EDN: GVQDEI
УДК 519.6

Incomplete Least Squared Regression Function Estimator Based on Wavelets

Ryma Douas*
Ilhem Laroussi†
Sciences Laboratory
Department of Mathematics
Mentouri Brothers University
Constantine, Algeria
Soumia Kharfouchi‡
Department of Mathematics
Laboratory of Mathematical Biostatistics Bioinformatics and
Methodology Applied to Health Sciences
Mentouri Brothers University
Constantine, Algeria

Received 10.07.2022, received in revised form 15.09.2022, accepted 20.10.2022

Abstract. In this paper, we introduce an estimator of the least squares regression function, for Y right censored by R and $\min(Y, R)$ left censored by L. It is based on ideas derived from the context of wavelet estimates and is constructed by rigid thresholding of the coefficient estimates of a series development of the regression function. We establish convergence in norm L_2. We give enough criteria for the consistency of this estimator. The result shows that our estimator is able to adapt to the local regularity of the related regression function and distribution.

Keywords: non-parametric regression, L_2 error, least squares estimators, orthogonal series estimates, convergence in the L_2-norm, twice censored data, regression estimation, hard thresholding.

Citation: R. Douas, I. Laroussi, S. Kharfouchi, Incomplete Least Squared Regression Function Estimator Based on Wavelets, J. Sib. Fed. Univ. Math. Phys., 2023, 16(2), 204–215. EDN: GVQDEI.

Regression is defined as being the set of statistical methods widely used to analyse the relationship between a variable and one or more others. For a long time, the regression of a random variable Y on a vector X of random variables designated the conditional mean of Y given X. Nowadays, the term regression designates any element of the conditional distribution of Y given X, as a function of X. We can for example be interested in the conditional mean, the conditional median, or the conditional variance. In presence of functional data, which are doubly infinite dimensional problems, the appeal to non parametric estimation is unavoidable. The starting point in this regards is a prediction problem that leads to the regression function due to the minimization of the mean squared error i.e., L_2 risk. In this setting, one can usually consider the model $Y = m(X) + \varepsilon$ where ε is centred and is independent of X with the explained variable fully observed. In the case of complete observation of $(X,\ Y)$, an abundant literature in this field can be found for instance in Györfi and al (2002) and references there in. However, in several situations the variable of interest X may be subject to randomly right and left censoring in the same sample. The lifetime Y is right censored by a variable R (which itself represents a survival

*rymadouas@yahoo.fr
†33laroussi@gmail.com
‡s_kharfouchi@yahoo.fr

time) and the minimum between Y and R is censored by a censorship variable on the left. A symbolical example of this model is the one given in Morales and al. (1991) that investigates the cause of death of trees on a farm. This kind of censoring model is exactly the Model one studied in Patilea and Rolin, for which local averaging estimates of $m(x) = \mathbf{E}(Y|X = x)$ has been introduced by Messaci (2010). In Kebabi and Messaci (2012), least squares estimator of $m(x)$ has been proposed and its L_2-norm convergence has been established. In this paper, we are mainly interested in least squares estimation approaches of the regression function for the Model I of Patilea and Rolin. Particularly, we investigate a least squares method based on wavelets. The use of a wavelets based approach is motivated by the possibility to achieve optimal convergence rates despite the high dimensionality of the problem. Moreover, wavelets are excellent approximators for signals with rapid local changes such as cusps, discontinuities, sharp spikes, etc. On the other hand, accurate wavelet decomposition, using only a few wavelet coefficient, can represent signals allowing dimensionality reduction and sparsity. So explicitly, the purpose of this paper is the construction of non-linear orthogonal series estimates by rigid transformation (thresholding) of the coefficients estimates of a regression function series development. The first part of our study is devoted to the introduction of the least squares estimators of the regression function for censored data and to some convergence properties. An important idea is introduced which consists in the estimation of orthogonal series of the regression function. Then, we present the estimation of the coefficients of these series, based on a wavelet system, is presented. In the second part, we list the proofs.

1. Model and recalls

Let (X, Y) be a random vector with values in $\mathbb{R}^d \times \mathbb{R}$ with $\mathbf{E}(Y)^2 < \infty$ and the dependence of Y on the value of X is of interest. Let R and L be censoring positive random variables. More specifically, the objective is to find a function $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(X)$ is a "good approximation" of Y.

1.1. Model

We introduce orthogonal series estimates of $m(x) = \mathbf{E}(Y|X = x)$ with respect to sample of iid $\mathcal{D}_n = \{X_i,\ Z_i = \max(\min(Y_i, R_i), L_i),\ A_i\}$ from the same distribution as (X, Z, A) or $Z = \max(\min(Y, R), L)$ and

$$A = \begin{cases} 0 & \text{if} \quad L < Y < R, \\ 1 & \text{if} \quad L < R \leqslant Y, \\ 2 & \text{if} \quad \min(Y, R) \leqslant L. \end{cases}$$

Indeed, let $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary (measurable) function and denote X distribution par μ then

$$\begin{aligned} \mathbf{E}|f(X) - Y|^2 &= \mathbf{E}|f(X) - m(X) + m(X) - Y|^2 = \\ &= \mathbf{E}|f(X) - m(X)|^2 + \mathbf{E}|m(X) - Y|^2 = \\ &= \mathbf{E}|m(X) - Y|^2 + \int |f(x) - m(x)|^2 \mu(dx). \end{aligned}$$

In the sequel we will denote by F_V the distribution function of the random variable V and by $S_V = 1 - F_V$ its survival function and $T_V = \sup\{t : F_V(t) < 1\}$ and $I_V = \inf\{t : F_V(t) \neq 0\}$ the end points of the support of the variable V. Assume that the variables X, Y, R et L satisfies the following hypotheses

H_1 : Y, R and L are independent.

H_2 : (L, R) is independent of (X, Y).

H_3 : $\exists T < T_R$ and $I > I_L$ such that, $\forall n \in \mathbb{N}, \forall i (1 \leqslant i \leqslant n) : A_i = 0 \Rightarrow I \leqslant Z_i \leqslant T$ a.s.

H_4 : F_L is continuous on $]0, \infty[$.

H_5 : $T_R \leqslant T_Y \leqslant T_L < \infty$ and $I_Y \leqslant I_L < I_R$.

H_1 is an inherent hypothesis of Patilea's et al . H_3 seems to be acceptable because $I \leqslant Z_i \leqslant T$ when $A_i = 0$. H_5 guarantees in particular that the model is identifiable.
Let h a mapping on $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, we introduce as unbiased estimator of $\mathbf{E}(h(X, Y))$ the amount

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{h(X_i, Z_i)}{S_R(Z_i)F_L(Z_i)}. \tag{1}$$

Indeed, under hypothesis H_1, H_2 and H_4. The problem is that functions S_R and F_L are generally unknown, we will replace them respectively with their estimators. Let $(Z'_j)_{1 \leqslant j \leqslant M}$, $(M \leqslant n)$ be the distinct values of Z_i listed in ascending order.

1.2. Estimation and proprieties

Set

$$D_{kj} = \sum_{i=1}^{n} 1_{\{Z_i = Z'_j, A_i = k\}}, \text{ and } \ N_j = \sum_{i=1}^{n} 1_{\{Z_i \leqslant Z'_j\}},$$

thus, [22] suggest estimating S_R by

$$\hat{S}_n(t) = \prod_{j/Z'_j \leqslant t} \left\{1 - \frac{D_{1j}}{U_{j-1} - N_{j-1}}\right\} \text{ and } U_{j-1} = n \prod_{j \leqslant l \leqslant M} \left\{1 - \frac{D_{2l}}{N_l}\right\}, \tag{2}$$

and by inverting time in the Kaplan et al estimator, we can deduce the estimator $\hat{F}_n$ from F_L (left censoring case) witch is

$$\hat{F}_n(t) = \prod_{j/Z'_j > t} \left\{1 - \frac{1_{\{A_j=2\}}}{j}\right\}. \tag{3}$$

Recall that under hypothesis H_1 and H_5, [22] have proven that

$$\sup_{t \in \mathbb{R}^+} \left|\hat{S}_n(t) - S_R(t)\right| \underset{n \longrightarrow \infty}{\longrightarrow} 0 \text{ a.s.} \tag{4}$$

And

$$\sup_{t \in \mathbb{R}^+} \left|\hat{F}_n(t) - F_L(t)\right| \underset{n \longrightarrow \infty}{\longrightarrow} 0 \text{ a.s.} \tag{5}$$

Note that hypothesis H_3 implies that

$$S_R(T) > 0 \text{ and } F_L(I) > 0. \tag{6}$$

In view of equations (4) – (6), we deduce that for n sufficiently large

$$\hat{S}_n(T) > 0 \text{ and } \hat{F}_n(I) > 0 \text{ a.s.}$$

If Y is uncensored , the regression function estimator of the least squares , obtained by minimizing the empirical risk L_2, is $\arg\min_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2$, where $\mathcal{F}_n$ is a class of functions that is

depending on the sample size n. Thus, in our context, according to the relation h and after having estimated S_R and F_L, the least squares estimator of $m(x)$ is given by

$$\widetilde{m}_n = \arg\min_{f\in\mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \left(\frac{0}{0} := 0\right). \tag{7}$$

$\mathcal{F}_n$ is a certain family of functions which will be clarified in the theorem. We see that $\hat{S}_n(Z_i)$ does not vanish in the expression of $\widetilde{m}_n$ if $A_i = 0$. It is easy to check that $\hat{F}_n(Z_i)$ does not vanish either if $A_i = 0$ but since Y is bounded, we are going to make some assumptions on our estimator. For that reintroduce the notation of the next use of truncation.

For $0 \leqslant t < \infty$ and $x \in \mathbb{R}$, define

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

and for $f : \mathbb{R}^d \to \mathbb{R}$, define $(\mathbf{T}_{[0,t]}f)(x) = \mathbf{T}_{[0,t]}(f(x))$. We can also use again the fact that this mapping verifies the following relation.

$$\forall b > a, \quad |\mathbf{T}_{[0,b]}(x) - \mathbf{T}_{[0,a]}(x)| \leqslant (b-a). \tag{8}$$

Y being limited and due to $M_n = \max(Z_1, \ldots, Z_n)$ with $M_n \underset{n\to+\infty}{\to} T_L$ a.s, we finally propose as an estimator of $m(x)$

$$m_n(x) = \mathbf{T}_{[0,M_n]}(\tilde{m}_n(x)). \tag{9}$$

1.3. Wavelet bases

Let $\mathcal{F}_n$ be the set of all piecewise polynomials of degree M (or less) with respect to some partition of $[0,1]$ consisting of $4n^{1-\alpha}$ intervals (or less). Let G_M be set of polynomials of degree M (or less), let P_n be an equidistant partition of $[0,1]$ in $\lceil \log(n) \rceil$ intervals. Denote $G_M \circ P_n$ the set of all piecewise polynomials of degree M (or less) with respect to P_n. We will also need the following notations

$$\mathcal{L}_n^{**} = \mathrm{T}_{\log n}(\mathcal{F}_n).$$

$$\mathcal{F}_n^{**} = \{\forall f \in G_M \circ P_n, \|f\|_\infty \ \leqslant \log(n)\}.$$

Now adapting the proofs given in Kohler et al [17], We get the following result concerning the convergence of the introduced estimators. We refer, for example to Györfi et al [7] for some definitions and results of the Vapnik et al [23] theory, used in this work.

We introduce orthogonal series estimates in the context of regression estimation with fixed, equidistant design, which is the field where they have been applied most successfully. Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be data according to the model $Y_i = m(x_i) + \varepsilon_i$ where x_i are fixed (non-random) equidistant points in $[0,1]$, ε_i are i.i.d. random variables with $\varepsilon_i = 0$ and $\mathbf{E}(\epsilon_i) < \infty$ and m is a regression function $f : [0,1] \to \mathbb{R}$.

Assume that $m \in L_2(\mu)$ where μ is Lebesgue measure on $[0,1]$; and $(f_j)_{j\in\mathbb{N}}$ is an orthonormal basis in $L_2(\mu)$, ie

$$\langle f_j, f_k\rangle_{L_2(\mu)} = \int f_j(x) f_k(x)\mu(dx) = \begin{cases} 1 & \text{if} \quad j = k \\ 0 & \text{if} \quad j \neq k \end{cases}.$$

Each function in $L_2(\mu)$ can be arbitrarily approximated by linear combinations of $(f_j)_{j\in\mathbb{N}}$. Then m can be represented by its Fourier series with respect to $(f_j)_{j\in\mathbb{N}}$,

$$m=\sum_{j=1}^{\infty}c_jf_j \quad \text{where} \quad c_j=\langle m,f_j\rangle_{L_2(\mu)}=\int m(x)f_j(x)\mu(dx). \tag{10}$$

Orthogonal series estimates use the estimates of coefficients of a series expansion $\mathbf{E}|f(X)-Y|^2=\mathbf{E}|m(X)-Y|^2+\int|f(x)-m(x)|^2\mu(dx)$ to reconstruct the regression function and in the model $Y_i=m(x_i)+\varepsilon_i$, where $x_1,\ldots,x_n$ are equidistant in $[0,1]$; coefficients c_j can be estimated by

$$\hat{c}_j=\frac{1}{n}\sum_{i=1}^{n}Y_if_j(x_i), j\in\mathbb{N}. \tag{11}$$

The traditional way to deal with these estimated coefficients to construct an estimate

$$m^1{}_n=\sum_{j=1}^{\tilde{K}}\hat{c}_jf_j,$$

m is to truncate the series expansion to an index $\tilde{K}$ and to inject the estimated coefficients.

Here, we try to choose $\tilde{K}$ such that the set of functions$\{f_1,\ldots,f_{\tilde{K}}\}$ is the "best" among all the sub-sets $\{f_1\},\{f_1,f_2\},\{f_1,f_2,\ldots\}$ of $\{f_j\}_{j\in\mathbb{N}}$ in view of the estimation error (7). This implicitly assumes that the most important information m is in the first coefficients $\tilde{K}$ of the series expansion $\mathbf{E}|f(X)-Y|^2=\mathbf{E}|m(X)-Y|^2+\int|f(x)-m(x)|^2\mu(dx)$.

[5] have proposed a way to overcome this hypothesis. This consists in contaminating the estimated coefficients, for example, we use all the coefficients whose absolute value is greater than a threshold δ_n (called hard thresholding). This leads to estimates of the form

$$m^2{}_n=\sum_{j=1}^{K}\eta_{\delta_n}(\hat{c}_j)f_j,$$

where K is generally much larger than $\tilde{K}$ in (7), $\delta_n>0$ is a threshold, and

$$\eta_{\delta_n}(\hat{c_j})=\begin{cases}\hat{c_j} & \text{if} \quad |\hat{c_j}|>\delta_n\\ o & \text{if} \quad |\hat{c_j}|\geqslant\delta_n\end{cases},$$

in the series expansion, we truncate the estimate at some data-independent height B_n, in other words, we define

$$\bar{m}_n(x)=(T_{B_n}\tilde{m}_n)(x)=\begin{cases}B_n & \text{if } \tilde{m}_n(x)>B_n,\\ \tilde{m}_n(x) & \text{if } -B_n\leqslant\tilde{m}_n(x)\leqslant B_n,\\ 0 & \text{if } \tilde{m}_n(x)<-B_n,\end{cases} \tag{12}$$

where $B_n>0$ and $B_n\to\infty$ $(n\to\infty)$.

In this paper, we study the consistency of our estimator of orthogonal series. for simplicity we will consider the case where $X\in[0;1]$ a.s. It is easy to modify the definition of our estimator so that we obtain a weakly and strongly universally consistent estimator for the univariate X. To prove the strong consistency of our estimator we need to make somme changes to its definition. Consider $\alpha\in(0;\frac{1}{2})$. Let functions f_j and coefficients $\hat{c}_j$ be as defined in (10) and (11). Write $(\hat{c}_{(1)};f_{(1)}),\ldots,(\hat{c}_{(K)};f_{(K)})$

switching $(\hat{c_1},f_1),\ldots,(\hat{c_K},f_K)$ and

$$|\hat{c}_1|\geqslant|\hat{c}_2|\geqslant\cdots\cdots\geqslant|\hat{c}_k| \tag{13}$$

let's define the estimator ${m^3}_n$ as

$$ {m^3}_n = \sum_{j=1}^{\min\{K, n^{\lfloor 1-\alpha \rfloor}\}} \eta_{\delta_n}(\hat{c}_j) f_j \tag{14} $$

This ensures that ${m^3}_n$ and a linear combination of no more than $n^{1-\alpha}$ functions f_j. And as in $\mathbf{E}|f(X) - Y|^2 = \mathbf{E}|m(X) - Y|^2 + \int |f(x) - m(x)|^2 \mu(dx)$ we can show that

$$ {m^3}_n = {m^3}_{n,J^*} \text{ with } J^* \subseteq \{1, \ldots\ldots, K\} \text{ where } J^* \text{ satisfies } |J^*| \leqslant n^{1-\alpha}. $$

finally we combine the notation of the two estimates to obtain as an estimate of $\tilde{m}_n$ the following formulas m_n^3 and m_n whith $T_L \leqslant B_n = \log(n)$. We will also need the following notations

$$ \mathcal{L}_n^* = \mathrm{T}_{T_L}(\mathcal{F}_n).\ \mathcal{F}_n^* = \{g : \exists f \in G_M \circ P_n, g = \mathrm{T}_{[0,T_L]} f\}. $$

2. Results

Theorem 2.1. *Under hypotheses H_1 – H_5, let $M \in \mathbb{N}$ be fixed, and m_n the m estimator defined by 9, 14 , with $T_L \leqslant B_n = \log(n)$ and $\delta_n \leqslant \dfrac{1}{(\log(n)+1)^2}$. Then*

$$ \int_{R^d} |m_n(x) - m(x)|^2 \mu(dx) \underset{n\to\infty}{\to} 0 \ a.s\ . $$

The following lemma will be used to establish our main result.

Lemma 2.2. *We set the quantity $\bar{m}_n(x) = \mathbb{T}_{[0,T_L]}(\tilde{m}_n(x))$ and with equations (2), (3), we have*

$$ \begin{aligned} &\int_{R^d} |\bar{m}_n(x) - m(x)|^2 \mu(dx) \leqslant \\ &\leqslant 2 \sup_{f \in \mathcal{L}_n^*} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{\{A_i = 0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \ E\,|f(X) - Y|^2 \right| + \\ &+ n\delta_n^2 2(M+1) \frac{(log(n)+1)^2}{n} + \inf_{f \in \mathcal{F}_n^*} \int_{R^d} |f(x) - m(x)|^2 \mu(dx). \end{aligned} \tag{15} $$

3. Proofs

We set the quantity $\bar{m}_n(x) = \mathbb{T}_{[0,T_L]}(\tilde{m}_n(x))$. We first show that the theorem is proved

$$ \int_{R^d} |m_n(x) - m(x)|^2 \mu(dx) \underset{n\to\infty}{\to} 0 \text{ a.s .} \iff \int_{R^d} |\bar{m}_n(x) - m(x)|^2 \mu(dx) \underset{n\to\infty}{\to} 0 \text{ a.s .} $$

Indeed, according to equation (8), we have $|m_n(x) - \bar{m}_n(x)|^2 \leqslant |T_L - M_n|$, which implies that

$$ \int_{\mathbb{R}^d} |m_n(x) - \bar{m}_n(x)|^2 \leqslant (T_L - M_n)^2 \to 0 \text{ a.s..} $$

Since by H_5 we have $\lim\limits_{n\to+\infty} M_n = T_L$ a.s Kebabi et al [12]. First, we prove the Lemma 2.2, and finally, we prove the theorem.

Proof Lemma 2.2. We start by proving, first we have

$$\begin{aligned}&\int_{R^d} |\bar{m}_n(x) - m(x)|^2 \mu(dx) = \\ &= \left\{ \mathbf{E}(|\bar{m}_n(X) - Y|^2 |\mathcal{D}_n) - \inf_{f \in \mathcal{F}_n^*} \mathbf{E}\,|f(X) - Y|^2 \right\} + \\ &+ \left\{ \inf_{f \in \mathcal{F}_n^*} \mathbf{E}\,|f(X) - Y|^2 - \mathbf{E}\,|m(X) - Y|^2 \right\}.\end{aligned}$$

In addition, the regression function satisfies

$$\inf_{f \in \mathcal{F}_n^*} E\,|f(X) - Y|^2 - \ E\,|m(x) - Y|^2 = \inf_{f \in \mathcal{F}_n^*} \int_{R^d} |f(x) - m(x)|^2 \mu(dx). \tag{16}$$

furthermore

$$\begin{aligned}& E\left(|\bar{m}_n(X) - Y|^2 |\mathcal{D}_n\right) - \inf_{f \in \mathcal{F}_n^*} E\,|f(X) - Y|^2 = \\ = & \sup_{f \in \mathcal{F}_n^*} \left\{ E\left(|\bar{m}_n(X) - Y|^2 |\mathcal{D}_n\right) - \ E\left(|f(X) - Y|^2 |\mathcal{D}_n\right) \right\} = \\ = & \sup_{f \in \mathcal{F}_n^*} \left\{ E\left(|\bar{m}_n(X) - Y|^2 |\mathcal{D}_n\right) - \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|\bar{m}_n(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} + \right. \\ & + \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|\bar{m}_n(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|\tilde{m}_n(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} + \\ & + \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|\tilde{m}_n(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} + \\ & \left. + \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \ E\,|f(X) - Y|^2 \right\} \leqslant \sum_{i=1}^4 Q_{n,i},\end{aligned}$$

where the $Q_{n,i}$ are explained below for all i, $1 \leqslant i \leqslant 4$.

- Since $\tilde{m} \in \mathcal{F}_n$,$\bar{m}_n \in \mathcal{F}_n^*$ and $\mathcal{F}_n^* \subset \mathcal{L}_n^*$, it is obvious that

$$\begin{aligned}Q_{n,1} &= \sup_{f \in \mathcal{F}_n^*} \left\{ E\left(|\bar{m}_n(X) - Y|^2 |\mathcal{D}_n\right) - \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|\bar{m}_n(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \right\} \leqslant \\ &\leqslant \sup_{f \in \mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \ E\,|f(X) - Y|^2 \right|,\end{aligned}$$

and

$$\begin{aligned}Q_{n,4} &= \sup_{f \in \mathcal{F}_n^*} \left\{ \left| \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \ E\,|f(X) - Y|^2 \right| \right\} \leqslant \\ &\leqslant \sup_{f \in \mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \ E\,|f(X) - Y|^2 \right|.\end{aligned}$$

- Since $\bar{m}_n(X_i) \leqslant T_L$ and $Z_i \leqslant T_L$ a.s., we obtain $1_{\{A_i=0\}}|\tilde{m}_n(X_i)-Z_i| \geqslant 1_{\{A_i=0\}}|\bar{m}_n(X_i)-Z_i|$, which implies

$$Q_{n,2} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|\bar{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|\tilde{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \leqslant 0$$

- As $\mathcal{F}_n^* \subset \mathcal{F}_n^{**}$ because of $T_L \leqslant \log(n)$ and fix $f \in G_M \circ P_n$. In view of P_n definition, Lemma 18.1 in Györfi et al [7] exist $\bar{J} \subset \{1,\ldots,n\}$ and $\bar{f} \in \mathcal{F}_{n,\bar{J}}$, such that $f(X_i) = \bar{f}(X_i)$ and $|\bar{J}| \leqslant 2(M+1)(log(n)+1)^2$ which implies that

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|\tilde{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} =$$

$$= \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|\tilde{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|\bar{f}(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \leqslant$$

$$\leqslant n\delta_n^2 2(M+1)\frac{(log(n)+1)^2}{n}.$$

From $\tilde{m}$ definition, it is obvious that

$$Q_{n,3} = \sup_{f\in\mathcal{F}_n^*} \left\{ \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|\tilde{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \right\} \leqslant$$

$$\leqslant n\delta_n^2 2(M+1)\frac{(log(n)+1)^2}{n}.$$

Inequality (15) is therefore proven. □

Proof Theorem 2.1. It remains to be proven that the three terms of Lemma 2.2 tend to zero almost surely when $n \to \infty$. To do this, we will proceed in three steps. In the first step, we show that

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

To do this, we use the following inequalities

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

$$\leqslant \sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)\hat{F}_n(Z_i)} \right| +$$

$$+ \sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)} \right| +$$

$$+ \sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)} - E\,|f(X)-Y|^2 \right| \leqslant \sum_{i=1}^{3} Q_{n,i}^*.$$

Since $f \in \mathcal{L}_n^*$ implies that $0 \leqslant f(x) \leqslant T_L$, we get – in view of – formulas (4)–(6)

$$Q_{n,1}^* = \sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)\hat{F}_n(Z_i)} \right| \leqslant$$

$$\leqslant \frac{T_L^2}{\hat{S}_n(T)S_R(T)\hat{F}_n(I)} \sup_{t\in\mathbb{R}^+} \left|\hat{S}_n(t) - S_R(t)\right| \underset{n\to\infty}{\to} 0, \text{ a.s.}$$

and

$$Q_{n,2}^* = \sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)} \right| \leqslant$$

$$\leqslant \frac{T_L^2}{F_L(I)S_R(T)\hat{F}_n(I)} \sup_{t\in\mathbb{R}^+} \left|\hat{F}_n(t) - F_L(t)\right| \underset{n\to\infty}{\to} 0 \text{ a.s.}$$

Let's 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 with the same distribution as V.

Define

$$\mathcal{H}_n = \Big\{ h : \mathbb{R}^d \times [0,T_L] \times \{0,1\} \to \mathbb{R}^+ : \exists f \in \mathcal{L}_n^* \text{ such as,}$$
$$h(x,z,1_A) = \frac{1_A\,|f(x)-z|^2}{S_R(z)F_L(z)}$$
$$\forall\,(x,z,1_A) \in \mathbb{R}^d \times [0,T_L] \times \{0,1\} \Big\}.$$

Functions of $\mathcal{H}_n$ are positive and bounded by $\dfrac{T_L^2}{S_R(T)F_L(I)}$, and

$$\mathbf{E}h(V) = \mathbf{E}\left(\frac{1_A\,|f(X)-Z|^2}{S_R(Z)F_L(Z)}\right) = \mathbf{E}\left[\mathbf{E}\left(\frac{1_A\,|f(X)-Z|^2}{S_R(Z)F_L(Z)} \mid X,Y\right)\right] = \mathbf{E}\left(|f(X)-Z|^2\right).$$

under H_1,H_2 et H_4. In addition we have

$$Q_{n,3}^* = \sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^n 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)} - E\,|f(X)-Y|^2 \right| =$$
$$= \sup_{f\in\mathcal{H}_n} \left| \frac{1}{n}\sum_{i=1}^n h(V) - \mathbf{E}h(V) \right|.$$

For all h_1 and $h_2 \in \mathcal{H}_n$, let f_1 and f_2 be their corresponding functions in $\mathcal{L}_n^*$ then

$$\frac{1}{n}\sum_{i=1}^n |h_1(V_i) - h_2(V_i)|$$
$$= \frac{1}{n}\sum_{i=1}^n \left| 1_{\{A_i=0\}} \frac{|f_1(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)} - 1_{\{A_i=0\}} \frac{|f_2(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)} \right| \leqslant$$
$$\leqslant \frac{1}{S_R(T)F_L(I)} \frac{1}{n}\sum_{i=1}^n |(f_1(X_i)+f_2(X_i)-2Z_i)\,(f_1(X_i)-f_2(X_i))| \leqslant$$
$$\leqslant \frac{2T_L}{S_R(T)F_L(I)} \frac{1}{n}\sum_{i=1}^n |f_1(X_i)-f_2(X_i)|,$$

which implies $\mathcal{N}(\varepsilon,\mathcal{H}_n,V_1^n) \leqslant \mathcal{N}\left(\varepsilon\frac{S_R(T)F_L(I)}{2T_L},\mathcal{L}_n^*,X_1^n\right)$, where $\mathcal{N}(\varepsilon,\mathcal{F}_n,Z_1^n)$ denotes the overlapping number. Theorem 9.1 in Györfi et al [7] gives, for all $\delta > 0$

$$p\left\{ \sup_{f\in\mathcal{H}_n} \left| \frac{1}{n}\sum_{i=1}^n h(V_i) - \mathbf{E}h(V) \right| > \delta \right\} \leqslant 8\,E\left\{ \mathcal{N}\left(\delta\frac{S_R(T)F_L(I)}{16T_L},\mathcal{L}_n^*,X_1^n\right)\right\} \exp\left(-\frac{n\delta^2 S_R^2(T)F_L^2(I)}{128T_L^4}\right),$$

which is, in view of Theorem 9.4, Theorem 9.5 and Lemma 13.1 in Györfi et al [7], we get

$$p\left\{\sup_{f\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n}h(V_i)-\mathbf{E}h(V)\right|>\delta\right\}\leqslant$$

$$\leqslant\ 8(5n)^{4n^{1-\alpha}}\left(-\frac{288eT_L^2}{\delta\left(S_R(T)F_L(I)\right)^4}\right)^{2(M+2)n^{1-\alpha}}\exp\left(-\frac{n\delta^2S_R^2(T)F_L^2(I)}{128T_L^4}\right).$$

The formula combined with the $V_{\mathrm{T}_{\log n}G_M^+}\leqslant V_{G_M^+}$ of the theorem where $V_{\mathrm{T}_{\log n}G_M^+}$ stands for the VC dimension of the set of graphs of function in G_M , allows to apply Borel Cantelli lemma, to get

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

In the second step, we get

$$n\delta_n^2 2(M+1)\frac{(log(n)+1)^2}{n}\underset{n\to\infty}{\to}0\text{ a.s because }\delta_n\leqslant\frac{1}{(\log(n)+1)^2}.$$

In the third step, we prove that

$$\inf_{f\in\mathcal{F}_n^*}\int_{R^d}\left|f(x)-m(x)\right|^2\mu(dx)\underset{n\to\infty}{\to}0\text{ a.s.}$$

Since m can be approximated arbitrarily closely by continuously differentiable functions, we may assume without loss of generality that m is continuously differentiable.For each $A\in P_n$ choose some $x_A\in A$ and set $f^*=\sum_{A\in P_n}m(x_A)I_A$.Then $f^*\in G_M\circ P_n$ and for n such that $\|m\|_\infty\ \leqslant T_L\leqslant\log(n)$ we get

$$\inf_{\forall f\in G_M\circ P_n,\|f\|_\infty\ \leqslant T_L}\int_{R^d}\left|f(x)-m(x)\right|^2\mu(dx)\leqslant\sup_{x\in[0,1]}\left|f^*(X)-m(x)\right|^2\leqslant\frac{c}{(\log(n))^2}\underset{n\to\infty}{\to}0.$$

where c is constant as a function of the first derivative of m. □

References

[1] L.Devroye, L.Györfi, A.Krzyżak, G.Lugosi, On the strong universal consistency of nearest neighbor regression function estimates, *Annals of Statistics*, **22**(1994), 1371–1385.

[2] L. Devroye, L.Györfi, G. Lugosi, A probabilist theory of pattern Recognition, Springer Verlag, 1996.

[3] N.R.Draper, H.Smith, Applied Regression Analysis, 2nd ed. Wiley, New York, 1981.

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

[5] D.Donoho, Adapting to unknown smoothness via wavelet shrinkage, *J. Am. Statist. Ass.*, **90**(1995), 1200–1224.

[6] D.Donoho, I.M.Johnstone, Minimax estimation via wavelet shrinkage, *Annals of Statistics*, **26**(1998), 879–921.

[7] L.Györfi, M.Kohler, A.Krzyżak, H.Walk, A Distribution Free theory of Non parametric Regression, Springer–Verlag, New York, Inc. 2002. DOI: 10.1007/b97848

[8] D.J.Hart, Non parametric Smoothing and Lack-of-Fit Tests, Springer–Verlag, New York, 1997.

[9] T.Hastie, R.J.Tibshirani, Generalized Additive Models, Chapman and Hall, London, UK, 1990.

[10] D.Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Inform. Comput.*, **100**(1992), 78–150.

[11] E.L.Kaplan, P.Meier, Nom parametric estimation from incomplete observations, *J. Amer. Statist. Assoc.*, **53**(1958), 457–481.

[12] K.Kebabi, I.Laroussi, F.Messaci, Least squares estimators of the regression function with twice censored data, *Statist. Probab. Lett.*, **81**(2011), 1588–1593.
DOI: 10.1016/j.spl.2011.06.010

[13] K.Kebabi, F.Messaci, Rate of the almost complete convergence of a kernel regression estimate with twice censored data, *Statist. Probab. Lett.*, **82**(2012), no. 11, 1908–1913.
DOI: 10.1016/j.spl.2012.06.026

[14] M.Kohler, On the universal consistency of a least squares spline regression estimator, *Math. Methods Statist.*, **6**(1997), 349–364.

[15] M.Kohler, Universally consistent regression function estimation using hierarchical b-splines, *J. Multivariate Anal.*, **67**(1999), 138–164.

[16] M.Kohler, A.Krzyżak, Nonparametric regression estimation using penalized least squares, *IEEE Trans. Inform. Theory*, **47**(2001), 3054–3058. DOI:10.1109/18.998089

[17] M.Kohler, K.Màthé, M. Pintér, Prediction from randomly right censored data, *J. Multivariate Anal.*, **80**(2002), 73–100. DOI: 10.1006/jmva.2000.1973

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

[19] D.Morales, L.Pardo, V.Quesada, Bayesian survival estimation for incomplete data when the life distribution is proportionally related to the censoring time distribution, *Comm. Statist. Theory Methods*, **20**(1991), 831–850. MR1131189.

[20] E.A.Nadaraya, On estimating regression, *Theory of Probability and Its Applications*, **9**(1964), no. 1, 141–142.

[21] A.Nobel, Histogram Regression Estimation Using Data-dependent Partitions, *Ann. Statist*, **24**(1996), 1084–1105.

[22] V.Patilea, J.M.Rolin, Product-limit estimators of the survival function with twice censored data, *Ann. Statist.*, **34**(2006), no. 2, 925–938. DOI: 10.1214/009053606000000065

[23] V.N.Vapnik, 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.

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

[25] V.N.Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

Неполная оценка функции регрессии методом наименьших квадратов на основе вейвлетов

Рима Дуас
Ильхем Ларуси
Сумия Харфуши
Кафедра математики
Университет братьев Ментури
Константин, Алжир

Аннотация. В этой статье мы вводим оценку функции регрессии методом наименьших квадратов для Y, цензурированного справа R, и $\min(Y, R)$, цензурированного слева L. Он основан на идеях, полученных из контекста вейвлет-оценок, и построен путем жесткой пороговой обработки оценок коэффициентов развития ряда функции регрессии. Устанавливаем сходимость по норме L_2. Мы даем достаточно критериев для непротиворечивости этой оценки. Результат показывает, что наша оценка способна адаптироваться к локальной регулярности соответствующей функции регрессии и распределения.

Ключевые слова: непараметрическая регрессия, ошибка L_2, оценки методом наименьших квадратов, оценки ортогональными рядами, сходимость в норме L_2, дважды цензурированные данные, оценка регрессии, жесткий порог.