Jordan Journal of Mathematics and Statistics (JJMS), 17(1), 2024, pp 179 - 197
DOI: https://doi.org/10.47013/17.1.11

B-SPLINE ESTIMATE OF THE REGRESSION FUNCTION UNDER GENERAL CENSORSHIP MODEL

ILHEM LAROUSSI [(1)]

Abstract. In a continuity reasoning of the different estimators proposed by de Kebabi et al. [21] and recently Douas et al. [11] and Laroussi [26]. The construction of the regression function estimator is based on three axes. The first one is the application of the non-parametric estimate, namely, the least-squares technique. The second axis represents the general censorship which combines all the existing types of censorship. Hence, empirical L_2-error estimates are constructed over data-dependent spaces of B-spline functions. The almost sure convergence of the proposed estimator is studied. Essentially, two models subject to twice or right censorship are assessed and this phenomena of censorship identified by the simulation shows the interest of this estimator.

1. Introduction

The fitting procedure is the main purpose of linear regression. The least squares, discovered independently by Legendre [27] and Gauss [15] and published in 1805 and 1809, is the most famous and used solution to such a problem. For parametric methods, rich literature is available, see, for instance, Rao [34], Seber [35], Draper and Smith [10], and the inside literature cited. The random aspect of this concept is marked by the minimization of the risk L_2 or mean squared error which in turn results to the regression function. Interest of such minimization is to construct an estimate with a predicted mean squared error which leads the minimum mean square error. Many general asymptotic empirical risk minimization properties have been proposed by influential papers such as Vapnik and Chervonenkis [39], Vapnik ([40],

Key words and phrases. Least squares regression, B-spline function, censored data, convergence, almost sure.

Received: Dec. 20, 2022 Accepted: Aug. 7, 2023 .

[41]) and Haussler [19], and more recently Montanari and Saeed [31]. The necessary conditions for the least squares estimates reliability are contained in Van de Geer and Wegkamp [38]. Various statistical applications, for penalized modelling, can be found, in Wahba[42] for the existence and computation concept, Green and Silverman [16] applied to generalized Linear Models, Eubank [13] for the L_2 theory, and Eggermont and LaRiccia [12] for the maximum likelihood estimation. Specifically, smoothing splines date back to Whittaker [44] which by using an analytical method called the numerical process of graduation obtains a more reliable approximation. Recently Mariati et al. [29], apply this parametric estimation method of the regression function to study poverty in the province of Papuasie. The definition of the penalized least squares estimates is considered by Mammen and van de Geer [28]. They using a penalty on the total variation of the function. The L_2 error criterion used throughout this item also applies to the non-parametric estimation. This estimator does not limit the class of possible relations. In 1947, Tukey[36] suggest the local mean estimate (partitioning) of the regression similar to " histogram " estimation method (classical partitioning) of the density. The best-known techniques for investigating non-parametric estimates are: classical local averaging including kernel, partitioning and nearest neighbour, least-squares. The latter uses function spline spaces, neural networks spaces, and radial basis function networks. Penalized least squares estimates, local polynomial kernel estimates, and orthogonal series estimates are also very famous. For an overview of these different technical estimates see, e.g., Gy*ö*rfi et al. [17] and more recently Pavel and Sadikoglu [33]. Existing survival analysis approaches can be listed, for example, in Fan and Gijbels [14]. Thus, Beran [3] introduced, in the case of right censoring to estimate conditional survival functions. He also proved that these estimates are consistent, for Nearest Neighbour, kernel, or recursive partitioning weights. The latter being extended by Dabrowska ([8], [9]). They assumed the conditionally independence between the variables of interest and right-censoring and the explanatory variable. In 2015, Casanova and Leconte [6] applied this result to estimate a cumulative distribution function in a finite population.

Kohler and Krzyzk [24] studied smoothing spline regression estimates and proved that the defined estimates are universally consistent.

Adaptive least squares estimates based on sample division were investigated by Györfi et al. [17] and showed an additional result regarding the universal consistency of the estimates. In the same year, Kohler et al. [25] simplified their proof and showed strong consistency for the randomly right-censored case. In the survey list of works, Wegman and Wright [43] and Agarwal [1], cite references relating to the application of splines. Consistency of least squares splines in supremum norm was establish in Zhu [45]. Agarwal and Studden [2] investigate the non-equidistant knots of a least-squares spline estimate in order to minimize the expected error of the estimate for fixed design regression.

The Kaplan-Meier method [20] or proportional regression proposed by Cox [7] have played in most of the above studies an important role in estimating the rate of events at any time and in the calculation of survival and hazard functions. On the other hand, for survival data where the failure time can be censored on the left or on the right said mixed censorship, Patilea and Rolin [32] have made great progress by proposing competing risk models. They obtained product limit estimators for the survival functions risks and derived the strong convergence of the proposed estimators.

Then, Messaci [30] innovates in the study of this model by proposing estimates of local means of $r(x) = E(Y|X = x)$. In evolution, Kebabi et al. [21] deduced least squares estimators for which they proved the convergence of the L_2 norm. Other works followed, including those of Kitouni et al. [22] who proposed a density estimator and established its almost complete convergence, and Boukeloua [4] where the mean square convergence of the estimator with the density rates has been proved. As a more general model comprising left, right, double, and twice censorship, Boukeloua and Messaci [5] proposed a generalized censored function of the interest variable. They also established the asymptotic normality of the density estimator. By coveting the L_2 convergence for a general censorship model similar to that of Boukeloua and Messaci [5] within the framework of the estimation approach proposed by Kebabi et al. [21], we propose to combine three approaches: that of Boukeloua and Messaci [5], that of Kebabi et al. [21], and the so-called B-spline-based approach which assumes that the base functions is as small as possible. In this case and inspired by Kohler [23], the knot sequences of the chosen B-splines depend locally on the data. The originality

in this idea is the exploitation of the generalized censorship function to modify the least-squares estimate and the introduction of the B-spline estimator. This leads to the desired L_2 convergence for a general censorship model in a less complicated way. To the best of our knowledge, this result is unprecedented. It opens the door to the establishment of the asymptotic normality of the least-squares estimator in the case of a general model that summarizes the four types of censorship (left, right, twice, and double). This paper is organized as follows: In Section 2, the data model as well as the tools used, some notations, and the construction of the estimators are given. The concomitant result is presented in section 3 with proof of the obtained theorem. The simulation study for estimators based on right and twice censorship are illustrated in Section 4.

2. B-spline estimator

First, note the B-spline space as the optimization space. Let $M \in \mathbb{N}^*$, and let the spline space $S_{t,M}([t_0, t_K[)$ for $x \in [t_0, t_K[$ by

$$S_{t,M}([t_0,t_K[) = \left\{ f : [t_0,t_K[\longrightarrow \mathbb{R} : \exists\ a_{-M},...,a_{K-1} \in \mathbb{R} \text{ such as } f(x) = \sum_{j=-M}^{K-1} a_j B_{j,M,t}(x) \right\},$$

and $t_{-M} \leq \cdots \leq t_0 \leq \cdots \leq t_K \leq \cdots \leq t_{K+M}$, here M is called the degree and $t = (t_j)_{j=-M,...,K-1}$ is called the knot sequence of $S_{t,M}$, where the B-splines $B_{j,M,t}$ are defined for $x \in \mathbb{R}$ by

- for $j = -M, ..., K + M - 1$ and for $j = -M, ...K - 1$,

$$B_{j,0,t} = \begin{cases} 1 & \text{if} \quad t_j \leq x < t_{j+1} \\ 0 & \text{otherwise} \end{cases}. \tag{2.1}$$

- For $j = -M, ..., K + M - l - 2,\ l = 0, ..., M - 1,\ x \in \mathbb{R}$.

$$B_{j,l+1,t}(x) = \frac{x - t_j}{t_{j+l+1} - t_j} B_{j,l,t}(x) + \frac{t_{j+l+2} - x}{t_{j+l+2} - t_{j+1}} B_{j+1,l,t}(x). \tag{2.2}$$

- And for $x \in \mathbb{R},\ j = -M, ..., K - 1$.

$$B_{j,M,t}(x) \geq 0, \text{ and } \sum_{j=-M}^{K-1} B_{j,M,t} = 1. \tag{2.3}$$

Second, let X be a real co-variable, the response variable Y is non-negative and bounded by $C < \infty$. We estimate $r(x) = E(Y|X = x)$ starting from one sample shaped by i. i. d. observations $\mathcal{D}_n = \{X_i, Z_i, \delta_i \ ; \ \ 1 \leq i \leq n\}$ with same law for (X, Z, δ) where $\delta = 1$ if the observation of Y is censored. If $\delta = 0$, the data Y is observed. Let $g(y) = P(\delta = 0|Y = y)$ the probability of the uncensored data. We write $Z = \max(\min(Y, R), L)$, when Y is right censored by real random variable R and $\min(Y, R)$ is left censored by L. Under the assumption that Y, L and R are positive and independent and that $\delta = 1_{\{L<R<Y\}} + 2 \times 1_{\{\min(Y,R)\leq L\}}$, Patilea and Rolin [32] obtained that $g(t) = F_L(t)S_R(t)$, it is a twice censored data model. If we assuming that $L = 0$ a.s., we come down to the right censored data model and we get $g(t) = S_R(t)$. Turnbull's [37] suggests doubly censored data model, where the lifetime Y is independent of the pair (L, R) and $P(0 \leq L \leq R) = 1$ and $\delta = 1_{\{Y>R\}} + 2 \times 1_{\{Y<L\}}$. It is easy to see that $g(t) = S_R(t) - S_L(t)$.
Finally, according to the estimator proposed by Kebabi et al. [21], for twice censorship model and for a general censored model proposed by Laroussi [26], the least squares B-spline estimator of $r(x)$ is first given by

$$\tilde{r}_n = \underset{f\in S_{t,M}}{\arg\, min} \frac{1}{n}\sum_{i=1}^{n} 1_{\{\delta_i=0\}} \frac{|f(X_i) - Z_i|^2}{g(Z_i)} \qquad \left(\frac{0}{0} := 0\right). \tag{2.4}$$

We identify the hypothesis H that includes the following conditions

- H: $\exists\ I > 0$ such that
  - $\forall n \in \mathbb{N},\ \forall i,\ (1 \leq i \leq n),\ \delta_i = 0 \implies I \leq Z_i \leq C$ a.s,
  - $g(I) = \inf\limits_{y\in\mathbb{R}^+} g(y) > 0.$

As Y is bounded, the estimator of $r(x)$ is given by its truncated version

$$r_n(x) = \mathbb{T}_{[0,M_n]}(\tilde{r}_n(x)), \tag{2.5}$$

where $M_n := \max\{Z_1, \ldots, Z_n\}$ with $M_n \underset{n\to\infty}{\to} C < \beta_n$ p.s. And we need the notation

$$\bar{r}_n(x) = \mathbb{T}_{\beta_n}(\tilde{r}_n(x)). \tag{2.6}$$

Where the truncation operator is defined for $0 \leq t < \infty$ and $x \in \mathbb{R}$, by

$$\mathbb{T}_{[0,t]}(x) = \begin{cases} t & \text{if } x > t \\ x & \text{if } 0 \leq x \leq t \\ 0 & \text{if } x < 0 \end{cases}.$$

The following theorem is the same obtained in Györfi et al. [17] for the complete data case, but our theorem wraps up the three types of censoring (right, twice and double) also our proof is less complicated that of mixed censorship introduced in Kebbabi et al. [21] since it just needs the hypothesis H.

3. Result

Theorem 3.1. [**?**] For $n \in \mathbb{N}$, $K_{\max}(n) \in \mathbb{N}^*$ and $M_{\max}(n) \in \mathbb{N}^*$. Lets $K, M \in \mathbb{N}^*$, $t_{-M}, ..., t_{K+M} \in \mathbb{R}$, such that $K \leq K_{max}(n)$, $M \leq M_{\max}(n)$ and $t_{-M} \leq ... \leq t_0 < ...t_K \leq ... \leq t_{K+M}$, with r_n who checks (2.5)

and we assume that

$$\beta_n \underset{n\to+\infty}{\longrightarrow} +\infty. \tag{3.1}$$

$$\frac{\beta_n^4}{n^{1-\sigma}} \underset{n\to+\infty}{\longrightarrow} 0, \forall \sigma > 0. \tag{3.2}$$

$$\frac{(K_{\max}(n)M_{\max}(n) + M_{\max}(n)^2)\beta_n^4 \log(n)}{n} \underset{n\to\infty}{\longrightarrow} 0. \tag{3.3}$$

In addition for each $C,\ \gamma > 0$ the distribution μ of X checked

$$\mu\left(\left\{(-\infty, t_0) \bigcup_{\substack{k=1,...,K \\ t_k - t_{k-M-1} > \gamma}} [t_{k-1}, t_k) \cup [t_k, \infty)\right\} \cap [-C, C]\right) \underset{n\to\infty}{\longrightarrow} 0 \ a.s. \tag{3.4}$$

$\forall\ C, \gamma > 0$. The empirical distribution μ_n de $X_1, ..., X_n$ checked

$$\mu_n\left(\left\{(-\infty, t_0) \bigcup_{\substack{k=1,...,K \\ t_k - t_{k-M-1} > \gamma}} [t_{k-1}, t_k) \cup [t_k, \infty)\right\} \cap [-C, C]\right) \underset{n\to\infty}{\longrightarrow} 0 \ a.s. \tag{3.5}$$

$\forall\, C, \gamma > 0$, then for $\mathbf{E}Y^2 < \infty$, we have

$$\int |r_n(x) - r(X)|^2 \mu(dx) \underset{n\to\infty}{\longrightarrow} 0 \; a.s.$$

Proof. Let's introduce the following set

$$\mathbb{T}_{\beta_n} S_{t,M} = \{g : \mathbb{R} \to \mathbb{R} : \exists f \in S_{t,M}, \forall x \in \mathbb{R} \quad g(x) = \mathbb{T}_{[0,\beta_n]} f(x)\}.$$

The proof of our theorem is based on the following inequality

$$\int |r_n(x) - r(X)|^2 \mu(dx) \tag{3.6}$$

$$\leq 2 \underset{f\in\mathbb{T}_{\beta_n} S_{t,M}}{sup} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=0\}} \frac{|f(X_i) - Z_i|}{g(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right| \tag{3.7}$$

$$+ \underset{f\in S_{t,M}, \|f\|_\infty < \beta_n}{inf} \int |f(x) - r(X)|^2 \mu(dx). \tag{3.8}$$

We start by showing this inequality.

- On one side, we have

$$\begin{aligned} &\int |r_n(x) - r(X)|^2 \mu(dx) \\ &= \left\{ E(\, |r_n(x) - Y|^2 | \mathcal{D}_n) \; - \inf_{f\in S_{t,M}, \|f\|_\infty < \beta_n} \mathbf{E}|f(X) - Y|^2 \right\} \\ &\quad + \left\{ \inf_{f\in S_{t,M}, \|f\|_\infty < \beta_n} \mathbf{E}(\, |f(X) - Y|^2 - \mathbf{E}(\, |r(X) - Y|^2 \right\}. \end{aligned}$$

- Moreover, the regression function checks

$$\inf_{f\in S_{t,M}, \|f\|_\infty < \beta_n} \mathbf{E}(\, |f(X) - Y|^2 - \mathbf{E}(\, |r(X) - Y|^2 \; = \inf_{f\in S_{t,M} \|f\|_\infty < \beta_n} \int |f(x) - r(X)|^2 \mu(dx).$$

- On another side

$$\begin{aligned} &= \mathbf{E}(\, |r_n(x) - Y|^2 | \mathcal{D}_n) \; - \inf_{f\in S_{t,M} \|f\|_\infty < \beta_n} \mathbf{E}|f(X) - Y|^2 \\ &\leq \sup_{f\in S_{t,M}, \|f\|_\infty < \beta_n} \left\{ \mathbf{E}(\, |r_n(x) - Y|^2 | \mathcal{D}_n) - \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=0\}} \frac{|r_n(X_i) - Z_i|^2}{g(Z_i)} \right. \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=0\}} \frac{|r_n(X_i) - Z_i|^2}{g(Z_i)} - \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=0\}} \frac{|\bar{r}_n(X_i) - Z_i|^2}{g(Z_i)} \\ &\quad \left. + \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=0\}} \frac{|\bar{r}_n(X_i) - Z_i|^2}{g(Z_i)} - \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=0\}} \frac{|f(X_i) - Z_i|^2}{g(Z_i)} \right. \end{aligned}$$

$$+\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\delta_i=0\}}\frac{|f(X_i)-Z_i|^2}{g(Z_i)}-\mathbf{E}|f(X)-Y|^2\Bigg\}\leq\sum_{i=1}^{4}Q_{n,i}.$$

$Q_{n,i}$ are explained below.

The fact that $\bar{r}_n\in\mathbb{T}_{\beta_n}S_{t,M}$, $r_n\in\mathbb{T}_{\beta_n}S_{t,M}$, it's clear that

$$Q_{n,1}=\sup_{f\in S_{t,M},\|f\|_\infty<\beta_n}\mathbf{E}(\ |r_n(x)-Y|^2|\mathcal{D}_n)-\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\delta_i=0\}}\frac{|r_n(X_i)-Z_i|^2}{g(Z_i)}$$

$$\leq\sup_{f\in\mathbb{T}_{\beta_n}S_{t,M}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\delta_i=0\}}\frac{|f(X_i)-Z_i|^2}{g(Z_i)}-\mathbf{E}|f(X)-Y|^2\right|,$$

and

$$Q_{n,4}=\sup_{f\in S_{t,M},\|f\|_\infty<\beta_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\delta_i=0\}}\frac{|f(X_i)-Z_i|^2}{g(Z_i)}-\mathbf{E}|f(X)-Y|^2\right|.$$

Since $r_n(X_i)\leq\beta_n$, we obtain

$$\mathbf{1}_{\{\delta_i=0\}}|\bar{r}_n(X_i)-Z_i|\geq\mathbf{1}_{\{\delta_i=0\}}|r_n(X_i)-Z_i|,$$

which implies

$$Q_{n,2}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\delta_i=0\}}\frac{|r_n(X_i)-Z_i|^2}{g(Z_i)}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\delta_i=0\}}\frac{|\bar{r}_n(X_i)-Z_i|^2}{g(Z_i)}\leq 0.$$

According to the definition of $\tilde{r}_n$, it's obvious that

$$Q_{n,3}=\sup_{f\in S_{t,M},\|f\|_\infty<\beta_n}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\delta_i=0\}}\frac{|\bar{r}_n(X_i)-Z_i|^2}{g(Z_i)}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\delta_i=0\}}\frac{|f(X_i)-Z_i|^2}{g(Z_i)}\right\}\leq 0.$$

The inequality (3.6) is therefore demonstrated.

It remains to prove that the two terms of the second member of the equation tend towards zero almost surely when $n\to+\infty$. Since the equation (3.8) does not depend on censorship and using the hypotheses (3.4) and (3.5), we proceed in the same way as in Györfi et al. [17], p271 to obtain the result.

To show that

$$\lim_{n\to\infty}\sup_{f\in\mathbb{T}_{\beta_n}S_{t,M}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\delta_i=0\}}\frac{|f(X_i)-Z_i|^2}{g(Z_i)}-\mathbf{E}|f(X)-Y|^2\right|=0\ a.s. \tag{3.9}$$

Let Π_n be the family of any partition of $\mathbb{R}$ formed by at least $K_{\max}(n)+2M_{\max}(n)+2$ intervals; and let $\mathcal{P}$ be the set of all polynomials of degree less than or equal to M, or $S_{t,M}\subset\mathcal{P}\circ\Pi_n$, so it suffices to show 3.9 with the elements of the set $S_{t,M}$ generated

by the elements of $\mathcal{P} \circ \Pi_n$.

$\mathcal{P}$ linear function space of dimension $M+1$ thus $V_{\mathcal{P}^+} \leq M_{\max}(n)+2$ such that

$$\mathcal{P}^+ = \{\{(z,t) \in \mathbb{R} \times \mathbb{R} : t \leq g(z) : g \in \mathcal{P}\}$$

$$\subseteq \{(z,t) \in \mathbb{R} \times \mathbb{R} : at + g(z) \geq 0,\ g \in \mathcal{P}, a \in \mathbb{R}\}\}.$$

Refer to Györfi et al. [17] (Theorem 9.5 p 152) for more details. From the example (13.1) Györfi et al. [17] p 236, the number of partitions of Π_n satisfies

$$\Delta_n(\Pi_n) \leq \begin{pmatrix} n + K_{\max}(n) + 2M_{\max}(n) + 1 \\ n \end{pmatrix}$$

$$\leq (n + K_{\max}(n) + 2M_{\max}(n) + 1)^{K_{\max}(n)+2M_{\max}(n)+1}.$$

For this purpose, let's use the following notations

$$V = (X, Z, 1_{\{\delta=0\}}), V_1 = (X_1, Z_1, 1_{\{\delta_1=0\}}, ..., V_n = (X_n, Z_n, 1_{\{\delta_n=0\}}),$$

n random vectors i. i. d. with the same distribution as V. Let's pose

$$\mathcal{H}_n = \Big\{ h : \mathbb{R} \times [0,C] \times \{0,1\} \to \mathbb{R}^+ : \exists f \in \mathbb{T}_{\beta_n} S_{t,M}$$

such as

$$h(x, z, 1_{\{\delta=0\}}) = \frac{1_{\{\delta=0\}} |f(x) - z|^2}{g(z)},$$

for all

$$(x, z, 1_{\{\delta=0\}}) \in \mathbb{R} \times [0,C] \times \{0,1\} \Big\}.$$

The functions of $\mathcal{H}_n$ are non-negative and bounded by $\frac{\beta_n^2}{g(I)}$, and in the same way that Laroussi [26], we have

$$\begin{aligned} Eh(V) &= \mathbf{E}\left[\mathbf{E}\left(\frac{1_{\{\delta=0\}}|f(X) - Z|^2}{g(Z)} | X, Y\right)\right] \\ &= E(|f(X) - Y|^2). \end{aligned}$$

Under the assumptions H. In addition

$$\sup_{f \in T_{\beta_n} S_{t,M}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=0\}} \frac{|f(X_i) - Z_i|^2}{g(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right|$$

$$= \sup_{f\in\mathcal{H}_n} \left| \frac{1}{n}\sum_{i=1}^{n} h(V) - Eh(V) \right|.$$

For all $h_1, h_2 \in \mathcal{H}_n$, lets f_1, f_2 their corresponding functions in $T_{\beta_n}S_{t,M}$, then

$$\frac{1}{n}\sum_{i=1}^{n} |h_1(V_i) - h_2(V_i)|$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left| \mathbf{1}_{\{\delta_i=0\}} \frac{|f_1(X_i) - Z_i|^2}{g(Z_i)} - \mathbf{1}_{\{\delta_i=0\}} \frac{|f_2(X_i) - Z_i|^2}{g(Z_i)} \right|$$

$$\leq \frac{1}{g(I)} \frac{1}{n}\sum_{i=1}^{n} |(f_1(X_i) + f_2(X_i) - 2Z_i)(f_1(X_i) - f_2(X_i))|$$

$$\leq \frac{2\beta_n}{g(I)} \frac{1}{n}\sum_{i=1}^{n} |f_1(X_i) - f_2(X_i)|,$$

which implies

$$\mathcal{N}(\epsilon, \mathcal{H}_n, V_1^n) \leq \mathcal{N}\left(\frac{2\beta_n}{g(I)}, T_{\beta_n}S_{t,M}, X_1^n\right) \leq \mathcal{N}\left(\frac{2\beta_n}{g(I)}, T_{\beta_n}\mathcal{P}\circ\pi_n, X_1^n\right).$$

In the same way as the proof of Theorem 13.1 (p 240) of Györfi et al. [17] the relations 3.2 and 3.3 allow to apply the Borel-Cantelli lemma, to have

$$\sup_{f\in T_{\beta_n}S_{t,M}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=0\}} \frac{|f(X_i) - Z_i|^2}{g(Z_i)} - E|f(X) - Y|^2 \right| \underset{n\to\infty}{\to} 0 \ p.s.$$

For $0 \leq C < \beta_n$

$$\mathbf{P}\left[\sup_{f\in T_{\beta_n}\mathcal{P}\circ\Pi_n} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=0\}} \frac{|f(X_i) - Z_i|^2}{g(Z_i)} - E|f(X) - Y|^2 \right| > t \right]$$

such as $t = \frac{g(I)}{2\beta_n^2}$

$$\leq 8(n + K_{\max}(n) + 2M_{\max}(n) + 1)^{(K_{\max}(n)+2M_{\max}(n)+1)}$$

$$\times \left(\frac{333e\beta_n^2}{t}\right)^{2(M_{\max}(n)+2)(K_{\max}(n)+2M_{\max}(n)+2)} \exp\left(-\frac{nt^2}{2048\beta_n^4}\right).$$

And from 3.2 and 3.3, we get the assertion by applying the Borel-Cantelli lemma.

□

4. Simulation study

In this section, to present the performances of the studied estimator for a finite-size sample, we carry out a simulation study. We give a visual impression of the quality of estimation by plotting the correspondent true curve together with the curve of the estimator, based on a sample obtained from two theoretical models: Weibull distribution denoted by $\mathcal{W}$ and Bertholon distribution denoted by $\mathcal{B}$. We propose two schemes of censorship models (right and twice). To assess the efficiency of the choice of the B-spline class of functions, the modelization of two curves is proposed, one linear and the other non-linear. This study is achieved through three sample sizes $(n = 100, 300, 500)$.

4.1. **Right censored case.** We will specifically study the performance of the estimator under two different structures. We will also consider the right-censorship model defined by $Z = \min(Y, R)$.

4.1.1. *Linear Weibull model.* The first model is defined by $Y = 2X + 1 + \varepsilon$, where $\varepsilon \simeq \mathcal{N}(0, 0.5)$, X is distributed as $\mathcal{W}(0.5, 2)$ and R is distributed as $\mathcal{W}(3, 3.5)$. For this model, the rate of right censoring is 16%. Figure 1 shows the obtained graphs for $r(x) = 2x + 1$ with 16% right censoring rate and Weibull model for $n = 100, 300, 500$.

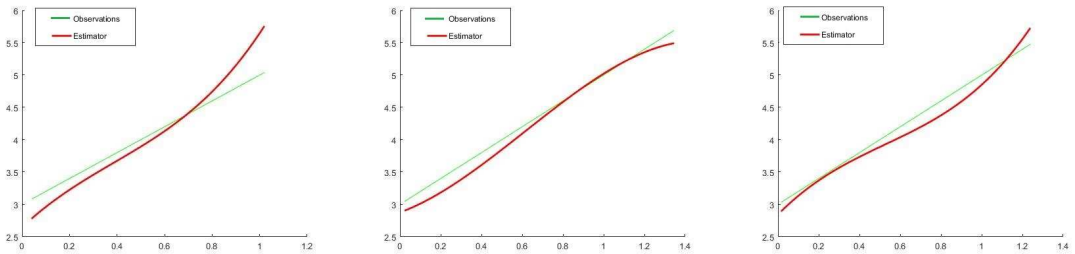

Figure 1. $r(x) = 2x + 1$ with 16% right censoring rate and Weibull model for $n = \{100, 300, 500\}$.

4.1.2. *Non linear Weibull model.* In the second model, $Y = \cos(2X + 3) + 4 + \varepsilon$. $\varepsilon \simeq \mathcal{N}(0, 0.5)$, X is distributed as $\mathcal{W}(0.5, 2)$ and R is distributed as $\mathcal{W}(5, 5.5)$. For

this model, the rate of right censoring is 16%. Figure 2 shows the obtained graphs for $r(x) = cos(2x + 3)$ with 16% right censoring rate and Weibull model for $n = 100, 300, 500$.

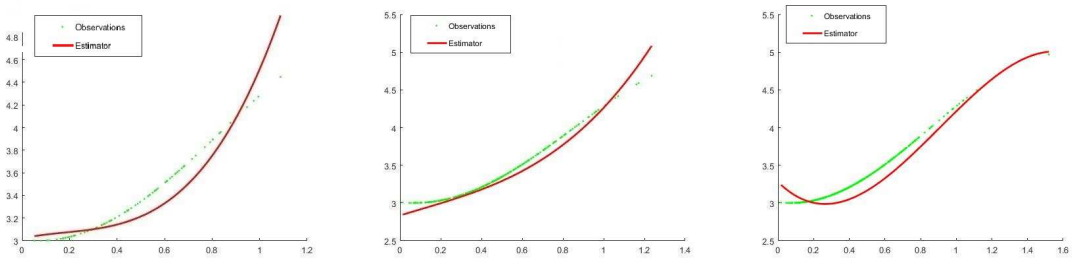

FIGURE 2. $r(x) = \cos(2X + 3) + 4$ with 16% right censoring rate and Weibull model for $n = \{100, 300, 500\}$.

4.1.3. *Linear Bertholon model.* The third visualization is done for $Y = 2X + 1 + \varepsilon$. Where $\varepsilon \simeq \mathcal{N}(0, 0.5)$, X is distributed as $\mathcal{B}(8.33, 1, 2)$ and R is distributed as $\mathcal{B}(10, 7, 7)$. For this model, the rate of right censoring is 16%. The results for $r(x) = 2x + 1$ with 16% right censoring rate and Bertholon model for $n = 100, 300, 500$ are shown in Figure 3.

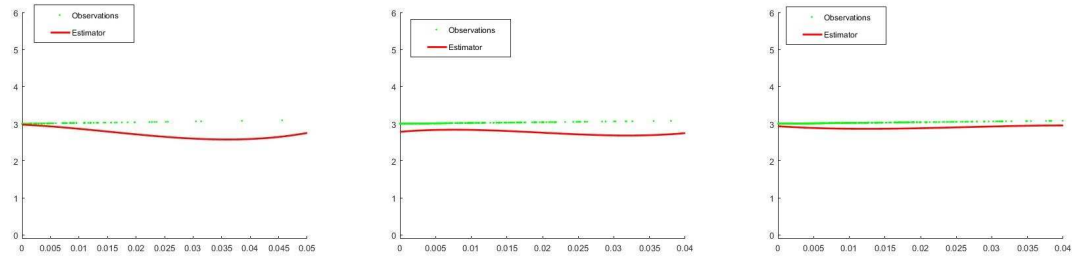

FIGURE 3. $r(x) = 2x+1$ with 16% right censoring rate and Bertholon model for $n = \{100, 300, 500\}$.

4.1.4. *Non linear Bertholon model.* For Y distributed as $Y = \cos(2X+3)+4+\varepsilon$. $\varepsilon \simeq \mathcal{N}(0, 0.5)$, X is distributed as $\mathcal{B}(10, 4, 4)$ and R is distributed as $\mathcal{B}(20, 10, 8)$. For this model, the rate of right censoring is 17%. The results for $r(x) = cos(2x+1)$ with 17%

right censoring rate and Bertholon model for $n = 100, 300, 500$ are shown in Figure 4.

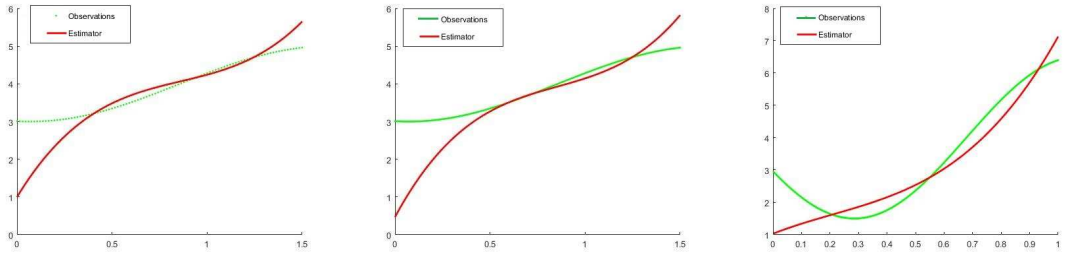

FIGURE 4. $r(x) = \cos(2X + 3) + 4$ with 17% right censoring rate and Bertholon model for $n = \{100, 300, 500\}$.

4.2. **Mixed censored case.** In this case, we will study the performance of the estimator under the same structures that above but with the twice-censorship model defined by $Z = \max(\min(Y, R), L)$. One consider

4.2.1. *Linear Weibull model.* The first model, is defined by $Y = 2X + 1 + \varepsilon$, where $\varepsilon \simeq \mathcal{N}(0, 0.5)$, X take values $\frac{1:n}{n}$, R is distributed as $\mathcal{W}(3.5, 4)$ and the left censoring random variable $L \simeq \mathcal{W}(0.001, 0.1)$. For $r(x) = 2x + 1$ with the rate of right (resp. left) censoring is 16% (resp. 10%) and Weibull model for $n = 100, 300, 500$, we obtained the following plots (see Figures 5).

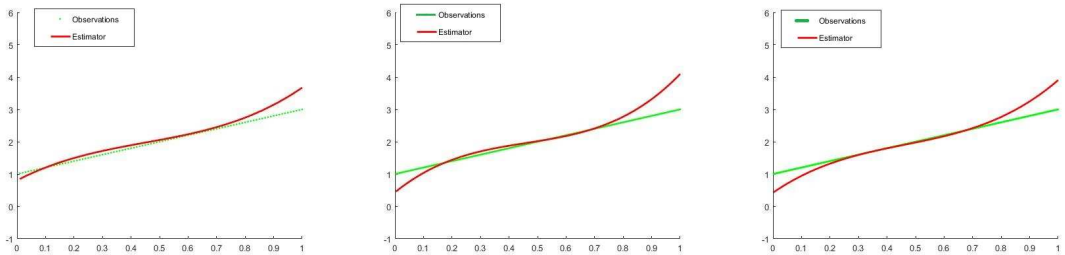

FIGURE 5. $r(x) = 2x + 1$ with the rate of right (resp. left) censoring is 16% (resp. 10%) and Weibull model for $n = \{100, 300, 500\}$.

4.2.2. *Non linear Weibull model.* In this case, Y is done by $Y = 5\cos(2X+1)^2 + 1.5 + \varepsilon$. Where, ε, L and X have the same distribution as the linear model and $R \simeq \mathcal{W}(6, 6.5)$. The modification in the parameters values allows to have an overall censorship rate of around 27%. Results are shown in Fgure 6.

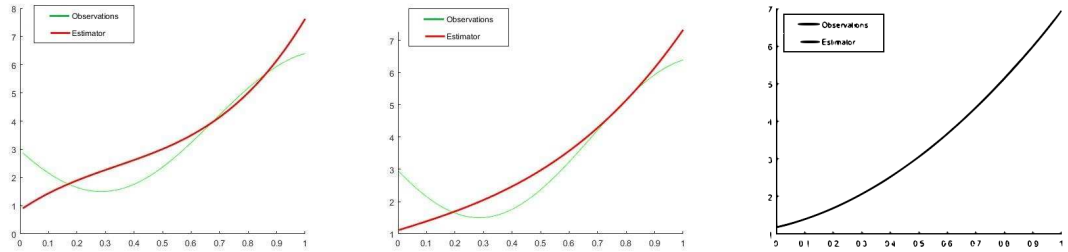

FIGURE 6. $r(x) = 5\cos(2X+1)^2 + 1.5$ with the rate of right (resp. left) censoring is 17% (resp. 10%) and Weibull model for $n = \{100, 300, 500\}$.

4.2.3. *Linear Bertholon model.* The third model, is realized for a model defined by $Y = 2X + 1 + \varepsilon$, where $\varepsilon \simeq \mathcal{N}(0, 0.5)$, $X \simeq \mathcal{B}(8, 1.25, 2)$, R is distributed as $\mathcal{B}(23, 7, 5.5)$ and the left-censoring random variable $L \simeq \mathcal{B}(3.5, 1.5, 3)$. For this model, the right-censorship rate is 17 % and the left-censorship rate is 10 %. Results for $r(x) = 2x + 1$, as $n = 100, 300, 500$, are shown in Figure 7.

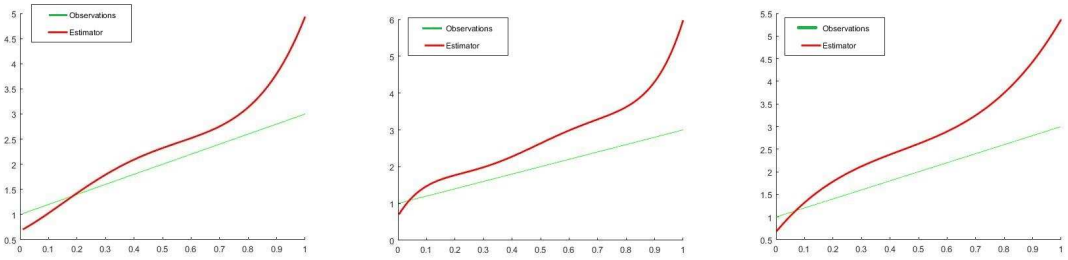

FIGURE 7. $r(x) = 2x + 1$ with the rate of right (resp. left) censoring is 16% (resp. 10%) and Bertholon model for $n = \{100, 300, 500\}$.

4.2.4. *Non linear Bertholon model.* The last model is obtained for $Y = 5\cos(2X + 1)^2 + 1.5 + \varepsilon$. Where, $\varepsilon \simeq \mathcal{N}(0, 0.5)$, $X \simeq \mathcal{B}(9, 1, 1)$, R is distributed as $\mathcal{B}(26, 2.5, 7)$

and the left-censoring random variable $L \simeq \mathcal{B}(3.5, 1.5, 3)$. For this model, the right-censorship rate is 17 % and the left-censorship rate is 10 %. Figure 8 shows the obtained plots for $r(x)$ as $n = 100, 300, 500$.

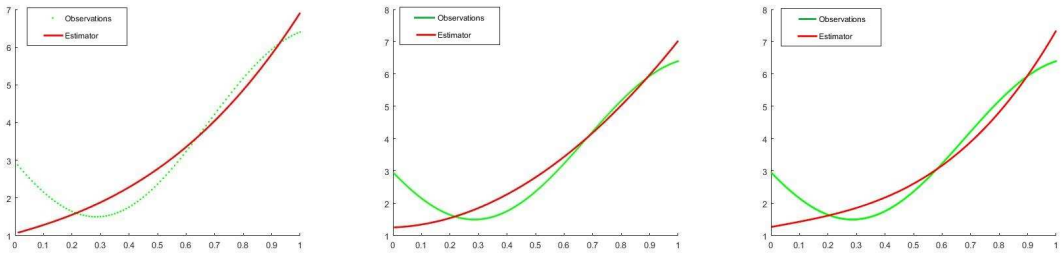

FIGURE 8. $r(x) = 5\cos(2X + 1)^2 + 1.5$ with the rate of right (resp. left) censoring is 17% (resp. 10%) and Bertholon model for $n = \{100, 300, 500\}$.

By inspecting the previous figures, as the sample sizes increase, the quality of fit increases for all the considered models, as expected. However, it can be clearly seen that the right censorship estimators are better than those of the mixed censorship models under both laws (Weibull and Bertholon). This is not unexpected since the censorship is well known for its influence on proximity of the estimators. To better interpret the simulation results, we calculate the mean square error (MSE) between the estimators and the real observations. Note here that, we have used the same complete sample for all the censoring schemes to observe the effect of censoring on the MSE. Table 1 summarizes the obtained results.

The results in Table 1 indicate that the root mean square error became smaller and smaller than sample size increases. Moreover, the quality of fit deteriorates under high levels of censorship in terms of higher MSE. Particularly, close inspection reveals that right-hand censorship has given more satisfactory results than double variance censorship. Furthermore, Bertholon and Weibull linear models perform better than the nonlinear ones for all sample sizes and for almost all the censoring schemes (see minimum variance values in bold in Table 1) except for the case of Bertholon non linear model under 26% twice censorship level where we notice the opposite.

TABLE 1. The mean square error MSE

		Right censorship		Twice censorship	
		Weibull	Bertholon	Weibull	Bertholon
Models	% censoring / Size	R≃ 16%	R≃ 16%	L≃ 10% R≃ 16%	L≃ 10% R ≃16%
Linear	100	**0.0319**	**0.0519**	0.095	0.7134
	300	**0.0302**	**0.049**	0.0941	0.4777
	500	**0.0191**	**0.0481**	0.0762	0.4269
Models	% censoring / Size	R≃ 16%	R≃ 17%	L ≃10% R≃17%	L ≃10% R≃ 17%
Non linear	100	0.0951	0.1561	0.4493	0.2862
	300	0.0941	0.1407	0.3529	0.2582
	500	0.0762	0.1398	0.3497	0.249

Acknowledgement

We would like to thank the editor and the referees.

REFERENCES

[1] G. G. Agarwal. Splines in statistics. *Bulletin of the Allahabad Mathematical Society*, **4**, (1989), 1-55.

[2] G. G. Agarwal and W. J. Studden. Asymptotic integrated mean square error using least squares and bias minimizing splines. *Annals of Statistics,* **8**, (1980), 1307-1325.

[3] R. Beran. *Non-parametric regression with randomly censored survival data.* Technical Report, University of California, Berkeley, (1981).

[4] M. Boukeloua. Rates of mean square convergence of density and failure rate estimators under twice censoring. *Statist. Probab. Lett.*, **106**, (2015), 121-128.

[5] M. Boukeloua and F. Messaci. Asymptotic normality of kernel estimators based upon incomplete data. *Journal of Nonparametric Statistics*, **28(3)**, (2016), 469-486.

[6] S. Casanova, E. Leconte. Nonparametric Model-Based Estimator for the Cumulative Distribution Function of Right-Censored Variable in a Finite Population. *Journal of Surveys : Statistics and Methodology* **3**, (2015), 317338.

[7] D. R. Cox. Regression models and life tables (with discussion). *J R Statist Soc* **B 34**, (1972), 187-220.

[8] D. M. Dabrowska. Nonparametric regression with censored data. *Scandinavian J. Statistics*, **14**, (1987),181-197.

[9] D. M. Dabrowska. Uniform consistency of the kernel conditional Kaplan- Meier estimate. *Annals of Statistics*, **17**, (1989), 1157-1167.

[10] N. R. Draper and H. Smith. *Applied Regression Analysis, 2nd ed. Wiley, New York, (1981).*

[11] R. Douas, I. Laroussi, S. Kharfouchi, Incomplete Least Squared Regression Function Estimator Based on Wavelets, *Journal of siberian federal university. mathematics and physics*, **16(2)**, (2023)-204-2015.

[12] P. P. B. Eggermont and V. N. LaRiccia. *Maximum Penalized Likelihood Estimation. Vol. I: Density Estimation*. Springer-Verlag, New York, (2001).

[13] R. L. Eubank. *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, New York, (1999).

[14] J. Fan and I. Gijebels. *Local Polynomial Modeling and its Applications*. Chapman and Hall, London, (1995).

[15] C. F. Gauss. *Book on celestial mechanics*, (1809)

[16] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman and Hall, London, (1994).

[17] L. Györfi, M. Kohler, A. Krzyak, H. Walk. *A Distribution Free theory of Nonparametric Regression*. Springer-Verlag New York, Inc. (2002).

[18] T. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, U. K. (1990).

[19] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, **100**, (1992), 78-150.

[20] E. L. Kaplan, P. Meier. Nomparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.* **53**, (1958), 457-481.

[21] K. Kebabi, I. Laroussi, F. Messaci . Least squares estimators of the regression function with twice censored data, *Statist. Probab. Lett.***81**, (2011), 1588-1593.

[22] A. Kitouni, M. Boukeloua and F. Messaci. Rate of strong consistency for nonparametric estimators based on twice censored data. *Statistics and Probability Letters*, **96**, (2015), 255-261.

[23] M. Kohler. Universally Consistent Regression Function Estimation Using Hierarchical B-Splines. *Journal of Multivariate Analysis*, **68**, (1999), 138-164.

[24] M. Kohler and A. Krzyzk. Nonparametric regression estimation using penalized least squares. *IEEE Transaction on Information Theory*, 47, (2001), 3054-3058.

[25] M. Kohler, K. Màthé, M. Pintér. Prediction from randomly right censored data, *J. Multivariate Anal.* **80**, (2002), 73-100.

[26] I. Laroussi A generalised censored least squares and smoothing spline estimators of regression function, *Int. J. Mathematics in Operational Research,* **Vol. 20,** No. 4, (2021), 506520.

[27] A. M. Legendre. *Nouvelles Mthodes pour la determination des orbites des cometes.* Firmim Didot, Libraire pour les Mathmatiques, (1805).

[28] E. Mammen and S. Van de Geer. Locally adaptive regression splines. *Annals of Statistics*, **25**, (1997), 387-413.

[29] Ni P. A. M. Mariati, I N. Budiantara, Vita Ratnasari. Locally adaptive regression splines. *Advances in Social Science, Education and Humanities Research*, **528**, (2020), 309-314.

[30] F. Messaci, Local averaging estimates of the regression function with twice censored data, *Statist. Probab. Lett.*, **80**, (2010) , 1508-1511.

[31] A. Montanari, Basil N. Saeed. Universality of empirical risk minimization. *Proceedings of Thirty Fifth Conference on Learning Theory, PMLR* **178**, (2022), 4310-4312.

[32] V. Patilea, J. M. Rolin, Product-limit estimators of the survival function with twice censored data, *Ann. Statist.* **34**, No 2, (2006) 925-938.

[33] Pavel Čižek and Serhan Sadikoglu . Robust nonparametric regression. *WIREs Computational Statistics*,(2019), 1-16.

[34] R. C. Rao. *Linear Statistical Inference and Its Applications*. Wiley, New York, 2nd edition, (1973).

[35] G. A. F. Seber. *Linear Regression Analysis*. Wiley, New York, (1977).

[36] J. W. Tukey. Nonparametric estimation II. Statistically equivalent blocks and tolerance regions. *Annals of Mathematical Statistics*, **18**, (1947), 529-539.

[37] B. W. Turnbull. Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. *Journal of the American Statistical Association.* **69**, No. 345, (1974), 169-173

[38] S. Van de Geer and M. Wegkamp. Consistency for the least squares estimator in nonparametric regression. *Annals of Statistics*, **24**, (1996), 2513-2523.

[39] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **16**, (1971), 264-280.

[40] V. N. Vapnik. *Estimation of Dependencies Based on Empirical Data, (1982)*. Springer-Verlag, New York.

[41] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, (1998).

[42] G. Wahba . *Spline Models for Observational Data*. SIAM, Philadelphia, PA, (1990).

[43] E. J. Wegman and I. W. Wright. Splines in statistics. *Journal of the American Statistical Association*, **78**, (1983), 351-365.

[44] E. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, **41**, (1923), 63-75.

[45] L. X. Zhu. A note on the consistent estimator of nonparametric regression constructed by splines. *Computers and Mathematics with Applications*, **24**, (1992), 65-70.

(1) Laboratory of Mathematics and Sciences of the Decision (LAMASD), Freres Mentouri University, 25017 Constantine, Algeria

Email address: 33laroussi@gmail.com