An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies

Peter Austin
University of Toronto

Multivariate Behavioral Research (2011): 46(3), 399-424.

Insper
Journal Club in Causal Inference

Motivation

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Table 5.1: Death rates per 1,000 person-years (Cochran 1968)

Motivation

Smoking group	Canada	British	US
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Table 5.2: Mean ages, years (Cochran 1968).

RCTs vs. Observational Studies

- Let us review the potential outcomes framework: $Y_i(0)$ and $Y_i(1)$.
- We can only observe one outcome for unit i:

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

  where Z is an indicator for treatment.

- Remember the Average Treatment Effect is:

$$ATE = E[Y_i(1) - Y_i(0)]$$

RCTs vs. Observational Studies

- Because of randomization, the ATT is equal to the ATE, or:

$$E[Y(1) - Y(0)] = E[Y(1) - Y(0)|Z = 1]$$

since $Z \perp\!\!\!\perp Y(0), Y(1)$.

- Thus, an *unbiased estimate* of the ATE can be directly computed from the study data.

RCTs vs. Observational Studies

- With observational data, in general:

$$E[Y(1)|Z = 1] \neq E[Y(1)]$$

  and similarly for the control.

- Thus, an *unbiased estimate* of the ATE **cannot** be obtained by directly comparing outcomes between the two groups.

Propensity Score and Propensity Score Methods

- Rosenbaum and Rubin (1983) defined the propensity score as the **probability of treatment assignment conditional on observed covariates**.

$$e_i = Pr(Z_i = 1|\mathbf{X}_i)$$

- **Balancing score**: conditional on the propensity score, the distribution of measured baseline covariates is similar between treated and untreated subjects.

Propensity Score and Propensity Score Methods

- In RCTs, the propensity score is known.
  - Defined by the study design.
- In observational studies we need to estimate it.
- Generally estimated using logistic regression.
  - Treatment status regressed on observed baseline characteristics.
  - Predicted probability of treatment derived from the fitted model.

Propensity Score and Propensity Score Methods

- 4 different methods for removing confounding with PS:
  - Propensity score matching.
  - Stratification.
  - Inverse probability of treatment weighting (IPTW).
  - Covariate adjustment.

- "No unmeasured confounders" assumption (Rosenbaum and Rubin, 1983):
  (a) $Y(1), Y(0) \perp\!\!\!\perp Z|X$
  (b) $0 < P(Z = 1|X) < 1$
- Conditioning on the propensity score $\implies$ unbiased estimates of the ATE.

Propensity Score Matching

- Match sets of treated and untreated who share a similar value of the propensity score.
- Treatment effect from comparing outcomes between subjects in the matched sample.
- Different types of matching (Gu and Rosenbaum, 1993).
  - one-to-one (most common).
  - many-to-one
  - full matching.
- One-to-one matching types:
  (i) Matching with replacement vs. without replacement.
  (ii) Greedy vs. optimal
- **Nearest neighbor:** Untreated subject with the closest propensity score to that of the treated.

Stratification on the Propensity Score

- Stratify subjects into mutually exclusive subsets based on their estimated propensity score.
- Subjects are ranked according to their propensity score.
- Common approach: divide subjects into five equal-size groups using quintiles of the estimated propensity score.
- Stratum-specific estimates of treatment effect can then be pooled across stratum to estimate an overall treatment effect.

Inverse Probability of Treatment Weighting Using the Propensity Score

- Weights to create a synthetic sample where the distribution of measured baseline covariates is independent of treatment assignment.

- If Z_i is the indicator for treatment and e_i denotes the propensity score, then:

$$w_i = \frac{Z_i}{e_i} + \frac{(1 - Z_i)}{1 - e_i}$$

- An estimate of the ATE is:

$$\frac{1}{n} \sum_{i=1}^{n} w_i Y_i$$

Covariate Adjustment Using the Propensity Score

- The outcome variable Y_i is regressed on Z_i and e_i.
- ATE is determined by regression coefficient from fitted model.
- For a linear model, ATE is an adjusted difference in means.
- Caveat: We need to assume that the model has been correctly specified.

Balance Diagnostics

- How do we know the propensity score model has been correctly specified?
- True propensity score is a **balancing score**: in strata of subjects that have the same ps, the distribution of measured covariates will be the same between treated and untreated.
- For a continuous covariate, we can use a standardized difference:

$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

- Higher order moments of covariates should also be compared (Austin, 2009; Ho, Imai, King and Stuart, 2007).

Balance Diagnostics

- Rosenbaum and Rubin (1984): iterative approach to specifying a PS model. One can modify the model by:
  - Including additional covariates,
  - Adding interactions.
  - Modeling the relationship between them.
- Rubin (2001): set of criteria based on comparing the distribution of the PS in a sample to determine if regression adjustment will eliminate bias.
- Statistical significance testing (but some caveats).
  - Significance may be confounded with sample size.
  - Balance is a property of a particular sample.

Variable Selection

- Lack of consensus in the literature to which variables to include in the PS model.
- Some theoretical arguments in favor of the inclusion of only variables that affect treatment assignment.
- (Austin, Grootendorst and Anderson, 2007): benefits to including only potential or true confounders.
- Look at published literature for guidance.
- Only include variables that are measured at *baseline* and not post-baseline covariates that may be modified by treatment.

Propensity Score versus Regression Adjustment

Conditional versus Marginal Estimates of Treatment Effect

- Conditional Treatment Effect: average effect on the individual.
- Marginal Treatment Effect: average effect on the population.
- Measure of treatment effect is *collapsible* if conditional and marginal effects coincide.
- PS models allow for estimation of the marginal effect (Rosenbaum, 2005).
- Marginal and conditional estimates coincide if:
  (a) no unmeasured confounding.
  (b) outcome is continuous.
  (c) the true outcome model is known.
- If outcome is binary this won't necessarily hold.

Propensity Score versus Regression Adjustment

Practical Concerns

- So why choose propensity score over regression?
- Simpler to determine if the model is right.
- Separate the design from the analysis.
  - Cannot modify the model to get your results.
- Increased flexibility when outcomes are rare and treatment is common.
- One can examine the degree of overlap in the distribution of baseline covariates and decide which is the best course of action.

Discussion

- Propensity is better than regression adjustment to remove bias from confounding.
- In economics it has gained some use with Dehejia and Wahba (2002).
- However, economists usually prefer RDD or Diff-in-Diff methods than Propensity Score.
- The main concern is with the Conditional Independence Assumption (Cunningham, 2020):

$$Y(1), Y(0) \perp\!\!\!\perp Z|X$$

- Economists are usually more worried with selection on unobservables.

Thank You!

ivangp@al.insper.edu.br