Science and Technology, 7 (2002) 311-322

Regression Estimator Using Double Ranked Set Sampling

Hani M. Samawi* and Eman M. Tawalbeh**

Department of Mathematics and Statistics, College of Science, Sultan Qaboos University, P.O.Box 36, Al Khod 123, Muscat, Sultanate of Oman,Department of Statistics, Yarmouk University, Irbid, Jordan 211-63, *Email: hsamawi@squ.edu.om.*

مقدر الانحدار باستخدام طريقة المعاينة المرتبة الثنائية المراحل

هاني سماوي وإيمان طوالبه

خلاصة : أداء طريقة المعاينة المرتبة الثنائية المراحل للتقدير الانحداري سوف يبحث هنا عندما يكون الوسط الحسابي للمتغير المساعد غير معروف. أن التحليلات الأولية وطرق المحاكاة ، دلت على أن مقدر الانحدار باستخدام طريقة المعاينة المرتبة الثنائية المراحل هو أكثر فاعلية من مقدرات الانحدار باستخدام أيّ من طرق المعاينة المرتبة وغير المرتبة الأخرى . وأيضاً هو أفضل وأكثر فاعلية من المقدرات البسيطة باستخدام الأوساط الحسابية لأي من طرق المعاينة المرتبة وغير المرتبة، أحادية أو ثنائية المراحل.

ABSTRACT: The performance of a regression estimator based on the double ranked set sample (DRSS) scheme, introduced by Al-Saleh and Al-Kadiri (2000), is investigated when the mean of the auxiliary variable X is unknown. Our primary analysis and simulation indicates that using the DRSS regression estimator for estimating the population mean substantially increases relative efficiency compared to using regression estimator based on simple random sampling (SRS) or ranked set sampling (RSS) (Yu and Lam, 1997) regression estimator. Moreover, the regression estimator using DRSS is also more efficient than the naïve estimators of the population mean using SRS, RSS (when the correlation coefficient is at least 0.4) and DRSS for high correlation coefficient (at least 0.91.) The theory is illustrated using a real data set of trees.

KEYWORDS: Double Extreme Ranked Set Sample; Double Ranked Set sample; Extreme Ranked Set Sample; Ranked Set Sample; Regression Estimator.

1. Introduction

In many applications, considerable cost savings can be achieved if the number of quantifications is only a small fraction of the number of available units, although all units contribute to the information content of the quantification. Ranked set sampling (RSS) is a method of sampling that can achieve this goal. RSS was first introduced by McIntyre (1952). It is highly powerful and much superior to the standard simple random sampling (SRS) for estimating some population parameters.

RSS can be applied in agricultural, environmental and human populations. For example, the level of bilirubin in the blood of infants can be ranked visually by observing: (*i*) Color of the face. (*ii*) Color of the chest. (*iii*) Color of lower part of the body. (*iv*) Color of terminal parts of the whole body. As the yellowish goes from (*i*) to (*iv*), the level of bilirubin in the blood goes higher (see Samawi and Al-Sakeer 2001).

Al-Saleh and Al-Kadiri (2000) showed that the efficiency of estimating the population mean could be improved even more by using double ranked set sampling (DRSS). Also, they proved that ranking in the second stage is easier than in the first stage. Moreover, as a variation of RSS Samawi *et al.* (1996) investigated extreme ranked set sample (ERSS) and also suggested double extreme

ranked set sampling (DERSS) Samawi (2002). More details about RSS can be found in Kaur *et al.,* (1995) and Patil *et al.* (1999). In this paper, we investigate the performance of DRSS for estimating the population mean using the regression estimator. Theoretical and numerical comparisons with other estimators will be considered. In section 2, notations, definitions and some basic results are introduced. The regression estimator using SRS and, RSS regression estimator (Yu and Lam, 1997) are introduced in section 3. Our proposed regression estimator using DRSS and its properties are given in section 4. In section 5, we illustrate the theory using a set of data representing a real life situation.

2. Sample Notation and Definition with Some Useful Results

2.1 One Stage Sampling

2.1.1 Univariate Population

RSS involves selecting r random sets each of size r from the target population. In the most practical situations, the size r will be 2, 3 or 4. Rank each set by a suitable method of ranking, for example, by using prior information or visual inspection. In sampling notation this implies:

$$\begin{bmatrix} X_{11}, & X_{12}, & \cdots, & X_{12} \\ X_{21}, & X_{22}, & \cdots, & X_{2r} \\ \vdots & & \vdots & \\ X_{r1}, & X_{r2}, & \cdots, & X_{rr} \end{bmatrix} \xrightarrow{\text{after ranking}} \begin{bmatrix} X_{1(1)}, & X_{1(2)}, & \cdots, & X_{1(r)} \\ X_{2(1)}, & X_{2(2)}, & \cdots, & X_{2(r)} \\ \vdots & & \vdots & \\ X_{r(1)}, & X_{r(2)}, & \cdots, & X_{r(r)} \end{bmatrix} \tag{2.1}$$

where X_{ij} denotes the i-*th* observation in the j-*th* set and $X_{j(i)}$ is the i-*th* ordered statistic in the j-*th* set. Only the elements $X_{1(1)}, X_{2(2)}, ..., X_{r(r)}$ are quantified i.e. the element with smallest rank from the first set, the second smallest from the second set, and so on until the largest unit from the r-*th* set is measured. This represents one cycle of RSS. We can repeat the whole procedure m times to get a RSS of size $n = mr$ (Takahasi and Wakimoto, 1968).

2.1.2 For bivariate population

Samawi and Muttlak (1996) modified the above procedure in the case of bivariate distributions to estimate the population ratio, $R = \mu_Y / \mu_X$. The procedure is described as follows: First choose r^2 independent bivariate elements from a population, with bivariate distribution function $F(x, y)$. Rank each set with respect to one of the variables Y or X. Suppose ranking is on variable X. Apply the same procedures as in case of univariate population but for each measured unit from the X's, the associated unit from the Y's is measured too. This may be repeated m times to get a bivariate sample of size $n = rm$. In sample notation: The sample $\{\left(X_{i(i)k}, Y_{i[i]k}\right),\ i = 1,2,...,r;\ k = 1,2,\ldots,m\}$ will denote the bivariate RSS.

2.2 Double Ranked Samples (Two stage sampling)

As a variation of RSS, Al-Saleh and Al-Kadiri (2000) introduced the DRSS procedure as follows:

1. Identify r^3 elements from the target population and divide these elements randomly into r sets each of size r^2 elements.
2. Apply the usual RSS procedure to each set to obtain r RSS, each of size r.
3. Employ again the RSS procedure in Step 2, to obtain the DRSS of size r.
4. We may repeat steps 1-3 m times to obtain a sample of size $n = rm$.

In sampling notation, after ranking each sample separately in each subset, we get:

$$\begin{bmatrix} X^1_{1(1)k} X^1_{1(2)k} \ldots\ldots X^1_{1(r)k} \\ X^1_{2(1)k} X^1_{2(2)k} \ldots\ldots X^1_{2(r)k} \\ \vdots \\ X^1_{r(1)k} X^1_{r(2)k} \ldots\ldots X^1_{r(r)k} \end{bmatrix}, \ldots, \begin{bmatrix} X^r_{1(1)k} X^r_{1(2)k} \ldots X^r_{1(r)k} \\ X^r_{2(1)k} X^r_{2(2)k} \ldots X^r_{2(r)k} \\ \vdots \\ X^r_{r(1)k} X^r_{r(2)k} \ldots X^r_{r(r)k} \end{bmatrix}, \tag{2.2}$$

k=1,2,…,m, where $X^{(l)}_{i(i)k}$ is the *i-th* ordered observation in the *i-th* sample of the *l-th* set in the *k-th* cycle. Use RSS scheme on each subset separately, to get

$$A_{ik} = \left\{X^1_{1(1)k}, X^1_{2(2)k}, \ldots X^1_{r(r)k}\right\}, \ldots . A_{ik} = \left\{X^r_{1(1)k}, X^r_{2(2)k}, \ldots X^r_{r(r)k}\right\} \quad .$$

Then in the second stage, let $W_{i(i)k}$ = *i-th* smallest observation in A_{ik}, then $\{W_{i(i)k}$, i=1,2,…,r, k=1,2,…,$m\}$ will denote the DRSS. Now let $W_{1(1)k}, \ldots, W_{r(r)k}$, k=1, 2, …., m, be a DRSS, where the mean and variance of $W_{i(i)k}$ are $\mu^{**}_{(i)}$ and $\sigma^{**^2}_{(i)}$, respectively. Al-Saleh and Al-Kadiri (2000) showed that:

$$\mu = \frac{1}{r}\sum_{i=1}^{r} \mu^{**}_{(i)} \qquad \text{and} \qquad \sigma^2 = \frac{1}{r}\left[\sum_{i=1}^{r} \sigma^{**^2}_{(i)} + \sum_{i=1}^{r} \left(\mu^{**}_{(i)} - \mu\right)^2\right]$$

where μ and σ^2 are the mean and the variance of the population, respectively. Also, it was shown that ranking in the second stage is easier than in the first stage.

3. Regression Estimators Using SRS and RSS

As in ratio estimation, the linear regression estimator is used to increase the precision of estimating the population mean by using extra information in an auxiliary variable X that is correlated with the survey variable Y. When the relation is approximately linear, and the line does not go through the origin, an estimate of the population mean based on the linear regression of Y on X is suggested rather than using the ratio of the two variables.

3.1 Regression estimator using SRS

Let (X_i, Y_i), i=1,2,…,r, be a bivariate sample from $F(x, y)$, and assume that

$$Y_i = \mu_y + \beta\left(X_i - \mu_x\right) + \varepsilon_i \tag{3.1}$$

where μ_x and μ_y are the means of X and Y respectively, and for fixed X_i, the ε_i's, i=1,2,…,r are *i.i.d* (independent and identically distributed) with mean zero and variance $\sigma^2_\varepsilon = \sigma^2_y\left(1-\rho^2\right)$.

Consider the case where μ_x is unknown. The method of double sampling can be used to obtain an estimate of μ_y. This involves drawing of a large random sample of size n', which is used to estimate μ_y. Then a subsample of size n is selected from the original selected units to study the primary characteristic of Y. Setting $n' = r^2 m$ and $n = rm$, the first and the second-phase samples are simple random samples. Then the double-sampling regression estimator $\bar{Y}_{ds}$ is given by

$$\bar{Y}_{ds} = \bar{Y}_{SRS} + \hat{\beta}\left(\bar{X}' - \bar{X}_{SRS}\right), \tag{3.2}$$

where $\bar{X}_{SRS} = \frac{1}{rm}\sum X_i$, $\bar{Y}_{SRS} = \frac{1}{rm}\sum Y_i$, $\bar{X}'$ is the sample mean of X based on r^2m observations of X in the first phase and

$$\hat{\beta} = \frac{\sum\left(X_i - \overline{X}_{SRS}\right)\left(Y_i - \overline{Y}_{SRS}\right)}{\sum\left(X_i - \overline{X}_{SRS}\right)^2}.$$

When the underlying distribution of (X, Y) is assumed to be bivariate normal, the regression estimator $\bar{Y}_{ds}$ is an unbiased estimator for μ_y and its variance is given by

$$Var\left(\bar{Y}_{ds}\right) = \frac{\sigma_\varepsilon^2}{n}\left(1 + \frac{r-1}{r(n-3)}\right) + \frac{1}{r^2 m}\rho^2\sigma_y^2 \tag{3.3}$$

(Sukhatme and Sukhatme, 1970). If the assumption of the linear relationship in (3.1) is invalid, then the SRS regression estimator in (3.2) is in general a biased estimator of μ_y.

3.2 Regression estimator using RSS

Consider the bivariate RSS. From (3.1) the relationship between $Y_{[i]k}$ and $X_{(i)k}$ is

$$Y_{[i]k} = \mu_y + \beta\left(X_{(i)k} - \mu_x\right) + \varepsilon_{[i]k}, \quad i=1,2,\ldots,r \text{ and } k=1,2,\ldots,m. \tag{3.4}$$

Again, when μ_x is unknown the method of double sampling (two-phase sampling) can be used to obtain an estimate of μ_x. Note that the first-phase sample is a simple random sample and the second-phase sample is a ranked set sample. Then the double-sampling regression estimator $\bar{Y}_{Rds}$ based on RSS as in Yu and Lam (1997) have given by:

$$\bar{Y}_{Rds} = \bar{Y}_{RSS} + \hat{\beta}\left(\bar{X}' - \bar{X}_{RSS}\right), \tag{3.5}$$

where $\bar{X}'$ is the sample mean of X based on the r^2m observations of the first phase. Furthermore, using the basic properties of conditional moments, Yu and Lam (1997) showed that $\bar{Y}_{Rds}$ is an unbiased estimator of μ_y under (3.4), and the variance is given by:

$$Var\left(\bar{Y}_{Rds}\right) = \frac{\sigma_\varepsilon^2}{n}\left(1 + E\left[\frac{\left(\bar{Z}_{RSS} - \bar{Z}\right)^2}{S_{zR}^2}\right]\right) + \frac{1}{r^2 m}\rho^2\sigma_y^2, \tag{3.6}$$

where

$$\bar{Z} = \left(\bar{X}' - \mu_x\right)/\sigma_x,\ Z_{(i)k} = \left[X_{(i)k} - \mu_x\right]/\sigma_x,\ \bar{Z}_{RSS} = \frac{1}{rm}\sum_i\sum_i Z_{(i)k},$$

$$S_{zR}^2 = \frac{1}{rm}\sum_k\sum_i\left(Z_{(i)k} - \bar{Z}_{RSS}\right)^2.$$

Again, if the assumption of linear relationship is invalid, the RSS regression estimator in (3.5) is in general a biased estimator of μ_y. Next we will propose our approach for using a regression estimator for estimating μ_y based on DRSS.

4. Regression Estimator Using DRSS

4.1 DRSS for regression estimator

In the two-phase regression estimator using DRSS, for the *k-th* cycle, in the first stage r quantified RSS samples each of size r are considered. The following will denote the first stage sampling:

$$A_{1k} = \left\{ X^1_{1(1)k}, X^1_{2(2)k}, ..., X^1_{r(r)k} \right\},$$
$$A_{2k} = \left\{ X^2_{1(1)k}, X^2_{2(2)k}, ..., X^2_{r(r)k} \right\},$$
$$A_{rk} = \left\{ X^r_{1(1)k}, X^r_{2(2)k}, ..., X^r_{r(r)k} \right\}, \quad k = 1, 2, ..., m.$$

These sets of quantified observations, of size mr^2, are used to estimate μ_x, the population mean of the variable *X*, which is assumed to be unknown. In the second stage a bivariate DRSS, of size $n=rm$, which is $\left\{ \left(W_{i(i)k}, Y_{i[i]k} \right) : \; i = 1, 2, ..., r, k = 1, 2, ..., m \right\}$ is measured.

Note that, rankings in the second stage on the variable *X* are based on the exact measures, i.e. perfect ranking. Also, we are not using the mr^3 observation from the first stage to estimate μ_x, because we quantified only mr^2 of them and not all the mr^3 observations and this will reduce the cost of the sampling unit in the study.

4.2 Regression Estimator of μ_y

If $W_{i(i)k}$ and $Y_{i[i]k}$ are, respectively, the i *-th* smallest value of X (from the second stage of DRSS), and the corresponding value of *Y* obtained from the i*-th* sample in the k*-th* set, then from (3.1), we have

$$Y_{i(i)k} = \mu_y + \beta \left(W_{i(i)k} - \mu_x \right) + \varepsilon_{ik}, \quad i = 1, 2, ..., r, \; k = 1, 2, ..., m, \tag{4.1}$$

where, β is the model slope, $\beta = \rho \sigma_y / \sigma_x$, and ε_{iik} are random errors as in (3.1). Let $\bar{X}^*_{RSS}$ be the sample mean based on the *r* RSS samples of size $r^2 m$, i.e., $\bar{X}^*_{RSS} = \frac{1}{r^2 m} \sum_{l=1}^{r} \sum_{k=1}^{m} \sum_{i=1}^{r} X^1_{i(i)k}$.

Note that,

(1) $$E\left(\bar{X}^*_{RSS} \right) = \frac{1}{r^2 m} \sum_{l=1}^{r} \sum_{k=1}^{m} \sum_{i=1}^{r} E\left(X^1_{i(i)k} \right) = \mu_x \tag{4.2}$$

(2) $$Var\left(\bar{X}^*_{RSS} \right) = \frac{\sigma^2}{r^2 m} - \frac{1}{r^3 m} \Sigma \left(\mu_{x(i)} - \mu_x \right)^2 \tag{4.3}$$

(see Dell and Clutter (1972)). Under DRSS, the regression estimator of the population mean μ_y can be defined as

$$\bar{Y}_{RegD} = \bar{Y}_{DRSS} + \hat{\beta}_D \left(\bar{X}^*_{RSS} - \bar{W} \right) \tag{4.4}$$

where,

$$\hat{\beta}_D = \frac{\sum_{k=1}^{m} \sum_{i=1}^{r} \left(W_{i(i)k} - \bar{W} \right) Y_{i[i]k}}{\sum_{k=1}^{m} \sum_{i=1}^{r} \left(W_{i(i)k} - \bar{W} \right)^2}, \quad \bar{W} = \frac{1}{rm} \sum_{k=1}^{m} \sum_{i=1}^{r} W_{i(i)k}, \quad \bar{Y}_{DRSS} = \frac{1}{rm} \sum_{k=1}^{m} \sum_{i=1}^{r} Y_{i[i]k}$$

and $\bar{X}^*_{RSS}$ as above.

4.2.1. Properties of the estimator

Again, using the basic properties of conditional moments and the above results, the following theorem will be proved.

Theorem 4.1: Under (4.1) assumptions:

(1) $E\left(\overline{Y}_{\mathrm{Re}gD}\right)=\mu_y$.

(2) $Var\left(\overline{Y}_{\mathrm{Re}\,gD}\right)=\left(1-\rho^2\right)\frac{\sigma_y^2}{rm}\left[1+E\left(\frac{\left(\overline{Z}^*_{RSS}-\overline{Z}_W\right)^2}{S_z^2}\right)\right]+\frac{\beta^2}{r^2m}\sum_{i=1}^{r}\sigma^2_{x(i)}$,

where

$$Z^*_{(i)k}=\frac{W_{i(i)k}-\mu}{\sigma_x},\qquad S_z^2=\frac{1}{rm}\sum_{k=1}^{m}\sum_{i=1}^{r}\left(Z^*_{(i)k}-\overline{Z}_{RSS}\right)^2,$$

$$\overline{Z}_W=\frac{\overline{W}-\mu_x}{\sigma_x}\quad\text{and}\quad\overline{Z}^{\,*}_{RSS}=\frac{\overline{X}^*_{RSS}-\mu_x}{\sigma_x}.$$

Proof: The prove of this theorem and two required propositions are in the Appendix.

4.2.2 Performance of $\overline{Y}_{\mathrm{Re}gD}$ with Respect to Naive Estimators

From the previous results, the relative precision of $\overline{Y}_{\mathrm{Re}gD}$ with respect to the naive estimators of μ_y, $\overline{Y}_{RSS}$ and $\overline{Y}_{DRSS}$ using RSS and DRSS respectively are as follows:

$$\begin{aligned}RP\left(\overline{Y}_{\mathrm{Re}gD},\overline{Y}_{RSS}\right)&=\frac{\left(1/r^2m\right)\sum_{i=1}^{r}\sigma^2_{y[i]}}{\left(1-\rho^2\right)\frac{\sigma_y^2}{rm}\left[1+E\left[\frac{\left(\overline{Z}^*_{RSS}-\overline{Z}_W\right)^2}{S_z^2}\right]\right]+\frac{\beta^2}{r^3m}\sum_{i=1}^{r}\sigma^2_{x(i)}}\\&=\frac{\sum\sigma^2_{y(i)}}{r\left(1-\rho^2\right)\sigma_y^2\left[1+E\left[\frac{\left(\overline{Z}^*_{RSS}-\overline{Z}_W\right)^2}{S_z^2}\right]\right]+\frac{\beta^2}{r}\sum_{i=1}^{r}\sigma^2_{x(i)}},\end{aligned}\tag{4.5}$$

$$RP\left(\overline{Y}_{\mathrm{Re}gD},\overline{Y}_{DRSS}\right)=\frac{\sum\sigma^{**2}_{y(i)}}{r\left(1-\rho^2\right)\sigma_y^2\left[1+E\left[\frac{\left(\overline{Z}^*_{RSS}-\overline{Z}_W\right)^2}{S_z^2}\right]\right]+\frac{\beta^2}{r}\sum_{i=1}^{r}\sigma^2_{x(i)}},\tag{4.6}$$

Note that these relative precisions are based on the variances of the estimators.

4.2.3 Performance of $\overline{Y}_{\mathrm{Re}gD}$ with respect to other Regression Estimators

When the sample is drown from a standard bivariate normal population, the relative precision of $\overline{Y}_{\mathrm{Re}gD}$ relative to the two-phase (double sampling) regression estimator $\overline{Y}_{ds}$ based on SRS (see Sukhatme and Sukhatme, 1970) and to the two-phase regression estimator $\overline{Y}_{Rds}$ based on RSS (see Yu and Lam, 1997) will be respectively as follows:

$$RP\left(\bar{Y}_{RegD},\bar{Y}_{ds}\right)=\frac{\left(1-\rho^2\right)\frac{\sigma_y^2}{rm}\left(1+\frac{r-1}{r\left(r\ m-3\right)}\right)+\rho^2\frac{\sigma_y^2}{r^2m}}{\left(1-\rho^2\right)\frac{\sigma_y^2}{rm}\left[1+E\left[\frac{\left(\bar{Z}^*_{RSS}-\bar{Z}_W\right)^2}{S_z^2}\right]\right]+\frac{\beta^2}{r^3m}\sum_{i=1}^{r}\sigma^2_{x(i)}}, \tag{4.7}$$

$$RP\left(\bar{Y}_{RegD},\bar{Y}_{Rds}\right)=\frac{\left(1-\rho^2\right)\frac{\sigma_y^2}{rm}\left\{1+E_x\left[\frac{\left(\bar{Z}^*_{RSS}-\bar{Z}_W\right)^2}{S_z^2}\right]\right\}+\rho^2\frac{\sigma_y^2}{r^2m}}{\left(1-\rho^2\right)\frac{\sigma_y^2}{rm}\left\{1+E_x\left[\frac{\left(\bar{Z}^*_{RSS}-\bar{Z}_W\right)^2}{S_z^2}\right]\right\}+\frac{\beta^2}{r^3m}\sum_{i=1}^{r}\sigma^2_{x(i)}}, \tag{4.8}$$

(see section 3). Note that these relative precisions are based on the variances of the estimators. Since *t* is not easy to find, the values of the above expressions simulation is used to calculate them.

4.3 Simulation Study

4.3.1 Design of the Simulation

A computer simulation is conducted to study the efficiency of the regression estimator. Using SRS, RSS, and DRSS bivariate normal random samples where generated when μ_x=2, $\mu_y=4$, $\sigma_x=1$, $\sigma=1$ and $\rho=\pm[0.0-0.99]$. The performance of the regression estimators are investigated for *r* = 4, 5, 6, 7 and 8 and *m* =1, 4 and 8. Using 5000 replications, estimates of the means and the mean square errors for the regression estimators were computed.

The efficiency of the regression estimator is defined by $RE\left(\hat{R}_i,\hat{R}_j\right)=MSE\left(\hat{R}_i\right)/MSE\left(\hat{R}_j\right)$ where *i* and *j* represent any type of the above sampling methods. The results for the simulation are in Tables 1 and 2.

Table 1: The efficiency of $\bar{Y}_{RegD}$ with respect to the naive estimators based on RSS and DRSS

m	*r*	ρ=0.99		ρ =0.95		ρ =0.93		ρ =0.9		ρ =0.8	
		RSS	DRSS	RSS	DRSS	RSS	DRSS	RSS	DRSS	RSS	DRSS
1	4	4.70	2.22	2.81	1.40	2.39	1.16	2.04	0.96	1.21	0.62
	5	6.92	2.40	3.55	1.32	3.06	1.09	2.53	0.92	1.29	0.58
	6	9.28	2.40	4.20	1.27	3.35	1.04	2.85	0.77	1.68	0.47
	7	12.95	2.60	5.68	1.17	4.71	0.91	3.18	0.70	2.05	0.39
	8	17.30	2.78	6.97	1.07	5.95	0.81	4.01	0.63	2.32	0.39
4	4	11.97	2.30	7.35	1.50	6.10	1.26	4.52	1.03	3.52	0.66
	5	22.78	2.55	12.04	1.38	9.84	1.16	8.73	0.92	5.06	0.55
	6	38.33	2.61	19.58	1.23	15.72	1.01	11.93	0.82	6.79	0.46
	7	67.94	2.68	28.73	1.18	22.37	0.97	17.13	0.72	10.23	0.40
	8	99.25	2.68	39.02	1.12	31.18	0.80	23.86	0.64	13.42	0.36
8	4	26.98	2.27	18.43	1.45	15.17	1.23	12.97	1.03	8.23	0.66
	5	52.75	2.50	31.41	1.31	26.66	1.11	23.61	0.89	13.89	0.54
	6	120.97	2.62	60.15	1.31	44.88	1.05	34.71	0.80	18.58	0.47
	7	207.09	2.67	90.35	1.17	73.47	0.93	53.73	0.70	30.22	0.41
	8	336.46	2.73	124.39	1.05	99.13	0.83	74.28	0.60	44.91	0.35

Notice that ρ takes only high positive values because the regression estimator for the population mean is used only when the correlation between the two variables is high. Also,

negative values are not considered since, from (4.6) and (4.7), the relative precision depends on the absolute values of ρ and β since they are squared.

4.3.2 Results of the Simulation

Our simulation (Table 1) shows that the efficiency is affected by the value of ρ. The regression estimator based on DRSS is more efficient than naive estimator using RSS whenever the absolute value of the correlation coefficient between X and Y (ρ) is more that 0.40. Moreover, this efficiency is increasing as the set size or the cycle size increases. Also, the regression estimator based on DRSS is more efficient than the naive estimator using DRSS whenever $|\rho|>0.90$. However, when $|\rho|<0.98$, the efficiency decreased as the set size increased and increased otherwise. Moreover, in this case the efficiency is not affected by the cycle size.

Table 2 shows that the double sampling regression estimator using DRSS was always superior to the double sampling regression estimators using SRS and RSS. However, the efficiency was affected by the value of ρ. The efficiency increased by increasing the value of ρ. Also, the efficiency decreases with increasing the set or the cycle size for small values of ρ. However, $\overline{Y}_{RegD}$ was still found to be more efficient than using other sampling methods.

Table 2: The efficiency of $\overline{Y}_{RegD}$ with respect to the regression estimators based on SRS and RSS.

M	r	ρ=0.99		ρ =0.9		ρ =0.8	
		SRS	RSS	SRS	RSS	SRS	RSS
1	4	2.25	2.18	1.94	1.64	1.85	1.48
	5	2.47	3.43	1.72	1.90	1.87	1.33
	6	2.35	2.60	1.43	1.27	1.49	1.27
	7	2.80	2.74	1.56	1.44	1.37	1.23
	8	2.93	2.89	1.51	1.40	1.32	1.21
4	4	2.14	2.13	1.46	1.45	1.26	1.24
	5	2.40	2.39	1.45	1.44	1.24	1.23
	6	2.58	2.57	1.23	1.22	1.22	1.20
	7	2.73	2.73	1.41	1.38	1.20	1.18
	8	2.89	2.87	1.38	1.36	1.19	1.17
8	4	2.13	2.13	1.44	1.43	1.24	1.23
	5	2.39	2.39	1.43	1.43	1.45	1.21
	6	2.57	2.60	1.22	1.21	1.20	1.19
	7	2.73	2.73	1.39	1.39	1.19	1.18
	8	2.87	2.87	1.36	1.36	1.17	1.16

5. Applications to Real Data Set

We illustrate the double ranked set sample mean estimation procedure using a real data set which consists of the height (Y) and the diameter (X) at breast height of 399 trees. See Platt *et al*. (1988) for a detailed description of the data set. The summary statistics for the data are reported in Table 3. Note that the correlation coefficient ρ = 0.908.

Table 3: Summary Statistics of trees data.

Variable	Mean	Variance
Height (X) in feet	52.36	325.14
Diameter (X) in cm	20.84	310.11
Population size N = 399 and the correlation coefficient between X and Y is ρ = 0.908.		

Using a set size *r=3* and the cycle size *m=3*, we draw bivariate SRS and DRSS, of size 9. Table 4 contains all the above proposed estimators and their estimated variances using the drawn samples.

Table 4: Results from the drawn samples

Sample	Naïve Estimator of Height (Y) in feet	Estimated Variance	Regression Estimator	Estimated Variance
SRS	52.49	408.88	51.19	176.04
DRSS	52.61	182.25	52.22	125.10

Although, Table 3 confirms our simulation results. It should be emphasized that the example is used as an illustration of the applicability of our proposed estimators.

6. Conclusions

In conclusion DRSS regression estimator is to be used to improve the population mean estimation whenever DRSS is possible to be conducted.

References

AL-SALEH, M.F. and AL-KADIRI, M.A. 2000. Double ranked set sampling. *Statistics and probability letters,* **48(2):** 205-212.

DELL, T.R. and CLUTTER, J.L. 1972. Ranked set sampling theory with order statistics background. *Biometrics,* **28:** 545-555.

KAUR, A., PATIL, G.P., SINHA, A.K. and TAILLIE, C. 1995. Ranked set sampling: an annotated bibliography. *Environmental and Ecological Statistics,* **2:** 25-54.

MCINTYRE, G.A. 1952. A method for unbiased selective sampling using ranked set. *Australian Journal of Agricultural Research*, **3:** 385-390.

PATIL, G.P., SINHA, A.K. and TAILLIE, C. 1999. Ranked set sampling: A bibliography. *Environ. Ecolog. Statist*., **6:** 91-98.

PLATT, W. J., EVANS, G.W., and RATHBUN, S.L. 1988. The population dynamics of a long-lived conifer. *The Amer. Naturalist,* **131:** 391-525.

SAMAWI, H.M. 2002. On double extreme ranked set sample with application to regression estimator. *Metron,* **LX n. 1-2:** 53-66.

SAMAWI, H.M. AHMED, M.S. and ABU DAYYEH, W. 1996. Estimating the population mean using extreme ranked set sampling. *Biometrical Journal,* **38 (5):** 577-586.

SAMAWI, H.M. and AL-SAGEER, O.A. 2001. On the estimation of the distribution function using extreme and median ranked set sampling. *Biometrical Journal,* **43 (3):** 357-373.

SAMAWI, H.M., and MUTTLAK, H.A. 1996. Estimation of ratio using ranked set sampling. *Biometrical Journal,* **38 (6):** 753-764.

SUKHATME, P.V. and SUKHATME, B.V. 1970. *Sampling theory of surveys with applications*. Ames: Iowa state university Press.

TAKAHASI, K. and WAKIMOTO, K. 1968. On unbiased estimates of the population mean based on the stratified sampling by means of ordering. *Ann. Inst. Statist. Math*., **20:** 1-31.

YU, L.H. and LAM, K. 1997. Regression estimator in ranked set sampling. *Biometrics*, **53:** 1070-1080.

Appendix

Proposition 1. Under (4.1) $E\left(\hat{\beta}_D\right) = \beta$.

Proof : Let

$$C_{ik} = \left(W_{(i)k} - \bar{W}\right) / \sum\sum \left(W_{(i)k} - \bar{W}\right)^2$$

then

$$E(\hat{\beta}_D) = E_x\left[E\left(\sum\sum C_{ik}\, Y_{i[i]k} \middle| X\right)\right] = E_x\left[\sum\sum C_{ik} E\left(Y_{i[i]k} \middle| X\right)\right]$$
$$= E_x\left[\sum\sum C_{ik} E\left(Y_{i[i]} \middle| X\right)\right] .$$

Note that since $E\left(Y_{i[i]}\right) = \beta_0 + \beta\ W_{i(i)k}$ and $\sum\sum C_{ik} = 0$, then

$$E\left(\hat{\beta}_D\right) = E_w\left(\sum_k \sum_i C_{ik}\left(\beta_0 + \beta\ W_{i(i)k}\right)\right)$$
$$= E_w\left(\beta_0 \sum\sum C_{ik} + \beta \sum\sum C_{ik} W_{i(i)k}\right) = \beta ,$$

where $\beta_0 = \mu_y - \beta\mu_x$.

Proposition 2. $Var\left(\hat{\beta}_D \middle| X\right) = \sigma_e^2 / \sum\sum \left(W_{i(i)k} - \bar{W}\right)^2$.

Proof:

$$Var\left(\hat{\beta}_D \middle| X\right) = Var\left(\frac{\sum\sum \left(W_{i(i)k} - \bar{W}\right) Y_{i[i]k}}{\sum\sum \left(W_{i(i)k} - \bar{W}\right)^2} \middle| X\right)$$
$$= Var\left(\sum_i \sum_k C_{ik}\, Y_{i[i]k} \middle| X\right) = \sum\sum C_{ik}^2 Var\left(Y_{i[i]k} \middle| X\right) .$$

Using $Y_{i[i]} = \beta_0 + \beta\ W_{i(i)k} + \varepsilon_i$, and since

$$E\left(Y_{i[i]} \middle| X\right) = \beta_0 + \beta\ W_{i(i)} \qquad Var\left(Y_{i[i]} \middle| X\right) = \sigma_e^2 ,$$

then

$$Var\left(\hat{\beta}_D \middle| X\right) = \sum_k \sum_i \left[\frac{\left(W_{i(i)k} - \bar{W}\right)}{\sum_k \sum_i \left(W_{i(i)k} - \bar{W}\right)^2}\right]^2 \sigma_e^2 = \frac{\sigma_e^2}{\sum\sum \left(W_{i(i)k} - \bar{W}\right)^2} .$$

Next we prove Theorem 4.1

Proof of Theorem 4.1:

(1) $E\left(\bar{Y}_{\mathrm{Reg}D}\right) = E_x\left(E_y\left(\bar{Y}_{\mathrm{Reg}D} \middle| X\right)\right)$

$$E_y\left(\bar{Y}_{\mathrm{Reg}D} \middle| X\right) = E_y\left(\bar{Y}_{DRSS} + \hat{\beta}_D\left(\bar{X}^*_{RSS} - \bar{W}\right) \middle| X\right)$$
$$= E_y\left(\frac{1}{rm}\sum_m \sum_r Y_{i[i]k} + \hat{\beta}_D\left(\bar{X}^*_{RSS} - \bar{W}\right) \middle| X\right)$$

Using (4.1), we have

$$E_y\left(\bar{Y}_{RegD}\middle|X\right)=E_y\left(\frac{1}{rm}\sum_{k1}^{m}\sum_{i=1}^{r}\left(\mu_y+\hat{\beta}_D\left(W_{(i)k}-\mu_x\right)\right)+\hat{\beta}_D\left(\bar{X}^*_{RSS}-\bar{W}\right)\middle|X\right)$$

$$=\mu_y+\beta\left(\bar{W}-\mu_x\right)+\beta\left(\bar{X}^*_{RSS}-\bar{W}\right).$$

Therefore,

$$E\left(\bar{Y}_{RegD}\right)=E_x\left(\mu_y+\beta\left(\bar{X}^*_{RSS}-\mu_x\right)\right)=\mu_y.$$

Hence $\bar{Y}_{RegD}$ is an unbiased estimator of μ_y.

(2) $$Var\left(\bar{Y}_{RegD}\right)=E_x\left[Var_y\left(\bar{Y}_{RegD}\middle|X\right)\right]+Var_x\left[E\left(\bar{Y}_{RegD}\middle|X\right)\right]$$

$$Var_x\left[E_y\left(\bar{Y}_{RegD}\middle|X\right)\right]=Var_x\left[E_y\left(\bar{Y}_{DRSS}+\hat{\beta}_D\left(\bar{X}^*_{RSS}-\bar{W}\right)\middle|X\right)\right].$$

From (1), we have $E\left(\bar{Y}_{RegD}\middle|X\right)=\mu_y+\beta\left(\bar{X}^*_{RSS}-\mu_x\right)$,
and then

$$\begin{aligned}Var_x\left(E_y\left(\bar{Y}_{RegD}\middle|X\right)\right) &= Var_x\left(\mu_y+\beta\left(\bar{X}^*_{RSS}-\mu_x\right)\right)\\ &=Var_x\left[\beta\bar{X}^*_{RSS}\right]=\beta^2Var_x\left(\bar{X}^*_{RSS}\right)\\ &=\beta^2\frac{1}{r^3m}\sum_{i=1}^{r}\sigma^2_{x(i)}.\end{aligned}$$

Also,

$$\begin{aligned}E_x\left[Var_y\left(\bar{Y}_{RegD}X\right)\right]&=E_x\left[Var_y\left(\bar{Y}_{DRSS}+\hat{\beta}_D\left(\bar{X}^*_{RSS}-\bar{W}\right)\middle|X\right)\right]\\ &=E_x\left[Var_y\left(\bar{Y}_{DRSS}\middle|X\right)+\left(\bar{X}^*_{RSS}-\bar{W}\right)^2Var_y\left(\hat{\beta}_D\middle|X\right)\right.\\ &\left.+2Cov\left(\bar{Y}_{DRSS},\ \hat{\beta}_D\left(\bar{X}^*_{RSS}-\bar{W}\right)\middle|X\right)\right],\end{aligned}$$

but,

$$\begin{aligned}Cov\left(\bar{Y}_{DRSS},\hat{\beta}_D\left(\bar{X}^*_{RSS}-\bar{W}\right)\middle|X\right)&=Cov\left(\frac{1}{rm}\sum_k\sum_i Y_{i[i]k}\ ,\ \sum_k\sum_i C_{ik}Y_{i[i]k}\left(\bar{X}^*_{RSS}-\bar{W}\right)\middle|X\right)\\ &=\frac{1}{rm}\left(\bar{X}^*_{RSS}-\bar{W}\right)\sum\sum C_{ik}Var\left(Y_{i[i]k}\middle|X\right)\\ &=\frac{1}{rm}\left(\bar{X}^*_{RSS}-\bar{W}\right)\sigma_e^2\sum\sum C_{ik}=0.\end{aligned}$$

Then,

$$E_x\left[Var_y\left(\bar{Y}_{RegD}\middle|X\right)\right]=E_x\left(Var_y\left(\bar{Y}_{DRSS}\middle|X\right)\right)+E_x\left(\left(\bar{X}^*_{RSS}-\bar{W}\right)^2Var_y\left(\hat{\beta}_D\middle|X\right)\right)$$

$$=E_x\left(\frac{1}{(rm)^2}\sum_{k=1}^{m}\sum_{i=1}^{r}\sigma_e^2\right)+E_x\left(\left(\bar{X}^*_{RSS}-\bar{W}\right)^2\frac{\sigma_e^2}{\sum_{k=1}^{m}\sum_{i=1}^{r}\sigma_e^2\left(W_{i[i]k}-\bar{W}\right)^2}\right),$$

clearly this implies that,

$$= \left(\frac{1}{rm} \left(\sigma_e^2 \right) \right) + E_x \left(\sigma_e^2 \frac{\left(\left(\bar{X}^*_{RSS} - \mu_x \right) - \left(\bar{W} - \mu_x \right) \right)^2 \Big/ \sigma_x^2}{\sum\sum \left(\left(W_{i[i]k} - \mu_x \right) - \left(\bar{W} - \mu_x \right) \right)^2 \Big/ \sigma_x^2} \right)$$

$$= \left(1 - \rho^2 \right) \sigma_y^2 \left[\frac{1}{rm} + E_x \left[\frac{\left(\bar{X}^*_{RSS} - \bar{Z}_W \right)^2}{(rm) \sum_k \sum_i \left(Z^*_{(i)k} - \bar{Z}_W \right)^2} \right] \right]$$

$$= \left(1 - \rho^2 \right) \sigma_y^2 \left[\frac{1}{rm} + E_x \left[\frac{\left(\bar{X}^*_{RSS} - \bar{Z}_W \right)^2}{(rm) S_z^2} \right] \right]$$

$$= \left(1 - \rho^2 \right) \frac{\sigma_y^2}{rm} \left[1 + E_x \left[\frac{\left(\bar{X}^*_{RSS} - \bar{Z}_W \right)}{S_z^2} \right] \right],$$

where

$$Z^*_{(i)k} = \frac{W_{i(i)k} - \mu_x}{\sigma_x} \quad , \quad \bar{Z}_W = \frac{\bar{W} - \mu_x}{\sigma_x} \quad ,$$

$$S_z^2 = \frac{1}{rm} \sum_{k=1}^{m} \sum_{i=1}^{r} \left(Z^*_{(i)k} - \bar{Z}_W \right)^2 \quad , \quad \bar{X}^*_{RSS} = \frac{\bar{X}^*_{RSS} - \mu_x}{\sigma_x},$$

therefore,

$$Var\left(\bar{Y}_{\mathrm{Re}gD} \right) = \left(1 - \rho^2 \right) \frac{\sigma_y^2}{rm} \left[1 + E_x \left[\frac{\left(\bar{X}^*_{RSS} - \bar{Z}_W \right)}{S_z^2} \right] \right] + \beta^2 \frac{1}{r^2 m} \sum_{i=1}^{r} \sigma^2_{x(i)} .$$

Received 8 November 2001
Accepted 31 October 2002