Matching As An Econometric Evaluation Estimator*

James J. Heckman† Hidehiko Ichimura‡ Petra Todd§

September 16, 1993

Revised July, 1997

Abstract

This paper develops the method of matching as an econometric evaluation estimator. A rigorous distribution theory for kernel-based matching is presented. The method of matching is extended to more general conditions than the ones assumed in the statistical literature on the topic. We focus on the method of propensity score matching and show that it is not necessarily better, in

*The work reported here is a distant outgrowth of numerous conversations with Ricardo Barros, Bo Honoré, and Richard Robb. We thank Manuel Arellano and three referees for helpful comments. An earlier version of this paper "Matching As An Evaluation Estimator: Theory and Evidence on Its Performance Applied to the JTPA Program, Part I. Theory and Methods" was presented at the Review of Economic Studies conference on evaluation research in Madrid in September, 1993. This paper was also presented by Heckman in his Harris Lectures at Harvard, November, 1995, at the Latin American Econometric Society meeting in Caracas, Venezuela, August, 1994; at the Rand Corporation, U.C. Irvine, U.S.C., U.C. Riverside, and U.C. San Diego September, 1994; at Princeton, October, 1994; at UCL London, November, 1994 and November 1996.and at Texas, Austin, March, 1996 and at the Econometric Society meetings in San Francisco, January, 1996.

†James Heckman is Henry Schultz Distinguished Service Professor in the Department of Economics, Director of the Center for Social Program Evaluation at the Irving B. Harris School of Public Policy, the University of Chicago and Senior Research Fellow, American Bar Foundation. Part of Heckman's contribution to the paper was presented at the Barcelona Lecture, Sixth World Congress of the Econometric Society Meetings, Barcelona, Spain, August 25, 1990. This was supported by NSF-SES-91-11455, NSF-SES-93-234118 and a grant from the Russell Sage Foundation.

‡Hidehiko Ichimura is Assistant Professor in the Department of Economics at the University of Pittsburgh and Research Associate of the Center For Social Program Evaluation, Irving B. Harris School of Public Policy, University of Chicago. His work is supported by NSF-93-234118.

§Petra Todd is Assistant Professor in the Department of Economics at the University of Pennsylvania and Research Associate at the Center for Social Program Evaluation, Irving B. Harris School of Public Policy, the University of Chicago.

the sense of reducing the variance of the resulting estimator, to use the propensity score method even if propensity score is known. We extend the statistical literature on the propensity score by considering the case when it is estimated both parametrically and nonparametrically. We examine the benefits of separability and exclusion restrictions in improving the efficiency of the estimator. Our methods also apply to the econometric selection bias estimator.

JEL Number: C10

1 Introduction

Matching is a widely-used method of evaluation. It is based on the intuitively attractive idea of contrasting the outcomes of program participants (denoted Y_1) with the outcomes of "comparable" nonparticipants (denoted Y_0). Differences in the outcomes between the two groups are attributed to the program.

Let $\mathbf{I_0}$ and $\mathbf{I_1}$ denote the set of indices for nonparticipants and participants, respectively. The following framework describes conventional matching methods as well as the smoothed versions of these methods analyzed in this paper. To estimate a treatment effect for each treated person $i \in \mathbf{I_1}$, outcome Y_{1i} is compared to an average of the outcomes Y_{0j} for matched persons $j \in \mathbf{I_0}$ in the untreated sample. Matches are constructed on the basis of observed characteristics X in R^d. Typically, when the observed characteristics of an untreated person are closer to those of the treated person $i \in \mathbf{I_1}$, using a specific distance measure, the untreated person gets a higher weight in constructing the match. The estimated gain for each person i in the treated sample is

$$Y_{1i} - \sum_{j \in \mathbf{I_0}} W_{N_0,N_1}(i,j) Y_{0j}, \tag{1}$$

where $W_{N_0,N_1}(i,j)$ is usually a positive valued weight function, defined so that for each $i \in \mathbf{I_1}$, $\sum_{j \in \mathbf{I_0}} W_{N_0,N_1}(i,j) = 1$, and N_0 and N_1 are the number of individuals in $\mathbf{I_0}$ and $\mathbf{I_1}$, respectively. The choice of a weighting function reflects the choice of a particular distance measure used in the matching method, and the weights are based on distances in the "X" space. For example, for each $i \in \mathbf{I_1}$ the

nearest-neighbor method selects one individual $j \in \mathbf{I_0}$ as the match whose X_j is the "closest" value to X_i, in some metric. The kernel methods developed in this paper construct matches using all individuals in the comparison sample and downweighting "distant" observations.

The widely-used evaluation parameter on which we focus in this paper is the mean effect of treatment on the treated for persons with characteristics X:

$$\mathrm{E}(Y_1 - Y_0 \mid D = 1, X), \tag{P-1}$$

where $D = 1$ denotes program participation. Heckman (1997) and Heckman and Smith (1997) discuss conditions under which this parameter answers economically interesting questions. For a particular domain $\mathcal{X}$ for X, this parameter is estimated by

$$\sum_{i \in \mathbf{I_1}} w_{N_0,N_1}(i)[Y_{1i} - \sum_{j \in \mathbf{I_0}} W_{N_0,N_1}(i,j) Y_{0j}] \tag{2}$$

where different values of $w_{N_0,N_1}(i)$ may be used to select different domains $\mathcal{X}$ or to account for heteroskedasticity in the treated sample. Different matching methods are based on different weighting functions $\{w_{N_0,N_1}(i)\}$ and $\{W_{N_0,N_1}(i,j)\}$.

The method of matching is intuitively appealing and is often used by applied statisticians, but not by economists. This is so for four reasons. First, it is difficult to determine if a particular comparison group is truly comparable to participants (*i.e.* would have experienced the same outcomes as participants had they participated in the program). An ideal social experiment creates a valid comparison group. But matching on the measured characteristics available in a typical nonexperimental study is not guaranteed to produce such a comparison group. The published literature presents conditional

independence assumptions under which the matched group is comparable, but these are far stronger than the mean-independence conditions typically invoked by economists. Moreover, the assumptions are inconsistent with many economic models of program participation in which agents select into the program on the basis of unmeasured components of outcomes unobserved by the econometrician. Even if conditional independence is achieved for one set of X variables, it is not guaranteed to be achieved for other sets of X variables including those that include the original variables as subsets. Second, if a valid comparison group can be found, the distribution theory for the matching estimator remains to be established for continuously distributed match variables X.[1]

Third, most of the current econometric literature is based on separability between observables and unobservables and on exclusion restrictions that isolate different variables that determine outcomes and program participation. Separability permits the definition of parameters that do not depend on unobservables. Exclusion restrictions arise naturally in economic models, especially in dynamic models where the date of enrollment into the program differs from the dates when consequences of the program are measured. The available literature on matching in statistics does not present a framework that incorporates either type of a priori restriction.

Fourth, matching is a data-hungry method. With a large number of conditioning variables, it is easy to have many cells without matches. This makes the method impractical or dependent on the use of arbitrary sorting schemes to select hierarchies of matching variables. (See, *e.g.* Westat, 1980, 1982, and 1984.) In an important paper, Rosenbaum and Rubin (1983) partially solve this problem. They establish that if matching on X is valid, so is matching solely on the probability of selection into the program $\Pr(D = 1|X) = P(X)$. Thus a multidimensional matching problem can be recast

[1]When the match variables are discrete, the matching estimator for each cell is a mean and consistency and asymptotic normality of the matching estimator are easily established.

as a one-dimensional problem and a practical solution of the curse of dimensionality for matching is possible.[2]

Several limitations hamper the practical application of their theoretical result. Their theorem assumes that the probability of selection is known and is not estimated. It is also based on strong conditional independence assumptions that are difficult to verify in any application and are unconventional in econometrics. They produce no distribution theory for their estimator.

In this paper we first develop an econometric framework for matching that allows us to incorporate additive separability and exclusion restrictions. We then provide a sampling theory for matching from a nonparametric vantage point. Our distribution theory is derived under weaker conditions than the ones currently maintained in the statistical literature on matching. We show that the fundamental identification condition of the matching method for estimating (P-1) is

$$\mathrm{E}(Y_0 \mid D = 1, X) = \mathrm{E}(Y_0 \mid D = 0, X)$$

whenever both sides of this expression are well defined. In order for both sides to be well defined simultaneously for all X it is usually assumed that $0 < P(X) < 1$ so that $\mathrm{Supp}(X \mid D = 1) = \mathrm{Supp}(X \mid D = 0)$. As Heckman, Ichimura, Smith, Todd (1994, revised 1996a), Heckman, Ichimura, Smith and Todd (1996c) and Heckman, Ichimura and Todd (1997) point out, this condition is not appropriate for important applications of the method. In order to meaningfully implement matching it is necessary to condition on the support common to both participant and comparison groups S,

[2]They term $P(X)$ the propensity score. For the relationship between propensity score methods and selection models, see Heckman and Robb (1986) or Heckman, Ichimura, Smith and Todd (1994, revised 1996a).

where

$$S = \text{Supp}(X \mid D = 1) \cap \text{Supp}(X \mid D = 0)$$

and to estimate the region of common support. Equality of the supports need not hold a priori although most formal discussions of matching assumes that it does. Heckman, Ichimura, Smith and Todd (1996) and Heckman, Ichimura and Todd (1997) report the empirical relevance of this point for evaluating job training programs. Invoking assumptions that justify the application of nonparametric kernel regression methods to estimate program outcome equations, maintaining weaker mean independence assumptions compared to the conditional independence assumptions used in the literature, and conditioning on S, we produce an asymptotic distribution theory for matching estimators when regressors are either continuous, discrete or both. This theory is general enough to make the Rosenbaum-Rubin theorem operational in the commonly-encountered case where $P(X)$ is estimated either parametrically or nonparametrically.

With a rigorous distribution theory in hand, we address a variety of important questions that arise in applying the method of matching: (1) We ask, if one knew the propensity score, $P(X)$, would one want to use it instead of matching on X? (2) What are the effects on asymptotic bias and variance if we use an estimated value of P? We address this question both for the case of parametric and nonparametric $P(X)$. Finally, we ask (3) what are the benefits, if any, of econometric separability and exclusion restrictions on the bias and variance of matching estimators?

The structure of this paper is as follows. Section 1 states the evaluation problem and the parameters identified by the analysis of this paper. Section 2 discusses how matching solves the evaluation problem. We discuss the propensity score methodology of Rosenbaum and Rubin (1983). We emphasize the importance of common support condition assumed in the literature and develop an approach

that does not require it. Section 3 contrasts the assumptions used in matching with the separability assumptions and exclusion restrictions conventionally used in econometrics. A major goal of this paper is to unify the matching literature with the econometrics literature. Section 4 investigates a central issue in the use of propensity scores. Even if the propensity score is known, is it better, in terms of reducing the variance of the resulting matching estimator, *i.e.* to condition on X or $P(X)$? There is no unambiguous answer to this question. Section 5 presents a basic theorem that provides the distribution theory for kernel matching estimators based on estimated propensity scores. In Section 6, these results are then applied to investigate the three stated questions. Section 7 summarizes the paper.

2 The Evaluation Problem and The Parameters of Interest

Each person can be in one of two possible states, 0 and 1, with associated outcomes (Y_0, Y_1), corresponding to receiving no treatment or treatment respectively. For example, "treatment" may represent participation in the social program, such as the job training program evaluated in our companion paper where we apply the methods developed in this paper. (Heckman, Ichimura and Todd, 1997). Let $D = 1$ if a person is treated; $D = 0$ otherwise. The gain from treatment is $\Delta = Y_1 - Y_0$. We do not know Δ for anyone because we observe only $Y = DY_1 + (1 - D)Y_0$, *i.e.* either Y_0 or Y_1.

This fundamental missing data problem cannot be solved at the level of any individual. Therefore, the evaluation problem is typically reformulated at the population level. Focusing on mean impacts for persons with characteristics X, a commonly-used parameter of interest for evaluating the mean impact of participation in social programs is (P-1). It is the average gross gain from participation in the program for participants with characteristics X. If the full social cost per participant is subtracted

from (P-1), and the no treatment outcome for all persons closely approximates the no program outcome then the net gain informs us of whether the program raises total social output compared to the no program state for the participants with characteristics X.[3]

The mean $\mathrm{E}(Y_1 \mid D = 1, X)$ can be identified from data on program participants. Assumptions must be invoked to identify the counterfactual mean $\mathrm{E}(Y_0 \mid D = 1, X)$, the no-treatment outcome of program participants. In the absence of data from an ideal social experiment, the outcome of self-selected nonparticipants $\mathrm{E}(Y_0 \mid D = 0, X)$ is often used to approximate $\mathrm{E}(Y_0 \mid D = 1, X)$. The selection bias that arises from making this approximation is

$$B(X) = \mathrm{E}(Y_0 \mid D = 1, X) - \mathrm{E}(Y_0 \mid D = 0, X).$$

Matching on X, or regression adjustment of Y_0 using X, is based on the assumption that $B(X) = 0$ so conditioning on X eliminates the bias.

Economists have exploited the idea of conditioning on observables using parametric or nonparametric regression analysis. (Barnow, Cain and Goldberger, 1980; Barros, 1986; Heckman and Robb, 1985, 1986.) Statisticians more often use matching methods, pairing treated persons with untreated persons of the same X characteristics (Cochrane and Rubin, 1973).

The literature on program evaluation gives two distinct responses to the problem of estimating (P-1) with continuous conditioning variables. The first borrows from the kernel regression literature. It uses a smoothing procedure that borrows strength from adjacent values of a particular value of $X = x$ and produces uniformly consistent estimators of (P-1) at all points of the support for the distributions

[3]See Heckman (1997) or Heckman and Smith (1997) for a precise statement of when this parameter answers an interesting economic evaluation question.

of X given $D = 1$ or $D = 0$. (See Heckman, Ichimura, Smith and Todd, 1994 revised 1996b or Heckman, Ichimura and Todd,1993, forthcoming 1997). Parametric assumptions about $\mathrm{E}(Y_0 \mid D = 1, X)$ play the same role as smoothing assumptions, and in addition allow analysts to extrapolate out of the sample for X. Unless the class of functions to which (P-1) may belong is restricted to be smaller than the finite-order continuously-differentiable class of functions, the convergence rate of an estimator of (P-1) is governed by the number of continuous variables included in X (Stone, 1982).

The second response to the problem of constructing counterfactuals abandons estimation of (P-1) at any point of X and instead estimates an average of (P-1) over an interval of X values. Commonly-used intervals include $\mathrm{Supp}(X|D = 1)$, or subintervals of the support corresponding to different groups of interest. The advantage of this approach is that the averaged parameter can be estimated with rate $N^{-1/2}$, where N is sample size, regardless of the number of continuous variables in X when the underlying functions are smooth enough. Averaging the estimators over intervals of X produces a consistent estimator of

$$M(S) = \mathrm{E}(Y_1 - Y_0 \mid D = 1, X \in S), \tag{P-2}$$

with a well-defined $N^{-1/2}$ distribution theory where S is a subset of $\mathrm{Supp}(X|D = 1)$. There is considerable interest in estimating impacts for groups so (P-2) is the parameter of interest in conducting an evaluation. In practice both pointwise and setwise parameters may be of interest. Historically, economists have focused on estimating (P-1) and statisticians have focused on estimating (P-2), usually defined over broad intervals of X values, including $\mathrm{Supp}(X|D = 1)$. In this paper, we invoke conditions sufficiently strong to consistently estimate both (P-1) and (P-2).

3 How Matching Solves the Evaluation Problem

Using the notation of Dawid (1979) let

$$(Y_0, Y_1) \perp\!\!\!\perp D|X \tag{A-1}$$

denote the statistical independence of (Y_0, Y_1) and D conditional on X. An equivalent formulation of this condition is

$$\Pr(D = 1 \mid Y_0, Y_1, X) = \Pr(D = 1 \mid X).$$

This is a non-causality condition that excludes the dependence between potential outcomes and participation that is central to econometric models of self selection. (See Heckman and Honoré, 1990.) Rosenbaum and Rubin (1983), henceforth denoted RR, establish that, when (A-1) and

$$0 < P(X) < 1 \tag{A-2}$$

are satisfied, $(Y_0, Y_1) \perp\!\!\!\perp D|P(X)$, where $P(X) = \Pr(D = 1|X)$. Conditioning on $P(X)$ balances the distribution of Y_0 and Y_1 with respect to D. The requirement (A-2) guarantees that matches can be made for all values of X. RR called condition (A-1) an "ignorability" condition for D, and they call (A-1) and (A-2) together a "strong ignorability" condition.

When the strong ignorability condition holds, one can generate marginal distributions of the counterfactuals:

$$F_0(y_0 \mid D = 1, X) \qquad \text{and} \qquad F_1(y_1 \mid D = 0, X)$$

but one cannot estimate the joint distribution of (Y_0, Y_1), $F(y_0, y_1 \mid D, X)$, without making further

assumptions about the structure of outcome and program participation equations.[4]

If $P(X) = 0$ or $P(X) = 1$ for some values of X, then one cannot use matching conditional on those X values to estimate a treatment effect. Persons with such X characteristics either always receive treatment or never receive treatment, so matches from both $D = 1$ and $D = 0$ distributions cannot be performed. Ironically, missing data give rise to the problem of causal inference, but missing data, *i.e.* the unobservables producing variation in D conditional on X, are also required to solve the problem of causal inference. The model predicting program participation should not be too good so that $P(X) = 1$ or 0 for any X. Randomness, as embodied in condition (A-2), guarantees that persons with the same characteristics can be observed in both states. This condition says that for any measurable set A, $\Pr(X \in A|D = 1) > 0$ if and only if $\Pr(X \in A|D = 0) > 0$, so the comparison of conditional means is well defined.[5] A major finding in Heckman, Ichimura, Smith, Todd (1994, revised 1996a,b,c) is that in their sample these conditions are not satisfied, so matching is only justified over the subset $\text{Supp}(X|D = 1) \cap \text{Supp}(X|D = 0)$.

Note that under assumption (A-1)

$$\begin{aligned} \mathrm{E}(Y_0|D = 1, X \in S) &= \mathrm{E}[\mathrm{E}(Y_0|D = 1, X)|D = 1, X \in S] \\ &= \mathrm{E}[\mathrm{E}(Y_0|D = 0, X)|D = 1, X \in S] \end{aligned}$$

so $\mathrm{E}(Y_0 \mid D = 1, X \in S)$ can be recovered from $\mathrm{E}(Y_0 \mid D = 0, X)$ by integrating over X using the distribution of X given $D = 1$, restricted to S. Note that, in principle, both $\mathrm{E}(Y_0 \mid X, D = 0)$

[4]Heckman, Smith and Clements (1993; revised, 1997) and Heckman and Smith (1997) analyze a variety of such assumptions.

[5]Thus, the implication of $0 < \Pr(D = 1|X) < 1$ is that conditional measures of X given $D = 0$ and that given $D = 1$ are absolutely continuous with respect to each other. These dominating measure conditions are standard in the matching literature.

and the distribution of X given $D = 1$ can be recovered from random samples of participants and nonparticipants.

It is important to recognize that unless the expectations are taken on the common support of S, the second equality does not necessarily follow. While $\mathrm{E}(Y_0 \mid D = 0, X)$ is always measurable with respect to the distribution of X given $D = 0$, $(\mu(X|D = 0))$, it may not be measurable with respect to the distribution of X given $D = 1$, $(\mu(X|D = 1))$. Invoking assumption (A-2) or conditioning on the common support S solves the problem because $\mu(X|D = 0)$ and $\mu(X|D = 1)$, restricted to S, are mutually absolutely continuous with respect to each other. In general, assumption (A-2) may not be appropriate in many empirical applications. (See Heckman, Ichimura, and Todd, 1997 or Heckman, Ichimura, Smith and Todd, 1996a,b,c.)

The sample counterpart to the population requirement that estimation should be over a common support arises when the set S is not known. In this case, we need to estimate S. Since the estimated set, $\hat{S}$, and S inevitably differ, we need to make sure that asymptotically the points at which we evaluate the conditional mean estimator of $\mathrm{E}(Y_0 \mid D = 0, X)$ are in S. We use the "trimming" method developed in our companion paper (Heckman, Ichimura, Smith, Todd, 1996b) to deal with the problem of determining the points in S. Instead of imposing (A-2), we investigate regions S, where we can reasonably expect to learn about $\mathrm{E}(Y_1 - Y_0|D = 1, S)$.

Conditions (A-1) and (A-2) which are commonly invoked to justify matching, are stronger than what is required to recover $\mathrm{E}(Y_1 - Y_0 \mid D = 1, X)$ which is the parameter of interest in this paper. We can get by with the weaker condition since our objective is construction of the counterfactual

$\mathrm{E}(Y_0|X, D = 1)$

$$Y_0 \perp\!\!\!\perp D \mid X, \tag{A-3}$$

which implies that $\Pr(Y_0 < t|D = 1, X) = \Pr(Y_0 < t|D = 0, X)$ for $X \in S$.

In this case, the distribution of Y_0 given X for participants can be identified using data only on nonparticipants provided that $X \in S$. From these distributions, one can recover the required counterfactual mean $\mathrm{E}(Y_0 \mid D = 1, X)$ for $X \in S$. Note that condition (A-3) does *not* rule out the dependence of D on Y_1 or on $\Delta = Y_1 - Y_0$ given X.[6]

For identification of the mean treatment impact parameter (P-1), an even weaker mean independence condition suffices:

$$\mathrm{E}(Y_0|D = 1, X) = \mathrm{E}(Y_0|D = 0, X) \qquad \text{for } X \in S. \tag{A-1$'$}$$

Under this assumption, we can identify $\mathrm{E}(Y_0|D = 1, X)$ for $X \in S$, the region of common support.[7] Mean independence conditions are routinely invoked in the econometrics literature.[8]

Under conditions (A-1) and (A-2), conceptually different parameters such as the mean effect of treatment on the treated, the mean effect of treatment on the untreated, or the mean effect of randomly assigning persons to treatment, all conditional on X, are the same. Under assumptions (A-3) or (A-1$'$), they are distinct.[9]

[6] By symmetric reasoning, if we postulate the condition $Y_1 \perp\!\!\!\perp D|X$ and (A-2), then $\Pr(D = 1|Y_1, X) = \Pr(D = 1|X)$, so selection could occur on Y_0 or Δ, and we can recover $\Pr(Y_1 < t|D = 0, X)$. Since $\Pr(Y_0 < t|D = 0, X)$ can be consistently estimated, we can recover $E(Y_1 - Y_0|D = 0, X)$.

[7] We can further identify $E(Y_1 - Y_0|D = 0)$ if we assume $E(Y_1|D = 1, X) = E(Y_1|D = 0, X)$ for X in S.

[8] See, for example, Barnow, Cain and Goldberger (1980) or Heckman and Robb (1985, 1986).

[9] See Heckman (1990) for a discussion of the three parameters. See also Heckman and Smith (1997).

Under these weaker conditions, we demonstrate below that it is not necessary to make assumptions about specific functional forms of outcome equations or distributions of unobservables that have made the empirical selection bias literature so controversial. What is controversial about these conditions is the assumption that the conditioning variables available to the analyst are sufficiently rich to justify application of matching. To justify the assumption, analysts implicitly make conjectures about what information goes into the decision sets of agents, and how unobserved (by the econometrician) relevant information is related to observables. (A-1) rules out dependence of D on Y_0 and Y_1 and so is inconsistent with the Roy model of self selection. See Heckman (1997) or Heckman and Smith (1997) for further discussion.

4 Separability and Exclusion Restrictions

In many applications in economics, it is instructive to partition X into two not-necessarily mutually exclusive sets of variables, (T, Z), where the T variables determine outcomes:

$$Y_0 = g_0(T) + U_0 \tag{3a}$$

$$Y_1 = g_1(T) + U_1 \tag{3b}$$

and the Z variables determine program participation

$$\Pr(D = 1 \mid X) = \Pr(D = 1 \mid Z) = P(Z). \tag{4}$$

Thus in a panel data setting Y_1 and Y_0 may be outcomes measured in periods after program participation decisions are made, so that Z and T may contain distinct variables although they may have

some variables in common. Different variables may determine participation and outcomes, as in the labor supply and wage model of Heckman (1974).

Additively-separable models are widely used in econometric research. A major advantage of such models is that any bias arising from observing Y_0 or Y_1 by conditioning on D is confined to the "error term" provided that one also conditions on T, *e.g.* $\mathrm{E}(Y_0|D=1,X) = g_0(T) + \mathrm{E}(U_0|D=1,Z)$ and $\mathrm{E}(Y_1|D=1,X) = g_0(T) + \mathrm{E}(U_0|D=1,Z)$. Another major advantage of such models is that they permit an operational definition of the effect of a change in T holding U constant. Such effects are derived from the g_0 and g_1 functions.

The Rosenbaum-Rubin Theorem (1983) does not inform us about how to exploit additive separability or exclusion restrictions. The evidence reported in Heckman, Ichimura, Smith and Todd (1994, revised 1996b), reveals that the no-training earnings of persons who chose to participate in a training program, Y_0, can be represented in the following way:

$$\mathrm{E}(Y_0 \mid D=1,X) = g_0(T) + \mathrm{E}(U_0 \mid P(Z)),$$

where Z and T contain some distinct regressors. This representation reduces the dimension of the matching or nonparametric regression problem if the dimension of Z is two or larger. Currently-available matching methods do not provide a way to exploit such information about the additive separability of the model or to exploit the information that Z and T do not share all of their elements in common.

This paper extends the insights of Rosenbaum and Rubin (1983) to the widely-used model of program participation and outcomes given by equations (3a) and (3b). Thus, instead of (A-1) or

(A-3), we consider the case where

$$U_0 \perp\!\!\!\perp D|X. \tag{A-4a}$$

Invoking the exclusion restrictions $P(X) = P(Z)$ and using an argument analogous to Rosenbaum and Rubin (1983), we obtain

$$\begin{aligned} \mathrm{E}\{D \mid U_0, P(Z)\} &= \mathrm{E}\{E(D \mid U_0, X) \mid U_0, P(Z)\} \\ &= \mathrm{E}\{P(Z) \mid U_0, P(Z)\} = P(Z) = \mathrm{E}\{D \mid P(Z)\} \end{aligned}$$

so that

$$U_0 \perp\!\!\!\perp D \mid P(Z). \tag{A-4b}$$

Under condition (A-4a) it is not necessarily true that (A-1) or (A-3) are valid but it is obviously true that

$$[Y_0 - g_0(T)] \perp\!\!\!\perp D|P(Z).$$

In order to identify the mean treatment effect on the treated, it is enough to assume that

$$\mathrm{E}(U_0 \mid D = 1, P(Z)) = \mathrm{E}(U_0 \mid D = 0, P(Z)) \tag{A-4b$'$}$$

instead of (A-4a) or (A-4b).

Observe that (A-4a), (A-4b), and (A-4b$'$) do not imply that $\mathrm{E}(U_0 \mid P(Z)) = 0$ or that $\mathrm{E}(U_1 \mid$

$P(Z)) = 0$. They only imply that the distributions of the unobservables are the same in populations of $D = 1$ and $D = 0$, once one conditions on $P(Z)$. Y_0 and Y_1 must be adjusted to eliminate the effects of T on outcomes. Only the residuals can be used to exploit the RR conditions. Thus $P(Z)$ is not, in general, a valid instrumental variable.

In order to place these results in the context of classical econometric selection models, consider the following index model setup

$$\begin{aligned} Y_0 &= g_0(T) + U_0 \\ D &= 1 \qquad \text{if } \psi(Z) - \nu \geq 0 \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

If Z and ν are independent, then $P(Z) = F_\nu(\psi(Z))$, where $F_\nu(\cdot)$ is the distribution function of ν. In this case identification condition (A-4b$'$) implies

$$\mathrm{E}\left[U_0 \mid D = 1, F_\nu(\psi(Z))\right] = \mathrm{E}\left[U_0 \mid D = 0, F_\nu(\psi(Z))\right] \qquad (*)$$

or when F_ν is strictly increasing,

$$\begin{aligned} & \int_{-\infty}^{\infty} \int_{-\infty}^{\psi(Z)} U_0 f(U_0, \nu \mid \psi(Z)) d\nu dU_0 / F_\nu(\psi(Z)) \\ = & \int_{-\infty}^{\infty} \int_{\psi(Z)}^{\infty} U_0 f(U_0, \nu \mid \psi(Z)) d\nu dU_0 / \left[1 - F_\nu(\psi(Z))\right]. \end{aligned}$$

If, in addition, $\psi(Z)$ is independent of (U_0, ν), and $\mathrm{E}(U_0) = 0$, condition (*) implies

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\psi(Z)} U_0 f(U_0, \nu) d\nu dU_0 = 0,$$

for any $\psi(Z)$, which in turn implies $\mathrm{E}(U_0 \mid \nu = s) = 0$ for any s when $\psi(Z)$ changes smoothly over the real line. Hence under these conditions our identification condition implies there is no selection on unobservables as defined by Heckman and Robb (1985, 1986). However, $\psi(Z)$ may not be statistically independent of (U_0, ν). Thus under the conditions assumed in the conventional selection model, the identification condition (A-4b$'$) may or may not imply selection on unobservables depending on whether $\psi(Z)$ is independent of (U_0, ν) or not.

5 Estimating The Mean Effect of Treatment: Should One Use the Propensity Score or Not?

Under (A-1$'$) with $S = \mathrm{Supp}(X|D = 1)$ and random sampling across individuals, if one knew $\mathrm{E}(Y_0|D = 0, X = x)$, a consistent estimator of (P-2) is

$$\hat{\Delta}_X = N_1^{-1} \sum_{i \in \mathbf{I_1}} [Y_{1,i} - \mathrm{E}(Y_0|D = 0, X = X_i)].$$

where $\mathbf{I}_1$ is the set of i indices corresponding to observations for which $D_i = 1$. If we assume

$$\mathrm{E}(Y_0|D = 1, P(X)) = \mathrm{E}(Y_0|D = 0, P(X)) \text{ for } X \in \mathrm{Supp}(P(X)|D = 1), \qquad \text{(A-1}''\text{)}$$

which is an implication of (A-1), and $\mathrm{E}(Y_0|D=0,P(X)=p)$ is known, the estimator:

$$\hat{\Delta}_P = N_1^{-1} \sum_{i \in \mathbf{I_1}} [Y_{1,i} - \mathrm{E}(Y_0 \mid D=0, P(X)=P(X_i))]$$

is consistent for $E(\Delta \mid D=1)$.

We compare the efficiency of the two estimators, $\hat{\Delta}_P$ and $\hat{\Delta}_X$. We show that neither is necessarily more efficient than the other. Neither estimator is feasible because both assume the conditional mean function and $P(X)$ are known whereas in practice they need to be estimated. However, the analysis of this case is of interest because the basic intuition from the simple theorem established below continues to hold when the conditional mean function and $P(X)$ are estimated.

Theorem 1 *Assume*

(i) *(A-1′) and (A-1″) hold for* $S = \mathrm{Supp}(X|D=1)$.

(ii) $\{Y_{1i}, X_i\}_{i \in \mathbf{I_1}}$ *are independent and identically distributed, and*

(iii) $0 < E(Y_0^2) \cdot E(Y_1^2) < \infty$.

Then $\hat{\Delta}_X$ *and* $\hat{\Delta}_P$ *are both consistent estimators of (P-2) with asymptotic distributions that are normal with mean 0 and asymptotic variances* V_X *and* V_P*, respectively, where*

$$V_X = \mathrm{E}[\mathrm{Var}(Y_1 \mid D=1, X) \mid D=1] + \mathrm{Var}[\mathrm{E}(Y_1 - Y_0 \mid D=1, X) \mid D=1]$$

and

$$V_P = \mathrm{E}[\mathrm{Var}(Y_1 \mid D=1, P(X)) \mid D=1] + \mathrm{Var}[\mathrm{E}(Y_1 - Y_0 \mid D=1, P(X)) \mid D=1] \quad \square.$$

The theorem directly follows from the central limit theorem for iid sampling with finite second moment and for the sake of brevity its proof is deleted.

Observe that

$$\mathrm{E}[\mathrm{Var}(Y_1 \mid D=1, X) \mid D=1] \leq \mathrm{E}[\mathrm{Var}(Y_1 \mid D=1, P(X)) \mid D=1]$$

because X is in general a better predictor than $P(X)$ but

$$\mathrm{Var}[\mathrm{E}(Y_1 - Y_0 \mid D=1, X) \mid D=1] \geq \mathrm{Var}[\mathrm{E}(Y_1 - Y_0 \mid D=1, P(X)) \mid D=1]$$

because vector X provides a finer conditioning variable than $P(X)$. In general, there are both costs and benefits of conditioning on a random vector X rather than $P(X)$. Using this observation, we can construct examples both where $V_X \leq V_P$ and where $V_X \geq V_P$.

Consider first the special case where the treatment effect is constant, that is $\mathrm{E}(Y_1 - Y_0 \mid D=1, X)$ is constant. An iterated expectation argument implies that $\mathrm{E}(Y_1 - Y_0 \mid D=1, P(X))$ is also constant. Thus, the first inequality, $V_X \leq V_P$ holds in this case. On the other hand, if $Y_1 = m(P(X)) + U$ for some measurable function $m(\cdot)$ and U and X are independent, then

$$V_X - V_P = \mathrm{Var}[\mathrm{E}(Y_0 \mid D=1, X) \mid D=1] - \mathrm{Var}[\mathrm{E}(Y_0 \mid D=1, P(X)) \mid D=1]$$

which is non-negative because vector X provides a finer conditioning variable than $P(X)$. So in this case $V_X \geq V_P$.

When the treatment effect is constant as in the conventional econometric evaluation models, there

is only an advantage of conditioning on X rather than on $P(X)$ and there is no cost.[10] When the outcome Y_1 depends on X only through $P(X)$, there is no advantage to conditioning on X over conditioning on $P(X)$.

Thus far we assume that $P(X)$ is known. In the next section, we investigate the more realistic situation where it is necessary to estimate both $P(X)$ and the conditional means. In this more realistic case, the trade-off between the two terms in V_X and V_P persists.[11]

When we need to estimate $P(X)$ or $\mathrm{E}(Y_0 \mid D = 0, X)$, the dimensionality of the X is a major drawback to the practical application of the matching method or to the use of conventional nonparametric regression. Both are data-hungry statistical procedures. For high dimensional X variables, neither method is feasible in samples of the size typically available to social scientists. Sample sizes per cell become small for matching methods with discrete X's. Rates of convergence slow down in high-dimensional nonparametric methods. In a parametric regression setting, one may evade this problem by assuming functional forms for $\mathrm{E}(U_0 \mid X)$ (see *e.g.* Barnow, Cain and Goldberger, 1980 and the discussion in Heckman and Robb, 1985, 1986), but this approach discards a major advantage of the matching method because it forces the investigator to make arbitrary assumptions about functional forms of estimating equations.

Conditioning on the propensity score avoids the dimensionality problem by estimating the mean function conditional on a one-dimensional propensity score $P(X)$. However, in practice one must

[10] Heckman (1992), Heckman and Smith (1993) and Heckman, Smith and Clements (1993, 1997) discuss the central role of the homogeneous response assumption in conventional econometric models of program evaluation.

[11] If we knew $\mathrm{E}(Y_1|D = 1, P(X) = p)$ as well, the estimator

$$N_1^{-1} \sum_{i \in \mathbf{I_1}} [E(Y_1|D = 1, P(X) = P(X_i)) - E(Y_0|D = 0, P(X) = P(X_i))]$$

would be more efficient than Δ_P. In practical applications, we don't know either $E(Y_1|D = 1, P(X) = p)$ or $E(Y_0|D = 0, P(X) = p)$ so this point is only a theoretical curiosum and is not investigated further.

estimate the propensity score. If it is estimated nonparametrically, we again encounter the curse of dimensionality. The asymptotic distribution theorem below shows that the bias and the asymptotic variance of the estimator of the propensity score affects the asymptotic distribution of the averaged matching estimator more the larger the effect of a change in the propensity score on the conditional means of outcomes.

6 Asymptotic Distribution Theory For Kernel-Based Matching Estimators

We present an asymptotic theory for our estimator of treatment effect (P-2) using either identifying assumption (A-3) or (A-4b$'$). The proof justifies the use of estimated P values under general conditions about the distribution of X.

We develop a general asymptotic distribution theory for kernel-regression-based and local-polynomial-regression-based matching estimators of (P-2). Let T and Z be not necessarily mutually exclusive subvectors of X, as before. When a function depends on a random variable, we use corresponding lower case letters to denote its argument, for example, $g(t,p) = \mathrm{E}(Y_0|D=1, T=t, P(Z)=p)$ and $P(z) = \Pr(D=1|Z=z)$. Although not explicit in the notation, it is important to remember that $g(t,p)$ refers to the conditional expectation conditional on $D=1$ as well as $T=t$ and $P(Z)=p$. We consider estimators of $g(t,P(z))$ where $P(z)$ must be estimated. Thus we consider an estimator $\hat{g}(t,\hat{P}(z))$, where $\hat{P}(z)$ is an estimator of $P(z)$. The general class of estimators of (P-2) that we analyze

are of the form:

$$\hat{\Delta} = \frac{N_1^{-1} \sum\limits_{i \in \mathbf{I_1}} [Y_{1,i} - \hat{g}(T_i, \hat{P}(Z_i))] I(X_i \in \hat{S})}{N_1^{-1} \sum\limits_{i \in \mathbf{I_1}} I(X_i \in \hat{S})}, \tag{6}$$

where $I(A) = 1$ if A holds and $= 0$ otherwise and $\hat{S}$ is an estimator of S, the region of overlapping support, where $S = \text{Supp}\{X|D=1\} \cap \text{Supp}\{X|D=0\}$.

To establish the properties of matching estimators of the form $\hat{\Delta}$ based on different estimators of $P(z)$ and $g(t, P(z))$, we use a class of estimators which we call *asymptotically linear estimators with trimming* . We analyze their properties by proving a series of lemmas and corollaries leading up to Theorem 2. With regard to the estimators $\hat{P}(z)$ and $\hat{g}(t,p)$, we only assume that they can be written as an average of some function of the data plus residual terms with appropriate convergence properties that are specified below. We start by defining the class of asymptotically linear estimators with trimming.

Definition 1 *An estimator $\hat{\theta}(x)$ of $\theta(x)$ is an asymptotically linear estimator with trimming $I(x \in \hat{S})$ if and only if there is a function $\psi_n \in \Psi_n$, defined over some subset of a finite-dimensional Euclidean space, and stochastic terms $\hat{b}(x)$ and $\hat{R}(x)$ such that*

(i) $[\hat{\theta}(x) - \theta(x)] I(x \in \hat{S}) = n^{-1} \sum_{i=1}^{n} \psi_n(X_i, Y_i; x) + \hat{b}(x) + \hat{R}(x)$,

(ii) $E\{\psi_n(X_i, Y_i; X) \mid X = x\} = 0$,

(iii) $plim_{n\to\infty} n^{-1/2} \sum_{i=1}^{n} \hat{b}(X_i) = b < \infty$,

(iv) $n^{-1/2} \sum_{i=1}^{n} \hat{R}(X_i) = o_p(1)$. $\square$

An estimator $\hat{\beta}$ of β is called asymptotically linear if

$$\hat{\beta} - \beta = n^{-1} \sum_{i=1}^{n} \psi(Z_i) + o_p(n^{-1/2})$$

holds.[12] Definition 1 is analogous to the conventional definition, but extends it in five ways to accommodate nonparametric estimators. First, since the parameter $\theta(x)$ that we estimate is a function evaluated at a point, we need a notation to indicate the point x at which we estimate it. Conditions (i)-(iv) are expressed in terms of functions of x. Second, for nonparametric estimation, asymptotic linearity only holds over the support of X - the region where the density is bounded away from zero. To define the appropriate conditions for this restricted region, we introduce a trimming function $I(x \in \hat{S})$ that selects observations only if they lie in $\hat{S}$ and discards them otherwise.

Third, nonparametric estimators depend on smoothing parameters and usually have bias functions that converge to zero for particular sequences of smoothing parameters. We introduce a subscript n to the ψ-function and consider it to be an element of a class of functions Ψ_n, instead of a fixed function, in order to accommodate smoothing parameters. For example, in the context of kernel estimators, if we consider a smoothing parameter of the form $\alpha(x) \cdot h_n$, different choices of h_n generate an entire class of functions Ψ_n indexed by a function $\alpha(\cdot)$ for any given kernel.[13] We absorb n^{-1} into the f-function to simplify the notation. We refer to the function ψ_n as a *score function*. The stochastic term $\hat{b}(x)$ is the bias term arising from estimation. For parametric cases, it often happens that $\hat{b}(x) = 0$.

Fourth, we change the notion of the residual term being "small" from $o_p(n^{-1/2})$ to the weaker condition (iv). We will demonstrate that this weaker condition is satisfied by some nonparametric

[12] See *e.g.* Newey and McFadden (1994) p. 2142, for example.
[13] See Ichimura (1996).

estimators when the stronger condition $o_p(n^{-1/2})$ is not. Condition (iii) is required to restrict the behavior of the bias term. The bias term has to be reduced to a rate $o(n^{-1/2})$ in order to properly center expression (i) asymptotically. For the case of a d-dimensional nonparametric model with p-times continuously differentiable functions, Stone (1982) proves that the optimal uniform rate of convergence of the nonparametric regression function with respect to mean square error is $(n/\log n)^{-p/(2p+d)}$. His result implies that some undersmoothing, compared to this optimal rate, is required to achieve the desired rate of convergence in the bias term alone. Note that the higher the dimension of the estimand, the more adjustment in smoothing parameters to reduce bias is required. This is the price that one must pay to safeguard against possible misspecifications of $g(t,p)$ or $P(z)$. It is straightforward to show that parametric estimators of a regression function are asymptotically linear under some mild regularity conditions. In the Appendix we establish that the local polynomial regression estimator of a regression function is also asymptotically linear.

A typical estimator of a parametric regression function $m(x;\beta)$ takes the form $m(x;\hat{\beta})$, where m is a known function and $\hat{\beta}$ is an asymptotically linear estimator, with $\hat{\beta}-\beta = n^{-1}\sum_{i=1}^n \psi(X_i,Y_i) + o_p(n^{-1/2})$. In this case, by a Taylor expansion,

$$\sqrt{n}[m(x,\hat{\beta})-m(x,\beta)] = n^{-1/2}\sum_{i=1}^n [\partial m(x,\beta)/\partial\beta]\psi(X_i,Y_i) + [\partial m(x,\bar{\beta})/\partial\beta - \partial m(x,\beta)/\partial\beta]n^{-1/2}\sum_{i=1}^n \psi(X_i,Y_i) + o_p(1),$$

where $\bar{\beta}$ lies between β and $\hat{\beta}$. When $E\{\psi(X_i,Y_i)\}=0$ and $E\{\psi(X_i,Y_i)\psi(X_i,Y_i)'\}<\infty$, under iid sampling, for example, $n^{-1/2}\sum_{i=1}^n \psi(X_i,Y_i) = O_p(1)$ and $\text{plim}_{n\to\infty}\hat{\beta}=\beta$ so that $\text{plim}_{n\to\infty}|\partial m(x,\bar{\beta})/\partial\beta -$

$\partial m(x,\beta)/\partial\beta| = o_p(1)$ if $\partial m(x,\beta)/\partial\beta$ is Hölder continuous at β.[14]

Under these regularity conditions,

$$\sqrt{n}[m(x,\hat{\beta}) - m(x,\beta)] = n^{-1/2}\sum_{i=1}^{n}[\partial m(x,\beta)/\partial\beta]\psi(X_i,Y_i) + o_p(1).$$

The bias term of the parametric estimator $m(x,\hat{\beta})$ is $\hat{b}(x) = 0$, under the conditions we have specified. The residual term satisfies the stronger condition that is maintained in the traditional definition of asymptotic linearity.

(a) Asymptotic Linearity of the Kernel Regression Estimator

We now establish that the more general kernel regression estimator for nonparametric functions is also asymptotically linear.The following corollary is a consequence of a more general theorem proved in the Appendix for local polynomial regression models used in Heckman, Ichimura, Smith and Todd (1994, revised, 1996a) and Heckman, Ichimura and Todd (1993, published 1997). We present a specialized result here to simplify notation and focus on main ideas. To establish this result we need to invoke the following assumptions.

Assumption 1 *Sampling of $\{X_i, Y_i\}$ is i.i.d., X_i takes values in R^d and Y_i in R, and $Var(Y_i) < \infty$.*

When a function is p-times continuously differentiable and its p-th derivative satisfies Hölder's condition, we call the function p-smooth. Let $m(x) = E\{Y_i|X_i = x\}$.

Assumption 2 *$m(x)$ is $\bar{p}$-smooth, where $\bar{p} > d$.*

[14]A function is Hölder continuous at $X = x_0$ with constant $0 < \alpha \leq 1$ if $|\varphi(x,\theta) - \varphi(x_0,\theta)| \leq C\cdot \parallel x - x_0 \parallel^{\alpha}$ for some $C > 0$ for all x and θ in the domain of the function $\varphi(\cdot,\cdot)$. Usually Hölder continuity is defined for a function with no second argument θ. We assume that usual Hölder continuity holds uniformly over θ whenever there is an additional argument.

We also allow for stochastic bandwidths:

Assumption 3 *Bandwidth sequence a_n satisfies $plim_{n\to\infty} a_n/h_n = \alpha_0 > 0$ for some deterministic sequence $\{h_n\}$ that satisfies $nh_n^d/\log n \to \infty$ and $nh_n^{2\bar{p}} \to c < \infty$ for some $c \geq 0$.*

This assumption implies $2\bar{p} > d$ but a stronger condition is already imposed in Assumption 3.[15]

Assumption 4 *Kernel function $K(\cdot)$ is symmetric, supported on a compact set, and is Lipschitz continuous.*

The assumption of compact support can be replaced by a stronger assumption on the distribution of X_i so that all relevant moments exist. Since we can always choose $K(\cdot)$, but we are usually not free to pick the distribution of X_i, we invoke compactness.

In this paper we consider trimming functions based on S and $\hat{S}$ that have the following structure. Let $f_X(x)$ be the Lebesgue density of X_i, $S = \{x \in R^d; f_X(x) \geq q_0\}$, and $\hat{S} = \{x \in R^d; \hat{f}_X(x) \geq q_0\}$, where $\sup_{x\in S}|\hat{f}_X(x) - f_X(x)|$ converges almost surely to zero, and $f_X(x)$ is p-smooth. We also require that $f_X(x)$ has a continuous Lebesgue density f_f in a neighborhood of q_0 with $f_f(q_0) > 0$. We refer to these sets S and $\hat{S}$ to be *p-nice on S*. The smoothness of $f_X(x)$ simplifies the analysis and hence helps to establish the equicontinuity results we utilize.

Assumption 5 *Trimming is $\bar{p}$-nice on S.*

In order to control the bias of the kernel regression estimator, we need to make additional assumptions. Certain moments of the kernel function need to be 0, the underlying Lebesgue density

[15] Assumption 3 implies $h_n \to 0$ and $\log n \cdot h_n^{2\bar{p}-d} \to 0$. These two imply $2\bar{p} > d$. Also notice that the assumption implies $a_n \to 0$.

of X_i, $f_X(x)$, needs to be smooth, and the point at which the function is estimated needs to be an interior point of the support of X_i. It is demonstrated in the Appendix that these assumptions are not necessary for $\bar{p}$-th order local polynomial regression estimator.

Assumption 6 *Kernel function $K(\cdot)$ has moments of order 1 through $\bar{p}-1$ that are zero.*

Assumption 7 *$f_X(x)$ is $\bar{p}$-smooth.*

Assumption 8 *A point at which $m(\cdot)$ is being estimated is an interior point of the support of X_i.*

The following characterization of the bias is a consequence of Therorem 3 that is proved in the Appendix.

Corollary 1 *Under Assumptions 1–7, if $K(u_1, \ldots, u_d) = k(u_1) \cdots k(u_d)$ where $k(\cdot)$ is a one dimensional kernel, the kernel regression estimator $\hat{m}_0(x)$ of $m(x)$ is asymptotically linear with trimming, where, writing $\varepsilon_i = Y_i - E\{Y_i|X_i\}$, and*

$\psi_n(X_i, Y_i; x) = (n\alpha_0 h_n^d)^{-1} \varepsilon_j K\left((X_i - x)/(\alpha_0 h_n)\right) I(x \in S)/f_X(x)$,

$$\hat{b}(x) = (\alpha_0 h_n)^{\bar{p}} \cdot [f_X(x) \cdot \int K(u)du]^{-1} \times \sum_{s=1}^{\bar{p}} [s!(\bar{p}-s)!]^{-1} \sum_{k=1}^{d} \left[[\int u_k^{\bar{p}} K(u)du][\partial^s m(x)/(\partial x_k)^s] \cdot [\partial^{(\bar{p}-s)} f_X(x)/(\partial x_k)^{(\bar{p}-s)}] \right]. \quad \square$$

Our use of an independent product form for the kernel function simplifies the expression for the bias function. For a more general expression without this assumption see the Appendix. Corollary 1 differs from previous analyses in the generality with which we characterize the residual term.

(b) Extensions To The Case Of Local Polynomial Regression

In the Appendix, we consider the more general case in which the local polynomial regression estimator for $\hat{g}(t,p)$ is asymptotically linear with trimming with a uniformly consistent derivative. The latter property is useful because, as the next lemma shows, if both $\hat{P}(z)$ and $\hat{g}(t,p)$ are asymptotically linear, and if $\partial\hat{g}(t,p)/\partial p$ is uniformly consistent, then $\hat{g}(t,\hat{P}(z))$ is also asymptotically linear under some additional conditions. We also verify in the Appendix that these additional conditions are satisfied for the local polynomial regression estimators.

Let $\bar{P}_t(z)$ be a function that is defined by a Taylor's expansion of $\hat{g}(t,\hat{P}(z))$ in the neighborhood of $P(z)$, i.e. $\hat{g}(t,\hat{P}(z)) = \hat{g}(t,P(z)) + \partial\hat{g}(t,\bar{P}_t(z))/\partial p \cdot [\hat{P}(z) - P(z)]$.

Lemma 1 *Suppose that*

(i) *Both $\hat{P}(z)$ and $\hat{g}(t,p)$ are asymptotically linear with trimming where*

$$[\hat{P}(z) - P(z)]I(x \in \hat{S}) = n^{-1}\sum_{j=1}^{n}\psi_{np}(D_j, Z_j; z) + \hat{b}_p(z) + \hat{R}_p(z),$$

$$[\hat{g}(t,p) - g(t,p)]I(x \in \hat{S}) = n^{-1}\sum_{j=1}^{n}\psi_{ng}(Y_j, T_j, P(Z_j); t, p) + \hat{b}_g(t,p) + \hat{R}_g(t,p).$$

(ii) *$\partial\hat{g}(t,p)/\partial p$ and $\hat{P}(z)$ are uniformly consistent and converge to $\partial g(t,p)/\partial p$ and $P(z)$, respectively and that $\partial g(t,p)/\partial p$ is continuous.*

(iii) *$plim_{n\to\infty} n^{-1/2}\sum_{i=1}^{n}\hat{b}_g(T_i, P(Z_i)) = b_g$ and*

$plim_{n\to\infty} n^{-1/2}\sum_{i=1}^{n}\partial g(T_i, P(Z_i))/\partial p \cdot \hat{b}_p(T_i, P(Z_i)) = b_{g_p}$,

(iv) *$plim_{n\to\infty} n^{-1/2}\sum_{i=1}^{n}[\partial\hat{g}(T_i, \bar{P}_{T_i}(Z_i))/\partial p - \partial g(T_i, P(Z_i))/\partial p] \cdot \hat{R}_p(Z_i) = 0$,*

(v) *$plim_{n\to\infty} n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}[\partial\hat{g}(T_i, \bar{P}_{T_i}(Z_i))/\partial p - \partial g(T_i, P(Z_i))/\partial p] \cdot \psi_{np}(D_j, Z_j; Z_i) = 0$.*

then $\hat{g}(t, \hat{P}(z))$ *is also asymptotically linear where*

$$
\begin{aligned}
& [\hat{g}(t, \hat{P}(z)) - g(t, P(z))] I(x \in \hat{S}) \\
= \; & n^{-1} \sum_{j=1}^{n} [\psi_{ng}(Y_j, T_j, P(Z_j); t, P(z)) + \partial g(t, P(z))/\partial p \cdot \psi_{np}(D_j, Z_j; z)] + \hat{b}(x) + \hat{R}(x),
\end{aligned}
$$

and $plim_{n\to\infty} n^{-1/2} \sum_{i=1}^{n} \hat{b}(X_i) = b_g + b_{g_p}$. □

An important property of this expression which we exploit below is that the effect of the kernel function $\psi_{np}(D_j, Z_j; z)$ always enters multiplicatively with $\partial g(t, P(z))/\partial p$. Thus both the bias and variance of $P(z)$ depend on the slope of g with respect to p. Condition (ii) excludes nearest-neighbor type matching estimators with a fixed number of neighbors. With conditions (ii)–(v), the proof of this lemma is just an application of Slutsky's theorem and hence the proof is omitted.

In order to apply the theorem, however, we need to verify the conditions. We sketch the main arguments for the case of parametric estimator $\hat{P}(z)$ of $P(z)$ here and present proofs and discussion of the nonparametric case to the Appendix.

Under the regularity conditions just presented, the bias function for a parametric $\hat{P}(z)$ is zero. Hence condition (iii) holds if $\hat{g}(t, p)$ is asymptotically linear and its derivative is uniformly consistent for the true derivative. Condition (iv) also holds since $|\hat{R}_p(Z_i)| = o_p(n^{-1/2})$ and the derivative of $\hat{g}(t, p)$ is uniformly consistent. Condition (v) can be verified by exploiting the particular form of score function obtained earlier. Observing that $\psi_p(D_j, Z_j; Z_i) = \psi_{p1}(Z_i) \cdot \psi_{p2}(D_j, Z_j)$, we obtain

$$
\begin{aligned}
& n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{n} [\partial \hat{g}(T_i, \bar{P}_{T_i}(Z_i))/\partial p - \partial g(T_i, P(Z_i))/\partial p] \cdot \psi_p(D_j, Z_j; Z_i) \\
& \qquad = n^{-1} \sum_{i=1}^{n} [\partial \hat{g}(T_i, \bar{P}_{T_i}(Z_i))/\partial p - \partial g(T_i, P(Z_i))/\partial p] \psi_{p1}(Z_i) n^{-1/2} \sum_{j=1}^{n} \psi_{p2}(D_j, Z_j),
\end{aligned}
$$

so condition (v) follows from an application of the central limit theorem and the uniform consistency of the derivative of $\hat{g}(t,p)$.

For the case of nonparametric estimators, ψ_{np} does not factor and the double summation does not factor as it does in the case of parametric estimation. For this more general case, we apply the equicontinuity results obtained by Ichimura (1996) for general U-statistics to verify the condition. We verify all the conditions for the local polynomial regression estimators in appendix. Since the derivative of $\hat{g}(t,p)$ needs to be defined we assume

Assumption 9 *$K(\cdot)$ is 1-smooth.*

Lemma 1 implies that the asymptotic distribution theory of $\hat{\Delta}$ can be obtained for those estimators based on asymptotically linear estimators with trimming with no loss of generality. Once this result is established, it can be used in lemma 1 to analyze the properties of two stage estimators of the form $\hat{g}(t,\hat{P}(z))$.

(c) Theorem 2: The Asymptotic Distribution of The Matching Estimator Under General Conditions

Theorem 2 enables us to produce the asymptotic distribution theory of a variety of estimators $\hat{\Delta}$ of the treatment effect under different identifying assumptions. It also produces the asymptotic distribution theory for matching estimators based on different estimators of $g(t,p)$ and $P(z)$. In sections 3 and 4, we presented various matching estimators for the mean effect of treatment on the treated invoking different identifying assumptions. An alternative to matching is the conventional index-sufficient selection estimators that can be used to construct estimators of $\mathrm{E}(Y_0 \mid D=1,X)$, as described in our companion paper Heckman, Ichimura and Todd (1994, 1997) and in Heckman,

Ichimura, Smith and Todd (1994, revised 1996a,b). Our analysis is sufficiently general to cover the distribution theory for that case as well.

Denote the conditional expectation or variance given that X is in S by $\mathrm{E}_S(\cdot)$ or $\mathrm{Var}_S(\cdot)$, respectively. Let the number of obsevations in sets $\mathbf{I_0}$ and $\mathbf{I_1}$ be N_0 and N_1, respectively, where $N = N_0 + N_1$ and that $0 < \lim_{N\to\infty} N_1/N_0 = \theta < \infty$.

Theorem 2 *Under the following conditions:*

(i) $\{Y_{0i}, X_i\}_{i\in\mathbf{I_0}}$ *and* $\{Y_{1i}, X_i\}_{i\in\mathbf{I_1}}$ *are independent and within each group they are i.i.d. and* Y_{0i} *for* $i \in \mathbf{I_0}$ *and* Y_{1i} *for* $i \in \mathbf{I_1}$ *each has a finite second moment.*

(ii) *The estimator* $\hat{g}(x)$ *of* $g(x) = E\{Y_{0i}|D_i = 1, X_i = x\}$ *is asymptotically linear with trimming, where*

$$[\hat{g}(x) - g(x)]I\{x \in \hat{S}\} = N_0^{-1} \sum_{i\in\mathbf{I_0}} \psi_{0N_0N_1}(Y_{0i}, X_i; x) + N_1^{-1} \sum_{i\in\mathbf{I_1}} \psi_{1N_0N_1}(Y_{1i}, X_i; x) + \hat{b}_g(x) + \hat{R}_g(x)$$

and the score functions $\psi_{dN_0N_1}(Y_d, X; x)$ *for* $d = 0$ *and* 1*, the bias term* $\hat{b}_g(x)$*, and the trimming function satisfy,*

(ii-a) $E\{\psi_{dN_0N_1}(Y_{di}, X_i; X)|D_i = d, X, D = 1)\} = 0$ *for* $d = 0$ *and* 1*, and*

$Var\{\psi_{dN_0N_1}(Y_{di}, X_i; X)\} = o(N)$ *for each* $i \in \mathbf{I_0} \cup \mathbf{I_1}$.

(ii-b) $plim_{N_1\to\infty} N_1^{-1/2} \sum_{i\in\mathbf{I_1}} \hat{b}(X_i) = b$

(ii-c) $plim_{N_1\to\infty} Var\{E[\psi_{0N_0N_1}(Y_{0i}, X_i; X) \mid Y_{0i}, D_i = 0, X_i, D = 1]|D = 1\} = V_0 < \infty$

$plim_{N_1\to\infty} Var\{E[\psi_{1N_0N_1}(Y_{1i}, X_i; X) \mid Y_{1i}, D_i = 1, X_i, D = 1]|D = 1\} = V_1 < \infty$,

and that

$$\lim_{N_1\to\infty} E\{[Y_{1i}-g(X_i)]I(X_i\in S)E[\psi_{1N_0N_1}(Y_{1i},X_i;X)\mid Y_{1i},D_i=1,X_i,D=1]|D=1\}=Cov_1,$$

(ii-d) *S and $\hat{S}$ are $\bar{p}$-nice on S, where $\bar{p}>d$, where d is the number of regressors in X and $\hat{f}(x)$ is a kernel density estimator that uses a kernel function that satisfies Assumption 6.*

Then under (A-1′), the asymptotic distribution of

$$N_1^{1/2}\left[\frac{N_1^{-1}\sum\limits_{i\in\mathbf{I_1}}[Y_{1i}-\hat{g}(X_i)]I(X_i\in\hat{S})}{N_1^{-1}\sum\limits_{i\in\mathbf{I_1}}I(X_i\in\hat{S})}-E_S(Y_1-Y_0|D=1)\right]$$

is normal with mean $b/\Pr(X\in S|D=1)$ and asymptotic variance

$$\begin{aligned}\Pr(X\in S|D=1)^{-1}\{&Var_S[E_S(Y_1-Y_0|T,P(Z),D=1)|D=1]\\&+E_S[Var_S(Y_1|T,P(Z),D=1)|D=1]\}\\&+\Pr(X\in S|D=1)^{-2}\{V_1+2\cdot Cov_1+\theta V_0\}.\quad\square\end{aligned}$$

Proof: See the Appendix

Theorem 2 shows that the asymptotic variance consists of five components. The first two terms are the same as those previously presented in Theorem 1. The latter three terms are the contributions to variance that arise from estimating of $g(x)=E\{Y_i|D_i=1,X_i=x\}$. The third and the fourth terms arise from using observations for which $D=1$ to estimate $g(x)$. If we use just observations for which $D=0$ to estimate $g(x)$, as in the case of the simple matching estimator, then these two terms

do not appear and we only acquire the fifth term.[16] We consider the more general case with all five terms. If N_0 is much larger than N_1, then the sampling variation contribution of $D = 0$ is small as θ is small.

Condition (i) covers both random and choice-based sampling and enables us to avoid degeneracies and to apply a central limit theorem. Condition (ii) elaborates the asymptotic linearity condition for the estimator of $g(x)$. We assume p-nice trimming. The additional condition on the trimming function is required to reduce the bias that arises in estimating the support.

Note that there is no need for $g(x)$ to be smooth. A smoothness condition on $g(x)$ is used solely to establish asymptotic linearity of the estimator of $g(x)$. Also note that the sampling theory above is obtained under mean independence:

$$\mathrm{E}_S(Y_0 \mid D = 1, X) = \mathrm{E}_S(Y_0 \mid X) = \mathrm{E}_S(Y_0 \mid D = 0, X).$$

Strong ignorability conditions given by (A-1), (A-2) or (A-3) while conventional in the matching literature are not needed but they obviously imply these equalities.

Theorem 2 can be combined with the earlier results to obtain an asymptotic distribution theory for estimators that use $\hat{g}(t, \hat{P}(z))$. One only needs to replace function $\psi_N(Y_{D_i i}, D_i, X_i; X)$ and the bias term by those obtained in Lemma 2.

[16] An earlier version of the paper assumed that only observations for which $D = 0$ are used to estimate $g(x)$.

7 Answers To The Three Questions of Section 1 and More General Questions Concerning the Value of Apriori Information

Armed with these results, we now investigate the three questions posed in the Section 1.

(1) Is it better to match on P(X) or X if you know P(X)?

Matching on X, Δ_X, involves d-dimensional nonparametric regression function estimation whereas matching on $P(X)$, $\hat{\Delta}_P$, only involves one dimensional nonparametric regression function estimation. Thus from the perspective of bias, matching on $P(X)$ is better in the sense that it allows $\sqrt{N}$-consistent estimation of (P-2) for a wider class of models than is possible if matching is performed directly on X. This is because estimation of higher-dimensional functions requires that the underlying functions be smoother for bias terms to converge to zero. If we specify parameteric regression models, the distinction does not arise if the model is correctly specified.

When we restrict consideration to models that permit$\sqrt{N}$-consistent estimation either by matching on $P(X)$ or on X, the asymptotic variance of $\hat{\Delta}_P$ is not necessarily smaller than that of $\hat{\Delta}_X$. To see this, consider the case where we use a kernel regression for the $D = 0$ observations *i.e.* those with $i \in \mathbf{I}_0$. In this case score function $\psi_{1N_0N_1}(Y_{1i}, X_i; X) = 0$ and

$$\psi_{0N_0N_1}(Y_{0i}, X_i; x) = \frac{\varepsilon_i K\left((X_i - x)/a_{N_0}\right) I(x \in S)}{a_{N_0}^d f_X(x|D=0) \int K(u)du},$$

where $\varepsilon_i = Y_{0i} - E\{Y_{0i}|X_i, D_i = 0\}$ and we write $f_X(x|D=0)$ for the Lebesgue density of X_i given $D_i = 0$. (We use analogous expressions to denote various Lebesgue densities.) Clearly V_1 and Cov_1 are zero in this case. Using the score function we can calculate V_0 when we match on X. Denoting

this variance by V_{0X},

$$V_{0X} = \lim_{N_0 \to \infty} Var\{E[\psi_{0N_0N_1}(Y_{0i}, X_i; X) \mid Y_{0i}, D_i = 0, X_i, D = 1]|D = 1\}$$
$$= \lim_{N_0 \to \infty} Var\left\{ E\left[\frac{\varepsilon_i K\left((X_i - X)/a_{N_0}\right) I(X \in S)}{a_{N_0}^d f_X(X|D=0) \int K(u)du} |Y_{0i}, D_i = 0, X_i, D = 1 \right] |D = 1 \right\}.$$

Now observe that conditioning on X_i and Y_{0i}, ε_i is given, so that we may write the last expression as

$$Var\left\{ \varepsilon_i E\left[\frac{K((X_i - X)/a_{N_0}) I(X \in S)}{a_{N_0}^d f_X(X \mid D=0) \int K(u)du} \mid D_i = 0, X_i, D = 1 \right] |D_i = 0, D = 1 \right\}.$$

Now

$$E\left[\frac{K((X_i - X)/a_{N_0}) I(X \in S)}{a_{N_0}^d f_x(X \mid D=0) \int K(u)du} \mid D_i = 0, X_i, D = 1 \right]$$

can be written in the following way, making the change of variable $(X_i - X)/a_{N_0} = w$:

$$\int \frac{K(w) I([X_i - a_{N_0} w] \in S)}{\int K(u)du} \frac{f(X_i - a_{N_0} w \mid D = 1)}{f(X_i - a_{N_0} w \mid D = 0)} dw.$$

Taking limits as $N_0 \to \infty$, and using assumptions 3, 4 and 7, so we can take limits inside the integral

$$\lim_{N_0 \to \infty} \mathrm{E}\left[\frac{K((X_i - X)/a_{N_0}) I(X \in S)}{a_{N_0}^d f_X(X \mid D=0) \int K(u)du} |D_i = 0, X_i, D = 1 \right] = \frac{f(X_i|D=1)}{f(X_i|D=0)} I(X_i \in S)$$

since $a_{N_0} \to 0$ and $\int K(w)dw / \int K(u)du = 1$. Thus

$$V_{0X} = E_S\left[\frac{\mathrm{Var}(Y_{0i}|X_i, D_i = 0) f_X^2(X_i|D_i = 1)}{f_X^2(X_i|D_i = 0)} |D_i = 0 \right] \Pr\{X_i \in S|D_i = 0\}.$$

Hence the asymptotic variance of $\hat{\Delta}_X$ is: writing $\lambda = \Pr\{X \in S|D=0\}/\Pr(X \in S|D=1)$,

$$\Pr(X \in S|D=1)^{-1}\{\mathrm{Var}_S[\mathrm{E}_S(Y_1 - Y_0|X, D=1)|D=1] + \mathrm{E}_S[\mathrm{Var}_S(Y_1|X, D=1)|D=1]$$
$$\lambda\theta E_S[Var(Y_0|X, D=0)f_X^2(X|D=1)/f_X^2(X|D=0)|D=0]\}.$$

Similarly for $\hat{\Delta}_P$, V_{0P} is

$$\Pr(X \in S|D=1)^{-1}\{\mathrm{Var}_S[\mathrm{E}_S(Y_1 - Y_0|P(X), D=1)|D=1] + \mathrm{E}_S[\mathrm{Var}_S(Y_1|P(X), D=1)|D=1]$$
$$+ \lambda\theta\mathrm{E}_S[\mathrm{Var}(Y_0|P(X), D=0)f_p^2(P(X)|D=1)/f_p^2(P(X)|D=0)|D=0]\}.$$

The first two terms for both variance expressions are the same as those that appear in V_X and V_P in Theorem 1. To see that one variance is not necessarily smaller than the other, consider the case where $f_X(X \mid D=1) = f_X(X \mid D=0)$ and $\theta = 1$. Clearly in this case $f_p(P(X) \mid D=1) = f_p(P(X) \mid D=0)$. Propensity score matching has smaller variance if and only if

$$\mathrm{E}_S\left\{E_S(Y_1 \mid P(X), D=1)\mathrm{E}_S(Y_0 \mid P(X), D=1) \mid D=1\right\}$$
$$> \mathrm{E}_S\left\{E_S(Y_1 \mid X, D=1)\mathrm{E}_S(Y_0 \mid X, D=1) \mid D=1\right\}.$$

Since the inequality does not necessarily hold, the propensity score matching estimator in itself does not necessarily improve upon the regular matching estimator.[17,18]

(2) What are the effects on asymptotic bias and variance if we use an estimated value of

[17]For example $E_S(Y_1 \mid X, D=1) = E_S(Y_0 \mid X, D=1)$ or $E_S(Y_1 \mid X, D=1) = -E_S(Y_0 \mid X, D=1)$ can hold.

[18]In an apparently independent analysis, Hahn (1996) considers a special case of models considered in this paper and shows in a model with no exclusion restrictions that when P is not known, the estimated propensity score estimator is efficient and that knowledge of P improves efficiency.

P?

When $P(x)$ is estimated nonparametrically, the smaller bias that arises from matching on the propensity score no longer holds true if estimation of $P(x)$ is a d-dimensional nonparametric estimation problem where $d > 1$. In addition, estimation of $P(x)$ increases the asymptotic variance. Lemma 1 informs us that the score, when we use estimated $P(z)$ but no other conditioning variables, is

$$\psi_{dN_0N_1g}(Y_{dj}, P(Z_j); P(z)) + \partial g(P(z))/\partial p \cdot \psi_{Np}(D_j, Z_j; z),$$

for $i \in \mathbf{I_d}$, $d = 0, 1$, where $\psi_{dN_0N_1g}$ are the scores for estimating $g(p)$ and ψ_{Np} is the score for estimating $P(z)$. By assumption (ii-a) they are not correlated with $\partial g(P(z))/\partial p \cdot \psi_{Np}(D_j, Z_j; z)$, and hence the variance of the sum of the scores is the sum of the variances of each score. So the variance increases by the variance contribution of the score $\partial g(P(z))/\partial p \cdot \psi_{Np}(D_j, Z_j; z)$ when we use estimated, rather than known, $P(z)$. Even with the additional term, however, matching on X does not necessarily dominate matching on $P(X)$ because the additional term may be arbitrarily close to zero when $g'(p)$ is close to zero.

(3) What are the benefits, if any, of econometric separability and exclusion restrictions on the bias and variance of matching estimators?

We first consider exclusion restrictions in the estimation of $P(x)$. Again we derive the asymptotic variance formulae explicitly using kernel regression estimator. Using Corollary 1, the score function for estimating $P(x)$ is

$$\psi_{Np}(X_j, D_j; x) = \frac{u_j K\left((X_j - x)/a_N\right) I(x \in S)}{a_N^d f_X(x) \int K(u)du},$$

where $u_j = D_j - E\{D_j|X_j\}$. Hence the variance contribution of estimation of $P(z)$ without imposing

exclusion restrictions is

$$
\begin{aligned}
V_{2X} = \lim_{N\to\infty} Var\{E[\partial g(P(Z))/\partial p \\
\times a_N^{-d} u_j K\left((X_j - X)/a_N\right) I(X \in S)/[f_X(X) \int K(u)du] \mid D_j, X_j, D = 1]\} \\
= \mathrm{E}_S[\mathrm{Var}(D_j|X_j)[\partial g(P(Z_j))/\partial p]^2 \cdot f_X^2(X_j|D=1)/f_X^2(X_j)][\Pr\{X_j \in S\}]^{-1}.
\end{aligned}
$$

Analogously, we define the variance contribution of estimating of $P(z)$ imposing exclusion restrictions by V_{2Z}. Observe that when Z is a subset of the variables in X, and when there are exclusion restrictions so $P(X) = P(Z)$ then one can show that $V_{2Z} \leq V_{2X}$. Thus, exclusion restrictions in estimating $P(X)$ reduce the asymptotic variance of the matching estimator—an intuitively obvious result.

To show this first note that in this case $\mathrm{Var}(D \mid X) = \mathrm{Var}(D \mid Z)$. Thus

$$
\begin{aligned}
&[V_{2X} - V_{2Z}] \cdot \Pr\{X \in S\} \\
&= \mathrm{E}_S\{\mathrm{Var}(D|Z)[\partial g(P(Z))/\partial p]^2 \cdot [f_X^2(X|D=1)/f_X^2(X) - f_Z^2(Z|D=1)/f_Z^2(Z)]\} \\
&= \mathrm{E}_S\{\mathrm{Var}(D|Z)[\partial g(P(Z))/\partial p]^2 \cdot \left[\frac{f_Z^2(Z|D=1)}{f_Z^2(Z)}\right] \cdot \left[E\left(\frac{f_X^2(X|Z,D=1)}{f_X^2(X \mid Z)}|Z\right) - 1\right]\} \\
&\geq \mathrm{E}_S\{\mathrm{Var}(D|Z)[\partial g(P(Z))/\partial p]^2 \cdot \left[\frac{f_Z^2(Z|D=1)}{f_Z^2(Z)}\right] \cdot \left[E\left(\frac{f_X(X|Z,D=1)}{f_X(X \mid Z)}|Z\right)^2 - 1\right]\} \\
&= 0.
\end{aligned}
$$

Since the other variance terms are the same, imposing the exclusion restriction helps to reduce the asymptotic variance by reducing the estimation error due to estimating the propensity score. The same is true when we estimate the propensity score by a parametric method. It is straightforward to show that, holding all other things constant, the lower the dimension of Z, the less the variance in the

matching estimator. Exclusion restrictions in T also reduce the asymptotic variance of the matching estimator.

By the same argument, it follows that $\mathrm{E}\{[f_X^2(X \mid D=1)/f_X^2(X \mid D=0)] - 1 \mid D=0\} \geq 0$. This implies that under homoskedasticity for Y_0, the case when $f(X \mid D=1) = f(X \mid D=0)$ yields the smallest variance.

We next examine the consequences of imposing an additive separability restriction on the asymptotic distribution. We find that imposing additive separability does not necessarily lead to a gain in efficiency. This is so, even when the additively separable variables are independent. We describe this using the estimators studied by Tjostheim and Auestad (1994) and Linton and Nielsen (1995).[19]

They consider estimation of $g_1(X_1)$, $g_2(X_2)$ in

$$E(Y \mid X_1 = x_1, X_2 = x_2) = g_1(X_1) + g_2(X_2)$$

where $x = (x_1, x_2)$. There are no overlapping variables among X_1 and X_2. In our context, $E(Y|X = x) = g(t) + K(P(z))$ and $\mathrm{E}(Y|X = x)$ is the parameter of interest. In order to focus on the effect of imposing additive separability, we assume $P(z)$ to be known so that we write P for $P(Z)$.

Their estimation method first estimates $E\{Y|T = t, P = p\} = g(t) + K(p)$ non-parametrically, say by $\hat{E}\{Y|T = t, P = p\}$, and then integrates $\hat{E}\{Y|T = t, P = p\}$ over p using an estimated marginal distribution of P. Denote the estimator by $\hat{g}(t)$. Then under additive separability, $\hat{g}(t)$ consistently estimates $g(t) + E\{K(P)\}$. Analogously one can integrate $\hat{E}\{Y|T = t, P = p\} - \hat{g}(t)$ over t using an estimated marginal distribution of T to obtain a consistent estimator of $K(p) - E\{K(P)\}$. We

[19]The derivative of $E\{Y|T = t, P = p\}$ with respect to p only depends on p if it is additively separable. Fan, Härdle, and Mammen (1996) exploits this property in their estimation. Using this estimator does not lead to an improvement in efficiency, either.

add the estimators to obtain the estimator of $E(Y \mid X = x)$ that imposes additive separability. The contribution of estimation of the regression function to asymptotic variance when T and P are independent and additive separability is imposed, is $\Pr(X \in S \mid D = 1)^{-1}$ times

$$\theta E_S \left\{ \mathrm{Var}_S(Y_0 \mid T, P, D = 0) \left[\frac{f(P \mid D = 1)}{f(P \mid D = 0)} + \frac{f(T \mid D = 1)}{f(T \mid D = 0)} - 1 \right]^2 \right\}.^{20}$$

When the additive separability is not used, it is $\Pr(X \in S \mid D = 1)^{-1}$ times

$$\theta E_S \left\{ \mathrm{Var}_S(Y_0 \mid T, P, D = 0) \left[\frac{f(P \mid D = 1) \cdot f(T \mid D = 1)}{f(P \mid D = 0) \cdot f(T \mid D = 0)} \right]^2 \right\}.$$

Note that the first expression is not necessarily smaller, since $f(P \mid D = 1) \cdot f(T \mid D = 1)$ can be small without both $f(P \mid D = 1)$ and $f(T \mid D = 1)$ being simultaneously small.[21]

Imposing additive separability per se does not necessarily improve efficiency. This is in contrast to the case of exclusion restrictions where imposing them always improved efficiency. Whether there exists a method that improves efficiency by exploiting additive separability is not known to us.

Note that when $f(P \mid D = 1) = f(P \mid D = 0)$ and $f(T \mid D = 1) = f(T \mid D = 0)$ both hold, the variance for the additively separable case and for the general case coincide. Under homoskedasticity of Y_0, the most efficient case arises when the distributions of $(T, P(Z))$ given $D = 1$ and $(T, P(Z))$ given $D = 0$ coincide. In the additively separable case, only the marginal distributions of $P(Z)$ and

[20] The derivation is straightforward but tedious. Use the asymptotic linear representation of the kernel regression estimator and then obtain the asymptotic linear expression using it.

[21] Let $a(P) = f(P \mid D = 1)/f(P \mid D = 0)$ and $b(T) = f(T \mid D = 1)/f(T \mid D = 0)$ and define an interval $H(T) = [[1 - b(T)]/[1 + b(T)], 1]$ when $b(T) < 1$. If whenever $b(T) > 1$, $a(P) > 1$ and whenever $b(T) < 1$, $a(P) \in H(T)$ holds. Then imposing additive separability improves efficiency. On the other hand, if whenever $b(T) > 1$, $a(P) < 1$ and whenever $b(T) < 1$, $a(P)$ lies outside the interval $H(T)$, then imposing additive separability using the available methods worsens efficiency even if the true model is additively.

T respectively have to coincide, but the basic result is the same.[22]

Note that nearest neighbor matching "automatically" imposes the restriction of balancing the distributions of the data whereas kernel matching does not. While our theorem does not justify the method of nearest neighbor matching, within a kernel matching framework we may be able to reweight the kernel to enforce the restrictions that the two distributions be the same. That is an open question which we will answer in our future research. Note that we clearly need to reweight so that the homoskedasticity condition holds.

8 Summary and Conclusion

This paper examines matching as an econometric method for evaluating social programs. Matching is based on the assumption that conditioning on observables eliminates selective differences between program participants and nonparticipants that are not correctly attributed to the program being evaluated.

We present a framework to justify matching methods that allows analysts to exploit exclusion restrictions and assumptions about additive separability. We then develop a sampling theory for kernel-based matching methods that allows the matching variables to be generated regressors produced from either from parametric or nonparameteric estimation methods. We show that the matching method based on the propensity score does not necessarily reduce the asymptotic bias or the variance of estimators of $M(S)$ compared to traditional matching methds.

The advantage of using the propensity score is simplicity in estimation. When we use the method

[22]The expression above implies that the same can be said for the estimator that is constructed without imposing additive separability. However that result is an artifact of assuming independence of $P(Z)$ and T.

of matching based on propensity scores, we can estimate treatment effects in two stages. First we build a model that describes the program participation decision. Then we construct a model that describes outcomes. In this regard, matching mimics features of the conventional econometric approach to selection bias. (Heckman and Robb, 1986 or Heckman, Ichimura, Smith and Todd, 1994, 1996a.)

A useful extension of our analysis would consider the small sample properties of alternative estimators. In samples of the usual size in economics, cells will be small if matching is made on a high-dimensional X. This problem is less likely to arise when matching is on a single variable like P. This small sample virtue of propensity score matching is not captured by our large sample theory. Intuitively, it appears that the less data hungry propensity score method would be more efficient than a high dimensional matching method.

Our sampling theory demonstrates the value of having the conditional distribution of the regressors the same for $D = 0$ and $D = 1$. This point is to be distinguished from the requirement of a common support that is needed to justify the matching estimator. Whether a weighting scheme can be developed to improve the asymptotic variance remains to be investigated.

References

[1] ARCONES, M. A. and GINÉ, (1993), "Limit Theorems for U-processes," *Annals of Probability*, **21**, 1494–1542.

[2] BARNOW, B., CAIN, G. and GOLDBERGER, A. (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies Review Annual, Volume 5*, ed. by E. Stromsdorfer and G. Farkas (San Francisco: Sage).

[3] BARROS, R. (1986), "Nonparametric Estimation of Causal Effects in Observational Studies," (University of Chicago, mimeo).

[4] COCHRANE, W. G. and RUBIN, D. B. (1973), "Controlling Bias In Observational Studies," *Sankyha*, **35**, 417-446.

[5] DAWID (1979), "Conditional Independence in Statistical Theory", *Journal of The Royal Statistical Society,* Series B, **41**, 1-31.

[6] FAN, J. (1993), "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, **21**, 196-216.

[7] FAN, J. , HÄRDLE, W., and MAMMEN, E. (1996), "Direct Estimation of Low Dimensional Components in Additive Models," working paper.

[8] HAHN, JINYOUNG (1996): "On the Role of the Propensity Score in the Efficient Semiparametric Estimation of Average Treatment Effects," unpublished manuscript, University of Pennsylvania.

[9] HECKMAN, J. , (1974), "Shadow Prices, Market Wages, and Labor Supply", *Econometrica*, **42**, 679-694.

[10] HECKMAN, J. , (1990), "Varieties of Selection Bias", *American Economic Review*, **80**, 313-318.

[11] HECKMAN, J. , (1997), "Instrumental Variables: A Study of the Implicit Assumptions Underlying One Widely Used Estimator for Program Evaluations", forthcoming *Journal of Human Resources*, Summer issue.

[12] HECKMAN, J., ICHIMURA, H., SMITH, J. and TODD, P (1994, revised 1996a), "Nonparametric Characterization of Selection Bias Using Experimental Data, Part I: Definitions, Applications and Empirical Results", unpublished manuscript, University of Chicago.

[13] HECKMAN, J., ICHIMURA, H., SMITH, J. and TODD, P (1994, revised 1996b), "Nonparametric Characterization of Selection Bias Using Experimental Data, Part II: Econometric Theory and Monte Carlo Evidence", unpublished manuscript, University of Chicago.

[14] HECKMAN, J., ICHIMURA, H., SMITH, J. AND TODD, P (1996), "Sources of Selection Bias in Evaluating Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching As A Program Evaluation Method", *Proceedings of The National Academy of Sciences*, 93(23), 13416-13420.

[15] HECKMAN, J., ICHIMURA, H., and TODD, P (1993, revised 1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program", forthcoming, *Review of Economic Studies*, October, 1997.

[16] HECKMAN, J. and ROBB, R. (1985), "Alternative Methods For Evaluating The Impact of Interventions", in James Heckman and Burton Singer, eds. *Longitudinal Analysis of Labor Market Data* (Cambridge, UK: Cambridge University Press).

[17] HECKMAN, J. and ROBB, R. (1986), "Alternative Method For Solving the Problem of Selection Bias in Evaluating The Impact of Treatments on Outcomes," in *Drawing Inferences from Self-Selected Samples*, ed. by Howard Wainer (New York: Springer-Verlag).

[18] HECKMAN, J. and SMITH, J. (1997), "Evaluating The Welfare State", Frisch Symposium Paper, Oslo Norway (1995), forthcoming in *Frisch Centenary*, Econometric Monograph Series, Cambridge: Cambridge University Press.

[19] HECKMAN, J., SMITH, J., and CLEMENTS N. (1993, revised 1997), "Making the Most Out of Social Experiments: Reducing the Intrinsic Uncertainty in Evidence From Randomized Trials With An Application to the National JTPA Experiment." first draft, March, 1993, forthcoming, *Review of Economic Studies*, October.

[20] ICHIMURA, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, **58**, 71-120.

[21] ICHIMURA, H. (1995), "Asymptotic Distribution of Nonparametric and Semiparametric Estimators with Data-Dependent Smoothing Parameters", unpublished manuscript, University of Chicago.

[22] KOLMOGOROV, A. N. and TIHOMIROV, V. M. (1961), "ϵ-Entropy and ϵ-Capacity of Sets in Functional Spaces", *American Mathematical Society Translations, series 2*, **17**, 277-364.

[23] LINTON, O. and NIELSON, J. P. (1995), "A Kernel Method of Estimating Structural Nonparametric Regression Based on Marginal Integration", *Biometrika*, **82**, 93-100.

[24] NOLAN, D. and POLLARD, D. (1987), "U-processes: Rates of Convergence", *Annal of Statistics*, **15**, 405-414.

[25] NEWEY, W. K. and McFadden, D. L. (1994) "Large Sample Estimation And Hypothesis Testing", in *Handbook of Econometrics* Vol IV, edited by R. F. Engle and D. L. McFadden, Elsevier Science B. V., Amsterdam.

[26] MASRY, E. (1995), "Multivariate Local Polynomial Regression for Time Series", unpublished manuscript.

[27] POLLARD, D. (1990), *Empirical Processes: Theory and Applications*, IMS, Hayward, CA.

[28] ROSENBAUM, P. and RUBIN, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, **70**, 41-55.

[29] RUPPERT, D. and WAND, M.P. (1994), "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, **22**, 1346-1370.

[30] SERFLING, R. J. (1980), *Approximation Theorems of Mathematical Statistic*s (New York: NY: Wiley).

[31] SHERMAN, R. P. (1994), "Maximal Inequalities for Degenerate U-processes with Applications to Optimization Estimators", *Annal of Statistics*, **22**, 439-459.

[32] STONE, C. (1982), "Optimal Global Rates of Convergence For Nonparametric Regression", *Annals of Statistics*, **10**, 1040-1053.

[33] TJOSTHEIM, D. and AUESTAD, B. H. (1994), "Nonparametric Identification of Nonlinear Time Series: Projections", *Journal of the American Statistical Association,* **89**, 1398-1409.

[34] WESTAT, INC. (1980), "Net Impact Report No. 1, Impact on 1977 Earnings of New FY 1976 CETA Enrollees in Selected Program Activities", Rockville, Maryland, Westat Inc.

[35] WESTAT, INC. (1982), "CLMS Follow up Report No. 7 (18 Months After Entry), Postprogram Experiences and Pre/Post Comparisons For Terminees Who Entered CETA During July 1976 Through September 1977", Rockville, Maryland, Westat, Inc.

[36] WESTAT, INC. (1984), "Summary of Net Impact Results", Rockville, Mayland, Westat, Inc.

A Appendix

In this Appendix we prove Lemma 1, verify the conditions of Lemma 2 for the case of a local polynomial regression estimator, and prove Theorem 2. We first establish the property that local polynomial regression estimators are asymptotically linear with trimming.

A.1 Proof of Theorem 3

We show that local polynomial regression estimators are asymptotically linear with trimming. Lemma 1 follows as a corollary.

The local polynomial regression estimator of a function and its derivatives is based on an idea of approximating the function at a point by a Taylor's series expansion and then to estimate the coefficients using data in a neighborhood of the point. In order to present the results, therefore, we first develop a compact notation to write a multivariate Taylor series expansion. Let $x = (x_1, ..., x_d)$ and $q = (q_1, ..., q_d) \in R^d$ where q_j $(j = 1, ..., d)$ are non-negative integers. Also let $x^q = x_1^{q_1} \cdots x_d^{q_d}/(q_1! \cdots q_d!)$. Note that we include $(q_1! \cdots q_d!)$ in the definition. This enables us to study the derivative of x^q without introducing new notation; for example, $\partial x^q/\partial x_1 = x^{\bar{q}}$ where $\bar{q} = (q_1 - 1, ..., q_d)$, if $q_1 \geq 1$ and 0 otherwise. When the sum of the elements of q is s, x^q corresponds to a Taylor series polynomial associated with the term $\partial^s m(x)/(\partial x_1^{q_1} \cdots \partial x_d^{q_d})$. In order to consider all polynomials that correspond to s-th order derivatives we next define a vector whose elements are themselves distinct vectors of nonnegative integers whose elements sum to s. We denote this row vector by $Q(s) = ((q_1, ..., q_d))_{q_1+...+q_d=s}$; that is $Q(s)$ is a row vector of length $(s + d - 1)!/[s!(d - 1)!]$ whose typical element is a row vector $(q_1, ..., q_d)$, which has arguments that sum to s. For concreteness we assume that $\{(q_1, ..., q_d)\}$ are ordered according to the magnitude of $\sum_{j=1}^{d} 10^{d-j} q_j$ from largest

to smallest. We define a row vector $x^{Q(s)} = (x^{(q_1,...,q_d)})_{q_1+...+q_d=s}$. This row vector corresponds to the polynomial terms of degree s. Let $x^{Q_p} = (x^{Q(s)})_{s\in\{1,...p\}}$. This row vector represents the whole polynomial up to degree p from lowest to the highest.

Also let $m^{(s)}(x)$ for $s \geq 1$ to denote a row vector whose typical element is $\partial^s m(x)/(\partial x_1^{q_1} \cdots \partial x_d^{q_d})$ where $q_1 + \cdots + q_d = s$ and the elements are ordered in the same way $\{(q_1, ..., q_d)\}$ are ordered. We also write $m^{(0)}(x) = m(x)$. Let $\beta_p^*(x_0) = (m^{(0)}(x_0), ..., m^{(p)}(x_0))'$. In this notation, Taylor's expansion of $m(x)$ at x_0 to order p without a remainder term can now be written as $(x - x_0)^{Q_p}\beta_p^*(x_0)$.

We now define the local polynomial regression estimator with a global smoothing parameter αh_n, where $\alpha \in [\alpha_0 - \delta, \alpha_0 + \delta]$ for some $\alpha_0 > 0$ and $\alpha_0 > \delta > 0$. We denote $\mathcal{A} = [\alpha_0 - \delta, \alpha_0 + \delta]$. Let $K_h(s) = (\alpha h_n)^{-d} K(s/(\alpha h_n))$ and let $\beta = (\beta_0', ..., \beta_p')'$, where β_t' is comformable with $m^{(t)}(x_0)$, for $t = 0, \ldots, p$. Also let $Y = (Y_1, ..., Y_n)'$, $W(x_0) = diag(K_h(X_1 - x_0), ..., K_h(X_n - x_0))$, and

$$X_p(x_0) = \begin{pmatrix} (X_1 - x_0)^{Q_p} \\ \vdots \\ (X_n - x_0)^{Q_p} \end{pmatrix}.$$

Then the local polynomial regression estimator is defined as the solution to

$$\min_{\beta} \sum_{i=1}^{n} [Y_i - (X_i - x_0)^{Q_p}\beta]^2 K_h(X_i - x_0)$$

or, more compactly

$$\hat{\beta}_p(x_0) = \arg\min (Y - X_p(x_0)\beta)' W(x_0)(Y - X_p(x_0)\beta).$$

Clearly the estimator equals $[X_p'(x_0)W(x_0)X_p(x_0)]^{-1}X_p'(x_0)W(x_0)Y$ when the inverse exists. When $p=0$, the estimator is the kernel regression estimator and when $p=1$ the estimator is the local linear regression estimator.[23]

Let $H = diag(1,(\alpha h_n)^{-1}\iota_d,...,(\alpha h_n)^{-p}\iota_{(p+d-1)!/[p!(d-1)!]})$, where ι_s denotes a row vector of size s with 1 in all arguments. Then $\hat{\beta}_p(x_0) = H\hat{M}_{pn}(x_0)^{-1}n^{-1}H'X_p'(x_0)W(x_0)Y$, where $\hat{M}_{pn}(x_0) = n^{-1}H'X_p'(x_0)W(x_0)X_p(x_0)H$.

Note that by Taylor's expansion of order $\bar{p} \geq p$ at x_0, $m(X_i) = (X_i - x_0)^{Q_{\bar{p}}}\beta^*(x_0) + r_{\bar{p}}(X_i,x_0)$, where $r_{\bar{p}}(X_i,x_0) = (X_i - x_0)^{Q(\bar{p})}[m^{(\bar{p})}(\bar{x}_i) - m^{(\bar{p})}(x_0)]$ and $\bar{x}_i$ lies on the line between X_i and x_0. Write

$$m = (m(X_1),\ldots,m(X_n))',\ r_{\bar{p}}(x_0) = (r_{\bar{p}}(X_1,x_0),\ldots,r_{\bar{p}}(X_n,x_0))',\text{ and } \varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)'.$$

Let $M_{pn}(x_0)$ be the square matrix of size $\sum_{q=0}^{p}(q+d-1)!/[q!(d-1)!]$ denoting the expectation of $\hat{M}_{pn}(x_0)$, where the s-th row, t-th column "block" of $M_{pn}(x_0)$ matrix be, for $0 \leq s,t \leq p$,

$$E_x\{[((X-x_0)/(\alpha h_n))^{Q(s-1)}]'[((X-x_0)/(\alpha h_n))^{Q(t-1)}]K_h(X-x_0)\}.$$

Let $\lim_{n\to\infty} M_{pn}(x_0) = M_p \cdot f(x_0)^{p+1}$. Note that M_p only depends on $K(\cdot)$ when x_0 is an interior point of the support of X. Also write $\hat{I}_i = I\{X_i \in \hat{S}\}$ and $I_i = I\{X_i \in S\}$. We prove the following theorem.

Theorem 3 *Suppose Assumptions 1–4 hold. If M_p is non-singular, then the local polynomial regres-*

[23] Ruppert and Wand (1994) develop multivariate version of the local linear estimator. Masry (1995) develops multivariate general order local polynomial regression.

sion estimator of order $\bar{p}$, $\hat{m}_{\bar{p}}(x)$, *satisfies,*

$$[\hat{m}_{\bar{p}}(x_0) - m(x_0)]\hat{I}_0 = n^{-1}\sum_{i=1}^{n}\varepsilon_i K^*_{\bar{p}h}(X_i - x_0)I_0 + \hat{b}(x_0) + \hat{R}(x_0),$$

where $\hat{b}(x_0) = o(h_n^{\bar{p}})$, $n^{-1/2}\sum\limits_{i=1}^{n}\hat{R}(X_i) = o_p(1)$, *and*

$$K^*_{\bar{p}h}(X_i - x_0) = (1, 0, \ldots, 0)\cdot M_{\bar{p}n}(x_0)^{-1}\cdot[((X_i - x_0)/(\alpha_0 h_n))^{Q_{\bar{p}}}]' K_{h0}(X_i - x_0).$$

Furthermore, suppose that Assumptions 5–7 hold. Then the local polynomial regression estimator of order $0 \le p < \bar{p}$, $\hat{m}_p(x)$, *satisfies,*

$$[\hat{m}_p(x_0) - m(x_0)]\hat{I}_0 = n^{-1}\sum_{i=1}^{n}\varepsilon_i K^*_{ph}(X_i - x_0)I_0 + \hat{b}(x_0) + \hat{R}(x_0),$$

where

$$b_n(x_0) = (\alpha_0 h_n)^{\bar{p}}\mathbf{e_1}\cdot[M_{pn}(x_0)]^{-1}\sum_{s=p+1}^{\bar{p}}\Bigg[\Big[\int u^{Q(0)}\cdot u^{Q(s)}m^{(s)}(x_0)\cdot u^{Q(\bar{p}-s)}K(u)du,$$
$$\ldots, \int u^{Q(p)}\cdot u^{Q(s)}m^{(s)}(x_0)\cdot u^{Q(\bar{p}-s)}K(u)du\Big]f^{(\bar{p}-s)}(x_0)'\Bigg],$$

and $n^{-1/2}\sum\limits_{i=1}^{n}\hat{R}(X_i) = o_p(1)$. $\square$

Results similar to Theorem 3 have been proved by Fan (1993), Ruppert and Wand (1994), and Masry (1995). They prove pointwise or uniform convergence property of the estimator. As any other nonparametric estimator the convergence rate in this sense is slower than $n^{-1/2}$-rate. We prove that

the averaged pointwise residuals converges to zero faster than $n^{-1/2}$-rate.[24]

We only specify that M_p is nonsingular because one can find different conditions on $K(\cdot)$ to guarantee it. For example, assuming that $K(u_1, \ldots, u_d) = k(u_1) \cdots k(u_d)$, if $\int s^{2p} k(s) ds > 0$, then M_p is nonsingular.

To prove the theorem, note first that $Y = m + \varepsilon = X_{\bar{p}}(x_0)\beta^*_{\bar{p}}(x_0) + r_{\bar{p}}(X_i, x_0) + \varepsilon$. We wish to consider the situation where the order of polynomial terms included, p, is less than the underlying smoothness of the regression function, $\bar{p}$. For this purpose let $X_{\bar{p}}(x_0) = [X_p(x_0), \tilde{X}_{\bar{p}}(x_0)]$ and note that

$$[\hat{\beta}_p(x_0) - \beta^*_p(x_0)]\hat{I}_0 = H[\hat{M}_{pn}(x_0)]^{-1} n^{-1} H' X_p'(x_0) W(x_0) \varepsilon \cdot \hat{I}_0 \tag{A-3}$$

$$+ H[\hat{M}_{pn}(x_0)]^{-1} n^{-1} H' X_p'(x_0) W(x_0) \tilde{X}_{\bar{p}}(x_0) \tilde{\beta}^*_{\bar{p}}(x_0) \cdot \hat{I}_0 \tag{B-3}$$

$$+ H[\hat{M}_{pn}(x_0)]^{-1} n^{-1} H' X_p'(x_0) W(x_0) r_{\bar{p}}(x_0) \cdot \hat{I}_0, \tag{C-3}$$

where $\beta^*_{\bar{p}}(x_0) = (\beta^*_p(x_0)', \tilde{\beta}^*_{\bar{p}}(x_0)')'$. Note that if we use a p-th order polynomial when $m(x)$ is p-th order continuously differentiable, that is $p = \bar{p}$, then there is no second term (B-3).

Denote the first element of $\hat{\beta}(x_0)$ by $\hat{m}_p(x_0)$. We establish asymptotic linearity of $\hat{m}_p(x_0)$ and uniform consistency of its derivative. Lemma 2 shows that the first term (A-3) determines the asymptotic distribution, Lemma 7 shows that the second term (B-3) determines the bias term, and Lemma 8 shows that the third term (C-3) is of sufficiently small order to be negligible.

[24] Masry (1995) allows ρ-mixing with $\sum_{j=1}^{\infty} \rho(j) < \infty$ or strong mixing with $\sum_{j=1}^{\infty} j^a [\alpha(j)]^{1-2/\nu} < \infty$ for some $\nu > 2$ and $a > 1 - 2/\nu$. We only consider i.i.d. sampling.

Lemma 2 (Term (A-3)) *Under the assumptions of Theorem 3,*

$$\text{(A-3)} = \mathbf{e_1} \cdot [M_{pn}(x_0)]^{-1} n^{-1} H' X_p'(x_0) W(x_0) \varepsilon \cdot I_0 + \hat{R}_1(x_0)$$

where $\mathbf{e_1} = (1, 0, \ldots, 0)$ *and* $n^{-1/2} \sum_{i=1}^{n} \hat{R}_1(X_i) = o_p(1)$.

Proof. We first define neighborhoods of functions $\mathbf{e_1} \cdot [M_{pn}(x)]^{-1}$, $f_X(x)$, $I(x \in S)$, and point α_0. We denote them by Γ_n, $\mathcal{H}$, $\mathcal{I}$, and $\mathcal{A}$, respectively, where

$$\Gamma_n = \{\gamma_n(x); \sup_{x \in S} |\gamma_n(x) - \mathbf{e_1} \cdot [M_{pn}(x)]^{-1}| \leq \varepsilon_\gamma\},$$

for some small $\varepsilon_\gamma > 0$, $\mathcal{H} = \{f(x); \sup_{x \in S} |f(x) - f_X(x)| \leq \varepsilon_f\}$ for some small $\varepsilon_f > 0$,

$$\mathcal{I} = \{I(x \in \tilde{S}); \tilde{S} = \{x; f(x) \geq q_0\} \text{ for some } f(x) \in \mathcal{H}(x)\},$$

and $\mathcal{A} = [\alpha_0 - \delta_\alpha, \alpha_0 + \delta_\alpha]$, where $0 < \delta_\alpha < \alpha_0$.[25]

Using the neighborhoods we next define a class of functions $\mathcal{G}_{1n}$ as follows:

$$\mathcal{G}_{1n} = \left\{g_n; g_n(\varepsilon_i, X_i, X_j) = n^{-3/2} \cdot \gamma_n(X_j) \cdot [[(X_i - X_j)/(\alpha h_n)]^{Q_p}]' \varepsilon_i K_h(X_i - X_j) \cdot \tilde{I}_j\right\}$$

where it is indexed by a row vector-valued function $\gamma_n(x) \in \Gamma_n$, $\alpha \in \mathcal{A}$, which is also implicit in $K_h(\cdot)$,

[25] Note that a calculation using change of variables and the Lebesgue dominated convergence theorem shows that on S, $M_{pn}(x)^{-1}$ converges to a nonsingular matrix which only depends on $K(\cdot)$ times $[1/f(x)]^{p+1}$. Hence, on S, each element of $M_{pn}(x)^{-1}$ is uniformly bounded. Thus use of the sup-norm is justified.

and an indicator function $\tilde{I}_j \in \mathcal{I}$. Let $\gamma_{n0}(X_j) = \mathbf{e_1} \cdot M_{pn}(X_j)^{-1}$, $\hat{\gamma}_n(X_j) = \mathbf{e_1} \cdot \hat{M}_{pn}(X_j)^{-1}$,

$$g_{n0}(\varepsilon_i, X_i, X_j) = n^{-3/2} \cdot \gamma_{n0}(X_j) \cdot [[(X_i - X_j)/(\alpha_0 h_n)]^{Q_p}]' \varepsilon_i K_{h0}(X_i - X_j) \cdot I_j,$$

and

$$\hat{g}_n(\varepsilon_i, X_i, X_j) = n^{-3/2} \cdot \hat{\gamma}_n(X_j) \cdot [[(X_i - X_j)/(\hat{\alpha} h_n)]^{Q_p}]' \varepsilon_i \hat{K}_h(X_i - X_j) \cdot \hat{I}_j,$$

where we denote $K_{h0}(X_i - X_j) = (\alpha_0 h_n)^{-d} K((X_i - X_j)/(\alpha_0 h_n))$ and $\hat{K}_h(X_i - X_j) = (\hat{\alpha} h_n)^{-d} K((X_i - X_j)/(\hat{\alpha} h_n))$. Then since $\hat{R}_1(x_0) = \hat{g}_n(\varepsilon_i, X_i, x_0) - g_{n0}(\varepsilon_i, X_i, x_0)$, the result follows if two conditions are met: (1) equicontinuity of the process $\sum_{j=1}^n \sum_{i=1}^n g_n(\varepsilon_i, X_i, X_j)$ over $\mathcal{G}_{1n}$ in a neighborhood of $g_{n0}(\varepsilon_i, X_i, X_j)$ and (2) that, with probability approaching 1, $\hat{g}_n(\varepsilon_i, X_i, X_j)$ lies within the neighborhood over which we establish equicontinuity. We use the $\mathcal{L}^2$-norm to examine (1). We verify both of these two conditions in turn.

We verify the equicontinuity condition (1) using a lemma in Ichimura (1996).[26] We first define some notation in order to state the lemma. For $r = 1$ and 2, let $\mathcal{X}^r$ denote the r-fold product space of $\mathcal{X} \subset R^d$ and define a class of functions $\mathcal{F}_n$ defined over $\mathcal{X}^r$. For any $\psi_n \in \Psi_n$, write $\psi_{n,\mathbf{i_r}}$ as a short hand for either $\psi_n(x_i)$ or $\psi_n(x_{i_1}, x_{i_2})$, where $i_1 \neq i_2$. We define $U_n \psi_n = \sum_{\mathbf{i_r}} \psi_{n,\mathbf{i_r}}$, where $\sum_{\mathbf{i_r}}$ denotes the summation over all permutations of r elements of $\{x_1, \ldots, x_n\}$ for $r = 1$ or 2. Then $U_n \psi_n$ is called a U-process over $\psi_n \in \Psi_n$. For $r = 2$ we assume that $\psi_n(X_i, X_j) = \psi_n(X_j, X_i)$. Note that a normalizing constant is included as a part of ψ_n. A U-process is called degenerate if all conditional expectations given other elements are zero. When $r = 1$, this condition is defined so that $\mathrm{E}(\psi_n) = 0$.

[26] The result extends Nolan and Pollard (1987), Pollard (1990), Arcones and Giné (1993), and Sherman (1994) by considering U-statistics of general order $r \geq 1$ under niid sampling and allowing $\mathcal{F}$ to depend on n. When $\mathcal{F}$ depends on n, we need to assume condition (ii) in the lemma below as noted by Pollard (1990) when $r = 1$.

We assume that $\Psi_n \subset \mathcal{L}^2(\mathcal{P}^r)$, where $\mathcal{L}^2(\mathcal{P}^r)$ denotes the $\mathcal{L}^2$-space defined over $\mathcal{X}^r$ using the product measure of $\mathcal{P}$, $\mathcal{P}^r$. We denote the covering number using $\mathcal{L}^2$-norm, $\|\cdot\|_2$, by $N_2(\varepsilon, \mathcal{P}, \Psi_n)$.[27]

Lemma 3 (Equicontinuity) *Let $\{X_i\}_{i=1}^n$ be an iid sequence of random variables generated by $\mathcal{P}$. For a degenerate U-process $\{U_n\psi_n\}$ over a separable class of functions $\Psi_n \subset \mathcal{L}^2(\mathcal{P}^r)$ suppose the following assumptions hold: let $\|\psi_n\|_2 = [\sum_{\mathbf{i_r}} E\{\psi_{n,\mathbf{i_r}}\}]^{1/2}$.*

(i) *There exists an $F_n \in \mathcal{L}^2(\mathcal{P}^r)$ such that for any $\psi_n \in \Psi_n$, $|\psi_n| < F_n$ such that $\limsup_{n\to\infty} \sum_{\mathbf{i_r}} E\{F_{n,\mathbf{i_r}}^2\} < \infty$.*

(ii) *For each $\varepsilon > 0$, $\lim_{n\to\infty} \sum_{\mathbf{i_r}} E\{F_{n,\mathbf{i_r}}^2 1\{F_{n,\mathbf{i_r}} > \varepsilon\}\} = 0$.*

(iii) *There exists $\lambda(\varepsilon)$ and $\bar{\varepsilon} > 0$ such that for each $\varepsilon > 0$ less than $\bar{\varepsilon}$,*

$$\sup_{\mathcal{P}} N_2(\varepsilon, \mathcal{P}, \Psi_n) \leq \lambda(\varepsilon)$$

and $\int_0^{\bar{\varepsilon}} [\log \lambda(x)]^{r/2} dx < \infty$.

Then for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\lim_{n\to\infty} \Pr\{\sup_{\|\psi_{1n} - \psi_{2n}\|_2 \leq \delta} |U_n(\psi_{1n} - \psi_{2n})| > \varepsilon\} = 0. \quad \square$$

Following the literature we call a function $F_{n,\mathbf{i_r}}$ an envelope function of $\mathcal{F}_n$ if for any $\psi_n \in \mathcal{F}_n$, $\psi_{n,\mathbf{i_r}} \leq F_{n,\mathbf{i_r}}$ holds.

[27] For each $\varepsilon > 0$, the covering number $N_r(\varepsilon, \mathcal{P}, \mathcal{F})$ is the smallest value of m for which there exist functions $g_1, \ldots, g_m$ (not necessarily in $\mathcal{F}$) such that $\min_j [E\{|f - g_j|^r\}^{1/r} \leq \varepsilon$ for each f in $\mathcal{F}$. If such m does not exist then set the number to be ∞. When the sup-norm is used to measure the distance in calculating the covering number, we write $N_\infty(\varepsilon, \mathcal{F})$.

In order to apply the lemma to the process $\sum_{j=1}^{n}\sum_{i=1}^{n} g_n(\varepsilon_i, X_i, X_j)$ over $\mathcal{G}_{1n}$ in a neighborhood of $g_{n0}(\varepsilon_i, X_i, X_j)$, we first split the process into two parts; the process $\sum_{\mathbf{i_r}} g_n(\varepsilon_i, X_i, X_j) = \sum_{\mathbf{i_r}} g_n^0(\varepsilon_i, X_i, \varepsilon_j, X_j)$, where

$$g_n^0(\varepsilon_i, X_i, \varepsilon_j, X_j) = [g_n(\varepsilon_i, X_i, X_j) + g_n(\varepsilon_j, X_j, X_i)]/2$$

and the process $\sum_{i=1}^{n} g_n(\varepsilon_i, X_i, X_i)$. Note that $g_n(\varepsilon_i, X_i, X_i) = n^{-3/2} \cdot \gamma_n(X_i) \cdot \mathbf{e_1'} \cdot \varepsilon_i \cdot (\alpha h_n)^{-d} K(0) \cdot \tilde{I}_i$ is a order one process and has mean zero, hence it is a degenerate process. On the other hand $g_n^0(\varepsilon_i, X_i, \varepsilon_j, X_j)$ is a order two process and not degenerate, although it has mean zero and is symmetric. Instead of studying $g_n^0(\varepsilon_i, X_i, \varepsilon_j, X_j)$ we study a sum of degenerate U-processes following Hoeffding (1961).[28] Write $Z_i = (\varepsilon_i, X_i)$, $\phi_n(Z_i) = E\{g_n^0(Z_i, z)|Z_i\} = E\{g_n^0(z, Z_i)|Z_i\}$, and

$$\tilde{g}_n^0(Z_i, Z_j) = g_n^0(Z_i, Z_j) - \phi_n(Z_i) - \phi_n(Z_j).$$

Then

$$\sum_{\mathbf{i_r}} g_n^0(Z_i, Z_j) = \sum_{\mathbf{i_r}} \tilde{g}_n^0(Z_i, Z_j) + \sum_{i=1}^{n} 2 \cdot (n-1) \cdot \phi_n(Z_i),$$

where $\tilde{g}_n^0(Z_i, Z_j)$ and $2 \cdot (n-1) \cdot \phi_n(Z_i)$ are degenerate U-processes. Hence we study the three degenerate U-processes: $\tilde{g}_n^0(Z_i, Z_j)$, $2 \cdot (n-1) \cdot \phi_n(Z_i)$, and $g_n(\varepsilon_i, X_i, X_i)$, by verifying the three conditions stated in the equicontinuity lemma.

We start by verifying conditions (i) and (ii). An envelope function for $g_n^0(Z_i, Z_j)$ can be constructed by the sum of envelope functions for $g_n(\varepsilon_i, X_i, X_i)$ and $g_n(\varepsilon_i, X_i, X_j)$. Similarly an envelope function

[28] See also Serfling (1980).

for $\tilde{g}_n^0(Z_i, Z_j)$ can be constructed by the sum of envelope functions for $g_n^0(Z_i, Z_j)$ and $2 \cdot \phi_n(Z_i)$. Thus we only need to construct envelope functions that satisfy conditions (i) and (ii) for $g_n(\varepsilon_i, X_i, X_i)$, $g_n(\varepsilon_i, X_i, X_j)$, and $2 \cdot n \cdot \phi_n(Z_i)$.

Let $I_i^* = 1\{f_X(X_i) \geq q_0 - 2\varepsilon_f\}$ for some $q_0 > 2\varepsilon_f > 0$. Since $\sup_{x \in S} |f(x) - f_X(x)| \leq \varepsilon_f$, $I_i^* \geq \tilde{I}_i$ holds for any $\tilde{I}_i \in \mathcal{I}$. Also for any neighborhood of $M_{pn}(x)^{-1}$ defined by the sup-norm, there exists a $C > 0$ such that $|M_{pn}(x)^{-1}| \leq C$ so that $|g_n(\varepsilon_i, X_i, X_i)| \leq n^{-3/2} \cdot C \cdot |\varepsilon_i| \cdot [(\alpha_0 - \delta_\alpha)h_n]^{-d} K(0) \cdot I_i^*$ and the second moment of the right hand side times n is uniformly bounded over n since the second moment of ε_i is finite and $nh_n^d \to \infty$. Hence condition (i) holds. Condition (ii) holds by an application of Lebesgue dominated convergence theorem since $nh_n^d \to \infty$.

Note that any element of $[[(X_i - X_j)/(\alpha h_n)]^{Q_p}]' K_h(X_i - X_j)$ is bounded by $C_1 \cdot [(\alpha_0 - \delta_\alpha)h_n]^{-d} I\{\|X_i - X_j\| \leq C_2 \cdot h_n\}$ for some C_1 and C_2. Thus

$$|g_n(\varepsilon_i, X_i, X_j)| \leq n^{-3/2} \cdot |\varepsilon_i| \cdot C \cdot C_1 \cdot [(\alpha_0 - \delta_\alpha)h_n]^{-d} I\{\|X_i - X_j\| \leq C_2 \cdot h_n\} \cdot I_j^*.$$

Therefore, analogous to the previous derivation, conditions (i) and (ii) hold for $g_n(\varepsilon_i, X_i, X_j)$.

Note further that since the density of x is bounded on S, by some constant, say, $C_3 > 0$, $|\phi_n(\varepsilon_i, X_i)| \leq n^{-3/2} \cdot |\varepsilon_i| \cdot C \cdot C_1 \cdot C_3$ and hence $2 \cdot n \cdot |\phi_n(\varepsilon_i, X_i)|$ has an envelope function $n^{-1/2} 2 \cdot |\varepsilon_i| \cdot C \cdot C_1 \cdot C_3$ that satisfies the two conditions.

To verify condition (iii), first note the following. Write $J_h(X_i - X_j) = [[(X_i - X_j)/(\alpha h_n)]^{Q_p}]' K_h(X_i -$

$X_j)$ and $J_{h0}(X_i - X_j) = [[(X_i - X_j)/(\alpha_0 h_n)]^{Q_p}]' K_{h0}(X_i - X_j)$. Using this notation

$$
\begin{aligned}
&|g_n(\varepsilon_i, X_i, X_j) - g_{n0}(\varepsilon_i, X_i, X_j)| \qquad \text{(L-3)}\\
&= n^{-3/2}|\varepsilon_i| \cdot |\gamma_n(X_j) \cdot J_h(X_i - X_j) \cdot \tilde{I}_j - \gamma_{n0}(X_j) \cdot J_{h0}(X_i - X_j) \cdot I_j|\\
&\leq n^{-3/2}|\varepsilon_i| \cdot |\gamma_n(X_j) - \gamma_{n0}(X_j)| \cdot C_1 \cdot [(\alpha_0 - \delta_\alpha)h_n]^{-d} I\{\|X_i - X_j\| \leq C_2 \cdot h_n\} \cdot I_j^*\\
&+ n^{-3/2}|\varepsilon_i| \cdot |\gamma_{n0}(X_j)| \cdot |J_h(X_i - X_j) - J_{h0}(X_i - X_j)| \cdot I_j^*\\
&+ n^{-3/2}|\varepsilon_i| \cdot |\gamma_{n0}(X_j)| \cdot |J_{h0}(X_i - X_j)| \cdot |\tilde{I}_j - I_j|.
\end{aligned}
$$

For $g_n(\varepsilon_i, X_i, X_i)$ right hand side is bounded by some $C > 0$,

$$
\begin{aligned}
n^{-3/2} h_n^{-d} |\varepsilon_i| \cdot C \cdot \Big[&|\gamma_n(X_i) - \gamma_{n0}(X_i)| \cdot (\alpha_0 - \delta_\alpha)^{-d} I_i^*\\
&+ |\gamma_{n0}(X_i)| \cdot |\alpha^{-d} - \alpha_0^{-d}| \cdot I_i^* + |\gamma_{n0}(X_i)| \cdot |\tilde{I}_i - I_i|] \Big].
\end{aligned}
$$

Since $nh_n^d \to \infty$, the $\mathcal{L}_2$-covering number for the class of functions denoted by $g_n(\varepsilon_i, X_i, X_i)$ for $g_n \in \mathcal{G}_{1n}$ can be bounded above by the product of the covering numbers of Γ_n, $\mathcal{A}$, and $\mathcal{I}$. Since it is the log of the covering number that needs to be integrable, if each of three spaces satisfy condition (iii), then this class will also. Clearly $\mathcal{A}$ satisfies condition (iii). To see that Γ_n and $\mathcal{I}$ do also, we use the following result by Kolmogorov and Tihomirov (1961).[29] First they define a class of functions for which the upper bound of the covering number is obtained.

Definition 2 *A function in $\Psi(K)$ has smoothness $q > 0$, where $q = p + \alpha$ with integer p and $0 < \alpha \leq 1$,*

[29] See pp.308–314. Kolmogorov and Tihomirov present their result using the concept of packing number instead of covering number.

if for any $x \in K$ and $x + h \in K$, we have

$$\psi(x+h) = \sum_{k=0}^{p} (k!)^{-1} B_k(h,x) + R_\psi(h,x),$$

where $B_k(h,x)$ is a homogenous form of degree k in h and $|R_\psi(h,x)| \leq C\|h\|^q$, where C is a constant.

□

Let $\Psi_q^K(C) = \{\psi \in \Psi(K) : |R_\psi(h,x)| \leq C\|h\|^q\}$.

If a function defined on K is p-times continuously differentiable and the p-th derivative satisfies Hölder continuity with the exponent $0 < \alpha \leq 1$, then a Taylor expansion shows that the function belongs to $\Psi_q^K(C)$ for some C, where $q = p + \alpha$.

Lemma 4 (K-T) *For every set $A \subset \Psi_q^K(C)$, where $K \subset R^d$, we have, for $0 < d, q < \infty$,*

$$\log_2 N_\infty(\varepsilon, A) \leq L(d,q,C,K)(1/\varepsilon)^{d/q}$$

for some constant $L(d,q,C,K) > 0$. □

Hence, because $d/q < 1$, condition (iii) holds for Γ_n and $\mathcal{I}$. Analogously we can verify condition (iii) for the remaining U-processes. Hence all three processes are equicontinuous.

The remaining task is to verify that $\hat{g}_n(\varepsilon_i, X_i, X_j)$ lies in the neighborhood of $g_{n0}(\varepsilon_i, X_i, X_j)$ over which we showed equicontinuity. By the inequality in Lemma 3 (L-3), this follows from Assumptions 3 and 5, and by verifying that almost surely

$$\sup_{x \in S, \alpha \in \mathcal{A}} ||\hat{M}_{pn}(x) - M_{pn}(x)|| \to 0,$$

where $\lim_{n\to\infty} \inf_{x\in S} \det(M_{pn}(x)) > 0$.[30] The latter follows directly from the nonsingularity of matrix M_p and the trimming rule. Hence the following lemma completes the proof. ■

Lemma 5 *Under the assumptions of Theorem 3, almost surely,*

$$\sup_{x\in S,\alpha\in\mathcal{A}} ||\hat{M}_{pn}(x) - M_{pn}(x)|| \to 0. \quad \square$$

Proof. Note that any element of the matrix difference $\hat{M}_{pn}(x) - M_{pn}(x)$ has the form

$$n^{-1}\sum_{i=1}^{n}\left[(\alpha h_n)^{-d}J((X_i - x_0)/(\alpha h_n)) - E\left\{(\alpha h_n)^{-d}J((X_i - x_0)/(\alpha h_n))\right\}\right],$$

where the kernel function $J(s) = s^q \cdot s^r K(s)$ for some vectors q and r whose elements are integers and they sum to p or less nonnegative integers, where q and r depend on which element we look at. To construct a proof, we use the following lemma of Pollard (1984).[31]

Lemma 6 (Pollard) *For each n, let Ψ_n be a separable class of functions whose covering numbers satisfy*

$$\sup_{\mathcal{P}} N_1(\varepsilon, \mathcal{P}, \Psi_n) \le A\varepsilon^{-W} \quad \textit{for} \quad 0 < \varepsilon < 1$$

with constants A and W not depending on n. Let $\{\xi_n\}$ be a non-increasing sequence of positive numbers for which $\lim_{n\to\infty} n\zeta_n^2\xi_n^2/\log n = \infty$. If $|\psi| \le 1$ and $(E\{\psi^2\})^{1/2} \le \zeta_n$ for each ψ in Ψ_n, then, almost surely,

$$\sup_{\Psi_n} |n^{-1}\sum_{i=1}^{n}[\psi(X_i) - E\{\psi(X_i)\}]|/(\zeta_n^2\xi_n) \to 0. \quad \square$$

[30] Recall from the discussion of Theorem 3 that $\hat{M}_{pn}(x)$ depends on the parameter α.

[31] His lemma only requires that $\mathcal{F}_n$ is a permissible class of functions. In our applications $\mathcal{F}_n$ is always separable. His Example 38, pp.35–36, gives a similar result. We provide a proof here for completeness and for later reference.

To use this lemma, we need to calculate $N_1(\varepsilon, \mathcal{P}, \Psi_n)$. Let $C_1 \geq \sup_{s \in R} J(s)$, C_2 is a Lipschitz constant for J, and C_3 is a number greater than the radius of a set that includes the support of J. In our application, recall from our proof of Theorem 3 that $\mathcal{A} = [\alpha_0 - \delta_\alpha, \alpha_0 + \delta_\alpha]$, where $0 < \delta_\alpha < \alpha_0$ so that

$$
\begin{aligned}
& E\{|\alpha_1^{-d} J((x - x_{01})/(\alpha_1 h_n)) - \alpha_2^{-d} J((x - x_{02})/(\alpha_2 h_n))|\} \\
\leq \; & |\alpha_1^{-d} - \alpha_2^{-d}| \cdot E\{|J((x - x_{01})/(\alpha_1 h_n))|\} \\
& + \alpha_2^{-d} \cdot E\{|J((x - x_{01})/(\alpha_1 h_n)) - J((x - x_{02})/(\alpha_2 h_n))|\} \\
\leq \; & |\alpha_1^{-d} - \alpha_2^{-d}| \cdot C_1 \\
& + (\alpha_0 - \delta_\alpha)^{-d} \cdot [C_2 \cdot (1 - \alpha_1/\alpha_2) \cdot [(\alpha_0 + \delta_\alpha)/(\alpha_0 - \delta_\alpha)] \cdot C_3\} \cdot \bar{h} \\
& + C_2 \cdot \|x_{01} - x_{02}\| / (\alpha_0 - \delta_\alpha)].
\end{aligned}
$$

The upper bound of the right hand side does not depend on $\mathcal{P}$. Moreover, the right hand side can be made less than $\varepsilon \cdot C$ for some $C > 0$ by choosing $|\alpha_1 - \alpha_2| \leq \varepsilon$ and $|x_{01} - x_{02}| \leq \varepsilon$. Since S and $\mathcal{A}$ are both bounded subsets of a finite dimensional Euclidean space, the uniform covering number condition holds. To complete the proof of Lemma 5, note that we are free to choose $\xi_n = 1$ and $\zeta_n = C \cdot h_n^{d/2}$ in our application of the lemma. ■

Next we examine the second term (B-3).

Lemma 7 (Term (B-3)) *Under the assumptions of Theorem 3,*

$$
\text{(B-3)} = b(x_0) + \hat{R}_2(x_0),
$$

where

$$b_n(x_0) = (\alpha_0 h_n)^{\bar{p}} \mathbf{e_1} \cdot [M_p(x_0)]^{-1} \sum_{s=p+1}^{\bar{p}} \Big[[\int u^{Q(0)} \cdot u^{Q(s)} m^{(s)}(x_0) \cdot u^{Q(\bar{p}-s)} K(u) du,$$
$$\ldots, \int u^{Q(p)} \cdot u^{Q(s)} m^{(s)}(x_0) \cdot u^{Q(\bar{p}-s)} K(u) du] f^{(\bar{p}-s)}(x_0)' \Big],$$

$n^{-1/2} \sum_{i=1}^n \hat{R}_2(X_i) = o_p(1)$, *and* $\hat{R}_2(x_0)$ *is defined as the difference between Term (B-3) and* $b_n(x_0)$.

Proof. Note that

$$\begin{aligned}
\text{(B-3)} &= \mathbf{e_1} \cdot [\hat{M}_{pn}(x_0)]^{-1} n^{-1} H' X_p'(x_0) W(x_0) \tilde{X}_{\bar{p}}(x_0) \tilde{\beta}^*_{\bar{p}}(x_0) \cdot \hat{I}_0 \\
&= \mathbf{e_1} \cdot [\hat{M}_{pn}(x_0)]^{-1} \sum_{s=p+1}^{\bar{p}} n^{-1} \sum_{i=1}^n [[(X_i - x_0)/(\alpha h_n)]^{Q_p}]'(X_i - x_0)^{Q(s)} m^{(s)}(x_0) K_h(X_i - x_0) \cdot \hat{I}_0 \\
&= \mathbf{e_1} \cdot [\hat{M}_{pn}(x_0)]^{-1} \sum_{s=p+1}^{\bar{p}} n^{-1} \sum_{i=1}^n \Big[[[(X_i - x_0)/(\alpha h_n)]^{Q_p}]'(X_i - x_0)^{Q(s)} K_h(X_i - x_0) \\
&\quad - E\{[[(X_i - x_0)/(\alpha h_n)]^{Q_p}]'(X_i - x_0)^{Q(s)} K_h(X_i - x_0) | x_0\} \Big] m^{(s)}(x_0) \cdot \hat{I}_0 \qquad \text{(L-7A)} \\
&\quad + \mathbf{e_1} \cdot [\hat{M}_{pn}(x_0)]^{-1} \sum_{s=p+1}^{\bar{p}} E\{[[(X_i - x_0)/(\alpha h_n)]^{Q_p}]'(X_i - x_0)^{Q(s)} K_h(X_i - x_0) | x_0\} m^{(s)}(x_0) \cdot \hat{I}_0. \\
&\qquad \text{(L-7B)}
\end{aligned}$$

Define term (L-7A) as $\hat{R}_{21}(x_0)$. We apply the same method as in Lemma 2 to show that $n^{-1/2} \sum_{i=1}^n \hat{R}_{21}(X_i) = o_p(1)$. Instead of $\mathcal{G}_{1n}$, define the class of functions $\mathcal{G}_{2n}$ as follows:

$$\begin{aligned}
\mathcal{G}_{2n} = \{ g_n; g_n(X_i, X_j) = n^{-3/2} \cdot \gamma_n(X_j) \cdot \Big[&[[(X_i - X_j)/(\alpha h_n)]^{Q_p}]'(X_i - X_j)^{Q(s)} K_h(X_i - X_j) \\
&- E\{[[(X_i - X_j)/(\alpha h_n)]^{Q_p}]'(X_i - X_j)^{Q(s)} K_h(X_i - X_j) | X_j\} \Big] m^{(s)}(X_j) \cdot \tilde{I}_j
\end{aligned}$$

where it is indexed by a row vector-valued function $\gamma_n(x) \in \Gamma_n$, $\alpha \in \mathcal{A}$, which is also implicit in $K_h(\cdot)$,

and an indicator function $\tilde{I}_j \in \mathcal{I}$. Let

$$g_{n0}(X_i, X_j) = n^{-3/2} \cdot \gamma_{n0}(X_j) \cdot [[(X_i - X_j)/(\alpha_0 h_n)]^{Q_p}]'(X_i - X_j)^{Q(s)} m^{(s)}(X_j) K_{h0}(X_i - X_j) \cdot I_j,$$

and

$$\hat{g}_n(X_i, X_j) = n^{-3/2} \cdot \hat{\gamma}_n(X_j) \cdot [[(X_i - X_j)/(\hat{\alpha} h_n)]^{Q_p}]'(X_i - X_j)^{Q(s)} m^{(s)}(X_j) \hat{K}_h(X_i - X_j) \cdot \hat{I}_j.$$

To prove that term (L-7B) equals $b_n(x_0) + o(h_n^{\bar{p}})$ we use the assumption that all the moments of $K(\cdot)$ of order $p+1$ and higher up to $\bar{p}$ have mean zero, that x_0 is an interior point of the support of x that is more than $(\alpha_0 + \delta)h_n C$ interior to the closest edge of the support, where C is the radius of the support of $K(\cdot)$, and the assumption that the density of x is $\bar{p}$-times continuously differentiable and its $\bar{p}$-th derivative satisfies a Hölder condition. Using a change of variable calculation and the Lebesgue dominated convergence theorem, and Lemma 5 the result follows. ■

Note that this is the term that converges to zero with the specified rate only if x_0 is an interior point. Thus for kernel regression estimators for higher dimensions, we need to introduce a special trimming method to guarantee this. Use of higher order local polynomial regression alleviates the problem for higher dimensional problems, but at the price of requiring more data locally.

Lemma 8 (Term (C-3)) *Under the assumptions of Theorem 3,*

$$\text{(C-3)} = o_p(h_n^{\bar{p}}).$$

Proof. Recall that the third term equals $\mathbf{e_1} \cdot [\hat{M}_{pn}(x_0)]^{-1} n^{-1} H' X_p'(x_0) W(x_0) r_{\bar{p}}(x_0) \cdot \hat{I}_0$. Note that

$$\begin{aligned}
& n^{-1} \left\| H' X_p'(x_0) W(x_0) r_{\bar{p}}(x_0) \cdot \hat{I}_0 \right\| \\
= \; & n^{-1} \left\| \sum_{i=1}^{n} [(X_i - x_0)/(\alpha h_n)]^{Q_p} [(X_i - x_0)/h_n]^{Q(\bar{p})} [m^{(\bar{p})}(\bar{x}_i) - m^{(\bar{p})}(x_0)] K_h(x_0 - X_i) h_n^{\bar{p}} \right\| \\
\leq \; & n^{-1} \sum_{i=1}^{n} \left\| [(X_i - x_0)/(\alpha h_n)]^{Q_p} | [(X_i - x_0)/h_n]^{Q(\bar{p})} | \cdot |K_h(x_0 - X_i)| \right\| o(h_n^{\bar{p}}) \\
= \; & o_p(h_n^{\bar{p}}),
\end{aligned}$$

where the inequality follows from the Hölder condition on $m^{(\bar{p})}(x_0)$ and the compact support condition on $K(\cdot)$, and the last equality follows from the same reasoning used to prove Lemma 5. The conclusion follows from Lemma 5 and the assumption of nonsingularity of $M_p(x_0)$. ■

A.2 Verifying the Assumptions of Lemma 1

Five assumptions in Lemma 1 are as follows:

(i) *Both $\hat{P}(z)$ and $\hat{g}(t,p)$ are asymptotically linear with trimming where*

$$[\hat{P}(z) - P(z)] I(x \in \hat{S}) = n^{-1} \sum_{j=1}^{n} \psi_{np}(D_j, Z_j; z) + \hat{b}_p(z) + \hat{R}_p(z),$$
$$[\hat{g}(t,p) - g(t,p)] I(x \in \hat{S}) = n^{-1} \sum_{j=1}^{n} \psi_{ng}(Y_j, T_j, P(Z_j); t, p) + \hat{b}_g(t,p) + \hat{R}_g(t,p).$$

(ii) *$\partial \hat{g}(t,p)/\partial p$ and $\hat{P}(z)$ are uniformly consistent to $\partial g(t,p)/\partial p$ and $P(z)$, respectively, and that $\partial g(t,p)/\partial p$ is continuous for all t and p.*

(iii) *$plim_{n\to\infty} n^{-1/2} \sum_{i=1}^{n} \hat{b}_g(T_i, X_i) = b_g$ and $plim_{n\to\infty} n^{-1/2} \sum_{i=1}^{n} \partial g(T_i, P(Z_i))/\partial p \cdot \hat{b}_p(T_i, P(Z_i)) = b_{g_p}$,*

(iv) $plim_{n\to\infty} n^{-1/2} \sum_{i=1}^{n} [\partial \hat{g}(T_i, \bar{P}_{T_i}(Z_i))/\partial p - \partial g(T_i, P(Z_i))/\partial p] \cdot \hat{R}_p(Z_i) = 0,$

(v) $plim_{n\to\infty} n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{n} [\partial \hat{g}(T_i, \bar{P}_{T_i}(Z_i))/\partial p - \partial g(T_i, P(Z_i))/\partial p] \cdot \psi_{np}(D_j, Z_j; Z_i) = 0.$

We verify these assumptions for the local polynomial regression estimator. Condition (i) is just shown. We next show that the derivative of the local polynomial regression estimator converges uniformly to the derivative of the limit of the local polynomial regression estimator. To our knowledge, our investigation of the derivatives of the local polynomial regression estimator is new.

Theorem 4 *If Assumptions 1–4, and 8 hold, then $\partial \hat{g}(t,p)/\partial p$ is uniformly consistent for $\partial g(t,p)/\partial p$.*

Proof. For convenience we drop the subscripts p, $\bar{p}$, and the argument x_0 of $X_p(x_0)$, $\hat{\beta}_p(x_0)$, $\beta_p^*(x_0)$, and $W(x_0)$ here so that $X = X_p(x_0)$, $\hat{\beta} = \hat{\beta}_p(x_0)$, $\beta^* = \beta_p^*(x_0)$, and $W = W(x_0)$. Also denote the derivative with respect to the first argument of x_0 by ∇. Note that $X'WY = X'WX\hat{\beta}$. Hence by the chain rule,

$$\begin{aligned}
\nabla\hat{\beta} &= (X'WX)^{-1}\{\nabla(X'W)Y - [\nabla(X'WX)]\hat{\beta}\} \\
&= (X'WX)^{-1}\{\nabla(X'W) - [\nabla(X'WX)](X'WX)^{-1}X'W\}Y \\
&= (X'WX)^{-1}\{\nabla(X'W) - [\nabla(X'WX)](X'WX)^{-1}X'W\}(X\beta^* + r + \varepsilon).
\end{aligned}$$

Since

$$\begin{aligned}
& \{\nabla(X'W) - [\nabla(X'WX)](X'WX)^{-1}X'W\}X\beta^* \\
= \; & \nabla(X'W)X\beta^* - \nabla(X'WX)\beta^* \\
= \; & \{(\nabla X')WX + X'(\nabla W)X - (\nabla X')WX - X'(\nabla W)X - X'W(\nabla X)\}\beta^* \\
= \; & -X'W(\nabla X)\beta^*,
\end{aligned}$$

we obtain

$$\begin{aligned}
\nabla\hat{\beta} \;=\; & -(X'WX)^{-1}(X'W\nabla X)\beta^* \\
& +(X'WX)^{-1}\{\nabla(X'W) - [\nabla(X'WX)](X'WX)^{-1}X'W\}r \\
& +(X'WX)^{-1}\{\nabla(X'W) - [\nabla(X'WX)](X'WX)^{-1}X'W\}\varepsilon.
\end{aligned}$$

Note that for $s \geq 1$,

$$\begin{aligned}
\nabla(x-x_0)^{Q(s)} &= (\nabla(x-x_0)^{(q_1,...,q_d)})_{q_1+\cdots+q_d=s} \\
&= \left((x-x_0)^{(q_1,...,q_d)}1(q_1 \geq 1)\right)_{q_1+\cdots+q_d=s-1} \\
&= \left[\left((x-x_0)^{(q_1,...,q_d)}\right)_{q_1+\cdots+q_d=s-1} 0,...,0\right],
\end{aligned}$$

where the second equality follows from our convention on the order of the elements.

Thus

$$\begin{aligned} X &= \begin{pmatrix} 1 & (x-x_0)^{Q(1)} & (x-x_0)^{Q(2)} & \dots & (x-x_0)^{Q(p)} \end{pmatrix} \\ \nabla X &= -\begin{pmatrix} 0 & 1 & 0 & \dots & 0 & (x-x_0)^{Q(1)} & 0 & \dots & 0 & \dots & (x-x_0)^{Q(p-1)} & 0 & \dots & 0 \end{pmatrix}. \end{aligned}$$

Note that each column of ∇X is either a column of X or the column with all elements being 0. Hence there exists a matrix J such that $\nabla X = -XJ$, where J is a matrix that selects appropriate column of X or the zero column. Without being more specific about the exact form of J we can see that $-(X'WX)^{-1}(X'W\nabla X)\beta^* = J\beta^*$, and also that

$$\begin{aligned} & (X'WX)^{-1}\{\nabla(X'W) - [\nabla(X'WX)](X'WX)^{-1}X'W\} \\ = \; & (X'WX)^{-1}\{(\nabla X')W + X'(\nabla W) \\ & -(\nabla X')WX(X'WX)^{-1}X'W - X'(\nabla W)X(X'WX)^{-1}X'W - X'W(\nabla X)(X'WX)^{-1}X'W\} \\ = \; & (X'WX)^{-1}\{-J'X'W + X'(\nabla W) \\ & +J'X'W - X'(\nabla W)X(X'WX)^{-1}X'W + X'WXJ(X'WX)^{-1}X'W\} \\ = \; & (X'WX)^{-1}\{X'(\nabla W) - [X'(\nabla W)X](X'WX)^{-1}X'W\} + J(X'WX)^{-1}X'W\}. \end{aligned}$$

Next, to simplify the last expression, we use some specific properties of matrix J. The key properties of J we use are that all elements of the first column are 0 and that the first element of the second column of J is 1. That is, the first column of AJ is always the 0-vector, regardless of A and the second column of AJ is the first column of A. Since the first column is chosen by J exactly once, the preceding observations also imply that the first row of J is $\mathbf{e_2}$, where $\mathbf{e_2} = (0, 1, 0, \dots, 0)$ if $p \geq 1$ and

0, if $p = 0$. Therefore

$$\begin{aligned}\mathbf{e_1} \cdot \nabla\hat{\beta} &= \nabla\hat{m}_p(x) = \mathbf{e_1} \cdot J\beta^* \\ &+ \Big(\mathbf{e_1} \cdot (X'WX)^{-1}\{(X'(\nabla W) - [X'(\nabla W)X](X'WX)^{-1}X'W\} - \mathbf{e_2} \cdot X'W\Big)\, r \\ &+ \Big(\mathbf{e_1} \cdot (X'WX)^{-1}\{(X'(\nabla W) - [X'(\nabla W)X](X'WX)^{-1}X'W\} - \mathbf{e_2} \cdot X'W\Big)\, \varepsilon,\end{aligned}$$

where $\mathbf{e_1} \cdot J\beta^* = \nabla m(x)$. That the remaining two terms converge uniformly to zero can be shown analogously as in Lemma 5. ■

Condition (iii) of Lemma 1 clearly holds under an i.i.d. assumption given the bias function obtained in Theorem 3. In order to verify condition (iv) of the lemma we need to go back to the defintion of the residual terms and then use the same equicontinuity argument.

Finally condition (v) can be verified by invoking the equicontinuity lemma. This is where the additional smoothness condition is required. ■

Armed with these results, we finally turn to the proof of key Theorem 2.

A.3 Proof of Theorem 2.

Note first that, writing $\hat{I}_i = I(X_i \in \hat{S})$,

$$N_1^{1/2}\left[\frac{N_1^{-1}\sum\limits_{i\in\mathbf{I_1}}[Y_{1i} - \hat{g}(X_i)]\hat{I}_i}{N_1^{-1}\sum\limits_{i\in\mathbf{I_1}}\hat{I}_i} - E_S(Y_1 - Y_0 \mid D = 1)\right] = \frac{N_1^{-1/2}\sum\limits_{i\in\mathbf{I_1}}[Y_{1i} - \hat{g}(X_i) - E_S(Y_1 - Y_0 \mid D = 1)]\hat{I}_i}{N_1^{-1}\sum\limits_{i\in\mathbf{I_1}}\hat{I}_i}. \quad \text{(T-1)}$$

We first consider the numerator and then turn to the denominator of the expression. Note that the numerator of the right-hand side of (T-1) is the sum of three terms (TR-1)–(TR-3): writing $g_1(x) = E_S(Y_1|D = 1, X = x)$,

$$\text{(TR-1)} = N_1^{-1/2} \sum_{i \in \mathbf{I_1}} [Y_{1i} - g_1(X_i)]\hat{I}_i,$$

$$\text{(TR-2)} = N_1^{-1/2} \sum_{i \in \mathbf{I_1}} \{[g_1(X_i) - g(X_i)] - E_S(Y_1 - Y_0|D = 1)\}\hat{I}_i,$$

$$\text{(TR-3)} = N_1^{-1/2} \sum_{i \in \mathbf{I_1}} [g(X_i) - \hat{g}(X_i)]\hat{I}_i.$$

Terms (TR-1) and (TR-2) are analogous to terms we examined in Theorem 1. Term (TR-3) is the additional term that arises from estimating $g(x)$. However, the first two terms from Theorem 1 have to be modified to allow for the trimming function introduced to control the impact of the estimation error of $\hat{g}(x)$.

Central limit theorems do not apply directly to the sums in (TR-1) and (TR-2) because the trimming function depends on all data and this creates correlation across all i. Instead, writing $I_i = I(X_i \in S)$, we show that these terms can be written as

$$\text{(TR-1)} = N_1^{-1/2} \sum_{i \in \mathbf{I_1}} [Y_{1i} - g_1(X_i)]I_i + o_p(1)$$

and

$$\text{(TR-2)} = N_1^{-1/2} \sum_{i \in \mathbf{I_1}} \{[g_1(X_i) - g(X_i)] - E_S(Y_1 - Y_0|D = 1)\}I_i + o_p(1),$$

respectively. One can use the equicontinuity lemma and our assumption of p-nice trimming to show the result for term (TR-1). The same method does not apply for term (TR-2), however. This is

because when we take an indicator function $\tilde{I}_i$ from $\mathcal{I}$, where $\tilde{I}_i \neq I_i$, then

$$E[\{[g_1(X_i) - g(X_i)] - E_S(Y_1 - Y_0|D = 1)\}\tilde{I}_i] \neq 0.$$

It is necessary to recenter this expression to adjust for the bias that arises from using $\tilde{I}_i$.

In order to achieve this we observe that, writing $\Delta_S(X_i) = [g_1(X_i) - g(X_i)] - E_S(Y_1 - Y_0|D = 1)$,

$$\begin{aligned}
&N_1^{-1/2} \sum_{i \in \mathbf{I_1}} \Delta_S(X_i) \cdot \hat{I}_i = N_1^{-1/2} \sum_{i \in \mathbf{I_1}} \Delta_S(X_i) \cdot I_i \\
&+ N_1^{-1/2} \sum_{i \in \mathbf{I_1}} \Delta_S(X_i) \cdot [\hat{\sigma}(X_i)]^{-1} \cdot \tilde{K}_- \left(\frac{f(X_i) - q_0}{\hat{\sigma}(X_i)} \right) [\hat{f}(X_i) - f(X_i)] I\{\hat{f}(X_i) > f(X_i)\} \\
&+ N_1^{-1/2} \sum_{i \in \mathbf{I_1}} \Delta_S(X_i) \cdot [\hat{\sigma}(X_i)]^{-1} \cdot \tilde{K}_+ \left(\frac{f(X_i) - q_0}{\hat{\sigma}(X_i)} \right) [\hat{f}(X_i) - f(X_i)] I\{\hat{f}(X_i) \leq f(X_i)\},
\end{aligned}$$

where $\hat{\sigma}(X_i) = |\hat{f}(X_i) - f(X_i)|$, $\tilde{K}_-(s) = 1$ if $-1 \leq s < 0$ and 0 otherwise, and $\tilde{K}_+(s) = 1$ if $0 \leq s < 1$ and 0 otherwise. Since $\hat{f}(X_i) = (N_0 a_n)^{-d} \sum_{j \in \mathbf{I_0}} K\left((X_j - X_i)/a_n\right)$, the latter two terms can be expressed as double sums. We then apply an equicontinuity argument to the expressions

$$\begin{aligned}
&N_1^{-1/2} (N_0 a_n)^{-d} \sum_{i \in \mathbf{I_1}} \sum_{j \in \mathbf{I_0}} \Delta_S(X_i) \cdot [\hat{\sigma}(X_i)]^{-1} \cdot \tilde{K}_- \left(\frac{f(X_i) - q_0}{\hat{\sigma}(X_i)} \right) \\
&\qquad \times \left[K\left((X_j - X_i)/a_n\right) - E\left\{ K\left((X_j - X_i)/a_n\right) | X_i \right\} \right] I\{\hat{f}(X_i) > f(X_i)\}
\end{aligned}$$

and

$$\begin{aligned}
&N_1^{-1/2} (N_0 a_n)^{-d} \sum_{i \in \mathbf{I_1}} \sum_{j \in \mathbf{I_0}} \Delta_S(X_i) \cdot [\hat{\sigma}(X_i)]^{-1} \cdot \tilde{K}_+ \left(\frac{f(X_i) - q_0}{\hat{\sigma}(X_i)} \right) \\
&\qquad \times \left[K\left((X_j - X_i)/a_n\right) - E\left\{ K\left((X_j - X_i)/a_n\right) | X_i \right\} \right] I\{\hat{f}(X_i) \leq f(X_i)\}
\end{aligned}$$

and control the bias by $\bar{p}$-smoothness of $f(X_i)$.

Finally, term (TR-3) can be written as the sum of three terms

$$N_1^{-1/2} N_0^{-1} \sum_{i \in \mathbf{I_1}} \sum_{j \in \mathbf{I_0}} \psi_{0N_0N_1}(Y_{0j}, X_j; X_i) + N_1^{-3/2} \sum_{i \in \mathbf{I_1}} \sum_{j \in \mathbf{I_1}} \psi_{1N_0N_1}(Y_{0j}, X_j; X_i), \tag{TR-3-1}$$

$$N_1^{-1/2} \sum_{i \in \mathbf{I_1}} \hat{b}_g(X_i), \tag{TR-3-2}$$

and

$$N_1^{-1/2} \sum_{i \in \mathbf{I_1}} \hat{R}_g(X_i). \tag{TR-3-3}$$

Terms (TR-3-2) and (TR-3-3) are $o_p(1)$ by the definition of asymptotic linearity of $\hat{g}(X_i)$. Term (TR-3-1) is a U-statistic and a central limit theorem can be obtained for it using the lemmas of Hoeffding (1948) and Powell, Stock, and Stoker (1989) and a two-sample extension of the projection lemma as in Serfling (1980). We first present the Hoeffding, Powell, Stock, and Stoker result:

Lemma 9 (H-P-S-S) *Suppose* $\{Z_i\}_{i=1}^n$ *is i.i.d.,* $U_n\psi_n = (n \cdot (n-1))^{-1} \sum_{\mathbf{i_r}} \psi_n(Z_i, Z_j)$, *where* $\psi_n(Z_i, Z_j) = \psi_n(Z_j, Z_i)$ *and* $E\{\psi_n(Z_i, Z_j)\} = 0$, *and* $\hat{U}_n\psi_n = n^{-1} \sum_{i=1}^n 2 \cdot p_n(Z_i)$, *where* $p_n(Z_i) = E\{\psi_n(Z_i, Z_j)|Z_i\}$. *If* $E\{\psi_n(Z_i, Z_j)^2\} = o(n)$, *then* $nE[(U_n\psi_n - \hat{U}_n\psi_n)^2] = o(1)$. □

We also make use of the results of Serfling (1980).

Lemma 10 (Serfling) *Suppose* $\{Z_{0,i}\}_{i \in \mathbf{I_0}}$ *and* $\{Z_{1,i}\}_{i \in \mathbf{I_1}}$ *are independent and within each group they are i.i.d.,* $U_{n_0,n_1}\psi_{n_0,n_1} = (n_0 \cdot n_1)^{-1} \sum_{i \in \mathbf{I_0}} \sum_{j \in \mathbf{I_1}} \psi_{n_0,n_1}(Z_{0i}, Z_{1j})$, *and* $E\{\psi_{n_0,n_1}(Z_{0i}, Z_{1j})\} = 0$,

and $\hat{U}_{n_0,n_1}\psi_{n_0,n_1} = n_0^{-1}\sum_{i\in\mathbf{I_0}} p_{0n_0,n_1}(Z_{0i}) + n_1^{-1}\sum_{j\in\mathbf{I_1}} p_{1n_0,n_1}(Z_{1j})$, *where for* $k = 0, 1$, $p_{kn_0,n_1}(Z_{ki}) = E\{\psi_{n_0,n_1}(Z_{0i}, Z_{1i})|Z_{ki}\}$. *If* $0 < \lim_{n\to\infty} n_1/n_0 = \eta < \infty$, *where* $n = n_0+n_1$ *and* $E\left\{\psi_{n_0,n_1}(Z_{0i}, Z_{1j})^2\right\} = o(n_0) + o(n_1)$, *then* $nE[(U_{n_0,n_1}\psi_{n_0,n_1} - \hat{U}_{n_0,n_1}\psi_{n_0,n_1})^2] = o(1)$. □

In order to apply the lemmas to term (TR-3-1), note that it can be written as

$$N_1^{-3/2}\sum_{i\in\mathbf{I_1}}\sum_{j\in\mathbf{I_1,j\neq i}}\psi_{1N_0N_1}(Y_{1j}, X_j; X_i) \tag{TR-3-1a}$$

$$+ N_1^{-3/2}\sum_{i\in\mathbf{I_1}}\psi_{1N_0N_1}(Y_{1i}, X_i; X_i) \tag{TR-3-1b}$$

$$+ N_1^{-1/2}N_0^{-1}\sum_{i\in\mathbf{I_1}}\sum_{j\in\mathbf{I_0}}\psi_{0N_0N_1}(Y_{0j}, X_j; X_i). \tag{TR-3-1c}$$

Term (TR-3-1a) can be rewritten as $N_1^{-1/2}\sum_{i\in\mathbf{I_1}}\sum_{j\in\mathbf{I_1,j\neq i}}\psi^0_{1N_0N_1}(Y_{1j}, X_j; X_i; Y_{1i}, X_i; X_j)$ where

$$\psi^0_{1N_0N_1}(Y_{1j}, X_j; X_i; Y_{1i}, X_i; X_j) = [\psi_{1N_0N_1}(Y_{1j}, X_j; X_i) + \psi_{1N_0N_1}(Y_{1i}, X_i; X_j)]/2.$$

Thus by Lemma 9 and assumption (ii-a), term (TR-3-1a) is asymptotically equivalent to

$$N_1^{-1/2}\sum_{i\in\mathbf{I_1}} E\{\psi_{1N_0N_1}(Y_{1i}, X_i; X_j)|Y_{1i}, X_i\}.$$

By assumption (ii-a), term (TR-3-1b) is $o_p(1)$. By Lemma 10 and assumption (ii-a), term (TR-3-1c) is asymptotically equivalent to

$$N_1^{1/2}N_0^{-1}\sum_{j\in\mathbf{I_0}} E\{\psi_{0N_0N_1}(Y_{0j}, X_j; X_i)|Y_{0j}, X_j\}.$$

Hence putting these three results together, term (TR-3-1) is asymptotically equivalent to

$$N_1^{-1/2} \sum_{i \in \mathbf{I_1}} E\{\psi_{1N_0N_1}(Y_{1i}, X_i; X_j)|Y_{1i}, X_i\} + N_1^{1/2} N_0^{-1} \sum_{j \in \mathbf{I_0}} E\{\psi_{0N_0N_1}(Y_{0j}, X_j; X_i)|Y_{0j}, X_j\}.$$

Collecting all these results, we have established the asymptotic normality of the numerator.

We next examine the denominator of (T-1). Note that $N_1^{-1} \sum_{i \in \mathbf{I_1}} \hat{I}_i = N_1^{-1} \sum_{i \in \mathbf{I_1}} I_i + N_1^{-1} \sum_{i \in \mathbf{I_1}} (\hat{I}_i - I_i)$. The first term on the right-hand side converges in probability to $\mathrm{E}(I)$ by the law of large numbers. To see that the second term is $o_p(1)$, note that $N_1^{-1} |\sum_{i \in \mathbf{I_1}} (\hat{I}_i - I_i)| \leq N_1^{-1} \sum_{i \in \mathbf{I_1}} |\hat{I}_i - I_i|$. The Markov inequality implies for any $\varepsilon > 0$, $\Pr\left\{N_1^{-1} \sum_{i \in \mathbf{I_1}} |\hat{I}_i - I_i| > \varepsilon\right\} \leq E\left\{|\hat{I}_i - I_i|\right\} / \varepsilon$. Hence assumption (ii-d) implies that the second term is $o_p(1)$. This result, in conjunction with our result for the denominator, proves Theorem 2. ■