SURVEY ESTIMATION UNDER INFORMATIVE NON-RESPONSE WITH FOLLOW-UP

SEPPO LAAKSONEN, RAY CHAMBERS

ABSTRACT

This paper deals with survey estimation when there is partial follow-up of sample non-response. Two different approaches that make use of the follow-up data are presented, the first based on weighting and the other on prediction, with appropriate variance estimators developed for each case. A simulation evaluation of these approaches using synthetic data and informative non-response is then used to contrast them with a basic weighting approach that does not take advantage of the follow-up survey. Our results indicate that the new approaches lead to significant improvement as far as estimation of the population total is concerned.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M05/02

University
of Southampton

Survey Estimation Under Informative Non-Response with Follow-up

Seppo Laaksonen, University of Helsinki, Finland

and

Ray Chambers, University of Southampton, United Kingdom

13 Sept 2004

Abstract

The paper deals with survey estimation when there is partial follow-up of sample non-response. Two different approaches that make use of the follow-up data are presented, the first based on weighting and the other on prediction, with appropriate variance estimators developed for each case. A simulation evaluation of these approaches using synthetic data and informative non-response is then used to contrast them with a basic weighting approach that does not take advantage of the follow-up survey. Our results indicate that the new approaches lead to significant improvement as far as estimation of the population total is concerned.

Key words: Imputation, Sample Weighting, Prediction Approach, Response Propensity Modelling

1. Introduction

In this paper we develop and evaluate estimation methods for sample surveys with follow-up. Our methods are applicable under follow-up of both the respondents and the non-respondents to the initial survey. However, for simplicity we will focus on the latter case in this paper. This situation is not uncommon in practice and has the potential to become widespread, due to increasing non-response rates for surveys in many countries. High non-response rates can have a dramatic impact on survey estimates, especially when auxiliary data that "explain" the non-response are not available for post-survey adjustments. In such situations a follow-up sub-sample of the non-respondents with a short questionnaire consisting of a few key survey questions can be helpful. At present, however, such additional data are typically used to check survey data quality, and are not exploited in estimation.

Innovation surveys are an important practical application of this follow-up approach. These surveys aim to collect information on uptake and/or development of new technology by businesses. However, they often have high non-response. It is not unreasonable in this situation to argue that it is the businesses that are not innovative, and hence see no value in the information being collected, who are less likely to respond to the survey. Consequently, following-up a sub-sample of these non-respondents with a short questionnaire containing a few key questions along the lines of ‘Is your business innovative, or has your business invested to innovative activities?’ can be a useful exercise. If the survey is based on a personal interview, it is possible, using the information collected in such a follow-up exercise, to clarify basic survey concepts. If a followed-up business is not innovative no further questions are asked, while if it is innovative, some further key questions are asked in order to assess the extent of the innovativeness of the business. Because of this structured approach, the non-response rate to the follow-up survey is usually very low. In what follows we therefore assume full response to this follow-up survey.

The data collected in this exercise can be represented by the layout in Table 1. Here X is an auxiliary variable or group of variables, known for the entire population (in our empirical example X corresponds to size-band), while Y denotes the variable(s) used to determine innovativeness status (these are collected from a sub-sample of the survey non-respondents via a follow-up survey); I_1 is an initial sample inclusion indicator; R is an initial sample response indicator, and I_2 is a sub-sample inclusion indicator (by definition, all initial respondents provide information allowing their value for Y to be computed, and so have their value of I_2 set to 1). Finally, we define R^* to be the response indicator restricted to those units with $I_2 = 1$. Note that '*obs*' means that values are observed while '*mis*' denotes non-observed values.

We assume that the initial sampling method is probability-based, with inclusion probabilities that depend only on X, and so is non-informative given X. Similarly we assume that the subsequent sub-sampling method is also probability-based, with inclusion probabilities that depend only on R and X, and so is non-informative given R and X in the sense that $pr(I_2 = 1 \mid R = 1) = 1$, while $pr(I_2 = 1 \mid R = 0)$ depends only on X. In the spirit of the discussion above, we do not assume the initial sample response R is non-informative given X. However, we assume that it is non-informative given X and Y.

Table 1: Data structure for a survey with partial follow-up of non-respondents

X	Y	I_2	R^*	R	I_1
obs	*obs*	= 1	= 1	= 1	= 1
obs	*obs*	= 1	= 0	= 0	
obs	*mis*	= 0	*mis*		
obs	*mis*				= 0

Our aim is to develop estimation methods for the population total of Y (plus the associated variances of these estimators) that fully exploit these observed data. To make the exposition straightforward, the development and empirical results presented below assume Y is a zero-one variable (e.g. corresponding to whether a business is innovative or not). However, our basic approach is quite general. In particular we consider two methods - weighting and prediction - of using the information described in Table 1 for estimating the population total of a survey variable.

2. The Weighting Approach

Compensating for sample survey non-response by re-weighting the sample respondents is a well-established approach. The basic idea is an application of response propensity modelling and has been discussed by Little (1986) among others. Ekholm and Laaksonen (1991) is an early application of this approach in the sample survey context. This paper develops this approach, extending it to the partial follow-up situation described in the previous section.

Ekholm and Laaksonen (1991) carried out respondent re-weighting at adjustment cell level. In this paper we follow Laaksonen (1999) in implementing the method at individual respondent level. There are two variants of this approach that we now describe. In both, the probability of non-response is explicitly modelled as a function of the survey variable Y. Since this value is only directly observed for the initial respondents and followed-up non-respondents (i.e. where $I_2 = 1$ in Table 1), the first variant estimates the probability of response by fitting a logistic regression model to the observed R^* values in Table 1, using both the auxiliary variable X and the survey variable Y as explanatory variables in this model. We denote the resulting fitted value of the probability of response (actually the probability that $R^* = 1$) by $\hat{\theta}^*(X,Y)$ below. This leads to the following re-weighted estimator for the population total of Y

$$\hat{T}^* = \sum_{A_i=1} Y_i \left[\pi_i \hat{\theta}^*(X_i, Y_i) \right]^{-1} \tag{1}$$

where π_i denotes the inclusion probability of population unit i and A_i denotes the indicator function for the respondents in the initial sample ($I_{1i} = 1$, $R_i = 1$). Note that there is nothing unique about the use of the logistic link in (1). In the simulation study reported in section 4 we also investigated the probit and complementary log-log with very similar results. Furthermore, unlike the situation faced by Ekholm and Laaksonen (1991) where there was little variation in the sample weights, these weights varied considerably in the business survey application we consider in this paper. Consequently the logistic model underlying the probability of response was fitted using the sample weights of the units contributing to the fit. The necessity for this weighting is made clear in section 4 where we also present results when the response model is estimated without weights. Since the model fitting process is restricted to respondents and followed-up non-respondents, these weights are scaled to sum to the total of the sample weights within each stratum prior to estimation of model parameters. Similarly, the adjusted weights derived from these model-based response probabilities that are used in (1) are also scaled to sum to the population size within each stratum.

The second variant attempts to model the actual response variable of interest (R) by imputing values for the unobserved Y values associated with the initial non-respondents who were not followed-up. Details of the imputation method used are set out in the next section. Treating these imputed values of Y as actual values, the probability of response is again modelled by fitting a (weighted) logistic regression model to the observed R-values on the entire sample, using both the auxiliary variable X and the survey variable Y as explanatory variables. We denote the resulting fitted value of the probability of response by $\hat{\theta}_{imp}(X,Y)$ with the corresponding re-weighted estimator of the population total of a survey variable Y given by

$$\hat{T}_{imp} = \sum\nolimits_{A_i=1} Y_i \left[\pi_i \hat{\theta}_{imp}(X_i, Y_i) \right]^{-1} . \tag{2}$$

Note that the adjusted weights used in (2) are re-scaled in the same way as in (1) prior to their use. Estimated sampling variances of these estimators can obtained using the approach described in Ekholm and Laaksonen (1991). In the case of stratified sampling this leads to an estimated variance of the form

$$\hat{V}(\hat{T}) = \sum\nolimits_a m_a \left\{ s_a^2(Y\pi^{-1}\hat{\theta}^{-1}) + \left(1 - \frac{m_a}{n_a} \right) \overline{(Y\pi^{-1}\hat{\theta}^{-1})_a}^{\,2} \right\} \tag{3}$$

where $\hat{\theta}$ can be either $\hat{\theta}^*(X,Y)$ or $\hat{\theta}_{imp}(X,Y)$ above and $\bar{U}_a$, $s_a^2(U)$ denote the mean and variance of the values of the variable *U*. When using (3) to estimate the variance of the estimator (1), n_a denotes the number of units that responded either in the initial survey or in the follow-up survey (I_2 = 1) in stratum *a* and m_a denotes the number of "non-missing" units in stratum *a* (i.e. those with R^* = 1), whereas when using (3) to estimate the variance of (2), n_a denotes the number of units initially selected in sample (I_2 = 1 or = 0) in stratum *a* and m_a denotes the number of units with I_2 = 1 in stratum *a*. It should be noted that the first term of (3) is a standard sampling variance, whereas the second shows the impact of the missingness on the variance of the estimator.

3. The Prediction Approach

The basic idea here is simple and is derived from the model-based approach to survey estimation. See Valliant, Dorfman and Royall (2000). However, its application to partial non-response follow-up is new, and so we develop it in more detail below. As in the previous section, we consider estimation of the population total *T* of the variable *Y*. Note that the minimum mean squared error (MMSE) predictor of this population total is its conditional expectation given the observed data.

Consequently we estimate it by approximating this optimal predictor. Assuming unit level independence, this MMSE predictor can be written

$$\tilde{T} = \sum_{A_i=1} Y_i + \sum_{B_i=1} Y_i + \sum_{C_i=1} E(Y_i \mid X_i, C_i = 1) + \sum_{I_{1i}=0} E(Y_i \mid X_i, I_{1i} = 0) \qquad (4)$$

where A_i was defined earlier as the indicator function for the respondents in the initial sample (I_{1i} = 1, R_i = 1), B_i is the indicator function for the followed-up non-respondents (I_{1i} = 1, R_i = 0, I_{2i} = 1) and C_i is the indicator function for the non-respondents who were not followed-up (I_{1i} = 1, R_i = 0, I_{2i} = 0).

In section 1 we assumed probability-based methods depending only on the population values of X are used to select both the initial sample and the follow-up sample. It is easy to see that then

$$E(Y_i \mid X_i, I_{1i} = 1, R_i = 0, I_{2i} = 0) = E(Y_i \mid X_i, I_{1i} = 1, R_i = 0, I_{2i} = 1) \qquad (5)$$

so the third term in the above MMSE predictor (4) can be approximated by the fitted regression of Y on X for the followed-up non-respondents. A similar approach can be used to approximate the fourth term of (4). In this case we can show that

$$\begin{aligned} E(Y_i \mid X_i, I_{1i} = 0) = E(Y_i \mid X_i, I_{1i} = 1, R_i = 0)\big(1 - pr(R_i = 1 \mid X_i, I_{1i} = 1)\big) \\ + E(Y_i \mid X_i, I_{1i} = 1, R_i = 1) pr(R_i = 1 \mid X_i, I_{1i} = 1). \end{aligned}$$

It is clear that we can estimate $E(Y_i \mid X_i, I_{1i} = 1, R_i = 1)$ from the initial respondents data. Denote this estimate by $\hat{\mu}_{1i}$. Similarly, we can estimate $E(Y_i \mid X_i, I_{1i} = 1, R_i = 0)$ from the followed-up non-respondents data. Denote this estimate by $\hat{\mu}_{0i}$. Let $\hat{\theta}(X_i, Y_i)$ denote an estimate of the response probability $pr(R_i = 1 \mid X_i, Y_i, I_{1i} = 1)$. An estimate $\hat{\theta}(X_i)$ of $pr(R_i = 1 \mid X_i, I_{1i} = 1)$ can then be obtained via a weighted average of the $\hat{\theta}(X_i, Y_i)$. For example, when X is continuous a nonparametric kernel estimate of this conditional probability is

$$\hat{\theta}(X_i) = \sum_{I_{1j}=1} K\left(b^{-1}(X_j - X_i)\right)\hat{\theta}(X_j, Y_j) \Big/ \sum_{I_{1j}=1} K\left(b^{-1}(X_j - X_i)\right)$$

where K denotes a suitable kernel (i.e. density) function and b is a bandwidth parameter. For discrete X a corresponding non-parametric approach leads us to replace K by the indicator $I(X_j = X_i)$. That is, $\hat{\theta}(X_i)$ is then just the average of $\hat{\theta}(X, Y)$ for those sample units with $X = X_i$,

$$\hat{\theta}(X_i) = \sum_{I_{1j}=1} I(X_j = X_i)\hat{\theta}(X_j, Y_j) \Big/ \sum_{I_{1j}=1} I(X_j = X_i).$$

Using (5), we can write down a "plug-in" estimator for T, based on $\tilde{T}$ (see (4)) as

$$\hat{T} = \sum_{A_i=1} Y_i + \sum_{B_i=1} Y_i + \sum_{C_i=1} \hat{\mu}_{0i} + \sum_{I_{1i}=0}\left(\hat{\mu}_{1i}\hat{\theta}(X_i) + \hat{\mu}_{0i}(1 - \hat{\theta}(X_i))\right). \qquad (6)$$

The problem therefore is one of determining $\hat{\theta}(X_i, Y_i)$. Since we do not have values of Y_1 for non-responding units that are not followed-up, this is not straightforward. We investigate an easy to implement but computer intensive method of doing this, based on regression hot deck imputation. The steps in this process are

1. Impute the missing value Y_i for a not followed-up non-responding unit (i.e. C_i = 1). In the case where Y is continuous, this is by $\hat{\mu}_{0i} + \varepsilon^*_{0i}$, where ε^*_{0i} is a random draw from the follow-up sub-sample residuals $\left\{Y_j - \hat{\mu}_{0j}; I_{2j} = 1, R_j = 0\right\}$. When Y is categorical, this is by a

random draw from the follow-up sub-sample units with the same X value as the unit being imputed - i.e. from $\{Y_{j;}\ X_j = X_i,\ I_{2j} = 1,\ R_j = 0,\ I_{1j} = 1\}$.

2. Using these imputed values of Y, apply logistic regression (or some other similar technique) to get estimates $\hat{\theta}(X_i, Y_i)$ for all the sampled units. Use these estimates to compute the values of $\hat{\theta}(X_i)$ for the non-sampled units.

3. Compute the "plug-in" estimator $\hat{T}$ using (6).

Estimation of the prediction mean squared error for (6) is not straightforward under this imputation approach. We therefore apply this technique below to the case where both Y and X are categorical and show how variance estimates can be computed when $\theta(X_i,Y_i)$ is estimated via imputation.

3.1 Imputation-Based Approach with Categorical Data

As noted above we assume both X and Y are categorical. In particular, we use $X_i = a$ to denote that the i^{th} population unit belongs to category a of X, and assume Y is a zero-one variable (e.g. denoting whether a business is not innovative/innovative respectively). We assume simple random sampling for both the initial and follow-up sample within each level of X.

A simple approach that makes minimal assumptions is to assume a saturated model for the $Y \times X \times R$ cross-classification. In this case we can use (5) to write down simple unbiased estimates for the various population parameters in (6). Define

$m_{yx} =$ # responding sample units ($I_1 = 1, R = 1$) with $X = x$ and $Y = y$

$k_{1yx} =$ # followed-up non-responding sample units ($I_1 = 1, R = 0, I_2 = 1$) with $X = x$ and $Y = y$

$k_{0yx} =$ # not followed-up non-responding sample units ($I_1 = 1, R = 0, I_2 = 0$) with $X = x$ and $Y = y$

$k_{1x} =$ # followed-up non-respondents with $X = x$

k_{0x} = # not followed-up non-respondents with $X = x$

m_x = # responding sample units with $X = x$

n_x = # selected sample units with $X = x$

In practise, k_{0yx} will not be known. We shall assume however that this value is available from the imputed values of Y for the not followed-up non-respondents. We denote this imputed value by k^*_{0yx} below. Then

$$\hat{\mu}_{1a} = \frac{m_{1a}}{m_a} \text{ = proportion of respondents with } Y = 1 \text{ and } X = \text{a}$$

$$\hat{\mu}_{0a} = \frac{k_{11a} + k^*_{01a}}{n_a - m_a} \text{ = proportion of non-respondents with } Y = 1 \text{ and } X = \text{a}$$

$$\hat{\theta}(a,1) = \frac{m_{1a}}{m_{1a} + k_{11a} + k^*_{01a}} \text{ = respondent proportion of units with } Y = 1 \text{ units and } X = \text{a}$$

$$\hat{\theta}(a,0) = \frac{m_{0a}}{m_{0a} + k_{10a} + k^*_{00a}} \text{ = respondent proportion of units with } Y = 0 \text{ units and } X = \text{a}$$

and so our estimator of

$$\begin{aligned} \theta_a &= pr(R = 1 \mid X = a, I_1 = 1) \\ &= pr(R = 1 \mid X = a, Y = 1, I_1 = 1) pr(Y = 1 \mid X = a, I_1 = 1) \\ &+ pr(R = 1 \mid X = a, Y = 0, I_1 = 1) pr(Y = 0 \mid X = a, I_1 = 1) \end{aligned}$$

is just the initial non-response rate for sample units with $X = a$,

$$\hat{\theta}_a = \hat{\theta}(a,1)\left(\frac{m_{1a} + k_{11a} + k^*_{01a}}{n_a}\right) + \hat{\theta}(a,0)\left(\frac{m_{0a} + k_{10a} + k^*_{00a}}{n_a}\right) = \frac{m_a}{n_a}.$$

The estimator (6) can then be written

$$\hat{T} = \sum\nolimits_a \left\{ m_a \hat{\mu}_{1a} + (n_a - m_a)\hat{\mu}_{0a} + (N_a - n_a)\left[\hat{\theta}_a \hat{\mu}_{1a} + (1-\hat{\theta}_a)\hat{\mu}_{0a} \right] \right\}. \qquad (7)$$

In order to estimate the prediction mean squared error $Var(\hat{T} - T)$ of (7) under the saturated model assumption, we use a sequence of iterated expectation arguments, first conditioning on the initial and follow-up sample data (thus obtaining the variability caused by the imputation process), then conditioning on the initial sample data (obtaining the variability due to the follow-up sampling process), and finally recovering the variability due to the initial sampling process. To start, we note that

$$Var(\hat{T} - T) = \sum\nolimits_a \mathrm{var}\left\{ k_{01a}^* - k_{01a} + (N_a - n_a)\left[\hat{\theta}_a \hat{\mu}_{1a} + (1-\hat{\theta}_a)\hat{\mu}_{0a} \right] - \sum_{\substack{I_{1i}=0 \\ X_i = a}} Y_i \right\}$$

$$= \sum\nolimits_a \left\{ \begin{array}{c} E\left[V^*\left(k_{01a}^* + (N_a - n_a)(1-\hat{\theta}_a)\hat{\mu}_{0a} \right) \right] + \\ \mathrm{var}\left[E^*(k_{01a}^*) - k_{01a} + (N_a - n_a)\left[\hat{\theta}_a \hat{\mu}_{1a} + (1-\hat{\theta}_a) E^*(\hat{\mu}_{0a}) \right] - \sum\limits_{\substack{I_{1i}=0 \\ X_i = a}} Y_i \right] \end{array} \right\}$$

where E^* and V^* denote expectation and variance with respect to the imputation process. In order to evaluate the above expression we observe that

$$k_{01a}^* = \sum\nolimits_{F_a} \Delta_i Y_i$$

where F_a denotes the followed-up non-responding sample units with $X = a$ and Δ_i is the number of times unit i is selected as a donor. Hence

$$E^*(k^*_{01a}) = k_{0a}k_{11a} / k_{1a}$$
$$V^*(k^*_{01a}) = k_{0a}k_{11a}k_{10a} / k^2_{1a}$$

and so

$$Var(\hat{T}-T) = \sum\nolimits_a \left\{ \begin{array}{c} E\left[k_{0a}\left(\dfrac{k_{11a}k_{10a}}{k^2_{1a}}\right)\left(1+\dfrac{(N_a-n_a)(1-\hat{\theta}_a)}{n_a-m_a}\right)^2 \right] \\ +Var\left[\begin{array}{c} \dfrac{k_{11a}k_{0a}}{k_{1a}} - k_{01a} \\ +(N_a-n_a)\left[\hat{\theta}_a\hat{\mu}_{1a} + (1-\hat{\theta}_a)\dfrac{k_{11a}(1+k_{0a}/k_{1a})}{n_a-m_a}\right] - \sum\nolimits_{I_{1i}=0,X_i=a} Y_i \end{array} \right] \end{array} \right\}$$

To proceed further, we note that the use of simple random sampling within each category of X implies that the number of successes in the respondent, non-respondent follow-up and non-respondent non-follow-up groups are mutually independent given the respective sizes of these groups, with k_{11a} distributed as binomial(k_{1a}, μ_{0a}), k_{01a} distributed as binomial(n_a–m_a–k_{1a}, μ_{0a}) and m_{1a} distributed as binomial(m_a, μ_{1a}). Hence, after some simplification we obtain

$$Var(\hat{T}-T) = \sum\nolimits_a \left\{ \begin{array}{c} E\left[k_{0a}\left(\dfrac{k_{11a}k_{10a}}{k^2_{1a}}\right)\left(\dfrac{N_a}{n_a}\right)^2 \right] \\ +\left(\dfrac{N_a-n_a}{n_a}\right)^2 (\mu_{1a}-\mu_{0a})^2 n_a\theta_a(1-\theta_a) \\ +E\left[\begin{array}{c} \mu_{0a}(1-\mu_{0a})\left[\dfrac{1}{k_{1a}}\left(k_{0a} + \left(\dfrac{N_a-n_a}{n_a}\right)(n_a-m_a)\right)^2 + k_{0a}\right] \\ +\left(\dfrac{N_a-n_a}{n_a}\right)^2 m_a\mu_{1a}(1-\mu_{1a}) \end{array} \right] \\ +(N_a-n_a)\big(\mu_{1a}\theta_a + \mu_{0a}(1-\theta_a)\big)\big(1-\mu_{1a}\theta_a - \mu_{0a}(1-\theta_a)\big) \end{array} \right\}. \quad (8)$$

An obvious "plug-in" estimator $\hat{V}$ of (8) then follows, where we replace unknown parameters in expression by their estimates, and expectations of random variables are replaced by realised values. That is

$$\hat{V} = \sum\nolimits_a \left\{ \begin{array}{c} k_{0a}\left(\dfrac{k_{11a}k_{10a}}{k_{1a}^2}\right)\left(\dfrac{N_a}{n_a}\right)^2 + \left(\dfrac{N_a - n_a}{n_a}\right)^2 \left(\hat{\mu}_{1a} - \hat{\mu}_{0a}\right)^2 n_a \hat{\theta}_a (1 - \hat{\theta}_a) \\ + \hat{\mu}_{0a}(1 - \hat{\mu}_{0a}) \left[\dfrac{1}{k_{1a}} \left(k_{0a} + \left(\dfrac{N_a - n_a}{n_a}\right)(n_a - m_a) \right)^2 + k_{0a} \right] \\ + \left(\dfrac{N_a - n_a}{n_a}\right)^2 m_a \hat{\mu}_{1a}(1 - \hat{\mu}_{1a}) \\ + (N_a - n_a)\left(\hat{\mu}_{1a}\hat{\theta}_a + \hat{\mu}_{0a}(1 - \hat{\theta}_a)\right)\left(1 - \hat{\mu}_{1a}\hat{\theta}_a - \hat{\mu}_{0a}(1 - \hat{\theta}_a)\right) \end{array} \right\}. \qquad (9)$$

3.2 Prediction Based on a Non-Saturated Model

In this case we apply logistic regression techniques to the sample data to fit an unsaturated model to $\theta(X_i,Y_i)$, again treating the imputed Y-values of the not followed-up non-respondents as "real" data. Let $\hat{\theta}(X_i,Y_i)$ denote the fitted values generated by this model. One estimator of θ_a is then

$$\hat{\theta}_a = \hat{\theta}(a,1)\left(\frac{m_{1a} + k_{11a} + k_{01a}^*}{n_a}\right) + \hat{\theta}(a,0)\left(\frac{m_{0a} + k_{10a} + k_{00a}^*}{n_a}\right). \qquad (10)$$

Note that (10) estimates $pr(Y = 1 \mid X = a, I_1 = 1)$ by the sample proportion of units with $X = a$ that also have $Y = 1$. However, a more sophisticated approach could easily be used here as well, modelling Y in terms of X. From the definition of θ_a, we see that

$$\theta_a = \frac{\theta(a,1)\mu_{0a} + \theta(a,0)(1 - \mu_{0a})}{[1 - \{\theta(a,1) - \theta(a,0)\}(\mu_{1a} - \mu_{0a})]}. \qquad (11)$$

An alternative to (10) is therefore to substitute the logistic model-based estimates $\hat{\theta}(a,1)$ and $\hat{\theta}(a,0)$, together with $\hat{\mu}_{1a} = \frac{m_{1a}}{m_a}$ and $\hat{\mu}_{0a} = \frac{k_{11a} + k^*_{01a}}{n_a - m_a}$, in (11).

Regardless of whether (10) or (11) forms the basis for estimation of θ_a, the final estimator of T is then given by (7).

3.3 Using Multiple Imputations

Suppose that we independently repeat the imputation process L times to define a "multiple imputation" estimator

$$\bar{T} = L^{-1}\sum_{l=1}^{L}\hat{T}_l \, . \tag{12}$$

Here $\hat{T}_l$ denotes the value of (7) based on the l^{th} set of imputed values. The average value (12) should then be more efficient than a single imputation value of (7). In order to estimate the prediction mean squared error of (12) we note that

$$Var(\bar{T} - T) = Var\left(L^{-1}\sum_{l=1}^{L}(\hat{T}_l - T) \right) = L^{-2}\sum_{l=1}^{L}\sum_{j=1}^{L} Cov\left(\hat{T}_l - T, \hat{T}_j - T\right)$$

where

$$\begin{aligned} Cov\left(\hat{T}_l - T, \hat{T}_j - T\right) &= E\left(\mathrm{cov}^*(\hat{T}_l - T, \hat{T}_j - T)\right) + Cov\left(E^*(\hat{T}_l - T), E^*(\hat{T}_j - T)\right) \\ &= Var\left(E^*(\hat{T} - T)\right). \end{aligned}$$

It follows $Var(\bar{T} - T) = Var\left(E^*(\hat{T} - T)\right)$ and so from (8) we obtain

$$Var(\bar{T}-T)=\sum\nolimits_a \left\{ \begin{array}{c} \left(\dfrac{N_a-n_a}{n_a}\right)^2 \left(\mu_{1a}-\mu_{0a}\right)^2 n_a\theta_a(1-\theta_a) \\ +E\left[\begin{array}{c} \mu_{0a}(1-\mu_{0a})\left[\dfrac{1}{k_{1a}}\left(k_{0a}+\left(\dfrac{N_a-n_a}{n_a}\right)(n_a-m_a)\right)^2+k_{0a}\right] \\ +\left(\dfrac{N_a-n_a}{n_a}\right)^2 m_a\mu_{1a}(1-\mu_{1a}) \end{array} \right] \\ +(N_a-n_a)\left(\mu_{1a}\theta_a+\mu_{0a}(1-\theta_a)\right)\left(1-\mu_{1a}\theta_a-\mu_{0a}(1-\theta_a)\right) \end{array} \right\}. \qquad (13)$$

Again, we see that an estimate of (13) is easily defined by substituting estimates for unknown parameters and replacing expectations by realised values. This leads to the prediction mean squared error estimator

$$\bar{V}=\sum\nolimits_a \left\{ \begin{array}{c} \left(\dfrac{N_a-n_a}{n_a}\right)^2 \left(\hat{\mu}_{1a}-\hat{\mu}_{0a}\right)^2 n_a\hat{\theta}_a(1-\hat{\theta}_a) \\ +\hat{\mu}_{0a}(1-\hat{\mu}_{0a})\left[\dfrac{1}{k_{1a}}\left(k_{0a}+\left(\dfrac{N_a-n_a}{n_a}\right)(n_a-m_a)\right)^2+k_{0a}\right] \\ +\left(\dfrac{N_a-n_a}{n_a}\right)^2 m_a\hat{\mu}_{1a}(1-\hat{\mu}_{1a}) \\ +(N_a-n_a)\left(\hat{\mu}_{1a}\hat{\theta}_a+\hat{\mu}_{0a}(1-\hat{\theta}_a)\right)\left(1-\hat{\mu}_{1a}\hat{\theta}_a-\hat{\mu}_{0a}(1-\hat{\theta}_a)\right) \end{array} \right\}. \qquad (14)$$

4. Empirical Results

The population data underpinning our simulations was generated from data collected in an innovation survey carried out in Finland in the 1990s. The population size was 4453 businesses, and *Y* was an indicator variable for whether a business is innovative or not. There were a total of 2474 such businesses in this population, and this number corresponds to the target population total of *Y* of interest. In each simulation a stratified random sample of size 1200 was selected from this population. The strata corresponded to size-bands based on the number of employees of each business. Random non-response was generated using a threshold model defined in terms of another variable "value added", which is strongly associated with innovation, as well as other variables correlated with the size of the business. There were an average of 800 respondents per sample, and since the non-response was informative, innovative businesses ($Y = 1$) were more likely to respond. For each sample of initial non-respondents, a sub-sample of 150 was followed up and values of Y obtained. There was no non-response associated with the follow-up sub-sample. Note that this is realistic in practice since the follow-up survey can be conducted by face-to-face interview and the questionnaire is typically reduced to a minimum.

A total of 200 independent simulations were carried out and values for various estimates of the population total of Y and associated estimates of variance were calculated. In addition to the “standard” estimator that ignores the non-response, we computed estimates based on the methods described in this paper. These estimates were as follows:

Weighting Approach

Estimators were defined by either (1) or (2), with estimated variance computed using (3) in both cases. Note that (2) was defined using a single imputation. We used two different response propensity models with these estimators. Model (A) corresponded to a logistic specification with main effects for size-band and value of Y, while Model (B) was the same as (A) but also included a size-band by Y interaction term.

Prediction Approach

A single imputation estimator based on (7) with variance estimator defined by (9) was computed. In addition, a multiple imputation estimator ($L = 8$) defined by (12) with variance estimator defined by (14) was also computed.

Table 2 shows the results from the 200 simulations. Here Mean denotes the average value of an estimator, MSE denotes the average of the squared difference between an estimator value and the true value of T (2474), Average(V) denotes the average of the corresponding variance estimator and 95% CI Coverage denotes the percentage of resulting confidence intervals that included the true value. All confidence intervals were generated as the estimate value plus or minus twice the square root of its estimated variance. All averaging is over the 200 simulations.

Table 2: Simulation Results. Each estimation strategy is identified by the equation number of the estimator + equation number of the corresponding variance estimator. In addition, for the weighting methods considered in the simulation, the specification includes the type of logistic model (A or B) used and whether the fit was weighted or not.

Estimation Strategy	**Mean** (T = 2474)	**MSE**	**Average(V)**	**95% CI Coverage**
Assuming that nonresponse is ignorable	2823.5	126494	5933	0
(1) + (3), unweighted Model A	2663.5	39676	8947	49.0
(1) + (3), weighted Model A	2469.9	5141	5686	95.5
(1) + (3), weighted Model B	2475.5	4942	5699	96.0
(2) + (3), weighted Model A	2474.9	5606	5998	95.5
(2) + (3), weighted Model B	2477.4	5611	5998	95.5
(7) + (9)	2477.3	5615	5970	95.5
(12) + (14)	2477.3	4718	5291	95.5

The first row in Table 2 clearly shows the bias associated with the unadjusted estimator. All other strategies considered in the table give better estimates than this one. The comparison between the use of unweighted and weighted logistic propensity modelling (second and third rows of the table) is also interesting, since the importance of weighting is very clear. As previously noted, there has been very little discussion of whether or not one should use sampling weights in response propensity modelling. From a design-based perspective the sample inclusion probability for unit i in (1) should be multiplied by pr(unit i is a respondent | unit i is in sample) to get the final probability of inclusion in the respondent sample. However, there are two ways we can define this conditional probability:

(a) When it corresponds to pr(R=1|Y,X) for a randomly chosen unit from the population. In this case it makes sense to weight when fitting the response propensity model since it is a model for the whole population.

(b) When it corresponds to pr(R=1|Y,X) for a randomly chosen unit from the selected sample. In this case weighting the response propensity fit is not appropriate.

Our interpretation accords with (a), and so we recommend weighting when carrying out response propensity modelling.

Comparing weighting method (1) with weighting method (2), we see that the former is preferable. However, there is little to choose between the different weighting estimators when we compare choice of propensity model, with estimators based on unsaturated Model A performing very similarly to those based on the saturated Model B. Note the methods that used Model B tended to give slightly higher estimates than those based on Model A. This lead to better estimates in the case of weighting method (1), but no essential difference in the case of weighting method (2). This may be interpreted as indicating that model B is slightly better fitting than model A, and can be generalised to indicate that it is important when using weighting methods to construct as well-fitting a response propensity model as possible.

For the prediction estimators (7) and (12), it is clear that using multiple imputations provides a substantial benefit. In fact, in MSE terms, the multiple imputations estimator (12) performed best of all estimators considered in our study. Also, it is interesting that the mean values of the last three methods in Table 2 are almost equal. This is because these methods assume the same (saturated) model B for the response. Note that we also investigated the behaviour of the prediction approach based on the non-saturated model A using the ideas described in Section 3.3. However, we saw very little change and so do not report these results.

Not surprisingly, the variance estimation methods investigated in the study show more variability than corresponding estimates of totals. In particular, it is interesting to see that the highly biased estimation method that ignored the non-response (row 1 in Table 2) gave very similar variance estimates to the much better performing methods that allowed for the non-response, leading to confidence intervals with substantial under-coverage. In contrast, the variance estimators (weighting and prediction based) that properly took account of this non-response (rows 3 to 8 in Table 2) tended to be somewhat conservative, with all achieving nominal coverage levels. In doing so, however, it should be noted that the variance estimates defined by (3) tended to be more positively correlated with the corresponding estimation errors than those defined by (9) or (14). Overall, however, our results indicate that a user will not be led astray by using these variance estimators.

5. Conclusion

In this paper we contrast two approaches to making use of partial follow-up information to adjust for non-ignorable non-response in survey estimation. The first approach is based on weighting by an estimate of the response propensity while the second uses the follow-up information to directly predict the population total of interest. Our simulation results show that, properly applied, both approaches are similar in performance and so the choice between them is matter of personal preference.

Note that we do not consider the case of survey variables in the main survey that are not measured in the follow up study. Both the weighting and prediction approaches can be extended to handle this situation, with the latter then depending on the conditional distribution of the not followed up *Y* variables given the values of the followed up *Y* variables. This remains a topic for further research, as does implementation of the prediction approach without recourse to imputation, which is technically possible but not explored in this paper.

Finally, we observe that both the weighting and prediction approached can be easily extended to multiple auxiliary variables. In practice, this should lead to better fitting response propensity models and hence better estimates.

References

Ekholm, A. and Laaksonen, S. (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. *Journal of Official Statistics,* 325-337.

Laaksonen, S. (1999). Weighting and Auxiliary Variables in Sample Surveys. In: G. Brossier and A-M. Dussaix (eds.). Enquêtes et Sondages. Méthodes, Modèles, Applications, Nouvelles Approches, 168-180. Paris: Dunod.

Little, R. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review* 54, 139-157.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). Finite Population Sampling and Inference. New York: John Wiley