Bernoulli **8**(3), 2002, 313–322

Adaptive ranked-set sampling with multiple concomitant variables: An effective way to observational economy

ZEHUA CHEN

Department of Statistics and Applied Probability, National University of Singapore, 3 Science Drive 2, Singapore 117543. E-mail: stachenz@nus.edu.sg

The notion of ranked-set sampling (RSS) proposed by McIntyre provides an effective means of achieving observational economy in certain particular situations. The use of concomitant variables broadens the range of application of RSS. In this paper, we deal with RSS with multiple concomitant variables and develop an adaptive RSS procedure in a general context. The particular case of multiple linear regression estimates with RSS is treated in detail. The procedure is illustrated with a real data set. Some simulation results are also given.

Keywords: adaptive sampling; concomitant variable; observational economy; ranked-set sampling; regression estimate

1. Introduction

The notion of ranked-set sampling (RSS) proposed by McIntyre (1952) has in recent years sparked an explosion of interest among statisticians; see Ross and Stokes (1999). One of the reasons for this is the cost-effective nature of the RSS procedure and its applicability in environmental and ecological sampling where methods providing observational economy are especially needed. The RSS scheme can be briefly described as follows. First, n sets of individual units, each of size k, are drawn at random from a population. Next, for each set, the units in the set are ranked according to some criterion, without measurement of the variable of interest, and then one and only one unit from the set with a specified rank is measured for the variable of interest. RSS is used in situations where measurement of the variable of interest for a sampled unit is costly or time-consuming, but the ranking of a set of units can be easily done by some means other than measurement of the variable of interest. The method, when applicable, reduces cost and improves efficiency.

In the original form of RSS, ranking is done by judgement with respect to the variable of interest itself. However, the variable according to which the ranking is done does not need to be the variable of interest. In particular, the ranking can be done according to some easily obtainable concomitant variable or function of concomitant variables. Situations where there are easily obtainable concomitant variables available often occur in areas such as medical studies, quantitative genetics, social sciences and environmental studies. For instance, as pointed out by Ross and Stokes (1999), in assessing the status of hazard waste

1350–7265

sites a great deal of knowledge about the sites can often be obtained from records, photos and physical characteristics. Or, to take another example, consider assessing the association of certain bio-markers with smoking in lung cancer studies; many more case or control individuals can be archived with their smoking records such as smoking history and daily consumption than can be taken for an expensive laboratory investigation on the bio-markers.

Concomitant variables, when available, can be used for the ranking in RSS as well as for the estimation of features of the variable of interest, especially when the variable of interest has a linear regression relationship with the concomitant variables. Several authors have considered the case of the existence of a single concomitant variable; see, for example, Patil *et al*. (1993), Yu and Lam (1997) and Chen (2001). In practical problems, however, the variable of interest is usually correlated with several concomitant variables. In this paper, we develop an adaptive RSS procedure to cope with multiple concomitant variables. The procedure can be applied whether or not the relationship between the variable of interest and the concomitant variables is linear. In the next section, we discuss ranking criteria and describe the adaptive RSS procedure. Some simulation results are also presented in the section. In Section 3, we present the adaptive RSS procedure combined with regression-type estimates. In Section 4, we illustrate the adaptive RSS procedure with a retrospective study of a data set obtained in the 1988 Test Population Census of Limeira in the state of São Paulo, Brazil.

2. The best ranking criterion and the adaptive RSS procedure

Let Y denote the variable of interest and $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ denote the vector of concomitant variables. When ranking is perfect, we denote the ranked statistic with rank r by $Y_{(r)}$, and otherwise by $Y_{[r]}$. This convention also applies to the concomitant variables and the related parameters.

It can be shown that, among all ranking mechanisms, the perfect ranking with respect to Y itself is the best ranking mechanism. In practice, however, the best ranking mechanism cannot be realized without actual measurements of Y. Since the concomitant variables are our only resort for the ranking, it is then sensible to take a function of the concomitant variables which has the highest correlation with Y as the ranking criterion. It is well known that the conditional expectation $\mathrm{E}[Y|\mathbf{X}]$ has the highest correlation with Y among all functions of $\mathbf{X}$ and hence is the best ranking function. Henceforth, we write $g(\mathbf{X}) = \mathrm{E}[Y|\mathbf{X}]$.

The conditional expectation $g(\mathbf{X})$ cannot yet be readily used for ranking, since it is an unknown function. To overcome this difficulty, we propose an adaptive RSS procedure. The procedure is a cyclical process described as follows. At an initial cycle, the RSS is conducted using any reasonable ranking criterion, say, a single concomitant variable. Subsequent cycles consist of an estimation step and a sampling step. In the estimation step, the available data are used to estimate (or update the estimate of) $g(\mathbf{X})$. In the sampling step, the updated estimate of $g(\mathbf{X})$ is used as a tentative ranking criterion and more ranked-set samples are collected. The procedure is described in more detail in the following algorithm:

Initial cycle. Choose a single concomitant variable with the highest possible correlation with Y. Denote this concomitant variable by X_{INI}. For $r = 1, \ldots, k$, draw a random sample of k units from the population, measure the value of X_{INI} for each unit and rank the units according to the orders of the X_{INI} values, and then measure the Y value for the unit ranked r. Denote the ranked-set sample obtained in this cycle by DATA_1.

Recursive cycles. For $j = 2, \ldots, m$:

(i) Compute the estimator $\hat{g}^{(j-1)}$ of g using DATA_{j-1}.
(ii) For $r = 1, \ldots, k$, draw a random sample of k units from the population, measure the value of $\mathbf{X}$, and compute $\hat{g}^{(j-1)}(\mathbf{X})$ for each unit, rank the units according to the orders of the $\hat{g}^{(j-1)}(\mathbf{X})$ values, and then measure the Y value for the unit ranked r.
(iii) Augment DATA_{j-1} to DATA_j.

The adaptive procedure described above is especially useful and can be easily implemented when sampling units can be archived for later investigations. Consider the following fishery study as an example. A good understanding of the age distribution of the fish in a fishing region is important in fishery studies. However, the process of determining the age of a fish is quite complicated and expensive. First, one of the otoliths of the fish is removed and cleaned; then the otolith is embedded onto a microscope slide with thermoplastic cement and polished with wet and dry sandpaper until the mid-plane is reached; finally, the otolith is viewed under immersion oil on a video screen attached to a microscope and the daily rings are counted along the longitudinal axis towards the posterior of the otolith to determine the age. The adaptive RSS procedure is particularly applicable in this case. First, since the age of a fish is closely associated with the fish's length and weight, which can be easily measured, the length and weight can be taken as concomitant variables. Second, the collected otoliths can be put into labelled plastic bags together with the records of the measurements on length and weight for the later laboratory investigation.

For the estimation of such features of Y as the mean, variance and quantiles, the adaptive RSS procedure is more efficient than any other RSS procedures using other ranking functions when the sample size is large. Although there is no theoretical guarantee that this should be true in the small-sample case, we have from simulation some evidence in favour of the adaptive procedure even if the sample size is small.

We present some simulation results for the estimation of the population mean of Y. The model $Y = g(\mathbf{X}) + \epsilon$ is considered in three examples in our simulation study, where ϵ is a random error with distribution $N(0, \sigma^2)$ and $\mathbf{X}$ is a random vector independent of ϵ; $g(\mathbf{X})$ is taken as a linear function of $\mathbf{X}$ in the first two examples and as nonlinear in the third example. The simulation study compares the adaptive RSS procedure with two ordinary RSS procedures: one using $g(\mathbf{X})$ as the ranking criterion and the other using a single concomitant variable. The ranking criterion g is used to illustrate the effectiveness of the adaptive procedure. In each of the three examples, $k = 5$ and the three sampling procedures are simulated 2000 times. The mean square errors of the estimates over the 2000 simulations are presented in Tables 1–3. In the tables, MSE_g, MSE_x and $\mathrm{MSE}_{\hat{g}}$ denote the mean square errors of the RSS estimates with ranking criteria g, X and $\hat{g}$, respectively.

Table 1. Simulated mean square errors and relative efficiencies of Example 1: $g(\mathbf{X}) = -2X_1 + 2X_2 + 2X_3$ and $\mathbf{X} \sim N_3(0, 0.75I + 0.35\mathbf{1}\mathbf{1}^{\mathrm{T}})$. See text for definitions of MSE and RE terms

σ	m	MSE_g	MSE_{X_3}	$\mathrm{MSE}_{\hat{g}}$	$\mathrm{RE}_{\hat{g}\cdot g}$	$\mathrm{RE}_{\hat{g}\cdot X_3}$	$\mathrm{MSE}_{\mathrm{REG}}$
	4	0.218	0.395	0.289	0.755	1.368	0.154
1	8	0.110	0.195	0.130	0.849	1.500	0.073
	16	0.055	0.089	0.057	0.963	1.574	0.036
	4	0.576	0.715	0.652	0.883	1.096	0.524
2.82	8	0.289	0.374	0.320	0.903	1.170	0.264
	16	0.141	0.188	0.149	0.943	1.257	0.121
	4	3.491	3.536	3.480	1.003	1.016	3.522
8	8	1.600	1.734	1.645	0.973	1.054	1.636
	16	0.824	0.896	0.836	0.986	1.072	0.809

Table 2. Simulated mean square errors and relative efficiencies of Example 2: $g(\mathbf{X}) = -2X_1 + 2X_2 + 2X_3$ and X_1, X_2, X_3 are independently and identically uniform on $[-2, 2]$. See text for definitions of MSE and RE terms

σ	m	MSE_g	MSE_{X_3}	$\mathrm{MSE}_{\hat{g}}$	$\mathrm{RE}_{\hat{g}\cdot g}$	$\mathrm{RE}_{\hat{g}\cdot X_3}$	$\mathrm{MSE}_{\mathrm{REG}}$
	4	0.332	0.663	0.451	0.737	1.469	0.225
1	8	0.153	0.326	0.180	0.853	1.813	0.103
	16	0.080	0.173	0.089	0.902	1.949	0.054
	4	0.640	1.002	0.825	0.776	1.214	0.590
2.82	8	0.341	0.517	0.377	0.903	1.370	0.292
	16	0.168	0.251	0.184	0.911	1.359	0.145
	4	3.458	3.953	3.644	0.949	1.085	3.576
8	8	1.790	1.858	1.805	0.992	1.029	1.728
	16	0.886	1.012	0.952	0.931	1.063	0.902

$\mathrm{RE}_{\hat{g}\cdot g}$ and $\mathrm{RE}_{\hat{g}\cdot x}$ are defined as $\mathrm{MSE}_g/\mathrm{MSE}_{\hat{g}}$ and $\mathrm{MSE}_x/\mathrm{MSE}_{\hat{g}}$ respectively, the relative efficiencies of the adaptive RSS with respect to the other two RSS procedures. Also given in the tables are the mean square errors of the regression estimator to be discussed in the next section.

3. Ranked-set sampling regression estimator with multiple concomitant variables

In this section, we consider the case where Y and $\mathbf{X}$ follow a linear regression model. In this case, the concomitant variables not only can be used for ranking in RSS but also can be incorporated into the estimation of the mean of Y. Suppose that

Table 3. Simulated mean square errors and relative efficiencies of Example 3: $g(X) = 2\sin(X) + 3\cos(X)$, X is uniform on $[0, 2\pi]$. See text for definitions of MSE and RE terms

σ	m	MSE_g	MSE_X	$\text{MSE}_{\hat{g}}$	$\text{RE}_{\hat{g}\cdot g}$	$\text{RE}_{\hat{g}\cdot x}$
	4	0.152	0.316	0.197	0.774	1.607
1	8	0.095	0.222	0.117	0.812	1.905
	16	0.075	0.160	0.088	0.858	1.817
	4	0.339	0.504	0.396	0.856	1.273
2.82	8	0.213	0.325	0.233	0.917	1.395
	16	0.167	0.250	0.185	0.906	1.355
	4	1.734	1.945	1.887	0.919	1.031
8	8	1.211	1.283	1.172	1.033	1.095
	16	0.777	0.931	0.890	0.873	1.046

$$Y = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} + \epsilon, \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of unknown constant coefficients, and ϵ is a random variable with mean zero and is independent of $\mathbf{X}$. Let $\boldsymbol{\Sigma}$ denote the variance–covariance matrix of $\mathbf{X}$ and $\boldsymbol{\sigma}_{\mathbf{X}y}$ denote the vector of covariances between the components of $\mathbf{X}$ and Y. Let σ_ϵ^2 and σ_Y^2 denote the variances of ϵ and Y, respectively. It can be verified that

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_{\mathbf{X}y}, \tag{2}$$

$$\sigma_\epsilon^2 = \sigma_Y^2 - \boldsymbol{\sigma}_{\mathbf{X}y}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_{\mathbf{X}y}. \tag{3}$$

Suppose that the sampling is implemented in m cycles. In a typical cycle i, for $r = 1, \ldots, k$, a simple random sample of size k with latent values $(Y_{1ri}, \mathbf{X}_{1ri}), \ldots, (Y_{kri}, \mathbf{X}_{kri})$ is drawn from the population. The values of the $\mathbf{X}$s are all measured. The k sampled units are ranked according to some function of $\mathbf{X}$. Then the Y value of the unit with rank r is measured. The $\mathbf{X}$ and Y values of the unit with rank r are denoted by $\mathbf{X}_{[r]i}$ and $Y_{[r]i}$, respectively. At the completion of the sampling, we have a data set which is represented as

$$(Y_{[1]i}, \mathbf{X}_{[1]i}, \mathbf{X}_{11i}, \ldots, \mathbf{X}_{k1i}), \ldots, (Y_{[k]i}, \mathbf{X}_{[k]i}, \mathbf{X}_{1ki}, \ldots, \mathbf{X}_{kki}), \qquad i = 1, \ldots, m. \tag{4}$$

(*i*) *The regression estimator.* Let

$$\overline{\mathbf{X}}_{\mathrm{RSS}} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m}\boldsymbol{X}_{[r]i},\ \overline{Y}_{\mathrm{RSS}} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m}Y_{[r]i},$$

$$\overline{X}_{\mathrm{T}} = \frac{1}{mk^2}\sum_{r=1}^{k}\sum_{s=1}^{k}\sum_{i=1}^{m}\boldsymbol{X}_{sri}.$$

Let $\hat{\alpha}_{\mathrm{RSS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{RSS}}$ denote, respectively, the least-squares estimates of α and $\boldsymbol{\beta}$ based on the ranked-set data, i.e.,

$$\hat{\alpha}_{\text{RSS}} = \overline{Y}_{\text{RSS}} - \hat{\boldsymbol{\beta}}_{\text{RSS}}^{\text{T}} \overline{X}_{\text{RSS}}, \tag{5}$$

$$\hat{\boldsymbol{\beta}}_{\text{RSS}} = \left[\mathbf{X}_{\text{RSS}}^{\text{T}}\left(I - \frac{\mathbf{11}^{\text{T}}}{mk}\right)\mathbf{X}_{\text{RSS}}\right]^{-1}\mathbf{X}_{\text{RSS}}^{\text{T}}\left(I - \frac{\mathbf{11}^{\text{T}}}{mk}\right)\boldsymbol{Y}_{\text{RSS}}, \tag{6}$$

where

$$\mathbf{X}_{\text{RSS}} = (\mathbf{X}_{[1]1}, \ldots, \mathbf{X}_{[k]1}, \ldots, \mathbf{X}_{[1]m}, \ldots, \mathbf{X}_{[k]m})^{\text{T}}$$

$$Y_{\text{RSS}} = (y_{[1]1}, \ldots, y_{[k]1}, \ldots, y_{[1]m}, \ldots, y_{[k]m})^{\text{T}}.$$

The RSS regression estimator of the mean of Y is defined as

$$\hat{\mu}_{\text{RSS}\cdot\text{REG}} = \overline{Y}_{\text{RSS}} + \hat{\boldsymbol{\beta}}_{\text{RSS}}^{\text{T}}(\overline{X}_{\text{T}} - \overline{X}_{\text{RSS}}). \tag{7}$$

The variance of the RSS regression estimator can be obtained as

$$\text{var}(\hat{\mu}_{\text{RSS}\cdot\text{REG}}) = \frac{\sigma_{\epsilon}^2}{mk}\{1 + \Delta_{\text{RSS}}\} + \frac{1}{mk^2}\boldsymbol{\beta}^{\text{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}, \tag{8}$$

where

$$\Delta_{\text{RSS}} = \text{E}\left[mk(\overline{\mathbf{X}}_{\text{T}} - \overline{X}_{\text{RSS}})^{\text{T}}\left[\mathbf{X}_{\text{RSS}}^{\text{T}}\left(I - \frac{\mathbf{11}^{\text{T}}}{mk}\right)\mathbf{X}_{\text{RSS}}\right]^{-1}(\overline{X}_{\text{T}} - \overline{X}_{\text{RSS}})\right].$$

If, in place of the ranked-set sample, we have a simple random sample (SRS), we will get an SRS regression estimator. The variance of the SRS regression estimator is also given by (8) but with Δ_{RSS} replaced by Δ_{SRS}, defined on the simple random sample instead of the ranked-set sample. It can be proved that, no matter which ranking criterion is used, we always have

$$\Delta_{\text{RSS}} < \Delta_{\text{SRS}}.$$

In other words, as in the case of a single concomitant variable, the RSS regression estimator is always more efficient than the SRS regression estimator.

(*ii*) *The ordinary RSS estimator or the RSS regression estimator?* In the case where Y and $\boldsymbol{X}$ follow a linear regression model, one faces a choice between an ordinary RSS estimator and an RSS regression estimator. In order to make the choice, we have to compare their respective variances. It can be shown that

$$\text{var}(\hat{\mu}_{\text{RSS}\cdot\text{REG}}) \approx \frac{\sigma_{\varepsilon}^2}{mk}\left[1 + \frac{1}{mk}\,\text{tr}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\text{RSS}} - \frac{p}{mk^2}\right] + \frac{1}{N}\boldsymbol{\beta}^{\text{T}}\boldsymbol{\Sigma}\boldsymbol{\beta},$$

where $\boldsymbol{\Sigma}_{\text{RSS}} = (1/k)\sum_{r=1}^{k}\boldsymbol{\Sigma}_{[r]}$, $\boldsymbol{\Sigma}_{[r]}$ denotes the covariance matrix of $\mathbf{X}_{[r]}$. Suppose that $\boldsymbol{\beta}^{\text{T}}\mathbf{X}$ is used as the ranking criterion. Then the variance of the ordinary RSS estimator $\overline{Y}_{\text{RSS}}$ can be obtained as

$$\text{var}(\overline{Y}_{\text{RSS}}) = \frac{1}{mk}\boldsymbol{\beta}^{\text{T}}\boldsymbol{\Sigma}_{\text{RSS}}\boldsymbol{\beta} + \frac{1}{mk}\sigma_{\varepsilon}^2.$$

If we further assume the normality of $\mathbf{X}$, we have

$$\boldsymbol{\Sigma}_{\mathrm{RSS}} = \boldsymbol{\Sigma} - \frac{\boldsymbol{\Sigma\beta\beta}^{\mathrm{T}}\boldsymbol{\Sigma}}{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma\beta}} D_k,$$

where

$$D_k = 1 - \frac{1}{k}\sum_{r=1}^{k} \sigma_{(r)}^2,$$

$\sigma_{(r)}^2$ being the variance of the rth order statistic of a standard normal sample of size k.

Define $R^2 = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma\beta}/\sigma_Y^2$. Recall that $\sigma_\varepsilon^2 = \sigma_Y^2 - \boldsymbol{\sigma}_{\mathbf{X}y}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_{\mathbf{X}y} = \sigma_{\mathrm{Y}}^2 - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma\beta}$. We then have

$$\mathrm{var}(\overline{Y}_{\mathrm{RSS}}) = \frac{\sigma_Y^2}{mk} R^2(1 - D_k) + \frac{1}{mk}\sigma_\varepsilon^2, \tag{9}$$

$$\mathrm{var}(\hat{\mu}_{\mathrm{RSS\cdot REG}}) = \frac{\sigma_Y^2}{mk}\left[\frac{1 - R^2}{mk}\left(\frac{k-1}{k}p - D_k\right) + \frac{1}{k}R^2\right] + \frac{1}{mk}\sigma_\varepsilon^2.$$

Therefore,

$$\mathrm{var}(\hat{\mu}_{\mathrm{RSS\cdot REG}}) \leqslant \mathrm{var}(\overline{Y}_{\mathrm{RSS}}) \Leftrightarrow \frac{1 - R^2}{mk}\left(\frac{k-1}{k}p - D_k\right) + \frac{1}{k}R^2 \leqslant R^2(1 - D_k)$$

$$\Leftrightarrow R^2 \geqslant \frac{p - k/(k-1)D_k}{p + mk - k(mk+1)/(k-1)D_k}. \tag{11}$$

As long as inequality (11) holds, the RSS regression estimator is more efficient than the ordinary RSS estimator. In fact, the quantity on the right-hand side of (11) is usually small. For example, under the normality assumption, when $k = 5$, $m = 10$ and $p = 2$ the quantity is only 0.1067. Notice that the R^2 is the proportion of the variation in Y explained by $\boldsymbol{X}$. Therefore, generally speaking, the RSS regression estimator is more efficient than the ordinary RSS estimator as long as the concomitant variables are reasonably related to the variable of interest. The mean square errors of the regression estimates in the first two examples of the previous section, denoted by $\mathrm{MSE}_{\mathrm{REG}}$, are given in the last column of Tables 1 and 2, which can be compared with the mean square errors of the ordinary RSS estimates.

4. A retrospective application to Brazil census data

In this section, we illustrate the adaptive RSS procedure with a retrospective study of a data set obtained in the 1988 Test Population Census of Limeira, São Paulo, Brazil. The test census was carried out in two stages. At the first stage, a population of about 44 000 households is administered a ‘short form’ questionnaire. At the second stage, a systematic sample of about 10% of the population size is surveyed with a ‘long form’ questionnaire. The ‘short form’ contains variables such as sex, age and education of the head of household, ownership of house, car and colour television, the number of rooms and bathrooms, a proxy to the monthly income of the head of household, etc. The ‘long form’ contains, in addition to the variables in the ‘short form’, the actual monthly income of the head of household and other variables. The data set which is used for our illustration consists of the sample records of the ‘long form’ for 426 heads of household. Details of the data set are described in Silva and Skinner (1997). Here, we take these 426 records as our population and consider the actual monthly income as the variable of interest. The following variables from among the variables in the ‘short form’ are taken as concomitant variables: total number of bathrooms in household, x_1; years of study of head of household, x_2; indicator of sex of head of household, x_3 (1 if male, 0 otherwise); an indicator variable, x_4, for age of head of household whether less than or equal to 35.

We simulated the adaptive RSS procedure 1000 times. We set $k = 5$ and $m = 10$, i.e., in the adaptive procedure, the set size is 5 and the number of cycles is 10. Each simulation is done as follows. At each cycle, five independent simple random samples of size 5 are taken from the 426 records with replacement. Except in the initial cycle, each of these simple random samples is ranked according to the first concomitant variables, the linear regression function of the first and second concomitant variables, the linear regression function of the first, second and third concomitant variables, and so on, thus generating four different ranked-set subsamples. The coefficients of the regression functions are estimated by using the data from previous cycles. In the initial cycle, the simple random samples are ranked according to the first concomitant variable and all four ranked-set subsamples are taken from the same units, differing only by the number of concomitant variables. A simple random subsample is generated in each cycle by taking the first record from the first simple random sample, the second record from the second simple random sample, and so on. At the completion of the data generation, the following 13 estimators of the population mean, whose actual value is 194.34, are computed: the mean monthly income of the simple random sample, $\hat{\mu}_1$; the mean monthly income of the ranked-set sample when ranking is done using the first concomitant variable, $\hat{\mu}_2$; the mean monthly income of the ranked-set sample when ranking is done using the first two concomitant variables, $\hat{\mu}_3$; the mean monthly income of the ranked-set sample when ranking is done using the first three concomitant variables, $\hat{\mu}_4$; the mean monthly income of the ranked-set sample when ranking is done using all the four concomitant variables, $\hat{\mu}_5$; the regression estimate based on the simple random sample using only the first concomitant variable, $\hat{\mu}_6$; the regression estimate based on the simple random sample using the first two concomitant variables, $\hat{\mu}_7$; the regression estimate based on the simple random sample using the first three concomitant variables, $\hat{\mu}_8$; the regression estimate based on the simple random sample using all the four concomitant variables, $\hat{\mu}_9$; the regression estimate based on the first ranked-set sample with the first concomitant variable, $\hat{\mu}_{10}$; the regression estimate based on the second ranked-set sample with the first two concomitant variables, $\hat{\mu}_{11}$; the regression estimate based on the third ranked-set sample with the first three concomitant variables, $\hat{\mu}_{12}$; the regression estimate based on the fourth ranked-set sample with all the four concomitant variables, $\hat{\mu}_{13}$. The mean square error of each of these estimators is approximated by

$$\mathrm{MSE}_i = \sum_{s=1}^{1000} [\hat{\mu}_i(s) - \mu_y]^2/1000.$$

These approximated mean square errors are given in Table 4.

Table 4. The approximate mean square errors of the 13 estimators of μ_y based on a simulation of size 1000

Estimators	MSE with concomitants (x_1, x_2, x_3, x_4)
$\hat{\mu}_1$	1634
$\hat{\mu}_2$	1430
$\hat{\mu}_3$	1315
$\hat{\mu}_4$	1361
$\hat{\mu}_5$	1333
$\hat{\mu}_6$	1349
$\hat{\mu}_7$	1266
$\hat{\mu}_8$	1253
$\hat{\mu}_9$	1255
$\hat{\mu}_{10}$	1289
$\hat{\mu}_{11}$	1191
$\hat{\mu}_{12}$	1195
$\hat{\mu}_{13}$	1175

The properties of the adaptive RSS procedure discussed previously are evident in the table. First, all the original RSS estimators are more efficient than the SRS sample mean, and the original RSS estimators with multiple concomitant variables are more efficient than the original RSS estimator with a single concomitant variable. Second, the regression estimators improve the sample means, and the RSS regression estimators improve their SRS counterparts. Third, the RSS regression estimators with multiple concomitant variables improve the RSS regression estimator with a single concomitant variable.

An interactive programme for the implementation of the adaptive RSS procedure is available from the author upon request.

Acknowledgement

The research for this paper was carried out during a visit to the Department of Statistics, Pennsylvania University. The author would like to express his appreciation to the referees for their constructive comments and suggestions which led to a substantial revision of the first draft.

References

Chen, Z. (2001) Ranked-set sampling with regression-type estimators. *J. Statist. Plann. Inference*, **92**, 181–192.

McIntyre, G.A. (1952) A method of unbiased selective sampling, using ranked sets. *Austral. J. Agriculture Res.*, **3**, 385–390.

Patil, G.P., Sinha, A.K. and Taillie, C. (1993) Observational economy of ranked-set sampling: comparison with the regression estimator. *Environmetrics*, **4**, 399–412.

Ross, N.P. and Stokes, S.L. (1999) Editorial: Special issue on statistical design and analysis with ranked-set samples. *Environ. Ecological Statist.*, **6**, 5–9.

Silva, P.L.D.N. and Skinner, C.J. (1997) Variable selection for regression estimation in finite populations. *Survey Methodol.*, **23**, 23–32.

Yu, P.L.H. and Lam, K. (1997) Regression estimator in ranked set sampling. *Biometrics*, **53**, 1070–1080.

Received February 2001 and revised September 2001