第62卷第1期 数学学报中文版 Vol.62, No.1
2019年1月 ACTA MATHEMATICA SINICA, CHINESE SERIES Jan., 2019

文章编号: 0583-1431(2019)01-0087-16 **文献标识码:** **A**

带相依终止事件的复发事件数据的可加可乘比率模型

孙 琴 曲连强

华中师范大学数学与统计学学院 武汉 430079
E-mail: 2443575357@qq.com; qulianq@mail.ccnu.edu.cn

摘 要 本文对带相依终止事件的复发事件数据提出了一个联合建模分析方法, 用一个带脆弱变量的可加可乘比率模型来刻画复发事件过程, 还用带脆弱变量的 Cox 风险率模型来刻画终止事件过程, 而且这两个过程的相依性由脆弱变量来刻画. 我们利用估计方程的方法, 对模型参数进行了估计, 给出了所得估计的渐近性质. 同时, 通过数值模拟分析验证了估计的渐近性质. 最后, 利用该方法分析了弗吉尼亚大学慢性心脏病病人医疗诊费数据.

关键词 可加可乘模型; 估计方程; 脆弱变量; 复发事件; 终止事件
MR(2010) 主题分类 62N01, 62N02
中图分类 O212.1

An Additive-Multiplicative Rates Model for Recurrent Event Data with a Terminal Event

Qin SUN Lian Qiang QU

School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, *P. R. China*
E-mail: 2443575357@*qq.com*; *qulianq@mail.ccnu.edu.cn*

Abstract We propose a joint modeling approach for the analysis of recurrent event data with a terminal event, where an additive-multiplicative rates model is specified for the recurrent event process, the Cox hazards frailty model is specified for the terminal event, and the shared frailty is used to account for the association between the two processes. An estimating equation approach is developed for estimating the model parameters. The asymptotic properties of the proposed estimators are established. Simulation studies are constructed to examine performances of the proposed estimators under finite samples. Finally, we use the proposed method to analyze a medical cost study of chronic heart failure patients.

收稿日期: 2017-02-20; 接受日期: 2018-03-08
基金项目: 中央高校基本科研业务费青年教师项目 (20205170465)
通讯作者: 曲连强

Keywords additive-multiplicative rates model; estimating equation; frailty; recurrent event; terminal event

MR(2010) Subject Classification 62N01, 62N02

Chinese Library Classification O212.1

1 引言

复发事件经常出现在生物学、公共健康学、工业及经济学等领域中. 在这些领域的一些研究中, 每个个体在一段时间内会多次经历某一特殊事件. 例如, 癌细胞复发, 多次医院就诊, 反复吸食毒品, 间断性经济衰退等 [5]. 研究者一般感兴趣的是如何分析协变量对事件复发时间的某些特点的效应. 基于此, 强度或者比率函数模型已被许多研究者采用, 并用以来分析复发事件数据 (见文 [2, 6, 12, 14, 30, 34, 35]). 对现存方法比较全面的回顾可参考 Cook 和 Lawless [5] 及 Kalbfleisch 和 Prentice [10] 等人的文献.

在许多研究中, 可能存在某一终止事件 (如死亡) 阻止试验的进一步观测. 例如, 在病人去医院就诊期间, 如果病人死亡, 那么病人的就诊过程也就被终止了. 而且, 终止事件与所感兴趣的复发事件存在很强的相关关系. 现在已有一些研究分析带终止事件的复发事件数据, 可以将之分为三大类: 强度模型, 边际模型和偏边际模型 [19, 22]. 强度模型通过利用脆弱变量来刻画复发事件和终止事件之间的相依性, 而且他们之间的相依性被可观测到的协变量和不可观测的脆弱变量所完全确定 [9, 20, 32, 36]. 强度模型方法一般假定在给定脆弱变量的条件下, 复发事件过程是一泊松过程, 而且估计方法对泊松假定十分敏感. 边际模型考虑的是复发事件与终止事件的边际关系, 两者之间的相关性没有明确假定 [6, 7, 27, 37]. 但在此情况下, 复发事件的比率是存活个体与死亡个体的平均, 使得模型参数的解释性很差 [11]. 偏边际模型考虑的是存活个体的事件复发比率, 而且可以引入一个脆弱变量来确定复发事件与终止事件的相关性. 有关此方法更加详细的讨论可参考文 [4, 19, 20, 22, 29, 33, 34]. 例如, Cook 和 Lawless [4] 提出了存活个体在某些时间点处的复发事件均值和比率模型. Kalbfleisch 等人 [11], Qu, Sun 和 Liu [25] 及 Ye, Kalbfleisch 和 Schaubel [33] 通过引入脆弱变量, 提出了一个联合半参数模型, 其中脆弱变量用以解释复发事件的条件比率和终止事件风险率的相依性. 偏边际模型在刻画协变量对存活个体复发事件的条件比率效应方面具有很好的灵活性, 而且模型参数可以解释为在给定个体存活条件下, 复发事件条件比率的边际效应.

对于偏边际模型, 大多数现存方法均假定协变量对复发事件条件比率和终止事件风险率具有乘性效应 [11, 33]; 加性模型作为乘性模型的重要替代, Pan 和 Schaubel [22] 对复发事件的条件比率建立了加性模型, 而对终止事件风险率构造了乘性模型; Zeng 和 Cai [34] 研究了可加比率模型, 并对终止事件风险率也构造了乘性模型. Chen, Shen 和 Chuang [3] 对复发事件比率提出了一个带有乘性脆弱变量的部分 Aalen 可加模型, 也对终止事件风险率假定了乘性模型. 但是他们并没有给出估计的大样本性质. Qu, Sun 和 Liu [25] 对复发事件比率和终止事件风险率均建立了加性模型, 并通过引入一个脆弱变量刻画两者之间的相关性, 给出估计的大样本性质. 然而, Lin 和 Ying [16] 指出: 在实际研究中, 协变量既可能存在乘性效应, 同时也可能存在加性效应. 尽管 Sun 和 Kang [29] 对复发事件比率提出了一个带脆弱变量的可加可乘模型, 但是由于脆弱变量是以非参数的形式存在于基准比函数中, 所以复发事件与终止事件的相关性在他们的模型中是作为冗余

参数对待的. 在许多实例中, 研究者往往十分关心两者之间的相关性.

本文对复发事件的条件比率提出了一个带乘性脆弱变量的可加可乘模型, 也对终止事件采用了一个带乘性脆弱变量的乘性模型. 用同一个脆弱变量来刻画复发事件与终止事件的相关性. 还利用广义估计方程的思想, 得到了回归参数和相关性参数的估计, 最后给出了估计的大样本性质.

2 联合模型

设 $\tilde{N}^R(t)$ 为时间区间 $(0,t]$ 上的复发次数, $X(t)=(Z(t)^T,W(t)^T)^T$ 为 $p+q$ 维外生协变量 [10]. 令 D 表示终止事件时间 (例如, 病人死亡时间), C 表示删失时间. 需要注意的是, 终止事件的发生阻止了复发事件的发生, 即当 $t>D$ 时, $d\tilde{N}^R(t)=0$. 令 $T=C\wedge D$ 及 $\Delta(t)=I\ (T\geq t)$, 其中 $a\wedge b=\min(a,b)$, $I(\cdot)$ 为示性函数. 由于删失的存在, $\tilde{N}^R(t)$ 并不能完全观测, 因此, 我们令 $N^R(t)=\tilde{N}^R(t\wedge T)$ 表示时间区间 $(0,t]$ 内实际观测到的复发事件个数. 进一步, 也令 $N^D(t)=\tilde{N}^D(t\wedge T)$ 表示实际观测到的终止事件个数, 其中 $\tilde{N}^D(t)=I(D\leq t)$. 假设我们观测到了 n 个独立同分布的样本, 数据组成为 $\{N_i^R(t),N_i^D(t),T_i,\Delta_i(t),X_i(t),\,0\leq t\leq T_i,\,i=1,\dots,n\}$.

设 υ 是一个与 $X(t)$ 独立且不可观测到的非负脆弱变量. 根据 Kalbfleisch 等人 [11] 及 Ye 等人 [33] 的文献, 给定 $Z(t)$, $D=s$ 及 υ, 对复发事件过程, 考虑如下的偏边际比率模型:

$$d\Lambda_R(t\,|\,\upsilon)=P\{d\tilde{N}^R(t)=1\,|\,X(t),\,D=s,\,\upsilon\},\quad s\geq t.$$

注意 $d\Lambda_R(t\,|\,\upsilon)$ 可能依赖于协变量 $X(t)$ 和脆弱变量 υ, 但不依赖于终止事件 $D=s\geq t$. 这也就意味着, 在给定协变量的条件下, 脆弱变量刻画了复发事件过程与终止事件过程的相关性. 而且我们还可以验证 $d\Lambda_R(t\,|\,\upsilon)=P\{d\tilde{N}^R(t)=1\,|\,X(t),\,D\geq t,\upsilon\}$, 这也说明在给定 $X(t)$ 和 υ 条件下, $d\Lambda_R(t\,|\,\upsilon)$ 可表示在 t 时刻存活个体将经历复发事件过程的边际比率. 为了便于分析, 对复发过程, 我们考虑如下的可加可乘边际比率模型:

$$d\Lambda_R(t\,|\,\upsilon)=\upsilon\big[\exp\{\gamma^TZ(t)\}d\Lambda_0^R(t)+\beta^TW(t)dt\big]. \tag{2.1}$$

设在给定 $X(t)$ 和 υ 下, $d\Lambda_D(t\,|\,\upsilon)=P\{d\tilde{N}^D(t)=1\,|\,X(t),\,D\geq t,\upsilon\}$ 表示终止事件风险函数. 考虑如下的 Cox 模型:

$$d\Lambda_D(t\,|\,\upsilon)=\upsilon\exp\{\alpha^TX(t)\}d\Lambda_0^D(t), \tag{2.2}$$

其中 α 为 $p+q$ 维回归参数, $d\Lambda_0^D(t)$ 为未知的基准风险率函数. 为了记号的方便, 我们假定模型 (2.1) 与模型 (2.2) 的协变量 $X(t)$ 相同. 然而, 下面所提出的估计方法可以直接推广到处理模型 (2.1) 和 (2.2) 协变量不同的情形. 另外, 根据 Ye, Kalbfleisch 和 Schaubel [33], 我们假定脆弱变量 υ 服从均值为 1, 方差为 θ 的伽玛分布, 其中 θ 未知. 注意这里假定 υ 的均值为 1, 主要是为了模型的可识别性. 在给定 $X(\cdot)$ 的条件下, 假定删失时间 C 与 $\{\tilde{N}^R(\cdot),\tilde{N}^D(\cdot),D,\upsilon\}$ 独立.

3 估计方法

为了符号方便, 令 $\eta=(\gamma^T,\beta^T)^T$. 如果脆弱变量 υ 能够被观测到, 那么对模型 (2.1) 和 (2.2) 分别利用 Liu 等人 [21] 及 Andersen 和 Gill [2] 的方法, 我们可以构造估计方程, 分别得到 η 和 α 的估计. 然而, 在实际中, υ 是不可观测的, 因此不能直接利用前面所述的方法. 下面在给定 $D\geq t$ 和 $X(t)$ 条件下, 通过对模型 (2.1) 和 (2.2) 中脆弱变量求条件期望, 可得一个有关 η 和 α 的边际

模型. 在 υ 服从伽玛分布的假定下, 经计算可得

$$d\Lambda_R(t) = P\{d\tilde{N}^R(t) = 1 \,|\, X(t),\, D \geq t\} = \psi(t)^{-1}\big[\exp\{\gamma^T Z(t)\}d\Lambda_0^R(t) + \beta^T W(t)dt\big], \tag{3.1}$$

$$d\Lambda_D(t) = P\{d\tilde{N}^D(t) = 1 \,|\, X(t), D \geq t\} = \psi(t)^{-1}\exp\{\alpha^T X(t)\}d\Lambda_0^D(t), \tag{3.2}$$

其中

$$\psi(t) = E\big[\upsilon \,|\, X(t),\, D \geq t\big] = \left[1 + \theta\int_0^t \exp\{\alpha^T X(t)\}d\Lambda_0^D(t)\right].$$

定义

$$\begin{aligned} dM_i^R(t) &= \psi_i(t)dN_i^R(t) - \Delta_i(t)\big[\exp\{\gamma^T Z_i(t)\}d\Lambda_0^R(t) + \beta^T W_i(t)dt\big], \\ dM_i^D(t) &= \psi_i(t)dN_i^D(t) - \Delta_i(t)\exp\{\alpha^T X_i(t)\}d\Lambda_0^D(t), \end{aligned}$$

其中

$$\psi_i(t) \equiv \psi_i(t;\theta,\alpha,\Lambda_0^D) = \left[1 + \theta\int_0^t \exp\{\alpha^T X_i(t)\}d\Lambda_0^D(t)\right].$$

在模型假设下, 利用 (3.1) 和 (3.2) 可以验证 $M_i^R(t)$ 和 $M_i^D(t)$ 是零均值过程. 因此, 给定 $\psi(t)$, 类似于文 [2, 21] 的方法, 可以对 $\Lambda_0^R(t)$, $\Lambda_0^D(t)$, η 和 α 分别构造如下估计方程

$$\begin{aligned} &\sum_{i=1}^n \big[\psi_i(t)dN_i^R(t) - \Delta_i(t)\big\{\exp\{\gamma^T Z_i(t)\}d\Lambda_0^R(t) + \beta^T W_i(t)dt\big\}\big] = 0, \quad 0 \leq t \leq \tau, \\ &\sum_{i=1}^n \big[\psi_i(t)dN_i^D(t) - \Delta_i(t)\exp\{\alpha^T X_i(t)\}d\Lambda_0^D(t)\big] = 0, \quad 0 \leq t \leq \tau, \\ &\sum_{i=1}^n \int_0^\tau \big[D_i(t;\eta) - \bar{D}(t;\eta)\big]\big[\psi_i(t)dN_i^R(t) - \Delta_i(t)\beta^T W_i(t)dt\big] = 0, \\ &\sum_{i=1}^n \int_0^\tau \big\{X_i(t) - \bar{X}(t)\big\}\,\psi_i(t)dN_i^D(t) = 0, \end{aligned}$$

其中, τ 是某一设定的常数且满足 $P(T_i \geq \tau) > 0$, $D_i(t;\eta)$ 是关于协变量和参数的光滑权重函数且与 $d\Lambda_0^R(t)$ 无关,

$$\bar{X}(t) = \frac{\sum_{i=1}^n \Delta_i(t)\exp\{\alpha^T X_i(t)\}X_i(t)}{\sum_{i=1}^n \Delta_i(t)\exp\{\alpha^T X_i(t)\}}.$$

不同的 $D_i(t;\eta)$ 会导致 η 估计的有效性不同, 根据 Lin 和 Ying [16] 的讨论, 我们选取

$$D_i(t;\eta) = \left(Z_i^T(t), \frac{W_i^T(t)}{\exp\{\gamma^T Z_i(t)\}}\right)^T.$$

定义 $\bar{D}(t;\eta) = \big(\bar{Z}^T(t;\eta), \bar{W}^T(t;\eta)\big)^T$, 其中

$$\bar{Z}(t;\eta) = \frac{\sum_{i=1}^n \Delta_i(t)\exp\{\gamma^T Z_i(t)\}Z_i(t)}{\sum_{i=1}^n \Delta_i(t)\exp\{\gamma^T Z_i(t)\}}, \quad \bar{W}(t;\eta) = \frac{\sum_{i=1}^n \Delta_i(t)W_i(t)}{\sum_{i=1}^n \Delta_i(t)\exp\{\gamma^T Z_i(t)\}}.$$

此时, η 的估计方程变为:

$$\begin{aligned} &\sum_{i=1}^n \int_0^\tau \big[Z_i(t) - \bar{Z}(t;\eta)\big]\big[\psi_i(t)dN_i^R(t) - \Delta_i(t)\beta^T W_i(t)dt\big], \\ &\sum_{i=1}^n \int_0^\tau \left[\frac{W_i(t)}{\exp\{\gamma^T Z_i(t)\}} - \bar{W}(t;\eta)\right]\big[\psi_i(t)dN_i^R(t) - \Delta_i(t)\beta^T W_i(t)dt\big]. \end{aligned}$$

然而, 函数 $\psi_i(t)$ 依然含有未知参数 θ, α 及 $\Lambda_0^D(t)$. 为了估计 θ, 令 $\tilde{N}_i^R(t)$ 和 D_i 如前所定义, 但是依赖于个体 $i, i=1,\dots,n$. 定义

$$\omega_{1i}(t)=E[\tilde{N}_i^R(t)\,|\,X_i(t),\,D_i=t],\quad \omega_{2i}(t)=E[\tilde{N}_i^R(t)\,|\,X_i(t),\,D_i>t].$$

在模型假定下, 经计算可得

$$\omega_{1i}(t)=(\theta+1)\psi_i(t)^{-1}\int_0^t\big[\exp\{\gamma^T Z_i(t)\}d\Lambda_0^R(u)+\beta^T W_i(u)du\big],$$

$$\omega_{2i}(t)=\psi_i(t)^{-1}\int_0^t\big[\exp\{\gamma^T Z_i(t)\}d\Lambda_0^R(u)+\beta^T W_i(u)du\big].$$

因此

$$\frac{\omega_{1i}(t)}{\omega_{2i}(t)}=\theta+1. \tag{3.3}$$

(3.3) 式可看作一种复发事件过程与终止事件过程的局部相依性度量 [11]. 这也就意味着: 恰在时刻 t 发生终止事件的个体所复发的次数, 要比在时刻 t 依然处于风险的个体的复发次数多. 根据 (3.3) 式, 利用 Kalbfleisch 等 [11] 的方法, 我们可以构造关于 θ 的估计方程如下:

$$\sum_{i=1}^n\int_0^\tau\big\{N_i^R(t)-(\theta+1)Q(t)\omega_{2i}(t)\big\}dN_i^D(t)=0,$$

其中

$$Q(t)=\frac{\sum_{i=1}^n\omega_{2i}(t)^{-1}\Delta_i^*(t)N_i^R(t)}{\sum_{i=1}^n\Delta_i^*(t)},$$

$\Delta_i^*(t)=\Delta_i(t)\{1-N_i^D(t)\}$ 表示个体 i 在 t 时刻依然处于风险集中, 并且在 t 时刻之后发生终止事件.

设 $\zeta=(\eta^T,\alpha^T,\theta,\Lambda_0^R,\Lambda_0^D)^T$. 定义方程 $U(\zeta)=(U_1^T,U_2^T,U_3,U_4,U_5)^T=0$ 的解作为 ζ 的估计, 其中

$$U_1=\sum_{i=1}^n\int_0^\tau\big\{D_i(t;\eta)-\bar{D}(t;\eta)\big\}\big\{\psi_i(t)dN_i^R(t)-\Delta_i(t)\beta^T W_i(t)dt\big\},$$

$$U_2=\sum_{i=1}^n\int_0^\tau\big\{X_i(t)-\bar{X}(t)\big\}\psi_i(t)dN_i^D(t),$$

$$U_3=\sum_{i=1}^n\int_0^\tau\big\{N_i^R(t)-(\theta+1)Q(t)\omega_{2i}(t)\big\}dN_i^D(t),$$

$$U_4=\sum_{i=1}^n\big\{\psi_i(t)dN_i^R(t)-\Delta_i(t)\big[\exp\{\gamma^T Z_i(t)\}d\Lambda_0^R(t)+\beta^T W_i(t)dt\big]\big\},\ 0\le t\le\tau,$$

$$U_5=\sum_{i=1}^n\big\{\psi_i(t)dN_i^D(t)-\Delta_i(t)\exp\{\alpha^T X_i(t)\}d\Lambda_0^D(t)\big\},\ 0\le t\le\tau.$$

设 $\hat{\zeta}=\big(\hat{\eta}^T,\hat{\alpha}^T,\hat{\theta},\hat{\Lambda}_0^R,\hat{\Lambda}_0^D\big)^T$ 为 $U(\zeta)=0$ 的解. 由于每一个参数的估计均依赖于其他的参数, 因此, 我们提出了一种稳健且有效的迭代解法:

步骤 0 选取初值 $\theta^{(0)}$, $\alpha^{(0)}$ 及 $\Lambda_0^{D(0)}(t)$;

步骤 1 设 $\psi_i^{(0)}(t)=\psi_i(t;\theta^{(0)},\alpha^{(0)},\Lambda_0^{D(0)})$. 将 $\psi_i^{(0)}(t)$ 代入 $U_1=0$, $U_2=0$, $U_4=0$ 和 $U_5=0$, 然后求解得到更新估计 $\eta^{(1)}$, $\alpha^{(1)}$, $\Lambda_0^{R(1)}(t)$ 及 $\Lambda_0^{D(1)}(t)$;

步骤 2 给定 $\eta^{(1)}$, $\Lambda_0^{R(1)}(t)$ 及 $\psi_i^{(1)}(t)=\psi_i(t;\theta^{(0)},\alpha^{(1)},\Lambda_0^{D(1)})$, 求解 $U_3=0$得到更新估计 $\theta^{(1)}$;

步骤 3 重复步骤 1 和步骤 2 直到收敛.

注 1 对初值 $\theta^{(0)}$, $\alpha^{(0)}$ 和 $\Lambda_0^{D(0)}(t)$ 的选取有多种不同的方法, 我们在下面的模拟研究及实例分析中, 取 $\theta^{(0)}=0$, $\alpha^{(0)}=0$, $\Lambda_0^{D(0)}(t)$ 为 Nelson–Aalen 估计. 我们也尝试了其他的初始值, 并不影响算法的收敛. 对于判断迭代过程收敛, 则采用了绝对偏差 $\leq 10^{-3}$ 为迭代终止准则.

设 $\xi=(\eta^T,\alpha^T,\theta)^T$, $\hat{\xi}=(\hat{\eta}^T,\hat{\alpha}^T,\hat{\theta})^T$, 并令 $\xi_0=(\eta_0^T,\alpha_0^T,\theta_0)^T$ 表示 ξ 的真值. 下面介绍所得估计的渐近性质.

定理 3.1 在附录的正则条件 (C1)–(C4) 下, $\hat{\xi}$, $\hat{\Lambda}_0^R(t)$ 及 $\hat{\Lambda}_0^D(t)$ 几乎处处存在且唯一. 同时, $\hat{\xi}$ 强相合于 ξ_0, 且在 $t\in[0,\tau]$ 上, $\hat{\Lambda}_0^R(t)$ 和 $\hat{\Lambda}_0^D(t)$ 几乎处处一致收敛到 $\Lambda_0^R(t)$ 和 $\Lambda_0^D(t)$.

定理 3.1 说明了估计 $\hat{\xi}$, $\hat{\Lambda}_0^R(t)$ 及 $\hat{\Lambda}_0^D(t)$ 的存在性, 唯一性和它们的强相合性. 接下来, 我们给出 $\hat{\xi}$, $\hat{\Lambda}_0^R(t)$ 和 $\hat{\Lambda}_0^D(t)$ 的渐近分布.

定理 3.2 在附录的正则条件 (C1)–(C4) 下, $\sqrt{n}(\hat{\xi}-\xi_0)$ 依分布收敛于一个均值为 0, 方差为 $\Gamma^{-1}\Sigma(\Gamma^T)^{-1}$ 的多元正态随机变量, 其中 Σ 和 Γ 的定义在附录中. 进一步地, $\big\{\sqrt{n}(\hat{\Lambda}_0^R(t)-\Lambda_0^R(t)),\ \sqrt{n}(\hat{\Lambda}_0^D(t)-\Lambda_0^D(t)),\ t\in[0,\tau]\big\}$ 弱收敛到一个零均值的高斯过程.

一般地, 利用 plug-in 方法可以得到方差的相合估计. 但是, Σ 中含有 Hadamard 导数, 而且其表达式十分复杂, 所以, 在小样本下利用 plug-in 方法估计 Σ 可能不稳定. 因此, 我们采用了 bootstrap 方法估计 $\hat{\xi}$ 的协方差矩阵. 当然, 重抽样的样本量越大和抽样次数越多, 估计会越精确. 在第 4 节的模拟研究中, 我们发现当样本为 200 和 400 时, 100 次 bootstrap 协方差矩阵估计已十分精确.

4 模拟研究

本节利用蒙特卡洛模拟验证所提方法的有限样本性质. 设协变量 Z 服从成功概率为 0.5 的伯努利分布, W 服从 $(0,1)$ 上的均匀分布. 令 $X=(Z,W)^T$. 脆弱变量 υ 产生于参数为 $(1,\theta)$ 的伽玛分布, 并设置 $\theta=0$ 或者 0.5.

给定脆弱变量 υ 和协变量 X, 终止时间 D 由模型 (2.2) 产生, 其中 $\Lambda_0^D(t)=0.5t$, 参数 $\alpha_0=(\alpha_{01},\alpha_{02})^T=(0.5,-0.5)^T$. 复发事件产生于强度为 $\upsilon\big(\lambda_0(t)\exp\{\gamma_0 Z\}+\beta_0 W\big)$ 的泊松过程, 其中 $\lambda_0(t)\equiv 2$, 参数 $\gamma_0=-0.5$, $\beta_0=0.5$. 删失时间产生于 $(0,8)$ 上的均匀分布, 并令 $\tau=6$. 此时删失率大约为 30%. 每个个体的平均复发次数约为 2.5. 重复模拟 500 次, 样本容量为 $n=200$ 或 400. 每次模拟重抽样 100 次估计渐近方差.

表 1 前半部分为 γ, β, α 和 θ 的所提方法的模拟结果, 其中包括参数的偏差 Bias (估计值与真值的差); 估计的标准差 SEE, 样本的标准差 SE 和基于正态分布计算的 95% 经验覆盖率 CP. 结果表明所提出的方法模拟结果表现良好: 所得估计是渐近无偏的, 渐近方差估计十分精确, 95% 经验覆盖率也比较合理.

为了比较, 我们也考虑了 naive 估计方法, 即估计时, 把终止事件作为与复发过程独立的删失时间处理 [21]. 数据设置与表 1 前半部分相同. 表 1 给出了比较结果, 从表 1 的结果可以看到, naive 估计方法只有在终止事件与复发过程独立的情况下 (即 $\theta=0$) 是相合的. 在此种情况下, 对参数 γ, β 和 α 的估计, 所提方法与 naive 方法均是合理的, 且前者的估计方差比后者略大. 这可能是由于后者利用了终止事件与复发过程独立的信息. 然而, 当独立条件不成立 (即 $\theta\neq 0$) 时, naive 估计方法是渐近有偏的, 但我们所提方法依然表现良好.

n	θ	参数	所提估计方法				naive 估计方法			
			Bias	SE	SEE	CP	Bias	SE	SEE	CP
200	0	γ	-0.0044	0.1229	0.1264	0.956	-0.0069	0.1195	0.1223	0.958
		β	-0.0162	0.2739	0.2768	0.946	-0.0150	0.2721	0.2712	0.940
		α_1	0.0085	0.1674	0.1742	0.960	0.0059	0.1633	0.1717	0.960
		α_2	-0.0231	0.2897	0.2996	0.956	-0.0219	0.2880	0.2961	0.956
		θ	0.0035	0.0569	0.0551	0.934				
	0.5	γ	-0.0282	0.2527	0.2851	0.972	-0.1879	0.2249	0.2496	0.950
		β	-0.0398	0.5659	0.5779	0.950	-0.0192	0.3450	0.3450	0.946
		α_1	0.0017	0.2558	0.2710	0.976	-0.1103	0.1749	0.1805	0.976
		α_2	-0.0374	0.4694	0.4775	0.960	0.0910	0.3216	0.3158	0.960
		θ	-0.0024	0.1277	0.1362	0.948				
400	0	γ	0.0016	0.0818	0.0849	0.968	0.0000	0.0820	0.0829	0.958
		β	-0.0222	0.1910	0.1924	0.938	-0.0214	0.1898	0.1903	0.932
		α_1	0.0046	0.1235	0.1217	0.944	-0.0033	0.1223	0.1203	0.944
		α_2	0.0042	0.2117	0.2062	0.928	0.0054	0.2106	0.2049	0.928
		θ	0.0027	0.0356	0.0371	0.960				
	0.5	γ	-0.0031	0.1724	0.1814	0.948	-0.1725	0.1536	0.1599	0.872
		β	0.0049	0.3874	0.4043	0.954	0.0154	0.2301	0.2421	0.950
		α_1	0.0171	0.1851	0.1817	0.954	-0.0997	0.1296	0.1263	0.954
		α_2	-0.0017	0.3091	0.3113	0.952	0.1093	0.2235	0.2190	0.952
		θ	-0.0033	0.0969	0.0919	0.932				

表 1　所提方法与 naive 法的比较结果

最后, 通过数值模拟来检验所提方法在伽玛分布假定错误时的表现. 考虑下面三种情况:

(i) υ 服从 $(0,2)$ 上的均匀分布;

(ii) υ 服从均值为 1, 方差为 0.5 的对数正态分布;

(iii) $\tilde{\upsilon}$ 服从均值为 10 的泊松分布, 令 $\upsilon=\tilde{\upsilon}/10$.

其他设置与表 1 前半部分相同, 其中

$$\gamma_0=-0.5,\quad \beta_0=0.5,\quad \alpha_0=(\alpha_{01},\quad \alpha_{02})^T=(0.5,-0.5)^T.$$

表 2 为模拟结果. 从结果可以看出, 所提出的方法在上面三种设置下依然表现合理, 这表明我们的方法在脆弱变量分布假定错误时 (至少在上面所考虑的情况下), 具有一定的稳健性.

脆弱变量	参数	Bias	SE	SEE	CP
均匀分布	γ	-0.0344	0.3079	0.3647	0.978
	β	0.0319	0.6674	0.6958	0.948
	α_1	0.0041	0.2845	0.3322	0.960
	α_2	-0.0031	0.4875	0.5030	0.952
对数正态分布	γ	0.0193	0.2082	0.2282	0.952
	β	-0.0673	0.4784	0.4762	0.948
	α_1	0.0173	0.2275	0.2380	0.952
	α_2	-0.0670	0.4197	0.4069	0.944
泊松分布	γ	-0.0073	0.1610	0.1639	0.946
	β	-0.0162	0.3600	0.3602	0.940
	α_1	0.0046	0.2009	0.1919	0.936
	α_2	-0.0167	0.3313	0.3309	0.948

表 2　脆弱变量分布假定错误时敏感性分析结果 ($n=200$) 所提方法与 naive 法的比较结果

5 实例分析

本节将所提出的方法运用到一个慢性心脏病病人的医疗花费的临床数据中. 此数据来自弗吉尼亚大学的医疗卫生系统, 共有 1475 个年龄在 60–89 岁之间的病人的信息. 这些病人在 2004 年被初次诊断出患有心脏病且接受了治疗. 每个病人的跟踪观测结束时间为此病人在 2006 年 7 月 31 日最后一次去医院就诊的时间或在此期间的死亡时间. 在观测期间, 有 297 个病人 (约 20%) 死亡, 其他病人均被删失了. 对每个病人均测量了三个指标: 性别, 种族和年龄. 初步研究表明, 去医院就诊次数越多的病人趋向于有更高的死亡率. 即病人死亡 (终止事件) 与去医院就诊 (复发事件) 有很强的相关性 [18, 28]. 我们用所提出的方法对死亡风险率与复发事件风险率进行综合分析, 探究性别, 种族和年龄三个协变量对病人死亡及去医院就诊的影响.

令 Z_{i1} 为病人的性别指标 (男性 = 1, 女性 = 0). 根据 Sun 等人 [28] 的描述, 定义 Z_{i2} 为年龄指标, 对 60–69, 70–79 和 80–89 年龄段分别取值为 0, 1 和 2. 所以 $Z_i = (Z_{i1}, Z_{i2})^T$. 令 W_i 为种族指标 (白种人 = 1, 非白种人 = 0). 设 τ 为病人最长跟踪观测时间. 渐近方差利用 100 次重抽样估计. 估计结果如表 3 所示. 结果表明: 种族和年龄均对病人去医院就诊过程的风险率有关, 但是性别并没有显著效应. 具体地, 白种人病人去医院就诊的风险率较低, 而老人却有很高的风险率. 而就病人的死亡过程而言, 性别和年龄在 0.05 水平下均有显著效应, 但是种族因子却没有显著效应. 具体地, 男性病人与老年病人均趋向于有更高的死亡风险率. 由表 3 中结果可知

$$\hat{\theta} = 0.3661(p\text{-value} < 0.0001),$$

即表明病人去医院的就诊过程与病人的死亡过程存在显著的相关性. 也就是说, 病人去医院就诊越频繁, 那么也就有较高的死亡风险率. 而且从 (3.3) 式可知, 在 t 时刻死亡病人的就诊次数大约是在 t 时刻仍处于风险病人的 1.37 倍.

	γ		β	α			θ
	性别	年龄	种族	性别	年龄	种族	相关性
				所提估计方法			
Est	-0.0094	0.1082	-0.0691	0.2769	0.5056	-0.2765	0.3661
SEE	0.0275	0.0201	0.0240	0.1337	0.0887	0.1425	0.0501
p-value	0.7352	< 0.0001	0.0040	0.0384	< 0.0001	0.0524	< 0.0001
				naive 估计方法			
Est	-0.0300	0.0729	-0.0511	0.2502	0.4707	-0.2597	
SEE	0.0256	0.0184	0.0201	0.1233	0.0814	0.1331	
p-value	0.2424	< 0.0001	0.0109	0.0425	< 0.0001	0.0510	

表 3 弗吉尼亚大学慢性心脏病病人医疗费用数据分析

注: Est 表示参数的估计值, SEE 表示方差的估计值

作为比较, 在不考虑两个过程相关性的条件下, 我们用 naive 估计方法对数据做了分析. 结果总结在表 3 的下半部分. 从结果中可以看出, 虽然 naive 方法所得估计的显著性与我们所提方法的相近, 但就病人的就诊过程而言, naive 方法对种族因子的估计比所提方法略小. 而就病人的死亡过程而言, naive 方法对性别因子的估计也比所提方法小. 与模拟研究中所展示的类似, 这可能是因为 naive 方法忽略了两个过程之间的相关性, 导致了估计存在偏差. 相反地, 我们所提出方法却用两个过程所共有的一个脆弱变量刻画了其相关性, 得到了比较合理的估计.

6 总结与展望

本文提出了一个复发事件与终止事件的联合模型, 并利用广义估计方程的思想, 得到了参数的样本估计, 并给出了估计得大样本性质. 模拟研究结果表明, 所提出的方法表现良好. 而且在脆弱变量分布假设错误的情况下, 所提方法也具有一定的稳健性. 最后, 利用所提方法对弗吉尼亚大学慢性心脏病病人医疗费用数据进行了实例分析.

需要注意的是, 模型 (2.1) 和 (2.2) 仅是刻画了复发事件与终止事件的正相关性. 然而在实际中, 两者之间可能是负相关的. 在这种情况下, 对复发事件过程, 可以考虑下面的模型:

$$d\Lambda_R(t\,|\,\upsilon)=\upsilon^{-1}\big[\exp\{\gamma^T Z(t)\}d\Lambda_0^R(t)+\beta^T W(t)dt\big],$$

且模型 (2.2) 不变, 其中 υ 服从均值为 1, 方差为 $\theta<1$ 的伽玛分布. 可以验证

$$\psi^*(t)=E[\upsilon^{-1}\,|\,X(t),\,D\geq t]=\psi(t)/(1-\theta),$$

且 $\omega_{1i}(t)/\omega_{2i}(t)=1-\theta$. 上式意味着, 恰在时刻 t 发生终止事件的个体所复发的次数, 要比在时刻 t 依然处于风险的个体的复发次数少. 因此, 复发事件与终止事件是负相关的. 另外, 将 U_1, U_3 和 U_4 中的 $\psi(t)$ 和 $1+\theta$ 分别替换为 $\psi^*(t)^{-1}$ 和 $1-\theta$, 同样地, 如上文所述, 我们可以构造参数的估计方程.

然而, 在实际数据分析中, 如何确定哪些协变量为 $Z(t)$ 和 $W(t)$ 仍然存在问题. 依靠数据的实际背景, 如果预判协变量对复发事件比率函数具有较大的效应, 则将其归为 $Z(t)$; 如果预判协变量对复发事件比率函数具有较大的绝对效应, 那么应将其归为 $W(t)$. 当然, 如果协变量的维数很小时, 可以通过拟合不同的模型判别协变量为乘性效应还是加性效应. 但是, 当数据的实际背景并不十分清晰时, 利用数据来判断协变量的不同效应是十分必要的, 这有待于进一步的研究.

本文采用了广义估计方程的方法, 因此所得到的估计可能不是半参有效的, 如何获得更有效的参数估计, 仍需要做进一步的研究. 另外, 为了评估模型拟合的精确性, 可以考虑一些基于残差 $d\hat{M}_i^R(t)$ 和 $d\hat{M}_i^D(t)$ 的绘图和数值方法 [13], 其中 $d\hat{M}_i^R(t)$ 和 $d\hat{M}_i^D(t)$ 可分别由 $dM_i^R(t)$ 和 $dM_i^D(t)$ 中的参数替换为其估计所得. 其理论性质和数值模拟也是值得将来要研究的问题.

7 附录: 渐近性质证明

为了推导估计的大样本性质, 我们假定下面的正则条件:

(C1) $\{N_i^R(\cdot),N_i^D(\cdot),T_i,X_i(\cdot)\}$, $i=1,\dots,n$ 是独立同分布的.

(C2) $E\{N_i^R(\tau)\}$ 是有界的, $\Pr\,(T\geq\tau)>0$.

(C3) $Z_i(t)$ 在 $[0,\tau]$ 上几乎处处为有界变差函数.

(C4) 存在 η_0 的一个紧致集 $\mathscr{B}$, 满足 $\eta\in\mathscr{B}$, $\Gamma(\eta)$ 是非奇异的, 其中 $\Gamma(\eta)$ 是 $-\partial\tilde{U}(\eta)/\partial\eta^T$ 的极限值, $\tilde{U}(\eta)$ 的定义在 (A.4) 中给出.

为了证明过程的简便, 我们假定协变量与时间 t 无关. 协变量与时间 t 有关的情况要较为繁琐, 但证明过程是类似的.

定理 3.1 的证明　设 $\psi_i(t;\Lambda^D)=\psi_i(t;\theta,\alpha,\Lambda_0^D)$ 及 $\Lambda_0^R(t;\Lambda_0^D)=\Lambda_0^R(t;\alpha,\theta,\Lambda_0^D)$. 对任意向量 γ^* 及协变量 Z^*, 定义

$$S_{kZ^*}(t;\gamma^*)=\frac{1}{n}\sum_{i=1}^n\Delta_i(t)\exp\{\gamma^{*T}Z_i^*(t)\}Z_i^{*\otimes k}(t),\quad k=0,1,$$

其中, 对任意向量 a, $a^{\otimes 0}=1$, $a^{\otimes 1}=a$. 设

$$\bar{N}^D(t)=n^{-1}\sum_{i=1}^n \psi_i(t;\tilde{\Lambda}_0^D)N_i^D(t),\quad \bar{N}^R(t)=n^{-1}\sum_{i=1}^n \psi_i(t;\tilde{\Lambda}_0^D)N_i^R(t),$$

及 $s_{kZ^*}(t;\gamma^*)$ 为 $S_{kZ^*}(t;\gamma^*)$ 的极限 $(k=0,1)$.

对给定的 ξ, 令 $\tilde{\Lambda}_0^R(t;\xi)$ 及 $\tilde{\Lambda}_0^D(t;\xi)$ 分别为 $U_4=0$ 和 $U_5=0$ 的解. 那么

$$\tilde{\Lambda}_0^R(t;\xi)=\frac{1}{n}\sum_{i=1}^n\int_0^t \frac{\psi_i(u;\tilde{\Lambda}_0^D)dN_i^R(u)-\Delta_i(u)\beta^TW_i(u)du}{S_{0Z}(u;\gamma)},$$

且 $\tilde{\Lambda}_0^D(t;\xi)$ 满足积分方程

$$\tilde{\Lambda}_0^D(t;\xi)=\frac{1}{n}\sum_{i=1}^n\int_0^t \frac{\psi_i(u;\tilde{\Lambda}_0^D)dN_i^D(u)}{S_{0X}(u;\alpha)}. \tag{A.1}$$

(A.1) 是一个 Volterra 积分方程, 且存在唯一的解 [24]. 给定 ξ, 设下列方程

$$\Lambda^D(t;\xi)=\int_0^t \frac{E\{d\bar{N}^D(u)\}}{s_{0X}(u;\alpha)} \tag{A.2}$$

的解为 $\Lambda^D(t;\xi)$. 注意到上述方程也是一个 Volterra 积分方程且有唯一的解满足

$$\Lambda^D(t;\xi_0)\equiv\Lambda_0^D(t).$$

定义

$$\Lambda(t;\xi)=\frac{1}{n}\sum_{i=1}^n\int_0^t \exp\{\alpha^TX_i(u)\}\frac{\theta dN_i^D(u)}{S_{0X}(u;\alpha)}.$$

利用 (A.1) 和 (A.2) 可得

$$\tilde{\Lambda}_0^D(t;\xi)-\Lambda^D(t;\xi)=\int_0^t \left[\tilde{\Lambda}_0^D(u;\xi)-\Lambda^D(u;\xi)\right]d\Lambda(t;\xi)+\mathcal{E}_n(t;\xi),$$

其中

$$\mathcal{E}_n(t;\xi)=\frac{1}{n}\sum_{i=1}^n\int_0^t \frac{dM_i^D(u)}{S_{0X}(u;\alpha_0)}.$$

利用一致强大数定律, 可知在 $t\in[0,\tau]$ 和 $\xi\in\mathscr{B}$ 上, 几乎必然一致地有 $\mathcal{E}_n(t;\xi)\to 0$. 由于上式也是一个 Volterra 积分方程, 求解可得:

$$\tilde{\Lambda}_0^D(t;\xi)-\Lambda^D(t;\xi)=\frac{1}{P_n(t)}\int_0^t P_n(u-)d\mathcal{E}_n(u;\xi), \tag{A.3}$$

其中

$$P_n(t)=\prod_{s\le t}\{1-d\Lambda(s;\xi)\}$$

表示关于 $\Lambda(s;\xi)$ 在 $[0,t]$ 上的乘积积分 (见文 [1, 定理 II. 6.3]) 及 $P_n(u-)$ 表示 $P_n(u)$ 的右连续函数. 利用乘积积分的渐近性质 [8], 一致强大数定律 [23] 及文 [17, 引理 A.1] 可知: 关于 $t\in[0,\tau]$ 和 $\xi\in\mathscr{B}$, $\tilde{\Lambda}_0^D(t;\xi)$ 几乎必然一致地收敛到 $\Lambda^D(t;\xi)$.

设

$$\Lambda^R(t;\xi)=\int_0^t \frac{E\{d\bar{N}^R(u)\}-\beta^Ts_{1W}^*(u)du}{s_{0Z}(u;\gamma)},$$

且 $\Lambda^R(t;\eta_0)\equiv\Lambda_0^R(t)$, 其中, $s_{1W}^*(u)$ 为 $S_{kW}(t;0)$ 的极限值. 类似地可知, 在 $t\in[0,\tau]$ 和 $\xi\in\mathscr{B}$ 上, $\tilde\Lambda_0^R(t;\xi)$ 几乎必然一致收敛到 $\Lambda^R(t;\xi)$.

故要想证明 $\hat\xi$, $\hat\Lambda_0^D(t)$ 及 $\hat\Lambda_0^R(t)$ 的存在性和唯一性, 只需要证明下述方程存在唯一的解即可

$$\tilde U(\xi)=\big(\tilde U_1(\xi)^T,\tilde U_2(\xi)^T,\tilde U_3(\xi)\big)^T=0, \tag{A.4}$$

其中

$$\tilde U_1(\xi)=\sum_{i=1}^n\int_0^\tau\big\{D_i(t;\eta)-\bar D(t;\eta)\big\}\big\{\psi_i(t;\tilde\Lambda_0^D)dN_i^R(t)-\Delta_i(t)\beta^TW_i(t)dt\big\},$$

$$\tilde U_2(\xi)=\sum_{i=1}^n\int_0^\tau\big\{X_i(t)-\bar X(t)\big\}\psi_i(t;\tilde\Lambda_0^D)dN_i^D(t),$$

$$\tilde U_3(\xi)=\sum_{i=1}^n\int_0^\tau\big\{N_i^R(t)-(\theta+1)\tilde Q(t)\tilde\omega_{2i}(t)\big\}dN_i^D(t),$$

以及

$$\tilde\omega_{2i}(t)=\psi_i^{-1}(t;\tilde\Lambda_0^D)\int_0^t\big[\exp\{\gamma^TZ_i(u)\}d\tilde\Lambda_0^R(u)+\beta^TW_i(u)du\big],$$

且 $\tilde Q(t)$ 为 $Q(t)$ 中的 $\omega_{2i}(t)$ 替换为 $\tilde\omega_{2i}(t)$ 所得.

设

$$\hat\Gamma(\xi)=-n^{-1}\partial\tilde U(\xi)/\partial\xi^T.$$

那么, 利用一致强大数定律及 $\tilde\Lambda_0^D(t;\hat\xi)$ 和 $\tilde\Lambda_0^R(t;\hat\xi)$ 的一致收敛性, 可得 $\hat\Gamma(\xi)$ 关于 $\xi\in\mathscr{B}$ 一致收敛到一个非随机函数 $\Gamma(\xi)$. 也可以验证几乎处处有

$$n^{-1}\tilde U(\xi_0)\to 0.$$

因此, $\hat\Gamma(\xi)$ 的一致相合性及条件 (C4) 意味着: 当 n 充分大时, 在 $\xi\in\mathscr{B}$ 上, $\hat\Gamma(\xi)$ 非奇异. 由逆函数定理 (见文 [26, 第 221 页]) 可知: 在 $\mathscr{B}$ 内, 当 n 充分大时, $\tilde U(\xi)=0$ 存在唯一的解 $\hat\xi$, 即存在唯一的估计

$$\hat\xi,\hat\Lambda_0^D(t)\equiv\tilde\Lambda_0^D(t;\hat\xi)\quad\text{和}\quad\hat\Lambda_0^R(t)\equiv\tilde\Lambda_0^R(t;\hat\xi)\quad(0\le t\le\tau).$$

对 (A.4) 利用泰勒展式, 可得

$$n^{-1}\tilde U(\hat\xi)-n^{-1}\tilde U(\xi_0)=-\hat\Gamma(\xi_0)(\hat\xi-\xi_0)+o(\|\hat\xi-\xi_0\|).$$

因此几乎处处有

$$\hat\Gamma(\xi_0)(\hat\xi-\xi_0)+o(\|\hat\xi-\xi_0\|)=o(1).$$

由于在 $\mathscr{B}$ 内, $\hat\Gamma(\xi)$ 的非奇异性, 上述等式意味着 $\hat\xi$ 是强相合的. 根据 $\tilde\Lambda_0^D(t;\xi)$ 和 $\tilde\Lambda_0^R(t;\xi)$ 的一致收敛性可知: 关于 $t\in[0,\tau]$, $\hat\Lambda_0^D(t)$ 和 $\hat\Lambda_0^R(t)$ 几乎处处一致地分别收敛到

$$\Lambda^D(t;\xi_0)\equiv\Lambda_0^D(t)\quad\text{和}\quad\Lambda^R(t;\xi_0)\equiv\Lambda_0^R(t).$$

证毕.

定理 3.2 的证明 令 $\tilde\Lambda_0^D(t)=\hat\Lambda_0^D(t;\xi_0)$. 可以验证

$$\tilde\Lambda_0^D(t)-\Lambda_0^D(t)=\sum_{i=1}^n\int_0^t[\tilde\Lambda_0^D(u)-\Lambda_0^D(u)]d\Lambda(u;\xi_0)+\frac1n\sum_{i=1}^n\int_0^t\frac{dM_i^D(u)}{S_{0X}(u;\alpha_0)}.$$

上式是一个 Volterra 积分方程, 求解可得

$$\tilde{\Lambda}_0^D(t)-\Lambda_0^D(t)=\frac{1}{\hat{P}(t;\xi_0)}\int_0^t \hat{P}(u-;\xi_0)\frac{\sum_{i=1}^n dM_i^D(u)}{nS_{0X}(u;\alpha_0)},$$

其中

$$\hat{P}_n(t;\xi_0)=\prod_{s\leq t}\{1-d\Lambda(s;\xi_0)\}$$

表示关于 $\Lambda(s;\xi_0)$ 在 $[0,t]$ 的乘积积分. 利用 $\tilde{\Lambda}_0^D(t)$ 的一致收敛性, 一致强大数定律 [23] 及 Lin 和 Ying 的文 [17, 引理 A.1] 可知, 在 $t\in[0,\tau]$ 上, 一致地有

$$\tilde{\Lambda}_0^D(t)-\Lambda_0^D(t)=\frac{1}{n}\sum_{i=1}^n\phi_{1i}(t)+o_p(n^{-1/2}), \tag{A.5}$$

其中

$$\phi_{1i}(t)=\frac{1}{P(t;\xi_0)}\int_0^t P(u-;\xi_0)\frac{dM_i^D(u)}{s_{0X}(u;\alpha_0)},$$

且 $P(t;\xi_0)$ 是 $\hat{P}_n(t;\xi_0)$ 的极限.

设

$$\Lambda^*(t;\xi_0)=\frac{1}{n}\sum_{i=1}^n\int_0^t \exp\{\alpha_0^TX_i(u)\}\frac{\theta_0 dN_i^R(u)}{S_{0Z}(u;\gamma_0)}.$$

与 (A.5) 类似可得

$$\begin{aligned}\tilde{\Lambda}_0^R(t)-\Lambda_0^R(t)&=\int_0^t\{\tilde{\Lambda}_0^D(u)-\Lambda_0^D(u)\}d\Lambda^*(u;\xi_0)+\frac{1}{n}\sum_{i=1}^n\int_0^t\frac{dM_i^R(u)}{S_{0Z}(u;\gamma_0)}\\&=\frac{1}{n}\sum_{i=1}^n\phi_{2i}(t)+o_p(n^{-1/2}),\end{aligned} \tag{A.6}$$

其中

$$\phi_{2i}(t)=\int_0^t\phi_{1i}(u)d\Lambda^*(u;\xi_0)+\int_0^t\frac{dM_i^R(u)}{s_{0Z}(u)}.$$

在 (A.4) 中, 对 $\tilde{U}_1(\xi_0)$ 可分解为

$$\begin{aligned}\tilde{U}_1(\xi_0)=&\sum_{i=1}^n\int_0^\tau\left\{D_i(t;\eta_0)-\bar{D}(t;\eta_0)\right\}\left\{\psi_i(t;\tilde{\Lambda}_0^D)dN_i^R(t)-\Delta_i(t)\beta^TW_i(t)dt\right\}\\=&\theta_0\sum_{i=1}^n\int_0^\tau\Delta_i(t)\exp\{\alpha_0^TX_i(t)\}\left\{D_i(t;\eta_0)-\bar{D}(t;\eta_0)\right\}\left\{\tilde{\Lambda}_0^D(t)-\Lambda_0^D(t)\right\}dN_i^R(t)\\&+\sum_{i=1}^n\int_0^\tau\left\{D_i(t;\eta_0)-\bar{D}(t;\eta_0)\right\}dM_i^R(t).\end{aligned}$$

令

$$\tilde{H}_1(t)=\frac{1}{n}\sum_{i=1}^n\int_0^t\theta_0\Delta_i(u)\exp\{\alpha_0^TX_i(t)\}\left\{D_i(u;\eta_0)-\bar{D}(u;\eta_0)\right\}dN_i^R(u),$$

且 $\bar{d}(t;\eta_0)$ 和 $H_1(t)$ 分别是 $\bar{D}(t;\eta_0)$ 和 $\tilde{H}_1(t)$ 的极限. 可以证

$$\tilde{U}_1(\xi_0)=\sum_{i=1}^{n}\vartheta_{1i}+o_p(n^{1/2}), \tag{A.7}$$

其中

$$\vartheta_{1i}=\int_0^\tau \big\{D_i(t;\eta_0)-\bar{d}(t;\eta_0)\big\}\,dM_i^R(t)+\int_0^\tau \phi_{1i}(t)dH_1(t).$$

类似地,

$$\tilde{U}_2(\xi_0)=\sum_{i=1}^{n}\vartheta_{2i}+o_p(n^{1/2}), \tag{A.8}$$

其中

$$\vartheta_{2i}=\int_0^\tau \{X_i(t)-\bar{x}(t)\}\,dM_i^D(t)+\int_0^\tau \phi_{1i}(t)dH_2(t),$$

$\bar{x}(t)$ 和 $H_2(t)$ 分别是 $\bar{X}(t)$ 和 $\tilde{H}_2(t)$ 的极限, 且

$$\tilde{H}_2(t)=\frac{1}{n}\sum_{i=1}^{n}\int_0^t \theta_0\Delta_i(u)\exp\{\alpha_0^TX_i(u)\}\big\{X_i(u)-\bar{X}(u)\big\}dN_i^D(u).$$

设

$$\begin{aligned}
\omega_{2i}^*(t)&=\psi_i^{-1}(t;\Lambda_0^D)\int_0^t\big[\exp\{\gamma^TZ_i(u)\}d\Lambda_0^R(u)+\beta^TW_i(u)du\big],\\
\tilde{Q}(t)&=\frac{\sum_{i=1}^n\tilde{\omega}_{2i}^{-1}(t)\Delta_i^*(t)N_i^R(t)}{\sum_{i=1}^n\Delta_i^*(t)},
\end{aligned}$$

且 $q(t)$ 为 $\hat{Q}(t)$ 的极限. 然后, 经计算可得

$$\begin{aligned}
\tilde{U}_3(\xi_0)=&\sum_{i=1}^{n}\int_0^\tau\big\{N_i^R(t)-(\theta_0+1)\omega_{2i}^*(t)q(t)\big\}dN_i^D(t)\\
&-(\theta_0+1)\sum_{i=1}^{n}\int_0^\tau\tilde{Q}(t)\big\{\tilde{\omega}_{2i}(t)-\omega_{2i}^*(t)\big\}dN_i^D(t)\\
&-(\theta_0+1)\sum_{i=1}^{n}\int_0^\tau\big\{\tilde{Q}(t)-q(t)\big\}\omega_{2i}^*(t)dN_i^D(t).
\end{aligned}\tag{A.9}$$

利用 (A.5) 和 (A.6), 类似 (A.7) 的证明可得: (A.9) 等式右手边的第二项为

$$\begin{aligned}
&-(\theta_0+1)\sum_{i=1}^{n}\int_0^\tau\big[\tilde{Q}(t)\big\{\psi_i^{-1}(t;\tilde{\Lambda}_0^D)-\psi_i^{-1}(t;\Lambda_0^D)\big\}\\
&\qquad\times\int_0^t\big[\exp\{\beta_0^TX_i(u)\}d\Lambda_0^R(u)+\beta^TW_i(u)du\big]\big]dN_i^D(t)\\
&\qquad-(\theta_0+1)\sum_{i=1}^{n}\int_0^\tau\tilde{Q}(t)\psi_i^{-1}(t;\tilde{\Lambda}_0^D)\exp\{\gamma_0^TZ_i(t)\}\big[\tilde{\Lambda}_0^R(t)-\Lambda_0^R(t)\big]dN_i^D(t)\\
&=(\theta_0+1)\sum_{i=1}^{n}\int_0^\tau\big[\phi_{1i}(t)dH_3(t)-\phi_{2i}(t)dH_4(t)\big]+o_p(n^{1/2}),
\end{aligned}\tag{A.10}$$

其中 $H_3(t)$ 和 $H_4(t)$ 分别是 $\tilde{H}_2(t)$ 和 $\tilde{H}_3(t)$ 的极限, 且

$$\tilde{H}_3(t)=\frac{1}{n}\sum_{i=1}^{n}\int_0^t\theta_0\big[\exp\{\alpha_0^TX_i(s)\}\tilde{Q}(s)\psi_i^{-1}(s;\tilde{\Lambda}_0^D)\omega_{2i}^*(s)\big]dN_i^D(s)$$

和

$$\tilde{H}_4(t)=\frac{1}{n}\sum_{i=1}^{n}\int_0^t\exp\{\gamma_0^TZ_i(s)\}\tilde{Q}(s)\psi_i^{-1}(s;\tilde{\Lambda}_0^D)dN_i^D(s).$$

设 $d\Phi(t)=E\{\omega_{2i}^*(t)\}dN_i^D(t)$,

$$\tilde{H}_5(t)=\frac{\theta_0\sum_{i=1}^n\exp\{\alpha_0^TX_i(t)\}\psi_i^{-1}(t;\tilde{\Lambda}_0^D)\tilde{\omega}_{2i}^{-1}(t)\Delta_i^*(t)N_i^R(t)}{\sum_{i=1}^n\Delta_i^*(t)},$$

$$\tilde{H}_6(t)=\frac{\sum_{i=1}^n\exp\{\gamma_0^TZ_i(t)\}\psi_i^{-1}(t;\tilde{\Lambda}_0^D)\tilde{\omega}_{2i}^{-1}(t)\omega_{2i}^{*-1}(t)\Delta_i^*(t)N_i^R(t)}{\sum_{i=1}^n\Delta_i^*(t)},$$

及 $H_4(t)$ 和 $H_5(t)$ 分别是 $\tilde{H}_4(t)$ 和 $\tilde{H}_5(t)$ 的极限. 类似地, (A.9) 等式右手边的第三项为

$$\begin{aligned}&(\theta_0+1)\sum_{i=1}^n\int_0^\tau\big[\phi_{1i}(t)H_5(t)+\phi_{2i}(t)H_6(t)\big]d\Phi(t)\\&\quad-(\theta_0+1)\sum_{i=1}^n\int_0^\tau\left[\frac{\omega_{2i}^{*-1}(t)\Delta_i^*(t)N_i^R(t)}{E\{\Delta_i^*(t)\}}-\frac{q(t)}{E\{\Delta_i^*(t)\}}\Delta_i^*(t)\right]d\Phi(t)+o_p(n^{1/2}).\end{aligned}\tag{A.11}$$

根据 (A.9)–(A.11), 可得

$$\tilde{U}_3(\xi_0)=\sum_{i=1}^n\vartheta_{3i}+o_p(n^{1/2}),\tag{A.12}$$

其中

$$\begin{aligned}\vartheta_{3i}=&\int_0^\tau\big\{N_i^R(t)-(\theta_0+1)\omega_{2i}^*(t)q(t)\big\}dN_i^D(t)\\&-(\theta_0+1)\int_0^\tau\big[\phi_{1i}(t)dH_3(t)+\phi_{2i}(t)dH_4(t)\big]\\&+(\theta_0+1)\int_0^\tau\big[\phi_{1i}(t)H_5(t)+\phi_{2i}(t)H_6(t)\big]d\Phi(t)\\&-(\theta_0+1)\int_0^\tau\left[\frac{\omega_{2i}^{*-1}(t)\Delta_i^*(t)N_i^R(t)}{E\{\Delta_i^*(t)\}}-\frac{q(t)\Delta_i^*(t)}{E\{\Delta_i^*(t)\}}\right]d\Phi(t).\end{aligned}$$

令

$$\vartheta_i=(\vartheta_{1i}^T,\vartheta_{2i}^T,\vartheta_{3i},\vartheta_{4i}^T)^T$$

及 $\Gamma=\Gamma(\xi_0)$. 根据 (A.7), (A.8), (A.12), 由泰勒展式可得

$$n^{1/2}(\hat{\xi}-\xi_0)=n^{-1/2}\Gamma^{-1}\sum_{i=1}^n\vartheta_i+o_p(1).\tag{A.13}$$

利用多元中心极限定理, $n^{1/2}(\hat{\xi}-\xi_0)$ 依分布收敛于均值为 0, 协方差为 $\Gamma^{-1}\Sigma(\Gamma^T)^{-1}$ 的正态随机变量, 其中 $\Sigma=E\{\vartheta_i\vartheta_i^T\}$.

下面证明 $n^{1/2}\{\hat{\Lambda}_0^D(t)-\Lambda_0^D(t)\}$ 和 $n^{1/2}\{\hat{\Lambda}_0^R(t)-\Lambda_0^R(t)\}$ 的弱收敛性质. 我们注意到

$$\hat{\Lambda}_0^D(t)-\Lambda_0^D(t)=\big\{\hat{\Lambda}_0^D(t;\hat{\xi})-\hat{\Lambda}_0^D(t;\xi_0)\big\}+\big\{\hat{\Lambda}_0^D(t;\xi_0)-\Lambda_0^D(t)\big\}.\tag{A.14}$$

设

$$\hat{\Upsilon}_1(t;\xi) = \partial\hat{\Lambda}_0^D(t,\xi)/\partial\xi.$$

利用一致强大数定律, 可证明 $\hat{\Upsilon}_1(t;\xi)$ 关于 $t \in [0,\tau]$ 和 $\xi \in \mathscr{B}$ 几乎处处一致收敛到一个非随机函数 $\Upsilon_1(t;\xi)$. 利用泰勒展式及 (A.5), (A.13) 和 (A.14),

$$n^{1/2}\{\hat{\Lambda}_0^D(t) - \Lambda_0^D(t)\} = n^{-1/2}\sum_{i=1}^{n}\Psi_{1i}(t) + o_p(1),$$

其中

$$\Psi_{1i}(t) = -\Upsilon_1(t;\xi_0)\Gamma^{-1}\vartheta_i + \phi_{1i}(t).$$

类似地, 利用泰勒展式及 (A.6) 和 (A.13) 可得:

$$n^{1/2}\{\hat{\Lambda}_0^R(t) - \Lambda_0^R(t)\} = n^{-1/2}\sum_{i=1}^{n}\Psi_{2i}(t) + o_p(1),$$

其中

$$\Psi_{2i}(t) = -\Upsilon_2(t;\xi_0)\Gamma^{-1}\vartheta_i + \phi_{2i}(t),$$

及 $\Upsilon_2(t;\xi_0)$ 为 $\partial\hat{\Lambda}_0^R(t;\xi)/\partial\xi$ 的极限. 令

$$\Psi_i(t) = (\Psi_{1i}(t), \Psi_{2i}^T(t))^T.$$

因为对于每一给定的 t, $\Psi_i(t)$ $(i = 1,\dots,n)$ 是独立同分布的零均值随机变量, 由多元中心极限定理可得 $n^{1/2}\{\hat{\Lambda}_0^D(t) - \Lambda_0^D(t)\}$ 和 $n^{1/2}\{\hat{\Lambda}_0^R(t) - \Lambda_0^R(t)\}$ 依任意有限维分布收敛到零均值的多元正态随机变量. 又因为 $\Psi_i(t)$ 可表示为关于 t 的单调函数的和或乘积的形式, 因此也是胎紧的 [31]. 由此知 $n^{1/2}\{\hat{\Lambda}_0^D(t) - \Lambda_0^D(t)\}$ 和 $n^{1/2}\{\hat{\Lambda}_0^R(t) - \Lambda_0^R(t)\}$ 也是胎紧的, 且弱收敛到一个零均值的高斯过程, 且在 (s,t) 处的协方差函数为 $E\{\Psi_i(s)\Psi_i(t)^T\}$. 证毕.

参　考　文　献

[1] Andersen P., Borgan O., Gill R., et al., Statistical Models Based on Counting Processes, Springer, New York, 1993.

[2] Andersen P., Gill R., Cox's regression model for counting processes: A large sample study, *Annals of Statistics*, 1982, **10**: 1100–1120.

[3] Chen C. M., Shen P. S., Chuang Y. W., The partly Aalen's model for recurrent event data with a dependent terminal event, *Statistics in Medicine*, 2016, **35**: 268–281.

[4] Cook R., Lawless J., Marginal analysis of recurrent events and a terminating event, *Statistics in Medicine*, 1997, **16**: 911–924.

[5] Cook R., Lawless J., The Statistical Analysis of Recurrent Events, Springer, New York, 2007.

[6] Cook R., Lawless J., Lakhal-Chaieb L., et al., Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: Application to skeletal complications in cancer metastatic to bone, *Journal of the American Statistical Association*, 2009, **104**: 60–75.

[7] Ghosh D., Lin D., Marginal regression models for recurrent and terminal events, *Statistica Sinica*, 2002, **12**: 663–688.

[8] Gill R., Johansen S., A survey of product-integration with a view toward application in survival analysis, *The Annals of Statistics*, 1990, **18**: 1501–1555.

[9] Huang C., Wang M., Joint modeling and estimation for recurrent event processes and failure time data, *Journal of the American Statistical Association*, 2004, **99**: 1153–1165.

[10] Kalbfleisch J., Prentice R., The Statistical Analysis of Failure Time Data, 2nd edition, Wiley, New York, 2002.

[11] Kalbfleisch J., Schaubel D., Ye Y., et al., An estimating function approach to the analysis of recurrent and terminal events, *Biometrics*, 2013, **69**: 366–374.

[12] Lin D., Wei L., Yang I., et al., Semiparametric regression for the mean and rate function of recurrent events, *Journal of the Royal Statistical Society B*, 2000, **62**: 711–730.

[13] Lin D., Wei L., Ying Z., Checking the Cox model with cumulative sums of martingale-based residuals, *Biometrika*, 1993, **80**: 557–572.

[14] Lin D., Wei L., Ying Z., Semiparametric transformation models for point processes, *Journal of the American Statistical Association*, 2001, **96**: 620–628.

[15] Lin D., Ying Z., Semiparametric analysis of the additive risk model, *Biometrika*, 1994, **81**: 61–71.

[16] Lin D., Ying Z., Semiparametric analysis of general additive-multiplicative hazard models for counting processes, *The Annals of Statistics*, 1995, **23**: 1712–1734.

[17] Lin D., Ying Z., Semiparametric and nonparametric regression analysis of longitudinal data, *Journal of the American Statistical Association*, 2001, **96**: 103–126.

[18] Liu L., Huang X., O'Quigley J., Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data, *Biometrics*, 2008, **64**: 950–958.

[19] Liu D., Schaubel D., Kalbfleisch J., Computationally efficient marginal models for clustered recurrent event data, *Biometrics*, 2012, **68**: 637–647.

[20] Liu L., Wolfe R., Huang X., Shared frailty model for recurrent events and a terminal event, *Biometrics*, 2004, **60**: 747–756.

[21] Liu Y., Wu Y., Can J., et al., Additive-multiplicative rates model for recurrent events, *Lifetime Data Analysis*, 2010, **16**: 353–373.

[22] Pan Q., Schaubel D., Flexible estimation of differences in treatment-specific recurrent event means in the presence of a terminating event, *Biometrics*, 2009, **65**: 753–761.

[23] Pollard D., Empirical Processes: Theory and Applications, Institute of Mathematical Statistics, Hayward, CA, 1990.

[24] Polyanin A., Manzhirov A., Handbook of Integral Equations, 2nd edition, Boca Chapman and Hall/CRC, Raton, FL, 2008.

[25] Qu L., Sun L., Liu L., Joint modeling of recurrent and terminal events using additive models, *Statistics and Its Interface*, 2017, **10**: 699–710.

[26] Rudin W., Principles of Mathematical Analysis, 3nd edition, McGraw-Hill, New York, 1976.

[27] Schaubel D., Zhang M., Estimating treatment effects on the marginal recurrent event mean in the presence of a terminating event, *Lifetime Data Analysis*, 2010, **16**: 451–477.

[28] Sun L., Song X., Zhou J., et al., Joint analysis of longitudinal data with informative observation times and a dependent terminal event, *Journal of the American Statistical Association*, 2012, **107**: 688–700.

[29] Sun L., Kang F., An additive-multiplicative rates model for recurrent event data with informative terminal event. *Lifetime Data Analysis*, 2013, **19**: 117–137.

[30] Sun L., Zhao X., Zhou J., A class of mixed models for recurrent event data, *The Canadian Journal of Statistics*, 2011, **39**: 578–590.

[31] van der Vaart A., Wellner J., Weak Convergence and Empirical Processes, Springer, New York, 1996.

[32] Wang M., Qin J., Chiang C., Analyzing recurrent event data with informative censoring, *Journal of the American Statistical Association*, 2001, **96**: 1057–1065.

[33] Ye Y., Kalbfleisch J., Schaubel D., Semiparametric analysis of correlated recurrent and terminal events. *Biometrics*, 2007, **63**: 78–87.

[34] Zeng D., Cai J., A semiparametric additive rate model for recurrent events with informative terminal event, *Biometrika*, 2010, **97**: 699–712.

[35] Zeng D., Lin D., Semiparametric transformation models with random effects for recurrent events, *Journal of the American Statistical Association*, 2007, **102**: 167–180.

[36] Zeng D., Lin D., Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events, *Biometrics*, 2009, **65**: 746–752.

[37] Zhao X., Zhou J., Sun L., Semiparametric transformation models with time-varying coefficients for recurrent and terminal events, *Biometrics*, 2011, **67**: 404–414.