Survey Methodology, June 2007
Vol. 33, No. 1, pp. 81-85
Statistics Canada, Catalogue No. 12-001

Handling survey nonresponse in cluster sampling

Jun Shao [1]

Abstract

In surveys under cluster sampling, nonresponse on a variable is often dependent on a cluster level random effect and, hence, is nonignorable. Estimators of the population mean obtained by mean imputation or reweighting under the ignorable nonresponse assumption are then biased. We propose an unbiased estimator of the population mean by imputing or reweighting within each sampled cluster or a group of sampled clusters sharing some common feature. Some simulation results are presented to study the performance of the proposed estimator.

Key Words: Nonignorable nonresponse; Random-effect-based nonresponse; Imputation; Collapsing clusters.

1. Introduction

Nonresponse exists in most survey problems. The probability of having a nonrespondent in a survey item (variable) y typically depends on the unobserved value of y, which creates a great challenge in handling nonrespondents. Commonly used procedures for handling nonresponse (such as reweighting and imputation) are all based on the assumption that nonresponse is ignorable conditional on an auxiliary variable. More precisely,

$$P(y \text{ is a respondent} \mid y, z) = P(y \text{ is a respondent} \mid z), \quad (1)$$

where z is an auxiliary variable whose values are observed for all sampled units in the survey. That is, conditional on z, the value of y and its response status are statistically independent. Assumption (1) is referred to as the unconfounded response mechanism by Lee, Rancourt and Särndal (1994). Using the terminology in Rubin (1976), nonresponse under (1) is ignorable conditional on z.

There are situations in which it is difficult to find a variable z to satisfy (1). The purpose of this article is to study a method of handling nonresponse when cluster sampling is used, assuming that a variable z satisfying (1) is not available. In cluster sampling, sampling is carried out in two stages; the first stage sampled units are clusters containing units that are sampled in the second stage. Cluster sampling is used because of economic considerations. It is necessary when no reliable list of the second stage units in the population is available (for example, there is no complete list of people but a list of households is available). Under cluster sampling, the variable of interest y may be decomposed as $y = \mu + b + e$, where μ is an unknown overall mean of y, b is a cluster level random effect (all units in the same cluster share the same random effect b), and e is a within-cluster random effect. In many cases, the dependence of the value of y and its response status is through the unobserved cluster level random effect b:

$$P(y \text{ is a respondent} \mid y, b) = P(y \text{ is a respondent} \mid b), \quad (2)$$

i.e., if b were observed, then we would have assumption (1) with $z = b$. For example, suppose that clusters are households and a single person completes survey forms for all sampled persons in a household. It is likely that the response probability depends on the household level variable b, not on the within household variable e.

Assumption (2) was first used by Wu and Carroll (1988) in a health problem where the clusters have a longitudinal (repeated-measure) structure. They called (2) informative censoring (missing) and proposed a method under some parametric assumptions on the probability $P(y$ is a respondent$\mid b)$ and the distribution of y. Later, Little (1995) called this type of missing mechanism the nonignorable random-coefficient-based missing mechanism. Thus, assumption (2) will be referred to as nonignorable random-effect-based response mechanism. Since b is not observed, response mechanism (2) is actually nonignorable.

For survey data, it is difficult to impose any parametric model on the distribution of y. Furthermore, it is also difficult to fit a parametric model for the response mechanism under (2), since b is not observed. After introducing some details on the sampling design and our assumptions, we propose in Section 2 a method for the estimation of the population mean of y under response mechanism (2), without requiring a parametric model for the response mechanism. It is assumed that y follows a random (cluster) effect model, but there is no parametric assumption on the distribution of y. Results from a simulation study are presented in Section 3 for examining the performance of the proposed estimator. Some discussions are given in the last section.

1. Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A.

2. Main results

Let S be a sample of clusters of size n from a population P. Within the i^{th} sampled cluster, let S_i be the second stage sample of size $m_i \geq 2$ from a population P_i. For sampled unit $j \in S_i$, a survey weight w_{ij} is constructed (from the specification of the sampling design) so that when there is no nonresponse, $\hat{Y} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}$ is an unbiased estimator of the population total Y on any variable y, *i.e.*, $E_s(\hat{Y} - Y) = 0$, where y_{ij} is the y-value of unit j in cluster i $Y = \sum_{i \in P} \sum_{j \in P_i} y_{ij}$, and E_s is the expectation with respect to repeated sampling.

Let y be the variable of interest. We adopt an imputation model approach, *i.e.*, we assume that each y_{ij} in the population is a random variable with

$$y_{ij} = \mu_i + b_i + e_{ij}, \tag{3}$$

where μ_i is an unknown parameter, b_i is an unobserved cluster level random effect with mean 0 and a finite variance, e_{ij} is an unobserved within cluster random effect with mean 0 and a finite variance, and b_i's and e_{ij}'s are independent. Note that the distribution of y_{ij} may vary with (i, j).

Let δ_{ij} be the response indicator for y_{ij} ($\delta_{ij} = 1$ if y_{ij} is a respondent and $\delta_{ij} = 0$ if y_{ij} is a nonrespondent). We adopt the approach in Shao and Steel (1999), *i.e.*, δ_{ij} is defined for every unit in the population and nonresponse mechanism is part of the model. Let $\boldsymbol{\delta}_i$ be the vector containing δ_{ij}, $j \in S_i$, and $\boldsymbol{y}_i$ be the vector containing y_{ij}, $j \in S_i$. We assume the following nonignorable random-effect-based response mechanism: for every sample,

$$P_m(\boldsymbol{\delta}_i \mid b_i, \boldsymbol{y}_i) = P_m(\boldsymbol{\delta}_i \mid b_i), \quad i \in S, \tag{4}$$

where P_m is the probability with respect to the model and $P_m(\xi \mid \eta)$ denotes the conditional distribution of ξ given η. That is, conditional on b_i, $\boldsymbol{y}_i$ and $\boldsymbol{\delta}_i$ are independent. (Unconditionally, they may be dependent.) We assume that the stochastic mechanism with respect to the model is independent of the sampling mechanism so that $E_s E_m(X) = E_m E_s(X)$ as long as X is integrable, where E_m is the expectation with respect to P_m.

Furthermore, we assume that

$$\text{for any } i \in S, \text{ at least one } \delta_{ij} \text{ is } 1. \tag{5}$$

That is, each cluster has at least one respondent. Without this assumption (or some other assumption), the population total Y may not be estimable. More discussion is given in Section 4.

If we assume ignorable nonresponse, *i.e.*, $P_m(\delta_{ij} = 1 \mid y_{ij}) = P_m(\delta_{ij} = 1)$, then a commonly used procedure is to impute each nonrespondent by the mean $\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} y_{ij} / \sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij}$, which leads to the following estimator of Y:

$$\hat{Y}_r = \sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \tilde{w}_{ij} y_{ij}, \; \tilde{w}_{ij} = w_{ij} \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \Big/ \sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} \right). \tag{6}$$

Under assumptions (3)-(5),

$$\begin{aligned} E_s E_m(\hat{Y}_r) &= E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \tilde{w}_{ij} (\mu_i + b_i + e_{ij}) \right) \\ &= E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \tilde{w}_{ij} \mu_i \right) + E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \tilde{w}_{ij} b_i \right), \end{aligned} \tag{7}$$

where the last equality follows from

$$\begin{aligned} E_m(\delta_{ij} \tilde{w}_{ij} e_{ij}) &= E_m[E_m(\delta_{ij} \tilde{w}_{ij} e_{ij} \mid b_i)] \\ &= E_m[E_m(\delta_{ij} \tilde{w}_{ij} \mid b_i) E_m(e_{ij} \mid b_i)] = 0 \end{aligned} \tag{8}$$

under (4). The first term in (7) is equal to

$$E_s E_m \left[\left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} \mu_i \right) \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \right) \Big/ \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} \right) \right]$$

which is approximately equal to (when n is large)

$$\frac{E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} \mu_i \right) E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \right)}{E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} \right)} = \frac{E_s \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \mu_i E_m(\delta_{ij}) \right) E_s \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \right)}{E_s \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} E_m(\delta_{ij}) \right)}.$$

Note that

$$E_s E_m(Y) = E_m(Y) = \sum_{i \in P} \sum_{j \in P_i} \mu_i = E_s \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \mu_i \right).$$

Hence, either $\mu_i = \mu$ for all i or $E_m(\delta_{ij})$ does not depend on (i, j) implies that the expectation of the first term in (7) is approximately equal to the expectation of Y. However, $E_m(\delta_{ij} \tilde{w}_{ij} b_i) \neq 0$ in general, because δ_{ij} and b_i

are dependent. Thus, the second term in (7) is not 0 and, hence, $\hat{Y}_r$ defined by (6) is biased under the nonignorable random-effect-based nonresponse. This bias does not go away asymptotically as $n \to \infty$ and/or $m_i \to \infty$ for all *i*.

Recognizing that the problem with $\hat{Y}_r$ is that imputation is done over the entire sample whereas the nonresponse depends on a cluster level random effect, we can find an unbiased estimator by performing imputation within each cluster. This would have been a natural way of imputing if the cluster random effect b_i were observed. If we impute a nonrespondent y_{ij} in cluster *i* by the cluster mean $\sum_{j\in S_i}\delta_{ij}w_{ij}y_{ij}/\sum_{j\in S_i}\delta_{ij}w_{ij}$, then the resulting estimator is

$$\hat{Y}_c = \sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\,\overline{w}_{ij}y_{ij},$$

with

$$\overline{w}_{ij} = w_{ij}\left(\sum_{j\in S_i} w_{ij}\Big/\sum_{j\in S_i}\delta_{ij}w_{ij}\right). \qquad (9)$$

Assumption (5) ensures that $\overline{w}_{ij}$ is well defined. Note that

$$\begin{aligned} E_sE_m(\hat{Y}_c) &= E_sE_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\,\overline{w}_{ij}\,\mu_i\right) + E_sE_m\left(\sum_{i\in S}\sum_{j\in S_i}\delta_{ij}\,\overline{w}_{ij}\,b_i\right) \\ &= E_sE_m\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\,\mu_i\right) + E_sE_m\left(\sum_{i\in S}\sum_{j\in S_i} w_{ij}\,b_i\right) \\ &= E_m(Y), \end{aligned}$$

where the first equality follows from assumption (3) and the fact that, under assumption (4), result (8) still holds with $\tilde{w}_{ij}$ replaced by $\overline{w}_{ij}$, the second equality follows from the definition of $\overline{w}_{ij}$ and the fact that μ_i and b_i do not depend on *j*, and the last equality follows from $E_m(b_i) = 0$. Hence, $\hat{Y}_c$ is an unbiased estimator of *Y*.

Since imputation is done within each cluster, the estimator defined by (9) seems inefficient when some cluster sample sizes m_i are very small. This worry, however, is not necessary in the case where $w_{ij} = w_i$ for all *j* (*e.g.*, the second stage sampling is with equal probability). When $w_{ij} = w_i$ for all *j*, imputation leading to $\hat{Y}_c$ in (9) is actually done in a much larger class, a group of clusters sharing something in common. Let $\overline{\delta}_i = m_i^{-1}\sum_{j\in S_i}\delta_{ij}$ be the response rate within cluster *i* and let

$$G_l = \{i\in S : m_i = m,\ \overline{\delta}_i = k/m\},\quad l = (k, m), k \le m. \qquad (10)$$

For each $l = (k, m), G_l$ in (10) is the group of sample clusters having the same $m_i = m$ and $\overline{\delta}_i = k$. If $w_{ij} = w_i$ for all *j*, then, for $i \in G_l$ with $l = (k, m)$,

$$\begin{aligned} \overline{w}_{ij} &= w_{ij}\left(\sum_{j\in S_i} w_{ij}\Big/\sum_{j\in S_i}\delta_{ij}w_{ij}\right) \\ &= w_i\left(\sum_{j\in S_i} w_i\Big/\sum_{j\in S_i}\delta_{ij}w_i\right) \\ &= w_i/\overline{\delta}_i \\ &= w_i/(k/m) \\ &= w_i\left(\sum_{i\in G_l} m_iw_i\right)\Big/\left(\sum_{i\in G_l}\frac{k}{m}m_iw_i\right) \\ &= w_i\left(\sum_{i\in G_l} m_iw_i\right)\Big/\left(\sum_{i\in G_l}\overline{\delta}_i m_iw_i\right) \\ &= w_i\left(\sum_{i\in G_l}\sum_{j\in S_i} w_i\right)\Big/\left(\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_i\right) \\ &= w_{ij}\left(\sum_{i\in G_l}\sum_{j\in S_i} w_{ij}\right)\Big/\left(\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_{ij}\right). \end{aligned}$$

Therefore, imputation leading to $\hat{Y}_c$ in (9) is actually done within each group G_l when $w_{ij} = w_i$ for all *j*, *i.e.*, a nonrespondent in S_i is imputed by the sample mean of the respondents in G_l, $\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_{ij}y_{ij}/\sum_{i\in G_l}\sum_{j\in S_i}\delta_{ij}w_{ij}$.

When w_{ij} varies with *j* for some *i*'s, some additional conditions are needed in order to combine clusters. A discussion is given in Section 4.

We end this section with a discussion of variance estimation, since most surveys require a variance estimator for each point estimator. A variance formula or its approximation (as $n \to \infty$) for $\hat{Y}_c$ can be derived, which may require more details on the sampling design. When the first stage sample size *n* is large, $m_i \le m$ for all *i* and a fixed integer *m*, and n/N is small, where *N* is the size of *P*, we can apply the adjusted jackknife method as described in Rao and Shao (1992). More precisely, we can follow the following steps.

1. Create *n* jackknife replicates, where the i^{th} replicate is obtained by deleting the i^{th} cluster and adjusting the weights to $w_{kj}^{(i)}, k \ne i, i = 1, \ldots, n$, according to the sampling design. For example, if the first stage sampling is a stratified sampling, then $w_{kj}^{(i)} = w_{kj}$ if *k* and *i* are not in the same stratum and $w_{kj}^{(i)} = n_hw_{kj}/(n_h - 1)$ if *k* and *i* are in the same stratum *h*, where n_h is the stratum size.

2. Re-impute the nonrespondents in the i^{th} jackknife replicate using the respondents in the i^{th} jackknife replicate, $i = 1, \ldots, n$.

3. Compute $\hat{Y}_{c,i}$ the same as $\hat{Y}_c$ but based on the i^{th} re-imputed jackknife replicate, $i = 1, \ldots, n$.

4. Compute the jackknife variance estimator for $\hat{Y}_c$ using a standard jackknife formula (*e.g.*, Shao and Tu 1995, Chapter 6). For example, if the first stage sampling is a stratified sampling with *H* strata, then a jackknife variance estimator is

$$v = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{i \in S_h} \left(\hat{Y}_{c,i} - \frac{1}{n} \sum_{k \in S} \hat{Y}_{c,k} \right)^2,$$

where S_h is the sample from the h^{th} stratum and n_h is the size of S_h.

3. Simulation results

We now present some results from a simulation study to examine the performance of the estimators $\hat{Y}_r$ and $\hat{Y}_c$.

We create a finite population similar to the elementary school teacher population in Maricopa County, Arizona (Lohr 1999, pages 446-447). The finite population contains 311 clusters (schools). In each cluster, the second stage units are teachers. The cluster size (the number of teachers) varies from 6 to 59 and, hence, the first stage sampling is an unequal probability sampling with probability proportional to cluster size. The first stage sampling is with replacement and the sample size is 31. The second stage sampling is a simple random sampling of size 6 (for any cluster) without replacement.

For each teacher, the variable of interest is the minutes spent per week in school on preparation. The values of y_{ij} for this variable in the simulation are generated according to model (3), where μ_i is the mean minutes spent per week in school on preparation for the i^{th} school, b_i is a random effect of the i^{th} school, and e_{ij} is a random effect of the j^{th} teacher in the i^{th} school. The values of μ_i's are the sample means in the data set in Lohr (1999, pages 446-447), which vary from 25.52 to 42.18 with a mean of 33.76 and a median of 33.47. The value of b_i is generated according to $b_i = 8.31(X_i - 2)$, where X_i has the gamma distribution with shape parameter 2 and scale parameter 1. The value of e_{ij} is generated from the normal distribution with mean 0 and standard deviation 2.27. The b_i's and e_{ij}'s are independently generated. The values of $y_{ij} = \mu_i + b_i + e_{ij}$ are generated in each simulation run so that we can evaluate the biases and standard errors of estimators using joint probability under sampling and models (3)-(5).

For sampled units, nonrespondents are generated according to (4) and (5). That is, each sampled cluster has one respondent and the response status of the rest of the sampled units in each cluster are independently determined by $P(y_{ij} \text{ is missing} \mid b_i) = e^{b_i - 1}/(1 + e^{b_i - 1})$. The mean nonresponse probability is 33.76%.

For the estimation of the finite population mean, a simulation of 1,000 runs shows that, when $\hat{Y}_r$ is used, the bias, standard error, and root mean squared error are -2.89, 1.32, and 3.17, respectively, and the relative bias $E(\hat{Y}_r - Y)/E(Y)$ is -8.5%; when $\hat{Y}_c$ is used, the bias, standard error, and root mean squared error are 0.12, 1.81, and 1.82, respectively, and the relative bias $E(\hat{Y}_c - Y)/E(Y)$ is 0.3%. This simulation result supports our theory, *i.e.*, $\hat{Y}_c$ is approximately unbiased but $\hat{Y}_r$ is biased. In this case, $\hat{Y}_c$ has a larger standard error than $\hat{Y}_r$, but $\hat{Y}_r$ has a much larger root mean squared error than $\hat{Y}_c$ due to its large bias.

4. Discussions

Without the assumption that each sampled cluster has at least one respondent, the population total may not be estimable unless some other assumption is added. Under the nonresponse mechanism (4), when all observations in a cluster are nonrespondents, no information in that cluster can be recovered from observed data in other clusters unless some additional assumption is made. For example, one may assume that the population of clusters with no respondent is similar to that of clusters with 1 respondent, in which case one can collapse clusters by distributing the weights of clusters with 0 respondent to the weights of clusters with 1 respondent. Another approach is to assume a model so that we can extrapolate results to clusters with no respondent.

The results in Section 2 are given for mean imputation. Extensions to some other imputation methods are straightforward. For example, if random hot deck imputation is considered, then our result leads to imputation within clusters (or G_l's). When there is a covariate x whose values are all observed, our result can be extended to regression imputation with model (3) modified to $y_{ij} = \alpha + \beta x_{ij} + b_i + e_{ij}$. For unit nonresponse, our result can also be applied to re-weighting, *i.e.*, adjusting weights within clusters (or G_l's).

Our method is imputation model based. We assume random-effect model (3) and random-effect-based response mechanism (4). If model (4) does not hold, then $E_m(\delta_{ij} \tilde{w}_{ij} e_{ij}) \neq 0$ and our estimator $\hat{Y}_c$ has a bias with a magnitude depending on the size of $|E_m(\delta_{ij} \tilde{w}_{ij} e_{ij})|$. Similarly, $\hat{Y}_c$ is not valid if model (3) does not hold.

It is shown in Section 2 that the condition $w_{ij} = w_i$ for all j ensures that imputation is done within each G_l that is the group of clusters with the same size and response rate. For two-stage sampling, this condition is satisfied when the last stage sampling is with equal probability (*e.g.*, simple random sampling without replacement). For three-stage sampling, model (3) should be replaced by $y_{ijk} = \mu_{ij} + b_{ij} + e_{ijk}$ and b_i in (4) should be replaced by b_{ij}. The survey weight w_{ijk} satisfies $w_{ijk} = w_{ij}$ as long as the last stage sampling is with equal probability and our result still holds. In two-stage sampling with w_{ij} varying with j, we may perform imputation within a group of clusters that have the same $E_m(\boldsymbol{y}_i \mid \boldsymbol{\delta}_i)$. For example, suppose that, in addition to (3)-(5), $\mu_i = \mu, b_i$'s are independent and identically distributed (iid), and conditional on b_i, the components of $\boldsymbol{\delta}_i$ are iid. Then $E_m(b_i \mid \boldsymbol{\delta}_i) = E_m(b_i \mid \bar{\delta}_i)$ depending only on the size of the cluster m_i and $\overline{\delta}_i$. Hence we can perform imputation within each G_l defined by (10).

Acknowledgments

This work was partially supported by the NCI Grant CA53786 and NSF Grant DMS-0404535. The author would like to thank Mr. Lei Xu for programming in the simulation study and two referees for their helpful comments.

References

Lee, H., Rancourt, E., and Särndal, C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.

Lohr, S.L. (1999). *Sampling*: *Design and Analysis*. Duxbury Press, New York.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Shao, J., and Steel, P. (1999). Variance estimation for imputed survey data with non-negligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

Wu, M.C., and Carroll, R.J. (1988). Estimation and comparisons of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175-188.