A Practical Guide to Implementing Matching Estimators*

October, 1999

I. *Introduction*

Matching estimators evaluate the effects of a treatment intervention by comparing outcomes for treated persons to those of similar persons in a comparison group. Treatment may represent, for example, participation in a training program, where the outcome is earnings or employment after the program intervention. Comparison group persons are determined to be suitable matches for treated persons if they have similar observed characteristics, as measured by some distance metric. Here we discuss different types of matching estimators that are proposed and considered in greater detail in Heckman, Ichimura, and Todd (1997, 1998) and in Heckman, Ichimura, Smith and Todd (1998). We give special consideration to implementation issues. At the end, some sample code is provided.

Notation:

- Let Y_1 denote the outcome for persons who receive the treatment.
- Let Y_0 denote the outcome without treatment.
- Let $D = 1$ if persons receive treatment, $D = 0$ if not.
- Let X denote other characteristics used as conditioning variables.
- Let $P(X) = Pr(D = 1|X)$

*This guide prepared by Petra Todd for IADB meeting in Santiago, Chile. Please direct any comments to petra@athena.sas.upenn.edu.

There are multiple types of matching estimators that differ in terms of the assumptions needed to justify their application and in terms of the estimation methods used in constructing the matches. They can be broadly classified into two main types:

- *cross-sectional (CS) matching estimators* compares the outcomes for treatments and comparison groups persons measured at some time period after the program.
- *difference-in-difference (DID) matching estimators* compares the change in outcomes for treatments to the change in outcomes for comparison group members, where the change is measured relative to some preprogram benchmark time period.

The advantage of using a difference-in-difference estimator instead of a cross-sectional estimator is that it allows for time-invariant unobservable differences between treatment and comparison group individuals. A major advantage of having baseline or preprogram data is that it allows a difference-in-difference strategy to be used.

The specific matching estimators that are discussed here are:

(a) nearest neighbor cross-sectional matching estimator
(b) nearest neighbor difference-in-difference (DID) matching estimator
(c) kernel and local linear versions of the above estimators.

In Heckman, Ichimura and Todd (1997, 1998) another kind of matching estimator, called *regression-adjusted matching* is developed. We do not consider it here, because it is more difficult to implement than the other methods.

II. *Identifying assumptions of different estimators*

A key parameter of interest in evaluations is the *mean impact of treatment on the treated*. It can be defined conditional on some characteristics X

$$\Delta_{D=1}(X) = E(Y_1 - Y_0|X, D = 1)$$

or an averaged parameter may be defined over some support of X, S_x :

$$\Delta_{D=1} = \frac{\int_{S_X} E(Y_1 - Y_0|X, D=1) f_x(X|D=1) dX}{\int_{S_X} f_x(X|D=1) dX},$$

where $f_x(X|D=1)$ is the density of X.

All estimators described below aim to estimate the overall mean impact of treatment on the treated, $\Delta_{D=1}$.[1]

A. Cross-sectional Matching Estimator

This estimator assumes:

(CS.1) $E(Y_0|P(X), D=1) = E(Y_0|P(X), D=0)$
(CS.2) $0 < \Pr(D=1|X) < 1$

Under these conditions, $\Delta_{D=1}$ can be estimated by

$$\hat{\Delta}^{CS}_{D=1} = n_1^{-1} \sum_{\substack{i=1 \\ \{D_i=1\}}}^{n_1} Y_{1i}(X_i) - \hat{E}(Y_{0i}|P(X_i), D_i=0),$$

where n_1 are the number of treated individuals with X values that satisfy CS.2. $\hat{E}(Y_{0i}|P(X_i), D_i=0)$ can be estimated by a nonparametrically by nearest neighbor, kernel, or local linear regression. These estimators are discussed below.

B. Difference-in-difference (DID) Matching Estimator

This estimator requires repeated cross-section data (or longitudinal data) on program participants and nonparticipants. Let t and t' be two time periods, one before the program start date and one after. Y_{0t} is the outcome observed at time t. Conditions needed to justify the application of the estimator are:

(DID.1) $E(Y_{0t} - Y_{0t'}|P(X), D=1) = E(Y_{0t} - Y_{0t'}|P(X), D=0)$
(DID.2) $0 < \Pr(D=1|X) < 1$

[1]Slightly modified version of the matching estimators considered here could be used to get at other parameters, such as the impact of treatment on a person randomly assigned to the program.

Under these conditions, $\Delta_{D=1}$ can be estimated by

$$\begin{aligned}\hat{\Delta}^{DID}_{D=1} &= n_{1t}^{-1} \sum_{\substack{i=1 \\ \{D_i=1\}}}^{n_{1t}} \{Y_{1ti}(X_i) - \hat{E}(Y_{0ti}|P(X_i), D_i = 0)\} - \\ & \quad n_{1t'}^{-1} \sum_{\substack{j=1 \\ \{D_j=1\}}}^{n_{1t'}} \{Y_{0t'j}(X_j) - \hat{E}(Y_{0t'j}|P(X_j), D_j = 0)\}\end{aligned}$$

where n_{1t} and $n_{1t'}$ are the number of observations in the two time periods.

III. *Implementation*

We next discuss how to implement the above estimators. A sample program (written in pseudo-code) is provided for a simple average nearest neighbor matching. Additional programs (written in Splus and Fortran) are provided in the appendix for the local linear estimator.

A. *Step One - Estimate a model for program participation.*

The conditional probability of participating in the program (also called the *propensity score*) plays an important role in implementing both matching and traditional econometric selection estimators. If the probability of participating in the program is estimated by a paramateric procedure (such as logit or probit), this provides a way of reducing the dimension of the conditioning problem in matching. That is, the problem of matching is reduced to a one-dimensional, nonparametric estimation problem – that of estimating $E(Y_0|D = 0, P(X))$ – instead of a k dimensional problem – that of estimating $E(Y_0|D = 0, X)$.

Estimating the propensity scores requires choosing a set X of conditioning variables. It is important to restrict the choice of X variables to ones that are not influenced by the program. Otherwise, the matching estimator will not correctly measure the effect of the program, because it will not capture changes in the distribution of the X variables induced by the program. For this reason, X variables are usually chosen to be characteristics of the persons prior to entering the program.[2] For example, X might include employment

[2]However, even these variables could be conceivably influenced by the program since individuals anticipate their entry into the program. For example, some individuals might quit their jobs in an attempt to qualify for entry into the program.

history in the year prior to entering a training program, which was found to be an important predictor of program participation in the Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998) papers.

In Heckman, Ichimura and Todd (1997, hereafter HIT), conditional probabilities of program participation are estimated by logistic regression. The set of matching variables is chosen to as the subset of variables that maximizes the equally weighted percentage of observations correctly classified under the logistic model. [3]In HIT, we found that the matching estimators performed best when a rich set of conditioning variables was used. The quality of the estimates deteriorated substantially (in terms of having greater bias) when a crude set of variables consisting only of general demographic characteristics was used.

What if the data is nonrandomly sampled?

In evaluation settings, the data is often gathered using a choice-based sampling design, where observations on the treatment group are either over or undersampled relative to their frequency in a random population. Usually data on treated persons is combined with data on comparison group persons and treatments are overrepresented relative to their frequency in a random population.

With choice-based sampled data, weighting of observations is required to obtain consistent estimates of the propensity scores. (See Amemiya, 1985) Often, though, the analyst does not know which weights to use. In this case, it is still possible to use matching estimators, but it requires matching on the log-odds ratio $\log(\hat{P}(X_i)/1 - \hat{P}(X_i))$ instead of on the estimated propensity scores directly.[4]

B. *Step Two - Construct the matched outcomes*

Constructing matched outcomes requires estimating $E(Y_{0i}|P(X_i), D_i = 0)$ for the cross-sectional matching estimator and $E(Y_{0ti}|P(X_i), D_i = 0)$ and

[3]However, there is no real justification for choosing the set in this way and it is useful to examine the sensitivity of impact estimates obtained by matching methods to alternative sets of conditioning variables.

[4]Heckman and Todd (1998) show that the log-odds ratio of the propensity scores estimated using the wrong weights (i.e. ignoring the fact that the data is choice-based sampled) is a scalar multiple of the true log-odds ratio.

$E(Y_{0t'i}|P(X_i), D_i = 0)$ for the difference-in-difference estimator. There are several different nonparametric estimators that could be used to estimate these conditional means. In HIT, we use local smoothing estimators that estimate the conditional mean by a weighted average of outcomes observed for $D_i = 0$ observations. A kernel estimator for

$$E(Y_{0i}|P(X_i), D_i = 0)$$

is given by

$$\hat{E}(Y_{0i}|P(X_i), D_i = 0) = \sum_{\substack{j=1 \\ \{D_j=0\}}}^{n_0} W_j(P(X_i))Y_{0j},$$

with weights

$$W_j(P(X_i)) = \frac{K(\frac{P(X_i)-P(X_k)}{h_n})}{\sum_{\substack{k=1 \\ \{D_k=0\}}}^{n_0} K(\frac{P(X_i)-P(X_k)}{h_n})}$$

K is a kernel function and h_n is a bandwidth, or smoothing parameter. (The choice of kernel function and bandwidth will be further discussed below)

Nearest neighbor, kernel, and local linear estimators of the conditional means can all be written in the same form as a weighted sum of the comparison group outcomes. The estimators differ only in the choice of weighting function $W_j(P(X_i))$.

(a) *Simple average nearest neighbor estimators*

The easiest estimator to implement is the *simple average nearest neighbor estimator*. First determine how many neighbors you want to use (one, two, five, ten, twenty, etc.). Then select the neighbors by their proximity to treatment group $P(X_i)$ values - i.e. for each $P(X_i)$ value observed for treatment group members, select the neighbors as the $D_i = 0$ observations with the closest propensity scores in terms of Euclidean distance. This can be done as follows:

(a) form $|P(X_i) - P(X_j)|$ for treatment observation i and for all comparison group observations j.

(b) sort the j observations in terms of $|P(X_i) - P(X_j)|$ from lowest to heighest.

(c) Let A_x index the set of x observations with the lowest values of $|P(X_i) - P(X_j)|$. These are the so-called *nearest neighbors.*

(d) Construct the matched outcome as a simple average over the outcomes for the nearest neighbors.

$$\hat{E}(Y_{0i}|P(X_i), D_i = 0) = \frac{1}{x} \sum_{\substack{j=1 \\ \{D_j \in A_x\}}}^{x} Y_{0j}$$

(b) *kernel regression matching estimator*

The *simple average nearest neighbor* estimator assigns either a weight of $\frac{1}{x}$ or zero to all comparison group observations. If a 5th nearest neighbor estimator is used, for example, the second and third nearest neighbors receive equal weight. A kernel regression estimator chooses the weights so that the observations closer in terms of the distances $|P(X_i) - P(X_j)|$ receive greater weight. This weighting is achieved through a kernel function. One kernel function that is often used is the "biweight kernel, " given by

$$\begin{aligned} K(s) &= \frac{15}{16}(s^2 - 1)^2 \text{ for } |s| < 1 \\ &= 0 \text{ else.} \end{aligned}$$

Kernel functions are usually chosen to satisfy $\int K(s)ds = 1$ and $\int K(s)sds = 0$.[5] The biweight kernel is symmetric and satisfies these properties.

Implementing a kernel estimator requires choosing a bandwidth h_n, which is analogous to the problem of choosing the number of neighbors in a nearest neighbor setting. The weights given to $D_j = 0$ observations depend on the values $K(\frac{P(X_i)-P(X_j)}{h_n})$.

There is a large literature on choosing bandwidths (or smoothing parameters) in nonparametric estimation.[6] Consistency of the nonparametric estimator requires that the bandwidth shrinks to zero as the sample size gets large, but not at too fast a rate. One simple and effective way of choosing

[5]These conditions are used in showing consistency of kernel density and regression estimators.

[6]See, for example, the survey article by Jones, Marron, and Sheather (1996).

the bandwidth to set the bandwidth equal to the absolute value of the distance to the xth nearest neighbor (i.e. $h_n = |P(X_i) - P(X_j)|$, where $P(X_j)$ is the propensity score for the X_j nearest neighbor). When the bandwidth is chosen in this way, it will vary from point to point of evaluation $P(X_i)$, with smaller bandwidths for points of evaluation where there is more data in the local neighborhood. Alternatively, a fixed bandwidth value could be specified (since $P(X)$ lies in between 0 and 1, an appropriate bandwidth choice might be 0.2 or 0.4). In any case, sensitivity of the estimates with respect to bandwidth choice should be examined.

(c) *local linear regression (LLR) estimator*

Local linear regression is a nonparametric regression technique that improves on the more traditional kernel regression estimator in two ways, as was shown by Fan (1992,1993).

(i) The bias of the local linear regression estimator does not depend on the design density of the data (i.e. on the density $f(P(X))$

(ii) The order of convergence of the bias of the local linear regression estimator is the same at boundary points as at interior points (it avoids the boundary bias problem associated with kernel regression estimators).

Local linear regression differs from kernel regression only in terms of weights, which for LLR are given by

$$W_j(P(X_i)) = \frac{K_{ij} \sum_{k=1}^{n_0} K_{ik}(P_k - P_i)^2 - [K_{ij}(P_j - P_i)][\sum_{k=1}^{n_0} K_{ik}(P_k - P_i)]}{\sum_{j=1}^{n_0} K_{ij} \sum_{k=1}^{n_0} K_{ik}(P_k - P_i)^2 - [\sum_{j=1}^{n_0} K_{ij}(P_j - P_i)]^2}$$

where $K_{ik} = K(\frac{P(X_i) - P(X_k)}{h_n})$.

Fan showed that the local linear estimator for $E(Y_{0i}|P(X_i), D_i = 0)$ can also be viewed as the solution $\hat{a}$ to the weighted regression problem

$$\min_{a,b} \sum_{\substack{j=1 \\ \{D_j=0\}}}^{n_0} (Y_{0j} - a - b(P(X_j) - P(X_i)))^2 K(\frac{P(X_i) - P(X_j)}{h_n}).$$

This provides another way of implementing the estimator. Namely, for each value $P(X_i)$ run a weighted least squares regression of Y_{0j} on a constant

term and on $P(X_j) - P(X_i)$ using data on $D_j = 0$ persons. The estimated intercept will be the estimate of $E(Y_{0i}|P(X_i), D_i = 0)$. (Note that a separate weighted least squares problem will need to be estimated for each point of evaluation $P(X_i)$, as changing the point of evaluation changes the weights.)

What if there are no close matches?

Nonparametric estimators of $E(Y_{0i}|P(X_i), D_i = 0)$ are only defined at points where the density $f(P(X_i)|D = 0) > 0$. These means, roughly speaking, that there should be $P(X_j)$ values for the $D = 0$ group in the vicinity of each $P(X_i)$ point of evaluation. $D_i = 1$ observations with $P(X_i)$ values for which there are no close matching $P(X_j)$ observations should be excluded in estimation.

We term the support of $P(X)$ for which both $f_x(P(X)|D = 1) > 0$ and $f_x(P(X)|D = 0) > 0$ the *region of overlapping support*. Implementing matching estimators requires determining which $P(X)$ values are in the overlapping support region. The mean program impact can only be obtained for treatment group persons in the overlap region.

One way of determining which observations lie in the region of overlapping support is simply to plot the histogram of the $P(X_i)$ values for both the treatment and comparison groups and then visually identify any ranges of $P(X_i)$ where there are no close matches. Another more rigorous way of determining the overlapping support region is to calculate the density $f(P(X_i)|D = 0)$ (using $D = 0$ comparison group data) at each of the $P(X_i)$ values observed for $D_i = 1$ observations. Nonparametric density estimators can be used to estimate these densities. The standard nonparametric density estimator is given by

$$\hat{f}(P(X_i)|D = 0) = \sum_{\substack{k=1 \\ \{D_k=0\}}}^{n_0} K(\frac{P(X_i) - P(X_k)}{h_n}),$$

where K is a kernel function and h_n the bandwidth parameter.[7]

After the estimates of the density at each point are obtained, rank the density estimates. Then find the 1 or 2% quantile of the positive density

[7]We do not recommend using a normal kernel in estimation because the normal kernel has infinite support, so it will give all positive density estimates at all points of evaluation. There are many different methods for choosing the bandwidth in density estimation. One commonly used method is the rule-of-thumb method, described in Silverman (1986).

estimates. All values of $P(X_i)$ for which the estimated density exceeds this threshhold are considered to be in the overlapping support region. Values below the threshold are outside the region and should be excluded in estimation.[8]

[8]It is possible that the majority of the treatment group data could be outside the region of overlap, if the model for participation predicted unusually well. If this were the case, then one might want to reestimate the propensity scores using an alternative set of X variables.

IV. Algorithm for a simple average nearest neighbor matching estimator (cross-sectional)

We next provide some pseudo-code for a program that implements the simpe average nearest neighbor matching estimator. This program assumes that treated group persons not in the overlapping support region have already been eliminated from the dataset, either by visual inspection of the histograms or by the trimming procedure based on the estimated densities, as described above.

```
#    Pseudo-code Program to implement simple, average, cross-sectional
#    matching estimator
#
#    Variable names
#    y1vec is the vector of y1 values for treatment group individuals

#    y0 vec is the vector of y2 values for comparison group individuals
#    p1vec is the vector of propensity scores for treatment group

#    p0vec is the vector of propensity scores for comparison group

#    n0 is the number of comparison group persons
#    n1 the number of treatment group persons
#    diffp is a vector of absolute value of the p differences


#    sortdiffp - the sorted diffp vector
#    neighbor is the number of neighbors to use in averaging
#    dist is the distance to the nearest neighbor
#    matchy0 is the vector that contains the matched y0 values


#    first read in the required data
call readdata(y1vec,y0vec,p1vec,p0vec,n1,n0)


# loop through vector of treatment group persons
# and construct a matches for each person.
```

```
neighor = 5
for i =1 to n1
        y1i = y1vec[i]

        p1i = p1vec[i]
        diffp = abs(p1vec[i]-p0vec)
        sortdiffp = sort(diffp)
        dist = diffp[neigbor]
        matchy0[i] = mean[y0vec[diffp<dist]]
end


# compute the average program impact as the mean over
#  the difference in y1 outcomes and the matched y0 outcomes


impact  = mean(y1vec-matchy0)
```

References

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

[2] Fan, J. (1993): "Local Linear Regression Smoothers and their Minimax Efficiencies," *The Annals of Statistics,* 21, 196-216.

[3] Heckman, J. and Todd, P. (1995): "Adapting Propensity Score Matching and Selection Models to Choice-based Samples," unpublished manuscript, University of Chicago.

[4] Heckman, J., H. Ichimura, J. Smith and P. Todd (1998): "Characterizing Selection Bias using Experimental Data" Econometrica, Vol. 66, September.

[5] Heckman, J., H. Ichimura, J. Smith and P. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program" with J. Heckman and H. Ichimura, Review of Economic Studies, Vol. 64(4), October.

[6] Heckman, J., H. Ichimura, J. Smith and P. Todd (1998): "Matching as an Econometric Evaluation Estimator" with J. Heckman and H. Ichimura, Review of Economic Studies, Vol. 65(2), April.

[7] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996): "A Brief Survey of Bandwidth Selection for Density Estimation" in *Journal of the American Statistical Association,* Vol. 91, No. 433, 401-407.

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