JIRSS (2021)
Vol. 20, No. 01, pp 1-26
DOI:10.52547/jirss.20.1.1

Kernel Ridge Estimator for the Partially Linear Model under Right-Censored Data

Syed Ejaz Ahmed [1], **Dursun Aydın** [2], **and Ersin Yılmaz** [2]

[1] Brock University, Faculty of Mathematics and Science, Department of Mathematics and Statistics, Niagara Region, 1812 Sir Isaac Brock Way, St. Catharines, ON, L2S 3A1, Canada.
[2] Mugla Sitki Kocman University, Faculty of Science, Department of Statistics, 48000, Mugla, Turkey.

Received: 08/12/2020, Revision received: 29/01/2021, Published online: 03/04/2021

Abstract. *Objective*: This paper aims to introduce a modified kernel-type ridge estimator for partially linear models under randomly-right censored data. Such models include two main issues that need to be solved: multi-collinearity and censorship. To address these issues, we improved the kernel estimator based on synthetic data transformation and kNN imputation techniques. The key idea of this paper is to obtain a satisfactory estimate of the partially linear model with multi-collinear and right-censored using a modified ridge estimator. *Results*: To determine the performance of the method, a detailed simulation study is carried out and a kernel-type ridge estimator for PLM is investigated for two censorship solution techniques. The results are compared and presented with tables and figures. Necessary derivations for the modified semiparametric estimator are given in appendices.

Keywords. Kernel Smoothing, KNN Imputation, Multi-Collinear Data, Partially Linear Model, Ridge Type Estimator, Right-Censored Data.

Corresponding Author: Syed Ejaz Ahmed (sahmed5@brocku.ca)
Dursun Aydın (duaydin@hotmail.com)
Ersin Yılmaz (yilmazersin13@hotmail.com)

MSC: 62G05, 62G08, 62N01.

1 Introduction

A partially linear model includes multicollinearity as follows:

$$y_i = \sum_{j}^{p} x_{ij}\beta_j + g(z_i) + \varepsilon_i, \quad 1 \le i \le n, \tag{1.1}$$

where y_i represents the response values, x_{ij} denotes the values of covariates correlated with each other, β_j represents unknown regression coefficients for the parametric component, $g(z_i)$ is the nonparametric part of the model with single nonparametric covariate z_i and an unknown smooth function $g(.)$, and ε_i denotes the random error terms with constant variance σ_ε^2 and mean zero. To simplify this notation, the matrix and vector form of the model (1.1) can be written as follows

$$\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{g} + \boldsymbol{\varepsilon}, \tag{1.2}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ is the $(n \times 1)$ vector of response values, $\boldsymbol{X}^T = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is the $(n \times p)$ covariate matrix formed by observations of independent variables and assumed to be a full-ranked matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the $(p \times 1)$ vector of regression coefficients to be estimated, $\mathbf{g} = (g(z_1), g(z_2), \ldots, g(z_n))^T$ is the $(n \times 1)$ dimensional vector of nonparametric component, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ is the $(n \times 1)$ vector of random error terms with $E(\varepsilon) = 0$ and $E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right) = \sigma_i^2\boldsymbol{I}_n$. A number of researchers have studied the fitness of the model defined in (1.2). Examples of such studies include Schimek (2000), Ruppert et al. (2003), and Liang (2006).

The paper discusses two important problems that arises in model (1.2), multi-collinearity and censored response values.

Multi-collinearity is an increasingly common problem in recent years, as data collection methods have been developed and data structures have evolved into high dimensionality. It is well-established that standard regression phenomenon assumes that covariates are linearly uncorrelated. However, in many application areas such as biology, economics, industry, finance, medical studies, and especially recent work in bioinformatics, this assumption often breaks down because there are many more explanatory variables to model the response variable. Therefore, multi-collinearity

is an inevitable problem for any linear, nonparametric or semi-parametric modeling procedures.

It is quite common to have observations with missing values for response variables in applied fields such as survival analysis. These missing observations are sometimes unavoidable, due to the fact that individuals can withdraw from a study at any time. In general, censoring occurs when there is incomplete information about the survival time of some individuals in the study. The censored data problem, where the observed value of a variable is partially known, is related to the missing data problem. Ordinary regression techniques cannot be used directly in the fitting procedure of model (1.1) because censored data leads to biased estimates. In such cases, one typical approach is to impute, or "fill in", the missing observations; another is to transform response observations.

It should be emphasized that when there are many independent variables affecting the right-censored response observations, the multi-collinearity problem may arise in the setting of regression analysis. Specifically, this article seeks a solution for situations where both right-censorship and collinearity appear simultaneously in the data set. This is because, based on our experience, datasets have frequently contained both these problems in recent years.

This paper solves the multi-collinearity problem by using the kernel-type ridge estimator for partially linear models studied by Yüzbaşı et al. (2017). To overcome censorship, we use two methods: the synthetic data transformation method described in Koul et al. (1981) and the kNN imputation method studied by Batista and Monard (2002) and Ahmed et al. (2020). Note also that kernel ridge regression under censored data has been previously studied by Shim (2005) in a nonparametric setting. As is known, in ridge-type estimators, the shrinkage parameter plays a crucial role in accuracy of the estimation. Moreover, the bandwidth parameter of the kernel smoothing method controls the amount of penalty term in minimization criterion given in equation (3.1). Determining both the ridge and bandwidth parameters are therefore two additional issues covered in this paper. To achieve improved AIC_c, the criterion proposed by Hurvich et al. (1998) is used in the determination of both parameters.

In the context of right-censored data, several authors have studied estimation of the partially linear model (1.2). Some of these include Orbe et. al. (2003) and Liang and Zhou (2008). In addition, the estimation of right-censored response values was proposed by Kaplan-Meier (1958), and later Miller (1976) proposed Kaplan-Meier (K-

M) weights for the linear regression model estimation using a K-M estimator.

According to given information above, the contributions of this paper can be summarized as follows:

- A modified kernel-type ridge estimator is introduced to estimate the partially linear model with right-censored data.
- Two solution techniques with different fundamentals are used to solve the censorship problem. One of these is a synthetic data transformation based on the Kaplan-Meier estimator of distribution of censoring variable (Koul et al., 1981). The other is kNN imputation, a machine learning method. It is a fully nonparametric model that provides different perspectives on the estimation process.

The paper is arranged as follows: In section 2, the right-censored data concept, synthetic data transformation and kNN imputation methods are introduced with their important aspects. Section 3 presents the ridge-type kernel estimator for a censored, partially linear model based on both synthetic data and kNN imputation. In Section 4, some performance evaluation metrics are defined. A detailed Monte-Carlo simulation study and results are given in Section 5. Finally, concluding remarks are made in Section 6.

2 Right-Censored Data and Solution Techniques

This paper focuses on estimating the vector $\boldsymbol{\beta}$ of regression coefficients and the unknown smooth function $\mathbf{g}$ in model (1.2), while the response values of y_i are observed incompletely and censored from the right by a random censoring variablec_i, $i = 1, \ldots, n$. However, x_{ij} and z_i variables are completely observed. Consequently, right-censored triplets $\{x_{ij}, z_i, y_i\}$ are turned into incomplete observations $\{x_i, z_i, t_i, \delta_i\}_{(i=1)}^{n}$ as follows:

$$t_i = \min(y_i, c_i) \text{ and } \delta_i = \boldsymbol{I}(y_i \leq c_i), 1 \leq i \leq n, \tag{2.1}$$

where t_i represents the right-censored response values that are updated according to existence of the censorship with distribution J, c_i denotes the values of the censoring variable with distribution G and δ_i values are binary scores of the censoring indicator, which carries the information of the existence of the censorship. Note also that it is assumed that y_i and c_i are independent random variables with unknown distributions F and G, respectively.

Now, it can be clearly seen that model (1.2) cannot be estimated directly using an ordinary semiparametric modelling procedure due to incomplete responses. Although

t_i's are obtained according to censorship, they do not involve the effect of censored observations which cause biased estimates. In order to overcome this problem, there are some solution methods in the literature. The most important methods can be divided into three categories: transformation (Koul et al. 1981), imputation (Batista and Monard 2002; Yenduri and Iyengar, (2007), and weights (Miller, 1976; Stute, 1993; Orbe et al. 2003). In this paper, the first two types are considered because, for the purposes of this paper, synthetic data transformation and imputation techniques are far easier, in terms of computation and formulation, than a weights-based method.

In the analysis of right-censored data, there are two fundamental assumptions (A1 and A2) which provide consistent and accurate estimates. These assumptions have previously been described in detail by Stute (1993), Koul et al. (1981) and Miller (1976). Therefore, they will be introduced only cursorily here:

A1. Completely observed y_i's and c_i's should be independent.

A2. $P(y_i < c_i|y_i, x_i, z_i) = P(y_i \leq c_i|y_i)$.

A1 is the ordinary assumption for survival analysis and allows acquisition of a meaningful model. If A1 is broken, then to obtain an accurate model, significantly more information is needed. On the other hand, A2 means that given time of failure, values of explanatory variables cannot provide any further information about response data points. Note that both A1 and A2 cannot assume the independency between explanatory variables, but in classical regression phenomenon, correlation between x_i's is not preferred, as explained in the introduction section. It should be emphasized that this paper's interests lie in two problems: right-censored responses and multi-collinearity. This section prepares us to handle censorship problem. Multi-collinearity under censorship is discussed in Section 3.

2.1 Synthetic Data Transformation

Synthetic data transformation is a widely used technique to overcome censorship problem in the data in the modelling context. There are a number of modified synthetic data transformations in the literature, some of which are mentioned in Section 1. However, this paper will obtain synthetic responses based on Koul et al. (1981). Accordingly, data transformation is computed by

$$t_{iG} = \frac{\delta_i t_i}{1 - G(t_i)} = \frac{\delta_i t_i}{\bar{G}(t_i)}, \tag{2.2}$$

where G is the distribution of c_i, as explained before. After computation of the synthetic responses, model (1.2) can be rewritten with synthetic data as follows:

$$\boldsymbol{t}_G = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{g} + \boldsymbol{\varepsilon}_G, \tag{2.3}$$

where $\boldsymbol{\varepsilon}_G = (\varepsilon_{1G}, \ldots, \varepsilon_{nG})$ is a vector of random error terms for known censoring distribution G and $E(\boldsymbol{\varepsilon}_G) = 0$ (see, for example, Aydın and Yılmaz, 2018, for a more detailed discussion). Because G is generally unknown and has to be estimated, G is replaced by its Kaplan-Meier estimation, as introduced by Koul et al. (1981), below:

$$\hat{G}(s) = 1 - \prod_{i=1}^{n}\left(\frac{n-i}{n-i+1}\right)^{I\left[t_{(i)}\leq s, \delta_{(i)}=0\right]}, \quad (s \geq 0), \tag{2.4}$$

where $\left\{\left(t_{(i)}, \delta_{(i)}\right)\right\}_{i=1}^{n}$ denotes ordered data pairs associated with $t_{(i)}$. Note that Koul et al. (1981) shows that for $n \to \infty, \hat{G}(s) \to G(s)$. Moreover, the main idea of using synthetic values is that t_{iG}'s have exactly the same expected value as the observed response variable: $E\left(t_{i\hat{G}}\right) \cong E\left(t_{iG}\right) \cong E\left(y_i\right)$, as discussed by Aydın and Yılmaz (2018).

2.2 The kNN Imputation Technique

This section has been prepared to introduce the kNN imputation technique to impute estimated scores instead of censored data points independent of distribution, which is the most important difference between the kNN imputation from synthetic data transformation. Advantages and disadvantages of this technique are given as follows:

Advantages
→ Imputed values are derived from actual values not synthetic or constructed values
→ kNN provides additional information using explanatory variables.
→ kNN imputation is a fully nonparametric method and it does not make any assumption about the relationship between x and y or x and z.
→ kNN can work for both discrete and continuous attributes. For discrete attributes,it uses the most frequently used value among k-nearest neighbors. For continuous attributes, it uses mean value of k-nearest neighbors
Disadvantages
→ It cannot be guaranteed to obtain the exact same expected values of imputed dataset (y_i^k) and completely observed response y_i
→ It is possible that the technique does not produce consistent results. Thus, estimates do not always get better when the sample size is larger.
→ Statistical properties such as bias and variance of the obtained estimator cannot be computed, because it is a fully nonparametric method.

The kNN is a similarity-based technique, that is, it depends on the distance between observations. There are various metrics to evaluate these distances. However, in this paper, the Euclidean norm, which is widely used in the literature, is preferred. Accordingly, computation of the Euclidean norm is given by

$$d(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}, \tag{2.5}$$

where $d(a,b)$ represents the value of the distance. An algorithm is developed to perform the kNN imputation for right-censored observations and is given in Algorithm 1. This algorithm is also used in Ahmed et al. (2020).

Algorithm 1 Algorithm of kNN imputation

1: **Input:** Right-censored dataset t_i
2: Censorship indicator δ_i
3: Number fo nearest neigbours k
4: Values of explanatory variable z_i
5: **Output:** Imputed dataset $\boldsymbol{y^{knn}} = \left(y_1^k, \ldots, y_n^k\right)^\tau$
6: **begin**
7: **for**($i = 1$ to n)
8: **if** ($\delta_i = 0$) **do** Step 4 (if data point is censored)
9: **for** ($j = 1$ to n)
10: Find the Euclidean distances (2.5) between z_j and z_i for each censored data point
11: **end** (If statement in Step 3)
12: sort the distances from small to large
13: **end** (for loop in Step 2)
14: **end** (for loop in Step 4)
15: **for** ($j = 1$ to k)
16: Take the first *uncensored* k values of t_i associated to sorted distances
17: Calculate the i^{th} imputed value (y_i^k) with average of nearest k records of y_i
18: Replace the imputed values (y_i^k) with censored data points ($z_i, \delta_i = 0$) in censored data set $\boldsymbol{t} = (t_1, \ldots, t_n)^T$
19: **end** (for loop in Step 10)
20: Return $\boldsymbol{y}^k = \left(y_1^k, \ldots, y_n^k\right)^T$
21: **end**.

See Ahmed et al. (2020) for a more detailed discussion on the kNN imputation

method based on right-censored data.

3 Censored Ridge-Type Kernel Estimator (CRK)

We first consider the nonparametric estimation of the unknown function $g(z)$ in (1.1) based on uncensored data. For simplicity, we assume that $\boldsymbol{\beta}$ in model (1.1) is known. Accordingly, the relationship between $y_i - \Sigma_j^p x_{ij}\beta_j$ and z_i can be defined as

$$y_i - \Sigma_j^p x_{ij}\beta_j = g(z_i) + \varepsilon_i, i = 1, 2, \ldots, n, \tag{3.1}$$

The purpose here is to approximate $g(z_i)$ closely. In this sense, an estimator is discussed by Nadaraya and Watson (1964), and it is also referred to as the kernel estimator, given by

$$g(z) = \sum_{i=1}^{n} w_{ih}(z_i)\left(y_i - \boldsymbol{x}_i\boldsymbol{\beta}\right)^2 = \boldsymbol{W}_h\left(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}\right) = \hat{\boldsymbol{g}}, \tag{3.2}$$

where h is a bandwidth (or smoothing) parameter, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, , \ldots, x_{ip})$ are known p-vectors, as defined in (1.1), and $\mathbf{W}_h$ is a kernel smoothing matrix with i^{th} entries w_{hi}, defined by

$$w_{hi} = K\left(\frac{z - z_i}{h}\right) / \sum_{i=1}^{n} K\left(\frac{z - z_i}{h}\right) = \frac{K(u)}{\sum K(u)} = \boldsymbol{W}_h, \tag{3.3}$$

where K(u) is a kernel function satisfying these properties: (i) $\int K(u)du = 1$,(ii) $K(u) = K(-u)$. The main characteristic of the kernel function is that it gives more weight to observations near z and less weight to observations far from z. Also, $K(u)$ controls the shape of the estimated nonparametric curve with bandwidth parameter h. For this study, bandwidth is determined by *GCV* criterion.

Using the equation (3.2) and the matrix-vector form of model (3.1), we can get the following partial residual:

$$\varepsilon = \boldsymbol{y} - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} - \hat{\boldsymbol{g}} = \left(\boldsymbol{I}_n - \boldsymbol{W}_h\right)\left(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}\right) = \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^T\boldsymbol{\beta}\right), \tag{3.4}$$

where $\tilde{\boldsymbol{X}} = \left(\boldsymbol{I}_n - \boldsymbol{W}_h\right)\boldsymbol{X}$ and $\tilde{\boldsymbol{y}} = \left(\boldsymbol{I}_n - \boldsymbol{W}_h\right)\boldsymbol{y}$. We thus obtain a transformed data set based on kernel residuals. Considering the partial kernel residuals given in (3.4), the estimator of the parameters vector $\boldsymbol{\beta}$ can be obtained by minimizing the weighted

residual sum of squares:

$$\begin{aligned} L_s(\boldsymbol{\beta}, h) &= \left[(\boldsymbol{I}_n - \boldsymbol{W}_h)\left(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}\right)\right]^T \left[(\boldsymbol{I}_n - \boldsymbol{W}_h)\left(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}\right)\right] \\ &= \left(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}^T\boldsymbol{\beta}\right)^T \left(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}^T\boldsymbol{\beta}\right). \end{aligned} \tag{3.5}$$

Differentiating with respect to $\boldsymbol{\beta}$ of (3.1), we obtain the normal equations by

$$\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta} = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{y}}. \tag{3.6}$$

To solve the normal equations for $\boldsymbol{\beta}$, we multiply both sides of (3.3) by $\left(\widetilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\right)^{-1}$. Thus, the weighted least square estimator of $\boldsymbol{\beta}$ is defined as

$$\widehat{\boldsymbol{\beta}} = \left(\widetilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}^T\tilde{\boldsymbol{y}}. \tag{3.7}$$

Using equation (3.2) for the unknown regression function vector $\mathbf{g}$, an updated equation can be written as follows:

$$\hat{\boldsymbol{g}} = \boldsymbol{W}_h\left(\boldsymbol{y} - \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}\right). \tag{3.8}$$

It should be emphasized that equations (3.7) and (3.8) are considered as if no perfect or exact relationship between the columns of the matrix $\mathbf{X}$ or the inverse matrix $\left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}\right)^{-1}$ exists. As noted in introduction, this paper is designed to overcome these problems. To this end, the standard remedy is to use a biased estimation method such as ridge regression, as proposed by Hoerl and Kennard (1970). For linear model $\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. For $k > 0$, the ridge estimate of the parameters vector $\boldsymbol{\beta}$ in a linear regression model is given by

$$\widehat{\boldsymbol{\beta}}_R = \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{3.9}$$

where $\mathbf{I}_p$ is $(p \times p)$ dimensional identity matrix and k is a ridge (or shrinkage) parameter to be selected using plug-in methods or common criteria such as generalized cross validation (GCV), Akaike information criterion, and so on. In this paper, it is selected by GCV criterion.

To fit model (1.1) to the data, a ridge procedure can be associated with the idea of hints due to Speckman (1988), where the ridge estimator is the minimizer of the

penalized least squares criterion

$$L_R(\boldsymbol{\beta}, \boldsymbol{g}; k) = \sum_{i=1}^{n} \left(\tilde{y}_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \boldsymbol{g}(z_i)\right)^2 + k\sum_{j=1}^{p}\beta_j^2$$
$$= \sum_{i=1}^{n} \left(\tilde{y}_i - \boldsymbol{X}_i^\tau\boldsymbol{\beta} - \boldsymbol{g}(z_i)\right)^2 + \sum_{j=1}^{p}\left(0 - k\beta_j\right)^2. \tag{3.10}$$

An equivalent way to write the criterion in (3.10) in matrix form is

$$L_R(\boldsymbol{\beta}, \boldsymbol{g}; k) = \left(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}^T\boldsymbol{\beta} - \boldsymbol{g}\right)^T \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^T\boldsymbol{\beta} - \boldsymbol{g}\right) + k\|0 - \boldsymbol{\beta}\|^2. \tag{3.11}$$

From (3.11), the ridge-type kernel estimator of $\boldsymbol{\beta}$ and $\mathbf{g}$ are defined respectively by

$$\widehat{\boldsymbol{\beta}}_R = \left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + k\boldsymbol{I}_p\right)^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{y}}, \tag{3.12}$$

and

$$\hat{\boldsymbol{g}}_R = \boldsymbol{W}_h\left(\boldsymbol{y} - \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_R\right). \tag{3.13}$$

See Aydın and Yılmaz (2018) for proofs, the statistical properties of the estimators given in (3.11), (3.12) and other details about using a ridge-type kernel estimator with a partially linear model.

In this section, a ridge-type kernel estimator is introduced to estimate right-censored responses under collinear data. As stated in Section 2, the problem of censored response observations is solved using two different techniques. We therefore have two new response variables, $t_{i\hat{G}}$ and y_i^k, obtained by synthetic data and imputation techniques, respectively. The core idea presented here is to replace the response vector $\mathbf{y}$ with the vector $t_{i\hat{G}}$ from synthetic data and the vector $\mathbf{y}^k$ obtained by kNN imputation method. Thus two different ridge-type kernel estimators will be obtained for partially linear model (1.2). With this context in mind, the following theorems establishes these procedures:

Theorem 3.1. *(CRK with synthetic data $t_{i\hat{G}}$): Let $\boldsymbol{t}_{\bar{G}} = \tilde{\boldsymbol{X}}^T\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$ where $\tilde{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{g}} + \tilde{\boldsymbol{\varepsilon}}_{\hat{G}}, \tilde{\boldsymbol{g}} = (\boldsymbol{I}_n - \boldsymbol{W}_h)\,\boldsymbol{g}$ and $\tilde{\boldsymbol{\varepsilon}}_{\boldsymbol{G}} = (\boldsymbol{I}_n - \boldsymbol{W}_h)\,\boldsymbol{\varepsilon}_{\hat{G}}$ Also, $\tilde{\boldsymbol{X}}$ is a $n \times p$ matrix and $\tilde{\boldsymbol{t}}_{\hat{G}} = (\boldsymbol{I}_n - \boldsymbol{W}_h)\,\boldsymbol{t}_{\hat{G}}$ in which $\boldsymbol{t}_{\hat{G}}$ is a $n \times 1$ synthetic data vector replaced with $\boldsymbol{y}$. If $\boldsymbol{W}_h$ is an arbitrary smoother matrix then the ridge regression estimates may be calculated by augmented data*

$$\boldsymbol{X} = \begin{pmatrix} \tilde{\boldsymbol{X}} \\ \sqrt{k}\boldsymbol{I}_p \end{pmatrix} \text{ and } \boldsymbol{t}_{\bar{G}} = \begin{pmatrix} \tilde{\boldsymbol{t}}_{\hat{G}} \\ \mathbf{0}_p \end{pmatrix}. \tag{3.14}$$

Then kernel-type ridge estimator for $\boldsymbol{\beta}$ *is denoted as* $\hat{\boldsymbol{\beta}}_{\hat{G}}$ *and given by*

$$\widehat{\boldsymbol{\beta}}_{\hat{G}} = \left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + k\boldsymbol{I}_p\right)^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{t}_{\hat{G}}. \tag{3.15}$$

Proof of the Theorem 3.1 is given in Appendix A1.

It should be noted that because of $t_{i\hat{G}} \to t_{iG}$ and $E(t_{iG}) \cong E(y_i)$ when $n \to \infty$. Thus, assumption of $E\left(\varepsilon_{i\hat{G}}\right) \cong E(\varepsilon) = 0$ will be ensured. Note also that when $k = 0$, the ridge-type estimate reduces to the estimate problem in the equation (3.7) (see Speckman, 1988). Also, it is seen that there is a formal similarity between the equation (3.11) and ridge estimator of the linear regression model. Combining equations (3.11) and (3.13) we get the estimator of the vector $\mathbf{g}$ as

$$\hat{\boldsymbol{g}}_{\hat{G}} = \boldsymbol{W}_h\left(\tilde{\boldsymbol{t}}_{\hat{G}} - \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_{\hat{G}}\right). \tag{3.16}$$

Theorem 3.2. *(CRK with imputed* $\boldsymbol{y}^k$*): Similar to Theorem 3.1, if* $\boldsymbol{y}_i$ *is replaced by imputed values* y_i^k*'s using kNN imputation, then ridge-type estimator for* $\boldsymbol{\beta}$ *and* $\boldsymbol{g}$ *in a partially linear model can be defined, respectively, as follows:*

$$\widehat{\boldsymbol{\beta}}_k = \left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + k\boldsymbol{I}_p\right)^{-1}\widetilde{\boldsymbol{X}}^T\tilde{\boldsymbol{y}}^k, \tag{3.17}$$

and

$$\hat{\boldsymbol{g}}_k = \boldsymbol{W}_h\left(\tilde{\boldsymbol{y}}^k - \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_R\right), \tag{3.18}$$

where $\tilde{\boldsymbol{X}} = (\boldsymbol{I}_n - \boldsymbol{W}_h)\boldsymbol{X}$, *and* $\tilde{\boldsymbol{y}}^k = (\boldsymbol{I}_n - \boldsymbol{W}_h)\boldsymbol{y}^k$. *Proof of equations* (3.17)-(3.18) *are given in Appendix A2.*

Theoretically, $E\left(\boldsymbol{y}^k\right) \cong E(\boldsymbol{y})$ for $n \to \infty$ cannot be proven for Theorem 3.2, because, kNN has a fully nonparametric nature. However, in practice, it can be seen that the larger the sample size, the better the results, although this improvement in the estimates cannot be guaranteed. For our results, see Section 5.

4 Performance Measures

This section prepared to introduce evaluation metrics of the proposed modified estimators. Bias and variance of the regression coefficients β_j, $j = 1, 2, \ldots, p$, variance of errors ($\hat{\sigma}_\varepsilon^2$), mean square error (MSE), and relative efficiencies (RE) are considered to determine the superiority of the estimators from the parametric and nonparametric components of the partially linear model. Note that in calculation of the metrics, for simplicity, we use only notation y_i to present response variable. One can use $t_{i\hat{G}}$ or y_i^k, instead of y_i, for synthetic data and kNN imputation methods, respectively.

4.1 Measures for Parametric Component

Equations (3.14)-(3.17) are rewritten as follows for y_{iG} equivalently:

$$\widehat{\boldsymbol{\beta}}_r = \left[\boldsymbol{I}_p + k\left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\right)^{-1}\right]^{-1}\boldsymbol{\beta}. \tag{4.1}$$

It can be seen from (4.1) that the ridge-type estimator $\hat{\boldsymbol{\beta}}_r$ is an obviously biased estimator because of $\left[\boldsymbol{I}_p + k\left(\widetilde{\boldsymbol{X}}^T\boldsymbol{X}\right)^{-1}\right]^{-1} \neq \boldsymbol{I}_p$. Note that shrinkage parameter $k \geq 0$ controls the size of estimated coefficients $\hat{\boldsymbol{\beta}}_r$ and, consequently, the bias of the estimates (See Hoerl and Kennard, 1970). From that, we can surmise that if $\boldsymbol{R}_r = \left[\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + k\boldsymbol{I}_p\right]^{-1}$, then the expected value, mean, and variance of $\hat{\boldsymbol{\beta}}_r$ can be obtained as follows:

$$E\left(\widehat{\boldsymbol{\beta}}_r\right) = \boldsymbol{R}_r\left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{g}}\right) = \boldsymbol{\beta} - k\boldsymbol{R}_r\boldsymbol{\beta} + \boldsymbol{R}_r\tilde{\boldsymbol{X}}^T, \tag{4.2}$$

$$\text{Bias}\left(\widehat{\boldsymbol{\beta}}_r\right) = \boldsymbol{B}\left(\widehat{\boldsymbol{\beta}}_r\right) = \boldsymbol{R}_r\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{g}} - k\boldsymbol{R}_r\boldsymbol{\beta}, \tag{4.3}$$

$$\text{Variance}\left(\widehat{\boldsymbol{\beta}}_r\right) = \boldsymbol{V}\left(\widehat{\boldsymbol{\beta}}_r\right) = \sigma^2\boldsymbol{R}_r\tilde{\boldsymbol{X}}^T\left(\boldsymbol{I}_n - \boldsymbol{W}_h\right)^T\left(\boldsymbol{I}_n - \boldsymbol{W}_h\right)\tilde{\boldsymbol{X}}\boldsymbol{R}_r. \tag{4.4}$$

Note that the given equations are obtained using joint notation $\hat{\boldsymbol{\beta}}_r$. To compute moments (4.2)-(4.4) for $\widehat{\boldsymbol{\beta}}_{\hat{G}}$ in (3.14) or $\widehat{\boldsymbol{\beta}}_k$ in (3.17), $\widehat{\boldsymbol{\beta}}_r$ should be replaced by those notations.

4.2 Measures for Nonparametric Component

The performance of the nonparametric component of the partially linear model is evaluated here by using mean squared error (MSE), which is a widely used metric in the nonparametric literature. MSE can be obtained for the synthetic data and the kNN imputation methods as follows:

$$MSE_{\hat{G}}\left(\hat{\boldsymbol{g}}_{\hat{G}}, \boldsymbol{g}\right) = E\left[\left\{\hat{g}_{\hat{G}}\left(z_i\right) - g\left(z_i\right)\right\}^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left[g_{\hat{G}}\left(z_i\right) - g\left(z_i\right)\right]^2, \tag{4.5}$$

and

$$MSE_k\left(\hat{\boldsymbol{g}}_k, \boldsymbol{g}\right) = E\left[\left\{\hat{g}_k\left(z_i\right) - g\left(z_i\right)\right\}^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{g}_k\left(z_i\right) - g\left(z_i\right)\right]^2, \tag{4.6}$$

where $MSE_{\hat{G}}$ and MSE_k denote the synthetic data and kNN methods, respectively.

4.3 Measures for the Model

In order to show the performance of the all estimated models using the mentioned methods, error variance of the models should be calculated. In this context, let σ_ε^2 and $\mathbf{y}$ be the joint notations. Accordingly, to estimate σ_ε^2, the residual sum of squares (RSS) are generally used. For the partially linear models, this can be obtained as follows:

$$\begin{aligned} RSS(\hat{\boldsymbol{y}}) &= (\hat{\boldsymbol{y}} - \boldsymbol{y})^T (\hat{\boldsymbol{y}} - \boldsymbol{y}) \text{ where } \left(\hat{\boldsymbol{y}} = \boldsymbol{X}^T \widehat{\boldsymbol{\beta}} + \hat{\boldsymbol{g}} = \boldsymbol{H}_h \boldsymbol{y}\right) \\ &= (\boldsymbol{y} - \boldsymbol{H}_h \boldsymbol{y})^T (\boldsymbol{y} - \boldsymbol{H}_h \boldsymbol{y}) = \boldsymbol{y}^T (\boldsymbol{I}_n - \boldsymbol{H}_h) \boldsymbol{y}, \end{aligned} \tag{4.7}$$

where $\boldsymbol{H}_h$ is known as a hat or smoother matrix, depending on bandwidth parameter $h \geq 0$, which is idempotent and used in the acquisition of fitted values for the linear models. Computation of $\boldsymbol{H}_h$ is given by

$$\boldsymbol{H}_h = \boldsymbol{W}_h + \left(\boldsymbol{I}_p - \boldsymbol{W}_h\right) \tilde{\boldsymbol{X}} \boldsymbol{R}_r \tilde{\boldsymbol{X}}^T. \tag{4.8}$$

To obtain an estimation of σ_ε^2 from that, the expected values of (4.7) are needed. This can be written as

$$E[RSS(\hat{\boldsymbol{y}})] = \sigma_\varepsilon^2 \left[n - tr\left(2\boldsymbol{H}_h - \boldsymbol{H}_h^T \boldsymbol{H}_h\right)\right] + E\left(\boldsymbol{y}^T\right) (\boldsymbol{I}_n - \boldsymbol{H}_h)^2 E(\boldsymbol{y}), \tag{4.9}$$

where $\sigma_\varepsilon^2 \left[n - \text{tr}\left(2\boldsymbol{H}_h - \boldsymbol{H}_h^T \boldsymbol{H}_h\right)\right]$ denotes the variance and $E\left(\boldsymbol{y}^T\right) (\boldsymbol{I}_n - \boldsymbol{H}_h)^2 E(\boldsymbol{y})$ represents the squared-bias of the model. In many applications, σ_ε^2 is unknown. Because its estimate $\hat{\sigma}_\varepsilon^2$ is used instead of σ_ε^2, it is calculated similar to OLS regression:

$$\hat{\sigma}_\varepsilon^2 = \frac{E(RSS)}{\text{tr}\left[(\boldsymbol{I}_n - \boldsymbol{H}_h)^2\right]} = \frac{1}{n} \frac{\left\|(\boldsymbol{I}_n - \boldsymbol{H}_h) \boldsymbol{y}\right\|^2}{n - p}, \tag{4.10}$$

where $\text{tr}\left[(\boldsymbol{I}_n - \boldsymbol{H}_h)^2\right] = (n - p)$ and p is the number of parameters. Note that $(n - p)$ is the degree of freedom and it therefore should be obvious that (4.10) fits the classical variance definition. In this context, it can be said that $\hat{\sigma}_\varepsilon^2$ is the bias estimate of σ_ε^2. This is a very common metric to evaluate the quality of estimated regression models.

As mentioned before, this paper considers two different estimators and equation (4.10) is therefore rearranged according to the kNN imputation and synthetic data methods with respect their response variables ($\boldsymbol{y}^k$ and $\boldsymbol{t}_{\hat{G}}$) and corresponding hat matrices. Also note that, in order to compare the two mentioned methods in the ridge-type estimators under censored data context, a relative measure is needed. Therefore, associated with equation (4.10), a new measurement is introduced in Definition 4.1.

Definition 4.1. The relative efficiency based on the model variances $\hat{\sigma}^2_{\varepsilon_G}$ and $\hat{\sigma}^2_{\varepsilon_k}$ is specified as follows:

$$RE\left(\boldsymbol{t}_{\hat{G}}, \hat{\boldsymbol{y}}^k\right) = \frac{\left\|\left(\boldsymbol{I}_n - \boldsymbol{H}_{h_{\hat{G}}}\right)\boldsymbol{t}_{\hat{G}}\right\|^2}{\left\|\left(\boldsymbol{I}_n - \boldsymbol{H}_{h_k}\right)\boldsymbol{y}^k\right\|^2} = \frac{\hat{\sigma}^2_{\varepsilon_G}}{\hat{\sigma}^2_{\varepsilon_k}}. \tag{4.11}$$

Note that if $RE\left(\boldsymbol{t}_{\hat{G}}, \hat{\boldsymbol{y}}^k\right) < 1$, then it can be said that fitted model based on the synthetic data technique is better than that of the kNN imputation.

5 Simulation Study

This section prepared to see behaviors of the introduced two *CRK* estimators under different conditions and scenarios. A detailed simulation experiments has been designed accordingly. Note that this paper has two main purposes: (i) to estimate right-censored data by partially linear models when multi-collinearity exists by using two different approaches, and (ii) to decide which approach is better under certain conditions and making inferences.

In order to obtain simulated data sets, a partially linear model is used and the values of the response variable are obtained using the following model:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \boldsymbol{g}(z_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{5.1}$$

where y_i represents completely observed values, $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4)$ are correlated explanatory variables, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T = (4, 2, 0.5, -1)^T$ and $\varepsilon \sim N\left(0, \sigma^2 \boldsymbol{I}\right)$. Here, the simulation experiment is repeated 1000 times for $\sigma^2 = 0.3$ and $\sigma^2 = 1$ in three sample sizes: $n = 30, 100$ and 250. In addition, to add censorship to the response variables, the censoring variable is generated randomly by using distribution of y_i's as $c_i \sim N(\mu_y, \sigma^2_y)$. $I(c_i > y_i)$ where μ_y and σ^2_y are the mean and variance of completely observed response variable y. Using equation (2.1), censored responses t_i's are obtained. Note that censoring levels are decided as CL=(5%, 35%). The nonparametric component of the model is generated as follows:

$$\boldsymbol{g}(z_i) = -0.1812 - 0.3221 z_i + 4 \sin\left(z_i^2\right) + \exp(z_i), \tag{5.2}$$

where $z_i = -2.4(i - 0.5)/n$ uniformly produced on interval $[0, 1]$. The covariates are obtained from normal distribution with two correlation levels. In this study, the level of multicollinearity (ρ) is set as $\rho = 0.85$ and $\rho = 0.99$. Note that we check existence of

the multicollinearity using a condition index (*CI*), which provides the mightiness of ρ between variables. *CI* is widely used to measure collinearity (see Belsley et al., 1980). The formulation of *CI* is given by

$$CI = \sqrt{\left[\lambda_{\max}\left(\boldsymbol{X}^T\boldsymbol{X}\right)/\lambda_{\min}\left(\boldsymbol{X}^T\boldsymbol{X}\right)\right]}. \tag{5.3}$$

In addition, to evaluate the performances of the models and their components, measurement tools given in Section 4 are used; these include bias, variances for parametric components, MSEs for nonparametric components, variance of the model, and REs for the all models. Results are given following tables and figures

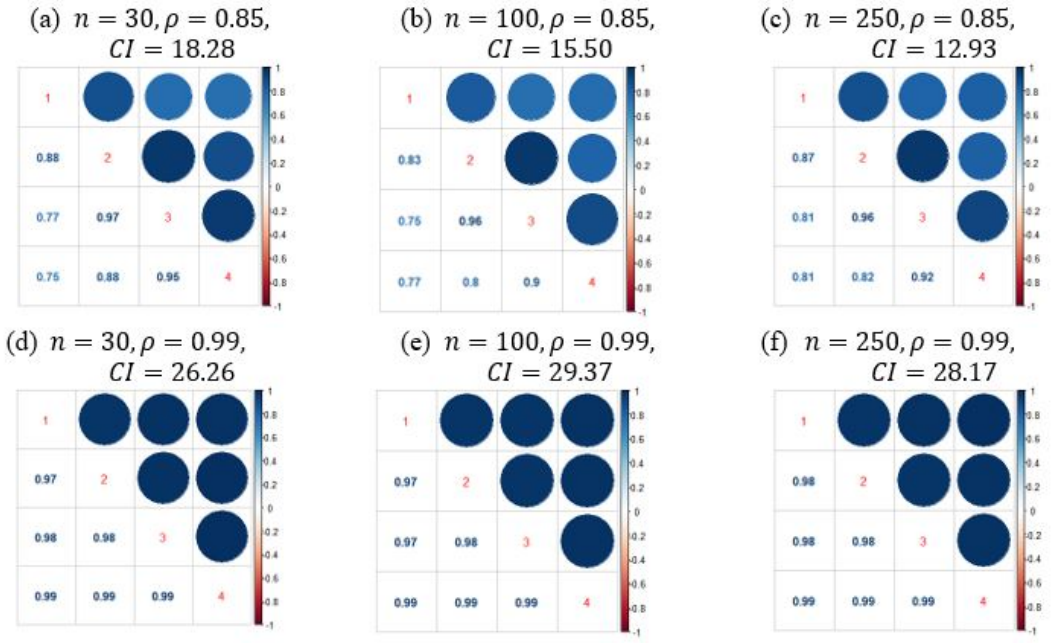

Figure 1: Correlation plots for correlation levels $\rho = 0.85$ and $\rho = 0.99$ for three sample sizes.

Figure 1 depicts correlation values between covariates that are possessed by the parametric component of the model. Also, condition index (CI) values are added to each panel of the figure. As can be seen, all CI values are larger than 10, which indicates serious multicollinearity. It is ensured that the two correlation levels ($\rho = 0.85, \rho = 0.99$) for explanatory variables are seen clearly in the sixth panel of the figure.

Table 1: Outcomes for the parametric components when $\rho = 0.85$

$\rho = 0.85$									
σ_ε		0.3				1			
Method		SDT		kNN		SDT		kNN	
CL	n	Var(β)	Bias(β)	Var(β)	Bias(β)	Var(β)	Bias(β)	Var(β)	Bias(β)
5%	30	**1.5299**	**0.2340**	2.1735	0.2346	**1.9954**	**0.2165**	2.5834	0.2281
	100	**0.8309**	0.2013	1.2315	**0.1732**	1.2610	0.2009	**0.8845**	**0.1763**
	250	0.4075	0.1555	**0.3626**	**0.1277**	0.3794	0.1657	**0.3156**	**0.1325**
35%	30	1.2576	0.2813	**1.2207**	**0.2641**	2.1904	0.2825	**1.7777**	**0.2636**
	100	**0.2500**	0.2580	0.4861	**0.2027**	**0.2765**	0.2582	0.7216	**0.2046**
	250	0.2900	0.1989	**0.2525**	**0.1445**	0.2885	0.1841	**0.2435**	**0.1421**

Performance scores of methods on estimating parametric component of the model are presented Tables 1 and 2 for $\rho = 0.85$ and $\rho = 0.99$, respectively. The best scores are indicated with bold font. It can be observed that the quality of estimations gets worse when censoring level and error variance increase, which is an expected situation. This case is also true of the correlation level. It is obvious that correlation level affects the performance of estimates, which can be monitored by Figures 3 and 4.

An interesting result from these tables is that it can be concluded that the kNN based model gives less bias than the synthetic data transformation (SDT) based model. Although both methods provide close scores in terms of $Var(\boldsymbol{\beta})$, kNN dominates the SDT method in $Bias(\boldsymbol{\beta})$, especially for high censoring levels. The reason for that can be explained by the nature of SDT, which causes spatial variation in the data. Because kNN works with data points individually, it manipulates data structure less than the SDT method.

Figures 2 and 3 are drawn to emphasize the bias of the regression coefficients. S1 and K1 in the x-axes in each panel represent SDT and kNN based biases for $n = 30$. Similarly, S2 and K2 denote biases for $n = 100$, S3 and K3 for $n = 250$. It can be seen from an examination of these figures that the boxplots are similar for both methods. In some cases, such as $\sigma = 0.3$, the SDT based estimated values of $\boldsymbol{\beta}$ are slightly better than those of kNN, but when variation in the generated data increases, creates better estimates. This can be seen in bottom panels of both figures. It can therefore be said that Figures 3 and 4 verify the results given in Tables 1 and 2.

Table 2: Outcomes for the parametric components when $\rho = 0.99$

$\rho = 0.99$									
σ_ε		0.3				1			
Method		SDT		kNN		SDT		kNN	
CL	n	Var(β)	Bias(β)	Var(β)	Bias(β)	Var(β)	Bias(β)	Var(β)	Bias(β)
5%	30	**9.0435**	**0.2978**	10.6518	0.315	**10.2537**	**0.2974**	11.1486	0.3018
	100	**4.1748**	0.2255	4.9727	**0.1889**	5.8733	0.2143	**5.0571**	**0.1799**
	250	2.1894	0.1538	**1.9546**	**0.1169**	**2.6385**	0.1524	2.7314	**0.1137**
35%	30	**15.0880**	**0.3115**	15.6767	0.3461	**11.7627**	0.3155	12.5868	**0.3144**
	100	6.8527	0.3042	**6.2235**	**0.2010**	6.7189	0.3074	**6.7090**	**0.2060**
	250	4.1108	0.2050	**4.0858**	**0.1274**	5.5155	0.2044	**3.1370**	**0.1279**

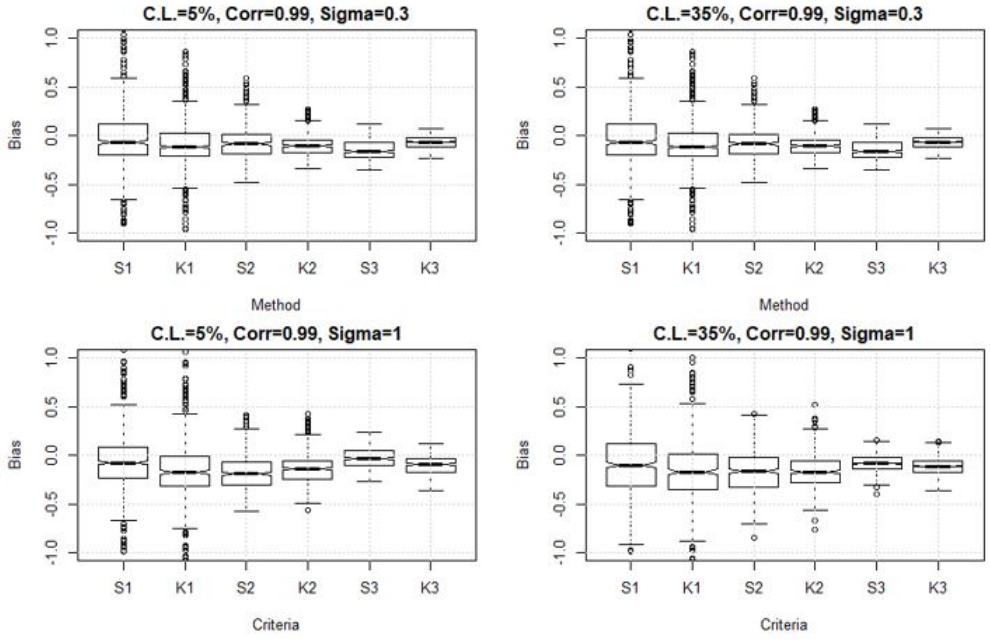

Figure 2: Boxplots for bias of regression coefficients

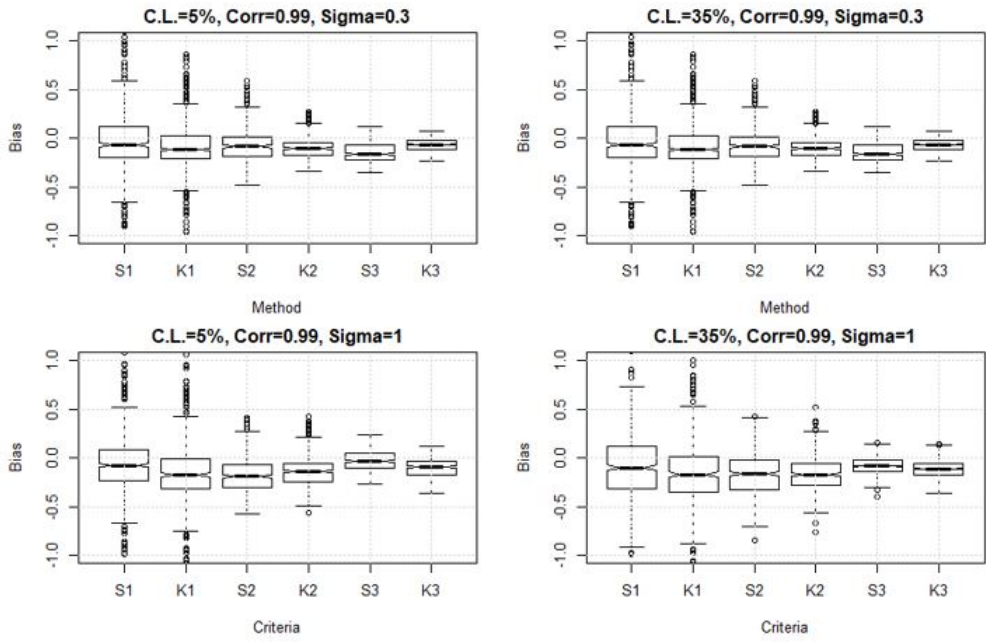

Figure 3: Boxplots for bias of regression coefficients

In Table 3, outcomes for the nonparametric component $g(z)$ are given. As noted above, MSE is commonly used as an evaluation metric for $g(z)$. From obtained MSEs in the Table 3, it can be seen that in most combinations, the scores are the same or close to each other. This can be confirmed from Figures 4-5. However, for different correlation levels and error variances, the outcomes change in favor of kNN based estimations. Differences between the methods become clear under hard conditions, such as heavy censoring or high correlation levels. From this, it can be said that although both methods provide satisfying performances on estimation of the nonparametric component, the kNN based model is more resistant to difficult conditions.

Figures 4 and 5 depict the fitted curves of SDT and kNN based estimators. Note that, because there are too many combinations to show, only some selections are presented. Figure 4 shows the estimated nonparametric functions for low correlation level $\rho = 0.85$ and Figure 5 is obtained from $\rho = 0.99$.

Table 3: Outcomes for the nonparametric components

		$\rho = 0.85$				$\rho = 0.99$			
σ		0.3		1		0.3		1	
CL	n	SDT	kNN	SDT	kNN	SDT	kNN	SDT	kNN
5%	30	0.0735	0.0735	0.0736	0.0736	0.0818	0.0818	0.0866	**0.0865**
	100	0.0718	0.0718	0.0768	0.0768	0.0853	0.0853	0.0903	0.0903
	250	0.0777	**0.0776**	**0.0771**	0.0776	0.0864	**0.0862**	0.0914	0.0914
35%	30	0.0920	**0.0915**	0.1052	**0.1046**	0.0981	**0.0973**	0.1131	**0.1123**
	100	0.0960	**0.0957**	0.1097	**0.1094**	0.1024	**0.1020**	0.1182	**0.1178**
	250	0.0971	**0.0970**	0.1110	**0.1108**	0.1036	**0.1034**	0.1195	**0.1193**

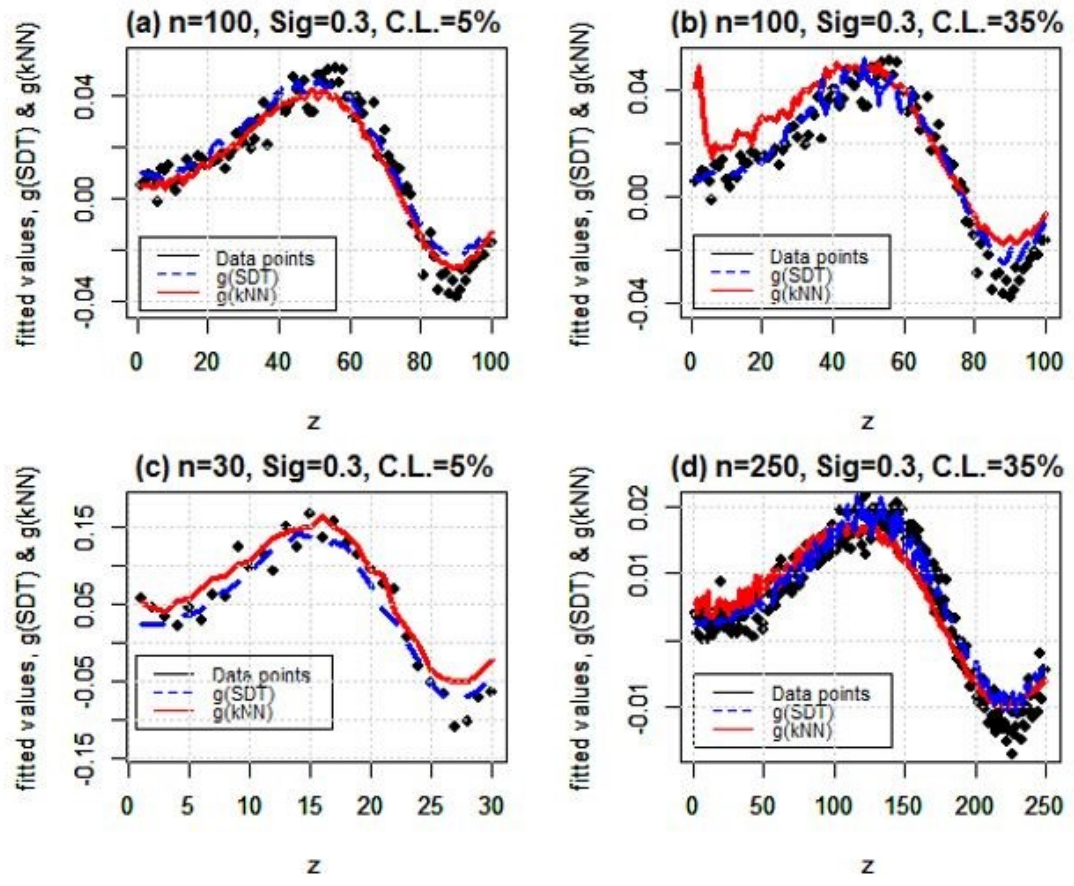

Figure 4: Fitted curves for synthetic data transformation (blue line), kNN imputation technique (red line) and real function (black line) when $\rho = 0.85$.

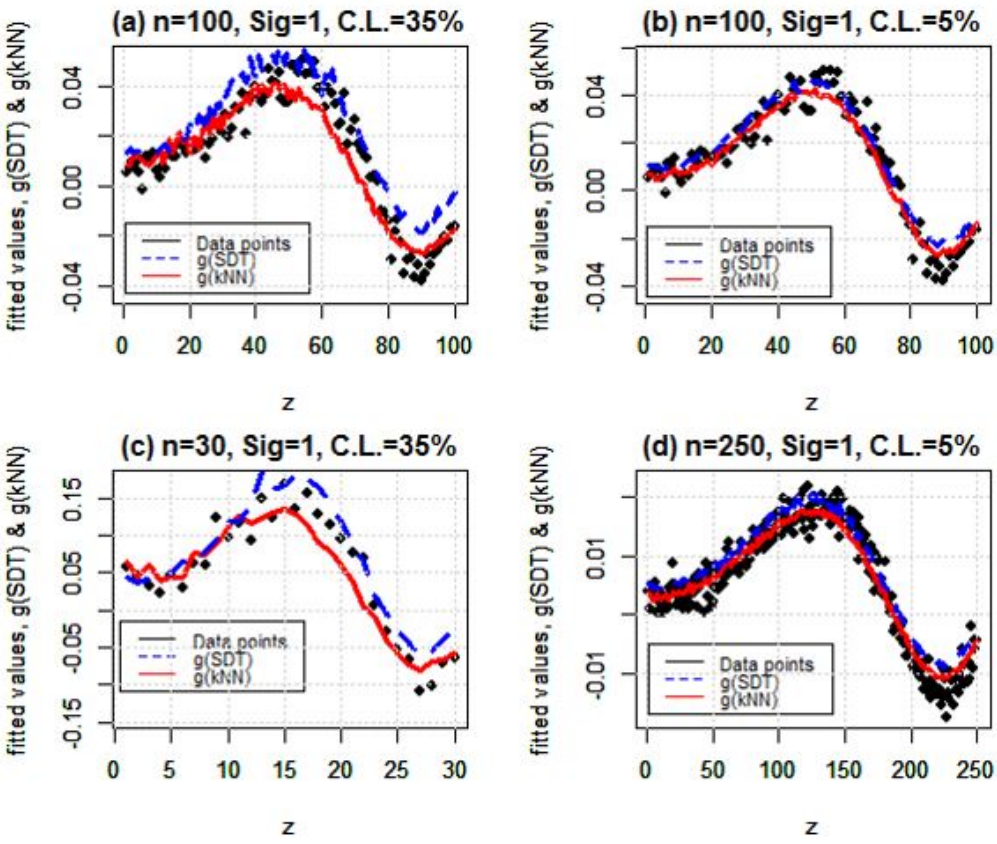

Figure 5: Fitted curves for synthetic data transformation (blue line), kNN imputation technique (red line) and real function (black line) when $\rho = 0.99$.

In the top-right panel of Figure 4, one can observe that the "weird" curve obtained by kNN seems to contradict the simulation experiments, because kNN has a better performance compared to SDT. This is because at this point, the nonparametric nature of the kNN method comes into play. Because of kNN's nature, it has the potential to give some outlier results. The plot given in top-right panel of Figure 4 can be counted as an example of such an outlier. The remaining plots in Figures 4-5 support the results in Table 3, depicting close-fitted curves and better kNN results under difficult conditions, as can be seen in the bottom-left of Figure 5.

Table 4: The REs and estimated model variances

$\rho = 0.85$									
σ		0.3				1			
Method		SDT		kNN		SDT		kNN	
CL	n	σ^2_ε	RE	σ^2_ε	RE	σ^2_ε	RE	σ^2_ε	RE
5%	30	0.9124	1.4067	**0.6486**	**0.7109**	**0.6159**	**0.9583**	0.6427	1.0435
	100	0.9020	1.4277	**0.6318**	**0.7004**	0.6420	1.0436	**0.6152**	**0.9583**
	250	0.9059	1.4282	**0.6343**	**0.7002**	0.6503	1.0448	**0.6224**	**0.9571**
35%	30	0.8906	1.4139	**0.6299**	**0.7073**	0.8840	1.4023	**0.6304**	**0.7131**
	100	0.8895	1.4347	**0.6200**	**0.6970**	0.8937	1.4391	**0.6210**	**0.6949**
	250	0.9037	1.4471	**0.6245**	**0.6910**	0.9001	1.4452	**0.6228**	**0.6919**
$\rho = 0.99$									
σ		0.3				1			
Method		SDT		kNN		SDT		kNN	
CL	n	σ^2_ε	RE	σ^2_ε	RE	σ^2_ε	RE	σ^2_ε	RE
5%	30	0.6467	1.0534	**0.6139**	**0.9493**	**0.6196**	**0.9487**	0.6531	1.0541
	100	0.6459	1.0595	**0.6096**	**0.9438**	**0.6065**	**0.9888**	0.6134	1.0114
	250	**0.5817**	**0.9632**	0.6039	1.0382	0.5743	1.0058	**0.5710**	**0.9943**
35%	30	0.8971	1.3913	**0.6448**	**0.7188**	0.9124	1.2188	**0.7486**	**0.8205**
	100	0.8763	1.3927	**0.6292**	**0.7180**	0.9020	1.2326	**0.7318**	**0.8113**
	250	0.8335	1.3763	**0.6056**	**0.7266**	0.8959	1.2369	**0.7243**	**0.8085**

Finally, Table 4 involves the model outcomes that are REs and estimated model variances. Note that because the nonparametric results are close to each other, and kNN has better scores in terms of parametric component of the model, in the model performance, the kNN method is indirectly given better model scores. Here, it can once more be seen that kNN reveals its difference for high correlation levels and high censoring levels. Otherwise, SDT has an acceptable performance for the estimation of a semiparametric model under censored multicollinear data. For this study, the kNN method demonstrates important success on modeling multicollinear censored data, which can be clearly seen in Table 4.

From the information obtained from the simulation experiments, it can be said that kernel ridge estimators based on SDT and kNN methods can handle multicollinear datasets. In addition, this study shows that kNN imputation gives more satisfying

results than SDT, which is an important conclusion. Details about the results are given in Section 6.

6 Conclusion

This paper considers the estimation of a partially linear model under multicollinear censored data. The aim of the paper is realized using a kernel-type ridge estimator based on two different censorship solution techniques: synthetic data transformation (SDT) and kNN imputation, respectively. This study therefore adds two main contributions to the literature. To show the behaviors of the introduced estimators, a detailed simulation study was designed and performed. From the results obtained from this simulation study (presented in Section 5), concluding remarks are given as follows:

- Tables 1 and 2 represent the outcomes of the parametric component of the model. It can be said that both methods, SDT and kNN, have satisfying results for estimating semiparametric models under data irregularities. On the other hand, the kNN based kernel ridge-type estimator performs better than the SDT based model, especially under heavy censoring levels and high correlation levels. This is because SDT manipulates the data set.

- Figures 2 and 3 show boxplots for the biases of the regression coefficients and confirm the results of Tables 1 and 2. The figures also demonstrate the effects of sample size, censoring level, correlation level, and error variance on the estimating of regression coefficients.

- Table 3 includes MSE values for the nonparametric component of the model for both estimators, kNN and SDT. The simulation study shows that, on the estimation of nonparametric function, the methods give highly similar results except for under high censoring and high correlation levels. The kNN method is more resistant to irregularities than SDT. Figures 4 and 5 show the fitted curves and clearly illustrate the inferences which can be drawn from Table 3.

- Finally, Table 4 shows the performance scores for model variance and relative efficiencies (RE) for the estimated models. The kNN based model has better estimates than SDT, although it should be noted that, for low censoring and correlation levels, SDT has also satisfying results. Thus, both estimators can be used under appropriate conditions.

Acknowledgement

The authors are thankful to the guest editor and anonymous reviewer for constructive comments. Prof. Dr. S. Ejaz Ahmed's research is supported by the Natural Sciences and the Engineering Research Council of Canada (NSERC).

References

Ahmed, S. E., Aydın, D., and Yılmaz, E. (2020), Nonparametric regression estimates based on imputation techniques for right-censored data. *Proceedings of the Thirteenth International Conference on Management Science and Engineering Management, ICMSEM 2019*, Advances in Intelligent Systems and Computing, Vol **1001**, Springer, Cham.

Ahmed, S. E. (2014), *Penalty, Shrinkage and Pretest Strategies: Variable Selection and Estimation*. Springer, New York, https://www.springer.com/gp/book/9783319031484.

Batista, G., and Monard, M. (2002), An analysis of four missing data treatment methods for supervised learning, *Applied Artificial Intelligence*, **17**, 519-533.

Hoerl, A. E., and Kennard, R. W. (1970), Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55-67.

Kaplan, E. L., and Meier, P. (1958), Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 457-481.

Koul, H., Susarla, V., and Van Ryzin, J. (1981), Regression analysis with randomly right censored data. *The Annals of Statistics*, **9**, 1276-88.

Liang, H. (2006), Estimation in partially linear models and numerical comparisons. *Computational Statistics and Data Analysis*, **50**(3), 675-685.

Liang, H., and Zhou, Y. (2008), Semiparametric Inference for ROC curves with censoring. *Scandinavian Journal of Statistics*, **35**(2), 212-227.

Miller, R. G. (1976), Least squares regression with censored data. *Biometrika*, **63**, 449-64.

Orbe, J., Ferreira, E., and Nunez-Anton, V. (2003), Censored Partial Regression. *Biostatistics*, **4**(1), 109-121.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), Semiparametric regression. *Cambridge University Press*, New York.

Schimek, M. G. (2000), Smoothing and Regression: Approaches, Computation and Application. *John Wiley & Sons*, Hoboken, NJ.

Shim, J. Y. (2005), Censored kernel ridge regression. *Journal of the Korean Data and Information Science Society*, **16**(4), 1045-1052.

Stute, W. (1993), Consistent Estimation Under Random Censorship When Covariables Are Present. *Journal Of Multivariate Analysis*, **4**, 89-103.

Yenduri, S., and Iyengar S. S. (2007), Performance evaluation of imputation methods for incomplete datasets. *International Journal of Software Engineering and Knowledge Engineering*, **17**, 127-152.

Yüzbaşı, B., Arashi, M., and Ahmed, S. E. (2017), Shrinkage estimation strategies in generalized ridge regression models under low/high-dimension regime. *International Statistical Review*, **88**(1), 229-251.

Appendix

A1. Proof of Theorem 3.1

As mentioned before, $\mathbf{W}_h$ is the kernel smoother matrix and $\tilde{\boldsymbol{X}}$ in (3.14) are the variables obtained by using $\mathbf{W}_h$, as mentioned in Speckman (1988). From that, let us consider the weighted minimization problem to obtain (3.14)-(3.16) as follows:

$$\begin{aligned} L_{\hat{G}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\varepsilon_{i\hat{G}}\right)^2 &= \sum_{i}^{n} \left(\tilde{t}_{i\hat{G}} - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}\right)^2 + \sum_{j=1}^{n} \left(0 - k\beta_j\right)^2 \\ &= \left(\tilde{\boldsymbol{t}}_{\hat{G}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}\right)^T \left(\tilde{\boldsymbol{t}}_{\hat{G}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\right) + k\|\mathbf{0} - \boldsymbol{\beta}\|^2. \end{aligned} \tag{A1.1}$$

In order to minimize (A1.1), augmented data sets are used. This can be shown as

$$\boldsymbol{X}_A = \begin{pmatrix} \widetilde{\boldsymbol{X}}_{n\times p} \\ \left(\sqrt{k}\boldsymbol{I}_p\right) \end{pmatrix} \text{ and } \boldsymbol{t}_A = \begin{pmatrix} \tilde{\boldsymbol{t}}_{\hat{G}_{n\times 1}} \\ \mathbf{0}_P \end{pmatrix}, \tag{A1.2}$$

where $\mathbf{I}_p$ is the $(p \times p)$ dimensional identity matrix and $\mathbf{0}_P$ is the $(n \times 1)$ zero matrix. For details of (A1.2), see Aydın et al. (2018). It should be differentiated with respect to $\boldsymbol{\beta}$ and set the equation to zero, which is given by

$$\begin{aligned} \frac{\partial}{\partial\beta} L_{\hat{G}}(\boldsymbol{\beta}) &= -2\left(\widetilde{\boldsymbol{X}}\tilde{\boldsymbol{t}}_{\hat{G}}\right) + 2\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}\boldsymbol{\beta} = \mathbf{0}, \\ \widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}\boldsymbol{\beta} &= \widetilde{\boldsymbol{X}}_{\hat{G}}, \\ \widehat{\boldsymbol{\beta}}_{\hat{G}} &= \left(\widetilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + k\boldsymbol{I}_p\right)^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{t}_{\hat{G}}. \end{aligned} \tag{A1.3}$$

From (A1.3), $\hat{\boldsymbol{g}}_{\hat{G}}$ is obtained as follows easily as in equation (3.16):

$$\hat{\boldsymbol{g}}_{\hat{G}} = \boldsymbol{W}_h\left(\tilde{\boldsymbol{t}}_{\hat{G}} - \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_{\hat{G}}\right). \tag{A1.4}$$

Proof of Theorem 3.2

Similar to Appendix A1, replacing synthetic response variable $\tilde{\boldsymbol{t}}_{\hat{G}}$ by the kNN based response variable $\mathbf{y}^k$ minimization criterion is written as follows:

$$\begin{aligned} L_{kNN}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (\varepsilon_i)^2 &= \sum_{i}^{n} \left(\tilde{y}_i^k - \tilde{X}_i\boldsymbol{\beta}\right)^2 + \sum_{j=1}^{n} \left(0 - k\beta_j\right)^2 \\ &= \left(\tilde{\boldsymbol{y}}^k - \widetilde{\boldsymbol{x}}\boldsymbol{\beta}\right)^T \left(\tilde{\boldsymbol{y}}^k - \widetilde{\boldsymbol{x}}\boldsymbol{\beta}\right) + k\|\mathbf{0} - \boldsymbol{\beta}\|^2. \end{aligned} \tag{A2.1}$$

Then, as in A1, augmented data sets are given by

$$\boldsymbol{X}_A = \begin{pmatrix} \tilde{\boldsymbol{X}}_{n\times p} \\ \left(\sqrt{k}\boldsymbol{I}_p\right) \end{pmatrix} \text{ and } \boldsymbol{t}_A = \begin{pmatrix} \tilde{\boldsymbol{y}}^k_{n\times 1} \\ \boldsymbol{0}_P \end{pmatrix}, \tag{A2.2}$$

where $\mathbf{I}_p$ is the $(p \times p)$ dimensional identity matrix and $\mathbf{0}_P$ is the $(n \times 1)$ zero matrix. According to (A2.2), derivation of $\boldsymbol{\beta}$ is given as

$$\begin{aligned} \frac{\partial}{\partial\beta} L_{kNN}(\boldsymbol{\beta}) &= -2\left(\widetilde{\boldsymbol{X}}\tilde{\boldsymbol{y}}^k\right) + 2\widetilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta} = \boldsymbol{0}, \\ \widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}\boldsymbol{\beta} &= \widetilde{\boldsymbol{X}}\tilde{\boldsymbol{y}}^k, \\ \widehat{\boldsymbol{\beta}}_k &= \left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + k\boldsymbol{I}_p\right)^{-1}\widetilde{\boldsymbol{X}}^T\tilde{\boldsymbol{y}}^k. \end{aligned} \tag{A2.3}$$

Thus equation (3.17) is derived. From (A2.3), $\hat{\boldsymbol{g}}_{\hat{G}}$ is obtained as follows easily as in equation (3.18):

$$\hat{\boldsymbol{g}}_{\hat{k}} = \boldsymbol{W}_h\left(\tilde{\boldsymbol{y}}^k - \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_k\right). \tag{A2.4}$$