Mashhad R. J. Math. Sci., Vol. 2(1)(2009) 53-60

Asymptotic normality of the truncation probability estimator for truncated dependent data*

S. Jomhoori†(⊠), V. Fakoor and H.A. Azarnoosh
Department of Statistics, Faculty of Mathematical Sciences,
Ferdowsi University of Mashhad, Iran

Abstract

In some long term studies, a series of dependent and possibly truncated lifetimes may be observed. Suppose that the lifetimes have a common marginal distribution function. In left-truncation model, one observes data (X_i, T_i) only, when $T_i \leq X_i$. Under some regularity conditions, we provide a strong representation of the $\hat{\beta}_n$ estimator of $\beta = P(T_i \leq X_i)$, in the form of an average of random variables plus a remainder term. This representation enables us to obtain the asymptotic normality for $\hat{\beta}_n$.

Keywords and phrases: α-mixing, left-truncation, product-limit estimator, strong representation, truncation probability.
AMS Subject Classification 2000: Primary 12J15, 26A03; Secondary 26E30.

1 Introduction

In medical follow-up or in engineering life testing studies, one may not be able to observe the variable of interest, referred to hereafter as the lifetime. Among the different forms in which incomplete data appears, right censoring and left-truncation are two common ones. Left truncation may occur if the time origin of

*Received: 18 October 2008; Revised: 26 May 2009
†e-mail: sa_jo67@stu-mail.um.ac.ir

the lifetime precedes the time origin of the study. Only the subjects which are failed after the start of the study are followed, otherwise they are left truncated. Woodroofe [9] reviews examples from astronomy and economy where such data may occur.

Let $X_1, X_2, \ldots$ be a sequence of the lifetime variables which may not be mutually independent, but have a common continuous marginal distribution function F. Let $T_1, T_2, \ldots$ be a sequence of independent and identically distributed random variables with continuous distribution function G. They are also assumed to be independent of the random variables X_i's. In the left-truncation model, (X_i, T_i) is observed when $T_i \leq X_i$. Let $(X_1, T_1), \ldots, (X_n, T_n)$ be only the sample one observes (i.e., $T_i \leq X_i$), and $\beta > 0$, where

$$\beta = P(T_1 \leq X_1) = \int_{-\infty}^{\infty} G(s)dF(s), \tag{1}$$

is the truncation probability (TP).

Assume, without loss of generality, that X_i and $T_i, i = 1, \ldots, n$, are non-negative random variables. For any distribution function H, we denote the left and right endpoints of its support by $a_H = \inf\{z : H(z) > 0\}$ and $b_H = \sup\{z : H(z) < 1\}$, respectively. Then under the current model, as discussed by Woodroofe [9], we assume that $a_G \leq a_F$ and $b_G \leq b_F$. Equation (1) suggests estimating β by

$$\beta_n = \int_{-\infty}^{\infty} G_n(s)dF_n(s), \tag{2}$$

provided good estimates F_n and G_n for F and G can be obtained.

For the case in which the lifetime observations are mutually independent, Woodroofe [9] proved that if F_n and G_n are product-limit estimates (given by (4) below), β_n converges in probability to β as $n \to \infty$. Under similar conditions as in Woodroofe [9], the asymptotic normality of $\sqrt{n}(\beta_n - \beta)$ has been investigated by several authors using different methods. Chao [7] used influence curves and Keiding and Gill [6] used finite Markov processes and the well-known delta method. Since F_n and G_n have complicated product-limit forms, the properties of β_n is generally not easy to study. Let $I(A)$ denotes the indicator function of

the event A. He and Yang [4] proposed, instead, another estimate of β as

$$\widehat{\beta}_n = \frac{G_n(x)(1 - F_n(x))}{C_n(x)},$$

for all x for which $C_n(x) > 0$, where

$$C_n(x) = n^{-1} \sum_{i=1}^{n} I(T_i \leq x \leq X_i),$$

is the empirical distribution of

$$C(x) = P(T_1 \leq x \leq X_1 | T_1 \leq X_1) = \beta^{-1}(1 - F(x))G(x).$$

Using $\widehat{\beta}_n$, He and Yang [4] proved the almost sure convergence of the estimate of β and obtained a manageable i.i.d. representation for $\widehat{\beta}_n$, hence the asymptotic normality of the estimate.

Our basic aim in this article is to express the TP estimator $\widehat{\beta}_n$ as an average of a sequence of bounded random variables plus a remainder of order $O(n^{-1/2}(\log n)^{-\delta})$ for some $\delta > 0$, for the case in which the underlying lifetimes are assumed to be α-mixing whose definition is given below. As a result, the asymptotic normality of TP estimator is obtained.

Let $\mathcal{F}_i^k$ denote the σ-field of events generated by $\{Y_j; i \leq j \leq k\}$. For easy reference, let us recall the following definition.

Definition. Let $\{Y_i, i \geq 1\}$ denote a sequence of random variables. Given a positive integer n, set

$$\alpha(n) = \sup_{k \geq 1} \{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_1^k, B \in \mathcal{F}_{k+n}^{\infty}\}. \tag{3}$$

The sequence is said to be α-mixing (strongly mixing) if the mixing coefficient $\alpha(n) \to 0$ as $n \to \infty$. Among various mixing conditions used in the literature, α-mixing is reasonably weak and has many practical applications (see, e.g. [1] for more details). In particular, the stationary autoregressive-moving average (ARMA) processes, which are widely applied in time series analysis, are α-mixing with exponential mixing coefficient, i.e., $\alpha(n) = e^{-\nu n}$, for some $\nu > 0$.

The rest of the present paper is organized as follows. In Section 2, we provide the strong representation results for the TP estimator. The proofs are given in Section 3.

2 Strong representation for the TP estimator

We first introduce some notation before stating the strong representation result. The random truncation model is defined by the joint distribution

$$H(x,t) = P(X_1 \leq x, T_1 \leq t | T_1 \leq X_1)$$

with marginal distributions,

$$F^*(x) = P(X_1 \leq x | T_1 \leq X_1) = \beta^{-1} \int_0^x G(s) dF(s),$$

and

$$G^*(x) = P(T_1 \leq x | T_1 \leq X_1) = \beta^{-1} \int_0^x (1 - F(s)) dG(s).$$

Let F_n^* and G_n^* be the empirical distributions of F^* and G^* defined by
$F_n^*(x) = n^{-1} \sum_{i=1}^n I(X_i \leq x)$ and $G_n^*(x) = n^{-1} \sum_{i=1}^n I(T_i \leq x)$.
The well-known product-limit (PL) estimates of F_n and G_n are defined by

$$\widehat{F}_n(x) = 1 - \prod_{i:X_i \leq x} (1 - \frac{1}{nC_n(X_i)}), \qquad \widehat{G}_n(x) = \prod_{i:T_i > x} (1 - \frac{1}{nC_n(T_i)}). \tag{4}$$

For construction of these estimates, see [9] or [7]. Suppose

$$\int \frac{dF^*(s)}{C^2(s)} < \infty \qquad \text{and} \qquad \int \frac{dG^*(s)}{C^2(s)} < \infty. \tag{5}$$

Let

$$\psi_1(x,t,y) = \frac{I(x \leq y)}{C(x)} - \int_0^y \frac{I(t \leq s \leq x)}{C^2(s)} dF^*(s),$$

and

$$\psi_2(x,t,y) = \frac{I(t > y)}{C(t)} - \int_y^\infty \frac{I(t \leq s \leq x)}{C^2(s)} dG^*(s).$$

Then, $E\psi_1(X_i, T_i, y) = E\psi_2(X_i, T_i, y) = 0$, and

$$Cov(\psi_1(X_i, T_i, y_1), \psi_1(X_i, T_i, y_2)) = \int_{a_{F^*}}^{y_1 \wedge y_2} \frac{dF^*(s)}{C^2(s)},$$

and

$$Cov(\psi_2(X_i,T_i,y_1),\psi_2(X_i,T_i,y_2)) = \int_{y_1\wedge y_2}^{b_{G^*}} \frac{dG^*(s)}{C^2(s)}.$$

The following theorem provides the strong representation for $\widehat{\beta}_n$.

Theorem 2.1. Suppose that $\{X_n; n \geq 1\}$ is a sequence of stationary α-mixing random variables with $\alpha(n) = O(n^{-\upsilon})$, for some $\upsilon > 3$. If $a_G < a_F$, then

$$\widehat{\beta}_n - \beta = -\beta\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,T_i) + R_n(y), \tag{6}$$

is uniformly in $0 \leq y \leq b < b_F$, where

$$\sup_{0\leq y\leq b}|R_n(y)| = O(n^{-1/2}(\log n)^{-\delta}) \qquad a.s.$$

for some $\delta > 0$ depending only on υ. We next present the asymptotic normality of the TP estimator based on our strong representation result.

Theorem 2.2. Under the assumptions of Theorem 2.1, if $a_G < a_F$, then for $0 \leq y \leq b < b_F$,

$$\sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} N(0,\sigma^2), \tag{7}$$

where

$$\sigma^2 = \beta^2\{Var(\psi(X_1,T_1,y)) + 2\sum_{i=2}^{\infty} Cov(\psi(X_1,T_1,y),\psi(X_i,T_i,y))\}.$$

3 Proofs

In order to prove Theorem 2.1, we need the following lemma which is Theorem 2.1 in Sun and Zhou [8]. Note that the proof of (9) is similar to that of (8) and is therefore omitted.

Lemma 3.1. Suppose that $\{X_n; n \geq 1\}$ is a sequence of α-mixing random variables with $\alpha(n) = O(n^{-\upsilon})$, for some $\upsilon > 3$. If $a_G < a_F$, then

$$\widehat{F}_n(y) = F(y) + (1-F(y))\frac{1}{n}\sum_{i=1}^{n}\psi_1(X_i,T_i,y) + R_{n1}(y) \quad a.s. \tag{8}$$

and

$$\widehat{G}_n(y) = G(y) - G(y)\frac{1}{n}\sum_{i=1}^{n}\psi_2(X_i, T_i, y) + R_{n2}(y) \quad a.s., \tag{9}$$

uniformly in $0 \leq y \leq b < b_F$, where

$$\sup_{0\leq y\leq b} |R_{n1}(y)| = O(n^{-1/2}(\log n)^{-\delta}) \qquad a.s.$$

and

$$\sup_{0\leq y\leq b} |R_{n2}(y)| = O(n^{-1/2}(\log n)^{-\delta}) \qquad a.s.,$$

for some $\delta > 0$ depending only on υ.

Proof of Theorem 2.1. Using Lemma 3.1, for $0 \leq y \leq b < b_F$, with probability 1 for large n, we have

$$\begin{aligned}
\widehat{\beta}_n - \beta &= \frac{G_n(y)(1-F_n(y))}{C_n(y)} - \frac{G(y)(1-F(y))}{C(y)} \\
&= \frac{(1-F(y))C(y)G(y)}{C_n(y)C(y)}\{-\frac{1}{n}\sum_{i=1}^{n}\psi_1(X_i,T_i,y) - \frac{1}{n}\sum_{i=1}^{n}\psi_2(X_i,T_i,y) \\
&- \frac{1}{nC(y)}\sum_{i=1}^{n}[I(T_i \leq y \leq X_i) - C(y)]\} + O(n^{-1/2}(\log n)^{-\delta}) \\
&= -\beta\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,T_i,y) + O(n^{-1/2}(\log n)^{-\delta}) \quad a.s.,
\end{aligned}$$

where

$$\begin{aligned}
\psi(X_i,T_i,y) &= \psi_1(X_i,T_i,y) + \psi_2(X_i,T_i,y) + \frac{1}{C(y)}[I(T_i \leq y \leq X_i) - C(y)] \\
&= \frac{1}{C(X_i)} - \int_0^{b_{F^*}} \frac{I(T_i \leq s \leq X_i)}{C^2(s)} dF^*(s) - 1 \quad a.s.
\end{aligned}$$

It is easy to see from Lemma 1 of Cai [1] that $\{\psi(X_i, T_i, y); T_i \leq X_i,\ i = 1, 2, \ldots\}$ is a sequence of stationary α-mixing bounded random variables. The random variable $\psi(X_i, T_i, y)$ does not depend on y, therefore, the proof of Theorem 2.1 is complete.

Proof of Theorem 2.2. An application of Theorem 18.5.4 of Ibragimov and Linnik Yu [5] and Theorem 2.1 gives (2.4). It can be easily checked that

$$Var(\psi(X_1,T_1,y) = \int_{a_{F^*}}^{x} \frac{dF^*(s)}{C^2(s)} + \int_x^{b_{G^*}} \frac{dG^*(s)}{C^2(s)} - \frac{1}{C(s)} + 2\alpha - 1,$$

which is finite under (5). On the other hand, $\alpha(n) = O(n^{-\nu}), \nu > 3$ implies $\sum \alpha(n) < \infty$ and therefore $\sum_{i=2}^{\infty} Cov(\psi(X_1, T_1, y), \psi(X_i, T_i, y)) < \infty$. So, σ^2 is a positive finite number and the proof of Theorem 2.2 is complete.

4 Acknowledgment

The authors would like to thank the referees for their careful constructive comments. The authors wish to acknowledge partial support from "Ordered and Spatial Data Centre of Excellence" at Ferdowsi University of Mashhad, Iran.

References

[1] Cai, Z., Kernel density and hazard rate estimation for censored dependent data, *J. Multivariate. Anal.* **67**(1998a), 23–34.

[2] Cai, Z., Asymptotic properties of Kaplan–Meier estimator for censored dependent data, *Statist. Probab. Lett.* **37**(1998b), 381–389.

[3] Chao, M.T., Influence curves for randomly truncated data, *Biometrika.* **74**(1987), 426–429.

[4] He, S. and Yang, G., Estimation of the truncation probability in the random truncation model, *Ann. Statist.* **26**(1998), 1011–1027.

[5] Ibragimov, I.A. and Linnik Yu, V., Independent and Stationary Sequences of Random Variables, Walters-Noordhoff, Groningen, The Netherlands, 1971.

[6] Keiding, N. and Gill, R.D., Random truncation models and Markov processes *Ann. Statist.* **18**(1990), 582–602.

[7] Wang, M.C., Jewell, N.P. and Tsai, W.Y., Asymptotic properties of the product limit estimate under random truncation, *Ann. Statist.* **14**(1986), 1597–1605.

[8] Sun, L. and Zhou, X., Survival function and density estimation for truncated dependent data, *Statist. Probab. Lett.* **52**(2001), 47–57.

[9] Woodroofe, M., Estimating a distribution function with truncated data, *Ann. Statist.* **13**(1985), 163–177.