 $A(\beta,t)$. □

引理 5　定义

$$\widehat{S}^{(k)}(\beta,t)=n^{-1}\sum_{l=1}^{n}Y_l(t)\widehat{Z}_l(t)^{\otimes k}\exp\{\beta^{\mathsf{T}}\widehat{Z}_l(t)\},\qquad k=0,1,2,$$

由假设条件 $(\mathrm{A}_1-\mathrm{A}_4)$, 可得

$$\widehat{S}^{(k)}(\beta,t)\xrightarrow{\mathsf{P}}S^{(k)}(\beta,t).$$

证明:　当 $j=0$ 时,

$$\widehat{S}^{(0)}(\beta,t)=\frac{1}{n}\sum_{l=1}^{n}Y_l(t)\exp\{\beta^{\mathsf{T}}\widehat{Z}_l(t)\},$$

由引理 4 可知

$$\widehat{S}^{(0)}(\beta,t)\xrightarrow{\mathsf{P}}S^{(0)}(\beta,t).$$

当 $k=1,2$ 时,

$$\begin{aligned}\widehat{S}^{(k)}(\beta,t)-S^{(k)}(\beta,t)&=\frac{1}{n}\sum_{l=1}^{n}\left[\widehat{Z}_l(t)^{\otimes k}Y_l(t)\exp\{\beta^{\mathsf{T}}\widehat{Z}_l(t)\}-\widehat{Z}_l(t)^{\otimes k}Y_l(t)\exp\{\beta^{\mathsf{T}}Z_l(t)\}\right]\\&=\frac{1}{n}\sum_{l=1}^{n}\left[\widehat{Z}_l(t)^{\otimes k}Y_l(t)\exp\{\beta^{\mathsf{T}}\widehat{Z}_l(t)\}-\widehat{Z}_l(t)^{\otimes k}Y_l(t)\exp\{\beta^{\mathsf{T}}Z_l(t)\}\right.\\&\quad\left.+\widehat{Z}_l(t)^{\otimes k}Y_l(t)\exp\{\beta^{\mathsf{T}}Z_l(t)\}-\widehat{Z}_l(t)^{\otimes k}Y_l(t)\exp\{\beta^{\mathsf{T}}Z_l(t)\}\right]\\&=\frac{1}{n}\sum_{l=1}^{n}\left\{\widehat{Z}_l(t)^{\otimes k}Y_l(t)[\exp\{\beta^{\mathsf{T}}\widehat{Z}_l(t)\}-\exp\{\beta^{\mathsf{T}}Z_l(t)\}]\right.\\&\quad\left.+[\widehat{Z}_l(t)^{\otimes k}-Z_l(t)^{\otimes k}]Y_l(t)\exp\{\beta^{\mathsf{T}}Z_l(t)\}\right\},\end{aligned}$$

由引理 3 和引理 4 可知

$$\widehat{Z}_l(t)^{\otimes k}-Z_l(t)^{\otimes k}=O_p(h+[nh/\ln(1/h)]^{-1/2}),\tag{16}$$

其中

$$\begin{aligned}\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(t)}-\mathrm{e}^{\beta^{\mathsf{T}}Z_l(t)}&\xlongequal{\text{泰勒展开}}[\widehat{Z}_l(t)-Z_l(t)]\cdot\mathrm{e}^{\beta^{\mathsf{T}}Z_l(t)}\cdot\beta+O_p(h+[nh/\ln(1/h)]^{-1})\\&=O_p(h+[nh/\ln(1/h)]^{-1/2}).\end{aligned}\tag{17}$$

则由 (16) 式和 (17) 式可知

$$\begin{aligned}&\frac{1}{n}\sum_{l=1}^{n}[\widehat{Z}_l(t)^{\otimes k}-Z_l(t)^{\otimes k}]Y_l(t)\mathrm{e}^{\beta^{\mathsf{T}}Z_l(t)}=O_p(h+[nh/\ln(1/h)]^{-1/2}),\\&\frac{1}{n}\sum_{l=1}^{n}\widehat{Z}_l(t)^{\otimes k}Y_l(t)(\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(t)}-\mathrm{e}^{\beta^{\mathsf{T}}Z_l(t)})=O_p(h+[nh/\ln(1/h)]^{-1/2}),\end{aligned}$$

则 $\widehat{S}^{(k)}(\beta,t)\xrightarrow{\mathsf{P}}S^{(k)}(\beta,t)$.　　□

定理 1 的证明: 参考文献 [26] 中的引理 1 可知, $\widetilde{D}(\beta,t)=n^{-1}\{A(\beta,t)-A(\beta_0,t)\}$ 与 $D(\beta,t)$ 依概率收敛于同一极限, 从而 $D(\beta,t)$ 一致收敛于

$$d(\beta)=\int_0^\tau\Big\{(\beta-\beta_0)^{\mathsf{T}}S^{(1)}(\beta_0,t)-\ln\Big[\frac{S^{(0)}(\beta,t)}{S^{(0)}(\beta_0,t)}\Big]S^{(0)}(\beta_0,t)\Big\}x_0(t)\mathrm{d}t,$$

其中 $d(\beta)$ 为关于 β 的连续的凸函数, 并且当且仅当 $\beta=\beta_0$ 时, $d(\beta)$ 取得最大值 [17], 即

$$d(\beta)\leqslant d(\beta_0). \tag{18}$$

运用反证法, 假设找到 $\widehat{\beta}_I$ 不依概率收敛于 β_0, 则存在一子列 $\widetilde{\beta}_{jn}$ 收敛于 β^* 但不等于 β_0, 因为 $\widetilde{\beta}_{jn}$ 为最值点, 则 $d(\widetilde{\beta}_{jn},\tau)\geqslant d(\widetilde{\beta}_0,\tau)$. 由极限的一致性和连续性, 当 $\beta^*\neq\beta_0$, 我们可得

$$d(\beta^*)\geqslant d(\beta_0).$$

这与 (18) 式相矛盾, 所以 $\widehat{\beta}_I$ 依概率收敛于 β_0. □

定理 2 的证明: 为了证明协方差矩阵, 首先引入以下定理. 由极大似然函数得到的得分方程为

$$F(\beta)=\sum_{i=1}^n\Delta_i\Big[Z_i-\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\Big]=0,$$

其极大似然函数估计 $\widehat{\beta}$ 可由求解得分方程 $u(\beta)=0$ 得到, 提出如下估计方程:

$$\widetilde{F}(\beta)=\sum_{i=1}^n\Delta_i\Big[\widetilde{Z}_i-\frac{\widetilde{S}^{(1)}(\beta,t)}{\widetilde{S}^{(0)}(\beta,t)}\Big]=0.$$

由 (4) 式可得,

$$\mathsf{E}\Big\{\Delta_i\Big[Z_i-\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\Big]\Big\}=0,\qquad \mathsf{E}\Big\{\Delta_i\Big[\widehat{Z}_i-\frac{\widehat{s}^{(1)}(\beta,t)}{\widehat{s}^{(0)}(\beta,t)}\Big]\Big\}=0,$$

$$\widehat{Z}_i=(\widehat{Z}_{i1},\widehat{Z}_{i2},\cdots,\widehat{Z}_{ip})^{\mathsf{T}}=\Big(\frac{\widetilde{Z}_{i1}}{\widehat{\phi}_1(U_i)},\frac{\widetilde{Z}_{i2}}{\widehat{\phi}_2(U_i)},\cdots,\frac{\widetilde{Z}_{ip}}{\widehat{\phi}_p(U_i)}\Big)^{\mathsf{T}}$$

$$=\begin{pmatrix}\widehat{\phi}_1^{-1}(U_i)&0&0&0&0&0\\0&\widehat{\phi}_2^{-1}(U_i)&0&0&0&0\\0&0&\cdot&0&0&0\\0&0&0&\cdot&0&0\\0&0&0&0&\cdot&0\\0&0&0&0&0&\widehat{\phi}_p^{-1}(U_i)\end{pmatrix}\begin{pmatrix}\widetilde{Z}_{i1}\\\widetilde{Z}_{i2}\\\cdot\\\cdot\\\cdot\\\widetilde{Z}_{ip}\end{pmatrix},$$

$$\mathsf{E}(\beta^{\mathsf{T}}\widehat{Z})=\mathsf{E}(\beta^{\mathsf{T}}\widehat{\phi}^{-1}\widetilde{Z})=\beta^{\mathsf{T}}\mathsf{E}(\widetilde{Z})=\beta^{\mathsf{T}}Z,$$

$$\widehat{S}^{(2)}(\beta,t)=\frac{1}{n}\sum_{l=1}^{n}\widehat{Z}_l(t)\widehat{Z}_l^{\mathsf{T}}(t)Y_l(t)\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(t)}=\frac{1}{n}\sum_{l=1}^{n}\phi_l^{-1}\widetilde{Z}_l(t)\widetilde{Z}_l^{\mathsf{T}}(t)\phi_l^{-1}Y_l(t)\mathrm{e}^{\beta^{\mathsf{T}}\phi_l^{-1}\widetilde{Z}_l(t)},$$

$$\widehat{S}^{(1)}(\beta,t)=\frac{1}{n}\sum_{l=1}^{n}\widehat{Z}_l(t)Y_l(t)\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(t)}=\frac{1}{n}\sum_{l=1}^{n}\phi_l^{-1}\widetilde{Z}_l(t)Y_l(t)\mathrm{e}^{\beta^{\mathsf{T}}\phi_l^{-1}\widetilde{Z}_l(t)},$$

$$\widehat{S}^{(0)}(\beta,t)=\frac{1}{n}\sum_{l=1}^{n}Y_l(t)\mathrm{e}^{\beta^{\mathsf{T}}\widehat{Z}_l(t)}=\frac{1}{n}\sum_{l=1}^{n}Y_l(t)\mathrm{e}^{\beta^{\mathsf{T}}\phi_l^{-1}\widetilde{Z}_l(t)},$$

$$S^{(2)}(\beta,t)=\frac{1}{n}\sum_{l=1}^{n}Z_l(t)Z_l^{\mathsf{T}}(t)Y_l(t)\mathrm{e}^{\beta^{T}Z_l(t)},$$

$$S^{(1)}(\beta,t)=\frac{1}{n}\sum_{l=1}^{n}Z_l(t)Y_l(t)\mathrm{e}^{\beta^{\mathsf{T}}Z_l(t)},$$

$$S^{(0)}(\beta,t)=\frac{1}{n}\sum_{l=1}^{n}Y_l(t)\mathrm{e}^{\beta^{\mathsf{T}}Z_l(t)},$$

$$\frac{\widehat{S}^{(2)}(\beta,t)}{\widehat{S}^{(0)}(\beta,t)}=\Phi^{-1}\frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)}\Phi^{-1},\qquad\frac{\widehat{S}^{(1)}(\beta,t)}{\widehat{S}^{(0)}(\beta,t)}=\Phi^{-1}\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)},$$

$$V=\frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)}-\mathsf{E}(\beta,t)^{\otimes 2}=\frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)}-\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\Big[\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\Big]^{\mathsf{T}},$$

$$\begin{aligned}\widehat{V}&=\frac{\widehat{S}^{(2)}(\beta,t)}{\widehat{S}^{(0)}(\beta,t)}-\frac{\widehat{S}^{(1)}(\beta,t)}{\widehat{S}^{(0)}(\beta,t)}\Big[\frac{\widehat{S}^{(1)}(\beta,t)}{\widehat{S}^{(0)}(\beta,t)}\Big]^{\mathsf{T}}\\&=\Phi^{-1}\frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)}\Phi^{-1}-\Phi^{-1}\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\Big[\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\Big]^{\mathsf{T}}\Phi^{-1}\\&=\Phi^{-1}\Big\{\frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)}-\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\Big[\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\Big]^{\mathsf{T}}\Big\}\Phi^{-1}\\&=\Phi^{-1}V\Phi^{-1},\end{aligned}$$

其中 $\Phi^{-1}=\sum_{l=1}^{n}\phi_l^{-1}$.

因为 $\widehat{\beta}_I$ 为 $\widehat{l}(\beta)$ 的驻点, 所以由 (8) 式可知

$$\nabla_\beta\widehat{l}(\widehat{\beta}_I)=0,\tag{19}$$

其中 $\nabla_\beta l(\beta)$ 表示对 $l(\beta)$ 中的 β 求一阶导, 运用一阶泰勒公式将 (19) 式左侧在 β_0 处展开, 可得

$$0=\nabla_\beta\widehat{l}(\beta_0)+\nabla_\beta^2\widehat{l}(\xi)(\widehat{\beta}_I-\beta_0),$$

其中 ξ 落在 $\widehat{\beta}_I$ 和 β_0 中间, 由 $\widehat{\beta}_I$ 收敛于 β_0 以及连续映射定理, 我们有

$$\nabla_\beta\widehat{l}(\beta_0)+\nabla_\beta^2\widehat{l}(\beta_0)(\widehat{\beta}_I-\beta_0)=o_p(1),$$

因此可得

$$\sqrt{n}(\widehat{\beta}_I-\beta_0)=\Big[-\frac{1}{n}\nabla_\beta^2\widehat{l}(\beta_0)\Big]^{-1}\Big[\frac{1}{\sqrt{n}}\nabla_\beta\widehat{l}(\beta_0)\Big]+o_p(1).$$

由假设条件 $(\mathrm{A}_1-\mathrm{A}_4)$ 和引理 3–引理 4, 我们有

$$-\frac{1}{n}\nabla_\beta^2\widehat{l}(\beta_0)\xrightarrow{\mathrm{P}}\Sigma(\beta_0),$$

$$\frac{1}{\sqrt{n}}\nabla_\beta\widehat{l}(\beta_0)\xrightarrow{\mathrm{D}}\mathrm{N}_p(0,\Sigma(\beta_0)),$$

则由 Slutsky 定理, 可得

$$\sqrt{n}(\widehat{\beta}_n-\beta_0)\xrightarrow{\mathrm{D}}\mathrm{N}_p(0,[\Sigma(\beta_0)]^{-1}).$$

由引理 4 可知

$$\widehat{S}^{(0)}(\beta,t)\xrightarrow{\mathrm{P}}S^{(0)}(\beta,t).$$

$$\Sigma(\beta)=\int_0^\tau\nu(\beta,t)s^{(0)}(\beta,t)x_0(t)\,\mathrm{d}t,$$

$$\widehat{\Sigma}(\beta)=\int_0^\tau\widehat{\nu}(\beta,t)\widehat{s}^{(0)}(\beta,t)x_0(t)\,\mathrm{d}t,$$

所以 $\widehat{\Sigma}(\beta)=\Phi^{-1}\Sigma(\beta)\Phi^{-1}$. □

参 考 文 献

[1] COX D R. Regression models and life-tables (with discussion) [J]. *J Roy Statist Soc Ser B*, 1972, **34(2)**: 187–220.

[2] CAO Y X, YU J C, LIU Y Y. Optimal generalized case-cohort analysis with Cox's proportional hazards model [J]. *Acta Math Appl Sin Engl Ser*, 2015, **31(3)**: 841–854.

[3] DING J L, TIAN G L, YUEN K C. A new MM algorithm for constrained estimation in the proportional hazards model [J]. *Comput Statist Data Anal*, 2015, **84**: 135–151.

[4] JI Y Y, WANG L M, ZHANG H H, et al. Semiparametric estimation of a Box-Cox transformation model with varying coefficients model [J]. *Sci China Math*, 2017, **60(5)**: 897–922.

[5] LIU W R, FANG J L, LU X W. Additive-multiplicative hazards model with current status data [J]. *Comput Statist*, 2018, **33(3)**: 1245–1266.

[6] HAMAD F, KACHOUIE N N. A hybrid method to estimate the full parametric hazard model [J]. *Comm Statist Theory Methods*, 2019, **48(22)**: 5477–5491.

[7] ROSMAN J B, TER WEE P M, MEIJER S, et al. Prospective randomised trial of early dietary protein restriction in chronic renal failure [J]. *Lancet*, 1984, **324(8415)**: 1291–1296.

[8] KLAHR S, LEVEY A S, BECK G J, et al. The effects of dietary protein restriction and blood-pressure control on the progression of chronic renal disease (modification of diet in renal disease study group) [J]. *N Engl J Med*, 1994, **330(13)**: 877–884.

[9] ŞENTÜRK D, MÜLLER H G. Covariate-adjusted regression [J]. *Biometrika*, 2005, **92(1)**: 75–89.

[10] CUI X, GUO W S, LIN L, et al. Covariate-adjusted nonlinear regression [J]. *Ann Statist*, 2009, **37(4)**: 1839–1870.

[11] LI F, LIN L, CUI X. Covariate-adjusted partially linear regression models [J]. *Comm Statist Theory Methods*, 2010, **39(6)**: 1054–1074.

[12] MA Y Y, LUAN Y H. Covariate-adjusted regression for time series [J]. *Comm Statist Theory Methods*, 2012, **41(3)**: 422–436.

[13] LI F, LU Y Q. Lasso-type estimation for covariate-adjusted linear model [J]. *J Appl Stat*, 2018, **45(1)**: 26–42.

[14] LU Y Q, LI F, FENG S Y. Local linear estimation for covariate-adjusted varying-coefficient models [J]. *Comm Statist Theory Methods*, 2019, **48(15)**: 3816–3835.

[15] BECKER M P, YANG I, LANGE K. EM algorithms without missing data [J]. *Stat Methods Med Res*, 1997, **6(1)**: 38–54.

[16] COX D R. Partial likelihood [J]. *Biometrika*, 1975, **62(2)**: 269–276.

[17] ANDERSEN P K, GILL R D. Cox's regression model for counting processes: a large sample study [J]. *Ann Statist*, 1982, **10(4)**: 1100–1120.

[18] CHIOU S H, KANG S, YAN J. Fast accelerated failure time modeling for case-cohort data [J]. *Stat Comput*, 2014, **24(4)**: 559–568.

[19] WANG S Y, HU T, XIANG L M, et al. Generalized M-estimation for the accelerated failure time model [J]. *Statistics*, 2016, **50(1)**: 114–138.

[20] ZHENG M, LIN R X, YU W. Competing risks data analysis under the accelerated failure time model with missing cause of failure [J]. *Ann Inst Statist Math*, 2016, **68(4)**: 855–876.

[21] LIN D Y, YING Z L. Semiparametric analysis of general additive-multiplicative hazard models for counting processes [J]. *Ann Statist*, 1995, **23(5)**: 1712–1734.

[22] CHEN Y Q, WANG M C. Analysis of accelerated hazards models [J]. *J Amer Statist Assoc*, 2000, **95(450)**: 608–618.

[23] ZENG D L, LIN D Y. Efficient estimation of semiparametric transformation models for counting processes [J]. *Biometrika*, 2006, **93(3)**: 627–640.

[24] KIM S, LI Y, SPIEGELMAN D. A semiparametric copula method for Cox models with covariate measurement error [J]. *Lifetime Data Anal*, 2016, **22(1)**: 1–16.

[25] FAN J Q, YAO Q W. *Nonlinear Time Series: Nonparametric and Parametric Methods* [M]. New York: Springer, 2003.

[26] SELF S G, PRENTICE R L. Asymptotic distribution theory and efficiency results for case-cohort studies [J]. *Ann Statist*, 1988, **16(1)**: 64–81.

Research on Covariate Adjustment Method Based on Proportional Hazards Model

RONG Guocai[1] WANG Yanan[2,3] WEI Chengdong[3] DENG Lifeng[1]

([1] *College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, 266590, China*)

([2] *Queshan No.1 Senior Middle School, Zhumadian, 463200, China*)

([3] *School of Mathematics and Statistics, Nanning Normal University, Nanning, 530029, China*)

Abstract: In actual data, especially medical data, the covariates are contaminated or interfered by certain factors, while the real covariates cannot be observed. This paper discusses how to adjust the disturbed covariates in the proportional risk model. Covariate existed in the adjustment methods cannot be directly used for survival data, in order to solve this problem, we use kernel functions to construct the interference factors of the distribution function, the interference of covariate smoothly get the estimate of the real covariate, again to get the parameters in the model of regression estimate, and completed the estimate satisfying consistency and asymptotic normality. We also proposed the use of Minorization-Maximization (MM) algorithm to obtain parameter estimates. The first M is to construct a surrogate function by the convexity of the exponential function and the negative logarithm function, which the Hessian matrix is a diagonal matrix; The second M is to obtain the estimators by maximizing the surrogate function. Finally, we demonstrate the feasibility of our proposed method through numerical simulation and real data research.

Keywords: covariate-adjusted; the proportional hazards model; asymptotic property; minorization-maximization algorithm; Bootstrap

2020 Mathematics Subject Classification: 62N01; 62N02; 62F12