Robust inference in two-phase sampling designs with application to unit nonresponse

Jean-François Beaumont[1], Cyril Favre Martinoz[2] and David Haziza[3,4]
[1] Statistics Canada, Ottawa, Canada.
[2] Laboratoire de Statistique d'Enquête, Crest-Ensai, Bruz, France
[3] Département de mathématiques et statistique, Université de Montréal, Montreal, Canada
[4] Corresponding author: David Haziza, e-mail: David.Haziza@umontreal.ca

Abstract

Influential units occur frequently in surveys, especially in the context of business surveys that collect economic variables whose distributions are highly skewed. A unit is said to be influential when its inclusion or exclusion from the sample has an important impact on the magnitude of survey statistics. We extend the results of Beaumont et al. (2013) to the case of two-phase sampling designs. We define the concept of conditional bias attached to a unit with respect to both phases and propose a robust version of the double expansion estimator, which depends on a tuning constant. Following Beaumont et al. (2013), we determine the tuning constant which minimizes the maximum estimated conditional bias. Our results can be naturally extended to the case of unit nonresponse, the set of respondents often being viewed as a second phase sample.

Key words: Conditional bias; influential value; two-pase sampling design; robust estimation; unit nonresponse.

1 Introduction

Two-phase sampling is often used in surveys when the sampling frame contains little or no auxiliary information. In this case, it may be wise to first select a large sample in order to collect data on variables that are inexpensive to obtain and that are related to the characteristics of interest. Using the variables observed in the first phase, an efficient sampling procedure can then be used to select a (typically small) subsample from the first-phase sample in order to collect the characteristics of interest. The theory behind inference for two-phase sampling design may also be helpful in the context of unit nonresponse since the set of respondents is often viewed as a second phase sample.

Influential units occur frequently in surveys, especially in the context of business surveys that collect economic variables whose distributions are highly skewed. The presence of influential units in the sample does not introduce a bias but lead generally to very unstable estimators. Methods for dealing with influential units include winsorization and M-estimation; see e.g., Beaumont and Rivest (2009) and Beaumont et al. (2013).

In this paper, we extend the results of Beaumont et al (2013) for uni-phase sampling designs, who suggested constructing robust estimators of population totals based on the concept of conditional bias of a unit; see also Moreno-Rebollo et al. (1999, 2002). The conditional bias of a unit can be viewed as an appropriate measure of influence in finite population sampling.

2 Set-up

Consider a population U of size N. We are interested in estimating the population total $Y = \sum_{i \in U} y_i$ of a characteristic of interest y. We select a sample according to a two-phase sampling design: in the first phase, a sample S_1, of size n_1, is selected from U according to a given sampling design $p(S_1)$. In the second phase, a sample, S_2, of size n_2, is selected from S_1 according to $p(S_2|S_1)$. We develop our results for invariant two-phase sampling design, which are those designs that satisfy $p(S_2|S_1) = p(S_2)$. Our results can be extended to cover non-invariant two-phase designs.

We adopt the following notation: let I_{1i} be a sample selection indicator attached to unit i such that $I_{1i} = 1$ if unit i is selected in S_1 and $I_{1i} = 0$, otherwise, and let $\mathbf{I}_1 = (I_1, \cdots, I_N)'$. Let I_{2i} be a sample selection indicator attached to unit i such that $I_{2i} = 1$ if unit i is selected in S_2 and $I_{2i} = 0$, otherwise. Let $\pi_{1i} = P(I_{1i} = 1)$ and $\pi_{1ij} = P(I_{1i} = 1, I_{1j} = 1)$ denote the first-order and second-order probabilities in S_1. Similarly, let $\pi_{2i} = P(I_{2i} = 1\,|I_{1i} = 1\,)$ and $\pi_{2ij} = P(I_{2i} = 1, I_{2j} = 1 \left| I_{1i} = 1, I_{1j} = 1\,\right)$ denote the first-order and second-order probabilities in S_2.

A basic estimator of Y is the double expansion estimator

$$\hat{Y}_{DE} = \sum_{i \in S_2} \pi_{1i}^{-1} \pi_{2i}^{-1} y_i. \tag{1}$$

To study the properties of (1) , we express its total error as :

$$\hat{Y}_{DE} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{DE} - \hat{Y}_E), \tag{2}$$

where $\hat{Y}_E = \sum_{i \in S_1} \pi_{1i}^{-1} y_i$ denotes the expansion estimator that one would have used had the design been a single phase design. The terms $\hat{Y}_E - Y$ and $\hat{Y}_{DE} - \hat{Y}_E$ on the right hand side of (2) denote the errors due to the first phase and second phase, respectively. Let $E_1(.)$ and $V_1(.)$ denote the expectation and variance with respect to the first phase and $E_2(.\,|\mathbf{I}_1\,)$ and $V_2(.\,|\mathbf{I}_1\,)$ denote the conditional expectation and conditional variance with respect to the second phase. Noting that $E_2(\hat{Y}_{DE}\,|\mathbf{I}_1\,) = \hat{Y}_E$ and $E_1(\hat{Y}_E) = Y$, it follows from (2) that $E_p(\hat{Y}_{DE}) \equiv E_1 E_2(\hat{Y}_{DE}\ \ |\mathbf{I}_1\,) = Y$; that is, $\hat{Y}_{DE}$ is design-unbiased for Y. The total variance of $\hat{Y}_{DE}$ is

$$V_p(\hat{Y}_{DE}) = V_1 E_2(\hat{Y}_{DE}\,|\mathbf{I}_1\,) + E_1 V_2(\hat{Y}_{DE}\,|\mathbf{I}_1\,) = \sum_{i \in U} \sum_{j \in U} \left(\frac{\pi_{ij}^*}{\pi_i^* \pi_j^*} - 1 \right) y_i y_j, \tag{3}$$

where $\pi_i^* = \pi_{1i}\pi_{2i}$ and $\pi_{ij}^* = \pi_{1ij}\pi_{2ij}$.

In the presence of influential units, the estimator (1) remains design-unbiased. However, its design variance may be very large. In other words, including or excluding an influential unit from the calculations may have an important impact on the magnitude of the total error, $\hat{Y}_{DE} - Y$. An influential unit may have a large impact on the first phase error, $\hat{Y}_E - Y$, and/or on the second-phase error, $\hat{Y}_{DE} - \hat{Y}_E$.

3 Measuring the influence: the conditional bias

For uni-phase sampling designs, Moreno-Rebollo et al. (1999, 2002) introduced the concept of conditional bias attached to a unit as a measure of influence; see also Beaumont et al. (2013). We extend this concept to the case of two-phase sampling designs. We distinguish between three types of units: (i) the sample units, i.e., the units for which $I_{1i} = 1$ and $I_{2i} = 1$; (ii) the units selected in the first-phase sample but not in the second phase, i.e., the units for which $I_{1i} = 1$ and $I_{2i} = 0$ and (iii) the non-selected units, i.e., the units for which $I_{1i} = 0$ and

$I_{2i}=0$. It is worth noting that each type of unit may have an influence on the total error. However, only the influence of the sample units can be reduced at the estimation stage. In other words, nothing can be done for (ii) and (iii) at this stage.

The conditional bias of sampled unit i is defined as :

$$\begin{aligned} B_i^{DE}(I_{1i}=1,I_{2i}=1) &= E_1E_2(\hat{Y}_{DE}-Y|\mathbf{I}_1,I_{1i}=1,I_{2i}=1) \\ &= E_1(\hat{Y}_E-Y|I_{1i}=1)+E_1E_2(\hat{Y}_{DE}-\hat{Y}_E|\mathbf{I}_1,I_{1i}=1,I_{2i}=1). \end{aligned}$$

For an arbitrary two-phase design, we obtain

$$\begin{aligned} B_i^{DE}(I_{1i}=1,I_{2i}=1) &= \sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1\right)y_j+\sum_{j\in U}\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}\left(\frac{\pi_{2ij}}{\pi_{2i}\pi_{2j}}-1\right)y_j \\ &= \sum_{j\in U}\left(\frac{\pi_{ij}^*}{\pi_i^*\pi_j^*}-1\right)y_j. \end{aligned} \tag{4}$$

Example 1 *For simple random sampling without replacement in both phases, (4) reduces to* $B_i^{DE}(I_{1i}=1,I_{2i}=1)=\frac{N}{(N-1)}\left(\frac{N}{n_2}-1\right)(y_i-\overline{Y})$, *where* $\overline{Y}=Y/N$. *The previous expression suggest that a unit has a large influence if its y-value is far from the population mean* $\overline{Y}$.

Example 2 *For Poisson sampling in both phases, (4) reduces to* $B_i^{DE}(I_{1i}=1,I_{2i}=1)=\left(\pi_i^{*-1}-1\right)y_i$. *Hence, a unit has a large influence if its "total weight"* π_i^{*-1} *is large and/or if its y-value is large.*

Example 3 *For an arbitrary design in the first phase and Poisson sampling in the second phase, (4) reduces to*

$$B_i^{DE}(I_{1i}=1,I_{2i}=1)=\sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1\right)y_j+\pi_{1i}^{-1}\left(\pi_{2i}^{-1}-1\right)y_i. \tag{5}$$

Expression (5) will be particularly useful in the context of unit nonresponse.

In general, the conditional bias (4) is unknown as it depends on population quantities. An estimator of $B_i^{DE}(I_{1i}=1,I_{2i}=1)$ is given by

$$\hat{B}_i^{DE}(I_{1i}=1,I_{2i}=1)=\sum_{j\in S_2}\frac{\pi_{1i}}{\pi_{1ij}}\frac{\pi_{2i}}{\pi_{2ij}}\left(\frac{\pi_{ij}^*}{\pi_i^*\pi_j^*}-1\right)y_j. \tag{6}$$

The estimator $\hat{B}_i^{DE}(I_{1i}=1,I_{2i}=1)$ is conditionally unbiased for $B_i^{DE}(I_{1i}=1,I_{2i}=1)$ in the sense that $E_1E_2\left\{\hat{B}_i^{DE}(I_{1i}=1,I_{2i}=1)|\mathbf{I}_1,I_{1i}=1,I_{2i}=1\right\}=B_i^{DE}(I_{1i}=1,I_{2i}=1)$.

4 Robustifying the double expansion estimator

Following Beaumont et al. (2013), we consider the robust version of $\hat{Y}_{DE}$

$$\hat{Y}_{DE}^R=\hat{Y}_{DE}-\sum_{i\in S_2}\left\{\hat{B}_i^{DE}(I_{1i}=1,I_{2i}=1)\right\}+\sum_{i\in S_2}\psi\left\{\hat{B}_i^{DE}(I_{1i}=1,I_{2i}=1);c\right\}, \tag{7}$$

where $\psi(.)$ is a function, which role consists of curbing the impact of influential units and c is a tuning constant whose value must be determined. We use the so-called Huber function given by $\psi(z;c) = \text{sign}(z) \times \min(|z|, c)$, where c is a positive tuning constant and $\text{sign}(z) = 1$, for $z \geq 0$, while $\text{sign}(z) = -1$, otherwise.

When $\pi_{2i} = 1$ for all $i \in S_2$ (i.e., the case of a single phase sampling design), the robust estimator (7) reduces to that proposed by Beaumont et al. (2013). A suitable value for c is sometimes determined by minimizing an estimator of the mean square error of the robust estimator (e.g., Kokic and Bell, 1994; and Rivest and Hurtubise, 1995). Following Beaumont et al. (2013), we consider an alternative method which consists of finding the value of c that minimizes $\max\{\hat{B}_i^{RDE}(I_{1i} = 1, I_{2i} = 1) : i \in S_2\}$, where $\hat{B}_i^{RDE}(I_{1i} = 1, I_{2i} = 1)$ is an estimator of the conditional bias of the robust double expansion estimator attached to unit i. Using (4), we obtain

$$\begin{aligned} B_i^{RDE}(I_{1i} = 1, I_{2i} = 1) &= E_1E_2(\hat{Y}_{DE}^R - Y|\mathbf{I}_1, I_{1i} = 1, I_{2i} = 1) \\ &= B_i^{DE}(I_{1i} = 1, I_{2i} = 1) + E_1E_2\left\{\Delta(c)|\mathbf{I}_1, I_{1i} = 1, I_{2i} = 1\right\}, \end{aligned}$$

where

$$\Delta(c) = \sum_{i \in S_2} \left[\psi\left\{\hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1); c\right\} - \hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1)\right].$$

As for $B_i^{DE}(I_{1i} = 1, I_{2i} = 1)$, the conditional bias $B_i^{RDE}(I_{1i} = 1, I_{2i} = 1)$ is generally unknown. We estimate it by

$$\hat{B}_i^{RDE}(I_{1i} = 1, I_{2i} = 1) = \hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) + \Delta(c),$$

which is conditionally unbiased for $B_i^{RDE}(I_{1i} = 1, I_{2i} = 1)$; i.e.,

$$E_1E_2\left\{\hat{B}_i^{RDE}(I_{1i} = 1, I_{2i} = 1)|\mathbf{I}_1, I_{1i} = 1, I_{2i} = 1\right\} = B_i^{RDE}(I_{1i} = 1, I_{2i} = 1).$$

Let $\hat{B}_{min}^{DE} = \min\left\{\hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) : i \in S_2\right\}$ and $\hat{B}_{max}^{DE} = \max\left\{\hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) : i \in S_2\right\}$. The value of $\Delta(c)$ that minimizes $\max\{\hat{B}_i^{RDE}(I_{1i} = 1, I_{2i} = 1) : i \in S_2\}$, denoted by $\Delta(c_{minmax})$, is given by

$$\Delta(c_{minmax}) = -\frac{1}{2}(\hat{B}_{min}^{DE} + \hat{B}_{max}^{DE}).$$

Noting that $\hat{Y}_{DE}^R(c) = \hat{Y}_{DE} + \Delta(c)$, the resulting robust estimator is given by

$$\hat{Y}_{DE}^R(c_{minmax}) = \hat{Y}_{DE} - \frac{1}{2}(\hat{B}_{min}^{DE} + \hat{B}_{max}^{DE}).$$

This estimator can be obtained without actually computing the value c_{minmax} so that no iterative process is required.

5 Application to unit nonresponse

In this section, we consider the problem of robust estimation in the context of unit nonresponse. In this context, S_1 denotes the sample selected from the population, whereas S_2 denotes the random set of respondents. The quantities I_{1i} and I_{2i} denote respectively the sample selection indicator and the response indicator attached to unit i. Also, π_{1i} and π_{2i} denote respectively the inclusion probability in the sample and the response probability for unit i. We assume that the units respond independently of one another; that is $\pi_{2ij} = \pi_{2i}\pi_{2j}$ for $i \neq j$. This is similar to Poisson sampling described in Example 3, except that the π_{2i}'s are now unknown. If the π_{2i}'s were known, a propensity score adjusted (PSA) estimator would be given by (1) and the conditional bias of a responding unit would be given (5). In practice, the response probabilities π_{2i} are unknown and must be estimated. We assume that they can be parametrically modeled by

$$\pi_{2i} = m(\mathbf{x}_i,\ \alpha), \tag{8}$$

where $m(.)$ is a known function, $\mathbf{x}$ is a vector of auxiliary variables available for all the sampled units (respondents and nonrespondents) and α is a vector of unknown parameters. A special case of (8) is the logistic regression model. Based on (8), an estimator of π_{2i} is given by $\hat{\pi}_{2i} = m(\mathbf{x}_i,\ \hat{\alpha})$, where $\hat{\alpha}$ denotes a suitable estimator of α (e.g., the maximum likelihood estimator). A PSA estimator of Y is thus given by

$$\hat{Y}_{PSA} = \sum_{i \in S_2} \frac{1}{\pi_{1i}\hat{\pi}_{2i}} y_i. \tag{9}$$

The total error of $\hat{Y}_{PSA}$ can be expressed as

$$\hat{Y}_{PSA} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{PSA} - \hat{Y}_E). \tag{10}$$

The terms $\hat{Y}_E - Y$ and $\hat{Y}_{PSA} - \hat{Y}_E$ in (10) denote the sampling error and the nonresponse error, respectively. Using a first-order Taylor expansion (Kim and Kim, 2007), we have

$$\hat{Y}_{PSA} - \hat{Y}_L = O_p\left(\frac{N}{n}\right), \tag{11}$$

where

$$\hat{Y}_L = \sum_{i \in S_1} \pi_{1i}^{-1} \left\{ k_i \pi_{1i} \pi_{2i} \mathbf{h}_i' \hat{\gamma} + \frac{I_{2i}}{\pi_{2i}} \left(y_i - k_i \pi_{1i} \pi_{2i} \mathbf{h}_i' \hat{\gamma} \right) \right\}$$

with $\mathbf{h}_i = \partial \left\{ logit(\pi_{2i}) \right\} / \partial \alpha$, $\hat{\gamma} = \left\{ \sum_{i \in S_1} k_i \pi_{2i} (1 - \pi_{2i}) \mathbf{h}_i \mathbf{h}_i' \right\}^{-1} \sum_{i \in S_1} \pi_{1i}^{-1} (1 - \pi_{2i}) \mathbf{h}_i y_i$ and k_i is a weight associated with unit i used in the estimation of α. Typically, $k_i = 1$ or $k_i = \pi_{1i}^{-1}$. Using (11) in (10), we obtain

$$\hat{Y}_{PSA} - Y = (\hat{Y}_E - Y) + (\hat{Y}_L - \hat{Y}_E) + O_p\left(\frac{N}{n}\right). \tag{12}$$

Ignoring the higher order terms in (12), the conditional bias of the PSA estimator attached to responding unit i can be approximated by

$$B_i^{PSA}(I_{1i} = 1, I_{2i} = 1) = E_1 E_2 (\hat{Y}_{PSA} - Y | \mathbf{I}_1, I_{1i} = 1, I_{2i} = 1).$$

After some tedious but relatively straightforward algebra, we obtain

$$\begin{aligned} B_i^{PSA}(I_{1i} = 1, I_{2i} = 1) \ &\doteq\ \sum_{j \in U} \left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_j - \pi_{1i}^{-1} (\pi_{2i}^{-1} - 1) \left(y_i - \mathbf{c}_i' \gamma \right) \\ &-\ \mathbf{c}_i' \mathbf{T}^{-1} \sum_{j \in U} \left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) (1 - \pi_{2j}) \left(y_j - \mathbf{c}_j' \gamma \right) \mathbf{h}_j, \end{aligned} \tag{13}$$

where $\mathbf{c}_i = k_i \pi_{1i} \pi_{2i} \mathbf{h}_i$ and $\gamma = \mathbf{T}^{-1} \sum_{i \in U} (1 - \pi_{2i}) \mathbf{h}_i y_i$ with $\mathbf{T} = \sum_{i \in U} k_i \pi_{1i} \pi_{2i} (1 - \pi_{2i}) \mathbf{h}_i \mathbf{h}_i'$. A robust version of $\hat{Y}_{PSA}$ is given by

$$\hat{Y}_{PSA}^R = \hat{Y}_{PSA} - \sum_{i \in S_2} \hat{B}_i^{PSA}(I_{1i} = 1, I_{2i} = 1) + \sum_{i \in S_2} \psi \left\{ \hat{B}_i^{PSA}(I_{1i} = 1, I_{2i} = 1); c \right\},$$

where $\hat{B}_i^{PSA}(I_{1i}=1,I_{2i}=1)$ is a suitable estimator of $B_i^{PSA}(I_{1i}=1,I_{2i}=1)$. Once again, we determine the value of c that minimizes $\max\{\hat{B}_i^{RPSA}(I_{1i}=1,I_{2i}=1) : i \in S_2\}$, where $\hat{B}_i^{RPSA}(I_{1i}=1,I_{2i}=1)$ is an estimator of the conditional bias of the robust PSA estimator attached to unit i.

In practice, it is customary to partition the population into weighting adjustment cells, $U_1,\ldots,U_G$. The response probability attached to unit i in cell g is estimated by realized response rate within the associated cell; that is,

$$\hat{\pi}_{2i} = \hat{\pi}_{2g} = \frac{\sum_{i\in S_2\cap U_g} \pi_{1i}^{-1}}{\sum_{i\in S_1\cap U_g} \pi_{1i}^{-1}}, \qquad \text{for } i \in U_g. \tag{14}$$

Assuming that nonresponse is uniform within cells, i.e., $\pi_{2i} = \pi_{2g}$ for $i \in U_g$, the PSA estimator (9) is an asymptotically unbiased estimator of the population total Y.

Note that the estimated response probabilities given by (14) can alternatively be obtained by fitting the parametric model (8) with $\mathbf{x}_i = (\delta_{1i},\ldots,\delta_{Gi})'$, where δ_{gi} is a class indicator such that $\delta_{gi}=1$ if unit $i \in U_g$ and $\delta_{gi}=0$, otherwise. Therefore, the conditional bias of the PSA estimator based on G weighting cells attached to unit i can be obtained as a special case of (13), which leads to

$$B_i^{PSA}(I_{1i}=1,I_{2i}=1) \doteq \sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1\right) y_j + \pi_{1i}^{-1}\left(\pi_{2g}^{-1} - 1\right)(y_i - \bar{Y}_g) \qquad \text{for } i \in U_g,$$

where $\bar{Y}_g = \sum_{i\in U_g} y_i / N_g$ with N_g denoting the size of U_g.

References

[1] Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. To appear in Biometrika.

[2] Beaumont, J.-F. and Rivest, L.-P. (2009). Dealing with outliers with survey data. Handbook of Statistics, Volume 29, Chapter 11, Sample Surveys: Theory Methods and Inference, Editors: C.R. Rao and D. Pfeffermann, 247–279.

[3] Kim, J.K. and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. The Canadian Journal of Statistics, 35, 501–514.

[4] Kokic, P.N., and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. Journal of Official Statistics, 10, 419–435.

[5] Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: conditional bias. Biometrika, 86, 923–968.

[6] Moreno-Rebollo, J.L., Muñoz-Reyez, A.M., Jimenez-Gamero, M.D. and Muñoz-Pichardo, J. (2002). Influence diagnostics in survey sampling: estimating the conditional bias. Metrika, 55, 209–214.

[7] Rivest, L.-P. and Hurtubise, D. (1995). On Searls Winsorized means for skewed populations. Survey Methodology, 21, 119–129.