

# **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  $\mathbb{E}[Y(1) - Y(0)|\cdot]$  that are easier to estimate?
- What average treatment effects are interesting? Internal validity versus external validity.
- Hypotheses on  $\mathbb{E}[Y(1) - Y(0)|X]$ :  
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  $i=1, \dots, N$ ,

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

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

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

$$e(x) = \mathbb{E}[W|X = x] = \Pr(W = 1|X = x) \text{ (propensity score, Rosenbaum and Rubin, 1983)}$$

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x] = \mu_1(x) - \mu_0(x) \text{ (conditional average treatment effect)}$$

## 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

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

$$\tau_C(\mathcal{A}) = \sum_{i|X_i \in \mathcal{A}} \tau(X_i) / \sum_{i|X_i \in \mathcal{A}} \mathbf{1}$$

(Subpopulation Average Treatment Effect)

$$\tau_{C,g} = \sum_{i=1}^N g(X_i) \cdot \tau(X_i) / \sum_{i=1}^N g(X_i)$$

(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

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

For all  $X$  there are treated and control units.

## Identification

$$\begin{aligned}\tau(X) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\ &= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x]\end{aligned}$$

By unconfoundedness this is equal to

$$\begin{aligned}\mathbb{E}[Y(1)|W = 1, X = x] - \mathbb{E}[Y(0)|W = 0, X = x] \\ = \mathbb{E}[Y|W = 1, X = x] - \mathbb{E}[Y|W = 0, X = x].\end{aligned}$$

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  $\tau_P$ ,  $\tau_T$ ,  $\tau_C$ ,  $\tau_C(\mathcal{A})$ , and  $\tau_{C,g}$ ?

**II** Which  $\mathcal{A}$  minimizes the variance of  $\tau_C(\mathcal{A})$ ?

**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 $X$ 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  $x = f, m$ .

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

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  $\hat{p}_m \cdot \tau_m + \hat{p}_f \cdot \tau_f$

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

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

$$V = \begin{cases} V(\hat{\tau}_f) & \text{if } \frac{e_m(1-e_m)}{e_f(1-e_f)} \leq \frac{1-p_m}{2-p_m}, \\ V(\hat{\tau}) & \text{if } \frac{1-p_m}{2-p_m} \leq \frac{e_m(1-e_m)}{e_f(1-e_f)} \leq \frac{1+p_m}{p_m}, \\ V(\hat{\tau}_m) & \text{if } \frac{1+p_m}{p_m} \leq \frac{e_m(1-e_m)}{e_f(1-e_f)}. \end{cases}$$

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

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

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

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

## 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)

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

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

$$V^{\text{eff}}(\tau_{C,g}) = \frac{1}{\mathbb{E}[g(X)]^2} \cdot \mathbb{E} \left[ g(X)^2 \cdot \left( \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \right) \right]$$

**Theorem 1**     *The Optimal Subpopulation ATE is  $\tau_C(\mathcal{A}^*)$ . If*

$$\begin{aligned} & \sup_{x \in \mathbb{X}} \frac{\sigma_1^2(x) \cdot (1 - e(x)) + \sigma_0^2(x) \cdot e(x)}{e(x) \cdot (1 - e(x))} \\ & \leq 2 \cdot \mathbb{E} \left[ \frac{\sigma_1^2(X) \cdot (1 - e(X)) + \sigma_0^2(X) \cdot e(X)}{e(X) \cdot (1 - e(X))} \right], \end{aligned}$$

*then  $\mathcal{A}^* = \mathbb{X}$ . Otherwise:*

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

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

## Special Case:

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

Then

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

where  $\gamma$  is the unique positive solution to

$$\gamma = 2 \cdot \mathbb{E} \left[ \frac{1}{e(X) \cdot (1 - e(X))} \mid \frac{1}{e(X) \cdot (1 - e(X))} < \gamma \right].$$

## 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  $V(\tau(\mathcal{A}^*)) / V(\tau(\mathbb{X}))$ .

Also calculate ratio of variances  $V(\tau(\mathcal{A}_q)) / V(\tau(\mathbb{X}))$  for

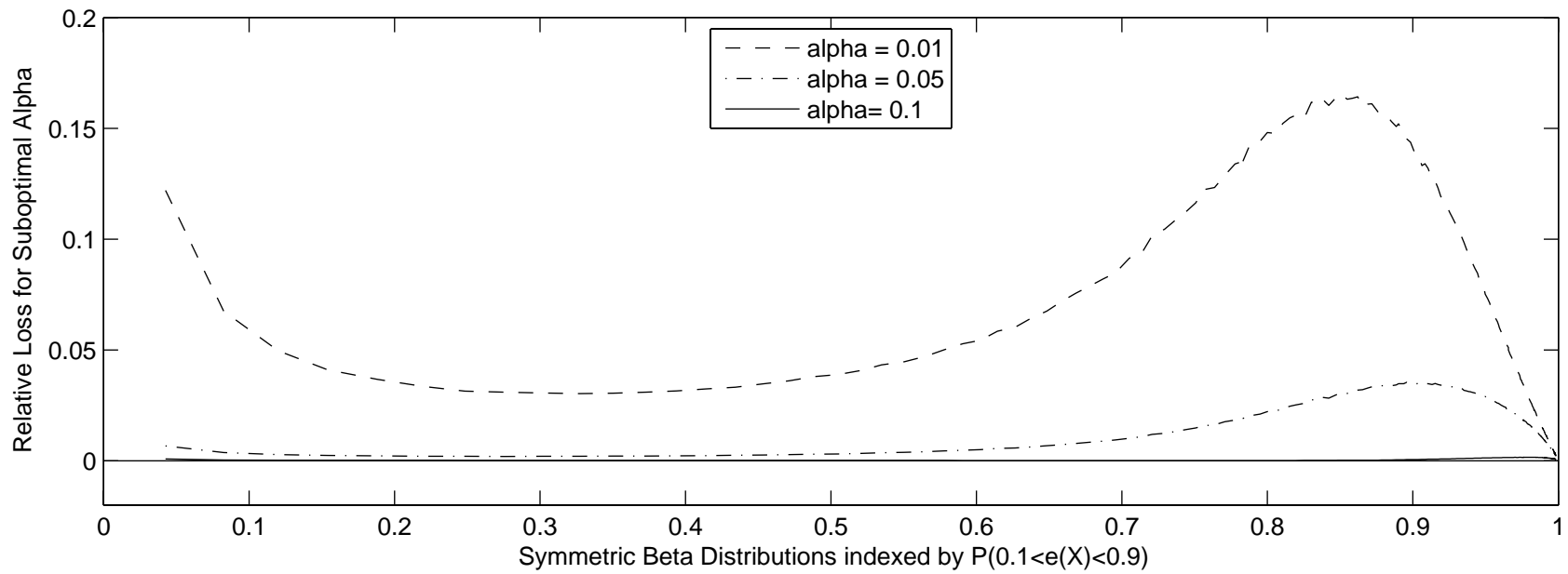
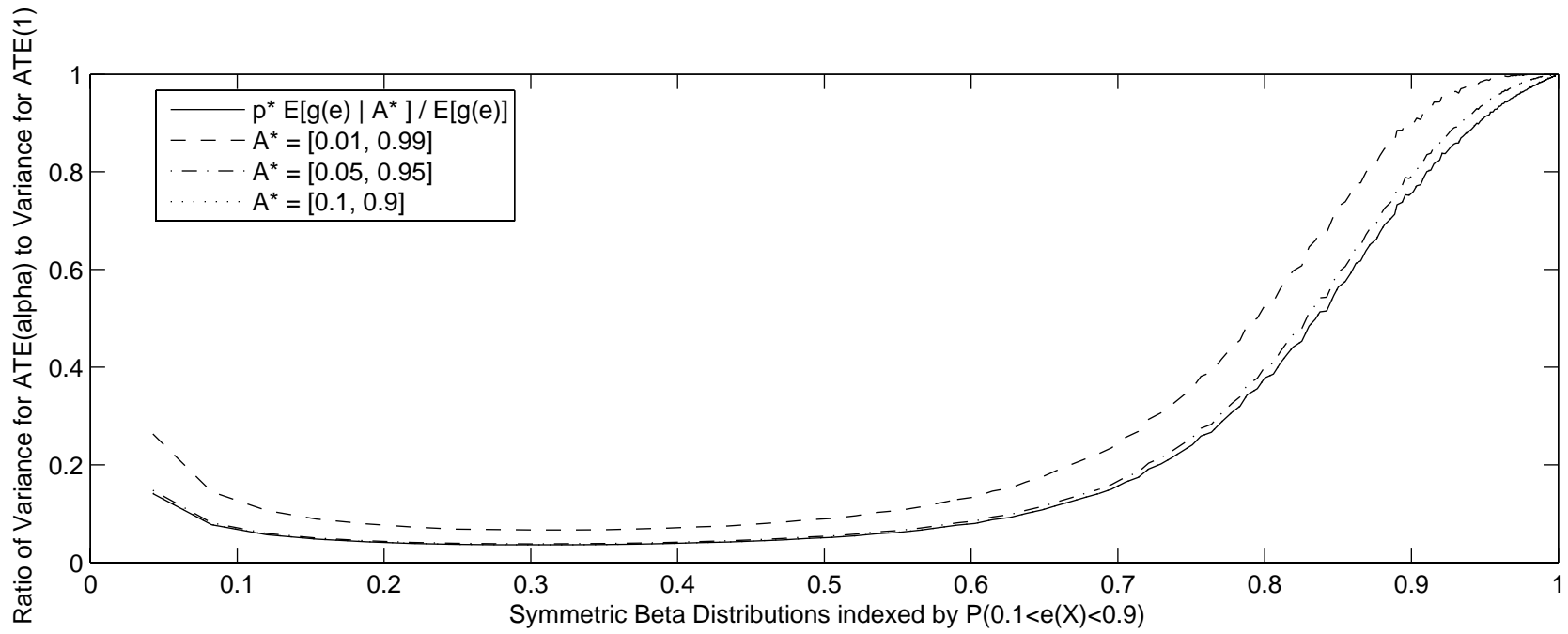
$$\mathcal{A}_q = \{X \in \mathbb{X} | q \leq e(x) \leq 1 - q\}$$

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  $q = 0.01, 0.05, 0.10$

$$(V(\tau(\mathcal{A}_q)) - V(\tau(\mathcal{A}^*))) / V(\tau(\mathbb{X})),$$



## 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  $\mathbb{X}$ .

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

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

Partial linear model (Robinson, 1988, Stock, 1989).

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

(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  $\mathbb{E}[Y(1) - Y(0)]$ .

## Covariate Balance for Lalonde Data

---

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

---

## Asymptotic Standard Errors for Lalonde Data

---

	ATE	ATT	OSATE	OWATE
ASE	636.58	2.58	1.62	1.29
Ratio to All	1.0000	0.0040	0.0025	0.0020

---

## Subsample Sizes for Lalonde Data: Propensity Score Threshold 0.0660

---

	$e(x) < a$	$a \leq e(x) \leq 1 - a$	$1 - a < e(x)$	all
controls	2302	183	5	2490
treated	9	129	47	185
all	2311	312	52	2675

---

## Testing:

The results concerning the importance of constant treatment effects suggests 3 null hypotheses of interest:

**I** (constant conditional average treatment effect)

$$H_0 : \exists \tau_0, \text{ such that } \forall x \in \mathbb{X}, \tau(x) = \tau_0.$$

**II** (zero conditional average treatment effect for all  $x$ )

$$H'_0 : \forall x \in \mathbb{X}, \tau(x) = 0.$$

**III** (OWATE is zero)

$$H''_0 : \tau_{C,g^*} = 0.$$

Last is easy and can be done using asymptotic normality for  $\hat{\tau}_{C,g^*}$ .

**Testing II:**  $\tau(x) = \tau$

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

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

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: Härdle 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

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

Then the test statistic is

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

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

---

---

	Zero CATE	Const. ATE	Zero ATE
	chi-sq (dof)	chi-sq (dof)	chi-sq (dof)
exp data	25.9 (10)	19.3 (9)	7.2 (1)
nonexper data	26.1 (10)	26.4 (9)	1.2 (1)

---

## **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.