Simulating Performance: A Comparative Study Between two Regression Estimators for Left Truncated Data

F. Hamrani[1]**, Z. Guessoum**[2]**, E. Ould-Saïd**[3]**, A. Tatachak**[4]

[1]Mouloud Mammeri University of Tizi Ouzou, Tizi Ouzou, Algeria

[2,4]University of Science and Technology Houari Boumediene, Algiers, Algeria

[3]University of the Littoral Opal Coast, Calais, France

farida.hamrani@ummto.dz, zguessoum@usthb.dz, atatachak@usthb.dz,
elias.ould-said@univ-littoral.fr

In this work, our focus is on the estimation of the regression function in a non-parametric setting, considering the presence of random left truncation in the target random variable. Our objective is to compare two estimators: the relative regression estimator, obtained by minimizing the mean squared relative error, and the estimator obtained by minimizing the mean squared error. To accomplish this, we conduct a simulation study to assess the performance of these estimators when the data exhibit dependence under various scenarios.

We aim to estimate the real random variable of interest, Y, based on the observed R^d-valued random vector of covariates, $\mathbf{X}$, $d \geq 1$. One popular approach for estimation is regression modeling, where we consider the relationship $Y = m(\mathbf{X}) + \epsilon$, with m being the unknown regression function and ϵ representing the random error variable.

Traditionally, the regression function m is estimated using the mean squared error as the loss function. However, this loss function is highly sensitive to outliers. To address this issue, an alternative approach involves utilizing a loss function based on the squared relative error. This technique, which has been employed in regression analysis, proves beneficial when analyzing data with positive responses, such as lifetimes (the focus of our work).

Another characteristic of lifetimes is that they are often incompletely observed. Incomplete data takes various forms, with censorship and truncation being the most common. In this study, our focus is on left truncated data, where the observation $(\mathbf{X}, Y)$ is affected by an independent random variable T. All three random quantities, Y, $\mathbf{X}$, and T, are observable only when $Y \geq T$. This model, that originally appeared in astronomy [5], was then applied in such areas as economics, epidemiology, demographics, actuarial mathematics.

Ould-Saïd and Lemdani [3] built a kernel estimator of the function $m(.)$ which take into account the truncation effect. The authors constructed a kernel regression function for $m(.)$ by minimizing the following mean squared loss function

$$\mathbb{E}\left[(Y - m(\mathbf{X}))^2 | \mathbf{X}\right]. \tag{1}$$

However, this kind of loss function is not suitable in some situations, especially when the data is affected by the presence of outliers. So, to circumvent this constraint, a robust against outliers approach was proposed to estimate $m(.)$ by minimizing the following mean squared relative error

$$\mathbb{E}\left[\left(\frac{Y - m(\mathbf{X})}{Y}\right)^2 \middle| \mathbf{X}\right], \; Y > 0. \tag{2}$$

Under the finiteness condition of the first two moments of Y given $\mathbf{X}$, Park and Stefanski [4] showed that the minimizer for the random function in (2) is expressed by

$$m(\mathbf{X}) = \frac{\mathbb{E}[Y^{-1}|\mathbf{X}]}{\mathbb{E}[Y^{-2}|\mathbf{X}]}. \tag{3}$$

Following the same arguments, we define the kernel estimator of the truncated relative error regression of m given in (3) and we study its asymptotic properties when the observations are weakly dependent. Specifically, we are interested in considering the concept of association, introduced by Esary, Proschan and Walkup [1].

A set of finite family of random variables $Y = (Y_1, Y_2, \dots, Y_N)$ is said to be associated if for every pair of functions $h_1(.)$ and $h_2(.)$ from $\mathbb{R}^N$ to $\mathbb{R}$, which are non decreasing componentwise,

$$Cov(h_1(Y), h_2(Y)) \geq 0$$

whenever this covariance exists. Hamrani and Guessoum [2] have considered the classical kernel regression estimation under association condition, for which they established the strong uniform convergence with a rate.

We conduct a simulation study to assess the performance of these estimators when the observations are associated. This study will show and compare the behavior of two estimators for :

- different percentages of truncation;
- several observed sample size;
- presence and absence of outliers.

1. Esary J. Proschan F. Walkup D. Association of random variables with applications. Ann. Math. Statist., 1967, 38, 1, 1466-1476.
2. Guessoum Z. Hamrani F. Convergence rate of the kernel regression estimator for associated and truncated data. J. Nonparametric Statist., 2017, 29, 2, 425-446.
3. Ould-Saïd E. Lemdani M. Asymptotic properties of a nonparametric regression function estimator with randomly truncated data. Ann.Inst.Statist.Math., 2006, 58, 357-378.
4. Park H., Stefanski L. A. Relative-error prediction. Statist. Probab. Lett., 1998, 40, 227–236.
5. Woodroofe M. Estimating a distribution function with truncated data. Ann. Statist., 1985, 13,1, 163-177.