The Stata Journal (2001) **1**, Number 1, pp. 1–18

Implementing Matching Estimators for Average Treatment Effects in Stata

Alberto Abadie, Harvard University
David Drukker, Stata Corporation
Jane Leber Herr, UC Berkeley
Guido W. Imbens[1], UC Berkeley

Abstract.

In this paper we discuss the implementation of matching estimators for average treatment effects in Stata. The `match` command provides a number of options, including the number of matches, the choice of estimating the average effect for all units or only for the treated or control units, the distance metric, a bias adjustment, and a various options for the variance.

Keywords: *Average Treatment Effects, Matching, Exogeneity, Unconfoundedness, Ignorability.*

1 Introduction

In this paper we provide a brief introduction to matching estimators for average treatment effects and describe the new Stata program `match` that implements these estimators. The program implements nearest neighbor matching estimators for average treatment effects either for the overall sample or for the subsample of treated or control units. While simple matching estimators have been widely used in the program evaluation literature, `match` implements the specific matching estimators developed in Abadie and Imbens (2002), including their bias-corrected matching estimator. The procedure `match` allows individuals to be used as a match more than once. Compared to matching without replacement this generally lowers the bias but increases the variance.

While `match` provides many options for fine tuning the estimators, a key feature of the program is that it can be used with few decisions by the researcher. The default settings are generally sufficient for many applications. Although theoretically matching on multi-dimensional covariates can lead to substantial bias, in many cases the combination of the matching with the bias adjustment implemented in `match` leads to estimators with little remaining bias.

This paper draws heavily on the more theoretical discussion about matching by Abadie and Imbens (2002), and the survey by Imbens (2003). See also Cochran and Rubin (1973), Rosenbaum and Rubin (1985), Rubin and Thomas (1992), Rosenbaum (1995) and Heckman et al. (1998). The reader is referred to those papers for more background on, and formal derivations of, some of the properties of the estimators described here.

2 Framework

We are interested in estimating the average effect of a binary treatment on a continuous scalar outcome. For individual i, $i = 1, \ldots, N$, with all units exchangeable, let $(Y_i(0), Y_i(1))$ denote the two potential outcomes, i.e. $Y_i(0)$ is the outcome of individual i when she is not exposed to the treatment and $Y_i(1)$ is the outcome of individual i when she is exposed to the treatment. For instance, the treatment could be participation in a job training program and the outcome could be income or wages. If both $Y_i(0)$ and $Y_i(1)$ were observable, then the effect of the treatment on i would be $Y_i(1) - Y_i(0)$. The root of the problem is that only one of the two outcomes is observed. Let the observed outcome be denoted by Y_i:

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1, \end{cases}$$

where W_i, for $W_i \in \{0, 1\}$ indicates the treatment received,

Ignore, for the moment, the problem that we can only observe one of the two outcomes. What would we want to know, if we could observe both outcomes? In general we are interested in Average Treatment Effects (ATE's). Of most interest are the population and sample average treatment effects, PATE and SATE,

$$\tau^{pop} = \mathbb{E}[Y(1) - Y(0)], \quad \text{and} \quad \tau^{sample} = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i(1) - Y_i(0) \right).$$

Whether one is interested in the average treatment effect in the population (PATE) or the sample (SATE) does not affect the choice of estimator: the same matching estimator will estimate both. The choice between PATE and SATE will affect the large sample variance, and thus the estimator for the variance as will be discussed in more detail later. In general the variance for SATE is smaller than for PATE. See Abadie and Imbens (2002) and Imbens (2003) for more details.

Similarly we can define the population and sample average treatment effect for subpopulation of the treated, PATT and SATT,

$$\tau^{pop,t} = \mathbb{E}[Y(1) - Y(0)|W = 1] \quad \text{and} \quad \tau^{sample,t} = \frac{1}{N_1} \sum_{i|W_i=1} \left(Y_i(1) - Y_i(0) \right),$$

and the population and sample average treatment effect for the controls, PATC and SATC,

$$\tau^{pop,c} = \mathbb{E}[Y(1) - Y(0)|W = 0], \quad \text{and} \quad \tau^{sample,c} = \frac{1}{N_0} \sum_{i|W_i=0} \left(Y_i(1) - Y_i(0) \right),$$

where $N_1 = \sum_i W_i$ and $N_0 = \sum_i (1 - W_i)$ are the number of treated and control units, respectively.

Now consider the problem of estimating the untreated outcome, $Y_i(0)$, for person i with covariates X_i who was exposed to the treatment. If the decision to take the

treatment is "purely random" for individuals with similar values of the pretreatment variables or covariates, then one approach would be to use the average outcome of some similar individuals who were not exposed to the treatment. This is the basic idea behind matching estimators. For each i, matching estimators impute the missing outcome by finding other individuals in the data whose covariates are similar but who were exposed to the other treatment. It is the process of "matching" similar individuals who chose the opposite treatment that causes these estimators to be known a "matching estimators".

The intuition above needs to be stated more rigorously. To ensure that the matching estimators identify and consistently estimate the treatment effects of interest, we assume that assignment to treatment is independent of the outcomes, conditional on the covariates, and that the probability of assignment is bounded away from zero and one. For details on the regularity conditions see Abadie and Imbens (2002).

Assumption 2.1 *For all x in the support of X,*
(i) W is independent of $(Y(0), Y(1))$ conditional on $X = x$;
(ii) $c < \mathbb{P}(W = 1|X = x) < 1 - c$, for some $c > 0$.

Part (i) is a rigorous definition of the restriction that the choice of participation be "purely random" for similar individuals. This assumption is also known as unconfoundedness, or "selection on observables".

Part (ii) is an identification assumption. If all the individuals with a given covariate pattern chose the treatment, then there would be no observations on similar inviduals who chose not to accept the treatment.

In their seminal article, Rosenbaum and Rubin (1983) define the treatment to be "strongly ignorable" when both parts of Assumption 2.1 are valid. In addition, Rosenbaum and Rubin (1983) provide intuition for these two conditions in terms of how they make data from a non-randomized experiment analyzable as if it had come from a randomized experiment. These conditions are strong, and in many cases may not be satisfied. In various studies, however, researchers have found it useful to consider estimators based on these or similar conditions. In addition, Imbens (2003) argues that most studies will want to proceed under the Assumption 2.1 at some during the analysis.

3 Estimators

The unit level treatment effect is $\tau_i = Y_i(1) - Y_i(0)$. However, as discussed above, only one of the potential outcomes $Y_i(0)$ or $Y_i(1)$ is observed for each individual and the other is unobserved or missing. The matching estimators we consider impute the missing potential outcome by using average outcomes for individuals with "similar" values for the covariates.

Let $\|x\|_V = (x'Vx)^{1/2}$ be the vector norm with positive definite weight matrix V. We define $\|z - x\|_V$ to be the distance between the vectors x and z. Let $d_M(i)$ be the

distance from the covariates for unit i, X_i, to the Mth nearest match with the opposite treatment. Allowing for the possibility of ties, this is the distance such that strictly fewer than M units are closer to unit i than $d_M(i)$, and at least M units are as close as $d_M(i)$. Formally, $d_M(i) > 0$ is the real number satisfying:

$$\sum_{l:W_l=1-W_i} 1\Big\{\|X_l - X_i\|_V < d_M(i)\Big\} < M \quad \text{and} \quad \sum_{l:W_l=1-W_i} 1\Big\{\|X_l - X_i\|_V \leq d_M(i)\Big\} \geq M,$$

where $1\{\cdot\}$ is the indicator function, equal to one if the expression in brackets is true and zero otherwise.

Let $\mathcal{J}_M(i)$ denote the set of indices for the matches for unit i that are at least as close as the Mth match:

$$\mathcal{J}_M(i) = \Big\{ l = 1, \ldots, N \Big| W_l = 1 - W_i, \|X_l - X_i\|_V \leq d_M(i) \Big\}.$$

If there are no ties the number of elements in $\mathcal{J}_M(i)$ is M. In general it may be larger. Let the number of elements of $\mathcal{J}_M(i)$ be denoted by $\#\mathcal{J}_M(i)$. Finally, let $K_M(i)$ denote the sum of the weights unit i has as a match for other units, and $K'_M(i)$ the sum of the squared weights in the matches:

$$K_M(i) = \sum_{l=1}^{N} 1\{i \in \mathcal{J}_M(l)\} \cdot \frac{1}{\#\mathcal{J}_M(l)}. \tag{1}$$

$$K'_M(i) = \sum_{l=1}^{N} 1\{i \in \mathcal{J}_M(l)\} \cdot \left(\frac{1}{\#\mathcal{J}_M(l)}\right)^2. \tag{2}$$

Note that $\sum_i K_M(i) = N$, $\sum_{i:W_i=1} K_M(i) = N_0$ and $\sum_{i:W_i=0} K_M(i) = N_1$.

3.1 The Simple Matching Estimator

The first estimator that we consider, the simple matching estimator, uses the following approach to estimate the pair of potential outcomes. For each individual i, there are two potential outcomes, one is observed and the other is not. The observed outcome is its own estimate. The unobserved outcome is estimated by averaging the outcomes of the other most similar individuals who did choose this outcome:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \dfrac{1}{\#\mathcal{J}_M(i)} \displaystyle\sum_{l \in \mathcal{J}_M(i)} Y_l & \text{if } W_i = 1, \end{cases}$$

and

$$\hat{Y}_i(1) = \begin{cases} \dfrac{1}{\#\mathcal{J}_M(i)} \displaystyle\sum_{l \in \mathcal{J}_M(i)} Y_l & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases}$$

TABLE 3.2
A MATCHING ESTIMATOR WITH SEVEN OBSERVATIONS, $m = 1$

i	W_i	X_i	Y_i	$\mathcal{J}_1(i)$	$\hat{Y}_i(0)$	$\hat{Y}_i(1)$	$K_M(i)$
1	0	2	7	{5}	7	8	3
2	0	4	8	{4,6}	8	$7\frac{1}{2}$	1
3	0	5	6	{4,6}	6	$7\frac{1}{2}$	0
4	1	3	9	{1,2}	$7\frac{1}{2}$	9	1
5	1	2	8	{1}	7	8	1
6	1	3	6	{1,2}	$7\frac{1}{2}$	6	1
7	1	1	5	{1}	7	5	0

The simple matching estimator we implement is

$$\hat{\tau}_M^{sm} = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{Y}_i(1) - \hat{Y}_i(0)\Big) = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\cdot(1 + K_M(i))\cdot Y_i. \tag{3}$$

The simple matching estimator can easily be modified to estimate the average treatment effect for the treated:

$$\tau_M^{sm,t} = \frac{1}{N_1}\sum_{i:W_i=1}\Big(Y_i - \hat{Y}_i(0)\Big) = \frac{1}{N_1}\sum_{i=1}^{N}(W_i - (1 - W_i)\cdot K_M(i))\cdot Y_i. \tag{4}$$

or the average treatment effect for the controls:

$$\tau_M^{sm,c} = \frac{1}{N_0}\sum_{i:W_i=0}\Big(\hat{Y}_i(1) - Y_i\Big) = \frac{1}{N_0}\sum_{i=1}^{N}(W_i\cdot K_M(i) - (1 - W_i))\cdot Y_i. \tag{5}$$

3.2 Some useful examples

In this section, we use a small artificial data set to illustrate the concepts in the previous sections. The variables for the seven observations are presented in Table 3.2. In this table we also present the predicted values for the potential outcomes and the set of matches for each unit for the case with $m = 1$ (single match). Note that although we search for the single closest match, for some units there is a tie. Consider the second unit, a control unit with $X_i = 4$. Treated units 4 and 6, both with $X_i = 3$ are equally close and so the predicted outcome given the treatment for this unit is equal to the average of the outcomes for units 4 and 6, namely $(9 + 6)/2 = 7.5$.

Since this dataset is small enough, one can compute that the ATE for this data is .14285714. Now, let's compute the estimates using the `match` command. While the

complete syntax for `match` is given in section 5.1, here we note that the basic syntax of `match` is given by

`match` *depvar treatvar varlist* [*weight*] [`if` *exp*] [`in` *range*] [`, m(#)`
`tc({satt|satc})`

Options

`m(#)` specifies the number of matches. The default is 1 implying a single match. If one is estimating the average treatment effect, then any integer less than or equal to the minimum of the number of treated and controls in the sample can be choosen: $M \leq \min(N_0, N_1)$. If the average effect on the treated is specified, then the limit is the number of controls in the sample: $M \leq N_0$. If estimating the average effect on the controls, then the limit is the number of treated in the sample, $M \leq N_1$.

`tc()` specifies the estimand. By default, `match` estimates the the average treatment effect, ATE. Specifying `tc(att)` cause `match` to estimate the sample average treatment effect for the treated, ATT. Specifying `tc(atc)` causes `match` to estimate the sample average effect for the controls ATC.

In the output below we estimate the ATE for the artificial data set.

```
. use artificial
. match y w x
Matching estimator for the average treatment effect
                                        Number of obs          =         7
                                        Number of matches  (m) =         1
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATE |    .1428571   .9407699     0.15   0.879    -1.701018    1.986732
------------------------------------------------------------------------------
Matching variables: x
------------------------------------------------------------------------------
```

We note that the standard Stata output says that our hand calculation was correct.

Now let's consider an example with real data. We use a subset of the data used by Lalonde (1986). The particular subset is the one constructed by Dehejia and Wahba (1999) and described there in more detail. The data set is available in STATA format at http://emlab.berkeley.edu/users/imbens.

In this example, we are interested in the possible effect of participation in a job training program on individuals' earnings in 1978, In this dataset, participation in the job training program is recorded in the variable **t** and the 1978 earnings of the individuals in the sample are recorded in the variable **re78** in terms of 1978 dollars. The observable covariates that we use to identify similar individuals are given in Figure 1.

Figure 1:

Variable Description	Variable Name
age	**age**
years of education	**educ**
indicator for afro-american	**black**
indicator for hispanic-american	**hisp**
indicator for married	**married**
indicator for more than grade school but less than high school education	**nodegree**
earnings in 1974(in thousands of 1978 $)	**re74**
earnings in 1975(in thousands of 1978 $)	**re75**
indicator for unemployed in 1974	**u74**
indicator for unemployed in 1975	**u75**

In the output below we estimate the ATT using this data.

```
. use lalonde
. match re78 t age educ black hisp married re74 re75 u74 u75, m(4)
Matching estimator for the average treatment effect
Weighting matrix: inverse variance          Number of obs          =       445
                                            Number of matches  (m) =         4
```

re78	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
SATE	1.903326	.7202149	2.64	0.008	.4917308	3.314921

```
Matching variables: age educ black hisp married re74 re75 u74 u75
```

Since 1978 earnings are recorded in thousands of 1978 dollars, this output implies that for the individuals in our sample, the average effect of participing in the job training program is an increase an individual's 1978 earnings by $1,903.

The population and sample average treatment effects are useful for answering different questions. For instance, the SATE is useful for judging whether or not this particular job training program was a success. In contrast, if we were considering launching another job training program in which we would obtain other sample from the same population, the PATE would be more useful. For the specification at hand we conclude that the sample average is significantly different from zero at the 1% level. Since the standard error of the SATE underestimates the standard error of the PATE, it is possible that the PATE might not be significantly different from zero at either the 5% nor the 1% level.

add results for PATE

We also need to point out that the size, as well as the statistical significance, is

important in interpreting the results in most treatment effects studies. For instance, if our earnings data was in terms of dollars instead of thousands of dollars, our results would indicate a statistically significant but economically unimportant impact of the job training program on the individuals in the current sample.

As discussed in Imbens (2003) and Heckman et al. (1998) the effect of the treatment on the subpopulation of treated units is frequently more important than the effect on the population as a whole when evaluating the importance of narrowly targeted labor market programs. For instance, when evaluating the importance of a program aimed at increasing the post-graduation earnings of youth from bad neighborhoods, the potential impact of the program on youth from good neighborhoods is not relevant.

In the output below we use `match` to estimate SATT using our extract from the Lalonde data.

```
. match re78 t age educ black hisp married re74 re75 u74 u75, tc(satt) m(4)
Matching estimator for the average treatment effect for the treated
Weighting matrix: inverse variance          Number of obs          =       445
                                            Number of matches  (m) =         4
```

re78	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
SATT	1.994622	.7127286	2.80	0.005	.5976995	3.391544

```
Matching variables: age educ black hisp married re74 re75 u74 u75
```

The output indicates that effect of the job training program on the participants in this sample is statistically different from zero.

In most of the examples that we do using the dataset, we use 4 matches. We chose 4 matches because it seemed to offer the benefit of not relying on too little information without incorporating observations that are not sufficiently similar. Like all smoothing parameters, the final inference can depend on the choice of the number of matches. For instance, in the output below we show that relying on a single match makes the ATT become statistically insignificant.

```
. match re78 t age educ black hisp married re74 re75 u74 u75, tc(satt)
Matching estimator for the average treatment effect for the treated
Weighting matrix: inverse variance          Number of obs          =       445
                                            Number of matches  (m) =         1
```

re78	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
SATT	1.223154	.8529323	1.43	0.152	-.4485624	2.894871

```
Matching variables: age educ black hisp married re74 re75 u74 u75
```

3.3 The Bias Corrected Matching Estimator

The simple matching estimator will be biased in finite samples when the matching is not exact. Abadie and Imbens (2002) show that with k continuous covariates the estimator will have a term corresponding to the matching discrepancies (the difference in covariates between matched units and their matches) that will be of the order $O_p(N^{-1/k})$. In practice one may therefore attempt to remove some of this bias term that remains after the matching. The bias-corrected matching estimator adjusts the difference within the matches for the differences in their covariate values. The adjustment is based on an estimate of the two regression functions $\mu_w(x) = \mathbb{E}[Y(w)|X = x]$. Following Rubin (1973) and Abadie and Imbens (2002) we approximate these regression functions by linear functions and estimate them using least squares on the matched observations.

First suppose we are estimating the average treatment effect. In this case we estimate the regression functions using only the data in the matched sample:

$$\hat{\mu}_w(x) = \hat{\beta}_{w0} + \hat{\beta}'_{w1}x,$$

for $w = 0, 1$, where

$$(\hat{\beta}_{w0}, \hat{\beta}_{w1}) = \text{argmin} \sum_{i:W_i=w} K_M(i) \cdot \left(Y_i - \beta_{w0} - \beta'_{w1}X_i\right)^2. \tag{6}$$

If we are interested in estimating the ATT, we only need estimate the regression function for the controls, $\mu_0(x)$ and if we are interested in ATC we only need estimate the regression function for the treated, $\mu_1(x)$.

We weight the observations in these regressions by $K_M(i)$, the number of times the unit is used as a match, because the weighted empirical distribution is closer to the distribution of covariates in which we are ultimately interested. For example, when we are estimating the SATT, control units that are not used as matches have potentially very different covariate values than the treated units we are trying to match. Hence using these controls to predict outcomes for the treated units can lead to results that can be very sensitive to the exact specification applied.

Given the estimated regression functions, we predict the missing potential outcomes as:

$$\tilde{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \dfrac{1}{\#\mathcal{J}_M(i)} \displaystyle\sum_{l \in \mathcal{J}_M(i)} (Y_l + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_l)) & \text{if } W_i = 1, \end{cases} \tag{7}$$

and

$$\tilde{Y}_i(1) = \begin{cases} \dfrac{1}{\#\mathcal{J}_M(i)} \displaystyle\sum_{l \in \mathcal{J}_M(i)} (Y_l + \hat{\mu}_1(X_i) - \hat{\mu}_1(X_l)) & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1, \end{cases} \tag{8}$$

with corresponding estimator for the ATE:

$$\hat{\tau}_M^{bcm} = \frac{1}{N}\sum_{i=1}^{N}\Big(\tilde{Y}_i(1) - \tilde{Y}_i(0)\Big). \tag{9}$$

The bias-adjusted matching estimators for ATT and ATC are

$$\hat{\tau}_M^{bcm,t} = \frac{1}{N_1}\sum_{i:W_i=1}\Big(Y_i - \tilde{Y}_i(0)\Big), \quad \text{and } \hat{\tau}_M^{bcm,c} = \frac{1}{N_0}\sum_{i:W_i=0}\Big(\tilde{Y}_i(1) - Y_i\Big).$$

Now let's return to our extract from the Lalonde data. In this example, we estimate the ATT with bias-adjustment.

```
. match re78 t age educ black hisp married re74 re75 u74 u75, tc(satt)    ///
>         m(4) bias(bias)
Matching estimator for the average treatment effect for the treated
Weighting matrix: inverse variance          Number of obs          =       445
                                            Number of matches  (m) =         4
```

re78	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
SATT	1.838424	.7160904	2.57	0.010	.434913	3.241936

```
Matching variables: age educ black hisp married re74 re75 u74 u75
Bias-adj variables: age educ black hisp married re74 re75 u74 u75
```

The above output indicates that while bias-adjustment reduces the size of the estimated SATE, it does not change our previous conclusion that treatment had an effect on its participants that is significant at the 5% but quite not at the 1% level.

4 Variance Estimation

In this section we describe the variance estimators implemented in `match`. First, it is important to note that the bias-adjustment does not affect the form of the estimator for the variance, although it may affect the numerical value. For the variance it does matter whether one is interested in the sample of population average treatment effect (or the average effect for the treated or controls). In addition there is an option for a robust estimator that allows for heteroskedasticity. Note that in general there is no theoretical justification for bootstrapping methods for variance estimation for matching estimators. For details on the theoretical justification for the various variance estimators see Abadie and Imbens (2002).

First the estimator for the variance of the sample average treatment effect (SATE):

$$\hat{V}^{sample} = \frac{1}{N^2}\sum_{i=1}^{N}\Big(1 + K_M(i)\Big)^2 \cdot \hat{\sigma}^2_{W_i}(X_i). \tag{10}$$

Below we return to the question of estimating the conditional error variance $\sigma_w^2(x)$. Similarly the variance for the estimator for SATT is

$$\hat{V}^{sample,t} = \frac{1}{N_1^2} \sum_{i=1}^{N} \left(W_i - (1-W_i)\cdot K_M(i)\right)^2 \cdot \sigma_{W_i}^2(X_i), \tag{11}$$

and for SATC,

$$V^{sample,c} = \frac{1}{N_0^2} \sum_{i=1}^{N} \left(W_i \cdot K_M(i) - (1-W_i)\right)^2 \cdot \sigma_{W_i}^2(X_i). \tag{12}$$

As an estimator for the variance of the matching estimator for the population average treatment effect we use:

$$\hat{V}^{POP} = \frac{1}{N^2} \sum_{i=1}^{N} \left[\left(\hat{Y}_i(1) - \hat{Y}_i(0) - \hat{\tau}\right)^2 + \left(K_M^2(i) + 2K_M(i) - K'_M(i)\right) \cdot \hat{\sigma}_{W_i}^2(X_i) \right] \tag{13}$$

In large samples this will be at least as large as the estimator for the variance of the matching estimator for SATE, with the difference an estimator for $\sum_i (\mu_1(X_i) - \mu_0(X_i))^2/N^2$. In small samples it need not be larger. In practice we therefore take the maximum of $\hat{V}^{pop}$ and $\hat{V}^{sample}$ as the estimator for the variance of the estimator for the PATE.

As an estimator for the variance of the matching estimator for the population average treatment effect for the treated we use:

$$\hat{V}^{POP,t} = \frac{1}{N_1^2} \sum_{i=1}^{N} \left[W_i \cdot \left(Y_i(1) - \hat{Y}_i(0) - \hat{\tau}^t\right)^2 + (1-W_i) \cdot \left(K_M^2(i) - K'_M(i)\right) \cdot \hat{\sigma}_{W_i}^2(X_i) \right]. \tag{14}$$

Finally, as an estimator for the variance of the matching estimator for the population average treatment effect for the controls we use:

$$\hat{V}^{POP,c} = \frac{1}{N_0^2} \sum_{i=1}^{N} \left[(1-W_i) \cdot \left(\hat{Y}_i(1) - Y_i(0) - \hat{\tau}^c\right)^2 + W_i \cdot \left(K_M^2(i) - K'_M(i)\right) \cdot \hat{\sigma}_{W_i}^2(X_i) \right]. \tag{15}$$

Estimating these variances requires estimation of the conditional outcome variance $\sigma_w^2(x)$. The matching program offers two options, either assuming this variance $\sigma_w^2(x)$ is constant for both treatment groups and all values of the covariates, or not.

4.1 Assuming a Constant Treatment Effect and Homoskedasticity

Here we discuss estimating the variance under two assumptions. First, the assumption that the treatment effect $Y_i(1) - Y_i(0)$ is constant. Second, the assumption that the

conditional variance of $Y_i(w)$ given X_i does not vary with either the covariates x nor the treatment w. In this case we can therefore estimate the outcome variance σ^2 as

$$\hat{\sigma}^2 = \frac{1}{2N}\sum_{i=1}^{N}\left(\frac{1}{\mathcal{J}_M(i)}\sum_{l\in\mathcal{J}_M(i)}\left(W_i\cdot(Y_i-Y_l-\hat{\tau})+(1-W_i)\cdot(Y_l-Y_i-\hat{\tau})\right)^2\right). \quad (16)$$

We then substitute $\hat{\sigma}^2$ for $\hat{\sigma}^2_{W_i}(X_i)$ into the relevant variance formula (10) or (13). For ATT we estimate σ^2 as

$$\hat{\sigma}^2_t = \frac{1}{2N_1}\sum_{i:W_i=1}\left(\frac{1}{\mathcal{J}_M(i)}\sum_{l\in\mathcal{J}_M(i)}(Y_i-Y_l-\hat{\tau})^2\right),$$

and substitute that for $\hat{\sigma}^2_{W_i}(X_i)$ into equation (11) or (14). Finally, for ATC we estimate σ^2 as

$$\hat{\sigma}^2_c = \frac{1}{2N_0}\sum_{i:W_i=0}\left(\frac{1}{\mathcal{J}_M(i)}\sum_{l\in\mathcal{J}_M(i)}(Y_l-Y_i-\hat{\tau})^2\right),$$

and substitute that for $\hat{\sigma}^2_{W_i}(X_i)$ into (12) or (15).

4.2 Variance Estimation Allowing for Heteroskedasticity

If $\sigma^2_w(x)$ difffers by w and x, one needs to estimate it for all sample points. In `match` this is implemented using a second matching procedure, now matching treated units to treated units and control units to control units. Define $d'_M(i)$ as the distance to the Mth closest unit with the same treatment indicator. Formally:

$$\sum_{l:W_l=W_i,l\neq i}1\Big\{\|X_l-X_i\|_V<d'_M(i)\Big\}<M \quad\text{and}\quad \sum_{l:W_l=W_i,l\neq i}1\Big\{\|X_l-X_i\|_V\leq d'_M(i)\Big\}\geq M.$$

Let $\mathcal{J}'_M(i)$ denote the set of indices for the first M matches for unit i:

$$\mathcal{J}'_M(i)=\Big\{j=1,\ldots,N\Big|W_j=W_i,j\neq i,\|X_j-X_i\|_V\leq d'_M(i)\Big\},$$

where the number of elements of $\mathcal{J}'_M(i)$ be denoted by $\#\mathcal{J}'_M(i)$. Then we estimate the conditional variance as the sample variance in this set augmented with the outcome for unit i itself $\mathcal{J}'_M(i)\cup\{i\}$:

$$\tilde{\sigma}^2_{W_i}(X_i)=\frac{1}{\#\mathcal{J}'_M(i)}\sum_{j\in(\mathcal{J}'_M(i)\cup\{i\})}\Big(Y_j-\bar{Y}_{\mathcal{J}'_M(i)\cup\{i\}}\Big)^2, \quad (17)$$

where

$$\bar{Y}_{\mathcal{J}'_M(i)\cup\{i\}}=\frac{1}{\#\mathcal{J}'_M(i)+1}\sum_{j\in(\mathcal{J}'_M(i)\cup\{i\})}Y_j,$$

is the average outcome in this set. The overall variance is estimated by plugging this unit-level variance estimate into the relevant variance expression (10)-(15)

Now we return to our Lalonde data extract. In the output below we re-estimate the ATT, but this time we estimate the standard error allowing for heteroskedasticity. We specify 4 matches in estimating the conditional variance functions for the same reason that we allow 4 matches in estimating the conditional mean functions; given our data 4 matches seems to include sufficient information without matching unlike individuals.

```
. match re78 t age educ black hisp married re74 re75 u74 u75, tc(satt)    ///
>         m(4) bias(bias) robust(4)

Matching estimator for the average treatment effect for the treated
Weighting matrix: inverse variance          Number of obs          =       445
                                            Number of matches  (m) =         4
                                            Number of matches,
                                              robust std. err. (h) =         4
------------------------------------------------------------------------------
        re78 |      Coef.   robust
             |             Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATT |   1.838424   .7526339     2.44   0.015      .363289     3.31356
------------------------------------------------------------------------------
Matching variables: age educ black hisp married re74 re75 u74 u75
Bias-adj variables: age educ black hisp married re74 re75 u74 u75
------------------------------------------------------------------------------
```

The output indicates that our estimated SATT remains at the 5% level but not at the 1% level when the standard error is estimated under these weaker conditions. Thus, in this sample the job training program appears to have had a significant impact on the 1978 earnings of it participants.

5 The `match` command

Here we discuss the formal syntax of the `match` command.

5.1 Syntax of `match`

The basic syntax is as follows:

`match` *depvar treatvar varlist* [*weight*] [`if` *exp*] [`in` *range*] [`, m(#)`
`tc({att|atc}) metric(maha|`*matrix*`) biasadj(bias |`$varlist_{adj}$`) robust(#)`
`population exact(`$varlist_{exa}$`) keep(`*filename*`) replace`

`pweights` are allowed. See the Stata 8 User's Guide [U] **14.1.6** for more information on weights. See 5.2 for how match handles weights.

Description

`match` estimates the sample average treatment effect, the sample average treatment effect for the treated and sample average treatment effect for the controls and their standard errors. The *depvar* is the outcome variable. The `treatvar` is a binary variable treatment indicator. The *varlist* specifies the covariates that are to be used in the matching.

Options

`m`(#) specifies the number of matches. The default is 1 implying a single match. If one is estimating the average treatment effect, then any integer less than or equal to the minimum of the number of treated and controls in the sample can be choosen: $M \leq \min(N_0, N_1)$. If the average effect on the treated is specified, then the limit is the number of controls in the sample: $M \leq N_0$. If estimating the average effect on the controls, then the limit is the number of controls in the sample.

In practice one should typically choose a fairly small number. In simulations in Abadie and Imbens (2002) using four matches was found to perform well in terms of mean-squared error.

`tc()` specifies the estimand. By default, `match` estimates the the average treatment effect, ATE. Specifying `tc(att)` cause `match` to estimate the sample average treatment effect for the treated, ATT. Specifying `tc(atc)` causes `match` to estimate the average effect for the controls ATC.

`biasadj(`[*varlist*`)` specifies that the bias-corrected matching estimator is to be used. By default, `match` uses the simple matching estimator. The first alternative, `biasadj(bias)`, uses the bias-corrected matching estimator using the same set of covariates as is used in the matching, entering linearly in the regression function. With many covariates one may wish to use only some of the covariates for covariate adjustment, thus the second alternative, `biasadj`(*v*arlist), is to use the bias-corrected matching estimator with a set of covariates distinct from the set used in matching.

`metric()` specifies the metric for measuring the distance between two vectors of covariates. Letting $\|x\|_V = (x'Vx)^{1/2}$ be the vector norm with positive definite weight matrix V, we define $\|z - x\|_V$ to be the distance between the vectors x and z. There are three choices for V. First, by default, V is the diagonal matrix constructed by putting the inverses of the variances of the covariates on the diagonal. Second, specifing `metric(maha)` causes `match` to use to use the Mahalanobis metric in which $V = S^{-1}$, where S is the sample covariance matrix of the of the covariates. Third, specifying `metric(`*matrixname*`)` causes `match` to use the matrix `matrixname`. This third option allows the user to flexibly choose any weight matrix.

`robust`(#$_v$) specifies that `match` estimate heteroskedasticty consistent standard errors using #$_v$ matches. The number of matches used in estimating the standard error, #, does not need to be the same as the number of matches used in estimating the treatment effect itself. By default, `match` uses the homoskedastic/constant variance

estimator.

`population` specifies whether the estimand is a sample or population average treatment effect. This only affects the choice of estimator for the variance. By default, `match` estimates the the sample average treatment effect, SATE, SATT, or SATC, using one of the variance estimators from (10)-(12). Specifying `pop` cause `match` to estimate the variance for the population average treatment effect, PATE, PATT, or PATC, using one from (13)-(15).

`exact`(*varlist*$_{exa}$) allows the user to specify some covariates that `match` will attemp to match exactly on. This option is useful if the user wants to ensure that all matches are exact on some variables. Most of the time this will be a single discrete covariate, e.g., gender, or employment status. If this option is used this list of variables is added to the full set of matching variables. A new weight matrix is then calculated with in the top left submatrix the original weight matrix, zeros on the off-diagonal parts and in the bottom right submatrix a diagonal matrix with the inverse of the variances of the variables in *varlist* multiplied by 1,000. As long as there are not too many covariates in this list, as long as the original weight matrix is not too far from the inverse variances, and as long as it is feasible, this will lead the program to match exactly on these variables. The program output will indicate what percentage of the matches are exact on these covariates. If one uses this option with a continuous covariate the result will be that the program attempts to match as well as possible on this covariate, without leading to exact matches.

5.2 How `match` handles weights

The procedure `match` allows probability weights. An observation represents a part of the population proportional to its weight. For example, if all observations have weight 1, other than observation i which has weight 2, the estimates are identical to those that would be obtained by using the unweighted estimator on an artificial sample created from the original sample with observation i duplicated once. The standard errors are updated to take account of the weighting. The weights are allowed to be non-integer, but have to be non-negative.

Formally, with the weight for individual i equal to ω_i, the estimator is calculated as follows. The distance $d_M(i) > 0$ is modified to ensure that the sum of the weights of the matches adds up to M:

$$\sum_{l:W_l=1-W_i} \omega_l \cdot 1\Big\{\|X_l - X_i\|_V < d_M^\omega(i)\Big\} < M$$

and

$$\sum_{l:W_l=1-W_i} \omega_l \cdot 1\Big\{\|X_l - X_i\|_V \leq d_M^\omega(i)\Big\} \geq M.$$

The definition of the set $\mathcal{J}_M(i)$ is unchanged. The estimated potential outcomes are

now:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \dfrac{1}{\sum_{l \in \mathcal{J}_M(i)} \omega_l} \displaystyle\sum_{l \in \mathcal{J}_M(i)} \omega_l \cdot Y_l & \text{if } W_i = 1, \end{cases}$$

and

$$\hat{Y}_i(1) = \begin{cases} \dfrac{1}{\sum_{l \in \mathcal{J}_M(i)} \omega_l} \displaystyle\sum_{l \in \mathcal{J}_M(i)} \omega_l \cdot Y_l & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases}$$

The $K_m(i)$ and $K'_M(i)$ are changed from equations (1) and (2) to

$$K_M^{\omega}(i) = \sum_{l=1}^{N} \omega_l \cdot 1\{i \in \mathcal{J}_M(l)\} \cdot \frac{\omega_i}{\sum_{j \in \mathcal{J}_M(l)} \omega_j}. \tag{18}$$

and

$$K_M^{\omega\prime}(i) = \sum_{l=1}^{N} \omega_l \cdot 1\{i \in \mathcal{J}_M(l)\} \cdot \left(\frac{\omega_i}{\sum_{j \in \mathcal{J}_M(l)} \omega_j} \right)^2. \tag{19}$$

The simple matching estimator is now defined as:

$$\hat{\tau}_M^{\omega,sm} = \frac{1}{\sum_{i=1}^{N} \omega_i} \sum_{i=1}^{N} \omega_i \cdot \left(\hat{Y}_i(1) - \hat{Y}_i(0) \right),$$

with a similar modification for ATT and ATC. The bias correction is as described above, with the one modification that in equation (6) the weighted version $K_M^{\omega}(i)$ is used.

The variance formulas change from (10)-(15) to

$$\hat{V}^{sample} = \frac{1}{(\sum_i \omega_i)^2} \sum_{i=1}^{N} \Big(\omega_i + K_M(i) \Big)^2 \cdot \hat{\sigma}^2_{W_i}(X_i).$$

$$\hat{V}^{sample,t} = \frac{1}{(\sum_i W_i \omega_i)^2} \sum_{i=1}^{N} \left(W_i \cdot \omega_i - (1 - W_i) \cdot K_M(i) \right)^2 \cdot \sigma^2_{W_i}(X_i),$$

$$V^{sample,c} = \frac{1}{(\sum_i (1 - W_i)\omega_i)^2} \sum_{i=1}^{N} \left(W_i \cdot K_M(i) - (1 - W_i)\omega_i \right)^2 \cdot \sigma^2_{W_i}(X_i).$$

$$\hat{V}^{POP} = \frac{1}{(\sum_i \omega_i)^2} \sum_{i=1}^{N} \left[\omega_i (\Big(\hat{Y}_i(1) - \hat{Y}_i(0) - \hat{\tau} \Big)^2 + \big((K_M^{\omega}(i))^2 + 2K_M^{\omega}(i) - K_M^{\omega\prime}(i) \big) \cdot \hat{\sigma}^2_{W_i}(X_i) \right]$$

$$\hat{V}^{POP,t} = \frac{1}{(\sum_i W_i \omega_i)^2} \sum_{i=1}^{N} \left[W_i \omega_i \cdot \Big(Y_i(1) - \hat{Y}_i(0) - \hat{\tau}^t \Big)^2 + (1 - W_i) \cdot \big((K_M^{\omega}(i))^2 - K_M^{\omega\prime}(i) \big) \cdot \hat{\sigma}^2_{W_i}(X_i) \right].$$

$$\hat{V}^{POP,c} = \frac{1}{(\sum_i (1 - W_i)\omega_i)^2}$$

$$\times \sum_{i=1}^{N} \left[(1-W_i)\omega_i \cdot \left(\hat{Y}_i(1) - Y_i(0) - \hat{\tau}^c \right)^2 + W_i \cdot \left((K_M^{\omega}(i))^2 - K_M^{\omega\prime}(i) \right) \cdot \hat{\sigma}^2_{W_i}(X_i) \right].$$

For $\hat{\sigma}^2_{W_i}(X_i)$ we modify the earlier estimators. In the homoskedastic case the error variance estimator is changed from (16) to

$$\hat{\sigma}^2 = \frac{1}{2\sum_i \omega_i} \sum_{i=1}^{N} \left(\frac{\omega_i}{\sum_{l \in \mathcal{J}_M(i)} \omega_l} \sum_{l \in \mathcal{J}_M(i)} \omega_l \left(W_i \cdot (Y_i - Y_l - \hat{\tau}) + (1-W_i) \cdot (Y_l - Y_i - \hat{\tau}) \right)^2 \right),$$

with similar modifications for the ATT and ATC cases.

In the heteroskedastic case the conditional variance $\sigma^2_w(x)$ is estimated using matching with some modification for the weights. First, $d^{\omega\prime}_M(x)$ is defined as:

$$\sum_{l:W_l=W_i, l\neq i} \omega_l 1\Big\{ \|X_l - X_i\|_V < d'_M(i) \Big\} < M,$$

and

$$\sum_{l:W_l=W_i, l\neq i} \omega_l 1\Big\{ \|X_l - X_i\|_V \leq d'_M(i) \Big\} \geq M,$$

with again the set $\mathcal{J}'_M(i)$ containing the indices for the matches for unit i:

$$\mathcal{J}'_M(i) = \Big\{ j = 1, \ldots, N \Big| W_j = W_i, j \neq i, \|X_j - X_i\|_V \leq d^{\omega\prime}_M(i) \Big\}.$$

Then we estimate the conditional variance $\sigma^2_{W_i}(X_i)$ as the sample variance in this set augmented with the outcome for unit i itself $\mathcal{J}'_M(i) \cup \{i\}$, taking account of the weights:

$$\tilde{\sigma}^2_{W_i}(X_i) = \sum_{j \in (\mathcal{J}'_M(i) \cup \{i\})} \omega_j \Big(Y_j - \bar{Y}_{\mathcal{J}'_M(i)\cup\{i\}} \Big)^2 \Big/ \sum_{j \in (\mathcal{J}'_M(i) \cup \{i\})} \omega_j, \tag{20}$$

where

$$\bar{Y}_{\mathcal{J}'_M(i)\cup\{i\}} = \sum_{j \in (\mathcal{J}'_M(i) \cup \{i\})} \omega_j Y_j \Big/ \sum_{j \in (\mathcal{J}'_M(i) \cup \{i\})} \omega_j,$$

is the average outcome in this set. The overall variance is estimated by plugging this unit-level variance estimate into the relevant variance expression (10)-(15)

6 References

Abadie, A. and G. Imbens. 2002. Simple and Bias-Corrected Matching Estimators. Tech. rep., Department of Economics, UC Berkeley. http://emlab.berkeley.edu/users/imbens/.

Cochran, W. and D. Rubin. 1973. Controlling Bias in Observational Studies: A Review. *Sankhya* 35: 417–46.

Dehejia, R. H. and S. Wahba. 1999. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association* 94: 1053–1062.

Heckman, J. J., H. Ichimura, and P. Todd. 1998. Matching as an Econometric Evaluation Estimator. *Review of Economic Studies* 65: 261–294.

Imbens, G. 2003. Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review. Tech. rep., Department of Economics, UC Berkeley. http://emlab.berkeley.edu/users/imbens/.

Lalonde, R. J. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review* 76: 604–620.

Rosenbaum, P. 1995. *Observational Studies*. New York: Springer–Verlag.

Rosenbaum, P. and D. Rubin. 1983. Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70: 41–55.

—. 1985. Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *American Statistician* 39: 33–38.

Rubin, D. 1973. The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies. *Biometrics* 29: 185–203.

Rubin, D. and N. Thomas. 1992. Affinely Invariant Matching Methods with Ellipsoidal Distributions. *Annals of Statistics* 20: 1079–1093.