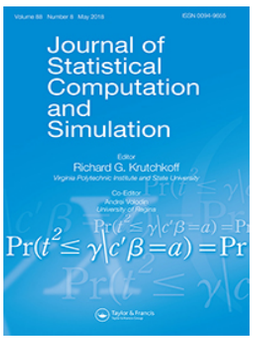

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Modified estimators in semiparametric regression models with right-censored data

Dursun Aydin & Ersin Yilmaz

To cite this article: Dursun Aydin & Ersin Yilmaz (2018) Modified estimators in semiparametric regression models with right-censored data, Journal of Statistical Computation and Simulation, 88:8, 1470-1498, DOI: 10.1080/00949655.2018.1439032

To link to this article: https://doi.org/10.1080/00949655.2018.1439032

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2018
VOL. 88, NO. 8, 1470–1498
https://doi.org/10.1080/00949655.2018.1439032

Taylor & Francis
Taylor & Francis Group

Modified estimators in semiparametric regression models with right-censored data

Dursun Aydin and Ersin Yilmaz

Faculty of Science, Department of Statistics, Mugla Sitki Kocman University, Muğla, Turkey

ABSTRACT

In this work we introduce different modified estimators for the vector parameter $\boldsymbol{\beta}$ and an unknown regression function g in semiparametric regression models when censored response observations are replaced with synthetic data points. The main idea is to study the effects of several covariates on a response variable censored on the right by a random censoring variable with an unknown probability distribution. To provide the estimation procedure for the estimators, we extend the conventional methodology to censored semiparametric regression using different smoothing methods such as smoothing spline (SS), kernel smoothing (KS), and regression spline (RS). In addition to estimating the parameters of the semiparametric model, we also provide a bootstrap technique to make inference on the parameters. A simulation study is carried out to show the performance and efficiency properties of the estimators and analyse the effects of the different censoring levels. Finally, the performance of the estimators is evaluated by a real right-censored data set.

ARTICLE HISTORY
Received 26 September 2017
Accepted 6 February 2018

KEYWORDS
Modified estimators; smoothing spline; kernel smoothing; regression spline; right-censored data

1. Introduction

In statistics literature, a problem commonly faced by statisticians is the analysis of censored survival data. Examples of this data arise in different applied fields such as medicine, biology, public health, epidemiology, engineering, economics, and demography. Observations in these fields are usually incomplete, especially in medical studies. For instance, some patients may still be alive, disease-free or die at the termination of a medical study. There are two main traditional statistical methods, parametric and nonparametric, used in analysis of the relationship between covariates and the censored response variable, known as lifetime. Instead, we focus on semiparametric estimation methods which do not require knowledge of the underlying distribution of the response variable.

Let $\{(y_i, x_i, t_i), \quad i = 1, 2, \ldots, n\}$ be independent and identically distributed (i.i.d) random variables satisfying the semiparametric regression model

$$y_i = x_i\beta + g(t_i) + \varepsilon_i, \tag{1}$$

CONTACT Ersin Yilmaz yilmazersin13@hotmail.com Faculty of Science, Department of Statistics, Mugla Sitki Kocman University, Muğla 48000, Turkey

where the y_i's are values of the response variable, $x_i = (x_{i1}, \ldots, x_{ip})$ are known p-vectors of explanatory variables with $p \leq n$, the t_i's are values of an extra univariate explanatory variable, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is an unknown vector of regression parameters, g is an unknown smooth regression function in $\mathbb{R}$, and ε_i's are random error terms with $E[\varepsilon_i | x_i, t_i] = 0$ and variance σ^2 and are independent of the data (x_i, t_i). To be specific, this model is also called a partially linear model due to the connection with the classical linear model.

In vector and matrix form, model (1) can be rewritten as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)'$, $\mathbf{X}' = [x_1, \ldots, x_n]$, $\mathbf{g} = (g(t_1), \ldots, g(t_n))'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. As in most literature, it is supposed that $\mathbf{X}$ has full column rank. Note also that the function g symbolizes the smooth part of the model and assume that it shows the unparameterized functional relationship. The model (1) was discussed by Engle et al. [1] based on the assumption of the completely observed y_i. For more details on the model (1), see Wahba [2,3], Speckman [4], Green and Silverman [5] among others.

In our study, we are interested in estimating the vector parameter $\boldsymbol{\beta}$ and an unknown function $g(.)$ in semiparametric model (1) when they_i's are observed incompletely and right censored by a random censoring variable $c_i, i = 1, 2, \ldots, n$, but x_i and t_i are observed completely. Therefore, instead of observing y_i, we now observe the pair of values (z_i, δ_i), $i = 1, \ldots, n$ where

$$z_i = \min(y_i, c_i) \quad \text{and} \quad \delta_i = \begin{cases} 1; & \text{if } y_i \leq c_i (i\text{th response is observed}), \\ 0; & \text{if } y_i > c_i (i\text{th response is censored}). \end{cases} \tag{3}$$

Here, we assume that c_i's are i.i.d random variables with a common distribution G (i.e. the distribution of the censoring observations c_i). Notice also that z_i and c_i are referred to as the lifetimes and the censoring time, respectively, for the ith survival subject. z_i's are the observed lifetimes, while δ_i stores the information on whether an observation is censored or uncensored. If an observation is not censored, we choose $z_i = y_i$ and $\delta_i = 1$; otherwise, we take $z_i = c_i$ and $\delta_i = 0$.

The model (1) is also said to be a right-censored semiparametric regression model when the response variable is observed incompletely and right censored by a random variable. The useful special cases of the censored partial linear model can be obtained using different methods and conditions. For example, if $g(t) = 0$ in Equation (1), the censored partially linear model reduces to the linear regression model with censored data. A number of authors have studied the case of the linear regression model with censored data. Examples of such studies include Miller [6], Buckley and James [7], Koul et al. [8], Zheng [9], Leurgans [10] and Lai et al. [11]. On the other hand, if $\boldsymbol{\beta} = 0$ in model (1), the mentioned model reduces to the nonparametric regression model with censored data. See studies by Dabrowska [12], Zheng [13], Fan and Gijbels [14], Guessoum and Ould-Saïd [15] for examples.

In this paper, the key idea is to estimate the vector parameter $\boldsymbol{\beta}$, an unknown function $g(.)$, and the mean vector $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g}$. Because the values of $\mathbf{Y}$ are the censored observations, the censoring distribution G is usually unknown. For this reason, traditional methods for estimating $\boldsymbol{\beta}$ and $g(.)$ cannot be applied directly here. To overcome this problem, Koul et al. [8] suggested replacing incomplete observations with synthetic data. In our

context, we introduce differently modified estimators for the components of the semiparametric model with right-censored data, especially when the censored response variable is replaced by synthetic data. The modified estimators are based on a generalization of the ordinary SS, KS, and RS methods in the case of unknown censoring distribution G. We also provide the bootstrapped confidence intervals for the estimators. It is worth noting that some authors have studied semiparametric regression with censored data. For example, the censored partial linear model in which the censoring distribution is supposed to be known is considered by Qin and Lawless [16] and Qin and Cai [17]. Asymptotic properties for the right-censored semiparametric regression are discussed by Wang and Zheng [18]. In the last decade, Qin and Jing [19] discussed the asymptotic properties for estimation of partial linear models with censored data. Orbe et al. [20] examined censored partial regression and proposed an estimation procedure based on penalized weighted least squares, using Kaplan–Meier weights.

The rest of the paper is outlined as follows. In Section 2, we present the censored partial regression model and provide the details for the estimation procedure. Section 4 describes a method that uses bootstrap resample techniques to make an inference. Section 6 provides some simulation results that support the adequacy of the methodology in different situations. Section 5 presents the results of the application of real data and, finally, Section 7 presents the conclusions.

2. Preliminaries and computation of estimators

Let F and G be the probability distribution functions of y_i and c_i, respectively. That is, the unknown distribution function of the response is $F(t) = P(y_i \leq k)$ and the censoring times are $G(t) = P(c_i \leq k)$. In order to ensure that the model is identifiable, one needs to make some specification assumptions on the response, censoring and explanatory variables and their dependence relationships. For this purpose, we have followed the identifiability assumptions of Stute [21,22]:

Assumption A: y_i and c_i are i.i.d, conditional on (x_i, t_i)

Assumption B: $P(y_i \leq c_i | y_i, x_i, t_i) = P(y_i \leq c_i | y_i)$

These assumptions are used in survival analysis applications. If we use the Kaplan–Meier estimator, assumption A is a standard independence condition to ensure identifiability of the model with censored data. If assumption A is violated, then we need more information about the censoring structure to construct a proper model. Assumption B will be required to allow for a dependency between $(x_i,\ t_i)$ and c_i. In other words, assumption B says that given time of death, covariates do not ensure any further information as to whether the observation is censored or not. See Stute [22,23,24], Tsiatis [25], Wei et. al. [26] and Zhou [27] for additional details on assumptions of lifetime data analysis.

As expressed in the previous section, a formal connection between a linear and partially linear model can be constructed through a right-censored response variable y. If $g(t) = 0$ in model (1), this model transforms into the linear model

$$y_i = x_i\boldsymbol{\beta} + \varepsilon_i,\ \ i = 1, \ldots, n. \tag{4}$$

Note that we assume only the response variable is censored. Since the y_i's are observed incompletely, standard methods which require knowledge of the completely observed y_i's cannot be used for the censored data. Under censorship, rather than a random variable

y_i, we observe $\{(z_i, \delta_i),\ i = 1, 2, \ldots, n\}$, as defined in (3). Under these conditions, Koul et al. [8] discussed that if G is continuous and known, it is possible to adjust the lifetime observations z_i to yield an unbiased modification

$$y_{iG} = \frac{\delta_i z_i}{1 - G(z_i)}, \quad i = 1, 2, \ldots, n. \tag{5}$$

The above assumptions A and B are also used to ensure that $E[y_{iG}|x_i] = E[y_i|x_i] = x_i\boldsymbol{\beta}$. Hence, the ordinary least squares (*OLS*) estimator of $\boldsymbol{\beta}$ in model (4) is defined by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_G, \tag{6}$$

where $\mathbf{Y}_G = (y_{1G}, \ldots, y_{nG})'$ is the n x 1 vector of adjusted responses. However, in most applications, G is usually unknown. To overcome this problem, Koul et al. [8] proposed replacing G in Equation (6) by its Kaplan–Meier [28] estimator $\hat{G}$, given by

$$1 - \hat{G}(k) = \prod_{i=1}^{n} \left(\frac{n-i}{n-i+1} \right)^{\mathrm{I}[z_{(i)} \le k, \delta_{(i)} = 0]}, \quad (k \ge 0), \tag{7}$$

where $z_{(1)} \le z_{(2)} \le \cdots \le z_{(n)}$ are the ordered observations of z and $\delta_{(i)}$ is the corresponding censoring indicator associated with $z_{(i)}$.

Hence, a feasible estimator of $\boldsymbol{\beta}$ can be obtained as

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_{\hat{G}}, \tag{8}$$

where

$$\mathbf{Y}_{\hat{G}} = (y_{1\hat{G}}, \ldots, y_{n\hat{G}})' = \frac{\delta_i z_i}{1 - \hat{G}(z_i)} = y_{i\hat{G}}, \quad i = 1, \ldots, n. \tag{9}$$

Here, it should be noted that $y_{i\hat{G}}$'s are also called synthetic observations since these values are synthesized from the data (z_i, δ_i) to fit the semiparametric model $E[y_{iG}|x_i, t_i] = x_i\boldsymbol{\beta} + g(t_i)$. In this case, in a similar fashion to the linear model based on synthetic data, the assumptions A and B provide that $E[y_{iG}|x_i, t_i] = E[y_i|x_i, t_i] = x_i\boldsymbol{\beta} + g(t_i)$.

A number of authors have studied synthetic data in dealing with the censored data problem (see, for example, Zheng [29], Leurgans [10], Qin and Jing [19], Delecroix et al. [30], and Guessoum and Ould-Saïd [15], Lemdani and Ould-Saïd [31]. The generalization of the well-known censored linear model to the partially linear model censored from the right will now be stated and discussed. As can be seen from (1), the partially linear models combine both parametric and nonparametric components; so they are much more flexible than the ordinary linear models. In this study, we adapted three different methods to fit (1) when the response variable is replaced by synthetic data. The first is smoothing spline method, which is suggested to estimate the vector of parameters $\boldsymbol{\beta}$ from partial residuals and unknown g function. The second method is kernel smoothing based on the Nadaraya [32] and Watson [33] kernel weighted average with kernel function. Finally, the third method is regression spline method which is based on penalized spline functions.

2.1. Smoothing spline

We first introduce the penalized least square estimates for $\boldsymbol{\beta}$ and $\mathbf{g}$ in the model (2) with right-censored data. Let the ordered distinct values among $t_1, t_2, \ldots, t_n$ be indicated by $r_1 < r_2 < \cdots < r_q$. The connection between $t_1, t_2, \ldots, t_n$ and $r_1, r_2, \ldots, r_q$ is provided via the $n \times q$ incidence matrix $\mathbf{N}$, with elements $N_{ij} = 1 \quad \text{if } t_i = r_j$ and $N_{ij} = 0 \quad \text{if } t_i \neq r_j$. It can be seen that $q \geq 2$ from the assumption that the t_i's are not all identical.

Let $\mathbf{g} = g(r_j) = (a_1, a_2, \ldots, a_q)'$ be a vector. Then, the estimates of the $\boldsymbol{\beta}$ and $\mathbf{g}$, based on the synthetic observations (9), are obtained by minimizing the penalized residuals sum of squares criterion

$$f_1(\boldsymbol{\beta}; \mathbf{g}) = (\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{g})'(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{g}) + \lambda \int_a^b g''(t)^2 \mathrm{d}t. \tag{10}$$

The first term on the right-hand side measures the goodness of fit to data, while the second term penalizes curvature in the function g. Note that the curvature (or smoothness of the function) is controlled by smoothing parameter $\lambda > 0$.

The idea behind penalized least square is to provide a minimum through the combination of the goodness of fit and the penalty terms. Using the properties of cubic splines, the resulting value of penalty is $\mathbf{g}'\mathbf{K}\,\mathbf{g} = \int g''(t)^2 \mathrm{d}t$ (see [5]). Thus, the criterion (10) can be rewritten as

$$f_2(\boldsymbol{\beta}; \mathbf{g}) = (\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{g})'(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{g}) + \lambda\, \mathbf{g}'\mathbf{K}\,\mathbf{g}, \tag{11}$$

where $\mathbf{K}$ is a symmetric $q \times q$ positive definite penalty matrix with a solution $\lambda\,\mathbf{K} = \mathbf{S}_\lambda^{-1} - \mathbf{I}$, and $\mathbf{S}_\lambda$ is a well-known positive-definite linear smoother matrix which depends on λ, as defined in Equation (16). Also, the elements of the matrix $\mathbf{K}$ are obtained by means of the knot points $r_1, \ldots, r_q$, and defined by

$$\mathbf{K} = \mathbf{Q}'\mathbf{R}^{-1}\mathbf{Q},$$

where $h_j = r_{j+1} - r_j, \quad j = 1, 2, \ldots, q-1$, $\mathbf{Q}$ is a tri-diagonal $(q-2) \times q$ matrix with entries $Q_{jj} = 1/h_j$, $Q_{j,j+1} = -(1/h_j + 1/h_{j+1})$, $Q_{j,j+2} = 1/h_{j+1}$, and $\mathbf{R}$ is a symmetric tri-diagonal $(q-2) \times (q-2)$ matrix with $R_{j-1,j} = R_{j,j-1} = h_j/6$, $R_{jj} = (h_j + h_{j+1})/3$.

By taking simple algebraic operations it is seen that the solution to the minimization problem $f_2(\boldsymbol{\beta}; \mathbf{g})$ in Equation (11) satisfies the block matrix equation:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{N} \\ \mathbf{N}'\mathbf{X} & (\mathbf{N}'\mathbf{N} + \lambda\mathbf{K}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{g} \end{pmatrix} = \begin{pmatrix} \mathbf{X}' \\ \mathbf{N}' \end{pmatrix} \mathbf{Y}_{\hat{G}}. \tag{12}$$

For some constant $\lambda > 0$, the corresponding estimators for $\boldsymbol{\beta}$ and $\mathbf{g}$, based on model (2) with censored data, can be easily obtained by (subscripts SS denotes the smoothing spline)

$$\hat{\boldsymbol{\beta}}_{SS} = [\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Y}_{\hat{G}} \tag{13}$$

and

$$\hat{\mathbf{g}}_{SS} = (\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'(\mathbf{Y}_{\hat{G}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{SS}). \tag{14}$$

Using Equations (13)–(14), the vector of fitted values is

$$\boldsymbol{\mu}_{SS} = (\mathbf{X}\hat{\boldsymbol{\beta}}_{SS} + \hat{\mathbf{g}}_{SS}) = (\mathbf{H}_{\lambda}^{SS}\mathbf{Y}_{\hat{G}}) = (\hat{\mathbf{Y}}_{\hat{G}} = E[Y|X,t]) \tag{15}$$

for

$$\mathbf{H}_{\lambda}^{SS} = \mathbf{S}_{\lambda} + (\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{X}[\mathbf{X}'(\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_{\lambda}), \tag{16}$$

where $\mathbf{S}_{\lambda} = \mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'$. Specifically, if t_i's are distinct and ordered, then the incidence matrix $\mathbf{N} = \mathbf{I}$. In this case, the matrix $\mathbf{S}_{\lambda}$ reduces to $\mathbf{S}_{\lambda} = (\mathbf{I} + \lambda\mathbf{K})^{-1}$. Note also that $\hat{\boldsymbol{\beta}}_{SS}$ and are called as modified smoothing spline regression estimators of the vectors $\boldsymbol{\beta}$ and $\mathbf{g}$, respectively.

Details on the derivation of the Equations (13–16) can be found in the Appendix A1.

2.2. Kernel smoothing

As shown in the previous sections, when the response variable is censored by a random variable, the model (1) reduces to the censored model. For simplicity, we shall use the $y_{i\hat{G}}$ defined in Section 2. Before starting, let us define $\varepsilon_{i\hat{G}} = y_{i\hat{G}} - (x_i\boldsymbol{\beta} + g(t_i))$, $i = 1, \ldots, n$. From this, we have

$$y_{i\hat{G}} = x_i\boldsymbol{\beta} + g(t_i) + \varepsilon_{i\hat{G}},\ i = 1, \ldots, n, \tag{17}$$

where ε_{iG}'s are identical but not independent random error observations with unknown constant variance. Conceptually, as $n \to \infty$, $\mathrm{E}(\varepsilon_{\hat{G}}) \cong 0$. This information will help us to define estimates for $\boldsymbol{\beta}$ and $\mathbf{g}$. For convenience, we assume that $\boldsymbol{\beta}$ is known. In this case, the relationship between $(y_{i\hat{G}} - x_i\boldsymbol{\beta})$ and t_i can be denoted by

$$(y_{i\hat{G}} - x_i\boldsymbol{\beta}) = g(t_i) + \varepsilon_{i\hat{G}}, \quad i = 1, \ldots, n.$$

This can be considered as equivalent to the first part of (17). Then, kernel smoothing can be used as an alternative nonparametric approach to spline to get a reasonable estimate of the function $g(.)$. In analogy with (14), this leads to the Nadarya–Watson estimator proposed by Nadaraya [32] and Watson [33] (subscripts KS denotes the smoothing spline)

$$\hat{\mathbf{g}}_{KS} = \sum_{j=1}^{n} w_{j\lambda}(t_j)((y_{j\hat{G}} - x_j\boldsymbol{\beta})) = \mathbf{W}_{\lambda}(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}), \tag{18}$$

where $\mathbf{W}_{\lambda}$ is a kernel smoother matrix with jth entries $w_{j\lambda}$, given by

$$w_{j\lambda}(t_j) = K\left(\frac{t - t_j}{\lambda}\right) \Bigg/ \sum_{j=1}^{n} K\left(\frac{t - t_j}{\lambda}\right) = K(u) / \sum K(u), \tag{19}$$

where λ is a smoothing (bandwidth) parameter as noted earlier and $K(u)$ is a kernel or weight function such that $\int K(u)\mathrm{d}u = 1$, and $K(u) = K(-u)$. The kernel function is selected to give the most weight to observations close to t and least weight to observations far from t. While the kernel $K(u)$ determines the shape of the regression curves, the smoothing parameter λ determines their width. In this study, the selection of the λ

is obtained by minimizing the generalized cross validation (GCV) criterion (see Craven and Wahba [34]).

Using the matrix and vector form of Equation (17), we can obtain the following partial residuals in matrix form,

$$\boldsymbol{\varepsilon}_{\hat{G}} = \mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta} - \hat{\mathbf{g}}_{KS} = (\mathbf{I} - \mathbf{W}_\lambda)(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}) = \tilde{\mathbf{Y}}_{\hat{G}} - \tilde{\mathbf{X}}\boldsymbol{\beta}, \tag{20}$$

where

$$\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X} = x_i - \sum_{j=1}^{n} w_{j\lambda}(t_j)x_j = \tilde{x}_i,$$

$$\tilde{\mathbf{Y}}_G = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{Y}_{\hat{G}} = y_{i\hat{G}} - \sum_{j=1}^{n} w_{j\lambda}(t_j)y_{j\hat{G}} = \tilde{y}_{i\hat{G}}.$$

Thus, we obtain a transformed set of data based on kernel residuals. Considering these partial residuals for the vector $\boldsymbol{\beta}$ yields the following weighted least squares criterion:

$$f_3(\boldsymbol{\beta}) = ||(\mathbf{I} - \mathbf{W}_\lambda)(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta})||^2. \tag{21}$$

The solution to the criterion $f_3(\boldsymbol{\beta})$ in Equation (21) is observed as

$$\hat{\boldsymbol{\beta}}_{KS} = \sum_{i=1}^{n} \tilde{x}_i\tilde{y}_{i\hat{G}} \Big/ \sum_{i=1}^{n} \tilde{x}_i^2 = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}_{\hat{G}}. \tag{22}$$

Replacing $\boldsymbol{\beta}$ in Equation (18) with $\hat{\boldsymbol{\beta}}_{KS}$ in Equation (22), we obtain an estimator of $\mathbf{g}$

$$\hat{\mathbf{g}}_{KS} = \sum_{j=1}^{n} w_{j\lambda}(t_j)((y_{j\hat{G}} - x_j\hat{\boldsymbol{\beta}}_{KS})) = \mathbf{W}_\lambda(\mathbf{Y}_{\hat{G}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{KS}) \tag{23}$$

and hence, from Equation (22) and (23) the vector of fitted values is

$$\boldsymbol{\mu}_{KS} = (\mathbf{X}\hat{\boldsymbol{\beta}}_{KS} + \hat{\mathbf{g}}_{KS}) = (\mathbf{H}_\lambda^{KS}\mathbf{Y}_{\hat{G}}) = (\hat{\mathbf{Y}}_{\hat{G}} = E[Y|X,t]), \tag{24}$$

where

$$\mathbf{H}_\lambda^{KS} = \mathbf{W}_\lambda + (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X}(\mathbf{X}'(\mathbf{I} - \mathbf{W}_\lambda)'(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{W}_\lambda)^2. \tag{25}$$

The vectors $\hat{\boldsymbol{\beta}}_{KS}$ and $\hat{\mathbf{g}}_{KS}$ are known as modified kernel regression estimators of the vectors $\boldsymbol{\beta}$ and $\mathbf{g}$, respectively.

The implementation details of Equations (22)–(25) are given in Appendix A2.

2.3. Regression spline

Smoothing spline becomes less practical when the sample size n is large because it uses n knot points. A regression spline is a piecewise polynomial function whose highest order non-zero derivative takes jumps at fixed 'knots'. Usually, regression splines are smoothed by deleting non-essential knots. When the knots are selected, regression spline can be fitted

by ordinary least squares. For further discussion on the selection of knots, see the study of Ruppert et al. [35].

As noted in previous sections, we fit partially linear model (1) with randomly right-censored data. For this purpose, regression spline can be used as an alternative approach to the others described above. By using the synthetic data in Equation (9) we will estimate components of the model (1) so that sum of squares of the differences between the censored response observations $y_{i\hat{G}}$ and $(x_i\boldsymbol{\beta} + g(t_i))$ is a minimum. Here, the unknown regression function $g(t_i)$ is approximated by a qth degree regression spline with a truncated power basis

$$g(t_i) = b_0 + b_1 t_{i1} + \ldots + b_q {t_{i1}}^q + \sum_{k=1}^{K} b_{q+k}(t_i - \kappa_k)_+^q + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{26}$$

where $\mathbf{b} = (b_0, b_1, \ldots, b_q, b_{q+1}, \ldots, b_{q+K})'$ is a vector of unknown coefficients to be estimated, $q \geq 1$ is an integer that indicates the degree of regression spline and $(t - \kappa_k)_+ = t_i$ when $(t - \kappa_k) > 0$ and $(t - \kappa_k)_+ = 0$ otherwise. Also, $\kappa_1 < \cdots < \kappa_K$ are the specifically selected knots $\{\min(t_i) \leq \kappa_1 <, \ldots, < \kappa_K \leq \max(t_i)\}$.

Substituting Equation (26) into Equation (17) we get an expression of the form

$$y_{i\hat{G}} = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + b_0 + b_1 t_{i1} + \cdots + b_q {t_{i1}}^q + \sum_{k=1}^{K} b_{q+k}(t_i - \kappa_k)_+^p + \varepsilon_{i\hat{G}}. \tag{27}$$

The equality (27) is a censored semiparametric model due to comprise of the parametric linear component and a nonparametric component. Using matrix and vector notation, Equation (27) can be rewritten as

$$\mathbf{Y}_{\hat{G}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{b} + \boldsymbol{\varepsilon}_{\hat{G}}, \tag{28}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p, b_0, b_1, \ldots, b_q)'$ represents the coefficients of the parametric component, while $\mathbf{b} = (b_{q+1}, \ldots, b_{q+K})$ denotes the coefficients of the nonparametric component of the model, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is a vector of random error, $\mathbf{X}$ and $\mathbf{U}$ are design matrices such that ith rows of them are defined as:

$$\mathbf{X}_i = \begin{bmatrix} 1 & x_{i1} & \ldots & x_{ip} & t_i & \ldots & t_i^q \end{bmatrix} \text{ and}$$
$$\mathbf{U}_i = [(t_i - \kappa_1)_+^p \quad \ldots \quad (t_i - \kappa_K)_+^p], \; 1 \leq i \leq n.$$

The fitted censored partially linear model is then

$$\hat{y}_{i\hat{G}} = x_{i1}\hat{\beta}_1 + \cdots + x_{ip}\hat{\beta}_p + \hat{b}_0 + \hat{b}_1 t_{i1} + \cdots + \hat{b}_q {t_{i1}}^q + (t_1, \ldots, t_K)(b_{q+1}, \ldots, b_{q+K})'. \tag{29}$$

Equation (29) gives a point estimate of the mean of $\mathbf{Y}_{\hat{G}}$ for a particular $\mathbf{X}$ and $\mathbf{U}$.

In the regression spline context, a penalty approach similar to smoothing spline is used to estimate the vector of $\boldsymbol{\beta}$ and an unknown regression function vector $\mathbf{g}$; but this approach has fewer knots and it applies a somewhat more general penalty. Regression spline estimators $(\hat{\boldsymbol{\beta}}_{RS} = (\hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{b}_0, \hat{b}_1, \ldots, \hat{b}_q)', \; \hat{\mathbf{g}}_{RS} = (\hat{b}_{q+1}, \ldots, \hat{b}_{q+K}))$ of $(\boldsymbol{\beta}, \mathbf{b})$ are obtained

by minimizing the penalized objective function

$$f_4(\boldsymbol{\beta};\lambda) = \sum_{i=1}^{n} (y_{i\hat{G}} - x_i\boldsymbol{\beta} - g(t_i))^2 + \lambda \sum_{k=1}^{K} \mathbf{b}_{p+k}^2 = ||(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\mathbf{b})||^2 + \lambda\, \mathbf{b}'\mathbf{D}\,\mathbf{b}, \tag{30}$$

where $\mathbf{g}$ is an unspecified regression function, $\lambda \sum_{j=1}^{K} \mathbf{b}_{p+j}^2$ is penalty term corresponding to the sum of squared coefficients of the truncated powers, $\mathbf{D} = \text{diag}(\mathbf{0}_{p+1},\ \mathbf{1}_K)$ – that is, $\mathbf{D}$ is a diagonal penalty matrix whose first $(p\ +\ 1)$ elements are 0, and the remaining elements are 1, and λ is a positive smoothing parameter that controls the influence of the penalty.

Minimization of the objective function $f_4(\boldsymbol{\beta};\lambda)$ in Equation (30) leads to the system of equations

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{U} \\ \mathbf{U}'\mathbf{X} & (\mathbf{U}'\mathbf{U} + \lambda\mathbf{D}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}' \\ \mathbf{U}' \end{pmatrix} \mathbf{Y}_{\hat{G}}. \tag{31}$$

From Equation (31) we can easily obtain (subscripts *RS* denotes the regression spline)

$$\hat{\boldsymbol{\beta}}_{RS} = (\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{Y}_{\hat{G}}, \tag{32}$$

where $\mathbf{A}^{-1} = \mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}$ and

$$\hat{\mathbf{g}}_{RS} = (\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'(\mathbf{Y}_{\hat{G}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{RS}). \tag{33}$$

Thus, the vector of fitted values is given by

$$\boldsymbol{\mu}_{RS} = (\mathbf{X}\hat{\boldsymbol{\beta}}_{RS} + \mathbf{U}\hat{\mathbf{g}}_{RS}) = (\mathbf{H}_{\lambda}^{RS}\mathbf{Y}_{\hat{G}}) = (\hat{\mathbf{Y}}_{\hat{G}} = E[Y|X,t]), \tag{34}$$

where

$$\mathbf{H}_{\lambda}^{RS} = \mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}' + (\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}')\mathbf{X}(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}. \tag{35}$$

Details on the derivation of Equations (32)–(35) can be found in Appendix A3.

The smoothing parameter (penalty parameter λ) and the number of knots $\{\kappa_i,\ i = 1,\ldots,K\}$ must be selected in implementing the regression spline. However, λ plays an essential role. (See Ruppert et al. [35] for a detailed discussion of the knot selection). A recent study was conducted by Aydın and Yılmaz [36] on the choice of optimum knots for regression spline under censored data.

3. Statistical properties of the estimator

In this section, we study the statistical characteristics of the estimators expressed in the previous sections. To see the computations of each estimator, we first expand $\hat{\boldsymbol{\beta}}_{SS}$ in Equation (13) by the matrix and vector form of (17) to find

$$\hat{\boldsymbol{\beta}}_{SS} = (\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}'\tilde{\mathbf{Y}}_{\hat{G}} = \boldsymbol{\beta} + (\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}'\tilde{\mathbf{g}} + (\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}'\,(\mathbf{I} - \mathbf{S}_{\lambda})\boldsymbol{\varepsilon}_{\hat{G}},$$

where $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{X}$, $\tilde{\mathbf{Y}}_{\hat{G}} = (\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{Y}_{\hat{G}}$ and $\tilde{\mathbf{g}} = (\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{g}$.

Thus, the bias and variance–covariance matrix of the $\boldsymbol{\beta}_{SS}$ can be expressed respectively as,

$$B(\hat{\boldsymbol{\beta}}_{SS}) = E(\hat{\boldsymbol{\beta}}_{SS}) - \boldsymbol{\beta} = (\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}'\tilde{\mathbf{g}}, \tag{36}$$

$$Var(\hat{\boldsymbol{\beta}}_{SS}) = \sigma^2(\mathbf{X}'\tilde{\mathbf{X}})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{X}(\mathbf{X}'\tilde{\mathbf{X}})^{-1}. \tag{37}$$

Similarly, expanded form of the $\hat{\boldsymbol{\beta}}_{KS}$ in Equation (22) is

$$\hat{\boldsymbol{\beta}}_{KS} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}_{\hat{G}} = \boldsymbol{\beta} + (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{g}} + (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{S}_\lambda)\boldsymbol{\varepsilon}_{\hat{G}}$$

with corresponding equations for bias and variance–covariance matrix:

$$B(\hat{\boldsymbol{\beta}}_{KS}) = E(\hat{\boldsymbol{\beta}}_{KS}) - \boldsymbol{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{g}}, \tag{38}$$

$$Var(\hat{\boldsymbol{\beta}}_{KS}) = \sigma^2(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)^2\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}. \tag{39}$$

Finally, as in the above expressions, the $\hat{\boldsymbol{\beta}}_{RS}$ in Equation (32) can be expanded as

$$\hat{\boldsymbol{\beta}}_{RS} = (\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{Y}_{\hat{G}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{g} + (\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}\boldsymbol{\varepsilon}_{\hat{G}}.$$

Hence, the bias and variance–covariance matrix of this estimator are,

$$B(\hat{\boldsymbol{\beta}}_{RS}) = E(\hat{\boldsymbol{\beta}}_{RS}) - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{g}, \tag{40}$$

$$Var(\hat{\boldsymbol{\beta}}_{RS}) = \sigma^2(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{X}\mathbf{A}^{-1}(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}. \tag{41}$$

As seen from Equations (37) and (39), the variance matrices are not practical because they depend on the unknown σ^2. It is seen that the estimate of σ^2 is needed to construct the mentioned variance–covariance matrices.

3.1. Estimating the variance of the error terms

As shown in the previous sections, although the smoothing methods provide estimates of the regression coefficients, they do not directly provide an estimate of the variance of the error terms (i.e. σ^2). As in linear regression, an estimate of σ^2 can be formed by residual sum of squares (*RSS*)

$$RSS = \sum_{i=1}^{n} \varepsilon_{i\hat{G}}^2 = \sum_{i=1}^{n} (y_{i\hat{G}} - \hat{y}_{i\hat{G}})^2 = (\mathbf{Y}_{\hat{G}} - \hat{\mathbf{Y}}_{\hat{G}})'(\mathbf{Y}_{\hat{G}} - \hat{\mathbf{Y}}_{\hat{G}}). \tag{42}$$

Substituting $\hat{\mathbf{Y}}_{\hat{G}} = \mathbf{H}_\lambda\mathbf{Y}_{\hat{G}}$, we have

$$RSS = (\mathbf{Y}_{\hat{G}} - \mathbf{H}_\lambda\mathbf{Y}_{\hat{G}})'(\mathbf{Y}_{\hat{G}} - \mathbf{H}_\lambda\mathbf{Y}_{\hat{G}}) = ||(\mathbf{I} - \mathbf{H}_\lambda)\mathbf{Y}_{\hat{G}}||^2,$$

where $\mathbf{H}_\lambda$ is a hat matrix for the semiparametric model (2) with censored data. Hence, as in linear regression, an estimate of σ^2 for smoothing spline method, as

$$\hat{\sigma}^2 = RSS \big/ tr(\mathbf{I} - \mathbf{H}_\lambda)^2 = ||(\mathbf{I} - \mathbf{H}_\lambda)\mathbf{Y}_{\hat{G}}||^2 \big/ tr((\mathbf{I} - \mathbf{H}_\lambda)'(\mathbf{I} - \mathbf{H}_\lambda)), \tag{43}$$

where $tr(\mathbf{I} - \mathbf{H}_\lambda)^2 = n - 2tr(\mathbf{H}_\lambda) + tr(\mathbf{H}_\lambda{}'\mathbf{H}_\lambda)$ is called the degrees of freedom (*DF*) for a λ pre-chosen with any smoothing parameter selection criteria. Note also that trace of a

square matrix $\mathbf{A}$, denoted by $tr(\mathbf{A})$, is the sum of the diagonal elements of $\mathbf{A}$. When we adopt the smoothing spline method, the computation of $\mathbf{H}_\lambda^{SS}$ in Equation (16) instead of $\mathbf{H}_\lambda$ as stated in Equations (42) and (43) is needed. In a similar fashion, for kernel smoothing and regression spline methods, we have to calculate the $\mathbf{H}_\lambda^{KS}$ in Equation (25) and $\mathbf{H}_\lambda^{RS}$ in Equation (35) matrices, respectively. The traces of the matrices, $tr(\mathbf{H}_\lambda^{SS})$, $tr(\mathbf{H}_\lambda^{RS})$, and $tr(\mathbf{H}_\lambda^{KS})$ can be found in O(n) algebraic operations, and hence, these matrices can be calculated in only a linear time.

The estimator of σ^2 in Equations (42) and (43) has a positive bias. However, also note that the Equation (42, 43) yields asymptotically negligible bias. Considering this point of view, it is noteworthy that $\hat{\sigma}^2$ is equivalent to mean square error (*MSE*), which is a widely used criterion for measuring the quality of estimation (see [4]).

3.2. Measuring the risk and efficiency

This section investigates the superiority of an estimator $\hat{\boldsymbol{\beta}}$ with respect to another estimator $\hat{\boldsymbol{\beta}}$. As indicated in the previous section, the estimators have bias and there is a need to measure the squared error loss of estimators. Generally, the expected loss of an estimator vector $\hat{\boldsymbol{\beta}}$ is measured by quadratic risk function. Our task now is to approximate the risk in the models with censored data. Such approximations have the advantage of being simpler in optimizing the practical selection of smoothing parameters. For convenience, we will work with the scalar valued version (*SMDE*) of mean distribution error.

In general, a comparison of estimators can be made via a mean distribution error (MDE) matrix. Let $\hat{\boldsymbol{\beta}}$ be an estimator of a p-dimensional parameters vector $\boldsymbol{\beta}$. With respect to a squared error loss, the MDE is defined as a matrix that consists of the sum of the variance–covariance matrix and the squared bias:

$$MDE(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = Var(\hat{\boldsymbol{\beta}}) + [E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}]^2. \tag{44}$$

Equation (44) gives detailed information about the quality of an estimator. In addition to the *MDE* matrix, the average or expected loss, which is referred to as the scalar valued version, can also be used for comparing the different estimators.

Definition 3.1: The quadratic risk of an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is defined as the mean distribution error matrix (SMDE), and given by

$$R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}, \mathbf{V}) = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{V}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = tr(\mathbf{V}(MDE(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}))) = SMDE,$$

where $\mathbf{V}$ is a $p \times p$ symmetric and non-negative definite matrix. Based on the risk above, it can be defined using the following criterion to compare estimators.

Definition 3.2: Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be two competing estimators of $\boldsymbol{\beta}$. It can be said that $\hat{\boldsymbol{\beta}}_2$ is superior to $\hat{\boldsymbol{\beta}}_1$ if the difference of their MDE matrices is non-negative definite, given by

$$\Delta(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) = MDE(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}) - MDE(\hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta}) \geq 0.$$

An important connection between the above definitions (36) and (37) is given by Theobald [37]. According to this, we have the following result.

Theorem 3.1: *Theobald* [37] *let* $\hat{\boldsymbol{\beta}}_1$ *and* $\hat{\boldsymbol{\beta}}_2$ *be estimator vectors of a parameter vector* $\boldsymbol{\beta}$. *As such, the following two statements are equivalent*

(a) $R(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}, \boldsymbol{V}) - R(\hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta}, \boldsymbol{V}) \geq 0$ *for all non-negative definite matrices* $\boldsymbol{V}$
(b) $MDE(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}) - MDE(\hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta})$ *is a non-negative definite matrix.*

An important consequence of the above theorem denotes that

$$E(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}) \leq E(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})$$

if and only if for non-negative definite matrices $\mathbf{V}$,

$$E(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta})'\mathbf{V}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}) \leq E(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}))'\mathbf{V}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}).$$

The results of Theorem 3.2 reveal that estimator $\hat{\boldsymbol{\beta}}_2$ has a smaller $MDE(\hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta})$. than $\hat{\boldsymbol{\beta}}_1$ if and only if the $R(\hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta}, \mathbf{V})$ of $\hat{\boldsymbol{\beta}}_2$ averaging over every quadratic risk is less than that of $\hat{\boldsymbol{\beta}}_1$. Thus, the superiority of $\hat{\boldsymbol{\beta}}_2$ over $\hat{\boldsymbol{\beta}}_1$ can be observed by comparing the MDE matrices.

Note also that Arnold and Katti [38] prove that $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta},\mathbf{V}) - MDE(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ is non-negative definite. According to the ideas stated above, if Theorem 3.1 and Definition 3.1 are considered together, we may write the expression $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta},\mathbf{V}) \geq MDE(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$. Non-negative definite also implies that the scalar valued version of the MDE matrix can be used for comparisons between different estimators.

Applying Equations (36) and (37), as described in Equation (44), the *MDE* matrix of the estimator $\hat{\boldsymbol{\beta}}_{SS}$ is obtained as

$$MDE(\hat{\boldsymbol{\beta}}_{SS},\boldsymbol{\beta}) = (\mathbf{X}'\tilde{\mathbf{X}})^{-1}(\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{X}(\sigma^2 + \mathbf{g}'\mathbf{g}))(\mathbf{X}'\tilde{\mathbf{X}})^{-1}. \tag{45}$$

Also, according to Definition 3.1, the quadratic risk (or *SMDE*) for $\hat{\boldsymbol{\beta}}_{SS}$ can be defined by

$$R(\hat{\boldsymbol{\beta}}_{SS},\boldsymbol{\beta}, \mathbf{V}) = tr(\mathbf{V}(MDE(\hat{\boldsymbol{\beta}}_{SS},\boldsymbol{\beta}))) = tr(MDE(\hat{\boldsymbol{\beta}}_{SS},\boldsymbol{\beta})) = SMDE. \tag{46}$$

Similarly, the *MDE* matrices of the estimators $\hat{\boldsymbol{\beta}}_{KS}$, and $\hat{\boldsymbol{\beta}}_{RS}$, are given by:

$$MDE(\hat{\boldsymbol{\beta}}_{KS},\boldsymbol{\beta}) = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{S}_\lambda)^2\tilde{\mathbf{X}}(\sigma^2 + \mathbf{g}'\mathbf{g}))(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} \tag{47}$$

and

$$MDE(\hat{\boldsymbol{\beta}}_{RS},\boldsymbol{\beta}) = (\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X}\mathbf{A}^{-1}(\sigma^2 + \mathbf{g}'\mathbf{g}))(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1} \tag{48}$$

with corresponding quadratic risks, similar to Equation (46).

Hence, using Theorem 3.1 and Definitions 3.2, for example, the superiority of the estimator $\hat{\boldsymbol{\beta}}_{KS}$ over $\hat{\boldsymbol{\beta}}_{SS}$ with respect to the difference in their MDE matrices can be observed as follows:

$$\Delta(\hat{\boldsymbol{\beta}}_{SS}, \hat{\boldsymbol{\beta}}_{KS}) = MDE(\hat{\boldsymbol{\beta}}_{SS}, \boldsymbol{\beta}) - MDE(\hat{\boldsymbol{\beta}}_{KS}, \boldsymbol{\beta}) \geq 0. \tag{49}$$

The remaining comparison scenarios can be made similar to Equation (49). As a result, it is possible to check whether a non-negative definite matrix condition is satisfied.

Furthermore, we can compare the quality of two estimators by looking at the ratio of their quadratic risks. This ratio provides the following definition concerning the asymptotic relative efficiencies of any two estimators.

Definition 3.3: The asymptotic relative efficiency (RE) of an estimator $\hat{\boldsymbol{\beta}}_1$ compared to another estimator $\hat{\boldsymbol{\beta}}_2$ is defined by the ratio,

$$RE(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \mathbf{V}) = R(\hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta}, \mathbf{V}) \Big/ R(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}, \mathbf{V}) = tr(MDE(\hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta})) \Big/ tr(MDE(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta})) \quad (50)$$

if $RE(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \mathbf{V}) < 1$, then $\hat{\boldsymbol{\beta}}_2$ is said to be more efficient than $\hat{\boldsymbol{\beta}}_1$.

4. Inference using bootstrap technique

Bootstrap is a computer-intensive technique that allows us to assess the statistical accuracy of the standard errors of regression estimates. The bootstrap technique was originally proposed by Efron [39]. It can also be used for creating nonparametric confidence intervals. In this paper, we focus on the semiparametric regression model with randomly right-censored data. A similar study was carried out by Orbe et. al. [20] where the authors used the Kaplan–Meier weights for estimating the semiparametric censored model based on bootstrap method. The difference is that we use synthetic data to generate the bootstrap samples with a replacement for the case of random censorship and the model. For each estimation method, the steps required to construct the bootstrap samples can be expressed as follows:

Step 1. Fit the semiparametric regression model (1) as described in Section 2 and compute its residuals.

Step 2. Select a random sample of n observations with replacement from the centred residuals and keep them in bootstrap residuals $\varepsilon_i^* = [\varepsilon_1^*, \ldots, \varepsilon_n^*]'$.

Step 3. Compute the bootstrap responses $y_i^* = x_i\hat{\boldsymbol{\beta}} + \hat{g}(t_i) + \varepsilon_i^*,\ i = 1, \ldots, n$. In other words, the bootstrap response observations y_i^* are calculated by adding the bootstrapped residuals ε_i^* to the predicted values $x_i\hat{\boldsymbol{\beta}} + \hat{g}(t_i)$.

Step 4. Generate a vector of Bernoulli variables $(\delta^*)_{i=1}^n$ where $P(\delta_i^* = 1 | y_i^*, x_i^*, t_i^*) = 1 - \hat{G}(y_i^*)$ and construct the bootstrapped censoring indicator. Here, $\hat{G}(y_i^*)$ is the Kaplan–Meier estimator of the probability distribution function of the censoring variable, as described in Equation (7).

Step 5. Generate the censoring variable c. If $z_i^* = y_i^*$ and $\delta_i^* = 1$, bootstrapped censoring variable c^* is taken from $\hat{G}$, which is restricted by the interval $[y_i^*, \infty)$, while if $z_i^* = c_i^*$ and $\delta_i^* = 0$, the c^* is drawn from $\hat{G}$, which is bounded by the interval $[0, z_i^*)$.

Step 6. Estimate the model (1) according to four estimation methods, as described in Section 2, and go back to step 2 and repeat the bootstrap procedure B times.

Note that the basic idea in step 2 is to randomly draw the bootstrap samples with replacement from the residuals. The bootstrap procedure is made B times, producing B bootstrap samples. We then refit the censored model to each of the bootstrap samples and examine the behaviour of the model fits over the $B = 1000$ replications. For more details about bootstrap procedure, see Efron [40], Akritas [41], Efron and Tibshirani [42].

5. Real data example

To illustrate the findings of our methods on real data, we consider the kidney data from a study by McGilchrist and Aisbett [43]. To estimate the recurrence times of infection for

kidney patients, they used a Cox regression model with the inclusion of a frailty term additive of five explanatory variables. The mentioned data consists of 76 kidney patients and four explanatory variables. The response variable is referred to as recurrence times of infection (*retime*) and explanatory variables are *age*, *sex* (1 = male, 2 = female), frailty (*frail*), and disease type (*distype*) coded as 0 = GN, 1 = AN, 2 = PKD, 3 = other. Definitions of these variables are presented in the study of McGilchrist and Aisbett [43]. Note that there is a censoring indicator (1 = infection occurs; 0 = censored) associated with the response variable. According to the censoring indicator variable, there are 58 censored recurrence time values. For that reason, it is clearly observed that the censoring percentage is about 76% and that the kidney data has been heavily censored.

To detect the nonparametric part of the semiparametric model, we used approximate *F*-test statistics proposed by Hastie and Tibshirani [44]. This *F*-test can also be extended to the semiparametric setting for the linear fit versus nonparametric fit. Suppose that we want to test the hypothesis $H_0 : E(y_i) = \mu$ (linear function) versus the alternative $H_1 : E(y_i) = g(t_i)$ (smooth function), when one uses this *F*-test statistics formula

$$F_{\mathrm{d}f_1-\mathrm{d}f_0,n-\mathrm{d}f_1} = \frac{\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2 - \sum_{i=1}^{n}\hat{v}_i^2\right)\Big/\left(\mathrm{d}f_1 - \mathrm{d}f_0\right)}{\sum_{i=1}^{n}\hat{v}_i^2\Big/\left(n - \mathrm{d}f_1\right)}, \tag{51}$$

where $\hat{\varepsilon}_i = (y_i - x_i'\hat{\boldsymbol{\beta}}_{OLS})$ and $\hat{v}_i = x_i'\hat{\boldsymbol{\beta}}_{SM} + \hat{g}(t_i) - x_i{}'\hat{\boldsymbol{\beta}}_{OLS}$. $\hat{\boldsymbol{\beta}}_{OLS}$ is the estimates of parameters from ordinary least squares (OLS) method, $\hat{\boldsymbol{\beta}}_{SM}$ is the estimates obtained by any smoothing method (*SM*) discussed in Section 2, $\mathrm{d}f_0$ is the number of the parameters in the OLS model and $\mathrm{d}f_1 = tr(2\mathbf{H}_\lambda - \mathbf{H}_\lambda\mathbf{H}_\lambda{}')$ where $\mathbf{H}_\lambda$is as described in Equations (42) and (43). To obtain the estimates in the *F*-test statistics (51), we considered only *SS* method here but, as we said before, any smoothing method can be used for this purpose. For this paper, the purpose of *F*-test statistics is to provide the right decision about the shape of the nonparametric component *g* in the semiparametric model (1) with censored data.

We use *frail* as a variable of the nonparametric component because it has the largest value of *F*-test statistics among all the other explanatory variables. The statistics of the above test for all explanatory variables can be found in Table 1. To display function *g* graphically, we follow the suggestions in the study of Ruppert et al. [35] by plotting the fitted model versus *frail* with *age*, *sex* and *distype* fixed at its average value. In this context, the plot of $\overline{age}\,\hat{\beta}_1 + \overline{sex}\,\hat{\beta}_2 + \overline{distype}\hat{\beta}_3 + \hat{g}_{SS}(frail)$ using the smoothing spline for modelling function *g* with 95% confidence intervals, together with partial residuals, is illustrated in Figure 1. Also, partial residuals in this figure are defined by

$$retime_i - [\hat{\beta}_1(age_i - \overline{age}) + \hat{\beta}_2(sex_i - \overline{sex}) + \hat{\beta}_3(distype_i - \overline{disytpe})].$$

Inspection of this figure indicates that the relationship between *retime* and *frail* may be nonlinear, especially when a smooth curve with 95% confidence intervals is added. Based

Table 1. The outcomes from the *F*-test.

Variable	$\mathrm{d}f_0$	$\mathrm{d}f_1$	*F*-Statistic	*p*-Value
age	5	32.999	1.516	0.213
distype	5	7.000	0.328	0.881
frail	5	16.425	16.129	0.000

Table 2. The outcomes from the model (52) by four different methods.

Methods	Variables	Est. Coeff ($\hat{\boldsymbol{\beta}}$)	95% C.I	*B*	*Var*	*SMDE*
KS	*age*	−0.833	[−0.851, −0.392]	3.945	21.452	37.015
	sex	102.857	[76.553, 105.382]			
	distype	−1.074	[−2.090, 1.227]			
SS	*age*	−0.731	[−0.826, −0.235]	3.955	25.446	41.088
	sex	139.174	[105.648, 141.521]			
	distype	−10.300	[−16.986, −8.356]			
RS	*age*	−1.094	[−1.120, −0.540]	2.249	18.383	23.441
	sex	70.138	[55.811, 75.124]			
	distype	−13.720	[−15.707, −6.015]			
OLS	*age*	−0.706	[−2.108, 1.374]	3.831	676.942	691.619
	sex	144.101	[55.830, 155.813]			
	distype	−14.896	[−30.261, 49.090]			

on these ideas, we consider the *frail* as a variable modelled nonparametrically. Hence, the nonparametric part of the semiparametric model is composed of a univariate variable *frail*, while the parametric part is constructed using three explanatory variables, *age*, *sex*, and *distype*, to estimate the *retime* of kidney patients. For this example, a semiparametric model with censored data is specified by

$$retime_i = \beta_{1i} age_i + \beta_{2i} sex + \beta_{3i} distype + g(frail_i) + \varepsilon_i, \quad i = 1, \ldots, 76. \tag{52}$$

In order to estimate the model (52), we first transformed the response variable $retime_i$ into synthetic variable $retime_{i\hat{G}}$, as described in Section 2. The estimate for the semiparametric regression model with censored data can then be obtained. In this context, the comparative outcomes come from parametric components of the model (52) by each of the smoothing methods summarized in Table 2. It should be emphasized that from left to right this table shows the method names, variable names, the estimated regression coefficients and 95% confidence intervals for the bootstrapped estimates of $\boldsymbol{\beta}$ (see Section 4). Moreover, bias, variance and quadratic risk values (see Section 3.2) for each estimator of $\boldsymbol{\beta}$ are expressed in Section 2. These outcomes show that the *RS* method performs better in the sense of having a smaller bias, variance and *SMDE* values. Continuing, using Equation (49) we obtained the difference between the $\hat{\boldsymbol{\beta}}_{RS}$ and other estimators:

$$\Delta(\hat{\boldsymbol{\beta}}_{KS}, \hat{\boldsymbol{\beta}}_{RS}) = 8.458, \ \Delta(\hat{\boldsymbol{\beta}}_{SS}, \hat{\boldsymbol{\beta}}_{RS}) = 8.528, \text{ and } \Delta(\hat{\boldsymbol{\beta}}_{OLS}, \hat{\boldsymbol{\beta}}_{RS}) = 648.506.$$

The above numerical results denote that the theoretical results and non-negative definite condition are satisfied.

After the parametric coefficients are estimated, the nonparametric component of the model (52) is calculated with the help of these coefficients. Since there is no possibility to express them parametrically, they are only shown graphically in Figure 2. Notice that Figure 2 compares nonlinear effects of the *frail* variable on *retime* for three smoothing methods. In addition to plotting the fitted curves, 95% bootstrap confidence intervals of the estimated curves and their relative bootstrapped errors are computed and illustrated in Figure 3.

The fitted lines in Figure 3 show the nonparametric component fits with the shaded regions' 95% bootstrapped confidence intervals by smoothing methods on the kidney data. For three smoothing methods, it is proposed to estimate the parameter λ by using the GCV. The value of λ is approximately found to be 0.500, 0.950 and 1.019 for the *KS*, *SS*

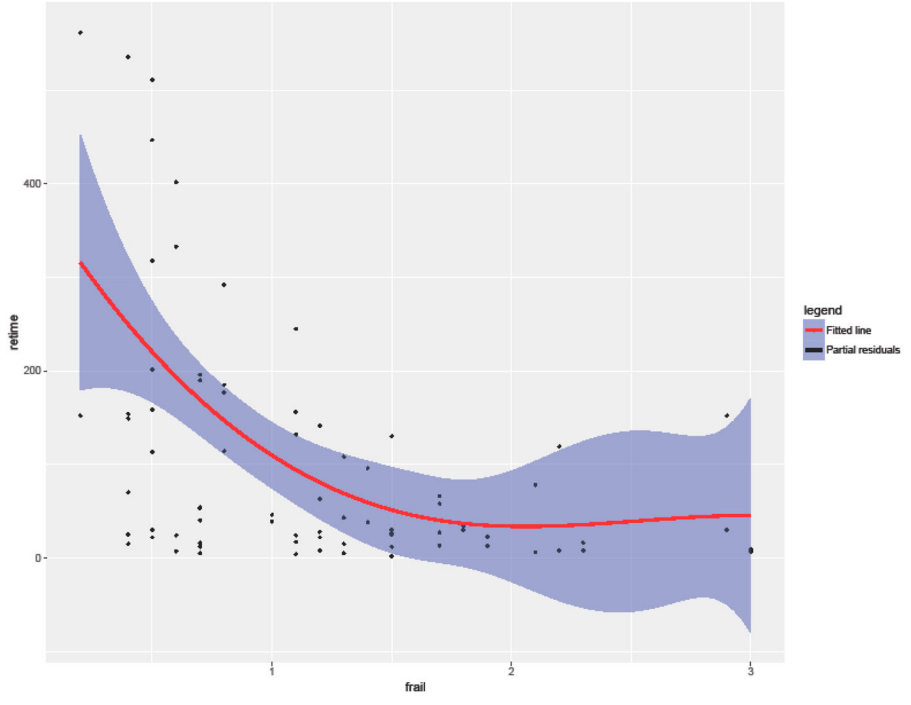

Figure 1. The kidney data: plot of the estimated model (52), modelling *g* as a smoothing spline with 95% confidence intervals. Also plotted are the *retime* partial residuals after regression on explanatory variables, *age*, *sex*, and *distype*.

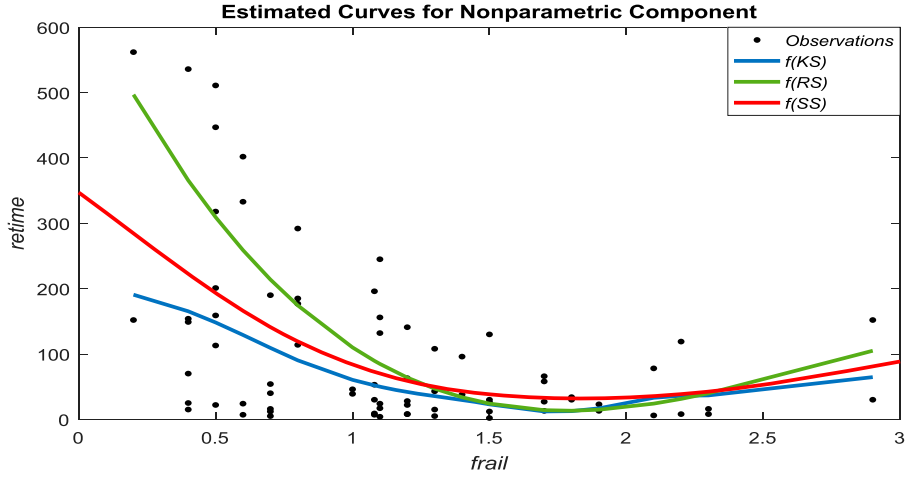

Figure 2. The estimates of the nonparametric component of the model (52).

and *RS* methods, respectively. In Figure 3, the shaded regions represent the confidence intervals based on the idea that bootstrapped standard errors tend to perform better than the pointwise confidence interval. The shaded the region in each panel is described by $\hat{g}(t) \pm 2.\text{se}(\hat{g}(t))$ where $\text{se}(\hat{g}(t))$ shows the bootstrapped standard error for function$\hat{g}(t)$.

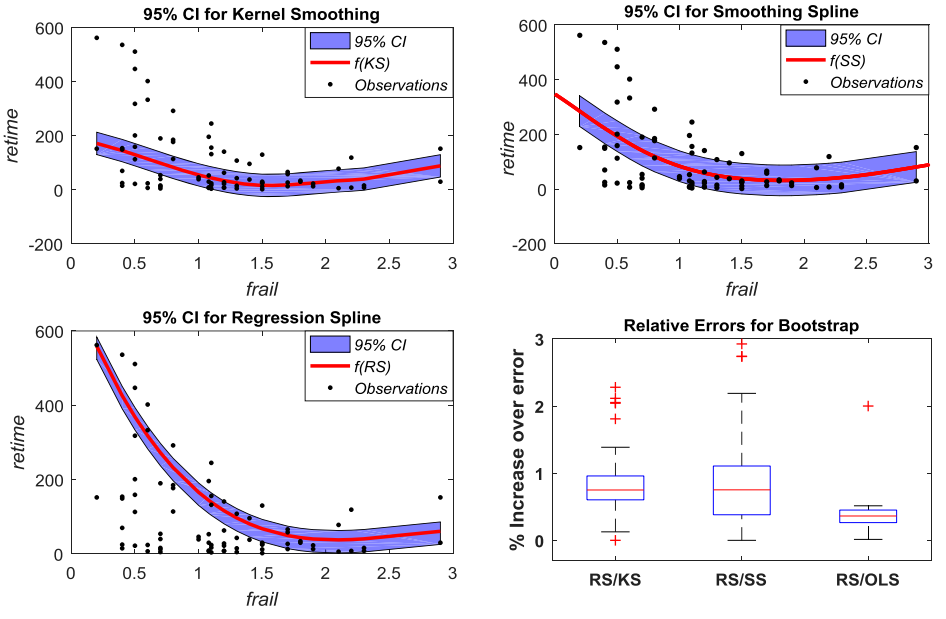

Figure 3. The upper panels and bottom left panel show the fitted curves with shaded regions' 95% bootstrapped confidence intervals for the nonparametric part of the model (52) by three different smoothing methods. Boxplots in the bottom right panel show the distribution of the relative bootstrapped errors over the six scenarios of the smoothing methods.

Note also that standard error bands are obtained by 1000 bootstrap repetitions. For *KS*, *SS* and *RS* methods, *MSE* values are 59.732, 42.088 and 32.373, respectively. The conclusion drawn from the information and figures is that *RS* curve provides a narrow standard error band and a good estimate of the regression function.

As shown in the previous results, we estimated the *RS* method as best among all the methods. For the kidney data set we computed the bootstrapped errors of *RS* and other possible methods. The boxplots in Figure 3 show the distribution of the errors using other possible methods relative to the favoured method. For each method (i.e. *SS*, *RS*, and *KS*) the distribution of the mentioned errors is constructed by

$$100 \times \left[\frac{Bootstrapped\,Error(Method) - min(Bootstrapped\,Error(Method))}{max(Bootstrapped\,Error(Method)) - min(Bootstrapped\,Error(Method))} \right]$$

over the three scenarios. When *KS* and *OLS* are compared to *RS*, it is seen that increase in the errors obtained from *RS* is decreasing. Since the range of the relative errors from the scenarios *RS*/*KS* and *RS*/OLS are lower and narrower than *RS*/*SS*, the *RS* seems to work well in all scenarios but *SS*.

6. Simulation study

A simulation study is carried out to demonstrate the impact of censoring and to assess the finite sample behaviours of the modified estimators. In our context, we first generate data

set(y_i, x_i, t_i), $i = 1, \ldots, n$ from the semiparametric regression model

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + g(t_i) + \varepsilon_i, \tag{53}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)' = (1, 2, -0.5)'$ and x_{i1}, x_{i2}, and x_{i3} are obtained by the standard normal distribution $N(0, 1)$; the ε_i's are generated from the standard normal distribution $N(0, 1)$; and the nonparametric component $g(.)$ is represented by two different functions:

$$g_1(t_i) = 3t_i \sin(t_i), \; t_i = 6((i - 0.5)/n)$$

and

$$g_2(t_i) = \sin(t_i) + 2\exp(-30t_i^2), t_i = ((i - 0.5)/n).$$

To introduce right censoring, we generate the censoring variable c_i from the exponential distribution with proportions at 5%, 25%, and 45%. For each censoring level (C.Ls) in the simulation, we generated 1000 random samples of size $n = 50,\ 100,$ and 200. Finally, from the censored model (53), we define ith indicator as $\delta_i = \mathrm{I}(y_i \leq c_i)$ and then the observed response as

$$z_i = \min(y_i, c_i).$$

Because of the censoring, the ordinary methods cannot be applied directly here to estimate the parameters of this model. For this reason, we consider transformed response (or synthetic) data points, as described in Section 2, to estimate the components of the model (53).

6.1. Results from the parametric component

Thousand estimates of $\boldsymbol{\beta} = (1, 2, -0.5)$'s forming the parametric part of model (53) for all sample sizes and censoring levels are obtained. The following figures and tables summarize the simulation results obtained from the semiparametric regression models using functions g_1 and g_2. For different sample sizes and censoring levels, the boxplots of the estimated regression coefficients are discussed in Appendix A4.

The main results from the simulation experiment are summarized in Table 3. The findings reported in this table are the *SMDE*s and variance values for the estimated coefficients by *KS*, *SS* and *RS* methods, as we illustrate in Section 3. Furthermore, the *RE* values of the mentioned smoothing methods with respect to the *OLS* are computed by (3.11). According to the findings in Table 3, the *RS* method performs better than the others, especially for the C.L = 25% and 45% under each sample size for two different models. It is also observed that for C.L = 45%, as the sample size increases, the *SMDE*s decrease for all methods.

In summary, the conclusion drawn from Table 3 is that the effect of the censoring tends to increase the variance of the estimators. Precision declines as the censoring level increases. In addition, precision is improved as the sample size increases. In order to examine the simulation results in detail, the simulated biases of parameters vector $\boldsymbol{\beta} = (1, 2, -0.5)$ are calculated for two semiparametric models given in Table 4. In general, the *RS* method provides the smallest bias, especially for C.L = 45%.

As in the real data example, simulation results show that the *RS* method usually performs better than the others in the sense of having smaller bias, variance, *SMDE* and *RE*

Table 3. Evaluation of parametric coefficients obtained by the proposed methods.

		$g_1(t_i) = 3t_i \sin(t_i)$								
		KS			*SS*			*RS*		
n	*C.Ls*	*SMDE*	$Var(\widehat{\beta})$	*RE*	*SMDE*	$Var(\widehat{\beta})$	*RE*	*SMDE*	$Var(\widehat{\beta})$	*RE*
50	5%	0.0006	0.0001	0.0069	0.0009	0.0003	0.0104	0.0009	0.0008	0.0104
	25%	0.0017	0.0016	0.0038	0.0126	0.0097	0.0283	0.0008	0.0008	0.0018
	45%	0.0026	0.0030	0.0019	0.0795	0.0791	0.0580	0.0013	0.0013	0.0009
100	5%	0.0001	0.0004	0.0025	0.0001	0.0001	0.0025	0.0004	0.0004	0.0100
	25%	0.0012	0.0010	0.0085	0.0087	0.0085	0.0619	0.0005	0.0005	0.0036
	45%	0.0019	0.0018	0.0059	0.0456	0.0454	0.1407	0.0010	0.0010	0.0031
200	5%	0.0001	0.0000	0.0053	0.0000	0.0000	0.0000	0.0002	0.0002	0.0106
	25%	0.0004	0.0000	0.0073	0.0034	0.0034	0.0623	0.0002	0.0002	0.0037
	45%	0.0015	0.0010	0.0122	0.0184	0.0184	0.1491	0.0006	0.0006	0.0049
					$g_2(t_i) = \sin(t_i) + 2\exp(-30t_i^2)$					
50	5%	0.0017	0.0020	0.0560	0.0041	0.0014	0.1311	0.0014	0.0004	0.0106
	25%	0.0087	0.0101	0.0356	0.0178	0.0129	0.7280	0.0049	0.0005	0.0201
	45%	0.0110	0.0139	0.0286	0.0380	0.0773	0.1031	0.0067	0.0057	0.0181
100	5%	0.0008	0.0005	0.0242	0.0010	0.0010	0.1084	0.0017	0.0002	0.0199
	25%	0.0072	0.0044	0.0144	0.0055	0.0093	0.1099	0.0021	0.0003	0.0416
	45%	0.0101	0.0052	0,1914	0.0215	0.0105	0.2280	0.0037	0.0003	0.0392
200	5%	0.0006	0.0005	0.0211	0.0002	0.0006	0.0779	0.0008	0.0001	0.0282
	25%	0.0008	0.0006	0.0490	0.0031	0.0010	0.1903	0.0009	0.0001	0.0519
	45%	0.0058	0.0056	0.0200	0.0079	0.0098	0.3017	0.0011	0.0002	0.0436

Table 4. Bias of estimated regression coefficients for all sample sizes and censoring levels.

		$g_1(t_i) = 3t_i \sin(t_i)$											
		$B(\widehat{\beta}_1)$				$B(\widehat{\beta}_2)$				$B(\widehat{\beta}_3)$			
n	*C.Ls*	*KS*	*SS*	*RS*	*OLS*	*KS*	*SS*	*RS*	*OLS*	*KS*	*SS*	*RS*	*OLS*
50	5%	0.003	0.032	0.001	0.003	0.017	0.008	0.002	0.004	0.012	0.000	0.000	0.006
	25%	0.115	0.050	0.007	0.014	0.056	0.011	0.004	0.028	0.014	0.016	0.007	0.089
	45%	0.074	0.013	0.000	0.040	0.001	0.013	0.006	0.027	0.009	0.018	0.005	0.027
100	5%	0.010	0.000	0.000	0.002	0.015	0.000	0.000	0.004	0.008	0.000	0.000	0.005
	25%	0.074	0.002	0.001	0.025	0.008	0.005	0.000	0.003	0.009	0.012	0.003	0.019
	45%	0.048	0.001	0.000	0.018	0.126	0.010	0.014	0.011	0.058	0.013	0.002	0.015
200	5%	0.002	0.001	0.000	0.005	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000
	25%	0.019	0.006	0.000	0.022	0.001	0.001	0.002	0.006	0.003	0.001	0.002	0.004
	45%	0.004	0.001	0.000	0.001	0.044	0.002	0.010	0.037	0.004	0.001	0.001	0.011
						$g_2(t_i) = \sin(t_i) + 2\exp(-30t_i^2)$							
50	5%	0.002	0.000	0.001	0.000	0.006	0.000	0.006	0.007	0.007	0.002	0.006	0.005
	25%	0.003	0.058	0.002	0.001	0.027	0.116	0.028	0.009	0.004	0.024	0.003	0.003
	45%	0.020	0.078	0.015	0.002	0.020	0.172	0.024	0.017	0.008	0.045	0.002	0.006
100	5%	0.000	0.006	0.000	0.000	0.003	0.006	0.003	0.007	0.000	0.000	0.000	0.000
	25%	0.001	0.035	0.002	0.010	0.006	0.063	0.004	0.023	0.004	0.015	0.003	0.011
	45%	0.006	0.054	0.003	0.012	0.036	0.129	0.049	0.001	0.020	0.041	0.020	0.010
200	5%	0.000	0.002	0.000	0.001	0.001	0.003	0.001	0.003	0.001	0.001	0.001	0.001
	25%	0.001	0.022	0.001	0.005	0.001	0.049	0.007	0.012	0.001	0.013	0.003	0.001
	45%	0.008	0.035	0.007	0.005	0.006	0.079	0.018	0.010	0.001	0.016	0.003	0.005

values. In our context, to show the performance of the *RS* method compared with others, the differences (49) between the *RS* and other estimators are summarized in Table 5.

The results reported in Table 5 denote that the numerical results are corresponding to the theoretical results and the non-negative definite condition is satisfied regarding the *RS* method. Also note that the mentioned estimator $\hat{\boldsymbol{\beta}}_{RS}$ maintains superiority. This means that the mentioned estimator $\hat{\boldsymbol{\beta}}_{RS}$ is also more efficient than the others in terms of having a minimum SMDE value.

Table 5. Differences of MDE matrices between the *RS* and others.

				$g_1(t_i) = 3t_i \sin(t_i)$					
	C.Ls for $n = 50$			*C.Ls for* $n = 100$			*C.Ls for* $n = 200$		
Differences	5%	25%	45%	5%	25%	45%	5%	25%	45%
$\Delta(\hat{\beta}_{KS}, \hat{\beta}_{RS})$	0.0000	0.0000	0.0020	0.0000	0.0010	0.0020	0.0000	0.0000	0.0040
$\Delta(\hat{\beta}_{SS}, \hat{\beta}_{RS})$	0.0000	0.0120	0.0790	0.0000	0.0080	0.0450	0.0000	0.0030	0.0180
$\Delta(\hat{\beta}_{OLS}, \hat{\beta}_{RS})$	0.0860	0.4450	1.3690	0.0400	0.0140	0.3240	0.0190	0.0540	0.1230
				$g_2(t_i) = \sin(t_i) + 2\exp(-30t_i^2)$					
$\Delta(\hat{\beta}_{KS}, \hat{\beta}_{RS})$	0.0032	0.0041	0.0056	0.0017	0.0020	0.0019	0.0008	0.0008	0.0011
$\Delta(\hat{\beta}_{SS}, \hat{\beta}_{RS})$	0.0030	0.0129	0.0313	0.0017	0.0034	0.0178	0.0008	0.0023	0.0067
$\Delta(\hat{\beta}_{OLS}, \hat{\beta}_{RS})$	0.0198	0.0195	0.0301	0.0869	0.0290	0.0570	0.0054	0.0078	0.0150

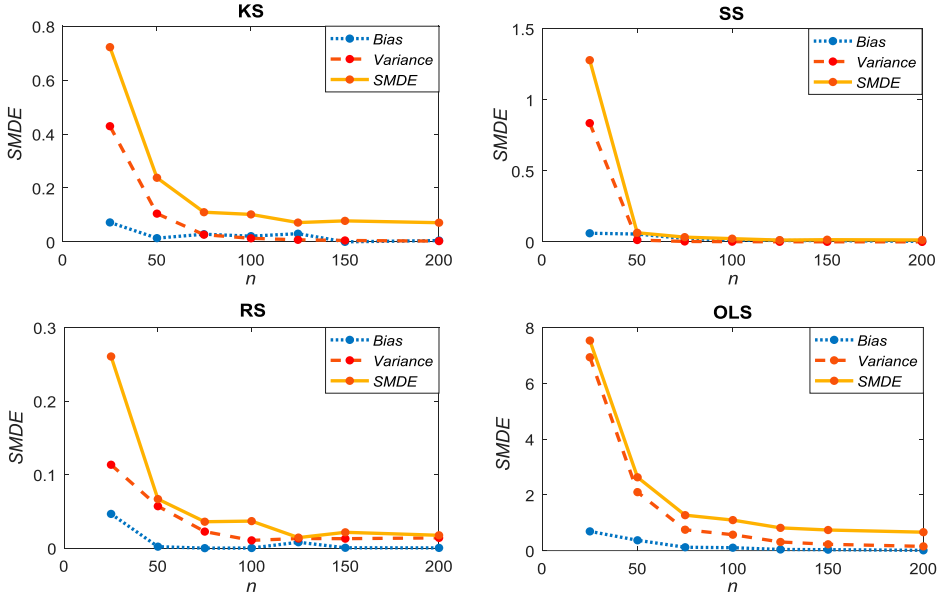

Figure 4. Panels show the *SMDE* (solid line), squared bias (dotted line), and variance values (dashed line) of the regression coefficients obtained by each method under different simulated data sets with censoring level 45%. Panels are for functio g_1.

6.2. Bias and variance decomposition

A useful way to assess the sources of estimation errors is to examine the bias and variance decomposition, as expressed in Equation (44). Figure 4 displays the bias and variance contributions to the *SMDE* values for the mentioned estimators (*KS*, *SS*, *RS*, and *OLS*) of regression coefficients for several samples of size 25, 50, 75, 100, 125, 150, and 200 under simulated data sets having a censorship rate of 45%. It is also noted that Figure 4 shows the results obtained by the model (53) using function g_1. The outcomes from the model with function g_2 are similar and, as such, the figure is not given here.

As can be seen from Figure 4, the values of both squared bias and variance can be reduced when the sample size is increased. It is also noted that both bias and variance contribute equally to the *SMDE* values as the size of the sample increases. From the right

Table 6. The estimated MSE values.

				$g_1(t_i) = 3t_i \sin(t_i)$					
	C.Ls for $n = 50$			C.Ls for $n = 100$			C.Ls for $n = 200$		
Methods	5%	25%	45%	5%	25%	45%	5%	25%	45%
KS	0.0593	1.1143	4.1445	0.0497	1.1023	3.8060	0.0425	1.0280	3.2565
SS	0.0513	1.1730	4.6388	0.0400	1.1905	4.4387	0.0351	1.0368	4.0637
RS	0.0418	1.1245	3.4443	0.0357	1.1727	3.3895	0.0344	1.0223	3.0016
				$g_2(t_i) = \sin(t_i) + 2\exp(-30t_i^2)$					
KS	0.0060	0.0752	0.1567	0.0090	0.0873	0.2002	0.0075	0.0890	0.1874
SS	0.0030	0.0990	0.1961	0.0137	0.0963	0.2307	0.0104	0.0988	0.2587
RS	0.0064	0.0746	0.1482	0.0032	0.0711	0.1891	0.0027	0.0778	0.1364

bottom panel of Figure 4, we see that the *OLS* performs extremely poorly due to high variance. It turns out that *OLS* method will provide a large *SMDE* if the variance is large. However, this variance can be reduced through smoothing in the defined estimators, such as *KS*, *RS* and *SS*, for the semiparametric regression models. In general, there is a tradeoff between bias and variance and this trade-off is governed by which smoothing parameter is selected to control bias and variance. In our context, under highly censored data, the *RS* gives better *SMDE* values than the *SS* and *KS* methods with regard to balancing both the squared bias and variance.

6.3. Results from the nonparametric component

As in the parametric components, we obtained 1000 estimates of the function g, which is the nonparametric component of model (53). For each method, 1000 replications were carried out and the estimated MSE values were computed for each estimator and corresponding each function g under the different censoring levels, given by

$$MSE(\hat{\mathbf{g}}, \mathbf{g}) = \frac{1}{1000} \sum_{j=1}^{1000} \sum_{i=1}^{n} \{\hat{g}(t_{ij}) - g(t_i)\}^2, \tag{54}$$

where $\hat{g}(t_{ij})$ shows the estimated value at the ith point of the function g in jth iterations. The results from Equation (54) are illustrated in Table 6.

We see in the second row of Table 3 that *SS* does poorly in terms of MSE values, especially for censoring levels 25% and 45%. In general, as in parametric component cases, *RS* method performs the best among all estimators.

In this simulation study, because 27 different configurations are made for nonparametric components g_1 and g_2, all of the configurations for two different nonparametric regression functions are given in Figures 5 and 6. The results appear to be quite reasonable for small ($n = 50$) sized samples under a low censoring level, C.L $= 5\%$. However, as shown in the right panels of the graphs, the estimated curves are not good, especially when the censorship rate is a high value for the same sized samples. In general, as sample sizes based on censored data sets get larger, estimates are get closer to each other and real function (see, the left panels of Figures 5 and 6 from top to bottom).

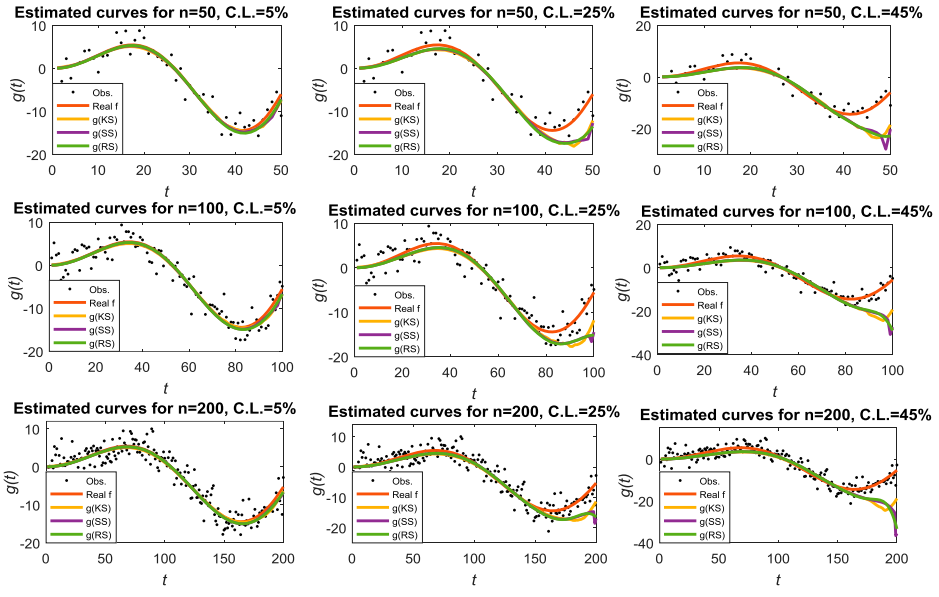

Figure 5. Panels show the observations, true regression function g_1, and three different estimated curves corresponding to the nonparametric part from *KS*, *RS,* and *SS*, respectively, for different sample sizes and censoring levels.

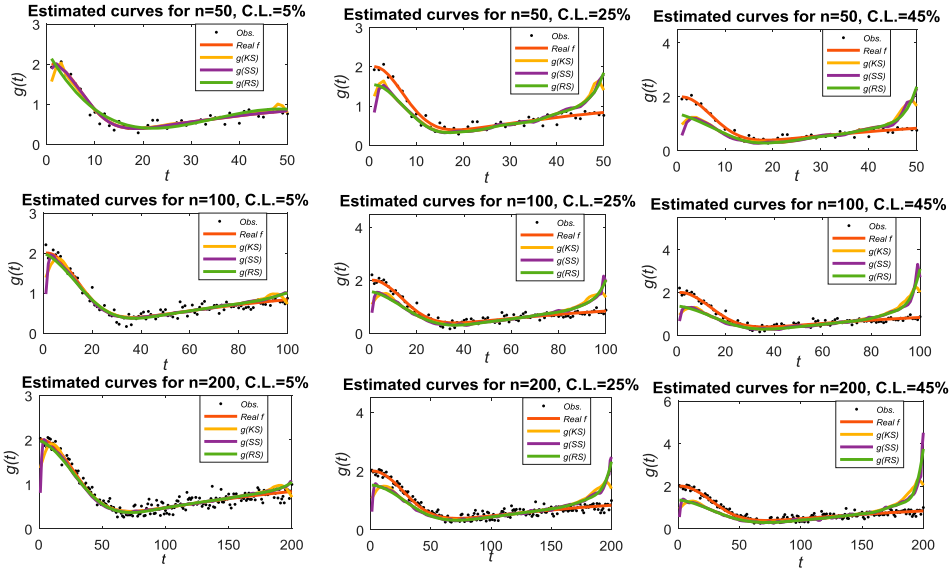

Figure 6. Similar to Figure 5 but for the function $g_2(t) = \sin(t) + 2\exp(-30t^2)$.

In Figures 5 and 6, it can be seen that estimated curves deviate mostly around the right end. As shown in Section 2, while response values are transformed to synthetic observations, they are sorted in an ascending way. Accordingly, great values of synthetic data would be around the right end of the axis. Therefore, major deviations of curves occur in the

right end. In detail, if Figure 5 and 6 are inspected, we see that curves in Figure 5 deviate downward and curves in Figure 6 upward. The reason for this is the negative values that relate to g_1 in the right end.

7. Summary and conclusions

Variously modified estimators are introduced to estimate the components of a semiparametic regression model when response observations are right censored. One of the estimators is defined by *OLS* method, while the remaining estimators are obtained with smoothing methods such as *SS*, *KS*, and *RS*, based on synthetic response observations. The mentioned synthetic response values have been defined with the appropriate modifications of the original observations (see [8]). After discussing the statistical estimation procedures required to obtain estimators, heavily censored, real kidney data and the Monte Carlo simulation experiments were used to compare the performances of the modified estimators.

The 1000 bootstrap samples were used to construct confidence intervals for heavily censored kidney data and the estimation results from this data are summarized in Tables 1 and 2 and Figures 1–3. Inspection of Table 1 and Figure 1 indicates that there is a nonlinear relationship between *retime* and *frail* variables. So, it is quite clear that a simple parametric model does not fit the censored data well while a semiparametric model fits observations more efficiently. As can be seen in Table 2 and Figure 2, the *RS* method usually leads to a better estimate of a censored semiparametric regression model. Furthermore, *OLS* provides the worst estimate for the parametric part of the semiparametric model (see Table 2 and bottom right panel of Figure 3). For the simulation studies, the outcomes of the numerical experiments are summarized in Tables 3–6 and Figures A1 and A2, and 4–6 . We conclude the following expressions from these tables and figures:

- For all the smoothing methods, the *SMDE*, variance, and bias values of the regression coefficients start to decrease as the sample size *n* gets larger.
- For small sample sizes, as expected the bias values of coefficients increase as the censoring levels increase.
- Also expected, when the sample size *n* increases, the *MSE* decreases even under higher censoring rates for all smoothing methods.
- The *RS* outperforms the *KS* and *SS* in the nonparametric part of the model for all simulation scenarios. However, the *SS* does quite poorly with regard to *MSE* values, especially for censoring levels 25% and 45% (see Table 4).
- Finally, when comparing the four methods, we see that the *RS* performs better than the other methods regarding the *SMDE*, *RE* variance and bias values of the estimates for all sample sizes, particularly when data sets are censored by rates of 25% and 45% (see Table 3 and 4).

Acknowledgements

We would like to thank the editor, the associate editor, and the anonymous referee for beneficial comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Engle RF, Granger WJ, Rice J, et al. Semiparametric estimates of the relation between weather and electricity sales. J Amer Statist Assoc. 1986;81:310–320.
[2] Wahba G. Cross validated spline methods for the estimation of multivariate functions from data on functionals. In: David HT, editor. Statistics: an appraisal proceedings 50th anniversary conference Iowa State Statistical Laboratory. Ames (IA): Iowa State University Press; 1984a. p. 205–235.
[3] Wahba G. Partial spline models for the semiparametric estimation of functions of several variables. Statistical Analysis of Time Series. Tokyo: Institute of Statistical Mathematics; 1984b. p. 319–329.
[4] Speckman P. Kernel smoothing in partial linear models. J R Stat Soc Ser B Methodol. 1988;50(3):413–436.
[5] Green PJ, Silverman BW. Nonparametric regression and generalized linear model. New York: Chapman & Hall; 1994.
[6] Miller RG. Least squares regression with censored data. Biometrika. 1976;63:449–464.
[7] Buckley J, James I. Linear regression with censored data. Biometrika. 1979;66(3):429–436.
[8] Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right-censored data. Ann Statist. 1981;9:1276–1288.
[9] Zheng Z. A class of estimators of the parameters in linear regression with censored data. Acta Math Appl Sin. 1987;3(3):231–241.
[10] Leurgans S. Linear models, random censoring and synthetic data. Biometrika. 1987;74:301–309.
[11] Lai TL, Ying Z, Zheng ZK. Asymptotic normality of a class of adaptive statistics with applications to synthetic data methods for censored regression. J Multivariate Anal. 1995;52:259–279.
[12] Dabrowska DM. Nonparametric quantile regression with censored data. Sankhya. 1992;54(2):252–259.
[13] Zheng Z. Strong consistency of nonparametric regression estimates with censored data. J Math Res Exposition. 1988;8:307–313.
[14] Fan J, Gijbels I. Censored regression: local linear approximations and their applications. J Amer Statist Assoc. 1994;89(426):560–570.
[15] Guessoum Z, Ould-Saïd E. On the nonparametric estimation of the regression function under censorship model. Statist Decisions. 2008;26:159–177.
[16] Qin J, Lawless J. Estimating equations, empirical likelihood and constraints on parameters. Canad J Statist. 1995;23(2):145–159.
[17] Qin GS, Cai L. Estimation for the asymptotic variance of parametric estimates in partial linear model with censored data. Acta Math Sci Engl Ed. 1996;16:192–208.
[18] Wang Q, Zheng Z. Asymptotic properties for the semiparametric regression model with randomly censored data. Sci China Ser A Math. 1997;40(9):945–957.
[19] Qin G, Jing B. Asymptotic properties for estimation of partial linear models with censored data. J Statist Plann Inference. 2000;84:95–110.
[20] Orbe J, Ferreira E, Núñez-Antón V. Censored partial regression. Biostatistics. 2003;4(1):109–121.
[21] Stute W. Distributional convergence under random censorship when covariables are present. Scand J Statist. 1996;23(4):461–471.
[22] Stute W. Nonlinear censored regression. Statist Sinica. 1999;9:1089–1102.
[23] Stute W. Consistent estimation under random censorship when covariables are present. J Multivariate Anal. 1993;45:89–103.
[24] Stute W. The central limit theorem under random censorship. Ann Statist. 1995;23:422–439.

[25] Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. Ann Statist. 1990;18(1):354–372.
[26] Wei LJ, Ying Z, Lin DY. Linear regression analysis of censored survival data based on rank tests. Biometrika. 1990;77(4):845–851.
[27] Zhou M. Asymptotic normality of the 'synthetic data' regression estimator for censored survival data. Ann Statist. 1992;20(2):1002–1021.
[28] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Amer Statist Assoc. 1958;53(282):457–481.
[29] Zheng ZK. Regression analysis with censored data [Ph.D. Dissertation]. Univ. of Colombia; 1984.
[30] Delecroix M, Lopez O, Patilea V. Nonlinear censored regression using synthetic data. Scand J Statist. 2008;35:248–265.
[31] Lemdani M, Ould-Saïd E. Exact asymptotic errors of the hazard rate kernel estimator under truncated and censored data. C R Acad Sci Ser I Math. 2001;333(11):1035–1040.
[32] Nadaraya EA. On estimating regression. Theory Probab Appl. 1964;10:186–190.
[33] Watson GS. Smooth regression analysis. Sankhya. 1964;26:359–372.
[34] Craven P, Wahba G. Smoothing noisy data with spline functions. Numer Math. 1979;31:377–403.
[35] Ruppert D, Wand MP, Carroll RJ. Semiparametric regression. New York: Cambridge University Press; 2003.
[36] Aydın D, Yılmaz E. Modified spline regression based on randomly right-censored data: a comparison study. Comm Statist Simulation Comput. 2017. doi: 10.1080/03610918.2017.1353615.
[37] Theobald CM. Generalizations of mean square error applied to ridge regression. J R Stat Soc Ser B. 1974;36(1):103–106.
[38] Arnold JC, Katti SK. An application of the Rao–Blackwell theorem in preliminary test estimators. J Multivariate Anal. 1972;2:236–238.
[39] Efron B. Bootstrap methods: another look at the jackknife. Ann Statist. 1979;7(1):1–26.
[40] Efron B. Censored data and the bootstrap. J Amer Statist Assoc. 1981;76(374):312–319.
[41] Akritas GA. Bootstrapping the Kaplan–Meier estimator. J Amer Statist Assoc. 1986;81(396):1032–1038.
[42] Efron B, Tibshirani RJ. Confidence intervals based on bootstrap tables An introduction to the bootstrap. New York: Chapman & Hall; 1993. p. 153–167.
[43] McGilchrist CA, Aisbett CW. Regression with frailty in survival analysis. Biometrics. 1991;47(2):461–466.
[44] Hastie T, Tibshirani R. Generalized additive models. New York: Chapman and Hall; 1990.
[45] Green P, Jennison C, Seheult A. Analysis of field experiments by least squares smoothing. J R Stat Soc Ser B. 1985;47:299–315.

Appendices. Supplemental technical materials

Appendix 1. Derivation of Equations (13)–(16)

First note that Equation (11) is minimized when $\boldsymbol{\beta}$ and $\mathbf{g}$ satisfy the following matrix equation:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{N} \\ \mathbf{N}'\mathbf{X} & (\mathbf{N}'\mathbf{N}+\lambda\mathbf{K}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{g} \end{pmatrix} = \begin{pmatrix} \mathbf{X}' \\ \mathbf{N}' \end{pmatrix} \mathbf{Y}_{\hat{G}}, \tag{A1}$$

where $\mathbf{K}$ is a $q \times q$ symmetric positive definite matrix, $\mathbf{X}$ and $\mathbf{N}$ are any matrices of dimension $n \times p$ and $n \times q$, respectively, as indicated before.

Equation (A1) can also be written as the pair of simultaneous matrix equations

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{N}\mathbf{g} = \mathbf{X}'\mathbf{Y}_{\hat{G}}, \tag{A2}$$

$$\mathbf{N}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{N}'\mathbf{N}+\lambda\mathbf{K})\mathbf{g} = \mathbf{N}'\mathbf{Y}_{\hat{G}}. \tag{A3}$$

From Equation (A3) we can easily obtain the smoothing spline estimator of the vector $\mathbf{g}$:

$$\hat{\mathbf{g}}_{ss} = (\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}). \tag{A4}$$

Substituting Equation (A4) into Equation (A2), we get

$$\begin{aligned}
&\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{N}((\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta})) = \mathbf{X}'\mathbf{Y}_{\hat{G}}\\
&\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'\mathbf{Y}_{\hat{G}} - \mathbf{X}'\mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}_{\hat{G}}\\
&\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}_{\hat{G}} - \mathbf{X}'\mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'\mathbf{Y}_{\hat{G}}\\
&\mathbf{X}'(\mathbf{I} - \mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}')\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'(\mathbf{I} - \mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}')\mathbf{Y}_{\hat{G}},
\end{aligned}$$

which provides the smoothing spline estimator for the vector $\boldsymbol{\beta}$, given by

$$\hat{\boldsymbol{\beta}}_{SS} = (\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Y}_{\hat{G}}, \tag{A5}$$

where $\mathbf{S}_\lambda = \mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'$ is a spline smoother matrix. Hence, replacing $\boldsymbol{\beta}$in Equation (A4) by the $\hat{\boldsymbol{\beta}}_{SS}$ will give the estimator for the unknown $\mathbf{g}$ defined by

$$\hat{\mathbf{g}}_{SS} = (\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'(\mathbf{Y}_{\hat{G}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{SS}) \tag{A6}$$

and hence fitted values

$$\hat{\mathbf{Y}}_{\hat{G}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{SS} + \mathbf{N}\hat{\mathbf{g}}_{SS}. \tag{A7}$$

Substituting $\hat{\boldsymbol{\beta}}_{SS}$ in Equation (A5) and $\hat{\mathbf{g}}_{SS}$in Equation (A6) into Equation (A7), we get

$$\begin{aligned}
\hat{\mathbf{Y}}_{\hat{G}} &= \mathbf{X}\hat{\boldsymbol{\beta}}_{SS} + \mathbf{N}\hat{\mathbf{g}} = \mathbf{X}[\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Y}_{\hat{G}} + (\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'(\mathbf{Y}_{\hat{G}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{SS})\\
&= \mathbf{X}[\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda) + \mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{X}[\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Y}_{\hat{G}}\\
&= \mathbf{S}_\lambda + \mathbf{X}[\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda) - \mathbf{S}_\lambda\mathbf{X}[\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Y}_{\hat{G}}\\
&= \mathbf{H}_\lambda^{SS}\mathbf{Y}_{\hat{G}},
\end{aligned}$$

where

$$\mathbf{H}_\lambda^{SS} = \mathbf{S}_\lambda + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}[\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda) \tag{A8}$$

as claimed.

Appendix 2. The implementation details of Equations (22)–(25)

Let $\mathbf{W}_\lambda$ be a smoother matrix defined by elements (21) and partial residuals of the variables $\mathbf{X}$ and $\mathbf{Y}_{\hat{G}}$ after adjusting the dependence on t,

$$\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X} = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip}),\ \ i = 1, 2, \ldots, n$$

and

$$\tilde{\mathbf{Y}}_G = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{Y}_{\hat{G}} = (\tilde{y}_{1\hat{G}}, \ldots, \tilde{y}_{n\hat{G}})'.$$

As suggested by Green et al. [45] , if $\mathbf{W}_\lambda$ is replaced by $(\mathbf{N}'\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}'$ in Equation (A4), it can be found to be a suitable estimator for the function g in Equation (21). given by

$$\hat{\mathbf{g}} = \mathbf{W}_\lambda(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}). \tag{A9}$$

Assuming that $\tilde{\mathbf{X}}$ has full rank, we can estimate the unknown vector $\boldsymbol{\beta}$ in Equation (A9) from the partial residuals $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}_G$. Considering the vector of the residuals in Equation (20), the minimizing

the weighted least square criterion (21) is

$$f_3(\boldsymbol{\beta}) = \sum_{i=1}^{n} (\boldsymbol{\varepsilon}_{i\hat{G}})^2 = \sum_{i=1}^{n} (\tilde{y}_{i\hat{G}} - \tilde{x}_i\boldsymbol{\beta})^2 = (\tilde{\mathbf{Y}}_{\hat{G}} - \tilde{\mathbf{X}}\boldsymbol{\beta})'(\tilde{\mathbf{Y}}_{\hat{G}} - \tilde{\mathbf{X}}\boldsymbol{\beta}).$$

Simplifying,

$$f_3(\boldsymbol{\beta}) = (\tilde{\mathbf{Y}}_{\hat{G}} - \tilde{\mathbf{X}}\boldsymbol{\beta})'(\tilde{\mathbf{Y}}_{\hat{G}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) = \tilde{\mathbf{Y}}'_{\hat{G}}\tilde{\mathbf{Y}}_{\hat{G}} - 2(\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}_{\hat{G}})\boldsymbol{\beta} + \boldsymbol{\beta}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta}. \tag{A10}$$

In order to minimize Equation (A10), we could differentiate with respect to $\boldsymbol{\beta}$ and set the derivative equal to zero:

$$\frac{\partial}{\partial \boldsymbol{\beta}} f_4(\boldsymbol{\beta}) = -2(\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}_{\hat{G}}) + 2\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} = 0. \tag{A11}$$

Setting Equation (A11) equal to zero and replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_{KS}$, the normal equations are obtained as

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{KS} = \tilde{\mathbf{X}}'\tilde{\mathbf{Y}}_{\hat{G}}. \tag{A12}$$

To solve for $\hat{\boldsymbol{\beta}}_{KS}$, multiply each side of Equation (A12) by $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$ to obtain the weighted least square regression solutions

$$\hat{\boldsymbol{\beta}}_{KS} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}_{\hat{G}} = \sum_{i=1}^{n} \tilde{x}_i\tilde{y}_{i\hat{G}} \Big/ \sum_{i=1}^{n} \tilde{x}_i^2, \tag{A13}$$

where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}_G$ denote the partial residuals, as defined in above. If the $\boldsymbol{\beta}$ in Equation (A9) is modified by $\hat{\boldsymbol{\beta}}_{KS}$ in Equation (A13), Equation (A9) can be re-expressed the following way:

$$\hat{\mathbf{g}}_{KS} = \sum_{j=1}^{n} w_{j\lambda}(t_j)((y_{j\hat{G}} - x_j\hat{\boldsymbol{\beta}}_{KS})) = \mathbf{W}_\lambda(\mathbf{Y}_{\hat{G}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{KS}) \tag{A14}$$

and the vector of fitted values is

$$\begin{aligned}
\hat{\mathbf{Y}}_{\hat{G}} &= \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{g}} = \mathbf{X}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}_{\hat{G}} + \mathbf{W}_\lambda(\mathbf{Y}_{\hat{G}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{KS}) \\
&= \mathbf{X}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{Y}_{\hat{G}} + \mathbf{W}_\lambda\mathbf{Y}_{\hat{G}} - \mathbf{W}_\lambda\mathbf{X}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{Y}_{\hat{G}} \\
&= \mathbf{W}_\lambda\mathbf{Y}_{\hat{G}} + \mathbf{X}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{Y}_{\hat{G}} \\
&= \mathbf{H}_\lambda^{KS}\mathbf{Y}_{\hat{G}},
\end{aligned}$$

where

$$\mathbf{H}_\lambda^{KS} = \mathbf{W}_\lambda + (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X}(\mathbf{X}'(\mathbf{I} - \mathbf{W}_\lambda)'(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{W}_\lambda)^2 \tag{A15}$$

as expressed in Section 2.3.

Appendix 3. The derivation details of Equations (32)–(35)

Let $\mathbf{D}$ be a diagonal penalty matrix, as in Section 2.3. For a given $\mathbf{D}$ matrix, the solution to the penalized objective function (30) provides the following system of equations:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{U} \\ \mathbf{U}'\mathbf{X} & (\mathbf{U}'\mathbf{U} + \lambda\mathbf{D}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}' \\ \mathbf{U}' \end{pmatrix} \mathbf{Y}_{\hat{G}}, \tag{A16}$$

where $\mathbf{X}$ and $\mathbf{U}$ are the design matrices, as defined in Section 2.3.

It is easy to show that Equation (A16) satisfies the following system of equations

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{U}\mathbf{b} = \mathbf{X}'\mathbf{Y}_{\hat{G}}, \tag{A17}$$

$$\mathbf{U}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})\mathbf{b} = \mathbf{U}'\hat{\mathbf{Y}}_G. \tag{A18}$$

Using Equation (A16), we obtain the modified regression spline estimator $\hat{\mathbf{g}}_{RS}$ of the coefficients vector $\mathbf{b}$ in Equation (28), given by

$$\hat{\mathbf{g}}_{RS} = (\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}) \tag{A19}$$

as expressed in Equation (33). Then, substituting this estimator of $\mathbf{b}$ into Equation (A17), we obtain

$$\begin{aligned}
&\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{U}((\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'(\mathbf{Y}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta})) = \mathbf{X}'\mathbf{Y}_{\hat{G}}\\
&\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'\mathbf{Y}_{\hat{G}} - \mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}_{\hat{G}}\\
&\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}_{\hat{G}} - \mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'\mathbf{Y}_{\hat{G}}\\
&(\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'\mathbf{X})\boldsymbol{\beta} = (\mathbf{X}' - \mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}')\mathbf{Y}_{\hat{G}}.
\end{aligned}$$

The last row of the above equation is solved for the vector $\boldsymbol{\beta}$, and the estimator $\hat{\boldsymbol{\beta}}_{RS}$ of the mentioned $\boldsymbol{\beta}$ vector in Equation (28) is obtained as

$$\hat{\boldsymbol{\beta}}_{RS} = (\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{Y}_{\hat{G}}, \tag{A20}$$

where $\mathbf{A}^{-1} = \mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'$.

Thus, fitted values

$$\hat{\mathbf{Y}}_{\hat{G}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{RS} + \mathbf{U}\hat{\mathbf{g}}_{RS}. \tag{A21}$$

Substituting $\hat{\boldsymbol{\beta}}_{RS}$ in Equation (A20) and $\hat{\mathbf{g}}_{RS}$in Equation (A19) into Equation (A21), it is obtained by

$$\begin{aligned}
\hat{\mathbf{Y}}_{\hat{G}} &= \mathbf{X}(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{Y}_{\hat{G}} + \mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'(\mathbf{Y}_{\hat{G}} - \mathbf{X}(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{Y}_{\hat{G}})\\
&= (\mathbf{X}(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1} + \mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}' - \mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'\mathbf{X}(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1})\mathbf{Y}_{\hat{G}}\\
&= (\mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}' + \mathbf{X}(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1} - \mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}'\mathbf{X}(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1})\mathbf{Y}_{\hat{G}}\\
&= \mathbf{H}_{\lambda}^{RS}\mathbf{Y}_{\hat{G}},
\end{aligned}$$

where

$$\mathbf{H}_{\lambda}^{RS} = \mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}' + (\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}')\mathbf{X}(\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}. \tag{A22}$$

This implies that the hat matrix in (35) is provided by (A21).

Appendix 4. Boxplots of regression coefficients

Figure A1 illustrates the boxplots of the estimated regression coefficients for the simulated data sets, which have a censorship rate of 5%. Figure A2 is similar to Figure A1 but for the observations with a censorship rate of 45%. The *x*-axis labels in the boxplots read as follows: In each panel 'K1, K2 and K3' denote the estimates of parameter vector $\boldsymbol{\beta}$ by *KS* method for $n = 50$, 100, and 200, respectively; similarly, 'S1, S2 and S3' indicate the cases using *SS* method for the sample sizes; 'R1, R2 and R3' denote the *RS* method cases; 'L1, L2 and L3' represent the cases of *OLS* method. The ordinate indicates the scale of regression coefficients vector $\boldsymbol{\beta}$.

As shown in Figures A1 and A2, when the sample size n gets larger, the range of estimates gets narrower. It also follows that the estimates under low censored (i.e. 5%) data are more stable than findings under highly censored (i.e. 45%) data. On the other hand, the outcomes corresponding to censoring level 25% are similar to the results displayed in Figure A1 and as such are not reported here. It is demonstrated that the sample size has an important effect on the quality of parametric estimates when high censoring levels are considered.

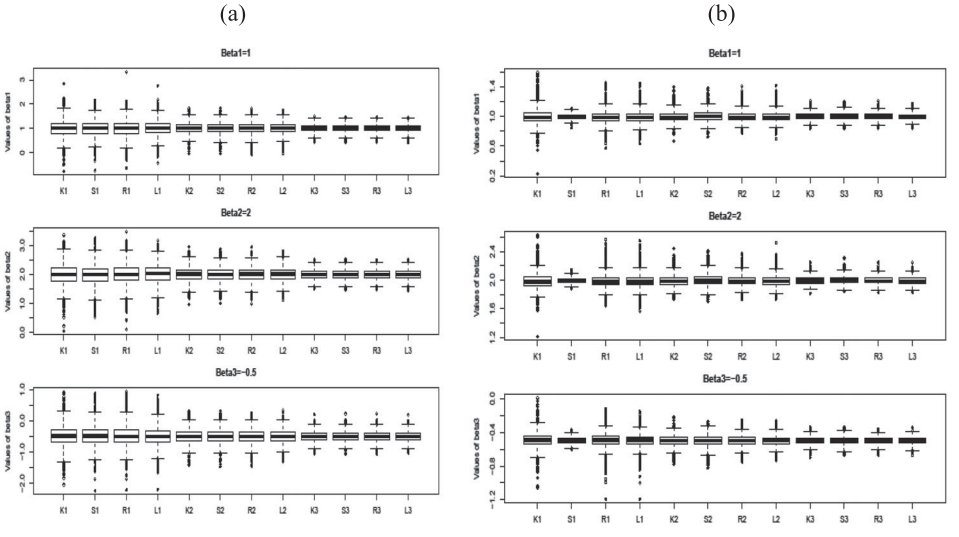

Figure A1. Boxplots of the estimates ($n = 50, 100, 200$) obtained from model (53) for censoring level 5%. Panels indicate the boxplots of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$; panel (a) is for $g_1(t) = 3t\sin(t)$ and panel (b) is for $g_2(t) = \sin(t) + 2\exp(-30t^2)$.

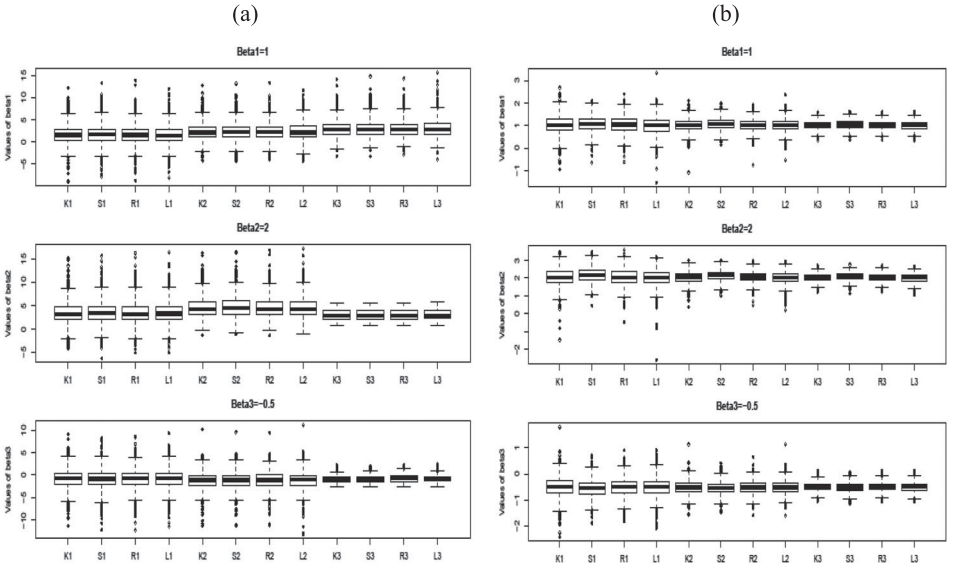

Figure A2. Similar to Figure 3 but for censoring levels 45%.