Implementing Matching Estimators for Average Treatment Effects in STATA

Guido W. Imbens - Harvard University

West Coast Stata Users Group meeting, Los Angeles

October 26th, 2007

General Motivation

Estimation of average effect of binary treatment, allowing for general heterogeneity.

Use matching to eliminate bias that is present in simple comparison of means by treatment status.

Economic Applications

Labor market programs:
Ashenfelter (1978), Ashenfelter and Card (1985), Lalonde (1986), Card and Sullivan (1989), Heckman and Hotz (1989), Friedlander and Robins (1995), Dehejia and Wahba (1999), Lechner (1999), Heckman, Ichimura and Todd (1998).

Effect of Military service on Earnings:
Angrist (1998)

Effect of Family Composition:
Manski, McLanahan, Powers, Sandefur (1992)

Many other applications.

Topics

1. General Set Up /Notation

2. Estimators for Ave Treatm Effect under Unconf.

3. Dealing with Lack of Overlap in Covariate Distributions

4. Implementation in STATA Using `nnmatch`

5. Illustration using Lalonde Data

1. Notation

N individuals/firms/units, indexed by i=1,. . . ,N,

$W_i \in \{0, 1\}$: Binary treatment,

$Y_i(1)$: Potential outcome for unit i with treatment,

$Y_i(0)$: Potential outcome for unit i without the treatment,

X_i: $k \times 1$ vector of covariates.

We observe $\{(X_i, W_i, Y_i)\}_{i=1}^{N}$, where

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

Fundamental problem: we never observe $Y_i(0)$ and $Y_i(1)$ for the same individual i.

Notation (ctd)

$\mu_w(x) = \mathbb{E}[Y(w)|X = x]$ (regression functions)

$\sigma^2_w(x) = \mathbb{E}[(Y(w) - \mu_w(x))^2|X = x]$ (conditional variances)

$e(x) = \mathbb{E}[W|X = x] = \Pr(W = 1|X = x)$ (propensity score, Rosenbaum and Rubin, 1983)

$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x] = \mu_1(x) - \mu_0(x)$ (conditional average treatment effect)

$\tau = \mathbb{E}[\tau(X)] = \mathbb{E}[Y(1) - Y(0)]$ Population Average Treatment Effect

Assumptions

I. Unconfoundedness

$$Y(0), Y(1) \perp W \mid X.$$

This form due to Rosenbaum and Rubin (1983). Like selection on observables, or exogeneity. Suppose

$$Y_i(0) = \alpha + \beta' X_i + \varepsilon_i, \quad Y_i(1) = Y_i(0) + \tau,$$

then

$$Y_i = \alpha + \tau \cdot W_i + \beta' X_i + \varepsilon_i,$$

and unconfoundedness $\iff \varepsilon_i \perp W_i | X_i$.

II. Overlap

$$0 < \Pr(W = 1|X) < 1.$$

For all X there are treated and control units.

Motivation for Assumptions

I. Descriptive statistics. After simple difference in mean outcomes for treated and controls, it may be useful to compare average outcomes adjusted for covariates.

II. Alternative: bounds (e.g., Manski, 1990)

III. Unconfoundedness follows from some economic models.

Suppose individuals choose treatment w to maximize expected utility, equal to outcome minus cost, $Y_i(w) - c_i \cdot w$, conditional on a set of covariates X:

$$W_i = \text{argmax}_w \ \mathbb{E}[Y_i(w)|X_i] - c_i \cdot w.$$

Suppose that costs c_i differ between individuals, indep. of potential outcomes. Then
(i) choices will vary between individuals with the same covariates, and
(ii) conditional on the covariates X the choice is independent of the potential outcomes.

Identification

$$\tau(X) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

$$= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x]$$

By unconfoundedness this is equal to

$$\mathbb{E}[Y(1)|W = 1, X = x] - \mathbb{E}[Y(0)|W = 0, X = x]$$

$$= \mathbb{E}[Y|W = 1, X = x] - \mathbb{E}[Y|W = 0, X = x].$$

By the overlap assumption we can estimate both terms on the righthand side.

Then

$$\tau = \mathbb{E}[\tau(X)].$$

Questions

How well can we estimate τ?

How do we estimate τ?

How do we do inference?

How do we assess assumptions (unconfoundedness/overlap)?

2. Estimation of Average Treatment Effect under Unconfoundedness

I. Regression estimators: estimate $\mu_w(x)$.

II. Propensity score estimators: estimate $e(x)$

III. Matching: match all units to units with similar values for covariates and opposite treatment.

IV. Combining Regression with Propensity score and Matching Methods.

Regression Estimators

Estimate $\mu_w(x)$ nonparametrically, and then

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)).$$

These estimators can reach efficiency bound.

Propensity Score Estimators

The key insight is that even with high-dimensional covariates, one can remove all bias by conditioning on a scalar function of the covariates, the propensity score. Formally, if

$$Y(0), Y(1) \perp W \mid X.$$

then

$$Y(0), Y(1) \perp W \mid e(X).$$

($e(x) = \Pr(W = 1 | X = x)$, Rosenbaum and Rubin, 1983)

Thus we can reduce the dimension of the conditioning set (if we know the propensity score) to one.

Propensity Score Estimators (ctd)

Estimate $e(x)$ nonparametrically, and then:

A. weighting (Hirano, Imbens, Ridder, 2003)

$$\widehat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{W_i \cdot Y_i}{\widehat{e}(X_i)} - \frac{(1 - W_i) \cdot Y_i}{1 - \widehat{e}(X_i)} \right).$$

This is based on the fact that

$$\mathbb{E}\left[\frac{W \cdot Y}{e(X)} \middle| X = x \right] = \mathbb{E}\left[\frac{W \cdot Y(1)}{e(X)} \middle| X = x \right]$$

$$= \mathbb{E}\left[\frac{W}{e(X)} \middle| X = x \right] \cdot \mathbb{E}\left[Y(1) \middle| X = x \right] = \mu_1(x).$$

Propensity Score Estimators (ctd)

B. Blocking (Rosenbaum and Rubin, 1983)

Divide sample in subsamples on the basis of the value of the (estimated) propensity score. Estimate average treatment effect within each block as the difference in average outcomes for treated and controls. Average within block estimates by the proportion of observations in each block.

Using five blocks reduces bias by about 90% (Cochran, 1968), under normality.

Matching

For each treated unit i, find untreated unit $\ell(i)$ with

$$\|X_{\ell(i)} - x\| = \min_{\{l:W_l=0\}} \|X_l - x\|,$$

and the same for all untreated observations. Define:

$$\widehat{Y}_i(1) = \begin{cases} Y_i & \text{if } W_i = 1, \\ Y_{\ell(i)} & \text{if } W_i = 0, \end{cases} \qquad \widehat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ Y_{\ell(i)} & \text{if } W_i = 1. \end{cases}$$

Then the simple matching estimator is:

$$\widehat{\tau}^{sm} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}_i(1) - \widehat{Y}_i(0) \right).$$

Note: since we match all units it is crucial that matching is done with replacement.

Matching (ctd)

More generally, let $\mathcal{J}_M(i) = \{\ell_1(i), ..., \ell_M(i)\}$ be the set of indices for the nearest M matches for unit i.

Define:

$$\widehat{Y}_i(1) = \begin{cases} Y_i & \text{if } W_i = 1, \\ \sum_{j \in \mathcal{J}_M(i)} Y_j / M & \text{if } W_i = 0, \end{cases}$$

$$\widehat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \sum_{j \in \mathcal{J}_M(i)} Y_j / M & \text{if } W_i = 1. \end{cases}$$

Matching is generally not efficient (unless $M \to \infty$), but efficiency loss is small (variance is less than $1 + 1/(2M)$ times the efficiency bound).

The bias is of order $O_p(N^{-1/k})$, where k is the dimension of the covariates.

Matching is consistent under weak smoothness conditions (does not require higher order derivatives).

Matching and Regression

Estimate $\mu_w(x)$, and modify matching estimator to:

$$\tilde{Y}_i(1) = \begin{cases} Y_i & \text{if } W_i = 1, \\ Y_{\ell(i)} + \widehat{\mu}_1(X_i) - \widehat{\mu}_1(X_{j(i)}) & \text{if } W_i = 0 \end{cases}$$

$$\tilde{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ Y_{\ell(i)} + \widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)}) & \text{if } W_i = 1 \end{cases}$$

Then the bias corrected matching estimator is:

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

Variance Estimation

Matching estimators have the form

$$\widehat{\tau} = \sum_{i=1}^{N} \Big(W_i \cdot \lambda_i \cdot Y_i - (1 - W_i) \cdot \lambda_i \cdot Y_i \Big),$$

(linear in outcomes) with weights

$$\lambda_i = \lambda(\mathbf{W}, \mathbf{X}).$$

$\lambda(\mathbf{W}, \mathbf{X})$ (known) is very non-smooth for matching estimators and bootstrap is **not** valid as a result. (not just no second order justification, but not valid asymptotically)

Variance conditional on $\mathbf{W}$ and $\mathbf{X}$ is

$$V\left(\widehat{\tau}|\mathbf{W},\mathbf{X}\right)=\sum_{i=1}^{N}\left(W_i\cdot\lambda_i^2\cdot\sigma_1^2(X_i)+(1-W_i)\cdot\lambda_i^2\cdot\sigma_0^2(X_i)\right).$$

All parts known other than $\sigma_w^2(x)$.

For each treated (control) find the closest treated (control) unit: $h(i)=\min_{j\neq i,W_j=W_i}\|X_i-X_j\|$. Then use the difference between their outcomes to estimate $\sigma^2(X_i)$ for this unit:

$$\widehat{\sigma}_{W_i}^2(X_i)=\frac{1}{2}(Y_i-Y_{h(i)})^2.$$

Substitute into variance formula.

Even though $\widehat{\sigma}_w^2(x)$ is not consistent, the estimator for $V(\widehat{\tau}|\mathbf{W},\mathbf{X})$ is because it averages over all $\widehat{\sigma}_w^2(X_i)$.

3. Assessing Overlap

The first method to detect lack of overlap is to plot distributions of covariates by treatment groups. In the case with one or two covariates one can do this directly. In high dimensional cases, however, this becomes more difficult.

One can inspect pairs of marginal distributions by treatment status, but these are not necessarily informative about lack of overlap. It is possible that for each covariate the distribution for the treatment and control groups are identical, even though there are areas where the propensity score is zero or one.

A more direct method is to inspect the distribution of the propensity score in both treatment groups, which can reveal lack of overlap in the multivariate covariate distributions.

Selecting a Subsample with Overlap

Define average effects for subsamples $\mathbb{A}$:

$$\tau(\mathbb{A}) = \sum_{i=1}^{N} 1\{X_i \in \mathbb{A}\} \cdot \tau(X_i) / \sum_{i=1}^{N} 1\{X_i \in \mathbb{A}\}.$$

The efficiency bound for $\tau(\mathbb{A})$, assuming homoskedasticity, as

$$\frac{\sigma^2}{q(\mathbb{A})} \cdot \mathbb{E}\left[\frac{1}{e(X)} + \frac{1}{1-e(X)} \middle| X \in \mathbb{A}\right],$$

where $q(\mathbb{A}) = \mathsf{Pr}(X \in \mathbb{A})$.

They derive the characterization for the set $\mathbb{A}$ that minimizes the asymptotic variance .

The optimal set has the form

$$\mathbb{A}^* = \{x \in \mathbb{X} | \alpha \leq e(X) \leq 1 - \alpha\},$$

dropping observations with extreme values for the propensity score, with the cutoff value α determined by the equation

$$\frac{1}{\alpha \cdot (1-\alpha)} =$$

$$2 \cdot \mathbb{E}\left[\frac{1}{e(X) \cdot (1-e(X))} \middle| \frac{1}{e(X) \cdot (1-e(X))} \leq \frac{1}{\alpha \cdot (1-\alpha)}\right].$$

Note that this subsample is selected solely on the basis of the joint distribution of the treatment indicators and the covariates, and therefore does not introduce biases associated with selection based on the outcomes.

Calculations for Beta distributions for the propensity score suggest that $\alpha = 0.1$ approximates the optimal set well in practice.

4. Implementation in STATA Using `nnmatch`

Syntax:

`nnmatch` *depvar treatvar varlist* [*weight*] [`if` *exp*] [`in` *range*] [, `tc`({`ate`|`att`|`atc`}) `m`(#) `metric`(`maha`|*matname*) `exact`(*varlist*$_{\text{ex}}$) `biasadj`(`bias`|*varlist*$_{\text{adj}}$) `robusth`(#) `population` `level`(#) `keep`(*filename*) `replace`]

`nnmatch` *depvar treatvar varlist* [*weight*] [`if` *exp*] [`in` *range*]

Basic command: `treatvar` must be binary variable

option 1

`tc({ate|att|atc})`

One can est. the overall average effect of the treatment (`ate`), or the average treatment effect for the treated units (`att`), or the average effect for those who were not treated (`atc`)

option 2

`m(#)`

The number of matches. In kernel matching estimators essentially the key difference is that the number of matches increases with the sample size. In practice there is little gain from using more than 3-4 matches. Under homoskedasticity the variance goes down proportional to $1+1/(2M)$, where M is the number of matches.

option 3

`metric(maha|`*matname*)

The distance metric. Two main options, the inverse of the variances or mahalanobis distance. It can also be prespecified.

option 4

ex`act`(*varlist*$_{\text{ex}}$)

A special list of covariates receives extra weight (1000 times the weight specified in `met`. Useful for binary covariates where one wishes to match exactly.

option 5

`biasadj(bias|`*varlist*$_{\text{adj}}$ `)`

In treatment and control group regression adjustment is used based on the variables in this option. If `bias(bias)` is used, all the variables used in the matching are used here again.

option 6 `robusth(#)`

heteroskedasticity consistent variance estimation.

option *7*

`population`

sample average treatment effect

$$\frac{1}{N}\sum_{i=1}^{N}\left(\mu_1(X_i)-\mu_0(X_i)\right),$$

versus population average treatment effect:

$$\mathbb{E}[\mu_1(X)-\mu_0(X)].$$

The former can be estimated more precisely if there is heterogeneity in $\mu_1(x)-\mu_0(x)$.

option 8

`level(#)`

standard STATA option that specifies the confidence level for confidence sets

option 9

`keep`(*filename*) `replace`

Allows the user to recover output beyond the estimate and its standard error.

A new data set is created with one observation per match, and covariate information is kept for control and treated unit in each match.

Examples

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

```
nnmatch re78 t age educ black hisp married re74 re75 u74 u75,
tc(att) m(4) exact(reo75) bias(bias) rob(4) keep(lalonde_temp1)
replace
```

```
nnmatch re78 t age educ black hisp married re74 re75 u74 u75,
tc(att) m(4) exact(pscore) bias(bias) rob(4) keep(lalonde_temp2)
replace
```

5. Illustration with Lalonde Data, CPS Control Group: Summary Statistics

```
bysort t:  summ re75 married
```

```
-------------------------------------------------------------------------------
--
      log:  c:\guido\match\program_06july\lalonde_nonexper_07oct24.log
   opened on:  24 Oct 2007, 18:04:32
.
. infile t age educ black hisp married nodegree re74 re75 re78  using
c:\data\l
> alonde\nonexper\nswcps.dat
(16177 observations read)

. gen u75=(re75==0)
.
. gen u74=(re74==0)
.
. logit t age educ black hisp married nodegree re74 re75 u74 u75
.
. bysort t: summ
```

-> t = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
t	15992	0	0	0	0
age	15992	33.22524	11.04522	16	55
educ	15992	12.02751	2.870846	0	18
black	15992	.0735368	.2610237	0	1
hisp	15992	.072036	.2585556	0	1
married	15992	.7117309	.4529712	0	1
nodegree	15992	.2958354	.4564316	0	1
re74	15992	14016.8	9569.796	0	25862.32
re75	15992	13650.8	9270.403	0	25243.55
re78	15992	14846.66	9647.392	0	25564.67
u75	15992	.1093047	.3120308	0	1
u74	15992	.1196223	.3245295	0	1
pscore	15992	.0078467	.0434384	4.70e-06	.637511

-> t = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
t	185	1	0	1	1
age	185	25.81622	7.155019	17	48
educ	185	10.34595	2.01065	4	16
black	185	.8432432	.3645579	0	1
hisp	185	.0594595	.2371244	0	1
married	185	.1891892	.3927217	0	1
nodegree	185	.7081081	.4558666	0	1
re74	185	2095.574	4886.62	0	35040.07
re75	185	1532.055	3219.251	0	25142.24
re78	185	6349.144	7867.402	0	60307.93
u75	185	.6	.4912274	0	1
u74	185	.7081081	.4558666	0	1
pscore	185	.3217095	.2258198	.0005119	.6333636

Estimation of Propensity Score

```
logit t age educ black hisp married re74 re75 reo74 reo75

predict pscore

bysort t:  summ pscore
```

Checking Plausability of Unconfoundedness

```
nnmatch re75 t age educ black hisp married re74 u74, tc(att)
m(4) bias(bias) rob(4)
```

re75_all.txt 10/25/2007

```
. nnmatch re75 t age educ black hisp married nodegree re74 u74, tc(att) m(1)
b
> ias(bias) rob(2) replace

Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: inverse variance    Number of obs           =    16177
                                      Number of matches  (m) =        1
                                      Number of matches,
                                      robust std. err. (h) =         2

------------------------------------------------------------------------------
  re75 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
  SATT |  -1344.544   314.9082    -4.27   0.000    -1961.753    -727.335
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married nodegree re74 u74
Bias-adj variables:  age educ black hisp married nodegree re74 u74

.
```

Selecting Sample

```
keep if (pscore>0.1)&(pscore<0.9)
```

```
.
. keep if (pscore>0.10)&(pscore<0.90)
(15725 observations deleted)

.
. bysort t: summ
```

-> t = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
t	308	0	0	0	0
age	308	26.46429	11.05429	16	55
educ	308	10.70779	2.623932	0	17
black	308	.9318182	.2524678	0	1
hisp	308	.0681818	.2524678	0	1
married	308	.2305195	.4218507	0	1
nodegree	308	.5974026	.4912191	0	1
re74	308	1917.193	4152.996	0	22321.93
re75	308	937.2397	1688.34	0	9554.952
re78	308	4259.504	5844.25	0	25564.67
u75	308	.5584416	.4973809	0	1
u74	308	.5844156	.4936245	0	1
pscore	308	.2624091	.1668767	.100889	.637511

-> t = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
t	144	1	0	1	1
age	144	25.88194	7.429385	17	48
educ	144	10.1875	2.085477	4	16
black	144	.9722222	.1649091	0	1
hisp	144	.0277778	.1649091	0	1
married	144	.1319444	.3396116	0	1
nodegree	144	.7569444	.4304255	0	1
re74	144	1284.277	3646.846	0	25929.68
re75	144	773.5342	1533.236	0	7867.916
re78	144	5931.443	8007.724	0	60307.93
u75	144	.6944444	.4622502	0	1
u74	144	.7986111	.4024377	0	1
pscore	144	.4053606	.1832125	.1019888	.6333636

.

Checking Plausability of Unconfoundedness in Selected Sample

```
nnmatch re75 t age educ black hisp married re74 u74, tc(att)
m(4) bias(bias) rob(4)
```

```
.
. nnmatch re75 t age educ black hisp married nodegree re74 u74, tc(att) m(1)
b
> ias(bias) rob(2) replace
note: hisp dropped due to collinearity

Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: inverse variance    Number of obs           =       452
                                      Number of matches  (m) =         1
                                      Number of matches,
                                      robust std. err. (h) =         2

------------------------------------------------------------------------------
  re75 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
  SATT |  -169.3418   248.5711    -0.68   0.496    -656.5322    317.8486
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married nodegree re74 u74
Bias-adj variables:  age educ black hisp married nodegree re74 u74

.
```

Final Estimates

```
nnmatch re78 t age educ black hisp married re74 re75 u74 u75,
tc(att) m(4) bias(bias) rob(4)
```

```
.
. nnmatch re78 t age educ black hisp married nodegree re74 re75 u74 u75,
tc(att
> ) m(1) bias(bias) rob(2) replace
note: hisp dropped due to collinearity

Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: inverse variance    Number of obs           =       452
                                      Number of matches  (m) =         1
                                      Number of matches,
                                      robust std. err. (h) =         2

------------------------------------------------------------------------------
  re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
  SATT |   2023.413   1045.263     1.94   0.053    -25.26369     4072.09
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married nodegree re74 re75 u74 u75
Bias-adj variables:  age educ black hisp married nodegree re74 re75 u74 u75

.
```

Cautionary note:

If there are few matching variables (all discrete), and many ties, `nnmatch` can be very slow, and memory intensive.

One solution is to add a continuous matching variable, even if it is unrelated to other things, to break the ties.