A Martingale Representation for Matching Estimators

Alberto ABADIE and Guido W. IMBENS

Matching estimators are widely used in statistical data analysis. However, the large sample distribution of matching estimators has been derived only for particular cases. This article establishes a martingale representation for matching estimators. This representation allows the use of martingale limit theorems to derive the large sample distribution of matching estimators. As an illustration of the applicability of the theory, we derive the asymptotic distribution of a matching estimator when matching is carried out without replacement, a result previously unavailable in the literature. In addition, we apply the techniques proposed in this article to derive a correction to the standard error of a sample mean when missing data are imputed using the "hot deck," a matching imputation method widely used in the Current Population Survey (CPS) and other large surveys in the social sciences. We demonstrate the empirical relevance of our methods using two Monte Carlo designs based on actual datasets. In these Monte Carlo exercises, the large sample distribution of matching estimators derived in this article provides an accurate approximation to the small sample behavior of these estimators. In addition, our simulations show that standard errors that do not take into account hot-deck imputation of missing data may be severely downward biased, while standard errors that incorporate the correction for hot-deck imputation perform extremely well. This article has online supplementary materials.

KEY WORDS: Hot-deck imputation; Martingales; Overt bias; Treatment effects.

1. INTRODUCTION

Matching methods provide simple and intuitive tools for adjusting the distribution of covariates among samples from different populations. Probably because of their transparency and intuitive appeal, matching methods are widely used in evaluation research to estimate treatment effects when all treatment confounders are observed (Rubin 1973a, 1977; Dehejia and Wahba 1999; Rosenbaum 2002; Hansen 2004). Matching is also used for the analysis of missing data, where it is often referred to as "hot-deck imputation" (Little and Rubin 2002). As a notorious example, missing weekly earnings are currently imputed using hot-deck methods for more than 30% of the records with weekly earnings data in the monthly U.S. Current Population Survey (CPS) files (Bollinger and Hirsch 2009).

In spite of the pervasiveness of matching methods, the asymptotic distribution of matching estimators has been derived only for special cases (Abadie and Imbens 2006). In the absence of large sample approximation results to the distribution of matching estimators, empirical researchers employing matching methods have sometimes used the bootstrap as a basis for inference. However, recent results have shown that, in general, the bootstrap does not provide valid large sample inference for matching estimators (Abadie and Imbens 2008). Similarly, the properties of statistics based on data imputed using sequential hot-deck methods, similar to those employed in the CPS and other large surveys, are not well understood, and empirical researchers using these surveys typically ignore missing data imputation issues when they construct standard errors. Andridge and Little (2010) provided a recent survey on hot-deck imputation methods.

The main contribution of this article is to establish a martingale representation for matching estimators. This representation allows the use of martingale limit theorems (Hall and Heyde 1980; Billingsley 1995; Shorack 2000) to derive the asymptotic distribution of matching estimators. Because the martingale representation applies to a large class of matching estimators, the applicability of the methods presented in this article is very broad. Despite its simplicity and immediate implications, the martingale representation of matching estimators described in this article seems to have been previously unnoticed in the literature. The use of martingale methods is attractive because the limit behavior of martingale sequences has been extensively studied in the statistics literature (see, e.g., Hall and Heyde 1980).

As an illustration of the usefulness of the theory, we apply the martingale methods proposed in this article to derive the asymptotic distribution of a matching estimator when matching is carried out without replacement, a result previously unavailable in the literature. In addition, we apply the techniques proposed in this article to derive a correction to the standard error of a sample mean when missing data are imputed using the hot deck.

Finally, we demonstrate the empirical relevance of our methods using two Monte Carlo designs based on actual datasets. In these Monte Carlo exercises, the large sample distribution of matching estimators derived in this article provides an accurate approximation to the small sample behavior of these estimators. In addition, our simulations show that standard errors that do not take into account hot-deck imputation of missing data may be severely downward biased, while standard errors that incorporate the correction for hot-deck imputation perform extremely well.

In this article, we reserve the term "matching" for procedures that use a small number of matches per unit. Heckman, Ichimura, and Todd (1998) proposed estimators that treat the number of matches as an increasing function of the sample size. Under certain conditions, these estimators have asymptotically linear representations, so their large sample distributions can be

Alberto Abadie, John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138 (E-mail: alberto_abadie@harvard.edu). Guido W. Imbens, Department of Economics, Harvard University, Cambridge, MA 02138 (E-mail: imbens@harvard.edu). We thank Rustam Ibragimov, Don Rubin, the editor (Leonard Stefanski), the associate editor, the referees, and seminar participants for useful comments and suggestions. We are also grateful to Greg Weyland, from the U.S. Census Bureau, for patiently explaining to us the intricacies of the hot-deck imputation algorithms employed in the Current Population Survey. This work was supported by the National Science Foundation under grant SES-0961707.

Journal of the American Statistical Association
June 2012, Vol. 107, No. 498, Theory and Methods
DOI: 10.1080/01621459.2012.682537

derived using the standard machinery for asymptotically linear estimators. In contrast, despite the pervasiveness of matching estimators that use a small number of matches (e.g., hot-deck imputation in the CPS), the previous literature does not provide a general framework for establishing their large sample properties.

The rest of the article is organized as follows. Section 2 describes matching estimators. Section 3 presents the main result of the article, which establishes a martingale representation for matching estimators. In Section 4, we apply martingale techniques to analyze the large sample properties of a matching estimator when matching is carried out without replacement. In Section 5, we apply martingale techniques to study hot-deck imputation. Section 6 describes the Monte Carlo simulation exercises and reports the results. Section 7 concludes. Proofs are presented in the Appendix.

2. MATCHING ESTIMATORS

Let W be a binary variable that indicates membership to a particular population of interest. Empirical researchers often compare the distributions of some variable, Y, between units with $W = 1$ and units with $W = 0$ after adjusting for the differences in a $(k \times 1)$ vector of observed covariates, $\mathbf{X}$. For example, in discrimination litigation research, W may represent membership in a certain demographic group, Y may represent labor wages, and $\mathbf{X}$ may represent a vector of variables describing job and/or worker characteristics. In evaluation research, W typically indicates exposure to an active treatment or intervention, Y is an outcome of interest, and $\mathbf{X}$ is a vector of observed confounders. As in that literature, we will say that units with $W = 1$ are "treated" and units with $W = 0$ are "untreated." Let

$$\tau = E[Y|W = 1] - E[E[Y|\mathbf{X}, W = 0]|W = 1]. \qquad (1)$$

In evaluation research, τ is given a causal interpretation as the "average treatment effect on the treated" under unconfoundedness assumptions (Rubin 1977). Applied researchers often use matching methods to estimate τ. Other parameters of interest that can be estimated by matching methods include (1) the "average treatment effect" on the entire population, which is of widespread interest in evaluation studies, (2) parameters that focus on features of the distribution of Y other than the mean, and (3) parameters estimated by hot-deck imputation methods in the presence of missing data. Rosenbaum (2002), Imbens (2004), and Rubin (2006) provided detailed surveys of the literature. For concreteness, and to avoid tedious repetition or unnecessary abstraction, in this section, we discuss matching estimation of τ only. While our main focus is on "treatment effect" parameters, in Section 5, we show that the techniques proposed in this article can be applied in the context of missing data imputation. The two contexts are intimately related, because estimating treatment effects can be seen as a missing data problem (Rubin 1974; Rosenbaum and Rubin 1983).

Also, to avoid notational clutter, we consider only estimators with a fixed number of matches, M, per unit. However, as it will be explained later, our techniques can also be applied to estimators for which the number of matches may differ across units (see, e.g., Hansen 2004).

Consider two random samples of sizes N_0 and N_1 of untreated and treated units, respectively. Pooling these two samples, we obtain a sample of size $N = N_0 + N_1$ containing both treated and untreated units. For each unit in the pooled sample, we observe the triple $(Y, \mathbf{X}, W)$. For each treated unit i, let $\mathcal{J}_M(i)$ be the indices of M untreated units, with values in the covariates similar to $\mathbf{X}_i$ (where M is some small positive integer). In other words, $\mathcal{J}_M(i)$ is a set of M matches for observation i. To simplify notation, we will assume that at least one of the variables in the vector $\mathbf{X}$ has a continuous distribution, so perfect matches happen with probability zero. Let $\|\cdot\|$ be some norm in $\mathbb{R}^k$ (typically the Euclidean norm). Let 1_A be the indicator function for the event A. For matching with replacement, $\mathcal{J}_M(i)$ consists of the indices of the M untreated observations with the closest covariate values to $\mathbf{X}_i$:

$$\mathcal{J}_M(i) = \Bigg\{ j \in \{1, \ldots, N\} \text{ s.t. } W_j = 0, \left(\sum_{k=1}^{N} (1 - W_k)\, 1_{\{\|\mathbf{X}_i - \mathbf{X}_j\| \le \|\mathbf{X}_i - \mathbf{X}_k\|\}} \right) \le M \Bigg\}.$$

For matching without replacement, the elements of $\{\mathcal{J}_M(i) \text{ s.t. } W_i = 1\}$ are nonoverlapping subsets of $\{j \in \{1, \ldots, N\} \text{ s.t. } W_j = 0\}$, chosen to minimize the sum of the matching discrepancies:

$$\sum_{i=1}^{N} W_i \sum_{j \in \mathcal{J}_M(i)} \|\mathbf{X}_i - \mathbf{X}_j\|.$$

In both cases, the matching estimator of τ is defined as

$$\widehat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} W_i \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j \right). \qquad (2)$$

Many other matching schemes are possible (see, e.g., Gu and Rosenbaum 1993; Rosenbaum 2002; Hansen 2004; Iacus, King, and Porro 2009; Diamond and Sekhon 2010), and the results in this article are of broad generality. However, as discussed above, our results pertain to matching estimators that employ a small number, M, of matches per unit. Heckman, Ichimura, and Todd (1998) proposed "kernel matching" estimators, which require that the number of matches increases with the sample size (with $M \to \infty$ as $N \to \infty$) to consistently estimate the conditional expectation function $E[Y|\mathbf{X}, W = 0]$ in Equation (1). In addition, the results of this article apply to estimators that match directly on the covariates, $\mathbf{X}$, and do not directly apply to matching on the estimated propensity score (Rosenbaum and Rubin 1983). Abadie and Imbens (2010) derived an adjustment to the distribution of the propensity score matching estimators for the case when the propensity score is not known, so matching is done on a first-step estimator of the propensity score.

3. A MARTINGALE REPRESENTATION FOR MATCHING ESTIMATORS

This section derives a martingale representation for matching estimators. For $w \in \{0, 1\}$, let $\mu_w(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}, W = w]$ and $\sigma_w^2(\mathbf{x}) = \text{var}(Y|\mathbf{X} = \mathbf{x}, W = w)$. Given Equation (2), we can write $\widehat{\tau} - \tau = D_N + R_N$, where

$$D_N = \frac{1}{N_1}\sum_{i=1}^{N} W_i(\mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i) - \tau) + \frac{1}{N_1}\sum_{i=1}^{N} W_i\left((Y_i - \mu_1(\mathbf{X}_i)) - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}(Y_j - \mu_0(\mathbf{X}_j))\right),$$

and

$$R_N = \frac{1}{N_1}\sum_{i=1}^{N} W_i\left(\mu_0(\mathbf{X}_i) - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\mu_0(\mathbf{X}_j)\right).$$

The term R_N is the conditional bias of matching estimators described by Abadie and Imbens (2006). This term is zero if all matches are perfect (i.e., if all matching discrepancies, $\mathbf{X}_i - \mathbf{X}_j$ for $j \in \mathcal{J}_M(i)$, are zero), or if the regression μ_0 is a constant function. In most cases of interest, however, this term is different from zero, as perfect matches happen with probability zero for continuous covariates. The order of magnitude of R_N depends on the number of continuous covariates, as well as the magnitude of N_0 relative to N_1. Under appropriate conditions, $\sqrt{N_1}R_N$ converges in probability to zero [see Section 4 for the case of matching without replacement, or Abadie and Imbens (2006) for the case of matching with replacement].

Next, it will be shown that the term D_N is a martingale array with respect to a certain filtration. First, note that

$$D_N = \frac{1}{N_1}\sum_{i=1}^{N} W_i(\mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i) - \tau) + \frac{1}{N_1}\sum_{i=1}^{N}\left(W_i - (1 - W_i)\frac{K_{N,i}}{M}\right)(Y_i - \mu_{W_i}(\mathbf{X}_i)),$$

where $K_{N,i}$ is the number of times that observation i (with $W_i = 0$) is used as a match:

$$K_{N,i} = \sum_{j=1}^{N} 1_{\{i\in\mathcal{J}_M(j)\}}.$$

Therefore, we can write

$$\sqrt{N_1}D_N = \sum_{k=1}^{2N}\xi_{N,k},$$

where

$$\xi_{N,k} = \begin{cases} \dfrac{1}{\sqrt{N_1}} W_k(\mu_1(\mathbf{X}_k) - \mu_0(\mathbf{X}_k) - \tau), & \text{if } 1 \le k \le N, \\ \dfrac{1}{\sqrt{N_1}}\left(W_{k-N} - (1 - W_{k-N})\dfrac{K_{N,k-N}}{M}\right) \times (Y_{k-N} - \mu_{W_{k-N}}(\mathbf{X}_{k-N})), & \text{if } N+1 \le k \le 2N. \end{cases}$$

Let $\mathcal{X}_N = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ and $\mathcal{W}_N = \{W_1, \ldots, W_N\}$. Consider the σ-fields $\mathcal{F}_{N,k} = \sigma\{\mathcal{W}_N, \mathbf{X}_1, \ldots, \mathbf{X}_k\}$ for $1 \le k \le N$ and $\mathcal{F}_{N,k} = \sigma\{\mathcal{W}_N, \mathcal{X}_N, Y_1, \ldots, Y_{k-N}\}$ for $N+1 \le k \le 2N$. Then, and this is the key insight in this article,

$$\left\{\sum_{j=1}^{i}\xi_{N,j}, \mathcal{F}_{N,i},\ 1 \le i \le 2N\right\}$$

is a martingale for each $N \ge 1$. As a result, the asymptotic behavior of $\sqrt{N_1}D_N$ can be analyzed using martingale methods. This martingale representation holds no matter whether matching is done with or without replacement, whether a fixed or a variable number of matches per unit are used, or which particular distance metric is employed to measure the matching discrepancies. Regardless of the choice of matching scheme, if matches depend only on the covariates $\mathbf{X}$, a martingale representation holds for $\sqrt{N_1}D_N$. The reason is that no matter how matching is implemented, (1) the number of times that unit k is used as a match, $K_{N,k}$, is a deterministic function of $\mathcal{X}_N$ and $\mathcal{W}_N$, and (2) $E[Y_k - \mu_{W_k}(\mathbf{X}_k) \mid \mathcal{X}_N, \mathcal{W}_N, Y_1, \ldots, Y_{k-1}] = 0$.

So far, we have considered the case where $K_{N,i}$ is fixed, given $\mathcal{X}_N$ and $\mathcal{W}_N$, for all $1 \le i \le N$. This assumption does not hold for certain matching schemes that break matching ties using randomization. Note, however, that any sequence of randomized tiebreakers can be included in the set of variables that span $\mathcal{F}_{N,k}$, for $N+1 \le k \le 2N$, to preserve the martingale representation of D_N. As a result, our derivations extend easily to randomized matching methods.

4. APPLICATION: MATCHING WITHOUT REPLACEMENT

In this section, we illustrate the usefulness of the martingale representation of matching estimators by deriving the asymptotic distribution of a matching estimator when matching is done without replacement, so $K_{N,i} \in \{0, 1\}$ for every unit i with $W_i = 0$. To simplify the exposition, we obviate some regularity conditions in the derivations. A precise statement of the result, including all regularity conditions, is provided at the end of the section.

For $1 \le k \le N$, the conditional variances of the martingale differences are given by

$$\begin{aligned} E\left[\xi^2_{N,k}|\mathcal{F}_{N,k-1}\right] &= \frac{1}{N_1}W_k E[(\mu_1(\mathbf{X}_k) - \mu_0(\mathbf{X}_k) - \tau)^2|\mathcal{F}_{N,k-1}] \\ &= \frac{1}{N_1}W_k E[(\mu_1(\mathbf{X}_k) - \mu_0(\mathbf{X}_k) - \tau)^2|W_k = 1]. \end{aligned}$$

For $N+1 \le k \le 2N$, the conditional variances of the martingale differences are given by

$$\begin{aligned} E\left[\xi^2_{N,k}|\mathcal{F}_{N,k-1}\right] &= \frac{1}{N_1}E\left[\left(W_{k-N} - (1 - W_{k-N})\frac{K_{N,k-N}}{M}\right)^2 \times (Y_{k-N} - \mu_{W_{k-N}}(\mathbf{X}_{k-N}))^2\Big|\mathcal{F}_{N,k-1}\right] \\ &= \frac{1}{N_1}\left(W_{k-N}\sigma^2_1(\mathbf{X}_{k-N}) + (1 - W_{k-N}) \times \frac{K_{N,k-N}}{M^2}\sigma^2_0(\mathbf{X}_{k-N})\right) \\ &= \frac{1}{N_1}W_{k-N}\left(\sigma^2_1(\mathbf{X}_{k-N}) + \frac{\sigma^2_0(\mathbf{X}_{k-N})}{M}\right) + r_{N,k-N}, \end{aligned}$$

where

$$r_{N,k-N} = \frac{1}{N_1}\left((1 - W_{k-N})\frac{K_{N,k-N}}{M^2}\sigma^2_0(\mathbf{X}_{k-N}) - W_{k-N}\frac{\sigma^2_0(\mathbf{X}_{k-N})}{M}\right).$$

Assume that the conditional variance function $\sigma_0^2(\mathbf{x})$ is Lipschitz continuous, with Lipschitz constant equal to $c_{\sigma_0^2}$. For $1 \leq i \leq N$ such that $W_i = 1$, let $\|\mathbf{U}_{N_0,N_1,i}^{(M,m)}\|$ be the mth matching discrepancy for treated unit i when untreated units are matched without replacement to treated units in such a way that the sum of the matching discrepancies is minimized. That is, if unit i is a treated observation, and unit j is the mth match for unit i, then $\|\mathbf{U}_{N_0,N_1,i}^{(M,m)}\| = \|\mathbf{X}_i - \mathbf{X}_j\|$. Lipschitz continuity of $\sigma_0^2(\mathbf{x})$ implies

$$\left|\sum_{k=N+1}^{2N} r_{N,k-N}\right| \leq \frac{c_{\sigma_0^2}}{M^2}\frac{1}{N_1}\sum_{i=1}^{N}\sum_{m=1}^{M} W_i \left\|\mathbf{U}_{N_0,N_1,i}^{(M,m)}\right\|.$$

Because the average matching discrepancy converges to zero in probability (see Proposition 1 in the Appendix for a stronger result), the weak law of large numbers implies

$$\sum_{k=1}^{2N} E\left[\xi_{N,k}^2 \middle| \mathcal{F}_{N,k-1}\right] \xrightarrow{p} \sigma^2,$$

where

$$\sigma^2 = E[(\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) - \tau)^2 | W = 1] + E\left[\sigma_1^2(\mathbf{X}) + \frac{\sigma_0^2(\mathbf{X})}{M}\middle| W = 1\right]. \quad (3)$$

In view of this result, to apply a martingale central limit theorem to D_N, it is sufficient to check the Lindeberg's condition,

$$\sum_{k=1}^{2N} E\left[\xi_{N,k}^2 1_{\{|\xi_{N,k}| \geq \varepsilon\}}\right] \to 0, \quad \text{for all } \varepsilon > 0$$

[see Billingsley (1995), Hall and Heyde (1980), and Shorack (2000) for alternative conditions]. Because for all $\delta > 0$, $|\xi_{N,k}|^2 1_{\{|\xi_{N,k}| \geq \varepsilon\}} \varepsilon^\delta \leq |\xi_{N,k}|^{2+\delta}$, it follows that Lindeberg's condition is implied by Lyapunov's condition:

$$\sum_{k=1}^{2N} E[|\xi_{N,k}|^{2+\delta}] \to 0, \quad \text{for some } \delta > 0.$$

For the matching estimators considered in this section, Lyapunov's condition can be established imposing regularity conditions on the existence of moments (such as condition (3) in the statement of Theorem 1). Then, the central limit theorem for triangular martingale arrays implies

$$\sqrt{N_1} D_N \xrightarrow{d} N(0, \sigma^2).$$

The proof concludes by showing that $\sqrt{N_1} R_N \xrightarrow{p} 0$. If μ_0 is Lipschitz continuous, then there exists a constant c_{μ_0} such that

$$\sqrt{N_1} R_N \leq c_{\mu_0} \frac{1}{\sqrt{N_1}} \frac{1}{M} \sum_{i=1}^{N}\sum_{m=1}^{M} W_i \left\|\mathbf{U}_{N_0,N_1,i}^{(M,m)}\right\|.$$

Proposition 1 in the Appendix shows that under some conditions, and if there exists $c > 0$ and $r > k$, where k is the number of (continuous) covariates, such that $N_1^r / N_0 \leq c$, then,

$$\frac{1}{\sqrt{N_1}} \sum_{i=1}^{N}\sum_{m=1}^{M} W_i \left\|\mathbf{U}_{N_0,N_1,i}^{(M,m)}\right\| \xrightarrow{p} 0,$$

so $\sqrt{N_1} R_N$ vanishes asymptotically.

We now collect in a theorem, the result of this section along with precise regularity conditions.

Theorem 1. Suppose that (1) $\{Y_i, \mathbf{X}_i, W_i\}_{i=1}^N$ is a pooled sample of N_1 treated and N_0 untreated observations obtained by random sampling from their respective population counterparts, (2) the support of $\mathbf{X}$ given $W = 1$, is a subset of the support of $\mathbf{X}$ given $W = 0$, (3) for some $\delta > 0$, and $w = 0, 1$, $E[|Y|^{2+\delta} | \mathbf{X} = \mathbf{x}, W = w]$ is bounded on the support of $\mathbf{X}$ given $W = w$, (4) the functions $\mu_0(\cdot)$ and $\sigma_0^2(\cdot)$ are Lipschitz continuous, and (5) $(1/\sqrt{N_1}) \sum_{i=1}^N \sum_{m=1}^M W_i \|\mathbf{U}_{N_0,N_1,i}^{(M,m)}\| \xrightarrow{p} 0$ as $N_1 \to \infty$. Then, $\sqrt{N_1}(\widehat{\tau} - \tau) \xrightarrow{d} N(0, \sigma^2)$ as $N_1 \to \infty$.

Assumption (5) in Theorem 1 is not primitive and Proposition 1 in the Appendix provides a set of primitive regularity conditions under which assumption (5) holds. The conditions of Proposition 1 assume that all covariates have continuous distributions. This is done without loss of generality for large enough samples. As sample sizes increase, discrete covariates with a finite number of support points are perfectly matched, so they can be easily dealt with by conditioning on their values, in which case k is equal to the number of continuous covariates in $\mathbf{X}$. In practice, however, discrete covariates may not be perfectly matched, and may therefore contribute to the bias of the matching estimator.

The proof of Proposition 1 indicates that the support conditions in this proposition can also be relaxed. However, the requirement that the size of the untreated group is of a larger order of magnitude than the size of the treated group (implied by $N_1^r / N_0 \leq c$ for $c > 0$ and $r > k$) is crucial to the result in the proposition. To see that $r = 1$ is not sufficient (even in the one-dimensional case where $k = 1$), consider the case with $M = 1$ and $N_0 = N_1$. Then, because matching is done without replacement and all treated units are matched, the matching estimator is equal to the difference in sample means of Y between treated and untreated units, regardless of the total sample size N.

Proposition 1 provides the conditions under which matching discrepancies are negligible in large samples. In practical terms, Proposition 1 demonstrates the benefits of having a large "donor pool" of control units for matching estimators. However, for any given sample, matching discrepancies are observed, and researchers can assess the quality of the matches directly from the data. If present, large matching discrepancies indicate that the magnitude of the bias term, R_N, is potentially large, in which case the large sample approximation of Theorem 1 is not warranted. Therefore, the size of the matching discrepancies should be conscientiously described (e.g., using the techniques in Abadie and Imbens 2011) before the result in Theorem 1 is applied.

When matching discrepancies are large, the resulting bias can be eliminated or reduced using the bias correction techniques in Rubin (1973b), Quade (1982), and Abadie and Imbens (2011). These authors proposed a bias-corrected matching estimator that adjusts each matched pair for its contribution to the conditional bias term:

$$\widehat{\tau}_{\text{bc}} = \frac{1}{N_1}\sum_{i=1}^{N} W_i \left((Y_i - \widehat{\mu}_0(\mathbf{X}_i)) - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} (Y_j - \widehat{\mu}_0(\mathbf{X}_j)) \right), \quad (4)$$

where $\widehat{\mu}_0(\cdot)$ is an estimator of $\mu_0(\cdot)$. Under certain conditions, Abadie and Imbens (2011) showed that this bias-correction technique eliminates the asymptotic bias of a matching with replacement estimator without affecting its asymptotic variance.

Straightforward calculations show that the variance estimator

$$\widehat{\sigma}^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N} W_i \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j - \widehat{\tau} \right)^2 \tag{5}$$

is consistent for σ^2. Despite the simplicity of this result, to our knowledge, the validity of $\widehat{\sigma}^2/N_1$ as an estimator of the variance of $\widehat{\tau}$ when matching is done without replacement has not been established previously. Conversely, it is known that $\widehat{\sigma}^2/N_1$ is not a valid estimator of the variance of $\widehat{\tau}$ when matching is done with replacement (Abadie and Imbens 2006).

5. APPLICATION: HOT-DECK IMPUTATION

In this section, we consider a "cell hot-deck" imputation scheme, where incomplete records of Y are imputed using complete observations within the same "cell" of the covariates, $\mathbf{X}$. That is, the support of the covariates is partitioned into T cells, $\mathcal{C}_1, \dots, \mathcal{C}_T$, and each incomplete record of Y is filled using a complete record from the same cell. Other hot-deck imputation procedures are possible (see, e.g., Little and Rubin 2002). However, the cell hot-deck method is probably the most widely used in practice, as it is the one used by the U.S. Census Bureau to impute missing data in the CPS, the decennial census, the Survey of Income and Program Participation (SIPP), and other large surveys. Derivations similar to the ones presented in this section can be applied to alternative hot-deck imputation schemes.

Let W be an indicator for a complete record, that is, $W = 1$ indicates that Y is observed. Cell hot-deck imputation methods, such as the one employed in the CPS, can be justified by the assumption that Y is independent of $(\mathbf{X}, W)$ conditional on $\mathbf{X} \in \mathcal{C}_t$, for $1 \leq t \leq T$. This is sometimes referred to as the cell mean model assumption (Brick, Kalton, and Kim 2004). This may be a strong assumption in many contexts where data are imputed using the cell hot deck. However, without this assumption or a similar one, in general, the cell hot deck will produce inconsistent estimators. Therefore, in our analysis, we adopt the cell mean model assumption. Also, we restrict our derivations to the case of simple random sampling. Let $\mu = E[Y]$, $\mu(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$, $\mu_t = E[Y|\mathbf{X} \in \mathcal{C}_t]$, and $\sigma_t^2 = \mathrm{var}(Y|\mathbf{X} \in \mathcal{C}_t)$. Let $j(i)$ be the index of the observation used to impute Y for observation i (if $W_i = 1$, then $j(i) = i$). Let

$$\begin{aligned} \bar{Y} &= \frac{1}{N} \sum_{i=1}^{N} Y_{j(i)} \\ &= \frac{1}{N} \sum_{i=1}^{N} W_i (1 + K_{N,i}) Y_i, \end{aligned} \tag{6}$$

where now $K_{N,i}$ is the number of times that observation i is used to impute an incomplete record. The variables $K_{N,i}$ depend on how imputations are chosen from the complete records within a cell. One possibility is the *random cell hot deck*, which imputes missing records using a record chosen at random among the complete observations in the same cell. The CPS and other large surveys use a more complicated procedure called the *sequential cell hot deck*. The sequential cell hot deck imputes missing records using the last complete record in the same cell. That is, unlike the random cell hot deck, the sequential cell hot deck uses information about the order of the observations in the sample.

Note that

$$\begin{aligned} \bar{Y} - \mu = & \frac{1}{N} \sum_{i=1}^{N} (\mu(\mathbf{X}_i) - \mu) \\ & + \frac{1}{N} \sum_{i=1}^{N} W_i (1 + K_{N,i})(Y_i - \mu(\mathbf{X}_i)) \\ & + \frac{1}{N} \sum_{i=1}^{N} \left(\mu\left(\mathbf{X}_{j(i)}\right) - \mu\left(\mathbf{X}_i\right) \right). \end{aligned}$$

Under the cell mean model assumption, $\mu(\mathbf{X}_{j(i)}) - \mu(\mathbf{X}_i) = 0$, for all i. Let N_t be the number of observations in cell $\mathcal{C}_t$. Assume that the second moment of $K_{N,i}$ exists, and that for each cell t, we have

$$\left| \frac{1}{N_t} \sum_{i=1}^{N} 1_{\{\mathbf{X}_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i})^2 - E\left[\frac{1}{N_t} \sum_{i=1}^{N} 1_{\{\mathbf{X}_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i})^2 \right] \right| \xrightarrow{p} 0, \tag{7}$$

which can be usually established using negative association properties of $\{K_{N,i} \text{ s.t. } W_i = 1, \mathbf{X}_i \in \mathcal{C}_t\}$ (Joag-Dev and Proschan 1983; see Proposition 2 in the Appendix). We can write

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{N}} = \sum_{k=1}^{2N} \xi_{N,k},$$

where

$$\begin{aligned} \sigma^2 = & E\left[\sum_{t=1}^{T} \left(\frac{N_t}{N} \right) (\mu_t - \mu)^2 \right] \\ & + E\left[\sum_{t=1}^{T} \left(\frac{N_t}{N} \right) \sigma_t^2 \frac{1}{N_t} \sum_{i=1}^{N} 1_{\{\mathbf{X}_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i})^2 \right], \end{aligned}$$

and

$$\xi_{N,k} = \begin{cases} \dfrac{1}{\sigma\sqrt{N}} (\mu(\mathbf{X}_k) - \mu), & \text{if } 1 \leq k \leq N, \\ \dfrac{1}{\sigma\sqrt{N}} W_{k-N} (1 + K_{N,k-N}) \\ \quad \times (Y_{k-N} - \mu(\mathbf{X}_{k-N})), & \text{if } N+1 \leq k \leq 2N. \end{cases}$$

Let $\mathcal{X}_N = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ and $\mathcal{W}_N = \{W_1, \dots, W_N\}$. Consider the σ-fields $\mathcal{F}_{N,k} = \sigma\{W_1, \dots, W_k, \mathbf{X}_1, \dots, \mathbf{X}_k\}$ for $1 \leq k \leq N$ and $\mathcal{F}_{N,k} = \sigma\{\mathcal{W}_N, \mathcal{X}_N, Y_1, \dots, Y_{k-N}\}$ for $N+1 \leq k \leq 2N$. Then,

$$\left\{ \sum_{j=1}^{i} \xi_{N,j}, \mathcal{F}_{N,i}, 1 \leq i \leq 2N \right\}$$

is a martingale for each $N \geq 1$. Equation (7) along with the central limit theorem for martingale arrays (e.g., theorem 35.12 in Billingsley 1995) imply

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{N}} \xrightarrow{d} N(0, 1). \tag{8}$$

We now present the result of this section in the form of a theorem, along with precise regularity conditions.

Theorem 2. Suppose that (1) $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}_{i=1}^N$ are sampled at random from the population of interest, (2) $\Pr(W = 1|\mathbf{X} \in \mathcal{C}_t) > 0$, for $t = 1, \ldots, T$, (3) Y is independent of $(W, \mathbf{X})$, given $\mathbf{X} \in \mathcal{C}_t$, for $t = 1, \ldots, T$, (4) $\text{var}(Y) > 0$, and (5) for some $\delta > 0$, $E[|Y|^{2+\delta}] < \infty$. Then, Equation (8) holds.

Consider now the usual variance estimator that ignores missing data imputation:

$$\widehat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{j(i)} - \bar{Y}\right)^2. \tag{9}$$

Note that

$$\left|\widehat{\sigma}^2 - \sum_{t=1}^{T}\left(\frac{N_t}{N}\right)(\mu_t - \mu)^2 - \sum_{t=1}^{T}\left(\frac{N_t}{N}\right)\sigma_t^2\right| \xrightarrow{p} 0.$$

In addition, because $\sum_{i=1}^N 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1 + K_{N,i}) = N_t$, then

$$\frac{1}{N_t}\sum_{i=1}^{N} 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1 + K_{N,i})^2 = 1 + \frac{1}{N_t}\sum_{i=1}^{N} 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i\left(K_{N,i}^2 + K_{N,i}\right).$$

This suggests using the following estimator of the variance of the rescaled estimator:

$$\begin{aligned}\widehat{\sigma}^2_{\text{adj}} &= \widehat{\sigma}^2 + \frac{1}{N}\sum_{t=1}^{T}\left(\sum_{i=1}^{N} 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i\left(K_{N,i}^2 + K_{N,i}\right)\right)\widehat{\sigma}_t^2 \\ &= \widehat{\sigma}^2 + \sum_{t=1}^{T}\left(\frac{N_t}{N}\right)\left(\frac{1}{N_t}\sum_{i=1}^{N} 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i\left(K_{N,i}^2 + K_{N,i}\right)\right)\widehat{\sigma}_t^2,\end{aligned} \tag{10}$$

where $\widehat{\sigma}_t^2$ is the sample variance of Y calculated from the complete observations in cell $\mathcal{C}_t$. Similar formulas of the estimator of the variance in contexts different from the one considered in this section have been previously derived by Hansen, Hurwitz, and Madow (1953, vol. II, pp. 139–140), Kalton (1983), and Brick, Kalton, and Kim (2004). Note that this formula applies no matter how imputation is done within the cells (e.g., randomized or based on the order of the observations in the sample) as long as Equation (7) holds.

6. MONTE CARLO ANALYSIS

This section reports the results of two Monte Carlo simulations based on actual data. Section 6.1 uses the Boston U.S. Home Mortgage Disclosure Act (HMDA) dataset, a dataset collected by the Federal Reserve Bank of Boston to investigate racial discrimination in mortgage credit markets, to assess the quality of the large sample approximation to the distribution of matching estimators derived in Section 4. Section 6.2 uses CPS data to investigate the performance of the standard error correction for missing data imputation derived in Section 5.

6.1 Matching Without Replacement in the Boston HMDA Dataset

To detect potential discriminatory practices of mortgage credit lenders against minority applicants, the HMDA of 1975 requires lenders to routinely disclose information on mortgage applications, including the race and ethnicity of the applicants. The information collected under the HMDA does not include, however, data on the credit histories of the applicants, and other loan and applicant characteristics that are considered to be important factors in determining the approval or denial of mortgage loans. The absence of such information has generated some skepticism about whether the HMDA data can effectively be used to detect discrimination in the mortgage credit market. To overcome these deficiencies in the HMDA data, the Federal Reserve Bank of Boston collected an additional set of 38 variables included in mortgage applications for a sample of applications in the Boston metropolitan area in 1990. The Boston HMDA dataset includes all mortgage applications by black and Hispanic applicants in the Boston metropolitan area in 1990, as well as a random sample of mortgage applications by white applicants in the same year and geographical area. Regression analysis of the Boston HMDA data indicated that minority applicants were more likely to be denied mortgage than white applicants with the same characteristics (Munnell et al. 1996).

In this section, we use the Boston HMDA dataset to evaluate the empirical performance of the large sample approximation to the distribution of matching estimators derived in Section 4. HMDA data provide a relevant context for this evaluation because the Federal Reserve System employs matching in HMDA data as an screening device for fair lending regulation compliance (Avery, Beeson, and Calem 1997; Avery, Canner, and Cook 2005). We restrict our sample to single-family residences and male applicants who are white non-Hispanic or black non-Hispanic, not self-employed, who were approved for private mortgage insurance, and who do not have a public record of default or bankruptcy at the time of the application. This leaves us with a sample of 148 black applicants and 1336 white applicants, for a total of 1484 applicants.

In the context of this application, the outcome variable, Y, is an indicator variable that takes value 1 if the mortgage application was denied, and 0 if the mortgage application was approved; W is a binary indicator that takes value 1 for black applicants; and $\mathbf{X}$ is a vector of six applicant and loan characteristics used in Munnell et al. (1996): housing expense to income ratio, total debt payments to income ratio, consumer credit history, mortgage credit history, regional unemployment rate in the applicant's industry, and loan amount to appraised value ratio [see Munnell et al. (1996) for a precise definition of these variables].

To run our simulations for samples sizes of N_1 black observations and N_0 white observations, we proceed in five steps. First, for the entire sample, we estimate a logistic model of the mortgage denial indicator on the black indicator and the covariates in $\mathbf{X}$. Second, we draw (with replacement) N_1 observations from the empirical distribution of $\mathbf{X}$ for black applicants and N_0 observations from the empirical distribution of $\mathbf{X}$ for white applicants. Third, for each individual in the simulated sample, we generate the mortgage denial indicator, Y, using the logistic model estimated in the first step. Fourth, for the simulated sample, we compute $\widehat{\tau}$, the matching estimator in Equation (2), matching without replacement, the bias-corrected version of this estimator, $\widehat{\tau}_{\text{bc}}$, in Equation (4), and the variance estimator, $\widehat{\sigma}^2$, in Equation (5). All covariates are normalized to have unit variance

Table 1. Boston HMDA data simulation results: black–white difference in mortgage denial probability for matched pairs (number of simulations =10,000)

Sample sizes		Bias		Variance		Coverage of 95% CI	
		(1) $\lvert E[\widehat{\tau}]-\tau\rvert$	(2) $\lvert E[\widehat{\tau}_{\text{bc}}]-\tau\rvert$	(3) $\text{var}(\widehat{\tau})$	(4) $E[\widehat{\sigma}^2/N_1]$	(5) $\widehat{\tau}\pm 1.96\,\widehat{\sigma}/\sqrt{N_1}$	(6) $\widehat{\tau}_{\text{bc}}\pm 1.96\,\widehat{\sigma}/\sqrt{N_1}$
$N_1=25$	$N_0=250$	0.0143	0.0012	0.0091	0.0091	0.9225	0.9348
	$N_0=500$	0.0106	0.0001	0.0092	0.0091	0.9244	0.9394
	$N_0=1000$	0.0077	0.0002	0.0090	0.0091	0.9263	0.9430
$N_1=50$	$N_0=500$	0.0106	0.0011	0.0045	0.0045	0.9427	0.9458
	$N_0=1000$	0.0073	0.0009	0.0044	0.0046	0.9427	0.9456
$N_1=100$	$N_0=1000$	0.0090	0.0001	0.0023	0.0023	0.9436	0.9468

prior to matching, and a logistic model is employed to calculate the bias correction. Finally, we repeat steps two to four for a total number of 10,000 simulations. That is, in this simulation, we sample from a population distribution of the covariates that is equal to the distribution of the covariates in the HMDA sample of 1484 applicants. The distribution of Y conditional on W and $\mathbf{X}$ in our simulation is given by a logistic model with parameters equal to those estimated in the HMDA sample of 1484 applicants. In this Monte Carlo design, the parameter τ in Equation (1) is equal to 0.099, which represents the difference in the probability of denial between black applicants and white applicants of the same characteristics in our simulation.

Table 1 reports the results of the simulation, for different sample sizes N_1 and N_0. Column (1) reports the bias of $\widehat{\tau}$ relative to τ. As suggested by the results in Section 4, our simulation results indicate that for a fixed N_1, the bias of $\widehat{\tau}$ decreases when N_0 increases. For small samples, however, the bias of $\widehat{\tau}$ may be substantial, reflecting the high dimensionality of the vector of matching variables. The bias-corrected estimator in column (2) generates much smaller biases. Columns (3) and (4) report the variance of $\widehat{\tau}$ across simulations and the average, also across simulations, of the variance estimator of $\widehat{\tau}$ in Equation (5). Even in fairly small samples ($N_1=25$ and $N_0=250$), $\widehat{\sigma}^2/N_1$ provides a very precise approximation to the variance of $\widehat{\tau}$. Finally, columns (5) and (6) report coverage rates of nominal 95% confidence intervals (CIs) constructed with $(\widehat{\tau},\widehat{\sigma}^2)$ and $(\widehat{\tau}_{\text{bc}},\widehat{\sigma}^2)$, respectively. The results indicate that, in this simulation, the normal approximation to the distribution of matching estimators derived in Section 4 is very accurate, especially when the bias of the matching estimator is corrected using the bias correction techniques in Rubin (1973b), Quade (1982), and Abadie and Imbens (2011).

Because we consider matching without replacement, the untreated sample should include a subset of N_1 units that are sufficiently close in their covariate values to the N_1 units in the treatment sample. This is important, as it precludes the use of matching without replacement in some important settings. In particular, in the National Supported Work Demonstration (NSW) dataset of LaLonde (1986) and Dehejia and Wahba (1999), which is prominently used to evaluate the performance of different matching methods (Dehejia and Wahba 1999; Dehejia 2005; Smith and Todd 2005; Diamond and Sekhon 2010; Abadie and Imbens 2011), there are severe imbalances in demographic characteristics between the treated sample and the untreated sample. This makes it impossible to obtain good matches across the treated and untreated samples when matching is done without replacement on the entire list of covariates included in propensity score specification by Dehejia and Wahba (1999). In an appendix available online, we repeat the analysis in this section using a subset of the NSW treated units and matching variables for which close matches exist when matching is carried out without replacement. The results are similar to those reported here for the HMDA data.

6.2 Hot-Deck Imputation in the CPS

Hot-deck methods have long been used to impute missing data in large surveys (see, e.g., Andridge and Little 2010). However, the sampling properties of complex hot-deck imputation methods, such as the sequential hot deck used by the Census Bureau in the CPS, are largely unknown. This void in the literature has become an object of serious concern in recent years, because the proportion of observations in the CPS with imputed values of weekly earnings has increased steadily: from around 16% in 1979, when weekly earnings were included in the monthly survey questionnaire, to more than 30% in recent years (Hirsch and Schumacher 2004; Bollinger and Hirsch 2009).

In this section, we investigate the performance of the approximation to the distribution of a sample mean proposed in Section 5, when data are imputed using a sequential hot deck as in the CPS. To make our exercise as realistic as possible, we base our Monte Carlo design on actual CPS data. However, as in Section 5 and in contrast to the CPS sampling scheme, we base our simulation on simple random sampling.

Hot-deck imputation in the CPS outgoing rotation groups is carried out through a series of steps, each one imputing a specific survey item. Here, we focus on imputation of missing earnings, because earnings are affected by imputation rates that are much higher than for other survey items. Like other missing survey items, the imputation of weekly earnings for nonhourly workers is implemented through a cell hot-deck procedure. Observations are assigned to cells defined by age, race, gender, education, occupation, hours worked, and receipt of overtime wages, tips, or commissions, for a total of 11,520 cells (see Bollinger and Hirsch 2009 for details). Then, each missing record is imputed using the value of weekly earnings of the last complete record in the same cell.

Table 2. CPS data simulation results: average log weekly earnings (number of simulations =50,000)

Sample size	Variance				Coverage of 95% CI	
	(1)	(2)	(3)	Ratio (4)	(5)	(6)
N	var($\bar{Y}$)	$E[\widehat{\sigma}^2_{\text{adj}}/N]$	$E[\widehat{\sigma}^2/N]$	(3)/(1)	$\bar{Y} \pm 1.96\,\widehat{\sigma}_{\text{adj}}/\sqrt{N}$	$\bar{Y} \pm 1.96\,\widehat{\sigma}/\sqrt{N}$
50	0.0072	0.0071	0.0052	0.7262	0.9436	0.8973
100	0.0039	0.0039	0.0026	0.6701	0.9476	0.8888
200	0.0021	0.0021	0.0013	0.6342	0.9492	0.8799
856	0.0005	0.0005	0.0003	0.5834	0.9482	0.8661

The imputation of weekly earnings in the CPS outgoing rotation groups cannot be perfectly reproduced with the CPS public use data files. The main reason is that the race variable used by the imputation algorithm is different from the one included in the public use data release. Nevertheless, the Monte Carlo exercise carried out in this section is designed to reproduce as closely as possible the imputation algorithm used by the Census Bureau for weekly earnings. In our simulation, we use data from the CPS monthly file of August 2009. To simplify the analysis, we first restrict our sample to male individuals working for a pay, who are white, aged 25–64, have a high school diploma or equivalent, hold one job only, have a tertiary occupation, do not receive overtime wages, tips, or commissions, and work 40 hr per week. In addition, we discard four observations with zero recorded weekly earnings. This leaves us with 856 observations in 30 of the 11,520 original hot-deck cells. The 30 hot-deck cells are defined by three categories of age, two of education, and five of occupation. The average number of observations per cell is 28.53, the minimum is 2, and the maximum is 149. In this sample, the percentage of observations with missing weekly earnings is 32.83, and each cell has at least two complete observations.

For a given number of observations, N, the simulation proceeds as follows. First, for each cell t, we simulate two observations of log weekly earnings, $Y^*_{t,1}$ and $Y^*_{t,2}$, from a normal distribution, with the same mean and variance as in the distribution of log weekly earnings for the complete CPS observations in the same cell. In our simulation, $Y^*_{t,1}$ and $Y^*_{t,2}$ represent the last two complete observations in cell t in previous CPS waves. Second, we sample N observations from the multinomial distribution of cell frequencies in the CPS sample. For each of these N observations, we simulate log weekly earnings using a normal distribution, with the same mean and variance as in the distribution of log weekly earnings for the complete CPS observations in the same cell. Then, for each observation, we mark weekly earnings as unrecorded, with probability equal to the proportion of missing weekly earnings in the same cell of the CPS sample. Third, in our simulated sample of N observations, we impute missing log weekly earnings using the last complete observation in the cell (which may possibly be $Y^*_{t,2}$). This creates a partially imputed sample with N values of log weekly earnings. Fourth, we calculate the sample average, $\bar{Y}$, in Equation (6), as well as the usual and adjusted variance estimators, $\widehat{\sigma}^2$ and $\widehat{\sigma}^2_{\text{adj}}$, in Equations (9) and (10), respectively. To compute the intracell variances, $\widehat{\sigma}^2_t$, of Equation (10), we use all the complete simulated observations in the cell plus $Y^*_{t,1}$ and $Y^*_{t,2}$. Simulating two complete observations per cell, $Y^*_{t,1}$ and $Y^*_{t,2}$, that correspond to the last two complete observations in the cell, in previous CPS waves, allows us to compute $\widehat{\sigma}^2_t$ even for cells with no other complete observations in the simulation. Finally, we repeat steps one to four for a total number of 50,000 simulations.

The results are reported in Table 2 for sample sizes 50, 100, 200, and 856, the actual number of observations in the CPS sample. The average of our adjusted variance estimator across simulations, in column (2), closely approximates the variance of $\bar{Y}$, in column (1), even for fairly small sample sizes. In contrast, columns (3) and (4) show that the usual variance estimator is severely downward biased, and that the bias of this estimator (as a percentage of the true variance) increases with the sample size. For 856 observations, that is, the actual size of the CPS data sample used in the simulation, the usual variance estimator is only 58% of the true variance of $\bar{Y}$. Large sample sizes make possible that some observations are repeatedly used for imputation, increasing the difference between the adjusted and the unadjusted variances in Equation (10). This happens when missing observations arrive consecutively to a cell, without the observation used for imputation being "refreshed" by another complete observation. Columns (5) and (6) report coverage rates of nominal 95% CIs constructed with $\widehat{\sigma}^2_{\text{adj}}$ and $\widehat{\sigma}^2$, respectively. The results show coverage rates close to nominal coverage in column (5), when the adjusted variance estimator is used to construct CI. In contrast, CIs calculated with the usual variance estimator suffer from severe undercoverage, as reported in column (6).

7. CONCLUSION

This article establishes a martingale array representation for matching estimators. This representation allows the use of well-known martingale limit theorems to determine the large sample distribution of matching estimators. Because the martingale representation applies to a large class of matching estimators, the applicability of the methods presented in this article is very broad. Specific applications include matching estimators of average treatment effects as well as "hot-deck" imputation methods for missing data. Two realistic simulations demonstrate the empirical relevance of the results of this article.

APPENDIX

Proposition 1. Let F_0 and F_1 be the distributions of $\mathbf{X}$ given $W = 0$ and $\mathbf{X}$ given $W = 1$, respectively. Assume that F_0 and F_1 have a common support that is a Cartesian product of intervals, and that the densities $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ are bounded and bounded away from zero: $\underline{f} \leq f_0 \leq \bar{f}$ and $\underline{f} \leq f_1 \leq \bar{f}$. Assume that there exists $c > 0$ and $r > k$, where k is the number of (continuous) covariates, such that $N_1^r/N_0 \leq c$. Then,

$$\frac{1}{\sqrt{N_1}} \sum_{i=1}^{N} \sum_{m=1}^{M} W_i \left\| \mathbf{U}_{N_0,N_1,i}^{(M,m)} \right\| \xrightarrow{p} 0.$$

Proof of Proposition 1. By changing units of measurement, we can always make the support of the covariates equal to the unit k-cube. (This only adds a multiplicative constant to our bounds.) Note that we can always divide a unit k-cube into N_1^k identical cubes, for $N_1 = 1, 2, 3, \ldots$.

Divide the support of F_0 and F_1 into N_1^k identical cubes. Let Z_{M,N_0,N_1} be the number of such cells where the number of untreated observations is less than M times the number of observations from the treated sample. Let M_{N_1} be the maximum number of observations from the treated sample in a single cell. Let m_{N_0,N_1} be the minimum number of untreated observations in a single cell. Note that for any series, $f(N_1)$, such that $1 \leq f(N_1) < N_1$, we have

$$\begin{aligned}
\Pr(Z_{M,N_0,N_1} > 0) &\leq \sum_{n=1}^{N_1} \Pr\left(m_{N_0,N_1} < Mn\right) \Pr\left(M_{N_1} = n\right) \\
&\leq \sum_{n=1}^{\lfloor f(N_1) \rfloor} \Pr\left(m_{N_0,N_1} < Mn\right) \Pr\left(M_{N_1} = n\right) \\
&\quad + \sum_{n=\lfloor f(N_1) \rfloor + 1}^{N_1} \Pr\left(m_{N_0,N_1} < Mn\right) \Pr\left(M_{N_1} = n\right) \\
&\leq f(N_1) \Pr\left(m_{N_0,N_1} < M f(N_1)\right) \\
&\quad + (N_1 - f(N_1)) \Pr(M_{N_1} > f(N_1)).
\end{aligned}$$

Let $D_{N_1,n}$ be the number of cells where the number of treated observations is larger than n. Let $0 < \alpha < \min\{r-k, 1\}$. Consider $f(N_1) = N_1^\alpha$. For N_1 large enough, $\bar{f}/N_1^k < 1$. Using Bonferroni's inequality, we obtain for N_1 large enough,

$$\begin{aligned}
\Pr\left(M_{N_1} > f(N_1)\right) &= \Pr\left(D_{N_1,N_1^\alpha} \geq 1\right) \\
&\leq N_1^k \Pr\left(B\left(N_1, \bar{f}/N_1^k\right) > N_1^\alpha\right),
\end{aligned}$$

where $B(N,p)$ denotes a binomial random variable with parameters (N,p). Using Bennett's bound for binomial tails (e.g., Shorack and Wellner 1986, p. 440), we obtain

$$\begin{aligned}
&\Pr\left(B\left(N_1, \bar{f}/N_1^k\right) > N_1^\alpha\right) \\
&\quad = \Pr\left(\frac{B\left(N_1, \bar{f}/N_1^k\right) - \bar{f}/N_1^{k-1}}{\sqrt{N_1}} > \frac{N_1^\alpha - \bar{f}/N_1^{k-1}}{\sqrt{N_1}}\right) \\
&\quad \leq \exp\left\{-\frac{\bar{f}/N_1^{k-1}}{1-\bar{f}/N_1^k}\left[\frac{N_1^{\alpha+k-1}}{\bar{f}}\left(\log\left(\frac{N_1^{\alpha+k-1}}{\bar{f}}\right) - 1\right) + 1\right]\right\} \\
&\quad = \exp\left\{-\frac{1}{1-\bar{f}/N_1^k}\left[N_1^\alpha\left(\log\left(\frac{N_1^{\alpha+k-1}}{\bar{f}}\right) - 1\right) + \frac{\bar{f}}{N_1^{k-1}}\right]\right\}.
\end{aligned}$$

Similarly, let $C_{N_0,N_1,m}$ be the number of cells with less than m untreated observations. Then, using Bonferroni's inequality, we obtain

$$\begin{aligned}
\Pr\left(m_{N_0,N_1} < m\right) &= \Pr\left(C_{N_0,N_1,m} \geq 1\right) \\
&\leq \sum_{n=1}^{N_1^k} \Pr(B(N_0, p_n) < m),
\end{aligned}$$

where p_n is the probability that an untreated observation falls in cell n. Then, because for all n, $p_n \geq \underline{f}/N_1^k$, we obtain

$$\Pr\left(m_{N_0,N_1} < m\right) \leq N_1^k \Pr\left(B\left(N_0, \underline{f}/N_1^k\right) < m\right).$$

Also, for large enough N_1, there exists δ such that $(Mc/\underline{f})/N_1^{r-\alpha-k} < \delta < 1$. Using Chernoff's bound for the lower tail of a sum of independent Poisson trials (e.g., Motwani and Raghavan 1995, p. 70), we obtain that for large enough N_1,

$$\begin{aligned}
&\Pr\left(B\left(N_0, \underline{f}/N_1^k\right) < MN_1^\alpha\right) \\
&\quad = \Pr\left(B\left(N_0, \underline{f}/N_1^k\right) < \underline{f}\frac{N_0}{N_1^k}\frac{MN_1^{\alpha+k}}{\underline{f}N_0}\right) \\
&\quad \leq \Pr\left(B\left(N_0, \underline{f}/N_1^k\right) < \underline{f}\frac{N_0}{N_1^k}\frac{Mc/\underline{f}}{N_1^{r-\alpha-k}}\right) \\
&\quad \leq \exp\left(-\left(\underline{f}N_0/N_1^k\right)\left(1 - (Mc/\underline{f})/N_1^{r-\alpha-k}\right)^2/2\right) \\
&\quad \leq \exp\left(-\underline{f}N_1^{r-k}(1-\delta)^2/2c\right).
\end{aligned}$$

This proves an exponential bound for $\Pr(Z_{M,N_0,N_1} > 0)$.

Rearrange the observations so the first N_1 observations in the sample are the treated observations. For $1 \leq i \leq N_1$, let $\|\mathbf{U}_{N_0,N_1,i}^{(M,m)}\|$ be the mth matching discrepancy for treated unit i when untreated units are matched without replacement to treated units in such a way that the sum of the matching discrepancies is minimized. For $1 \leq i \leq N_1$, let $\|\mathbf{V}_{N_0,N_1,i}^{(M,m)}\|$ be the mth matching discrepancy for treated unit i when untreated units are matched without replacement to treated units in such a way that the matches are first done within cells and, after all possible within-cell matches are exhausted, untreated units that were not previously used as a match are matched without replacement to previously unmatched treated units in other cells. Note that

$$\sum_{i=1}^{N_1}\sum_{m=1}^{M}\left\|\mathbf{U}_{N_0,N_1,i}^{(M,m)}\right\| \leq \sum_{i=1}^{N_1}\sum_{m=1}^{M}\left\|\mathbf{V}_{N_0,N_1,i}^{(M,m)}\right\|.$$

Let $d_{N_1,k}$ be the diameter of the cells. Let C_k be the diameter of the unit k-cube. Note that if the unit k-cube is divided into N_1^k identical cells, then $C_k = N_1 d_{N_1,k}$. For $1 \leq n \leq N_1^k$, let $\mathcal{A}_{N_1,n}$ be the nth cell. Then,

$$\begin{aligned}
E\left[\left\|\mathbf{V}_{N_0,N_1,i}^{(M,m)}\right\| \mid Z_{M,N_0,N_1} = 0\right] &\leq \sum_{n=1}^{N_1^k} d_{N_1,k} \Pr\left(\mathbf{X}_{1,i} \in \mathcal{A}_{N_1,n} | Z_{N_0,N_1} = 0\right) \\
&\leq d_{N_1,k} \\
&= \frac{C_k}{N_1}.
\end{aligned}$$

Now,

$$\begin{aligned}
&E\left[\frac{1}{\sqrt{N_1}}\sum_{i=1}^{N_1}\sum_{m=1}^{M}\left\|\mathbf{U}_{N_0,N_1,i}^{(M,m)}\right\|\right] \\
&\quad \leq E\left[\frac{1}{\sqrt{N_1}}\sum_{i=1}^{N_1}\sum_{m=1}^{M}\left\|\mathbf{V}_{N_0,N_1,i}^{(M,m)}\right\|\right] \\
&\quad = E\left[\frac{1}{\sqrt{N_1}}\sum_{i=1}^{N_1}\sum_{m=1}^{M}\left\|\mathbf{V}_{N_0,N_1,i}^{(M,m)}\right\| \Big| Z_{M,N_0,N_1} = 0\right] \Pr\left(Z_{M,N_0,N_1} = 0\right) \\
&\qquad + E\left[\frac{1}{\sqrt{N_1}}\sum_{i=1}^{N_1}\sum_{m=1}^{M}\left\|\mathbf{V}_{N_0,N_1,i}^{(M,m)}\right\| \Big| Z_{M,N_0,N_1} > 0\right] \Pr\left(Z_{M,N_0,N_1} > 0\right) \\
&\quad \leq M\frac{C_k}{\sqrt{N_1}} + \sqrt{N_1}MC_k \Pr\left(Z_{M,N_0,N_1} > 0\right) \longrightarrow 0.
\end{aligned}$$

Markov's inequality produces the desired result. □

Proof of Theorem 1. Note that condition (3) in Theorem 1 implies that for $w = 0, 1$, $\mu_w(\mathbf{x})$ and $\sigma_w^2(\mathbf{x})$ are bounded on the support of $\mathbf{X}$, given $W = w$. Then, the result of the theorem follows easily from the derivations in Section 4. □

Before proving Theorem 2, it is useful to prove the following proposition.

Proposition 2. Let

$$A_{N,t} = \frac{1}{N}\sum_{i=1}^{N} 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^2 - E\left[\frac{1}{N}\sum_{i=1}^{N} 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^2\right].$$

Under the conditions of Theorem 2, we have $A_{N,t} \xrightarrow{p} 0$, for all $t = 1, 2, \ldots, T$.

Proof of Proposition 2. Given the nature of the sequential hot deck, it is easy to check that for any N and i, the positive moments of $K_{N,i}$ conditional on $\mathbf{X}_i \in \mathcal{C}_t$ are bounded by the corresponding moments of a geometric distribution with parameter $\Pr(W = 1|\mathbf{X} \in \mathcal{C}_t)$. Therefore, we obtain that for any $r > 0$, there exists a constant c_r such that $E[K_{N,i}^r] \le c_r$ for all N and i.

Because $E[A_{N,t}] = 0$, Markov's inequality implies that if $\operatorname{var}(A_{N,t}) \to 0$, then $A_{N,t} \xrightarrow{p} 0$.

$$\begin{aligned}\operatorname{var}(A_{N,t}) &= \operatorname{var}\left(\frac{1}{N}\sum_{i=1}^{N} 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^2\right)\\ &= \frac{1}{N^2}\sum_{i=1}^{N}\operatorname{var}\left(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^2\right)\\ &\quad + \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j>i}\operatorname{cov}\Big(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^2,\\ &\quad 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j})^2\Big).\end{aligned}$$

To show that $\operatorname{var}(A_{N,t})$ converges to zero, we will first prove the following intermediate result: for all $i = 1, \ldots, N-1$, all $j = i+1, \ldots, N$, and all $p \ge 0$, $\Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1 \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p) \ge \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)$. To prove this result, note that

$$\Pr((1+K_{N,i}) > p|W_i = 1, \mathbf{X}_i \in \mathcal{C}_t) = \Pr(W = 0|\mathbf{X}\in\mathcal{C}_t)^p \Pr\left(\sum_{k=i+1}^{N} 1_{\{\mathbf{X}_k\in\mathcal{C}_t\}} \ge p\right).$$

Therefore,

$$\begin{aligned}&\Pr\left(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) > p\right)\\ &\quad= \Pr(W = 0|\mathbf{X}\in\mathcal{C}_t)^p \Pr\left(\sum_{k=i+1}^{N} 1_{\{\mathbf{X}_k\in\mathcal{C}_t\}} \ge p\right)\\ &\qquad\times \Pr(W_i = 1|\mathbf{X}_i\in\mathcal{C}_t)\Pr(\mathbf{X}_i\in\mathcal{C}_t).\end{aligned}$$

Similarly,

$$\begin{aligned}&\Pr(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) > p \mid 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)\\ &\quad= \Pr(W = 0|\mathbf{X}\in\mathcal{C}_t)^p \Pr\left(\sum_{k=i+1}^{j-1} 1_{\{\mathbf{X}_k\in\mathcal{C}_t\}} \ge p\right)\\ &\qquad\times \Pr(W_i = 1|\mathbf{X}_i\in\mathcal{C}_t)\Pr(\mathbf{X}_i\in\mathcal{C}_t).\end{aligned}$$

Now, because

$$\Pr\left(\sum_{k=i+1}^{j-1} 1_{\{\mathbf{X}_k\in\mathcal{C}_t\}} \ge p\right) \le \Pr\left(\sum_{k=i+1}^{N} 1_{\{\mathbf{X}_k\in\mathcal{C}_t\}} \ge p\right),$$

we obtain that

$$\begin{aligned}&\Pr(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) > p \mid 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)\\ &\qquad\le \Pr(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) > p),\end{aligned}$$

or equivalently,

$$\begin{aligned}&\Pr(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p \mid 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)\\ &\qquad\ge \Pr(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p).\end{aligned}$$

By Bayes' theorem,

$$\frac{\Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1 \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p)}{\Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)} = \frac{\Pr(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p \mid 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)}{\Pr(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p)}$$

and we therefore obtain the desired result,

$$\Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1 \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p) \ge \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1). \tag{A.1}$$

We will now show that, for all $i = 1, \ldots, N-1$ and all $j = i+1, \ldots, N$, $\operatorname{cov}(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^2, 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j})^2) \le 0$. Consider two units i and j, with $j > i$. Note that because of the sequential nature of hot-deck imputation, $K_{N,j}$ is independent of $(W_i, K_{N,i})$ conditional on W_j. Therefore,

$$\begin{aligned}&\Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j}) \le q \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p)\\ &\quad= \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j}) \le q \mid 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1, 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}}\\ &\qquad\times W_i(1+K_{N,i}) \le p)\ \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1 \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}}\\ &\qquad\times W_i(1+K_{N,i}) \le p)\\ &\qquad+ \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j}) \le q \mid 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 0, 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}}\\ &\qquad\times W_i(1+K_{N,i}) \le p)\ \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 0 \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}}\\ &\qquad\times W_i(1+K_{N,i}) \le p)\\ &\quad= \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j}) \le q \mid 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)\\ &\qquad\times \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1 \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p)\\ &\qquad+ \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 0 \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p)\\ &\quad= 1 - \Big(1 - \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j}) \le q \mid 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)\Big)\\ &\qquad\times \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1 \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p).\end{aligned}$$

Now, because $\Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1 \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p) \ge \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)$ [Equation (A.1)], we obtain

$$\begin{aligned}&\Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j}) \le q \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) \le p)\\ &\quad\le 1 - (1 - \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j}) \le q \mid 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)\\ &\qquad\times \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j = 1)\\ &\quad= \Pr(1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j}) \le q).\end{aligned}$$

As a result, the variables $1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j})$ and $1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j})$ are negative quadrant dependent and, therefore, negatively associated (Joag-Dev and Proschan 1983). Furthermore, because increasing transformations of negatively associated random variables are also negatively associated (Joag-Dev and Proschan 1983), we obtain

$$\operatorname{cov}(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^2, 1_{\{\mathbf{X}_j\in\mathcal{C}_t\}} W_j(1+K_{N,j})^2) \le 0,$$

for all $i = 1, \ldots, N$ and all $j = i+1, \ldots, N$. This result implies

$$\operatorname{var}(A_{N,t}) \le \frac{1}{N^2}\sum_{i=1}^{N}\operatorname{var}\left(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^2\right). \tag{A.2}$$

To finish the proof, we will show that $\operatorname{var}(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^2)$ is uniformly bounded in (i, N). Because

$$\begin{aligned}\operatorname{var}\Big(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^2\Big) &\le E\Big[1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i})^4\Big]\\ &= E\Big[(1+K_{N,i})^4 \mid 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i = 1\Big]\\ &\quad\times \Pr(1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i = 1),\end{aligned}$$

and because $E[K_{N,i}^4 | 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i = 1]$ is uniformly bounded in (i, N), we obtain $\operatorname{var}(A_{N,t}) \to 0$. □

Proof of Theorem 2. First, note that, because $(1+K_{N,i})^2 \geq (1+K_{N,i})$ and $\sum_{i=1}^{N} 1_{\{\mathbf{X}_i\in\mathcal{C}_t\}} W_i(1+K_{N,i}) = N_t$, we obtain

$$\sigma^2 \geq E\left[\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)(\mu_t-\mu)^2\right] + E\left[\sum_{t=1}^{T}\sigma_t^2\frac{1}{N}\sum_{i=1}^{N}1_{\{\mathbf{X}_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})\right] = E\left[\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)(\mu_t-\mu)^2\right] + E\left[\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)\sigma_t^2\right] = \mathrm{var}(Y) > 0,$$

and the sequence $\{\xi_{N,k}\}_{k=1}^{2N}$ is well defined. Now, applying Proposition 2, we obtain

$$\sum_{k=1}^{2N} E\left[\xi_{N,k}^2|\mathcal{F}_{N,k-1}\right] = \frac{1}{\sigma^2 N}\sum_{k=1}^{N}E[(\mu(\mathbf{X}_k)-\mu)^2] + \frac{1}{\sigma^2 N}\sum_{k=N+1}^{2N}\sum_{t=1}^{T}1_{\{\mathbf{X}_{k-N}\in\mathcal{C}_t\}}W_{k-N}(1+K_{N,k-N})^2\sigma_t^2 = \frac{1}{\sigma^2}E\left[\frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}1_{\{\mathbf{X}_k\in\mathcal{C}_t\}}(\mu_t-\mu)^2\right] + \frac{1}{\sigma^2}\sum_{t=1}^{T}\sigma_t^2\frac{1}{N}\sum_{k=1}^{N}1_{\{\mathbf{X}_k\in\mathcal{C}_t\}}W_k(1+K_{N,k})^2 \xrightarrow{p} 1.$$

Jensen's inequality implies $E[|\mu(\mathbf{X}_i)|^{2+\delta}] \leq E[|Y_i|^{2+\delta}] < \infty$. Because $E[|Y_i-\mu(\mathbf{X}_i)|^{2+\delta}] < \infty$ and because all positive moments of $K_{N,i}$ are bounded (uniformly in N and i), Hölder's inequality implies that $E[W_i(1+K_{N,i})^{2+\delta/2}|Y_i-\mu(\mathbf{X}_i)|^{2+\delta/2}]$ is bounded (uniformly in N and i). As a result, we obtain the Lyapunov's condition

$$\sum_{k=1}^{2N}E\left[\xi_{N,k}^{2+\delta/2}\right] \to 0.$$

The result of Theorem 2 follows now from Theorem 35.12 in Billingsley (1995).

□

SUPPLEMENTARY MATERIALS

In the online supplementary materials we repeat the analysis of Section 6.1 using a subset of the NSW data from Dehejia and Wahba (1999).

[Received April 2010. Revised January 2012.]

REFERENCES

Abadie, A., and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267. [833,835,837]

——— (2008), "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, 76, 1537–1558. [833]

——— (2010), "Matching on the Estimated Propensity Score," NBER Working Paper No. 15301, National Bureau of Economic Research. [834]

——— (2011), "Bias Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics*, 29, 1–11. [836,839]

Andridge, R. R., and Little, R. J. A. (2010), "A Review of Hot Deck Imputation for Survey Non-Response," *International Statistical Review*, 78, 40–64. [833,839]

Avery, R. B., Beeson, P. E., and Calem, P. S. (1997), "Using HMDA Data as a Regulatory Screen for Fair Lending Compliance," *Journal of Financial Services Research*, 11, 9–42. [838]

Avery, R. B., Canner, G. B., and Cook, R. E. (2005), "New Information Reported Under HMDA and Its Application in Fair Lending Enforcement," *Federal Reserve Bulletin*, 91, 344–394. [838]

Billingsley, P. (1995), *Probability and Measure* (3rd ed.), New York: Wiley. [833,836,837,843]

Bollinger, C. R., and Hirsch, B. T. (2009), "Wage Gap Estimation With Proxies and Nonresponse," *Proceedings of the 2009 North American Summer Meetings of the Econometric Society*; mimeo. [833,839]

Brick, J. M., Kalton, G., and Kim, J. K. (2004), "Variance Estimation With Hot Deck Imputation Using a Model," *Survey Methodology*, 30, 57–66. [837,838]

Dehejia, R. (2005), "Practical Propensity Score Matching: A Reply to Smith and Todd," *Journal of Econometrics*, 125, 355–364. [839]

Dehejia, R., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062. [833,839]

Diamond, A., and Sekhon, J. S. (2010), "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," *Review of Economics and Statistics*, forthcoming. [834,839]

Gu, X. S., and Rosenbaum, P. R. (1993), "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms," *Journal of Computational and Graphical Statistics*, 2, 405–420. [834]

Hall, P., and Heyde, C. C. (1980), *Martingale Limit Theory and its Applications*, New York: Academic Press. [833,836]

Hansen, B. B. (2004), "Full Matching in an Observational Study of Coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618. [833,834]

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), *Sample Survey Methods and Theory*, New York: Wiley. [838]

Heckman, J., Ichimura, H., and Todd, P. (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294. [833,834]

Hirsch, B. T., and Schumacher, E. J. (2004), "Match Bias in Wage Gap Estimates Due to Earnings Imputation," *Journal of Labor Economics*, 22, 689–722. [839]

Iacus, S. M., King, G., and Porro, G. (2009), "Causal Inference Without Balance Checking: Coarsened Exact Matching," *Political Analysis*, forthcoming. [834]

Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4–29. [834]

Joag-Dev, K., and Proschan, F. (1983), "Negative Association of Random Variables With Applications," *The Annals of Statistics*, 11, 286–295. [837,842]

Kalton, G. (1983), "Compensating for Missing Survey Data," Research Report Series, Institute for Social Research, The University of Michigan, Ann Arbor, MI. [838]

LaLonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *American Economic Review*, 76, 604–620. [839]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley-Interscience. [833,837]

Motwani, R., and Raghavan, P. (1995), *Randomized Algorithms*, New York: Cambridge University Press. [841]

Munnell, A. H., Tootell, G. M. B., Browne, L. E., and McEneaney, J. (1996), "Mortgage Lending in Boston: Interpreting HMDA Data," *American Economic Review*, 86, 25–53. [838]

Quade, D. (1982), "Nonparametric Analysis of Covariance by Matching," *Biometrics*, 38, 597–611. [836,839]

Rosenbaum, P. R. (2002), *Observational Studies* (2nd ed.), New York: Springer. [833,834]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [834]

Rubin, D. B. (1973a), "Matching to Reduce Bias in Observational Studies," *Biometrics*, 29, 159–183. [833]

——— (1973b), "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies," *Biometrics*, 29, 185–203. [836,839]

——— (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [834]

——— (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26. [833,834]

——— (2006), *Matched Sampling for Causal Effects*, New York: Cambridge University Press. [834]

Shorack, G. R. (2000), *Probability for Statisticians*, New York: Springer. [833,836]

Shorack, G. R., and Wellner, J. A. (1986), *Empirical Processes With Applications to Statistics*, New York: Wiley. [841]

Smith, J. A., and Todd, P. E. (2005), "Does Matching Overcome LaLondes Critique of Nonexperimental Estimators?," *Journal of Econometrics*, 125, 305–353. [839]