Received: 21 December 2018 | Revised: 8 September 2019 | Accepted: 30 October 2019

DOI: 10.1111/sjos.12429

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework

Shu Yang[1] | **Jae Kwang Kim**[2]

[1]Department of Statistics, North Carolina State University

[2]Department of Statistics, Iowa State University

Correspondence
Jae Kwang Kim, Department of Statistics, Iowa State University, 1208 Snedecor Hall, Ames, IA 50011.
Email: jkim@iastate.edu

Funding information
US National Science Foundation (US-NSF), 1733572

Abstract

Predictive mean matching imputation is popular for handling item nonresponse in survey sampling. In this article, we study the asymptotic properties of the predictive mean matching estimator for finite-population inference using a superpopulation model framework. We also clarify conditions for its robustness. For variance estimation, the conventional bootstrap inference is invalid for matching estimators with a fixed number of matches due to the nonsmoothness nature of the matching estimator. We propose a new replication variance estimator, which is asymptotically valid. The key strategy is to construct replicates directly based on the linear terms of the martingale representation for the matching estimator, instead of individual records of variables. Simulation studies confirm that the proposed method provides valid inference.

KEYWORDS

hot deck imputation, Jackknife variance estimation, martingale central limit theorem, missing at random

1 | INTRODUCTION

Predictive mean matching imputation (Heitjan & Little, 1991; Little, 1988; Rubin, 1986) is used widely to compensate for item nonresponse in survey sampling. First, it is a hot deck method

© 2019 Board of the Foundation of the Scandinavian Journal of Statistics

(Ford, 1983), because the donors for a missing value are actually observed values from respondents. Using real values instead of artificial values for imputation is often preferred in government statistical agencies. Second, it is a special version of nearest neighbor imputation. In nearest neighbor imputation, the vector of the auxiliary variables is directly used in determining the nearest neighbor for each nonrespondent, whereas in predictive mean matching imputation, a scalar predictive mean function is used in determining the nearest neighbor. Schenker and Taylor (1996) and Horton and Lipsitz (2001) advocated predictive mean matching imputation for its robustness against model misspecification.

Although these imputation methods have a long history of application, there are relatively few articles that investigate their theoretical properties. Chen and Shao (2000, 2001) have developed a nice set of asymptotic theories for the nearest neighbor imputation estimator. Beaumont and Bocci (2009) developed a model-based variance estimator for the nearest neighbor imputation estimator. Kim, Fuller, and Bell (2011) studied nearest neighbor imputation with an application to the U.S. census long form data. Yang and Kim (2019) provided theoretical investigation of the nearest neighbor imputation estimator for general population parameters, including population means, proportions, and quantiles. Vink, Frank, Pannekoek, and Buuren (2014) and Morris, White, and Royston (2014) adopted predictive mean matching as a tool for multiple imputations. In econometrics, Abadie and Imbens (2006, 2008, 2011, 2016) studied the matching estimator for causal effects from observational studies. To the best of our knowledge, a theoretical investigation of predictive mean matching for finite-population inference in survey sampling seems to be lacking.

Predictive mean matching is implemented in two steps. First, the predictive mean function is specified and estimated from the respondents. Second, for each nonrespondent, the nearest neighbor is identified among the respondents based on the estimated predictive mean function. Then, the observed outcome value of the nearest neighbor is used for imputation. Because the predictive mean function is estimated prior to matching, it is necessary to account for the uncertainty due to parameter estimation. The typical Taylor expansion technique is not applicable, because of the nonsmooth nature of matching. Our proposal is based on the technique developed by Andreou and Werker (2012), which offers a general approach for deriving the limiting distribution of statistics that involve estimated nuisance parameters. This technique has been successfully used by Abadie and Imbens (2016) for the matching estimators of the average causal effects based on the estimated propensity score. We extend their results to the predictive mean matching estimator in survey sampling. Abadie and Imbens (2016) clarified the seemingly paradoxical phenomenon that matching on the estimated propensity score improves the estimation of the average causal effect compared with matching on the true propensity score. However, we note that matching on the estimated predictive mean function can either increase or decrease the estimation efficiency compared with matching on the true predictive mean function. This is because the propensity score is not pertinent to the true values of the population mean parameters, while the predictive mean function is. In addition, we provide a condition for the robustness of the predictive mean matching estimator that allows the predictive mean function to be misspecified to some extent.

For variance estimation of the predictive mean matching estimator, Morris et al. (2014) suggested using multiple imputation without theoretical justification. We consider an alternative route for variance estimation based on replication methods (Efron, 1979; Wolter, 2007). Lack of smoothness makes the conventional replication methods invalid for the predictive mean matching estimator. If the number of matches increases with the sample size, such as in kernel matching and local linear matching (Heckman, Ichimura, Smith, & Todd, 1998; Heckman, Ichimura, & Todd, 1997), the matching estimator is asymptotically smooth, which enables the conventional replication methods for inference. When the number of matches remains fixed, Abadie and Imbens (2008) demonstrated the failure of the bootstrap for matching estimators in the setting with independently and identically distributed data. This is because the nonparametric bootstrap cannot preserve the distribution of the number of times that each unit is used as a match. To overcome this issue, Otsu and Rai (2017) proposed a wild bootstrap procedure for the matching estimator when matching is directly based on the covariates. Following the two-step procedure for the predictive mean matching estimator, the variability of the matching estimator results from three sources: sampling, estimation of the predictive mean function, and matching. We propose a new replication variance estimation procedure, which faithfully takes these sources of variability into account. Toward that end, we construct replicates of the estimator following the steps for the predictive mean matching estimator. First, we construct replicates of the nuisance parameter estimators in the predictive mean function. Second, based on the martingale representation of the predictive mean matching estimator, we construct replicates of the matching estimator directly based on its linear terms with the replicated nuisance parameters. In this way, the distribution of the number of times that each unit is used as a match can be retained, which leads to valid variance estimation. Utilizing the parallelism between the replication procedure and the predictive mean matching procedure, we demonstrate the consistency of the proposed replication variance estimator by extending the technique of Andreou and Werker (2012) to the replication process. Furthermore, our replication method is flexible and can accommodate bootstrap and jackknife, among others.

The rest of this article is organized as follows. In Section 2, we introduce the basic setup for the survey sample data with nonresponse and the predictive mean matching procedure. In Section 3, we establish and compare the asymptotic distributions of the predictive mean matching estimator when the predictive mean function is known or estimated. In Section 4, we propose the new replication variance estimators and establish their consistency. In Section 5.1, we evaluate the finite sample performance of the proposed methods via a simulation study. In Section 5.2, we apply the proposed methods to a real survey example estimating the academic performance of schools in California based on the academic performance index (API) program. We end with a brief discussion in Section 6. All proofs are deferred to Appendix.

2 | BASIC SETUP

2.1 | Notation and assumptions

Let $\mathcal{F}_N = \{(x_i, y_i, \delta_i) : i = 1, \dots, N\}$ denote a finite population, where N is known, a vector of covariates x_i is always observed, y_i is the study variable, which is subject to missingness, and δ_i is the response indicator of y_i, that is, $\delta_i = 1$ if y_i is observed and 0 if it is missing. The δ_i's are defined throughout the finite population as in Fay (1991), Shao and Steel (1999), and Kim, Navarro, and Fuller (2006). To fix ideas, we focus on estimating the finite population mean $\mu = N^{-1} \sum_{i=1}^{N} y_i$, although our framework can be extended to general population quantities; see Section 6. Let A denote an index set of the sample selected under a probability sampling design. Let I_i be the sampling indicator, that is, $I_i = 1$ if unit i is selected into the sample, and $I_i = 0$ otherwise. We assume that π_i, the probability of selection of i, is positive and known throughout the sample.

To facilitate imputation and theoretical derivation, we consider a *superpopulation* framework, where we assume that $\mathcal{F}_N$ is a random sample from a superpopulation model ζ. We first make the following assumption for the missing data process.

Assumption 1 (Missing at random and positivity). (i) The missing data process satisfies $\text{pr}(\delta = 1 | x, y) = \text{pr}(\delta = 1 | x)$, denoted by $p(x)$; and (ii) with probability 1, $p(x) > \epsilon$ for a constant $\epsilon > 0$.

Assumption 1 (i) states that the missingness depends only on x but not on the unobserved value y (Rubin, 1976). Although this assumption is not testable, researchers can collect a rich set of covariates that are associated with both the outcome and the missingness process to ensure this assumption holds. Assumption 1 (ii) indicates that for any possible value x, there is a positive probability for the unit with such characteristic to respond. Positivity is testable based on the empirical distribution of x between the respondents and the nonrespondents. When the positivity assumption is violated, the distribution of x from the respondents and that from the nonrespondents may not fully overlap. Then the matching discrepancy of the nonrespondents in the nonoverlapping region and their nearest neighbors among the respondents would be large even when the sample size grows, and therefore any matching estimators would be biased. In this case, one may consider trimming the sample (Yang & Ding, 2018; Yang, Imbens, Cui, Faries, & Kadziola, 2016), which however, changes the target population and the parameter of interest. In this article, we focus on the case when positivity holds and leaves the issue of limited overlap to our future work.

2.2 | Nearest neighbor imputation

Nearest neighbor imputation hinges on imputing the missing outcome for each nonrespondent by matching directly on the covariate among the respondents. Let $A = A_R \cup A_M$, where $A_R = \{i \in A : \delta_i = 1\}$ and $A_M = \{i \in A : \delta_i = 0\}$ are the index sets of respondents and nonrespondents, respectively. Nearest neighbor imputation can be described in the following steps:

Step 1. For each unit i with $\delta_i = 0$, find the nearest neighbor from the respondents with the minimum distance between x_j and x_i. Let $\nu(i)$ be the index set of its nearest neighbor, which satisfies $d(x_{\nu(i)}, x_i) \leq d(x_j, x_i)$, for $j \in A_R$. For the distance function, we can use the Euclidean norm $d(x_i, x_j) = ||x_i - x_j||$, where $||x|| = (x^{\mathrm{T}}x)^{1/2}$. Other norms of the form $||x||_D = (x^{\mathrm{T}}Dx)^{1/2}$, where D is a positive definite symmetric matrix D, are equivalent to the Euclidean norm, because $||x||_D = \{(Qx)^{\mathrm{T}}(Qx)\}^{1/2} = ||Qx||$ with $Q^{\mathrm{T}}Q = D$. In particular, Mahalanobis distance is commonly used, where $D = \hat{\Sigma}^{-1}$ with $\hat{\Sigma}$ the empirical covariance matrix of x.

Step 2. The nearest neighbor imputation estimator of μ is

$$\hat{\mu}_{\mathrm{NNI}} = \frac{1}{N} \sum_{i \in A} \frac{1}{\pi_i} \{\delta_i y_i + (1 - \delta_i) y_{\nu(i)}\} = \frac{1}{N} \sum_{i \in A} \frac{1}{\pi_i} \delta_i (1 + k_i) y_i, \tag{1}$$

where $k_i = \sum_{j \in A} \pi_i \pi_j^{-1} (1 - \delta_j) d_{ij}$, and $d_{ij} = 1$ if $\nu(j) = i$, that is, unit i is used as a donor for unit $j \in A_M$, and $d_{ij} = 0$ otherwise.

Under simple random sampling, $k_i = \sum_{j \in A} (1 - \delta_j) d_{ij}$ becomes the number of times that unit i is used as a match for nonrespondents.

To study the asymptotic properties of Equation (1), define $m(x) = E(y|x)$. Following Yang and Kim (2019), we assume the following conditions hold.

Assumption 2. (i) The matching variable x has a compact and convex support, with its density bounded and bounded away from zero. Let $g_1(x_i)$ and $g_0(x_i)$ be the conditional density of x_i given

$\delta_i = 1$ and $\delta_i = 0$, respectively. There exist constants C_{1L} and C_{1U} such that $C_{1L} \leq g_1(x_i)/g_0(x_i) \leq C_{1U}$ almost surely; and (ii) $m(x)$ is Lipschitz continuous in x; that is, there exists a constant C_3 such that $d\{m(x_i), m(x_j)\} \leq C_3 d(x_i, x_j)$, for any i, j.

Denote $E_p(\cdot)$ and $\text{var}_p(\cdot)$ to be the expectation and the variance under the sampling design, respectively. For the asymptotics, we consider both the sample size n and the population size N to go to infinity. We impose the following regularity conditions on the sampling design.

Assumption 3. (i) There exist positive constants C_1 and C_2 such that $C_1 \leq \pi_i N n^{-1} \leq C_2$, for $i = 1, \dots, N$; (ii) the sampling fraction is negligible, $nN^{-1} = o(1)$; (iii) let z_i represent x_i or y_i, and let the corresponding population quantity and the Horvitz–Thompson estimator be $\mu_z = N^{-1} \sum_{i=1}^{N} z_i$ and $\hat{\mu}_{z,\text{HT}} = N^{-1} \sum_{i \in A} \pi_i^{-1} z_i$, respectively, which satisfy that $\text{var}_p(\hat{\mu}_{z,\text{HT}}) = O(n^{-1})$ and $\{\text{var}_p(\hat{\mu}_{z,\text{HT}})\}^{-1/2} \times (\hat{\mu}_{z,\text{HT}} - \mu_z) | \mathcal{F}_N \to \mathcal{N}(0, 1)$ in distribution, as $n \to \infty$.

The conditions in Assumption 3 are widely accepted in survey sampling (Fuller, 2009, chapter 1). Assumption 3 (ii) is a convenient condition to simplify theoretical derivations, which can be relaxed; see Remark 1 and the discussion in Section 6.

We write $n^{1/2}(\hat{\mu}_{\text{NNI}} - \mu) = D_N + B_N$, where

$$D_N = n^{1/2} \left(\frac{1}{N} \sum_{i \in A} \frac{1}{\pi_i} [m(x_i) + \delta_i (1 + k_i) \{y_i - m(x_i)\}] - \mu \right),$$

and

$$B_N = \frac{n^{1/2}}{N} \sum_{i \in A} \frac{1}{\pi_i} (1 - \delta_i) \{m(x_{\nu(i)}) - m(x_i)\}. \tag{2}$$

The difference $m(x_{\nu(i)}) - m(x_i)$ in Equation (2) accounts for the matching discrepancy, and B_N contributes to the asymptotic bias of the matching estimator. In general, for a p-dimensional matching variable x, Abadie and Imbens (2006) showed that $d(x_{\nu(i)}, x_i) = O_p(n^{-1/p})$. Therefore, if matching is directly based on a p-vector covariate with $p \geq 2$, the bias is

$$\begin{aligned} B_N &= \frac{n^{1/2}}{N} \sum_{i \in A} \frac{1}{\pi_i} (1 - \delta_i) \{m(x_{\nu(i)}) - m(x_i)\} \\ &= O_p \left\{ \frac{n^{1/2}}{N} \sum_{i \in A} \frac{1}{\pi_i} (1 - \delta_i) d(x_{\nu(i)}; x_i) \right\} = O_p(n^{1/2 - 1/p}) \neq o_p(1), \end{aligned} \tag{3}$$

where the second line of Equation (3) follows from Assumption 2 (ii) and Assumption 3 (i). Herein, the probability distribution is the joint distribution of the sampling distribution and the superpopulation model ζ.

For bias correction, let $\hat{m}(x)$ be a consistent estimator of $m(x)$, for example, using sieve estimation (Chen, 2007; Newey, 1997). Then, we can estimate B_N by $\hat{B}_N = n^{1/2} N^{-1} \sum_{i \in A} \pi_i^{-1} (1 - \delta_i) \{\hat{m}(x_{\nu(i)}) - \hat{m}(x_i)\}$. Under certain regularity conditions imposed on $\hat{m}(x)$ (Abadie & Imbens, 2011), $\hat{B}_N$ is consistent for B_N, that is, $\hat{B}_N - B_N = o_p(1)$. A bias-corrected nearest neighbor imputation estimator of μ is

$$\tilde{\mu}_{\text{NNI}} = \frac{1}{N} \sum_{i \in A} \frac{1}{\pi_i} \{\delta_i y_i + (1 - \delta_i) y_i^*\}, \tag{4}$$

where $y_i^* = \hat{m}(x_i) + y_{\nu(i)} - \hat{m}(x_{\nu(i)})$. Although $\tilde{\mu}_{\text{NNI}}$ is asymptotically unbiased, the imputed value y_i^* may not be an actual realized value and therefore bias-corrected nearest neighbor imputation is no longer a hot deck imputation method. To overcome the curse of dimensionality and meanwhile retain the hot deck imputation mechanism, we investigate predictive mean matching and its asymptotic properties in the following section.

3 | PREDICTIVE MEAN MATCHING

To reduce the dimension of the matching variable, we assume that

$$E(y_i|x_i) = m(x_i;\beta^*) \tag{5}$$

holds for every unit in the population, where $m(\cdot;\beta^*)$ is a function of x known up to β^*. Under Assumption 1, let the normalized estimating equation for β^* be

$$S_N(\beta) = \frac{n^{1/2}}{N}\sum_{i\in A}\frac{1}{\pi_i}\delta_i g(x_i;\beta)\{y_i - m(x_i;\beta)\} = 0, \tag{6}$$

where $g(x;\beta)$ is any function that ensures that the solution to Equation (6) exists and is unique. To simplify the presentation, let $g(x;\beta)$ be $\dot{m}(x;\beta) = \partial m(x;\beta)/\partial\beta$. General functions for $g(x;\beta)$ can be considered at the expense of heavier notation. Under certain regularity conditions specified in Appendix, the solution to Equation (6), $\hat{\beta}$, converges to β^* in probability. Adjusting by the sampling weight π_i^{-1} in Equation (6) guarantees a consistent estimator of β^* even under informative sampling (Berg, Kim, & Skinner, 2016).

Under the model in Equation (5), predictive mean matching can be described as follows:

Step 1. Obtain a consistent estimator of β, denoted by $\hat{\beta}$, by solving Equation (6). For each unit i, obtain a predicted value of y_i as $\hat{m}_i = m(x_i;\hat{\beta})$. Find the nearest neighbor of unit i with $\delta_i = 0$ from the respondents with the minimum distance between $\hat{m}_j$ and $\hat{m}_i$. As a slight abuse of notation, let $\nu(i)$ be the index of the nearest neighbor of unit i in A_R, where determination of the nearest neighbor is based on the estimated predictive mean function $m(x_i;\hat{\beta})$, which satisfies $d(\hat{m}_{\nu(i)},\hat{m}_i) \le d(\hat{m}_j,\hat{m}_i)$, for any $j \in A_R$.

Step 2. The imputation estimator based on predictive mean matching is

$$\hat{\mu}_{\text{PMM}} = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}\{\delta_i y_i + (1-\delta_i)y_{\nu(i)}\}. \tag{7}$$

In Equation (7), the imputed values are real observations. The imputation model is only used for identifying the nearest neighbor, but not for creating the imputed values.

To study the asymptotic properties of the predictive mean matching estimator, we write $\hat{\mu}_{\text{PMM}} = \hat{\mu}_{\text{PMM}}(\hat{\beta})$ to reflect its dependence on $\hat{\beta}$, where

$$\begin{aligned}\hat{\mu}_{\text{PMM}}(\beta) &= \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}\{\delta_i y_i + (1-\delta_i)y_{\nu(i)}\}\\ &= \frac{1}{N}\left(\sum_{i\in A}\frac{1}{\pi_i}\delta_i y_i + \sum_{j\in A}\frac{1-\delta_j}{\pi_j}\sum_{i\in A}\delta_i d_{ij} y_i\right) = \frac{1}{N}\sum_{i\in A}\frac{\delta_i}{\pi_i}(1+k_{\beta,i})y_i,\end{aligned} \tag{8}$$

with

$$k_{\beta,i} = \sum_{j\in A} \frac{\pi_i}{\pi_j}(1-\delta_j)d_{ij}. \tag{9}$$

We first consider the case when β^* and hence $m(x_i) = m(x_i;\beta^*)$, denoted by m_i for abbreviation, are known. Suppose that the superpopulation model satisfies the following assumption.

Assumption 4. (i) The matching variable $m(x)$ has a compact and convex support, with its density bounded and bounded away from zero. Let $g_1(m_i)$ and $g_0(m_i)$ be the conditional density of m_i given $\delta_i = 1$ and $\delta_i = 0$, respectively. There exist constants C_{1L} and C_{1U} such that $C_{1L} \leq g_1(m_i)/g_0(m_i) \leq C_{1U}$ almost surely; and (ii) there exists $\delta > 0$ such that $E(|y_i|^{2+\delta}|x_i)$ is uniformly bounded for any x_i.

Assumption 4 (i) is a convenient regularity condition for ease of notation (Abadie & Imbens, 2006). Assumption 4 (ii) is a moment condition for establishing the central limit theorem. For discrete x, Assumption 4 (i) does not hold; however, our discussion below still applies. In this case, when the predictive mean function includes all discrete x and their interactions, units with the same x will have the same predictive mean and will be matched. In the presence of ties, we can choose one of the matched units at random. This is similar to random hot deck imputation with imputation cells defined through different levels of x.

We write

$$n^{1/2}\{\hat{\mu}_{\text{ PMM}}(\beta) - \mu\} = D_N(\beta) + B_N(\beta), \tag{10}$$

where

$$D_N(\beta) = \frac{n^{1/2}}{N}\left(\sum_{i\in A}\frac{1}{\pi_i}[m(x_i;\beta) + \delta_i(1+k_{\beta,i})\{y_i - m(x_i;\beta\}] - \mu\right), \tag{11}$$

and

$$B_N(\beta) = \frac{n^{1/2}}{N}\sum_{i\in A}\frac{1}{\pi_i}(1-\delta_i)\{m(x_{\nu(i)};\beta) - m(x_i;\beta)\}. \tag{12}$$

As in Equation (2), the difference $m(x_{\nu(i)};\beta^*) - m(x_i;\beta^*)$ in Equation (12) accounts for the matching discrepancy, and $B_N(\beta^*)$ contributes to the asymptotic bias of the matching estimator. For predictive mean matching, the matching variable is a scalar function $m(x)$, and hence the bias is $B_N(\beta^*) = O_p(n^{-1/2}) = o_p(1)$ and therefore is negligible. We establish the asymptotic distribution of $\hat{\mu}_{\text{PMM}}(\beta^*)$ as follows.

Theorem 1. *Under Assumptions 1,* 3, *and* 4, *suppose that* $m(x) = E(y|x) = m(x;\beta^*)$ *and* $\sigma^2(x) = \text{var}(y|x)$. *Then,* $n^{1/2}\{\hat{\mu}_{\text{PMM}}(\beta^*) - \mu\} \to \mathcal{N}(0, V_1)$ *in distribution, as* $n \to \infty$, *where*

$$V_1 = V^m + V^e \tag{13}$$

with

$$V^m = \lim_{n\to\infty} \frac{n}{N^2} E\left[\mathrm{var}_p\left\{\sum_{i\in A} \pi_i^{-1} m(x_i)\right\}\right],$$

$$V^e = \lim_{n\to\infty} \frac{n}{N^2} E\left\{\sum_{i=1}^{N} \pi_i^{-1}(1-\pi_i)\delta_i(1+k_{\beta^*,i})^2\sigma^2(x_i)\right\},$$

*and $k_{\beta,i}$ is defined in Equation (*9*).*

In Equation (13), we follow the same approach of Fay (1991) and Shao and Steel (1999) for expressing the total variance by assuming that we first have a census with nonrespondents, and then a sample is taken from the census.

Remark 1. The variance Equation (13) is based on the assumption that the sampling fraction is negligible in the sense that $n/N = o(1)$. Extension to the setting with nonnegligible sampling fraction is possible at the expense of heavier notation. For example, as we show in Appendix, the asymptotic variance Equation (13) has an extra term of order $O(n/N)$, $nN^{-2}\sum_{i=1}^{N} E[\{\delta_i(1+\kappa_{\beta^*,i})-1\}^2\sigma^2(x_i)]$, which cannot be ignored unless n/N is negligible.

In practice, β^* is unknown and therefore has to be estimated prior to matching. Following Abadie and Imbens (2016), the following theorem presents the approximate distribution of $\hat{\mu}_{\mathrm{PMM}}(\hat{\beta})$.

Theorem 2. *Under Assumptions 1,* 3, 4, *and regularity conditions specified in Appendix, the approximate distribution of $n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}(\hat{\beta}) - \mu\}$ is $\mathcal{N}(0, V_2)$, as $n \to \infty$, where $\hat{\beta}$ is the solution to Equation (*6*) and*

$$V_2 = V_1 - \gamma_1^{\mathrm{T}} V_s^{-1}\gamma_1 + \gamma_2^{\mathrm{T}}(\tau_{\beta^*}^{-1} V_s \tau_{\beta^*}^{-1})\gamma_2, \tag{14}$$

*V_1 is defined in Equation (*13*), $V_s = \mathrm{var}\{S_N(\beta^*)\}$, $\gamma_1 = \lim_{n\to\infty} nN^{-2}E\{\sum_{i=1}^{N}\pi_i^{-1}(1-\pi_i)\delta_i(1+k_{\beta^*,i})g(x_i;\beta^*)\sigma^2(x_i)\}$, $\gamma_2 = E\{\dot{m}(x;\beta^*)\}$, and $\tau_\beta = E\{p(x)\dot{m}(x;\beta)\,\dot{m}(x;\beta)^{\mathrm{T}}\}$.*

Comparing the asymptotic variances of the predictive mean matching estimator in Theorems 1 and 2, the difference between V_2 and V_1, $-\gamma_1^{\mathrm{T}} V_s^{-1}\gamma_1 + \gamma_2^{\mathrm{T}}(\tau_{\beta^*}^{-1} V_s \tau_{\beta^*}^{-1})\gamma_2$, can be either positive, zero, or negative. Thus, the estimation error in the predictive mean function should not be ignored. This is different from the result in Abadie and Imbens (2016) that matching on the estimated propensity score always improves the estimation efficiency when matching on the true propensity score. To explain the difference, we note that the propensity score is not pertinent to the true population mean of outcome; whereas the predictive mean function is, which is reflected through the dependence of μ on β^* and a nondegenerate γ_2. It is worth discussing the two variance modification terms. The variance reduction term $-\gamma_1^{\mathrm{T}} V_s^{-1}\gamma_1$ is due to the projection of the matching estimator onto the space spanned by the score function of β^*. Therefore, if y is strongly associated with the score function of β^*, then the efficiency gain by estimating β^* instead of using the true β^* is large. On the other hand, if the predictive mean function changes quickly as β changes in the sense that it has a large derivative with respect to β, $\dot{m}(x;\beta^*)$, then matching on $m(x;\hat{\beta})$ will contribute a large increase of variance, $\gamma_2^{\mathrm{T}}(\tau_{\beta^*}^{-1} V_s \tau_{\beta^*}^{-1})\gamma_2$, compared with matching on $m(x;\beta^*)$.

3.1 | Robustness for the predictive mean function specification

To discuss the robustness of the predictive mean matching estimator with respect to specification for the predictive mean function, let $m(x; \beta)$ be a working model for $E(y|x)$, $\hat{\beta}$ be the estimator of β solving Equation (6), and β^* be its probability limit. As a slight abuse of notation, we also use $m = m(x; \beta^*)$ for shorthand. We require the following assumption hold for the working model.

Assumption 5. $E(y|m)$ is Lipschitz continuous in m; that is, there exists a constant C_3 such that $|E(y|m_i) - E(y|m_j)| \leq C_4|m_i - m_j|$, for any i, j.

For example, suppose that a scalar y given x follows a generalized linear model with $E(y|x) = g(x^{\mathrm{T}}\beta^*)$, where $g(\cdot)$ is an unknown function. If one assumes the mean function $m = x^{\mathrm{T}}\tilde{\beta}^*$, then $E(y|m) = g(\gamma^* m)$ with $\gamma^* = \beta^{*\mathrm{T}}(\tilde{\beta}^*\tilde{\beta}^{*\mathrm{T}})^{-1}\tilde{\beta}^*$. Thus, Assumption 5 holds if $g(\cdot)$ is differentiable and Assumption 4 holds. It is important to note that we cannot omit variables in x because we require Assumption 1 hold to ensure missingness at random.

Theorem 3. *Under Assumptions 1, 3, 4, and 5, the predictive mean matching estimator based on the working model $m(x; \beta^*)$ is consistent for μ.*

The result can be obtained directly from the decomposition Equation (10) by replacing $m(x; \beta)$ in $D_N(\beta)$ and $B_N(\beta)$ with $E\{y|m(x; \beta)\}$. The new term $D_N(\beta^*)$ is still consistent for zero; by Assumption 5, the new bias term becomes

$$\begin{aligned}|B_N(\beta^*)| &= \left| \frac{n^{1/2}}{N} \sum_{i \in A} \frac{1}{\pi_i}(1 - \delta_i) \left[E\left\{ y|m(x_{\nu(i)}; \beta^*) \right\} - E\left\{ y|m(x_{(i)}; \beta^*) \right\} \right] \right| \\ &\leq \frac{n^{1/2}}{N} C_4 \sum_{i \in A} \frac{1}{\pi_i}(1 - \delta_i) \left| m(x_{\nu(i)}; \beta^*) - m(x_i; \beta^*) \right| = O_p\left(n^{-1/2}\right).\end{aligned}$$

We end this section by summarizing the theoretical findings of predictive mean matching. Predictive mean matching provides a remedy to the curse of dimensionality by summarizing the vector of covariates into a scalar predictive mean function as the matching variable, so that the bias due to matching discrepancy is negligible. The predictive mean function is not necessarily the correct conditional mean function and therefore provides some robustness against model misspecification.

4 | REPLICATION VARIANCE ESTIMATION

In this section, we propose replication variance estimation (Mashreghi, Haziza, & Léger, 2016; Rust & Rao, 1996; Wolter, 2007) for valid inference based on predictive mean matching.

We first describe replication variance estimation for the case when y_i is observed throughout the sample. Let $\hat{\mu} = \sum_{i \in A} \omega_i y_i$ with $\omega_i = (N\pi_i)^{-1}$ be the Horvitz–Thompson estimator of μ. The replication variance estimator of $\hat{\mu}$ takes the form of

$$\hat{V}_{\mathrm{rep}}(\hat{\mu}) = \sum_{k=1}^{L} c_k (\hat{\mu}^{(k)} - \hat{\mu})^2, \tag{15}$$

where L is the number of replicates, c_k is the kth replication factor, and $\hat{\mu}^{(k)}$ is the kth replicate of $\hat{\mu}$. When $\hat{\mu} = \sum_{i \in A} \omega_i y_i$, we can write the replicate of $\hat{\mu}$ as $\hat{\mu}^{(k)} = \sum_{i \in A} \omega_i^{(k)} y_i$ for some $\omega_i^{(k)}$ for $i \in A$. The replications are constructed such that $E\{\hat{V}_{\text{rep}}(\hat{\mu})\} = \text{var}(\hat{\mu})\{1 + o(1)\}$.

We illustrate the replication weights in the following examples.

Example 1. Suppose that Sample A was selected using simple random sampling. For the nonparametric bootstrap, we have $L = B$, which is the number of bootstrap replicates, $c_k = B^{-1}$, $\omega_i^{(k)} = n^{-1} m_i^{(k)}$, where $(m_1^{(k)}, \dots, m_n^{(k)})$ is a multinomial random vector with n draws on n equal probability cells.

Example 2. Suppose that Sample A was selected using probability proportional to size sampling with $\omega_i = N^{-1}\pi_i^{-1}$. For the delete-1 jackknife, we have $L = n$, $c_k = (n-1)/n$, and $\omega_i^{(k)} = n\omega_i/(n-1)$ if $i \neq k$, and $\omega_k^{(k)} = 0$.

We now propose a new replication variance estimation for the predictive mean matching estimator. We first consider $\hat{\mu}_{\text{PMM}}(\beta^*)$ with a known β^* given in Equation (8). For simplicity, we suppress the dependence of quantities on β^*. Let the individual linearized term be

$$\psi_i = m(x_i) + \delta_i(1 + k_i)\{y_i - m(x_i)\}, \tag{16}$$

and the corresponding population quantity and the Horvitz–Thompson estimator be $\mu_\psi = N^{-1}\sum_{i=1}^N \psi_i$ and $\hat{\psi}_{\text{HT}} = \sum_{i \in A} \omega_i \psi_i$, respectively. We can write

$$\hat{\mu}_{\text{PMM}} - \mu = (\hat{\mu}_{\text{PMM}} - \hat{\psi}_{\text{HT}}) + (\hat{\psi}_{\text{HT}} - \mu_\psi) + (\mu_\psi - \mu),$$

where $\mu_{\text{PMM}} - \hat{\psi}_{\text{HT}} = o_p(n^{-1/2})$ by Theorem 1, and $\mu_\psi - \mu = O_p(N^{-1/2})$. Given that the sample fraction is negligible, that is, $nN^{-1} = o(1)$, we have $\hat{\mu}_{\text{PMM}} - \mu = \hat{\psi}_{\text{HT}} - \mu_\psi + o_p(n^{-1/2})$. It is then sufficient to estimate the variance of $\hat{\psi}_{\text{HT}} - \mu_\psi$. Because $E_p(\hat{\psi}_{\text{HT}} - \mu_\psi) = 0$, we have $\text{var}(\hat{\psi}_{\text{HT}} - \mu_\psi) = E\{\text{var}_p(\hat{\psi}_{\text{HT}} - \mu_\psi)\}$, which is essentially the sampling variance of $\hat{\psi}_{\text{HT}}$. This suggests that we can treat $\{\psi_i : i \in A\}$ as pseudo observations in applying the replication variance estimator. It is important to note that k_i is treated as an intrinsic characteristic with unit i in Equation (16). In this way, the distribution of k_i can be preserved, unlike the naive bootstrap methods. Otsu and Rai (2017) used a similar idea to develop a wild bootstrap technique (Wu, 1986) for a matching estimator by resampling the response variable based on the residual values. More specifically, we construct the kth replicate of $\hat{\psi}_{\text{HT}}$ as follows: $\hat{\psi}_{\text{HT}}^{(k)} = \sum_{i \in A} \omega_i^{(k)} \psi_i$, where $\omega_i^{(k)}$ is the replication weight that accounts for complex sampling design. The replication variance estimator of $\hat{\psi}_{\text{HT}}$ is obtained by applying $\hat{V}_{\text{rep}}(\cdot)$ in Equation (15) for the above replicates $\hat{\psi}_{\text{HT}}^{(k)}$. It follows that $E\{\hat{V}_{\text{rep}}(\hat{\psi}_{\text{HT}})\} = \text{var}(\hat{\psi}_{\text{HT}} - \mu_\psi)\{1 + o(1)\} = \text{var}(\hat{\mu}_{\text{PMM}} - \mu)\{1 + o(1)\}$.

We now consider $\hat{\mu}_{\text{PMM}}(\hat{\beta})$, which can be expressed as

$$\hat{\mu}_{\text{PMM}}(\hat{\beta}) = \sum_{i \in A} \omega_i[m(x_i; \hat{\beta}) + \delta_i(1 + k_{\hat{\beta},i})\{y_i - m(x_i; \hat{\beta})\}] + o_p(n^{-1/2}).$$

Motivated by the two-step procedure for the predictive mean matching estimator, we propose a parallel two-step procedure to construct the replicates of $\hat{\mu}_{\text{PMM}}(\hat{\beta})$:

Step 1. Obtain the kth replicate of $\hat{\beta}$, denoted as $\hat{\beta}^{(k)}$, by solving

$$S_N^{(k)}(\beta) = n^{1/2} \sum_{i \in A} \omega_i^{(k)} \delta_i g(x_i; \beta)\{y_i - m(x_i; \beta)\} = 0.$$

Step 2. Obtain the kth replicate as

$$\hat{\mu}_{\mathrm{PMM}}^{(k)}(\hat{\beta}^{(k)}) = \sum_{i \in A} \omega_i^{(k)} [m(x_i; \hat{\beta}^{(k)}) + \delta_i (1 + k_{\overline{\beta}^*, i})\{y_i - m(x_i; \hat{\beta}^{(k)})\}], \tag{17}$$

where $\overline{\beta}^* = L^{-1} \sum_{l=1}^{L} \hat{\beta}^{(l)}$.

If β^* were known, we would not need to replicate the estimation for β^*, and the above procedure would reduce to Step 2 only. As β^* is estimated, Step 1 is necessary, because as shown in Theorem 2, the predictive mean matching estimators by matching on the true and estimated predictive mean function may have different asymptotic distributions.

To show the consistency of the proposed replication variance estimator, we cannot apply the usual linearization technique (Kim & Rao, 2009) due to lack of smoothness. Utilizing the parallelism between the replication procedure and the predictive mean matching procedure, we demonstrate the consistency by extending the technique of Andreou and Werker (2012) to the replication process. See Appendix for details. The consistency of the replication variance estimator is presented in the following theorem.

Theorem 4. *Under the assumptions in Theorem 2, for the Horvitz–Thompson estimator $\hat{\mu}$, suppose that $\hat{V}_{\mathrm{rep}}(\hat{\mu})$ in Equation (15) is consistent for $\mathrm{var}_p(\hat{\mu})$. Then, the replication variance estimators for $\hat{\mu}_{\mathrm{PMM}}(\hat{\beta})$ is consistent, that is, $n\hat{V}_{\mathrm{rep}}\{\hat{\mu}_{\mathrm{PMM}}(\hat{\beta})\}/V_2 \to 1$ in probability, as $n \to \infty$, where the replicates of $\hat{\mu}_{\mathrm{PMM}}(\hat{\beta})$ are given in Equation (17), and V_2 is given in Equation (14).*

5 | EMPIRICAL STUDIES

5.1 | A simulation study

In this simulation study, we investigate the performance of the proposed replication variance estimator. For generating finite populations of size $N = 50,000$: first, let x_{1i}, x_{2i}, and x_{3i} be generated independently from Uniform$[0, 1]$, and x_{4i}, x_{5i}, x_{6i}, and e_i be generated independently from $\mathcal{N}(0, 1)$; then, let y_i be generated from (P1) $y_i = -1 + x_{1i} + x_{2i} + e_i$, (P2) $y_i = -1.167 + x_{1i} + x_{2i} + (x_{1i} - 0.5)^2 + (x_{2i} - 0.5)^2 + e_i$, and (P3) $y_i = -1.5 + x_{1i} + \ldots + x_{6i} + e_i$. The parameter of interest is $\mu = N^{-1} \sum_{i=1}^{N} y_i$. The covariates are fully observed, but y_i is not. The response indicator of y_i, δ_i, is generated from Bernoulli(p_i) with $\mathrm{logit}\{p(x_i)\} = 0.2 + x_{1i} + x_{2i}$. This results in the average response rate about 75%. To generate samples, we consider two sampling designs: (S1) simple random sampling with $n = 400$; (S2) probability proportional to size sampling. In (S2), for each unit in the population, we generate a size variable s_i as $\log(|y_i + v_i| + 4)$, where $v_i \sim \mathcal{N}(0, 1)$. The selection probability is specified as $\pi_i = 400 s_i / \sum_{i=1}^{N} s_i$. Therefore, (S2) is informative, where units with larger y_i values have larger probabilities to be selected into the sample.

For estimation, we consider predictive mean matching imputation, nearest neighbor imputation, and stochastic regression imputation. In stochastic regression imputation, for units with $\delta_i = 0$, the imputation of y_i is obtained as $y_i^* = \hat{y}_i + \hat{e}_i^*$, where $\hat{y}_i = m(x_i; \hat{\beta})$ and $\hat{e}_i^*$ is randomly selected from the observed residuals $\{\hat{e}_i = y_i - \hat{y}_i : \delta_i = 1\}$. For (P1) and (P2), we specify the predictive mean function to be $m(x;\beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Note that for (P1), $m(x;\beta)$ is correctly specified; whereas for (P2), $m(x;\beta)$ is misspecified. For (P3), we specify the mean function to be $m(x;\beta) = \beta_0 + \beta^{\mathrm{T}} x$, where $x = (x_1, \dots, x_6)$. We construct 95% confidence intervals (CIs) using $(\hat{\mu}_I - z_{0.975}\hat{V}_I^{1/2}, \hat{\mu}_I + z_{0.975}\hat{V}_I^{1/2})$, where $\hat{\mu}_I$ is the point estimate and $\hat{V}_I$ is the variance estimate obtained by the proposed jackknife variance estimation. For stochastic regression imputation, the kth replicate of μ is given by $\hat{\mu}^{(k)}_{\mathrm{REG}}(\hat{\beta}^{(k)}) = \sum_{i\in A}\omega_i^{(k)}[m(x_i;\hat{\beta}^{(k)}) + \delta_i(1+k_i)\{y_i - m(x_i;\hat{\beta}^{(k)})\}]$, where $\hat{\beta}^{(k)}$ is obtained from the estimating equation of β based on the replication weights, and k_i is the number of times that $\hat{e}_i$ is selected to impute the missing values of y based on the original data.

Table 1 presents the simulation results based on 2,000 Monte Carlo samples. When the covariate is two dimensional, all three imputation estimators have small biases, even when the mean function is misspecified. In addition, the proposed jackknife method provides valid coverage of CIs for the predictive mean matching and stochastic regression imputation estimators in all scenarios. This suggests that the proposed replication method can be used widely even for stochastic regression imputation. When the covariate is six dimensional, nearest neighbor imputation presents large biases and low coverage rates. This is consistent with Equation (3).

5.2 | Real-life data analysis

To illustrate the methods proposed, we analyze data from the API program in California (http://api.cde.ca.gov/). The API is computed for all California schools based on standardized testing of students. The full population data consist of $N = 6,194$ observations for all schools with at least 100 students on various academic measures including the API for years 2000 and 1999 (api99, x_1, and api00, y), percentage of students eligible for subsidized meals (meals, x_2), percentage of English language learners (ell, x_3), average parental education level (avg.ed, x_4), percentage of fully qualified teachers (full, x_5), and number of students enrolled (enroll, x_6). We are interested in the population average of the API, $\mu = N^{-1}\sum_{i=1}^{N} y_i$, in California in year 2000. The original data have full observations on the API, and therefore the population parameter of interest can be

TABLE 1 Simulation results: Bias (×10^2) and *SE* (×10^2) of the point estimator, relative bias (RB) of jackknife variance estimates (×10^2) and coverage rate (CR%) of 95% confidence intervals

	PMM		NNI		SRI		PMM		NNI		SRI	
	Bias	***SE***	**Bias**	***SE***	**Bias**	***SE***	**RB**	**CR**	**RB**	**CR**	**RB**	**CR**
	(S1) Simple random sampling											
(P1)	−0.15	6.46	−0.21	6.54	−0.23	6.44	4	95.2	3	95.1	5	95.8
(P2)	−0.22	6.54	−0.25	6.55	−0.37	6.46	6	95.5	3	95.3	5	95.6
(P3)	1.90	11.85	18.59	11.06	0.11	11.17	5	95.1	4	63.8	4	95.5
	(S2) Probability proportional to size sampling											
(P1)	0.05	6.46	0.13	6.37	0.18	6.53	3	95.3	3	94.8	2	94.9
(P2)	0.30	6.52	0.12	6.47	0.16	6.60	2	95.3	0	95.3	3	94.9
(P3)	1.33	10.99	17.53	10.70	0.40	11.10	6	95.6	3	65.5	−3	95.6

Abbreviations: NNI, nearest neighbor imputation; PMM, predictive mean matching; SRI, stochastic regression imputation.

TABLE 2 Bias and *SE* ($\times 10^2$) of the point estimator, and coverage rate (CR%) of 95% confidence intervals for the population mean of the API with the true value $\mu = 664.7$

Method	Bias	*SE*	CR	Bias	*SE*	CR
	(P1) MCAR & (S1) SRS			(P1) MCAR & (S2) PPS		
CC	0.37	16.63	94.30	−0.29	16.08	95.10
PMM	0.49	13.06	94.95	0.27	12.14	95.30
	(P2) MAR & (S1) SRS			(P2) MAR & (S2) PPS		
CC	43.66	15.01	17.75	43.34	14.97	19.55
PMM	1.48	13.31	94.70	1.04	12.40	95.45

Abbreviations: CC, complete case analysis; MAR, missingness at random; MCAR, missingness completely at random; PMM, predictive mean matching; PPS, probability proportional to size sampling; SRS, simple random sampling.

calculated with the true value $\mu = 664.7$. So, the API is uniquely placed for demonstration of the proposed methods.

In the population, we create artificial missingness. The response indicator of y_i, δ_i, is generated from Bernoulli(p_i) with (P1) $p_i \equiv 0.65$, representing missingness completely at random (Rubin, 1976), and (P2) $\text{logit}\{p(x_i)\} = 1 + 2x_{1i} + x_{2i} + x_{3i} + x_{4i} + x_{5i} + x_{6i}$, representing missingness at random, where all covariates are standardized with mean 0 and SD 1. This results in the average response rate 65%. We generate samples under two sampling designs: (S1) simple random sampling with $n = 200$; (S2) probability proportional to size sampling which is the same as in the simulation study in Section 5.1.

For estimation, we consider complete case estimation and predictive mean matching imputation. The complete case estimator is simply the weighted average of observed y_i's weighted by the sample weights. For the predictive mean matching estimator, we specify the mean function to be $m(x; \beta) = \beta_0 + \beta^{\mathrm{T}} x$, where $x = (x_1, \dots, x_6)'$. We construct 95% CIs using $(\hat{\mu}_I - z_{0.975}\hat{V}_I^{1/2}, \hat{\mu}_I + z_{0.975}\hat{V}_I^{1/2})$, where $\hat{\mu}_I$ is the point estimate and $\hat{V}_I$ is the variance estimate obtained by the proposed jackknife variance estimation.

Table 2 summarizes the results based on 2,000 Monte Carlo samples. Under (P1) with missingness completely at random, both the complete case method and the predictive mean matching method have small biases and coverage rates close to the 95% nominal coverage. Moreover, the predictive mean matching method is more efficient than the complete case method with smaller standard errors. Under (P2) with missingness at random, the complete case method has large biases and low converge rates; whereas the predictive mean matching inference still has small biases and good coverage rates.

6 | DISCUSSION

Predictive mean matching is used widely to impute missing values to facilitate full-sample analysis. Variance estimation for predictive mean matching has been an important research gap in survey sampling. We addressed this problem by proposing a simple two-step replication procedure, which can faithfully reflect variability of the predictive mean matching estimator. In this article, we assumed that the sampling fraction is negligible (see Assumption 3, ii). Following Shao and Steel (1999) and Mashreghi, Léger, and Haziza (2014), we will extend the variance estimation to handle the case of nonnegligible sampling fractions.

We focused on inference of the population mean using predictive mean imputation, a hot-deck type of imputation. The superiority of the hot deck imputation methods over the mean, ratio, and regression imputation methods is that the hot-deck imputation methods provide asymptotically valid distribution and quantile estimators (Andridge & Little, 2010; Chen & Shao, 2000). Yang and Kim (2019) considered nearest neighbor imputation for the finite population parameter defined through $\mu_g = N^{-1}\sum_{i=1}^{N} g(y_i)$ for some known $g(\cdot)$, or $\xi_N = \inf\{\xi : S_N(\xi) \geq 0\}$, where $S_N(\xi) = N^{-1}\sum_{i=1}^{N} s(y_i - \xi)$, and $s(\cdot)$ is a univariate real function. For example, let $g(y) = y$, $\mu_g = N^{-1}\sum_{i=1}^{N} y_i$ is the population mean of y. Let $g(y) = I(y < c)$ 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. Let $s(y_i - \xi) = I(y_i - \xi \leq 0) - \alpha$, ξ_N is the population αth quantile. Extending the current framework using predictive mean matching to general parameter estimation is feasible and will be pursued elsewhere.

Propensity score matching has been recently proposed for inferring causal effects of treatments in the context of survey data; however, their asymptotic properties are underdeveloped (Lenis, Nguyen, Dong, & Stuart, 2017). Because causal inference is inherently a missing data problem (Ding & Li, 2018), the proposed methodology here can be easily generalized to investigate the asymptotic properties of propensity score matching estimators with survey weights.

ACKNOWLEDGEMENTS

The authors are grateful to anonymous referees and the AE for very constructive comments. Dr. Yang is partially supported by *the National Science Foundation grant DMS 1811245, National Cancer Institute grant P01 CA142538, and Oak Ridge Associated Universities*. Dr. Kim is partially supported by *the US National Science Foundation grant MMS-1733572*.

ORCID

Jae Kwang Kim https://orcid.org/0000-0002-0246-6029

REFERENCES

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

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, *76*, 1537–1557.

Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, *29*, 1–11.

Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, *84*, 781–807.

Andreou, E., & Werker, B. J. (2012). An alternative asymptotic analysis of residual-based statistics. *The Review of Economics and Statistics*, *94*, 88–99.

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*, 40–64.

Beaumont, J.-F., & Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *The Canadian Journal of Statistics*, *37*, 400–416.

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

Bickel, P. J., Klaassen, C., Ritov, Y., & Wellner, J. (1993). *Efficient and adaptive inference in semiparametric models*. Baltimore, Maryland: Johns Hopkins University Press.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York, NY: Wiley.

Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, *16*, 113–131.

Chen, J., & Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, *96*, 260–269.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, *6*, 5549–5632.

Ding, P., & Li, F. (2018). Causal inference: A missing data perspective. *Statistical Science*, *33*, 214–237.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Fay, R. (1991). *A design-based perspective on missing data variance*. In *Proceedings of the 1991 Annual Research Conference* (pp. 429–440). Washington, DC: *US Bureau of the Census*.

Ford, B. L. (1983). An overview of hot-deck procedures. *Incomplete Data in Sample Surveys*, *2*(Part IV), 185–207.

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

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*, 1017–1098.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, *64*, 605–654.

Heitjan, D. F., & Little, R. J. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics*, *40*, 13–29.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, *55*, 244–254.

Kim, J. K., Fuller, W. A., & Bell, W. R. (2011). Variance estimation for nearest neighbor imputation for US Census long form data. *The Annals of Applied Statistics*, *5*, 824–842.

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

Kim, J. K., & Rao, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, *96*, 917–932.

Le Cam, L., & Yang, G. L. (1990). *Asymptotics in statistics: Some basic concepts*. Berlin, Germany: Springer.

Lenis, D., Nguyen, T. Q., Dong, N., & Stuart, E. A. (2017). It's all about balance: Propensity score matching in the context of complex survey data. *Biostatistics*, *20*, 147–163.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*, 287–296.

Mashreghi, Z., Haziza, D., & Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, *10*, 1–52.

Mashreghi, Z., Léger, C., & Haziza, D. (2014). Bootstrap methods for imputed data from regression, ratio and hot-deck imputation. *The Canadian Journal of Statistics*, *42*, 142–167.

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*, 75.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, *79*, 147–168.

Otsu, T., & Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, *112*, 1720–1732.

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

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*, 87–94.

Rust, K. F., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, *5*, 283–310.

Schenker, N., & Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, *22*, 425–446.

Shao, J., & Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, *94*, 254–265.

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

Vink, G., Frank, L. E., Pannekoek, J., & Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, *68*, 61–90.

Wolter, K. (2007). *Introduction to variance estimation* (2nd ed.). New York, NY: Springer.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, *14*, 1261–1295.

Yang, S., & Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, *105*, 487–493.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., & Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, *72*, 1055–1065.

Yang, S., & Kim, J. K. (2019). *Nearest neighbor imputation for general parameter estimation in survey sampling*. In K. P. Huynh, D. T. Jacho-Chavez, & G. Tripathi (Eds.), *The econometrics of complex survey data: Theory and applications* (pp. 209–234). Bingley, West Yorkshire, England: Emerald Publishing Limited.

How to cite this article: Yang S, Kim JK. Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. *Scand J Statist*. 2020;47:839–861. https://doi.org/10.1111/sjos.12429

APPENDIX

A1 Consistency of $\hat{\beta}$

We provide regularity conditions and proof for the consistency of $\hat{\beta}$ solving Equation (6).

Assumption A1 (Uniform weak convergence). $\sup_{\beta\in\mathcal{B}}|n^{-1/2}S_N(\beta)-S(\beta)|\to 0$ in probability for some nonstochastic function $S(\beta)$, as $n\to\infty$, where $\mathcal{B}$ is a closed set.

Assumption A1 is a high-level condition. Sufficient conditions for Assumption A1 include Assumption 3 for sampling and assumption that $m(x;\beta)$ and $g(x;\beta)$ are continuous for $\beta\in\mathcal{B}$.

Assumption A2 (Identifiability). $S(\beta)=0$ has a unique solution β^*. That is, for any $\epsilon>0$, there exists a $\delta>0$ such that $\beta\notin\mathcal{B}_\epsilon(\beta^*)$ implies $|S(\beta)|\geq\delta$, where $\mathcal{B}_\epsilon(\beta^*)=\{\beta\in\mathcal{B}:||\beta-\beta^*||<\epsilon\}$.

To show that $\hat{\beta}$ solving Equation (6) is consistent for β^*, for any $\epsilon>0$, we find $\delta>0$ such that

$$\begin{aligned}0\leq P\{\hat{\beta}\notin\mathcal{B}_\epsilon(\beta^*)\} &\leq P\{|S(\hat{\beta})-S(\beta^*)|\geq\delta\}\\ &= P\{|S(\hat{\beta})-n^{-1/2}S_N(\hat{\beta})+n^{-1/2}S_N(\hat{\beta})-S(\beta^*)|\geq\delta\}\\ &= P\{|S(\hat{\beta})-n^{-1/2}S_N(\hat{\beta})|\geq\delta\}\\ &\leq P\left\{\sup_{\beta\in\mathcal{B}}|S(\hat{\beta})-n^{-1/2}S_N(\hat{\beta})|\geq\delta\right\}\to 0,\end{aligned}$$

as $n\to\infty$.

A2 Proof for Theorem 1

Based on the decomposition in Equation (10), we write

$$n^{1/2}\{\hat{\mu}_{\text{PMM}}(\beta^*)-\mu\}=D_N(\beta^*)+B_N(\beta^*), \tag{A1}$$

where $D_N(\beta)$ and $B_N(\beta)$ are defined in Equations (11) and (12), respectively. For simplicity, we introduce the following notation: $m_i=m(x_i;\beta^*)$ and $e_i=y_i-m_i$.

Under Assumption 4, for the predictive mean matching estimator, $m_{\nu(i)}-m_i=O_p(1)$. Together with Assumption 3, we evaluate the order of $B_N(\beta^*)$ as

$$B_N(\beta^*)=\frac{n^{1/2}}{N}\sum_{i\in A}\frac{1}{\pi_i}(1-\delta_i)(m_{\nu(i)}-m_i)=O_p(n^{-1/2})=o_p(1).$$

Therefore, Equation (A1) reduces to $n^{1/2}\{\hat{\mu}_{\text{PMM}}(\beta^*)-\mu\}=D_N(\beta^*)+o_p(1)$. Then, to study the asymptotic properties of $n^{1/2}\{\hat{\mu}_{\text{PMM}}(\beta^*)-\mu\}$, we only need to study the asymptotic properties of $D_N(\beta^*)$. We express

$$
\begin{aligned}
D_N(\beta^*) &= \frac{n^{1/2}}{N}\left[\sum_{i\in A}\frac{1}{\pi_i}\{m_i+\delta_i(1+k_{\beta^*,i})e_i\}-\mu\right]\\
&= \frac{n^{1/2}}{N}\sum_{i=1}^{N}\left(\frac{I_i}{\pi_i}-1\right)m_i+\frac{n^{1/2}}{N}\sum_{i=1}^{N}\left(\frac{I_i}{\pi_i}-1\right)\delta_i(1+k_{\beta^*,i})e_i\\
&\quad+\frac{n^{1/2}}{N}\sum_{i=1}^{N}(m_i-\mu)+\frac{n^{1/2}}{N}\sum_{i=1}^{N}\delta_i(1+k_{\beta^*,i})e_i \qquad \text{(A2)}\\
&= \frac{n^{1/2}}{N}\sum_{i=1}^{N}\left(\frac{I_i}{\pi_i}-1\right)m_i+\frac{n^{1/2}}{N}\sum_{i=1}^{N}\left(\frac{I_i}{\pi_i}-1\right)\delta_i(1+k_{\beta^*,i})e_i+o_p(1), \qquad \text{(A3)}
\end{aligned}
$$

given that $nN^{-1}=o(1)$. Using the conditioning argument, we can verify that the covariance of the two terms in Equation (A3) is zero. Thus, the asymptotic variance of $D_N(\beta^*)$ is

$$
\operatorname{var}\left\{\frac{n^{1/2}}{N}\sum_{i=1}^{N}\left(\frac{I_i}{\pi_i}-1\right)m_i\right\}+\operatorname{var}\left\{\frac{n^{1/2}}{N}\sum_{i=1}^{N}\left(\frac{I_i}{\pi_i}-1\right)\delta_i(1+k_{\beta^*,i})e_i\right\}.
$$

The first term, as $n\to\infty$, becomes

$$
V^m=\lim_{n\to\infty}\frac{n}{N^2}E\left\{\operatorname{var}_p\left(\sum_{i\in A}\frac{m_i}{\pi_i}\right)\right\},
$$

and the second term, as $n\to\infty$, becomes

$$
V^e=\operatorname{plim}\frac{n}{N^2}\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}\delta_i(1+k_{\beta^*,i})^2\operatorname{var}(e_i|x_i).
$$

The remaining is to show that $V^e=O(1)$. To do this, the key is to show that the moments of $k_{\beta^*,i}$ are bounded. Under Assumption 4, some algebra yields

$$
\underline{\omega}\tilde{k}_{\beta^*,i}\le k_{\beta^*,i}\le\overline{\omega}\tilde{k}_{\beta^*,i}, \qquad \text{(A4)}
$$

for some constants $\underline{\omega}$ and $\overline{\omega}$, where $\tilde{k}_{\beta^*,i}=\sum_{j=1}^{n}(1-\delta_j)d_{ij}$ is the number of unit i used as a match for the nonrespondents. Under Assumption 4, $\tilde{k}_{\beta^*,i}=O_p(1)$ and $E(\tilde{k}_{\beta^*,i})$ and $E(\tilde{k}^2_{\beta^*,i})$ are uniformly bounded over n (Abadie & Imbens, 2006, lemma 3); therefore, together with Equation (A4), we have $k_{\beta^*,i}=O_p(1)$ and $E(k_{\beta^*,i})$ and $E(k^2_{\beta^*,i})$ are uniformly bounded over n. Therefore, simple algebra yields $V^e=O(1)$.

Combining all results, the asymptotic variance of $n^{1/2}\{\hat{\mu}_{\text{PMM}}(\beta^*)-\mu\}$ is V^m+V^e. By the central limit theorem, the result in Theorem 1 follows.

A3 Proof for Remark 1

If the sampling fraction is asymptotically nonnegligible, the asymptotic variance of the terms in Equation (A2) is

$$
\operatorname{var}\left[\frac{n^{1/2}}{N}\sum_{i=1}^{N}\{\delta_i(1+k_{\beta^*,i})-1\}e_i\right]=nN^{-2}\sum_{i=1}^{N}E[\{\delta_i(1+\kappa_{\beta^*,i})-1\}^2\sigma^2(x_i)]=O(n/N).
$$

Because $I_i/\pi_i - 1$ given $\mathcal{F}_N$ is design unbiased, by the conditioning argument, the covariance of $n^{1/2}N^{-1}\sum_{i=1}^{N}\{\delta_i(1+k_{\beta^*,i})-1\}e_i$ and the other terms in Equation (A3) is zero. This completes the proof for the statement in Remark 1.

A4 Le Cam's third Lemma

Consider two sequences of probability measures $(Q^{(N)})_{N=1}^{\infty}$ and $(P^{(N)})_{N=1}^{\infty}$. Assume that under $P^{(N)}$, a statistic T_N and the likelihood ratios $\mathrm{d}Q^{(N)}/\mathrm{d}P^{(N)}$ satisfy

$$\begin{pmatrix} T_N \\ \log(\mathrm{d}Q^{(N)}/\mathrm{d}P^{(N)}) \end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix} 0 \\ -\sigma^2/2 \end{pmatrix}, \begin{pmatrix} \tau^2 & c \\ c & \sigma^2 \end{pmatrix}\right\}$$

in distribution, as $N \to \infty$. Then, under $Q^{(N)}$, $T_N \to \mathcal{N}(c, \tau^2)$ in distribution, as $N \to \infty$. See Le Cam and Yang (1990), Bickel, Klaassen, Ritov, and Wellner (1993), and van der Vaart (2000) for textbook discussions.

A5 Proof for Theorem 2

Let P be the distribution of $(x_i, y_i, \delta_i, I_i)$, for $i = 1, \ldots, N$, induced by the marginal distribution of x_i, the conditional distribution of y_i given x_i, the conditional distribution of δ_i given (x_i, y_i), and the conditional distribution of I_i given (x_i, y_i, δ_i). Consider P to be restricted by the moment condition through the predictive mean function Equation (5) with the true parameter value β^*. The consistent estimator $\hat{\beta}$ is the solution to the normalized estimating equation

$$S_N(\beta) = \frac{n^{1/2}}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\delta_i g(x_i;\beta)\{y_i - m(x_i;\beta)\} = 0. \tag{A5}$$

To discuss the asymptotic properties of $\hat{\mu}_{\mathrm{PMM}}(\hat{\beta})$, we rely on Le Cam's third lemma and consider a parametric model P^{β} defined locally around β^* with a density

$$\frac{\exp\{n^{1/2}(\beta-\beta^*)^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}S_N(\beta^*) - 2^{-1}n(\beta-\beta^*)^{\mathrm{T}}\Lambda^{-1}(\beta-\beta^*)\}}{E[\exp\{n^{1/2}(\beta-\beta^*)^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}S_N(\beta^*) - 2^{-1}n(\beta-\beta^*)^{\mathrm{T}}\Lambda^{-1}(\beta-\beta^*)\}]}. \tag{A6}$$

Because under P^{β^*}, $S_N(\beta^*) \to \mathcal{N}(0, V_s)$ in distribution, the normalizing constant in the denominator converges to 1 as $n \to \infty$. The Fisher information under the parametric model Equation (A6) is $n\Lambda^{-1}$. Therefore, $\hat{\beta}$ is efficient under Equation (A6).

We now consider sequences that are local to β^*, $\beta_N = \beta^* + n^{-1/2}h$, indexed by N. In our context, we have the population size N goes to infinity with sample size n. Consider $(x_i, y_i, \delta_i, I_i)$, for $i = 1, \ldots, N$, with the local shift P^{β_N} (Bickel et al., 1993). We make the following regularity assumption:

Assumption A3. (i) The superpopulation model is regular (Bickel et al., 1993, pp. 12–13); (ii) under P^{β_N}: $S_N(\beta_N) \to \mathcal{N}(0, V_s)$ in distribution, as $n \to \infty$; (iii) τ_β is nonsingular around β^*, and $n^{1/2}(\hat{\beta} - \beta_N) = \tau_{\beta^*}^{-1}S_N(\beta_N) + o_p(1)$; (iv) for all bounded continuous functions $h(x, y, \delta, I)$, the conditional expectation $E_{\beta_N}\{h(x, y, \delta, I) \mid x, \delta = 1\}$ converges in distribution to $E\{h(x, y, \delta, I) \mid x, \delta = 1\}$, where E_{β_N} denotes the expectation taken with respect to P^{β_N}.

We now sketch a heuristic proof for Theorem 2.

Under Equation (A6), the likelihood ratio under P^{β_N} is

$$\begin{aligned}\log(\mathrm{d}P^{\beta^*}/\mathrm{d}P^{\beta_N}) &= -h^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}S_N(\beta^*) + \frac{1}{2}h^{\mathrm{T}}\Lambda^{-1}h + o_p(1)\\ &= -h^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}S_N(\beta_N) - \frac{1}{2}h^{\mathrm{T}}\Lambda^{-1}h + o_p(1),\end{aligned}$$

where the second equality follows by the Taylor expansion of $S_N(\beta^*)$ at β_N.

We can derive that under P^{β_N},

$$\begin{pmatrix} n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}(\beta_N) - \mu(\beta_N)\} \\ n^{1/2}(\hat{\beta} - \beta_N) \\ \log(dP^{\beta^*}/dP^{\beta_N}) \end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix} 0 \\ 0 \\ \frac{-1}{2}h^{\mathrm{T}}\Lambda^{-1}h \end{pmatrix}, \begin{pmatrix} V_1 & \gamma_1^{\mathrm{T}}\tau_{\beta^*}^{-1} & -\gamma_1^{\mathrm{T}}V_s^{-1}\tau_{\beta^*}h \\ \tau_{\beta^*}^{-1}\gamma_1 & \Lambda & -h \\ -h^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}\gamma_1 & -h^{\mathrm{T}} & h^{\mathrm{T}}\Lambda^{-1}h \end{pmatrix}\right\} \quad \text{(A7)}$$

in distribution, as $n \to \infty$. Herein, we write $\mu = \mu(\beta_N)$ to reflect its dependence on β_N. We then express $\mu(\beta_N) = \mu(\beta^*) + \gamma_2^{\mathrm{T}}(n^{-1/2}h) + o(n^{-1/2})$, and use the shorthand μ for $\mu(\beta^*)$.

By Le Cam's third lemma, under P^{β^*}, we have

$$\begin{pmatrix} n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}(\beta_N) - \mu\} \\ n^{1/2}(\hat{\beta} - \beta_N) \end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix} -\gamma_1^{\mathrm{T}}V_s^{-1}\tau_{\beta^*}h - \gamma_2^{\mathrm{T}}h \\ -h \end{pmatrix}, \begin{pmatrix} V_1 & \gamma_1^{\mathrm{T}}\tau_{\beta^*}^{-1} \\ \tau_{\beta^*}^{-1}\gamma_1 & \Lambda \end{pmatrix}\right\}$$

in distribution, as $n \to \infty$. Replacing β_N by $\beta^* + n^{-1/2}h$ yields that under P^{β^*},

$$\begin{pmatrix} n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}(\beta^* + n^{-1/2}h) - \mu\} \\ n^{1/2}(\hat{\beta} - \beta^*) \end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix} -\gamma_1^{\mathrm{T}}V_s^{-1}\tau_{\beta^*}h - \gamma_2^{\mathrm{T}}h \\ 0 \end{pmatrix}, \begin{pmatrix} V_1 & \gamma_1^{\mathrm{T}}\tau_{\beta^*}^{-1} \\ \tau_{\beta^*}^{-1}\gamma_1 & \Lambda \end{pmatrix}\right\}$$

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

Heuristically, if the normal distribution was exact, then

$$n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}(\beta^* + n^{-1/2}h) - \mu\} | n^{1/2}(\hat{\beta} - \beta^*) = h \sim \mathcal{N}(-\gamma_2^{\mathrm{T}}h, V_1 - \gamma_1^{\mathrm{T}}V_s^{-1}\gamma_1). \quad \text{(A8)}$$

Given $n^{1/2}(\hat{\beta} - \beta^*) = h$, we have $\beta^* + n^{-1/2}h = \hat{\beta}$, and hence $\hat{\mu}_{\mathrm{PMM}}(\beta^* + n^{-1/2}h) = \hat{\mu}_{\mathrm{PMM}}(\hat{\beta})$. Integrating Equation (A8) over the asymptotic distribution of $n^{1/2}(\hat{\beta} - \beta^*)$, we derive

$$n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}(\hat{\beta}) - \mu\} \sim \mathcal{N}(0, V_1 - \gamma_1^{\mathrm{T}}V_s^{-1}\gamma_1 + \gamma_2^{\mathrm{T}}\Lambda\gamma_2). \quad \text{(A9)}$$

The formal technique to derive Equation (A9) can be find in Andreou and Werker (2012). Equation (A9) gives the result in Theorem 2.

In the following, we provide the proof to Equation (A7). Asymptotic normality of $n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}(\beta_N) - \mu\}$ under P^{β_N} follows from Theorem 1. Asymptotic joint normality of $n^{1/2}(\hat{\beta} - \beta_N)$ and $\log(\mathrm{d}P^{\beta^*}/\mathrm{d}P^{\beta_N})$ follows from Assumption A3. Therefore, the remaining is to show that, under P^{β_N}:

$$\begin{pmatrix} D_N(\beta_N) \\ S_N(\beta_N) \end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_1 & \gamma_1^T \\ \gamma_1 & V_s \end{pmatrix}\right\} \quad \text{(A10)}$$

in distribution, as $n \to \infty$. To prove Equation (A10), consider the linear combination $c_1D_N(\beta_N) + c_2^{\mathrm{T}}S_N(\beta_N)$, which has the same limiting distribution as

$$\begin{aligned} C_N &= c_1 \frac{n^{1/2}}{N} \sum_{i=1}^{N} \left(\frac{I_i}{\pi_i} - 1 \right) m(x_i; \beta_N) + c_1 \frac{n^{1/2}}{N} \sum_{i=1}^{N} \left(\frac{I_i}{\pi_i} - 1 \right) \delta_i (1 + k_{\beta_N, i}) \{ y_i - m(x_i; \beta_N) \} \\ &+ c_2^{\mathrm{T}} \frac{n^{1/2}}{N} \sum_{i=1}^{N} \left(\frac{I_i}{\pi_i} - 1 \right) \delta_i g(x_i; \beta_N) \{ y_i - m(x_i; \beta_N) \}, \end{aligned}$$

given that $nN^{-1} = o(1)$.

We analyze C_N using the martingale theory. First, we rewrite $C_N = \sum_{k=1}^{N} \xi_{N,k}$, where

$$\begin{aligned} \xi_{N,k} &= c_1 \frac{n^{1/2}}{N} \left(\frac{I_k}{\pi_k} - 1 \right) m(x_k; \beta_N) + c_1 \frac{n^{1/2}}{N} \left(\frac{I_k}{\pi_k} - 1 \right) \delta_k (1 + k_{\beta_N, k}) \{ y_k - m(x_k; \beta_N) \} \\ &+ c_2^{\mathrm{T}} \frac{n^{1/2}}{N} \left(\frac{I_k}{\pi_k} - 1 \right) \delta_k g(x_k; \beta_N) \{ y_k - m(x_k; \beta_N) \}. \end{aligned}$$

Consider the σ-fields $\mathcal{F}_{N,k} = \sigma\{x_1, \dots, x_N, \delta_1, \dots, \delta_N, y_1, \dots, y_k, I_1, \dots, I_k\}$ for $1 \le k \le N$. Then,

$$\left\{ \sum_{k=1}^{i} \xi_{N,k}, \mathcal{F}_{N,i}, 1 \le i \le N \} \right\}$$

is a martingale for each $N \ge 1$. Therefore, the limiting distribution of C_N can be studied using the martingale central limit theorem (Billingsley, 1995, theorem 35.12). Under Assumption 4, and the fact that $k_{\beta_N,k}$ has uniformly bounded moments, it follows that $\sum_{k=1}^{N} E_{\beta_N}(|\xi_{N,k}|^{2+\delta}) \to 0$ for some $\delta > 0$. It then follows that Lindeberg's condition in Billingsley's theorem holds. As a result, we obtain that under P^{β_N}, $C_N \to \mathcal{N}(0, \sigma^2)$ in distribution, as $n \to \infty$, where $\sigma^2 = \text{plim} \sum_{k=1}^{N} E_{\beta_N}(\xi_{N,k}^2 | \mathcal{F}_{N,k-1})$. Assumption A3 further implies the following expressions:

$$\begin{aligned} \sigma^2 &= \text{plim} \sum_{k=1}^{N} E_{\beta_N}(\xi_{N,k}^2 | \mathcal{F}_{N,k-1}) \\ &= c_1^2 \text{plim} \frac{n}{N^2} \sum_{k=1}^{N} E_{\beta_N} \left[\left\{ \left(\frac{I_k}{\pi_k} - 1 \right) m(x_k; \beta_N) \right\}^2 | \mathcal{F}_{N,k-1} \right] \\ &+ c_1^2 \text{plim} \frac{n}{N^2} \sum_{k=1}^{N} E_{\beta_N} \left(\left[\left(\frac{I_k}{\pi_k} - 1 \right) \delta_k (1 + k_{\beta_N, k}) \{ y_k - m(x_k; \beta_N) \} \right]^2 | \mathcal{F}_{N,k-1} \right) \\ &+ 2 c_2^{\mathrm{T}} \text{plim} \frac{n}{N^2} \sum_{k=1}^{N} E_{\beta_N} \left[\left(\frac{I_k}{\pi_k} - 1 \right)^2 \delta_k (1 + k_{\beta_N, k}) g(x_k; \beta_N) \{ y_k - m(x_k; \beta_N) \}^2 | \mathcal{F}_{N,k-1} \right] c_1 \\ &+ c_2^{\mathrm{T}} \text{plim} \frac{n}{N^2} \sum_{k=1}^{N} E_{\beta_N} \left[\left(\frac{I_k}{\pi_k} - 1 \right)^2 \delta_k g(x_k; \beta_N) g(x_k; \beta_N)^{\mathrm{T}} \{ y_k - m(x_k; \beta_N) \}^2 | \mathcal{F}_{N,k-1} \right] c_2 \\ &= c_1^2 \text{plim} \frac{n}{N^2} \text{var}_p \left(\sum_{k \in A} \frac{m_k}{\pi_k} \right) + c_1^2 \text{plim} \frac{n}{N^2} \sum_{k=1}^{N} \frac{1 - \pi_k}{\pi_k} \delta_k (1 + k_{\beta^*, k})^2 \sigma^2(x_k) \\ &+ 2 c_2^{\mathrm{T}} \text{plim} \frac{n}{N^2} \sum_{k=1}^{N} \frac{1 - \pi_k}{\pi_k} \delta_k (1 + k_{\beta^*, k}) g(x_k; \beta^*) \sigma^2(x_k) c_1 \end{aligned}$$

$$
\begin{aligned}
&+c_2^{\mathrm{T}}\operatorname{plim}\frac{n}{N^2}\sum_{k=1}^{N}\frac{1-\pi_k}{\pi_k}\delta_k g(x_k;\beta^*)g(x_k;\beta^*)^{\mathrm{T}}\sigma^2(x_k)c_2\\
&= c_1^2V^m + c_1^2V^e + 2c_2^{\mathrm{T}}\gamma_1 c_1 + c_2^{\mathrm{T}}V_s c_2.
\end{aligned}
$$

By the martingale central limit theorem, under P^{β_N}, Equation (A10) follows.

A6 Proof for Theorem 4

The replication method implicitly induces replication weights ω_i^* and random variables u_i such that $E^*(\omega_i^* u_i) = N^{-1}\pi_i^{-1}$ and $\operatorname{var}^*(\omega_i^* u_i) = N^{-2}(1-\pi_i)\pi_i^{-2}$, for $i = 1, \dots, N$, where $E^*(\cdot)$ and $\operatorname{var}^*(\cdot)$ denote the expectation and variance for the resampling given the observed data. For example, in delete-1 jackknife under probability proportional to size sampling with $nN^{-1} = o(1)$, we have $\omega_i^{(k)} = (n-1)^{-1}n\omega_i$ if $i \neq k$, and $\omega_k^{(k)} = 0$. Then, the induced random variables u_i follows a two-point mass distribution as

$$
u_i = \begin{cases} 1, & \text{with probability } \frac{n-1}{n},\\ 0, & \text{with probability } \frac{1}{n}, \end{cases}
$$

and weights $\omega_i^* = (n-1)^{-1}n\omega_i$. It is straightforward to verify that $E^*(\omega_i^* u_i) = \omega_i = N^{-1}\pi_i^{-1}$ and $\operatorname{var}^*\{(\omega_i^* u_i)^2\} = (n-1)^{-1}\omega_i^2 \approx n^{-1}N^{-2}(1-\pi_i)\pi_i^{-2}$.

The kth replicate of $\hat{\beta}$, $\hat{\beta}^{(k)}$, can be viewed as one realization of $\hat{\beta}^*$ which is the solution to the estimating equation

$$
S_N^*(\beta) = n^{1/2}\sum_{i\in A}\omega_i^* u_i \delta_i g(x_i;\beta)\{y_i - m(x_i;\beta)\} = 0. \tag{A11}
$$

Let P^* be the distribution of $z_i^* = (x_i, y_i, \delta_i, I_i, \omega_i^* u_i)$, for $i = 1, \dots, N$, given the observed data induced by bootstrap resampling satisfying

$$
\begin{aligned}
E^*\{S_N^*(\hat{\beta})\} &= n^{1/2}E^*\left[\sum_{i\in A}\omega_i^* u_i \delta_i g(x_i;\hat{\beta})\{y_i - m(x_i;\hat{\beta})\}\right]\\
&= \frac{n^{1/2}}{N}\sum_{i\in A}\frac{1}{\pi_i}\delta_i g(x_i;\hat{\beta})\{y_i - m(x_i;\hat{\beta})\} = 0,
\end{aligned}
$$

and

$$
\begin{aligned}
E^*\{S_N^*(\hat{\beta})S_N^*(\hat{\beta})^{\mathrm{T}}\} &= E^*[\{S_N^*(\hat{\beta}) - S_N(\hat{\beta})\}\{S_N^*(\hat{\beta}) - S_N(\hat{\beta})\}^{\mathrm{T}}]\\
&= nE^*\left[\sum_{i\in A}\left(\omega_i^* u_i - \frac{1}{N\pi_i}\right)^2 \delta_i g(x_i;\hat{\beta})g(x_i;\hat{\beta})^{\mathrm{T}}\{y_i - m(x_i;\hat{\beta})\}^2\right]\\
&= \frac{n}{N^2}\sum_{i\in A}\frac{1-\pi_i}{\pi_i^2}\delta_i g(x_i;\hat{\beta})g(x_i;\hat{\beta})^{\mathrm{T}}\{y_i - m(x_i;\hat{\beta})\}^2.
\end{aligned}
$$

We consider an auxiliary parametric model P^{β} defined locally around $\hat{\beta}$ with a density

$$
\frac{\exp\{n^{1/2}(\beta-\hat{\beta})^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}S_N^*(\hat{\beta}) - 2^{-1}n(\beta-\hat{\beta})^{\mathrm{T}}\Lambda^{-1}(\beta-\hat{\beta})\}}{E^*[\exp\{n^{1/2}(\beta-\hat{\beta})^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}S_N^*(\hat{\beta}) - 2^{-1}n(\beta-\hat{\beta})^{\mathrm{T}}\Lambda^{-1}(\beta-\hat{\beta})\}]}. \tag{A12}
$$

Consider sequences that are local to $\hat{\beta}$, $\beta_N^* = \hat{\beta} + n^{-1/2}h$, indexed by N, and z_i^*, for $i = 1, \dots, N$, with the local shift $P^{\beta_N^*}$. We make the following regularity assumptions:

Assumption A4. (i) Model Equation (A12) is regular; (ii) under $P^{\beta_N^*}$: $S_N^*(\beta_N^*) \to \mathcal{N}(0, V_s)$ in distribution, as $n \to \infty$; (iii) $n^{1/2}(\hat{\beta}^* - \beta_N^*) = \tau_{\beta^*}^{-1} S_N^*(\beta_N^*) + o_p(1)$; (iv) for all bounded continuous functions $h(z_i^*)$, the conditional expectation $E_{\beta_N^*}^*\{h(z_i^*)\}$ converges in distribution to $E_{\hat{\beta}}^*\{h(z_i^*)\}$, where $E_{\beta_N^*}$ is the expectation under $P^{\beta_N^*}$.

Under Equation (A12), the likelihood ratio under $P^{\beta_N^*}$ is

$$\begin{aligned}\log(dP^{\hat{\beta}}/dP^{\beta_N^*}) &= -h^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}S_N^*(\hat{\beta}) + \frac{1}{2}h^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}\tau_{\beta^*}h + o_p(1)\\ &= -h^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}S_N^*(\beta_N^*) - \frac{1}{2}h^{\mathrm{T}}\tau_{\beta^*}V_s^{-1}\tau_{\beta^*}h + o_p(1),\end{aligned}$$

where the second equality follows by the Taylor expansion of $S_N^*(\hat{\beta})$ at β_N^*.

The kth replication of $\hat{\mu}_{\mathrm{PMM}}(\hat{\beta})$, $\hat{\mu}_{\mathrm{PMM}}^{(k)}(\hat{\beta}^{(k)})$, can be viewed as one realization of

$$\hat{\mu}_{\mathrm{PMM}}^*(\hat{\beta}^*) = \sum_{i\in A}\omega_i^* u_i[m(x_i;\hat{\beta}^*) + \delta_i(1 + k_{\overline{\beta}^*,i})\{y_i - m(x_i;\hat{\beta}^*)\}]. \tag{A13}$$

We can derive that under $P^{\beta_N^*}$, the sequence $[n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}^*(\beta_N^*) - \hat{\mu}_{\mathrm{PMM}}(\beta_N^*)\}\ n^{1/2}(\hat{\beta}^* - \beta_N^*)^{\mathrm{T}}$ $\log(dP^{\hat{\beta}}/dP^{\beta_N^*})]^{\mathrm{T}}$ has the same limiting distribution as in Equation (A7). Then, following the same argument in the Proof of Theorem 2, we can obtain that the asymptotic conditional variance of $n^{1/2}\hat{\mu}_{\mathrm{PMM}}^*(\hat{\beta}^*)$, given the observed data, is V_2.

The remaining is to show that, under $P^{\beta_N^*}$ given the observed data:

$$\begin{pmatrix} n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}^*(\beta_N^*) - \hat{\mu}_{\mathrm{PMM}}(\beta_N^*)\} \\ S_N^*(\beta_N^*)\end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} V_1 & \gamma_1^T \\ \gamma_1 & V_s\end{pmatrix}\right\} \tag{A14}$$

in distribution, as $n \to \infty$. To prove Equation (A14), given the observed data, consider the linear combination $c_1 n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}^*(\beta_N^*) - \hat{\mu}_{\mathrm{PMM}}(\beta_N^*)\} + c_2^{\mathrm{T}}S_N^*(\beta_N^*)$, which has the same limiting distribution as

$$\begin{aligned}C_N^* = {} & c_1 n^{1/2}\sum_{i=1}^N I_i\left(\omega_i^* u_i - \frac{1}{N\pi_i}\right)m(x_i;\beta_N^*)\\ &+ c_1 n^{1/2}\sum_{i=1}^N I_i\left(\omega_i^* u_i - \frac{1}{N\pi_i}\right)\delta_i(1 + k_{\beta_N^*,i})\{y_i - m(x_i;\beta_N^*)\}\\ &+ c_2^{\mathrm{T}} n^{1/2}\sum_{i=1}^N I_i\left(\omega_i^* u_i - \frac{1}{N\pi_i}\right)\delta_i g(x_i;\beta_N^*)\{y_i - m(x_i;\beta_N^*)\}.\end{aligned}$$

This is because under $P^{\beta_N^*}$, the extra term in C_N^* compared with $c_1 n^{1/2}\{\hat{\mu}_{\mathrm{PMM}}^*(\beta_N^*) - \hat{\mu}_{\mathrm{PMM}}(\beta_N^*)\} + c_2^{\mathrm{T}}S_N^*(\beta_N^*)$ is

$$\begin{aligned} n^{1/2}\sum_{i=1}^N \frac{I_i}{N\pi_i}\delta_i g(x_i;\beta_N^*)\{y_i - m(x_i;\beta_N^*)\} &= \frac{n^{1/2}}{N}\sum_{i=1}^N \frac{I_i}{\pi_i}\delta_i g(x_i;\hat{\beta})\{y_i - m(x_i;\hat{\beta})\} + O_p(\beta_N^* - \hat{\beta})\\ &= 0 + O_p(n^{-1/2}) = o_p(1).\end{aligned}$$

We analyze C_N^* using the martingale theory. First, we rewrite $C_N^* = \sum_{k=1}^N \xi_{N,k}^*$, where

$$\begin{aligned}\xi_{N,k}^* &= c_1 n^{1/2} I_k \left(\omega_k^* u_k - \frac{1}{N\pi_i}\right) m(x_k;\beta_N^*) + c_1 n^{1/2} I_k \left(\omega_k^* u_k - \frac{1}{N\pi_i}\right) \delta_k(1+k_{\beta_N^*,k})\{y_k - m(x_k;\beta_N^*)\} \\ &\quad + c_2^{\mathrm{T}} n^{1/2} I_k \left(\omega_k^* u_k - \frac{1}{N\pi_i}\right) \delta_k g(x_k;\beta_N^*)\{y_k - m(x_k;\beta_N^*)\},\end{aligned}$$

for $1 \le k \le N$. Consider the σ-fields

$$\mathcal{F}_{N,k}^* = \sigma\{x_1,\ldots,x_N, I_1,\ldots,I_N, \delta_1,\ldots,\delta_N, y_1,\ldots,y_N, \omega_1^* u_1,\ldots,\omega_k^* u_k\}$$

for $1 \le k \le N$. Then, $\left\{\sum_{k=1}^i \xi_{N,k}^*, \mathcal{F}_{N,i}^*, 1 \le i \le N\right\}$ is a martingale for each $N \ge 1$. As a result, we obtain that under $P^{\beta_N^*}$, $C_N^* \to \mathcal{N}(0,\tilde{\sigma}^2)$ in distribution, as $n \to \infty$, where

$$\begin{aligned}\tilde{\sigma}^2 &= \mathrm{plim} \sum_{k=1}^N E_{\beta_N^*}^*(\xi_{N,k}^{*2} | \mathcal{F}_{N,k-1}) \\ &= c_1^2 \mathrm{plim} n \sum_{k=1}^N E_{\beta_N^*}^* \left[\left\{I_k \left(\omega_k^* u_k - \frac{1}{N\pi_i}\right) m(x_k;\beta_N^*)\right\}^2 | \mathcal{F}_{N,k-1}\right] \\ &\quad + c_1^2 \mathrm{plim} n \sum_{k=1}^N E_{\beta_N^*}^* \left(\left[I_k \left(\omega_k^* u_k - \frac{1}{N\pi_i}\right) \delta_k(1+k_{\beta_N^*,k})\{y_k - m(x_k;\beta_N^*)\}\right]^2 | \mathcal{F}_{N,k-1}\right) \\ &\quad + 2c_2^{\mathrm{T}} \mathrm{plim} n \sum_{k=1}^N E_{\beta_N^*}^* \left[I_k \left(\omega_k^* u_k - \frac{1}{N\pi_i}\right)^2 \delta_k(1+k_{\beta_N^*,k}) g(x_k;\beta_N^*)\{y_k - m(x_k;\beta_N^*)\}^2 c_1 | \mathcal{F}_{N,k-1}\right] \\ &\quad + c_2^{\mathrm{T}} \mathrm{plim} n \sum_{k=1}^N E_{\beta_N^*}^* \left[I_k \left(\omega_k^* u_k - \frac{1}{N\pi_i}\right)^2 \delta_k g(x_k;\beta_N^*) g(x_k;\beta_N^*)^{\mathrm{T}} \{y_k - m(x_k;\beta_N^*)\}^2 | \mathcal{F}_{N,k-1}\right] c_2 \\ &= c_1^2 \mathrm{plim} \frac{n}{N^2} \sum_{k=1}^N \frac{I_k(1-\pi_k)}{\pi_k^2} m(x_k;\hat{\beta})^2 + c_1^2 \mathrm{plim} \frac{n}{N^2} \sum_{k=1}^N \frac{I_k(1-\pi_k)}{\pi_k^2} \delta_k(1+k_{\hat{\beta},k})^2 \{y_k - m(x_k;\hat{\beta})\}^2 \\ &\quad + 2c_2^{\mathrm{T}} \mathrm{plim} \frac{n}{N^2} \sum_{k=1}^N \frac{I_k(1-\pi_k)}{\pi_k^2} \delta_k(1+k_{\hat{\beta},k}) g(x_k;\hat{\beta})\{y_k - m(x_k;\hat{\beta})\}^2 c_1 \\ &\quad + c_2^{\mathrm{T}} \mathrm{plim} \frac{n}{N^2} \sum_{k=1}^N \frac{I_k(1-\pi_k)}{\pi_k^2} \delta_k g(x_k;\hat{\beta}) g(x_k;\hat{\beta})^{\mathrm{T}} \{y_k - m(x_k;\hat{\beta})\}^2 c_2 \\ &= c_1^2 \mathrm{plim} \frac{n}{N^2} \sum_{k=1}^N \frac{1-\pi_k}{\pi_k} m(x_k;\beta^*)^2 + c_1^2 \mathrm{plim} \frac{n}{N^2} \sum_{k=1}^N \frac{1-\pi_k}{\pi_k} \delta_k(1+k_{\beta^*,k})^2 \sigma^2(x_k) \\ &\quad + 2c_2^{\mathrm{T}} \mathrm{plim} \frac{n}{N^2} \sum_{k=1}^N \frac{1-\pi_k}{\pi_k} \delta_k(1+k_{\beta^*,k}) g(x_k;\beta^*) \sigma^2(x_k) c_1 \\ &\quad + c_2^{\mathrm{T}} \mathrm{plim} \frac{n}{N^2} \sum_{k=1}^N \frac{1-\pi_k}{\pi_k} \delta_k g(x_k;\beta^*) g(x_k;\beta^*)^{\mathrm{T}} \sigma^2(x_k) c_2.\end{aligned}$$

Therefore, by the martingale central limit theorem, conditional on the observed data under $P^{\beta_N^*}$, Equation (A14) follows. This completes the proof.