The Variance of Non-Parametric Treatment Effect Estimators in the Presence of Clustering

NOTES

THE VARIANCE OF NON-PARAMETRIC TREATMENT EFFECT ESTIMATORS IN THE PRESENCE OF CLUSTERING*

Samuel G. Hanson and Adi Sunderam*

Abstract—Nonparametric estimators of treatment effects are often applied in settings where clustering may be important. We provide a general methodology for consistently estimating the variance of a large class of nonparametric estimators, including the simple matching estimator, in the presence of clustering. Software for implementing our variance estimator is available in Stata.

I. Introduction

AVERAGE treatment effects (ATEs) can be estimated using a variety of nonparametric techniques, including matching and propensity score–based estimators. These methods are usually applied in settings, such as program evaluation, where cross-sectional data have been collected. With such data, geographic shocks or omitted common factors may induce correlation across observations, even after controlling for treatment status and the covariates used in the research design. In other words, there may be a clustering problem. While a large literature has studied the consistency of nonparametric ATE estimators (see Imbens, 2004), there has been little discussion of the effects of clustering on the variance of these estimators. This is surprising given the significant attention devoted to clustering in parametric settings. Given the popularity of nonparametric estimators, and matching estimators in particular, the ability to compute cluster-robust standard errors for them is important.

In this note, we consider a setting where both residuals and treatment effects may be correlated within clusters. We provide a methodology for estimating the variance of matching estimators in this setting. There are two main challenges. First, matching estimators are highly nonsmooth functionals of the data and, as discussed further below, standard asymptotic arguments for smooth functionals (e.g., method-of-moments estimators) cannot be applied (Abadie & Imbens, 2006). Second, since they do not rely on consistent estimation of the underlying regression functions, matching methods do not generate estimated residuals, which are crucial to standard clustering adjustments (Moulton, 1990; Wooldridge, 2002). We surmount these challenges by extending the approach of Abadie and Imbens (2006, 2008). Our methodology generates quasi-residuals, which we use to compute a cluster-robust variance estimator that is consistent as the number of clusters grows large. While we focus on the matching estimator in this note, our methodology can easily be extended to a broader class of nonparametric treatment effect estimators.

The remainder of the paper is organized as follows. Section II defines the relevant class of matching estimators for the ATE and derives the clustering correction. Section III details our methodology for cluster robust variance estimation. In section IV, we explore the finite sample behavior of our variance estimator using a short Monte Carlo study. Section V briefly discusses extensions of our basic methodology and section VI concludes.

Received for publication January 9, 2009. Revision accepted for publication March 22, 2011.

* Both authors: Harvard Business School.

We thank Sergey Chernenko, Judd Kessler, participants at the Harvard Econometrics Seminar, Alberto Abadie (the editor), an anonymous referee, and especially Guido Imbens for extremely helpful comments and suggestions.

An online appendix is available at http://www.mitpressjournals.org/doi/suppl/10.1162/REST_a_00211.

II. Preliminaries

A. Setup and Notation

We consider matching estimators for the average effect of a binary treatment on some outcome. Let $j = 1, \ldots, J$ index clusters, $i = 1, \ldots, I_j$ index individuals in cluster j, and $N = \sum_j I_j$ be the total number of individuals. Let X_{ij} denote the vector of covariates and $W_{ij} \in \{0, 1\}$ the treatment status of individual ij. Also, let $\mathbf{X}$ denote the matrix of covariates and $\mathbf{W}$ the vector of treatment indicators for all N individuals. Let $(Y_{ij}(0), Y_{ij}(1))$ denote the potential outcomes given the control treatment or the active treatment, respectively, for individual ij. Of course, only $Y_{ij} = Y_{ij}(W_{ij})$ is observed. Under the standard unconfoundedness assumption that W_{ij} is independent of $(Y_{ij}(0), Y_{ij}(1))$ conditional on X_{ij}, we can write the conditional expectation of the outcome given treatment w and covariates $X = x$ as $\mu_w(x) = E[Y(w)|X = x]$. The average treatment effect for the subpopulation with $X = x$ is $\tau(x) = E[Y(1) - Y(0)|X = x] = \mu_1(x) - \mu_0(x)$. The population average treatment effect (PATE) is $\tau = E[\tau(X)]$, and the sample average treatment effect (SATE), conditional on $\mathbf{X}$, is $\overline{\tau(\mathbf{X})} = N^{-1} \sum_{i,j} \tau(X_{ij})$.

The matching estimator imputes unobserved potential outcomes for each individual by matching that individual with M individuals of the opposite treatment status. Specifically, let $\mathcal{J}_M(ij)$ be the set of indices of the M closest matches to unit ij with the opposite treatment status: $\#\mathcal{J}_M(ij) = M$ and for all $st \in \mathcal{J}_M(ij)$, $W_{st} = 1 - W_{ij}$ and $||X_{ij} - X_{st}|| \leq ||X_{ij} - X_{s't'}||$ for all $s't' \notin \mathcal{J}_M(ij)$. The estimator imputes missing outcomes as

$$\widehat{Y}_{ij}(0) = \begin{cases} Y_{ij} & \text{if } W_{ij} = 0 \\ \frac{1}{M} \sum_{st \in \mathcal{J}_M(ij)} Y_{st} & \text{if } W_{ij} = 1 \end{cases}$$

$$\text{and } \widehat{Y}_{ij}(1) = \begin{cases} \frac{1}{M} \sum_{st \in \mathcal{J}_M(ij)} Y_{st} & \text{if } W_{ij} = 0 \\ Y_{ij} & \text{if } W_{ij} = 1 \end{cases}.$$

When we define $K_M(ij) = \sum_{st} \mathbf{1}\{i, j \in \mathcal{J}_M(st)\}$ as the number of times observation ij is used as a match, the simple matching estimator is

$$\begin{aligned} \widehat{\tau}_M &= N^{-1} \sum\nolimits_{i,j} (\widehat{Y}_{ij}(1) - \widehat{Y}_{ij}(0)) \\ &= N^{-1} \sum\nolimits_{i,j} (2W_{ij} - 1) \left(1 + \frac{K_M(ij)}{M}\right) Y_{ij}. \end{aligned}$$

We consider a random-effects type setting in the sense that individual treatment effects, $Y_{ij}(1) - Y_{ij}(0)$, are assumed to be independent of the cluster-level shocks. Specifically, we assume that all members of cluster j are subject to the same cluster-level shock regardless of treatment status. Thus, potential outcomes can be decomposed as $Y_{ij}(w) = \mu_w(X_{ij}) + \eta_j + \omega_{ij}(w)$, and we have $Y_{ij}(1) - Y_{ij}(0) = \tau(X_{ij}) + \omega_{ij}(1) - \omega_{ij}(0)$, which is independent of η_j. In what follows, we suppress the $\omega_{ij}(w)$ notation and simply write observed outcomes

as $Y_{ij} = \mu_{W_{ij}}(X_{ij}) + \varepsilon_{ij} = \mu_{W_{ij}}(X_{ij}) + \eta_j + \omega_{ij}$. For simplicity, we assume that the cluster-level shocks are homoskedastic but the individual-specifics shocks may be heteroskedastic. That is, we assume $\varepsilon_{ij} = \eta_j + \omega_{ij}$, where $\eta_j \overset{iid}{\sim} (0, \sigma^2_\eta)$ and $\omega_{ij} \overset{iid}{\sim} (0, \sigma^2_\omega(X_{ij}, W_{ij}))$.[1] In addition to cluster-level correlation of the residuals, in many empirical settings it is likely that the individual treatment effects are correlated within clusters. Therefore, we also allow the covariates X_{ij}, and hence the treatment effects $\tau(X_{ij})$, to be correlated within clusters, but we assume they are independent across clusters.

Many empirical settings fit the assumptions laid out here. For instance, it would be sensible to assume clustering of both treatment effects and shocks at the school level for program evaluation exercises with school-level pay-for-grades treatments within school districts (Fryer, 2010). Clustering at the county level would be expected when evaluating individual specific job training treatments within counties (Hotz, Imbens, & Klerman, 2006).

B. *The Variance of Matching Estimators with Clustering*

Following Abadie and Imbens (2006), we write the difference between $\widehat{\tau}_M$ and τ as $\widehat{\tau}_M - \tau = (\overline{\tau(\mathbf{X})} - \tau) + E + B$, where $\overline{\tau(\mathbf{X})} = N^{-1}\sum_{i,j}\tau(X_{ij})$ is the sample average treatment effect conditional on $\mathbf{X}$ and

$$E = N^{-1}\sum_{i,j}(2W_{ij}-1)\left(1+\frac{K_M(ij)}{M}\right)\varepsilon_{ij}$$

$$B = N^{-1}\sum_{i,j}\left[(2W_{ij}-1)\left(1+\frac{K_M(ij)}{M}\right)\mu_{W_{ij}}(X_{ij}) - (\mu_1(X_{ij})-\mu_0(X_{ij}))\right].$$

Here E is a weighted sum of the residuals, and B is a conditional bias term.

As Imbens (2004) pointed out, the variance of the matching estimator depends on the quantity we are trying to estimate. If the SATE is the estimand of interest, then the normalized variance of the estimator is given by the conditional variance: $V^E \equiv \text{Var}[\sqrt{N}\widehat{\tau}_M|\mathbf{X},\mathbf{W}] = \text{Var}[\sqrt{N}E|\mathbf{X},\mathbf{W}]$. If the PATE ($\tau = E[\tau(X)]$) is the estimand of interest, then the normalized variance of the estimator is the marginal variance $V = V^E + V^{\tau(X)}$, where $V^{\tau(X)} \equiv \text{Var}[\sqrt{N}\,\overline{\tau(\mathbf{X})}]$ is the normalized variance of the SATE. In this note, we provide cluster-robust estimators for both the conditional and marginal variance.

Under our assumed error structure, we can write the conditional variance as

$$V^E = N^{-1}\sum_{i,j}\left(1+\frac{K_M(ij)}{M}\right)^2\sigma^2_\varepsilon(X_{ij},W_{ij}) + \underbrace{N^{-1}\sum_j\left[\sum_i\sum_{i'\neq i}(2W_{ij}-1)(2W_{i'j}-1)\times\left(1+\frac{K_M(ij)}{M}\right)\left(1+\frac{K_M(i'j)}{M}\right)\sigma^2_\eta\right]}_{\text{Clustering correction}} \quad (1)$$

where $\sigma^2_\varepsilon(X_{ij},W_{ij}) = \sigma^2_\eta + \sigma^2_\omega(X_{ij},W_{ij})$. The first term represents the conditional variance if we ignore the impact of clustering. The second term is the contribution of error clustering to the conditional variance. From equation (1) we see that the clustering correction is largest when all units in a given cluster have the same treatment status (when $W_{ij} = W_{i'j}$ for all i and i' in cluster j) so that all of the terms in the correction are positive. This parallels the linear OLS case where clustering matters more when covariates are more highly correlated within clusters (Greenwald, 1983; & Moulton, 1986).

The marginal variance of the matching estimator also depends on $V^{\tau(X)} = \text{Var}[\sqrt{N}\,\overline{\tau(\mathbf{X})}]$. If the X_{ij}, and hence the $\tau(X_{ij})$, were drawn independently within each cluster, then we would have $V^{\tau(X)} = E[(\tau(X_{ij})-\tau)^2]$. However, under our assumption that the X_{ij} are correlated within cluster, the identity $\sqrt{N}(\overline{\tau(\mathbf{X})}-\tau) = N^{-1/2}\sum_j\sum_i(\tau(X_{ij})-\tau)$ implies that $V^{\tau(X)} = N^{-1}J\cdot E[(\sum_i(\tau(X_{ij})-\tau))^2]$ where the expectation is taken across clusters.[2]

III. Cluster-Robust Variance Estimators

A. *Why Standard Variance Estimation Methods Fail*

Before outlining our estimation approach, we explain why traditional techniques for estimating cluster-robust variances, such as those from the literature on the generalized method of moments (GMM) (Bhattacharya, 2005; Wooldridge, 2006), are not applicable here. GMM is based on the assumption that we have a population moment condition such that $E[g(z_i,\theta)] = 0 \Leftrightarrow \theta = \theta_0$ for some known function $g(\mathbf{z}_i,\theta)$ that is specified a priori and does not depend on the sample under consideration. The GMM estimator is then defined using the sample analog of the population moment condition: $N^{-1}\sum_i g(\mathbf{z}_i,\hat{\theta}) = 0$. Traditional GMM asymptotics and cluster-robust variance estimators are based on the assumption that $g(z,\theta)$ is continuously differentiable in θ. The literature has extended these results to settings where $E[g(z,\theta)]$ is continuously differentiable even though $g(z,\theta)$ may not be. In these cases, the estimator will typically have an asymptotically linear representation, and existing techniques can be used to consistently estimate a cluster-robust variance.

Can we use these methods to estimate a cluster-robust variance for the matching estimator? For instance, one might reason that $\widehat{\tau}_M$ satisfies the sample moment condition $N^{-1}\sum_{i,j}[(2W_{i,j}-1)(1+M^{-1}K_M(ij|\mathbf{X},\mathbf{W}))Y_{i,j}-\widehat{\tau}_M] = 0$. Here we write $K_M(ij|\mathbf{X},\mathbf{W})$ to emphasize that $K_M(ij)$ depends on all the covariates and treatment assignments, not just those for unit ij. As we add observations, $K_M(ij|\mathbf{X},\mathbf{W})$ can rise or fall discretely, so the $K_M(ij|\mathbf{X},\mathbf{W})$, and hence $\widehat{\tau}_M$, are highly nonsmooth functionals of the data (Abadie & Imbens, 2006).[3] As a result, this condition is not based on the sample average of some known function, so matching estimators are not traditional GMM estimators.

However, one might still wonder whether existing techniques for asymptotically linear estimators might be used in this setting. Applied researchers often favor matching estimators with small, fixed M due to concerns about the conditional bias of estimators with large M. When the number of matches M is fixed, there is no evidence that matching estimators become asymptotically linear, which may explain the failure of standard bootstrapping methods for inference (Abadie & Imbens,

[1] While we assume $\eta_j \overset{iid}{\sim} (0,\sigma^2_\eta)$ for simplicity, our approach can accomodate certain forms of heteroskedasticity for the η_j. For instance, suppose that $X_{ij} = (X_j^{1\prime}, X_{ij}^{2\prime})'$, where X_j^1 is a vector of covariates that is constant within clusters and X_{ij}^2 is a vector of covariates that varies within clusters. Our approach is robust to forms of heteroskedasticity where the variance of the group-level shock is a function of the X_j^1 and the variance individual specific errors depends on both X_{ij} and W_{ij} (i.e., where $\sigma^2_\varepsilon(X_{ij},W_{ij}) = \sigma^2_\eta(X_j^1) + \sigma^2_\omega(X_{ij},W_{ij})$).

[2] Although we do not explore such an extension here for simplicity, our framework could easily be extended to allow for multiway clustering as in Cameron, Gelbach, and Miller (2011).

[3] This should be contrasted with nonparametric kernel regressions considered by Bhattacharya (2005), which are typically smooth functionals of the data.

2008). Thus, in our setting, standard techniques may not be valid and hence should not be used.

B. Estimating the Conditional Variance

The main difficulty in estimating the conditional variance (1) of the matching estimator is that we do not have estimated residuals since we are not directly estimating the regression functions $\mu_0(x)$ and $\mu_1(x)$. We follow the approach of Abadie and Imbens (2006, 2008) who generate quasi-residuals by matching each individual to the most similar individual with the same treatment status. Specifically, they define $\widehat{\varepsilon}_{ij} = Y_{ij} - Y_{g(ij)h(ij)}$, where $(g(ij), h(ij)) = \arg\min_{g,h|W_{gh}=W_{ij}} \|X_{ij} - X_{gh}\|$. Note that this is a different matching from that used to compute $\widehat{\tau}_M$ above. Abadie and Imbens then show that these quasi-residuals can be used to compute a heteroskedasticity-robust variance estimator in the absence of clustering.

The problem is more difficult in our setting because the presence of clustering means that we also need a consistent estimate of σ^2_η. Furthermore, following the clustering literature, we seek an estimator that requires only $J \to \infty$ for consistency. Therefore, we consider matching across clusters. Specifically, let $l(ij)$ and $k(ij)$ index the individual and cluster, respectively, of the closest match to ij with the same treatment status in a different cluster: $(l(ij), k(ij)) = \arg\min_{l,k\neq j|W_{lk}=W_{ij}} \|X_{ij} - X_{lk}\|$. Define the quasi-residuals

$$\widehat{\varepsilon}_{ij} = Y_{ij} - Y_{l(ij)k(ij)} = \underbrace{\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)})}_{\text{Matching discrepancy}} + \omega_{ij} - \omega_{l(ij)k(ij)} + \eta_j - \eta_{k(ij)}. \tag{2}$$

The matching discrepancy, $\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}X_{l(ij)k(ij)}$, will vanish as the number of potential matches grows. Thus, we can ignore these terms as $J \to \infty$.[4]

Expanding a within-cluster cross-product of these quasi-residuals, $\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j}$, and ignoring the matching discrepancy terms yields

$$\begin{aligned}\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j} = \eta_j^2 + \eta_j(&\omega_{ij} - \omega_{l(ij)k(ij)} \\ &- \eta_{k(ij)} + \omega_{i'j} - \omega_{l(i'j)k(i'j)} - \eta_{k(i'j)}) \\ &+ (\omega_{ij} - \omega_{l(ij)k(ij)} - \eta_{k(ij)})(\omega_{i'j} - \omega_{l(i'j)k(i'j)} - \eta_{k(i'j)}).\end{aligned}$$

Since we match across clusters and the ηs are independent of the ωs, it follows that

$$\begin{aligned}E[\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j}|X_{ij}, X_{i'j}] = \sigma^2_\eta &+ E[\eta_{k(ij)}\eta_{k(i'j)}] \\ &+ E[\omega_{l(ij)k(ij)}\omega_{l(i'j)k(i'j)}] + E[\omega_{ij}\omega_{i'j}].\end{aligned}$$

If $i = i'$, we have $E[\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j}|X_{ij}] = 2(\sigma^2_\eta + \sigma^2_\omega(X_{ij}, W_{ij})) = 2\sigma^2_\varepsilon(X_{ij}, W_{ij})$. For $i \neq i'$, $E[\eta_{k(ij)}\eta_{k(i'j)}|X_{ij}, X_{i'j}] = 0$ if i and i' are matched to units in distinct clusters (i.e., $k(ij) \neq k(i'j)$); by contrast, $E[\eta_{k(ij)}\eta_{k(i'j)}|X_{ij}, X_{i'j}] = \sigma^2_\eta$ if $k(ij) = k(i'j)$. Similarly, $E[\omega_{l(ij)k(ij)}\omega_{l(i'j)k(i'j)}] = 0$ unless i and i' are matched to the exact same unit, in which case $E[\omega_{l(ij)k(ij)}\omega_{l(i'j)k(i'j)}|X_{ij}, X_{i'j}] = \sigma^2_\omega(X_{l(ij)k(ij)}, W_{l(ij)k(ij)})$. Thus, in the absence of these duplicative matchings, the cross-product $\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j}$ for $i \neq i'$ is an unbiased estimator for σ^2_η.

[4] Note that if we were to allow matches within clusters (i.e., if $k(ij) = j$), the η terms would drop out, leaving $\widehat{\varepsilon}_{ij} = \mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)}) + \omega_{ij} - \omega_{l(ij)k(ij)}$. Although such matches would generally vanish and could be ignored as $J \to \infty$, they would impart a downward bias on the variance estimator in small samples. As a result, we would want to keep track of and correct for the occurence of within-cluster matching, which would unnecessarily complicate our methodology.

We can correct for duplicative matchings by defining

$$\widehat{\sigma}^2(X_{ij}, X_{i'j}) = \begin{cases} \widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j} & \text{if } k(ij) \neq k(i'j) \\ \widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j}/2 & \text{if } k(ij) = k(i'j) \text{ and if } l(ij) \neq l(i'j) \\ \widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j} - (\widehat{\varepsilon}_{i'j})^2/2 & \text{if } k(ij) = k(i'j) \text{ and if } l(ij) = l(i'j) \end{cases}. \tag{3}$$

If we ignore the matching discrepancy (i.e., assume that both ij and $i'j$ are perfectly matched), it follows that

$$E[\widehat{\sigma}^2(X_{ij}, X_{i'j})|X_{ij}, X_{i'j}] = \begin{cases} \sigma^2_\varepsilon(X_{ij}, W_{ij}) & \text{if } i = i' \\ \sigma^2_\eta & \text{if } i \neq i' \end{cases},$$

so we sometimes write $\widehat{\sigma}^2(X_{ij}, X_{ij}) = (\widehat{\varepsilon}_{ij})^2/2 = \widehat{\sigma}^2_\varepsilon(X_{ij}, W_{ij})$.

Let $P = \lim_{J\to\infty} J^{-1}\sum_j[(I_j(I_j-1))^{-1}\sum_i\sum_{i'\neq i} 1\{k(ij) = k(i'j)\}]$ be the probability of duplicative matchings as $J \to \infty$. It is worth noting that under many sampling arrangements, $P = 0$. In such circumstances, correcting for duplicative matchings is unnecessary asymptotically. However, the probability of duplicative matchings need not vanish as $J \to \infty$.[5] By defining $\widehat{\sigma}^2(X_{ij}, X_{i'j})$ as we have above, we ensure that our estimator is robust for any limiting probability of duplicative matchings.

Formally, we need the following assumptions:

Assumption 1 (Unconfoundedness)**.** *W is independent of $(Y(1), Y(0))$ conditional on X. (Overlap) $0 < \Pr(W = 1|X) < 1$.*

Assumption 2. *The X_{ij} are chosen from some bounded set $\mathbb{X} \subset \mathbb{R}^m$, and there is an upper bound $\bar{I}$ on cluster size.*

Assumption 3. *The conditional expectation and conditional variance functions are Lipschitz on $\mathbb{X}$: $|\mu_W(X) - \mu_W(X')| \leq C_{\mu,W}\|X - X'\|$ and $|\sigma^2_\omega(X, W) - \sigma^2_\omega(X', W)| \leq C_{\sigma,W}\|X - X'\|$ for $W \in \{0, 1\}$.*

Assumption 4. *$E[\eta^4]$ and $E[\omega^4]$ are bounded.*

Assumption 1 is needed to ensure that $\widehat{\tau}_M \xrightarrow{p} \tau$. As shown in the appendix, assumptions 2 and 3 ensure that the average matching discrepancy vanishes as the number of clusters increases. Assumption 4 ensures that the variances of η^2 and ω^2 are defined.

Proposition 1. *Suppose assumptions 1 through 5 hold, and let*

$$\hat{V}^E = N^{-1}\sum_j\left[\sum_i\sum_{i'}(2W_{ij}-1)(2W_{i'j}-1) \times\left(1 + \frac{K_M(ij)}{M}\right)\left(1 + \frac{K_M(i'j)}{M}\right)\widehat{\sigma}^2(X_{ij}, X_{i'j})\right]. \tag{4}$$

Then, holding fixed cluster sizes, as $J \to \infty$, we have $\hat{V}^E \xrightarrow{p} V^E$.

While both the clustering correction term in V^E in equation (1) and its estimate $\hat{V}^E$ in equation (4) are similar in form to the standard case (i.e., a weighted average of the cross-product of residuals), it is worth emphasizing that the construction of these residuals is quite distinct.

[5] For instance, suppose that there is a finite number of cluster types, each associated with a nondegenerate distribution of continuous covariates. If clusters are sampled i.i.d. from this set of types, then as $J \to \infty$, there will be many clusters of each type, and the probability of duplicative matchings will vanish (i.e., $P = 0$). Now suppose that there is a single continuous covariate that is constant within clusters. Ignoring ties, all units in a cluster j will be matched to units in a single cluster k (i.e., $P = 1$).

As a result, the consistency proof underlying proposition 1 is quite different from the proof in the standard case. However, we believe it is a strength of our approach that similar formulas are shown to apply in this nonstandard setting.

C. Estimating the Marginal Variance

In this section we show how to estimate $V^{\tau(X)} = N^{-1}J \times E[(\sum_i(\tau(X_{ij}) - \tau))^2]$. Obviously if we knew the $\tau(X_{ij})$, an estimate of $V^{\tau(X)}$ could be based on $N^{-1}\sum_j[\sum_i(\tau(X_{ij}) - \overline{\tau(\mathbf{X})})]^2$. For the matching estimator, the imputed outcome $\widehat{Y}_{ij}(1) - \widehat{Y}_{ij}(0)$ can serve as an estimator of $\tau(X_{ij})$. As $J \to \infty$, the matching discrepancy vanishes, and we have

$$
\begin{aligned}
&E\left[\left(\sum\nolimits_i(\widehat{Y}_{ij}(1) - \widehat{Y}_{ij}(0) - \tau)\right)^2\right] \\
&\quad\approx E\left[\left(\sum\nolimits_i(\tau(X_{ij}) - \tau)\right)^2\right] \\
&\qquad + E\left[\left(\sum\nolimits_i(2W_{ij} - 1)\varepsilon_{ij}\right)^2\right] \\
&\qquad - 2E\left[\left(\sum\nolimits_i(2W_{ij} - 1)\varepsilon_{ij}\right)\right. \\
&\qquad \left.\times\left(M^{-1}\sum\nolimits_{i'}\sum\nolimits_{st\in\mathcal{J}_M(i'j)}(2W_{i'j} - 1)\varepsilon_{st}\right)\right] \\
&\qquad + E\left[\left(M^{-1}\sum\nolimits_i\sum\nolimits_{st\in\mathcal{J}_M(ij)}(2W_{ij} - 1)\varepsilon_{st}\right)^2\right].
\end{aligned}
$$

In the presence of clustering, this expectation is somewhat more complicated than the case considered in Abadie and Imbens (2006). First and most important, the ε_{ij} within a given cluster will be correlated due to the shared component, η_j. In addition, the issue of duplicative matchings we saw above also arises in estimating $V^{\tau(X)}$.[6] Specifically, the ε_{ij} will be correlated with the ε_{st} if units in cluster j are used to impute missing outcomes for other units in cluster j (i.e., if any element of $\mathcal{J}_M(i'j)$ is in cluster j for some i'). Moreover, the ε_{st} will be correlated with each other if multiple units in some different cluster $j' \neq j$ are used to impute missing outcomes for units in cluster j.

If $P = 0$, the last two sources of correlation will vanish asymptotically, but the first will always be present. Below, we present a simple estimator of $V^{\tau(X)}$ that is consistent in this case. However, for certain empirical applications, it will be important to have an estimator that is valid even if the probability of duplicative matchings does not vanish as $J \to \infty$. This estimator, which contains additional terms to correct for duplicative matchings, is given by equation (1) in the appendix.

Proposition 2. *Suppose assumptions 1 through 4 hold and that $P = 0$, and let*

$$
\begin{aligned}
\hat{V}^{\tau(X)} &= N^{-1}\sum\nolimits_j\left[\sum\nolimits_i(\widehat{Y}_{ij}(1) - \widehat{Y}_{ij}(0) - \widehat{\tau}_M)\right]^2 \\
&\quad - N^{-1}\sum\nolimits_{i,j}\frac{K_M(ij)}{M^2}\widehat{\sigma}^2_\varepsilon(X_{ij}, W_{ij}) \\
&\quad - N^{-1}\sum\nolimits_j\left[\sum\nolimits_i\sum\nolimits_{i'}(2W_{ij} - 1)(2W_{i'j} - 1)\widehat{\sigma}^2(X_{ij}, X_{i'j})\right]. \qquad (5)
\end{aligned}
$$

[6] Here we are referring to the probability of duplicative matchings for the matching used in the matching estimator (indexed *st*), not the matching used to compute the quasi-residuals (indexed *lk*). However, the asymptotic probability of duplicative matchings is a function of the distribution of covariates, so the probability of duplicative matchings will generally be either zero or nonzero for both matchings.

TABLE 1.—COVERAGE PROBABILITIES OF 95% CI

		J		
		10	20	50
I	2	0.90	0.93	0.94
	10	0.91	0.92	0.95
	50	0.92	0.94	0.95

Then, holding fixed cluster sizes, as $J \to \infty$, $\hat{V}^{\tau(X)} \xrightarrow{p} V^{\tau(X)}$. An estimate of the marginal variance can then be computed as $\hat{V} = \hat{V}^{\tau(X)} + \hat{V}^E$.

Combining the results in propositions 1 and 2, we have

$$
\begin{aligned}
\hat{V} &= N^{-1}\sum\nolimits_{i,j}(\widehat{Y}_{ij}(1) - \widehat{Y}_{ij}(0) - \widehat{\tau}_M)^2 \\
&\quad + N^{-1}\sum\nolimits_{i,j}\left(\left(\frac{K_M(ij)}{M}\right)^2 + \frac{2M-1}{M}\frac{K_M(ij)}{M}\right) \\
&\quad \times\widehat{\sigma}^2_\varepsilon(X_{ij}, W_{ij}) \\
&\quad + N^{-1}\sum\nolimits_{i,j}\sum\nolimits_{i'\neq i}\left[(\widehat{Y}_{ij}(1) - \widehat{Y}_{ij}(0) - \widehat{\tau}_M)(\widehat{Y}_{i'j}(1)\right. \\
&\quad \left. - \widehat{Y}_{i'j}(0) - \widehat{\tau}_M)\right] \\
&\quad + N^{-1}\sum\nolimits_{i,j}\sum\nolimits_{i'\neq i}(2W_{ij}-1)(2W_{i'j}-1) \\
&\quad \times\left(\frac{K_M(ij) + K_M(i'j)}{M} + \frac{K_M(ij)}{M}\frac{K_M(i'j)}{M}\right) \\
&\quad \times\widehat{\sigma}^2(X_{ij}, X_{i'j}). \qquad (6)
\end{aligned}
$$

The first two terms are the estimator of $V^E + V^{\tau(X)}$ given in Abadie and Imbens (2006), which is valid in the absence of clustering. The second pair of terms is the combined clustering correction, which is valid in the case where $P = 0$.

IV. A Short Monte Carlo Study

To get a sense of the finite sample properties of our variance estimator, we examine the confidence intervals it generates. We assume there is a single covariate, $X_{ij} \overset{iid}{\sim} N(0, 5)$, and that $\mu_0(x) = 0$ and $\mu_1(x) = x$. It follows that $\tau(x) = x$, and the observed outcome is $Y_{ij} = W_{ij}X_{ij} + \varepsilon_{ij}$. We assume that half the clusters are treated and that $\sigma^2_\eta = 1.5$ and $\sigma^2_\omega = 1.5$. For each replication, we draw a new set of covariates X_{ij}, as well as error components η_j and ω_{ij}. We then estimate the PATE using the simple matching estimator, $\widehat{\tau}_1$ with $M = 1$ matches to impute unobserved outcomes. Clustering-corrected 95% confidence intervals for the estimator are computed as $(\widehat{\tau}_1 - 1.96(\hat{V}/N)^{1/2}, \widehat{\tau}_1 + 1.96(\hat{V}/N)^{1/2})$, where $\hat{V} = \hat{V}^{\tau(X)} + \hat{V}^E$, $\hat{V}^E$ is given by equation (4), and $\hat{V}^{\tau(X)}$ is given by equation (1) in the appendix. We carry out $R = 1,000$ replications and report the fraction of replications where $E[\tau(x)] = 0$ is covered by the resulting confidence intervals.

The results are shown in table 1. Our marginal variance estimator has very good finite sample performance. Although it shows slight undercoverage in very small samples, its coverage reaches 95% quickly as J increases. Furthermore, while the coverage of an estimator that did not correct for clustering would decline to approximately 0.70 for $I = 10$ and 0.40 for $I = 50$, the performance of our estimator is nearly constant as a function of I. Similar coverage probabilities obtain if we allow the X_{ij} to be correlated within clusters.

V. Other Nonparametric Treatment Effect Estimators

First, as discussed in the appendix, it is straightforward to use our approach to compute cluster-robust variances for matching estimators of the average treatment effect for the treated (ATT). Second, while this note has focused on the matching estimators, the methodology we have described can be extended to a broader class of nonparametric treatment effect estimators. A number of estimators for average treatment effects, including propensity score–based estimators, can be written as

$$\begin{aligned}\widehat{\tau} &= \sum\nolimits_{W_{ij}=1} \gamma_{ij}(\mathbf{X},\mathbf{W})Y_{ij} - \sum\nolimits_{W_{ij}=0} \gamma_{ij}(\mathbf{X},\mathbf{W})Y_{ij} \\ &= \sum\nolimits_{i,j} (2W_{ij}-1)\gamma_{ij}(\mathbf{X},\mathbf{W})Y_{ij}, \end{aligned} \tag{7}$$

where $\gamma_{ij}(\mathbf{X},\mathbf{W})$ is a set of data-dependent weights. For estimators of the form (7), the conditional variance can be estimated using equation (4) and replacing $N^{-1}(1 + K_M(ij)/M)$ with γ_{ij}. Similarly, for estimators that impute treatment effects $\widehat{Y}_{ij}(1) - \widehat{Y}_{ij}(0)$ for each unit, we can estimate $V^{\tau(X)}$ using an expression analogous to equation (5).

VI. Conclusion

In this note, we develop a methodology to estimate the conditional and marginal variances of the matching estimator in the presence of clustering. Our cluster-robust variance estimators are consistent as the number of clusters grows large, holding the size of clusters fixed. Furthermore, our methodology can easily be extended to a broader class of nonparametric treatment effect estimators.

REFERENCES

Abadie, Alberto, and Guido Imbens, "Large Sample Properties of Matching Estimators," *Econometrica* 74 (2006), 235–276.

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

Bhattacharya, Debopam, "Asymptotic Inference from Multi-Stage Samples," *Journal of Econometrics* 126 (2005), 145–171.

Cameron, A. Colin, Jonah Gelbach, and Douglas Miller, "Robust Inference with Multiway Clustering," *Journal of Business and Economic Statistics* 29 (2011), 238–249.

Fryer, Roland, "Financial Incentives and Student Achievement: Evidence from Randomized Trials," unpublished paper, Harvard University (2010).

Greenwald, Bruce, "A General Analysis of Bias in the Estimated Standard Errors of Least Squares Coefficients," *Journal of Econometrics* 22 (1983), 323–338.

Hotz, V. Joseph, Guido Imbens, and Jacob Klerman, "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Re-Analysis of the California GAIN Program," *Journal of Labor Economics* 24 (2006), 521–566.

Imbens, Guido, "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," this REVIEW 86 (2004), 4–29.

Moulton, Brent, "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics* 32 (1986), 385–397.

——— "An Illustration of a Pitfalls in Estimating the Effects of Aggregate Variables on Micro Units," this REVIEW 72 (1990), 334–338.

Wooldridge, Jeffrey, *Econometric Analysis of Cross Section and Panel Data* (Cambridge, MA: MIT Press, 2002).

——— "Cluster-Sample Methods in Applied Econometrics: An Extended Analysis," unpublished paper, Michigan State University (2006).