文章编号:1000-5862(2014)06-0578-04

信息区间删失数据的参数估计及敏感性分析

李文静,邓文丽*,章婷婷
(江西师范大学数学与信息科学学院,江西 南昌 330022)

摘要:基于连接函数构造了信息区间删失数据的似然函数,研究了信息区间删失的分布函数问题. 连接函数的假定会对估计结果产生一定的影响,通过模拟计算对这种影响进行了敏感性分析.

关键词:信息区间删失数据;连接函数;敏感性分析
中图分类号:O 212.1　　**文献标志码**:A

0 引言

在临床实验或医学研究中,由于客观因素的限制,失效时间常常不能直接观测到,而只能知道它在某一个区间内,这类数据在统计学上被称为区间删失数据(interval-censored data). 如在一些传染性疾病感染时间的研究中,实验对象被放入传染源后,只能知道在某个观察点实验对象是否已染上疾病,染上疾病的具体时间却无法观测到,所以只能推测出从接触传染源到染上传染病所经历的时间落在某个区间内.

区间删失数据存在于许多应用领域中,因此,这引发了一些统计学者对相关问题的研究. Huang Jian 等[1]对区间删失数据的分类及对应的统计方法进行了较为详细地描述. Sun Jian-guo[2]较为全面和系统地概括了区间删失数据分析中涉及到的基本概念和方法. 吕秋萍等[3]运用无偏转换思想构造了区间删失数据函数的均值估计,并在此基础上对所构造的估计量方差进行了研究. 在区间删失数据的研究中,许多学者都是基于失效时间变量 T 和删失时间变量 C 相互独立的假定进行研究的,称这种删失情况为独立删失或非信息删失(Independent Censoring,Noninformative Censoring). 然而,在实际问题中,这个假定常常会遭到质疑. 如在对某种疾病的治疗中,由于病情恶化或者是已接受的治疗方案不奏效,从而导致病人退出治疗,这种情况通常预示着该病人的存活时间会比较短,即删失的个体对应的生存时间更短. 相反地,有些病人的退出可能因为病情好转,不需要进一步治疗,这种情况的删失个体的生存时间可能会较长. 和独立删失相反的是非独立删失(Dependent Censoring),或称为信息删失(Informative Censoring). 如果对信息删失数据仍采用独立删失下的统计分析方法,则可能会得到有偏或者无效的结论.

在信息删失数据的研究中,对失效时间和删失时间相依性的假定是至关重要的. 正确的假定可以提高估计的效率,得到更好的统计结论;不合适的假定可能会导致错误的结论. 在实际应用中,由于造成信息删失的应用背景和原因的不同,失效时间和删失时间相依的形式和程度也变得非常复杂,很难准确估计. 敏感性分析可以评价相依关系的假定对统计分析结果造成的影响. 王纯杰[4]基于 Copula 函数的一些性质,给出了非参数模型下的信息区间删失数据分布函数的相合估计. F. Siannis 等[5]对失效时间和删失时间的相依关系进行了假定,引入了标示相关程度的参数和偏度函数,且对参数估计受相依程序的影响进行了敏感性分析. Y. Park 等[6]在单个总体和 2 个总体的情形下,分别对独立删失和信息右删失混合数据下的相关估计问题进行了敏感性分析. Zhang Zhi-gang 等[7]在正态脆弱模型假定下对 I 型信息区间删失数据的比例风险模型进行了敏感性分析. Huang Xue-lin 等[8]基于连接函数(Copula)对信息右删失数据下的比例风险模型的估计问题进行了敏感性分析.

本文拟基于连接函数对信息区间删失数据下失效时间的生存函数进行估计,并在 3 种不同连接函数的情形下关于相依关系对参数估计所造成的影响

收稿日期:2014-09-20
基金项目:国家自然科学青年基金(71001046)资助项目.
通信作者:邓文丽(1974-),女,江西南昌人,副教授,博士,主要从事数理统计研究.

进行敏感性分析.

1　方法

记 T_i 为第 i 个个体的失效时间,C_i 为第 i 个个体的删失时间. 试验中得到的独立同分布观测值为 $\{(c_i,\delta_i),i=1,2,\cdots,n\}$,其中

$$\delta_i=\begin{cases}1, T_i\leqslant C_i\\ 0, T_I>C_i.\end{cases}$$

假设 T_i 和 C_i 的边际分布函数分别为 $F(\cdot)$ 和 $G(\cdot)$,$g(\cdot)$ 是 C_i 的边际密度函数,这里 $i=1,2,\cdots,n$.

基于观测样本,可以构造似然函数:

$$L(F)=\prod_{i=1}^{n}[P(C_i=c_i,\delta_i=0)]^{1-\delta_i}\cdot[P(C_i=c_i,\delta_i=1)]^{\delta_i}.$$

当 T_i 和 C_i 相互独立时,基于观测样本的似然函数为

$$L(F)=\prod_{i=1}^{n}[F(C_i)]^{\delta_i}[1-F(C_i)]^{1-\delta_i},$$

当 T_i 和 C_i 不相互独立时,给定 1 个有参数 α 的连接函数 $H(u,v,\alpha)$,假设 T_i 和 C_i 的联合分布函数为 $J(t,c)=P(T_i\leqslant t,C_i\leqslant c)=H(F(t),G(c),\alpha)$,联合生存函数为

$$S(t,c)=P(T_i>t,C_i>c)=1-F(t)-G(c)H(F(t),G(c)),$$

则第 i 个个体被删失的概率为

$$P(C_i=c_i,\delta_i=0)=P(C_i=c_i,T_i>c_i)=\lim_{\Delta c\to 0}\frac{P(C_i\in(c_i,c_i+\Delta c),T_i>c_i)}{\Delta c}=\frac{\mathrm{d}}{\mathrm{d}c_i}\int_0^{c_i}\int_v^{+\infty}\frac{\partial^2 S(u,v)}{\partial u\partial v}\mathrm{d}u\mathrm{d}v=\int_{c_i}^{+\infty}\frac{\partial^2 S(u,v)}{\partial u\partial v}\Big|_{v=c_i}\mathrm{d}u=-\frac{\partial S(u,v)}{\partial v}\Big|_{u=c_i,v=c_i}.$$

同理可得,第 i 个个体失效的概率为

$$P(C_i=c_i,\delta_i=1)=\int_0^{c_i}\frac{\partial^2 J(u,v)}{\partial u\partial v}\Big|_{v=c_i}\mathrm{d}u=\frac{\partial J(u,v)}{\partial v}\Big|_{u=c_i,v=c_i}.$$

由此可知,

$$\frac{\partial S(t,c)}{\partial c}=-g(c)+\frac{\partial H(F(t),G(c))}{\partial G(c)}g(c),$$

$$\frac{\partial J(t,c)}{\partial c}=\frac{\partial H(F(t),G(c))}{\partial G(c)}g(c).$$

综上所述,当失效时间和删失时间不相互独立时,样本的观测似然函数可以表示为

$$L(F)=\prod_{i=1}^{n}\left(1-\frac{\partial H(F(c_i),v)}{\partial v}\Big|_{v=G(c_i)}\right)^{1-\delta_i}\cdot\left(\frac{\partial H(F(c_i),v)}{\partial v}\Big|_{v=G(c_i)}\right)^{\delta_i}.\quad(1)$$

在信息区间删失数据中,删失时间能够完全观测到,所以 $G(\cdot)$ 可以直接用它的经验分布函数 $\hat{G}(\cdot)$ 代替,其中 $\hat{G}(c)=\frac{1}{n}\sum_{i=1}^{n}I_{C_i\leqslant c}$.

似然函数(1)可以转化为

$$\hat{L}(F)=\prod_{i=1}^{n}\left(1-\frac{\partial H(F(c_i),v)}{\partial v}\Big|_{v=\hat{G}(c_i)}\right)^{1-\delta_i}\cdot\left(\frac{\partial H(F(c_i),v)}{\partial v}\Big|_{v=\hat{G}(c_i)}\right)^{\delta_i}.\quad(2)$$

如果对失效时间的分布形式掌握的信息不多,则通常会考虑用非参数模型直接估计失效时间的分布函数. 类似于文献[2]给出的独立区间删失数据非参数极大似然估计的方法,可以在(2)式中利用邓文丽等[9]提出的一类保序最优化问题的迭代算法得到分布函数 F 的估计. 当已知一些影响失效时间的协变量时,比例风险模型和加速失效模型是广泛接受的 2 类半参数模型. 如果假定协变量的影响满足比例风险模型,则在似然函数的表达式中,边际分布函数 $F(\cdot)$ 可以用含回归系数和基准风险函数的分布函数表达式代替,然后在似然函数(2)中,通过迭代的方法得到相关的估计; 如果假定协变量的影响满足加速失效模型,则在似然函数的表达式中,边际分布函数 $F(\cdot)$ 可以用含回归系数和随机误差项的分布函数表达式代替,然后在似然函数(2)中,通过迭代的方法得到相关的估计. 张连增等[10]基于极大似然法研究了 Copula 的参数和半参数方法的估计效果. 如果失效时间的分布函数形式已知,而只待估其中包含的参数,则利用似然函数(2)就可以得出参数的极大似然估计.

2　数值模拟

在实际应用中由于失效时间和删失时间相依的形式和程度非常复杂,很难准确估计,所以通过敏感性分析来评价相依关系的假定对统计分析结果造成的影响.

模拟计算中失效时间 T 采用威布尔分布随机生成,因为它的危险率不是常数,所以,与指数分布相比,它有较广阔的应用,将其用于调查深槽轮滚珠轴承的疲劳寿命,或将其用于描写电子管的失效. 威布尔的分布函数为 $F(t)=1-\mathrm{e}^{-(\lambda t)^{\gamma}}$,其中,$\gamma$ 是分布

曲线的形状参数,λ 是尺度参数. 模拟计算中选取了 $\gamma = 2, \lambda = 0.5$. 删失时间 C 的边际分布选取的是 $(0, A)$ 上的均匀分布,调整 A 的大小可以改变删失的比例.

失效时间和删失时间的相关性选用阿基米德连接函数来描述. 这里选取了 Clayton、Gumbe-Hougard 和 Frank 3 种连接函数.

D. G. Clayton[11] 给出在 $\tau = 1/(1 + 2\alpha)$ 下的 Copula 函数:

$$H(u, v, \alpha) = u + v - 1 + \{(1 - u)^{-1/\alpha} + (1 - v)^{-1/\alpha} - 1\}^{-\alpha} (\alpha > 0). \quad (3)$$

E. J. Gumbel 等[12] 给出 $\tau = 1 - 1/\alpha$ 下的 Copula 函数:

$$H(u, v, \alpha) = \exp\{-[(-\log u)^{\alpha} + (-\log v)^{\alpha}]^{1/\alpha}\} (\alpha \geqslant 1).$$

M. J. Frank[13] 给出的 Copula 函数:

$$H(u, v, \alpha) = \log\{1 + (\alpha^{u} - 1)(\alpha^{v} - 1)/(\alpha - 1)\} (\alpha > 0, \alpha \neq 1),$$

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

R. B. Nelsen[14] 对于连接函数的相关性质和特殊的连接函数进行了详细介绍.

下面主要是通过数值计算分析连接函数的选取对参数 γ 和 λ 的估计产生的影响. 这里的稳健性分析包括参数敏感性分析和连接函数敏感性分析.

2.1 参数的敏感性分析

分别取 $\tau = 0.8$、$\tau = 0.5$、$\tau = 0.2$ 的 Frank Copula 作为连接函数,T 服从 $\gamma = 2, \lambda = 0.5$ 的威布尔分布,C 服从 $(0, 37)$ 上的均匀分布,生成容量为 200 的样本,删失比例 $P(T < C)$ 为 0.5. 在本文方法中,选取 Frank 连接函数,$\tau = 0.8$. 模拟次数为 1 000,得到上述情况下 $\tilde{\lambda}$ 和 $\tilde{\gamma}$ 的均值、标准差和偏差的估计值(见表 1).

表 1 总体的参数 τ 变化下参数 γ 和 λ 的估计

τ	参数	真实值	Frank 连接函数 $\tau = 0.8$			独立删失		
			估计值	标准差	偏差	估计值	标准差	偏差
0.8	γ	2.000	2.018	0.022	0.018	1.108	0.183	0.892
	λ	0.500	0.503	0.153	0.003	0.436	0.046	0.064
0.5	γ	2.000	2.025	0.156	0.025	1.167	0.197	0.833
	λ	0.500	0.492	0.021	0.008	0.468	0.045	0.032
0.2	γ	2.000	2.214	0.183	0.214	1.597	0.249	0.403
	λ	0.500	0.494	0.021	0.006	0.489	0.035	0.011

由表 1 可以看出:如果生成样本的 Frank 连接函数的参数 τ 为 0.2、0.5、0.8,采用独立删失的估计方法,得到的 γ 和 λ 估计量的偏差都较大,特别是 γ 估计值的偏差很大. 而采用本文方法(选取 Frank 连接函数,参数 $\tau = 0.8$)得到的估计量都比较理想,其估计值的偏差远远小于独立删失下估计值的偏差. 由此可见,在失效时间和删失时间不相互独立的情况下,采用独立删失方法进行估计会得到不理想的估计结果,因此,应该采用带相关性假定的模型进行分析.

其次,在参数 τ 的选取对估计量的影响方面,对参数 τ 为 0.5 和 0.8 的总体,似然方法采用 $\tau = 0.8$ 都能够得到偏差较小的估计量;然而,对参数 τ 为 0.2 的总体,采用 $\tau = 0.8$ 得到偏差较大的估计量. 但总的来说,在总体参数 τ 发生改变的条件下本文方法能够得到较稳健的估计量.

2.2 改变连接函数下参数估计的敏感性分析

分别取 Clayton Copula 和 Gumbel Copula 作为连接函数,$\tau = 0.8$,T 服从 $\gamma = 2, \lambda = 0.5$ 的威布尔分布,C 服从 $(0, 37)$ 上的均匀分布,生成容量为 200 的样本,删失比例 $P(T < C)$ 为 0.5. 在估计方法中采用 Frank 连接函数,$\tau = 0.8$. 2 种数据集下 $\tilde{\lambda}$ 和 $\tilde{\gamma}$ 的均值、标准差如表 2 所示.

表 2 总体的连接函数形式变化时参数 γ 和 λ 的估计

连接函数	参数	真实值	Frank 连接函数 $\tau = 0.8$			独立删失		
			估计值	标准差	偏差	估计值	标准差	偏差
Clayton Copula	γ	2.000	1.959	0.154	0.041	0.980	0.167	1.020
	λ	0.500	0.521	0.024	0.021	0.511	0.057	0.011
Grumbel Copula	γ	2.000	2.031	0.154	0.031	1.163	0.197	0.837
	λ	0.500	0.506	0.022	0.006	0.183	0.047	0.317

由表 2 可以看出：当 $\tau = 0.8$ 时，如果生成样本的连接函数分别选取 Clayton 和 Gumbel 连接函数，采用独立删失的估计方法，得到的 γ 和 λ 估计量的偏差都很大. 而采用本文方法（选取 Frank 连接函数，参数 $\tau = 0.8$）得到的估计量都比较理想，估计量的偏差比较小. 由此可见，在失效时间和删失时间不相互独立的情况下，采用独立删失方法进行估计可能会得到不理想的估计，所以应该采用带相关性假定的模型进行分析.

其次，在连接函数的选取对估计量的影响方面，当连接函数的假定和总体不一致时本文方法能够得到较稳健的估计量.

2.3　不同删失比例下参数估计的敏感性分析

选取 C 服从均匀 $U(0,3.3)$ 和 $U(0,4.0)$，删失比例 $P(T < C)$ 分别为 0.3、0.7. T 服从 $\gamma = 2$，$\lambda = 0.5$ 的威布尔分布，连接函数为 Frank，$\tau = 0.8$，生成容量为 200 的数据集，估计不同数据集下 $\tilde{\lambda}$ 和 $\tilde{\gamma}$ 的均值、标准差，估计结果如表 3 所示.

表 3　不同删失比例下参数 γ 和 λ 的估计

删失比例	参数	真实值	Frank 连接函数 $\tau = 0.8$			独立删失		
			估计值	标准差	偏差	估计值	标准差	偏差
0.3	γ	2.000	2.022	0.164	0.022	1.008	0.235	0.992
	λ	0.500	0.502	0.020	0.002	0.245	0.056	0.255
0.7	γ	2.000	2.020	0.142	0.020	1.171	0.197	0.829
	λ	0.500	0.504	0.023	0.004	0.184	0.068	0.316

由表 3 可以看出：如果连接函数及其参数 τ 的假定都是正确的，则在不同的删失比例下，本文方法都能够得到较好的估计，但采用独立删失方法得到的估计却不理想.

综合上述模拟计算的结果，可以得出：1）本文方法的参数估计效果比独立删失方法的参数估计效果更好；2）由连接函数不能准确识别或者参数 τ 不能正确识别所导致的偏差小于由独立删失错误假设所引起的偏差；3）当连接函数或参数 τ 的假定发生偏差时，本文方法依然能够较稳健.

3　讨论

通过上面的敏感性分析，可以看出在带信息的删失数据下估计参数的标准差小于在独立情况下估计参数的标准差. 且对于不同连接函数、相关系数以及删失比例，效果都较稳健. 另外，除了假设分布函数为威布尔分布外，还关于对数正态、指数分布等分布函数作了估计，效果也较好. 因此，文本提供的方法具有一定的实用价值.

本文的工作还有许多地方可以进一步深入地研究，如在本文方法的框架下，继续解决半参数模型的分布函数估计[15]；考虑有协变量影响下基于连接函数的带信息删失的非参数估计等.

4　参考文献

[1] Huang Jian, Wellner J A. Interval censored survival data: a review of recent progress [M]. New York: Springer-Verlag, 1997: 123-169.

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

[3] 吕秋萍，邓文丽. 区间删失数据函数的均值估计 [J]. 江西师范大学学报：自然科学版，2011，35(1)：96-100.

[4] 王纯杰. 基于 Copula 函数的相依删失数据的非参数统计推断 [D]. 长春：吉林大学，2012.

[5] Siannis F, Copas J, Lu Baobing. Sensitivity analysis for informative censoring in parametric survival models [J]. Biostatistics, 2005, 6(1): 77-91.

[6] Park Y, Lee Jenwei. One-and two-sample nonparametric inference procedures in the presence of a mixture of independent and dependent censoring [J]. Biostatistics, 2006, 7(2): 252-267.

[7] Zhang Zhigang, Sun Liuquan, Sun Jianguo, et al. Regression analysis of failure time data with informative interval censoring [J]. Statistics in Medicine, 2007, 26(12): 2533-2546.

[8] Huang Xuelin, Zhang Nan. Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach [J]. Biometrics, 2008, 64(4): 1090-1099.

[9] 邓文丽，朱莹莹. 一类保序最优化问题的迭代算法 [J]. 统计与决策，2011(14)：10-11.

[10] 张连增，胡祥. Copula 的参数与半参数估计方法的比较 [J]. 统计研究，2012，31(2)：91-95.

[11] Clayton D G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence [J]. Biometrika, 1978, 65(1): 141-151.

（下转第 592 页）

The Analysis and Measure of Allocation Efficiency of Financial Resources in China under Fiscal Decentralization

——Based on SE-DEA-TOBIT Two-Step Method

YANG Hai-wen[1,2], CHENG Li-wen[2], XU Ye[2], QI Ya-wei[2]

(1. School of Mathematics and Physics, Jinggangshan University, Ji'an Jiangxi 343009, China;
2. School of Accountancy, Jiangxi University of Finance & Economics, Nanchang Jiangxi 330013, China)

Abstract: Based on the model of SE-DEA-TOBIT of panel data, the efficiency of financial resources of 29 provinces in china during 1994—2011 are investigated. Firstly, the allocation efficiency of financial resources through SE-DEA based on the input-output structure of fiscal decentralization are measured. Then, the reason of inefficiency with the decomposition of Malmquist index is analyzed. Lastly, the relationship between the factors and the efficiency scores is investigated by TOBIT model. The experiential result shows that there are significant regional differences between financial resources allocation efficiency and the fiscal policy in 2002 expands the differences. The importance of "degree" of fiscal decentralization is proved.

Key words: fiscal decentralization; allocation efficiency of financial resources; SE-DEA-TOBIT; influencing factor

(责任编辑: 曾剑锋)

(上接第 581 页)

[12] Hougaard P. A class of multivariate failure time distributions [J]. Biometrika, 1986, 73(3): 671-678.

[13] Frank M J. On the sumultaneous association of $F(x, y)$ and $x + y - F(X, Y)$ [J]. Aequationes Mathematicae, 1979, 21(41): 37-38.

[14] Nelsen R B. An introduction to copulas [M]. 2nd ed. New York: Springer-Verlag, 2006.

[15] 杨金英, 赵培信. 缺失数据下 $\tilde{\rho}$ 混合误差线性模型的参数估计 [J]. 西南大学学报: 自然科学版, 2012, 34(9): 35-37.

The Parametric Estimation and Sensitivity Analysis for Information Interval Censoring

LI Wen-jing, DENG Wen-li*, ZHANG Ting-ting

(College of Mathematics and Informatics, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: The maximum likelihood function with information interval-censored data is constructed by copula function, and the distribution with informative interval censoring is studied. Different assumption about copula function will have different influence on the estimation result. Thus, sensitivity analysis of the influence is made by simulation.

Key words: information interval censoring; copula; sensitivity analysis

(责任编辑: 曾剑锋)