Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/100124

Median and Extreme Ranked Set Sampling for penalized spline estimation

*Ali Algarni**

Statistics Department, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

Received: 3 Mar. 2015, Revised: 16 May 2015, Accepted: 1 Aug. 2015
Published online: 1 Jan. 2016

Abstract: This paper improves and demonstrates two approaches of Ranked Set Sampling (RSS) method for penalized spline models which are Median and Extreme RSS. These improved methods increase the efficiency of the estimated parameters in the targeted model with comparing to usual RSS and Simple Random Sampling (SRS). Moreover, in practical studies, our improved methods can reduce sampling expenses dramatically. The paper approaches are illustrated using a simulation study as well as a practical example.

Keywords: Ranked set sampling, Penalized spline model, Penalized least squares, efficiency.

1 Introduction

Linear regression models concern, with much attention, in procedures that can accommodate data in smooth fashion appropriately. In many practical situations, trends of the underlying model has curvilinear shapes which need improved fitting procedures. A well developed smoothing procedure is penalized spline models. This model type is an advance version of spline models where it has a new term, called penalty term, that can penalize trends with much rough to appear smooth.

Within last few decades, penalized spline approach developed in the literature significantly. It was originally introduced by [1]. [2] discussed characteristics for penalized least squares estimators. Improvements added to this method by [3] made it popular. Ruppert et.al. [4] summarized this approach magnificently and presented it in an easy way. The model approach was improved in the context of mixed models for random effect curves where the spline basis functions reduced number of knots to moderate level. This made computation of the penalized spline approach advantageous. Theoretical background with asymptotic properties for low rank approximation shown that estimators of penalized spline models are efficient as smoothing splines, [5] and [6]. Number of knots as well as smoothing parameter are two important elements that can shape the degree of smoothness. Ruppert [7] provided a new technique to select number of knots whilst, recently, Takuma Yoshida [8] provided a direct method to choose the smoothing parameter.

The common method to choose sampling units when fitting penalized spline models is the Simple Random Sampling (SRS) method. However, because it is more efficient, as well as its other procedures, RSS starts an increasingly influence in literature beside SRS for model fitting. This paper introduces Median Ranked Set Sampling (MRSS) as well as Extreme Ranked Set Sampling (ERSS) to fit these models as two important RSS procedures. Ranked sampling procedure was originally investigated by [9] to estimate mean population zone yields. He choose m simple random subsamples each of size m from the population. Then, and after ranked each subsamples separately, he selected the i^{th} smallest unit from the i^{th} subsample. Generally, this method can be repeated r times, where each repetition called a cycle, to generate the wanted RSS of size $n = rm$, where n is the SRS sample size. Quantification for the selected units is now available. This RSS procedure called balanced RSS.

Mathematical infrastructure of the RSS procedure was investigated by [10] and [11]. Significant articles have been published, later on, on the improvements of this sampling method. Importantly, Wolfe [12] reviewed the literature where he summarized RSS and its other procedures.

Most popular RSS procedures are MRSS and ERSS. The first procedure was proposed by [13] where the

* Corresponding author e-mail: ahalgarni@kau.edu.sa

median of each subsample is chosen rather than usual RSS units. While the ERSS was proposed by [14]. These two approaches verified their effectiveness in practical applications, see for example [15], [16] and [17].

The procedure of RSS for simple linear regression is summarized as in the following steps. Taking into account that ordering sample units is achieved according to the dependent variable.

1. From the targeted population, select m SRS subsamples each with size m. Assume these subsamples as $\{(x_1, y_1)_1, (x_2, y_2)_1, \ldots (x_m, y_m)_1\}$, $\{(x_1, y_1)_2, (x_2, y_2)_2, \ldots, (x_m, y_m)_2\}$, $\ldots$, $\{(x_1, y_1)_m, (x_2, y_2)_m, \ldots, (x_m, y_m)_m\}$.
2. Order, without any unit quantification, each subsample separately with respect to the dependent variable. The produced ranked-subsamples can be notated as $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_1, \ldots, (x_{[m]}, y_{(m)})_1\}$, $\{(x_{[1]}, y_{(1)})_2, (x_{[2]}, y_{(2)})_2, \ldots, (x_{[m]}, y_{(m)})_2\}$, $\ldots$, $\{(x_{[1]}, y_{(1)})_m, (x_{[2]}, y_{(2)})_m, \ldots, (x_{[m]}, y_{(m)})_m\}$. Ordering can be achieved visually by the analyst or by a skilled person or by any other relatively cheap method. In general, the pair $(x_{[i]}, y_{(i)})_j$ means, the i^{th} independent value correspond to the i^{th} minimum dependent value in the j^{th} subsample.
3. From the first ordered subsample select the first minimum pair with respect to the dependent variable, from the second ordered subsample select the second minimum pair with respect to the dependent variable. Continue by this way until choosing the maximum pair from the last subsample. The produced RSS set of size m is $\{(x_{[1]},y_{(1)})_1, (x_{[2]},y_{(2)})_2, \ldots (x_{[m]},y_{(m)})_m\}$. The hall process can be repeated r cycles to achieve equality with the SRS size n, *i.e* $n = rm$.

Actual quantification can now be done for these selected units and used to estimate parameters in the regression model. Consequently, the same RSS procedure can be applied after ranking the independent variable instead the dependent variable as alternative method.

Though this paper considers MRSS and ERSS, derivation of sampling units is similarly as in the above procedure. Exclusively, the third step need to be modified as follows. Firstly, to produce MRSS units with the case of odd sample size *i.e.m* is odd, the unit $y^{th}_{(\frac{m+1}{2})}$ is selected from each ordered subsample associated with the correspond independent variable. While, for the case of even sample size, *i.e.m* is even, two units need to be selected from successive ordered subsample which are $y^{th}_{(\frac{m}{2})}$ and $y^{th}_{(\frac{m}{2}+1)}$ associated with the correspond independent variables. This means, we select $(X_{[\frac{m}{2}]},y_{(\frac{m}{2})})$ from the first subsample then we select $(X_{[\frac{m}{2}+1]},y_{(\frac{m}{2})+1})$ from the second subsample and so on until reach the last subsample. Secondly, to produce ERSS units, the two extreme ordered units $y_{(1)}$ and $y_{(m)}$ are selected, with their correspondence independent variable values, from each ordered subsample.

A spline model, that use n SRS data points $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$, can be expressed as follows

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

where y is the response variable, x is the predictor, $\beta_0,\beta_1,\beta_{2j}$ are the model coefficients, e is the error term and K_j are the model knots and 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.

Settling the spline model in (1) in matrix form gives

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \quad (2)$$

where the design matrices of this model are

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

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

Model assumptions during this research propose $E(\varepsilon) = \mathbf{0}$, uncorrelated error terms with constant variance, say σ^2.

Using the least squares method produces a piecewise fitting that join at different value of knots however, in a more developed model fitting, these piecewise line segments that have much variability can be penalized to produce smooth fitting. These penalties introduce to quadratic form of the least squares method by adding a new term that can applied penalty to the rough estimates. So, the quadratic form with the penalty term can be written as

$$||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda^2\beta^{\mathbf{T}}\mathbf{D}\beta \quad (3)$$

where λ is a non-negative smoothing parameter, the matrix $\mathbf{D}$ is diagonal such that $\mathbf{D}$ = diag $\{\mathbf{0}_{2\times 2}, \mathbf{1}_{q\times q}\}$ and $||\mathbf{A}||$ equals $\sqrt{\mathbf{A}^T\mathbf{A}}$. Commonly, the last term in (3) is called the penalty term.

Minimizing (3) via penalized least squares method generate the exact solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \sigma^2\lambda^2\mathbf{D})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \quad (4)$$

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 $\mathbf{X}(\mathbf{X}^T\mathbf{X} + \sigma^2\lambda^2\mathbf{D})^{-1}\mathbf{X}^T$.

The covariance matrix of fitted coefficient $\hat{\beta}$ can be expressed as

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2\left[(\mathbf{X}^T\mathbf{X} + \sigma^2\lambda^2\mathbf{D})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \sigma^2\lambda^2\mathbf{D})^{-1}\right]. \quad (5)$$

In the penalized spline context, the constant variance σ^2 and the smoothing parameter λ need to be estimated. A common estimator for the variance σ^2, that proposed by [18], uses the Residual Sum of Squares (SSE) to produce an unbiased estimator such that

$$\hat{\sigma}^2 = \frac{||\mathbf{y}-\hat{\mathbf{y}}||^2}{n-tr(\mathbf{S}_\lambda)}. \quad (6)$$

where $tr(.)$ is the trace of a matrix.

Accordingly, the smoothing parameter is often chosen by minimizing the generalized cross-validation (GCV), [19], such that

$$\text{GCV}(\lambda) = \frac{||\mathbf{y}-\hat{\mathbf{y}}||^2}{[1-n^{-1}tr(\mathbf{S}_\lambda)]^2} = \sum_{i=1}^{n}\left[\frac{\{(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{y}\}_i}{1-n^{-1}tr(\mathbf{S}_\lambda)}\right]^2 \quad (7)$$

where $\mathbf{I}$ is the identity matrix. A single variance component in the $\text{Cov}(\hat{\beta})$ matrix in (5) can be estimated by

$$\widehat{Var}(\hat{\beta}_i^*) = \hat{\sigma}^{*2}\ [\text{the } i^{th} \text{ diagonal entry of } (\mathbf{X}_{[RSS]}^T\mathbf{X}_{[RSS]})^{-1}]. \quad (8)$$

This paper considers MRSS and ERSS to fit penalized spline models under simple linear model settings. Firstly, we propose ranking the dependent variable to achieve sampling unit selection then, we propose ranking the independent variable to achieve the same target. These selected units are used to estimate the penalized spline model after achieve measurements. These scenarios are explained in the next section.

2 Penalized spline fitting using MRSS and ERSS

Because the two method, MRSS and ERSS, are more efficient than usual RSS and SRS [17] and [20], illustrations of these two method to select sampling units and fit the penalized spline models are discussed in this section. The next subsection contains improvements of the penalized spline models when achieve ranking on the dependent variable using MRSS and ERSS while the subsection (2.2) improves these RSS illustrations when ranking the independent variable.

2.1 Penalized spline fitting when order the dependent variable

In this section, the penalized spline model uses the MRSS and ERSS after we rank the dependent variable. We consider model with general set up to cover both MRSS and ERSS parallel. Assume the produced MRSS sampling units that described in section (**??**), when m is odd, are give as: $\{(x_{[\frac{m}{2}]},y_{(\frac{m}{2})})_1,\ldots,(x_{[\frac{m}{2}]},y_{(\frac{m}{2})})_m\}$. We can regenerate this set r cycles to satisfy $n = rm$ where n is the SRS size. Accordingly, assume the produced ERSS sampling units that described in section (**??**) as: $\{(x_{[1]},y_{(1)})_1,(x_{[m]},y_{(m)})_1,\ldots,$ $(x_{[1]},y_{(1)})_m,(x_{[m]},y_{(m)})_m\}$. We can reproduce this set r cycles to satisfy $n = 2rm$.

Introducing the above generated units, under MRSS and ERSS, to the penalized spline model in (2) gives

$$\mathbf{y}_{(RSS)} = \mathbf{X}_{[RSS]}\beta^* + \varepsilon_{(RSS)} \quad (9)$$

where the design matrices under MRSS setting, with odd m, are

$$\mathbf{y}_{(RSS)} = \begin{bmatrix} y_{(\frac{m}{2})1_1} \\ \vdots \\ y_{(\frac{m}{2})m_1} \\ \vdots \\ y_{(\frac{m}{2})1_r} \\ \vdots \\ y_{(\frac{m}{2})m_r} \end{bmatrix};$$

$$\mathbf{X}_{[RSS]} = \begin{bmatrix} 1 & x_{[\frac{m}{2}]1_1} & (x_{[\frac{m}{2}]1_1}-K_1)_+ & \cdots & (x_{[\frac{m}{2}]1_1}-K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[\frac{m}{2}]m_1} & (x_{[\frac{m}{2}]m_1}-K_1)_+ & \cdots & (x_{[\frac{m}{2}]m_1}-K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[\frac{m}{2}]1_r} & (x_{[\frac{m}{2}]1_r}-K_1)_+ & \cdots & (x_{[\frac{m}{2}]1_r}-K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[\frac{m}{2}]m_r} & (x_{[\frac{m}{2}]m_r}-K_1)_+ & \cdots & (x_{[\frac{m}{2}]m_r}-K_q)_+ \end{bmatrix}$$

$\beta^* = \begin{bmatrix}\beta_0^* & \beta_1^* & \beta_{21}^* \cdots \beta_{2q}^*\end{bmatrix}^T$; $\varepsilon_{(RSS)} = \begin{bmatrix} e_{[\frac{m}{2}]1_1}^* & \cdots e_{[\frac{m}{2}]m_r}^*\end{bmatrix}^T$ where $y_{[\frac{m}{2}]i_j}$ is the median dependent variable in the i^{th} subsample from the j^{th} cycle, $x_{[\frac{m}{2}]i_j}$ is the correspondence independent variable value, β_0^* ; β_1^* ; $\beta_{21}^*;\cdots;\beta_{2q}^*$ are the model coefficients and $e_{[\frac{m}{2}]i_j}$ is the correspondence error term in the model.

By assuming the ERSS procedure, the design matrices in (9) become where

$$\mathbf{y}_{(RSS)} = \begin{bmatrix} y_{(1)1_1} \\ y_{(m)1_1} \\ \vdots \\ y_{(1)m_1} \\ y_{(m)m_1} \\ \vdots \\ y_{(1)1_r} \\ y_{(m)1_r} \\ \vdots \\ y_{(1)m_r} \\ y_{(m)m_r} \end{bmatrix};$$

$$\mathbf{X}_{[RSS]} = \begin{bmatrix} 1 & x_{[1]1_1} & (x_{[1]1_1}-K_1)_+ & \cdots & (x_{[1]1_1}-K_q)_+ \\ 1 & x_{[m]1_1} & (x_{[m]1_1}-K_1)_+ & \cdots & (x_{[m]1_1}-K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[1]m_1} & (x_{[1]m_1}-K_1)_+ & \cdots & (x_{[1]m_1}-K_q)_+ \\ 1 & x_{[m]m_1} & (x_{[m]m_1}-K_1)_+ & \cdots & (x_{[m]m_1}-K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[1]1_r} & (x_{[1]1_r}-K_1)_+ & \cdots & (x_{[1]1_r}-K_q)_+ \\ 1 & x_{[m]1_r} & (x_{[m]1_r}-K_1)_+ & \cdots & (x_{[m]1_r}-K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[1]m_r} & (x_{[1]m_r}-K_1)_+ & \cdots & (x_{[1]m_r}-K_q)_+ \\ 1 & x_{[m]m_r} & (x_{[m]m_r}-K_1)_+ & \cdots & (x_{[m]m_r}-K_q)_+ \end{bmatrix}$$

$\beta^* = \left[\beta_0^* \ \beta_1^* \ \beta_{21}^* \cdots \beta_{2q}^*\right]^T$; $\varepsilon_{(RSS)} = \left[e^*_{(1)1_1} \ \cdots e^*_{(m)m_r}\right]^T$.

The least squares criterion can fit the model in (9) by minimizing the penalized sum of square errors

$$(\mathbf{y_{(RSS)}} - \mathbf{X_{[RSS]}}\beta^*)^{\mathbf{T}}(\mathbf{y_{(RSS)}} - \mathbf{X_{[RSS]}}\beta^*) + \lambda^{*\mathbf{2}}\beta^{*\mathbf{T}}\mathbf{D}\beta^*$$

with respect to β^* and for some smoothing parameter λ^* where the matrix $\mathbf{D} = \text{diag}\{\mathbf{0}_{2\times 2}, \mathbf{1}_{q\times q}\}$. The penalized least squares method gives the following linear smoother $\hat{\mathbf{y}}_{(RSS)} = \mathbf{H}_{\lambda^*}\mathbf{y}_{(RSS)}$ where the smoothing matrix is $\mathbf{H}_{\lambda^*} = \mathbf{X}_{[RSS]}(\mathbf{X}^T_{[RSS]}\mathbf{X}_{[RSS]} + \sigma^{*2}\lambda^{*2}\mathbf{D})^{-1}\mathbf{X}^T_{[RSS]}$, here σ^{*2} is the constant variance for error term under model assumption.

Consequently, the estimated model coefficient matrix can be written in the form

$$\hat{\beta}^* = (\mathbf{X}^T_{[RSS]}\mathbf{X}_{[RSS]} + \sigma^{*2}\lambda^{*2}\mathbf{D})^{-1}\mathbf{X}^T_{[RSS]}\mathbf{y}_{(RSS)} \quad (10)$$

where the covariance matrix of the estimator can be written as

$$\text{Cov}(\hat{\beta}^*) = \sigma^{*2}\left[(\mathbf{X}^T_{[RSS]}\mathbf{X}_{[\mathbf{RSS}]} + \sigma^{*\mathbf{2}}\lambda^{*\mathbf{2}}\mathbf{D})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}_{[\mathbf{RSS}]}\mathbf{X}_{[\mathbf{RSS}]} \times (\mathbf{X}^{\mathbf{T}}_{[\mathbf{RSS}]}\mathbf{X} + \sigma^{*\mathbf{2}}\lambda^{*\mathbf{2}}\mathbf{D})^{-\mathbf{1}}\right], \quad (11)$$

To accomplish estimating model coefficient and its covariance above, both smoothing parameter λ^* and variance σ^{*2} need to be estimated. The smoothing parameter λ^* can be estimated using GCV concepts as follows

$$\text{GCV}(\lambda) = \sum_{i=1}^{n}\left[\frac{\{(\mathbf{I}-\mathbf{H}_{\lambda^*})\mathbf{y}_{(RSS)}\}_i}{1-n^{-1}tr(\mathbf{H}_{\lambda^*})}\right]^2 \quad (12)$$

while the variance σ^{*2} can be estimated using the following formula

$$\hat{\sigma}^{*2} = \frac{||\mathbf{y}_{(RSS)} - \hat{\mathbf{y}}_{(RSS)}||^2}{n - tr(\mathbf{H}_{\lambda^*})} \quad (13)$$

which is unbiased estimator for the variance of the error term in the penalized spline model in (9).

Depending on (11), the variance component for a single fitted model parameter $\hat{\beta}_i^*$ can be estimated as

$$\widehat{Var}(\hat{\beta}_i^*) = \hat{\sigma}^{*2} \ [\text{the } i^{th} \text{ diagonal entry of } (\mathbf{X}^T_{[RSS]}\mathbf{X}_{[RSS]})^{-1}]. \quad (14)$$

The main inference result of model fitting that is $\hat{\beta}^*$ is unbiased estimator and $\text{Cov}(\hat{\beta}^*) \geq \text{Cov}(\hat{\beta})$. Unbiasedness property can be proved straightforwardly while Covariance property is proved numerically as seen in Table (1). More details come at the simulation study in section (3).

2.2 Penalized spline fitting when order the independent variable

Parallel to the previous subsection, both MRSS and ERSS sampling units are generated after ranking the independent variable and then used to fit the penalized spline model. Consider the produced MRSS sampling units, when m is odd, are give as: $\{(x_{(\frac{m}{2})}, y_{[\frac{m}{2}]})_1, \ldots, (x_{(\frac{m}{2})}, y_{[\frac{m}{2}]})_m\}$. We can regenerate this set r cycles to satisfy $n = rm$ where n is the SRS size. Similarly, assume the produced ERSS sampling units are: $\{(x_{(1)}, y_{[1]})_1, (x_{(m)}, y_{[m]})_1, \ldots, (x_{(1)}, y_{[1]})_m, (x_{(m)}, y_{[m]})_m\}$. We can reproduce this set r cycles to satisfy $n = 2rm$. Settled these produced sampling units in design matrices gives the following penalized spline model

$$\mathbf{y}_{[RSS]} = \mathbf{X}_{(RSS)}\beta^* + \varepsilon_{[RSS]} \quad (15)$$

other model matrices can be defined similarly.

Building model inference is parallel to the previous subsection. Minimizing the penalized least squares

$$(\mathbf{y_{[RSS]}} - \mathbf{X_{(RSS)}}\beta^*)^{\mathbf{T}}(\mathbf{y_{[RSS]}} - \mathbf{X_{(RSS)}}\beta^*) + \lambda^{*\mathbf{2}}\beta^{*\mathbf{T}}\mathbf{D}\beta^*$$

with respect to β^* and for some smoothing parameter λ^* gives the estimated model coefficient which can be written in the form

$$\hat{\beta}^* = (\mathbf{X}^T_{(RSS)}\mathbf{X}_{(RSS)} + \sigma^{*2}\lambda^{*2}\mathbf{D})^{-1}\mathbf{X}^T_{(RSS)}\mathbf{y}_{[RSS]} \quad (16)$$

where the covariance matrix of this estimator can be written as

$$\mathrm{Cov}(\hat{\beta}^*) = \sigma^{*2}\Big[(\mathbf{X}^T_{(RSS)}\mathbf{X}_{(\mathbf{RSS})} + \sigma^{*\mathbf{2}}\lambda^{*\mathbf{2}}\mathbf{D})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}_{(\mathbf{RSS})}\mathbf{X}_{(\mathbf{RSS})} \times (\mathbf{X}^{\mathbf{T}}_{(\mathbf{RSS})}\mathbf{X} + \sigma^{*\mathbf{2}}\lambda^{*\mathbf{2}}\mathbf{D})^{-\mathbf{1}}\Big], \tag{17}$$

The smoothing parameter λ^* can be estimated using GCV concepts as follows

$$\mathrm{GCV}(\lambda) = \sum_{i=1}^{n}\left[\frac{\{(\mathbf{I}-\mathbf{H}_{\lambda^*})\mathbf{y}_{[RSS]}\}i}{1-n^{-1}tr(\mathbf{H}_{\lambda^*})}\right]^2 \tag{18}$$

while the variance σ^{*2} can be estimated using the following formula

$$\hat{\sigma}^{*2} = \frac{||\mathbf{y}_{[RSS]} - \hat{\mathbf{y}}_{[RSS]}||^2}{n - tr(\mathbf{H}_{\lambda^*})} \tag{19}$$

which is unbiased estimator for the variance of the error term in the penalized spline model in (15).

Depending on (17), the variance component for a single fitted model parameter $\hat{\beta}^*_i$ can be estimated as

$$\widehat{Var}(\hat{\beta}^*_i) = \hat{\sigma}^{*2}\ [\text{the } i^{th} \text{ diagonal entry of } (\mathbf{X}^T_{(RSS)}\mathbf{X}_{(RSS)})^{-1}]. \tag{20}$$

The main inference result of model fitting that is $\hat{\beta}^*$ is unbiased estimator and $\mathrm{Cov}(\hat{\beta}^*) \geq \mathrm{Cov}(\hat{\beta})$. Unbiasedness property can be proved straightforwardly while Covariance property is proved numerically as seen in Table (2).

3 Simulation study

A simulation study was illustrated to clarify the practical improvements of estimating the penalized spline models when using MRSS and ERSS procedures. Data were generated from the following smooth function: $y_i = f(x_i) + e_i$, such that $f(x) = 1 + \frac{1}{2}\Phi(\frac{x-36}{5})$ and $x \sim Uniform(0,1)$. The error terms e_i were assumed uncorrelated with 0 mean and constant variance that equals to 0.122. Here, Φ is the standard normal density function. We proposed MRSS and ERSS samples with sizes $m = 3$ and 5 units with specific number of cycles r to perform the relation $n = rm$ (the case when using MRSS) or $n = 2rm$ (the case when using ERSS), where n is the SRS size. Without any loss of generality in our method, all models were proposed in this section used three knots, i.e. $q = 3$. This is due to handle presentation of the simulation results in comfortable tables.

After selecting MRSS and ERSS sampling units from the generated samples above, the penalized spline model, established in (9) and (15), were fitted. Ranking units were attained either by ranking the dependent variable or the independent variable.

For the sake of comparison, the smooth function that considered earlier in this section was used to generate SRS samples of size $n = 9, 12, 20$ and 25. The produced SRS samples were used to estimate penalized spline models with 3 knots. This small number of knots is to allow comparison with the simulated RSS procedures that have the same number of knots. Last point to mention that all settings in this simulation study were achieved with 10000 replicates.

In what follows, ordering sample units were done according to dependent and independent variables. Firstly, simulation and model fitting were done when ordering the dependent variable. The penalized spline model coefficients in (9) were estimated using (10) once by implementing MRSS and later on by implementing ERSS. Variance of these estimated coefficients was computed using (14). Simulation results are summarized in Table (1).

Comparison of variances of the estimated model coefficients were attained by using "Relative Efficiency" concept. This concept, which mainly depends on comparing variance of the estimated model parameter under RSS procedures with variance under SRS, can be defined as

$$eff(\hat{\beta}^*_i, \hat{\beta}_i) = \frac{\mathrm{Var}(\hat{\beta}_i)}{\mathrm{Var}(\hat{\beta}^*_i)} \tag{21}$$

where $\mathrm{Var}(\hat{\beta}^*_i)$ and $\mathrm{Var}(\hat{\beta}_i)$ are defined in (14) and (8) respectively. Table(1) shows that both MRSS and ERSS are more efficient than RSS and SRS in estimating penalized spline models when order the dependent variable.

Secondly, we conducted simulation study and model fitting when we ordered the independent variable of the penalized spline model (15). Model coefficients were estimated using (16) while both MRSS and ERSS procedures were used to handle the model design matrices. Variance components were estimated using (20). We summarized results in Table(2) by computing relative efficiencies. This table shows that MRSS and ERSS are more efficient than both RSS and SRS when fitting penalized spline models after rank the independent variable.

4 Real data example

In this example, we investigated performance of our new sampling procedures, that are MRSS and ERSS, when estimating penalized spline regression models.

The real life application called "Air Pollution". The data set presents daily measurements 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 data set can be found in [21]. We investigated in this study the efficiency

Table 1: Relative efficiencies of the estimated penalized spline model coefficients using MRSS and ERSS w.r.t. SRS in the simulated data study. We assumed ranking of the dependent variable.

	MRSS		ERSS		RSS	
	m=3 r=3	m=5 r=5	m=3 r=2	m=5 r=2	m=3 r=3	m=5 r=5
	n = 9	n = 25	n = 12	n = 20	n = 9	n = 25
$\hat{\beta}_0^*$	1.728	1.909	1.614	1.813	1.231	1.401
$\hat{\beta}_1^*$	1.735	1.917	1.635	1.837	1.242	1.398
$\hat{\beta}_{21}^*$	1.713	1.915	1.607	1.794	1.250	1.414
$\hat{\beta}_{22}^*$	1.714	1.913	1.628	1.805	1.219	1.411
$\hat{\beta}_{23}^*$	1.720	1.908	1.623	1.815	1.238	1.437

Table 2: Relative efficiencies of the estimated penalized spline model coefficients using MRSS and ERSS w.r.t. SRS in the simulated data study. We assumed ranking of the independent variable.

	MRSS		ERSS		RSS	
	m=3 r=3	m=5 r=5	m=3 r=2	m=5 r=2	m=3 r=3	m=5 r=5
	n = 9	n = 25	n = 12	n = 20	n = 9	n = 25
$\hat{\beta}_0^*$	1.765	1.903	1.637	1.863	1.135	1.492
$\hat{\beta}_1^*$	1.739	1.899	1.658	1.857	1.131	1.484
$\hat{\beta}_{21}^*$	1.727	1.911	1.631	1.871	1.129	1.485
$\hat{\beta}_{22}^*$	1.740	1.912	1.622	1.839	1.133	1.490
$\hat{\beta}_{23}^*$	1.752	1.910	1.670	1.841	1.128	1.486

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

In both MRSS and ERSS samples were chosen from the Air Pollution data set with size $m = 3$ where number of cycles were proposed $r = 8$ and $r = 4$, respectively. Firstly, we achieved ranking units with respect to the dependent variable. After this achievement, we ranked sampling units with respect to the independent variable. The produced ranked units were used to estimate the underlying model via penalized spline fitting that are in (9) and (15). All produced models above were compared to the estimated penalized spline models when using RSS and SRS sample units in (2). The sample size in the RSS method was considered $m = 3$ with $r = 8$ cycles while the SRS size was considered $n = 24$. In all models, we proposed number of knots $q = 3$ and we selected the optimal smoothing parameter using GCV method.

Results of these estimated models are summarized in Table(3) and Table(4). As seen in both tables, MRSS and ERSS are more efficient than RSS and SRS. Moreover, MRSS seems to be the most efficient at all.

The two illustrations of the RSS method, that are MRSS and ERSS, were developed in this research for penalized spline models. As seen in table (1) and (2) of the simulation study as well as seen in table (3) and (4) of the practical study, both MRSS and ERSS are more efficient in estimating the linear penalized models than RSS and SRS. Also, it seems that MRSS is better than ERSS.

This research used balanced RSS where the i^{th} ranked sampling unit selected from i^{th} subsample. We can improve our method for other methods of unit selection that called allocation methods. Also, in this research, we considered perfect (actual) ranking method which can be developed to other ranking procedures to add more challenge to the method.

Acknowledgments

This work was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant No. (130-532-D 1435). The authors, therefore, acknowledge with thanks DSR technical and financial support.

Table 3: Relative efficiencies of the estimated penalized spline model coefficients using MRSS and ERSS in the Air Pollution data set. We assume ranking of the dependent variable.

	efficiency w.r.t RSS, m = 3 r = 8			efficiency w.r.t SRS, n = 24		
	MRSS	ERSS		MRSS	ERSS	RSS
	m = 3 r = 8	m =3 r = 4		m = 3 r = 8	m = 3 r = 4	m = 3 r = 8
$\hat{\beta}_0^*$	1.541	1.508		1.984	1.941	1.287
$\hat{\beta}_1^*$	1.522	1.501		1.963	1.935	1.289
$\hat{\beta}_{21}^*$	1.517	1.517		1.951	1.921	1.286
$\hat{\beta}_{22}^*$	1.565	1.546		1.970	1.947	1.259
$\hat{\beta}_{23}^*$	1.491	1.503		1.899	1.915	1.274

Table 4: Relative efficiencies of the penalized spline model coefficients using MRSS and ERSS in the Air Pollution data set. We assume ranking of the independent variable.

	efficiency w.r.t RSS, m = 3 r =8			efficiency w.r.t SRS, n =24		
	MRSS	ERSS		MRSS	ERSS	RSS
	m = 3 r = 8	m = 3 r = 4		m = 3 r = 8	m = 3 r = 4	m = 3 r = 8
$\hat{\beta}_0^*$	1.644	1.569		1.989	1.899	1.210
$\hat{\beta}_1^*$	1.557	1.495		1.901	1.825	1.221
$\hat{\beta}_{21}^*$	1.642	1.565		1.893	1.890	1.208
$\hat{\beta}_{22}^*$	1.599	1.539		1.947	1.874	1.217
$\hat{\beta}_{23}^*$	1.582	1.509		1.992	1.901	1.259

References

[1] F. O'Sullivan, A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, (1986) **1**, 502—518.

[2] R. Eubank and R. Gunst, Diagnostics for penalized least-squares estimators, *Statistics and Probability Letters*, **4** (1986), 265 — 72.

[3] P. Eilers , B. Marx, Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder), *Statistical Science*, **11** (1996), 89 – 121.

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

[5] Y. Li, D. Ruppert, On the asymptotics of penalized splines, *Biometrika* (2008) **95**, 415—36.

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

[7] D. Ruppert, Selecting the number of knots for penalized splines, *Journal of Computational and Graphical Statistics*, (2002) **11**, 735–757.

[8] Takuma Yoshida, Direct Determination of Smoothing Parameter for Penalized Spline Regression,*Journal of Probability and Statistics* (2014) **2014**. http://dx.doi.org/10.1155/2014/203469.

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

[10] K. Takahasi and K.Wakimoto, On unbiased estimates of the population mean based on the sample stratified by means of ordering, *Annals of the Institute of Statistical Mathematics*, **20** (1968), 1 — 31.

[11] T. R. Dell and J. L. Clutter, Ranked set sampling theory with order statistics background, *Biometrics*, **28** (1972), 545 — 555.

[12] D. Wolfe, Ranked set sampling: Its relevance and impact on statistical inference, *International Scholarly Research Network , Probability and Statistics*, **10** (2012) 1070 – 1080. http://dx.doi.org/10.5402/2012/568385

[13] H. Muttlak, Median ranked set sampling, *Journal of Applied Statistical Science*, **6** (1997), 245 — 255.

[14] H. M. Samawi, M. S. Ahmed, andW. Abu-Dayyeh, Estimating the population mean using extreme ranked set sampling, *Biometrical Journal*, **38** (1996), 577 — 586.

[15] H. A. Muttlak, Median ranked set sampling with concomitant variables and a comparison with ranked set sampling and regression estimators, *Environmetrics*, **9** (1998), 255 — 267.

[16] Y. A. Ozdemir and F. Gokpinar, "A new formula for inclusion probabilities in median-ranked set sampling, *Communications in Statistics*, **37** (2008), 2022 — 2033.

[17] A. I. Al-Omari, Ratio estimation of the population mean using auxiliary information in simple random sampling and median ranked set sampling, *Statistics & Probability Letters*, **82** (2012), 1883 — 1890.

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

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

[20] M. T. Alodat, M. Y. Al-Rawwash, I. M. Nawajah, Inference about the regression parameters using median-ranked set sampling,*Communications in Statistics*, **39**,(2010), 2604—2616.
[21] Y. Cohen, J. Cohen, *Statistics and Data with R: An Applied Approach Through Examples*, Wiley, New York, 2008.

Ali Algarni has obtained his PhD degree at 2013 from School of Mathematics and Applied Statistics, Wollongong University, Australia. He is a Investigator and Data Analyst at center of survey methodology and data analysis,Wollongong. Australia. He is a member, Saudi Association of Mathematics and Applied Statistics. He is presently employed as assistant professor in Statistics Department, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia.