Mass Imputation for Two-phase Sampling

Seho Park, Jae Kwang Kim[1,*]

Department of Statistics, Iowa State University, Ames, IA, 50011, USA

Abstract

Two-phase sampling is a cost-effective method of data collection using outcome-dependent sampling for the second-phase sample. In order to make efficient use of auxiliary information and to improve domain estimation, mass imputation can be used in two-phase sampling. Rao & Sitter (1995) introduce mass imputation for two-phase sampling and its variance estimation under simple random sampling in both phases. In this paper, we extend the Rao-Sitter method to general sampling design. The proposed method is further extended to mass imputation for categorical data. A limited simulation study is performed to examine the performance of the proposed methods.

Keywords: Auxiliary information, categorical data, domain estimation, outcome-dependent sampling

2010 MSC: 62D05, 62D05

1. Introduction

Two-phase sampling, first introduced by Neyman (1938), is a convenient and economical sampling design where the sample selection is conducted in two phases. In phase one, a large sample is collected from the target population and a relatively inexpensive auxiliary variable x is measured. In phase two, a

*Corresponding author
Email address: jkim@iastate.edu (Jae Kwang Kim)

smaller sample is drawn from the first-phase sample and the study variable y, which is expensive to measure, is collected.

Two-phase sampling or double sampling increases the precision of estimates by using auxiliary information available from the first-phase sample. Two-phase sampling is also called outcome-dependent sampling since the second-phase sampling design depends on the observations from the first-phase sampling. Hidiroglou (2001) and Legg & Fuller (2009) provided comprehensive overviews of two-phase sampling.

The structure of two-phase sampling can be seen as a missing data problem. Since y's are observed only in the second-phase sample and are missing in the remaining part of the first-phase sample, we can regard the two-phase sample as a planned missing data problem and apply methods for handling missing data. One popular technique is to create imputation for the missing values in the first-phase sample. It is also called as mass imputation (Kim & Rao, 2012) since it requires generating a large number of imputed values.

In large-scale surveys, it is sometimes convenient or requested to produce estimates for various domains. Estimates for domains, or small area, can be computed using various techniques, including mass imputation (Moore & Robbins, 2004). Breidt et al. (1996) also considered using imputation method for domain estimation under two-phase sampling and showed that the estimates obtained using mass imputation provide better estimates at finer levels of detail.

Mass imputation is also applicable to survey data integration problem, in which two surveys are combined for enhanced estimation. Chipperfield et al. (2012) developed mass imputation for data integration combining two independent surveys with common measurements. They considered the composite estimation after mass imputation for improved estimation. Kim & Rao (2012)

also discussed mass imputation under non-nested two-phase sampling and the conditions for design consistency.

Rao & Sitter (1995) introduced a mass imputation method for two-phase sampling when both phases use the simple random sampling design. In this paper, we extend it to the complex sampling designs in each of the two phases. We propose mass imputation using a "working" regression model and replication variance estimation method for the mass imputation estimator. In addition, we extend the proposed method to cover categorical data mass imputation.

The rest of the paper is organized as follows. In Section 2, we introduce notation used throughout the paper and introduce two-phase regression estimator and its known properties. In Section 3, we present the proposed mass imputation estimator with its asymptotic properties. In Section 4, replication variance estimation for the proposed mass imputation estimator is discussed. In Section 5, an extension to categorical data mass imputation is discussed and in Section 6, an illustrative example is provided. Results from a simulation study is presented in Section 7 and concluding remarks are made in Section 8.

2. Basic Setup

To discuss the setup for two-phase sampling, consider a finite population, denoted by $\mathcal{F}_N = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}$ is a column vector of dimension p and y is a scalar. Let A_1 denote the index set of the first-phase sample of size n_1 collected from the finite population. For the first-phase sample A_1, we assume that the first-order inclusion probability of unit i, denoted by $\pi_{1i} = P(i \in A_1)$, is known for all element $i \in A_1$. From the first-phase sample, we select a second-phase sample by a probability sampling design with known conditional first-order inclusion probability $\pi_{2i|1i} = P(i \in A_2 | i \in A_1)$. The conditional first-order inclusion probability is random in the sense that it

depends on the observations from the first-phase sample. We assume that $\pi_{2i|1i}$ are available throughout the first-phase sample.

Let w_{1i} denote the sampling weight for the first-phase sample and it is the reciprocal of the first-order inclusion probability for the first-phase sample; $w_{1i} = \pi_{1i}^{-1}$. Also, $w_{2i|1i}$ is defined as the conditional sampling weight for the second-phase sample that is the reciprocal of the conditional inclusion probability of the second-phase sample, that is $w_{2i|1i} = \pi_{2i|1i}^{-1}$.

We are interested in estimating the finite population total of y, denoted by $Y = \sum_{i=1}^{N} y_i$. When the study variable y is observed in the second-phase sample, the population total Y can be estimated using the two-phase regression estimator defined by

$$\hat{Y}_{tp,reg} = \hat{Y}_2 + (\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_2)'\hat{\boldsymbol{\beta}} \ , \tag{1}$$

where $\hat{\mathbf{X}}_1 = \sum_{i \in A_1} w_{1i}\mathbf{x}_i$, $(\hat{\mathbf{X}}_2, \hat{Y}_2) = \sum_{i \in A_2} w_{1i}w_{2i|1i}(\mathbf{x}_i, y_i)$, and $\hat{\boldsymbol{\beta}}$ is obtained using the observations from the second-phase sample. Note that $\boldsymbol{\beta}$ is a column vector of dimension p and notation $\mathbf{x}'$ denotes the transpose of $\mathbf{x}$. To study the asymptotic properties of the two-phase regression estimator in (1), we assume a sequence of finite populations and samples defined in Fuller (2009) with bounded fourth moments of (x_i, y_i). Under some regularity conditions, we can establish that

$$\begin{aligned} \hat{Y}_{tp,reg} &= \hat{Y}_2 + (\hat{X}_1 - \hat{X}_2)'\boldsymbol{\beta}_N + (\hat{X}_1 - \hat{X}_2)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) \\ &= \hat{Y}_2 + (\hat{X}_1 - \hat{X}_2)'\boldsymbol{\beta}_N + O_p(n_2^{-1}N), \end{aligned}$$

where $\boldsymbol{\beta}_N$ is the probability limit of $\hat{\boldsymbol{\beta}}$. Thus, the two-phase regression estimator $\hat{Y}_{tp,reg}$ is design-consistent for Y regardless of the form of $\hat{\boldsymbol{\beta}}$.

3. Proposed Method

In this section, we present a new approach for mass imputation under two-phase sampling. Mass imputation estimator for the population total Y is composed of the observed y values of the second-phase sample and the imputed values for the rest of the first-phase sample. Thus, a mass imputation estimator for population total using a regression model is written by

$$\hat{Y}_{imp} = \sum_{i \in A_2} w_{1i} y_i + \sum_{i \in \tilde{A}_2} w_{1i} \hat{y}_i, \tag{2}$$

where $\tilde{A}_2 = A_1 \bigcap A_2^c$, $\hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ is to be determined later. The first component is a weighted sum of the real observations in A_2 and the second term is a weighted sum of imputed values in $\tilde{A}_2$.

Our goal is to find a sufficient condition that makes the imputation estimator (2) algebraically equivalent to the two-phase regression estimator in (1).

Lemma 1. *If $\hat{\boldsymbol{\beta}}$ satisfies*

$$\sum_{i \in A_2} w_{1i}(w_{2i|1i} - 1)(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}) = 0, \tag{3}$$

then the mass imputation estimator $\hat{Y}_{imp}$ in (2) is algebraically equivalent to the two-phase regression estimator defined in (1).

Proof. Condition (3) can be expressed as

$$\sum_{i \in A_2} w_{1i} w_{i2|1}(y_i - \hat{y}_i) = \sum_{i \in A_2} w_{1i}(y_i - \hat{y}_i).$$

Thus,

$$\begin{aligned}
\hat{Y}_{imp} &= \sum_{i\in A_2} w_{1i} y_i + \sum_{i\in \tilde{A}_2} w_{1i}\hat{y}_i \\
&= \sum_{i\in A_1} w_{1i}\hat{y}_i + \sum_{i\in A_2} w_{1i}(y_i - \hat{y}_i) \\
&= \sum_{i\in A_1} w_{1i}\hat{y}_i + \sum_{i\in A_2} w_{1i} w_{i2|1}(y_i - \hat{y}_i) \\
&= \hat{Y}_2 + (\hat{X}_1 - \hat{X}_2)'\hat{\boldsymbol{\beta}},
\end{aligned}$$

which establishes the equivalence between the mass imputation estimator and the two-phase regression estimator. ■

Note that condition (3) is satisfied if $\hat{\boldsymbol{\beta}}$ is of the form

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\in A_2} w_{1i}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i\in A_2} w_{1i}\mathbf{x}_i y_i \tag{4}$$

and $w_{2i|1i}-1$ is included in the column space of $\mathbf{x}_i$, which means that $w_{2i|1i}-1 = \mathbf{x}_i'\mathbf{a}$ for some p-dimensional vector $\mathbf{a}$. Under condition (3), the mass imputation estimator (2) is also design-consistent for the population total Y. Condition (3) is similar in spirit to internal bias calibration (IBC) condition of Firth & Bennett (1998).

The mass imputation using $\hat{y}_i$ as the imputed values for y_i can be called deterministic imputation. We can also apply the idea of fractional imputation (Fuller & Kim, 2005) for mass imputation. To do this, we can write

$$\hat{Y}_{FI} = \sum_{i\in A_2} w_{1i} y_i + \sum_{i\in \tilde{A}_2} w_{1i}(\hat{y}_i + \sum_{j\in A_2} w_{ij}^* \hat{e}_j), \tag{5}$$

where $\hat{e}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ and w_{ij}^* is the fractional weight assigned to $\hat{e}_j$ in unit $i \in \tilde{A}_2$.

If we choose

$$w_{ij}^{*} = w_{1j}(w_{2j|1j} - 1) / \sum_{j \in A_2} [w_{1j}(w_{2j|1j} - 1)],$$

then, by (3), we have $\sum_{j \in A_2} w_{ij}^{*} \hat{e}_j = 0$ and (5) is algebraically equivalent to the mass imputation estimator (2). By including the residual terms in the fractional imputation, we can estimate other parameters such as percentiles or distribution functions. However, it leads to have an aggregated dataset as it requires to impute n_2 values for one unit, so the dataset can be huge after the fractional imputation.

Note that we can express (5) as

$$\hat{Y}_{FEFI} = \sum_{i \in A_2} w_{1i} y_i + \sum_{i \in \tilde{A}_2} w_{1i} \sum_{j \in A_2} w_{ij}^{*} y_{ij}^{*}, \tag{6}$$

where $y_{ij}^{*} = \hat{y}_i + \hat{e}_j$. Because (6) uses all possible imputed values for imputation, it can be called fully efficient fractional imputation (FEFI) estimator (Fuller & Kim, 2005).

4. Replication Variance Estimation

In this section, we consider replication variance estimation of the mass imputation estimator in (2). Let the replicate variance estimator for the first-phase sample estimator of total be

$$\hat{V}_1(\hat{T}_1) = \textstyle\sum_{k=1}^{L} c_k (\hat{T}_1^{(k)} - \hat{T}_1)^2 \tag{7}$$

where $\hat{T}_1^{(k)} = \sum_{i \in A_1} w_{1i}^{(k)} y_i$ is the k-th replicate of estimated total $\hat{T}_1 = \sum_{i \in A_1} w_{1i} y_i$, L is the number of replications, and c_k is the replication factor.

The jackknife variance estimator for the mass imputation estimator using

the second-phase sample can be written as

$$\hat{V}(\hat{Y}_{imp}) = \sum_{k=1}^{L} c_k (\hat{Y}_{imp}^{(k)} - \hat{Y}_{imp})^2, \tag{8}$$

where

$$\hat{Y}_{imp}^{(k)} = \sum_{i \in A_2} w_{1i}^{(k)} y_i + \sum_{i \in \tilde{A}_2} w_{1i}^{(k)} \mathbf{x}_i' \hat{\boldsymbol{\beta}}^{(k)} \tag{9}$$

and $\hat{\boldsymbol{\beta}}^{(k)} = (\sum_{i \in A_2} w_{1i}^{(k)} \mathbf{x}_i \mathbf{x}_i')^{-1} \sum_{i \in A_2} w_{1i}^{(k)} \mathbf{x}_i y_i$. Note that $\hat{Y}_{imp}^{(k)}$ is the k^{th} replicate of $\hat{Y}_{imp}$ using the k^{th} replicated weight of w_{1i}. We can show that the jackknife variance estimator is consistent for the variance of the mass imputation estimator. For simplicity we now assume that a Poisson sampling is used in the second-phase. Fuller (1998) argued that Poisson sampling for second-phase sample is a good approximation and has little impact on the variance estimation of the mean under two-phase sampling.

Theorem 1. *Assume that a finite population of $z_i = (x_i, y_i)$ is a random sample from an infinite population with $4 + \delta$, $\delta > 0$, moments and $E(\pi_{2i|1i}) = \kappa_i$. Assume that $w_{2i|1i} - 1$ is in the column space of $\mathbf{x}_i$ in computing $\hat{\boldsymbol{\beta}}$ in (4). Denote $n_1 = |A_1|$, $n_2 = |A_2|$ and $\hat{T}_{1z} = \sum_{i \in A_1} w_{1i} z_i$ is a total estimators of variable z obtained from the first-phase. Assume that*

$$E[|\hat{T}_{1z} - T_{1z}|^2 | \mathcal{F}_N] = O(n_1^{-1} N^2),$$

and

$$V(\hat{T}_{1y} | \mathcal{F}_N) \leq K_M V(\hat{T}_{y,SRS} | \mathcal{F}_N), \tag{10}$$

for a fixed K_M, where $V(\hat{T}_{y,SRS} | \mathcal{F}_N)$ is the variance of the Horvits-Thompson estimator based on a simple random sample of size n_1. Assume that the variance of a linear estimator of a total is a quadratic function of y and assume that

$$n_1 N^{-2} V(\sum_{i \in A_1} w_{1i} y_i | \mathcal{F}_N) = \sum_{i=1}^{N} \sum_{j=1}^{N} \Omega_{ij} y_i y_j \tag{11}$$

where the coefficients Ω_{ij} *satisfy*

$$\sum_{i=1}^{N} |\Omega_{ij}| = O(N^{-1}). \tag{12}$$

Let $\hat{V}_1(\hat{T}_1)$ *be the first-phase sample replicate estimator of the variance of* $\hat{T}_1$ *given in (7) and satisfy*

$$E\left[\left(\frac{\hat{V}_1(\hat{T}_1)}{V(\hat{T}_1|\mathcal{F}_N)} - 1\right)^2 \middle| \mathcal{F}_N\right] = o(1) \tag{13}$$

for any y *with bounded fourth moments. Assume that the replicates for the first-phase sample estimator of a total,* $\hat{T}_1$*, satisfy*

$$\max_k E[\{c_k(\hat{T}_1^{(k)} - \hat{T}_1)^2\}^2 | \mathcal{F}_N] < K_T L^{-2} [V(\hat{T}_1|\mathcal{F}_N)]^2 \tag{14}$$

for some constant K_T*, uniformly in* N*. Also, assume that*

$$\max_k c_k = O(1). \tag{15}$$

Then, the jackknife variance estimator of form (8) satisfies

$$\hat{V}_{JK}(\hat{Y}_{imp}) = V(\hat{Y}_{imp} \mid \mathcal{F}_N) - \textstyle\sum_{i=1}^{N} \kappa_i^{-1}(1-\kappa_i)e_i^2 + o_p(n_2^{-1}N^2), \tag{16}$$

where $e_i = y_i - \bar{Y}_N - (x_i - \bar{X}_N)\boldsymbol{\beta}_N$.

The proof of Theorem 1 is presented in Appendix. Assumptions (10)-(12) are the regularity conditions for the variance of the Horvitz-Thompson estimator in the first-phase sample. Assumption (13) implies that the first-phase replication variance estimator is consistent and assumption (14) implies that all components of the replication variance estimator are of the same order, uniformly contribute and no component dominates the others. Assumption (15) is about the order of the replication factor and it is satisfied for the Jackknife variance estimator.

These assumptions are quite standard in two-phase sampling literature and can be found in Kim et al. (2006).

From (16), the bias of $\hat{V}_{JK}(\hat{Y}_{imp})$ is $O(N)$ and it can be estimated unbiasedly by $\sum_{i\in A_2} w_{1i}\pi_{2i|1i}^{-2}(1-\pi_{2i|1i})\hat{e}_i^2$, where $\hat{e}_i = y_i - x_i'\hat{\boldsymbol{\beta}}$. The bias term in (16) is negligible if the first-phase sampling rate, n_1/N, is negligible. Then, the replicate variance estimator in (8) can be used directly for the variance of mass imputation estimator under two-phase sampling.

We now consider variance estimation of the FEFI estimator in (6). The k-th replicate of the FEFI estimator is

$$\hat{Y}_{FEFI}^{(k)} = \sum_{i\in A_2} w_{1i}^{(k)} y_i + \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \sum_{j\in A_2} w_{ij}^{*(k)} y_{ij}^*, \tag{17}$$

where

$$w_{ij}^{*(k)} = w_{1j}^{(k)}(w_{2j|1j}-1) / \sum_{j\in A_2} [w_{1j}^{(k)}(w_{2j|1j}-1)] \tag{18}$$

is the k-th replicate of fractional weight. Note that the imputed values are not changed for each replication, only the fractional weights are changed. The following theorem provides the asymptotic property of the replicate variance estimator of the FEFI estimator.

Theorem 2. *Assume that*

$$\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}} = O_p(n_2^{-1}). \tag{19}$$

Then, the jackknife variance estimator of the FEFI estimator, which has a form of $\hat{V}_{FEFI} = \sum_{k=1}^{L} c_k(\hat{Y}_{FEFI}^{(k)} - \hat{Y}_{FEFI})^2$, *satisfies*

$$\hat{V}_{FEFI} = V(\hat{Y}_{FEFI}) - \sum_{i=1}^{N} \kappa_i^{-1}(1-\kappa_i)e_i^2 + o_p(n_2^{-1}N^2), \tag{20}$$

where κ_i *and* e_i *are defined in the Theorem 1.*

The proof of Theorem 2 is presented in Appendix.

Theorem 2 establishes the asymptotic equivalence of the FEFI variance estimator using (17) and the variance estimator (8) for mass imputation. By Theorem 2, the proposed FEFI variance estimator is design-consistent under two-phase sampling.

5. Categorical Data Mass Imputation

We now extend the proposed mass imputation method to handle categorical data. Note that the regression imputation using $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$ does not necessary produce imputed values belonging to the range of y values and cannot be used directly to handle categorical y-values. To discuss the problem, let y take values on $\{1, 2, \cdots, K\}$. We assume a "working" model for $P(Y = l \mid \mathbf{x})$:

$$P(Y = l \mid \mathbf{x}) = p_l(\mathbf{x}; \boldsymbol{\beta})$$

with $\sum_{l=1}^{K} p_l(\mathbf{x}; \boldsymbol{\beta}) = 1$. For example, for binary y, we may use a logistic regression model

$$P(Y = 1 \mid \mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}.$$

Now, suppose that we are interested in estimating $\theta_l = P(Y = l)$ from the survey data. The sampling design is the same two-phase sampling in Section 2. The only difference is that the study variabe y is categorical. The two-phase regression estimator of θ_l can be defined as

$$\hat{\theta}_{l,tp,reg} = \sum_{i \in A_1} w_{1i} p_l(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) + \sum_{i \in A_2} w_{1i}\pi_{2i|1i}^{-1} \left\{ I(y_i = l) - p_l(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) \right\}, \qquad (21)$$

for some $\hat{\boldsymbol{\beta}}$. Note that $\hat{\theta}_{l,tp,reg}$ is design-consistent for θ_l, regardless of whether the working model is true or not.

Similarly to (2), we can construct mass imputation estimator of θ_l as follows.

$$\hat{\theta}_{l,I,reg} = \sum_{i \in A_2} w_{1i} I(y_i = l) + \sum_{i \in \tilde{A}_2} w_{1i} p_l(\mathbf{x}_i; \hat{\boldsymbol{\beta}}). \tag{22}$$

Two estimators, (21) and (22), are algebraically equivalent if the following condition holds:

$$\sum_{i \in A_2} w_{1i} \left(\pi_{2i|1i}^{-1} - 1 \right) \left\{ I(y_i = l) - p_l(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) \right\} = 0. \tag{23}$$

More generally, we can use

$$\sum_{i \in A_2} w_{1i} \left(\pi_{2i|1i}^{-1} - 1 \right) S(\hat{\boldsymbol{\beta}}; \mathbf{x}_i, y_i) = 0 \tag{24}$$

as the pseudo score equation for model parameter $\boldsymbol{\beta}$ in the working model $f(y \mid \mathbf{x}; \boldsymbol{\beta})$, where $S(\boldsymbol{\beta}; \mathbf{x}, y) = \partial \log f(y \mid \mathbf{x}; \boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ is the score function of $\boldsymbol{\beta}$ in the parametric working model $f(y \mid \mathbf{x}; \boldsymbol{\beta})$. Condition (24) is the IBC condition of Firth & Bennett (1998) for the parametric model approach in two-phase sampling. It is also related to doubly robust imputation in the context of missing data imputation (Kim & Haziza, 2014).

The mass imputation estimator in (22) is design-consistent under (23) but the imputed value $\hat{y}_i = p_l(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ is not necessarily categorical. To create categorical imputed values and achieve design consistency, we can apply parametric fractional imputation of Kim (2011) adopted to two-phase sampling. In fractional imputation for categorical data, for each unit $i \in \tilde{A}_2$, we create K values of (y_{ij}^*, w_{ij}^*), for $j = 1, \cdots, K$, where y_{ij}^* is the j-th imputed value of y_i and w_{ij}^* is the fractional weight assigned to y_{ij}^* satisfying $\sum_{j=1}^K w_{ij}^* = 1$. In the proposed method, we use $y_{ij}^* = j$ and $w_{ij}^* = p_j(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ satisfies (23). Using the

fractionally imputed data, we can estimate θ_l by

$$\hat{\theta}_{l,FI} = \sum_{i \in A_2} w_{1i} I(y_i = l) + \sum_{i \in \tilde{A}_2} \sum_{j=1}^{K} w_{1i} w_{ij}^* I(y_{ij}^* = l). \tag{25}$$

Note that, since $y_{ij}^* = j$, the fractionally imputed estimator (25) with $w_{ij}^* = p_j(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ is algebraically equivalent to the mass imputation estimator in (22). Because all the imputed values are categorical, the fractionally imputed dataset can be used for estimating many different parameters such as proportions.

For variance estimation, we develop replication method for fractional imputation. In fractional imputation, only the fractional weights are replicated and the imputed values are not changed for each replication. To construct the k-th replicate of $w_{ij}^* = p_j(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$, we first compute $\hat{\boldsymbol{\beta}}^{(k)}$, the k-th replicate of $\hat{\boldsymbol{\beta}}$, by solving (24) with w_{1i} replaced by $w_{1i}^{(k)}$. The replication fractional weights are given by $w_{ij}^{*(k)} = p_j(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(k)})$.

Using the replicated fractional weights, the k-th replicate of $\hat{\theta}_{l,FI}$ is obtained as

$$\hat{\theta}_{l,FI}^{(k)} = \sum_{i \in A_2} w_{1i}^{(k)} I(y_i = l) + \sum_{i \in \tilde{A}_2} \sum_{j=1}^{K} w_{i1}^{(k)} w_{ij}^{*(k)} I(y_{ij}^* = l)$$

and applied to (7) to compute the variance estimator of $\hat{\theta}_{l,FI}$.

6. An Illustrative Example

In this section, we use a toy example to illustrate the mass imputation estimator and its variance estimation under two-phase sampling. The data are tabulated in Table 1, which is modified from Table 3.6 of Fuller (2009).

Suppose that the data were obtained by two-phase sampling where the first-phase sample contains of 26 elements in two strata ($h = 1, 2$) and the second-phase sample contains 14 elements in three groups ($g = 1, 2, 3$). Simple random

Table 1: Data for the illustrative example

Element ID	Phase 1 Stratum	Phase 1 Weight	Phase 2 Group	Phase 2 Weight	y
1	1	300	1		
2	1	300	1	600	7.2
3	1	300	1	600	6.8
4	1	300	2		
5	1	300	2	525	8.6
6	1	300	2		
7	1	300	2	525	8.0
8	1	300	3	550	6.2
9	1	300	3	550	6.5
10	1	300	3		
11	1	300	3	550	5.9
12	1	300	3		
13	2	200	1		
14	2	200	1	400	5.2
15	2	200	1		
16	2	200	1	400	5.5
17	2	200	1		
18	2	200	2		
19	2	200	2	350	5.7
20	2	200	2	350	6.3
21	2	200	3		
22	2	200	3		
23	2	200	3	366.7	5.3
24	2	200	3		
25	2	200	3	366.7	4.9
26	2	200	3	366.7	5.0

sampling is used in each group for selecting the second-phase sample and the second-phase sampling rate is four-in-eight, four-in-seven, and six-in-eleven, for groups 1, 2, and 3, respectively. Weights for both phases are also presented in Table 1.

Let $\mathbf{x}_i$ be the vector of covariate variable, x_{gi}, which is an indicator variable having either 1 if element i is in group g, or 0 otherwise. A study variable y is continuous and observed only in the second-phase sample (A_2), whereas it is missing in the remaining part of the first-phase sample ($\tilde{A}_2$).

We are interested in estimating the population mean of Y, $\theta = N^{-1}\sum_{i=1}^{N} y_i$. In order to obtain the mass imputation estimator for θ written by

$$\hat{\theta}_{imp} = N^{-1}\left(\sum_{i\in A_2} w_{1i}y_i + \sum_{i\in \tilde{A}_2} w_{1i}\hat{y}_i\right), \tag{26}$$

the missing values of y_i in $\tilde{A}_2$ should be filled in by imputed values, which are $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$. Using equation (4), we can calculate $\hat{\boldsymbol{\beta}}$ from the second-phase sample given by

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\in A_2} w_{1i}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i\in A_2} w_{1i}\mathbf{x}_i y_i = (6.34, 7.38, 5.75).$$

Then the y_i's in $\tilde{A}_2$ can be replaced by imputed values, $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$, which are tabulated in Table 2. Note that y^* in Table 2 is defined as

$$y_i^* = \begin{cases} y_i & \text{if } i \in A_2 \\ \hat{y}_i & \text{if } i \in \tilde{A}_2. \end{cases}$$

Then, we can obtain the mass imputation estimator by (26), which is

$$\hat{\theta}_{imp} = N^{-1}\left(\sum_{i\in A_2} w_{1i}y_i + \sum_{i\in \tilde{A}_2} w_{1i}\hat{y}_i\right) = 6.382.$$

Note that only the first-phase sample weights are used for computation of the mass imputation estimate. On the other hand, the direct expansion estimator (DEE) of θ is

$$\hat{\theta}_{DEE} = \left(\sum_{i\in A_2} w_{2i}\right)^{-1}\sum_{i\in A_2} w_{2i}y_i = 6.369.$$

Variance of $\hat{\theta}_{imp}$ can be estimated using Jackknife variance estimator given in (8), where the k^{th} replicates of $\hat{\theta}_{imp}$ are calculated by (9). That is, leave-

Table 2: Data for the illustrative example with imputed values

Element ID	Phase 1 Stratum	Phase 1 Weight	Phase 2 Group	y^*
1	1	300	1	6.34
2	1	300	1	7.20
3	1	300	1	6.80
4	1	300	2	7.38
5	1	300	2	8.60
6	1	300	2	7.38
7	1	300	2	8.00
8	1	300	3	6.20
9	1	300	3	6.50
10	1	300	3	5.75
11	1	300	3	5.90
12	1	300	3	5.75
13	2	200	1	6.34
14	2	200	1	5.20
15	2	200	1	6.34
16	2	200	1	5.50
17	2	200	1	6.34
18	2	200	2	7.38
19	2	200	2	5.70
20	2	200	2	6.30
21	2	200	3	5.75
22	2	200	3	5.75
23	2	200	3	5.30
24	2	200	3	5.75
25	2	200	3	4.90
26	2	200	3	5.00

one-out procedure is repeated for $n_1 = |A_1|$ times and the $\hat{\theta}_{imp}^{(k)}$ is computed for each replicate, which is

$$\hat{\theta}_{imp}^{(k)} = N^{-1} \left(\sum_{i \in A_2} w_{1k}^{(k)} y_i + \sum_{i \in \tilde{A}_2} w_{1i}^{(k)} \hat{y}_i^{(k)} \right),$$

where $\hat{y}_i^{(k)} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)} = \left(\sum_{i \in A_2} w_{1i}^{(k)} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in A_2} w_{1i}^{(k)} \mathbf{x}_i y_i$. Note that, for each replicate, the first-phase sample weights are changed.

Then the Jackknife variance estimate of the mass imputation estimator is

obtained by

$$\hat{V}_{JK}(\hat{\theta}_{imp}) = \sum_{k=1}^{n_1} c_k(\hat{\theta}_{imp}^{(k)} - \hat{\theta}_{imp})^2 = 0.057.$$

The variance estimate of the DEE estimator is 0.075.

7. Simulation Study

A limited simulation study is performed to study the finite sample performance of the proposed mass imputation estimator and the replication variance estimator.

We consider two types of study variable $\mathbf{Y} = (Y_1, Y_2)$, where Y_1 is continuous and Y_2 is categorical.

1. Two artificial finite populations for Y_1 is considered: linear model $y_{1i} = 0.8 + 0.5x_i + z_i + e_i$ where $x_i \sim N(2,1)$, $e_i \sim N(0,1)$ and ratio model $y_{1i} = 0.3x_i + z_i + u_i$ where $x_i \sim N(2,1)$ and $u_i \sim N(0, |x_i|)$. For both models, $z_i \sim \exp(1) + 2$ is used as the size measure for the unequal probability sampling in the second-phase sampling.

2 Categorical variable of Y_2: we consider a binary variable of $Y_2 \sim \text{Bernoulli}(p_i)$ where $\text{logit}(p_i) = -1.8 + x_i + 0.4y_{1i}$ using the y_{1i} values generated from either of the artificial finite populations.

A finite population of size $N = 100,000$ is generated from each model. From each of the finite population, first-phase samples of size $n_1 = 500$ are independently generated by simple random sampling. Then, second-phase samples of size $n_2 = 80$ are selected from the first-phase sample using the three different sampling designs as follows:

1) Simple random sampling without replacement of size $n_2 = 80$.

2) Poisson sampling:

Define δ_i for selecting unit i as $\delta_i | I_i = 1 \sim \text{Bernoulli}(\pi_{2i|1i})$, where I_i is

an indicator variable having 1 if unit i is included in the first-phase, and having 0 otherwise. We use the conditional first-order inclusion probability of second-phase sample as $\pi_{2i|1i} = n_2 z_i / \sum_{i \in A_1} z_i$, which depends on the first-phase sample, where $n_2 = 80$.

3) Randomized systematic PPS sampling (RSPPS) of size $n_2 = 80$: We follow the procedure introduced in Thompson & Wu (2008).

   a. Arrange units in the first-phase sample in a random order.

   b. Denote $q_i = z_i / \sum_{i \in A_1} z_i$ and let $A_j = \sum_{i=1}^{j} n_2 q_i$ be the cumulative totals of $n_2 q_i$. Note that $A_0 = 0$ and we have the order of $0 = A_0 < A_1 < \cdots < A_{n_1} = n_2$.

   c. Let u be a uniform random number over $[0, 1]$.

   d. Units with indices j satisfying $A_{j-1} \leq u+k < A_j$ for $k = 0, 1, \cdots, n_2 - 1$ to be included in the second-phase sample.

   Note that the first-order inclusion probability of second-phase sample $\pi_{2i|1i}$ obtained by the randomized systematic PPS sampling procedure satisfies $\pi_{2i|1i} = n_2 z_i / \sum_{i \in A_1} z_i$, for $i \in A_1$.

Once the two-phase samples are generated, we compute four estimators for the population mean $\theta = N^{-1} \sum_{i=1}^{N} y_i$; 1) direct estimator, 2) classical two-phase regression estimator, 3) classical two-phase regression estimator including $\pi_{2i|1i} - 1$ as a covariate, and 4) mass imputation estimator. These estimates are defined as follows:

1. Direct estimator: $\hat{\theta}_{dir} = \sum_{i \in A_2} w_{1i} w_{2i|1i} y_i / \sum_{i \in A_2} (w_{1i} w_{2i|1i})$.

2. Two-phase regression estimator: $\hat{\theta}_{tp,reg} = \bar{y}_2 + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \hat{\boldsymbol{\beta}}$,

where

$$\begin{aligned}
\bar{\mathbf{x}}_1 &= \textstyle\sum_{i\in A_1} w_{1i}\mathbf{x}_i / \sum_{i\in A_1} w_{1i}, \\
(\bar{\mathbf{x}}_2, \bar{y}_2) &= \textstyle\sum_{i\in A_2} w_{1i} w_{2i|1i}(\mathbf{x}_i, y_i) / \sum_{i\in A_2}(w_{1i} w_{2i|1i}), \\
\hat{\boldsymbol{\beta}} &= \left(\textstyle\sum_{i\in A_2} w_{1i} w_{2i|1i}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i\in A_2} w_{1i} w_{2i|1i}\mathbf{x}_i y_i,
\end{aligned}$$

and $\mathbf{x}_i = (1, x_i)'$.

3. Two-phase regression estimator including $\pi_{2i|1i}-1$ as a covariate: $\hat{\theta}_{tp,reg2} = \bar{y}_2+(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\hat{\boldsymbol{\beta}}$, where all estimators $(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2, \hat{\boldsymbol{\beta}})$ are defined as the same with estimators in $\hat{\theta}_{tp,reg}$ except for $\mathbf{x}_i = (1, \pi_{2i|1i}, x_i)'$.

4. Mass imputation estimator: $\hat{\theta}_{imp} = N^{-1}(\sum_{i\in A_2} w_{1i} y_i + \sum_{i\in \tilde{A}_2} w_{1i}\hat{y}_i)$, where $\hat{y}_i = \mathbf{x}_i^{'}\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} = \left(\sum_{i\in A_2} w_{1i}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i\in A_2} w_{1i}\mathbf{x}_i y_i$, and $\mathbf{x}_i = (1, \pi_{2i|1i}^{-1}, x_i)'$.

Further, we compute the proposed replication variance estimator for the mass imputation estimator. The replication variance estimator of the mass imputation estimator was computed using the replication number $L = n_1$. Since the first-phase sample is selected from simple random sampling of size n_1, the kth replicate weight is given by

$$w_{1i}^{(k)} = \begin{cases} w_{1i} n_1/(n_1-1) & \text{if } i \neq k \\ 0 & \text{otherwise,} \end{cases}$$

and the replication factor is $c_k = (1 - n_1/N)(1 - 1/n_1)$. This procedure was repeated 1,000 times and Monte Carlo bias and variance of the four estimators, Monte Carlo coverage rate of the mass imputation estimator and Monte Carlo mean and relative bias of the replication variance estimator are computed.

Table 3 and Table 4 present the Monte Carlo bias and variance of the four estimators for continuous case (Y_1) and categorical case (Y_2), respectively, and

Table 3: Continuous Case: Monte Carlo bias and Monte Carlo variance of the four estimators: Direct estimator ($\hat{\theta}_{dir}$); Two-phase regression estimator ($\hat{\theta}_{tp,reg}$); Two-phase regression estimator with extended covariates ($\hat{\theta}_{tp,reg2}$); Mass imputation estimator ($\hat{\theta}_{imp}$)

Population	Second-phase Sampling	Estimator	Bias	Variance
Linear	SRS	$\hat{\theta}_{dir}$	0.00	0.029
		$\hat{\theta}_{tp,reg}$	0.00	0.026
		$\hat{\theta}_{tp,reg2}$	0.00	0.026
		$\hat{\theta}_{imp}$	0.00	0.026
	Poisson	$\hat{\theta}_{dir}$	0.00	0.027
		$\hat{\theta}_{tp,reg}$	0.00	0.020
		$\hat{\theta}_{tp,reg2}$	0.00	0.019
		$\hat{\theta}_{imp}$	0.00	0.017
	RSPPS	$\hat{\theta}_{dir}$	0.00	0.022
		$\hat{\theta}_{tp,reg}$	0.00	0.018
		$\hat{\theta}_{tp,reg2}$	0.00	0.017
		$\hat{\theta}_{imp}$	0.00	0.016
Ratio	SRS	$\hat{\theta}_{dir}$	0.00	0.040
		$\hat{\theta}_{tp,reg}$	0.00	0.038
		$\hat{\theta}_{tp,reg2}$	0.00	0.038
		$\hat{\theta}_{imp}$	0.00	0.038
	Poisson	$\hat{\theta}_{dir}$	0.00	0.047
		$\hat{\theta}_{tp,reg}$	0.00	0.038
		$\hat{\theta}_{tp,reg2}$	0.00	0.031
		$\hat{\theta}_{imp}$	0.00	0.030
	RSPPS	$\hat{\theta}_{dir}$	0.00	0.032
		$\hat{\theta}_{tp,reg}$	0.00	0.031
		$\hat{\theta}_{tp,reg2}$	0.00	0.030
		$\hat{\theta}_{imp}$	0.00	0.030

table 5 presents the Monte Carlo coverage rate of the mass imputation estimator for all cases. We can check that all four point estimators are unbiased for the population mean regardless of sampling design and specified population model type. The Monte Carlo coverage rates are about 95% for all cases. Moreover, the variances of classical two-phase regression estimator, two-phase regression estimator with extended $\mathbf{x}_i$ and mass imputation estimator for the sample selected using simple random sampling for both phases are the same, because $\pi_{2i|1i}^{-1}$ is constant under simple random sampling for the second-phase sampling.

Table 4: Categorical Case: Monte Carlo bias and Monte Carlo variance of the four estimators: Direct estimator ($\hat{\theta}_{dir}$); Two-phase regression estimator ($\hat{\theta}_{tp,reg}$); Two-phase regression estimator with extended covariates ($\hat{\theta}_{tp,reg2}$); Mass imputation estimator ($\hat{\theta}_{imp}$)

Population	Second-phase Sampling	Estimator	Bias	Variance($\times 10^5$)
Linear	SRS	$\hat{\theta}_{dir}$	0.00	181
		$\hat{\theta}_{tp,reg}$	0.00	157
		$\hat{\theta}_{tp,reg2}$	0.00	157
		$\hat{\theta}_{imp}$	0.00	157
	Poisson	$\hat{\theta}_{dir}$	0.00	359
		$\hat{\theta}_{tp,reg}$	0.00	232
		$\hat{\theta}_{tp,reg2}$	0.00	206
		$\hat{\theta}_{imp}$	0.00	181
	RSPPS	$\hat{\theta}_{dir}$	0.00	256
		$\hat{\theta}_{tp,reg}$	0.00	198
		$\hat{\theta}_{tp,reg2}$	0.00	197
		$\hat{\theta}_{imp}$	0.00	184
Ratio	SRS	$\hat{\theta}_{dir}$	0.00	223
		$\hat{\theta}_{tp,reg}$	0.00	189
		$\hat{\theta}_{tp,reg2}$	0.00	189
		$\hat{\theta}_{imp}$	0.00	189
	Poisson	$\hat{\theta}_{dir}$	0.00	397
		$\hat{\theta}_{tp,reg}$	0.00	257
		$\hat{\theta}_{tp,reg2}$	0.00	246
		$\hat{\theta}_{imp}$	0.00	216
	RSPPS	$\hat{\theta}_{dir}$	0.00	289
		$\hat{\theta}_{tp,reg}$	0.00	234
		$\hat{\theta}_{tp,reg2}$	0.00	233
		$\hat{\theta}_{imp}$	0.00	216

For other designs, the mass imputation estimator has smaller variance compared with the classical two-phase regression estimator as the auxiliary variable used for the mass imputation estimator contains the additional information in $\pi_{2i|1i}^{-1}$. Because the mass imputation estimator is based on the augmented regression model, augmented by $\pi_{2i|1i}^{-1}$, it is more efficient in the sense of reducing the variance. The mass imputation estimator is slightly more efficient than the two-phase regression estimator with extended covariates. The mass imputation estimator uses only w_{1i} in computing $\hat{\boldsymbol{\beta}}$ while the two-phase regression estimator

Table 5: Monte Carlo coverage rate of the mass imputation estimator

Case	Population	Second-phase Sampling	Coverage Rate
Continuous	Linear	SRS	0.953
		Poisson	0.951
		RSPPS	0.949
	Ratio	SRS	0.951
		Poisson	0.950
		RSPPS	0.951
Categorical	Linear	SRS	0.948
		Poisson	0.949
		RSPPS	0.949
	Ratio	SRS	0.950
		Poisson	0.949
		RSPPS	0.951

uses $w_{1i}w_{2i|1i}$, which creates extra variability in the final estimation.

Table 6: Monte Carlo mean and relative bias (R.B.) of the replication variance estimator of the mass imputation estimator

Case	Population	Second-phase Sampling	Mean	R.B.
Continuous	Linear	SRS	0.026	0.001
		Poisson	0.017	0.003
		RSPPS	0.016	0.002
	Ratio	SRS	0.039	0.006
		Poisson	0.032	0.057
		RSPPS	0.031	0.017
Categorical	Linear	SRS	0.0015	0.002
		Poisson	0.0018	0.016
		RSPPS	0.0018	-0.005
	Ratio	SRS	0.0019	0.028
		Poisson	0.0022	0.048
		RSPPS	0.0022	0.033

Table 6 presents Monte Carlo mean and relative bias of the replication variance estimator of the mass imputation estimator. The relative bias of the variance estimator is obtained by dividing Monte Carlo bias of the variance estimator by the Monte Carlo variance of the point estimator. All Monte Carlo means of the replication variance estimators are consistent for the variance of the mass imputation estimator given in Table 3 and Table 4, and it leads to small relative

biases of the replication variance estimator in Table 6. This result supports the Theorem 1, as the bias term in (16) can be safely ignored since the first-phase sampling rate is $500/100,000 = 0.005$, which is small enough.

8. Conclusion

We treat two-phase sampling as a missing data problem and propose a mass imputation estimator that is equivalent to the two-phase regression estimator. The proposed replication variance estimation is simple to implement since it does not require computing replicates of the conditional inclusion probability for the second-phase sample, which may be complicated or impossible to compute depending on the sampling designs. The proposed method is further extended to categorical data mass imputation.

In mass imputation, to achieve design consistency, we have used an augmented regression model for imputation by including the inverse of the conditional inclusion probability for the second-phase sample into the covariates. Thus, the proposed method is applicable only when the conditional inclusion probabilities are available throughout the first-phase sample. If all the design information for the second-phase sampling is available at the imputation stage, then the conditional inclusion probability can be constructed for all the elements in the first-phase sample. If such design information is not available, the proposed method is not applicable. This is one limitation of our proposed method.

Acknowledgements

We are grateful to referees and the associate editor for comments that have helped to improve this paper. The research of the second author was partially supported by the U.S. National Science Foundation.

References

Breidt, F. J., McVey, A., & Fuller, W. A. (1996). Two-phase estimation by imputation. *Journal of the Indian Society of Agricultural Statistics*, *49*, 79–90.

Chipperfield, J., Chessman, J., & Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals. *Australian & New Zealand Journal of Statistics*, *54*, 223–238.

Fay, R. (1991). A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference, US Bureau of the census* (p. 440). volume 429.

Firth, D., & Bennett, K. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*, 3–21.

Fuller, W. A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, (pp. 1153–1164).

Fuller, W. A. (2009). *Sampling Statistics*. John Wiley & Sons.

Fuller, W. A., & Kim, J. K. (2005). Hot deck imputation for the response model. *Survey Methodology*, *31*, 139.

Hidiroglou, M. (2001). Double sampling. *Survey methodology*, *27*, 143–154.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, *98*, 119–132.

Kim, J. K., & Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, *24*, 375–394.

Kim, J. K., Navarro, A., & Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American statistical association*, *101*, 312–320.

Kim, J. K., & Rao, J. N. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, *99*, 85–100.

Legg, J. C., & Fuller, W. A. (2009). Two-phase sampling. *Handbook of statistics*, *29*, 55–70.

Moore, R., & Robbins, N. (2004). A study of mass imputation in small-area estimation. In *Joint Statistical Meeting, Toronto, Canada*.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, *33*, 101–116.

Rao, J. N., & Sitter, R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, *82*, 453–460.

Thompson, M. E., & Wu, C. (2008). Simulation-based randomized systematic pps sampling under substitution of units. *Survey Methodology*, *34*, 3.

Appendix A Proof of Theorem 1

Proof. By Lemma 1, we have

$$\hat{Y}_{imp} \quad = \quad \hat{Y}_2 + (\hat{X}_1 - \hat{X}_2)'\hat{\boldsymbol{\beta}}.$$

Since we assume that $w_{2i|1i} - 1$ is in the column space of $\mathbf{x}_i$, we have

$$\sum_{i \in A_2} w_{1i}^{(k)}(w_{2i|1i} - 1)(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{(k)}) = 0, \tag{A.1}$$

where $w_{1i}^{(k)}$ is a replicate weight for the first-phase sample for unit i. It follows from (A.1) that

$$\hat{Y}_{imp}^{(k)} = \hat{Y}_2^{(k)} + (\hat{X}_1^{(k)} - \hat{X}_2^{(k)})'\hat{\boldsymbol{\beta}}^{(k)},$$

where

$$\hat{\boldsymbol{\beta}}^{(k)} = \left(\sum_{i \in A_2} w_{1i}^{(k)} \mathbf{x}_i \mathbf{x}_i^{'} \right)^{-1} \sum_{i \in A_2} w_{1i}^{(k)} \mathbf{x}_i y_i$$

and $(\hat{Y}_2^{(k)}, \hat{X}_2^{(k)})$ are computed from the second-phase replicate using $w_{1i}^{(k)}$. Let a_i be the indicator function of the inclusion for the second-phase sample such that $a_i = 1$ if unit i is selected in A_2 and $a_i = 0$ otherwise. Using defined indicator variable for the second-phase sample, a_i, we can write

$$\left(\hat{X}_1^{(k)}, \hat{X}_2^{(k)}, \hat{Y}_2^{(k)} \right) = \sum_{i \in A_1} w_{1i}^{(k)} (x_i, \pi_{2i|1i}^{-1} a_i x_i, \pi_{2i|1i}^{-1} a_i y_i)$$

and

$$\hat{\boldsymbol{\beta}}^{(k)} = \left(\sum_{i \in A_1} w_{1i}^{(k)} a_i \mathbf{x}_i \mathbf{x}_i^{'} \right)^{-1} \sum_{i \in A_1} w_{1i}^{(k)} a_i \mathbf{x}_i y_i.$$

Note that, by assumption (14) and (15),

$$\begin{aligned} c_k^{1/2}\left(\hat{X}_1^{(k)} - \hat{X}_1\right) &= O_p(n_1^{-1/2}NL^{-1/2}) \\ c_k^{1/2}\left(\hat{X}_2^{(k)} - \hat{X}_2, \hat{Y}_2^{(k)} - \hat{Y}_2\right) &= O_p(n_2^{-1/2}NL^{-1/2}). \end{aligned}$$

Also, it can be shown that

$$\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}} + O_p(n_2^{-1/2}L^{-1/2}).$$

Next, we write the $\hat{Y}_{imp}^{(k)} - \hat{Y}_{imp}$ as

$$\begin{aligned} \hat{Y}_{imp}^{(k)} - \hat{Y}_{imp} &= \hat{Y}_2^{(k)} + \left(\hat{X}_1^{(k)} - \hat{X}_2^{(k)}\right)' \hat{\boldsymbol{\beta}}^{(k)} - \hat{Y}_2 - (\hat{X}_1 - \hat{X}_2)'\hat{\boldsymbol{\beta}} \\ &= \hat{Y}_2^{(k)} - \hat{Y}_2 + \left(\hat{X}_1^{(k)} - \hat{X}_1\right)' \left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right) - \left(\hat{X}_2^{(k)} - \hat{X}_2\right)' \left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right) \\ &+ \left(\hat{X}_1^{(k)} - \hat{X}_1\right)' \hat{\boldsymbol{\beta}} - \left(\hat{X}_2^{(k)} - \hat{X}_2\right)' \hat{\boldsymbol{\beta}} + \left(\hat{X}_1 - \hat{X}_2\right)' \left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right). \end{aligned}$$

Since

$$\begin{aligned} \left(\hat{X}_1^{(k)} - \hat{X}_1\right)' \left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right) &= O_p(n_1^{-1/2}L^{-1/2}N)O_p(n_2^{-1/2}L^{-1/2}) \\ &= O_p(n_1^{-1/2}n_2^{-1/2}L^{-1}N), \\ \left(\hat{X}_2^{(k)} - \hat{X}_2\right)' \left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right) &= O_p(n_2^{-1/2}L^{-1/2}N)O_p(n_2^{-1/2}L^{-1/2}) \\ &= O_p(n_2^{-1}L^{-1}N), \\ \left(\hat{X}_1^{(k)} - \hat{X}_1\right)' \hat{\boldsymbol{\beta}} &= \left(\hat{X}_1^{(k)} - \hat{X}_1\right)' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\right) + \left(\hat{X}_1^{(k)} - \hat{X}_1\right)' \boldsymbol{\beta}_N \\ &= \left(\hat{X}_1^{(k)} - \hat{X}_1\right)' \boldsymbol{\beta}_N + O_p(n_1^{-1/2}n_2^{-1/2}L^{-1/2}N), \\ \left(\hat{X}_2^{(k)} - \hat{X}_2\right)' \hat{\boldsymbol{\beta}} &= \left(\hat{X}_2^{(k)} - \hat{X}_2\right)' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\right) + \left(\hat{X}_2^{(k)} - \hat{X}_2\right)' \boldsymbol{\beta}_N \\ &= \left(\hat{X}_2^{(k)} - \hat{X}_2\right)' \boldsymbol{\beta}_N + O_p(n_2^{-1}L^{-1/2}N), \\ \left(\hat{X}_1 - \hat{X}_2\right)' \left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right) &= O_p(n_2^{-1/2}N)O_p(n_2^{-1/2}L^{-1/2}) \\ &= O_p(n_2^{-1}L^{-1/2}N), \end{aligned}$$

we have

$$\begin{aligned}\hat{Y}_{imp}^{(k)} - \hat{Y}_{imp} &= \hat{Y}_2^{(k)} - \hat{Y}_2 - \left(\hat{X}_1^{(k)} - \hat{X}_1\right)' \boldsymbol{\beta}_N - \left(\hat{X}_2^{(k)} - \hat{X}_2\right)' \boldsymbol{\beta}_N + O_p(n_2^{-1}L^{-1/2}N) \\ &:= \hat{e}_2^{(k)} - \hat{e}_2 - \left(\hat{X}_1^{(k)} - \hat{X}_1\right)' \boldsymbol{\beta}_N + O_p(n_2^{-1}L^{-1/2}N),\end{aligned}$$

where $e_i = y_i - \bar{Y}_N - (\mathbf{x}_i - \bar{X}_N)\boldsymbol{\beta}_N$. Hence, we can write

$$c_k^{1/2}(\hat{Y}_{imp}^{(k)} - \hat{Y}_{imp}) = c_k^{1/2}\left[\hat{e}_2^{(k)} - \hat{e}_2 - (\hat{X}_1^{(k)} - \hat{X}_1)'\boldsymbol{\beta}_N\right] + O_p(n_2^{-1}L^{-1/2}N). \tag{A.2}$$

by (15) and it follows from (A.2) that

$$\textstyle\sum_{k=1}^{L} c_k(\hat{Y}_{imp}^{(k)} - \hat{Y}_{imp})^2 = \sum_{k=1}^{L} c_k[\hat{e}_2^{(k)} - \hat{e}_2 + (\hat{X}_1^{(k)} - \hat{X}_1)'\boldsymbol{\beta}_N]^2 + O_p(n_2^{-3/2}N^2). \tag{A.3}$$

Order in (A.3) follows from that the order of the first term in (A.2) is $n_2^{-1/2}L^{-1/2}N$ by (14) and (10), and note that $O_p(n_2^{-3/2}N^2)$ is $o_p(n_2^{-1}N^2)$.

We now extend the definition of the second-phase sample indicator a_i that is defined throughout the population and this concept has been discussed by Fay (1991) and used by Kim et al. (2006). It means that a_i is defined for every unit in the population. Then, we can see the sample selection process as selecting the first-phase sample from the population of $(a_i, \mathbf{x}_i, a_i y_i)$ vectors. Hence, the main term of the right side of (A.3) can be written by

$$\begin{aligned}\hat{e}_2^{(k)} - \hat{e}_2 + (\hat{X}_1^{(k)} - \hat{X}_1)'\boldsymbol{\beta}_N &= \textstyle\sum_{i\in A_1}(w_{1i}^{(k)} - w_{1i})\kappa_i^{-1}a_i e_i + (\hat{X}_1^{(k)} - \hat{X}_1)'\boldsymbol{\beta}_N \\ &= \textstyle\sum_{i\in A_1}(w_{1i}^{(k)} - w_{1i})(\mathbf{x}_i'\boldsymbol{\beta}_N + \kappa_i^{-1}a_i e_i) \\ &\equiv \textstyle\sum_{i\in A_1}(w_{1i}^{(k)} - w_{1i})\eta_i,\end{aligned}$$

where $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}_N + \kappa_i^{-1}a_i e_i$. Thus, we can express the main tern of right side of (A.3) as a linear function form of η_i. Then, we are interested in the linearization

form for the variance estimation of $\hat{Y}_{imp}$.

Let $\tilde{Y}_{imp} = \sum_{i \in A_1} w_{1i}\eta_i$. By assumption (10) and (13), conditional on a_i, the replicate variance estimator of $\tilde{Y}_{imp}$ satisfies

$$\hat{V}(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N) = V(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N) + o_p(n_1^{-1}N^2). \tag{A.4}$$

It implies that the replicate variance estimator of $\tilde{Y}_{imp}$ is a consistent estimator of conditional variance of $\tilde{Y}_{imp}$. We now want to show that the replicate variance estimator is also consistent for the unconditional variance of $\tilde{Y}_{imp}$, $V(\tilde{Y}_{imp}|\mathcal{F}_N)$. The variance of the mass imputation estimator can be written by

$$V(\tilde{Y}_{imp}|\mathcal{F}_N) = E[V(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N] + V[E(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N]. \tag{A.5}$$

We next show that $\hat{V}(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)$ is a consistent estimator of the first term of (A.5). For this, we must show that $V(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)$ converges to $E[V(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N]$ and it is sufficient to demonstrate that

$$V(n_1 N^{-2} V(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N) = o(1).$$

Since we assumed that $a_i \sim$ Bernoulli$(\pi_{2i|1i})$, we have $Cov(a_i a_j, a_k a_l|\mathcal{F}_N) = \kappa_i\kappa_j(1-\kappa_i\kappa_j)$ where if $(i,j) = (k,l)$ or $(i,j) = (l,k)$ and $Cov(a_i a_j, a_k a_l|\mathcal{F}_N) = 0$

otherwise. By assumption (11) and (12), we have

$$\begin{aligned}
& V(n_1 N^{-2} V(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N) \\
= \; & V[n_1 N^{-2} V(\textstyle\sum_{i\in A_1} w_{1i}\eta_i|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N] \\
= \; & V[\textstyle\sum_{i=1}^N \sum_{j=1}^N \Omega_{ij} w_{1i}\eta_i w_{1i}\eta_j|\mathcal{F}_N] \\
= \; & \textstyle\sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N \sum_{l=1}^N \Omega_{ij}\Omega_{kl} Cov(\eta_i\eta_j, \eta_k\eta_l|\mathcal{F}_N) \\
= \; & 2\textstyle\sum_{i=1}^N \sum_{j=1}^N \Omega_{ij}^2 \kappa_i\kappa_j(1-\kappa_i\kappa_j)\eta_i^2\eta_i^2 \\
\leq \; & 2\max_{i,j}\left\{\kappa_i\kappa_j(1-\kappa_i\kappa_j)\eta_i^2\eta_i^2\right\}(\max_{i,j}|\Omega_{ij}|)\textstyle\sum_{i=1}^N\sum_{j=1}^N|\Omega_{ij}| \\
= \; & O(N^{-1}).
\end{aligned}$$

345 Therefore, $\hat{V}(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)$ is consistent for $E[V(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N]$.

Finally, the last term of (A.5) is

$$\begin{aligned}
V[E(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N] &= V[E(\textstyle\sum_{i\in A_1} w_{1i}\eta_i|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N] \\
&= V[\textstyle\sum_{i=1}^N \eta_i|\mathcal{F}_N] \\
&= \textstyle\sum_{i=1}^N\sum_{i=1}^N Cov(\eta_i, \eta_j|\mathcal{F}_N) \\
&= \textstyle\sum_{i=1}^N\sum_{j=1}^N Cov(\kappa_i^{-1}a_ie_i, \kappa_j^{-1}a_je_j) \\
&= \textstyle\sum_{i=1}^N \kappa_i^{-1}(1-\kappa_i)e_i^2.
\end{aligned}$$

Therefore, by combining all the results, we have

$$\hat{V}(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N) = V(\tilde{Y}_{imp}|\mathcal{F}_N) - \textstyle\sum_{i=1}^N \kappa_i^{-1}(1-\kappa_i)e_i^2 + o_p(n_2^{-1}N^2), \qquad \text{(A.6)}$$

which, by (A.4) and (A.6), establishes (16). ∎

Appendix B Proof of Theorem 2

Proof. First define $\tilde{Y}_{FEFI}^{(k)}$ as

$$\tilde{Y}_{FEFI}^{(k)} = \sum_{i\in A_2} w_{1i}^{(k)} y_i + \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \sum_{j\in A_2} w_{ij}^{*(k)} y_{ij}^{*(k)} \tag{B.1}$$

where $y_{ij}^{*(k)} = \hat{y}_i^{(k)} + \hat{e}_j^{(k)} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{(k)} + (y_j - \mathbf{x}_j'\hat{\boldsymbol{\beta}}^{(k)})$ is the k^{th} replicate of y_{ij}^*. Now,

$$\begin{aligned}
&\tilde{Y}_{FEFI}^{(k)} - \hat{Y}_{FEFI}^{(k)} \\
&= \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \sum_{j\in A_2} w_{ij}^{*(k)} y_{ij}^{*(k)} - \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \sum_{j\in A_2} w_{ij}^{*(k)} y_{ij}^{*} \\
&= \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{(k)} + \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \frac{\sum_{j\in A_2} w_{1i}^{(k)}(w_{2j|1j}-1)(y_j - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{(k)})}{\sum_{j\in A_2} w_{1i}^{(k)}(w_{2j|1j}-1)} \\
&- \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \frac{\sum_{j\in A_2} w_{1i}^{(k)}(w_{2j|1j}-1)(y_j - \mathbf{x}_i'\hat{\boldsymbol{\beta}})}{\sum_{j\in A_2} w_{1i}^{(k)}(w_{2j|1j}-1)} \\
&= \left[\sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \mathbf{x}_i - \frac{\sum_{i\in \tilde{A}_2} w_{1i}^{(k)}}{\sum_{j\in A_2} w_{1i}^{(k)}(w_{2j|1j}-1)} \sum_{j\in A_2} w_{1i}^{(k)}(w_{2j|1j}-1)\mathbf{x}_j\right]' (\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}) \\
&= \left[\sum_{i\in A_1} (1-a_i) w_{1i}^{(k)} \mathbf{x}_i - \frac{\sum_{i\in A_1} (1-a_i) w_{1i}^{(k)}}{\sum_{j\in A_1} a_i w_{1i}^{(k)}(w_{2j|1j}-1)} \sum_{j\in A_1} a_i w_{1i}^{(k)}(w_{2j|1j}-1)\mathbf{x}_j\right]' \times (\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}).
\end{aligned} \tag{B.2}$$

Define

$$\begin{aligned}
\hat{X}_{2c}^{(k)} &= \sum_{i\in A_1} (1-a_i) w_{1i}^{(k)} \mathbf{x}_i, \\
\hat{X}_{2}^{(k)} &= \sum_{i\in A_1} a_i w_{1i}^{(k)} (w_{2i|1i}-1)\mathbf{x}_i,
\end{aligned}$$

and

$$\hat{X}_{1c}^{(k)} = \sum_{i \in A_1} (1 - \pi_{2i|1i}) w_{1i}^{(k)} \mathbf{x}_i.$$

Further, let $\hat{N}_{2c}^{(k)}$, $\hat{N}_{1c}^{(k)}$ and $\hat{N}_{2}^{(k)}$ be defined similarly using 1 instead of $\mathbf{x}_i$. Then, (B.2) can be written by

$$\left[\hat{X}_{2c}^{(k)} - \hat{X}_{2}^{(k)} \hat{N}_{2c}^{(k)} / \hat{N}_{2}^{(k)}\right]' (\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}). \tag{B.3}$$

Note that

$$E(\hat{X}_{2c}^{(k)}) = \hat{X}_{1c}^{(k)} = E(\hat{X}_{2}^{(k)}) \tag{B.4}$$

and

$$E(\hat{N}_{2c}^{(k)}) = \hat{N}_{1c}^{(k)} = E(\hat{N}_{2}^{(k)}). \tag{B.5}$$

Also, we have $\hat{N}_{2c}^{(k)} = \hat{N}_{1c}^{(k)} + O_p(n_2^{-1/2}N)$ and $\hat{N}_{2}^{(k)} = \hat{N}_{1c}^{(k)} + O_p(n_2^{-1/2}N)$. Using the Taylor expansion, the ratio term in (B.3) can be expressed as

$$\begin{aligned}
\frac{\hat{N}_{2c}^{(k)}}{\hat{N}_{2}^{(k)}} &= [N^{-1}\hat{N}_{1c}^{(k)} + O_p(n_2^{-1/2})] \left[\frac{N}{\hat{N}_{1c}^{(k)}} - \frac{N^{-1}(\hat{N}_{2}^{(k)} - \hat{N}_{1c}^{(k)})}{(N^{-1}\hat{N}_{1c}^{(k)})^2} + o_p(n_2^{-1/2})\right] \\
&= \frac{\hat{N}_{1c}^{(k)}}{\hat{N}_{1c}^{(k)}} - \frac{\hat{N}_{1c}^{(k)}(\hat{N}_{2}^{(k)} - \hat{N}_{1c}^{(k)})}{(\hat{N}_{1c}^{(k)})^2} + o_p(n_2^{-1/2}) \\
&= 1 + O_p(n_2^{-1/2}),
\end{aligned}$$

based on (B.5). Hence, the first term in (B.3) can be expressed as

$$\begin{aligned}
\hat{X}_{2c}^{(k)} - \hat{X}_{2}^{(k)} \hat{N}_{2c}^{(k)} / \hat{N}_{2}^{(k)} &= [\hat{X}_{1c}^{(k)} + O_p(n_2^{-1/2}N)] - [1 + O_p(n_2^{-1/2})][\hat{X}_{1c}^{(k)} + O_p(n_2^{-1/2}N)] \\
&= [\hat{X}_{1c}^{(k)} + O_p(n_2^{-1/2}N)] - [\hat{X}_{1c}^{(k)} + O_p(n_2^{-1/2}N)] \\
&= O_p(n_2^{-1/2}N),
\end{aligned} \tag{B.6}$$

by (B.4). By combining (19) and (B.6), we have

$$\hat{Y}_{FEFI}^{(k)} = \tilde{Y}_{FEFI}^{(k)} + o_p(n_2^{-1}N). \tag{B.7}$$

With the choice of $w_{ij}^{*(k)}$ given by (18), we can show that $\tilde{Y}_{FEFI}^{(k)}$ in (B.1) is algebraically equivalent to the k^{th} replicate of $\hat{Y}_{imp}$ in (9). That is,

$$\begin{aligned}
\tilde{Y}_{FEFI}^{(k)} &= \textstyle\sum_{i\in A_2} w_{1i}^{(k)} y_i + \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \sum_{j\in A_2} w_{ij}^{*(k)} (\mathbf{x}_i' \hat{\boldsymbol{\beta}}^{(k)} + (y_j - \mathbf{x}_j' \hat{\boldsymbol{\beta}}^{(k)})) \\
&= \textstyle\sum_{i\in A_2} w_{1i}^{(k)} y_i + \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \mathbf{x}_i' \hat{\boldsymbol{\beta}}^{(k)} + \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \sum_{j\in A_2} w_{ij}^{*(k)} (y_j - \mathbf{x}_j' \hat{\boldsymbol{\beta}}^{(k)}) \\
&= \textstyle\sum_{i\in A_2} w_{1i}^{(k)} y_i + \sum_{i\in \tilde{A}_2} w_{1i}^{(k)} \mathbf{x}_i' \hat{\boldsymbol{\beta}}^{(k)}, \qquad \text{(B.8)}
\end{aligned}$$

where the last equality follows from $\sum_{j\in A_2} w_{ij}^{*(k)} \hat{e}_j^{(k)} = 0$. Since the FEFI estimator (6) is equivalent to the mass imputation estimator (2), we have

$$\begin{aligned}
\hat{Y}_{FEFI}^{(k)} - \hat{Y}_{FEFI} &= \tilde{Y}_{FEFI}^{(k)} - \hat{Y}_{FEFI} + \hat{Y}_{FEFI}^{(k)} - \tilde{Y}_{FEFI}^{(k)} \\
&= \hat{Y}_{imp}^{(k)} - \hat{Y}_{imp} + o_p(n_2^{-1}N),
\end{aligned}$$

where the second equality holds by (B.7) and (B.8). Therefore, by Theorem 1, the result (20) follows. ■