ESTIMATION OF CONDITIONAL SURVIVAL FUNCTION UNDER DEPENDENT RANDOM CENSORED DATA

A.A. Abdushukurov
Branch of Moscow State University in Tashkent named after M.V.Lomonosov
Tashkent, UZBEKISTAN
e-mail: a_abdushukurov@rambler.ru

Abstract

The aim of paper is considering the problem of estimation of conditional survival function in the case of right random censoring with presence of covariate.
Keywords: data science, conditional survival function, right random censoring

Let us consider the case when the support of covariate C is the interval $[0,1]$ and we describe our results on fixed design points $0 \leq x_1 \leq x_2 \leq ... \leq x_n \leq 1$ at which we consider responses (survival or failure times) $X_1, ..., X_n$ and censoring times $Y_1, ..., Y_n$ of identical objects, which are under study. These responses are independent and nonnegative random variables (r.v.-s) with conditional distribution function (d.f.) at x_i,$F_{x_i}(t) = P(X_i \leq t/C_i = x_i)$. They are subjected to random right censoring, that is for X_i there is a censoring variable Y_i with conditional d.f. $G_{x_i}(t) = P(Y_i \leq t/C_i = x_i)$ and at $n-$th stage of experiment the observed data is $S^{(n)} = \{(Z_i, \delta_i, C_i), 1 \leq i \leq n\}$, where $Z_i = min(X_i, Y_i), \delta_i = I(X_i \leq Y_i)$ with $I(A)$ denoting the indicator of event A. Note that in sample S^n r.v. X_i is observed only when $\delta_i = 1$. Commonly, in survival analysis independence between the r.v.-s X_i and Y_i conditional on the covariate C_i is assumed. But, in some practical situations, this assumption does not hold. Therefore, in this article we consider a dependence model in which dependence structure is described through copula function. So let

$$S_x(t_1, t_2) = P(X_x > t_1, Y_x > t_2),\ t_1, t_2 \geq 0,$$

the joint survival function of the response X_x and the censoring variable Y_x at x. Then the marginal survival functions are ${S_x}^X(t) = 1 - F_x(t) = S_x(t, 0)$ and ${S_x}^Y(t) = 1 - G_x(t) = S_x(0, t), t \leq 0$. We suppose that the marginal d.f.-s F_x and G_x are continuous. Then according to the Theorem of Sclar (see, [1]), the joint survival function $S_x(t_1, t_2)$ can be expressed as

$$S_x(t_1, t_2) = C_x(S_x^X(t_1), S_x^X(t_2))\,,\ t_1, t_2 \geq 0, \tag{1}$$

where $C_x(u, v)$ is a known copula function depending on x, S_x^X and S_x^Y in a general way. We consider estimator of d.f. F_x which is equivalent to the relative-risk power estimator [2,3] under independent censoring case.

Assume that at the fixed design value $x \in (0, 1), C_x$ in (1) is Archimedean copula, i.e.

$$S_x(t_1, t_2) = \varphi_x^{-1}(\varphi_x(S_x^X(t_1)) + \varphi_x(S_x^Y(t_2))),\ t_1, t_2 \geq 0, \tag{2}$$

where, for each $x, \varphi_x : [0,1] \to [0,+\infty]$ is a known continuous, convex, strictly decreasing function with $\varphi_x(1) = 0$. We assume that copula generator function φ_x is strict, i.e. $\varphi_x(0) = \infty$ and φ_x^{-1} is a inverse of φ_x. From (2), it follows that

$$P(Z_x > t) = 1 - H_x(t) = \overline{H_x(t)} = S_x{}^Z(t) = S_x(t,t) =$$

$$= \varphi_x^{-1}(\varphi_x(S_x^X(t)) + \varphi_x(S_x^Y(t))),\ t \geq 0, \tag{3}$$

Let $H_x^{(1)}(t) = P(Z_x \leq t, \delta_x = 1)$ be a subdistribution function and $\Lambda_x(t)$ is crude hazard function of r.v. X_x subjecting to censoring by Y_x,

$$\Lambda_x(dt) = \frac{P(X_x \in dt, X_x \leq Y_x)}{P(X_x \geq t, Y_x \geq t)} = \frac{H_x^{(1)}(dt)}{S_x^Z(t-)}. \tag{4}$$

From (4) one can obtain following expression of survival function S_x^X :

$$S_x^X(t) = \varphi_x^{-1}[-\int\limits_0^t \varphi_x'(S_x^Z(u))dH_x^{(1)}(u)],\ \ t \geq 0. \tag{5}$$

In order to constructing the estimator of S_x^X according to representation (5), we introduce smoothed estimators of $S_x^Z, H_x^{(1)}$ and regularity conditions for them. We use the Gasser-Müller weights

$$w_{ni}(x,h_n) = \frac{1}{q_n(x,h_n)} \int_{x_{i-1}}^{x_i} \frac{1}{h_n}\pi(\frac{x-z}{h_n})dz,\ \ i = 1,...,n, \tag{6}$$

with

$$q_n(x,h_n) = \int_0^{x_n} \frac{1}{h_n}\pi(\frac{x-z}{h_n})dz,$$

where $x_0 = 0$, π is a known probability density function (kernel) and $\{h_n, n \geq 1\}$ is a sequence of positive constants, tending to zero as $n \to \infty$, called bandwidth sequence. Let's introduce the weighted estimators of H_x, S_x^Z and $H_x^{(1)}$ respectively as

$$H_{xh}(t) = \sum_{i=1}^n w_{ni}(x,h_n)I(Z_i \leq t),\ \ S_{xh}^Z(t) = 1 - H_{xh}(t),$$

$$H_{xh}^{(1)}(t) = \sum_{i=1}^n w_{ni}(x,h_n)I(Z_i \leq t, \delta_i = 1). \tag{7}$$

Then by pluggin estimators (6) and (7) in (5) we obtained the following intermediate estimator of S_x^X :

$$S_{xh}^X(t) = 1 - F_{xh}(t) = \varphi_x^{-1}[-\int\limits_0^t \varphi_x'(S_x^Z(u))dH_x^{(1)}(u)],\ \ t \geq 0.$$

In this work we propose the next extended analogue of estimator introduced in [2,3]:

$$\widehat{S}_{xh}^{X}(t)=\varphi_x^{-1}[\varphi(S_{xh}^{Z}(t))\cdot\mu_{xh}(t)]=1-\widehat{F}_{xh}(t), \tag{8}$$

where $\mu_{xh}(t) = \varphi(S_{xh}^{X}(t))/\varphi(\widetilde{S}_{xh}^{Z}(t))$, $\varphi(S_{xh}^{X}(t)) = -\int\limits_0^t \varphi_x'(S_{xh}^{Z}(u))dH_{xh}^{(1)}(u)$, $\varphi(\widetilde{S}_{xh}^{Z}(t)) = -\int\limits_0^t \varphi_x'(S_{xh}^{Z}(u))dH_{xh}(u)$. In order to investigate the estimate (8) we introduce some conditions. For the design points $x_1,...,x_n$, denote $\underline{\Delta_n} = \min\limits_{1\le i\le n}(x_i - x_{i-1})$, $\overline{\Delta_n} = \max\limits_{1\le i\le n}(x_i - x_{i-1})$.

For the kernel π, let $\|\pi\|_2^2=\int_{-\infty}^{\infty}\pi^2(u)du,\ m_\nu(\pi)=\int_{-\infty}^{\infty}u^\nu\pi(u)du,\ \nu=1,2.$

Moreover, we use next assumptions on the design and on the kernel function:
(A1) As $n\to\infty, x_n\to 1, \underline{\Delta_n}=O(\frac{1}{n}), \overline{\Delta_n}-\underline{\Delta_n}=o(\frac{1}{n})$.
(A2) π is a probability density function with compact support $[-M,M]$ for some $M>0$, with $m_1(\pi)=0$ and $|\pi(u)-\pi(u')|\le C(\pi)|u-u'|$, where $C(\pi)$ is some constant.

Let $T_{H_x}=inf\{t\ge 0: H_x(t)=1\}$. Then $T_{H_x}=min(T_{F_x},T_{G_x})$. For our results we need some smoothnees conditions on functions $H_x(t)$ and $H_x^{(1)}(t)$. We formulate them for a general (sub)distribution function $N_x(t), 0\le x\le 1, t\in R$ and for a fixed $T>0$.
(A3) $\frac{\partial^2}{\partial x^2}N_x(t)=\ddot{N}_x(t)$ exists and is continuous in $(x,t)\in[0,1]\times[0,T]$.
(A4) $\frac{\partial^2}{\partial t^2}N_x(t)=N_x''(t)$ exists and is continuous in $(x,t)\in[0,1]\times[0,T]$.
(A5) $\frac{\partial^2}{\partial x\partial t}N_x(t)=\dot{N}'_x(t)$ exists and is continuous in $(x,t)\in[0,1]\times[0,T]$.
(A6) $\frac{\partial\varphi_x(u)}{\partial u}=\varphi_x'(u)$ and $\frac{\partial^2\varphi_x(u)}{\partial u^2}=\varphi_x''(u)$ are Lipschitz in the $x-$direction with a bounded Lipschitz constant and $\frac{\partial^3\varphi_x(u)}{\partial u^3}=\varphi_x'''(u)$ exists and is continuous in $(x,u)\in[0,1]\times(0,1]$.

We derive an almost sure representation result with rate.

Theorem 1. Assume (A1), (A2), $H_x(t)$ and $H_x^{(1)}(t)$ satisfy (A3)-(A5) in $[0,T]$ with $T<T_{H_x}$, φ_x satisfies (A6) and $h_n\to 0, \frac{logn}{nh_n}\to 0, \frac{nh_n^5}{logn}=O(1)$. Then, as $n\to\infty$,

$$\widehat{F}_{xh}(t)-F_x(t)=\sum_{i=1}^{n}w_{ni}(x,h_n)\Psi_{tx}(Z_i,\delta_i)+r_n(t),$$

where

$$\Psi_{tx}(Z_i,\delta_i)=\frac{-1}{\varphi_x'(S_x^X(t))}[\int_0^t\varphi_x''(S_x^Z(u))(I(Z_i\le u)-H_x(u))dH_x^{(1)}(u)-$$
$$-\varphi_x'(S_x^Z(t))(I(Z_i\le t,\delta_i=1)-H_x^{(1)}(t))-$$
$$-\int_0^t\varphi_x''(S_x^Z(u))(I(Z_i\le u,\delta_i=1)-H_x^{(1)}(u))dH_x(u)],$$

and

$$\sup_{0\le t\le T}|r_n(t)|\stackrel{a.s.}{=}O((\frac{\log n}{nh_n})^{3/4}).$$

The weak convergence of the empirical process $(nh_n)^{1/2}\{\widehat{F}_{xh}(\cdot)-F_x(\cdot)\}$ in the space $l^\infty[0,T]$ of uniformly bounded functions on $[0,T]$, endowed with the uniform topology is the contents of the next theorem.

Theorem 2. Assume (A1), (A2), $H_x(t)$ and $H_x^{(1)}(t)$ satisfy (A3)-(A5) in $[0,T]$ with $T < T_{H_x}$, and that φ_x satisfies (A6).
(I) If $nh_n^5 \to 0$ and $\frac{(logn)^3}{nh_n} \to 0$, then, as $n \to \infty$,

$$(nh_n)^{1/2}\{\widehat{F}_{xh}(\cdot) - F_x(\cdot)\} \Rightarrow \mathbf{W}_x(\cdot) \ in\ l^\infty[0,T].$$

(II)If $h_n = Cn^{-1/5}$ for some $C > 0$, then, as $n \to \infty$,

$$(nh_n)^{1/2}\{\widehat{F}_{xh}(\cdot) - F_x(\cdot)\} \Rightarrow \mathbf{W}^*_x(\cdot) \ in\ l^\infty[0,T],$$

where $\mathbf{W}_x(\cdot)$ and $\mathbf{W}^*_x(\cdot)$ are Gaussian processes with means

$$E\mathbf{W}_x(t) = 0, E\mathbf{W}^*_x(t) = a_x(t),$$

and same covariance

$$Cov(\mathbf{W}_x(t), \mathbf{W}^(_x s)) = Cov(\mathbf{W}^*_x(t), \mathbf{W}^*_x(s)) = \Gamma_x(t,s),$$

with

$$a_x(t) = \frac{-C^{5/2}m_2(\pi)}{2\varphi_x'(S_x^X(t))}\int_0^t [\varphi_x''(S_x^Z(u))\ddot{H}_x(u)dH_x^{(1)}(u) - \varphi_x'(S_x^Z(u))d\ddot{H}_x^{(1)}(u)],$$

and

$$\Gamma_x(t,s) = \frac{\|\pi\|_2^2}{\varphi_x'(S_x^X(t))\varphi_x'(S_x^X(s))}\{\int_0^{\min(t,s)} (\varphi_x'(S_x^Z(z)))^2 dH_x^{(1)}(z)+$$

$$+\int_0^{\min(t,s)}[\varphi_x''(S_x^Z(w))S_x^Z(w) + \varphi_x'(S_x^Z(w))]\int_0^w \varphi_x''(S_x^Z(y))dH_x^{(1)}(y)dH_x^{(1)}(w)+$$

$$+\int_0^{\min(t,s)}\varphi_x''(S_x^Z(w))\int_w^{\max(t,s)}(\varphi_x''(S_x^Z(y))S_x^Z(y) + \varphi_x'(S_x^Z(y)))dH_x^{(1)}(y)dH_x^{(1)}(w)-$$

$$-\int_0^t[\varphi_x''(S_x^Z(y))S_x^Z(y) + \varphi_x'(S_x^Z(y))]dH_x^{(1)}(y)\cdot$$

$$\cdot\int_0^s[\varphi_x''(S_x^Z(w))S_x^Z(w) + \varphi_x'(S_x^Z(w))]dH_x^{(1)}(w)\}.$$

References

[1] Nelsen R.B. (1999). *An Introduction to Copulas*. Springer, New York.

[2] Abdushukurov A.A. (1998). Nonparametric estimation of distribution function based on relative risk function. *Commun. Statist.: Theory and Methods*. Vol. **27**, No.**8**, pp. 1991-2012.

[3] Abdushukurov A.A. (1999). On nonparametric estimation of reliability indices by censored samples. *Theory Probab. Appl.* Vol. bf 43, No. 1, pp. 3-11.