ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2019, Vol: 20, pp. 1 - 12, DOI: 10.18038/estubtda.632694

SEMIPARAMETRIC REGRESSION ESTIMATES BASED ON SOME TRANSFORMATION TECHNIQUES FOR RIGHT-CENSORED DATA

Dursun AYDIN [1,*], Ersin YILMAZ [1]

[1] Department of Statistics, Faculty of Science, Muğla Sıtkı Koçman University, Muğla, Turkey

ABSTRACT

In this paper, we introduce three different data transformation approaches such as synthetic data transformation ([1]; [2]; [3]), Kaplan-Meier weights ([4]; [5]; [6]) and k-nearest neighbor (kNN) imputation method ([7]) which are commonly used in censored data applications. The aforementioned approaches are particularly useful when one deals with censored data. The key idea expressed here is to find the smoothing spline estimates for the parametric and nonparametric components of a semiparametric regression model with right censored data. The estimation is then carried out based on the modified (or transformed) data set obtained via these transformation techniques. In order to compare the outcomes of three approaches in semi-parametric regression setting, we carried out a simulation study. According to the results of the simulation, it can be said that the Kaplan-Meier weights has been very successful in dealing with censored observations.

Keywords: Right-censored data, Semi-parametric regression, Imputation, synthetic data, Kaplan-Meier weights

1. INTRODUCTION

Right-censored data is a common notion that reveals in various applied fields. In general, datasets are composed of missing or censored observations for many different reasons such as sudden death, width drawn from the study, equipment failure, etc. In statistics literature, it is possible to encounter this type of data, especially in medical fields. It can be said that one of the most important problems that disrupts the quality of data is censored observations. Moreover, ordinary statistical methods cannot be applied directly to these censored data sets.

As indicated before, the focus of this article is to compare the methods used to deal with censored datasets. In the semi-parametric regression setting, three different transformation techniques are used to deal with censoring observations. After overcoming the censorship problem, the components of the semi-parametric regression model defined in the model (1.1) are estimated by smoothing spline method. Therefore, three different estimates for the components of the model (1.1) are obtained and a comparison of these estimates is also provided in this paper.

Among the data transformation techniques considered in this article, Kaplan Meier weights method is proposed by [5] and developed by Stute ([6], [8] and [9]); synthetic data transformation is discussed by various authors such as [1], [2] and [3]; the *k*NN method is proposed [7]. Note also that [10] studied the *k*NN for imputation of microarray data.

Considering the semiparametric regression model

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + f(t_i) + \varepsilon_i, \;\; i = 1, \dots, n \tag{1.1}$$

where Y_i's are the values of response variable, $\boldsymbol{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ is known p-dimensional explanatory variables vector, t_i's are values of an extra univariate explanatory variable, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a vector of regression coefficients to be estimated, $f(.)$ is an unknown smooth function

*Corresponding Author: duaydin@hotmail.com
Received: 24.09.2019 Published: 16.12.2019

and $\varepsilon_i's$ are the independent random errors with mean zero and finite variance σ^2. In a matrix and vector form, the model (1) is described as

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \tag{1.2}$$

where $\mathbf{Y} = (y_1, \dots, y_n)^T$, $\mathbf{X} = \left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_p\right)^T$, $\mathbf{f} = \left(f(t_1), \dots, f(t_n)\right)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$. It should also be that the first term on the right side of the semiparametric regression model stated in the equation (1.2) shows the parametric component ($\mathbf{X\beta}$), while the second term shows the nonparametric component ($\mathbf{f}$). For more details on the model (1.1), see [11] among others

In this paper, our primary interest is in estimating the vector parameter $\boldsymbol{\beta}$ and function $f(.)$ in the model (1.1) when the Y_i's are observed incompletely and right censored by a random variable C_i, i = 1, 2, . . . , n, but $\boldsymbol{x}_i^T$ and t_i are observed completely. In this case, instead of observing Y_i, we now observe the pair of values (Z_i, δ_i), i = 1, . . . , n such that

$$Z_i = \min(Y_i, C_i) \text{ and } \delta_i = I(Y_i \leq C_i) \tag{1.3}$$

where Z_i's are the values of observed response variable with unknown distribution K and δ_i's are the values of censoring indicator function that contains the censoring information. If ith observation is censored, $\delta_i = 0$ otherwise $\delta_i = 1$. Then, the model (1.2) can be rewritten as follows

$$\mathbf{Z} = \mathbf{X\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \tag{1.4}$$

where $\mathbf{Z} = (z_1, \dots, z_n)^T$ denotes the values of observed response variable according to censorship. Throughout this paper we assume that model (1.4) is considered. To ensure identifiability of the model (1.4), we need to make some specification assumptions on the response, censoring and explanatory variables and their dependence relationships. For this purpose, we consider the assumptions expressed in the study of [6]:

A1. Y_i and C_i are independent.
A2. $P(Y \leq C|Y, X) = P(Y \leq C|Y)$.

It should be emphasized that Y and C also have unknown distributions F and G, respectively. A1 is the assumption of a standard independence condition to provide identifiability of the model (4). Assumption A2 is required to let for a dependency between explanatory variables and censoring variable C

This paper organized as follows. Smoothing spline method is introduced in section 2. In section 3, solutions for censorship are expressed which are synthetic data transformation, Kaplan-Meier weights, kNN imputation methods respectively. In section 4, statistical properties of the estimators are given and evaluation measures used for comparison are presented. A simulation study and real data work are shown and presented the results in section 5 and 6 respectively. In section 7, promotion of the web application is made and usage information is given. Finally, conclusions are presented in section 8.

2. SMOOTHING SPLINES

Smoothing splines is a commonly used method to estimate the vector $\boldsymbol{\beta}$ and the unknown function $f(.)$ in a semiparametric regression model with right censored data. Let the ordered unique values among $t_1, \dots, t_n$ be denoted by $v_1, \dots, v_q$. The connection between v_i's and t_i's also provide a $(n \times q)$ dimensional incidence matrix $\mathbf{N}$ with entries $N_{ij} = 0$, if $v_j \neq t_i$ and $N_{ij} = 1$ otherwise (where $i = 1, \dots, n$ and $j = 1, \dots, q$). Note that the values of t_i are not identical so that we need focus on $q \geq 2$. Then, fitting a semiparametric model with censored data can make use of a univariate smoothing spline

method. Thus, use of a penalized residuals sum of squares approach would lead to choosing the vector parameter $\boldsymbol{\beta}$ and function $f(.)$ that minimize the criterion ([13]):

$$L(\boldsymbol{\beta};\mathbf{f}) = (\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\mathbf{N}\mathbf{f})^T(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\mathbf{N}\mathbf{f}) + \lambda\int_a^b f''(t)^2dt, \tag{2.1}$$

for a constant $\lambda > 0$, called a smoothing parameter. The first part in the right hand of the equation (2.1) penalizes the luck of fit, whereas the second part ($\int_a^b f''(t)^2dt$) puts a penalty on the roughness of the function.

Using properties of the smoothing splines, in the light of Green and Silverman (1994), the penalty term $\int_a^b f''(t)^2dt$ can be defined as $\mathbf{f}^T\mathbf{K}\mathbf{f}$. Consequently, the penalized criterion in the equation (2.1) can be rewritten as

$$L_0(\boldsymbol{\beta};\mathbf{f}) = (\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\mathbf{N}\mathbf{f})^T(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\mathbf{N}\mathbf{f}) + \lambda\mathbf{f}^T\mathbf{K}\mathbf{f} \tag{2.2}$$

where $\mathbf{K} = \mathbf{Q}^{\mathrm{T}}\boldsymbol{R}^{-1}\mathbf{Q}$ is a $(q \times q)$ dimensional symmetric and positive definite penalty matrix, $\mathbf{Q}$ and $\mathbf{R}$ are tri-diagonal matrices with dimensions $(q-2)\times(q)$ and $(q-2)\times(q-2)$ respectively. Elements of these matrices are given by

$$Q_{jj} = \frac{1}{h_j}, Q_{j,j+1} = -\left(\frac{1}{h_j}+\frac{1}{h_{j+1}}\right), Q_{j,j+2} = \frac{1}{h_{j+1}}$$
$$R_{j-1,j} = R_{j,j-1} = \frac{h_j}{6},\ R_{jj} = \left(h_j + h_{j+1}\right)/3$$

where $h_j = v_{j+1} - v_j, j = 1,\dots,q-1$. After some algebraic calculations, the smoothing spline estimators $\widehat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ of the vectors $\boldsymbol{\beta}$ and $\mathbf{f}$ that minimize the criterion (2.2) can be obtained as follows

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{Z} \tag{2.3}$$

and

$$\hat{\mathbf{f}} = \mathbf{S}_\lambda\left(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{2.4}$$

where $\mathbf{S}_\lambda = \mathbf{N}(\mathbf{N}^T\mathbf{N}+\lambda\mathbf{K})^{-1}\mathbf{N}^T$ is a positive definite smoothing matrix. If, also, the values of t_i are distinct and already ordered, $\mathbf{N} = \mathbf{I}$, and in this case $\mathbf{S}_\lambda = (\mathbf{I}+\lambda\mathbf{K})^{-1}$.

Using equations (2.3)-(2.4), the fitted values can be obtained by

$$\hat{\mathbf{Z}} = \left(\mathbf{X}\widehat{\boldsymbol{\beta}}+\hat{\mathbf{f}}\right) = \mathbf{H}\mathbf{Z} = E[Z|x,t] \tag{2.5}$$

for

$$\mathbf{H} = \mathbf{S}_\lambda + (\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X}(\mathbf{X}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X})^{-1}\boldsymbol{X}^T(\mathbf{I}-\mathbf{S}_\lambda) \tag{2.6}$$

See, Aydin and Yılmaz, (2018) for more detailed discussions.

As mentioned earlier, this article is designed to estimate the observations of the right censored response. If the aforementioned observations are directly estimated by the equations (2.3) - (2.4), this situation leads to a larger bias in terms of parameter estimates. Therefore, in practice, it is not appropriate to use these equations for estimating the components of the semi-parametric model. In this sense, the censoring effect should be included in the estimation process. The basic idea is that ono needs to adjust for the censoring effect by transforming the data in an unbiased way. To achieve this aim, three transformation methods are discussed in the following section.

3. DATA TRANSFORMATION TECHNIQUES

The key idea of this section is that in order to handle censored responses one needs to include the effect of censored data into the estimation procedure. In this context, to establish the regression model (1.2) based on transformed data, we consider three alternative approaches that take censoring into account, such as Kaplan-Meier weights, synthetic data and *k*NN imputation, as expressed in the previous sections.

3.1. Kaplan-Meier Weights

We begin by adapting the smoothing spline method based on censored response observations. To handle censored observations we use Kaplan-Meier (K-M) weights discussed in the study of [6]. In the context of smoothing spline, the penalized residuals sum of squares minimizing expression in (2.2) is modified by the following way

$$L_1(\boldsymbol{\beta};\mathbf{f}) = (\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\mathbf{N}\mathbf{f})^T\boldsymbol{W}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\mathbf{N}\mathbf{f}) + \lambda\mathbf{f}^T\mathbf{K}\mathbf{f} \tag{3.1}$$

where $\mathbf{Z} = (z_{(1)},\ldots,z_{(n)})'$ is a vector that contains the ordered values of observed response variable $Z = \min(Y, C)$ where C is a proper censoring variable, as stated before, $\mathbf{X}$ is a matrix of parametric explanatory variables related to ordered vector $\mathbf{Z}$, and $\boldsymbol{W}$ is a $n \times n$ diagonal matrix that denotes the K-M weights associated to $Z_{(i)}$. The diagonal elements of this matrix are computed by

$$w_{(i)} = \hat{F}\big(Z_{(i)}\big) - \hat{F}\big(Z_{(i-1)}\big) = \frac{\delta_{(i)}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(i)}} \tag{3.2}$$

where $\delta_{(i)}$ denotes the value of censoring indicator associated with ordered values $Z_{(i)}$'s. It should be emphasized that the K-M weights defined in (3.2) can also be computed as the contribution of Kaplan-Meier estimator [4] $\hat{F}$ of the distribution function F of response observations Y_i's at each ordered value $Z_{(i)}$.

Hence, the estimators $\widehat{\boldsymbol{\beta}}_{KM}$ and $\hat{\mathbf{f}}_{KM}$ minimizing the equation (3.1) can be obtained, respectively, as

$$\widehat{\boldsymbol{\beta}}_{KM} = (\mathbf{X}^T(\mathbf{I}-\mathbf{S}_{W\lambda})\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{S}_{W\lambda})\mathbf{W}\mathbf{Z} \tag{3.3}$$

and

$$\hat{\mathbf{f}}_{KM} = \mathbf{S}_{W\lambda}\big(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\beta}}_W\big) \tag{3.4}$$

where $\mathbf{S}_{W\lambda} = \mathbf{N}(\mathbf{N}^T\mathbf{W}\mathbf{N}+\lambda\mathbf{K})^{-1}\mathbf{N}^T\mathbf{W}$ is a smoothing matrix for incidence $\mathbf{N}$ and the K-M weights based on censored observations.

According to the equations (3.3)-(3.4), the fitted values of censored response variable can be calculated as follows

$$\hat{\mathbf{Z}}_{KM} = \big(\mathbf{X}\widehat{\boldsymbol{\beta}}_W + \hat{\mathbf{f}}_W\big) = \mathbf{H}_{KM}\mathbf{Z} = E[Z|x,t] \tag{3.5}$$

for

$$\mathbf{H}_{KM} = \mathbf{S}_{W\lambda} + (\mathbf{I}-\mathbf{S}_{W\lambda})\mathbf{X}[\mathbf{X}'(\mathbf{I}-\mathbf{S}_{W\lambda})\mathbf{W}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(\mathbf{I}-\boldsymbol{S}_{W\lambda}) \tag{3.6}$$

As can be seen herein, the smoothing parameter λ plays a critical role for estimating the vectors $\boldsymbol{\beta}$ and $\mathbf{f}$. For these purposes, the improved version of the Akaike information criterion (AIC_c) proposed by [12] is used, given by

$$AIC_c(\lambda) = 1 + \log[\|(\mathbf{S}_\lambda-\mathbf{I})\mathbf{Z}\|^2/n] + [\{2tr(\mathbf{S}_\lambda)+1\}/n - tr(\mathbf{S}_\lambda) - 2]. \tag{3.7}$$

Hence, the value of λ that minimizes the AIC_c is determined as optimum smoothing parameter.

3.2. Synthetic Data

The main problem in the case of censored data is that the observed variable Z and the variable of interest Y does not have the same expectation value. In order to provide this, we will use some unbiased transformation based on (Z, δ). In general, this transformation procedure in an unbiased way is referred to as the synthetic data approach (see, [14]). Also, several transformation approaches in the literature are discussed by authors including [1], [2]. In this study, we consider the transformation suggested by [2], which is given by

$$Z_{iG} = \frac{\delta_i Z_i}{1-G(Z_i)} = \frac{\delta_i Z_i}{\bar{G}(Z_i)} \tag{3.8}$$

where G is the distribution function of censoring variable C, as defined in assumption A2. However, in practical applications, G is unknown, and therefore transformation stated in (3.8) can not be calculated without estimating the distribution G. To overcome this problem, Koul et al., (1981) suggested replacing G in (3.8) by its Kaplan–Meier estimator [4] $\hat{G}$, given by

$$\hat{G}(t) = 1 - \prod_{i=1}^{n} \left(\frac{n-i}{n-i+1}\right)^{I[Z_{(i)} \le s, \delta_{(i)}=0]}, \ (t \ge 0) \tag{3.9}$$

where $(Z_{(i)}, \delta_{(i)}), i = 1,2, \dots, n$ are the ordered observations of $(Z_{(i)}, \delta_{(i)})$ ordered on the $Z_{(i)}$. That is, $z_{(1)} \le z_{(2)} \le, \dots, \le z_{(n)}$ are the ordered observations of Z and $\delta_{(i)}$ is the censoring indicator associated with $Z_{(i)}$.

Since statistical inferences for the parametric and nonparametric components of the semiparametric regression model expressed in (1.1) or (1.2) based on transformed data (Z_G, X, t), the basic requirement is that $E(Z_G|X,t) = E(Y|X,t) = X\boldsymbol{\beta} + \mathbf{f}$, ensuring that we estimate right components ([13]). Accordingly, we employ the transformed data (Z_G, X, t) instead of (Y, X, t) to estimate the parameters $\boldsymbol{\beta}$ and the function $f(.)$ can be estimate based on smoothing spline.

Using the equations (2.3)-(2.4), and replacing $\mathbf{Z}_{\hat{G}}$ by $\mathbf{Z}$, leads to the following smoothing spline estimators ($\widehat{\boldsymbol{\beta}}_{SD}$ and $\hat{\mathbf{f}}_{SD}$) based on synthetic data (*SD*) of ($\boldsymbol{\beta}$ and $\mathbf{f}$), respectively, as

$$\widehat{\boldsymbol{\beta}}_{SD} = (\mathbf{X}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Z}_{\hat{G}} \tag{3.10}$$

and

$$\hat{\mathbf{f}}_{SD} = \mathbf{S}_\lambda\left(\mathbf{Z}_{\hat{G}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\hat{G}}\right) \tag{3.11}$$

where $\mathbf{S}_\lambda$ is the spline smoother matrix, as defined in (2.4). Note also that in equations (3.10)-(3.11), $\mathbf{Z}_{\hat{G}}$ denotes the estimated *SD* vector. In other words, $\mathbf{Z}_{\hat{G}} = (z_{1\hat{G}}, \dots, z_{n\hat{G}}) = \frac{\delta_i Z_i}{1-\hat{G}(Z_i)} = Z_{i\hat{G}}$. Hence, according to the equations (3.10)-(3.11), the fitted values based on synthetic data can be obtained as

$$\hat{\mathbf{Z}}_{SD} = \left(\mathbf{X}\widehat{\boldsymbol{\beta}}_{\hat{G}} + \hat{\mathbf{f}}_{\hat{G}}\right) = \mathbf{H}_{SD}\mathbf{Z}_{\hat{G}} = E[Z|x,t] \tag{3.12}$$

for

$$\mathbf{H}_{SD} = \mathbf{S}_\lambda + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}[\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \boldsymbol{S}_\lambda) \tag{3.13}$$

3.3. *k*-NN Imputation

In this section we consider the use of *k*-NN imputation algorithm to estimate and substitute censoring data. The main idea behind using *k*-NN is that censoring values in a sample can be imputed (or approximated) by using values computed from the *k*-NN method. It should be noted that in this method, a censoring value is imputed by either a value measured for the neighbor or the average of measured values for multiple neighbors. Some important benefits of this technique are:

- Method is a free technique from distribution. This feature provides an important advantage for dealing with data that does not fit any distribution family.
- Right-censored data points are filled with actual observations, not synthetic or constructed values.
- Unlike synthetic data transformation and K-M weights, the k-NN method uses explanatory variables to provide additional information in completing censored data points.
- One of the most important properties of *k*-NN imputation is completely nonparametric method and it does not include any assumptions about the relationship between observation pairs (X_i, Y_i) or (X_i, Z_i), $i = 1, \dots, n$.

The *k*-NN method can work with discrete and continuous variables. It uses most frequently used data point among *k*-closest neighbors. For continuous attributes, it uses the average value of *k*- closest neighbours. In this study, Minkowski norm, which is a distance measurement, is used to evaluate the distance between the related observation and neighbors. The distance measure expressed this norm is given by

$$d_M(X,Y) = \left(\sum_{i=1}^{n} |X_i - Y_i|^p\right)^{\frac{1}{p}} \tag{3.14}$$

Note that Minkowski distance is also known as a *p-norm* and it turns into Euclidean distance when $p = 2$ and Manhattan distance when $p = 1$. Here, Euclidean distance is used to decide similarity between instances.

Table 2. Algorithm for *k*-NN imputation method

Algorithm1: *k*-NN imputation for right censored data

Input: Right − censored data set S
Censoring indicator δ_i associated with S
Number of nearest neigbours k
Values of predictor variable x_i related with S
Output: Imputed dataset $\mathbf{y}^{knn}$

```
begin
for (i = 1 to n) do
    if(δ_i = 0) do  (if data point is censored)
        for(j = 1 to n) do
            Find the distances between x_j and x_i for each censored data point with (3.14)
    Sort the distances from small to large
    for (j = 1 to k) do
        Take the first uncensored k − values of z_i associated to sorted distances
        Calculate the ith imputed value (y_i^kNN) with average of nearest k − records of y_i
Replace the imputed values (y_i^kNN) with censored data points (z_i, δ_i = 0) in censored data set z = (z_1, ..., z_n)
Return Y_kNN = (y_1^kNN, ..., y_n^kNN)^T
end
```

As stated in the previous section, using the equations (2.3)-(2.4) and changing $\mathbf{Y}_{kNN}$ by $\mathbf{Z}$, we obtain the smoothing spline estimators $\widehat{\boldsymbol{\beta}}_{kNN}$ and $\hat{\mathbf{f}}_{kNN}$, based on *k*-NN imputation method, respectively, as

$$\widehat{\boldsymbol{\beta}}_{kNN} = (\mathbf{X}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Y}_{kNN} \tag{3.15}$$

and

$$\hat{\mathbf{f}}_{kNN} = (\mathbf{N}^T\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}^T\left(\mathbf{y}_{knn} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{kNN}\right) \tag{3.16}$$

From the equations (3.15)-(3.16), the fitted values on *k*-NN imputation can be computed as

$$\widehat{\mathbf{Y}}_{kNN} = \left(\mathbf{X}\widehat{\boldsymbol{\beta}}_{kNN} + \hat{\mathbf{f}}_{kNN}\right) = \mathbf{H}_{kNN}\mathbf{Y}_{kNN} = E[Z|x,t] \tag{3.17}$$

for

$$\mathbf{H}_{kNN} = \mathbf{S}_\lambda + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}[\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \boldsymbol{S}_\lambda) \tag{3.18}$$

4. STATISTICAL PROPERTIES OF THE ESTIMATORS

This section provides the some statistical properties of the estimators explained herein. To see the calculations of each estimator, we first expand $\widehat{\boldsymbol{\beta}}_{KM}$ in Equation (3.3) by the matrix and vector form of $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{f}} + \tilde{\boldsymbol{\varepsilon}}$ such that

$$\widehat{\boldsymbol{\beta}}_{KM} = \left(\mathbf{X}^T\mathbf{W}\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\mathbf{W}\widetilde{\mathbf{Y}} = \boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{W}\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\mathbf{W}\tilde{\mathbf{f}} + \left(\mathbf{X}^T\mathbf{W}\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\mathbf{W}\tilde{\boldsymbol{\varepsilon}} \tag{4.1}$$

where $\widetilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{S}_{W\lambda})\mathbf{Z}$, $\widetilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S}_{W\lambda})\mathbf{X}$, $\tilde{\mathbf{f}} = (\mathbf{I} - \mathbf{S}_{W\lambda})\mathbf{f}$, and $\tilde{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{S}_{W\lambda})\boldsymbol{\varepsilon}$. Accordingly, the bias and variance-covariance matrix of the estimator (4.1) can be expressed, respectively, as

$$Bias\left(\widehat{\boldsymbol{\beta}}_{KM}\right) = E\left(\widehat{\boldsymbol{\beta}}_{KM}\right) - \boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{W}\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\mathbf{W}\tilde{\mathbf{f}} \tag{4.2}$$

$$Var\left(\widehat{\boldsymbol{\beta}}_{KM}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{W}\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{I} - \mathbf{S}_{W\lambda})^2\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\widetilde{\mathbf{X}}\right)^{-1} \tag{4.3}$$

In a similar manner, expanded form of the $\widehat{\boldsymbol{\beta}}_{SD}$ in (3.10) is

$$\widehat{\boldsymbol{\beta}}_{SD} = \left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\widetilde{\mathbf{Y}} = \boldsymbol{\beta} + \left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\tilde{\mathbf{f}} + \left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\tilde{\boldsymbol{\varepsilon}} \tag{4.4}$$

where $\widetilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{Z}_{\hat{G}}$ and $\widetilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{X}$. According to these ideas, the bias and variance of the estimator $\widehat{\boldsymbol{\beta}}_{SD}$ in (4.4) can be defined, respectively, as

$$Bias\left(\widehat{\boldsymbol{\beta}}_{SD}\right) = E\left(\widehat{\boldsymbol{\beta}}_{SD}\right) - \boldsymbol{\beta} = \left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\tilde{\mathbf{f}} \tag{4.5}$$

$$Var\left(\widehat{\boldsymbol{\beta}}_{SD}\right) = \sigma^2\left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S}_{\lambda})^2\mathbf{X}\left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1} \tag{4.6}$$

Finally, as defined in the above statements, the estimator $\widehat{\boldsymbol{\beta}}_{kNN}$ based on *k*-NN imputation, given in (3.15), can be expanded as follows

$$\widehat{\boldsymbol{\beta}}_{kNN} = \left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\widetilde{\mathbf{Y}} = \boldsymbol{\beta} + \left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\tilde{\mathbf{f}} + \left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\tilde{\boldsymbol{\varepsilon}} \tag{4.7}$$

where $\widetilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{Y}_{kNN}$ and $\widetilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{X}$. Hence, the bias and variance–covariance matrix of the estimator $\widehat{\boldsymbol{\beta}}_{kNN}$ stated in (4.7) are

$$Bias\left(\widehat{\boldsymbol{\beta}}_{kNN}\right) = E\left(\widehat{\boldsymbol{\beta}}_{kNN}\right) - \boldsymbol{\beta} = \left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\tilde{\mathbf{f}} \tag{4.8}$$

$$Var\left(\widehat{\boldsymbol{\beta}}_{kNN}\right) = \sigma^2\left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S}_{\lambda})^2\mathbf{X}\left(\mathbf{X}^T\widetilde{\mathbf{X}}\right)^{-1} \tag{4.9}$$

As can be seen from the equations stated above, the variance matrices are not practical due to the quantity of unknown σ^2. Therefore, we only need to derive the estimate of σ^2 in order to construct the aforementioned variance-covariance matrices. As in linear regression analysis, an estimate of variance can be computed by residual sum of squares (*RSS*)

$$RSS = (\mathbf{Z} - \mathbf{H}_{\lambda}\mathbf{Z})^T(\mathbf{Z} - \mathbf{H}_{\lambda}\mathbf{Z}) = \|(\mathbf{I} - \mathbf{H}_{\lambda})\mathbf{Z}\|^2 \tag{4.10}$$

where $\mathbf{H}_{\lambda}\mathbf{Z} = \hat{\mathbf{Z}}$ is the fitted values for censored response observations, and $\mathbf{H}_{\lambda}$ is the hat matrix obtained from smoothing spline based on any data transformation data technique. Hence, the estimate of σ^2 can be obtained as

$$\hat{\sigma}^2 = \frac{RSS}{trace(\mathbf{I}-\mathbf{H}_{\lambda})^2} = \frac{\|(\mathbf{I}-\mathbf{H}_{\lambda})\mathbf{Z}\|^2}{n-2\times trace(\mathbf{H}_{\lambda})-trace(\mathbf{H}_{\lambda}{}^T\times\mathbf{H}_{\lambda})} \tag{4.11}$$

where$trace(\mathbf{I} - \mathbf{H}_{\lambda})^2$ is called as degrees of freedom (*df*). It should be emphasized that if we use the K-M weights method, the computation of $\mathbf{H}_{KM}$ stated in (3.6) instead of $\mathbf{H}_{\lambda}$ pointed in equations (4.10)

and (4.11) is needed. In a similar fashion, when we adopt the synthetic data and *k*-NN method, we need to compute the $\mathbf{H}_{SD}$ in (3.13) and $\mathbf{H}_{kNN}$ in (3.18) matrices, respectively.

Especially in simulation experiments, to evaluate the outputs and make inferences about parametric component of the model, bias and variances of estimators stated in this study are illustrated in equations (4.2-4.3), (4.5-4.6) and (4.8-4.9). Also, variance of the model in terms of errors is given in (4.11). In addition to ideas, to measure performance of the nonparametric component of the model, the mean square error (MSE) is used that can be calculated as follows

$$MSE = n^{-1}\sum_{i=1}^{n}\left(f(t_i) - \hat{f}(t_i)\right)^2 \tag{4.12}$$

5. SIMULATION STUDY

A Monte Carlo simulation study is conducted to indicate the impact of censoring and to assess the finite sample behaviours of the estimators based on three data transformation approaches. In our context, we first generate dataset $(Y_i, \mathbf{x}_i, t_i)$ from the following semiparametric regression model

$$Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + f(t_i) + \varepsilon_i, \;\; i = 1, \ldots, n \tag{5.1}$$

where $\mathbf{x}_i = (x_{1i}, x_{2i}, x_{3i})$ is generated from uniform distribution, $\boldsymbol{\beta} = (-2,\ 0.5,\ 3)^T$, the nonparametric function $f(t_i)$ is obtained by $f(t_i) = t_i \sin(t_i)$, where $t_i = 10(i - 0.5)/n$, and $\varepsilon_i \sim N(0, \sigma^2)$. Censoring variable C_i is determined by normal distribution with censoring levels (*C.L*) at 5%, 20%, and 40%. The censoring indicator is then defined as $\delta_i = I(Y_i \leq C_i)$ and the observations of censored response variable are created as $Z_i = \min(Y_i, C_i)$. For each *C.L* in simulation experiments, we generated 1000 random samples of size $n = 75, 150, \text{and } 200$.

5.1. Evaluation of Parametric Component

Estimates of $\boldsymbol{\beta} = (-2,\ 0.5,\ 3)'$s constructing the parametric components of the model (5.1) are summarized in the following tables and figures for each censoring levels and samples. The basic outcomes from the simulation experiments are summarized in Table 1.

Table 1 shows the averaged biases and variances of parametric coefficients and variances of errors obtained from smoothing spline estimators based on tree data transformations techniques. As can be seen from the Table 1, the best scores obtained from estimators are marked with bold color. Clearly, the estimator based on K-M weights has smaller bias, while the estimator based on SD has larger bias for all simulation configurations. From the Table 1, we also conclude that the estimator using K-M weights performs reasonably well with all sample size and censoring levels at 20% and 40% considered in terms of small bias and accurate inference based on parametric components of the model (5.1).

Table 1. Assessment of statistical estimations related to the parametric regression coefficients

		SD			K-M			*k*-NN		
C.L	*n*	$Bias(\hat{\beta})$	$Var(\hat{\beta})$	$\hat{\sigma}^2$	$Bias(\hat{\beta})$	$Var(\hat{\beta})$	$\hat{\sigma}^2$	$Bias(\hat{\beta})$	$Var(\hat{\beta})$	$\hat{\sigma}^2$
5	75	0.4826	0.3746	0.3745	**0.3648**	**0.1996**	0.4480	0.4579	0.3264	**0.3301**
	150	0.3318	0.1684	0.3666	**0.2497**	**0.0915**	0.3305	0.3178	0.1519	**0.3250**
	200	0.2755	0.1262	0.3392	**0.2080**	**0.0670**	0.3620	0.2677	0.1123	**0.3002**
20	75	0.8114	1.0423	1.0500	**0.4120**	**0.2463**	**0.6542**	0.6549	0.6626	0.6686
	150	0.5505	0.4740	1.0395	**0.2712**	**0.1111**	**0.5462**	0.4352	0.3039	0.6540
	200	0.4726	0.3491	0.9644	**0.2293**	**0.0807**	**0.5007**	0.3727	0.4128	0.6082
40	75	1.2688	2.3278	2.3824	**0.5100**	**0.3472**	**0.9356**	0.7632	0.8895	0.9435
	150	0.8350	1.0869	2.3518	**0.3492**	**0.1525**	**0.8929**	0.5135	0.4128	0.8980
	200	0.7258	0.7925	2.2147	**0.3062**	**0.1107**	**0.7869**	0.4457	0.3054	0.8293

In addition to the ideas given above, the k-NN has some good results for variance of the model with low censoring level, but when the censorship level is high, the K-M weights method gives better results. Note also that the SD has the worst performance scores in estimating the parametric component. Some common characteristics of three methods can be listed as follows: For larger sample sizes and lower censoring levels, the data transformation approaches work better, while the approaches perform worse for higher censoring levels and smaller samples.

Figure 1 shows different box plots displayed in the two panels. In each panel, "SD(5%), K-M(5%), and *k*-NN(5%) demonstrate the parametric biases of $\widehat{\boldsymbol{\beta}}$ from the semiparametric regression using smoothing spline based on data transformed by SD, K-M, and k-NN methods for censoring level at 5% and samples of size n=75 and 200, respectively. Similarly, the remaining box plots denote the parametric biases, but for censoring levels at 20% and 40%, respectively. The ordinate indicates the scale of the biases of regression coefficients.

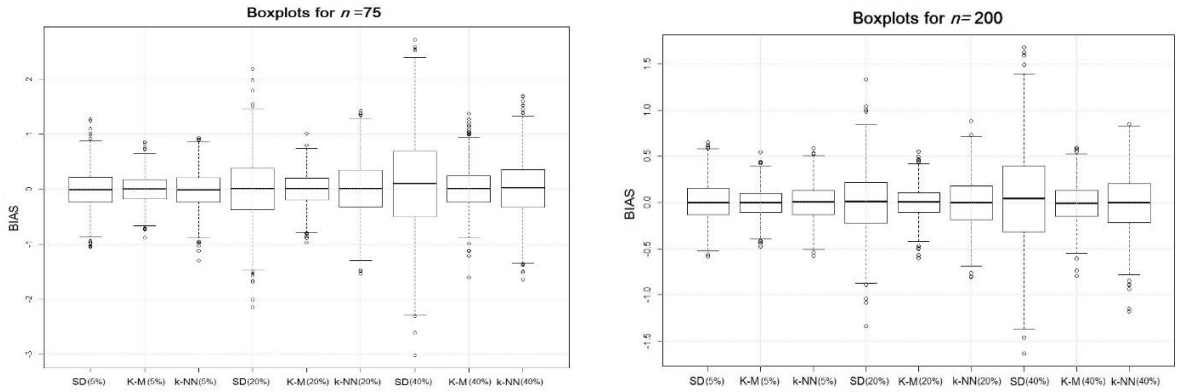

Figure 1. Boxplots of estimated regression coefficients for all censoring levels and samples of size n=75 and 200.

The results presented in Figure 1 are consistent with the results in Table 1. As shown in Figure 1, the two panels have similar box plots, but the boxes in the right panel represent values smaller biases than those in the left panel. The K-M weight method always has boxes that are narrower than others. In both panels, the SD does not perform well, especially at the high censoring level 40%. From the results of the parametric component of the semiparametric regression model, it can be said that the smoothing spline method can solve the censoring regression problems better by applying the K-M instead of applying the *k*-NN and SD methods to transform data. Furthermore, when the data are heavily censored (i.e., 40%), it seems that the variances from the K-M and *k*-NN methods remain stable, while the variance values obtained by the SD increase.

5.2. Evaluation of Nonparametric Component

As in the parametric components, we obtained 1000 estimates of the function f, which is the nonparametric part of model (5.1). In this context, 1000 replications are performed for each data transformation techniques and the MSE values are computed by (4.12) for each techniques and corresponding each $f(.)$ under the different censoring levels. The findings are summarized in Table 2.

According to Table 2, the K-M weights method gives the best performance to estimate the non-parametric component. In addition, at medium and high censorship levels, the K-M weights are significantly lower than the other two methods. In addition, the k-NN method has the worst performance in estimating the non-parametric component.

Table 2. Outcomes from the nonparametric component

C.L.(%)	*n*	SD	K-M	*k*-NN
5	75	1.2615	**1.0461**	1.2888
	150	1.1231	**1.0192**	1.1545
	200	1.0828	**1.0090**	1.1279
20	75	2.1807	**1.1377**	2.3944
	150	1.5267	**1.0418**	1.9642
	200	1.3771	**1.0296**	1.9047
40	75	4.0265	**1.3021**	4.5609
	150	2.2942	**1.1301**	4.0157
	200	2.0365	**1.1083**	3.8744

In our simulation study, because 27 different configurations are carried out, it is very hard to illus-trate all of them. Therefore, only four different configurations will be presented in Figure 2. The panels in this figure represent the smoothed curves together with a real function $f(t)$. In each graph, the smoothed curves, $f(SD)$, f(K-M), and $f(kNN)$, respectively, are estimates of $f(t)$ using smoothing spline based on SD, K-M, and *k*-NN approaches for different censoring levels and sample size.

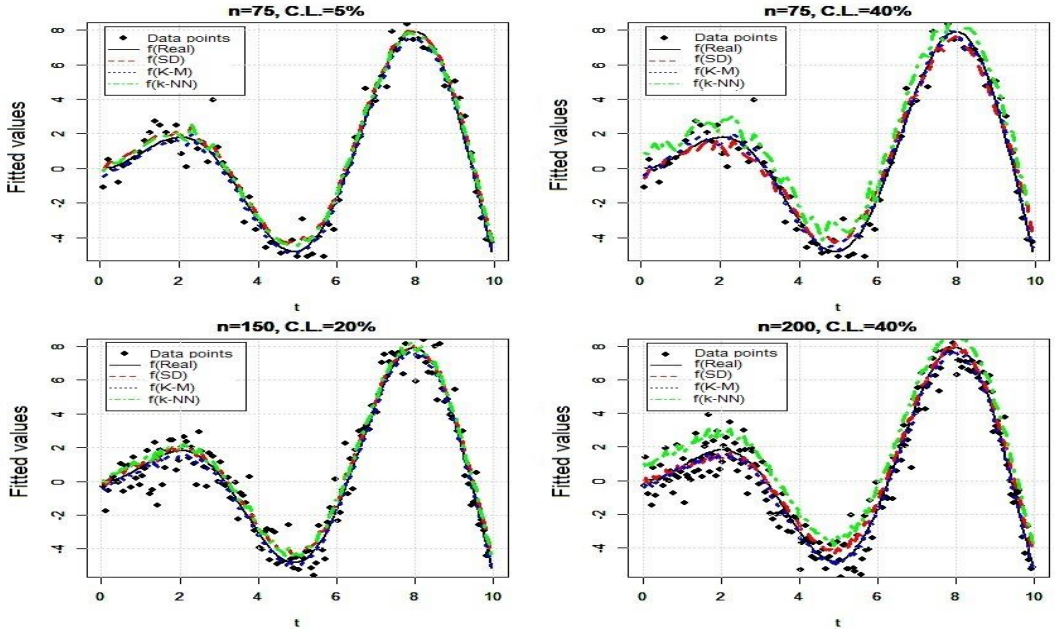

Figure 2. Fitted curves obtained from three methods for low (5%) and high (40%) censoring levels for all sample sizes.

The two upper panels are obtained by the same sample of size n = 75 to denote how the fitted curves are affected by the different censoring rates. From the Figure 2 we observe that *k*-NN (green line) is highly influenced by censored observations in the response variable. In the contrast, the K-M weights method gives a fitted curve really close to the real function for all simulation configurations. Of these methods, the SD displays quite poorly performance in estimating function $f(t_i)$. Note also that the fitted values gives a smoother curve than the small ones, especially for large sample sizes.

6. CONCLUSIONS AND RECOMMENDATIONS

In this paper, three common methods, SD, K-M weights and k-NN valuation method, were used to deal with right censored data and to see their contribution to the estimation procedure of the components of a semiparametric regression model with censored data. A Simulation study was conducted to compare the performance of these three methods. From the simulation results, the K-M weights method provides

better performance than the other two methods in terms of estimating both components of the semiparametric model.

A detailed assessment of the simulation results provides the significant findings for these data transformation methods. Specifically, the K-M weights method provides satisfactory performance for different simulation configurations, especially at high censoring levels. In this context, the K-M is followed by *k*-NN and SD methods, respectively, in terms of performance ranking. Conceptually, because the censoring values are given zero, and the expected values of real and censored responses will be different. However, the SD method only changes the non-censored response values to ensure that the actual (Y_i) and censored responses ($Y_{i\hat{G}}$) are equal in terms of the expected value, $\{E(Y_i|x_i, t_i) \cong E(Y_{i\hat{G}}|x_i, t_i)\}$ (also, see Koul *et al*., 1981). Probably, the main reason for the failure of the SD method compared to others may be that it only changes the observations of non-censored observations. When data points are inspected individually, it can be realized that synthetic data points are highly different from original response values.

In summary, the outcomes of this study show that one way to overcome the right-censored problems is to use the K-M weight method in semiparametric regression setting. In addition, *k*-NN valuation method can be used for low censoring levels. It should be noted that only simulation studies are conducted in this paper. To provide general validation of the results, the simulation study can be expanded and some real data sets can be applied.

REFERENCES

[1] J. Buckley, I. James, Linear Regression with Censored Data, Biometrika, 66(3), 429-436, 1979

[2] H. Koul, V. Susarla, J. Van Ryzin, Regression Analysis with Randomly Right-Censored Data, The Annals of Statistics, 1276-1285, 1981.

[3] S. Leurgans, Linear Models, Random Censoring and Synthetic Data, Biometrika, 74, 301-309, 1987.

[4] E.L. Kaplan, P. Meier, Nonparametric Estimation from Incomplete Observations, Journal of the American Statistical Association, 53(282), 457-481, 1958.

[5] R.G. Miller, Least Squares Regression with Censored Data, Biometrika, 63, 449-464, 1976.

[6] W. Stute, Consistent Estimation under Random Censorship when Covariables are Present, Journal of Multivariate Analysis, 45, 89-103, 1993.

[7] G.E.A.P.A. Batista and M.C. Monard, K-Nearest Neighbour as an Imputation Method: Experimental Results, Technical Report, ICMC-USP, ISSN-0103-2569, 2002.

[8] W. Stute, The Central Limit Theorem under Random Censorship. The Annals of Statistics, 2, 422-439, 1995.

[9] W. Stute, Nonlinear Censored Regression, Statistica Sinica, 9, 1089-1102, 1999.

[10] O. Troyanskaya, M. Cantor, Sherlock, G., P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics, 17(6), 520-525, 2001.

[11] P.J. Green, B.W. Silverman, Nonparametric Regression and Generalized Linear Model, Chapman & Hall, 1994.

[12] C.M. Hurvich, S. Simonoff, S. and C.L. Tsai, Smoothing Parameter Selection in Nonparametric Regression using an Improved Akaike Information Criterion. Journal of Royal Statistical Society B, 60(2), 271-293, 1998.

[13] D. Aydın, E. Yılmaz, Modified Spline Regression Based on Randomly Right-Censored Data: A Comparison Study, Communication in Statistics-Simulation and Computation, 47(9), 2587-2611, 2018.

[14] M. Talamakrouni, A.E. Gouch, I. Van Keilegom, Guided Censored Regression, Scandinavian Journal of Statistics-Theory and Applications, 42(1), 214-233, 2015.