

# The Central Role of the Propensity Score in Observational Studies for Causal Effects Paper Review [Rosenbaum and Rubin, 1983]

Jai-Hua Kevin Yen

Department of Biostatistics  
University at Buffalo

October 23, 2025

# Table of Contents

- 1 Introduction
- 2 Definitions
- 3 Theoretical Backgrounds
- 4 Propensity Score Matching
- 5 Subclassification on Propensity Score
- 6 Propensity Scores and Covariance Adjustment
- 7 References

# Introduction

# Introduction

- In observational studies, groups are not randomized. This leads to bias when computing treatment effect.
  - ▶ Example: Sicker patient more likely get treatment → if result shows that higher risk on treatment, we cannot conclude anything
  - ▶ Cause by **confounding** or **selection bias**
- **Target:** Although our target is to estimate  $E(r_{1i} - r_{0i})$ , we cannot directly observe  $r_{1i}$  and  $r_{0i}$ . Instead we can estimate the average treatment effect (ATE), i.e., estimating

$$E(r_1 - r_0)$$

# Introduction

- **Challenge:** We only observe one potential outcome for each unit. Direct comparison of treated ( $z = 1$ ) and control ( $z = 0$ ) groups is misleading because the groups may differ systematically based on their pre-treatment covariates,  $x$ .
- **Solution:** Using the propensity score  $e(x)$  to obtain unbiased estimate of ATE.

# Definitions

# Definitions

## Definition (Balancing Score)

A balancing score  $b(x)$  is a function of observed covariates  $x$  such that

$$x \perp z | b(x)$$

Note that a trivial balancing score is  $b(x) = x$ .

## Definition (Propensity Score)

A propensity score  $e(x)$  is the probability assigning to treatment given covariates  $x$ , which is denoted by

$$e(x) = P(z = 1 | x)$$

# Randomized Trial vs. Nonrandomized Trial

Randomized trials and nonrandomized trials differ in two distinct ways:

- ① In randomized trials, the propensity score is known. In nonrandomized trials, the propensity score is unknown but can be estimated by observed data.
- ② In randomized trials, it meets the condition of strongly ignorable.

# Definitions

## Definition (Exchangeability)

In a randomized trial, if  $r_1$  and  $r_0$  has exchangeability given  $x$ , then

$$(r_1, r_0) \perp z | x$$

Note: This means no other (unobserved) covariates that are associated with both treatment assignment and outcome.

## Definition (Positivity)

If all every subgroup of the population, has a non-zero probability of receiving any given treatment is called having the property of positivity. i.e.,

$$0 < P(z = 1 | x) < 1$$

Note: Although here introduces the term strong ignorable, it is just the combination of exchangeability and positivity.

# Theoretical Backgrounds

## Theorem (Theorem 1)

*Treatment assignment and the observed covariates are conditionally independent given the propensity score. i.e.,*

$$x \perp z | e(x)$$

Notes:

- This is just a special case of Theorem 2.
- This theory implies that if a matched treatment-control pair is homogenous in  $b(x)$ , then the treated and control units in that matched pair will have the same distribution of  $x$ .
- Here the propensity score acts like a sufficient statistic.

# Theories

## Definition (Finer Function)

A function  $b(x)$  is said to be finer than  $e(x)$  if

$$e(x) = f(b(x))$$

for some function  $f$ .

## Theorem (Theorem 2)

*Let  $b(x)$  be a function of  $x$ . Then  $b(x)$  is a balancing score, that is,*

$$x \perp z | b(x)$$

*if and only if  $b(x)$  is finer than  $e(x)$ .*

## Proof of Theorem 2.

See Appendix 1.1. □

## Theorem (Theorem 3)

*If treatment assignment is strongly ignorable given  $x$ , then it is strongly ignorable given any balancing score  $b(x)$ ; that is,*

$$(r_1, r_0) \perp z | x \quad \text{and} \quad 0 < P(z = 1 | x) < 1$$

*for all  $x$  imply*

$$(r_1, r_0) \perp z | b(x) \quad \text{and} \quad 0 < P(z = 1 | b(x)) < 1$$

*for all  $b(x)$ .*

## Proof of Theorem 3.

Can be proved by using Theorem 2. See Appendix 1.2. □

# Theories

- We can only compute  $E(r_1|x, z = 1) - E(r_0|x, z = 0)$ , which is obviously not ATE.
- It can be computed by  $E_x\{E(r_1|x, z = 1) - E(r_0|x, z = 0)\}$  and by strong ignorable assumption, equal to  $E_x\{E(r_1|x) - E(r_0|x)\}$  which is ATE.
- Here we can use the balancing score  $b(x)$  and the assumption of strong ignorable to compute the ATE

$$\begin{aligned} E_{b(x)}\{E[r_1|b(x), z = 1] - E[r_0|b(x), z = 0]\} = \\ E_{b(x)}[r_1|b(x)] - E[r_0|b(x)] = E[r_1 - r_0] \end{aligned}$$

- Under strongly ignorable treatment assignment, units with the same value of the balancing score  $b(x)$  but different treatments can act as controls for each other.

## Theorem (Theorem 4)

*Suppose treatment assignment is strongly ignorable and  $b(x)$  is a balancing score. Then the expected difference in observed responses to the two treatments at  $b(x)$  is equal to the ATE at  $b(x)$ , that is,*

$$E\{r_1|b(x), z = 1\} - E\{r_0|b(x), z = 0\} = E\{r_1 - r_0|b(x)\}.$$

## Proof of Theorem 4.

Under strong ignorable assumption,  $E\{r_1|b(x), z = 1\} = E(r_1|b(x))$  and  $E\{r_0|b(x), z = 0\} = E(r_0|b(x))$ . Then the rest is trivial.  $\square$

## Corollary (Corollary 4.1: Pair matching on balancing scores)

*Suppose treatment assignment is strongly ignorable. Further suppose that a value of a balancing score  $b(x)$  is randomly sampled from the population of units, and then one treated,  $z = 1$ , unit and one control,  $z = 0$ , unit are sampled with this value of  $b(x)$ . Then the expected difference in response to the two treatments for the units in the matched pair equals the ATE at  $b(x)$ . Moreover, the mean of matched pair differences obtained by this two-step sampling process is unbiased for the ATE.*

Note: This corollary shows that the average of outcome difference with all **matched pair (1 to 1)** provides an unbiased estimate of ATE.

## Corollary (Corollary 4.2: Subclassification on balancing scores)

*Suppose treatment assignment is strongly ignorable. Suppose further that a group of units is sampled using  $b(x)$  such that:*

- ❶  *$b(x)$  is constant for all units in the group*
- ❷ *at least one unit in the group received each treatment*

*Then, for these units, the expected difference in treatment means equals the ATE at that value of  $b(x)$ . Moreover, the weighted average of such differences, that is, the directly adjusted difference, is unbiased for the ATE, when the weights equal the fraction of the population at  $b(x)$ .*

Note: This corollary shows that the treatment is calculated **separately inside each subclass**, which is defined by the balancing score  $b(x)$ .

## Corollary (Corollary 4.3: Covariance adjustment on balancing scores)

*Suppose treatment assignment is strongly ignorable, so that in particular,  $E\{r_t|z = t, b(x)\} = E\{r_t|b(x)\}$  for balancing score  $b(x)$ . Further suppose that the conditional expectation of  $r_t$  given  $b(x)$  is linear:*

$$E\{r_t|z = t, b(x)\} = \alpha_t + \beta_t b(x) \quad (t = 0, 1).$$

*Then the estimator*

$$(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)b(x)$$

*is conditionally unbiased given  $b(x_i)$  ( $i = 1, \dots, n$ ) for the treatment effect at  $b(x)$*

## Corollary (Corollary 4.3: Covariance adjustment on balancing scores (continued))

*Namely*

$$E\{r_1 - r_0 | b(x)\},$$

*if  $\hat{\alpha}_t$  and  $\hat{\beta}_t$  are conditionally unbiased estimators of  $\alpha_t$  and  $\beta_t$ , such as least squares estimators. Moreover,*

$$(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)\bar{b},$$

*where  $\bar{b} = n^{-1} \sum b(x_i)$ , is unbiased for the ATE  $(1 \cdot 1)$  if the units in the study are a simple random sample from the population.*

Note: This collary shows that we can use a simple linear regression to have an unbiased estimate of the ATE

# Theories

Usually the propensity score must be estimated from available data, so here we assume that having a random sample  $(z_i, x_i)$  ( $i = 1, \dots, N$ ).

## Theorem (Theorem 5)

*Suppose  $0 < \hat{e}(a) < 1$ . Then*

$$\begin{aligned} \text{prop}\{z = 0, x = a | \hat{e}(x) = \hat{e}(a)\} &= \text{prop}\{z = 0 | \hat{e}(x) = \hat{e}(a)\} \\ \text{prop}\{x = a | \hat{e}(x) = \hat{e}(a)\}. \end{aligned}$$

Note: In practice, except when  $x$  takes on only a few values,  $e(a)$  will be either zero or one for most values of  $a$ . Consequently, in order to estimate propensity scores, some modelling will be required ([Cox, 1970] suggests using a logistic model to estimate it).

# Propensity Score Matching

# Propensity Score Matching

- Intuition: Small treatment group vs. large control reservoirs group. Want to compare the difference of some properties. Direct comparison will have some issue of "systematic difference", i.e., different initial properties, that would cause bias
- Method: For every single individual in treatment group, find one (or simetime many) individual in the control group that has the similar properties (i.e., similar propensity score) of the individual in treatment group (in mathematical language, the distance of these points are closer compare to others)
- Benefits:
  - ① Reduces bias
  - ② Increase clarity: can focus on the actual treatment effect without other systematic bias
  - ③ Improves statistical analysis: if distribution of two groups become more similar, statistical tools become more reliable and less sensitive to model assumptions

## Example Using R: Lalonde (1986) Dataset

The simulation will use the [LaLonde, 1986] dataset, a famous dataset in the causal inference literature. The goal is to estimate the effect of the job training program ( $treat = 1$ ) on real earnings in 1978 ( $re78$ ). The covariates include age, education ( $educ$ ), race, marital status ( $married$ ), high school degree status ( $nodegree$ ), and real earnings in 1974 and 1975 ( $re74$ ,  $re75$ ). The dataset combines two groups:

- Treatment Group: Individuals from the National Supported Work (NSW) Demonstration, an experimental job training program conducted in the mid-1970s
- Control Group: A non-equivalent comparison group constructed from the Panel Survey of Income Dynamics (PSID), a large-scale public survey

See Appendix 2.1

# Propensity Score Matching

We want to have the initial bias (existing difference that makes "unfair" comparison, covariate differences), which is

$$B = E(x|z = 1) - E(x|z = 0)$$

to be small as possible. However, we cannot have this result and the dataset is biased if directly used. Then after matching and obtain the matched sample dataset, we have the expected bias

$$B_m = E(x|z = 1) - E_m(x|z = 0)$$

# Propensity Score Matching

- Exact matches even on a scalar balancing score are often impossible to obtain, so methods which seek approximate matches must be used.
- Equal Percent Bias Reduction Assumption (EPBR):
  - ① Equal per cent bias reducing if the bias in each coordinate of  $x$  is reduced by the same percentage
  - ②  $B_m = \gamma B$  for some scalar  $\gamma$ , with  $0 < \gamma < 1$

Note: This ensures groups are made more similar in a more balanced and consistent way. If this assumption does not hold, for example having a dataset with age and blood pressure collected, a young person with high blood pressure will match with the old person with low blood pressure.

# Simulation Result

**Table 1:** Per cent reduction in bias due to matched sampling based on the sample and population propensity scores. 50 control units vs.  $50R$  potential control units. Study from [Rubin, 1979].

$R$	Type of score	Initial bias			
		0.25	0.50	0.75	1.00
2	Sample	92	85	77	67
	Population	92	87	78	69
3	Sample	101	96	91	83
	Population	96	95	91	84
4	Sample	97	98	95	90
	Population	98	97	94	89

# Simulation Result in R

Table 2: Percent Reduction in Bias (Replication of [Rosenbaum and Rubin, 1983] Table 1)

Ratio	Type of Score	Initial Bias			
		0.25	0.50	0.75	1.00
2	Sample	94.88	91.90	82.39	67.02
2	Population	97.45	93.13	83.28	67.83
3	Sample	95.26	95.35	91.76	82.45
3	Population	98.71	96.89	92.51	83.04
4	Sample	95.53	96.54	94.76	88.65
4	Population	99.11	98.28	95.56	89.15

See Appendix 2.2.

## Subclassification on Propensity Score

# Subclassification on Propensity Score

- **Simplicity and persuasiveness:** direct comparisons of groups that are visibly similar (can use bar charts of means)
- As the number of confounding variables ( $P$ ) increases, the number of subclasses needed to control for them grows exponentially ( $2^P$ : two category per variable case, even higher in other cases).
- Instead of creating subclasses based on every individual covariate, you can create a small number of subclasses based on a single variable: **the estimated propensity score.**

## Case Study: Chronary Artery Disease

- **Study Type:** Observational study of therapies for chronary artery disease
- **Treatments:**  $z = 1$ : Chronary artery bypass surgery;  $z = 0$ : Drug therapy
- **Covariates:**  $x$ : Clinical, haemodynamic, and demographic measurements on each patient made prior to treatment assignment
- **Note:** Covariates have quite different distributions in the two treatment groups, within each of the five subclasses, the surgical and drug patients will be seen to have similar sample distributions of  $x$ .
- **Estimation of Propensity Score:** Logistic regression for  $z$  given  $x$  (covariates and interactions were selected by stepwise procedure)

# Case Study: Chronary Artery Disease

Table 3: Data Summary of Chronary Artery Disease Case

Propensity Score Level	Surgical Patients	Drug Patients
1 (Highest)	234	69
2	164	139
3	98	205
4	68	235
5	26	277

# Case Study: Chronary Artery Disease

**Table 4:** Example of increased balance using subclassification on estimated propensity score as summarized by distributions of  $F$  statistics for 74 covariates

	Minimum	Lower quartile	Median	Upper quartile	Maximum
Treatment main effect without subclassification	4.0	6.8	10.9	16.8	51.8
Treatment main effect with subclassification	0.0	0.1	0.2	0.6	3.6
Treatment by subclass interaction	0.0	0.4	0.8	1.2	2.9

Note: This showed that within each subclass, the surgical and drug patients now had very similar distributions of the 74 covariates, creating a much fairer comparison.

# Subclassification on Propensity Score

- Issue: Generally in practice, subclasses will not be exactly homogeneous in the balancing score  $b(x)$  that was used in subclassification, so the directly adjusted estimate may contain some residual bias due to  $x$ .
- Solution: Direct adjustment based on a balancing score  $b = b(x)$  can be expected to reduce bias in each coordinate of  $x$  providing the adjustment reduces the bias in  $b$ .

# Subclassification on Propensity Score

## Theorem (Theorem 7)

*Let  $I_s$  be the set of values of a balancing score which make up subclass  $s$  ( $s = 1, \dots, S$ ), so that  $b(a) \in I_s$  implies that units with  $x = a$  fall in subclass  $s$ . Suppose the weight applied to subclass  $s$  in direct adjustment is  $w_s$ . The bias in  $x$  after direct adjustment for the subclasses ( $I_s, s = 1, \dots, S$ ) is*

$$B_s = \sum_{s=1}^S w_s \int E\{x|b\} \{pr\{b|z = 1, b \in I_s\} - pr\{b|z = 0, b \in I_s\}\} db,$$

*where  $b = b(x)$ .*

# Subclassification on Propensity Score

## Corollary (Corollary 7.1)

*If  $E\{x|b\} = \alpha + \beta f(b)$  for some vectors  $\alpha$  and  $\beta$  and some scalar valued function  $f(\cdot)$  of  $b$ , and if the subclasses are formed using  $b$ , then the subclassification is equal per cent bias reducing in the sense that the per cent of bias in  $x$  remaining after adjustment is the same for each coordinate of  $x$ , namely,  $100\gamma$ , where*

$$\gamma = \frac{\sum_s w_s \int f(b) \{pr\{b|z = 1, b \in I_s\} - pr\{b|z = 0, b \in I_s\}\} db}{\int f(b) \{pr\{b|z = 1\} - pr\{b|z = 0\}\} db},$$

*where the sum is over  $s = 1, \dots, S$ .*

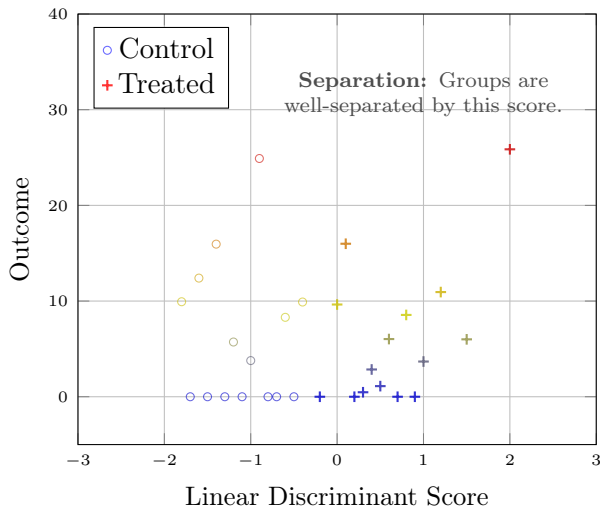
# Propensity Scores and Covariance Adjustment

# Propensity Scores and Covariance Adjustment

- **A Useful Diagnostic Plot:** Plotting the outcomes (or residuals) against the linear discriminant is a useful 2D diagnostic tool to check for problems like non-linear relationships or reliance on extrapolation, which could distort the treatment effect estimate
- **Propensity Score is a Better Alternative:** It is generally more appropriate to plot outcomes against the propensity score  $e(x)$  rather than the linear discriminant
  - ▶ From Corollary 4.3, at any specific value of the propensity score, the observed difference in outcomes between treated and control units is an unbiased estimate of the average treatment effect for that specific group
- Standard covariance adjustment cannot be fully relied upon unless the linear discriminant is highly correlated with the true propensity score

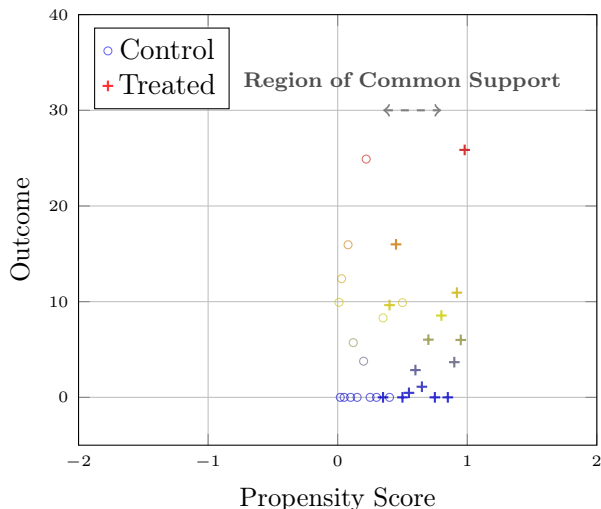
# Propensity Scores and Covariance Adjustment

**Plot 1: Outcome vs. Linear Discriminant Score**



# Propensity Scores and Covariance Adjustment

**Plot 2: Outcome vs. Propensity Score**



See Appendix 2.3 for more details.

# References

# References I



Cochran, W. G. (1968).

The effectiveness of adjustment by subclassification in removing bias in observational studies.

*Biometrics*, pages 295–313.



Cox, D. (1970).

The analysis of binary data.



Ding, P. (2024).

*A first course in causal inference*.

Chapman and Hall/CRC.



LaLonde, R. J. (1986).

Evaluating the econometric evaluations of training programs with experimental data.

*The American economic review*, pages 604–620.

## References II



MA, H. and JM, R. (2020).  
*Causal Inference: What If*.  
Chapman and Hall/CRC.



Rosenbaum, P. R. and Rubin, D. B. (1983).  
The central role of the propensity score in observational studies for  
causal effects.  
*Biometrika*, 70(1):41–55.



Rubin, D. B. (1978).  
Using multivariate matched sampling and regression adjustment to  
control bias in observational studies.  
*ETS Research Bulletin Series*, 1978(2):i–33.

## References III



Rubin, D. B. (1979).

Using multivariate matched sampling and regression adjustment to control bias in observational studies.

*Journal of the American Statistical Association*, 74(366a):318–328.