عدم الاستجابة في معاينة مزدوجة للانحدار

زين العابدين عبدالرحيم البشير

أستاذ مشارك، قسم الأساليب الكمية، كلية العلوم الإدارية، جامعة الملك سعود، الرياض، المملكة العربية السعودية

(قُدم للنشر في ١٤١٣/٩/٦هـ وقُبل للنشر في ١٤١٤/٤/١٦هـ)

ملخص البحث. بعض النتائج التي توصل إليها سارندال وسونيسون للمعاينة المزدوجة ـ وهي طريقة معاينة تسحب فيها عينة أولى كبيرة يليها سحب عينة ثانية أصغر تمثل مجموعة جزئية فعلية من الأولى ـ عممت لتشمل الحالة التي تسحب فيها عينة ثالثة من العينة الثانية.

هذه النتائج النظرية استخدمت بعد ذلك في معالجة مشكلة عدم الاستجابة في معاينة مزدوجة للانحدار.

and $\hat{S}_e^2$ taking the same form as $\hat{S}_y^2$ with y replaced by e. Again, the quantity within square brackets in (6.3) is the usual estimate of variance of the mean in double sampling with simple random sampling in both phases and complete response (e.g. Cochran [4, pp. 343]).

In the above example, the first-phase and second-phase designs are assumed simple random. However, the preceeding theory applies to any type of design. In particular the design in one or both phases may be a multistage design of some complex form.

References

[1] Särndal, C.E. and Swensson, B. "A General View of Estimation for Two-Phases of Selection with Applications to Two-Phase Sampling and Nonresponse." *Int. Statist. Rev.*, 55, (1987), pp.279-294.

[2] Cassel, C.M.; Särndal, C.E. and Wretman, J.H. "Some uses of statistical models in connection with the nonresponse problem" (1983). In: W.G. Madow and I. Olkin. *Incomplete Data in Sample Surveys*, 3 rd ed. pp. 143-160. New York: Academic Press.

[3] Little, R.J.A. "Survey Nonresponse Adjustments for Estimates of Means." *Int. Statist. Rev.*, 54,(1986), (139-157).

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

[5] Platek, R. and Gray, G.B. *"Imputation Methodology"*. (1983), In: W.G. Madow, I. Olkin and D.B. Rubin, *Incomplete Data in Sample Surveys*, 2nd ed. pp. 255-293. New York: Academic Press.

[6] Oh, H.L. and Scheuren, F.J. *"Weighting Adjustment for Unit Nonresponse"* (1983), In: W.G. Madow, I. Olkin and D.B. Rubin, *Incomplete Data in Sample Surveys, 2nd* ed. pp. 143-184. New York: Academic Press.

[7] Chaudhuri, A. and Adhikary, A.K. "On optimality of double sampling strategies with varying probabilities". *J. Statist. Plan. Inf.*, 8, (1983), 257-265.

[8] David, M., Little, R., Samuhel, M. and Triest, R. "Nonrandom nonresponse models based on the propensity to respond." *Proc. Bus. Econ. Statist. Sect. Am. Statist. Assoc.*, (1983), pp. 168-173.

[9] Little, R.J.A. Models for nonresponse in sample surveys. *J. Am. Statist. Assoc.*, 77,(1982), 327-350.

[10] Thomsen, I. "A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analysing survey data." *Statist Tidskrift*, 11, (1973) 278-285.

In a similar manner, expressions for the variance and estimator of variance can be obtained by substituting values of the various inclusion probabilities. Thus using (5.1b) and noting that for $k \neq l$

$$\pi_{akl} = \frac{n_1}{N} \cdot \frac{(n_1 - 1)}{(N-1)} \qquad ;\pi_{kl|s_1} = \frac{n_2}{n_1} \cdot \frac{(n_2 - 1)}{(n_1 - 1)}$$

$$\Delta_{akl} = \frac{-f_1^*(1 - f_1^*)}{N-1}$$

and for $k = l$

$$\Delta_{akl} = f_1^*(1 - f_1^*)$$

the approximate estimate of variance (5.2c) takes the form:

$$v(\hat{t}_{cR^*}) \cong \sum_r f_1^{*-2} f_2^{*-1}(1-f_1) y_k^2 / \pi_{k|s_1,s_2,\underline{m}} - \sum\sum_r f_1^{*-2} f_2^{*-1}(1-f_1) y_k y_l / (n_2 - 1)\pi_{kl|s_1,s}$$

$$\sum_r f_1^{*-2} f_2^{*-2}(1-f_2^*) e_k^2 / \pi_{k|s_1,s_2,\underline{m}} - \sum\sum_{k \neq l} \sum_r f_1^{*-2} f_2^{*-2}(1-f_2^*) e_k e_l / (n_2 - 1)\pi_{kl|s_1,s_2,\underline{m}}$$

$$\sum_h^{H_{s_2}} n_h^2 (1-f_h) S_{e_{r_h}}^2 / m_h \qquad (6.2)$$

where $\pi_{k|s_1,s_2,\underline{m}}$ and $\pi_{kl|s_1,s_2,\underline{m}}$ are as in §5.

When response is complete, the last term vanishes because $f_h = 1$. Also, since in this case $r = s_2$ it follows that for all k and L:

$$\pi_{k|s_1,s_2,\underline{m}} = \pi_{kl|s_1,s_2,\underline{m}} = 1$$

The expression (6.2) then reduces to the simple form

$$\hat{V}(\hat{t}_{cR^*}) \cong N^2 \left[\frac{\hat{S}_y^2}{n_1} - \frac{\hat{S}_y^2}{N} + \frac{\hat{S}_e^2}{n_2} - \frac{\hat{S}_e^2}{n_1} \right] \qquad (6.3)$$

with

$$\hat{S}_y^2 = \left(\sum_{s_2} y_k^2 - \left(\sum_{s_2} y_k \right)^2 / n_2 \right) / (n_2 - 1)$$

and the overall respondents set is r. Measurement is made of characteristic y on all units in the responding set r, and a linear relation is assumed to exist between x and y.

We then have:

$$\pi_{ak} = \frac{n_1}{N} = f_1^* \qquad k \in U; \qquad \pi_{k|s_1} = \frac{n_2}{n_1} = f_2^* \quad (k \in S_1)$$

Also for $k,l \in S_{2h}$:

$$\pi_{k|s_1,s_2 m} = \frac{m_h}{n_h} = f_h$$

The regression estimator (5.2a) becomes:

$$\hat{t}_{cR^*} = \frac{N}{n_1}\sum\nolimits_{s_1} \hat{y}_k + \frac{N}{n_2}\sum\nolimits_h f_h^{-1} \sum\nolimits_{r_h} (y_k - \hat{y}_k) \tag{6.1}$$

where the regression estimate b used in obtaining $\hat{y}_k$ is defined as in §4 with vectors replaced by scalars. If response is complete $f_h = 1$, $r = S_2$ and b will be estimated from the second sample. We then have (using $\hat{y}_k = x_k b$):

$$\frac{N}{n_1}\sum\nolimits_{s_1} \hat{y}_k = Nb\bar{x}'$$

and

$$\frac{N}{n_2}\sum\nolimits_h f_h^{-1} \sum\nolimits_{r_h} (y_k - \hat{y}_k) = \frac{N}{n_2} \sum\nolimits_{r=s_2} (y_k - \hat{y}_k) = N\bar{y} - Nb\bar{x}$$

where $\bar{x}'$ and $\bar{x}$ is the mean of the x's from the first-phase and second-phase sample respectively. Hence we have the familiar regression estimator for the total in double sampling with simple random sampling at each phase and complete response.

$$\hat{t}_{cR^*} = N\left[\bar{y} + b(\bar{x}' - \bar{x})\right]$$

The quantity within square brackets is the estimator of the mean (e.g. Cochran [4, pp. 339]).

$$\hat{V}_2\left(\hat{t}_{cR^*}\right) = \sum\sum_r \Delta_{kl|s_1} e_k^{11} e_l^{11} / \pi_{kl|s_1} \quad \pi_{kl|s_1,s_2,\underline{m}}$$

and

$$\hat{V}_3\left(\hat{t}_{cR^*}\right) = \sum_h^{H_{s_2}} n_h^2 (1-f_h) S^2_{e^{11}_{r_h}} / m_h$$

The first component $\hat{V}_1$ (t_{cR^*}) represents the contribution to variance of the first-phase sample, the second $\hat{V}_2$ (t_{cR^*}) the contribution of the second-phase or intended sample. The component $\hat{V}_3$ (t_{cR^*}) gives variation due to nonresponse under the ERHG model. When response is complete, $f_h = 1$ and $\hat{V}_3$ (t_{cR^*}) = 0.

On the other hand, assuming the ERHG model holds, an approximate 100 (1–α)% confidence interval for the population total t that takes into account nonresponse can be constructed as $t_{cR^*} \pm Z_{1-\alpha/2} \sqrt{\hat{V}(t_{cR^*})}$. Here $Z_{1-\alpha/2}$ is the value preceeded by 100 (1–α/2)% of the area under the standardized normal curve.

Finally, in line with Särndal and Swensson [1] we interpret the variance of t_{cR^*} through an imaginary repeated three-step sampling process as follows. First, samples S_1 ($S_1 \subset U$) are obtained according to the fixed probability distribution P_1 (S_1). Every time a given S_1 is obtained, repeated samples S_2 ($2_2 \subset s_1$) can be selected according to the distribution P_2 (S_2). Finally, given S_2, repeated realizations of r can be obtained according to the ERHG model. It is assumed that the same stratified design i.e. the same number of strata, same stratification principle, ... etc. is used everytime the same second phase sample S_2 is selected, but that this need not be so for different samples.

Example

Double sampling with simple Random sampling in both phases.

As an example of the above let us consider a double sampling situation of the type frequently encountered in sampling texts. We have a population U of size N. A first-phase sample S_1 of size n_1 is taken by simple random sampling (without replacement). Measurement is made of characteristic x on all units in S_1. Given S_1, a second phase sample S_2 of size n_2 is selected also by simple random sampling. The objective is to estimate the population total t. Now, suppose some of the units failed to respond. Suppose further that the sampler managed to divide S_2 into H_{s_2} groups S_{2h} ($h = 1..., H_{s_2}$) such that units within each group respond with approximately the same probability. The response set in group S_{2h} and its size are r_h and m_h respectively

randomization schemes of some forms. Once the second-phase sample S_2 is drawn, it is partitioned into H_{s_2} disjoint groups the h^{th}, S_{2h} say, of size n_h. Let r_h be the subset of responding units in S_{2h} and m_h its size. Also, denote by r the union of r_h (h = $1,\ldots,H_{s_2}$). Furthermore, let us assume that for any group h:

$$P\left(k \in r \middle| S_1, S_2\right) = \pi_{k|s_1,s_2} = \Theta_{hs_2} \qquad k \in S_{2h}$$

$$P\left(k,l, \in r \middle| S_1, S_2\right) = \pi_{kl|s_1,s_2} = P\left(k \in r \middle| S_1, S_2\right) P\left(l \in r \middle| S_1, S_2\right) \qquad k \neq l$$

This model, which is a direct extension, to the double sampling case, of the RHG model simply assumes that units in a given group, S_{2h} say, respond independently with the same probability Θ_{hs_2}. A valuable feature of the model is that the structure of the groups as well as their number can depend on the actually selected sample S_2. This enables the sampling operations to which units in the actually selected sample are exposed to be reflected in the response model. We refer to the extended model by ERHG.

Now, the only difference between the ERGH model and the situation considered in §5 is that the former is merely an assumption. Hence, the theory given in §5 is directly applicable with $\underline{m}$ now interpreted as the vector of response counts and f_h as the response rate. Assuming the ERHG model holds, a regression estimator in double sampling in the presence of nonresponse is then given by t_{cR^*}. Its variance and estimate of variance, that take into account nonresponse, are given by (5.2b) and (5.2c) respectively. As indicated in §5, t_{cR^*} and $\hat{V}(t_{cR^*})$ are approximately unbiased for t and $V(t_{cR^*})$ respectively.

Some remarks are in order here. First we note that the estimate of the regression coefficient $\underline{B}$ is calculated from responding units only whereas the fitted values $\hat{y}_k$ used in t_{cR^*} are obtained for all $k \varepsilon S_1$. This is so since the data points $(y_k, \underline{x}_k)$ are observed only for $k \varepsilon r$ while $\underline{x}_k$ is observed for all $k \varepsilon S_1$.

The second remark has to do with the composition of the approximate estimate of variance $\hat{V}(t_{cR^*})$ given by (5.2c). This may be written as

$$\hat{V}\left(\hat{t}_{cR^*}\right) \cong \hat{V}_1\left(\hat{t}_{cR^*}\right) + \hat{V}_2\left(\hat{t}_{cR^*}\right) + \hat{V}_3\left(\hat{t}_{cR^*}\right)$$

where:

$$\hat{V}_1\left(\hat{t}_{cR^*}\right) = \sum \sum_r \Delta_{akl} y_k^1 y_1^1 / \pi_{akl} \; \pi_{kl|s_1} \quad \pi_{kl|s_1,s_2,\underline{m}}$$

6. Application to Nonresponse in Double Sampling for Regression

The problem of nonresponse in single-phase sampling has been approached in a number of different ways. One such approach is to accept the incomplete data set as it is and estimate parameters using methods based on a model for incomplete data. This may, however, pause a number of practical difficulties particularly with complex sample designs. Another approach is to try to make adjustment either by imputation or weighting so that a rectangular data set is achieved. This usually requires using adjustment cells (groups) formed on the basis of information recorded for all units in the sample. Little [3] compared three adjustment methods. Several models are suggested in the literature for the analysis of estimators based on adjustment cell techniques. A classical model in this respect (e.g. Cochran [4]) is the deterministic model of a dicotomized population in which the population is thought of as divided into two strata: one consisting of all units that would respond if they are selected in the sample, and the other of all units that would not respond if selected in the sample. This restricts the response probability to the values of zero and one. It is thus more realistic to consider a model in which nonresponse behaviour is looked at as probabilistic. The stochastic approach to nonresponse is adopted by a number of authors including Platek and Gray [5], Cassel et al. [2] and Oh and Scheuren [6]. In particular Oh and Scheuren employed a stochastic model in which the population is divided into a fixed and disjoint set of sub-populations with units in each subpopulation responding independently according to a Bernoulli Law. They described their assumption as quasi-randomization. A basic feature of this model is that the subpopulations (groups) are fixed. In general, this renders the model incapable of taking into account the survey operations to which units in the selected sample are exposed. To avoid this weakness, Särndal and Swensson [1] suggested what they called "the response homogeneiety groups (RHG) model". In the RHG model the selected sample, S say, is partitioned into H_s groups. This gives the sampler the opportunity to perform the partition with an eye given to such sampling variables as interviewer's skills, age,... etc.

The RHG model, with necessary modifications, will, in what follows, be applied to the problem of nonresponse in double sampling for regression. In this respect, the preceeding notation will be adhered to as much as possible.

We assume the following situation. We have the population $U = \{1,...,k,...N\}$. A double sample is to be taken to provide a regression estimate for the population total t. The first-phase sample S_1 ($S_1 \subset U$) is drawn as in §2.

The intended second phase sample S_2 ($S_2 \subset S_1$) is selected also as in §2. As before, the sample designs in the two phases $P_1(S_1)$ and $P_2(S_2)$ are actually imposed

if k and l belong to the same stratum but $f_h f_{h'}$, if k and l belong to different strata h and h'. Let $m_h \geq 1$ for all h. Then, using (2.9) and (2.10), we have by substituting (5.1) in (4.1a) and (4.1b):

$$\hat{t}_{cR^*} = \sum\nolimits_{s_1} \hat{y}_k / \pi_{ak} + \sum\nolimits_h f_h^{-1} \sum\nolimits_{r_h} \left(y_k - \hat{y}_k\right) / \pi_{ak} \pi_{k|s_1} \tag{5.2a}$$

and

$$V\left(\hat{t}_{cR^*}\right) \cong \sum\sum\nolimits_u \Delta_{akl} y_k^1 y_l^1 + E_a \left\{ \sum\sum\nolimits_{s_1} \Delta_{kl|s_1} E_k^{11} E_l^{11} \right\} + E_a E E_{\underline{m}} \left\{ \sum_h^{H_{s_2}} n_h^2 \left(1 - f_h\right) S^2_{E^{11}_{s_{2h}}} / m_h \middle| S_1 \right\} \tag{5.2b}$$

respectively.

The approximate estimator of variance, assuming all $m_h \geq 2$ is:

$$\hat{V}\left(\hat{t}_{cR^*}\right) \cong \sum\sum\nolimits_r \Delta_{akl} y_k^1 y_l^1 / \pi_{akl} \pi_{kl|s} \pi_{kl|s_1, s_2, \underline{m}} + \sum\sum\nolimits_r \Delta_{kl|s_1} e_k^{11} e_l^{11} / \pi_{kl|s_1} \pi_{kl|s_1, s_2, \underline{m}} + \sum_h^{H_{s_2}} n_h^2 \left(1 - f_h\right) S^2_{e^{11}_{r_h}} / m_h \tag{5.2c}$$

where

$$S^2_{E^{11}_{s_{2h}}} = \left[\sum\nolimits_{s_{2h}} E_k^{11^2} - \frac{\left(\sum\nolimits_{s_{2h}} E_k^{11} \right)^2}{n_h} \right] / \left(n_h - 1\right)$$

and

$$S^2_{e^{11}_{r_h}} = \left[\sum\nolimits_{r_h} e_k^{11^2} - \frac{\left(\sum\nolimits_{r_h} e_h^{11} \right)^2}{m_h} \right] / \left(m_h - 1\right)$$

and where e_k^{11} again replaces the unknown E_k^{11}. Estimators $t_{\pi^{**}}$, t_{R^*} and t_{cR^*} are the three-phase analogue of estimators obtained by Särndal and Swensson [1] for two-phase sampling.

We can write

$$\hat{t}^{\circ}_{R^*} - t = \left(\sum_{s_1} y_k^1 - \sum_u y_k\right)\left[\sum_{s_2} E_k^{11} - \sum_{s_1} E_k^1\right]$$

$$+\left(\sum_r E_k^{111} - \sum_{s_2} E_k^{11}\right) = A_{s_1} + A^*_{s_2} + A^*_r$$

say. Noting that $A^*_{s_2}$ and A^*_r differ, respectively, from A_{s_2} and A_r only in E_k replacing y_k we can with arguments identical to those used in proving (3.1a) – (3.1c), immediately see that $t^{\circ}_{R^*}$ is unbiased hence (4.1a) approximately unbiased. Also (4.1b) and (4.1c) can similarly be shown to hold. It should be noted, however, that we must replace E_k in (4.1c) by e_k since the former depends on the unknown $\underline{\hat{B}}$.

5. Regression Estimation with Stratification at Phase Three

In the treatment of nonresponse in double sampling to be discussed in §6 we shall need to consider regression estimation in three phase sampling with stratification in phase three. Hence we envisage the problem considered in §4 with the added assumption that stratified random sampling is employed in phase three. The following situation is assumed. Once the second phase sample S_2, of size n_{s_2} say, is drawn, information is recorded for all units in it. This information is then used to stratify S_2 in H_{s_2} strata the h^{th}, S_{2h} ($h = 1,\ldots H_{s_2}$), of size n_h. From S_{2h} a subsample r_h is realized. Whether unit k ($k \varepsilon S_{2h}$) is included in r_h or not is decided by a Bernoulli law in which the probability of inclusion is Θ_{hs_2} and in which inclusions are independent. The size of r_h, m_h say, is random and the Θ_{hs_2} are unknown and have to be estimated from the sample. Note that although information is available on all units in S_1, stratification has to be based on S_2 the set to be sampled in the third phase.

Our problem now is to arive at a regression estimator for this case. Let $\underline{m} = (m_1, m_2,\ldots,m_{H_{s_2}})$ be the vector of realized counts. Then $\underline{m}$ is a random vector. For fixed S_2 and $\underline{m}$, n_h ($h = 1,\ldots,H_{s_2}$) will also be fixed and r_h will be a simple random sample of m_h units from n_h units. Hence for fixed S_2 and $\underline{m}$ we have for k, $l \varepsilon S_{2h}$ ($S_{2h} \subset S_2$):

$$P\left(k \in r_h \middle| S_1, S_2, \underline{m}\right) = \pi_{k|s_1,s_2\underline{m}} = \frac{m_h}{n_h} = f_h \tag{5.1a}$$

and for $k \neq l$

$$P\left(k,l \in r_h \middle| S_1, S_2, \underline{m}\right) = \pi_{kl|s_1,s_2\underline{m}} = f_h\left(m_h - 1\right)/\left(n_h - 1\right) \tag{5.1b}$$

However $(y_k, \underline{x}_k)$ is observed only for $k \in r$ and each k carries a weight π^{**-1}. Hence we may estimate $\underline{B}$ by $\underline{b}$ where

$$\underline{b} = \left(\sum_r \underline{x}_k \underline{x}'_k / \left(\sigma_k^2 \pi_k^{**}\right)\right)^{-1} \sum_r x_k y_k / \left(\sigma_k^2 \pi_k^{**}\right) \qquad k \in r$$

and the residuals will then be $e_k = y_k - \underline{x}'_k \underline{b}$ $\qquad k \epsilon r$

We now establish the following result.

Result (2)

In three-phase sampling where information is recorded on the auxiliary variables $x_1, x_2, \ldots, x_q$ for all $k \in S_1$, and on the characteristic y for all $k \in r$ the regression estimator

$$\hat{t}_{R*} = \sum_{s_1} \hat{y}_k / \pi_{ak} + \sum_r \left(y_k - \hat{y}_k\right) / \pi_k^{**} \tag{4.1a}$$

where $\hat{y}_k = \underline{x}'_k \underline{b}$ $(k \epsilon s_1)$ is approximately unbaised for the total t with variance

$$V\left(\hat{t}_{R*}\right) \cong V\left(\hat{t}^{\circ}_{R*}\right) = \sum\sum_u \Delta_{akl} y_k^1 y_1^1 + E_a\left[\sum\sum_{s_1} \Delta_{kl|s_1} E_k^{11} E_1^{11}\right] \tag{4.1b}$$

$$+ \; E_a E\left[\sum\sum_{s_2} \Delta_{kl|s_1,s_2} E_k^{111} E_i^{111} \middle| S_1\right]$$

where $t^{\circ}_{R\bullet}$ is obtained from $t_{R\bullet}$ by replacing $\underline{b}$ in $\hat{y}_k$ by $\hat{\underline{B}}$. An unbiased estimator of $V(t^{\circ}_{R\bullet})$ and hence an approximately unbiased estimator of $V(t_{R\bullet})$ is

$$\hat{V}\left(\hat{t}^{o}_{R*}\right) = \sum\sum_r \Delta_{akl} y_k^1 y_1 1 / \pi_{kl}^{**}$$

$$+ \sum\sum_r \Delta_{kl|s_1,s_2} e_k^{11} e_1^{11} / \pi_{kl|s_1,s_2} \tag{4.1c}$$

$$+ \sum\sum_r \Delta_{kl|s_1,s_2} e_k^{111} e_1^{111} / \pi_{kl|s_1,s_2}$$

Proof

Replacing $\hat{y}_k$ in (4.1a) by $y^{\circ}_k = \underline{x}'_k \hat{\underline{B}}$ we have

$$\hat{t}^{\circ}_{R*} = \sum_{s_1} y_k^1 - \sum_{s_1} E_k^1 + \sum_{s_1,s_2} E_k^{111}$$

where V_a denotes variance in phase one. Letting c_k take the value 1 if k is drawn in phase one but zero otherwise and using (2.1) - (2.3) we have

$$V\left(A_{s_1}\right) = \sum\sum_u y_k^1 y_l^1 \operatorname{cov}(c_k, c_l) = \sum\sum_u \Delta_{akl} y_k^1 y_l^1 \tag{3.4}$$

By similar arguments it can be shown that

$$V\left(A_{s_2}\right) = E_a\left[\sum\sum_{s_1} \Delta_{kl|s_1} y_k^{11} y_l^{11}\right] \tag{3.5}$$

and

$$V(A_r) = E_a E\left[\sum\sum_{s_2} \Delta_{kl|s_1,s_2} y_k^{111} y_l^{111} | S_1\right] \tag{3.6}$$

substituting (3.4) - (3.6) in (3.3) completes the proof.

That the three terms of (3.1c) are unbiased for the respective terms in (3.1b) can be shown by a repeated use of the indicator variable technique and the utilization of the rule for expected value in three phases of selection mentioned above.

4. Regression Estimation in Three-Phase Sampling

In this section we employ result (1) to arrive at a regression estimator for the population total together with expressions of its variance and estimator of variance in three-phase sampling. Discussion is confined to the following situation. A three-phase sample of the type described in §2 is drawn. Information is recorded, on each of q auxiliary vriables $x_1, x_2, \ldots, x_q$ for each unit $k \in S_1$. The relation between the characteristic y and the auxiliary variables can be represented by the finite population regression model [2]:

$$E(y_k) = \underline{x}'\underline{B} \qquad V(y_k) = \sigma_k^2 \qquad k \in U$$

and all y_k's independent. Here $\underline{x}_k = (x_{k1}, \ldots, x_{kq}')'$. Our problem is to estimate t through a regression estimator. Had data points $(y_k, \underline{x}_k)$ been available for all $k \in U$ the weighted least squares estimator $\hat{\underline{B}}$ and residuals E_k would have been

$$\hat{\underline{B}} = \left(\sum_u \underline{x}_k \underline{x}_k' / \sigma_k^2\right)^{-1} \sum_u \underline{x}_k y_k / \sigma_k^2 \quad , \quad E_k = y_k - \underline{x}_k' \hat{\underline{B}} \qquad k \in U$$

so that $E(A_{s_2}) = 0$. Also

$$E\left(\sum_r y_k^{111}\right) = E_a E\left[\sum_{s_2} y_k^{11} | S_1\right] = E_a\left(\sum_{s_1} y_k^1\right) = \sum_u y_k = E\left(\sum_{s_2} y_k^{11}\right)$$

Hence $E(A_r) = 0$. So that

$$E(\hat{t}_{\pi^{**}} - t) = 0$$

meaning that $\hat{t}_{\pi^{**}}$ is unbiased for t.

To prove the variance result (3.1b) we first note that

$$V(\cdot) = E_a E[V(\cdot | S_1, S_2 | S_1] + V^*(E(\cdot | S_1, S_2)) \tag{3.2}$$

where: $V^*\left(E(\cdot | S_1, S_2)\right) = E_a E\left[\left(E(\cdot | S_1, S_2)\right)^2 | S_1\right] - \left[E_a E\left(E(\cdot | S_1, S_2) | S_1\right)\right]^2$

and where $V(\cdot \mid S_1, S_2)$ denotes variance in phase three given S_1 and S_2. Now:

$$V(\hat{t}_{\pi^{**}}) = V(\hat{t}_{\pi^{**}} - t) = V(A_{s_1}) + V(A_{s_2}) + V(A_r) \tag{3.3}$$

since all covariances vanish because

$$E(A_{s_1}) = E(A_{s_2}) = E(A_1) = 0$$

and

$$E(A_{s_1} A_{s_2} | S_1) = 0;\ E(A_{s_1} A_r | S_1, S_2) = E(A_{s_2} A_r | S_1, S_2) = 0$$

On the other hand using (3.2) and noting that:

$$V(A_{s_1} | S_1, S_2) = 0, E(A_{s_1} | S_1, S_2) = A_{s_1}, E(A_{s_1}^2 | S_i) = A_{s_1}^2, E_a(A_{s_1}) = 0$$

we have

$$V(A_{s_1}) = E_a E\left[(E(A_{s_1} | S_1, S_2))^2 | S_1\right] = V_a\left(\sum_{sl} y_k^1\right)$$

is a design unbiased estimator of the population total t with variance

$$V(\hat{t}_{\pi^{**}}) = \sum\sum_{u} \Delta_{akl} y_k^{I} y_l^{I} + E_a\left[\sum\sum_{s_1} \Delta_{kl|s_1} y_k^{II} y_l^{II}\right] + E_a E\left[\sum\sum_{s_2} \Delta_{kl|s_1,s_2} y_k^{III} y_l^{III} | s_1\right] \tag{3.1b}$$

where $\Sigma\Sigma_u$ is used for $\Sigma_{k\in u}\Sigma_{l\in u}$, $E_a(\cdot)$ for expectation with respect to sampling design in phase one, and $E(\cdot \mid s_1)$ for expectation with respect to sampling design in phase two given S_1. Furthermore, the estimator $\hat{V}(t_{\pi^{**}})$ where:

$$\hat{V}(\hat{t}_{\pi^{**}}) = \sum\sum_{r} \Delta_{akl} y_k^{I} y_l^{I} / \pi_{kl}^{**} + \sum\sum_{r} \Delta_{kl|s_1} y_k^{II} y_l^{II} / \pi_{kl|s_1}^{**} \pi_{kl|s_1,s_2} + \sum\sum_{r} \Delta_{kl|s_1,s_2} y_k^{III} y_l^{III} / \pi_{kl|s_1 s_2} \tag{3.1c}$$

is design unbiased for the variance.

Proof

To show that (3.1a) is unbiased consider the identity:

$$\hat{t}_{\pi^{**}} - t = \left(\sum_{s_1} y_k^{I} - \sum_{u} y_k\right) + \left(\sum_{s_2} y_k^{II} - \sum_{s_1} y_k^{I}\right) + \left(\sum_{r} y_k^{III} - \sum_{s_2} y^{II}\right)$$
$$= A_{s_1} + A_{s_2} + A_r$$

Now utilizing the result $E(\cdot) = E_a E[E(\cdot \mid S_1, S_2) \mid S_1]$ where $E(\cdot \mid S_1, S_2)$ denotes expectation over the sampling distribution in the third phase given S_1 and S_2, we have

$$E\left(\sum_{s_1} y_k^{I}\right) = E_a E\left[E\left(\sum_{s_1} y_k^{I} | S_1, S_2\right) | S_1\right] = E_a E\left[\sum_{s_1} y_k^{I} | S_1\right]$$
$$= E_a\left(\sum_{s_1} y_k^{I}\right) = \sum_{u} y_k$$

Hence $E(A_{s_1}) = 0$. Similarly

$$E\left(\sum_{s_2} y_k^{II}\right) = E_a E\left[\sum_{s_2} y_k^{II} | S_1\right] = E_a\left(\sum_{s_1} y_k^{I}\right) = \sum_{u} y_k = E\left(\sum_{s_1} y_k^{I}\right)$$

where $\pi_{kk|s_1,s_2} = \pi_{k|s_1,s_2}$ It is assumed that $\pi_{k|s_1,s_2} > 0$ for all $k \varepsilon s_2$. and $\pi_{kl|s_1,s_2} > 0$ for all $k \neq l \in s_2$. We also put

$$\Delta_{kl|s_1,s_2} = \pi_{kl|s_1,s_2} - \pi_{k|s_1,s_2}\pi_{l|s_1,s_2} \tag{2.9}$$

Characteristic y is observed only on units $k\varepsilon r$.

It should be noted that in all three phases the sampling design is arbitrary and the sample size not neessarily fixed.

For ease of reference let us set

$$\begin{aligned}
&\pi_k^* = \pi_{ak} \quad \pi_{k|s_1} && \pi_{kl}^* = \pi_{akl} \quad \pi_{kl|s_1} \\
&\pi_k^{**} = \pi_{ak} \quad \pi_{k|s_1}\pi_{k|s_1,s_2} && \pi_{kl}^{**} = \pi_{akl} \quad \pi_{kl|s_1} \quad \pi_{kl|s_1,s_2} \\
&\Delta_{kl}^{**} = \pi_{kl}^{**} - \pi_k^{**} \quad \pi_l^{**}
\end{aligned} \tag{2.10}$$

We also use $^{(1)}$, $^{(11)}$ and $^{(111)}$ to denote first, second and third phase expansion. Thus for example

$$\begin{aligned}
&y_k^1 = y_k / \pi_{ak} \ , \ y_k^{11} = y_k^1 / \pi_{k|s_1} = y_k / \pi_k^* \\
&y_k^{111} = y_k^{11} / \pi_{k|s_1s_2} = y_k / \pi_k^{**}
\end{aligned} \tag{2.11}$$

and so on.

3. A General Estimator for Three Phases of Selection

In this section we provide a general estimator for three phases of selection together with an expression for its variance and estimate of variance. The estimator is a π^{**} - expanded sum estimator that is the three-phase analogue of the π^* - expanded sum estimator of Särndal & Swensson [1] in the two-phase case.

Result (1)

In three-phase sampling

$$\hat{t}_{\pi^{**}} = \sum_r y_k^{111} \tag{3.1a}$$

that of units k & l

$$\pi_{akl} = \sum_{s_1 \ni k,l} P_1(S_1) \tag{2.2}$$

with $\pi_{akk} = \pi_{ak}$, we put

$$\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al} \tag{2.3}$$

It is assumed that $\pi_{akl} > 0$ for all $k \neq l$ and $\pi_{ak} > 0$ for all k.

Phase two

Given S_1 is drawn in phase one, a sample S_2 ($S_2 \subset S_1$) of size n_{s_2} is drawn according to the conditional probability distribution $\{P_2(S_2 \mid S_1\}, S_2 \subset S_1\}$

The conditional inclusion probability of unit k given s_1 is

$$\pi_{k|s_1} = \sum_{s_2 \ni k} P_2(S_2|S_1) \tag{2.4}$$

that of units k & l is

$$\pi_{k|s_1} = \sum_{s_2 \ni k,l} P_2(S_2|S_1) \tag{2.5}$$

with $\pi_{kk|s_1} = \pi_{k|s_1}$. We set

$$\Delta_{kl|s_1} = \pi_{kL|s_1} - \pi_{k|s_1}\pi_{l|s_1} \tag{2.6}$$

It is assumed that $\pi_{k|s_1} > 0$ for all $k \varepsilon s_1$ and $\pi_{kl|s_1} > 0$ for all $k \neq l \varepsilon S_1$.

Phase three

Given that S_1 and S_2 are selected in phase one and phase two respectively, a third sample r ($r \subset S_2$) is selected in accordance with the conditional probability distribution $\{P_3(r \mid S_1, S_2), S_1 \subset U, S_2 \subset S_1\}$. The conditional inclusion probability of unit k given S_1 and S_2 is

$$\pi_{k|s_1,s_2} = \sum_{r \ni k} P_3(r|S_1,S_2) \tag{2.7}$$

that of k & l is

$$\pi_{kl|s_1,s_2} = \sum_{r \ni k,l} P_3(r|S_1,S_2) \tag{2.8}$$

(usually much smaller) sample is taken and observations are made on the variate of the study, y say, for every unit in the second sample. The variates x and y are assumed correlated. Variate x is usually chosen such that information about it is relatively easy to obtain. It may, for example, be a rough estimate or a characteristic, like age and sex, that is available in files or registration lists.

Frequently, however, the researcher is faced with the problem that some of the units in the second sample (the 'intended' sample) fail to respond. Ignoring the non-respondents' set altogether may lead to bias in the estimate and underestimation of variance. Attempts in the literature to provide estimation theory that allows for non-response are restricted to single-phase sampling. It is this point that motivated the present paper.

We first discuss estimation of the population total in a three-phase sampling set-up with arbitrary design admitted at each phase. This is done in §3. In §4 and §5 results in §3 are used to develop a regression estimator without stratification and a regression estimator with stratification at phase three respectively. Finally in §6 we apply results of §5 to the treatment of nonresponse in double sampling for regression. It should be pointed out that although discussion §3 – §5 can be looked at as providing a theory for three phases of selection, it is primarily meant to pave the way for the treatment of nonresponse in double sampling. Three-phase sampling can hardly be justified in a practical situation.

2. Preliminaries

Consider the population $U = \{1,2,\ldots,k,\ldots,N\}$. Let a characteristic y be observed on each unit in U with y_k the value of the characteristic for unit **k**. Our objective is to estimate the population total $t = \Sigma_u y_k$ (where we write Σ_u for $\Sigma_{k\varepsilon u}$). The estimation is to be based on a random sample r obtained through three phases of selection as follows:

Phase one

A sample S_1 ($S_1 \subset U$) of size n_{s_1} is drawn in accordance with the probability distribution $\{P_1(S_1), S_1 \subset U\}$.

The inclusion probability of unit k is

$$\pi_{ak} = \sum_{s_1 \ni k} P_1(S_1) \tag{2.1}$$

J. King Saud Univ., Vol. 7, *Admin. Sci.* (1), pp. 19-35 (A.H. 1415/1995)

Nonresponse in Double Sampling for Regression

Z.A. El-Beshir
Department of Quantitative Methods
College of Administrative Sciences
King Saud University, Riyadh, Saudi Arabia

(Received 6/9/1413; accepted for publication 16/4/1414 A.H.)

Abstract. Some of the results derived by Särndal & Swensson for two-phase sampling are generalized to incorporate a third phase of selection. The resulting theory is then applied to the treatment of nonresponse in double sampling for regression.

Introduction

Särndal & Swensson [1] discussed estimation in two-phase (or double) sampling assuming arbitrary design in both phases. An attractive feature of their approach is that it can, with some modifications, be applied to the problem of nonresponse in single-phase sampling. Thus by regarding the respondents' group as a subsample 'drawn' from the intended sample by a certain probabilistic response mechanism. Särndall & Swensson [1] showed that the nonresponse problem can be viewed as a two-phase sampling problem to which their general two-phase sampling theory is directly applicable. In the present paper, it is shown that, by extending their approach to allow for a third phase of selection, it is possible to generalize their results to include the treatment of nonresponse in two-phase sampling.

Our interest will be focused only in two-phase sampling for regression. In practice, this situation may arise when a researcher aims at a regression estimate based on, say the population mean $\bar{X}$ of an auxiliary variate x. If no prior information is available on x, it may pay to devote some of the survey resources to a preliminary large sample in which measurement on x alone is taken. From this sample, a second