文章编号:1671-6833(2004)02-0091-03

截断情形下污染数据半参数回归模型估计方法

胡玉萍[1], 陆宜清[2]

(1.郑州大学系统科学与数学系,河南 郑州 450052;2.郑州牧业工程高等专科学校,河南 郑州 450008)

摘 要:对半参数回归模型 $y_j = x_j\beta + g(t_j) + \varepsilon_j, j = 1,2,\cdots,n$ 进行分析.其中,(x_j, t_j) 为取值于 $R\times[0,1]$ 上的固定设计,β 为未知参数,g 是定义在 $[0,1]$ 上的未知函数,ε 为随机误差,$E\varepsilon = 0, E\varepsilon^2 = \sigma_1^2$.但 $y_j, \cdots, y_n$ 受到另一独立同分布的随机变量序列 $u_1,\cdots,u_n$ 两种不同方式的污染,u_j 与 y_j 独立,同时,它们被另一独立同分布的随机变量序列截断.最后,利用最小二乘法及矩估计方法给出随机截断情形下两种污染方式的 α,β 和污染参数的估计.

关键词:污染参数;截断数据;半参数回归模型;污染数据

中图分类号:O 212.1 **文献标识码**:A

0 引言

同删失数据一样,在实际工作中也经常遇到一些关于污染数据的统计分析问题.目前的研究中,对于删失数据已得到了一系列较为成熟的研究成果,但对污染数据问题却研究得甚少,Huber[1] 于 1964 年首次提出了一类"被污染的正态分布族":

$$F_{N,\nu} = \{f: f = (1-\nu)N(0.1) + \nu g, g \in F_s\},$$

其中:ν 为污染参数,F_s 为一切关于原点对称的一维概率密度函数族.YU[2] 利用矩方法对污染参数进行了估计,郑祖康[3] 已对污染数据的回归分析问题讨论了两类污染数据参数 ν 与回归参数 α,β 的点估计.胡玉萍[4] 针对污染数据的回归分析问题,讨论了回归参数 α,β 的区间估计.潘建敏[5] 对污染数据半参数回归分析的估计问题进行了讨论,而关于随机截断情形下污染数据半参数回归分析的估计问题迄今鲜见讨论.本文现考虑此种问题的参数估计,推广了文献[3]中的结果.

1 第Ⅰ类污染模型的估计方法

郑祖康在文献[3]中提出第Ⅰ类简单线性模型:

$$y_i = \alpha + \beta x_j + \varepsilon, j = 1,2\cdots,n,$$

式中:ε 相互独立,服从 $N(0,\sigma_1^2)$;$\{y_i\}$ 受到另一串与之独立的随机变量$\{u_j\}$的干扰;u_j 相互独立,服从 $N(0,\sigma_2^2)$;σ_1^2, σ_2^2 均已知,仅能观察到:

$$y_j^* = (1-\nu)y_j + \nu u_j, \quad 0 \leqslant \nu < 1,$$

且要求

$$\sigma_1^2/\sigma_2^2 > \nu/(1-\nu), \quad 0 \leqslant \nu < 1 \qquad (1)$$

现考虑半参数回归模型

$$y_j = \beta x_j + g(t_j) + \varepsilon, \quad j = 1,2,\cdots,n \qquad (2)$$

式中:(x_j, t_j) 为取值于 $R\times[0,1]$ 上的固定设计,β 为未知数,g 是定义在 $I=[0,1]$ 上的未知函数.我们本应观察到:

$$y_j^* = (1-\nu)y_j + \nu u_j =$$
$$(1-\nu)\beta x_j + (1-\nu)g(t_j) + \eta_j,$$

其中,

$$\begin{cases} \eta_j = (1-\nu)\varepsilon + \nu u_j, j = 1,2,\cdots,n, \\ i.i.d \sim N(0,(1-\nu)^2\sigma_1^2 + \nu^2\sigma_2^2) \end{cases} \qquad (3)$$

但在一些实际问题中,如可靠性寿命试验、医药追踪试验及对生存分析等领域的研究中,y_j^* 常常因随机右截断而不能被完全观察,我们仅能观察到:

$Z_j = \min(Y_j^*, C_j)$, $\delta_j = I[Y_j^* \leqslant C_j]$, $j = 1,2,\cdots,n$,

式中:$I(\cdot)$ 表示某事件的示性函数;$C_1, C_2, \cdots, C_n$ 表示截断的随机变量列.

以下均假定 $C_1, C_2, \cdots, C_n$ 独立同分布,有共同的连续分布函数 G,$Y_1^*, Y_2^*, \cdots, Y_n^*$ 由前面假

收稿日期:2004-01-06;**修订日期**:2004-04-01

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.

设显然是彼此独立的,且现设 Y_j^* 的分布函数用 F_j 表示($j=1,2,\cdots,n$).易见 Z_j 的分布函数 $H_j=1-(1-F_j)(1-G)$,$j=1,2,\cdots,n$.此外,为方便计,对任何分布函数 $V(\cdot)$,定义 $V_s(\cdot)=1-V$,$\tau_v=\inf\{t:V(t)=1\}$,对任何 $r>0$,定义 $V^{-r}(\cdot)=\{V(\cdot)\}^{-r}$.

以下始终假定 $\tau_{F_j}\leqslant\tau_G$,$j=1,2,\cdots,n$,$g(t)$ 在 I 上连续.注意到:

$$E\delta Z_jG_s^{-1}(Z_j)=\int_{-\infty}^{\tau_{F_j}}\int_{y}^{\tau_G}\frac{y}{1-G(y)}\mathrm{d}G(t)\mathrm{d}F_j(y)=EY_j^*=(1-v)\beta x_j+(1-v)g(t_j),$$

由此我们认为 $\{\delta Z_jG_s^{-1}(Z_j),j=1,2,\cdots,n\}$ 遵从如下模型:

$$\delta Z_jG_s^{-1}(Z_j)=(1-v)\beta x_j+(1-v)g(t_j)+\eta_j \tag{4}$$

记 $Z_{jG}=\delta Z_jG_s^{-1}(Z_j)$,$j=1,2,\cdots,n$,于是 G 已知时,仿文献[6]中的方法,在此构造 $g(\cdot)$ 的估计:

$$\hat{g}_n(t)\triangleq g_n(t,\beta)=\sum_{j=1}^{n}W_{nj}(t)(Z_{jG}-(1-v)\beta x_j)/(1-v)\triangleq\frac{1}{1-v}\hat{g}_{1n}^*(t)-\beta\hat{g}_{2n}(t) \tag{5}$$

其中,$\hat{g}_{1n}^*(t)=\sum_{j=1}^{n}W_{nj}(t)Z_{jG}$,$\hat{g}_{2n}(t)=\sum_{j=1}^{n}W_{nj}(t)x_j$,$0\leqslant w_{nj}(t)\leqslant1$ 为权函数,$j=1,2,\cdots,n$.

则基于式(4),以 $\hat{g}_n(t)$ 代替 $g(t)$.由最小二乘估计可知

$$(1-\hat{v})\beta=\sum_{j=1}^{n}\tilde{x}_j\tilde{Z}_{jG}/\sum_{j=1}^{n}\tilde{x}_j^2 \tag{6}$$

其中,

$$\tilde{Z}_{jG}=Z_{jG}-\sum_{i=1}^{n}W_{ni}(t_j)Z_{iG},$$

$$\tilde{x}_j=x_j-\sum_{i=1}^{n}W_{ni}(t_j)x_i,$$

又

$$Z_{jG}\sim N((1-v)(\beta x_j+g(t_j)),(1-v)^2\sigma_1^2+v^2\sigma_2^2),$$

则 Z_{jG} 的方差估计为

$$R_n=\frac{1}{n-2}\sum_{j=1}^{n}\{\tilde{Z}_{jG}-(1-\hat{v})\beta\tilde{x}_j\}^2.$$

令

$$R_n=(1-\hat{v})^2\sigma_1^2+\hat{v}^2\sigma_2^2,$$

再注意到式(1),可得

$$\hat{v}=\{\sigma_1^2-\sqrt{(\sigma_1^2+\sigma_2^2)R_n-\sigma_1^2\sigma_2^2}\}/(\sigma_1^2+\sigma_2^2),$$

从而

$$\beta=\sum_{j=1}^{n}\tilde{Z}_{jG}\tilde{x}_j/\{(1-\hat{v})\sum_{j=1}^{n}\tilde{x}_j^2\} \tag{7}$$

$$\hat{g}_n^*(t)=\frac{1}{1-\hat{v}}\hat{g}_{1n}^*(t)-\beta\hat{g}_{2n}(t) \tag{8}$$

2 第Ⅱ类污染模型的估计方法

郑祖康在文献[3]中提出第Ⅱ类简单线性模型:

$$y_j=\alpha+\beta x_j+\varepsilon_j,j=1,2,\cdots,n \tag{9}$$

式中:ε_j 相互独立,服从 $N(0,\sigma_1^2)$,$\{y_j\}$ 受到另一串与之独立的随机变量 $\{u_j\}$ 的干扰,u_j 相互独立,服从 $N(0,\sigma_2^2)$,仅能观察到 $\{y_j^*\}$,y_j^* 的分布函数为

$$F_{y_j^*}(y)=(1-v)F_{y_j}(y)+vF_{u_j}(y),0\leqslant v\leqslant1 \tag{10}$$

式中:$F_{y_j}(y)$ 与 $F_{u_j}(y)$ 分别为 y_j 与 u_j 的分布函数.

现仍考虑半参数回归模型(2),且仍是仅能观察到 Z_j.又由于 $\{y_j\}$ i.i.d $\sim N(x_j\beta+g(t_j),\sigma_1^2)$,故 y_j^* 的密度函数为

$$f_{y_j^*}(y)=\frac{1-v}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{(y-\beta x_j-g(t_j))^2}{2\sigma_1^2}\right\}+\frac{v}{\sqrt{2\pi\sigma_2^2}}\exp\left\{-\frac{y^2}{2\sigma_2^2}\right\}.$$

则可计算得到

$$Ey_j^*=(1-v)(\beta x_j+g(t_j)) \tag{11}$$

$$E(y_j^*)^2=(1-v)\{(\beta x_j+g(t_j))^2+\sigma_1^2\}+v\sigma_2^2 \tag{12}$$

$$E(y_j^*)^3=(1-v)(\beta x_j+g(t_j))^3+3(1-v)\cdot(\beta x_j+g(t_j))\sigma_1^2 \tag{13}$$

$$E(y_j^*)^5=(1-v)(\beta x_j+g(t_j))^5+10(1-v)\cdot(\beta x_j+g(t_j))^3\sigma_1^2+15(1-v)(\beta x_j+g(t_j))\sigma_1^4 \tag{14}$$

令

$$Z_{jG}^k=Z_j^k\delta G_s^{-1}(Z_j),$$

则显然有

$$E(Z_{jG})=E(\delta Z_jG_s^{-1}(Z_j))=E(Y_j^*) \tag{15}$$

$$E(Z_{jG}^k)=E(Z_j^k\delta G_s^{-1}(Z_j))=\int_{-\infty}^{\tau_{F_j}}\int_{y}^{\tau_G}\frac{y^k}{1-G(y)}\mathrm{d}G(t)\mathrm{d}F_j(y)=E(Y_j^*)^k \tag{16}$$

由于 $E(Y_j)=\beta x_j+g(t_j)$,并注意到式(11)、式(15),再次利用非参数的权函数估计法,以 $\frac{1}{1-v}Z_{jG}$ 代替 y_j,得 g 估计量同式(5),以 $\hat{g}_n(t)$ 代替 $g(t)$,首先考虑 σ_1^2,σ_2^2 已知的情况下,β,v 及 g 的估计.由矩估计原理得

$$\sum_{j=1}^{n}Z_{jG}=(1-v)\left\{\beta\sum_{j=1}^{n}\tilde{x}_j+\frac{1}{1-v}\sum_{j=1}^{n}\hat{g}_{1n}^*(t_i)\right\} \tag{17}$$

$$\sum_{j=1}^{n}Z_{jG}^2=(1-v)\left\{\sum_{j=1}^{n}(\beta\tilde{x}_j+\frac{1}{1-v}\hat{g}_{1n}^*(t_j))^2+\sigma_1^2\right\}+v\sigma_2^2 \tag{18}$$

由式(17)得

$$(1-v)\beta=\sum_{j=1}^{n}Z_{jG}/\sum_{j=1}^{n}\tilde{x}_j \tag{19}$$

将式(19)代入式(18),有

$$(\sigma_1^2-\sigma_2^2)(1-v)^2-\left[\sum_{j=1}^{n}Z_{jG}^2-\sigma_2^2\right](1-v)+d_0=0 \tag{20}$$

其中,

$$d_0=\left[\sum_{j=1}^{n}Z_{jG}\right]^2\sum_{j=1}^{n}\tilde{x}_j^2\Big/\left[\sum_{j=1}^{n}\tilde{x}_j\right]^2+2\sum_{j=1}^{n}Z_{jG}\cdot\sum_{j=1}^{n}(\tilde{x}_j\tilde{g}_{1n}^{*}(t_i))\Big/\sum_{j=1}^{n}\tilde{x}_j+\sum_{j=1}^{n}\tilde{g}_{1n}^{*}(t_j),$$

若 $\sigma_1^2=\sigma_2^2\triangleq\sigma^2$,则

$$\hat{v}=1-d_0/\left[\sum_{j=1}^{n}Z_{jG}^2-\sigma^2\right] \tag{21}$$

$$\beta=\sum_{j=1}^{n}Z_{jG}/\left[(1-\hat{v})\sum_{j=1}^{n}\hat{x}_j\right] \tag{22}$$

而 g 的最终估计同式(7),但若 $\sigma_1^2\neq\sigma_2^2$,则

$$\hat{v}=1-\left\{\left(\sum_{j=1}^{n}Z_{jG}^2-\sigma_2^2\right)\pm\sqrt{\left(\sum_{j=1}^{n}Z_{jG}^2-\sigma_2^2\right)-4d_0(\sigma_1^2-\sigma_2^2)}\right\}/2(\sigma_1^2-\sigma_2^2),$$

正负号的选择服从 $0<v<1$,而 β 与 $\hat{g}_n(t)$ 同式(22)与式(7).考虑 σ_2^2 未知的情形,β,v,σ_1^2 及 g 的矩估计可通过解以下方程组

$$\sum_{j=1}^{n}Z_{jG}=(1-v)\left\{\beta\sum_{j=1}^{n}\tilde{x}_j+\frac{1}{1-v}\sum_{j=1}^{n}\hat{g}_{1n}^{*}(t_j)\right\} \tag{23}$$

$$\sum_{j=1}^{n}Z_{jG}^3=(1-v)\sum_{j=1}^{n}\left\{\beta\tilde{x}_j+\frac{1}{1-v}\hat{g}_{1n}^{*}(t_j)\right\}^3+3(1-v)\cdot\sum_{j=1}^{n}\left\{\beta\tilde{x}_j+\frac{1}{1-v}\hat{g}_{1n}^{*}(t_j)\right\}\sigma_1^2 \tag{24}$$

$$\sum_{j=1}^{n}Z_{jG}^5=(1-v)\sum_{j=1}^{n}\left\{\beta\tilde{x}_j+\frac{1}{1-v}\hat{g}_{1n}^{*}(t_j)\right\}^5+10(1-v)\sum_{j=1}^{n}\left\{\beta\tilde{x}_j+\frac{1}{1-v}\hat{g}_{1n}^{*}(t_j)\right\}^3\sigma_1^2+15(1-v)\cdot\sum_{j=1}^{n}\left\{\beta\tilde{x}_j+\frac{1}{1-v}\hat{g}_{1n}^{*}(t_j)\right\}\sigma_1^4 \tag{25}$$

若记 $C_0=\sum_{j=1}^{n}Z_{jG}/\sum_{j=1}^{n}\tilde{x}_j$,$h_j=C_0\tilde{x}_j+\hat{g}_{1n}^{*}(t_j)$,则由式(23)~(25)可得 σ_1^2 的一元二次方程为

$$A\sigma_1^4+B\sigma_1^2+C=0 \tag{26}$$

其中,

$$A=9\left[\sum_{j=1}^{n}h_j\right]^2\left[\sum_{j=1}^{n}h_j^5\right]\left[\sum_{j=1}^{n}h_j^3\right]^{-2}-15\sum_{j=1}^{n}h_j;$$

$$B=10\left[\sum_{j=1}^{n}Z_{jG}^2\right]-6\left[\sum_{j=1}^{n}Z_{jG}^3\right]\left[\sum_{j=1}^{n}h_j\right]\left[\sum_{j=1}^{n}h_j^5\right]\left[\sum_{j=1}^{n}h_j^3\right]^{-2};$$

$$C=\left[\sum_{j=1}^{n}Z_{jG}^3\right]^2\left[\sum_{j=1}^{n}h_j^5\right]\left[\sum_{j=1}^{n}h_j^3\right]^{-2}-\sum_{j=1}^{n}Z_{jG}^5;$$

判别式

$$\triangle=B^2-4AC=4\left[\sum_{j=1}^{n}h_j^3\right]^{-2}\left\{5\left[\sum_{j=1}^{n}Z_{jG}^3\right]^2-3\left[\sum_{j=1}^{n}Z_{jG}^5\right]\cdot\left[\sum_{j=1}^{n}h_j\right]\right\}\times\left\{5\left[\sum_{j=1}^{n}h_j^3\right]^2-3\left[\sum_{j=1}^{n}h_j^5\right]\left[\sum_{j=1}^{n}h_j\right]\right\}.$$

则由式(26)得到

$$\hat{\sigma}_1^2=\frac{5\left[\sum_{j=1}^{n}Z_{jG}^3\right]-3\left[\sum_{j=1}^{n}Z_{jG}^3\right]\left[\sum_{j=1}^{n}h_j\right]\left[\sum_{j=1}^{n}h_j^5\right]\left[\sum_{j=1}^{n}h_j^3\right]^{-2}\pm\sqrt{\triangle}/2}{15\left[\sum_{j=1}^{n}h_j\right]-9\left[\sum_{j=1}^{n}h_j\right]^2\left[\sum_{j=1}^{n}h_j^5\right]\left[\sum_{j=1}^{n}h_j^3\right]^{-2}},$$

且可推得,若

$$\left[\sum_{j=1}^{n}h_j^3\right]\left\{5\left[\sum_{j=1}^{n}h_j^3\right]^2-3\left[\sum_{j=1}^{n}h_j^5\right]\left[\sum_{j=1}^{n}h_j\right]\right\}\geqslant 0,$$

则 $\hat{\sigma}_1^2$ 计算公式中取负号,反之则取正号,然后由式(23),(24)便可解得 $\beta,\hat{v};\hat{g}_n^{*}(t)$ 仍由式(8)计算.

参考文献:

[1] HUBER Peter J.Robust estimation of a location parameter [J].Ann Math Staist,1964,35:73~101.

[2] YU K F.A note on the estimation of the mixing parameter in mixture of two distribution[R].Carolina:University of South Carolina,1990.

[3] 郑祖康,丁邦俊,杨　瑛,等.关于两类污染数据回归分析的参数估计[J].高校应用教学学报,1996,11A(1):31~39.

[4] 胡玉萍,王　霞,李学相.污染数据回归分析参数的区间估计[J].郑州大学学报(工学版),2003,24(4):99~101.

[5] 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.

[6] 王启华.随机截断下半参数回归模型中的相合估计[J].中国科学,1995,25(8):819~832.

(下转第 100 页)

plating solution are compared, reaching the conclusion that the effects of Y and Eu on the stability of plating solution are more remarkable, and those of NH_4F and Na_2CO_3 on the rate of deposition are more notable especially the stability of plating solution is enhanced to the most extent when the added concentration of Y^{3+} is 0.02 g/L while effect of acceleration is the obvious when the added concentration of Na_2CO_3 is about 19 g/L. Also this paper analyzes the mechanism of effect of the additives on the rate of deposition and the stability of plating solution.

Key words: magnesium alloy; electroless plating; additives; stability; rate of deposition

(上接第93页)

Estimation Method of Semiparametric Regression Model with Contaminated and Censored Data

HU Yu-ping[1], LU Yi-qing[2]

(1. Department of System Science & Mathematics, Zhengzhou University, Zhengzhou 450052, China; 2. Zhengzhou College of Animal Husbandry Engineering, Zhengzhou 450008, China)

Abstract: This paper studies the semiparametric regression model $y_1 = x_j\beta + g(t_j) + \xi_j, j = 1, 2, \cdots, n$, where $E\xi_j = 0, E\xi_j^2 = \sigma_1^2$. But $y_1, \cdots, y_n$ are contaminated by another i.i.d random variable sequence $u_1, \cdots, u_n$ in two different ways. And $\{u_i\}$ is independent of $\{y_i\}$ while they are censored by another i.i.d random variable sequence. This paper also presents the estimations of β, g and contamination parameter respectively.

Key words: Censored data; semiparametric regression model; contamination parameter; contamination data