Variance estimation under monotone non-response for a panel survey

Hélène Juillard, Guillaume Chauvet

Variance estimation under monotone non-response for a panel survey

Hélène Juillard*
Guillaume Chauvet†

December 18, 2016

Abstract

Panel surveys are frequently used to measure the evolution of parameters over time. Panel samples may suffer from different types of unit non-response, which is currently handled by estimating the response probabilites and by reweighting respondents. In this work, we consider estimation and variance estimation under unit non-response for panel surveys. Extending the work by Kim and Kim (2007) for several times, we consider an expansion estimator accounting for initial non-response and attrition, and propose a suitable variance estimator. It is then extended to cover most estimators encountered in surveys, including calibrated estimators, complex parameters and longitudinal estimators. The properties of the proposed variance estimator and of a simplified variance estimator are estimated through a simulation study. An illustration of the proposed methods on data from the ELFE survey is also presented.

1 Introduction

Surveys are not only used to produce estimators for one point in time (cross-sectional estimations), but also to measure the evolution of parameters (longitudinal estimations), and are thus repeated over time. Kalton (2009) distinguishes three broad families of sampling designs for such surveys: the repeated cross-sectional surveys, in which estimations are produced through samples selected independently at each time; the panel surveys, in which

*INED, 133 boulevard Davout, 75020 Paris, France
†ENSAI/IRMAR, Campus de Ker Lann, 35170 Bruz, France

measures are repeated over time for units in a same sample; the rotating panel surveys, which correspond to panel surveys with a sub-sample of units being replaced at each time by another incoming sub-sample. In this paper, we are interested in estimation and variance estimation for panel surveys.

Among the panel surveys (a.k.a. longitudinal surveys, see Lynn, 2009), cohort surveys are particular cases where the units in the sample are linked by a common original event, such as being born on the same year for children in the ELFE survey (Enquête Longitudinale Française depuis l'Enfance), which is the motivating example for this work. ELFE is the first longitudinal study of its kind in France, tracking children from birth to adulthood (Pirus et al., 2010). Covering the whole metropolitan France, it was launched in 2011 and consists of more than 18,000 children whose parents consented to their inclusion. It will examine every aspect of these children's lives from the perspectives of health, social sciences and environmental health. The ELFE survey suffers from unit non-response, which needs to be accounted for by using available auxiliary information, so as to limit the bias of estimators. Though the ELFE survey will be used for illustration in this paper, non-response occurs in virtually any panel survey so that the proposed methods are of general interest; see for example Laurie et al. (1999) for the treatment of non-response of the British Household Panel Survey, or Vandecasteele and Debels (2007) for the European Community Household Panel.

Non-response is currently handled by modeling the response probabilities (Kim and Kim, 2007) and by reweighting respondents with the inverse of these estimated probabilities. A panel sample may suffer from three types of unit non-response (Hawkes and Plewis, 2009): initial non-response refers to the original absence of selected units; wave non-response occurs when some units in the panel sample temporarily do not answer at some point in time, while attrition occurs when some units in the panel sample permanently do not answer from some point in time. Wave non-response was fairly uncommon in the first waves of the ELFE survey which were at our disposal. We therefore simplify this set-up by assuming monotone non-response, where only initial non-response and attrition occur.

There is a vast literature on the treatment of unit non-response for surveys over time, see Ekholm and Laaksonen (1991), Fuller et al. (1994), Rizzo et al. (1996), Clarke and Tate (2002), Laaksonen and Chambers (2006), Hawkes and Plewis (2009), Rendtel and Harms (2009), Laaksonen (2007), Slud and Bailey (2010), Zhou and Kim (2012). Variance estimation for longitudinal estimators is considered in Tam (1984), Laniel (1988), Nordberg (2000), Berger

(2004), Skinner and Vieira (2005), Qualité and Tillé (2008) and Chauvet and Goga (2016), but with focus on the sampling variance only. Variance estimation in case of non-response weighting adjustments on cross-sectional surveys is considered in Kim and Kim (2007). To the best of our knowledge, and despite the interest for applications, variance estimation accounting for non-response for panel surveys has not been treated in the literature, with the exception of Zhou and Kim (2012).

The paper is organized as follows. Basic notations are given in Section 2. In Section 3, we define the expansion estimator for a total and a corresponding variance estimator when the response probabilities are assumed to be known at each time. Though this case appears unrealistic in most applications, it is common practice in some surveys that the response probabilities are assumed to be known without error to simplify variance estimation. Consequently, the simplified set-up in Section 3 enables to propose a first simplified variance estimator. In Section 4, we consider the usual case when the response probabilities are unknown. A parametric model is postulated leading to estimated response probabilities and to a reweighted estimator, and a variance estimator is derived by following the approach in Kim and Kim (2007). Some illustrations on the particular important case of the response homogeneity groups are also given. The proposed variance estimator is extended to cover calibrated estimators and complex parameters in Section 5. Longitudinal estimation is discussed in Section 6, and the proposed variance estimator is used to cover such cases. The variance estimators are compared in Section 7 through a simulation study, and an illustration on the ELFE data is proposed in Section 8. We draw some conclusions in Section 9. Some technical conditions useful to derive the properties of some estimators that we consider and the proofs are given in Appendix.

2 Notation

We are interested in a finite population U. A sample s_0 is first selected according to some sampling design $p(\cdot)$, and we assume that the first-order inclusion probabilities π_i are strictly positive for any $i \in U$. This first sampling phase corresponds to the original inclusion of units in the sample. For example, the 2011 ELFE survey involved the selection of a sample of babies according to a cross-classified sampling design (CCS), where a sample of maternity units and a sample of days were selected independently, the survey being performed in the maternity units selected on the days selected (Juillard et al., 2016). Also, we note π_{ij} for the probability that units i and j are

selected jointly in the sample, and $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$.

We consider the case of a panel survey in which the sole units in the original sample s_0 are followed over time, without reentry or late entry units at subsequent times to represent possible newborns. We are therefore interested in estimating some parameter defined over the population U, for some study variable y taking the value y_i for the unit i. The units in the sample s_0 are then followed at subsequent times $\delta = 1, \ldots, t$, and the sample is prone to unit non-response at each time. We note r_i^δ for the response indicator for unit i at time δ, and s_δ for the subset of respondents at time δ.

We assume monotone non-response resulting in the nested sequence

$$s_0 \supset s_1 \supset ... \supset s_t. \tag{2.1}$$

For $\delta = 1, \ldots, t$, we note

$$p_i^\delta \quad = \quad Pr\left(i \in s_\delta | s_{\delta-1}\right) \tag{2.2}$$

for the response probability of some unit i to be a respondent at time δ. We assume that the non-response mechanisms are ignorable, in the sense that the response probability p_i^δ at time δ can be explained by the variables observed at times $0, \ldots, \delta - 1$. Also, we assume that at any time δ the units answer independently of one another, and we note

$$p_{ij}^\delta \quad = \quad Pr\left(i, j \in s_\delta | s_{\delta-1}\right) = p_i^\delta p_j^\delta \tag{2.3}$$

for the probability that two distinct units i and j answer jointly at time δ.

3 Estimation with known response probabilities

3.1 Expansion estimator

We are interested in estimating the total $Y = \sum_{i \in U} y_i$. In a situation of full response, the Horvitz-Thompson estimator

$$\tilde{Y}_0 \quad = \quad \sum_{i \in s_0} \frac{y_i}{\pi_i} \tag{3.1}$$

is design-unbiased for Y. In the situation of unit non-response, the subsample s_t only is observed at time t. Assuming that the response probabilities at each time are known, the expansion estimator at time t is

$$\tilde{Y}_t = \sum_{i \in s_t} \frac{y_i}{\pi_i p_i^{1 \to t}} \quad \text{with} \quad p_i^{1 \to t} = \prod_{\delta=1}^{t} p_i^{\delta}. \tag{3.2}$$

Under some conditions on the variable of interest, the sampling design and the response mechanisms, the expansion estimator $\tilde{Y}_t$ is design-unbiased and consistent for Y, see Appendix B. Here and elsewhere, the subscript δ will be used when the sample observed at time δ is used for estimation. The superscript δ will be used when we account for non-response at time δ, like for the probability p_i^{δ} of unit i to be a respondent at time δ.

3.2 Variance computation

At time t, we have

$$\begin{aligned} V(\tilde{Y}_t) &= VE(\tilde{Y}_t|s_{t-1}) + EV(\tilde{Y}_t|s_{t-1}) & (3.3) \\ &= V(\tilde{Y}_{t-1}) + EV(\tilde{Y}_t|s_{t-1}). & (3.4) \end{aligned}$$

Using a proof by induction, we obtain

$$V(\tilde{Y}_t) = V(\tilde{Y}_0) + E\left\{\sum_{\delta=1}^{t} V(\tilde{Y}_\delta|s_{\delta-1})\right\}. \tag{3.5}$$

The first term in the right-hand side of (3.5) is the variance due to the sampling design, that we note as $V^p(\tilde{Y}_t)$, and that may be rewritten as

$$V^p(\tilde{Y}_t) = \sum_{i,j \in U} \Delta_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{3.6}$$

The second term in the right-hand side of (3.5) is the variance due to non-response, that we note as $V^{nr}(\tilde{Y}_t)$ and that may be rewritten as

$$V^{nr}(\tilde{Y}_t) = E\left\{\sum_{\delta=1}^{t} V^{nr\delta}(\tilde{Y}_t)\right\} \tag{3.7}$$

with

$$V^{nr\delta}(\tilde{Y}_t) = \sum_{i \in s_{\delta-1}} p_i^{\delta}(1 - p_i^{\delta}) \left(\frac{y_i}{\pi_i p_i^{1 \to \delta}}\right)^2. \tag{3.8}$$

3.3 Variance estimation

At time t, an estimator for the variance due to the sampling design $V^p(\tilde{Y}_t)$ is

$$\hat{V}_t^p(\tilde{Y}_t) = \sum_{i,j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{p_{ij}^{1 \to t}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}, \tag{3.9}$$

where $p_{ij}^{1 \to t} \equiv \prod_{\delta=1}^{t} p_{ij}^{\delta}$. This estimator is unbiased for $V^p(\tilde{Y}_t)$, provided that $\pi_{ij} > 0$ for any units $i \neq j \in U$. An unbiased estimator for the variance due to non-response $V^{nr}(\tilde{Y}_t)$ is

$$\hat{V}_t^{nr}(\tilde{Y}_t) = \sum_{\delta=1}^{t} \hat{V}_t^{nr\delta}(\tilde{Y}_t) \tag{3.10}$$

with

$$\hat{V}_t^{nr\delta}(\tilde{Y}_t) = \sum_{i \in s_t} \frac{p_i^{\delta}(1 - p_i^{\delta})}{p_i^{\delta \to t}} \left(\frac{y_i}{\pi_i p_i^{1 \to \delta}} \right)^2. \tag{3.11}$$

By using the writing

$$\hat{V}_t^{\mathrm{nr}\delta}(\tilde{Y}_t) = \sum_{i \in s_t} \left(\frac{y_i}{\pi_i} \right)^2 \times \frac{1}{p_i^{1 \to t}} \times \left(\frac{1}{p_i^{1 \to \delta}} - \frac{1}{p_i^{1 \to \delta - 1}} \right), \tag{3.12}$$

and by summing for $\delta = 1, \ldots, t$, the estimator for the variance due to non-response simplifies as

$$\hat{V}_t^{nr}(\tilde{Y}_t) = \sum_{i \in s_t} \frac{1 - p_i^{1 \to t}}{\left(p_i^{1 \to t}\right)^2} \left(\frac{y_i}{\pi_i} \right)^2. \tag{3.13}$$

This leads to the global variance estimator at time t

$$\hat{V}_t(\tilde{Y}_t) = \hat{V}_t^p(\tilde{Y}_t) + \hat{V}_t^{nr}(\tilde{Y}_t). \tag{3.14}$$

This variance estimator can be shown to be unbiased and consistent for $V(\tilde{Y}_t)$, see Appendix C.

3.4 Application to Response Homogeneity Groups

For the purpose of illustration, we consider the model of Response Homogeneity Groups (RHG) which is often used in practice. More precisely, we assume that at each time $\delta = 1, \ldots, t$, the sub-sample $s_{\delta - 1}$ may be partitioned into

$C(\delta-1)$ groups $s^c_{\delta-1},\ c=1,\ldots,C(\delta-1)$, such that the response probability p^δ_i is constant inside a group. In such case, we simplify the notation as

$$p^\delta_i = p^\delta_c \quad \text{for any} \quad i \in s^c_{\delta-1}. \tag{3.15}$$

Note that the number of groups, and the groups themselves, may vary over time.

If the expansion estimator is computed at time $t=1$, the estimator in (3.13) for the variance due to non-response may be rewritten as

$$\hat{V}^{nr}_1(\tilde{Y}_1) \;=\; \sum_{c=1}^{C(0)} \frac{1-p^1_c}{\left(p^1_c\right)^2} \sum_{i\in s_1\cap s^c_0} \left(\frac{y_i}{\pi_i}\right)^2, \tag{3.16}$$

and the global variance estimator at time 1 is

$$\hat{V}_1(\tilde{Y}_1) \;=\; \sum_{i,j\in s_1} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{p^1_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} + \sum_{c=1}^{C(0)} \frac{1-p^1_c}{\left(p^1_c\right)^2} \sum_{i\in s_1\cap s^c_0} \left(\frac{y_i}{\pi_i}\right)^2. \tag{3.17}$$

If the expansion estimator is computed at time $t=2$, the estimator in (3.13) for the variance due to non-response may be rewritten as

$$\hat{V}^{nr}_2(\tilde{Y}_2) \;=\; \sum_{c=1}^{C(0)} \sum_{d=1}^{C(1)} \frac{1-p^1_c p^2_d}{\left(p^1_c p^2_d\right)^2} \sum_{i\in s_2\cap s^d_1\cap s^c_0} \left(\frac{y_i}{\pi_i}\right)^2. \tag{3.18}$$

A simple case occurs when the same system of RHGs is kept over time. In this case, the number of groups at each time is equal to $C(0)$, and we obtain a nested sequence of sub-samples

$$s^c_0 \supset s^c_1 \supset \ldots \supset s^c_t \quad \text{for any} \quad c=1,\ldots,C(0). \tag{3.19}$$

The variance estimator in (3.18) simplifies as

$$\hat{V}^{nr}_2(\tilde{Y}_2) \;=\; \sum_{c=1}^{C(0)} \frac{1-p^{1\to 2}_c}{\left(p^{1\to 2}_c\right)^2} \sum_{i\in s_2\cap s^c_1} \left(\frac{y_i}{\pi_i}\right)^2, \tag{3.20}$$

with $p^{1\to 2}_c = \prod_{\delta=1}^2 p_c$ for $c=1,\ldots,C(0)$.

4 Estimation with unknown response probabilities

4.1 Reweighted estimator

In practice, the response probabilities at each time are unknown and need to be estimated. We assume that at each time δ the probability of response is parametrically modeled as

$$p_i^\delta = p^\delta(z_i^\delta, \alpha^\delta) \tag{4.1}$$

for some known function $p^\delta(\cdot,\cdot)$, where z_i^δ is a vector of auxiliary variables observed for all the units in the subsample $s_{\delta-1}$, and α^δ denotes some unknown parameter. Following the approach in Kim and Kim (2007), we assume that the true parameter is estimated by $\hat{\alpha}^\delta$, which is the solution of the estimating equation

$$\frac{\partial}{\partial\alpha}\sum_{i\in s_{\delta-1}} k_i^\delta \left\{r_i^\delta \ln(p_i^\delta) + (1-r_i^\delta)\ln(1-p_i^\delta)\right\} = 0, \tag{4.2}$$

with k_i^δ some weight of unit i in the estimating equation. Customary choices for these weights include $k_i^\delta = 1$ and $k_i^\delta = \pi_i^{-1}$, see Fuller and An (1998), Beaumont (2005) and Kim and Kim (2007).

The estimated response probability at time δ is

$$\hat{p}_i^\delta = p^\delta(z_i^\delta, \hat{\alpha}^\delta). \tag{4.3}$$

The reweighted estimator at time t is

$$\hat{Y}_t = \sum_{i\in s_t} \frac{y_i}{\pi_i \hat{p}_i^{1\to t}} \quad \text{with} \quad \hat{p}_i^{1\to t} = \prod_{\delta=1}^{t} \hat{p}_i^\delta. \tag{4.4}$$

It is obtained by substituting in (3.2) each unknown response probability p_i^δ with its estimator in (4.3).

4.2 Variance computation

Under some regularity assumptions on the response mechanisms and some regularity conditions on the $p^\delta(\cdot,\cdot)$'s, we obtain from Theorem 1 in Kim and Kim (2007) that we can write

$$\hat{Y}_t = \hat{Y}_{lin,t} + O_p(Nn^{-1}), \tag{4.5}$$

where

$$\hat{Y}_{lin,t} = \sum_{i \in s_{t-1}} \frac{1}{\pi_i \hat{p}_i^{1 \to t-1}} \left\{ k_i^t \pi_i \hat{p}_i^{1 \to t-1} p_i^t (h_i^t)^\top \gamma^t + \frac{r_i^t}{p_i^t} (y_i - k_i^t \pi_i \hat{p}_i^{1 \to t-1} p_i^t (h_i^t)^\top \gamma^t) \right\}, \quad (4.6)$$

and where for any $\delta = 1, \ldots, t$ we denote by h_i^δ the value of $h_i^\delta(\alpha) = \partial \, \text{logit}(p_i^\delta)/\partial\alpha$ evaluated at $\alpha = \alpha^\delta$, and

$$\gamma^\delta = \left\{ \sum_{i \in s_{\delta-1}} k_i^\delta p_i^\delta (1 - p_i^\delta) h_i^\delta (h_i^\delta)^\top \right\}^{-1} \sum_{i \in s_{\delta-1}} \frac{1 - p_i^\delta}{\hat{p}_i^{1 \to \delta-1}} h_i^\delta \frac{y_i}{\pi_i}. \quad (4.7)$$

From (4.6), we obtain that

$$E\left(\hat{Y}_{lin,t} \,|s_{t-1}\right) = \hat{Y}_{t-1}. \quad (4.8)$$

Using a proof by induction, it follows from (4.5) and (4.8) that $\hat{Y}_t$ is approximately unbiased for Y. Also, the variance of $\hat{Y}_t$ may be asymptotically approximated by

$$V_{app}(\hat{Y}_t) = V(\tilde{Y}_0) + E\left\{ \sum_{\delta=1}^{t} V(\hat{Y}_{lin,\delta} | s_{\delta-1}) \right\}. \quad (4.9)$$

The first term in the right-hand side of (4.9) is the variance due to the sampling design, that we note as $V^p(\hat{Y}_t)$. It is identical to the variance due to the sampling design for the expansion estimator. The second term in the right-hand side of (4.9) is the variance due to non-response, that we note as $V^{nr}(\hat{Y}_t)$. From (4.6), this asymptotic variance is given by

$$V^{nr}(\hat{Y}_t) = E\left\{ \sum_{\delta=1}^{t} V^{nr\delta}(\hat{Y}_t) \right\}, \quad (4.10)$$

where

$$V^{nr\delta}(\hat{Y}_t) = \sum_{i \in s_{\delta-1}} p_i^\delta (1 - p_i^\delta) \left(\frac{y_i}{\pi_i \hat{p}_i^{1 \to \delta-1} p_i^\delta} - k_i^\delta (h_i^\delta)^\top \gamma^\delta \right)^2. \quad (4.11)$$

We now compare the variance due to non-response for the reweighted estimator with estimated response probabilities $\hat{Y}_t$, given in equation (4.10), and

the variance due to non-response for the expansion estimator with known response probabilities $\tilde{Y}_t$, given in equation (3.7). For each of its component $\delta = 1, \ldots, t$, the term $V^{nr\delta}(\hat{Y}_t)$ in (4.11) includes a centering term $k_i^\delta (h_i^\delta)^\top \gamma^\delta$, which is essentially a prediction of $(\pi_i \hat{p}_i^{1\to\delta-1} p_i^\delta)^{-1} y_i$ by means of regressors h_i^δ. This centering is due to the estimation of the response probabilities, and therefore does not appear in equation (3.7). It usually leads to a smaller variance than that of $\tilde{Y}_t$; see also Beaumont (2005), equation (5.7) and Kim and Kim (2007), equation (17), for the case $t = 1$.

4.3 Variance estimation

At time t, an approximately unbiased estimator for the variance due to the sampling design $V^p(\hat{Y}_t)$ is

$$\hat{V}_t^p(\hat{Y}_t) = \sum_{i,j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}, \tag{4.12}$$

where $\hat{p}_{ij}^{1\to t} \equiv \prod_{\delta=1}^t \hat{p}_{ij}^\delta$, and where

$$\hat{p}_{ij}^\delta = \begin{cases} \hat{p}_i^\delta & \text{if } i = j, \\ \hat{p}_i^\delta \hat{p}_j^\delta & \text{otherwise.} \end{cases} \tag{4.13}$$

Following equation (25) in Kim and Kim (2007), $V^{nr}(\hat{Y}_t)$ may be approximately unbiasedly estimated at time t by

$$\hat{V}_t^{nr}(\hat{Y}_t) = \sum_{\delta=1}^t \hat{V}_t^{nr\delta}(\hat{Y}_t) \tag{4.14}$$

where

$$\hat{V}_t^{nr\delta}(\hat{Y}_t) = \sum_{i\in s_t} \frac{\hat{p}_i^\delta (1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \left(\frac{y_i}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^\delta (\hat{h}_i^\delta)^\top \hat{\gamma}_t^\delta \right)^2, \tag{4.15}$$

$$\hat{h}_i^\delta = h(z_i, \hat{\alpha}^\delta), \tag{4.16}$$

$$\hat{\gamma}_t^\delta = \left\{ \sum_{i\in s_t} k_i^\delta \frac{\hat{p}_i^\delta (1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \hat{h}_i^\delta (\hat{h}_i^\delta)^\top \right\}^{-1} \sum_{i\in s_t} \frac{1-\hat{p}_i^\delta}{\hat{p}_i^{1\to t}} \hat{h}_i^\delta \frac{y_i}{\pi_i}. \tag{4.17}$$

This leads to the global variance estimator at time t

$$\hat{V}_t(\hat{Y}_t) = \hat{V}_t^p(\hat{Y}_t) + \hat{V}_t^{nr}(\hat{Y}_t). \tag{4.18}$$

A simplified estimator of the variance due to non-response is obtained by ignoring the prediction terms $k_i^\delta(\hat{h}_i^\delta)^\top\hat{\gamma}_t^\delta$ for each of the $\delta = 1,\ldots,t$ variance components. Mimicking the reasoning in Section 3.3, this leads to the simplified variance estimator

$$\begin{aligned}\hat{V}_{t,simp}^{nr}(\hat{Y}_t) &= \sum_{\delta=1}^{t}\sum_{i\in s_t}\frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}}\left(\frac{y_i}{\pi_i\hat{p}_i^{1\to\delta}}\right)^2 \\ &= \sum_{i\in s_t}\frac{1-\hat{p}_i^{1\to t}}{(\hat{p}_i^{1\to t})^2}\left(\frac{y_i}{\pi_i}\right)^2. \end{aligned} \tag{4.19}$$

This simplified variance estimator is computed as if in the reweighted estimator $\hat{Y}_t$, the response probabilities were known. It will tend to overestimate the variance due to non-response of $\hat{Y}_t$ if the prediction term $k_i^\delta(h_i^\delta)^\top\gamma^\delta$ partly explains $(\pi_i\hat{p}_i^{1\to\delta-1}p_i^\delta)^{-1}y_i$.

4.4 Application to the logistic regression model

In the particular case when a logistic regression model is used at each time δ, the model (4.1) may be rewritten as

$$\text{logit}(p_i^\delta) = (z_i^\delta)^\top\alpha^\delta. \tag{4.20}$$

We obtain $\hat{h}_i^\delta = z_i^\delta$, and the estimator for the variance due to non-response is given by (4.14), with

$$\hat{V}_t^{nr\delta}(\hat{Y}_t) = \sum_{i\in s_t}\frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}}\left(\frac{y_i}{\pi_i\hat{p}_i^{1\to\delta}} - k_i^\delta(z_i^\delta)^\top\hat{\gamma}_t^\delta\right)^2, \tag{4.21}$$

$$\hat{\gamma}_t^\delta = \left\{\sum_{i\in s_t}k_i^\delta\frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}}z_i^\delta(z_i^\delta)^\top\right\}^{-1}\sum_{i\in s_t}\frac{1-\hat{p}_i^\delta}{\hat{p}_i^{1\to t}}z_i^\delta\frac{y_i}{\pi_i}. \tag{4.22}$$

If the reweighted estimator is computed at time $t = 1$, the estimator in (4.14) for the variance due to non-response may be rewritten as

$$\begin{aligned}\hat{V}_1^{nr}(\hat{Y}_1) &= \hat{V}_1^{nr,1}(\hat{Y}_1) \\ &= \sum_{i\in s_1}(1-\hat{p}_i^1)\left(\frac{y_i}{\pi_i\hat{p}_i^1} - k_i^1(z_i^1)^\top\hat{\gamma}_1^1\right)^2. \end{aligned} \tag{4.23}$$

If the reweighted estimator is computed at time $t = 2$, the estimator in (4.14)

for the variance due to non-response may be rewritten as

$$\begin{aligned}\hat{V}_2^{nr}(\hat{Y}_2) &= \hat{V}_2^{nr,1}(\hat{Y}_2) + \hat{V}_2^{nr,2}(\hat{Y}_2) \\ &= \sum_{i\in s_2} \frac{(1-\hat{p}_i^1)}{\hat{p}_i^2}\left(\frac{y_i}{\pi_i \hat{p}_i^1} - k_i^1 (z_i^1)^\top \hat{\gamma}_2^1\right)^2 \\ &+ \sum_{i\in s_2} (1-\hat{p}_i^2)\left(\frac{y_i}{\pi_i \hat{p}_i^1 \hat{p}_i^2} - k_i^2 (z_i^2)^\top \hat{\gamma}_2^2\right)^2. \end{aligned} \tag{4.24}$$

4.5 Application to Response Homogeneity Groups

We consider the model of Response Homogeneity Groups introduced in Section 3.4. At each time $\delta = 1, \ldots, t$, the sub-sample $s_{\delta-1}$ is partitioned into $C(\delta-1)$ groups $s_{\delta-1}^c,\ c = 1, \ldots, C(\delta-1)$. The response probabilities are assumed to be constant within the groups.

This model is equivalent to the logistic regression model in (4.20), with

$$z_i^\delta = \left[1\left\{i \in s_{\delta-1}^1\right\}, \ldots, 1\left\{i \in s_{\delta-1}^{C(\delta-1)}\right\}\right]^\top. \tag{4.25}$$

Solving the estimating equation (4.2) leads to the estimated response probabilities

$$\hat{p}_i^\delta = \frac{\sum_{i\in s_{\delta-1}^c} k_i^\delta r_i^\delta}{\sum_{i\in s_{\delta-1}^c} k_i^\delta} \quad \text{for} \quad i \in s_{\delta-1}^c. \tag{4.26}$$

That is, the response probability is estimated by the weighted response rate inside the RHG.

We first consider the case when the reweighted estimator is computed at time $t = 1$. In the estimator of the variance due to non-response given in (4.23), the vector $\hat{\gamma}_1^1$ simplifies as

$$\hat{\gamma}_1^1 = \left(\frac{\sum_{i\in s_1\cap s_0^1} \frac{y_i}{\pi_i}}{\hat{p}_1^1 \sum_{i\in s_1\cap s_0^1} k_i^1}, \ldots, \frac{\sum_{i\in s_1\cap s_0^{C(0)}} \frac{y_i}{\pi_i}}{\hat{p}_{C(0)}^1 \sum_{i\in s_1\cap s_0^{C(0)}} k_i^1}\right)^\top. \tag{4.27}$$

After some algebra, the variance estimator in (4.23) may be rewritten as

$$\hat{V}_1^{nr}(\hat{Y}_1) = \sum_{c=1}^{C(0)} \frac{(1-\hat{p}_c^1)}{(\hat{p}_c^1)^2} \sum_{i\in s_1\cap s_0^c} \left(\frac{y_i}{\pi_i} - k_i^1 \frac{\sum_{j\in s_1\cap s_0^c} \frac{y_j}{\pi_j}}{\sum_{j\in s_1\cap s_0^c} k_j^1}\right)^2. \tag{4.28}$$

We now consider the case when the reweighted estimator is computed at time $t = 2$. We focus on the simpler case when the same system of RHGs is kept over time. In the estimator of the variance due to non-response given in (4.24), the vectors $\hat{\gamma}_2^1$ and $\hat{\gamma}_2^2$ simplify as

$$\hat{\gamma}_2^1 = \left(\frac{\sum_{i \in s_2 \cap s_1^1} \frac{y_i}{\pi_i}}{\hat{p}_1^1 \sum_{i \in s_2 \cap s_1^1} k_i^1}, \ldots, \frac{\sum_{i \in s_2 \cap s_1^{C(0)}} \frac{y_i}{\pi_i}}{\hat{p}_{C(0)}^1 \sum_{i \in s_2 \cap s_1^{C(0)}} k_i^1} \right)^\top, \quad (4.29)$$

$$\hat{\gamma}_2^2 = \left(\frac{\sum_{i \in s_2 \cap s_1^1} \frac{y_i}{\pi_i}}{\hat{p}_1^1 \hat{p}_1^2 \sum_{i \in s_2 \cap s_1^1} k_i^2}, \ldots, \frac{\sum_{i \in s_2 \cap s_1^{C(0)}} \frac{y_i}{\pi_i}}{\hat{p}_{C(0)}^1 \hat{p}_{C(0)}^2 \sum_{i \in s_2 \cap s_1^{C(0)}} k_i^2} \right)^\top. \quad (4.30)$$

After some algebra, the variance estimator in (4.24) may be rewritten as

$$\begin{aligned} \hat{V}_2^{nr}(\hat{Y}_2) &= \sum_{c=1}^{C(0)} \frac{(1 - \hat{p}_c^1)}{\hat{p}_c^2} \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_i}{\pi_i \hat{p}_c^1} - k_i^1 \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_j}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j^1} \right)^2 \\ &+ \sum_{c=1}^{C(0)} (1 - \hat{p}_c^2) \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_i}{\pi_i \hat{p}_c^1 \hat{p}_c^2} - k_i^2 \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_j}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j^2} \right)^2. \quad (4.31) \end{aligned}$$

If we further assume that k_i^δ is constant over times $\delta = 1, 2$, and may thus be rewritten as k_i, the expression in (4.31) simplifies as

$$\hat{V}_2^{nr}(\hat{Y}_2) = \sum_{c=1}^{C(0)} \frac{(1 - \hat{p}_c^{1 \to 2})}{(\hat{p}_c^{1 \to 2})^2} \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_i}{\pi_i} - k_i \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_j}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j} \right)^2. \quad (4.32)$$

with $\hat{p}_c^{1 \to 2} = \prod_{\delta=1}^2 \hat{p}_c^\delta$ for $c = 1, \ldots, C(0)$. This simplification of the variance estimator can be extended to the reweighted estimator computed at time t. Assuming that the RHGs are kept over time, and that $k_i^\delta = k_i$ for any $\delta = 1, \ldots, t$, the variance estimator in (4.14) may be written after some algebra as

$$\hat{V}_t^{nr}(\hat{Y}_t) = \sum_{c=1}^{C(0)} \frac{(1 - \hat{p}_c^{1 \to t})}{(\hat{p}_c^{1 \to t})^2} \sum_{i \in s_t \cap s_{t-1}^c} \left(\frac{y_i}{\pi_i} - k_i \frac{\sum_{j \in s_t \cap s_{t-1}^c} \frac{y_j}{\pi_j}}{\sum_{j \in s_t \cap s_{t-1}^c} k_j} \right)^2 \quad (4.33)$$

with $\hat{p}_c^{1 \to t} = \prod_{\delta=1}^t \hat{p}_c^\delta$ for $c = 1, \ldots, C(0)$.

5 Calibrated estimators and complex parameters

In most surveys, a calibration step is used to obtain adjusted weights which enable to improve the accuracy of total estimates. Such calibrated estimators

are considered in Section 5.1. Also, more complex parameters than totals are frequently of interest, and a linearization step can be used for variance estimation. This is the purpose of Section 5.2. The estimation of complex parameters with calibrated weights is treated in Section 5.3. In each case, explicit formulas for variance estimation and simplified variance estimation are derived, and the bias of the simplified variance estimator is discussed.

5.1 Variance estimation for calibrated total estimators

Assume that a vector x_i of auxiliary variables is available for any unit $i \in s_t$, and that the vector of totals X on the population U is known. Then an additional calibration step (Deville and Särndal, 1992) is usually applied to $\hat{Y}_t$. It consists in modifying the weights $d_{ti} = \pi_i^{-1}(\hat{p}_i^{1\to t})^{-1}$ to obtain calibrated weights w_{ti} which enable to match the real total X, in the sense that

$$\sum_{i \in s_t} w_{ti} x_i \quad = \quad X. \tag{5.1}$$

The new calibrated weights are chosen so as to minimize a distance function with the original weights, while satisfying (5.1). This leads to the calibrated estimator

$$\hat{Y}_{wt} \quad = \quad \sum_{i \in s_t} w_{ti} y_i. \tag{5.2}$$

Under some mild conditions on the chosen distance function, on the sampling design and on the response mechanisms, it can be shown that the calibrated estimator $\hat{Y}_{wt}$ is approximately unbiased for Y.

The estimated residual for the weighted regression of y_i on x_i is denoted by

$$e_i \quad = \quad y_i - \hat{b}_t x_i \tag{5.3}$$

$$\text{with } \hat{b}_t \quad = \quad \left(\sum_{i \in s_t} \frac{1}{\pi_i \hat{p}_i^{1\to t}} x_i x_i^\top \right)^{-1} \sum_{i \in s_t} \frac{1}{\pi_i \hat{p}_i^{1\to t}} x_i y_i. \tag{5.4}$$

Replacing in (4.12) the variable y_i with e_i yields the estimator of the variance due to the sampling design

$$\hat{V}_t^p(\hat{Y}_{wt}) \quad = \quad \sum_{i,j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{e_i}{\pi_i} \frac{e_j}{\pi_j}. \tag{5.5}$$

Similarly, replacing in (4.14) the variable y_i with e_i yields the estimator of the variance due to the non-response

$$\hat{V}_t^{nr}(\hat{Y}_{wt}) = \sum_{\delta=1}^{t}\sum_{i\in s_t}\frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}}\left(\frac{e_i}{\pi_i\hat{p}_i^{1\to\delta}}-k_i^\delta(\hat{h}_i^\delta)^\top\hat{\gamma}_{te}^\delta\right)^2 \tag{5.6}$$

$$\hat{\gamma}_{te}^\delta = \left\{\sum_{i\in s_t}k_i^\delta\frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}}\hat{h}_i^\delta(\hat{h}_i^\delta)^\top\right\}^{-1}\sum_{i\in s_t}\frac{1-\hat{p}_i^\delta}{\hat{p}_i^{1\to t}}\hat{h}_i^\delta\frac{e_i}{\pi_i}. \tag{5.7}$$

The global variance estimator for $\hat{Y}_{wt}$ is

$$\hat{V}_t(\hat{Y}_{wt}) = \hat{V}_t^p(\hat{Y}_{wt})+\hat{V}_t^{nr}(\hat{Y}_{wt}). \tag{5.8}$$

The simplified estimator of the variance due to non-response is

$$\begin{aligned}\hat{V}_{t,simp}^{nr}(\hat{Y}_{wt}) &= \sum_{\delta=1}^{t}\sum_{i\in s_t}\frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}}\left(\frac{e_i}{\pi_i\hat{p}_i^{1\to\delta}}\right)^2\\ &= \sum_{i\in s_t}\frac{1-\hat{p}_i^{1\to t}}{(\hat{p}_i^{1\to t})^2}\left(\frac{e_i}{\pi_i}\right)^2.\end{aligned} \tag{5.9}$$

Here again, this simplified variance estimator ignores the prediction terms $k_i^\delta(\hat{h}_i^\delta)^\top\hat{\gamma}_{te}^\delta$. If all the auxiliary variables that are explanatory for y_i are included in the calibration, which means that the underlying calibration model is appropriate, then e_i is essentially a white noise. The explanatory power of $\hat{h}_i^\delta$ for e_i is then expected to be small, as well as the prediction term $k_i^\delta(\hat{h}_i^\delta)^\top\hat{\gamma}_{te}^\delta$. In such case, we expect the bias of the simplified variance estimator to be small. If it is believed that some important auxiliary variables are not included in the calibration, then there may remain in e_i some significant part of y_i that may not been explained by the sole x_i. In such case, there may remain some explanatory power for $\hat{h}_i^\delta$ on e_i, and the bias of the simplified variance estimator may be non-negligible. The same problem may occur in case of domain estimation, when the calibration variables do not include any auxiliary information specific of the domain. In such case, the calibration model is not appropriate for domain estimation and the bias of the simplified variance estimator may be non-negligible.

5.2 Variance estimation for complex parameters

We may be interested in estimating more complex parameters than totals. Suppose that the variable of interest y is q-multivariate, and that the parameter of interest is $\theta = f(Y)$ with $f(\cdot)$ a known function. At time t,

substituting $\hat{Y}_t$ into θ yields the plug-in estimator $\hat{\theta}_t = f(\hat{Y}_t)$. Under some mild regularity conditions on the function f, on the sampling design and on the response mechanisms (see Deville, 1999; Goga et al., 2009), the plug-in estimator $\hat{\theta}_t$ is approximately unbiased for θ.

The estimated linearized variable of θ is

$$u_i = \{f'(\hat{Y}_t)\}^\top y_i, \tag{5.10}$$

with $f'(\hat{Y}_t)$ the q-vector of first derivatives of f at point $\hat{Y}_t$. Replacing in (4.12) the variable y_i with u_i yields the estimator of the variance due to the sampling design

$$\hat{V}_t^p(\hat{\theta}_t) = \sum_{i,j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{u_i}{\pi_i} \frac{u_j}{\pi_j}. \tag{5.11}$$

Similarly, replacing in (4.14) the variable y_i with u_i yields the estimator of the variance due to the non-response

$$\hat{V}_t^{nr}(\hat{\theta}_t) = \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^\delta (1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \left(\frac{u_i}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^\delta (\hat{h}_i^\delta)^\top \hat{\gamma}_{t\theta}^\delta \right)^2 \tag{5.12}$$

$$\hat{\gamma}_{t\theta}^\delta = \left\{ \sum_{i \in s_t} k_i^\delta \frac{\hat{p}_i^\delta (1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \hat{h}_i^\delta (\hat{h}_i^\delta)^\top \right\}^{-1} \sum_{i \in s_t} \frac{1-\hat{p}_i^\delta}{\hat{p}_i^{1\to t}} \hat{h}_i^\delta \frac{u_i}{\pi_i}. \tag{5.13}$$

The global variance estimator for $\hat{\theta}_t$ is

$$\hat{V}_t(\hat{\theta}_t) = \hat{V}_t^p(\hat{\theta}_t) + \hat{V}_t^{nr}(\hat{\theta}_t). \tag{5.14}$$

The simplified estimator of the variance due to non-response is

$$\begin{aligned} \hat{V}_{t,simp}^{nr}(\hat{\theta}_t) &= \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^\delta (1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \left(\frac{u_i}{\pi_i \hat{p}_i^{1\to\delta}} \right)^2 \\ &= \sum_{i \in s_t} \frac{1-\hat{p}_i^{1\to t}}{(\hat{p}_i^{1\to t})^2} \left(\frac{u_i}{\pi_i} \right)^2. \end{aligned} \tag{5.15}$$

The bias of this simplified variance estimator will depend on the explanatory power for $\hat{h}_i^\delta$ on the linearized variable u_i.

5.3 Variance estimation for complex parameters under calibration

The calibrated weights w_{ti} may also be used to obtain an estimator of the parameter θ at time t. Substituting $\hat{Y}_{wt}$ into $\theta = f(Y)$ yields the calibrated plug-in estimator $\hat{\theta}_{wt} = f(\hat{Y}_{wt})$. So as to obtain a variance estimator for $\hat{\theta}_{wt}$, we first compute the estimated linearized variable $u_i = \{f'(\hat{Y}_t)\}^\top y_i$. Then, we compute

$$e_{\theta i} = u_i - \hat{b}_{\theta t} x_i \tag{5.16}$$

$$\text{with } \hat{b}_{\theta t} = \left(\sum_{i \in s_t} \frac{1}{\pi_i \hat{p}_i^{1 \to t}} x_i x_i^\top \right)^{-1} \sum_{i \in s_t} \frac{1}{\pi_i \hat{p}_i^{1 \to t}} x_i u_i. \tag{5.17}$$

Replacing in (4.12) the variable y_i with $e_{\theta i}$ yields the estimator of the variance due to the sampling design

$$\hat{V}_t^p(\hat{\theta}_{wt}) = \sum_{i,j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1 \to t}} \frac{e_{\theta i}}{\pi_i} \frac{e_{\theta j}}{\pi_j}. \tag{5.18}$$

Similarly, replacing in (4.14) the variable y_i with $e_{\theta i}$ yields the estimator of the variance due to the non-response

$$\hat{V}_t^{nr}(\hat{\theta}_{wt}) = \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^\delta (1 - \hat{p}_i^\delta)}{\hat{p}_i^{\delta \to t}} \left(\frac{e_{\theta i}}{\pi_i \hat{p}_i^{1 \to \delta}} - k_i^\delta (\hat{h}_i^\delta)^\top \hat{\gamma}_{te\theta}^\delta \right)^2 \tag{5.19}$$

$$\hat{\gamma}_{te\theta}^\delta = \left\{ \sum_{i \in s_t} k_i^\delta \frac{\hat{p}_i^\delta (1 - \hat{p}_i^\delta)}{\hat{p}_i^{\delta \to t}} \hat{h}_i^\delta (\hat{h}_i^\delta)^\top \right\}^{-1} \sum_{i \in s_t} \frac{1 - \hat{p}_i^\delta}{\hat{p}_i^{1 \to t}} \hat{h}_i^\delta \frac{e_{\theta i}}{\pi_i}. \tag{5.20}$$

The global variance estimator for $\hat{\theta}_{wt}$ is

$$\hat{V}_t(\hat{\theta}_{wt}) = \hat{V}_t^p(\hat{\theta}_{wt}) + \hat{V}_t^{nr}(\hat{\theta}_{wt}). \tag{5.21}$$

The simplified estimator of the variance due to non-response is

$$\begin{aligned} \hat{V}_{t,simp}^{nr}(\hat{\theta}_{wt}) &= \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^\delta (1 - \hat{p}_i^\delta)}{\hat{p}_i^{\delta \to t}} \left(\frac{e_{\theta i}}{\pi_i \hat{p}_i^{1 \to \delta}} \right)^2 \\ &= \sum_{i \in s_t} \frac{1 - \hat{p}_i^{1 \to t}}{(\hat{p}_i^{1 \to t})^2} \left(\frac{e_{\theta i}}{\pi_i} \right)^2. \end{aligned} \tag{5.22}$$

The bias of this simplified variance estimator will depend on the explanatory power for $\hat{h}_i^\delta$ on $e_{\theta i}$. Since the variable $e_{\theta i}$ is obtained as the residual in the

regression of the linearized variable u_i on the calibration variables x_i, the explanatory power for $\hat{h}_i^\delta$ on $e_{\theta i}$ is expected to be small in practice, and the bias of the simplified variance estimator is expected to be small as well.

As an illustration, we consider the model of Response Homogeneity Groups, and the simple case when RHGs are kept over time and when $k_i^\delta = k_i$ for any $\delta = 1, \ldots, t$. In such case, the estimator of the variance due to the non-response in (5.19) may be rewritten as

$$\hat{V}_t^{nr}(\hat{\theta}_{wt}) = \sum_{c=1}^{C(0)} \frac{(1-\hat{p}_c^{1\to t})}{(\hat{p}_c^{1\to t})^2} \sum_{i \in s_t \cap s_{t-1}^c} \left(\frac{e_{\theta i}}{\pi_i} - k_i \frac{\sum_{j \in s_t \cap s_{t-1}^c} \frac{e_{\theta j}}{\pi_j}}{\sum_{j \in s_t \cap s_{t-1}^c} k_j} \right)^2 . \quad (5.23)$$

6 Longitudinal estimators

We may be interested in a change in parameters, such as the difference between the totals of a variable of interest measured at two different times $u < t$. Denoting by y_{ui} and y_{ti} the value of this variable of interest for unit i at times u and t, respectively, and denoting by $Y(u) = \sum_{i \in U} y_{ui}$ and $Y(t) = \sum_{i \in U} y_{ti}$ their totals, the parameter of interest is

$$\Delta(u \to t) \quad = \quad Y(t) - Y(u). \quad (6.1)$$

Since the variable y_{ui} is measured on all sub-samples $s_{u'}$ for $u' = u, \ldots, t$, there are several possible estimators for $\Delta(u \to t)$. For $u' = u, \ldots, t$, we denote by

$$\hat{\Delta}_{u't}(u \to t) \quad = \quad \sum_{i \in s_t} \frac{y_{ti}}{\pi_i \hat{p}_i^{1\to t}} - \sum_{i \in s_{u'}} \frac{y_{ui}}{\pi_i \hat{p}_i^{1\to u'}} \quad (6.2)$$

the estimator which makes use of the sample s_t for the estimation of $Y(t)$, and of the sample $s_{u'}$ for the estimation of $Y(u)$. The case $u' = u$ corresponds to the estimation of $Y(u)$ on the largest available sub-sample, s_u. The case $u' = t$ corresponds to the estimation of $Y(u)$ and $Y(t)$ on the common sub-sample, s_t.

In the context of full response, several authors have recommended the estimator $\hat{\Delta}_{tt}(u \to t)$ which makes use of the common sample only, if the variables y_{ui} and y_{ti} are strongly positively correlated; see Caron and Ravalet (2000), Qualité and Tillé (2008), Goga et al. (2009), Chauvet and Goga (2016). In our context, this choice may be heuristically justified as follows. For $u' < t$,

and by conditioning on the sub-sample $s_{u'}$, we obtain

$$V\left\{\hat{\Delta}_{u't}(u\rightarrow t)\right\} \simeq V\left\{\sum_{i\in s_{u'}}\frac{y_{ti}-y_{ui}}{\pi_i\hat{p}_i^{1\rightarrow u'}}\right\}+EV\left\{\sum_{i\in s_t}\frac{y_{ti}}{\pi_i\hat{p}_i^{1\rightarrow t}}\middle| s_{u'}\right\} \quad (6.3)$$

$$V\left\{\hat{\Delta}_{tt}(u\rightarrow t)\right\} \simeq V\left\{\sum_{i\in s_{u'}}\frac{y_{ti}-y_{ui}}{\pi_i\hat{p}_i^{1\rightarrow u'}}\right\}+EV\left\{\sum_{i\in s_t}\frac{y_{ti}-y_{ui}}{\pi_i\hat{p}_i^{1\rightarrow t}}\middle| s_{u'}\right\} \quad (6.4)$$

In equations (6.3) and (6.4), the first term in the right-hand side is identical. If the variables y_{ui} and y_{ti} are positively correlated, then the difference $y_{ti}-y_{ui}$ is expected to be smaller than y_{ti}, so that the second term in the right-hand side of (6.4) is expected to be smaller than the second term in the right-hand side of (6.3). Therefore, the estimator $\hat{\Delta}_{tt}(u\rightarrow t)$ based on the common sample is expected to be more efficient in terms of variance.

The results of a small simulation study in Section 7.2 support this heuristic reasoning. Therefore, we focus only in this Section on the estimator $\hat{\Delta}_{tt}(u\rightarrow t)$ for the estimation of $\Delta(u\rightarrow t)$. Replacing in (4.12) the variable y_i with $y_{ti}-y_{ui}$ yields the estimator of the variance due to the sampling design

$$\hat{V}_t^p\left\{\hat{\Delta}_{tt}(u\rightarrow t)\right\} = \sum_{i,j\in s_t}\frac{\Delta_{ij}}{\pi_{ij}}\frac{1}{\hat{p}_{ij}^{1\rightarrow t}}\frac{(y_{ti}-y_{ui})}{\pi_i}\frac{(y_{tj}-y_{uj})}{\pi_j}. \quad (6.5)$$

Similarly, replacing in (4.14) the variable y_i with $y_{ti}-y_{ui}$ yields the estimator of the variance due to the non-response

$$\hat{V}_t^{nr}\left\{\hat{\Delta}_{tt}(u\rightarrow t)\right\} = \sum_{\delta=1}^{t}\sum_{i\in s_t}\frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\rightarrow t}}\left(\frac{y_{ti}-y_{ui}}{\pi_i\hat{p}_i^{1\rightarrow\delta}}-k_i^\delta(\hat{h}_i^\delta)^\top\hat{\gamma}_{t\Delta}^\delta\right)^2 \quad (6.6)$$

with

$$\hat{\gamma}_{t\Delta}^\delta = \left\{\sum_{i\in s_t}k_i^\delta\frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\rightarrow t}}\hat{h}_i^\delta(\hat{h}_i^\delta)^\top\right\}^{-1}\sum_{i\in s_t}\frac{1-\hat{p}_i^\delta}{\hat{p}_i^{1\rightarrow t}}\hat{h}_i^\delta\frac{y_{ti}-y_{ui}}{\pi_i}. \quad (6.7)$$

The global variance estimator for $\hat{\Delta}_{tt}(u\rightarrow t)$ is

$$\hat{V}_t\left\{\hat{\Delta}_{tt}(u\rightarrow t)\right\} = \hat{V}_t^p\left\{\hat{\Delta}_{tt}(u\rightarrow t)\right\}+\hat{V}_t^{nr}\left\{\hat{\Delta}_{tt}(u\rightarrow t)\right\}. \quad (6.8)$$

Variance estimation for measures of change is also considered in Berger (2004), Qualité and Tillé (2008), Goga et al. (2009), Chauvet and Goga

(2016), among others.

The simplified estimator of the variance due to non-response is

$$\begin{aligned}\hat{V}^{nr}_{t,simp}(\hat{\Delta}_{tt}(u \to t)) &= \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta \to t}} \left(\frac{y_{ti} - y_{ui}}{\pi_i \hat{p}_i^{1 \to \delta}} \right)^2 \\ &= \sum_{i \in s_t} \frac{1-\hat{p}_i^{1 \to t}}{(\hat{p}_i^{1 \to t})^2} \left(\frac{y_{ti} - y_{ui}}{\pi_i} \right)^2. \qquad (6.9)\end{aligned}$$

If the variables y_{ti} and y_{ui} are strongly positively correlated, the explanatory power for $\hat{h}_i^{\delta}$ on $y_{ti} - y_{ui}$ is expected to be small in practice. In such case, the bias of the simplified variance estimator is also expected to be small.

7 A simulation study

In this Section, several artificial populations are generated according to some superpopulation model described in Section 7.1. In Section 7.2, we consider several estimators for a change between totals, which illustrates the heuristic reasoning in Section 6. A Monte Carlo experiment is then presented in Section 7.3, and several variance estimators for estimating a total, a ratio or a parameter change are compared. The results from Tables 1 and 2 are readily reproducible using the R code provided in the supplementary material of the present paper.

7.1 Simulation set-up

We consider seven populations of size $10,000$, each containing three variables of interest y_{1i}, y_{2i} and y_{3i} observed at times $t = 1, 2$ and 3, respectively. The variables of interest are generated according to the superpopulation model

$$\begin{aligned} y_{1i} &= \alpha^0 + \alpha^a x_{ai} + \alpha^b x_{bi} + \sigma u_{1i}, \qquad (7.1) \\ y_{2i} &= \rho y_{1i} + \sigma u_{2i}, \qquad (7.2) \\ y_{3i} &= \rho y_{2i} + \sigma u_{3i}. \qquad (7.3) \end{aligned}$$

The auxiliary variables x_{ai} and x_{bi} are independently generated from a Gamma distribution with shape and scale parameters 2 and 1. Two other auxiliary variables x_{ci} and x_{di} are also independently generated from a Gamma distribution with shape and scale parameters 2 and 1. These two last variables are not related to the variables of interest. The variables u_{1i}, u_{2i} and u_{3i} are independently generated according to a standard normal distribution. We

use $\alpha^0 = 10$, $\alpha^a = \alpha^b = 5$ and $\sigma = 10$, which leads to a coefficient of determination (R^2) in model (7.1) approximately equal to 0.50. The parameter ρ is set to 0 for population 1, 0.2 for population 2, 0.4 for population 3, 0.6 for population 4, 0.8 for population 5, 1.0 for population 6 and 1.2 for population 7.

For each population, a simple random sample s_0 of size $n = 1,000$ is selected. Three non-response phases are then successively simulated. At each phase $\delta = 1, 2, 3$, the sub-sample of respondents s_δ is obtained by Poisson sampling with a response probability p_i^δ for unit i, defined as

$$\text{logit}(p_i^\delta) \quad = \quad \beta^{\delta 0} + \beta^{\delta a} x_{ai} + \beta^{\delta b} x_{bi}. \tag{7.4}$$

We use $\beta^{\delta 0} = -1$ at each phase $\delta = 1, 2, 3$. For $\delta = 1$, we use $\beta^{1a} = \beta^{1b} = 0.60$, which corresponds to an average response rate of 0.75. For $\delta = 2, 3$, we use $\beta^{\delta a} = \beta^{\delta b} = 0.75$, which corresponds to an average response rate of 0.81. Inside each sub-sample s_δ, the estimated response probabilities $\hat{p}_i^\delta$ are obtained by means of an unweighted logistic regression.

7.2 Comparison of estimators for a difference of totals

In this Section, we are interested in comparing the accuracy of two estimators for a difference of totals $\Delta(u \to t)$ for $u = 1$ and $t = 2$, for $u = 1$ and $t = 3$, and for $u = 2$ and $t = 3$. We consider the estimator $\hat{\Delta}_{ut}(u \to t)$, which makes use of the whole appropriate sub-samples for variables y_{ui} and y_{ti}, and the estimator $\hat{\Delta}_{tt}(u \to t)$, which makes use of the common sub-sample only. These two estimators are compared through the relative difference (RD) of their variances, which are defined as follows:

$$RD(1 \to 2) \quad = \quad 100 \times \frac{V\left\{\hat{\Delta}_{12}(1 \to 2)\right\} - V\left\{\hat{\Delta}_{22}(1 \to 2)\right\}}{V\left\{\hat{\Delta}_{22}(1 \to 2)\right\}}, \tag{7.5}$$

$$RD(1 \to 3) \quad = \quad 100 \times \frac{V\left\{\hat{\Delta}_{13}(1 \to 3)\right\} - V\left\{\hat{\Delta}_{33}(1 \to 3)\right\}}{V\left\{\hat{\Delta}_{33}(1 \to 3)\right\}}, \tag{7.6}$$

$$RD(2 \to 3) \quad = \quad 100 \times \frac{V\left\{\hat{\Delta}_{23}(2 \to 3)\right\} - V\left\{\hat{\Delta}_{33}(2 \to 3)\right\}}{V\left\{\hat{\Delta}_{33}(2 \to 3)\right\}}. \tag{7.7}$$

The true variances are replaced by their Monte Carlo approximation, obtained by repeating $B = 100,000$ times the sample selection and the non-

response phases.

The results are presented in Table 1. A positive RD indicates that the use of the common sample only leads to a more accurate estimator. As could be expected, the RD increases in all cases with ρ, that is, when the correlation between y_{ti} and y_{ui} increases. For $u = 1$ and $t = 2$, and for $u = 2$ and $t = 3$, the estimator $\hat{\Delta}_{tt}(u \to t)$ is more accurate for ρ greater than 0.6. For $u = 1$ and $t = 3$, $\hat{\Delta}_{tt}(u \to t)$ is more accurate for ρ greater than 0.8.

ρ	$RD(1 \to 2)$	$RD(1 \to 3)$	$RD(2 \to 3)$
0.0	-12	-27	-13
0.2	-09	-25	-11
0.4	-04	-20	-03
0.6	05	-09	11
0.8	17	11	39
1.0	30	33	83
1.2	40	46	127

Table 1: Relative Difference (RD) between two estimators for a difference of totals

7.3 Performances of the variance estimators

In this Section, we consider the artificial population 5 ($\rho = 0.8$) generated as described in Section 7.1. The sample selection by means of simple random sampling of size $n = 1,000$ and the three non-response phases are applied $B = 5,000$ times. We are interested in evaluating the variance estimators and the simplified variance estimators, in case of estimating a total, a ratio or a change in totals.

As for the total Y, we consider at each time $t = 1, 2, 3$, three estimators. The estimator $\hat{Y}_t$ makes use of the weights $d_{ti} = \pi_i^{-1}(\hat{p}_i^{1\to t})^{-1}$. The estimator $\hat{Y}_{wt}$ makes use of the weights w_i, obtained by calibrating the weights d_{ti} on the population size and on the totals of the auxiliary variables x_{ai} and x_{bi}. In view of model (7.1), the working model underlying this calibration is well-specified. Finally, the estimator $\hat{Y}_{\tilde{w}t}$ makes use of the weights $\tilde{w}_i$, obtained by calibrating the weights d_{ti} on the population size and on the totals of the auxiliary variables x_{ci} and x_{di}. In view of model (7.1), the working model underlying this calibration is therefore not correctly specified. The proposed variance estimator for $\hat{Y}_t$ is obtained from equation (4.18), and the simplified

variance estimator is obtained by plugging in (4.18) the simplified variance estimator for non-response given in (4.19). The proposed variance estimators for $\hat{Y}_{wt}$ and $\hat{Y}_{\tilde{w}t}$ are obtained from equation (5.8), and the simplified variance estimators are obtained by plugging in (5.8) the simplified variance estimator for non-response given in (5.9).

We are also interested in estimating the ratio $R_t = Y_t/Y_1$ for $t = 2, 3$. At each time t, we consider three estimators. The estimator $\hat{R}_t$ makes use of the weights d_i. The proposed variance estimator is obtained from equation (5.14), by using the estimated linearized variable $u_i = (\hat{Y}_1)^{-1}(y_{ti} - \hat{R}_t y_{1i})$. The simplified variance estimator is obtained by plugging in (5.14) the simplified variance estimator for non-response given in (5.15). The estimators $\hat{R}_{wt}$ and $\hat{R}_{\tilde{w}t}$ make use of the calibrated weights w_i and $\tilde{w}_i$. The proposed variance estimators are obtained from equation (5.21). The simplified variance estimators are obtained by plugging in (5.21) the simplified variance estimator for non-response given in (5.22).

Finally, we are interested in estimating the change in totals $\Delta(1 \to t)$ for $t = 2, 3$. At each time t, we consider three estimators. The estimator $\hat{\Delta}_{tt}(1 \to t)$ makes use of the weights d_i. The proposed variance estimator is obtained from equation (6.8), and the simplified variance estimator is obtained by plugging in (6.8) the simplified variance estimator for non-response given in (6.9). The estimators $\hat{\Delta}_{tt,w}(1 \to t)$ and $\hat{\Delta}_{tt,\tilde{w}}(1 \to t)$ make use of the calibrated weights w_i and $\tilde{w}_i$. The proposed variance estimators are obtained from equation (6.8), by replacing $y_{ti} - y_{ui}$ by the estimated residual for the weighted regression of $y_{ti} - y_{ui}$ on the calibration variables. The simplified variance estimators are obtained by plugging in (6.8) the simplified variance estimator for non-response given in (6.9).

For a proposed variance estimator $\hat{V}$, we computed the Monte Carlo Percent Relative Bias

$$\mathrm{RB}_{\mathrm{MC}}(\hat{V}) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}^{(b)} - V}{V}$$

where the global variance V was approximated through an independent set of $100,000$ simulations. So as to evaluate the contribution of some component $\hat{V}_a$ into the proposed variance estimator $\hat{V}$, we also computed the contribution (in percent)

$$\mathrm{CONTR}_{\mathrm{MC}}(\hat{V}_a) = 100 \times \frac{\frac{1}{B}\sum_{b=1}^{B}\hat{V}_a^{(b)}}{\frac{1}{B}\sum_{b=1}^{B}\hat{V}^{(b)}}.$$

So as to evaluate the simplified variance estimator for the non-response $\hat{V}^{nr}_{simp}$, we also computed the Monte Carlo Percent Relative Bias

$$\mathrm{RB}_{\mathrm{MC}}(\hat{V}^{nr}_{simp}) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}^{(b)}_{simp} - V^{nr}}{V^{nr}},$$

where the variance V^{nr} due to non-response was approximated through an independent set of $100,000$ simulations.

The simulation results are presented in Table 2. The proposed variance estimator is almost unbiased in all cases. As could be expected, the contribution of the variance due to the sampling design decreases with time, as the number of respondents decreases and as the variance due to non-response becomes larger. The simplified variance estimator is highly biased for the variance due to non-response in case of $\hat{Y}_t$. The bias decreases quickly with time, but remains large at time $t = 3$. The simplified variance estimator is almost unbiased for a calibrated estimator when the working model is adequately specified, but is severely biased otherwise. This is consistent with our reasoning in Section 5.1. The simplified variance estimator is almost unbiased for the three estimators of the ratio, and for the calibrated estimators of the change in totals. In case of the non-calibrated estimator for the change in totals, the bias can be as high as 30 % .

8 Illustration

In this Section, we aim at illustrating the results previously obtained on a real data set from the ELFE survey. Covering the whole metropolitan France, it was launched in 2011 and consists of more than 18,000 children whose parents consented to their inclusion. The population of inference consists of infants born in one of the 544 French maternity units during 2011, except very premature infants.

An original sample s_0 of about $35,600$ infants was originally selected when the babies were just a few days old and were still at the maternity unit. The sample was selected using a cross-classified sampling design (Skinner, 2015; Juillard et al., 2016). A sample of days and a sample of maternity units were independently selected, and both sample selections may be approximated by stratified simple random sampling (STSI). The sample sizes inside strata are provided in Tables 3 and 4. The sample consisted in all the infants born during one of the 25 selected days in one of the 320 selected maternity units.

	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$
	$\hat{Y}_t$			$\hat{Y}_{wt}$			$\hat{Y}_{\tilde{w}t}$		
$\mathrm{RB_{MC}}(\hat{V})$	-0	-1	-2	-1	-1	-2	-1	-1	-3
$\mathrm{CONTR_{MC}}(\hat{V}_t^p)$	81	57	35	69	49	32	80	56	35
$\mathrm{CONTR_{MC}}(\hat{V}_t^{nr1})$	19	19	13	31	22	15	20	18	13
$\mathrm{CONTR_{MC}}(\hat{V}_t^{nr2})$	-	25	18	-	28	19	-	25	17
$\mathrm{CONTR_{MC}}(\hat{V}_t^{nr3})$	-	-	34	-	-	34	-	-	34
$\mathrm{RB_{MC}}(\hat{V}_{simp}^{nr})$	559	188	80	0	-1	-2	83	34	15
	$\hat{R}_t$			$\hat{R}_{wt}$			$\hat{R}_{\tilde{w}t}$		
$\mathrm{RB_{MC}}(\hat{V})$	-	-0	-2	-	-1	-2	-	-1	-2
$\mathrm{CONTR_{MC}}(\hat{V}_t^p)$	-	49	32	-	49	32	-	50	33
$\mathrm{CONTR_{MC}}(\hat{V}_t^{nr1})$	-	22	15	-	22	15	-	22	15
$\mathrm{CONTR_{MC}}(\hat{V}_t^{nr2})$	-	28	19	-	28	19	-	28	19
$\mathrm{CONTR_{MC}}(\hat{V}_t^{nr3})$	-	-	34	-	-	34	-	-	34
$\mathrm{RB_{MC}}(\hat{V}_{simp}^{nr})$	-	0	0	-	-1	-2	-	-1	-1
	$\hat{\Delta}_{tt}(1 \to t)$			$\hat{\Delta}_{tt,w}(1 \to t)$			$\hat{\Delta}_{tt,\tilde{w}}(1 \to t)$		
$\mathrm{RB_{MC}}(\hat{V})$	-	-0	-2	-	-0	-2	-	-1	-3
$\mathrm{CONTR_{MC}}(\hat{V}_t^p)$	-	50	33	-	49	32	-	50	33
$\mathrm{CONTR_{MC}}(\hat{V}_t^{nr1})$	-	22	14	-	22	15	-	22	14
$\mathrm{CONTR_{MC}}(\hat{V}_t^{nr2})$	-	28	18	-	28	19	-	28	18
$\mathrm{CONTR_{MC}}(\hat{V}_t^{nr3})$	-	-	34	-	-	34	-	-	34
$\mathrm{RB_{MC}}(\hat{V}_{simp}^{nr})$	-	19	30	-	-1	-2	-	3	5

Table 2: Relative bias of a global variance estimator, relative contribution to the estimators of variance components and relative bias of a simplified variance estimator for the variance due to non-response for the estimation of a total, a ratio or a change in totals with three sets of weights

Strata g	Strata size N_{Mg}	Sample size n_{Mg}
1	108	21
2	108	41
3	109	55
4	108	80
5	111	90
Total	544	287

Table 3: Population and sample strata sizes for the maternity units design.

Strata h	Strata size N_{Dh}	Sample size n_{Dh}
1	91	4
2	91	6
3	91	7
4	92	8
Total	365	25

Table 4: Population and sample strata sizes for the days design.

Among the $35,600$ infants originally selected, a total of $18,329$ face-to-face interviews were completed with their families, which represents a response rate of 51 % . This led to the subsample s_1 after accounting for non-response. The weights at time $t = 1$ were computed on the basis of the original sampling weights, adjusted in two steps. First, response probabilities were estimated by means of a model of Response Homogeneity Groups (RHGs), with 20 RHGs defined by using a logistic regression model with explanatory variables *Age of the mother*, *Gemellary identity* and *Season of birth*. Then, a calibration by means of the raking ratio method was performed on the binary variables *Born within marriage*, *Immigrant mother* and *Gemellary identity*.

When the children reached the age of two months, the parents had the first telephone interview with a response rate of 87 % . This leads to the subsample s_2. The weights at time $t = 2$ were computed on the basis on the weight obtained at time $t = 1$, with a two-step adjustment. First, response probabilities were estimated by means of 20 RHGs, defined by using a logistic regression with explanatory variables *Age of the mother*, *Mother nationality* and *Father present at childbirth*. Then, a calibration by means of the raking ratio method was performed on the same calibration variables as at time $t = 1$.

When the children were one year old, the parents were contacted by phone with a response rate of 77 % . This led to the subsample s_3. The weights at time $t = 3$ were computed on the basis on the weights obtained at time $t = 2$, with a two-step adjustment similar to that realized at time $t = 2$. The parents were expected to be also interviewed when the infants would reach the age of two, three and half years and five and half years, but at the time when the paper was written, the three first waves only were available.

We considered three variables of interest: *Breastfeeding exclusivity at the childbirth, at two month, at one year*. For each of these variables, we computed the estimator $\hat{Y}_t$ from equation (4.4) and the estimated variance $\hat{V}_t(\hat{Y}_t)$ from the equation (4.18). We also computed the estimated coefficient of variation (in percent), defined as

$$\widehat{\mathrm{CV}}_t\left(\hat{Y}_t\right) = 100 \times \frac{\sqrt{\hat{V}_t\left(\hat{Y}_t\right)}}{\hat{Y}_t}. \tag{8.1}$$

For each component $\hat{V}_{ta}$ in the estimated variance $\hat{V}_t$, we computed its contribution (in percent) defined as

$$\mathrm{CONTR}(\hat{V}_{ta}) = 100 \times \frac{\hat{V}_{ta} - \hat{V}_t}{\hat{V}_t}. \tag{8.2}$$

We also computed the simplified variance estimator for non-response $\hat{V}^{nr}_{t,simp}$ given in (4.19), and the relative difference (in percent) with the approximately unbiased variance estimator $\hat{V}^{nr}$ defined as

$$\mathrm{RD}(\hat{V}^{nr}_{simp}) = 100 \times \frac{\hat{V}^{nr}_{simp} - \hat{V}^{nr}_t}{\hat{V}^{nr}_t}. \tag{8.3}$$

The results are given in the upper left of Table 5. For each of the three variables of interest, we also computed the calibrated estimator $\hat{Y}_{wt}$, and the same indicators. They are given in the upper right of Table 5. Finally, for each variable interest, we computed the estimator $\hat{R}_t$ and the calibrated estimator $\hat{R}_{wt}$ for the percentage of breastfeeding among all the children. The same indicators were computed. They are presented in the lower part of Table 5. As observed in the simulation study, the RD of the simplified variance estimator for non-response can be large in case of the estimator of the total without calibration, but the bias decreases with time. The bias appears as negligible for the calibrated estimator of the total, and for both estimators of the ratio.

Breastfeeding exclusivity	$t=1$ maternity	$t=2$ 2 months	$t=3$ 1 year	$t=1$ maternity	$t=2$ 2 months	$t=3$ 1 year
	without calibration			with calibration		
$\hat{Y}_t$	402409	209009	22658	415272	214262	23276
$\hat{V}_t(\cdot)$	8.51E+07	2.32E+07	1.58E+06	5.95E+06	6.93E+06	1.21E+06
$\hat{\mathrm{CV}}_t(\cdot)$ (%)	2.3	2.3	5.6	0.6	1.2	4.7
CONTR($\hat{V}_t^p$)	94	78	42	28	34	25
CONTR($\hat{V}_t^{nr1}$)	6	17	32	72	51	42
CONTR($\hat{V}_t^{nr2}$)	-	5	10	-	15	13
CONTR($\hat{V}_t^{nr3}$)	-	-	16	-	-	21
RD($\hat{V}_{simp}^{nr}$)	91	31	3	1	2	0
$\hat{R}_t$ (%)	59.0	30.6	3.3	59.4	31.0	3.4
$\hat{V}(\hat{R}_t)$	1.34E-05	1.50E-05	2.58E-06	1.28E-05	1.48E-05	2.60E-06
$\hat{\mathrm{CV}}(\hat{Y}_t)$ (%)	0.6	1.3	4.8	0.6	1.2	4.7
CONTR($\hat{V}_t^p$)	31	34	24	28	34	25
CONTR($\hat{V}_t^{nr1}$)	69	51	42	72	51	41
CONTR($\hat{V}_t^{nr2}$)	-	15	13	-	15	13
CONTR($\hat{V}_t^{nr3}$)	-	-	21	-	-	21
RD($\hat{V}_{simp}^{nr}$)	2	2	0	1	2	0

Table 5: Estimates for a total and a ratio, variance estimates, estimated coefficient of variation, relative contributions of variance components and relative difference of a simplified variance estimator for some variables in the ELFE survey

9 Conclusion

In this paper, we considered variance estimation accounting for weighting adjustments in panel surveys. We proposed both an approximately unbiased variance estimator and a simplified variance estimator for estimators of totals, complex parameters and measures of change, which covers most cases that may be encountered in practice. Our simulation results indicate that the proposed variance estimator performs well in all cases considered. The simplified variance estimator tends to overestimate the variance of the expansion estimator for totals, and to overestimate the variance for calibrated estimators of totals when the calibration variables lack of explanatory power for the variable of interest. However, the simplified variance estimator performs well for the estimation of ratios and change in totals with calibrated weights, even if the underlying calibration model is not appropriate for the study variable.

The assumption of independent response behaviour is usually not tenable for multi-stage surveys, since units within clusters tend to be correlated with respect to the response behaviour. In this context, estimation of response probabilities based upon conditional logistic regression in the context of correlated responses has been studied by Skinner and D'Arrigo (2011), see also Kim et al. (2016). Extending the present work in the context of correlated response behaviour is a challenging problem for further research.

10 Supplementary Materials

The three supplemental files are contained in a single archive.

readme: description of the supplemental files. (txt file)

CodeR_Functions: basic functions required to calculate estimators. (R file)

CodeR_Tables: commands that calculate and display the results in Table 1 and Table 2 (call the CodeR_Functions). (R file)

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society B*, 67:445–458.

Berger, Y. (2004). Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, 32(4):451–467.

Boistard, H., Lopuhaa, H., and Ruiz-Gazen, A. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electron. J. Statist.*, 6:1967–1983.

Breidt, F. and Opsomer, J. (2000). Local polynomial regresssion estimators in survey sampling. *Ann. Statist.*, 28(4):1026–1053.

Cardot, H., Goga, C., and Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electron. J. Statist.*, 7:562–596.

Caron, N. and Ravalet, P. (2000). Estimation dans les enquêtes répétées : application à l'enquête emploi en continu. *Technical report INSEE, Paris.*

Chauvet, G. and Goga, C. (2016). Linearization versus bootstrap for variance estimation of the change between Gini indexes. *In revision for Survey Methodology.*

Clarke, P. and Tate, P. (2002). An application of non-ignorable non-response models for gross flows estimation in the British labour force survey. *Australian and New Zealand Journal of Statistics*, 4:413–425.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25:193–203.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Ekholm, A. and Laaksonen, S. (1991). Weighting via response modeling in the Finnish household budget survey. *Journal of Official Statistics*, 7:325–327.

Fuller, W. and An, A. (1998). Regression adjustment for non-response. *Journal of the Indian Society of Agricultural Statistics*, 51:331–342.

Fuller, W. A., Loughin, M. M., and Baker, H. D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 national food consumption survey. *Survey Methodology*, 20:75–85.

Goga, C., Deville, J.-C., and Ruiz-Gazen, A. (2009). Composite estimation and linearization method for two-sample survey data. *Biometrika*, 96:691–709.

Hawkes, D. and Plewis, I. (2009). Modelling nonresponse in the national child development study. *Journal of the royal Statistical Society Series A*, 169:479–491.

Juillard, H., Chauvet, G., and Ruiz-Gazen, A. (2016). Estimation under cross-classified sampling with application to a childhood survey. *To appear in Journal of the American Statistical Association*.

Kalton, G. (2009). Design for surveys over time. *Handbook of Statistics*, 29:89–108.

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

Kim, J. K., Kwon, Y., and Park, M. (2016). Calibrated propensity score method for survey nonresponse in cluster sampling. *Biometrika*, 103:461–473.

Laaksonen, S. (2007). Weighting for two-phase surveyed data. *Survey Methodology*, 33:121–130.

Laaksonen, S. and Chambers, R. L. (2006). Survey estimation under informative nonresponse with follow-up. *Journal of Official Statistics*, 22:81–95.

Laniel, N. (1988). Variances for a rotating sample from a changing population. *Proceedings of the Business and Economics Statistics Section, American Statistical Association*, pages 246–250.

Laurie, H., Smith, R., and Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, 15:269–282.

Lynn, P. (2009). Methods for longitudinal surveys. *Methodology of Longitudinal Surveys*, pages 1–19.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16:363–378.

Pirus, C., Bois, C., Dufourg, M., Lanoë, J., Vandentorren, S., Leridon, H., and the Elfe team (2010). Constructing a cohort: Experience with the French Elfe project. *Population*, 65:637–670.

Qualité, L. and Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34:173–181.

Rendtel, U. and Harms, T. (2009). Weighting and calibration for household panels. *Methodology of Longitudinal Surveys*, pages 265–286.

Rizzo, L., Kalton, G., and Brick, J. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22:43–53.

Skinner, C. (2015). Cross-classified sampling: some estimation theory. *Statistics and Probability Letters*, 104:163–168.

Skinner, C. and D'Arrigo, J. (2011). Inverse probability weighting for clustered non-response. *Biometrika*, 98:953–966.

Skinner, C. and Vieira, M. (2005). Design effects in the analysis of longitudinal survey data. *S3RI Methdology Working Papers, M05/13. Southampton, UK: Southampton Statistical Sciences Research Institute.*

Slud, E. V. and Bailey, L. (2010). Evaluation and selection of models for attrition nonresponse adjustment. *Journal of Official Statistics*, 26:1–18.

Tam, S. (1984). On covariance from overlapping samples. *The American Statistician*, 38:1–18.

Vandecasteele, L. and Debels, A. (2007). Attrition in panel data: The effectiveness of weighting. *European Sociological Review*, 23(1):81–97.

Zhou, M. and Kim, J. (2012). An efficient method of estimation for longitudinal surveys with monotone missing data. *Biometrika*, 99:631–648.

A Some technical conditions

We make the following assumptions:

H1: There exists some constant $f \in]0,1[$ such that $N^{-1}n \to f$.

H2: There exists some constants $0 < C_1 \leq C_2$ such that for any $i \in U$: $C_1 \leq Nn^{-1}\pi_i \leq C_2$.

H3: There exists some constant $C_3 > 0$ such that $\sup_{i \neq j \in U} |\pi_{ij}^{-1} \Delta_{ij}| \leq C_3 n^{-1}$.

H4: There exists some constant C_4 such that $N^{-1} \sum_{i \in U} y_i^4 \leq C_4$.

H5: There exists some constant C_5 such that at any time $\delta = 1, \ldots, t$ and for any unit i we have $(p_i^\delta)^{-1} \geq C_5$.

H6: There exists some constant $C_6 > 0$ such that $V(\tilde{Y}_0) \geq C_6 N^2 n^{-1}$, and we have $(V(\tilde{Y}_0))^{-1} \hat{V}_0^p(\tilde{Y}_0) \to_{Pr} 1$, where $\hat{V}_0^p(\tilde{Y}_0)$ is defined in equation (3.9) and $\to_{Pr}$ stands for the convergence in probability.

H7: There exists some constant $C_7 > 0$ such that $V^{nr}(\tilde{Y}_t) \geq C_7 N^2 n^{-1}$, where $V^{nr}(\tilde{Y}_t)$ is defined in equation (3.7).

The assumptions (H1), (H2) and (H3) are classical in survey sampling, see for example Cardot et al. (2013). It is assumed in (H5) that at any time, the response probabilities are bounded below; this assumption is similar to condition (R.3) in Kim and Kim (2007). It is assumed in (H6) that the variance of $\tilde{Y}_0$ does not vanish and has the usual order of magnitude $N^2 n^{-1}$. It is also assumed in (H6) that the variance estimator $\hat{V}_0^p(\tilde{Y}_0)$ is consistent for $V(\tilde{Y}_0)$; this second part of the assumption could be avoided by additional assumptions on the higher order inclusion probabilities, see for example Breidt and Opsomer (2000) and Boistard et al. (2012). It is assumed in (H7) that the variance of $\tilde{Y}_t$ due to non-response has the usual order of magnitude $N^2 n^{-1}$.

B Consistency of the expansion estimator $\tilde{Y}_t$

In this Section, we prove that under assumptions (H1)-(H5) we have

$$E(\tilde{Y}_t - Y) = 0 \tag{B.1}$$

and

$$V\left\{N^{-1}(\tilde{Y}_t - Y)\right\} = O(n^{-1}). \tag{B.2}$$

Equation (B.1) follows from the fact that inclusion probabilities and response probabilities are bounded below from 0. Therefore, we focus on (B.2). From equation (3.5), we have $V(N^{-1}\tilde{Y}_t) = V(N^{-1}\tilde{Y}_0) + V^{nr}(N^{-1}\tilde{Y}_t)$. Also, we have

$$V(N^{-1}\tilde{Y}_0) = N^{-2} \sum_{i \in U} \pi_i (1 - \pi_i) \left(\frac{y_i}{\pi_i}\right)^2 + N^{-2} \sum_{i \neq j \in U} \Delta_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{B.3}$$

It follows from Assumptions (H1)-(H4) that there exists some constant $C > 0$ such that $V(N^{-1}\tilde{Y}_0) \leq Cn^{-1}$. Also, it follows from Assumption (H1)-(H5) that there exists some constant C such that

$$V^{nr}(N^{-1}\tilde{Y}_t) = E\left[N^{-2}\sum_{\delta=1}^{t}\left\{\left(\frac{y_i}{\pi_i p_i^{1\to t}}\right)^2 p_i^{\delta}(1-p_i^{\delta})\right\}\right] \quad \text{(B.4)}$$

is lower than Cn^{-1}. This leads to (B.2).

C Consistency of the variance estimator $\hat{V}_t(\tilde{Y}_t)$

In this Section, we prove that under assumptions (H1)-(H7) we have

$$E\left[\hat{V}_t(\tilde{Y}_t) - V(\tilde{Y}_t)\right] = 0 \quad \text{(C.1)}$$

and

$$\frac{\hat{V}_t(\tilde{Y}_t)}{V(\tilde{Y}_t)} \to_{Pr} 1. \quad \text{(C.2)}$$

Equation (C.1) follows from the fact that from (H1)-(H3), the second-order inclusion probabilities are bounded below from 0 and from (H5), the response probabilities are bounded below from 0. Therefore, we focus on (C.2) for which it is sufficient to prove that

$$\frac{\hat{V}_t^p(\tilde{Y}_t)}{V^p(\tilde{Y}_t)} \to_{Pr} 1 \quad \text{and} \quad \frac{\hat{V}_t^{nr}(\tilde{Y}_t)}{V^{nr}(\tilde{Y}_t)} \to_{Pr} 1. \quad \text{(C.3)}$$

We first focus on the first part of equation (C.3). We can write

$$\begin{aligned}\hat{V}_t^p(\tilde{Y}_t) - V^p(\tilde{Y}_t) &= \sum_{\delta=1}^{t}\left\{\hat{V}_\delta^p(\tilde{Y}_\delta) - \hat{V}_{\delta-1}^p(\tilde{Y}_{\delta-1})\right\} \quad \text{(C.4)}\\ &+ \left\{\hat{V}_0^p(\tilde{Y}_0) - V^p(\tilde{Y}_t)\right\}.\end{aligned}$$

From assumption (H6), we obtain

$$\frac{\hat{V}_0^p(\tilde{Y}_0) - V^p(\tilde{Y}_t)}{V^p(\tilde{Y}_t)} \to_{Pr} 0. \quad \text{(C.5)}$$

Also, we have

$$E\left[\left\{\sum_{\delta=1}^{t}\hat{V}_\delta^p(\tilde{Y}_\delta) - \hat{V}_{\delta-1}^p(\tilde{Y}_{\delta-1})\right\}^2\right] = E\sum_{\delta=1}^{t} V\left[\hat{V}_\delta^p(\tilde{Y}_\delta)\,|s_{\delta-1}\right]. \quad \text{(C.6)}$$

After some algebra, we obtain that

$$
\begin{aligned}
V\left[\hat{V}_\delta^p(\tilde{Y}_\delta)\,|s_{\delta-1}\right] &= \sum_{i\in s_{\delta-1}} \frac{p_i^\delta(1-p_i^\delta)}{(p_i^{1\to\delta})^2}(1-\pi_i)^2\left(\frac{y_i}{\pi_i}\right)^4 \\
&+ 4\sum_{i\neq j\in s_{\delta-1}} \frac{p_i^\delta(1-p_i^\delta)}{(p_i^{1\to\delta})^2}(1-\pi_i)\left(\frac{y_i}{\pi_i}\right)^3 \times \frac{p_j^\delta}{p_j^{1\to\delta}}\frac{\Delta_{ij}}{\pi_{ij}}\frac{y_j}{\pi_j} \\
&+ 4\sum_{i\neq j\neq k\in s_{\delta-1}} \frac{p_i^\delta(1-p_i^\delta)}{(p_i^{1\to\delta})^2}\left(\frac{y_i}{\pi_i}\right)^2 \times \frac{p_j^\delta}{p_j^{1\to\delta}}\frac{\Delta_{ij}}{\pi_{ij}}\frac{y_j}{\pi_j} \times \frac{p_k^\delta}{p_k^{1\to\delta}}\frac{\Delta_{ik}}{\pi_{ik}}\frac{y_k}{\pi_k} \qquad \text{(C.7)}
\end{aligned}
$$

Under assumptions (H1)-(H5), we obtain from (C.6) and (C.7) that there exists some constant C such that

$$
E\left[\left\{\sum_{\delta=1}^{t}\hat{V}_\delta^p(\tilde{Y}_\delta)-\hat{V}_{\delta-1}^p(\tilde{Y}_{\delta-1})\right\}^2\right] \leq CN^2n^{-1}. \qquad \text{(C.8)}
$$

From assumption (H6), it follows that

$$
\frac{\sum_{\delta=1}^{t}\{\hat{V}_\delta^p(\tilde{Y}_\delta)-\hat{V}_{\delta-1}^p(\tilde{Y}_{\delta-1})\}}{V^p(\tilde{Y}_t)} \to_{Pr} 0, \qquad \text{(C.9)}
$$

which, along with (C.5), leads to the first part of (C.3).

We now consider the second part of (C.3). We have

$$
E\left[\hat{V}_t^{nr}(\tilde{Y}_t)-V^{nr}(\tilde{Y}_t)\right]^2 = V\left[\sum_{i\in s_t}\frac{1-p_i^{1\to t}}{(p_i^{1\to t})^2}\left(\frac{y_i}{\pi_i}\right)^2\right]. \qquad \text{(C.10)}
$$

From Assumptions (H1), (H2), (H4) and (H5), we may find some constant C such that

$$
E\left[\hat{V}_t^{nr}(\tilde{Y}_t)-V^{nr}(\tilde{Y}_t)\right]^2 \leq CN^5n^{-4}, \qquad \text{(C.11)}
$$

which, along with Assumption (H7), leads to the second part of (C.3). This completes the proof.