Sankhyā : The Indian Journal of Statistics
2004, Volume 66, Part 4, pp 791-802

Extending Sitter's Mirror-Match Bootstrap to Cover Rao-Hartley-Cochran Sampling in Two-stages with Simulated Illustrations

Arijit Chaudhuri
Indian Statistical Institute, Kolkata, India
Amitava Saha
Directorate General of Mines Safety, Dhanbad, India

Abstract

A practical problem of estimating the total area under cultivation in Indian districts is addressed by two-stage sampling with unequal selection-probabilities. To assess the accuracy in estimation bootstrap technique is employed in constructing confidence intervals and simulation-based performance criteria are evaluated from live-data as are shown for competitive procedures. Rao-Hartley-Cochran's (RHC, 1962) scheme is employed in both stages of sampling. Sitter's mirrormatch bootstrap procedure is employed suitably modifying it to cover the two-stages.

AMS (2000) *subject classification.* 62D05.
Keywords and phrases. Bootstrap, confidence intervals, correlation and regression estimation, generalized regression estimation, mirror-match, two-stage sampling, unequal selection-probabilities.

1 Introduction

In large-scale surveys a population is usually split up into a number of non-overlapping strata. From each stratum sampling is then implemented in practice mostly in two stages. In estimating a population total one is then interested to assess the accuracy by evaluating the coefficient of variation of an estimator, the standard error of the latter and also the length of a confidence interval around the point estimator. A handy way to construct a confidence interval (CI) with a pre-assigned confidence coefficient (CC) for it is to regard the pivotal, which is the 'estimator minus the population total' divided by the 'estimated standard error' as a standard normal deviate. An alternative course, bypassing this normality assumption, is to employ the bootstrap technique which involves repeated sampling from the realized sample itself in a suitable manner so as to visualize the sampling distribution of the original statistic taken as the initial point estimator. Using the repeatedly drawn samples called bootstrap samples it is possible to evaluate standard error and confidence intervals with pre-assigned CC's even starting with a nonlinear initial point estimator for a total. We illustrate here mainly with Cassel, Särndal and Wretman's (CSW, 1976) generalized regression (greg) estimator in what follows. In Section 2 below, we present details of sampling, estimation and bootstrap methods employed here. In Section 3, simulation-based performance characteristics evaluated are shown to indicate competitive estimation procedures. In Section 4, we present our concluding remarks. In our presentation we shall skip the stratification step because each stratum may be treated as a population avoiding complexities in illustration.

Sitter's (1992) mirror-match technique is useful to construct bootstrap samples from an initial sample chosen by Rao-Hartley-Cochran's (RHC, 1962) scheme. We slightly extend it here to RHC scheme applied in two stages. In Indian National Sample Surveys (NSS) the first stage units are chosen by circular systematic sampling (CSS) with probabilities proportional to sizes (PPS) but the second stage sampling is by "equal probability CSS"-scheme. So there is a curiosity on how RHC scheme in two stages may fare in NSS. The Indian Statistical Institute (ISI) employs RHC schemes in two stages in many relatively moderate surveys. A latest example from ISI surveys (ISI, 2003) employed a five-stage sampling with RHC sampling in the first stage followed by simple random sampling without replacement (SRSWOR) in the later four stages.

2 Sampling, Bootstrapping and Estimation

Let $U = (1, \ldots, i, \ldots, N)$ denote a finite population of N first stage units (fsu), with the ith fsu in its turn composed of M_i second stage units (ssu). Let the jth ssu in the ith fsu have the value y_{ij} on a real variable of interest y with $y_i = \sum_2 y_{ij}$ as the total of y for the ith fsu, denoting by $\sum_2$ the sum over the second subscript j of y over its range from 1 to M_i. Also, let $\sum_1 = \sum$ denote the sum over the first subscript of y over the range of i from 1 to N. Then, letting $Y = \sum_1 \sum_2 y_{ij} = \sum_1 y_i$ our problem addressed here is to estimate Y.

We suppose that $p_i(0 < p_i < 1, \sum p_i = 1)$ are the normed size-measures known for the fsu's $i = 1, \ldots, N$. In order to choose a sample of n fsu's from U by Rao, Hartley and Cochran's (RHC, 1962) scheme first certain positive integers N_i are to be specified subject to $\sum_n N_i = N$, writing $\sum_n$ as the sum over n non-overlapping groups into which U is randomly divided taking N_i fsu's in the ith group, $i = 1, \ldots, n$. Writing Q_i as the sum of the p_i-values for the N_i units falling in the ith group one unit is chosen from the ith group with a probability equal to its p_i-value divided by Q_i. Writing p_i, y_i as the normed size-measure and the y-value for the unit chosen from the ith group Y is unbiasedly estimated by the RHC estimator

$$t_r = \sum\nolimits_n y_i \frac{Q_i}{p_i}$$

if y_i's were ascertainable.

Here we assume that y_i's are not ascertainable. So, from the M_i ssu's in the ith fsu, if sampled, we decide to take a sample of m_i ssu's again employing the RHC scheme assuming that normed size-measures

$$p_{ij} \left(0 < p_{ij} < 1, j = 1, \ldots, M_i, \sum_{j=1}^{M_i} p_{ij} = 1 \right)$$

are known for $i, j (j = 1, \ldots, M_i, i = 1, \ldots, N)$. Then $\sum_{m_i}, M_{ij}, y_{ij}, Q_{ij}, p_{ij}$'s are the notations to be used paralleling respectively $\sum_n, N_i, y_i, Q_i, P_i$. Then following RHC and Chaudhuri, Adhikary and Dihidar (2000) we may employ the unbiased estimator

$$e_R = \sum\nolimits_n \frac{Q_i}{p_i} \sum\nolimits_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij}$$

for which the variance is

$$V(e_R) = A \left[\left(\sum \frac{y_i^2}{p_i} - Y^2 \right) + \sum \frac{V_i}{p_i} \right] + (1 - A) \sum V_i$$

where

$$A = \frac{\sum_n N_i^2 - N}{N(N-1)}, V_i = \frac{\sum_{m_i} M_{ij}^2 - M_i}{M_i(M_i - 1)} \left(\sum_{j=1}^{M_i} \frac{y_{ij}^2}{p_{ij}} - y_i^2 \right)$$

and an unbiased estimator for $V(e_R)$ is

$$v(e_R) = B \sum\nolimits_n Q_i \left(\frac{\hat{y}_i}{p_i} - e_R \right)^2 + \sum\nolimits_n \frac{Q_i}{p_i} v_i$$

writing

$$\begin{aligned} B &= \left(\sum\nolimits_n N_i^2 - N\right) / \left(N^2 - \sum\nolimits_n N_i^2\right), \\ \hat{y}_i &= \sum\nolimits_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij} \\ \text{and } v_i &= \left(\frac{\sum_{m_i} M_{ij}^2 - M_i}{M_i^2 - \sum_{m_i} M_{ij}^2}\right) \sum\nolimits_{m_i} Q_{ij} \left(\frac{y_{ij}}{p_{ij}} - \hat{y}_i\right)^2 . \end{aligned}$$

Suppose there is an auxiliary variable x well-correlated with y and its values x_{ij} are known for every $j = 1, \ldots, M_i$ and for each $i = 1, \ldots, N$. In such a case to estimate Y following CSW (1976) it is suitable to employ the method of generalized regression estimation. In the present case it seems reasonable to estimate y_i by

$$\hat{y}_{gi} = \sum\nolimits_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij} + b_{Ri} \left(x_i - \sum\nolimits_{m_i} \frac{Q_{ij}}{p_{ij}} x_{ij}\right)$$

taking $b_{Ri} = \frac{\sum_{m_i} y_{ij} x_{ij} R_{ij}}{\sum_{m_i} x_{ij}^2 R_{ij}}$ and hence $\hat{y}_{gi}$ as $\hat{y}_{gi} = \sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij} g_{ij}$ with

$$g_{ij} = 1 + \left(x_i - \sum_{m_i} \frac{Q_{ij}}{p_{ij}} x_{ij}\right) \frac{x_{ij} R_{ij} \frac{p_{ij}}{Q_{ij}}}{\sum_{m_i} x_{ij}^2 \frac{p_{ij}}{Q_{ij}}}$$

with a suitable choice of positive constants R_{ij}, e.g.,

$$R_{ij} = \frac{1}{x_{ij}}, \frac{1}{\frac{p_{ij}}{Q_{ij}} x_{ij}} \text{ and } \frac{1 - \frac{p_{ij}}{Q_{ij}}}{\frac{p_{ij}}{Q_{ij}} x_{ij}}.$$

These are similar to $\frac{1}{x_i}, \frac{1}{\pi_i x_i}, \frac{1-\pi_i}{\pi_i x_i}$ in CSW's (1976) original greg estimator. Then, finally Y may be estimated by the generalized regression (greg) estimator

$$e_{gR} = \sum\nolimits_n \frac{Q_{ij}}{p_{ij}} \hat{y}_{gi} + b_R \left(X - \sum\nolimits_n \frac{Q_i}{p_i} x_i\right); \text{ writing } b_R = \frac{\sum_n \hat{y}_{gi} x_i R_i}{\sum_n x_i^2 R_i}$$

with R_i as a suitably chosen positive constant. We shall take R_i as one of

$$\frac{1}{x_i}, \frac{1}{\frac{p_i}{Q_i} x_i}, \frac{1 - \frac{p_i}{Q_i}}{\frac{p_i}{Q_i} x_i},$$

these too as in the use of CSW's (1976) greg estimator.

Since $\hat{y}_{gi}$ and e_{gR} are non-linear functions involving y_{ij},s we consider it proper to assess the accuracy of e_{gR} as an estimator for Y on applying bootstrap technique of repeated sampling from the initially realized two-stage sample. For this we first note the following:

(i) $\hat{y}_{gi}$ is a non-linear function of RHC estimators of the fsu-totals of the four variable quantities namely

$$\sum_1^{M_i} y_{ij}, \sum_1^{M_i} x_{ij}, \sum_1^{M_i} x_{ij} R_{ij} \frac{p_{ij}}{Q_{ij}} \text{ and } \sum_1^{M_i} x_{ij}^2 R_{ij} \frac{p_{ij}}{Q_{ij}};$$

so we may denote $\hat{y}_{gi}$ by $f_i\left(y, x, xR\frac{p}{Q}, x^2 R\frac{p}{Q}\right)$. Likewise,

(ii) e_{gR} is a non-linear function of the RHC estimators of the population totals of the four quantities namely

$$\sum_1^{N} y_i, \sum_1^{N} x_i, \sum_1^{N} \hat{y}_{gi} x_i R_i \frac{p_j}{Q_j} \text{ and } \sum_1^{N} x_i^2 R_i \frac{p_j}{Q_j}.$$

So we may write

$$\begin{aligned} e_{gR} &= f\left(y, x, \hat{y}_g x R\frac{p}{Q}, x^2 R\frac{p}{Q}\right) \\ &= \sum\nolimits_n \frac{Q_i}{p_i} \hat{y}_{gi} g_i, \\ g_i &= 1 + \left(X - \sum\nolimits_n \frac{Q_i}{p_i} x_i\right) \frac{x_i R_i \frac{p_i}{Q_i}}{\sum_n x_i^2 R_i \frac{p_i}{Q_i}}. \end{aligned}$$

In view of (i) and (ii) if appropriate bootstrap samples may be drawn, then it is easy to find (a) bootstrap standard error for e_{gR}, (b) bootstrap confidence interval for Y with a pre-assigned confidence coefficient (CC) by the well-known percentile method and associated measures of relative accuracies for the alternative choices of the greg estimators through various R_{ij}'s and R_i's.

An extension of Sitter's mirror-match bootstrap procedure to cover sampling by RHC technique in the two stages seems to be appropriate here to be employed in the following way.

First consider the already chosen sample in the second stage from the ith fsu already selected in the first stage-both by the RHC scheme in manners explained above. From the m_i ssu's selected, let a sample of size $m_i^*(1 \leq$

$m_i^* \leq m_i$) be selected again by the RHC scheme with Q_{ij}'s ($j = 1, \ldots, m_i$) as the normed size-measures recalling that $\sum_{m_i} Q_{ij}$ =1 for every $i \in U$. In drawing m_i^* ssu's by RHC scheme out of m_i ssu's sampled from the M_i ssu's in the ith fsu suppose m_{ij} ssu's are randomly assigned to the jth group ($j = 1, \ldots, m_i^*$), p_{ij}^* denoting the normed size-measures (now, one of the Q_{ij}'s) of the single ssu selected from the jth group with Q_{ij}^* denoting the sum of these over the m_{ij} units falling in the jth group.

Then, $\sum_{m_i^*} \frac{Q_{ij}^*}{p_{ij}^*} \left(\frac{Q_{ij}}{p_{ij}} y_{ij}\right)$ is an unbiased estimator for $\sum_{M_i} \frac{Q_{ij}}{p_{ij}} y_{ij} = \hat{y}_i$ for this sampling. Such a sampling is a bootstrap sampling. Let this be independently repeated a number of times, say, $\ell_i, i \in U$. Also, let E_2^*, V_2^* denote the operators for expectation and variance respectively, with respect to this bootstrap sampling. Then,

$$E_2^* \left[\frac{1}{\ell_i} \sum_1^{\ell_i} \left(\sum_{m_i^*} \frac{Q_{ij}^*}{p_{ij}^*} \left(\frac{Q_{ij}}{p_{ij}} y_{ij} \right) \right) \right] = \hat{y}_i$$

and

$$\begin{aligned} & V_2^* \left[\frac{1}{\ell_i} \sum_1^{\ell_i} \left(\sum_{m_i^*} \frac{Q_{ij}^*}{p_{ij}^*} \left(\frac{Q_{ij}}{p_{ij}} y_{ij} \right) \right) \right] \\ = \quad & \frac{1}{\ell_i} V_2^* \left[\sum_{m_i}^* \frac{Q_{ij}^*}{p_{ij}^*} w_{ij} \right], \text{ writing } w_{ij} = \frac{Q_{ij}}{p_{ij}} y_{ij} \\ = \quad & \frac{1}{\ell_i} \left[\left\{ \frac{\sum_{m_i}^* m_{ij}^2 - m_i}{m_i(m_i - 1)} \right\} \left(\sum_{m_i} \frac{w_{ij}^2}{Q_{ij}} - \left(\sum_{m_i} w_{ij} \right)^2 \right) \right] \\ = \quad & \frac{1}{\ell_i} \left[\left\{ \frac{\sum_{m_i}^* m_{ij}^2 - m_i}{m_i(m_i - 1)} \right\} \left(\sum_{m_i} Q_{ij} \frac{y_{ij}^2}{p_{ij}^2} - \hat{y}_i^2 \right) \right] \end{aligned} \tag{2.1}$$

Recalling that

$$v_i = \left(\frac{\sum_{m_i} M_{ij}^2 - M_i}{M_i^2 - \sum_{m_i} M_{ij}^2} \right) \left(\sum_{m_i} Q_{ij} \frac{y_{ij}^2}{p_{ij}^2} - \hat{y}_i^2 \right) \tag{2.2}$$

is an unbiased estimator of the variance of $\hat{y}_i$ Sitter (1992) recommends choosing ℓ_i such that the variance in (2.1) may equal v_i in (2.2) provided our concern was only about sampling in the second stage. But such a choice of ℓ_i is to be modified as follows to take account of sampling in two stages.

From the original sample of n fsu's already selected by the RHC scheme let a sample of $n^*(2 \leq n^* < n)$ fsu's be selected again employing an RHC scheme in the following way. Noting that $\sum_n Q_i = 1$, let Q_i's be now treated as the normed size-measures for the n fsu's already chosen. Let n_i's be positive integers chosen subject to $\sum_{n^*} n_i = n$ so that n_i fus's are assigned to the ith group when the sample of n fsu's is split up randomly into n^* groups. Now let Q_i^* denote the sum of the n_i values of Q_i falling in the ith group thus formed. Also let p_i^* denote the Q_i-value of the unit chosen in a sample of size one drawn from this ith group with a probability equal to its Q_i-value divided by Q_i^*. When this is independently repeated for all the n^* groups we have a bootstrap sample from the initial sample of n fsu's. Let this bootstrap sampling be independently repeated k times; k is a positive integer to be fixed up in a manner described below. Let E_1^*, V_1^* denote the operators for expectation and variance respectively, with respect to this bootstrap sampling. Then, we have, for any variable quantity u_i for $i \in U$,

$$\begin{aligned} E_1^* \left[\frac{1}{k} \sum_1^k \left(\sum\nolimits_{n^*} \frac{Q_i^*}{p_i^*} u_i \right) \right] &= \sum\nolimits_n \frac{Q_i}{p_i} u_i \text{ and} \\ V_1^* \left(\frac{1}{k} \sum_1^k \left(\sum\nolimits_{n^*} \frac{Q_i^*}{p_i^*} u_i \right) \right) &= \frac{1}{k} \left(\frac{\sum_{n_i^*} n_i^2 - n}{n(n-1)} \right) \left[\sum\nolimits_n \frac{u_i^2}{Q_i} - \left(\sum_1^N u_j \right)^2 \right]. \end{aligned}$$

If we implement the bootstrap sampling in the above manners in the two stages such that the second stage units are selected from the respective fsu's selected, then if we take

$$e^* = \frac{1}{k} \sum_1^k \left[\sum\nolimits_{n^*} \frac{Q_i^*}{p_i^*} \frac{Q_i}{p_i} \left(\frac{1}{\ell_i} \sum_1^{\ell_l} \left(\sum\nolimits_{m_i^*} \frac{Q_{ij}^*}{p_{ij}^*} \left(\frac{Q_{ij}}{p_{ij}} y_{ij} \right) \right) \right) \right]$$

we have $E^*(e^*) = E_1^* E_2^*(e^*) = \sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij} = e_R$, on taking $E^* = E_1^* E_2^* = E_2^* E_1^*$. If following Chaudhuri, Adhikary and Dihidar (2000) we write $V^* = E_1^* V_2^* + V_1^* E_2^* = E_2^* V_1^* + V_2^* E_1^*$ then following them we can check easily that

$$V^*(e^*) = \frac{1}{k} \left[\frac{\sum_{n^*} n_i^2 - n}{n(n-1)} \left(\sum\nolimits_n Q_i \left(\frac{\hat{y}_i}{p_i} \right)^2 - e_R^2 \right) + \frac{1}{\ell_i} \right.$$

$$\left. \left\{ \frac{\sum_{n^*} n_i^2 - n}{n(n-1)} \sum\nolimits_n \frac{a_i}{p_i} + \left(1 - \frac{\sum_{n^*} n_i^2 - n}{n(n-1)} \right) \sum\nolimits_n a_i \right\} \right],$$

writing

$$a_i = \left(\frac{\sum_{m^*} m_{ij}^2 - m_j}{m_i(m_i - 1)} \right) \left(\sum\nolimits_{m_i} Q_{ij} \left(\frac{y_{ij}}{P_{ij}} \right)^2 - (\hat{y}_i)^2 \right) \frac{Q_i^2}{p_i^2}.$$

Taking the cue from Sitter (1992) let us choose k and ℓ_i such that $V^*(e^*)$ may equal $v(e_R)$. This equation yields

$$\begin{aligned} k &= \frac{\sum_{n^*} n_i^2 - n}{n(n-1)} \cdot \frac{N^2 - \sum_n N_i^2}{\sum_n N_i^2 - N} \quad \text{and} \\ \ell_i &= \frac{1}{k}\left\{\frac{\sum_{n^*} n_i^2 - n}{n(n-1)}\frac{1}{p_i} + \left(1 - \frac{\sum_{n^*} n_i^2 - n}{n(n-1)}\right)\right\} \\ & \quad \cdot \frac{\sum_{m_i^*} m_{ij}^2 - m_i}{m_i(m_i-1)} \cdot \frac{M_i^2 - \sum_{m_i} M_{ij}^2}{\sum_{m_i} M_{ij}^2 - M_i}\,\frac{p_i}{Q_i}. \end{aligned}$$

In practice k and ℓ_i are to be replaced by integers nearest to them. Thus this finally specified bootstrap sampling procedure may be announced as the version of Sitter's (1992) mirror-match bootstrap sampling in two-stages starting with an RHC scheme in the first stage followed by one also in the second stage.

Now e^* is a bootstrap estimator for $Y = \sum_1^N \sum_i^{M_i} y_{ij}$. This bootstrap sampling is to be repeated usually a large number B of times. Suppose for the bth ($b = 1, \ldots, B$) bootstrap sample this e^* be denoted as e_b^*. Then, $e^*(.) = \frac{1}{B}\sum_{b=1}^B e_b^*$ is the bootstrap average version of e^* and

$$v(e^*) = \frac{1}{B-1}\sum_{b=1}^{B}(e_b^* - e^*(.))^2$$

is taken as the bootstrap variance estimator for the original estimator e_R.

Since e^* is an estimator for $\sum_1^N \ \sum_i^{M_i} y_{ij}$. we may denote it as $e^*(y)$, the bootstrap version of e_R. So we may write $e^*(y, x, \hat{y}_g xRp/Q, x^2Rp/Q)$ and $e_i^*(y, x, xRp/Q, x^2Rp/Q)$ as the bootstrap versions of $f(y, x, \hat{y}_g xRp/Q, x^2Rp/Q)$ and $f_i(y, x, xRp/Q, x^2Rp/Q)$ respectively. So, for the bth ($b = 1, \ldots, B$) bootstrap we evaluate $e_b^*(y, x, \hat{y}_g xRp/Q, x^2Rp/Q)$ on evaluating earlier $e_{i_b}^*(y, x, xRp/Q, x^2Rp/Q)$ on rightly choosing (R_{ij}, R_i). Then bootstrap variance estimators for the greg estimators are easily derived employing

$$v\left(e_b^*\left(y, x, \hat{y}_g xR\frac{p}{Q}, x^2R\frac{p}{Q}\right) = \frac{1}{B-1}\sum_{b=1}^{B}(e_b^*(.,.,.,.) - e^*(.)\right)^2.$$

Incidentally writing the finite population correlation coefficient between y

and x as

$$R_N = \frac{\sum_1^N M_i \sum_1^N \sum_i^{M_i} y_{ij} x_{ij} - \left(\sum_1^N \sum_i^{M_i} y_{ij}\right)\left(\sum_1^N \sum_i^{M_i} x_{ij}\right)}{\sqrt{\sum_1^N M_i \sum_1^N \sum_1^{M_i} y_{ij}^2 - \left(\sum_1^N \sum_1^{M_i} y_{ij}\right)^2} \sqrt{\sum_1^N M_i \sum_1^N \sum_1^{M_i} x_{ij}^2 - \left(\sum_1^N \sum_1^{M_i} x_{ij}\right)^2}}$$

and the coefficient of regression of y on x as

$$B_N = \frac{\sum_1^N M_i \sum_1^N \sum_i^{M_i} y_{ij} x_{ij} - \left(\sum_1^N \sum_i^{M_i} y_{ij}\right)\left(\sum_1^N \sum_i^{M_i} x_{ij}\right)}{\sum_1^N M_i \sum_1^N \sum_1^{M_i} x_{ij}^2 - \left(\sum_1^N \sum_1^{M_i} x_{ij}\right)^2}$$

we may estimate them respectively by

$$r = \frac{\left(\sum_n \frac{Q_i}{p_i} M_i\right)\left(\sum_n \frac{Q_i}{p_i} \sum_n \frac{Q_{ij}}{p_{ij}} y_{ij} x_{ij}\right) - \left(\sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij}\right)\left(\sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} x_{ij}\right)}{\sqrt{\left(\sum_n \frac{Q_i}{p_i} M_i\right)\left(\sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij}^2\right) - \left(\sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij}\right)^2}}$$
$$\times \frac{1}{\sqrt{\left(\sum_n \frac{Q_i}{p_i} M_i\right)\left(\sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} x_{ij}^2\right) - \left(\sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} x_{ij}\right)^2}}$$

and

$$b = \frac{\left(\sum_n \frac{Q_i}{p_i} M_i\right)\left(\sum_n \frac{Q_i}{p_i} \sum_n \frac{Q_{ij}}{p_{ij}} y_{ij} x_{ij}\right) - \left(\sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij}\right)\left(\sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} x_{ij}\right)}{\left(\sum_n \frac{Q_i}{p_i} M_i\right)\left(\sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij}^2\right) - \left(\sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij}\right)^2}$$

Both r and b are non-linear estimators of totals of $(1, yx, y, x, y^2$ and $x^2)$ and of $(1, yx, y, x,$ and $x^2)$. So, to assess their accuracies bootstrap sampling as above is useful in deriving bootstrap versions of r and b, denoted respectively by r^* and b^* and their bootstrap variance estimators denoted by $v(r^*), v(b^*)$, say, to assess accuracies in estimation respectively of R_N and B_N.

3 Numerical Illustration of Relative Efficacies

We consider estimation of total area under cultivation (y) in a particular district composed of 34 blocks as fsu's with numbers of villages (ssu's) respectively therein as 36, 22, 28, 69, 72, 23, 38, 31, 18, 43, 43, 35, 22, 30, 32, 39, 19, 18, 26, 42, 34, 23, 36, 75, 41, 46, 60, 50, 38, 44, 53, 55, 48 and 35. We consider surveying $n = 14$ blocks and 1/3rd of the villages (rounded below to integers) within the sampled blocks. A size-measure of a block is taken as its total population counted in the 1991 census and that of a village as its area (in hectares) also determined by 1991 census. As auxiliary variable x we take the area under irrigation, which is already, ascertained each village-wise. Bootstrap block sample-size is taken as n^*=7 and the village bootstrap sample-size is taken as m_i^* equal to half of m_i rounded downward to an integer. For a greg estimator we consider 3 alternative choices of (R_{ij}, R_i) equal to

$$\text{(i)}\ \left(\frac{1}{x_{ij}}, \frac{1}{x_i}\right),\ \text{(ii)}\ \left(\frac{1}{\frac{p_{ij}}{Q_{ij}}x_{ij}}, \frac{1}{\frac{p_i}{Q_i}x_i}\right)\ \text{and (iii)}\ \left(\frac{1-\frac{p_{ij}}{Q_{ij}}}{\frac{p_{ij}}{Q_{ij}}x_{ij}}, \frac{1-\frac{p_i}{Q_i}}{\frac{p_i}{Q_i}x_i}\right).$$

The respective greg estimators are denoted as e_{g1}, e_{g2}, e_{g3} and their bootstrap versions as $e^*_{g1}, e^*_{g2}, e^*_{g3}$, which we shall refer to as bootstrap estimators.

For the live data at hand we have $Y = 123639.03$ (hectare), $X = 983175.67$ (hectare), $R_N = 0.7019$ and $B_N = 0.0822$.

We take $100 \times \sqrt{v(e^*)}/e^*$ as a coefficient of variation (CV) as a measure of accuracy of e^* taken generically as any bootstrap estimator. Using the calculated values of e^*_b for $b = 1, \ldots, B$ (taken as 200) from their histogram 95% confidence interval for the parameter θ of which e is a point-estimator is found on working out e_L and e_U the lower and upper 2.5% points of the histogram, along with its length equal to $e_U - e_L$. Repeating this entire exercise a large number of times T taken as equal to 1000 we calculate (1) ACV, the average over the T replicates of the values of CV, (2) ACP, the percent of T replicates for which the CI namely e_L, e_U) covers θ and (3) AL, the average length $(e_U - e_L.)$ of the CI over the T replicates for the respective initial estimators – to be treated as criteria for comparative performances of several estimators.

Table 1. Showing the comparative performances of estimators

Bootstrap	Estimated values of the	For the bootstrap estimator e^*		
estimators	bootstrap estimators	ACV (in %)	ACP (in %)	AL
e^*_R	125185.07	8.61	86.10	20920.26
e^*_{gl}	127796.47	7.99	91.20	24182.61
e^*_{g2}	127858.89	7.53	93.00	24443.98
e^*_{g3}	127916.33	7.08	95.10	24631.83
r^*	0.6975	10.87	92.15	0.2103
b^*	0.0874	10.35	93.12	0.0287

Following the recommendations of a referee, for the above-mentioned bootstrap estimators we calculate the relative bias, say, $R.B.$ defined as

$R.B. = \left|\frac{1}{T}\sum_T e^*(.) - \theta\right|/\theta$, writing $E(\hat{\theta}) = \frac{1}{T}\sum_T e^*(.)$ where $\sum_T$ denotes the sum over T replicates. Taking $T = 100$, we illustrate these relative biases in Table 2 below.

Table 2. Showing percent relative biases of the bootstrap estimators

Estimator	Average of $T = 100$	% relative
	bootstrap estimates	bias
e_R^*	124885.69	1.0
e_{gl}^*	126873.89	2.6
e_{g2}^*	126902.89	2.6
e_{g3}^*	127001.33	2.7
r^*	0.6925	1.3
b^*	0.0848	3.2

4 Concluding Remarks

In inferring about a survey population parameter like a total or mean, correlation coefficient, regression coefficient etc., estimators that are nonlinear functions of sample-based random variables often have to be employed for the sake of efficiency. For efficiency comparisons bootstrap procedures may be helpful especially when the estimand parameter is unknown. In the present case if Y were unknown we could not evaluate ACP but ACV and AL values may be regarded as good enough criteria. Even if by alternative procedures measures of errors of the point estimates illustrated above could be evaluated and comparative study could be made in terms of the values of CV and AL for a realized sample better summary measures may be recognized as ACV and AL obtained through a bootstrap procedure. An alternative to a bootstrap demands assumption of normality in obtaining CI's and their assessments by ACP's. In the present case, however, our findings vindicate the use of a greg estimator as a better performer than the original RHC estimator for a population total.

Acknowledgement. The suggestions from the editor and a referee are gratefully acknowledged as useful in improving upon earlier drafts. The research of the first author is partially supported by CSIR through grant no. 21(0539)/02/EMR-II.

References

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, **63**, 615-620.

Chaudhuri, A., Adhikary, A.K. and Dihidar, S. (2000). Mean square error estimation in multi-stage sampling. *Metrika*, **52**, 115-131.

Indian Statistical Institute (2003). *Report on Audit Sampling.* Applied Statistics Unit, Indian Statistical Institute, Kolkata.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *J. Roy. Statist. Soc. B.* **24**, 482-491.

Sitter, R. (1992). A re-sampling procedure for complex survey data. *J. Amer. Statist. Assoc.* **87**, 755-765.

Arijit Chaudhuri
Applied Statistics Unit
Indian Statistical Institute
203 Barrackpore Trunk Road
Kolkata-700108, India
E-mail: achau@isical.ac.in

Amitava Saha
Directorate General of Mines Safety
Dhanbad-826001
Jharkhand, India
E-mail: saha_amitava@hotmail.com

Paper received: July 2003; revised August 2004.