Lecture X:
Matching

Stanislao Maldonado

Universidad del Rosario

Impact Evaluation
April 25th, 2017

1. Introduction

- Matching offers a way to estimate ATE when:
  - Controlled randomization is impossible
  - There are not convincing natural experiments
- Key idea: to compute treatment effects using carefully selected "matches" between treatment and control units
- Problem: selection is based on observables, so depends on a strong assumption (some form of exogeneity):
  - Selection into treatment is completely determined by variables that can be observed by the researcher

- “conditioning” on these observable variables the assignment to treatment is random

- Some “names” for this assumption:
  - Unconfoundedness (Rosembaum and Rubin 1983, Imbens 2004)
  - Selection on observables
  - Conditional independence

2. Identification

- **Assumption A.1**: **Unconfoundedness** (Imbens 2004)

  Assignment to treatment is unconfounded given pre-treatment variables if:

$$(1)\ \{Y_i(1), Y_i(0)\} \perp D_i / X$$

- Equivalent to say:
  - Within each cell defined by X treatment is random
  - The selection onto treatment depends only on the observables X

- **Assumption A.2**: **Overlap** (Imbens 2004)

$$(2)\ 0<\Pr\{D_i = 1/X\}<1$$

- Are these assumptions plausible? (what do you think?)
- Given A.1 and A.2:

$$(3)\ E\left[Y_i(0)/D_i=0,X\right]=E\left[Y_i(0)/D_i=1,X\right]=E\left[Y_i(0)/X\right]$$

$$(4)\ E\left[Y_i(1)/D_i=1,X\right]=E\left[Y_i(1)/D_i=0,X\right]=E\left[Y_i(1)/X\right]$$

- Then:

$$\begin{aligned}(5)\ ATE_X &= E\left[\Delta_i/X\right]=E\left[Y_i(1)-Y_i(0)/X\right]\\ &= E\left[Y_i(1)/X\right]-E\left[Y_i(0)/X\right]\\ &= E\left[Y_i(1)/D_i=1,X\right]-E\left[Y_i(0)/D_i=0,X\right]\\ &= E\left[Y_i/X,D_i=1\right]-E\left[Y_i/X,D_i=0\right]\end{aligned}$$

- Why A.2 is needed?
  - We need to be able to estimate: $E\left[Y_i / X = x, D_i = d\right]$ for all values of x and d in the support of these variables
  - This ensures that there will be treated (control) units to be compared to control (treated) units in each cell
- These assumptions suggest the following estimation strategy for ATE:
  - Stratify the data into cells defined by particular values of X
  - Within each cell, compute ATE comparing treated and controls
  - Average these differences with respect to the distribution of X in the population of treated units

3. The “curse of dimensionality” and the propensity score

- When sample size is small, the set of covariates is large and some of them are multi-valued or (even worse) continue, you are in trouble (“the curse of dimensionality”)
  - With K binary variables the number of cells is 2^k
  - If some X are multivalued, then number of cells increases further
  - It could be the case that cells will contain only treated (control) units. Treatment effects can not be computed
- Solution: use the propensity score for solving the dimensionality problem (Rosenbaum and Rubin 1983)

- **Definition 1**: **Propensity score** (Rosenbaum and Rubin 1983)

  The propensity score is the conditional probability of receiving the treatment given the pre-treatment variables:

  $$(6)\ p(\mathrm{X})=\Pr(D_i=1/X)=E(D_i/X)$$

- Two important properties:
  - Lemma 1: *Balancing of pre-treatment characteristics given the propensity score* (See proof in Rosenbaum and Rubin 1983)

    $$(6')\ D_i \perp X \mid p(X)$$

- Lemma 2: *Unconfoundedness given the propensity score* (See proof in Rosenbaum and Rubin 1983)

  Suppose that assignment to treatment is unconfounded, i.e

  $$(6'') \quad \{Y_i(1), Y_i(0)\} \perp D_i / X$$

  Then, assignment to treatment is unconfounded given the propensity score, i.e

  $$(6''') \quad \{Y_i(1), Y_i(0)\} \perp D_i / p(X)$$

- Using p(X), we have:

$$(7) \; E\left[Y_i(0)/D_i = 0, p(X)\right] = E\left[Y_i(0)/D_i = 1, p(X)\right] = E\left[Y_i(0)/p(X)\right]$$

$$(8) \; E\left[Y_i(1)/D_i = 1, p(X)\right] = E\left[Y_i(1)/D_i = 0, p(X)\right] = E\left[Y_i(1)/p(X)\right]$$

- Therefore, we compute ATE as follows:

$$
\begin{aligned}
(9)\ ATE_{p(X)} &= E\left[\Delta_i / p(X)\right] = E\left[Y_i(1) - Y_i(0) / p(X)\right] \\
&= E\left[Y_i(1) / p(X)\right] - E\left[Y_i(0) / p(X)\right] \\
&= E\left[Y_i(1) / D_i = 1,\ p(X)\right] - E\left[Y_i(0) / D_i = 0,\ p(X)\right] \\
&= E\left[Y_i / p(X), D_i = 1\right] - E\left[Y_i / p(X), D_i = 0\right]
\end{aligned}
$$

Estimation strategy with the p(X)

1. Estimation of the propensity score

   True p(X) is unknown!

   - Thanks to Lemma 1, we have:
     - ✓ Observation with the same p(X) have the same distribution of observable covariates independently of treatment status
     - ✓ For a given propensity assignment to treatment is random and therefore treated and controls units are on average observationally identical
   - Any standard probability model can be used to estimate the p(X), e.g. logit model:

$$(10)\ p(\mathrm{X})=\Pr(D_i=1/X)=\frac{e^{\lambda h(X_i)}}{1+e^{\lambda h(X_i)}}$$

- Where h(X) is a function of covariates with linear and high order terms

2. Estimation of the treatment effect (ATE, ATT or ATU) given the propensity score
   a. Match treated and controls units with the same (computed) p(X)
   b. Compute the treatment effect for each value of the (computed) p(X)
   c. Obtain the average of these conditional effects

- Problems: it is unfeasible to compute ATE for each value since it is hard to find two units with the same p(X)

- Recall: the goal is balance of covariates rather than p(X) model validity!
- Are propensity scores the way to go?
  - If model is misspecified, it can increase bias even if CIA holds (Drake 1993)
  - Covariates imbalance may be increased, specially if these covariates don't have ellipsoidal distributions (Rubin and Thomas 1992)
  - Solution: revise specification of propensity score until balance is achieved (Rosenbaum and Rubin 1984). Alternative: genetic matching (Diamond and Sekhon 2010)

4. Implementing matching estimators

- 4 key steps (Stuart 2010):

1. Defining “closeness”: the distance measure to define whether an individual is a good match for another
2. Implementing a matching method: based on step 1.
3. Assessing the quality of the resulted matched sample: if balance is not good, re-do steps 1 and 2 until balance is achieved.
4. Analysis of the outcome and estimation of the treatment effect: based on step 3.

4.1. Closeness

- Two questions:

  1. Under which dimensions we are going to define whether two individuals are close each other? (**Selection of variables**)

  2. How are we going to measure the distance between these individuals? (**Distance measures**)

- **Selection of variables**

  - Include all variables known to be related both to the treatment assignment and the outcome (Heckman et al 1998).

  - Small sample issues: Priority to critical variables according previous knowledge and economic theory

- Critical: not to choose variables potentially affected by the treatment

- **Distance measures**:
  1. Exact
  2. Metric-based
     a. Euclidean
     b. Mahalanobis
     c. Abadie-Imbens
  3. Propensity score-based

- Exact:

$$d_{ij} = \begin{cases} 0, & \text{if } X_i = X_j \\ \infty, & \text{if } X_i \neq X_j \end{cases}$$

- Euclidean metric:

$$d_{ij} = (X_i - X_j)'(X_i - X_j)$$

- Mahalanobis (See Robin and Thomas 1992):

$$d_{ij} = (X_i - X_j)'\Sigma^{-1}(X_i - X_j)$$

Σ^{-1} : Covariance matrix of covariates

- Abadie and Imbens (2010):

$$d_{ij} = (X_i - X_j)' \operatorname{diag}(\Sigma^{-1})(X_i - X_j)$$

- Propensity score:

$$d_{ij} = | p_i(x) - p_j(x) |$$

- Linear propensity score (See Rubin 2001):

$$d_{ij} = | \operatorname{logit}[p_i(x)] - \operatorname{logit}[p_j(x)] |$$

- See Zhao (2004) for a discussion of alternative metrics
- Metrics can be combined: Mahalanobis and propensity score (Rubin and Thomas 2000)

- How to choose a metric?
  - Exact and Mahalanobis don't work well when X is high dimensional
  - Exact matching can lead to a larger bias than inexact matching if many treated individuals are not matched to control ones (Rosenbaum and Rubin 1985). Solution: Coarsened exact matching (Iacus et al 2011)
  - When X<8, Mahalanobis works well (Zhao 2004)
  - Little experience with other metric than Mahalanobis (Imbens 2004) but not clear winner in simulations (Zhao 2004).
  - Propensity score is the solution?

4.2. Matching methods

- Matching comes in many flavors:

1. Exact matching
2. Nearest neighbor matching
3. Sub-classification and full matching
4. Weighting adjustments

Generic matching

- Some extra notation:

  N_T : Treated units

  N_C : Control units

- Define N_T sets of weights with N_C weights in each set:

  $w(i, j)$: $(i : 1, 2, \ldots N_T, j : 1, 2, \ldots N_C)$

- Where: $\sum w(i, j) = 1$

- Define:

$$(11)\ Y_i(0) = \begin{cases} Y_i \text{ if } D_i = 0 \\ \sum_j w(i, j) Y_j \text{ if } D_i = 1 \end{cases}$$

$$(12)\ Y_i(1) = \begin{cases} \sum_j w(i,j) Y_j \ \text{ if } D_i = 0 \\ Y_i \ \text{ if } D_i = 1 \end{cases}$$

- The **generic matching estimators** are:

$$(13)\ \hat{\tau}_{ATE} = \frac{1}{N} \sum_i \left[Y_i(1) - Y_i(0) \right]$$

$$\hat{\tau}_{ATT} = \frac{1}{N_T} \sum_{i \in \{D=1\}} \left[Y_i - Y_i(0) \right]$$

$$\hat{\tau}_{ATU} = \frac{1}{N_C} \sum_{i \in \{D=0\}} \left[Y_i(1) - Y_i \right]$$

- Another way to write ATT:

$$(14)\ \tau_{ATT} = \frac{1}{N_T} \sum_{i \in \{D=1\}} \left[Y_i - Y_i(0) \right] = \frac{1}{N_T} \sum_{i \in \{D=1\}} \left[y_{1i} - \sum_{j \in \{D=0\}} w(i,j) y_{0j} \right]$$

a. Exact matching

- The **exact matching estimator** for ATT:

$$(15)\ \hat{\tau}_{exact,ATT} = \frac{1}{N_T} \sum_{i \in \{D=1\}} \left[y_{1i}(X_i) - \sum_{j \in \{D=0\}} w(i,j) y_{0j}(X_j) \right]$$

- Where:

$$w(i,j) = \begin{cases} \frac{1}{k} & \text{if } X_i = X_j \\ 0 & \text{if } X_i \neq X_j \end{cases}$$

- Problems: sample size issues and loss of information

b. Nearest neighbor matching

- Define:

$$(16)\ Y_i(0) = \begin{cases} Y_i \text{ if } D_i = 0 \\ \frac{1}{M}\sum_j Y_j \text{ if } D_i = 1 \end{cases}; \quad Y_i(1) = \begin{cases} \frac{1}{M}\sum_j Y_j \text{ if } D_i = 0 \\ Y_i \text{ if } D_i = 1 \end{cases}$$

- The **simple nearest neighbor matching estimator** for M matches for unit with replacement is given by:

$$(17)\ \hat{\tau}_{ATE} = \frac{1}{N}\sum_i \left[Y_i(1) - Y_i(0) \right]$$

- For M=1, does this level of M reduce power?

- Not necessarily. Power reduction is often minimal:
  - Depends on size of the smaller group. If treatment group stays the same, this won't matter that much
  - Power increases when groups are more similar
- On the other hand, matching with M>1 can lead to some poor matches
- Some NN-matching methods variants:
  - Optimal matching
    - Order in which controls are matched matters for quality
    - Solution: take into account all set of matches minimizing a global distance measure (Rosenmbaum 2002)

- Some alternative using propensity score procedures:
  - Stratification on the score
  - Nearest neighbor matching on the score
  - Radius matching on the score
  - Kernel matching on the score
  - Weighting on the basis of the score

5. Applications: Arcenaux et al (2006)

- They use data from a large scale field experiment (60,000 treated vs. 2 million of control units!) and a rich set of covariates in order to assess the performance of matching estimators
- Matching estimators are severely biased!
- The experiment:
  - Conducted in Iowa and Michigan before 2002 election
  - Treatment: Get-out-the-Vote campaign

> Hello, may I speak with (name of person) please? Hi. This is (caller's name) calling from Vote 2002, a nonpartisan effort working to encourage citizens to vote. We just wanted to remind you that elections are being held this Tuesday. The success of our democracy depends on whether we exercise our right to vote or not, so we hope you'll come out and vote this Tuesday. Can I count on you to vote next Tuesday?

Table 4 Experimental benchmark estimates of the effect of phone calls on turnout in Iowa and Michigan, 2002

	All observations				*Sample containing exact matches only*	
	Sample excludes unlisted numbers		*Sample includes unlisted numbers*		*Excluding unlisted numbers*	*Including unlisted numbers*
Covariates	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*
Phone contact	0.4 (0.5)	0.5 (0.4)	−0.0 (0.6)	0.3 (0.5)	0.5 (0.6)	0.2 (0.7)
State dummy(1=Iowa)	7.4 (0.4)	2.6 (1.3)	3.3 (0.3)	3.6 (1.1)		
Competitiveness dummy in Michigan	4.9 (0.1)	−1.8 (0.3)	5.0 (0.1)	−1.4 (0.3)		
Competitiveness dummy in Iowa	6.1 (0.2)	−0.7 (0.7)	5.8 (0.1)	−1.6 (0.6)		
Household size		7.0 (0.1)		8.0 (0.1)		
Age		0.3 (0.002)		0.3 (0.002)		
Female		−1.2 (0.1)		−1.2 (0.1)		
Newly registered		5.5 (0.1)		8.1 (0.1)		
Vote in 2000		37.1 (0.1)		38.2 (0.1)		
Vote in 1998		21.7 (0.1)		22.2 (0.1)		
Missing values in female dummy		−32.1 (0.2)		−29.2 (0.2)		
Constant	46.1 (0.1)	[a]	43.9 (0.1)	[a]		
N	1,905,320	1,905,320	2,474,927	2,474,927	499,836	781,780
F	5,649.20	4,128.26	3,855.41	5,786.29	1,085.00	1,319.74
Adjusted R^2	0.01	0.29	0.01	0.30	0.01	0.01

Note. Two-stage least squares estimates: Vote 2002 = a + β_1 contact + β_2 MI competitiveness + β_3 IA competitiveness + β_4 state dummy + $\Sigma\gamma_i$ covariates$_i$.
Instrument: Random assignment to treatment group.
[a]Dummy variables for state-house district, state-senate district, and county are included but not shown to save space.

Table 5 Biased OLS estimates of the effect of actual contact on turnout in Iowa and Michigan, 2002

	All observations				*Sample containing exact matches only*	
	Sample excludes unlisted numbers		*Sample includes unlisted numbers*		*Excluding unlisted numbers*	*Including unlisted numbers*
*Covariates**	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*	*Coefficient (robust SE)*
Phone contact	6.2 (0.3)	2.7 (0.3)	10.7 (0.3)	4.4 (0.3)	2.7 (0.3)	4.2 (0.3)
State dummy(1=Iowa)	6.7 (0.3)	2.4 (1.3)	2.5 (0.3)	3.3 (0.1)		
Competitiveness dummy in Michigan	4.8 (0.1)	−1.8 (0.3)	4.9 (0.1)	−1.5 (0.3)		
Competitiveness dummy in Iowa	6.4 (0.2)	−0.6 (0.7)	6.1 (0.1)	−1.5 (0.6)		
Household size		7.0 (0.1)		8.0 (0.1)		
Age		0.3 (0.002)		0.3 (0.002)		
Female		−1.2 (0.1)		−1.2 (0.1)		
Newly registered		5.5 (0.1)		8.1 (0.1)		
Vote in 2000		37.1 (0.1)		38.2 (0.1)		
Vote in 1998		21.7 (0.1)		22.2 (0.1)		
Missing values in female dummy		−32.1 (0.2)		−29.2 (0.2)		
Constant	46.1 (0.1)	[a]	44.0 (0.1)	[a]		
N	1,905,320	1,905,320	2,474,927	2,474,927	243,736	309,535
F	5,745.62	4,129.75	4,141.81	5,791.01	51.45	52.44
Adjusted R^2	0.01	0.29	0.01	0.30	0.01	0.01

Note. Entries are OLS estimates.

[a]Dummy variables for state-house district, state-senate district, and county are included but not shown to save space.

Table 6 Sensitivity of OLS and Matching Estimates to Changes in Sample Definition

	OLS	*Exact matching*	*Inexact matching*
Sample excludes unlisted numbers			
Treatment effect[a]	2.7	2.8	2.9
(SE)	(0.3)	(0.3)[b]	(0.1)[b]
(SE)		(0.4)[c]	(0.3)[c]
N[d]	1,905,320	22,711	25,028
Matched[e]		90.7%	99.9%
R^2	0.28		
Sample includes unlisted numbers			
Treatment effect[a]	4.4	4.4	4.4
(SE)	(0.3)	(0.3)[b]	(0.1)[b]
(SE)		(0.3)[c]	(0.3)[c]
N[d]	2,474,927	23,467	25,034
Matched[e]		93.7%	99.9%
R^2	0.30		

[a]For the OLS results, the treatment effect is the slope coefficient on the contact variable included in the regression: Vote 2002 $= a + \beta_1$ contact $+ \Sigma\gamma_i$ covariates$_i$.

[b]Standard errors for treatment effects estimated with Abadie and Imbens (2004) method (see appendix for details).

[c]Standard errors for treatment effects estimated with Becker and Ichino (2002) method (see appendix for details).

[d]For the matching analysis, this indicates the number of contact group individuals who were matched to the control group; for the OLS analysis it indicates the total number of observations.

[e]Percent of contacted group with at least one identical match in the control group.