Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/050102

Spline Model Estimation Using Double Ranked Set Sampling

*M. A. Al Kadiri**

Department of Statistics, Yarmouk University, Irbid, Jordan

Received: 13 Sep. 2015, Revised: 30 Oct. 2015, Accepted: 2 Nov. 2015
Published online: 1 Mar. 2016

Abstract: This paper introduces and investigates the Double Ranked Set Sampling (DRSS) to estimate spline models. Properties of the new estimated parameters are examined and compared to both Simple Random Sampling (SRS) and regular Ranked Set Sampling (RSS). Moreover, this paper demonstrates efficiency of the new sampling method using artificial and real data examples.

Keywords: Generalized least square method,Ranked Set Sampling, Double Ranked Set Sampling, Spline regression models, Efficiency.

1 Introduction

Recent linear regression researches concern, with much attention, on fitting approaches that can accommodate data sets adequately. 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.

Crucially, spline models play a central role in regression because their computational properties and ability to gain appropriate fit, [7]. At early stages of investigation, researchers developed spline models to scatter plot smoothing (e.g. [8]). Later on, they treated spline model as polynomial which can be improved in a frame of knot selection (e.g. [20]) and basis functions (e.g. [9]). Introducing spline models to multivariate regression (e.g. [10]), nonparametric regression (e.g.[8]) and Bayesian models (e.g. [6]) took a wide range of interest in the literature. [16] made a considerable comparison between spline models.

Availability of various sampling methods is indeed a major challenge for researchers. This is because they need to investigate appropriateness of these methods to gain better model estimates. A classical sampling method to fit spline models considers Simple Random Sampling (SRS). However, since it is practically more efficient, Ranked Set Sampling (RSS) has an increasing attractiveness when estimating regression models, [18]. This method can minimize sampling costs and furthermore, it can improve efficiency of the estimated parameters in the underlying model, [17]. For these reasons, this research investigates DRSS technique, as an improved method of RSS, for spline models and compare it with RSS and SRS techniques.

McIntyer [12], who firstly introduced RSS method, used it to estimate the population mean of forage yields. [15] provided the mathematical theory of this method. They proved that the estimated mean using RSS method is an unbiased estimator to the population mean and it has less variance than usual SRS estimated mean. The recent monograph by [18] summarized most of research linked to RSS method until that date. He presented the dramatic increase of using RSS method in different statistical fields as well as its practical efficiency in various research fields. Importantly, [2] introduced a new RSS procedure called Double RSS (DRSS). This procedure depends mainly on repeating the usual RSS twice where the produced sampling units can increase the efficiency of the estimated mean dramatically.

The RSS procedure was introduced to regression by [19] and [4]. In a recent paper, [1] extended the RSS method to estimate spline regression models. In order to compare the fitted models, [1] found the new estimated parameters using RSS have less variance than the estimated parameters using SRS. To enhance improvement, this paper develops spline

* Corresponding author e-mail: alkadiri-m@yu.edu.jo

model fitting by using DRSS as an alternative procedure to RSS and SRS. Mainly, this research proposes that the new sampling procedure can improve model fitting by reduce variance of the estimated parameters.

1.1 Spline models with simple random sampling

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

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

where y_i is the response variable, x_i 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; $j = 1,...,q$, where q is number of knots. These knots are usually selected from the dense set of the predictor variable. The mathematical expression $(a)_+$ means the non-negative part of a; i.e. $max(0,a)$. Here we call the term $(x_i - K)_+$ by a linear spline basis function.

Settling the spline model (1) in matrix form gives

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{2}$$

where the design matrices of this model are

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix};\ \mathbf{X} = \begin{bmatrix} 1\ x_1\ (x_1-K_1)_+ \cdots (x_1-K_q)_+ \\ 1\ x_2\ (x_2-K_1)_+ \cdots (x_2-K_q)_+ \\ \vdots\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ 1\ x_n\ (x_n-K_1)_+ \cdots (x_n-K_q)_+ \end{bmatrix};\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{21} \\ \vdots \\ \beta_{2q} \end{bmatrix};\ \varepsilon = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

General model assumptions over the random error term ε assume that $\mathrm{E}(\varepsilon) = \mathbf{0}$ and $\mathrm{Cov}(\varepsilon) = \Sigma$. During this research we keep the random error term independent of the predictor variable.

Applying the generalized least square method on the above spline model produces the fit

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \tag{3}$$

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

$$||\mathbf{y} - \mathbf{X}\beta||^2 = (\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\beta) \tag{4}$$

with closed solution $\hat{\beta} = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y}$. where the produced least square estimate is unbiased ; i.e $\mathrm{E}(\hat{\beta}) = \beta$, with covariance $\mathrm{Cov}(\hat{\beta}) = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-\mathbf{1}}$. Simply, one can realized that variance of the model coefficients $\hat{\beta}_i$ take the form $\mathrm{Var}(\hat{\beta}_i) = [\text{the } i^{th} \text{ diagonal element of } (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}]$.

Alternative simple model assumption consider uncorrelated errors with constant variance such that $\mathrm{Cov}(\varepsilon) = \sigma^2\mathbf{I}$; where $\mathbf{I}$ is the identity matrix, which gives the least square estimate

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{5}$$

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

$$\mathrm{Var}(\hat{\beta}_i) = \sigma^2[\text{the } i^{th} \text{ diagonal element of } (\mathbf{X}^T\mathbf{X})^{-1}]. \tag{6}$$

Model fitting needs to estimate σ^2. Implementing Sum Square Errors (SSE) is a common approach to produce an unbiased estimator for σ^2 as

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{n-p} = \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{n-p} \tag{7}$$

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

The previous model fitting approach considered SRS method when select sampling units. However in this paper, we investigate the DRSS method to estimate spline models and compare them with the regular RSS and SRS. The RSS method, which selects sampling units after spread them in a proceeding manner, verified its quality in many practical modeling situations, [18]. In what follows, we describe the regular RSS method for general statistics and for simple linear regression in specific. Then, the DRSS procedure is also described.

1.2 Ranked set sampling procedure

When the RSS procedure was firstly established, [12] divided sample units to distinguished subsamples then each subsample had been ordered in a proceeding manner separately. The following Particular steps can summarize the procedure.

–Step 1: Randomly select m^2 units from the target population.
–Step 2: Allocate the m^2 selected units as randomly as possible into m sets, each of size m.
–Step 3:Rank the units within each set and select the i^{th} ranked unit from the i^{th} sample.
–Step 4:The whole process can be repeated r cycles if needed to increase the sample size.

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, RSS method is practically effective when sampling units are expensive or hard to measure, however rank few units, without real quantification, is relatively cheaper. Attain order of sampling units can be made by an expert or an analyst judgment visually or by any other relatively cheap method.

To generate DRSS units, procedure can be described as follows: Identify m^3 units from the target population and divide these units randomly into m sets each of size m^2. The procedure of ranked set sampling is applied on these sets to obtain m ranked set sampling each of size m, and again apply the ranked set sampling procedure on the m ranked set sampling sets obtained in the first stage to obtain a DRSS of size m. The yielded DRSS set can be presented as $\{x_{(11)1}, x_{(22)2}, ..., x_{(mm)m}\}$.

Introducing RSS methods to regression can be extended similarly as above. The only note to mention is that RSS units can be yielded either by ordering the response variable or the predictor variable. In the following regression example, we consider the case of ordering the response variable y to generate RSS units. 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 unit linked with the correspondence predictor; 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.

This paper implements the DRSS procedure to fit spline models where either ranking the response variable or the predictor variable can be achieved. Demonstration of the RSS to DRSS can be achieved straightforwardly. Simply we need to regenerate m different RSS samples each of size m as above to produce the following DRSS units: $\{(x_{[11]}, y_{(11)})_1, (x_{[22]}, y_{(22)})_2, ..., (x_{[mm]}, y_{(mm)})_m\}_1$,
$\{(x_{[11]}, y_{(11)})_1, (x_{[22]}, y_{(22)})_2, ..., (x_{[mm]}, y_{(mm)})_m\}_2$,...,$\{(x_{[11]}, y_{(11)})_1, (x_{[22]}, y_{(22)})_2, ..., (x_{[mm]}, y_{(mm)})_m\}_r$ where r is number of cycles that DRSS need to be repeated to achieve equality $n = rm$.

The next two sections define the RSS and DRSS procedures, that have been described above, for spline models. Model's parameters are estimated using the new DRSS method and the efficiency of these parameters are compared with the estimated parameters concluded by RSS and SRS.

2 Spline model estimation using RSS techniques

Demonstrations of RSS procedures to select sample units and fit spline models are achieved in this section. Firstly, in subsection (2.1), the RSS and DRSS 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. At the end of this section, we investigate the new sampling schemes. We realized that RSS and DRSS, achieved better performance than SRS scheme when fitting spline models. Moreover, the DRSS scheme has the best performance.

2.1 Spline models with a ranked response

Mainly in this subsection, spline model fitting is achieved using RSS and DRSS units after order the response variable. We illustrate the method described at the end of the introduction to produce the DRSS units. Now, the produced sample is available to estimate the proposed spline model.

The linear spline model, after implement DRSS units, can be written similar to model (1) as follows

$$y_{(ii)j} = \beta_0^* + \beta_1^* x_{[ii]j} + \sum_{l=1}^{q} \beta_{2l}^* (x_{[ii]j} - K_l)_+ + e_{(ii)j}^*;\ i = 1,\ldots,m;\ j = 1,\ldots,r. \tag{8}$$

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

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

where $\mathbf{y}_{(DRSS)} = \begin{bmatrix} y_{(11)1} \\ \vdots \\ y_{(mm)1} \\ \vdots \\ y_{(11)r} \\ \vdots \\ y_{(mm)r} \end{bmatrix}$; $\mathbf{X}_{[DRSS]} = \begin{bmatrix} 1 & x_{[11]1} & (x_{[11]1}-K_1)_+ & \cdots & (x_{[11]1}-K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[mm]1} & (x_{[mm]1}-K_1)_+ & \cdots & (x_{[mm]1}-K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[11]r} & (x_{[11]r}-K_1)_+ & \cdots & (x_{[11]r}-K_q)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{[mm]r} & (x_{[mm]r}-K_1)_+ & \cdots & (x_{[mm]r}-K_q)_+ \end{bmatrix}$

$\beta^* = \begin{bmatrix} \beta_0^* & \beta_1^* & \beta_{21}^* \cdots \beta_{2q}^* \end{bmatrix}^T$; $\varepsilon_{(DRSS)} = \begin{bmatrix} e_{(11)_1}^* & e_{(22)_1}^* \cdots e_{(mm)r}^* \end{bmatrix}^T$.

Similarly, we can produce the following RSS units $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_2, ..., (x_{[m]}, y_{(m)})_m\}_1$, $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_2, ..., (x_{[m]}, y_{(m)})_m\}_2$,..., $\{(x_{[1]}, y_{(1)})_1, (x_{[2]}, y_{(2)})_2, ..., (x_{[m]}, y_{(m)})_m\}_r$ and introduce them to the model in (9) however the new model matrices contain RSS units.

Model assumptions assume uncorrelated errors with non-constant variance and zero mean which means

$$E(\varepsilon_{(DRSS)}) = \mathbf{0} \text{ and } \mathrm{Cov}(\varepsilon_{(DRSS)}) = \mathrm{diag}\{\sigma_{(1)}^{*2}, ..., \sigma_{(m)}^{*2}, ..., \sigma_{(1)}^{*2}, ..., \sigma_{(m)}^{*2}\}_{mr\times mr} \equiv \Sigma^*. \tag{10}$$

Keeping the non-constant variance assumption in (10) 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), [13]. Computer statistical softwares are rich with packages that can 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

$$E(\varepsilon_{(DRSS)}) = \mathbf{0} \text{ and } \mathrm{Cov}(\varepsilon_{(DRSS)}) \equiv \Sigma^* = \sigma^{*2}\mathbf{I}. \tag{11}$$

Under these assumptions, using the generalized least square method to minimize $||\mathbf{y}_{(DRSS)} - \mathbf{X}_{[DRSS]}\beta^*||^2$ produces

$$\hat{\beta}^* = (\mathbf{X}_{[DRSS]}^T \Sigma^{*-1} \mathbf{X}_{[DRSS]})^{-1} \mathbf{X}_{[DRSS]}^T \Sigma^{*-1} \mathbf{y}_{(DRSS)} \tag{12}$$

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

$$\widehat{Var}(\hat{\beta}_i^*) = \ [\text{the } i^{th} \text{ diagonal entry of } (\mathbf{X}_{[DRSS]}^T \widehat{\Sigma}^{*-1} \mathbf{X}_{[DRSS]})^{-1}] \tag{13}$$

where $\widehat{\Sigma}^*$ is the estimated covariance matrix.

Considerably, the produced estimator $\hat{\beta}^*$, either by using RSS or DRSS procedures, is an unbiased estimator for the model parameter β and their covariances satisfies $\mathrm{Cov}(\hat{\beta}^*) \leq \mathrm{Cov}(\hat{\beta})$ where $\hat{\beta}$ is the least square estimate of β when using SRS as defined in (5). Proof of the first property is given next while proof of the second property is attained numerically as seen in the simulation study Table (1).

To prove unbiasedness of the $\hat{\beta}^*$, we need to prove that $E(\hat{\beta}^*) = \beta$. Note that

$$\begin{aligned} E(\hat{\beta}^*) &= E((\mathbf{X}_{[DRSS]}^T \Sigma^{*-1} \mathbf{X}_{[DRSS]})^{-1} \mathbf{X}_{[DRSS]}^T \Sigma^{*-1} \mathbf{y}_{(DRSS)}) \\ &= (\mathbf{X}_{[DRSS]}^T \Sigma^{*-1} \mathbf{X}_{[DRSS]})^{-1} \mathbf{X}_{[DRSS]}^T \Sigma^{*-1} E(\mathbf{y}_{(DRSS)}) \\ &= (\mathbf{X}_{[DRSS]}^T \Sigma^{*-1} \mathbf{X}_{[DRSS]})^{-1} \mathbf{X}_{[DRSS]}^T \Sigma^{*-1} \mathbf{X}_{[DRSS]} \beta \\ &= \beta \end{aligned} \tag{14}$$

To demonstrate improvement of our new procedures, we define the relative efficiency concept of $\hat{\beta}_i^*$ with respect to $\hat{\beta}_i$ as follows

$$eff(\hat{\beta}_i^*, \hat{\beta}_i) = \frac{\widehat{\text{Var}}(\hat{\beta}_i)}{\widehat{\text{Var}}(\hat{\beta}_i^*)} \tag{15}$$

which indicates how much one of the estimators is better than the other one.

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

2.2 Spline models with ranked predictor variable

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

When the ranking is performed on the predictor variable, the suggested spline model using DRSS is given by

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

where $x_{(ii)j}$ is i^{th} smallest unit of the predictor variable from the i^{th} DRSS subsample in the j^{th} cycle, $y_{[ii]j}$ is the response variable that associate with $x_{(ii)j}$; β_0^*, β_1^* and β_{2l}^* are the model parameters,$K_1, ..., K_q$ are the model knots and $e_{[ii]j}^*$ is the random error term.

Settle the above model in matrix form produces

$$\mathbf{y}_{[DRSS]} = \mathbf{X}_{(DRSS)} \beta^* + \varepsilon_{(DRSS)}. \tag{16}$$

Matrices of the above model can be defined similarly as in model (9) with the same model assumptions.

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

$$\hat{\beta}^* = (\mathbf{X}_{(DRSS)}^T \Sigma^{*-1} \mathbf{X}_{(DRSS)})^{-1} \mathbf{X}_{(DRSS)}^T \Sigma^{*-1} \mathbf{y}_{[DRSS]}. \tag{17}$$

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

$$\text{Cov}(\hat{\beta}^*) = (\mathbf{X}_{(DRSS)}^T \Sigma^{*-1} \mathbf{X}_{(DRSS)})^{-1}. \tag{18}$$

Importantly, the produced estimator $\hat{\beta}^*$ in (17) is unbiased estimator for the model parameter β and its covariance satisfies $\text{Cov}(\hat{\beta}^*) \leq \text{Cov}(\hat{\beta})$ where $\hat{\beta}$ is the least square estimate of β when using SRS as defined in (5).

The second property with support of Table (2), can show that the fitted spline model using both RSS and DRSS are more efficient than ones that fitted using SRS where, $\text{eff}(\hat{\beta}^*, \hat{\beta}) \geq \mathbf{1}$. Magnificently, DRSS gained the best efficiency at all.

3 Simulation study

To illustrate the practical performance of estimating spline models using RSS and DRSS schemes, computer artificial studies were conducted with the following general set up. Data sets were generated from the curvilinear 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

Table 1: Relative efficiencies of the simulation study when using DRSS and RSS spline models comparing to the SRS spline models. Rank the response variable was achieved.

		$m=2$		$m=3$		$m=4$	
		$r=2$	$r=3$	$r=2$	$r=3$	$r=3$	$r=6$
		$n=4$	$n=6$	$n=6$	$n=9$	$n=12$	$n=24$
RSS							
	$\hat{\beta}_0^*$	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
DRSS							
	$\hat{\beta}_0^*$	1.727	1.701	1.892	1.821	2.105	2.010
	$\hat{\beta}_1^*$	1.718	1.709	1.851	1.818	2.137	1.976
	$\hat{\beta}_{21}^*$	1.721	1.711	1.864	1.823	2.116	2.017
	$\hat{\beta}_{22}^*$	1.726	1.713	1.859	1.830	2.097	1.965
	$\hat{\beta}_{23}^*$	1.720	1.710	1.872	1.814	2.125	1.971

uncorrelated with 0 mean and 0.122 constant variance. We proposed RSS and 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. Both methods RSS and DRSS sample were used to estimate the spline model with 3 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 where SRS, RSS and DRSS are compared based on the same number of measured units. The yielded SRS samples were used to estimate spline models with 3 knots. This small number of knots is to allow comparison with the simulated RSS and DRSS that have the same number of knots. Last point to mention that all configurations in this simulation study were ran with 10000 replicates.

3.1 Simulated spline models

According to the above simulation arrangements, RSS and DRSS samples were produced after ranking the response variable as discussed in subsection (2.1). Then the generated data sets were used to estimate the following spline model

$$y_{(ii)j}=\beta_0^*+\beta_1^*x_{[ii]j}+\beta_{21}^*(x_{[ii]j}-K_l)_+ +\beta_{22}^*(x_{[ii]j}-K_2)_+ +\beta_{23}^*(x_{[ii]j}-K_3)_+ +e_{(ii)j}^*;\ \ i=1,\ldots,m;\ \ j=1,\ldots,r$$

where $y_{(ii)j}$ is i^{th} smallest response unit that has been selected from i^{th} RSS subsample in the j^{th} cycle , $x_{[ii]j}$ is the predictor variable that is associated with $y_{(ii)j}$; β_0^*,β_1^* ,..., β_{23}^* are model parameters and e is the random error term. Here K_1,K_2,K_3 are the model knots. We fitted this model under the proposed assumptions in (10) by using (12).

To enhance model comparison, the generated SRS samples in the above section (3) were used to estimate the spline model (2) via (3).

Outputs of these simulation trails are summarized in terms relative efficiency of the model parameters in Table (1). Relative efficiencies $eff(\hat{\beta}_i^*,\hat{\beta}_i)$ were computed using (15) for $i=1,2,3,4,5$. These outputs show, with all sampling sizes, that the RSS and DRSS spline models are more efficient than SRS spline models. Also, we can note that the DRSS gained petter performance since it is more efficient than SRS and RSS.

To show effectiveness of extension of the RSS methods, samples were generated after ranking the predictor variable as mentioned in subsection (2.2) to estimate the spline model. Then the produced RSS and DRSS samples were used to fit the spline model in (16) by using the estimated parameters in (17).

We compared the estimated RSS and DRSS spline models, after ranking the predictor variable, with the 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 has been ordered. Consequently, DRSS gained the best model efficiency. Also, the efficiency of the estimator has been increased according to the increment in the ordered sample size m.

Table 2: Relative efficiencies of the simulation study when using DRSS and RSS spline models comparing to the SRS spline models. Order the predictor variable was achieved.

	$m=2$		$m=3$		$m=4$	
	$r=2$	$r=3$	$r=2$	$r=3$	$r=3$	$r=6$
	$n=4$	$n=6$	$n=6$	$n=9$	$n=12$	$n=24$
RSS						
$\hat{\beta}_0^*$	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
DRSS						
$\hat{\beta}_0^*$	1.714	1.713	1.853	1.811	2.105	2.011
$\hat{\beta}_1^*$	1.721	1.701	1.871	1.807	2.114	1.998
$\hat{\beta}_{21}^*$	1.709	1.700	1.862	1.814	2.117	2.100
$\hat{\beta}_{22}^*$	1.711	1.711	1.867	1.813	2.120	2.037
$\hat{\beta}_{23}^*$	1.715	1.712	1.859	1.820	2.109	1.986

Table 3: Relative efficiencies of the estimated parameters in the DRSS and RSS spline model comparing to the SRS model in the practical example. Order the response and predictor variables were achieved.

$m=3 \quad r=8 \quad n=24$

order the response variable	RSS	DRSS	order the predictor variable	RSS	DRSS
$\hat{\beta}_0^*$	1.273	1.887	$\hat{\beta}_0^{**}$	1.205	1.853
$\hat{\beta}_1^*$	1.256	1.891	$\hat{\beta}_1^{**}$	1.211	1.846
$\hat{\beta}_{21}^*$	1.251	1.876	$\hat{\beta}_{21}^{**}$	1.196	1.861
$\hat{\beta}_{22}^*$	1.244	1.895	$\hat{\beta}_{22}^{**}$	1.213	1.858

4 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. 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 [5].

Our investigations on this study is mainly to show efficiency of using DRSS when fitting 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 and DRSS 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 estimates. Later on, we ranked sampling units with respect to the predictor variable to estimate appropriate spline models. And for the purpose of comparison, we selected a SRS of size $n=24$ and then we estimated 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, and from practical point of view, 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 $q=2$. Table(3) shows the relative efficiencies for the estimated spline model parameters by using DRSS and RSS sample units when ordering either the response or the predictor.

As seen in this table, both spline models that were fitted using DRSS and RSS methods are more efficient than model that was fitted using SRS method. Imperatively, the estimated parameters in the DRSS spline model have a superior efficiency more than RSS and SRS.

5 Conclusions and Discussion

This paper defined DRSS procedure for spline model fitting. It proved that estimators of the underlying spline model using DRSS units are more efficient than both RSS and SRS spline models. Tables of the simulation as well as practical studies supported this claim.

Practically, in real data applications where sampling units are difficult or expensive to measure, RSS method and its extension to DRSS are more beneficial than SRS when selecting sampling units because they can reduce sampling costs. 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 relatively cheapest predictor variable to rank among all other expensive predictors.

This paper establishes a paradigm for future research under general linear model scenarios. Applying DRSS 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. [14], Chapter 3, summarized these spline models which prepares an appropriate infrastructure to implement RSS techniques. Moreover, statistical inferences for our improved models can be investigated.

Acknowledgment

The author is grateful to the anonymous referees for significant comments and careful checking the first draft of this paper.

References

[1] M. Al Kadiri, Linear Penalized Spline Model estimation Using Ranked Set Sampling, paper is submitted (2015).
[2] M. Al-Saleh and M. Al Kadiri, Double Ranked Set Sampling, Statistics and Probability Letters **48**, 205–212 (2000).
[3] M. Al Kadiri, R. Carroll and M. Wand, Marginal Longitudinal Semiparametric Regression via Penalized Splines, Statistics & Probability Letters **80**, 1242–1252 (2010).
[4] Z. Chen, Ranked-set sampling with regression-type estimators, Journal of Statistical Planning Inference **92**, 181-192 (2001).
[5] Y. Cohen and J. Cohen, Statistics and Data with R: An Applied Approach Through Examples, Wiley, New York, 2008.
[6] I. DiMatteo, C. Genovese and R. Kass, Bayesian curve-fitting with free-knot splines, Biometrika **88**, 1055–1072 (2001).
[7] P. Eilers and B. Marx, Splines, Knots, and Penalties, Wiley Interdisciplinary Reviews: Computational Statistics **2**, 637–653 (2010).
[8] R. Eubank, The hat matrix for smoothing splines, Statistics & Probability Letters **2**, 9–14 (1984).
[9] R. Eubank, A simple smoothing spline, American Statistician **48**, 103–106 (1994).
[10] C. Gu, Multivariate spline regression, In Schimek M. G. (Eds.), Smoothing and Regression: Approaches, Computation and Application, Wiley, NewYork, 2000,
[11] Y. Li and D. Ruppert, On the Asymptotics of Penalized Splines, Biometrika **95**, 415–436 (2008).
[12] G. McIntyre, A method for Unbiased Selective Sampling, Using Ranked Sets, Australian Journal of Agricultural Research **3**, 385–390 (1952).
[13] R. Phillips, Iterated Feasible Generalized Least-Squares Estimation of Augmented Dynamic Panel Data, Journal of Business and Economic Statistics **28**, 410–422 (2010).
[14] D. Ruppert, M. Wand and R. Carroll R Semiparametric Regression, Cambridge university Press, New York, 2003.
[15] 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 Statistical Mathematics **20**, 421–428 (1968).
[16] M. Wand, A comparison of regression spline smoothing procedures, Computational Statistics **15**, 443462 (2000).
[17] D. Wolfe, Ranked Set Sampling: An Approach to More Efficient Data Collection, Statistical Science **19**, 636–643 (2004).
[18] D. Wolfe, Ranked set sampling: Its relevance and impact on statistical inference, International Scholarly Research Network , Probability and Statistics (2012), DOI: 10.5402/2012/568385.
[19] P. Yu and K. Lam, Regression Estimator in Ranked Set Sampling, Biometrics **53** 1070–1080 (1997).
[20] S. Zhou and X. Shen, Spatially adaptive regression splines and accurate knot selection schemes, Journal of the American Statistical Association **96**, 247–259 (2001).

M. A. Al Kadiri is currently assistant prof. of statistics at Dept of Statistics, Yarmouk University, Jordan. He finished his PhD from Federation University Australia, Victoria, Australia. His research areas of interest are: Sampling methods, Mixed Models, Smoothing models, Statistical Computing, Semiparametric regression, and inference.