7-8-2018

Integration of survey data and big observational data for finite population inference using mass imputation

Shu Yang
North Carolina State University

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

Integration of survey data and big observational data for finite population inference using mass imputation

Abstract
Multiple data sources are becoming increasingly available for statistical analyses in the era of big data. As an important example in finite-population inference, we consider an imputation approach to combining a probability sample with big observational data. Unlike the usual imputation for missing data analysis, we create imputed values for the whole elements in the probability sample. Such mass imputation is attractive in the context of survey data integration (Kim and Rao, 2012). We extend mass imputation as a tool for data integration of survey data and big non-survey data. The mass imputation methods and their statistical properties are presented. The matching estimator of Rivers (2007) is also covered as a special case. Variance estimation with mass-imputed data is discussed. The simulation results demonstrate the proposed estimators outperform existing competitors in terms of robustness and efficiency.

Keywords
Calibration weighting, Data fusion, Generalized additive model, Nearest neighbor imputation, Post stratification, Statistical matching

Disciplines
Statistical Methodology | Statistical Models

Comments
This pre-print is made available through arxiv: https://arxiv.org/abs/1807.02817.

Integration of survey data and big observational data for finite population inference using mass imputation

Shu Yang* and Jae Kwang Kim†

Abstract

Multiple data sources are becoming increasingly available for statistical analyses in the era of big data. As an important example in finite-population inference, we consider an imputation approach to combining a probability sample with big observational data. Unlike the usual imputation for missing data analysis, we create imputed values for the whole elements in the probability sample. Such mass imputation is attractive in the context of survey data integration (Kim and Rao; 2012). We extend mass imputation as a tool for data integration of survey data and big non-survey data. The mass imputation methods and their statistical properties are presented. The matching estimator of Rivers (2007) is also covered as a special case. Variance estimation with mass-imputed data is discussed. The simulation results demonstrate the proposed estimators outperform existing competitors in terms of robustness and efficiency.

Keywords: Calibration weighting; Data fusion; Generalized additive model; Nearest neighbor imputation; Post stratification; Statistical matching.

1 Introduction

In finite population inference, probability sampling is the gold standard for obtaining a representative sample from the target population. Because the selection probability is known, the subsequent inference from a probability sample is often design-based and respect the way in which the data were collected. However, large-scale survey programs continually face heightened demands coupled with reduced resources. Demands include requests for estimates for domains with small sample sizes and desires for more timely estimates. Simultaneously, program budget cuts force reductions in sample sizes, and decreasing response rates make nonresponse bias an important concern. Baker et al. (2013) and Keiding and Louis (2016) address the current challenges in using probability samples for finite population inferences.

*Department of Statistics, North Carolina State University, North Carolina 27695, U.S.A. Email: syang24@ncsu.edu

†Department of Statistics, Iowa State University, Iowa 50011, U.S.A.

To meet the new challenges, statistical offices face the increasing pressure to utilize convenient but often uncontrolled big data sources, such as web survey panels and satellite information. While such data sources provide timely data for a large number of variables and population elements, they often fail to represent the target population of interest because of inherent selection biases.

To address new objectives and utilize modern data sources in statistically defensible ways, it is important to develop statistical tools for data integration for combining a probability sample with big observational data. To achieve this goal, one can apply mass imputation, where the imputed values are created for the whole elements in the probability sample. In the usual imputation for missing data analysis, the respondents in the sample provide a training dataset for developing an imputation model. In the proposed mass imputation, an independent big data sample is used as a training dataset, and mass imputation is applied to the probability sample. While the mass imputation idea for incorporating information from big data is very natural, the literature on mass imputation itself is very sparse. Breidt et al. (1996) discuss mass imputation for two-phase sampling. Kim and Rao (2012) develop a rigorous theory for mass imputation using two independent probability samples. Chipperfield et al. (2012) discuss composite estimation when one of the surveys is mass imputed. Rivers (2007) proposes a mass imputation approach using nearest neighbor imputation but the theory is not fully developed. Recently, Kim and Wang (2018), a technique report available by request from the authors, develop a theory for mass imputation for big data using a parametric model approach. However, the parametric model assumptions do not necessarily hold in practice. In order for mass imputation to be more useful and practical, the assumptions should be as weak as possible.

In this paper, we first develop a formal framework for mass imputation incorporating information from big data into a probability sample and present rigorous asymptotic results. Unlike Kim and Wang (2018), we do not make strong parametric model assumptions for mass imputation. Thus, the proposed method is appealing to survey practitioners. Our framework covers the nearest neighbor imputation estimator of Rivers (2007). In § 4, we investigate two strategies for improving the nearest neighbor imputation estimator, one using k nearest neighbor imputation and the other using generalized additive models. Secondly, using a novel calibration weighting idea, we propose an efficient mass imputation estimator and develop its asymptotic results. The efficiency gain is justified under a purely design-based framework and no model assumptions are used. The proposed methods are evaluated through extensive simulation studies based on artificial data and real-life data from U.S. Census Bureau's Monthly Retail Trade Survey.

2 Basic Setup

2.1 Notation: two data sources

Let $\mathcal{F}_N = \{(X_i, Y_i) : i \in U\}$ with $U = \{1, \ldots, N\}$ denote a finite population, where $X_i = (X_i^1, \ldots, X_i^p)$ is a p-dimensional vector of covariates, and Y_i is the study variable. We assume that $\mathcal{F}_N$ is a random sample from a superpopulation model ζ, and N is known. Our objective is

Table 1: Two data sources. "$\surd$" and "?" indicate observed and unobserved data, respectively.

		Sample weight $d = \pi^{-1}$	Covariate X	Study Variable Y
Probability Sample	1	$\surd$	$\surd$	?
$\mathcal{O}_A$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n	$\surd$	$\surd$	?
Big Data Sample	1	?	$\surd$	$\surd$
$\mathcal{O}_B$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	N_B	?	$\surd$	$\surd$

Sample A is a probability sample, and Sample B is a big data but may have selection biases.

to estimate the general finite population parameter $\mu_g = N^{-1}\sum_{i=1}^{N} g(Y_i)$ for some known $g(\cdot)$. For example, if $g(Y) = Y$, $\mu_g = N^{-1}\sum_{i=1}^{N} Y_i$ is the population mean of Y. If $g(Y) = I(Y < c)$ for some constant c, $\mu_g = N^{-1}\sum_{i=1}^{N} I(Y_i < c)$ is the population proportion of Y less than c.

Suppose that there are two data sources, one from a probability sample, referred to as Sample A, and the other from a big data source, referred to as Sample B. Table 1 illustrates the observed data structure. Sample A contains observations $\mathcal{O}_A = \{(d_i = \pi_i^{-1}, X_i) : i \in A\}$ with sample size $n = |A|$, where $\pi_i = P(i \in A)$ is known throughout Sample A, and Sample B contains observations $\mathcal{O}_B = \{(X_i, Y_i) : i \in B\}$ with sample size $N_B = |B|$. Although the big data source has a large sample size, the sampling mechanism is often unknown, and we cannot compute the first-order inclusion probability for Horvitz-Thompson estimation. The naive estimators without adjusting for the sampling process are subject to selection biases. On the other hand, although the probability sample with sampling weights represents the finite population, it does not observe the study variable.

2.2 Assumptions

Let $f(Y \mid X)$ be the conditional distribution of Y given X in the superpopulation model ζ. We define δ_B to be the indicator of selection to Sample B. We first make the following assumption.

Assumption 1 (Ignorability) *Conditional on X, the distribution of Y in Sample B follows the superpopulation model; i.e., $f(Y \mid X; \delta_B = 1) = f(Y \mid X)$.*

Assumption 1 states the ignorability of the selection mechanism to Sample B conditional upon the covariates. This assumption is also a missingness at random assumption (Rubin; 1976).

Now, let $f(X)$ and $f(X \mid \delta_B = 1)$ be the density function of X in the finite population and Sample B, respectively. We also require the following assumption.

Assumption 2 (Common support) *The vector of covariates X has a compact and convex support, with its density bounded and bounded away from zero. There exist constants C_l and C_u such that $C_l \leq f(X)/f(X \mid \delta_B = 1) \leq C_u$ almost surely.*

Assumption 2 implies that the support of X in Sample B is the same as that in the finite population. This assumption can also be formulated as a positivity assumption that $P(\delta_B = 1 \mid X) > 0$ for all X. This is necessary, because if the probability of selection into Sample B given some X is zero, then Sample B cannot provide adequate Y information for the units in this region without extrapolation.

3 Methodology

3.1 Nearest neighbor imputation

For estimation, if Y_i were observed throughout Sample A, the Horvitz–Thompson estimator $\hat{\mu}_{g,\mathrm{HT}} = N^{-1}\sum_{i\in A}\pi_i^{-1}g(Y_i)$ can be used. Our primary focus will be on the imputation estimator of μ_g, given by $\hat{\mu}_{g,I} = N^{-1}\sum_{i\in A}\pi_i^{-1}g(Y_i^*)$, where Y_i^* is an imputed value for Y_i. Creating imputed values for the whole data is called mass imputation (Chipperfield et al.; 2012; Kim and Rao; 2012).

To find suitable imputed values, we consider nearest neighbor imputation; that is, find the closest matching unit from Sample B based on the X values and use the corresponding Y value from this unit as the imputed value. This approach has been called statistical matching by Rivers (2007). To investigate the theoretical properties, we first consider matching with replacement with single imputation; the discussion on k nearest neighbor imputation is presented in § 4.

The nearest neighbor approach to mass imputation can be described in the following steps:

Step 1. For each unit $i \in A$, find the nearest neighbor from Sample B with the minimum distance between X_j and X_i. Let $i(1)$ be the index of its nearest neighbor, which satisfies $d(X_{i(1)}, X_i) \leq d(X_j, X_i)$, for $j \in B$, where $d(X_i, X_j)$ is a distance function between X_i and X_j. Without loss of generality, we use the Euclidean distance, $d(X_i, X_j) = ||X_i - X_j||$, where $||X|| = (X^{\mathrm{T}}X)^{1/2}$, to determine neighbors; our theoretical development applies to other distances (Abadie and Imbens; 2006).

Step 2. The nearest neighbor imputation estimator of μ_g is

$$\hat{\mu}_{g,\mathrm{nni}} = \frac{1}{N}\sum_{i\in A}\pi_i^{-1}g(Y_{i(1)}). \tag{1}$$

The matching estimator is attractive in practice because it does not require parametric model assumptions. Secondly, it does not require Sample A and Sample B to have common units, but requires only Assumption 2. Assumption 2 ensures that for any X_i in Sample A, we can find a value $X_{i(1)}$ in Sample B that is arbitrarily close to X_i as $N_B \to \infty$. Then, by Assumption 1, $g(Y_{i(1)})$ has the same distribution of $g(Y_i)$, given X_i. Moreover, for the same imputed dataset, one can estimate different parameters by choosing reasonable $g(\cdot)$. The main weakness of nearest neighbor imputation is that it is subject to the curse of dimensionality when X is a vector, but such weakness is not applicable when the size of the matching donor pool is huge as in our big data setup.

3.2 Asymptotic results

To study the asymptotic properties of $\hat{\mu}_{g,\text{nni}}$, we impose the following regularity conditions on the functional continuity and finite moments (e.g., Mack; 1981) and the sampling design for Sample A (Fuller; 2009, Ch. 1).

Assumption 3 *(i) $f(X)$ and $\mu_g(X) = E\{g(Y) \mid X\}$ are continuously differentiable for any continuous and bounded $g(Y)$, and (ii) $E\{g(Y)^\beta \mid X\}$ is bounded for $\beta = 0, 1, 2$.*

Assumption 4 *(i) There exist positive constants C_1 and C_2 such that $C_1 \leq Nn^{-1}\pi_i \leq C_2$, for $i = 1, \ldots, N$; (ii) the sampling fraction for Sample A is negligible, $nN^{-1} = o(1)$; and (iii) the sequence of the Horvitz-Thompson estimators $\hat{\mu}_{g,\text{HT}}$ satisfies $\text{var}_p(\hat{\mu}_{g,\text{HT}}) = O(n^{-1})$ and $\{\text{var}_p(\hat{\mu}_{g,\text{HT}})\}^{-1/2}(\hat{\mu}_{g,\text{HT}} - \mu_g) \mid \mathcal{F}_N \to \mathcal{N}(0,1)$ in distribution, as $n \to \infty$, where $\text{var}_p(\cdot)$ is the variance under the sampling design for Sample A.*

We derive the asymptotic theory for $\hat{\mu}_{g,\text{nni}}$ in the following theorem and defer its proof to the Supplementary Material.

Theorem 1 *Under Assumptions 1–3 and $NN_B^{-1} = O(1)$, $\hat{\mu}_{g,\text{nni}}$ has the same distribution as $\hat{\mu}_{g,\text{HT}}$ as $N_B \to \infty$. Furthermore, under Assumption 4, $\hat{\mu}_{g,\text{nni}}$ is consistent for μ_g, and*

$$n^{1/2}(\hat{\mu}_{g,\text{nni}} - \mu_g) \to \mathcal{N}(0, V_{\text{nni}}), \tag{2}$$

where

$$V_{\text{nni}} = \lim_{n\to\infty} \frac{n}{N^2} E\left[\text{var}_p\left\{\sum_{i\in A} \pi_i^{-1} g(Y_i)\right\}\right].$$

Theorem 1 implies that the standard point estimator can be applied to the imputed data $\{(X_i, Y_{i(1)}) : i \in A\}$ as if the $Y_{i(1)}$'s were observed values. Let π_{ij} be the joint inclusion probability for units i and j. We show in the Supplementary Material that the direct variable estimator based on the imputed data

$$\hat{V}_{\text{nni}} = \frac{n}{N^2} \sum_{i\in A} \sum_{j\in A} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} \frac{g(Y_{i(1)})}{\pi_i} \frac{g(Y_{j(1)})}{\pi_j}.$$

is consistent for V_{nni}.

4 Other techniques for mass imputation

4.1 k-nearest neighbor imputation

Instead of using a single imputed value, we now consider fractional imputation with k imputed values for each missing outcome. Fractional imputation is designed to reduce the variance of the final estimator due to imputation (Kalton and Kish; 1984; Kim and Fuller; 2004).

Assume no matching ties, let $\mathcal{J}_k(i)$ be the set of k nearest neighbors for unit i

$$\mathcal{J}_k(i) = \left\{ l \in B : \sum_{j \in B} 1_{\{d(X_j, X_i) \leq d(X_l, X_i)\}} \leq k \right\} = \{i(1), \ldots, i(k)\}.$$

The k nearest neighbor approach to mass imputation can be described in the following steps:

Step 1. For each unit $i \in A$, find the k nearest neighbors from Sample B, $\mathcal{J}_k(i)$. Impute the Y value for unit i by $\hat{\mu}_g(X_i) = k^{-1} \sum_{j=1}^{k} g(Y_{i(j)})$.

Step 2. The k nearest neighbor imputation estimator of μ_g is

$$\hat{\mu}_{g,\text{knn}} = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} \hat{\mu}_g(X_i).$$

In the nonparametric estimation literature, researchers have investigated the asymptotic properties of the k nearest neighbor imputation estimators extensively. See, e.g., Mack and Rosenblatt (1979) and Mack (1981) for early references. Cheng (1994) established root-n consistency of the k nearest neighbor imputation estimator of the outcome mean when the outcome is subject to missingness. We derive the asymptotic theory for $\hat{\mu}_{g,\text{knn}}$ in the context of mass imputation combining a probability sample and a big data sample in the following theorem and defer its proof to the Supplementary Material.

Theorem 2 *Under Assumptions 1–4, $n\,(k/N)^{4/p} \to 0$, $k/n \to 0$, and $k^2/n \to \infty$,*

$$n^{1/2}(\hat{\mu}_{g,\text{knn}} - \mu_g) \to \mathcal{N}(0, V_{\text{knn}}), \tag{3}$$

where

$$V_{\text{knn}} = \lim_{n \to \infty} \frac{n}{N^2} \left(E\left[\text{var}_p \left\{ \sum_{i \in A} \pi_i^{-1} \mu_g(X_i) \right\} \right] + E\left\{ \frac{1 - \pi_B(X)}{\pi_B(X)} \sigma_g^2(X) \right\} \right),$$

and $\pi_B(X) = P(\delta_B = 1 \mid X)$.

If $\pi_B(X)$ goes to 1, V_{knn} reduces to $\lim_{n \to \infty} \left(n/N^2\right) E\left[\text{var}_p \left\{\sum_{i \in A} \pi_i^{-1} \mu_g(X_i)\right\}\right]$. In this case, V_{knn} is smaller than V_{nni}, suggesting that $\hat{\mu}_{g,\text{knn}}$ gains efficiency over $\hat{\mu}_{g,\text{nni}}$. In finite samples, Beretta and Santaniello (2016) conduct a simulation study to compare nearest neighbor imputation and k nearest neighbor imputation in the setting with independent and identically distributed data. They found that k nearest neighbor imputation with a small k outperforms nearest neighbor imputation in terms of mean squared error.

4.2 Generalized additive models

Nearest neighbor imputation methods are nonparametric. On the other hand, parametric models especially linear models are sensitive to model misspecification. We now consider semiparametric methods for mass imputation. Among semiparametric methods, generalized additive models

(Hastie and Tibshirani; 1990) are flexible regarding model specification of the dependence of Y on X by specifying the model only through smooth functions rather than assuming a parametric relationship. We apply generalized additive models to leverage the predictive power of the big data sample to produce a predictive model for Y given X, so as to facilitate mass imputation for the probability sample.

We assume that $g(Y_i)$ given X_i follows some exponential family distribution, and

$$h^{-1}\{\mu_g(X_i)\} = f_1(X_i^1) + f_2(X_i^2) + \cdots f_p(X_i^p), \tag{4}$$

where $h(\cdot)$ is an inverse link function, and each $f_k(\cdot)$ is a smooth function of X^k, for $k = 1, \ldots, p$. Model (4) allows for rather flexible specification of the dependence of Y on X. The estimated function $f_k(X^k)$ can reveal possible nonlinearities of the relationship of Y and X^k.

There are several challenges in fitting model (4). First, $f_k(x)$ is an infinite-dimensional parameter, estimation of which often relies on some approximation. Second, we need to decide how smooth the $f_k(x)$ should be to balance the trade-off between model complexity and overfitting to the data at hand.

To solve the first issue, a common way to approximate $f_k(x)$ using splines. Let $B_m(x)$ be the basis spline functions for $m = 1, \ldots, M$ (Ruppert et al.; 2009). We approximate $f_k(x)$ by $f_k(x) = \sum_{m=1}^{M} \gamma_m^k B_m(x)$ with spline coefficients γ_m^k. This leads to an approximation of model (4):

$$h^{-1}[\hat{E}\{g(Y_i) \mid X_i\}] = \sum_{k=1}^{p} \sum_{m=1}^{M} \gamma_m^k B_m(X_i^k). \tag{5}$$

In (5), a large M allows for increased model complexity and also an increased chance of overfitting; while a small M may result in an inadequate model. This trade-off is balanced by choosing a relative large M and then penalizing the model complexity in the estimation stage (Eilers and Marx; 1996). Let the vector of spline coefficients be $\gamma_k^{\mathrm{T}} = (\gamma_1^k, \ldots, \gamma_m^k)$ and $\gamma^{\mathrm{T}} = (\gamma_1^{\mathrm{T}}, \ldots, \gamma_p^{\mathrm{T}})$. The estimate $\hat{\gamma}$ is obtained by maximizing the penalized likelihood:

$$-2l(\gamma) + \sum_{k=1}^{p} \lambda_k \gamma_k^{\mathrm{T}} S_k \gamma_k \tag{6}$$

where $l(\gamma)$ is the log likelihood function of γ, S_k is a matrix with the (m, l)th component $\int B_m''(x) B_l''(x) \mathrm{d}x$, $\gamma_k^{\mathrm{T}} S_k \gamma_k$ regularizes f_k to be smooth for which the degree of smoothness is controlled by λ_k. Given the smoothing parameter $\lambda^{\mathrm{T}} = (\lambda_1, \ldots, \lambda_p)$, the penalized likelihood function in (6) is optimized by a penalized version of the iteratively reweighted least squares algorithm (Nelder and Baker; 1972; McCullagh; 1984) to obtain $\hat{\gamma}$. Regarding the choice of λ, we note that λ controls the trade-off between model complexity and overfitting, which can be estimated separately from other model coefficients using generalized cross-validation or estimated simultaneously using restricted maximum likelihood estimation (Wood; 2006). In practice, the model performance is not sensitive to the choice of the number of basis functions, but rather estimation of the smoothing parameter is critical to

control the model complexity.

Once fitting the model, we can create an imputed value for each element i in Sample A as

$$\hat{\mu}_{g,\mathrm{GAM}}(X_i) = h\{\hat{f}_1(X_i^1) + \hat{f}_2(X_i^2) + \cdots \hat{f}_p(X_i^p)\},$$

where $\hat{f}_k(x) = \sum_{m=1}^{M} \hat{\gamma}_m^k B_m(x)$ for $k = 1, \ldots, p$. The mass imputation estimator based on the generalized additive model is

$$\hat{\mu}_{g,\mathrm{GAM}} = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} \hat{\mu}_{g,\mathrm{GAM}}(X_i).$$

Because in our context, the sample size of Sample B is much larger than that of Sample A, the estimation error in the imputation model can be negligible compared to the sampling variability of $\hat{\mu}_{g,\mathrm{GAM}}$.

To close this subsection, it is worth commenting on the assumption of additive effects of X in model (4). This assumption may be fairly strong one. To relax the additivity assumption, we can extend model (4) to include interactions through using the tensor product basis. For example, we can include a bivariate interaction surface $f_{12}(X^1, X^2) = \sum_{m=1}^{M} \sum_{l=1}^{L} \gamma_{ml} B_m(X^1) B_l(X^2)$.

5 Regression calibration

In practice, especially for government agencies, one nearest neighbor may be preferred because of its simplicity in implementation and data storage. We now consider another strategy to improve the efficiency for $\hat{\mu}_{g,\mathrm{nni}}$ when additionally the membership to Sample B can be determined throughout Sample A with the indicator δ_B. We can obtain δ_B by matching or directly asking about the membership to Sample B. The key insight is that the subsample of units in Sample A with $\delta_B = 1$ constitutes a second-phase sample from Sample B, where Sample B acts as a new population.

Let $h(\delta_B, X, Y)$ be a multi-dimensional function of δ_B, $\delta_B X$ and $\delta_B Y$, e.g., $h(\delta_B, X, Y) = (\delta_B, 1 - \delta_B, \delta_B X, \delta_B Y)^{\mathrm{T}}$. For simplicity of notation, we use h_i to denote $h(\delta_{Bi}, X_i, Y_i)$. We can calculate the population quantity $H = N^{-1} \sum_{i=1}^{N} h_i$ from Sample B. This insight enables the typical calibration weighting in survey sampling with known marginal totals. In Sample A, we treat the imputed values as observed values, and the design weighted estimator of H is $\hat{H}_A = N^{-1} \sum_{i \in A} \pi_i^{-1} h_i$. In general, $\hat{H}_A$ is not equal to H. We can use the known information H to improve the efficiency of $\hat{\mu}_{g,\mathrm{nni}}$.

This suggests the following calibration strategy. We modify the original design weights $\{d_i : i \in A\}$ in $\hat{\mu}_{g,\mathrm{nni}}$ to a new set of weights $\{\omega_i : i \in A\}$ by minimizing a distance function

$$\sum_{i \in A} G(\omega_i, d_i) = \sum_{i \in A} d_i \left(\frac{\omega_i}{d_i} - 1 \right)^2, \tag{7}$$

subject to the calibration constraints $N^{-1} \sum_{i \in A} \omega_i h_i = H$. The resulting weights $\{\omega_i : i \in A\}$ can

be called generalized regression weights.

The proposed estimator utilizing the new set of weights is

$$\hat{\mu}_{g,\mathrm{RC}} = \frac{1}{N}\sum_{i\in A}\omega_i g(Y_{i(1)}), \tag{8}$$

which is asymptotically equivalent to a generalized regression estimator (Park and Fuller; 2012).

We derive the asymptotic theory for $\hat{\mu}_{g,\mathrm{RC}}$ in the following theorem and defer its proof to the Supplementary Material.

Theorem 3 *Under Assumptions 1–4,*

$$n^{1/2}(\hat{\mu}_{g,\mathrm{RC}} - \mu_g) \to \mathcal{N}(0, V_{\mathrm{RC}}), \tag{9}$$

in distribution, as $n \to \infty$, where

$$V_{\mathrm{RC}} = \lim_{n\to\infty}\frac{n}{N^2}E\left(\mathrm{var}_p\left[\sum_{i\in A}\pi_i^{-1}\left\{g(Y_i) - h_i^{\mathrm{T}}\beta_N\right\}\right]\right),$$

and $\beta_N = \left(\sum_{i=1}^N h_i h_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^N h_i g(Y_i)$.

The calibrated estimator $\hat{\mu}_{g,\mathrm{RC}}$ improves the efficiency of $\hat{\mu}_{g,\mathrm{nni}}$ in the sense that V_{RC} is at most as large as V_{nni} given in Theorem 1. Moreover, $\hat{\mu}_{g,\mathrm{RC}}$ is robust in the sense that we do not require any modeling assumption.

Remark 1 (Choice of distance functions) *Different distance functions in (7) can be considered. If we choose $G(\omega_i, d_i) = -d_i\log(\omega_i/d_i)$, it leads to empirical likelihood estimation (Newey and Smith; 2004). If we choose the Kullback–Leibler distance function $G(\omega_i, d_i) = \omega_i\log(d_i/\omega_i)$, it leads to exponential tilting estimation (Kitamura and Stutzer; 1997; Imbens et al.; 1998; Schennach; 2007). Under mild conditions, these procedures provide a set of weights that is asymptotically equivalent to the set of regression weights (Deville and Särndal; 1992; Breidt and Opsomer; 2017).*

For variance estimation, by Theorem (3), we construct a consistent variance estimator for $\hat{\mu}_{g,\mathrm{RC}}$ as $\widehat{V}_{\mathrm{RC}}/n$, where

$$\widehat{V}_{\mathrm{RC}} = \frac{n}{N^2}\sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{\hat{e}_i}{\pi_i}\frac{\hat{e}_j}{\pi_j},$$

with $\hat{e}_i = g(Y_{i(1)}) - h_i^{\mathrm{T}}\hat{\beta}$, and

$$\hat{\beta} = \left(\sum_{i=1}^N h_i h_i^{\mathrm{T}}\right)^{-1}\begin{pmatrix}\sum_{i=1}^N \delta_{Bi}g(Y_i)\\ \sum_{i\in A}\pi_i^{-1}(1-\delta_{Bi})g(Y_{i(1)})\\ \sum_{i=1}^N \delta_{Bi}X_i g(Y_i)\\ \sum_{i=1}^N \delta_{Bi}Y_i g(Y_i)\end{pmatrix}.$$

6 Empirical experiments

In this section, we evaluate the finite sample performance of the proposed estimator using simulation studies, one based on artificial data and the other based on a synthetic population file from a single month sample of the U.S. Census Bureau's Monthly Retail Trade Survey.

6.1 A simulation study

We generate the data according to the following mechanism. We first generate a finite population $\mathcal{F}_N = \{X_i = (X_{1i}, X_{2i}), Y_i = (Y_{1i}, Y_{2i}) : i = 1, \dots N\}$ with size $N = 1,000,000$, where Y_{1i} is a continuous outcome and Y_{2i} is a binary outcome. From the finite population, we select a big data sample B where the inclusion indicator $\delta_{Bi} \sim$Ber(p_i) with p_i the inclusion probability for unit i, and we obtain a representative sample A of size $n = 1,000$ using simple random sampling. The parameters of interest are the population mean $N^{-1}\sum_{i=1}^{N} Y_i$ and the conditional population mean of Y_1 given $Y_2 = 1$.

For generating the finite population, we consider linear models

$$\begin{aligned} Y_{1i} &= 1 + X_{1i} + X_{2i} + \alpha_i + \epsilon_i, \\ P(Y_{2i} = 1 \mid X_{1i}, X_{2i}; \alpha_i) &= \text{logit}(1 + X_{1i} + X_{2i} + \alpha_i), \end{aligned} \tag{10}$$

and nonlinear models

$$\begin{aligned} Y_i &= 0.5(X_{1i} - 1.5)^2 + X_{2i}^2 + \alpha_i + \epsilon_i, \\ P(Y_{2i} = 1 \mid X_{1i}, X_{2i}; \alpha_i) &= \text{logit}\left\{0.5(X_{1i} - 1.5)^2 + X_{2i}^2 + \alpha_i\right\}, \end{aligned} \tag{11}$$

where $X_{1i} \sim \mathcal{N}(1, 1)$, $X_{2i} \sim$Exp(1), $\alpha_i \sim \mathcal{N}(0, 1)$, $\epsilon_i \sim \mathcal{N}(0, 1)$, and X_{1i}, X_{2i}, α_i and ϵ_i are mutually independent. The variables α_i induce the dependence of Y_{1i} and Y_{2i} even adjusting for X_{1i} and X_{2i}. For the big-data inclusion probability, we also consider a logistic linear model

$$\text{logit}(p_i) = X_{2i}, \tag{12}$$

and a nonlinear logistic model

$$\text{logit}(p_i) = -3 + (X_{1i} - 1.5)^2 + (X_{2i} - 2)^2. \tag{13}$$

We consider the following combinations: I. (10) and (12); II. (10) and (13); II. (11) and (12); and IV. (11) and (13) for data generating mechanisms. Therefore, the simulation setup is a 2×2 factorial design with two levels in each factor.

Kim and Wang (2018) proposed the inverse propensity score weighting estimator using the estimated probability of selection into Sample B and the double robust estimator which further incorporates an outcome regression model. To evaluate the robustness and efficiency, we compare

the following estimators:

1. $\hat{\mu}_{\mathrm{HT}}$, the Horvitz–Thompson estimator assuming the Y_i's were observed in Sample A for the purpose of benchmark comparison;

2. $\hat{\mu}_{\mathrm{ipw}}$, the inverse propensity score weighting estimator,

$$\hat{\mu}_{\mathrm{ipw}} = \frac{1}{N}\sum_{i\in B}\frac{1}{p_i(\hat{\eta})}Y_{i(1)},$$

   where $p_i(\eta) = P(\delta_{Bi} = 1 \mid X_i; \eta)$ is a logistic regression model with a linear predictor X_{2i} as a working model, and $\hat{\eta}$ is an estimator of η based on Sample A;

3. $\hat{\mu}_{\mathrm{dr}}$, the double robust estimator,

$$\hat{\mu}_{\mathrm{dr}} = \frac{1}{N}\sum_{i\in B}\frac{1}{p_i(\hat{\eta})}\left(Y_{i(1)} - X_i^{\mathrm{T}}\hat{\beta}\right) + \frac{1}{n}\sum_{i\in A}X_i^{\mathrm{T}}\hat{\beta},$$

   where $\hat{\beta}$ is the estimated regression coefficients using (10) as the working outcome regression model based on Sample B;

4. $\hat{\mu}_{\mathrm{nni}}$, the nearest neighbor imputation estimator;

5. $\hat{\mu}_{\mathrm{knn}}$, the k nearest neighbor imputation estimator with $k = 5$;

6. $\hat{\mu}_{\mathrm{GAM}}$, the generalized additive model imputation estimator;

7. $\hat{\mu}_{\mathrm{RC}}$, the regression calibration estimator based on $\hat{\mu}_{\mathrm{nni}}$ with calibration variables $H(\delta_B, X, Y) = (\delta_B, 1-\delta_B, \delta_B X, \delta_B Y)^{\mathrm{T}}$.

All simulation results are based on $1,000$ Monte Carlo runs. Table 2 summarizes the simulation results with biases, standard errors, and coverage rates of 95% confidence intervals using asymptotic normality of the point estiamtors. The following observations can be made from Table 2. $\hat{\mu}_{\mathrm{ipw}}$ has large biases when the propensity score is misspecified. $\hat{\mu}_{\mathrm{dr}}$ gains robustness over $\hat{\mu}_{\mathrm{ipw}}$ if one of the outcome regression model or the propensity score is correctly specified. However, if both models are misspecified, $\hat{\mu}_{\mathrm{dr}}$ has a larger bias. $\hat{\mu}_{\mathrm{nni}}$ has small biases across four scenarios, which shows its robustness. Importantly, the performance of $\hat{\mu}_{\mathrm{nni}}$ is close to that of $\hat{\mu}_{\mathrm{HT}}$ in terms of standard errors and coverage rates, which is consistent with our theory in Theorem 1. Moreover, as predicted by our theoretical results, $\hat{\mu}_{\mathrm{knn}}$ improves $\hat{\mu}_{\mathrm{nni}}$ in terms of efficiency. Also, $\hat{\mu}_{\mathrm{GAM}}$ shows robustness because of the flexibility of the model specification. The regression calibration estimator $\hat{\mu}_{\mathrm{RC}}$ has small biases across all scenarios and therefore shows robustness against model specifications for sampling score and outcome. Moreover, it has smaller standard errors than both $\hat{\mu}_{\mathrm{nni}}$ and $\hat{\mu}_{\mathrm{knn}}$. The coverage rates are all close to the nominal level.

Table 2: Simulation results: bias, standard error, and coverage rate of 95% confidence intervals under four scenarios based on $1,000$ Monte Carlo samples. OM: outcome model; PS: propensity score model

	Bias $\times 10^2$	S.E. $\times 10^2$	C.R. $\times 10^2$	Bias $\times 10^2$	S.E. $\times 10^2$	C.R. $\times 10^2$	Bias $\times 10^2$	S.E. $\times 10^2$	C.R. $\times 10^2$	Bias $\times 10^2$	S.E. $\times 10^2$	C.R. $\times 10^2$
	Scenario I			Scenario II			Scenario III			Scenario IV		
OM	linear			linear			nonlinear			nonlinear		
PS	linear			nonlinear			linear			nonlinear		
Population Mean of Y_1												
$\hat{\mu}_{\text{HT}}$	0.2	6.5	96.0	-0.2	6.4	94.5	0.61	15.2	95.7	-0.5	15.6	93.5
$\hat{\mu}_{\text{ipw}}$	-0.1	1.6	95.5	25.0	47.0	97.6	-0.1	4.1	95.5	465.1	427.0	76.8
$\hat{\mu}_{\text{dr}}$	0.1	4.6	95.7	0.0	4.5	96.6	0.7	14.0	95.6	266.7	460.2	37.9
$\hat{\mu}_{\text{nni}}$	0.2	6.5	95.1	-0.3	6.4	94.7	0.7	15.2	94.6	-0.6	15.6	93.7
$\hat{\mu}_{\text{knn}}$	0.2	4.9	96.1	-0.3	4.9	95.6	0.5	14.5	94.6	-0.6	14.9	93.8
$\hat{\mu}_{\text{GAM}}$	0.1	4.5	95.7	-0.2	4.5	96.0	0.5	14.3	94.9	-0.6	14.8	93.4
$\hat{\mu}_{\text{RC}}$	0.0	3.2	95.5	-0.2	4.1	95.3	-0.1	4.8	95.0	0.1	6.7	95.5
Population Mean of Y_2												
$\hat{\mu}_{\text{HT}}$	-0.0	1.5	96.2	-0.0	1.6	95.1	-0.1	1.6	95.2	0.1	1.6	94.4
$\hat{\mu}_{\text{ipw}}$	0.0	0.2	95.5	-12.3	3.7	0.0	0.0	0.4	95.9	2.8	4.0	70.6
$\hat{\mu}_{\text{dr}}$	-0.0	0.8	95.9	-0.8	3.8	69.8	-0.0	0.7	95.4	9.7	7.1	10.0
$\hat{\mu}_{\text{nni}}$	0.0	1.4	95.3	-0.0	1.6	95.3	-0.1	1.6	94.6	0.1	1.6	95.3
$\hat{\mu}_{\text{knn}}$	0.0	1.0	95.8	-0.0	1.1	95.8	-0.0	1.0	95.2	0.0	0.9	96.1
$\hat{\mu}_{\text{GAM}}$	-0.0	0.9	95.3	-0.0	0.9	94.8	-0.0	0.8	96.2	0.0	0.8	94.5
$\hat{\mu}_{\text{RC}}$	0.0	1.2	95.5	-0.1	1.4	94.2	-0.0	1.4	94.1	0.1	1.5	95.6
Conditional Mean of Y_1 given $Y_2 = 1$												
$\hat{\mu}_{\text{HT}}$	0.0	7.3	95.1	-0.3	7.2	95.2	0.2	9.3	95.3	-0.1	9.8	94.1
$\hat{\mu}_{\text{ipw}}$	-0.1	1.5	95.5	-8.4	12.2	70.4	-0.1	1.3	95.7	20.6	2.5	0.0
$\hat{\mu}_{\text{dr}}$	0.1	4.7	94.8	2.3	4.5	93.3	0.8	5.5	94.5	24.2	5.5	0.8
$\hat{\mu}_{\text{nni}}$	-0.0	7.3	95.0	-0.3	7.3	95.3	0.1	9.2	95.4	-2.2	9.5	95.2
$\hat{\mu}_{\text{knn}}$	-0.1	4.7	96.8	-0.3	4.6	96.5	0.1	6.0	94.8	0.0	6.4	93.6
$\hat{\mu}_{\text{GAM}}$	0.0	4.8	94.2	-0.3	4.5	96.0	-0.1	6.5	95.5	-0.6	6.8	94.8
$\hat{\mu}_{\text{RC}}$	-0.0	3.9	94.8	-0.2	5.0	96.0	-0.2	5.4	95.1	-0.1	5.4	96.7

Table 3: The stratum size, sample allocation, mean and standard error of the inventory data on the log scale extracted from the 2014 Monthly Retail Trade Survey

Stratum h	1	2	3	4	5	6	7	8
N_h	366	20	2,015	4,646	7,402	700	12,837	17,080
n_h	37	5	34	57	74	7	103	115
$\mu_{X,h}$	16.8	16.7	16.6	16.4	16.1	15.6	16.0	15.7
$\sigma_{X,h}$	1.1	0.8	0.4	0.3	0.4	0.6	0.4	0.4
Stratum h	9	10	11	12	13	14	15	16
N_h	29,808	2,400	41,343	57,518	83,465	95,244	115,028	342,893
n_h	116	12	184	196	218	200	220	336
$\mu_{X,h}$	15.6	15.5	15.4	15.1	14.8	14.5	13.9	11.5
$\sigma_{X,h}$	0.4	0.3	0.4	0.4	0.3	0.7	0.5	1.1

6.2 Monthly retail trade survey

To demonstrate the practical relevance, we consider the U.S. Census Bureau's 2014 Monthly Retail Trade Survey (Mulry et al.; 2014). The Monthly Retail Trade Survey is an economic indicator survey whose monthly estimates are inputs to the Gross Domestic Product estimates. This survey selects a sample of about $12,000$ retail businesses each month with paid employees to collect data on sales and inventories. It employs an one-stage stratified sample with stratification based on major industry, further substratified by the estimated annual sales referred to as the size variable.

For simulation purpose, according to the 2014 Monthly Retail Trade Survey, we generate a finite population of $N = 812,765$ retail businesses with 16 strata with a stratum identifier h, sales Y, inventories X, and a size variable Z on the log scale. Table 3 reports some summary statistics extracted from the actual survey. We generate the inventory and size data from $X_{hi} \sim N(\mu_{X,h}, \sigma^2_{X,h})$ and $Z_{hi} \sim N(\mu_{X,h}, \sigma^2_{X,h})$, for $i = 1, \ldots, N_h$ and $h = 1, \ldots, 16$, and the sales data from a linear model

$$Y_{hi} \quad = \quad \beta_0 + X_{hi} + Z_{hi} + \epsilon_{hi}, \tag{14}$$

and a nonlinear model

$$Y_{hi} \quad = \beta_0 + \quad X^2_{hi} + Z^2_{hi} + \epsilon_{hi}, \tag{15}$$

where $\epsilon_{hi} \sim \mathcal{N}(0, 0.52)$. In (14) and (15), we specify different values for β_0 so that the parameter of interest, $\mu = N^{-1} \sum_{h=1}^{16} \sum_{i=1}^{N_h} Y_{hi}$, matches with the true population mean 12.73.

We also generate a big data sample $\mathcal{S}_B$ where the inclusion indicator δ_{hi} $\sim$Ber(p_{hi}) with the inclusion probability p_{hi} for unit i in stratum h. The big data sample in practice is often available from E-commercial companies who monitor inventories and sales for retail businesses. For the big data inclusion probability, we consider a logistic linear model

$$\text{logit}(p_{hi}) = \alpha_0 + Z_{hi}, \tag{16}$$

Table 4: Simulation results: bias, standard error, and coverage rate of 95% confidence intervals under four scenarios based on 1,000 Monte Carlo runs for the 2014 Monthly Retail Trade Survey. OM: outcome model; PS: propensity score model

	Bias $\times 10^2$	S.E. $\times 10^2$	C.R. $\times 10^2$	Bias $\times 10^2$	S.E. $\times 10^2$	C.R. $\times 10^2$	Bias $\times 10^2$	S.E. $\times 10^2$	C.R. $\times 10^2$	Bias $\times 10^2$	S.E. $\times 10^2$	C.R. $\times 10^2$
	Scenario I			Scenario II			Scenario III			Scenario IV		
OM	linear			linear			nonlinear			nonlinear		
PS	linear			nonlinear			linear			nonlinear		
$\hat{\mu}_{\text{HT}}$	0.0	4.2	96.1	0.0	4.2	96.1	3.3	85.4	96.2	3.3	85.4	96.2
$\hat{\mu}_{\text{ipw}}$	0.1	12.8	96.4	-25.7	7.7	10.3	4.1	340.3	96.3	-850	202.1	1.8
$\hat{\mu}_{\text{dr}}$	0.1	3.6	96.2	0.1	3.6	96.4	4.2	92.5	96.7	-220	93.4	38.6
$\hat{\mu}_{\text{nni}}$	0.0	4.2	96.0	0.1	4.1	96.9	2.6	85.4	96.2	2.1	85.3	96.1
$\hat{\mu}_{\text{knn}}$	0.1	3.7	96.6	0.1	3.7	96.1	1.9	85.3	96.0	1.2	85.2	96.0
$\hat{\mu}_{\text{GAM}}$	0.1	3.6	96.3	0.1	3.60	96.6	-2.7	85.5	96.1	-19.0	85.8	95.7
$\hat{\mu}_{\text{RC}}$	0.0	3.7	95.8	0.1	3.89	96.6	3.2	76.0	96.0	1.0	83.6	96.3

and a nonlinear logistic model

$$\text{logit}(p_{hi}) = \alpha_0 + X_{hi} + Z_{hi}^2, \tag{17}$$

where we specify different values for α_0 so that the mean inclusion probability is about 30%. Lastly, we generate a representative sample $\mathcal{S}_A$ by stratified sampling with simple random sampling within strata without replacement; see Table 3 for the sample allocation.

We consider the seven estimators in § 6.1 adopted for stratified sampling. In each mass imputed dataset, we apply the following point estimator and variance estimator: $\hat{\mu} = N^{-1} \sum_{h=1}^{H} N_h \bar{y}_{n_h}$ with $\bar{y}_{n_h}$ is the sample mean of y in the hth stratum, $\hat{V}(\hat{\mu}) = N^{-2} \sum_{h=1}^{H} N_h^2 (1 - n_h/N_h) s_{n_h}^2 / n_h$ with $s_{n_h}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_{n_h})^2$.

Table 4 summarizes the simulation results. A similar discussion to § 6.1 applies. $\hat{\mu}_{\text{ipw}}$ is sensitive to misspecification of the selection model; while $\hat{\mu}_{\text{dr}}$ has double robustness feature, which still relies on at least one model to be correctly specified. Mass imputation based on nearest neighbor imputation, k nearest neighbor imputation and generalized additive model shows good performances by leveraging the representativeness of the survey sample and the predictive power of the big data sample. In addition, if the big data membership is known throughout the survey data, the regression calibration estimator gains efficiency while maintaining the robustness against model misspecification.

7 Discussion

Mass imputation is an important technique for survey data integration. When the training dataset for imputation is obtained from a probability sample, the theory of Kim and Rao (2012) can be directly applied. If the training dataset is a non-probability sample and its size is huge, we have shown in this article that various nonparametric methods can be used for mass imputation, and

the estimation error in the imputation model can be safely ignored, under the assumption that the sampling mechanism for training data is missing at random in the sense of Rubin (1976). If the sampling mechanism is believed to be not missing not at random, imputation techniques can be developed under the strong model assumptions for the sampling mechanism (e.g. Riddles et al.; 2016; Morikawa and Kim; 2018). Also, when the training dataset has a hierarchical structure, multilevel models can be used to develop mass imputation. This is closely related to unit-level small area estimation in survey sampling (Rao and Molina; 2015). These are topics for future research.

Supplementary Material

Supplementary material includes the proofs for three theorems.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects, *Econometrica* **74**: 235–267.

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.

Beretta, L. and Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation, *BMC Med. Inform. Decis. Mak.* **16**: 198–208.

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.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques, *Statist. Sci.* **32**: 190–205.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random, *Journal of the American Statistical Association* **89**: 81–87.

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

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling, *J. Am. Stat. Assoc.* **87**: 376–382.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties, *Statistical Science* **11**: 89–102.

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

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, NY: Chapman and Hall, Inc.

Imbens, G., Johnson, P. and Spady, R. H. (1998). Information theoretic approaches to inference in moment condition models, *Econometrika* **66**: 333–357.

Kalton, G. and Kish, L. (1984). Some efficient random imputation methods, *Communications in Statistics-Theory and Methods* **13**: 1919–1939.

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**: 319–376.

Kim, J. K. and Fuller, W. (2004). Fractional hot deck imputation, *Biometrika* **91**: 559–578.

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

Kitamura, Y. and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation, *Econometrica* **65**: 861–874.

Klenke, A. (2006). *Probability Theory*, Springer-Verlag: Heidelberg.

Mack, Y.-P. (1981). Local properties of k-NN regression estimates, *SIAM Journal on Algebraic Discrete Methods* **2**: 311–323.

Mack, Y. and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates, *Journal of Multivariate Analysis* **9**: 1–15.

McCullagh, P. (1984). Generalized linear models, *Eur. J. Oper. Res.* **16**: 285–292.

Morikawa, K. and Kim, J. K. (2018). A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse, *Statistics & Probability Letters* **140**: 1–6.

Mulry, M. H., Oliver, B. E. and Kaputa, S. J. (2014). Detecting and treating verified influential values in a Monthly Retail Trade Survey, *J. Offic. Stat.* **30**: 721–747.

Nelder, J. A. and Baker, R. J. (1972). *Generalized Linear Models*, Wiley Online Library.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators, *Econometrica* **72**: 219–255.

Park, M. and Fuller, W. A. (2012). Generalized regression estimators, *Encyclopedia of Environmetrics* **2**: 1162–1166.

Rao, J. N. and Molina, I. (2015). *Small Area Estimation*, Wiley: New York.

Riddles, M. K., Kim, J. K. and Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse, *Journal of Survey Statistics and Methodology* **4**: 215–245.

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

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

Ruppert, D., Wand, M. P. and Carroll, R. J. (2009). Semiparametric regression during 2003–2007, *Electronic Journal of Statistics* **3**: 1193–1256.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood, *Ann. Statist.* **35**: 634–672.

van der Vaart, A. W. (2000). *Asymptotic Statistics*, Cambridge University Press, Cambridge, MA.

Wood, S. N. (2006). *Generalized Additive Models: an Introduction with R*, Chapman and Hall/CRC.

Supplementary material

by Shu Yang and Jae Kwang Kim

S1 Proofs

S1.1 Proof for Theorem 1

For a given $X_i = x$ in Sample A, we show that $X_{i(1)}$ converges to x in probability as $N_B \to \infty$. To show this, consider for any $\epsilon > 0$,

$$\begin{aligned} P\{d(X_{i(1)}, x) > \epsilon\} &= P\{d(X_j, x) > \epsilon \ \forall j \in B\} \\ &= [P\{d(X_j, x) > \epsilon\}]^{N_B} . \end{aligned} \tag{S1}$$

By Assumption 2, x is in the support of X in Sample B. This leads to $P\{d(X_j, x) < \epsilon\} > 0$ and $P\{d(X_j, x) > \epsilon\} < 1$. Therefore, (S1) converges to zero, and $X_{i(1)}$ converges to x in probability as $N_B \to \infty$.

Given $X_i = x$, for any continuous and bounded $g(y)$,

$$\begin{aligned} E\{g(Y_{i(1)}) \mid X_i = x\} &= E[E\{g(Y_{i(1)}) \mid X_{i(1)}\} \mid X_i = x] \\ &= E\{\mu_g(X_{i(1)}) \mid X_i = x\} \\ &\to E\{\mu_g(X_i) \mid X_i = x\} \\ &= E\left\{g(Y_i) \mid X_i = x\right\}, \end{aligned}$$

in probability as $N_B \to \infty$, where $\to$ follows from the fact that $\mu_g(x)$ is bounded and continuous. Then, by Portmanteau Lemma (Klenke; 2006), $Y_{i(1)} \to Y_i \mid (X_i = x)$ in distribution as $N_B \to \infty$. By Assumption 1, $g(Y_{i(1)}) \mid X_i \to \mu_g(X_i) + e_g^*(X_i)$ in distribution as $N_B \to \infty$, where $e_g^*(X_i)$ has the same distribution as $\{g(Y_i) \mid X_i\} - \mu_g(X_i)$.

We now show that for $i \neq j \in A$, $e_g^*(X_i)$ and $e_g^*(X_j)$ are conditionally independent, given data $\mathcal{O}_A$. It is sufficient to show that $P\left\{i(1) = j(1)\right\} \to 0$ as $N_B \to \infty$; in other words, the same unit can not be matched for unit i and unit j with probability 1. This can be shown using (S1) with $\epsilon = \min_{i \neq j \in A} ||X_i - X_j||$.

Therefore, conditional on data $\mathcal{O}_A$,

$$\hat{\mu}_{g,\text{nni}} = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} g(Y_{i(1)}) \to \frac{1}{N} \sum_{i \in A} \pi_i^{-1} g(Y_i) = \hat{\mu}_{g,\text{HT}}$$

in distribution as $N_B \to \infty$. This completes the proof for Theorem 1.

Let

$$\tilde{V}_{\text{nni}} = \frac{n}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{g(Y_i)}{\pi_i} \frac{g(Y_j)}{\pi_j}. \tag{S2}$$

Then, $\tilde{V}_{\text{nni}}$ is consistent for V_{nni}.

Similar to the above argument, for $i, j \in A$, conditional on data $\mathcal{O}_A$, $g(Y_{i(1)})g(Y_{j(1)}) \to g(Y_i)g(Y_j)$ as $N_B \to \infty$. Therefore, conditional on data $\mathcal{O}_A$,

$$\hat{V}_{\text{nni}} = \frac{n}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{g(Y_{i(1)})}{\pi_i} \frac{g(Y_{j(1)})}{\pi_j} \to \tilde{V}_{\text{nni}}, \tag{S3}$$

in distribution as $N_B \to \infty$. Combining (S2) and (S3), $\hat{V}_{\text{nni}}$ is consistent for V_{nni}.

S1.2 Proof for Theorem 2

To investigate the asymptotic properties of $\hat{\mu}_{g,\text{knn}}$, we re-express

$$\hat{\mu}_g(x) = \frac{\sum_{j \in B} K_{R_x}(x - X_j) g(Y_j)}{\sum_{j \in B} K_{R_x}(x - X_j)},$$

where

$$K_h(u) = \frac{1}{h^p} K\left(\frac{u}{h}\right), \quad K(u) = 0.5 I(||u|| \leq 1),$$

and the bandwidth $h = R_x$ is the random distance between x and its furthest among the k nearest neighbors. Therefore, $\hat{\mu}_{g,\text{knn}}$ can be viewed as a kernel estimator incorporating a data-driven bandwidth.

In the literature, asymptotic properties of the k nearest neighbor imputation estimator have been studied extensively. The result shown in the following lemma on k nearest neighbor imputation is extracted from Mack (1981).

Lemma S1 *Under Assumptions 1–3,*

$$N^{-1} \sum_{j=1}^{N} \delta_{B,j} K_{R_x}(x - X_j) g(Y_j) = f(x) \pi_B(x) \mu_g(x) + O_p\left\{ \left(\frac{k}{N}\right)^{2/p} + \frac{1}{k} \right\}. \tag{S4}$$

We now express

$$\hat{\mu}_{g,\text{knn}} \;=\; \frac{1}{N} \sum_{i=1}^{N} \pi_i^{-1} \delta_{A,i} \mu_g(X_i) + \frac{1}{N} \sum_{i=1}^{N} \pi_i^{-1} \delta_{A,i} \left\{ \hat{\mu}_g(X_i) - \mu_g(X_i) \right\}.$$

Let $T_N = N^{-1} \sum_{i=1}^{N} \pi_i^{-1} \delta_{A,i} \{\hat{\mu}_g(X_i) - \mu_g(X_i)\}$. To study the properties for T_N, we first look at $\hat{\mu}_g(x)$, which can be expressed as

$$\hat{\mu}_g(x) = \frac{h_N(x)}{f_N(x)},$$

where $h_N(x) \equiv N^{-1}\sum_{j=1}^N \delta_{B,j} K_{R_x}(x - X_j) g(Y_j)$ and $f_N(x) \equiv N^{-1}\sum_{j=1}^N \delta_{B,j} K_{R_x}(x - X_j)$. By the result in Lemma S1, we obtain

$$\begin{aligned} h_N(x) &= f(x)\pi_B(x)\mu_g(x) + O_p\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\} \\ f_N(x) &= f(x)\pi_B(x) + O_p\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}. \end{aligned}$$

Now, by a Taylor expansion, we obtain

$$\begin{aligned} \hat{\mu}_g(x) - \mu_g(x) &= \frac{h_N(x)}{f_N(x)} - \mu_g(x) \\ &= \frac{1}{f(x)\pi_B(x)}\{h_N(x) - f(x)\pi_B(x)\mu_g(x)\} \\ &\quad -\frac{f(x)\pi_B(x)\mu_g(x)}{\{f(x)\pi_B(x)\}^2}\{f_N(x) - f(x)\pi_B(x)\} + O_p\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\} \\ &= \frac{1}{f(x)\pi_B(x)}\{h_N(x) - f_N(x)\mu_g(x)\} + O_p\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}. \end{aligned}$$

Therefore, we obtain

$$T_N = \frac{1}{N^2}\sum_{i=1}^N \frac{\delta_{A,i}}{\pi_i}\frac{1}{f(X_i)\pi_B(X_i)}\sum_{j=1}^N \delta_{B,j} K_{R_{X_i}}(X_i - X_j)\{g(Y_j) - \mu_g(X_i)\} + O_p\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.$$

Under the assumption in Theorem 2, it is easy to drive that $(k/N)^{2/p} + 1/k = o(n^{-1/2})$, and therefore,

$$T_N = \frac{1}{N^2}\sum_{i=1}^N \frac{\delta_{A,i}}{\pi_i}\frac{1}{f(X_i)\pi_B(X_i)}\sum_{j=1}^N \delta_{B,j} K_{R_{X_i}}(X_i - X_j)\{g(Y_j) - \mu_g(X_i)\} + o_p(n^{-1/2}).$$

We then express T_N in a form of U-statistics (van der Vaart; 2000; Ch. 12):

$$T_N = \frac{1}{N(N-1)}\sum_{i=1}^N \sum_{j\neq i} h(Z_i, Z_j) + o_p(n^{-1/2}),$$

where $Z_i = (X_i, Y_i, \delta_{A,i}, \delta_{B,i})$ and

$$\begin{aligned} h(Z_i, Z_j) &= \frac{1}{2}\left[\frac{\delta_{A,i}\delta_{B,j}}{\pi_i}\frac{1}{f(X_i)\pi_B(X_i)}K_{R_{X_i}}(X_i - X_j)\{g(Y_j) - \mu_g(X_i)\}\right. \\ &\quad \left. + \frac{\delta_{A,j}\delta_{B,i}}{\pi_j}\frac{1}{f(X_j)\pi_B(X_j)}K_{R_{X_j}}(X_j - X_i)\{g(Y_i) - \mu_g(X_j)\}\right] \\ &\equiv \frac{1}{2}(\zeta_{ij} + \zeta_{ji}). \end{aligned}$$

Now, by Lemma S1, we obtain

$$\begin{aligned} E(\zeta_{ij} \mid Z_i) &= E\left[\frac{\delta_{A,i}\delta_{B,j}}{\pi_i}\frac{1}{f(X_i)\pi_B(X_i)}K_{R_{X_i}}(X_i - X_j)\{g(Y_j) - \mu_g(X_i)\} \mid Z_i\right] \\ &= \frac{\delta_{A,i}}{\pi_i}\frac{1}{f(X_i)\pi_B(X_i)}E\left[\delta_{B,j}K_{R_{X_i}}(X_i - X_j)\{g(Y_j) - \mu_g(X_i)\} \mid Z_i\right] \\ &= O\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}, \end{aligned}$$

and

$$\begin{aligned} E(\zeta_{ji} \mid Z_i) &= E\left[\frac{\delta_{A,j}\delta_{B,i}}{\pi_j}\frac{1}{f(X_j)\pi_B(X_j)}K_{R_{X_j}}(X_j - X_i)\{g(Y_i) - \mu_g(X_j)\} \mid Z_i\right] \\ &= \delta_{B,i}E\left(E\left[\frac{\delta_{A,j}}{\pi_j}\frac{1}{f(X_j)\pi_B(X_j)}K_{R_{X_j}}(X_j - X_i)\{g(Y_i) - \mu_g(X_j)\} \mid R_{X_j}, Z_i\right] \mid Z_i\right) \\ &= \frac{\delta_{B,i}}{\pi_B(X_i)}\{g(Y_i) - \mu_g(X_i)\} + O\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}. \end{aligned}$$

Therefore, by the theory of U-statistics, we obtain

$$\begin{aligned} T_N &= \frac{2}{N}\sum_{i=1}^{N} E\{h(Z_i, Z_j) \mid Z_i\} + o_p(n^{-1/2}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_{B,i}}{\pi_B(X_i)}\{g(Y_i) - \mu_g(X_i)\} + o_p(n^{-1/2}). \end{aligned}$$

Combining the above results leads to

$$\begin{aligned} \hat{\mu}_{g,\mathrm{knn}} - \mu_g &= \frac{1}{N}\sum_{i=1}^{N}\left\{\pi_i^{-1}\delta_{A,i}\mu_g(X_i) - \mu_g(X_i)\right\} \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\delta_{B,i}}{\pi_B(X_i)} - 1\right\}\{g(Y_i) - \mu_g(X_i)\} + o_p(n^{-1/2}). \end{aligned} \tag{S5}$$

Then, the asymptotic results in Theorem 2 follow by Assumptions 1–4 and (S5).

S1.3 Proof for Theorem 3

The consistency and asymptotic normality of $n^{1/2}\hat{\mu}_{g,\mathrm{nni}}$ follow by the standard arguments under Assumptions 1–4. The remaining is to show that the asymptotic variance of $n^{1/2}\hat{\mu}_{g,\mathrm{nni}}$ is V_{nni}.

Using the distance function $G(\omega_i, d_i) = d_i(\omega_i/d_i - 1)^2$ in (7), the minimum distance estimation

leads to generalized regression estimation (Park and Fuller; 2012). Therefore, we express

$$\begin{aligned} n^{1/2}\hat{\mu}_g &= \frac{n^{1/2}}{N}\sum_{i\in A}\omega_i g(Y_{i(1)}) \\ &= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1} g(Y_{i(1)}) - \frac{n^{1/2}}{N}\left(\sum_{i\in A}\pi_i^{-1}h_i^{\mathrm{T}}\beta_N - \sum_{i=1}^{N}h_i^{\mathrm{T}}\beta_N\right) + o_p(n^{-1/2}). \end{aligned} \tag{S6}$$

Similar to the argument in the proof for Theorem 1, we express

$$\begin{aligned} n^{1/2}\hat{\mu}_g &= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1} g(Y_{i(1)}) - \frac{n^{1/2}}{N}\left(\sum_{i\in A}\pi_i^{-1}h_i^{\mathrm{T}}\beta_N - \sum_{i=1}^{N}h_i^{\mathrm{T}}\beta_N\right) + o_p(n^{-1/2}) \\ &= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}\left\{g(Y_{i(1)}) - h_i^{\mathrm{T}}\beta_N\right\} + \frac{n^{1/2}}{N}\sum_{i=1}^{N}h_i^{\mathrm{T}}\beta_N + o_p(n^{-1/2}). \end{aligned} \tag{S7}$$

It is straightforward to show the variance of the second term in (S7) is negligible given $nN^{-1} = o(1)$. Following the arguments in the proof for Theorems 1 and 2, $g(Y_{i(1)})$ has the asymptotic distribution as $g(Y_i)$ given the data $\mathcal{O}_A$ from Sample A. Therefore, the asymptotic variance of $n^{1/2}\hat{\mu}_g$ is

$$V_{\mathrm{RC}} = \lim_{n\to\infty}\mathrm{var}\left[\frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}\left\{g(Y_i) - h_i^{\mathrm{T}}\beta_N\right\}\right].$$