Implementing matching estimators for average treatment effects in Stata

Alberto Abadie[1]   David Drukker   Jane Leber Herr   Guido W. Imbens[2]
Harvard University   StataCorp   UC Berkeley   UC Berkeley

Abstract.   This paper presents an implementation of matching estimators for average treatment effects in Stata. The `nnmatch` command allows you to estimate the average effect for all units or only for the treated or control units; to choose the number of matches; to specify the distance metric; to select a bias adjustment; and to use heteroskedastic-robust variance estimators.

Keywords: st0072, nnmatch, average treatment effects, matching, exogeneity, unconfoundedness, ignorability

1   Introduction

This paper provides a brief introduction to matching estimators for average treatment effects and describes the new Stata command `nnmatch` which implements these estimators. The command implements nearest-neighbor matching estimators for average treatment effects for either the overall sample or a subsample of treated or control units. While simple matching estimators have been widely used in the program evaluation literature, `nnmatch` implements the specific matching estimators developed in Abadie and Imbens (2002), including their bias-corrected matching estimator. Following Abadie and Imbens (2002), `nnmatch` allows individual observations to be used as a match more than once. Compared with matching without replacement, this method generally lowers the bias but increases the variance.

While `nnmatch` provides many options for fine-tuning the estimators, a key feature of the program is that it requires few decisions by the researcher. The default settings are generally sufficient for many applications. Although theoretically matching on multidimensional covariates can lead to substantial bias, the matching approach combined with the bias adjustment implemented in `nnmatch` often leads to estimators with little remaining bias.

This paper draws heavily on the more theoretical discussion about matching in Abadie and Imbens (2002) and the survey by Imbens (2004). See also Cochran and Rubin (1973), Rosenbaum and Rubin (1985), Rubin and Thomas (1992), Rosenbaum (1995), Heckman, Ichimura, and Todd (1998), and Wooldridge (2002) for more background.

---

st0072

2 Framework

We are interested in estimating the average effect of a binary treatment on a continuous or discrete 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: $Y_i(1)$ is the outcome of individual i when exposed to the treatment, and $Y_i(0)$ is the outcome of individual i when not 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, the effect of the treatment on unit i would be directly observable as $Y_i(1) - Y_i(0)$. We could then use this information for the full sample N to either estimate the population average treatment effect (PATE)

$$\tau^{\mathrm{pop}} = E\big\{Y(1) - Y(0)\big\}$$

or calculate the sample average treatment effect (SATE)

$$\tau^{\mathrm{sample}} = \frac{1}{N} \sum_{i=1}^{N} \big\{Y_i(1) - Y_i(0)\big\}$$

Whether the average treatment effect in the population (PATE) or the sample (SATE) is of interest depends on the research question. Inference on the PATE is made with respect to another sample drawn from the same population; inference on the SATE is made conditional on the sample at hand. (Although this does not affect the choice of estimator, for this reason, the standard error of the estimated PATE is generally larger than the standard error of the estimated SATE.) See Imbens (2004) for a discussion of the difference between these two estimands.

An example may help to illustrate the meaning of the two estimands. While the SATE is useful for judging how a job-training program has affected a particular group of participants, the PATE can be used to evaluate whether another group of participants drawn from the same population is likely to benefit from the program. We can also define the population and sample average treatment effect for the subpopulation of the treated, PATT and SATT,

$$\tau^{\mathrm{pop},t} = E\big\{Y(1) - Y(0)|W = 1\big\} \ \ \text{and} \ \ \tau^{\mathrm{sample},t} = \frac{1}{N_1} \sum_{i|W_i=1} \big\{Y_i(1) - Y_i(0)\big\}$$

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

$$\tau^{\mathrm{pop},c} = E\big\{Y(1) - Y(0)|W = 0\big\} \ \ \text{and} \ \ \tau^{\mathrm{sample},c} = \frac{1}{N_0} \sum_{i|W_i=0} \big\{Y_i(1) - Y_i(0)\big\}$$

where W_i, for $W_i \in \{0, 1\}$ indicates the treatment received, and $N_1 = \sum_i W_i$ and $N_0 = \sum_i (1 - W_i)$ are the number of treated and control units, respectively. See Abadie and Imbens (2002) and Imbens (2004) for more details.

When we estimate average treatment effects, only one of the two outcomes is observed. Let the observed outcome be denoted by Y_i:

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

To estimate the average treatment effect, we will estimate the unobserved potential outcome for each observation in the sample. Consider 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, we could use the average outcome of some similar individuals who were not treated to estimate the untreated outcome. 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. The process of matching similar individuals who chose the opposite treatment causes these estimators to be known as "matching estimators".

To ensure that the matching estimators identify and consistently estimate the treatment effect 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 all x in the support of X,
> (i) W is independent of $(Y(0), Y(1))$ conditional on $X = x$;
> (ii) $c < \Pr(W = 1 | X = x) < 1 - c$, 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, there would be no observations on similar individuals who chose not to accept the treatment against which to compare them.

For details on these regularity conditions, see Abadie and Imbens (2002).

In their seminal article, Rosenbaum and Rubin (1983) define the treatment to be "strongly ignorable" when both parts of assumption 2 are true. They also provide an explanation for these two conditions, showing how "strong ignorability" allows you to analyze a nonrandomized experiment 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. Imbens (2004) argues that most studies will want to proceed under assumption 2 at some stage of the analysis.

3   Estimators

For each observation i, 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, 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.

Considering the set of observed covariates for an individual i, X_i, let $\|x\|_V = (x'Vx)^{1/2}$ be the vector norm with positive definite matrix V. We define $\|z - x\|_V$ as the distance between the vectors x and z, where z represents the covariate values for a potential match for observation i. 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, at this distance 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, which is 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 but 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 number of times i is used as a match for all observations l of the opposite treatment group, each time weighted by the total number of matches for observation l. Let $K'_M(i)$ represent a comparable measure in which the square of the number of matches is used as the weight.

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

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

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:

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

and

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

Namely, given that only one potential outcome is observed for each observation i, the observed outcome $Y_i = Y_i(0)$ or $Y_i(1)$ represents one potential outcome. The unobserved outcome is estimated by averaging the observed outcomes for the observations l of the opposite treatment group that are chosen as matches for i.

Using these estimates of the potential outcomes, the simple matching estimator is

$$
\widehat{\tau}_M^{\mathrm{sm}} = \frac{1}{N} \sum_{i=1}^{N} \left\{ \widehat{Y}_i(1) - \widehat{Y}_i(0) \right\} = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1)\left\{1 + K_M(i)\right\} Y_i
$$

This estimator can be modified easily to estimate the average treatment effect for the treated

$$
\tau_M^{\mathrm{sm},t} = \frac{1}{N_1} \sum_{i:W_i=1} \left\{ Y_i - \widehat{Y}_i(0) \right\} = \frac{1}{N_1} \sum_{i=1}^{N} \left\{ W_i - (1 - W_i)K_M(i) \right\} Y_i
$$

or the average treatment effect for the controls

$$
\tau_M^{\mathrm{sm},c} = \frac{1}{N_0} \sum_{i:W_i=0} \left\{ \widehat{Y}_i(1) - Y_i \right\} = \frac{1}{N_0} \sum_{i=1}^{N} \left\{ W_i K_M(i) - (1 - W_i) \right\} Y_i
$$

3.2   Some useful examples

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

Table 1: A matching estimator with seven observations, $m = 1$

i	W_i	X_i	Y_i	$\mathcal{J}_1(i)$	$\widehat{Y}_i(0)$	$\widehat{Y}_i(1)$	$K_M(i)$
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

Since this dataset is small enough, we can calculate by hand that the SATE for this data is .14285714. Now let's compute the estimates using the `nnmatch` command. While the complete syntax for `nnmatch` is given in section 5.1, the basic syntax of `nnmatch` is

`nnmatch` *depvar* *treatvar* *varlist* [*weight*] [`if` *exp*] [`in` *range*]
   [`,` `tc(`{ `att` | `atc` }`)` `m(`#`)`]

Options

`tc(`{ `att` | `atc` }`)` specifies the estimand. By default, `nnmatch` estimates the average treatment effect, ATE. Specifying `tc(att)` causes `nnmatch` to estimate the average treatment effect for the treated, ATT. Specifying `tc(atc)` causes `nnmatch` to estimate the average effect for the controls, ATC.

`m(`#`)` specifies the number of matches.

See section 5.1 for a more detailed discussion of the syntax and options available.

Examples

In the output below, we estimate the SATE for the artificial dataset.

```
. use artificial
. nnmatch y w x
Matching estimator:  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
```

Note that `nnmatch` confirms our hand calculation.

Now let's consider an example with real data. We use a subset of the experimental dataset used by Lalonde (1986). The particular subset is the one constructed by Dehejia and Wahba (1999) and described there in more detail.

In this example, we are interested in the possible effect of participation in a job-training program on individuals' earnings in 1978. In the data, 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 1978 dollars). The observable covariates that we use to match similar individuals are given in table 2.

Table 2

Variable description	Variable name
age	age
years of education	educ
indicator for African-American	black
indicator for Hispanic	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 SATE using this data:

```
. use ldw_exper
. nnmatch re78 t age educ black hisp married re74 re75 u74 u75, m(4)
Matching estimator:  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
```

When more than one covariate is specified, the matching depends on the weighting matrix used to define a vector norm. The output reflects our choice of the inverse variance weighting matrix, which accounts for differences in the scale of the covariates.

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

As noted above, the population and sample average treatment effects are useful for answering different questions. For instance, the SATE is useful for judging whether this particular job-training program was successful. In contrast, if we were considering launching another job-training program in which we would obtain a second 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 is in large samples less than or equal to the standard error of the PATE, the PATE might not be significantly different from zero at either the 5% or the 1% level. In the output below, we estimate the PATE:

```
. nnmatch re78 t age educ black hisp married re74 re75 u74 u75, m(4) pop
Matching estimator: Population 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]	
PATE	1.903326	.7132952	2.67	0.008	.5052932	3.301359

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

As expected, the point estimate is exactly the same as for the SATE. We also see that the standard error of the PATE is slightly smaller than that of the SATE, so we can still reject the null hypothesis of no effect; however, our conclusion is different. Instead of concluding that the job-training program had a significant effect on this group of participants, we now conclude that the job-training program is likely to have an effect on another group of individuals drawn from the same population.

We also need to point out that the effect size, as well as the statistical significance, is important in interpreting the results in most treatment-effect studies. For instance, if our earnings data were 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.

As discussed in Imbens (2004) and Heckman, Ichimura, and Todd (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 a narrowly targeted labor-market program. For instance, when evaluating the importance of a program aimed at increasing the postgraduation 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 therefore use **nnmatch** to estimate the SATT using our extract from the Lalonde data:

```
. nnmatch re78 t age educ black hisp married re74 re75 u74 u75, tc(att) m(4)
Matching estimator:  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 the effect of the job-training program on the participants in this sample is statistically different from zero but, perhaps surprisingly, is not much larger than the SATE.

In most of the examples given in this paper, we use four matches because it offers 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 SATT statistically insignificant:

```
. nnmatch re78 t age educ black hisp married re74 re75 u74 u75, tc(att)
Matching estimator:  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
```

Note, however, that the lack of statistical significance is due largely to the change in the point estimate, rather than to the change in the standard error. It is not entirely clear which estimate is more reliable in this case.

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, we may therefore attempt to remove some of this bias term that remains after matching.

The bias-corrected matching estimator (nnmatch using the biasadj() option) 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, $\mu_w(x) = E\{Y(w)|X = x\}$ for $w = 0$ or 1. 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 that we are estimating the average treatment effect. In this case, we estimate the regression functions using only the data in the matched sample

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

for $w = 0, 1$, where

$$(\widehat{\beta}_{w0}, \widehat{\beta}_{w1}) = \mathrm{argmin}_{\{\beta_{w0}, \beta_{w1}\}} \sum_{i:W_i=w} K_M(i) \left(Y_i - \beta_{w0} - \beta'_{w1}X_i\right)^2 \qquad (3)$$

If we are interested in estimating the SATT or PATT, we need only estimate the regression function for the controls, $\mu_0(x)$. Similarly, for the SATC and PATC we need only estimate $\mu_1(x)$.

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 this same reason, we use the matched sample only in this step; using the full sample would include observations sufficiently dissimilar from our sample of interest. 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 leads to results that can be very sensitive to the exact specification applied.

Given the estimated regression functions, for the bias-corrected matching estimator we predict the missing potential outcomes as

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

and

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

with the corresponding estimator for the ATE

$$\widehat{\tau}_M^{\mathrm{bcm}} = \frac{1}{N} \sum_{i=1}^{N} \left\{ \tilde{Y}_i(1) - \tilde{Y}_i(0) \right\}$$

The bias-adjusted matching estimators for the ATT and ATC are

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

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

```
. nnmatch re78 t age educ black hisp married re74 re75 u74 u75, tc(att)
> m(4) bias(bias)
Matching estimator:  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 using bias adjustment decreases the size of the estimated SATE slightly, thus reducing the significance level. We can conclude that the treatment had an effect on its participants that is significant at the 5% but not at the 1% level.

4   Variance estimation

In this section, we describe the variance estimators implemented in **nnmatch**. Note that the bias adjustment does not affect the form of the estimator for the variance, although it may affect the numerical value. In contrast, the form of the estimator of the variance differs between the sample and population average treatment effects (or the average effect for the treated or controls). In addition, **nnmatch** offers an option for a robust estimator that allows for heteroskedasticity. Bootstrapping methods for estimating the variance of matching estimators do not necessarily give correct results. For details on the theoretical justification for the various variance estimators, see Abadie and Imbens (2002).

The estimator for the variance of the sample average treatment effect (SATE) is

$$\widehat{V}^{\text{sample}} = \frac{1}{N^2} \sum_{i=1}^{N} \left\{ 1 + K_M(i) \right\}^2 \widehat{\sigma}_{W_i}^2(X_i) \tag{4}$$

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

$$\widehat{V}^{\text{sample},t} = \frac{1}{N_1^2} \sum_{i=1}^{N} \{W_i - (1 - W_i)K_M(i)\}^2 \, \widehat{\sigma}_{W_i}^2(X_i) \tag{5}$$

and for the SATC

$$\widehat{V}^{\text{sample},c} = \frac{1}{N_0^2} \sum_{i=1}^{N} \{W_i K_M(i) - (1 - W_i)\}^2 \, \widehat{\sigma}_{W_i}^2(X_i) \tag{6}$$

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

$$\widehat{V}^{\text{pop}} = \frac{1}{N^2} \sum_{i=1}^{N} \left[\left\{ \widehat{Y}_i(1) - \widehat{Y}_i(0) - \widehat{\tau} \right\}^2 \right.$$
$$\left. + \left\{ K_M^2(i) + 2K_M(i) - K_M'(i) \right\} \widehat{\sigma}_{W_i}^2(X_i) \right] \tag{7}$$

In large samples, this will be at least as large as the estimator for the variance of the matching estimator for the SATE. Asymptotically, $\widehat{V}^{\text{pop}}$ exceeds $\widehat{V}^{\text{sample}}$ by $\sum_i \{\mu_1(X_i) - \mu_0(X_i)\}^2/N^2$. However, in small samples, $\widehat{V}^{\text{pop}}$ need not be larger than $\widehat{V}^{\text{sample}}$. In practice, we may therefore want to take the maximum of $\widehat{V}^{\text{pop}}$ and $\widehat{V}^{\text{sample}}$ as the estimator for the variance of the estimator for the PATE.

As an estimator for the variance of the matching estimator for the PATT, we use

$$\widehat{V}^{\text{pop},t} = \frac{1}{N_1^2} \sum_{i=1}^{N} \left[W_i \left\{ Y_i(1) - \widehat{Y}_i(0) - \widehat{\tau}^t \right\}^2 \right.$$
$$\left. + (1 - W_i) \left\{ K_M^2(i) - K_M'(i) \right\} \widehat{\sigma}_{W_i}^2(X_i) \right] \tag{8}$$

(Continued on next page)

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

$$
\widehat{V}^{\mathrm{pop},c} = \frac{1}{N_0^2} \sum_{i=1}^{N} \left[(1 - W_i) \left\{ \widehat{Y}_i(1) - Y_i(0) - \widehat{\tau}^t \right\}^2 \right.
$$
$$
\left. + W_i \left\{ K_M^2(i) - K_M'(i) \right\} \widehat{\sigma}_{W_i}^2(X_i) \right]
\tag{9}
$$

Estimating these variances requires estimating the conditional outcome variance $\sigma_w^2(x)$. The matching program allows to assume that this variance, $\sigma_w^2(x)$, is constant for both treatment groups and all values of the covariates or that it is not.

4.1 Assuming a constant treatment effect and homoskedasticity

Here we discuss estimating the variance under two assumptions: first, that the treatment effect, $Y_i(1) - Y_i(0)$, is constant; and second, that the conditional variance of $Y_i(w)$ given X_i does not vary with either the covariates x or the treatment w (homoskedasticity). Under these two assumptions, for the constant average treatment effect, we estimate the outcome variance σ^2 as

$$
\widehat{\sigma}^2 = \frac{1}{2N} \sum_{i=1}^{N} \left[\frac{1}{\mathcal{J}_M(i)} \sum_{l \in \mathcal{J}_M(i)} \left\{ W_i(Y_i - Y_l - \widehat{\tau}) + (1 - W_i)(Y_l - Y_i - \widehat{\tau}) \right\}^2 \right]
\tag{10}
$$

We then substitute $\widehat{\sigma}^2$ for $\widehat{\sigma}_{W_i}^2(X_i)$ into the relevant variance formula, (4) or (7). For the ATT, we estimate σ^2 as

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

and substitute this for $\widehat{\sigma}_{W_i}^2(X_i)$ into (5) or (8). Finally, for the ATC, we estimate σ^2 as

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

and substitute this for $\widehat{\sigma}_{W_i}^2(X_i)$ into (6) or (9).

4.2 Variance estimation allowing for heteroskedasticity

If $\sigma_w^2(x)$ instead varies by w and x, we must estimate it for all sample points. `nnmatch` does this using a second matching procedure, matching treated units to treated units and control units to controls. Define $d_M'(i)$ as the distance to the Mth closest unit with the same treatment indicator. More formally, let

$$\sum_{l:W_l=W_i,l\neq i} 1\left\{\|X_l - X_i\|_V < d_M'(i)\right\} < M$$

$$\sum_{l:W_l=W_i,l\neq i} 1\left\{\|X_l - X_i\|_V \leq d_M'(i)\right\} \geq M$$

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

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

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

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

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 expressions (4) through (9).

Now we return to our Lalonde–Dehejia–Wahba data extract. In the output below, we re-estimate the SATT, but this time we estimate the standard error allowing for heteroskedasticity. We specify four matches in estimating the conditional variance functions (`robusth(4)`) for the same reason that we allow four matches in estimating the conditional mean functions: given our data, four matches seem to include sufficient information without matching unlike individuals.

```
. nnmatch re78 t age educ black hisp married re74 re75 u74 u75, tc(att)
> m(4) robusth(4)
Matching estimator:  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.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
SATT	1.994622	.7526339	2.65	0.008	.5194864	3.469757

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

The output indicates that our estimated SATT remains significant at the 1% level, even 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 its participants.

5   The nnmatch command

Here we discuss the formal syntax of the nnmatch command.

5.1   Syntax of nnmatch

The complete syntax of nnmatch is as follows:

nnmatch *depvar treatvar varlist* $\big[$*weight*$\big]$ $\big[$if *exp*$\big]$ $\big[$in *range*$\big]$

$\big[$, tc($\{$ate$\,|\,$att$\,|\,$atc$\}$) m($\#$) <u>metric</u>(maha$\,|\,$*matname*) <u>exact</u>(*varlist*$_{\text{ex}}$)

<u>bias</u>adj(bias$\,|\,$*varlist*$_{\text{adj}}$) <u>r</u>obusth($\#$) population <u>level</u>($\#$) <u>keep</u>(*filename*)

replace$\big]$

pweights are allowed. See [U] **14.1.6 weight** for more information about weights. See section 5.2 for information about how nnmatch handles weights.

Description

nnmatch estimates the average treatment effect for the treated, the controls, or the sample as a whole and their standard errors. *depvar* is the outcome variable. *treatvar* is a binary variable treatment indicator. *varlist* specifies the covariates to be used in the matching.

Options

tc({ ate | att | atc }) specifies the estimand. By default, nnmatch estimates the average treatment effect, ATE. Specifying tc(att) causes nnmatch to estimate the average treatment effect for the treated, ATT. Specifying tc(atc) causes nnmatch to estimate the average effect for the controls, ATC.

m(#) specifies the number of matches to be made per observation. If two observations are equally close to that being matched, both will be used. Thus, the number of matches per observation will be greater than or equal to the number specified in m().

The default number of matches is 1, a single match. If we are estimating the average treatment effect, any integer less than or equal to the minimum of the number of treated and controls in the sample can be chosen, $M \leq \min(N_0, N_1)$. If tc(att) is specified, the limit is the number of controls in the sample, $M \leq N_0$. If tc(atc) is specified, the limit is the number of treated in the sample, $M \leq N_1$.

Typically, we should choose a small number. In simulations in Abadie and Imbens (2002), using four matches performed well in terms of mean-squared error.

metric(maha | *matname*) 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 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 of the inverses of the variances of each element of X_i (the elements of *varlist*). Second, specifying metric(maha) causes nnmatch to use to use the Mahalanobis metric, in which $V = S^{-1}$, where S is the sample covariance matrix of the covariates X_i. Third, specifying metric(*matname*) causes nnmatch to use the user-defined matrix *matname*. This third option allows you to choose any positive definite matrix.

exact(*varlist*$_{\text{ex}}$) allows you to specify exact matching (or as exact as possible) on one or more variables. As constructed, the exact-matching variables need not overlap with the elements of *varlist*. In practice, however, the exact() option adds these variables to the original $k \times k$ *varlist* matrix and multiplies the corresponding elements in the weight matrix by 1,000 relative to the weights placed on the elements of *varlist*. (Regardless of the metric() option chosen for the *varlist* variables, the exact-match variables are normalized via the default option—the inverse sample variance of the exact-matching variables.) Because for each matched observation there may not exist a member of the opposite treatment group with equal value, matching may not be exact across the full dataset. The output therefore lists the percentage of matches (across the paired sample of observations, greater than or equal to $N \times M$ in number) that match exactly.

biasadj(bias | *varlist*$_{\text{adj}}$) specifies that the bias-corrected matching estimator be used. By default, nnmatch uses the simple matching estimator. The first alternative, biasadj(bias), uses the bias-corrected matching estimator, using the same set of matching covariates, *varlist*, entered linearly in the regression function. However, you might wish to use only a subset of the covariates for this bias adjustment, or even an alternate set, so the second alternative, biasadj(*varlist*$_{\text{adj}}$), is to use the bias-corrected matching estimator with a set of covariates distinct from the set used in matching.

robusth(#) specifies that nnmatch estimate heteroskedasticity-consistent standard errors using # matches in the second matching stage (across observations of the same treatment level). The number of matches used in estimating the standard error, #, need not be the same as the number of matches used in estimating the treatment effect itself. By default, nnmatch uses the homoskedastic/constant variance estimator.

population specifies whether the estimand is a sample or population average treatment effect. This affects only the choice of estimator for the variance. By default, nnmatch estimates a sample average treatment effect—SATE, SATT, or SATC—using one of the variance estimators from (4) through (6). Specifying population causes nnmatch to estimate the variance for the population average treatment effect—PATE, PATT, or PATC—using one of the estimators from (7) through (9), as appropriate.

level(#) specifies the confidence level, as a percentage, for confidence intervals. The default is level(95) or as set by set level.

keep(*filename*) saves the temporary matching dataset in the file *filename*.dta. In the estimation process, nnmatch creates a temporary dataset holding, for each observation i being matched, a new observation containing the values of i's outcome variable, the matching variable(s), and the outcome and matching variable(s) for its M closest matches. The new dataset will hold at least $N \times M$ observations. If biasadj(*varlist*$_{\text{adj}}$) or exact(*varlist*$_{\text{ex}}$) are selected, the temporary dataset will also hold these values for each observation i and its match(es). The variables in this dataset are discussed in greater detail in section (6).

replace replaces the dataset specified in keep(*filename*) if it already exists.

5.2 How nnmatch handles weights

nnmatch allows probability weights. Under this assumption, 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 weighted nnmatch estimates are identical to those that would be obtained by using the unweighted estimator on a sample in which observation i is duplicated once. The standard errors are updated to account for the weighting. The weights are allowed to be noninteger but must be nonnegative.

Formally, with the weight for individual i equal to ω_i, the estimator is calculated as follows. The definition of the distance $d_M(i)$ to the Mth closest match is modified to ensure that the sum of the weights of the matches adds up to M. Thus, $d_M^\omega(i)$ satisfies

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

and

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

The definition of the set $\mathcal{J}_M(i)$ is unchanged. The $K_M(i)$ and $K_M'(i)$ are changed from (1) and (2) to

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

and

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

The simple matching estimator is now defined as

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

with a similar modification for the ATT and ATC. The bias correction is as described above, except that in (3) the weighted version $K_M^\omega(i)$ is used.

The variance formulas change from (4) through (9) to

$$\widehat{V}^{\text{sample}} = \frac{1}{(\sum_i \omega_i)^2} \sum_{i=1}^N \left\{ \omega_i + K_M^\omega(i) \right\}^2 \widehat{\sigma}_{W_i}^2(X_i) \tag{4'}$$

$$\widehat{V}^{\text{sample},t} = \frac{1}{(\sum_i W_i \omega_i)^2} \sum_{i=1}^N \left\{ W_i \omega_i - (1 - W_i) K_M^\omega(i) \right\}^2 \widehat{\sigma}_{W_i}^2(X_i) \tag{5'}$$

$$V^{\text{sample},c} = \frac{1}{(\sum_i (1 - W_i) \omega_i)^2} \sum_{i=1}^N \left\{ W_i K_M^\omega(i) - (1 - W_i) \omega_i \right\}^2 \widehat{\sigma}_{W_i}^2(X_i) \tag{6'}$$

$$\widehat{V}^{\text{pop}} = \frac{1}{(\sum_i \omega_i)^2} \sum_{i=1}^N \Bigg(\omega_i \left\{ \widehat{Y}_i(1) - \widehat{Y}_i(0) - \widehat{\tau} \right\}^2$$
$$+ \left[\{K_M^\omega(i)\}^2 + 2K_M^\omega(i) - K_M^{\omega\prime}(i) \right] \widehat{\sigma}_{W_i}^2(X_i) \Bigg) \tag{7'}$$

$$\widehat{V}^{\text{pop},t} = \frac{1}{(\sum_i W_i \omega_i)^2} \sum_{i=1}^N \Bigg(W_i \omega_i \left\{ Y_i(1) - \widehat{Y}_i(0) - \widehat{\tau}^c \right\}^2$$
$$+ (1 - W_i) \left[\{K_M^\omega(i)\}^2 - K_M^{\omega\prime}(i) \right] \widehat{\sigma}_{W_i}^2(X_i) \Bigg) \tag{8'}$$

$$\widehat{V}^{\text{pop},c} = \frac{1}{(\sum_i (1 - W_i) \omega_i)^2} \sum_{i=1}^N \Bigg((1 - W_i) \omega_i \left\{ \widehat{Y}_i(1) - Y_i(0) - \widehat{\tau}^c \right\}^2$$
$$+ W_i \left[\{K_M^\omega(i)\}^2 - K_M^{\omega\prime}(i) \right] \widehat{\sigma}_{W_i}^2(X_i) \Bigg) \tag{9'}$$

As always, we must pay attention to the interpretation of these weighted estimators. In the population variance estimators, the weights correct any bias introduced by the sampling design so that inference is made with respect to the underlying population of interest. The weighted sample variances can thus be interpreted as providing inference about another sample of identical individuals drawn from the population that corresponds to the weights.

For $\widehat{\sigma}_{W_i}^2(X_i)$, we modify the earlier estimators. In the homoskedastic case, the error-variance estimator is changed from (10) to

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

with similar modifications for the ATT and ATC cases.

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

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

and

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

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

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

We estimate the conditional variance, $\sigma_{W_i}^2(X_i)$, as the sample variance in this set, augmented with the outcome for unit i itself, $\mathcal{U} = \mathcal{J}_M'(i) \cup \{i\}$, taking into account the weights

$$\tilde{\sigma}_{W_i}^2(X_i) = \frac{\sum_{j\in\mathcal{U}} \omega_j \Big\{Y_j - \bar{Y}_{\mathcal{U}}\Big\}^2}{\sum_{j\in\mathcal{U}} \omega_j}$$

where

$$\bar{Y}_{\mathcal{U}} = \frac{\sum_{j\in\mathcal{U}} \omega_j Y_j}{\sum_{j\in\mathcal{U}} \omega_j}$$

is the average outcome in this set. The overall variance is estimated by plugging this unit-level variance estimate into the relevant variance expression $(4')$ through $(9')$.

6 The matching dataset

If keep(*filename*) is selected, each observation of *filename*.dta will hold the following variables:

t	The treatment group indicator for the observation being matched.
y	The observed outcome variable for observation i.
x	The values of the covariates, *varlist*, for observation i.
id	The identification code for the observation being matched. (When the command nnmatch is given, the program creates a temporary variable, $\mathtt{id} = 1, 2, \ldots, N$, based on the original sort order.)
index	The identification code for j, the match for observation i.
dist	The estimated distance between observation i and its match j, based on the values of the *varlist* and the selected weight matrix.
k_m	The number of times observation i is used as a match for any observation l of the opposite treatment level, each time weighted by the total number of matches for the given observation l. For example, if observation i is one of three matches for observation l, it receives a value of $1/3$ for that match. k_m is the sum across all observations l of this value. Thus the sum of k_m across all observations will equal N (or N_0 or N_1, if the ATC or ATT, respectively, are estimated). Note that this value refers to i's use as a match, not to its matches j, so the value of k_m is equal across all observations in the temporary dataset that pertain to the matching of observation i.
w_id	Weight for observation i, if weights are selected.
w_index	Weight of observation j, the match for observation i, if weights are selected.
y_0	The estimated value of the dependent variable if observation i were in the control group. If observation i is a control observation, $\mathtt{y\_0} = \mathtt{y}(i)$. If i is a treated observation, $\mathtt{y\_0} = \mathtt{y}(j)$.
y_1	Estimated value of the dependent variable if i were in the treated group.
x_0m	Values of *varlist* for i's control observation. Namely, if i is a control observation, $\mathtt{x\_0m} = x_i$ for each element x of *varlist*. If i is a treated observation, $\mathtt{x\_0m} = x_j$.
x_1m	Values of *varlist* for i's treatment observation.
b_0b	Values of the bias-adjustment variables, if biasadj() is selected, for i's control observation, where b represents each element of the bias-adjustment variables.
b_1b	Bias-adjustment variables for i's treatment observation.
e_0e	Values of the exact-matching variables, if exact() is selected, for i's control observation, where e represents each element of the exact-matching variables.
e_1e	Exact-matching variables for i's treatment observation.

7   References

Abadie, A. and G. Imbens. 2002. Simple and bias-corrected matching estimators. Technical report, Department of Economics, University of California, Berkeley. *http://emlab.berkeley.edu/users/imbens/*.

Cochran, W. and D. B. Rubin. 1973. Controlling bias in observational studies. *Sankhyā* 35: 417–446.

Dehejia, R. H. and S. Wahba. 1999. Causal effects in nonexperimental studies: Re-evaluation of the evaluation of training programs. *Journal of the American Statistical Association* 94: 1053–1062.

Heckman, J., H. Ichimura, and P. Todd. 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65(2): 261–294.

Imbens, G. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86(1): 4–29.

Lalonde, R. J. 1986. Evaluating the econometric evaluations of training programs. *American Economic Review* 76(4): 604–620.

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

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

—. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* 39: 33–38.

Rubin, D. B. 1973. The use of matched sampling and regression adjustments to remove bias in observational studies. *Biometrics* 29: 185–203.

Rubin, D. B. and N. Thomas. 1992. Affinely invariant matching methods with ellipsoidal distributions. *Annals of Statistics* 20: 1079–1093.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

About the Authors

Alberto Abadie is Associate Professor of Public Policy at Harvard University.

David Drukker is Director of Econometrics at StataCorp.

Jane Leber Herr is a doctoral student in the Department of Economics University of California at Berkeley.

Guido Imbens is Professor of Economics, Department of Economics, and Professor of Agricultural Economics, Department of Agricultural and Resource Economics, University of California at Berkeley.