Matching Estimators

Petra E. Todd

October, 2006

1 Introduction

Matching is a widely-used nonexperimental method of evaluation that can be used to estimate the average effect of a treatment or program intervention. The method compares the outcomes of program participants with those of matched nonparticipants, where matches are chosen on the basis of similarity in observed characteristics. One of the main advantages of matching estimators is that they typically do not require specifying the functional form of the outcome equation and are therefore not susceptible to misspecification bias along that dimension. Traditional matching estimators pair each program participant with a single matched nonparticipant (see, e.g., Rosenbaum and Rubin, 1983), whereas more recently developed estimators pair program participants with multiple nonparticipants and use weighted averaging to construct the matched outcome.

We next define some notation and discuss how matching estimators solve the evaluation problem. Much of the treatment effect literature is built on the potential outcomes framework of Fisher (1935), exposited more recently in Rubin (1974, 1977) and Holland (1986). The framework assumes that there are two potential outcomes, denoted (Y_0, Y_1), that represent the states of being without and with treatment. An individual can only be in one state at a time, so only one of the outcomes is observed. The outcome that is not observed is termed a counterfactual outcome. The treatment impact for an individual is

$$\Delta = Y_1 - Y_0,$$

which is not directly observable. Assessing the impact of a program intervention requires making an inference about what outcomes would have been observed in the no-program state. Let $D = 1$ for persons who participate in the program and $D = 0$ for persons who do not. The $D = 1$ sample often represents a select group of persons who were deemed eligible for a program, applied to it, got accepted into it and decided to participate in it. The outcome that is observed is $Y = DY_1 + (1 - D)Y_0$.

Before considering different parameters of interest and their estimation, we first consider what is available directly from the data. The conditional distributions $F(Y_1|X, D = 1)$ and $F(Y_0|X, D = 0)$ can be recovered from the observations on Y_1 and Y_0, but not the joint distributions $F(Y_0, Y_1|X, D = 1)$, $F(Y_0, Y_1|X)$, or the impact distribution, $F(\Delta|X, D = 1)$. Because of this missing data problem, researchers often aim instead on recovering some features of the impact distribution, such as its mean. The parameter that is most commonly the focus of evaluation studies is the *mean impact of treatment on the treated*, $TT = E(Y_1 - Y_0|D = 1)$, which gives the benefit of the program to program participants.[1]

Matching estimators typically assume that there exist a set of observed characteristics Z such that outcomes are independent of program participation conditional on Z. That is, it is assumed that the outcomes (Y_0, Y_1) are independent of participation status D conditional on Z,[2]

$$(Y_0, Y_1) \perp\!\!\!\perp D \ |Z \ . \tag{1}$$

[1] If the outcome were earnings and the TT parameter exceeded the average cost of the program, then the program might be considered to at least cover its costs.

[2] In the terminology of Rosenbaum and Rubin (1983) treatment assignment is "strictly ignorable" given Z. The independence condition can be equivalently represented as $\Pr(D = 1|Y_0, Y_1, Z) = \Pr(D = 1|Z)$, or $E(D|Y_0, Y_1, Z) = E(D|Z)$.

It is also assumed that for all Z there is a positive probability of either participating ($D = 1$) or not participating ($D = 0$) in the program, i.e.,

$$0 < \Pr(D = 1|Z) < 1. \tag{2}$$

This assumption is required so that matches for $D = 0$ and $D = 1$ observations can be found. If assumptions (1) and (2) are satisfied, then the problem of determining mean program impacts can be solved by substituting the Y_0 distribution observed for matched on Z non-participants for the missing participant Y_0 distribution.

The above assumptions are overly strong if the parameter of interest is the mean impact of treatment on the treated (TT), in which case a weaker conditional mean independence assumption on Y_0 suffices[3]: $E(Y_0|Z, D = 1) = E(Y_0|Z, D = 0) = E(Y_0|Z)$. Furthermore, when TT is the parameter of interest, the condition $0 < \Pr(D = 1|Z)$ is also not required, because that condition is only needed to guarantee a participant analogue for each non-participant. The TT parameter requires only $\Pr(D = 1|Z) < 1$.

Under these assumptions, the mean impact of the program on program participants can be written as

$$\begin{aligned} \Delta_{TT} &= E(Y_1 - Y_0|D = 1) \\ &= E(Y_1|D = 1) - E_{Z|D=1}\{E_Y(Y|D = 1, Z)\} \\ &= E(Y_1|D = 1) - E_{Z|D=1}\{E_Y(Y|D = 0, Z)\}, \end{aligned}$$

where the second term can be estimated from the mean outcomes of the matched on Z comparison group.[4]

Assumption (**??**) implies that D does not help predict values of Y_0 conditional on Z, which rules out selection into the program directly on values of Y_0. However, there is no similar restriction imposed on Y_1, so the method does allow individuals who expect to experience higher levels of Y_1 to select into the program on the basis of that information. For estimating the TT parameter, matching methods allow selection into treatment to be based on possibly unobserved components of the anticipated program impact, but only insofar that the program participation decisions are based on the unobservable determinants of Y_1 and not those of Y_0.

Secondly, the matching method also requires that the distribution of the matching variables, Z, not be affected by whether the treatment is received. For example, age, gender, and race would generally be valid matching variables, but marital status may not be if it is were directly affected by receipt of the program. To see why this assumption is necessary, consider the term

$$E_{Z|D=1}\{E_Y(Y|D = 0, Z)\} = \int_{z\in Z}\int_{y\in Y} y\ f(y|D = 0, z)f(z|D = 1)dz.$$

It uses the $f(z|D = 1)$ conditional density to represent the density that would also have been observed in the no treatment ($D = 0$) state, which rules out the possibility that receipt of treatment changes the density of Z. Variables that are likely to be affected by the treatment or program intervention cannot be used in the set of matching variables.

With nonexperimental data, there may or may not exist a set of observed conditioning variables for which (1) and (2) (or (3) and (4)) hold. A finding of Heckman, Ichimura and Todd (1997) and HIST (1996,1998) in their application of matching methods to JTPA data is that (2) was not satisfied, meaning that no match could be found for a fraction of the nonparticipants. If there are regions where the support of Z does not overlap for the $D = 1$ and $D = 0$ groups, then matching is only justified when performed over the *region of common support*. The estimated treatment effect must then be defined conditionally on the region of overlap. Some methods for empirically determining the overlap region are described below.

Matching estimators can be difficult to implement when the set of conditioning variables Z is large. If Z are discrete, small cell problems may arise. If Z are continuous and the conditional mean $E(Y_0|D = 0, Z)$ is estimated nonparametrically, then convergence rates will be slow due to the so-called *curse of dimensionality*

[3] See Heckman, Ichimura, and Todd (1998).

[4] The notation $E_{Z|D=1}$ denotes that the expectation is taken with respect to the $f(Z|D = 1)$ density.

problem. Rosenbaum and Rubin (1983) provide a theorem that can be used to address this dimensionality problem. They show that for random variables Y and Z and a discrete random variable D

$$E(D|Y, P(D=1|Z)) = E(E(D|Y,Z)|Y, \Pr(D=1|Z)),$$

so that

$$E(D|Y,Z) = E(D|Z) \Longrightarrow E(D|Y, \Pr(D=1|Z) = E(D|\Pr(D=1|Z)).$$

This result implies that when Y_0 outcomes are independent of program participation conditional on Z, they are also independent of participation conditional on the probability of participation, $P(Z) = \Pr(D=1|Z)$. That is, when matching on Z is valid, matching on the summary statistic $\Pr(D=1|Z)$ (the *propensity score*) is also valid. Provided that $P(Z)$ can be estimated parametrically (or semiparametrically at a rate faster than the nonparametric rate), matching on the propensity score reduces the dimensionality of the matching problem to that of a univariate problem. For this reason, much of the literature on matching focuses on propensity score matching methods.[5] Using the Rosenbaum and Rubin (1983) theorem, the matching procedure can be broken down into two stages. In the first stage, the propensity score $\Pr(D=1|Z)$ is estimated, using a binary discrete choice model.[6] In the second stage, individuals are matched on the basis of their predicted probabilities of participation.

1.1 Justifying Matching Within a Model of Program Participation

We next describe a simple model of the program participation decision to illustrate the kinds of assumptions needed to justify matching.[7] Assume that individuals choose whether to apply to a training program on the basis of the expected benefits. He/she compares the expected earnings streams with and without participating, taking into account opportunity costs and net of some random training cost ε, which may include a psychic component expressed in monetary terms. The participation decision is made at time $t=0$ and the training program lasts for periods 1 through τ. The information set used to determine expected earnings is given by W, which might include, for example, earnings and employment history. The participation model is

$$D = 1 \text{ if } \mathrm{E}\left(\sum_{j=\tau}^{T} \frac{Y_{1j}}{(1+r)^j} - \sum_{k=1}^{T} \frac{Y_{0k}}{(1+r)^k} | W\right) > \varepsilon + Y_{00}, \text{ else } D = 0.$$

The terms of the right hand side of the inequality are assumed to be known to the individual but not to the econometrician.

If $f(Y_{0k}|\varepsilon + Y_{00}, X) = f(Y_{0k}|X)$, then

$$E(Y_{0k}|X, D=1) = E(Y_{0k}|X, \varepsilon + Y_{00} < \eta(W)) = E(Y_{0k}|X)$$

which would justify application of matching. This assumption places restrictions on the correlation structure of the earnings residuals. For example, the assumption would not be plausible if $X = W$ and $Y_{00} = Y_{0k}$, because then knowing that a person selected into the program ($D=1$) would be informative about subsequent earnings. We could assume, however, a model for earnings such as

$$Y_{0k} = \phi(X) + v_{0k},$$

where v_{0k} follows an MA(q) process with $q < k$, which would imply that Y_{0k} and Y_{00} are uncorrelated conditional on X. The matching method does not require that everything in the information set be known, but it does require sufficient information to make the selection on observables assumption plausible.

[5] Heckman, Ichumura and Todd (1998) and Hahn (1998) consider whether it is better in terms of efficiency to match on $P(X)$ or on X directly.

[6] Options for first the stage estimation include, for example, a parametric logit or probit model or a semiparametric estimator, such as semiparametric least squares (Ichimura, 1993), maximum score (Manski, 1973), smoothed maximum score (Horowitz, 1992), or semiparametric maximum likelihood (Klein and Spady, 1993). If $P(Z)$ were estimated using a fully nonparametric method, then the curse of dimensionality problem would reappear.

[7] This model is similar to an example given in Lalonde, Heckman and Smith (1999).

2 Cross-Sectional Matching Methods

For notational simplicity, let $P = P(Z)$. A prototypical propensity score matching estimator takes the form

$$\hat{\alpha}_M = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} [Y_{1i} - \hat{E}(Y_{0i}|D=1, P_i)] \qquad ((*))$$

$$\hat{E}(Y_{0i}|D=1, P_i) = \sum_{j \in I_0} W(i,j) Y_{0j},$$

where I_1 denotes the set of program participants, I_0 the set of non-participants, S_P the region of common support (see below for ways of constructing this set). n_1 is the number of persons in the set $I_1 \cap S_P$. The match for each participant $i \in I_1 \cap S_P$ is constructed as a weighted average over the outcomes of non-participants, where the weights $W(i,j)$ depend on the distance between P_i and P_j. Define a neighborhood $C(P_i)$ for each i in the participant sample. Neighbors for i are non-participants $j \in I_0$ for whom $P_j \in C(P_i)$. The persons matched to i are those people in set A_i where $A_i = \{j \in I_0 \mid P_j \in C(P_i)\}$. We describe a number of alternative matching estimators below, that differ in how the neighborhood is defined and in how the weights $W(i,j)$ are constructed.

2.1 Alternative Ways of Constructing Matched Outcomes

Nearest Neighbor matching Traditional, pairwise matching, also called *nearest-neighbor matching*, sets

$$C(P_i) = \min_j \|P_i - P_j\|, \ j \in I_0.$$

That is, the non-participant with the value of P_j that is closest to P_i is selected as the match and A_i is a singleton set. The estimator can be implemented either matching with or without replacement. When matching is performed with replacement, the same comparison group observation can be used repeatedly as a match. A drawback of matching without replacement is that the final estimate will usually depend on the initial ordering of the treated observations for which the matches were selected.

Caliper matching (Cochran and Rubin, 1973) is a variation of nearest neighbor matching that attempts to avoid "bad" matches (those for which P_j is far from P_i) by imposing a tolerance on the maximum distance $\|P_i - P_j\|$ allowed. That is, a match for person i is selected only if $\|P_i - P_j\| < \varepsilon$, $j \in I_0$, where ε is a pre-specified tolerance. Treated persons for whom no matches can be found within the caliper are excluded from the analysis, which is one way of imposing a common support condition. A drawback of caliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.

Stratification or Interval Matching In this variant of matching, the common support of P is partitioned into a set of intervals, and average treatment impacts are calculating through simple averaging within each interval. A weighted average of the interval impact estimates, using the fraction of the $D = 1$ population in each interval for the weights, provides an overall average impact estimate. Implementing this method requires a decision on how wide the intervals should be. Dehejia and Wahba (1999) implement interval matching using intervals that are selected such that the mean values of the estimated P_i's and P_j's are not statistically different from each other within intervals.

Kernel and Local Linear matching More recently developed matching estimators construct a match for each program participant using a weighted average over multiple persons in the comparison group. Consider, for example, the nonparametric *kernel matching estimator,* given by

$$\hat{\alpha}_{KM} = \frac{1}{n_1} \sum_{i \in I_1} \left\{ Y_{1i} - \frac{\sum_{j \in I_0} Y_{0j} G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)} \right\}.$$

where $G(\cdot)$ is a kernel function and a_n is a bandwidth parameter.[8] In terms of equation (*), the weighting function, $W(i,j)$, is equal to $\frac{G\left(\frac{P_j-P_i}{a_n}\right)}{\sum_{k\in I_0} G\left(\frac{P_k-P_i}{a_n}\right)}$. For a kernel function bounded between -1 and 1, the neighborhood is $C(P_i) = \{|\frac{P_i-P_j}{a_n}| \leq 1\}$, $j \in I_0$. Under standard conditions on the bandwidth and kernel , $\frac{\sum_{j\in I_0} Y_{0j} G\left(\frac{P_j-P_i}{a_n}\right)}{\sum_{k\in I_0} G\left(\frac{P_k-P_i}{a_n}\right)}$ is a consistent estimator of $E(Y_0|D=1,P_i)$.[9]

Heckman, Ichimura and Todd (1997) also propose a generalized version of kernel matching, called local linear matching.[10] The local linear weighting function is given by

$$W(i,j) = \frac{G_{ij} \sum\limits_{k\in I_0} G_{ik}(P_k-P_i)^2 - [G_{ij}(P_j-P_i)][\sum\limits_{k\in I_0} G_{ik}(P_k-P_i)]}{\sum\limits_{j\in I_0} G_{ij} \sum\limits_{k\in I_0} G_{ij}(P_k-P_i)^2 - \left(\sum\limits_{k\in I_0} G_{ik}(P_k-P_i)\right)^2}. \tag{9}$$

As demonstrated in research by Fan (1992a,b), local linear estimation has some advantages over standard kernel estimation. These advantages include a faster rate of convergence near boundary points and greater robustness to different data design densities. (See Fan, 1992a,b.) Thus, local linear regression would be expected to perform better than kernel estimation in cases where the nonparticipant observations on P fall on one side of the participant observations.

To implement the matching estimator given by equation (*), the region of common support S_P needs to be determined. The common support region can be estimated by

$$\hat{S}_P = \{P : \hat{f}(P|D=1) > 0 \text{ and } \hat{f}(P|D=0) > c_q\},$$

where $\hat{f}(P|D=d)$, $d \in \{0,1\}$ are standard nonparametric density estimators. To ensure that the densities are strictly greater than zero, it is required that the densities be strictly positive density (i.e. exceed zero by a certain amount), determined using a "trimming level" q. That is, after excluding any P points for which the estimated density is zero, an additional small percentage of the remaining P points are excluded for which the estimated density is positive but very low. The set of eligible matches is thus given by

$$\hat{S}_q = \{P \in \hat{S}_P : \hat{f}(P|D=1) > c_q \text{ and } \hat{f}(P|D=0) > c_q\},$$

where c_q is the density cut-off level that satisfies:

$$\sup_{c_q} \frac{1}{2J} \sum_{\{i\in I_1\cap \hat{S}_P\}} \{1(\hat{f}(P|D=1) < c_q + 1(1(\hat{f}(P|D=0) < c_q\} \leq q.$$

Here, J is the cardinality of the set of observed values of P that lie in $I_1 \cap \hat{S}_P$. That is, matches are constructed only for the program participants for which the propensity scores lie in $\hat{S}_q$.

The above estimators are representations of matching estimators and are commonly used. They can be easily adapted to estimate other parameters of interest, such as the Average Effect of Treatment on the Untreated (UT=$E(Y_1 - Y_0|D=0,X)$), or the average treatment effect (ATE=$E(Y_1 - Y_0|X)$), which is just a weighted average of treatment on the treated (TT) and treatment on the untreated (UT).

The recent literature has also developed alternative matching estimators that employ different weighting schemes to increase efficiency. See, for example, Hahn (1998) and Hirano, Imbens and Ridder (2003) for

[8] See Heckman, Ichimura and Todd (1997, 1998) and Heckman, Ichimura, Smith and Todd (1998).

[9] Specifically, we require that $G(\cdot)$ integrates to one, has mean zero and that $a_n \to 0$ as $n \to \infty$ and $na_n \to \infty$. One example of a kernel function is the quartic kernel, given by $G(s) = \frac{15}{16}(s^2-1)^2$ if $|s| < 1, G(s) = 0$ otherwise.

[10] Recent research by Fan (1992a,b) demonstrated advantages of local linear estimation over more standard kernel estimation methods. These advantages include a faster rate of convergence near boundary points and greater robustness to different data design densities. See Fan (1992a,b).

estimators that attain the semiparametric efficiency bound. The methods are not described in detail here, because those studies focus on the average treatment effect estimator (ATE) and not on the average effect of treatment on the treated (TT) parameter. Heckman, Ichimura and Todd (1998) develop a regression-adjusted version of the matching estimator, which replaces Y_{0j} as the dependent variable with the residual from a regression of Y_{0j} on a vector of exogenous covariates. The estimator uses a Robinson (1988)-type estimation approach to incorporate exclusion restrictions, i.e. that some of the conditioning variables in an equation for the outcomes do not enter into the participation equation or vice versa. In principal, imposing exclusions restrictions can increase efficiency. In practice, though, researchers have not observed much gain from using the regression-adjusted matching estimator. Some alternatives to propensity score matching are discussed in Diamond and Sekhon (2005).

2.2 When Does Bias Arise in Matching?

The success of a matching estimator depends on the availability of observable data to construct the conditioning set Z, such that (1) and (2) are satisfied. Suppose only a subset $Z_0 \subset Z$ of the required variables is observed. The propensity score matching estimator based on Z_0 then converges to

$$\alpha'_M = E_{P(Z_0)|D=1}\left(E(Y_1|P(Z_0), D=1) - E(Y_0|P(Z_0), D=0)\right). \tag{3}$$

The bias for the parameter of interest, $E(Y_1 - Y_0|D=1)$, is:

$$bias_M = E(Y_0|D=1) - E_{P(Z_0)|D=1}\{E(Y_0|P(Z_0), D=0)\}.$$

There is no way of a priori choosing the set of Z variables to satisfy the matching condition or of testing whether a particular set meets the requirements. In rare cases, where data are available on a randomized social experiment, it is sometimes possible to ascertain the bias.[11]

3 Difference-in-difference matching estimators

The estimators described above assume that after conditioning on a set of observable characteristics, outcomes are conditionally mean independent of program participation. However, for a variety of reasons there may be systematic differences between participant and nonparticipant outcomes, even after conditioning on observables, that could lead to a violation of the identification conditions required for matching. Such differences may arise, for example, because of program selectivity on unmeasured characteristics or because of levels differences in outcomes that might arise when participants and nonparticipants reside in different local labor markets or if the survey questionnaires used to gather the data differ in some ways across groups.

A difference-in-differences (DID) matching strategy, as defined in Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998), allows for temporally invariant differences in outcomes between participants and nonparticipants. This type of estimator matches on the basis of differences in outcomes using the same weighting functions described above. The propensity score DID matching estimator requires that

$$E(Y_{0t} - Y_{0t'}|P, D=1) = E(Y_{0t} - Y_{0t'}|P, D=0),$$

where t and t' are time periods after and before the program enrollment date. This estimator also requires the support condition given above, which must now hold in both periods t and t'. The local linear difference-in-difference estimator is given by

$$\hat{\alpha}_{DM} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} \left\{ (Y_{1ti} - Y_{0t'i}) - \sum_{j \in I_0 \cap S_P} W(i,j)(Y_{0tj} - Y_{0t'j}) \right\},$$

[11] See, for example, Heckman, Ichimura, and Todd (1997), Dehejia and Wahba (1998, 1999), and Smith and Todd (2004).

where the weights correspond to the local linear weights defined above. If repeated cross-section data are available, instead of longitudinal data, the estimator can be implemented as

$$\hat{\alpha}_{DM} = \frac{1}{n_{1t}} \sum_{i \in I_{1t} \cap S_P} \left\{ (Y_{1ti} - \sum_{j \in I_{0t} \cap S_P} W(i,j) Y_{0tj} \right\} - \frac{1}{n_{1t'}} \sum_{i \in I_{1t'} \cap S_P} \left\{ (Y_{1t'i} - \sum_{j \in I_{0t'}} W(i,j) Y_{0t'j} \right\},$$

where $I_{1t}, I_{1t'}, I_{0t}, I_{0t'}$ denote the treatment and comparison group datasets in each time period.

Finally, the DID matching estimator allows selection into the program to be based on anticipated gains from the program in the sense that D can help predict the value of Y_1 given P. However, the method assumes that D does not help predict changes $Y_{0t} - Y_{0t'}$ conditional on a set of observables (Z) used in estimating the propensity score. In their analysis of the effectiveness of matching estimators, Smith and Todd (2004) found difference-in-difference matching estimators to perform much better than cross-sectional methods in cases where participants and nonparticipants were drawn from different regional labor markets and/or were given different survey questionnaires.

4 Matching when the Data are Choice-based Sampled

The samples used in evaluating the impacts of programs are often choice-based, with program participants oversampled relative to their frequency in the population of persons eligible for the program. Under choice-based sampling, weights are generally required to consistently estimate the probabilities of program participation.[12] When the weights are unknown, Heckman and Todd (1995) show that with a slight modification, matching methods can still be applied, because the odds ratio $(P/(1-P))$ estimated using a logistic model with incorrect weights (i.e., ignoring the fact that samples are choice-based) is a scalar multiple of the true odds ratio, which is itself a monotonic transformation of the propensity scores. Therefore, matching can proceed on the (misweighted) estimate of the odds ratio (or of the log odds ratio).

5 Using Balancing Tests to Check the Specification of the Propensity Score Model

As described earlier, the propensity score matching estimator requires the outcome variable to be mean independent of the treatment indicator conditional on the propensity score, $P(Z)$. An important consideration in implemention is how to choose Z. Unfortunately, there is no theoretical basis for how to choose a particular set Z to satisfy the identifying assumptions and the set is not necessarily the most inclusive one.

To guide in the selection of Z, there is some accumulated empirical evidence on how bias estimates depended on the choice of Z in particular applications. For example, Heckman Ichimura Smith and Todd (1998), Heckman Ichimura and Todd (1997) and Lechner (2001) show that which variables are included in Z can make a substantial difference to the esimator's performance. These papers found that biases tended to be higher when the participation equation was estimated using a cruder set of conditioning variables. One approach adopted is select the set Z to maximize the percent of people correctly classified under the model. Another finding in these papers is that the matching estimators performed best when the treatment and control groups were located in the same geographic area and when the same survey instrument was administered to both treatments and controls to ensure comparable measurement of outcomes.

Rosenbaum and Rubin (1983) suggest a method to aid in the specification of the propensity score model. The method does not provide guidance in choosing which variables to include in Z, but can help to determine which interactions and higher order terms to include in the model for a given Z set. They note that for the true propensity score, the following holds:

$$Z \perp\!\!\!\perp D | \Pr(D = 1|Z),$$

[12] See, e.g., Manski and Lerman (1977) for discussion of weighting for logistic regressions.

or equivalently $E(D|Z, \Pr(D=1|Z)) = E(D|\Pr(D=1|Z))$. The basic intuition is that after conditioning on $\Pr(D=1|Z)$, additional conditioning on Z should not provide new information about D. If after conditioning on the estimated values of $P(D=1|Z)$ there is still dependence on Z, this suggests misspecification in the model used to estimate $\Pr(D=1|Z)$. The theorem holds for any Z, including sets Z that do not satisfy the conditional independence condition required to justify matching. As such, the theorem is not informative about what set of variables to include in Z.

This result motivates a specification test for $\Pr(D=1|Z)$, i.e. a test whether or not there are differences in Z between the $D=1$ and $D=0$ groups after conditioning on $P(Z)$. The test has been implemented in the literature a number of ways (see, e.g. Eichler and Lechner (2001), Dehijia and Wahba (1999,2001), Smith and Todd (2001), Diamond and Sekohn (2005)).

6 Assessing the Variability of Matching Estimators

The distribution theory for the cross-sectional and difference-in-difference kernel and local linear matching estimators given above is derived in Heckman, Ichimura and Todd (1998). However, implementing the asymptotic standard error formulae can be cumbersome, so standard errors for matching estimators are often instead generating using bootstrap resampling methods.[13] A recent paper by Abadie and Imbens (2004a) shows that standard bootstrap resampling methods are not valid for assessing the variability of nearest neighbor estimators but can be applied to assess the variability of kernel or local linear matching estimators. Abadie and Imbens (2004b) present alternative standard error formulae for assessing the variability of nearest neighbor matching estimators.

7 Applications

There have been numerous evaluations of matching estimators in recent decades. For a survey of many applications in the context of evaluating the effects of labor market programs, see Heckman, Lalonde and Smith (1999). More recently, propensity score matching estimators have been used in evaluating the impacts of a variety of program interventions in developing countries. Jyotsna and Ravaillon (1999) assess the impact of a workfare program in Argentina (the *Trabajar* program), and Jyotsna and Ravaillon (2003) study the effects of public investments in piped water on child health outcomes in rural India. Galiani, Gertler, and Schargrodsky (2005) use difference-in-difference matching methods to analyze the effects of privatization of water services on child mortality in Argentina. Other applications include Gertler, Levine and Ames (2004) in a study of the effects of parental death on child outcomes, Lavy (2004) in a study of the effects of a teacher incentive program in Israel on student performance, Angrist and Lavy (2001) in a study of the effects of teacher training on children's test scores in Israel, and Chen and Ravaillon (2003) in a study of a poverty reduction project in China.

Behrman, Cheng and Todd (2004) use a modified version of a propensity score matching estimator to evaluate the effects of a preschool program in Bolivia on child health and cognitive outcomes. They identify program effects by comparing children with different lengths of duration in the program, using matching to control for selectivity into atlernative durations. Also, see Imbens (2000) and Hirano and Imbens (2004) for an analysis of the role of the propensity score with continuous treatments. Lechner (2001) extends propensity score analysis for the case of multiple treatments.

References

[1] Abadie, Alberto and Guido Imbens (2004a): "On the Failure of the Bootstrap for Matching Estimators," manuscript, Harvard University.

[13] See Efron and Tibshirani (1993) for an introduction to boostrap methods, and Horowitz (2001) for a recent survey of bootstrapping in econometrics.

[2] Abadie, Alberto and Guido Imbens (2004b): "Large Sample Properties of Matching Estimators for Average Treatment Effects," manuscript, Harvard University.

[3] Angrist, J. and Lavy, V. (2001): "Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools," *Journal of Labor Economics*, 19(2), 343-369.

[4] Behrman, Jere, Yingmei Cheng and Petra Todd (2004): "Evaluating Preschool Programs when Length of Exposure to the Program Varies: A Nonparametric Approach," in Review of Economics and Statistics, 86,1, 108-132.

[5] Chen, Shaohua and Martin Ravallion, 2003. "Hidden Impact? Ex-Post Evaluation of an Anti-Poverty Program," Policy Research Working Paper Series 3049, The World Bank.

[6] Cochran, W. and Donald Rubin (1973): "Controlling Bias in Observational Studies," *Sankyha*, 35, 417-446.

[7] Dehejia, Rajeev and Sadek Wahba (1998): "Propensity Score Matching Methods for Nonexperimental Causal Studies," NBER Working Paper No. 6829.

[8] Dehejia, Rajeev and Sadek Wahba (1999): "Causal Effects in Noexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94(448), 1053-1062.

[9] Diamond, Alexis and Jasjeet S. Sekhon (2006): "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," working paper, Dept of Political Science, Berkeley.

[10] Efron, Bradley and Robert Tibshirani (1993): *An Introduction to the Bootstrap*, Chapman and Hall, New York: New York.

[11] Eichler, Martin and Michael Lechner (2001): "An Evaluation of Public Employment Programmes in the East German State of Sachsen-Anhalt," *Labour Economics*.

[12] Fan, J. (1992a): "Design Adaptive Nonparametric Regression, " *Journal of the American Statistical Association,* 87, 998-1004.

[13] Fan, J. (1992b): : "Local Linear Regression Smoothers and their Minimax Efficiencies," *The Annals of Statistics,* 21, 196-216.

[14] Fisher, R. A. (1935): *Design of Experiments,* New York: Hafner.

[15] Friedlander, Daniel and Philip Robins (1995): "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *American Economic Review*, 85(4), 923-937.

[16] Galiani, Sebastian, Gertler, Paul, and Ernesto Schargrodsky "Water for Life: The Impact of the Privatization of Water Services on Child Mortality in Argentina," *Journal of Political Economy.*

[17] Gertler, Paul, Levine, David and Minnie Ames (2004): "Schooling and Parental Death," *Review of Economics and Statistics, 86(1).*

[18] Hahn, Jinyong (1998): "On the Role of the Propensity Score in Efficient Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315-331.

[19] Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1996): "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evdience on the Effectiveness of Matching as a Program Evaluation Method," *Proceedings of the National Academy of Sciences*, 93(23), 13416-13420.

[20] Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica* , 66(5), 1017-1098.

[21] Heckman, James, Hidehiko Ichimura and Petra Todd (1997): "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies,* 64(4), 605-654.

[22] Heckman, James, Hidehiko Ichimura and Petra Todd (1998), "Matching As An Econometric Evaluation Estimator," *Review of Economic Studies,* 65(2), 261-294.

[23] Heckman, James, Robert Lalonde and Jeffrey Smith (1999): "The Economics and Econometrics of Active Labor Market Programs" in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics Volume 3A* (Amsterdam: North-Holland), 1865-2097.

[24] Heckman, James and Jeffrey Smith, with Nancy Clements (1997): "Making the Most Out of Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64(4), 487-536.

[25] Heckman, James and Petra Todd (1995): "Adapting Propensity Score Matching and Selection Models to Choice-based Samples," manuscript, University of Chicago.

[26] Hirano, Keisuke, Imbens, Guido and Geert Ridder (2000): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," manuscript, UCLA.

[27] Hirano, Keisuke, and Guido Imbens (2004): "The propensity score with continuous treatments," *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives*: 73-84 (A. Gelman & XL Meng, Eds.) New York: Wiley.

[28] Holland, P. W. (1986): "Statistics and Causal Inference (with discussion)," *Journal of the American Statistical Association*, 81, 945-970.

[29] Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model, " in Econometrica, Vol. 60, No. 3, 505-532.

[30] Ichimura, Hidehiko (1993): "Semiparametric Least Squares and Weighted SLS Estimation of Single Index Models," in Journal of Econometrics, 58, 71-120.

[31] Imbens, Guido (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," Biometrika, Vol. 87, No. 3, 706-710.

[32] Jalan, Jyotsna and Martin Ravaillon (1999): "Income Gains to the Poor from Workfare: Evidence for Argentina's Trabajar Program," Policy Research Working Paper Series, the World Bank.

[33] Jalan, Jyotsna and Martin Ravaillon (2001): "Does Piped Water Reduce Diarrea for Children in Rural India," *Journal of Econometrics*, 112, 153-173.

[34] Klein, R.W. and R.H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Reponse Models," in Econometrica, Vol. 61, No.2, 387-422.

[35] LaLonde, Robert (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review,* 76, 604-620.

[36] Lavy, Victor (2004): "Performance Pay and Teachers' Effort, Productivity and Grading Ethics," NBER Working Paper #10622, National Bureau of Economic Research.

[37] Lavy, Victor (2002): "Evaluating the Effects of Teachers' Group Performance Incentives on Pupil Achievement," Journal of Political Economics, 110(6), 1286-1387.

[38] Lechner, Michael (2001): "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in Lechner and Pfeiffer (eds), Econometric Evaluations of Active Labor Market Policies in Europe, *Heidelberg*, Physica.

[39] Manski, Charles (1973): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, Volume 3, 205-228.

[40] Manski, Charles and Steven Lerman (1977): "The Estimation of Choice Probabilities from Choice-Based Samples," *Econometrica*, 45(8), 1977-1988.

[41] Robinson, Peter (1988): "Root-N Consistent Nonparametric Regression," *Econometrica*, 56, 931-954.

[42] Rosenbaum, Paul and Donald Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects, "*Biometrika,* 70,41-55.

[43] Rosenbaum, Paul and Donald Rubin (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *American Statistician,* 39, 33-38.

[44] Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688-701.

[45] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis* (London: Chapman and Hall).

[46] Smith, Jeffrey and Petra Todd (2005a): "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125 (1-2), p. 305-353.