On Regression Analysis Using Bivariate Extreme Ranked Set Sampling

Atsu S. S. Dorvlo
Department of Mathematics and Statistics
Sultan Qaboos University
AlKhoud Sultanate of Oman
atsu@squ.edu.om

Walid Abu-Dayyeh
Department of Mathematics and Statistics
Sultan Qaboos University
AlKhoud Sultanate of Oman
walidan@squ.edu.om

Obaid Alsaidy
Department of Mathematics and Statistics
Sultan Qaboos University
AlKhoud Sultanate of Oman
obaidalsaidy@gmail.com

Abstract- Many forms of ranked set samples have been introduced recently for estimating the population mean and other parameters. The ranked set samples are proven to be more efficient than simple random sample in many situations. For multiple characteristics estimation, bivariate ranked set sample was shown to be more efficient than bivariate simple random sample and the usual univariate rank set sample for estimating population means. In this paper, bivariate extreme ranked set sample is introduced and its effect on regression analysis, when the regressors are assumed to be random, is investigated. A simulation study is used to compare the efficiency of the estimators and study the impact of these sampling plans on regression analysis in general. It is shown that the bivariate extreme ranked set sample gives unbiased and more efficient estimators of the regression model parameters, than those obtained by using bivariate simple random sample, univariate ranked set sample and the bivariate ranked set sample, using the same number of quantified observations.

Keywords- Bivariate Extreme Ranked Set Sampling, Bivariate Ranked Set Sampling, Bivariate Simple Random Sampling, Random Regressors, Ranked Set Sampling, Regression Analysis.

I. INTRODUCTION

Linear Regression model attempts to explain the relationship between two or more variables using a straight line. For example, in a chemical process where the yield is to be regressed on the temperature, the temperature cannot be assumed to be constant. When both X and Y are considered random variables, the usual results on estimation, testing and prediction as obtained from the constant X case still hold if (i) the conditional distributions of the Y_i given X_i are normal and independent, with conditional means \( \alpha + \beta X_i \) and conditional variance \( \sigma^2 \) and (ii) the X_i are independent random variables, whose probability distribution does not involve the parameters, \( \alpha \), \( \beta \) and \( \sigma^2 \) [6]. Furthermore, the major modification occurs in the interpretation of confidence coefficients and specified risks of error. When X is random, these refer to repeated
sampling of pairs of \((X_i, Y_i)\) values where the \(X_i\) values as well as the \(Y_i\) values change from sample to sample. In any case, the observed data are used to obtain estimates of the unknown parameters. The method of estimation used in this paper is the ordinary least squares method.

In practice, data are collected on each of a number of units or cases on these variables using simple random sampling (SRS) technique. As an alternative for SRS, the rank set sampling (RSS) procedure, as introduced by McIntyre [3], has been shown to improve the efficiency of estimating a population mean. The RSS procedure can be described as follows: Identify a group of sampling units from the target population. Then, randomly partition the group into disjoint subsets each having a pre-assigned size \(m\). In the most practical situations, the size \(m\) will be 2, 3 or 4. Then, rank each subset by a suitable method of ranking like prior information, visual inspection or by the experimenter himself . . . etc. Then the \(i^{th}\) judgment ordered statistic from the \(i^{th}\) subset, \(i = 1, \ldots, m\), will be quantified, say \(X_{(i)}\). Therefore, \(X_{(1)}\), \(X_{(2)}\), . . . , \(X_{(m)}\) denote a RSS. This represents one cycle. We can repeat the whole procedure \(c\)-times to get a RSS of size \(m^2\). Thus we have \(m^2\) observations denote by : \((X_{(i)} , Y_{(i)})\), \(i=1,2\ldots m\) and \(j=1,2\ldots m\).

5. The process continues until the label \((m, m)\) is resembled from the \((m^2)\)th (last) pool. The above procedure produces a BVRSS of size \(m^2\). Thus we have \(m^2\) observations denote by : \((X_{(i)} , Y_{(i)})\), \(i=1,2\ldots m\) and \(j=1,2\ldots m\).

6. The procedure can be repeated \(r\) times to obtain a sample of size \(n = m^2r\) which will be denoted by \((X_{(i)} , Y_{(i)})\), \(i=1,2\ldots m\) and \(j=1,2\ldots r\).

Now we will introduce bivariate extreme ranked set sample. For any \(m\), obtain

\[
(X_{(1)} , Y_{(1)}) , (X_{(1)} , Y_{(m)}) , (X_{(m)} , Y_{(m)})
\]

and repeat \(q\) times so that \(n=4q\).

Denote this sample by

\[
(X_{(i)} , Y_{(i)}) \quad , \quad i,j = 1,m , k = 1, 2,\ldots, q.
\]

This will be our bivariate extreme ranked set sample BVERSS.

Muttlak [4] used RSS to estimate the parameters of the simple regression model assuming that \(X\) values are known constants. However, under his model assumptions there is no improvement in estimating the model parameters. To achieve any improvement of estimated model parameters, it is required to assume that both \(Y\) and \(X\) are random variables. Moreover, he assumed that ranking is done on one variable (say \(X\)) only. Therefore, \(X\) can be perfectly ranked but \(Y\) will be in error in ranking. Samawi and Ababneh [8], used RSS, ranking only on the
variable $X$, to investigate the effect of RSS on regression analysis in general. They found that the efficiency of using RSS for model parameter estimation is just slightly better than using SRS. However, Samawi and Abu-Dayyeh [10] showed that using extreme ranked set sampling (ERSS) scheme will improve the performance of estimating the model parameters better than using RSS [9]. Also, Samawi and Al-Saleh [11] used BVRSS to investigate its effect on regression analysis in general. They found that the efficiency of using BVRSS for model parameter estimation is just slightly better than using SRS and RSS.

In this paper, the effect of BVRSS on regression analysis assuming that both $Y$ and $X$ are random variables is investigated. Note that, in BVRSS and BVERSS ranking is done on both variable $X$ and $Y$ simultaneously. In Section 2, the general set up of the problem, notation and simple linear regression using BVERSS is discussed. Inference for the simple regression model parameters ($\alpha$ and $\beta$) using asymptotic results is discussed in Section 3. In Section 4 the effect of using BVERSS on residual analysis and model diagnosis is illustrated using real data.

II. LINEAR REGRESSION

A. Some Notations

The following notations will be used throughout this paper. $E(X) = \mu_X$, $E(Y) = \mu_Y$, $\text{Var}(X) = \sigma_X^2$, $\text{Var}(Y) = \sigma_Y^2$, $\rho = \text{Cov}(X, Y)$, $E(X_{(i,j)}Y_{(j)}) = \mu_{X_{(i,j)}Y_{(j)}}$, $E(X_{(i,j)}^2) = \mu_{X_{(i,j)}^2}$, $E(Y_{(j)}^2) = \mu_{Y_{(j)}^2}$, $E(Y_{(j)}) = \mu_Y$, $\text{Cov}(X_{(i,j)}, Y_{(j)}) = \sigma_{X_{(i,j)}Y_{(j)}}$, $\text{Var}(X_{(i,j)}) = \sigma^2_{X_{(i,j)}}$, $\text{Var}(Y_{(i,j)}) = \sigma^2_{Y_{(i,j)}}$, $\text{Cov}(X_{(i,j)}Y_{(j)}) = \sigma_{X_{(i,j)}Y_{(j)}}$

B. Simple Linear Regression Using BVERSS

$$\text{var}(\hat{Y}_{(i,j)k}) = \frac{\sigma^2_{e}}{n} \left[ 1 + \frac{E\left(\frac{Y_{(i,j)k} - \bar{X}_{\text{bverss}}}{S^2_{X_{\text{bverss}}}}\right)^2}{\bar{X}^2_{\text{bverss}}} \right]$$

$$+ \beta^2 \sigma^2_{X_{(i,j)k}}$$

be a BVERSS from the population with size $n=4q$. The simple regression model of the two variables $Y$ and $X$ is defined by: $Y_{(i,j)k} = \alpha + \beta X_{(i,j)k} + \epsilon_{ijk}$

where $\alpha$ is the model intercept, $\beta$ is the model slope and $\epsilon_{ijk}$ is the random error. The assumptions needed here for the purpose of parameters estimation are the mean of the error is zero, its variance is finite and they are uncorrelated. Also $X_i$ and $\epsilon_i$ are independent. Then the least squares estimators of $\alpha$ and $\beta$ are given by

$$\hat{\alpha}_{\text{bverss}} = \bar{Y}_{\text{bverss}} - \bar{X}_{\text{bverss}} \bar{X}_{\text{bverss}}$$

$$\hat{\beta}_{\text{bverss}} = \frac{\sum_{i,j,k}(X_{(i,j)k} - \bar{X}_{\text{bverss}})(Y_{(i,j)k} - \bar{Y}_{\text{bverss}})}{\sum_{i,j,k}(X_{(i,j)k} - \bar{X}_{\text{bverss}})^2}$$

Then the fitted model is

$$\hat{Y}_{(i,j)k} = \hat{\alpha}_{\text{bverss}} + \hat{\beta}_{\text{bverss}} X_{(i,j)k}$$

$$e_{(i,j)k} = Y_{(i,j)k} - \hat{Y}_{(i,j)k}$$

Also, a consistent unbiased estimator for

$$\sigma^2_e = \frac{\sum_{i,j,k} e_{(i,j)k}^2}{n-2}.$$
Using the basic properties of conditional moments and the results of Samawi and Al-Saleh [11], we have the following theorem. The proof is straightforward and will not be presented here.

**Theorem:** Assuming the conditions of the regression model above, then

1. \( E(\hat{\alpha}_{bverss}) = \alpha \) and \( E(\hat{\beta}_{bverss}) = \beta \)
2. \( \text{Var}(\hat{\alpha}_{bverss}) = \frac{\sigma^2}{n} \left[ 1 + E\left( \frac{\bar{X}^2_{\text{bverss}}}{S^2_{X,bverss}} \right) \right] \)

where

\[
S^2_{X,bverss} = \frac{\sum (X_{ij})^2 - \bar{X}^2_{\text{bverss}})^2}{n}
\]

3. \( \text{Var}(\hat{\beta}_{bverss}) = \frac{\sigma^2}{n} \left[ E\left( \frac{1}{S^2_{X,bverss}} \right) \right] \)
4. \( E\left( \frac{\hat{Y}_{ij,k}}{X_i} = x_i \right) = \alpha + \beta x_{ij,k} \)
5. \( \text{Var}(\hat{Y}_{ij,k}) = \frac{\sigma^2}{n} \left[ 1 + E\left( \frac{(x_{ij,k} - \bar{X}_{\text{bverss}})^2}{S^2_{X,bverss}} \right) \right] \)

\[
+ \beta^2 \sigma^2_{X}\]

6. \( E\left( e_{ij,k} \right) = 0 \)
7. \( \text{Var}(e_{ij,k}) = \sigma^2 \left[ 1 + E\left( \frac{(x_{ij,k} - \bar{X}_{\text{bverss}})^2}{nS^2_{X,bverss}} \right) \right] \)

8. \( \text{cov}(\hat{\alpha}_{bverss}, \hat{\beta}_{bverss}) = -\sigma^2 \frac{\bar{X}_{\text{bverss}}}{E(\bar{Y}_{\text{bverss}})} \)
9. \( E(\hat{\sigma}^2) = \sigma^2 \)
10. \( \text{var}(\hat{\sigma}^2) = \frac{2\sigma^4}{n-2} \) (assuming normality)
11. \( \hat{\sigma}^2 \to \sigma^2 \)

**III. NUMERICAL COMPUTATIONS**

From Theorem 1 and the above results, we can derive the efficiencies of the estimators of \( \alpha \) and \( \beta \) using BVERSS relative to the estimators using BVSRS as follows:

\[
\text{eff}(\hat{\alpha}_{bverss}, \hat{\alpha}_{bverss}) = \frac{[1 + E\left( \frac{\bar{X}^2_{\text{bverss}}}{S^2_{X,bverss}} \right) - E\left( \frac{\bar{X}^2_{\text{bverss}}}{S^2_{X,bverss}} \right)]}{\left[ 1 + E\left( \frac{\bar{X}^2_{\text{bverss}}}{S^2_{X,bverss}} \right) \right]} \]

and \( \text{eff}(\hat{\beta}_{bverss}, \hat{\beta}_{bverss}) = \frac{1}{E(\bar{S}^2_{X,bverss})} \)

Definitions of the efficiencies of the estimators of \( \alpha \) and \( \beta \) using RSS, BVRSS and BVERSS relative to the estimators using BVSRS can be defined as above.

Assume that \((X, Y)\) follow a bivariate normal distribution, the performance of simple regression parameter estimators using BVSRS, RSS, BVRSS and BVERSS, for \( m = 2, 3, 4; r=1 \) to 4 and \( \rho = \pm 0.2, \pm 0.5, \pm 0.6, \pm 0.8, \pm 0.9 \) is investigated based on an extensive simulation study to evaluate the expressions efficiencies. In this case, the total sample size is \( n = m^2 r \) for all sample schemes used. Ranking on the concomitant variable \( X \) only is considered when using RSS. The simulation results are given in Tables 1 and 2.

**IV. RESULTS and CONCLUSION**

From the simulation results we may conclude the following:

1. The RSS, BVRSS and BVERSS are all more efficient than BVSRS.
2. BVERSS is the best followed by BVRSS which is followed RSS.
On Regression Analysis Using Bivariate Extreme Ranked Set Sampling

### TABLE I
THE RELATIVE PRECISION OF THE INTERCEPT AND SLOPE ESTIMATORS

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( r )</th>
<th>( m )</th>
<th>( \text{RSS} )</th>
<th>( \text{BVRSS} )</th>
<th>( \text{BVERS} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.8</td>
<td>1</td>
<td>2</td>
<td>1.33</td>
<td>1.46</td>
<td>1.46</td>
</tr>
<tr>
<td>-0.8</td>
<td>1</td>
<td>4</td>
<td>1.05</td>
<td>1.07</td>
<td>1.32</td>
</tr>
<tr>
<td>-0.8</td>
<td>4</td>
<td>2</td>
<td>1.02</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>-0.8</td>
<td>4</td>
<td>4</td>
<td>1.01</td>
<td>1.02</td>
<td>1.23</td>
</tr>
<tr>
<td>-0.2</td>
<td>1</td>
<td>2</td>
<td>1.17</td>
<td>1.24</td>
<td>1.21</td>
</tr>
<tr>
<td>-0.2</td>
<td>1</td>
<td>4</td>
<td>1.05</td>
<td>1.05</td>
<td>1.26</td>
</tr>
<tr>
<td>-0.2</td>
<td>4</td>
<td>2</td>
<td>1.02</td>
<td>1.03</td>
<td>1.03</td>
</tr>
<tr>
<td>-0.2</td>
<td>4</td>
<td>4</td>
<td>1.01</td>
<td>1.01</td>
<td>1.19</td>
</tr>
<tr>
<td>Slope</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.8</td>
<td>1</td>
<td>2</td>
<td>1.20</td>
<td>1.34</td>
<td>1.36</td>
</tr>
<tr>
<td>-0.8</td>
<td>1</td>
<td>4</td>
<td>1.07</td>
<td>1.10</td>
<td>1.96</td>
</tr>
<tr>
<td>-0.8</td>
<td>4</td>
<td>2</td>
<td>1.02</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>-0.8</td>
<td>4</td>
<td>4</td>
<td>1.01</td>
<td>1.02</td>
<td>1.77</td>
</tr>
<tr>
<td>-0.2</td>
<td>1</td>
<td>2</td>
<td>1.14</td>
<td>1.22</td>
<td>1.19</td>
</tr>
<tr>
<td>-0.2</td>
<td>1</td>
<td>4</td>
<td>1.07</td>
<td>1.07</td>
<td>1.73</td>
</tr>
<tr>
<td>-0.2</td>
<td>4</td>
<td>2</td>
<td>1.02</td>
<td>1.03</td>
<td>1.03</td>
</tr>
<tr>
<td>-0.2</td>
<td>4</td>
<td>4</td>
<td>1.01</td>
<td>1.01</td>
<td>1.60</td>
</tr>
</tbody>
</table>

### TABLE II
THE RELATIVE PRECISION OF THE INTERCEPT AND SLOPE ESTIMATORS

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( r )</th>
<th>( m )</th>
<th>( \text{RSS} )</th>
<th>( \text{BVRSS} )</th>
<th>( \text{BVERS} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>1</td>
<td>2</td>
<td>1.20</td>
<td>1.24</td>
<td>1.15</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>4</td>
<td>1.05</td>
<td>1.05</td>
<td>1.26</td>
</tr>
<tr>
<td>0.2</td>
<td>4</td>
<td>2</td>
<td>1.02</td>
<td>1.03</td>
<td>1.02</td>
</tr>
<tr>
<td>0.2</td>
<td>4</td>
<td>4</td>
<td>1.01</td>
<td>1.01</td>
<td>1.19</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>2</td>
<td>1.21</td>
<td>1.37</td>
<td>1.36</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>4</td>
<td>1.05</td>
<td>1.07</td>
<td>1.32</td>
</tr>
<tr>
<td>0.8</td>
<td>4</td>
<td>2</td>
<td>1.03</td>
<td>1.04</td>
<td>1.03</td>
</tr>
<tr>
<td>0.8</td>
<td>4</td>
<td>4</td>
<td>1.01</td>
<td>1.02</td>
<td>1.23</td>
</tr>
<tr>
<td>Slope</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>2</td>
<td>1.19</td>
<td>1.24</td>
<td>1.15</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>4</td>
<td>1.07</td>
<td>1.07</td>
<td>1.73</td>
</tr>
<tr>
<td>0.2</td>
<td>4</td>
<td>2</td>
<td>1.02</td>
<td>1.03</td>
<td>1.02</td>
</tr>
<tr>
<td>0.2</td>
<td>4</td>
<td>4</td>
<td>1.01</td>
<td>1.02</td>
<td>1.60</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>2</td>
<td>1.10</td>
<td>1.35</td>
<td>1.29</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>4</td>
<td>1.07</td>
<td>1.10</td>
<td>1.96</td>
</tr>
<tr>
<td>0.8</td>
<td>4</td>
<td>2</td>
<td>1.03</td>
<td>1.05</td>
<td>1.05</td>
</tr>
<tr>
<td>0.8</td>
<td>4</td>
<td>4</td>
<td>1.02</td>
<td>1.02</td>
<td>1.77</td>
</tr>
</tbody>
</table>

(3) The efficiencies of BVRSS and RSS decrease while BVERS increase for fixed \( r \) and fixed \( \rho \) as \( m \) increases.

(4) The efficiencies decrease for fixed \( m \) and fixed \( \rho \) as \( r \) increases.

### REFERENCES


