# Matching Methods

Michael R. Roberts

Department of Finance
The Wharton School
University of Pennsylvania

July 28, 2009

# Matching Intuition

- Matching estimates the missing counterfactual by using the information of subjects from the control group that are "close" in some sense.
- E.g., Estimate weight loss effect of a new diet
  1. For each person who followed the diet, find a "similar" person who didn't.
     1. Similar on height, weight, occupation, health, etc.
  2. Difference between the average weight loss for the dieters and non-dieters is the weight loss (gain?) effect of the diet.
- This talk will follow closely the review article by Imbens (2004)

# Statistical Software

- Stata & Matlab: "match" (Abadie et al. (2001, 2003))
- Stata: "psmatch" (Sianesi (2001),
- Stata: "psmatch2" (Sianesi and Leuven (2001), Todd (2001))
  - http://econpapers.repec.org/software/bocbocode/S432001.htm
  - http://athena.sas.upenn.edu/ petra/copen/statadoc.pdf
- Stata: "pscare", "att*" (Becker and Ichino (2002))
- SAS:
  - Kawabata et al.: http://www2.sas.com/proceedings/sugi29/173-29.pdf
  - Perraillon: http://www2.sas.com/proceedings/forum2007/185-2007.pdf
  - Mandrekar: http://www2.sas.com/proceedings/sugi29/208-29.pdf
  - Several macros (gmatch, match, vmatch): http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm

# Notation 1

- Random sample: $N$ units (e.g., firms) indexed by $i = 1, ..., N$
- For each unit $i$
  - Treatment indicator (observed): $D_i \in \{0, 1\}$
  - Pair of **Potential Outcomes** (unobserved):

$$\begin{aligned} Y_i(0) \quad & \text{if } D_i = 0 \text{ (outcome under treatment)} \\ Y_i(1) \quad & \text{if } D_i = 1 \text{ (outcome under NO treatment)} \end{aligned}$$

  - Realized outcome (observed): $Y_i$:

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

    which can be written as:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

  - Treatment effect or impact (estimable): $\tau$

$$\tau = Y(1) - Y(0)$$

# Notation 2

- For each unit $i$
  - Vector of characteristics, $X_i$, *unaffected* by treatment (e.g., variables measured prior to treatment)
  - Propensity Score (estimable): $ps(x)$

$$ps(x) \equiv Pr(D = 1|X = x) = E(D|X = x)$$

- Observed triple is $(Y_i, D_i, X_i) \implies$ the following distributions are (are not) recoverable from the data

$$\begin{aligned} \text{Recoverable} &: F(Y(0)|X, D = 0); F(Y(1)|X, D = 1) \\ \text{Unrecoverable} &: F(Y(0), Y(1)|X, D) \\ \text{Unrecoverable} &: F(Y(0), Y(1)|X) \\ \text{Unrecoverable} &: F(\tau|X, D) \end{aligned}$$

- So, we estimate a moment, typically mean, of impact dist.

# Covariates

- Why must covariates be unaffected by treatment? Consider ATT

$$\begin{aligned} & E[Y(1) - Y(0)|D = 1] \\ = \; & E[Y(1)|D = 1] - E[Y(0)|D = 1] \\ = \; & E[Y(1)|D = 1] - E[E[Y(0)|D = 1, X = x]|D = 1] \text{ tower} \\ = \; & E[Y(1)|D = 1] - E[E[Y(0)|D = 0, X = x]|D = 1] \text{ unconf} \end{aligned}$$

  Note

$$\begin{aligned} & E[E[Y(0)|D = 0, X = x]|D = 1] \\ = \; & \int_{x \in X} \int_{y \in Y} y f(y|D = 0, x) f(x|D = 1) dx \end{aligned}$$

- $f(x|D = 1)$ represents the density that *would* have been observed in the no treatment state ($D = 0$).

$\therefore$ Receipt of treatment better not change density of $Z$

# Population Treatment Effects

- Average Treatment Effect (ATE)

$$E[Y(1) - Y(0)]$$

  Effect of treatment on *entire population*
- Average Treatment Effect for the Treated (ATT) (Rubin (1977), Heckman & Robb (1984))

$$E[Y(1) - Y(0)|D = 1]$$

  Effect of treatment on *treated subpopulation*
  - Could be more relevant when a program is aimed at a subpopulation, such as disadvantaged individuals, small firms, etc.
- Average Treatment Effect for the Untreated or Controls (ATU, ATC)

$$E[Y(1) - Y(0)|D = 0]$$

  Effect of treatment on *control subpopulation*

# Key Assumption #1: Unconfoundedness

- Unconfoundedness assumption (let $\perp\!\!\!\perp$ denote independence)

$$(Y(0), Y(1)) \perp\!\!\!\perp D|X \tag{1}$$

  - "ignorable treatment assignment" (Rosenbaum and Rubin (1983))
  - "conditional independence" (Lechner (1999, 2002))
  - "selection on observables" (Barnow, Cain, and Goldberger (1908))
- This assumption says outcomes $(Y(0), Y(1))$ are independent of participation status $(D)$ conditional on $X$.
- Equivalent expressions of condition (1)

$$\begin{aligned} Pr(D = 1|Y(0), Y(1), X) &= Pr(D = 1|X), \text{ or} \\ E(D = 1|Y(0), Y(1), X) &= E(D = 1|X) \end{aligned}$$

# Unconfoundedness & Exogeneity

- Similar to standard regression exogeneity assumption.
- If treatment effect ($\tau$) is constant $\forall i$ and

$$Y_i(0) = \alpha + X_i'\beta + \varepsilon_i$$

with $\varepsilon \perp\!\!\!\perp X_i$, then

$$Y_i = \alpha + \tau D_i + X_i'\beta + \varepsilon_i$$

- Unconfoundedness $\equiv$ to independence of $D_i$ and $\varepsilon_i$ conditional on $X_i$. (i.e., $D_i$ is exogenous)
- Without constant treatment effect assumption, unconfoundedness doesn't imply linear relation with mean independent errors

# Key Assumption #2: Overlap

- Overlap is an assumption on the joint distribution of treatments ($D$) and covariates ($X$)

$$0 < Pr(D = 1|X) < 1$$

- Intuition: For each $X$, $\exists$ strictly positive probability of being in the treatment group ($Pr(D = 1|X)$) and the control group ($1 - Pr(D = 1|X)$)
- Why is this important?
  - Imagine a value of $X$, $x'$, for which this didn't hold (i.e., $Pr(D = 1|X = x') = 1$)
  - This means there are only treatment units with $X = x'$, no controls with this value, and so no controls that are really comparable.
  - Therefore, no good obs to estimate counterfactual

# Unconfoundedness & Overlap

- If assumptions #1 and #2 hold we can substitute the $Y(0)$ distribution observed for matched on $X$ non-participants for the missing participant $Y(0)$ distribution.
- I.e., we can treat the outcome of the non-participants that have similar covariates as the participants as if it were the counterfactual outcome for the participants.

# Academic Debate Over Uncounfoundedness & Overlap

- Agent's optimizing behavior *precludes* choices being independent of potential outcomes, regardless of covariate conditioning
  - Agent's select into programs for many reasons $\Longrightarrow$ unconfoundedness is inherently violated
- Still several reasons to investigate ATE
  1. Data-description...nocausality
  2. Unconfoundedness requires that all variables that need to be adjusted for are observed by researcher
     - Strong assumption but economic theory can help identify the vars
  3. Even if agents choose treatment optimally, agents with same observables can differ in treatment choices without invalidating unconfoundedness if choices driven by unobserved differences unrelated to outcomes.
  4. If we restrict how individuals form expectations about unknown potential outcomes, unconfoundedness may hold (Heckman, Lalonde, and Smith (2000))

# Useful Facts

- Recall that the observed outcome $Y$ can be written

$$Y = DY(1) + (1 - D)Y(0)$$

- This implies

$$E[Y|D = 0] = E[DY(1) + (1 - D)Y(0)|D = 0] = E[Y(0)|D = 0]$$
$$E[Y|D = 1] = E[DY(1) + (1 - D)Y(0)|D = 1] = E[Y(1)|D = 1]$$

# Identification of ATE 1

- Write the ATE for a subpopulation with a certain $X = x$, ATE(x), in terms of observables.

$$
\begin{aligned}
ATE(x) &= E[Y(1) - Y(0)|X = x] \text{ def.} \\
&= E[Y(1) - Y(0)|X = x, D = d] \text{ unconf.} \\
&= E[Y(1)|X = x, D = 1] - E[Y(0)|X = x, D = 0] \\
&= E[DY(1) + (1 - D)Y(0)|X = x, D = 1] \\
&- E[DY(1) + (1 - D)Y(0)|X = x, D = 0] \\
&= E[Y|X = x, D = 1] - E[Y|X = x, D = 0] \text{ def of Y}
\end{aligned}
$$

# Identification of ATE 2

- We need to be able to estimate the expectations comprising ATE(x)

$$E[Y(1)|X = x, D = 1] \text{ and } E[Y(0)|X = x, D = 0]$$

- This is where we need overlap. If overlap violated at $X = x$ then we couldn't estimate *both* of the expectations since there wouldn't be any observations to estimate one of them.
- We need to both unconfoundedness *and* overlap for identification of the ATE

# Weakening Unconfoundedness 1

- Mean Independence Assumption

$$E(Y(d)|D,X) = E(Y(d)|X)$$

  for d = 0,1.
- Weaker version of unconfoundedness (Heckman, Ichimura, and Todd (1998))
- In practice, hard to make a case for this assumption without also making one for unconfoundedness.
- Mean independence intrinsically tied to functional-form assumptions,
  - $\Longrightarrow$ one cannot identify average effects on transformations of original outcome (e.g., logarithms) without unconfoundedness

# Weakening Unconfoundedness & Overlap for ATT

- If interest only in ATT, can weaken both key assumptions (Heckman et al. (1997))
- Unconfoundedness for controls:

$$Y(0) \perp\!\!\!\perp D|X$$

- Overlap for controls:

$$Pr(D = 1|X) < 1$$

- These assumptions sufficient for identification of ATT because moments of distribution of $Y(1)$ for treated are observable.

$$E(Y(1)|D = 1) = E(Y|D = 1)$$

# Identification of ATT

- Write ATT in terms of observables

$$\begin{aligned} ATT(x) &= E[Y(1) - Y(0)|X = x, D = 1] \\ &= E[Y(1)|X = x, D = 1] - E[Y(0)|X = x, D = 1] \\ &= E[Y|X = x, D = 1] - E[Y(0)|X = x, D = 1] \\ &= E[Y|X = x, D = 1] - E[Y(0)|X = x, D = 0] \text{ (unconf.)} \\ &= E[Y|X = x, D = 1] - E[Y|X = x, D = 0] \text{ (unconf.)} \\ &= ATE(x) \end{aligned}$$

- To get unconditional ATT, need to average over appropriate distribution of $X$ *conditioning on treatment*
- Overlap is only needed at points $x$ where there is a treatment unit

# Propensity Score

- All biases due to observable covariates can be removed by conditioning on propensity score (Rosenbaum and Rubin (1983))

$$(Y(0), Y(1)) \perp\!\!\!\perp D|X \implies (Y(0), Y(1)) \perp\!\!\!\perp D|ps(X)$$

- Intuition, conditioning on propensity score, $ps(X)$, has same effect as conditioning on all covariates $X$.
- So, when matching on $X$ is valid (under key assumptions #1 and #2) so too is matching on $ps(X)$

# Propensity Score Result Proof

- $Pr(D = 1|Y(0), Y(1), ps(X)) =$

$$
\begin{aligned}
&= E(D = 1|Y(0), Y(1), ps(X)) \text{ def} \\
&= E[E(D = 1|Y(0), Y(1), ps(X), X)|Y(0), Y(1), ps(X)] \text{ tower} \\
&= E[E(D = 1|Y(0), Y(1), X)|Y(0), Y(1), ps(X)] \\
&= E[E(D = 1|X)|Y(0), Y(1), ps(X)] \text{ unconf.} \\
&= E[ps(X)|Y(0), Y(1), ps(X)] \text{ def.} \\
&= ps(X)
\end{aligned}
$$

- This shows that

$$Pr(D = 1|Y(0), Y(1), ps(X)) = Pr(D = 1|ps(X)) = ps(X)$$

implying independence of $(Y(0), Y(1))$ and $D$ conditional on $ps(X)$

# Intuition for Propensity Score Results

- Consider regression model

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2' X_i + \varepsilon_i$$

- Bias on $\beta_1$ from omitting $X$ equals $\beta_2'\delta$, where $\delta$ is vector of coefficients on $D$ in regressions of each element of $X$ on $D$.
- By conditioning on propensity score, we remove correlation between $X$ and $D$ because $X \perp\!\!\!\perp D|ps(X)$
- So, omitting $X$ no longer leads to bias (but may lead to inefficiency).

# Estimators

1. **Regression Estimators** rely on consistent estimation of conditional regression functions

$$E(Y_d|X = x) \text{ for } d = 0, 1$$

2. **Matching Estimators** compare outcomes across pairs of matched treated and control units
   - each unit matched to fixed # of obs with opposite treatment
   - As $N \to 0$, bias of within-pair ests $\to 0$, but variance doesn't because # of matches constant
3. **Propensity Score (PS)** estimators
   1. Weighting by reciprocal of PS
   2. Blocking on the PS
   3. Regression on the PS
   4. Matching on the PS
4. **Mixed Methods**

# Estimation of ATE

- Recall to estimate ATE (and ATT) we need to estimate conditional expectations of potential outcomes

$$\hat{\mu}_1(x) \rightarrow \mu_1(x) \equiv E(Y(1)|X = x)$$
$$\hat{\mu}_0(x) \rightarrow \mu_0(x) \equiv E(Y(0)|X = x)$$

- ATE estimated by averaging difference over empirical distribution of covariates

$$\begin{aligned} \hat{ATE} &= \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} D_i\left[Y_i - \hat{\mu}_0(x_i)\right] + (1 - D_i)\left[\hat{\mu}_1(x_i) - Y_i\right] \end{aligned}$$

# Interpretation

- From previous slide

$$\begin{aligned} \hat{ATE} &= \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} D_i\left[Y_i - \hat{\mu}_0(x_i)\right] + (1 - D_i)\left[\hat{\mu}_1(x_i) - Y_i\right] \end{aligned}$$

- First term is avg outcome of treated obs $(D_i Y_i)$ minus avg estimated counterfactual for treated obs $(D_i \hat{\mu}_0(X_i))$
- Second term is avg outcome of control obs $((1 - D_i)Y_i)$ minus avg estimated counterfactual for control obs $((1 - D_i)\hat{\mu}_0(X_i))$

# Estimation of ATT

- For ATT only control regression function is estimated $\Longrightarrow$ only need to predict control outcomes for treatment obs

$$\hat{ATT} = \frac{1}{N_T} \sum_{i=1}^{N} D_i \left[ Y_i - \hat{\mu}_0(x_i) \right]$$

  - First term is avg outcome of treated obs ($D_i Y_i$) minus avg estimated counterfactual for treated obs ($D_i \hat{\mu}_0(X_i)$)

# OLS Estimation of Regression Function $\mu_d(x)$

- OLS with regression function: $\mu_d(x) = \beta' x + \tau d$
  - ATE $= \tau$ since

$$\hat{ATE} = \hat{\mu}_1(x) - \hat{\mu}_0(x) = [\beta' x + \tau] - [\beta' x] = \tau$$

  - OLS on $Y_i = \alpha + \beta' X_i + \tau D_i + \varepsilon_i$
- OLS with regression functions: $\mu_d(x) = \beta_d' x$
  - Estimate 2 separate regressions on 2 subsamples (treatment & control)
  - Substitute predicted values into $\hat{ATE}$ equation
- Nonparametric regression (Rubin (1977))

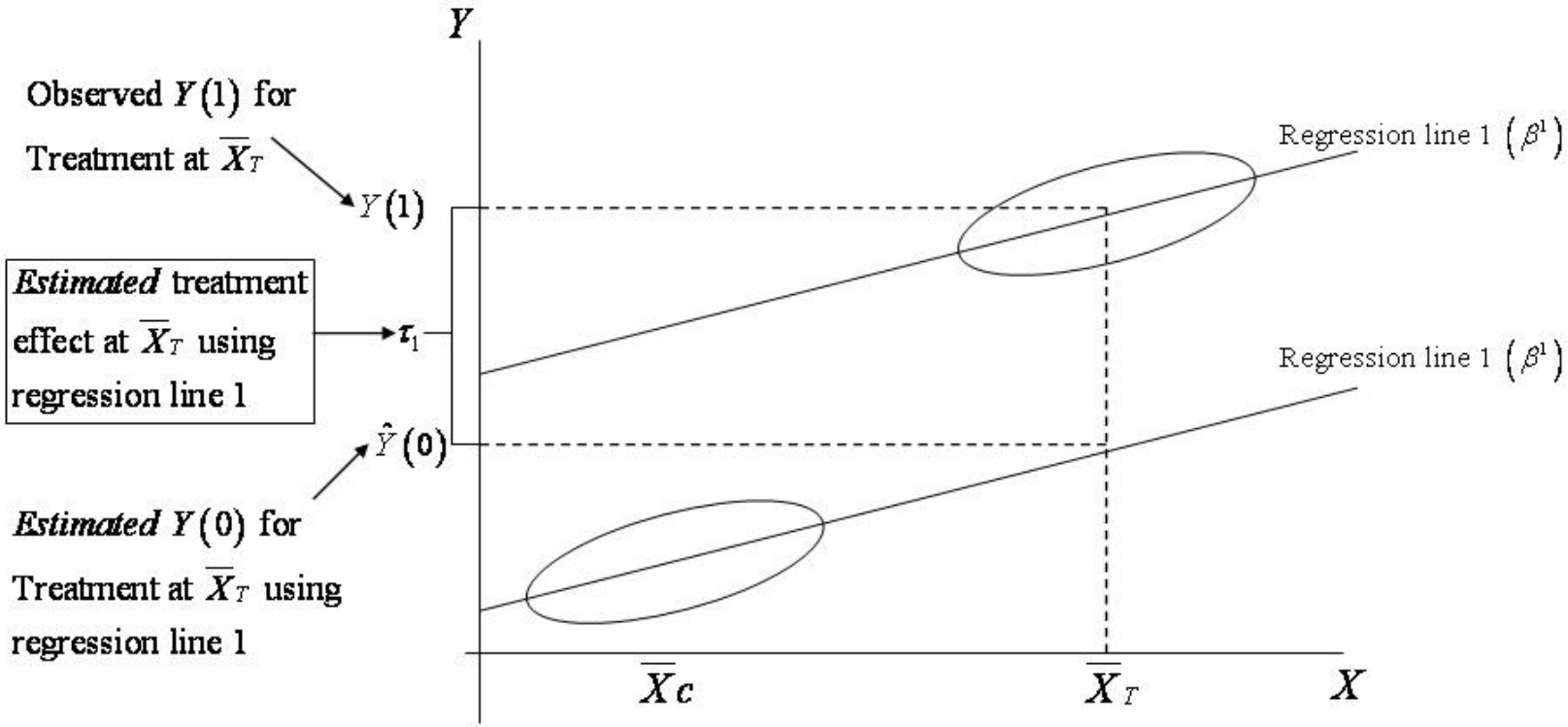
# Potential Concerns

- Regression estimators rely heavily on extrapolation
  - $\Longrightarrow$ estimates can be very sensitive to differences in covariate distributions for treated and control units
- Intuition:
  - Estimate missing outcomes for treated using regression function for the controls (& vice versa)
  - On avg., want to predict the control outcome at $\bar{x}_1$, the avg. covariate value for the treated
  - With linear regression, avg prediction is

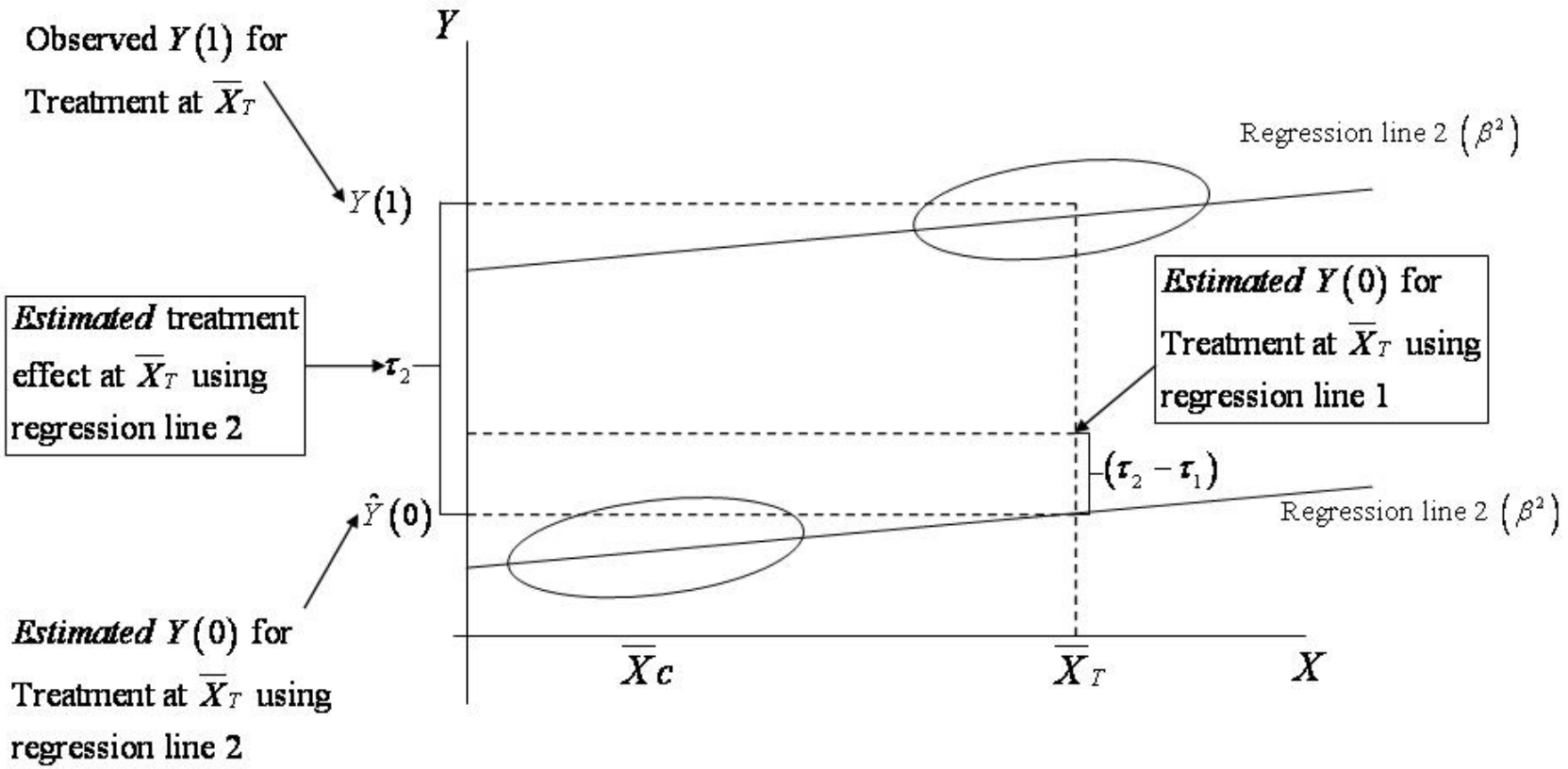
$$\bar{Y}(0) + \beta'(\bar{X}(1) - \bar{X}(0))$$

  - When covariate avgs are close, coefficient has little impact
  - When covariate avgs not close, prediction based on linear regression can be very sensitive to changes in specification

# Dissimilar Group Characteristics 1



# Dissimilar Group Characteristics 2

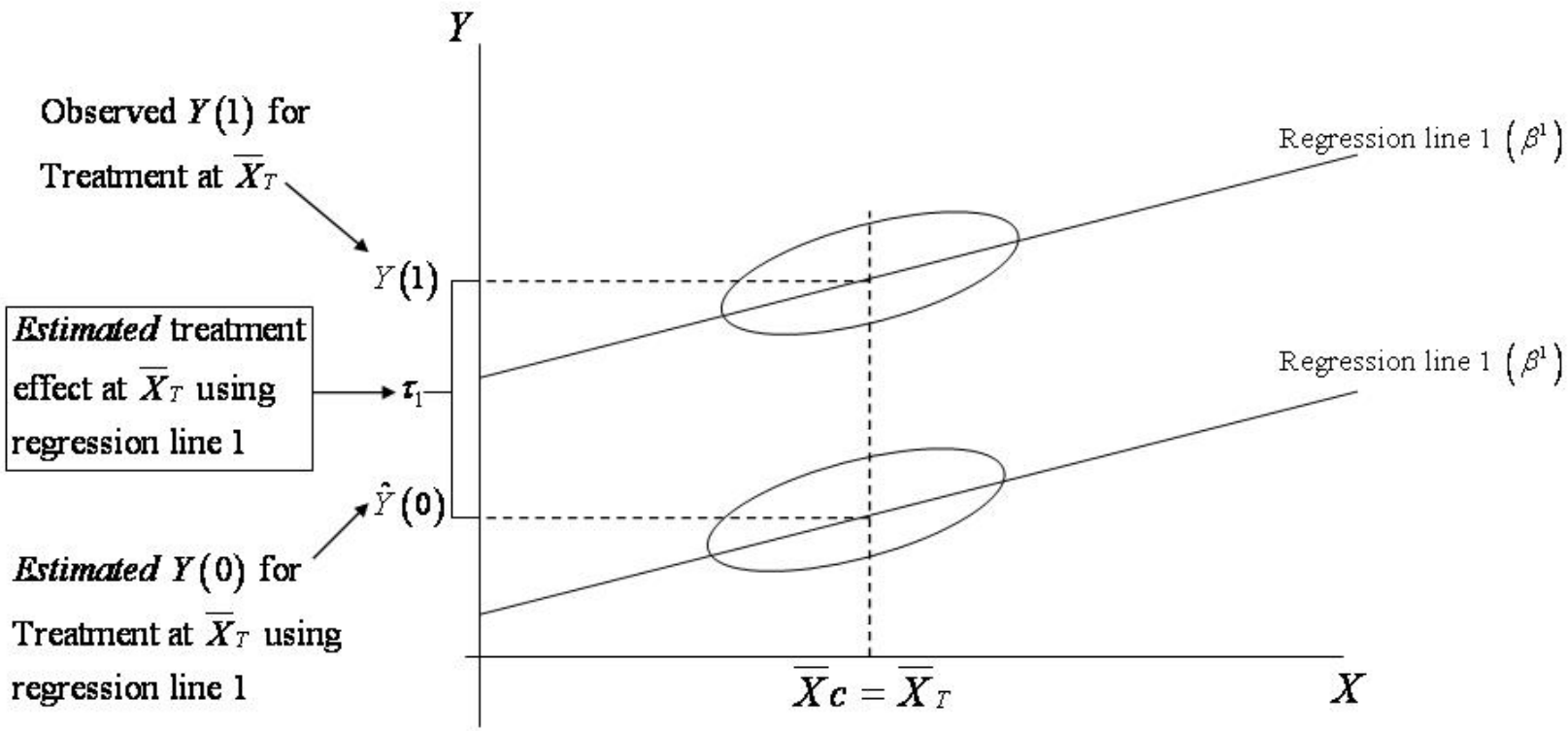


# Dissimilar Group Characteristics Discussions

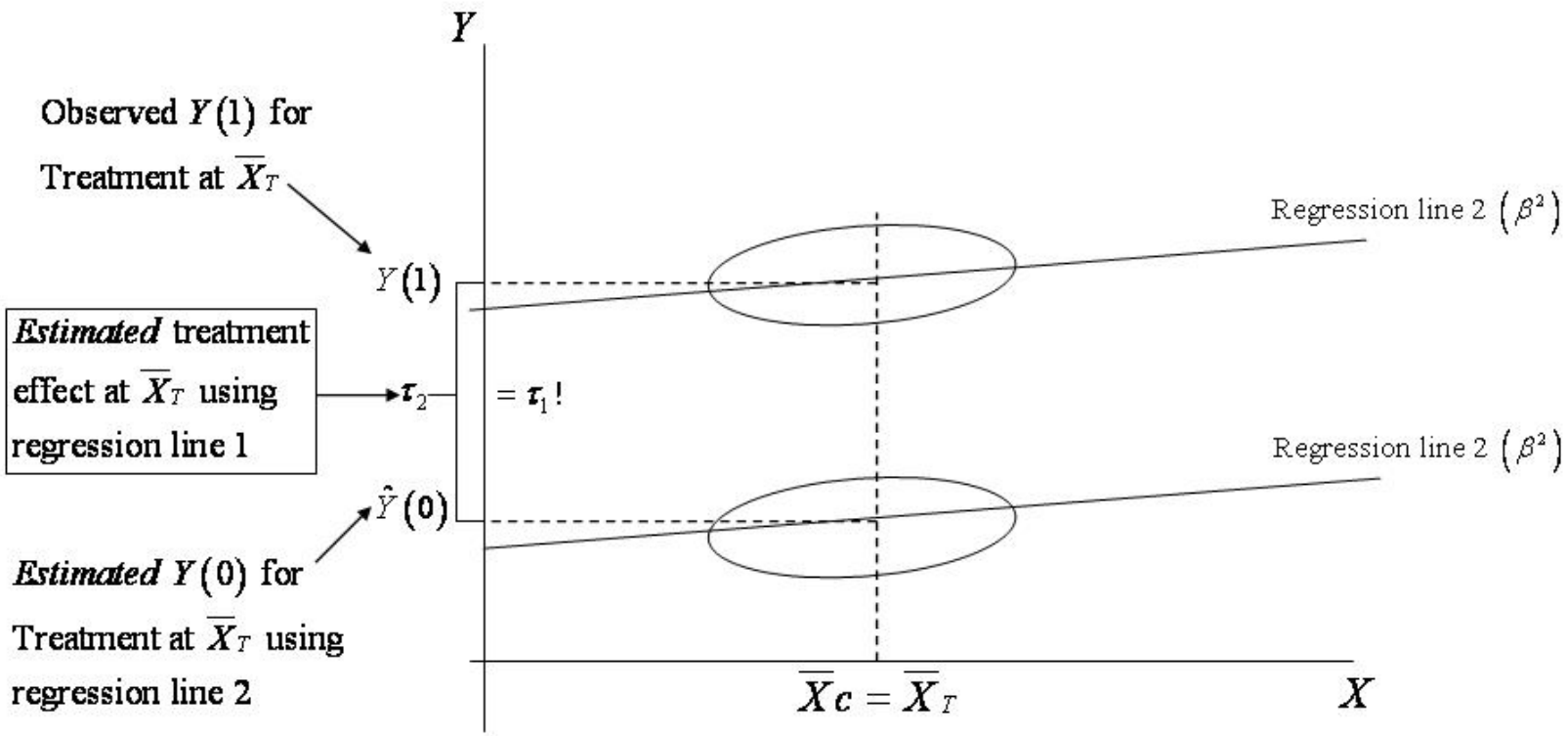
- Slight change in slope of estimated regression equation, $\beta_1$ to $\beta_2$, leads to a large change in the estimated effect from $\tau_1$ to $\tau_2$
- This sensitivity is due to the dissimilarity of the treatment and control groups along the $X$ dimension

$$\bar{Y}(1) + \beta' \underbrace{(\bar{X}(1) - \bar{X}(0))}$$

# Similar Group Characteristics 1



# Similar Group Characteristics 2



# Similar Group Characteristics Discussions

- No change in estimated effect
- Regression lines rotate through point of averages
- Recall

$$\bar{Y}(0) + \beta' \underbrace{(\bar{X}(1) - \bar{X}(0))}$$

  so $\bar{X}(1) = \bar{X}(0) \implies$ second term is 0
- Lesson: Treatment and Control groups better be similar along observables!
  - Means are important but other moments matter as well.

# Nonparametric Estimators - Hahn (1998)

- Estimate nonparametrically using series methods the following:

$$\begin{aligned} g_1(x) &= E(DY|X) \\ g_0(x) &= E((1-D)Y|X) \\ ps(x) &= E(D|X) \end{aligned}$$

- With $g_1, g_2, e$ we can estimate the regression functions $\mu_1(x)$ and $\mu_0(x)$ as follows

$$\hat{\mu}_1(x) = \frac{\hat{g}_1(x)}{\hat{ps}(x)}$$

$$\hat{\mu}_0(x) = \frac{\hat{g}_0(x)}{1-\hat{ps}(x)}$$

- See Imbens, Newey, and Ridder (2003) for refinement

# Nonparametric Estimators - Heckman et al. (1997,1998)

- Simple kernel estimator:

$$\hat{\mu}_d(x) = \frac{\sum_{i:D_i=d} Y_i \cdot K\left(\frac{X_i - x}{h}\right)}{\sum_{i:D_i=d} K\left(\frac{X_i - x}{h}\right)}$$

  with kernel $K(\cdot)$, bandwidth $h$.

- Local linear kernel regression. Regression function $\mu_d(x)$ estimated by $\beta_0$ in

$$\min_{\beta_0,\beta_1} \sum_{i:D_i=d} [Y_i - \beta_0 - \beta_1'(X_i - x)]^2 \cdot K\left(\frac{X_i - x}{h}\right)$$

# Nonparametric Estimators - Loose Ends

- Choice of kernel less important than bandwidth (i.e., smoothing parameter)
- Choice of smoothing parameter?
  - In Hahn, # of terms in the series
  - In Heckman et al., bandwidth
- Not a lot of guidance here. Robustness should be overarching concern.

# Overview

- Regression based estimators impute missing potential value (i.e., counterfactual) using the estimated regression function, $\hat{\mu}_d(x)$.
- Matching based estimators impute missing potential value using only the outcomes of nearest neighbors of the opposite treatment group
  - # of neighbors is like bandwidth in nonparametric regression
  - Matching estimators are unbiased but inconsistent (# of matches doesn't change as sample size grows)
  - Regression estimators rely on consistency of $\hat{\mu}_d(x)$
  - Less neighbors $\implies$ less bias and less precision
  - Given matching metric, only need choose # of neighbors
  - Given matched pairs, treatment effect within pair is difference in outcomes. ATT estimator is average of within pair differences.
  - Matching examples: Gu & Rosenbaum (1993); Rosenbaum (1989,1995,2002), Rubin (1973, 1979), Heckman et al. (1998), Dehejia & Wahba (1999), Abadie & Imbens (2002)

# Abadie & Imbens (2002) Estimator

- Loose descendant of Dehejia and Wahba (1999)
- Sample $(Y_i, X_i, D_i), i = 1, ..., N$
- Let $l_m(i)$ be the index $l : D_l \neq D_i$ and

$$\sum_{j|D_j \neq D_i} I(||X_j - X_i|| \leq ||X_l - X_i||) = m$$

- Intuition: $l$ is the index of the unit in the opposite group that is the $m^{th}$ closest to unit $i$ in terms of the distance measure based on the norm $||\cdot||$.
  - $l_1(i)$ is the nearest match for unit $i$

# Abadie & Imbens (2002) Estimator (Cont)

- Let $Ł_M(i)$ be the set of indices for the first $M$ matches to unit $i$. $Ł_M(i) = \{l_1(i), ..., l_M(i)\}$
- Imputed potential outcomes for $i$ are

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if} \quad D_i = 0, \\ \frac{1}{M}\sum_{j \in Ł_M(i)} Y_j & \text{if} \quad D_i = 1, \end{cases}$$

$$\hat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j \in Ł_M(i)} Y_j & \text{if} \quad D_i = 0, \\ Y_i & \text{if} \quad D_i = 1, \end{cases}$$

- Simple matching estimator is:

$$\frac{1}{N}\sum_{i=1}^{N}[\hat{Y}_i(1) - \hat{Y}_i(0)]$$

# Abadie & Imbens (2002) Estimator - Loose Ends

- Since estimator is just difference between two sample means, can use standard methods to compute SEs
- This estimator is biased and bias doesn't disappear with large $N$ and at least 3 covariates!
- How do we choose the # of matches?
- What is the distance metric?

$$\begin{aligned} \text{Euclidean} &: d_E(x,z) = (x-z)'(x-z) \\ \text{Standardized} &: d_S(x,z) = (x-z)'diag(\Sigma_X^{-1})(x-z) \\ \text{Mahalanobis} &: d_M(x,z) = (x-z)'\Sigma_X^{-1}(x-z) \end{aligned}$$

where $\Sigma_X$ is the covariance matrix of the covariates, and $diag(\cdot)$ is the matrix consisting of zero everywhere but the diagonal.

# Overview

- Matching requires "adjusting" directly for all covariates
- Propensity Score Matching requires "adjusting" only for the propensity score.
- Several different Propensity Score (PS) based estimators:
  1. Weighting by reciprocal of PS
  2. Blocking on the PS
  3. Regression on the PS
  4. Matching on the PS

# Estimating the Propensity Score

- A number of options
- Key consideration is accuracy and robustness
  1. OLS
  2. Discrete Choice Model (e.g., Logit, Probit)
  3. Nonparametric approach (e.g., series estimator, kernel regression, sieve estimator)

# Weighting Estimators 1

- Weighting estimators use PSs as weights to balance sample of treatment and controls
- Note difference in avg outcomes for treatment and control groups

$$\hat{ATE} = \frac{\sum D_i Y_i}{\sum D_i} - \frac{\sum (1 - D_i) Y_i}{\sum 1 - W_i}$$

  is **not** unbiased estimator for ATE $= E(Y_1 - Y_0)$
- Problem is conditional on treatment indicator, distributions of covariates differ.
- Formally

$$E\left[\frac{DY}{ps(X)}\right] = E\left[\frac{DY_1}{ps(X)}\right] = E\left[E\left[\frac{DY_1}{ps(X)}\middle| X\right]\right]$$

# Weighting Estimators Problem

- Formally, the problem is:

$$
\begin{aligned}
E\left[\frac{DY}{ps(X)}\right] &= E\left[\frac{DY(1)}{ps(X)}\right] = E\left[E\left[\frac{DY(1)}{ps(X)}\middle| X\right]\right] \text{ (tower)} \\
&= E\left[\frac{1}{ps(X)}E\left[DY(1)\middle| X\right]\right] \\
&= E\left[\frac{1}{ps(X)}E[D|X]E[Y(1)|X]\right] \text{ (unconf)} \\
&= E\left[\frac{ps(X)}{ps(X)}E[Y(1)|X]\right] = E[Y(1)]
\end{aligned}
$$

- Similarly,

$$
E\left[\frac{(1-D)Y}{1-ps(X)}\right] = E[Y(0)]
$$

# ATE Weighting Estimator

- The ATE is equal to

$$ATE = E\left[\frac{DY}{ps(X)} - \frac{(1-D)Y}{1-ps(X)}\right]$$

- The weighting propensity score estimator of ATE is equal to

$$\hat{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i Y_i}{\hat{ps}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{ps}(X_i)}\right]$$

# Normalizing the Weights

- Problem: Weights don't sum to 1 (only in expectation). So, normalize

$$\begin{aligned} \hat{ATE} &= \sum_{i=1}^{N} \frac{D_i Y_i}{\hat{ps}(X_i)} \Bigg/ \sum_{i=1}^{N} \frac{D_i}{\hat{ps}(X_i)} \\ &- \sum_{i=1}^{N} \frac{(1-D_i) Y_i}{1-\hat{ps}(X_i)} \Bigg/ \sum_{i=1}^{N} \frac{1-D_i}{1-\hat{ps}(X_i)} \end{aligned}$$

# ATT Weighting Estimator

- The ATT estimator is:

$$\begin{aligned} \hat{ATT} &= \left[ \frac{1}{N_1} \sum_{i:D_i=1} Y_i \right] \\ &- \left[ \sum_{i:D_i=0} Y_i \cdot \frac{\hat{ps}(X_i)}{1-\hat{ps}(X_i)} \Big/ \sum_{i:D_i=0} \cdot \frac{\hat{ps}(X_i)}{1-\hat{ps}(X_i)} \right] \end{aligned}$$

# Weighting Estimators Loose Ends

- Choice of smoothing parameters
  - Hirano, Imbens & Ridder (2003) use series estimators $\Longrightarrow$ need to choose # of terms
  - Ichimura and Linton (2001) use kernel version $\Longrightarrow$ need to choose bandwidth

# Blocking on the Propensity Score

- Originally suggested by Rosenbaum & Rubin (1983)
- The recipe/intuition
  1. Estimate propensity score (parametrically or nonparametrically)
  2. Divide sample into $M$ blocks of units of approximately equal probability of treatment
     - Implement by dividing unit interval into $M$ blocks with boundary values equal to $m/M$ for $m = 1, ..., M-1$ so

$$J_{im} = I\left\{\frac{m-1}{M} < ps(X_i) \leq \frac{m}{M}\right\}$$

     $J_{im}$ is indicator for unit $i$ being in block $m$
  3. Within each block, $N_{dm}$ obs with treatment $= d$

$$N_{dm} = \sum_i I\left\{D_i = d, J_{im} = 1\right\}$$

# Blocking on the PS 2

- Estimate within each block, avg. treatment effect as if random assignment held

$$\hat{ATE}_m = \frac{1}{N_{1m}} \sum_{i=1}^{N} J_{im} D_i Y_i - \frac{1}{N_{0m}} \sum_{i=1}^{N} J_{im}(1 - D_i) Y_i$$

Intuition:

- $J_{im}$ identifies the units in bloc $m$,
- $D_i$ identifies the treated units, and
- $U_i$ is the outcome.
- $1^{st}$ sum is average outcome of treatment units in block $m$
- $2^{nd}$ sum is average outcome of control units in block $m$

# Blocking on the PS ATE & ATT

- Overall ATE is:

$$\hat{ATE} = \sum_{m=1}^{M} \hat{ATE}_m \cdot \frac{N_{1m} + N_{0m}}{N}$$

- Overall ATT is:

$$\hat{ATE} = \sum_{m=1}^{M} \hat{ATE}_m \cdot \frac{N_{1m}}{N_T}$$

where $N_T$ is the number of treated units

# Blocking on the PS Loose Ends

- Akin to nonparametric regression where unknown fxn is approximated by step fxn with fixed jump points
- How many blocks to use in practice?
  - Rule of thumb suggest 5 blocks (Cochran (1968))
  - Should check balance of covariates within each block
  - If true PS is constant within a block, the distribution of the covariates among treatment and control should be identical (i.e., covariates should be balanced)
  - Assess adequacy of model by comparing distribution of covariates among treated and controls within blocks
  - If distributions are different, we can
    1. split blocks into subblocks, or
    2. generalize specification of PS
  - If within-block PS is unbalanced, blocks too large & need to be split
  - If within-block PS is balanced but covariates unbalanced, PS specification is inadequate

# Regression on the PS

- This method estimates conditional expectation of $Y$ given $D$ and $ps(X)$

$$E[Y(d)|ps(X) = e] \tag{2}$$

- Unconfoundedness implies

$$E[Y(d)|ps(X) = e] = E[Y|D = d, ps(X) = e]$$

- If we have an estimator for eqn 2, $\hat{v}_d(e)$, then we can estimate ATE as

$$\hat{ATE} = \frac{1}{N}\sum_{i=1}^{N}[\hat{v}_1(e(X_i)) - \hat{v}_0(e(X_i))]$$

- Heckman et al. (1998) consider local linear version estimator.
- Hahn (1998) consider series estimator (less efficient than local linear).

# Matching on the PS

- Recall: it's sufficient to adjust solely for differences in the PS between treatment and control units.
- One way to adjust for differences in covariates is matching.
- Therefore, we can use the propensity score to match treatment and control units
- Problem: matching on the estimated PS produces an estimated ATE (or ATT) for which there is no known variance formula.

# Overview

- Mixed methods combine two of the three (regression, matching, PS) methods.
- One method alone is sufficient to obtain consistent (or even) efficient estimates, incorporating regression may eliminate remaining bias and improve precision
- E.g., Robins and Ritov (1997) mix weighting and regression to produce *double robustness*
- Methods that combine matching & regression are robust against misspecification of the regression function

# Weighting and Regression

- Recall weighting estimator above:

$$\hat{ATE} = \sum_{i=1}^{N} \frac{D_i Y_i}{\hat{ps}(X_i)} \Big/ \sum_{i=1}^{N} \frac{D_i}{\hat{ps}(X_i)} - \sum_{i=1}^{N} \frac{(1-D_i)Y_i}{1-\hat{ps}(X_i)} \Big/ \sum_{i=1}^{N} \frac{1-D_i}{1-\hat{ps}(X_i)}$$

- We can rewrite this as estimating the following regression fxn by weighted least squares

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$

with weights

$$\lambda_i = \sqrt{\frac{D_i}{e(X_i)} + \frac{1-D_i}{1-e(X_i)}}$$

- The weights ensure that the covariates are uncorrelated with the treatment indicator $\Longrightarrow$ WLS estimator is consistent.

# Refining Precision

- We can add covariates to the regression fxn to improve precision

$$Y_i = \alpha + \beta' X_i + \tau D_i + \varepsilon_i$$

  with the same weights
- Other references: Robins and Roznitzky (1995), Robins, Roznitzky, and Zhao (1995), Robins and Ritov (1997), Hirano and Imbens (2001).
- If either the regression model *or* the propensity score correctly specified, the estimator is consistent (i.e., *doubly robust*)

# Blocking and Regression

- Rosenbaum and Rubin (1983b) modify blocking by using least squares regression within the blocks
- Note, the estimated treatment effect within blocks can be written as a least squares estimator, $\tau_m$, for the regression fxn

$$Y_i = \alpha + \tau_m D_i + \varepsilon_i$$

using only units in block $m$

- We can also add covariates

$$Y_i = \alpha\beta' X_i + \tau_m D_i + \varepsilon_i$$

using only units in block $m$

# Matching and Regression 1

- Abadie and Imbens (2002) show that bias of simple matching estimator can dominate variance if dimension of covariates is too large.
- Additional bias corrections through regression can be help
- Recall from the Matching estimators section

$$\hat{Y}_{i0} = \begin{cases} Y_i & \text{if} \quad D_i = 0, \\ \frac{1}{M} \sum_{j \in \text{Ł}_M(i)} Y_j & \text{if} \quad D_i = 1, \end{cases}$$

$$\hat{Y}_{i1} = \begin{cases} \frac{1}{M} \sum_{j \in \text{Ł}_M(i)} Y_j & \text{if} \quad D_i = 0, \\ Y_i & \text{if} \quad D_i = 1, \end{cases}$$

$\hat{Y}_{i0}$ and $\hat{Y}_{i1}$ are observed or imputed potential outcome for $i$

- Bias arises because covariates $X_i$ and $X_{l(i)}$ (the covariates of $i$'s match) aren't equal, though they're close from matching

# Matching and Regression 2

- Consider single match case and for each unit define

$$\hat{X}_{i0} = \begin{cases} X_i & \text{if} \quad D_i = 0, \\ X_{l_1(i)} & \text{if} \quad D_i = 1, \end{cases}$$

$$\hat{X}_{i1} = \begin{cases} X_{l_1(i)} & \text{if} \quad D_i = 0 \\ X_i & \text{if} \quad D_i = 1, \end{cases}$$

- If match is exact, $\hat{X}_{i0} = \hat{X}_{i1}$ for each unit
- If not, discrepancies may lead to bias
- Difference $\hat{X}_{i1} - \hat{X}_{i0}$ can be used to reduce bias of simple matching estimator

# Matching and Regression 3

- Assume for unit $i$, $D_i = 1$, which $\implies \hat{Y}_{i1} = Y_i(1)$ and $\hat{Y}_{i0}$ is an imputed value for $Y_i(0)$
- Imputed value unbiased for $\hat{\mu}_0(X_{l_1(i)}) \equiv E(Y(0)|X_{l_1(i)})$ since $\hat{Y}_{i0} = Y_{l_1(i)}$, but not necessarily $\hat{\mu}_0(X_i) = E(Y(0)|X_i)$
- We can adjust $\hat{Y}_{i0}$ by an estimate of $\mu_0(X_i) = \mu_0(X_{l_1(i)})$
- Typically assume corrections are linear in difference in covariates of unit $i$ and its match

$$\beta_0'[\hat{X}_{i1} - \hat{X}_{i0}] = \beta_0'[X_i - X_{l_1(i)}]$$

- Rubin (1973b) suggests 3 corrections differing in how $\beta_0$ is estimated

# Matching and Regression: Bias Correction 1

- We can write matching estimator as OLS estimator for regression fxn

$$\hat{Y}_{i1} - \hat{Y}_{i0} = \tau + \varepsilon_i$$

- Simple modification to the regression fxn

$$\hat{Y}_{i1} - \hat{Y}_{i0} = \tau + [\hat{X}_{i1} - \hat{X}_{i0}]'\beta + \varepsilon_i$$

which we can estimate via OLS

# Matching and Regression: Bias Correction 2

- Estimate $\mu_0(x)$ directly by taking all control unites and estimating a linear regression

$$Y_i = \alpha_0 + \beta_0' X_i + \varepsilon_i$$

by OLS

- If unit $i$ is a control unit, the correction will be done using an estimator for the regression fxn $\mu_1(x)$ based on a linear specification

$$Y_i = \alpha_1 + \beta_1' X_i + \varepsilon_i$$

estimated on the treated units

- See Abadie and Imbens (2002) for further details

# Matching and Regression: Bias Correction 3

- Estimate same regression fxn for controls using only those controls that are used as matches for treated units, with weights corresponding to # of times a control observation is used as a match. a linear regression
- Potentially inefficient (discards some control obs and weights some more than others) but only uses the *relevant* matches
- Discarded controls may be outliers relative to treated obs and may unduly affect OLS estimates
- See Abadie and Imbens (2002) for further details

# Variance of ATE

- Variance of efficient estimators of ATE is:

$$V = E\left[\frac{\sigma_1^2(X)}{ps(X)} + \frac{\sigma_0^2(X)}{1-ps(X)} + (\mu_1(X) - \mu_0(X) - \tau)^2\right]$$

- Three ways to estimate this
  1. Brute force: estimate all five components using kernel methods or series.
  2. ??
  3. Bootstrapping: Seems ok for regression and PS methods but matching may cause problems because it introduces discreteness in the distribution that will lead to ties in the matching algorithm (Politis and Romano (1999)

# Unconfoundedness

- To be clear: **the unconfoundedness assumption is not directly testable**
- Unconfoundedness says: the conditional distribution of the outcome under the control treatment, $Y(0)$, given receipt of the active treatment and given covariates $(D = 1, X = x)$, is *identical* to the distribution of the control outcome given receipt of the control treatment and given covariates $(D = 0, X = x)$
  - Same is assumed for the distribution of the active treatment outcome, $Y(1)$
- Problem is we don't observe counterfactual so we can never directly reject unconfoundedness
- Two broad groups of tests to indirectly assess unconfoundedness based on falsification tests (Heckman and Hotz (1989) and Rosenbaum (1987)

# Falsification Tests using Multiple Control Groups

- Estimate causal effect of a treatment known not to have an affect, relying on presence of multiple control groups (Rosenbaum (1987))
- For example, we can replace the treatment group with one of the control groups and use the other control group for comparison
- Not rejecting the test doesn't imply the unconfoundedness assumption is valid (both control groups could suffer the same bias), but nonrejection where the 2 control groups are likely to have different potential biases makes it more plausible that unconfoundedness assumption holds.

# Falsification Tests using Unaffected Variables

- Estimate causal effect of treatment on a variable known not to be affected by treatment (e.g., variable whose value is determined prior to treatment)
- E.g., consider effect of treatment on a lagged outcome relying on presence of multiple control groups (Rosenbaum (1987))
- For example, we can replace the treatment group with one of the control groups and use the other control group for comparison
- Not rejecting the test doesn't imply the unconfoundedness assumption is valid (both control groups could suffer the same bias), but nonrejection where the 2 control groups are likely to have different potential biases makes it more plausible that unconfoundedness assumption holds.

# Issues

- There may be some variables for which we should *not* adjust for.
- We may be better off in finite samples ignoring variables with weak correlation with the treatment indicator and the outcomes because they reduce precision
- Unfortunately no hard fast rules
- Big concern is including covariates affected by the treatment (e.g., intermediate outcomes). This is a no no!
- Make sure covariates are measured before the treatment was chosen
- Think hard about what covariates to include
- See Rosenbaum (1984b) and Angrist and Kruger (2000)

# Propensity Score

- Recall overlap requires that the PS be between 0 and 1
- This assumption raises several questions:
  1. How to detect a lack of overlap in the covariate distribution
  2. How to deal with lack of overlap given a lack exists
  3. How individual estimation methods address lack of overlap
- Matching is valid only in the region of common support!

# Detecting Lack of Overlap

- Plot distributions of covariates by treatment groups
- In 1,2 dimensions, this is easy
- In higher dimensions, can look at pairs of marginals but they may not be informative about lack of overlap
- More useful is to inspect distribution of PSs in treatment and control groups
  - Need to estimate PS nonparametrically
  - But, misspecification may lead to failure in detecting a lack of overlap
  - May wish to undersmooth the estimation of the PS, either by choosing a bandwidth smaller than optimal for by including higher-order terms in a series expansion
- Inspect quality of worst matches. For each component $k$ of covariates $X$ inspect $max_i|x_{i,k} - X_{l_1(i),k}]$ (max over all obs of matching discrepancy) If this difference is large relative to SD of $k^{th}$ component of covariates, be worried

# Addressing Lack of Overlap

- Given lack of overlap
  1. conclude that ATE cannot be estimated with sufficient precision, or
  2. decide to focus on ATE that is estimable with greater accuracy
- For 2, we can:
  - discard some of the bad matches or treatment and control obs with PSs above/below a certain value (remember PS must be between 0 and 1)
  - Only accept matches where difference in PS is below a certain value
  - Drop matches where individual covariates are severely mismatched

# Handling Lack of Overlap with each Method

- Assume we have data with sufficient overlap and we want to estimate ATT
- Now add a few treated obs with outlying values. Doing so $\Longrightarrow$
  - we can't estimate ATE as precisely, because we lack suitable controls against which to compare additional units
  - Variance estimate should increase
- Now add a few control obs with outlying values. Doing so $\Longrightarrow$
  - little effect since outlying controls are irrelevant for ATT (not for ATE!)
  - Methods appropriately dealing with limited overlap should show estimates approx unchanged in bias and precision

# Handling Lack of Overlap with Regression

- Adding obs with outlying values of regressors $\Longrightarrow$ more precise estimates
  - If added obs are treated units, precision of estimated control regression fxn at these outlying values will be lower (few control units are in this outlying region) $\Longrightarrow$ variance increases
  - If added obs are control units, then precision of control regression fxn will increase *spuriously*.
- Punchline: Regression methods can be misleading in cases with limited overlap

# Handling Lack of Overlap with Matching

- Adding controls with outlying obs has little affect on results since they won't be used as matches
- Adding treated units with outlying obs will alter results because these obs would have poor matches leading to possibly biased estimates
- SEs would be largely unaffected

# Handling Lack of Overlap with PS

- Adding obs with outlying values will lead to PSs close to 0 and 1
- Values close to 0 for the control obs cause little problem because these units would get close to 0 weight in the regression
- Values close to 1 for the control obs would receive high weights leading to increases in the variance of ATE
- Recall

$$\hat{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i Y_i}{\hat{ps}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{ps}(X_i)}\right]$$

- Blocking on the PS leads to similar conclusions

# Handling Lack of Overlap Conclusions

- PS and Matching methods are better designed to cope with limited overlap in the covariate distributions relative to parametric or semi-parametric (series) regression models
- Bottom line: Inspect the histograms of the estimated PS in both groups to assess whether limited overlap is an issue

# References


- *Imbens, Guido W., 2004, Nonparametric estimation of average treatment effects under exogeneity: A Review, *The Review of Economics and Statistics*, 86, 4-29.
- *Todd, Petra E. 2006, Matching Estimators, *Working Paper*, University of Pennsylvania