CEDE

DOCUMENTO CEDE 2006-13
ISSN 1657-7191 (Edición Electrónica)
FEBRERO DE 2006

A PRIMER ON PROPENSITY SCORE MATCHING ESTIMATORS

KATJA VINHA[1]

Abstract

Nonparametric matching estimators are frequently applied in evaluation studies. The general idea of the methodology is to determine the impact of treatment on the treated using information from treated and from similar non-treated observations to build a counterfactual of no treatment. I discuss the methodology for both the binary treatment case as well as for the multiple treatment case.

Key Words: Propensity score matching, binary treatment, multiple treatments.

JEL Classification: C14

[1] The primer is based on Chapter 2 of my dissertation *The impact of the Washington Metro on development patterns* written in partial fulfillment of the degree of Doctor of Philosophy at the University of Maryland, College Park.

METODOLOGÍA DE ESTIMADORES DE PAREO: UNA INTRODUCCIÓN

Resumen

Los estimadores de pareo no paramétricos son utilizados frecuentemente en estudios de evaluaciones de impacto. La idea general de la metodología es determinar el impacto de tratamiento sobre los tratados utilizando información de tratados y de individuos similares que no recibieron el tratamiento. A partir de esta información se construye el contrafactual de no tratamiento. El documento discute esta metodología para el caso en que existe un único tipo de tratamiento y para el caso de varios tratamientos.

Palabras clave: Estimadores de pareo, tratamiento binario, tratamiento múltiple.

Clasificación JEL: C14.

There are several reasons why traditional parametric regression analysis may not be suited for analyzing the impacts of endogenous (non-random) policies such as training programs where the participants self-select into treatment or policies on urban planning based on characteristics such as population density or geographical attributes of particular areas. Berhman, Cheng and Todd (2004) discuss the shortcomings of parametric regression analyses in these cases. First, it is possible that the participants in a particular program are quite different from the average non-participant. If these differences are important in affecting the desired outcomes then the non-participant group as a whole does not provide good information on the outcome of the participants had the participants not received the treatment. Using methods such as matching these differences are reduced since the control group individuals, that is those individuals used to build the counterfactual of no treatment, are re-weighted to better match the treatment group. Second, the true relationship between explanatory variables and the outcome variable may be very nonlinear. Since nonparametric estimation methods do not make any assumptions on the functional form (as do parametric analyses) it is not necessary to know the exact relationship between the explanatory and the outcome variables.[1] Third, in traditional regression analysis there may be problems of nonoverlapping support. It is possible that only treatment observations are found over certain ranges of x, and only control observations over other ranges. Traditional parametric regression analyses extrapolate the results to these regions where there are no observations. Non-parametric methods restrict the analysis to only those ranges that are similar. The objective of this

[1] Strictly speaking, propensity score matching is a quasi-parametric approach. Propensity scores used in constructing control groups are estimated parametrically, but treatment effects are nonparametrically determined.

paper is to outline one non-parametric methodology, the propensity score matching, both when there is only one type of treatment and when multiple treatments are available.

The propensity score matching methodology has been applied in various types of policy analyses. The methodology has been used extensively to evaluate the impacts of employment programs (for example, Dehejia & Wahba, 2002; Heckman, Ichimura & Todd, 1997; Sianesi 2004; Smith & Todd, 2005) and the impacts of education or training programs (for example, Black & Smith, 2004; Lechner, 2000; Saiz & Zoido, 2005). Also the methodology has been used to assess the impact of anti-poverty programs (Jalan & Ravallion, 2003b), of infrastructure (Jalan & Ravallion, 2003a) and of environmental policies (Greenstone, 2004). Recently it has also been applied in spatial context to determine the impact of zoning on land values (McMillen & McDonald, 2002), on job creation (O'Keefe, 2004) and on urban development patterns (Vinha, 2005). In general the above studies evaluate the impacts of a single treatment option. There are, however, also several studies that have analyzed the impacts of programs with multiple modalities (Frölich, Heshmati & Lechner, 2004; Lechner, 2002) or varying doses (Lu, Zanutto, Hornik & Rosenbaum 2001; Vinha, 2005).

Binary propensity score matching estimator

One objective of program evaluation is to calculate the mean impact of the program on those treated. If the treatment condition is denoted by $T=1$ for those who received the treatment and $T=0$ otherwise, and the impact variable of interest is denoted by y^1 if treatment was received and y^0 if not, then the objective is to estimate the

equation $E\left(y^1|T=1\right)-E\left(y^0|T=1\right)$, the difference in the outcome with and without the treatment for the treated group. Unfortunately the second term is not observable, since for an individual in the treatment no outcome without the treatment exists. Thus, the challenge is to be able to say something about the unobserved counterfactual for those who have been part of the program.

When the treatment is assigned *randomly*, then it can be assumed that the covariates and unobservables do not differ in any systematic way between the treated and non-treated groups. That is, they come from the same distribution. In this case, to estimate the average treatment effect on the treated, φ, of the outcome variable, y, one can compare the after treatment outcome levels of the two groups. The average treatment impact in a randomized experiment can be calculated as:

$$\varphi = E(y^1 \mid T=1) - E(y^0 \mid T=0) \qquad (1)$$

where the assumption is that $E(y^0 \mid T=1) = E(y^0 \mid T=0)$. That is, those in the treatment group would have had, on average, the same outcome level as the control group participants had they been assigned to the control group.

In the case of a non-randomized program, such as the building of a subway system, the treatment and control groups may vary in a systematic way, and it no longer can be assumed that $E(y^0 \mid T=1) = E(y^0 \mid T=0)$. Therefore, the treatment outcome measure for the non-participant group is not a valid counterfactual for the treatment group without treatment. As a specific example, if the location of subway stations is based on some characteristics of the area, such as population and employment densities,

then one would expect outcome measures in the treatment areas to be quite different from those in an average control area, even without the treatment. In this case, the outcome levels of the average non-treated areas are not good proxies for unobserved outcomes of the treatment group.

When there are no experimental data available, when assignment to the treatment group is non-random, and the treatment status is determined by some set of covariates, *x*, then an alternative mechanism needs to be employed to determine the treatment impact. One such mechanism is to establish a control group that is similar in x to the treatment group. The set of x ought to capture both the variables that affect the treatment decision, as well as those that influence the outcome measure.[2] The average treatment effect is based on the difference in the average outcomes of the individuals in the treatment group and this "matched" control group with similar set of x.

Matching on the covariates guarantees that the two groups have similar distributions of covariates and a treatment impact mimics that of a randomized experiment. Formally, the treatment impact is captured by

$$\varphi = E(y^1 \mid x, T=1) - E(y^0 \mid x, T=0) \qquad (2)$$

[2] The covariates do not need to include variables that are strictly from the pre-treatment period. That is, if the objective is to analyze the impact of a job-training program on wages or unemployment rates, it is not necessary to have information on the wages or unemployment status prior to entering the program. It is assumed that by matching on factors such as formal education, age, etc. that determine wages and unemployment status these pre-treatment conditions are also captured. However, the measures included in x that may influence the outcome measure should not have been affected by the treatment. For example, if the objective is to evaluate the impact of subway stations on the distribution of employment within a metropolitan area, it would not be possible to use current population density in the set of covariates that explain current employment density given that population density may have also been affected by the treatment (proximity to a subway station).

where the outcomes are conditioned on the covariates that determine treatment participation.

The above approach works only when (i) outcomes, conditional on the set of covariates, are independent of the group to which the individual belongs; and (ii) there is no covariate that unequivocally decides the treatment assignment. Mathematically, these conditions of strong ignorability[3] can be represented as:

$$(y^0, y^1) \perp\!\!\!\perp T \mid x \quad \text{and} \quad 0 < Pr\,(T=1) \mid x < 1. \qquad (3)$$

When the above conditions apply, the control group can be used to infer information about the treatment group. If there are any unobservables that influence the treatment decision and the first condition of strong ignorability does not hold, then the control group does not provide the necessary counterfactual information. The second condition rules out the possibility that any particular condition or characteristic unequivocally determines inclusion in or exclusion from the treatment.[4]

Rosenbaum and Rubin (1983) show that it is not necessary to match individuals based on the vector of observable characteristics, x, per se; matching on balancing scores, such as the propensity score, $b(x)$, is sufficient. The propensity score is, in effect, the conditional probability of being assigned to the treatment group given the individual's covariates. In Theorem 3, the authors demonstrate that when treatment assignment is strongly ignorable in x then it is also strongly ignorable in $b(x)$. That is if

[3] Strong ignorability is the same as conditional independence, unconfoundedness, or selection on observables.

[4] For example, it cannot be the case that all areas within x miles from the CBD are within a subway station treatment zone and no areas farther than x miles are outside station treatment zones.

$$(y^0, y^1) \perp\!\!\!\perp T \mid x \quad \text{and} \quad 0 < Pr\,(T=1) \mid x < 1$$

then also

$$(y^0, y^1) \perp\!\!\!\perp T \mid b(x) \quad \text{and} \quad 0 < Pr\,(T=1) \mid b(x) < 1. \qquad (4)$$

The above theorem greatly aids in the assignment of individuals into the control group since a univariate score can be used instead of a vector of individual covariates (or subclassification of the observations based on the covariates). Therefore, it is not necessary to match the observations based on multiple dimensions but only on a "summary" measure.

Rosenbaum and Rubin (1983) further show that if the treatment assignment is strongly ignorable, the average treatment effect can be obtained by comparing the treatment and matched control groups solely conditioned on the propensity score. Therefore, the treatment impact, $\hat{E}$, is given by:

$$\hat{E} = E\{y^1 \mid b(x), T = 1\} - E\{y^0 \mid b(x), T = 0\} = E\{y^1 - y^0 \mid b(x)\}. \qquad (5)$$

The average treatment effect is the average outcome level of those in the treatment group minus the average outcome level of those in the control group after conditioning on the propensity score. The methodology, besides determining the appropriate control group to use and reducing the bias in the treatment impacts, is also desirable because it allows for the control of covariates when the sample size is small (Rosenbaum and Rubin, 1983).

The impact, however, is valid only for the observations within the common support—that is, the range of propensity scores for which there are both control and treatment observations. For example, if there are no observations with high propensity

scores in the control group, then those observations with high propensity scores are outside of the region of support. Common support, *CS*, is defined as the set of propensity scores for which the distributions of *T=1* and *T=0* have positive values, such that $CS = pdf(T=1) > 0 \cap pdf(T=0) > 0$. That is, the common support is the range of propensity scores for which there is a positive probability of observing both treatment and control observations. It is possible that there is no exact match for a treatment observation's propensity score. As long as within a pre-specified interval of propensity scores (i.e. within 0.05 points) there is a control observation then the two observations are said to be within the same support.

In practice, the first step is to estimate a binary choice model (logit or probit) where the dependent variable is whether or not the observation is in the treatment group and the covariates include all the variables that influence the treatment condition as well as those that may affect the impact measures. These probabilities, $\hat{P}(x)$, are then used to construct the counterfactual of no treatment for the treated based on the non-treated individuals. There are several ways to construct the counterfactual, or several methods of matching the observations. These include counterfactuals based on one control observation per treatment observation, as well as counterfactuals based on some weighted average of several control observations.

In choosing the matching algorithm, the first decision is to determine the number of control observations. On the one hand, choosing only one control observation per

treatment observation[5] reduces the bias that is introduced when the matched pairs are less similar in their probability of receiving treatment. On the other hand, with a greater number of comparison observations the precision of the estimates, or the magnitude of the standard errors, is better. As often the case in empirical work, the trade-off is between unbiasedness and precision.

After determining the number of observations, it is necessary to define the matching estimator, or the manner in which the counterfactual is determined for each treatment observation. The generic matching estimator for observation i in the treatment group is given by

$$E\left(y^0 \middle| \hat{P}(x_i)\right) = \sum_{j=1}^{N_0} W\left(\hat{P}(x_i), \hat{P}(x_j)\right) y_j^0 \tag{6}$$

where $W(\cdot)$ determines the weight of each control observation j in the counterfactual for observation i. The various matching algorithms differ in the weights they place on the control observations to build the counterfactual.

If only one control is used per treatment observation, then the logical match for each treatment observation is the control observation with the closest propensity score, or nearest neighbor matching. In this case a weight of one is given to the control observation with the closest propensity score. That is, the treatment impact is given by $\frac{1}{N_1}\sum_{i=1}^{N_1}\left(y_i^1 - y_j^0\right)$ where N_1 is the number of treatment observations, y_i^1 is the outcome for

[5] The control observation in a pairwise-matching will be the observation with the closest propensity score to the treatment group observation.

treatment group observation *i*, and y_j^0 is the outcome for the control group observation *j* which has the closest propensity score to observation *i*. The nearest neighbor to observation *i* is defined as observation *j* such that $\left\|\hat{P}_i(x) - \hat{P}_j(x)\right\| \leq \left\|\hat{P}_i(x) - \hat{P}_k(x)\right\| \quad \forall \quad k \in I_o \neq j$ where I_0 is the set of all possible control observations. For nearest neighbor matching, it is also possible to set a maximum value, *d*, (often called a caliper) for the difference, such that $\left\|\hat{P}_i(x) - \hat{P}_j(x)\right\| \leq d$ in order to limit the differences between treatment and control observations. A caliper can also serve as a measure for observations to be within a common support. In this case, it is possible that not all treatment observations have a control observation within this maximum difference and that particular treatment observations will thus be dropped from the analysis.[6] As noted by Smith and Todd (2005) there is no way of determining, *a priori*, an acceptable size for *d*.

With nearest neighbor matching, one also needs to determine whether or not to match with replacement. When matching with replacement each control observation can serve as the counterfactual for more than one treatment observation. Dehejia and Wahba (2002) show that without replacement (and without imposing a caliper) the later matched pairs can differ considerably in their propensity scores. This is especially the case when there are relatively few possible controls for some range of propensity scores. Allowing replacement, the number of "better" matches increases. However, the variance of the

[6] That is, observations are not used since they do not fulfill the common support condition.

estimator increases given that less control group information is used and it possible that several control group observations are relied upon very heavily.

When multiple controls are assigned to a given treatment observation, then it is necessary to determine how to weight the control group observations to construct the counterfactual. Adapting the notation of Heckman, Ichimura and Todd (1997, 1998), the general form to calculate the average treatment impact, $\hat{M}(T)$, can be given as:

$$\hat{M}(T) = \sum_{i \in I_1 \subset CS} \omega(i)[y_i^1 - \sum_{j \in I_0} W(i,j)\, y_j^0] \tag{7}$$

where y_i^1 is the outcome with treatment for observation *i*, y_j^0 is the outcome for the control observation *j*, and *W(i,j)* is the weight that appears in equation (2.6). $W(i,j)$ is the weight given to observation *j* in the control group when comparing with observation i in the treatment group, such that $\sum_{j \in I_0} W(i,j) = 1$. That is, for each treatment observation, the weights of the controls used sum to one. I_0 and I_1 are the sets of observations in the control group and the treatment group, respectively. Only those treatment observations within the common support are used.[7] Finally, $\omega(i)$ is the weight of each treatment observation, *i*, in the construction of the average treatment impact. In general $\omega(i)$ is $1/N_1$, such that each treatment observation is weighted equally in the average treatment impact.

[7] Certain matching estimators impose the common support condition "automatically." In other cases it needs to be explicitly defined and thus the set of observations for which the weights are determined may not include all the treatment observations. An example of the first is the kernel matching estimator and of the second the local linear matching estimator.

The different matching algorithms differ in the way that the W matrix is determined. The simpler algorithms include N-neighbor matching and radial matching. In the first, the counterfactual outcome is made up of the average of the N control group observations closest in their propensity score to the treatment observation.[8] The average can be a simple average of the control group observations or an average weighted by the distance of the control group observation from the treatment observation. In radial matching an average of all the control observations with a propensity score within a certain distance, *d*, from the propensity score of the treatment observation is calculated. That is, the number of control observations used for each treatment observation may differ. Again, it is possible to use a weighted average instead of weighting all observations equally.

Additionally, when multiple controls are used other, more complex, matching algorithms are possible. Heckman, Ichimura and Todd (1997, 1998) propose two alternative estimators – kernel matching and local linear matching estimators – that build the counterfactual using additional information from the control group observations.

In a kernel estimator the matrix W is determined by a kernel function, $K(\cdot)$.[9] Following the general notation of Smith and Todd (2005), *W(i,j)* in this case is given by

[8] The formula is a generalized formula for the matching estimator. For example, for the case of 10-neighbor matching algorithm with simple weights, the *W(i,j)* matrix is such that for row *i*, the matrix has a value of $1/10$ in the columns for the ten control observations, *j*, with the closest propensity score to treatment observation *i*, and 0 otherwise.

[9] In essence the kernel function, $K(\cdot)$, is a histogram but instead of determining the frequency of observations in non-overlapping intervals, the kernel estimator estimates the density using *overlapping* intervals. Kernel functions used are symmetric and $\int K(z)dz = 1$.

$$W(i,j)=\frac{K\left(\frac{\hat{P}_j-\hat{P}_i}{h}\right)}{\sum_{k\in I_0}K\left(\frac{\hat{P}_k-\hat{P}_i}{h}\right)} \text{ if } |z|<\bar{Z} \text{ and 0 otherwise, where } z=\frac{\hat{P}_k-\hat{P}_i}{h} \qquad (8)$$

where h is the bandwidth of the kernel, and $\hat{P}_i$ and $\hat{P}_j$ are the probabilities of receiving treatment for treatment observation i and control observation j, respectively, and $\bar{Z}$ is some upper limit for a kernel value. This upper limit depends on the kernel used. There are several choices for the kernel function. They differ in the way they assign weight to observations depending on the distance of the two probabilities. For example, the rectangular kernel, which gives the same weight to all control observations (within a particular bandwidth), is

$$K(z)=0.5 \text{ if } |z|<1 \text{ and 0 otherwise, where } z=\frac{\hat{P}_k-\hat{P}_i}{h} .$$

The Epanechnikov kernel, which gives more weight to control observations with similar propensity scores, is

$$K(z)=0.75\left(1-0.2z^2\right)/\sqrt{5} \text{ if } |z|<\sqrt{5} \text{ and 0 otherwise}$$

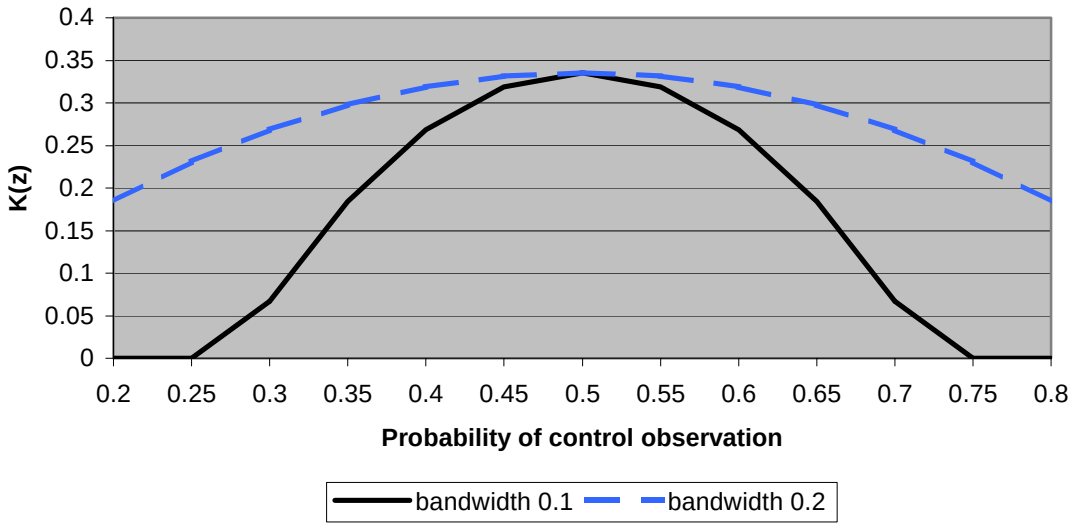
where $z=\frac{\hat{P}_k-\hat{P}_i}{h}$.

In general the choice of the kernel has been shown to have little impact on the estimated weight matrix; the choice of the bandwidth, however, does typically impact the weights (DiNardo & Tobias, 2001).

The bandwidth, *h*, in the kernel functions determines the interval over which positive weights are given to the control observations. A kernel with a small bandwidth

will use only control observations with very similar propensity scores to that of the treatment observation. A kernel with a larger bandwidth gives weight to less similar observations. [10] A sufficiently small bandwidth may not find any matches for the treatment observations. A sufficiently large bandwidth will give weight to all of the control observations such that the weight vector for a particular treatment observation will take on the shape of the kernel function.[11] Given the above property of the kernel estimator, it also limits the analysis to only those observations within a common support. That is, only observations with a probability of existing in both the treatment and the control groups, given the distribution of probabilities, are included. Figure 1 shows an example of the weights given to various control observations when the treatment observation has a propensity score of 0.5.

Figure 1: Countefactual using an Epanechnikov Kernel



[10] For the same difference in the propensity scores, $\hat{P}_j - \hat{P}_i$, a larger h will decrease the numerator quotient, z, of W such that more of the control observations will fulfill the $K(\cdot)$ rule.

[11] For example, if the rectangular kernel is used with a sufficiently large bandwidth, then all the control observations are used to calculate the counterfactual for each of the treatment observations, and the weight matrix would be a matrix of $1/N_0$, where N_0 is the number of control observations.

Given that the weight matrix (and therefore also the estimated treatment impacts) is in general sensitive to the choice of bandwidth, it is important to objectively determine the bandwidth. This can be done in several ways. The easiest way is to visually inspect the data and determine which bandwidth gives a good fit. However, one would like to determine more objectively, and in an automated manner, a good bandwidth (Härdle, 1990; Pagan and Ullah, 1999). The procedures to objectively determine the optimal bandwidth take as the basis the minimization of some global error.

The method that has been used in the evaluation literature is that of cross-validation, or the leave-one-out method (Black and Smith, 2004; Frölich, 2004a, 2004b). The objective of cross validation is to minimize the mean squared error when estimating the outcome measure y_j based on the information from the rest of the observations y_k such that $k \neq j$. That is, the mean squared error, $\frac{1}{N_0}\sum_{j \in I_0}\left(y_j - \bar{y}_{(j)}\right)^2$, is calculated for various bandwidths, where y_j is the outcome for observation j and $\bar{y}_{(j)}$ is the predicted outcome using the kernel estimator when observation j is not part of the sample. Given that the outcome without treatment only exists for the non-treated group, the measure is based on the non-treated sample. Efron and Gong (1983) summarize the methodology as consisting of the following:

> (a) deleting the points x_i from the data set one at a time; (b) recalculating the prediction rule on the basis of the remaining *n-1* points; (c) seeing how well the recalculated rule predicts the

deleted point; and (d) averaging these predictions of all n deletions of an x_i. (pg. 37)

Of the different bandwidths tested the one that minimizes the mean squared error is chosen as the "optimal" bandwidth.

The other matching estimator proposed by Heckman, Ichimura and Todd (1997, 1998) is the local linear estimator. Adapting the notation in Smith and Todd (2005) the estimator is given by:

$$W(i,j)=\frac{K_{ij}\sum_{k\in I_0}K_{ik}\left(\hat{P}_k-\hat{P}_i\right)^2-\left[K_{ij}\left(\hat{P}_k-\hat{P}_i\right)\right]\left[\sum_{k\in I_0}K_{ik}\left(\hat{P}_k-\hat{P}_i\right)\right]}{\sum_{j\in I_0}K_{ij}\sum_{k\in I_0}K_{ij}\left(\hat{P}_k-\hat{P}_i\right)^2-\left(\sum_{k\in I_0}K_{ik}\left(\hat{P}_k-\hat{P}_i\right)\right)^2}$$

where $K_{ij}=K((P_j-P_i/h)$. Again, any kernel function can be used.

Asymptotically all of the matching estimators will converge since in asymptotically large datasets the matches will be perfect. However, in finite samples there are differences. There are several studies that have compared the various matching estimators. The first set uses randomized experiments where $E(y^0 \mid T=1) = E(y^0 \mid T=0)$ and compares the impacts obtained with those derived from various matching algorithms on another dataset with non-participants that were not part of the experiment. Using this methodology, Dehejia and Wahba (2002) do not find any significant differences between nearest neighbor matching and radial matching. Smith and Todd (2005) also compare different matching estimators and similarly do not find any consistent results as to the superiority between nearest neighbor matching and local linear matching with reasonable

bandwidths. Based on the asymptotic properties of various estimators, Heckman, Ichimura and Todd (1997) advocate the use of local linear weights given that the estimator converges faster than kernel estimators. Frölich (2004a) using Monte Carlo studies finds, however, that ridge matching[12] and kernel matching are in general superior to pair-wise matching, and that local linear matching, multiple-neighbor estimators generally perform the poorest.[13] He finds that the local linear matching estimator does not perform as well as the other estimators, even if it asymptotically converges faster, when there are regions with low density of propensity scores. He finds that when the ratio of control observations to treatment observations is large, kernel matching is a good option.

When matching is done using a propensity score measure it is also necessary to determine whether or not the resulting non-treated sample is similar in the observables to the treated sample. That is, whether or not the two samples are balanced in the observables after the appropriate matching algorithm has been applied to obtain the counterfactuals for each treatment observation. The common support condition guarantees that only observations within the range of positive probabilities for both treatment and control groups are included. Balancing tests check via the use of t-tests that the means of the covariates, x, are statistically similar in the two groups (after weighting the control group observations by the weights used to construct the counterfactual). If the two samples are not similar then additional higher order terms,

[12] A weighted average of the local linear regression estimator and the Nadaraya Watson estimator.
[13] Furthermore Frölich finds that the weighting estimator is sensitive to trimming and states that there currently is no way to determine the optimal trimming level. Trimming is one method of imposing the common support condition, by excluding from the analysis the tails of the probability distributions of propensity scores.

such as squares of the covariates used, or interaction terms of the covariates need to be included in the construction of propensity scores until the two samples are similar (Dehejia and Wahba, 2002).

In order to obtain a confidence interval on the estimated treatment impact, bootstrapping methods are used. The standard errors are calculated by resampling the data with replacement and recalculating the treatment impact using the chosen estimator, N_B number of times. Each of the N_B samples is (potentially) different since a particular treatment observation may appear more than once. The distribution of the N_B different average treatment impacts are used to calculate the standard error or confidence intervals.

There are three options for determining the interval. If the underlying distribution is symmetric then either the standard error of the normal distribution or the percentile based confidence interval can be used. Ordering the treatment impacts, $\hat{\theta}_i$, from the lowest to highest, the percentile based confidence interval uses the $\hat{\theta}_{N_B \cdot x/100}$ and $\hat{\theta}_{N_B \cdot (1-x)/100}$ treatment impacts as the limits for a *(100-2x)%* confidence interval. When the underlying distribution is asymmetric then the bias-corrected bootstrap confidence intervals yield more accurate coverage probabilities (Efron and Tibshirani, 1998). In the bias-adjusted confidence intervals the percentile based confidence limits are adjusted by a factor taking into account the proportion of times in the true estimated impact using the full sample, $\hat{\theta}$, is greater than the bootstrapped replication (Efron and Tibshirani, 1998).

In effect, the confidence interval is adjusted for the difference in the median and mean impact values.[14]

Multiple treatment matching propensity score estimator

In some cases, the treatment is not a binary condition; there may be varying doses of treatment or a set of different treatment options. Joffe and Rosenbaum (1999), Imbens (2000) and Lechner (1999, 2002) expand the analysis the use of propensity score matching estimators when there are multiple mutually exclusive treatments.

In the multiple treatment case, it is necessary to determine for the M possible treatments the M theoretically possible outcomes, Y^1, Y^2,…, Y^M for each individual. Again, only one of the possible outcomes is realized for each individual and the other outcomes are "missing." The challenge is to be able to determine the counterfactual for all of those treatments that the individual did not experience.[15]

Imbens (2000) weakens the initial conditions imposed by Joffe and Rosenbaum (1999) for obtaining the average treatment impact in the multiple treatment case. He shows that it is not necessary for the treatment type to be independent of all the potential

[14] When there is no bias, that is, 50 percent of the replications are below the true estimated impact, the bias corrected and the percentile confidence intervals are the same.

[15] Lechner (1999) identifies three different average impacts that can be obtained. Namely, the expected average treatment effect of being in treatment t relative to treatment s for:

(1) a randomly chosen individual from the whole population, $\gamma_0^{t,s} = E(y^t) - E(y^s)$,

(2) a randomly chosen individual who received either treatment t or s, $\alpha_0^{t,s} = E\left(y^t \mid T = t, s\right) - E\left(y^s \mid T = t, s\right)$, and

(3) a randomly chosen individual who was in treatment t, $\theta_0^{t,s} = E\left(y^t \mid T = t\right) - E\left(y^s \mid T = t\right)$.

outcomes. The average treatment impacts can be estimated if there is only pairwise independence. This weaker condition (*weak* unconfoundedness) requires that the treatment type t is independent of the outcome, Y^t, when subjected to treatment t conditional on the covariates. Using the notation of Imbens (2000), if $D_i(t)$ is an indicator for each individual i such that:

$$D_i(t) = 1 \quad if \ T_i = t$$
$$D_i(t) = 0 \quad otherwise$$

then weak unconfoundedness can be expressed as

$$D(t) \perp Y^t \mid x \qquad \forall \ t.$$

The outcome Y^t is independent of whether or not treatment t is applied rather than of the treatment level per se.

Furthermore, Imbens (2000) shows that, as in the binary case, the propensity score can be used to condition the outcomes instead of the vector of observables, x. When the treatments are weakly unconfounded, then the average treatment effects are equal whether conditioning on the covariates or on the propensity score. Theorem 1 of Imbens (2000) states that

(i) $\beta(t, r) \equiv E\{Y^t \mid r(t, x) = r\} = E\{Y \mid T = t, r(T, x) = r\}$

(ii) $E\{Y^t\} = E\{\beta(t, r(t, x))\}$

where $r(t, x)$ is the generalized propensity score. That is, the conditional expectation of the impact evaluated at a particular treatment level, $\beta(t, r)$, is equal to the average treatment impact, $E\{Y^t\}$.

Given that there is a propensity score associated with each of the M treatments, more than one propensity score needs to be determined for each individual. That is, each individual needs to be evaluated for her propensity to receive each of the different treatments. Lechner (2002) describes two different ways – a structural approach and a reduced approach – of calculating the propensity scores. The first estimates the probabilities using a multinomial, or ordered, discrete choice model. The predicted probabilities from the model are used to calculate the conditional probabilities

$$\hat{P}_i^{s|ts}(x) = \frac{\hat{P}_i^s(x)}{\hat{P}_i^s(x) + \hat{P}_i^t(x)}. \tag{9}$$

where $\hat{P}_i^s(x)$ is the predicted probability of receiving treatment s given the vector of characteristics x. The conditional probabilities are required since the comparisons to be made are between two different groups and not all groups at the same time.

In the reduced approach separate binary choice equations are estimated for each of the possible $M*(M-1)/2$ pairs[16] of treatments in order to obtain $\hat{P}_i^{s|ts}(x)$. That is, only observations that received either treatment t or s are included in the calculation of the conditional probability. Lechner (2002) advocates the use of this second approach on two counts. First, in the ordered multinomial probit "if one choice equation is misspecified all conditional probabilities could be misspecified" (pg. 210), given that the probabilities are all evaluated at the same time. Second, it is easier to estimate binary models than ordered models. Lechner (2002) finds that the estimated conditional probabilities are highly correlated across the two approaches and thus the treatment

[16] Where M is the number of different groups, including the no treatment group.

impacts are very similar regardless of which approach is used to estimate the propensity scores.[17]

For the multiple-treatment case the common support set is in general determined by the minima of the maximum and the maxima of the minimum participation probabilities for the various treatment options (Frölich, Heshmati & Lechner, 2004). Equations 10 and 11 give the common support conditions for the lower bound and the upper bound, respectively.

$$Lower\ bound = \max\left(\min\left(\hat{P}_i^{s|ts}(x)\right) \forall\ t, s \in T\right) \tag{10}$$

$$Upper\ bound = \min\left(\max\left(\hat{P}_i^{s|ts}(x)\right) \forall\ t, s \in T\right) \tag{11}$$

For example, if there are three distinct treatment groups and the lowest probability of receiving treatment C is 0.1 in among those observations belonging to treatment group A and it is 0.05 among observations in group B, and 0.01 for those in treatment group C, then all those observations with a probability of receiving treatment C less than 0.1 are dropped from the sample. The procedure is applied to all of the different treatments.

Becuase with multiple treatments it is necessary to match on more than one conditional probability, in general, the matching is done using a nearest neighbor algorithm. The treatment impact is given by

$$\hat{M}(T) = \frac{1}{N_t} \sum_{i \in I_t \in CS} [y_i^t - \sum_{j \in I_s} W(i,j)\, y_j^s] \tag{12}$$

[17] All correlation coefficients for his sample were greater than 0.98.

where the $W(i,j)$ is 1 for observation j in treatment s that is the $\min\left(d(i,j)\,\forall\, j \in I_s\right)$, where $d(i, j)$ is the closeness of the two conditional probabilities $\hat{P}_k^{s|ts}(x)$ and $\hat{P}_k^{t|ts}(x)$ for $\forall k \in \{I_t, I_s\}$.

The distance metric generally used in the literature is the Mahalanobis distance.[18] Formally, the Mahalanobis distance $d(i, j)$, between observations i and j is defined as:

$$d(i,j) = \left(P_i^t - P_j^s\right)' V^{-1} \left(P_i^t - P_j^s\right)$$

where P_i^t is a vector of propensity scores for treatments t and s for observation i in treatment group t, P_j^s is the same vector of propensity scores for observation j in the alternative treatment group s. V is the covariance matrix based on the all the subset of observations from I_t and I_s such that,

$$V = \left\{\left(N_t - 1\right)V_t + \left(N_s - 1\right)V_s\right\} / \left(N_t + N_s - 2\right)$$

where N_k is the number of observations in treatment k, and V_k is the sample covariance of the relevant propensity scores, P, in group k, $k = t,s$ (Rubin, 1980).

As a summary, the algorithm proposed by Lechner (1999) for calculating the impact of different treatments is given in Table 1.

[18] There are not many applications of multiple treatment matching. Frölich, Heshmati and Lechner (2004), Lechner (2002) use the Mahalanobis distance as the metric to determine the nearest neighbor. Behrman, Cheng and Todd (2004) use local linear regression estimators, where the weights are given by the closeness of the observations in terms of the observable characteristics and dose.

Table 1: Algorithm for calculating multiple treatment impacts

Step 1	Specify and estimate a multinomial choice model to obtain $\left[\hat{P}_N^0(X), \hat{P}_N^1(X), ..., \hat{P}_N^M(X)\right]$
Step 2	Estimate the expectations of the outcome variables condition on the respective balancing scores. For a given value of m and l the following steps are performed: a) Compute $\hat{P}_N^{l\|ml}(X) = \frac{\hat{P}_i^l(X)}{\hat{P}_N^l(X) + \hat{P}_{Ni}^m(X)}$ or use $\left[\hat{P}_N^m, \hat{P}_N^l(X)\right]$ directly. Alternatively step 1 may be omitted and the conditional probabilities may be directly modeled (as in the binary case). b) Choose one observation in the subsample defined by participation in m and delete it from that pool. c) Find an observation in the subsample of participants in l that is as close as possible to the one chose in step a) in terms of $\hat{P}_N^{l\|ml}(X)$ or $\left[\hat{P}_N^m, \hat{P}_N^l(X)\right]$. In the case of using $\left[\hat{P}_N^m, \hat{P}_N^l(X)\right]$ "closeness" can be based on the Mahalanobis distance. Do not remove that observation, so that it can be used again. d) Repeat a) and b) until no participant in m is left. e) Using the matched comparison group formed in c) compute the respective conditional expectation by the sample mean. Note that the same observations may appear more than once in that group.
Step 3	Repeat step 2 for all combinations of m and l.
Step 4	Compute the estimate of the treatment effects using the results of step 3.
Source: Lechner, 1999.	

List of references

Behrman, J.R., Cheng, Y., & Todd, P.E. (2004). Evaluating preschool programs when length of exposure to the program varies: a nonparametric approach, *The Review of Economics and Statistics*, 86 (1), 108-132.

Black, D., & Smith, J. (2004). How robust is the evidence on the effects of college quality? Evidence from matching, *Journal of Econometrics*, 121 (1-2), 99-124.

Dehejia, R., & Wahba, S. (2002). Propensity score-matching methods for nonexperimantal causal studies, *The Review of Economics and Statistics*, 84 (1), 151-161.

DiNardo, J., & Tobias, J.L. (2001). Nonparametric density and regression estimation, *The Journal of Economic Perspectives*, 15 (4), 11-28.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation, *The American Statistician*, 37 (1), 36-48.

Efron, B., & Tibshirani, R.J. (1998). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall.

Frölich, M. (2004a). Finite-sample properties of propensity-score matching and weighting estimators, *The Review of Economics and Statistics*, 86 (1), 77-90.

Frölich, M. (2004b). Matching estimators and optimal bandwidth choice, Department of Economics, University of St. Gallen, version July 24.

Frölich, M., Heshmati, A., & Lechner, M. (2004). A microeconometric evaluation of rehabilitation of long-term sickness in Sweden, *Journal of Applied Econometrics*, 19 (3), 375-396.

Greenstone, M., (2004). Did the Clean Air Act cause the remarkable decline in sulfur dioxide concentrations?, *Journal of Environmental Economics and Management*, 47 (3), 585-611.

Härdle, W. (1990). *Applied nonparametric regression*, Cambridge, UK: Cambridge University Press.

Heckman, J., Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from a job training programme, *Review of Economic Studies*, 64, 605-654.

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

Imbens, G.W. (2000). The role of the propensity score in estimating dose-response functions, *Biometrika*, 87 (3), 706-710.

Jalan, J., & Ravallion, M. (2003a). Does piped water reduce diarrhea for children in rural India?, *Journal of Econometrics*, 112 (1), 153-173.

Jalan, J., & Ravallion, M. (2003b). Estimating the benefit incidence of an antipoverty program by propensity-score matching. *Journal of Business & Economic Statistics*, 21 (1), 19-30.

Joffe, M.M., & Rosenbaum, P.R. (1999). Propensity Scores, *American Journal of Epidemiology*, 150, 327-333.

Lechner, M. (1999). Identification & estimation of causal effects of multiple treatments under the conditional independence assumption, IZA Discussion Paper No. 91. (Later published in M. Lechner & F. Pfeiffer (Eds.) *Econometric evaluations of active labor market policies in Europe*, Heidelberg: Physica/Springer, 2001).

Lechner, M. (2000). An evaluation of public-sector-sponsored continuous vocational training programs in East Germany, *Journal of Human Resources*, 35 (2), 347-75.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies, *The Review of Economics and Statistics*, 84 (2), 205-220.

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P.R. (2001). Matching with doses in an observational study of a media campaign against drug abuse, *Journal of the American Statistical Association*, 96 (456), 1245-53.

McMillen, D.P., & McDonlad, J.F. (2002). Land values in a newly zoned city, *The Review of Economics and Statistics*, 84 (1), 62-72.

O'Keefe, S. (2004). Job creation in California's enterprise zones: A comparison using a propensity score matching model, *Journal of Urban Economics*, 55 (1), 131-150.

Pagan, A., & Ullah, A. (1999). *Nonparametric Econometrics*, Cambridge, UK: Cambridge University Press.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrica*, 70 (1), 41-55.

Rubin, D.B. (1980). Bias reduction using Mahalanobis-metric matching, *Biometrics*, 36 (2), 293-298.

Saiz, A., & Zoido, E. (2005). Listening to what the world says: Bilingualism and earnings in the United States, *The Review of Economics and Statistics*, 87 (3), 523-538.

Sianesi, B. (2004). An evaluation of the Swedish system of active labor market programs in the 1990s, *The Review of Economics and Statistics*, 86 (1), 133-155.

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators?, *Journal of Econometrics*, 125 (1-2), 305-353.

Vinha, K. (2005). The impact of the Washington Metro on development patterns. Unpublished doctoral dissertation. University of Maryland, College Park.