PROPENSITY SCORE MATCHING
A PRACTICAL TUTORIAL

Cody Chiuzan, PhD
Biostatistics, Epidemiology and Research Design (BERD) Lecture

March 19, 2018

Outline

- Experimental vs Non-Experimental Study
- WHEN and WHY using Propensity Score Matching (PSM)?
- Methods for performing Propensity Score Analysis (PSA)
- Frequently asked questions (FAQs)
- Statistical software for implementing PSM
- Illustrative example(s)

Propensity Score Matching (PSM)

MAIN IDEA: Propensity score methods allow causal inferences from non-experimental (observational) studies.

Observational Studies: data are observed and collected on each subject, with the goal of understanding a cause-effect relationship

- Unlike experiments, the independent variable is not under the control of the researcher

A Causal Relationship must meet the following criteria:

- Temporal order – the cause must precede the effect in time
- The two variables are highly correlated
- The correlation/association is not by coincidence alone

Importance of Randomization

- The randomized clinical trials are now the 'norm' for outcomes evaluation in experimental studies:
  - Participants are randomly assigned to one of the arms: intervention(s) or control
  - Randomization guards against bias such as selection bias, but accidental bias is also minimized
  - Randomization makes the groups 'comparable' with respect to any confounding variables, so that any difference in response can be attributed to the explanatory variable

Options for Non-Experimental Studies

- Stratification
  - Make different categories and group subjects based on common values
  - Problem: limited sample size
- Regression Models
  - Linear regression models with lots of covariates are not always a solution!
  - Problems: model assumptions are violated, linearity is questionable, confounders have different distributions across the intervention groups.
- Matching Methods (PSM)
  - Trying to 'fix' the lack of randomization
  - The modeling of the propensity scores can be conducted independently of the outcome and provide some protection to fit the model

Steps in PSM

1. Identify appropriate data (large sample size)
2. Define the treatment (and control) and outcome
3. Select the covariates of interest
4. Estimate the propensity scores
5. Use the propensity score to 'match' the groups: matching, weighting, stratification, etc.
6. Assess the 'matching' using balance diagnostics methods
7. Run the analysis of the outcome on the propensity score-adjusted sample

Treatment and Outcome

- Define the "treatment or intervention" of interest
  - Examples: "being male" or "heavy drug user"
- Defined in reference to some control, condition of interest (sometimes difficult to define)
  - Examples: "being female" or "no/light drug user"
- Define the potential outcome under treatment/control:
  - Observed outcome if a unit (subject) gets the treatment, $Y(T = 1) = Y(1)$
  - Observed outcome if a unit (subject) gets the control, $Y(T = 0) = Y(0)$
  - E.g., assess the effect of drug use on mental health outcomes

Treatment and Outcome: Setting

- "Treatment" T is measured at a particular point in time
- Covariate(s) X are observed on all individuals, measured (or applicable to) before treatment T
- Outcome(s) Y are also observed on all individuals
  - Ideally have X measured before T measured before Y
- Treatment can be administered at individual-level, but group-level treatments (group as a unit) also work
- We aim to estimate the average causal effects of the treatment, but not the individual effects

Selecting the Covariates

- Ideally, include covariates/variables X that are related to (1) treatment and (2) outcomes
  - Continuous covariates are difficult to match, so categorize
- Trade-off problem:
  - Too many covariates might be difficult to match and increase the variability
  - Excluding important confounders might lead to biased results
- Do not include variables that might be affected by treatment
  - Exclude from matching, but consider later in the outcome analysis

Estimate the Propensity Scores

- Fit a model of "treatment" assignment given the set of covariates (measured at baseline)
  - Most common model: logistic regression (binary treatment)
  - Non-parametric options: classification and regression trees
- In a logistic regression model, the propensity scores are the predicted probabilities of receiving treatment, given the set of covariates:

$$e_i = P(T_i = 1|X_i)$$

- Model predictability and diagnostics are not important in PSM

Create Matched Samples

- Use the predicted probabilities to create balanced matched samples
  - Match a 'treated' subject with one that did not receive treatment
- Most common methods for matching:
  - Nearest neighbor 1:1 (or 1 to many) matching
    - Greedy or optimal matching
  - Inverse probability of treatment weighting
  - Stratification/subclassification on the propensity scores (equally-size groups using quintiles)
- We use statistical software to 'match' the data!

Nearest Neighbor Matching

- 1:1 matching – one 'treated' subject matched with a 'control'
  - If lots of controls, take more than one control for each 'treated' subject
- Match with or without replacement?
- Can use 'exact' matching or a caliper to limit matches within some range of propensity score values
  - 0.25 or 0.5 of the propensity score standard deviation
- Criticism – this matching process throws away data

Assess PSM Matching

- Main goal is to achieve similar distributions in the matched groups
  - Model estimates are not of interest
- Balance diagnostics for covariates:
  - Use standardized differences to quantify differences in means and proportions between the two groups
  - Histograms and cumulative density plots to compare the distributions of continuous covariates used in the model
  - Do not reply solely on statistical significance tests and p-values, but look at multiple measures of balance

Balance Diagnostics: Standardized Differences

- For continuous covariates, the standardized differences are given by:

$$d = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{s_1^2 + s_2^2}{2}}},$$

where $\overline{X_1}, \overline{X_2}$ represent the sample means and s_1^2, s_2^2 denote the sample variances.

- For binary covariates, the standardized differences are given by:

$$d = \frac{(\widehat{p_1} - \widehat{p_2})}{\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1}) + \widehat{p_2}(1-\widehat{p_2})}{2}}},$$

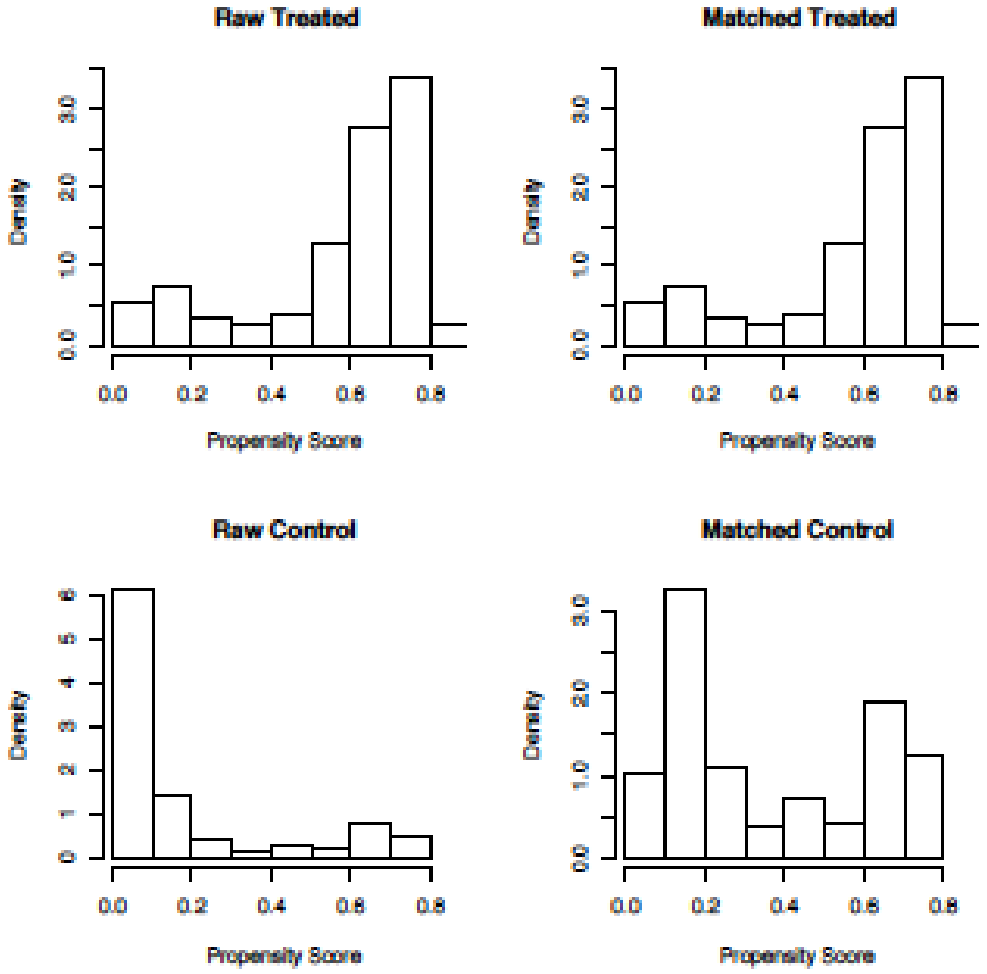
where $\widehat{p_1}, \widehat{p_2}$ represent the estimated treatment and control proportions.

Balance Diagnostics: Standardized Differences

- The standardized differences (effect sizes) can measure the non-overlap between two populations
- For example: a standardized difference d = 0.2 (small effect size) indicates that there is approximately 15% of non-overlap in the two distributions
- Cohen (1988) suggested that:
  - d = 0.2 represents a small effect size
  - d = 0.5 represents a moderate effect size
  - d = 0.8 represents a large effect size
- For balance diagnosis, standardized differences (absolute value) greater than 0.2 indicate group imbalance
  - Those covariates should be included in the model used for generating propensity scores
  - Check balance before and after matching!!

Balance Diagnostics: Graphs

Compare the histograms of propensity scores between the two groups:



Outcome Analysis

- Perform the outcome analysis using matched data
  - Should we use matched data methods? Debatable.

  No, since pairs are not selected on the basis of the outcome.

  Yes, since selected pairs are similar.
- Matching: run regression models on matched samples
- Weighting: run regression models adjusting for weights
- Subclassification: estimate effects within subclasses and then combine, or include subclass category and its interaction with treatment
- 'Adjust' the outcome model using the propensity scores

Outcome Analysis

- Should you include the same covariates in both models (propensity score and outcome)?
  - Yes, if not interested in the coefficient of that covariate in the outcome model
  - Do not include in the propensity score model if you are explicitly interested in that coefficient
- Covariate balance should always be checked!

Missing Data

- Missing covariate values
  - Discard that covariate
  - Create missing data indicator variables for each variable
  - Include the variables and missing data indicators in the propensity score model
- Missing outcomes and treatment/control
  - Nightmare!
  - Multiple imputation methods should be used with caution for missing outcomes and NOT recommended for treatment

FAQs

- Can PSM be used for longitudinal studies?
  - Usually used for cross-sectional data, but longitudinal models can be employed
- What is the unconfoundedness assumption?
  - There are no unobserved differences between the treatment and control groups, given the observed variables
- What if you have controls without similar treated units, and viceversa?
  - Might need to discard some observations or estimate the effect only on a subset
- Matching throws away data – does this affect power?
  - Not exactly as the pairs are similar and an effect (if one exists) might be easier to detect

FAQs

- What is a reasonable sample size?
  - At least 200 subjects in total
  - Remember logistic regression rule: at least 10 events should be observed for every covariate that is entered into the model
- Is PSM always a solution to observational studies?
  - PSM summarizes all covariates into one number (score)
  - One can more easily assess whether observed confounding (only) has been adequately eliminated
  - PSM should not be used to give a 'once and for all' definitive indication of a treatment effect

Statistical Software for PSM

- R and Stata have the most functions for performing propensity score matching:
  - R functions: MatchIt or Optmatch
  - STATA functions: PSmatch2
- SAS has some well-written macros:
  - See reference link for a SAS macro to check covariates balance

PSM Application

- Study Title: "The robotic approach significantly reduces length of stay after colectomy: a propensity score-matched analysis"**

- Design/Data: Retrospective study using patients' information from the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) 2012–2014 datasets.

- Methods: A propensity score-matched analysis to create comparable risk groups in the laparoscopic and robotic colectomy cohorts with respect to demographics, comorbidities, and operative characteristics.

**Ahmed M. Al-Mazrou, Codruta Chiuzan, Ravi P. Kiran, The robotic approach significantly reduces length of stay after colectomy: a propensity score-matched analysis, International Journal of Colorectal Disease, 2017, Volume 32, Number 10, Page 1415.

PSM Application

- PSM Details: Standardized differences were calculated to compare patients' features before and after matching with imbalance being defined as an absolute value greater than 0.10 (small effect size). Matching was performed using the nearest neighbor algorithm with a caliper distance of 0.0001.

- All perioperative variables with > 90% missing data were excluded from the analysis (e.g., alcohol use, intraoperative complications).

PSM Application

Table 1 Propensity score matching (PSM) of the two groups

Patients' characteristics	Unmatched comparisons				Matched comparisons			
	Laparoscopic $N = 34,130$	Robotic $N = 1709$	p value	Standardized difference[a]	Laparoscopic $N = 1241$	Robotic $N = 1241$	p value	Standardized difference[a]
Age, mean ± SD	60.5 ± 15.1	59.6 ± 13.4	0.04	0.06	60.7 ± 14.1	60.1 ± 13.7	0.07	0.08
Sex (male)	16,473 (48.3)	847 (49.6)	0.3	−0.17	630 (50.8)	595 (48.0)	0.2	0.04
Race (white)	27,079 (86.6)	1394 (86.2)	0.8	0.01	976 (86.7)	1008 (85.7)	0.5	0.03
Body mass index								
Underweight	778 (2.3)	33 (1.9)			19 (1.5)	21 (1.7)		
Normal	9897 (29)	453 (26.6)			318 (25.8)	354 (28.6)		
Overweight	11,510 (33.7)	583 (34.2)			457 (37.0)	409 (33.0)		
Obese	11,770 (34.5)	636 (37.2)	0.04	0.07	441 (35.7)	454 (36.7)	0.2	0.09
Smoking history	5440 (15.9)	261 (15.3)	0.5	0.02	235 (18.9)	188 (15.2)	0.02	−0.01
Functional status								
Independent	33,442 (98.4)	1686 (99.1)			1212 (97.7)	1218 (98.8)		
Partially dependent	470 (1.4)	13 (0.8)			22 (1.8)	13 (1.1)		
Totally dependent	81 (0.2)	2 (0.1)	0.06	0.07	6 (0.5)	2 (0.2)	0.12	0.08
ASA classification								
I	1056 (3.1)	45 (2.6)			31 (2.5)	34 (2.7)		
II	17,615 (51.7)	907 (53.1)			649 (52.3)	639 (51.5)		
III	14,307 (41.9)	725 (42.4)			518 (41.7)	544 (43.8)		
IV	1122 (3.3)	32 (1.9)			43 (3.5)	24 (1.9)		
V	4 (0.1)	0 (0)	0.02	0.10	0 (0)	0 (0)	0.1	0.09
Wound classification								
Clean	294 (0.9)	15 (0.9)			6 (0.5)	12 (0.9)		
Clean/contaminated	28,406 (83.2)	1443 (84.4)			1016 (81.9)	1039 (83.7)		
Contaminated	3576 (10.5)	173 (10.1)			132 (10.6)	129 (10.4)		
Dirty/infected	1854 (5.4)	78 (4.6)	0.4	0.04	87 (7.0)	61 (4.9)	0.08	0.10
Underlying diagnosis								
Colon cancer	13,663 (40.0)	761 (44.5)			572 (46.1)	543 (43.8)		
Diverticular disease	8251 (24.2)	493 (28.9)			305 (24.6)	361 (29.1)		
Inflammatory bowel disease	2054 (6.0)	35 (2.1)			35 (2.8)	30 (2.4)		
Other diagnosis	10,162 (29.8)	420 (24.6)	<0.001	0.25	329 (26.5)	307 (24.7)	0.09	0.10

PSM Application

Table 2. Thirty day postoperative complications. Comparisons performed on matched data using two-sample t-test or Fisher's Exact test.

Variable	Laparoscopic $N = 1241$	Robotic $N = 1241$	p value
Surgical site infection (Superficial, deep, or organ space)	98 (7.9)	96 (7.7)	0.9
Wound disruption	5 (0.4)	0	0.06
Sepsis and septic shock	50 (4.0)	29 (2.3)	0.02
Postoperative ileus	131 (10.6)	122 (9.8)	0.6
Anastomotic leak	35 (2.8)	38 (3.1)	0.7
Pneumonia	20 (1.6)	9 (0.7)	0.4
Cardiac arrest	4 (0.3)	1 (0.1)	0.4
Myocardial infarction	8 (0.6)	5 (0.4)	0.6
Acute renal failure	3 (0.2)	8 (0.6)	0.2
Urinary tract infection	26 (2.1)	16 (1.3)	0.1
Deep vein thrombosis /thrombophlebitis	14 (1.1)	9 (0.7)	0.6
Pulmonary embolism	5 (0.4)	7 (0.6)	0.8
Unplanned reoperation	55 (4.4)	48 (3.9)	0.5
Unplanned intubation	13 (1.1)	9 (0.7)	0.4
Cerebrovascular accident	2 (0.2)	0 (0)	0.5
30-day mortality	10 (0.8)	7 (0.6)	0.6

Selected References

1. Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41-55.
2. Cohen, J. (1988). Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd ed). Lawrence Erlbaum Associates Publishers: Hillsdale, NJ.
3. Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. Annals of Internal Medicine 127, 757-763.
4. Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. Pharmacoepidemiology and drug safety 13, 855-857.
5. Morgan, S.L, & Winship, C. (2007). Counterfactuals and Causal Inference: Methods and Principles for Social Research. New York: Cambridge University Press.
6. Peter C. Austin. (2009) Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statist. Med. 2009; 28:3083–3107
7. Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer.
8. Guo, S. & Fraser, W.M. (2014). Propensity Score Analysis: Statistical Methods and Applications, Second Edition. Thousand Oaks, CA: Sage Publications.

Useful links:

SAS macro for checking balance: http://support.sas.com/resources/papers/proceedings12/335-2012.pdf

Elizabeth Stuart, Summer Institute course on PSM:

http://www.jhsph.edu/dept/mh/summer_institute/courses.html

Thank you

Cody Chiuzan: cc3780@cumc.columbia.edu

BERD EDU link:
http://irvinginstitute.columbia.edu/resources/biostat_educational_initiatives.html