DOI：10.19789/j.1004-9398.2024.01.004
文献引用：刘娟，蔡婷婷，胡涛.左截断右删失数据下OLLGG分布的参数估计[J].首都师范大学学报（自然科学版），2024，45（1）：29-36. LIU J，CAI T T，HU T. Parameter estimation of the Odd Log-Logistic Generalized Gompertz distribution under left-truncated and right-censored data[J]. Journal of Capital Normal University（Natural Science Edition），2024，45（1）：29-36.

左截断右删失数据下OLLGG分布的参数估计*

刘　娟，蔡婷婷**，胡　涛**
（首都师范大学数学科学学院，北京　100048）

摘要：为了探索Odd Log-Logistic Generalized Gompertz（OLLGG）分布在生物学、社会学、计算机和市场营销等领域的广泛应用，本文研究左截断右删失数据下OLLGG分布的多参数回归模型，利用极大似然思想对参数进行估计。通过数值模拟，检验估计方法的有效性，并使用赤池信息量准则来选择适合的分布形式。最后，将本文的回归模型应用于channing house数据集，展示其良好的拟合效果。

关键词：左截断右删失；OLLGG分布；极大似然估计；多参数回归
中图分类号：O212.1　　**文献标识码**：A

Parameter estimation of the Odd Log-Logistic Generalized Gompertz distribution under left-truncated and right-censored data*

LIU Juan，CAI Tingting**，HU Tao**
（School of Mathematical Sciences，Capital Normal University，Beijing　100048）

Abstract：In order to explore the widespread applications of Odd Log-Logistic Generalized Gompertz（OLLGG）distribution in fields such as biology，sociology，computer science and marketing，this paper studies a multi-parameter regression model for OLLGG distribution under left-truncated and right-censored data，and estimates the parameters using maximum likelihood thinking. Through numerical simulation，the effectiveness of the estimation method is verified，and the Akaike information criterion is used to select the appropriate distribution form. Finally，the regression model proposed in this article is applied to the channing house dataset to demonstrate its good fitting performance.

Keywords：left-truncated and right-censored；OLLGG distribution；maximum likelihood estimation；multi-parameter regression
CLC：O212.1　　**DC**：A

收稿日期：2023-08-08
* 北京市自然科学基金项目（Z210003）
** 通信作者：2220501029@cnu.edu.cn；hutao@cnu.edu.cn

0 引 言

在医疗卫生、经济金融和可靠性工程等领域中,经常遇到左截断右删失数据,即只能观测到一部分事件发生的时间。左截断右删失数据的分析具有一定的挑战性。近年来,许多学者对该数据类型进行了深入的研究。文献[1-7]中利用期望最大化(expectation-maximum,EM)算法估计了左截断右删失数据的多种分布参数:Emura和Shiu[8]使用赤池信息量准则(Akaike information criterion,AIC)讨论了左截断右删失数据下威布尔分布、对数正态分布和指数分布的模型选择与参数估计问题。除此之外,还有很多文献探讨了左截断右删失数据下的半参数估计问题,如Kundu等[9]在假设潜在失效时间模型的前提下,对左截断右删失的竞争风险数据进行了分析;Burke和Mackenzie[10]从分布参数回归的角度,提出了新的多参数回归(multi-parameter regression,MPR)模型概念,以提高回归模型的灵活性和可解释性。

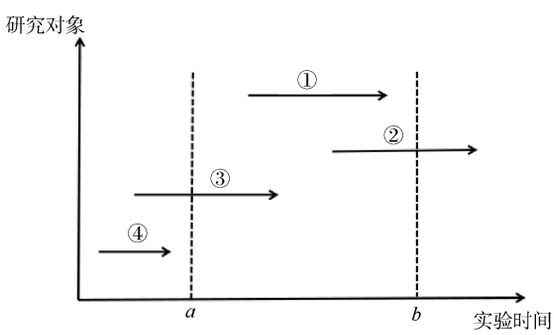
然而,现有的分布往往不能很好地拟合不同类型的风险函数,如常数风险、单调增加或减少风险以及浴缸形风险等。因此,近年来,许多学者提出了一些新的分布来扩展已有的分布,增加其参数和形态的多样性。如:Gleaton和Lynch[11]利用一种新的分布函数转换,提出了Odd Log-Logistic-Gompertz(OLLG)族;Valdemiro等[12]在此基础上,针对区间删失数据,提出了Generalized Odd Log-Logistic-Gompertz(GOLLG)分布;Prataviera等[13]基于OLLG族和威布尔分布提出了一种四参数的Generalized Odd Log-Logistic Flexible Weibull(GOLLFW)回归模型。在这些新的分布中,最具有代表性的是Alizadeh等[14]提出的Odd Log-Logistic Generalized Gompertz(OLLGG)分布,该分布具有以下特点:(1)有多个参数和丰富的形态,可以拟合多种风险函数;(2)可以灵活地转化为多种已知的分布,如指数分布、Gompertz分布等;(3)其多参数模型可以提高回归模型的灵活性和可解释性。

目前,针对左截断右删失数据的分析方法还不够完善,缺乏一个通用且有效的分布来描述该数据类型。由于OLLGG分布的灵活性,本文采用OLLGG分布的多参数模型对左截断右删失数据进行建模及预测。首先,给出OLLGG分布的多参数回归模型,推导其在左截断右删失数据下的似然函数和极大似然估计;其次,通过数值模拟验证估计方法的有效性和稳健性;最后,将所提出的模型应用于Channing House数据集,并与其他常用分布进行拟合优度检验。本文旨在探索OLLGG分布在处理左截断右删失数据方面的优势和适用性,并为生存数据分析提供一种新的思路和方法。

1 多参数OLLGG回归模型

1.1 生存数据

假设实验研究时间从点a开始,到点b实验结束,图1展示了生存数据可以分成以下4种类型。第1种类型是完全数据,即感兴趣的事件从开始到终结发生在a与b之间,而且能够观察到;第2种类型是数据发生了右删失,即直到实验研究结束后,感兴趣的事件仍在继续;第3种类型是数据发生了左截断,意味着感兴趣的事件在实验开始之前已经发生;第4种类型是完全左截断数据,即感兴趣的事件在实验开始之前已经结束了,这是左截断数据的特殊情况,由于无法收集到这种情况下的数据,所以本文主要关注前3种类型的截断删失数据。



注:①~④代表4种类型。

图1 生存数据类型

1.2 OLLGG分布

OLLGG分布是Alizadeh等[14]基于Generalized Gompertz(GG)分布提出的一类特别灵活的四参数分布。假设随机变量X服从GG分布,则分布函数表示为

$$G(x;a,b,c)=\left(1-\mathrm{e}^{\frac{-b}{a}\left(\mathrm{e}^{ax}-1\right)}\right)^{c},\ x>0,a>0,b>0,c>0, \tag{1}$$

式中:a、b为尺度参数,c为形状参数。

相应的随机变量X的密度函数表示为

$$g(x;a,b,c)=cb\mathrm{e}^{ax}\mathrm{e}^{\frac{-b}{a}\left(\mathrm{e}^{ax}-1\right)}\left(1-\mathrm{e}^{\frac{-b}{a}\left(\mathrm{e}^{ax}-1\right)}\right)^{c-1}, \tag{2}$$

当$c=1$时,GG分布退化为Gompertz分布[15]。该分

布在生物学[16]、社会学[17]、计算机[18]和市场营销[19]等研究领域中有重要的应用。

Gleaton和Lynch[11]利用一种新的分布函数转换,构造了一类带有形状参数$\alpha>0$的OLLG族。累积分布函数表示为

$$F(x;d,\varphi)=\frac{G(x;\varphi)^d}{G(x;\varphi)^d+\bar{G}(x;\varphi)^d},\quad(3)$$

式中,$\bar{G}(x;\varphi)=1-G(x;\varphi)$,$\varphi$为分布函数$G(x;\varphi)$的参数,相应的概率密度函数为

$$f(x;d,\varphi)=\frac{dg(x;\varphi)G(x;\varphi)^{d-1}\bar{G}(x;\varphi)^{d-1}}{\left[G(x;\varphi)^d+\bar{G}(x;\varphi)^d\right]^2}。\quad(4)$$

如果随机变量X的密度函数形式为式(4),则可以表示为$X\sim\mathrm{OLLGG}(d,\varphi)$。基于Gleaton和Lynch[11]的思想,将式(3)和(4)中的分布函数和密度函数用式(1)和(2)重新替换可得OLLGG分布,参数$a>0,b>0,c>0,d>0$,密度函数表示为

$$f(x;a,b,c,d)=\frac{dcb\mathrm{e}^{ax}\mathrm{e}^{\frac{-b}{a}(\mathrm{e}^{ax}-1)}\left[1-\mathrm{e}^{\frac{-b}{a}(\mathrm{e}^{ax}-1)}\right]^{cd-1}\left\{1-\left[1-\mathrm{e}^{\frac{-b}{a}(\mathrm{e}^{ax}-1)}\right]^c\right\}^{d-1}}{\left\{\left[1-\mathrm{e}^{\frac{-b}{a}(\mathrm{e}^{ax}-1)}\right]^{cd}+\left\{1-\left[1-\mathrm{e}^{\frac{-b}{a}(\mathrm{e}^{ax}-1)}\right]^c\right\}^d\right\}^2},\quad(5)$$

相应地,分布函数为

$$F_0(x;a,b,c,d)=\frac{\left[1-\mathrm{e}^{\frac{-b}{a}(\mathrm{e}^{ax}-1)}\right]^{cd}}{\left[1-\mathrm{e}^{\frac{-b}{a}(\mathrm{e}^{ax}-1)}\right]^{cd}+\left\{1-\left[1-\mathrm{e}^{\frac{-b}{a}(\mathrm{e}^{ax}-1)}\right]^c\right\}^d},\quad(6)$$

式中:a、b表示尺度参数,c、d表示形状参数。则生存函数表示为

$$S_0(x;a,b,c,d)=1-F_0(x;a,b,c,d)。\quad(7)$$

特别地,OLLGG分布可以灵活地转换为各种分布形式,进一步可以拟合不同形状的数据集。当形状参数$c=1,d=1$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$d=1$,该分布退化为GG分布;若同时尺度参数$a\to0$,则该分布退化为Generalized Exponential(GE)分布。此外,若尺度参数$c=1$,则该分布退化为OLLG分布;若同时尺度参数$a\to0$,则该分布退化为OLL-Exponential(OLLE)分布;若仅有尺度参数$a\to0$,则分布退化为OLL-Generalized Exponential(OLLGE)分布。

1.3　多参数OLLGG回归模型

Burke和Mackenzie[10]从分布参数回归的角度,提出了新的MPR模型概念,该模型的基本思想是对参数分布的每一个参数建立回归模型,从而提高了回归模型的灵活性与可解释性;王淑影等[20]讨论了OLLGG分布的多参数模型,假设分布模型中的尺度参数e,f和形状参数g,h受协变量$\boldsymbol{X}_1\in\mathbf{R}^{p_1}$,$\boldsymbol{X}_2\in\mathbf{R}^{p_2}$,$\boldsymbol{X}_3\in\mathbf{R}^{p_3}$和$\boldsymbol{X}_4\in\mathbf{R}^{p_4}$的影响,则尺度参数回归和形状参数回归方程如下:

$$e=\exp\left(\alpha_0+\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{\alpha}_1\right),\quad f=\exp\left(\beta_0+\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{\beta}_1\right),$$

$$g=\exp\left(\lambda_0+\boldsymbol{X}_3^{\mathrm{T}}\boldsymbol{\lambda}_1\right),\quad h=\exp\left(\gamma_0+\boldsymbol{X}_4^{\mathrm{T}}\boldsymbol{\gamma}_1\right),$$

式中:α_0、β_0、λ_0、γ_0表示截距项;$\boldsymbol{\alpha}_1\in\mathbf{R}^{p_1}$,$\boldsymbol{\beta}_1\in\mathbf{R}^{p_2}$分别表示尺度参数向量;$\boldsymbol{\lambda}_1\in\mathbf{R}^{p_3}$,$\boldsymbol{\gamma}_1\in\mathbf{R}^{p_4}$分别表示形状参数向量。在实际问题分析中,允许协变量相等,即$\boldsymbol{X}_1=\boldsymbol{X}_2=\boldsymbol{X}_3=\boldsymbol{X}_4$。

假设$\boldsymbol{\alpha}=\left(\alpha_0,\boldsymbol{\alpha}_1^{\mathrm{T}}\right)^{\mathrm{T}}$,$\boldsymbol{\beta}=\left(\beta_0,\boldsymbol{\beta}_1^{\mathrm{T}}\right)^{\mathrm{T}}$,$\boldsymbol{\lambda}=\left(\lambda_0,\boldsymbol{\lambda}_1^{\mathrm{T}}\right)^{\mathrm{T}}$,$\boldsymbol{\gamma}=\left(\gamma_0,\boldsymbol{\gamma}_1^{\mathrm{T}}\right)^{\mathrm{T}}$。将参数线性回归模型带入到分布函数(6)中,则给定协变量$\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3$和$\boldsymbol{X}_4$时,OLLGG分布下的多参数回归模型的分布函数为

$$F(x;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})=\frac{(A(x;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}))^{\exp\left(\gamma_0+\boldsymbol{X}_4\boldsymbol{\gamma}_1\right)}}{(A(x;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}))^{\exp\left(\gamma_0+\boldsymbol{X}_4\boldsymbol{\gamma}_1\right)}+(1-A(x;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda}.\boldsymbol{\gamma}))^{\exp\left(\gamma_0+\boldsymbol{X}_4\boldsymbol{\gamma}_1\right)}},\quad(8)$$

式中

$$A(x;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})=\left[1-\mathrm{e}^{-\exp\left(\beta_0+\boldsymbol{X}_2\boldsymbol{\beta}_1\right)\left(\mathrm{e}^{\exp\left(\alpha_0+\boldsymbol{X}_1\boldsymbol{\alpha}_1\right)x}-1\right)/\exp\left(\alpha_0+\boldsymbol{X}_1\boldsymbol{\alpha}_1\right)}\right]^{\exp\left(\lambda_0+\boldsymbol{X}_3\boldsymbol{\lambda}_1\right)}。\quad(9)$$

生存函数$S(x;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})=1-F(x;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})$,注意到,当$\boldsymbol{\alpha}_1=\boldsymbol{\beta}_1=\boldsymbol{\lambda}_1=\boldsymbol{\gamma}_1=\mathbf{0}$时,多参数回归模型变为传统的OLLGG分布模型。

2　参数估计

为了构建似然,采用了Emura和Shiu[8]的参数似然方法。假设随机变量T表示感兴趣事件发生的时间,C表示右删失时间,令$Z=\min\{T,C\}$,δ表示删失的示性函数,即

$$\delta=\begin{cases}0, & \text{数据发生右删失},\\ 1, & \text{数据未发生右删失}。\end{cases}$$

Y表示左截断时间,d表示左截断的示性函数,即

$$d=\begin{cases}0, & \text{数据发生左截断},\\ 1, & \text{数据未发生左截断}。\end{cases}$$

$\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3$和$\boldsymbol{X}_4$为协变量,则n个独立同分布的左截断右删失数据的观测样本结构为

$$D=\left\{x_{1i},x_{2i},x_{3i},x_{4i},z_i,y_i,\delta_i,d_i,i=1,\cdots,n\right\}。$$

设S_1为左截断数据集,S_2为没有出现左截断,由右删失数据和完全数据所构成的集合。分别记为$S_1=\{i:d_i=0\}$,$S_2=\{i:d_i=1\}$。则在左截断右删失数据的情况下,参数的似然函数为

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma};D)=\prod_{i\in S_2}\left[f\left(z_i\right)\right]^{\delta_i}\left[1-F\left(z_i\right)\right]^{1-\delta_i}\times\prod_{i\in S_1}\left[\frac{f\left(z_i\right)}{1-G\left(y_i\right)}\right]^{\delta_i}\left[\frac{1-F\left(z_i\right)}{1-G\left(y_i\right)}\right]^{1-\delta_i},\tag{10}$$

式中:G为左截断时间的分布函数,F为感兴趣事件发生的时间的分布函数,f为相应的密度函数。

在左截断右删失数据下,选择左截断点的分布与感兴趣的事件时间的分布相同[6,8,21],都是OLLGG分布,即$G(\cdot)=F(\cdot)$,所以在左截断右删失数据的情况下,OLLGG分布的多参数回归模型的似然函数表示为

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma};D)=\prod_{i\in S_2}\left[f\left(z_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}\right)\right]^{\delta_i}\cdot\left[1-F\left(z_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}\right)\right]^{1-\delta_i}\times\prod_{i\in S_1}\left[\frac{f\left(z_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}\right)}{1-F\left(y_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}\right)}\right]^{\delta_i}\left[\frac{1-F\left(z_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}\right)}{1-F\left(y_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}\right)}\right]^{1-\delta_i},\tag{11}$$

所以,对数似然函数表示为

$$\begin{aligned}l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma};D)&=\log L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma};D)=\\&\sum_{i\in S_2}\{\delta_i\log[f(z_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})]+(1-\delta_i)\log[1-F(z_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})]\}+\sum_{i\in S_1}\{\delta_i[\log f(z_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})-\\&\log[1-F(y_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})]]+(1-\delta_i)[\log[1-F(z_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})]-\log[1-F(y_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})]]\}。\end{aligned}\tag{12}$$

利用极大似然法求解模型的参数,基于对数似然函数式(12)计算各个参数的偏导数,可以获得得分方程

$$\begin{aligned}U(\boldsymbol{\alpha})&=\partial l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma};D)/\partial\boldsymbol{\alpha}=0,\\U(\boldsymbol{\beta})&=\partial l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma};D)/\partial\boldsymbol{\beta}=0,\\U(\boldsymbol{\lambda})&=\partial l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma};D)/\partial\boldsymbol{\lambda}=0,\\U(\boldsymbol{\gamma})&=\partial l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma};D)/\partial\boldsymbol{\gamma}=0。\end{aligned}$$

联立方程即可获得模型的极大似然估计量$\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\gamma}}$。根据文献[22]的定理2.13可知,在参数模型假设下,极大似然估计量$\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\gamma}}$具有相合性和渐近正态性。假设$\boldsymbol{\alpha}_0$、$\boldsymbol{\beta}_0$、$\boldsymbol{\lambda}_0$、$\boldsymbol{\gamma}_0$是模型参数的真值,则$\sqrt{n}\left(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0,\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0,\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\right)$依分布收敛到均值为0,协方差为$\boldsymbol{I}^{-1}(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0,\boldsymbol{\lambda}_0,\boldsymbol{\gamma}_0)$的正态分布,即

$$\sqrt{n}\left(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0,\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0,\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\right)\xrightarrow{L}N(0,\boldsymbol{I}^{-1}(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0,\boldsymbol{\lambda}_0,\boldsymbol{\gamma}_0)),$$

其中,

$$\boldsymbol{I}(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0,\boldsymbol{\lambda}_0,\boldsymbol{\gamma}_0)=-E\left(\begin{array}{cccc}\frac{\partial U(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}&\frac{\partial U(\boldsymbol{\alpha})}{\partial\boldsymbol{\beta}}&\frac{\partial U(\boldsymbol{\alpha})}{\partial\boldsymbol{\lambda}}&\frac{\partial U(\boldsymbol{\alpha})}{\partial\boldsymbol{\gamma}}\\\frac{\partial U(\boldsymbol{\beta})}{\partial\boldsymbol{\alpha}}&\frac{\partial U(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}&\frac{\partial U(\boldsymbol{\beta})}{\partial\boldsymbol{\lambda}}&\frac{\partial U(\boldsymbol{\beta})}{\partial\boldsymbol{\gamma}}\\\frac{\partial U(\boldsymbol{\lambda})}{\partial\boldsymbol{\alpha}}&\frac{\partial U(\boldsymbol{\lambda})}{\partial\boldsymbol{\beta}}&\frac{\partial U(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}&\frac{\partial U(\boldsymbol{\lambda})}{\partial\boldsymbol{\gamma}}\\\frac{\partial U(\boldsymbol{\gamma})}{\partial\boldsymbol{\alpha}}&\frac{\partial U(\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}}&\frac{\partial U(\boldsymbol{\gamma})}{\partial\boldsymbol{\lambda}}&\frac{\partial U(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\end{array}\right)_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0,\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\lambda}=\boldsymbol{\lambda}_0,\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}。$$

3 数值模拟

由OLLGG的分布函数可以得到分位数函数为

$$Q(u)=\frac{1}{a}\ln\left\{1-\frac{a}{b}\ln\left\{1-\left[\frac{u^{\frac{1}{d}}}{u^{\frac{1}{d}}+(1-u)^{\frac{1}{d}}}\right]^{\frac{1}{c}}\right\}\right\},\quad u\in(0,1)。\tag{13}$$

通过逆变换方法生成遵循OLLGG分布的随机变量。即$T=Q(U)\sim$OLLGG,U服从均匀分布$U(0,1)$。假设协变量$X_{1i}=X_{2i}=X_{3i}=X_{4i}$,并服从成功概率为0.5的二项分布。设置参数如下(p表示右删失比例,q表示左截断比例)。

(1)$\alpha_0=0.7$,$\alpha_1=-0.2$,$\beta_0=0.2$,$\beta_1=-0.3$,$\lambda_0=0.9$,$\lambda_1=0.2$,$\gamma_0=-0.8$,$\gamma_1=-0.2$。

① $p=0.15$,$q=0.50$;

② $p=0.30$,$q=0.50$。

(2)$\alpha_0=0.8$,$\alpha_1=-0.2$,$\beta_0=-0.7$,$\beta_1=0.3$,$\lambda_0=0.8$,$\lambda_1=0.3$,$\gamma_0=-0.8$,$\gamma_1=-0.2$。

① $p = 0.15, q = 0.50$;

② $p = 0.30, q = 0.50$。

(3) $\alpha_0 = 0.7, \alpha_1 = -0.2, \beta_0 = 0.2, \beta_1 = -0.3, \lambda_0 = 1.1, \lambda_1 = 0.2, \gamma_0 = -1.1, \gamma_1 = -0.2$。

① $p = 0.60, q = 0.30$;

② $p = 0.60, q = 0.50$。

在基于OLLGG分布的多参数回归模型假设下,根据上述3组参数设置生成失效时间T_i。根据Balakrishnan和Mitra[23]中生成左截断右删失数据的方法,并使用R软件进行随机模拟。检验不同样本量$n = 200$和$n = 400$,并重复1 000次参数估计,模拟结果如表1所示。不同n时,估计值的偏差较小,且都<0.1,估计值基本接近真实值,样本标准差(sample standard error,SSE)与估计值均数的标准差(mean of estimated standard error,MESE)比较接近,置信水平为95%置信区间的覆盖概率(coverage rate,CP)在0.95左右波动。并且不同p与q时,低删失比例情况下估计值的偏差和SSE略低于高删失比例的情况,低截断比例情况下估计值的偏差和SSE略低于高截断比例的情况,但是差异均不大。因此本文的模型可以适用于不同删失比例以及不同截断比例的数据。

同时,使用忽略左截断的方法来分析数值模拟参数设置(1)生成的左截断数据,结果如表2所示。在不同删失比例条件下,使用忽略左截断的方法所得到的估计值均比考虑左截断方法得到的估计值的偏差以及标准差高。

4　实例分析

本文分析了Channing House数据集[24-26]。该数据集记载了位于加州帕洛阿尔托1964年1月—1975年7月加入Channing House退休中心的462名老年居民的死亡时间。因为只有实际寿命超过实验起始点的个体才可以被观测到,所以本文对所有进入实验时实际年龄$<$ 70岁的个体进行了左截断处理,截断比为18.2%,用Y表示截断时间,示性函数d表示截断状态,$d = 0$表示截断,$d = 1$表示未截断;同时,由于有些个体存活到实验结束时间之后,对这些个体进行了右删失处理,删失比例为61.9%,示性函数δ表示删失状态,$\delta = 0$表示删失,$\delta = 1$表示未删失。该数据集中进入退休中心的女性有365名,男性有97名,用协变量X表示性别,$X = 1$表示进入退休中心的女性,$X = 2$表示进入退休中心的男性。

使用Kaplan-Meier(K-M)估计和Nelson-Altschuler(N-A)估计量来估计进入退休中心男性的生存函数并画出图像,如图2所示。K-M估计值在近65岁左右骤降为0,且K-M估计是累乘,一旦降为0,则永远为0。因为在早期,男性的风险集很小,进而迫使生存曲线为0。N-A估计依赖于指数的累积和,没有一直为0的情况,但也一开始就骤降。这样都无法准确估计男性的生存分布。所以本文考虑年龄超过70岁的左截断,可以看出,绿色曲线有比较好的表现,能较好的解决男性早期风险集较小的问题。

同时,将本文方法分别与只处理右删失不处理左截断、只处理左截断不处理右删失、既不处理左截断也不处理右删失进行比较,相关结果如图3所示。与进入退休中心居民真实的生存曲线相比,本文所采用的方法最为接近,既不处理右删失也不处理左截断方法的生存曲线相距最远。所以应当充分考虑左截断和右删失数据的情况。

为了选择合适的分布来拟合数据,比较OLLGG分布、威布尔分布、指数分布、广义指数分布、Gompertz分布的AIC值。AIC值定义为

$$A_{\mathrm{IC}} = -2\log L(\hat{\theta}) + 2k, \tag{14}$$

式中:$L(\hat{\theta})$为对数似然的最大值,k为模型的参数个数。上述5种分布的AIC值分别为1 324.400、1 414.577、1 325.902、1 341.823和1 331.939。可知OLLGG分布的AIC值最小,因此选择OLLGG分布来进行数据分析。

基于OLLGG分布和前面提出的模型,对数据进行拟合和估计,得到分析结果如表3所示。协变量系数α_1、β_1、λ_1、γ_1对应的检验P>0.05,说明性别因素对老年居民生存概率的影响并不显著,换言之,在这个特定群体中,性别因素并未对加州帕洛阿尔托市退休中心老年居民的生存概率造成明显的影响。

5　结论与展望

左截断右删失数据处理起来比较困难,如何有效利用左截断数据和右删失数据的信息显得尤为重要。本文基于OLLGG分布的多参数回归模型形式,充分利用了左截断数据和右删失数据的信息,进而构造似然函数,并用R软件进行参数估计和随机模拟实验。结果表明参数估计值具有良好的稳

表1 参数设置(1~3)的模拟结果

右删失比例/%	左删失比例/%	参数	样本量为200				样本量为400			
			样本偏差	标准误差	标准差	覆盖概率	样本偏差	标准误差	标准差	覆盖概率
15	50	α_0=0.7	0.004 0	0.247 6	0.254 9	0.949	0.023 6	0.170 5	0.166 3	0.952
		α_1=−0.2	0.015 3	0.343 9	0.344 3	0.949	0.003 4	0.241 5	0.239 6	0.947
		β_0=0.2	0.069 6	0.433 4	0.445 1	0.945	0.013 2	0.282 6	0.277 6	0.946
		β_1=−0.3	0.069 6	0.571 1	0.542 9	0.956	0.009 4	0.388 4	0.388 9	0.939
		λ_0=0.9	0.069 6	0.459 9	0.467 0	0.944	0.065 1	0.340 8	0.339 2	0.936
		λ_1=0.2	0.010 1	0.615 0	0.635 7	0.944	0.018 1	0.474 7	0.455 8	0.938
		γ_0=−0.8	0.063 9	0.374 6	0.378 6	0.934	0.053 7	0.266 4	0.267 6	0.931
		γ_1=−0.2	0.036 9	0.536 6	0.529 3	0.947	0.017 7	0.389 0	0.377 7	0.941
30	50	α_0=0.7	0.021 9	0.244 7	0.238 3	0.953	0.019 7	0.206 7	0.204 2	0.945
		α_1=0.7	0.024 3	0.338 5	0.345 4	0.944	0.023 6	0.268 3	0.274 2	0.949
		β_0=0.2	0.047 8	0.446 0	0.445 4	0.949	0.024 6	0.398 8	0.389 3	0.951
		β_1=−0.3	0.017 3	0.548 3	0.575 3	0.945	0.014 6	0.388 8	0.389 2	0.949
		λ_0=0.9	0.068 8	0.510 3	0.493 4	0.932	0.044 2	0.331 3	0.346 7	0.943
		λ_1=0.2	0.062 6	0.646 1	0.666 5	0.946	0.025 3	0.466 5	0.468 4	0.943
		γ_0=−0.8	0.039 5	0.376 0	0.383 1	0.932	0.039 1	0.294 3	0.292 3	0.942
		γ_1=−0.2	0.092 1	0.567 9	0.573 4	0.946	0.001 6	0.400 4	0.400 0	0.944
15	50	α_0=0.8	0.035 6	0.208 1	0.206 4	0.951	0.011 0	0.141 3	0.139 8	0.947
		α_1=−0.2	0.028 5	0.285 1	0.288 0	0.945	0.020 9	0.201 6	0.199 6	0.945
		β_0=−0.7	0.080 8	0.445 7	0.447 9	0.953	0.005 5	0.338 4	0.336 1	0.954
		β_1=0.3	0.038 5	0.591 3	0.597 9	0.939	0.013 4	0.396 0	0.377 3	0.954
		λ_0=0.8	0.086 7	0.473 6	0.494 4	0.938	0.068 8	0.357 8	0.350 9	0.945
		λ_1=0.3	0.064 1	0.702 1	0.710 6	0.947	0.006 0	0.471 2	0.471 1	0.937
		γ_0=−0.8	0.084 9	0.416 0	0.420 5	0.932	0.050 2	0.296 2	0.290 1	0.945
		γ_1=−0.2	0.053 3	0.590 8	0.601 4	0.945	0.021 8	0.403 4	0.405 1	0.941
30	50	α_0=0.8	0.031 7	0.227 8	0.221 4	0.954	0.016 7	0.169 8	0.172 7	0.945
		α_1=−0.2	0.023 8	0.352 8	0.343 7	0.954	0.005 9	0.250 2	0.246 7	0.948
		β_0=−0.7	0.060 4	0.517 3	0.510 4	0.953	0.032 6	0.354 7	0.339 2	0.947
		β_1=0.3	0.035 1	0.679 4	0.682 8	0.944	0.002 3	0.452 2	0.445 5	0.947
		λ_0=0.8	0.090 9	0.502 4	0.504 5	0.947	0.059 6	0.403 2	0.405 2	0.948
		λ_1=0.3	0.045 3	0.663 1	0.652 8	0.945	0.027 4	0.520 7	0.520 5	0.944
		γ_0=−0.8	0.084 8	0.416 4	0.426 4	0.949	0.048 8	0.328 1	0.332 8	0.945
		γ_1=−0.2	0.051 1	0.570 1	0.562 5	0.948	0.028 6	0.446 7	0.458 3	0.945
60	30	α_0=0.7	0.010 8	0.545 9	0.554 2	0.953	0.008 3	0.357 9	0.368 8	0.960
		α_1=−0.2	0.097 8	0.877 8	0.863 9	0.939	0.077 7	0.704 0	0.682 5	0.961
		β_0=0.2	0.062 9	0.582 3	0.559 3	0.946	0.034 8	0.396 9	0.386 9	0.950
		β_1=−0.3	0.063 9	0.892 3	0.902 3	0.951	0.041 3	0.569 2	0.577 3	0.944
		λ_0=1.1	0.058 2	0.690 2	0.671 3	0.939	0.072 1	0.509 3	0.509 1	0.946
		λ_1=0.2	0.078 1	0.889 3	0.952 1	0.938	0.025 7	0.665 7	0.644 8	0.939
		γ_0=−1.1	0.042 1	0.617 2	0.612 1	0.929	0.066 4	0.457 7	0.440 4	0.938
		γ_1=−0.2	0.095 3	0.822 9	0.824 7	0.933	0.032 0	0.636 2	0.607 2	0.939
60	30	α_0=0.7	0.063 6	0.606 9	0.577 8	0.956	0.008 4	0.326 5	0.312 8	0.949
		α_1=−0.2	0.099 8	0.899 9	0.894 9	0.947	0.087 9	0.604 1	0.604 1	0.952
		β_0=0.2	0.073 0	0.5914	0.594 5	0.949	0.053 5	0.367 7	0.367 7	0.944
		β_1=−0.3	0.065 8	0.899 9	0.912 4	0.947	0.042 8	0.597 8	0.597 8	0.951
		λ_0=1.1	0.092 1	0.681 1	0.676 6	0.943	0.081 2	0.514 3	0.514 3	0.946
		λ_1=0.2	0.082 7	0.948 6	0.935 1	0.941	0.056 4	0.659 7	0.659 7	0.948
		γ_0=−1.1	0.096 9	0.605 1	0.607 2	0.939	0.071 8	0.473 9	0.473 9	0.943
		γ_1=−0.2	0.099 1	0.885 7	0.887 1	0.946	0.046 1	0.634 2	0.634 2	0.939

表2　样本量为200时参数设置(1)下考虑左截断和忽略左截断的结果比较

右删失比例/%	参数	考虑左截断				忽略左截断			
		样本偏差	标准误差	标准差	覆盖概率	样本偏差	标准误差	标准差	覆盖概率
15	α_0=0.7	0.004 0	0.247 6	0.254 9	0.949	0.052 3	0.268 1	0.266 9	0.912
	α_1=−0.2	0.015 3	0.343 9	0.344 3	0.949	0.025 6	0.380 8	0.378 8	0.927
	β_0=0.2	0.069 6	0.433 4	0.445 1	0.945	0.083 4	0.449 8	0.450 3	0.916
	β_1=−0.3	0.035 9	0.571 1	0.542 9	0.956	0.076 5	0.595 7	0.586 4	0.909
	λ_0=0.9	0.085 7	0.459 9	0.467 0	0.944	0.134 3	0.519 4	0.534 8	0.906
	λ_1=0.2	0.010 1	0.615 0	0.635 7	0.944	0.035 7	0.659 7	0.638 9	0.918
	γ_0=−0.8	0.063 9	0.374 6	0.378 6	0.934	0.110 3	0.430 9	0.405 4	0.901
	γ_1=−0.2	0.036 9	0.536 6	0.529 3	0.947	0.041 3	0.561 4	0.570 7	0.891
30	α_0=0.7	0.021 9	0.244 7	0.238 3	0.953	0.052 6	0.326 3	0.319 9	0.910
	α_1=−0.2	0.024 3	0.338 5	0.345 4	0.944	0.031 1	0.544 1	0.546 6	0.919
	β_0=0.2	0.047 8	0.446 0	0.445 4	0.949	0.081 7	0.486 4	0.519 6	0.911
	β_1=−0.3	0.017 3	0.548 3	0.575 3	0.945	0.076 9	0.726 9	0.721 4	0.849
	λ_0=0.9	0.068 8	0.510 3	0.493 4	0.932	0.137 6	0.569 1	0.555 8	0.903
	λ_1=0.2	0.062 6	0.646 1	0.666 5	0.946	0.075 2	0.725 5	0.725 2	0.900
	γ_0=−0.8	0.039 5	0.376 0	0.383 1	0.932	0.122 3	0.475 1	0.497 1	0.912
	γ_1=−0.2	0.092 1	0.567 9	0.573 4	0.946	0.092 5	0.655 4	0.653 5	0.894

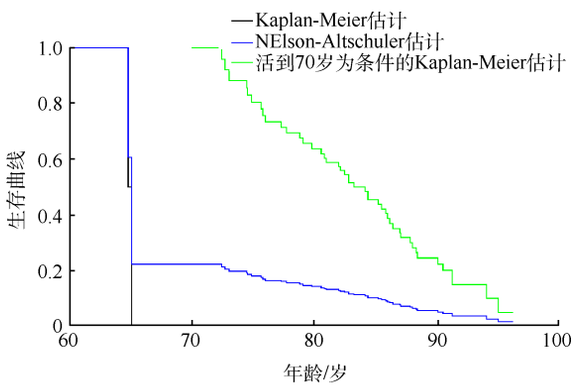


图2　进入退休中心男性的生存函数的估计曲线

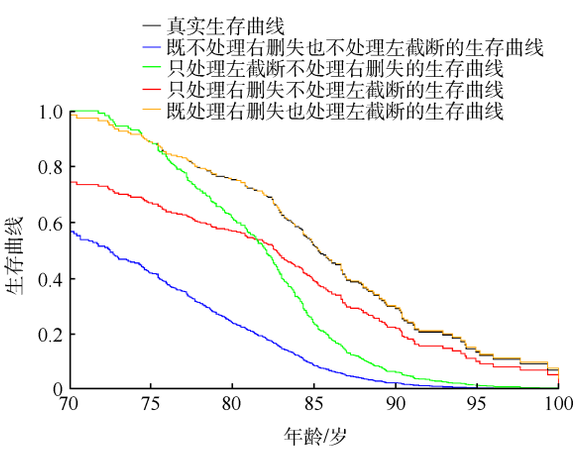


图3　使用不同方法处理老年居民存活时间分布的估计曲线

表3　Channing House 数据集的估计结果

参数	估计值	标准差	*P*值
α_0	−12.465 5	0.228 0	<0.000 1
α_1	5.699 5	0.222 6	0.981 7
β_0	−3.244 8	0.057 8	<0.000 1
β_1	0.155 9	0.107 7	0.209 8
λ_0	−0.383 5	0.055 4	<0.000 1
λ_1	−0.120 1	0.128 9	0.099 9
γ_0	0.551 1	0.037 3	<0.000 1
γ_1	0.048 0	0.074 0	0.755 6

定性和有效性。多参数回归形式可以灵活地拟合各种数据模型,弥补了单一回归模型的局限性。本文还将提出的模型应用在美国加州帕洛阿尔托市退休中心的居民的数据集上,并用AIC信息准则进行了模型选择。协变量性别对老年居民的生存概率没有显著影响。本文的主要贡献是提出一种左截断右删失数据的OLLGG分布多参数回归模型,并进行参数估计。未来的工作可以考虑将本文的模型扩展到半参数情形。

参 考 文 献

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

[2] 热比古力,周菊玲.基于左截断右删失数据下瑞利分布的参数估计[J].数学的实践与认识,2018,48(22):202-206.
[3] 赵江南,周菊玲.基于左截断右删失数据的单参数Pareto分布参数估计[J].数学的实践与认识,2017,47(15):254-257.
[4] 何朝兵,郭梦雪,高唱,等.左截断右删失数据下离散型寿命失效率变点模型的Bayes参数估计[J].数学的实践与认识,2016,46(24):194-204.
[5] 胡隽,曹显兵.左截断右删失下Weibull分布参数估计及其应用[J].数学的实践与认识,2015,45(17):228-234.
[6] 胡隽,曹显兵.基于左截断右删失数据的指数分布参数估计[J].统计与决策,2016(16):71-73.
[7] 何朝兵.右删失左截断数据下离散威布尔分布的参数估计[J].郑州大学学报(理学版),2016,48(2):18-23.
[8] EMURA T, SHIU S K. Estimation and model selection for left-truncated and right-censored lifetime data with application to electric power transformers analysis[J]. Communications in Statistics-Simulation and Computation,2016,45(9):3171-3189.
[9] KUNDU D, MITRA D, GANGULY A. Analysis of left truncated and right censored competing risks data[J]. Computational Statistics and Data Analysis, 2017, 108: 12-26.
[10] BURKE K, MACKENZIE G. Multi-parameter regression survival modeling: an alternative to proportional hazards[J]. Biometrics, 2017, 73(2): 678-686.
[11] 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.
[12] VALDEMIRO P V, EDWIN M M O, ADRIANO K S, et al. The generalized odd log-logistic-G regression with interval-censored survival data[J]. Journal of Applied Statistics, 2023: 1-22.
[13] PRATAVIERA F, ORTEGA E M M, CORDEIRO G M, et al. A new generalized odd log-logistic flexible weibull regression model with applications in repairable systems[J]. Reliability Engineering & System Safety, 2018, 176: 13-26.
[14] 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.
[15] GOMPERTZ B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies[J]. Proceedings of the Royal Society of London, 1815: 252-253.
[16] ECONONMOS A C. Rate of aging, rate of dying and the mechanism of mortality[J]. Archives of Gerontology and Geriatrics, 1982, 1(1): 3-27.
[17] WILLEMSE J W, KOPPELAAR H. Knowledge elicitation of gompertz' law of mortality[J]. Scandinavian Actuarial Journal, 2000(2): 168-179.
[18] OHISHI K, OKAMURA H, DOHI T. Gompertz software reliability model: estimation algorithm and empirical validation[J]. The Journal of Systems & Software, 2009, 82(3): 535-543.
[19] BEMMAOR A C, GLADY N. Modeling purchasing behavior with sudden "death": a flexible customer lifetime model[J]. Management Science, 2012, 58(5): 1012-1021.
[20] 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.
[21] BALAKRISHNAN N, MITRA D. Some further issues concerning likelihood inference for left truncated and right censored lognormal data[J]. Communications in Statistics(Simulation and Computation), 2014, 43(2): 400-416.
[22] 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.
[23] BAlAKRISHNAN N, MITRA D. Likelihood inference for lognormal data with left truncation and right censoring with an illustration[J]. Journal of Statistical Planning and Inference, 2011, 141(11): 3536-3553.
[24] MOORE D F. Applied survival analysis using R[M]. New York: Springer International Publishing, 2016.
[25] 彭非,王伟.生存分析[M].北京:中国人民大学出版社,2004.
[26] 赵波,王纯杰,李群,等.左截断右删失数据下尺度参数与协变量相关时广义指数分布的估计及应用[J].吉林大学学报(理学版),2018,56(4):864-870.

(责任编辑:马田田)