2018

Survey data integration using mass imputation

Seho Park
Iowa State University

Survey data integration using mass imputation

by

Seho Park

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Jae Kwang Kim, Major Professor
Emily Berg
Wayne A. Fuller
Dan Nettleton
Lily Wang

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

DEDICATION

I would like to dedicate this dissertation to my family whose eternal love, support and encouragement have inspired me throughout this doctoral work.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my major advisor Dr. Jae Kwang Kim for his encouragement and support throughout this research and scholarship. Without his expertise, guidance and persistent help, this dissertation would not have been possible. I also express my warmest gratitude to my committee members, Dr. Wayne Fuller, Dr. Dan Nettleton, Dr. Lily Wang and Dr. Emily Berg, for their encouragement and advice on this dissertation. My thanks and appreciation also go to my friends and colleagues for being supportive and providing valuable assistance and time during this research.

ABSTRACT

Survey sampling has been considered a scientific method of collecting data that represent the target population. Statistical inference using survey data can be improved by incorporating information from existing external data sources. The auxiliary information from other sources can be incorporated into either the design or the estimation stage. In some cases, the original survey data can be augmented with extra data. The data integration can be viewed as a missing data problem and a mass imputation approach can be used for data integration. By filling in the missing values for the study variable in one sample with imputed values incorporating information from the other sample, we can obtain an improved estimator integrating information from two samples.

This dissertation addresses the development of procedures that incorporate auxiliary information or data for three different situations. Three corresponding papers constitute the dissertation and each paper deals with some aspect of incorporation of auxiliary information with survey data that enables us to gain efficiency in inference.

The first paper considers the propensity score weighting method that incorporates auxiliary information from paradata. Paradata are automatically obtainable data about a survey process, which are generated as by-product, and they can be used to handle nonresponse biases. Conditions that are necessary to obtain efficiency gain by incorporating auxiliary information from paradata into the propensity score are considered.

The second paper introduces a new approach to combine two independent probability samples that are selected from the same target population. Augmenting two surveys increases the amount of information about the quantities of our interest and enhances precision in estimation. We introduce the survey data integration method using the measurement error model approach.

The third paper deals with the integration of a two-phase sample where the two samples can be nested or non-nested. We first present the two-phase sampling using the mass imputation method,

which can provide an efficient method to combine two samples where one is nested within the other. A special case of non-nested two-phase sampling where the second-phase sample is a non-probability sample is also investigated.

CHAPTER 1. OVERVIEW

This dissertation addresses methods that enable more accurate and efficient estimation by incorporating information from other sources. Many studies investigating various research questions have used survey data for answering such questions, typically collected using a probability sampling. Advantages of probability sampling are an efficiency in data collection process and theoretical justification resulting from decades of study. Theories for the probability sampling have been developed since 1920s (Neyman, 1934) and support its use in providing accurate and efficient collection methods with assess to the error due to sampling (Lohr and Raghunathan, 2017).

With a long history of probability samples, various methods exist for incorporating information from other sources for estimation. The additional information is called auxiliary information, and it can be augmented into the design or estimation process of sampling for efficiency gain. We can consider a stratification or balanced sampling in order to account for the auxiliary information used in the design stage. Otherwise, the auxiliary information can be used in the estimation process to improve the precision of estimates using poststratification or regression estimation.

However, missingness in survey data is frequently encountered in practice and one of the main reasons is nonresponse. There are two categories of nonresponse: unit nonresponse and item nonresponse. A propensity score weighting method is often used to handle unit nonresponse. The propensity score weighting method compensates for nonresponse and undercoverage by producing appropriate weights on observed units. Auxiliary information can be used to compute the propensity score when the response mechanism is missing at random.

Combining one survey with another, where the two surveys are collected from the same target population, also can be considered as incorporating auxiliary information to obtain improved estimates. Augmenting data by combining two independent surveys can induce a sample with increased size containing more detailed information. Two-phase sampling is one of the classical

setups that can be considered for survey data integration, where the measurement in covariate x is observed in both surveys while the study variable y is observed from only one survey. By combining the auxiliary information in x from one survey with data from the other survey and filling in the missing y's, more efficient estimates can be obtained.

In the meantime, due to the challenges in probability sampling, such as decreasing response rate and increasing expenses, there is an increasing demand to utilize large amount of data that are no longer probability samples. Administrative sources, electronic health records, credit card records or big data sources from web survey panels are examples of the other type of data that are not collected using the probability sampling. These data can be obtained more easily, quickly and cheaply than probability samples and can provide detailed information for subpopulation of interest (Lohr and Raghunathan, 2017). While such data sources can provide valuable auxiliary information, they are often not representative of the target population due to inherent selection biases. Such non-probability samples can be considered more carefully when we try to incorporate them with survey data for increased estimation precision.

In this dissertation, we cover three topics in regard to efficiency gain by incorporating auxiliary information. In Chapter 2, the propensity score weighting method using auxiliary information often believed to increase the precision in estimation is investigated. Chapter 2 is devoted to improve efficiency of estimation using the propensity score weighting that includes surrogate variable when paradata is available. Paradata is often considered as auxiliary information that is collected during the survey process to monitor the quality of the survey response. One such useful type of paradata is the respondent behavior, which can be used to construct response models. The propensity score weight using the respondent behavior information can be applied to the final analysis to reduce the nonresponse bias. However, we discover that including the surrogate variable in the propensity score weighting does not always guarantee increased efficiency. We show that the surrogate variable is useful only when it is correlated with the study variable. Results from a limited simulation study confirm the finding. A real data application using the Korean Workplace Panel Survey data is also presented.

In Chapter 3, we propose a method to combine two data sets using measurement error model approach. Combining information from several surveys from the same target population is an important practical problem in survey sampling. The paper is motivated by the real problem that the authors addressed in a project sponsored by the Food and Nutrition Technical Assistance III (FANTA) Project, with funding from the U.S. Agency for International Development (USAID) Bureau of Food Security (BFS). In the project, two surveys were conducted independently for some areas and we present a measurement error model approach to integrate mean estimates obtained from the two surveys. The predicted values for the counterfactual outcome are used to create composite estimates for the overlapped areas. The proposed method is applied to the real data from the FANTA Project.

Chapter 4 is devoted to estimation of parameters under two-phase sampling in terms of integrating information from two samples. 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 and Sitter (1995) introduce mass imputation for two-phase sampling and its variance estimation under simple random samples in both phases. In this paper, we extend the Rao-Sitter method to the general sampling design. In addition, we also consider a special case of non-nested two-phase sampling where the second-phase sample is a non-probability sample. The proposed method requires the outcome model be correctly specified. Two simulation studies are performed to examine the performance of the proposed methods.

Bibliography

Lohr, S. L. and Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2):293–312.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Rao, J. N. and Sitter, R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82(2):453–460.

CHAPTER 2. A NOTE ON PROPENSITY SCORE WEIGHTING METHOD USING PARADATA IN SURVEY SAMPLING

Submitted to Survey Methodology (revision invited)

Jae Kwang Kim[1], Seho Park[1] and Kimin Kim[2]

Abstract

Paradata is often collected during the survey process to monitor the quality of the survey response. One such paradata is the respondent behavior, which can be used to construct response models. The propensity score weight using the respondent behavior information can be applied to the final analysis to reduce the nonresponse bias. However, including the surrogate variable in the propensity score weighting does not always guarantee the efficiency gain. We show that the surrogate variable is useful only when it is correlated with the study variable. Results from a limited simulation study confirm the finding. A real data application using the Korean Workplace Panel Survey data is also presented.

Key Words: Unit Nonresponse, Smoothed weight, Surrogate variable

2.1 Introduction

Paradata provides additional information on the quality of the collected survey data. The term paradata was coined by Couper (1998) to refer to the process data automatically generated from the data collection. It has been expanded to include various type of data about the data collection process in sample surveys (Kreuter, 2013).

One possibly useful paradata is the respondent behavior during the survey interview. Response time to survey can be one of the respondent behaviors. Knowles and Condon (1999) and Bassili

[1]Department of Statistics, Iowa State University
[2]Korea Labor Institute, South Korea

(2003) found that response time has a negative correlation with the tendency of positive answer. It is called acquiescence bias (Couper and Kreuter, 2013). Longer response times were found to be an indicator of uncertainty and response error (Draisma and Dijkstra, 2004). Such paradata is helpful when we want to build a model for non-responses. Increasing non-response may cause non-response bias and has become a serious problem in recent years. Using the paradata that may be related to response model, non-response adjustment can be used to handle unit nonresponse effectively (Kott, 2006).

In addition to the auxiliary variables, Data Collection Process (DCP) variables are considered for estimation of non-response propensity (Beaumont, 2005). The DCP variable is treated as fixed in Holt and Elliot (1991) and the DCP variable, sometimes refer to the paradata, is used for non-response adjustment. On the other hand, Beaumont (2005) suggests to use DCP variable as a random variable and to be included in the non-response model. They show that using the paradata does not introduce additional bias and variance. Moreover, if the paradata variable is related to the study variable and the non-response, it reduces the non-response bias when the study variable is related to the non-response mechanism directly.

In our study, we show that using the paradata when it is conditionally independent with study variable given auxiliary variables, it inflates the variance as it brings unnecessary noise. While such phenomenon has been recognized in the literature (Little and Vartivarian, 2005), up to the knowledge of authors, it is not fully investigated theoretically. We investigate more rigorously whether using the paradata always improves data analysis by augmenting the nonresponse model.

This paper is motivated from a real survey data from Korean Workplace Panel Survey (KWPS). In the KWPS data, the reaction of the interviewee at the first contact was recorded during the data collection. We investigate possible use of such paradata to enhance the quality of the data analysis.

The paper is organized as follows. In Section 2, basic setup is introduced and the main theoretical results are presented in Section 3. In Section 4, results of simulation studies are presented and a real data application is presented in Section 5. Concluding remarks are made in Section 6.

2.2 Basic Setup

Consider a finite population of size N, where N is known. The finite population $\mathcal{F}_N = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N\}, \mathbf{u}_i = (x_i, y_i)$ is assumed to be a random sample from a superpopulation distribution $F(x, y)$. In addition, we assume that x is always observed and y is subject to missingness. Let δ be the response indicator function that takes the value one if y is observed and takes the value zero otherwise. Note that x, y, z and δ are all considered as random.

Suppose a sample of size n is drawn from the finite population according to a probability sampling design, where inclusion in the sample is represented by the indicator variables I_i, with $I_i = 1$ if unit i is included in the sample and $I_i = 0$ otherwise. Let A be the index set of the sample and $w_i = \pi_i^{-1}$ be the design weight, where π_i is the first-order inclusion probability.

We are interested in estimating parameter θ that is implicitly defined through an estimating equation $E\{U(\theta; X, Y)\} = 0$. Under complete response, an estimator of θ is obtained by solving

$$\frac{1}{N}\sum_{i\in A} w_i U(\theta; x_i, y_i) = 0.$$

In the presence of missing data, the propensity-score adjusted estimator is obtained by solving

$$\frac{1}{N}\sum_{i\in A} w_i \frac{\delta_i}{p_i} U(\theta; x_i, y_i) = 0.$$

where p_i is the response probability of unit i, which is unknown.

Now suppose that there exists additional variable z obtained from paradata, which is always observed and satisfies

$$P(\delta_i = 1 \mid x_i, y_i, z_i) = P(\delta_i = 1 \mid x_i, z_i). \tag{2.1}$$

Then, we can use z to make inference about θ under nonresponse. Such variable z is sometimes called surrogate variable (Chen et al., 2008). By including a suitable surrogate variable, we can make the response mechanism missing at random (MAR) in the sense of Rubin (1976). We call assumption (2.1) as the Augmented MAR (AMAR) since MAR holds only under the augmented model that includes surrogate variable z.

Under (2.1), we can build a parametric model for the response mechanism and construct a propensity score weighted (PSW) estimator that is obtained from

$$\frac{1}{N}\sum_{i\in A} w_i \frac{\delta_i}{\hat{\pi}(x_i, z_i)} U(\theta; x_i, y_i) = 0,$$

where $\hat{\pi}(x_i, z_i)$ is a consistent estimator of $\pi(x_i, z_i) = P(\delta_i = 1 \mid x_i, z_i)$. Such PSW approach incorporating z variable has been discussed in Peress (2010) and Kreuter and Olson (2013).

In survey sampling, the surrogate variable z can be obtained from paradata when we are not directly interested in making inferences about the distribution of z. The information on z, however, can be helpful in making inferences about the joint distribution of x and y. In some cases, the surrogate variable z can satisfy

$$f(y \mid x, z) = f(y \mid x). \tag{2.2}$$

Condition (2.2) means that the surrogate variable z is not related to the study variable y that is subject to missingness. The model satisfying (2.2) can be called the reduced outcome model. If condition (2.2) does not hold, we call $f(y \mid x, z)$ the full outcome model.

If condition (2.2) holds in addition to condition (2.1), we can use this information to obtain a more efficient PSW estimator. Note that, by (2.1) and (2.2), we can establish

$$\begin{aligned}
P(\delta = 1 \mid x, y) &= \int P(\delta = 1 \mid x, y, z) f(z \mid x, y) dz \\
&= \int P(\delta = 1 \mid x, z) f(z \mid x, y) dz \\
&= \frac{\int P(\delta = 1 \mid x, z) f(y \mid x, z) f(z \mid x) dz}{\int f(y \mid x, z) f(z \mid x) dz} \\
&= \frac{\int P(\delta = 1 \mid x, z) f(y \mid x) f(z \mid x) dz}{\int f(y \mid x) f(z \mid x) dz} \\
&= P(\delta = 1 \mid x),
\end{aligned}$$

where the second equality follows from assumption (2.1) and the fourth equality follows from assumption (2.2). Thus, assumption (2.1) and (2.2) imply

$$f(y \mid x, \delta = 1) = f(y \mid x). \tag{2.3}$$

Under the reduced model assumption (2.2), then we can use another type of PSW estimator of the form

$$\frac{1}{N}\sum_{i\in A} w_i \frac{\delta_i}{\hat{\pi}_1(x_i)} U(\theta; x_i, y_i) = 0, \tag{2.4}$$

where $\hat{\pi}_1(x_i) = \int \hat{\pi}(x_i, z_i)\hat{f}(z_i \mid x_i)dz_i$ and $\hat{f}(z \mid x)$ is an estimated conditional density of z given x. Let the estimator from (2.4) be the smoothed PSW estimator. Note that $\hat{\pi}_1(x)$ is the smoothed version of $\hat{\pi}(x, z)$ averaged over the conditional distribution $f(z \mid x)$.

The smoothed PSW estimator obtained by solving the equation (2.4) is justified under MAR condition in (2.3). In this case, use of paradata for nonresponse adjustment is not necessarily useful, which will be justified in Section 3.

2.3 Main Result

We now establish the main result of the paper. We assume that the response indictor functions δ_i are independent to each other. To avoid unnecessary details, we assume that $P(\delta_i = 1 \mid x_i, z_i) = \pi(x_i, z_i)$ is a known function of (x_i, z_i). Let $\hat{\theta}_{PSW}$ be the PSW estimator of θ obtained from

$$U_1(\theta) \equiv \frac{1}{N}\sum_{i\in A} w_i \frac{\delta_i}{\pi(x_i, z_i)} U(\theta; x_i, y_i) = 0.$$

Also, let $\hat{\theta}_{PSW2}$ be the smoothed PSW estimator of θ obtained from

$$U_2(\theta) \equiv \frac{1}{N}\sum_{i\in A} w_i \frac{\delta_i}{\pi_1(x_i)} U(\theta; x_i, y_i) = 0,$$

where $\pi_1(x_i) = P(\delta_i = 1 \mid x_i)$ for $i = 1, \cdots, n$.

Theorem 1 *Under the assumptions (2.1) and (2.2) hold, the smoothed PSW estimator $\hat{\theta}_{PSW2}$ is asymptotically unbiased and has asymptotic variance smaller than that of $\hat{\theta}_{PSW}$. That is,*

$$V(\hat{\theta}_{PSW} \mid \mathcal{F}_N) \geq V(\hat{\theta}_{PSW2} \mid \mathcal{F}_N). \tag{2.5}$$

Proof. First note that

$$E(U_2 \mid \boldsymbol{\delta}_N, \mathcal{F}_N) = \frac{1}{N}\sum_{i=1}^{N} \frac{\delta_i}{\pi_1(x_i)} U(\theta; x_i, y_i),$$

where $\boldsymbol{\delta}_N = \{\delta_1, \cdots, \delta_N\}$. Thus, asymptotic unbiasedness of $\hat{\theta}_{PSW2}$ can be easily established by

$$\begin{aligned} E(U_2 \mid \mathcal{F}_N) &= E\left\{E(U_2 \mid \boldsymbol{\delta}_N, \mathcal{F}_N) \mid \mathcal{F}_N\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left\{\frac{\delta_i}{\pi_1(x_i)} U(\theta; x_i, y_i) \mid x_i, y_i\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{E\left(\delta_i \mid x_i, y_i\right)}{\pi_1(x_i)} U(\theta; x_i, y_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{\pi_1(x_i)}{\pi_1(x_i)} U(\theta; x_i, y_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} U(\theta; x_i, y_i). \end{aligned}$$

For (2.5), it is enough to show that

$$V(U_1 \mid \mathcal{F}_N) \geq V(U_2 \mid \mathcal{F}_N). \tag{2.6}$$

Note that

$$\begin{aligned} V(U_1) &= V\{E(U_1|\boldsymbol{\delta}_N, \mathcal{F}_N) \mid \mathcal{F}_N\} + E\{V(U_1|\boldsymbol{\delta}_N, \mathcal{F}_N) \mid \mathcal{F}_N\} \\ &= V\left\{\frac{1}{N}\sum_{i=1}^{N} \frac{\delta_i}{\pi(x_i, z_i)} U(\theta; x_i, y_i) \mid \mathcal{F}_N\right\} \\ &+ E\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j Cov(I_i, I_j) \frac{\delta_i}{\pi(x_i, z_i)} \frac{\delta_j}{\pi(x_j, z_j)} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \mid \mathcal{F}_N\right\} \\ &:= V_1 + V_2. \end{aligned}$$

Now, since δ_i are independent,

$$V_1 = E\left[\frac{1}{N^2}\sum_{i=1}^{N}\left\{\frac{1}{\pi(x_i, z_i)} - 1\right\} U(\theta; x_i, y_i)^{\otimes 2} \mid \mathcal{F}_N\right],$$

where $B^{\otimes 2} = BB'$. Also, writing $\Delta_{ij} = Cov(I_i, I_j)$,

$$\begin{aligned} V_2 &= E\left\{ \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \Delta_{ij} \frac{\delta_i}{\pi(x_i, z_i)} \frac{\delta_j}{\pi(x_j, z_j)} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \mid \mathcal{F}_N \right\} \\ &= E\left\{ \frac{1}{N^2} \sum_{i=1}^{N} w_i^2 \Delta_{ii} \frac{E(\delta_i \mid x_i, y_i, z_i)}{\pi(x_i, z_i)^2} U(\theta; x_i, y_i)^{\otimes 2} \mid \mathcal{F}_N \right\} \\ &+ E\left\{ \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} w_i w_j \Delta_{ij} \frac{E(\delta_i \mid x_i, y_i, z_i)}{\pi(x_i, z_i)} \frac{E(\delta_j \mid x_j, y_j, z_j)}{\pi(x_j, z_j)} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \mid \mathcal{F}_N \right\} \\ &= E\left\{ \frac{1}{N^2} \sum_{i=1}^{N} (w_i - 1) \frac{1}{\pi(x_i, z_i)} U(\theta; x_i, y_i)^{\otimes 2} + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} w_i w_j \Delta_{ij} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \right\}. \end{aligned}$$

Thus, combining the two results, we obtain

$$\begin{aligned} V(U_1) &= \frac{1}{N^2} E\left[\sum_{i=1}^{N} w_i \left\{ \frac{1}{\pi(x_i, z_i)} - 1 \right\} U(\theta; x_i, y_i)^{\otimes 2} \mid \mathcal{F}_N \right] \\ &+ \frac{1}{N^2} E\left[\sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \Delta_{ij} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \mid \mathcal{F}_N \right]. \end{aligned} \tag{2.7}$$

Similarly, we can establish that

$$\begin{aligned} V(U_2) &= \frac{1}{N^2} E\left[\sum_{i=1}^{N} w_i \left\{ \frac{1}{\pi(x_i)} - 1 \right\} U(\theta; x_i, y_i)^{\otimes 2} \mid \mathcal{F}_N \right] \\ &+ \frac{1}{N^2} E\left[\sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \Delta_{ij} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \mid \mathcal{F}_N \right]. \end{aligned} \tag{2.8}$$

Comparing (2.7) with (2.8), in order to show (2.6), we have only to show that

$$E\left[\frac{1}{\pi(x, z)} \mid x, y \right] \geq \frac{1}{E[\pi(x, z) \mid x, y]}, \tag{2.9}$$

where $E[\pi(x, z) \mid x, y] = \pi_1(x)$. To show (2.9), note that $f(x) = 1/x$ is a convex function of $x \in (0, 1)$ and $\pi(\cdot)$ take values on (0,1). We can apply Jensen's inequality to get

$$E[f(\pi)] \geq f\left(E(\pi)\right),$$

which justifies (2.9) if we use the expectation with respect to the conditional distribution of $\pi(x, z)$ given x and y. ■

By Theorem 1, under some conditions, using the smoothed PSW estimator $\hat{\theta}_{PSW2}$ leads to more efficient data analysis. Beaumont (2008) proposed the smoothed weighting for the efficient estimation with survey data in a slightly different context, but the weight smoothing method of Beaumont (2008) matches with our finding in the sense that z is the design variable and δ is the sample indicator function. In this case, $P(\delta = 1 \mid x, z)$ is the first order inclusion probability while $P(\delta = 1 \mid x)$ is a smoothed version of the first order inclusion probability. Thus, if the sampling design is non-informative in the sense that $P(\delta = 1 \mid x, z, y) = P(\delta = 1 \mid x)$, then it is better to use the smoothed weight that uses $\tilde{w}_i = \{P(\delta = 1 \mid x)\}^{-1}$, which is consistent with the claim of Beaumont (2008) and Kim and Skinner (2013).

Under the reduced model (2.2), adding the surrogate variable z can be regarded as including unnecessary noise and thus it generates inefficient estimators. For the case when the condition (2.2) is unsatisfied, which can be common in practice, we can still use the smoothed PSW estimator using the weight obtained by weight smoothing conditioning on x_i, y_i, and $\delta_i = 1$. So, if the surrogate variable does not satisfy the condition (2.2) and partially or weakly correlated with the study variable given covariate variables, using the smoothing weight conditioning on x_i, y_i, and $\delta_i = 1$ without condition (2.2) can provide the result equivalent to solving (2.4).

2.4 Simulation Study

To test our theory, we perform a limited simulation study. In the simulation, we set a situation when the augmented MAR assumption holds to see if including the surrogate variable in data analysis improves the efficiency of the estimation.

We generate B=2,000 Monte Carlo samples of size $n = 200$ from the outcome model

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{2.10}$$

where $e_i \sim N(0, 1)$, $(\beta_0, \beta_1) = (1.2, 2.6)$, and $X_i \sim N(2, 1)$ for $i = 1, \cdots, n$.

We consider two scenarios: (i) surrogate variable is uncorrelated with study variable; (ii) surrogate variable is correlated with study variable. In scenario 1, condition (2.2) is satisfied, while it is

not satisfied in scenario 2. In scenario 1, we generate a surrogate variable Z from $z_i \sim \text{Unif}(-10, 10)$. In Scenario Two, the paradata model is $z_i = 1.2y_i + \epsilon_i$ where $\epsilon_i \sim N(0, 1)$ for $i = 1, \cdots, n$.

For the response probability, we consider the response model

$$\delta_i \sim \text{Bernoulli}(\pi_i)$$

where

$$\pi_i = \frac{\exp(\phi_0 + \phi_1 x_i + \phi_2 z_i)}{1 + \exp(\phi_0 + \phi_1 x_i + \phi_2 z_i)} \tag{2.11}$$

and $(\phi_0, \phi_1, \phi_2) = (-1.2, 0.5, 0.4)$.

Parameter of interest are regression coefficients in the outcome model (2.10). We compare four methods for estimation of the parameters using Monte Carlo root mean squared errors for the estimates. The four methods considered are as follows:

1. Complete case method (CC): Use the complete observations of (x_i, y_i) and estimate the parameters by the ordinary least squares method. That is, solve

$$\sum_{i=1}^{n} \delta_i U(\theta; x_i, y_i) = 0.$$

2. Propensity score weighting model method (PSW1): Use the estimated response rates as weights in estimating equation and solve the equation to estimate the parameters.

   (a) Fit a logistic regression model (2.11) for the response probability $\pi_i = \pi_i(x_i, z_i; \phi)$ and estimate $\phi = (\phi_0, \phi_1, \phi_2)$ by using the maximum likelihood method.

   (b) Parameter estimates are obtained by solving the estimating equation:

$$\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} U(\theta; x_i, y_i) = 0,$$

   where $\hat{\pi}_i = \hat{\pi}(x_i, z_i; \hat{\phi})$ and $\hat{\phi}$ is computed from Step (a).

3. Smoothed propensity score weighting model method (PSW2): Use the same procedure of PSW1, but the response probability is a function of explanatory variable (x) only. A response probability $\pi(x_i)$ is estimated as

$$\hat{\pi}_1(x_i) = \int \hat{\pi}(x_i, z_i) \hat{f}(z_i \mid x_i) dz_i,$$

where $\hat{\pi}(x_i, z_i)$ is the estimated response probability in PSW1 method. Since the estimated conditional density of z given x, $\hat{f}(z|x)$, is unknown, we use a nonparametric regression method for estimating $\hat{f}(z|x)$. Let $K_h(\cdot)$ be the kernel function satisfying certain regularity conditions and h be the bandwidth. Then, $\hat{\pi}_1(x_i)$ is obtained by

$$\hat{\pi}_1(x_i) = \frac{\sum_{j=1}^{n} \hat{\pi}(x_j, z_j) K_h(x_i, x_j)}{\sum_{j=1}^{n} K_h(x_i, x_j)}.$$

We used the Gaussian Kernel for K_h with bandwidth $h = 1.06\hat{\sigma}n^{-1/5}$ chosen by the rule-of-thumb method of Silverman (1986).

4. Smoothed propensity score weighting estimator (PSW3) using logistic regression for estimating $\hat{\pi}_1(x_i)$: Use the same procedure of PSW1, but the response probability is estimated by a logistic regression model using only x_i.

   (a) Fit a logistic regression model for the response probability $\pi_i^* = \pi_i(x_i; \phi^*)$ as a function of explanatory variable (x_i) only and estimate $\phi^* = (\phi_0^*, \phi_1^*)$ by using the maximum likelihood method.

   (b) Parameter estimates are obtained by solving the estimating equation:

$$\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i^*} U(\theta; x_i, y_i) = 0,$$

   where $\hat{\pi}_i^* = \hat{\pi}(x_i; \hat{\phi}^*)$ and $\hat{\phi}^*$ is computed from Step (a).

Table 2.1 presents the Monte Carlo biases and Monte Carlo root mean squared errors of the estimator of the three parameters under the scenario 1, where the surrogate variable is uncorrelated with the study variable. Monte Carlo bias can be obtained by the difference between Monte Carlo mean and the true mean divided by the true mean. Monte Carlo root squared mean squared error is the squared value of Monte Carlo mean squared error, which is sum of squared Monte Carlo bias and Monte Carlo variance. As discussed in Section 3, the Monte Carlo root mean squared errors obtained using the smoothed propensity score weighting method (PSW2 and PSW3) are smaller than the standard error of the propensity score weighting method (PSW1) as condition (2.2) is

satisfied. The result confirms our theory that including the surrogate variable that is uncorrelated with the study variable may cause unnecessary noise for estimating parameters and decrease the efficiency.

Table 2.1 Monte Carlo biases (Bias) and Monte Carlo root mean squared errors (RMSE) of point estimators for scenario 1.

Parameter	Method	Bias	RMSE
β_0	CC	0.000	0.248
	PSW1	0.001	0.451
	PSW2	0.001	0.263
	PSW3	0.000	0.252
β_1	CC	-0.001	0.105
	PSW1	-0.001	0.178
	PSW2	-0.001	0.115
	PSW3	-0.001	0.107

The simulation results for the scenario 2 are presented in Table 2.2. The results show that the estimators obtained by the propensity score weighting method (PSW1) are unbiased, but the estimators obtained by CC method, PSW2 method and PSW3 method are biased. The PSW2 and PSW3 method is biased because the surrogate condition (2.2) is not satisfied. It implies that the surrogate variable contains useful information for estimation that cannot be overlooked. The PSW2 and PSW3 methods provide estimates with smaller standard errors than PSW1 method in this simulation setup as they take account the surrogate variable for estimation of propensity score weights.

2.5 Application

2.5.1 Data Description

The research is motivated by real data analysis in Korean workplace panel survey data, which is a biennial panel survey of the workplaces in Korea, sponsored by Korean Labor Institute. We used the KWPS data collected in 2007, 2009, and 2011 for our analysis.

Table 2.2 Monte Carlo biases (Bias) Monte Carlo standard errors (Std Error) and Monte Carlo root mean squared errors (RMSE) of point estimators for scenario 2.

Parameter	Method	Bias	Std Error	RMSE
β_0	CC	-0.145	0.206	0.252
	PSW1	-0.004	0.236	0.236
	PSW2	-0.738	0.356	0.819
	PSW3	-0.167	0.240	0.292
β_1	CC	0.045	0.088	0.098
	PSW1	0.004	0.101	0.101
	PSW2	0.200	0.131	0.239
	PSW3	0.054	0.104	0.117

Target population of the survey is all company in South Korea with size of greater than 30 people in the company except agriculture, forestry, fishing and hunting. Of all companies in the target population, which is of size 37,644 companies, 1400 companies were selected using complex sampling design. Sample size is allocated for the companies that have employees less than 500 people and the companies with size of greater than 500 people were all selected and interviewed. This is because there are less number of companies whose size is greater than 500.

The sampling design used for the survey is a stratified sampling using the company as both a sampling unit and an experimental unit. The stratification variable is formed using 3 variables: the size of the company, the type of the company and the area where it was located. A combination of the three variables resulted in 200 strata since there are 5 levels of area, 4 levels of size of company, and 10 levels of type of company results.

In order to determine the sample from the target population, two-stage sampling is used. The procedure is as follows: In the first stage, the target population is divided into 10 strata using the type of the company variable. Within the stratum, the size of company is used to form substrata and the Kish method was applied to allocate samples within the stratum. In the second stage, within a stratum formed by two variables, the area variable was used to make substrata and the proportional allocation method was used to assign a sample size according to the size of the area.

From the KWPS data, we are interested in fitting a regression model for the regression of the log-scaled sales per person (Y=log(Sales)/Person) on two covariates of the company: size of

company (X_1) and type of company (X_2). In the dataset, variable Y is not completely observed for all targets of the survey; they contain some missing values. However, the explanatory variables are completely observed as the size and type of company are the characteristics that do not change easily in every two years.

The response variable (Y), the log-scaled sales per person, is a continuous variable. The two explanatory variables are categorical. The size of company variable (X_1) has four categories; 30-99 people, 100-299 people, 300-499 people, and more than 500 people. The type of company variable (X_2) contains twelve categories: Light industry, chemical industry, electric/electronic industry, etc.

In the KWPS data, the variable regarding the reaction of interviewees at the first contact has been collected during the survey process and is considered a surrogate variable in our analysis. The reaction at the first contact is categorical with three categories:

1. Friendly response ($Z = 1$): the interviewee accepts the survey or answers the pre-questionnaire or fixes the visit date.

2. Moderate response ($Z = 2$): the interviewee cannot complete the survey immediately, but allows the follow-up survey.

3. Negative response ($Z = 3$): the interviewee who completes the survey uncooperatively or responded negatively.

Table 2.3 shows the response rates for each category of the first contact reaction. In friendly and moderate responses, response rates are 0.71 and 0.67, respectively, but the response rate for negative response is 0.45. This suggests that the surrogate variable is an important predictor for the response model.

Table 2.3 Response rate corresponding each level of reaction of interviewees

	Friendly Response	Moderate Response	Negative Response
Response Rate	0.71	0.67	0.45

From the dataset, we are interested in estimating the parameters in the regression model

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

2.5.2 Analysis

We first check whether condition (2.2) is satisfied. Using the idea of Fuller (1984), we test the hypothesis $H_0 : \gamma = 0$ in the following model

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + e, \tag{2.12}$$

where $\mathbf{X} = (1, x_1, x_2)$ is a vector of explanatory variables, $\mathbf{Z}$ is a vector of surrogate variables, and e is a random error following $N(0, \sigma^2)$. Under H_0, we can roughly say that surrogate condition (2.2) is satisfied. Table 2.4 presents the result of the hypothesis testing. The F-statistic of the test is 0.3508 and its p-value is 0.7041, suggesting strong evidence in favor of the null hypothesis that the surrogate variables are not significant in the augmented regression model (2.11). Thus, we can safely assume that the vector of surrogate variables $\mathbf{Z}$ can be treated as conditionally independent with the response variable Y given the explanatory variable X and condition (2.2) is satisfied.

Table 2.4 Test of the significance of the surrogate variable in the model (2.11)

	F statistic	p-value
$H_0 : \gamma = 0$	0.3508	0.7041

Figure 2.1 confirms the surrogacy condition (2.2). The median of three boxes seems to be almost the same around 0 and supports the result of the test that the surrogate variable is uncorrelated with response variable given explanatory variables. Hence, all of these results imply that the assumption (2.2) holds for the data.

We now compare the three methods for estimating the parameters of the outcome model in (2.11), which are CC method, PSW1 method and PSW2 method. Estimated coefficients and their standard errors are presented in Table 2.5. The standard errors are calculated using bootstrap method. Since the company is an experimental unit as well as a sampling unit, a bootstrap sample of 1,400 companies is randomly sampled with replacement from the original sample and coefficients are estimated from the bootstrap sample. The estimates are bootstrap replicate of the coefficients. We repeat the procedure B=1,000 times and obtain 1,000 bootstrap replicates. Then, a bootstrap standard error is obtained from the bootstrap replicates.

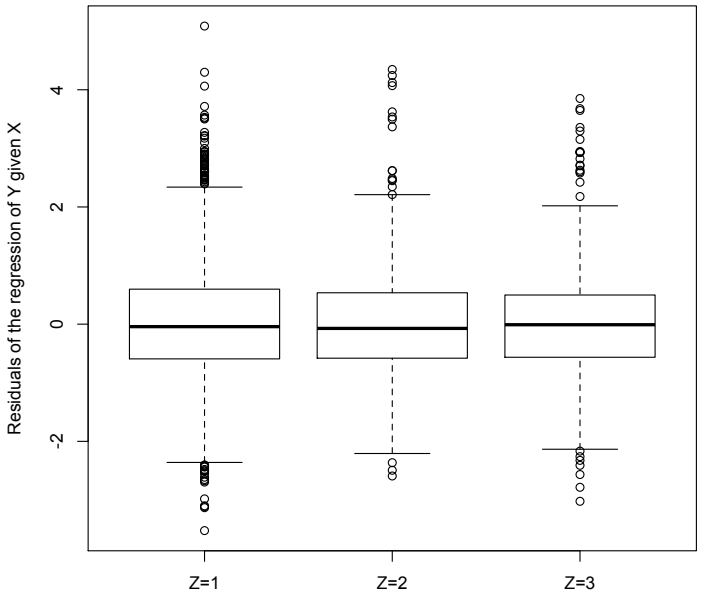

Figure 2.1 Boxplots of residuals of the regression of Y given $\mathbf{X}$ across each category of Z

Since two explanatory variables are categorical with 4 and 12 levels, respectively, there are 15 coefficient parameters to be estimated. Table 2.5 presents the parameter estimates and their standard errors. We can see that the estimates obtained by using three methods are similar, but the standard errors obtained by using PSW2 method are smaller across all levels of variables than the standard errors calculated by using PSW1 method. The gain of efficiency by using PSW2 method rather than PSW1 method is not large, we can check there exists an decrease in standard errors for PSW2 method. The reason why PSW1 method produces larger standard errors compares to PSW2 method is that it incorporates the additional information from the paradata through the surrogate variable z. As indicated before, including the surrogate variable in calculating the propensity score weight generated unnecessary noise in estimation as the surrogate variable is uncorrelated with the study variable. Thus, we conclude that it is desirable to avoid including the paradata's information into the data analysis.

Table 2.5 Estimated coefficient (standard error) from the real data analysis. (CC, complete case; PSW1, propensity score weighting method 1; PSW2, smoothed propensity score weighting method 2)

	CC	PSW1	PSW2
Intercept	5.395 (0.212)	5.396 (0.217)	5.414 (0.211)
100-299 people	0.155 (0.200)	0.154 (0.205)	0.138 (0.199)
300-499 people	0.432 (0.206)	0.379 (0.212)	0.378 (0.205)
$>$ 500 people	0.618 (0.216)	0.565 (0.218)	0.556 (0.215)
Chemical	0.379 (0.242)	0.374 (0.246)	0.371 (0.242)
Metal/Auto	0.259 (0.221)	0.260 (0.223)	0.257 (0.221)
Elec/Electronic	-0.026 (0.236)	-0.006 (0.239)	-0.019 (0.235)
Construction	0.207 (0.282)	0.194 (0.286)	0.189 (0.282)
Personal Services	0.339 (0.242)	0.383 (0.245)	0.356 (0.241)
Transportation	-1.219 (0.269)	-1.195 (0.279)	-1.207 (0.268)
Communication	0.090 (0.351)	0.145 (0.356)	0.104 (0.350)
Financial Insur	1.145 (0.299)	1.194 (0.334)	1.152 (0.298)
Business Services	-1.155 (0.069)	-1.094 (0.070)	-1.113 (0.069)
Social Services	-0.869 (0.256)	-0.841 (0.259)	-0.840 (0.256)
Elec/Gas	2.114 (0.261)	2.106 (0.263)	2.099 (0.261)

Chemical, Chemical Industry; Metal/Auto, Metal and Automobile Industry; Elec/Electronic, Electrical and Electronical Industry; Financial Insur, Finance and Insurance Services; Elec/Gas, Electric and Gas Services

Although the fact in Theorem 1 can be found in the example, that is PSW1 method produces larger standard errors because of addition of unnecessary noise, there is no real gain using PSW2 method compared with CC method. That is, disregarding missing values and using only observed values produces the most efficient estimates in the example. The possible reasons of no real gain using PSW2 method compared with CC method are 1) the missing pattern in Y is completely random; or 2) the lack of covariate variables used in the construction of the propensity score weights. That is, it seems the pattern in missingness in the sales per person (Y) is completely independent with the size or type of company in the example. In this case, constructing and using the propensity score weight for adjustment of the missingness brings no gain in estimation.

2.6 Conclusion

Motivated by the real survey project, we have investigated the propensity score approach incorporating the information from paradata into the response propensity model. Use of paradata in the propensity model has been advocated in the literature. However, it is not always the case. We find that using more information can decrease the efficiency of analysis, which is justified in Theorem 1. The claim is confirmed in the simulation study and the real data analysis using KWPS data. When the surrogate variable in the paradata is conditionally independent with the study variable, conditional on the explanatory variable, it is better not to include the surrogate variable because the smoothed propensity score weight can provide more efficient estimation. In other words, it is useful to include the information from paradata only when the surrogate is correlated with the variable of interest.

Bibliography

Bassili, J. N. (2003). The minority slowness effect: Subtle inhibitions in the expression of views not shared by others. *Journal of Personality and Social Psychology*, 84(2):261.

Beaumont, J. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, 31(2):227.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95(3):539–553.

Chen, S. X., Leung, D. H., and Qin, J. (2008). Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):803–823.

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 48:743–772.

Couper, M. P. and Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1):271–286.

Draisma, S. and Dijkstra, W. (2004). Response latency and (para) linguistic expressions as indicators of response error. *Methods for testing and evaluating survey questionnaires*, pages 131–147.

Fuller, W. A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10:97–118.

Holt, D. and Elliot, D. (1991). Methods of weighting for unit non-response. *The Statistician*, pages 333–342.

Kim, J. K. and Skinner, C. J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, 100(2):385–398.

Knowles, E. S. and Condon, C. A. (1999). Why people say" yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, 77(2):379.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133–142.

Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information*, volume 581. John Wiley & Sons.

Kreuter, F. and Olson, K. (2013). Paradata for nonresponse error investigation. In *Improving surveys with paradata: Analytic uses of process information*, volume 581, pages 13–42. John Wiley & Sons.

Little, R. J. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2):161.

Peress, M. (2010). Correcting for survey nonresponse using variable response propensity. *Journal of the American Statistical Association*, 105(492):1418–1430.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

CHAPTER 3. A MEASUREMENT ERROR MODEL APPROACH TO SURVEY DATA INTEGRATION: COMBINING INFORMATION FROM TWO SURVEYS

Published in Metron, Volume 75(3), 345-357

Seho Park[1], Jae Kwang Kim[1] and Diana Stukel[2]

Abstract

Combining information from several surveys from the same target population is an important practical problem in survey sampling. The paper is motivated by work that authors undertook, sponsored by the Food and Nutrition Technical Assistance III Project (FANTA), with funding from the U.S. Agency for International Development (USAID) Bureau of Food Security (BFS). In the project, two surveys were conducted independently for some areas and we present a measurement error model approach to integrate mean estimates obtained from the two surveys. The predicted values for the counterfactual outcome are used to create composite estimates for the overlapped areas. An application of the technique to the project is provided.

Key Words: counterfactual outcome; composite estimate; variance estimation.

3.1 Introduction

Survey integration is an emerging research area of statistics, which concerns combining information from two or more independent surveys to get improved estimates for various parameters of interest for the target population. One of the early applications of survey integration is the Consumer Expenditure Survey (Zieschang, 1990), where two survey vehicles (a Diary survey and a quarterly interview survey) were used to obtain improved estimates for the Diary survey items.

[1]Department of Statistics, Iowa State University
[2]FANTA III Project, FHI 360, Washington, DC, USA

Renssen and Nieuwenbroek (1997), Merkouris (2004, 2010), Wu (2004) and Ybarra and Lohr (2008) considered the problem of combining data from two independent surveys to estimate totals at the population and domain levels.

Combining information from two or more independent surveys is a problem frequently encountered in survey sampling. One of the classical setups used to combine information is two-phase sampling, where the measurement x is observed in both surveys and the study variable y is observed only from one survey, say, in Survey A. There is no measurement for y in survey B. In this case, we can treat the union of Survey A and Survey B samples as a phase one sample and treat the Survey A sample as a phase two sample. Hidiroglou (2001) formulated this problem and developed efficient estimation using a two-phase regression estimation method. Fuller (2003), Legg and Fuller (2009), and Kim and Rao (2012) considered this problem as a missing data problem and developed mass imputation to obtain improved estimation for the total as well as domain totals. Our setup is different from the two-phase sampling approach in the sense that we have a different measurement of y from two surveys.

We consider a situation where two surveys have common measurement for x but different measurements for y. For example, x can be demographic information that does not suffer from measurement errors but y can suffer from survey-specific measurement errors. The survey-specific difference can occur due to differences in survey questions or survey modes (e.g. Dillman et al. (2009)). In Table 3.1, for example, the Survey A sample contains observations in x and y_1 while the Survey B sample contains observations in x and y_2. In the case of y_1 being the study variable of interest, if we can assume that y_2 is a measurement for y_1 with measurement errors, then at issue is the estimation of the population mean of y_1 combining two surveys.

Table 3.1 Data structure for combining two surveys with measurement errors

	x	y_1	y_2
Survey A	o	o	
Survey B	o		o

Our research is motivated by work sponsored by The Food and Nutrition Technical Assistance III Project (FANTA) with funding from U.S. Agency for International Development (USAID), to produce integrated estimates from two independent surveys conducted in Guatemala where the geographic areas covered by the two surveys have substantial overlap.

Section 2 provides background on the projects and data descriptions and Section 3 introduces the proposed method for survey integration. In Section 4, we illustrate the estimation process and results of the work sponsored by FANTA, and Section 5 provides concluding remarks.

3.2 The Food and Nutrition Technical Assistance III Project

3.2.1 Background

FANTA is a 5-year cooperative agreement between the USAID and FHI 360. FANTA aims to improve the health and well-being of vulnerable groups through technical support in the areas of maternal and child health and nutrition in development and emergency contexts, HIV and other infectious diseases, food security and livelihood strengthening, agriculture and nutrition linkages and emergency assistance in nutrition crises.

USAID is the lead U.S. government agency that works to end extreme global poverty and enable resilient, democratic societies to realize their potential. The Feed the Future Initiative (FTF) was launched in 2010 by the United States government to address global hunger and food insecurity. The Initiative is coordinated primarily by the USAID and is housed within the Bureau of Food Security (BFS), but includes the Office of Food for Peace (FFP). The main objectives of the FTF initiative are the advancement of global agricultural development, increased food production and food security, and improved nutrition particularly for vulnerable populations such as women and children. The FTF initiative is active in 19 focus developing countries in Africa, Asia and Latin America. One of these focus countries is Guatemala.

Both BFS (through the FTF initiative) and FFP sponsor periodic baseline, interim and end-line household surveys to gauge the extent of progress towards achieving the goals of the FTF initiative. In 2013, FFP engaged a third party contractor, ICF International, to conduct a baseline household

survey in 5 departments of the Western Highlands of Guatemala. In the same year, BFS/FTF (henceforth referred to as FTF) engaged a third party contractor, UNC MEASURE, to conduct an interim household survey in the same 5 departments in Guatemala. Although the surveys were conducted in the same 5 departments, the geography of the two surveys did not exactly coincide; however, there was substantial geographic overlap. The union of the geography covered by the two surveys represents the FTF Zone of Influence (ZOI), where some of the most food insecure parts of the population in the country reside. Because, FTF was interested in obtaining ZOI-level estimates for a number of key indicators using data from the two independent surveys, they provided funding to FANTA, who in turn, engaged the authors to undertake the work. Because of the overlapped geography from the two surveys, it was necessary to use data integration methods to produce overall ZOI-level estimates.

Guatemala has 22 departments, which are geographic entities, divided into 334 municipalities. The two surveys were each conducted in the following five departments of the Western Highlands of Guatemala: San Marcos, Totonicapan, Quiche, Quezaltenango, and Huehuetenango. Thus, two surveys were conducted in the areas and the survey data from the two samples are ready to be combined for survey integration. More details of this project can be found from the reference provided by USAID (USAID, 2013).

3.2.2 Common Indicators

ICF International (FFP) and UNC MEASURE (FTF) used their own questionnaire for the surveys, and among the indicators in the questionnaires, there were 11 common indicators in both surveys indicating maternal and child health status. Among the 11 common indicators, 4 were collected at the household-level and the remaining 7 were collected at the individual-level. Five indicators of the 7 individual-level indicators pertained to children and remaining 2 to women. Table 3.2 presents the common indicators and their descriptions.

Most indicator variables are dichotomous, taking the values of either 0 or 1 in both data sets, but the other two indicator variables, which are 'PCE' and 'WDDS,' are numeric in both data sets.

Table 3.2 Eleven Common Indicators

Level	Indicator
Household	Daily Per Capita Expenditures (PCE)
	Prevalence of Households with Hunger (HHS)
	Prevalence of Poverty (PP)
	Mean Depth Poverty (MDP)
Individual(Children)	Prevalence of Stunted Children
	Prevalence of Wasted Children
	Prevalence of Underweight Children
	Prevalence of Children Receiving a Minimum Acceptable Diet (MAD)
	Prevalence of Exclusive Breastfeeding (EBF)
Individual(Women)	Prevalence of Underweight Women
	Women's Dietary Diversity Score (WDDS)

In this paper, we focus on the 'PCE' and the 'HHS' indicators for analysis as examples of a numeric variable and a dichotomous variable, respectively.

3.2.3 Survey Design

3.2.3.1 FFP Survey

The survey for the FFP project used a three-stage sampling design. In the first stage, the primary sampling unit is the village, where the village population for five departments is divided into two substrata in each department. Each department has two substrata except for Quetzaltenango which has one stratum. So, we have 9 strata and the first stage sample selection probability is based on the number of villages in the sampling frame and the size of the village within each stratum. The sampling frame for the first stage sampling included all the villages identified for program implementation. Table 3.3 shows the summary of sample clusters in each stratum.

In the second stage sampling, sample households were selected randomly from each sampled village. The target number of households selected for each village was 40. The second stage sample selection probability is based on the number of households selected for each village divided by the total number of households in each village.

The third stage sampling was done at the individual level to select woman and children in households. The third stage sample selection probability is based on the total number of individuals selected for each interview module and the number of eligible individuals in the household. Only one eligible woman was randomly selected using the Kish grid (Kish, 1949), but all children were selected to be interviewed.

The final sampling weights are computed as the inverse of products of the three stage first-order inclusion probabilities.

Table 3.3 Survey Design of the FFP Project

Department	Strata	Total No. of Clusters	No. of Selected Clusters
1. San Marcos	11	89	17
	12	30	17
2. Totonicapan	21	85	22
	22	22	19
3. Quiche	31	62	22
	32	19	13
4. Huehuetenango	41	48	12
	42	18	12
5. Quetzaltenango	51	24	16

3.2.3.2 FTF Survey

The survey for the FTF project also used a three-stage sampling design using census sectors as the primary sampling units. In the first stage, the census areas (urban/rural) were formed in each department and census sectors were sampled within the census area. From the sampled census sectors, the sample households were randomly selected in the second stage sampling. For the third stage sampling, data on individual-level women and children were collected. All women and children in a household are included in the sample, but the weights associated with women and children are adjusted for nonresponse. Table 3.4 shows the summary of sample clusters in each stratum.

Table 3.4 Survey Design of the FTF Project

Department	Strata	Total No. of Clusters	No. of Selected Clusters
1. San Marcos	Rural	192	25
	Urban	99	3
2. Totonicapan	Rural	237	5
	Urban	128	1
3. Quiche	Rural	284	33
	Urban	97	7
4. Huehuetenango	Rural	336	39
	Urban	80	8
5. Quetzaltenango	Rural	117	1
	Urban	190	1

3.3 Survey Data Integration

We present the proposed method in the context of measurement error models. In a classical measurement error model problem, the interest lies in estimating the regression coefficient for the regression of y on x and the covariate x is subject to measurement errors (Fuller, 2009). In our problem, the measurement error occurs in y for one survey (Survey B) and we are interested in combining two surveys to estimate the population mean of y more efficiently. Thus, we still consider the data structure in Table 3.1. We treat y_1 as the gold standard, $y_1 = y$, in the sense that there is no measurement error in y_1.

Let $f_1(y_1 \mid x; \theta_1)$ be the density for the conditional distribution of y_1 on x, characterized by parameter θ_1. Model for $f_1(y_1 \mid x; \theta_1)$ can be called a structural equation model (Fornell and Larcker, 1981). Let $f_2(y_2 \mid x, y_1; \theta_2)$ be the density for the conditional distribution of y_2 on (x, y_1), characterized by parameter θ_2. For parameter identifiability, we assume that

$$f_2(y_2 \mid x, y_1) = f_2(y_2 \mid y_1). \tag{3.1}$$

Such assumption is sometimes called the nondifferential measurement error assumption (Buonaccorsi, 2010, p.7) in the measurement error model literature. That is, x is an instrumental variable for y_1. The nondifferential measurement error assumption is used to obtain a reduced model.

Given the sample with the data structure in Table 3.1, the imputed values for y_1 in sample B are used to obtain the composite estimator that combines direct observations in the sample A and synthetic values in the sample B. The imputed values are the best predicted values of the counterfactual outcome variable y_1 in sample B, which correct for measurement errors in observed valued of y_2. The imputed values are generated using the prediction model for y_1, $f(y_1 \mid x, y_2)$.

For the parameter estimation, the (pseudo) maximum likelihood estimator of θ_1 and θ_2 can be obtained by using the full EM algorithm as follows:

[E-step] Compute

$$\begin{aligned} Q_1(\theta_1|\theta_1^{(t)},\theta_2^{(t)}) &= \sum_{i\in S_a} w_{ia}\log f_1(y_{1i}|x_i;\theta_1) \\ &+ \sum_{i\in S_b} w_{ib}\mathrm{E}\left[\log f_1(y_{1i}|x_i;\theta_1) \mid x_i, y_{2i}; \theta_1^{(t)},\theta_2^{(t)}\right] \end{aligned}$$

and

$$\begin{aligned} Q_2(\theta_2|\hat{\theta}_1^{(t)},\theta_2^{(t)}) &= \sum_{i\in S_a} w_{ia}\mathrm{E}\left[\log f_2(y_{2i}|y_{1i};\theta_2) \mid x_i, y_{1i}; \hat{\theta}_1^{(t)},\theta_2^{(t)}\right] \\ &+ \sum_{i\in S_b} w_{ib}\mathrm{E}\left[\log f_2(y_{2i}|y_{1i};\theta_2) \mid x_i, y_{2i}; \hat{\theta}_1^{(t)},\theta_2^{(t)}\right], \end{aligned}$$

where S_a and S_b are the index sets for the Survey A sample and the Survey B sample, respectively. Also, w_{ia} and w_{ib} are the sampling weight for unit $i \in S_a$ and for unit $i \in S_b$, respectively. The conditional expectation in Q_1 is taken with respect to

$$f(y_1|x,y_2;\theta_1,\theta_2) = \frac{f_1(y_1|x;\theta_1)f_2(y_2|y_1;\theta_2)}{\int f_1(y_1|x;\theta_1)f_2(y_2|y_1;\theta_2)dy_1}$$

evaluated at $\theta_1 = \theta_1^{(t)}$ and $\theta_2 = \theta_2^{(t)}$ for Q_1 and at $\theta_1 = \hat{\theta}_1^{(t)}$ and $\theta_2 = \theta_2^{(t)}$. For Q_2, the first conditional expectation is taken with respect to $f(y_{2i}|x_i, y_{1i}) = f(y_{2i}|y_{1i})$ by assumption (3.1), evaluated at $\theta_2 = \theta_2^{(t)}$.

[M-step] Update θ_1 by maximizing $Q_1(\theta_1|\theta_1^{(t)},\theta_2^{(t)})$ with respect to θ_1 and update θ_2 by maximizing $Q_2(\theta_2|\hat{\theta}_1^{(t)},\theta_2^{(t)})$ with respect to θ_2.

Based on the estimated parameters $\hat{\theta}_1$ and $\hat{\theta}_2$, the best predictor of y_1 of the Survey B sample is obtained as the expectation of the predictive distribution, which is the conditional distribution of y_1 given x and y_2. That is, the best predictor of y_{1i} is

$$\hat{y}_{1i}^{*} = \mathrm{E}\left(y_{1i}|x_i, y_{2i}; \hat{\theta}_1, \hat{\theta}_2\right). \tag{3.2}$$

The parametric fractional imputation of Kim (2011) can be used to generate fractionally imputed values for y_1 in sample B under the general parametric models (Park et al., 2016). When $f_1(y_1|x;\theta_1)$ and $f_2(y_2|x,y_1;\theta_2)$ have general parametric models, the prediction model may not have a closed form. In this case, the parametric fractional imputation can be used following two-step method:

1. For each $i \in S_b$, generate $y_{1i}^{*(j)}$ from $f_1(y_{1i} \mid x_i; \hat{\theta}_1)$ for $j = 1, \cdots, m$.

2. Let $y_{1i}^{*(j)}$ be the j-th imputed value of y_{1i} obtained from Step 1. The fractional weight assigned to $y_{1i}^{*(j)}$ is
$$w_i^{*(j)} = \frac{f_2(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_2)}{\sum_{k=1}^{m} f_2(y_{2i} \mid x_i, y_{1i}^{*(k)}; \hat{\theta}_2)}.$$

Once we use the parametric fractional imputation, the conditional expectation in (3.2) can by computed by a Monte Carlo approximation. That is, the conditional expectation can be written by

$$\hat{y}_{1i}^{*} \cong \sum_{j=1}^{m} w_i^{*(j)} y_{1i}^{*(j)}.$$

Using the counterfactual values (3.2) of the Sample B and observations of the Survey A sample, we can construct a composite estimator that combines two values. The combined estimator is

$$\bar{y}_{com}^{*} = \frac{\sum_{i\in S_a} w_{ia} y_{1i} + \sum_{i\in S_b} w_{ib} \hat{y}_{1i}^{*}}{\sum_{i\in S_a} w_{ia} + \sum_{i\in S_b} w_{ib}}.$$

Kim et al. (2016) have investigated the parametric fractional imputation of Kim (2011) in the context of statistical matching where the main interest lies in estimating θ_2 in $f_2(y_2 \mid x, y_1; \theta_2)$. In their simulation study, the imputation model is based on the nondifferential measurement error assumption, but they noticed that departure from the assumption does not affect the validity of

the imputation estimator for the population mean of y_1, even though it leads to biased estimation of the regression parameters. Note that if the assumption does not hold, then the imputation model (based on the assumption) is incorrectly specified. Under the incorrectly specified model, the imputed estimator is still unbiased for the mean estimation, as long as an intercept term is included in the model (Kim and Rao, 2012).

3.4 Application of Methodology to USAID surveys in Guatemala

Based on the two estimates obtained from the two independent surveys on the overlap areas, we can improve the efficiency of the estimation by combining the two estimates.

3.4.1 Survey Data Integration

In this section, we use a measurement error model approach to integrate two surveys, the FFP and the FTF, presented in Section 3. In the view of the measurement error model approach, we treat one sample as a gold standard and the other sample containing measurement errors.

Throughout this study, the FFP sample was used as a benchmark and we predicted the counterfactual outcomes of the FTF sample, which is the value that would have obtained when the FTF sample was collected by ICF International who conducted the FFP project. This is based on the idea that measurement errors between two surveys are diminished when we consider the predicted values of the counterfactual values instead of the original values from the survey. We chose the FFP sample as a reference point since it has a smaller residual sum of squares compares to the one from the FTF sample.

3.4.1.1 Case 1: Continuous Study Variable

Since the PCE indicator has continuous values, we treat a structural equation model and a measurement error model both follow normal distributions. Assume that a structural equation model for y_1 is

$$y_{1i} = \boldsymbol{\beta}_1 \mathbf{x}_{1i} + \beta_2 x_{2i} + e_i, \tag{3.3}$$

where $\mathbf{x}_{1i}$ is a department indicator and x_{2i} is a variable indicating the total number of household members, and $e_i \sim N(0, \sigma_e^2)$. Also, a measurement error model for y_2 is

$$y_{2i}|y_{1i} = \alpha_0 + \alpha_1 y_{1i} + u_i,$$

where $u_i \sim N(0, \sigma_u^2)$. By using the Bayes theorem, the predictive distribution can be derived as

$$y_{1i}|y_{2i}, \mathbf{x}_i \sim N(\mu_i, v^2) \tag{3.4}$$

where $\mathbf{x}_i = (\mathbf{x}_{1i}, x_{2i})$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \beta_2)$ and

$$\mu_i = c_i \boldsymbol{\beta} \mathbf{x}_i + (1 - c_i)\alpha_1^{-1}(y_{2i} - \alpha_0)$$

with

$$c_i = \frac{1/\sigma_e^2}{1/\sigma_e^2 + \alpha_1^2/\sigma_u^2}$$

and

$$v^2 = \frac{\sigma_e^2 \sigma_u^2/\alpha_1^2}{\sigma_e^2 + \sigma_u^2/\alpha_1^2}.$$

For the analysis of the PCE indicator, we assumed a linear regression model (3.3). The model diagnostics for the model assumptions are given in Figure 3.1. Two plots show that the normality assumption and the homogeneity of variance assumption are appropriate. Residual plot also shows no particular pattern in residuals so the model assumptions in (3.3) are regarded as reasonable.

For the parameter estimation, we write $\theta_1 = (\boldsymbol{\beta}_1, \beta_2, \sigma_e^2)$ and $\theta_2 = (\alpha_0, \alpha_1, \sigma_u^2)$. The best estimator of θ_1 and θ_2 can be obtained by the full EM algorithm as explained in Section 3. In this example, the Q_1 and Q_2 are as follows:

[E-step] Compute

$$\begin{aligned} Q_1(\theta_1|\theta_1^{(t)}, \theta_2^{(t)}) &= \sum_{i \in S_a} w_{ia} \left\{ -\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}(y_{1i} - \boldsymbol{\beta}\mathbf{x}_i)^2 \right\} \\ &+ \sum_{i \in S_b} w_{ib} \mathrm{E}\left[-\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}(y_{1i} - \boldsymbol{\beta}\mathbf{x}_i)^2 \,\middle|\, \mathbf{x}_i, y_{2i}; \theta_1^{(t)}, \theta_2^{(t)} \right] \end{aligned}$$

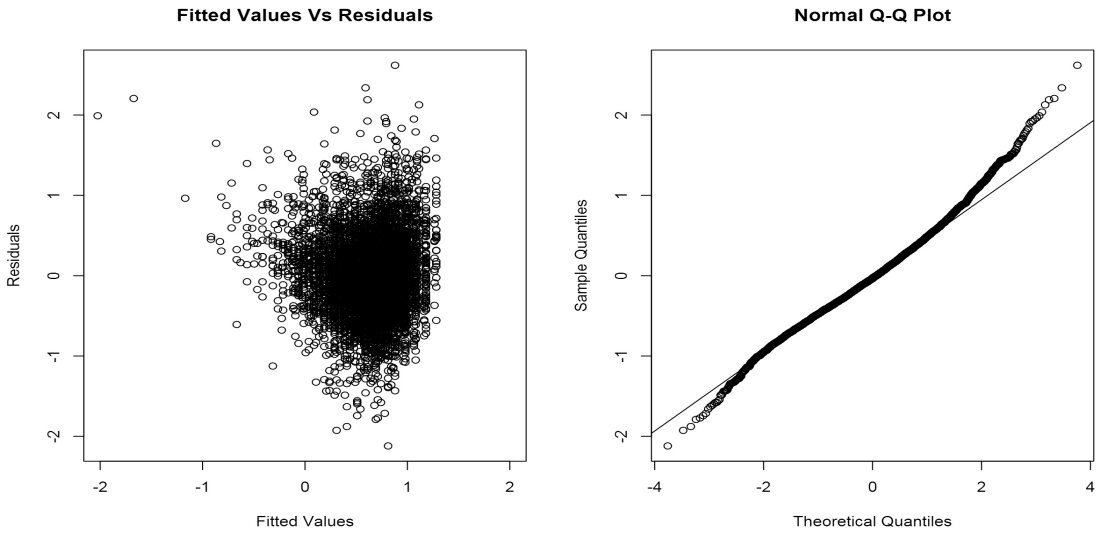

Figure 3.1 Model Diagnostics of Model (3.3)

and

$$\begin{aligned} Q_2(\theta_2|\hat{\theta}_1^{(t)},\theta_2^{(t)}) &= \sum_{i\in S_a} w_{ia}\mathrm{E}\left[-\frac{1}{2}\log(\sigma_u^2)-\frac{1}{2\sigma_u^2}\left(y_{2i}-\alpha_0-\alpha_1 y_{1i}\right)^2 \,\middle|\, \mathbf{x}_i, y_{1i}; \hat{\theta}_1^{(t)},\theta_2^{(t)}\right] \\ &+ \sum_{i\in S_b} w_{ib}\mathrm{E}\left[-\frac{1}{2}\log(\sigma_u^2)-\frac{1}{2\sigma_u^2}\left(y_{2i}-\alpha_0-\alpha_1 y_{1i}\right)^2 \,\middle|\, \mathbf{x}_i, y_{2i}; \hat{\theta}_1^{(t)},\theta_2^{(t)}\right], \end{aligned}$$

where the conditional distribution for

$$f(y_1|\mathbf{x},y_2;\theta_1,\theta_2)=\frac{f_1(y_1|\mathbf{x};\theta_1)f_2(y_2|y_1;\theta_2)}{\int f_1(y_1|\mathbf{x};\theta_1)f_2(y_2|y_1;\theta_2)dy_1}$$

is also normal as in (3.4), evaluated at $\theta_1=\hat{\theta}_1^{(t)}$ and $\theta_2=\hat{\theta}_2^{(t)}$.

[M-step] Update θ_1 by maximizing $Q_1(\theta_1|\theta_1^{(t)},\theta_2^{(t)})$ with respect to θ_1 and update θ_2 by maximizing $Q_2(\theta_2|\hat{\theta}_1^{(t)},\theta_2^{(t)})$ with respect to θ_2.

Based on the estimated parameters $\hat{\theta}_1$ and $\hat{\theta}_2$, the best predictor of y_1 of the FTF sample is obtained as a mean of the predictive distribution, which is a conditional expectation of y_1 given $\mathbf{x}$ and y_2. That is,

$$\hat{y}_{1i}^{*}=\widehat{E}\left(y_{1i}|\mathbf{x}_i,y_{2i}\right)=\frac{\hat{\boldsymbol{\beta}}\mathbf{x}_i/\hat{\sigma}_e^2+\hat{\alpha}_1(y_{2i}-\hat{\alpha}_0)/\hat{\sigma}_u^2}{1/\hat{\sigma}_e^2+\hat{\alpha}_1^2/\hat{\sigma}_u^2}$$

is the best prediction of y_{1i} in the FTF sample that correct for measurement errors in y_{2i}.

Using the counterfactual values of the FTF sample and observations of the FFP sample, we can construct a composite estimator that combines two values. The combined estimator is

$$\bar{y}^*_{com} = \frac{\sum_{i\in S_a} w_{ia} y_{1i} + \sum_{i\in S_b} w_{ib}\hat{y}^*_{1i}}{\sum_{i\in S_a} w_{ia} + \sum_{i\in S_b} w_{ib}}, \tag{3.5}$$

where S_a and S_b denote the FFP sample and the FTF sample, respectively.

3.4.1.2 Case 2: Dichotomous Study Variable

When a study variable is dichotomous, such as the HHS indicator in the project, the normal distribution assumption does not hold for both the structural equation model and the measurement error model. In this case, we consider a logistic regression model for the structural equation model and the misclassification model is used instead of the measurement error model (Buonaccorsi, 2010). The structural equation model for y_1 is

$$y_{1i}|\mathbf{x}_i \sim \text{Ber}(r_i)$$

where $\mathbf{x}_i = (\mathbf{x}_{1i}, x_{2i})$ and

$$r_i = \frac{\exp(\boldsymbol{\beta}_1\mathbf{x}_{1i} + \beta_2 x_{2i})}{1+\exp(\boldsymbol{\beta}_1\mathbf{x}_{1i} + \beta_2 x_{2i})},$$

where $\mathbf{x}_{1i}$ is a department indicator and x_{2i} is a variable indicating total number of household members. The misclassification model is given

$$f(y_{2i}|y_{1i}) = p^{y_{1i}y_{2i}}(1-p)^{y_{1i}(1-y_{2i})}q^{(1-y_{1i})y_{2i}}(1-q)^{(1-y_{1i})(1-y_{2i})}$$

where $p = P(y_{2i}=1|y_{1i}=1)$ and $q = P(y_{2i}=1|y_{1i}=0)$ are the misclassification parameters.

Denote the parameters $\theta_1 = (\boldsymbol{\beta}_1, \beta_2)$ and $\theta_2 = (p, q)$. Then, the implementation of the EM algorithm via parametric fractional imputation involves the following steps:

[E-step]

$$\begin{aligned} Q_1(\theta_1|\theta_1^{(t)}, \theta_2^{(t)}) &= \sum_{i\in S_a} w_{ia}\left[y_{1i}(\boldsymbol{\beta}_1\mathbf{x}_{1i} + \beta_2 x_{2i}) - \log\left\{1+\exp(\boldsymbol{\beta}_1\mathbf{x}_{1i} + \beta_2 x_{2i})\right\}\right] \\ &+ \sum_{i\in S_b} w_{ib}\sum_{j=1}^{2} w_{1i}^{*(j)}\left[y_{1i}^{*(j)}(\boldsymbol{\beta}_1\mathbf{x}_{1i} + \beta_2 x_{2i}) - \log\left\{1+\exp(\boldsymbol{\beta}_1\mathbf{x}_{1i} + \beta_2 x_{2i})\right\}\right] \end{aligned}$$

and

$$
\begin{aligned}
Q_2(\theta_2|\hat{\theta}_1^{(t)},\theta_2^{(t)}) &= \sum_{i\in S_a} w_{ia} \sum_{j=1}^{2} w_{2i}^{*(j)} \left[y_{2i}^{*(j)} \{ y_{1i} \log p + (1-y_{1i}) \log q \} \right] \\
&+ \sum_{i\in S_a} w_{ia} \sum_{j=1}^{2} w_{2i}^{*(j)} \left[(1-y_{2i}^{*(j)}) \{ y_{1i} \log(1-p) + (1-y_{1i}) \log(1-q) \} \right] \\
&+ \sum_{i\in S_b} w_{ib} \sum_{j=1}^{2} w_{1i}^{*(j)} \left[y_{1i}^{*(j)} \{ y_{2i} \log p + (1-y_{2i}) \log(1-p) \} \right] \\
&+ \sum_{i\in S_b} w_{ib} \sum_{j=1}^{2} w_{1i}^{*(j)} \left[(1-y_{1i}^{*(j)}) \{ y_{2i} \log q + (1-y_{2i}) \log(1-q) \} \right],
\end{aligned}
$$

where $y_{ki}^{*(1)} = 1$ and $y_{ki}^{*(2)} = 0$ for $k = 1, 2$ and

$$
\begin{aligned}
w_{1i}^{*(j)} &= P(y_{1i}^{*(j)}|y_{2i}, \mathbf{x}_i) \\
&\propto f(y_{1i}^{*(j)}|\mathbf{x}_i) P(y_{2i}|y_{1i}^{*(j)}) \\
w_{2i}^{*(j)} &= P(y_{2i}^{*(j)}|y_{1i}, \mathbf{x}_i) \\
&= P(y_{2i}^{*(j)}|y_{1i}),
\end{aligned}
$$

where $\sum_j w_{1i}^{*(j)} = 1$ and $\sum_j w_{2i}^{*(j)} = 1$.

[M-step] Update θ_1 by maximizing $Q_1(\theta_1|\theta_1^{(t)},\theta_2^{(t)})$ with respect to θ_1 and update θ_2 by maximizing $Q_2(\theta_2|\hat{\theta}_1^{(t)},\theta_2^{(t)})$ with respect to θ_2.

The best predictor of y_{1i} of the FTF sample can be written by

$$
\hat{y}_{1i}^{*} = \widehat{E}(y_{1i}|\mathbf{x}_i, y_{2i}) = \sum_{j=1}^{2} w_{1i}^{*(j)} y_{1i}^{*(j)} \tag{3.6}
$$

and the composite estimator combining two samples can be calculated as (3.5) using (3.6).

3.4.2 Variance Estimation of the Combined Estimator

For variance estimation of the combined estimator, replicate variance estimation method is applied. More precisely, we used the bootstrap method of Rao and Wu (1988). For each bootstrap

dataset $D_{(b)}$, $b = 1, \cdots, B$, we can calculate estimates for the specific bootstrap sample, say $\hat{\mu}_{(b)}$. Then, the bootstrap approach computes the estimated variance of estimator $\bar{y}$ by

$$\hat{V}(\bar{y}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\mu}_{(b)} - \hat{\bar{\mu}}\right)^2,$$

where $\hat{\bar{\mu}} = B^{-1} \sum_{b=1}^{B} \hat{\mu}_{(b)}$ is the mean of B bootstrap estimates. We used $B = 500$ in this study.

3.4.3 Results

In this section, results of the two examples in Section 2.4 are presented: the PCE indicator's result is shown in Table 3.5 and the HHS indicator's result is shown in Table 3.6. Both tables contain mean estimates of the FFP project (FFP), mean estimates of the FTF project (FTF) and combined mean estimates (Combined) using the original estimate of the FFP project and the new FTF mean estimates. Also, standard errors of each mean estimate are also reported.

Table 3.5 PCE Indicator: Mean Estimates (Standard Errors) of the FFP Project, Mean Estimates (Standard Errors) of the FTF Project, and Combined Mean Estimates (Standard Errors)

Department	FFP	FTF	Combined
San Marcos	0.558	1.165	0.563
	(0.030)	(0.038)	(0.026)
Totonicapan	0.388	0.895	0.331
	(0.030)	(0.085)	(0.028)
Quiche	0.382	1.045	0.396
	(0.030)	(0.031)	(0.026)
Huehuetenango	0.456	1.140	0.479
	(0.044)	(0.036)	(0.027)
Quetzaltenango	0.695	1.325	0.795
	(0.044)	(0.232)	(0.043)

Mean estimates of the FFP sample and the new mean estimates of the FTF sample are combined using (3.5) in order to obtain the composite estimates and the result is listed in the last column of the both tables. From the results in Table 3.5 and Table 3.6, we find that the combined estimator provides reasonable estimates for the population mean with smaller standard errors.

Table 3.6 HHS Indicator: Proportion Estimates (Standard Errors) of the FFP Project, Proportion Estimates (Standard Errors) of the FTF Project, and Combined Proportion Estimates (Standard Errors) (%)

Department	FFP	FTF	Combined
San Marcos	3.76	15.35	3.77
	(1.01)	(2.22)	(1.00)
Totonicapan	11.79	15.01	12.08
	(1.70)	(6.00)	(1.60)
Quiche	7.13	9.73	7.19
	(1.50)	(1.57)	(1.42)
Huehuetenango	8.91	15.58	8.75
	(1.90)	(2.00)	(1.90)
Quetzaltenango	6.84	9.94	6.85
	(1.80)	(8.25)	(1.70)

Estimates of parameters of the measurement error model for PCE variable are $(\hat{\alpha}_0, \hat{\alpha}_1) = (0.261, 0.732)$. The $\hat{\alpha}_0 = 0.261$ can be thought of as the mean of the measurement error model and it can explain why some combined estimates are outside the confidence interval of the estimate from the FTF.

In some cases, the combined estimate is not in between the FFP and the FTF. For example, the combined estimate of PCE in Totonicapan and the combined estimate of HHS in Huehuetenango are smaller than the FFP and the FTF. The new estimate of the FTF, which was adjusted for measurement errors, is even smaller than the FFP and it leads to the combined estimate that is not between the two original values. The new FTF estimate is not tabulated in the result, but the new estimate of PCE in Totonicapan is 0.275 and the new one of HHS in Huehuetenango is 8.70, which are smaller than the FFP for both cases.

3.5 Discussion

This study suggests a new approach to combine information from two surveys using the measurement error model approach and it can be generalized to combine more than two sources of information. Using a structural equation model and a measurement error model, we present a

guidance on data integration with illustration of the work sponsored by FANTA. The results shown in Table 3.5 and Table 3.6 indicate that the reference estimate and the counterfactual predicted values of the other sample can be used to produce the combined estimates.

The choice of a benchmark among several surveys can be decided in various ways. We considered a smaller mean squared error as a criterion in our study. If we have auxiliary information, such as previous experiences on the surveys, it can be used to determine a gold standard among several surveys.

The proposed approach can be applied to combine more than two survey data. Similarly, we can implement the method as follows: set one survey data as a benchmark, remove measurement errors existing in the remaining survey data and calculate the composite estimator using the estimates from the surveys. Also, multivariate modeling for the structural equation model can provide a more efficient estimation. Such extension will be a topic for future research.

Bibliography

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. CRC Press.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., and Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (ivr) and the internet. *Social Science Research*, 38(1):1–18.

Fornell, C. and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, pages 39–50.

Fuller, W. A. (2003). Estimation for multiple phase samples. In R.L. Chambers and C.J. Skinner (eds.). *Analysis of Survey Data*, pages 307–322.

Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.

Hidiroglou, M. (2001). Double sampling. *Survey methodology*, 27(2):143–154.

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

Kim, J. K., Berg, E., and Park, T. (2016). Statistical matching using fractional imputation. *Survey Methodology*, 42:19–40.

Kim, J. K. and Rao, J. N. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99(1):85–100.

Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44(247):380–387.

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

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99(468):1131–1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):27–48.

Park, S., Kim, J. K., and Park, S. (2016). An imputation approach for handling mixed-mode surveys. *The Annals of Applied Statistics*, 10(2):1063–1085.

Rao, J. N. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the american statistical association*, 83(401):231–241.

Renssen, R. H. and Nieuwenbroek, N. J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92(437):368–374.

USAID (2013). *Baseline study of Food For Peace Title II development food assistance program in Guatemala*. https://www.usaid.gov/data/dataset/beafc8ed-c5cf-41a0-84a4-19303c309516.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Canadian Journal of Statistics*, 32(1):15–26.

Ybarra, L. M. and Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4):919–931.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, 85(412):986–1001.

CHAPTER 4. MASS IMPUTATION FOR TWO-PHASE SAMPLING

A paper to be submitted to Statistica Sinica

Seho Park[1] and Jae Kwang Kim[1]

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 and Sitter (1995) introduce mass imputation for two-phase sampling and its variance estimation under simple random sample in both phases. In this paper, we extend the Rao-Sitter method to the general sampling design. In addition, we also consider a special case of non-nested two-phase sampling where the second-phase sample is a non-probability sample. The proposed method requires the outcome model be correctly specified. Two simulation studies are performed to examine the performance of the proposed methods.

Key Words: outcome-dependent sampling, auxiliary information, domain estimation, non-nested two-phase sampling, non-probability sample.

4.1 Introduction

Two-phase sampling, which was first proposed by Neyman (1938), is a convenient and economical sampling design when the sample selection is conducted in two phases. In phase one, a large sample is collected from population and a relatively inexpensive auxiliary variable x is measured. In phase two, incorporating the auxiliary information obtained from the first-phase into the second-phase sampling design, a smaller sample is drawn and a variable of interest y, which is expensive to measure, is collected.

[1]Department of Statistics, Iowa State University

Two-phase sampling or double sampling is effective in increasing the precision of estimates by using auxiliary information and is a cost-effective technique that enables us to consider two layers of information rather than one layer of information available from single-phase sampling (Hidiroglou and Särndal, 1998). Two-phase sampling is also called outcome-dependent sampling since the second phase sampling depends on the observations from the first phase sampling. Hidiroglou (2001) and Legg and Fuller (2009) provide comprehensive overviews of two-phase sampling. Moreover, two-phase sampling can efficiently sample a relatively rare and not easily identifiable population (Blair, 1999) or a highly clustered population (Blair and Czaja, 1982).

In addition to the above mentioned advantages, two-phase sampling is powerful as it is applicable to various situations. Separate sample collections from sub-population of responses and non-responses for the cases with rare or low responses is an example of two-phase sampling. For example, case-control studies in epidemiology or choice-based sampling in econometrics are examples of two-phase sampling (Breslow and Holubkov, 1997). For the case-control studies, most of disease cases and relatively small portion of the control cases are sampled separately in order to examine effects of potential risk factors to the disease (Breslow, 2014).

Structure of two-phase sample can be seen as a missing data problem; some are observed and the others are missing. 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 planned missing data and apply an imputation method. This technique, termed synthetic imputation, generates synthetic values for the missing values and use the imputed values for the estimation. It is also called as mass imputation (Kim and Rao, 2012) since it requires generating a large number of synthetic values. In terms of methodological advantage, the mass imputation is more efficient than when we use naive weighting since auxiliary information is used and it is also practically advantageous as of weighting is not necessary to produce estimates (De Waal, 2000).

In the large scale survey, it is sometimes convenient or requested to produce estimates for various domains from the first phase sample when it is very large. Sometimes the first phase sample is hard to handle entirely because of its large size, and estimates of detailed or finer-level domains are

of interest. Estimates for domains, or small area, can be computed using various techniques and mass imputation is one of them (Moore and Robbins, 2004). Breidt et al. (1996) also considered using imputation method for domain estimation and they showed that estimates obtained using mass imputation provide better estimates at finer levels of detail.

Mass imputation for two-phase sampling when both phases use the simple random sampling design has been introduced by Rao and Sitter (1995). In this paper we extend it to the arbitrary sampling designs in each of the two phases. We propose a mass imputation estimator and a replication variance estimation for the estimator for the two-phase sample collected using complex sampling design. Further, we also consider a non-nested two-phase sampling as an extension when the second phase sample is not necessarily selected from a probability sampling. If two samples are selected independently from the same target and one sample observes the auxiliary variable only and the other sample observes the study variable as well as the auxiliary variable, it is called non-nested two-phase sample. Filling in the missing values of study variable in one sample with imputed values, which incorporates information from the other sample, and obtaining improved estimator integrating information from two samples is presented in this paper.

The rest of the paper is organized as follows. In section 2, we introduce notations used throughout the paper and two-phase regression estimator and its known properties. In section 3, we present a 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, non-nested two-phase sampling is considered. Simulation study result is presented in section 6 and we conclude in section 7.

4.2 Basic Setup

To discuss the setup for two-phase sampling, consider a finite population, denoted by $\mathcal{F}_N = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$. 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 $\pi_{2i|1i} = P(i \in A_2 | i \in A_1)$ for $i \in A_2$. The conditional first-order inclusion probability is random in the sense that it depends on the observations from 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; $w_{1i} = \pi_{1i}^{-1}$. 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 estimation of the finite population total of y, denoted as $Y = \sum_{i=1}^{N} y_i$. When the study variable y is observed in the second-phase sample, the population total Y is estimated by two-phase regression estimator defined by

$$\hat{Y}_{tp,reg} = \hat{Y}_2 + (\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_2)\hat{\boldsymbol{\beta}}, \tag{4.1}$$

where

$$\begin{aligned} \hat{\mathbf{X}}_1 &= \sum_{i \in A_1} w_{1i}\mathbf{x}_i, \\ (\hat{\mathbf{X}}_2, \hat{Y}_2) &= \sum_{i \in A_2} w_{1i}w_{2i|1i}(\mathbf{x}_i, y_i), \end{aligned}$$

and $\hat{\boldsymbol{\beta}}$ is obtained using the observations from the second-phase sample. To study the asymptotic properties of the two-phase regression estimator, we assume a sequence of finite populations and samples defined in Isaki and Fuller (1982) with bounded fourth moments of (x_i, y_i). Define $X_N = \sum_{i=1}^{N} \mathbf{x}_i$ and $n_i = |A_i|$ for $i = 1, 2$.

Lemma 4.2.1 *Suppose that the sequence of finite population and samples satisfies the following assumptions:*

A1. $E\left[|\hat{X}_1 - X_N|^2 \mid \mathcal{F}_N\right] = O(n_1^{-1}N^2)$

A2. $E\left[|(\hat{X}_2, \hat{Y}_2) - (X_N, Y)|^2 \mid \mathcal{F}_N\right] = O(n_2^{-1}N^2)$

A3. $E\left[|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N|^2 \mid \mathcal{F}_N\right] = O(n_2^{-1})$ *for some* $\boldsymbol{\beta}_N$.

Then,

$$N^{-1}(\hat{Y}_{tp,reg} - \tilde{Y}_{tp,reg} \mid \mathcal{F}_N) = o_p(n_2^{-1/2}), \tag{4.2}$$

where

$$\tilde{Y}_{tp,reg} = \hat{Y}_2 + (\hat{X}_1 - \hat{X}_2)\boldsymbol{\beta}_N.$$

Note that

$$E(\tilde{Y}_{tp,reg}) = Y$$

and

$$\begin{aligned} V(\tilde{Y}_{tp,reg} - Y) &= V\left[E\left(\tilde{Y}_{tp,reg} \mid A_1\right)\right] + E\left[V\left(\tilde{Y}_{tp,reg} \mid A_1\right)\right] \\ &= V(\hat{Y}_1) + E\left[V(\hat{e}_2 \mid A_1)\right], \end{aligned}$$

where $\hat{e}_2 = \sum_{i \in A_2} w_{1i} w_{2i|1i}(y_i - \mathbf{x}_i \boldsymbol{\beta}_N)$.

Proof. From assumption A3, we can obtain that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_N + O_p(n_2^{-1/2}). \tag{4.3}$$

Using (4.3) and $\hat{X}_1 - \hat{X}_2 = O_p(n_2^{-1/2}N)$, we can obtain that

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

which proves (4.2). ■

Since $\tilde{Y}_{tp,reg}$ is design-unbiased for Y, Lemma 1 implies that the two-phase regression estimator $\hat{Y}_{tp,reg}$ is design-consistent for Y regardless of the form of $\hat{\boldsymbol{\beta}}$.

4.3 Proposed method

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

the second-phase sample to develop a model generating imputed values for unobserved variables using the observed relationships among the variables (Fetter, 2001).

Denote $A_2 \cup A_2^c = A_1$. Then, a mass imputation estimator for population total is written by

$$\hat{Y}_{imp} = \sum_{i \in A_2} w_{1i} y_i + \sum_{i \in A_2^c} w_{1i} \hat{y}_i, \tag{4.4}$$

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

Our goal is to find a mass imputation method for two-phase sampling that makes the imputation estimator (4.4) algebraically equivalent to the two-phase regression estimator in (4.1).

Lemma 4.3.1 *If $\hat{\boldsymbol{\beta}}$ satisfies*

$$\sum_{i \in A_2} w_{1i}(w_{2i|1i} - 1)(y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}) = 0, \tag{4.5}$$

then the mass imputation estimator

$$\hat{Y}_{imp} = \sum_{i \in A_2} w_{1i} y_i + \sum_{i \in A_2^c} w_{1i} \hat{y}_i$$

where $\hat{y}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}}$ is algebraically equivalent to the two-phase regression estimator defined in (4.1).

Proof. Condition (4.5) can be expressed as

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

and so,

$$\sum_{i \in A_2} w_{1i} w_{i2|1}(y_i - \hat{y}_i) + \sum_{i \in A_2} w_{1i} \hat{y}_i = \sum_{i \in A_2} w_{1i} y_i \tag{4.6}$$

Substituting (4.6) into (4.4), we have

$$\begin{aligned}
\hat{Y}_{imp} &= \sum_{i \in A_2} w_{1i} y_i + \sum_{i \in A_2^c} w_{1i} \hat{y}_i \\
&= \sum_{i \in A_1} w_{1i} \hat{y}_i + \sum_{i \in A_2} w_{1i} w_{i2|1}(y_i - \hat{y}_i) \qquad (4.7) \\
&= \sum_{i \in A_2} w_{1i} w_{2i|1i} y_i + \left(\sum_{i \in A_1} w_{1i} \mathbf{x}_i - \sum_{i \in A_2} w_{1i} w_{2i|1i} \mathbf{x}_i \right) \hat{\boldsymbol{\beta}} \\
&= \hat{Y}_2 + (\hat{X}_1 - \hat{X}_2)' \hat{\boldsymbol{\beta}},
\end{aligned}$$

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

To discuss condition (4.5), note that if $\hat{\boldsymbol{\beta}}$ is of the form

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

and $w_{2i|1i} - 1$ is included in the column space of $\mathbf{x}_i$, then condition (4.5) is satisfied. Thus, the mass imputation estimator (4.4) is also design-consistent for the population total Y. Condition (4.5) is similar in spirit to internal bias calibration (IBC) condition of Firth and Bennett (1998).

Note that the first term on the right side of (4.7) is defined as a projection estimator or synthetic estimator in Kim and Rao (2012) and the second term in equation (4.7) can be considered as a bias-correction term of the projection estimator. The projection estimator is asymptotically design-unbiased if the intercept is included in the column space of $\mathbf{x}_i$.

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 and Kim, 2005) for mass imputation. To do this, we can write

$$\hat{Y}_{FI} = \sum_{i \in A_2} w_{1i} y_i + \sum_{i \in A_2^c} w_{1i} (\hat{y}_i + \sum_{j \in A_2} w_{ij}^* \hat{e}_j), \tag{4.8}$$

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 $i \in A_2^c$. By (4.5), if we choose

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

then we have $\sum_{j \in A_2} w_{ij}^* \hat{e}_j = 0$ and (4.8) is algebraically equivalent to (4.4).

Note that we can express (4.8) as

$$\hat{Y}_{FEFI} = \sum_{i \in A_2} w_{1i} y_i + \sum_{i \in A_2^c} w_{1i} \sum_{j \in A_2} w_{ij}^* y_{ij}^*, \tag{4.9}$$

where $y_{ij}^* = \hat{y}_i + \hat{e}_j$. Because (4.9) uses all possible imputed values for imputation, it can be called fully efficient fractional imputation (FEFI) estimator (Fuller and Kim, 2005).

4.4 Replication Variance Estimation

In this section, we consider a replication variance estimation of the mass imputation estimator in (4.4). Jackknife variance estimation is considered by Rao and Sitter (1995) under the two-phase sampling with simple random sampling for both phases and it can be extended to general designs by augmenting $\mathbf{x}_i$ to include $w_{2i|1i} - 1$. Let the replicate variance estimator for the first-phase sample estimator of total is

$$\hat{V}_1(\hat{T}_1) = \sum_{k=1}^{L} c_k \left(\hat{T}_1^{(k)} - \hat{T}_1\right)^2 \tag{4.10}$$

where $\hat{T}_1^{(k)} = \sum_{i\in A_1} w_{1i}^{(k)} y_i$ is the k-th replicate of estimated total $\hat{T}_1 = \sum_{i\in A_1} w_{1i} y_i$, $k = 1, \cdots, L$, 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 has a form of

$$\hat{V}(\hat{Y}_{imp}) = \sum_{k=1}^{L} c_k \left(\hat{Y}_{imp}^{(k)} - \hat{Y}_{imp}\right)^2, \tag{4.11}$$

where

$$\hat{Y}_{imp}^{(k)} = \sum_{i\in A_2} w_{1i}^{(k)} y_i + \sum_{i\in A_2^c} w_{1i}^{(k)} \mathbf{x}_i \hat{\boldsymbol{\beta}}^{(k)} \tag{4.12}$$

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 $\hat{Y}_{imp}^{(k)}$ is the k^{th} replicate of $\hat{Y}_{imp}$ using 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 in the second-phase. That is, we assume that the second-phase sampling is Bernoulli with $\pi_{2i|1i}$. Kim and Yu (2011) also consider the unequal probability Poisson sampling within the second-phase strata and Fuller (1998) observes that Poisson sampling for second-phase sample is a good approximation and has little impact on the variance estimation of the mean.

Theorem 2 *Assume that a finite population is a 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* $\mathbf{x}_i$. *Assume*

that

$$E\left[|\hat{T}_{1z} - T_{1z}|^2|\mathcal{F}_N\right] = O(n_1^{-1}N^2)$$
$$E\left[|\hat{T}_{2z} - T_{2z}|^2|\mathcal{F}_N\right] = O(n_2^{-1}N^2) \tag{4.13}$$

where

$$\left(\hat{T}_{1z}, \hat{T}_{2z}\right) = \left(\sum_{i\in A_1} w_{1i}z_i, \sum_{i\in A_2} w_{1i}w_{2i|1i}z_i\right).$$

are total estimators of variable z for the first-phase and the second-phase sample, respectively. Assume that

$$V(\hat{T}_{1y}|\mathcal{F}_N) \leq K_M V(\hat{T}_{y,SRS}|\mathcal{F}_N), \tag{4.14}$$

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

$$n_1 N^{-2} V\left(\sum_{i\in A_1} w_{1i}y_i|\mathcal{F}_N\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\Omega_{ij}y_i y_j \tag{4.15}$$

where the coefficients Ω_{ij} satisfy

$$\sum_{i=1}^{N}|\Omega_{ij}| = O(N^{-1}). \tag{4.16}$$

Let $\hat{V}_1(\hat{T}_1)$ be the first-phase sample replicate estimator of the variance of $\hat{T}_1$ given in (4.10) and assume

$$E\left\{\left[\frac{\hat{V}_1(\hat{T}_1)}{V(\hat{T}_1|\mathcal{F}_N)} - 1\right]^2|\mathcal{F}_N\right\} = o(1) \tag{4.17}$$

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

$$E\left\{\left[c_k\left(\hat{T}_1^{(k)} - \hat{T}_1\right)^2\right]^2|\mathcal{F}_N\right\} < K_T L^{-2}\left[V(\hat{T}_1|\mathcal{F}_N)\right]^2 \tag{4.18}$$

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

$$c_{kN} = O(1). \tag{4.19}$$

Then, the jackknife variance estimator of form (4.11) satisfies

$$\hat{V}_{JK}(\hat{Y}_{imp}) = V(\hat{Y}_{imp} \mid \mathcal{F}_N) - \sum_{i=1}^{N} \kappa_i^{-1}(1-\kappa_i)e_i^2 + o_p(n_2^{-1}N^2), \tag{4.20}$$

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

For the proof see appendix A.

The bias of $\hat{V}_{JK}(\hat{Y}_{imp})$ has order of $O(N)$ and it can be estimated unbiasedly by

$$\sum_{i\in A_2} w_{1i}\pi_{2i|1i}^{-1}(1-\pi_{2i|1i})\hat{e}_i^2,$$

where $\hat{e}_i = y_i - x_i\hat{\boldsymbol{\beta}}$. The second term in (4.20) is small relative to the first term if the first-phase sampling rate, n_1/N, is small. Then, replicate variance estimator (4.11) can be used for the variance of mass imputation estimator of two-phase sample.

We consider the replication method for the variance estimation of the FEFI estimator. A k^{th} replicate for the FEFI estimator is

$$\hat{Y}_{FEFI}^{(k)} = \sum_{i\in A_2} w_{1i}^{(k)} y_i + \sum_{i\in A_2^c} w_{1i}^{(k)} \sum_{j\in A_2} w_{ij}^{*(k)} y_{ij}^*, \tag{4.21}$$

where

$$w_{ij}^{*(k)} = \frac{w_{1j}^{(k)}(w_{2j|1j}-1)}{\sum_{j\in A_2} w_{1j}^{(k)}(w_{2j|1j}-1)} \tag{4.22}$$

is a k^{th} replicate of fractional weight. The following theorem provides the asymptotic property of the replicate variance estimator of the FEFI estimator.

Theorem 3 *Assume that*

$$\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}} = O_p(n_2^{-1}). \tag{4.23}$$

Then, the jackknife variance estimator of the FEFI estimator, which has a form of

$$\hat{V}_{FEFI} = \sum_{k=1}^{L} c_k \left(\hat{Y}_{FEFI}^{(k)} - \hat{Y}_{FEFI}\right)^2,$$

satisfies

$$\hat{V}_{FEFI} = V(\hat{Y}_{FEFI}) - \sum_{i=1}^{N} \kappa_i^{-1}(1-\kappa_i)e_i^2 + o_p(n_2^{-1}N^2), \tag{4.24}$$

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

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

$$\tilde{Y}^{(k)}_{FEFI} = \sum_{i\in A_2} w^{(k)}_{1i} y_i + \sum_{i\in A_2^c} w^{(k)}_{1i} \sum_{j\in A_2} w^{*(k)}_{ij} y^{*(k)}_{ij} \tag{4.25}$$

when $y^{*(k)}_{ij} = \hat{y}^{(k)}_i + \hat{e}^{(k)}_j = \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(k)} + \left(y_j - \mathbf{x}_j\hat{\boldsymbol{\beta}}^{(k)}\right)$ is a k^{th} replicate of y^*_{ij}. Then, a difference between (4.21) and (4.25) is

$$\begin{aligned}
&\tilde{Y}^{(k)}_{FEFI} - \hat{Y}^{(k)}_{FEFI} \\
&= \sum_{i\in A_2^c} w^{(k)}_{1i} \sum_{j\in A_2} w^{*(k)}_{ij} y^{*(k)}_{ij} - \sum_{i\in A_2^c} w^{(k)}_{1i} \sum_{j\in A_2} w^{*(k)}_{ij} y^*_{ij} \\
&= \sum_{i\in A_2^c} w^{(k)}_{1i} \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(k)} + \sum_{i\in A_2^c} w^{(k)}_{1i} \frac{\sum_{j\in A_2} w^{(k)}_{1i}(w_{2j|1j}-1)\left(y_j - \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(k)}\right)}{\sum_{j\in A_2} w^{(k)}_{1i}(w_{2j|1j}-1)} \\
&- \sum_{i\in A_2^c} w^{(k)}_{1i} \mathbf{x}_i\hat{\boldsymbol{\beta}} + \sum_{i\in A_2^c} w^{(k)}_{1i} \frac{\sum_{j\in A_2} w^{(k)}_{1i}(w_{2j|1j}-1)\left(y_j - \mathbf{x}_i\hat{\boldsymbol{\beta}}\right)}{\sum_{j\in A_2} w^{(k)}_{1i}(w_{2j|1j}-1)} \\
&= \left[\sum_{i\in A_2^c} w^{(k)}_{1i}\mathbf{x}_i - \frac{\sum_{i\in A_2^c} w^{(k)}_{1i}}{\sum_{j\in A_2} w^{(k)}_{1i}(w_{2j|1j}-1)} \sum_{j\in A_2} w^{(k)}_{1i}(w_{2j|1j}-1)\mathbf{x}_j\right]\left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right) \\
&= \left[\sum_{i\in A_1} (1-\delta_i) w^{(k)}_{1i}\mathbf{x}_i - \frac{\sum_{i\in A_1}(1-\delta_i) w^{(k)}_{1i}}{\sum_{j\in A_1} \delta_i w^{(k)}_{1i}(w_{2j|1j}-1)} \sum_{j\in A_1} \delta_i w^{(k)}_{1i}(w_{2j|1j}-1)\mathbf{x}_j\right] \times \left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right),
\end{aligned} \tag{4.26}$$

where the last equation is expressed using δ_i that is defined by

$$\delta_i = \begin{cases} 1 & \text{if } i \in A_2 \text{ when } i \in A_1 \\ 0 & \text{if } i \notin A_2 \text{ when } i \in A_1. \end{cases}$$

Let's denote $\hat{X}^{(k)}_{2c}$, $\hat{X}^{(k)}_2$ and $\hat{X}^{(k)}_{1c}$as

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

and

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

Further, let $\hat{N}_{2c}^{(k)}$, $\hat{N}_{1c}^{(k)}$ and $\hat{N}_{2}^{(k)}$ be denoted similarly using 1 instead of $\mathbf{x}_i$. Then, (4.26) can be written by

$$\left[\hat{X}_{2c}^{(k)} - \frac{\hat{N}_{2c}^{(k)}}{\hat{N}_{2}^{(k)}}\hat{X}_{2}^{(k)}\right]\left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right). \tag{4.27}$$

Note that

$$E\left(\hat{X}_{2c}^{(k)}\right) = \hat{X}_{1c}^{(k)} = E\left(\hat{X}_{2}^{(k)}\right) \tag{4.28}$$

and

$$E\left(\hat{N}_{2c}^{(k)}\right) = \hat{N}_{1c}^{(k)} = E\left(\hat{N}_{2}^{(k)}\right) \tag{4.29}$$

Also, we have

$$\frac{1}{N}\hat{N}_{2c}^{(k)} = \frac{1}{N}\hat{N}_{1c}^{(k)} + O_p(n_2^{-1/2})$$

and

$$\frac{1}{N}\hat{N}_{2}^{(k)} = \frac{1}{N}\hat{N}_{1c}^{(k)} + O_p(n_2^{-1/2}).$$

Using the Taylor expansion, the ratio term in (4.27) can be expressed as

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

based on (4.29). Hence, the first term in (4.27) can be expressed as

$$\begin{aligned}
\hat{X}_{2c}^{(k)} - \frac{\hat{N}_{2c}^{(k)}}{\hat{N}_{2}^{(k)}}\hat{X}_{2}^{(k)} &= \left[\hat{X}_{1c}^{(k)} + O_p(n_2^{-1/2}N)\right] - \left[1 + O_p(n_2^{-1/2})\right]\left[\hat{X}_{1c}^{(k)} + O_p(n_2^{-1/2}N)\right] \\
&= \left[\hat{X}_{1c}^{(k)} + O_p(n_2^{-1/2}N)\right] - \left[\hat{X}_{1c}^{(k)} + O_p(n_2^{-1/2}N)\right] \\
&= O_p(n_2^{-1/2}N),
\end{aligned} \tag{4.30}$$

based on (4.28). By combining (4.23) and (4.30), we have

$$\hat{Y}_{FEFI}^{(k)} = \tilde{Y}_{FEFI}^{(k)} + o_p(n_2^{-1}N). \tag{4.31}$$

With choice of $w_{ij}^{*(k)}$ given by (4.22), we can show that $\tilde{Y}_{FEFI}^{(k)}$ in (4.25) is algebraically equivalent to k^{th} replicate of $\hat{Y}_{imp}$ in (4.12). That is,

$$\begin{aligned}
\tilde{Y}_{FEFI}^{(k)} &= \sum_{i\in A_2} w_{1i}^{(k)} y_i + \sum_{i\in A_2^c} w_{1i}^{(k)} \sum_{j\in A_2} w_{ij}^{*(k)} \left(\mathbf{x}_i\hat{\boldsymbol{\beta}}^{(k)} + (y_j - \mathbf{x}_j\hat{\boldsymbol{\beta}}^{(k)})\right) \\
&= \sum_{i\in A_2} w_{1i}^{(k)} y_i + \sum_{i\in A_2^c} w_{1i}^{(k)} \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(k)} + \sum_{i\in A_2^c} w_{1i}^{(k)} \sum_{j\in A_2} w_{ij}^{*(k)} \left(y_j - \mathbf{x}_j\hat{\boldsymbol{\beta}}^{(k)}\right) \\
&= \sum_{i\in A_2} w_{1i}^{(k)} y_i + \sum_{i\in A_2^c} w_{1i}^{(k)} \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(k)}, \qquad (4.32)
\end{aligned}$$

where the last equality follows by $\sum_{j\in A_2} w_{ij}^{*(k)} \hat{e}_j^{(k)} = 0$. Since the FEFI estimator (4.9) is equivalent to the mass imputation estimator (4.4), we have

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

based on (4.31) and (4.32). By Theorem 1, the result (4.24) follows. ■

4.5 Non-nested two-phase sampling

We now extend the idea of two-phase sampling to data integration, which is an area of research on combining information from multiple sources. We will consider the case of combining two data sources, where the first one, sample A, observes the auxiliary variable (X) only and the second one, sample B, observes the study variable (Y) in addition to the auxiliary variable, and the two samples are selected independently from the same target population. If the two samples are independently selected, then it is called non-nested two-phase sampling (Hidiroglou, 2001).

Under the non-nested two-phase sampling, our goal is to combine information from two sources to get an improved estimator over the naive approach using sample B only. Merkouris (2004, 2010) presents methods for optimal estimation under this setup, and Kim and Rao (2012) considered mass imputation for sample A using information from sample B observations. The approach in Kim and Rao (2012) is design-based and the two samples are probability samples.

Data integration for non-nested two-phase sampling can be extended to the case where sample B, observing (X, Y), is a non-probability sample, which is subject to inherent selection bias. By

assuming that the prediction model for Y, denoted by $f(Y|X)$, can be estimated from sample B, we can obtain the same mass imputation estimator for the sample A. Unlike the setup of Kim and Rao (2012), the proposed estimator is no longer design-consistent, but is still justified under the design-model framework, where the model refers to the superpopulation model corresponding to $f(Y|X)$.

Under this setup, Rivers (2007) considers a mass imputation estimator based on nearest neighbor imputation. Nearest neighbor imputation is a nonparametric method of imputation that does not require any parametric model assumptions. In this paper, we make a parametric moment assumption,

$$E(Y|\mathbf{x}) = m(\mathbf{x}_i; \boldsymbol{\beta}) \tag{4.33}$$

for some $\boldsymbol{\beta}$ with known function $m(\cdot)$ and assume that model (4.33) holds for sample B.

4.5.1 Proposed Estimator

We now formally introduce the setup and notation for data integration of survey sample data and non-probability sample data. Let A denote the index set of survey sample data and B denote that of the non-probability sample. Let $n_A = |A|$ and $n_B = |B|$. Table 4.1 presents the setup of our data integration problem.

Table 4.1 Data Structure

Data	X	Y	Representativeness
A	✓		Yes
B	✓	✓	No

Under this setup, we are interested in estimating population mean $\theta_N = N^{-1}\sum_{i=1}^{N} y_i$. To achieve this goal, we use a model, such as (4.33), to create an imputed value $\hat{y}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ for each $i \in A$ and construct an imputed estimator of Y given by

$$\hat{\theta}_I = \sum_{i \in A} w_i \hat{y}_i \tag{4.34}$$

where $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ obtained from sample B. Instead of deterministic imputation $\hat{y}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$, one can also consider a stochastic imputation $y_i^* = \hat{y}_i + \hat{e}_i^*$ where $\hat{e}_i^*$ is randomly selected from $\{\hat{e}_i; i \in B\}$.

To justify the mass imputation estimator in (4.34), we first assume missing at random condition of Rubin (1976) for sample B. That is,

$$f(y|\mathbf{x}, \delta = 1) = f(y|\mathbf{x}) \tag{4.35}$$

where δ is defined by

$$\delta_i = \begin{cases} 1 & \text{if } i \in B \\ 0 & \text{if } i \notin B. \end{cases}$$

We assume $p-$dimensional $\mathbf{x}$ but a scalar y. Also, we assume that $\hat{\boldsymbol{\beta}}$ is the unique solution to

$$\sum_{i \in B} \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta})\} h(\mathbf{x}_i; \boldsymbol{\beta}) = 0$$

for some p-dimensional vector $h(\mathbf{x}_i; \boldsymbol{\beta})$. If the variance function $V(y_i|\mathbf{x}_i)$ is specified as $V(y_i|\mathbf{x}_i) = \sigma^2 a(\mathbf{x}_i, \boldsymbol{\beta})$ for some known function $a(\cdot)$, then $h(\mathbf{x}_i; \boldsymbol{\beta}) = \dot{m}(\mathbf{x}_i; \boldsymbol{\beta})/a(\mathbf{x}_i, \boldsymbol{\beta})$ where $\dot{m}(\mathbf{x}_i; \boldsymbol{\beta}) = \partial(\mathbf{x}_i; \boldsymbol{\beta})/\partial\boldsymbol{\beta}$.

Theorem 4 *Assume a sequence of finite populations and samples with bounded fourth moments of* $(\mathbf{x}_i, y_i, m(\mathbf{x}_i; \boldsymbol{\beta}), \dot{m}(\mathbf{x}_i; \boldsymbol{\beta}), h(\mathbf{x}_i; \boldsymbol{\beta}))$. *Assume that* $\boldsymbol{\beta}_0$ *satisfies (4.33) and (4.35) holds. Further assume that:*

(1) Under sample B,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p(n_B^{-1}); \tag{4.36}$$

(2) For each i, $m(\mathbf{x}_i; \boldsymbol{\beta})$ *and* $h(\mathbf{x}_i; \boldsymbol{\beta})$ *are continuous functions of* $\boldsymbol{\beta}$ *in a compact set B containing* $\boldsymbol{\beta}_0$ *as an interior point;*

(3) For each i, $m(\mathbf{x}_i; \boldsymbol{\beta})$ *is differentiable with continuous partial derivatives* $\dot{m}(\mathbf{x}_i; \boldsymbol{\beta})$ *in a compact set containing* $\boldsymbol{\beta}_0$.

Then, the mass imputation estimator (4.34) satisfies

$$\hat{\theta}_I = \tilde{\theta}_I(\boldsymbol{\beta}_0, \mathbf{c}^*) + o_p(n_B^{-1/2}), \tag{4.37}$$

where

$$\tilde{\theta}_I(\boldsymbol{\beta}, \mathbf{c}) = \frac{1}{N}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}) + \frac{1}{n_B}\sum_{i\in B}\{y_i - m(\mathbf{x}_i;\boldsymbol{\beta})\}\,\mathbf{c}^{'} h(\mathbf{x}_i;\boldsymbol{\beta}) \tag{4.38}$$

and

$$\mathbf{c}^* = \left[\frac{1}{n_B}\sum_{i\in B}\dot{m}(\mathbf{x}_i;\boldsymbol{\beta}_0)h'(\mathbf{x}_i;\boldsymbol{\beta}_0)\right]^{-1}\frac{1}{N}\sum_{i=1}^{N}\dot{m}(\mathbf{x}_i;\boldsymbol{\beta}_0). \tag{4.39}$$

Also,

$$E\{\tilde{\theta}_I(\boldsymbol{\beta}_0, \mathbf{c}^*) - \theta_N | B\} = 0, \tag{4.40}$$

where the expectation in (4.40) is with respect to the joint distribution of the sampling design for sample A and the superpopulation model, treating sample B fixed, and

$$\begin{aligned} V\left[\tilde{\theta}_I(\boldsymbol{\beta}_0, \mathbf{c}^*) - \theta_N \middle| B\right] &= V\left[\frac{1}{N}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - \frac{1}{N}\sum_{i=1}^{N} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right] \\ &+ V\left[\frac{1}{n_B}\sum_{i\in B} e_i h(\mathbf{x}_i;\boldsymbol{\beta}_0)'\mathbf{c}^* \middle| B\right]. \end{aligned} \tag{4.41}$$

Proof. To prove (4.37), we consider the class of estimators $\tilde{\theta}_I(\boldsymbol{\beta}, \mathbf{c})$ given in (4.38). Note that mass imputation estimator (4.34) can be expressed by $\hat{\theta}_I = \tilde{\theta}_I(\hat{\boldsymbol{\beta}}, \mathbf{c})$ for all p-dimensional vectors $\mathbf{c}$. Now we wish to find a particular choice of $\mathbf{c}$, say $\mathbf{c}^*$, that satisfies

$$\tilde{\theta}_I(\hat{\boldsymbol{\beta}}, \mathbf{c}^*) = \tilde{\theta}_I(\boldsymbol{\beta}_0, \mathbf{c}^*) + o_p(n_B^{-1/2}). \tag{4.42}$$

Since $\hat{\theta}_I = \tilde{\theta}_I(\hat{\boldsymbol{\beta}}, \mathbf{c})$, (4.42) implies that (4.37). Since we have

$$\tilde{\theta}_I(\hat{\boldsymbol{\beta}}, \mathbf{c}^*) - \tilde{\theta}_I(\boldsymbol{\beta}_0, \mathbf{c}^*) = E\left[\frac{\partial\tilde{\theta}_I(\boldsymbol{\beta}, \mathbf{c}^*)}{\partial\boldsymbol{\beta}}\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(n_B^{-1/2}),$$

by (4.36), we can show (4.42) if the limiting mean function has a zero differential at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ (Randles, 1982), that is

$$E\left[\frac{\partial\tilde{\theta}_I(\boldsymbol{\beta}, \mathbf{c}^*)}{\partial\boldsymbol{\beta}}\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = 0. \tag{4.43}$$

It implies that

$$\begin{aligned} E\left[\frac{\partial\tilde{\theta}_I(\boldsymbol{\beta},\mathbf{c}^*)}{\partial\boldsymbol{\beta}}\right] &= E\left[\frac{\partial}{\partial\boldsymbol{\beta}}\frac{1}{N}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}) + \frac{\partial}{\partial\boldsymbol{\beta}}\frac{1}{n_B}\sum_{i\in B}\{y_i - m(\mathbf{x}_i;\boldsymbol{\beta})\}\, h(\mathbf{x}_i;\boldsymbol{\beta})'\mathbf{c}^*\right] \\ &= E\left[\frac{1}{N}\sum_{i\in A} w_i \dot{m}(\mathbf{x}_i;\boldsymbol{\beta})'\right] - E\left[\frac{1}{n_B}\sum_{i\in B}\dot{m}(\mathbf{x}_i;\boldsymbol{\beta})h(\mathbf{x}_i;\boldsymbol{\beta})'\mathbf{c}^*\right] \\ &+ E\left[\frac{1}{n_B}\sum_{i\in B}\{y_i - m(\mathbf{x}_i;\boldsymbol{\beta})\}\frac{\partial h(\mathbf{x}_i;\boldsymbol{\beta})'}{\partial\boldsymbol{\beta}}\mathbf{c}^*\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\dot{m}(\mathbf{x}_i;\boldsymbol{\beta}) - \frac{1}{n_B}\sum_{i\in B}\dot{m}(\mathbf{x}_i;\boldsymbol{\beta})h(\mathbf{x}_i;\boldsymbol{\beta})'\mathbf{c}^* = 0, \end{aligned}$$

where $E(\cdot)$ denotes the design-model expectation and we used $E(y_i - m(\mathbf{x}_i;\boldsymbol{\beta}_0)) = 0$ in the third equality. Thus, we show (4.43) with $\mathbf{c}^*$ defined in (4.39) and so result (4.37) follows.

Note that

$$\begin{aligned} \tilde{\theta}_I(\boldsymbol{\beta}_0,\mathbf{c}^*) - \theta_N &= \frac{1}{N}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) + \frac{1}{n_B}\sum_{i\in B}\{y_i - m(\mathbf{x}_i;\boldsymbol{\beta}_0)\}\, h(\mathbf{x}_i;\boldsymbol{\beta}_0)'\mathbf{c}^* - \frac{1}{N}\sum_{i=1}^{N} y_i \\ &= \frac{1}{N}\left[\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - \sum_{i=1}^{N} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right] + \left[\frac{1}{n_B}\sum_{i\in B} e_i h(\mathbf{x}_i;\boldsymbol{\beta}_0)'\mathbf{c}^* - \frac{1}{N}\sum_{i=1}^{N} e_i\right], \end{aligned} \tag{4.44}$$

where $e_i = y_i - m(\mathbf{x}_i;\boldsymbol{\beta}_0)$. Since the expectation is design-model expectation treating sample B fixed, it follows that

$$\begin{aligned} E\{\tilde{\theta}_I(\boldsymbol{\beta}_0,\mathbf{c}^*) - \theta_N|B\} &= \frac{1}{N}\left[E\left\{\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\} - \sum_{i=1}^{N} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right] \\ &+ \left[\frac{1}{n_B}\sum_{i\in B} E(e_i|B)h(\mathbf{x}_i;\boldsymbol{\beta}_0)'\mathbf{c}^* - \frac{1}{N}\sum_{i=1}^{N} E(e_i|B)\right] \\ &= 0. \end{aligned}$$

Based on (4.44), the variance of $\tilde{\theta}_I(\boldsymbol{\beta}_0,\mathbf{c}^*)$ is obtained by

$$V\left[\tilde{\theta}_I(\boldsymbol{\beta}_0,\mathbf{c}^*) - \theta_N\middle|B\right] = V_A + V_B + C,$$

where

$$V_A = V\left[\frac{1}{N}\left\{\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - \sum_{i=1}^{N} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\}\right],$$
$$V_B = V\left[\frac{1}{n_B}\sum_{i\in B} e_i h(\mathbf{x}_i;\boldsymbol{\beta}_0)'\mathbf{c}^* \middle| B\right],$$

and

$$\begin{aligned} C &= Cov\left[\frac{1}{N}\left\{\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - \sum_{i=1}^{N} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\}, \frac{1}{n_B}\sum_{i\in B} e_i h(\mathbf{x}_i;\boldsymbol{\beta}_0)'\mathbf{c}^* \middle| B\right] \\ &= \frac{1}{n_B N} E\left[\left\{\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - \sum_{i=1}^{N} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\}\sum_{i\in B} e_i h(\mathbf{x}_i;\boldsymbol{\beta}_0)'\mathbf{c}^* \middle| B\right]. \end{aligned}$$

Note that $\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - \sum_{i=1}^{N} m(\mathbf{x}_i;\boldsymbol{\beta}_0)$ is a function of $\mathbf{x}_i$ and based on a parametric moment assumption (4.33), we have

$$E\left[\sum_{i\in B}\{y_i - m(\mathbf{x}_i;\boldsymbol{\beta}_0)\}\, t(\mathbf{x}_i) \middle| B, X_N\right] = 0,$$

where $t(\mathbf{x}_i)$ is any function of $\mathbf{x}_i$ and $X_N = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$. Hence we have $C = 0$, so (4.41) follows.

■

4.5.2 Variance Estimation

For the variance estimation of the mass imputation estimator (4.34), we are interested in the variance estimation of the approximating random variable in (4.37). Since variance of $\tilde{\theta}_I(\boldsymbol{\beta}_0, \mathbf{c}^*)$ can be treated separately given by (4.41), we estimate V_A and V_B separately for the estimation of $V(\tilde{\theta}_I(\boldsymbol{\beta}_0, \mathbf{c}^*))$. If $\boldsymbol{\beta}_0$ is known, then $m(\mathbf{x}_i;\boldsymbol{\beta}_0)$ is observable for all $i \in A$ and we can use the design-unbiased variance estimator, that is

$$\hat{V}_A = \frac{1}{N^2}\sum_{i\in A}\sum_{i\in A} \Omega_{ij} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) w_j m(\mathbf{x}_j;\boldsymbol{\beta}_0),$$

where $\Omega_{ij} = \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}$ and π_{ij} is a joint inclusion probability for unit i and j assumed to be positive. If $\boldsymbol{\beta}_0$ is unknown, it can be replaced by $\hat{\boldsymbol{\beta}}$ and the estimated variance estimator has a form of

$$\hat{V}_A = \frac{1}{N^2}\sum_{i\in A}\sum_{i\in A} \Omega_{ij} w_i m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) w_j m(\mathbf{x}_j;\hat{\boldsymbol{\beta}}).$$

For the variance estimation of V_B, note that the variance V_B can be written by

$$V_B = \frac{1}{n_B^2} \sum_{i \in B} V(e_i|B) \left\{ h(\mathbf{x}_i; \boldsymbol{\beta}_0)' \mathbf{c}^* \right\}^2 .$$

If we assume a variance function of y, such as $V(y_i|\mathbf{x}_i) = \sigma^2 a(\mathbf{x}_i; \boldsymbol{\beta}_0)$, then we can estimate V_B by

$$\hat{V}_B = \frac{1}{n_B^2} \sum_{i \in B} \hat{\sigma}^2 a(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) \left\{ h(\mathbf{x}_i; \hat{\boldsymbol{\beta}})' \hat{\mathbf{c}}^* \right\}^2 ,$$

where $\hat{\sigma}^2$ is an unbiased estimator of σ^2 and $\hat{\mathbf{c}^*} = \mathbf{c}^*(\hat{\boldsymbol{\beta}})$. If we do not assume any variance function, then we can use $\widehat{V(e_i|B)} = \widehat{E(e_i^2|B)} = e_i^2$ as unbiased estimates and the estimated variance is given by

$$\hat{V}_B = \frac{1}{n_B^2} \sum_{i \in B} \hat{e}_i^2 \left\{ h(\mathbf{x}_i; \hat{\boldsymbol{\beta}})' \hat{\mathbf{c}}^* \right\}^2 ,$$

where $\hat{e}_i = y_i - m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$. Hence, variance of $\tilde{\theta}_I(\boldsymbol{\beta}_0, \mathbf{c}^*)$ can be estimated by

$$\hat{V}(\tilde{\theta}_I(\boldsymbol{\beta}_0, \mathbf{c}^*)) = \hat{V}_A + \hat{V}_B.$$

If $n_A/n_B = o(1)$, then V_B is smaller order than V_A and total variance is dominated by V_A. Otherwise, the two variances both contribute to the total variance. In the big data application, n_B is huge and V_B can be safely ignored.

4.6 Simulation Study

4.6.1 Nested Two-phase Sampling

A small simulation study is presented to study the finite sample performance of the regression imputation estimator and of the replication variance estimator.

The simulation setup has 2×3 factorial structure with three factors, which are as follows:

1. Two artificial finite populations: linear model $y_i = 0.8 + 0.5x_i + z_i + e_i$ where $x_i \sim N(2, 1)$, $e_i \sim N(0, 1)$ and ratio model $y_i = 0.7x_i + z_i + u_i$ 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 a size measure for the unequal probability sampling in the second phase sampling.

2. One sampling design for the first-phase sample: simple random sampling of size $n_1 = 500$.

3. Three sampling designs for the second-phase sample: 1) Simple random sampling of size $n_2 = 80$, 2) Poisson sampling with expected sample size $n_2 = 80$ and 3) Randomized systematic probability proportional to size (PPS) sampling of size $n_2 = 80$.

A finite population of size $N = 100,000$ is generated from each of models. From each of the finite population, the first phase sample of size $n_1 = 500$ was collected 1,000 times by simple random sampling. Then, the second phase sample was collected from the first phase sample using the three sampling designs as follows.

1) Simple random sampling without replacement.

2) Poisson sampling:

   Define δ_i for selecting unit i as follows:

$$\delta_i | I_i = 1 \sim \text{Bernoulli}(\pi_{2i|1i}),$$

   where I_i is an indicator variable having 1 if unit i is included in the first-phase, and having 0 otherwise. We consider a conditional first-order inclusion probability of second phase sample as

$$\pi_{2i|1i} = n_2 \frac{z_i}{\sum_{i \in A_1} z_i},$$

   which depends on the first phase sample.

3) Randomized systematic PPS sampling (RSPPS): We follow the procedure introduced in Thompson and Wu (2008).

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

   b. Denote $q_i = \frac{z_i}{\sum_{i \in A_1} z_i}$ 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

$$\pi_{2i|1i} = n_2 \frac{z_i}{\sum_{i \in A_1} z_i},$$

for $i \in A_1$.

To check performance of our proposed method, we compare three estimators for the population mean $\theta = N^{-1} \sum_{i=1}^{N} y_i$; 1) direct estimator, 2) classical two-phase regression estimator, and 3) mass imputation estimator. These are defined as follows:

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

2. Two-phase regression estimator:
$$\hat{\theta}_{tp,reg} = \hat{\bar{Y}}_2 + (\hat{\bar{\mathbf{X}}}_1 - \hat{\bar{\mathbf{X}}}_2)\hat{\boldsymbol{\beta}},$$
where
$$\begin{aligned} \hat{\bar{\mathbf{X}}}_1 &= \frac{\sum_{i \in A_1} w_{1i} \mathbf{x}_i}{\sum_{i \in A_1} w_{1i}}, \\ \left(\hat{\bar{\mathbf{X}}}_2, \hat{\bar{Y}}_2\right) &= \frac{\sum_{i \in A_2} w_{1i} w_{2i|1i} (\mathbf{x}_i, y_i)}{\sum_{i \in A_2} w_{1i} w_{2i|1i}}. \end{aligned}$$

3. Mass imputation estimator:
$$\hat{\theta}_{imp} = \frac{1}{N} \left(\sum_{i \in A_2} w_{1i} y_i + \sum_{i \in A_2^c} w_{1i} \hat{y}_i \right).$$

Further, we investigate 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 $L = n_1$. 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 three estimators and Monte Carlo mean and relative bias of the replication variance estimator are computed.

Table 4.2 Monte Carlo bias and variance of the three estimators: Direct estimator ($\hat{\theta}_{dir}$); Two-phase regression estimator ($\hat{\theta}_{tp,reg}$); Mass imputation estimator ($\hat{\theta}_{imp}$)

Population	Second-phase Sampling	Estimator	Bias	Variance
Linear	SRS	$\hat{\theta}_{dir}$	-0.001	0.029
		$\hat{\theta}_{tp,reg}$	-0.001	0.026
		$\hat{\theta}_{imp}$	-0.001	0.026
	Poisson	$\hat{\theta}_{dir}$	0.000	0.028
		$\hat{\theta}_{tp,reg}$	0.001	0.025
		$\hat{\theta}_{imp}$	0.000	0.018
	RSPPS	$\hat{\theta}_{dir}$	-0.001	0.027
		$\hat{\theta}_{tp,reg}$	0.000	0.025
		$\hat{\theta}_{imp}$	-0.002	0.018
Ratio	SRS	$\hat{\theta}_{dir}$	0.000	0.045
		$\hat{\theta}_{tp,reg}$	0.000	0.040
		$\hat{\theta}_{imp}$	0.000	0.040
	Poisson	$\hat{\theta}_{dir}$	0.002	0.045
		$\hat{\theta}_{tp,reg}$	0.002	0.040
		$\hat{\theta}_{imp}$	0.000	0.034
	RSPPS	$\hat{\theta}_{dir}$	0.001	0.044
		$\hat{\theta}_{tp,reg}$	0.002	0.039
		$\hat{\theta}_{imp}$	-0.001	0.033

Table 4.2 presents the Monte Carlo bias and variance of the three estimators and we can check that all three estimators are unbiased for the population mean regardless of sampling design and specified population model type. Moreover, variances of two-phase regression estimator and mass imputation estimator for the sample selected using simple random sampling for both phases are the same, which is 0.026, since the estimators are equivalent with each other under the linear regression model that is demonstrated in Lemma 2. Further, the mass imputation estimator has smaller variance compared with a variance of two-phase regression estimator, as the mass imputation

estimator uses more information for the estimation; auxiliary variable $\mathbf{x}_i$ in A_1 is used for the mass imputation estimator whereas only $\mathbf{x}_i$ in A_2 is used for the two-phase regression estimator.

Table 4.3 Monte Carlo mean and relative bias (R.B.) of the replication variance estimator of the mass imputation estimator

Population	Second-phase Sampling	Mean	R.B.
Linear	SRS	0.026	0.002
	Poisson	0.018	-0.003
	RSPPS	0.018	-0.004
Ratio	SRS	0.040	-0.007
	Poisson	0.034	0.006
	RSPPS	0.033	-0.004

Table 4.3 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 4.2, and it leads to small relative biases of the replication variance estimator in Table 4.3. This result supports the Theorem 1, as the bias term in (4.20) can be safely ignored since the first-phase sampling rate is $500/100,000 = 0.005$, which is small enough.

4.6.2 Non-nested Two-phase Sampling

In this section, we present a simulation study in order to check the performance of the proposed method under non-nested two-phase sampling.

We can consider two finite populations of size $N = 50,000$. One is from a linear regression model

$$y_i = 0.3 + 1.2x_i + e_i,$$

where $x_i \sim N(2,1)$ and $e_i \sim N(0,1)$. The other is from a logistic regression model

$$y_i \sim \text{Bernoulli}(p_i),$$

where $\text{logit}(p_i) = -0.8 + 0.7x_i$ and $x_i \sim N(2, 1)$.

From each of the two populations, we generate two independent samples. We use a simple random sample of size $n_A = 500$ for sample A. In selecting sample B of size $n_B = 500$, we consider two cases as follows:

Case 1. We create two strata where Stratum 1 consists of elements with $x_i \leq 2$ and Stratum 2 consists of elements with $x_i > 2$. Within each stratum, we select n_h elements by simple random sampling independently, where $n_1 = 300$ and $n_2 = 200$. We assume that the stratum information is unavailable at the time of data analysis.

Case 2. We select sample B with expected size of $n_B = 500$ using Poisson sampling where

$$\delta_i | I_i = 1 \sim \text{Bernoulli}(\pi_{2i|1i}),$$

with

$$\pi_{2i|1i} = [1 + \exp(-2.8 - 1.2x_i)]^{-1}.$$

From the two samples, we compute three estimators:

1) Sample mean estimator from sample A ($\hat{\theta}_A$)

2) Naive mean estimator from sample B

3) Mass imputation estimator given in (4.34) of sample A ($\hat{\theta}_I$)

Sample mean estimator of sample A works as a gold standard estimator. For Monte Carlo simulation, we repeat this procedure $B = 5,000$ times.

Table 4.4 presents the Monte Carlo mean, Monte Carlo variance, and root mean squared error of the three point estimators. Sample mean from sample A is unavailable in practice, but is computed for the comparison with other estimators as a gold standard. We can see that the mass imputation estimator is unbiased for the population mean, but naive mean estimator of sample B underestimates the population mean for all population model types and sampling designs for the sample B. Table 4.4 also shows that the bias for the naive mean estimator of sample B is increased

Table 4.4 Monte Carlo mean, Monte Carlo variance, and root mean squared error (RMSE) of three point estimators: Sample mean estimator from sample A(Mean A); Mean estimator from sample B(Naive B); Mass imputation estimator (M.I.)

Case	Population	Estimator	Mean	Var($\times 10^{-2}$)	RMSE
Case 1	Linear Regression	Mean A	2.700	0.480	0.069
		Naive B	2.498	0.298	0.208
		M.I.	2.700	0.491	0.070
	Logistic Regression	Mean A	0.632	0.046	0.021
		Naive B	0.607	0.044	0.033
		M.I.	0.632	0.048	0.022
Case 2	Linear Regression	Mean A	2.700	0.475	0.069
		Naive B	1.320	0.415	1.381
		M.I.	2.700	0.736	0.086
	Logistic Regression	Mean A	0.632	0.048	0.022
		Naive B	0.453	0.044	0.180
		M.I.	0.632	0.082	0.028

under case 2, which uses Poisson sampling for the sample B, but the mass imputation estimator still provides unbiased estimates regardless of population model types and sampling designs. Moreover, the variance of the mass imputation estimator for case 1 with linear regression population in Table 4.4 can be computed by

$$\begin{aligned} V(\hat{\theta}_I) &\cong \left(\frac{1}{n_A}-\frac{1}{N}\right)\beta_1^2\sigma_x^2+\left(\frac{1}{n_B}+\frac{1}{\sum_{i\in B}(x_i-\bar{x}_B)^2}\left\{(\mu_x-\bar{x}_B)^2+\frac{\sigma_x^2}{N}\right\}\right)\sigma_e^2 \\ &= \left(\frac{1}{500}-\frac{1}{50,000}\right)\times 1.2^2+\left(\frac{1}{500}+\frac{1}{488.313}\left\{0.0276+\frac{1}{50,000}\right\}\right) \\ &= 0.0049, \end{aligned}$$

where μ_x is a population mean of $\mathbf{x}$ and $\bar{x}_B$ is a sample mean of sample B. Similarly, the variance of the mass imputation estimator for case 2 with linear regression population in Table 4.4 can be computed by

$$\begin{aligned} V(\hat{\theta}_I) &\cong \left(\frac{1}{n_A}-\frac{1}{N}\right)\beta_1^2\sigma_x^2+\left(\frac{1}{n_B}+\frac{1}{\sum_{i\in B}(x_i-\bar{x}_B)^2}\left\{(\mu_x-\bar{x}_B)^2+\frac{\sigma_x^2}{N}\right\}\right)\sigma_e^2 \\ &= \left(\frac{1}{500}-\frac{1}{50,000}\right)\times 1.2^2+\left(\frac{1}{500}+\frac{1}{521.313}\left\{1.3345+\frac{1}{50,000}\right\}\right) \\ &= 0.0074. \end{aligned}$$

Since a variance of sample mean of sample A ($\hat{\theta}_A$) with linear regression population is

$$V(\hat{\theta}_A) = \frac{1}{n_A}\sigma_y^2 = \frac{1}{n_A}\left(\beta_1^2\sigma_x^2 + \sigma_e^2\right)$$

and $n_A = n_B = 500$ in our setup, differences between the Monte Carlo variance of the mass imputation estimator and the Monte Carlo variance of sample mean of sample A in Table 4.4 are due to the bias of $\mathbf{x}$ variable in sample B. That is,

$$\left(\frac{1}{\sum_{i\in B}(x_i - \bar{x}_B)^2}\left\{(\mu_X - \bar{x}_B)^2 + \frac{\sigma_x^2}{N}\right\}\right)\sigma_e^2$$

constitutes for the difference between two estimated variances under linear population model.

Table 4.5 Monte Carlo mean and relative bias (R.B.) of proposed variance estimator of mass imputation estimator

Case	Population	Mean ($\times 10^{-2}$)	R.B.
Case 1	Linear Regression	0.490	-0.002
	Logistic Regression	0.049	0.007
Case 2	Linear Regression	0.734	-0.003
	Logistic Regression	0.082	-0.007

Table 4.5 presents Monte Carlo mean and relative bias of the replication variance estimator of the mass imputation estimator. We can check that Monte Carlo means of mass imputation estimators are consistent for the Monte Carlo variances of the mass imputation estimator in Table 4.4, so as to produce corresponding relative biases, which are small, presented in Table 4.5.

4.7 Conclusion

We treat two-phase sampling as a missing data problem and propose the 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 replicates of conditional inclusion probability for the second phase sample, which may be complicated or impossible to compute depending on sampling designs, are not necessary for the replication variance estimation.

In addition, we consider a mass imputation when the second-phase sample is a non-probability sample, which is subject to selection bias. For the data integration, the proposed method is developed based on the parametric moment assumption. Instead of the parametric model approach, we can consider a non-parametric model approach such as Kernel regression or nearest neighbor imputation. Furthermore, the proposed method can be extended to data integration handling bigdata, which has a relatively large sample size with inherent selection bias.

Bibliography

Blair, J. (1999). A probability sample of gay urban males: The use of two-phase adaptive sampling. *Journal of Sex Research*, 36(1):39–44.

Blair, J. and Czaja, R. (1982). Locating a special population using random digit dialing. *Public Opinion Quarterly*, 46(4):585–590.

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

Breslow, N. E. (2014). Case-control studies. In *Handbook of epidemiology*, pages 293–323. Springer.

Breslow, N. E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):447–461.

De Waal, T. (2000). A brief overview of imputation methods applied at statistics netherlands. *Netherlands Official Statistics*, 15:23–27.

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

Fetter, M. (2001). Mass imputation of agricultural economic data missing by design: a simulation study of two regression based techniques. In *Federal Conference on Survey Methodology*.

Firth, D. and Bennett, K. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):3–21.

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

Fuller, W. A. and Kim, J. K. (2005). Hot deck imputation for the response model. *Survey Methodology*, 31(2):139.

Hidiroglou, M. (2001). Double sampling. *Survey methodology*, 27(2):143–154.

Hidiroglou, M. and Särndal, C. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24:11–20.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Kim, J. K., Navarro, A., and Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American statistical association*, 101(473):312–320.

Kim, J. K. and Rao, J. N. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99(1):85–100.

Kim, J. K. and Yu, C. L. (2011). Replication variance estimation under two-phase sampling. *Survey methodology*, 37(1):67.

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

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99(468):1131–1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):27–48.

Moore, R. and 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(201):101–116.

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, pages 462–474.

Rao, J. N. and Sitter, R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82(2):453–460.

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Thompson, M. E. and Wu, C. (2008). Simulation-based randomized systematic pps sampling under substitution of units. *Survey Methodology*, 34(1):3.

APPENDIX A. PROOF OF THEOREM 1

By Lemma 1, we have

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

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

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

and $(\hat{Y}_2^{(k)}, \hat{X}_2^{(k)})$ are computed from the second-phase replicate using $w_{1i}^{(k)}$. Using the 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 (4.13) and (4.19),

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

We now determine the order of $\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}$, which is written by

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(k)} &= \left(N^{-1}\sum_{i\in A_1} w_{1i}^{(k)} a_i \mathbf{x}_i^{'}\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i\in A_1} w_{1i}^{(k)} a_i \mathbf{x}_i^{'} y_i\right)\\ &:= \left[\hat{M}_x^{(k)}\right]^{-1}\left[\hat{M}_y^{(k)}\right].\end{aligned}$$

Let $\hat{M}_x$, $\hat{M}_y$, Δ_x and Δ_y be

$$\begin{aligned}(\hat{M}_x, \hat{M}_y) &= N^{-1}\sum_{i\in A_1} w_{1i} a_i \mathbf{x}_i^{'}(\mathbf{x}_i, y_i),\\ \Delta_x &= \frac{\hat{M}_x^{(k)} - \hat{M}_x}{\hat{M}_x},\\ \Delta_y &= \frac{\hat{M}_y^{(k)} - \hat{M}_y}{\hat{M}_y}.\end{aligned}$$

Since $\hat{M}_x^{(k)} = \hat{M}_x + O_p(n_2^{-1/2}L^{-1/2})$ and $\hat{M}_x = O_p(1)$, we have

$$\begin{aligned}\Delta_x &= O_p(1)O_p(n_2^{-1/2}L^{-1/2})\\ &= O_p(n_2^{-1/2}L^{-1/2})\end{aligned}$$

and similarly $\Delta_y = O_p(n_2^{-1/2}L^{-1/2})$. Therefore, we can determine the order of $\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}$ as

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(k)} &= \left[\hat{M}_x(1+\Delta_x)\right]^{-1}\left[\hat{M}_y(1+\Delta_y)\right]\\ &= \hat{M}_x^{-1}\hat{M}_y(1+\Delta_y)\left[1 - \Delta_x + (\Delta_x)^2 + O_p(n_2^{-3/2}L^{-3/2})\right]\\ &= \hat{M}_x^{-1}\hat{M}_y\left[1 + O_p(n_2^{-1/2}L^{-1/2})\right]\\ &= \hat{\boldsymbol{\beta}} + O_p(n_2^{-1/2}L^{-1/2}),\end{aligned}$$

since $\hat{\boldsymbol{\beta}} = \hat{M}_x^{-1}\hat{M}_y$.

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

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

by (4.19) and it follows from (A.2) that

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

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 (4.18) and (4.14), 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 generated 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

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

where $\eta_i = \mathbf{x}_i\boldsymbol{\beta}_N + \kappa_i^{-1}a_i e_i$ is defined as a pseudo value and 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 $\tilde{Y}_{imp} = \sum_{i\in A_1} w_{1i}\kappa_i^{-1}a_i e_i - \hat{X}_1\boldsymbol{\beta}_N$. By assumption (4.14) and (4.17), 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

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

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 $E\left[V(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N\right]$ 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 $a_i \sim \text{Bernoulli}(\pi_{2i|1i})$, 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 (4.15) and

(4.16), we have

$$
\begin{aligned}
& V(n_1 N^{-2} V(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N) \\
= \; & V\left[n_1 N^{-2} V\left(\sum_{i \in A_1} w_{1i} \kappa_i^{-1} a_i e_i - \hat{X}_1 \boldsymbol{\beta}_N | \mathbf{a}, \mathcal{F}_N \right) | \mathcal{F}_N \right] \\
= \; & V\left[\sum_{i=1}^{N} \sum_{j=1}^{N} \Omega_{ij} (w_{1i} \kappa_i^{-1} a_i e_i - x_i \boldsymbol{\beta}_N)(w_{1i} \kappa_i^{-1} a_i e_i - x_i \boldsymbol{\beta}_N) | \mathcal{F}_N \right] \\
= \; & \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \Omega_{ij} \Omega_{kl} Cov(\kappa_i^{-1} a_i e_i \kappa_j^{-1} a_j e_j, \kappa_k^{-1} a_k e_k \kappa_l^{-1} a_l e_l | \mathcal{F}_N) \\
= \; & 2 \sum_{i=1}^{N} \sum_{j=1}^{N} \Omega_{ij}^2 (\kappa_i^{-1} \kappa_j^{-1} - 1) e_i^2 e_j^2 \\
\leq \; & 2 \max_{i,j} \left\{ (\kappa_i^{-1} \kappa_j^{-1} - 1) e_i^2 e_j^2 \right\} \left(\max_{i,j} |\Omega_{ij}| \right) \sum_{i=1}^{N} \sum_{j=1}^{N} |\Omega_{ij}| \\
= \; & O(N^{-1}).
\end{aligned}
$$

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

Now, last term of (A.5) is

$$
\begin{aligned}
V\left[E(\tilde{Y}_{imp}|\mathbf{a}, \mathcal{F}_N)|\mathcal{F}_N \right] &= V\left[E\left(\sum_{i \in A_1} w_{1i} (\kappa_i^{-1} a_i e_i - \hat{X}_1 \boldsymbol{\beta}_N) | \mathbf{a}, \mathcal{F}_N \right) | \mathcal{F}_N \right] \\
&= V\left[\sum_{i=1}^{N} (\kappa_i^{-1} a_i e_i - \hat{X}_1 \boldsymbol{\beta}_N) | \mathcal{F}_N \right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} Cov(\kappa_i^{-1} a_i e_i, \kappa_j^{-1} a_j e_j) \\
&= \sum_{i=1}^{N} \kappa_i^{-1} (1 - \kappa_i) e_i^2.
\end{aligned}
$$

Therefore, by combining all the results, we have

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

and by (A.3) and (A.6), we have conclusion (4.20).

Bibliography

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

Kim, J. K., Navarro, A., and Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American statistical association*, 101(473):312–320.