Informative Sampling on Two Occasions: Estimation and Prediction

Abdulhakeem A.H. Eideh
Department of Mathematics
Faculty of Science and Technology
Al-Quds University, Abu-Deis Campus
Palestine
msabdul@science.alquds.edu

Abstract

The sample distribution is defined as the distribution of the sample measurements given the selected sample. Under informative sampling, this distribution may be different from the corresponding population distribution. Sampling on two occasions under informative sampling design, utilizing the sample and sample-complement distributions for occasion one, the matched sample and unmatched sample distributions, and matched sample-complement and unmatched sample-complement for occasion two, is proposed for predicting finite population total of a variable under study for the current (second) occasion, viewing information collected on the first (previous) occasion as auxiliary information. An interesting result of the present analysis is that known predictors in common use are shown to be special cases of the present predictors obtained under informative sampling, thus providing them a new justification.

Keywords: Matched distribution, Sample-complement distribution, Unmatched distribution.

1. Introduction

The practice of relying on samples for the collection of important series of data, published at regular intervals has become common. In most surveys, interest centers on the current total or average. For discussions of repeated sampling in general and on sampling on two occasions, in particular, under noninformative sampling, see Cochran (1977). However if the design is informative, in the sense that the study or response variable is correlated with design variables not included in the model, even after conditioning on the model covariates, standard estimates of the model parameters can be severely biased, leading possibly to false inference. For example, see Pfeffermann, Krieger and Rinott (1998) and Eideh (2010).

In this paper we propose to deal with the prediction of finite population total of a variable under study for the second occasion, using information collected on the first occasion as an auxiliary variable, and under informative sampling, by combining two separate statistical methodologies: sampling on two occasions and methods of analysis under complex informative sampling.

As an example of situation where the sampling design is informative on both occasions is: if the same sampling design is used on every occasion and the sampling scheme is informative on one occasion then it is informative on every occasion. (Corresponding talk with Danny Pfeffermann).

Methods of prediction under informative sampling have been investigated by Sverchkov and Pfeffermann (2004), and Pfeffermann and Sverchkov (2007) in the context of analytical inference from complex surveys for cross-sectional analysis based on data from a single occasion. Later, Eideh and Nathan (2009) investigated the effects of informative two-stage cluster sampling on estimation and prediction with application in small area estimation.

Previous work in this area deals with sampling on two occasions under equal and unequal probability of selection sampling designs. See for example Arnab (1998), and Prasad and Graham (1994). In particular, none of the above studies extract the sample matched and sample un-matched distributions, for sampling on two occasions, from the population distribution and first order inclusion probabilities. The key reference about effects of informative design on sample distributions, applied here to the case of repeated sampling, is Sverchkov and Pfeffermann (2004).

As pointed out by Sverchkov and Pfeffermann (2004) in Section 8 Concluding Remarks “Further experimentation with this kind of predictors and MSE (mean square error) estimation is therefore highly recommended”.

Thus, the aims of the present study are then to extend and develop the methods of prediction of finite population totals under informative sampling by utilizing the sample distribution and sample-complement distributions for sampling on two occasions. In Section 2 a review of known results on the sample and sample-complement probability density functions (pdfs) is given. In Section 3 we introduce the marginal distributions of matched and un-matched sample observations for occasion two. Marginal distributions of complement-matched and complement-un-matched samples are discussed in Section 4. In Section 5 we present the results of prediction of finite population totals of a variable under study for the second occasion, under informative sampling. Prediction methods are discussed in Section 6. In Section 7 we give examples. Section 8 presents the estimation of mean square error and Section 9 provides a discussion of the results.

It should be noted that this paper is based completely on model-based inference (rather than randomization based).

2. Review of results on sample and sample-complement probability density functions

In this article we assume sampling on two occasions from a finite population that is composed of the same elements at the two different occasions. The study or response variable Y is observed at each occasion, but not necessarily for the same set of elements. The response variable will be denoted Y_1 at the first occasion and Y_2 at the second occasion. At the first occasion, a sample s_1 is drawn by the sampling design $P_1(\cdot)$ such that $P_1(s_1)$ is the probability that s_1 is

chosen. The corresponding inclusion probabilities are denoted by π_{1i} for element $i \in U$. Let $I_{1i} = 1$ if $i \in s_1$ and $I_{1i} = 0$, otherwise, be sample membership indicator of element $i \in U$ for the first occasion. At the first occasion, we assume that the population values $(y_{11}, x_{11}), \cdots, (y_{1N}, x_{1N})$ are independent realizations of random variables Y_1 and X_1 with continuous joint pdf $f_p(y_{1i}, x_{1i})$. In the application, the variable Y_1 is the variable of interest for period one and is observed only for the sample on the first occasion. (In practice, the variable of interest for period one may often be observed also for the sample of the second occasion (retrospective studies)). The variable X_1 represents auxiliary variable and its values are assumed known for the whole population. Let $\mathbf{z} = \{z_1, ..., z_N\}$ be the values of a known design variable, used for the sample selection process but not included in the working model under consideration. In what follows we consider a sampling design with selection probabilities $\pi_i = \Pr(i \in s)$ where $i = 1, ..., N$. In practice, the π_i's may depend on the population values $(\mathbf{x}, \mathbf{y}, \mathbf{z})$. We consider single stage sampling, with inclusion probabilities:

$$\pi_{1i} = \Pr(i \in s_1 | y_{i1}, x_{1i}, z_i) = g(y_{1i}, x_{1i}, z_i) \tag{2.1}$$

for some function g and all units $i \in U$. See Eideh (2010) for further discussion on examples of g.

Since $\pi_1, ..., \pi_N$ are defined by the realizations $(y_{1i}, x_{1i}, z_i), i = 1, ..., N$, therefore they are random realizations defined on the space of possible populations.

According to Pfeffermann, Krieger and Rinott (1998), the conditional marginal sample pdf of Y_{1i} is defined as:

$$\begin{aligned} f_{s_1}(y_{1i} | x_{1i}) &= f_p(y_{1i} | x_{1i}, I_{1i} = 1) \\ &= \frac{\Pr(I_{1i} = 1 | y_{1i}, x_{1i}) f_p(y_{1i} | x_{1i})}{\Pr(I_{1i} = 1 | x_{1i})} \end{aligned} \tag{2.2}$$

with the second equality obtained by application of Bayes theorem.

Note that the conditional marginal sample pdf is different from the super population pdf generating the finite population values, unless $\Pr(I_{1i} = 1 | y_{1i}, x_{1i}) = \Pr(I_{1i} = 1 | x_{1i})$ for all possible values y_{1i}, in which case the sampling process or scheme is noninformative or can be ignored conditional on x_{1i}.

Denote by E_p and E_s the expectations under the population and sample pdfs, respectively. Then according to Pfeffermann, Krieger and Rinott (1998), (2.2) can be written as:

$$f_{s_1}(y_{1i} \mid x_{1i}) = \frac{E_p(\pi_{1i} \mid y_{1i}, x_{1i}) f_p(y_{1i} \mid x_{1i})}{E_p(\pi_{1i} \mid x_{1i})} \tag{2.3}$$

where

$$E_p(\pi_{1i} \mid x_{1i}) = \int E_p(\pi_{1i} \mid y_{1i}, x_{1i}) f_p(y_{1i} \mid x_{1i}) dy_{1i} \tag{2.4}$$

It follows from (2.3) that the population and sample pdf's are different, unless $E_p(\pi_{1i}|y_{1i},x_{1i}) = E_p(\pi_{1i}|x_{1i})$ for all y_{1i}, in which case the sampling process can be ignored for inferences that condition on the x_1.

Comment 1. Note that $E_p(\pi_{1i} \mid y_{1i}, x_{1i}) = E_{z_i|y_{1i},x_{1i}} E_p(\pi_i \mid y_{1i}, x_{1i}, z_i)$, so that z_i is integrated out in (2.3). See Remark 1 in Sverchkov and Pfeffermann (2004) for further discussion.

Let $w_{1i} = 1/\pi_{1i}$ define the sampling weight of unit $i \in U$. According to Pfeffermann and Sverchkov (1999), the following relationships hold:

$$E_p(\pi_{1i}|y_{1i},x_{1i}) = \frac{1}{E_{s_1}(w_{1i}|y_{1i},x_{1i})} \tag{2.5a}$$

$$E_p(\pi_{1i}|x_{1i}) = \frac{1}{E_{s_1}(w_{1i}|x_{1i})} \tag{2.5b}$$

$$E_p(y_{1i}|x_{1i}) = \frac{E_{s_1}(w_{1i}y_{1i}|x_{1i})}{E_{s_1}(w_{1i}|x_{1i})} \tag{2.5c}$$

$$E_p(y_{1i}) = \frac{E_{s_1}(w_{1i}y_{1i})}{E_{s_1}(w_{1i})} \tag{2.5d}$$

$$E_p(\pi_{1i}) = \frac{1}{E_{s_1}(w_{1i})} \tag{2.5e}$$

Similar to (2.2), the conditional marginal sample-complement pdf (for units not in s_1, denoted by s_1^c) is defined as:

$$\begin{aligned} f_{s_1^c}(y_{1i}|x_{1i}) &= f_p(y_{1i}|x_{1i}, I_{1i} = 0) \\ &= \frac{\Pr(I_{1i} = 0|y_{1i},x_{1i}) f_p(y_{1i}|x_{1i})}{\Pr(I_{1i} = 0|x_{1i})} \end{aligned} \tag{2.6a}$$

It follows from Sverchkov and Pfeffermann (2004) that this pdf can be written as:

$$\begin{aligned} f_{s_1^c}(y_{1i} \mid \mathrm{x}_{1\mathrm{i}}) &= \frac{E_p(1-\pi_{1i} \mid y_{1i}, x_i) f_p(y_{1i} \mid x_{1i})}{E_p(1-\pi_{1i} \mid x_i)} \\ &= \frac{E_{s_1}((w_{1i}-1) \mid y_{1i}, x_{1i}) f_{s_1}(y_{1i} \mid x_{1i})}{E_{s_1}((w_{1i}-1) \mid x_{1i})} \end{aligned} \tag{2.6b}$$

Now, using (2.6a), (2.5b) and (2.5c), we have:

$$
\begin{aligned}
E_{s_1^c}\left(y_{1i}|x_{1i}\right) &= \int y_{1i} f_{s_1^c}\left(y_{1i} \mid x_{1i}\right) dy_{1i} \\
&= \int y_{1i} \frac{E_p\left((1-\pi_{1i}) \mid y_{1i}, x_i\right) f_p\left(y_{1i} \mid x_{1i}\right)}{E_p\left(1-\pi_{1i} \mid x_i\right)} dy_{1i} \\
&= \int \frac{E_p\left((1-\pi_{1i}) y_{1i} \mid y_{1i}, x_i\right) f_p\left(y_{1i} \mid x_{1i}\right)}{E_p\left(1-\pi_{1i} \mid x_{1i}\right)} dy_{1i} \\
&= \frac{E_p E_p\left((1-\pi_{1i}) y_{1i} \mid y_{1i}, x_i\right)}{E_p\left(1-\pi_{1i} \mid x_{1i}\right)} = \frac{E_p\left((1-\pi_{1i}) y_{1i} | x_{1i}\right)}{E_p\left((1-\pi_{1i}) | x_{1i}\right)} \\
&= \frac{E_{s_1}\left((w_{1i}-1) y_{1i} | x_{1i}\right)}{E_{s_1}\left((w_{1i}-1) | x_{1i}\right)} = E_{s_1}\left(\frac{(w_{1i}-1) y_{1i}}{E_{s_1}\left((w_{1i}-1) | x_{1i}\right)} | x_{1i}\right)
\end{aligned}
\tag{2.7}
$$

where $E_{s_1^c}$ denotes the expectation under the sample-complement pdf.

Using (2.3), (2.5a), (2.5b) and (2.6), we have the following results:

$$
\frac{f_{s_1}\left(y_{1i}|x_{1i}\right)}{f_p\left(y_{1i}|x_{1i}\right)} = \frac{E_{s_1}\left(w_{1i}|x_{1i}\right)}{E_{s_1}\left(w_{1i}|y_{1i},x_{1i}\right)} \tag{2.8}
$$

$$
\frac{f_{s_1}\left(y_{1i}|x_{1i}\right)}{f_{s_1^c}\left(y_{1i}|x_{1i}\right)} = \frac{E_{s_1}\left((w_{1i}-1)|x_{1i}\right)}{E_{s_1}\left((w_{1i}-1)|y_{1i},x_{1i}\right)} \tag{2.9}
$$

and

$$
\frac{f_{s_1^c}\left(y_{1i}|x_{1i}\right)}{f_p\left(y_{1i}|x_{1i}\right)} = \frac{E_{s_1}\left(w_{1i}|x_{1i}\right) E_{s_1}\left((w_{1i}-1)|y_{1i},x_{1i}\right)}{E_{s_1}\left(w_{1i}|y_{1i},x_{1i}\right) E_{s_1}\left((w_{1i}-1)|x_{1i}\right)} \tag{2.10}
$$

According to (2.8), the sample and population pdfs are different unless $E_{s_1}\left(w_{1i} \mid y_{1i}, x_{1i}\right) = E_{s_1}\left(w_{1i} \mid x_{1i}\right)$ for all y_{1i}, in which case the sampling process can be ignored for inference that conditions on the x_1.

The key references to the relationships between the population and sample distributions and their applications are the articles by Eideh and Nathan (2006, 2009), Eideh (2007, 2008, 2009), Krieger and Pfeffermann (1997), Pfeffermann, Krieger and Rinott (1998), Pfeffermann and Sverchkov (1999, 2003, 2007), Skinner (1994), and Sverchkov and Pfeffermann (2004).

3. Marginal distributions of matched and unmatched sample observations for occasion two

To the sample s_1, of size n, drawn at the first occasion corresponds a complementary sample $s_1^c = U - s_1$. The complementary sample is not surveyed at the first occasion, but we need the probabilities of inclusion in the

complementary sample induced by the design $P_1(\cdot)$. When sampling on the second occasion we have more information than at the first occasion. For every $i \in s_1$, we know the values $(y_{1i}, x_{1_i}), i = 1,...,n$ and x_{1_i}, $i = n+1,...,N$. For the second sample, we can consider situations of no overlap, complete overlap, or partial overlap with the first sample. Cochran (1977) considers, under noninformative sampling, the optimal designs for the estimation of different parameters at the second occasion. It is intuitively clear that there are cases in which the information from the first occasion may be used to improve the current estimates. Hence we opt for dealing with the situation of partial overlap, for which the other situations can be considered as special cases.

At the second occasion, two independent samples are drawn, a matched sample and an unmatched sample. The matched sample, s_{2m}, of size n_m, is drawn from s_1 by the design $P_m(\cdot \mid s_1)$ such that $P_m(s_{2m} \mid s_1)$ is the conditional probability of choosing s_{2m} on the second occasion, given that s_1 was selected on the first occasion. The inclusion probabilities under this design are denoted $\pi_{2i}^m = \Pr(i \in s_{2m} \mid i \in s_1)$ for elements $i \in s_1$. The unmatched sample, s_{2u}, of size $n_u = n - n_m$, is drawn from $s_1^c = U - s_1$ according to the design $P_u(\cdot \mid s_1^c)$ such that $P_u(s_{2u} \mid s_1^c)$ is the conditional probability of choosing s_{2u}, given the complementary sample s_1^c. The inclusion probabilities under this design are denoted $\pi_{2i}^u = \Pr(i \in s_{2u} \mid i \in s_1^c)$. Note that s_{2m} and s_{2u} are disjoint and are chosen independently. The total sample at the second occasion is $s_2 = s_{2m} \bigcup s_{2u}$. Let $\underline{y}_i = (y_{1i}, y_{2i}), \underline{x}_i = (x_{1i}, x_{2i})$ are the values of Y and X for unit i for the two occasions. The variable Y_{2i} is observed for all elements in the second sample. We assume that inclusion probabilities may depend on the values of Y_{1i}, Y_{2i}, X_{1i} and X_{2i} for the same unit:

$$\pi_{2i}^m = \Pr(i \in s_{2m} | i \in s_1, \underline{y}_i, \underline{x}_i) = g_{2m}(\underline{y}_i, \underline{x}_i) \tag{3.1}$$

for some function g_{2m} and all elements $i \in s_1$, and

$$\pi_{2i}^u = \Pr(i \in s_{2u} \mid \mathrm{i} \in \mathrm{s}_1^{\mathrm{c}}, \underline{y}_i, \underline{x}_{\mathrm{i}}) = g_{2u}(\underline{y}_i, \underline{x}_i) \tag{3.2}$$

for some function g_{2u} and all units $i \in s_1^c$.

Let $I_{2i}^m = 1$ if $i \in s_{2m}$ and $I_{2i}^m = 0$ if $i \in s_1 - s_{2m}$. Also, let $I_{2i}^u = 1$ if $i \in s_{2u}$ and $I_{2i}^u = 0$ if $i \in s_1^c - s_{2u}$. To find the marginal distribution of the matched sample observations, we treat the first sample as if it were a population. Assume that Y_{2i} denotes the value of a response variable Y_2, associated with unit i that belongs to the new 'population' $s_1 = \{1,...,n\}$. If Y_{2i} depends on Y_{1i} and $\underline{x}_i$, then the conditional sample pdf of Y_{2i} is defined analogously to (2.2) and (2.3) as:

$$f_{s_{2m}}(y_{2i} \mid y_{1i}, \underline{x}_i) = \frac{\Pr(i \in s_{2m} \mid \underline{y}_i, \underline{x}_i) f_{s_1}(y_{2i} \mid y_{1i}, \underline{x}_i)}{\Pr(i \in s_{2m} \mid y_{1i}, \underline{x}_i)} \tag{3.3}$$

which can be written as:

$$f_{s_{2m}}\left(y_{2i} \mid y_{1i}, \underline{x}_i\right) = \frac{E_{s_1}\left(\pi_{2i}^m \mid \underline{y}_i, \underline{x}_i\right) f_{s_1}\left(y_{2i} \mid y_{1i}, \underline{x}_i\right)}{E_{s_1}\left(\pi_{2i}^m \mid y_{1i}, \underline{x}_i\right)} \tag{3.4}$$

where

$$E_{s_1}\left(\pi_{2i}^m | y_{1i}, \underline{x}_i\right) = \int E_{s_1}\left(\pi_{2i}^m | \underline{y}_i, \underline{x}_i\right) f_{s_1}\left(y_{2i} | y_{1i}, \underline{x}_i\right) dy_{2i} \tag{3.5}$$

and E_{s_1} denotes the expectation under the first sample pdf f_{s_1}.

Similar to (1.5), we have:

$$E_{s_1}\left(\pi_{2i}^m | \underline{y}_i, \underline{x}_i\right) = \frac{1}{E_{s_{2m}}\left(w_{2i}^m | \underline{y}_i, \underline{x}_i\right)} \tag{3.6a}$$

$$E_{s_1}\left(\pi_{2i}^m | y_{1i}, \underline{x}_i\right) = \frac{1}{E_{s_{2m}}\left(w_{2i}^m | y_{1i}, \underline{x}_i\right)} \tag{3.6b}$$

$$E_{s_1}\left(y_{2i} | y_{1i}, \underline{x}_i\right) = \frac{E_{s_{2m}}\left(w_{2i}^m y_{2i} | y_{1i}, \underline{x}_i\right)}{E_{s_{2m}}\left(w_{2i}^m | y_{1i}, \underline{x}_i\right)} \tag{3.6c}$$

$$E_{s_1}\left(y_{2i}\right) = \frac{E_{s_{2m}}\left(w_{2i}^m y_{2i}\right)}{E_{s_{2m}}\left(w_{2i}^m\right)} \tag{3.6d}$$

$$E_{s_1}\left(\pi_{2i}^m\right) = \frac{1}{E_{s_{2m}}\left(w_{2i}^m\right)} \tag{3.6e}$$

where $w_{2i}^m = 1/\pi_{2i}^m$ and $E_{s_{2m}}$ denotes the expectation under the matched sample pdf.

Using (3.4) and (3.6, a, b), we have the following:

$$\frac{f_{s_{2m}}\left(y_{2i} | y_{1i}, \underline{x}_i\right)}{f_{s_1}\left(y_{2i} | y_{1i}, \underline{x}_i\right)} = \frac{E_{s_{2m}}\left(w_{2i}^m | y_{1i}, \underline{x}_i\right)}{E_{s_{2m}}\left(w_{2i}^m | \underline{y}_i, \underline{x}_i\right)} \tag{3.7}$$

In order to find the distribution of the unmatched sample, we treat the sample-complement units in the first occasion as if it were a population. In the same way as in the matched sample one can obtain the conditional marginal unmatched sample pdf of Y_{2i} which is given by:

$$f_{s_{2u}}\left(y_{2i} \mid \underline{x}_i\right) = \frac{E_{s_1^c}\left(\pi_{2i}^u \mid y_{2i}, \underline{x}_i\right) f_{s_1^c}\left(y_{2i} \mid \underline{x}_i\right)}{E_{s_1^c}\left(\pi_{2i}^u \mid \underline{x}_i\right)} \tag{3.8}$$

where

$$E_{s_1^c}\left(\pi_{2i}^u \mid \underline{x}_i\right) = \int E_{s_1^c}\left(\pi_{2i}^u \mid y_{2i}, \underline{x}_i\right) f_{s_1^c}\left(y_{2i} \mid \underline{x}_i\right) dy_{2i} \tag{3.9}$$

and $E_{s_1^c}$ denotes the expectation under $f_{s_1^c}$.

Analogously to (3.6), we have the following relationships:

$$E_{s_1^c}\left(\pi_{2i}^u|y_{2i},\underline{x}_{2i}\right)=\frac{1}{E_{s_{2u}}\left(w_{2i}^u|y_{2i},\underline{x}_{2i}\right)} \tag{3.10a}$$

$$E_{s_1^c}\left(\pi_{2i}^u|x_i\right)=\frac{1}{E_{s_{2u}}\left(w_{2i}^u|\underline{x}_{2i}\right)} \tag{3.10b}$$

$$E_{s_1^c}\left(y_{2i}|\underline{x}_{2i}\right)=\frac{E_{s_{2u}}\left(w_{2i}^u y_{2i}|\underline{x}_{2i}\right)}{E_{s_{2u}}\left(w_{2i}^u|\underline{x}_{2i}\right)} \tag{3.10c}$$

$$E_{s_1^c}\left(y_{2i}\right)=\frac{E_{s_{2u}}\left(w_{2i}^u y_{2i}\right)}{E_{s_{2u}}\left(w_{2i}^u\right)} \tag{3.10d}$$

$$E_{s_1^c}\left(\pi_{2i}^u\right)=\frac{1}{E_{s_{2u}}\left(w_{2i}^u\right)} \tag{3.10e}$$

where $; w_{2i}^u = 1/\pi_{2i}^u$.

Using (3.8) and (3.10a) and (3.10b), we have the following:

$$\frac{f_{s_{2u}}\left(y_{2i}|\underline{x}_{2i}\right)}{f_{s_1^c}\left(y_{2i}|\underline{x}_{2i}\right)}=\frac{E_{s_{2u}}\left(w_{2i}^u|y_{2i},\underline{x}_{2i}\right)}{E_{s_{2u}}\left(w_{2i}^u|\underline{x}_{2i}\right)} \tag{3.11}$$

4. Marginal distributions of matched-complement and unmatched-complement samples

The matched-complement and unmatched-complement sample pdf's are needed to predict finite population totals for the second occasion using sample data for both occasions under informative sampling.

Similar to (2.6), the conditional marginal matched sample-complement pdf, i.e., the pdf for units $i \notin s_{2m}$ is defined as:

$$\begin{aligned} f_{s_{2m}^c}\left(y_{2i}|y_{1i},\underline{x}_i\right) &= f_{s_1}\left(y_{2i}|y_{1i},\underline{x}_i,I_{2i}^m=0\right) \\ &= \frac{\Pr\left(I_{2i}^m=0|\underline{y}_i,\underline{x}_i\right)f_{s_1}\left(y_{2i}|y_{1i},\underline{x}_i\right)}{\Pr\left(I_{2i}^m=0|y_{1i},\underline{x}_i\right)} \\ &= \frac{E_{s_{2m}}\left(\left(w_{2i}^m-1\right)\underline{y}_i,\underline{x}_i\right)f_{s_{2m}}\left(y_{2i}|y_{1i},\underline{x}_i\right)}{E_{s_{2m}}\left(\left(w_{2i}^m-1\right)y_{1i},\underline{x}_i\right)} \end{aligned} \tag{4.1}$$

Also, the following relationship holds:

$$\begin{aligned} E_{s_{2m}^c}\left(y_{2i}|y_{1i},\underline{x}_i\right) &= \frac{E_{s_2}\left(\left(1-\pi_{2i}^m\right)y_{2i}|y_{1i},\underline{x}_i\right)}{E_{s_2}\left(\left(1-\pi_{2i}^m\right)y_{1i},\underline{x}_i\right)} \\ &= \frac{E_{s_{2m}}\left(\left(w_{2i}^m-1\right)y_{2i}|y_{1i},\underline{x}_i\right)}{E_{s_{2m}}\left(\left(w_{2i}^m-1\right)y_{1i},\underline{x}_i\right)} \end{aligned} \tag{4.2}$$

Application of (3.7) and (4.1) yields the following ratios:

$$\frac{f_{s_{2m}}\left(y_{2i}|y_{1i},\underline{x}_i\right)}{f_{s_{2m}^c}\left(y_{2i}|y_{1i},\underline{x}_i\right)}=\frac{E_{s_{2m}}\left(\left(w_{2i}^m-1\right)y_{1i},\underline{x}_i\right)}{E_{s_{2m}}\left(\left(w_{2i}^m-1\right)\underline{y}_i,\underline{x}_i\right)} \tag{4.3}$$

and

$$\frac{f_{s_{2m}^c}\left(y_{2i}|y_{1i},\underline{x}_i\right)}{f_{s_1}\left(y_{2i}|y_{1i},\underline{x}_i\right)}=\frac{E_{s_{2m}}\left(w_{2i}^m|y_{1i},\underline{x}_i\right)E_{s_{2m}}\left(\left(w_{2i}^m-1\right)\underline{y}_i,\underline{x}_i\right)}{E_{s_{2m}}\left(w_{2i}^m|\underline{y}_i,\underline{x}_i\right)E_{s_{2m}}\left(\left(w_{2i}^m-1\right)y_{1i},\underline{x}_i\right)} \tag{4.4}$$

Analogously to (4.1), the conditional marginal unmatched sample-complement pdf, i.e., the pdf for units $i\notin s_{2u}$ is defined as:

$$\begin{aligned}f_{s_{2u}^c}\left(y_{2i}|\underline{x}_i\right)&=f_{s_1^c}\left(y_{2i}|\underline{x}_i,I_{2i}^u=0\right)\\&=\frac{\Pr\left(I_{2i}^u=0|y_{2i},\underline{x}_i\right)f_{s_1^c}\left(y_{2i}|\underline{x}_i\right)}{\Pr\left(I_{2i}^u=0|\underline{x}_i\right)}\\&=\frac{E_{s_{2u}}\left(\left(w_{2i}^u-1\right)y_{2i},\underline{x}_i\right)f_{s_{2u}}\left(y_{2i}|\underline{x}_i\right)}{E_{s_{2u}}\left(\left(w_{2i}^u-1\right)\underline{x}_i\right)}\end{aligned} \tag{4.5}$$

Also, we have the relationship:

$$\begin{aligned}E_{s_{2u}^c}\left(y_{2i}|\underline{x}_i\right)&=\frac{E_{s_1^c}\left(\left(1-\pi_{2i}^u\right)y_{2i}|\underline{x}_i\right)}{E_{s_1^c}\left(\left(1-\pi_{2i}^u\right)\underline{x}_i\right)}\\&=\frac{E_{s_{2u}}\left(\left(w_{2i}^u-1\right)y_{2i}|\underline{x}_i\right)}{E_{s_{2u}}\left(\left(w_{2i}^u-1\right)\underline{x}_i\right)}\end{aligned} \tag{4.6}$$

Using (3.11) and (4.5) we obtain the following ratios:

$$\frac{f_{s_{2u}}\left(y_{2i}|\underline{x}_i\right)}{f_{s_{2u}^c}\left(y_{2i}|\underline{x}_i\right)}=\frac{E_{s_{2u}}\left(\left(w_{2i}^u-1\right)\underline{x}_i\right)}{E_{s_{2u}}\left(\left(w_{2i}^u-1\right)y_{2i},\underline{x}_i\right)} \tag{4.7}$$

and

$$\frac{f_{s_{2u}^c}\left(y_{2i}|\underline{x}_i\right)}{f_{s_1^c}\left(y_{2i}|\underline{x}_i\right)}=\frac{E_{s_{2u}}\left(w_{2i}^u|\underline{x}_i\right)E_{s_{2u}}\left(\left(w_{2i}^u-1\right)y_{2i},\underline{x}_i\right)}{E_{s_{2u}}\left(w_{2i}^u|y_{2i},\underline{x}_i\right)E_{s_{2u}}\left(\left(w_{2i}^u-1\right)\underline{x}_i\right)} \tag{4.8}$$

5. Prediction of finite population totals under informative sampling

Sverchkov and Pfeffermann (2004) develop various methods for prediction of finite population totals under informative sampling for a single occasion using only information obtained from that occasion. In this and subsequent sections we extend these methods to predict finite population totals under informative sampling using data obtained from sampling on two occasions.

Let $T_2=\sum_{i=1}^N y_{2i}$ define the population total that we want to predict using the sample data from two occasions and possibly population values of auxiliary

variables that may contain some or all of the design variables. For the prediction process we have the following available information:

(a) The information that comes from the first occasion denoted by:

$$\Re_1 = \{(y_{1i}, \pi_{1i}) : i \in s_1, (x_{1i}, I_{1i}) : i \in U\} \tag{5.1}$$

(b) The information that comes from the second occasion denoted by:

$$\Re_2 = \left\{\left(y_{2i}, \pi_{2i}^m\right) : i \in s_{2m}; \left(x_{2i}, I_{2i}^m\right) : i \in s_1; \left(y_{2i}, \pi_{2i}^u\right) : i \in s_{2u}, \left(\underline{x}_i, I_{2i}^u\right), i \in s_1^c\right\} \tag{5.2}$$

Thus the available information for the prediction process is $\Re = \Re_1 \bigcup \Re_2$.

Let $\hat{T}_2 = \hat{T}_2(\Re)$ define the predictor. The mean square error (MSE) of $\hat{T}_2$ given $\Re$ with respect to the population pdf is defined by:

$$\begin{aligned} MSE\left(\hat{T}_2\right) &= E_p\left(\left(\hat{T}_2 - T_2\right)^2 \mid \Re\right) \\ &= \left(\hat{T}_2 - E_p(T_2 \mid \Re)\right)^2 + V_p(T_2 \mid \Re) \end{aligned} \tag{5.3}$$

Note that $V_p(T_2 \mid \Re)$ does not depend on $\hat{T}_2$, thus $MSE\left(\hat{T}_2\right)$ is minimized when $\hat{T}_2 = E_p(T_2 \mid \Re)$.

Now $E_p(T_2|\Re)$ can be composed as:

$$\begin{aligned} E_p(T_2|\Re) &= E_p\left(\sum_{i=1}^{N} y_{2i}|\Re\right) = \sum_{i\in U} E_p\left(y_{2i}|\Re\right) \\ &= \sum_{i\in s_{2m}} E_p\left(y_{2i}|\Re, I_{2i}^m = 1\right) + \sum_{i\in s_{2m}^c} E_p\left(y_{2i}|\Re, I_{2i}^m = 0\right) \\ &\quad + \sum_{i\in s_{2u}} E_p\left(y_{2i}|\Re, I_{2i}^u = 1\right) + \sum_{i\in s_{2u}^c} E_p\left(y_{2i}|\Re, I_{2i}^u = 0\right) \\ &= \sum_{i\in s_{2m}} E_{s_1}\left(y_{2i}|\Re_1\right) + \sum_{i\in s_{2m}^c} E_{s_1}\left(y_{2i}|\Re_1\right) + \sum_{i\in s_{2u}} E_{s_1^c}\left(y_{2i}|\Re_1\right) + \sum_{i\in s_{2u}^c} E_{s_1^c}\left(y_{2i}|\Re_2\right) \end{aligned} \tag{5.4}$$

where in the last equality we assume that $\{y_{2j}, j \notin s_{2m}\}$ and $\left\{\left(y_{2i}, \pi_{2i}^m\right) : i \in s_{2m}\right\}$ are independent given $y_{1j}, \underline{x}_j$, also $\{y_{2j} : j \notin s_{2u}\}$ and $\left\{\left(y_{2i}, \pi_{2i}^u\right) : i \in s_{2u}\right\}$ are independent given $\underline{x}_j$.

But we know the values $\{y_{2i} : i \in s_{2m}\}$ and $\{y_{2i} : i \in s_{2u}\}$, so that to predict T_2, we need to predict the Y_2 values not in s_{2m} and not in s_{2u}. Equation (4.4) can be written as:

$$E_p(T_2 \mid \Re) = \sum_{i\in s_{2m}} y_{2i} + \sum_{i\in s_{2u}} y_{2i} + \sum_{j\in s_{2m}^c} E_{s_{2m}^c}\left(y_{2j} \mid y_{1j}, \underline{x}_j\right) + \sum_{j\in s_{2u}^c} E_{s_{2u}^c}\left(y_{2j} \mid \underline{x}_j\right) \tag{5.5}$$

Thus the prediction of T_2 reduces to the prediction of $E_{s_{2m}^c}\left(y_{2j} \mid y_{1j}, \underline{x}_j\right)$ and $E_{s_{2u}^c}\left(y_{2j} | \underline{x}_j\right)$.

6. Prediction methods

In this section we consider the non-parametric and semi-parametric prediction of T_2.

6.1 Non-parametric prediction

In this subsection we consider estimation of the expectations $E_{s_{2m}^c}\left(y_{2j} \mid y_{1j}, \underline{x}_j\right)$ and $E_{s_{2u}^c}\left(y_{2j} \mid \underline{x}_j\right)$, and hence prediction of T_2, based on estimation of only sample expectations. The key to this method are the relationships (4.2) and (4.6). These relationships suggest the following two-step procedure:

Step-one:

(a) Estimate $E_{s_{2m}}\left(w_{2i}^m \mid y_{1i}, \underline{x}_i\right)$ and hence $q_{mi} = \dfrac{w_{2i}^m - 1}{E_{s_{2m}}\left(\left(w_{2i}^m - 1\right) \mid y_{1i}, \underline{x}_i\right)}$ by regressing w_{2i}^m against $\left(y_{1i}, \underline{x}_i\right), i \in s_{2m}$. Denote the resulting estimate by:

$$\hat{q}_{mi} = \frac{w_{2i}^m - 1}{\hat{E}_{s_{2m}}\left(w_{2i}^m \mid y_{1i}, \underline{x}_i\right) - 1} \tag{6.1}$$

and let $y_{2i}^m = \hat{q}_{mi} y_{2i}$.

For further discussion on estimation of this conditional expectation, under single stage informative sampling, see Eideh (2010).

(b) Estimate $E_{s_{2u}}\left(w_{2i}^u \mid \underline{x}_{2i}\right)$ and hence $q_{ui} = \dfrac{w_{2i}^u - 1}{E_{s_{2u}}\left(\left(w_{2i}^u - 1\right) \mid \underline{x}_{2i}\right)}$ by regressing w_{2i}^u against $\underline{x}_{2i}, i \in s_{2u}$. Denote the resulting estimate by:

$$\hat{q}_{ui} = \frac{w_{2i}^u - 1}{\hat{E}_{s_{2u}}\left(w_{2i}^u \mid \underline{x}_{2i}\right) - 1} \tag{6.2}$$

and let $y_{2i}^u = \hat{q}_{ui} y_{2i}$.

Step-two:

(a) Estimate $E_{s_{2m}}\left(y_{2i}^m \mid y_{1i}, \underline{x}_i\right)$ by regressing y_{2i}^m against $\left(y_{1i}, \underline{x}_i\right)$and substitute in (4.2) to get the estimate of $E_{s_{2m}^c}\left(y_{2i} \mid y_{1i}, \underline{x}_i\right)$.

(b) Estimate $E_{s_{2u}}\left(y_{2i}^u \mid \underline{x}_{2i}\right)$ by regressing y_{2i}^u against $\underline{x}_i$ and substitute in (4.6) to get the estimate of $E_{s_{2u}^c}\left(y_{2i} \mid \underline{x}_{2i}\right)$.

Thus by (5.5) the prediction of T_2 is given by:

$$\hat{T}_{2,1} = \sum_{i\in s_{2m}} y_{2i} + \sum_{i\in s_{2u}} y_{2i} + \sum_{j\in s_{2m}^c} \hat{E}_{s_{2m}}\left(\hat{q}_{mj} y_{2j} \mid y_{1j}, \underline{x}_{2j}\right) + \sum_{j\in s_{2u}^c} \hat{E}_{s_{2u}}\left(\hat{q}_{uj} y_{2j} \mid \underline{x}_{2j}\right) \quad (6.3)$$

This predictor depends on the models holding for the matched sample observations $y_{2i}^m = q_{2i}^m y_{2i}$, $i \in s_{2m}$ and the unmatched sample observations $y_{2i}^u = q_{2i}^u y_{2i}$, $i \in s_{2u}$.

Another predictor can be introduced which bases on the estimation of $E_{s_{2m}}\left(y_{2j} \mid y_{1j}, \underline{x}_{2j}\right)$ from the matched sample data, and the estimation of $E_{s_{2u}}\left(y_{2j} \mid \underline{x}_{2j}\right)$ from the unmatched sample data. The estimator depends on the relationship:

$$\begin{aligned}
&\sum_{j\in s_{2m}^c} E_{s_{2m}^c}\left(y_{2j}|y_{1j},\underline{x}_j\right) + \sum_{j\in s_{2u}^c} E_{s_{2u}^c}\left(y_{2j}|\underline{x}_j\right) \\
&= \sum_{j\in s_{2m}^c} \left(E_{s_{2m}}\left(y_{2j}|y_{1j},\underline{x}_j\right) + E_{s_{2m}^c}\left(y_{2j}|y_{1j},\underline{x}_j\right) - E_{s_{2m}}\left(y_{2j}|y_{1j},\underline{x}_j\right)\right) + \\
&\sum_{j\in s_{2u}^c} \left(E_{s_{2u}}\left(y_{2j}|\underline{x}_j\right) + E_{s_{2u}^c}\left(y_{2j}|\underline{x}_j\right) - E_{s_{2u}}\left(y_{2j}|\underline{x}_j\right)\right) \\
&= \sum_{j\in s_{2m}^c} E_{s_{2m}}\left(y_{2j}|y_{1j},\underline{x}_j\right) + \sum_{j\in s_{2m}^c} E_{s_{2m}^c}\left(\left(y_{2j} - E_{s_{2m}}\left(y_{2j}|y_{1j},\underline{x}_j\right)\right)y_{1j},\underline{x}_j\right) + \\
&\sum_{j\in s_{2u}^c} E_{s_{2u}}\left(y_{2j}|\underline{x}_j\right) + \sum_{j\in s_{2u}^c} E_{s_{2u}^c}\left(\left(y_{2j} - E_{s_{2u}}\left(y_{2j}|\underline{x}_j\right)\right)\underline{x}_j\right) \\
&\approx \sum_{j\in s_{2m}^c} E_{s_{2m}}\left(y_{2j}|y_{1j},\underline{x}_j\right) + (n-n_m)\frac{1}{(n-n_m)} \sum_{j\in s_{2m}^c} E_{s_{2m}^c}\left(\left(y_{2j} - E_{s_{2m}}\left(y_{2j}|y_{1j},\underline{x}_j\right)\right)\right) + \\
&\sum_{j\in s_{2u}^c} E_{s_{2u}}\left(y_{2j}|\underline{x}_j\right) + (N-n-n_u)\frac{1}{(N-n-n_u)} \sum_{j\in s_{2u}^c} E_{s_{2u}^c}\left(\left(y_{2j} - E_{s_{2u}}\left(y_{2j}|\underline{x}_j\right)\right)\right)
\end{aligned} \quad (6.4)$$

where

n_m is the size of the matched sample and n_u is the size of the unmatched sample. The nature of this approximation is based on Sverchkov and Pfeffermann (2004), equation (4.10).

Using (4.2) and (4.6), the matched sample-complement and the unmatched sample-complement means in the last two rows of (6.4) can be estimated, respectively, by:

$$\begin{aligned}
\hat{E}_{s_{2m}^c}\left(\left(y_{2j} - \hat{E}_{s_{2m}}\left(y_{2j} \mid y_{1j}, \underline{x}_j\right)\right)\right) &= \hat{E}_{s_{2m}}\left(\frac{w_{2j}^m - 1}{\hat{E}_{s_{2m}}\left(w_{2j}^m\right) - 1}\left(y_{2j} - \hat{E}_{s_{2m}}\left(y_{2j} \mid y_{1j}, \underline{x}_j\right)\right)\right) \\
&= \frac{1}{n_m} \sum_{i\in s_{2m}} \frac{w_{2i}^m - 1}{\frac{1}{n_m}\sum_{l\in s_{2m}} w_{2l}^m - 1}\left(y_{2i} - \hat{E}_{s_{2m}}\left(y_{2i} \mid y_{1i}, \underline{x}_i\right)\right)
\end{aligned} \quad (6.5)$$

and

$$\hat{E}_{s_{2u}^c}\left(\left(y_{2j}-\hat{E}_{s_{2u}}\left(y_{2j}\mid \underline{x}_j\right)\right)\right)=\hat{E}_{s_{2u}}\left(\frac{w_{2j}^u-1}{\hat{E}_{s_{2u}}\left(w_{2j}^u\right)-1}\left(y_{2j}-\hat{E}_{s_{2u}}\left(y_{2j}\mid \underline{x}_j\right)\right)\right)$$
$$=\frac{1}{n_u}\sum_{i\in s_{2u}}\frac{w_{2i}^u-1}{\frac{1}{n_u}\sum_{l\in s_{2u}}w_{2l}^u-1}\left(y_{2i}-\hat{E}_{s_{2u}}\left(y_{2i}\mid \underline{x}_i\right)\right) \quad (6.6)$$

Thus we have the following predictor for T_2:

$$\hat{T}_{2,2}=\sum_{i\in s_{2m}}y_{2i}+\sum_{i\in s_{2u}}y_{2i}+\sum_{j\in s_{2m}^c}\hat{E}_{s_{2m}}\left(y_{2j}\mid y_{1j},\underline{x}_j\right)+\sum_{j\in s_{2u}^c}\hat{E}_{s_{2u}}\left(y_{2j}\mid \underline{x}_j\right)+$$
$$\left(n-n_m\right)\frac{1}{n_m}\sum_{i\in s_{2m}}\left(\frac{w_{2i}^m-1}{\frac{1}{n_m}\sum_{l\in s_{2m}}w_{2l}^m-1}\left(y_{2i}-\hat{E}_{s_{2m}}\left(y_{2i}\mid y_{1i},\underline{x}_i\right)\right)\right)+ \quad (6.7)$$
$$\left(N-n-n_u\right)\frac{1}{n_u}\sum_{i\in s_{2u}}\left(\frac{w_{2i}^u-1}{\frac{1}{n_u}\sum_{l\in s_{2u}}w_{2l}^u-1}\left(y_{2i}-\hat{E}_{s_{2u}}\left(y_{2i}\mid \underline{x}_i\right)\right)\right)$$

This predictor is fully determined by estimating only the conditional expectations: $E_{s_{2m}}\left(y_{2j}\mid y_{1j},\underline{x}_j\right)$ and $E_{s_{2u}}\left(y_{2j}\mid \underline{x}_j\right)$, which can be carried out using an appropriate regression analysis.

6.2 Semi-parametric estimation under given matched sample-complement and given unmatched sample-complement models

Following Sverchkov and Pfeffermann (2004), in this section we show that if the models holding for units outside the matched and unmatched samples can be identified and estimated properly from the matched and unmatched data, it is possible to estimate the unknown parameters of these models without having to estimate the regressions $E_{s_{2m}}\left(w_{2i}^m\mid y_{1i},\underline{x}_i\right)$and $E_{s_{2u}}\left(w_{2i}^u\mid \underline{x}_i\right)$.

Assume that the matched sample-complement model for units outside the matched sample is:

$$y_{2j}=C_{\beta_m}\left(y_{1j},\underline{x}_j\right)+\varepsilon_{2j}^m$$
$$E_{s_{2m}^c}\left(\varepsilon_{2j}^m\mid y_{1j},\underline{x}_j\right)=0,\ E_{s_{2m}^c}\left(\varepsilon_{2j}^{m\,2}\mid y_{1j},\underline{x}_j\right)=\sigma^2 r_m\left(y_{1j},\underline{x}_j\right),\ j\notin s_{2m} \quad (6.8)$$

where $C_{\beta_m}\left(y_{1j},\underline{x}_j\right)$ is a known function of $\left(y_{1j},\underline{x}_{2j}\right)$ that depends on unknown vector parameter β_m and $r_m\left(y_{1j},\underline{x}_j\right)$ is a known function of $\left(y_{1j},\underline{x}_j\right)$ with σ^2 unknown.

The vector parameter β_m can be estimated by:

$$\begin{aligned}\beta_m &= \arg\min_{\tilde{\beta}_m} E_{s_{2m}^c}\left(\frac{\left(y_{2j}-C_{\tilde{\beta}_m}\left(y_{1j},\underline{x}_j\right)\right)^2}{r_m\left(y_{1j},\underline{x}_j\right)} \mid y_{1j},\underline{x}_j\right)\\ &= \arg\min_{\tilde{\beta}_m} E_{s_{2m}}\left(\left(\frac{w_{2j}^m-1}{E_{s_{2m}}\left(\left(w_{2j}^m-1\right)\mid y_{1j},\underline{x}_j\right)}\right)\frac{\left(y_{2j}-C_{\tilde{\beta}_m}\left(y_{1j},\underline{x}_j\right)\right)^2}{r_m\left(y_{1j},\underline{x}_j\right)} \mid y_{1j},\underline{x}_j\right)\end{aligned} \tag{6.9}$$

where the second row of (6.9) is obtained using (4.2).

Thus the vector parameter β_m can be estimated, based on only the matched sample data, by:

$$\hat{\beta}_{m1} = \arg\min_{\tilde{\beta}} \sum_{i\in s_{2m}}\left(\hat{q}_{mi}\frac{\left(y_{2i}-C_{\tilde{\beta}_m}\left(y_{1i},\underline{x}_i\right)\right)^2}{r_m\left(y_{1i},\underline{x}_i\right)}\right) \tag{6.10}$$

where $\hat{q}_{mi}=\dfrac{w_{2i}^m-1}{\hat{E}_{s_{2m}}\left(w_{2i}^m \mid y_{1i},\underline{x}_i\right)-1}$.

Similarly suppose that the unmatched sample-complement model for units outside the unmatched sample is of the form:

$$\begin{aligned} y_{2j} &= C_{\beta_u}\left(\underline{x}_j\right)+\varepsilon_{2j}^u\\ E_{s_{2u}^c}\left(\varepsilon_{2j}^u \mid \underline{x}_j\right)&=0,\ E_{s_{2u}^c}\left({\varepsilon_{2j}^u}^2 \mid \underline{x}_j\right)=\sigma^2 r_u\left(\underline{x}_j\right),\ j\notin s_{2u}\end{aligned} \tag{6.11}$$

where $C_{\beta_u}\left(\underline{x}_j\right)$ is a known function of $\underline{x}_j$ that depends on an unknown vector parameter

β_u and $r_u\left(\underline{x}_j\right)$is a known function of $\underline{x}_j$ with σ^2 unknown.

The vector parameter β_u can be estimated by:

$$\begin{aligned}\beta_u &= \arg\min_{\tilde{\beta}_u} E_{s_{2u}^c}\left(\frac{\left(y_{2j}-C_{\tilde{\beta}_u}\left(\underline{x}_j\right)\right)^2}{r_u\left(\underline{x}_j\right)} \mid \underline{x}_j\right)\\ &= \arg\min_{\tilde{\beta}_u} E_{s_{2u}}\left(\left(\frac{w_{2i}^u-1}{E_{s_{2u}}\left(\left(w_{2i}^u-1\right)\mid \underline{x}_i\right)}\right)\frac{\left(y_{2i}-C_{\tilde{\beta}_u}\left(\underline{x}_i\right)\right)^2}{r_u\left(\underline{x}_i\right)} \mid \underline{x}_i\right)\end{aligned} \tag{6.12}$$

where the second row of (6.12) is obtained using (4.6).

Thus the vector parameter β_u can be estimated, based only on the unmatched sample data, by:

$$\hat{\beta}_{u1} = \arg\min_{\tilde{\beta}_u} \sum_{i\in s_{2u}}\left(\hat{q}_{ui}\frac{\left(y_{2i}-C_{\tilde{\beta}_u}\left(\underline{x}_i\right)\right)^2}{r_u\left(\underline{x}_i\right)}\right) \tag{6.13}$$

where $\hat{q}_{ui} = \dfrac{w_{2i}^{u} - 1}{\hat{E}_{s_{2u}}\left(w_{2i}^{u} \mid \underline{x}_i\right) - 1}$.

In our situation we have the following estimates of $E_{s_{2m}^{c}}\left(y_{2j} \mid y_{1j}, \underline{x}_j\right)$ and $E_{s_{2u}^{c}}\left(y_{2j} | \underline{x}_j\right)$:

$$\hat{E}_{s_{2m}^{c}}\left(y_{2j} \mid y_{1j}, \underline{x}_j\right) = \mathrm{C}_{\hat{\beta}_{m1}}\left(y_{2j}, \underline{x}_j\right) \tag{6.14}$$

and

$$\hat{E}_{s_{2u}^{c}}\left(y_{2j} \mid \underline{\mathrm{x}}_j\right) = C_{\hat{\beta}_{u1}}\left(\underline{x}_j\right) \tag{6.15}$$

Thus the predictor of the finite population total, T_2, is given by:

$$\begin{aligned}\hat{T}_{2,3} &= \sum_{i \in s_{2m}} y_{2i} + \sum_{i \in s_{2u}} y_{2i} + \sum_{j \in s_{2m}^{c}} \hat{E}_{s_{2m}^{c}}\left(y_{2j} | y_{1j}, \underline{x}_j\right) + \sum_{j \in s_{2u}^{c}} \hat{E}_{s_{2u}^{c}}\left(y_{2j} | \underline{x}_j\right) \\ &= \sum_{i \in s_{2m}} y_{2i} + \sum_{i \in s_{2u}} y_{2i} + \sum_{j \in s_{2m}^{c}} C_{\hat{\beta}_{m1}}\left(y_{2j}, \underline{x}_j\right) + \sum_{j \in s_{2u}^{c}} C_{\hat{\beta}_{u1}}\left(\underline{x}_j\right)\end{aligned} \tag{6.16}$$

Now we can base our prediction of the finite population total, T_2, without conditioning on the y_{1i} and $\underline{x}_i$ or $\underline{x}_i$, because from (6.8) and (6.11) we can deduce that:

$$E_{s_{2m}^{c}}\left(\frac{\left(y_{2j} - C_{\beta_m}\left(y_{1j}, \underline{x}_j\right)\right)^2}{r_m\left(y_{1j}, \underline{x}_j\right)} \mid y_{1j}, \underline{x}_j\right) = E_{s_{2m}^{c}}\left(\frac{\left(y_{2j} - C_{\beta_m}\left(y_{1j}, \underline{x}_j\right)\right)^2}{r_m\left(y_{1j}, \underline{x}_j\right)}\right) \tag{6.17}$$

and

$$E_{s_{2u}^{c}}\left(\frac{\left(y_{2j} - C_{\beta_u}\left(\underline{x}_j\right)\right)^2}{r_u\left(\underline{x}_j\right)} \mid \underline{x}_j\right) = E_{s_{2u}^{c}}\left(\frac{\left(y_{2i} - C_{\beta_u}\left(\underline{x}_i\right)\right)^2}{r_u\left(\underline{x}_i\right)}\right) \tag{6.18}$$

Thus, application of (4.2) and (4.6) to the right hand sides of (6.17) and (6.18) but without conditioning on y_{1i} and $\underline{x}_i$ or $\underline{x}_i$ and since $E_{s_{2m}}\left(w_{2i}^{m}\right)$and $E_{s_{2u}}\left(w_{2i}^{u}\right)$ are constants, leads to the following estimates:

$$\hat{\beta}_{m2} = \arg\min_{\tilde{\beta}} \sum_{i \in s_{2m}}\left(\left(w_{2i}^{m} - 1\right)\frac{\left(y_{2i} - C_{\tilde{\beta}_m}\left(y_{1i}, \underline{x}_i\right)\right)^2}{r_m\left(y_{1i}, \underline{x}_i\right)}\right) \tag{6.19}$$

and

$$\hat{\beta}_{u2} = \arg\min_{\tilde{\beta}_u} \sum_{i \in s_{2u}}\left(\left(w_{2i}^{u} - 1\right)\frac{\left(y_{2i} - C_{\tilde{\beta}_u}\left(\underline{x}_i\right)\right)^2}{r_u\left(\underline{x}_i\right)}\right) \tag{6.20}$$

Hence the following predictor of the finite population total, T_2:

$$\hat{T}_{2,4} = \sum_{i \in s_{2m}} y_{2i} + \sum_{i \in s_{2u}} y_{2i} + \sum_{j \notin s_{2m}^{c}} C_{\hat{\beta}_{m2}}\left(y_{2j}, \underline{x}_j\right) + \sum_{j \notin s_{2u}^{c}} C_{\hat{\beta}_{u2}}\left(\underline{x}_j\right) \tag{6.21}$$

Note that the predictor $\hat{T}_{2,4}$ does not require the identification and estimation of the expectations: $E_{s_{2m}}\left(w_{2i}^{m} \mid y_{1i}, \underline{x}_i\right)$and $E_{s_{2u}}\left(w_{2i}^{u} \mid \underline{x}_i\right)$, while $\hat{T}_{2,3}$ requires that.

7. Examples

7.1 Prediction with no auxiliary variables

In this section we assume that there are no auxiliary variables x_2 and y_1, and in the next section we assume the auxiliary variable y_1. So the predictor is given by:

$$\hat{T}_2 = \sum_{i\in s_{2m}} y_{2i} + \sum_{i\in s_{2u}} y_{2i} + \sum_{j\notin s_{S_{2m}}} \hat{E}_{s_{2m}^c}\left(y_{2j}\right) + \sum_{j\notin s_{2u}} E_{s_{2u}^c}\left(y_{2j}\right) \tag{7.1}$$

It follows from (4.2) and (4.6) that:

$$\begin{aligned}\hat{T}_2 &= \sum_{i\in s_{2m}} y_{2i} + \sum_{i\in s_{2u}} y_{2i} + \sum_{j\notin s_{S_{2M}}} \hat{E}_{s_{2m}}\left(\frac{w_{2i}^{m}-1}{\hat{E}_{s_{2m}}\left(w_{2i}^{m}-1\right)} y_{2i}\right) + \sum_{j\notin s_{2U}} \hat{E}_{s_{2u}}\left(\frac{w_{2i}^{u}-1}{\hat{E}_{s_{2u}}\left(w_{2i}^{u}-1\right)} y_{2j}\right) \\ &= \sum_{i\in s_{2m}} y_{2i} + \sum_{i\in s_{2u}} y_{2i} + \left(n-n_m\right)\hat{E}_{s_{2m}}\left(\frac{w_{2i}^{m}-1}{\hat{E}_{s_{2m}}\left(w_{2i}^{m}-1\right)} y_{2i}\right) + \left(N-n-n_u\right)\hat{E}_{s_{2u}}\left(\frac{w_{2i}^{u}-1}{\hat{E}_{s_{2u}}\left(w_{2i}^{u}-1\right)} y_{2i}\right)\end{aligned} \tag{7.2}$$

Estimating the four unknown expectations in (7.2) by the respective sample means yields the following estimate:

$$\begin{aligned}\hat{T}_{2new} &= \sum_{i\in s_{2m}} y_{2i} + \sum_{i\in s_{2u}} y_{2i} + \left(n-n_m\right)\frac{1}{n_m}\sum_{i\in s_{2m}} \frac{\left(w_{2i}^{m}-1\right)}{\frac{1}{n_m}\sum_{l\in s_{2m}}\left(w_{2l}^{m}-1\right)} y_{2i} \\ &\quad + \left(N-n-n_u\right)\frac{1}{n_u}\sum_{i\in s_{2u}} \frac{\left(w_{2i}^{u}-1\right)}{\frac{1}{n_u}\sum_{l\in s_{2u}}\left(w_{2l}^{u}-1\right)} y_{2i} \\ &= \sum_{i\in s_{2m}} y_{2i} + \sum_{i\in s_{2u}} y_{2i} + \left(n-n_m\right)\sum_{i\in s_{2m}} \frac{\left(w_{2i}^{m}-1\right)}{\sum_{l\in s_{2m}}\left(w_{2l}^{m}-1\right)} y_{2i} \\ &\quad + \left(N-n-n_u\right)\sum_{i\in s_{2u}} \frac{\left(w_{2i}^{u}-1\right)}{\sum_{l\in s_{2u}}\left(w_{2l}^{u}-1\right)} y_{2i} \\ &= \sum_{i\in s_{2m}} y_{2i} + \sum_{i\in s_{2u}} y_{2i} + \frac{n-n_m}{\sum_{i\in s_{2m}}\left(w_{2i}^{m}-1\right)}\sum_{i\in s_{2m}}\left(w_{2i}^{m}-1\right) y_{2i} + \frac{N-n-n_u}{\sum_{l\in s_{2u}}\left(w_{2l}^{u}-1\right)}\sum_{i\in s_{2u}}\left(w_{2i}^{u}-1\right) y_{2i}\end{aligned} \tag{7.3}$$

For sampling design such that $\sum_{i\in s_{2m}} w_{2i}^{m} = n$ and $\sum_{i\in s_u} w_{2i}^{u} = N-n$ for all s_{2m} and s_{2u} or if we estimate the expectations by: $\hat{E}_{s_{2m}}\left(w_{2i}^{m}\right) = \frac{n}{n_m}$ and $\hat{E}_{s_{2u}}\left(w_{2i}^{u}\right) = \frac{N-n}{n_u}$, the estimator becomes the Horvitz-Thompson:

$$\hat{T}_{2,H-T} = \sum_{i\in s_{2m}} w_{2i}^{m} y_{2i} + \sum_{i\in s_{2u}} w_{2i}^{u} y_{2i} \tag{7.4}$$

For given s_1, this is the estimator obtained for T_2 in the case of stratified sampling with two strata, the first stratum is s_1 and the second stratum is s_1^c. So that, the sampling units of the first sample is used to divide the population into two subpopulations. However, for noninformative sampling and if $w_{2i}^{m} = \frac{n}{n_m}, i \in s_1$ and $w_{2i}^{u} = \frac{N-n}{n_u}$, $i \in s_1^c$, the estimator reduces to:

$$\hat{T}_{2,st} = n\sum_{i\in s_{2m}} \frac{y_{2i}}{n_m} + (N-n)\sum_{i\in s_u} \frac{y_{2i}}{n_u} = n\bar{y}_{2m} + (N-n)\bar{y}_{2u} \tag{7.5}$$

Now, conditionally given s_1, this is the stratified estimator in the case of stratified simple random sampling without replacement.

Comment 2. Note that the estimator given in (7.5) which does not use auxiliary variable y_1, is not equivalent to the composite estimator given in Cochran (1977), because the composite estimator takes advantage of the correlation between y_1 and y_2 unlike the estimator given in (7.5). So the composite estimator is more efficient than (7.5).

Now, rather than predicting $\sum_{j\in s_{2m}^c} y_{2j}$ and $\sum_{j\in s_{2u}^c} y_{2j}$, we predict $\sum_{j\in s_1} y_{2j}$ and $\sum_{j\in s_1^c} y_{2j}$ by the corresponding expectations $\sum_{i\in s_1} E_{s_1}\left(y_{2i}\right)$ and $\sum_{i\in s_1^c} E_{s_1^c}\left(y_{2i}\right)$. By (3.6e) and (3.10e), the finite population total is predicted as:

$$\begin{aligned}
\hat{T}_{2,t} &= \sum_{i\in s_1} \frac{\hat{E}_{s_{2m}}\left(w_{2i}^{m} y_{2i}\right)}{\hat{E}_{s_{2m}}\left(w_{2i}^{m}\right)} + \sum_{i\in s_1^c} \frac{\hat{E}_{s_{2u}}\left(w_{2i}^{u} y_{2i}\right)}{\hat{E}_{s_{2u}}\left(w_{2i}^{u}\right)} \\
&= \sum_{i\in s_1} \frac{\sum_{i\in s_{2m}} w_{2i}^{m} y_{2i}}{\sum_{i\in s_{2m}} w_{2i}^{m}} + \sum_{i\in s_1^c} \frac{\sum_{i\in s_{2u}} w_{2i}^{u} y_{2i}}{\sum_{i\in s_{2u}} w_{2i}^{u}} \\
&= n\frac{\sum_{i\in s_{2m}} w_{2i}^{m} y_{2i}}{\sum_{i\in s_{2m}} w_{2i}^{m}} + (N-n)\frac{\sum_{i\in s_{2u}} w_{2i}^{u} y_{2i}}{\sum_{i\in s_{2u}} w_{2i}^{u}}
\end{aligned} \tag{7.6}$$

Note that the predictors $\hat{T}_{2,H-T}$ and $\hat{T}_{2,t}$ are coincide for sampling designs such that $\sum_{i\in s_{2m}} w_{2i}^{m} = n$ and $\sum_{i\in s_u} w_{2i}^{u} = N-n$ for all s_{2m} and s_{2u}.

7.2 Prediction with auxiliary variables

In this section we consider prediction of the finite population total for the second occasion utilizing the data collected on the study variable, in the first occasion, as an auxiliary variable, and assuming that $\underline{x}_i = 1 \text{ for all } i$. Then:

$$\hat{T}_2 = \sum_{i\in s_{2m}} y_{2i} + \sum_{i\in s_{2u}} y_{2i} + \sum_{j\in s_{2m}^c} \hat{E}\left(y_{2j} \mid y_{1j}\right) + \sum_{j\in s_{2u}^c} \hat{E}\left(y_{2j}\right) \tag{7.8}$$

So to find $\hat{T}_2$ we need to identify and estimate $E_{s_{2m}^c}\left(y_{2j} \mid y_{1j}\right)$ and $E\left(y_{2j}\right)$. But $E\left(y_{2j}\right)$ is estimated in Section (7.1), so as it require to estimate $E_{s_{2m}^c}\left(y_{2j} \mid y_{1j}\right)$. In order to do this, following Sverchkov and Pfeffermann (2004), let the population model for the first sample be:

$$\begin{aligned} &y_{2i} = H_\beta\left(y_{1i}\right) + \varepsilon_i, E_{s_1}\left(\varepsilon_i \mid y_{1i}\right) = 0 \\ &E_{s_1}\left(\varepsilon_i^2 \mid y_{1i}\right) = r\left(y_{1i}\right), E\left(\varepsilon_i \varepsilon_j \mid y_{1i}, y_{1j}\right) = 0 \text{ for } i \neq j = 1,\ldots,n \end{aligned} \tag{7.9}$$

and suppose that the matched sample inclusion probabilities can be modeled as:

$$\pi_{2i}^{m} = k\left(y_{2i} g\left(y_{1i}\right) + \delta_i\right), E_{s_1}\left(\delta_i \mid \underline{y}_i\right) = 0 \tag{7.10}$$

where $H_\beta\left(y_1\right), r\left(y_1\right) \text{and } g\left(y_1\right)$ are positive functions and k is a normalizing constant. See Eideh (2010) for the effect of k on estimation.

Under (7.9), $\pi_{2i}^{m}\left(y_{1i}\right) = kH_\beta\left(y_{1i}\right) g\left(y_{1i}\right)$. Hence by (2.7) and (7.10):

$$\begin{aligned} E_{s_{2m}^c}\left(y_{2i} \mid y_{1i}\right) &= E_{s_1}\left(\frac{1-\pi_{2i}^{m}}{1-\pi_{2i}^{m}\left(y_{1i}\right)} y_{2i} \mid y_{1i}\right) \\ &= E_{s_1}\left(\frac{1-\pi_{2i}^{m}\left(y_{1i}\right) - k\varepsilon_i g\left(y_{1i}\right) - k\delta_i}{1-\pi_{2i}^{m}\left(y_{1i}\right)} y_{2i} \mid y_{1i}\right) \\ &= E_{s_1}\left(y_{2i} \mid y_{1i}\right) - \frac{kg\left(y_{1i}\right) r\left(y_{1i}\right)}{1-\pi_{2i}^{m}\left(y_{1i}\right)} \end{aligned} \tag{7.11}$$

As a special case of (7.9), suppose that $H_\beta\left(y_1\right) = y_1\beta$ and $r\left(y_1\right) = \sigma^2 y_1$. Let in (7.10) $g\left(y_1\right) = 1$ for all y_1 so that $\pi_{2i}^{m} = \dfrac{m\left(y_{2i}+\delta_i\right)}{\sum_{i=1}^{n}\left(y_{2i}+\delta_i\right)}$ which for sufficiently large n and under some regularity conditions can be approximated as

$\pi_{2i}^{m} = \dfrac{m(y_{2i}+\delta_i)}{n\beta\bar{y}_1}$,where $\bar{y}_1 = \dfrac{1}{n}\sum_{i=1}^{n} y_{1i}$, implying that $E_{s_1}\left(\pi_{2i}^{m} \mid y\right) \approx \dfrac{my_{1i}}{n\bar{y}_1}$. By (7.11) w e have:

$$E_{s_{2m}^c}\left(y_{2j} \mid y_{1j}\right) = y_{1j}\beta - \frac{\sigma^2 y_{1j}}{\beta\left(\dfrac{m}{n}\bar{y}_1 - y_{1j}\right)} \tag{7.12}$$

So that for known β and σ^2 and using the results of Section (7.1), the predictor of T_2 is given by:

$$\begin{aligned}\hat{T}'_{2,new} = &\sum_{i\in s_{2m}} y_{2i} + \sum_{i\in s_{2u}} y_{2i} + \beta\sum_{j\in s_{2m}^c} y_{1j} - \frac{\sigma^2}{\beta}\sum_{j\in s_{2m}^c}\frac{y_{1j}}{\left(\dfrac{n}{m}\right)\bar{y}_1 - y_{1j}} + \\ &N - n - n_u\sum_{i\in s_{2u}}\frac{\left(w_{2i}^{u}-1\right)}{\sum_{l\in s_{2u}}\left(w_{2l}^{u}-1\right)}y_{2i}\end{aligned} \tag{7.13}$$

8. Mean square error estimation

Let $\Re_1 = \left\{\left(y_{1i},\pi_{1i}\right): i\in s_1, \left(x_{1i},I_{1i}\right): i\in U\right\}$ be the information that comes from the first occasion and $\Re_2 = \left\{\left(y_{2i},\pi_{2i}^{m}\right): i\in s_{2m}; \left(x_{2i},I_{2i}^{m}\right): i\in s_1; \left(y_{2i},\pi_{2i}^{u}\right) i\in s_{2u}, \left(\underline{x}_i,I_{2i}^{u}\right): i\in s_1^c\right\}$ be the information that comes from the second occasion, so that the available information for the prediction process is $\Re = \Re_1 \bigcup \Re_2$.

Let $\hat{T}_2 = \hat{T}_2(\Re)$ define the predictor. The mean square error (MSE) of $\hat{T}_2$ given $\Re$ with respect to the population pdf is defined by:

$$\begin{aligned}MSE\left(\hat{T}_2 \middle| \Re\right) &= E_p\left(\left(\hat{T}_2 - T_2\right)^2 \mid \Re\right) \\ &= \left(\hat{T}_2 - E_p\left(T_2 \mid \Re\right)\right)^2 + V_p\left(T_2 \mid \Re\right)\end{aligned} \tag{8.1}$$

Following Sverchkov and Pfeffermann (2004), estimation of $MSE\left(\hat{T}_2\right)$ for the predictors $\hat{T}_2$ considered in Sections 4, 5, 6 and 7, requires strict model assumptions that could be hard to validate. This is largely due to the conditioning on the design information $\Re = \Re_1 \bigcup \Re_2$. In order to deal with this problem, we propose to estimate instead the unconditional mean square error:

$$MSE\left(\hat{T}_2\right) = E\left(\hat{T}_2 - T_2\right)^2 = E_{\Re}\left\{E_p\left(\left(\hat{T}_2 - T_2\right)^2 \mid \Re\right)\right\} \tag{8.2}$$

where $E_{\Re} = E_{\Re'}E_s$ defines the expectation over the sample distribution (given the selected sample) and over all possible sample selections. By changing the order of expectations, the unconditional mean square error can be expressed as:

$$MSE\left(\hat{T}_2\right) = E_s E_p E_{\Re'}\left(\left(\hat{T}_2 - T_2\right)^2 \mid \mathbf{y}_2\right) \tag{8.3}$$

where $\mathbf{y}_2 = \{y_{2i} : i \in U\}$. Estimating the unconditional mean square error of any predictor $\hat{T}_2$ can be carried out therefore by estimating its randomization mean square error. See Pfeffermann (1993) and Sverchkov and Pfeffermann (2004) for further discussion.

9. Conclusions

In this paper we use the sample and sample-complement distributions for occasion one, the matched sample and unmatched sample distributions, and matched sample-complement and unmatched sample-complement for occasion two, for deriving predictors of finite population totals under informative probability sampling using data obtained from sampling on two occasions. Known predictors in common use are shown to be special cases of the present predictors obtained under informative sampling. According to Sverchkov and Pfeffermann (2004) and Pfeffermann and Sverchkov (2007) the mean square error estimation of the new predictors can be obtained by a combination of an inverse sampling algorithm and a resampling method. Hence further experimentation with this kind of predictors and MSE (mean square error) estimation is therefore highly recommended. The paper is purely mathematical. The performance of likewise predictors for single occasion based on simulation study and empirical data can be found in Sverchkov and Pfeffermann (2004). I hope that the new mathematical results obtained will encourage further theoretical, empirical and practical research in these directions.

Acknowledgments

The author is grateful to the Associate Editor, to the referees, and to Gad Nathan, and Danny Pfeffermann for their valuable comments.

References

1. Arnab, R. (1998). Sampling on Two Occasions: Estimation of Population Total. *Survey Methodology*, Vol. 24, No.2, pp. 185-192.
2. Choudhry, G.H., and Graham, J.E. (1983). Sampling on Two Occasions with PPSWOR. *Survey Methodology*, Vol. 9, No.1, pp. 139-151.
3. Eideh, A. H. (2007). A Correction Note on Small Area Estimation. *International Statistical Review*. Volume 75, Issue 1, pp. 122-123.
4. Eideh A.H. (2008). Estimation and Prediction of Random Effects Models for Longitudinal Survey Data under Informative Sampling. *Statistics in Transition – New Series .Volume 9, Number 3, December 2008,* pp. 485 – 502.
5. Eideh A.H. (2009). On the use of the Sample Distribution and Sample Likelihood for Inference under Informative Probability Sampling. *DIRASAT (Natural Science)*, *Volume 36 (2009), Number 1*, pp. 18-29.

6. Eideh, A. H. (2010). Analytic Inference of Complex Survey Data under Informative Probability Sampling. *Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume I. The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010*). Edited by: Zeinab Amin and Ali S. Hadi. The American University in Cairo: pp 507–536.

7. Eideh, A. H. and Nathan, G. (2006). Fitting Time Series Models for Longitudinal Survey Data under Informative Sampling. *Journal of Statistical Planning and Inference*, 136, pp. 3052-3069.

8. Eideh, A. H. and Nathan, G. (2009). Two-Stage Informative Cluster Sampling with application in Small Area Estimation. *Journal of Statistical Planning and Inference.139*, pp. 3088-3101.

9. Krieger, A.M, and Pfeffermann, D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*. 13: pp. 123-142.

10. Pfeffermann, D. (1993). The role of sampling weight when modeling survey data. *International Statistical Review* 61: 317-337.

11. Pfeffermann, D., Krieger, A. M, and Rinott, Y. (1998). Parametric Distributions of Complex Survey Data under Informative Probability Sampling. *Statistica Sinica*, 8, pp. 1087- 1114.

12. Pfeffermann, D. and Sverchkov, M. (1999). Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data. *Sankhya*, 61, Ser. B, pp. 66-186.

13. Pfeffermann, D. and Sverchkov, M. (2003). Fitting Generalized Linear Models under Informative Probability Sampling. *Analysis of Survey Data*. (eds. R. Chambers and C. J. Skinner), pp. 175-195. New York: Wiley.

14. Pfeffermann, D. and Sverchkov, M. (2007). Small Area Estimation under Informative Probability Sampling of Areas and within the Selected Areas. *Journal of the American Statistical Association*, Vol. 102, No. 480, pp. 1427-1438.

15. Prasad, N.G.N., and Graham, J.E. (1994). PPS Sampling over Two Occasions. *Survey Methodology*, Vol. 20, No.1, pp. 59-64.

16. Skinner, C.J.(1994). Sample models and weights. *American Statistical Association Proceedings of the Section on Survey Research Methods*, pp. 133-142.

17. Sverchkov, M. and Pfeffermann, D. (2004). Prediction of Finite Population Totals based on the Sample Distribution. *Survey Methodology*, Vol. 30, No. 1, pp. 79-92.