协变量带误差的随机删失数据线性模型的一类半参数估计

刘网定[1,2], 王海康[2], 周秀轻[2]

(1. 南京陆军指挥学院, 江苏 南京 210045)
(2. 南京师范大学数学科学学院, 江苏 南京 210097)

[摘要] 考虑线性模型 $Y_i = X_i^{\mathrm{T}} \quad + \quad _i$, $i = 1, 2, \quad , n$, 其中 Y_i 为随机右删失因变量, X_i 难以观测, 或需要较高成本才能得到其精确观测值, 故转而观测与 X_i 相关的相对易得的随机变量 X_i. 利用数据集 $\{(X_j, X_j)\}_{j=n+1}^{n+N}$ 估计 $E[X|X]$ 的同时, 给出了参数向量 的一类半参数估计, 并且证明了估计量的渐近正态性.

[关键词] 随机删失数据, 线性回归模型, 协变量带误差, 渐近正态性
[中图分类号] O212 [文献标识码] A [文章编号] 1001-4616(2009)02-0031-05

A Class of Semiparametric Estimators for Linear Error-in-Covariables Models Under Random Censorship

Liu Wangding[1,2], Wang Haikang[2], Zhou Xiuqing[2]

(1. Nanjing Army Command College, Nanjing 210045, China)
(2. School of Mathematical Science, Nanjing Normal University, Nanjing 210097, China)

Abstract Consider the linear models $Y_i = X_i^{\mathrm{T}} \quad + \quad _i$, $i = 1, 2, \quad , n$, with the response Y_i censored randomly on the right and X_i hard or costly to measure. A relative variable X is to be measured instead of X. When estimating $E[X|X]$ with the help of validation data $\{(X_j, X_j)\}_{j=n+1}^{n+N}$, a class of semiparametric estimators for vector are given and the asymptotic normality is also proved.

Key words random censoring, linear models, errors in covariables, asymptotic normality

回归模型是一个常用的重要模型. 考虑线性回归模型

$$Y_i = X_i^{\mathrm{T}} \quad + \quad _i, \quad i = 1, 2, \quad , n, \tag{1}$$

其中 Y_i 为数值型因变量, X_i 为 p 维随机协变量, X_i^{T} 是 X_i 的转置, $= (_1, _2, \quad , _p)^{\mathrm{T}}$ 是 p 维待估参数, $_i$ 是随机误差, 并且在给定 X_i 的条件下, 误差项 $_i = Y_i - X_i^{\mathrm{T}}$ 独立同分布.

当因变量发生右删失时, 某些 Y_i 无法观测到, 观测到的是 $Z_i = \min(Y_i, T_i)$ 和 $_i = I(Y_i \quad T_i)$, 其中 $I(A)$ 是集合 A 的指示函数, T_i 为随机删失变量. 此时, 对模型(1)参数的 LSE估计见文献[1-5]. 在有些情况下, 删失是由某些自然原因或突发事件而产生的, 此时, 删失变量的分布是未知的. 例如, 在医学研究中, 由于病人的中途退出, 或在观测结束时仍然存活等等, 造成对病人生存时间的观测往往是不完全的. 但在其它一些情况下, 删失是由于人为的设定与控制而造成. 例如, 在工程技术方面, 对一些物件进行(毁坏性)测试, 但不希望无休止地等待下去直至所有的物件都毁坏, 因为那将耗费太大的人力财力, 所以可以设定试验的时间不能超过某个界限. 在这种情况下, 删失变量的分布是完全已知的. 有不少作者对删失分布已知情况下的统计模型作过研究, 例如文献[6, 7].

另外, 在许多研究背景中, 存在某种变量(如 X)难以观测到, 或者对 X 的精确观测需要较高的成本, 从而转向观测一个与 X 相关的且相对容易观测到的变量(如 X), 用 $E[X|X]$ 来代替 X, 这样的例子见文献[8-12]. 此时称协变量 X 带误差. 一般情况下, 并不能指出 X 与 X 之间的确切关系, 但是可以通过某种代

收稿日期: 2008-07-12
通讯联系人: 刘网定, 助教, 研究方向: 生存分析和回归分析. E-mail: memoryliu@163.com

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

价比较大的途径来获得一些$\boldsymbol{X}$和$\boldsymbol{X}$的观测值(记为$\{(\boldsymbol{X}_j, \boldsymbol{X}_j)\}_{j=n+1}^{n+N}$),通过$\{(\boldsymbol{X}_j, \boldsymbol{X}_j)\}_{j=n+1}^{n+N}$来估计$\boldsymbol{X}$与$\boldsymbol{X}$之间的关系,详见文献[10-14].

当Y发生右删失,$\boldsymbol{X}$带误差时,模型(1)可以改写为

$$\begin{cases} Y_i = \ ^{\mathrm{T}}(\boldsymbol{X}_i) + e_i, \\ e_i = \ _i + (\boldsymbol{X}_i - (\boldsymbol{X}_i))^{\mathrm{T}} , \\ (\boldsymbol{X}_i) = E[\boldsymbol{X}_i \mid \boldsymbol{X}_i]. \end{cases} \tag{2}$$

文献[14]考虑了模型(2)的参数估计问题.该文利用Kaplan-Meier估计法来估计删失变量的分布函数$G(\ )$,利用核估计法来估计$\boldsymbol{X}$与$\boldsymbol{X}$的关系$E[\boldsymbol{X} \mid \boldsymbol{X}]$,最后使用Koul-Susarla-Van Ryzin估计的方法调整删失数据,得到 的估计量$\hat{\ }_{n,N}$.众所周知,在协变量$\boldsymbol{X}$能准确观测到时,Koul-Susarla-Van Ryzin估计是将所有观测到的Z_i都进行了调整,这使得所有未被截断的Y_i均被提高,因此在进行随机模拟时,Koul-Susarla-Van Ryzin估计的效果不是很理想[6,15].

本文中,我们在假定$G(\ )$已知的情况下,对模型(2)的参数估计问题作进一步的研究.给出 的一类半参数估计方法,并且得到了估计量$\hat{\ }_{n,N}$的渐近正态性.

1 估计方法

对于模型(2),设$\boldsymbol{X}$为d维随机变量,删失变量T的分布$G(\ )$为已知,且具有密度函数$g(\ ) > 0$,随机误差满足$E[\ \mid \boldsymbol{X}] = 0$并且假设在给定$\boldsymbol{X}$的条件下,$Y$与$T$相互独立,且$T$与$\boldsymbol{X}$相互独立.观测数据$\{(Z_i, \ _i, \boldsymbol{X}_i)\}_{i=1}^{n}$独立同分布.根据文献[15]中Class K估计的思想方法,我们将删失数据调整为

$$Y_i^* = \ _i \ _1(Z_i) + (1 - \ _i) \ _2(Z_i), \tag{3}$$

此处 $_1$、 $_2$为连续函数,且满足

$$\begin{cases} [1 - G(y)] \ _1(y) + \int_{-}^{y} \ _2(t)\,\mathrm{d}G(t) = y, \\ \ _1、\ _2 \text{与} Y_i\text{的分布无关,但有可能依赖于} G. \end{cases}$$

由条件$E[\ _i \mid \boldsymbol{X}_i] = 0$,可以验证$E[Y_i^* \mid \boldsymbol{X}_i] = \ ^{\mathrm{T}}(\boldsymbol{X}_i)$, $i = 1, 2, \ , n$.即对于模型

$$Y_i^* = \ ^{\mathrm{T}}(\boldsymbol{X}_i) + e_i, \tag{4}$$

有$E[e_i \mid \boldsymbol{X}_i] = 0$, $i = 1, 2, \ , n$.

因此对模型(4)可以利用最小二乘估计,得到 的估计量为

$$\ _n = \ _n^{-1}\boldsymbol{A}_n, \tag{5}$$

其中$(\ )$为已知, $\ _n = \dfrac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i)\ ^{\mathrm{T}}(\boldsymbol{X}_i)$, $\boldsymbol{A}_n = \dfrac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i)Y_i^*$.

由于$\boldsymbol{X}$与$\boldsymbol{X}$之间的具体关系$(\ )$通常是未知的,类似于文献[12]和[13],我们可以通过某种代价比较大的途径来得到一组独立同分布的观测值$\{(\boldsymbol{X}_j, \boldsymbol{X}_j)\}_{j=n+1}^{n+N}$,并且与前面的$\{(Z_i, \ _i, \boldsymbol{X}_i)\}_{i=1}^{n}$相独立,利用$\{(\boldsymbol{X}_j, \boldsymbol{X}_j)\}_{j=n+1}^{n+N}$对$(\ )$进行核估计.因此对于$\boldsymbol{x}\ \ \boldsymbol{x}$, $(\boldsymbol{X}_i)$的估计量定义为

$$\hat{\ }_N(\boldsymbol{x}) = \frac{\sum_{j=n+1}^{n+N} \boldsymbol{X}_j K\left(\dfrac{\boldsymbol{x} - \boldsymbol{X}_j}{h_N}\right)}{\sum_{j=n+1}^{n+N} K\left(\dfrac{\boldsymbol{x} - \boldsymbol{X}_j}{h_N}\right)}, \tag{6}$$

其中$K(\ )$是一个核函数,h_N是一列趋于0的正常数,$\boldsymbol{x}$是$\boldsymbol{X}$的支撑.参数 的估计量定义为

$$\hat{\ }_{n,N} = \hat{\ }_{n,N}^{-1}\hat{\boldsymbol{A}}_{n,N}, \tag{7}$$

其中

$$\hat{\ }_{n,N} = \frac{1}{n}\sum_{i=1}^{n} \hat{\ }_N(\boldsymbol{X}_i)\hat{\ }_N^{\mathrm{T}}(\boldsymbol{X}_i), \tag{8}$$

$$\hat{\boldsymbol{A}}_{n,N} = \frac{1}{n}\sum_{i=1}^{n} \hat{\ }_N(\boldsymbol{X}_i)Y_i^* = \frac{1}{n}\sum_{i=1}^{n} \hat{\ }_N(\boldsymbol{X}_i)[\ _i \ _1(Z_i) + (1 - \ _i) \ _2(Z_i)]. \tag{9}$$

在下节中,我们将对$\hat{\ }_{n,N} -$ 进行分解,进而给出$\hat{\ }_{n,N} -$ 的渐近正态性.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2 假设条件和主要结论

首先假定:

(1) $\mathscr{D}^m$ 为定义在 R^d 上的所有连续函数 f 的集合, 并且 f 的微商满足

$$\frac{i_1}{x_1^{i_1}} \frac{i_2}{x_2^{i_2}} \quad \frac{i_d}{x_p^{i_d}} f(x_1, \quad , x_d)$$

对于 $0 \quad i_1 + i_2 + \quad + i_d \quad m$ 一致有界.

(2) $F(t \mid \boldsymbol{x}) = P(Y \quad t \mid \boldsymbol{X} = \boldsymbol{x}), H(t \mid \boldsymbol{x}) = P(Z \quad t \mid \boldsymbol{X} = \boldsymbol{x})$.

(3) 对于任意分布函数 $W(\ )$ 以及 $r > 0, W(\ ) = 1 - W(\ ), \ W^{-r}(\ ) = (W(\ ))^{-r}$.

(4) $\boldsymbol{X}_{ir}$、 $_r(\boldsymbol{x})$ 分别为 $\boldsymbol{X}_i$、 $(\boldsymbol{x})$ 的第 r 个分量.

(5) 对于任意向量 $\boldsymbol{a}, \boldsymbol{b}, \quad \boldsymbol{a} - \boldsymbol{b} \quad = \quad | a_i - b_i |$, 其中 a_i, b_i 分别为 $\boldsymbol{a}, \boldsymbol{b}$ 的第 i 个分量.

为了证明本文的定理, 并给出如下假设条件:

(C.) $E[\quad | \boldsymbol{X} = \boldsymbol{x}] = 0$

(C.) 的真实值 $_0$ 为有界常向量, 在 $_0$ 的有界区域内.

(C. $\boldsymbol{X}$) $\sup_{x \quad x} E[X_{1r}^4 | \boldsymbol{X} = \boldsymbol{x}] < \quad , \ r = 1, 2, \quad , p$.

(C. $\boldsymbol{X}$) (i) $\boldsymbol{X}$ 的密度函数 $f_X(x)$ 存在, 并且存在 $_n > 0$, 使得 $nP(f_X(x) < \ _n) \quad 0$;

(ii) 存在 $m > 2d$, 使得 $f_X(x) \quad \mathscr{D}^m$.

(C.) ( ) $\mathscr{D}^m$.

(C. Y) $\sup_{x \quad x} E[(\quad _1(Z) + (1 - \quad) \ _2(Z))^4 | \boldsymbol{X} = \boldsymbol{x}] < \quad$.

(C. K) (i) $K(\ )$ 是一个具有有界支撑的有界核函数; (ii) $K(\ )$ 具有有界变差; (iii) $|K(\ )|$ 具有有界变差; (iv) $K(\ )$ 为 m 阶核函数.

(C. h_N) (i) $(N h_N^{2d} \ _n)^{-1/2} \sqrt{\log\log N} \quad 0$ (ii) $N h_N^{4d} \quad$. (iii) 对于 (C. $\boldsymbol{X}$) 中的 m 有 $N h_N^{2m} \quad 0$ (iv) 对于 (C. $\boldsymbol{X}$) (i) 中的 $_n$, 有 $h_N / \ _n \quad 0$

(C. N_n) $\frac{n}{N}$, 其中 为非负常数.

在下面的定理中, 我们给出了估计量 $\hat{\ }_{n,N}$ 的渐近表达式, 进而给出了它的渐近正态性. 对于 $i, j = 1, 2, \quad , n, k = n + 1, n + 2, \quad , n + N$, 设

$$\boldsymbol{h}_0(\boldsymbol{X}_i, Z_i, \quad _i; \boldsymbol{X}_j, Z_j, \quad _j) = \quad (\boldsymbol{X}_i)(Y_i^* - \quad ^{\mathrm{T}}(\boldsymbol{X}_i) \quad) + \quad (\boldsymbol{X}_j)(Y_j^* - \quad ^{\mathrm{T}}(\boldsymbol{X}_j) \quad), \tag{10}$$

$$_0(\boldsymbol{X}_k, \boldsymbol{X}_k) = - \quad (\boldsymbol{X}_k)(\boldsymbol{X}_k - \quad (\boldsymbol{X}_k))^{\mathrm{T}} ,$$

$$= E \quad (\boldsymbol{X}) \quad ^{\mathrm{T}}(\boldsymbol{X}),$$

$$h(\boldsymbol{X}_i, Z_i, \quad _i; \boldsymbol{X}_j, Z_j, \quad _j) = \quad ^{-1}\boldsymbol{h}_0(\boldsymbol{X}_i, Z_i, \quad _i; \boldsymbol{X}_j, Z_j, \quad _j),$$

$$(\boldsymbol{X}_k, \boldsymbol{X}_k) = \quad ^{-1} \quad _0(\boldsymbol{X}_k, \boldsymbol{X}_k). \tag{11}$$

定理 1 在条件 (C.), (C. $\boldsymbol{X}$), (C. K) (i, ii, iv), (C. $\boldsymbol{X}$), (C. Y), (C.), (C. h_N) 和 (C. N_n) 下, 可以将 $\hat{\ }_{n,N} -$ 分解为

$$\hat{\ }_{n,N} - \quad = \boldsymbol{U}_n + \boldsymbol{S}_N + o_P(n^{-\frac{1}{2}}),$$

其中 $\boldsymbol{U}_n = \frac{1}{n^2} \ _{1 \quad i<j \quad n} \boldsymbol{h}(\boldsymbol{X}_i, Z_i, \quad _i; \boldsymbol{X}_j, Z_j, \quad _j), \boldsymbol{S}_N = \frac{1}{N} \ _{k=n+1}^{n+N} (\boldsymbol{X}_k, \boldsymbol{X}_k)$.

定理 1 将 $\hat{\ }_{n,N} -$ 分解成 3 部分, 其中前两部分相互独立, 第三部分为余项. 如果我们能证明前两项的渐近正态性, 则 $\hat{\ }_{n,N} -$ 的渐近正态性就可得证. 根据这种思路, 我们有如下定理.

定理 2 在条件 (C.), (C. $\boldsymbol{X}$), (C. K), (C. $\boldsymbol{X}$), (C. Y), (C.), (C. h_N) 和 (C. N_n) 下,

$$\sqrt{n}(\hat{\ }_{n,N} - \quad) \xrightarrow{d} N(0, \boldsymbol{V}),$$

其中 d 表示依分布收敛,

$$\boldsymbol{V} = \quad ^{-1}(\boldsymbol{V}_1 + \quad \boldsymbol{V}_2)(\quad ^{-1})^{\mathrm{T}},$$

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\boldsymbol{V}_1 = E\{ (\boldsymbol{X}_1)^{\mathrm{T}}(\boldsymbol{X}_1)(Y_1^* - {}^{\mathrm{T}}(\boldsymbol{X}_1))^2\},$$

$$\boldsymbol{V}_2 = E\{ (\boldsymbol{X}_1)^{\mathrm{T}}(\boldsymbol{X}_1)[(\boldsymbol{X}_1^{\mathrm{T}} - {}^{\mathrm{T}}(\boldsymbol{X}_1))]^2\}.$$

在利用定理 2对 进行区间估计和假设检验时,由于 $\boldsymbol{V}$是未知的,因此下面给出 $\boldsymbol{V}$的连续估计量:

$$\hat{\boldsymbol{V}}_n = \hat{}_{n,N}^{-1}(\hat{\boldsymbol{V}}_{1n} + \hat{\boldsymbol{V}}_{2n})(\hat{}_{n,N}^{-1})^{\mathrm{T}},$$

其中 $\hat{}_{n,N}$ 如式(8)所定义,

$$\hat{\boldsymbol{V}}_{1n} = \frac{1}{n}\sum_{i=1}^{n} \hat{}_N(\boldsymbol{X}_i)\hat{}_N^{\mathrm{T}}(\boldsymbol{X}_i)(Y_i^* - \hat{}_N^{\mathrm{T}}(\boldsymbol{X}_i)\hat{}_{n,N})^2\},$$

$$\hat{\boldsymbol{V}}_{2n} = \frac{1}{n}\sum_{i=1}^{n} \hat{}_N(\boldsymbol{X}_i)\hat{}_N^{\mathrm{T}}(\boldsymbol{X}_i)(\boldsymbol{X}_i - \hat{}_N^{\mathrm{T}}(\boldsymbol{X}_i))\hat{}_{n,N})^2\}.$$

3 定理的证明

由 $\hat{}_{n,N}$的定义,可得

$$\hat{}_{n,N} - = \hat{}_{n,N}^{-1}\boldsymbol{A}_{n,N}(), \tag{12}$$

其中 $\boldsymbol{A}_{n,N}() = \frac{1}{n}\sum_{i=1}^{n}\hat{}_N(\boldsymbol{X}_i)({}_i\,{}_1(Z_i) + (1-{}_i)\,{}_2(Z_i) - \hat{}_N^{\mathrm{T}}(\boldsymbol{X}_i))$.

首先,为了证明定理 1,我们需要下面的 3个引理:引理 1、引理 2和引理 3.这 3个引理的详细证明过程类似于文献[14]中的引理 4.1、引理 4.2和引理 4.3,在此我们略去了其证明过程.

引理 1 在条件(C.$\boldsymbol{X}$),(C.K)(i,ii,iv),(C.$\boldsymbol{X}$)(i,ii),(C.h_N)(i)和(C.)下,我们有

$$\hat{}_{n,N} \xrightarrow{P} ,$$

其中 在第二部分式(10)已有定义.

引理 2 在条件(C.$\boldsymbol{X}$),(C.Y),(C.$\boldsymbol{X}$),(C.),(C.K)(i,iv),(C.h_N)(ii,iii)和(C.N_n)下,可以将 $\boldsymbol{A}_{n,N}()$化简为

$$\boldsymbol{A}_{n,N}() = \frac{1}{n^2}\sum_{1\le i<j\le n}\boldsymbol{h}_0(\boldsymbol{X}_i, Z_i, {}_i; \boldsymbol{X}_j, Z_j, {}_j) + \frac{1}{nN}\sum_{i=1}^{n}\sum_{j=n+1}^{n+N}{}_n(\boldsymbol{X}_i, Z_i, {}_i; \boldsymbol{X}_j, \boldsymbol{X}_j) + o_p(n^{-\frac{1}{2}}),$$

其中

$$ {}_n(\boldsymbol{X}_i, Z_i, {}_i; \boldsymbol{X}_j, \boldsymbol{X}_j) = \frac{(\boldsymbol{X}_j - (\boldsymbol{X}_i))(Y_i^* - {}^{\mathrm{T}}(\boldsymbol{X}_i))K\left(\frac{\boldsymbol{X}_i - \boldsymbol{X}_j}{h_N}\right)}{h_N^d f_X(\boldsymbol{X}_i)} - \frac{(\boldsymbol{X}_i)(\boldsymbol{X}_j - (\boldsymbol{X}_i))^{\mathrm{T}} K\left(\frac{\boldsymbol{X}_i - \boldsymbol{X}_j}{h_N}\right)}{h_N^d f_X(\boldsymbol{X}_i)},$$

$\boldsymbol{h}_0()$如式(10)所定义,Y_i^* 如式(3)所定义,$i = 1, 2, \cdots, n$, $j = N+1, N+2, \cdots, n+N$.

引理 3 在条件(C.),(C.K)(i,iv),(C.$\boldsymbol{X}$),(C.$\boldsymbol{X}$),(C.Y)下,我们有

$$\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=n+1}^{n+N}{}_n(\boldsymbol{X}_i, Z_i, {}_i; \boldsymbol{X}_j, \boldsymbol{X}_j) = -\frac{1}{N}\sum_{j=n+1}^{n+N}(\boldsymbol{X}_j)(\boldsymbol{X}_j - (\boldsymbol{X}_j))^{\mathrm{T}} + o_p(n^{-\frac{1}{2}}).$$

定理 1的证明 由式(12)以及引理 1~引理 3,定理 1结论显然.

定理 2的证明 由于 $\boldsymbol{U}_n$ 与 $\boldsymbol{S}_N$ 相互独立,因此只需要证明

$$\sqrt{n}\boldsymbol{U}_n \quad N(0, {}^{-1}\boldsymbol{V}_1(2^{-1})^{\mathrm{T}}), \tag{13}$$

$$\sqrt{n}\boldsymbol{S}_N \quad N(0, K2^{-1}\boldsymbol{V}_2(2^{-1})^{\mathrm{T}}). \tag{14}$$

根据 $\boldsymbol{U}_n$ 的定义,可见 $(n^2/C_n^2)\boldsymbol{U}_n$ 为以 $\boldsymbol{h}(\boldsymbol{X}_i, Z_i, D_i; \boldsymbol{X}_j, Z_j, D_j)$ 为对称核的 U统计量.

对于任何 p维向量 $A \neq 0$,设

$$h_1(\boldsymbol{X}_i, Z_i, D_i; \boldsymbol{X}_j, Z_j, D_j) = A^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{X}_i, Z_i, D_i; \boldsymbol{X}_j, Z_j, D_j),$$

则

$$E(h_1(\boldsymbol{X}_i, Z_i, D_i; \boldsymbol{X}_j, Z_j, D_j)) = 0, \quad i \neq j.$$

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

由计算可得

$$Eh_1^2 = E(A^{\mathrm{T}}{}_2{}^{-1}[L(\boldsymbol{X}_i)(Y_i^* - L^{\mathrm{T}}(\boldsymbol{X}_i)B) + L(\boldsymbol{X}_j)(Y_j^* - L^{\mathrm{T}}(\boldsymbol{X}_j)B)]$$
$$[L^{\mathrm{T}}(\boldsymbol{X}_i)(Y_i^* - L^{\mathrm{T}}(\boldsymbol{X}_i)B) + L^{\mathrm{T}}(\boldsymbol{X}_j)(Y_j^* - L^{\mathrm{T}}(\boldsymbol{X}_j)B)](2^{-1})^{\mathrm{T}}\mathrm{A}) =$$
$$2A^{\mathrm{T}}{}_2{}^{-1}E(L(\boldsymbol{X}_1)L^{\mathrm{T}}(\boldsymbol{X}_1)E[(Y_1^* - L^{\mathrm{T}}(\boldsymbol{X}_1)B)^2 \mid \boldsymbol{X}_1])(2^{-1})^{\mathrm{T}}\mathrm{A},$$

由条件 $(C.\boldsymbol{X})$, $(C.Y)$ 可知 $Eh_1^2 <]$. 设

$$h_2(\boldsymbol{X}_1, Z_1, D_1) := E[h_1(\boldsymbol{X}_1, Z_1, D_1; \boldsymbol{X}_2, Z_2, D_2) \mid \boldsymbol{X}_1, Z_1, D_1] = A^{\mathrm{T}}{}_2{}^{-1}\mathrm{L}(\boldsymbol{X}_1)(Y_1^* - L^{\mathrm{T}}(\boldsymbol{X}_1)B),$$

易证

$$\mathrm{Var}[h_2(\boldsymbol{X}_1, Z_1, D_1)] = A^{\mathrm{T}}{}_2{}^{-1}E(L(\boldsymbol{X}_1)L^{\mathrm{T}}(\boldsymbol{X}_1)E[(Y_1^* - L^{\mathrm{T}}(\boldsymbol{X}_1)B)^2 \mid \boldsymbol{X}_1])(2^{-1})^{\mathrm{T}}\mathrm{A} =$$
$$A^{\mathrm{T}}{}_2{}^{-1}\boldsymbol{V}_1(2^{-1})^{\mathrm{T}}\mathrm{A} > 0,$$

其中 $\boldsymbol{V}_1$ 如定理 2中定义.

根据 U统计量中的 Hoeffding定理 (见文献 [16]), 有

$$\sqrt{n}A^{\mathrm{T}}\left(\frac{n^2}{C_n^2}\right)\boldsymbol{U}_n \xrightarrow{d} N(0, 4A^{\mathrm{T}}{}_2{}^{-1}\boldsymbol{V}_1(2^{-1})^{\mathrm{T}}\mathrm{A}),$$

则

$$\sqrt{n}A^{\mathrm{T}}\boldsymbol{U}_n \xrightarrow{d} N(0, A^{\mathrm{T}}{}_2{}^{-1}\boldsymbol{V}_1(2^{-1})^{\mathrm{T}}\mathrm{A}).$$

由于 $\boldsymbol{S}_N$ 是一系列独立同分布的随机向量之和, 则对于 $\mathrm{A}^{\mathrm{T}}\boldsymbol{S}_N$, 由中心极限定理得:

$$\sqrt{n}A^{\mathrm{T}}\boldsymbol{S}_N \xrightarrow{d} N(0, A^{\mathrm{T}}\mathrm{K}{}_2{}^{-1}\boldsymbol{V}_2(2^{-1})^{\mathrm{T}}\mathrm{A}),$$

其中 $\boldsymbol{V}_2$ 如定理 2中定义.

式 (13)、(14) 得证.

由于 A的任意性, 从而定理 2成立.

[参考文献]

[1] Buckley J James. Linear regression with censored data[J]. Biometrika, 1979, 66(3): 429-436

[2] Chatterjee S, McLeish D L. Fitting linear regression models to censored data by least squares and maximum likelihood methods[J]. Comm Statist-Theor Meth, 1986, 15(11): 3 227-3 243.

[3] Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right-censored data[J]. Ann Stat, 1981, 9(6): 1 276-1 288.

[4] Leurgans S. Linear models, random censoring and synthetic data[J]. Biometrika, 1987, 74(2): 301-309.

[5] Miller R G. Least squares regression with censored data[J]. Biometrika, 1976, 63(6): 449-464.

[6] Zheng Z K. A class of estimator for the parameter in linear regression with censored data[J]. Acta Mathematic Applicatae Sinica, 1987, 3: 231-241.

[7] 秦更生. 随机删失场合部分线性模型中的核光滑方法 [J]. 数学年刊, 1995, A16(4): 441-453.

[8] Prentice R L. Surrogate endpoints in clinical trials: definition and operational[J]. Criteria Statist, 1989, 8: 431-440.

[9] Bound J, Krunt A B. The extend of measurement error in longitudinal earning data: Do two wrongs make a right? [J]. J Labor Economics, 1989, 9: 1-24.

[10] Pepe M S, Fleming T R. A general nonparametric method for dealing with errors in missing or surrogate covariate data[J]. J Amer Statist Assoc, 1991, 86: 108-113.

[11] Pepe M S. Inference using surrogate outcome data and a validation sample[J]. Biometrika, 1992, 79(2): 355-365.

[12] Sepanski J H, Lee L F. Semiparametric estimation of nonlinear error-in-variable models with validation study[J]. J Nonparametr Statist 4, 1995, 4(4): 365-394.

[13] Carroll R J, Wand M P. Semiparametric estimation in logistic measure error models[J]. J Roy Statist Soc B, 1991, 53(3): 573-585.

[14] Wang Qihua. Estimation of linear error-in-covariables models with validation data under random censorship[J]. J Multivariate Anal, 2000, 74(2): 245-266.

[15] Tze Leung lai, Zheng Zukang. Survival Analysis[M]. Hangzhou: Zhejiang Publishing House of Science and Technology, 1993: 105-116.

[16] 陈希儒. 非参数统计 [M]. 上海: 上海科学技术出版社, 2000: 58-67.

[责任编辑: 丁　蓉]

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net