Metrika (1994) 41:355–362

Metrika

Model Assisted Survey Sampling Strategy in Two Phases

ARIJIT CHAUDHURI

Indian Statistical Institute. 203, B. T. Road, Calcutta 700035, India

DEBESH ROY

Residency College, Calcutta 700073, India

Abstract: Postulating a super-population regression model connecting a size variable, a cheaply measurable variable and an expensively observable variable of interest, an asymptotically optimal double sampling strategy to estimate the survey population total of the third variable is specified. To render it practicable, unknown model-parameters in the optimal estimator are replaced by appropriate statistics. The resulting generalized regression estimator is then shown to have a model-cum-asymptotic design based expected square error equal to that of the asymptotically optimum estimator itself. An estimator for design variance of the estimator is also proposed.

Key Words and Phrases: Asymptotic analysis; double sampling; generalised regression estimator; optimal strategy, survey population, variance estimation.

1 Introduction

We consider double sampling with varying probability in both phases from a finite population in order to estimate the total of a variable y of interest. Earlier Rao and Bellhouse (1978) considered, taking a first phase sample utilizing known size-measures, w's and observing the values for it on an auxiliary variable x. Their second phase sample is a sub-sample from the first and its selection probability utilizes w's but not x's. Postulating a linear regression model connecting y, x and w of which the first two are treated as stochastic and the third as non-stochastic, they derived an optimal estimator for the total of y as a generalised difference estimator involving unknown parameters within a class of non-homogeneous linear unbiased estimators. Chaudhuri and Adhikary (1983, 1985) treated the same design set-up and superpopulation model but obtained a model parameter-free optimal estimator within a severely restricted class of design unbiased estimators. Mukerjee and Chaudhuri (1990) extended the class of designs permitting the second phase sample selection-probabilities to depend on both x and w-values, the latter known for the entire population. They consider the regression estimator which is design-biased and hence resorted to studying its asymptotic design-based properties. Also postulating an appropriate super-population model slightly different from the earlier one they established certain model-cum-asymptotic design-based properties of the regression estimator and specified optimal two-phase designs. Their study is based on the asymptotics of Robinson and Särndal (1983) which is a follow-up of Isaki and Fuller's (1982) and Fuller and Isaki's (1981) asymptotic approach. In the present work Särndal's (1980) generalised regression (greg) estimator in one phase sampling is extended to two-phase sampling with Rao-Bellhouse (1978) design, free of x's. Its properties following Brewer's (1979) asymptotic approach applicable to greg estimators are examined postulating linear regression model as in Mukerjee and Chaudhuri (1990) and extending exact theories of Godambe and Joshi (1965) and Godambe and Thompson (1977) a model-cum-asymptotic design-based optimum estimator is first derived. This involves model parameters. A class of optimal two-phase designs is noted and then shown that the double sampling regression estimator itself shares the same property with the optimal one. Finally we present a method of estimating the approxiate design variance of the two-phase regression estimator following Särndal's (1982) and Särndal, Swensson and Wretman's (1992) technique of estimating the approximate variance of greg estimator in single-phase sampling. The main reason for this presentation is to take advantage of simpler analysis by Brewer's (1979) approach compared to the much tougher one treated by Mukerjee and Chaudhuri (1990).

2 Asymptotically Optimal Double Sampling Strategy

We consider a finite population $U = (1, \ldots, i, \ldots, N)$ of N units labelled i $(= 1, \ldots, N)$. On it are defined three variables w, x and y with values w_i known and positive with total W, x_i unknown but ascertainable at little cost with total X and Y_i unknown and ascertainable at a high cost with total Y. The problem is to estimate Y. For this sample s_1 of size n_1 with probability $p_1(s_1)$ is taken from U and from s_1 sample s_2 of size n_2 is to be drawn with a conditional probability $p_2(s_2|s_1)$. The over-all two-phase sample $s = (s_1, s_2)$ has then the selection probability $p(s) = p_1(s_1).\ p_2(s_2|s_1)$. The designs p_1, p_2, p are permitted to involve elements of $\underline{W} = (w_1, \ldots, w_i, \ldots, w_N)$ but not those of $\underline{X} = (x_1, \ldots, x_i, \ldots, x_N)$ or $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$.

For design p_1, we suppose the inclusion probabilities $\sum_{s_1 \ni i} p_1(s_1) = \pi_{1i}$ and $\sum_{s_1 \ni ij} p_1(s_1) = \pi_{1ij}$ are positive. For design p_2 also the conditional inclusion probabilities, for each s_1 held fixed, namely $\sum_{s_2 \ni i} p_2(s_2|s_1) = \pi_{2i}(s_1)$ and $\sum_{s_2 \ni i,j} p_2(s_2|s_1) = \pi_{2ij}(s_1)$ are assumed to be positive. The survey data are denoted by $d = (s, x_i, y_j | i \in s_1, j \in s_2)$. By an estimator we mean a real-valued function of d which is free of any x_i for $i \notin s_1$ and of any y_j for $j \notin s_2$. It is to be noted throughout that s_2 is a subset of s_1. To employ a serviceable estimator $t = t(d)$ for Y we "first" postulate a model $\underline{M}$ connecting $\underline{W}$, $\underline{X}$, $\underline{Y}$ of which the first one is a vector of known positive constants but the latter are treated as random vectors. By $E_m(V_m, C_m)$ we denote the operator of expectation (variance, covariance) over the joint distribution of $\underline{Y}$ and $\underline{X}$. By $E_1(V_1, C_1)$ we denote the operator of expectation (variance, covariance) over the conditional distribution of $\underline{Y}$ given $\underline{X}$ and by $E_2(V_2, C_2)$ that over the distribution of $\underline{X}$. For $\underline{M}$ we postulate the following:

$$E_1(y_i|\underline{X}, \underline{W}) = \beta_1 x_i + \beta_2 w_i$$

$$V_1(y_i|\underline{X}, \underline{W}) = \sigma_{1i}^2 \ , \qquad E_2(\sigma_{1i}^2) = \psi_i^2 \ ,$$

$$E_2(x_i|\underline{W}) = \beta_3 W_i, \ V_2(x_i|\underline{W}) = \sigma_{2i}^2 \ , \qquad i = 1, \ldots, N$$

Further, y_i's are 'independent' conditionally on $\underline{X}$ and x_i's are 'independent'. We shall write $\theta = \beta_1 \beta_3$. "Secondly", we adopt the asymptotic analysis of Brewer (1979). This stipulates that U along with $\underline{W}$, $\underline{X}$, $\underline{Y}$, may be supposed to reporduce itself T (> 1) times leading to the following entities:

$$U_T = (\underline{U}(1), \ldots, \underline{U}(j), \ldots, \underline{U}(T)) \ ,$$

where

$$\underline{U}(j) = ((j-1)N + 1, \ldots, (j-1)N + N) \ ,$$

$$\underline{Y}_T = (\underline{Y}(1), \ldots, \underline{Y}(j), \ldots, \underline{Y}(T)) \ ,$$

where

$$\underline{Y}(j) = (y_{(j-1)N+1}, \ldots, y_{(j-1)N+i}, \ldots, y_{(j-1)N+N}) \ , \qquad j = 1, \ldots, T$$

such that units i, $N + i, \ldots, (j-1)N + i, \ldots, (T-1)N + i$ are same separately for every $i = 1, \ldots, N$ and $y_i = y_{N+i} = \cdots = y_{(j-1)N+i} = \cdots = y_{(T-1)N+i}$ for $i = 1, \ldots, N$ and similarly for $\underline{W}_T$, $\underline{X}_T$. From each $\underline{U}(j)$, samples of the type s are independently drawn according to p and amalgamated. From such pooled samples, estimators t are calculated so as to estimate TY rather than Y itself. Calculating the expectations with respect to the resulting designs limits are taken allowing T to tend to infinity. Such limits are denoted by $\lim E_p$. Utilizing Slutzky's theorem (cf. Cramér (1966)) about limits of functions of several sequences highly convenient simplifications are available under such limiting operations. With these preliminaries we shall seek estimators $t = t(d)$ for Y which are asymptotically design unbiased (ADU) satisfying

$$\lim E_p(t - Y) = 0 \tag{2.1}$$

For such ADU estimators of Y we have the following theorem.

Theorem 1: Assuming that E_p and E_m commute, under $\underline{M}$,

$$\begin{aligned} M(t) &= E_m \lim E_p(t - Y)^2 \\ &\geq \sum_1^N \left(\frac{1}{\pi_{1i}^2} \sum_{s_1 \ni i} \frac{p_1(s_1)}{\pi_{2i}(s_1)} - 1 \right) \psi_i^2 + \beta_1^2 \sum_1^N \left(\frac{1}{\pi_{1i}} - 1 \right) \sigma_{2i}^2 \\ &= E_m \lim E_p(t_0 - Y)^2 \\ &= M(t_0), \text{ say, where} \end{aligned}$$

$$\begin{aligned} t_0 &= t_0(d) \\ &= \sum_{j \in s_2} (y_j - \beta_1 x_i - \beta_2 w_j)/\pi_{1j}\pi_{2j}(s_1) + \beta_1 \sum_{i \in s_1} (x_i - \beta_3 w_i)/\pi_{1i} \\ &\quad + (\theta + \beta_2) \sum_1^N W_i \end{aligned}$$

Sketch of a proof: Writing $\Delta_m(t) = E_m(t - Y)$ and extending Godambe and Thompson's (1977) analysis relevant to $t(d)$ subject to $E_p(t(d) - Y) = 0$, we easily simplify, using commutativity of E_p, E_m, to get

$$M(t) = \lim E_p V_m(t) + \lim E_p \Delta_m^2(t) - V_m(Y) \ . \tag{2.2}$$

Next let $t_1 = t_1(d) = \sum_{j \in s_2} \frac{y_j}{\pi_{1j}\pi_{2j}(s_1)} + \beta_1 \left(\sum_{i \in s_1} \frac{x_i}{\pi_{1i}} - \sum_{j \in s_2} \frac{x_j}{\pi_{1j}\pi_{2j}(s_1)} \right)$. Then $E_p(t_1) = Y$. Hence $\lim E_p(t_1 - Y) = 0$. Let $h = h(d)$ be an "ADU estimator for 0" i.e. $\lim E_p(h) = 0$. Then we may write $t = t_1 + h$. Then, extending Godambe and Joshi's (1965) approach so as to verify that $\lim E_p C_m(t_1, h) = 0$, the inequality in Theorem 1 easily follows, on further noting that (i) $V_m(Y) =$

$\sum_1^N V_m(y_i) = \sum_1^N (\psi_i^2 + \beta_1^2\sigma_{2i}^2)$, (ii) $C_m(y_i - \beta_1 x_i, \beta_1 x_i) = 0$ and (iii) $\lim E_p V_m(t_1) = \sum_1^N \frac{\psi_i^2}{\pi_{1i}^2}\left(\sum_{s_1 \ni i} \frac{p_1(s_1)}{\pi_{2i}(s_1)}\right) + \beta_1^2 \sum_1^N \left(\frac{\sigma_{2i}^2}{\pi_{1i}}\right)$.

To get rid of the second term in $M(t)$ in (2.2), following Godambe and Thompson (1977) again an easy way is to choose $h = h(d)$ as $h_0 = h_0(d)$ where

$$h_0(d) = \theta\left(\sum_1^N W_i - \sum_{i\in s_1}\frac{w_i}{\pi_{1i}}\right) + \beta_2\left(\sum_1^N W_i - \sum_{j\in s_2}\frac{w_j}{\pi_{1j}\pi_{2j}(s_1)}\right)$$

yielding $t_0(d) = t_1(d) + h_0(d)$. Then, noting that $V_m(h_0) = 0$ and $\Delta_m(t_0) = 0$, the Theorem 1 finally follows.

The optimum design p for employing t_0, the optimal estimator for Y, is given by the following theorem.

Theorem 2: $M_0(t) \geq M_1 + \rho$ where $M_1 = \left(\frac{1}{n_2}\left(\sum_1^N \psi_i\right)^2 - \sum_1^N \psi_i^2\right) + \beta_1^2\left(\frac{1}{n_1}\left(\sum_1^N \sigma_{2i}\right)^2 - \sum_1^N \sigma_{2i}^2\right)$ and $\rho \to 0$ asymptotically in Brewer's sense and this lower bound is attained for a sampling design for which (i) $\pi_{2i}(s_1) = n_2(\psi_i/\pi_{1i}) \Big/ \sum_{i\in s_1}(\psi_i/\pi_{1i})$ for i in s_1 and (ii) $\pi_{1i} = n_1\sigma_{2i} \Big/ \sum_1^N \sigma_{2i}$

Sketch of a proof: To choose $\pi_{2i}(s_1)$ so as to minimise $M_0(t)$ we require to minimise $\sum_1^N \frac{\psi_i^2}{\pi_{1i}^2}\sum_{s_1\ni i}\frac{p_1(s_1)}{\pi_{2i}(s_1)}$ subject to $\sum_{i\in s_1}\pi_{2i}(s_1) = n_2$. The choice (i) is immediate. To choose π_{1i} we then need to minimise

$$\begin{aligned} A &= \frac{1}{n_2}\sum_1^N \frac{\psi_i}{\pi_{1i}}\sum_{s_1\ni i} p_1(s_1)\left(\sum_{i\in s_1}\frac{\psi_i}{\pi_{1i}}\right) + \beta_1^2\left(\sum_1^N \frac{\sigma_{2i}^2}{\pi_{1i}}\right) \\ &= \frac{1}{n_2}\left(\sum_1^N \psi_i\right)^2 + \beta_1^2\left(\sum_1^N \frac{\sigma_{2i}^2}{\pi_{1i}}\right) + \rho \\ &= A_0 + \rho, \text{ say.} \end{aligned}$$

Here $\rho = \frac{1}{n_2}\sum_1^N \frac{\psi_i}{\pi_{1i}}\sum_{s_1\ni i} p_1(s_1)\left(\sum_{i\in s_1}\frac{\psi_i}{\pi_{1i}} - \sum_1^N \psi_i\right)$

Since $|\rho| \leq \underset{s_1}{\text{Max.}}\left|\sum_{i\in s_1}\frac{\psi_i}{\pi_{1i}} - \sum_1^N \psi_i\right|\left(\frac{1}{n_2}\sum_1^N \psi_i\right)$

asymptotically in Brewer's sense, $\rho \to 0$. So it is desirable to choose π_{1i}'s so that A_0 is minimal. Now using Cauchy-Schwarz inequality $A_0 \geq \frac{1}{n_2}\left(\sum_1^N \psi_i\right)^2 + \frac{\beta_1^2}{n_1}\left(\sum_1^N \sigma_{2i}\right)^2$, with equality if and only if $\pi_{1i} = n_1\sigma_{2i}\Big/\sum_1^N \sigma_{2i}$. Hence follows the Theorem 2.

3 Asymptotic Optimality of Regression Estimator

The optimum design is of course not applicable because σ_{2i}, ψ_i cannot be known in practice. For any design with pre-assigned π_{1i}, $\pi_{2i}(s_1)$ also t_0 is not practicable because β_j ($j = 1, 2, 3$) are unknowable. So let $s_{zu} = \sum_{i \in s_2} z_i u_i$, where z_i, u_i stand for w_i, x_i, y_i, $i \in U$. Then, we suggest estimating $\underline{\beta} = (\beta_1, \beta_2)'$ by $\underline{b} = (b_1, b_2)'$ where

$$\underline{b} = \begin{pmatrix} s_{xx} & s_{xw} \\ s_{wx} & s_{ww} \end{pmatrix}^{-1} \begin{pmatrix} s_{xy} \\ s_{wy} \end{pmatrix}$$

and θ by $\hat{\theta} = \frac{s_{wy}}{s_{ww}} - b_2$. Then it may be checked that $E_m(\underline{b}) = \underline{\beta}$, $E_m(\hat{\theta}) = \theta$ and so $E_m(t_0^* - Y) = 0$. Here

$$t_0^* = \sum_{j \in s_2} \frac{y_j}{\pi_{1j}\pi_{2j}(s_1)} + b_1\left(\sum_{i \in s_1} \frac{x_i}{\pi_{1i}} - \sum_{j \in s_2} \frac{x_j}{\pi_{1j}\pi_{2j}(s_1)}\right)$$
$$+ b_2\left(\sum_1^N w_i - \sum_{j \in s_2} \frac{w_j}{\pi_{1j}\pi_{2j}(s_1)}\right) + \hat{\theta}\left(\sum_1^N w_i - \sum_{i \in s_1} \frac{w_i}{\pi_{1i}}\right)$$

So it is easy to check that $\lim E_p(t_0^* - Y) = 0$ i.e. t_0^* fulfils (2.1). Further it is not difficult to check applying Slutzky's theorem according to requirements, that

$$E_m \lim E_p(t_0^* - Y)^2 = E_m \lim E_p(t_0 - Y)^2 \ .$$

Details are omitted to save space and may be obtained on request from the authors if one needs.

4 Variance Estimation

We may recall that in one-phase sampling the well-known Horvitz – Thompson (1952) estimator has Yates – Grundy (1953) variance estimator. If the former is replaced by Särndal's (1980) greg estimator a Yates – Grundy type variance estimator for the latter is available from Särndal (1982) and Särndal, Swensson and Wretman (1992). Drawing analogy to these in the present double sampling situation, we propose, following these estimators, an estimator for an approximate design variance of the regression estimator t_0^* as

$$\hat{V} = \sum_{i<j\in s_2}\sum \left(\frac{\pi_{1i}\pi_{1j} - \pi_{1ij}}{\pi_{1ij}\pi_{2ij}(s_1)}\right)\left(\frac{y_i - (b_2 + \hat{\theta})w_i}{\pi_{1i}} - \frac{y_j - (b_2 + \hat{\theta})w_j}{\pi_{1j}}\right)^2$$

$$+ \sum_{i<j\in s_2}\sum \frac{\pi_{2i}(s_1)\pi_{2j}(s_1) - \pi_{2ij}(s_1)}{\pi_{2ij}(s_1)}\left(\frac{y_i - b_1 x_i}{\pi_{1i}} - \frac{y_j - b_1 x_j}{\pi_{1j}}\right)^2$$

Acknowledgement: The authors are grateful to a referee for constructive suggestions that lead to this improved and abridged version of an earlier draft.

References

Brewer KRW (1979) A class of robust sampling designs for large-scale surveys. Jour Amer Statist Assoc 74:911–915

Chaudhuri A, Adhikari AK (1983) On optimality of double sampling strategies with varying probabilities. Jour Statist Plan and Inf 8:257–265

Chaudhuri A, Adhikari AK (1985) Some results on admissibility and uniform admissibility in double sampling. Jour Statist Plan and Inf 12:199–202

Cramer H (1966) Mathematical methods of statistics. Princepon Univ Press

Fuller WA, Isaki CT (1981) Survey design under superpopulation models. In: Current topics in survey sampling. Krewski D, Platek R, Rao JNK (Eds) Academic Press NY 199–226

Godambe VP, Joshi VM (1965) Admissibility and bayes estimation in sampling finite populations. I. Ann Math Statist 36:1707–1722

Godambe VP, Thompson ME (1977) Robust near optimal estimation in survey practice. Bull Int Statist Inst 47, 3:129–146

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Jour Amer Statist Assoc 47:663–685

Isaki CT, Fuller WA (1982) Survey design under the regression superpopulation model. Jour Amer Statist Assoc 7:89–96

Mukerjee R, Chaudhuri A (1990) Asymptotic optimality of doulbe sampling plans employing generalized regression estimators. Jour Statist Plan and Inf 26:173–183

Rao JNK, Bellhouse DR (1978) Optimal estimation of a finite population mean under generalized random permutation models. Jour Statist Plan and Inf 2:125–141

Robinson PM, Särndal CE (1983) Asymptotic properties of the generalized regression estimator in probability sampling. Sankhya B, 45: 240–248

Särndal CE (1980) On π-inverse weighting versus best linear weighting in probability sampling. Biometrika 67: 634–650

Särndal CE (1982) Implications of survey design for generalized regression estimation of lienar functions. Jour Statist Plan and Inf 7: 155–170

Särndal CE, Swensson BE, Wretman JH (1992) Model assisted survey sampling. Springer Verlag

Yates F, Grundy PM (1953) Selection without replacement from within strata with probability proportional to size. Jour Roy Statist Soc B, 15:253–261

Received 02.10.1992
Revised version 09.11.1993