

相依删失下基于连接函数的参数模型的统计分析 *

邓文丽^{*} 吴子星

蔡 明

(江西师范大学数学与信息科学学院, 南昌, 330022) (江西师范大学图书馆, 南昌, 330022)

摘要: 在临床数据的收集中, 由于竞争性风险或者病人的退出可能导致数据删失. 删失数据的统计分析大多是基于独立删失的假定进行的. 而实际情况中, 数据的删失往往是非独立的, 即删失变量和失效时间变量是相关的. 相依删失使得原本复杂的删失数据处理变得更加困难. 在本文中, 假定删失变量和失效时间变量的联合分布可以用它们边际分布的连接函数表示, 在给定连接函数下, 得到了比例风险模型的极大似然估计. 模拟计算显示, 如果删失假定成立, 本文所采用方法比独立删失假定下的估计方法更准确.

关键词: 右删失数据; 比例风险模型; 连接函数

中图分类号: O212.1

英文引用格式: DENG W L, WU Z X, CAI M. The statistical analysis of parameter model with an assumed copula for dependent censoring data [J]. Chinese J Appl Probab Statist, 2018, 34(5): 492-500. (in Chinese)

§1. 引言

在生存分析和可靠性分析中, 寿命数据的收集中常常会出现删失的情况. 数据删失增加了统计分析的难度, 在独立删失的假定下, 相关的统计研究工作已经取得了丰富的成果. 在很多实际情况中, 由于竞争风险相依或者病人的退出和后来的死亡时间有关联等原因, 独立删失的假设并不成立. 相依删失使得原本复杂的删失数据处理变得更加困难. Tsiatis^[1] 在竞争风险模型下, 研究了相依结构模型的不可识别性. Williams 和 Lagakos^[2] 以指数分布为例, 说明了采用独立删失的统计分析方法处理相依删失数据, 可能会得到有偏或低效的统计结论, 并证明了, 对相依的右删失数据, 如果不做相依结构的假定, 相依性结构模型是不可估计的. 假定删失变量和失效时间变量的联合分布可以用它们边际分布的连接函数 (Copula) 表示出来, 既可以解决可估性问题, 也可以充分利用连接函数的优良性质, 因此用连接函数对相依结构进行假定, 这一做法被很多统计学者采纳. 关于连接函数的详细介绍, 可以参考文献 [3].

Zheng 和 Klein^[4] 利用连接函数对失效时间变量和删失时间变量的联合分布进行假定, 得到了失效时间变量生存函数的 Copula-graph 估计, 并证明了把相依删失退化为独立删

*国家自然科学基金项目 (批准号: 71001046、11171112、11101114、11201190) 资助.

*通讯作者, E-mail: wldfudan@126.com.

本文 2017 年 1 月 3 日收到, 2017 年 11 月 8 日收到修改稿.

失时, 所得的估计就是右删失数据的乘积极限估计. 在此基础上, Rivest 和 Wells^[5] 选用了阿基米德连接函数来刻画相依性, 得到了上述 Copula-graph 估计的解析解, 并证明了估计的一致相合性和渐近正态性. Wang^[6] 在阿基米德连接函数的假定下, 对相依结构的可估性问题进行了研究.

Siannis^[7], Siannis 等^[8], Zhang 和 Heitjan^[9] 对参数生存模型进行了敏感性分析. Park 等^[10] 在没有协变量和非参数的设定下提出了敏感性分析的方法. Huang 和 Zhang^[11] 在二元比例风险模型假定下研究了阿基米德连接函数对偏似然估计结果的影响. 文中分别针对失效可能性和删失可能性列出了似然函数, 并假定这两个似然函数之间是相互独立的, 得到了一个联合似然函数. 本文拟沿用独立删失假定下比例风险模型似然函数的构造思想, 在连接函数假定下, 对二元比例风险模型进行统计推断.

文章的第二节主要是问题介绍和模型假定. 第三节建立了似然函数的表达式, 以及对回归参数和基准风险函数进行估计的思路和推导公式, 求似然估计的迭代步骤. 第四节通过模拟计算验证了本文所述方法的可行性. 从模拟计算结果可以看出, 相依删失的假定成立时, 本文所述的方法可以比独立删失方法得到更准确的估计结果.

§2. 主要问题和模型假定

在生存分析研究中, 收集到一批独立样本 $(T_i, \delta_i, Z_i, W_i)$, $i = 1, 2, \dots, n$, 其中 $T_i = \min(X_i, C_i)$, X_i 是失效时间变量, C_i 是删失时间变量, $\delta_i = I(X_i \leq C_i)$, $i = 1, 2, \dots, n$. Z_i, W_i , $i = 1, 2, \dots, n$ 分别是对失效时间变量和删失时间变量有影响的 p 维和 q 维协变量, 它们可能相同, 也可能部分相同或完全不同.

假定协变量对失效时间变量 X_i 和删失时间变量 C_i 的危险率函数的影响均符合比例风险模型:

$$\lambda(t | Z_i) = \lambda_0(t) \exp(Z'_i \beta), \quad \psi(t | W_i) = \psi_0(t) \exp(W'_i \gamma), \quad i = 1, 2, \dots, n,$$

其中 β, γ 分别是 p 维和 q 维未知参数, $\lambda_0(\cdot), \psi_0(\cdot)$ 分别为未知的基准危险率函数. X_i 和 C_i 的边际分布函数分别记为 $F_i(\cdot), G_i(\cdot)$, $i = 1, 2, \dots, n$,

$$F_i(x) = 1 - \exp[-\Lambda_0(x) \exp(Z'_i \beta)], \quad G_i(c) = 1 - \exp[-\Psi_0(c) \exp(W'_i \gamma)], \quad i = 1, 2, \dots, n,$$

其中 $\Lambda_0(x) = \int_0^x \lambda_0(t) dt$, $\Psi_0(c) = \int_0^c \psi_0(t) dt$.

根据 Sklar 定理^[3], 存在一个连接函数 $H(u, v)$, 满足 $H(u, 0) = H(0, v) = 0$, $H(u, 1) = u$ 和 $H(1, v) = v$, 使得 X_i 和 C_i 的联合分布可以用它们边际分布的连接函数表示:

$$J_i(x, c) = \mathbb{P}(X_i \leq x, C_i \leq c) = H(F(x, \beta), G(c, \gamma)),$$

如果 F_i 和 G_i 连续, 那么 H 是唯一的. 在这里假定连接函数 $H(\cdot, \cdot)$ 已知. 类似地, X_i 和 C_i 的联合生存函数可表示为

$$S_i(x, c) = \mathbb{P}(X_i > x, C_i > c) = 1 - F_i(x) - G_i(c) + H(F_i(x), G_i(c)).$$

基于以上观测数据和模型假定, 本文将对比例风险模型中的基准危险率函数和回归参数进行估计.

§3. 极大似然估计

为表述方便, 假设样本已经过排序, 使得 $t_1 \leq t_2 \leq \cdots \leq t_n$, 基于观测样本, 可以构造似然函数

$$L = \prod_{i=1}^n [\mathbb{P}(T_i = t_i, \delta_i = 0)]^{1-\delta_i} [\mathbb{P}(T_i = t_i, \delta_i = 1)]^{\delta_i}.$$

在相依删失下, 第 i 个个体在时刻 t 瞬间被删失的概率可表示为

$$\begin{aligned} \mathbb{P}(T_i = t, \delta_i = 0) &= -\frac{\partial S_i(t, v)}{\partial v} \Big|_{v=t} \\ &= G'_i(t) - H_v(F_i(t), G_i(t))G'_i(t), \quad i = 1, 2, \dots, n, \end{aligned}$$

其中 $H_v(u, v) = \partial H(u, v)/\partial v$.

第 i 个个体在时刻 t 瞬间发生失效的概率可表示为

$$\begin{aligned} \mathbb{P}(T_i = t, \delta_i = 1) &= -\frac{\partial S_i(u, t)}{\partial u} \Big|_{u=t} \\ &= F'_i(t) - H_u(F_i(t), G_i(t))F'_i(t), \quad i = 1, 2, \dots, n, \end{aligned}$$

其中 $H_u(u, v) = \partial H(u, v)/\partial u$.

样本的似然函数可以表示为

$$L = \prod_{i=1}^n [G'_i(t_i) - H_v(F_i(t_i), G_i(t_i))G'_i(t_i)]^{1-\delta_i} [F'_i(t_i) - H_u(F_i(t_i), G_i(t_i))F'_i(t_i)]^{\delta_i}.$$

把 $F_i(\cdot)$ 和 $G_i(\cdot)$ 的表达式代入似然函数表达式, 可得

$$\begin{aligned} L &= \prod_{i=1}^n \left\{ \left\{ \psi_0(t_i) e^{W'_i \gamma} \exp[-\Psi_0(t_i) e^{W'_i \gamma}] \right. \right. \\ &\quad \times \left\{ 1 - H_v(1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]) \right\} \right\}^{1-\delta_i} \\ &\quad \times \left\{ \lambda_0(t_i) e^{Z'_i \beta} \exp[-\Lambda_0(t_i) e^{Z'_i \beta}] \right. \\ &\quad \times \left. \left. \left\{ 1 - H_u(1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]) \right\} \right\} \right\}^{\delta_i}. \end{aligned}$$

记 $\lambda_i = \lambda_0(t_i)$, $\psi_i = \psi_0(t_i)$, 那么

$$\Lambda_0(t_i) = \sum_{t_j \leq t_i} \lambda_j, \quad \Psi_0(t_i) = \sum_{t_j \leq t_i} \psi_j,$$

似然函数可以转化为

$$\begin{aligned} L = \prod_{i=1}^n & \left\{ \left\{ \psi_i e^{W'_i \gamma} \exp[-\Psi_0(t_i) e^{W'_i \gamma}] \right. \right. \\ & \times \left\{ 1 - H_v(1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]) \right\}^{1-\delta_i} \\ & \times \left\{ \lambda_0(t_i) e^{Z'_i \beta} \exp[-\Lambda_0(t_i) e^{Z'_i \beta}] \right. \\ & \left. \left. \times \left\{ 1 - H_u(1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]) \right\}^{\delta_i} \right\}. \right. \end{aligned}$$

当 $\delta_k = 0$, 即第 k 个观测点为删失点, 分析 λ_k 对似然函数的影响. 在似然函数中, 当 $i = k$ 时, 与 λ_k 有关的项为

$$I_1 = 1 - H_v(1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]),$$

$\Lambda_0(t_k) = \sum_{t_j \leq t_k} \lambda_j$ 关于 λ_k 递增, 因此 I_1 关于 λ_k 递减.

在似然函数中, 当 $i \neq k$ 时, 与 λ_k 有关的项为

$$\begin{aligned} I_2 = \prod_{i \neq k}^n & \left\{ \left\{ 1 - H_u(1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]) \right\}^{1-\delta_i} \right. \\ & \left. \times \left\{ \exp[-\Lambda_0(t_i) e^{Z'_i \beta}] \left\{ 1 - H_v(1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]) \right\} \right\}^{\delta_i} \right\}. \end{aligned}$$

I_2 关于 λ_k 递减.

因此当第 k 个观测点为删失点时, 取 $\lambda_k = 0$ 可使似然函数达到最大. 类似地, 当 $\delta_k = 1$, 即第 k 个观测点为失效点时, 取 $\psi_k = 0$ 可使似然函数达到最大.

下面讨论当 $\delta_k = 1$ 时 λ_k 的极大似然估计, 以及 $\delta_k = 0$ 时 ψ_k 的极大似然估计. 显然, 似然函数 L 是关于 $\lambda_k, \psi_k, k = 1, 2, \dots, n$ 的有界连续函数, 定义域为有界闭区域, 在偏导数存在的条件下, 极值点在驻点取到.

为表述方便, 记

$$\Lambda_0(t_i) = \sum_{t_j \leq t_i, j \in D} \lambda_j, \quad \Psi_0(t_i) = \sum_{t_j \leq t_i, j \in C} \psi_j,$$

其中 D 为失效点下标集, C 为删失点下标集. 对数似然函数可表示为

$$\begin{aligned} \ln L = \sum_{i=1}^n (1 - \delta_i) & \left\{ \ln \psi_i + W'_i \gamma - \Psi_0(t_i) e^{W'_i \gamma} \right. \\ & \left. + \ln \left\{ 1 - H_v(1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]) \right\} \right\} \end{aligned}$$

$$\begin{aligned}
& + \sum_{i=1}^n \delta_i \left\{ \ln \lambda_i + Z'_i \beta - \Lambda_0(t_i) e^{Z'_i \beta} \right. \\
& \quad \left. + \ln \left\{ 1 - H_u (1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]) \right\} \right\} \\
& = \sum_{i \in C} \left\{ \ln \psi_i - \Psi_0(t_i) e^{W'_i \gamma} \right. \\
& \quad \left. + \ln \left\{ 1 - H_v (1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]) \right\} \right\} \\
& \quad + \sum_{i \in D} \left\{ \ln \lambda_i - \Lambda_0(t_i) e^{Z'_i \beta} \right. \\
& \quad \left. + \ln \left\{ 1 - H_u (1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}]) \right\} \right\}.
\end{aligned}$$

对任意的 $k \in D$, λ_k 的极大似然估计满足方程

$$\frac{1}{\lambda_k} = \sum_{i \in D, i \geq k} e^{Z'_i \beta} \left\{ 1 + \frac{H_{uu} \exp[-\Lambda_0(t_i) e^{Z'_i \beta}]}{1 - H_u} \right\} + \sum_{i \in C, i \geq k} \frac{H_{vu} \exp[-\Lambda_0(t_i) e^{Z'_i \beta}] e^{Z'_i \beta}}{1 - H_v}, \quad (1)$$

类似地, 可以得到对任意的 $k \in C$, ψ_k 的极大似然估计满足方程

$$\frac{1}{\psi_k} = \sum_{i \in C, i \geq k} e^{W'_i \gamma} \left\{ 1 + \frac{H_{vv} \exp[-\Psi_0(t_i) e^{W'_i \gamma}]}{1 - H_v} \right\} + \sum_{i \in D, i \geq k} \frac{H_{uv} \exp[-\Psi_0(t_i) e^{W'_i \gamma}] e^{W'_i \gamma}}{1 - H_u}, \quad (2)$$

其中

$$\begin{aligned}
H_u &\hat{=} \frac{\partial H(u, v)}{\partial u} \Big|_{u=u_0, v=v_0}, & H_v &\hat{=} \frac{\partial H(u, v)}{\partial v} \Big|_{u=u_0, v=v_0}, \\
H_{uu} &\hat{=} \frac{\partial^2 H(u, v)}{\partial u^2} \Big|_{u=u_0, v=v_0}, & H_{vu} &\hat{=} \frac{\partial^2 H(u, v)}{\partial u \partial v} \Big|_{u=u_0, v=v_0}, \\
H_{uv} &\hat{=} \frac{\partial^2 H(u, v)}{\partial v \partial u} \Big|_{u=u_0, v=v_0}, & H_{vv} &\hat{=} \frac{\partial^2 H(u, v)}{\partial v^2} \Big|_{u=u_0, v=v_0}, \\
u_0 &= 1 - \exp[-\Lambda_0(t_i) e^{Z'_i \beta}], & v_0 &= 1 - \exp[-\Psi_0(t_i) e^{W'_i \gamma}].
\end{aligned}$$

在独立删失情况下, $H(u, v) = uv$, 显然 $H_{uu} = 0$, $H_{vu} = 1$, $H_v = u$, $H_u = v$, 方程 (1) 简化为

$$\begin{aligned}
\frac{1}{\lambda_k} &= \sum_{i \in D, i \geq k} e^{Z'_i \beta} + \sum_{i \in C, i \geq k} e^{Z'_i \beta}, \\
\hat{\lambda}_k &= 1 / \sum_{i \geq k} e^{Z'_i \beta}.
\end{aligned} \quad (3)$$

这个表达式对第 k 个观测点为失效点成立; 如果第 k 个观测点为删失点, $\hat{\lambda}_k = 0$.

方程 (2) 简化为

$$\begin{aligned}
\frac{1}{\psi_k} &= \sum_{i \in C, i \geq k} e^{W'_i \gamma} + \sum_{i \in D, i \geq k} e^{W'_i \gamma}, \\
\hat{\psi}_k &= 1 / \sum_{i \geq k} e^{W'_i \gamma}.
\end{aligned} \quad (4)$$

这个表达式对第 k 个观测点为删失点成立; 如果第 k 个观测点为失效点, $\hat{\psi}_k = 0$.

为了表述方便, 用 $t_{(1)} < t_{(2)} < \dots < t_{(m)}$ 表示 $t_1 \leq t_2 \leq \dots \leq t_n$ 中的互不相同的元素, 假设 $t_{(i)}$ 结的大小记为 n_i , 显然 $\sum_{i=1}^m n_i = n$, 用 $Z_{(i)j}$, $j = 1, 2, \dots, n_i$, $W_{(i)j}$, $j = 1, 2, \dots, n_i$ 表示观测值为 $t_{(i)}$ 的个体的协变量, 独立删失下回归系数 β 和 γ 的估计可以由下面的表达式得到:

$$\hat{\beta} = \arg \max \prod_{i=1}^m \left[\exp \left(\sum_{j=1}^{n_i} Z'_{(i)j} \beta \right) / \sum_{k \geq i} \exp \left(\sum_{j=1}^{n_k} Z'_{(k)j} \beta \right) \right], \quad (5)$$

$$\hat{\gamma} = \arg \max \prod_{i=1}^m \left[\exp \left(\sum_{j=1}^{n_i} W'_{(i)j} \gamma \right) / \sum_{k \geq i} \exp \left(\sum_{j=1}^{n_k} W'_{(k)j} \gamma \right) \right]. \quad (6)$$

详细的推导过程可参考文献 [12].

下面在 (1) 和 (2) 的基础上, 对相依删失情况求 β 和 γ 的极大似然估计. 对数似然函数关于 β 和 γ 求偏导数, 得到似然方程

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta} = & \sum_{i=1}^n \left\{ \delta_i [1 - \Lambda_0(t_i) e^{Z'_i \beta}] \right. \\ & \left. - \left[\frac{\delta_i H_{uu}}{1 - H_u} + \frac{(1 - \delta_i) H_{vu}}{1 - H_v} \right] \exp[-\Lambda_0(t_i) e^{Z'_i \beta}] \Lambda_0(t_i) e^{Z'_i \beta} \right\} Z_i = 0, \end{aligned} \quad (7)$$

$$\begin{aligned} \frac{\partial \ln L}{\partial \gamma} = & \sum_{i=1}^n \left\{ (1 - \delta_i) [1 - \Psi_0(t_i) e^{W'_i \gamma}] \right. \\ & \left. - \left[\frac{(1 - \delta_i) H_{vv}}{1 - H_v} + \frac{\delta_i H_{uv}}{1 - H_u} \right] \exp[-\Psi_0(t_i) e^{W'_i \gamma}] \Psi_0(t_i) e^{W'_i \gamma} \right\} W_i = 0. \end{aligned} \quad (8)$$

下面采用迭代的方法求 $\{\lambda_i\}_{i=1}^n$, $\{\psi_i\}_{i=1}^n$, β , γ 的极大似然估计的近似解.

第一步, 在独立删失假定下, 对两个比例风险模型, 利用 (3), (4), (5) 和 (6) 求出 $\Lambda_0(\cdot)$, $\Psi_0(\cdot)$, β , γ 的初始值, 分别记为 $\hat{\Lambda}_0^{(0)}(\cdot)$, $\hat{\Psi}_0^{(0)}(\cdot)$, $\hat{\beta}^{(0)}$, $\hat{\gamma}^{(0)}$, 令 $m = 0$.

第二步, 把 $\hat{\Lambda}_0^{(m)}(\cdot)$, $\hat{\Psi}_0^{(m)}(\cdot)$ 代入到方程 (7) 和 (8) 中, 得到的 β 和 γ 估计值, 分别记为 $\hat{\beta}^{(m+1)}$, $\hat{\gamma}^{(m+1)}$.

第三步, 把 $\hat{\beta}^{(m+1)}$, $\hat{\gamma}^{(m+1)}$ 代入方程 (1) 和 (2), 得到 $\{\lambda_i\}_{i=1}^n$ 和 $\{\psi_i\}_{i=1}^n$ 的估计值, 从而得到 $\Lambda_0(\cdot)$ 和 $\Psi_0(\cdot)$ 的第 $m + 1$ 步估计值 $\hat{\Lambda}_0^{(m+1)}(\cdot)$, $\hat{\Psi}_0^{(m+1)}(\cdot)$.

第四步, 令 $m = m + 1$, 返回到第二步, 直到收敛.

算法收敛后, 得到估计值 $\hat{\Lambda}_0(\cdot)$, $\hat{\Psi}_0(\cdot)$, $\hat{\beta}$, $\hat{\gamma}$.

上述估计就是 Efron^[13] 所定义的自相合估计. Tsai 和 Crowley^[14] 证明了上述迭代算法的收敛性, 并证明了迭代算法得到的自相合估计就是广义极大似然估计. 广义似然估计的相合性可参见文献 [15].

§4. 模拟计算

在模拟计算中采用了下面的模型, 对 T 和 C 的边际分布作如下假定, 假设它们的失效

率函数分别为

$$\lambda(t) = 0.5t \exp(-0.5Z_1 + 0.1Z_2), \quad \psi(t) = 0.2 \exp(0.3Z_1 + 0.2Z_2),$$

其中 $Z_1 \sim b(1, 0.5)$ 和 $Z_2 \sim U(-10, 10)$.

基于连接函数对相依删失数据进行模型假定, 比较多的研究工作主要集中在连接函数的一个子类, 阿基米德连接函数类. 在以往的研究中发现, 在阿基米德连接函数类中, 不同的连接函数假定下, 估计量还是比较稳健的. 因此在本文的模拟计算中, 选用了 Frank copula 函数来对 T 和 C 的联合分布进行假定. 这个连接函数是 Frank^[16] 提出的, 函数表达式为

$$H(u, v; \alpha) = \ln \left[1 + \frac{(\alpha^u - 1)(\alpha^v - 1)}{\alpha - 1} \right], \quad \alpha > 0, \alpha \neq 1,$$

对于 Frank 连接函数, kendall - τ 相关系数可以用参数 α 表示.

$$\tau = 1 + 4\gamma^{-1}[D_1(\gamma) - 1], \quad \gamma = -\ln \alpha, \quad D_1(\gamma) = \gamma^{-1} \int_0^\gamma t/(e^t - 1) dt.$$

另外, 在模拟计算中还考虑了一个独立删失变量 $A \sim U(0, 10)$.

模拟计算进行了 500 次, 并利用本文所提出的估计方法对模型参数进行了估计. 表 1 中 T 前面的百分数表示所有观测样本中发生失效的样本点所占的比重, 如 48.5% T 表示样本中失效时间点所占的比例为 46.3%. 类似地, C 前面的百分数表示所有观测样本中发生删失的样本点所占的比重, 如 35.2% C 表示样本中删失时间点所占的比例为 36.4%.

表 1 不同样本大小下的参数估计 (Frank copula, $\tau = 0.8$)

真实值	估计值	偏差	标准差	估计值	偏差	标准差
数据的产生: $n = 100, 46.3\%T, 34.4\%C$						
估计方法: Frank copula, $\tau = 0.8$ 估计方法: 独立删失模型						
$\hat{\beta}_1$	-0.5	-0.478	0.022	0.658	-0.701	0.201
$\hat{\beta}_2$	0.1	0.155	0.055	0.460	0.189	0.089
$\hat{\gamma}_1$	0.3	0.375	0.075	0.254	0.384	0.084
$\hat{\gamma}_2$	0.2	0.191	0.009	0.242	0.175	0.025
数据的产生: $n = 300, 43.3\%T, 35.9\%C$						
估计方法: Frank copula, $\tau = 0.8$ 估计方法: 独立删失模型						
$\hat{\beta}_1$	-0.5	-0.518	0.018	0.556	-0.632	-0.132
$\hat{\beta}_2$	0.1	0.130	0.030	0.336	0.168	0.068
$\hat{\gamma}_1$	0.3	0.269	0.031	0.223	0.362	0.065
$\hat{\gamma}_2$	0.2	0.190	0.010	0.154	0.231	0.031

在表 1 中, 用 Frank copula 构造 T 和 C 的联合分布函数, $\tau = 0.8$, 分别选取了样本大小 $n = 100$ 和 $n = 300$ 下的计算结果进行比较. 所有模型参数的估计值都很接近真实值.

随着样本大小的增加, 估计量的偏差值和标准差都明显减小。模拟计算的结果同时也显示出, 当总体满足相依删失模型时, 本文所用的方法能够得到比独立删失假定下更精确的估计量。

参 考 文 献

- [1] TSIATIS A. A nonidentifiability aspect of the problem of competing risks [J]. *Proc Nat Acad Sci USA*, 1975, **72**(1): 20–22.
- [2] WILLIAMS J S, LAGAKOS S W. Models for censored survival analysis: constant-sum and variable-sum models [J]. *Biometrika*, 1977, **64**(2): 215–224.
- [3] NELSEN R B. *An Introduction to Copulas* [M]. New York: Springer, 1999.
- [4] ZHENG M, KLEIN J P. Estimates of marginal survival for dependent competing risks based on an assumed copula [J]. *Biometrika*, 1995, **82**(1): 127–138.
- [5] RIVEST L P, WELLS M T. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring [J]. *J Multivariate Anal*, 2001, **79**(1): 138–155.
- [6] WANG A T. Properties of the marginal survival functions for dependent censored data under an assumed Archimedean copulaa [J]. *J Multivariate Anal*, 2014, **129**: 57–68.
- [7] SIANNIS F. Applications of a parametric model for informative censoring [J]. *Biometrics*, 2004, **60**(3): 704–714.
- [8] SIANNIS F, COPAS J, LU G B. Sensitivity analysis for informative censoring in parametric survival models [J]. *Biostatistics*, 2005, **6**(1): 77–91.
- [9] ZHANG J M, HEITJAN D F. A simple local sensitivity analysis tool for nonignorable coarsening: application to dependent censoring [J]. *Biometrics*, 2006, **62**(4): 1260–1268.
- [10] PARK Y, TIAN L, WEI L J. One- and two-sample nonparametric inference procedures in the presence of a mixture of independent and dependent censoring [J]. *Biostatistics*, 2006, **7**(2): 252–267.
- [11] HUANG X L, ZHANG N. Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach [J]. *Biometrics*, 2008, **64**(4): 1090–1099.
- [12] KLEIN J P, MOESCHBERGER M L. *Survival Analysis: Techniques for Censored and Truncated Data* [M]. 2nd ed. New York: Springer-Verlag, 2003.
- [13] EFRON B. The two sample problem with censored data [C] // LE CAM L M, NEYMAN J. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Biology and Problems of Health*, Berkeley, CA: University of California Press, 1967: 831–853.
- [14] TSAI W Y, CROWLEY J. A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency [J]. *Ann Statist*, 1985, **13**(4): 1317–1334.
- [15] KIEFER J, WOLFOWITZ J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters [J]. *Ann Math Statist*, 1956, **27**(4): 887–906.
- [16] FRANK M J. On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$ [J]. *Aequationes Math*, 1979, **19**(1): 194–226.

The Statistical Analysis of Parameter Model with an Assumed Copula for Dependent Censoring Data

DENG Wenli WU Zixing

(School of Mathematics and Information Science, Jiangxi Normal University, Nanchang,
330022, China)

CAI Ming

(Library, Jiangxi Normal University, Nanchang, 330022, China)

Abstract: In collecting clinical data, data would be censored due to competing risks or patient withdrawal. The statistical inference for censoring data is always based on the assumption that the failure time and censoring time is independent. But in practice the failure time and censoring time are often dependent. Dependent censoring make the job to deal with censoring data more complicated. In this paper, we assume that the joint distribution of the failure time variable and censoring time variable is a function of their marginal distributions. This function is called a copula. Under prespecified copulas, the maximum likelihood estimators for cox proportional hazards models are worked out. Statistical analysis results are carried by simulations. When dependent censoring happens, the proposed method will do better than the traditional method used in independent situations. Simulation results show that the proposed method can get efficient estimations.

Keywords: right-censored data; cox proportion hazard model; copula

2010 Mathematics Subject Classification: 62N02; 62N01; 62F12