The Evaluation of Social Programs: Some Practical Advice

Guido W. Imbens - Harvard University

2nd IZA/IFAU Conference on Labor Market Policy Evaluation

Bonn, October 11th, 2008

Background Reading

Guido Imbens and Jeffrey Wooldridge (2008), "Recent Developments in the Econometrics of Program Evaluation," under revision for *Journal of Economic Literature*

Guido Imbens and Donald Rubin, (2009), *Statistical Methods for Causal Inference in Biomedical and Social Sciences* forthcoming, Cambridge University Press.

These slides and Imbens-Wooldridge paper are available on Imbens home page at Harvard University, Department of Economics.

http://www.economics.harvard.edu/faculty/imbens

General Motivation

In seventies empirical work on evaluation of social programs (Ashenfelter & Card).

Since then much empirical and methodological work, in both statistics and econometrics literature.

Here: my views on current state of the art, with emphasis on recommendations for empirical researchers in this area.

Based on work with various coauthors, Abadie, Angrist, Athey, Chamberlain, Crump, Hirano, Hotz, Manski, Mitnik, Newey, Ridder, Rubin, Rosenbaum, Wooldridge (without implicating them)

Many Applications in Social Sciences

Labor market programs (main focus in economics):
Ashenfelter (1978), Ashenfelter and Card (1985), Lalonde (1986), Card and Sullivan (1989), Heckman and Hotz (1989), Friedlander and Robins (1995), Dehejia and Wahba (1999), Lechner (1999), Heckman, Ichimura and Todd (1998).

Effect of Military service on Earnings:
Angrist (1998)

Effect of Family Composition:
Manski, McLanahan, Powers, Sandefur (1992)

Effect of Regulations:
Minimum wage laws (Card and Krueger, 1995), Tax reforms (Eissa and Liebman, 1998), Information Disclosure (restaurant hygiene, Leslie and Jin, 2004)

Literature

1. Large theoretical literature:

   (a) efficiency bounds: Hahn, 1998.

   (b) estimators: Hahn, 1998; Heckman, Ichimura and Todd, 1998; Hirano, Imbens and Ridder, 2003, Abadie & Imbens, 2006; Robin and Rotnitzky, 1995.

2. Few concrete recommendations for finite samples

   (a) Simulations: Frölich, 2004, Zhao, 2004.

   (b) Covariate selection: Caliendo, 2006; De Luna, Richardson, Waernbaum, 2008.

Outline for Talk

1. Introduction & Notation

2. Two Illustr: Lottery Data, Lalonde (Dehejia-Wahba) Data

3. Key Assumptions: Unconfoundedness and Overlap

4. Prelim: Specifying and Estimating the Propensity Score

5. Design Stage: Creating a Sample with Overlap

6. Estimation, Variance Estimation & and the Bootstrap

7. Assessing Unconfoundedness

1.A Introduction

Steps for Estimating Ave Treatm Effects Given Unconf

1. Estimate the propensity score

2. Assess Overlap: If necessary create balanced sample

3. Re-estimated the propensity score in the balanced sample.

4. Estimate ate by blocking/matching on cov, with regression

5. Estimate conditional variance

6. If possible, assess unconf assump using pseudo outcomes

1.B Rubin Causal Model Notation

N individuals/firms/units, indexed by i=1,. . . ,N,

$W_i \in \{c, t\}$: Binary treatment,

$Y_i(t)$: Potential outcome for unit i given active treatment,

$Y_i(c)$: Potential outcome for unit i given control treatment,

X_i: $k \times 1$ vector of covariates.

We observe $\{(X_i, W_i, Y_i)\}_{i=1}^N$, where

$$Y_i = \begin{cases} Y_i(c) & \text{if } W_i = c, \\ Y_i(t) & \text{if } W_i = t. \end{cases}$$

Fundamental problem of Causal Inference (Holland, 1986): we never observe $Y_i(c)$ and $Y_i(t)$ for the same individual i, so unit-level causal effect $\tau_i = Y_i(t) - Y_i(c)$ is never observed.

Notation (ctd)

$\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x]$ (conditional means)

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

$e(x) = \text{pr}(W_i = t|X_i = x)$ (propensity score, Rosenbaum and Rubin, 1983)

$\tau(x) = \mathbb{E}[Y_i(t) - Y_i(c)|X_i = x] = \mu_t(x) - \mu_c(x)$ (conditional average treatment effect)

$\tau_{\text{SP}} = \mathbb{E}[Y_i(t)-Y_i(c)]$ Super Population Average Treatment Effect

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

2.A Illustr 1: The Imbens-Rubin-Sacerdote Lottery Data

IRS (American Economic Review, 2000) were interested in estimating the effect of unearned income on behavior including labor supply, consumption and savings.

(also on children's outcomes and happiness, but too few observations for answerin first question, and wrong survey questions for answering second)

They surveyed individiuals who had played and won large sums of money in the lottery in Massachusetts (400K on average). As a comparison group they collected data on a second set of individuals who also played the lottery but who had not won big prizes (the "losers")

Summary Statistics Lottery Data

Covariates	All		Losers	Winners		
	(N=496)		(N=259)	(N=237)		**Norm.**
	mean	(s.d.)	mean	mean	[t-stat]	**Dif.**
Year Won	6.23	(1.18)	6.38	6.06	[-3.0]	0.19
# Tickets	3.33	(2.86)	2.19	4.57	[9.9]	0.64
Age	50.2	(13.7)	53.2	47.0	[-5.2]	0.33
Male	0.63	(0.48)	0.67	0.58	[-2.1]	0.13
Education	13.7	(2.2)	14.4	13.0	[-7.8]	0.50
Working Then	0.78	(0.41)	0.77	0.80	[0.9]	0.06
Earn Y-6	13.8	(13.4)	15.6	12.0	[-3.1]	0.19
⋮						
Earn Y-1	16.3	(15.7)	18.0	14.5	[-2.5]	0.16
Pos Earn Y-6	0.69	(0.46)	0.69	0.70	[0.3]	0.02
⋮						
Pos Earn Y-1	0.71	(0.45)	0.69	0.74	[1.2]	0.07

Normalized difference is

$$\text{norm. dif.} = \left| \frac{\overline{X}_t - \overline{X}_c}{\sqrt{S_t^2 + S_c^2}} \right|$$

If the normalized difference is greater than 0.2, linear regression based estimates for average treatment effect are likely to be sensitive to specification.

Recall: OLS estimator for average control outcomes for treated units has the form

$$\widehat{\mathbb{E}}[Y_i(c)|W_i = t] = \overline{Y}_c^{\text{obs}} + \widehat{\beta}'_{\text{ols},c} \left(\overline{X}_t - \overline{X}_c \right)$$

so if $\overline{X}_t - \overline{X}_c$ is far from zero, the estimator relies heavily on extrapolation.

Summary Statistics Lalonde/Dehejia/Wahba Data (CPS Comparison Group)

Covariates	All (N=16,177) mean	(s.d.)	Contr (15,992) mean	Trainees (N=185) mean	[t-stat]	**Norm. Dif.**
black	0.08	(0.27)	0.07	0.84	[28.6]	1.72
hispanic	0.07	(0.26)	0.07	0.06	[-0.7]	0.04
age	33.1	(11.0)	33.2	25.8	[-13.9]	0.56
married	0.71	(0.46)	0.71	0.19	[-18.0]	0.87
nodegree	0.30	(0.46)	0.30	0.71	[12.2]	0.64
education	12.0	(2.9)	12.0	10.4	[-11.2]	0.48
earn '74	13.9	(9.6)	14.0	2.1	[-32.5]	1.11
unempl '74	0.13	(0.33)	0.12	0.71	[17.5]	1.05
earn '75	13.5	(9.3)	13.7	1.5	[-48.9]	1.23
unempl '75	0.11	(0.32)	0.11	0.60	[13.6]	0.84

3. Key Assumptions

3.A Unconfoundedness

$$Y_i(c), Y_i(t) \perp\!\!\!\perp W_i \;\Big|\; X_i.$$

This form due to Rosenbaum and Rubin (1983). Like selection on observables, or exogeneity. Suppose

$$Y_i(c) = \alpha + \beta' X_i + \varepsilon_i, \qquad Y_i(t) = Y_i(c) + \tau,$$

then

$$Y_i = \alpha + \tau \cdot 1_{W_i = t} + \beta' X_i + \varepsilon_i,$$

and unconfoundedness $\Longleftrightarrow \varepsilon_i \perp\!\!\!\perp W_i | X_i$.

3.B Overlap

$$0 < \text{pr}(W_i = t | X_i = x) < 1$$

For all x in the support of X_i there are treated and control units.

In practice we typically use stronger version

$$c < \text{pr}(W_i = t | X_i = x) < 1 - c$$

for some positive c.

3.C Motivation for Unconfoundedness Assumption

I. Descriptive statistics. After simple difference in mean outcomes for treated and controls, it may be useful to compare average outcomes adjusted for covariates.

II. Dropping Unconfoundedness Assumption loses identification, and leads to bounds (e.g., Manski, 1990)

III. Unconfoundedness is consistent some optimizing behavior.

3.D Identification

$$\tau(x) = \mathbb{E}[Y_i(t) - Y_i(c)|X_i = x]$$

$$= \mathbb{E}[Y_i(t)|X_i = x] - \mathbb{E}[Y_i(c)|X_i = x]$$

By unconfoundedness this is equal to

$$\mathbb{E}[Y_i(t)|W_i = t, X_i = x] - \mathbb{E}[Y_i(c)|W_i = c, X_i = x]$$

$$= \mathbb{E}[Y_i|W_i = t, X_i = x] - \mathbb{E}[Y_i|W_i = c, X_i = x]$$

By the overlap assumption we can estimate both terms on the righthand side.

Then

$$\tau_{\mathsf{SP}} = \mathbb{E}[\tau(X_i)] = \mathbb{E}[Y_i(t) - Y_i(c)]$$

4. Preliminaries: Estimating the Propensity Score

For a number of analyses we need an estimate of the propensity score.

General problem: given $(X_i, W_i)_{i=1}^N$, estimate

$$e(x) = \text{pr}(W_i = t | X_i = x)$$

Here: use logit model

$$e(x) = \frac{\exp(h(x)'\gamma)}{1 + \exp(h(x)'\gamma)}$$

for some selected vector of functions $h(x)$ (no loss of generality yet).

Given specification estimate γ by maximum likelihood.

4.A Selecting Covariates

Implementation: $h(x)$ contains linear and second order terms.

Step 1: Select some covariates to be included linearly (*a priori* important covariates, e.g., lagged outcomes). Could be none.

Step 2: Select additional linear terms sequentially, adding those with the largest value for likelihood ratio test for zero coefficient, till the largest likelihood ratio test statistic is less than $C_{\text{lin}} = 1$.

Step 3: Select second order terms sequentially (based on included linear terms), adding those with the largest value for likelihood ratio test for zero coefficient, till the largest likelihood ratio test statistic is less than $C_{\text{qua}} = 2.72$.

4.B Illustration 2: The Lottery Data

Covariates in Propensity Score: 18 covariates

1. 4 Pre-selected: `# Tickets`, `Education`, `Working Then`, `Earn Y-1`

2. 4 Additional Linear Terms Selected: `Age`, `Year Won`, `Pos Earn Y -5`, `Male`

3. 10 Second Order Terms Selected: `Year Won`×`Year Won`, `Earn Y-1`×`Male`, `# Tickets`×`# Tickets`, `# Tickets`×`Working Then`, `Education`×`Education`, `Education`×`Earn Y-1`, `# Tickets`×`Education`, `Earn Y-1`×`Age`, `Age`×`Age`, `Year Won`×`Male`

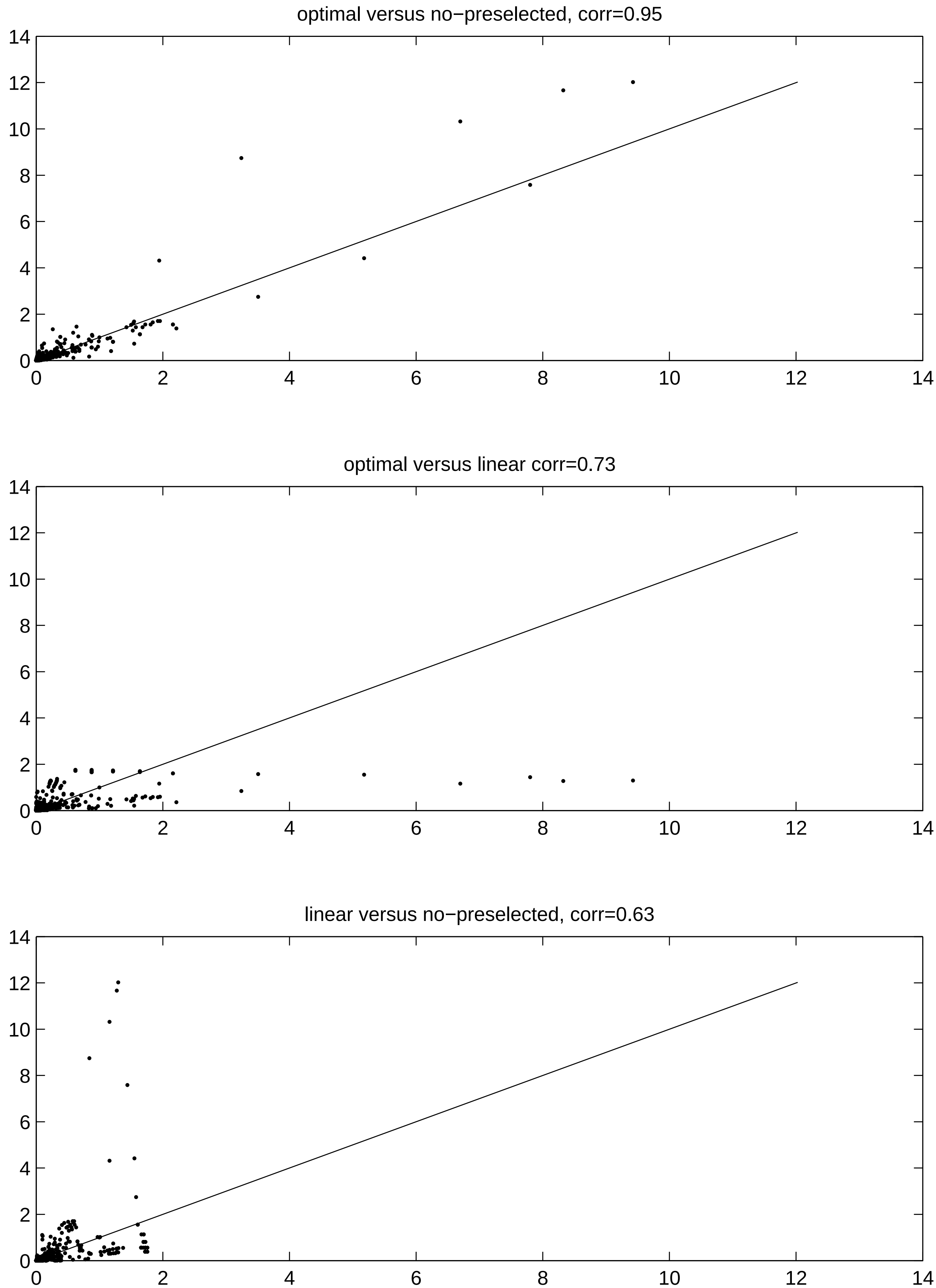
4.C Illustration 2: The Lalonde Data

Covariates in Propensity Score: 10 covariates

1. 4 Pre-selected: `earn'74`, `unempl'74`, `earn'75`, `unempl'75`

2. 5 Additional Linear Terms Selected: `black`, `married`, `nodegree`, `hispanic`, `age` (only `educ` is not selected)

3. 5 Second Order Terms Selected: `age`×`age`, `unempl'74`×`unempl'75`, `earn'74`×`age`, `earn'75`×`married`, `unempl'74`×`earn'75`


optimal versus no−preselected, corr=0.95
0
2
4
6
8
10
12
14
optimal versus linear corr=0.73
linear versus no−preselected, corr=0.63

4.D Assessing Adequacy of Specification

Given estimated propensity score, construct subclasses. Then for each covariate test whether the average value of the covariate in each subclass is the same for treated and controls.

Compare the K p-values (one for each covariate) to a uniform distribution.

Construction of subclasses

Start with single subclass. Calculate the t-statistic for the null-hypothesis that the average value of the propensity score for treated in the subclass is the same as the average value for the controls in the same subclass.

If the subclass were to be split, it would be split by the median value of the estimated propensity score. Count how many treated and control units would be in each new subclass, $N_{c,low}$, $N_{t,low}$, $N_{c,high}$, and $N_{t,high}$.

Split the block if:
1. $|t| \geq \text{tmax} = 1$ and
2. $\min(N_{c,low}, N_{t,low}, N_{c,high}, N_{t,high}) \geq nbmin1 = 3$, and
3. $\min(N_{c,low} + N_{t,low}, N_{c,high} + N_{t,high}) \geq nbmin2 = dim(X_i) + 2$

Fig 1: lottery balance in covariates (C_L=1,C_Q=2.78)

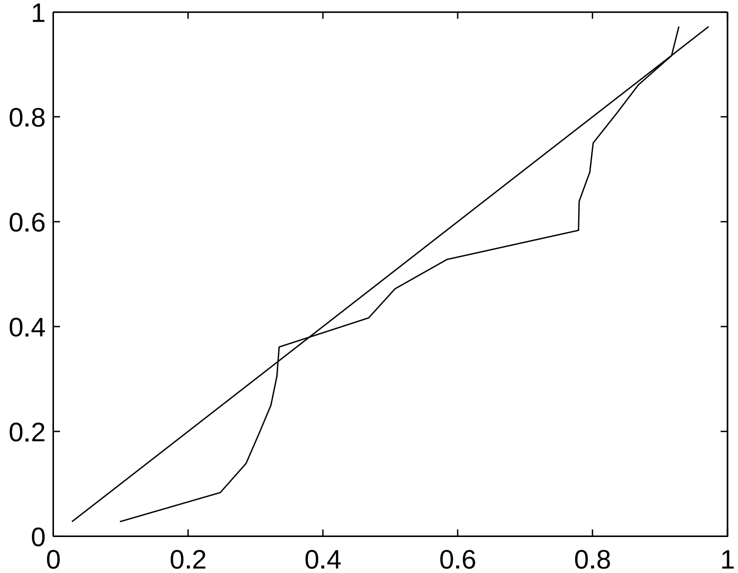


Fig 3: lalonde balance in covariates (C_L=1,C_Q=2.78)

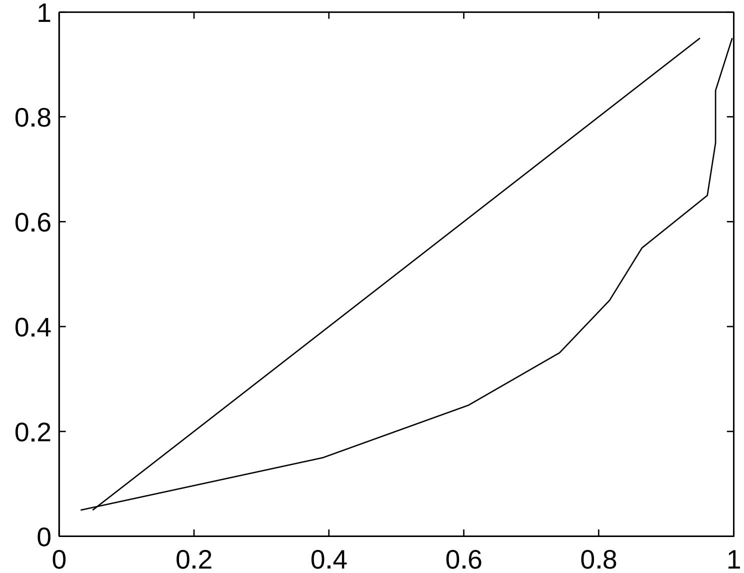


Fig 2: lottery balance in covariates (C_L=0,C_Q=infty)

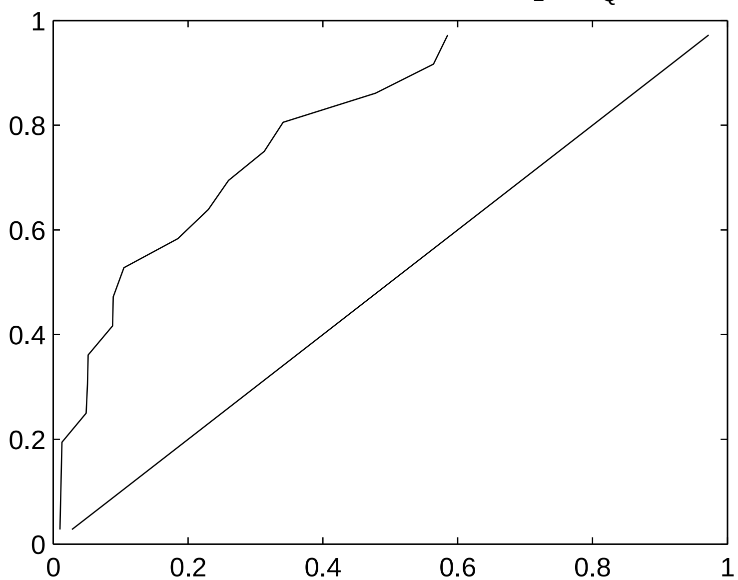
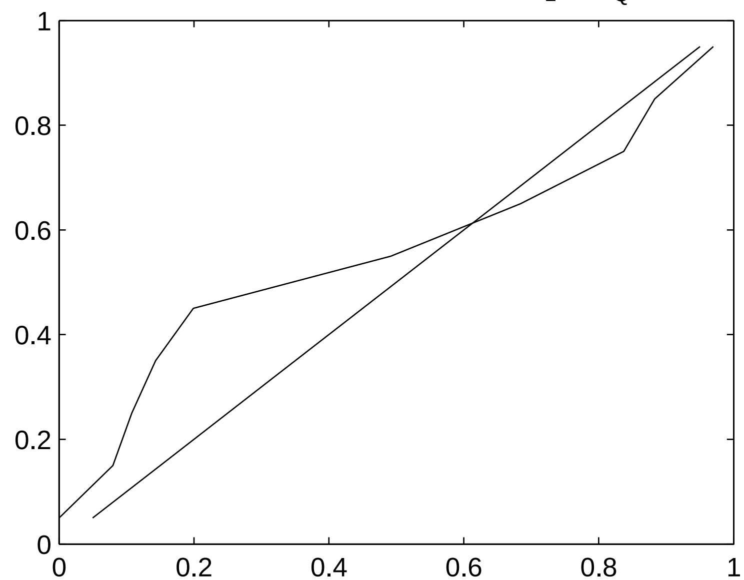


Fig 4: lalonde balance in covariates (C_L=0,C_Q=infty)



Note that this specification of the propensity score leads to better covariate balance than simply including all covariates linearly.

5. Design: What Do We Want to Learn / What Can We Learn?

Population average effect

$$\tau_{\text{SP}} = \mathbb{E}[Y_i(t) - Y_i(c)], \qquad \tau_{\text{SP},T} = \mathbb{E}[Y_i(t) - Y_i(c)|W_i = t]$$

may be difficult to estimate if overlap is limited. Instead we focus on "easier" estimands: conditional average treatment effect:

$$\tau_{\text{cond}} = \frac{1}{N}\sum_{i=1}^{N} \tau(X_i), \qquad \tau_{\text{cond},T} = \frac{1}{N_t}\sum_{i:W_i=t} \tau(X_i)$$

and, for subsets $\mathbb{A}$ of the covariate space $\mathbb{X}$,

$$\tau_{\text{cond}}(\mathbb{A}) = \frac{1}{\#(\mathbb{A})}\sum_{i:X_i\in\mathbb{A}} \tau(X_i)$$

5.A Preliminary Result: Efficiency Bounds

Efficiency bound for τ (Hahn, 1998). For $\hat{\tau}$ with

$$\sqrt{N}(\hat{\tau} - \tau_{\text{SP}}) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}),$$

$$\mathbb{V} \geq \mathbb{E}\left[\frac{\sigma_t^2(X_i)}{e(X_i)} + \frac{\sigma_c^2(X_i)}{1 - e(X_i)} + (\tau(X) - \tau)^2\right].$$

Bound for $\tau_{\text{cond}}(\mathbb{A})$ can be much lower:

$$\sqrt{N}(\hat{\tau} - \tau_{\text{cond}}(\mathbb{A})) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}),$$

$$\mathbb{V} \geq \frac{1}{\text{pr}(X_i \in \mathbb{A})} \mathbb{E}\left[\frac{\sigma_t^2(X_i)}{e(X_i)} + \frac{\sigma_c^2(X_i)}{1 - e(X_i)} \middle| X_i \in \mathbb{A}\right].$$

5.B Matching Without Replacement on the Propensity Score to Create Balanced Sample

- Appropriate when interest in average effect for treated, and large pool of potential controls.

Step 1. Rank treated by decreasing value of estimated propensity score (those with highest propensity score are most difficult to find good matches for)

Step 2. For each treated find closest control in terms of estimated propensity score, without replacement.

(not optimal, but jointly matching is considerably more difficult and little gain)

5.C Summary Statistics Lalonde/Dehejia/Wahba Data (CPS Comparison Group), Matched Sample

Covariates	All (N=370) mean	(s.d.)	Comp. (N=185) mean	Trainees (N=185) mean	**Full Norm. Dif.**	**Matched Norm. Dif.**
black	0.85	(0.36)	0.85	0.84	1.72	0.02
hispanic	0.06	(0.24)	0.06	0.06	0.04	0.02
age	25.9	(7.4)	25.9	25.8	0.56	0.01
married	0.22	(0.41)	0.25	0.19	0.87	0.10
nodegree	0.64	(0.48)	0.57	0.71	0.64	0.20
education	10.6	(2.5)	10.9	10.4	0.48	0.16
earn '74	2.45	(5.27)	2.81	2.10	1.11	0.10
unempl '74	0.69	(0.46)	0.66	0.71	1.05	0.07
earn '75	1.68	(3.52)	1.82	1.53	1.23	0.06
unempl '75	0.55	(0.50)	0.50	0.60	0.84	0.14

Re-estimating the Propensity Score

Still substantial differences in covariates by treatment group: e.g., `earn '74` differs by $700. Need for further adjustment, partly based on propensity score.

Covariates in Propensity Score in Matched Sample:

1. 4 Pre-selected: `earn'74`, `unempl'74`, `earn'75`, `unempl'75`

2. 2 Additional Linear Terms Selected: `nodegree`, `married`

3. 2 Second Order Terms Selected: `unempl'75`×`married`, `earn'74`×`earn'75`

5.C Crump, Hotz, Imbens and Mitnik (2007): Trim units that create problems for estimation.

Choose $\mathbb{A} = \mathbb{A}^*$ to minimize variance bound for $\tau_{\text{cond}}(\mathbb{A})$:

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

where γ solves

$$\gamma = 2 \cdot \mathbb{E}\left[\frac{\sigma_t^2(X_i) \cdot (1 - e(X_i)) + \sigma_c^2(X_i) \cdot e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \right|$$

$$\left. \frac{\sigma_t^2(X_i) \cdot (1 - e(X_i)) + \sigma_c^2(X_i) \cdot e(X_i)}{e(X_i) \cdot (1 - e(X_i))} < \gamma \right]$$

Optimal Subset Under Homoskedasticity: (relevant case in practice, because $\sigma_w^2(x)$ is difficult to estimate)

Suppose $\sigma_c^2(x) = \sigma_t^2(x) = \sigma^2$ for all x.

Then

$$\mathbb{A}^* = \mathbb{A}_\alpha = \{x \in \mathbb{X} \,|\alpha \leq e(x) \leq 1 - \alpha\}\,,$$

where α is the solution to

$$\frac{1}{\alpha \cdot (1-\alpha)} = 2 \cdot \mathbb{E}\left[\frac{1}{e(X_i) \cdot (1 - e(X_i))}\right| \alpha < e(X_i) < 1 - \alpha\bigg].$$

Simple, Good Approximation to Optimal Set:

$$\mathbb{A}_{0.10} = \{x \in \mathbb{X} \,|0.10 \leq e(x) \leq 0.90\}\,,$$

5.E Trimming for Lottery Data

Sample Sizes for Selected Subsamples with the Propensity Score between α and $1-\alpha$ (Estimated α =0.0891)

	low	middle	high	All
	$e(x) < \alpha$	$\alpha \leq e(X) \leq 1-\alpha$	$1-\alpha < e(X)$	
Losers	82	172	5	259
Winners	4	151	82	237
All	86	323	87	496

Summary Statistics Lottery Data, Trimmed Sample

Covariates	All (N=323) mean	All (N=323) (s.d.)	Losers (N=172) mean	Winners (N=151) mean	**Full Norm. [t-stat]**	**Trimm. Norm. Dif.**
Year Won	6.36	1.15	6.40	6.32	0.19	0.04
# Tickets	2.99	2.52	2.40	3.67	0.64	0.36
Age	51.0	13.3	51.5	50.4	0.33	0.06
Male	0.63	0.48	0.65	0.60	0.13	0.08
Education	13.6	2.1	14.0	13.0	0.50	0.33
Working Then	0.80	0.40	0.79	0.80	0.06	0.02
Earn Y-6	14.3	13.3	15.5	13.0	0.19	0.13
⋮						
Earn Y-1	17.0	15.7	18.4	15.5	0.16	0.13
Pos Earn Y-6	0.71	0.45	0.71	0.71	0.02	0.00
⋮						
Pos Earn Y-1	0.71	0.45	0.72	0.71	0.07	0.01

Re-estimating the Propensity Score in the Trimmed Lottery Sample:

1. 4 Pre-selected: `# Tickets`, `Education`, `Working Then`, `Earn Y-1`

2. 4 Additional Linear Terms Selected: `Age`, `Pos Earn Y -5`, `Year Won`, `Earn Y -5`

3. 4 Second Order Terms Selected: `Year Won`×`Year Won`, `# Tickets`×`Year Won`, `# Tickets`×`# Tickets`, `Working Then`×`Year Won`

6.A Estimation of and Inference for Average Treatment Effect under Unconfoundedness

I. Regression estimators: estimate $\mu_w(x)$, and average $\hat{\mu}_t(X_i - \hat{\mu}_c(X_i)$. **Unreliable when covariate distributions far apart.**

II. Propensity score estimators: estimate $e(x)$, and block on $\hat{e}(x)$, or do inverse probability weighting. (Horvitz and Thompson, 1951; Rosenbaum and Rubin, 1984; Robins and Rotznitzky, 1995; Hirano, Imbens and Ridder, 2003) **sensitive to specification of propensity score**

III. Matching: match all units to units with similar values for covariates and opposite treatment, and average within pair difs (Abadie and Imbens, 2006) **considerable bias remains**

IV. Combining Regression with Propensity score blocking or Matching Methods. **Recommended**

6.B Blocking and Regression

1. Estimate propensity score $\hat{e}(x)$, using logistic regression, select main effects and interactions/quadratic terms.

2. Construct blocks so that within blocks the propensity score does not vary too much, and number of treated and control observations is not too small. (tuning parameters: tmax = 1.96, $nbmin1 = 3$, $nbmin2 = dim(X_i) + 2$.)

3. Within blocks estimate average treatment effect using regression on all covariates.

4. Average within-block estimates of average effect.

6.C Lalonde/Dehejia/Wahba Data: Subclassification

Optimal Subclassification with LDW Data in Matched Sample

	Pscore		# of Obs		Ave Pscore		
Subclass	Min	Max	Contr	Treat	Contr	Treat	t-stat
1	0.00	0.39	53	22	0.32	0.34	0.7
2	0.39	0.50	40	35	0.44	0.44	0.7
3	0.50	1.00	92	128	0.57	0.58	1.1
all	0.00	1.00	185	185	0.47	0.53	4.6

6.C Lalonde/Dehejia/Wahba Data: Normalized Differences After Subclassification in Matched Sample

Covariates	Comp. (N=185) mean	Trainees (N=185) mean	**Full Norm. Dif.**	**Matched Norm. Dif.**	**Blocked Norm. Dif.**
black	0.85	0.84	1.72	0.02	0.02
hispanic	0.06	0.06	0.04	0.02	0.03
age	25.9	25.8	0.56	0.01	0.01
married	0.25	0.19	0.87	0.10	0.04
nodegree	0.57	0.71	0.64	0.20	0.06
education	10.9	10.4	0.48	0.16	0.08
earn '74	2.81	2.10	1.11	0.10	0.05
unempl '74	0.66	0.71	1.05	0.07	0.00
earn '75	1.82	1.53	1.23	0.06	0.02
unempl '75	0.50	0.60	0.84	0.14	0.01

6.D Lottery Data: Optimal Subclassification

	Pscore		# of Obs		Ave Pscore		
Subclass	Min	Max	Contr	Treat	Contr	Treat	t-stat
1	0.00	0.24	67	13	0.16	0.16	-0.14
2	0.24	0.32	32	8	0.28	0.28	0.92
3	0.32	0.44	24	17	0.37	0.39	1.66
4	0.44	0.69	34	47	0.54	0.57	1.95
5	0.69	1.00	15	66	0.80	0.83	1.58
All	0.00	1.00	172	151	0.34	0.61	10.9

6.D Lottery Data: Normalized Differences After Subclassification in Trimmed Sample

Covariates	Comp. (N=172) mean	Trainees (N=151) mean	**Full Norm. Dif.**	**Trimmed Norm. Dif.**	**Blocked Norm. Dif.**
Year Won	6.40	6.32	0.19	0.04	0.04
# Tickets	2.40	3.67	0.64	0.36	0.06
Age	51.48	50.44	0.33	0.06	0.03
Male	0.65	0.60	0.13	0.08	0.10
Education	14.01	13.03	0.50	0.33	0.08
Working Then	0.79	0.80	0.06	0.02	0.03
Earn Y-6	15.49	13.02	0.19	0.13	0.02
⋮					
Earn Y-1	18.38	15.45	0.16	0.13	0.01
Pos Earn Y-6	0.71	0.71	0.02	0.00	0.05
⋮					
Pos Earn Y-1	0.72	0.71	0.07	0.01	0.05

6.C Matching (see Abadie and Imbens, 2006)

For each treated unit i, find untreated unit $\ell(i)$ with

$$\|X_{\ell(i)} - x\| = \min_{\{l:W_l=c\}} \|X_l - x\|,$$

and the same for all untreated observations. Define:

$$\widehat{Y}_i(t) = \begin{cases} Y_i & \text{if } W_i = t, \\ Y_{\ell(i)} & \text{if } W_i = c, \end{cases} \qquad \widehat{Y}_i(c) = \begin{cases} Y_i & \text{if } W_i = c, \\ Y_{\ell(i)} & \text{if } W_i = t. \end{cases}$$

Then the simple matching estimator is:

$$\widehat{\tau}^{sm} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}_i(t) - \widehat{Y}_i(c) \right).$$

Note: since we match all units it is crucial that matching is done with replacement.

6.D Matching and Regression

Estimate $\mu_w(x)$, and modify matching estimator to:

$$\tilde{Y}_i(t) = \begin{cases} Y_i & \text{if } W_i = t, \\ Y_{\ell(i)} + \hat{\mu}_t(X_i) - \hat{\mu}_t(X_{j(i)}) & \text{if } W_i = c \end{cases}$$

$$\tilde{Y}_i(c) = \begin{cases} Y_i & \text{if } W_i = c, \\ Y_{\ell(i)} + \hat{\mu}_c(X_i) - \hat{\mu}_c(X_{j(i)}) & \text{if } W_i = t \end{cases}$$

Then the bias corrected matching estimator is:

$$\hat{\tau}^{bcm} = \frac{1}{N} \sum_{i=1}^{N} \left(\tilde{Y}_i(t) - \tilde{Y}_i(c) \right)$$

6.E Variance Estimation

Most (all sensible) other estimators for Average Treatment Effects have the form

$$\hat{\tau} = \sum_{i:W_i=t} \lambda_i \cdot Y_i - \sum_{i:W_i=c} \lambda_i \cdot Y_i,$$

(linear in outcomes) with weights

$$\lambda_i = \lambda(\mathbf{W}, \mathbf{X}, i), \qquad \sum_{i:W_i=c} \lambda_i = 1, \qquad \sum_{i:W_i=t} \lambda_i = 1$$

$\lambda(\mathbf{W}, \mathbf{X}, i)$ (known) is very non-smooth for matching estimators and bootstrap is **not** valid as a result. (not just no second order justification, but not valid asymptotically – see Abadie & Imbens 2008). For other estimators leads to variance for $\hat{\tau} - \tau_{\text{SP}}$.

Variance conditional on $\mathbf{W}$ and $\mathbf{X}$ (relative to $\hat{\tau} - \tau_{\text{cond}}$) is

$$\mathbb{V}\left(\hat{\tau}|\mathbf{W},\mathbf{X}\right) = \sum_{i=1}^{N} \lambda_i^2 \cdot \sigma_{W_i}^2(X_i)$$

All components of variance known other than $\sigma_{W_i}^2(X_i)$.

Abadie-Imbens Variance Estimator: for each unit find the closest unit with same treatment:

$$h(i) = \min_{j \neq i, W_j = W_i} \|X_i - X_j\|.$$

Then

$$\hat{\sigma}_{W_i}^2(X_i) = \frac{1}{2}\left(Y_i - Y_{h(i)}\right)^2.$$

Even though $\hat{\sigma}_w^2(x)$ is not consistent for $\mathbb{V}(Y_i(w)|X_i = x)$, the estimator for $\mathbb{V}(\hat{\tau}|\mathbf{W},\mathbf{X})$ is because it averages over all $\hat{\sigma}_w^2(X_i)$.

6.F Estimates

Lalonde/Dehejia/Wahba Data

Estimator	Estimate	(s.e.)
Subclassification	1.89	0.76
Matching (single nearest neighbor)	1.74	0.76

Lottery Data (outcome rescaled by average prize, 55.2K)

Estimator	Estimate	(s.e.)
Subclassification	-0.104	0.021
Matching (single nearest neighbor)	-0.080	0.019

Subclassification and matching estimates similar.

7.A Assessing Unconfoundedness by Estimating Effect on Pseudo Outcome

We consider the unconfoundedness assumption,

$$Y_i(c), Y_i(t) \perp\!\!\!\perp W_i \mid X_i.$$

We partition the vector of covariates X_i into two parts, a (scalar) pseudo outcome, denoted by X_i^{p}, and the remainder, denoted by X_i^{r}, so that $X_i = (X_i^{\mathsf{p}}, X_i^{\mathsf{r}\prime})'$.

Now we assess whether the following conditional independence relation holds:

$$X_i^{\mathsf{p}} \perp\!\!\!\perp W_i \mid X_i^{\mathsf{r}}.$$

The two issues are
(i) the interpretation of this indendepence condition
(ii) the implementation of the test.

7.B Link between the conditional independence relation and unconfoundedness

This link is indirect, as unconfoundedness cannot be tested directly.

First consider a related condition:

$$Y_i(c), Y_i(t) \perp\!\!\!\perp W_i \mid X_i^{\mathsf{r}}$$

If this modified unconfoundedness condition were to hold, one could use the adjustment methods using only the subset of covariates X_i^{r}. In practice this is a stronger condition than the original unconfoundedness condition.

The modified condition is not testable. We use the pseudo outcome X_i^{p} as a proxy variable for $Y_i(c)$, and test

$$X_i^{\mathsf{p}} \perp\!\!\!\perp W_i \mid X_i^{\mathsf{r}}$$

A leading example is that where X_i contains multiple lagged measures of the outcome.

In the lottery example we have observations on earnings for six, years prior to winning. Denote these lagged outcomes by $Y_{i,-1},\ldots,Y_{i,-6}$, where $Y_{i,-1}$ is the most recent and $Y_{i,-6}$ is the most distant pre-winning earnings measure.

One could implement the above ideas using $X_i^{\mathsf{p}} = Y_{i,-1}$ and $X_i^{\mathsf{r}} = (Y_{i,-2},\ldots,Y_{i,-6})$ (ignoring additional pre-treatment variables).

7.C Implementation of Assessment of Unconfoundedness

One approach to testing the conditional independence assumption is to estimate the average difference in X_i^{p} by treatment status, after adjusting for differences in X_i^{r}.

This is exactly the same problem as estimating the average effect of the treatment, using X_i^{p} as the pseudo outcome and X_i^{r} as the vector of pretreatment variables.

7.D Lottery Data

subclassification estimator

Pseudo Outcome	Remaining Covariates	est	(s.e.)
Y_{-1}	$X, Y_{-6:-2}, Y_{-6:-2} > 0$	-0.010	(0.010)
$\frac{Y_{-1}+Y_{-2}}{2}$	$X, Y_{-6:-3}, Y_{-6:-3} > 0$	-0.021	(0.013)
$\frac{Y_{-1}+Y_{-2}+Y_{-3}}{3}$	$X, Y_{-6:-4}, Y_{-6:-4} > 0$	-0.013	(0.018)

Actual Outcome			
Y	$X, Y_{-1:-6}, Y_{-6:-1} > 0$	**-0.104**	**(0.021)**

Conclusion for Lottery Data

Unconfoundedness is plausible (based on analysis without outcome data)

Estimates are robust, -0.09 is reduction in yearly labor earnings per dollar of yearly unearned income.

7.E Lalonde-Dehejia-Wahba Data

subclassification estimator

Pseudo Outcome	Remaining Covariates	est	(s.e.)
`earn '75`	X,`earn'74,unempl'74`	-1.51	(0.33)
Actual Outcome			
`earn '78`	X,`earn'74,unempl'74, earn'75,unempl'75`	**1.89**	**(0.76)**

Conclusion for Lalonde/Dehejia/Wahba Data

It is not clear whether unconfoundedness holds based on analysis without outcome data.

Estimates are robust, and in fact close to experimental estimates, but that they are accurate would have been difficult to ascertain without experimental data.

8. Overall Conclusion

Progam Evaluation methods under unconfoundedness are well established now.

Work well in practice.

Subclassification with regression, and matching with regression are more robust and credible than regression alone.