Implementing Matching Estimators for Average Treatment Effects in STATA

Guido W. Imbens - Harvard University

Stata User Group Meeting, Boston

July 26th, 2006

General Motivation

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

Use matching to eliminate bias thatis 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. Implementation in STATA Using `nnmatch`

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)

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

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

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 $\Longleftrightarrow \varepsilon_i \perp W_i | X_i$.

II. Overlap

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

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

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

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)

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

This is based on the fact that

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

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

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:

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

Then the simple matching estimator is:

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

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:

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

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

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:

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

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

Then the bias corrected matching estimator is:

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

Variance Estimation

Matching estimators have the form

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

(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

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

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:

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

Substitute into variance formula.

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

3. 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

`exact`(*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.

3 Examples

```
nnmatch re78 t age educ black hisp married re74 re75 reo74 reo75,
tc(att)

nnmatch re78 t age educ black hisp married re74 re75 reo74 reo75,
tc(att) m(4) exact(reo75) bias(bias) rob(4) keep(lalonde_temp1)
replace

nnmatch re78 t age educ black hisp married re74 re75 reo74 reo75,
tc(att) m(4) exact(pscore) bias(bias) rob(4) keep(lalonde_temp2)
replace
```

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.