文章编号:1671-6833(2004)03-0102-03

截断情形下污染系数的估计

胡玉萍

(郑州大学系统科学与数学系,河南 郑州 450052)

摘 要:设 $X_1, X_2, \cdots, X_n$ 为一列非负独立同分布的随机变量,其分布为:$F_\alpha(x)=(1-\alpha)F_1(x)+\alpha F_2(x)$,其中 $\alpha\in[0,1]$,$F_1(x)$,$F_2(x)$ 都是定义在 R^+ 上的分布函数,现 $Y_1, Y_2, \cdots, Y_n$ 为非负 i.i.d $\sim G(t)$ 的截断随机变量列,并且 X_i 与 Y_i 也相互独立,在仅能观察到:$Z_i=\min(X_i, Y_i)$,$\delta_i=I(X_i\leqslant Y_i)(i=1,2,\cdots,n)$ 的情况下,给出了污染系数 α 的估计,并在 $G(t)$ 已知的情况下证明了其相合性.

关键词:污染数据;截断数据;污染系数.

中图分类号:O 212.7 **文献标识码**:A

0 引言

在实际问题中,我们经常遇到一类所谓污染数据(contaminated data),早在 1952 年,Davis[1] 就注意到在寿命试验中,元件寿命分布函数可能为两个分布函数的混合,考虑 $X_1, X_2, \cdots, X_n$ 为一列非负独立同分布的随机变量,具有分布函数 $F_\alpha(x)$

$$F_\alpha(x)=(1-\alpha)F_1(x)+\alpha F_2(x) \quad (1)$$

其中,$\alpha\in[0,1]$;$F_1(x)$,$F_2(x)$ 都是定义在半直线 $R^+=[0,\infty)$ 的分布函数.试验所观察到的元件寿命数据以概率 $1-\alpha$ 来自分布 $F_1(x)$,以概率 α 来自分布 $F_2(x)$,通常我们关心 $F_1(x)$,认为数据本应该服从 $F_1(x)$,但却受到了少量来自分布 $F_2(x)$ 的数据的污染,我们称 α 为污染系数.它衡量了数据受污染的程度,是我们所关心的.郑祖康[2] 已对两类污染数据回归分析的参数估计进行了讨论,潘建敏[3] 考虑了污染数据半参数回归分析的估计问题,任哲[4] 研究了文献[2]中提出的第Ⅰ类污染数据回归模型的参数的最小一乘估计,随着近几年研究成果的涌现,使得人们对有关污染数据的问题也更加重视起来.

在不少涉及污染模型的领域中,存在着 X_i 不能直接观察到,而是被另一组随机变量 Y_i 截断的情况,例如考虑 n 个元件的寿命试验,已知寿命分布服从式(1),由于受试验时间、费用等的限制,不可能将寿命试验进行到所有元件都失效,笔者[5] 对随机截断情形下污染数据半参数回归分析的估计问题进行了讨论,本文现考虑污染数据被截断的情况下污染系数 α 的估计问题.

1 截断模型下污染系数的估计

设 $X_1, X_2, \cdots, X_n$ 为独立同分布(i.i.d)的非负随机变量,共同的分布为 $F_\alpha(X)$,X_i 因右截断而不能完全观察,仅能观察到:

$$Z_i=\min(X_i, Y_i),\ \delta_i=I(X_i\leqslant Y_i)$$
$$(i=1,2,\cdots,n)$$

其中 $I(\cdot)$ 表示某事件的示性函数;$Y_1, Y_2, \cdots, Y_n$ 为非负 i.i.d $\sim G(t)$ 的截断随机变量列.现在的问题是如何利用上面所获得的截断数据 (Z_i, δ_i),$i=1,2,\cdots,n$ 来估计污染系数 α,为方便,本文使用下列记号:

$S=1-F_\alpha$;

$Z_i=\min(X_i, Y_i)$ 的分布函数为 $H(t)$;

$\tau_{F_\alpha}=\inf(t>0, F_\alpha(t)=1)$;

$\tau_G=\inf(t>0, G(t)=1)$;

$\tau_H=\inf(t>0, H(t)=1)$.

且假设 $\tau_{F_\alpha}<\tau_G\leqslant\infty$,显见 $1-H(t)=[1-F_\alpha(t)][1-G(t)]$,$\tau_H=\min(\tau_{F_\alpha}, \tau_G)=\tau_{F_\alpha}$.

关于 F_α 的估计量到目前为止,讨论较多的是 Kaplan-Meier 估计[6] 它的定义为

收稿日期:2004-04-06;**修订日期**:2004-06-11

基金项目:河南省自然科学基金资助项目(0211011000)

作者简介:胡玉萍(1971-),女,河南省尉氏县人,郑州大学讲师,硕士,主要从事数理统计的研究.

$$F_{KM}(t)=1-\prod_{Z_{(i)}\leqslant t}\left(1-\frac{1}{n-i+1}\right)^{\delta_{(i)}} \quad (2)$$

其中：$Z_{(1)}$，$Z_{(2)}$，…，$Z_{(n)}$ 为 Z_1，Z_2，…，Z_n 的次序统计量；$\delta_{(i)}$ 为与 $Z_{(i)}$ 相应的截断示性函数，当没有截断发生时（$\delta_{(i)}=1$，$i=1,2,\cdots,n$），$F_{KM}(t)$ 恰好为经验分布函数. 因此，KM 估计可以看作经验分布函数在截断情况下的推广，但此估计量没有使用 $G(t)$ 已知这一重要信息，因此从信息论的角度考虑，$F_{KM}(t)$ 不是最好的估计量.

当 $G(t)$ 已知时，下面我们引进 F_α 的另一个估计量，它利用了已知 $G(t)$ 这个假设和样本（Z_i，δ_i），$i=1,2,\cdots,n$ 的信息，具有很好的性质，下面的引理 1 和引理 2 由郑祖康（1995）首先提出并证明（见文献[7]），为方便，先给出如下定义：

定义 1　设函数 Φ_1，Φ_2 满足

$$\begin{cases}[1-G(x)]I(x>t)\Phi_1(t,x)+\int_0^x\Phi_2(t,y)I(y>t)\mathrm{d}G(y)=I(x>t) & (3)\\ \Phi_1,\ \Phi_2 \text{ 与 } F_\alpha \text{ 独立（但可依赖 } G)\end{cases}$$

我们称（Φ_1，Φ_2）属于类 K^*，记为（Φ_1，Φ_2）$\in K^*$（易见 $\Phi_1(t,x)=\Phi_2(t,x)=[1-G(t)]^{-1}$ 满足式（3），故 $K^*\neq\phi$）.

取（Φ_1，Φ_2）$\in K^*$，定义：

$$S_n(t)=\frac{1}{n}\sum_{i=1}^{n}[I(Z_i>t)\delta_i\Phi_1(t,Z_i)+I(Z_i>t)(1-\delta_i)\Phi_2(t,Z_i)] \quad (4)$$

取 $\Phi_1(t,x)=\Phi_2(t,x)=[1-G(t)]^{-1}$ 代入式（4）式，得

$$S_n(t)=\frac{1}{n}\sum_{i=1}^{n}I(Z_i>t)[1-G(t)]^{-1} \quad (5)$$

定义 2　记 $\mu_k\triangleq\int_0^{\tau_{F_\alpha}}kx^{k-1}S_n(x)\mathrm{d}x$

记 $F_\alpha(x)$，$F_1(x)$，$F_2(x)$ 的 k 阶矩分别为 μ_k，$\mu_k^{(1)}$，$\mu_k^{(2)}$，进一步我们假定可以选择到这样的 k，使得

$$\mu_k^{(1)}\neq\mu_k^{(2)}.$$

很自然，我们可构造 α 的估计为

$$\hat{\alpha}=\frac{\mu_k-\mu_k^{(1)}}{\mu_k^{(2)}-\mu_k^{(1)}}.$$

当 $G(t)$ 未知时，以 $1-G$ 的 Kaplan-Meier 估计

$$1-G_n(t)=\prod_{j=1}^{n}\left(\frac{N^+(Z_j)}{1+N^+(Z_j)}\right)^{I[Z_j\leqslant t,\ \delta_j=0]}\quad(t\geqslant 0)$$

作为 $1-G(t)$ 的估计，其中

$$N^+(Z_j)=\sum_{i=1}^{n}I[Z_i>Z_j],$$

记　$\bar{S}_n(t)=\frac{1}{n}\sum_{i=1}^{n}I(Z_i>t)[1-G_n(t)]$

定义 3　记 $\bar{\mu}_k\triangleq\int_0^{\tau_{F_\alpha}}kx^{k-1}\bar{S}_n(x)\mathrm{d}x$，假设可以选择到这样的 k，使得

$$\mu_k^{(1)}\neq\mu_k^{(2)},$$

则我们可构造 α 的估计为

$$\tilde{\alpha}=\frac{\bar{\mu}_k-\mu_k^{(1)}}{\mu_k^{(2)}-\mu_k^{(1)}}.$$

2　估计的强相合性

关于 α 的估计，在 $G(t)$ 已知时，我们可给出下面的定理.

定理　在模型式（1）下，若 $F_1(x)$，$F_2(x)$ 的某个 k 阶矩 $\mu_k^{(1)}$，$\mu_k^{(2)}$ 存在，且 $\mu_k^{(1)}\neq\mu_k^{(2)}$，则有：

（1）$\hat{\alpha}$ 为 α 的无偏估计；

（2）$\hat{\alpha}\to\alpha$，a.s.

为了给出上面定理的证明，我们先给出几个引理.

引理 1[7]　$ES_n(t)=S(t)=1-F_\alpha(t)$

引理 2[7]　$\mathrm{Var}[S_n(t)]=\inf_{[\Phi_1,\Phi_2]\in K^*}\mathrm{Var}[S_n(t)]$；

引理 1 表明式（4）定义的统计量是 $1-F_\alpha$ 无偏估计，引理 2 表明（5）式定义的统计量为类 K^* 中方差最小的一个.

引理 3　$E\mu_k=\mu_k$。

证明：由引理 1 可知：

$$\begin{aligned}E\mu_k&=E\int_0^{\tau_{F_\alpha}}kx^{k-1}S_n(x)\mathrm{d}x\\&=\int_0^{\tau_{F_\alpha}}k\cdot x^{k-1}S(x)\mathrm{d}x\\&=\int_0^{\tau_{F_\alpha}}k\cdot x^{k-1}(1-F_\alpha(x))\mathrm{d}x\\&=\mu_k.\end{aligned}$$

定理的证明：

设　$W_i=\int_0^{\tau_{F_\alpha}}\frac{I(Z_i>x)}{1-G(x)}k\cdot x^{k-1}\mathrm{d}x$；

$$\begin{aligned}\mu_k&=\int_0^{\tau_{F_\alpha}}k\cdot x^{k-1}S_n(x)\mathrm{d}x\\&=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_{F_\alpha}}\frac{kx^{k-1}I(Z_i>x)}{1-G(x)}\mathrm{d}x\\&=\frac{1}{n}\sum_{i=1}^{n}W_i,\end{aligned}$$

由引理 3 知，$E\mu_k=\mu_k$，故

$$E(\hat{\alpha})=\frac{E\mu_k-\mu_k^{(1)}}{\mu_k^{(2)}-\mu_k^{(1)}}=\frac{\mu_k-\mu_k^{(1)}}{\mu_k^{(2)}-\mu_k^{(1)}}.$$

又由式(1)可得

$$\mu_k = (1-\alpha)\mu_k^{(1)} + \alpha\mu_k^{(2)},$$

因此 $E(\hat{\alpha}) = \alpha.$

即 $\hat{\alpha}$ 为 α 的无偏估计.

由 $\tau_{F_\alpha} = \tau_H$,得

$$\begin{aligned} W_i &= \int_0^{\tau_{F_\alpha}} \frac{kx^{k-1}I(Z_i > x)}{1-G(x)}\mathrm{d}x \\ &= \int_0^{\min(\tau_{F_\alpha}, Z_i)} \frac{kx^{k-1}}{1-G(x)}\mathrm{d}x \\ &= \int_0^{Z_i} \frac{kx^{k-1}}{1-G(x)}\mathrm{d}x; \\ E(W_i) &= \int_0^{\tau_{F_\alpha}} E\left(\frac{kx^{k-1}I(Z_i > x)}{1-G(x)}\right)\mathrm{d}x \\ &= \int_0^{\tau_{F_\alpha}} kx^{k-1}E\left(\frac{I(Z_i > x)}{1-G(x)}\right)\mathrm{d}x \\ &= \int_0^{\tau_{F_\alpha}} kx^{k-1}(1-F_0(x))\mathrm{d}x \\ &= \int_0^{\tau_{F_\alpha}} kx^{k-1}S(x)\mathrm{d}x \\ &= \mu_k, \end{aligned}$$

由柯尔莫哥洛夫定理,可得

$$\mu_k \to \mu_k, \text{ a.s.}$$

因此, $\hat{\alpha} \to \alpha$, a.s.

证毕.

参考文献:

[1] DAVIS D J. An analysis of some failure data [J]. Journal of American Statistics Association, 1952, 47: 113~150.

[2] 郑祖康,杨邦俊,杨 瑛,等. 关于两类污染数据回归分析的参数估计[J]. 高校应用数学学报(A辑), 1996, (11): 31~40.

[3] 潘建敏. 污染数据半参数回归模型的估计方法[J]. 工程数学学报, 1997, 14(3): 81~84.

[4] 任 哲,陈明华. 污染数据回归分析中参数的最小一乘估计[J]. 应用概率统计, 2000, (16): 262~268.

[5] 胡玉萍, 陆宜清. 截断情形下污染数据半参数回归模型估计方法[J]. 郑州大学学报(工学版), 2004, 25(2): 91~93.

[6] KAPLAN E L, MEIER P. Nonparametric estimation from incomplete observations [J]. Journal of American Statistics Association, 1958, (53): 457~481.

[7] ZHENG Zu-kang. The estimation of σ^2 in linear regressin with censored data [J]. Chin Ann of Math, 1993, (14B): 319~326.

Estimation of Contamination Coefficients of Censored Data

HU Yu-ping

(Department of System Science & Mathematics, Zhengzhou University, Zhengzhou 450052, China)

Abstract: $X_1, X_2, \cdots, X_n$ are a sequence of i.i.d random variables with the distribution function of $F_0(x) = (1-\alpha)F_1(x) + \alpha F_2(x)$, where $\alpha \in [0,1]$, and $F_1(x)$ and $F_2(x)$ are distribution functions. $Y_1, Y_2, \cdots, Y_n$ are assumed nonnegative, independent, and identically distributed random variables with a common distribution function $G(t)$, the observations are pairs (Z_i, δ_i), $i = 1, 2, \cdots, n$ with $Z_i = \min(X_i, Y_i)$ and δ_i. This paper presents an estimation of α and proves its strong consistency.

Key words: contaminated data; censored data; contamination coefficient