应用概率统计　第二十六卷
第六期　2010年12月

Chinese Journal of Applied Probability
and Statistics Vol.26 No.6 Dec. 2010

Uniform Consistency of the Nonparametric Error Density Estimation in Regression with Censored Data *

Zhou Xiuqing
(*School of Mathematical Sciences, Nanjing Normal University, Nanjing, 210046*)
Shi Ningzhong
(*KLAS and School of Mathematics and Statistics, Northeast Normal University, Changchun, 130024*)
Zhao Jin
(*Department of Mathematics, Nanjing University, Nanjing, 210093*)

Abstract

Nonparametric estimation of the density function of unobservable regression errors is a fundamental issue in regression analysis, because there are many practical applications of error density estimation. This problem for regression models with complete observed data has been studied by several authors. But in many application fields, the corresponding variables are not complete observable because of censoring. In this case, the density function of the unobservable regression errors can be estimated by the kernel type estimator based on the censored regression residuals. In this paper, the asymptotic property of the kernel type estimator is considered and the uniform consistency of the estimator is established.

Keywords: Nonparametric estimation, censored data, regression residuals, uniform consistency.

AMS Subject Classification: 62N02.

§1. Introduction

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ are n independent and identically distributed (i.i.d.) realizations of the random vector (X, Y), with X taking values in $[0, 1]$. Suppose X and Y are related via the regression function $m(x) := \mathsf{E}(Y|X = x)$ which satisfy

$$Y_i = m(X_i) + e_i, \qquad i = 1, \cdots, n, \tag{1.1}$$

where $e_1, \cdots, e_n$ are unobservable random errors which are independent and identically distributed with zero mean.

*The project supported by NSFC foundation (10626028).
Received August 15, 2008. Revised April 4, 2010.

Estimation of the density function of the unobservable regression errors e_i is a fundamental issue in regression problem (1.1), because there are many practical applications of error density estimation, such as hypothesis testing, prediction, and so on. Estimation of the error density function in regression models with complete observed data has been studied by Cheng[1, 2], and Efromovich[3] among others.

But in many fields, such as life time data analysis, econometrics, environmental science and other applications, the response variables Y_i may not be completely observable because of censoring. In such case, one observes not $\{(X_i, Y_i) : i = 1, \cdots, n\}$, but $\{(X_i, \min(Y_i, C_i), \delta_i) : i = 1, \cdots, n\}$ only, where $C_1, \cdots, C_n$ are i.i.d. censoring variables independent of $X_1, \cdots, X_n$, $\delta_i = I(Y_i \le C_i)$ and $I(A)$ is the indicate function of set A.

The regression problem (1.1) with censored data has been studied by Buckley and James[4], Tsiatis[5], Lai and Ying[6], Fan and Gijbels[7], Stute[8, 9] and others. Let $m_n(x)$ be a regression estimator, then $\widehat{e}_i = Y_i - m_n(X_i)$ are called regression residuals. Note that because of the censoring, to estimate the density function of e_i, only $\{(\widehat{z}_i, \delta_i) : i = 1, \cdots, n\}$ can be used, where $\widehat{z}_i = \min(Y_i, C_i) - m_n(X_i)$ are the censored regression residuals, and δ_i are the same as defined above.

Suppose $e_1, \cdots, e_n$ have the same distribution function F and density function f. The K-M type estimator of F based on the censored residuals is

$$\widehat{F}_n(t) = 1 - \begin{cases} \prod_{i=1}^{n} \Big(\frac{N_n(\widehat{z}_i)}{N_n(\widehat{z}_i)+1}\Big)^{I(\widehat{z}_i \le t, \delta_i = 1)}, & \text{if } t \le \max(\widehat{z}_1, \cdots, \widehat{z}_n); \\ 0, & \text{elsewhere,} \end{cases}$$

where $N_n(t) = \sum_{i=1}^{n} I(\widehat{z}_i > t)$. The kernel type estimator of f based on $\widehat{F}_n$ is

$$\widehat{f}_n(t) = \frac{1}{h(n)} \int_R K\Big(\frac{t-s}{h(n)}\Big) \mathrm{d}\widehat{F}_n(s), \tag{1.2}$$

where $h(n)$ is a sequence of positive numbers such that $h(n) \to 0$ as $n \to \infty$, and K is a density function.

$\widehat{f}_n(t)$ is proposed by Van Keilegom and Veraverbeke[10]. The main purpose of this paper is to investigate the uniform consistency of the estimator (1.2). In next section, we will state the assumptions and some notations needed in this paper. In Section 3, the uniform consistency of the estimator is established, and the proofs are given in detail. The simulation results are shown in Section 4.

§2. Assumptions and Some Notations

Denote

$$d_n(x) = m_n(x) - m(x), \qquad x \in [0, 1].$$

To obtain the uniform consistency of estimator (1.2), a basic assumption about $d_n(x)$ is given first:

(C1) There is a sequence of positive numbers β_n which satisfy $\beta_n \to 0$ and $n\beta_n^2 \to \infty$ as $n \to \infty$ such that

$$\sup_{x\in[0,1]} |d_n(x)| = O_p(\beta_n) \qquad \text{as } n \to \infty. \tag{2.1}$$

(2.1) indicate that for any $\epsilon > 0$, there is a constant $M_\epsilon < \infty$, such that

$$\mathsf{P}(A_{n,\epsilon}) > 1 - \epsilon \tag{2.2}$$

holds for all n large enough, where

$$A_{n,\epsilon} = \Big\{\beta_n^{-1} \sup_{x\in[0,1]} |d_n(x)| \le M_\epsilon\Big\}.$$

There are many examples of the case that assumption (C1) is satisfied.

Censored linear regression model, i.e. the censored regression model with $m(x) = \theta_0^T x$ has been studied by many authors. For example, Tsiatis[5], Lai and Ying[6] and Stute[8] proposed different estimators $\widehat{\theta}_n$ of θ_0 and, respectively, proved that under some conditions, $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ converge to normal distribution with zero mean and finite variance. Let $m_n(x) = \widehat{\theta}_n^T x$. In such cases, we have

$$\sup_{x\in[0,1]} |m_n(x) - m(x)| = \sup_{x\in[0,1]} |\widehat{\theta}_n^T x - \theta_0^T x| \le |\widehat{\theta}_n - \theta_0| = O_p\Big(\frac{1}{\sqrt{n}}\Big).$$

Therefore, (2.1) holds for any sequence of β_n satisfying $n\beta_n^2 \to +\infty$.

Stute[9] proposed an estimator $\widehat{\theta}_n$ for the parameter in nonlinear censored regression model, i.e. the censored regression model with $m(x) = f(x, \theta_0)$. From Theorem 1.2 in [9] we know that under some conditions

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = O_p(1).$$

Suppose that there is a positive constant $C > 0$ such that $|f(x,\theta_1) - f(x,\theta_2)| \le C|\theta_1 - \theta_2|$ holds for all $x \in [0,1]$ and any θ_1, θ_2. Let $m_n(x) = f(x, \widehat{\theta}_n)$. Since

$$\sup_{x\in[0,1]} |m_n(x) - m(x)| = \sup_{x\in[0,1]} |f(x,\widehat{\theta}_n) - f(x,\theta_0)| \le C|\widehat{\theta}_n - \theta_0| = O_p\Big(\frac{1}{\sqrt{n}}\Big),$$

(2.1) holds for any sequence of β_n satisfying $n\beta_n^2 \to +\infty$.

Ould-Saïd and Cai[11] proposed an estimator $\xi_n(y|x)$ for $\xi(y|x)$, which is the conditional probability density function of Y given $X = x$ under censoring, and proved that when the support D_Y of Y is bounded,

$$\sup_{x\in[0,1]} \sup_{y\in D_Y} |\xi_n(y|x) - \xi(y|x)| = O\Big(\max\Big(\frac{\log n}{na_n}, a_n^2\Big)\Big)$$

holds with probability 1 as $n \to \infty$ under some conditions, where a_n is a sequence of positive numbers satisfy $a_n \to 0$ and $na_n^4/(\log\log n) \to \infty$. Let $m_n(x) = \int_{D_Y} y\xi_n(y|x)\mathrm{d}y$. Then we can get that

$$\begin{aligned}\sup_{x\in[0,1]}|m_n(x)-m(x)| &= \sup_{x\in[0,1]}\Big|\int_{D_Y} y[\xi_n(y|x)-\xi(y|x)]\mathrm{d}y\Big| \\ &\le \sup_{x\in[0,1]}\int_{D_Y}|y||\xi_n(y|x)-\xi(y|x)|\mathrm{d}y \\ &\le \sup_{x\in[0,1]}\sup_{y\in D_Y}|\xi_n(y|x)-\xi(y|x)|\int_{D_Y}|y|\mathrm{d}y \\ &= O\Big(\max\Big(\frac{\log n}{na_n}, a_n^2\Big)\Big)\end{aligned}$$

holds with probability 1. Let $a_n = \sqrt{\log n}/\sqrt[4]{n}$ and $\beta_n = \log n/\sqrt{n}$, then condition (C1) holds obviously.

Other assumptions needed in this paper are imposed as following:

(A1) $e_1, \cdots, e_n$ are i.i.d. random variables with same distribution function F and bounded and continuous density function f. $\{C_i - m(X_i) : i = 1, \cdots, n\}$ are i.i.d. random variables which are also independent of $\{e_i : i = 1, \cdots, n\}$. The distribution function G of $C_i - m(X_i)$ has a bounded and continuous density function g. Moreover, $\tau_F \le \tau_G \le +\infty$, where $\tau_F = \inf\{t : F(t) = 1\}$, $\tau_G = \inf\{t : G(t) = 1\}$.

(A2) $h(n) \to 0$, $nh(n)^2 \to \infty$, $nh(n)/\log n \to \infty$, $\beta_n/h(n) \to 0$.

(A3) K is a continuous and bounded density function with support $[-1, 1]$, and there is a positive constant $L > 0$ such that

$$|K(x_1) - K(x_2)| \le L|x_1 - x_2|$$

holds for any $x_1 \in [-1, 1]$ and $x_2 \in [-1, 1]$.

Note that the regression error e_i is censored at $C_i - m(X_i)$. Define $z_i = \min(Y_i - m(X_i), C_i - m(X_i))$ and denote its distribution function by H. To simplify the proofs in Section 3, some notations are defined beforehand:

$$\widehat{F}_n^*(t) = 1 - \begin{cases}\prod_{i=1}^n \Big(\frac{N_n(\widehat{z}_i)+1}{N_n(\widehat{z}_i)+2}\Big)^{I(\widehat{z}_i\le t,\delta_i=1)}, & \text{if } t \le \max(\widehat{z}_1, \cdots, \widehat{z}_n); \\ 0, & \text{elsewhere.}\end{cases}$$

$$H_n^u(z_i;t) = \frac{1}{n}\sum_{i=1}^n I(z_i \le t, \delta_i = 1), \qquad H_n^u(\widehat{z}_i;t) = \frac{1}{n}\sum_{i=1}^n I(\widehat{z}_i \le t, \delta_i = 1),$$

$$H_n(z_i;t) = \frac{1}{n}\sum_{i=1}^n I(z_i \le t), \qquad H_n(\widehat{z}_i;t) = \frac{1}{n}\sum_{i=1}^n I(\widehat{z}_i \le t),$$

$$H^u(t) = \mathsf{P}(e_i \le t, \delta_i = 1) = \int_{-\infty}^{t} (1 - G(s))\mathrm{d}F(s),$$
$$I_n^*(t) = \int_{-\infty}^{t} \frac{1}{1 - H_n(\widehat{z}_i; s)} \mathrm{d}H_n^u(\widehat{z}_i; s),$$
$$I^*(t) = \int_{-\infty}^{t} \frac{1}{1 - H(s)} \mathrm{d}H^u(s) = -\log(1 - F(t)).$$

For any distribution function S let $\overline{S} = 1 - S$ denote the corresponding survival function.

§3. Uniform Consistency of the Estimator

In order to establish the uniform consistency of $\widehat{f}_n$, firstly, we give the following two lemmas.

Lemma 3.1 Under assumptions (C1) and (A1) we have that for any closed interval $D = [T_1, T_2] \subset (-\infty, \tau_F)$,

$$\beta_n^{-1} \sup_{t \in D} |H_n(\widehat{z}_i; t) - H(t)| = O_p(1), \tag{3.1}$$
$$\beta_n^{-1} \sup_{t \in D} |H_n^u(\widehat{z}_i; t) - H^u(t)| = O_p(1) \tag{3.2}$$

hold true as $n \to +\infty$.

Proof To prove (3.1), note that

$$\sup_{t \in D} |H_n(\widehat{z}_i; t) - H(t)| \le \sup_{t \in D} |H_n(\widehat{z}_i; t) - H_n(z_i; t)| + \sup_{t \in D} |H_n(z_i; t) - H(t)|.$$

The second term on the right-hand side of the above inequality is $O(\sqrt{\log\log n/n})$ with probability 1 in view of the LIL for the Kolmogorov-Smirnov distance (see, e.g., [12]). So, by (2.2) it is suffices to show that for any $\epsilon > 0$,

$$\sup_{t \in D} |H_n(\widehat{z}_i; t) - H_n(z_i; t)| = O_p(\beta_n) \tag{3.3}$$

holds on $A_{n,\epsilon}$. Check that on $A_{n,\epsilon}$,

$$\begin{aligned} \sup_{t \in D} |H_n(\widehat{z}_i; t) - H_n(z_i; t)| &= \sup_{t \in D} \Big| \frac{1}{n} \sum_{i=1}^{n} [I(z_i \le t - d_n(X_i)) - I(z_i \le t)] \Big| \\ &\le \sup_{t \in D} \frac{1}{n} \sum_{i=1}^{n} I(t - M_\epsilon \beta_n < z_i \le t + M_\epsilon \beta_n) \end{aligned}$$

holds for all n large enough. For any $\epsilon' > 0$, by Corollary 1 of Massart[13] and the fact

that $n\beta_n \to \infty$ we have

$$\begin{aligned}
& \mathsf{P}\Big(\sup_{t\in D}\frac{1}{M_\epsilon\beta_n}\Big|\sum_{i=1}^{n}\frac{1}{n}I(t-M_\epsilon\beta_n<z_i\le t+M_\epsilon\beta_n) \\
& -H(t+M_\epsilon\beta_n)+H(t-M_\epsilon\beta_n)\Big|>\epsilon'\Big) \\
\le\ & \mathsf{P}\Big(\sup_{t\in D}\frac{1}{M_\epsilon\beta_n}\Big|\sum_{i=1}^{n}\frac{1}{n}I(z_i\le t+M_\epsilon\beta_n)-H(t+M_\epsilon\beta_n)\Big|>\frac{\epsilon'}{2}\Big) \\
& +\mathsf{P}\Big(\sup_{t\in D}\frac{1}{M_\epsilon\beta_n}\Big|\sum_{i=1}^{n}\frac{1}{n}I(z_i\le t-M_\epsilon\beta_n)-H(t-M_\epsilon\beta_n)\Big|>\frac{\epsilon'}{2}\Big) \\
\le\ & 4\exp\Big(-\frac{n(\epsilon' M_\epsilon\beta_n)^2}{2}\Big) \\
\to\ & 0.
\end{aligned}$$

Noting that

$$\sup_{t\in D}\Big|\frac{1}{M_\epsilon\beta_n}[H(t+M_\epsilon\beta_n)-H(t-M_\epsilon\beta_n)]-2[f(t)\overline{G}(t)+g(t)\overline{F}(t)]\Big|\overset{\text{a.s.}}{\to}0$$

holds in view of the uniform continuity of $f(t)\overline{G}(t)+g(t)\overline{F}(t)$ on D, we have

$$\sup_{t\in D}\Big|\frac{1}{nM_\epsilon\beta_n}\sum_{i=1}^{n}I(t-M_\epsilon\beta_n<z_i\le t+M_\epsilon\beta_n)-2[f(t)\overline{G}(t)+g(t)\overline{F}(t)]\Big|=o_p(1).$$

So (3.3) holds directly. Whence equation (3.1) comes.

By the similar way used above and the fact that

$$\sup_{t\in D}\Big|\frac{1}{M_\epsilon\beta_n}[H^u(t+M_\epsilon\beta_n)-H^u(t-M_\epsilon\beta_n)]-2f(t)\overline{G}(t)\Big|\overset{\text{a.s.}}{\to}0,$$

we can prove (3.2) holds. □

Though the uniform consistency of $\widehat{F}_n(t)$ under some conditions has been embedded in [10], for the purpose of this paper, we will prove this property of $\widehat{F}_n(t)$ in Lemma 3.2 under different conditions. The assumptions imposed in Lemma 3.2 is looser in some sense than that imposed in [10]. For example, $H(t)$ is not required to be twice continuous differentiable here.

Lemma 3.2 Under assumptions (C1) and (A1) we have that for any closed interval $D=[T_1,T_2]\subset(-\infty,\tau_F)$,

$$\beta_n^{-1}\sup_{t\in D}|\widehat{F}_n(t)-F(t)|=O_p(1)$$

holds true as $n\to+\infty$.

Proof Since $H(T_2) < 1$ and $H^u(T_2) \le F(T_2) < 1$, by Lemma 3.1 we know that for any $\epsilon > 0$,

$$\mathsf{P}\Big(H_n(\widehat{z}_i; T_2) \le \frac{1 + H(T_2)}{2} < 1\Big) > 1 - \epsilon, \tag{3.4}$$

$$\mathsf{P}\Big(H_n^u(\widehat{z}_i; T_2) \le \frac{1 + H^u(T_2)}{2} < 1\Big) > 1 - \epsilon \tag{3.5}$$

hold when n large enough.

Note that

$$\sup_{t\in D} |\widehat{F}_n(t) - F(t)| \le \sup_{t\in D} |\widehat{F}_n(t) - \widehat{F}_n^*(t)| + \sup_{t\in D} |\widehat{F}_n^*(t) - F(t)|. \tag{3.6}$$

For the first term on the right-hand side of (3.6), by the fact that the inequality $\big|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i\big| \le \sum_{i=1}^n |a_i - b_i|$ holds for any $0 \le a_i \le 1$, $0 \le b_i \le 1$, we have that

$$\begin{aligned}
\sup_{t\in D} |\widehat{F}_n(t) - \widehat{F}_n^*(t)| &\le \sup_{t\in D} \sum_{i=1}^n \Big| \Big(\frac{N_n(\widehat{z}_i)}{N_n(\widehat{z}_i + 1)}\Big)^{I(\widehat{z}_i \le t, \delta_i = 1)} - \Big(\frac{N_n(\widehat{z}_i) + 1}{N_n(\widehat{z}_i + 2)}\Big)^{I(\widehat{z}_i \le t, \delta_i = 1)} \Big| \\
&\le \sup_{t\in D} \sum_{i=1}^n \frac{I(\widehat{z}_i \le t, \delta_i = 1)}{N_n(\widehat{z}_i + 1)^2} \\
&\le \frac{1}{n} \sup_{t\in D} \int_{-\infty}^t \frac{1}{[1 - H_n(\widehat{z}_i; s)]^2} \mathrm{d}H_n^u(\widehat{z}_i; s) \\
&\le \frac{1}{n} \frac{1}{[1 - H_n(\widehat{z}_i; T_2)]^2} H_n^u(\widehat{z}_i; T_2).
\end{aligned}$$

Then by (3.4) and (3.5) we get

$$\sup_{t\in D} |\widehat{F}_n(t) - \widehat{F}_n^*(t)| \le O_p\Big(\frac{1}{n}\Big).$$

For the second term on the right-hand side of (3.6), noting that $|e^a - e^{-b}| \le |a + b|$ holds for any $a \le 0$, $b \ge 0$, we can obtain

$$\begin{aligned}
\sup_{t\in D} |\widehat{F}_n^*(t) - F(t)| &\le \sup_{t\in D} \big[\big|e^{\log \overline{\widehat{F}}_n^*(t)} - e^{-I_n^*(t)}\big| + |e^{-I_n^*(t)} - e^{-I^*(t)}|\big] \\
&\le \sup_{t\in D} \big[\big|\log \overline{\widehat{F}}_n^*(t) + I_n^*(t)\big| + |I_n^*(t) - I^*(t)|\big].
\end{aligned}$$

Rewrite $\big|\log \overline{\widehat{F}}_n^*(t) + I_n^*(t)\big|$ as

$$\Big| \int_{-\infty}^t \Big\{ n \log\Big[1 - \frac{1}{2 + n(1 - H_n(\widehat{z}_i; s))}\Big] + \frac{1}{1 - H_n(\widehat{z}_i; s)} \Big\} \mathrm{d}H_n^u(\widehat{z}_i; s) \Big|.$$

Note the fact that

$$\log\Big(1 - \frac{1}{x}\Big) = -\sum_{i=1}^{+\infty} \frac{1}{ix^i} \qquad \text{and} \qquad \log\Big(1 - \frac{1}{x}\Big) + \frac{1}{x} + \frac{1}{x^2} \ge 0$$

hold for any $x \geq 2$. Then using (3.4) and (3.5) again we can show

$$\begin{aligned}
&\sup_{t\in D}\big|\log\overline{\widehat{F}}^{*}_n(t)+I^*_n(t)\big|\\
=\;&\sup_{t\in D}\Big|\int_{-\infty}^{t}\Big\{n\Big[-\sum_{l=1}^{+\infty}\frac{1}{l(2+n(1-H_n(\widehat{z}_i;s)))^l}\Big]+\frac{1}{1-H_n(\widehat{z}_i;s)}\Big\}\mathrm{d}H^u_n(\widehat{z}_i;s)\Big|\\
\leq\;&\sup_{t\in D}\Big\{\int_{-\infty}^{t}\Big|\frac{1}{1-H_n(\widehat{z}_i;s)}-\frac{n}{2+n(1-H_n(\widehat{z}_i;s))}\Big|\mathrm{d}H^u_n(\widehat{z}_i;s)\\
&+\int_{-\infty}^{t}\Big[n\sum_{l=2}^{+\infty}\frac{1}{l(2+n(1-H_n(\widehat{z}_i;s)))^l}\Big]\mathrm{d}H^u_n(\widehat{z}_i;s)\Big\}\\
\leq\;&\sup_{t\in D}\Big\{\frac{2}{n}\int_{-\infty}^{t}\frac{1}{[1-H_n(\widehat{z}_i;s)]^2}\mathrm{d}H^u_n(\widehat{z}_i;s)+\int_{-\infty}^{t}\frac{n}{[2+n(1-H_n(\widehat{z}_i;s))]^2}\mathrm{d}H^u_n(\widehat{z}_i;s)\Big\}\\
\leq\;&\sup_{t\in D}\frac{3}{n}\int_{-\infty}^{t}\frac{1}{[1-H_n(\widehat{z}_i;s)]^2}\mathrm{d}H^u_n(\widehat{z}_i;s)\\
\leq\;&\frac{3}{n[1-H_n(\widehat{z}_i;T_2)]^2}H^u_n(\widehat{z}_i;T_2)\\
=\;&O_p\Big(\frac{1}{n}\Big).
\end{aligned}$$

Since

$$\begin{aligned}
&\sup_{t\in D}|I^*_n(t)-I^*(t)|\\
\leq\;&\sup_{t\in D}\Big[\Big|\int_{-\infty}^{t}\frac{\mathrm{d}H^u_n(\widehat{z}_i;s)}{1-H_n(\widehat{z}_i;s)}-\int_{-\infty}^{t}\frac{\mathrm{d}H^u_n(\widehat{z}_i;s)}{1-H(s)}\Big|+\Big|\int_{-\infty}^{t}\frac{\mathrm{d}H^u_n(\widehat{z}_i;s)}{1-H(s)}-\int_{-\infty}^{t}\frac{\mathrm{d}H^u(s)}{1-H(s)}\Big|\Big]\\
\leq\;&\sup_{t\in D}\Big[\int_{-\infty}^{t}\frac{|H_n(\widehat{z}_i;s)-H(s)|}{(1-H(s))(1-H_n(\widehat{z}_i;s))}\mathrm{d}H^u_n(\widehat{z}_i;s)+\Big|\int_{-\infty}^{t}\frac{\mathrm{d}(H^u_n(\widehat{z}_i;s)-H^u(s))}{1-H(s)}\Big|\Big],
\end{aligned}$$

by (3.4), (3.5) and Lemma 3.1 we have that

$$\begin{aligned}
\sup_{t\in D}|I^*_n(t)-I^*(t)| &\leq O_p(1)\sup_{t\in D}[|H_n(\widehat{z}_i;t)-H(t)|+|H^u_n(\widehat{z}_i;t)-H^u(t)|]\\
&= O_p(\beta_n).
\end{aligned}$$

Therefore we have completed the proof of Lemma 3.2. □

Now, we are ready to state and prove the uniform consistency of $\widehat{f}_n$ for f.

Theorem 3.1 Under assumptions (C1) and (A1)-(A3), we have that for any closed interval $D=[T_1,T_2]\subset(-\infty,\tau_F)$,

$$\sup_{t\in D}|\widehat{f}_n(t)-f(t)|=o_p(1)$$

holds true as $n\to\infty$.

Proof From the argument of Mielniczuk[14] and Theorem A of Silverman[15] we know that under assumptions (A1)-(A3)

$$\sup_{t\in D}\Big|\frac{1}{\overline{G}(t)h(n)}\int_R K\Big(\frac{t-s}{h(n)}\Big)\mathrm{d}H_n^u(z_i;s)-f(t)\Big|=o(1)$$

holds with probability 1 as $n\to\infty$. So, it is enough to prove that

$$\sup_{t\in D}\Big|\widehat{f}_n(t)-\frac{1}{\overline{G}(t)h(n)}\int_R K\Big(\frac{t-s}{h(n)}\Big)\mathrm{d}H_n^u(z_i;s)\Big|=o_p(1).$$

Let $a_n(\widehat{z}_i)$ denotes the jump size of $\widehat{F}_n$ at point $\widehat{z}_i$ and $S(t,r)=\{y:|y-t|\le r\}$. Then

$$\begin{aligned}
&\Big|\widehat{f}_n(t)-\frac{1}{\overline{G}(t)h(n)}\int_R K\Big(\frac{t-s}{h(n)}\Big)\mathrm{d}H_n^u(z_i;s)\Big|\\
=\;&\frac{1}{h(n)}\Big|\int_R K\Big(\frac{t-s}{h(n)}\Big)\mathrm{d}\widehat{F}_n(s)-\frac{1}{\overline{G}(t)}\int_R K\Big(\frac{t-s}{h(n)}\Big)\mathrm{d}H_n^u(z_i;s)\Big|\\
=\;&\frac{1}{h(n)}\Big|\sum_{i=1}^n\delta_i a_n(\widehat{z}_i)K\Big(\frac{t-\widehat{z}_i}{h(n)}\Big)-\frac{1}{\overline{G}(t)}\sum_{i=1}^n\delta_i\frac{1}{n}K\Big(\frac{t-z_i}{h(n)}\Big)\Big|\\
\le\;&\frac{1}{h(n)}\Big|\sum_{i=1}^n\delta_i a_n(\widehat{z}_i)\Big[K\Big(\frac{t-\widehat{z}_i}{h(n)}\Big)-K\Big(\frac{t-z_i}{h(n)}\Big)\Big]\Big|+\frac{1}{h(n)}\sum_{i=1}^n\delta_i K\Big(\frac{t-z_i}{h(n)}\Big)\Big|a_n(\widehat{z}_i)-\frac{1}{n\overline{G}(t)}\Big|\\
\le\;&I_{n1}+I_{n2}+I_{n3}+I_{n4},
\end{aligned}$$

where

$$\begin{aligned}
I_{n1}&=\frac{1}{h(n)}\Big|\sum_{i=1}^n\delta_i a_n(\widehat{z}_i)\Big[K\Big(\frac{t-\widehat{z}_i}{h(n)}\Big)-K\Big(\frac{t-z_i}{h(n)}\Big)\Big]I(\widehat{z}_i\in S(t,h(n)),z_i\in S(t,h(n)))\Big|,\\
I_{n2}&=\frac{1}{h(n)}\Big|\sum_{i=1}^n\delta_i a_n(\widehat{z}_i)K\Big(\frac{t-\widehat{z}_i}{h(n)}\Big)I(\widehat{z}_i\in S(t,h(n)),z_i\in S^c(t,h(n)))\Big|,\\
I_{n3}&=\frac{1}{h(n)}\Big|\sum_{i=1}^n\delta_i a_n(\widehat{z}_i)K\Big(\frac{t-z_i}{h(n)}\Big)I(\widehat{z}_i\in S^c(t,h(n)),z_i\in S(t,h(n)))\Big|,\\
I_{n4}&=\frac{1}{h(n)}\sum_{i=1}^n\delta_i K\Big(\frac{t-z_i}{h(n)}\Big)\Big|a_n(\widehat{z}_i)-\frac{1}{n\overline{G}(t)}\Big|,
\end{aligned}$$

and S^c denotes the complement of set S.

For I_{n1}, by assumptions (C1), (A2) and (A3) we have that

$$\begin{aligned}
&\sup_{t\in D}\max_i\Big\{\Big|K\Big(\frac{t-\widehat{z}_i}{h(n)}\Big)-K\Big(\frac{t-z_i}{h(n)}\Big)\Big|I(\widehat{z}_i\in S(t,h(n)),z_i\in S(t,h(n)))\Big\}\\
\le\;&\sup_{t\in D}\max_i\Big\{L\Big|\frac{d_n(X_i)}{h(n)}\Big|\Big\}\le L\cdot\sup_{x\in[0,1]}\frac{|d_n(x)|}{h(n)}\\
=\;&o_p(1).
\end{aligned}\tag{3.7}$$

On the other hand, by the similar way as that used in the proof of Lemma 3.1, we can show

$$\sup_{t\in D}\Big|\frac{1}{2nh(n)}\sum_{i=1}^n\delta_i I(z_i\in S(t,h(n)))-f(t)\overline{G}(t)\Big|\xrightarrow{\text{a.s.}}0.\tag{3.8}$$

Note that

$$\sup_{t\in D}\max_i\{na_n(\widehat{z}_i)\delta_i I(\widehat{z}_i\in S(t,h(n)))\}\le \sup_{t\in D}\sup_{s\in S(t,2h(n))}\frac{\overline{\widehat{F}}_n(s)}{\overline{H}_n(\widehat{z}_i;s)}=O_p(1) \tag{3.9}$$

holds in view of Lemma 3.1, assumption (A2) and the fact (see [16])

$$na_n(\widehat{z}_i)\delta_i=\frac{\overline{\widehat{F}}_n(\widehat{z}_i-0)}{\overline{H}_n(\widehat{z}_i;\widehat{z}_i-0)}\delta_i. \tag{3.10}$$

Then by (3.7), (3.8) and (3.9) together we have

$$\sup_{x\in D}I_{n1}=o_p(1)$$

holds true as $n\to\infty$.

For I_{n2} and any $\epsilon>0$, using (3.9) and the similar way used in the proof of Lemma 3.1 again we have that on $A_{n,\epsilon}$,

$$\begin{aligned}\sup_{t\in D}I_{n2} &\le \frac{L}{h(n)}\cdot\sup_{t\in D}\Big|\sum_{i=1}^n\delta_i a_n(\widehat{z}_i)I(\widehat{z}_i\in S(t,h(n)),z_i\in S^c(t,h(n)))\Big|\\ &\le O_p\Big(\frac{1}{h(n)}\Big)\cdot\sup_{t\in D}\Big\{\frac{1}{n}\sum_{i=1}^n[I(t+h(n)\le z_i\le t+h(n)+M_\epsilon\beta_n)\\ &\quad +I(t-h(n)-M_\epsilon\beta_n\le z_i\le t-h(n))]\Big\}\\ &= o_p(1).\end{aligned}$$

Combining this with (2.2) we get

$$\sup_{t\in D}I_{n2}=o_p(1)$$

holds true as $n\to\infty$.

By the way similar as that used to I_{n2} we can show that

$$\sup_{t\in D}I_{n3}=o_p(1)$$

holds as $n\to\infty$.

For I_{n4}, by (3.8) and assumption (A3) together we have that

$$\begin{aligned}\sup_{t\in D}I_{n4} &= \frac{1}{h(n)}\cdot\sup_{t\in D}\sum_{i=1}^n\delta_i\Big|K\Big(\frac{t-z_i}{h(n)}\Big)\Big|\Big|a_n(\widehat{z}_i)-\frac{1}{n\overline{G}(t)}\Big|\\ &\le 2L\cdot\sup_{t\in D}\left[\frac{\sum_{i=1}^n\delta_i I(z_i\in S(t,h(n)))}{nh(n)}\right]\cdot\sup_{t\in D}\max_{i:z_i\in S(t,h(n))}\Big[\delta_i\Big|na_n(\widehat{z}_i)-\frac{1}{\overline{G}(t)}\Big|\Big]\\ &= O_p(1)\sup_{t\in D}\max_{i:z_i\in S(t,h(n))}\Big[\delta_i\Big|na_n(\widehat{z}_i)-\frac{1}{\overline{G}(t)}\Big|\Big].\end{aligned}$$

Using (3.10) again, we can obtain that

$$\sup_{t\in D} I_{n4} \leq O_p(1)\cdot \sup_{t\in D}\Big\{ \sup_{s\in S(t,2h(n))}\Big|\frac{\overline{\widehat{F}}_n(s)}{\overline{H}_n(\widehat{z}_i;s)}-\frac{\overline{\widehat{F}}_n(s)}{\overline{H}(s)}\Big| + \sup_{s\in S(t,2h(n))}\Big|\frac{\overline{\widehat{F}}_n(s)}{\overline{H}(s)}-\frac{\overline{F}(s)}{\overline{H}(s)}\Big| + \sup_{s\in S(t,2h(n))}\Big|\frac{\overline{F}(s)}{\overline{H}(s)}-\frac{1}{\overline{G}(t)}\Big|\Big\}.$$

Noting the fact $h(n)\to 0$, it follows that when n lager enough

$$\sup_{t\in D} I_{n4} \leq O_p(1)\cdot \Big\{ \sup_{t\in[T_1-1,(T_2+\tau_F)/2]}\Big|\frac{\overline{\widehat{F}}_n(t)}{\overline{H}_n(\widehat{z}_i;t)}-\frac{\overline{\widehat{F}}_n(t)}{\overline{H}(t)}\Big| + \sup_{t\in[T_1-1,(T+\tau_F)/2]}\Big|\frac{\overline{\widehat{F}}_n(t)}{\overline{H}(t)}-\frac{\overline{F}(t)}{\overline{H}(t)}\Big| + \sup_{t\in D}\sup_{s\in S(t,2h(n))}\Big|\frac{\overline{F}(s)}{\overline{H}(s)}-\frac{1}{\overline{G}(t)}\Big|\Big\}.$$

The first two terms in the braces in the proceeding inequality are $O_p(\beta_n)$ in view of Lemma 3.1 and Lemma 3.2. For the last term note that

$$\sup_{s\in S(t,2h(n))}\Big|\frac{\overline{F}(s)}{\overline{H}(s)}-\frac{1}{\overline{G}(t)}\Big| \leq \frac{4h(n)}{\overline{G}^2(t+2h(n))}\cdot\frac{G(t+2h(n))-G(t-2h(n))}{4h(n)}.$$

Since

$$\sup_{t\in D}\Big|\frac{G(t+2h(n))-G(t-2h(n))}{4h(n)}-g(t)\Big|\to 0,$$

by the fact that

$$\sup_{t\in D}\overline{G}(t+2h(n)) \geq \overline{G}\Big(\frac{T+\tau_F}{2}\Big) > 0$$

holds when n large enough, we have

$$\sup_{t\in D}\sup_{s\in S(t,2h(n))}\Big|\frac{\overline{F}(s)}{\overline{H}(s)}-\frac{1}{\overline{G}(t)}\Big| = o_p(1).$$

Then we have

$$\sup_{t\in D} I_{n4} = o_p(1)$$

holds as $n\to\infty$.

Thus the proof of Theorem 3.1 is completed. □

§4. Simulations

In this section we do simulations to assess the performance of the proposed method. For the censored regression model, data are generated as follows:

$$t_i = \min(\theta_0+\theta_1 x_i+e_i, C_i), \tag{4.1}$$

where $(\theta_0, \theta_1)^T = (1, 1)^T$, $e_i \sim N(0, 1)$ and x_i are distributed as the uniform distribution with support $[0, 1]$. The censoring variables are set as $C_i \sim E(C)$, where $E(C)$ stands for the exponential distribution with mean $1/C$ and C is a positive constant which is varied to achieve certain censoring percent.

With the data generated by above procedure, we estimate (θ_0, θ_1) by the method proposed by Stute[8] firstly and then the density function of e_i are estimated by the method proposed in Section 1 where $h(n) = n^{-1/10}$ and K is chosen as the density function of the uniform distribution with support $[-1, 1]$. Let I_n stands for the maximum of $|\widehat{f}_n(t) - f(t)|$. To investigate the performance of the proposed method, for each scenario, 500 random samples are drawn. The simulation results including the various values of C, the average shares (in percent) of censored observations depending on constant C, the different sample size n and the mean and standard deviation of I_n are summarized in Table 1.

Table 1 Simulation results. All values in columns 4 and 5 are multiplied by 10^5.

C	Sample size	Censoring percent	Mean	Standard deviation
0.25	50	29.956	8385	4200
	100	29.838	6620	3258
	200	30.126	5363	3578
	500	29.987	3791	1891
	1000	29.823	3054	1380
	10000	30.014	1389	579
0.45	50	44.664	9898	5027
	100	44.896	7826	4013
	200	44.914	5782	2875
	500	45.102	4352	2115
	1000	45.081	3364	1590
	10000	45.023	1463	617

From the simulation results in Table 1, we can get the conclusion that as the sample size n converges to infinite, both mean and standard deviation of I_n converge to zero. As an example, the true density function $f(t)$ of the error terms, and its estimator $\widehat{f}_n(t)$ with $C = 0.25$ and $n = 50, 200, 1000, 10000$ respectively, are shown in Figure 1. From Figure 1 we can see that the maximum distance between $\widehat{f}_n(t)$ and $f(t)$ is converges to zero as the sample size n converges to infinite. Results both in Table 1 and in Figure 1 are consistent with Theorem 3.3 in Section 3.

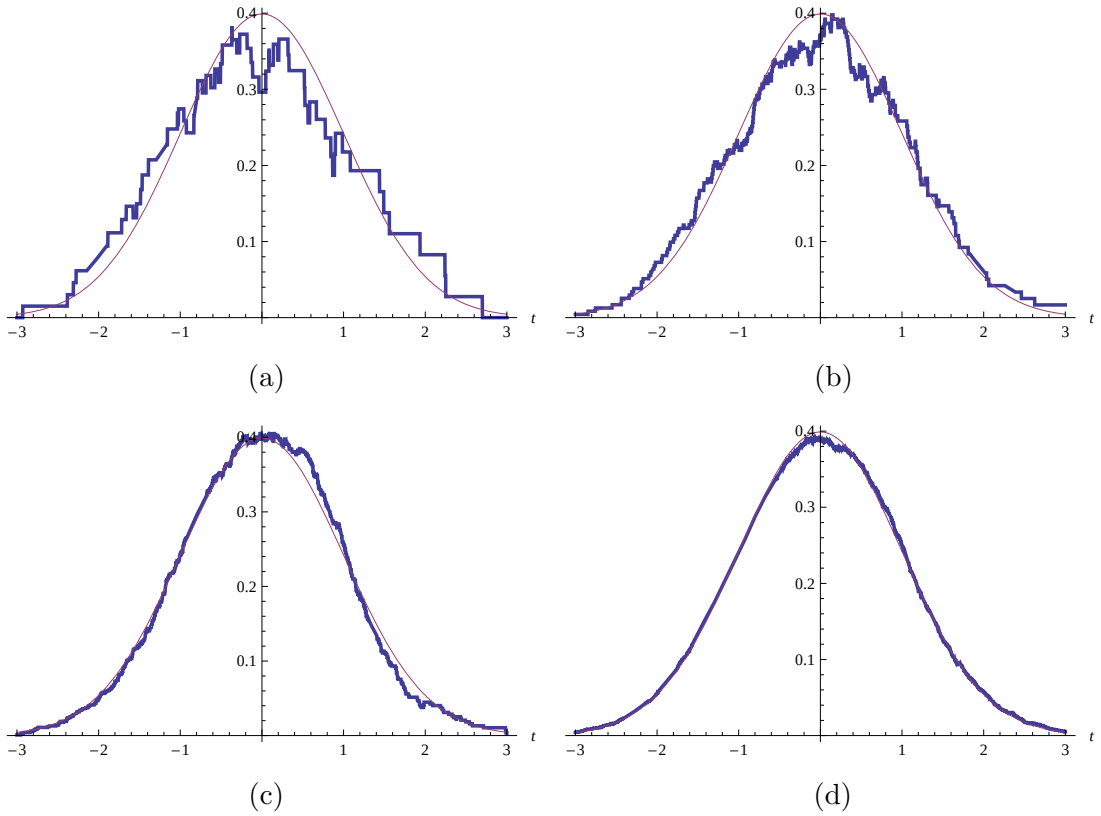

Figure 1 Plot of the true density function $f(t)$ (thinner lines) of the error terms and its estimator $\widehat{f}(t)$ (thicker lines) proposed in Section 1. What shown in (a), (b), (c), (d) are $f(t)$ versus $\widehat{f}(t)$ with $C = 0.25$ and $n = 50, 200, 1000, 10000$ respectively.

Acknowledegments The authors would like to thank the referees for their helpful comments.

References

[1] Cheng, F., Consistency of error density and distribution function estimators in nonparametric regression, *Statistics and Probability Letters*, **59**(2002), 257–270.

[2] Cheng, F., Weak and strong uniform consistency of a kernel error density estimator in nonparametric regression, *J. Statist. Plann. Infer.*, **119**(2004), 95–107.

[3] Efromovich, S., Estimation of the density regression errors, *Ann. Statist.*, **33**(2005), 2194–2227.

[4] Buckley, J. and James, I., Linear regression with censored data, *Biometrika*, **66**(1979), 429–436.

[5] Tsiatis, A.A., Estimation regression parameters using linear rank tests for censored data, *Ann. Statist.*, **18**(1990), 354–372.

[6] Lai, T.L. and Ying, Z., Large sample theory of a modified Buckley-James estimator for regression analysis with censored data, *Ann. Statist.*, **19**(1991), 1370–1402.

[7] Fan, J. and Gijbels, I., Censored regression: Local linear approximations and their applications, *Journal of the American Statistical Association*, **89**(1994), 560–570.

[8] Stute, W., Consistent estimation under random censorship when covariables are present, *J. Multivariate Anal.*, **45**(1993), 89–103.

[9] Stute, W., Nonlinear censored regression, *Statistica Sinica*, **9**(1999), 1089–1102.

[10] Van Keilegom, I. and Veraverbeke, N., Density and hazard estimation in censored regression models, *Bernoulli*, **8**(2002), 607–625.

[11] Ould-Saïd, E. and Cai, Z., Strong uniform consistency of nonparametric estimation of the censored conditional mode function, *Journal of Nonparametric Statistics*, **17**(2005), 797–806.

[12] Serfling, R.J., *Approximation Theorems of Mathematical Statistics*, New York: Wiley, 1980.

[13] Massart, P., The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality, *Ann. Probab.*, **18**(1990), 1269–1283.

[14] Mielniczuk, J., Some asymptotic properties of kernel estimators of a density function in case of censored data, *Ann. Statist.*, **14**(1986), 766–773.

[15] Silverman, B.W., Weak and strong uniform consistency of the kernel estimates of a density and its derivatives, *Ann. Statist.*, **6**(1978), 177–184.

[16] Efron, B., The two-sample problem with censored data, *Proc. Fifth Bekeley Symp. Math. Statist. Prob.*, **4**(1967), 831–853.

删失数据回归误差项的非参数密度估计的一致相合性

周秀轻

(南京师范大学仙林校区数学科学学院, 南京, 210046)

史宁中

(东北师范大学数学与统计学院, 长春, 130024)

赵 进

(南京大学数学系, 南京, 210093)

回归误差项是不可观测的. 由于回归误差项的密度函数在实际中有许多应用, 故使用非参数方法对其进行估计就成为回归分析中的一个基本问题. 针对完全观测数据回归模型, 曾有作者对此问题进行了研究. 然而在实际应用中, 经常会有数据被删失的情况发生, 在此情况下, 可以利用删失回归残差, 并使用核估计的方法对回归误差项的密度函数进行估计. 本文研究了该估计的大样本性质, 并证明了估计量的一致相合性.

关键词: 非参数估计, 删失数据, 回归残差, 一致相合性.

学科分类号: O212.7.