Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand

Richard K. Crump - UC Berkeley
V. Joseph Hotz - UC Los Angeles
Guido W. Imbens - UC Berkeley
Oscar Mitnik - U Miami

Johns Hopkins University, Symposium on Causality
January 10th, 2006
Problem:

Under unconfoundedness (selection on observables), if overlap in covariates between treated and controls is limited, the population average treatment effect is difficult to estimate.

Questions:

• Are there other average treatment effects of the form $E[Y(1) - Y(0)|.]$ that are easier to estimate?

• What average treatment effects are interesting? Internal validity versus external validity.

• Hypotheses on $E[Y(1) - Y(0)|X]$: Zero? Constant?
Example

Suppose are interested in the average effect of a new treatment.

Experimental data, with both men and women in sample.

women: 50% gets treatment, 50% gets control
men: 0% gets treatment, 100% gets control

Options:
I estimate bounds on average effect (Manski, 1990)
II focus on average effect for women

Now suppose: women as before,
men: 1% gets treatment, 99% gets control

What to do?
Specific Questions:

I Which subpopulation (defined in terms of covariates) leads to the most precisely estimated average treatment effect? (Optimal Subpopulation Average Treatment Effect, OSATE)

II What is the weight function (of covariates) that maximizes the precision for the weighted average treatment effect? (Optimally Weighted Average Treatment Effect, OWATE)

III Explore implications homogeneity of treatment effect:  
A. Estimation under constant treatment effect  
B. Link to partial linear model (Robinson, 1988, Stock, 1989)

IV Testing:  
A. Testing for zero conditional average treatment effect  
B. Testing for constant conditional average treatment effect
**Notation** (Potential Outcome Framework)

\(N\) individuals/firms/units, indexed by \(i=1,\ldots,N\),

\(W_i \in \{0, 1\}\): 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)

\[ \mu_w(x) = \mathbb{E}[Y(w)|X = x] \] (conditional means)

\[ \sigma^2_w(x) = \mathbb{E}[(Y(w) - \mu_w(x))^2|X = x] \] (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)
Standard Estimands (in econometrics)

\[ \tau_P = \mathbb{E}[Y(1) - Y(0)] \]
(Average Treatment Effect)

\[ \tau_T = \mathbb{E}[Y(1) - Y(0)|W = 1] \]
(Average Treatment Effect for the Treated)

New Estimands

\[ \tau_C = \frac{1}{N} \sum_{i=1}^{N} \tau(X_i) \] (Average Conditional Treatment Effect)

\[ \tau_C(A) = \sum_{i|X_i \in A} \frac{\tau(X_i)}{\sum_{i|X_i \in A} 1} \]
(Subpopulation Average Treatment Effect)

\[ \tau_{C,g} = \frac{\sum_{i=1}^{N} g(X_i) \cdot \tau(X_i)}{\sum_{i=1}^{N} g(X_i)} \]
(Weighted Average Treatment Effect)
Assumptions

I. Unconfoundedness
(Selection-on-Observables, Exogeneity)

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

This form due to Rosenbaum and Rubin (1983).

II. Overlap

\[ 0 < \Pr(W = 1 \mid X) < 1. \]

For all \( X \) there are treated and control units.
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

\[ \tau_P = \mathbb{E}[\tau(X)]. \]
Problem: \( \tau_P \) can be difficult to estimate (variance and bias) when there are values \( x \in \mathbb{X} \) with \( e(x) \) close to zero or one.

Previous Solutions: (all focus on \( \tau_T \))

- Dehejia & Wahba (1999): Drop control units \( i \) with \( e(X_i) < \min_{j:W_j=1} e(X_j) \).

- Heckman, Ichimura, Todd (1998): Estimate \( f_w(x) = f(X|W = w) \), \( w = 0, 1 \). Drop unit \( i \) if \( \hat{f}_w(X_i) \leq q_w \).

- Ho, Imai, King, & Stuart (2004): first match all observations and discard those that are not used as match.

- King (2005): construct convex hull around \( X_i \) for treated and discard controls outside this set.
Specific Questions

I How well can we estimate $\tau_P$, $\tau_T$, $\tau_C$, $\tau_C(A)$, and $\tau_C,g$?

II Which $A$ minimizes the variance of $\tau_C(A)$?

III Which $g(\cdot)$ minimizes the variance of $\tau_C,g$?

IV Test zero conditional average treatment effect $H_0$: $\tau(x) = 0$

V Test constant average treatment effect $H_0$: $\tau(x) = c$ for some $c$. 
**Binary Case**  $X \in \{f, m\}$

$N_x$ is sample size for the subsample with $X = x$

$p_x = N_x/N$ be the population share of type $x$.

$\tau_x$ is average treatment effect conditional on the covariate

$$\tau = p_m \cdot \tau_m + p_f \cdot \tau_f.$$  

$N_{xw}$ is number of observations with covariate $X_i = x$ and treatment indicator $W_i = w$.

$e_x = N_{x1}/N_x$ is propensity score for $x = f, m$.

$$\bar{y}_{xw} = \sum_{i=1}^{N} Y_i \cdot 1\{X_i = x, W_i = w\}/N_{xw}$$

Assume that the variance of $Y(w)$ given $X_i = x$ is $\sigma^2$ for all $x$. 
\[ \hat{\tau}_x = \bar{y}_{x1} - \bar{y}_{x0}, \quad V(\hat{\tau}_x) = \frac{\sigma^2}{N \cdot p_x} \cdot \frac{1}{e_x \cdot (1 - e_x)} \]

The estimator for the population average treatment effect is

\[ \hat{\tau} = \hat{p}_m \cdot \hat{\tau}_m + \hat{p}_f \cdot \hat{\tau}_f. \]

with variance relativ to \( \hat{p}_m \cdot \tau_m + \hat{p}_f \cdot \tau_f \)

\[ V(\hat{\tau} - \hat{p}_m \cdot \tau_m - \hat{p}_f \cdot \tau_f) = \frac{\sigma^2}{N} \cdot E \left[ \frac{1}{e_X \cdot (1 - e_X)} \right]. \]

Define \( V = \min(V(\hat{\tau}), V(\hat{\tau}_f), V(\hat{\tau}_m)). \) Then

\[
V = \begin{cases} 
V(\hat{\tau}_f) & \text{if } \frac{1-p_m}{2-p_m} \leq \frac{e_m(1-e_m)}{e_f(1-e_f)} \leq \frac{1-p_m}{2-p_m}, \\
V(\hat{\tau}) & \text{if } \frac{1-p_m}{2-p_m} \leq \frac{e_m(1-e_m)}{e_f(1-e_f)} \leq \frac{1+p_m}{p_m}, \\
V(\hat{\tau}_m) & \text{if } \frac{1+p_m}{p_m} \leq \frac{e_m(1-e_m)}{e_f(1-e_f)}. 
\end{cases}
\]
One can also consider weighted average treatment effects

\[ \tau_\lambda = \lambda \cdot \tau_m + (1 - \lambda) \cdot \tau_f \]

\[ V(\hat{\tau}_\lambda) = \frac{\sigma^2 \lambda^2}{N p_m e_m (1 - e_m)} + \frac{\sigma^2 (1 - \lambda)^2}{N p_f e_f (1 - e_f)}. \]

This variance is minimized at

\[ \lambda^* = \frac{p_m \cdot e_m \cdot (1 - e_m)}{p_f \cdot e_f \cdot (1 - e_f) + p_m \cdot e_m \cdot (1 - e_m)}. \]

\[ V(\tau_{\lambda^*}) = \frac{\sigma^2}{N} \cdot \frac{1}{\mathbb{E}[e_X \cdot (1 - e_X)]}. \]

\[ V(\tau_C)/V(\tau_{\lambda^*}) = \mathbb{E} \left[ \frac{1}{V(e_X)} \right] / \mathbb{E}[V(e_X)]. \]
Efficiency Bounds

\[ V^{\text{eff}}(\tau_P) = \mathbb{E} \left[ \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} + (\tau(X) - \tau)^2 \right] \]

(Hahn, 1998, Robins and Rotznitzky, 1995)

\[ V^{\text{eff}}(\tau_C) = \mathbb{E} \left[ \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \right] \]

\[ V^{\text{eff}}(\tau_C(\mathcal{A})) = \frac{1}{\Pr(X \in \mathcal{A})} \cdot \mathbb{E} \left[ \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \bigg| X \in \mathcal{A} \right] \]

\[ V^{\text{eff}}(\tau_{C,g}) = \frac{1}{\mathbb{E}[g(X)]^2} \cdot \mathbb{E} \left[ g(X)^2 \cdot \left( \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \right) \right] \]
Theorem 1  The Optimal Subpopulation ATE is $\tau_C(A^*)$. If

$$\sup_{x \in X} \frac{\sigma_1^2(x) \cdot (1 - e(x)) + \sigma_0^2(x) \cdot e(x)}{e(x) \cdot (1 - e(x))}$$

$$\leq 2 \cdot \mathbb{E} \left[ \frac{\sigma_1^2(X) \cdot (1 - e(X)) + \sigma_0^2(X) \cdot e(X)}{e(X) \cdot (1 - e(X))} \right],$$

then $A^* = X$. Otherwise:

$$A^* = \left\{ x \in X \middle| \frac{\sigma_1^2(x) \cdot (1 - e(x)) + \sigma_0^2(x) \cdot e(x)}{e(x) \cdot (1 - e(x))} \leq \gamma \right\},$$

$$\gamma = 2 \cdot \mathbb{E} \left[ \frac{\sigma_1^2(X) \cdot (1 - e(X)) + \sigma_0^2(X) \cdot e(X)}{e(X) \cdot (1 - e(X))} \right].$$

$$\frac{\sigma_1^2(X) \cdot (1 - e(X)) + \sigma_0^2(X) \cdot e(X)}{e(X) \cdot (1 - e(X))} < \gamma.$$
Special Case:

Suppose $\sigma_0^2(x) = \sigma_1^2(x) = \sigma^2$ for all $x \in X$.

Then

$$A^* = \left\{ x \in X \left| \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{\gamma}} \leq e(x) \leq \frac{1}{2} + \sqrt{\frac{1}{4} - \frac{1}{\gamma}} \right. \right\},$$

where $\gamma$ is the unique positive solution to

$$\gamma = 2 \cdot \mathbb{E} \left[ \frac{1}{e(X) \cdot (1 - e(X))} \left| \frac{1}{e(X) \cdot (1 - e(X))} < \gamma \right. \right].$$
How much difference does this make?

Suppose for illustration $e(X) \sim \mathcal{B}(c, c)$ (symm Beta dist.)

For difference values of $c$ one can calculate the optimal value for $\gamma$ and the cutoff point $\alpha = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{\gamma}}$

We then calculate the ratio of the variances $V(\tau(A^*)) / V(\tau(X))$.

Also calculate ratio of variances $V(\tau(A_q)) / V(\tau(X))$ for $A_q = \{X \in \mathbb{X} | q \leq e(x) \leq 1 - q\}$ for fixed cutoff points $q = 0.01$, $q = 0.05$, and $q = 0.10$.

We plot the var ratios against the prob $\Pr(0.1 < e(X) < 0.9)$.

Also relative difference in variances, for $q = 0.01, 0.05, 0.10$

$$(V(\tau(A_q)) - V(\tau(A^*))) / V(\tau(X)),$$
Symmetric Beta Distributions indexed by $P(0.1 < X < 0.9)$

**Ratio of Variance for ATE(\(\alpha\)) to Variance for ATE(1)**

- $p^* E[g(e) | A^*] / E[g(e)]$
- $A^* = [0.01, 0.99]$
- $A^* = [0.05, 0.95]$
- $A^* = [0.1, 0.9]$

**Relative Loss for Suboptimal \(\alpha\)**

- $p^* E[g(e) | A^*] / E[g(e)]$
- $\alpha = 0.01$
- $\alpha = 0.05$
- $\alpha = 0.1$
**Theorem 2**

*The Optimally Weighted Average Treatment Effect (OWATE) is* \( \tau_{g^*} \), *where*

\[
g^*(x) = \left( \frac{\sigma_1^2(x)}{e(x)} + \frac{\sigma_0^2(x)}{1 - e(x)} \right)^{-1},
\]

\[
V_{\text{eff}}(\tau_C, g^*) = \left( \mathbb{E} \left[ \left( \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \right)^{-1} \right] \right)^{-1}
\]

*Special case with* \( \sigma_0^2(x) = \sigma_1^2(x) = \sigma^2:*

\[
g^*(x) = e(x) \cdot (1 - e(x)),
\]

\[
V_{\text{eff}}(\tau_C, g^*) = \sigma^2 \cdot \frac{1}{\mathbb{E} [e(X) \cdot (1 - e(X))]}.
\]
Remark 1

\( V^{\text{eff}}(\tau_C) > V^{\text{eff}}(\tau_C, g^*) \) by Jensen’s inequality if \( \sigma_1^2(x)/e(x) + \sigma_0^2(x)/(1 - e(x)) \) varies over \( X \).

Recall:

\[ V^{\text{eff}}(\tau_C) = \mathbb{E}[\sigma_1^2(X)/e(X) + \sigma_0^2(X)/(1 - e(X))] \]

Special case with \( \sigma_0^2(x) = \sigma_1^2(x) = \sigma^2 \):

\[
\frac{V^{\text{eff}}(\tau_C)}{V^{\text{eff}}(\tau_C, g^*)} = \mathbb{E}[e(X) \cdot (1 - e(X))] \cdot \mathbb{E}\left[ \frac{1}{e(X) \cdot (1 - e(X))} \right]
\]
Remark 2: Suppose $\tau(x) = \tau$, then

$$E[Y|X, W] = \mu_0(X) + \tau \cdot W,$$


$$V^{\text{eff}}(\tau) = \left( E \left[ \left( \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \right)^{-1} \right] \right)^{-1}$$

(Robins, Mark and Newey, 1992) which is equal to $V^{\text{eff}}(\tau_{C,g^*})$.

Comments:

I $\tau_{C,g^*}$ is efficient estimator for $\tau$ under assump that $\tau(x) = \tau$.

II $\hat{\tau}_{C,g^*}$ is most precisely estimable average treatment effect under treatment effect heterogeneity.

III Potentially large price to pay for treatment effect heterogeneity if focus is on $E[Y(1) - Y(0)]$. 21
Covariate Balance for Lalonde Data

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>stand. dev.</th>
<th>mean contr.</th>
<th>mean treat.</th>
<th>mean all</th>
<th>Normalized [t-stat]</th>
<th>Dif $a &lt; e(x)$</th>
<th>$a &lt; e(x) &lt; 1 - a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>34.2</td>
<td>10.5</td>
<td>34.9</td>
<td>25.82</td>
<td>-0.86</td>
<td>[-16.0]</td>
<td>-0.18</td>
<td></td>
</tr>
<tr>
<td>educ</td>
<td>12.0</td>
<td>3.1</td>
<td>12.1</td>
<td>10.35</td>
<td>-0.58</td>
<td>[-11.1]</td>
<td>-0.04</td>
<td></td>
</tr>
<tr>
<td>black</td>
<td>0.29</td>
<td>0.45</td>
<td>0.25</td>
<td>0.84</td>
<td>1.30</td>
<td>[21.0]</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>hispanic</td>
<td>0.03</td>
<td>0.18</td>
<td>0.03</td>
<td>0.06</td>
<td>0.15</td>
<td>[1.5]</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>married</td>
<td>0.82</td>
<td>0.38</td>
<td>0.87</td>
<td>0.19</td>
<td>-1.76</td>
<td>[-22.8]</td>
<td>-0.81</td>
<td></td>
</tr>
<tr>
<td>u ’74</td>
<td>0.13</td>
<td>0.34</td>
<td>0.09</td>
<td>0.71</td>
<td>1.85</td>
<td>[18.3]</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>u ’75</td>
<td>0.13</td>
<td>0.34</td>
<td>0.10</td>
<td>0.60</td>
<td>1.46</td>
<td>[13.7]</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>earn ’74</td>
<td>18.2</td>
<td>13.7</td>
<td>19.4</td>
<td>2.10</td>
<td>-1.26</td>
<td>[-38.6]</td>
<td>-0.20</td>
<td></td>
</tr>
<tr>
<td>earn ’75</td>
<td>17.9</td>
<td>13.9</td>
<td>19.1</td>
<td>1.53</td>
<td>-1.26</td>
<td>[-48.6]</td>
<td>-0.14</td>
<td></td>
</tr>
<tr>
<td>l odds ratio</td>
<td>-7.87</td>
<td>4.91</td>
<td>-8.53</td>
<td>1.08</td>
<td>1.96</td>
<td>[53.6]</td>
<td>0.42</td>
<td></td>
</tr>
</tbody>
</table>
Asymptotic Standard Errors for Lalonde Data

<table>
<thead>
<tr>
<th></th>
<th>ATE</th>
<th>ATT</th>
<th>OSATE</th>
<th>OWATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE</td>
<td>636.58</td>
<td>2.58</td>
<td>1.62</td>
<td>1.29</td>
</tr>
<tr>
<td>Ratio to All</td>
<td>1.0000</td>
<td>0.0040</td>
<td>0.0025</td>
<td>0.0020</td>
</tr>
</tbody>
</table>

Subsample Sizes for Lalonde Data: Propensity Score Threshold 0.0660

<table>
<thead>
<tr>
<th></th>
<th>e(x) &lt; a</th>
<th>a ≤ e(x) ≤ 1 − a</th>
<th>1 − a &lt; e(x)</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>controls</td>
<td>2302</td>
<td>183</td>
<td>5</td>
<td>2490</td>
</tr>
<tr>
<td>treated</td>
<td>9</td>
<td>129</td>
<td>47</td>
<td>185</td>
</tr>
<tr>
<td>all</td>
<td>2311</td>
<td>312</td>
<td>52</td>
<td>2675</td>
</tr>
</tbody>
</table>
Testing:
The results concerning the importance of constant treatment effects suggests 3 null hypotheses of interest:

I (constant conditional average treatment effect)
\[ H_0 : \exists \tau_0, \text{ such that } \forall x \in X, \tau(x) = \tau_0. \]

II (zero conditional average treatment effect for all x)
\[ H'_0 : \forall x \in X, \tau(x) = 0. \]

III (OWATE is zero)
\[ H''_0 : \tau_{C,g^*} = 0. \]

Last is easy and can be done using asymptotic normality for \( \tilde{\tau}_{C,g^*} \).
Testing II: \( \tau(x) = \tau \)

Not same as null \( Y(1) - Y(0) = 0 \), or null \( Y(1)|X \sim Y(0)|X \).

\[
T = \frac{1}{N} \sum_{i=1}^{N} \left( \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right)^2
\]

We use series estimation for \( \mu_w(x) \):

\[
\hat{\mu}_w(x) = R_K(x)' \hat{\gamma}_{w,K}
\]

where \( \hat{\gamma}_{w,K} \) are least squares estimators.

Alternative is kernels: H"ardle looks at parametric restrictions between nonparametric regression functions but only gives results for scalar case.
Define:

\[
\hat{\Omega}_{w,K} = \left( R'_{w,K} R_{w,K} / N_w \right)
\]

and

\[
\hat{V}_K \equiv (\hat{\sigma}^2_{0,K} \cdot \hat{\Omega}^{-1}_{0,K} + \hat{\sigma}^2_{1,K} \cdot \hat{\Omega}^{-1}_{1,K}).
\]

Then the test statistic is

\[
T_2 \equiv \frac{N/2}{\sqrt{2K}} \left( (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})' \cdot \hat{V}^{-1}_K \cdot (\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K \right).
\]

Asymptotic distribution \( \mathcal{N}(0,1) \) (use result from Götze (1991) on rate of convergence in multivariate central limit theorem)
## Tests for Zero and Constant Average Treatment Effects

<table>
<thead>
<tr>
<th></th>
<th>Zero CATE (dof)</th>
<th>Const. ATE (dof)</th>
<th>Zero ATE (dof)</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp data</td>
<td>25.9 (10)</td>
<td>19.3 (9)</td>
<td>7.2 (1)</td>
</tr>
<tr>
<td>nonexper data</td>
<td>26.1 (10)</td>
<td>26.4 (9)</td>
<td>1.2 (1)</td>
</tr>
</tbody>
</table>
Conclusion

Even if p-score is strictly between zero and one, there can be areas where the treatment effect cannot be estimated precisely.

Options:

I Choose an optimal subsample to estimate OSATE

II Estimate a weighted average treatment effect (OWATE)

Gains:

Precision gains can be large, depending on var in the p-score.

Remark:

Costs of allowing for heterogeneous treatment effects can be very large.