DOI: 10.16452/j.cnki.sdkjzk.2024.01.008 文章编号:1672-3767(2024)01-0073-09

区间删失数据下 Weibull 比例优势模型的参数估计

王淑影,郭祥道,李红伟,赵 波
(长春工业大学 数学与统计学院,吉林 长春 130012)

摘 要:基于区间删失数据建模是当前复杂数据分析的热点之一。本研究在两类区间删失数据下建立 Weibull 比例优势模型,基于极大似然估计给出了模型参数,进一步讨论了估计量的渐近性质。数值模拟验证了模型参数的估计效果,并将提出的模型及方法应用到艾滋病临床试验数据和肺肿瘤试验数据中,给出了感兴趣事件的生存函数曲线,通过与生存函数的非参数极大似然估计比较,表明所提方法具有较好的拟合效果。

关键词:区间删失数据;比例优势模型;Weibull 分布;极大似然估计
中图分类号:O212 **文献标志码**:A

Parameter estimation of Weibull proportional odds model under interval-censored data

WANG Shuying, GUO Xiangdao, LI Hongwei, ZHAO Bo
(School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, China)

Abstract: Modeling based on interval-censored data is one of the hotspots of complex data analysis at present. In this paper, the Weibull proportional odds model was established under the two types of interval-censored data. The parameter estimation of the model was given based on the maximum likelihood estimation and the asymptotic normality of the estimator was further discussed. Numerical simulation verified the estimation effect of model parameters. Then the proposed model was applied to AIDS clinical trial data and lung tumor experimental data, and the smooth survival function curve was drawn. Compared with the nonparametric maximum likelihood estimation of survival function, the proposed method has been proven to have better fitting effect.

Key words: interval-censored data; proportional odds model; Weibull distribution; maximum likelihood estimation

区间删失数据常出现于生物医学、工业工程学、经济学、人口统计学等众多交叉领域中[1],分析及推断此类数据的内在规律是发展学科交叉融合的重要基础和前提。尽管比例风险模型在这类数据的分析处理中有着极大的优势,但是许多实际问题中,风险率函数经常会收敛为一个常数,并不满足比例风险模型的假设,Bennett 建立的比例优势模型,能很好地解决这类问题,较比例风险模型更有优势[2-4]。

许多学者研究了区间删失数据下各种模型的估计问题。文献[5-6]讨论了Ⅰ型区间删失数据半参数比例风险模型的估计问题以及估计结果渐近正态性;文献[7]讨论了Ⅰ型区间删失数据下线性变换模型的估计问题;文献[8-9]在Ⅰ型区间删失数据下建立加性风险模型,并证明了估计的大样本性质;文献[10]提出Ⅰ型区间删失数据(现状数据)下含有潜变量的加性风险回归模型,给出了两步估计方法;文献[11]讨论了Ⅰ型区间删失数据下广义极值分布回归的贝叶斯估计;文献[12-14]基于不同的估计方法建立Ⅱ型区间删失数据的加性风险模型;文献[15]基于比例风险模型提出Ⅱ型区间删失数据下的 BAR 惩罚变量选择方法,证明了估计的 Orcale 性质;文献[16]考虑了含有治愈亚组的区间删失数据的变量选择方法等。然而,已有文献主要

收稿日期:2022-09-02
基金项目:国家自然科学基金青年基金项目(11901054);中国博士后科学基金项目(2021M700536);吉林省自然科学基金优秀青年基金项目(20230101371JC)
作者简介:王淑影(1990—),女,吉林榆树人,副教授,博士,主要从事生物统计研究. E-mail:wangshuying0601@163.com

集中于传统风险模型假设,感兴趣事件发生时间的危险率满足乘积假设或者加法假设。但在实际问题中,这些假设可能不成立,例如危险率函数收敛到常数时,经典模型的风险函数假设并不满足。因此,本研究充分考虑Ⅰ型区间删失数据与Ⅱ型区间删失数据之间的关系,应用参数模型假设的便利性,建立 Weibull 比例优势模型,并基于极大似然估计给出两种数据类型下参数的估计量。

1 数据及符号

假设变量 T 表示感兴趣事件发生的时间,变量 $\boldsymbol{Z}$ 表示 p 维度的协变量。在有限的资源下,为了记录受试个体的更多信息,研究人员在实验中对受试者进行两次检测 U 和 $V(U\leqslant V)$,感兴趣事件发生的时间会出现 3 种情况:①感兴趣事件发生在第一次检测时间 U 之前,即 $T\leqslant U$;②感兴趣事件发生在两次检测时间 U 和 V 的中间,即 $U<T\leqslant V$;③感兴趣事件发生在第二次检测时间 V 以后,这类数据称为Ⅱ型区间删失数据[1]。数据结构可表示为:

$$\{U,V,\boldsymbol{Z},\delta_1=I(T\leqslant U),\delta_2=I(U<T\leqslant V),\delta_3=1-\delta_1-\delta_2\}。\tag{1}$$

式中:I 为示性函数,δ_1、δ_2、δ_3 为示性变量。

当所有受试个体只在时间 C 被观测一次,则Ⅱ型区间删失数据退化为Ⅰ型区间删失数据(或者现状数据),即 $C=U=V$,此时,所记录到的感兴趣事件发生时间只知道发生在某个观测时间 C 之前或之后,此时数据结构表示为:

$$\{C,\boldsymbol{Z},\Delta\}。\tag{2}$$

式中,示性变量 $\Delta=I(T\leqslant C)$。

2 Weibull 比例优势模型

同比例风险模型类似,比例优势模型也是刻画感兴趣事件发生的时间与协变量关系的模型,但两个模型有所不同。比例风险模型假设个体的危险率是成比例的,而比例优势模型假设个体危险率收敛到常数,能很好地补充比例风险模型的假设缺陷[3,17]。因此本研究将在区间删失数据下研究比例优势模型。

假设给定协变量 $\boldsymbol{Z}$ 时,失效时间 T 与观测过程条件独立,则比例优势模型可表示为[2,17-18]:

$$\frac{F(t\mid\boldsymbol{Z})}{1-F(t\mid\boldsymbol{Z})}=\Lambda_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}),\tag{3}$$

或

$$\mathrm{logit}(F(t\mid\boldsymbol{Z}))=\mathrm{logit}(F_0(t))+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}。$$

式中:未知参数 $\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$ 为回归系数向量;$\Lambda_0(t)$ 为单调递增的基准比率函数,且 $\Lambda_0(t)=\dfrac{F_0(t)}{1-F_0(t)}=\dfrac{1-S_0(t)}{S_0(t)}$;$F_0(t)$ 为 $\boldsymbol{Z}=0$ 时失效时间 T 的累积分布函数;$S_0(t)$ 为 $\boldsymbol{Z}=0$ 时失效时间 T 的生存函数;函数 $\mathrm{logit}(s)=\dfrac{s}{1-s}$。因此在给定的协变量 $\boldsymbol{Z}$,生存时间 T 在比例优势模型的假设下,条件分布函数为:

$$F(t\mid\boldsymbol{Z})=\frac{\Lambda_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})}{1+\Lambda_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})}。\tag{4}$$

同理可得,生存时间 T 在比例优势模型的假设下的生存函数为:

$$S(t\mid\boldsymbol{Z})=\frac{1}{1+\Lambda_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})}。$$

在模型中非参数比例优势函数 $\Lambda_0(t)$ 的估计是困难的,考虑一种参数分布模型可以简化估计过程。Weibull 分布是当前数据中主要的参数模型之一,其分布特征形状参数与尺度参数灵活性在可靠性分析以及生存分析中有着极其重要的作用[19]。综上,讨论 Weibull 分布下区间删失数据的建模并解释数据蕴含的实际规律,能极大提高估计的效率。

本研究提出将 Weibull 分布融入经典的比例优势模型中,即将式(4)中的基准比率函数 $\Lambda_0(t)=$

$\frac{F_0(t)}{1-F_0(t)}$中的分布数,替换为 Weibull 分布。具体地,假设 $F_0(t)$是尺度参数 $\alpha(\alpha>0)$、形状参数 $\gamma(\gamma>0)$的 Weibull 分布的分布函数[17],有

$$F_0(t)=1-\exp(-\alpha t^{\gamma}), \tag{5}$$

生存函数为:

$$S_0(t)=1-F_0(t)=\exp(-\alpha t^{\gamma})。$$

特别地,当形状参数 $\gamma=1$ 时,Weibull 分布退化为指数分布。

基于式(4)和式(5),建立 Weibull 比例优势模型为:

$$\frac{F(t|\boldsymbol{Z})}{1-F(t|\boldsymbol{Z})}=\frac{1-\exp(-\alpha t^{\gamma})}{\exp(-\alpha t^{\gamma})}\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})=(\exp(\alpha t^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})。$$

分布函数重新写作:

$$F(t|\boldsymbol{Z})=\frac{(\exp(\alpha t^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})}{1+(\exp(\alpha t^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})}, \tag{6}$$

或将 Weibull 比例优势模型写作线性形式,即

$$\operatorname{logit}(F(t|\boldsymbol{Z}))=\operatorname{logit}(1-\exp(-\alpha t^{\gamma}))+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}。 \tag{7}$$

3 区间删失数据下模型的参数估计

在区间删失数据结构下,假设 $\boldsymbol{D}_i=\{U_i,V_i,\boldsymbol{Z}_i,\delta_{i1},\delta_{i2},\delta_{i3}\}$, $i=1,2,\cdots,n$ 是观测样本,设 T_i 表示第 i 个个体的失效时间。在给定协变量 $\boldsymbol{Z}_i$ 时,观测时间 U_i、V_i 与失效时间 T_i 独立,则在Ⅱ型区间删失数据下 Weibull 比例优势模型的似然函数可表示为:

$$\begin{aligned}L(\alpha,\gamma,\boldsymbol{\beta})=&\prod_{i=1}^{n}F(U_i\mid\boldsymbol{Z}_i)^{\delta_{i1}}[F(V_i\mid\boldsymbol{Z}_i)-F(U_i\mid\boldsymbol{Z}_i)]^{\delta_{i2}}[1-F(V_i\mid\boldsymbol{Z}_i)]^{\delta_{i3}}\\=&\prod_{i=1}^{n}\left[\frac{(\exp(\alpha U_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}{1+(\exp(\alpha U_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}\right]^{\delta_{i1}}\times\\&\prod_{i=1}^{n}\left[\frac{(\exp(\alpha V_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}{1+(\exp(\alpha V_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}-\frac{(\exp(\alpha U_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}{1+(\exp(\alpha U_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}\right]^{\delta_{i2}}\times\\&\prod_{i=1}^{n}\left[\frac{1}{1+(\exp(\alpha V_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}\right]^{\delta_{i3}}。\end{aligned} \tag{8}$$

对数似然函数为:

$$\begin{aligned}l(\alpha,\gamma,\boldsymbol{\beta})=&\sum_{i=1}^{n}\delta_{i1}\lg\left[\frac{(\exp(\alpha U_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}{1+(\exp(\alpha U_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}\right]+\\&\sum_{i=1}^{n}\delta_{i2}\lg\left[\frac{(\exp(\alpha V_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}{1+(\exp(\alpha V_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}-\frac{(\exp(\alpha U_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}{1+(\exp(\alpha U_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}\right]+\\&\sum_{i=1}^{n}\delta_{i3}\lg\left[\frac{1}{1+(\exp(\alpha V_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}\right]。\end{aligned} \tag{9}$$

在实际问题中,Ⅱ型区间删失数据的一个特殊例子是Ⅰ型区间删失数据或者现状数据。该类型数据表示对每个个体 i 仅观测一次,此时观测样本表示为 $\boldsymbol{D}_i=\{C_i,\boldsymbol{Z}_i,\Delta_i\}$, $i=1,2,\cdots,n$,Weibull 比例优势模型的似然函数为:

$$L(\alpha,\gamma,\boldsymbol{\beta})=\prod_{i=1}^{n}F(C_i\mid\boldsymbol{Z}_i)^{\Delta_i}(1-F(C_i\mid\boldsymbol{Z}_i))^{1-\Delta_i}=\prod_{i=1}^{n}\frac{[(\exp(\alpha C_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)]^{\Delta_i}}{1+(\exp(\alpha C_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)}, \tag{10}$$

式中:C_i 表示第 i 个个体的观测时间,示性函数 $\Delta_i=I(T_i\leqslant C_i)$。

对数似然函数为:

$$l(\alpha,\gamma,\boldsymbol{\beta})=\sum_{i=1}^{n}\Delta_i[\lg((\exp(\alpha C_i^{\gamma})-1))+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i]-\sum_{i=1}^{n}\lg[1+(\exp(\alpha C_i^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)]。 \tag{11}$$

为了估计模型参数 $\boldsymbol{\theta}=(\alpha,\gamma,\boldsymbol{\beta})$,最直观的方法就是直接极大化对数似然函数 $l(\alpha,\gamma,\boldsymbol{\beta})$,从而获得参数的极大似然估计 $\hat{\boldsymbol{\theta}}=(\hat{\alpha},\hat{\gamma},\hat{\boldsymbol{\beta}})$。具体地,可以通过求解得分方程组 $U(\boldsymbol{\theta})=(U(\alpha),U(\gamma),U(\boldsymbol{\beta}))^{\mathrm{T}}$ 获得参数解 $\hat{\boldsymbol{\theta}}$,其中:

$$U(\alpha)=\frac{\partial l(\alpha,\gamma,\boldsymbol{\beta})}{\partial\alpha}=0,U(\gamma)=\frac{\partial l(\alpha,\gamma,\boldsymbol{\beta})}{\partial\gamma}=0,U(\boldsymbol{\beta})=\frac{\partial l(\alpha,\gamma,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=0。\tag{12}$$

为了讨论 Weibull 比例优势模型极大似然估计量的极限性质,设 $\boldsymbol{\theta}_0=(\alpha_0,\gamma_0,\boldsymbol{\beta}_0)$ 表示模型的真实参数,$\hat{\boldsymbol{\theta}}=(\hat{\alpha},\hat{\gamma},\hat{\boldsymbol{\beta}})$ 表示模型的极大似然估计量,$l(\boldsymbol{\theta})=l(\alpha,\gamma,\boldsymbol{\beta})$,下面是一些正则性条件。

条件 1) 对每个样本 i,存在一个 $\varepsilon_i>0$,使得 $P(V_i-U_i>\varepsilon_i)=1$。

条件 2) 参数 $\boldsymbol{\theta}\in R^+\times R^+\times R^p$,且对数似然函数 $l(\boldsymbol{\theta})$ 的一、二、三阶导数存在。

条件 3) 在参数真值 $\boldsymbol{\theta}_0$,有 $\mathrm{E}_{\boldsymbol{\theta}_0}[U(\boldsymbol{\theta}_0)]=0$,矩阵 $\mathrm{E}_{\boldsymbol{\theta}_0}[U(\boldsymbol{\theta}_0)U(\boldsymbol{\theta}_0)^{\mathrm{T}}]$ 正定。

条件 4) 协变量 $\boldsymbol{Z}$ 有界,即存在常数 $z^0>0$,满足 $P(|\boldsymbol{Z}|<z^0)=1$。

定理 1 假设条件 1)~4)成立,当 $n\to\infty$ 时,$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ 以分布收敛到正态分布 $N(\boldsymbol{0},\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0))$,亦即:$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{L}N(\boldsymbol{0},\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0))$。其中,$\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)$ 表示 Fisher 信息矩阵 $\boldsymbol{I}(\boldsymbol{\theta}_0)$ 的逆矩阵,$\boldsymbol{I}(\boldsymbol{\theta}_0)=\mathrm{E}_{\boldsymbol{\theta}_0}\left(\frac{\partial l(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0}\left(\frac{\partial l(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0}\right)^{\mathrm{T}}\right)$。

证明:在假设条件 1)~3)满足时,根据参考文献[21]的定理 2.13 可知 $\hat{\boldsymbol{\theta}}$ 是 $\boldsymbol{\theta}_0$ 的相合估计。在 Weibull 比例优势模型假设下可知,删失数据下对数似然函数是连续可微的,且满足得分函数为零,为了证明渐近正态性,将对数似然函数在真实参数处进行 Taylor 展开,有

$$\frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=\frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\theta_0}+(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\bigg|_{\theta_0}+o_p(1)。$$

又因为在独立样本假设下,$l(\boldsymbol{\theta})$、$\frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$、$\frac{\partial l^2(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}$ 为独立同分布的随机变量的和,则由中心极限定理可以证明:

$$\sqrt{n}\frac{1}{n}\frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\theta_0}\xrightarrow{L}N(\boldsymbol{0},\boldsymbol{I}(\boldsymbol{\theta}_0))。$$

注意到:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=\frac{-\sqrt{n}\dfrac{\partial l(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\bigg|_{\theta_0}}{\dfrac{\partial^2 l(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}^2}\bigg|_{\theta_0}+o_p(1)},$$

结合正则性条件 1)~4),即可获得 $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{L}N(\boldsymbol{0},\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0))$,$I(\boldsymbol{\theta}_0)=-E_{\boldsymbol{\theta}_0}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0^{\mathrm{T}}}l(\boldsymbol{\theta}_0)\right)$。其中:位于 (k,j) 处的元素为 $I_{kj}=-\mathrm{E}_{\boldsymbol{\theta}_0}\left(\frac{\partial^2}{\partial\theta_{0k}\partial\theta_{0j}}l(\boldsymbol{\theta}_0)\right)$,$k,j=1,2,\cdots,p+2$;$\theta_{0k}$、$\theta_{0j}$ 分别为参数 $\boldsymbol{\theta}_0$ 的第 k、j 个元素。

4 数值模拟

为验证所提方法的有效性,分别对Ⅱ型区间删失数据和Ⅰ型区间删失数据进行数值模拟。同时,对于协变量 $\boldsymbol{Z}$ 考虑三种形式:① $\boldsymbol{Z}$ 服从成功概率为 0.5 的伯努利分布;②$\boldsymbol{Z}$ 服从标准正态分布;③$\boldsymbol{Z}$ 是二维协变量,第一个分量协变量服从标准正态分布,第二个分量协变量服从成功概率为 0.5 的伯努利分布。在式(6)的假设下,基于逆变换方法反解生成失效时间 T,即在区间(0,1)生成均匀分布 L,令

$$F(T|\boldsymbol{Z})=\frac{(\exp(\alpha t^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})}{1+(\exp(\alpha t^{\gamma})-1)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})}=L, \tag{13}$$

可得:

$$T=\left[-\frac{1}{\alpha}\lg\left(\frac{(1-L)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})}{L+(1-L)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})}\right)\right]^{\frac{1}{\gamma}}。 \tag{14}$$

对Ⅱ型区间删失数据,为了生成观测时间,本研究首先分别在区间(a_u,b_u)和区间(a_v,b_v)($b_u\leqslant a_v$)生成服从均匀分布的观测时间 U 和 V,比较失效时间 T 和观测时间 U 和 V 的大小,生成示性变量 $\delta_1=I(T\leqslant U)$,$\delta_2=I(U<T\leqslant V)$和 $\delta_3=1-\delta_1-\delta_2$。协变量分别在一维和二维情况下,不同参数真值设置时,样本容量 n 分别为 200 和 600,并重复实验 1 000 次得到一维(表 1)和二维(表 2)的结果,包括模型参数的真实值、平均估计值减去真实值的偏差、估计值的样本标准差、估计值的样本标准差均值以及重复实验 1 000 次所得到的经验覆盖概率。

从表 1 和表 2 可以看出,估计的偏差较小,接近于 0,样本标准差接近于估计值的样本标准差均值,覆盖概率接近 0.95。随着样本量的增加,效果更加明显。因此,本研究方法获得估计量是相合的、渐近有效的。

为了更好地验证所提出方法的有效性,下面再讨论Ⅱ型区间删失数退化为Ⅰ型区间删失数据下的例子。为生成区间Ⅰ型区间删失数据,首先在区间(a_c,b_c)生成均匀 C 作为观测时间,通过调整 a_c、b_c 的取值控制左删失与右删失的比例。比较失效时间 T 与观测时间 C 生成示性变量 $\Delta=I(T\leqslant C)$。表 3 和表 4 分别给出协变量在一维和二维情形下,在不同参数真值设置时,样本容量 200 和 600 时的模拟结果。结果表明,在Ⅰ型区间删失数据下,得到结论与Ⅱ型区间删失数据的结论是一致的,获得的估计量仍然是相合的、渐近有效的。

表 1 Ⅱ型区间删失数据下单个协变量时的模拟结果

Table 1 Simulation results of single covariate under Case Ⅱ interval-censored data

分量	真实值	样本量 $n=200$				样本量 $n=600$			
		偏差	标准差	标准差均值	覆盖概率	偏差	标准差	标准差均值	覆盖概率
第一分量	$\alpha=0.8$	0.002 1	0.070 6	0.072 0	0.960	−0.000 5	0.040 6	0.041 3	0.956
	$\gamma=0.8$	0.008 8	0.071 5	0.070 5	0.953	0.004 8	0.041 4	0.040 3	0.948
	$\beta=0.5$	0.011 1	0.149 6	0.147 1	0.945	0.003 1	0.081 5	0.084 1	0.954
	$\alpha=0.8$	−0.000 3	0.075 6	0.073 2	0.940	−0.000 3	0.041 7	0.042 1	0.947
	$\gamma=1.5$	0.025 8	0.144 3	0.141 7	0.948	0.009 0	0.080 4	0.080 6	0.954
	$\beta=-1.5$	−0.024 4	0.198 7	0.199 7	0.956	−0.007 4	0.113 8	0.113 7	0.954
	$\alpha=1.0$	0.004 0	0.088 5	0.087 2	0.952	0.001 9	0.048 5	0.050 1	0.965
	$\gamma=1.0$	0.014 5	0.091 3	0.091 3	0.962	0.002 9	0.051 1	0.051 9	0.952
	$\beta=1.0$	0.024 3	0.171 4	0.166 6	0.951	0.004 9	0.094 8	0.094 8	0.955
第二分量	$\alpha=0.8$	0.007 3	0.105 7	0.102 6	0.943	0.001 9	0.059 1	0.058 7	0.946
	$\gamma=0.8$	0.007 5	0.075 0	0.073 4	0.947	0.001 1	0.041 7	0.041 9	0.952
	$\beta=0.5$	0.001 8	0.287 5	0.284 3	0.953	−0.002 9	0.161 5	0.163 0	0.946
	$\alpha=0.8$	0.000 8	0.101 0	0.103 4	0.956	0.001 3	0.060 2	0.059 6	0.944
	$\gamma=1.5$	0.019 5	0.132 5	0.132 9	0.954	0.001 0	0.076 8	0.075 7	0.947
	$\beta=-1.5$	−0.012 7	0.323 1	0.317 1	0.946	−0.005 0	0.176 6	0.181 0	0.954
	$\alpha=1.0$	0.008 8	0.126 0	0.126 8	0.949	0.005 1	0.076 6	0.072 6	0.944
	$\gamma=1.0$	0.010 0	0.095 6	0.093 7	0.951	0.006 3	0.053 4	0.053 7	0.960
	$\beta=1.0$	−0.000 9	0.297 2	0.296 1	0.955	0.006 2	0.170 7	0.170 0	0.946

表 2 Ⅱ型区间删失数据下两个协变量时的模拟结果

Table 2 Simulation results for two covariates under Case Ⅱ interval-censored data

真实值	样本量 $n=200$				样本量 $n=600$			
	偏差	标准差	标准差均值	覆盖概率	偏差	标准差	标准差均值	覆盖概率
$\alpha=0.8$	0.008 5	0.108 4	0.107 0	0.954	0.000 6	0.060 3	0.061 0	0.959
$\gamma=0.8$	0.019 8	0.082 7	0.078 6	0.944	0.003 2	0.043 7	0.044 2	0.953
$\beta_1=0.5$	0.017 9	0.154 7	0.151 5	0.958	0.003 6	0.086 1	0.085 9	0.949
$\beta_2=1.5$	0.041 3	0.323 1	0.322 1	0.961	0.012 0	0.179 3	0.182 7	0.959
$\alpha=0.8$	0.005 3	0.102 4	0.105 8	0.956	0.001 4	0.061 1	0.060 5	0.946
$\gamma=0.8$	0.014 5	0.078 5	0.079 3	0.965	0.004 2	0.042 9	0.044 9	0.959
$\beta_1=-1.5$	−0.026 7	0.207 7	0.202 2	0.955	−0.011 4	0.114 5	0.114 8	0.956
$\beta_2=0.5$	0.011 6	0.314 4	0.313 5	0.948	0.001 6	0.180 3	0.178 8	0.951
$\alpha=1.0$	0.014 3	0.132 2	0.130 1	0.946	0.002 0	0.072 7	0.074 3	0.948
$\gamma=1.0$	0.014 2	0.097 8	0.096 6	0.958	0.005 2	0.054 9	0.055 0	0.951
$\beta_1=1.0$	0.020 6	0.170 0	0.170 7	0.954	0.006 3	0.094 2	0.097 1	0.959
$\beta_2=1.0$	0.005 6	0.318 3	0.309 7	0.950	0.006 8	0.171 3	0.177 0	0.953

表 3 Ⅰ型区间删失数据下单个协变量时的模拟结果

Table 3 Simulation results of single covariate under Case Ⅰ interval-censored data

分量	真实值	样本量 $n=200$				样本量 $n=600$			
		偏差	标准差	标准差均值	覆盖概率	偏差	标准差	标准差均值	覆盖概率
第一分量	$\alpha=0.8$	0.010 9	0.160 1	0.160 9	0.944	0.007 3	0.092 6	0.091 9	0.951
	$\gamma=1.5$	0.030 6	0.483 7	0.473 9	0.945	−0.003 8	0.266 2	0.270 3	0.951
	$\beta=0.8$	0.029 6	0.202 2	0.195 8	0.947	0.008 3	0.110 6	0.110 7	0.958
	$\alpha=0.8$	0.017 6	0.198 9	0.190 9	0.939	0.003 1	0.106 6	0.107 8	0.948
	$\gamma=1.5$	0.035 9	0.516 0	0.513 4	0.956	0.014 1	0.288 4	0.294 1	0.952
	$\beta=-1.5$	−0.038 3	0.246 0	0.255 7	0.955	−0.009 0	0.147 0	0.145 4	0.946
	$\alpha=1.5$	0.018 2	0.168 5	0.164 9	0.953	0.008 4	0.094 0	0.093 4	0.946
	$\gamma=0.8$	0.013 6	0.326 6	0.328 8	0.955	−0.005 5	0.192 1	0.187 6	0.948
	$\beta=-1.5$	−0.044 0	0.264 1	0.256 4	0.965	−0.007 5	0.142 9	0.144 2	0.953
第二分量	$\alpha=0.8$	0.001 8	0.134 2	0.130 0	0.945	0.002 4	0.076 5	0.074 7	0.949
	$\gamma=1.5$	0.041 4	0.455 3	0.438 9	0.944	0.012 8	0.250 7	0.248 7	0.952
	$\beta=0.8$	0.033 8	0.338 5	0.339 2	0.946	−0.002 3	0.196 7	0.196 7	0.945
	$\alpha=0.8$	0.024 6	0.203 7	0.196 6	0.949	0.009 0	0.111 5	0.111 3	0.951
	$\gamma=1.5$	0.015 0	0.297 9	0.289 7	0.948	0.003 6	0.167 4	0.166 0	0.952
	$\beta=-1.5$	−0.035 3	0.313 7	0.321 5	0.955	−0.015 8	0.178 7	0.183 1	0.956
	$\alpha=1.5$	0.047 5	0.274 8	0.274 9	0.955	0.013 0	0.152 6	0.154 3	0.952
	$\gamma=0.8$	0.005 7	0.362 5	0.356 5	0.947	0.007 1	0.197 3	0.205 1	0.950
	$\beta=-1.5$	−0.047 6	0.356 9	0.362 1	0.951	−0.013 2	0.204 6	0.205 1	0.954

表 4 Ⅰ型区间删失数据下两个协变量时的模拟结果

Table 4 Simulation results of two covariates under Case Ⅰ interval-censored data

真实值	样本量 $n=200$				样本量 $n=600$			
	偏差	标准差	标准差均值	覆盖概率	偏差	标准差	标准差均值	覆盖概率
$\alpha=1.5$	0.046 5	0.250 1	0.243 4	0.951	0.015 7	0.135 2	0.136 0	0.950
$\gamma=1.5$	0.038 6	0.329 3	0.316 8	0.953	0.003 7	0.177 3	0.179 0	0.956
$\beta_1=0.5$	0.015 1	0.202 3	0.193 3	0.938	0.006 6	0.110 0	0.108 9	0.952
$\beta_2=1.5$	0.030 9	0.401 8	0.390 6	0.962	0.001 3	0.215 3	0.220 7	0.955
$\alpha=1.5$	0.045 1	0.248 1	0.234 3	0.949	0.022 0	0.134 9	0.132 0	0.950
$\gamma=1.5$	0.025 6	0.377 3	0.367 1	0.941	0.002 0	0.209 1	0.208 7	0.949
$\beta_1=1.5$	0.048 2	0.272 7	0.264 2	0.954	0.026 9	0.154 2	0.148 9	0.951
$\beta_2=-0.5$	−0.039 2	0.409 5	0.390 6	0.941	−0.015 0	0.230 4	0.221 1	0.946
$\alpha=1.5$	0.046 1	0.260 2	0.250 1	0.950	0.011 7	0.144 4	0.139 7	0.944
$\gamma=1.5$	0.069 6	0.353 6	0.357 5	0.966	0.007 1	0.201 8	0.201 2	0.948
$\beta_1=1.5$	0.061 9	0.277 6	0.272 6	0.963	0.019 5	0.153 9	0.151 9	0.955
$\beta_2=1.5$	0.067 0	0.416 5	0.428 7	0.965	0.016 2	0.246 1	0.240 7	0.947

5 实例分析

将上述模型分别应用到 Gómez 等[20]研究的艾滋病临床试验数据和 Sun[1]介绍的肺肿瘤实验数据中,验证所提方法的有效性。

5.1 艾滋病临床试验数据

艾滋病临床试验数据集[17]包括 271 名艾滋病患者,分为初始核糖核酸(ribonucleic acid,RNA)病毒拷贝数量大于 20 000 copies/mL 和初始 RNA 病毒拷贝数量低于 20 000 copies/mL 的两个分组。为了探索患者的初始 RNA 病毒拷贝数量对于患者 RNA 病毒拷贝数量首次低于 500 copies/mL 时的影响,每位患者进行不同次数的血样采集,次数最多的采集记录 8 次,最少采集记录 1 次。根据患者观察记录的时间,可获得患者 RNA 病毒拷贝数量首次低于 500 copies/mL 时的观测区间。若第一次观测时发现 RNA 病毒拷贝数量 500 copies/mL 以下,则此患者数据可转化为左删失数据;若 RNA 病毒拷贝数量在 500 copies/mL 以下出现在某两次观测之间,则此患者数据可转化为区间删失数据;若在最后一次观测 RNA 病毒拷贝数量在 500 copies/mL 以上,则此患者数据为右删失数据,因此数据是Ⅱ型区间删失数据。

假设 T_i 表示实验中患者 RNA 病毒拷贝数量首次低于 500 copies/mL 时的时间;定义协变量 Z_i,若 $Z_i=0$,则表示为初始 RNA 病毒拷贝数量低于 20 000 copies/mL 的患者,反之 $Z_i=1$,则表示为初始 RNA 病毒拷贝数量高于 20 000 copies/mL 的患者。分析结果如表 5 所示。

表 5 艾滋病临床试验数据估计结果

Table 5 Estimation results of AIDS clinical trial data

参数	估计值	标准误差	t 值	p 值
α	0.642 3	0.088 0	7.292	<0
γ	0.432 7	0.047 4	9.120	<0
β	−1.767 9	0.248 1	−7.125	<0

由表 5 可知,回归系数 β 的估计值 $\hat{\beta}=-1.767\ 9$,估计标准误差等于 0.248 1,检验回归系数 $\beta=0$ 的 p 值近似于 0。因此,初始 RNA 病毒拷贝数对 RNA 拷贝数下降到 500 copies/mL 以下的时间有显著性影响,初始 RNA 病毒拷贝数量较低的患者的 RNA 病毒拷贝数下降到 500 copies/mL 以下的时间比初始 RNA 病

毒拷贝数较高的患者早得多。这一结论与文献[1]的非参数极大似然方法给出的结果一致。本研究方法与文献[1]给出的生存函数曲线如图 1 所示。由图 1 可见,本研究提出的参数方法与 NPMLE 方法估计的结果一致,且获得的生存函数更为光滑。

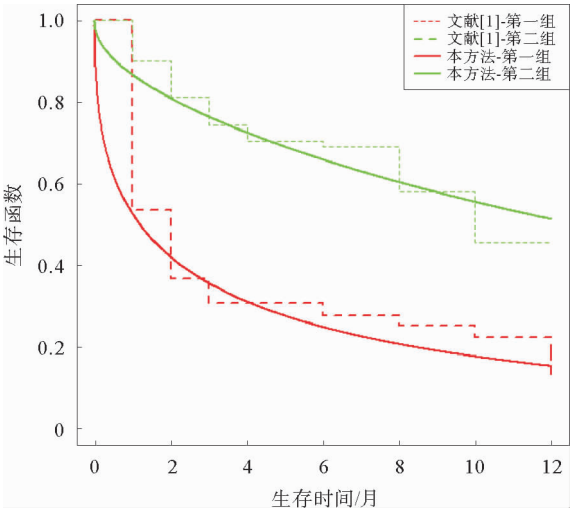

图 1 RNA 数据生存函数的估计

Fig. 1 Estimation of RNA data survival function

5.2 肺肿瘤试验数据

肺肿瘤试验数据集[1]有 144 只小白鼠试验样本,以天为单位记录了每只小白鼠的准确死亡时间以及死亡时刻小白鼠的肺肿瘤发病情况,是一个经典Ⅰ型区间删失数据的例子。为了探索环境对肺肿瘤发病情况的影响,将 96 只小白鼠放置在常规环境,另外 48 只小白鼠放置在无菌环境。假设时间 T_i 表示小白鼠肺肿瘤发病时间,定义 $Z_i=0$ 表示小白鼠被放置在常规环境,$Z_i=1$ 表示小白鼠被放置在无菌环境。分析结果如表 6 和图 2 所示。

表 6 肺肿瘤感染数据的估计结果

Table 6 Estimation results of lung tumor infection Data

参数	估计值	标准误差	t 值	p 值
α	0.104 9	0.053 4	1.963	0.049 7
γ	2.260 2	0.820 8	2.753	0.005 9
β	1.138 0	0.485 6	2.343	0.019 1

由表 6 可以看出,回归系数 β 的估计值 $\hat{\beta}=1.138\ 0$,估计标准误差等于 0.485 6,对应的回归系数 $\beta=0$ 的检验 p 值为 0.019 1,因此环境对肺肿瘤感染时间有着显著影响。更具体地说,无菌环境下的小白鼠肺肿瘤发病率明显高于常规环境下的小白鼠肺肿瘤发病率。该分析结果与文献[11]的结论基本一致。

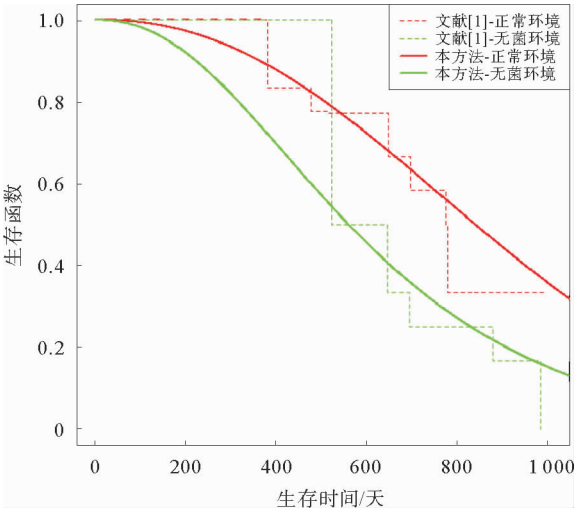

图 2 肺肿瘤发病时间的生存函数的估计

Fig. 2 Estimation of survival function of the onset time of lung tumor

6 结论

本研究提出Ⅰ型区间删失数据和Ⅱ型区间删失数据下的 Weibull 比例优势模型,通过极大似然估计获得了模型的参数估计,并讨论了估计的渐近性质。所提方法能较好地解决风险率函数收敛到常数的情况,同时方法简单易行。由于 Weibull 分布的灵活性,可以近似各种形态的风险函数,使建立的模型具有较强的灵活性。数值模拟结果表明估计是相合的、渐近有效的。艾滋病临床试验数据与肺肿瘤试验数据的分析结果表明所提方法具有较好的拟合效果。

参考文献:

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

[2] BENNETT S. Analysis of survival data by the proportional odds model[J]. Statistics in Medicine,1983,2(2):273-277.

[3] WANG L,WANG L M. Regression analysis of arbitrarily censored survival data under the proportional odds model[J]. Statistics in Medicine. 2021,40(16):3724-3739.

[4] WANG L,WANG L M. An EM algorithm for analyzing right-censored survival data under the semiparametric proportional odds model[J]. Communications in Statistics-Theory and Methods,2022,51(15):5284-5297.

[5] HUANG J. Efficient estimation for the proportional hazards model with interval censoring[J]. The Annals of Statistics, 1996,24(2):540-568.

[6] HU T,ZHOU Q N,SUN J G. Regression analysis of bivariate current status data under the proportional hazards model[J]. Canadian Journal of Statistics,2017,45(4):410-424.

[7] XU D,ZHAO S S,HU T,et al. Regression analysis of informative current status data with the semiparametric linear transformation model[J]. Journal of Applied Statistics,2019,46(2):187-202.

[8] LI S W,TIAN T,HU T,et al. A simulation-extrapolation approach for regression analysis of misclassified current status data with the additive hazards model[J]. Statistics in Medicine,2021,40(28):6309-6320.

[9] LI H Q,ZHANG H,SUN J G. Estimation of the additive hazards model with current status data in the presence of informative censoring[J]. Statistics and its Interface,2019,12(2):321-330.

[10] WANG C J,ZHAO B,LUO L L,et al. Regression analysis of current status data with latent variables[J]. Lifetime Data Analysis,2021,27(3):413-436.

[11] 孙烨,蒋京京,王纯杰. 广义极值回归模型下现状数据的贝叶斯估计[J]. 广西师范大学学报(自然科学版),2022,40(1):82-90.
SUN Ye,JIANG Jingjing,WANG Chunjie. Bayesian estimation of current status data with generalized extreme value regression model[J]. Journal of Guangxi Normal University (Natural Science Edition),2022,40(1):82-90.

[12] HE B H,LIU Y Y,WU Y S,et al. Semiparametric efficient estimation for additive hazards regression with case II interval-censored survival data[J]. Lifetime Data Analysis,2020,26(4):708-730.

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

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

[15] ZHAO H,WU Q W,LI G,et al. Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression[J]. Journal of the American Statistical Association,2020; 115(529):204-216.

[16] XU Y,ZHAO S S,HU T,et al. Variable selection for generalized odds rate mixture cure models with interval-censored failure time data[J]. Computational Statistics & Data Analysis,2021,156:107-115.

[17] MURPHY S A,ROSSINI A J,VAN DER VAART A W. Maximum likelihood estimation in the proportional odds model [J]. Journal of the American Statistical Association,1997,92(439):968-976.

[18] ZHU L,TONG X W,CAI D J,et al. Maximum likelihood estimation for the proportional odds model with mixed interval-censored failure time data[J]. Journal of Applied Statistics,2020,48(8):1496-1512.

[19] BDAIR O M,ABU AWWAD R R,ABUFOUDEH G K,et al. Estimation and prediction for flexible Weibull distribution based on progressive type II censored data[J]. Communications in Mathematics and Statistics,2020,8(3):255-277.

[20] GÓMEZ G,ESPINAL A,LAGAKOS S W. Inference for a linear regression model with an interval-censored covariate[J]. Statistics in Medicine,2003,22(3):409-425.

[21] 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,2006:119-120.

(责任编辑:齐敏华)