Journal of Official Statistics, Vol. 22, No. 1, 2006, pp. 81–95

Survey Estimation Under Informative Nonresponse with Follow-up

Seppo Laaksonen[1] *and Ray Chambers*[2]

This article deals with survey estimation when there is partial follow-up of sample nonresponse. 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 using synthetic data and informative nonresponse is then used to compare these two approaches, as well as to contrast them with a simpler weighting approach that ignores the information obtained by the follow-up survey and treats the nonresponse as missing at random. 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 article we develop and evaluate estimation methods for sample surveys with follow-up. This situation is not uncommon in practice and has the potential to become widespread, owing to increasing nonresponse rates for surveys in many countries. A widely used strategy for dealing with this problem is to reweight the respondents' data to account for their different response propensities. However, this depends on the availability of auxiliary information that "explains" the nonresponse, and implicitly assumes that the nonrespondents' data are missing at random (MAR) given this auxiliary information. Clearly, the MAR assumption cannot be tested just using the respondents' data. In situations where the MAR assumption seems unjustified, therefore, one option is to carry out a follow-up of a subsample of the nonrespondents, with a short questionnaire consisting of a few key survey questions. This is not unlike the strategy of multiple callbacks used in some social surveys in order to improve the overall response rate (Groves 1989; Elliott et al. 2000). Here, however, the aim is to collect data from a sample of the nonrespondents that can be used either to build a "better" overall model for response propensity (e.g., by incorporating information obtained from the sampled nonrespondents) or to allow a more sophisticated method of estimation that uses the data obtained from both respondents and sampled nonrespondents. Unfortunately, however our experience is that in many cases where

[1] University of Helsinki, Department of Statistics, Helsinki, Finland. Email: seppo.laaksonen@helsinki.fi
[2] University of Wollongong, Centre for Statistical and Survey Methodology, Wollongong NSW 2522, Australia. Email: ray@uow.edu.au

Acknowledgment: The authors would like to thank the Associate Editor and the referees for their valuable suggestions, which lead to a considerable improvement of this article.

follow-up samples are taken, the additional data collected are used to check the quality of the information obtained for core survey data items and are not exploited in estimation.

Innovation surveys are an important practical application of this follow-up approach. These surveys are designed to collect information on the uptake and/or development of new technology by businesses. However, they often have high nonresponse (e.g., the nonresponse rate in the EU 1998–2001 innovation survey was more than 40% in many countries; see Eurostat 2004, p. 287). 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, that are less likely to respond to the survey. Consequently, following-up a subsample of these nonrespondents with a short questionnaire containing a few key questions along the lines of "Is your business innovative, or has your business invested in 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. In particular, this can facilitate a respondent's understanding of what the survey is about, and hence increase the probability of a response. 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. Because of this structured approach, the nonresponse rate for the follow-up survey is usually very low. In what follows we therefore assume full response to this follow-up survey. This seems realistic in business surveys where small enterprises are the main source of nonresponse. In household surveys, however, this assumption may not be realistic, and our approach would then need to be extended to allow for this extra source of nonresponse.

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, while Y is a target variable for the survey (in our empirical example X corresponds to size-band and Y corresponds to the variable(s) used to determine innovativeness status); I_1 is an initial sample inclusion indicator; R_1 is an initial sample response indicator, and I_2 is a subsample inclusion indicator (by definition, all initial respondents have their value of I_2 automatically set to 1, i.e., if $R_1 = 1$, then $I_2 = 1$). Finally, we define R_2 to be the response indicator restricted to those units with $I_2 = 1$. It immediately follows that if $R_1 = 1$ then $R_2 = 1$. Note that '*obs*' means that Y-values are observed while '*mis*' denotes nonobserved values. By definition, Y-values are only observed for units with $R_2 = 1$. However, some of these units will have $R_1 = 0$.

We assume that the initial sampling method is probability-based, with inclusion probabilities that depend only on X, and so is noninformative given X. Similarly we assume that the subsequent subsampling method is also probability-based, with inclusion probabilities that depend only on R_1 and X, and so is noninformative given R_1 and X.

Table 1. Data structure for a survey with partial follow-up of nonrespondents

X	I_1	R_1	I_2	R_2	Y
obs	$= 1$	$= 1$	$= 1$	$= 1$	*obs*
obs		$= 0$	$= 1$	$= 1$	*obs*
obs			$= 0$	*mis*	*mis*
obs	$= 0$	*mis*	*mis*	*mis*	*mis*

This follows since $pr(I_2 = 1|R_1 = 1) = 1$, while $pr(I_2 = 1|R_1 = 0)$ depends only on X. In the spirit of the discussion above, we do *not* assume that the initial sample response R_1 is noninformative given X. However, we assume that it is noninformative given *X and Y*.

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 are based on the assumption that Y is a one–zero 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 nonresponse by reweighting 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, and in a more general framework using a two-phase sampling approach proposed by Särndal and Swensson (1987). Ekholm and Laaksonen (1991) made an early application of this approach in the sample survey context. This article develops this approach, extending it to the partial follow-up situation described in the previous section.

Ekholm and Laaksonen (1991) carried out respondent reweighting at the adjustment cell level. In this article we follow Laaksonen (1999) in implementing the method at the individual respondent level. There are two variants of this approach that we now describe. In both, the probability of nonresponse 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 nonrespondents (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 indicator variable R_1 based on the data from those units with $R_2 = 1$ in Table 1, using both the auxiliary variable X and the survey variable Y as explanatory variables in this model. We denote the fitted value under this model by $\hat{\theta}_A(X, Y)$ below. This fitted value is assumed to be an estimate of the probability $\theta(X, Y)$ that $R_1 = 1$ given $I_1 = 1$ and leads to the reweighted estimator for the population total of Y

$$\hat{T}_A = \sum_{i \in s_1} Y_i[\pi_i \hat{\theta}_A(X_i, Y_i)]^{-1} \tag{1}$$

where π_i denotes the inclusion probability of population unit i and s_1 denotes the set of respondents in the initial sample, i.e., the collection of units with ($I_{1i} = 1, R_{1i} = 1$). 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 article. Consequently the logistic model for $\theta(X, Y)$ was fitted using the sample weights π_i^{-1} of the units with $R_2 = 1$ that contributed 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 nonrespondents, 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 constitutes an attempt to model the actual response variable of interest (R_1) by imputing values for the unobserved Y values associated with the initial nonrespondents who were not followed up (i.e., those with $I_2 = 0$ in Table 1). 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 initial response $\theta(X, Y)$ is again modelled by the (weighted) logistic regression of the observed R_1-values for the entire sample on both the auxiliary variable X and the survey variable Y. We denote the resulting fitted value of the probability of response by $\hat{\theta}_B(X, Y)$, with the corresponding reweighted estimator of the population total of a survey variable Y given by

$$\hat{T}_B = \sum_{i \in s_1} Y_i [\pi_i \hat{\theta}_B (X_i, Y_i)]^{-1} \tag{2}$$

Note that the adjusted weights used in (2) are rescaled in the same way as in (1) prior to their use. Estimated sampling variances of these estimators can be obtained using the approach described by Ekholm and Laaksonen (1991). In the case of stratified sampling this leads to an estimated variance of the form (this is a slight upward approximation as they mention, p. 333)

$$\hat{V}(\hat{T}) = \sum_h m_{1h} \left\{ V_{1h} + \left(1 - \frac{m_{1h}}{n_{1h}} \right) E_{1h}^2 \right\} \tag{3}$$

where h indexes the strata and $E_{1h} = n_{1h}^{-1} \sum_{i \in s_{1h}} Y_i \pi_i^{-1} \hat{\theta}_i^{-1}$ and $V_{1h} = (n_{1h} - 1)^{-1} \sum_{i \in s_{1h}} (Y_i \pi_i^{-1} \hat{\theta}_i^{-1} - E_{1h})^2$. Here $\hat{\theta}_i$ can be either $\hat{\theta}_A(X_i, Y_i)$ or $\hat{\theta}_B(X_i, Y_i)$. When using (3) to estimate the variance of (1), n_{1h} denotes the number of units that responded either in the initial survey or in the follow-up survey ($R_2 = 1$) in Stratum h and m_{1h} denotes the number of initial respondents in Stratum h (i.e., those with $R_1 = 1$), whereas when using (3) to estimate the variance of (2), n_{1h} denotes the number of units initially selected in Sample ($I_1 = 1$) in Stratum h and m_{1h} denotes the number of units with $I_2 = 1$ in Stratum h. It should be noted that the first term of (3) is a standard sampling variance, whereas the second shows the effect 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 nonresponse 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. Since this "best" predictor will depend on unknown parameters, we approximate it by replacing these parameters by suitable sample-based estimates, leading to what is sometimes referred to as the "Empirical Best" (EB) predictor.

Assuming that distinct population units have independent Y-values, the MMSE predictor can be written

$$\tilde{T} = \sum_{i:A_i=1} Y_i + \sum_{i:B_i=1} Y_i + \sum_{i:C_i=1} E(Y_i|X_i) + \sum_{i:I_{1i}=0} E(Y_i|X_i) \tag{4}$$

where A_i is the indicator function for the respondents in the initial sample ($I_{1i} = 1, R_{1i} = 1$), B_i is the indicator function for the followed-up nonrespondents ($I_{1i} = 1, R_{1i} = 0, I_{2i} = 1$) and C_i is the indicator function for the nonrespondents who were not followed up ($I_{1i} = 1$, $R_{1i} = 0, I_{2i} = 0$).

In Section 1 we assumed that 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

$$E(Y_i|X_i, I_{1i} = 1, R_{1i} = 0, I_{2i} = 0) = E(Y_i|X_i, I_{1i} = 1, R_{1i} = 0, I_{2i} = 1) \tag{5}$$

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

$$\begin{aligned} E(Y_i|X_i, I_{1i} = 0) = {} & E(Y_i|X_i, I_{1i} = 1, R_{1i} = 0)(1 - pr(R_{1i} = 1|X_i, I_{1i} = 1)) \\ & + E(Y_i|X_i, I_{1i} = 1, R_{1i} = 1)pr(R_{1i} = 1|X_i, I_{1i} = 1) \end{aligned}$$

It is clear that we can estimate $E(Y_i|X_i, I_{1i} = 1, R_{1i} = 1)$ from the initial respondents' data. Denote this estimate by $\hat{\mu}_{1i}$. Similarly, we can estimate $E(Y_i|X_i, I_{1i} = 1, R_{1i} = 0)$ from the followed-up nonrespondents' data. Denote this estimate by $\hat{\mu}_{0i}$. Suppose now that we can also construct an estimate $\hat{\theta}(X_i, Y_i)$ of the response probability $pr(R_{1i} = 1|X_i, Y_i, I_{1i} = 1)$. An estimate $\hat{\theta}(X_i)$ of $pr(R_{1i} = 1|X_i, I_{1i} = 1)$ can then be calculated as a suitably weighted average of the $\hat{\theta}(X_i, Y_i)$ values generated by the initial sample. For example, if X is discrete we can define $\hat{\theta}(X_i)$ to be the average of $\hat{\theta}(X, Y)$ for those initial sample units with $X = X_i$:

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

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

$$\hat{T}_C = \sum_{i:A_i=1} Y_i + \sum_{i:B_i=1} Y_i + \sum_{i:C_i=1} \hat{\mu}_{0i} + \sum_{i:I_{1i}=0} (\hat{\mu}_{1i}\hat{\theta}(X_i) + \hat{\mu}_{0i}(1 - \hat{\theta}(X_i))) \tag{6}$$

The problem therefore is one of determining $\hat{\theta}(X_i, Y_i)$. Since we do not have values of Y for nonresponding 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 imputation. The steps in this process are

- Impute the missing value Y_i of a not-followed-up nonresponding unit (i.e., one with $C_i = 1$). In the case where Y is continuous, this could be by $\hat{\mu}_{0i} + \varepsilon^*_{0i}$, where ε^*_{0i} is a random draw from the follow-up subsample residuals $\{Y_j - \hat{\mu}_{0j}; I_{2j} = 1, R_{1j} = 0\}$. When Y is categorical, this is by a random draw from follow-up subsample units with the same X value as the unit being imputed – i.e., from $\{Y_j; X_j = X_i, I_{2j} = 1, R_{1j} = 0\}$.

- Using these imputed values of Y, calculate 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 nonsampled units.
- 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 $\hat{\theta}(X_i, Y_i)$ is based on simple moment-type estimation.

3.1. Imputation-based Approach with Categorical Data

As noted above, we assume that both X and Y are categorical. In particular, we use $X_i = a$ to denote that the *ith* population unit belongs to category a of X, and take Y to be a zero–one variable (e.g., denoting whether a business is not innovative/innovative, respectively). We assume that the population is stratified on the levels of X and that the initial sample is randomly selected from these strata. We also assume that the follow-up subsample is obtained by randomly selecting units from the initial sample nonrespondents within each stratum.

In order to apply a model-based approach, we need to specify a model for the joint population distribution of Y, X and R. A simple approach that makes minimal assumptions is to assume a saturated model for the $Y \times X \times R$ population cross-classification. In this case we can use (5) to write down simple unbiased moment-type estimates for the parameters of this model. Define

m_{yx} = # responding sample units ($I_1 = 1, R_1 = 1$) with $X = x$ and $Y = y$
k_{1yx} = # followed-up nonresponding sample units ($I_1 = 1, R_1 = 0, I_2 = 1$) with $X = x$ and $Y = y$
k_{0yx} = # not followed-up nonresponding sample units ($I_1 = 1, R_1 = 0, I_2 = 0$) with $X = x$ and $Y = y$
k_{1x} = # followed-up nonrespondents with $X = x$
k_{0x} = # not followed-up nonrespondents with $X = x$
m_x = # responding sample units with $X = x$
n_x = # selected sample units with $X = x$

In practice, 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 nonrespondents. We denote this imputed value by k^*_{0yx} below. Then

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

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

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

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

and so our estimator of

$$
\begin{aligned}
\theta_a &= pr(R_1 = 1|X = a, I_1 = 1) \\
&= pr(R_1 = 1|X = a, Y = 1, I_1 = 1)pr(Y = 1|X = a, I_1 = 1) \\
&\quad + pr(R_1 = 1|X = a, Y = 0, I_1 = 1)pr(Y = 0|X = a, I_1 = 1)
\end{aligned}
$$

is just the initial nonresponse 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 therefore be written

$$\hat{T}_C = \sum_a \{m_a\hat{\mu}_{1a} + (n_a - m_a)\hat{\mu}_{0a} + (N_a - n_a)[\hat{\theta}_a\hat{\mu}_{1a} + (1 - \hat{\theta}_a)\hat{\mu}_{0a}]\} \tag{7}$$

In order to estimate the prediction mean squared error $Var(\hat{T}_C - T)$ 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

$$
\begin{aligned}
Var(\hat{T}_C - T) &= \sum_a Var\left\{k^*_{01a} - k_{01a} + (N_a - n_a)[\hat{\theta}_a\hat{\mu}_{1a} + (1 - \hat{\theta}_a)\hat{\mu}_{0a}] - \sum_{i:\{I_{1i}=0, X_i=a\}} Y_i\right\} \\
&= \sum_a \Bigg\{E\left[Var^*\left(k^*_{01a} + (N_a - n_a)(1 - \hat{\theta}_a)\hat{\mu}_{0a}\right)\right] + Var\Bigg[E^*(k^*_{01a}) - k_{01a} \\
&\qquad + (N_a - n_a)[\hat{\theta}_a\hat{\mu}_{1a} + (1 - \hat{\theta}_a)E^*(\hat{\mu}_{0a})] - \sum_{i:\{I_{1i}=0, X_i=a\}} Y_i\Bigg]\Bigg\}
\end{aligned}
$$

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

$$k^*_{01a} = \sum_{i \in F_a} \Delta_i Y_i$$

where F_a denotes the followed-up nonresponding sample units with $X = a$ and Δ_i is the number of times unit i is selected as a donor. Hence $E^*(k^*_{01a}) = k_{0a}k_{11a}/k_{1a}$ and $Var^*(k^*_{01a}) = k_{0a}k_{11a}k_{10a}/k^2_{1a}$, so

$$Var(\hat{T}_C - T) = \sum_a (E_a + V_a)$$

where

$$E_a = E\left[k_{0a}\left(\frac{k_{11a}k_{10a}}{k_{1a}^2}\right)\left(1+\frac{(N_a-n_a)(1-\hat{\theta}_a)}{n_a-m_a}\right)^2\right]$$

$$V_a = Var\left[\frac{k_{11a}k_{0a}}{k_{1a}} - k_{01a} + (N_a-n_a)\left\{\hat{\theta}_a\hat{\mu}_{1a} + (1-\hat{\theta}_a)\frac{k_{11a}(1+k_{0a}/k_{1a})}{n_a-m_a}\right\} - \sum_{i:\{I_{1i}=0,X_i=a\}} Y_i\right]$$

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, nonrespondent follow-up and nonrespondent 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

$$Var(\hat{T}_C - T) = \sum_a \{M_{1a} + M_{2a} + M_{3a} + M_{4a}\} \tag{8}$$

where

$$M_{1a} = E\left[k_{0a}\left(\frac{k_{11a}k_{10a}}{k_{1a}^2}\right)\left(\frac{N_a}{n_a}\right)^2\right]$$

$$M_{2a} = \left(\frac{N_a-n_a}{n_a}\right)^2(\mu_{1a}-\mu_{0a})^2 n_a\theta_a(1-\theta_a)$$

$$M_{3a} = E\left[\mu_{0a}(1-\mu_{0a})\left\{\frac{1}{k_{1a}}\left(k_{0a} + \left(\frac{N_a-n_a}{n_a}\right)(n_a-m_a)\right)^2 + k_{0a}\right\}\right.$$
$$\left. + \left(\frac{N_a-n_a}{n_a}\right)^2 m_a\mu_{1a}(1-\mu_{1a})\right]$$

$$M_{4a} = (N_a-n_a)(\mu_{1a}\theta_a + \mu_{0a}(1-\theta_a))(1-\mu_{1a}\theta_a - \mu_{0a}(1-\theta_a))$$

An obvious "plug-in" estimator of (8) then follows, where we replace unknown parameters in the expression by their estimates, and expectations of random variables are replaced by realised values. This is

$$\hat{V}(\hat{T}_C) = \sum_a \{\hat{M}_{1a} + \hat{M}_{2a} + \hat{M}_{3a} + \hat{M}_{4a}\} \tag{9}$$

where

$$\hat{M}_{1a} = k_{0a}\left(\frac{k_{11a}k_{10a}}{k_{1a}^2}\right)\left(\frac{N_a}{n_a}\right)^2$$

$$\hat{M}_{2a} = \left(\frac{N_a - n_a}{n_a}\right)^2(\hat{\mu}_{1a} - \hat{\mu}_{0a})^2 n_a\hat{\theta}_a(1 - \hat{\theta}_a)$$

$$\hat{M}_{3a} = \hat{\mu}_{0a}(1 - \hat{\mu}_{0a})\left[\frac{1}{k_{1a}}\left(k_{0a} + \left(\frac{N_a - n_a}{n_a}\right)(n_a - m_a)\right)^2 + k_{0a}\right] + \left(\frac{N_a - n_a}{n_a}\right)^2 m_a\hat{\mu}_{1a}(1 - \hat{\mu}_{1a})$$

$$\hat{M}_{4a} = (N_a - n_a)(\hat{\mu}_{1a}\hat{\theta}_a + \hat{\mu}_{0a}(1 - \hat{\theta}_a))(1 - \hat{\mu}_{1a}\hat{\theta}_a - \hat{\mu}_{0a}(1 - \hat{\theta}_a))$$

3.2. *Prediction Based on a Nonsaturated Model*

The saturated model assumed in Section 3.1 will typically be over-specified, and so we can expect that parameter estimation will not be fully efficient. For small sample sizes this may be of some concern. In such cases we can 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 nonrespondents as "real" data. Let $\hat{\theta}_L(X_i, Y_i)$ denote the fitted values generated by this model. One estimator of θ_a is then

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

Note that (10) estimates $pr(Y = 1|X = a, I_1 = 1)$ 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

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

An alternative to (10) is therefore to substitute the logistic model-based estimates $\hat{\theta}_L(a,1)$ and $\hat{\theta}_L(a,0)$, together with $\hat{\mu}_{1a} = (m_{1a})/(m_a)$ and $\hat{\mu}_{0a} = (k_{11a} + k^*_{01a})/(n_a - m_a)$, into (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). Variance estimation for this nonsaturated model-based version of (7) is complex and will depend on the actual specification of the model. In the empirical results reported in the next section, we therefore adopt a conservative variance estimation strategy, replacing $\hat{\theta}_a$ by (10) in the saturated model-based variance estimator (9).

3.3. Using Multiple Imputations

Clearly we can independently repeat the imputation process L times to define a "multiple imputations" estimator

$$\bar{T}_C = L^{-1}\sum_{l=1}^{L}\hat{T}_C(l) \tag{12}$$

Here $\hat{T}_C(l)$ denotes the value of (7) based on the lth 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

$$Var(\bar{T}_C - T) = L^{-2}\left[\sum_{l=1}^{L} Var(\hat{T}_C(l) - T) + \sum_{l=1}^{L}\sum_{\substack{j=1 \\ j\neq l}}^{L} Cov(\hat{T}_C(l) - T, \hat{T}_C(j) - T)\right] \tag{13}$$

As in Section 3.1, we use a "star" superscript to denote moments with respect to the distribution induced by the simulation process. Independence of the repeated imputations then implies that for $l \neq j$, $Cov^*(\hat{T}_C(l) - T, \hat{T}_C(j) - T) = 0$. Thus for $l \neq j$

$$\begin{aligned} Cov(\hat{T}_C(l) - T, \hat{T}_C(j) - T) &= E(Cov^*(\hat{T}_C(l) - T, \hat{T}_C(j) - T)) \\ &\quad + Cov(E^*(\hat{T}_C(l) - T), E^*(\hat{T}_C(j) - T)) = Var(E^*(\hat{T}_C - T)) \end{aligned}$$

since $E^*(\hat{T}_C(l) - T) = E^*(\hat{T}_C(j) - T)$ and $E^*(\hat{T}_C(l) - T) = E^*(\hat{T}_C - T)$. Similarly

$$Var(\hat{T}_C(l) - T) = E(Var^*(\hat{T}_C(l) - T)) + Var(E^*(\hat{T}_C - T))$$

Substituting these expressions into (13) and simplifying implies that $Var(\bar{T}_C - T) = Var(E^*(\hat{T}_C - T)) + L^{-1}E(Var^*(\hat{T}_C - T))$. From (8) we obtain

$$Var(\bar{T}_C - T) = \sum_a \{L^{-1}M_{1a} + M_{2a} + M_{3a} + M_{4a}\} \tag{14}$$

where the components M_{1a}, M_{2a}, M_{3a} and M_{4a} are defined following (8). An estimate of (14) is easily defined by substituting estimates for unknown parameters and replacing expectations by realised values. This leads to the prediction mean squared error estimator

$$\hat{V}(\bar{T}_C) = \sum_a \{L^{-1}\hat{M}_{1a} + \hat{M}_{2a} + \hat{M}_{3a} + \hat{M}_{4a}\} \tag{15}$$

where again the components $\hat{M}_{1a}$, $\hat{M}_{2a}$, $\hat{M}_{3a}$ and $\hat{M}_{4a}$ on the right-hand side of (15) are defined following (9). It should be noted that (15) is not the same as the usual multiple imputation variance estimator, since it is based on a plug-in estimate of the actual prediction variance of $\bar{T}_C$. Also, it is clear that the larger the value of L, the smaller the actual prediction variance (14) as well as its estimate (15).

4. Empirical Results

The population data underpinning our simulations were generated from data collected in an innovation survey carried out in Finland in the 1990s. The population size was 4,453 businesses, and Y was an indicator variable that identified whether a business is innovative $(Y = 1)$ or not $(Y = 0)$. There were a total of $T = 2,474$ such businesses in this population. In each simulation a stratified random sample of size 1,200 was selected from this population (note that in this type of survey the sampling fractions tend to be rather high; see Eurostat 2004, p. 287). Table 2 shows the strata used in the sample design, defined by size-bands based on the number of employees of each business.

Random nonresponse 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 nonresponse was informative, innovative businesses $(Y = 1)$ were more likely to respond. For each sample of initial nonrespondents, a stratified subsample of 150 was followed up and values of Y obtained. There was no nonresponse associated with the follow-up subsample.

A total of 650 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" weighted estimator that assumes ignorable nonresponse within strata, we computed estimates on the basis of the methods described in this article. These estimates were as follows:

4.1. Weighting Approach

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

Table 2. Population and sample sizes by stratum. Note that micro enterprises are excluded from the target population

Size-band (number of employees)	Population size	Sampling fraction, %
5.0–9.9	1,046	20.0
10.0–19.9	1,305	18.4
20.0–29.9	595	28.1
30.0–49.9	546	30.2
50.0–99.9	443	40.0
100.0–199.9	229	40.2
200.0–499.9	180	50.0
500.0–999.9	56	50.0
1,000.0+	53	60.4
All	4,453	26.9

4.2. Prediction Approach

The single imputation estimator (7) based on the saturated Model II and with variance estimator defined by (9) was computed. In addition, the multiple imputations estimator (12) given the same saturated model and with variance estimator defined by (15) was computed. This was based on $L = 8$ independent imputations for each missing value of Y. We also computed these estimators using the unsaturated Model I, but observed no significant difference in performance. These results are therefore omitted.

Table 3 shows the results from the 650 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, 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 squared root of its estimated variance. All averaging is over the 650 simulations. We also carried out a similar simulation exercise, but with a smaller sampling fraction (20%) and consequently with a smaller subsample size. We do not present these results since they are essentially the same as those in Table 3, the only difference being that variances are higher because of the smaller sample size.

The first row in Table 3 clearly shows that the weighted estimator based on ignorable nonresponse assumption within strata is heavily biased. All other strategies considered in the table give better estimates than this one. The comparison between the use of

Table 3. Simulation results. Each estimation strategy is identified by the equation number of the estimator + the equation number of the corresponding variance estimator. In addition, for the two weighting methods considered in the simulation, the specification includes the type of logistic model (I or II) used and whether the logistic fit was weighted or not. The figure in parentheses in the Mean and Average(V) columns is the Monte Carlo standard error of the corresponding row entry

Estimation strategy	Mean (True = 2,474)	MSE	Average(V)	95% CI Coverage
Reweighting for ignorable nonresponse within strata	2,821.9 (2.6)	12,5746	5,921 (6.8)	0.3
(1) + (3), weighting (A), Model I – unweighted logistic fit	2,661.6 (2.6)	39,415	8,917 (13.6)	48.0
(1) + (3), weighting (A), Model I	2,467.7 (2.8)	5,291	5,671 (10.7)	95.2
(1) + (3), weighting (A), Model II	2,473.8 (2.8)	5,153	5,687 (11.0)	96.3
(2) + (3), weighting (B), Model I	2,478.3 (2.9)	5,672	6,003 (12.9)	96.0
(2) + (3), weighting (B), Model II	2,480.7 (2.9)	5,623	6,000 (12.7)	96.3
(7) + (9), prediction (C) single imputation, Model II	2,480.6 (2.9)	5,623	5,933 (12.9)	96.3
(12) + (15), prediction (C) multiple imputations, Model II	2,480.7 (2.8)	5,141	5,367 (10.8)	95.4

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 probability of inclusion for unit *i* in the respondent sample is defined by the product of the sample inclusion probability for this unit and the probability that this unit is a respondent given that it is in sample. However, there are two ways we can define this conditional probability:

- When it corresponds to pr $(R_1 = 1|Y, X)$ 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.
- When it corresponds to pr $(R_1 = 1|Y, X)$ for a randomly chosen unit from the selected sample. In this case weighting the response propensity fit is not appropriate.

Our interpretation accords with the first dot point above, and so we recommend weighting when carrying out response propensity modelling for use in "model-assisted" estimation methods like weighting (A) and weighting (B).

Comparing weighting (A) with weighting (B), we see that the former is preferable. However, there is little to choose between the weighting estimators when we compare choice of propensity model, with estimators based on unsaturated Model I performing very similarly to those based on the saturated Model II. This may be interpreted as indicating that Model II fits the data only marginally better than Model I, and indicates that it is important when using weighting-based methods to construct as well-fitting a response propensity model as possible.

On the other hand, we also note that the weighting methods that used Model II tended to give slightly higher estimates than those based on Model I. This leads to better estimates in the case of weighting (A), but not in the case of weighting (B) where both estimates are slightly larger than the true value.

The prediction (C) strategy (12) based on multiple imputations under Model II performed best in terms of MSE of all estimation strategies considered in our study, although the advantage is minor when compared with weighting (A) based on Model II. However, we also note that the weighting (B) strategy and the single imputation prediction (C) strategy based on the same saturated model lead to virtually identical MSEs. This raises the possibility that a multiple imputations version of weighting (B) might also lead to significant MSE gains. Development of such an estimator (as well as an estimator of its variance) remains a topic for further research. As noted earlier, we also investigated the behaviour of the prediction approach based on the nonsaturated Model I 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 do not behave like the corresponding estimates of totals. In particular, it is interesting to see that the highly biased estimation method that assumed ignorable nonresponse within strata (Row 1 in Table 3) gave very similar variance estimates to the much better performing methods that allowed for nonignorable nonresponse, leading to confidence intervals with substantial under-coverage. In contrast, the variance estimators (weighting and prediction based) that properly took account of this nonresponse (Rows 3 to 8 in Table 3) tended to be somewhat conservative, with all achieving close to nominal coverage levels. Note that the variance estimator (3) underpinning weighting (A) and weighting (B) does not include any type of finite population correction (fpc), and so some of this conservatism may be due to this fact. Unfortunately, it is not straightforward to include an fpc in (3). In contrast, the mean squared error estimator (15) used under prediction (C) has "built-in" fpc since the expression (8) for the prediction variance is zero when the entire population is sampled and all nonrespondents are followed up. The conservative behaviour for (15) observed in Table 3 is therefore probably more a consequence of Model II being somewhat over-specified. It should also 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, our results indicate that a user will not be led astray by using these variance estimators.

5. Conclusion

In this article we contrast two approaches to making use of partial follow-up information to adjust for nonignorable nonresponse 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, the two approaches are rather similar in performance and so the choice between them is a matter of personal preference. If small design bias is a primary consideration then the weighting (A) strategy is simple to apply and returned the smallest design bias in our simulations. If, on the other hand, small mean squared error is the aim then the prediction (C) multiple imputations strategy performed well in the same simulations.

For reasons of simplicity of exposition, the development in this article has been based on a simple dichotomous specification for Y. In practice one would expect to also encounter situations where Y is polychotomous, or even continuous. The theory developed in this article can be readily extended to these situations, and we anticipate that applications based on use of either a weighting type estimation methodology or a prediction approach will eventually appear.

Another extension that we do not address in this article is 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 here.

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

6. References

Ekholm, A. and Laaksonen, S. (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. Journal of Official Statistics, 7, 325–337.

Elliot, M.R., Little, R.J.A., and Lewitzky, S. (2000). Subsampling Callbacks to Improve Survey Efficiency. Journal of the American Statistical Association, 95, 730–838.

EUROSTAT (2004). Innovation in Europe. Results for the EU, Iceland and Norway. Data 1998–2001. Theme 9: Science and Technology.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley and Sons.

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.

Särndal, C.-E. and Swensson, B. (1987). A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. International Statistical Review, 55, 279–294.

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

Received September 2004
Revised September 2005