王淑影,张贺楠,赵波,等. 区间删失数据下基于 OLLGG 分布的多参数回归模型的参数估计 [J]. 江西师范大学学报(自然科学版),2023,47(2):199-205.
WANG Shuying,ZHANG He′nan,ZHAO Bo,et al. The estimation of the multiple parameter regression model with interval-censored data based on the odd log-logistic generalized Gompertz distribution [J]. Journal of Jiangxi Normal University (Natural Science),2023,47(2):199-205.

文章编号:1000-5862(2023)02-0199-07

区间删失数据下基于 OLLGG 分布的多参数回归模型的参数估计

王淑影,张贺楠,赵　波,程云飞
(长春工业大学数学与统计学院,吉林 长春　130012)

摘要:该文基于Ⅱ型区间删失数据,在 OLLGG 分布下提出多参数回归模型,通过线性回归刻画分布参数与协变量之间的关系,并通过极大似然方法给出了模型的参数估计,数值模拟验证了模型参数的估计有良好的性质,将提出的模型应用到血友病患者 HIV 感染的数据中,发现提出的模型对数据有灵活的拟合效果.

关键词:多参数回归模型;Ⅱ型区间删失;OLLGG 分布;极大似然估计
中图分类号:O 212.1　　**文献标志码**:A　　**DOI**:10.16357/j.cnki.issn1000-5862.2023.02.13

0　引言

在临床医学的随访中,兴趣事件发生的时间经常不能被精确记录,只能记录随访时间以及在随访时兴趣事件发生的状态,这类数据被定义为区间删失失效时间数据[1]. 关于这类数据的分析与建模一直是众多学者研究的热点. 因此,本文将基于更加灵活的 4 参数 odd log-logistic generalized Gompertz (OLLGG)分布[2]建立在区间删失数据下的多参数回归模型,分析兴趣事件发生时间的分布模型参数与协变量之间的线性关系.

目前对于区间删失数据的建模与分析已有很多文献[1,3-10],但大多是基于指数分布、威布尔分布、Gompertz 分布等建立的参数模型[4],这些模型尽管也具有较为灵活的形式,但是受参数个数与参数形式的限制,并不能完全拟合出事件发生时间的曲线. 如指数分布描述了常数风险模型、广义 Gompertz 模型的风险函数只能是递增函数或者是常数,不能提供适合于浴缸型风险函数的建模现象等[4]. Sun Jianguo[1]讨论在指数分布、威布尔分布下区间删失数据的风险回归模型;E. M. Hashimoto 等[5]提出了一种基于对数广义 Gamma 分布的位置-尺度回归模型;Peng Defen 等[6]主要研究了全参数、非比例风险的多参数威布尔回归生存模型;文献[7-8]建立了在广义指数分布与威布尔分布下区间删失数据的加速失效时间回归模型. 此外,还有许多文献讨论了在区间删失数据下的半参数回归分析,并给出了推断方法. 如 Hu Tao 等[9]讨论了在Ⅰ型区间删失数据下半参数比例风险模型的极大似估计方法,并证明了极大似然估计的大样本性质;文献[10-11]提出了在Ⅰ型区间删失数据下加性风险模型的估计方程方法;Zeng Donglin 等[12]提出了建立Ⅱ型区间删失数据的半参数加性风险模型,并给出了半参数极大似然估计;Wang Lianming 等[13]基于加性风险模型提出了在区间删失数据下的估计方程方法,并运用

收稿日期:2023-01-10
基金项目:中国博士后基金面上课题(2021M700536)和吉林省自然科学基金优秀青年课题(20230101371JC)资助项目.
作者简介:王淑影(1990—),女,吉林省榆树人,副教授,博士,博士生导师,主要从事数理统计、生物统计研究. E-mail:wangshuying@ccut.edu.cn

计数过程的鞅理论证明了参数估计的渐近性质;R. A. Betensky 等[14]建立了 II 型区间删失数据加速失效时间模型等. 这些已有文献主要讨论了失效时间风险函数的回归建模问题. 2017 年 K. Burke 等[15]从分布参数回归的角度提出了在右删失数据下多参数回归(MPR)模型概念,增加了回归模型的灵活性与可解释性. 因此,本文将基于区间删失数据建立在 4 参数 OLLGG 分布下的多参数回归模型,分析协变量与分布参数之间的线性关系,并利用极大似然估计给出在 OLLGG 分布下多参数回归模型参数的估计量.

本文先介绍了 II 型区间删失数据的数据结构以及在 OLLGG 分布下多参数回归模型形式;然后给出了在Ⅱ型区间删失数据的 OLLGG 分布下多参数回归模型的似然函数与极大似然估计;再运用数值模拟验证了估计的性质;最后将提出的模型运用到血友病患者 HIV-1 感染的数据集中.

1 数据和模型

1.1 Ⅱ型区间删失数据

假设随机变量 T 表示兴趣事件发生的时间. 在随访过程中,不能记录事件发生的精确时间,仅仅记录从实验开始到实验结束的有效随访时间 U 和 V $(U<V)$,事件发生的时间 T 在随访时间 U 和 V 构成的最小时间区间内. 因此,定义示性变量 $\delta_1=I(0<T\leqslant U)$、$\delta_2=I(U<T\leqslant V)$、$\delta_3=I(V<T\leqslant\infty)$ 分别表示兴趣事件发生时间的区间,$\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3$、$\boldsymbol{X}_4$ 分别是维度为 p_1、p_2、p_3、p_4 的 4 类协变量,则区间删失数据的数据结构表示为

$$D=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3,\boldsymbol{X}_4,U,V,\delta_1,\delta_2,\delta_3\},$$

其中示性变量满足 $\delta_1+\delta_2+\delta_3=1$.

1.2 OLLGG 分布

OLLGG 分布是 M. Alizadeh 等[2]基于广义 Gompertz 分布提出的一类特别灵活的 4 参数分布. 假设正随机变量 T 服从广义 Gompertz 分布函数,则分布函数表示为

$$G(t;a,b,c)=(1-\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a})^c,\ t>0,a>0,b>0,c>0, \tag{1}$$

其中 a、b 表示尺度参数,c 表示形状参数. 基于微分可得随机变量 T 的密度函数表示为

$$g(t;a,b,c)=cb\mathrm{e}^{at}\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a}(1-\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a})^{c-1},\ t>0,a>0,b>0,c>0. \tag{2}$$

基于 J. U. Gleaton 等[16]提出 odd log-logistic-Gompertz(OLLG)分布函数思想,用式(1)和式(2)重新替换在 OLLG 分布中的密度函数和分布函数可得 odd log-logistic generalized Gompertz 分布,其密度函数表示为

$$f_0(t;a,b,c,d)=dcb\mathrm{e}^{at}\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a}(1-\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a})^{cd-1}\cdot(1-(1-\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a})^c)^{d-1}/((1-\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a})^{cd}+(1-(1-\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a})^c)^d)^2,$$

OLLGG 分布的分布函数为

$$F_0(t;a,b,c,d)=(1-\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a})^{cd}/((1-\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a})^{cd}+(1-(1-\mathrm{e}^{-b(\mathrm{e}^{at}-1)/a})^c)^d), \tag{3}$$

其中 a、b 表示尺度参数,c、d 表示形状参数,则生存函数为

$$S_0(t;a,b,c,d)=1-F_0(t;a,b,c,d).$$

特别地,当形状参数 $c=1$ 和 $d=1$ 时,该分布退化为 Gompertz 分布,若同时尺度参数 $a\to0$,则该分布退化为指数分布;若形状参数 $d=1$,则分布退化为 generalized Gompertz 分布,若同时尺度参数 $a\to0$,则分布退化为广义指数分布;此外,若尺度参数 $c=1$,则该分布为 OLLG 分布,若同时满足 $a\to0$,则该分布退化为 odd log-logistic 广义指数分布;若仅有 $a\to0$,则该分布退化为 odd log-logistic 指数分布. 因此,OLLGG 分布可以灵活地转换成各种分布形式.

1.3 多参数 OLLGG 回归模型

在已有生存分析建模中,协变量与失效时间所建立的模型一般是加速失效时间模型、加性风险模型以及比例风险模型等. 2017 年,K. Burke 等[15]从分布参数回归的角度提出的一种了新的回归方式——多参数回归(MPR)模型,其基本思想是基于参数分布的每个参数建立回归模型,这种建模方式的优点是多个参数回归增强了回归模型的灵活性与可解释性. 本文将基于文献[15]的方法建立在 OLLGG 分布下的多参数回归模型. 因此,基于 OLLGG 分布的风险函数,假设在分布模型中尺度参数 a、b 和形状参数 c、d 分别受协变量 $\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3$、$\boldsymbol{X}_4$ 的影响,则尺度参数回归 $a=\exp(\tilde{\alpha}_0+\boldsymbol{X}_1\boldsymbol{\alpha}_1)$、$b=\exp(\tilde{\beta}_0+\boldsymbol{X}_2\boldsymbol{\beta}_1)$,形状参数回归 $c=\exp(\tilde{\lambda}_0+\boldsymbol{X}_3\boldsymbol{\lambda}_1)$、$d=\exp(\tilde{\gamma}_0+\boldsymbol{X}_4\boldsymbol{\gamma}_1)$,其中 $\tilde{\alpha}_0$、$\tilde{\beta}_0$、$\tilde{\lambda}_0$、$\tilde{\gamma}_0$ 表示截距项,系数向量 $\boldsymbol{\alpha}_1$、$\boldsymbol{\beta}_1$ 分别表示 2 个尺度参数的回归参数向量,系数向量 $\boldsymbol{\lambda}_1$、$\boldsymbol{\gamma}_1$ 分别表示 2 个形状参数的回归参数向量. 在实际问题分析中,协变量 $\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3$、$\boldsymbol{X}_4$ 允许完全一致,即 $\boldsymbol{X}_1=\boldsymbol{X}_2=\boldsymbol{X}_3=\boldsymbol{X}_4$.

假设 $\boldsymbol{\alpha}=(\tilde{\alpha}_0,\boldsymbol{\alpha}_1^{\mathrm{T}})^{\mathrm{T}}$,$\boldsymbol{\beta}=(\tilde{\beta}_0,\boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$,$\boldsymbol{\lambda}=(\tilde{\lambda}_0,$

$\boldsymbol{\lambda}_1^{\mathrm{T}})^{\mathrm{T}}$,$\boldsymbol{\gamma}=(\tilde{\gamma}_0,\boldsymbol{\gamma}_1^{\mathrm{T}})^{\mathrm{T}}$. 将参数线性回归模型代入到分布函数(3)中,则当给定协变量 $\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3$、$\boldsymbol{X}_4$ 时,OLLGG 分布的多参数回归模型的分布函数为

$$F(t;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})=(A(t;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}))^{\exp(\tilde{\gamma}_0+\boldsymbol{X}_4\boldsymbol{\gamma}_1)}/((A(t;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}))^{\exp(\tilde{\gamma}_0+\boldsymbol{X}_4\boldsymbol{\gamma}_1)}+(1-A(t;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}))^{\exp(\tilde{\gamma}_0+\boldsymbol{X}_4\boldsymbol{\gamma}_1)}),$$

其中 $A(t;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})=1-(1-\exp(-\exp(\tilde{\beta}_0+X_2\boldsymbol{\beta}_1)(\exp(\exp(\tilde{\alpha}_0+\boldsymbol{X}_1\boldsymbol{\alpha}_1)^t)-1)/\exp(\tilde{\alpha}_0+\boldsymbol{X}_1\boldsymbol{\alpha}_1)))^{\exp(\tilde{\lambda}_0+\boldsymbol{X}_3\boldsymbol{\lambda}_1)}$.

生存函数 $S(t;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})=1-F(t;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})$,注意到,当 $\boldsymbol{\alpha}_1=\boldsymbol{\beta}_1=\boldsymbol{\lambda}_1=\boldsymbol{\gamma}_1=\boldsymbol{0}$ 时,模型变为传统 OLLGG 分布模型.

2　极大似然估计

假设 $D_i=\{\boldsymbol{X}_{1i},\boldsymbol{X}_{2i},\boldsymbol{X}_{3i},\boldsymbol{X}_{4i},U_i,V_i,\delta_{1i},\delta_{2i},\delta_{3i}\}$ $(i=1,2,\cdots,n,\delta_{1i}+\delta_{2i}+\delta_{3i}=1)$ 是数据 $D=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3,\boldsymbol{X}_4,U,V,\delta_1,\delta_2,\delta_3\}$ 的 n 个独立同分布观测样本,T_i 表示第 i 个样本的精确失效时间. 在给定协变量 $\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3$、$\boldsymbol{X}_4$ 时,样本失效时间 T_i 与随访过程时间 U_i、V_i 独立,因此,在给定观测数据时,在 OLLGG 分布下的多参数回归模型的似然函数表示为

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})=\prod_{i=1}^{n}(1-S(U_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}))^{\delta_{1i}}(S(U_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})-S(V_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}))^{\delta_{2i}}(S(V_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}))^{\delta_{3i}},$$

则对数似然函数表示为

$$l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})=\ln L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})=\sum_{i=1}^{n}\delta_{1i}\ln(1-S(U_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}))+\sum_{i=1}^{n}\delta_{2i}\ln(S(U_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})-S(V_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}))+\sum_{i=1}^{n}\delta_{3i}\ln S(V_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}).\tag{4}$$

为了求解模型的参数,一种最直接的方法是极大化对数似然函数,因此,基于对数似然函数(4)计算各个参数的偏导数,给出得分方程分别为

$$\tilde{U}(\boldsymbol{\alpha})=\partial l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})/\partial\boldsymbol{\alpha}=0,$$

$$\tilde{U}(\boldsymbol{\beta})=\partial l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})/\partial\boldsymbol{\beta}=0,$$

$$\tilde{U}(\boldsymbol{\lambda})=\partial l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})/\partial\boldsymbol{\lambda}=0,$$

$$\tilde{U}(\boldsymbol{\gamma})=\partial l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}=0.$$

联立方程并求解获得模型的极大似然估计量 $\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\gamma}})$. 根据文献[17]的定理 2.13 可知,在参数模型假设下极大似然估计量 $\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\gamma}})$ 具有相合性和渐近正态性. 因此,假设 α_0、β_0、λ_0、γ_0 是模型参数的真值,则 $\sqrt{n}(\hat{\boldsymbol{\alpha}}-\alpha_0,\hat{\boldsymbol{\beta}}-\beta_0,\hat{\boldsymbol{\lambda}}-\lambda_0,\hat{\boldsymbol{\gamma}}-\gamma_0)$ 收敛到均值为 0、方差为 Σ 的正态分布,即

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}-\alpha_0,\hat{\boldsymbol{\beta}}-\beta_0,\hat{\boldsymbol{\lambda}}-\lambda_0,\hat{\boldsymbol{\gamma}}-\gamma_0)\xrightarrow{L}N(0,\Sigma),$$

其中方差 Σ 的维度为 $(p_1+p_2+p_3+p_4+4)\times(p_1+p_2+p_3+p_4+4)$. 在模拟与实际问题的分析中,$\Sigma$ 相合估计通过 Fisher 信息矩阵估计,则

$$\Sigma^{-1}=-\begin{pmatrix}\frac{\partial\tilde{U}(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} & \frac{\partial\tilde{U}(\boldsymbol{\alpha})}{\partial\boldsymbol{\beta}} & \frac{\partial\tilde{U}(\boldsymbol{\alpha})}{\partial\boldsymbol{\lambda}} & \frac{\partial\tilde{U}(\boldsymbol{\alpha})}{\partial\boldsymbol{\gamma}}\\ \frac{\partial\tilde{U}(\boldsymbol{\beta})}{\partial\boldsymbol{\alpha}} & \frac{\partial\tilde{U}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} & \frac{\partial\tilde{U}(\boldsymbol{\beta})}{\partial\boldsymbol{\lambda}} & \frac{\partial\tilde{U}(\boldsymbol{\beta})}{\partial\boldsymbol{\gamma}}\\ \frac{\partial\tilde{U}(\boldsymbol{\lambda})}{\partial\boldsymbol{\alpha}} & \frac{\partial\tilde{U}(\boldsymbol{\lambda})}{\partial\boldsymbol{\beta}} & \frac{\partial\tilde{U}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}} & \frac{\partial\tilde{U}(\boldsymbol{\lambda})}{\partial\boldsymbol{\gamma}}\\ \frac{\partial\tilde{U}(\boldsymbol{\gamma})}{\partial\boldsymbol{\alpha}} & \frac{\partial\tilde{U}(\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}} & \frac{\partial\tilde{U}(\boldsymbol{\gamma})}{\partial\boldsymbol{\lambda}} & \frac{\partial\tilde{U}(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\end{pmatrix}_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}},\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\\ \boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}},}}$$

其中 $\hat{\Sigma}^{-1}$ 表示矩阵 $\hat{\Sigma}$ 的逆矩阵.

3　模拟研究

本部分将通过数值模拟验证在 OLLGG 分布下的多参数回归模型极大似然估计的有效性. 在模拟计算中,根据协变量情况,考虑了 2 种模型设置.

(Ⅰ)假设维度 $p_1=p_2=p_3=p_4=1$,且协变量 $X_{1i}=X_{2i}=X_{3i}=X_{4i}$,并假设服从均值为 0、方差为 1 的标准正态分布,并设置 2 组参数真值:

(A1) $\tilde{\alpha}_0=-0.4,\alpha_1=0.2,\tilde{\beta}_0=0.2,\beta_1=-0.4,\tilde{\lambda}_0=0.8,\lambda_1=0.2,\tilde{\gamma}_0=-0.8,\gamma_1=-0.2$;

(A2) $\tilde{\alpha}_0=0.5,\alpha_1=-0.3,\tilde{\beta}_0=-0.4,\beta_1=0.3,\tilde{\lambda}_0=0.8,\lambda_1=0.2,\tilde{\gamma}_0=-0.8,\gamma_1=-0.2.$

或者假设 $X_{1i}=X_{2i}=X_{3i}=X_{4i}$ 且服从成功概率为 0.5 的伯努利分布(二项分布),并设置 2 组模型参数真值:

(A3) $\tilde{\alpha}_0=0.8,\alpha_1=-0.2,\tilde{\beta}_0=-0.7,\beta_1=0.2,\tilde{\lambda}_0=0.6,\lambda_1=0.3,\tilde{\gamma}_0=-0.7,\gamma_1=-0.2$;

(A4) $\tilde{\alpha}_0=0.8,\alpha_1=0.2,\tilde{\beta}_0=-0.7,\beta_1=-0.2,\tilde{\lambda}_0=0.6,\lambda_1=0.3,\tilde{\gamma}_0=-0.7,\gamma_1=-0.2.$

(Ⅱ)假设维度 $p_1=p_2=p_3=p_4=2$,协变量 $\boldsymbol{X}_{1i}=(X_{11i},X_{12i})$,$\boldsymbol{X}_{2i}=(X_{21i},X_{22i})$,$\boldsymbol{X}_{3i}=(X_{31i},X_{32i})$,

$\boldsymbol{X}_{4i}=(X_{41i},X_{42i})$,且 $X_{11i}=X_{12i}=X_{13i}=X_{14i}$并服从成功概率为 0.5 的伯努利分布,$X_{21i}=X_{22i}=X_{23i}=X_{24i}$并服从均值为 0、方差为 1 的标准正态分布. 与情形(Ⅰ)一致,同样设置下列 2 组模型参数真值:

(B1) $\tilde{\alpha}_0=-0.3,\alpha_1=-0.4,\alpha_2=-0.2,\tilde{\beta}_0=0.4,\beta_1=0.2,\beta_2=0.2,\tilde{\lambda}_0=-0.5,\lambda_1=0.2,\lambda_2=-0.2,\tilde{\gamma}_0=0.3,\gamma_1=-0.2,\gamma_2=0.2$;

(B2) $\tilde{\alpha}_0=0.3,\alpha_1=-0.4,\alpha_2=0.2,\tilde{\beta}_0=-0.4,\beta_1=0.2,\beta_2=-0.2,\tilde{\lambda}_0=-0.5,\lambda_1=0.2,\lambda_2=-0.2,\tilde{\gamma}_0=0.3,\gamma_1=-0.2,\gamma_2=0.2$.

在基于 OLLGG 分布的多参数回归模型假设下,根据上述 2 种参数设置生成失效时间 T_i. 为了生成随访过程时间,首先在区间$(0,\tau_1)$内生成服从均匀分布的观测时间 U_i,并在区间$(U_i+0.01,\tau_2)$内生成服从均匀分布的观测时间 V_i,其中在模型设置(Ⅰ)中,考虑在服从二项分布的协变量时,设置$\tau_1=1.2,\tau_2=2.0$,在服从正态分布的协变量时,设置 $\tau_1=1.0,\tau_2=5.0$;在模型设置(Ⅱ)中,设置$\tau_1=0.2,\tau_2=2.0$. 比较失效时间 T_i 与观测时间 U_i 和 V_i 的关系,生成示性变量 $\delta_{1i}=I(T_i\leqslant U_i)$、$\delta_{2i}=I(U_i<T_i\leqslant V_i)$和 $\delta_{3i}=1-\delta_{1i}-\delta_{2i}$.

在上述 2 种模型设置下,分别考虑了样本容量 $n=400$ 和 $n=600$ 的情况,并重复 1 000 随机实验获得表 1 ~ 表 3 的结果. 表中结果包含有模型参数的真实设置(Para)、极大似然估计值减去参数真实值获得的估计的平均偏差(BIAS)、极大似然估计量的样本标准差(SSE)、估计值的样本标准差均值(ESE)和以重复 1 000 次随机实验所得到的经验覆盖概率(CP). 根据表中计算的结果可知:在不同样本量下,估计量的平均偏差较小,且接近于 0,估计值基本接近真实值,样本标准差(SSE)与估计值的样本标准差均值(ESE)近似相等,覆盖概率的估计值在 0.95 左右波动. 随样本量的增大,估计量的平均偏差和标准差均减小,因此,提出的方法获得估计量是相合的和渐近有效的. 比较表 1 和表 2 可知:在改变协变量的分布假设时,提出模型的估计结果表现一致且稳定. 表 3 表明:随着协变量维度增加,估计结果仍然是稳健的.

表 1 在模型设置(Ⅰ)情形下正态分布设置的参数估计的模拟结果

参数	Para	BIAS	ESE	SSE	CP	BIAS	ESE	SSE	CP
		$n=400$				$n=600$			
$\tilde{\alpha}_0$	−0.4	−0.016 8	0.289 7	0.354 6	0.938	−0.010 8	0.168 8	0.235 8	0.941
α_1	0.2	−0.019 2	0.293 7	0.314 4	0.936	−0.001 8	0.124 0	0.135 5	0.944
$\tilde{\beta}_0$	0.2	−0.006 0	0.687 3	0.597 1	0.958	−0.003 4	0.581 1	0.572 9	0.956
β_1	−0.4	−0.022 8	0.537 3	0.530 9	0.968	−0.012 1	0.390 9	0.388 3	0.961
$\tilde{\lambda}_0$	0.8	0.047 4	0.551 1	0.584 0	0.926	0.040 3	0.407 2	0.411 5	0.936
λ_1	0.2	−0.017 3	0.627 7	0.640 6	0.930	−0.009 2	0.515 0	0.521 3	0.946
$\tilde{\gamma}_0$	−0.8	−0.016 9	0.526 9	0.539 0	0.926	−0.008 5	0.465 6	0.475 2	0.948
γ_1	−0.2	−0.012 6	0.467 2	0.490 4	0.922	−0.006 7	0.323 2	0.332 3	0.946
$\tilde{\alpha}_0$	0.5	−0.045 6	0.368 9	0.365 6	0.958	−0.041 9	0.157 1	0.166 3	0.943
α_1	−0.3	0.029 5	0.297 8	0.291 6	0.969	0.028 1	0.125 1	0.127 8	0.927
$\tilde{\beta}_0$	−0.4	0.003 0	0.390 0	0.409 2	0.929	0.002 6	0.296 2	0.297 9	0.949
β_1	0.3	−0.028 1	0.429 1	0.393 2	0.976	−0.016 5	0.272 5	0.278 3	0.944
$\tilde{\lambda}_0$	0.8	−0.015 4	0.461 0	0.501 3	0.936	−0.015 0	0.351 1	0.359 5	0.938
λ_1	0.2	−0.018 0	0.409 1	0.397 1	0.966	−0.013 5	0.331 2	0.328 1	0.968
$\tilde{\gamma}_0$	−0.8	0.022 3	0.460 3	0.455 7	0.966	0.024 0	0.285 8	0.294 0	0.933
γ_1	−0.2	0.018 4	0.267 7	0.285 4	0.962	0.014 7	0.234 6	0.241 2	0.937

注:Para 表示真实值;BIAS 表示估计的样本偏差;SSE 表示标准误差;ESE 表示标准差均值;CP 表示置信水平为 95% 的置信区间的覆盖概率. 下文同.

表 2　在模型设置(Ⅰ)情形下二项分布设置的参数估计的模拟结果

参数	Para	BIAS	ESE	SSE	CP	BIAS	ESE	SSE	CP
		$n=400$				$n=600$			
$\tilde{\alpha}_0$	0.8	−0.016 8	0.492 3	0.483 6	0.953	−0.017 5	0.276 9	0.290 4	0.952
α_1	−0.2	0.018 3	0.586 3	0.541 6	0.931	0.017 7	0.419 9	0.438 9	0.929
$\tilde{\beta}_0$	−0.7	−0.030 8	0.877 3	0.889 3	0.944	0.010 5	0.718 0	0.690 4	0.950
β_1	0.2	−0.013 3	0.923 5	0.918 6	0.938	−0.012 4	0.859 2	0.818 9	0.940
$\tilde{\lambda}_0$	0.6	0.064 2	0.905 3	0.876 2	0.963	0.051 3	0.736 8	0.741 8	0.942
λ_1	0.3	−0.014 4	0.949 9	0.972 4	0.938	−0.017 4	0.821 5	0.867 4	0.943
$\tilde{\gamma}_0$	−0.7	−0.011 9	0.666 7	0.651 8	0.962	−0.003 7	0.549 7	0.608 4	0.945
γ_1	−0.2	0.014 3	0.771 9	0.770 5	0.938	0.010 7	0.701 5	0.636 3	0.951
$\tilde{\alpha}_0$	0.8	0.032 8	0.275 6	0.284 8	0.938	0.018 8	0.198 6	0.202 2	0.946
α_1	0.2	−0.010 3	0.441 5	0.448 3	0.925	−0.009 2	0.399 1	0.401 2	0.927
$\tilde{\beta}_0$	−0.7	−0.017 4	0.563 5	0.573 9	0.937	−0.010 4	0.445 0	0.451 8	0.944
β_1	−0.2	0.044 4	0.707 6	0.712 5	0.938	0.032 7	0.560 5	0.562 6	0.948
$\tilde{\lambda}_0$	0.6	0.031 6	0.555 2	0.566 6	0.942	0.023 6	0.451 3	0.460 1	0.947
λ_1	0.3	−0.013 2	0.645 8	0.650 7	0.926	−0.011 8	0.514 9	0.517 3	0.929
$\tilde{\gamma}_0$	−0.7	0.027 7	0.451 3	0.459 7	0.946	0.015 6	0.356 3	0.357 1	0.948
γ_1	−0.2	0.003 4	0.775 0	0.782 4	0.941	0.001 4	0.672 6	0.678 8	0.942

表 3　在模型设置(Ⅱ)情形下正态分布设置的参数估计的模拟结果

参数	Para	BIAS	ESE	SSE	CP	BIAS	ESE	SSE	CP
		$n=400$				$n=600$			
$\tilde{\alpha}_0$	−0.3	−0.025 9	0.574 0	0.565 0	0.963	−0.015 7	0.451 4	0.446 1	0.960
α_1	−0.4	0.019 1	0.592 3	0.670 8	0.935	0.007 4	0.481 8	0.538 6	0.941
α_2	−0.2	0.020 4	0.415 0	0.485 7	0.924	0.013 5	0.364 1	0.423 2	0.937
$\tilde{\beta}_0$	0.4	−0.015 8	0.543 6	0.523 1	0.966	−0.012 4	0.493 2	0.486 0	0.954
β_1	0.2	−0.001 9	0.274 9	0.282 5	0.941	−0.000 9	0.258 4	0.262 4	0.946
β_2	0.2	−0.015 9	0.239 9	0.240 6	0.937	−0.014 1	0.182 2	0.182 5	0.948
$\tilde{\lambda}_0$	−0.5	−0.008 3	0.366 2	0.368 0	0.936	−0.005 5	0.304 8	0.305 0	0.940
λ_1	0.2	−0.007 7	0.674 1	0.684 9	0.924	−0.006 7	0.583 0	0.587 2	0.927
λ_2	−0.2	0.021 8	0.185 6	0.178 3	0.972	0.015 8	0.161 5	0.158 2	0.961
$\tilde{\gamma}_0$	0.3	−0.007 8	0.319 3	0.328 0	0.940	−0.005 4	0.259 7	0.261 8	0.948
γ_1	−0.2	0.014 7	0.546 1	0.553 0	0.936	0.013 2	0.501 5	0.508 7	0.938
γ_2	0.2	−0.020 7	0.165 4	0.167 9	0.938	−0.017 4	0.127 4	0.129 0	0.941
$\tilde{\alpha}_0$	0.3	−0.051 2	0.471 3	0.490 7	0.930	−0.025 6	0.232 8	0.249 5	0.926
α_1	−0.4	0.010 8	0.816 8	0.826 8	0.937	0.012 8	0.624 8	0.633 3	0.933
α_2	0.2	−0.011 5	0.367 1	0.388 8	0.920	−0.010 4	0.243 1	0.255 3	0.921

表 3(续)

参数	Para	BIAS	ESE	SSE	CP	BIAS	ESE	SSE	CP
		$n=400$				$n=600$			
$\tilde{\beta}_0$	−0.4	−0.016 2	0.417 5	0.406 3	0.938	−0.009 8	0.308 2	0.309 1	0.942
β_1	0.2	−0.031 2	0.891 0	0.801 7	0.963	−0.025 4	0.571 7	0.580 9	0.939
β_2	−0.2	0.011 1	0.418 5	0.399 5	0.976	0.009 2	0.263 5	0.259 7	0.977
$\tilde{\lambda}_0$	−0.5	0.019 0	0.277 8	0.283 3	0.926	0.014 9	0.189 9	0.191 5	0.940
λ_1	0.2	−0.008 6	0.419 0	0.398 4	0.937	−0.008 9	0.377 5	0.385 0	0.947
λ_2	−0.2	0.013 8	0.308 2	0.314 7	0.946	0.006 7	0.177 7	0.182 7	0.948
$\tilde{\gamma}_0$	0.3	−0.015 7	0.294 5	0.312 7	0.924	−0.005 2	0.204 9	0.206 2	0.940
γ_1	−0.2	0.024 8	0.539 9	0.482 4	0.972	0.013 2	0.370 1	0.367 6	0.965
γ_2	0.2	−0.016 7	0.233 4	0.214 8	0.957	−0.014 8	0.178 9	0.174 6	0.954

4 实例分析

本部分将提出的方法应用到血友病患者感染 HIV-1 的风险数据集[1]中,该数据集包含 368 名血友病患者. 为了分析血液药物制品Ⅷ浓缩物与血友病患者感染时间的关系,368 名血友病患者被分成 2 组,其中一组注射含有Ⅷ浓缩物的血液制品,另一组注射不含有Ⅷ浓缩物的血液药物制品. 在整个随访过程中,患者血液样品的收集和采集是间隔的,因此不能记录精确的 HIV-1 感染时间,只能记录患者采集血液样本的时间以及患者 HIV-1 感染的时间,因此该数据集是一个经典区间删失数据. 对于血友病患者 i,假设 T_i 表示患者 HIV-1 感染的时间,定义协变量 $X_{1i}=X_{2i}=X_{3i}=X_{4i}=0$ 表示血友病患者注射不含有Ⅷ浓缩物的血液药物制品,$X_{1i}=X_{2i}=X_{3i}=X_{4i}=1$ 表示血友病患者注射含有Ⅷ浓缩物的血液药物制品,分析结果如表 4 所示.

由表 4 可知,当检验协变量系数 $\alpha_1=0$,$\beta_1=0$,$\lambda_1=0$,$\gamma_1=0$ 时,对应的 p 值很小,几乎近似于 0,因此协变量对模型参数有着显著性影响. 2 个尺度参数的回归系数分别为 $\alpha_1=0.757\ 0$、$\beta_1=2.328\ 6$,且对应的标准差分别为 0.018 7、0.200 4,因此,与注射不含有Ⅷ浓缩物的血液药物制品相比,注射含有Ⅷ浓缩物的血液药物制品收缩了模型尺度参数的变化. 同时,2 个形状参数的回归系数分别为 $\lambda_1=3.798\ 1$、$\gamma_1=-0.699\ 1$,对应的标准差分别为 0.302 0、0.200 4,因此,注射含有Ⅷ浓缩物的血液药物制品收缩了模型形状参数 c,同时扩大了模型形状参数 d,且收缩程度高于扩大程度. 综合 4 参数 OLLGG 分布的性质以及上述回归结果可知,注射含有Ⅷ浓缩物的血液药物制品能在短时间内降低血友病患者感染 HIV-1 的风险,这与目前存在的单参数模型[1]获得的结论一致. 然而,随着时间的推移,这种效果会逐渐减弱.

表 4 HIV-1 感染风险数据的估计结果

参数	估计值	标准差	p 值
$\tilde{\alpha}_0$	−1.832 7	0.018 7	<0.0001
α_1	0.757 0	0.018 7	<0.0001
$\tilde{\beta}_0$	−3.400 1	0.201 3	<0.0001
β_1	2.328 6	0.200 4	<0.0001
$\tilde{\lambda}_0$	2.986 9	0.302 1	<0.0001
λ_1	3.798 1	0.302 0	<0.0001
$\tilde{\gamma}_0$	−2.400 4	0.201 3	<0.0001
γ_1	−0.699 1	0.200 4	0.0005

5 结论与展望

本文提出了在区间删失数据下基于 4 参数 OLLGG 分布的多参数回归模型,通过极大似然估计方法给出了模型的参数估计量,并通过 Fisher 信息矩阵计算了估计的标准差. 与其他模型相比,多参数回归不再是在标准风险函数回归形式下单一参数与协变量之间的回归关系,它提供了多种参数回归建模的思路,产生了灵活的回归模型. 此外,它也可以通过在失效时间模型假设下建立失效时间分布参数与协变量之间的模型来放松比例风险假设,弥补单一回归模型假设的不足. 因为该分布假设具有便利性,所以估计过程简单且更容易实现. 尽管本文使用了 4 参数 OLLGG 分布模型,但在实际问题分析中,数据本身的分布假设仍然存在假设偏离的

情形,因此可以通过其他分布形式替代 OLLGG 分布,建立新的多参数回归模型,提高模型拟合实际数据的效果. 随着数据形式的复杂,该种建模方法也可以推广到其他更加复杂的数据结构中或者应用于高维数据下的变量选择等. 值得注意的是:由于变量选择过程会涉及到 4 类参数的选择问题,所以计算过程会变得更加复杂一些.

6　参考文献

[1] SUN Jianguo. The statistical analysis of interval-censored failure time data [M]. New York:Springer,2006.

[2] ALIZADEH M,BENKHELIFA L,RASEKHI M,et al. The odd log-logistic generalized Gompertz distribution: properties,applications and different methods of estimation [J]. Communications in Mathematics and Statistics, 2020, 8(3):295-317.

[3] CHEN Yurong,LUO Ji,FENG Jie. Regression analysis of case II interval-censored data with auxiliary covariates [J]. Communications in Statistics: Theory and Methods, 2021,50(17):4022-4038.

[4] EL-GOHARY A, ALSHAMRANI A, AL-OTAIBI A N. The generalized Gompertz distribution [J]. Applied Mathematical Modelling,2013,37(1/2):13-24.

[5] HASHIMOTO E M,ORTEGA E M M,CANCHO V G. On estimation and diagnostics analysis in log-generalized Gamma regression model for interval-censored data [J]. Statistics,2013,47(2):379-398.

[6] PENG Defen,MACKENZIE G,BURKE K. A multiparameter regression model for interval-censored survival data [J]. Statistics in Medicine,2020,39(14):1903-1918.

[7] 赵波. 带潜变量的区间删失数据的两种联合建模方法研究 [D]. 长春:长春工业大学,2018.

[8] 张红. 加速失效回归模型基于广义指数分布的统计推断 [D]. 长春:长春工业大学,2017.

[9] HU Tao, ZHOU Qingning, SUN Jianguo. Regression analysis of bivariate current status data under the proportional hazards model [J]. Canadian Journal of Statistics, 2017,45(4):410-424.

[10] LIN Danyu, OAKES D, YING Zhiliang. Additive hazards regression with current status data [J]. Biometrika,1998, 85(2):289-298.

[11] MARTINUSSEN T,SCHEIKE T H. Efficient estimation in additive hazards regression with current status data [J]. Biometrika, 2002,89(3):649-658.

[12] ZENG Donglin, CAI Jianwen, SHEN Yu. Semiparametric additive risks model for interval-censored data [J]. Statistica Sinica,2006,16(1):287-302.

[13] WANG Lianming, SUN Jianguo, TONG Xingwei. Regression analysis of case II interval-censored failure time data with the additive hazards model [J]. Statistica Sinica. 2010,20(4):1709-1723.

[14] BETENSKY R A,RABINOWITZ D,TSIATIS A A. Computationally simple accelerated failure time regression for interval censored data [J]. Biometrika,2001,88(3):703-711.

[15] BURKE K, MACKENZIE G. Multi-parameter regression survival modeling: an alternative to proportional hazards [J]. Biometrics,2017,73(2):678-686.

[16] GLEATON J U,LYNCH J. Properties of generalized log-logistic families of lifetime distributions [J]. Journal of Probability and Statistical Science,2006,4(1):51-64.

[17] 茆诗松,王静龙,濮晓龙. 高等数理统计[M]. 2 版. 北京:高等教育出版社,2006.

The Estimation of the Multiple Parameter Regression Model with Interval-Censored Data Based on the Odd Log-Logistic Generalized Gompertz Distribution

WANG Shuying,ZHANG He′nan,ZHAO Bo,CHENG Yunfei

(School of Mathematics and Statistics,Changchun University of Technology,Changchun Jilin 130012,China)

Abstract: The multi-parameter regression model is made to describe the relationship between distribution parameters and covariates through linear regression based on case II interval censoring failure time data and OLLGG distribution. The parameter estimation is given by using the maximum likelihood method and the numerical simulation shows that the estimator has good properties. The proposed model is applied to HIV infection data set in hemophilia patients,and it is found that the proposed model has a more flexible fitting effect.

Key words: multiple parameter regression model; case II interval-censored data; odd log-logistic generalized Gompertz distribution; maximum likelihood estimation

(责任编辑:曾剑锋)