

PARAMETER ESTIMATION IN MULTIPLE LINEAR REGRESSION MODELS USING RANKED SET SAMPLING

YAPRAK ARZU ÖZDEMİR, A. ALPTEKİN ESİN

ABSTRACT. In statistical surveys, if the measurements of sampling units according to the variable under consideration is expensive in all sense, and if it is possible to rank sampling units according to the same variable by means of a method which is not expensive at all, in those cases, Ranked Set Sampling (RSS) is a more efficient sampling method than the Simple Random Sampling (SRS) to estimate the population mean. In this study, the effects of using RSS in multiple linear regression analysis are considered in terms of estimation of model parameters. Firstly, according to RSS and SRS the estimates of multiple regression model parameters are obtained and then the effects concerning the variances of the estimators are investigated by Monte Carlo simulation study based on Relative Efficiency (RE) measure. It is shown that the estimators obtained based on RSS are more efficient than the estimators based on SRS when the sample size is small.

1. INTRODUCTION

Recently, especially in the studies relating to areas of ecology, agriculture and medicine, it is a widely encountered situation that the measurement of the variable under consideration is too costly or too difficult in aspects of time and labor. So, in such areas, it is needed to prefer a sampling method in which the population is represented by the smallest sample size in a best way possible. Such a sampling method aiming at this is initially suggested by McIntyre [7] with the name of Ranked Set Sampling (RSS). McIntyre indicates that RSS is a more efficient sampling method than Simple Random Sampling (SRS) method for estimation of the population mean.

In RSS, the sample selection is composed of two stages. At the first stage of sample selection, m simple random samples of size m are selected from an infinite population and each sample is called as set. Equivalently, m^2 units may be drawn from the population and randomly partitioned into m equal samples. Then, each

Received by the editors Nov. 3, 2006; Rev: Feb.3, 2007; Accepted: Feb. 26, 2007.

2000 *Mathematics Subject Classification.* Primary 62D05; Secondary 62J05.

Key words and phrases. Simple Random Sampling, Ranked Set Sampling, Order Statistics, Relative Efficiency, Regression Analysis.

of observations are ranked from the smallest to the largest according to variable of interest, say Y , in each set. This measurement is such a low level measurement which does not cost too much and the ranking of the units can be done by the help of previous experiences, a visual ranking or by the help of a concomitant variable. At the second stage, the first observation unit from the first set, the second observation unit from the second set and going on like this m^{th} observation unit from the m^{th} set are taken and measured according to the variable Y with a high level of measurement satisfying the desired sensitivity. To provide necessary measurements for inference, the entire process (or cycle) can be replicated r times, thus yielding $n = mr$ measured units out of m^2r selected units. Under the assumption that there is no ranking error, these n measured units constitute the ranked set sample which is denoted by $\{Y_{(i)j}; i = 1, 2, \dots, m, j = 1, 2, \dots, r\}$. So, the unbiased estimator of the population mean is defined as

$$\bar{Y}_{RSS} = \frac{1}{mr} \sum_{i=1}^m \sum_{j=1}^r Y_{(i)j}$$

[7]. Dell and Clutter [6] showed that, even if there are an errors in ranking, \bar{Y}_{RSS} is an unbiased estimator of the population mean and furthermore that,

$$\frac{Var(\bar{Y}_{SRS})}{Var(\bar{Y}_{RSS})} \geq 1$$

where \bar{Y}_{SRS} is the SRS sample mean with size n and equality occurs only if the ranking is so poor as to yield a random sample.

RSS is preferred for obtaining efficient parameter estimations in many statistical analyses. The use of RSS in the linear regression model for the first time is introduced by Stokes [12]. Stokes estimated the population mean by modeling the relation between concomitant variable and variable of interest. Muttalak [8, 9] considered RSS for simple and multiple regression models. For both cases, he adopted the unequal variance assumption for the regression errors. However; he did the parameter estimation by using the Least Squares Estimation (LSE) method. So, under his model assumptions there is no improvement in estimating the model parameters. Barnett and Barreto [1] considers simple linear regression model with replicated observations obtained from RSS. Optimal-L estimators of model parameters are obtained under the assumption that dependent variable has normal distribution and the ranking is perfect. They showed that the RE gains can be high for the simple linear regression coefficients. Samawi and Ababneh [11] examined RSS method in simple linear regression model and they assume that variables in the model (X, Y) have a bivariate normal distribution. Also, parameter estimation is done with LSE method under the assumption that the error term of the model is constant variance like SRS. They conclude that, RSS provides a more efficient way to do regression analysis. Chen and Wang [5] considered optimal sampling schemes

with different optimality criteria in the application of RSS for efficient regression analysis.

In this study, the study of Samawi and Ababneh [11] is benefited and multiple linear regression model is considered in case that units are selected by RSS with ranking according to one of independent variables. Estimators of model parameters (regression coefficients) and their variances are examined under the assumption that error terms have constant variance based on the Lemma given by Bhattacharya [2]. Then, for the comparison of the efficiency of the estimated model parameters obtained by using RSS with those obtained by using SRS, we use RE measure. Also, a Monte Carlo simulation study is realized to obtain RE values for different sample size and correlation coefficients under the assumption that the dependent variable and independent variables have multivariate normal distribution.

2. ESTIMATION OF MODEL PARAMETERS USING RSS IN A MULTIPLE REGRESSION MODEL

In regression analysis, since the variation of dependent variable is tried to be explained with the help of independent variables, in the application phase, such situations can be faced as that the measurement of dependent variable can be quite costly or so difficult in aspects of time and labor. In those cases, RSS can be used for obtaining the sample units; however, the method used in ranking process of the units according to dependent variable is important. Chen [3] suggested the adaptive RSS method in derivation of the regression estimator of population mean by ranking with multiple concomitant variables. However, the method needs a ranking criterion function for using all independent variables in ranking. Ranking of the units can be done according to dependent variable, for instance visual ranking, however model assumptions can be violated and also the obligation of expressing the regression model in terms of order statistics can arise. On the other hand, the most appropriate ranking related to dependent variable can be done by using one of the independent variables. For example; the ages of animals need to be determined in animal growth studies but aging an animal is usually time consuming and costly. However, variables on the physical size of an animal, which are closely related to age, can be measured easily and cheaply [5]. So, independent variable which will be used in ranking will also take as a concomitant variable. If sensitive measurement of the units associated with the independent variable is cheap and easy, the units are ranked according to independent variable X with sensitive measurement techniques and if not, with low-level measurement techniques are used which do not bring about ranking error. Here, the primary aim is to reach the most effective information about the dependent variable of which the measurement is the most difficult and expensive, by the help of the independent variable of which measurement is the easiest and cheapest one among independent variables in the regression model and to estimate the more efficient estimators of model parameters.

The regression model between dependent variable Y and independent variables (X_1, X_2, \dots, X_p) can be written as;

$$E(\mathbf{Y}/\mathbf{X}) = \mathbf{XB} \quad (2.1)$$

where \mathbf{B} and \mathbf{X} are p -dimensional parameters and random independent variables vectors, respectively.

If sample selection is done with RSS method for the model 2.1, let X_k be the independent variable which will be used as a concomitant variable for ranking. At the first stage of the sample selection, random sample of size m^2 are selected from the population as $(Y, X_1, X_2, \dots, X_k, \dots, X_p)$ and each unit is partitioned randomly into m sets. Then, sample units in each set are ranked according to X_k variable. Thus, dependent variable Y which is still not measured sensitively and remaining independent variables $(X_1, X_2, \dots, X_{k-1}, X_{k+1}, \dots, X_p)$ will have been ranked according to X_k . At the second stage, the unit which has the smallest X_k value from the first set, the unit which has the second smallest X_k value from the second set and going on like this the unit which has the m^{th} largest X_k value from the m^{th} set are selected and measured sensitively according both to independent variables included in the model and dependent variable Y . This cycle can be replicated r times until the desired sample size is obtained. Thus, multiple linear regression model for ranked set sample of size $n = mr$ is defined as

$$\mathbf{Y}_R = \mathbf{X}_R \boldsymbol{\beta} + \boldsymbol{\epsilon}_R$$

where,

$$\mathbf{Y}_R = \begin{bmatrix} Y_{[1]1} \\ Y_{[2]1} \\ \vdots \\ Y_{[m]1} \\ \vdots \\ Y_{[1]r} \\ Y_{[2]r} \\ \vdots \\ Y_{[m]r} \end{bmatrix}, \mathbf{X}_R = \begin{bmatrix} 1 & X_{1[1]1} & \cdots & X_{k(1)1} & \cdots & X_{p[1]1} \\ 1 & X_{1[2]1} & \cdots & X_{k(2)1} & \cdots & X_{p[2]1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & X_{1[m]1} & \cdots & X_{k(m)1} & \cdots & X_{p[m]1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & X_{1[1]r} & \cdots & X_{k(1)r} & \cdots & X_{p[1]r} \\ 1 & X_{1[2]r} & \cdots & X_{k(2)r} & \cdots & X_{p[2]r} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & X_{1[m]r} & \cdots & X_{k(m)r} & \cdots & X_{p[m]r} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon}_R = \begin{bmatrix} \epsilon_{[1]1} \\ \epsilon_{[2]1} \\ \vdots \\ \epsilon_{[m]1} \\ \vdots \\ \epsilon_{[1]r} \\ \epsilon_{[2]r} \\ \vdots \\ \epsilon_{[m]r} \end{bmatrix}$$

In this model; \mathbf{X}_R is a $nx(p+1)$ -dimensional independent variables matrix the elements of which are $(1, X_{1[i]j}, \dots, X_{k(i)j}, \dots, X_{p[i]j})(i = 1, 2, \dots, m, j = 1, 2, \dots, r)$. $X_{k(i)j}$ denotes i^{th} order statistic of X_k in the i^{th} set and j^{th} cycle and $X_{p[i]j}$ is the X_k - induced i^{th} order statistic of X_p in the i^{th} set and j^{th} cycle. \mathbf{Y}_R is a $nx1$ -dimensional dependent variable vector the elements of which are $Y_{[i]j}$. $Y_{[i]j}$ is the value of the dependent variable for i^{th} independent variables $(1, X_{1[i]j}, \dots, X_{k(i)j}, \dots, X_{p[i]j})$ in the i^{th} set and j^{th} cycle. $\boldsymbol{\beta}$ is a $(p+1) \times 1$ -dimensional parameter vector. $\boldsymbol{\epsilon}_R$ is a $nx1$ -dimensional random error vector with $E(\boldsymbol{\epsilon}_R) = 0$, $Var(\boldsymbol{\epsilon}_R) = \sigma^2 \mathbf{I}$ and $Cov(\boldsymbol{\epsilon}_R, \mathbf{X}_R) = 0$. Constant variance assumption related to error vector can be written from the Lemma about induced order statistics given by Bhattacharya [2]. From this Lemma, denote by $f_{Y/X}(y/x)$ the conditional distribution of Y given X . If (X_i, Y_i) independent and identical distributed random variables then the conditional probability density function of Y given $X_{(i)} = x$ is written as

$$f_{Y_{[i]}/X_{(i)}}(y/X_{(i)} = x) = f_{Y/X}(y/x)$$

where, $X_{(i)}$ is the i^{th} order statistic and $Y_{[i]}$ is the X -induced i^{th} order statistic of Y . This lemma can be also extended to multivariate case [4]. In this study, this structure is used in the estimation of model parameters.

Under these model assumptions, $\boldsymbol{\beta}$ parameter vector is estimated with RSS by using LSE method as below

$$\widehat{\boldsymbol{\beta}}_{RSS} = (\mathbf{X}'_R \mathbf{X}_R)^{-1} \mathbf{X}'_R \mathbf{Y}_R$$

The variance of this estimator is

$$Var(\widehat{\boldsymbol{\beta}}_{RSS}) = \sigma^2 (\mathbf{X}'_R \mathbf{X}_R)^{-1} \quad (2.2)$$

The unbiased estimator of σ^2 is written as

$$\widehat{\sigma}_{RSS}^2 = \frac{\mathbf{Y}'_R \mathbf{Y}_R - \widehat{\boldsymbol{\beta}}_{RSS} \mathbf{X}'_R \mathbf{Y}_R}{mr - (p+1)}$$

3. RELATIVE EFFICIENCIES OF ESTIMATORS

In regression analysis, generally, sample selection is done with SRS. In this study, it is assumed that the units are randomly selected with SRS without any pre-determination of independent variables and the measurement of the values of dependent variable and independent variables taken from each observation unit. By taking a simple random sample of size n from a multivariate normal distribution of $(Y, X_1, X_2, \dots, X_p)$, the model 2.1 can be denoted in matrix notation as

$$\mathbf{Y}_S = \mathbf{X}_S \boldsymbol{\beta} + \boldsymbol{\epsilon}_S \quad (3.1)$$

where,

$$\mathbf{Y}_S = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X}_S = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon}_S = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

In this model; \mathbf{Y}_S is a $nx1$ -dimensional dependent variable vector obtained with SRS. \mathbf{X}_S is a $nx(p+1)$ -dimensional independent variables matrix obtained with SRS. $\boldsymbol{\beta}$ is a $(p+1)x1$ -dimensional parameter vector. $\boldsymbol{\epsilon}_S$ denoted as $nx1$ -dimensional random error vector with $E(\boldsymbol{\epsilon}_S) = 0$, $Var(\boldsymbol{\epsilon}_S) = \sigma^2 \mathbf{I}$ and $Cov(\boldsymbol{\epsilon}_S, \mathbf{X}_S) = 0$.

When sample selection is done with SRS, $\boldsymbol{\beta}$ parameter vector is estimated with LSE method from the model 3.1, like that

$$\hat{\boldsymbol{\beta}}_{SRS} = (\mathbf{X}'_S \mathbf{X}_S)^{-1} \mathbf{X}'_S \mathbf{Y}_S$$

The variance of this estimator is written as

$$Var(\hat{\boldsymbol{\beta}}_{SRS}) = \sigma^2 (\mathbf{X}'_S \mathbf{X}_S)^{-1} \quad (3.2)$$

The unbiased estimator of σ^2 according to SRS is

$$\hat{\sigma}_{SRS}^2 = \frac{\mathbf{Y}'_S \mathbf{Y}_S - \hat{\boldsymbol{\beta}}_{SRS} \mathbf{X}'_S \mathbf{Y}_S}{nr - (p+1)}$$

As it can be seen from equations 2.2 and 3.2, the variances of estimators depend on the expressions $(\mathbf{X}'_R \mathbf{X}_R)^{-1}$ and $(\mathbf{X}'_S \mathbf{X}_S)^{-1}$. So, to compare RSS and SRS method based on RE measure, it is necessary to consider matrix of independent variables \mathbf{X}_R and \mathbf{X}_S as random matrices. Since all possible values of independent variables are considered in this way, what effects RSS creates in parameter estimation can be explained better. In this condition, the variances of estimators according to SRS

and RSS are respectively can be written as;

$$\begin{aligned}
Var(\widehat{\beta}_{SRS}) &= E \left[Var(\widehat{\beta}_{SRS}/\mathbf{X}_S) \right] + Var \left[E(\widehat{\beta}_{SRS}/\mathbf{X}_S) \right] \\
&= E \left[E(\widehat{\beta}_{SRS} - \beta)(\widehat{\beta}_{SRS} - \beta)' / \mathbf{X}_S \right] \\
&= E \left[(\mathbf{X}'_S \mathbf{X}_S)^{-1} \mathbf{X}'_S \sigma^2 \mathbf{I} \mathbf{X}_S (\mathbf{X}'_S \mathbf{X}_S)^{-1} \right] \\
&= \sigma^2 E \left[(\mathbf{X}'_S \mathbf{X}_S)^{-1} \right]
\end{aligned}$$

$$\begin{aligned}
Var(\widehat{\beta}_{RSS}) &= E \left[Var(\widehat{\beta}_{RSS}/\mathbf{X}_R) \right] + Var \left[E(\widehat{\beta}_{RSS}/\mathbf{X}_R) \right] \\
&= E \left[E(\widehat{\beta}_{RSS} - \beta)(\widehat{\beta}_{RSS} - \beta)' / \mathbf{X}_R \right] \\
&= E \left[(\mathbf{X}'_R \mathbf{X}_R)^{-1} \mathbf{X}'_R \sigma^2 \mathbf{I} \mathbf{X}_R (\mathbf{X}'_R \mathbf{X}_R)^{-1} \right] \\
&= \sigma^2 E \left[(\mathbf{X}'_R \mathbf{X}_R)^{-1} \right]
\end{aligned}$$

and RE is defined as

$$RE(\widehat{\beta}_{(l-1)}) = \frac{Var(\widehat{\beta}_{SRS})_{ll}}{Var(\widehat{\beta}_{RSS})_{ll}} = \frac{\sigma^2 E \left[(\mathbf{X}'_S \mathbf{X}_S)^{-1} \right]_{ll}}{\sigma^2 E \left[(\mathbf{X}'_R \mathbf{X}_R)^{-1} \right]_{ll}} \quad l = 1, 2, \dots, p + 1$$

To examine RE values, matrices of $E \left[(\mathbf{X}'_S \mathbf{X}_S)^{-1} \right]$ and $E \left[(\mathbf{X}'_R \mathbf{X}_R)^{-1} \right]$ should be obtained theoretically. However, derivation of these expected values seems to be possible with asymptotic theory. Also, it is not appropriate to put forward any opinion by using asymptotic theory due to small sample sizes. Thus, RE values for regression coefficients in multiple linear regression model with two independent variables are observed by the help of the simulation study.

4. MONTE CARLO SIMULATION STUDY TO ESTIMATE THE RELATIVE EFFICIENCY VALUES

The simulation study is done with Matlab package program with the algorithm given below:

Sample selection for RSS;

1. Data of (Y, X_1, X_2) are generated from a multivariate normal distribution according to different R^2 coefficient of determination values and different ρ_{x_1y} , ρ_{x_2y} and $\rho_{x_1x_2}$ correlation coefficients which satisfy these R^2 values. They are given in Table 1.
2. Sample units with size m^2 are randomly selected from the generated data.
3. Sample is partitioned randomly into m sets including m sample units.
4. Sample units in each set are ranked according to variable X_1 which represents a concomitant variable in practice. In the process of ranking first the unit with the smallest X_1 value from the first set is taken. Then, the unit with the

second smallest X_1 value from the second set is taken and this process continues until the unit with the largest X_1 value is taken from the m^{th} set.

5. When the cycle number is $r > 1$, relatively 2^{nd} , 3^{rd} and 4^{th} steps of the program are replicated r times.

6. At the end of r cycles, ranked set sample of size $n = mr$ is obtained.

Sample selection for SRS;

7. Sample of size $n = mr$ is selected randomly from the population which is generated in Step 1th. This sample constitute of simple random sample of size $n = mr$.

Estimation of model parameters;

8. $\widehat{\beta}_{RSS}$ vector is obtained by using ranked set sample of size n which is derived in Step 6th.

9. $\widehat{\beta}_{SRS}$ vector is obtained by using simple random sample of size n which is derived in Step 7th.

10. All steps from 1st one to 10th one are repeated 300000 times and after 300000 replications, means and variances of estimations obtained in 8th and 9th steps are calculated.

11. The variance values of $\widehat{\beta}_{0(SRS)}, \widehat{\beta}_{1(SRS)}, \widehat{\beta}_{2(SRS)}$ calculated with SRS are divided by the variance values of $\widehat{\beta}_{0(RSS)}, \widehat{\beta}_{1(RSS)}, \widehat{\beta}_{2(RSS)}$ calculated with RSS and RE values of estimators are obtained.

12. At the 4th step, ranking is done according to X_2 and remaining steps are repeated.

In this simulation study, for all possible (m, r) pairs satisfying the sample size $n = 5, 6, 10, 16$ are taken. Correlation values which will be used in the simulation study are defined according to coefficient of determination R^2 . Different ρ_{x_1y} , ρ_{x_2y} and $\rho_{x_1x_2}$ combinations which have the same value of R^2 are used to explore the source of the main effect on RE. R^2 is defined as

$$R^2 = \frac{\rho_{x_1y}^2 + \rho_{x_2y}^2 - 2\rho_{x_1y}\rho_{x_2y}\rho_{x_1x_2}}{1 - \rho_{x_1x_2}^2}$$

ρ_{x_1y} , ρ_{x_2y} and $\rho_{x_1x_2}$ values satisfying 3 different R^2 values are taken in 15 different correlation cases. These cases are given in Table 1. The results related to RE values obtained from the simulation study are given at Figure 1-6.

Table 1 Correlation coefficient values which is used in Monte Carlo simulation study for multiple regression model with two independent variables.

| Number of Cases | R^2 | ρ_{x_1y} | ρ_{x_2y} | $\rho_{x_1x_2}$ |
|-----------------|-------|---------------|---------------|-----------------|
| 1 | 0,50 | 0,50 | 0,50 | 0 |
| 2 | 0,50 | 0,10 | 0,70 | 0 |
| 3 | 0,50 | 0,70 | 0,10 | 0 |
| 4 | 0,50 | 0,50 | 0,60 | 0,25 |
| 5 | 0,50 | 0,60 | 0,50 | 0,25 |
| 6 | 0,72 | 0,60 | 0,60 | 0 |
| 7 | 0,72 | 0,10 | 0,84 | 0 |
| 8 | 0,72 | 0,84 | 0,10 | 0 |
| 9 | 0,72 | 0,60 | 0,80 | 0,45 |
| 10 | 0,72 | 0,80 | 0,60 | 0,45 |
| 11 | 0,98 | 0,70 | 0,70 | 0 |
| 12 | 0,98 | 0,14 | 0,98 | 0 |
| 13 | 0,98 | 0,98 | 0,14 | 0 |
| 14 | 0,98 | 0,80 | 0,95 | 0,60 |
| 15 | 0,98 | 0,95 | 0,80 | 0,60 |

As seen from Figure 1-6, $RE(\hat{\beta}_0)$, $RE(\hat{\beta}_1)$ and $RE(\hat{\beta}_2)$ gets close to 1 as n increases and the highest values of REs are obtained when $m = 5$ and $r = 1$ with the desired sample size of $n = 5$. For the fixed sample size n ,

as set size m increases and cycle size r decreases, the RE values of estimated regression coefficient related to independent variable used in ranking and $RE(\hat{\beta}_0)$ increase (see Figure 1,2,4,6) but the RE values of estimated regression coefficient related to independent variable which was not used in ranking approximates to 1 (see Figure 3,5). For the fixed (m, r) values, REs are not affected from the coefficient of determination R^2 and correlation coefficients ρ_{x_1y} and ρ_{x_2y} . However, when $|\rho_{x_1x_2}| > 0,25$ and the sample size is small, $RE(\hat{\beta}_1)$ and $RE(\hat{\beta}_2)$ values decrease when the ranking is done according to X_1 and X_2 respectively (see Figure 2, 6). $RE(\hat{\beta}_0)$ is not affected from whether the ranking is done according either to X_1 or X_2 and takes similar values for the fixed (m, r) and for different correlation coefficient cases taken into consideration (see Figure 1,4). For the fixed (m, r) and all correlation coefficient cases, $RE(\hat{\beta}_1)$ values obtained from the ranking done according to X_1 and $RE(\hat{\beta}_2)$ values obtained from the ranking done according to X_2 give closer results (see Figure 2 ,6).

As it can be seen in Table 1, in this simulation study just the positive values of correlation coefficients are examined. The reason is that simulation studies done for negative values give the same results with positive ones. For more detail, see Özdemir [10].

5. CONCLUSION

In this study, regression coefficients of multiple linear regression model are estimated by using RSS and RE values of these estimations according to SRS are investigated. By the help of simulation study, multiple linear regression model with two independent variables is taken into consideration and ranking is done according to each independent variable in RSS. Also, different correlation cases for specified R^2 are considered. So, the effects of ranking and correlation cases on the parameter estimation are examined based on RE measure. Based on the simulation results, RSS method in the estimation of model parameters β_0 and β_1 when the ranking is done according to X_1 and in the estimation of model parameters β_0 and β_2 when the ranking is done according to X_2 is more effective than SRS method under the small sample size. However, when the sample size increases, RE values approximate to 1. So, it is expected that the difference between RSS and SRS disappears by means of RE. When $|\rho_{x_1x_2}| < 0.25$ and sample size n is small, RE will take the highest values. In regression analysis, for the high values of $|\rho_{x_1x_2}|$ the collinearity problem may arise. So, in general, it is desired that just one of the independent variables which have high relationship between each other is included in the model. In this case, ranking can be done by the help of the remaining independent variable. In a regression model at which $|\rho_{x_1x_2}|$ does not have a high value, if the sensitive measurement of both of the independent variables is cheap and easy, the independent variable which has a higher correlation with the dependent variable may be preferred in the ranking process. In this way, RE values of estimated regression coefficient related to independent variable used in ranking process will increase, and so, the significance of the related regression coefficient will be supported since the variance of the estimated regression coefficient related to independent variable which has the highest contribution in explaining the model decreases.

In case when studying with a small sample size is desired due to the budget limitations, estimations of model parameters in regression analysis with RSS are more efficient than the estimations obtained with SRS. However, after the selection of the maximum set size m and the minimum repetition number r according to the method which will be used in the ranking process of the units with RSS to satisfy n sample size needed for the application, the efficiencies relating to parameter estimations will take their highest values.

ÖZET: İstatistiksel arařtırmalarda, örnek birimlerinin ilgilenilen deęişken bakımından ölçülmesinin, emek, zaman ve maliyet anlamında pahalı olduęu durumlarda, örnek birimleri eđer pahalı ölçüm gerektirmeyen bazı tekniklerle aynı deęişken bakımından sıralanabiliyorsa, Sıralı Küme Örnekleme (SKÖ), yığın ortalamasını tahmin etmede Basit Tesadüfi Örneklemeye (BTÖ) göre oldukça etkin bir örnekleme yöntemidir. Bu çalışmada, çoklu linear

regresyon analizinde SKÖ kullanımının model parametrelerinin tahmini üzerindeki etkileri ele alınmaktadır. Çoklu doğrusal regresyon model parametreleri, hem SKÖ hem de BTÖ'ye göre tahmin edilerek, tahmin edicilerin varyansları bakımından ortaya çıkan etkiler, Göreli Etkinlik (RE) ölçüsüne dayalı olarak Monte Carlo simülasyon çalışması ile incelenmiştir. SKÖ'ne göre elde edilen tahmin edicilerin örnek çapı küçük iken BTÖ'ye göre elde edilen tahmin edicilerden daha etkin olduğu gösterilmiştir.

REFERENCES

- [1] Barnett, V., Barreto, M.C.M., Best linear unbiased estimators for the simple linear regression model using ranked set sampling, *Environmental and Ecological Statistics*. 6 (1999), 119-133.
- [2] Bhattacharya, P.K., Convergence of sample paths of normalized sums of induced order statistics, *The Annals of Statistics*. 2 (1974), 1034-1039.
- [3] Chen, Z., Adaptive ranked-set sampling with multiple concomitant variables, *Bernoulli*. 8 (2002), 313-322.
- [4] Chen, Z., Bai, Z., Sinha, B.K., *Ranked Set Sampling Theory and Applications*, 2004, New York: Springer-Verlag.
- [5] Chen, Z., Wang, Y., Efficient regression analysis with ranked-set sampling, *Biometrics*. 60 (2004), 994-997.
- [6] Dell, D.R., Clutter, J.L., Ranked set sampling theory with order statistics background, *Biometrics*. 28 (1972), 545-555.
- [7] McIntyre, G.A., A method of unbiased selective sampling using ranked sets, *Australian Journal of Agricultural Research*. 3 (1952), 385-390.
- [8] Muttalak, H.A., Parameters estimation in a simple linear regression using ranked set sampling, *Biometrical Journal*. 37 (1995), 799-810.
- [9] Muttalak, H.A., Estimation of parameters in a multiple regression model using ranked set sampling, *Information and Optimization Sciences*. 17 (1996), 521-533.
- [10] Özdemir, Y.A., Investigation of Parameter Estimation in Linear Regression Models Using Ranked Set Sampling, Gazi Üniversitesi Fen Bilimleri Enstitüsü, unpublished PhD thesis, (2005).
- [11] Samawi, H.M., Ababneh, F.M., On regression analysis using ranked set sample, *Journal of Statistical Research*. 35 (2001), 93-105.
- [12] Stokes, L.S., Ranked set sampling with concomitant variables, *Commun. Statist. Theor. Meth.* A6 (1977), 1207-1211.

Current address: Department of Statistics, Faculty of Science and Arts, University of Gazi, Beşevler, 06500, Ankara, Turkey

E-mail address: yaprak@gazi.edu.tr, alpesin@gazi.edu.tr

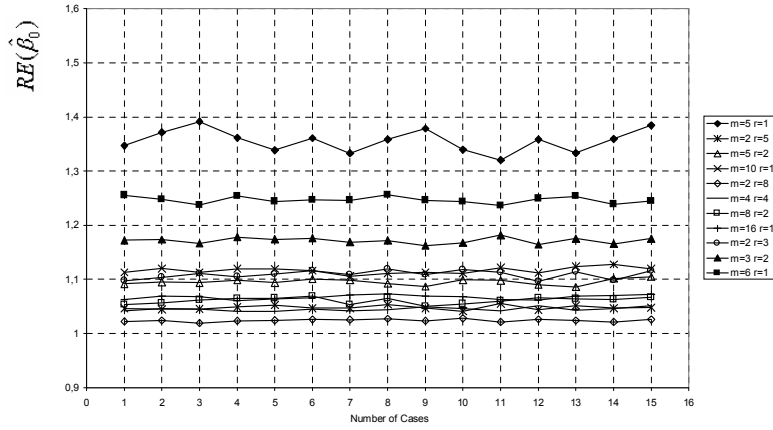


FIGURE 1. Plots of $RE(\hat{\beta}_0)$ values for all considered correlation cases and all possible (m, r) values which satisfy $n = 5, 6, 10, 16$ when the ranking is done according to X_1

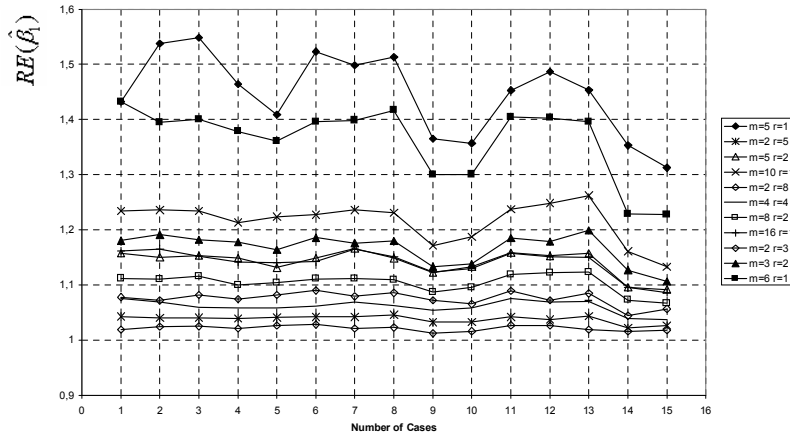


FIGURE 2. Plots of $RE(\hat{\beta}_1)$ values for all considered correlation cases and all possible (m, r) values which satisfy $n = 5, 6, 10, 16$ when the ranking is done according to X_1

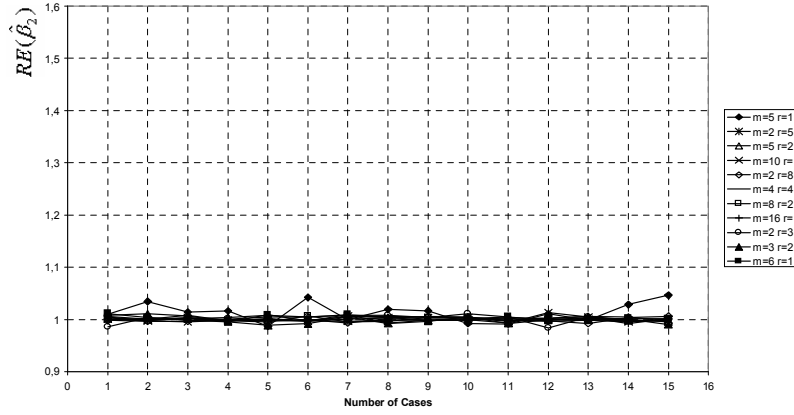


FIGURE 3. Plots of $RE(\hat{\beta}_2)$ values for all considered correlation cases and all possible (m, r) values which satisfy $n = 5, 6, 10, 16$ when the ranking is done according to X_1

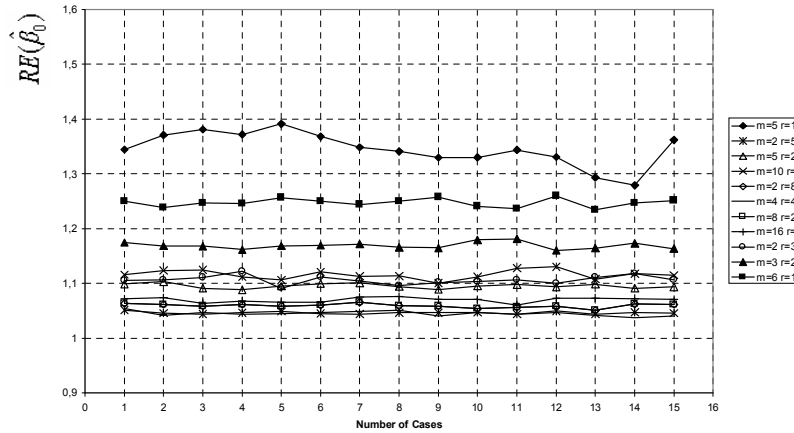


FIGURE 4. Plots of $RE(\hat{\beta}_0)$ values for all considered correlation cases and all possible (m, r) values which satisfy $n = 5, 6, 10, 16$ when the ranking is done according to X_2

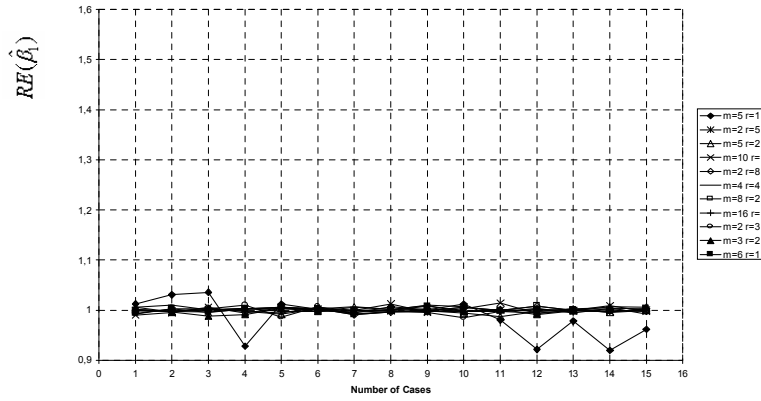


FIGURE 5. Plots of $RE(\hat{\beta}_1)$ values for all considered correlation cases and all possible (m, r) values which satisfy $n = 5, 6, 10, 16$ when the ranking is done according to X_2

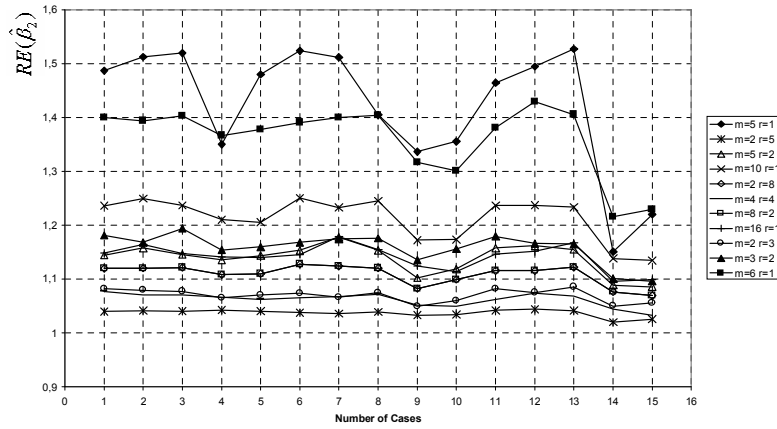


FIGURE 6. Plots of $RE(\hat{\beta}_2)$ values for all considered correlation cases and all possible (m, r) values which satisfy $n = 5, 6, 10, 16$ when the ranking is done according to X_2