Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 15** : 1–12.
ISSN-0037-8712
doi:10.5269/bspm.81650

Forecasting in Nonparametric Regression Models with Double Censoring

Ilhem Laroussi and Ranya Boustila

ABSTRACT: In this work, we study the nonparametric estimation of the regression function using the least squares method in the presence of double censoring. We construct an estimator by replacing unknown survival functions with self-consistent estimators in the spirit of Turnbull (1974). We prove that this estimator is strongly consistent, converging almost surely to the optimal regression function. Finally, we illustrate our theoretical findings with a simulation study under linear and nonlinear regression models.

Keywords: Regression function, least squares, double censorship.

Contents

1. Introduction

Nonparametric regression plays a central role in modern mathematical statistics due to its flexibility and its ability to model complex relationships without imposing restrictive structural assumptions. Given a random pair (X, Y), the objective is to estimate the regression function

$$\mathcal{R}(x) = \mathbb{E}(Y \mid X = x),$$

which characterizes the conditional mean of the response variable given the covariate. Classical nonparametric regression methods include kernel estimators, such as those proposed by Nadaraya and Watson [8,11] in 1964, nearest-neighbor procedures introduced by Stone [9] in 1977, local polynomial methods developed by Fan and Gijbels [3] in 1996, and wavelet-based techniques proposed by Donoho and Johnstone [1] in 1994.

In many applied fields, such as survival analysis, reliability theory, economics, and biomedical studies, the response variable is often incompletely observed due to censoring. Censoring occurs when the exact value of a variable is not available and only partial information is recorded. While regression models under right censoring have been extensively studied, considerably fewer results are available for the more general and challenging cases of double or general censoring.

The statistical analysis of doubly censored data was initiated by Turnbull [10] in 1974, who introduced a self-consistent estimator of the underlying distribution function. This estimator has since become a fundamental tool in inference problems involving interval and double censoring and has motivated numerous extensions in regression settings.

2020 *Mathematics Subject Classification:* 35B40, 35L70.
Submitted January 29, 2026. Published April 30, 2026.

In this context, several contributions have investigated least squares–based nonparametric regression estimators under censoring. Kebabi, Laroussi, and Messaci [5] studied least squares estimation of the regression function in the presence of twice-censored data and established its asymptotic properties. Laroussi [6] proposed a generalized censored least squares framework combined with smoothing spline techniques, providing a unified approach to regression estimation under censoring. More recently, wavelet-based incomplete least squares estimators were introduced by Douas, Laroussi, and Kharfouchi [2], highlighting the benefits of multiresolution analysis in censored regression models. Spline-based methods under general censoring schemes were further developed in [7], where a B-spline estimator of the regression function was investigated.

Beyond classical nonparametric approaches, modern machine learning tools have also been explored in censored regression settings. In particular, Idiou *et al.* [4] studied the combination of neural networks and least squares estimation, opening new perspectives for flexible regression modeling under incomplete data.

Motivated by these works, the present study focuses on nonparametric regression estimation in the presence of double censoring. We propose an estimation procedure that builds upon self-consistent techniques and nonparametric smoothing methods. The theoretical properties of the proposed estimator are established, and its finite-sample performance is illustrated through simulation experiments.

2. Framework of the Study

Let Y denote the response variable of interest, which is subject to both left and right censoring. Let L and R denote the censoring variables, with observed data consisting of triplets (X, L, R) along with censoring indicators. Formally, we observe

$$Z = \max(L, \min(Y, R)),$$

with indicators specifying whether Y is left-censored, right-censored, or observed exactly. Let Y be a real-valued random variable of interest, whose exact value is not always observed. Instead, two random variables L and R are observed such that

$$L \leq Y \leq R,$$

where L and R represent the lower and upper censoring bounds, respectively.

The standard assumptions on these variables are as follows:

- **Order of bounds:** almost surely, $L \leq R$ and $\mathbb{P}(L < R) > 0$.
- **Non-informative censoring:** the variable of interest Y is independent of the censoring mechanism, i.e.,
$$Y \perp\!\!\!\perp (L, R) \mid X.$$
- **Non-degeneracy:** the probability that Y is effectively censored is strictly positive,
$$\mathbb{P}(L < Y < R) > 0.$$
- **Common support:** the support of Y is included in the support of the observed intervals,
$$\forall y \in \text{supp}(Y), \quad \mathbb{P}(L \leq y \leq R) > 0.$$
Let $T_R = \inf\{t \in supp(R) : \forall y \in \text{supp}(Y), \quad \mathbb{P}(y \leq R) > 0\}$.
- **Regularity:** the distributions of Y, L, and R are continuous, with
$$\mathbb{P}(Y = L) = \mathbb{P}(Y = R) = 0,$$
ensuring the necessary regularity conditions for asymptotic analysis.

These assumptions guarantee the identifiability of the double censoring model as well as the validity of estimators based on the observed intervals $[L, R]$.

3. Least Squares Estimation

In the classical case without censoring, the regression function $\mathcal{R}$ minimizes the quadratic risk

$$\mathcal{R}(f) = \mathbb{E}\big[(Y - f(X))^2\big],$$

over a suitable class of candidate functions. The minimizer is the conditional expectation $\mathcal{R}(x) = \mathbb{E}(Y|X = x)$ from the data $\mathcal{O}_n = \{X_i, Z_i, A_i \ ; \ \ 1 \leq i \leq n\}$ which is independent and of the same law as (X, Z, A) as

$$A = \begin{cases} 0 & if \quad L < Y \leq R \\ 1 & if \quad Y > R \\ 2 & if \quad Y \leq L \end{cases} .$$

.

When Y is doubly censored, direct minimization is not feasible since Y is not always observed. We thus replace the unknown survival distributions S_Y, S_L, S_R by their self-consistent estimators $\hat{S}_Y, \hat{S}_L, \hat{S}_R$, as proposed by Turnbull (1974). Given by

$$\hat{S}_Y(t) = Q^{(n)}(t) - \int_{\{u \leq t\}} \frac{\hat{S}_Y(t)}{\hat{S}_Y(u)} dQ_1^{(n)}(u) + \int_{\{t < u\}} \frac{1 - \hat{S}_Y(t)}{1 - \hat{S}_Y(u)} dQ_2^{(n)}(u).$$

$$\hat{S}_R(t) = 1 + \int_{\{u \leq t\}} \frac{dQ_1^{(n)}(u)}{\hat{S}_Y(u)}, \ t < B_n.$$

$$\hat{S}_L(t) = - \int_{\{t < u\}} \frac{dQ_2^{(n)}}{1 - \hat{S}_Y(u)}, \ t \geq A_n.$$

These estimators converge uniformly and almost surely to the survival functions S_R and S_L, in this case the estimator will be as follows This leads to an empirical least squares criterion which is obtained by minimizing the empirical risk $\mathcal{L}_2$

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i \left(Z_i - f(X_i)\right)^2,$$

where

$$\hat{w_i} = \frac{\mathbb{1}_{(A_i = 0)}}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)}$$

are weights derived from the censoring mechanism. The regression estimator is then defined as

$$\hat{r}_n = \arg \min_{f \in \mathcal{F}_n} \mathcal{R}_n(f).$$

Where $\mathcal{F}_n$ is the class of functions.

4. Consistency of the Estimator

We establish that the proposed regression estimator converges almost surely to the optimal regression function. Let $\mathcal{R}^*$ given above, denote the minimizer of the true quadratic risk $\mathcal{R}$. Then, under regularity assumptions on the censoring distributions and the complexity of $\mathcal{F}_n$, $\hat{\mathcal{R}}_n(x) \xrightarrow[n \to \infty]{a.s.} \mathcal{R}^*(x)$.
Assume that the following conditions are checked
$\bullet H_1 : \ A_1 \neq 2$ so that $S_R(t) - S_L(t) > 0, \forall t \geq 0$. This assumption is verified by choosing a sample with the first data non censoring.
$\bullet H_2 :$ If $T_R < \infty$ then $M_n = \max\{Z_1, ..., Z_n\} \xrightarrow[n \to \infty]{\text{a.s}} T_R$.
$\bullet H_3 : \exists \ b \in \mathbb{R}_*^+ : b = \inf_{t \leq T_R}(S_R(t) - S_L(t))$.

Then the proof relies on

- Uniform convergence of the self-consistent estimators $\hat{S}_L, \hat{S}_R$ to their true counterparts are given by

$$\forall t \in \mathbb{R}_+ \quad \hat{S}_R(t) \xrightarrow[n\to\infty]{\text{a.s}} S_R(t) \tag{4.1}$$

$$\forall t \in \mathbb{R}_+ \quad \hat{S}_L(t) \xrightarrow[n\to\infty]{\text{a.s}} S_L(t). \tag{4.2}$$

- Approximation properties of the function class $\mathcal{F}_n$.
- An application of empirical process techniques to show almost sure convergence.

Since Y is bounded ($0 \leq Y \leq T_R$), the estimator $\hat{\mathcal{R}}_n(x)$ must be truncated by

$$\mathcal{R}^* * n(x) = \mathbb{T} * [0, T_R](\hat{\mathcal{R}}_n(x)). \tag{4.3}$$

But in the truth, we will work with the second version of tronque estimator which is defined by

$$\mathcal{R} * n(x) = \mathbb{T} * [0, M_n](\hat{\mathcal{R}}_n(x)). \tag{4.4}$$

Such as, For any real x and any real strictly positive t, $\mathbb{T}_{[0,t]}(x)$ is called the truncation operator which is defined by

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

and checks the following properties

- $\forall b > 0$, $\forall a > 0$ and $\forall x \in \mathbb{R}$, $|\mathbb{T} * [0, b](x) - \mathbb{T} * [0, a](x)| \leq |b - a|$.

In the following we will present the consistency of our estimator by a theorem, for this we will define the following sets

$$\mathcal{B}_n^* \mathcal{F}_n = \{f \in \mathcal{F}_n : 0 \leq f(x) \leq T_R \, (x \in \mathbb{R})\}.$$

$$\mathcal{F}_n^* \mathcal{F}_n = \left\{g : \mathbb{R} \to \mathbb{R}^+ / \exists f \in \mathcal{F}_n, \forall x \in \mathbb{R} : g(x) = \mathbb{T}_{[0,T_R]}(f(x))\right\}.$$

$$\mathcal{F}_n^+ = \{\{(x, y) \in \mathbb{R} \times \mathbb{R} : f(x) \geq y\}, f \in \mathcal{F}_n\}$$

4.1. Result and proof

The convergence of the estimator $\mathcal{R}_n(x)$ is equivalent to the convergence of the estimator $\mathcal{R}_n^*(x)$

Theorem 4.1 *Under the assumptions* (H_1, H_2, H_3),

$$\int_{\mathbb{R}} |\mathcal{R}_n(x) - \mathcal{R}(x)|^2 \, \mu(dx) \underset{n\to\infty}{\to} 0, a.s.$$

for a family choice $\mathcal{F}_n$ *of functions* $f : \mathbb{R} \to \mathbb{R}$ *satisfying*

$$\frac{\mathcal{V}_{F_n^+}}{n} \underset{n\to\infty}{\to} 0. \tag{4.5}$$

Where $\mathcal{V}_{F_n+}$ *denotes the V.C (Vapnik-Chervonenkis) dimension of all function graphs in* $\mathcal{F}_n$ *and*

$$\inf_{f \in B_n^* \mathcal{F}_n} \sup_{x \in [-A, A]} |f(x) - g(x)| \underset{n\to\infty}{\to} 0, \tag{4.6}$$

for all $A \in \mathbb{R}^+$ *and for any continuous function* $g : \mathbb{R} \to \mathbb{R}$, *which cancels out of* $[-A, A]$ *and which is bounded by* T_R.

Proof: The proof of this theorem is divided into three steps

- The first step: It is shown that

$$\int_{\mathbb{R}} |\mathcal{R}_n^*(x) - \mathcal{R}(x)|^2 \mu(dx)$$
$$\leq 2 \sup_{f \in \mathcal{F}_n^* \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(A_i=0)} \frac{|f(X_i) - Z_i|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)} - \mathbf{E}(|f(X) - Y|^2) \right|$$
$$+ \inf_{f \in \mathcal{B}_n^* \mathcal{F}_n} \int_{\mathbb{R}} |f(x) - r(x)|^2 \mu(dx).$$

We have

$$\begin{aligned} & \int_{\mathbb{R}} |\mathcal{R}_n^*(x) - \mathcal{R}(x)|^2 \mu(dx) \\ = & \left\{ \mathbf{E}[|\mathcal{R}_n^*(X) - Y|^2 \mid \mathcal{O}_n] - \inf_{f \in \mathcal{B}_n^* \mathcal{F}_n} \mathbf{E}[|f(X) - Y|^2] \right\} \\ + & \left\{ \inf_{f \in \mathcal{B}_n^* \mathcal{F}_n} \mathbf{E}[|f(X) - Y|^2] - \mathbf{E}[|\mathcal{R}(X) - Y|^2] \right\}. \end{aligned}$$

1. The regression function checks that

$$\inf_{f \in \mathcal{B}_n^* \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 - \mathbf{E}|\mathcal{R}(X) - Y|^2 = \inf_{f \in \mathcal{B}_n^* \mathcal{F}_n} \int_{\mathbb{R}} |f(x) - \mathcal{R}(x)|^2 \mu(dx). \tag{4.7}$$

2. On the other hand

$$\mathbf{E}\left(|\mathcal{R}_n^*(X) - Y|^2 \mid \mathcal{O}_n \right) - \inf_{f \in \mathcal{B}_n^* \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2$$
$$= \sup_{f \in \mathcal{B}_n^* \mathcal{F}_n} \left\{ \mathbf{E}\left(|\mathcal{R}_n^*(X) - Y|^2 \mid \mathcal{O}_n \right) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{A_i=0\}} \frac{|\mathcal{R}_n^*(X_i) - Z_i|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)} \right.$$
$$+ \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{A_i=0\}} \frac{|\mathcal{R}_n^*(X_i) - Z_i|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{A_i=0\}} \frac{\left|\hat{\mathcal{R}}_n(X_i) - Z_i\right|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)}$$
$$+ \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{A_i=0\}} \frac{\left|\hat{\mathcal{R}}_n(X_i) - Z_i\right|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)}$$
$$\left. + \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right\}.$$

We treat each term independently to get our result.

- We have $f \in \mathcal{B}_n^* \mathcal{F}_n \subseteq \mathcal{F}_n$, therefore

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{A_i=0\}} \frac{\left|\hat{\mathcal{R}}_n(X_i) - Z_i\right|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)} \leq 0.$$

- Also, for all $i = 1, ..., n$

$$\begin{aligned} [\hat{\mathcal{R}}_n(X_i) - Z_i] & = [\hat{\mathcal{R}}_n(X_i) - \mathcal{R}_n^*(X_i)] + [\mathcal{R}_n^*(X_i) - Z_i] \geq [\mathcal{R}_n^*(X_i) - Z_i] \\ & \Longrightarrow [\hat{\mathcal{R}}_n(X_i) - Z_i] \geq [\mathcal{R}_n^*(X_i) - Z_i]. \end{aligned}$$

Then

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{A_i=0\}} \frac{|\mathcal{R}_n^*(X_i) - Z_i|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{A_i=0\}} \frac{\left|\hat{\mathcal{R}}_n(X_i) - Z_i\right|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)} \leq 0.$$

Therefore

$$\mathbf{E}\left(\left|\mathcal{R}_n^*(X)-Y\right|^2 \mid \mathcal{O}_n\right) - \inf_{f\in\mathcal{B}_n^*\mathcal{F}_n} \mathbf{E}|f(X)-Y|^2$$

$$\leq \sup_{f\in\mathcal{B}_n^*\mathcal{F}_n}\left\{\mathbf{E}\left(\left|\mathcal{R}_n^*(X)-Y\right|^2 \mid \mathcal{O}_n\right) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{A_i=0\}}\frac{\left|\mathcal{R}_n^*\left(X_i\right)-Z_i\right|^2}{\hat{S}_R\left(Z_i\right)-\hat{S}_L\left(Z_i\right)}\right.$$

$$\left.+\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{A_i=0\}}\frac{\left|f\left(X_i\right)-Z_i\right|^2}{\hat{S}_R\left(Z_i\right)-\hat{S}_L\left(Z_i\right)} - \mathbf{E}|f(X)-Y|^2\right\}.$$

- Of the fact that $f\in\mathcal{B}_n^*\mathcal{F}_n$, $\mathcal{R}_n^*\in\mathcal{B}_n^*\mathcal{F}_n$ and $\mathcal{B}_n^*\mathcal{F}_n\subseteq\mathcal{F}_n^*\mathcal{F}_n$, it's clear that

$$\sup_{f\in\mathcal{B}_n^*\mathcal{F}_n}\left\{\mathbf{E}\left(\left|\mathcal{R}_n^*(X)-Y\right|^2 \mid \mathcal{O}_n\right) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{A_i=0\}}\frac{\left|\mathcal{R}_n^*\left(X_i\right)-Z_i\right|^2}{\hat{S}_R\left(Z_i\right)-\hat{S}_L\left(Z_i\right)}\right\}$$

$$\leq \sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{A_i=0\}}\frac{\left|f\left(X_i\right)-Z_i\right|^2}{\hat{S}_R\left(Z_i\right)-\hat{S}_L\left(Z_i\right)} - \mathbf{E}|f(X)-Y|^2\right|\right\},$$

then

$$\mathbf{E}\left(\left|\mathcal{R}_n^*(X)-Y\right|^2 \mid \mathcal{O}_n\right) - \inf_{f\in\mathcal{B}_n^*\mathcal{F}_n} \mathbf{E}|f(X)-Y|^2$$

$$\leq 2\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{A_i=0\}}\frac{\left|f\left(X_i\right)-Z_i\right|^2}{\hat{S}_R\left(Z_i\right)-\hat{S}_L\left(Z_i\right)} - \mathbf{E}|f(X)-Y|^2\right|\right\}. \tag{4.8}$$

So according (4.7) et (4.8), we have

$$\int_{\mathbb{R}}|\mathcal{R}_n^*(x)-r(x)|^2\mu(dx)$$

$$\leq 2\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(Ai=0)}\frac{|f(X_i)-Z_i|^2}{\hat{S}_R(Z_i)-\hat{S}_L(Z_i)} - \mathbf{E}(|f(X)-Y|^2)\right|$$

$$+\inf_{f\in\mathcal{B}_n^*\mathcal{F}_n}\int_{\mathbb{R}}|f(x)-\mathcal{R}(x)|^2\mu(dx).$$

- The second stage : shows that

$$\inf_{f\in\mathcal{B}_n^*\mathcal{F}_n}\int_{\mathbb{R}}|f(x)-\mathcal{R}(x)|^2\mu(dx)\xrightarrow[n\to\infty]{\text{a.s}}0.$$

Either $\mathrm{C}^0(\mathbb{R})$ all continuous functions with compact support. As $\mathrm{C}^0(\mathbb{R})$ is dense in $\mathcal{L}_2$, for all $\varepsilon>0$ there is a function h of $\mathrm{C}^0(\mathbb{R})$ verifying

$$\int_{\mathbb{R}}|h(x)-\mathcal{R}(x)|^2\mu(dx)\leq\varepsilon \text{ p.s.}.$$

Let's choose $A>0$ such as $h(x)=0$, if $x\notin[-A,A]$ and

$$\mu\left([-A,A]^{\mathrm{c}}\right)\leq\frac{\varepsilon}{T_R^2}.$$

Define $\bar{h}(x)=T_{[0,T_R]}(h(x)),\ x\in\mathbb{R}$, then $\bar{h}(x)\in\mathrm{C}^0(\mathbb{R}),0\leq\bar{h}(x)\leq T_R$ and $0\leq\mathcal{R}(x)\leq T_R$ involve

$$\int_{\mathbb{R}}|\bar{h}(x)-\mathcal{R}(x)|^2\mu(dx)\leq\int_{\mathbb{R}}|h(x)-\mathcal{R}(x)|^2\mu(dx)\leq\varepsilon.$$

For all $f\in\mathcal{B}_n^*\mathcal{F}_n$, we have

$$|f(x)-\mathcal{R}(x)|^2\leq T_R^2\quad(x\in\mathbb{R}).$$

Therefore

$$\begin{aligned}
&\inf_{f\in\mathcal{B}_n^*\mathcal{F}_n}\int_{\mathbb{R}}|f(x)-\mathcal{R}(x)|^2\mu(dx)=\inf_{f\in\mathcal{B}_n^*\mathcal{F}_n}\int_{[-A,A]}|f(x)-\mathcal{R}(x)|^2\mu(dx)\\
+\quad&\inf_{f\in\mathcal{B}_n^*\mathcal{F}_n}\int_{[-A,A]^c}|f(x)-\mathcal{R}(x)|^2\mu(dx)\\
\leq\quad&\inf_{f\in\mathcal{B}_n^*\mathcal{F}_n}\int_{[-A,A]}|f(x)-\mathcal{R}(x)|^2\mu(dx)+T_R^2\mu\left(\mathbb{R}\setminus[-A,A]\right)\\
\leq\quad&2\inf_{f\in\mathcal{B}_n^*\mathcal{F}_n}\int_{[-A,A]}|f(x)-\bar{h}(x)|^2\mu(dx)+2\inf_{f\in\mathcal{B}_n^*\mathcal{F}_n}\int_{[-A,A]}|\bar{h}(x)-\mathcal{R}(x)|^2\mu(dx)+\varepsilon\\
\leq\quad&2\inf_{f\in\mathcal{B}_n^*\mathcal{F}_n}\int_{[-A,A]}|f(x)-\bar{h}(x)|^2\mu(dx)+3\varepsilon.
\end{aligned}$$

According (4.6) and $\varepsilon\to 0$ therefore

$$\inf_{f\in\mathcal{B}_n^*\mathcal{F}_n}\int_{\mathbb{R}}|f(x)-\mathcal{R}(x)|^2\mu(dx)\xrightarrow[n\to\infty]{\text{a.s}}0.$$

- The third step: shows that

$$\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{\hat{S}_R(Z_i)-\hat{S}_L(Z_i)}-\mathbf{E}(|f(X)-Y|^2)\right|\xrightarrow[n\to\infty]{\text{a.s}}0.$$

The following breakdown is used

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{\hat{S}_R(Z_i)-\hat{S}_L(Z_i)}-\mathbb{E}(|f(X)-Y|^2)\right|\\
\leq&\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i)|^2}{\hat{S}_R(Z_i)-\hat{S}_L(Z_i)}-\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{S_R(Z_i)-S_L(Z_i)}\right|\\
+&\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{S_R(Z_i)-S_L(Z_i)}-\mathbf{E}(|f(X)-Y|^2)\right|.
\end{aligned}$$

So to prove

$$\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{\hat{S}_R(Z_i)-\hat{S}_L(Z_i)}-\mathbf{E}(|f(X)-Y|^2)\right|\xrightarrow[n\to\infty]{\text{a.s}}0$$

just show that

$$\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i)|^2}{\hat{S}_R(Z_i)-\hat{S}_L(Z_i)}-\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{S_R(Z_i)-S_L(Z_i)}\right|\xrightarrow[n\to\infty]{\text{a.s}}0\tag{4.9}$$

and

$$\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{S_R(Z_i)-S_L(Z_i)}-\mathbf{E}(|f(X)-Y|^2)\right|\xrightarrow[n\to\infty]{\text{a.s}}0.\tag{4.10}$$

1. We start with the first term

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i)|^2}{\hat{S}_R(Z_i)-\hat{S}_L(Z_i)}-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{S_R(Z_i)-S_L(Z_i)}\right|\\
&=\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(A_i=0)}|f(X_i)-Z_i|^2\left[\frac{1}{\hat{S}_R(Z_i)-\hat{S}_L(Z_i)}-\frac{1}{S_R(Z_i)-S_L(Z_i)}\right]\right|\\
&\leq T_R^2\sup_{t\geq 0}\left|\frac{(S_R(t)-S_L(t))-(\hat{S}_R(t)-\hat{S}_L(t))}{(S_R(t)-S_L(t))(\hat{S}_R(t)-\hat{S}_L(t))}\right|\\
&\leq T_R^2\sup_{t\geq 0}\left|\frac{(S_R(t)-\hat{S}_R(t))+(\hat{S}_L(t)-S_L(t))}{(S_R(t)-S_L(t))(\hat{S}_R(t)-\hat{S}_L(t))}\right|\\
&\leq T_R^2\sup_{t\geq 0}\left[\frac{1}{(S_R(t)-S_L(t))(\hat{S}_R(t)-\hat{S}_L(t))}\right]\\
&\times\sup_{t\geq 0}\left[|S_R(t)-\hat{S}_R(t)|+|\hat{S}_L(t)-S_L(t)|\right].
\end{aligned}$$

we have

$$\begin{aligned}
\left(\hat{S}_R(t)-\hat{S}_L(t)\right) &= \left(\hat{S}_R(t)-S_R(t)+S_R(t)-\hat{S}_L(t)+S_L(t)-S_L(t)\right)\\
&= \left(\hat{S}_R(t)-S_R(t)\right)+\left(S_L(t)-\hat{S}_L(t)\right)+\left(S_R(t)-S_L(t)\right),
\end{aligned}$$

based on the convergence of $\hat{S}_R(t)$ and $\hat{S}_L(t)$, as given by equations (4.1) and (4.2) we find then

$$\lim_{n\to\infty}\left(\hat{S}_R(t)-\hat{S}_L(t)\right)=S_R(t)-S_L(t)\quad\text{a.s.}$$

Using the hypothesis H_3 therefore

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i)|^2}{\hat{S}_R(Z_i)-\hat{S}_L(Z_i)}-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{S_R(Z_i)-S_L(Z_i)}\right|\\
&\leq\frac{T_R^2}{b^2}\sup_{t\geq 0}\left[|(S_R(t)-\hat{S}_R(t)|+|\hat{S}_L(t)-S_L(t)|\right].
\end{aligned}$$

This gives

$$\lim_{n\to\infty}\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i)|^2}{\hat{S}_R(Z_i)-\hat{S}_L(Z_i)}-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{S_R(Z_i)-S_L(Z_i)}\right|=0\ \text{a.s.}\tag{4.11}$$

2. It remains to be proven that

$$\sup_{f\in\mathcal{F}_n^*\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(A_i=0)}\frac{|f(X_i)-Z_i|^2}{S_R(Z_i)-S_L(Z_i)}-\mathbf{E}(|f(X)-Y|^2)\right|\xrightarrow[n\to\infty]{\text{a.s}}0.$$

Let us introduce the following notations $V=(X,Z,1_A),V_1=(X_1,Z_1,1_{A_1}),\ldots,V_n=(X_n,Z_n,1_{A_n})$, n i.i.d random vectors of the same distribution as V. Let's Pose

$$\begin{aligned}
\mathcal{H}_n=\big\{&h:\mathbb{R}\times[0,T_R]\times\{0,1\}\to\mathbb{R}^+:\exists f\in\mathcal{F}_n^*\mathcal{F}_n\ \text{ such as}\\
&h(x,z,\mathbb{1}_A)=\frac{1_A|f(x)-z|^2}{S_R(z)-S_L(z)}\ \text{pour tout}\ \ (x,z,\mathbb{1}_A)\in\mathbb{R}\times[0,T_R]\times\{0,1\}\big\}.
\end{aligned}$$

Since $0 \leq Z \leq T_R$ a.s and $0 \leq f(x) \leq T_R$, for all $x \in \mathbb{R}$ and all $f \in \mathcal{F}_n^* \mathcal{F}_n$, functions in $\mathcal{H}_n$ are non-negative and bounded by $\frac{T_R^2}{b}$. In addition, we have

$$\mathbf{E}\left(h(V)\right) = \mathbb{E}\left(\frac{\mathbb{1}_A |f(X) - Z|^2}{S_R(Z) - S_L(Z)}\right) = \mathbf{E}\left(|f(X) - Y|^2\right).$$

And

$$\begin{aligned}
&\sup_{f \in \mathcal{F}_n^* \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{S_R(Z_i) - S_L(Z_i)} - \mathbf{E}(|f(X) - Y|^2) \right| \\
= &\sup_{h \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^{n} h(V_i) - \mathbf{E}h(V) \right|.
\end{aligned}$$

For any $h_1, h_2 \in \mathcal{H}_n$, be f_1, f_2 their corresponding functions in $\mathcal{F}_n^* \mathcal{F}_n$, then

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} |h_1(V_i) - h_2(V_i)| \\
&= \frac{1}{n} \sum_{i=1}^{n} \left| \mathbb{1}_{\{A_i=0\}} \frac{|f_1(X_i) - Z_i|^2}{S_R(Z_i) - S_L(Z_i)} - \mathbb{1}_{\{A_i=0\}} \frac{|f_2(X_i) - Z_i|^2}{S_R(Z_i) - S_L(Z_i)} \right| \\
\leq &\frac{1}{nb} \sum_{i=1}^{n} |(f_1(X_i) + f_2(X_i) - 2Z_i)(f_1(X_i) - f_2(X_i))| \\
&\leq \frac{2T_R}{nb} \sum_{i=1}^{n} |f_1(X_i) - f_2(X_i)|,
\end{aligned}$$

which implies that

$$\mathcal{N}(\varepsilon, \mathcal{H}_n, V_1^n) \leq \mathcal{N}\left(\varepsilon \frac{b}{2T_R}, \mathcal{F}_n^* \mathcal{F}_n, X_1^n\right)$$

where $\mathcal{N}(\varepsilon, \mathcal{F}_n, Z_1^n)$ denotes the recovery number.
We get for everything $\varepsilon > 0$,

$$\begin{aligned}
&\mathbf{P}\left\{ \sup_{h \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^{n} h(V_i) - \mathbf{E}h(V) \right| > \varepsilon \right\} \\
\leq &8\mathbf{E}\left\{ \mathcal{N}\left(\varepsilon \frac{b}{16T_R}, \mathcal{F}_n^* \mathcal{F}_n, X_1^n\right) \right\} \exp\left(-\frac{n\varepsilon^2 b^2}{128 T_R^4}\right) \\
\leq &16 \left(\frac{64 e T_R^3}{\varepsilon b^2}\right)^{2\mathcal{V}_{\mathcal{F}_n^* \mathcal{F}_n^+}} \exp\left(-\frac{n\varepsilon^2 b^2}{128 T_R^4}\right).
\end{aligned}$$

Since $\mathcal{V}_{\mathcal{F}_n^* \mathcal{F}_n^+} \leq \mathcal{V}_{\mathcal{F}_n^+}$ and the condition (4.5) of the theorem, the probability above is summable. And using the lemma of Borel-Cantelli, we have

$$\sup_{f \in \mathcal{F}_n^* \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(A_i=0)} \frac{|f(X_i) - Z_i|^2}{S_R(Z_i) - S_L(Z_i)} - \mathbf{E}(|f(X) - Y|^2) \right| \xrightarrow[n \to \infty]{\text{a.s}} 0.$$

So according to the relation (4.10) and (4.9), we find

$$\sup_{f \in \mathcal{F}_n^* \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(A_i=0)} \frac{|f(X_i) - Z_i|^2}{\hat{S}_R(Z_i) - \hat{S}_L(Z_i)} - \mathbf{E}(|f(X) - Y|^2) \right| \xrightarrow[n \to \infty]{\text{a.s}} 0.$$

□

5. Simulation Study

This section presents a simulation study illustrating the finite-sample performance of the proposed nonparametric least squares estimator under double censoring. We investigate the impact of the sample size ($n = 50$, $n = 100$, and $n = 500$) and compare the estimator with the true regression function under both linear and nonlinear models. In all experiments, the censoring rate does not exceed 30%.

5.1. Linear Model

In the linear setting, we generate data from

$$X = 0.1Y + 100 + \varepsilon,$$

where Y takes integer values from 1 to n, and $\varepsilon \sim \mathcal{N}(0, 0.5)$. The censoring variables are defined by $L \sim \mathcal{W}(7.1, 100)$ and $R = L + V$, with $V \sim \mathcal{W}(1, 80)$, where $\mathcal{W}$ denotes the Weibull distribution.

The following figures illustrate the comparison between the estimator and the true regression function for different sample sizes:

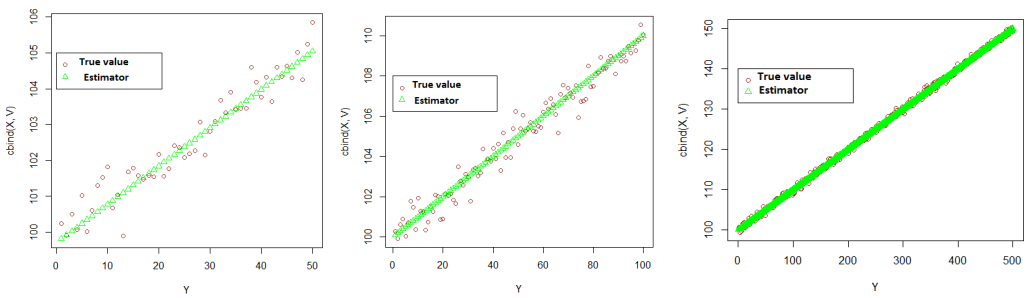

Figure 1: Estimator versus true regression function in the linear model.

For small sample sizes, the estimator exhibits noticeable variability and deviates from the true regression curve, especially in regions with heavier censoring. As the sample size increases, the estimator approaches the true regression function more closely, and the variability is substantially reduced. This behavior is consistent with the theoretical consistency result established in Section 4.1.

5.2. Nonlinear Model

In the nonlinear setting, we consider the quadratic model

$$X = 0.0001Y^2 + 0.1Y + 100 + \varepsilon,$$

with Y again ranging from 1 to n and $\varepsilon \sim \mathcal{N}(0, 0.5)$. The censoring variables are $L \sim \mathcal{W}(5, 99)$ and $R = L + V$, with $V \sim \mathcal{W}(3, 70)$.

The resulting estimator and true regression function are shown below: The results show that the estimator is able to capture the nonlinear structure of the regression function. As the sample size increases, the approximation error decreases uniformly over the support of Y. Compared to the linear model, the nonlinear estimator achieves a smaller approximation error, particularly for large samples, highlighting its greater flexibility in modeling complex relationships.

5.3. Model Comparison

To further assess performance, we consider the linear generating model

$$X = 0.1Y + 100 + \varepsilon,$$

and evaluate both the linear and nonlinear estimators. The following figures display the comparison across different sample sizes:

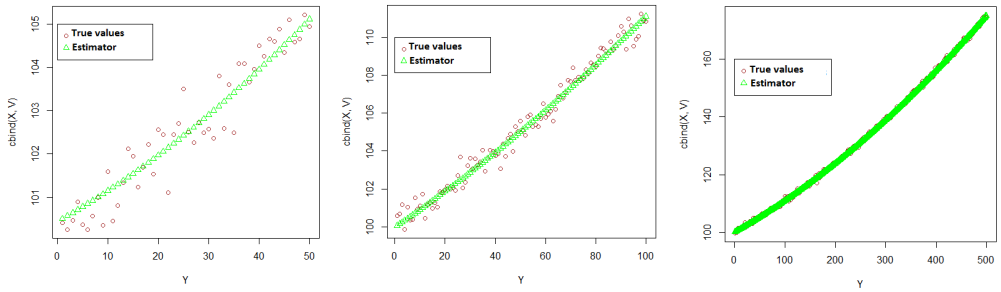

Figure 2: Estimator versus true regression function in the nonlinear model.

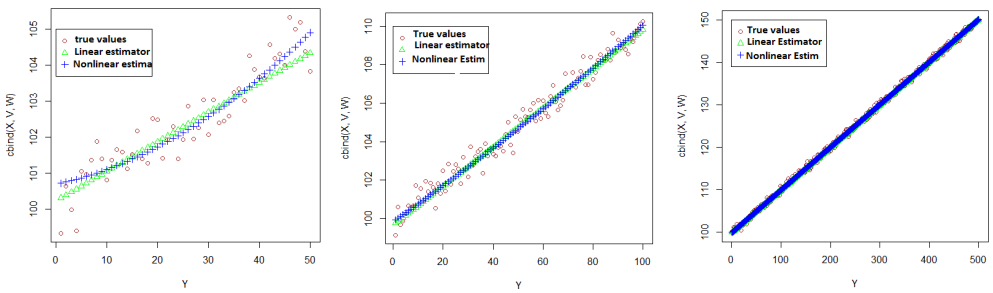

Figure 3: Comparison between linear and nonlinear estimators when the true model is linear.

The approximation errors are reported in Table 5.3:

n	Error_1 (linear estimator)	Error_2 (nonlinear estimator)
50	0.3494	0.3060
100	0.2681	0.2627
500	0.2175	0.2164

Table 5.3: Approximation errors of linear vs. nonlinear estimators.

A comparison between the linear and nonlinear estimators reveals that, while both estimators improve with increasing sample size, the nonlinear estimator consistently provides a better fit to the true regression function. This improvement becomes more pronounced for larger samples, where the nonlinear estimator benefits from its ability to adapt to curvature in the regression structure. Overall, these findings confirm that the least squares–based estimator performs more accurately in the nonlinear setting when sufficient data are available.

6. Conclusion

We developed a nonparametric least squares estimator for regression under double censoring. The construction relies on self-consistent estimators of the survival functions, and the resulting estimator is shown to be consistent. Simulation studies corroborate the theoretical results and illustrate the practical performance of the method.

References

1. Donoho, D. L., Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425–455.

2. Douas, R., Laroussi, I., Kharfouchi, S. Incomplete least squared regression function estimator based on wavelets. *Journal of Siberian Federal University. Mathematics & Physics*, 16(2), 204–215, (2023).

3. Fan, J., Gijbels, I. . *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London. (1996)

4. Idiou, G., Bourezaz, H., Laroussi, I. Neural network and least square estimator. *Gulf Journal of Mathematics*, 20(2), 147–161. (2025).
DOI: https://doi.org/10.56947/gjom.v20i2.325510.56947/gjom.v20i2.3255

5. Kebabi, K., Laroussi, I., Messaci, F. Least squares estimators of the regression function with twice censored data. *Statistics and Probability Letters*, 81, 1588–1593. (2011).

6. Laroussi, I. A generalised censored least squares and smoothing spline estimators of regression function. *International Journal of Mathematics in Operational Research*, 20(4), 506–520. (2021).

7. Laroussi, I. B-spline estimate of the regression function under general censorship model. *Jordan Journal of Mathematics and Statistics*, 17(1), 179–197. (2024).

8. Nadaraya, E. A. On estimating regression. *Theory of Probability and Its Applications*, 9, 141–142. (1964).

9. Stone, C. J. Consistent nonparametric regression. *The Annals of Statistics*, 5, 595–620. (1977).

10. Turnbull, B. W. The empirical distribution for interval-censored data. *Journal of the American Statistical Association*, 69, 290–295. (1974).

11. Watson, G. S. Smooth regression analysis. *Sankhyā, Series A*, 26, 359–372. (1964).

Ilhem Laroussi and Ranya Boustila,
Laboratory of Mathematics and Sciences of Decision,
Mentouri University,
Constantine 1,
Algeria
E-mail address: ilhem.laroussi@umc.edu.dz
E-mail address: ranya.boustila@doc.umc.edu.dz