带有辅助信息的多元失效时间数据的估计标准部分似然方法

杨青龙^①, 马学敏^②, 袁中尚^{③*}

① 中南财经政法大学统计与数学学院, 武汉 430073;

② 武汉大学数学与统计学院, 武汉 430072;

③ 山东大学公共卫生学院, 济南 250012

E-mail: qinglongyang728@gmail.com, maxuemind@sohu.com, yuanzhongshang@163.com

收稿日期: 2011-10-17; 接受日期: 2012-02-27; * 通信作者

国家自然科学基金 (批准号: 11171263) 和山东大学自主创新基金 (批准号: 2011GN041) 资助项目

摘要 在许多实际研究中, 由于预算限制, 主协变量值只能对某一个有效集进行准确测量, 但同时对应此主协变量的辅助信息则对全部个体均可以观测. 利用这些辅助协变量的信息有助于提高统计研究的效率. 本文在基于共同基准危险率的边际模型框架下, 我们提出了一些统计推断方法来分析多元失效时间数据. 对于回归参数, 我们提出标准的估计部分似然方程来估计它, 同时也给出了累积基准危险率函数的 Breslow 型估计. 得到的估计可以证明是相合的和渐近正态的. 利用模拟分析结果来表明了提出的方法在有限样本下的可行性.

关键词 有效集 辅助协变量 边际危险模型 多元失效时间

MSC (2010) 主题分类 62G05, 62N01

1 引言

在现代生物医学研究中, 测量暴露变量成本可能会非常高, 有时只能测量全队列的一个子集的主暴露变量 (例如生物标记和基因性状的化验分析), 这个子集就称为有效集. 通常会同时对全队列测到一个成本相对较低的辅助变量, 如何利用辅助变量信息提高统计效率一直是非常重要的问题. 对带有辅助信息的一元失效时间数据分析已经有了很多方法, 例如 [1–6]. 然而, 带有辅助信息的多元失效时间数据的分析方法却非常少.

分析多元失效时间数据的模型主要有两种: 脆性模型 (例如 [7]) 和边际危险模型 (例如 [8–10]). 当对多元时间内部相关性不感兴趣时, 边际危险模型可以有效地发现总体平均的协变量效应. Gorfine 等^[11] 处理了成对相关失效时间数据在 copula 模型下的测量误差. Hu 和 Lin^[12] 在假定辅助变量和主协变量有相同均值的条件下, 通过修正得分的方法提出了一类相合估计量. Liu 等^[13] 在辅助变量为离散的假定下提出了一个估计的伪部分似然方法. Liu 等^[14] 使用核估计的方法将此方法推广到连续辅助变量的情形. 所有这些方法都是在假定不同的失效类型具有不同的基准危险函数的边际危险模型下进行研究的 (见 [8]). 在很多实际问题中, 假定不同的失效类型具有相同的基准危险函数将会更加合理, Lee 等^[9] 提出了很多具体的医学例子. 基于具有相同基准危险函数的边际危险模型, Liu 和

英文引用格式: Yang Q L, Ma X M, Yuan Z S. Estimated standard partial likelihood method for multivariate failure time data with auxiliary information (in Chinese). *Sci Sin Math*, 2012, 42(6): 563–577, doi: 10.1360/012011-791

Yuan^[15] 从每一组中随机选取一个个体组成一个随机序列, 用来估计相对风险函数, 提出了估计的广义估计方程, 他们方法的核心是传统的鞅论就足以能够证明渐近性质, 该方法具有良好的可行性, 比只使用有效集得到的部分似然估计量更有效. 但是当每组内个体数目很多时, 该方法必然包含更多的序列, 计算会带来很大的负担.

本文针对只测量一个有效集的主协变量的多元失效时间数据, 我们在具有共同基准危险率函数模型下提出了一种新的统计方法. 首先, 通过在总体中平均相应的量, 然后通过估计标准部分似然的方法进行统计推断. 该方法从似然出发并且计算方便. 由于鞅的方法失效, 我们采用经验过程的理论来证明相应估计量的渐近性质.

本文组织如下: 第 2 节给出数据结构和方法, 第 3 节给出估计量的渐近性质, 第 4 节通过模拟验证估计量在有限样本下的可行性, 第 5 节给出总结说明及下一步可能进行的研究, 所有的证明将在附录中给出.

2 模型和估计

2.1 记号和模型

假定在队列研究中有 n 个独立的组, 每组中包含 K 个相关个体. 考虑独立失效时间向量 $T_i = (T_{i1}, \dots, T_{iK})'$, $i = 1, \dots, n$. 假定每组中的个体同分布但不独立. 令 C_{ik} 表示 T_{ik} 对应的删失时间, 观测时间 $X_{ik} = \min(T_{ik}, C_{ik})$, $Y_{ik}(t) = I(X_{ik} \geq t)$ 和 $\Delta_{ik} = I(T_{ik} \leq C_{ik})$ 分别来表示风险过程和删失指示变量, 其中 $I(\cdot)$ 表示示性函数. 记 $\lambda_{ik}(t)$ 为相应的边际风险函数, τ 表示研究结束时间. 假定删失是独立删失.

记 $E_{ik}(t)$ 和 $Z_{ik}(t)$ ($0 \leq t \leq \tau$) 分别为 q 维和 p 维的协变量, 这些协变量在有效集中都能观测到, 但在非有效集中只有 $Z_{ik}(t)$ 可被观测. 假定对每一个 k , 随机抽取一个子队列 (有效集) V_k , 该集合元素个数为 v_k . 令 η_{ik} 为个体 (i, k) 是否被选取到有效集 V_k 中的示性变量, $A_{ik}(t)$ 表示 $E_{ik}(t)$ 的辅助变量且在全部队列中都能观测到. 进一步, 假设给定 E_{ik} , A_{ik} 对回归模型不提供任何额外信息, 即 $\lambda_{ik}(t|E_{ik}(t), Z_{ik}(t), A_{ik}(t)) = \lambda_{ik}(t|E_{ik}(t), Z_{ik}(t))$.

因此, 观测数据可表示为:

$$\begin{aligned} &\{X_{ik}, \Delta_{ik}, Z_{ik}(t), E_{ik}(t), A_{ik}(t), \eta_{ik}, 0 \leq t \leq X_{ik}\}, \quad \text{当 } \eta_{ik} = 1, \\ &\{X_{ik}, \Delta_{ik}, Z_{ik}(t), A_{ik}(t), \eta_{ik}, 0 \leq t \leq X_{ik}\}, \quad \text{当 } \eta_{ik} = 0. \end{aligned}$$

假设 T_{ik} 服从具有共同基准危险率的边际模型:

$$\lambda_{ik}(t|Z_{ik}(t), E_{ik}(t)) = \lambda_0(t) \exp\{\beta_2' Z_{ik}(t) + \beta_1' E_{ik}(t)\}, \quad (2.1)$$

其中 $\lambda_0(t)$ 是待估共同基准危险率函数, $\beta = (\beta_1', \beta_2')'$ 为待估参数.

当个体 (i, k) 在无效集 $\bar{V}_k$ 中时, 只能观测 $Z_{ik}(t)$ 和 $A_{ik}(t)$. 此时, 由 Prentice^[16] 的思想, 可以导出个体 (i, k) 的危险函数:

$$\begin{aligned} \lambda_{ik}(t|Z_{ik}(t), A_{ik}(t)) &= \lambda_0(t) e^{\beta_2' Z_{ik}(t)} E\{e^{\beta_1' E_{ik}(t)} | Y_{ik}(t) = 1, A_{ik}(t), Z_{ik}(t)\} \\ &= \lambda_0(t) e^{\beta_2' Z_{ik}(t)} E\{e^{\beta_1' E_{ik}(t)} | Y_{ik}(t) = 1, A_{ik}^*(t)\}, \end{aligned} \quad (2.2)$$

其中 A_{ik}^* 包含了辅助变量 A_{ik} 以及协变量 Z_{ik} 在给定 A_{ik} 下仍然与 E_{ik} 有关的那一部分信息. 即 A_{ik}^* 满足 $f(E_{ik}(t)|X_{ik}(t) \geq t, Z_{ik}(t), A_{ik}(t)) = f(E_{ik}(t)|X_{ik}(t) \geq t, A_{ik}^*(t))$. 注意到 A^* 仍然满足辅助协变量的假定, 给定 E 和 Z , A^* 对回归模型没有贡献, 即 $\lambda(t|Z(t), E(t), A^*(t)) = \lambda(t|Z(t), E(t))$.

2.2 提出的估计

方程 (2.1) 和 (2.2) 表明相对风险函数是:

$$r_{ik}(\beta, t) = [\exp\{\beta'_1 E_{ik}(t)\} \eta_{ik} + \phi_{ik}(\beta_1, t)(1 - \eta_{ik})] \exp(\beta'_2 Z_{ik}(t)),$$

其中 $\phi_{ik}(\beta_1, t) = E\{e^{\beta'_1 E_{ik}(t)} | Y_{ik}(t) = 1, A_{ik}^*(t)\}$.

如果数据全部被观测到, 则模型 (2.1) 的回归参数 β 可以通过极大化标准部分似然 (见 [9]) 来估计, 即

$$\text{SPL}(\beta) = \prod_{i=1}^n \prod_{k=1}^K \left[\frac{r_{ik}(\beta, X_{ik})}{\sum_{q=1}^K \sum_{l=1}^n Y_{lq}(X_{ik}) r_{lq}(\beta, X_{ik})} \right]^{\Delta_{ik}}. \quad (2.3)$$

但由于非有效集中 E_{ik} 不能被观测, (2.3) 中的 r_{ik} 无法计算, 因此首先需要估计 r_{ik} , 本质上是估计 $\phi_{ik}(\beta_1, t)$. 特别地, 当 $A_{ik}^*(t)$ 是分类变量时, 可通过下式估计 $\phi_{ik}(\beta_1, t)$:

$$\hat{\phi}_{ik}(\beta_1, t) = \frac{\sum_1^K \sum_1^n \eta_{js} Y_{js}(t) I_{(A_{js}^*(t)=A_{ik}^*(t))} \exp(\beta'_1 E_{js}(t))}{\sum_1^K \sum_1^n \eta_{js} Y_{js}(t) I_{(A_{js}^*(t)=A_{ik}^*(t))}}.$$

$\hat{\phi}_{ik}(\beta_1, t)$ 是在时刻 t 还活着, 并且与个体 (i, k) 有着相同分类变量取值的那些个体风险的样本均值. 进一步, 可得到相对风险函数 $r_{ik}(\beta, t)$ 的估计为:

$$\hat{r}_{ik}(\beta, t) = \exp\{\beta'_1 E_{ik}(t)\} \exp\{\beta'_2 Z_{ik}(t)\} \eta_{ik} + \hat{\phi}_{ik}(\beta_1, t) \exp\{\beta'_2 Z_{ik}(t)\} (1 - \eta_{ik}).$$

在 (2.3) 中, 用 $\hat{r}_{ik}$ 代替 r_{ik} , 即可得到估计的标准部分似然函数:

$$\text{ESPL}(\beta) = \prod_{i=1}^n \prod_{k=1}^K \left[\frac{\hat{r}_{ik}(\beta, X_{ik})}{\sum_{q=1}^K \sum_{l=1}^n Y_{lq}(X_{ik}) \hat{r}_{lq}(\beta, X_{ik})} \right]^{\Delta_{ik}}. \quad (2.4)$$

估计 $\hat{\beta}_E$ 就定义为使 (2.4) 最大化的值. $\hat{\beta}_E$ 可以通过求解下面的估计的标准部分似然得分方程而得到, $\hat{U}(\beta) = (\partial/\partial\beta)\{\log \text{ESPL}(\beta)\} = 0$.

令 $\hat{r}_{ik}^{(d)}(\beta, u)$, $d = 0, 1, 2$, 表示 $\hat{r}_{ik}(\beta, u)$ 关于 β 的 d 阶导数. 定义

$$\hat{S}^{(d)}(\beta, t) = n^{-1} \sum_{q=1}^K \sum_{l=1}^n Y_{lq}(u) \hat{r}_{lq}^{(d)}(\beta, u).$$

通过简单计算可得:

$$\hat{U}(\beta) = \sum_{k=1}^K \sum_{i=1}^n \int_0^\tau \frac{\hat{r}_{ik}^{(1)}(\beta, u)}{\hat{r}_{ik}(\beta, u)} dN_{ik}(u) - \int_0^\tau \frac{\hat{S}^{(1)}(\beta, u)}{\hat{S}^{(0)}(\beta, u)} d\bar{N}(u),$$

其中 $N_{ik}(t) = I(X_{ik} \leq t, \Delta_{ik} = 1)$ 是标准的计数过程, $\bar{N}(t) = \sum_{k=1}^K \sum_{i=1}^n N_{ik}(t)$.

基准累积危险率函数 $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ 自然可通过下式估计:

$$\hat{\Lambda}_0(t) = \int_0^t \frac{d\bar{N}(s)}{\sum_{q=1}^K \sum_{i=1}^n Y_{iq}(s) \hat{r}_{iq}(\hat{\beta}_E, s)}.$$

3 渐近性质

为了研究得到的估计的渐近性质, 首先来引入一些记号, 令 $\beta_0 = (\beta'_{10}, \beta'_{20})'$ 表示回归参数的真值. 对于向量 $a = (a_1, \dots, a_p)'$, 令 $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, $a^{\otimes 2} = aa'$.

对 $k = 1, \dots, K$, $r_{ik}^{(d)}(\beta, t)$, $d = 0, 1, 2$, 表示 $r_{ik}(\beta, t)$ 关于 β 的 d 阶导数. 定义

$$\begin{aligned} s_k^{(d)}(\beta, t) &= E\{Y_{ik}(t)r_{ik}^{(d)}(\beta, t)\}, \quad d = 0, 1, 2, \\ s_k^{(3)}(\beta, t) &= E\left(Y_{ik}(t)\left(\frac{r_{ik}^{(1)}(\beta, t)}{r_{ik}(\beta, t)}\right)^{\otimes 2} r_{ik}(\beta_0, t)\right), \\ s_k^{(4)}(\beta, t) &= E\left(Y_{ik}(t)\frac{r_{ik}^{(2)}(\beta, t)}{r_{ik}(\beta, t)}r_{ik}(\beta_0, t)\right), \\ s^{(d)}(\beta, t) &= \sum_{k=1}^K s_k^{(d)}(\beta, t), \quad d = 0, 1, 2, 3, 4, \\ M_{ik}(t) &= N_{ik}(t) - \int_0^t Y_{ik}(u)r_{ik}(\beta_0, u)\lambda_0(u)du. \end{aligned}$$

易证明 $M_{ik}(t)$ 是一个鞅, 简单代数计算可以得到:

$$\begin{aligned} n^{-1/2}\hat{U}(\beta) &= n^{-1/2}\sum_{k=1}^K\sum_{i=1}^n\int_0^\tau\left[\frac{\hat{r}_{ik}^{(1)}(\beta, t)}{\hat{r}_{ik}(\beta, t)} - \frac{\hat{S}^{(1)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)}\right]dM_{ik}(t) \\ &\quad + n^{-1/2}\sum_{k=1}^K\sum_{i=1}^n\int_0^\tau\left[\frac{\hat{r}_{ik}^{(1)}(\beta, t)}{\hat{r}_{ik}(\beta, t)} - \frac{\hat{S}^{(1)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)}\right]r_{ik}(\beta_0, t)Y_{ik}(t)\lambda_0(t)dt. \end{aligned} \quad (3.1)$$

利用经验过程理论, 在附录中我们证明了 $n^{-\frac{1}{2}}\hat{U}(\beta_0)$ 渐近等于

$$\begin{aligned} &n^{-\frac{1}{2}}\sum_{k=1}^K\sum_{i\in\bar{V}_k}\int_0^\tau\left(\frac{r_{ik}^{(1)}(\beta_0, t)}{r_{ik}(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)}\right)dM_{ik}(t) \\ &\quad + n^{-\frac{1}{2}}\sum_{k=1}^K\sum_{j\in V_k}\left[\int_0^\tau\left(\frac{r_{jk}^{(1)}(\beta_0, t)}{r_{jk}(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)}\right)dM_{jk}(t) - \rho\begin{pmatrix} Q_{jk}(\beta_0) \\ H_{jk}(\beta_0) \end{pmatrix}\right], \end{aligned} \quad (3.2)$$

其中

$$\begin{aligned} Q_{ik}(\beta) &= \int_0^\tau\left(\frac{\phi_{ik}^{(1)}(\beta_1, t)}{\phi_{ik}(\beta_1, t)} - \frac{s^{(11)}(\beta, t)}{s^{(0)}(\beta, t)}\right)Y_{ik}(t)(e^{\beta'_1 E_{ik}} - \phi_{ik}(\beta_1, t))\delta_{ik}^*(\beta, t)\lambda_0(t)dt, \\ H_{ik}(\beta) &= \int_0^\tau Y_{ik}(t)(e^{\beta'_1 E_{ik}} - \phi_{ik}(\beta_1, t))\delta_{ik}^{**}(\beta, t)\lambda_0(t)dt, \\ \delta_{ik}^*(t, \beta) &= E(e^{\beta'_2 Z_{ik}(t)} | Y_{ik}(t) = 1, A_{ik}^*(t)), \\ \delta_{ik}^{**}(t, \beta) &= E\left(\left[Z_{ik}(t) - \frac{s^{(12)}(\beta, t)}{s^{(0)}(\beta, t)}\right]e^{\beta'_2 Z_{ik}(t)} \middle| Y_{ik}(t) = 1, A_{ik}^*(t)\right), \\ \rho &= \frac{\sum_s^K (1 - \rho_s)}{\sum_s^K \rho_s}, \\ \rho_k &= \lim_{n \rightarrow \infty} (v_k/n), \end{aligned}$$

$s^{(11)}(\beta, t)$ 和 $s^{(12)}(\beta, t)$ 分别表示 $s^{(0)}(\beta, t)$ 关于 β_1 和 β_2 的一阶导数. 因此,

$$s^{(1)}(\beta, t) = \begin{pmatrix} s^{(11)}(\beta, t) \\ s^{(12)}(\beta, t) \end{pmatrix}.$$

(3.2) 式的两项分别为有效集和无效集中独立同分布的变量的和, 又因 $E(Q_{ik}(\beta_0)) = 0$, $E(H_{ik}(\beta_0)) = 0$, $M_{ik}(t)$ 是个鞅,

$$\frac{r_{jk}^{(1)}(\beta_0, t)}{r_{jk}(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)}$$

可料, 可以得到, 这两项分别依分布收敛到零均值的正态随机向量, 这两项显然是独立的, 因此 $n^{-1/2} \hat{U}(\beta_0)$ 依分布收敛到零均值正态随机向量, 渐近方差为前面提到两部分渐近方差的和.

通过 Taylor 展式, 可以证明 $n^{1/2}(\hat{\beta}_E - \beta_0)$ 的渐近正态性. $\hat{\beta}_E$ 的渐近性质可通过下面定理中给出:

定理 1 在附录给定的正则条件下, $\hat{\beta}_E$ 是 β_0 的一个相合估计. 并且 $n^{1/2}(\hat{\beta}_E - \beta_0)$ 渐近正态的, 其极限分布的均值为 0, 方差阵形式为 $\Sigma_{\text{ESPL}}(\beta_0) = \Sigma^{-1}(\beta_0)[\Sigma_1(\beta_0) + \Sigma_2(\beta_0)]\Sigma^{-1}(\beta_0)'$, 其中,

$$\begin{aligned} \Sigma(\beta) &= - \int_0^\tau \left[\frac{(s^{(1)}(\beta, t))^{\otimes 2}}{s^{(0)}(\beta, t)} - s^{(3)}(\beta, t) \right] \lambda_0(t) dt, \\ \Sigma_1(\beta) &= E \left\{ \sum_{k=1}^K \sqrt{1 - \rho_k} g_{ik}(\beta) \right\}^{\otimes 2}, \quad \Sigma_2(\beta) = E \left\{ \sum_{k=1}^K \sqrt{\rho_k} h_{ik}(\beta) \right\}^{\otimes 2}, \end{aligned}$$

这里,

$$\begin{aligned} g_{ik}(\beta) &= \int_0^\tau \left[\begin{pmatrix} \frac{\phi_{ik}^{(1)}(\beta_1, t)}{\phi_{ik}(\beta_1, t)} \\ Z_{ik}(t) \end{pmatrix} - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right] dM_{ik}(t), \\ h_{ik}(\beta) &= \int_0^\tau \left[\begin{pmatrix} E_{ik}(t) \\ Z_{ik}(t) \end{pmatrix} - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right] dM_{ik}(t) - \rho \begin{pmatrix} Q_{ik}(\beta) \\ H_{ik}(\beta) \end{pmatrix}. \end{aligned}$$

注 当主协变量全部能观测时, 即对任意 k , 有效百分比 $\rho_k = 1$ 时, 方差阵 Σ_{ESPL} 与 Lee 等^[9]给出的方差阵相同.

为了研究累积危险率函数估计 $\hat{\Lambda}_0(t)$ 的渐近性质, 定义如下的度量空间, 令 $\mathcal{D}[0, \tau]$ 表示所有右连左极函数 $\{f(t)\} : [0, \tau] \rightarrow R$ 所组成的空间, 在该空间中定义度量 $\rho(f, g) = \max_{t \in [0, \tau]} |f(t) - g(t)|$, 对于 $f, g \in \mathcal{D}[0, \tau]$. 下面定理给出了 $\hat{\Lambda}_0(t)$ 的渐近结果.

定理 2 在附录给定正则条件下, $\hat{\Lambda}_0(t)$ 在 $t \in [0, \tau]$ 中一致依概率收敛到 $\Lambda_0(t)$, $\sqrt{n}(\hat{\Lambda}_0(t) - \Lambda_0(t))$ 在 $\mathcal{D}[0, \tau]$ 中弱收敛到一个零均值的 Gauss 随机域 $\mathcal{W}(t)$, $\mathcal{W}(s)$ 和 $\mathcal{W}(t)$ 之间的协方差函数是 $C(s, t) = \mathbf{1}'_K A(s, t) \mathbf{1}_K$, 其中 $\mathbf{1}_K$ 是所有元素为 1 的 K 维向量, $A(s, t)$ 是一个 $K \times K$ 矩阵, 其元素为

$$A_{lk}(s, t)(\beta_0) = \sqrt{1 - \rho_l} \sqrt{1 - \rho_k} E(u_{1l}(s, \beta_0) u_{1k}(t, \beta_0)) + \sqrt{\rho_l} \sqrt{\rho_k} E(v_{1l}(s, \beta_0) v_{1k}(t, \beta_0)),$$

其中

$$\begin{aligned} u_{ik}(t, \beta_0) &= \int_0^t \frac{dM_{ik}(s)}{s^{(0)}(\beta_0, s)} - \left[\int_0^t \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)} \lambda_0(s) ds \right]' \Sigma^{-1}(\beta_0) g_{ik}(\beta_0), \\ v_{jk}(t, \beta_0) &= \int_0^t \frac{dM_{jk}(s)}{s^{(0)}(\beta_0, s)} - \left[\int_0^t \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)} \lambda_0(s) ds \right]' \Sigma^{-1}(\beta_0) h_{jk}(\beta_0) \end{aligned}$$

$$-\rho \int_0^t \frac{1}{s^{(0)}(\beta_0, s)} Y_{jk}(t) [e^{\beta'_{10} E_{jk}(t)} - \phi_{jk}(\beta_{10}, t)] \delta_{jk}^*(\beta_0, t) \lambda_0(s) ds.$$

我们可以用相应的经验值代替表达式的值给出渐近方差的相合估计, 在附录中给出了所需的正则条件以及定理 1 和 2 的具体证明.

4 模拟计算

我们通过模拟计算来检验所提出的估计量在有限样本下的可行性. 相关失效时间通过具有指数边际的多元 Clayton 和 Cuzick 联合模型生成^[7]. 给定协变量 $(Z_1, \dots, Z_K, E_1, \dots, E_K)$, $(T_1, \dots, T_K)$ 的联合生存函数为:

$$S(t_1, \dots, t_K | Z_1, \dots, Z_K, E_1, \dots, E_K) = \left\{ \sum_{k=1}^K \exp(\theta^{-1} \lambda_0 t_k e^{\beta'_1 E_k + \beta'_2 Z_k}) - (K-1) \right\}^{-\theta},$$

考虑 $K=2$ 和 $\lambda_0=1$ 的情形, 参数 θ 用来控制 T_1 与 T_2 间的相关性, $\theta \rightarrow \infty$ 时表明二者趋于独立, $\theta \rightarrow 0$ 时表明二者的正相关性越来越强. 简单计算表明, (T_1, T_2) 可由 $(0, 1)$ 上的独立均匀分布变量 u_1, u_2 通过下式产生:

$$\begin{aligned} t_1 &= -\ln(1 - u_1) \exp(-(\beta_1 E_1 + \beta_2 Z_1)), \\ t_2 &= \theta \ln\{(1 - a) + a(1 - u_2)^{-(1+\theta)^{-1}}\} \exp(-(\beta_1 E_2 + \beta_2 Z_2)), \end{aligned}$$

其中 $a = (1 - u_1)^{-1/\theta}$. 部分观测的协变量 (E_1, E_2) 来自于多元正态分布, 其中边际均值为 0, 方差为 1, E_1 和 E_2 的相关系数为 0.2. (Z_1, Z_2) 通过标准正态分布来得到. θ 通过取 0.25, 1.50 和 5.70, 分别刻画生存时间之间很强、中等和很弱的三种相关性. 删失时间用 $(0, \tau)$ 上的均匀分布产生, 我们选择 $\tau = 5.96, 1.57$ 或者 0.39 用来产生近似 20%, 50% 或者 80% 的删失率.

辅助协变量 A 可以通过如下方式产生: 首先产生 $W_k = E_k + e_k$, $k = 1, 2$, 其中, $e_k \sim N(0, \sigma^2)$, a_1, a_2, a_3 分别表示 W_k 的 25%, 50%, 75% 分位数, 当 W_k 分别落在区间 $(-\infty, a_1]$, $(a_1, a_2]$, $(a_2, a_3]$, 或者 (a_3, ∞) 时, 令 A_k 分别取值 0, 1, 2, 或者 3, 这里 σ 用来控制 E 和 W 的相关程度, 也就是 E 和 A 的相关程度. σ 取值小表明 E 和 A 的相关性强. 我们取 $\sigma = 0.1$ 或者 0.7. 在整个模拟中产生 300 个样本, 模拟 1000 次, 通过简单随机抽样的方法使得有效比 ρ 为 0.3 或者 0.5.

将提出的估计量 $\hat{\beta}_E$ 与另外两个估计量 $\hat{\beta}_F$ 和 $\hat{\beta}_V$ 作比较, 其中这两个估计量分别表示用全部个体和有效集的数据的标准部分似然估计 (2.3) 的极大值点. 很明显, $\hat{\beta}_F$ 只能在模拟计算中算出, 因为在实际问题中 E 只能在有效集中观测.

当 $\beta_1 = 0, \beta_2 = -0.2$, 删失率为 50% 时, 表 1 给出了在不同的有效百分比和不同相关性下的模拟结果, 其中生存时间之间的不同相关性是通过参数 θ 的不同取值来刻画的, 模拟表明所有的估计是无偏的, 本文所提出的方差的估计量对参数 β_1 和 β_2 的真正的标准差估计良好, 95% 的置信区间也具有好的覆盖率. $\hat{\beta}_E$ 的标准差总是比 $\hat{\beta}_V$ 的小. 对 β_2 来说, 由于协变量 Z 能一直被观测. $\hat{\beta}_E$ 和 $\hat{\beta}_F$ 具有类似的标准差. 对 $\hat{\beta}_1$ 而言, 当 θ 变大时, $\hat{\beta}_E$ 的标准差变小, 这点说明当生存时间间的相关性变弱时, 所提出的方法更加有效. 和所期望的一样, 当辅助变量包含 E 的信息增多时, 即 σ 变小时, 对参数 β_1 而言, 所提出的估计量具有较小的标准差.

当 $(\beta_1, \beta_2) = (0.693, -0.2)$ 时, 表格 2 和表格 3 给出了删失率为 20% 和 50% 时的模拟结果, 我们同样可以发现和表 1 类似的现象. 对 β_1 , 当有效百分比比较低 (如 30%) 和辅助变量包含 E 的信息比

较少 (如 $\sigma = 0.7$) 时, 所提估计量就变的不是很精确了. 然而, 当样本增大到 500 时, 这种偏差就将会变小. 随着删失率的提高, $\hat{\beta}_E$ 对 $\hat{\beta}_V$ 的相对效率将会变大.

当 $\beta_1 = 0.693$, $\beta_2 = -0.2$, $\theta = 1.5$ 和 $\sigma = 0.1$ 时, 图 1 给出了 $\hat{\beta}_E$ 对 $\hat{\beta}_V$ 的相对效率的结果, 很明显, 当有效百分比变小和删失率变大时, 相对效率也会增大. 这个结果说明我们提出的估计比通过有效集得到的估计更好, 特别是在有效集比例比较低和删失率比较高时.

5 总结

本文基于共同基准危险率的边际风险模型, 我们考虑了一种估计的标准部分似然方法, 并且给出

表 1 当 $\beta_1 = 0, \beta_2 = -0.2$, 删失率为 50% 时, 在模型 $\lambda_{ik}(t; E_{ik}, Z_{ik}) = \exp\{\beta_1 E_{ik} + \beta_2 Z_{ik}\}$ ($k = 1, 2$) 下的模拟结果

r	θ	σ	Est.	$\hat{\beta}_1$					$\hat{\beta}_2$				
				Mean	SD	SE	CI	RE	Mean	SD	SE	CI	RE
30%	0.25		$\hat{\beta}_F$	-0.001	0.063	0.061	0.936	3.191	-0.200	0.060	0.059	0.450	3.218
			$\hat{\beta}_V$	-0.001	0.112	0.109	0.941	1.000	-0.200	0.107	0.108	0.951	1.000
		0.1	$\hat{\beta}_E$	0.001	0.066	0.063	0.931	2.870	-0.200	0.060	0.059	0.945	3.212
		0.7	$\hat{\beta}_E$	-0.001	0.073	0.071	0.935	2.347	-0.200	0.060	0.059	0.944	3.216
	1.50		$\hat{\beta}_F$	0.000	0.061	0.059	0.946	3.445	-0.199	0.060	0.059	0.950	3.519
			$\hat{\beta}_V$	-0.004	0.112	0.108	0.934	1.000	-0.199	0.112	0.109	0.942	1.000
		0.1	$\hat{\beta}_E$	-0.000	0.064	0.062	0.938	3.099	-0.199	0.060	0.059	0.947	3.509
		0.7	$\hat{\beta}_E$	-0.002	0.071	0.070	0.940	2.497	-0.199	0.060	0.059	0.950	3.516
	5.70		$\hat{\beta}_F$	0.000	0.060	0.058	0.945	3.462	-0.199	0.060	0.058	0.949	3.552
			$\hat{\beta}_V$	-0.005	0.111	0.108	0.937	1.000	-0.198	0.113	0.108	0.940	1.000
		0.1	$\hat{\beta}_E$	-0.001	0.063	0.062	0.946	3.125	-0.199	0.060	0.058	0.948	3.554
		0.7	$\hat{\beta}_E$	0.002	0.070	0.069	0.943	2.538	-0.199	0.060	0.058	0.948	3.555
50%	0.25		$\hat{\beta}_F$	0.001	0.063	0.061	0.936	1.877	-0.200	0.060	0.059	0.945	1.864
			$\hat{\beta}_V$	0.000	0.086	0.085	0.951	1.000	-0.201	0.082	0.083	0.960	1.000
		0.1	$\hat{\beta}_E$	0.001	0.065	0.062	0.930	1.732	-0.200	0.060	0.059	0.944	1.863
		0.7	$\hat{\beta}_E$	0.000	0.071	0.067	0.932	1.478	-0.200	0.060	0.059	0.944	1.864
	1.50		$\hat{\beta}_F$	0.000	0.061	0.059	0.946	1.942	-0.199	0.060	0.059	0.950	1.953
			$\hat{\beta}_V$	-0.002	0.084	0.084	0.940	1.000	-0.201	0.083	0.083	0.956	1.000
		0.1	$\hat{\beta}_E$	-0.000	0.063	0.061	0.942	1.806	-0.199	0.060	0.059	0.947	1.951
		0.7	$\hat{\beta}_E$	-0.001	0.068	0.066	0.939	1.541	-0.199	0.060	0.059	0.949	1.953
	5.70		$\hat{\beta}_F$	0.000	0.060	0.058	0.945	1.968	-0.199	0.060	0.058	0.949	2.025
			$\hat{\beta}_V$	-0.003	0.084	0.083	0.944	1.000	-0.199	0.060	0.058	0.948	1.000
		0.1	$\hat{\beta}_E$	-0.000	0.062	0.061	0.949	1.833	-0.199	0.060	0.058	0.941	2.028
		0.7	$\hat{\beta}_E$	-0.001	0.067	0.066	0.949	1.577	-0.199	0.060	0.058	0.946	2.027

r 表示有效百分比, Mean 和 SD 表示参数估计量得均值和标准差, SE 是估计标准差的平均值, CI 表示 95% 置信区间的覆盖率, RE 表示对 $\hat{\beta}_V$ 的相对效率.

表 2 当 $\beta_1 = 0.693, \beta_2 = -0.2$, 删失率为 20% 时, 在模型 $\lambda_{ik}(t; E_{ik}, Z_{ik}) = \exp\{\beta_1 E_{ik} + \beta_2 Z_{ik}\} (k = 1, 2)$ 下的模拟结果

r	θ	σ	Est.	$\hat{\beta}_1$					$\hat{\beta}_2$				
				Mean	SD	SE	CI	RE	Mean	SD	SE	CI	RE
30%	0.25		$\hat{\beta}_F$	0.697	0.061	0.059	0.939	2.963	-0.199	0.048	0.046	0.944	3.530
			$\hat{\beta}_V$	0.705	0.106	0.099	0.930	1.000	-0.200	0.090	0.085	0.945	1.000
		0.1	$\hat{\beta}_E$	0.693	0.072	0.064	0.919	2.129	-0.1923	0.048	0.047	0.939	3.460
			$\hat{\beta}_E$	0.674	0.083	0.076	0.918	1.602	-0.181	0.048	0.047	0.920	3.447
	1.50		$\hat{\beta}_F$	0.696	0.058	0.054	0.933	2.991	-0.199	0.047	0.046	0.945	3.847
			$\hat{\beta}_V$	0.701	0.100	0.097	0.945	1.000	-0.198	0.092	0.085	0.934	1.000
		0.1	$\hat{\beta}_E$	0.691	0.068	0.062	0.917	2.150	-0.193	0.047	0.046	0.949	3.760
			$\hat{\beta}_E$	0.671	0.079	0.075	0.933	1.611	-0.181	0.047	0.047	0.932	3.747
	5.70		$\hat{\beta}_F$	0.694	0.056	0.053	0.936	3.072	-0.199	0.047	0.046	0.945	3.839
			$\hat{\beta}_V$	0.699	0.099	0.096	0.937	1.000	-0.198	0.091	0.085	0.928	1.000
		0.1	$\hat{\beta}_E$	0.690	0.066	0.061	0.927	2.210	-0.192	0.047	0.046	0.943	3.748
			$\hat{\beta}_E$	0.671	0.078	0.074	0.928	1.603	-0.181	0.047	0.047	0.929	3.773
50%	0.25		$\hat{\beta}_F$	0.697	0.061	0.059	0.939	1.825	-0.199	0.048	0.046	0.944	1.912
			$\hat{\beta}_V$	0.700	0.082	0.078	0.928	1.000	-0.200	0.066	0.066	0.945	1.000
		0.1	$\hat{\beta}_E$	0.695	0.068	0.060	0.923	1.486	-0.194	0.048	0.046	0.942	1.868
			$\hat{\beta}_E$	0.687	0.074	0.067	0.917	1.248	-0.186	0.049	0.046	0.926	1.856
	1.50		$\hat{\beta}_F$	0.696	0.058	0.054	0.933	1.836	-0.199	0.047	0.046	0.945	2.145
			$\hat{\beta}_V$	0.697	0.079	0.075	0.936	1.000	-0.199	0.068	0.065	0.944	1.000
		0.1	$\hat{\beta}_E$	0.693	0.064	0.058	0.937	1.534	-0.194	0.047	0.046	0.944	2.096
			$\hat{\beta}_E$	0.685	0.069	0.066	0.939	1.284	-0.186	0.048	0.046	0.939	2.145
	5.70		$\hat{\beta}_F$	0.694	0.056	0.053	0.936	1.863	-0.199	0.047	0.046	0.945	2.192
			$\hat{\beta}_V$	0.696	0.077	0.074	0.938	1.000	-0.199	0.069	0.065	0.933	1.000
		0.1	$\hat{\beta}_E$	0.692	0.062	0.058	0.933	1.551	-0.194	0.047	0.046	0.947	2.155
			$\hat{\beta}_E$	0.684	0.067	0.065	0.945	1.304	-0.186	0.047	0.046	0.935	2.155

了回归参数的一个相合估计. 我们没有对缺失变量和辅助变量之间的关系作任何假定, 从这个意义上来说, 该方法是非参数的. 利用现代经验过程理论, 我们证明了回归参数和累积风险函数估计的渐近性质. 模拟结果表明所提出的方法在中等样本数量下具有良好的可行性, 并且比只使用有效集得到的部分似然估计量更有效. 图 1 表明, 当有效百分比很小并且删失率很高时, 效率将会提高的更多.

然而, 该方法也有很多缺点, 例如辅助变量 A 是离散的, 并且当有效样本很少时 (低于 60), 分类变量数目也不能太大. 当辅助协变量是连续的情况下, 可以考虑将其先离散化, 然后再利用本文中的统计方法进行推断. 其他的模拟也表明当辅助变量 A 包含较少的 E 的信息时, 该方法易于低估参数. 增加有效百分比或者全部的样本将会缓解这个问题.

该方法的核心是通过在对有效集的风险作平均从而估计无效集中未知的相对风险函数, 我们在此只考虑了当组内个体数目相等的情形, 即 K 固定. 当每组内个体数目不相同, 我们同样可以估计相对风险函数, 然后做统计推断. 直觉上, 因为我们用别的全队列个体作平均, 当组内相关性很强时, 该

表 3 当 $\beta_1 = 0.693, \beta_2 = -0.2$, 删失率为 50% 时, 在模型 $\lambda_{ik}(t; E_{ik}, Z_{ik}) = \exp\{\beta_1 E_{ik} + \beta_2 Z_{ik}\} (k = 1, 2)$ 下的模拟结果

r	θ	σ	Est.	$\hat{\beta}_1$					$\hat{\beta}_2$				
				Mean	SD	SE	CI	RE	Mean	SD	SE	CI	RE
30%	0.25		$\hat{\beta}_F$	0.695	0.071	0.069	0.948	3.005	-0.199	0.059	0.058	0.946	3.557
			$\hat{\beta}_V$	0.703	0.123	0.119	0.940	1.000	-0.199	0.111	0.107	0.945	1.000
		0.1	$\hat{\beta}_E$	0.693	0.082	0.076	0.937	2.242	-0.195	0.060	0.058	0.936	3.474
			$\hat{\beta}_E$	0.680	0.095	0.089	0.932	1.680	-0.187	0.059	0.059	0.942	3.486
		1.50	$\hat{\beta}_F$	0.693	0.067	0.065	0.946	3.212	-0.198	0.059	0.058	0.939	3.753
			$\hat{\beta}_V$	0.699	0.121	0.117	0.934	1.000	-0.199	0.114	0.107	0.931	1.000
	0.7		$\hat{\beta}_E$	0.690	0.078	0.074	0.933	2.411	-0.193	0.060	0.058	0.939	3.653
			$\hat{\beta}_E$	0.677	0.091	0.087	0.932	1.761	-0.185	0.059	0.059	0.937	3.699
		1.50	$\hat{\beta}_F$	0.693	0.067	0.064	0.944	3.136	-0.198	0.060	0.058	0.939	3.716
			$\hat{\beta}_V$	0.697	0.118	0.116	0.939	1.000	-0.198	0.115	0.107	0.930	1.000
		0.1	$\hat{\beta}_E$	0.691	0.077	0.0734	0.941	2.355	-0.193	0.060	0.058	0.933	3.620
			$\hat{\beta}_E$	0.678	0.089	0.087	0.933	1.788	-0.185	0.060	0.059	0.928	3.684
50%	0.25		$\hat{\beta}_F$	0.695	0.071	0.069	0.948	1.853	-0.199	0.059	0.058	0.946	1.940
			$\hat{\beta}_V$	0.698	0.097	0.094	0.947	1.000	-0.201	0.082	0.083	0.952	1.000
		0.1	$\hat{\beta}_E$	0.695	0.078	0.072	0.940	1.551	-0.196	0.059	0.058	0.939	1.925
			$\hat{\beta}_E$	0.690	0.083	0.079	0.937	1.341	-0.190	0.059	0.058	0.942	1.926
		1.50	$\hat{\beta}_F$	0.693	0.067	0.065	0.946	1.933	-0.198	0.059	0.058	0.939	2.051
			$\hat{\beta}_V$	0.695	0.094	0.091	0.932	1.000	-0.200	0.085	0.083	0.947	1.000
	0.7		$\hat{\beta}_E$	0.692	0.073	0.070	0.941	1.630	-0.195	0.060	0.058	0.938	2.008
			$\hat{\beta}_E$	0.686	0.080	0.078	0.942	1.386	-0.189	0.059	0.058	0.936	2.026
		1.50	$\hat{\beta}_F$	0.693	0.067	0.064	0.944	1.901	-0.198	0.060	0.058	0.939	2.060
			$\hat{\beta}_V$	0.694	0.092	0.090	0.936	1.000	-0.199	0.085	0.083	0.944	1.000
		0.1	$\hat{\beta}_E$	0.693	0.073	0.069	0.934	1.611	-0.195	0.060	0.058	0.933	2.024
			$\hat{\beta}_E$	0.687	0.078	0.077	0.943	1.400	-0.189	0.060	0.058	0.933	2.043

图 1 当 $\beta_1 = 0.693, \beta_2 = -0.2, \theta = 1.5, \sigma = 0.1$ 时, 对 $\hat{\beta}_1$, 在各种有效百分比和删失率下, $\hat{\beta}_E$ 对 $\hat{\beta}_V$ 的相对效率. 实线对应删失率 20% 的情形, 删失率 50% 和 80% 的情形分别用点虚线和长虚线表示.

方法用起来可能会比较差. 模拟结果同时也表明, 当组内相关性变大. 即 θ 变小时, 该方法的效率会降低. Cai 和 Prentice^[10] 指出, 可以在伪部分似然得分方程中引入适当的权来获得更有效的统计量. 如何在该方法中引入合适的权来提高效率是我们以后的工作.

参考文献

- 1 Hughes M D. Regression dilution in the propotional hazards model. *Biometrics*, 1993, 49: 1056–1066
- 2 Zhou H, Pepe M S. Auxiliary covariate data in failure time regression analysis. *Biometrika*, 1995, 58: 352–360
- 3 Wang C Y, Hsu L, Feng Z D, et al. Regression calibration in failure time regression. *Biometrics*, 1997, 53: 131–145
- 4 Kulich M, Lin D Y. Additive hazards regression with covariate measurement error. *J Amer Statist Assoc*, 2000, 95: 238–248
- 5 Zhou H, Wang C Y. Failure time regression with continuous covariates measured with error. *J R Statist Soc B*, 2000, 62: 657–665
- 6 Jiang J, Zhou H. Additive hazards regression with auxiliary covariates. *Biometrika*, 2007, 94: 359–369
- 7 Clayton D, Cuzick J. Multivariate generalizations of the propotional hazard model (with discussion). *J R Statist Soc A*, 1985, 54: 168–184
- 8 Wei L J, Lin D Y, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Amer Statist Assoc*, 1989, 84: 1065–1073
- 9 Lee E W, Wei L J, Amato D A. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: Klein J P, Goel P K, eds. *Survival Analysis: State of the Art*. Netherlands: Kluwer Academic Publishers, 1992, 237–247
- 10 Cai J, Prentice R L. Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, 1995, 82: 151–164
- 11 Gorfine M, Hsu L, Prentice R L. Estimation of dependence between paired correlated failure times in the presence of covariate measurement error. *J R Statist Soc B*, 2003, 65: 643–661
- 12 Hu C, Lin, D Y. Cox regression with covariate measurement error. *Scand J Stat*, 2002, 29: 637–655
- 13 Liu Y Y, Zhou H, Cai J. Estimated pseudo-partial-likelihood method for correlated failure time data with auxiliary covariates. *Biometrics*, 2009, 65: 1184–1193
- 14 Liu Y Y, Wu Y, Zhou H. Multivariate failure time regression with a continuous auxiliary covariates. *J Multivariate Anal*, 2010, 101: 679–691
- 15 Liu Y Y, Yuan Z. Estimated generalized estimating equation for correlated failure time data with auxiliary covariates. *Comm Statist Theory Methods*, in press
- 16 Prentice R L. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 1982, 69: 331–342
- 17 Kosorok M R. *Introduction to Empirical Process and Semiparametric Inference*. New York: Springer, 2008
- 18 Foutz R V. On the unique consistent solution to the likelihood equations. *J Amer Statist Assoc*, 1977, 72: 147–148
- 19 Andersen P K, Gill R D. Cox's regression model for counting processes: A large samle study. *Ann Statist*, 1982, 10: 1100–1120

附录 定理 1 和定理 2 的条件和证明

假设 A^* 是分类变量, 其分布为 $P(A^* = a_m) = p_m$, $m = 1, \dots, L$. 对向量 $a = (a_1, \dots, a_p)'$, 定义 $a^{\otimes 4} = aa'(aa')'$, $\|a\| = \sup_{1 \leq i \leq p} |a_i|$. 对矩阵 $A = (a_{ij})$, 定义 $A^{\otimes 2} = AA'$, $\|A\| = \sup_{i,j} |a_{ij}|$. 进一步, 假定下面的正则条件成立:

(C1) 有限区间: $\int_0^\tau \lambda_0(t)dt < \infty$;

(C2) 对 $A^*(t)$ 的任意取值 a_m , $\Pr(Y_{ik}(t) = 1 | A_{ik}^*(t) = a_m) > 0$;

(C3) 对任意 $k = 1, \dots, K$, $i = 1, \dots, n$, $(E_{ik}(t), Z_{ik}(t), A_{ik}^*(t))'$ 在区间 $[0, \tau]$ 上几乎处处具有一致有界变差;

(C4) 对任意 $k = 1, \dots, K$, 存在 β_0 的邻域 $\mathcal{B} = \mathcal{B}_1 \times \mathcal{B}_2$, 使得 $\phi_{ik}(\beta_1, t)$ 在 $\mathcal{B}_1 \times [0, \tau]$ 上有界并且大于 0, $\Sigma(\beta_0)$ 正定;

(C5) 对 $d = 0, 1$, $\sup_{t \in [0, \tau]} |L^{(d)}(t)| = O_p(1)$, 其中

$$L^{(d)}(t) = \left[\frac{1}{\sqrt{\sum_{s=1}^K v_s}} \sum_{s=1}^K \sum_{j \in V_s} Y_{js}(t) I_{(A_{js}^*(t)=a_m)} [\gamma_{js}^{(d)}(\beta_1, t) - E(\gamma_{js}^{(d)}(\beta_1, t) | Y_{js}(t) = 1, A_{js}^*(t) = a_m)] \right],$$

$\gamma_{js}(\beta_1, t) = e^{\beta_1' E_{js}(t)}$. 定义 $\phi_{a_m s}(\beta_1, t) = E(e^{\beta_1' E_{js}(t)} | Y_{js}(t) = 1, A_{js}^* = a_m)$. 对 $d = 0$, 该条件等价于 $\sqrt{\sum_{s=1}^K v_s}(\hat{\phi}_{a_m s} - \phi_{a_m s})$ 为 $O_p(1)$, 因为 $\hat{\phi}_{a_m s}$ 是 $\phi_{a_m s}$ 对应的样本量, 由中心极限定理, 该条件的满足是合理的.

下面引理来自于 [17, 54页, 引理4.2], 在整个定理的证明中经常用到.

引理 1 设 $B_n \in D[a, b]$, $A_n \in l^\infty[a, b]$ 是左连右极或者右连左极函数, 其中 $l^\infty[a, b]$ 表示 $[a, b]$ 上的有界函数全体, 假定 $\sup_{t \in [0, \tau]} |A_n(t)| \rightarrow_p 0$. A_n 具有一致有界变差, 并且 B_n 弱收敛到一个紧的, 零均值的过程, 其样本轨道在空间 $D[a, b]$ 里, 那么我们有 $\int_a^b A_n(s) dB_n(s) \rightarrow_p 0$.

定理 1 的证明 定义

$$\begin{aligned} \hat{S}_k^{(d)}(\beta, t) &= n^{-1} \sum_{i=1}^n Y_{ik} \hat{r}_{ik}^{(d)}(\beta, t), \quad d = 0, 1, 2, \\ \hat{S}_k^{(3)}(\beta, t) &= n^{-1} \sum_{i=1}^n Y_{ik}(t) \left(\frac{\hat{r}_{ik}^{(1)}(\beta, t)}{\hat{r}_{ik}(\beta, t)} \right)^{\otimes 2} r_{ik}(\beta_0, t), \\ \hat{S}_k^{(4)}(\beta, t) &= n^{-1} \sum_{i=1}^n Y_{ik}(t) \frac{\hat{r}_{ik}^{(2)}(\beta, t)}{\hat{r}_{ik}(\beta, t)} r_{ik}(\beta_0, t), \\ \hat{S}^{(d)}(\beta, t) &= \sum_{k=1}^K \hat{S}_k^{(d)}(\beta, t), \quad d = 0, 1, 2, 3, 4, \\ \Delta \hat{r}_{ik}(\beta, t) &= \frac{\hat{r}_{ik}^{(2)}(\beta, t)}{\hat{r}_{ik}(\beta, t)} - \left(\frac{\hat{r}_{ik}^{(1)}(\beta, t)}{\hat{r}_{ik}(\beta, t)} \right)^{\otimes 2}, \\ \Delta \hat{S}(\beta, t) &= \frac{\hat{S}^{(2)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} - \left(\frac{\hat{S}^{(1)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} \right)^{\otimes 2}, \\ \Delta s(\beta, t) &= \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} - \left(\frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right)^{\otimes 2}, \\ \hat{A}(\beta) &= \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^n \int_0^\tau [\Delta \hat{r}_{ik}(\beta, t) - \Delta \hat{S}(\beta, t)] Y_{ik}(t) r_{ik}(\beta_0, t) \lambda_0(t) dt. \end{aligned}$$

类似于 Liu 等在 [13] 中的推断, 可以证明, 对 $d = 0, 1, 2, 3, 4$,

$$\sup_{B \times [0, \tau]} \|\hat{S}_k^{(d)}(\beta, t) - s_k^{(d)}(\beta, t)\| \rightarrow 0, \quad \text{a.s.}, \quad (\text{A1})$$

$s_k^{(d)}$ 有界连续, $s_k^{(0)}$ 有界且严格大于 0. 因为 $\Delta \hat{S}(\beta, t) - \Delta s(\beta, t)$ 连续并且具有一致有界变差, $M_{ik}(t)$ 是一个 Donsker 类, $n^{-1/2} \sum_{i=1}^n M_{ik}(t)$ 弱收敛到一个胎紧的零均值过程, 由引理 1 和 (A1), 我们有

$$n^{-1/2} \sum_{i=1}^n \int_0^\tau \Delta \hat{S}(\beta, t) dM_{ik}(t) = n^{-1/2} \sum_{i=1}^n \int_0^\tau \Delta s(\beta, t) dM_{ik}(t) + o_p(1). \quad (\text{A2})$$

根据 Foutz^[18] 的结果, 只要保证下面 4 条成立, 就能说明 $\hat{\beta}_E$ 是 β_0 的相合估计:

- (I) $n^{-1}\partial\hat{U}(\beta)/\partial\beta$ 在 β_0 的一个开邻域内 $\mathcal{B}$ 存在且连续;
- (II) $n^{-1}\partial\hat{U}(\beta_0)/\partial\beta_0$ 以概率 1 负定;
- (III) $n^{-1}\partial\hat{U}(\beta)/\partial\beta$ 在 β_0 的某个开邻域内依概率一致收敛到 $A(\beta)$;
- (IV) $n^{-1}\hat{U}(\beta_0)$ 依概率收敛到 0.

显然, 由 $\hat{r}_{ik}^{(d)}(\beta, t)$ 和 $\hat{S}^{(d)}(\beta, t)$ 的连续性, (I) 成立.

类似于 [19], 可以将 $n^{-1}\partial\hat{U}(\beta)/\partial\beta$ 分解成如下两部分:

$$n^{-1}\partial\hat{U}(\beta)/\partial\beta = n^{-1}\sum_{k=1}^K\sum_{i=1}^n\int_0^\tau [\Delta\hat{r}_{ik}(\beta, t) - \Delta\hat{S}(\beta, t)]dM_{ik}(t) + \hat{A}(\beta). \quad (\text{A3})$$

由 (A2), 可以证明 (A3) 的右边第一项渐近等于

$$n^{-1}\sum_{k=1}^K\sum_{i=1}^n\int_0^\tau [\Delta r_{ik}(\beta, t) - \Delta s(\beta, t)]dM_{ik}(t),$$

上式为鞅, 并且由 Lenglart 不等式易知依概率收敛到 0.

由条件 (C1) 和 (A1), 对 $\beta \in \mathcal{B}$, 可以证明 $\hat{A}(\beta) = A(\beta) + o_p(1)$, 其中

$$A(\beta) = \int_0^\tau \left[s^{(4)}(\beta, t) - \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} s^{(0)}(\beta_0, t) - s^{(3)}(\beta, t) + \left(\frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right)^{\otimes 2} s^{(0)}(\beta_0, t) \right] \lambda_0(t) dt.$$

当 $\beta = \beta_0$ 时, $s_k^{(4)}(\beta_0, t) = s_k^{(2)}(\beta_0, t)$, $A(\beta_0) = -\Sigma(\beta_0)$ 为负定的. 因此 (II) 和 (III) 满足.

进一步地, $n^{-1}\hat{U}(\beta_0)$ 与下式有相同的极限

$$\hat{B}(\beta_0) = \frac{1}{n}\sum_{k=1}^K\sum_{i=1}^n\int_0^\tau \left[\frac{\hat{r}_{ik}^{(1)}(\beta_0, t)}{\hat{r}_{ik}(\beta_0, t)} - \frac{\hat{S}^{(1)}(\beta_0, t)}{\hat{S}^{(0)}(\beta_0, t)} \right] Y_{ik}(t) r_{ik}(\beta_0, t) \lambda_0(t) dt.$$

类似于 [13], 可以证明

$$\frac{1}{n}\sum_{i=1}^n \frac{\hat{r}_{ik}^{(1)}(\beta_0, t)}{\hat{r}_{ik}(\beta_0, t)} Y_{ik}(t) r_{ik}(\beta_0, t) \rightarrow_p s_k^{(1)}(\beta_0, t).$$

所以

$$\hat{B}(\beta_0) \rightarrow_p \int_0^\tau \left(\sum_{k=1}^K s_k^{(1)}(\beta_0, t) - s^{(1)}(\beta_0, t) \right) \lambda_0(t) dt = 0,$$

和 (IV) 成立.

对正态性的证明, 我们只需要证明 $n^{-1/2}\hat{U}(\beta_0)$ 的渐近正态性, 定义 $S_k^{(d)}(\beta, t)$ 为 $r_{ik}(\beta, t)$ 代替 $\hat{S}_k^{(d)}(\beta, t)$ 中的 $\hat{r}_{ik}(\beta, t)$ 而得到相应的函数.

由 (A2), 可以证明 (3.1) 的第一项在 β_0 点等于

$$n^{-1/2}\sum_{k=1}^K\sum_{i=1}^n\int_0^\tau \left[\frac{r_{ik}^{(1)}(\beta_0, t)}{r_{ik}(\beta_0, t)} - \frac{\sum_{q=1}^K s_q^{(1)}(\beta_0, t)}{\sum_{q=1}^K s_q^{(0)}(\beta_0, t)} \right] dM_{ik}(t) + o_p(1).$$

对 (3.1) 的第二项在 β_0 点的值, 我们对 $\hat{r}_{ik}^{(1)}/\hat{r}_{ik}$ 和 $\hat{S}^{(1)}/\hat{S}^{(0)}$ 分别在 $r_{ik}^{(1)}/r_{ik}$ 和 $S^{(1)}/S^{(0)}$ 应用一阶展开, $x/y = x_0/y_0 + (x - x_0)/y_0 - (y - y_0)x_0/y_0^2 + o\{(x - x_0)^2 + (y - y_0)^2\}$, 可以证明 (3.1) 的第二项在 β_0 处渐近等于:

$$\begin{aligned} & -n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i=1}^n \int_0^\tau \left[\frac{r_{ik}^{(1)}(\beta_0, t)}{r_{ik}(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right] Y_{ik}(t) (\hat{r}_{ik}(\beta_0, t) - r_{ik}(\beta_0, t)) \lambda_0(t) dt \\ &= -n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \bar{V}_k} \int_0^\tau \left[\left(\frac{\phi_{ik}^{(1)}(\beta_{10}, t)}{\phi_{ik}(\beta_{10}, t)} - \frac{s^{(11)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right) \right. \\ &\quad \times Y_{ik}(t) e^{\beta'_{20} Z_{ik}(t)} (\hat{\phi}_{ik}(\beta_{10}, t) - \phi_{ik}(\beta_{10}, t)) \lambda_0(t) dt \quad (A4) \\ &\quad \left. \times \begin{pmatrix} \Psi_k^{(1)}(t) \\ \Psi_k^{(2)}(t) \end{pmatrix} \right], \end{aligned} \quad (A4)$$

其中 $\Psi_k^{(1)}(t)$ 和 $\Psi_k^{(2)}(t)$ 是 (A4) 式的两项.

由条件 (C1)–(C5), 有

$$\begin{aligned} \sum_{k=1}^K \Psi_k^{(1)}(t) &= -\sum_{k=1}^K \int_0^\tau \sum_{m=1}^L \left(\frac{\phi_{amk}^{(1)}(\beta_{10}, t)}{\phi_{amk}(\beta_{10}, t)} - \frac{s^{(11)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right) \frac{1}{(\sum_{s=1}^K \rho_s) P(Y_{ik}(t) = 1, A_{ik}^* = a_m)} \\ &\quad \times \sqrt{n} \frac{1}{n} \sum_{l=1}^n \sum_{q=1}^K Y_{lq}(t) \eta_{lq} I(A_{lq}^* = a_m) [\exp\{\beta'_{10} E_{lq}(t)\} - \phi_{amk}(\beta_{10}, t)] \\ &\quad \times \frac{1}{n} \sum_{i \in \bar{V}_k} Y_{ik}(t) e^{\beta'_{20} Z_{ik}(t)} I(A_{ik}^*(t) = a_m) \lambda_0(t) dt + o_p(1) \\ &= -\sum_{k=1}^K \int_0^\tau \sum_{m=1}^L \left(\frac{\phi_{amk}^{(1)}(\beta_{10}, t)}{\phi_{amk}(\beta_{10}, t)} - \frac{s^{(11)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right) \frac{1 - \rho_k}{\sum_{s=1}^K \rho_s} E(e^{\beta'_{20} Z_{ik}(t)} | Y_{ik}(t) = 1, A_{ik}^*(t)) \\ &\quad \times (-n^{-1/2}) \sum_{q=1}^K \sum_{l \in V_q} Y_{lq}(t) I(A_{lq}^* = a_m) [\exp\{\beta'_{10} E_{lq}(t)\} - \phi_{amq}(\beta_{10}, t)] \lambda_0(t) dt + o_p(1) \\ &= -\rho n^{-1/2} \int_0^\tau \sum_{q=1}^K \sum_{l \in V_q} \sum_{m=1}^L \left(\frac{\phi_{amk}^{(1)}(\beta_{10}, t)}{\phi_{amk}(\beta_{10}, t)} - \frac{s^{(11)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right) E(e^{\beta'_{20} Z_{ik}(t)} | Y_{ik}(t) = 1, A_{ik}^*(t)) \\ &\quad \times Y_{lq}(t) I(A_{lq}^* = a_m) [\exp\{\beta'_{10} E_{lq}(t)\} - \phi_{amq}(\beta_{10}, t)] \lambda_0(t) dt + o_p(1) \\ &= -\rho n^{-1/2} \int_0^\tau \sum_{q=1}^K \sum_{l \in V_q} \left(\frac{\phi_{lq}^{(1)}(\beta_{10}, t)}{\phi_{lq}(\beta_{10}, t)} - \frac{s^{(11)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right) \\ &\quad \times Y_{lq}(t) [\exp\{\beta'_{10} E_{lq}(t)\} - \phi_{lq}(\beta_{10}, t)] E(e^{\beta'_{20} Z_{ik}(t)} | Y_{ik}(t) = 1, A_{ik}^*(t)) \lambda_0(t) dt + o_p(1) \\ &= -\rho n^{-1/2} \int_0^\tau \sum_{q=1}^K \sum_{l \in V_q} Q_{lq}(\beta_{10}, t) \lambda_0(t) dt + o_p(1), \end{aligned}$$

上式渐近等于 $-n^{-\frac{1}{2}} \sum_{q=1}^K \rho \sum_{l \in V_q} Q_{lq}(\beta_0)$, 其中 $Q_{lk}(\beta_0)$ 如定理 1 的定义. 类似地, $\sum_{k=1}^K \Psi_k^{(2)}(t)$ 渐近等于 $-n^{-\frac{1}{2}} \times \sum_{q=1}^K \rho \sum_{l \in V_q} H_{lq}(\beta_0)$, 其中 $H_{lk}(\beta_0)$ 如定理 1 的定义.

因此 $n^{-\frac{1}{2}} \hat{U}(\beta_0)$ 渐近等于 (3.2), 而 (3.2) 依分布收敛到零均值正态向量, 其方差为 $\Sigma_1(\beta_0) + \Sigma_2(\beta_0)$.

对 $\hat{U}(\beta_0)$ Taylor 展开, 我们有

$$n^{-\frac{1}{2}}\hat{U}(\beta_0) = \left\{ -n^{-1} \frac{\partial}{\partial \beta^T} \hat{U}(\beta^*) \right\} n^{\frac{1}{2}}(\hat{\beta}_E - \beta_0), \quad (\text{A5})$$

其中 β^* 介于 $\hat{\beta}_E$ 和 β_0 之间.

由 (III) 和 $n^{-\frac{1}{2}}\hat{U}(\beta_0)$ 的渐近正态性, 定理 1 得证. $\square$

定理 2 的证明 注意到

$$\begin{aligned} \sup_{t \in [0, \tau]} |\hat{\Lambda}_0(t) - \Lambda_0(t)| &\leq \sup_{t \in [0, \tau]} \left| \int_0^t \frac{\frac{1}{n} \sum_{k=1}^K \sum_{i=1}^n dM_{ik}(s)}{\sum_{k=1}^K \hat{S}_k^{(0)}(\hat{\beta}_E, s)} \right| \\ &\quad + \sup_{t \in [0, \tau]} \left| \int_0^t \frac{\sum_{k=1}^K S_k^{(0)}(\beta_0, s) - \sum_{k=1}^K \hat{S}_k^{(0)}(\hat{\beta}_E, s)}{\sum_{k=1}^K \hat{S}_k^{(0)}(\hat{\beta}_E, s)} \lambda_0(s) ds \right|. \end{aligned}$$

由 $\hat{\beta}_E$ 的相合性, 以及 (A2) 和鞅性质, 我们可以证明上面不等式的右边第一项收敛到 0, 由 (A1), 第二项可以忽略, 因此 $\hat{\Lambda}_0(t)$ 的一致相合性得证.

进一步,

$$\begin{aligned} \sqrt{n}(\hat{\Lambda}_0(t) - \Lambda_0(t)) &= \frac{1}{\sqrt{n}} \int_0^t \frac{\sum_{k=1}^K \sum_{i=1}^n dM_{ik}(s)}{\hat{S}^{(0)}(\hat{\beta}_E, s)} + \sqrt{n} \int_0^t \frac{\hat{S}^{(0)}(\beta_0, s) - \hat{S}^{(0)}(\hat{\beta}_E, s)}{\hat{S}^{(0)}(\hat{\beta}_E, s)} \lambda_0(s) ds \\ &\quad + \sqrt{n} \int_0^t \frac{S^{(0)}(\beta_0, s) - \hat{S}^{(0)}(\beta_0, s)}{\hat{S}^{(0)}(\hat{\beta}_E, s)} \lambda_0(s) ds. \end{aligned}$$

由 (A2), 第一项等于

$$n^{-1/2} \sum_{k=1}^K \sum_{i=1}^n \int_0^t \frac{dM_{ik}(s)}{s^{(0)}(\beta_0, s)} + o_p(1).$$

由 Taylor 展式和 (A1), 第二项等于

$$\left[\int_0^t \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)} \lambda_0(s) ds \right]' \sqrt{n}(\hat{\beta}_E - \beta_0) + o_p(1).$$

进一步由 (A5) 和 (III), 上式等于

$$\left[\int_0^t \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)} \lambda_0(s) ds \right]' \Sigma^{-1}(\beta_0) n^{-\frac{1}{2}} \hat{U}(\beta_0) + o_p(1).$$

类似于 $\sum_{k=1}^K \Psi_k^{(1)}(t)$ 近似的证明, 第三项渐近等于

$$-n^{-1/2} \sum_{k=1}^K \rho \sum_{i \in v_k} \int_0^t \frac{1}{s^{(0)}(\beta_0, s)} Y_{ik}(t) [e^{\beta'_{10} E_{ik}(t)} - \phi_{ik}(\beta_{10}, s)] \delta_{ik}^*(s) \lambda_0(s) ds.$$

因此

$$\begin{aligned} \sqrt{n}(\hat{\Lambda}_0(t) - \Lambda_0(t)) &= n^{-1/2} \sum_{k=1}^K \sum_{i=1}^n \int_0^t \frac{dM_{ik}(s)}{s^{(0)}(\beta_0, s)} - \left[\int_0^t \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)} \lambda_0(s) ds \right]' \Sigma^{-1}(\beta_0) n^{-\frac{1}{2}} \hat{U}(\beta_0) \\ &\quad - n^{-1/2} \sum_{k=1}^K \rho \sum_{i \in v_k} \int_0^t \frac{1}{s^{(0)}(\beta_0, s)} Y_{ik}(t) [e^{\beta'_{10} E_{ik}(t)} - \phi_{ik}(\beta_{10}, s)] \delta_{ik}^*(s) \lambda_0(s) ds. \end{aligned}$$

$$-n^{-1/2} \sum_{k=1}^K \rho \sum_{i \in v_k} \int_0^t \frac{1}{s^{(0)}(\beta_0, s)} Y_{ik}(t) [e^{\beta'_{10} E_{ik}(t)} - \phi_{ik}(\beta_{10}, s)] \delta_{ik}^*(s) \lambda_0(s) ds + o_p(1).$$

可以将上式的重新写成两个独立和如下,

$$\sqrt{n}(\hat{\Lambda}_0(t) - \Lambda_0(t)) = n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \bar{V}_k} u_{ik}(t, \beta_0) + n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{j \in V_k} v_{jk}(t, \beta_0) + o_p(1),$$

其中 $u_{ik}(t, \beta_0)$ 和 $v_{jk}(t, \beta_0)$ 如定理 2 的定义. 易证 $E(u_{ik}(t, \beta_0)) = 0$, $E(v_{jk}(t, \beta_0)) = 0$. 由多元中心极限定理, 定理 2 得证. $\square$

Estimated standard partial likelihood method for multivariate failure time data with auxiliary information

YANG QingLong, MA XueMin & YUAN ZhongShang

Abstract In some practical studies, the main covariate is only precisely measured in a validation set due to financial limitations, while as an auxiliary information for the main covariate are collected for the full cohort. Taking use of these auxiliary information will increase the study efficiency. In this paper, based on the framework of marginal proportional hazards model, we develop some inference procedures to analyze such multivariate failure time data. We propose an estimated standard partial likelihood equation for the regression parameters and a Breslow-type estimator for cumulative hazard function. The resultant estimators are shown to be consistent and asymptotically normally distributed. Simulation studies are conducted to evaluate the performance of the proposed estimators.

Keywords validation set, auxiliary covariates, marginal hazard model, multivariate failure time

MSC(2010) 62G05, 62N01

doi: 10.1360/012011-791