Model assisted survey sampling strategies with randomized response

Arijit Chaudhuri[a,*], Debesh Roy[b]

[a] Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700035, India
[b] Presidency College, Calcutta, India

Received 26 October 1992

Abstract

For estimating survey population totals with direct responses, use of QR-predictors motivated by postulated linear regression models and their asymptotic design-based analysis is rapidly becoming common. Their extension with randomized responses, to cover sensitive issues, is illustrated providing asymptotically optimal predictors along with variance estimators. Exact design unbiasedness requirement is replaced by asymptotic design unbiasedness restriction.

AMS classification: 62 D05

Keywords: Asymptotic optimality; QR-predictor; Randomized response; Survey populations; Variance estimation

1. Introduction

We consider a survey population $U = (1, \ldots, i, \ldots, N)$ of N individuals bearing the unknown values y_i of a stigmatizing variable y. Writing $\sum$ for sum over i in U, to estimate the total $Y = \sum y_i$, a sample s of distinct units, n in number, is supposed to be drawn according to a design p with probability $p(s)$ and positive inclusion probabilities π_i, π_{ij}, $i, j = 1, \ldots, N$. Supposing an auxiliary variable x with known positive values x_i, with a total X, is available, a practice is to assume plausibility of the following 'model' M, say, for which we may write

$$y_i = \beta x_i + \varepsilon_i, \quad i = 1, \ldots, N. \tag{1.1}$$

Here β is an unknown constant, ε_i are random variables with model expectations, variances and covariances, respectively, as

$$E_m(\varepsilon_i) = 0, \qquad V_m(\varepsilon_i) = \sigma_i^2, \qquad C_m(\varepsilon_i, \varepsilon_j) = \sigma_{ij}, \quad i \neq j.$$

For simplicity, occasionally we shall write $\mu_i = \beta x_i$. Writing $E_p\,(V_p)$ as design expectation (variance) operator, t as an estimator for Y and demanding

$$E_p(t) = Y, \tag{1.2}$$

procedures for choosing (p, t) to control $E_m V_p(t)$ and their modifications, when randomized responses (RR) rather than direct responses (DR) are only available, are well known to have been documented in the literature. One may consult for example, Cassel et al. (1977), Chaudhuri and Vos (1988) and Chaudhuri and Mukerjee (1988). An alternative approach is to replace (1.2) by a requirement of 'asymptotic design unbiased' (ADU)-ness of an estimator t demanding

$$\lim E_p(t) = T \tag{1.3}$$

and to be motivated to seek a strategy (p, t) so as to control the magnitude of

$$\lim E_p E_m (t - Y)^2. \tag{1.4}$$

The concepts of ADU-ness and $\lim E_p(\cdot)$ will be briefly indicated in Section 2 below, following Brewer (1979). In what follows, our plan is to extend this approach, elaborately narrated in the recent book by Särndal, Swensson and Wretman (SSW) (1992) to the case when instead of the DR y_i, only RR, say, r_i is available for i in s. In the sequel, we shall derive results concerning appropriate estimators for Y based on $(s, r_i \mid i \in s)$. Their corollaries covering estimators based on $(s, y_i \mid i \in s)$ will emerge as new results about DR not yet recorded in the literature. So, this paper presents new optimality results covering RR as well as DR. This is achieved adopting Brewer's (1979) asymptotic method. Variance estimators for the former are derived and noted to coincide with the corresponding ones already available for the latter when r_i are replaced by y_i, $i \in s$.

Following Chaudhuri and Mukerjee (1988), we assume that devices are available to produce r_i independently of r_j for $i \neq j$, satisfying, say,

$$E_R(r_i) = y_i, \qquad V_R(r_i) = \alpha_i y_i^2 + \beta_i y_i + \theta_i = V_i, \tag{1.5}$$

with known constants $\alpha_i, \beta_i, \theta_i$, $i = 1, \ldots, N$. By E_R, V_R, C_R we denote the operators for expectation, variance and covariance with respect to randomization experiment.

Chaudhuri and Mukerjee (1988) also gave an estimator for V_i as

$$\hat{V}_i = \frac{1}{1 + \alpha_i}(\alpha_i r_i^2 + \beta_i r_i + \theta_i) \tag{1.6}$$

satisfying $E_R(\hat{V}_i) = V_i$, $i \in s$, provided $1 + \alpha_i \neq 0$.

The celebrated QR-predictor for Y introduced by Wright (1983) and further studied by Särndal and Wright (1984) and SSW (1992) based on DR is

$$t_{\mathrm{QR}} = X\hat{\beta}_Q + \sum{}' R_i(y_i - x_i\hat{\beta}_Q). \tag{1.7}$$

We write $\sum{}'$ for sum over i in s and

$$\hat{\beta}_Q = \frac{\sum{}' y_i x_i Q_i}{\sum{}' x_i^2 Q_i}; \quad Q_i(>0) \text{ and } R_i \text{ are constants.} \tag{1.8}$$

Later we shall write $\sum\sum$ for sum over $i, j\ (i \neq j)$ in U and $\sum{}'\sum{}'$ for that in s.

Corresponding to any estimator t for Y based on y_i, $i \in s$, we shall consider an estimator for Y based on RR as e which is t evaluated with each y_i in t replaced by r_i, $i \in s$. The RR-version of t_{QR} will be denoted by e_{QR}. Following Särndal and Wright (1984) we choose Q_i, R_i subject to

$$\frac{1 - R_i\pi_i}{Q_i\pi_i x_i} = \text{a constant for each } i \text{ in } U, \tag{1.9}$$

to ensure that $\lim E_p(t_{\mathrm{QR}}) = Y$. This also ensures that $\lim E_p(e_{\mathrm{QR}}) = \sum r_i$.

First we shall derive a lower bound $M_0(R)$, say, for $M(R) = \lim E_p E_m E_R(e - Y)^2$ with e subject to

$$\lim E_p(e) = \sum r_i \tag{1.10}$$

and shall illustrate choice of Q_i, R_i for which 'this bound will be attained' for 'e as e_{QR}', irrespective of the values of $\beta, \sigma_i, \sigma_{ij}$. Then, analogous to the variance estimators for t_{QR} given by Särndal (1982) we shall propose two variance estimators for e_{QR} based on a fixed-size design and show them to suitably match when y_i and r_i are interchanged.

2. Asymptotically optimal QR-predictor

We shall adopt Brewer's (1979) approach in toto in our asymptotic analysis. Following him we suppose that U is reproduced $T\ (>1)$ times yielding $U(j) = ((j-1)N+1, \ldots, (j-1)N+i, \ldots, (j-1)N+N)$, the units $(j-1)N+i$ for $j = 1, \ldots, T$ and fixed i being the same for respective $i = 1, \ldots, N$. Consequently, $y_{(j-1)N+i} = y_i$ for every $j = 1, \ldots, T$ for each respective $i = 1, \ldots, N$. From each $U(j)$ a sample is then drawn according to p 'independently' for every $j = 1, \ldots, T$. The samples so drawn are amalgamated into a pooled sample, say, s_T which has a size Tn. The latter is in fact drawn from the population $U_T = (U(1), \ldots, U(j), \ldots, U(T))$ of size TN. The selection probability of s_T is, say,

$$p(s(1)) \cdots p(s(T)) = p_T(s_T),$$

p_T denoting the resulting design. The inclusion probability for every $(j-1)N+i$ obviously equals π_i for every $j=1,\ldots,T$ for a fixed $i\,(=1,\ldots,N)$. The (a) identity of $(j-1)N+i$ with i for every $j=1,\ldots,T$ for each separate $i\,(=1,\ldots,N)$ and (b) 'independence' of $s(j)$ over $j=1,\ldots,T$ are the two crucial assumptions in Brewer's (1979) approach. The total of y based on U_T is $Y_T=\sum_{j=1}^{T}(\sum_{i=1}^{N}y_{(j-1)N+i})=TY$. If $t=t(s)$, say, based on $(s,y_i\,|\,i\in s)$ is intended to estimate Y, then $t(s_T)$ based on $(s_T,y_i\,|\,i\in s_T)$ should be a natural estimator of TY. Consequently, if for t,

$$\lim_{T\to\infty}E_{pT}\left(\frac{1}{T}t(s_T)-Y\right)=0, \tag{2.1}$$

written, in brief, as

$$\lim E_p(t-Y)=0,$$

then such a t is ADU for Y.

The above limiting operation is analogously defined for an estimator e based on RR and is conveniently applicable making use of Slutzky's theorem, vide Cramèr (1946) in easily deriving several asymptotic results illustrated below.

Following Brewer (1979) again, the limiting design-based mean square error (MSE) of t as an estimator for Y will be taken as

$$\lim E_p(t-Y)^2=\lim_{T\to\infty}E_{pT}\left[\frac{1}{T}(t(s_T)-Y_T)^2\right].$$

It is worth noting that $Y_T^2=T(\sum y_i^2+\sum\sum y_iy_j)=TY^2$. Further, we follow Brewer (1979) to define

$$\lim E_pE_m(t-Y)^2=\lim_{T\to\infty}E_{pT}E_m\left[\frac{1}{T}(t(s_T)-Y_T)^2\right]$$

and extend this likewise to the definition of

$$M(R)=\lim E_pE_mE_R(e-Y)^2=\lim_{T\to\infty}E_{pT}E_mE_R\left[\frac{1}{T}(e(s_T)-Y_T)^2\right].$$

Chaudhuri and Stenger (1992) discuss how, with Brewer's (1979) approach above, an estimator t which is ADU for $Y=\lim E_p(t)$ also 'converges in probability' to '$\lim E_p(t)$'. This facilitates application of Slutzky's theorem concerning well-behaved functions of $(s,y_i\,|\,i\in s)$ and $(s,r_i\,|\,i\in s)$.

Essentially following Godambe and Joshi (1965) and Godambe and Thompson (1977) we get, assuming that E_p,E_m,E_R commute and using (1.5) and (1.10),

$$\begin{aligned}M(R)&=\lim E_pE_mE_R(e-Y)^2\\&=\lim E_pE_mE_R[e-E_R(e)+(E_R(e)-E_mE_R(e))\\&\quad+(E_mE_R(e)-E_mY)-(Y-E_mY)]^2\end{aligned}$$

$$
\begin{aligned}
&= \lim E_p E_m V_R(e) + \lim E_p V_m(E_R(e)) \\
&\quad + \lim E_p [E_m E_R(e) - E_m Y]^2 - V_m(Y).
\end{aligned}
\tag{2.2}
$$

Let $\bar{e} = \sum' (r_i/\pi_i)$ and $h = e - \bar{e}$ so that

$$\lim E_p h = 0. \tag{2.3}$$

Then,

$$\lim E_p E_m V_R(e) = \lim E_p E_m V_R(\bar{e}) + \lim E_p E_m V_R(h), \tag{2.4}$$

$$\lim E_p V_m(E_R(e)) = \lim E_p V_m(E_R(\bar{e})) + \lim E_p V_m(E_R(h)). \tag{2.5}$$

Eqs. (2.4) and (2.5) follow from (i) and (ii) below.

$$
\begin{aligned}
\text{(i)}\qquad \lim E_p C_R(\bar{e}, h) &= \lim E_p E_R\left[h \sum{}' \frac{(r_i - y_i)}{\pi_i}\right] \\
&= E_R \lim E_p \left[h \sum{}' \frac{(r_i - y_i)}{\pi_i}\right] \\
&= E_R\left[(\lim E_p h) \lim E_p \sum{}' \frac{(r_i - y_i)}{\pi_i}\right], \\
&\qquad \text{applying Slutzky's theorem} \\
&= 0, \quad \text{by (2.3), and}
\end{aligned}
$$

$$
\begin{aligned}
\text{(ii)}\qquad \lim E_p C_m[E_R(\bar{e}), E_R(h)] &= \lim E_p E_m \left[E_R(h) \sum{}' \frac{(y_i - \mu_i)}{\pi_i}\right] \\
&= E_m \lim E_p \left[E_R(h) \sum{}' \frac{(y_i - \mu_i)}{\pi_i}\right] \\
&= E_m \left[\lim E_p(E_R h) \lim E_p \sum{}' \frac{(y_i - \mu_i)}{\pi_i}\right] \\
&\qquad \text{by Slutzky's theorem} \\
&= E_m \left[E_R(\lim E_p h) \lim E_p \sum{}' \frac{(y_i - \mu_i)}{\pi_i}\right] \\
&= 0, \quad \text{by (2.3).}
\end{aligned}
$$

So,

$$
\begin{aligned}
M(R) \geqslant{}& \lim E_p E_m V_R(\bar{e}) + \lim E_p V_m(E_R \bar{e}) - V_m(Y) \\
&- \sum \frac{E_m(V_i)}{\pi_i} + \sum \sigma_i^2 \left(\frac{1}{\pi_i} - 1\right) + \sum\sum \sigma_{ij}\left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) = M_O(R), \text{ say.}
\end{aligned}
\tag{2.6}
$$

By an analogous analysis, for t subject to (1.3) it is easy to note the corollary to (2.6) concerning t that

$$\begin{aligned} M(D) &= \lim E_p E_m (t - Y)^2 \\ &\geqslant \sum \sigma_i^2 \left(\frac{1}{\pi_i} - 1\right) + \sum\sum \sigma_{ij} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) \\ &= M_0(D), \text{ say.} \end{aligned} \tag{a}$$

Referring to Section 1, we have

$$E_m E_R(e_{QR}) = E_m(Y), \tag{2.7}$$

$$\begin{aligned} \lim E_p E_m V_R(e_{QR}) &= E_m \lim E_p V_R \left[\sum{}' \left\{ R_i + \frac{(X - \sum' R_i x_i)}{\sum' x_i^2 Q_i} x_i Q_i \right\} r_i \right] \\ &= \sum \pi_i \left[R_i + \frac{(X - \sum R_i x_i \pi_i)}{\sum x_i^2 Q \pi_i} x_i Q_i \right]^2 E_m(V_i), \end{aligned} \tag{2.8}$$

$$\begin{aligned} &\lim E_p V_m(E_R e_{QR}) = \lim E_p V_m(t_{QR}) \\ &\quad = \sum \pi_i \left[R_i + \left(\frac{X - \sum R_i x_i \pi_i}{\sum x_i^2 Q_i \pi_i} \right) x_i Q_i \right]^2 \sigma_i^2 \\ &\quad\quad + \sum\sum \pi_{ij} \left[R_i - \left(\frac{X - \sum R_i x_i \pi_i}{\sum x_i^2 Q_i \pi_i} \right) x_i Q_i \right] \\ &\quad\quad \times \left[R_j + \left(\frac{X - \sum R_i x_i \pi_i}{\sum x_i^2 Q_i \pi_i} \right) x_j Q_j \right] \sigma_{ij}. \end{aligned} \tag{2.9}$$

Let $R_i = 1/\pi_i$ in e_{QR} with Q_i (>0) arbitrary as before. Such an e_{QR} will be denoted by e_G. Then, using (2.2) and (2.7)–(2.9) it follows that

$$\lim E_p E_m E_R (e_G - Y)^2 = M_0(R). \tag{2.10}$$

Thus, for any e satisfying (1.10), including e_{QR} subject to (1.9),

$$\lim E_p E_m E_R (e - Y)^2 \geqslant \lim E_p E_m E_R (e_G - Y)^2.$$

So, whatever design p may be employed, an 'asymptotically optimal' predictor or estimator for Y under the model M is e_G which is e_{QR} with $R_i = 1/\pi_i$ and Q_i as any positive constants. The corresponding t_{QR}, to be written t_G, is called by Särndal (1980) the generalized regression (Greg) predictor and its RR version e_G may also be so labelled. It may be noted that for $Q_i = 1/(\pi_i x_i)$, t_G is called Hájek's (1971) predictor and for $Q_i = (1 - \pi_i)/(\pi_i x_i)$, t_G is called Brewer's (1979) predictor. The latter is also identical with t_{QR} with $R_i = 1$, $Q_i = (1 - \pi_i)/(\pi_i x_i)$. As a corollary to (2.10), it follows with an analogous analysis that

$$\lim E_p E_m (t_G - Y)^2 = M_0(D). \tag{b}$$

Then, combining (a) and (b), for any t which is ADU for Y, under $\underline{M}$,

$$M(D) = \lim E_p E_m(t - Y)^2 \geqslant \sum \sigma_i^2 \left(\frac{1}{\pi_i} - 1\right) + \sum\sum \sigma_{ij} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)$$
$$- \lim E_p E_m(t_G - Y)^2, \tag{2.13}$$

a result not yet recorded in the literature.

3. Variance estimation

For a Greg predictor t_G an approximate variance formula using Taylor's expansion was given by Särndal (1982) as

$$V = \frac{1}{2} \sum\sum \Delta_{ij} \left(\frac{E_i}{\pi_i} - \frac{E_j}{\pi_j}\right)^2. \tag{3.1}$$

Here $E_i = y_i - x_i B_Q$, $B_Q = \sum y_i x_i Q_i \pi_i / (\sum X_i^2 Q_i \pi_i)$, $\Delta_{ij} = \pi_i \pi_j - \pi_{ij}$. He also gave two estimators for it as

$$v_{Gk} = \frac{1}{2} \sum{}' \sum{}' \frac{\Delta_{ij}}{\pi_{ij}} \left(a_{ki} \frac{e_i}{\pi_i} - a_{kj} \frac{e_j}{\pi_j}\right)^2, \quad k = 1, 2;$$

here $e_i = y_i - x_i \hat{\beta}_Q$, $a_{1i} = 1$, $a_{2i} = 1 + (X - \sum' (x_i/\pi_i))[x_i Q_i \pi_i / \sum' x_i^2 Q_i]$. V is approximately equal to $E_p(t_G - Y)^2$. Motivated by these we may write

$$\hat{\beta}_Q(r) = \frac{\sum' r_i x_i Q_i}{\sum' x_i^2 Q_i}, \qquad e_i(r) = r_i - x_i \hat{\beta}_Q(r),$$

$$v_{Gk}(r) = \frac{1}{2} \sum{}' \sum{}' \frac{\Delta_{ij}}{\pi_{ij}} \left(a_{ki} \frac{e_i(r)}{\pi_i} - a_{kj} \frac{e_j(r)}{\pi_j}\right)^2, \quad k = 1, 2.$$

Let us note that

$$E_R(e_i(r) - e_i)(e_j(r) - e_j) = V_i \delta_{ij} - x_i x_j \left[\frac{Q_i V_i + Q_j V_j}{\sum' x_i^2 Q_i} - \frac{\sum' x_i^2 Q_i^2 V_i}{(\sum' x_i^2 Q_i)^2}\right]$$
$$= F_{ij}, \text{ say;}$$

hence $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. We shall write $\hat{F}_{ij}$ for F_{ij} on replacing V_i, V_j in the latter throughout by $\hat{V}_i, \hat{V}_j$.

$$\text{Then, } E_p E_R(e_G - Y)^2 = E_p(t_G - Y)^2 + E_p\left[\sum{}' \left(\frac{a_{2i}}{\pi_i}\right)^2 V_i\right]$$
$$\simeq V + E_p\left[\sum{}' \left(\frac{a_{2i}}{\pi_i}\right)^2 V_i\right], \tag{3.2}$$

$$E_R v_{Gk}(r) = v_{Gk} + \frac{1}{2} \sum{}' \sum{}' \frac{\Delta_{ij}}{\pi_{ij}} \left[\left(\frac{a_{ki}}{\pi_i}\right)^2 F_{ii} + \left(\frac{a_{kj}}{\pi_j}\right)^2 F_{jj} - 2 a_{ki} a_{kj} \frac{F_{ij}}{\pi_i \pi_j}\right]. \tag{3.3}$$

From these we propose two estimators, for $E_pE_R(e_G - Y)^2$, a proposed measure of error of e_G as an estimator of Y, as

$$v^*_{Gk}(r) = v_{Gk}(r) - \frac{1}{2}\sum{}'\sum{}' \frac{\Delta_{ij}}{\pi_{ij}}\left[\left(\frac{a_{ki}}{\pi_i}\right)^2 \hat{F}_{ii} + \left(\frac{a_{kj}}{\pi_j}\right)^2 \hat{F}_{jj} - 2a_{ki}a_{kj}\frac{\hat{F}_{ij}}{\pi_i\pi_j}\right] + \sum{}'\left(\frac{a_{2i}}{\pi_i}\right)^2 \hat{V}_i, \quad k = 1, 2. \tag{3.4}$$

If y_i, $i \in s$ are available, then $v^*_{Gk}(r)$ reduces to v_{Gk} on substituting y_i for r_i, $i \in s$, and no new formula is required. This facilitates computer-based calculations.

Acknowledgements

The authors are grateful to three anonymous referees, the Associate Editor and the Co-ordinating Editor for their helpful comments which led to this vastly improved version of several earlier drafts.

References

Brewer, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *J. Amer. Statist. Assoc.* **74**, 911–915.

Cassel, C.M., C.E. Särndal and J.H. Wretman (1977). *Foundations of Inference in Survey Sampling*. Wiley, New York.

Chaudhuri, A. and R. Mukerjee (1988). *Randomized Response: Theory and Techniques*. Marcel Dekker, New York.

Chaudhuri, A. and H. Stenger (1992). *Survey Sampling: Theory and Methods*. Marcel Dekker, New York.

Chaudhuri, A. and J.W.E. Vos (1988). *Unified Theory and Strategies of Survey Sampling*. North-Holland, Amsterdam.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

Godambe, V.P. and V.M. Joshi (1965). Admissibility and Bayes estimation in sampling finite populations. I. Ann. Math. Statist. **36**, 1707–1722.

Godambe, V.P. and M.E. Thompson (1977). Robust near optimal estimation in survey practice. *Bull. Internat. Statist. Inst.* **47** (3), 129–146.

Hájek, J. (1971). Comment on a paper by Basu, D. In: V.P. Godambe and D.A. Sprott, Eds., *Foundations of Statistical Inference*. Holt, Rinehart & Winston, Toronto.

Horvitz, D.G. and D.J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663–685.

Särndal, C.E. (1980). On π-inverse weighting versus best linear weighting in probability sampling. *Biometrika* 67, 639–650.

Särndal, C.E. (1982). Implications of survey design for generalized regression estimation of linear functions. *J. Statist. Plann. Inference* **7**, 155–170.

Särndal, C.E., B.E. Swensson and J.H. Wretman (1992). Model assisted survey sampling. Springer, New York.

Särndal, C.E. and R.L. Wright (1984). Cosmetic form of estimators in survey sampling. *Scand. J. Statist.* **11**, 146–156.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. *J. Amer. Statist. Assoc.* **78**, 879–884.