Taisuke Otsu and Yoshiyasu Rai

Bootstrap inference of matching estimators for average treatment effects

Article (Accepted version)
(Refereed)

Original citation:
Otsu, Taisuke and Rai, Yoshiyasu (2016) *Bootstrap inference of matching estimators for average treatment effects*. Journal of the American Statistical Association ISSN 0162-1459

DOI: 10.1080/01621459.2016.1231613

© 2017 Taylor & Francis

This version available at: http://eprints.lse.ac.uk/84170/

Available in LSE Research Online: September 2017

BOOTSTRAP INFERENCE OF MATCHING ESTIMATORS FOR AVERAGE TREATMENT EFFECTS

TAISUKE OTSU AND YOSHIYASU RAI

Abstract. Abadie and Imbens (2008) showed that the naive bootstrap is not asymptotically valid for a matching estimator of the average treatment effect with a fixed number of matches. In this article, we propose asymptotically valid inference methods for matching estimators based on the weighted bootstrap. The key is to construct bootstrap counterparts by resampling based on certain linear forms of the estimators. Our weighted bootstrap is applicable for the matching estimators of both the average treatment effect and its counterpart for the treated population. Also, by incorporating a bias correction method in Abadie and Imbens (2011), our method can be asymptotically valid even for matching based on a vector of covariates. A simulation study indicates that the weighted bootstrap method is favorably comparable with the asymptotic normal approximation by Abadie and Imbens (2006). As an empirical illustration, we apply the proposed method to the National Supported Work data.

1. Introduction

The method of matching is widely applied in empirical research for treatment effects and program evaluations. In a series of papers, Abadie and Imbens (2006, 2008, 2011, 2012) studied various aspects of matching estimators for average treatment effects with fixed numbers of matches. In contrast to other studies on nonparametric estimation of treatment effects (e.g., Heckman, Ichimura and Todd, 1998, and Hirano, Imbens and Ridder, 2003), Abadie and Imbens analyzed rather nonstandard behaviors of the matching

We thank Joachim Freyberger, Jack Porter, Xiaoxia Shi and Christopher Taber for thoughtful comments. We also thank Martin Huber for a computer program. We acknowledge substantial comments from an anonymous referee regarding the general weighted bootstrap scheme. Financial support from the ERC Consolidator Grant (SES-0720961) is gratefully acknowledged (Otsu).

estimators due to lack of smoothness of the functional forms caused by the fixed numbers of matches. Abadie and Imbens (2006) showed that the matching estimators are not $\sqrt{N}$-consistent in general, where N is the sample size. Abadie and Imbens (2011) proposed a bias correction method based on nonparametric series regression. In addition to the non-standard asymptotic behavior of the point estimator, Abadie and Imbens (2008) provided an example to show that the standard naive bootstrap (i.e., resampling from observations with equal weights) fails to provide an asymptotically valid standard error and quantiles for a matching estimator. As Abadie and Imbens (2008) argued, the main reason for this failure is that the naive bootstrap fails to reproduce the distribution of the number of times each unit is used as a match.[1] Given this negative result on the bootstrap, Abadie and Imbens (2008) recommended using the asymptotic standard error derived in Abadie and Imbens (2006) or subsampling (Politis and Romano, 1994) for inference.

In this paper, we propose an alternative inference method for the matching estimators based on the weighted bootstrap (e.g., Mason and Newton, 1992, and Pauly, 2011). We show that even though the naive bootstrap is not valid for the matching estimator, the weighted bootstrap approach based on a linear form of the estimator can still be valid for inference based on the estimator. The intuition for the validity of our weighted bootstrap is that we treat the number of times unit i is used for a match (i.e., $K_M(i)$ defined in (2.1)) as one of the characteristics of unit i and directly resample it. Such resampling preserves the distribution of $K_M(i)$ in bootstrap resamples and circumvents the problem in Abadie and Imbens (2008). The weights for bootstrapping are flexible: they can be multinomial, Dirichlet (as in the Bayesian bootstrap by Rudin, 1981), or some two-point distribution (as in the wild bootstrap by Mammen, 1993). We note that the weighted bootstrap using multinomial weights (which resamples from the linear form of the estimator) is

[1] Denoted by $K_M(i)$ in (2.1) below. It means how many times unit i's observation appears to estimate potential outcomes by imputations.

different from the naive bootstrap (which resamples from the original data). Indeed, Abadie and Imbens (2008, p. 1546) mentioned the possibility of using the wild bootstrap for valid inference. As a special case, this paper formally confirms their conjecture. Also, Abadie and Imbens (2012) provided a representation of the matching estimator based on a martingale sequence of length $2N$. We argue that although it is possible to conduct the weighted bootstrap based on this martingale representation (where we draw weights of size $2N$), it is indeed enough to resample from the linear form of the estimator (where we draw weights of only size N).

We show the asymptotic validity of the weighted bootstrap for both the average treatment effect and its counterpart on the treated population. Also, by incorporating a bias correction method in Abadie and Imbens (2011), our method can be asymptotically valid even for matching based on a vector of covariates. A small simulation study indicates that our weighted bootstrap method is favorably comparable with the asymptotic normal approximation by Abadie and Imbens (2006). Finally, the proposed method is illustrated by an empirical analysis using the National Supported Work data.

The paper is organized as follows. Section 2 introduces our basic setup and notation. In Section 3, we present the weighted bootstrap method and show its asymptotic validity. Section 4 presents some simulation results. In Section 5, we apply the weighted bootstrap method to the National Supported Work data. Section 6 concludes. In Appendix A, we list assumptions for the main theorem and provide some remarks. All proofs are contained in Appendices B and C. Tables are presented in Appendix D. In the web appendix, we present an analogous result of bootstrap validity for the case of the average treatment effect on the treated population and sketch how our weighted bootstrap approach may be applied to the average derivative estimator under the small bandwidth asymptotics of Cattaneo, Crump and Jansson (2010, 2014).

2. Setup

Let us introduce the basic setup. For each unit $i = 1, \ldots, N$, we observe an indicator variable D_i for a treatment ($D_i = 1$ if treated and $D_i = 0$ otherwise), and outcome

$$Y_i = \begin{cases} Y_i(0) & \text{if } D_i = 0 \\ Y_i(1) & \text{if } D_i = 1 \end{cases},$$

where $Y_i(0)$ and $Y_i(1)$ are potential outcomes for $D_i = 0$ and 1, respectively. Also we observe a vector of covariates X_i for each unit. Based on the non-experimental observations $\{Y_i, D_i, X_i\}_{i=1}^N$ of size N, we wish to conduct inference on the average treatment effect $\tau = E[Y_i(1) - Y_i(0)]$. Let $\mathbb{I}\{A\}$ be the indicator function for an event A and $\|x\|$ be the Euclidean norm. To estimate τ, we consider the matching estimator based on the distance measured by the covariates,

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \{\hat{Y}_i(1) - \hat{Y}_i(0)\},$$

where $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ are estimates of the potential outcomes by imputations defined as

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } D_i = 0, \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } D_i = 1, \end{cases} \qquad \hat{Y}_i(1) = \begin{cases} \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } D_i = 0, \\ Y_i & \text{if } D_i = 1, \end{cases}$$

and $\mathcal{J}_M(i)$ is the set of indices of the first M matches for unit i,

$$\mathcal{J}_M(i) = \left\{ j \in \{1, \ldots, N\} : D_j = 1 - D_i, \sum_{l : D_l = 1 - D_i} \mathbb{I}\{\|X_l - X_i\| \leq \|X_j - X_i\|\} \leq M \right\}.$$

For the estimator $\hat{\tau}$, each unit may be used as a match more than once (matching with replacement). Here, the distance $\|\cdot\|$ for matching may be replaced with the weighted

Euclidean distance. Let $K_M(i)$ denote the number of times unit i is used as a match

$$K_M(i) = \sum_{l=1}^{N} \mathbb{I}\{i \in \mathcal{J}_M(l)\}. \tag{2.1}$$

In practice, it is common that the number of matches M can be small (could be one) even though the sample size N is large. To characterize behaviors of the matching estimators in such a practical scenario, Abadie and Imbens (2006) analyzed asymptotic properties of $\hat{\tau}$ as N increases to infinity with fixed M (called fixed-M asymptotics).

Let $\mu(d, x) = E[Y|D = d, X = x]$ and $\sigma^2(d, x) = Var(Y|D = d, X = x)$. Note that Assumption M (iii) in Appendix A guarantees $\mu(d, x) = E[Y(d)|X = x]$. Under Assumption M, Abadie and Imbens (2006, Theorems 3 and 4) showed that $\hat{\tau}$ is consistent and asymptotically normal, i.e.,

$$\frac{\sqrt{N}(\hat{\tau} - B_N - \tau)}{\sigma_N} \xrightarrow{d} N(0, 1), \tag{2.2}$$

as $N \to \infty$ (but M is fixed), where B_N and σ_N^2 are asymptotic bias and variance terms, respectively, defined as

$$\begin{aligned}
B_N &= \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)\left(\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\{\mu(1 - D_i, X_i) - \mu(1 - D_i, X_j)\}\right), \\
\sigma_N^2 &= \sigma_{1N}^2 + \sigma_2^2, \\
\sigma_{1N}^2 &= \frac{1}{N}\sum_{i=1}^{N}\left\{1 + M^{-1}K_M(i)\right\}^2 \sigma^2(D_i, X_i), \\
\sigma_2^2 &= E[\{(\mu(1, X_i) - \mu(0, X_i)) - \tau\}^2].
\end{aligned} \tag{2.3}$$

In empirical applications, researchers typically choose a small M even for large samples. The fixed-M asymptotics in (2.2) provides a useful approximation for the distributions of the matching estimators in such situations. A key feature of the asymptotic distribution in (2.2) is the presence of the bias term B_N that depends on M. As shown in Abadie and Imbens (2006, Theorems 1 and 2), this bias term satisfies $B_N = O_p(N^{-1/k})$, where k is the dimension of X. Therefore, if $k \geq 2$, $\hat{\tau}$ is not $\sqrt{N}$-consistent for τ.

To deal with this problem, Abadie and Imbens (2011) estimated B_N by

$$\hat{B}_N = \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)\left(\frac{1}{M}\sum_{j \in \mathcal{J}_M(i)}\{\hat{\mu}(1-D_i, X_i) - \hat{\mu}(1-D_i, X_j)\}\right), \tag{2.4}$$

where $\hat{\mu}(d,x)$ is a nonparametric estimator of $\mu(d,x)$. Requirements on $\hat{\mu}(d,x)$ are presented in Assumption R in Appendix A. Abadie and Imbens (2011, Theorem 2) employed a series estimator for $\hat{\mu}(d,x)$ and showed a remarkable result: $\sqrt{N}(\hat{B}_N - B_N) \xrightarrow{p} 0$ under certain regularity conditions allowing X to be a vector.[2] As clarified in Abadie and Imbens (2011), this surprisingly fast convergence rate follows from the fact that $\hat{B}_N$ basically estimates the contrast $\mu(d,\dot{x}) - \mu(d,x)$ with $\dot{x} - x \to 0$. However, in contrast to the number of matches M, the series length should increase to infinity to guarantee the fast convergence rate.[3] Due to the fast convergence property of $\hat{B}_N$, the bias corrected estimator $\tilde{\tau} = \hat{\tau} - \hat{B}_N$ satisfies $\sqrt{N}(\tilde{\tau} - \tau)/\sigma_N \xrightarrow{d} N(0,1)$ (Abadie and Imbens, 2011, Theorem 2). Since the asymptotic variance σ_N^2 can be consistently estimated (see, Theorem

[2] More specifically, Abadie and Imbens (2011) proposed a power series regression estimator for $\hat{\mu}(d,x)$ with k-dimensional x. For a vector of nonnegative integers $\lambda = (\lambda_1, \ldots, \lambda_k)$, let $|\lambda| = \sum_{h=1}^{k}\lambda_h$ and $x^\lambda = \prod_{h=1}^{k} x_h^{\lambda_h}$, where x_h is the h-th element of x. Define a series $\{\lambda(l)\}_{l=1}^{\infty}$ for all distinct vectors of λ such that $|\lambda(l)|$ is nondecreasing. Based on this series, consider an L-vector $p^L(x) = (x^{\lambda(1)}, \ldots, x^{\lambda(L)})'$ of power functions of x. The power series estimator of $\mu(d,x)$ is given by

$$\hat{\mu}(d,x) = p^L(x)'\left(\sum_{i:D_i=d} p^L(X_i)p^L(X_i)'\right)^{-}\sum_{i:D_i=d} p^L(X_i)Y_i,$$

for $d \in \{0,1\}$ and $x \in \mathbb{X}$, where $(\cdot)^-$ is a generalized inverse. Suppose that $\mu(d,x)$ is infinitely differentiable at $x \in \mathbb{X}$ and that the series length L grows with the sample size N and satisfies $L = O(N^v)$ with $v \in (0, \min\{2/(4k+3), 2/(4k^2-k)\})$. Then, as shown in Abadie and Imbens (2011, Theorem 2), the power series estimator $\hat{\mu}(d,x)$ satisfies Assumption R in Appendix A and achieves the fast convergence $\sqrt{N}(\hat{B}_N - B_N) \xrightarrow{p} 0$ for the bias term.

[3] A major requirement on the choice of the series length is to achieve $|\hat{\mu}(d,\cdot) - \mu(d,\cdot)|_{k-1} = o_p(N^{-1/2+1/k})$ (the last condition of Assumption R in Appendix A). When the dimension k of X is large, this condition typically requires more stringent smoothness on $\mu(d,x)$ in x. For example, Abadie and Imbens (2011, Theorem 2) imposed infinite differentiability of $\mu(d,x)$ in x.

6 of Abadie and Imbens, 2006), this asymptotic normality result on $\tilde{\tau}$ yields a confidence interval for τ, which is valid under fixed-M asymptotics.

Alternatively, one may consider bootstrap inference based on some resampling scheme. However, Abadie and Imbens (2008) provided an example showing that the naive bootstrap method (i.e., resampling from the observations $\{Y_i, D_i, X_i\}_{i=1}^N$ with uniform weights) is not valid to estimate the standard error of a matching estimator under fixed-M asymptotics. This failure is due to the fact that the naive bootstrap is not able to reproduce the distribution of $K_M(i)$ in (2.1), the number of times each unit is used as a match. To the best of our knowledge, currently there is no valid bootstrap procedure to approximate the distribution of $\sqrt{N}(\tilde{\tau} - \tau)$ under fixed-M asymptotics. Indeed, Abadie and Imbens (2008, p. 1546) conjectured that a wild bootstrap method may be asymptotically valid because of the fact that $\tilde{\tau}$ can be written by some linear forms. In this paper, we confirm and generalize their conjecture by developing a new weighted bootstrap method that is valid under fixed-M asymptotics and contains the wild bootstrap as a special case.

3. Weighted Bootstrap

In this section, we present a valid bootstrap method for the bias corrected matching estimator $\tilde{\tau} = \hat{\tau} - \hat{B}_N$ under fixed-M asymptotics. By the definitions of $\hat{\tau}$ and $\hat{B}_N$, and the fact that $\hat{\mu}(1 - D_i, X_j) = \hat{\mu}(D_j, X_j)$ for $j \in \mathcal{J}_M(i)$, the estimator $\tilde{\tau}$ can be written as a linear form

$$\begin{aligned}
\tilde{\tau} &= \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)\left(\{Y_i - \hat{\mu}(1 - D_i, X_i)\} - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)}\{Y_j - \hat{\mu}(1 - D_i, X_j)\}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)\left(Y_i - \hat{\mu}(1 - D_i, X_i) + M^{-1}K_M(i)\{Y_i - \hat{\mu}(D_i, X_i)\}\right) \\
&\equiv \frac{1}{N}\sum_{i=1}^{N}\tilde{\tau}_i, \qquad (3.1)
\end{aligned}$$

where $K_M(i)$ is defined in (2.1). Let $\hat{e}_i = Y_i - \hat{\mu}(D_i, X_i)$ and $\hat{\xi}_i = (2D_i - 1)\{\hat{\mu}(D_i, X_i) - \hat{\mu}(1 - D_i, X_i)\} - \tilde{\tau}$. Then we are able to write the i-th 'residual' as

$$\tilde{\tau}_i - \tilde{\tau} = (2D_i - 1)\{1 + M^{-1}K_M(i)\}\hat{e}_i + \hat{\xi}_i. \tag{3.2}$$

This expression is insightful. If we consider the population counterparts $e_i = Y_i - \mu(D_i, X_i)$ and $\xi_i = (2D_i - 1)\{\mu(D_i, X_i) - \mu(1 - D_i, X_i)\} - \tau$ of $\hat{e}_i$ and $\hat{\xi}_i$, respectively, then the variance components σ_{1N}^2 and σ_2^2 appearing in (2.3) can be written as

$$\begin{aligned} \sigma_{1N}^2 &= Var\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (2D_i - 1)\{1 + M^{-1}K_M(i)\}e_i \middle| \mathbf{D}, \mathbf{X} \right), \\ \sigma_2^2 &= Var\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i \right), \end{aligned}$$

respectively, where $\mathbf{D} = (D_1, \ldots, D_N)$ and $\mathbf{X} = (X_1, \ldots, X_N)$. This suggests that valid bootstrap inference may be possible if we treat $\{\tilde{\tau}_i\}_{i=1}^N$ like 'observations' and resample them. Precisely speaking, we construct the weighted bootstrap counterpart of $\sqrt{N}(\tilde{\tau} - \tau)$ as follows

$$\sqrt{N}T^* = \sum_{i=1}^{N} W_i^* (\tilde{\tau}_i - \tilde{\tau}), \tag{3.3}$$

where $\{W_i^*\}_{i=1}^N$ is a sequence of random variables satisfying Assumption W in Appendix A. Assumption W is general enough to include several popular bootstrap methods. For example, the nonparametric bootstrap (Efron, 1979) sets the weight as $W_i^* = M_i^*/\sqrt{N}$, where $(M_1^*, \ldots, M_N^*)$ is a multinomial random vector with N draws on N equal probability cells. The Bayesian bootstrap (Rubin, 1981) sets the weight as $W_i^* = \delta_i^*/\sqrt{N}$, where $(\delta_1^*, \ldots, \delta_N^*)$ is drawn from a Dirichlet distribution. The wild bootstrap (Wu, 1986, and Mammen, 1993) sets the weight as $W_i^* = \epsilon_i^*/\sqrt{N}$, where $\{\epsilon_i^*\}_{i=1}^N$ is an i.i.d. sequence with $E[\epsilon_i^*] = 0$ and $E[\epsilon_i^{*2}] = 1$.[4] It should be noted that the nonparametric bootstrap with the weight $W_i^* = M_i^*/\sqrt{N}$ is different from the naive bootstrap (i.e., draw from the original observations $\{Y_i, D_i, X_i\}_{i=1}^N$ with equal weights) investigated by Abadie and Imbens (2008). See Remark 2 below for a detailed discussion.

Our main theorem, asymptotic validity of the weighted bootstrap under fixed-M asymptotics, is presented as follows (see Appendices B and C for a proof).

Theorem. *Under Assumptions M, W, and R in Appendix A,*

$$\sup_r |\Pr\{\sqrt{N}T^* \leq r|\mathbf{Z}\} - \Pr\{\sqrt{N}(\tilde{\tau} - \tau) \leq r\}| \xrightarrow{p} 0,$$

as $N \to \infty$ with fixed M.

Remarks. 1. This theorem says that the distribution of the bootstrap statistic $\sqrt{N}T^*$ consistently estimates that of the target object $\sqrt{N}(\tilde{\tau}-\tau)$ under the Kolmogorov distance. For example, let $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$ be the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantiles of T^* respectively. These quantiles can be estimated by simulating T^*. Then based on this theorem, the $100(1-\alpha)\%$ bootstrap confidence interval of τ is obtained as $[\tilde{\tau} - q_{1-\alpha/2}^*, \tilde{\tau} - q_{\alpha/2}^*]$. In our simulation study below, we employ this confidence interval.

2. As we mentioned above, our bootstrap method allows the nonparametric bootstrap weights (i.e., $W_i^* = M_i^*/\sqrt{N}$ with multinomial M_i^*) while Abadie and Imbens (2008) showed failure of the naive bootstrap. This is not a contradiction because these methods resample different objects. The naive bootstrap draws a resample $\{Y_i^*, D_i^*, X_i^*\}_{i=1}^N$ from

[4] In the simulation and empirical studies below, we employ Mammen's (1993) two point distribution

$$\epsilon_i^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability} \quad (\sqrt{5}+1)/2\sqrt{5} \\ (\sqrt{5}+1)/2 & \text{with probability} \quad (\sqrt{5}-1)/2\sqrt{5} \end{cases}.$$

the original observations $\{Y_i, D_i, X_i\}_{i=1}^N$ and then computes the bootstrap counterpart

$$\tilde{\tau}^* = \frac{1}{N}\sum_{i=1}^{N}(2D_i^* - 1)\left(Y_i^* - \hat{\mu}(1 - D_i^*, X_i^*) + M^{-1}K_M^*(i)\{Y_i^* - \hat{\mu}(D_i^*, X_i^*)\}\right).$$

Note that $K_M^*(i)$ is computed by the bootstrap resample. As argued by Abadie and Imbens (2008), this approach causes a problem because in the naive bootstrap resample the same unit may appear multiple times, which occurs with probability 0 in the population. This event affects the number of times unit i is used as a match. As a result, the distribution of $K_M^*(i)$ fails to approximate that of $K_M(i)$, and the naive bootstrap counterpart $\tilde{\tau}^*$ fails to recover the distribution of $\tilde{\tau}$.

In contrast, the nonparametric bootstrap with $W_i^* = M_i^*/\sqrt{N}$ treats $K_M(i)$ as one of the characteristics of unit i. Indeed our bootstrap draws a resample from $\{Y_i, D_i, X_i, K_M(i)\}_{i=1}^N$ and then computes the bootstrap counterpart without recomputing $K_M(i)$. Thus, our bootstrap circumvents the above problem.

3. An insightful paper by Abadie and Imbens (2012) pointed out that the (bias corrected) matching estimators can be written as martingale sequences. For instance, consider the triangular array

$$\xi_{N,i} = \begin{cases} \frac{1}{\sqrt{N}}(2D_i - 1)\{\mu(D_i, X_i) - \mu(1 - D_i, X_i)\} - \tau & \text{for } i = 1, \dots, N \\ \frac{1}{\sqrt{N}}(2D_i - 1)\{1 + M^{-1}K_M(i)\}(Y_i - \mu(D_i, X_i)) & \text{for } i = N+1, \dots, 2N \end{cases}$$

and filtration $\mathcal{F}_{N,\ell} = \sigma\{D_1, \dots, D_\ell, X_1, \dots, X_\ell\}$ for $\ell = 1, \dots, N$ and $\mathcal{F}_{N,\ell} = \sigma\{\mathbf{D}, \mathbf{X}, Y_1, \dots, Y_{\ell-N}\}$ for $\ell = N+1, \dots, 2N$. Abadie and Imbens (2012) argued that $\left\{\sum_{i=1}^{\ell} \xi_{N,i}, \mathcal{F}_{N,\ell},\ \ell = 1, \dots, 2N\right\}$ is a martingale sequence satisfying $\sqrt{N}(\hat{\tau} - B_N - \tau) = \sum_{i=1}^{2N} \xi_{N,i}$, and that asymptotic normality of the estimator directly follows from the martingale central limit theorem. This finding is closely related to our bootstrap statistics. Indeed, since $\tilde{\tau}_i = \hat{\xi}_{N,i} + \hat{\xi}_{N,N+i}$, our bootstrap statistic in (3.3) can be written as $\sqrt{N}T^* = \sum_{i=1}^{N} W_i^*(\hat{\xi}_{N,i} + \hat{\xi}_{N,N+i})$, where $\hat{\xi}_{N,i}$ is an estimated counterpart of $\xi_{N,i}$. Also we are able to utilize the martingale structure more directly by constructing a bootstrap statistic $\sqrt{N}\tilde{T}^* = \sum_{i=1}^{2N} W_i^* \hat{\xi}_{N,i}$. Basically the same argument shows asymptotic validity of $\tilde{T}^*$.

4. Abadie and Imbens (2012) considered two other applications of the above martingale representation approach, matching without replacements and hot-deck imputation. It is easy to modify our weighted bootstrap method for those applications.

5. Instead of the weighted bootstrap method described above, one may use subsampling (Politis and Romano, 1994) as an alternative inference method. In our setting, the subsampling confidence interval for τ can be obtained as follows.

(i): Compute $\tilde{\tau}$, $\hat{\sigma}_N$, $\hat{\mu}(1, x)$, and $\hat{\mu}(0, x)$ from the entire sample.

(ii): Draw subsamples $\{Y_i^*, D_i^*, X_i^*\}_{i=1}^{S}$ without replacement from the entire sample.

(iii): Compute $\hat{\sigma}_S$ and the bias corrected estimator $\tilde{\tau}_S = \hat{\tau}_S - \hat{B}_S$ from the subsample drawn in (ii), where

$$\hat{B}_S = \frac{1}{S}\sum_{i=1}^{S}(2D_i^* - 1)\left(\frac{1}{M}\sum_{j \in \mathcal{J}_M^*(i)} \{\hat{\mu}(1 - D_i^*, X_i^*) - \hat{\mu}(1 - D_i^*, X_j^*)\}\right).$$

(iv): Repeat (ii)-(iii) for different subsamples and report $[\tilde{\tau} - \sqrt{S/N}\hat{\sigma}_N q_{1-\alpha/2}^*, \tilde{\tau} - \sqrt{S/N}\hat{\sigma}_N q_{\alpha/2}^*]$ as the $100(1-\alpha)\%$ subsampling confidence interval, where $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$ are the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantiles of $\sqrt{S}(\tilde{\tau}_S - \tilde{\tau})/\hat{\sigma}_S$, respectively.

By Politis and Romano (1994), this subsampling confidence interval is asymptotically valid under mild conditions if $S \to \infty$ and $S/N \to 0$ as $N \to \infty$. In the special case where the estimator $\tilde{\tau}$ is $\sqrt{N}$-consistent without bias adjustment, we do not need to subtract $\hat{B}_S$. Thus, unlike our weighted bootstrap approach, the subsampling method does not involve any nonparametric estimator of $\mu(d, x)$. On the other hand, subsampling requires the choice of the size of subsamples S, which may be a difficult task. Our simulation results below suggest the subsampling confidence interval is sensitive to the choice of S when the sample size N is moderate. Also, we note that the computational cost of our weighted bootstrap is significantly cheaper than that of subsampling (see, Section 4 for some example). This is because the weighted bootstrap does not require computation of matches in the bootstrap resamples.

6. As suggested by an Associate Editor, it is an interesting direction for future research to see whether the basic idea of our bootstrap approach (i.e., resample properly linearized objects by the weighted bootstrap) can be applied to other contexts, where the naive bootstrap is invalid due to lack of asymptotic linearity. In the Web Appendix, we sketch and conjecture how our weighted bootstrap approach for matching estimators may be applied to the average derivative estimator under the small bandwidth asymptotics of Cattaneo, Crump and Jansson (2010, 2014).

4. Simulation

In this section, we evaluate the finite sample performance of our weighted bootstrap method by Monte Carlo simulation. Based on Frölich (2004) and Busso, DiNardo and McCrary (2014), we consider the following data generating process for $\{Y_i, D_i, X_i\}_{i=1}^N$,

$$\begin{aligned}
Y_i(1) &= \tau + m(\|X_i\|) + \epsilon_i, \qquad Y_i(0) = m(\|X_i\|) + \epsilon_i, \\
D_i &= \mathbb{I}\{P(X_i) \geq \nu_i\}, \qquad \nu_i \sim U[0,1], \\
P(X_i) &= \gamma_1 + \gamma_2 \|X_i\|, \qquad X_i = (X_{1i}, \ldots, X_{ki})', \\
X_{ji} &= \xi_i |\zeta_{ji}| / \|\zeta_i\| \quad \text{for } j = 1, \ldots, k, \\
\xi_i &\sim U[0,1], \qquad \zeta_i \sim N(0, I_k), \qquad \epsilon_i \sim N(0, 0.2^2),
\end{aligned}$$

where $(\epsilon_i, \nu_i, \xi_i, \zeta_i)$ are mutually independent. In this simulation design, the average treatment effect is τ, which is set as $\tau = 0$. For the parameters of $P(X_i)$, we set $\gamma_1 = 0.15$ and $\gamma_2 = 0.7$.[5] For the function $m(\cdot)$, we consider six curves presented in Table 1. For $k = \dim(X)$, we consider $k = 1, 2, \ldots, 5$.

For all cases, we set the number of matches as $M = 8$,[6] and consider the bias corrected estimator $\tilde{\tau} = \hat{\tau} - \hat{B}_N$, where $\hat{B}_N$ is given by the OLS of the linear regression $\hat{\mu}_i = \hat{\alpha} + X_i\hat{\beta}$.[7] We compare five inference methods: (i) wild bootstrap (weighted bootstrap with $W_i^* = \epsilon_i^*/\sqrt{n}$, where ϵ_i^* is drawn from Mammen's (1993) two point distribution), (ii) nonparametric bootstrap (weighted bootstrap with $W_i^* = M_i^*/\sqrt{n}$), (iii) asymptotic t using Abadie and Imbens' (2006) standard error, (iv) naive bootstrap (i.e., resample from $\{Y_i, D_i, X_i\}_{i=1}^n$ with uniform weights), and (v) subsampling. Except for the naive bootstrap, all methods are valid under fixed-M asymptotics. Methods (i) and (ii) are proposed in this paper. For Method (v), we report the cases of $S = 20$, 30, 50, and a data-driven method based on Bickel and Sakov (2008) for the subsample size.[8] For each simulation design, we report coverage rates of the 95% confidence intervals by the above methods. We also report average lengths (over Monte Carlo replications) of the confidence intervals. The sample size is set as $N = 100$. Simulation results based on 10,000 replications are presented in Tables 2 and 3.

Our findings are summarized as follows. First, for all cases, the coverage probabilities of the naive bootstrap confidence intervals are below the nominal level. Since the naive

[5] We tried several different combinations of (γ_1, γ_2) and the case where ϵ_i follows the centered lognormal with the standard deviation 0.2. Results are similar overall.

[6] In a preliminary simulation study, results are similar for different values of M including $M = 1$,4, and 16.

[7] The linear regression is understood to be a special case of the power series estimator of $\mu(d, x)$ presented in Footnote 2. In a preliminary simulation study, we consider different series lengths and find that results are similar overall.

[8] More precisely, the data-driven subsample size is chosen by $S^* = \arg\min_{S_j}\{\sup_x |L_{S_j}(x) - L_{S_{j+1}}(x)|\}$, where $L_{S_j}(\cdot)$ is the subsampling cumulative distribution function of $\sqrt{S_j}(\tilde{\tau}_{S_j} - \tilde{\tau})/\hat{\sigma}_{S_j}$ and $S_j = 50 - 2j$ for $j = 0, 1, \ldots, 20$.

bootstrap is asymptotically invalid as shown by Abadie and Imbens (2008), this result is reasonable.[9] Thus, we need to employ other asymptotically valid inference methods. Second, Tables 3 shows that the coverage rates of the subsampling confidence intervals are always below the nominal level. The data-driven choice of S does not improve the under-coverage in general. Furthermore, the results are sensitive to the choice of the subsample size S. Thus, even though subsampling is asymptotically valid, it does not work well to approximate the finite sample distribution of the matching estimator in this setup. Third, across all cases, our weighted bootstrap (both wild and nonparametric) and the asymptotic t confidence intervals by Abadie and Imbens (2006) are reasonably close to the nominal level. Overall these coverage rates slightly decrease as $k = \dim(X)$ increases, but the results are not sensitive to k. For example, for Curve 1, the coverage of the wild bootstrap decreases from 0.9342 (for $k = 1$) to 0.9152 (for $k = 5$). Fourth, the average lengths of the weighted bootstrap (wild and nonparametric) are similar to those of the asymptotic t confidence intervals except for some cases where the asymptotic t shows under-coverage. Finally, for Curve 6, our weighted bootstrap methods (wild and nonparametric) show better coverage than the asymptotic t method across all cases. For example, if $k = 1$, the coverage of the wild bootstrap confidence interval is 0.9500 but for the asymptotic t is 0.8731.

Overall our simulation results are encouraging: the weighted bootstrap inference method developed in this paper is favorably comparable to the asymptotic t method in Abadie and Imbens (2006, 2011).

We close this section with a remark on the computational cost. For Curve 1 with $k = 1$ and $N = 100$, we implemented 1,000 Monte Carlo replications of the confidence intervals based on the wild and nonparametric bootstrap (with 999 bootstrap replications),

[9] Abadie and Imbens (2008) showed that the naive bootstrap can produce both under- and over-covarage. In our simulation designs, we only observe under-coverage.

asymptotic t, and subsampling (with $S = 20$) by a laptop with Intel Core i7-3540M 3.0GHz. In terms of CPU seconds, it takes 11.11 for the wild bootstrap, 11.28 for the nonparametric bootstrap, 8.85 for the asymptotic t, and 8859.17 for the subsampling. Also, for $N = 8000$, one Monte Carlo replication takes 11.16 for the wild bootstrap (with 29,999 bootstrap repetitions), 7.45 for the nonparametric bootstrap (with 29,999 bootstrap repetitions), 1.70 for the asymptotic t, and 592.70 for the subsampling (with $S = 300$). Therefore, except for subsampling, the computational costs to implement these methods are small.

5. Empirical Application

In this section, we apply our weighted bootstrap method to an empirical analysis based on the National Supported Work (NSW) data. The NSW demonstration was a program which aimed to provide subsidized work experience to individuals with longstanding employment problems. The NSW dataset was first analyzed by Lalonde (1986) and subsequently by Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002) (hereafter, DW), and Smith and Todd (2005) among others. Here we use the dataset uploaded by Rajeev Dehejia (http://users.nber.org/~rdehejia/nswdata2.html). In particular, we analyze two experimental samples and two non-experimental ones. The first experimental sample is also used by Lalonde (1986) (labelled NSW-L). It includes a male sample with ex-addict, ex-offender, and/or high school dropout having complete ex-ante and ex-post earning data. The second experimental sample is also used by DW (labelled NSW-DW). This is a subset of the first experimental sample for males who were randomly assigned in either January to April of 1976 or October of 1976 to August of 1977, and had zero earnings in the 13-24 months before the random assignment. Two non-experimental samples were also analyzed by Lalonde (1986) and DW. The first non-experimental sample is drawn from the Current Population Survey (labelled CPS) with matched Social Security earnings data. It includes all males less than 55 years old. The second non-experimental sample is drawn from the Panel Study Income Dynamics (labelled PSID). It contains all male household heads who were less than 55 years old and did not classify themselves as retired in 1975. We follow the previous studies and focus on the average treatment effect $\tau^t = E[Y_i(1) - Y_i(0)|D_i = 1]$ for the treated population. The weighted bootstrap method for τ^t and its validity are presented in the web appendix (Section A). From the experimental samples, we can obtain unbiased estimates for τ^t. We then compute the matching estimators from the experimental participants and non-experimental control groups from the CPS and PSID.

We first summarize the data. Table 6 reports the means, standard errors (in parentheses), and normalized distances between the experimental and non-experimental samples of the control groups defined as $(\bar{X}_1 - \bar{X}_0)/\sqrt{(S_0^2 + S_1^2)/2}$, where $\bar{X}_d = \sum_{i:D_i=d} X_i/N_d$ and $S_w = \sum_{i:D_i=d}(X_i - \bar{X}_d)^2/(N_d - 1)$.[10] The two experimental samples (NSW-L and NSW-DW) have similar characteristics except for real earnings in 1975 (RE75). This difference in RE75 is due to the fact that NSW-DW contains more observations with zero earnings. On the other hand, the characteristics of the experimental and non-experimental samples (particularly age, marital status, ethnicity, and earnings) are very different. This difference is confirmed by the large normalized distances of the covariates between the experimental treated and non-experimental control groups. Therefore, it is generally difficult to find good matches from these samples.

For each pair of samples for the treated and control groups, we implement the matching estimator for τ^t and then compute its standard error by Methods (i)-(iv) in Section 4.[11]

[10] This normalized distance is employed in Abadie and Imbens (2011). Note that this distance is different from the t-statistic $(\bar{X}_1 - \bar{X}_0)/\sqrt{S_0^2/N_0 + S_1^2/N_1}$ to test the null hypothesis $E[X_1] = E[X_0]$.

[11] The matching is based on the Mahalanobis distance. The estimator $\hat{\mu}(d, x)$ for the bias correction in (2.4) is given by the OLS of the linear regression $\hat{\mu}_i = \hat{\alpha} + X_i\hat{\beta}$, which is understood to be a special case

Recall that Methods (i) and (ii) are proposed in this paper, and that Method (iv) (naive bootstrap) is invalid under fixed-M asymptotics (Abadie and Imbens, 2008). For the number of matches M, we consider the cases $M = 1$, 4, and 16.

Table 7 summarizes the results for the matching estimates and standard errors. The first row presents the mean differences between the treatment and control groups. The unbiased estimates of τ^t are 886.30 and 1794.34 for the NSW-L and NSW-DW samples, respectively. The simple covariate matching works well for the NSW-DW sample, but not for the NSW-L sample. For the standard errors, we find that our weighted bootstrap methods (wild and nonparametric) provide similar values as the asymptotic method by Abadie and Imbens (2006). On the other hand, the naive bootstrap standard errors take somewhat different values compared to the others. This result suggests that statistical inference based on the naive bootstrap can be misleading and we recommend employing asymptotic or weighted bootstrap inference, which are valid under fixed-M asymptotics.

6. Conclusion

This paper proposes a weighted bootstrap inference method for matching estimators of treatment effects. In contrast to the naive bootstrap, our weighted bootstrap method is valid under asymptotics with a fixed number of matches. Our method is applicable to both the average treatment effect and its counterpart for the treated population. Simulation results indicate that the weighted bootstrap method works well in finite samples and is favorably comparable with the asymptotic normal approximation. Although it is beyond the scope of this paper, it would be interesting to investigate higher-order properties of the weighted bootstrap method under fixed-M asymptotics (see, Kline and Santos, 2012, for higher-order analysis under standard asymptotics). Also an extension of the weighted

of the power series estimator in Footnote 2. Since results are not sensitive to the length of series, we report only the case of linear regression.

bootstrap method to propensity score matching is currently under investigation by the authors.

Appendix A. Assumptions

Recall $\mu(d,x) = E[Y|D=d, X=x]$, $\sigma^2(d,x) = Var(Y|D=d,X=x)$, and $N_0 = N - N_1$. Assumptions for the main theorem are listed as follows. All limits are taken as $N \to \infty$ while M is fixed.

Assumption M. (Conditions for $\hat{\tau}$)

- **(i):** $\{Y_i, D_i, X_i\}_{i=1}^N$ *is an i.i.d. sample of* (Y, D, X).
- **(ii):** X *is continuously distributed on a compact and convex support* $\mathbb{X} \subset \mathbb{R}^k$. *The density of* X *is bounded and bounded away from zero on* $\mathbb{X}$.
- **(iii):** D *is independent of* $(Y(0), Y(1))$ *conditional on* $X = x$ *for almost every* x. *There exists a positive constant* c *such that* $\Pr\{D = 1|X = x\} \in (c, 1-c)$ *for almost every* x.
- **(iv):** *For each* $d \in \{0,1\}$, $\mu(d,x)$ *and* $\sigma^2(d,x)$ *are Lipschitz in* $\mathbb{X}$, $\sigma^2(d,x)$ *is bounded away from zero on* $\mathbb{X}$, *and* $E[Y^4|D=d,X=x]$ *is bounded uniformly on* $\mathbb{X}$.

Assumption W. (Conditions for W^*)

- **(i):** $(W_1^*, \dots, W_N^*)$ *is exchangeable and independent of* $\mathbf{Z} = (\mathbf{Y}, \mathbf{D}, \mathbf{X})$.
- **(ii):** $\sum_{i=1}^N (W_i^* - \bar{W}^*)^2 \xrightarrow{p} 1$, *where* $\bar{W}^* = N^{-1}\sum_{i=1}^N W_i^*$.
- **(iii):** $\max_{i=1,\dots,N} |W_i^* - \bar{W}^*| \xrightarrow{p} 0$.
- **(iv):** $E[W_i^{*2}] = O(N^{-1})$ *for all* $i = 1, \dots, N$.

Let $\lambda = (\lambda_1, \dots, \lambda_k)'$ be a k-dimensional vector of non-negative integers and $\partial^\lambda a(x) = \partial^{\sum_{l=1}^k \lambda_l} a(x)/\partial x_1^{\lambda_1} \cdots \partial x_k^{\lambda_k}$. Define $|a(\cdot)|_m = \max_{\sum_{l=1}^k \lambda_l \le m} \sup_{x \in \mathbb{X}} |\partial^\lambda a(x)|$.

Assumption R. (Conditions for $\mu(d,x)$)

For each $d \in \{0,1\}$ and λ satisfying $\sum_{l=1}^{k} \lambda_l = k$, the derivative $\partial^{\lambda}\mu(d,x)$ exists and satisfies $\sup_{x\in\mathbb{X}} |\partial^{\lambda}\mu(d,x)| \leq C$ for some $C > 0$. Furthermore, $\hat{\mu}(d,x)$ satisfies $|\hat{\mu}(d,\cdot) - \mu(d,\cdot)|_{k-1} = o_p(N^{-1/2+1/k})$ for each $d \in \{0,1\}$.

Remarks. 1. Assumption M, employed in Abadie and Imbens (2006), is used for the estimator $\hat{\tau}$. Assumption M (i) is on the sampling process. Assumption M (ii) is on the distributional form of the covariates X. The assumption that X is continuously distributed can be relaxed. Discrete covariates with finite support can be accommodated by using subsamples which does not change our main result. On the other hand, for continuous and high dimensional X, the assumption that the density of X is bounded away from zero can be restrictive. In practice, we can trim observations where the estimated densities of X are too low. Assumption M (iii) contains standard unconfoundedness and overlap conditions to identify the average treatment effect τ. Assumption M (iv) lists boundedness and smoothness conditions for the conditional mean and variance functions. Although Assumption M (iv) is relatively mild, Assumption R typically requires more stringent conditions on the smoothness of $\mu(d,x)$ (particularly for high dimensional X).

2. Assumption W (i)-(iii) are standard in the literature of the weighted bootstrap (e.g., Mason and Newton, 1992). Assumption W (iv) is imposed to deal with the estimation error of $\hat{\mu}$.

3. Assumption R is imposed to guarantee a sufficiently fast convergence rate on the bias estimator $\hat{B}_N$ in (2.4), i.e., $\sqrt{N}(\hat{B}_N - B_N) \xrightarrow{p} 0$. For example, $\hat{\mu}(d,\cdot)$ can be a series estimator with a suitable choice of series length (see, Footnote 2). Other candidates of $\hat{\mu}(d,\cdot)$ are the kernel estimator and nearest neighborhood estimator with adequate trimming (Stone, 1977). For any choice of $\hat{\mu}(d,\cdot)$, we need to select a tuning constant that varies with N to guarantee fast convergence of the bias estimator. As mentioned in Footnote 3, when $k = \dim(X)$ is large, the last condition typically calls for stringent smoothness of $\mu(d, x)$ in x, such as infinite (or very higher-order) differentiability.

Appendix B. Proof of Theorem

Here we present only the proof of part (i) of the theorem. The proof of part (ii) is similar. By definition, the bootstrap counterpart T^* is decomposed as

$$\begin{aligned}
\sqrt{N}T^* &= \sum_{i=1}^{N} W_i^*(\tilde{\tau}_i - \tilde{\tau}) = \sum_{i=1}^{N}(W_i^* - \bar{W}^*)(\tilde{\tau}_i - \tilde{\tau}) \\
&= \sum_{i=1}^{N}(W_i^* - \bar{W}^*)\left[(2D_i - 1)\{1 + M^{-1}K_M(i)\}\hat{e}_i + \hat{\xi}_i\right] \\
&= \sqrt{N}(T_N^* + R_{1N}^* + R_{2N}^*),
\end{aligned}$$

where

$$\begin{aligned}
\sqrt{N}T_N^* &= \sum_{i=1}^{N}(W_i^* - \bar{W}^*)\left[(2D_i - 1)\{1 + M^{-1}K_M(i)\}e_i + \xi_i\right], \\
\sqrt{N}R_{1N}^* &= \sum_{i=1}^{N}(W_i^* - \bar{W}^*)(2D_i - 1)\{1 + M^{-1}K_M(i)\}\{\mu(D_i, X_i) - \hat{\mu}(D_i, X_i)\}, \\
\sqrt{N}R_{2N}^* &= \sum_{i=1}^{N}(W_i^* - \bar{W}^*)(\hat{\xi}_i - \xi_i).
\end{aligned}$$

Thus, it is enough for the conclusion to show that

$$\Pr\{\sqrt{N}|R_{1N}^*| > \epsilon|\mathbf{Z}\} \xrightarrow{p} 0, \qquad \Pr\{\sqrt{N}|R_{2N}^*| > \epsilon|\mathbf{Z}\} \xrightarrow{p} 0, \tag{B.1}$$

for any $\epsilon > 0$, and

$$\sup_r |\Pr\{\sqrt{N}T_N^* \le r|\mathbf{Z}\} - \Pr\{\sqrt{N}(\tilde{\tau} - \tau) \le r\}| \xrightarrow{p} 0. \tag{B.2}$$

For (B.1), the definition of R^*_{1N}, $(2D_i-1)^2=1$, and Assumption W (i) imply

$$\begin{aligned}
& E[(\sqrt{N}R^*_{1N})^2|\mathbf{Z}] \\
= \;& NE[(W^*_1-\bar{W}^*)^2]\frac{1}{N}\sum_{i=1}^{N}\{1+M^{-1}K_M(i)\}^2\{\mu(D_i,X_i)-\hat{\mu}(D_i,X_i)\}^2 \\
& +N(N-1)E[(W^*_1-\bar{W}^*)(W^*_2-\bar{W}^*)] \\
& \quad\times\frac{1}{N(N-1)}\sum_{i\neq j}\left[\begin{array}{c}(2D_i-1)\{1+M^{-1}K_M(i)\}\{\mu(D_i,X_i)-\hat{\mu}(D_i,X_i)\} \\ \times(2D_j-1)\{1+M^{-1}K_M(j)\}\{\mu(D_j,X_j)-\hat{\mu}(D_j,X_j)\}\end{array}\right].
\end{aligned}$$

Assumption W (iv) guarantees $NE[(W^*_1-\bar{W}^*)^2]=O(1)$ and $N(N-1)E[(W^*_1-\bar{W}^*)(W^*_2-\bar{W}^*)]=O(1)$. Therefore, by $|\hat{\mu}(d,\cdot)-\mu(d,\cdot)|_{k-1}=o_p(N^{-1/2+1/k})$ (Assumption R) and the fact that $E[K_M(i)^q]$ is uniformly bounded over N for any $q>0$ (Lemma 3 of Abadie and Imbens, 2006), we obtain $E[(\sqrt{N}R^*_{1N})^2|\mathbf{Z}]\xrightarrow{p}0$. Then the Markov inequality implies $\Pr\{\sqrt{N}|R^*_{1N}|>\epsilon|\mathbf{Z}\}\xrightarrow{p}0$. A similar argument yields $\Pr\{\sqrt{N}|R^*_{2N}|>\epsilon|\mathbf{Z}\}\xrightarrow{p}0$. Therefore, we obtain (B.1).

We now show (B.2). Define

$$\eta_i=\left[(2D_i-1)\{1+M^{-1}K_M(i)\}e_i+\xi_i\right]/\sqrt{N},$$

so that $T^*_N=\sum_{i=1}^N W^*_i\eta_i$. From Polya's theorem and $\Pr\left\{\frac{\sqrt{N}}{\sigma_N}(\tilde{\tau}-\tau)\leq r\right\}\to\Phi(r)$ for all $r\in\mathbb{R}$ (Abadie and Imbens, 2006, Theorem 4), it is enough for (B.2) to verify

$$\Pr\left\{\sqrt{N}\sum_{i=1}^{N}\frac{(W^*_i-\bar{W}^*)\eta_i}{\sigma_N}\leq r\middle|\mathbf{Z}\right\}-\Phi(r)\xrightarrow{p}0\quad\text{for all } r\in\mathbb{R}, \tag{B.3}$$

where $\Phi(r)$ is the standard normal distribution function. By Lemma (i) in Appendix C and $\sum_{i=1}^N(W^*_i-\bar{W})^2\xrightarrow{p}1$, (Assumption W (ii)), it is enough for (B.3) to show

$$\Pr\{Z_N \leq r|\mathbf{Z}\} - \Phi(r) \xrightarrow{p} 0 \quad \text{for all } r \in \mathbb{R}, \tag{B.4}$$

where

$$Z_N = \sqrt{N}\sum_{i=1}^{N} \frac{(W_i^* - \bar{W}^*)\eta_i}{\sqrt{\sum_{i=1}^{N}(W_i^* - \bar{W}^*)^2}\sqrt{\sum_{i=1}^{N}(\eta_i - \bar{\eta})^2}},$$

and $\bar{\eta} = N^{-1}\sum_{i=1}^{N}\eta_i$. To show, (B.4), we adapt the argument in Mason and Newton (1992) to our setup. In particular, let $(R_1, \ldots, R_N)$ be a random vector which takes each permutation of $(1, \ldots, N)$ with equal probability and is independent from $\mathbf{Z}$ and $(W_1^*, \ldots, W_N^*)$. Define

$$Z_N^* = \sqrt{N}\sum_{i=1}^{N} \frac{(W_{R_i}^* - \bar{W}^*)\eta_i}{\sqrt{\sum_{i=1}^{N}(W_i^* - \bar{W}^*)^2}\sqrt{\sum_{i=1}^{N}(\eta_i - \bar{\eta})^2}}.$$

Since $(W_1^*, \ldots, W_N^*)$ is exchangeable (Assumption W (i)), Z_N and Z_N^* follow the same distribution. Thus, it is enough for (B.3) to show

$$\Pr\{Z_N^* \leq r|\mathbf{Z}\} - \Phi(r) \xrightarrow{p} 0 \quad \text{for all } r \in \mathbb{R},$$

as $N \to \infty$, or equivalently, every subsequence $\{N_k\}_{k\in\mathbb{N}} \subset \mathbb{N}$ contains a further subsequence $\{N_{k(l)}\}_{l\in\mathbb{N}} \subset \{N_k\}_{k\in\mathbb{N}}$ such that

$$\Pr\{Z_{N_{k(l)}}^* \leq r|\mathbf{Z}\} - \Phi(r) \xrightarrow{a.s.} 0 \quad \text{for all } r \in \mathbb{R}, \tag{B.5}$$

as $l \to \infty$.

Pick any $r \in \mathbb{R}$. Now define

$$\begin{aligned} V_i &= \frac{W_i^* - \bar{W}^*}{\sqrt{\sum_{i=1}^N (W_i^* - \bar{W}^*)^2}}, \qquad U_i = \frac{\eta_i - \bar{\eta}}{\sqrt{\sum_{i=1}^N (\eta_i - \bar{\eta})^2}}, \\ d_N(\delta) &= \sum_{i=1}^N \sum_{j=1}^N U_i^2 V_j^2 \mathbb{I}\{N U_i^2 V_j^2 > \delta\}. \end{aligned}$$

From Assumption W and Lemmas (ii)-(iii), we have

$$\max_{1,\dots,N} |V_i| \xrightarrow{p} 0, \qquad \max_{1,\dots,N} |U_i| \xrightarrow{p} 0, \qquad d_N(\delta) \xrightarrow{p} 0, \tag{B.6}$$

as $N \to \infty$. Pick any subsequence $\{N_k\}_{k \in \mathbb{N}} \subset \mathbb{N}$. By (B.6), there exists a further subsequence $\{N_{k(l)}\}_{l \in \mathbb{N}} \subset \{N_k\}_{k \in \mathbb{N}}$ such that

$$\max_{1,\dots,N_{k(l)}} |V_i| \overset{a.s.}{\to} 0, \qquad \max_{1,\dots,N_{k(l)}} |U_i| \overset{a.s.}{\to} 0, \qquad d_{N_{k(l)}}(\delta) \overset{a.s.}{\to} 0, \tag{B.7}$$

as $l \to \infty$. Notice that $Z^*_{N_{k(l)}}$ is a simple linear rank statistic conditional on $\mathbf{Z}$ and $(W_1^*, \dots, W^*_{N_{k(l)}})$. Thus, under (B.7), we can apply the rank central limit theorem (Hájek, 1961), that is

$$\Pr\{Z^*_{N_{k(l)}} \le r | \mathbf{Z}, W_1^*, \dots, W^*_{N_{k(l)}}\} - \Phi(r) \overset{a.s.}{\to} 0,$$

as $l \to \infty$. Furthermore, the bounded convergence theorem implies

$$\Pr\{Z^*_{N_{k(l)}} \le r | \mathbf{Z}\} - \Phi(r) = E\left[\Pr\{Z^*_{N_{k(l)}} \le r | \mathbf{Z}, W_1^*, \dots, W^*_{N_{k(l)}}\} - \Phi(r) \middle| \mathbf{Z}\right] \overset{a.s.}{\to} 0,$$

as $l \to \infty$. Therefore, we obtain (B.5) and the conclusion follows.

Appendix C. Lemmas

Lemma. *Use the same notation in Appendix B. Under Assumptions M, W, and R in Appendix A, (i)* $\sum_{i=1}^{N}(\eta_i - \bar{\eta})^2 - \sigma_N^2 \xrightarrow{p} 0$*, (ii)* $\max_{1,\dots,N} |U_i| \xrightarrow{p} 0$*, and (iii)* $d_N(\delta) \xrightarrow{p} 0$*, as* $N \to \infty$*.*

Proof of (i). First, we show $\sum_{i=1}^{N} \eta_i^2 - \sigma_N^2 \xrightarrow{p} 0$. By definition, decompose $\sum_{i=1}^{N} \eta_i^2 = \hat{\sigma}_{1N}^2 + \hat{\sigma}_{2N}^2 + 2C_N$, where

$$\begin{aligned}
\hat{\sigma}_{1N}^2 &= \frac{1}{N}\sum_{i=1}^{N}\{1 + M^{-1}K_M(i)\}^2 e_i^2, \qquad \hat{\sigma}_{2N}^2 = \frac{1}{N}\sum_{i=1}^{N}(\tau_i - \tau)^2, \\
C_N &= \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)\{1 + M^{-1}K_M(i)\} e_i(\tau_i - \tau).
\end{aligned}$$

The law of large numbers guarantees $\hat{\sigma}_{2N}^2 \xrightarrow{p} \sigma_2^2$. For $\hat{\sigma}_{1N}^2$, note that

$$\begin{aligned}
E[(\hat{\sigma}_{1N}^2 - \sigma_N^2)^2] &= \frac{1}{N} E\left[\{1 + M^{-1}K_M(i)\}^4 E[\{e_i^2 - \sigma^2(D_i, X_i)\}^2 | D_i, X_i]\right] \\
&\le \frac{1}{N} E[\{1 + M^{-1}K_M(i)\}^4] \sup_{d\in\{0,1\}, x\in\mathbb{X}} E[e_i^4 | D_i = d, X_i = x] \to 0,
\end{aligned}$$

where the convergence follows from Assumption M (iv) and boundedness of $E[K_M(i)^q]$ for all $q > 0$ uniformly over N (Lemma 3 of Abadie and Imbens, 2006). Thus, the Markov inequality implies $|\hat{\sigma}_{1N}^2 - \sigma_{1N}^2| \xrightarrow{p} 0$. Similarly, by the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
E[C_N^2] &= \frac{1}{N} E\left[\{1 + M^{-1}K_M(i)\}^2 e_i^2 (\tau_i - \tau)^2\right] \\
&\le \frac{1}{N} E[\{1 + M^{-1}K_M(i)\}^4]^{1/2} \left(E[(\tau_i - \tau)^4] \sup_{d\in\{0,1\}, x\in\mathbb{X}} E[e_i^4 | D_i = d, X_i = x] \right)^{1/2} \to 0,
\end{aligned}$$

and thus $C_N \xrightarrow{p} 0$.

Next, we show

$$\bar{\eta} \xrightarrow{p} 0. \tag{C.1}$$

Since (a) $\sup_{d\in\{0,1\},x\in\mathbb{X}} E[|e_i||D_i = d, X_i = x]$, (b) $\sup_{x\in\mathbb{X}} E[|\xi_i||X_i = x]$, and (c) $E[\{1 + M^{-1}K_M(i)\}]$ is uniformly bounded over N, we have

$$E|\bar{\eta}| \leq \frac{1}{\sqrt{N}} E[\{1 + M^{-1}K_M(i)\}(|e_i| + |\xi_i|)] \to 0.$$

Thus, the Markov inequality implies (C.1).

Finally, by using $\sum_{i=1}^{n} \eta_i/\sigma_N \xrightarrow{d} N(0,1)$ (Theorem 4 of Abadie and Imbens, 2006) and $E[\sigma_N] = O(1)$ (Lemma 3 of Abadie and Imbens, 2006), we obtain $N\bar{\eta} = O_p(1)$. Therefore, combining all these results, we obtain the conclusion as

$$\sum_{i=1}^{N}(\eta_i - \bar{\eta})^2 - \sigma_N^2 = \left(\sum_{i=1}^{N}\eta_i^2 - \sigma_N^2\right) - N\bar{\eta}^2 \xrightarrow{p} 0.$$

Proof of (ii). By Lemma (i) and (C.1), it is enough to show $\max_{1,\dots,N}|\eta_i| \xrightarrow{p} 0$. This follows by

$$\Pr\left\{\max_{1,\dots,N}|\eta_i| > \epsilon\right\} \leq N\Pr\{|\eta_i| > \epsilon\} \leq \frac{1}{N\epsilon^4}E[(\{1 + M^{-1}K_M(i)\}e_i + \xi_i)^4] \to 0,$$

for any $\epsilon > 0$, where the first inequality follows from a set inclusion relationship and the fact that $\{\eta_i\}_{i=1}^N$ are identically distributed, the second inequality follows from the Markov inequality, and the convergence follows from Assumption M (iv) and uniform boundedness of $E[K_M(i)^q]$ over N and $q > 0$ (Lemma 3 of Abadie and Imbens, 2006).

Proof of (iii). Pick any $\epsilon > 0$. It holds that with probability approaching one,

$$\begin{aligned} d_N(\delta) &\leq \sum_{i=1}^{N} U_i^2 \mathbb{I}\{NU_i^2 > \delta/\epsilon\} \leq \frac{\epsilon}{\delta} N \sum_{i=1}^{N} U_i^4 \\ &= \frac{\epsilon}{\delta} \left(\sum_{i=1}^{N} (\eta_i - \bar{\eta})^2 \right)^{-2} \frac{1}{N} \sum_{i=1}^{N} (\{1 + M^{-1}K_M(i)\}e_i + \xi_i)^4, \end{aligned}$$

where the first inequality follows from $\max_{1,\ldots,N} |V_i| \xrightarrow{p} 0$ (by Assumption W (ii)-(iii)) and $\sum_{j=1}^{N} V_j^2 = 1$. Note that $\left(\sum_{i=1}^{N}(\eta_i - \bar{\eta})^2\right)^{-2} = O_p(1)$ by Lemma (i) and that σ_N^2 is uniformly bounded from below ($\sigma_N^2 \geq \inf_{d\in\{0,1\},x\in\mathbb{X}} \sigma^2(d,x)$). Thus, if we can show that

$$\frac{1}{N} \sum_{i=1}^{N} (\{1 + M^{-1}K_M(i)\}e_i + \xi_i)^4 = O_p(1), \tag{C.2}$$

then we obtain the conclusion by taking ϵ arbitrarily small.

It remains to show (C.2). Let $\gamma_i = \{1 + M^{-1}K_M(i)\}e_i + \xi_i$. Note that

$$\begin{aligned} \gamma_i^4 &= \{1 + M^{-1}K_M(i)\}^4 e_i^4 + 4\{1 + M^{-1}K_M(i)\}^3 e_i^3 \xi_i + 6\{1 + M^{-1}K_M(i)\}^2 e_i^2 \xi_i^2 \\ &\quad + 4\{1 + M^{-1}K_M(i)\} e_i \xi_i^3 + \xi_i^4. \end{aligned}$$

By Assumption M (iv), $\max_{q=1,\ldots,4} E|e_i|^q < C$ for some $C > 0$. By Abadie and Imbens (2006, Lemma 3), $E[K_M(i)^q]$ is uniformly bounded for any q. By the Lipschitz continuity of μ (Assumption M (iv)) and the compact support of X (Assumption M (ii)), it holds $|\xi_i| < C'$ a.s. for some $C' > 0$. Combining these results, we can see that $E\left[\left|\frac{1}{N}\sum_{i=1}^{N}\gamma_i^4\right|\right]$ is bounded uniformly over N. Therefore, the Markov inequality implies (C.2), and the conclusion follows.

Appendix D. Tables

Table 1. Simulation designs for $m(\cdot)$

Curves	$m(z)$
1	$0.15 + 0.7z$
2	$0.1 + z/2 + \exp(-200(z-0.7)^2)/2$
3	$0.8 - 2(z-0.9)^2 - 5(z-0.7)^3 - 10(z-0.6)^{10}$
4	$0.2 + \sqrt{1-z} - 0.6(0.9-z)^2$
5	$0.2 + \sqrt{1-z} - 0.6(0.9-z)^2 - 0.1z\cos(30z)$
6	$0.4 + 0.25\sin(8z-5) + 0.4\exp(-16(4z-2.5)^2)$

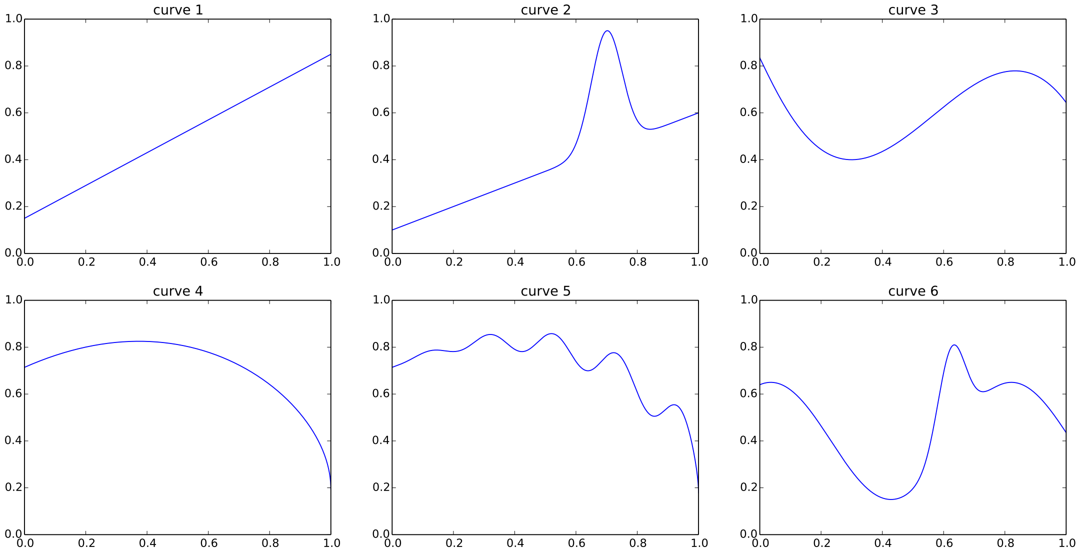

Table 2. Simulation results for bootstrap and asymptotic t methods

		Wild		Nonparametric		Asymptotic t		Naive	
			Average		Average		Average		Average
k	Curve	95% CI	CI length	95% CI	CI length	95% CI	CI length	95% CI	CI length
1	1	0.9342	0.1696	0.9342	0.1701	0.9423	0.1760	0.8575	0.1435
	2	0.9396	0.2063	0.9397	0.2071	0.9068	0.1901	0.8317	0.1649
	3	0.9298	0.1936	0.9307	0.1941	0.9026	0.1786	0.8236	0.1556
	4	0.9395	0.1837	0.9406	0.1841	0.9309	0.1775	0.8642	0.1518
	5	0.9424	0.1855	0.9426	0.1859	0.9325	0.1799	0.8719	0.1533
	6	0.9500	0.2400	0.9503	0.2406	0.8731	0.1904	0.8576	0.1890
2	1	0.9257	0.1689	0.9268	0.1693	0.9390	0.1775	0.8572	0.1420
	2	0.9282	0.1982	0.9282	0.1987	0.9256	0.1980	0.8472	0.1634
	3	0.9394	0.1911	0.9394	0.1915	0.9257	0.1817	0.8659	0.1570
	4	0.9210	0.1807	0.9205	0.1811	0.9154	0.1784	0.8398	0.1506
	5	0.9264	0.1818	0.9268	0.1822	0.9211	0.1800	0.8459	0.1515
	6	0.9473	0.2381	0.9476	0.2385	0.9119	0.2092	0.8819	0.1942
3	1	0.9243	0.1692	0.9247	0.1696	0.9396	0.1799	0.8561	0.1426
	2	0.9234	0.1956	0.9242	0.1960	0.9286	0.1994	0.8478	0.1627
	3	0.9323	0.1920	0.9329	0.1924	0.9205	0.1847	0.8627	0.1585
	4	0.9073	0.1810	0.9083	0.1814	0.9003	0.1794	0.8143	0.1515
	5	0.9154	0.1816	0.9163	0.1820	0.9094	0.1806	0.8215	0.1520
	6	0.9282	0.2388	0.9296	0.2393	0.8979	0.2145	0.8535	0.1956
4	1	0.9184	0.1695	0.9191	0.1699	0.9360	0.1819	0.8501	0.1431
	2	0.9151	0.1941	0.9154	0.1945	0.9226	0.2002	0.8413	0.1622
	3	0.9300	0.1932	0.9313	0.1937	0.9201	0.1871	0.8597	0.1598
	4	0.9029	0.1812	0.9024	0.1816	0.8980	0.1800	0.8111	0.1520
	5	0.9061	0.1817	0.9076	0.1821	0.9025	0.1811	0.8145	0.1525
	6	0.9228	0.2396	0.9239	0.2401	0.8918	0.2177	0.8264	0.1972
5	1	0.9152	0.1698	0.9150	0.1702	0.9370	0.1835	0.8439	0.1437
	2	0.9127	0.1933	0.9126	0.1938	0.9227	0.2012	0.8375	0.1622
	3	0.9270	0.1940	0.9281	0.1944	0.9181	0.1884	0.8528	0.1607
	4	0.8961	0.1817	0.8960	0.1821	0.8901	0.1807	0.7932	0.1528
	5	0.8976	0.1827	0.8979	0.1831	0.8919	0.1824	0.7920	0.1535
	6	0.9103	0.2404	0.9111	0.2409	0.8782	0.2201	0.8069	0.1986

Table 3. Simulation results for subsampling

		$S = 20$		$S = 30$		$S = 50$		Bickel-Sakov	
			Average		Average		Average		Average
k	Curve	95% CI	CI length	95% CI	CI length	95% CI	CI length	95% CI	CI length
1	1	0.8920	0.1549	0.8630	0.1462	0.8080	0.1269	0.8562	0.1422
	2	0.8775	0.1700	0.8430	0.1619	0.8020	0.1371	0.7909	0.1642
	3	0.8603	0.1735	0.8150	0.1628	0.7787	0.1375	0.8015	0.1688
	4	0.8740	0.1725	0.8320	0.1613	0.7920	0.1365	0.8766	0.1590
	5	0.8788	0.1722	0.8392	0.1609	0.7968	0.1363	0.8889	0.1614
	6	0.8713	0.1805	0.8285	0.1686	0.7942	0.1421	0.7773	0.1967
2	1	0.8720	0.1769	0.8324	0.1653	0.7939	0.1396	0.8347	0.1382
	2	0.8709	0.1777	0.8346	0.1661	0.7934	0.1402	0.8185	0.1602
	3	0.8749	0.1780	0.8391	0.1659	0.7954	0.1400	0.8742	0.1615
	4	0.8751	0.1771	0.8409	0.1650	0.7945	0.1393	0.8426	0.1561
	5	0.8745	0.1765	0.8415	0.1642	0.7940	0.1387	0.8469	0.1578
	6	0.8768	0.1802	0.8449	0.1679	0.7958	0.1414	0.8607	0.1914
3	1	0.8768	0.1782	0.8453	0.1661	0.7960	0.1402	0.8401	0.1385
	2	0.8754	0.1785	0.8441	0.1665	0.7951	0.1404	0.8294	0.1604
	3	0.8765	0.1786	0.8470	0.1664	0.7961	0.1404	0.8949	0.1636
	4	0.8750	0.1781	0.8446	0.1657	0.7934	0.1399	0.8341	0.1566
	5	0.8746	0.1778	0.8440	0.1652	0.7913	0.1395	0.8291	0.1586
	6	0.8753	0.1802	0.8460	0.1677	0.7918	0.1414	0.8557	0.1980
4	1	0.8748	0.1788	0.8465	0.1665	0.7919	0.1405	0.8384	0.1386
	2	0.8751	0.1791	0.8471	0.1667	0.7922	0.1407	0.8179	0.1619
	3	0.8754	0.1791	0.8477	0.1667	0.7921	0.1407	0.8762	0.1655
	4	0.8736	0.1787	0.8448	0.1661	0.7889	0.1403	0.8123	0.1571
	5	0.8724	0.1784	0.8433	0.1657	0.7874	0.1400	0.8115	0.1579
	6	0.8728	0.1800	0.8441	0.1675	0.7871	0.1414	0.8240	0.2020
5	1	0.8722	0.1790	0.8440	0.1666	0.7866	0.1408	0.8330	0.1375
	2	0.8715	0.1790	0.8437	0.1667	0.7852	0.1408	0.8447	0.1627
	3	0.8720	0.1790	0.8447	0.1667	0.7860	0.1408	0.8467	0.1647
	4	0.8702	0.1787	0.8423	0.1663	0.7837	0.1405	0.7693	0.1587
	5	0.8685	0.1784	0.8404	0.1660	0.7816	0.1402	0.7976	0.1593
	6	0.8677	0.1799	0.8404	0.1674	0.7806	0.1413	0.8111	0.2029

Table 4. Summary statistics

	# of obs	Age	Education	Black	Hispanic
NSW-L					
Treated	297	24.63	10.38	.80	.09
		(6.69)	(1.82)	(.40)	(.29)
Control	425	24.45	10.19	.84	.11
		(6.59)	(1.62)	(.40)	(.32)
NSW-DW					
Treated	185	24.45	10.19	.84	.06
		(7.16)	(2.01)	(.36)	(.24)
Control	260	25.05	10.09	.83	.11
		(7.06)	(1.61)	(.38)	(.31)
PSID	2,490	34.85	12.11	.25	.03
		(10.44)	(3.08)	(.43)	(.18)
CPS	15,992	33.22	12.02	.07	.07
		(11.05)	(2.87)	(.26)	(.25)
Normalized difference					
NSW-L T/C		.03	.11	.003	-.06
NSW-L/PSID		-1.17	-.69	1.32	-1.95
NSW-L/CPS		-0.94	-.69	2.16	-1.31
NSW-DW T/C		.11	.14	.04	-.17
NSW-DW/PSID		-1.01	-.68	1.48	.13
NSW-DW/CPS		-.80	-.68	2.43	-.05

	No degree	Married	RE74	RE75
NSW-L				
Treated	.73	.17		3,066
	(.37)	(.44)		(4,874)
Control	.81	.16		3,026
	(.36)	(.39)		(5,201)
NSW-DW				
Treated	.71	.19	2,095	1,532
	(.39)	(.46)	(4,886)	(3,219)
Control	.83	.15	2,107	1,267
	(.36)	(.37)	(5,687)	(3,102)
PSID	.31	.87	19,429	19,063
	(.46)	(.34)	(13,406)	(13,596)
CPS	.29	.71	14,016	13,650
	(.45)	(.46)	(9,570)	(9,270)
Normalized difference				
NSW-L T/C	-.20	.03		.008
NSW-L/PSID	.26	.94		-1.57
NSW-L/CPS	.08	.97		-1.43
NSW-DW T/C	-.30	.09	-.002	.08
NSW-DW/PSID	.88	-1.84	-1.72	-1.77
NSW-DW/CPS	.90	-1.23	-1.57	-1.75

* Earnings (RE74 and RE75) are expressed in 1982 dollars.

Table 5. Covariate matching

		NSW-L			NSW-DW	
	NSW-L	CPS	PSID	NSW-DW	CPS	PSID
Mean difference	886.30			1794.34		
	(488.20)			(671.00)		
Covariate matching						
$M = 1$	361.80	19.80	-638.54	2046.21	2186.92	742.54
Wild bootstrap	(590.08)	(655.89)	(1018.55)	(762.77)	(891.21)	(1262.48)
Nonparametric bootstrap	(593.25)	(668.62)	(1001.16)	(738.83)	(864.54)	(1303.98)
Asymptotic t	(619.24)	(687.29)	(1139.77)	(739.40)	(924.67)	(1262.48)
Naive bootstrap	(613.46)	(674.26)	(977.71)	(805.07)	(849.30)	(1055.76)
$M = 4$	881.31	-340.29	-1122.87	1908.52	1688.37	1460.62
Wild bootstrap	(500.76)	(546.06)	(984.54)	(704.27)	(734.14)	(831.73)
Nonparametric bootstrap	(519.47)	(560.89)	(987.79)	(676.26)	(744.72)	(860.74)
Asymptotic t	(511.36)	(560.47)	(1008.49)	(678.94)	(735.24)	(902.77)
Naive bootstrap	(534.03)	(605.76)	(925.64)	(699.86)	(763.08)	(898.42)
$M = 16$	952.86	-427.62	-928.56	1959.27	1430.83	1687.08
Wild bootstrap	(502.91)	(530.95)	(777.75)	(668.78)	(657.65)	(800.84)
Nonparametric bootstrap	(499.09)	(529.05)	(783.63)	(670.20)	(669.93)	(799.79)
Asymptotic t	(487.19)	(508.27)	(785.43)	(664.68)	(654.88)	(846.48)
Naive bootstrap	(522.84)	(535.95)	(774.82)	(712.53)	(663.70)	(790.79)

References

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

[2] Abadie, A. and G. W. Imbens (2008) On the failure of the bootstrap for matching estimators, *Econometrica*, 76, 1537-1557.

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

[4] Abadie, A. and G. W. Imbens (2012) A Martingale representation for matching estimators, *Journal of the American Statistical Association*, 107, 833-843.

[5] Bickel, P. J. and A. Sakov (2008) On the choice of m in the m out of n bootstrap and confidence bounds for extrema, *Statistica Sinica*, 18, 967-985.

[6] Busso, M., DiNardo, J. and J. McCrary (2014) New evidence on the finite sample properties of propensity score matching and reweighting estimators, *Review of Economics and Statistics*, 96, 885-897.

[7] Cattaneo, M. D., Crump, R. K. and M. Jansson (2010) Robust data-driven inference for density-weighted average derivatives, *Journal of the American Statistical Association*, 105, 1070-1083.

[8] Cattaneo, M. D., Crump, R. K. and M. Jansson (2014) Bootstrapping density-weighted average derivatives, *Econometric Theory*, 30, 1135-1164.

[9] Dehejia, R. and S. Wahba (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs, *Journal of the American Statistical Association*, 94, 1053-1062.

[10] Dehejia, R. and S. Wahba (2002) Propensity score-matching method for nonexperimental causal studies, *Review of Economics and Statistics*, 84, 151-161.

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

[12] Frölich, M. (2004) Finite-sample properties of propensity-score matching and weighting estimators, *Review of Economics and Statistics*, 86, 77-90.

[13] Hájek, J. (1961) Some extensions of the Wald-Wolfowitz-Noether theorem, *Annals of Mathematical Statistics*, 32, 506-523.

[14] Heckman, J. and J. Hotz (1989) Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training, *Journal of the American Statistical Association*, 84, 862-874.

[15] Heckman, J., Ichimura, H. and P. Todd (1998) Matching as an econometric evaluation estimator, *Review of Economic Studies*, 65, 261-294.

[16] Hirano, K., Imbens, G. and G. Ridder (2003) Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica*, 71, 1161-1189.

[17] Kline, P. and A. Santos (2012) Higher order properties of the wild bootstrap under misspecification, *Journal of Econometrics*, 171, 54-70.

[18] Lalonde, R. J. (1986) Evaluation the econometric evaluations of training programs with experimental data, *American Economic Review*, 76, 604-620.

[19] Mammen, E. (1993) Bootstrap and wild bootstrap for high dimensional linear models, *Annals of Statistics*, 21, 255-285.

[20] Mason, D. M. and M. A. Newton (1992) A rank statistics approach to the consistency of a general bootstrap, *Annals of Statistics*, 20, 1611-1624.

[21] Pauly, M. (2011) Weighted resampling of martingale difference arrays with applications, *Electronic Journal of Statistics*, 5, 41-52.

[22] Politis, N. and J. Romano (1994) Large sample confidence regions based on subsamples under minimal assumptions, *Annals of Statistics*, 22, 2031-2050.

[23] Rubin, D. (1981) The Bayesian bootstrap, *Annals of Statistics*, 9, 130-134.

[24] Smith, J. and P. Todd (2005) Does matching address Lalonde's critique of nonexperimental estimators?, *Journal of Econometrics*, 125, 305-353.

[25] Stone, C. J. (1977) Consistent nonparametric regression, *Annals of Statistics*, 5, 595-620.

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

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

E-mail address: t.otsu@lse.ac.uk

Department of Economics, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, WI 53706, USA.

E-mail address: yrai@wisc.edu