Hacettepe Journal of Mathematics and Statistics
Volume 46 (4) (2017), 669 – 683

Linear penalized spline model estimation using ranked set sampling technique

Al Kadiri M. A. *

Abstract

Benefits of using Ranked Set Sampling (RSS) rather than Simple Random Sampling (SRS) are indeed significant when estimating population mean or estimating linear models. Significance of this sampling method clearly appears since it can increase efficiency of the estimated parameters and decrease sampling costs. This paper investigates and introduces RSS method to fit spline and penalized spline models parametrically. It shows that the estimated parameters using RSS are more efficient than the estimated parameters using SRS for both spline and penalized spline models. The superiority of RSS approach is demonstrated using a simulation study as well as the "Air Pollution"environmental real data study. The approach in this paper can be illustrated for general smoothing spline models; for example B-spline,Radial spline etc, straightforwardly.

Keywords: Penalized generalized least square method, Ranked Set Sampling, Mallow criterion, efficiency.

2000 AMS Classification: 62D05

Received : 31.08.2015 *Accepted* : 22.11.2015 *Doi* : 10.15672/HJMS.201510314219

1. Introduction

Recent linear regression researches concern, with much attention, on fitting approaches that can accommodate data sets adequately as well as show the fitted model in a smooth fashion. A most popular regression approach, which will be discussed in this paper, is spline models. This model approach can accommodate the underlying trends of the data, which in some cases are curvilinear, in a linear regression model. It consists of piecewise lines that join at "knots"which gives a precise data representation than a single straight regression line. Furthermore, the piecewise lines with much rough can be penalized to appear smooth.

*Department of Statistics, Yarmouk University,Irbid, Jordan,
E-mail:alkadiri-m@yu.edu.jo (or) Alkadiri_moh@yahoo.com

Spline models although play a central role in regression because their computational properties and ability to gain appropriate fit, [7]. At early stages of their improvement, researchers developed spline models to scatter plot smoothing (e.g. [9]). They treated spline model as polynomial that can be improved in frame of knot selection (e.g. [25]) and basis functions (e.g. [10]). Introducing spline models to multivariate regression (e.g. [11]), nonparametric regression (e.g.[9]) and Bayesian models (e.g. [6]) took a wide range of interest in the literature. [21] made a considerable comparison between spline models.

Availability of various sampling methods challenge researchers to investigate appropriateness of these methods to gain better model estimates. A classical sampling method to fit spline models considers Simple Random Sampling (SRS) when selecting units. However, since it is practically more efficient, Ranked Set Sampling (RSS) has an increasing popularity when estimating regression models, [23]. This is because it can minimize sampling costs and furthermore, it can improve efficiency of the estimated parameters in the underlying model, [22].

[15], who firstly introduced RSS method, used it to estimate the population mean of yields in some fields. [19] provided the mathematical theory of this method. They proved that the estimated mean using RSS method is an unbiased estimator to the population mean as well as has less variance than usual SRS estimated mean. The recent monograph by [23] summarized all research linked to RSS method until that date. He presented the dramatic increase of using RSS method in different statistical fields (e.g. estimation, statistical testing, regression etc) as well as its practical efficiency in various research fields (e.g. environment, health science, epidemiology, agriculture etc). The RSS procedure was introduced to regression by [24] and [2]. By comparing the estimated model, researchers found the new estimated parameters using RSS had less variance than the estimated parameters using SRS, i.e. more efficient. This paper improves spline model fitting by using RSS as an alternative procedure to SRS. It improves parameter estimation because RSS decrease estimators' variances.

A simple spline model for n data points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ that have been selected by SRS method, can be expressed as follows

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{q} \beta_{2j}(x_i - K_j)_+ + e_i;\ i = 1, \dots, n. \tag{1.1}$$

where y is the response variable, x is the predictor variable, $\beta_0, \beta_1, \beta_{2j}$ are the model coefficients, e is the error term and K_j are the model knots where q is number of knots. The mathematical expression $(a)_+$ means the non-negative part of a; i.e. $max(0, a)$. Here we call the term $(x - K)_+$ by a linear spline basis function. Simply we can note that the spline model in (1.1) is a linear combination of these spline basis functions $1, x, (x - K_1)_+, ..., (x - K_q)_+$.

The set of knots are usually selected from the dense set of the predictor variable. A possible scenario for the selection method, which will be used in this paper, is equally-spaced with sufficiently large number of knots. Sufficiently large means, number of knots is around 35 as in most literature, see for example [14].

Settling the spline model (1.1) in matrix form gives

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.2}$$

where the design matrices of this model are

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \ \mathbf{X} = \begin{bmatrix} 1 & x_1 & (x_1 - K_1)_+ & \cdots & (x_1 - K_q)_+ \\ 1 & x_2 & (x_2 - K_1)_+ & \cdots & (x_2 - K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n - K_1)_+ & \cdots & (x_n - K_q)_+ \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{21} \\ \vdots \\ \beta_{2q} \end{bmatrix};$$

and $\boldsymbol{\varepsilon} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$.

A general model assumptions over the random error term $\boldsymbol{\varepsilon}$ assume that $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \Sigma$. During this research we keep the random error term independent of the predictor variable. Applying the generalized least square method yields the model fitting

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{1.3}$$

where $\hat{\boldsymbol{\beta}}$ is the minimizer of the quadratic form

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^{\mathbf{2}} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{1.4}$$

with closed solution $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$. The produced least square estimate is unbiased ; i.e $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, moreover, its covariance is $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-\mathbf{1}}$. Simply, one can note that the variances of model coefficients $\hat{\beta}_i$ are $\mathrm{Var}(\hat{\beta}_i) =$ [the i^{th} diagonal element of $(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}$]. An alternative simple model assumption considers uncorrelated errors with constant variance such that $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$; $\mathbf{I}$ is the identity matrix, which gives the least square estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{1.5}$$

Also, this leads the covariance matrix to be $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ which simply means that

$$\mathrm{Var}(\hat{\beta}_i) = \sigma^2[\text{the } i^{th} \text{ diagonal element of } (\mathbf{X}^T\mathbf{X})^{-1}]. \tag{1.6}$$

Building model inference, as we will see in next sections, needs to estimate σ^2. Implementing Sum Square Errors (SSE) is a common approach to produce an unbiased estimator for σ^2 as

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{n-p} = \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{n-p} \tag{1.7}$$

where n is the sample size and p is number of terms in the candidate model.

For spline model selection or goodness of fit, Mallows constant C_p is used in this paper accordingly. The C_p statistic can be attained as

$$C_p = ||\mathbf{y} - \hat{\mathbf{y}}||^2 + 2\hat{\sigma}^2 p. \tag{1.8}$$

A considerable improvement can be made to the fitted spline model in (1.2), which includes piecewise line segments that join at specific set of knots, is by handle a smooth fit. One possible method is by applying the penalized spline approach. [18] built a rich infrastructure for this model type as a smoothing regression technique. They investigated this technique under irregularly spaced knots, a basis of truncated power functions and a ridge penalty for spline coefficients. While the paper by [8] studied penalized spline technique under regularly spaced knots, a set of B-spline basis functions and a penalty on first or second order differences between neighboring spline coefficients. A well written paper discussed this approach and its theory is by [16]. [17] examined effect of number of knots on the degree of model smoothness. [3] studied asymptotic properties of the penalized estimators. [1] introduced penalized spline models to marginal models with application to longitudinal data.

Penalized spline approach mainly depends on avoid data over-fitting. Therefore, and in ordered to optimize the fitted model, the parameters β_{2j} in (1.2) are constrained to some conditions. Fortunately, there are quit a few choices for the penalization criteria, see for example [18]; chapter 3, but the easiest constraint to implement is to choose a constant C such as $\Sigma\beta_{2j}^2 \leq C$. Then the optimization problem becomes

$$(1.9) \qquad \min \ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^{\mathbf{2}} \ \text{subject to} \ \boldsymbol{\beta}^{\mathbf{T}}\mathbf{D}\boldsymbol{\beta} \ \leq \mathbf{C}$$

where the matrix $\mathbf{D}$ is a diagonal such that $\mathbf{D}$ = diag $\{\mathbf{0}_{2\times2}, \mathbf{1}_{q\times q}\}$ and q is number of knots in the candidate model.

Solving this minimization problem introduces a penalty term to the equation in (1.4) to penalize fits that have much rough and hence, produces a smoother fit. This produces penalized residual sum of square criterion

$$(1.10) \qquad ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^{\mathbf{2}} + \lambda^{\mathbf{2}}\boldsymbol{\beta}^{\mathbf{T}}\mathbf{D}\boldsymbol{\beta}$$

where λ is a non-negative smoothing parameter. The last term in (1.10) is called the penalty term.

Minimize (1.10) using penalized generalized least square method attains the following solution

$$(1.11) \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} + \lambda^{\mathbf{2}}\mathbf{D})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{y}$$

and therefore, the fitted penalized spline model can be written as $\hat{\mathbf{y}} = \mathbf{S}_\lambda\mathbf{y}$ where the "**smoothing matrix**"$\mathbf{S}_\lambda$ equals $\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} + \lambda^2\mathbf{D})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}$. The covariance matrix of model coefficient can be expressed as

$$(1.12) \qquad \text{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} + \lambda^{\mathbf{2}}\mathbf{D})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{X}(\mathbf{X}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{X} + \lambda^{\mathbf{2}}\mathbf{D})^{-\mathbf{1}}.$$

In the penalized spline context, two parameters need to be estimated. The smoothing parameter λ and the covariance matrix Σ. The smoothing parameter λ is often chosen by minimizing the generalized cross-validation (GCV), [5], such that

$$(1.13) \qquad \text{GCV}(\lambda) \quad = \quad \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{[1 - n^{-1}tr(\mathbf{S}_\lambda)]^2} = \sum_{i=1}^{n}\left[\frac{\{(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}\}_i}{1 - n^{-1}tr(\mathbf{S}_\lambda)}\right]^2$$

where $tr(.)$ is the trace of a matrix. While the covariance matrix Σ is estimated by [20]. He proposed an unbiased estimator for the covariance matrix Σ under the simple assumption which considers $\text{Cov}(\boldsymbol{\varepsilon}) \equiv \Sigma = \sigma^2\mathbf{I}$.

The proposed estimator for σ^2 can be expressed as follows

$$(1.14) \qquad \hat{\sigma}^2 = \frac{||\mathbf{y} - \hat{\mathbf{y}}||^{\mathbf{2}}}{n - tr(\mathbf{S}_\lambda)}.$$

This produces the following estimator for the covariance matrix $\hat{\Sigma} = \hat{\sigma}^2\mathbf{I}$. Consequently, the C_p criterion can be calculated after using $\hat{\sigma}^2$ as

$$(1.15) \qquad C_p = ||\mathbf{y} - \hat{\mathbf{y}}||^2 + 2\hat{\sigma}^2 tr(\mathbf{S}_\lambda).$$

In the previous method model fitting, we consider SRS method when selecting sampling units. However in this paper, we introduce a RSS method to estimate spline and penalized spline models. This sampling method, which selects sampling units after spread them in a proceeding manner, verified its quality in many practical modeling situations, [23]. In what follows, we describe a RSS method for general statistics and for simple linear regression in specific.

When the RSS procedure was firstly established, [15] divided sample units in distinguished subsamples then each subsample had been ordered in a proceeding manner separately. Particularly, he selected m simple random subsamples each of size m from the target population, say $\{x_1, x_2, ..., x_m\}_1; \{x_1, x_2, ..., x_m\}_2; ...;$ $\{x_1, x_2, ..., x_m\}_m$. Then he ordered each subsample separately to produce ranked subsets $\{x_{(1)}, x_{(2)}, ..., x_{(m)}\}_1$; $\{x_{(1)}, x_{(2)}, ..., x_{(m)}\}_2$;...;$\{x_{(1)}, x_{(2)}, ..., x_{(m)}\}_m$. Note that till this step, no actual quantification has been made for the selected units. Finally, McIntyre selected and measured the i^{th} smallest unit from the i^{th} subsample, $x_{(i)i}$. This means the produced RSS units are: $\{x_{(1)1},\ x_{(2)2}, ..., x_{(m)m}\}$. This procedure popularly known in the literature by balanced RSS. Generally, this procedure can be repeated r times, where each repetition called a cycle, to generate the desired RSS size $n = rm$, where n is the SRS sample size.

Essentially, McIntyre's new method is practically effective when sampling units are expensive or hard to measure, however rank few units, without real quantification, is relatively cheaper. Attain ordering for sampling units can be made by an expert or an analyst judgment visually or by any other relatively cheap method.

For regression, the RSS procedure can be extended similarly as the above procedure. Only, we need to note that ordering sample units can be achieved either on the response or on the predictor variables. In the following example, we consider the case of ordering the response variable y. Also, and for simple presentation, we assume a simple regression model (i.e. the model has one predictor variable x). The SRS sample units can be denoted as $(x_i, y_i); i = 1, 2, \cdots, n$.

In this example, assume the desired RSS size is $m = 3$. For this purpose, consider we have the following 3 subsamples each of size 3 pairs: $\{(x_1, y_1)_1$, $(x_2, y_2)_1$, $(x_3, y_3)_1\}$,$\{(x_1, y_1)_2$,$(x_2, y_2)_2, (x_3, y_3)_2\}$ and $\{(x_1, y_1)_3, (x_2, y_2)_3, (x_3, y_3)_3\}$. Before measuring any sample unit, order these subsamples separately according to the response variable. Ordering can be performed by any relatively cheap method. Then from the first subsample, choose the first minimum-response value linked with the correspondence predictor value; which can be denoted by $(x_{[1]}, y_{(1)})_1$, from the second subsample choose the second minimum-response pair $(x_{[2]}, y_{(2)})_2$ and finally, from the last subsample choose the maximum-response pair $(x_{[3]}, y_{(3)})_3$. Generally in this research, the pair $(x_{[i]}, y_{(i)})_j$ means that i^{th} predictor value $x_{[i]}$ corresponds to the i^{th} minimum-response value $y_{(i)}$ from the j^{th} subsample. So, the yielded RSS set of size 3 is $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_2, (x_{[3]}, y_{(3)})_3\}$ which can be used to estimate the regression model.

It is important to mention for general applications and in order to achieve comparison, we need to increase number of the RSS sampling units to satisfy equality with SRS sample size. This can be produced if we repeat the above RSS samples r times or cycles, i.e $n = rm$, where n is the SRS sample size. Thus, the produced RSS of size 3 can be denoted as $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_2, (x_{[3]}, y_{(3)})_3\}_1$, $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_2, (x_{[3]}, y_{(3)})_3\}_2$, ..., $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_2, (x_{[3]}, y_{(3)})_3\}_r$.

Equivalently , the above RSS procedure can be demonstrated when the predictor variable need to be ranked rather than the response variable.

This paper introduces the above RSS procedure to fit spline and penalized spline models where ranking the response variable or the predictor variable is achieved over sampling units. The efficiency of the new estimators in the spline models are compared with SRS estimators. Finally, a simulation study as well as a practical example are illustrated to verify our results. A considerable note should be mentioned here that is, this paper investigates improvements that can be made for parameters efficiencies in the penalized spline models however, the degree of smoothness is not our target.

The next two sections define the RSS procedure, that has been described above, for spline and penalized spline models in which models' parameters are estimated using the new sampling units and the efficiency for these parameters are compared with SRS procedure.

2. Spline model estimation using RSS

Demonstrations of RSS procedure to select sample units and fit spline models are achieved in this section. Firstly, in subsection (2.1), the RSS sampling units are gained after rank the response variable and illustrated to estimate the spline models. Then, in a similar fashion, in subsection (2.2), the entire process is applied again however this time after rank a predictor variable. Evaluation of the RSS method is made with comparing to SRS as well as after computing C_p goodness of fit criterion. At the end of this section, we found the new sampling scheme, RSS, achieved better performance than SRS scheme when fitting spline models.

2.1. Spline models with ranked response variable. Mainly in this subsection, spline model fitting is achieved using RSS units after order the response variable. We illustrated the method described at the end of the introduction to produce the following RSS units: $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_2, ..., (x_{[m]}, y_{(m)})_m\}_1$, $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_2, ..., (x_{[m]}, y_{(m)})_m\}_2$, ..., $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_2, ...,(x_{[m]}, y_{(m)})_m\}_r$ where r is number of cycles that RSS need to be repeated to achieve equality $n = rm$, n is the SRS size. Now, the produced RSS sample is available to estimate the proposed spline model.

The spline linear model, after implement RSS units, can be written similar to model (1.1) as follows

$$(2.1)\qquad y_{(i)j} = \beta_0^* + \beta_1^* x_{[i]j} + \sum_{l=1}^{q} \beta_{2l}^*(x_{[i]j} - K_l)_+ + e_{(i)j}^*;\ i = 1, \dots, m;\ j = 1, \dots, r.$$

where $y_{(i)j}$ is i^{th} smallest response unit that has been selected from i^{th} subsample in the j^{th} cycle , $x_{[i]j}$ is the predictor variable that is associated with $y_{(i)j}$; β_0^*, β_1^* and β_{2l}^* are model parameters. Here $K_1, ..., K_q$ are model knots; for a suitable number of knots q; and $e_{(i)j}$ is the random error term. The produced model in matrix entity can be written as

$$(2.2)\qquad \mathbf{y}_{(RSS)} = \mathbf{X}_{[RSS]}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_{(RSS)}$$

where

$$\mathbf{y}_{(RSS)} = \begin{bmatrix} y_{(1)1} \\ \vdots \\ y_{(m)1} \\ \vdots \\ y_{(1)r} \\ \vdots \\ y_{(m)r} \end{bmatrix};\ \mathbf{X}_{[RSS]} = \begin{bmatrix} 1 & x_{[1]1} & (x_{[1]1} - K_1)_+ & \cdots & (x_{[1]1} - K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[m]1} & (x_{[m]1} - K_1)_+ & \cdots & (x_{[m]1} - K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[1]r} & (x_{[1]r} - K_1)_+ & \cdots & (x_{[1]r} - K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[m]r} & (x_{[m]r} - K_1)_+ & \cdots & (x_{[m]r} - K_q)_+ \end{bmatrix}$$

$\boldsymbol{\beta}^* = \begin{bmatrix}\beta_0^* & \beta_1^* & \beta_{21}^* \cdots \beta_{2q}^*\end{bmatrix}^T$; $\boldsymbol{\varepsilon}_{(RSS)} = \begin{bmatrix}e_{(1)_1}^* & e_{(2)_1}^* \cdots e_{(m)r}^*\end{bmatrix}^T$.

Model assumptions assume uncorrelated errors with non-constant variance and zero mean. Re-writing these assumptions in matrix notation produces

$$(2.3)\qquad E(\boldsymbol{\varepsilon}_{(RSS)}) = \mathbf{0} \text{ and}$$

$$\text{Cov}(\boldsymbol{\varepsilon}_{(RSS)}) = \text{diag}\{\sigma_{(1)}^{*2}, ..., \sigma_{(m)}^{*2}, ..., \sigma_{(1)}^{*2}, ..., \sigma_{(m)}^{*2}\}_{mr \times mr} \equiv \Sigma^*.$$

Keeping the non-constant variance assumption in (2.3) needs an appropriate method to estimate variance components. A popular method to achieve this goal, when the likelihood is general, is by using Feasible Generalized Least Square algorithm (FGLS), [16]. Computer statistical softwares are rich with packages to compute this algorithm. For example, the package RFGLS in R software is a direct algorithm. If a simple assumption is proposed by assuming constant variance hence, model assumptions becomes

$$(2.4)\qquad E(\boldsymbol{\varepsilon}_{(RSS)}) = \mathbf{0} \text{ and } \mathrm{Cov}(\boldsymbol{\varepsilon}_{(RSS)}) \equiv \Sigma^* = \sigma^{*2}\mathbf{I}.$$

Using the generalized least square method to minimize $||\mathbf{y}_{(RSS)} - \mathbf{X}_{[RSS]}\boldsymbol{\beta}^*||^2$ produces

$$(2.5)\qquad \hat{\boldsymbol{\beta}}^* = (\mathbf{X}^T_{[RSS]}\Sigma^{*-1}\mathbf{X}_{[RSS]})^{-1}\mathbf{X}^T_{[RSS]}\Sigma^{*-1}\mathbf{y}_{(RSS)}$$

where the covariance matrix of these estimated coefficients is
$\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = (\mathbf{X}^T_{[RSS]}\Sigma^{*-1}\mathbf{X}_{[RSS]})^{-1}$. This generates the following estimated variance for the model coefficient $\hat{\beta}^*_i$

$$(2.6)\qquad \widehat{Var}(\hat{\beta}^*_i) = \text{ [the } i^{th} \text{ diagonal entry of } (\mathbf{X}^T_{[RSS]}\widehat{\Sigma}^{*-1}\mathbf{X}_{[RSS]})^{-1}]$$

where $\widehat{\Sigma}^{*-1}$ is the estimated covariance matrix. Accordingly, C_p can be calculated similar to (1.8).

Considerably, the produced estimator $\hat{\boldsymbol{\beta}}^*$ is unbiased estimator for the model parameter $\boldsymbol{\beta}$ and its covariances satisfies $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) \leq \mathrm{Cov}(\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the least square estimate of $\boldsymbol{\beta}$ when using SRS as defined in (1.5). Proof of the first property is straightforward whilst proof of the second property was attained numerically as seen in the simulation study Table (1).

To demonstrate improvement of our new procedure, we compute the relative efficiency concept using the following definition

$$(2.7)\qquad eff(\hat{\beta^*}_i, \hat{\beta}_i) = \frac{\widehat{\mathrm{Var}}(\hat{\beta}_i)}{\widehat{\mathrm{Var}}(\hat{\beta}^*_i)}$$

which can indicate which estimator is better.

The second property with support of Table (1), can show that the fitted spline model using RSS is more efficient than the fitted spline models using SRS where, $\mathrm{eff}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\beta}}) \geq \mathbf{1}$.

2.2. Spline models with ranked predictor. In the same imperative manner that has been improved in the previous subsection, RSS can be easily extended to fit spline models where RSS sampling units are produced after order the predictor variable.

Analogous to model (1.1), but this time we order the predictor variable to produce RSS units, the following spline model is produced

$$y_{[i]j} = \beta^*_0 + \beta^*_1 x_{(i)j} + \sum_{l=1}^{q}\beta^*_{2l}(x_{(i)j} - K_l)_+ + e^*_{[i]j}\ ;\quad i = 1, \cdots, m;\quad j = 1, \cdots, r.$$

where $x_{(i)j}$ is i^{th} smallest unit of the predictor variable from the i^{th} subsample in the j^{th} cycle, $y_{[i]j}$ is the response variable that associate with $x_{(i)j}$; β^*_0, β^*_1 and β^*_{2l} are the model parameters,$K_1, ..., K_q$ are the model knots and $e^*_{[i]j}$ is the random error term.

Settle the above model in matrix form produces

$$(2.8)\qquad \mathbf{y}_{[RSS]} = \mathbf{X}_{(RSS)}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_{(RSS)}.$$

Matrices of the above model can be defined similarly as in model (2.2) with the same model assumptions.

Minimizing the least square criterion of $||\mathbf{y}_{[RSS]} - \mathbf{X}_{(RSS)}\boldsymbol{\beta}^*||^2$ gives the solution

$$(2.9)\qquad \hat{\boldsymbol{\beta}}^* = (\mathbf{X}^T_{(RSS)}\Sigma^{*-1}\mathbf{X}_{(RSS)})^{-1}\mathbf{X}^T_{(RSS)}\Sigma^{*-1}\mathbf{y}_{[RSS]}.$$

The covariance matrix for the above estimated coefficient can be defined as follows

$$(2.10)\qquad \mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = (\mathbf{X}^T_{(RSS)}\Sigma^{*-1}\mathbf{X}_{(RSS)})^{-1}.$$

Importantly, the produced estimator $\hat{\boldsymbol{\beta}}^{*}$ in 2.9 is unbiased estimator for the model parameter $\boldsymbol{\beta}$ and its covariance satisfies $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^{*}) \leq \mathrm{Cov}(\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the least square estimate of $\boldsymbol{\beta}$ when using SRS as defined in (1.5). A proof for the second property of the above estimator was gained numerically as realized from Table (2) in the simulation study section (4). This proof depends mainly on the definition of the relative efficiency which is defined in (2.7). Another model evaluation principle is C_p criterion which is defined in (1.8) but after introducing the RSS units to the spline model.

The second property with support of Table (2), can show that the fitted spline model using RSS are more efficient than ones that fitted using SRS where,
$\mathrm{eff}(\hat{\boldsymbol{\beta}}^{*}, \hat{\boldsymbol{\beta}}) \geq \mathbf{1}$.

3. Penalized spline model estimation using RSS

A more convenient model than simple spline model is smooth spline regression model. This is because smooth models can fit data appropriately and capture the underlying relation efficiently. A specific smoothing spline model is penalized spline model which will be considered in this section. Suppose a general spline model with sufficiently large number of knots; q, as established in (2.2). Also, suppose the same design matrices as in (2.2) and the same model assumption as in (2.3). Note that the generated design matrices consider ranking the response variable.

The least square criterion can fit this model by minimizing the penalized sum of square errors

$$(\mathbf{y}_{(\mathbf{RSS})} - \mathbf{X}_{[\mathbf{RSS}]}\boldsymbol{\beta}^{**})^{\mathbf{T}}(\mathbf{y}_{(\mathbf{RSS})} - \mathbf{X}_{[\mathbf{RSS}]}\boldsymbol{\beta}^{**}) + \lambda^{*\mathbf{2}}\boldsymbol{\beta}^{**\mathbf{T}}\mathbf{D}\boldsymbol{\beta}^{**}$$

with respect to $\boldsymbol{\beta}^{**}$ and for some smoothing parameter λ^{*}. The matrix $\mathbf{D} = \mathrm{diag}\{\mathbf{0}_{2\times 2}, \mathbf{1}_{q\times q}\}$. The penalized least square method gives the following linear smoother $\hat{\mathbf{y}}_{(RSS)} = \mathbf{H}_{\lambda^{*}}\mathbf{y}_{(RSS)}$ where the smoothing matrix is
$\mathbf{H}_{\lambda^{*}} = \mathbf{X}_{[RSS]}(\mathbf{X}_{[RSS]}^{T}\Sigma^{**-1}\mathbf{X}_{[RSS]} + \lambda^{*2}\mathbf{D})^{-1}\mathbf{X}_{[RSS]}^{T}\Sigma^{**-1}$. Consequently, the estimated model coefficient matrix can be written in the form

$$\hat{\boldsymbol{\beta}}^{**} = (\mathbf{X}_{[RSS]}^{T}\Sigma^{**-1}\mathbf{X}_{[RSS]} + \lambda^{*2}\mathbf{D})^{-1}\mathbf{X}_{[RSS]}^{T}\Sigma^{**-1}\mathbf{y}_{(RSS)}. \tag{3.1}$$

Accordingly, the smoothing parameter λ^{*} can be estimated using GCV concepts as defined in (1.13).

Parallel to (1.12), the covariance matrix of the estimated model coefficients can be written as

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}^{**}) \equiv \widehat{MSE}(\hat{\boldsymbol{\beta}}^{**}) = \tag{3.2}$$

$$\left[(\mathbf{X}_{[RSS]}^{T}\hat{\Sigma}^{**-1}\mathbf{X}_{[RSS]} + \lambda^{*2}\mathbf{D})^{-1}\mathbf{X}_{[RSS]}^{T}\hat{\Sigma}^{**-1}\mathbf{X}_{[RSS]}(\mathbf{X}_{[RSS]}^{T}\hat{\Sigma}^{**-1}\mathbf{X}_{[RSS]}^{T} + \lambda^{*2}\mathbf{D})^{-1}\right].$$

where $\widehat{\Sigma}^{**}$ is the estimated covariance matrix and MSE is Mean Square Error. Also, the C_p constant can be computed using (1.15).

Substantially, the covariance of generated estimator $\hat{\boldsymbol{\beta}}^{**}$ satisfies $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^{**}) \leq \mathrm{Cov}(\hat{\boldsymbol{\beta}})$ where $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ is the least square estimate of for $\mathrm{Cov}(\boldsymbol{\beta})$ when using SRS as defined in (1.5). This property was realized as seen in the simulation study, Table (3). We identified these variance components using definition of the relative efficiency as in (3.3).

The relative efficiency of the estimated model coefficients under RSS scheme, comparing to SRS scheme, can be computed in terms of MSE as

$$eff(\hat{\beta}_i^{**}, \hat{\beta}_i) = \frac{\widehat{MSE}(\hat{\beta}_i)}{\widehat{MSE}(\hat{\beta}_i^{**})} \tag{3.3}$$

where $\widehat{MSE}(\hat{\beta}_i^{**})$ is the i^{th} diagonal element of the matrix (3.2) when RSS sampling units are considered in the penalized spline model. Similarly, $\widehat{MSE}(\hat{\beta}_i)$ is the estimated diagonal element of the covariance matrix (1.12) under the SRS penalized spline model.

Demonstration of the RSS method to fit penalized spline models after ordering the predictor variable is straightforward. To achieve this goal, we follow the steps of this section and model construction after providing the design matrices by new RSS units. Model evaluation conducted by computing the relative efficiencies of the estimated parameters which are summarized in Table (4).

Regarding to the above results with support of the simulation study, Table (3) and Table (4), we can conclude that the fitted penalized spline models using RSS are more efficient than ones that fitted using SRS where, eff$(\hat{\boldsymbol{\beta}}^{**}, \hat{\boldsymbol{\beta}}) \geq 1$.

Next section, a simulation study is conducted to show main properties of our new sampling method that has been discussed in this paper. Tables of the relative efficiencies of the model parameters are presented.

4. Simulation study

To illustrate the practical performance of estimating spline and penalized spline models using RSS scheme, computer artificial studies were conducted with the following general set up. Data sets were generated from the smooth relation: $y_i = f(x_i) + e_i$, such that $f(x) = 2\ sin(x)\ exp(-x^2)$ and x has $Uniform(-2, 2)$ distribution. The error terms e_i were assumed uncorrelated with 0 mean and 0.122 constant variance. We proposed RSS samples of size $m = 2, 3$ and 4 units with specific number of cycles r to perform the relation $n = rm$, where n is the SRS size. Each yielded RSS sample was used to estimate a spline model with 3 knots then it was used to estimate a penalized spline model with the same number of knots.

Our specific selection for small number of knots; $i.e\ q = 3$, is to enhance comfortable visibility for the produced tables where each table will only have 5 estimated parameters $\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_{21}^*, \hat{\beta}_{22}^*$ and $\hat{\beta}_{23}^*$. Despite the small number of knots we used, performance of our method when we increase number of knots to be large is statistically indistinguishable.

For sake of comparison, the same smoothing model above was used to generate SRS samples of size $n = 4, 6, 9, 12$ and 24. The yielded SRS samples were used to estimate spline and penalized spline models with 3 knots. This small number of knots is to allow comparison with the simulated RSS that have same number of knots. Last point to mention that all configurations in this simulation study were ran with 10000 replicates.

4.1. Simulated spline models. According to the above simulation arrangements, RSS samples were produced after ranking the response variable as discussed in subsection (2.1). Then the generated data sets were used to estimate the spline model in (2.2) under the proposed assumptions in (2.3) by using (2.5).

To enhance model comparison, the generated SRS samples in the above section (4) were used to estimate the spline model (1.2) via (1.3).

Outputs of these simulation trails are summarized relative efficiency of the model parameters in Table (1). Relative efficiencies $eff(\hat{\beta}_i^*, \hat{\beta}_i)$ were computed using (2.7) for $i = 1, 2, 3, 4, 5$. These outputs show, with all RSS sizes, that the spline models which used RSS sampling units are more efficient than spline models used SRS sampling units.

To show effectiveness of extension of the RSS method, samples were generated after ranking the predictor variable as mentioned in subsection (2.2) to estimate the spline model. Then the produced RSS samples were used to fit the spline model in (2.8) by using the estimated parameters in (2.9).

Table 1. Relative efficiencies of the RSS spline models comparing to the SRS spline models when ranking the response variable.

	$m = 2$		$m = 3$		$m = 4$	
	$r = 2$	$r = 3$	$r = 2$	$r = 3$	$r = 3$	$r = 6$
	$n = 4$	$n = 6$	$n = 6$	$n = 9$	$n = 12$	$n = 24$
$\hat{\beta}_0^*$	1.151	1.107	1.208	1.187	1.482	1.410
$\hat{\beta}_1^*$	1.149	1.093	1.208	1.188	1.469	1.396
$\hat{\beta}_{21}^*$	1.150	1.117	1.210	1.176	1.417	1.389
$\hat{\beta}_{22}^*$	1.152	1.153	1.194	1.179	1.431	1.390
$\hat{\beta}_{23}^*$	1.147	1.126	1.211	1.181	1.416	1.397

Table 2. Relative efficiencies of the RSS spline models comparing to the SRS spline models when ranking the predictor variable.

	$m = 2$		$m = 3$		$m = 4$	
	$r = 2$	$r = 3$	$r = 2$	$r = 3$	$r = 3$	$r = 6$
	$n = 4$	$n = 6$	$n = 6$	$n = 9$	$n = 12$	$n = 24$
$\hat{\beta}_0^*$	1.138	1.131	1.196	1.176	1.412	1.378
$\hat{\beta}_1^*$	1.137	1.130	1.201	1.180	1.396	1.329
$\hat{\beta}_{21}^*$	1.140	1.129	1.189	1.175	1.402	1.363
$\hat{\beta}_{22}^*$	1.132	1.137	1.193	1.173	1.378	1.337
$\hat{\beta}_{23}^*$	1.139	1.141	1.185	1.182	1.381	1.345

We compared the estimated RSS spline models, after ranking the predictor variable, with the above SRS spline models. The results for these simulation experiments are summarized in Table (2).

A general conclusion can be summarized from both Tables (1) and (2) that RSS method is more efficient than SRS when it used to fit spline models either the response variable or the predictor variable was ordered. Also, it can be realized that ranking on the response variable is more efficient than ranking on the predictor variable. Adding to this, it can be noted that the efficiency in all tables increasing with RSS size, m. This note cab be clarified as proved by [19] where the upper bound of the efficiency is $\frac{m+1}{2}$ which is increasing as m increasing.

4.2. Simulated penalized spline models. In the previous subsection (4.1), we fitted the spline regression models using RSS approach however, in this section, we illustrated our method to penalized spline models. By returning to the above simulation configurations in section (4), simulated RSS samples were generated after ranking the response variable. The same model and assumptions as in subsection (4.1) were considered where, however, estimates in (3.1) were used to produce smooth model fitting. Variance components were estimated using (3.2).

To make the comparison applicable, we generated SRS samples as discussed in section (4) and then the produced samples were used to fit penalized spline models. This smooth model fitting was achieved using (1.11). Results of these simulation experiments are summarized in Table (3) where relative efficiencies of model parameters are given. Relative efficiencies $eff(\hat{\beta}_i^{**}, \hat{\beta}_i)$ were computed using (3.3).

To present a further advantage of the RSS method, samples were simulated after ranking the predictor variable as discussed at the end of section (3). Then, the generated data were settled in the design matrices and fed into model (2.8). The penalized spline model,

Table 3. Relative efficiencies of the RSS penalized spline models comparing to the SRS spline models when ranking the response variable.

	$m=2$		$m=3$		$m=4$	
	$r=2$	$r=3$	$r=2$	$r=3$	$r=3$	$r=6$
	$n=4$	$n=6$	$n=6$	$n=9$	$n=12$	$n=24$
$\hat{\beta}_0^*$	1.145	1.113	1.313	1.263	1.401	1.399
$\hat{\beta}_1^*$	1.171	1.107	1.296	1.217	1.418	1.371
$\hat{\beta}_{21}^*$	1.190	1.116	1.308	1.219	1.398	1.400
$\hat{\beta}_{22}^*$	1.186	1.125	1.271	1.231	1.413	1.385
$\hat{\beta}_{23}^*$	1.192	1.131	1.259	1.205	1.418	1.412

Table 4. Relative efficiencies of the RSS penalized spline models comparing to the SRS spline models when ranking the predictor variable.

	$m=2$		$m=3$		$m=4$	
	$r=2$	$r=3$	$r=2$	$r=3$	$r=3$	$r=6$
	$n=4$	$n=6$	$n=6$	$n=9$	$n=12$	$n=24$
$\hat{\beta}_0^*$	1.168	1.137	1.203	1.135	1.481	1.412
$\hat{\beta}_1^*$	1.157	1.128	1.196	1.112	1.426	1.408
$\hat{\beta}_{21}^*$	1.183	1.117	1.173	1.117	1.398	1.391
$\hat{\beta}_{22}^*$	1.171	1.129	1.190	1.125	1.401	1.400
$\hat{\beta}_{23}^*$	1.176	1.133	1.199	1.131	1.417	1.397

with the same assumption, were estimated using (3.1) using the new design matrices. Covariance matrix was computed using (3.2). We compared these estimated penalized spline models (i.e models after ranking the predictor variable) with above SRS penalized spline models. Results of these simulation trails are summarized in Table (4).

A superior and general result, according to both tables (3) & (4), that is RSS method increased the competency of the selected sampling units to capture the underlying spline and penalized spline fittings. This is clearly shown by the computed efficiencies of the method. Also, it can be notified that ranking on the response variable is more efficient (in most cases) than ranking on the predictor variable. Additionally, it can be seen that the efficiency increased as the RSS size, m, increased which matches with the results in the previous section.

Finally, we run a small simulation study to compare penalized spline models under RSS and SRS methods in term of bias. The model in (1.1) was used to generate data with 3 knots. We propose the SRS size $n = 45$ and RSS size $m = 5$ which means that number of cycles is $r = 9$. To implement RSS methodology, a ranking mechanism was applied to order the predictor variable. After estimate penalized spline model under two sampling approaches, we summarized the results in table (5). As seen in this table, bias is higher in penalized spline SRS model estimators than penalized spline RSS model estimators.

5. Practical study

To illustrate the method that has been improved in this paper to real life applications, the environment study “Air Pollution”data set was used in this section. The data set shows daily readings of air quality components in New York city from May 1, 1973 to September 30, 1973. The data set have 154 observations with 6 variables. More details about this study can be found in [4]. Our investigations on this study is mainly to show

Table 5. Bias of the estimated parameters in penalized spline models under RSS and SRS methods. Ordering observations were achieved on the response variable.

	$m = 5$	$r = 9$	$n = 45$
Parameter	Exact value	Bias under RSS	Bias under SRS
β_0	1	0.0759	0.1102
β_1	2	0.0583	0.0901
β_{21}	1	0.0843	0.1273
β_{22}	1.5	0.0861	0.1165
β_{23}	2	0.0533	0.0971

efficiency of using RSS when fitting spline and penalized spline models. We studied two variables of this study which are Ozone (which represent the mean ozone parts per billion from 1300 to 1500 hours) as the response variable and Solar Radiation (which represent solar radiation in Langleys in the frequency band 4000-7700 Angstroms from 0800 to 1200 hours) as the predictor variable. The transformation Ozone$^{(1/3)}$ was considered in this paper.

Using set size $m = 3$, RSS samples were drawn from the Air Pollution data set with $r = 8$ cycles. In the first step, we ranked sample units with respect to the response variable to estimate the underlying relationship using spline and penalized spline estimates. Later on, we ranked sampling units with respect to the predictor variable to estimate appropriate spline and penalized spline models. And for the purpose of comparison, we selected a SRS of size $n = 24$ and then we estimated spline and penalized spline models as regular.

A note to mention here is that variables of the study were ranked based on exactly measured values. This method of ranking called "perfect ranking". We used this method because observations of this example were already measured. However, practically, the interesting attribute of RSS method is to use a relatively cheap ranking method to order subsamples then measure a few units of these subsamples which reduces sampling costs.

In all above models, we considered number of knots $q = 2$ and we chose optimal value of the smoothing parameter using GCV approach. Table (6) shows the relative efficiencies of the estimated spline and penalized spline models by using RSS sample units when ordering the response variable. While to enhance visual evaluation of our fitting models, we plot these estimated models as shown in Figure (1).

Table (7) presents the relative efficiencies of the estimated spline and penalized spline models when ordering the predictor variable. As seen in tables of this practical study, both spline and penalized spline models that were fitted using RSS method are more efficient than models that were fitted using SRS method. Adding to this, by comparing C_p plots of the estimated models is looking superior where the small C_p the better model.

6. Conclusion

The main conclusion drawn from this research is that the RSS methodology is more efficient when its sampling units were used to fit penalized spline models. The improvement of using our method is illustrated through parameters efficiencies which is clearly shown in all tables of the simulation study as well as outputs of the practical study. In spite of this paper presented better performance in estimating penalized spline models, we are not improving degree of smoothness of the targeted model. This is because we are keen about minimizing MSE of model coefficients.

Table 6. Relative efficiencies of the RSS models comparing to the SRS models of the practical example. Ordering observations were achieved on the response variable. $C_{p_{RSS}}$ is the Mallow constant computed for the RSS estimated models and $C_{p_{SRS}}$ is the Mallow constant computed for the SRS estimated models.

	$m = 3$	$r = 8$	$n = 24$	
spline model			penalized spline model	
$\hat{\beta}_0^*$	1.273		$\hat{\beta}_0^{**}$	1.289
$\hat{\beta}_1^*$	1.256		$\hat{\beta}_1^{**}$	1.293
$\hat{\beta}_{21}^*$	1.251		$\hat{\beta}_{21}^{**}$	1.272
$\hat{\beta}_{22}^*$	1.244		$\hat{\beta}_{22}^{**}$	1.286
$C_{p_{RSS}}$	15.23		$C_{p_{RSS}}$	15.01
$C_{p_{SRS}}$	16.83		$C_{p_{SRS}}$	16.03

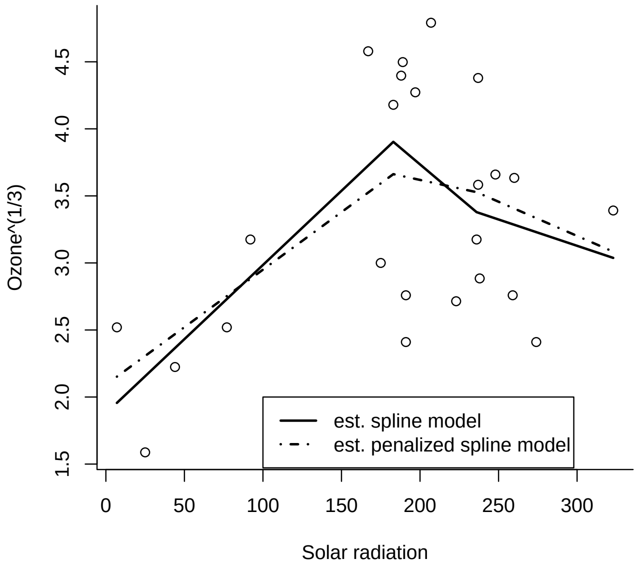

Figure 1. Estimated models for "Air Pollution"data set via RSS method when ranking the response variable.

In real data applications where sampling units are difficult or expensive to measure, another advantage can be appeared for this sampling method that is cost efficient attribute. This means, ranking a small number of units, before measuring a subset, can reduce time and sampling expenditure. Another practical point of view when ranking sampling units, analyst can consider a negligibly cost variable to achieve ranking, so he can select either the response or the predictor variable which is cheaper. Also, he can select the cheapest predictor variable to rank among all other expensive predictors.

Table 7. Relative efficiencies of the RSS models comparing to the SRS models of the practical example. Ordering observations were achieved on the predictor variable. $C_{p_{RSS}}$ is the Mallow constant computed for the RSS estimated models and $C_{p_{SRS}}$ is the Mallow constant computed for the SRS estimated models.

	$m = 3$	$r = 8$	$n = 24$	
spline model			penalized spline model	
$\hat{\beta}_0^*$	1.199		$\hat{\beta}_0^{**}$	1.205
$\hat{\beta}_1^*$	1.192		$\hat{\beta}_1^{**}$	1.211
$\hat{\beta}_{21}^*$	1.201		$\hat{\beta}_{21}^{**}$	1.196
$\hat{\beta}_{22}^*$	1.195		$\hat{\beta}_{22}^{**}$	1.213
$C_{p_{RSS}}$	15.23		$C_{p_{RSS}}$	15.10
$C_{p_{SRS}}$	16.93		$C_{p_{SRS}}$	16.33

This paper establishes a paradigm for future research under general linear model scenarios. Applying RSS procedure to other spline models like B-spline, natural cubic spline etc., to produce smooth regression models can be extended in the same simple manner. [18], Chapter 3, summarized these spline models which can prepare for general setup to use RSS method. Moreover, statistical inferences for our improved models can be investigated.

Also, and because the penalized spline RSS estimators are biased, further research can investigate bias reduction procedures. A possible scenario can improve methods discussed in [13]. The authors improved two weighting methods that can reduce estimators bias of the least square estimates. One method by using probability and the second is to smooth the weights. They also extended their method to pseudo maximum likelihood estimation for generalized linear models.

Acknowledgment

The author is grateful to the two referees for their crucial suggestions and comments which improves this paper.

References

[1] Al Kadiri, M. Carroll, R. and Wand, M. *Marginal Longitudinal Semiparametric Regression via Penalized Splines*, Statistics and Probability Letters **80**, 1242–1252, 2010.

[2] Chen, Z. *Ranked-set sampling with regression-type estimators*, Journal of Statistical Planning Inference **92**, 181-192, 2001.

[3] Claeskens, A. Krivobokova, T. and Opsomer, D. *Asymptotic Properties of Penalized Spline Estimators*, Biometrika **96**, 529–544, 2009.

[4] Cohen, Y. and Cohen, J. *Statistics and Data with R: An Applied Approach Through Examples* (New York: Wiley, 2008).

[5] Craven, P. and Wahba, G. *Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation*, Numerische Mathematik **31**, 377–403, 1979.

[6] DiMatteo, I. Genovese, C. and Kass, R. *Bayesian curve-fitting with free-knot splines*, Biometrika **88**, 1055–1072, 2001.

[7] Eilers, P. and Marx, B. *Splines, Knots, and Penalties*, Wiley Interdisciplinary Reviews: Computational Statistics **2**, 637–653, 2010.

[8] Eilers, P. and Marx, B. *Flexible smoothing with B-splines and penalties (with discussion)*, Statistical Sciences **11**, 89–121, 1996.

[9] Eubank, R. *The hat matrix for smoothing splines*, Statistics and Probability Letters **2**, 9–14, 1984.

[10] Eubank, R. *A simple smoothing spline*, American Statistician **48**, 103–106, 1994.

[11] Gu, C. *Multivariate spline regression. InM. G. Schimek (eds.)*, Smoothing and Regression: Approaches, Computation and Application (NewYork: Wiley, 2000), 329–355.

[12] Hall, P. and Opsomer, J. *Theory for Penalized Spline Regression*, Biometrika **92**, 105–118, 2005.

[13] Kim, K. and Skinner, J. *Weighting in Survey Analysis Under Informative Sampling*, Biometrika **100**, 385–398, 2013.

[14] Li, Y. and Ruppert, D. *On the Asymptotics of Penalized Splines*, Biometrika **95**, 415–436, 2008.

[15] McIntyre, G. *A method for Unbiased Selective Sampling, Using Ranked Sets*, Australian Journal of Agricultural Research **3**, 385–390, 1952.

[16] Phillips, R. *Iterated Feasible Generalized Least-Squares Estimation of Augmented Dynamic Panel Data*, Journal of Business and Economic Statistics **28**, 410–422, 2010.

[17] Ruppert, D. *Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation*, Journal of the American Statistical Association **92**, 1049–1062, 1997.

[18] Ruppert, D. Wand, M. and Carroll, R. *Semiparametric Regression.* New York: Cambridge university Press, 2003.

[19] Takahasi, K. and Wakimoto, K. *On Unbiased Estimates of the Population Mean Based on the Sample Stratified by Means of Ordering*, Annals of the Institute Statistical Mathematics **20**, 421–428, 1968.

[20] Wahba, G. *A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem*, Annals Statistics **13**, 1378–1402, 1985.

[21] Wand, M. *A comparison of regression spline smoothing procedures*, Computational Statistics **15**, 443–462, 2000.

[22] Wolfe, D. *Ranked Set Sampling: An Approach to More Efficient Data Collection*, Statistical Science **19**, 636–643, 2004.

[23] Wolfe, D. *Ranked set sampling: Its relevance and impact on statistical inference*, International Scholarly Research Network , Probability and Statistics, DOI: 10.5402/2012/568385, 2012.

[24] Yu, P. and Lam, K. *Regression Estimator in Ranked Set Sampling*, Biometrics **53** 1070–1080, 1997.

[25] Zhou, S. and Shen, X. *Spatially adaptive regression splines and accurate knot selection schemes*, Journal of the American Statistical Association **96**, 247–259, 2001.