SQU Journal For Science, 9 (2004) 67-86

On Regression Estimators Using Extreme Ranked Set Samples

Hani M. Samawi*, Ahmed Y.A. Al-Samarraie* and Obaid M. Al-Saidy**

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

حول تقديرات الإنحدار باستخدام العينات القصوى المرتبة

هاني سماوي، أحمد السامرائي و عبيد السعيدي

خلاصة : أسلوب الانحدار قد استخدم لتقدير الوسط الحسابي للمجتمع للمتغير المعتمد (ص) في حالتين. في الحالة الأولى عندما يكون الوسط الحسابي للمتغير المساعد المستقل (س) معلوماً وفي الحالة الثانية عندما يكون غير معلوم . أيضا في الحالة الثانية لقد استخدمنا طريقة العينة المزدوجة لتقدير الوسط الحسابي للمتغير المستقل (س). لقد تحققنا من أداء الطريقتين باستخدام العينات القصوى المرتبة كما جاء في بحث سماوي وزملاءه (1996). لقد تم عرض الناحية النظرية والرقمية بواسطة المحاكاة والتطبيق في هذا البحث. ولقد أظهرت النتائج أنه في حالة التوزيعات المتماثلة فان طريقة استخدام العينات القصوى المرتبة لتقريرات الانحدار هي اكثر فعالية من طريقة العينات المرتبة العادية والعينات البسيطة.

ABSTRACT: Regression is used to estimate the population mean of the response variable, Y , in the two cases where the population mean of the concomitant (auxiliary) variable, X , is known and where it is unknown. In the latter case, a double sampling method is used to estimate the population mean of the concomitant variable. We invesitagate the performance of the two methods using extreme ranked set sampling (ERSS), as discussed by Samawi *et al*. (1996). Theoretical and Monte Carlo evaluation results as well as an illustration using actual data are presented. The results show that if the underlying joint distribution of X and Y is symmetric, then using ERSS to obtain regression estimates is more efficient than using ranked set sampling (RSS) or simple random sampling (SRS).

KEYWORDS: Extreme ranked set sample, ranked set sample, relative efficiency, regression estimators, two-phase sampling.

1. Introduction

In many experimental situations the response variable Y is related to a non-stochastic concomitant variable, X . For instance, let Y be the Bilirubin level in jaundice babies who stay in neonatal intensive

care and let X be the weight of the baby at birth. By obtaining simultaneous observations on X and Y, we can use information contained in the X-measurements to estimate the mean value of Y. This can be done by using either ratio estimation or regression estimation.

Herein, we are interested in the regression estimation method used to obtain increased precision in estimating the population means or totals of the variable of interest, Y, by taking advantage of its correlation with the auxiliary variable X. The two cases where the mean, μ_x, of X is known and where it is unknown are considered.

In many cases the sampling units in a study are easier ranked than actually quantified. McIntyre (1952) proposed to use the mean of n units obtained from a ranked set sample (RSS) to estimate a population mean. Patil *et al*. (1993) compared the precision of ranked set sampling with the regression estimator. They showed that using RSS is superior to regression estimator under SRS in most of the cases. Yu and Lam (1997) used the RSS regression estimation method to estimate the population mean and showed that using RSS provides a more efficient estimator than using SRS. For more details on RSS see, for example, Kaur *et al*. (1995) and Patil *et al*. (1999). Samawi *et al*. (1996) investigated the use of extreme ranked set sampling (ERSS) in reducing the ranking error and in improving the precision in estimating the population mean in the case of a symmetric underlying distribution. They showed that if the underlying distribution is the uniform distribution, then the highest magnitude of the relative savings occur when only the extreme ordered units are measured with equal proportion. However, in the case of other unimodal symmetric distributions the highest gain is achieved when the units possessing the middle rank are measured. For this reason, Yanagawa and Chen (1980) did not consider the uniform distribution while investigating various symmetric distributions to develop a better ranked set sample estimator of the population mean.

As in Samawi *et al*. (1996) we obtain an extreme rank set sample by first choosing *r* independent sets, each of which contains r bivariate elements drawn randomly from an infinite population. Rank the elements in each set with respect to one of the variables Y or X. Suppose that the ranking is done on the variable X. From the first set an actual measurement is taken of the X element with the smallest rank, together with the value of Y associated with this smallest element of X. From the second set an actual measurement is taken of the element with the largest rank of X, together with the associated Y value. From the third set an actual measurement is taken of the element with the smallest rank of X, together with the associated Y value, and so on. In this way we obtain the first $r-1$ measured elements using the first $r-1$ sets, together with the associated values of the Y variable. The choice of the $r-\text{th}$ element from the $r-\text{th}$ (i.e., the last) set depends on whether r is even or odd :

(a) If r is even the largest ranked X element is measured, together with the value of the associated variable Y. ERSSa will denote such a sample.

(b) If r is odd we measure the median of X, together with the value of variable Y associated with the median of X. ERSSb will denote such a sample.

The cycle may be repeated *m* times until $n = rm$ bivariate elements have been measured.

In this paper we propose to use ERSS to improve the precision of the two methods of regression estimation. We study the properties of these estimators and compare them under different settings. In Section 2, we obtain the regression estimator of the mean of Y using extreme ranked set sampling when μ_X is known. The mean and variance of the estimator are derived. Comparisons between the various estimators are discussed in terms of efficiencies. In Section 3, we obtain the regression estimator using extreme ranked set sampling when μ_X is unknown using a double sampling method. Again, we derive the mean and variance of the estimator and some comparisons between the various estimators are discussed in terms of efficiencies. An illustration of the methods using real data about the Bilirubin level in jaundice babies is given in Section 4.

2. Regression Estimators when μ_x is Known

Like ratio estimation, linear regression estimation of the mean is designed to increase the precision of the estimator by using an auxiliary variable X that is correlated with Y. When the relationship between Y and X is examined, it may be found that although the relation is approximately linear, the line does not go through the origin. This suggests that an estimator based on the linear regression of Y on X is better than an estimator that is based on the ratio of the two variables.

2.1 Regression Estimator Using SRS

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

$$Y_i = \mu_y + \beta(X_i - \mu_x) + \varepsilon_i \tag{2.1}$$

where μ_x and μ_y are the means of X and Y respectively, and for a fixed X_i, the ε_i's, $i = 1, 2, ..., n$ are *i.i.d.* with mean zero and variance $\sigma_\varepsilon^2 = \sigma_y^2(1-\rho^2)$, where ρ is the correlation coefficient between X and Y.

When the population mean μ_x is known, the regression estimator of the mean of Y is given by:

$$\overline{Y}_{reg} = \overline{Y} + \hat{\beta}(\mu_x - \overline{X}) \tag{2.2}$$

where $\overline{X} = \frac{1}{n}\sum X_i$, $\overline{Y} = \frac{1}{n}\sum Y_i$, $\hat{\beta} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$, and $n = m\,r$.

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

$$Var(\overline{Y}_{reg}) = \frac{\sigma_y^2}{n}(1-\rho^2)\left(1 + \frac{1}{n-3}\right) \tag{2.3}$$

(see Tikkiwal (1960) or Sukhatme and Sukhatme, 1970.) However, if the assumption of the linear relationship in (2.1) is invalid, then the SRS regression estimator in (2.2) is in general a biased estimator of μ_y.

2.2 Regression Estimator Using RSS

Consider a bivariate RSS where the relationship between $Y_{[i]k}$ and $X_{(i)k}$ is

$$Y_{[i]k} = \mu_y + \beta(X_{(i)k} - \mu_x) + \varepsilon_{[i]k}, \quad i = 1, 2, ..., r \text{ and } k = 1, 2, ..., m. \tag{2.4}$$

Then the regression estimator $\overline{Y}_{Reg}$ based on RSS as in Yu and Lam (1997) is given by

$$\overline{Y}_{Reg} = \overline{Y}_{RSS} + \hat{\beta}(\mu_x - \overline{X}_{RSS}). \tag{2.5}$$

Using basic properties of conditional moments, Yu and Lam (1997) showed that under (2.4), $\bar{Y}_{Reg}$ is an unbiased estimator of μ_y and its variance is

$$Var\left(\bar{Y}_{Reg}\right)=\frac{\sigma_y^2}{n}\left(1-\rho^2\right)\left(1+E\left(\frac{\bar{Z}_{RSS}^2}{S_{ZR}^2}\right)\right), \tag{2.6}$$

where, $\bar{Z}_{RSS}=\frac{1}{mr}\sum_i\sum_i Z_{(i)k}$, $Z_{(i)k}=\frac{X_{(i)k}-\mu_x}{\sigma_x}$, and $S_{ZR}^2=\frac{1}{rm}\sum_k\sum_i\left(Z_{(i)k}-\bar{Z}_{RSS}\right)^2$.

Again, if the assumption of the linear relationship is invalid, the RSS regression estimator in (2.5) is in general a biased estimator for μ_y.

2.3 Regression Estimator Using ERSS

Assuming that both variables, X and Y, have symmetric underlying distributions, let $X_{(i)jk}$, $Y_{[i]jk}$ be respectively, the $i-th$ smallest value of X and the corresponding value of Y obtained from the $j-th$ sample and the $k-th$ cycle. Then regressing $Y_{[i]jk}$ on $X_{(i)jk}$ we have

$$Y_{[i]jk}=\mu_y+\rho\frac{\sigma_y}{\sigma_x}\left(X_{(i)jk}-\mu_X\right)+\varepsilon_{ijk}, \tag{2.7}$$

where $i=1,\ r;\ j=1,2,...,r/2$ and $k=1,2,...,m,$ when r is even, $\left(i=1,r,\ j=1,2,...,\ \frac{r-1}{2}\right.$ and $\left.k=1,2,...,m\right)$ when r is odd, and ε_{ijk} has the same distributional assumptions as in (2.1). In what follows we discus in details the case when r is even. The case when *r* is odd is similar and it will only be presented in the numerical results.

When the population mean μ_x is known, we have the difference estimator,

$$\bar{Y}_{Da}=\bar{Y}_{ERSSa}+\beta\left(\mu_x-\bar{X}_{ERSSa}\right) \tag{2.8}$$

where, $\bar{Y}_{ERSSa}=\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{r/2}\left(Y_{[1](j-1)k}+Y_{[r]2jk}\right)$, $\bar{X}_{ERSSa}=\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{r/2}\left(X_{(1)(j-1)k}+X_{(r)2jk}\right)$, and β is a constant to be determined. Under the assumption of symmetric underlying distribution functions of X and Y, Y_{ERssa} and X_{ERssa} are unbiased estimators for μ_y and μ_x respectively, see Samawi *et al*. (1996). Therefore, it can easily be shown that $\bar{Y}_{Da}$ is an unbiased estimator of μ_y. Furthermore, $\operatorname{var}\left(\bar{Y}_{Da}\right)=\beta^2\operatorname{var}\left(\bar{X}_{ERSSa}\right)-2\beta\rho\frac{\sigma_Y}{\sigma_X}\operatorname{var}\left(\bar{X}_{ERSSa}\right)+\operatorname{var}\left(\bar{Y}_{ERSSa}\right)$

where,

$\mathrm{var}\left(\bar{X}_{ERSSa}\right)=\frac{\sigma^2_{X_{(1)}}}{n}$, $\mathrm{var}\left(\bar{Y}_{ERSSa}\right)=\frac{\sigma^2_{Y_{[1]}}}{n}$, $\mathrm{var}\left(X_{(1)}\right)=\sigma^2_{X_{(1)}}$, $\mathrm{var}\left(Y_{[1]}\right)=\sigma^2_{Y_{[1]}}$ and $n=r\,m$. Note that by the symmetry of the underlying distributions, $\sigma^2_{X_{(1)}}=\sigma^2_{X_{(n)}}$ and $\sigma^2_{Y_{[1]}}=\sigma^2_{Y_{[n]}}$, see Samawi *et al*. (1996).

Since for any value of β, $\bar{Y}_{D_a}$ is an unbiased estimator of μ_y the optimal value of β can be obtained by minimizing the variance of $\bar{Y}_{D_a}$. Doing so gives $\beta^*=\rho\frac{\sigma_Y}{\sigma_X}$ as the optimal value of β. However, β^* is unknown but can be estimated by

$$\beta^*_a=\frac{\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}\left(X_{(i)jk}-\bar{X}_{ERSSa}\right)\left(Y_{[i]jk}-\bar{Y}_{ERSSa}\right)}{\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}\left(X_{(i)jk}-\bar{X}_{ERSSa}\right)^2}=\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}C_{(i)jk}Y_{[i]jk},$$

$i=1$ and r, where

$$C_{(i)jk}=\frac{\left(X_{(i)jk}-\bar{X}_{ERSSa}\right)}{\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}\left(X_{(i)jk}-\bar{X}_{ERSSa}\right)^2}.$$

Now, define the ERRSa regression estimator for μ_y as

$$\bar{Y}_{Ereg}=\bar{Y}_{ERSSa}+\hat{\beta}_a\left(\mu_x-\bar{X}_{ERSSa}\right). \tag{2.9}$$

Then using basic properties of conditional moments, we have the following theorem:

Theorem 2.1*:* Under (2.2) and assuming that the underlying marginals distributions of X and of Y are symmetric, the regression estimator of μ_y as defined in (2.9) has the following properties:

(a) $\mathrm{E}\left(\bar{Y}_{Ereg}\right)=\mu_Y$

(b) $Var\left(\bar{Y}_{Ereg}\right)=\frac{\sigma^2_Y}{n}\left(1-\rho^2\right)\left[1+E\left(\frac{\bar{Z}^2_{ERSS}}{S^2_{Z_E}}\right)\right]$

where,

$$\bar{Z}_{ERSS}=\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{r/2}\left(Z_{(1)2j-1k}+Z_{(r)2jk}\right),$$

and

$$S^2_{Z_E}=\frac{1}{n}\sum_{k=1}^{m}\left[\sum_{j=1}^{r/2}\left(Z_{(1)2j-1k}-\bar{Z}_{ERSS}\right)^2+\sum_{j=1}^{r/2}\left(Z_{(r)2jk}-\bar{Z}_{ERSS}\right)^2\right],$$

with

$$Z_{(i)jk} = \frac{X_{(i)jk} - \mu_x}{\sigma_X}, \quad i = 1,\ r; j = 1,..., r/2 \text{ and } k = 1,..., m,$$

Proof: To prove Theorem 2.1, we first show that

(1) $E\left(\hat{\beta}_a \mid X\right) = \beta$ and $Var\left(\hat{\beta}_a \mid X\right) = \dfrac{\sigma_e^2}{\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}\left(X_{(i)jk} - \bar{X}_{ERSSa}\right)^2}$

(2)

Proof of (1): From the definition of $\hat{\beta}_a = \sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i} C_{(i)jk} Y_{[i]jk}$, we have that $E_x[E_y(\hat{\beta}_a \mid X)] = E_x[E_y(\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i} C_{(i)jk} Y_{[i]jk} \mid X)]$. Since $E_y\left(Y_{[i]jk} \mid X\right) = \mu_y + \beta(X_{(i)jk} - \mu_x)$, $\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i} C_{(i)jk} = 0$ and $\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i} C_{(i)jk} X_{(i)jk} = 1$,

then it is clear that $E_x(E_y(\hat{\beta}_a \mid X)) = E_x[\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i} C_{(i)jk}(\mu_y + \beta(X_{(i)jk} - \mu_x))] = E_x[0 + \beta\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i} C_{(i)jk} X_{(i)jk} - 0]$

$$= E_x(\beta) = \beta.$$

Proof of (2): Similarly, since $Var_y\left(Y_{[i]jk} \mid X\right) = \sigma_e^2$ and $\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i} C_{(i)jk}^2 = \dfrac{1}{\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}(X_{(i)jk} - \bar{X}_{ERSSa})^2}$, then

$Var\left(\hat{\beta}_a \mid X\right) = \sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i} C_{(i)jk}^2\ Var(Y_{[i]jk} \mid X) = \dfrac{\sigma_e^2}{\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}(X_{(i)jk} - \bar{X}_{ERSSa})^2}.$

Proof of Theorem 2.1 (a): Using (2.7) and the proof of (1), we have that

$$E_y\left(\bar{Y}_{Ereg} \middle| X\right) = E_y\left(\bar{Y}_{ERSSa} + \hat{\beta}_a\left(\mu_x - \bar{X}_{ERSSa}\right) \middle| X\right)$$

$$= E_y\left(\frac{1}{rm}\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i} Y_{[i]jk} + \hat{\beta}_a\left(\mu_x - \bar{X}_{ERSSa}\right) \middle| X\right),$$

$$= E_y\left(\frac{1}{rm}\sum_{k=1}^{m}\sum_{j=1}^{k}\sum_{i}\left(\mu_y + \beta\left(X_{(i)k} - \mu_x\right)\right) + \hat{\beta}_a\left(\mu_x - \bar{X}_{ERSSa}\right) \middle| X\right)$$

$$= \mu_y + \beta\left(\bar{X}_{ERSSa} - \mu_x\right) + \beta\left(\mu_x - \bar{X}_{ERSSa}\right).$$

$$E\left(\bar{Y}_{Ereg}\right) = E_x\left(E_y\left(\bar{Y}_{Ereg} \middle| X\right)\right)$$

Therefore, $E\left(\bar{Y}_{Ereg}\right)=\mu_y,$ and hence $\bar{Y}_{Ereg}$ is an unbiased estimator of μ_y.

Proof of Theorem 2.1 (b): Using properties of conditional moments,

$$Var\left(\bar{Y}_{Ereg}\right)=E_x\left[Var_y\left(\bar{Y}_{Ereg}\middle|X\right)\right]+Var_x\left[E_y\left(\bar{Y}_{Ereg}\middle|X\right)\right].$$

First note that, $Var_x\left[E_y\left(\bar{Y}_{Ereg}\middle|X\right)\right]=Var_x\left[E_y\left(\bar{Y}_{ERSSa}+\hat{\beta}_a\left(\mu_x-\bar{X}_{ERSSa}\right)\middle|X\right)\right]$ and from the proof of part (a), $E\left(\bar{Y}_{Ereg}\middle|X\right)=\mu_y$ $Var_x\left[E_y\left(\bar{Y}_{Ereg}\middle|X\right)\right]=Var_x\left[\mu_y\right]=0$, then

$$E_x\left[Var_y\left(\bar{Y}_{Ereg}\middle|X\right)\right]=E_x\left[Var_y\left(\bar{Y}_{ERSSa}+\hat{\beta}_a\left(\mu_x-\bar{X}_{ERSSa}\right)\middle|X\right)\right].$$

Also,

$$=E_x\left[Var_y\left(\bar{Y}_{ERSSa}\middle|X\right)+\left(\mu_x-\bar{X}_{ERSSa}\right)^2Var_y\left(\hat{\beta}_a\middle|X\right)\right.$$

$$\left.+2Cov\left(\bar{Y}_{ERSSa},\ \hat{\beta}_a\left(\mu_x-\bar{X}_{ERSSa}\right)\ \middle|X\right)\right],$$

$$Cov\left(\bar{Y}_{ERSSa}\ ,\hat{\beta}_a\left(\mu_x-\bar{X}_{ERSSa}\right)\middle|X\right)=$$

$$\left(\mu_x-\bar{X}_{ERSSa}\right)Cov\left(\frac{1}{rm}\sum_{k=1}^{m}\sum_{\mathrm{j}=1}^{\mathrm{r/2}}\sum_{\mathrm{i}}Y_{[i]jk}\ \ ,\sum_{k=1}^{m}\sum_{\mathrm{j}=1}^{\mathrm{r/2}}\sum_{\mathrm{i}}C_{(i)jk}Y_{[i]jk}\middle|X\right)$$

but,

$$=\frac{1}{rm}\left(\mu_x-\bar{X}_{ERSSa}\right)\ \sum_{k=1}^{m}\sum_{\mathrm{j}=1}^{\mathrm{r/2}}\sum_{\mathrm{i}}C_{(i)jk}Var(Y_{[i]jk}\mid X),$$

$$=\frac{1}{rm}\left(\mu_x-\bar{X}_{ERSSa}\right)\ \sigma_e^2\sum_{k=1}^{m}\sum_{\mathrm{j}=1}^{\mathrm{r/2}}\sum_{\mathrm{i}}C_{(i)jk},=0,$$

therefore,

$$E_x\left[Var_y\left(\bar{Y}_{Ereg}\middle|X\right)\right]=E_x\left(Var_y\left(\bar{Y}_{ERSSa}\middle|X\right)\right)+E_x\left(\left(\mu_x-\bar{X}_{ERSSa}\right)^2Var_y\left(\hat{\beta}_a\middle|X\right)\right),$$

$$E_x\left[Var_y\left(\bar{Y}_{Ereg}\middle|X\right)\right]=E_x\left(\frac{1}{(rm)^2}\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}Var_y(Y_{[i]jk}\right)$$

and from the proof of (2) above,

$$+E_x\left(\left(\mu_x-\bar{X}_{ERSSa}\right)^2\frac{\sigma_e^2}{\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}(X_{(i)jk}-\bar{X}_{ERSSa})^2}\right).$$

$$E_x\left[Var_y\left(\bar{Y}_{Ereg}\middle|X\right)\right] = E_x\left(\frac{1}{(rm)^2}\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}\sigma_e^2\right)$$

Clearly this implies that,

$$+\sigma_e^2 E_x\left(\frac{\left(\bar{X}_{ERSSa}-\mu_x\right)^2/\sigma_x^2}{\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}\{(X_{(i)jk}-\mu_x)-(\bar{X}_{ERSSa}-\mu_x)\}^2/\sigma_x^2}\right)$$

and hence,

$$Var\left(\bar{Y}_{Ereg}\right)=\frac{\sigma_y^2}{n}\left(1-\rho^2\right)\left[1+E_x\left[\frac{\bar{Z}_{ERSS}^2}{S_{z_E}^2}\right]\right].$$

2.4 Comparison with Naïve Estimators

Using Theorem (2.1) and the above results, the relative precision of the ERSS regression estimator, $\bar{Y}_{Ereg}$, relative to the ERSS naive estimator, $\bar{Y}_{ERSS}$, is

$$RP\left(\bar{Y}_{Ereg},\bar{Y}_{ERSS}\right)=\frac{Var\left(\bar{Y}_{ERSS}\right)}{Var\left(\bar{Y}_{Ereg}\right)}=\frac{\frac{\sigma_{[.]}^2}{n}}{\frac{\sigma_Y^2}{n}\left(1-\rho^2\right)\left[1+\mathrm{E}\left(\frac{\bar{Z}_{ERSS}^2}{S_{Z_E}^2}\right)\right]} \tag{2.10}$$

whereas the relative precision of ERSS regression estimator, $\bar{Y}_{Ereg}=\frac{1}{r}\sum_{i=1}^{r}\bar{Y}_{[i]}^2$ relative to the RSS naive estimator $\bar{Y}_{RSS}$ is given by

$$RP\left(\bar{Y}_{Ereg},\bar{Y}_{RSS}\right)=\frac{Var\left(\bar{Y}_{RSS}\right)}{Var\left(\bar{Y}_{Ereg}\right)}=\frac{1}{\sigma_Y^2\left(1-\rho^2\right)\left[1+\mathrm{E}\left(\frac{\bar{Z}_{ERSS}^2}{S_{Z_E}^2}\right)\right]}. \tag{2.11}$$

For the variances of the naïve RSS and ERSS estimators, see for example Samawi *et al*. (1996). As it is known that $Var\left(\bar{Y}_{ERSS}\right)<Var\left(\bar{Y}_{SRS}\right)$, Samawi *et al*. (1996), we only compare $\bar{Y}_{Ereg}$ to $\bar{Y}_{ERSS}$ and $\bar{Y}_{RSS}$. Using (2.10), $\bar{Y}_{Ereg}$ has the a greater precision than $\bar{Y}_{[.]}^2$ whenever

$$|\rho|\geq\left[1-\frac{\frac{1}{2}}{\sigma_y^2\left[1+\mathrm{E}\left(\frac{\bar{Z}_{ERSS}^2}{S_{Z_E}^2}\right)\right]}\right]^{\frac{1}{2}}.$$

Therefore, the regression method of estimating μ_y based on ERSS is most preferable if ρ is large. Similarly, from (2.11), $\bar{Y}_{Ereg}$ has a greater precision than $\bar{Y}_{RSS}$ whenever

$$|\rho| \geq \left[1 - \frac{\frac{1}{r}\sum_{i=1}^{r} \sigma^2_{Y_{[i]}}}{\sigma_y^2 \left[1 + \mathrm{E}\left(\frac{\bar{Z}^2_{ERSS}}{S^2_{Z_E}} \right) \right]} \right]^{\frac{1}{2}}.$$

2.5 Comparisons with Regression Estimators

2.5.1 Comparisons with SRS Regression Estimator

We consider the relative precision of our proposed ERSS regression estimator relative to the SRS regression estimator. Table 2.1 presents the relative precision when (X,Y) has a bivariate normal distribution with a correlation coefficient of zero. From the table we see that the relative precision is always greater than 1 when $\rho = 0$. Since the relative precision as given in (2.12) is independent of ρ, the ERSS regression estimator is always superior to the SRS regression estimator, regardless of the value of ρ.

Table 2.1. Relative precision of ERSS regression estimator relative to the SRS regression estimator.

$RP\left(\bar{Y}_{Erg}, \bar{Y}_{reg}\right)$ when $\rho = 0$					
m/r	4	5	6	7	8
1	1.771401	1.396518	1.282631	1.213074	1.17554
4	1.054236	1.043639	1.038408	1.032832	1.02938
8	1.023997	1.019787	1.017815	1.015426	1.01390
∞	1	1	1	1	1

2.5.2 Comparisons with RSS Regression Estimator

Finally, we consider the relative precision of our proposed ERSS regression estimator relative to the RSS regression estimator, as presented by Yu and Lam (1997). Following, Yu and Lam (1997), since $\bar{Y}_{ERSS}$ does not utilize any information on the concomitant variable X, it is fair to compare ERSS regression estimator, $\bar{Y}_{Ereg}$, with the regression estimator, $\bar{Y}_{reg}$, based on a SRS, (see Hedayat and Sinha, (1992)) and with the regression estimator, $\bar{Y}_{Reg}$ based on RSS. When the sample is drawn from a bivariate normal population the relative precision of $\bar{Y}_{Ereg}$ relative to $\bar{Y}_{reg}$ is

$$RP\left(\bar{Y}_{Ereg},\bar{Y}_{reg}\right)=\frac{Var\left(\bar{Y}_{reg}\right)}{Var\left(\bar{Y}_{Ereg}\right)}=\frac{1+\frac{1}{n-3}}{1+\mathrm{E}\left(\frac{\bar{Z}^2_{ERSS}}{S^2_{Z_E}}\right)}. \tag{2.12}$$

and the relative precision of $\bar{Y}_{Ereg}$ relative to $\bar{Y}_{Reg}$ is

$$RP\left(\bar{Y}_{Ereg},\bar{Y}_{Reg}\right)=\frac{Var\left(\bar{Y}_{Reg}\right)}{Var\left(\bar{Y}_{Ereg}\right)}=\frac{1+\mathrm{E}\left(\frac{\bar{Z}^2_{RSS}}{S^2_{Z_R}}\right)}{1+\mathrm{E}\left(\frac{\bar{Z}^2_{ERSS}}{S^2_{Z_E}}\right)}. \tag{2.13}$$

Table 2.2 presents the relative precision for a bivariate normal distribution with zero correlation coefficient. The table shows that the relative precision is always greater than 1 when $\rho=0$. Since the relative precision given in (2.13) is independent of ρ, we can again conclude that the ERSS regression estimator is always superior to the RSS regression estimator regardless of the value of ρ.

Table 2.2. Relative precision of ERSS regression estimator relative to the RSS regression estimator

$RP\left(\bar{Y}_{Erg},\bar{Y}_{reg}\right)$ when $\rho=0$					
m/r	4	5	6	7	8
1	1.096072	1.038646	1.029965	1.018206	1.015733
4	1.008527	1.004899	1.004976	1.003545	1.003144
8	1.003801	1.002274	1.00236	1.001684	1.001516
∞	1	1	1	1	1

2.6 Evaluation of Departure from the Linearity Assumption

Generally, if the assumption of the linear relationship in (2.7) is invalid, the ERSS regression estimator is a biased estimator. In such a case, we define the relative precision to be the ratio of the MSEs of the estimators compared. As in Yu and Lam (1997), we evaluate the performance of the regression estimator under the departure from the linearity assumption by using Plackett's class of bivariate distributions with fixed marginal distribution functions $F(x)$ and $G(y)$. The joint cdf is given by

$$\mathrm{H}\left(x,y\right)=\begin{cases}\dfrac{s\left(x,y\right)-\left[s^2\left(x,y\right)-4\psi\left(\psi-1\right)F\left(x\right)G\left(y\right)\right]^{\frac{1}{2}}}{2\left(\psi-1\right)} & \text{if } \psi\neq 1\\ F\left(x\right)G\left(y\right) & \text{if } \psi=1,\end{cases}$$

where $s\left(x,y\right)=1+\left(\psi-1\right)\left[F\left(x\right)+G\left(y\right)\right]$ and the parameter ψ governs the dependence between X and Y.

Table 2.3. Relative precision of ERSS regression estimator relative to ERSS naive estimator when the linearity assumption is violated (bold numbers indicate $RP < 1$).

$r = 4$							
		Y					
		$N(\theta,1)$			$U(0,1)$		
		M			**M**		
X	ψ	1	4	8	1	4	8
$N(\theta,1)$	0.05	1.3437	1.4061	1.5112	1.2469	1.3043	1.3604
	0.3	**0.9467**	1.0188	1.0351	**0.9099**	1.0178	1.0343
	1	**0.8878**	**0.9735**	**0.9897**	**0.8741**	**0.9786**	**0.9909**
	3	**0.9444**	1.0183	1.0382	**0.914**	1.0149	1.0294
	10	1.1241	1.2085	1.2466	1.0167	1.1686	1.1649
$U(0,1)$	0.05	1.3481	1.4636	1.4963	1.3333	1.4565	1.4647
	0.3	**0.9589**	1.0303	1.0452	**0.9511**	1.0305	1.0464
	1	**0.9127**	**0.9913**	**0.9919**	**0.8908**	**0.9839**	**0.9929**
	3	**0.9717**	1.0289	1.0316	**0.9411**	1.0255	1.0438
	10	1.1652	1.1947	1.2164	1.1018	1.1886	1.2483
$r = 5$							
		Y					
		$N(\theta,1)$			$U(0,1)$		
		M			**M**		
X	ψ	1	4	8	1	4	8
$N(\theta,1)$	0.05	1.3485	1.3797	1.4031	1.2057	1.2991	1.2878
	0.3	**0.9692**	1.0241	1.0395	**0.975**	1.0261	1.0355
	1	**0.9305**	**0.9836**	**0.9959**	**0.9336**	**0.9833**	**0.9962**
	3	**0.9473**	1.0165	1.0262	**0.9627**	1.0261	1.0336
	10	1.1455	1.1612	1.1627	1.0535	1.1489	1.1671
$U(0,1)$	0.05	1.3668	1.3565	1.3876	1.3086	1.4015	1.4203
	0.3	1.0039	1.0363	1.0383	**0.9848**	1.036	1.0529
	1	**0.9452**	**0.9885**	**0.9937**	**0.9508**	**0.9884**	**0.9956**
	3	**0.9973**	1.0291	1.028	**0.9834**	1.0306	1.0387
	10	1.1845	1.1734	1.2056	1.1509	1.2088	1.2089

The reason for choosing this class of bivariate distributions is that it covers the full range of dependence**:**

(a) $\psi \to 0 \Rightarrow F(x) = 1 - G(y)$

(b) $\psi = 1 \Rightarrow$ X and Y are independent

(c) $\psi \to \infty \Rightarrow F(x) = G(y)$.

In general, the relationship between X and Y is not linear. However, their relationship might be close to linear when ψ is close to 0 or ∞ and their marginal distributions are the same and symmetric if

ψ is close to 0. For a more detailed description of Plackett's distribution and its random generation, see Johnson (1987), (P. 191-197).

First, we fix the set size *r* to be 4 and 5, and examine *m* = 1, 4, 8. Five types of dependence from strongly negative to strongly positive corresponding to ψ = 0.05, 0.3, 1, 3, 10, and two marginal distributions, normal (θ, 1), uniform (0,1), are considered here. Table 2.3 gives the relative precision of the ERSS regression estimator relative to the ERSS naive estimator based on simulations of size 100,000.

The main conclusions from Table 2.3 are:

1. Clearly, if both *X* and *Y* have symmetric marginal distributions and ψ is 0.05 or 10, the ERSS regression estimator is superior to the ERSS naive estimator since the Plackett's distribution in these cases is close to a bivariate distribution with linearly related marginal.

2. The efficiency decreases as the value of ψ increases from 0.05 to 1, and starts to increase as ψ increases from 1 to 10 for any given value of *m* and for *r* = 4 and 5.

3. For any fix ψ and any value of *r*, we note that as *m* increases the efficiency increases.

In general when ψ is close to 1, the performance of the ERSS regression estimator is poor. This may be due to the fact that when ψ is close to 1, the two variables X and Y are independent.

3. Regression Estimators when μ_x is Unknown

In this Section, we discuss how to obtain the extreme ranked set sample regression estimator by using the method of double sampling (or two-phase sampling), when μ_x is unknown.

3.1 Regression Estimation Using Two-phase Sampling

The regression estimators $\overline{Y}_{Ereg}$, $\overline{Y}_{Reg}$ and $\overline{Y}_{reg}$ involve the population mean μ_x of the concomitant variable X, which is usually unknown in practical settings. If μ_x is unknown, the method of double sampling can be used to obtain an estimate of μ_x. This involves the drawing of a large random sample of size n', which is used to estimate μ_x. A sub-sample of sizeis then selected from the n'' original (n') selected units to study the primary characteristics of Y. Under an Extreme Ranked Set Sampling setting,phase sampling is SRS and the second -. Note that the first $n'' = rm$ and $n' = r^2 m$ phase sampling is ERSS.

Let $\bar{X}'$ be the sample mean of X based on $r^2 m$ observation of X in the first-phase. Clearly, $\bar{X}'$ is an unbiased estimator for μ_x. If ERSS is the second phase sampling, the double sampling regression estimator of the population mean μ_y is defined as

$$\bar{Y}_{Eds} = \bar{Y}_{ERSSa} + \hat{\beta}_a \left(\bar{X}' - \bar{X}_{ERSSa} \right), \tag{3.1}$$

where,

$\bar{Y}_{ERSSa} = \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{r/2}\left(Y_{[1`]2\,j-1k} + Y_{[r]2\,jk}\right)$, $\bar{\boldsymbol{X}}_{\boldsymbol{ERSSa}} = \frac{1}{\boldsymbol{n}}\sum_{\boldsymbol{k}=1}^{\boldsymbol{m}}\sum_{\boldsymbol{j}=1}^{\boldsymbol{r}/2}\left(\boldsymbol{X}_{(1)2\,\boldsymbol{j}\text{-}1\boldsymbol{k}} + \boldsymbol{X}_{(\boldsymbol{r})2\,\boldsymbol{jk}}\right)$,

$\bar{X}' = \frac{\sum_{k=1}^{m}\sum_{j=1}^{r^2} X_{\,jk}}{nr}$, $\hat{\beta}_a$ is as in (2.9) and $n = mr$.

Again, using basic properties of conditional moments, we have the following theorem.

Theorem 3.1: Assume that the model in (2.7) is satisfied and that the underlying marginals distribution functions of Y and X are symmetric. Then the double sampling regression estimator for μ_y defined in (3.1) has the following properties:

(a) $\mathrm{E}\left(\mathrm{Y}_{Eds}\right) = \mu_Y$

(b) $Var\left(\bar{Y}_{\boldsymbol{Eds}}\right) = \frac{\sigma_{\boldsymbol{Y}}^2}{n}\left(1-\rho^2\right)\left[1+\mathrm{E}\left(\frac{\left(\bar{Z}_{\boldsymbol{ERSS}}-\bar{Z}\right)^2}{S_{\boldsymbol{Z_E}}^2}\right)\right]+\frac{\sigma_{\boldsymbol{Y}}^2}{rn}\rho^2$,

where, $Z_{(i)_{jk}}$, $\bar{Z}_{ERSS}$ and $S_{\boldsymbol{E_E}}^2$ as in Section 2

and $\bar{Z} = \frac{\bar{X}'-\mu_X}{\sigma_X}$.

Proof of Theorem 3.1: From the proof of Theorem 2.1, we have

(1) $E\left(\hat{\beta}_a\right) = \beta$ and $Var\left(\hat{\beta}_a \mid \bar{X}\right) = \frac{\sigma_e^2}{\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}\left(X_{(i)jk}-\bar{X}_{ERSSa}\right)^2}$

(2) .

$$E_y\left(\bar{Y}_{Eds}\,\middle|\,X\right) = E_y\left(\bar{Y}_{ERSSa} + \hat{\beta}_a\left(\bar{X}'-\bar{X}_{ERSSa}\right)\,\middle|\,X\right)$$

Proof of (a): $E\left(\bar{Y}_{Eds}\right) = E_x\left(E_y\left(\bar{Y}_{Eds}\,\middle|\,X\right)\right)$

$$= E_y\left(\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}Y_{[i]jk} + \hat{\beta}_a\left(\bar{X}'-\bar{X}_{ERSSa}\right)\,\middle|\,X\right),$$

$$= E_y\left(\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{k}\sum_{i}\left(\mu_y + \beta\left(X_{(i)k}-\mu_x\right)\right) + \hat{\beta}_a\left(\bar{X}'-\bar{X}_{ERSSa}\right)\,\middle|\,X\right),$$

then by the proof of part (1) of Theorem 2.1, we have that

$$E_y\left(\bar{Y}_{Eds}\,\middle|\,X\right) = \mu_y + \beta\left(\bar{X}_{ERSSa}-\mu_x\right) + \beta\left(\bar{X}'-\bar{X}_{ERSSa}\right).$$

Since $\bar{X}_{ERSSa}$ is an unbiased estimator for μ_x (under the symmetry assumption, see Samawi *et al.* (1996)) and $\bar{X}'$ is also an unbiased estimators for μ_x, then

$$E_x E_y(\bar{Y}_{Eds} \mid X) = \mu_y + E_x\{\beta\left(\bar{X}_{ERSSa} - \mu_x\right) + \beta\left(\bar{X}' - \bar{X}_{ERSSa}\right)\} = \mu_y,$$

and hence $\bar{Y}_{Eds}$ is an unbiased estimator of μ_y.

Proof of Theorem 3.1 (B): Similar to the proof of Theorem 2.1,

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

First from the proof of part (a) above,

$$Var_x\left[E_y\left(\bar{Y}_{Eds}\middle|X\right)\right] = Var_x\left[E_y\left(\bar{Y}_{ERSSa} + \hat{\beta}_a\left(\bar{X}' - \bar{X}_{ERSSa}\right)\middle|X\right)\right].$$

$$Var_x\left(E_y\left(\bar{Y}_{Eds}\middle|X\right)\right) = Var_x\{\mu_y + \beta\left(\bar{X}_{ERSSa} - \mu_x\right) + \beta\left(\bar{X}' - \bar{X}_{ERSSa}\right)\}$$

$$= Var_x\left[\beta\bar{X}'\right] = \beta^2\frac{\sigma_x^2}{rm}.$$

From (1) we know that $E_x\left[Var_y\left(\bar{Y}_{Eds}\middle|X\right)\right] = E_x\left[Var_y\left(\bar{Y}_{ERSSa} + \hat{\beta}_a\left(\bar{X}' - \bar{X}_{ERSSa}\right)\middle|X\right)\right]$.

Also,

$$= E_x\left[Var_y\left(\bar{Y}_{ERSSa}\middle|X\right) + \left(\bar{X}' - \bar{X}_{ERSSa}\right)^2 Var_y\left(\hat{\beta}_a\middle|X\right) + 2Cov\left(\bar{Y}_{ERSSa},\ \hat{\beta}_a\left(\bar{X}' - \bar{X}_{ERSSa}\right)\ \middle|X\right)\right].$$

Similar to the proof of Theorem 2.1, we can show that

$$Cov\left(\bar{Y}_{ERSSa}, \hat{\beta}_a\left(\bar{X}' - \bar{X}_{ERSSa}\right)\middle|X\right) = 0, \text{ and hence}$$

$$E_x\left[Var_y\left(\bar{Y}_{Eds}\middle|X\right)\right] = E_x\left(Var_y\left(\bar{Y}_{ERSSa}\middle|X\right)\right) + E_x\left(\left(\bar{X}_{ERSSa} - \mu_x\right)^2 \sigma_y^2 \Big/ \sum_{k=1}^{m}\sum_{j=1}^{r/2}\sum_{i}\left(\left(X_{(i)jk} - \mu_x\right) - \left(\bar{X}_{ERSSa} - \mu_x\right)\right)^2 \Big/ \sigma_x^2\right) Var_y\left(\hat{\beta}_a\middle|X\right)\Bigg),$$

$$= \left(1-\rho^2\right)\sigma_y^2\left[\frac{1}{rm} + E_x\left[\frac{\left(\bar{Z}_{ERSS} - \bar{Z}\right)^2}{(rm)S_{z_E}^2}\right]\right].$$

$$Var\left(\bar{Y}_{Eds}\right) = \frac{\sigma_y^2}{n}\left(1-\rho^2\right)\left[1+E_x\left[\frac{\left(\bar{Z}_{ERSS}-\bar{Z}\right)^2}{S_{z_E}^2}\right]\right]+\rho^2\frac{\sigma_y^2}{rn}.$$

Therefore,

For the double sampling regression estimator $\bar{Y}_{ds}$ based on SRS, Sukhatme and Sukhatme (1970) showed that, when (X,Y) follows a bivariate normal distribution, $\bar{Y}_{ds}$ is an unbiased estimator of μ_y with variance

$$Var\left(\bar{Y}_{ds}\right)=\frac{\sigma_\varepsilon^2}{n}\left[1+\frac{r-1}{r}\frac{1}{n-3}\right]+\frac{1}{rn}\rho^2\sigma_Y^2.$$

3.2 Relative Efficiency

Again since $\bar{Y}_{ERSS}$ did not use any information on the concomitant variable X, we can compare the two-phase ERSS regression estimator, $\bar{Y}_{Eds}$, to the two-phase regression estimator $\bar{Y}_{ds}$ based on SRS, and to the two-phase regression estimator $\bar{Y}_{Rds}$ based on RSS. The relative precision of $\bar{Y}_{Eds}$ relative to $\bar{Y}_{ds}$ when (X,Y) has a bivariate normal distribution (see Tikkiwal, 1960) is

$$RP\left(\bar{Y}_{Eds},\bar{Y}_{ds}\right)=\frac{Var\left(\bar{Y}_{ds}\right)}{Var\left(\bar{Y}_{Eds}\right)}=\frac{\left(1-\rho^2\right)\left[1+\frac{r-1}{r}\frac{1}{n-3}\right]+\frac{\rho^2}{r}}{\left(1-\rho^2\right)\left[1+\mathrm{E}\left(\frac{\left(\bar{Z}_{ERSS}-\bar{Z}\right)^2}{S_{Z_E}^2}\right)\right]+\frac{\rho^2}{r}} \tag{3.2}$$

and the relative precision of $\bar{Y}_{Eds}$ relative to $\bar{Y}_{Rds}$ is

$$RP\left(\bar{Y}_{Eds},\bar{Y}_{Rds}\right)=\frac{Var\left(\bar{Y}_{Rds}\right)}{Var\left(\bar{Y}_{Eds}\right)}=\frac{\left(1-\rho^2\right)\left[1+\mathrm{E}\left(\frac{\left(\bar{Z}_{RSS}-\bar{Z}\right)^2}{S_{Z_R}^2}\right)\right]+\frac{\rho^2}{r}}{\left(1-\rho^2\right)\left[1+\mathrm{E}\left(\frac{\left(\bar{Z}_{ERSS}-\bar{Z}\right)^2}{S_{Z_E}^2}\right)\right]+\frac{\rho^2}{r}}. \tag{3.3}$$

3.3 Numerical Comparison

Assuming that (X,Y) has a bivariate normal distribution, we compute various expressions for the relative efficiencies obtained in the previous section. The set sizes examined are r = 4, 5, 6, 7 and 8 with cycles of m = 1, 4, 8 and ∞. A simulation size of 100,000 is used to evaluate the values of $\mathrm{E}\left(\left(\bar{Z}_{ERSS}-\bar{Z}\right)^2/S_{Z_E}^2\right)$ and $\mathrm{E}\left(\left(\bar{Z}_{RSS}-\bar{Z}\right)^2/S_{Z_R}^2\right)$. In the case of double sampling, note that the

relative precision is less than the relative precision of the case when μ_x is known. This is due to the extra variation introduced when estimating the mean μ_x.

Table 3.1 shows the relative precision of $\bar{Y}_{Eds}$ relative to $\bar{Y}_{ds}$ for an underlying bivariate normal distribution. From the table we see that all the relative precision values are at least 1 indicating again in precision when using ERSS instead of SRS.

The main conclusions from Table 3.1 are:

1. When ranking is done on the variable X, the relative precision is best at $\rho = 0$. The efficiency increases as the value of $|\rho|$ decreases from .99 to 0.
2. For a fixed value of the set size, r, we note that as m increases the efficiency converges rapidly to 1.
3. The efficiency decreases with increasing set size (r), for any given value of m.
4. For a given value of r, there is no change in the efficiency when the cycle is repeated more than 8, (Efficiency stability). This may be due to the fact that when the sample size is large enough to represent the population, the ranking has less impact on the regression estimator.
5. The double sampling ERSS regression estimator is always superior to the double sampling SRS regression estimator no mater how large the correlation coefficient, ρ is.

Table 3.1. The relative precision of double sampling ERSS regression estimator relative to double sampling SRS regression estimator.

$RP\left(\bar{Y}_{Eds}, \bar{Y}_{Rds}\right)$ when $\rho = 0$					
m/r	4	5	6	7	8
1	1.63416	1.34607	1.24611	1.19055	1.15995
4	1.04718	1.03945	1.03428	1.03009	1.02665
8	1.02100	1.01802	1.01588	1.01399	1.01269
∞	1	1	1	1	1
$RP\left(\bar{Y}_{Eds}, \bar{Y}_{Rds}\right)$ when $\rho = 0.9$					
m/r	4	5	6	7	8
1	1.3124	1.189	1.1501	1.1206	1.1029
4	1.0226	1.0212	1.0197	1.0186	1.0176
8	1.0100	1.0096	1.0093	1.0087	1.0082
∞	1	1	1	1	1

Table 3.2 presents the relative precision under the assumption of an underlying bivariate normal distribution. Again, the table shows that the relative precisions are all at least 1. We also note that the double sampling ERSS regression estimator is always slightly better than the double sampling RSS regression estimator no mater how large the correlation coefficient, ρ is.

Table 3.2. The relative precision of double sampling ERSS regression estimator relative to double sampling RSS regression estimator.

$RP\left(\bar{Y}_{Eds}, \bar{Y}_{Rds}\right)$ when $\rho = 0$					
m/r	4	5	6	7	8
1	1.02771	1.00935	1.00749	1.0046	1.00405
4	1.00192	1.00111	1.0011	1.00079	1.00071
8	1.00081	1.00053	1.00052	1.00035	1.00034
∞	1	1	1	1	1
$RP\left(\bar{Y}_{Eds}, \bar{Y}_{Rds}\right)$ when $\rho = 0.9$					
m/r	4	5	6	7	8
1	1.0083	1.0063	1.0047	1.0029	1.0025
4	1.0008	1.00064	1.00061	1.0004	1.0004
8	1.0003	1.0002	1.0002	1.0002	1.0002
∞	1	1	1	1	1

4. Application to Bilirubin level in Jaundice Babies

We illustrate the methods discussed above using real data on bilirubin level in jaundice babies who stay in neonatal intensive care. Hyper Bilirubinemia is defined as a total serum Bilirubin above 1.5 mg/dl while neonatal jaundice is defined as yellowish discoloration of skin and sclera and it occurs if Bilirubin level is more than 5 mg/dl. (see Nelson *et al.*, 1994). Jaundice is observed during the first week of life in approximately 60% of term infants (from 37 to less than 42 completed weeks) and 80% of pre-term infants (less than 37 completed weeks) (see Nelson *et al.*, 1994).

Neonatal jaundice is a common problem in full-term infants (42 completed weeks or more (294 days or more)) and pre-term babies. It is possible that the generally accepted levels are too high and may produce some high tone hearing loss. Most experts accept that 18.82 mg/dl to 20 mg/dl should not be exceeded in full-term babies, who are less than three days of age, but that a mature baby can tolerate levels of up to 21.18 mg/dl or 22.35 mg/dl by the fifth day without evidence of damage. Pre-mature babies are probably more susceptible and 17.64mg/dl should not be exceeded. Since most cases of neonatal jaundice appear on the second day of life and most of normal newborn babies leave the hospital after 24 hours of life, our primary concern will be on babies staying in neonatal intensive care.

Physicians are interested in jaundice because of its importance and risk on hearing, brain and death. It will be really helpful to the physicians if we can estimate the populations mean of the amount of Bilirubin in the blood for jaundice pre-term, mature, and full term babies. However, estimating the population mean can be expensive and time consuming. Therefore, there is a need for a sampling scheme which can give more accurate population mean estimates with a smaller sample size, and hence results in saving money and time.

All babies who appear significantly jaundiced on clinical examination should have their plasma Bilirubin estimated. This is done in a laboratory test that needs about half an hour or more to find the level of Bilirubin in the blood. This test is expensive and time consuming. However, by using the regression estimator calculated based on extreme rank set sample, we will show that the population mean of plasma Bilirubin for babies who stay in neonatal intensive care, can be estimated with more precision without measuring all units.

4.1 Data Collection

The data were collected by Samawi and Al-Sagheer (2001) from five hospitals in Jordan. These hospitals are Al-Qawasmeh Hospital, Prince Rahma Hospital, Irbid Specialty Hospital, Ibin al-Nafies Hospital, and Queen Zein Al-Sharaf Hospital.

The data were limited to deliveries in the first six months of 1997. Herein, we find the population mean estimate for the Bilirubin level for neonatal jaundice. Jaundice is measured by the level of Bilirubin in the blood. This level is determined via a blood test (tsb). The unit of measurement is mg/dl. The test is conducted on neonatal infants twice daily during the period of the neonatal in the intensive care. One hundred and twenty cases are included in the study. The weight at birth is taken as the concomitant variable. Since ranking on the concomitant variable X (weight) is easier and measuring X is less expensive than ranking and measuring Y (tsb), we will rank on the variable X.

4.2 Parameters

The following are the exact population values of the data:

$\mu_X = 2.87$, $\sigma_X = 0.71$, $\sum_{i=1}^{120} X_i = 344.73$, $\sum_{i=1}^{120} X_i^2 = 1049.62$, $\mu_Y = 11.18$, $\sigma_Y = 5.08$,

$\sum_{i=1}^{120} Y_i = 1341.06$, $\sum_{i=1}^{120} Y_i^2 = 18062.12$, $\sum_{i=1}^{120} XY = 3877.27$, $\rho = 0.06$.

4.3 Using ERSS, RSS and SRS

ERSS and RSS and SRS sampling methods are used to obtain the samples shown in Table 4.1. The following results are obtained from the samples:

1) Based on the ERSS sample, the regression estimate is $\hat{\mu}_y = 11.46$, with $\hat{Var}(\bar{Y}_{Ereg}) = 0.675$ and the naïve estimate is $\bar{Y}_{ERSS} = 11.47$ with $\hat{Var}(\bar{Y}_{ERSS}) = 0.634$.

2) Based on the RSS sample the regression estimate is $\hat{\mu}_Y = 11.44$ with $\hat{Var}(\bar{Y}_{Reg}) = 0.685$ and the naïve estimate is $\bar{Y}_{RSS} = 11.81$ with $\hat{Var}(\bar{Y}_{RSS}) = 0.560$.

3) Based on the SRS sample, the regression estimate is $\hat{\mu}_y = 11.67$ with $\hat{Var}(\bar{Y}_{reg}) = 0.962$ and the naïve estimate is $\bar{Y}_{SRS} = 11.42$ with $\hat{Var}(\bar{Y}_{RSS}) = 0.746$. Note that $\hat{Var}(\bar{Y}_{reg}) \geq \hat{Var}(\bar{Y}_{Ereg})$ also $\hat{Var}(\bar{Y}_{Reg}) \geq \hat{Var}(\bar{Y}_{Ereg})$.

For the data at hand, the naïve estimators are doing better than the regression estimators. This may be due to the fact that the correlation between the weight and TSB is very small. Although this is only an illustration of the computations, the results confirm our earlier conclusions: $eff(\bar{Y}_{Ereg}, \bar{Y}_{reg},) = 1.42$, $eff(\bar{Y}_{Ereg}, \bar{Y}_{Reg},) = 1.01$.

Table 4.1. The drawn samples.

	ERSS		RSS		SRS	
Cycle	Wt	Tsb	Wt	Tsb	Wt	Tsb
1	2.83	6.67	2.83	6.67	2.83	6.67
	3.50	11.94	2.45	8.71	2.45	8.71
	1.50	8.51	4.15	2.06	1.80	16.94
	3.45	8.00	3.45	8.00	3.00	5.50
2	2.00	10.94	2.00	10.94	2.50	10.58
	2.60	16.76	2.50	16.60	2.50	19.79
	1.50	5.90	2.75	5.60	2.75	5.60
	3.50	12.59	3.50	12.59	3.50	12.59
3	1.95	15.76	1.95	15.76	4.40	16.60
	3.70	12.28	3.40	8.00	3.70	12.82
	2.50	25.12	3.25	5.60	2.85	15.20
	3.00	6.90	3.00	6.90	2.70	14.20
4	1.80	22.94	1.80	22.94	3.00	22.94
	3.60	7.20	2.70	14.20	2.00	10.94
	1.90	8.00	2.70	15.47	2.50	15.19
	3.70	5.50	3.70	5.50	2.50	10.58
5	2.45	13.76	2.45	13.76	2.45	13.76
	3.30	9.53	3.10	12.30	3.15	7.80
	1.95	15.76	2.83	6.67	1.95	15.76
	4.40	10.94	4.40	10.94	1.90	11.88
6	2.50	12.76	2.50	12.76	3.25	5.60
	3.60	16.46	3.20	11.60	3.20	11.60
	1.85	9.20	2.60	22.52	2.60	22.52
	3.15	11.53	3.15	11.53	3.15	11.53
7	2.75	5.60	2.75	5.60	4.45	2.06
	2.85	15.20	2.45	8.71	2.45	13.76
	1.75	8.53	2.30	18.29	1.75	8.53
	3.6	16.46	3.60	16.46	2.20	7.60
8	2.00	11.00	2.00	11.00	3.40	8.00
	3.50	11.94	2.70	7.45	2.85	13.94
	3.00	5.90	3.25	8.90	3.65	7.50
	2.60	22.52	2.60	22.52	1.80	16.94
9	2.70	7.45	2.70	7.45	2.70	7.45
	3.75	8.20	3.40	16.50	3.40	16.50
	1.50	5.90	3.50	22.12	3.10	10.18
	3.40	16.50	3.40	16.50	2.10	14.59
10	1.20	8.76	1.20	8.76	3.20	11.60
	3.85	14.27	2.50	7.06	3.85	14.27
	3.00	12.3	3.20	8.53	3.20	8.53
	3.30	3.30	3.30	3.30	3.00	5.50

5. References

HEDAYAT, A.S. and SINHA, B.K. 1992. *Design and inference in finite population sampling*. New York: Wiley.

KAUR, A., PATIL, G.P., SINHA A.K. and TAILLIE, C. 1995. Ranked set sampling: An annotated bibliography. *Environmental and Ecological Statistics* **2:** 25-53

JOHNSON, M.E. 1987. *Multivariate Statistical Simulation*. New York: Wiley.

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

NELSON, W.E. BEHRMAN, R.E., KLIEGMAN, R.M. and VANGHAN, V.C. 1994. *Textbook of pediatrics.* 4th edn. W. B. Saunders Company Harcourt Barance Jovanovich, Inc.

PATIL, G.P., SINHA, A.K. and TAILLIE 1999. Ranked set sampling: a bibliography. *Environmental and Ecological Statistics*. **6 (1):** 91-98.

PATIL, G.P., SINHA, A.K. and TAILLIE 1993. Relative precision of ranked set sampling: Comparison with regression estimator. *Environmetrics.* **4 (4):** 399-412.

SAMAWI, H.M. and Al-SAGHEER, O.A. 2001. On the estimation of the distribution function using extreme and median ranked set sampling. *Biom. J.* **43 (3):** 357-373.

SAMAWI, H.M., MOHMMAD, S. and ABU-DAYYEH, W. 1996. Estimating the population means using extreme ranked set sampling. *Biom. Journal*. **38:** 577-586.

SUKHATME. P.V. and SUKHATME, B.V. 1970. *Sampling Theory of Surveys with Applications*. Ames: Iowa State University Press.

TIKKIWAL. B.D. 1960. On the theory of classical regression and double sampling estimation. *Journal of the Royal Statistical Society*, Series B **22:** 131-138.

YANAGAWA, T. and CHEN, S.H. 1980. The MG procedure in ranked set sampling: Comparision with the regression estimator.

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

Received 12 March 2004
Accepted 15 December 2004