Journal of Official Statistics, Vol. 22, No. 3, 2006, pp. 541–563

Alternative Designs for Regression Estimation

Mingue Park[1]

Restricted random samples partially balanced on sample moments are compared with a two-per-stratum design as designs for use with regression estimation. The shape of the average weight for the best linear unbiased predictor as a function of the auxiliary variable is similar to the shape of the inclusion probability for a sample partially balanced on the auxiliary variable. In the simulation study the MSE of the regression estimator with the stratified random sample is comparable to the regression estimator with the balanced sample when the assumed linear model holds. When estimating the population cumulative distribution function at selected points, two-per-stratum stratified random sampling shows better performance than partially balanced samples.

Key words: Best linear unbiased predictor; regression estimator; regression superpopulation model; balanced sampling.

1. Introduction

Design and estimation in survey sampling involve the use of information about the study population, sometimes called *auxiliary information*, to construct efficient procedures. If the auxiliary variables are available at the design stage for every element in the population, they can be used in sample selection and in estimation. Stratification based on the size of auxiliary variables is a commonly used design when such auxiliary information is available. Stratified random sampling is covered in standard texts such as Cochran (1977) and Särndal et al. (1992).

Under a regression superpopulation model, the model variance of the regression estimator for a simple random sample is smaller than that of the sample mean if the multiple correlation coefficient is larger than the ratio of the number of auxiliary variables to the sample size. The efficiency of the regression estimator relative to the Horvitz-Thompson estimator in a design based approach has been addressed by, for example, Cochran (1977) and Särndal et al. (1992). Design consistency of the regression estimator has been discussed by Isaki and Fuller (1982) and Robinson and Särndal (1983). Under a regression model, Fuller and Park (2002) give conditions under which the regression estimator is the best linear model unbiased predictor (BLUP) and is also design consistent.

[1] Korea University, Department of Statistics, 1 Anam-Dong, Seoul 136-701, Korea. Email: mpark2@korea.ac.kr
Acknowledgments: This research was part of my Ph.D. dissertation at Iowa State University and was partially supported by Cooperative Agreement 13-3AEU-0-80064 between Iowa State University, the U.S. National Agricultural Statistics, and the U.S. Bureau of the Census. I am deeply indebted to Professor Wayne Fuller for numerous suggestions made throughout the research.

Royall (1992) gave results for a particular kind of balanced sample in which the sample is selected so that the sample moments of the auxiliary variables are equal to the population moments. He showed that, for a population satisfying the regression model, the balanced sample is optimal in the sense that it minimizes the model variance of the BLUP of the population total. Dorfman and Valliant (2000) considered the stratified balanced sample constructed by selecting balanced samples from each stratum in a set of strata. Using prediction theory, they concluded that an unstratified, balanced sample yields essentially the same model variance for the BLUP as a stratified balanced sample.

Our motivation is estimation for a large-scale survey in which a large number of analyses of a large number of variables is anticipated. In such a situation, a single model usually fails to explain the relationship between all study variables and a set of auxiliary variables. The set of sample weights that gives the BLUP for a particular variable need not give the BLUP for other variables. Therefore it is desirable to construct robust strategies. In this article, we consider several selection strategies for the regression estimator. By deriving the approximate inclusion probabilities of samples partially balanced on auxiliary variables, we compare the BLUP with a simple random sample to the BLUP for a partially balanced sample. Through a simulation study, we compare stratified random sampling with the partially balanced samples for regression estimation.

2. Balanced Samples and Restricted Random Sampling

Royall (1992) presented a theorem that identifies an optimal design for a particular model and such that the design provides robustness against certain models. Royall assumed

$$\mathbf{E}\{\mathbf{y}_U\} = \mathbf{X}_U\beta, \quad \mathbf{V}\{\mathbf{y}_U\} = \boldsymbol{\Sigma}_{UU} \tag{1}$$

where $\mathbf{X}_U$ is an $N \times p$ matrix of regressors, $\mathbf{y}_U$ is an N-dimensional column vector of study variables, $\boldsymbol{\Sigma}_{UU}$ is a diagonal matrix with diagonal elements σ_{UUii}, $i = 1, \cdots, N$ and N is the population size. The matrix $\mathbf{X}_U$ and the matrix $\boldsymbol{\Sigma}_{UU}$ are known. The vector β is unknown. For a given sample A of n units, let

$$\mathbf{y}_U = \begin{pmatrix} \mathbf{y}_A \\ \mathbf{y}_{\bar{A}} \end{pmatrix}, \quad \mathbf{X}_U = \begin{pmatrix} \mathbf{X}_A \\ \mathbf{X}_{\bar{A}} \end{pmatrix}, \quad \boldsymbol{\Sigma}_{UU} = \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & 0 \\ 0 & \boldsymbol{\Sigma}_{\bar{A}\bar{A}} \end{pmatrix}$$

where $\mathbf{y}_{\bar{A}}$, $\mathbf{X}_{\bar{A}}$ and $\boldsymbol{\Sigma}_{\bar{A}\bar{A}}$ are the quantities corresponding to nonsampled elements and $\mathbf{y}_A$, $\mathbf{X}_A$ and $\boldsymbol{\Sigma}_{AA}$ are the quantities corresponding to sampled elements. Let $\mathbf{J}_n$ and $\mathbf{J}_{(N-n)}$ denote columns of ones of length n and $(N-n)$, respectively. BLUP of the population mean $\bar{y}_N = N^{-1}(\mathbf{J}_n'\mathbf{y}_A + \mathbf{J}_{(N-n)}'\mathbf{y}_{\bar{A}})$ is

$$\bar{y}_{BLUP} = N^{-1}(\mathbf{J}_n'\mathbf{y}_A + \mathbf{J}_{(N-n)}'\mathbf{X}_{\bar{A}}\hat{\beta}_{wls}) \tag{2}$$

where $\hat{\beta}_{wls} = (\mathbf{X}_A'\boldsymbol{\Sigma}_{AA}^{-1}\mathbf{X}_A)^{-1}(\mathbf{X}_A'\boldsymbol{\Sigma}_{AA}^{-1}\mathbf{y}_A)$. The model variance is

$$\mathbf{V}\{\bar{y}_{BLUP} - \bar{y}_N\} = N^{-2}\left[\mathbf{J}_{(N-n)}'(\mathbf{X}_{\bar{A}}\mathbf{G}_A^{-1}\mathbf{X}_{\bar{A}}' + \boldsymbol{\Sigma}_{\bar{A}\bar{A}})\mathbf{J}_{(N-n)}\right] \tag{3}$$

where $\mathbf{G}_A = \mathbf{X}_A'\boldsymbol{\Sigma}_{AA}^{-1}\mathbf{X}_A$. Under the assumption that both $\boldsymbol{\Sigma}_{UU}\mathbf{J}_N$ and $\boldsymbol{\Sigma}_{UU}^{\frac{1}{2}}\mathbf{J}_N$ are in the column space of $\mathbf{X}_U$, the BLUP of the population mean, defined in (2), can be expressed as

$$\bar{y}_{BLUP} = \bar{\mathbf{x}}_N \hat{\beta}_{wls} \tag{4}$$

and the variance of the BLUP satisfies the inequality

$$\mathbf{V}\{\bar{y}_{BLUP} - \bar{y}_N\} \geq N^{-2}\left\{n^{-1}\left(\mathbf{J}'_N \mathbf{\Sigma}^{\frac{1}{2}}_{UU}\mathbf{J}_N\right)^2 - \mathbf{J}'_N \mathbf{\Sigma}_{UU}\mathbf{J}_N\right\} \tag{5}$$

The bound in (5) is attained if and only if the design satisfies

$$\frac{1}{n}\mathbf{J}'_n \mathbf{\Sigma}^{-\frac{1}{2}}_{AA}\mathbf{X}_A = \frac{\mathbf{J}'_N \mathbf{X}_U}{\mathbf{J}'_N \mathbf{\Sigma}^{\frac{1}{2}}_{UU}\mathbf{J}_N} \tag{6}$$

and in which case

$$\bar{y}_{BLUP} = \frac{1}{N}\left(\frac{1}{n}\mathbf{J}'_N \mathbf{\Sigma}^{\frac{1}{2}}_{UU}\mathbf{J}_N\right)\left(\mathbf{J}'_n \mathbf{\Sigma}^{-\frac{1}{2}}_{AA}\mathbf{y}_A\right) \tag{7}$$

See, Royall (1992).

A sample which satisfies the condition (6) is called a *weighted balanced sample*. See Royall and Herson (1973) and Royall (1992). When $\mathbf{\Sigma}_{UU}$ is the identity matrix of dimension N, $\mathbf{I}_N$, a sample satisfying the condition, $\bar{\mathbf{x}}_n = \bar{\mathbf{x}}_N$, is called a *(simple) balanced sample*, where $\bar{\mathbf{x}}_n$ is the sample mean and $\bar{\mathbf{x}}_N$ is the population mean. The condition (6) is a condition on the weighted sample mean. To see this, let the first element of $\mathbf{x}$ be equal to one and denote the vector of auxiliary variables by $\mathbf{x}_i = (1,\ \mathrm{x}_{1,i})$. Then the condition (6) on the first element of $\mathbf{x}$ is

$$\frac{\sum_{i\in A}\sigma_{ii}^{-\frac{1}{2}}}{n} = \frac{N}{\sum_{i\in U}\sigma_{ii}^{\frac{1}{2}}}$$

Thus, the sample size n of a weighted balanced sample satisfies

$$n = \frac{1}{N}\left(\sum_{i\in A}\sigma_{ii}^{-\frac{1}{2}}\right)\left(\sum_{i\in U}\sigma_{ii}^{\frac{1}{2}}\right)$$

and a weighted balanced sample satisfies the condition

$$\left(\sum_{i\in A}\sigma_{ii}^{-\frac{1}{2}}\right)^{-1}\left(\sum_{i\in A}\sigma_{ii}^{-\frac{1}{2}}\mathbf{x}_{1,i}\right) = \bar{\mathbf{x}}_{1,N} \tag{8}$$

The condition (8) implies that the weighted sample mean in which weights are proportional to $\sigma_{ii}^{-\frac{1}{2}}$ is equal to the population mean of $\mathbf{x}_1$.

If the model is misspecified, especially if the model for the study variable fails to include a set of important auxiliary variables, say $\mathbf{Z}$, Royall (1992) suggested the selection of a balanced sample which is balanced on the variable **Z** as well as on the original set of auxiliary variables **X** so that the BLUP under the model (1) is still model unbiased under the more general model,

$$\mathbf{E}\{\mathbf{y}_U\} = \mathbf{X}_U\beta + \mathbf{Z}_U\gamma, \quad \mathbf{V}\{\mathbf{y}_U\} = \mathbf{\Sigma}_{UU} \tag{9}$$

Several authors have studied selecting a sample that has preferred properties. Goodman and Kish (1950) suggested a *controlled selection* by which the probabilities of selection for preferred samples are increased. Hájek (1964) introduced *rejective sampling* in which samples that do not meet a specified condition are rejected. Herson (1976) discussed the selection of a balanced sample in which simple random samples are drawn and any that are not sufficiently close to being balanced are rejected.

Valliant, Dorfman, and Royall (2000, Section 3.4.4) defined a *restrictive random sampling plan* using standardized measures of imbalance for a given sample A. Suppose we seek balance on p auxiliary variables $x_1, \cdots, x_p$. Define

$$\Delta_j(A) = \frac{|\sqrt{n}(\bar{x}_{j,n} - \bar{x}_{j,N})|}{S_{x_j,N}} \quad j = 1, \cdots, p$$

where $S^2_{x_j,N} = (N-1)^{-1}\sum_{i\in U}(x_{ij} - \bar{x}_{j,N})^2$. A sample is considered sufficiently close to balance if, for a prescribed constant δ_j,

$$\Delta_j(A) \leq \delta_j \quad \text{for all} \quad j = 1, \cdots, p \tag{10}$$

The steps of balanced sampling are

1. Specify δ_j for $j = 1, \cdots, p$.
2. Select a simple random sample without replacement.
3. Retain the sample if (10) is satisfied; otherwise replace the sample into the population and repeat Step 2.

We will refer the sample that satisfies the condition (10) as a restricted random sample partially balanced on $x_1, \cdots, x_p$. The choice of δ is somewhat arbitrary. Royall and Cumberland (1981) discussed reasonable choices for δ.

3. Inclusion Probabilities for a Restricted Random Sample

The first order inclusion probability π_i for a restricted random sample partially balanced on auxiliary variables is the probability that the i-th element is in the sample conditional on the sample means of auxiliary variables satisfying the specified constraints. For simplicity, assume we have one auxiliary variable and assume that the population mean of the auxiliary variable is equal to zero. Let $\delta > 0$ and assume the sample is rejected unless $|\bar{x}_n| \leq \delta$. The first order inclusion probability for element i is

$$\pi_i = \frac{n}{N}\frac{\Pr\{-n^{-1}x_i - \delta < n^{-1}(n-1)\bar{x}_{(n-1)} < -n^{-1}x_i + \delta\}}{\Pr\{-\delta < \bar{x}_n < \delta\}} \tag{11}$$

where $\bar{x}_{(n-1)}$ is the mean of $n-1$ observations selected from a population of $N-1$ elements with the i-th observation deleted. If we assume approximate normality of the auxiliary variable we can approximate the inclusion probability using the normal distribution function.

$$\hat{\pi}_i = \frac{n}{N}\frac{F_z(b_{i,2}) - F_z(b_{i,1})}{F_z(a_2) - F_z(a_1)} \tag{12}$$

where

$$a_1 = -\left[\sigma_x^2(1-f)\right]^{-\frac{1}{2}} n^{\frac{1}{2}}\delta$$

$$a_2 = \left[\sigma_x^2(1-f)\right]^{-\frac{1}{2}} n^{\frac{1}{2}}\delta$$

$$b_{i,1} = \left[\sigma_x^2(1-f)\right]^{-\frac{1}{2}} (n-1)^{-\frac{1}{2}} n(-n^{-1}x_i - \delta)$$

$$b_{i,2} = \left[\sigma_x^2(1-f)\right]^{-\frac{1}{2}} (n-1)^{\frac{1}{2}} n(-n^{-1}x_i + \delta)$$

σ_x^2 is the population variance of x, $F_z(\cdot)$ is the distribution function of the standard normal distribution. For a given δ and x_i, the approximate inclusion probability in (12) approaches the sampling fraction as $n \to \infty$.

A second approximation for the first order inclusion probability for a restricted random sample can be derived by using the conditional probability suggested by Tillé (1998). Under the approximate normality of $\bar{x}_n$, an approximate unconditional inclusion probability for element i in a sample with sample mean $\bar{x}_n$ is

$$\tilde{\pi}_i = \frac{n}{N}\frac{f(\bar{x}_n | i \in A)}{f(\bar{x}_n)} \quad = \frac{n}{N}\frac{\sigma_{\bar{x},(i)}^{-1}}{\sigma_{\bar{x}}^{-1}} \exp\left(-\frac{d_i}{2}\right) \tag{13}$$

where

$$\sigma_{\bar{x}}^2 = V\{\bar{x}_n | \mathscr{F}\} = (1-f)\frac{S_{x,N}^2}{n}$$

$$\sigma_{\bar{x},(i)}^2 = V\{\bar{x}_n | i \in A, \mathscr{F}\} = \frac{(N-n)(n-1)}{n^2(N-2)}\left\{S_{x,N}^2 - \frac{N(x_i - \bar{x}_N)^2}{(N-1)^2}\right\}$$

$$d_i = \sigma_{\bar{x},(i)}^{-2}\left\{\frac{(N-n)(x_i - \bar{x}_N)}{n(N-1)}\right\}^2$$

$S_{x,N}^2 = (N-1)^{-1}\sum_{j=1}^{N}(x_j - \bar{x}_N)^2, f(\bar{x}_n)$ and $f(\bar{x}_n | i \in A)$ are normal density functions that have means $\bar{x}_N$ and $\bar{x}_N + [n(N-1)]^{-1}(N-n)(x_i - \bar{x}_N)$, and variances $\sigma_{\bar{x}}^2$ and $\sigma_{\bar{x},(i)}^2$, respectively, and $\mathscr{F} = (x_1, \cdots, x_N)$ is the set of values of the auxiliary variable for the finite population. The variance $\sigma_{\bar{x}}^2$ is the design variance of the sample mean and $\sigma_{\bar{x},(i)}^2$ is the conditional design variance of the sample mean conditional on the i-th element being in the sample. If $x_i = \bar{x}_N + c\sqrt{r}$ for a nonzero constant c, then the d_i defined in (13) goes to infinity and the corresponding approximate inclusion probability of x_i approaches zero as r increases. That is, an observation that is far away from the population mean has a small inclusion probability.

By analogy to (12), an approximate second order inclusion probability for elements i and j in a restricted random sample partially balanced on x, for a given $\delta > 0$, is

$$\begin{aligned}\hat{\pi}_{ij} &= \frac{n(n-1)}{N(N-1)} \frac{\Pr\{-n^{-1}x_i - n^{-1}x_j - \delta < n^{-1}(n-2)\bar{x}_{(n-2)} < -n^{-1}x_i - n^{-1}x_j + \delta\}}{\Pr\{-\delta < \bar{x}_n < \delta\}} \\ &= \frac{n(n-1)}{N(N-1)} \frac{F_z(b_{ij,2}) - F_z(b_{ij,1})}{F_z(a_2) - F_z(a_1)}\end{aligned} \tag{14}$$

where

$$b_{ij,1} = \left[\sigma_x^2(1-f)\right]^{-\frac{1}{2}}(n-2)^{-\frac{1}{2}}n(-n^{-1}x_i - n^{-1}x_j - \delta)$$

$$b_{ij,2} = \left[\sigma_x^2(1-f)\right]^{-\frac{1}{2}}(n-1)^{-\frac{1}{2}}n(-n^{-1}x_i - n^{-1}x_j + \delta)$$

and σ_x^2, a_1 and a_2 are defined in (12).

The conditional design expectation and variance of the sample mean given that the i-th and j-th elements are in the sample are

$$\mathrm{E}\{\bar{x}_n | (i,j) \in A, \mathscr{F}\} = \bar{x}_N + \frac{N-n}{n(N-2)}\{(x_i - \bar{x}_N) + (x_j - \bar{x}_N)\} \tag{15}$$

and

$$\sigma^2_{\bar{x},(ij)} = \mathrm{V}\{\bar{x}_n | (i,j) \in A, \mathscr{F}\} = \frac{(n-2)(N-n)(N-1)}{n^2(N-2)(N-3)} \times C \tag{16}$$

where

$$C = S^2_{x,N} - \frac{1}{N-2}\left[(x_i - \bar{x}_N)^2 + (x_j - \bar{x}_N)^2 + \frac{2}{N-1}(x_i - \bar{x}_N)(x_j - \bar{x}_N)\right]$$

and $S^2_{x,N}$ is defined in (13). Then, an approximate second order inclusion probability for a restricted random sample is

$$\tilde{\pi}_{ij} = \frac{n(n-1)}{N(N-1)} \frac{\sigma^{-1}_{\bar{x},(ij)}}{\sigma^{-1}_{\bar{x}}} \exp\left(-\frac{d_{ij}}{2}\right) \tag{17}$$

where

$$d_{ij} = \sigma^{-2}_{\bar{x},(ij)}\left\{\frac{N-n}{n(N-2)}[(x_i - \bar{x}_N) + (x_j - \bar{x}_N)]\right\}^2$$

and $\sigma^2_{\bar{x}}$ and $\sigma^2_{\bar{x},(ij)}$ are as defined in (13) and (16), respectively.

To investigate approximations for the first and second order inclusion probabilities for a restricted random sample, we generated a population of size 1,050 from $N(5, 1)$ and selected 50,000 samples of size 30 using the restricted random sampling plan with $\delta = 0.126$. The $\delta = 0.126$ is chosen so that the sampling procedure rejects about 90% of samples, and provides reasonable balance. See Herson (1976) and Royall and Cumberland (1981). 440,417 samples were rejected to obtain 50,000 samples. That is, the fraction of samples rejected is around 0.90. Figure 1 shows the estimated first order inclusion probabilities and the two approximations plotted against the values of x, where the approximations are defined in (12) and (13). The two approximations for the first order inclusion probabilities are almost the same and both approximate reasonably well the true inclusion probabilities. The average selection probability is 1/35, which is 0.02857. The standard error of the mean of a sample of 50,000 binomial random variables with mean 0.02857 is 0.0007451. In Figure 1, we also plot the approximation plus and minus 1.96 standard errors. About 4.2% of the sample probabilities fall outside the bound.

In a regression of the estimated probability on the approximation without an intercept, weighted by the variance of the estimated probability, the regression coefficient was 1.00 and the standardized residual mean square was 0.97. This result was true for both approximations. If an intercept is included in the regression, we obtain

$$\begin{array}{ccccc} \hat{p} & = & 0.0018 & + & 0.937A_{13} \\ & & (0.0010) & & (0.034) \end{array}$$

for approximation (13), denoted by A_{13}, and

$$\begin{array}{ccccc} \hat{p} & = & 0.0037 & + & 0.871A_{12} \\ & & (0.0009) & & (0.032) \end{array}$$

for approximation (12), denoted by A_{12}. In both cases the standardized residual mean square is 0.97. Approximation (13) is superior in this example, but both approximations perform well. Approximations tend to overestimate small inclusion probabilities.

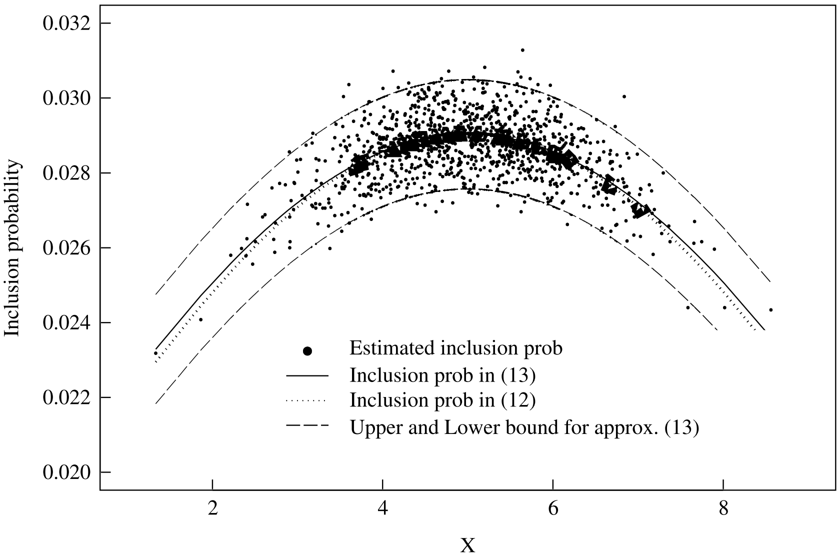

Fig. 1. Estimated first order inclusion probabilities for restricted random samples and their approximations

Figure 2 shows the estimated second order inclusion probabilities corresponding to pairs of elements such that $x_i < x_j$. The second order inclusion probabilities for the pair of elements in which both x-values are quite small or both quite large relative to the population mean are small. For a pair of elements in which one is much smaller than the population mean and the other is larger than the population mean, the second order inclusion probability is relatively large. This is because the balanced sampling scheme forces the sample mean to be close to the population mean.

Regressing the estimated second order inclusion probability on the approximations without intercept gives an estimated regression coefficient of 1.00 for both approximations. Like the approximations of the first order inclusion probability, both approximations (14) and (17) tend to overestimate small second order inclusion probabilities. Among the 5% smallest estimated inclusion probabilities, 99% are overestimated by both approximations.

By assuming approximate multivariate normality for the mean vector of auxiliary variables, approximations of the inclusion probabilities in (13) and (17) can be extended to the case of multiple auxiliary variables. Let the vector of auxiliary variables for the i-th element, $\mathbf{x}_i$, be available for all elements in the population. Approximations for the first and second order inclusion probabilities for a balanced sample are

$$\tilde{\pi}_i = \frac{n}{N}|\boldsymbol{\Sigma}_{\bar{x}\bar{x}}|^{\frac{1}{2}}|\boldsymbol{\Sigma}_{\bar{x}\bar{x},(i)}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(G_{\bar{x}\bar{x},(i)} - G_{\bar{x}\bar{x}})\right\} \tag{18}$$

and

$$\tilde{\pi}_{ij} = \frac{n(n-1)}{N(N-1)}|\boldsymbol{\Sigma}_{\bar{x}\bar{x}}|^{\frac{1}{2}}|\boldsymbol{\Sigma}_{\bar{x}\bar{x},(ij)}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(G_{\bar{x}\bar{x},(ij)} - G_{\bar{x}\bar{x}})\right\} \tag{19}$$

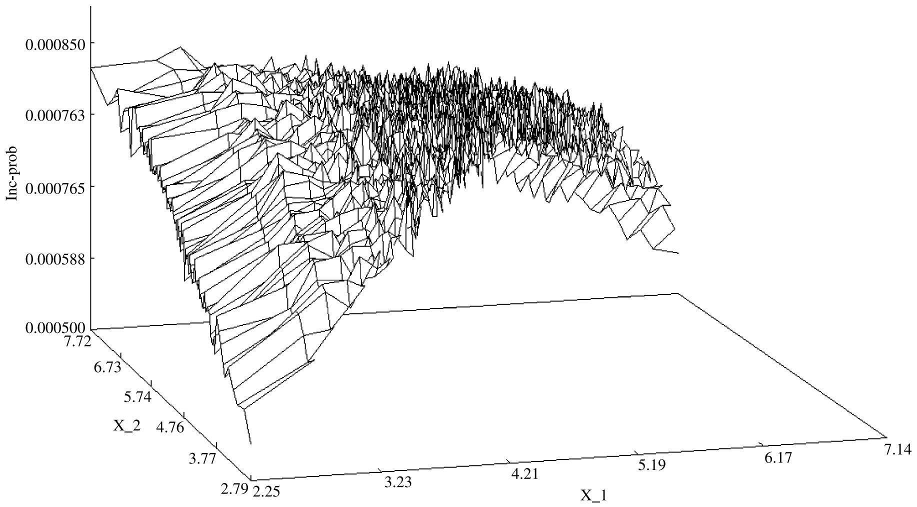

Fig. 2. Estimated second order inclusion probabilities for restricted random samples

respectively, where

$$G_{\bar{x}\bar{x}} = (\bar{\mathbf{x}}_n - \bar{\mathbf{x}}_N)\boldsymbol{\Sigma}_{\bar{x}\bar{x}}^{-1}(\bar{\mathbf{x}}_n - \bar{\mathbf{x}}_N)'; \quad G_{\bar{x}\bar{x},(i)} = (\bar{\mathbf{x}}_{(n-1)} - \bar{\mathbf{x}}_{(N-1)})\boldsymbol{\Sigma}_{\bar{x}\bar{x},(i)}^{-1}(\bar{\mathbf{x}}_{(n-1)} - \bar{\mathbf{x}}_{(N-1)})';$$

$$G_{\bar{x}\bar{x},(ij)} = (\bar{\mathbf{x}}_{(n-2)} - \bar{\mathbf{x}}_{(N-2)})\boldsymbol{\Sigma}_{\bar{x}\bar{x},(ij)}^{-1}(\bar{\mathbf{x}}_{(n-2)} - \bar{\mathbf{x}}_{(N-2)})'$$

$\bar{\mathbf{x}}_{(N-1)} = (N-1)^{-1}(N\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_i)$, $\bar{\mathbf{x}}_{(N-2)} = (N-2)^{-1}(N\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$, $\boldsymbol{\Sigma}_{\bar{x}\bar{x}}$ is the covariance matrix of $\bar{\mathbf{x}}_n$, and $\boldsymbol{\Sigma}_{\bar{x}\bar{x},(i)}$ and $\boldsymbol{\Sigma}_{\bar{x}\bar{x},(ij)}$ are the conditional covariance matrices of the sample mean conditional on $i \in A$ and $i, j \in A$, respectively. Using the constructed approximate inclusion probabilities of a restricted random sample partially balanced on auxiliary variables, the design related properties of an estimator such as design unbiasedness, design consistency and unconditional variance under the design and model can be evaluated approximately.

4. Regression Weights and Inclusion Probabilities for Balanced Sample

In this section, we compare the weights of a regression estimator for a simple random nonreplacement sample and the regression weights for a balanced sample. The regression estimator is the BLUP of (2) under the model (1) and is design consistent if there exists a vector **c** such that

$$\boldsymbol{\Sigma}_{AA}(\mathbf{L}_\pi - \mathbf{J}_n) = \mathbf{X}_A\mathbf{c} \tag{20}$$

where $\mathbf{L}_\pi$ is the vector of the inverse of inclusion probabilities. See also Fuller and Park (2002). With a simple random nonreplacement sample of size n and a single auxiliary variable, the regression estimator can be expressed as

$$\bar{y}_{reg} = \sum_{i \in A} w_{i,reg} y_i \tag{21}$$

where

$$\begin{aligned} w_{i,reg} &= \frac{1}{n} + (\bar{x}_N - \bar{x}_n)\left[\sum_{j=1}^{n}(x_j - \bar{x}_n)^2\right]^{-1}(x_i - \bar{x}_n) \\ &= \frac{1}{n}\left\{1 + (\bar{x}_N - \bar{x}_n)\frac{(x_i - \bar{x}_n)}{n^{-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2}\right\} \end{aligned} \tag{22}$$

The regression weight for an element in the sample defined in (22) can be approximated by a function of the x_i.

Theorem 1. Let $\mathscr{F}_N = \{x_1, \cdots, x_N\}$ be a sequence of finite populations, where $\mathscr{F}_N$ is a random sample of size N from a superpopulation with finite fourth moments. Assume a simple random nonreplacement sample is selected from each $\mathscr{F}_N$. Let μ and σ^2 be the superpopulation mean and variance of x. Then the regression weight for x_i of (22), given that x_i is in the sample, satisfies

$$nw_{i,reg} = 1 + \frac{(x_i - \mu)}{\sigma^2}(\bar{x}_N - \bar{x}_n) + \frac{1}{\sigma^2}(\bar{x}_N - \bar{x}_n)(x_i - \bar{x}_n)$$
$$- \frac{(x_i - \mu)}{\sigma^4}(\bar{x}_N - \bar{x}_n)\left[n^{-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2\right] + O_p\left(n^{-\frac{3}{2}}\right)$$

If, in addition, $\mathscr{F}_N$ is a random sample from the normal distribution, then the conditional expectation of the regression weight given that x_i is in the sample, satisfies

$$\mathrm{E}\{nw_{i,reg}|x_i\} = 1 + \frac{(1-f)(n-1)}{n^2} - \frac{(1-f)(n^2+3n-3)}{n^3\sigma^2}(x_i - \mu)^2$$
$$+ \frac{(1-f)(n-1)}{n^3\sigma^4}(x_i - \mu)^4 + O\left(n^{-\frac{3}{2}}\right) \quad (23)$$

If the regression estimator is constructed for a simple random sample, the weight applied to an observation in the sample is a function of the x-value and of the x-values in the sample. Let $w_{i,reg}$ be the regression weight for element i with $w_{i,reg} = 0$ if $i \notin A$. The conditional expected value of the regression estimator is

$$E\{\bar{y}_{reg}|\mathscr{F}\} = \sum_{i \in U} E\{w_{i,reg}\}y_i \quad (24)$$

We call $E\{w_{i,reg}\}$ the average weight.

To compare the weights for a simple random nonreplacement sample with the regression estimator to the strategy of a balanced sample with the regression estimator, we consider the ratio of the average weight for an observation in the estimator for the population mean to be n^{-1}. We are comparing averages over all possible samples for a particular procedure. For a simple random nonreplacement sample with the regression estimator, the ratio is

$$R_{i,\mathrm{SI}} = n\mathrm{E}\{w_{i,reg}\} \quad (25)$$

where $\mathrm{E}\{w_i|x_i\}$ is defined in (23). For a balanced sample with the regression estimator, the ratio is

$$R_{i,\mathrm{BAL}} = n^{-1}N\hat{\pi}_i \quad (26)$$

where $\hat{\pi}_i$ is as defined in (12). The regression weight for a balanced sample under the model (1) with the identity covariance matrix and a single auxiliary variable is n^{-1} because $\bar{x}_n = \bar{x}_N$ for a balanced sample.

Figure 3 shows the two ratios (25) and (26) plotted against x for the population used to generate Figure 1 of Section 3. In both cases, the weights for the elements near the center are increased and the weights corresponding to extreme x values are decreased relative to n^{-1}. The shapes of the two relative weight functions are similar but a simple random nonreplacement sample with the regression estimator gives a wider range in the ratios than the strategy of a balanced sample with the regression estimator.

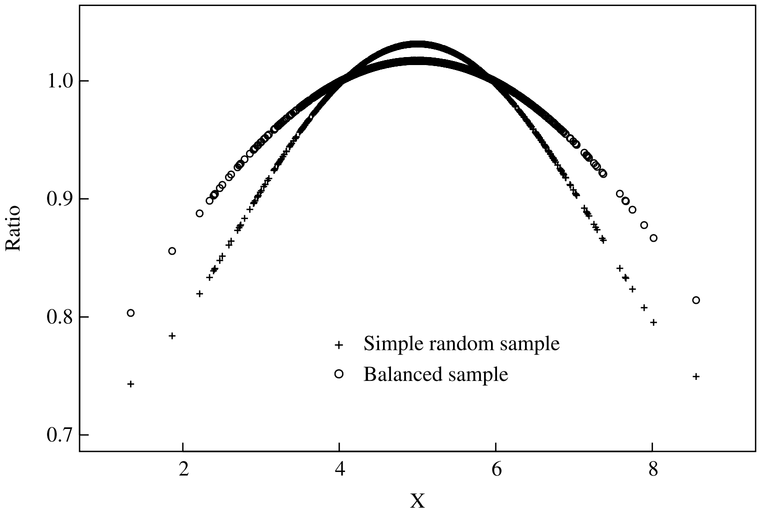

Fig. 3. Approximate relative average weights for a simple random nonreplacement sample with the regression estimator and for a balanced sample with the regression estimator

5. Restricted Random Sampling and Stratified Random Sampling

In Section 2, we introduced the result due to Royall (1992) that the model variance of the BLUP under model (1) is minimized by selecting a sample with sample mean of x equal to the population mean of x. Balancing on additional variables as well as on the regressors of the assumed model provides model bias-robustness against the linear effects of the additional variables.

Stratified random sampling with strata formed on the basis of the auxiliary variable is another method of selection that produces an approximately balanced sample. The deviation of the selected sample mean from the population mean is not perfectly controlled in stratified random sampling. Thus, the model variance of the regression estimator depends on the imbalance of the stratified simple random sample. But with a stratified simple random sample, we can obtain the exact inclusion probabilities and we can construct a model based regression estimator that is the BLUP under the assumed model and is design consistent (see Fuller and Park 2002). A design consistent estimator has the property of robustness to model failure in the sense that the estimator approaches the true population characteristic as the sample and population sizes increase.

Let the population be sorted on the auxiliary variable x and let strata h, $h = 1, \cdots, H$, be formed, equalizing the number of units $N_1 = \cdots = N_H = N_0$ in each stratum. Assume that simple random nonreplacement samples of size $n_1 = \cdots = n_H = n_0$ are selected from each stratum. Assume the model

$$\mathbf{E}\{\mathbf{y}_U\} = \mathbf{X}_U\beta, \quad \mathbf{V}\{\mathbf{y}_U\} = \sigma^2\mathrm{I}_N \tag{27}$$

where

$$\mathbf{X}_U = (\mathbf{J}_N, \mathbf{x}_{1,U}), \quad \mathbf{x}_{1,U} = (x_{1,1}, \cdots, x_{1,N})', \quad \beta = (\beta_0, \beta_1)'$$

$\mathbf{y}_U$ is the vector of y for the population, $\mathbf{I}_N$ is the identity matrix of dimension N and $\mathbf{J}_N$ is the column of ones with length N.

Under model (27), the BLUP for the population mean is

$$\bar{y}_{BLUP} = \bar{\mathbf{x}}_N\hat{\beta} \quad = \bar{y}_n + (\bar{x}_{1,N} - \bar{x}_{1,n})\hat{\beta}_1 \quad = \bar{y}_{reg} \tag{28}$$

where

$$\bar{\mathbf{x}}_N = N^{-1}\sum_{i=1}^{N}(1, x_{1,i}), \quad (\bar{y}_n,\, \bar{x}_{1,n}) = n^{-1}\sum_{i=1}^{n}(y_i, x_{1,i}),$$

$$\hat{\beta} = (\hat{\beta}_0,\, \hat{\beta}_1)' = \left[\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i\right]^{-1}\sum_{i=1}^{n}\mathbf{x}_i'y_i$$

and $n = n_0H$ is the sample size. Because the model and stratified random sampling satisfy condition (20), the regression estimator is equivalent to the BLUP. With a perfectly balanced sample under the model (27), the BLUP is the sample mean.

Because the first element of $\mathbf{x}$ is equal to one, the conditional model variance of the BLUP is

$$\begin{aligned}\mathbf{V}\{\bar{y}_{reg} - \bar{y}_N|\mathbf{X}_U\} &= \bar{\mathbf{x}}_N\mathbf{V}\{\hat{\beta}|\mathbf{X}_U\}\bar{\mathbf{x}}_N' + \mathbf{V}\{\bar{y}_N|\mathbf{X}_U\} - 2\mathrm{Cov}\{\bar{\mathbf{x}}_N\hat{\beta}, \bar{y}_N|\mathbf{X}_U\}\\ &= \sigma^2\bar{\mathbf{x}}_N\left(\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\bar{\mathrm{x}}_N' - \frac{\sigma^2}{N}\\ &= \sigma^2\left[\frac{1}{n} - \frac{1}{N} + \frac{(\bar{x}_{1,n} - \bar{x}_{1,N})^2}{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_{1,n})^2}\right]\end{aligned} \tag{29}$$

The conditional model variance of the BLUP for a perfectly balanced sample is

$$\mathbf{V}\{\bar{y}_{reg} - \bar{y}_N|\mathbf{X}_U, \bar{x}_n = \bar{x}_N\} = \sigma^2\left[\frac{1}{n} - \frac{1}{N}\right] \quad =: V_B \tag{30}$$

because $\bar{x}_{1,n} = \bar{x}_{1,N}$ for a balanced sample.

The model relative efficiency of the BLUP with stratified simple random sampling to the one with a perfectly balanced sample is

$$\frac{\mathrm{V}\{\bar{y}_{reg} - \bar{y}_N|\mathbf{X}_U\}}{V_B} = 1 + \frac{(\bar{x}_{1,n} - \bar{x}_{1,N})^2}{(1-f)\left[n^{-1}\sum_{i=1}^{n}(x_{1,i} - \bar{x}_{1,n})^2\right]} \quad =: 1 + \gamma_n \tag{31}$$

where $f = N^{-1}n < 1$ and a perfectly balanced sample has $\bar{x}_{1,n} = \bar{x}_{1,N}$. Under stratified simple random sampling, γ is design consistent for zero in that

$$\gamma_n|\mathscr{F}_N = O_p(n^{-1})$$

Therefore, using this stratified sampling design, the strategy of a stratified simple random sample with the regression estimator, that is the BLUP under the model, is approximately as model efficient as the strategy of a balanced sample with the BLUP.

To illustrate the difference between selection strategies, we consider the population that was generated in Section 3. To select a stratified simple random sample, the population was sorted on x and 15 equal-sized strata of size 70 were formed. We selected 50,000 stratified simple random samples of size 30 by selecting a simple random nonreplacement sample of size 2 from each stratum. We chose this simple procedure, but one could use a more complicated allocation and (or) stratum definition procedure based on within stratum variance.

We also selected 50,000 samples of size 30 partially balanced on $\sqrt{x}$, x and x^2 by the restricted random sampling plan with $\delta = 0.126$. The same procedure was used in Dorfman and Valliant (2000) so that at least 10% of the best-balanced samples are obtained. That is we selected samples which satisfied the conditions

$$\left| \frac{\sqrt{30}\left(\bar{x}_n^{(0.5)} - \bar{x}_N^{(0.5)}\right)}{S_{(0.5)}} \right| < 0.126$$

$$\left| \frac{\sqrt{30}(\bar{x}_n - \bar{x}_N)}{S} \right| < 0.126$$

$$\left| \frac{\sqrt{30}\left(\bar{x}_n^{(2)} - \bar{x}_N^{(2)}\right)}{S_{(2)}} \right| < 0.126$$

where

$$\bar{x}_n^{(j)} = n^{-1}\sum_{i=1}^{n} x_i^j, \quad \bar{x}_N^{(i)} = N^{-1}\sum_{i=1}^{N} x_i^j$$

and

$$S_{(j)} = \left[(N-1)^{-1} \sum_{i=1}^{N} \left(x_i^j - \bar{x}_N^{(j)} \right)^2 \right]^{\frac{1}{2}}$$

for $j = 0.5, 1, 2$. The fraction of samples rejected is 0.96. 1,126,087 samples were rejected to obtain 50,000 samples. The means of the 50,000 stratified simple random samples and 50,000 restricted random samples partially balanced on $\sqrt{x}$, x and x^2 are 5.0000366 and 5.0003616, respectively. The simulation variance of the sample mean of x for stratified simple random sampling is 0.00089 and the variance of the sample mean for restricted random sampling is 0.00010.

Figure 4 shows the estimated inclusion probabilities for a restricted random sample partially balanced on $\sqrt{x}$, x and x^2. The elements that have large absolute deviation, $|x_i - \bar{x}_N|$, have small inclusion probabilities, as we observed in Figure 1. The range of estimated inclusion probabilities for a restricted random sample partially balanced on $\sqrt{x}$, x and x^2 is $(0.00516, 0.03172)$, which is much wider than that of inclusion probabilities for a restricted random sample partially balanced on x only, $(0.02318, 0.03128)$. By balancing on the additional auxiliary variables $\sqrt{x}$ and x^2, inclusion probabilities for the elements that are far from the population mean are

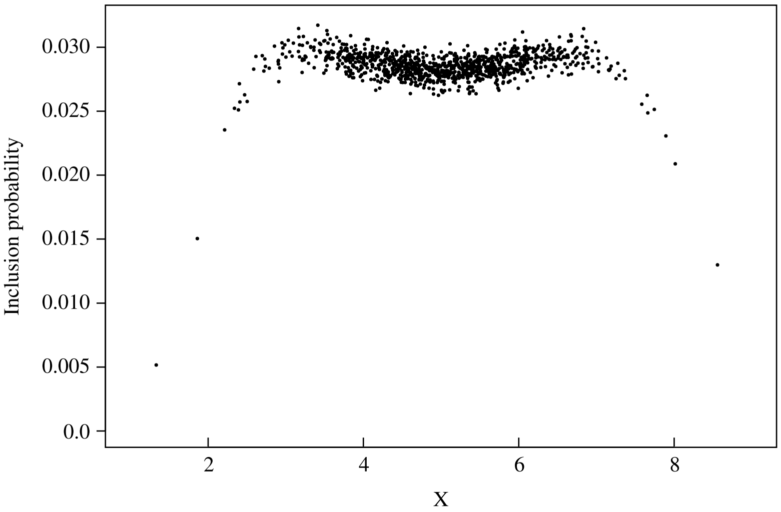

Fig. 4. Estimated inclusion probabilities for a restricted random sample partially balanced on x, $\sqrt{x}$ and x^2

extremely small. Ten percent of estimated probabilities differ from the mean probability of 0.02857 by more than 0.0015.

Table 1 shows the summary statistics for 50,000 γ-values for stratified simple random samples where γ is defined in (31). The maximum loss of efficiency under the assumed model due to selecting a stratified simple random sample rather than a perfectly balanced sample is 1.4%. Because the sample selected by restricted random sampling does not satisfy the condition $\bar{x}_n = \bar{x}_N$, the actual γ for the BLUP with the restricted random sampling is not zero. The maximum γ for the BLUP with restricted random sampling is 0.06%.

In our finite population, the possible maximum model variance for a stratified sample defined in (29) is 0.03355 and occurs if the two observations selected are the largest x-values in each stratum. The corresponding γ is 0.035222. Thus, the maximum possible loss is 3.5%. The BLUP for the stratified simple random samples has approximately the same efficiency as the BLUP with a balanced sample.

One can apply restricted sampling to the stratified design. For example, rejecting samples with an efficiency loss greater than 0.5% would result in rejection of 1.5% of the stratified samples.

In multipurpose surveys, it is common practice to use regression to construct a single set of weights to be used in all analyses. In such cases, the set of weights that are BLUP for a particular variable need not give the BLUP for another variable. We study the performance of the regression estimator under alternative designs for the estimation of points on the cumulative distribution function of a variable y closely related to x.

Table 1. Summary statistics of γ for 50,000 stratified simple random samples

	Minimum	.25 Quantile	Median	Mean	.75 Quantile	Maximum
$\gamma \times 100$	0.000	0.008	0.036	0.083	0.108	1.400

Let $y_i = x_i + e_i$, where $e_i \sim N(0, 0.05)$ and x_i is the observation from the population of auxiliary variable generated in Section 3. Let q_j be the value satisfying

$$P_j = \Pr\{y \leq q_j\} = 0.01j$$

for $j = 5$, 10, 25, 50, 75, 90, and 95. The regression estimator with a stratified random sample and the regression estimator with restricted random samples are considered. For the restricted random sample, we consider a restricted random sample partially balanced on x only and a restricted random sample partially balanced on $\sqrt{x}$, x and x^2.

The regression estimator of the population parameter P_j is

$$\hat{P}_{j,reg} = \sum_{i=1}^{n} \left\{ \frac{1}{n} + (\bar{x}_N - \bar{x}_n) \left[\sum_{k=1}^{n} (x_k - \bar{x}_n)^2 \right]^{-1} (x_i - \bar{x}_n) \right\} I_{j,i} \tag{32}$$

where

$$I_{j,i} = \begin{cases} 1 & \text{if } y_i \leq q_j \\ 0 & \text{otherwise.} \end{cases}$$

We consider the following three strategies:

1. Regression estimator with a stratified simple random nonreplacement sample. (Stratified random sample)
2. Regression estimator with a restricted random sample partially balanced on x. (Partially balanced on x)
3. Regression estimator with a restricted random sample partially balanced on $\sqrt{x}$, x and x^2. (Partially balanced on $\sqrt{x}$, x, x^2)

Figure 5 shows the estimated relative design biases of the estimated cumulative distribution function for the three strategies where the relative bias is

$$\text{Relative Bias} = \frac{E\{\hat{P}_j | \mathscr{F}\} - P_j}{\min(P_j, 1 - P_j)} \tag{33}$$

For all P_j, the stratified random sample has the smallest absolute bias. The restricted random sample partially balanced on x severely underestimates the true values for P_5, P_{10}, and P_{25}. As we observed in Section 3, for a restricted random sample partially balanced on x, the observations far from the mean have small inclusion probability so that the regression estimator with a restricted random sample partially balanced on x underestimates the small values of P.

Figure 6 shows the relative design MSE of the restricted random sample partially balanced on x and of the restricted random sample partially balanced on $\sqrt{x}$, x and x^2 relative to the MSE of the stratified random sample. With respect to MSE, the stratified random sample has better performance than the restricted random samples for all P_j. One reason for this phenomenon is the smaller variability of the weights for the stratified sample relative to that for the restricted random samples partially balanced on auxiliary

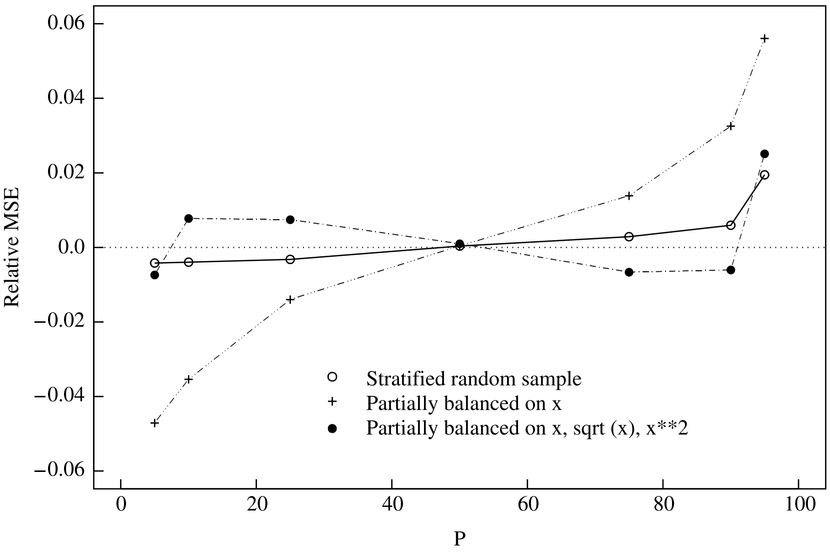

Fig. 5. Monte Carlo relative design biases of the estimated proportion for the three strategies

variables. That is, the ratio of the regression weight to the sampling weight within each stratum, is relatively stable.

To compare the three strategies for a skewed finite population, we generated a finite population of size 1,050 as a sample from the chi-squared distribution with two degrees of freedom. Samples were selected as described in Section 5 and the same three strategies compared. Figure 7 shows the estimated relative design biases of the estimated cumulative distribution functions of y for the three strategies for the finite population, where $y_i = x_i + e_i$, $e_i \sim N(0, 0.05)$ and x_i is from the chi-squared distribution with two

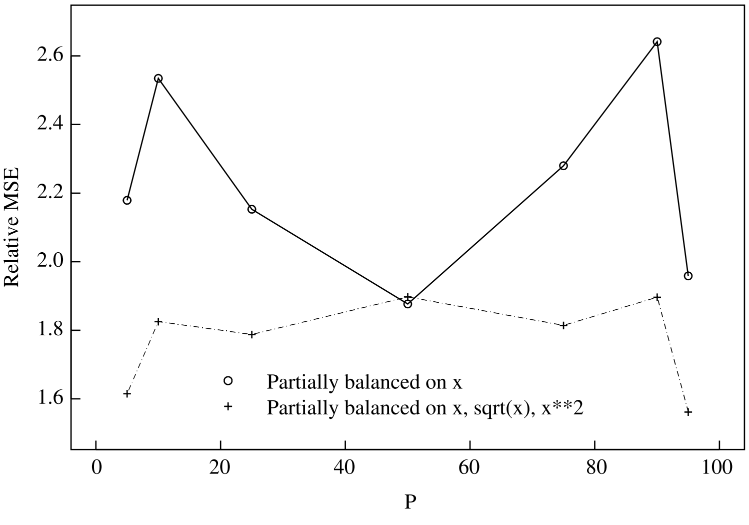

Fig. 6. Monte Carlo relative design MSE of the estimated proportions for two design strategies relative to stratified random sample

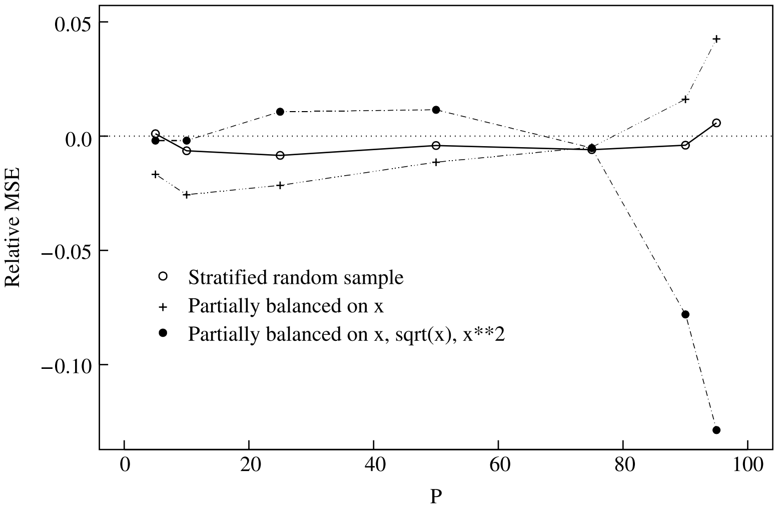

Fig. 7. Monte Carlo relative design biases of the estimated proportion for the three strategies: Skewed population

degrees of freedom. The relative bias is defined in (33). Except for P_{10} and P_{75}, the stratified random sample has the smallest absolute bias. For P_{10}, the stratified random sample and the restricted random sample partially balanced on $\sqrt{x}$, x and x^2 have smaller absolute bias than the restricted random sample partially balanced on x. All three strategies are comparable for P_{75}. Unlike other strategies, the restricted random sample partially balanced on $\sqrt{x}$, x, and x^2 severely underestimates the true population values P_{90} and P_{95}. This is because, for this skewed finite population, the chance that elements with large x-values are selected is extremely small for a restricted random sample partially balanced on $\sqrt{x}$, x, and x^2.

Figure 8 shows the design MSE of the restricted random sample partially balanced on x and of the restricted random sample partially balanced on $\sqrt{x}$, x and x^2 relative to the stratified random sample for the skewed population. With respect to MSE, the stratified random sample has much better performance for the skewed population, with relative efficiencies 114% to 310%.

Means, variances, biases and MSEs of the regression estimator corresponding to different designs, population parameters are given in the Appendix for the two finite populations.

6. Discussion

Restricted random sampling partially balanced on auxiliary variables and stratified random sampling as designs for the regression estimator are compared. Through simulation, we found that the median loss in efficiency for the mean of y made by selecting a stratified sample instead of a perfectly balanced sample is 0.036% for a normal population. The maximum possible loss for the mean is 3.5% and the mean loss is 0.083% for the illustrative finite population selected from a normal population. In estimating the population distribution function of a variable y, strongly correlated with the auxiliary

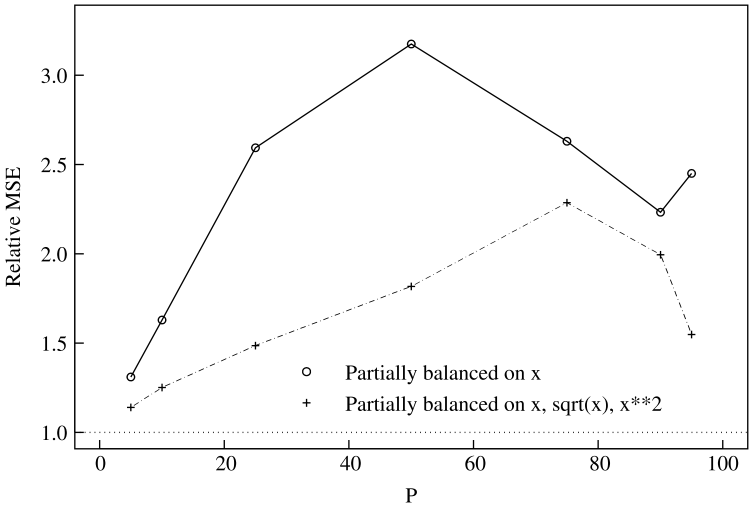

Fig. 8. Monte Carlo relative design MSE of the estimated proportions for two strategies relative to stratified random sample: Skewed population

variable, the stratified random sample shows much better performance than the restricted random samples partially balanced on auxiliary variables. This was true for both symmetric and skewed finite populations.

In a large-scale survey, many variables of interest are obtained, and one set of weights typically is used. It is difficult to construct a model that is appropriate for all possible variables, especially for dichotomous variables. For such a situation, the stratified random sample is more robust for the regression estimation than the restricted random sampling.

It seems that some restrictions are placed on samples in practice. That is, few practitioners would retain a stratified sample that contained the largest elements in each stratum. Our investigation suggests that the imposition of modest restriction on the stratified random sample will have modest effects on the selection probabilities. For example, to reduce the largest loss associated with a stratified sample from 3.5% to 0.5% requires rejection of 1.5% of the stratified samples.

Appendix

Table 2. Properties of estimated proportion for the regression estimator with a stratified simple random nonreplacement sample of size 30 for finite population generated from the normal distribution

Variable	Mean	Bias × 10^3	Variance × 10^3	MSE × 10^3
P5	0.0498	−0.209	0.548	0.548
P10	0.0996	−0.397	0.783	0.783
P25	0.2492	−0.802	1.422	1.423
P50	0.5002	0.187	1.639	1.639
P75	0.7507	0.716	1.307	1.307
P90	0.9006	0.595	0.743	0.744
P95	0.9510	0.973	0.605	0.606

Table 3. Properties of estimated proportion for the regression estimator with a restricted random sample of size 30 partially balanced on x for finite population generated from the normal distribution

Variable	Mean	Bias × 10^3	Variance × 10^3	MSE × 10^3
P5	0.0476	−2.356	1.188	1.193
P10	0.0965	−3.539	1.972	1.984
P25	0.2465	−3.507	3.051	3.064
P50	0.5002	0.195	3.075	3.076
P75	0.7535	3.455	2.969	2.981
P90	0.9033	3.255	1.954	1.964
P95	0.9528	2.803	1.180	1.188

Table 4. Properties of estimated proportion for the regression estimator with a restricted random sample of size 30 partially balanced on $\sqrt{x}$, x and x^2 for finite population generated from the normal distribution

Variable	Mean	Bias × 10^3	Variance × 10^3	MSE × 10^3
P5	0.0496	−0.371	0.884	0.885
P10	0.1008	0.777	1.428	1.429
P25	0.2519	1.855	2.540	2.543
P50	0.5005	0.472	3.109	3.109
P75	0.7483	−1.663	2.368	2.371
P90	0.8994	−0.606	1.410	1.410
P95	0.9513	1.255	0.945	0.947

Table 5. Properties of estimated proportion for the regression estimator with a stratified simple random nonreplacement sample of size 30 for finite population generated as a sample from the exponential distribution

Variable	Mean	Bias × 10^3	Variance × 10^3	MSE × 10^3
P5	0.0501	0.054	1.104	1.104
P10	0.0994	−0.639	1.588	1.588
P25	0.2479	−2.111	1.789	1.793
P50	0.4979	−2.064	1.319	1.323
P75	0.7485	−1.482	0.742	0.745
P90	0.8996	−0.400	0.573	0.574
P95	0.9503	0.290	0.366	0.366

Table 6. Properties of estimated proportion for the regression estimator with a restricted random sample of size 30 partially balanced on x for finite population generated as a sample from the exponential distribution

Variable	Mean	Bias × 10^3	Variance × 10^3	MSE × 10^3
P5	0.0492	−0.835	1.446	1.446
P10	0.0974	−2.562	2.581	2.588
P25	0.2446	−5.384	4.622	4.651
P50	0.4943	−5.710	4.169	4.201
P75	0.7488	−1.218	1.957	1.958
P90	0.9016	1.616	1.278	1.280
P95	0.9521	2.127	0.892	0.897

Table 7. Properties of estimated proportion for the regression estimator with a restricted random sample of size 30 partially balanced on $\sqrt{x}$, x and x^2 for finite population generated as a sample from the exponential distribution

Variable	Mean	Bias × 10^3	Variance × 10^3	MSE × 10^3
P5	0.0499	−0.097	1.258	1.258
P10	0.0998	−0.197	1.987	1.987
P25	0.2527	2.671	2.657	2.664
P50	0.5058	5.769	2.372	2.405
P75	0.7487	−1.307	1.700	1.702
P90	0.8922	−7.803	1.083	1.144
P95	0.9436	−6.432	0.525	0.567

Proof of Theorem 1. The regression weight of (22) is a function of $(\bar{x}_N - \bar{x}_n)$, $(x_i - \bar{x}_n)$ and $n^{-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2$. These statistics can be formulated as

$$\begin{aligned} \bar{x}_N - \bar{x}_n &= (1-f)\left[\bar{x}_{(N-n)} - \bar{x}_{(n-1)} - \frac{1}{n}(x_i - \bar{x}_{(n-1)})\right] \\ x_i - \bar{x}_n &= \left(1 - \frac{1}{n}\right)(x_i - \bar{x}_{(n-1)}) \end{aligned} \tag{34}$$

and

$$n^{-1}\left[\sum_{j=1}^{n}(x_j - \bar{x}_n)^2\right] = \frac{n-1}{n^2}(x_i - \bar{x}_{(n-1)})^2 + \frac{1}{n}\sum_{\substack{j=1 \\ j \neq i}}^{n}[x_j - \bar{x}_{(n-1)}]^2$$

where

$$\bar{x}_{(n-1)} = \frac{1}{n-1}\sum_{\substack{j=1 \\ j \neq i}}^{n} x_j$$

and

$$\bar{x}_{(N-n)} = \frac{1}{N-n}(N\bar{x}_N - n\bar{x}_n)$$

For a superpopulation with finite fourth moments, we have

$$\bar{x}_N - \bar{x}_n | x_i = O_p(n^{-\frac{1}{2}}), \quad x_i - \bar{x}_n | x_i = (x_i - \mu) + O_p(n^{-\frac{1}{2}}),$$

and

$$n^{-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2 \Big| x_i = \sigma^2 + O_p(n^{-\frac{1}{2}})$$

because $\mathrm{E}\{(\bar{x}_N - \bar{x}_n)^2 | x_i\} = O(n^{-1})$, $\mathrm{E}\{[(x_i - \bar{x}_n) - (x_i - \mu)]^2 | x_i\} = O(n^{-1})$, $\mathrm{E}\{[n^{-1}\sum_{\substack{j=1 \\ j \neq i}}^{n}(x_j - \bar{x}_{(n-1)})^2 - \sigma^2]^2 | x_i\} = O(n^{-1})$ and $n^{-2}(n-1)(x_i - \bar{x}_{(n-1)})^2 = O_p(n^{-1})$. By applying the Taylor expansion to $n(x_i - \bar{x}_n)[\sum_{j=1}^{n}(x_j - \bar{x}_n)^2]^{-1}$ as a function of $x_i - \bar{x}_n$

and $n^{-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2$ about $(x_i - \mu, \sigma^2)$, we obtain

$$\frac{n(x_i - \bar{x}_n)}{\sum_{j=1}^{n}(x_j - \bar{x}_n)^2}\Bigg|x_i = \frac{(x_i - \bar{x}_n)}{\sigma^2} - \frac{(x_i - \mu)}{\sigma^4}\left[n^{-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2 - \sigma^2\right] + O_p(n^{-1})$$

and

$$\begin{aligned} nw_i|x_i = 1 &+ \frac{(x_i - \mu)}{\sigma^2}(\bar{x}_N - \bar{x}_n) + \frac{(\bar{x}_N - \bar{x}_n)(x_i - \bar{x}_n)}{\sigma^2} \\ &- \frac{(x_i - \mu)}{\sigma^4}(\bar{x}_N - \bar{x}_n)\left[n^{-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2\right] + O_p\left(n^{-\frac{3}{2}}\right) \end{aligned} \tag{35}$$

Under the normality assumption, the conditional expectation of the regression weight is

$$\begin{aligned} \mathrm{E}\{nw_i|x_i\} = 1 &+ \mathrm{E}\left\{\frac{(x_i - \mu)}{\sigma^2}(\bar{x}_N - \bar{x}_n)|x_i\right\} + \mathrm{E}\left\{\frac{(\bar{x}_N - \bar{x}_n)(x_i - \bar{x}_n)}{\sigma^2}|x_i\right\} \\ &- \mathrm{E}\left\{\frac{(x_i - \mu)}{\sigma^4}(\bar{x}_N - \bar{x}_n)\left[n^{-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2\right]|x_i\right\} + O\left(n^{-\frac{3}{2}}\right) \end{aligned} \tag{36}$$

because x and $(x_j - \bar{x}_n)^2$ have finite r-th moments for all integers $0 < r < \infty$ by the normality assumption on x and because w_i is a continuous and differentiable function of the means of x and $(x_j - \bar{x}_n)^2$. See Fuller (1996, Theorem 5.4.3). The conditional expectation of $(\bar{x}_N - \bar{x}_n)(x_i - \bar{x}_n)$ is

$$\mathrm{E}\{(\bar{x}_N - \bar{x}_n)(x_i - \bar{x}_n)|x_i\} = -\frac{1-f}{n}\left(1 - \frac{1}{n}\right)[(x_i - \mu)^2 - \sigma^2] \tag{37}$$

By utilizing the moments of the normal and χ^2 distributions, we obtain

$$\begin{aligned} &\mathrm{E}\left\{(\bar{x}_N - \bar{x}_n)\left[n^{-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2\right]|x_i\right\} \\ &\quad = -\frac{(1-f)(n-1)}{n^3}[(x_i - \mu)^3 + (n-3)(x_i - \mu)\sigma^2] \end{aligned} \tag{38}$$

because

$$\mathrm{E}\{\bar{x}_{(n-1)}(x_i - \bar{x}_{(n-1)})^2 | x_i\} = \mu(x_i - \mu)^2 - \frac{\sigma^2}{n-1}(2x_i - 3\mu)$$

$$\mathrm{E}\{\bar{x}_{(N-n)}(x_i - \bar{x}_{(n-1)})^2 | x_i\} = \mu\left[(x_i - \mu)^2 + \frac{\sigma^2}{n-1}\right]$$

$$\mathrm{E}\{x_i(x_i - \bar{x}_{(n-1)})^2 | x_i\} = x_i\left[(x_i - \mu)^2 + \frac{\sigma^2}{n-1}\right]$$

$$\mathrm{E}\left\{\bar{x}_{(N-n)}\sum_{\substack{j=1\\ j\neq i}}^{n}(x_j - \bar{x}_{(n-1)})^2 | x_i\right\} = (n-2)\mu\sigma^2$$

$$\mathrm{E}\left\{x_i\sum_{\substack{j=1\\ j\neq i}}^{n}(x_j - \bar{x}_{(n-1)})^2 | x_i\right\} = (n-2)x_i\sigma^2 \quad \text{and}$$

$$\mathrm{E}\left\{\bar{x}_{(n-1)}\sum_{\substack{j=1\\ j\neq i}}^{n}(x_j - \bar{x}_{(n-1)})^2 | x_i\right\} = (n-2)\mu\sigma^2$$

The result (23) follows from (36), (37), and (38).

7. References

Cochran, W.G. (1977). Sampling Techniques (3rd ed.) New York: Wiley.

Dorfman, A.H. and Valliant, R. (2000). Stratification by Size Revisited. Journal of Official Statistics, 16, 139–154.

Fuller, W.A. (1996). Introduction to Statistical Time Series. (2nd ed.) New York: Wiley.

Fuller, W.A. and Park, M. (2002). Model Weights for Regression Estimation. Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM], Alexandria, VA.

Goodman, R. and Kish, L. (1950). Controlled Selection – A Technique in Probability Sampling. Journal of the American Statistical Association, 45, 350–372.

Hájek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population. Annals of Mathematical Statistics, 35, 1491–1523.

Herson, J. (1976). An Investigation of Relative Efficiency of Least-squares Prediction to Conventional Probability Sampling Plans. Journal of the American Statistical Association, 71, 700–703.

Isaki, C.T. and Fuller, W.A. (1982). Survey Design Under the Regression Superpopulation Model. Journal of the American Statistical Association, 77, 89–96.

Robinson, P.M. and Särndal, C.E. (1983). Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling. Sankhyā B, 45, 240–248.

Royall, R.M. (1992). Robust and Optimal Design Under Prediction Models for Finite Populations. Survey Methodology, 18, 179–185.

Royall, R.M. and Cumberland, W.G. (1981). An Empirical Study of the Ratio Estimator and Estimators of Its Variance. Journal of the American Statistics Association, 76, 66–77.

Royall, R.M. and Herson, J. (1973). Robust Estimation in Finite Populations. Journal of the American Statistical Association, 68, 880–889.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Tillé, Y. (1998). Estimation in Surveys Using Conditional Probabilities: Simple Random Sampling. International Statistical Review, 66, 303–322.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). Finite Population Sampling and Inference – A Prediction Approach. New York: Wiley.

Received September 2003
Revised March 2006