

Reducing Selection Bias via Propensity Score Approach

Bo Lu, Ph.D.
Division of Biostatistics
College of Public Health
The Ohio State University
U.S.A

[Outline]

- Overview: causal inference
- Real example: evaluation of training program
- Theory on propensity score
- Simulation example
- Implementation of propensity score matching
- Revisit the real example

[Overview: Causal Inference]

- Potential Outcome

$Y \sim$ outcome

$T \sim$ treatment indicator

$X \sim$ covariate (pretreatment)

What would have happened to those who, in fact, received treatment, if they have not received treatment (or vice versa)?

Y_{1i} denotes the outcome of individual i given being treated

Y_{0i} denotes the outcome of individual i given being control

$\Delta_i = Y_{1i} - Y_{0i}$ is the treatment effect on i

Sub.	Y_1	Y_0	Δ
A	15		
B	13		
C		8	
D		4	

Y_{1i} denotes the outcome of individual i given being treated

Y_{0i} denotes the outcome of individual i given being control

$\Delta_i = Y_{1i} - Y_{0i}$ is the treatment effect on i

Sub.	Y_1	Y_0	Δ
A	15	10	5
B	13	8	5
C	13	8	5
D	9	4	5

Suppose we also know the covariate X , which is associated with the treatment reception

Sub.	X	Y_1	Y_0	Δ
A	40	15	10	5
B	30	13	8	5
C	30	13	8	5
D	20	9	4	5

[]

In a perfect world, we can observe both Y_{1i} and Y_{0i} .

Individual treatment effect: $Y_{1i} - Y_{0i}$

Average treatment effect: $E(Y_{1i} - Y_{0i})$

Subgroup treatment effect: $E(Y_{1i} - Y_{0i} | X)$

However, in reality, we can never observe both Y_{1i} and Y_{0i} .

$$Y_{\text{obs},i} = (1-T) \times Y_{0i} + T \times Y_{1i}$$

The best we can do is to find an approximation for the potential outcome.

[]

- Randomization

RCT is the best available study design to explore causal effect

$$(Y_1, Y_0) \perp\!\!\!\perp T$$

$$E(Y_{1i} - Y_{0i}) = E(Y_{1i} - Y_{0i} | T)$$

$$= E(Y_{1i} | T) - E(Y_{0i} | T)$$

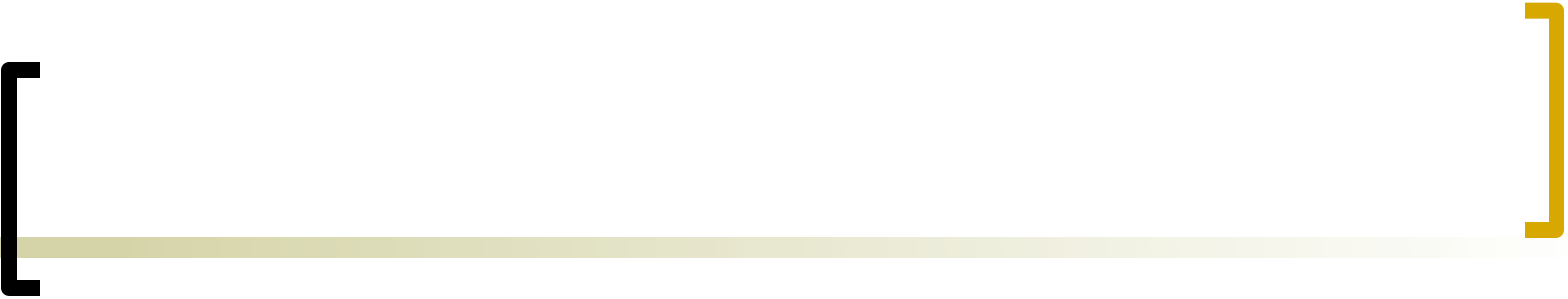
$$= E(Y_i | T=1) - E(Y_i | T=0)$$

No confounding effect in RCT



RCT has its own limitations:

- RCT is not feasible for all causal effect studies—unethical, legal issues, etc
- Small RCT may still suffer from unbalanced covariate distribution
- Large RCT could be both costly and time-consuming

- 
- A large black left bracket and a yellow right bracket are positioned at the top of the slide, with a horizontal line passing through them.
- The units in observational studies are usually more representative, since randomized studies generally have to be conducted in a restricted environment.
 - Many studies designed as randomized experiments become more like observational studies when protocols are broken

A large black left bracket and a large yellow right bracket are positioned at the top of the slide, with a thin yellow horizontal line between them.

■ Observational Studies

- treatment assignment is not random
- the study could be carried out in a time- and money-efficient manner
- traditional statistical analysis may provide biased results due to the self-selection into treatment

Example: Evaluation of Training Program

- National Supported Work (NSW) Demonstration

A randomized study implemented in the mid-1970s to provide work experience for a period of 6-18 months to individual who had faced economic and social problems.

Outcome: difference in annual earning between pre- and post-intervention

A large black left bracket and a large yellow right bracket are positioned at the top of the slide, with a horizontal olive-green line passing through them.

■ Lalonde's Analysis (1986)

- Estimate from experimental data
- Estimate from non-experimental data: combine treated subjects in NSW with control subjects from PSID or CPS
- Pre-intervention characteristics:
 - age, education, Black, Hispanic, no-degree, married, earnings in 74 and 75

A large black left bracket '[' is on the left, and a large yellow right bracket ']' is on the right. A horizontal line with a light yellow gradient runs across the top of the slide, positioned between the two brackets.

- linear regression, fixed-effects and latent variable selection model

- Dehejia and Wahba's analysis (1999)
 - re-analyze the data with propensity score matching/stratification

[Theory on Propensity Scores]

- First established in the seminal paper by Rosenbaum and Rubin (1983)

- Assumptions

- Stable unit treatment value assumption (SUTVA)

The response of subject i to the treatment T does not depend on the treatment given to subject j .

[]

- Strongly ignorable treatment assignment assumption

$(Y_1, Y_0) \perp\!\!\!\perp T | X$, conditional independence

$0 < P(T=1|X) < 1$, common support

$e(x) = P(T=1|x)$ is defined as the propensity score, which is a scalar summary of all observed covariates

A large black left bracket and a yellow right bracket are positioned at the top of the slide, with a horizontal line passing through them.

■ Key results

- Propensity score is a balancing score

$$X \perp T \mid e(X)$$

$$P(T=1 \mid X, e(X)) = P(T=1 \mid e(X))$$

- Average treatment effect at $e(X)$ is the average difference between the observed responses in each treatment group at $e(X)$

$$E(Y_1 - Y_0 \mid e(X)) = E(Y \mid e(X), T=1) - E(Y \mid e(X), T=0)$$

- []
- The overall average treatment effect is the individual treatment effect averaged over the distribution of $e(X)$

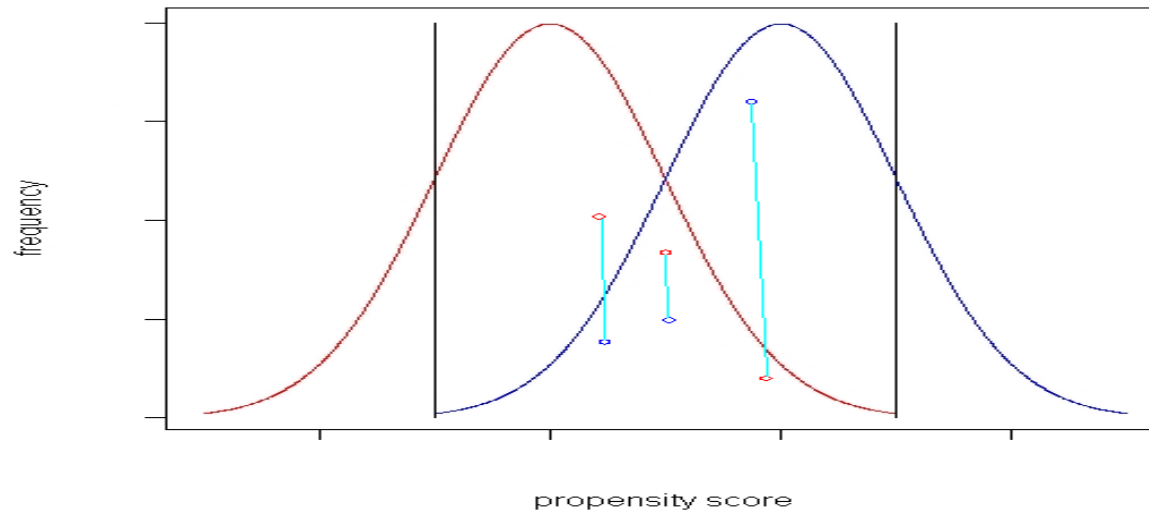
$$E(Y_1 - Y_0) = E[E(Y_1 - Y_0 | e(X))]$$

[]

- Analytical use of propensity score

- Matching

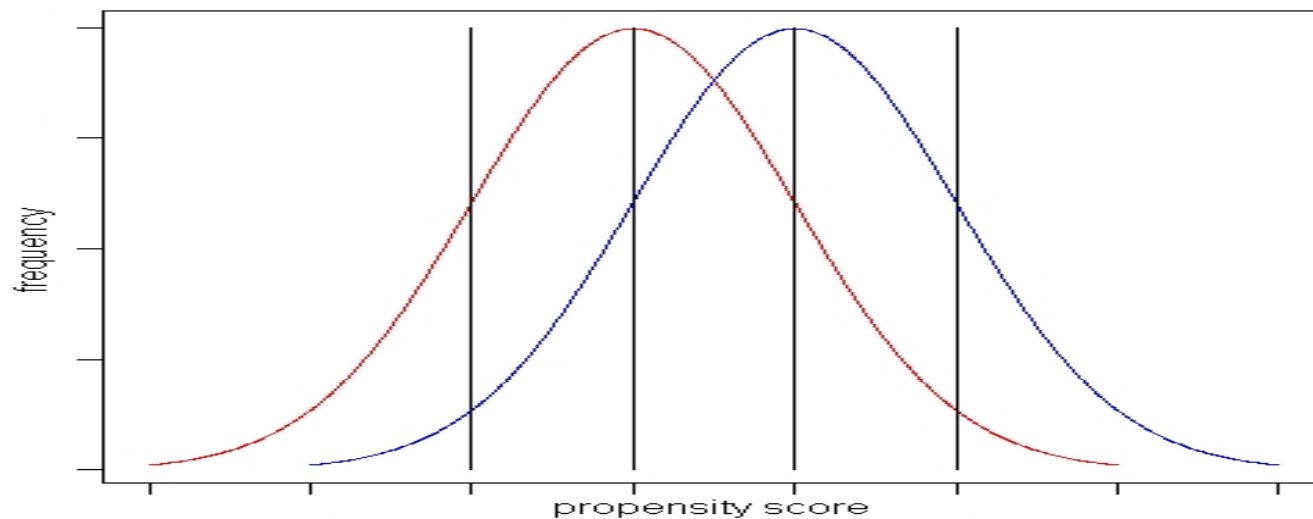
Subsets consisting both treated and control subjects with the same propensity score



[

- Stratification

The data is divided into several strata based on propensity score, then regular analysis carried out within each strata



- [
- Used as weight
propensity score is considered as the
sampling weight
-]

$$E(Y_1 - Y_0) = E\left[\frac{T \times Y}{e(X)} - \frac{(1 - T) \times Y}{1 - e(X)}\right]$$

[Simulation Example]

- Simulation Setup

T: treatment indicator (1,0)

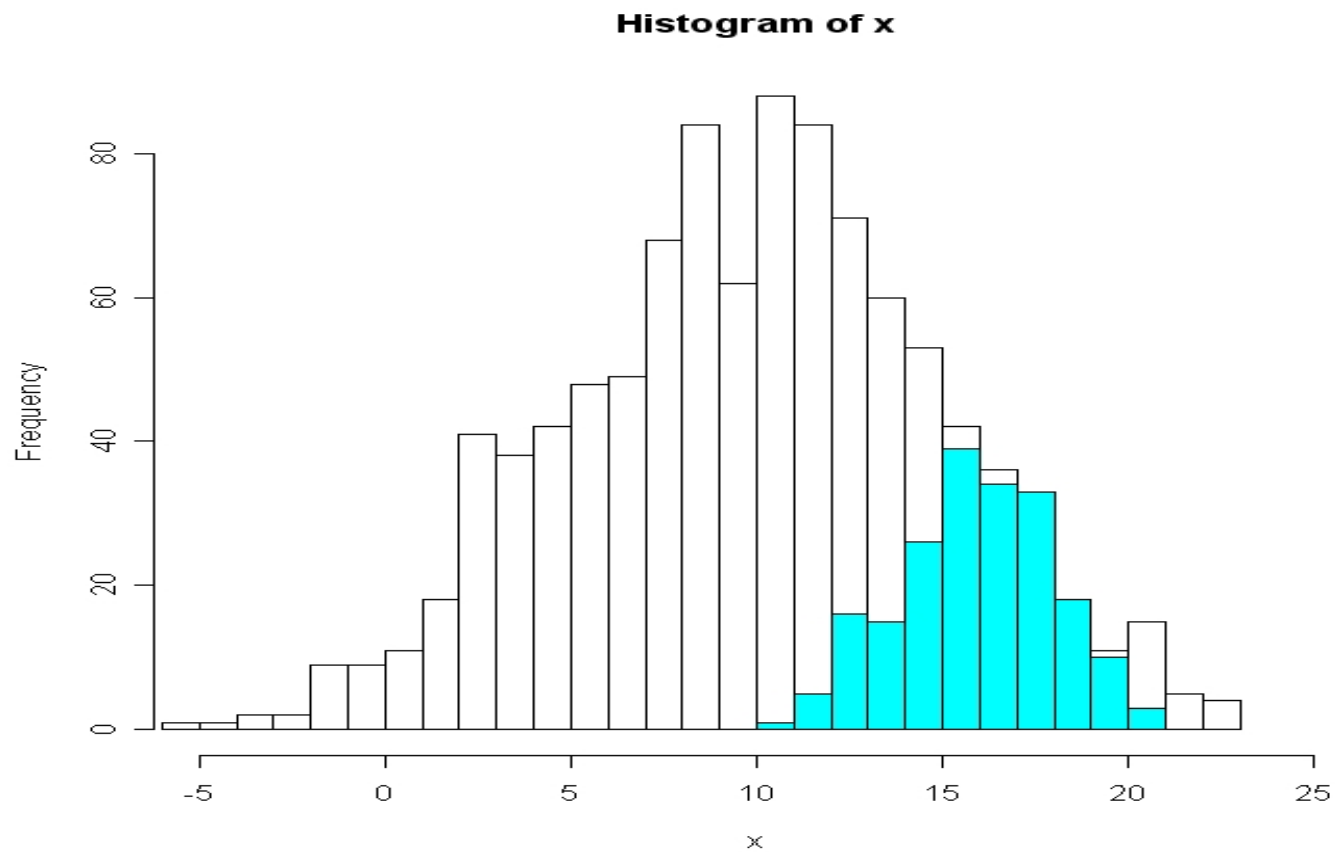
X: covariate, normally distributed

T=1, $N(16,4)$, $n_1=200$

T=0, $N(10,25)$, $n_0=1000$

Y: outcome, determined by
hypothetical models

Histogram of x in treated and control groups



[]

- Linear model

$$Y1 = b0 + c \times T + b1 \times X + \text{err}$$

$c=5$: treatment effect

$b0=3$: intercept

$b1=1$: covariate effect

err: random noise $N(0,9)$

[]

- Treatment effect

Naive path: think subjects randomly selected into treatment

	Estimate	Std. Error	t value	Pr(> t)
T	11.1549	0.4294	25.98	<2e-16 ***

Overestimate the treatment effect.
Why?

[

]

- Problem

Subjects with high covariate value tend to select treatment

```
> t.test(x1,x0)
```

```
t = 27.6809, df = 745.829, p-value < 2.2e-16
```

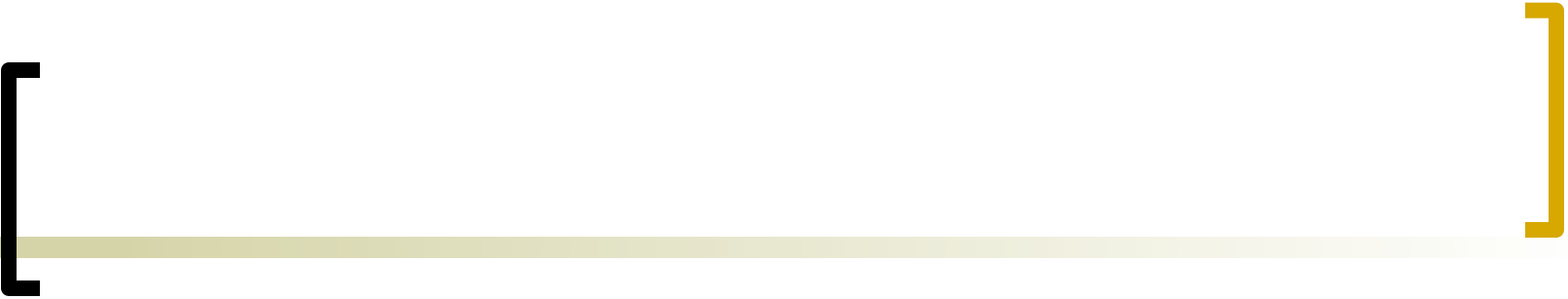
```
95 percent confidence interval:
```

```
 5.515863 6.357964
```

```
sample estimates:
```

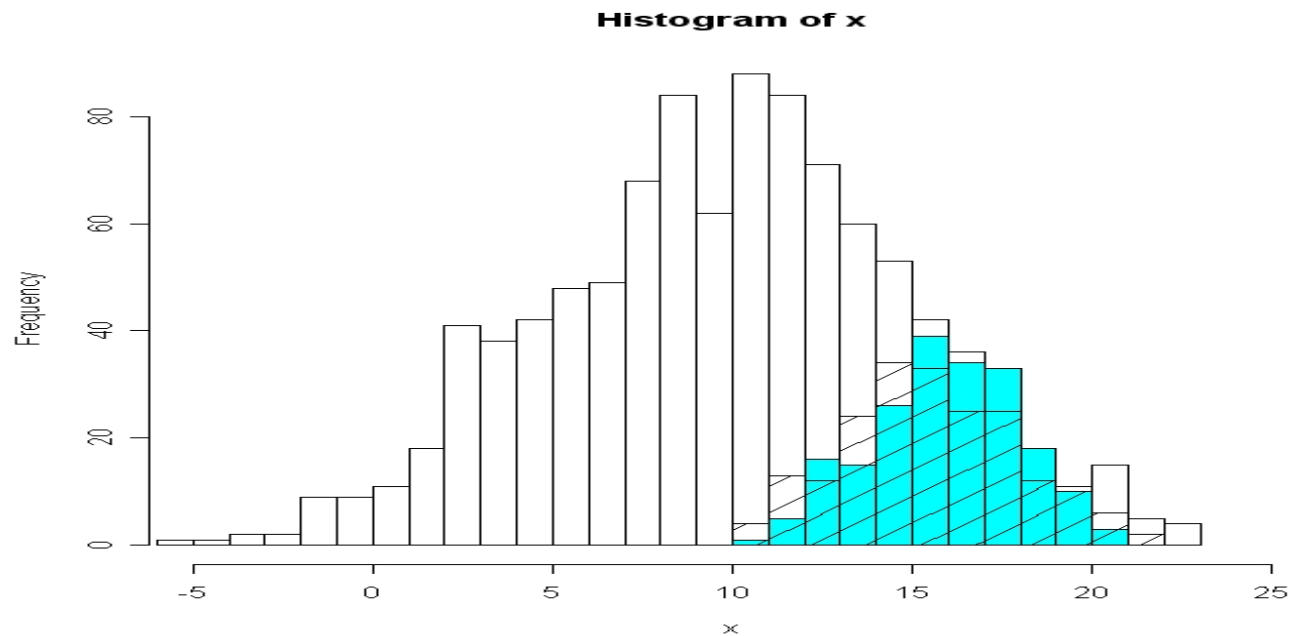
```
mean of x mean of y
```

```
15.885911 9.948997
```

- 
- A large black left bracket and a yellow right bracket are positioned at the top of the slide, with a thin yellow horizontal line extending between them.
- Matching to balance the covariate distribution
 - To make the treated and control subject look alike before treatment
 - To produce a study regime which resembles a randomized experiment most, in terms of the observed covariates

■ Pair matching

Select 200 subjects in the control group, which resemble the treated most.



[]

- Checking the balance on x

```
> t.test(x1,x.m1m[,2])
```

```
t = 1.5153, df = 388.681, p-value = 0.1305
```

```
95 percent confidence interval:
```

```
-0.1004297  0.7755485
```

```
sample estimates:
```

```
mean of x mean of y
```

```
15.88591  15.54835
```

- [
-]
- Treatment effect (pair matching)
Comparison only made within the
matched subgroups (n=400)

	Estimate	Std. Error	t value	Pr(> t)
T	5.4159	0.3792	14.28	<2e-16 ***

[Implementation of PSM]

- Estimation of propensity score

Unlike randomized trials, the propensity scores are usually unknown in observational studies, so it has to be estimated.

Usually, propensity scores are estimated by logistic regression models given the nature of the data

The ultimate goal is to balance the pretreatment covariates distribution



- Inclusion of covariates

Include as many observed pretreatment variables as possible; the statistical significance of individual terms are not as important

- Function form of covariates

Consider higher order polynomials and interaction terms to achieve better balance

- Selection of the model

Depends on the real scenario: logistic, probit, survival function



■ Matching algorithms

- Nearest Neighbor algorithm

Iteratively find the pair of subjects with the shortest distance

Easy to understand and implement; Offers good results in practice; fast running time; Rarely offers the best matching results (compared to optimal matching)



- Optimal algorithm

To minimize the total distance for the overall population

Offers the “best” matching results overall;
Runs reasonably fast; Implementation is not easy; Not readily to extend to n-cube matching ($n > 2$)

[]

- Heckman's difference-in-difference matching

$$E(Y_{1t} - Y_{0t} | X, T=1) - E(Y_{0t} - Y_{0t'} | X, T=0)$$

it requires repeated observations for the same subjects before and after the treatment applied; it accommodates multiple matching by weighting

$$ATE = \{ \sum (Y_{1ti} - Y_{0t'i}) - \sum w_{ij} (Y_{0tj} - Y_{0t'j}) \} / n_1$$

weights decided by kernel or other methods

A large black left bracket and a yellow right bracket are positioned at the top of the slide, with a horizontal olive-green line extending between them.

- Choices of distance

exact match not possible, use one distance measure to summarize the information

- * Mahalanobis distance
- * Propensity score
- * Mahalanobis distance with propensity score caliper
- * Any distance with the requirement of exact match on a specific variable

A large black left bracket and a large yellow right bracket are positioned at the top of the slide, with a horizontal olive-green line passing through them.

■ Matching Design

- Bi-partite matching

Pair matching: used when the numbers of the treated and control are comparable

1-K matching: used when control group is huge compared to the treated

Variable matching: more flexible than 1-K, the matched control is set to be between a and b for each treated subject



Full matching: a way of sub-classification generalizing variable matching; each matched group contains one treated and multiple control or one control with multiple treated



- Non-bipartite Matching

When there are multiple treatment groups or one treatment group with several control groups or treatment status changing over time

Pair matching

1-K matching

Require special algorithm (complicated!)

A large black left bracket and a large yellow right bracket are positioned at the top of the slide, with a horizontal line between them. The line is light yellow and spans most of the width of the slide.

■ Available Software

- SAS procs by Bergstralh, Kosanke, Jacobsen (1996) “Software for Optimal Matching in Observational Studies”, Epidemiology, 7, 331-332

<http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm>

Bipartite matching: pair, 1-k, variable



- R functions by Ben Hansen

<http://www.stat.lsa.umich.edu/~bbh/#>

full matching

- STATA function by Abadie, et al.

<http://elsa.berkeley.edu/~imbens/statamatching.pdf>

Nearest Neighbor matching, estimating treatment effect proposed in econ literatures: SATE, PATE, SATT, ATT, etc

[]

- STATA functions by Leuven & Sianesi
psmatch2()

[http://athena.sas.upenn.edu/~petra/copen/
statadoc.pdf](http://athena.sas.upenn.edu/~petra/copen/statadoc.pdf)

Mahalanobis or propensity score distance matching for various designs : pair, kernel, local linear and spline matching (based on nearest-neighbor matching)

[]

- R functions for optimal matching (under development)

Core algorithms:

- a. FORTRAN codes for optimal non-bipartite matching by Derigs (1988)
- b. C codes for optimal weighted matching

<http://elib.zib.de/pub/Packages/mathprog/matching/weighted/>



Loaded into R with `.Fortran()/.C()`

No Fortran or C compiler needed

Downloadable on line

Works under both UNIX and WINDOWS

- UNIX: load `.so` file
- WINDOWS: load `.dll` file

Further questions, contact Bo Lu blu@cph.osu.edu

A large black left bracket '[' is on the left, and a large yellow right bracket ']' is on the right. A horizontal line with a light yellow gradient runs across the top of the slide, positioned between the two brackets.

■ Practical issues

- Matching vs. Covariance adjustment modeling

Matching: always reduce the bias; no worry about the true regression equation; easy post-matching analysis; restricted to common support

Covariance adjustment modeling: has to guess the true regression equation (prone to bias); apply to the full range of the data; may lead to smaller variance estimation

[]

- Exact matching vs. Complete matching

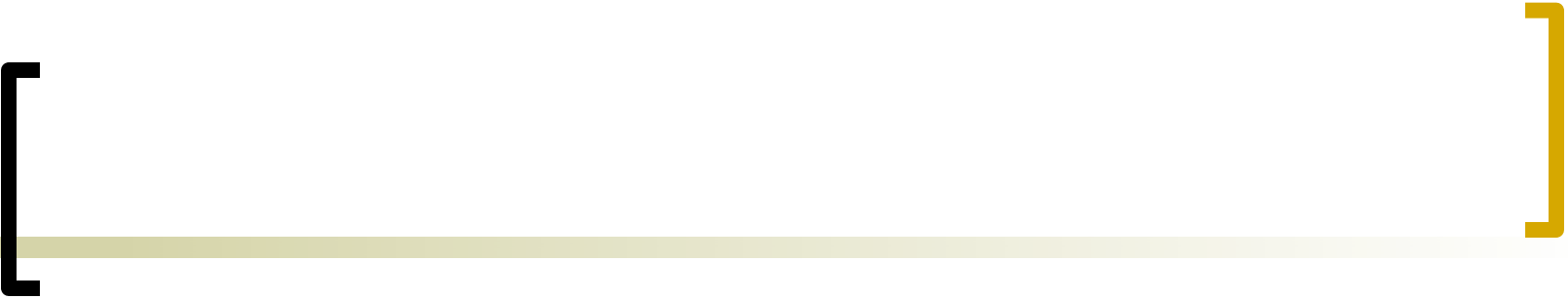
Exact: balanced treated and control group; less usable data (treated cases may be excluded)

Complete: try to use all data; less balanced covariate distribution; may need post-matching regression adjustment

A large black left bracket and a large yellow right bracket are positioned at the top of the slide, with a horizontal line passing through them.

■ Procedure for PSM

- Identify propensity score model
- Estimate the propensity score with all data
- Compute the distance between any two subjects
- Create matched pair/group using a specific matching algorithm

- 
- A large black left square bracket and a large yellow right square bracket are positioned at the top of the slide, with a thin yellow horizontal line extending between them.
- Check covariate balance between the treated and control among the matched subjects; If not good enough, go back to improve propensity score model
 - Contrast between the treated and control subjects within each pair/group
 - Obtain the average treatment effect by averaging over all pairs/groups

[Revisit The Example]

The goal is to investigate the credibility of the conventional analytical results from non-experimental data

So, the authors compared the results from the experimental data and the results from non-experimental data by combining the treated with an existing comparable control dataset.

- Pre-treatment covariates distribution

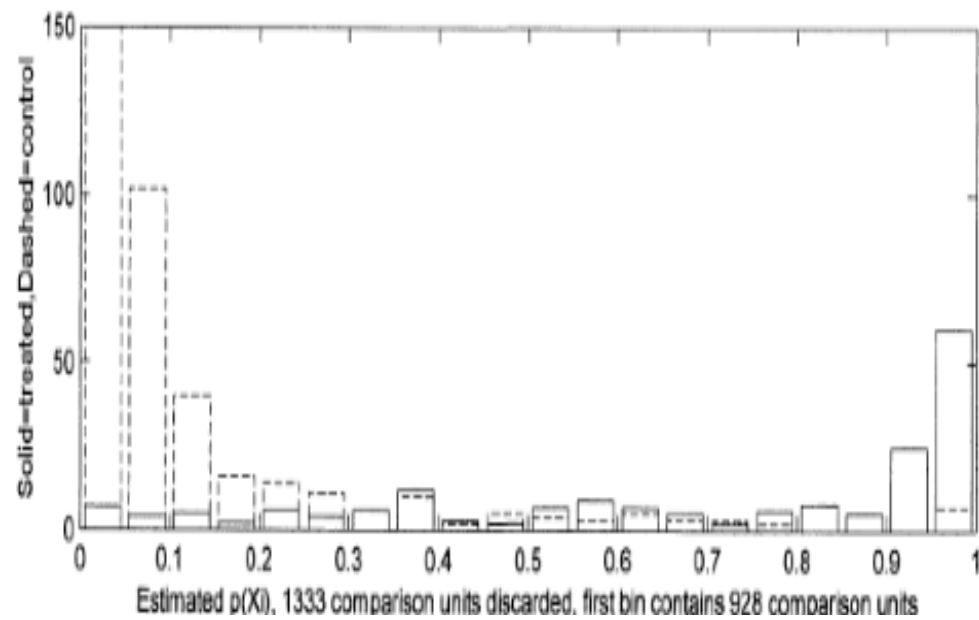
Table 1. Sample Means of Characteristics for NSW and Comparison Samples

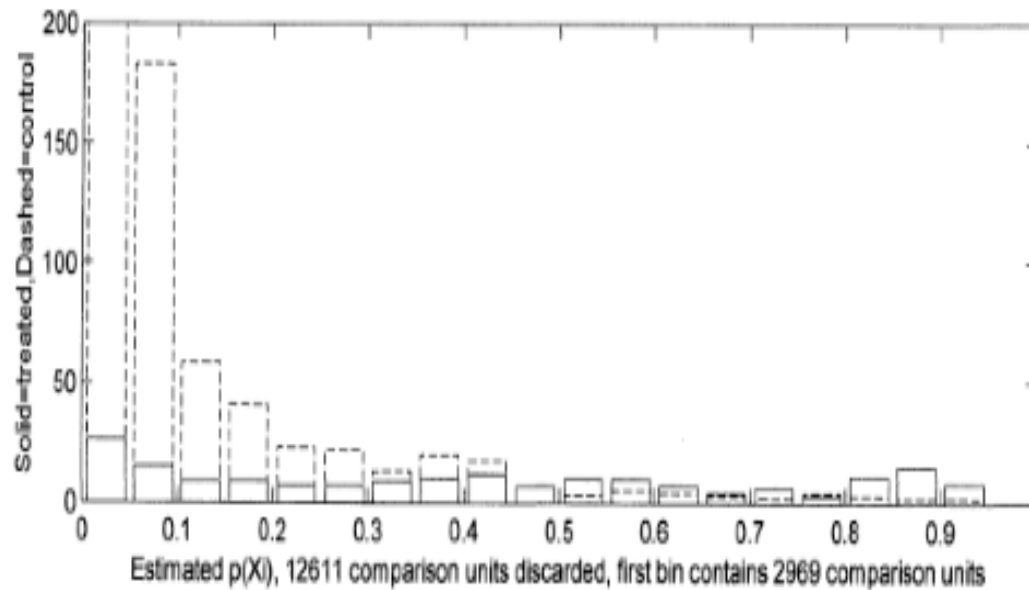
	No. of observations	Age	Education	Black	Hispanic	No degree	Married	RE74 (U.S. \$)	RE75 (U.S. \$)
NSW/Lalonde: ^a									
Treated	297	24.63 (.32)	10.38 (.09)	.80 (.02)	.09 (.01)	.73 (.02)	.17 (.02)		3,066 (236)
Control	425	24.45 (.32)	10.19 (.08)	.80 (.02)	.11 (.02)	.81 (.02)	.16 (.02)		3,026 (252)
RE74 subset: ^b									
Treated	185	25.81 (.35)	10.35 (.10)	.84 (.02)	.059 (.01)	.71 (.02)	.19 (.02)	2,096 (237)	1,532 (156)
Control	260	25.05 (.34)	10.09 (.08)	.83 (.02)	.1 (.02)	.83 (.02)	.15 (.02)	2,107 (276)	1,267 (151)
Comparison groups: ^c									
PSID-1	2,490	34.85 [.78]	12.11 [.23]	.25 [.03]	.032 [.01]	.31 [.04]	.87 [.03]	19,429 [991]	19,063 [1,002]
PSID-2	253	36.10 [1.00]	10.77 [.27]	.39 [.04]	.067 [.02]	.49 [.05]	.74 [.04]	11,027 [853]	7,569 [695]
PSID-3	128	38.25 [1.17]	10.30 [.29]	.45 [.05]	.18 [.03]	.51 [.05]	.70 [.05]	5,566 (686)	2,611 [499]
CPS-1	15,992	33.22 [.81]	12.02 [.21]	.07 [.02]	.07 [.02]	.29 [.03]	.71 [.03]	14,016 [705]	13,650 [682]
CPS-2	2,369	28.25 [.87]	11.24 [.19]	.11 [.02]	.08 [.02]	.45 [.04]	.46 [.04]	8,728 [667]	7,397 [600]
CPS-3	429	28.03 [.87]	10.23 [.23]	.21 [.03]	.14 [.03]	.60 [.04]	.51 [.04]	5,619 [552]	2,467 [288]

[

]

- Checking the propensity score overlap





Many observations in control need to be discarded

- Checking the balance after the matching

Table 4. Sample Means of Characteristics for Matched Control Samples

Matched samples	No. of observations	Age	Education	Black	Hispanic	No degree	Married	RE74 (U.S. \$)	RE75 (U.S. \$)
NSW	185	25.81	10.35	.84	.06	.71	.19	2,096	1,532
MPSID-1	56	26.39	10.62	.86	.02	.55	.15	1,794	1,126
		[2.56]	[.63]	[.13]	[.06]	[.13]	[.12]	[1,406]	[1,146]
MPSID-2	49	25.32	11.10	.89	.02	.57	.19	1,599	2,225
		[2.63]	[.83]	[.14]	[.08]	[.16]	[.16]	[1,905]	[1,228]
MPSID-3	30	26.86	10.96	.91	.01	.52	.25	1,386	1,863
		[2.97]	[.84]	[.13]	[.08]	[.16]	[.16]	[1,680]	[1,494]
MCPS-1	119	26.91	10.52	.86	.04	.64	.19	2,110	1,396
		[1.25]	[.32]	[.06]	[.04]	[.07]	[.06]	[841]	[563]
MCPS-2	87	26.21	10.21	.85	.04	.68	.20	1,758	1,204
		[1.43]	[.37]	[.08]	[.05]	[.09]	.08	[896]	[661]
MCPS-3	63	25.94	10.69	.87	.06	.53	.13	2,709	1,587
		[1.68]	[.48]	[.09]	[.06]	[.10]	[.09]	[1,285]	[760]

NOTE: Standard error on the difference in means with NSW sample is given in brackets.

MPSID1-3 and MCPS1-3 are the subsamples of PSID1-3 and CPS1-3 that are matched to the treatment group.

- Comparison of the analytical results

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
	(1) Unadjusted	(2) Adjusted ^a	Quadratic in score ^b (3)	Stratifying on the score			Matching on the score	
				(4) Unadjusted	(5) Adjusted	(6) Observations ^c	(7) Unadjusted	(8) Adjusted ^d
NSW	1,794 (633)	1,672 (638)						
PSID-1 ^e	-15,205 (1,154)	731 (886)	294 (1,389)	1,608 (1,571)	1,494 (1,581)	1,255	1,691 (2,209)	1,473 (809)
PSID-2 ^f	-3,647 (959)	683 (1,028)	496 (1,193)	2,220 (1,768)	2,235 (1,793)	389	1,455 (2,303)	1,480 (808)
PSID-3 ^f	1,069 (899)	825 (1,104)	647 (1,383)	2,321 (1,994)	1,870 (2,002)	247	2,120 (2,335)	1,549 (826)
CPS-1 ^g	-8,498 (712)	972 (550)	1,117 (747)	1,713 (1,115)	1,774 (1,152)	4,117	1,582 (1,069)	1,616 (751)
CPS-2 ^g	-3,822 (670)	790 (658)	505 (847)	1,543 (1,461)	1,622 (1,346)	1,493	1,788 (1,205)	1,563 (753)
CPS-3 ^g	-635 (657)	1,326 (798)	556 (951)	1,252 (1,617)	2,219 (2,082)	514	587 (1,496)	662 (776)

KNAW, March 29, 2007

A large black left bracket and a yellow right bracket are positioned at the top of the slide, with a horizontal line passing through them.

- Observations

- The results after propensity score matching/stratification are much closer to the truth (experimental data analysis)
- The variances seem to be larger due to the loss of the data
- The results are not very sensitive to the function form of the chosen covariates in propensity score model; however, they are sensitive to the selection of covariates to be included in the propensity score model

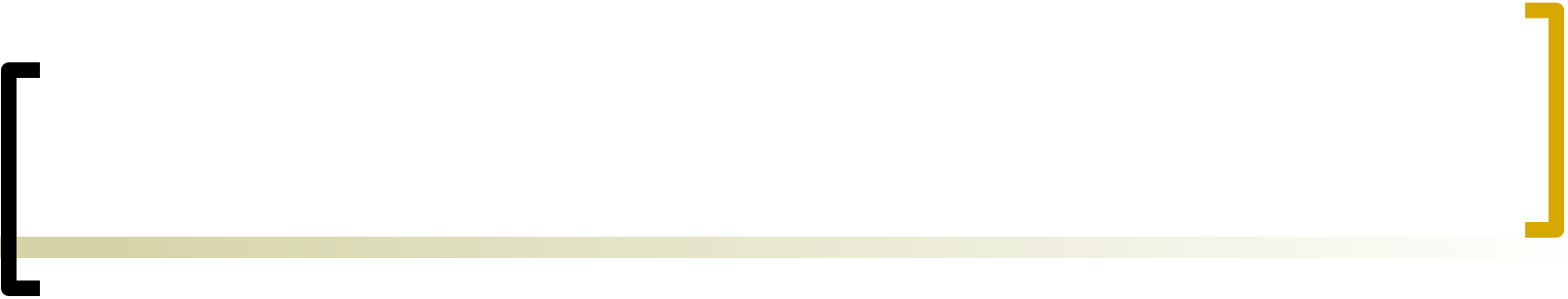
[Limitations and new advances]

- Limitation of PS method

- Rely on a unverifiable assumption: strongly ignorable treatment assignment given the observed covariates

- Unlike the randomized studies, it has no control over the unobserved confounders

- One possible solution is to use sensitivity analysis to evaluate to what degree the results will change given a hypothesized unknown covariate

A large black left bracket '[' is on the left, and a large yellow right bracket ']' is on the right. A horizontal line with a light green-to-yellow gradient runs across the top of the slide, passing between the two brackets.

- Need substantial overlap between the treated and the control groups, otherwise, it may result in significant loss of the data in analysis

One possible solution is to use regression-like technique to extrapolate; however, such extrapolation might not be reliable

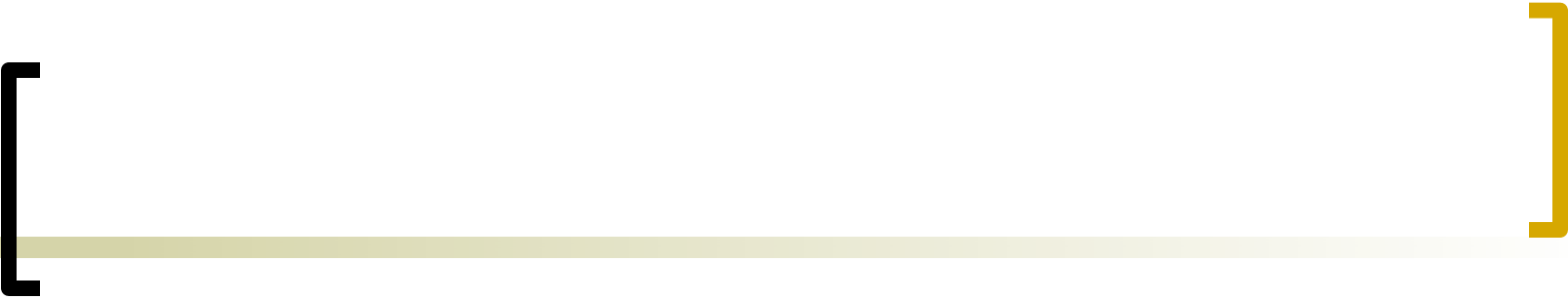
- 
- A large black left bracket and a yellow right bracket are positioned at the top of the slide, with a horizontal line passing through them.
- Apply propensity score in longitudinal studies

construct time-dependent propensity score

- sequential matching
- inverse-probability-of treatment weighted (IPTW) estimator

[Reference]

- D'Agostino (1998), "Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group", *Stat. Med.* 17, 2265-2281.
- Dehejia & Wahba (1999), "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs", *JASA*, 94, 1053-1062.
- Dergis (1988), "Solving non-bipartite matching problem via shortest path techniques", *Annals of Operations Research*, 13, 225-261.
- Joffe & Rosenbaum (1999), "Propensity scores", *American J. of Epi.*, 150, 327-333.
- Imbens (2004), "Nonparametric estimation of average treatment effects under exogeneity", *Review of economics and statistics*, 86, 4-29.

- 
- A large black left bracket and a yellow right bracket are positioned at the top of the slide, with a horizontal olive-green line passing through them.
- Lalonde (1986), “Evaluating the econometric evaluations of training programs”, *American Economic Review*, 76, 604-620.
 - Lu (2005), “Propensity score matching with time dependent covariates”, *Biometrics*, 61, 721-728.
 - Lu, Zanutto, Hornik & Rosenbaum (2001), “Matching with doses in an observational study of a media campaign against drug abuse”.
 - Robins, Hernan & Brumback (2000), “Marginal structural models and causal inference in epidemiology”, *Epidemiology*, 11, 550-560.
 - Rosenbaum (1987), “Model-based direct adjustment”, *JASA*, 82, 387-394.



- Rosenbaum (1989), “Optimal matching for observational studies”, *JASA*, 84, 1024-1032.
- Rosenbaum (2002), *Observatioal Studies*, 2nd Edition, Springer.
- Rosenbaum & Rubin (1983), “The central role of the propensity score in observational studies for causal effect”, *Biometrika*, 70, 41-55.
- Rosenbaum & Rubin (1985), “The bias due to incomplete matching”, *Biometrics*, 41, 103-116.
- Rubin (2005), “Causal inference using potential outcomes: Design, Modeling, Decision”, *JASA*, 100, 322-331.

[Acknowledgement]

My research on optimal nonbipartite matching was partially supported by funding provided by a seed grant from OSU's Initiative in Population Research Center Grant (NICHD, R21, HD-47943-03)