Econometrica, Vol. 84, No. 2 (March, 2016), 781–807

NOTES AND COMMENTS

MATCHING ON THE ESTIMATED PROPENSITY SCORE

BY ALBERTO ABADIE AND GUIDO W. IMBENS[1]

Propensity score matching estimators (Rosenbaum and Rubin (1983)) are widely used in evaluation research to estimate average treatment effects. In this article, we derive the large sample distribution of propensity score matching estimators. Our derivations take into account that the propensity score is itself estimated in a first step, prior to matching. We prove that first step estimation of the propensity score affects the large sample distribution of propensity score matching estimators, and derive adjustments to the large sample variances of propensity score matching estimators of the average treatment effect (ATE) and the average treatment effect on the treated (ATET). The adjustment for the ATE estimator is negative (or zero in some special cases), implying that matching on the estimated propensity score is more efficient than matching on the true propensity score in large samples. However, for the ATET estimator, the sign of the adjustment term depends on the data generating process, and ignoring the estimation error in the propensity score may lead to confidence intervals that are either too large or too small.

KEYWORDS: Matching estimators, propensity score matching, average treatment effects, causal inference, program evaluation.

1. INTRODUCTION

PROPENSITY SCORE MATCHING ESTIMATORS (Rosenbaum and Rubin (1983)) are widely used to estimate treatment effects.[2] Rosenbaum and Rubin (1983) defined the propensity score as the conditional probability of assignment to a treatment given a vector of covariates. Suppose that adjusting for a set of covariates is sufficient to eliminate confounding. The key insight of Rosenbaum and Rubin (1983) is that adjusting only for the propensity score is also sufficient to eliminate confounding. Relative to matching directly on the covariates, propensity score matching has the advantage of reducing the dimensionality of matching to a single dimension. This greatly facilitates the matching process

[1]We are grateful to the editor and three referees for helpful comments, to Ben Hansen, Judith Lok, James Robins, Paul Rosenbaum, Donald Rubin, and participants in many seminars for comments and discussions, and to Jann Spiess for expert research assistance. Financial support by the NSF through Grants SES 0820361 and SES 0961707 is gratefully acknowledged.

[2]Following the terminology in Abadie and Imbens (2006), the term "matching estimator" is reserved in this article to estimators that match each unit (or each unit of some sample subset, e.g., the treated) to a small number of units with similar characteristics in the opposite treatment arm. Thus, our discussion does not refer to regression imputation methods, like the kernel matching method of Heckman, Ichimura, and Todd (1998), which use a large number of matches per unit and nonparametric smoothing techniques to consistently estimate unit-level regression values under counterfactual treatment assignments. See Hahn (1998), Heckman, Ichimura, and Todd (1998), Imbens (2004), and Imbens and Wooldridge (2009) for a discussion of such estimators.

 DOI: 10.3982/ECTA11293

because units with dissimilar covariate values may nevertheless have similar values for their propensity scores.

In observational studies, propensity scores are not known, so they have to be estimated prior to matching. In spite of the great popularity that propensity score matching methods have enjoyed since they were proposed by Rosenbaum and Rubin in 1983, their large sample distribution has not yet been derived for the case when the propensity score is estimated in a first step.[3] A possible reason for this void in the literature is that matching estimators are non-smooth functionals of the distribution of the matching variables, which makes it difficult to establish an asymptotic approximation to the distribution of matching estimators when a matching variable is estimated in a first step. This has motivated the use of bootstrap standard errors for propensity score matching estimators. However, recently it has been shown that the bootstrap is not, in general, valid for matching estimators (Abadie and Imbens (2008)).[4]

In this article, we derive large sample approximations to the distribution of propensity score matching estimators. Our derivations take into account that the propensity score is itself estimated in a first step. We show that propensity matching estimators have approximately Normal distributions in large samples. We demonstrate that first step estimation of the propensity score affects the large sample distribution of propensity score matching estimators, and derive adjustments to the large sample variance of propensity score matching estimators that correct for first step estimation of the propensity score. We do this for estimators of the average treatment effect (ATE) and the average treatment effect on the treated (ATET). The adjustment for the ATE estimator is negative (or zero in some special cases), implying that matching on the estimated propensity score is more efficient than matching on the true propensity score in large samples. As a result, treating the estimated propensity score as it was the true propensity score for estimating the variance of the ATE estimator leads to conservative confidence intervals. However, for the ATET estimator, the sign of the adjustment depends on the data generating process, and ignoring the estimation error in the propensity score may lead to confidence intervals that are either too large or too small.

2. MATCHING ESTIMATORS

The setup in this article is a standard one in the program evaluation literature, where the focus of the analysis is often the effect of a binary treatment,

[3]Influential papers using matching on the estimated propensity score include Heckman, Ichimura, and Todd (1997), Dehejia and Wahba (1999), and Smith and Todd (2005).

[4]In contexts other than matching, Heckman, Ichimura, and Todd (1998), Hirano, Imbens, and Ridder (2003), Abadie (2005), Wooldridge (2007), and Angrist and Kuersteiner (2011) derived large sample properties of statistics based on a first step estimator of the propensity score. In all these cases, the second step statistics are smooth functionals of the propensity scores and, therefore, standard stochastic expansions for two-step estimators apply (see, e.g., Newey and McFadden (1994)).

represented in this paper by the indicator variable W, on some outcome variable, Y. More specifically, $W = 1$ indicates exposure to the treatment, while $W = 0$ indicates lack of exposure to the treatment. Following Rubin (1974), we define treatment effects in terms of potential outcomes. We define $Y(1)$ as the potential outcome under exposure to treatment, and $Y(0)$ as the potential outcome under no exposure to treatment. Our goal is to estimate the average treatment effect,

$$\tau = E\big[Y(1) - Y(0)\big],$$

where the expectation is taken over the population of interest. Alternatively, the goal may be estimation of the average effect for the treated,

$$\tau_t = E\big[Y(1) - Y(0)|W = 1\big].$$

Estimation of these average treatment effects is complicated by the fact that for each unit in the population, we observe at most one of the potential outcomes:

$$Y = \begin{cases} Y(0) & \text{if } W = 0, \\ Y(1) & \text{if } W = 1. \end{cases}$$

Let X be a vector of covariates of dimension k. The propensity score is $p(X) = \Pr(W = 1|X)$, and $p^* = \Pr(W = 1)$ is the probability of being treated. The following assumption is often referred to as "strong ignorability" (Rosenbaum and Rubin (1983)). It means that adjusting for X is sufficient to eliminate all confounding.

ASSUMPTION 1: (i) $Y(1), Y(0) \perp\!\!\!\perp W|X$ *almost surely*; (ii) $\underline{p} \leq p(X) \leq \overline{p}$ *almost surely, for some* $\underline{p} > 0$ *and* $\overline{p} < 1$.

Assumption 1(i) uses the conditional independence notation in Dawid (1979). This assumption is often referred to as "unconfoundedness." It will hold, for example, if all confounders are included in X, so that after controlling for X, treatment exposure is independent of the potential outcomes. Hahn (1998) derived asymptotic variance bounds and studied asymptotically efficient estimation under Assumption 1(i). Assumption 1(ii) implies that, for almost all values of X, the population includes treated and untreated units. Moreover, Assumption 1(ii) bounds the values of the propensity score away from zero and 1. Khan and Tamer (2010) have shown that this condition is necessary for root-N consistent estimation of the average treatment effect.

Let $\mu(w, x) = E[Y|W = w, X = x]$ and $\sigma^2(w, x) = \text{var}(Y|W = w, X = x)$ be the conditional mean and variance of Y given $W = w$ and $X = x$. Similarly, let $\bar{\mu}(w, p) = E[Y|W = w, p(X) = p]$ and $\bar{\sigma}^2(w, p) = \text{var}(Y|W = w, p(X) = p)$ be the conditional mean and variance of Y given $W = w$ and

$p(X) = p$. Under Assumption 1,

$$\tau = E\big[\mu(1, X) - \mu(0, X)\big]$$

and

$$\tau_t = E\big[\mu(1, X) - \mu(0, X)|W = 1\big]$$

(see Rubin (1974)). Therefore, adjusting for differences in the distribution of X between treated and nontreated removes all confounding and, therefore, allows identification of ATE and ATET. Rosenbaum and Rubin (1983) proved that W and X are independent conditional on the propensity score, $p(X)$, which implies that under Assumption 1:

$$\tau = E\big[\bar{\mu}\big(1, p(X)\big) - \bar{\mu}\big(0, p(X)\big)\big]$$

and

$$\tau_t = E\big[\bar{\mu}\big(1, p(X)\big) - \bar{\mu}\big(0, p(X)\big)|W = 1\big].$$

In other words, under Assumption 1, adjusting for the propensity score only is enough to remove all confounding. This result motivates the use of propensity score matching estimators. A propensity score matching estimator for the average treatment effect can be defined as

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

where M is a fixed number of matches per unit and $\mathcal{J}_M(i)$ is the set of matches for unit i.[5] (The superscript $*$ on $\widehat{\tau}_N^*$ indicates that matching is done on the true propensity score.) For concreteness, in this article we will consider matching with replacement, so each unit in the sample can be used as a match multiple times. In the absence of matching ties, the set of matches $\mathcal{J}_M(i)$ can formally

[5]In typical applications, M is small, often $M = 1$. Choosing a small M reduces finite sample biases caused by matches of poor quality, that is, matches between individuals with substantial differences in their propensity score values. Larger values of M produce lower large sample variances (see Abadie and Imbens (2006, Section 3.4)), and one could consider increasing M in a particular application if such increase has a small effect on the size of the matching discrepancies, which are observed in the data (Abadie and Imbens (2011)). Similarly to Yatchew's (1997) work on semiparametric differencing estimators, we derive a large sample approximation to the distribution of matching estimators for fixed values of M. Large sample approximations based on fixed values of smoothing parameters have been shown to increase accuracy in other contexts (see, in particular, Kiefer and Vogelsang (2005)).

be defined as

$$\mathcal{J}_M(i) = \Bigg\{ j = 1, \ldots, N : W_j = 1 - W_i, \\ \Bigg(\sum_{k: W_k = 1 - W_i} 1_{[|p(X_i) - p(X_k)| \leq |p(X_i) - p(X_j)|]} \Bigg) \leq M \Bigg\},$$

where $1_{[\cdot]}$ is a binary indicator that takes value 1 if the event inside brackets is true, and value zero if not. For the average effect on the treated, τ_t, the corresponding estimator is

$$\widehat{\tau}^*_{t,N} = \frac{1}{N_1} \sum_{i=1}^{N} W_i \Bigg(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j \Bigg),$$

where $N_1 = \sum_{i=1}^{N} W_i$ is the number of treated units in the sample.

The large sample distributions of $\widehat{\tau}^*_N$ and $\widehat{\tau}^*_{t,N}$ can be easily derived from results on matching estimators in Abadie and Imbens (2006), applied to the case where the only matching variable is the known propensity score. However, one of the assumptions in Abadie and Imbens (2006) requires that the density of the matching variables is bounded away from zero. Although this assumption may be appropriate in settings where the matching is carried out directly on the covariates, X, it is much less appealing for propensity score matching estimators. For example, if the propensity score has the form $p(X) = F(X'\theta)$, then even if the density of X is bounded away from zero on its support, the density of $F(X'\theta)$ will generally not be bounded away from zero on its support. We therefore generalize the results in Abadie and Imbens (2006) to allow the density of propensity score to take values that are arbitrarily close to zero.

ASSUMPTION 2: (i) *The propensity score* $p(X)$ *is continuously distributed, has interval support* $[\underline{p}, \overline{p}]$, *and has a density that is a continuous function on* $[\underline{p}, \overline{p}]$; (ii) *for* $w = 0, 1$, $\bar{\mu}(w, p)$ *and* $\bar{\sigma}^2(w, p)$ *are Lipschitz-continuous and continuous in* p, *respectively*; (iii) *for* $w = 0, 1$, *there exists* $\delta > 0$ *such that* $E[|Y|^{2+\delta} | W = w, p(X) = p]$ *is uniformly bounded*.

ASSUMPTION 3: $\{(Y_i, W_i, X_i)\}_{i=1}^N$ *are independent draws from the distribution of* (Y, W, X).

The next proposition presents the large sample distributions of $\widehat{\tau}^*_N$ and $\widehat{\tau}^*_{t,N}$ under Assumptions 1–3.

PROPOSITION 1: *Suppose Assumptions 1–3 hold. Then*, (i)

$$\sqrt{N}\big(\widehat{\tau}^*_N - \tau\big) \xrightarrow{d} N\big(0, \sigma^2\big),$$

where

$$\sigma^2 = E\big[\big(\bar{\mu}\big(1, p(X)\big) - \bar{\mu}\big(0, p(X)\big) - \tau\big)^2\big] + E\left[\bar{\sigma}^2\big(1, p(X)\big)\left(\frac{1}{p(X)} + \frac{1}{2M}\left(\frac{1}{p(X)} - p(X)\right)\right)\right] + E\left[\bar{\sigma}^2\big(0, p(X)\big) \times \left(\frac{1}{1-p(X)} + \frac{1}{2M}\left(\frac{1}{1-p(X)} - \big(1-p(X)\big)\right)\right)\right],$$

and (ii)

$$\sqrt{N}\big(\widehat{\tau}^*_{t,N} - \tau_t\big) \xrightarrow{d} N\big(0, \sigma_t^2\big),$$

where

$$\sigma_t^2 = \frac{1}{E\big[p(X)\big]^2} E\big[p(X)\big(\bar{\mu}\big(1, p(X)\big) - \bar{\mu}\big(0, p(X)\big) - \tau_t\big)^2\big] + \frac{1}{E\big[p(X)\big]^2} E\big[\bar{\sigma}^2\big(1, p(X)\big)p(X)\big] + \frac{1}{E\big[p(X)\big]^2} E\left[\bar{\sigma}^2\big(0, p(X)\big) \times \left(\frac{p^2(X)}{1-p(X)} + \frac{1}{M}p(X) + \frac{1}{2M}\frac{p^2(X)}{1-p(X)}\right)\right].$$

The proof of this proposition is available in the Supplemental Material (Abadie and Imbens (2016)).

Motivated by the fact that, in observational studies, propensity scores are not known, we are interested in the case where matching is not on the true propensity score $p(X)$, but on an estimate of the propensity score. Following Rosenbaum and Rubin (1983) and most of the empirical literature, we consider a generalized linear specification for the propensity score, $p(x) = F(x'\theta)$. In empirical research, the link function F is usually specified as Logit or Probit. It is straightforward to extend our results to more general parametric models for the propensity score.[6] For unit i, and for arbitrary values for θ, let $\mathcal{J}_M(i, \theta)$

[6]Misspecification of the propensity score typically leads to inconsistency of the treatment effect estimator, unless the misspecified propensity score constitutes a balancing score, that is, a function, $b(X)$, of the covariates such that $X \perp\!\!\!\perp W | b(X)$ (see Rosenbaum and Rubin (1983)). Moti-

denote the set of M matches where we match on $F(X'\theta)$:

$$\mathcal{J}_M(i,\theta) = \Bigg\{ j = 1,\ldots,N : W_j = 1 - W_i, \\ \Bigg(\sum_{k:W_k=1-W_i} 1_{[|F(X_i'\theta)-F(X_k'\theta)|\le|F(X_i'\theta)-F(X_j'\theta)|]} \Bigg) \le M \Bigg\}.$$

The matching estimator for the average treatment effect where we match on $F(X'\theta)$ is then

$$\widehat{\tau}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\Bigg(Y_i - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i,\theta)} Y_j \Bigg).$$

Let θ^* denote the true value of the propensity score model parameter vector, so that $p(X) = F(X'\theta^*)$. Then, the estimator based on matching on the true propensity score can be written as $\widehat{\tau}_N^* = \widehat{\tau}_N(\theta^*)$. We are interested in the case where $\widehat{\tau}_N(\theta)$ is evaluated at an estimator $\widehat{\theta}_N$ of θ^*, based on a sample $\{Y_i, W_i, X_i\}_{i=1}^N$. We focus on the case where $\widehat{\theta}_N$ is the maximum likelihood estimator of θ:[7]

$$\widehat{\theta}_N = \arg\max_{\theta} L(\theta|W_1, X_1, \ldots, W_N, X_N),$$

where the log-likelihood function is

$$L(\theta|W_1, X_1, \ldots, W_N, X_N) \\ = \sum_{i=1}^{N} W_i \ln F\big(X_i'\theta\big) + (1 - W_i)\ln\big(1 - F\big(X_i'\theta\big)\big).$$

The propensity score matching estimator of τ that matches on the estimated propensity score can now be written as

$$\widehat{\tau}_N = \widehat{\tau}_N(\widehat{\theta}_N) = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\Bigg(Y_i - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i,\widehat{\theta}_N)} Y_j \Bigg).$$

vated by this consideration, empirical researchers routinely use measures of balance in the distribution of the covariates between treated and nontreated, conditional on the estimated propensity score, to perform specification searches on the propensity score (see, e.g., Dehejia and Wahba (1999)). An alternative safeguard against misspecification of the propensity score is the use "doubly robust" matching estimators, like the bias-corrected matching estimator of Abadie and Imbens (2011).

[7]It is straightforward to extend our results to other asymptotically linear estimators of θ^*.

Similarly, the propensity score matching estimator of τ_t that matches on the estimated propensity score can be written as

$$\widehat{\tau}_{t,N} = \widehat{\tau}_{t,N}(\widehat{\theta}_N) = \frac{1}{N_1}\sum_{i=1}^{N} W_i\Bigg(Y_i - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i,\widehat{\theta}_N)} Y_j\Bigg).$$

Whenever confusion is possible, we will be explicit in the dependence of the matching estimators on θ. If the argument is omitted, $\widehat{\tau}_N$ and $\widehat{\tau}_{t,N}$ are used as shorthand for $\widehat{\tau}_N(\widehat{\theta}_N)$ and $\widehat{\tau}_{t,N}(\widehat{\theta}_N)$, respectively.

The two main questions addressed in this article are (i) do the estimators based on matching on the estimated propensity score have Normal large sample distributions, and (ii) if so, how does their large sample variance compare to that of the estimators that match on the true propensity score, given in Proposition 1? In the next section, we answer these two questions and derive the large sample distribution of $\widehat{\tau}_N(\widehat{\theta}_N)$ and $\widehat{\tau}_{t,N}(\widehat{\theta}_N)$. Conventional linearization methods for two-step statistics are difficult to apply in the context of matching estimators because matching estimators are complicated functionals of the distribution of the data. We therefore follow a different route, building on work by Andreou and Werker (2012) on residual based statistics, and the martingale representations for matching estimators derived in Abadie and Imbens (2012).

3. LARGE SAMPLE DISTRIBUTION

In the first part of this section, we derive the large sample approximation to the sampling distribution of $\widehat{\tau}_N(\widehat{\theta}_N)$, and in the second part, we present the results for $\widehat{\tau}_{t,N}(\widehat{\theta}_N)$.

Let P^θ be the distribution of $Z = \{Y, W, X\}$ induced by the propensity score, $F(X'\theta)$, the marginal distribution of X, and the conditional distribution of Y given X and W. We index this distribution P^θ by θ, and will consider properties of estimators for different values of θ, under the same marginal distribution for X, and the same conditional distribution for Y given W and X. Given Assumption 1, the average treatment effect is equal to

$$\tau = E\big[Y(1) - Y(0)\big] = E\big[E[Y|W=1,X] - E[Y|W=0,X]\big].$$

From this equation, it can be seen that ATE does not depend on the propensity score; it only depends on the conditional distribution of Y given W and X and the marginal distribution of X. The average treatment effect for the treated is

$$\begin{aligned}\tau_t &= E\big[Y(1) - Y(0)|W=1\big]\\ &= E\big[E[Y|W=1,X] - E[Y|W=0,X]|W=1\big]\\ &= \frac{E\big[F\big(X'\theta^*\big)\big(E[Y|W=1,X] - E[Y|W=0,X]\big)\big]}{E\big[F\big(X'\theta^*\big)\big]}.\end{aligned}$$

In contrast to the average treatment effect, τ, the average treatment effect for the treated, τ_t, depends on the propensity score, and we make this dependence explicit by indexing τ_t by θ wherever appropriate. In particular, $\tau_t = \tau_t(\theta^*)$ is the average effect of the treatment on the treated.

To derive the large sample distribution of $\widehat{\tau}_N$ and $\widehat{\tau}_{t,N}$, we invoke some additional regularity conditions. First, we extend Assumption 2 to hold for all θ in a neighborhood of θ^*.

ASSUMPTION 4: (i) $\theta^* \in \text{int}(\Theta)$ *with* Θ *compact*, X *has a bounded support, and* $E[XX']$ *is nonsingular*; (ii) $F : \mathbb{R} \mapsto (0, 1)$ *is continuously differentiable with strictly positive and bounded derivative* f; (iii) *there exists a component of* X *that is continuously distributed, has nonzero coefficient in* θ^*, *and admits a continuous density function conditional on the rest of* X; *and* (iv) *there exists* $\varepsilon > 0$ *such that, for all* θ *with* $\|\theta - \theta^*\| \leq \varepsilon$, $E[Y|W = w, F(X'\theta) = p]$ *is Lipschitz-continuous in* p, $\text{var}(Y|W = w, F(X'\theta) = p)$ *is continuous in* p, *and there is* $\delta > 0$ *such that* $E[|Y|^{2+\delta}|W = w, F(X'\theta) = p]$ *is uniformly bounded*.

The case of only discrete regressors is left out from Assumption 4, but, as noted in Abadie and Imbens (2006), it is a simple case to treat separately. With only discrete regressors, each observation (or each treated observation in the case of ATET) can be matched to every observation with identical estimated propensity score value in the opposite treatment arm. In this case, the propensity score matching estimator is identical to the subclassification estimator in Cochran (1968) and Angrist (1998), and valid analytical and bootstrap standard errors can be easily derived.

We will study the behavior of certain statistics under sequences θ_N that are local to θ^*. Consider $Z_{N,i} = \{Y_{N,i}, W_{N,i}, X_{N,i}\}$ with distribution given by the local "shift" P^{θ_N} with $\theta_N = \theta^* + h/\sqrt{N}$, where h is a conformable vector of constants. Let $\Lambda_N(\theta|\theta')$ be the difference between the value of the log-likelihood function evaluated at θ and the value of the log-likelihood function evaluated at θ':

$$\Lambda_N\big(\theta|\theta'\big) = L(\theta|Z_{N,1}, \ldots, Z_{N,N}) - L\big(\theta'|Z_{N,1}, \ldots, Z_{N,N}\big).$$

Let $\Delta_N(\theta)$ be the normalized score function, or central sequence,

$$\begin{aligned}\Delta_N(\theta) &= \frac{1}{\sqrt{N}}\frac{\partial}{\partial\theta}L(\theta|Z_{N,1}, \ldots, Z_{N,N}) \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_{N,i}\frac{W_{N,i} - F\big(X'_{N,i}\theta\big)}{F\big(X'_{N,i}\theta\big)\big(1 - F\big(X'_{N,i}\theta\big)\big)}f\big(X'_{N,i}\theta\big).\end{aligned}$$

Finally, let

$$I_\theta = E\left[\frac{f(X'\theta)^2}{F(X'\theta)(1-F(X'\theta))}XX'\right]$$

be the Fisher Information Matrix for θ. The expectation in this equation is taken over the marginal distribution of X, which does not depend on θ, so the indexing by θ solely reflects the value of θ where $f(X'\theta)$ and $F(X'\theta)$ are evaluated. The following intermediate lemma derives some important regularity properties of the propensity score model that will be needed for our derivations.

LEMMA 1: *Suppose that Assumptions 3, 4(i), and 4(ii) hold. Then, under* P^{θ_N},

$$(1) \qquad \Lambda_N(\theta^*|\theta_N) = -h'\Delta_N(\theta_N) - \frac{1}{2}h'I_{\theta^*}h + o_p(1),$$

$$(2) \qquad \Delta_N(\theta_N) \xrightarrow{d} N(0, I_{\theta^*}),$$

where I_{θ^*} *is not singular, and*

$$(3) \qquad \sqrt{N}(\widehat{\theta}_N - \theta_N) = I_{\theta^*}^{-1}\Delta_N(\theta_N) + o_p(1).$$

The proof of this proposition is available in the Supplemental Material (Abadie and Imbens (2016)). For regular parametric models, equation (1) can be established using Proposition 2.1.2 in Bickel, Klaassen, Ritov, and Wellner (1998). Also for regular parametric models, equation (2) is derived in the proof of Proposition 2.1.2 in Bickel et al. (1998). Equation (3) can be established using the same set of results in combination with classical conditions for asymptotic linearity of maximum likelihood estimators (see, e.g., van der Vaart (1998, Theorem 5.39), Lehmann and Romano (2005, Theorem 12.4.1)).

Let E_θ be the expectation operator with respect to the distributions P^θ. The following assumption is a regularity condition that will be used later in this section.

ASSUMPTION 5: $E_{\theta_N}[r(Y,W,X)|W, F(X'\theta_N)]$ *converges to* $E[r(Y,W,X)|W, F(X'\theta^*)]$ *almost surely, for any* $\mathbb{R}^{k+2}$*-to-*$\mathbb{R}$ *bounded and measurable function,* $r(y,w,x)$*, continuous in* x*, and any sequence,* $\theta_N \to \theta^*$.

Primitive conditions for this assumption are provided in the Supplemental Material.

3.1. *Large Sample Distribution for* $\widehat{\tau}_N(\widehat{\theta}_N)$

Our derivation of the limit distribution of $\sqrt{N}(\widehat{\tau}_N - \tau)$ is based on the techniques developed in Andreou and Werker (2012) to analyze the limit distribution of residual-based statistics. We proceed in three steps. First, we derive the joint limit distribution of $(\sqrt{N}(\widehat{\tau}_N(\theta_N) - \tau), \sqrt{N}(\widehat{\theta} - \theta_N), \Lambda_N(\theta^*|\theta_N))$ under P^{θ_N}.

PROPOSITION 2: *Suppose that Assumptions 1–5 hold. Then, under* P^{θ_N},

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_N(\theta_N) - \tau\big) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \\ \Lambda_N\big(\theta^*|\theta_N\big) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ -h' I_{\theta^*} h/2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' I_{\theta^*}^{-1} & -c'h \\ I_{\theta^*}^{-1} c & I_{\theta^*}^{-1} & -h \\ -h'c & -h' & h' I_{\theta^*} h \end{pmatrix}\right),$$

where

$$c = E\left[\left(\frac{\operatorname{cov}\big(X, \mu(1,X)|F\big(X'\theta^*\big)\big)}{F\big(X'\theta^*\big)} + \frac{\operatorname{cov}\big(X, \mu(0,X)|F\big(X'\theta^*\big)\big)}{1 - F\big(X'\theta^*\big)}\right) f\big(X'\theta^*\big)\right]. \tag{4}$$

All proofs for the results in this section are provided in the Appendix.

Asymptotic Normality of the first component, $\sqrt{N}(\widehat{\tau}_N(\theta_N) - \tau)$, under P^{θ^*} follows from Proposition 1. Asymptotic joint Normality of the last two components, $\sqrt{N}(\widehat{\theta}_N - \theta_N)$ and $\Lambda_N(\theta^*|\theta_N)$, follows from Lemma 1. Proposition 2 derives the joint large sample distribution of the three components under P^{θ_N}. The proof extends the martingale techniques of Abadie and Imbens (2012) to derive the result of the proposition.

In the second step of our argument, we use Le Cam's third lemma (e.g., van der Vaart (1998, p. 90)). Given the result of Proposition 2, Le Cam's third lemma implies that, under P^{θ^*},

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_N(\theta_N) - \tau\big) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} -c'h \\ -h \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' I_{\theta^*}^{-1} \\ I_{\theta^*}^{-1} c & I_{\theta^*}^{-1} \end{pmatrix}\right).$$

Substituting $\theta_N = \theta^* + h/\sqrt{N}$, this implies that (still under P^{θ^*}), for any $h \in \mathbb{R}^k$,

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_N\big(\theta^* + h/\sqrt{N}\big) - \tau\big) \\ \sqrt{N}\big(\widehat{\theta}_N - \theta^*\big) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} -c'h \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' I_{\theta^*}^{-1} \\ I_{\theta^*}^{-1} c & I_{\theta^*}^{-1} \end{pmatrix}\right). \tag{5}$$

If equation (5) was an exact result rather than an approximation based on convergence in distribution, it would directly lead to the result of interest. In that case, it would follow that

$$\sqrt{N}\big(\widehat{\tau}_N\big(\theta^* + h/\sqrt{N}\big) - \tau\big)\big|\sqrt{N}\big(\widehat{\theta}_N - \theta^*\big) = h \sim N\big(0, \sigma^2 - c' I_{\theta^*}^{-1} c\big)$$

(see, e.g., Goldberger (1991, p. 197)). Because $\sqrt{N}(\widehat{\theta}_N - \theta^*) = h$ implies $\theta^* + h/\sqrt{N} = \widehat{\theta}_N$, and thus implies that $\widehat{\tau}_N(\theta^* + h/\sqrt{N}) = \widehat{\tau}_N(\widehat{\theta}_N) = \widehat{\tau}_N$, the last displayed equation can also be written as

$$\sqrt{N}(\widehat{\tau}_N - \tau)\big|\sqrt{N}(\widehat{\theta}_N - \theta) = h \sim N\big(0, \sigma^2 - c' I_{\theta^*}^{-1} c\big).$$

Because this conditional distribution does not depend on h, this in turn implies that, under P^{θ^*}, unconditionally,

$$\sqrt{N}(\widehat{\tau}_N - \tau) \sim N\big(0, \sigma^2 - c' I_{\theta^*}^{-1} c\big),$$

which is the result we are looking for: the distribution of the matching estimator based on matching on the estimated propensity score.

A challenge formalizing this argument is that convergence of $\sqrt{N}(\widehat{\tau}_N - \tau)|\sqrt{N}(\widehat{\theta}_N - \theta^*) = h$ involves convergence in a conditioning event. To overcome this challenge, in the third step of the argument, we employ a Le Cam discretization device, as proposed in Andreou and Werker (2012). Consider a grid of cubes in $\mathbb{R}^k$ with sides of length $d/\sqrt{N}$, for arbitrary positive d. Then $\bar{\theta}_N$ is the discretized estimator, defined as the midpoint of the cube $\widehat{\theta}_N$ belongs to. If $\widehat{\theta}_{N,j}$ is the jth component of the k-vector $\widehat{\theta}_N$, then the jth component of the k-vector $\bar{\theta}_N$ is $\bar{\theta}_{N,j} = (d/\sqrt{N})[\sqrt{N}\widehat{\theta}_{N,j}/d]$, where $[\cdot]$ is the nearest integer function. Now we can state the main result of the paper.

THEOREM 1: *Suppose Assumptions 1–5 hold. Then, under* P^{θ^*},

$$\lim_{d \downarrow 0} \lim_{N \to \infty} \Pr\big(\sqrt{N}\big(\sigma^2 - c' I_{\theta^*}^{-1} c\big)^{-1/2}\big(\widehat{\tau}_N(\bar{\theta}_N) - \tau\big) \le z\big) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx.$$

An implication of Theorem 1 is that we can approximate the distribution of $\sqrt{N}(\widehat{\tau}_N(\widehat{\theta}_N) - \tau)$ by a Normal distribution with mean zero and variance $\sigma^2 - c' I_{\theta^*}^{-1} c$. The result in Theorem 1 indicates that the adjustment to the stan-

dard error of the propensity score matching estimator for first step estimation of the propensity score is always negative, or zero in some special cases as discussed below. This implies that using matching on the estimated propensity score, rather than on the true propensity score, to estimate ATE increases precision in large samples. As we will see later, this gain in precision from using the estimated propensity score does not necessarily hold for the estimation of parameters different than ATE.

Equation (4) and the fact that X and W are independent conditional on the propensity score imply that if the covariance of X and $\mu(W, X)$ given $F(X'\theta^*)$ and W is equal to zero, then $c = 0$ and first step estimation of the propensity score does not affect the large sample variance of $\sqrt{N}(\widehat{\theta}_N - \theta^*)$. This would be the case if the propensity score provides no "dimension reduction," that is, if the propensity score is a bijective function of X. In that case, each value of the propensity score corresponds to only one value of X, so $\operatorname{cov}(X, \mu(W, X)|W, F(X'\theta^*)) = 0$ and, therefore, $c = 0$.

For concreteness, our derivations focus on the case of matching with replacement. However, as shown in Abadie and Imbens (2012), martingale representations analogous to the one employed in the proof of Proposition 2 exist for alternative matching estimators (e.g., estimators that construct the matches without using replacement). An inspection of the proof of Proposition 2 reveals that the adjustment term, $-c' I_{\theta^*} c$, does not depend on the type of matching employed to obtain the estimators. Therefore, the result in Theorem 1 translates easily to other matching settings. A different type of matching scheme may change the form of σ^2 in the result of Theorem 1, but not the adjustment term, $-c' I_{\theta^*} c$.

3.2. *Large Sample Distribution for* $\widehat{\tau}_{t,N}$

In this section, we consider the asymptotic distribution for $\sqrt{N}(\widehat{\tau}_{t,N} - \tau_t(\theta^*))$. The derivations are similar to those for the case of ATE, so we relegate details to the Supplemental Material. The following theorem provides the result.

THEOREM 2: *Suppose Assumptions 1–5 hold, and let*

$$
\begin{aligned}
c_t = {} & \frac{1}{E\big[F\big(X'\theta^*\big)\big]} E\big[Xf\big(X'\theta^*\big)\big(\bar{\mu}\big(1, F\big(X'\theta^*\big)\big) - \bar{\mu}\big(0, F\big(X'\theta^*\big)\big) - \tau_t\big)\big] \\
& + \frac{1}{E\big[F\big(X'\theta^*\big)\big]} E\Bigg[\Bigg(\operatorname{cov}\big(X, \mu(1, X)|F\big(X'\theta^*\big)\big) \\
& + \frac{F\big(X'\theta^*\big)}{1 - F\big(X'\theta^*\big)} \operatorname{cov}\big(X, \mu(0, X)|F\big(X'\theta^*\big)\big)\Bigg) f\big(X'\theta^*\big)\Bigg].
\end{aligned}
$$

Then, under P^{θ^*},

$$\lim_{d\downarrow 0}\lim_{N\to\infty}\Pr\left(\sqrt{N}\left(\sigma_t^2 - c_t' I_{\theta^*}^{-1} c_t + \frac{\partial\tau_t(\theta^*)}{\partial\theta}' I_{\theta^*}^{-1}\frac{\partial\tau_t(\theta^*)}{\partial\theta}\right)^{-1/2} \times \left(\widehat{\tau}_{t,N}(\bar{\theta}_N) - \tau_t\right) \le z\right) = \int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)dx.$$

Notice that, in contrast to the ATE case, the adjustment for first step estimation of the propensity score for the ATET estimator may result in a decrease *or* an increase in the standard error. The adjustment to the standard error of the ATET estimator will be positive, for example, if the propensity score does not provide "dimension reduction," so $c_t = 0$, and $\partial\tau_t(\theta^*)/\partial\theta \neq 0$ (which will typically be the case if the average effect of the treatment varies with the covariates, X). In contrast, if $c_t \neq 0$ and $\partial\tau_t(\theta^*)/\partial\theta = 0$, the adjustment is negative. Like for the case of ATE, it can be shown that the adjustment term does not depend on the particular type of matching estimator of ATET.

4. ESTIMATION OF THE ASYMPTOTIC VARIANCE

In this section, we discuss estimation of the large sample variances of ATE and ATET adjusting for first step estimation of the propensity score. As shown in the previous section, the asymptotic variance for $\sqrt{N}(\widehat{\tau}_N - \tau)$ is

$$\sigma_{\text{adj}}^2 = \sigma^2 - c' I_{\theta^*}^{-1} c, \tag{6}$$

and the asymptotic variance for $\sqrt{N}(\widehat{\tau}_{t,N} - \tau_t)$ is

$$\sigma_{t,\text{adj}}^2 = \sigma_t^2 - c_t' I_{\theta^*}^{-1} c_t + \frac{\partial\tau_t(\theta^*)}{\partial\theta}' I_{\theta^*}^{-1}\frac{\partial\tau_t(\theta^*)}{\partial\theta}. \tag{7}$$

To estimate σ_{adj}^2 and $\sigma_{t,\text{adj}}^2$, we define estimators of each of the components of the right-hand sides of equations (6) and (7). First, estimation of the information matrix, I_{θ^*}, is standard:

$$\widehat{I}_{\theta^*} = \frac{1}{N}\sum_{i=1}^{N}\frac{f(X_i'\widehat{\theta}_N)^2}{F(X_i'\widehat{\theta}_N)(1 - F(X_i'\widehat{\theta}_N))}X_i X_i'.$$

Consider next estimation of the variances corresponding to matching on the true propensity score, σ^2 and σ_t^2. For these components, we use estimators

that are based on those in Abadie and Imbens (2006). Let $K_{M,\theta}(i)$ be the number of times that observation i is used as a match (when matching on $F(X'\theta)$):

$$K_{M,\theta}(i) = \sum_{j=1}^{N} 1_{[i \in \mathcal{J}_M(j,\theta)]}, \tag{8}$$

and let $\widehat{\bar{\sigma}}^2(W_i, F(X_i'\theta^*))$ be an asymptotically unbiased (but not necessarily consistent) estimator of $\bar{\sigma}^2(W_i, F(X_i'\theta^*))$. The Abadie and Imbens (2006) variance estimators are

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left((2W_i-1)\left(Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i,\widehat{\theta})} Y_j\right) - \widehat{\tau}_N\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\left(\left(\frac{K_{M,\widehat{\theta}}(i)}{M}\right)^2 + \frac{2M-1}{M}\left(\frac{K_{M,\widehat{\theta}}(i)}{M}\right)\right)\widehat{\bar{\sigma}}^2\big(W_i, F\big(X_i'\theta^*\big)\big)$$

and

$$\widehat{\sigma}_t^2 = \frac{N}{N_1^2}\sum_{i=1}^{N} W_i\left(Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i,\widehat{\theta})} Y_j - \widehat{\tau}_{t,N}\right)^2 + \frac{N}{N_1^2}\sum_{i=1}^{N}(1-W_i)\left(\frac{K_{M,\widehat{\theta}}(i)\big(K_{M,\widehat{\theta}}(i)-1\big)}{M^2}\right)\widehat{\bar{\sigma}}^2\big(W_i, F\big(X_i'\theta^*\big)\big).$$

To obtain the estimator $\widehat{\bar{\sigma}}^2(W_i, F(X_i'\theta^*))$, let $\mathcal{H}_L(i,\theta)$ be the set of units in the same treatment arm as unit i that have the closest L values of $F(X'\theta)$ to $F(X_i'\theta)$,

$$\mathcal{H}_L(i,\theta) = \left\{ j = 1,\ldots,N : W_j = W_i, \left(\sum_{k:W_k=W_i} 1_{[|F(X_i'\theta)-F(X_k'\theta)| \le |F(X_i'\theta)-F(X_j'\theta)|]}\right) \le L \right\},$$

where L is generic notation for a (small) positive integer. Later, we will also use the set $\mathcal{H}_L^{(-i)}(i,\theta)$, which is similarly defined but excludes i:

$$\mathcal{H}_L^{(-i)}(i,\theta) = \left\{ j = 1,\ldots,N : i \ne j, W_j = W_i, \left(\sum_{k:W_k=W_i, k\ne i} 1_{[|F(X_i'\theta)-F(X_k'\theta)| \le |F(X_i'\theta)-F(X_j'\theta)|]}\right) \le L \right\}.$$

The sets $\mathcal{H}_L(i,\theta)$, $\mathcal{H}_L^{(-i)}(i,\theta)$, and $\mathcal{J}_L(i,\theta)$ (this last one defined in Section 2) will be used to estimate the different components of σ^2_{adj} and $\sigma^2_{t,\text{adj}}$. The value of L can vary for different components.

For $L \geq 2$ (typically, $L = 2$), consider the following matching estimator of $\bar{\sigma}^2(W_i, F(X_i'\theta^*))$:

$$\widehat{\bar{\sigma}}^2\big(W_i, F\big(X_i'\theta^*\big)\big) = \frac{1}{L-1} \sum_{j \in \mathcal{H}_L(i,\widehat{\theta}_N)} \left(Y_j - \frac{1}{L} \sum_{k \in \mathcal{H}_L(i,\widehat{\theta}_N)} Y_k \right)^2.$$

That is, $\widehat{\bar{\sigma}}^2(W_i, F(X_i'\theta^*))$ is a local variance estimator that uses information only from units with the same value of W as unit i and with similar values of $F(X'\widehat{\theta}_N)$. Because $\widehat{\theta}_N - \theta^*$ converges in probability to zero, $\widehat{\bar{\sigma}}^2(W_i, F(X_i'\theta^*))$ becomes asymptotically unbiased as $N \to \infty$. But because L is fixed, the variance of $\widehat{\bar{\sigma}}^2(W_i, F(X_i'\theta^*))$ does not converge to zero and $\widehat{\bar{\sigma}}^2(W_i, F(X_i'\theta^*))$ is not consistent for $\bar{\sigma}^2(W_i, F(X_i'\theta^*))$. The objects of interest, however, are σ^2 and σ_t^2, rather than $\bar{\sigma}^2(W_i, F(X_i'\theta^*))$. The estimators $\widehat{\sigma}^2$ and $\widehat{\sigma}_t^2$ average terms that are bounded in probability and asymptotically unbiased. As a result, as shown in Abadie and Imbens (2006), $\widehat{\sigma}^2$ and $\widehat{\sigma}_t^2$ are consistent for σ^2 and σ_t^2, respectively.

Next consider estimation of c and c_t. Notice first that

$$\begin{aligned}
&\operatorname{cov}\big(X, Y | F\big(X'\theta^*\big), W = 1\big) \\
&\quad = \operatorname{cov}\big(X, \mu(1,X) | F\big(X'\theta^*\big), W = 1\big) \\
&\qquad + \operatorname{cov}\big(X, Y - \mu(1,X) | F\big(X'\theta^*\big), W = 1\big) \\
&\quad = \operatorname{cov}\big(X, \mu(1,X) | F\big(X'\theta^*\big)\big) \\
&\qquad + \operatorname{cov}\big(X, Y - \mu(1,X) | F\big(X'\theta^*\big), W = 1\big).
\end{aligned}$$

Using the Law of Iterated Expectations,

$$\begin{aligned}
&\operatorname{cov}\big(X, Y - \mu(1,X) | F\big(X'\theta^*\big), W = 1\big) \\
&\quad = E\big[X\big(Y - \mu(1,X)\big) | F\big(X'\theta^*\big), W = 1\big] \\
&\qquad - E\big[X | F\big(X'\theta^*\big), W = 1\big] \\
&\qquad \times E\big[Y - \mu(1,X) | F\big(X'\theta^*\big), W = 1\big] \\
&\quad = E\big[X\big(\mu(1,X) - \mu(1,X)\big) | F\big(X'\theta^*\big), W = 1\big] \\
&\qquad - E\big[X | F\big(X'\theta^*\big), W = 1\big] \\
&\qquad \times E\big[\mu(1,X) - \mu(1,X) | F\big(X'\theta^*\big), W = 1\big] \\
&\quad = 0.
\end{aligned}$$

Therefore,

$$\operatorname{cov}\bigl(X, \mu(1, X)|F\bigl(X'\theta^*\bigr)\bigr) = \operatorname{cov}\bigl(X, Y|F\bigl(X'\theta^*\bigr), W = 1\bigr),$$

and the analogous result is valid conditional on $W_i = 0$:

$$\operatorname{cov}\bigl(X, \mu(0, X)|F\bigl(X'\theta^*\bigr)\bigr) = \operatorname{cov}\bigl(X, Y|F\bigl(X'\theta^*\bigr), W = 0\bigr).$$

If $W_i = w$, $\operatorname{cov}(X_i, \mu(w, X_i)|F(X_i'\theta^*))$ can be estimated as

$$\widehat{\operatorname{cov}}\bigl(X_i, \mu(w, X_i)|F\bigl(X_i'\theta^*\bigr)\bigr) = \frac{1}{L-1}\sum_{j\in\mathcal{H}_L(i,\widehat{\theta}_N)}\Biggl(X_j - \frac{1}{L}\sum_{k\in\mathcal{H}_L(i,\widehat{\theta}_N)} X_k\Biggr)\Biggl(Y_j - \frac{1}{L}\sum_{k\in\mathcal{H}_L(i,\widehat{\theta}_N)} Y_k\Biggr),$$

for $L \geq 2$ (typically, $L = 2$). If $W_i \neq w$, then

$$\widehat{\operatorname{cov}}\bigl(X_i, \mu(w, X_i)|F\bigl(X_i'\theta^*\bigr)\bigr) = \frac{1}{L-1}\sum_{j\in\mathcal{J}_L(i,\widehat{\theta}_N)}\Biggl(X_j - \frac{1}{L}\sum_{k\in\mathcal{J}_L(i,\widehat{\theta}_N)} X_k\Biggr)\Biggl(Y_j - \frac{1}{L}\sum_{k\in\mathcal{J}_L(i,\widehat{\theta}_N)} Y_k\Biggr),$$

also for $L \geq 2$ (typically, $L = 2$). Like $\widehat{\widehat{\sigma}}^2(W_i, F(X_i'\theta^*))$, the estimators $\widehat{\operatorname{cov}}(X_i, \mu(w, X_i)|F(X_i'\theta^*))$ are asymptotically unbiased and bounded in probability. This allows us to construct a consistent analog estimator of c that averages $\widehat{\operatorname{cov}}(X_i, \mu(w, X_i)|F(X_i'\theta^*))$ over the sample:

$$\widehat{c} = \frac{1}{N}\sum_{i=1}^{N}\Biggl(\frac{\widehat{\operatorname{cov}}\bigl(X_i, \mu(1, X_i)|F\bigl(X_i'\theta^*\bigr)\bigr)}{F\bigl(X_i'\widehat{\theta}_N\bigr)} + \frac{\widehat{\operatorname{cov}}\bigl(X_i, \mu(0, X_i)|F\bigl(X_i'\theta^*\bigr)\bigr)}{1 - F\bigl(X_i'\widehat{\theta}_N\bigr)}\Biggr) f\bigl(X_i'\widehat{\theta}_N\bigr).$$

For c_t, we propose separate estimators for the two components, $c_{t,1}$ and $c_{t,2}$, where

$$c_{t,1} = \frac{1}{E\bigl[F\bigl(X'\theta^*\bigr)\bigr]}E\bigl[Xf\bigl(X'\theta^*\bigr) \times \bigl(\bar{\mu}\bigl(1, F\bigl(X'\theta^*\bigr)\bigr) - \bar{\mu}\bigl(0, F\bigl(X'\theta^*\bigr)\bigr) - \tau_t\bigr)\bigr],$$

$$c_{t,2} = \frac{1}{E\bigl[F\bigl(X'\theta^*\bigr)\bigr]}E\Biggl[\Biggl(\operatorname{cov}\bigl(X, \mu(1, X)|F\bigl(X'\theta^*\bigr)\bigr) + \frac{F\bigl(X'\theta^*\bigr)}{1 - F\bigl(X'\theta^*\bigr)}\operatorname{cov}\bigl(X, \mu(0, X)|F\bigl(X'\theta^*\bigr)\bigr)\Biggr) f\bigl(X'\theta^*\bigr)\Biggr],$$

and $c_t = c_{t,1} + c_{t,2}$. The second component, $c_{t,2}$, is similar to c_t, and our proposed estimator for $c_{t,2}$ is correspondingly similar to the estimator for c:

$$\widehat{c}_{t,2} = \frac{1}{N_1}\sum_{i=1}^{N}\Bigg(\widehat{\mathrm{cov}}\big(X_i, \mu(1, X_i)|F\big(X_i'\theta^*\big)\big) + \frac{F\big(X_i'\widehat{\theta}_N\big)}{1 - F\big(X_i'\widehat{\theta}_N\big)}\widehat{\mathrm{cov}}\big(X_i, \mu(0, X_i)|F\big(X_i'\theta^*\big)\big)\Bigg) f\big(X_i'\widehat{\theta}_N\big). \tag{9}$$

The first component, $c_{t,1}$, involves the regression functions $\bar{\mu}(w, F(X'\theta^*))$. We estimate these regression functions using matching:

$$\widehat{\bar{\mu}}\big(0, F\big(X_i'\theta^*\big)\big) = \begin{cases} \dfrac{1}{L}\displaystyle\sum_{j\in\mathcal{H}_L^{(-i)}(i,\widehat{\theta}_N)} Y_j & \text{if } W_i = 0, \\ \dfrac{1}{L}\displaystyle\sum_{j\in\mathcal{J}_L(i,\widehat{\theta}_N)} Y_j & \text{if } W_i = 1, \end{cases}$$

and

$$\widehat{\bar{\mu}}\big(1, F\big(X_i'\theta^*\big)\big) = \begin{cases} \dfrac{1}{L}\displaystyle\sum_{j\in\mathcal{J}_L(i,\widehat{\theta}_N)} Y_j & \text{if } W_i = 0, \\ \dfrac{1}{L}\displaystyle\sum_{j\in\mathcal{H}_L^{(-i)}(i,\widehat{\theta}_N)} Y_j & \text{if } W_i = 1, \end{cases}$$

for $L \geq 1$ (typically, $L = 1$). Our proposed estimator for $c_{t,1}$ is

$$\widehat{c}_{t,1} = \frac{1}{N_1}\sum_{i=1}^{N} X_i f\big(X_i'\widehat{\theta}_N\big)\big(\widehat{\bar{\mu}}\big(1, F\big(X_i'\theta^*\big)\big) - \widehat{\bar{\mu}}\big(0, F\big(X_i'\theta^*\big)\big) - \widehat{\tau}_t\big).$$

Notice that if $W_i = w$, the estimator of $\bar{\mu}(w, F(X_i'\theta^*))$ is an average over $\mathcal{H}_L^{(-i)}(i, \widehat{\theta}_N)$, rather than over $\mathcal{H}_L(i, \widehat{\theta}_N)$. If observation i was not excluded to estimate $\bar{\mu}(W_i, F(X_i'\theta^*))$, then Y_i would be one of the terms of the average $\widehat{\bar{\mu}}(W_i, F(X_i'\theta^*))$. Therefore, if observation i was not excluded, the estimator $\widehat{c}_{t,1}$ would contain terms of the type $X_i f(X_i'\widehat{\theta}_N)(Y_i - \widehat{\bar{\mu}}(0, F(X_i'\theta^*)) - \widehat{\tau}_t)$ when $W_i = 1$ and terms of the type $X_i f(X_i'\widehat{\theta}_N)(\widehat{\bar{\mu}}(1, F(X_i'\theta^*)) - Y_i - \widehat{\tau}_t)$ when $W_i = 0$. These terms estimate $E[Xf(X'\theta^*)(\mu(1, X) - \bar{\mu}(0, F(X'\theta^*)) - \tau_t)]$ and $E[Xf(X'\theta^*)(\bar{\mu}(1, F(X'\theta^*)) - \mu(0, X) - \tau_t)]$, respectively, rather than $E[Xf(X'\theta^*)(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*)) - \tau_t)]$. To avoid this problem, we exclude observation i for the estimation of $\bar{\mu}(W_i, F(X_i'\theta^*))$.

For the remaining variance component, $\partial\tau_t(\theta^*)/\partial\theta$, notice that

$$\frac{\partial\tau_t}{\partial\theta}(\theta^*) = \frac{1}{E\big[F\big(X'\theta^*\big)\big]}E\big[Xf\big(X'\theta^*\big)\big(Y(1)-Y(0)-\tau_t\big)\big]$$
$$= \frac{1}{E\big[F\big(X'\theta^*\big)\big]}E\big[Xf\big(X'\theta^*\big)\big(\mu(1,X)-\mu(0,X)-\tau_t\big)\big].$$

To estimate this component, we need to estimate the regression functions $\mu(1,X)$ and $\mu(0,X)$, which is done by matching on the covariates rather than on the propensity score. Define the matching set (on covariates):

$$\mathcal{J}_L^X(i) = \Bigg\{ j=1,\ldots,N : W_j = 1-W_i,$$
$$\Bigg(\sum_{k:W_k=1-W_i} 1_{[\|X_i-X_k\|\le\|X_i-X_j\|]} \Bigg) \le L \Bigg\}.$$

Our estimator of $\partial\tau_t(\theta^*)/\partial\theta$ is

$$\widehat{\frac{\partial\tau_t}{\partial\theta}} = \frac{1}{N_1}\sum_{i=1}^{N} X_i f\big(X_i'\widehat{\theta}_N\big)\Bigg((2W_i-1)\Bigg(Y_i - \frac{1}{L}\sum_{j\in\mathcal{J}_L^X(i)} Y_j\Bigg) - \widehat{\tau}_t\Bigg).$$

Putting these results together, our estimator of the large sample variance of the propensity score matching estimator of the average treatment effect, adjusted for first step estimation of the propensity score, is

$$\widehat{\sigma}^2_{\text{adj}} = \widehat{\sigma^2} - \widehat{c}'\widehat{I}^{-1}_{\theta^*}\widehat{c}.$$

The corresponding estimator for the variance for the estimator for the average effect for the treated is

$$\widehat{\sigma}^2_{\text{adj},t} = \widehat{\sigma}^2_t - \widehat{c}_t'\widehat{I}^{-1}_{\theta^*}\widehat{c}_t + \widehat{\frac{\partial\tau_t}{\partial\theta}}'\widehat{I}^{-1}_{\theta^*}\widehat{\frac{\partial\tau_t}{\partial\theta}}.$$

Consistency of these estimators for fixed L can be shown using the results in Abadie and Imbens (2006) and the arguments employed in Section 3.

5. CONCLUSIONS AND EXTENSIONS

In this article, we derive the large sample distribution of propensity score matching estimators for the case where the propensity score is unknown and needs to be estimated in a first step prior to matching. We show that first step estimation of the propensity score generally affects the asymptotic variance of matching estimators, and derive adjustments for propensity score matching

estimators of ATE and ATET. These results allow, for the first time, valid large sample inference for estimators that use matching on the estimated propensity score.

For concreteness, we frame the article within the context of estimation of average treatment effects under the assumption that treatment assignment is independent of potential outcomes conditional on a set of covariates, X (Assumption 1(i)). Without this assumption, the results of this article apply still to the estimation of the "controlled comparison" parameters

$$E\big[E[Y|X,W=1]-E[Y|X,W=0]\big]$$

and

$$E[Y|W=1]-E\big[E[Y|X,W=0]|W=1\big],$$

which have the same form as ATE and ATET parameters but lack a causal interpretation in the absence of Assumption 1(i). Controlled contrasts are the building blocks of Oaxaca–Blinder-type decompositions, commonly applied in economics (Oaxaca (1973), Blinder (1973), DiNardo, Fortin, and Lemieux (1996)). The ideas and results in this article can easily be applied to other contexts where it is required to adjust for differences in the distribution of covariates between two samples. An important example is estimation with missing data when missingness is random conditional on a set of covariates (see, e.g., Little and Rubin (2002), Wooldridge (2007)).

APPENDIX

Before proving Proposition 2, we introduce some additional notation. Using the definition of $K_{M,\theta}(i)$ in (8), the estimator $\widehat{\tau}_N(\theta)$ can be written as

$$\widehat{\tau}_N(\theta)=\frac{1}{N}\sum_{i=1}^{N}(2W_i-1)\left(1+\frac{K_{M,\theta}(i)}{M}\right)Y_i.$$

Define $\bar{\mu}_\theta(w,p)=E_\theta[Y|W=w,F(X'\theta)=p]$, where E_θ is the expectation operator under P^θ,

$$\begin{aligned}D_N(\theta)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\big(\bar{\mu}_\theta\big(1,F\big(X_i'\theta\big)\big)-\bar{\mu}_\theta\big(0,F\big(X_i'\theta\big)\big)-\tau\big)\\ +\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(2W_i-1)\left(1+\frac{K_{M,\theta}(i)}{M}\right)\big(Y_i-\bar{\mu}_\theta\big(W_i,F\big(X_i'\theta\big)\big)\big)\end{aligned}$$

and

$$R_N(\theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (2W_i - 1) \times \left(\bar{\mu}_\theta\big(1 - W_i, F\big(X_i'\theta\big)\big) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \bar{\mu}_\theta\big(1 - W_i, F\big(X_j'\theta\big)\big) \right).$$

Now the normalized estimator can be written as

$$\sqrt{N}\big(\widehat{\tau}_N(\theta) - \tau\big) = D_N(\theta) + R_N(\theta).$$

PROOF OF PROPOSITION 2: It can be seen that the result of Lemma S.1 in the Supplemental Material (Abadie and Imbens (2016)) holds uniformly in θ for $\|\theta - \theta^*\| \leq \varepsilon$. This implies $R_N(\theta_N) \xrightarrow{p} 0$. Therefore, in order to prove the result in the proposition, it suffices to prove that, under P^{θ_N},

$$\begin{pmatrix} D_N(\theta_N) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \\ \Lambda_N\big(\theta^* | \theta_N\big) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ -h' I_{\theta^*} h/2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' I_{\theta^*}^{-1} & -c'h \\ I_{\theta^*}^{-1} c & I_{\theta^*}^{-1} & -h \\ -h'c & -h' & h' I_{\theta^*} h \end{pmatrix} \right).$$

By Lemma 1, under P^{θ_N},

$$\Lambda_N\big(\theta^* | \theta_N\big) = -h' \Delta_N(\theta_N) - \frac{1}{2} h' I_{\theta^*} h + o_p(1)$$

and

$$\sqrt{N}(\widehat{\theta}_N - \theta_N) = I_{\theta^*}^{-1} \Delta_N(\theta_N) + o_p(1).$$

Therefore, it suffices to prove that, under P^{θ_N},

$$\text{(A.1)} \qquad \begin{pmatrix} D_N(\theta_N) \\ \Delta_N(\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' \\ c & I_{\theta^*} \end{pmatrix} \right).$$

To prove (A.1), we extend the martingale representation of matching estimators (Abadie and Imbens (2012)) to allow for estimation of the propensity

score. Consider the linear combination $C_N = z_1 D_N(\theta_N) + z_2' \Delta_N(\theta_N)$:

$$
\begin{aligned}
C_N = {} & z_1 \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bigl(\bar{\mu}_{\theta_N}\bigl(1, F\bigl(X_{N,i}'\theta_N\bigr)\bigr) - \bar{\mu}_{\theta_N}\bigl(0, F\bigl(X_{N,i}'\theta_N\bigr)\bigr) - \tau\bigr) \\
& + z_1 \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (2W_{N,i} - 1)\left(1 + \frac{K_{M,\theta_N}(i)}{M}\right) \\
& \times \bigl(Y_{N,i} - \bar{\mu}_{\theta_N}\bigl(W_{N,i}, F\bigl(X_{N,i}'\theta_N\bigr)\bigr)\bigr) \\
& + z_2' \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_{N,i} \frac{W_{N,i} - F\bigl(X_{N,i}'\theta_N\bigr)}{F\bigl(X_{N,i}'\theta_N\bigr)\bigl(1 - F\bigl(X_{N,i}'\theta_N\bigr)\bigr)} f\bigl(X_{N,i}'\theta_N\bigr).
\end{aligned}
$$

We analyze C_N using martingale methods. First, notice that

$$
C_N = \sum_{k=1}^{3N} \xi_{N,k},
$$

where

$$
\begin{aligned}
\xi_{N,k} = {} & z_1 \frac{1}{\sqrt{N}} \bigl(\bar{\mu}_{\theta_N}\bigl(1, F\bigl(X_{N,k}'\theta_N\bigr)\bigr) - \bar{\mu}_{\theta_N}\bigl(0, F\bigl(X_{N,k}'\theta_N\bigr)\bigr) - \tau\bigr) \\
& + z_2' \frac{1}{\sqrt{N}} E\bigl[X_{N,k} | F\bigl(X_{N,k}'\theta_N\bigr)\bigr] \\
& \times \frac{W_{N,k} - F\bigl(X_{N,k}'\theta_N\bigr)}{F\bigl(X_{N,k}'\theta_N\bigr)\bigl(1 - F\bigl(X_{N,k}'\theta_N\bigr)\bigr)} f\bigl(X_{N,k}'\theta_N\bigr)
\end{aligned}
$$

for $1 \le k \le N$,

$$
\begin{aligned}
\xi_{N,k} = {} & z_2' \frac{1}{\sqrt{N}} \bigl(X_{N,k-N} - E\bigl[X_{N,k-N} | F\bigl(X_{N,k-N}'\theta_N\bigr)\bigr]\bigr) \\
& \times \frac{\bigl(W_{N,k-N} - F\bigl(X_{N,k-N}'\theta_N\bigr)\bigr) f\bigl(X_{N,k-N}'\theta_N\bigr)}{F\bigl(X_{N,k-N}'\theta_N\bigr)\bigl(1 - F\bigl(X_{N,k-N}'\theta_N\bigr)\bigr)} \\
& + z_1 \frac{1}{\sqrt{N}} (2W_{N,k-N} - 1)\left(1 + \frac{K_{M,\theta_N}(k-N)}{M}\right) \\
& \times \bigl(\mu(W_{N,k-N}, X_{N,k-N}) - \bar{\mu}_{\theta_N}\bigl(W_{N,k-N}, F\bigl(X_{N,k-N}'\theta_N\bigr)\bigr)\bigr)
\end{aligned}
$$

for $N+1 \le k \le 2N$, and

$$\xi_{N,k} = z_1 \frac{1}{\sqrt{N}}(2W_{N,k-2N}-1)\left(1+\frac{K_{M,\theta_N}(k-2N)}{M}\right) \times \big(Y_{N,k-2N} - \mu(W_{N,k-2N}, X_{N,k-2N})\big),$$

for $2N+1 \le k \le 3N$. Consider the σ-fields $\mathcal{F}_{N,k} = \sigma\{W_{N,1}, \ldots, W_{N,k}, X'_{N,1}\theta_N, \ldots, X'_{N,k}\theta_N\}$ for $1 \le k \le N$, $\mathcal{F}_{N,k} = \sigma\{W_{N,1}, \ldots, W_{N,N}, X'_{N,1}\theta_N, \ldots, X'_{N,N}\theta_N, X_{N,1}, \ldots, X_{N,k-N}\}$ for $N+1 \le k \le 2N$, and $\mathcal{F}_{N,k} = \sigma\{W_{N,1}, \ldots, W_{N,N}, X_{N,1}, \ldots, X_{N,N}, Y_{N,1}, \ldots, Y_{N,k-N}\}$ for $2N+1 \le k \le 3N$. Then,

$$\left\{\sum_{j=1}^{i} \xi_{N,j}, \mathcal{F}_{N,i}, 1 \le i \le 3N\right\}$$

is a martingale for each $N \ge 1$. Therefore, the limiting distribution of C_N can be studied using a Martingale Central Limit Theorem (e.g., Theorem 35.12 in Billingsley (1995, p. 476); importantly, notice that this theorem allows that the probability space varies with N). Because of Assumption 4, and because $K_{M,\theta}(i)$ has uniformly bounded moments (see Abadie and Imbens (2016)), it follows that

$$\sum_{k=1}^{3N} E_{\theta_N}\big[|\xi_{N,k}|^{2+\delta}\big] \to 0 \quad \text{for some } \delta > 0.$$

Lindeberg's condition in Billingsley's theorem follows easily from the last equation (Lyapunov's condition). As a result, we obtain that, under P^{θ_N},

$$C_N \xrightarrow{d} N\big(0, \sigma_1^2 + \sigma_2^2 + \sigma_3^2\big),$$

where

$$\sigma_1^2 = \operatorname{plim} \sum_{k=1}^{N} E_{\theta_N}\big[\xi_{N,k}^2 | \mathcal{F}_{N,k-1}\big],$$

$$\sigma_2^2 = \operatorname{plim} \sum_{k=N+1}^{2N} E_{\theta_N}\big[\xi_{N,k}^2 | \mathcal{F}_{N,k-1}\big],$$

$$\sigma_3^2 = \operatorname{plim} \sum_{k=2N+1}^{3N} E_{\theta_N}\big[\xi_{N,k}^2 | \mathcal{F}_{N,k-1}\big].$$

Assumption 5 implies

$$\sigma_1^2 = z_1^2 E\big[\big(\bar{\mu}\big(1, F\big(X'\theta^*\big)\big) - \bar{\mu}\big(0, F\big(X'\theta^*\big)\big) - \tau\big)^2\big] + z_2' E\Bigg[\frac{f^2\big(X'\theta^*\big)}{F\big(X'\theta^*\big)\big(1 - F\big(X'\theta^*\big)\big)} \times E\big[X|F\big(X'\theta^*\big)\big]E\big[X'|F\big(X'\theta^*\big)\big]\Bigg]z_2.$$

Expectations of the sums of terms involving $(1 + K_{M,\theta_N}(i)/M)^2$ can be calculated as in Abadie and Imbens (2016). We obtain

$$\begin{aligned}\sigma_2^2 = {} & z_2' E\Bigg[\frac{f^2\big(X'\theta^*\big)}{F\big(X'\theta^*\big)\big(1 - F\big(X'\theta^*\big)\big)} \operatorname{var}\big(X|F\big(X'\theta^*\big)\big)\Bigg]z_2 \\ & + z_1^2 E\Bigg[\frac{\operatorname{var}\big(\mu(1, X)|F\big(X'\theta^*\big)\big)}{F\big(X'\theta^*\big)} + \frac{\operatorname{var}\big(\mu(0, X)|F\big(X'\theta^*\big)\big)}{1 - F\big(X'\theta^*\big)}\Bigg] \\ & + z_1^2 \frac{1}{2M} E\Bigg[\Bigg(\frac{1}{F\big(X'\theta^*\big)} - F\big(X'\theta^*\big)\Bigg) \operatorname{var}\big(\mu(1, X)|F\big(X'\theta^*\big)\big)\Bigg] \\ & + z_1^2 \frac{1}{2M} E\Bigg[\Bigg(\frac{1}{1 - F\big(X'\theta^*\big)} - \big(1 - F\big(X'\theta^*\big)\big)\Bigg) \\ & \times \operatorname{var}\big(\mu(0, X)|F\big(X'\theta^*\big)\big)\Bigg] \\ & + 2z_2' E\Bigg[\Bigg(\frac{\operatorname{cov}\big(X, \mu(1, X)|F\big(X'\theta^*\big)\big)}{F\big(X'\theta^*\big)} \\ & + \frac{\operatorname{cov}\big(X, \mu(0, X)|F\big(X'\theta^*\big)\big)}{1 - F\big(X'\theta^*\big)}\Bigg) f\big(X'\theta^*\big)\Bigg]z_1.\end{aligned}$$

Here we use the fact that, conditional on the propensity score, X is independent of W. Finally, notice that

$$\frac{1}{N}\sum_{i=1}^{N}\Bigg(1 + \frac{K_{M,\theta_N}(i)}{M}\Bigg)^2 \times \big(\operatorname{var}(Y_i|W_i, X_i) - E\big[\operatorname{var}(Y_i|W_i, X_i)|W_i, F\big(X_i'\theta_N\big)\big]\big)$$

is a sum of martingale differences

$$\zeta_{N,i} = \frac{1}{N}\left(1 + \frac{K_{M,\theta_N}(i)}{M}\right)^2 \times \left(\operatorname{var}(Y_i|W_i, X_i) - E\left[\operatorname{var}(Y_i|W_i, X_i)|W_i, F\left(X_i'\theta_N\right)\right]\right)$$

with respect to the filtration $\mathcal{F}_{N,i} = \sigma\{W_1, \ldots, W_N, X_1'\theta_N, \ldots, X_1'\theta_N, X_1, \ldots, X_i\}$. As a result, we obtain that, for $i > j$, $E[\zeta_{N,i}\zeta_{N,j}|\mathcal{F}_{N,i-1}] = E[\zeta_{N,i}|\mathcal{F}_{N,i-1}]\zeta_{N,j} = 0$. Therefore, using the Law of Iterated Expectations to eliminate the cross-products, we obtain

$$E\left[\left(\frac{1}{N}\sum_{i=1}^{N}\zeta_{N,i}\right)^2\right] = \frac{1}{N^2}E\left[\sum_{i=1}^{N}\zeta_{N,i}^2\right] \to 0.$$

Therefore,

$$\begin{aligned}
\sigma_3^2 &= z_1^2 \operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}\left(1 + \frac{K_{M,\theta_N}(i)}{M}\right)^2 E\left[\operatorname{var}(Y_i|W_i, X_i)|W_i, F\left(X_i'\theta_N\right)\right] \\
&= z_1^2 E\left[\frac{E\left[\operatorname{var}(Y|X, W=1)|F\left(X'\theta^*\right)\right]}{F\left(X'\theta^*\right)}\right. \\
&\quad \left. + \frac{E\left[\operatorname{var}(Y|X, W=0)|F\left(X'\theta^*\right)\right]}{1 - F\left(X'\theta^*\right)}\right] \\
&\quad + z_1^2\frac{1}{2M}E\left[\left(\frac{1}{F\left(X'\theta^*\right)} - F\left(X'\theta^*\right)\right)\right. \\
&\quad \left. \times E\left[\operatorname{var}(Y|X, W=1)|F\left(X'\theta^*\right)\right]\right] \\
&\quad + z_1^2\frac{1}{2M}E\left[\left(\frac{1}{1 - F\left(X'\theta^*\right)} - \left(1 - F\left(X'\theta^*\right)\right)\right)\right. \\
&\quad \left. \times E\left[\operatorname{var}(Y|X, W=0)|F\left(X'\theta^*\right)\right]\right].
\end{aligned}$$

Collecting terms and applying the fact that W is independent of X given $F(X'\theta)$, we obtain

$$\sigma_1^2 + \sigma_2^2 + \sigma_3^2 = z_1^2\sigma^2 + z_2'I_{\theta^*}z_2 + 2z_2'cz_1.$$

Hence, by the Martingale Central Limit Theorem and the Cramer–Wold device, under P^{θ_N},

$$\begin{pmatrix} D_N(\theta_N) \\ \Delta_N(\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' \\ c & I_{\theta^*} \end{pmatrix}\right),$$

proving (A.1) and thus Proposition 2. *Q.E.D.*

PROOF OF THEOREM 1: Given our preliminary results, Theorem 1 follows from Andreou and Werker (2012). *Q.E.D.*

The proof of Theorem 2 can be found in the Supplemental Material.

REFERENCES

ABADIE, A. (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72 (1), 1–19. [782]

ABADIE, A., AND G. W. IMBENS (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74 (1), 235–267. [781,784,785,789,795,796,799]

——— (2008): "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, 76 (6), 1537–1557. [782]

——— (2011): "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics*, 29 (1), 1–11. [784,787]

——— (2012): "A Martingale Representation for Matching Estimators," *Journal of the American Statistical Association*, 107 (498), 833–843. [788,791,793,801]

——— (2016): "Supplement to 'Matching on the Estimated Propensity Score'," *Econometrica Supplemental Material*, 84, http://dx.doi.org/10.3982/ECTA11293. [786,790,801,803,804]

ANDREOU, E., AND B. J. M. WERKER (2012): "An Alternative Asymptotic Analysis of Residual-Based Statistics," *Review of Economics and Statistics*, 94 (1), 88–99. [788,791,792,806]

ANGRIST, J. D. (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66 (2), 249–288. [789]

ANGRIST, J. D., AND G. M. KUERSTEINER (2011): "Causal Effects of Monetary Shocks: Semiparametric Conditional Independence Tests With a Multinomial Propensity Score," *Review of Economics and Statistics*, 93 (3), 725–747. [782]

BICKEL, P. J., C. A. KLAASSEN, Y. RITOV, AND J. A. WELLNER (1998): *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer. [790]

BILLINGSLEY, P. (1995): *Probability and Measure*. New York: Wiley. [803]

BLINDER, A. S. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8 (4), 436–455. [800]

COCHRAN, W. G. (1968): "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24 (2), 295–313. [789]

DAWID, A. P. (1979): "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society, Series B*, 41 (1), 1–31. [783]

DEHEJIA, R., AND S. WAHBA (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94 (448), 1053–1062. [782,787]

DINARDO, J., N. M. FORTIN, AND T. LEMIEUX (1996): "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach," *Econometrica*, 64 (5), 1001–1044. [800]

GOLDBERGER, A. S. (1991): *A Course in Econometrics*. Cambridge, MA: Harvard University Press. [792]

HAHN, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66 (2), 315–331. [781,783]

HECKMAN, J., H. ICHIMURA, AND P. TODD (1997): "Matching as an Econometric Evaluation Estimator: Evidence From a Job Training Programme," *Review of Economic Studies*, 64 (4), 605–654. [782]

——— (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65 (2), 261–294. [781,782]

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71 (4), 1161–1189. [782]

IMBENS, G. (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86 (1), 1–29. [781]

IMBENS, G., AND J. WOOLDRIDGE (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47 (1), 5–86. [781]

KHAN, S., AND E. TAMER (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78 (6), 2021–2042. [783]

KIEFER, N. M., AND T. J. VOGELSANG (2005): "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory*, 21, 1130–1164. [784]

LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing Statistical Hypothesis*. New York: Springer. [790]

LITTLE, R. J., AND D. B. RUBIN (2002): *Statistical Analysis With Missing Data* (Second Ed.). Hoboken, NJ: Wiley-Interscience. [800]

NEWEY, W. K., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. McFadden. Amsterdam: Elsevier Science. [782]

OAXACA, R. (1973): "Male–Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14 (3), 693–709. [800]

ROSENBAUM, P., AND D. B. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70 (1), 41–55. [781,783,784,786]

RUBIN, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66 (5), 688–701. [783,784]

SMITH, J., AND P. TODD (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125 (1–2), 305–353. [782]

VAN DER VAART, A. (1998): *Asymptotic Statistics*. New York: Cambridge University Press. [790,791]

WOOLDRIDGE, J. M. (2007): "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, 141 (2), 1281–1301. [782,800]

YATCHEW, A. (1997): "An Elementary Estimator of the Partial Linear Model," *Economics Letters*, 75, 135–143. [784]

John F. Kennedy School of Government, 79 John F. Kennedy Street, Cambridge, MA 02138, U.S.A. and NBER; alberto_abadie@harvard.edu

and

Stanford Graduate School of Business, 655 Knight Way, Stanford, CA 94305-7298, U.S.A. and NBER; imbens@stanford.edu.

Co-editor Elie Tamer handled this manuscript.

Manuscript received December, 2012; final revision received August, 2015.