1-1-2019

Combining Non-probability and Probability Survey Samples Through Mass Imputation

Jae Kwang Kim
Iowa State University, jkim@iastate.edu

Seho Park

Yilin Chen

Changbao Wu

Combining Non-probability and Probability Survey Samples Through Mass Imputation

Abstract
This paper presents theoretical results on combining non-probability and probability survey samples through mass imputation, an approach originally proposed by Rivers (2007) as sample matching without rigorous theoretical justification. Under suitable regularity conditions, we establish the consistency of the mass imputation estimator and derive its asymptotic variance formula. Variance estimators are developed using either linearization or bootstrap. Finite sample performances of the mass imputation estimator are investigated through simulation studies and an application to analyzing a non-probability sample collected by the Pew Research Centre.

Keywords
Data integration, bootstrap, missing at random, selection bias

Disciplines
Design of Experiments and Sample Surveys | Probability | Statistical Methodology

Comments
This pre-print is made available through arxiv: https://arxiv.org/abs/1812.10694v2.

Combining Non-probability and Probability Survey Samples Through Mass Imputation

Jae Kwang Kim Seho Park Yilin Chen Changbao Wu

January 1, 2019

Abstract. This paper presents theoretical results on combining non-probability and probability survey samples through mass imputation, an approach originally proposed by Rivers (2007) as sample matching without rigorous theoretical justification. Under suitable regularity conditions, we establish the consistency of the mass imputation estimator and derive its asymptotic variance formula. Variance estimators are developed using either linearization or bootstrap. Finite sample performances of the mass imputation estimator are investigated through simulation studies and an application to analyzing a non-probability sample collected by the Pew Research Centre.

Key words: Data integration, bootstrap, missing at random, selection bias.

1 Introduction

Probability sampling is a classical tool for obtaining a representative sample from a target population. Because the first-order inclusion probabilities are known, probability sampling can provide design unbiased estimators and also construct valid statistical inferences for finite population parameters. Although probability samples are known to achieve the representativeness of the target population but they are usually expensive and do not provide up-to-date information on variables of specific studies.

On the other hand, non-probability samples, such as web panels, are increasingly popular in spite of its potential danger of selection biases (Baker et al., 2013). New challenges and nonstandard data sources generate objectives that traditional sampling techniques cannot easily address. To utilize modern data sources in statistically defensible ways, it is important, and many times critical, to develop better statistical tools to combine information from two data sources, one from a probability sample and the other from a non-probability sample. Combining the up-to-date information from a non-probability sample and auxiliary information from a probability sample is an area of data integration, which is an emerging area of research in survey sampling (Lohr and Raghunathan, 2017).

For data integration, a popular approach is to use an independent probability sample as a benchmark for calibration weighting. Such calibration weighting method is based on the assumption that the selection mechanism for the non-probability sample is ignorable after adjusting for the auxiliary variables used for calibration weighting. Such assumption is essentially the missing at random (MAR) assumption of Rubin (1976). Use of calibration weighting for non-probability samples has been discussed in Dever and Valliant (2016) and Elliott et al. (2017), among others.

Instead of using calibration weighting, the mass imputation method can also be developed under the same MAR assumption. If there is no measurement on the study variable of interest in the probability sample, we can view the probability sample as a missing data with 100% missingness in the study variable and apply imputation techniques using the non-probability sample as the training data for developing an

imputation model. Mass imputation has been developed in the context of two-phase sampling (Breidt et al., 1996; Kim and Rao, 2012), but it is not fully investigated in the context of survey integration for combining the non-probability sample with a probability survey sample. One notable exception is the sample matching method of Rivers (2007), but he did not provide a theoretical justification for the proposed method.

In this paper, we aim to fill this important research gap in survey sampling and develop the mass imputation method for a probability sample using observations from a non-probability sample. Even though the observations in the non-probability sample are not necessarily representative of the target population, the relationships among variables in the non-probability sample may be used to develop a predictive model for mass imputation. Thus, the non-probability sample can be used as a training data for developing an imputation model. If the training data for the imputation model were a probability sample, then the theory of Kim and Rao (2012) could be directly applicable. Under some mild assumptions, we show that the method of Kim and Rao (2012) can be applied to non-probability samples for the training data. The main contribution of the current paper is to develop a valid statistical inference procedure through mass imputation which integrates probability and non-probability survey samples. Rigorous asymptotic theory for the mass imputation estimator is developed and a linearization variance estimator is proposed. Furthermore, a bootstrap variance estimator that does not require access to the training data for users is also developed.

The basic setting is described in Section 2. Main results on consistency and the asymptotic variance formula are presented in Section 3. A practically useful bootstrap variance estimator is proposed in Section 4. Extensions to more general parameters defined through estimating equations are given in Section 5. Results from simulation studies on the finite sample performances of the mass imputation estimator are reported in Section 6. The mass imputation technique is applied in Section 7 to analyze a non-probability survey sample collected by the Pew Research Centre using two different probability samples: the Behavioral Risk Factor Surveillance System survey data and the Volunteer Supplement survey data from the Current Population

Survey. Some additional remarks are given in Section 8, and proofs are relegated to the appendix.

2 Basic Setup

Suppose that we have two data sources, where the first sample A observes the vector of auxiliary variables ($\mathbf{X}$) only and the second 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. We further assume that sample B with observations on both $\mathbf{X}$ and Y is a non-probability sample and is subject to inherent selection bias.

Let A and B denote respectively the set of units included in the probability and non-probability samples. Let $n_A = |A|$ and $n_B = |B|$ be the sample sizes. Table 1 presents the general setup of the two sample structure for data integration.

Table 1: Data Structure

Sample	$\mathbf{X}$	Y	Representativeness
A	✓		Yes
B	✓	✓	No

Let δ_B be the indicator variable for the unit being included in the non-probability sample B. The ignorability assumption for the sample B is specified as

$$P(\delta_B = 1 \mid \mathbf{X}, Y) = P(\delta_B = 1 \mid \mathbf{X})\,. \tag{1}$$

We further assume that each unit in the population has a non-zero probability to be included in the sample B, i.e.,

$$P(\delta_B = 1 \mid \mathbf{X} = \mathbf{x}) > 0 \tag{2}$$

for all $\mathbf{x}$ in the support of $\mathbf{X}$. Under assumptions (1) and (2), the prediction model

$f(y \mid \mathbf{x})$ can be estimated by using observed $(Y, \mathbf{X})$ from sample B since

$$f(y \mid \mathbf{x}, \delta_B = 1) = f(y \mid \mathbf{x}) \,. \tag{3}$$

The prediction model $f(y \mid \mathbf{x})$ can then be used for creating mass imputation for the probability sample A. Condition (3) is sometimes called the transportability condition in the sense that the imputation model obtained from sample B is transportable to sample A. Assumption (2) implies that the sample support of $\mathbf{X}$ in sample B coincides with the support of $\mathbf{X}$ in the population. If assumption (2) is not satisfied, then we do not necessarily have (3) for all values of $\mathbf{X}$ in the support of $\mathbf{X}$.

Under assumptions (1)-(2), it is possible to consider a mass imputation estimator based on nearest neighbor imputation as suggested by Rivers (2007). Nearest neighbor imputation is a nonparametric method that does not require any parametric model assumptions. While nonparametric imputation methods can provide robust estimation, it suffers from curse of dimensionality, and the asymptotic bias of the nearest neighbor imputation is not negligible if the dimension of x is greater than one (Yang and Kim, 2018). In this paper, we consider a semi-parametric model for sample B with the first moment specified as

$$E(Y \mid \mathbf{X} = \mathbf{x}) = m(\mathbf{x}; \boldsymbol{\beta}) \tag{4}$$

for some unknown $p \times 1$ vector $\boldsymbol{\beta}$ and a known function $m(\cdot; \cdot)$. Let $(y_i, \mathbf{x}_i)$ be the observed values of $(Y, \mathbf{X})$ for unit i. We assume that $\hat{\boldsymbol{\beta}}$ is the unique solution to

$$\hat{U}(\boldsymbol{\beta}) = \frac{1}{n_B} \sum_{i \in B} \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta})\} \mathbf{h}(\mathbf{x}_i; \boldsymbol{\beta}) = 0 \tag{5}$$

for some p-dimensional vector of functions $\mathbf{h}(\mathbf{x}_i; \boldsymbol{\beta})$. The estimator $\hat{\boldsymbol{\beta}}$ is first obtained by using data from sample B and then used to obtain the predicted value $\hat{y}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ for all $i \in A$. The mass imputation estimator for the finite population mean $\theta_N =$

$N^{-1}\sum_{i=1}^{N} y_i$ is computed as

$$\hat{\theta}_I = \frac{1}{N}\sum_{i\in A} w_i \hat{y}_i \,, \tag{6}$$

where $w_i = \pi_i^{-1}$ is the sampling weight and $\pi_i = P(i \in A)$ is the inclusion probability of unit $i \in A$. Rigorous asymptotic properties of the mass imputation estimator (6) are presented in the next section.

3 Main Theoretical Results

We now discuss asymptotic properties of the mass imputation estimator given in (6). For the asymptotic framework, we assume a sequence of finite populations and and a sequence of samples A and B as discussed in Fuller (2009). Let $\boldsymbol{\beta}_0$ be the true value of the parameters $\boldsymbol{\beta}$ for model (4). The following theorem establishes the consistency and asymptotic variance formula of the mass imputation estimator under the joint randomization of the probability sampling design for sample A and the prediction model for sample B.

Theorem 1. *Suppose that $(\mathbf{x}, y)$ has bounded fourth moments over the sequence of finite populations and that assumptions (1)-(2) hold. Under the regularity conditions stated in Appendix A, the mass imputation estimator (6) satisfies $\hat{\theta}_I = \tilde{\theta}_I + o_p\big(n_B^{-1/2}\big)$, where*

$$\tilde{\theta}_I = N^{-1}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) + n_B^{-1}\sum_{i\in B}\left\{y_i - m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\}\mathbf{c}'\mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}_0) \tag{7}$$

with

$$\mathbf{c} = \left[n_B^{-1}\sum_{i\in B}\dot{\mathbf{m}}(\mathbf{x}_i;\boldsymbol{\beta}_0)\mathbf{h}'(\mathbf{x}_i;\boldsymbol{\beta}_0)\right]^{-1} N^{-1}\sum_{i=1}^{N}\dot{\mathbf{m}}(\mathbf{x}_i;\boldsymbol{\beta}_0) \tag{8}$$

and $\dot{\mathbf{m}}(\mathbf{x};\boldsymbol{\beta}) = \partial m(\mathbf{x};\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. The quantity $\tilde{\theta}_I$ satisfies $E(\tilde{\theta}_I - \theta_N) = 0$ and $V(\tilde{\theta}_I - \theta_N) = V_A + V_B$, where

$$V_A = V\left\{N^{-1}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\} \tag{9}$$

is the design-based variance under the probability sampling design for sample A and

$$V_B = E\left[n_B^{-2}\sum_{i\in B} E\Big(e_i^2 \mid \mathbf{x}_i\Big)\Big\{\mathbf{c}'\mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}_0)\Big\}^2\right] \tag{10}$$

is the variance component for sample B under the prediction model with $e_i = y_i - m(\mathbf{x}_i;\boldsymbol{\beta}_0)$.

Proof of Theorem 1 is presented in Appendix A. The asymptotic variance $V(\tilde{\theta}_I - \theta_N)$ consists of two parts. The first term V_A is of order $O(n_A^{-1})$ and the second term V_B is of order $O(n_B^{-1})$. If $n_A/n_B = o(1)$, i.e., the sample size n_B is much larger than n_A, the term V_B is of smaller order and the leading term of the total variance is V_A. Otherwise the two variance components both contribute to the total variance. In big data applications where n_B is very large, the term V_B can be safely ignored.

Example 1. *Under the linear regression model* $Y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i$ *with* $e_i \sim (0, \sigma_e^2)$, *independent among all i, we have* $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ *where* $\hat{\boldsymbol{\beta}} = \left(\sum_{i\in B}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i\in B}\mathbf{x}_i y_i$. *The mass imputation estimator of (6) under the regression model is given by* $\hat{\theta}_{I,reg} = \left(N^{-1}\sum_{i\in A} w_i\mathbf{x}_i\right)'\hat{\boldsymbol{\beta}}$. *If the probability sample A is selected by simple random sampling, the asymptotic variance of* $\hat{\theta}_{I,reg}$ *is given by*

$$V\left(\hat{\theta}_{I,reg} - \theta_N\right) \approx V\left(n_A^{-1}\sum_{i\in A}\mathbf{x}_i\boldsymbol{\beta}\right) + V\left(n_B^{-1}\sum_{i\in B} e_i\mathbf{x}_i'\mathbf{c}\right), \tag{11}$$

where $\mathbf{c} = \left(n_B^{-1}\sum_{i\in B}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\bar{x}_N$ *and* $\bar{x}_N = N^{-1}\sum_{i=1}^N \mathbf{x}_i$. *If* $\mathbf{x}_i = (1, x_i)'$ *and* $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, *then the asymptotic variance reduces to*

$$V\left(\hat{\theta}_{I,reg} - \theta_N\right) \approx \frac{1}{n_A}\beta_1^2\sigma_x^2 + \frac{1}{n_B}\sigma_e^2 + E\left[\frac{(\bar{x}_N - \bar{x}_B)^2}{\sum_{i\in B}(x_i - \bar{x}_B)^2}\right]\sigma_e^2,$$

where $\sigma_x^2 = V(X)$ *and* $\bar{x}_B = n_B^{-1}\sum_{i\in B} x_i$. *If sample B is a random sample from the population, then the third term is of order* $O(n_B^{-2})$ *and becomes negligible. However, since sample B is a non-probability sample, the third term might not be negligible.*

Variance estimation for the mass imputation estimator (6) requires the estimation

of the two components V_A and V_B. The first component can be estimated by

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

where $\pi_{ij} = P(i, j \in A)$ are the joint inclusion probabilities and are assumed to be positive. The second component can be estimated by

$$\hat{V}_B = \frac{1}{n_B^2} \sum_{i \in B} \hat{e}_i^2 \left\{ \hat{\mathbf{c}}' \mathbf{h}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) \right\}^2 , \tag{12}$$

where $\hat{e}_i = y_i - m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ and $\hat{\mathbf{c}} = \left(n_B^{-1} \sum_{i \in B} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} N^{-1} \sum_{i \in A} w_i \mathbf{x}_i$. The total variance of $\hat{\theta}_I$ can be estimated by $\hat{V}(\hat{\theta}_I - \theta_N) = \hat{V}_A + \hat{V}_B$.

4 Bootstrap Variance Estimation

The variance estimator presented in Section 3 is based on the linearization method. The closed-form formula for the asymptotic variance is simple to implement. However, to compute $\hat{V}_B$ in (12), we need to use individual observations of $(\mathbf{x}_i, y_i)$ in sample B, which is not necessarily available when only the sample A with mass imputed responses is released to the public data users. Note that the goal of mass imputation is to produce a representative sample A with synthetic observations on the response variable using sample B as a training dataset. Once the mass imputation is performed, the training data is no longer necessary in computing point estimators. It is therefore desirable to develop a variance estimation method that does not require access to observations in sample B.

To achieve this goal, we propose a bootstrap method for variance estimation that creates a replicated set of synthetic data $\{\hat{y}_i^{(k)}, i \in A\}$ corresponding to each set of bootstrap weights $\{w_i^{(k)}, i \in A\}$, $k = 1, \cdots, L$ associated with sample A only. The method enables users to correctly estimate the variance of the mass imputation estimator $\hat{\theta}_I$ without access to the training data $\{(y_i, \mathbf{x}_i) : i \in B\}$ from sample B. The data file will contain additional columns of $\{y_i^{(k)} : i \in A\}$ associated with the columns

of bootstrap weights $\{w_i^{(k)}; i \in A\}$, $k = 1, \cdots, L$, where L is the number of replicates created from sample A. Kim and Rao (2012) also considered a similar method in the context of survey integration from non-nested two-phase sampling.

In order to develop a valid bootstrap method for the mass imputation estimator $\hat{\theta}_I$ in (6), it is critical to develop a valid bootstrap method for estimating $V(\hat{\boldsymbol{\beta}})$ when $\hat{\boldsymbol{\beta}}$ is computed from (5). Note that, under assumptions (3) and (4), we can obtain

$$V(\hat{\boldsymbol{\beta}}) \doteq J^{-1}\Omega J^{-1'} \tag{13}$$

where $J = E\left\{n_B^{-1}\sum_{i\in B}\dot{\mathbf{m}}_i\mathbf{h}_i'\right\}$, $\Omega = E\left\{n_B^{-2}\sum_{i\in B}E(e_i^2 \mid \mathbf{x})\mathbf{h}_i\mathbf{h}_i'\right\}$ with $\dot{\mathbf{m}}_i = \dot{\mathbf{m}}(\mathbf{x}_i;\boldsymbol{\beta}_0)$ and $\mathbf{h}_i = \mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}_0)$. The reference distribution for (13) is the joint distribution of the superpopulation model (4) and the unknown sampling mechanism for the non-probability sample B. Interestingly, the variance formula in (13) equals exactly to the variance of $\hat{\boldsymbol{\beta}}$ when sample B is selected by simple random sampling (SRS). That is, even though the sampling design for sample B is not SRS, its effect on the variance of $\hat{\boldsymbol{\beta}}$ is essentially the same with SRS. This is due to the MAR assumption in (3) which makes the effect of the sampling design for estimating $\boldsymbol{\beta}$ ignorable even though it is still not ignorable for $\theta = E(Y)$. Therefore, we can safely ignore the sampling design for sample B when estimating $\boldsymbol{\beta}$ and develop a valid bootstrap method for variance estimation of $\hat{\boldsymbol{\beta}}$ using the bootstrap method for SRS.

Our proposed bootstrap method can be described as the following four steps:

Step 1. Create the kth set of replication weights $\{w_i^{(k)}, i \in A\}$ based on the sampling design for the probability sample A.

Step 2. Generate the kth bootstrap sample of size n_B from sample B using simple random sampling with replacement and compute $\hat{\boldsymbol{\beta}}^{(k)}$ using the same estimation equations (5) applied to the bootstrap sample.

Step 3. Use $\hat{\boldsymbol{\beta}}^{(k)}$ obtained from Step 2 to compute $\hat{y}_i^{(k)} = m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}^{(k)})$ for each $i \in A$ and create a new column $\{\hat{y}_i^{(k)}, i \in A\}$ alongside $\{w_i^{(k)}, i \in A\}$ for the sample A dataset.

Step 4. Repeat Steps 1-3, independently, for $k = 1, \cdots, L$ for a pre-chosen L.

Step 1 needs to follow standard practice in survey sampling on creating replication weights for design-based variance estimation. The final sample A dataset contains additional columns for the replicate versions of mass imputed values $\{\hat{y}_i^{(k)}, i \in A\}$ and the sets of replication weights $\{w_i^{(k)}, i \in A\}$, $k = 1, \cdots, L$. The replicate versions of the mass imputation estimator $\hat{\theta}_I$ are computed as

$$\hat{\theta}_I^{(k)} = \frac{1}{N} \sum_{i \in A} w_i^{(k)} \hat{y}_i^{(k)}, \qquad k = 1, \cdots, L, \tag{14}$$

and the resulting bootstrap variance estimator of $\hat{\theta}_I$ is computed as

$$\hat{V}_b(\hat{\theta}_I) = \frac{1}{L} \sum_{k=1}^{L} \left(\hat{\theta}_I^{(k)} - \hat{\theta}_I \right)^2. \tag{15}$$

The following theorem establishes the consistency of the proposed bootstrap variance estimator. Its proof is presented in Appendix B.

Theorem 2. *Suppose that the assumptions of Theorem 1 hold. Under the additional assumptions stated in Appendix B, the bootstrap variance estimator given by (15) satisfies*

$$\hat{V}_b\big(\hat{\theta}_I\big) = V\big(\hat{\theta}_I - \theta_N\big) + o_p\big(n_B^{-1}\big). \tag{16}$$

5 Extension to More General Parameters

The discussions in previous sections focus on the estimation of the finite population mean. We now consider an extension to more general parameters θ_N defined as the unique solution to the census estimating equation $U_N(\theta) = \sum_{i=1}^{N} g(\theta; \mathbf{x}_i, y_i) = 0$ for some estimating function $g(\theta; \mathbf{x}, y)$. If the dataset $\{(\mathbf{x}_i, y_i), i \in A\}$ is fully observed for the probability sample A, we can use

$$\hat{U}(\theta) = \sum_{i \in A} w_i g(\theta; \mathbf{x}_i, y_i) = 0 \tag{17}$$

to obtain an estimator of θ_N. Since the y_i's are not observed in sample A, we can adapt the technique of mass imputation to develop an estimator when training data from sample B are available.

We consider fractional imputation for the estimation of a general parameter θ_N. If $f(y \mid \mathbf{x})$ follows a parametric model $f(y \mid \mathbf{x}) = f(y \mid \mathbf{x}; \boldsymbol{\beta})$ for some $\boldsymbol{\beta}$, we can use parametric fractional imputation of Kim (2011) to develop a parametric fractional mass imputation (PFMI) estimator of θ_N. If Y is categorical with J categories with support $y \in \{z_1, \cdots, z_J\}$, we can create $M = J$ imputed values with $y_i^{*(j)} = z_j$ with fraction weight $w_{ij}^* = P(Y = z_j \mid \mathbf{x}_i; \hat{\boldsymbol{\beta}})$, $j = 1, \cdots, J$. For continuous Y, the proposed PFMI can be described as follows:

1. Use sample B as the training data to obtain $\hat{\boldsymbol{\beta}}$.

2. For each i, generate M imputed values $y_i^{*(1)}, \cdots, y_i^{*(M)}$ from $f(y_i \mid \mathbf{x}_i; \hat{\boldsymbol{\beta}})$. The jth imputed value is generated from $f(y \mid \mathbf{x}; \hat{\beta})$ as $y_i^{*(j)} = \hat{F}^{-1}(u_j \mid \mathbf{x}_i)$, where $\hat{F}(y \mid \mathbf{x}) = \int_{-\infty}^{y} f(t \mid \mathbf{x}; \hat{\boldsymbol{\beta}}) dt$ and $u_1, \cdots, u_M$ are M systematic samples from the $U(0, 1)$ distribution.

3. The fractional weight $w_{ij}^* = 1/M$ is assigned to $y_i^{*(j)}$, $j = 1, \cdots, M$.

The proposed PFMI method creates M imputed values for each y_i, $i \in A$. The mass imputation estimator of θ_N is then computed as the solution to

$$\sum_{i \in A} \sum_{j=1}^{M} w_i w_{ij}^* g(\theta; \mathbf{x}_i, y_i^{*(j)}) = 0\,. \tag{18}$$

Since $\sum_{j=1}^{M} w_{ij}^* g(\theta; \mathbf{x}_i, y_i^{*(j)}) \approx E\{g(\theta; \mathbf{x}_i, Y) \mid \mathbf{x}_i; \hat{\boldsymbol{\beta}})$, the mass imputation estimator obtained from solving (18) is approximately unbiased.

For variance estimation, we can use a bootstrap method similar to the procedures described in Section 4. More specifically, we first compute $\hat{\boldsymbol{\beta}}^{(k)}$ using the kth bootstrap sample from sample B and then compute $w_{ij}^{*(k)} = P(Y = z_j \mid \mathbf{x}_i; \hat{\beta}^{(k)})$ if Y is a

categorical variable or compute

$$w_{ij}^{*(k)} \propto \frac{f(y_i^{*(j)} \mid \mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(k)})}{f(y_i^{*(j)} \mid \mathbf{x}_i; \hat{\boldsymbol{\beta}})} \tag{19}$$

such that $\sum_{j=1}^{M} w_{ij}^{*(k)} = 1$ if Y is continuous. Computing the replication fractional weights using (19) is based on the idea of importance sampling and it has been discussed in Berg et al. (2016). The kth replicate of $\hat{\theta}$ can be obtained by solving

$$\sum_{i \in A} \sum_{j=1}^{M} w_i^{(k)} w_{ij}^{*(k)} g(\theta; \mathbf{x}_i, y_i^{*(j)}) = 0\,. \tag{20}$$

Note that the imputed values used in (20) are not changed for each replication. Only the survey weights and the fractional weights are replicated. Once the solution $\hat{\theta}_I^{(k)}$ to (20) is obtained, then the same bootstrap variance formula (15) can be used.

6 Simulation Studies

In this section, we perform two simulation studies to evaluate the finite sample performance of the proposed method for mass imputation using a non-probability sample. In the first simulation, we consider a continuous study variable and use regression imputation. In the second simulation, we consider a binary study variable and use parametric fractional imputation.

6.1 Simulation study one

The setup for simulation one employed a 3×3 factorial structure with two factors. The first factor is the superpopulation model that generates the finite population. The second factor is the sample size for sample B. We generated the following three models for finite populations of size $N = 100,000$.

1. Model I: The y_i's are independently generated from $N(0.3+2x_i, 1)$, where $x_i \overset{i.i.d}{\sim} N(2,1)$.

2. Model II: The y_i's are independently generated from $N(0.3 + x_i, 2^2)$, where $x_i \overset{i.i.d}{\sim} N(2, 1)$.

3. Model III: The y_i's are independently generated from $N(0.3 + 0.5x_i^2, 1)$, where $x_i \overset{i.i.d}{\sim} N(2, 1)$.

Model I generates a finite population with a high correlation between x and y ($r^2 = 0.8$), Model II generates a finite population with a low correlation ($r^2 = 0.2$), and Model III generates a finite population where the linear relationship fails. Model III is included to check the effect of model mis-specification in the imputation model.

From each of the three populations, we generated two independent samples. We use simple random sampling of size $n_A = 500$ to obtain sample A. In selecting sample B of size n_B, where $n_B \in \{300, 500, 1000\}$, 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, independent between the two strata, where $n_1 = 0.7n_B$ and $n_2 = 0.3n_B$. We assume that the stratum information is unavailable at the time of data analysis. Using the two samples A and B, we compute four estimators of $\theta_N = N^{-1}\sum_{i=1}^{N} y_i$:

1) The sample mean from sample A: $\hat{\theta}_A = n_A^{-1}\sum_{i\in A} y_i$.

2) The naive estimator (sample mean) from sample B: $\hat{\theta}_B = n_B^{-1}\sum_{i\in B} y_i$.

3) The mass imputation estimator from sample A given in (6) using $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where $(\hat{\beta}_0, \hat{\beta}_1)$ are the estimated regression coefficients obtained from sample B.

4) The inverse probability weighted (IPW) estimator proposed by Chen et al. (2018): $\hat{\theta}_{IPW} = N^{-1}\sum_{i\in B} \hat{\pi}_i^{-1} y_i$, where the propensity scores, $\pi_i = \pi(\mathbf{x}_i; \boldsymbol{\phi}) = \{1+\exp(-\phi_0 - \phi_1 x_i)\}^{-1}$ with $\mathbf{x}_i = (1, x_i)'$ and $\boldsymbol{\phi} = (\phi_0, \phi_1)'$, are estimated by using $\hat{\boldsymbol{\phi}}$ which solves the following score equations:

$$U(\boldsymbol{\phi}) = \sum_{i\in B} \mathbf{x}_i - \sum_{i\in A} w_i \pi(\mathbf{x}_i; \boldsymbol{\phi})\mathbf{x}_i = \mathbf{0}. \tag{21}$$

The sample mean of sample A serves as a gold standard estimator. Results are based on $2,000$ repeated simulation runs. Table 2 presents the Monte Carlo bias,

the Monte Carlo variance, and the relative mean squared error of the four point estimators. The relative mean squared error of estimator $\hat{\theta}$ is defined as

$$ReMSE = \frac{MSE(\hat{\theta})}{MSE(\hat{\theta}_A)}.$$

Note that the sample mean $\hat{\theta}_A$ from sample A is not available in practice but is computed here as the gold standard. Table 2 shows that, with models I and II where the linear regression model holds, the mass imputation estimator is unbiased for the population mean. The naive mean estimator of sample B underestimates the population mean for all scenarios considered in the simulation. When the size of sample B for training data is larger than the size of sample A ($n_B = 1,000$), it is possible that the mass imputation estimator has a smaller MSE than the gold standard. Under the simple regression model, the asymptotic variance of the mass imputation estimator is

$$V\left(\hat{\theta}_I - \theta_N\right) \approx \frac{1}{n_A}\sigma_x^2\beta_1^2 + \frac{1}{n_B}\sigma_e^2\left\{1 + \frac{(\bar{x}_N - \bar{x}_B)^2}{s_{x,B}^2}\right\},$$

where $s_{x,B}^2 = (n_B - 1)^{-1}\sum_{i\in B}(x_i - \bar{x}_B)^2$, while the variance of sample mean of sample A is $V(\hat{\theta}_A) = \sigma_y^2/n_A = (\beta_1^2\sigma_x^2 + \sigma_e^2)/n_A$. Thus, if n_B is much larger than n_A, the mass imputation estimator can be more efficient than the sample mean of A. The IPW estimator is less efficient than the mass imputation estimator for all cases considered in the current simulation setup.

The imputation model using the simple linear regression is incorrectly specified for model III. The mass imputation estimator is modestly biased. Nonetheless, the performance in terms of MSE is better than IPW estimator because the mass imputation estimator has much smaller variance than the IPW estimator, even though the absolute bias is larger, when the linear relationship fails.

Table 3 presents Monte Carlo mean and relative bias of the two variance estimators of the mass imputation estimator using linearization and bootstrap. Both variance estimators show negligible relative biases. In particular, the bootstrap vari-

ance estimator shows good performance even under model III.

6.2 Simulation study two

The second simulation study uses the same setup of the first simulation study except for different population models. We consider binary Y variables with the same X variable as in study one and use the following two models to generate a finite population of size $N = 100,000$ for each model.

1. Model I: The y_i's are independently generated from a Bernoulli distribution with $P(Y = 1 \mid x) = \{1 + \exp(-x)\}^{-1}$.

2. Model II: The y_i's are independently generated from a Bernoulli distribution with $P(Y = 1 \mid x) = \{1 + \exp(-0.5x^2)\}^{-1}$.

Model II is considered to check the effect of model mis-specification in the imputation model. We use the same sampling methods of the simulation study one to select two independent samples A and B and then compute the four estimators of θ_N which is the finite population proportion for $Y = 1$: The sample mean from sample A, the naive estimator from sample B, the mass imputation estimator using parametric fractional imputation, and the IPW estimator using the same method from the first simulation study. For fractional imputation, we impute two values 1 and 0 for each unit $i \in A$ along with the fractional weight:

$$(y_i^{*(j)}, w_{ij}^*) = \begin{cases} (1, \hat{p}_i) & \text{if } j = 1 \\ (0, 1 - \hat{p}_i) & \text{if } j = 2\,, \end{cases}$$

where $\text{logit}(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $(\hat{\beta}_0, \hat{\beta}_1)$ are obtained by the maximum likelihood method from observations in sample B and the logistic regression model.

Table 4 presents the Monte Carlo bias, the Monte Carlo variance, and the relative mean squared error of the four point estimators. The main conclusions from the simulation results are similar to study one: The mass imputation estimator is nearly unbiased even under the incorrectly specified model and is more efficient than the IPW

estimator. If the sample size for the training data is large ($n_B = 1,000$), the mass imputation estimator can be more efficient than the gold standard. The bootstrap variance estimator, not presented here to save space, shows negligible biases for all cases.

7 An Application

In this section, we illustrate the proposed mass imputation method through the analysis of a non-probability survey sample collected by the Pew Research Centre in 2015. The dataset is referred to as PRC. The PRC dataset contains a total of 9301 cases and 56 variables and is provided by eight different vendors with unknown sampling and data collection strategies. We treat the PRC dataset as a non-probability survey sample with the sample size $n_B = 9301$. The PRC dataset aims to study the relation between people and community. We choose 9 variables, among them 8 are binary and 1 is continuous, as response variables in our analysis. We consider 24 variables listed in Table 5 as possible covariates for building prediction models for the response variables.

An important aspect of the mass imputation method is the availability of a probability survey sample with information on covariates. We consider two such samples. The first is the Behavioral Risk Factor Surveillance System (BRFSS) survey data and the second is the Volunteer Supplement survey data from the Current Population Survey (CPS), both collected in 2015. The two datasets have no measurements on the response variables but share a rich set of common covariates with the PRC dataset as shown in Table 5. The BRFSS dataset contains a very large number 441456 cases. The CPS dataset contains 80075 cases with measurements on volunteering tendency, which is highly relevant to the response variables considered in the PRC dataset. Both the BRFSS and the CPS datasets contain a separate column of the survey weights.

We first examine marginal distributions of the covariates from three datasets. Table 5 contains the estimated population mean using each of the three datasets, where $\hat{\mu}_{PRC}$ is the simple sample mean using the PRC dataset while $\hat{\mu}_{BRFSS}$ and

$\hat{\mu}_{CPS}$ are respectively the survey weighed estimates obtained from the BRFSS and the CPS datasets. There are noticeable differences between the naive estimates from the PRC sample and the estimates from the two probability samples for covariates such as Origin (Hispanic/Latino), Education (High school or less), Household (with children), Health (Smoking) and Volunteer works. It is strong evidence that the PRC dataset is not a representative sample for the population.

Estimates of the population mean for each of the 9 response variables using the proposed mass imputation method are presented in Tables 6 and 7. The second column indicates whether BRFSS or CPS is used as the probability sample. The results presented in Table 6 use a common set of covariates which are available in all three datasets, and the results reported in Table 7 are obtained by using two different sets of covariates, one between PRC and BRFSS and the other between PRC and CPS, depending on the availability, as shown in Table 5.

Computation of the mass imputation estimator $\hat{\theta}_I$ requires a prediction model. We use a logistic regression model for each of the 8 binary responses and a linear regression model for the continuous response. The naive sample mean estimator $\hat{\theta}_B$ is listed for comparisons. The two variance estimators v_l and v_b for $\hat{\theta}_I$ are also computed. The linearization variance estimator v_l is based on the formula $V\left(\tilde{\theta}_I - \theta_N\right) = V_A + V_B$ given by Theorem 1, where V_A is the designed based variance component under the probability sampling design for sample A. Unfortunately, detailed design information other than the survey weights is not available for either BRFSS or CPS. We use an approximate variance formula for V_A by assuming that the survey design is single-stage PPS sampling with replacement, a strategy often used by survey data analyst for the purpose of variance estimation. The bootstrap variance estimator v_b is computed based on the procedure described in Section 4 using $L = 5000$ bootstrap samples.

There are three major observations from the results presented in Table 6 where a single common set of covariates is used: (i) there are substantial discrepancies between the mass imputation estimator and the naive estimator in most cases; (ii) the mass imputation estimates obtained with two different probability samples are comparable for all cases; and (iii) the two variance estimators obtained by using the linearization

and the bootstrap methods generally agree with each other.

Results in Table 7 are obtained by including additional covariates in the prediction model for each of the two probability samples, as shown in the bottom part of Table 5. We observe that the discrepancies between the mass imputation estimator $\hat{\theta}_I$ and the naive estimator $\hat{\theta}_B$ become more pronounced. More importantly, there are also noticeable differences between the two mass imputation estimators, especially for the response variables "Participated in school groups", "Participated in service organizations" and "Participated in sports organizations". This is likely attributed to the strong association between the response variables and the additional covariate "Volunteer works" available in the CPS sample but not in BRFSS. Similarly, the covariates "Smoke everyday", "Smoke never" and "No money to see doctors", which are available in BRFSS but not in CPS, likely explain the opposite change of estimates found in the two response variables "Days had at least one drink last month" and "No money to buy food".

8 Additional Remarks

The use of non-probability survey samples as an efficient and cost-effective data source has become increasingly popular in recent years. Theoretical developments on analysis of non-probability samples, however, severely lag behind the need of making valid inference from such datasets. Non-probability survey samples are biased and do not represent the target population. Valid inferences require supplementary information on the population. The mass imputation approach relies on the availability of a probability survey sample from the same target population with information on covariates. The covariates are also measured for the non-probability sample and need to possess two crucial features for the framework discussed in this paper: (a) they characterize the inclusion/exclusion mechanism for units in the non-probability sample; and (b) they are relevant to the response variable in terms of prediction power.

Statistics Canada has been implementing the modernization initiatives in recent years, which call for a culture switch from the traditional survey-centric approach by

the agency to using data from multiple sources. One of the questions arising from the discussions is the role of traditional probability-based surveys, and there are even questions on the necessity of their existence in the future. Our theoretical results presented in this paper call for probability survey samples with rich information on auxiliary variables. A few large scale high quality probability surveys representing the target population can play significant roles in analyzing data from non-probability survey samples.

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1:90–143.

Berg, E., Kim, J.-K., and Skinner, C. (2016). Imputation under informative sampling. *Journal of Survey Statistics and Methodology*, 4(4):436–462.

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.

Chen, Y., Li, P., and Wu, C. (2018). Doubly robust inference with non-probability survey samples. *arXiv preprint arXiv:1805.06432*.

Dever, J. A. and Valliant, R. (2016). General regression estimation adjusted for undercoverage and estimated control totals. *Journal of Survey Statistics and Methodology*, 4:289–318.

Elliott, M. R., Valliant, R., et al. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2):249–264.

Fuller, W. A. (2009). *Sampling Statistic*. Wiley, Hoboken, NJ.

Keiding, N. and Louis, T. A. (2016). Perils and potentials of self-selected entry to

epidemiological studies and surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):319–376.

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

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

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

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *Annals of Statistics*, 10:462–474.

Rivers, D. (2007). Sampling for web surveys. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*.

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

Yang, S. and Kim, J. K. (2018). Predictive mean matching imputation in survey sampling. *arXiv preprint arXiv:1703.10256*.

Appendix

A. Proof of Theorem 1

We assume the following regularity conditions:

(1) The solution $\hat{\boldsymbol{\beta}}$ to (5) satisfies

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p(n_B^{-1/2}). \tag{A.1}$$

(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 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 $\boldsymbol{\beta}_0$.

To prove $\hat{\theta}_I = \tilde{\theta}_I + o_p\big(n_B^{-1/2}\big)$, we consider the class of estimators

$$\tilde{\theta}_I(\boldsymbol{\beta},\mathbf{c}) = N^{-1}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}) + \mathbf{c}'\hat{U}(\boldsymbol{\beta}),$$

indexed by p-dimensional vectors $\boldsymbol{\beta}$ and $\mathbf{c}$. Since $\hat{\boldsymbol{\beta}}$ satisfies $\hat{U}(\boldsymbol{\beta}) = 0$, the mass imputation estimator (6) can be expressed by $\hat{\theta}_I = \tilde{\theta}_I(\hat{\boldsymbol{\beta}},\mathbf{c})$ for any p-dimensional vector $\mathbf{c}$. Now we wish to find a particular choice of $\mathbf{c}$, say $\mathbf{c}^*$, that satisfies

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

Using the theory of Randles (1982), a sufficient condition for (A.2) is

$$E\left[\frac{\partial\tilde{\theta}_I(\boldsymbol{\beta},\mathbf{c}^*)}{\partial\boldsymbol{\beta}}\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = 0. \tag{A.3}$$

Since

$$E\left[\partial\tilde{\theta}_I(\boldsymbol{\beta},\mathbf{c})/\partial\boldsymbol{\beta}\right] \;=\; N^{-1}\sum_{i=1}^{N}\dot{\mathbf{m}}(\mathbf{x}_i;\boldsymbol{\beta}) - n_B^{-1}\sum_{i\in B}\dot{\mathbf{m}}(\mathbf{x}_i;\boldsymbol{\beta})\mathbf{c}'\mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}) = 0,$$

we can show that $\mathbf{c}$ specified in (8) satisfies (A.3) and hence is the choice for $\mathbf{c}^*$. This completes the proof of the first part of the theorem.

Let $\mathbf{c}$ be given by (8). To derive the asymptotic variance formula, we write

$$\begin{aligned}
&\tilde{\theta}_I(\boldsymbol{\beta}_0,\mathbf{c}) - \theta_N \\
=&N^{-1}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) + n_B^{-1}\sum_{i\in B}\left\{y_i - m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\}\mathbf{c}'\mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}_0) - N^{-1}\sum_{i=1}^{N} y_i \\
=&N^{-1}\left[\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - \sum_{i=1}^{N} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right] + \left[n_B^{-1}\sum_{i\in B} e_i\mathbf{c}'\mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}_0) - N^{-1}\sum_{i=1}^{N} e_i\right],
\end{aligned} \tag{A.4}$$

where $e_i = y_i - m(\mathbf{x}_i; \boldsymbol{\beta}_0)$. By treating sample B as fixed, we have

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

The second part of the theorem on the asymptotic variance of $\hat{\theta}_I$ follows from the fact that the two terms in (A.4) are uncorrelated.

B. Proof of Theorem 2

In addition to the assumptions of Theorem 1, we further assume that the bootstrap variance estimator of $\hat{\theta} = N^{-1}\sum_{i\in A} w_i y_i$ is design consistent under the sampling design for sample A, i.e., the estimator $\hat{V}_b(\hat{\theta}) = L^{-1}\sum_{k=1}^{L}\left(\hat{\theta}^{(k)} - \hat{\theta}\right)^2$ satisfies

$$\frac{\hat{V}_b(\hat{\theta})}{V(\hat{\theta})} \longrightarrow 1 \tag{B.1}$$

in probability, as $n \to \infty$ and $L \to \infty$, where $\hat{\theta}^{(k)} = N^{-1}\sum_{i\in A} w_i^{(k)} y_i$ and $\{w_i^{(k)}, i = 1, \cdots, n_A\}$ is the kth set of bootstrap replication weights.

Now, to show (15), we first define

$$\hat{U}^{(k)}(\boldsymbol{\beta}) = n_B^{-1}\sum_{i=1}^{n_B}\{y_i^{(k)} - m(\mathbf{x}_i^{(k)}; \boldsymbol{\beta})\}\mathbf{h}(\mathbf{x}_i^{(k)}; \boldsymbol{\beta})$$

where $\{(\mathbf{x}_i^{(k)}, y_i^{(k)}), i = 1, \cdots, n_B\}$ is the kth bootstrap sample for sample B selected by simple random sampling with replacement. Note that $\hat{\boldsymbol{\beta}}^{(k)}$ is the solution to $\hat{U}^{(k)}(\boldsymbol{\beta}) = 0$. Since

$$L^{-1}\sum_{k=1}^{L}\left\{\hat{U}^{(k)}(\boldsymbol{\beta}) - \hat{U}(\boldsymbol{\beta})\right\}^2 = O_p(n_B^{-1})\,,$$

we must have $\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}} = o_p(1)$. Using the Taylor linearization method, we have

$$\begin{aligned}
\hat{\theta}_I^{(k)} &= N^{-1}\sum_{i\in A} w_i^{(k)} m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}^{(k)}) \\
&= N^{-1}\sum_{i\in A} w_i^{(k)} m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) + \left\{N^{-1}\sum_{i\in A} w_i \dot{\mathbf{m}}(\mathbf{x}_i;\hat{\boldsymbol{\beta}})\right\}\left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right) \\
&\quad + N^{-1}\left\{\sum_{i\in A} w_i^{(k)} \dot{\mathbf{m}}(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) - \sum_{i\in A} w_i \dot{\mathbf{m}}(\mathbf{x}_i;\hat{\boldsymbol{\beta}})\right\}\left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right) \\
&= N^{-1}\sum_{i\in A} w_i^{(k)} m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) + \left\{N^{-1}\sum_{i\in A} w_i \dot{\mathbf{m}}(\mathbf{x}_i;\hat{\boldsymbol{\beta}})\right\}\left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right) + o_p(n_A^{-1}) \qquad \text{(B.2)}
\end{aligned}$$

We also have

$$\begin{aligned}
0 &= \hat{U}^{(k)}(\hat{\boldsymbol{\beta}}^{(k)}) \\
&= \hat{U}^{(k)}(\hat{\boldsymbol{\beta}}) + \left\{\frac{\partial}{\partial \boldsymbol{\beta}'}\hat{U}(\hat{\boldsymbol{\beta}})\right\}\left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}\right) + o_p(n_B^{-1}). \qquad \text{(B.3)}
\end{aligned}$$

Combining (B.2) with (B.3) and ignoring the smaller order terms, we obtain

$$\begin{aligned}
\hat{\theta}_I^{(k)} &= N^{-1}\sum_{A} w_i^{(k)} m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) + n_B^{-1}\sum_{B}\{y_i^{(k)} - m(\mathbf{x}_i^{(k)};\hat{\boldsymbol{\beta}})\}\hat{\mathbf{c}}'\mathbf{h}(\mathbf{x}_i^{(k)};\hat{\boldsymbol{\beta}}) \\
&:= \hat{P}_A^{(k)} + \hat{Q}_B^{(k)}, \qquad \text{(B.4)}
\end{aligned}$$

where

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

is a consistent estimator of $\mathbf{c}$ in (8). Noting that we can rewrite $\hat{\theta}_I = N^{-1}\sum_A w_i\hat{y}_i$ as

$$\begin{aligned}
\hat{\theta}_I &= N^{-1}\sum_{A} w_i m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) + n_B^{-1}\sum_{B}\{y_i - m(\mathbf{x}_i;\hat{\boldsymbol{\beta}})\}\hat{\mathbf{c}}'\mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}_0) \\
&:= \hat{P}_A + \hat{Q}_B, \qquad \text{(B.5)}
\end{aligned}$$

we can re-express $\hat{V}_b = L^{-1}\sum_{k=1}^{L}\left(\hat{\theta}_I^{(k)} - \hat{\theta}_I\right)^2$ as

$$\begin{aligned} & L^{-1}\sum_{k=1}^{L}\left(\hat{P}_A^{(k)} - \hat{P}_A\right)^2 + L^{-1}\sum_{k=1}^{L}\left(\hat{Q}_B^{(k)} - \hat{Q}_B\right)^2 + 2L^{-1}\sum_{k=1}^{L}\left(\hat{P}_A^{(k)} - \hat{P}_A\right)\left(\hat{Q}_B^{(k)} - \hat{Q}_B\right) \\ := \quad & \hat{V}_b(\hat{P}_A) + \hat{V}_b(\hat{P}_B) + 2\hat{C}_b(\hat{P}_A, \hat{Q}_B). \end{aligned}$$

Note that, by assumption (B.1), we have

$$\hat{V}_b(\hat{P}_A) = V(\hat{P}_A) + o_p(n_A^{-1}).$$

By the construction of the bootstrap sample for sample B, we have

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

which is equal to $\hat{V}_B$ in (12). The last term $\hat{C}_b(\hat{P}_A, \hat{Q}_B)$ estimates zero because the two bootstrap samples are selected independently.

Table 2: Monte Carlo mean, Monte Carlo variance, and relative mean square error (ReMSE) of the four point estimators in Simulation One, based on 2,000 Monte Carlo samples

n_B	Estimator	Model I			Model II			Model III		
		Bias	Var ($\times 10^3$)	ReMSE ($\times 10^2$)	Bias	Var ($\times 10^3$)	ReMSE ($\times 10^2$)	Bias	Var ($\times 10^3$)	ReMSE ($\times 10^2$)
300	$\hat{\theta}_A$	0.00	9.69	100	0.00	10.25	100	0.00	10.76	100
	$\hat{\theta}_B$	-0.64	8.27	4,311	-0.32	15.12	1,131	-0.64	7.89	3,883
	$\hat{\theta}_I$	0.00	11.56	119	0.00	1.76	172	-0.06	12.80	153
	$\hat{\theta}_{IPW}$	-0.03	16.81	185	-0.01	18.91	186	-0.04	27.04	263
1,000	$\hat{\theta}_A$	0.00	9.69	100	0.00	10.25	100	0.00	10.76	100
	$\hat{\theta}_B$	-0.64	2.26	4,252	-0.32	4.13	1,035	-0.64	2.25	3,837
	$\hat{\theta}_I$	0.00	8.92	92	0.00	6.44	63	-0.06	8.44	112
	$\hat{\theta}_{IPW}$	-0.03	12.21	137	-0.02	7.25	73	-0.03	16.79	166

Table 3: Monte Carlo means and relative biases (R.B.) of two variance estimation methods: Linearization and Bootstrap

Case	Model	Linearization		Bootstrap	
		Mean	R.B.	Mean	R.B.
Case 1 ($n_B = 300$)	I	0.0117	0.012	0.0117	0.012
	II	0.0168	-0.047	0.0169	-0.040
	III	0.0121	-0.054	0.0131	0.027
Case 2 ($n_B = 1,000$)	I	0.0091	0.020	0.0091	0.019
	II	0.0064	0.001	0.0065	0.005
	III	0.00830	-0.017	0.0086	0.020

Table 4: Monte Carlo mean, Monte Carlo variance, and relative mean square error (ReMSE) of the four point estimators in the second simulation study, based on 2,000 Monte Carlo samples

n_B	Estimator	Model I			Model II		
		Bias	Var ($\times 10^4$)	ReMSE ($\times 10^2$)	Bias	Var ($\times 10^4$)	ReMSE ($\times 10^2$)
300	$\hat{\theta}_A$	0.00	4.15	100	0.00	4.88	100
	$\hat{\theta}_B$	-0.06	6.90	1,050	-0.11	4.84	2,474
	$\hat{\theta}_I$	0.00	6.44	155	0.00	7.54	157
	$\hat{\theta}_{IPW}$	-0.01	6.78	171	-0.02	8.85	261
1,000	$\hat{\theta}_A$	0.00	4.15	100	0.00	4.88	100
	$\hat{\theta}_B$	-0.06	2.19	920	-0.11	1.48	2,412
	$\hat{\theta}_I$	0.00	2.47	59	0.00	3.47	74
	$\hat{\theta}_{IPW}$	0.00	2.55	67	-0.02	4.05	162

Table 5: Estimated Population Mean of Covariates from the Three Samples

		$\hat{\mu}_{PRC}$	$\hat{\mu}_{BRFSS}$	$\hat{\mu}_{CPS}$
Age category	<30	0.183	0.209	0.212
	>=30,<50	0.326	0.333	0.336
	>=50,<70	0.387	0.327	0.326
	>=70	0.104	0.131	0.126
Gender	Female	0.544	0.513	0.518
Race	White only	0.823	0.750	0.786
Race	Black only	0.088	0.126	0.125
Origin	Hispanic/Latino	0.093	0.165	0.156
Region	Northeast	0.200	0.177	0.180
Region	South	0.275	0.383	0.373
Region	West	0.299	0.232	0.235
Marital status	Married	0.503	0.508	0.528
Employment	Working	0.521	0.566	0.589
Employment	Retired	0.243	0.179	0.143
Education	High school or less	0.216	0.427	0.407
Education	Bachelor's degree and above	0.416	0.263	0.309
Education	Bachelor's degree	0.221	NA	0.198
Education	Postgraduate	0.195	NA	0.111
Household	Presence of child in household	0.289	0.368	NA
Household	Home ownership	0.654	0.672	NA
Health	Smoke everyday	0.157	0.115	NA
Health	Smoke never	0.798	0.833	NA
Financial status	No money to see doctors	0.207	0.133	NA
Financial status	Having medical insurance	0.891	0.878	NA
Financial status	Household income < 20K	0.161	NA	0.153
Financial status	Household income >100K	0.199	NA	0.233
Volunteer works	Volunteered	0.510	NA	0.248

Table 6: Estimated Population Mean Using A Single Set of Common Covariates

Binary Response y		$\hat{\theta}_B$	$\hat{\theta}_{I(v_l \backslash v_b)}$	$\hat{\theta}_{IPW(v_p)}$
Talked with	BRFSS	0.461	$0.457_{(4.323 \backslash 4.187)}$	$0.447_{(4.160)}$
neighbours frequently	CPS		$0.458_{(4.195 \backslash 4.055)}$	$0.451_{(4.282)}$
Tended to trust	BRFSS	0.590	$0.553_{(4.200 \backslash 4.221)}$	$0.546_{(4.115)}$
neighbours	CPS		$0.557_{(4.070 \backslash 4.044)}$	$0.551_{(4.207)}$
Expressed opinions	BRFSS	0.265	$0.240_{(2.858 \backslash 2.881)}$	$0.238_{(2.817)}$
at a government level	CPS		$0.243_{(2.878 \backslash 2.925)}$	$0.242_{(2.911)}$
Voted local	BRFSS	0.750	$0.707_{(3.687 \backslash 3.498)}$	$0.699_{(3.730)}$
elections	CPS		$0.716_{(3.447 \backslash 3.258)}$	$0.709_{(3.775)}$
Participated in	BRFSS	0.210	$0.200_{(2.599 \backslash 2.615)}$	$0.198_{(2.526)}$
school groups	CPS		$0.206_{(2.602 \backslash 2.607)}$	$0.206_{(2.660)}$
Participated in	BRFSS	0.141	$0.133_{(1.910 \backslash 1.886)}$	$0.130_{(1.762)}$
service organizations	CPS		$0.135_{(1.922 \backslash 1.930)}$	$0.134_{(1.867)}$
Participated in	BRFSS	0.168	$0.165_{(2.278 \backslash 2.221)}$	$0.160_{(2.102)}$
sports organizations	CPS		$0.170_{(2.262 \backslash 2.257)}$	$0.166_{(2.199)}$
No money	BRFSS	0.251	$0.289_{(3.681 \backslash 3.562)}$	$0.281_{(3.599)}$
to buy food	CPS		$0.286_{(3.516 \backslash 3.457)}$	$0.285_{(3.708)}$
Continuous Response y		$\hat{\theta}_B$	$\hat{\theta}_{I(v_l \backslash v_b)}$	$\hat{\theta}_{IPW(v_p)}$
Days had at least	BRFSS	5.301	$4.931_{(1.010 \backslash 0.996)}$	$4.857_{(0.9603)}$
one drink last month	CPS		$4.986_{(0.978 \backslash 0.952)}$	$4.921_{(0.9436)}$

Estimated variance for binary variables have been multiplied by 10^5, and estimated variance for continuous variable have been multiplied by 10^2.

Table 7: Estimated Population Mean Using Separate Sets of Common Covariates

Binary Response y		$\hat{\theta}_B$	$\hat{\theta}_{I(v_l \backslash v_b)}$	$\hat{\theta}_{IPW(v_p)}$
Talked with	BRFSS	0.461	$0.446_{(4.687 \backslash 4.608)}$	$0.435_{(4.830)}$
neighbours frequently	CPS		$0.404_{(4.623 \backslash 4.539)}$	$0.395_{(4.753)}$
Tended to trust	BRFSS	0.590	$0.561_{(4.567 \backslash 4.480)}$	$0.552_{(5.001)}$
neighbours	CPS		$0.530_{(4.824 \backslash 4.814)}$	$0.528_{(5.350)}$
Expressed opinions	BRFSS	0.265	$0.223_{(2.828 \backslash 2.783)}$	$0.220_{(2.955)}$
at a government level	CPS		$0.199_{(2.548 \backslash 2.540)}$	$0.198_{(2.715)}$
Voted local	BRFSS	0.750	$0.715_{(3.896 \backslash 3.597)}$	$0.707_{(4.592)}$
elections	CPS		$0.681_{(4.431 \backslash 4.222)}$	$0.680_{(5.208)}$
Participated in	BRFSS	0.210	$0.198_{(2.789 \backslash 2.830)}$	$0.194_{(2.874)}$
school groups	CPS		$0.133_{(1.428 \backslash 1.337)}$	$0.135_{(1.416)}$
Participated in	BRFSS	0.141	$0.121_{(1.842 \backslash 1.864)}$	$0.118_{(1.793)}$
service organizations	CPS		$0.087_{(0.977 \backslash 0.955)}$	$0.088_{(0.981)}$
Participated in	BRFSS	0.168	$0.158_{(2.395 \backslash 2.419)}$	$0.155_{(2.368)}$
sports organizations	CPS		$0.116_{(1.514 \backslash 1.500)}$	$0.115_{(1.363)}$
No money	BRFSS	0.251	$0.239_{(2.974 \backslash 2.995)}$	$0.233_{(3.410)}$
to buy food	CPS		$0.259_{(3.773 \backslash 3.700)}$	$0.244_{(3.613)}$

Continuous Response y		$\hat{\theta}_B$	$\hat{\theta}_{I(v_l \backslash v_b)}$	$\hat{\theta}_{IPW(v_p)}$
Days had at least	BRFSS	5.301	$4.812_{(1.028 \backslash 0.984)}$	$4.705_{(0.988)}$
one drink last month	CPS		$5.059_{(1.241 \backslash 1.217)}$	$4.965_{(1.290)}$

Estimated variance for binary variables have been multiplied by 10^5, and estimated variance for continuous variable have been multiplied by 10^2.