Kaplan-Meier Estimator for Associated Random Variables Under Left Truncation and Right Censoring

Jabbari, H. [1]

Department of Statistics, Ferdowsi University of Mashhad

Abstract

It is assumed that in long term studies the lifetimes are positively (negatively) associated random variables. Under some regular conditions, the strong convergence rates of Kaplan-Meier estimator of marginal distribution function F and cumulative hazard function Λ are obtained. In order to demonstrate the empirical performance of the results, simulation studies are done.

Keywords: Censored data, Kaplan-Meier estimator, Negative association, Positive association, Strong consistency, Truncation.

1 Introduction

Let $\{X_n, n \geq 1\}$ be sequence of the lifetime variables which may not be mutually independent, but have a common continuous marginal distribution function (df) F. Let $\{T_n, n \geq 1\}$ be a sequence of iid rv's with continuous df G. Suppose that the rv's X_i be censored on the right by the rv's Y_i, so that one observe only $Z_i = X_i \wedge Y_i$ and $\delta_i = I(X_i \leq Y_i)$ where $\wedge$ denotes minimum and $I(.)$ is the indicator function. In this random censorship model, the censoring times $Y_i,\ i = 1, ..., n$ are assumed to be iid rv's with df H and be independent of the X_i's and T_i's. The problem at hand is that of drawing nonparametric inference about F, based on the right censored and left truncated observations $(Z_i, T_i, \delta_i),\ i = 1, ..., n$. In the left truncated model, (Z_i, T_i) is observed only when $Z_i \geq T_i$. Let $\gamma \equiv P(T_1 \leq Z_1) > 0$. Assume, without loss of generality, that X_i, T_i and Y_i are nonnegative rv's, $i = 1, ..., n$. For any df L denotes the left and right endpoints of its support by $a_L = \inf\{x;\ L(x) > 0\}$ and $b_L = \sup\{x;\ L(x) < 1\}$, respectively. Then under the current model, we assume that $a_G \leq a_W$ and $b_G \leq b_W$, where W be the df of Z. Let $\Lambda(x)$ denotes the cumulative hazard function of F, $C(x) = P(T_1 \leq x \leq Z_1 | T_1 \leq Z_1) = \gamma^{-1} P(T_1 \leq x \leq Y_1) \times (1 - F(x))$

[1] Jabbarinh@um.ac.ir

and $W_1(x) = P(Z_1 \leq x, \delta_1 = 1|T_1 \leq Z_1)$. Let $C_n(x)$ and $W_{1n}(x)$ be the empirical estimators of $C(x)$ and $F_1(x)$, respectively, i.e. $C_n(x) = n^{-1}\sum_{i=1}^n I(T_i \leq x \leq Z_i)$ and $W_{1n}(x) = n^{-1}\sum_{i=1}^n I(Z_i \leq x,\ \delta_i = 1)$, where $F_1(x) = P(Z_1 \leq x,\ \delta_1 = 1)$. Then, the PL estimator of F and $\Lambda(x)$ are $\hat{F}_n(x) = \begin{cases} 1 - \prod_{Z_i<x}(1 - \frac{1}{nC_n(Z_i)})^{\delta_i} & ; x < Z_{(n)} \\ 1 & ; x \geq Z_{(n)} \end{cases}$ and $\hat{\Lambda}_n(x) = \sum_{i=1}^n \frac{I(Z_i \leq x,\ \delta_i=1)}{nC_n(Z_i)}$, respectively.

For independent failure time observations, the PL estimator has been studied extensively by many investigators. However, there are preciously few results available for the dependent case. Our focus in present paper is to study asymptotic properties of PL estimator for the right censored and left truncated data under PA (NA) failure times. So in Section 2, we introduce preliminary results and discuss strong uniform consistency and rates of convergence for the estimators $\hat{F}_n$ and $\hat{\Lambda}_n(x)$. Finally in Section 3, we use a simulation study to show the convergence rates. We now introduce general assumptions to be used throughout the article.

(A1). $\{X_n,\ n \geq 1\}$ is stationary sequence of PA (NA) rv's having bounded density function and finite second moment.

(A2). The censoring time variables $\{Y_n,\ n \geq 1\}$ and truncated time variables $\{T_n,\ n \geq 1\}$ are iid rv's with bounded density and are independent of $\{X_n,\ n \geq 1\}$.

(A3). $\sum_{j=2}^{\infty} j^{-2}\sum_{i=1}^{j-1}|Cov(X_i, X_j)|^{1/3} < \infty$.

(A4). $\sum_{j=n+1}^{\infty}|Cov(X_1, X_j)|^{1/3} = O(n^{-(r-2)/2})$, for some $r > 2$.

For the dfs F, G and H (the possibly infinite) times τ_F, τ_G and τ_H by $\tau_F = \inf\{y;\ F(y) = 1\}$, $\tau_G = \inf\{y;\ G(y) = 1\}$, $\tau_H = \inf\{y;\ H(y) = 1\}$, $a_F = \sup\{y;\ F(y) = 0\}$, $a_G = \sup\{y;\ G(y) = 0\}$ and $a_H = \sup\{y;\ H(y) = 0\}$. Then for the marginal df W of the Z_i's, it holds $\tau_W = \tau_F \wedge \tau_H$ and $a_W = a_F \wedge a_H$.

2 Strong uniform consistency with rates

In this section, we introduce preliminary and main results. Let $\{X_n,\ n \geq 1\}$ be a stationary sequence of rv's. Then:

i) If rv's are PA under (A3), $\frac{1}{n}\sum_{i=1}^n (X_i - EX_i) \to 0$ *a.s.*

ii) If rv's are NA with finite first moment, the convergence of (i) holds. Suppose that (A1) and (A2) hold. Then

i) If the rv's $\{X_i,\ i \geq 1\}$ are PA and (A3) is fulfilled, it holds

$$\sup_{a_w \leq x \leq \tau_W} |C_n(x) - G(x)[1 - W(x)]| \longrightarrow 0 \quad a.s., \tag{1}$$

$$\sup_{a_w \leq x \leq \tau_W} |W_{1n}(x) - F_1(x)| \longrightarrow 0 \quad a.s. \tag{2}$$

ii) If the rv's $\{X_n,\ n \geq 1\}$ are NA, (1) and (2) hold true.

Theorem 1. *Under (A1) and (A2) and for any $a_W < \tau < \tau_W$,*
i) If the rv's $\{X_i,\ i \geq 1\}$ are PA and (A3) is satisfied, it holds

$$\sup_{a_W \leq x \leq \tau_W} |\hat{\Lambda}_n(x) - \Lambda(x)| \longrightarrow 0. \qquad a.s. \tag{3}$$

ii) If the rv's $\{X_i,\ i \geq 1\}$ are NA, then (3) holds.

Theorem 2. *Under (A1) and (A2) and the additional assumptions either in part (i) or part (ii) of Theorem 1, it holds*

$$\sup_{a_W \le x \le \tau_W} |\hat{F}_n(x) - F(x)| \longrightarrow 0 \quad a.s., \qquad \sup_{Z_{1:n} \le x \le Z_{n:n}} |\hat{F}_n(x) - F(x)| \longrightarrow 0 \quad a.s.,$$

where $Z_{n:n} = \max_{i \le n} Z_i$ *and* $Z_{1:n} = \min_{i \le n} Z_i$.

Theorem 3. *Suppose that (A1), (A2) and (A4) hold. Then, for any* $a_W \le a < \tau \le \tau_W$,

$$\sup_{a \le x \le \tau} |\hat{\Lambda}_n(x) - \Lambda(x)| = O(n^{-\theta}) \quad a.s., \tag{4}$$

where $0 < \theta < (r-2)/(2r+2+\delta)$, *for any* $\delta > 0$ *and* r *in (A4).*

Theorem 4. *Under the assumptions of Theorem 3, it follows*

$$\sup_{a \le x \le \tau} |\hat{F}_n(x) - F(x)| = O(n^{-\theta}). \quad a.s. \tag{5}$$

3 Simulation study

In this section, we intend to compare our results with simulation of such generated NA (PA) data of size n=10(1)1000 to check the goodness convergence of the estimators. For generating NA data as introduced by Cai and Roussas (1998), we could use n-variate normal distribution with $\mu' = (12, 12, ..., 12)$ and

$$\Sigma = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & -\rho^2 \dots & -\rho^{n-1} \\ -\rho & 1 & -\rho \dots & -\rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho^{n-1} & -\rho^{n-2} & -\rho & 1 \end{pmatrix}. \tag{6}$$

We have a vector with NA property when $\rho > 0$. We set $\rho = 0.2$ and the censored and truncation samples are generated from $N(13,1)$ and $N(11,1)$, respectively. So, we calculate $\hat{F}_n(x)$, $\hat{\Lambda}_n(x)$, $d_{F_n} = \sup_{a_W \le x \le \tau_W} |\hat{F}_n(x) - F(x)|$ and $d_{\Lambda_n} = \sup_{a_W \le x \le \tau_W} |\hat{\Lambda}_n(x) - \Lambda(x)|$ for some n. Figure 1 shows the results for this two functions against n and the green line is the convergence rates (4) and (5) using θ is equal to 0.27 and 0.12, respectively. In

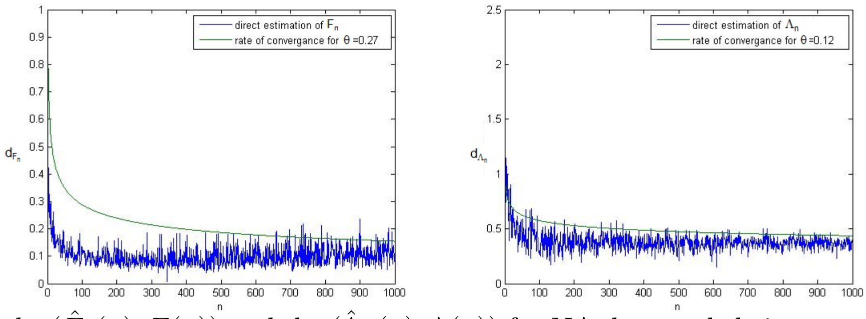

Figure 1: $d_{F_n}(\hat{F}_n(x), F(x))$ and $d_{\Lambda_n}(\hat{\Lambda}_n(x), \Lambda(x))$ for NA data and their convergence rates (green line).

both graphs of Figure 1, we can see that the convergence rates are reasonable in NA case i.e.:

(a) in the left graph the convergence rate could get sharper and this graph shows that the convergence behavior of $\hat{F}_n(x)$ is good.
(b) in the right graph however the convergence rate isn't reasonable as well as left graph, but it is good enough to present. Since $d_{\Lambda(x)} \in [0, +\infty)$, the differences more than one could be reasonable.
For generating PA sample, we follow the same way as in NA case with

$$\Sigma = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 \dots & \rho^{n-1} \\ \rho & 1 & \rho \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho & 1 \end{pmatrix}. \tag{7}$$

We have a vector with PA property when $\rho > 0$. We set $\rho = 0.2$ and generate the censored and truncation samples as NA case. Figure 2 shows the trend of d_{F_n} and d_{Λ_n} with respect to n and the green line is the convergence rates (4) and (5) using θ is equal to 0.27 and 0.12, respectively. Figure 2 shows the same results as in NA case.

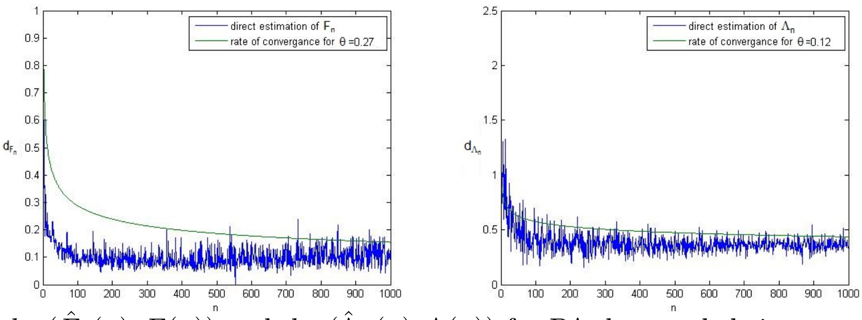

Figure 2: $d_{F_n}(\hat{F}_n(x), F(x))$ and $d_{\Lambda_n}(\hat{\Lambda}_n(x), \Lambda(x))$ for PA data and their convergence rates (green line).

References

[1] Cai, Z.W. and Roussas, G.G., (1998). Kaplan-Meier estimator under association, *J. Multivariate Anal.*, 67, 318-348.