Matching Estimators

Ethan Kaplan

March 2, 2009

1 Matching Estimators: Motivation

- What are matching estimators?
  - Individual Matching: Match observations and estimate between individually matched observations.

$$\sum_{i=1}^{N} w_i \left[Y_i \left(T = 1 \right) - Y_{M(i)} \left(T = 0 \right) \right]$$

where N is the number of treated observations, $Y_i \left(T = 1 \right)$ is the outcome for the i^{th} treated observation, $Y_{M(i)} \left(T = 0 \right)$ is the matched observation for the i^{th} treated observation, and w_i is the population weight of the i^{th} treated observation. (Note that each treated observation is matched to at most one untreated observation)

  - Block Matching: Match groups of similar observations (on covariates):

$$\sum_{i=1}^{N} w_i \left[\frac{\sum_{j=1}^{M} Y_{ij} \left(T = 1 \right)}{Z_{i,T=1}} - \frac{\sum_{j=1}^{M} Y_{M(i)} \left(T = 0 \right)}{Z_{i,T=0}} \right]$$

where Y_{ij} is the j^{th} observation in the i^{ih} group and $Z_{i,T=1}$ is the number of observations in the i^{th} treatment group (and similarly for $Z_{i,T=0}$)

- Main questions with matching estimators:
  - How to do matching?
  - How to compute the weights (w_i)?
  - How to computer the standard errors?
- Many methods:
  - Criterion
    * Exact covariate matching
    * Propensity score matching
    * Mahalanobis matching

$$\min_{\{ik,ij\}} \left(X_{ik} - X_{jk}\right) S^{-1} \left(X_{ik} - X_{jk}\right)$$

- Matching Techniques
  - Nearest neighbor matching (For every treatment, find the nearest control - with or without replacement)
  - Genetic algorithm matching
  - Many others!

- Why use matching estimators? Under Gauss Markov Assumptions, OLS is BLUE (Best Linear Unbiased Estimator)
  - $Y_{it} = \alpha + \beta X_{it} + \epsilon_{it}$
  - $cov\left(X_{it}, \epsilon_{it}\right) = 0$
  - $cov\left(\epsilon_{it}, \epsilon_{jt}\right) = 0$ where $i \neq j$
  - $V\left(\epsilon_{it}\right) = \sigma^2$

- Matching Estimators are linear in treatment (and thus in variables since matching estimators usually only include treatment variables as RHS variables). So why use matching estimators as opposed to OLS?

- Tradeoff: Efficiency vs. Robustness
  - What does robustness mean?
    * Must know ALL the relevant covariates
    * Lack of knowledge of functional form
    * Must have alot of data
    * Then conditional on covariates, only average difference in outcome comes from treatment. However, any specific functional form for OLS or NLS may be mis-specified. So, look only within covariate groupings.
  - Problem: No models (that Im aware of) of robustness

– Other possibilities of more robust, local estimators with less variance (greater efficiency): non-parametric and semi-parametric matching

2 Matching Estimators: Estimation

- Use Rubin's Potential Outcomes Model

- Assume:

  1. Unconfoundedness: $(Y_i(1), Y_i(0)) \amalg W_i | X_i$
     - also called selection on observables assumption
     - twins example
  2. Overlap: $0 < \Pr(W_i = 1 | X_i) < 1$
  3. (1) & (2) together are called "strongly ignorable treatment"

- Given strongly ignorable treatment:

$$E[Y_i(1) - Y_i(0) | X_i = x] =$$

$$E[Y_i(1) | X_i = x] - Y[Y_i(0) | X_i = x] =$$

$$E\left[Y_i\left(1\right)|X_i = x, W_i = 1\right] - E\left[Y_i\left(0\right)|X_i = x, W_i = 0\right] =$$

$$E\left[Y_i|X_i, W_i = 1\right] - E\left[Y_i|X_i, W_i = 0\right]$$

- How Different is Matching from OLS really?
  - Defintion: saturated models are models where there is a dummy variable for every covariate realization.
  - Example: LHS: Wages, RHS: Education (University Completion, HS Completion, Less Than High School), Race (Black, White), Sex (Female, Male). Transform variables into dummies (11 dummy variables with one category left out as the constant):
    1. University Completion, Black, Female
    2. University Completion, Black, Male
    3. University Completion, White, Female

4. University Completion, White, Male
5. HS Completion, Black, Female
6. HS Completion, Black, Male
7. HS Completion, White, Female
8. HS Completion, White, Male
9. $<$ HS, Black, Female
10. $<$ HS, Black, Male
11. $<$ HS, White, Female
12. $<$ HS, White, Male

– This replicates exact covariate matching though with different weights than matching estimators:

– Matching: (From Mostly Harmless Econometric,

Angrist and Pischke):

$$\widehat{\beta}_T^{Match} = \frac{E_X R_x}{\sum_x R_x}$$

where

$$\begin{aligned} R_x &= P(W_i = 1|X_i = x)\, P(X_l = x) \\ E_x &= E[Y_{i1}|X_i, W_i = 1] - E[Y_{i1}|X_i, W_i = 0] \end{aligned}$$

- OLS:

$$\widehat{\beta}_T^{OLS} = \frac{E_x R_x [1 - P(W_i = 1|X_i = x)]}{\sum_x R_x [1 - P(W_i = 1|X_i = x)]}$$

- So, OLS weights by variance of the observations, matching estimators by their population frequency.

- Is this a fair characterization of the differences between OLS and Matching? So, is matching a weighted OLS?

- Covariate Balance (Rosenbaum and Rubin, 1985)

- With discrete variables and large samples, we may be able to look within actual covariate bins

- With small sample or continuous variables, we will not be able to match exactly. But:

  * Matching should lead to covariate balance across treatment and control: $X_{i,T} - X_{M(i),T} \tilde{} N\left(0, \sigma^2\right)$

  * In other words, covariates should be randomly distributed across treatment and control for matched observations

- The Search Problem

  - Many search problems have an exponential asymptotic

  - Finding the optimal set of matches is such a problem.

  - E.G., we want to estimate ATT and do matching without replacement:

- With 10 treated and 20 control obs: 184,756 possible matches
- With 20 treated and 40 control obs: 13,784,652,8820 possible matches
- With 40 treated and 80 control obs: 1.075e+23
- With 185 treated and 260 control: 1.633e+69 with 185 treated and 4000 control: computer infinity
- Matching with replacement makes the search problem explode even more quickly.

- Propensity Score Matching (Rosenbaum and Rubin, 1985) : Dimensional Reduction
  - Suppose that unconfoundness holds, then:

$$\left(Y_i\left(1\right), Y_i\left(0\right)\right) \amalg W_i | p\left(X_i\right)$$

    where $p\left(X_i\right)$ is the probability of treatment (estimated via linear probability, probit, or logit)

- Intuition: by leaving out covariates, we introduce ommitted variables bias. However, since $X_i \amalg W_i | p(X_i)$ (covariate balance), adding X_i will not change the estimate of Y_i on X_i

- Steps for propensity score estimation

  1. First estimate selection equation

  $$T_i = F(X_i) + \epsilon_i$$

  2. Estimate fitted probabilites of selection

  $$\hat{F}(X_i)$$

  3. Create bins of a given width (or do nearest neighbor matching)

  4. Check for covariate balance across treatment and control within bins

5. Estimate difference between treatment and control within bins

$$\frac{\sum_{j=1}^{M} Y_{ij}\left(T=1\right)}{Z_{i,T=1}} - \frac{\sum_{j=1}^{M} Y_{M(i)}\left(T=0\right)}{Z_{i,T=0}}$$

6. Choose weights

   (a) Homogeneous treatment effect: weight by size of bin or other measures of variance of estimate

   (b) Heterogeneous treatment effect: weight using population weights

7. Estimate average treatment effect by weighting across bins:

$$\sum_{i=1}^{N} w_i \left[\frac{\sum_{j=1}^{M} Y_{ij}\left(T=1\right)}{Z_{i,T=1}} - \frac{\sum_{j=1}^{M} Y_{M(i)}\left(T=0\right)}{Z_{i,T=0}}\right]$$

8. Estimate standard errors for average treatment effect

   (a) Estimate component by component

$$\left[\text{from } \sigma^2_{X,T=1},\ \sigma^2_{X,T=0},\ \bar{Y}_{T=1},\ \bar{Y}_{T=0},\ p\left(x\right)\right]$$

(b) Bootstrap

- Problems with bootstrapping IV estimators (at least nearest neighbor matching) due to non-linearities: Abadie, Imbens (2006)

- Some problems with FE: Interlude on Measurement Error (From Ashenfelter and Krueger, 1994)

  - Attenuation:

    True:

    $$y_i = \beta x_i + \epsilon_i$$

    Observed

    $$\bar{x}_i = x_i + \delta_i$$

$$\hat{\beta} = \frac{\sum_{t=1}^{t} x_i y_i}{\sum_{t=1}^{t} x_i^2} =$$

$$\frac{\sum_{t=1}^{t} \bar{x}_i \left(\beta x_i + \epsilon_i\right)}{\sum_{t=1}^{t} \bar{x}_i^2} = \frac{\sum_{t=1}^{t} \left(\beta x_i^2 + \beta \delta_i x_i + x_i \epsilon_i + \delta_i \epsilon_i\right)}{\sum_{t=1}^{t} \left(x_i^2 + 2 x_i \delta_i + \delta_i^2\right)}$$

$$E\hat{\beta} = \frac{\beta \sum_{t=1}^{t} x_i^2}{\sum_{t=1}^{t} \left(x_i^2 + \delta_i^2\right)} = \frac{\beta \sigma_x^2}{\sigma_x^2 + \sigma_\delta^2} = \beta \left(1 - \frac{\sigma_\delta^2}{\sigma_x^2 + \sigma_\delta^2}\right) < \beta$$

- – $\frac{\sigma_\delta^2}{\sigma_x^2 + \sigma_\delta^2}$ is called the reliability ratio

– Measurement Error Tradeoff

Suppose T=2;

$$y_{1i} = \alpha Z_i + \beta X_{1i} + \mu_i + \epsilon_{1i}$$

$$y_{2i} = \alpha Z_i + \beta X_{2i} + \mu_i + \epsilon_{2i}$$

$$\mu_i = \gamma X_{1i} + \gamma X_{2i} + \delta Z_i + \omega_i$$

Then $\hat{\beta}_{FE} = \hat{\beta}_{FD}$ comes from the regression $y_{1i} - y_{2i} = \beta (X_{1i} - X_{2i}) + \epsilon_{1i} - \epsilon_{2i}$

It can be shown that

$$\hat{\beta}_{FE} = \beta \left(1 - \frac{\sigma_\delta^2}{\left[\sigma_x^2 + \sigma_\delta^2\right](1 - \rho_X)} \right)$$

where ρ_X is the correlation coefficient of X within the "fixed effect" group:$\frac{cov(X_{1i}X_{2i})}{\sigma_X^2}$

- – So there is a tradeoff: bias from exclusion of the fixed effect versus bias due to exacerbation of the attenuation in the presence of measurement error with highly correlated X's.

  – Intuition: if the X's are highly correlated, then when using fixed effects, most of the variation left is measurement error.

  – Relavance to matching: One can think of matching as a type of fixed effect. Could be exacerbation of attenuation in the presence of measurement error in treatment. What about measurement error in covariates? Non-classical measurement error? Not yet studied!

- Tradeoff: Matching often allows for better controls, less bias but at the cost of efficiency.

- Many observations thrown away.

- Also since emphasis is on average treatment effect, in the presence of heterogeneity, must use population weights as opposed to weighting be inverse variance: efficiency loss.

- Propensity Score can help when overlap is low (Angrist and Han, 2004)

  * Dont have to throw away observations with low overlap

  * Can gain in efficiency even without gain from less observations thrown out due to greater comparisons per bin

- Previews of Things to Come: Comparison with IV

  - IV estimates ATE if

    1. Homogeneous treatment effect

2. Set of compliers is the entire population

- Matching Estimators estimate ATE if

1. Homogeneous Treatment Effect

2. Overlap satisfied at all parts of distribution of covariates (i.e. over full support of covariates)

- Understanding Matching: Future Econometric Research
  - Does matching help with omitted variable bias? No!
  - Can matching help when there is functional form uncertainty and no omitted variable bias (Robustness)? Yes!
  - Can matching help with specification bias? Yes (functional form) and no (variable selection)!

- Unknown: Matching and Measurement Error (In Treatment and in Covariates)
- Unknown: Constructing SEs for Matching Estimators
- Unknown: Balancing Bias (Due to Functional Form) with Efficiency

3 Bootstrapping

- Method for estimating standard errors when techniques don't exist for estimating SEs from econometric theory (for example small sample distributions-i.e. IV), when SE computation is to computationally intensive.

- Conisder

$$Y = XB + \epsilon$$

- Non-Parametric Bootstrapping
  - 1. Estimate true $\hat{\beta}$ from the full sample
    2. Choose N observations at random (with replacement)
    3. Estimate $\hat{\beta}_j$

4. Estimate J of the $\hat{\beta}_j$

5. Either

   (a) Test $\hat{\beta}$ relative to the non-parametric distribution of $\hat{\beta}_j$

   (b) or compute the variance of the $\hat{\beta}_j$: $\sum_{j \in j} \frac{\left(\bar{\beta} - \hat{\beta}_j\right)^2}{|J|}$ where $\bar{\beta} = \sum_{j \in J} \frac{\hat{\beta}_j}{|J|}$; test using the normality assumption with $\sqrt{V\left(\hat{\beta}_j\right)}$

- Paramtetric Bootstrapping

  – 1. Estimate $\hat{\beta}$ from the full sample

  2. Calculate the residuals: $\epsilon_i = Y_i - X_i \hat{\beta}$

  3. Take the full sample of X_i; for each X_i, resample a residuals ϵ_{i_j} at random

4. Create a sample of N pairs $\left(X_i, Y_{i_j}\right)$ where $Y_{i_j} = \hat{\beta} X_i + \epsilon_{i_j}$

5. Run a regression for each sample and obtain a distribution $\hat{\beta}_m$

6. Either

   (a) Test $\hat{\beta}$ relative to the non-parametric distribution of $\hat{\beta}_m$

   (b) or compute the variance of the $\hat{\beta}_m$: $\sum_{m \in M} \frac{\left(\bar{\beta} - \hat{\beta}_m\right)^2}{|M|}$ where $\bar{\beta} = \sum_{m \in M} \frac{\hat{\beta}_m}{|M|}$ and test using the normality assumption with $\sqrt{V\left(\hat{\beta}_m\right)}$

- Which bootstrap method is preferable?
  - With parametric, you resample ϵ for the different X, which is what you want to do so non-parametric is preferable but:

- If $cov\left(X, \epsilon\right) \neq 0$, then your $\hat{\beta}$ which you use to compute Y_{i_j} will be tainted; in this case, it is better to use the parametric bootstrapping

- Block Bootstrap
  - Suppose $cov\left(\epsilon_i, \epsilon_j\right) \neq 0$ for $i \neq j$
  - Then you can block bootstrap (i.e. randomly pick K sequential observations at a time)
  - This way, you randomly sample blocks of data which keeps the error structure in tact

4 Jackknife

- Similar to bootstrap
- Estimate $\hat{\beta}$ from the full sample
- Define $\hat{\beta}_j$ = estimate without the j^{th} observation (could exclude more than one)
- Like boostrapping but
  1. Without replacement
  2. Constructed by excluding variables rather than including them
- Then: either
  1. Test $\hat{\beta}$ relative to the non-parametric distribution of $\hat{\beta}_j$

2. or compute the variance of the $\hat{\beta}_j$: $\sum_{j \in j} \frac{\left(\bar{\beta} - \hat{\beta}_j\right)^2}{|J|}$ where $\bar{\beta} = \sum_{j \in J} \frac{\hat{\beta}_j}{|J|}$; test using the normality assumption with $\sqrt{V\left(\hat{\beta}_j\right)}$