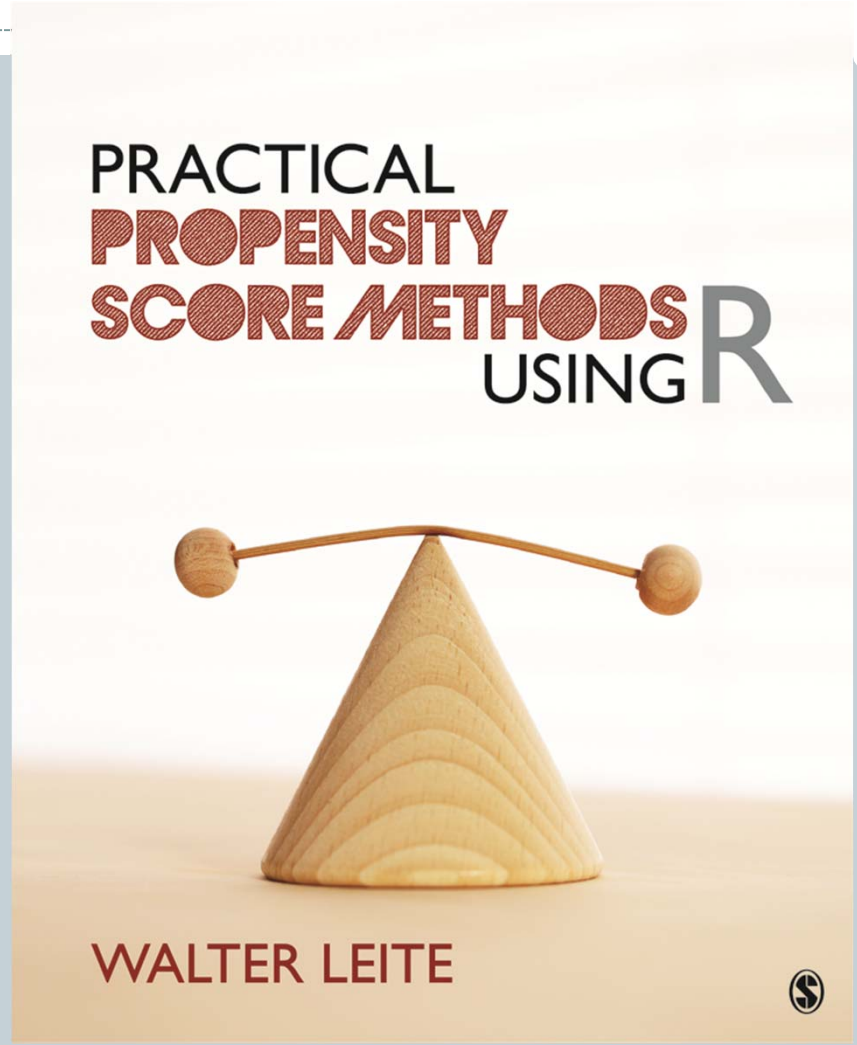
Overview of Propensity Score Analysis

WALTER L. LEITE, PH.D.

RESEARCH AND EVALUATION METHODOLOGY PROGRAM
UNIVERSITY OF FLORIDA

WALTER.LEITE@COE.UFL.EDU
PHONE: +1-352-273-4302
WEBSITE: HTTP://EDUCATION.UFL.EDU/LEITE/

Source of examples and R code:



Leite, W. L. (2016). Practical Propensity Score Methods Using R. Thousand Oaks, CA: Sage Publications.

https://us.sagepub.com/en-us/nam/practical-propensity-score-methods-using-r/book241054

Experimental Designs

- The objective is to estimate the effect of a condition (e.g. treatment, program, intervention) on outcomes.
- Participants are randomly assigned to conditions.
- Participants DO NOT need to be randomly sampled.
- Conditions are manipulated;

Quasi-experimental/Observational Designs

- The objective is to estimate the effect of a condition (e.g. treatment, program, intervention) on outcomes.
- No random assignment to the condition of interest is possible.
- Manipulation of the condition may be possible.

Description of Example

Objective: Estimate the effect of high school student participation on a career academy on future income.

Treatment: Career academies are programs within high schools that integrate academic preparation and workplace experiences through a career-focused curriculum.

Dataset: Educational Longitudinal Study (ELS) of 2002 and 2006.

Example – Effect of Career Academy on Income

- Research question: Does participating in a career academy in high school affect future income?
- Sampling design of ELS: two-stage stratified sampling method where schools were sampled with probability proportional to size (PPS) sampling, with oversampling of Asian and Hispanic students.
- Treatment measure: "Have you ever been in any of the following kinds of courses or programs in high school?", where option *k* is "Career Academy"
- Treated proportion: 1,371 (8.5%) students participated, 14,826 (91.5%) who did not.

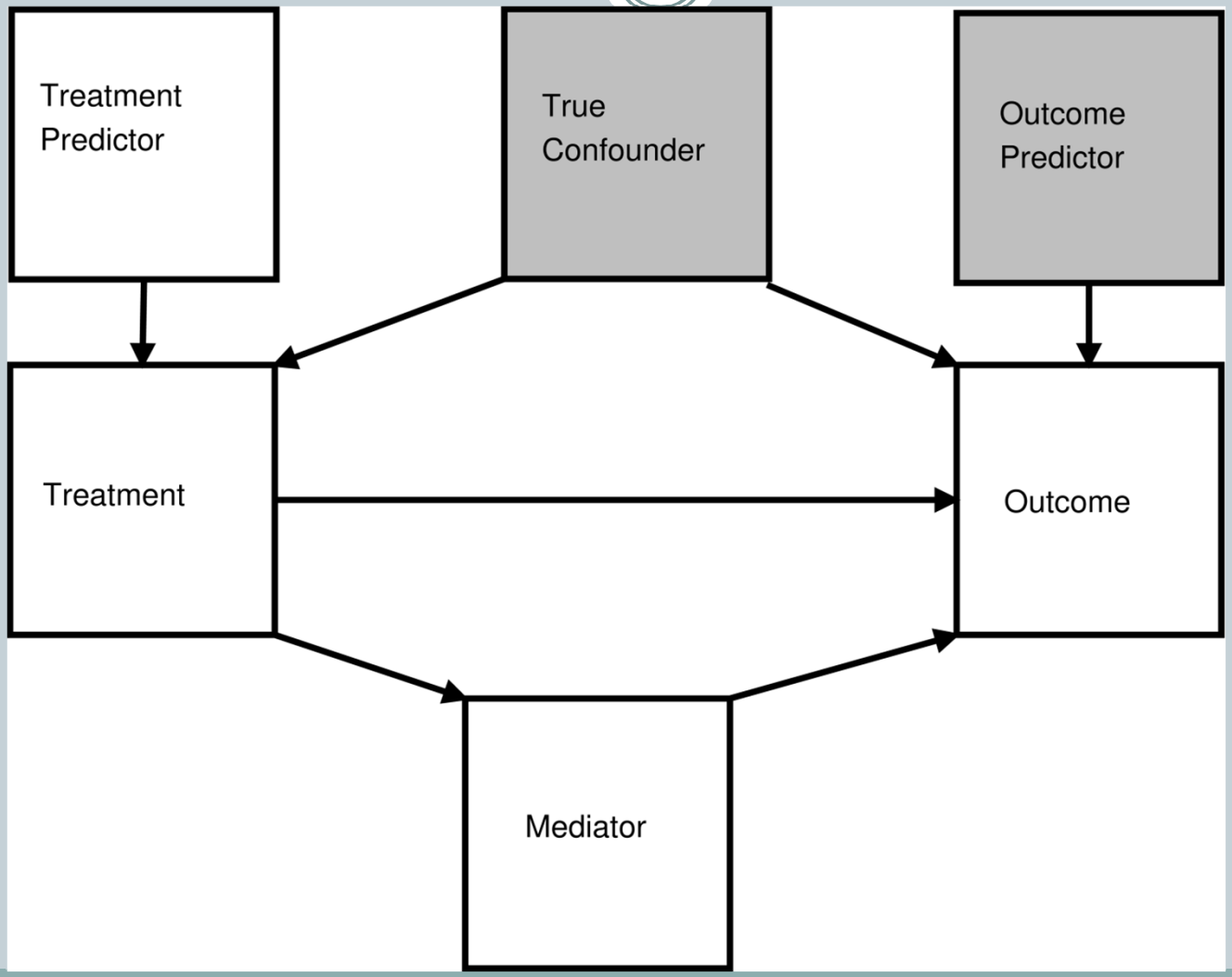
Internal Validity of Research Designs

- It is the validity of statements that can be made about whether there is a causal relationship from one variable to another in the form in which the variables were manipulated or measured.

- Internal validity is strongest when the study's design effectively controls possible sources of error (confounding variables).

- It is difficult to find clear evidence of internal validity on non-experimental designs, because there are many factors not under the control of the researchers.

Confounding variables



Example potential confounders of the relationship between career academy participation and future income

Variable Name in ELS Data	Variable Description
bypared	Parents' highest level of education
byhomlit	Home literacy resources
byriskfc	Number of academic risk factors in 10th grade
bystexp	How far in school student thinks will get—composite
BYS27I	Parents expect success in school
byurban	*Urbanicity of school*
byregion	*Geographic region of school*

Rubin's Causal Model

- All individuals in the population have potential outcomes and in the presence of the treatment and control conditions.
- The outcomes of the treatment group *T* are only observed in the presence of the treatment condition *t*;
- The outcomes of the control group *C* are only observed in the presence of the control condition *c*.

	Potential Income if participant	Potential Income if non-participant
Participants	$Y_i^1 \mid Z_i = 1$	$Y_i^0 \mid Z_i = 1$
Non – participants	$Y_i^1 \mid Z_i = 0$	$Y_i^0 \mid Z_i = 0$

Average Treatment Effect (ATE)

- It is the difference between the expected values of the outcomes of the individuals in the treated and untreated conditions.

$$ATE = E[Y_i^1] - E[Y_i^0]$$

Conventional estimation of the ATE

Sample mean of treatment group

Sample mean of control group

$$\hat{\hat{\gamma}} = \hat{\bar{Y}}^{1}_{i \in T} - \hat{\bar{Y}}^{0}_{i \in C}$$

- In randomized designs, this estimator is valid because:

$$E\left[Y_i^1 \mid Z = 1\right] = E\left[Y_i^1 \mid Z = 0\right]$$

$$E\left[Y^0 \mid Z = 0\right] = E\left[Y^0 \mid Z = 1\right]$$

Assumptions required for valid treatment effect estimates

- Strong ignorability of treatment assignment
- Stable Unit Treatment Value Assumption (SUTVA)
- Full treatment adherence (compliance).
- No attrition from post-test measurement.

Strong Ignorability of Treatment Assignment

- The treatment assignment is independent of the potential outcome distributions, given observed covariates.

$$\left[Y_i^1, Y_i^0\right] \perp Z \mid \mathrm{X}$$

- It also requires that for every value of Z, the probability of treatment assignment is neither zero nor one

$$0 < P(\mathrm{Z} = 1 \mid \mathrm{X}) < 1$$

The Stable Unit Treatment Value Assumption (SUTVA)

- SUTVA (Rubin, 1986):there is a unique potential outcome for each treatment version.
  - SUTVA requires that the distribution of potential outcomes for one individual is independent of the potential treatment status of another individual and there are no unrepresented versions of the treatment.

Average Treatment Effect for the Treated

- It is the difference between the expected value of the outcomes of the treated individuals and the expected value of the potential outcomes of the treated individuals.

$$ATT = E(Y_i^1 \mid Z_i = 1) - E(Y_i^0 \mid Z_i = 1)$$

Average Treatment Effect for the Control

- It is the difference between the expected value of the outcomes of the control individuals and the expected value of the potential outcomes of the control individuals.

$$ATC = E(Y_i^1 \mid Z_i = 0) - E(Y_i^0 \mid Z_i = 0)$$

Propensity scores

- The propensity score is defined as a conditional probability of treatment assignment, given observed covariates (Rosembaum & Rubin, 1983);

$$e(x_i) = P(Z_i = 1 | \mathbf{X})$$

- If the propensity score was correctly specified, balancing the treatment and control groups with respect to propensity score also balances them with respect to distributions of covariates;

Propensity Scores (PS)

- The propensity score reduces all the information in the predictors to one number
  - This can make it easier to do matching or stratifying when there are multiple matching variables available.
- In a randomized experiment, the true propensity score is .50 for each person
- In a quasi-experiment, the true propensity score is unknown.

Strong Ignorability and Propensity Scores

- If treatment selection is *strongly ignorable* given an observed set of individual covariates **X**, then it is also strongly ignorable when these individual covariates are combined into a propensity score $e(\mathbf{X})$, (proved Rosenbaum & Rubin 1983):

$$if\ (Y^0, Y^1) \perp Z \mid X \quad then$$

$$(Y^0, Y^1) \perp Z \mid e(X) \text{ and } Z \perp X \mid e(X)\ with$$

$$e(x_i) = P(Z_i = 1 \mid X_i = x_i)\ \ and$$

$$0 < e(X) < 1$$

Advantages of Propensity Score Methods over Conditioning on Covariates

- Smaller models where fewer parameters are estimated;
- Linearity assumptions are not made;
- Problem of differences in distributions of covariates for treatment and control groups is eliminated.

Steps of Propensity Score Analysis

Step	Objective
Data preparation	Obtain complete data that is ready for analysis
Propensity score estimation	Obtain propensity scores for treated and untreated individuals
Propensity score method implementation	Implement a strategy to balance treated and untreated covariate distributions using propensity scores
Covariate balance evaluation	Determine the degree to which balance of covariate distributions between treated and untreated was achieved
Treatment effect estimation	Estimate the treatment effect and its standard error
Sensitivity analysis	Determine how strong the effect of an omitted covariate would have to be for the significance test of the treatment effect to change

What covariates to include in the propensity score model

- Including variables related to exposure and the outcome in the PS model (True Confounders) decreases bias and variance.
- Including variables related to the outcome but not to exposure in the PS model does not affect bias but decreases variance.
- Including variables related to exposure but not the outcome does not affect bias but increases variance.

Estimation of Propensity Scores

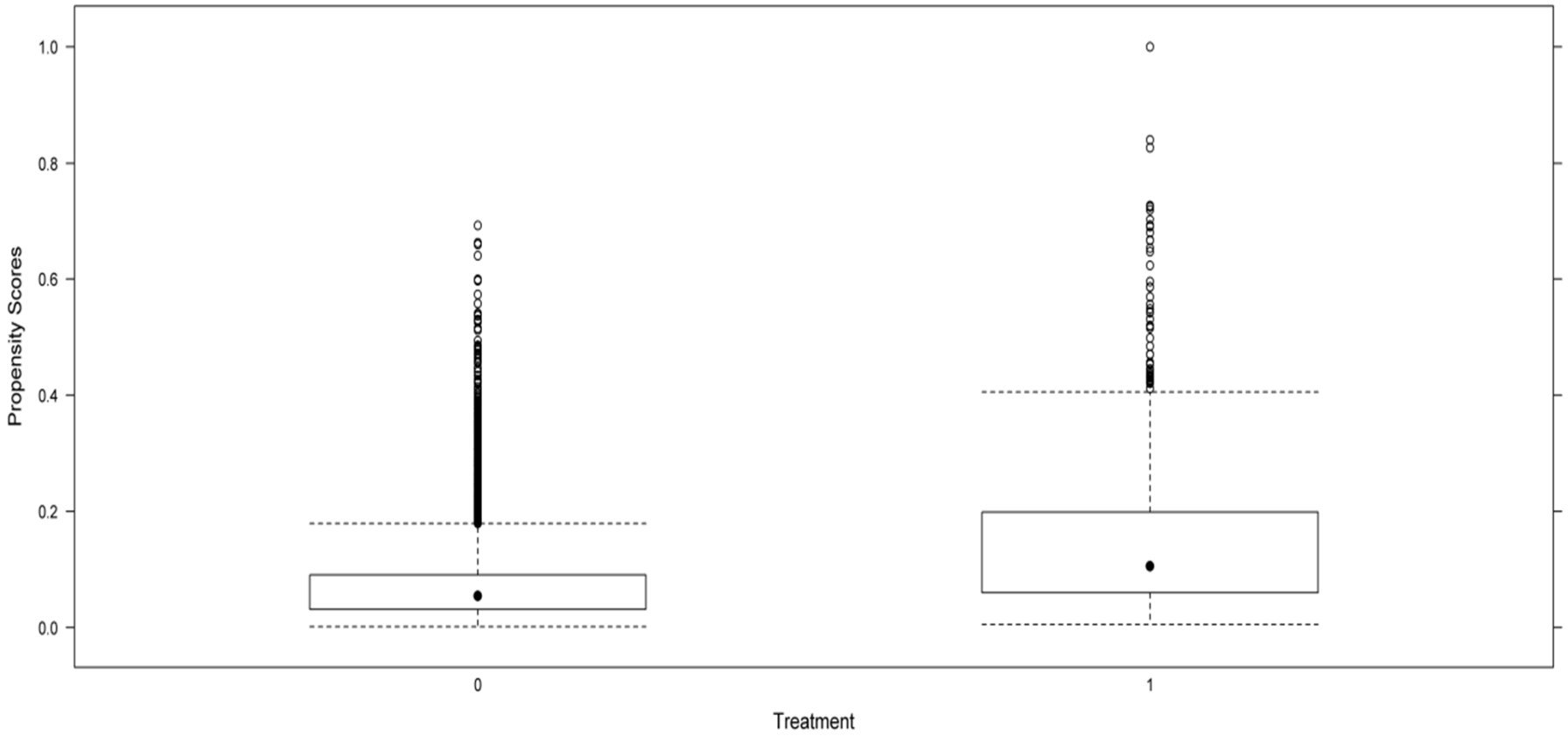
- Different methods for estimating PS can be used:
  - *Statistical models*: logistic regression, probit regression
  - *Statistical learning algorithms* (data mining methods):
    classification trees, boosting, bagging, random forests

R functions for estimating propensity scores

Package	Function	Method of propensity score estimation
stats	glm	Logistic regression
lme4	glmer	Multilevel logistic regression with clustered data
survey	svyglm	Logistic regression with complex survey data
party	cforest	Random forests
twang	ps	Generalized boosted modeling

Common support

- *Common support: The area of the propensity score distribution where values exist for both groups.*



Propensity Score Method Implementation

- Propensity scores methods:
  - Weighting
  - Stratification
  - Matching

Weighting

- If the treatment assignment is random, the distribution of a covariate for the treatment group will be similar to the distribution for the control group.
- In experimental studies, the treated and untreated outcomes have equal probabilities of being sampled ($P(C=c|X)=0.5$).
- In quasi-experimental designs, weighting adjusts the distribution of covariates so they are similar across groups.
- In quasi-experimental studies, the probability of sampling the treated outcome is the propensity score $P(C=1|X)$, and the probability of sampling the untreated outcome is $1 - P(C=1|X)$

Propensity Score Stratification

- If the sample is divided into strata such that all units within a stratum have the same e(x) and at least one unit in the stratum is exposed to each condition, the expected difference in treatment means within each stratum is the ATE at that e(x), and the weighted average of the stratum differences is an unbiased estimated of the ATE.

- Stratification can be considered as a coerce form of matching, or as a non-parametric form of weighting.

Approaches to propensity score stratification

- Divide propensity score distribution of treated and untreated into strata, typically five.

Option 1) pool strata-specific treatment effects;

Option 2) obtain a marginal treatment effect across strata (marginal mean weighting through stratification).

Propensity Scores for matching

- It is advantageous to match on the linear propensity score (i.e., the logit of the propensity score) rather than the propensity score itself, because it avoids compression around zero and one.

$$\log(\mathrm{e}(\mathrm{X})) = \log\left(\frac{e(X)}{1-e(X)}\right)$$

Matching methods taxonomy

Replacement:

- Matching with replacement
- Matching without replacement

Algorithms:

- Greedy matching:
- Optimal matching
- Genetic matching

Ratio:

- *Pair matching (1 to 1)*
- 1 to k
- variable ratio
- Full

Weights for treatment effect estimation

- One-to-k or variable ratio: weighs are the inverse of the total number of matches unit received.

- Matching with replacement: weights for each untreated unit are summed across the multiple matched groups it was included in. Then, weights of the matched cases are multiplied by the ratio of the total number of matched units and total number of treated units.

R packages for propensity score method implementation

Package	Function	Type of propensity score method
MatchIt	matchit	Stratification, greedy matching, and interface for genetic, optimal, and full matching
Matching	GenMatch Match	Obtain weights for genetic matching Genetic matching and greedy matching
twang	ps	Calculate propensity score weights

Evaluation of covariate balance

- *Balance* refers to the equivalence of the treatment and control groups' joint distribution on all observed covariates:

- In practice, balance on the joint distribution is hard to check (curse of dimensionality), therefore the focus is on each covariates distribution separately;

- Covariate balance can be checked by visual, descriptive or inferential methods.

Checking for Balance

- *Visual inspection*: Comparison of histograms or kernel density estimates; QQ-plots

- *Descriptive criteria* that compare the covariate distribution of the treatment and control group:

- *Standardized mean difference*

$$d = (\bar{x}_t - \bar{x}_c) / \sqrt{(s_t^2 + s_c^2)/2} \qquad \text{or} \qquad d = (\bar{x}_t - \bar{x}_c) / s_t$$

Common balance criteria:

$|d| < 0.1$ or

$|d| < 0.05$ or $|d| < 0.25$ + covariate adjustment

What if adequate common support and covariate balance cannot be achieved?

- Balance could be improved by change propensity score estimation method or model;
- If imbalance is not too severe, *additional covariance adjustment* may remove the residual bias;
- If balance is still not achieved, this indicates that groups are too *heterogeneous* for estimating a causal treatment effect;

Example Comparison of Absolute Standardized Effect Sizes

Propensity Score Estimation	Minimum	Mean	Maximum
Logistic regression	0.000	0.010	0.057
Random forest	0.000	0.035	0.157
GBM	0.000	0.033	0.125

Estimators of the treatment effect

Horvitz and Thompson Estimator (for weighting)

Abadie and Imbens Estimators (for matching)

Regression Estimation

R packages for treatment effect estimation

Package	Function	Type of propensity score method
survey	svymean	Estimate weighted means
	Svycontrast	Estimate ATE, ATT with Horvitz and Thompson estimator
survey	svyglm	Fit weighted regression
Matching	Match	Estimate ATT and ATE with Abadie and Imbens estimator

Sensitivity Analysis

- It asks the question: Would the conclusion change if an important covariate was omitted?

- Goals:
  - Determine how strong the effect of an omitted covariate would have to be for the significance test of the treatment effect to change

  - Determine the degree of robustness of treatment effects to hidden bias, which is the part of the selection bias due to omitted confounders.

R package for sensitivity analysis

Package	Function	Objective
treatSens	treatSens	Obtain a sensitivity analysis with Carnegie, Harada and Hill's (2016) method
rbounds	*psens*	Obtain a sensitivity analysis with Rosenbaum's (1987) method