The Econometrics of Complex Survey Data

Nearest Neighbor Imputation for General Parameter Estimation in Survey Sampling

Shu Yang, [a] Jae Kwang Kim, [b]

Article information:

To cite this document: Shu Yang, [a] Jae Kwang Kim, [b] . "Nearest Neighbor Imputation for General Parameter Estimation in Survey Sampling" *In* The Econometrics of Complex Survey Data. Published online: 26 Mar 2019; 209-234.
Permanent link to this document:
https://doi.org/10.1108/S0731-905320190000039012

NEAREST NEIGHBOR IMPUTATION FOR GENERAL PARAMETER ESTIMATION IN SURVEY SAMPLING

Shu Yang[a] and Jae Kwang Kim[b]
[a]*Department of Statistics, North Carolina State University, USA*
[b]*Department of Statistics, Iowa State University, USA*

ABSTRACT

Nearest neighbor imputation has a long tradition for handling item nonresponse in survey sampling. In this article, we study the asymptotic properties of the nearest neighbor imputation estimator for general population parameters, including population means, proportions and quantiles. For variance estimation, we propose novel replication variance estimation, which is asymptotically valid and straightforward to implement. The main idea is to construct replicates of the estimator directly based on its asymptotically linear terms, instead of individual records of variables. The simulation results show that nearest neighbor imputation and the proposed variance estimation provide valid inferences for general population parameters.

Keywords: Bahadur representation; bootstrap; jackknife variance estimation; matching; missing at random; quantile estimation

The Econometrics of Complex Survey Data: Theory and Applications
Advances in Econometrics, Volume 39, 209–234

ISSN: 0731-9053/doi:10.1108/S0731-905320190000039012

1. INTRODUCTION

In survey sampling, nearest neighbor imputation is popular for dealing with item nonresponse. In nearest neighbor imputation, for each unit with missing data, the nearest neighbor is identifie among respondents based on the vector of fully observed covariates and then is used as a donor for hot deck imputation (Little & Rubin, 2002). Although nearest neighbor imputation has a long history of application, there are relatively few papers on investigating its statistical properties. Sande (1979) used nearest neighbor imputation in business surveys. Lee and Särndal (1994) studied different methods of nearest neighbor imputation by simulation. Chen and Shao (2000, 2001) developed asymptotic properties for the nearest neighbor imputation estimator of population means. Shao and Wang (2008) proposed methods for constructing confidenc intervals for population means and quantiles with nearest neighbor imputation. Kim et al. (2011) applied nearest neighbor imputation for the US Census long form data. However, most of these studies focused on mean estimation or a one-dimensional covariate in the context of a simple random sample, which is restrictive both theoretically and practically.

In the empirical economics literature, nearest neighbor imputation (also known as matching) has been widely used in evaluation research for adjusting the distribution of covariates among different treatment groups; see Stuart (2010) for a survey of matching estimators. Abadie and Imbens (2006, 2008, 2011, 2012, 2016) systematically studied the asymptotic properties of the matching estimators for the average treatment effects with a finit number of matches. In particular, Abadie and Imbens (2006, 2012) derived the asymptotic distribution for the matching estimators that match directly on the covariates using a martingale representation. Abadie and Imbens (2016) and Yang et al. (2016) further showed that the matching estimators that match on the estimated propensity score are consistent and asymptotically normal. However, these studies are restricted to mean estimation and non-survey data.

Empirical researchers are often interested in various finit population quantities, such as the population means, proportions and quantiles, to name a few (Francisco and Fuller, 1991; Wu and Sitter, 2001; Berger and Skinner, 2003). Some corresponding sample estimators should be treated differently than others. For example, estimators of population quantiles involve nondifferentiable functions of estimated quantities. Moreover, there often are more than one covariate available to facilitate nearest neighbor imputation for survey data. The current framework of nearest neighbor imputation does not fully cover inferences in these settings.

In this article, we provide a framework of nearest neighbor imputation for general parameter estimation in survey sampling. In general, the nearest neighbor imputation estimator is not root-n consistent Abadie and Imbens (2006), where n

is the sample size. Based on a scalar matching variable summarizing all covariates information, we show that nearest neighbor imputation can provide consistent estimators for a fairly general class of parameters. If the matching variable is chosen to be the mean function of the outcome given the covariates, our method resembles predictive mean matching imputation (Rubin, 1986; Little, 1988; Heitjan and Little, 1991). However, unlike predictive mean matching imputation, nearest neighbor imputation does not require the mean function be correctly specified Its consistency only requires the matching variable satisfy certain Lipschitz continuity conditions; see Section 3 for details.

The asymptotic results suggest that variance estimation can proceed based on a large sample approximation to the normal distribution but requires additional estimation for the variance function of the outcome given the covariates. To avoid such complication, we consider replication variance estimation (Rust and Rao, 1996; Wolter, 2007; Mashreghi et al., 2016), which has gained popularity in practice because of its intuitive appeal. Intrinsically, the nearest neighbor imputation estimator with fi ed number of matches is not smooth. The lack of smoothness makes the conventional replication methods invalid for variance estimation (Abadie and Imbens, 2008). This is because the conventional replication method distorts the distribution of the number of times each unit is used as a match, k_i. We provide a heuristic illustration using an unrealistic but insightful example. Suppose in a sample of size $2n$, let Sequence 1 be the firs n observations, and let Sequence 2 be the last n observations. Further, suppose that each observation in Sequence 1 matches to that of Sequence 2. Therefore, the distribution of k_i is degenerated to 1 with probability 1. On the other hand, for the conventional bootstrap, the distribution of k_i^*, where k_i^* is the number of times each unit is used as a match in the bootstrapping sample, would have a different distribution from k_i. Therefore, the conventional bootstrap fails to preserve the distribution of k_i. If the number of matches increases with the sample size, such as in the "kernel matching" estimators of Heckman et al. (1998), both k_i and k_i^* are infinit in the original and conventional bootstrapping samples, and therefore the conventional bootstrap works in this setting. To address the non-smoothness due to the fi ed number of matches, subsampling (Politis et al., 1999) and m out of n bootstrap (Bickel et al., 2012) can be used; however, their consistency relies critically on the choice of the size for subsampling. Unfortunately, there is no clear guidance on how to choose these values in practice. Alternatively, Otsu and Rai (2016) proposed a wild bootstrap method for the matching estimator based on the full vector of covariates in the context of non-survey data. Adusumilli (2017) developed a novel bootstrap procedure for the matching estimator based on the estimated propensity score, built on the notion of "potential errors." His simulation study also demonstrated the superior performance of the bootstrap method relative to using the asymptotic distribution for inference.

We propose new replication variance estimation for nearest neighbor imputation for general parameters in the context of survey data. To address the non-smoothness of the matching estimator, we construct replicates of the estimator directly based on the linear representation of the nearest neighbor imputation estimator. In this way, the distribution of k_i can be preserved, which leads to valid variance estimation. Furthermore, our replication variance method is fl xible, which can accommodate bootstrap and jackknife, among others. To assess the performance of the proposed replicate variance estimator, we run a Monte Carlo simulation study. The simulation results show that the proposed estimator outperforms the conventional replication estimator under various data-generating mechanisms and sampling schemes.

The rest of the article is organized as follows. In Section 2, we introduce the setup and notation and describe the nearest neighbor imputation estimators for general parameters from survey data. In Section 3, we present the main results of the article, which establish asymptotic distributions for the nearest neighbor imputation estimators. In Section 4, we propose a replication method for variance estimation and establish its consistency. In Section 5, we evaluate the finit sample properties of the proposed procedure via Monte Carlo simulation studies under different sampling schemes. Section 6 concludes. Technical details are deferred to the Appendices.

2. BASIC SETUP

Let $\mathcal{F}_N = \{(x_i, y_i, \delta_i) : i = 1, \ldots, N\}$ denote a finit population of size N, where x_i is a p-dimensional vector of covariates, which is always observed, y_i is the outcome that is subject to missingness, and δ_i is the response indicator of y_i, i.e., $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ if it is missing. The δ_i's are define throughout the finit population, as in Shao and Steel (1999) and Kim et al. (2006). We assume that $\mathcal{F}_N$ is a random sample from a superpopulation model ζ, and N is known. Our objective is to estimate the finit population parameter define 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. These parameters are fairly general, which cover many parameters of interest in survey sampling. For example, let $g(y) = y$, $\mu_g \equiv N^{-1} \sum_{i=1}^{N} y_i$ is the population mean of y. Let $g(y) = I(y < c)$ for some constant c, $\mu_g \equiv N^{-1} \sum_{i=1}^{N} I(y_i < c)$ 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.

Let A denote an index set of the sample selected by a probability sampling design. Let I_i be the sampling indicator function, i.e., $I_i = 1$ if unit i is selected into the sample, and $I_i = 0$ otherwise. The sample size is $n = \sum_{i=1}^{N} I_i$. Suppose that π_i,

the first-orde inclusion probability of unit i, is positive and known throughout the sample. If y_i were fully observed throughout the sample, the sample estimator of μ_g and ξ_N are $\hat{\mu}_g = N^{-1}\sum_{i\in A}\pi_i^{-1}g(y_i)$ and $\hat{\xi} = \inf\{\xi : \hat{S}_N(\xi) \geq 0\}$ with $\hat{S}_N(\xi) = \hat{N}^{-1}\sum_{i\in A}\pi_i^{-1}s(y_i - \xi)$ and $\hat{N} = \sum_{i\in A}\pi_i^{-1}$ is an estimator for N. Even with a known N, it is necessary to use $\hat{N}$; we articulate this point in Example 3.

We make the following assumption for the missing data process.

Assumption 1.
(Missing at random and positivity)*The missing data process satisfies* $P(\delta = 1 \mid x, y) = P(\delta = 1 \mid x)$*, denoted by* $p(x)$*. With probability* 1*,* $p(x) > \epsilon$ *for a constant* $\epsilon > 0$.

We focus on the imputation estimators of μ_g and ξ_N given by $\hat{\mu}_{g,I} = N^{-1}\sum_{i\in A}\pi_i^{-1}\left\{\delta_i g(y_i) + (1-\delta_i)g(y_i^*)\right\}$ and $\hat{\xi}_I = \inf\{\xi : \hat{S}_I(\xi) \geq 0\}$, respectively, where $\hat{S}_I(\xi) = \hat{N}^{-1}\sum_{i\in A}\pi_i^{-1}\left\{\delta_i s(y_i - \xi) + (1-\delta_i)s(y_i^* - \xi)\right\}$, and y_i^* is an imputed value of y_i for unit i with $\delta_i = 0$.

To fin suitable imputed values, we use nearest neighbor imputation. Let $d(x_i, x_j)$ be a distance function between x_i and x_j. For example, $d(x_i, x_j) = ||x_i - x_j||$, 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 definit 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.

The classical nearest neighbor imputation can be described in the following steps:

Step 1. For each unit i with $\delta_i = 0$, fin the nearest neighbor from the respondents with the minimum distance between x_i and x_j, for $j \in A_R \equiv \{j \in A : \delta_i = 1\}$. Let $i(1)$ be the index set of its nearest neighbor, which satisfie $d(x_{i(1)}, x_i) \leq d(x_j, x_i)$, for all $j \in A_R$.

Step 2. The nearest neighbor imputation estimators of μ_g and ξ_N are computed by

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

and $\hat{\xi}_{\mathrm{NNI}} = \inf\{\xi : \hat{S}_{\mathrm{NNI}}(\xi) \geq 0\}$, respectively, with

$$\hat{S}_{\mathrm{NNI}}(\xi) = \frac{1}{\hat{N}}\sum_{i\in A}\frac{1}{\pi_i}\left\{\delta_i s(y_i - \xi) + (1-\delta_i)s(y_{i(1)} - \xi)\right\}. \tag{2}$$

In (1) and (2), the imputed values are real observations obtained from the current sample.

3. MAIN RESULTS

For asymptotic inference, we use the framework of Isaki and Fuller (1982), where the asymptotic properties of estimators are established under a f xed sequence of populations and a corresponding sequence of random samples. Specificall , let a sequence of nested finit populations be given by $\mathcal{F}_{N_1} \subset \mathcal{F}_{N_2} \subset \mathcal{F}_{N_3} \subset \cdots$. Also, let a sequence of samples of sizes $\{n_t : t = 1, 2, 3, \ldots\}$ be constructed from the sequence of populations with an increasing sample size $n_1 < n_2 < n_3 < \cdots$. For the ease of exposition, we omit the dependence of N_t and n_t on t. Denote $E_P(\cdot)$ and $\mathrm{var}_P(\cdot)$ to be the expectation and the variance under the sampling design, respectively. We impose the following regularity conditions on the sampling design.

Assumption 2.
(1) There exist positive constants C_1 and C_2 such that $C_1 \le Nn^{-1}\pi_i \le C_2$, for $i = 1, \ldots, N$; (2) the sampling fraction is negligible; i.e., $nN^{-1} = o(1)$; (3) the sequence of the Horvitz–Thompson estimators $\hat{\mu}_{g,\mathrm{HT}} = N^{-1}\sum_{i\in A}\pi_i^{-1}g(y_i)$ satisfies $\mathrm{var}_P(\hat{\mu}_{g,\mathrm{HT}}) = O(n^{-1})$ and $\{\mathrm{var}_P(\hat{\mu}_{g,\mathrm{HT}})\}^{-1/2}(\hat{\mu}_{g,\mathrm{HT}} - \mu_g) \mid \mathcal{F}_N \to \mathcal{N}(0,1)$ in distribution, as $n \to \infty$.

Assumption 2 is widely accepted in survey sampling (Fuller, 2009).

We introduce additional notation. Let $A = A_R \cup A_M$, where A_R and A_M are the sets of respondents and nonrespondents, respectively. Defin $d_{ij} = 1$ if $y_{j(1)} = y_i$, i.e., unit i is used as a donor for unit $j \in A_M$ and $d_{ij} = 0$ otherwise. We write $\hat{\mu}_{g,\mathrm{NNI}}$ in (1) as

$$\begin{aligned}\hat{\mu}_{g,\mathrm{NNI}} &= \frac{1}{N}\left\{\sum_{i\in A}\frac{1}{\pi_i}\delta_i g(y_i) + \sum_{j\in A}\frac{1-\delta_j}{\pi_j}\sum_{i\in A}\delta_i d_{ij} g(y_i)\right\} \\ &= \frac{1}{N}\sum_{i\in A}\frac{\delta_i}{\pi_i}(1+k_i)g(y_i), \qquad (3)\end{aligned}$$

with

$$k_i = \sum_{j\in A}\frac{\pi_i}{\pi_j}(1-\delta_j)d_{ij}. \qquad (4)$$

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

We firs study the asymptotic properties of $\hat{\mu}_{g,\mathrm{NNI}}$. Let $\mu_g(x) \equiv E\{g(y) \mid x\}$ and $\sigma_g^2(x) \equiv \mathrm{var}\{g(y) \mid x\}$, where the expectation and variance are taken with respect to the superpopulation model. We use the following decomposition:

$$n^{1/2}(\hat{\mu}_{g,\mathrm{NNI}} - \mu_g) = D_N + B_N, \tag{5}$$

where

$$D_N = n^{1/2}\left[\frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}\left\{\mu_g(x_i) + \delta_i(1+k_i)\{g(y_i) - \mu_g(x_i)\}\right\} - \mu_g\right], \tag{6}$$

and

$$B_N = \frac{n^{1/2}}{N}\sum_{i\in A}\frac{1}{\pi_i}(1-\delta_i)\{\mu_g(x_{i(1)}) - \mu_g(x_i)\}. \tag{7}$$

The difference $\mu_g(x_{i(1)}) - \mu_g(x_i)$ accounts for the matching discrepancy, and B_N contributes to the asymptotic bias of the matching estimator. In general, if x is p-dimensional, Abadie and Imbens (2006) showed that $d(x_{i(1)}, x_i) = O_P(n^{-1/p})$. Therefore, for nearest neighbor imputation with $p \geq 2$, the asymptotic bias is $B_N = O_P(n^{1/2-1/p}) \neq o_P(1)$. Abadie and Imbens (2011) proposed a bias-adjustment using a nonparametric estimator $\hat{\mu}_g(x)$ that renders matching estimators $n^{1/2}$-consistent. This approach may not be convenient for general parameter estimation.

To address for the matching discrepancy due to a non-scalar x, we propose an alternative method. We firs summarize the covariate information into a scalar matching variable $m = m(x)$ and then apply nearest neighbor imputation based on this matching variable. For simplicity of notation, we may suppress the dependence of m on x if there is no ambiguity. Let $f_1(m)$ and $f_0(m)$ be the conditional density of m given $\delta = 1$ and $\delta = 0$, respectively. We assume the superpopulation model ζ and the matching variable m satisfy the following assumption.

Assumption 3.
(1) The matching variable m has a compact and convex support, with density bounded and bounded away from zero. Suppose that there exist constants C_{1L} and C_{1U} such that $C_{1L} \leq f_1(m)/f_0(m) \leq C_{1U}$; (2) $\mu_g(x)$ and $\mu_s(\xi, x) \equiv E\{s(y-\xi) \mid x\}$ satisfy a Lipschitz continuous condition: there exists a constant C_2 such that $|\mu_g(x_i) - \mu_g(x_j)| < C_2|m_i - m_j|$ and $|\mu_s(\xi, x_i) - \mu_s(\xi, x_j)| < C_2|m_i - m_j|$ for any i and j; (3) there exists $\delta > 0$ such that $E\left\{|g(y)|^{2+\delta} \mid x\right\}$ is uniformly bounded for any x, and $E\left\{|s(y-\xi)|^{2+\delta} \mid x\right\}$ is uniformly bounded for any x and ξ in the neighborhood of ξ_N.

Assumption 3 (1) a convenient regularity condition (Abadie and Imbens 2006). Assumption 3 (2) imposes a smoothness condition for $\mu_g(x)$, $\mu_s(\xi, x)$ and $m(x)$, which is not restrictive (Chen and Shao 2000). One simple example is when the outcome distribution follows a single index model as $E\{g(y) \mid x\} = \phi_g(\beta_0^{\mathrm{T}} x)$, where ϕ_g is a smooth function. There exists some nonparametric estimator $\hat{\beta}$ that is root-n consistent for β_0; see Li and Racine (2007) for a textbook discussion. In this case, $m(x)$ can be taken as the linear predictor $\hat{\beta}^{\mathrm{T}} x$. By a judicious choice, the scalar matching variable should ensure that Assumption 3 holds. If the mean function of the outcome given the covariates is feasible, we can choose the matching variable to be the conditional mean function. We note that in this case the proposed nearest neighbor imputation reduces to the predictive mean matching imputation. However, our method is more general than predictive mean matching imputation, because the latter requires the mean function to be correctly specified Assumption 3 (3) is a moment condition for establishing the central limit theorem.

We derive the asymptotic distribution of $\hat{\mu}_{g,\mathrm{NNI}}$ in the following theorem, with the proof deferred to the Appendices.

Theorem 1. *Under Assumptions 1–3,* $n^{1/2}\left(\hat{\mu}_{g,\mathrm{NNI}} - \mu_g\right) \to \mathcal{N}(0, V_g)$ *in distribution, as* $n \to \infty$*, where*

$$V_g = V_g^{\mu} + V_g^{e} \tag{8}$$

with

$$V_g^{\mu} = \lim_{n\to\infty} \frac{n}{N^2} E\left[\mathrm{var}_P\left\{\sum_{i\in A} \frac{1}{\pi_i}\mu_g(x_i)\right\}\right],$$

$$V_g^{e} = \lim_{n\to\infty} \frac{n}{N^2} E\left[\sum_{i=1}^{N}\left\{\frac{I_i}{\pi_i}\delta_i(1+k_i) - 1\right\}^2 \sigma_g^2(x_i)\right],$$

and k_i *is defined in (4).*

We now establish a similar result for $\hat{\xi}_{\mathrm{NNI}}$, with the proof deferred to the Appendices.

Theorem 2. *Under Assumptions 1–3, suppose the population parameter* ξ_N *and the population estimating function* $S_N(\cdot)$ *satisfy regularity conditions in Assumptions B.1 and B.2. We obtain the following asymptotic linearization representation of* $\hat{\xi}_{\mathrm{NNI}}$*:*

$$n^{1/2}(\hat{\xi}_{\mathrm{NNI}} - \xi_N) = -n^{1/2} S'(\xi_N)^{-1}\{\hat{S}_{\mathrm{NNI}}(\xi_N) - S_N(\xi_N)\} + o_P(1), \tag{9}$$

where $S'(\xi_N) = \mathrm{d}S(\xi_N)/\mathrm{d}\xi$. *It follows that* $n^{1/2}(\hat{\xi}_{\mathrm{NNI}} - \xi_N) \to \mathcal{N}(0, V_\xi)$ *in distribution, as* $n \to \infty$, *where*

$$V_\xi = S'(\xi_N)^{-2}\mathrm{var}\{\hat{S}_{\mathrm{NNI}}(\xi_N)\}, \tag{10}$$

$$\mathrm{var}\{\hat{S}_{\mathrm{NNI}}(\xi_N)\} = \lim_{n\to\infty} \frac{n}{N^2} E\left(\mathrm{var}_P\left[\sum_{i\in A} \frac{E\{s(y_i - \xi_N) \mid x_i\}}{\pi_i}\right]\right)$$
$$+ \lim_{n\to\infty} \frac{n}{N^2} \sum_{i=1}^{N} E\left(\left\{\frac{I_i}{\pi_i}\delta_i(1+k_i) - 1\right\}^2 \mathrm{var}\left[s(y_i - \xi_N) \mid x_i\right]\right), \tag{11}$$

and k_i *is defined in (4).*

For illustration, we use quantile estimation as an example.

Example 1: (Quantile estimation) The estimating function for the αth quantile is $s(y_i - \xi) = I(y_i - \xi \le 0) - \alpha$, and the population estimating equation $S_{\alpha,N}(\xi) = F_N(\xi) - \alpha$, where $F_N(\xi) = N^{-1}\sum_{i=1}^{N} I(y_i \le \xi)$. The nearest neighbor imputation estimator $\hat{\xi}_{\alpha,\mathrm{NNI}}$ is define as

$$\hat{\xi}_{\alpha,\mathrm{NNI}} = \inf\{\xi : \hat{S}_{\alpha,\mathrm{NNI}}(\xi) \ge 0\},$$

where $\hat{S}_{\alpha,\mathrm{NNI}}(\xi) = \hat{F}_{\mathrm{NNI}}(\xi) - \alpha$, $\hat{F}_{\mathrm{NNI}}(\xi) = \hat{N}^{-1}\sum_{i\in A}\pi_i^{-1}\delta_i(1+k_i)I(y_i \le \xi)$, $\hat{N} = \sum_{i\in A}\pi_i^{-1}$, and k_i is define in (4). Let $F(\xi) = \mathrm{P}(y \le \xi)$ be the cumulative distribution function of y. Then, $\hat{F}_{\mathrm{NNI}}(\xi)$ is a Hajek estimator for $F(\xi)$, which is asymptotically equivalent to the one using N instead of $\hat{N}$. Even with a known N, it is necessary to use $\hat{N}$ because $\hat{F}_{\mathrm{NNI}}(\xi)$ for $\xi = \infty$ should be 1. The limiting function of $S_{\alpha,N}(\xi)$ is $S_\alpha(\xi) = F(\xi) - \alpha$. The asymptotic linearization representation of $\hat{\xi}_{\alpha,\mathrm{NNI}}$ is

$$\hat{\xi}_{\alpha,\mathrm{NNI}} - \xi_N = -\frac{\hat{F}_{\mathrm{NNI}}(\xi_N) - F_N(\xi_N)}{f(\xi_N)} + o_P(n^{-1/2}), \tag{12}$$

where $f(\xi) = \mathrm{d}F(\xi)/\mathrm{d}\xi$. Expression (12) is called the Bahadur-type representation for $\hat{\xi}_{\alpha,\mathrm{NNI}}$ (Francisco and Fuller, 1991). The asymptotic variance of $\hat{\xi}_{\alpha,\mathrm{NNI}}$ is then given by (10) with $S'(\xi_N)$ and $\hat{S}_{\mathrm{NNI}}(\xi_N)$ replaced by $f(\xi_N)$ and $\hat{F}_{NNI}(\xi_N)$, respectively.

4. REPLICATION VARIANCE ESTIMATION

Theorems 1 and 2 suggest that variance estimation for the nearest neighbor imputation estimators can be obtained using the sample analogues of the asymptotic variance formulas. This approach involves estimation of the variance function of the outcome given the covariates. Alternatively, we consider replication variance estimation (Rust and Rao, 1996; Wolter, 2007).

Let $\hat{\mu}_g$ be the Horvitz–Thompson estimator of μ_g. The replication variance estimator of $\hat{\mu}_g$ takes the form of

$$\hat{V}_{\text{rep}}(\hat{\mu}_g) = \sum_{k=1}^{L} c_k(\hat{\mu}_g^{(k)} - \hat{\mu}_g)^2, \tag{13}$$

where L is the number of replicates, c_k is the kth replication factor and $\hat{\mu}_g^{(k)}$ is the kth replicate of $\hat{\mu}_g$. For $\hat{\mu}_g = \sum_{i\in A} \omega_i g(y_i)$, we can write the replicate of $\hat{\mu}_g$ as $\hat{\mu}_g^{(k)} = \sum_{i\in A} \omega_i^{(k)} g(y_i)$, where $\omega_i^{(k)}$ is the replication weight that account for the complex sampling design. The replicates are constructed such that $E_P\{\hat{V}_{\text{rep}}(\hat{\mu}_g)\} = \text{var}_P(\hat{\mu}_g)\{1 + o(1)\}$.

Example 2: In the delete-1 jackknife method, we have $L = n$, $c_k = (n-1)n^{-1}$,

$$\omega_i^{(k)} = \begin{cases} (n-1)^{-1} & \text{if } i \neq k, \\ 0 & \text{if } i = k, \end{cases}$$

under simple random sampling.

We now propose a new replication variance estimation for $\hat{\mu}_{g,\text{NNI}}$. Let $\psi_i = \mu_g(x_i) + \delta_i(1 + k_i)\{g(y_i) - \mu_g(x_i)\}$ and $\mu_\psi = N^{-1}\sum_{i=1}^{N} \psi_i$. Then, the Horvitz–Thompson estimator for μ_ψ is $\hat{\psi}_{\text{HT}} = \sum_{i\in A} \omega_i \psi_i$, where $\omega_i = N^{-1}\pi_i^{-1}$. By Theorem 1, we have $\hat{\mu}_{g,\text{NNI}} - \hat{\psi}_{\text{HT}} = o_P(n^{-1/2})$. Moreover, we have $\mu_\psi - \mu_g = O_P(N^{-1/2})$. Therefore,

$$\begin{aligned} \hat{\mu}_{g,\text{NNI}} - \mu_g &= (\hat{\mu}_{g,\text{NNI}} - \hat{\psi}_{\text{HT}}) + (\hat{\psi}_{\text{HT}} - \mu_\psi) + (\mu_\psi - \mu_g), \\ &= o_P(n^{-1/2}) + (\hat{\psi}_{\text{HT}} - \mu_\psi) + O_P(N^{-1/2}). \end{aligned}$$

With negligible sampling fractions, i.e., $nN^{-1} = o(1)$, $\hat{\mu}_{g,\text{NNI}} - \mu_g = \hat{\psi}_{\text{HT}} - \mu_\psi + o_P(n^{-1/2})$. Then, it is sufficien to estimate $\text{var}(\hat{\psi}_{\text{HT}} - \mu_\psi) = E\{\text{var}_P(\hat{\psi}_{\text{HT}} - \mu_\psi)\}$,

which is essentially the sampling variance of $\hat{\psi}_{\mathrm{HT}}$. This suggests that we can treat $\{\psi_i : i \in A\}$ as pseudo observations in applying the replication variance estimator. Otsu and Rai (2016) used a similar idea to develop a wild bootstrap technique for the matching estimators for the average treatment effects. To be specific we construct replicates of $\hat{\psi}_{\mathrm{HT}}$ as follows: $\hat{\psi}_{\mathrm{HT}}^{(k)} = \sum_{i \in A} \omega_i^{(k)} \psi_i$. The replication variance estimator of $\hat{\psi}_{\mathrm{HT}}$ is obtained by applying $\hat{V}_{\mathrm{rep}}(\cdot)$ in (13) for the above replicates $\hat{\psi}_{\mathrm{HT}}^{(k)}$. It follows that $E\{\hat{V}_{\mathrm{rep}}(\hat{\psi}_{\mathrm{HT}})\} = \mathrm{var}(\hat{\psi}_{\mathrm{HT}} - \mu_\psi)\{1 + o(1)\} = \mathrm{var}(\hat{\mu}_{g,\mathrm{NNI}} - \mu_g)\{1 + o(1)\}$. Because the pseudo observations $\psi_i's$ involve unknown $\mu_g(x)$, we use a nonparametric estimator $\hat{\mu}_g(x)$. Concretely, we adopt sieves estimators (Geman and Hwang, 1982; Chen, 2007) which includes power series estimators as examples; see the Appendices for details.

In summary, the new replication variance estimation for $\hat{\mu}_{g,\mathrm{NNI}}$ proceeds as follows:

Step 1. Obtain a sieves estimator for $\mu_g(x)$, denoted by $\hat{\mu}_g(x)$.

Step 2. Construct replicates of $\hat{\mu}_{g,\mathrm{NNI}}$ as

$$\hat{\mu}_{g,\mathrm{NNI}}^{(k)} = \sum_{i \in A} \omega_i^{(k)} [\hat{\mu}_g(x_i) + \delta_i(1 + k_i)\{g(y_i) - \hat{\mu}_g(x_i)\}], \tag{14}$$

where $\omega_i^{(k)}$ is the kth replication weight for unit i.

Step 3. Apply $\hat{V}_{\mathrm{rep}}(\cdot)$ in (13) for the above replicates to obtain the replication variance estimator of $\hat{\mu}_{g,\mathrm{NNI}}$.

We now consider a replication variance estimator for $\hat{\xi}_{\mathrm{NNI}}$. Following the previous section, we obtain the asymptotic variance of $\hat{\xi}_{\mathrm{NNI}}$ using var $\left\{\hat{S}_{\mathrm{NNI}}(\xi)\right\}$ and $S'(\xi)$. First, to estimate var $\left\{\hat{S}_{\mathrm{NNI}}(\xi)\right\}$, we can use the similar replication variance estimation earlier in this section by considering $I(y < \xi)$ and $\mu_s(\xi, x)$ instead of y and $\mu_g(x)$. Second, to estimate $S'(\xi)$, we follow the kernel-based derivative estimation of Deville (1999):

$$\hat{S}'(\xi) = \frac{1}{Nh} \sum_{i \in A} \frac{1}{\pi_i} \int s(y_i - x) K'\left(\frac{\xi - x}{h}\right) \mathrm{d}x \tag{15}$$

where $K(\cdot)$ is a kernel function, $K'(x) = \mathrm{d}K(x)/\mathrm{d}x$, and h is the bandwidth. Under Assumption C.1 for the kernel function and bandwidth and previously stated regularity conditions on the superpopulations and sampling designs, the kernel-based estimator (15) is consistent for $S'(\xi)$.

In summary, the new replication variance estimation for $\hat{\xi}_{\text{NNI}}$ proceeds as follows:

Step 1. Obtain a sieves logit estimator for $\mu_s(\hat{\xi}_{\text{NNI}}, x)$, denoted by $\hat{\mu}_s(\hat{\xi}_{\text{NNI}}, x)$; see the Appendices for details.

Step 2. Construct replicates of $\hat{S}_{\text{NNI}}(\hat{\xi}_{\text{NNI}})$ as

$$\hat{S}^{(k)}_{\text{NNI}}(\hat{\xi}_{\text{NNI}}) = \sum_{i \in A} \omega_i^{(k)} [\hat{\mu}_s(\hat{\xi}_{\text{NNI}}, x_i) + \delta_i(1 + k_i)\{s(y_i - \hat{\xi}_{\text{NNI}}) - \hat{\mu}_s(\hat{\xi}_{\text{NNI}}, x_i)\}]. \tag{16}$$

Step 3. Apply $\hat{V}_{\text{rep}}(\cdot)$ in (13) for the above replicates to obtain the variance estimator of $\hat{S}_{\text{NNI}}(\hat{\xi}_{\text{NNI}})$, denoted as $\hat{V}_{\text{rep}}\{\hat{S}_{\text{NNI}}(\hat{\xi}_{\text{NNI}})\}$.

Step 4. Obtain the kernel-based derivative estimator $\hat{S}'(\hat{\xi}_{\text{NNI}})$, where $\hat{S}'(\xi)$ is define in (15).

Step 5. Calculate the variance estimator of $\hat{\xi}_{\text{NNI}}$ as $\hat{S}'(\hat{\xi}_{\text{NNI}})^{-2}\hat{V}_{\text{rep}}\{\hat{S}_{\text{NNI}}(\hat{\xi}_{\text{NNI}})\}$.

For illustration, we continue with Example 3.

Example 3: (Quantile estimation (Cont.)) Obtain a sieves logit estimator for $F(\xi) = P(y \le \xi)$ and a kernel-based estimator for $f(\xi)$, denoted as $\hat{F}(\xi)$ and $\hat{f}(\xi)$, respectively. Construct replicates of $\hat{F}_{\text{NNI}}(\hat{\xi}_{\alpha,\text{NNI}})$ as

$$\hat{F}^{(k)}_{\text{NNI}}(\hat{\xi}_{\alpha,\text{NNI}}) = \sum_{i \in A} \omega_i^{(k)} [\hat{F}(\hat{\xi}_{\alpha,\text{NNI}}) + \delta_i(1 + k_i)\{I(y_i \le \hat{\xi}_{\alpha,\text{NNI}}) - \hat{F}(\hat{\xi}_{\alpha,\text{NNI}})\}].$$

Apply $\hat{V}_{\text{rep}}(\cdot)$ in (13) for the above replicates to obtain the replication variance estimator of $\hat{F}_{\text{NNI}}(\hat{\xi}_{\alpha,\text{NNI}})$, denoted as $\hat{V}_{\text{rep}}\{\hat{F}_{\text{NNI}}(\hat{\xi}_{\alpha,\text{NNI}})\}$. Calculate the variance estimator of $\hat{\xi}_{\alpha,\text{NNI}}$ as $\hat{f}(\hat{\xi}_{\alpha,\text{NNI}})^{-2}\hat{V}_{\text{rep}}\{\hat{F}_{\text{NNI}}(\hat{\xi}_{\alpha,\text{NNI}})\}$.

We present the consistency results for the proposed replication variance estimators, with the proof presented in the Appendices.

Theorem 3. *Suppose assumptions in Theorem 2 and Assumptions D.1 and D.2 for the sieves estimators hold. Suppose further that $\hat{V}_{\text{rep}}(\hat{\mu}_g)$ in (13) is consistent for $\text{var}_p(\hat{\mu}_g)$. Then, the replication variance estimator for $\hat{\mu}_{g,\text{NNI}}$ is consistent, i.e., $n\hat{V}_{\text{rep}}\{\hat{\mu}_{g,\text{NNI}}\}/V_g \to 1$ in probability, as $n \to \infty$, where the replicates of $\hat{\mu}_{g,\text{NNI}}$ are given in (14), and V_g is given in (8).*

Given that the kernel-based estimator $\hat{S}'(\xi)$ in (15) is consistent for $S'(\xi)$, the replication variance estimator for $\hat{\xi}_{\text{NNI}}$ is consistent, i.e., $n\hat{V}_{\text{rep}}\{\hat{\xi}_{\text{NNI}}\}/V_\xi \to 1$ in probability, as $n \to \infty$, where the replicates of $\hat{S}_{\text{NNI}}(\hat{\xi}_{\text{NNI}})$ are given in (16), and V_ξ is given in (11).

5. SIMULATION STUDY

In this section, we investigate the finite-sampl performance of the proposed replication method for variance estimation and constructing confidenc intervals and comparing them to conventional competitors.

For generating finit 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} and x_{6i} and e_i be generated independently from $\mathcal{N}(0, 1)$; then, let y_i be generated under six mechanisms: (P1) $y_i = -1 + x_{1i} + x_{2i} + e_i$, (P2) $y_i = -1.5 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + e_i$, (P3) $y_i = -1.5 + x_{1i} + \cdots + x_{6i} + e_i$, (P4) $y_i = -1 + x_{1i} + x_{2i} + x_{1i}^2 + x_{2i}^2 - 2/3 + e_i$, (P5) $y_i = -1.5 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + x_{1i}^2 + x_{2i}^2 - 2/3 + e_i$ and (P6) $y_i = -1.5 + x_{1i} + \cdots + x_{6i} + x_{1i}^2 + x_{2i}^2 - 2/3 + e_i$. 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 $\text{logit}\{p(x_i)\} = x_i^{\mathrm{T}}\mathbf{1}$, where x_i includes all corresponding covariates under each data-generating mechanism and $\mathbf{1}$ is a vector of 1 with a compatible length. This results in a 75% response rate, on average. The parameters of interest are $\mu = N^{-1}\sum_{i=1}^{N} y_i$, $\eta = N^{-1}\sum_{i=1}^{N} I(y_i < c)$, where c is the 80th percentile such that the true value of η is 0.8, and the median ξ. To generate samples, we consider two sampling designs: (S1) simple random sampling with $n = 800$ and (S2) probability proportional to size sampling. In (S2), for each unit in the population, we generate a size variable s_i as $\log(|y_i + \nu_i| + 4)$, where $\nu_i \sim \mathcal{N}(0, 1)$ and specify the selection probability as $\pi_i = 400 s_i / \sum_{i=1}^{N} s_i$. Therefore, (S2) is endogenous (also known as informative), where units with larger y_i values have larger probabilities to be selected into the sample.

For nearest neighbor imputation, the matching scalar variable m is set to be the conditional mean function of y given x, $m(x)$, approximated by power series estimation. For investigating the effect of the matching variable, we consider the power series including all firs and second order terms under (P1)–(P3) and only firs order terms under (P4)–(P6), so that $m(x)$ is correctly specifie for the mean function under (P1)–(P3) but misspecifie under (P4)–(P6). We construct 95% confidenc intervals 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 conventional and proposed jackknife variance estimation. In the conventional jackknife variance estimation, the whole procedure of nearest neighbor imputation is repeated on the replicated data sets for obtaining the replicates for the estimators. In the proposed jackknife variance estimation, the kth replicates of $\hat{\mu}_{\text{NNI}}$, $\hat{\eta}_{\text{NNI}}$ and $\hat{\xi}_{\text{NNI}}$ are given by

$$\hat{\mu}_{\text{NNI}}^{(k)} = \sum_{i=1}^{n} \omega_i^{(k)}[\hat{\mu}(x_i) + \delta_i(1 + k_i)\{y_i - \hat{\mu}(x_i)\}],$$

$$\hat{\eta}_{\mathrm{NNI}}^{(k)} = \sum_{i=1}^{n} \omega_i^{(k)} [\hat{\mu}_\eta(x_i) + \delta_i(1+k_i)\{I(y_i < c) - \hat{\mu}_\eta(x_i)\}],$$

$$\hat{\xi}_{\mathrm{NNI}}^{(k)}(\hat{\xi}_{\mathrm{NNI}})$$

$$= \hat{f}(\hat{\xi}_{\mathrm{NNI}})^{-2} \sum_{i=1}^{n} \omega_i^{(k)} [\hat{\mu}_s(\hat{\xi}_{\mathrm{NNI}}, x_i) + \delta_i(1+k_i)\{I(y_i \le \hat{\xi}_{\mathrm{NNI}}) - \hat{\mu}_s(\hat{\xi}_{\mathrm{NNI}}, x_i)\}],$$

where $\hat{\mu}_\eta(x)$, $\hat{\mu}_s(\xi, x)$ and $\hat{f}(x)$ are nonparametric estimators of $\mu_\eta(x) = P(y < c \mid x)$, $\mu_s(\xi, x) = P(y < \xi \mid x)$ and $f(\xi)$, respectively. These are obtained by kernel regression using a Gaussian kernel with bandwidth $h = 1.5n^{-1/5}$. We note that k_i is the number of times that y_i is selected to impute the missing values of y based on the original data and therefore is kept the same across replicated data sets. The variance estimators are compared in terms of empirical coverage rate and relative bias, $\{E(\hat{V}_I) - V\}/V$, where V is the true variance estimated from Monte Carlo samples.

Tables 1 and 2 present the simulation results under simple random sampling and probability proportional to size sampling, respectively, based on 2,000 Monte Carlo samples. Under both sampling designs, the nearest neighbor imputation estimator has small biases for all parameters μ, η and ξ, under (P1)–(P3) with $m(x)$ correctly specifie for the mean function and (P4)–(P6) with $m(x)$ misspecifie for the mean function. For variance estimation, as expected, the conventional jackknife variance estimator is severely biased, indicating that the lack of smoothness of the matching estimator needs to be taken into account in variance estimation. In contrast, the proposed jackknife variance estimators provide satisfactory results under both sampling designs and for all parameters. The relative biases are small and the empirical coverage rates are close to the nominal coverage of 95% of confidence Overall, the simulation results suggest that the proposed replication variance estimation works reasonably well under the settings we considered.

6. CONCLUDING REMARKS

We focus on inference of general population parameters when the outcome is missing at random in survey data using nearest neighbor 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 not only asymptotically valid mean estimators but also valid distribution and quantile estimators. This article establishes asymptotic properties of the nearest neighbor imputation estimators based on a scalar variable summarizing all covariate information. Because of the non-smooth nature of nearest neighbor

Table 1. Simulation Results for The Population Mean μ, the Population Proportion $\eta = 0.8$ and the Population Median ξ Under Simple Random Sampling: Bias ($\times 10^2$) and Standard Error (SE $\times 10^2$) of the Point Estimator, Relative Bias of Jackknife Variance Estimates (RB $\times 10^2$) and Coverage Rate (CR %) of 95% Confidenc Intervals.

Simple Random Sampling								
					Prop JK		Conv JK	
		$m(x)$	Bias	SE	RB	CR	RB	CR
μ	(P1)	c	0.00	4.87	0.1	94.9	>1,000	100
	(P2)	c	0.12	6.08	0.5	95.3	>1,000	100
	(P3)	c	1.09	8.42	2.2	95.3	>1,000	100
	(P4)	m	−0.10	5.41	3.6	96.0	>1,000	100
	(P5)	m	0.20	6.59	0.1	95.4	>1,000	100
	(P6)	m	1.17	8.81	0.3	94.8	>1,000	100
η	(P1)	c	0.00	1.77	0.4	95.0	>1,000	100
	(P2)	c	0.00	1.53	−0.1	94.9	>1,000	100
	(P3)	c	−0.01	1.50	−5.1	94.7	>1,000	100
	(P4)	m	0.03	1.63	6.1	95.4	>1,000	100
	(P5)	m	0.05	1.48	4.3	95.5	>1,000	100
	(P6)	m	−0.01	1.47	−0.7	94.9	>1,000	100
ξ	(P1)	c	−0.25	6.15	2.7	94.8	>1,000	100
	(P2)	c	−0.40	7.60	2.5	94.7	>1,000	100
	(P3)	c	−0.37	10.19	4.0	94.6	>1,000	100
	(P4)	m	−0.25	7.09	3.2	94.6	>1,000	100
	(P5)	m	−0.35	8.17	7.2	96.0	>1,000	100
	(P6)	m	−0.54	10.78	1.8	94.1	>1,000	100

Prop JK: Proposed jackknife variance estimation; Conv JK: conventional jackknife variance estimation. c: correctly specifie and m: misspecified

imputation, we propose a novel replication method for variance estimation based on linearization of the estimator, which is asymptotically valid, while the conventional replication methods are not. Simulation results show that, under various scenarios, the proposed method outperforms the conventional counterparts. Coupled with the proposed replication procedure, the nearest neighbor imputation inference is straightforward to implement requiring only software routines for existing estimators.

In the empirical economic literature, as an important example in evaluation research, causal inference of treatment effects can be viewed from a missing data perspective (e.g., Ding and Li, 2018). Propensity score matching has been recently proposed for inferring causal effects of treatments in the context of survey data;

Table 2. Simulation Results for the Population Mean μ, the Population Proportion $\eta = 0.8$ and the Population Median ξ Under Probability Proportional to Size Sampling: Bias ($\times 10^2$) and Standard Error (SE $\times 10^2$) of the Point Estimator, Relative Bias of Jackknife Variance estimates (RB $\times 10^2$) and Coverage Rate (CR %) of 95% Confidenc Intervals.

Probability Proportional to Size								
					Prop JK		Conv JK	
		$m(x)$	Bias	SE	RB	CR	RB	CR
μ	(P1)	c	0.07	4.71	1.8	95.4	>1,000	100
	(P2)	c	0.20	5.71	6.1	95.9	>1,000	100
	(P3)	c	0.73	7.71	6.0	96.1	>1,000	100
	(P4)	m	−0.06	5.29	2.4	95.5	>1,000	100
	(P5)	m	0.22	6.08	7.0	95.9	>1,000	100
	(P6)	m	0.99	8.23	5.4	95.1	>1,000	100
η	(P1)	c	−0.01	1.89	−6.0	94.5	>1,000	100
	(P2)	c	0.02	1.63	−1.9	95.3	>1,000	100
	(P3)	c	0.08	1.66	−5.5	94.4	>1,000	100
	(P4)	m	0.02	1.79	−4.0	95.2	>1,000	100
	(P5)	m	0.03	1.60	1.8	95.2	>1,000	100
	(P6)	m	0.08	1.67	−8.7	93.7	>1,000	100
ξ	(P1)	c	−0.31	6.34	6.2	94.8	>1,000	100
	(P2)	c	−0.06	8.30	0.8	94.5	>1,000	100
	(P3)	c	−0.42	11.36	5.4	94.6	>1,000	100
	(P4)	m	−0.32	7.57	4.1	94.0	>1,000	100
	(P5)	m	−0.34	8.91	7.0	94.8	>1,000	100
	(P6)	m	−0.49	12.22	2.2	94.4	>1,000	100

Prop JK: Proposed jackknife variance estimation; Conv JK: conventional jackknife variance estimation; c: correctly specifie and m: misspecified

however, their asymptotic properties are underdeveloped (Lenis et al., 2017). The proposed methodology here can be easily generalized to investigate the asymptotic properties of propensity score matching estimators with survey weights.

Our methodology and theoretical results for nearest neighbor imputation represent an important building block for future developments. Such developments can follow three lines. First, extending the current theory to non-negligible sampling fractions is possible; see, e.g., Mashreghi et al. (2014). For non-negligible sampling fraction, note that

$$\text{var}\left(\hat{\mu}_{g,\text{NNI}} - \mu_g\right) = \text{var}\left(\hat{\psi}_{\text{HT}} - \mu_\psi\right) + \text{var}\left(\mu_\psi - \mu_g\right) + o(n^{-1})$$

and var $\left(\mu_\psi - \mu_g\right) = O(N^{-1})$. Thus, we can add a model-based estimator of var $\left(\mu_\psi - \mu_g\right)$ in addition to the replication variance estimator for var$(\hat{\psi}_{\text{HT}} - \mu_\psi)$. Second, instead of choosing the nearest neighbor as a donor for missing items, we can consider fractional imputation (Kim and Fuller, 2004; Yang et al., 2013; Kim and Yang, 2014; Yang and Kim, 2016) using K $(K > 1)$ nearest neighbors. Third, writing $y_i = x_i R_i$ and using the fact that x_i is always observed, we can apply nearest neighbor imputation only to impute R_i, which can be called nearest neighbor ratio imputation.

ACKNOWLEDGMENTS

Dr. Yang is partially supported by NSF grant DMS 1811245, NCI grant P01 CA142538, and Ralph E. Powe Junior Faculty Enhancement Award from Oak Ridge Associated Universities. Dr. Kim is partially supported by NSF grant MMS 1733572.

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. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, *107*, 833–843.

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

Adusumilli, K. (2017). Bootstrap inference for propensity score matching. Retrieved from https://economics.sas.upenn.edu/events/bootstrap-inference-propensity-score-matching.

Berger, Y. G., & Skinner, C. J. (2003). Variance estimation for a low income proportion. *Journal of the Royal Statistical Society: Series C*, *52*, 457–468.

Bickel, P. J., Götze, F., & van Zwet, W. R. (2012). Resampling fewer than n observations: Gains, losses, and remedies for losses. In *Selected works of Willem van Zwet* (pp. 267–297). New York, NY: Springer.

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.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, *25*, 193–204.

Ding, P., & Li, F. (2018). Causal inference: A missing data perspective. *Statistical Science*, *33*, 214–237.
Francisco, C. A., & Fuller, W. A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, *19*, 454–469.
Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons.
Geman, S., & Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, *10*, 401–414.
Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*, 1017–1098.
Heitjan, D. F., & Little, R. J. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics*, *40*, 13–29.
Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficien estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1161–1189.
Ichimura, H., & Linton, O. B. (2005). Asymptotic expansions for some semiparametric program evaluation estimators. In D. Andrews & J. Stock (Eds.), *Identification and inference in econometric models: Essays in honor of Thomas J. Rothenberg*. Cambridge: Cambridge University Press.
Isaki, C. T., & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, *77*, 89–96.
Kim, J. K., & Fuller, W. (2004). Fractional hot deck imputation. *Biometrika*, *91*, 559–578.
Kim, J. K., Fuller, W. A., Bell, W. R., et al. (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 stratifie sampling. *Journal of the American Statistical Association*, *101*, 312–320.
Kim, J. K., & Yang, S. (2014). Fractional hot deck imputation for robust inference under item nonresponse in survey sampling. *Survey Methodology*, *40*, 211–230.
Lee, H., & Särndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, *10*, 231–243.
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*. doi:10.1093/biostatistics/kxx063.
Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.
Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*, 287–296.
Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.
Mashreghi, Z., Haziza, D., & Léger, C. (2016). A survey of bootstrap methods in finit 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. *Canadian Journal of Statistics*, *42*, 142–167.
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
Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. New York, NY: Springer-Verlag.
Rubin, D. B. (1986). Statistical matching using fil 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.
Sande, I. G. (1979). A personal view of hot deck imputation procedures. *Survey Methodology*, *5*, 238–258.

Serfling R. J. (1980). *Approximation theorems of mathematical statistics*. Hoboken, NJ: John Wiley & Sons.

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.

Shao, J., & Wang, H. (2008). Confidenc intervals based on survey data with nearest neighbor imputation. *Statistica Sinica*, *18*, 281–297.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21.

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

Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, *96*, 185–193.

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

Yang, S., & Kim, J. K. (2016). Fractional imputation in survey sampling: A comparative review. *Statistical Science*, *31*, 415–432.

Yang, S., Kim, J. K., & Shin, D. W. (2013). Imputation methods for quantile estimation under missing at random. *Statistics and its Interface*, *6*, 369–377.

APPENDICES

The Appendices include proofs of Theorems 1–3 and additional technical details.

APPENDIX A: PROOF FOR THEOREM 1

With a scalar matching variable m, we have

$$\begin{aligned} B_N &= \frac{n^{1/2}}{N}\sum_{i\in A}\frac{1}{\pi_i}(1-\delta_i)\{\mu_g(x_{i(1)})-\mu_g(x_i)\} \\ &\le \frac{n^{1/2}}{N}\sum_{i\in A}\frac{1}{\pi_i}(1-\delta_i)\,|\,m_{i(1)}-m_i\,|=o_P(1), \end{aligned}$$

where $\le$ in the second line followed by Assumption 3 (2). Based on the decomposition in (5), we can write

$$n^{1/2}(\hat{\mu}_{g,\mathrm{NNI}}-\mu_g)=D_N+o_P(1), \tag{A.1}$$

where D_N is define in (6). Then, to study the asymptotic properties of $n^{1/2}(\hat{\mu}_{g,\mathrm{NNI}}-\mu_g)$, we only need to study the asymptotic properties of D_N. For simplicity, we introduce the following notation: $\mu_{g,i}\equiv\mu_g(x_i)$ and $e_i\equiv g(y_i)-\mu_{g,i}$. We express

$$\begin{aligned} D_N &= \frac{n^{1/2}}{N}\left[\sum_{i\in A}\frac{1}{\pi_i}\left\{\mu_{g,i}+\delta_i(1+k_i)e_i\right\}-\sum_{i=1}^{N}g(y_i)\right] \\ &= \frac{n^{1/2}}{N}\sum_{i=1}^{N}\left(\frac{I_i}{\pi_i}-1\right)\mu_{g,i}+\frac{n^{1/2}}{N}\sum_{i=1}^{N}\left\{\frac{I_i}{\pi_i}\delta_i(1+k_i)-1\right\}e_i, \end{aligned} \tag{A.2}$$

and we can verify that the covariance of the two terms in (A.2) is zero. Thus,

$$\mathrm{var}(D_N)=\mathrm{var}\left\{\frac{n^{1/2}}{N}\sum_{i=1}^{N}\left(\frac{I_i}{\pi_i}-1\right)\mu_{g,i}\right\}+\mathrm{var}\left[\frac{n^{1/2}}{N}\sum_{i=1}^{N}\left\{\frac{I_i}{\pi_i}\delta_i(1+k_i)-1\right\}e_i\right].$$

As $n\to\infty$, the firs term becomes

$$V_g^{\mu}=\lim_{n\to\infty}\frac{n}{N^2}E\left\{\mathrm{var}_P\left(\sum_{i\in A}\frac{\mu_{g,i}}{\pi_i}\right)\right\},$$

and the second term becomes

$$V_g^e = \text{plim}\frac{n}{N^2}\sum_{i=1}^{N}\left\{\frac{I_i}{\pi_i}\delta_i(1+k_i)-1\right\}^2 \text{var}(e_i \mid x_i).$$

The remaining is to show that $V_g^e = O(1)$. To do this, the key is to show that the moments of k_i are bounded. Under Assumption 2, it is easy to verify that

$$\underline{\omega}\tilde{k}_i \le k_i \le \bar{\omega}\tilde{k}_i, \tag{A.3}$$

for some constants $\underline{\omega}$ and $\bar{\omega}$, where $\tilde{k}_i = \sum_{j\in A}(1-\delta_j)d_{ij}$ is the number of unit i used as a match for the nonrespondents. Under Assumption 3, $\tilde{k}_i = O_P(1)$ and $E(\tilde{k}_i)$ and $E(\tilde{k}_i^2)$ are uniformly bounded over n (Abadie and Imbens 2006, Lemma 3); therefore, together with (A.3), we have $k_i = O_P(1)$ and $E(k_i)$ and $E(k_i^2)$ are uniformly bounded over n. Therefore, a simple algebra yields $V_g^e = O(1)$.

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

APPENDIX B: PROOF FOR THEOREM 2

Assumption B.1.
The following conditions hold for the population parameter ξ_N and the population estimating function $S_N(\cdot)$:

1. *The population parameter ξ_N lies in a closed interval $\mathcal{I}_\xi$;*
2. *the function $s(\cdot)$ is bounded;*
3. *the population estimating function $S_N(\xi)$ converges to $S(\xi)$ uniformly on $\mathcal{I}_\xi$ as $N\to\infty$, and the equation $S(\xi)=0$ has a unique root in the interior of $\mathcal{I}_\xi$;*
4. *the limiting function $S(\xi)$ is strictly increasing and absolutely continuous with finite first derivative in $\mathcal{I}_\xi$, and the derivative $S'(\xi)$ is bounded away from 0 for ξ in $\mathcal{I}_\xi$; and*
5. *the population quantities*

$$\sup_{\xi\in\mathcal{I}_s} N^\alpha |S_N(\xi_N + N^{-\alpha}\xi) - S_N(\xi_N) - S(\xi_N + N^{-\alpha}\xi) - S(\xi_N)| \to 0,$$

and

$$\sup_{\xi\in\mathcal{I}_s} N^{-1}\sum_{i=1}^{N} |s(y_i - \xi_N - N^{-\alpha}\xi) - s(y_i - \xi_N)| = O_P(N^{-\alpha}),$$

where $\mathcal{I}_s$ is a large enough compact set and $\alpha \in (1/4, 1/2]$.

Assumption B.1 (5) holds with probability one under suitable assumptions on the probability mechanism generating the y_i's and on the function $s(\cdot)$, and therefore it is justifiable Under Assumption B.1, by the standard arguments from the theory on M-estimators (Serfling 1980), $\hat{\xi}_{\text{NNI}}$ is consistent for ξ_N. We further make the following assumption.

Assumption B.2.
The nearest neighbor imputation estimator $\hat{\xi}_{\text{NNI}}$ is root-n consistent for ξ_N.

Now, we give proof for Theorem 2. Under Assumptions B.1 and B.2, we can write

$$\hat{S}_{\text{NNI}}(\hat{\xi}_{\text{NNI}}) - S_N(\xi_N) = \{\hat{S}_{\text{NNI}}(\xi_N) - S_N(\xi_N)\} + S'(\xi_N)(\hat{\xi}_{\text{NNI}} - \xi_N) + o_P(n^{-1/2}). \tag{B.1}$$

By Assumption B.1 (4), $S(\xi)$ is smooth, and therefore $S_N(\xi_N) = O_P(N^{-1})$, $\hat{S}_{\text{NNI}}(\hat{\xi}_{\text{NNI}}) = O_P(n^{-1})$, and the left hand side of (B.1) is $o_P(n^{-1/2})$. Therefore, we can obtain a linearization for $\hat{\xi}_{\text{NNI}}$ as in (9).

Based on the linearization (9), the asymptotic variance is

$$V_\xi = S'(\xi_N)^{-2}\text{var}\{\hat{S}_{\text{NNI}}(\xi_N)\}.$$

Following a similar derivation in the proof for Theorem 1, it is easy to show that

$$\begin{aligned}\text{var}\{\hat{S}_N(\xi)\} \;=\; & \lim_{n\to\infty} \frac{n}{N^2} E\left(\text{var}_P\left[\sum_{i\in A} \frac{E\{s(y_i-\xi)\,|\,x_i\}}{\pi_i}\right]\right) \\ & + \lim_{n\to\infty} \frac{n}{N^2} \sum_{i=1}^{N} \left\{\frac{I_i}{\pi_i}\delta_i(1+k_i) - 1\right\}^2 \text{var}\left[s(y_i-\xi)\,|\,x_i\right].\end{aligned}$$

APPENDIX C: ASSUMPTIONS FOR KERNEL FUNCTIONS

Assumption C.1.
The following conditions hold for kernel function $K(\cdot)$ and bandwidth h:

1. *the kernel function $K(\cdot)$ is absolutely continuous with nonzero finite derivative $K'(\cdot)$ and $\int K(x)\mathrm{d}x = 1$;*
2. *the bandwidth $h \to 0$ and $nh \to \infty$ as $n \to \infty$;*
3. *there exists a constant c, such that $|h^{-1}K'(x_1/h) - h^{-1}K'(x_2/h)| \le c|x_1 - x_2|$ for any x_1, x_2 and an arbitrarily small h.*

Assumption C.1 states conditions on the smoothness and tail behavior of the kernel functions. Popular kernel functions, including Epanechnikov, Gaussian and triangle kernels, satisfy the required conditions.

APPENDIX D: SIEVES ESTIMATION

The method of sieves (Geman and Hwang, 1982) offers a powerful tool for estimation for nonparametric or semiparametric models. See Chen (2007) for a textbook discussion. In particular, the sieves can be constructed using linear spans of power series. For illustration, we describe the power series estimator for $\mu_g(x)$ (Newey, 1997) and the series logit estimator for $s_\xi(x) = I(y - \xi \le 0) - \alpha$ (Hirano et al., 2003; Ichimura and Linton, 2005).

Power Series Estimator for $\mu_g(x)$

We consider continuous $g(y)$ and power series estimation for $\mu_g(x)$ with K terms in the series, where K increases with n. Let p be the dimension of X. Consider a sequence of power functions

$$p^K(x) = (p_1(x), \ldots, p_K(x))^{\mathrm{T}}, \tag{D.1}$$

where $p_j(x) = x^{\lambda_j} \equiv x_1^{\lambda_{j1}} \times \cdots \times x_p^{\lambda_{jp}}$ with $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jp})$, and $|\lambda_j| = \sum_{k=1}^{p} \lambda_{jk}$ is nondecreasing in j.

For simplicity of the presentation, let the firs r units be the respondents, i.e., $\delta_i = 1$ for $i = 1, \ldots, r$. From the observations $\{(x_i, y_i) : i = 1, \ldots, r\}$, the power series estimator of $\mu_g(x)$ can be calculated as the predicted value obtained from a weighted regression of $g(y_i)$ on $p^K(x_i)$. To be precise, let $P = \left(p^K(x_1), \ldots, p^K(x_r)\right)^{\mathrm{T}}$ and $G_r = (g(y_1), \ldots, g(y_r))^{\mathrm{T}}$. A power series estimator of $\mu_g(x)$ takes the form

$$\hat{\mu}_g(x) = p^K(x)^{\mathrm{T}} (P^{\mathrm{T}} W P)^{-} P^{\mathrm{T}} W G_r, \tag{D.2}$$

where W is a diagonal matrix with the ith diagonal element π_i^{-1}, and $(P^{\mathrm{T}} W P)^{-}$ denotes a generalized inverse of a matrix $P^{\mathrm{T}} W P$.

Suppose the following assumption holds for establishing the fast convergence rate of $\hat{\mu}_g(x)$ in (D.2).

Assumption D.1.

1. *The support of x is a Cartesian product of compact intervals;*
2. *$\mu_g(x)$ is s-times continuously differentiable at x with $s/p > 1$;*
3. *the number of series $K = O(n^\nu)$ with $0 < \nu < 1/3$.*

Assumption D.1 (2) requires $\mu_g(x)$ to be sufficientl smooth, depending on the dimension of x and the number of derivatives of $\mu_g(x)$. Assumption D.1 (3) requires the number of series increases at a certain rate.

Lemma D.1. *Under Assumption D.1, the power series estimator $\hat{\mu}_g(x)$ in (D.2) satisfies that* $\sup_x |\hat{\mu}_g(x) - \mu_g(x)| = O_P\left(\sqrt{K^3/n} + K^{1-s/p}\right) = o_P(1)$.

The proof of Lemma D.1 can be found in Newey (1997).

Series Logit Estimator for $\mu_s(\xi, x) = E\{I(y - \xi \le 0) \mid x\} - \alpha$

Denote $p_\xi(x) = E\{I(y \le \xi) \mid x\}$ and $\text{logit}(a) = \{1 + \exp(-a)\}^{-1}$. The series logit estimator for $p_\xi(x)$ can be obtained as

$$\hat{p}_\xi(x) = \text{logit}\{p^K(x)^\top \hat{\pi}_K\}, \tag{D.3}$$

where $p^K(x)$ is define in (D.1), and

$$\hat{\pi}_K = \arg\max_\pi \sum_{i \in A} \pi_i^{-1} \Big(I(y_i - \xi \le 0)\text{logit}\{p^K(x_i)^\top \pi\} + I(y_i - \xi > 0)[1 - \text{logit}\{p^K(x_i)^\top \pi\}]\Big).$$

Suppose that the following assumption holds for establishing the fast convergence rate of the series logit estimator $\hat{p}_\xi(x)$ in (D.3).

Assumption D.2.

1. *The support of x is a Cartesian product of compact intervals;*
2. *$p_\xi(x)$ is s times continuously differentiable with $s/p \ge 3$;*
3. *$p_\xi(x)$ is bounded away from zero and one on the support of x;*
4. *the density of x is bounded away from zero on the support of x;*
5. *the number of series $K = O(n^\nu)$ with $\nu < 1$.*

Lemma D.2. *Under Assumption D.2, the series logit estimator $\hat{p}_\xi(x)$ in (D.3) satisfies that* $\sup_x |\hat{p}_\xi(x) - p_\xi(x)| = O_P\left(\sqrt{K/n} + K^{1-(s/2p)}\right)$

The proof of Lemma D.2 can be found in Hirano et al. (2003).

Remark 4. *When the dimension of x, p, becomes larger, Assumption D.1 (2) and Assumption D.2 (2) typically require more stringent smoothness on $\mu_g(x)$ and $p_\xi(x)$ in x. Alternatively, we can estimate $\mu_g(x)$ and $p_\xi(x)$ by applying the power series constructed based on $m_i = m(x_i)$, i.e., using the 1-dimensional variable m_i.*

APPENDIX E: PROOF FOR THEOREM 3

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

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

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

In what follows, we use $P^*(\cdot)$ to denote the probability mass or density function induced from resampling given the observed data and use the supscript $*$ to indicate the random variables resulting from one replication sampling. Then, the kth replication of $\hat{\mu}_{g,\mathrm{NNI}}$, $\hat{\mu}_{g,\mathrm{NNI}}^{(k)}$, can be viewed as one realization of

$$\begin{aligned} \hat{\mu}_{g,\mathrm{NNI}}^* &= \sum_{i\in A} \omega_i^* [\hat{\mu}_g(x_i) + \delta_i(1+k_i)\{g(y_i) - \hat{\mu}_g(x_i)\}]u_i \\ &= \sum_{i\in A} \omega_i^* [\mu_g(x_i) + \delta_i(1+k_i)\{g(y_i) - \mu_g(x_i)\}]u_i \\ &\quad + \sum_{i\in A} \omega_i^* \{(1-\delta_i) + \delta_i k_i\}\{\hat{\mu}_g(x_i) - \mu_g(x_i)\}u_i \\ &= \sum_{i\in A} \omega_i^* \psi_i u_i + R_N^*, \end{aligned} \tag{E.1}$$

where $R_N^* = \sum_{i\in A} \omega_i^* \{(1-\delta_i) + \delta_i k_i\}\{\hat{\mu}_g(x_i) - \mu_g(x_i)\}u_i$.

We now show $E^*\left\{\left(n^{1/2}R_N^*\right)^2\right\} \to 0$ in probability. We write

$$
\begin{aligned}
&E^*\left\{\left(n^{1/2}R_N^*\right)^2\right\} \\
&= nNE\left\{(\omega_1^* u_1)^2\right\}\frac{1}{N}\sum_{i\in A}\{(1-\delta_i)+\delta_i k_i\}^2\{\hat{\mu}_g(x_i)-\mu_g(x_i)\}^2 \\
&\quad +2nN(N-1)E^*\left(\omega_1^*\omega_2^* u_1 u_2\right)\frac{1}{N(N-1)}\sum_{i\neq j\in A}\{(1-\delta_i)+\delta_i k_i\} \\
&\quad \times\{(1-\delta_j)+\delta_j k_j\}\{\hat{\mu}_g(x_i)-\mu_g(x_i)\}\{\hat{\mu}_g(x_j)-\mu_g(x_j)\}.
\end{aligned}
$$

Because of Assumption 2 (1), and the facts that $nNE\left\{(\omega_1^* u_1)^2\right\} = O(1)$, $nN(N-1)E^*\left(\omega_1^*\omega_2^* u_1 u_2\right) = O(1)$, the uniform convergence of $\hat{\mu}_g(x)$ to $\mu_g(x)$ in Lemma D.1, and $E(k_i^l)$ is uniformly bounded over n and for any $l > 0$, we obtain $E^*\left\{\left(n^{1/2}R_N^*\right)^2\right\} \to 0$ in probability. Then, by the Markov inequality, we obtain for any ϵ, $P^*\left(n^{1/2}|R_N^*| > \epsilon\right) \to 0$ in probability.

It then becomes straightforward to verify that $\hat{V}_{\text{rep}}(\cdot)$ applied to $\hat{\mu}_{g,\text{NNI}}^{(k)}$ is consistent for $\text{var}(\bar{\psi}_n)$ and therefore for $\text{var}(\hat{\mu}_{g,\text{NNI}})$.

The proof for the second part of Theorem 3 is similar and therefore omitted.