

半变系数伽马脆弱模型惩罚部分似然估计

张中文^{1,2}, 王晓光^{*1}, 宋立新¹

(1. 大连理工大学 数学科学学院, 辽宁 大连 116024;
2. 滨州医学院 公共卫生与管理学院, 山东 烟台 264003)

摘要: 为了更好地分析对数风险函数与协变量之间复杂的非线性关系, 提出一种半变系数伽马脆弱模型并给出其估计方法. 首先, 应用 B-样条将半变系数伽马脆弱模型近似转化为线性伽马脆弱模型, 然后运用惩罚部分似然法估计转化后模型的线性参数, 随后采用近似轮廓似然法并运用黄金搜索算法估计随机效应的参数; 在通过迭代获得转化后的线性系数以及随机效应参数的估计以后, 运用 B-样条得到变系数函数的估计. 经蒙特卡罗模拟研究发现, 该方法可以给出协变量的线性参数以及变系数函数较为精准、稳定的估计, 是分析协变量对于风险率影响的有效方法. 最后, 应用所提出的方法分析了 NCCTG 肺癌数据.

关键词: 伽马脆弱模型; B-样条; 变系数模型; 惩罚部分似然估计; 黄金搜索算法
中图分类号: O212 **文献标识码:** A **doi:** 10.7511/dllgxb201806015

0 引言

脆弱模型是比例风险模型的推广, 是考虑随机效应的生存分析模型. 其中, 随机效应(即脆弱)一般用于描述对应于不同分类(例如个体或家庭)的额外风险或者脆弱, 其基本思想是不同的个体具有不同的脆弱, 相对比较脆弱的个体与其他个体相比更容易发病或死亡. 近年来, 脆弱模型被广泛应用于研究对象之间存在不可观测的组间异质性的非独立生存时间问题的研究; 同时, 多种脆弱模型以及拟合这些模型的数值技术被广泛研究, 例如: 为了增加模型的灵活程度、扩大模型的应用范围, Du 等提出了一种非参数带脆弱项的危险率模型^[1]; Yu 等将多元对数正态脆弱模型推广到可加半参数情形用以描述协变量对于对数危险率的非线性影响, 并提出了一种双惩罚部分似然法用于模型的估计^[2], 该模型在增加了模型适应性的同时, 也避免了多元非参数函数的估计问题.

变系数模型是近年来发展起来的具有广泛应

用背景的回归模型, 该模型通过假设回归系数是其他协变量的未知函数而增加模型的灵活性, 因为系数函数通常被假设为某个协变量的一元函数, 所以维数灾难问题得到了有效避免. 如 Zhang 等研究了半变系数多元脆弱模型的估计问题, 用以描述某些协变量对于危险率的影响受其他协变量的影响, 并通过数值模拟和实例分析说明了方法的有效性, 其中脆弱的分布假定为对数多元正态分布^[3]. 而在实际的脆弱模型应用过程中, 假定脆弱服从伽马分布更为常见, 这是因为伽马分布的变量为正数, 十分适合脆弱分布无符号改变的特性; 伽马分布可以通过 Laplace 变换获得导数, 从而使得整个模型求导具备相对的简便性. 本文提出一种半变系数伽马脆弱模型, 以进一步丰富脆弱模型的模型结构, 用以描述聚集生存数据或者复发型生存数据分析中协变量效应受其他协变量的影响, 从而为分析生存时间与协变量更准确、更复杂的关系提供方法学支持.

收稿日期: 2018-08-15; 修回日期: 2018-09-25.

基金项目: 国家自然科学基金资助项目(11471065, 11371077, 81502891); 全国统计科学研究项目(2018LY57); 山东省统计科研重点课题资助项目(KT16244); 山东省医药卫生科技发展计划资助项目(2016WS0009); 山东省软科学研究计划项目(2018RKB14103).

作者简介: 张中文(1983-), 男, 博士生, E-mail: zhangzhongwen994@163.com; 王晓光*(1977-), 男, 副教授, 博士生导师, E-mail: wangxg@dlut.edu.cn.

1 半变系数伽马脆弱模型

设 T_{ij} 为表示第 i 个聚类中第 j 个个体的随机变量, C_{ij} 表示删失时间, 则观测时间 $Y_{ij} = \min(T_{ij}, C_{ij})$, 其中 $i = 1, 2, \dots, s, j = 1, 2, \dots, n_i$, 令 $\delta_{ij} = I[T_{ij} \leq C_{ij}]$, 第 i 个聚类中观测到结局事件的个数 $d_i = \sum_{j=1}^{n_i} \delta_{ij}$. 进一步假设 $\mathbf{x}_{ij} = (x_{1,ij} \ x_{2,ij} \ \dots \ x_{p_1,ij})^T$ 以及 $\mathbf{w}_{ij} = (w_{1,ij} \ w_{2,ij} \ \dots \ w_{p_2,ij})^T$ 表示协变量向量, 函数系数 $\boldsymbol{\beta}(u) = (\beta_1(u) \ \beta_2(u) \ \dots \ \beta_{p_1}(u))^T$ 是关于协变量 u 的 p_1 维二次可微未知函数向量, $\boldsymbol{\alpha}$ 为 p_2 维的协变量线性回归系数向量. 则半变系数伽马脆弱模型定义为

$$\lambda_{ij}(t; \mathbf{x}_{ij}, u_{ij}, \mathbf{w}_{ij}, \nu_i) = \lambda_0(t) \nu_i \exp(\boldsymbol{\beta}^T(u_{ij}) \mathbf{x}_{ij} + \boldsymbol{\alpha}^T \mathbf{w}_{ij}) \quad (1)$$

式中: $\lambda_0(t)$ 是基准危险率函数; 同时 $\nu_i (i = 1, 2, \dots, s)$ 表示第 i 个聚类中的脆弱, 并且服从单参数伽马分布, 其概率密度函数为

$$f(\nu) = \frac{\nu^{1/\theta-1} \exp(-\nu/\theta)}{\theta^{1/\theta} \Gamma(1/\theta)} \quad (2)$$

若令 $r_i = \log \nu_i$, 可称 r_i 为随机效应, 则半变系数伽马脆弱模型亦可改写为

$$\lambda_{ij}(t; \mathbf{x}_{ij}, u_{ij}, \mathbf{w}_{ij}, r_i) = \lambda_0(t) \exp(\boldsymbol{\beta}^T(u_{ij}) \mathbf{x}_{ij} + \boldsymbol{\alpha}^T \mathbf{w}_{ij} + r_i) \quad (3)$$

假设生存时间 T_{ij} 与删失时间 C_{ij} 关于协变量、随机效应 r_i 条件独立, 不同个体的生存时间关于随机效应条件独立, 同时假设随机效应与删失时间相互独立.

2 模型估计方法

2.1 B-样条

变系数函数向量 $\boldsymbol{\beta}(u)$ 可以通过基函数为 $\{B_1(u), B_2(u), \dots, B_m(u)\}$ 的 B-样条进行估计, 其中 m 指样条基函数的个数, 样条基函数的数量和形状是由节点个数和位置决定的. 本研究在模拟和实例分析过程中选择 $m = 5$.

令 $\beta_l(u) = \sum_{i=1}^m \eta_{l,i} B_i(u) = \boldsymbol{\eta}_l^T \mathbf{B}(u)$, 其中 $l = 1, 2, \dots, p_1$, $\mathbf{B}(u) = (B_1(u) \ B_2(u) \ \dots \ B_m(u))^T$, $\boldsymbol{\eta}_l = (\eta_{l,1} \ \eta_{l,2} \ \dots \ \eta_{l,m})^T$, 记 $\boldsymbol{\eta} = (\eta_{1,1} \ \eta_{1,2} \ \dots \ \eta_{1,m} \ \eta_{2,1} \ \eta_{2,2} \ \dots \ \eta_{2,m} \ \dots \ \eta_{p_1,1} \ \eta_{p_1,2} \ \dots \ \eta_{p_1,m})^T$, 并将协变量取值与对应的样条基函数相乘记为 $\mathbf{g}_{ij} = (x_{1,ij} \mathbf{B}^T(u_{ij}) \ x_{2,ij} \mathbf{B}^T(u_{ij}) \ \dots \ x_{p_1,ij} \mathbf{B}^T(u_{ij}))^T$, 则第 i 个聚类中第 j 个个体在给定 ν_i 以及其他协变量条件下的风险函数可以近似转化为

$$\lambda_{ij}(t; \mathbf{x}_{ij}, u_{ij}, \mathbf{w}_{ij}, \nu_i) \approx \lambda_0(t) \nu_i \exp(\boldsymbol{\eta}^T \mathbf{g}_{ij} + \boldsymbol{\alpha}^T \mathbf{w}_{ij}) = \lambda_0(t) \exp(\boldsymbol{\eta}^T \mathbf{g}_{ij} + \boldsymbol{\alpha}^T \mathbf{w}_{ij} + r_i) \quad (4)$$

2.2 惩罚部分似然估计

本文首先在假定 θ 已知的条件下, 采用惩罚部分似然法给出协变量参数的估计, 同时随机效应也假定为回归参数进行估计^[4]. 其中, 惩罚函数选择随机效应的负对数似然函数, 由 ν_i 的分布可知, r_i 服从对数伽马分布, 密度函数为

$$f(r) = \frac{(\exp(r))^{1/\theta} \exp(-\exp(r)/\theta)}{\theta^{1/\theta} \Gamma(1/\theta)} \quad (5)$$

显然, 随机效应密度函数的分母中并不包含随机效应, 因而在将随机效应视为额外参数进行估计时, 分母中对应的项并不起作用, 因此可以去掉负对数似然函数中与 r 无关的项, 将惩罚函数表示为 $l_{\text{pen}} = -\frac{1}{\theta} \sum_{i=1}^s (r_i - \exp(r_i))$.

进一步可以给出半变系数伽马脆弱模型的惩罚部分似然函数:

$$\begin{aligned} l_{\text{PPL}}(\boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{r}) &= l_{\text{part}} - l_{\text{pen}} = \\ &= \sum_{i=1}^s \sum_{j=1}^{n_i} \delta_{ij} \left((\boldsymbol{\beta}^T(u_{ij}) \mathbf{x}_{ij} + \boldsymbol{\alpha}^T \mathbf{w}_{ij} + \log \nu_i) - \log \left(\sum_{kq \in R(y_{ij})} \nu_k \exp(\boldsymbol{\beta}^T(u_{kq}) \mathbf{x}_{kq} + \boldsymbol{\alpha}^T \mathbf{w}_{kq}) \right) \right) + \frac{1}{\theta} \sum_{i=1}^s (r_i - \exp(r_i)) \approx \\ &= \sum_{i=1}^s \sum_{j=1}^{n_i} \delta_{ij} \left((\boldsymbol{\eta}^T \mathbf{g}_{ij} + \boldsymbol{\alpha}^T \mathbf{w}_{ij} + r_i) - \log \left(\sum_{kq \in R(y_{ij})} \exp(\boldsymbol{\eta}^T \mathbf{g}_{kq} + \boldsymbol{\alpha}^T \mathbf{w}_{kq} + r_k) \right) \right) + \frac{1}{\theta} \sum_{i=1}^s (r_i - \exp(r_i)) \quad (6) \end{aligned}$$

类似于线性伽马脆弱模型^[5], 对于固定的 θ , 可以通过最大化 $l_{\text{PPL}}(\boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{r})$ 获得 $\boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{r}$ 的估计值. 估计过程中, 首先假定 \mathbf{r} 为固定效应的回归系数, 然后分别关于 $\boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{r}$ 求 $l_{\text{PPL}}(\boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{r})$ 的得分方程:

$$\frac{\partial(l_{\text{PPL}}(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{r}))}{\partial \boldsymbol{\eta}} = \sum_{i=1}^s \sum_{j=1}^{n_i} \delta_{ij} \left(\mathbf{g}_{ij} - \sum_{kq \in R(y_{ij})} (\mathbf{g}_{kq} \cdot \exp(\boldsymbol{\eta}^T \mathbf{g}_{kq} + \boldsymbol{\alpha}^T \mathbf{w}_{kq} + r_k)) \right) / \sum_{kq \in R(y_{ij})} \exp(\boldsymbol{\eta}^T \mathbf{g}_{kq} + \boldsymbol{\alpha}^T \mathbf{w}_{kq} + r_k) = 0 \quad (7)$$

$$\frac{\partial(l_{\text{PPL}}(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{r}))}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^s \sum_{j=1}^{n_i} \delta_{ij} \left(\mathbf{w}_{ij} - \sum_{kq \in R(y_{ij})} (\mathbf{w}_{kq} \cdot \exp(\boldsymbol{\eta}^T \mathbf{g}_{kq} + \boldsymbol{\alpha}^T \mathbf{w}_{kq} + r_k)) \right) / \sum_{kq \in R(y_{ij})} \exp(\boldsymbol{\eta}^T \mathbf{g}_{kq} + \boldsymbol{\alpha}^T \mathbf{w}_{kq} + r_k) = 0 \quad (8)$$

$$\frac{\partial(l_{\text{PPL}}(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{r}))}{\partial r_h} = \sum_{i=1}^s \sum_{j=1}^{n_i} \delta_{ij} \left(z_{ij,h} - \sum_{kq \in R(y_{ij})} (z_{kq,h} \cdot \exp(\boldsymbol{\eta}^T \mathbf{g}_{kq} + \boldsymbol{\alpha}^T \mathbf{w}_{kq} + r_k)) \right) / \sum_{kq \in R(y_{ij})} \exp(\boldsymbol{\eta}^T \mathbf{g}_{kq} + \boldsymbol{\alpha}^T \mathbf{w}_{kq} + r_k) + \frac{1 - \exp(r_h)}{\theta} = 0 \quad (9)$$

其中 $h = 1, 2, \dots, s$, 且当 $i = h$ 时, $z_{ij,h} = 1$, 否则 $z_{ij,h} = 0$. 通过调整各项的排列方式, 式(9)可以改写为

$$\frac{\partial(l_{\text{PPL}}(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{r}))}{\partial r_h} = \sum_{j=1}^{n_h} \delta_{hj} - \sum_{j=1}^{n_h} \exp(\boldsymbol{\eta}^T \mathbf{g}_{hj} + \boldsymbol{\alpha}^T \mathbf{w}_{hj} + r_h) \Lambda_0(y_{hj}) + \frac{1 - \exp(r_h)}{\theta} = 0 \quad (10)$$

其中

$$\Lambda_0(y_{hj}) = \sum_{y_{(l)} \leq y_{hj}} \left(N_{(l)} / \sum_{kq \in R(y_{(l)})} \exp(\boldsymbol{\eta}^T \mathbf{g}_{kq} + \boldsymbol{\alpha}^T \mathbf{w}_{kq} + r_k) \right) \quad (11)$$

由牛顿迭代法求解得分方程, 可以给出估计 $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{r}}$, 进而可以采用 Nelson-Aalen 法得到基准危险率的估计量 $\hat{\lambda}_0$.

2.3 随机效应参数的估计

在假设 $\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{r}$ 已知的条件下, 本文采用近似轮廓似然法估计 θ . 首先, 对于固定的 $\boldsymbol{\alpha}, \boldsymbol{\eta}$, 建立边际似然函数如下:

$$L_{\text{marg}} = \prod_{i=1}^s \left(\int_0^{\infty} \prod_{j=1}^{n_i} (\lambda_0(t_{ij}) \nu_i \exp(\boldsymbol{\eta}^T \mathbf{g}_{ij} + \boldsymbol{\alpha}^T \mathbf{w}_{ij}))^{\delta_{ij}} \exp(-\Lambda_0(t_{ij}) \nu_i \exp(\boldsymbol{\eta}^T \mathbf{g}_{ij} + \boldsymbol{\alpha}^T \mathbf{w}_{ij})) \frac{\nu_i^{1/\theta-1} \exp(-\nu_i/\theta)}{\theta^{1/\theta} \Gamma(1/\theta)} d\nu_i \right) \quad (12)$$

利用伽马函数的性质(或者应用 Laplace 变换)^[6], 经过适当计算、整理, 均可将边际似然函数改写为

$$L_{\text{marg}} = \prod_{i=1}^s \left(\prod_{j=1}^{n_i} (\lambda_0(t_{ij}) \exp(\boldsymbol{\eta}^T \mathbf{g}_{ij} + \boldsymbol{\alpha}^T \mathbf{w}_{ij}))^{\delta_{ij}} \cdot (1 + \theta \Lambda_i)^{-(1/\theta + d_i)} (\theta)^{d_i} \frac{\Gamma(1/\theta + d_i)}{\Gamma(1/\theta)} \right) \quad (13)$$

其中 $\Lambda_i = \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\boldsymbol{\eta}^T \mathbf{g}_{ij} + \boldsymbol{\alpha}^T \mathbf{w}_{ij})$, $d_i = \sum_{j=1}^{n_i} \delta_{ij}$. 将 2.2 中估计得到的 $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{r}}$, 以及 $\hat{\lambda}_0(t_{ij})$ 代入边际似然函数, 得对数边际似然函数为^[4,6]

$$l_{\text{marg}} = l_{\text{PPL}} + \sum_{i=1}^s \frac{1}{\theta} \exp(\hat{r}_i) - \left(\frac{1}{\theta} + d_i \right) \cdot \log\left(\frac{1}{\theta} + d_i\right) + \frac{1}{\theta} \log\left(\frac{1}{\theta}\right) + \log\left(\frac{\Gamma(1/\theta + d_i)}{\Gamma(1/\theta)}\right) \quad (14)$$

接下来考虑 2.2 中的 $l_{\text{part}}(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{r})$, 类似于 cox 回归, 在利用得分函数求解参数估计值时, 由于分子和分母中都会出现 $\hat{\boldsymbol{r}}$, 故对各分量均相等的常数向量 \mathbf{c} , 有 $l_{\text{part}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{r}}) = l_{\text{part}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{r}} + \mathbf{c})$ 成立, 再由简单的优化计算可知, 在 $\sum_{i=1}^s \exp(\hat{r}_i) = s$ 条件下, l_{pen} 取最小值, 故不妨将 $\sum_{i=1}^s \exp(\hat{r}_i) = s$ 作为 $l_{\text{part}}(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{r})$ 优化过程中的一个约束条件, 将约束条件代入式(14)可得

$$l_{\text{marg}} = l_{\text{PPL}} + \sum_{i=1}^s \frac{1}{\theta} - \left(\frac{1}{\theta} + d_i \right) \log\left(\frac{1}{\theta} + d_i\right) + \frac{1}{\theta} \log\left(\frac{1}{\theta}\right) + \log\left(\frac{\Gamma(1/\theta + d_i)}{\Gamma(1/\theta)}\right) \quad (15)$$

利用黄金搜索算法可以得到式(15)的最大值, 从而给出随机效应方差成分的估计 $\hat{\theta}$.

2.4 模型估计流程归纳

现将整个估计流程归纳如下: 第 1 步, 运用 B-样条生成新的协变量 G_i ; 第 2 步, 在给定 θ 初始值的条件下, 利用 Newton-Raphson 算法求解惩罚对数部分似然函数的最大值, 从而给出 $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{r}}$; 第 3 步, 对于上一步得到的 $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{r}}$, 通过极

大化式(15)得到 $\hat{\theta}$; 接下来不断重复第 2 步和第 3 步直到收敛, 最后给出 α, η, θ 最终的估计.

$\beta(u) = (\beta_1(u) \ \beta_2(u) \ \cdots \ \beta_{p_1}(u))^T$ 的估计可由 $\hat{\eta}$ 和 B-样条基函数近似得到, 即 $\hat{\beta}_l(u) = \sum_{i=1}^m \hat{\eta}_{(l-1)m+i} B_i(u)$, 其中 $l=1, 2, \dots, p_1$, $\hat{\eta}_{(l-1)m+i}$ 是 $\hat{\eta}$ 的第 $(l-1)m+i$ 分量.

3 数值模拟

本文通过数值模拟的方式对提出的模型及其估计方法进行评价. 模拟分为 2 个部分: (1) 所有协变量的系数均是常数系数情形生成的数据集; (2) 协变量系数部分为变系数、部分为常数系数情形生成的数据集. 考察的模型是半变系数伽马脆弱模型, 其模型结构为

$$\lambda_{ij}(t; x_{ij}, u_{ij}, w_{ij}, \nu_i) = \lambda_0(t) \nu_i \exp(\beta(u_{ij}) \cdot x_{ij} + \alpha \cdot w_{ij});$$

$$i=1, 2, \dots, s,$$

$$j=1, 2, \dots, n_i$$

模拟中, 每个聚类中个体的个数 (n_i) 为 5, 样本量分别设定为 100、300、500, 则对应的聚类个数 (s) 分别为 20、60、100; 删失率分别设定为 10%、30%、50%, 实际删失率的误差控制在 0.5% 以内; 不同情形下, 每种模拟的次数设定为 500 次; 删失时间被设定为服从指数分布, 模拟中通过调整指数分布的参数控制删失率; 另设 $\lambda_0(t) = t$, 脆弱项 $\nu_i \sim \Gamma(1, 1)$, 即方差 (θ) 为 1 的伽马分布, $x_{ij} \sim \exp(1)$, $w_{ij} \sim N(1, 1)$; 常数系数 $\alpha=1$, 协变量 u 在区间 $(1, 3]$ 内按样本量大小等距取点, 在模拟 1 中 $\beta(u)=1$, 在模拟 2 中 $\beta(u) = \cos(2u) + 1$.

3.1 模拟 1

模拟 1 用于说明本文提出的方法是否适用于所有协变量系数均为常数的情形. 不同条件下 α 、 θ 的模拟结果见表 1.

表 1 $\beta(u)=1$ 条件下模型参数的模拟结果

Tab. 1 Simulation results of the parameters in the condition of $\beta(u)=1$

序号	删失率/%	样本量	α 真值	$\hat{\alpha}$	$S_{est}(\hat{\alpha})$	$S_{emp}(\hat{\alpha})$	95%覆盖率/%	θ 真值	$\hat{\theta}$	$S_{emp}(\hat{\theta})$
1	10	100	1	1.026	0.163	0.165	94.4	1	0.932	0.399
2	10	300	1	0.990	0.088	0.095	94.2	1	0.970	0.215
3	10	500	1	1.005	0.068	0.074	93.6	1	0.982	0.156
4	30	100	1	1.040	0.185	0.217	92.6	1	0.899	0.443
5	30	300	1	1.008	0.100	0.109	93.2	1	0.963	0.230
6	30	500	1	1.001	0.076	0.078	94.8	1	0.975	0.175
7	50	100	1	1.034	0.218	0.240	93.6	1	0.874	0.553
8	50	300	1	1.006	0.118	0.128	92.8	1	0.957	0.293
9	50	500	1	1.009	0.090	0.093	94.8	1	0.979	0.221

由表 1 可见, 不同条件下 α 的估计误差均非常小, 即使在删失率达到 50%, 而聚类个数仅为 20 个时, 估计偏差也仅为 0.034; α 标准误的估计方面, 经验标准误 S_{emp} 均略高于估计标准误 S_{est} , 这与相关文献中的研究结果一致^[2-3,7]. 不同删失率条件下, 标准误均随着样本量的增大而减小, 而相同样本量条件下, 标准误随着删失率的升高而增大. 在样本量较小时, θ 的估计误差相对较大, 删失率的提高会造成 θ 估计误差的增大, 在样本量为 500 时, 删失率的提高对于 θ 估计的影响较小.

模拟 1 中, $\beta(u)$ 的估计及其 95% 置信带见图

1. 篇幅原因, 未将样本量为 300 或者删失率为 30% 的情况显示, 其各自的估计效果介于对应的不同样本量和删失率之间.

由图 1 可见, 不同条件下, $\beta(u)$ 的估计效果均较好, 特别是当样本量为 500 时, $\beta(u)$ 和 $\hat{\beta}(u)$ 的曲线几乎是重合的; 95% 置信带的曲线形状与 $\beta(u)$ 基本一致, 只是边界有略大的波动, 这与文献结果一致^[2-3,8], 置信带的宽度随着样本量的增大而变窄; $\hat{\beta}(u)$ 的估计偏差同样受样本量的影响, 样本量越大, 估计偏差越小; 在模拟过程中考察的删失率范围内, 即 10%~50%, $\hat{\beta}(u)$ 的估计偏差无明显变化, 这体现出了本文方法对于不同的删

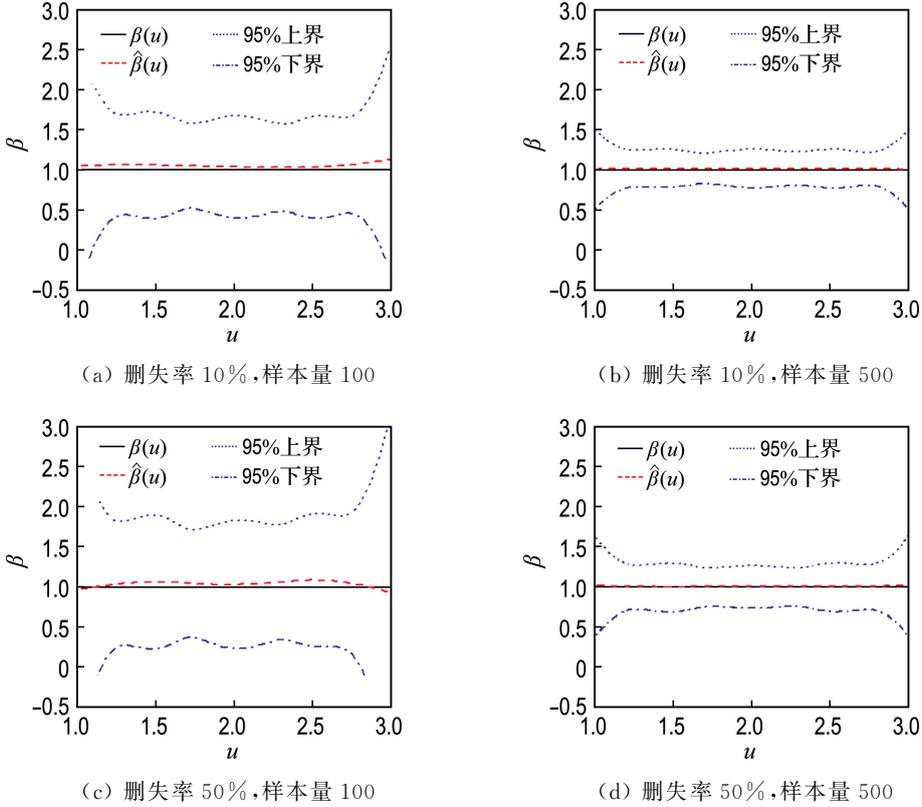


图 1 $\beta(u)=1$ 时不同样本量和删失率条件下 $\beta(u)$ 的估计

Fig. 1 Estimation of $\beta(u)$ in the condition of $\beta(u)=1$ based on different sample number and censor rate

失率具备一定的稳健性。

3.2 模拟 2

模拟 2 用于评价本文提出的方法在半变系数伽马脆弱模型条件下的拟合效果. 不同条件下 α 、 θ 的模拟结果见表 2. $\beta(u)$ 的估计及其 95% 置信带见图 2.

由表 2 可知,与模拟 1 的结果类似,不同条件下 α 的估计误差仍然都比较小,即使在删失率达到 50%,而聚类个数仅为 20 个时,估计偏差仍旧不大; α 标准误的估计方面,经验标准误也均略高于估计标准误,这与相关文献中的研究结果一致^[2-3,7]. 对于固定的删失率水平,样本量的增大可

表 2 $\beta(u)=\cos(2u)+1$ 条件下模型参数的模拟结果

Tab. 2 Simulation results of the parameters in the condition of $\beta(u)=\cos(2u)+1$

序号	删失率/%	样本量	α 真值	$\hat{\alpha}$	$S_{est}(\hat{\alpha})$	$S_{emp}(\hat{\alpha})$	95%覆盖率/%	θ 真值	$\hat{\theta}$	$S_{emp}(\hat{\theta})$
1	10	100	1	1.018	0.163	0.188	0.920	1	0.920	0.405
2	10	300	1	1.001	0.088	0.098	0.922	1	0.970	0.209
3	10	500	1	0.998	0.068	0.070	0.946	1	0.981	0.153
4	30	100	1	1.020	0.185	0.196	0.944	1	0.878	0.428
5	30	300	1	1.004	0.100	0.102	0.954	1	0.944	0.228
6	30	500	1	1.000	0.077	0.080	0.952	1	0.977	0.172
7	50	100	1	1.012	0.219	0.251	0.928	1	0.861	0.516
8	50	300	1	1.004	0.118	0.132	0.930	1	0.945	0.281
9	50	500	1	1.006	0.091	0.092	0.946	1	0.967	0.220

以带来估计标准误和经验标准误的减小;与此同时,对于固定的样本量,估计标准误和经验标准误也随着删失率的提高而略有增大.在样本量较小同时删失率又较高时, θ 的估计误差相对较大,模拟中最大平均误差达到近0.14;在样本量较小时,删失率的提高会造成 θ 估计误差较明显的增大,而在样本量较大时,删失率的提高对于 θ 估计的影响不再显著.

由图2可知,对应于不同的样本量和删失率,

$\beta(u)$ 的平均估计均比较准确,尤其是在样本量较大(500)时, $\beta(u)$ 和 $\hat{\beta}(u)$ 的图像几乎是重合的.与模拟1类似,95%置信带的曲线形状与 $\beta(u)$ 基本一致,只是在边界处有相对较大的波动,置信带的宽度随着样本量的增大而变窄; $\hat{\beta}(u)$ 的偏差同样受样本量的影响,样本量越大,偏差越小;在模拟过程中考察的删失率范围内,即10%~50%, $\hat{\beta}(u)$ 偏差的变化并不显著,模拟2和模拟1共同体现出了本文提出方法是比较稳健的.

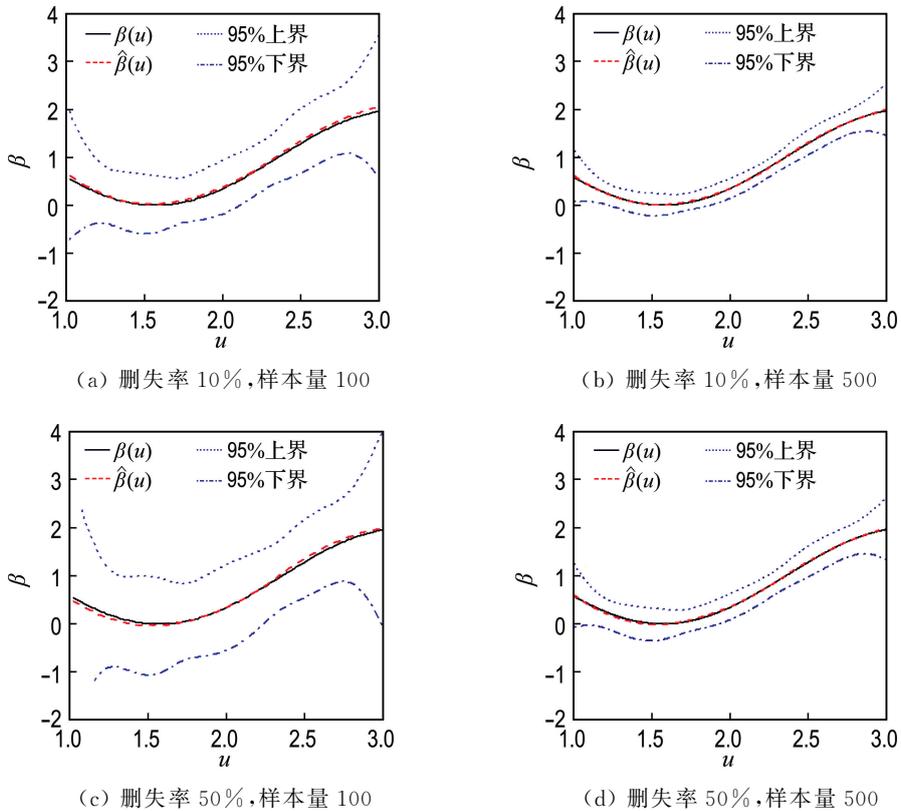


图2 $\beta(u) = \cos(2u) + 1$ 时不同样本量和删失率条件下 $\beta(u)$ 的估计

Fig. 2 Estimation of $\beta(u)$ in the condition of $\beta(u) = \cos(2u) + 1$ based on different sample number and censor rate

4 实例分析

本文通过分析美国北部癌症治疗中心(North Central Cancer Treatment Group, NCCTG)的晚期肺癌数据来评价本文提出的模型与方法的应用效果.调查涉及167名晚期肺癌患者,删失率为28%,文献中已经有关于本数据集的一些分析^[9-10],本研究重点考察病人Karnofsky自评分对于危险率的影响受其他因素的影响情况,了解晚期肺癌的预后因素,从而为医师以及病

人制订更合理的治疗方案提供参考.所谓预后是指预测疾病的可能病程和结局,它既包括判断疾病的特定后果,如康复,某种症状、体征和并发症等其他异常的出现或消失及死亡,也包括提供时间线索,如预测某段时间内发生某种结局的可能性.

纳入本研究的评价指标包括机构代码(I)、生存时间(T)、删失指标(C)、病人年龄(U)、病人的Karnofsky自评分(X)、性别(W_1)、ECOG得分

(W_2)、卡路里摄入量(W_3)。考虑到不同医疗机构的治疗水平存在差别,即考虑病人的生存时间在就医机构方面表现出聚集性,即具有不可观测的随机效应,因而将这些变量代入半变系数伽马脆弱模型,模型结构如下:

$$\lambda_{ij}(t; \nu_i, u_{ij}, x_{ij}, \omega_{1,ij}, \omega_{2,ij}, \omega_{3,ij}) = \lambda_0(t) \nu_i \exp(\beta(u_{ij}) x_{ij} + \alpha_1 \omega_{1,ij} + \alpha_2 \omega_{2,ij} + \alpha_3 \omega_{3,ij}) \quad (16)$$

采用本文提出方法给出各协变量的估计,同时采用 bootstrap 方法给出 95% 置信带的估计,结果显示,性别以及 ECOG 得分对于危险率的影响具有统计学意义,男性相比女性危险率更高, ECOG 得分越高,危险率也越高,详见表 3。

表 3 NCCTG 数据回归参数的估计

Tab. 3 Estimation of the regression parameters of NCCTG data

变量	回归系数	标准误	χ^2	p
W_1	-0.530 6	0.204 3	6.742 6	0.009 4
W_2	0.388 7	0.168 3	5.334 2	0.020 9
W_3	0	0.000 2	0.020 0	0.887 5

图 3 显示,不同年龄段晚期肺癌患者的 Karnofsky 自评分对对数危险率的影响大小非常数,而是一个非线性函数。

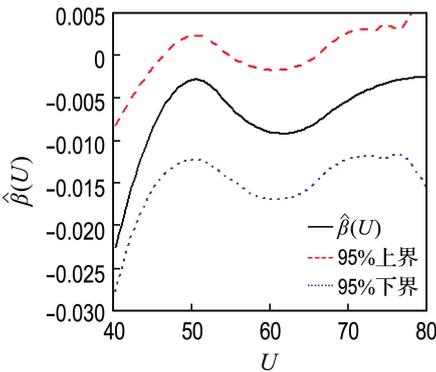


图 3 NCCTG 数据分析中变系数函数及其置信带的估计

Fig. 3 Estimation of the varying coefficient function and confidence belt of NCCTG data

5 结 语

经模拟研究和实例分析发现,在样本量不大

的条件下,本文提出的方法即可给出模型线性回归系数非常准确的估计,删失率的提高也不会对参数的估计效果造成明显影响.在样本量较小同时删失率又较高的条件下,随机效应参数的估计误差较大,这提示在实际的应用过程中本方法有低估随机效应方差的倾向,当样本量较大时,随机效应参数的估计误差则会明显减小.本文提出的方法可以给出常数函数非常准确的估计,即适用于所有系数均为线性回归系数的场合,也即本文的方法包含传统的线性伽马脆弱模型作为其特殊形式;同时,本文提出的模型可以给出非线性函数系数非常准确的估计,这也扩展了伽马脆弱模型的适用范围和应用领域。

综上所述,本文提出的方法对传统的伽马脆弱模型进行了有效的扩展,方法稳定、计算速度也较快,常数回归系数以及函数回归系数的估计对样本量和删失率的要求均不高,适宜在实际问题中推广使用.当然,本研究中也存在一些不足,例如:虽然在模拟研究中给出了变系数函数的置信带,在实例分析中也应用 bootstrap 方法给出了变系数函数的置信带,但未能就函数系数的假设检验等问题进行探讨,这也有待于接下来更深入的研究。

参 考 文 献:

- [1] DU Pang, MA Shuangge. Frailty model with spline estimated nonparametric hazard function [J]. *Statistica Sinica*, 2010, **20**(2):561-580.
- [2] YU Zhangsheng, LIN Xihong, TU Wanzhu. Semiparametric frailty models for clustered failure time data [J]. *Biometrics*, 2012, **68**(2):429-436.
- [3] ZHANG Z, SONG L, WANG X, *et al.* Estimation of multivariate frailty models with varying coefficients [J]. *Communication in Statistics-Theory and Methods*, 2018(2):1-12.
- [4] THERNEAU T M, GRAMBSCH P M, PANKRATZ V S. Penalized survival models and frailty [J]. *Journal of Computational & Graphical Statistics*, 2003, **12**(1):156-175.
- [5] GRAMBSCH P M. *Modeling Survival Data: Extending the Cox Model* [M]. New York: Springer-Verlag, 2000.

- [6] DUCHATEAU L, JANSSEN P. **The Frailty Model** [M]. New York: Springer, 2008:199-233.
- [7] RIPATTI S, PALMGREN J. Estimation of multivariate frailty models using penalized partial likelihood [J]. **Biometrics**, 2000, **56** (4): 1016-1022.
- [8] YU Zhangsheng, LIU Lei, BRAVATA D M, *et al.* A semiparametric recurrent events model with time-varying coefficients [J]. **Statistics in Medicine**, 2013, **32**(6):1016-1026.
- [9] LOPRINZI C L, LAURIE J A, WIEAND H S, *et al.* Prospective evaluation of prognostic variables from patient-completed questionnaires [J]. **Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology**, 1994, **12**(3): 601-607.
- [10] WANG Xiaoguang, SHI Xinyong. Robust estimation for survival partially linear single-index models [J]. **Computational Statistics & Data Analysis**, 2014, **80**(4):140-152.

Penalized partial likelihood estimation of semi-varying coefficient Gamma frailty models

ZHANG Zhongwen^{1,2}, WANG Xiaoguang^{*1}, SONG Lixin¹

(1. School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China;

2. School of Public Health and Management, Binzhou Medical University, Yantai 264003, China)

Abstract: To analyze more complex nonlinear relationships between the logarithmic risk function and covariants, a set of semi-varying coefficient Gamma frailty models and their estimation method are proposed. Firstly, the semi-varying coefficient Gamma frailty models are approximatively transformed to linear Gamma frailty models using B-spline. Secondly, the linear parameters of transformed models are estimated by the penalized partial likelihood. Thirdly, the profile likelihood method is adopted to estimate the parameter of random effect using the golden section search method. After the estimations of linear parameters and random effect parameters are gotten from the iterative algorithm, the estimations of varying coefficient functions can be obtained taking advantage of B-spline. The finite sample performance of the proposed method is assessed by Monte Carlo simulation studies, the method can give fully precise and stabilized estimation of the linear parameters and varying coefficient function, and can be used to analyze the influence of the covariants on hazard rates. At last, the proposed method is demonstrated by the analysis of NCCTG lung cancer data.

Key words: Gamma frailty model; B-spline; varying coefficient models; penalized partial likelihood estimation; golden section search method