Propensity Score Analysis

Hideo Yasunaga

Department of Clinical Epidemiology and Health Economics, School of Public Health, The University of Tokyo

ABSTRACT

Propensity score is defined as the probability of each individual being assigned to the treatment group. Propensity score analysis has recently become the sine qua non of comparative effectiveness studies using retrospective observational data. The present report provides useful information on how to use propensity score analysis as a tool for estimating treatment effects with observational data, including (i) assumptions for propensity score analysis, (ii) how to estimate propensity scores and evaluate propensity score distribution, and (iii) four methods of using propensity scores to control covariates: matching, adjustment, stratification, and inverse probability of treatment weighting.

KEY WORDS

propensity score analysis, propensity score matching, inverse probability of treatment weighting

1. INTRODUCTION

There has been increasing interest in using observational studies to estimate the association between treatments and outcomes. However, observational studies are limited because baseline characteristics often differ significantly between treated and untreated patients, resulting in failure of the robust comparison.

Historically, regression adjustment has been utilized to account for such differences. Recently, various advanced methods for controlling confounding in observational studies have been developed and are being used, including propensity score analysis.

Propensity score analysis is increasingly applied in clinical epidemiology. However, some researchers may be unfamiliar with how to apply propensity score analysis and its limitations.

The aims of the present article are to introduce clinicians and researchers to the concept of propensity scores and to provide essential information on how to use propensity score analysis as a useful tool for analyzing observational data to estimate treatment effects.

2. RANDOMIZED CONTROLLED TRIALS

In comparative effectiveness studies, it is essential to ensure that the groups are comparable and to avoid “comparing apples and oranges”. For instance, assume there is a study comparing outcomes between surgery and conservative therapy for infective endocarditis. Patients with severely deteriorated cardiac function will not receive surgery; surgeons may select easy-to-treat patients who are expected to have better outcomes through surgery. Due to differences in patient backgrounds between the surgical and non-surgical patients, surgery would be reported for better outcomes, which merely reflect differences in patient severity rather than true difference in treatment effect. This kind of bias is generally called “confounding by indication”.

Randomization is the most effective method to address inter-group comparability and to balance covariates between groups. That is, random treatment assignment ensures that unmeasured confounders as well as measured confounders are equally distributed between groups.

For this reason, randomized clinical trials (RCTs) have high internal validity and are called ‘gold standard’ [1]. In
RC Ts, random treatment assignment allows investigators to establish causal relationships between treatment and outcome and to obtain an unbiased assessment of the average treatment effect (ATE) by directly comparing outcomes between treated and untreated patients. However, RCTs are not always feasible due to ethical considerations and high costs.

**3. OVERVIEW OF PROPENSITY SCORE ANALYSIS**

In observational studies, several alternative methods are used to balance covariates between groups. Multiple regression analysis can partly account for confounding variables, but cannot always address the issue of intergroup comparability. Furthermore, when a lot of covariates exist, multivariable regression analysis may not have sufficient power to demonstrate a significant association between treatment and outcome and may provide misleading results due to overfitting.

Propensity score analysis can effectively adjust for confounders and offer investigators the ability to balance patient backgrounds between two groups across all putative risk factors. Propensity score analysis was first introduced by Rosenbaum and Rubin in 1983 [2].

This chapter explains the overview of propensity score analysis including (i) assumptions of propensity score analysis, (ii) estimating propensity score, and (iii) evaluating propensity score distribution.

**3-1. Assumptions of Propensity Score Analysis**

Applying propensity score has several assumptions. One is the "strongly ignorable treatment assignment" assumption. That is, if treatment assignment is strongly ignorable and there is no unmeasured confounders, conditioning on the propensity score can result in unbiased estimates of average treatment effects. In fact, this assumption is untestable and hard to realize; confounding due to unmeasured covariates cannot be completely avoided. Thus, including as many potential confounders as the propensity score model as possible are recommended [3].

Another assumption is the stable unit treatment value assumption (SUTVA). That is, the treatment effect for one individual should be unaffected by the treatment status of another.

**3-2. Estimating Propensity Scores**

Propensity score is defined as the probability of each individual being assigned to the treatment group. Among patients with the same propensity score, the distribution of measured baseline covariates will be the same.

Propensity scores are generally estimated using a multivariable logistic regression model, where treatment status is regressed against potential confounders. The model can incorporate a large number of background covariates to estimate a single number of propensity scores, ranging from 0 to 1 for each patient.

C-statistics (the area under the receiver operating characteristic curve of the logistic regression model) is used to assess the ability to correctly differentiate between the two groups.

Several studies tried to use various methods, other than logistic regression analysis, to estimate propensity scores, such as random forests [4] and neural networks [5]. However, I recommend researchers use only logistic regression analysis, because other methods are not popular and do not have a big advantage.

Variables in the propensity score model must be present prior to the treatment assignment. Variables that were present after the treatment assignment should not be included in the propensity score model.

There is some controversy on variable selection for constructing the propensity score model [6]. The following variables should possibly be included: (i) all measured covariates, (ii) measured covariates associated with treatment assignment, (iii) measured covariates associated with outcome, and (iv) measured covariates associated with both treatment assignment and outcome (true confounders). However, it is difficult to clarify whether each covariate should be classified as a true confounder. Including all measured covariates is the simplest way that can satisfy ignorable treatment assignment, and mimic a randomized controlled trial by removing all sources of incomparability between the groups [7]. Precision of differences in outcomes between the groups may be reduced, when we include only variables related to treatment assignment but not to outcomes [8].

**3-3. Evaluating Propensity Score Distribution**

Evaluating the distribution of propensity scores can assess their overlap between the two groups. Very different distributions of propensity scores and little overlap indicate that the two groups are not comparable.

For evaluating propensity score distribution, the standardized difference can be used to compare the mean of continuous and binary variables between the treatment and control groups.

For a continuous covariate, the standardized difference is defined as:
where $Mt$ and $Mc$ denote the mean of the covariate, whereas $Vt$ and $Vc$ denote the variance of the covariate in treated and control groups, respectively.

For dichotomous variables, the standardized difference is defined as:

$$\text{Standardized difference} = \frac{Pt - Pc}{\sqrt{\frac{Pt(1 - Pt)}{2} + \frac{Pc(1 - Pc)}{2}}}$$

where $Pt$ and $Pc$ denote the proportions of the dichotomous variable in treated and control groups, respectively.

The standardized difference is unaffected by sample size. A standardized difference <0.1 indicates a negligible difference in the mean or proportion of a covariate between the treatment and control groups [9].

Showing $p$ values with t tests or chi-square tests is not recommended to check the balance of covariates, because failure to reject the null hypothesis does not guarantee successful balance of covariates between the groups.

### 4. METHODS OF USING PROPENSITY SCORES

There are four methods of using propensity scores to control covariates: matching, adjustment, stratification, and inverse probability of treatment weighting (IPTW).

#### 4-1. Propensity Score Matching

**1:1 or 1:n matching**

Propensity score matching pairs each patient in the treatment group with a patient in the control group who shares a similar propensity score value. Propensity score matching can estimate the average treatment effect on the treated (ATT) [10].

When there is partial overlap in the distribution of propensity scores between groups, a small portion of the entire sample is selected for the final analysis, and generalizability of the results to the whole population may be limited.

One-to-one matching, where pairs of treated and untreated patients have similar propensity score values, is most commonly implemented [11]. However, when the control group includes many more patients, 1:n matching approach can also be used.

*With or without replacement*

Matching can be implemented "with replacement" or "without replacement". Matching with replacement can be applied when the numbers of control patients are limited [12]. In fact, propensity score matching is generally performed “without replacement”. That is, once a control patient is selected to be matched to a given treated patient, the control patient is no longer selected. Each control patient is included only once at most.

**Nearest neighbor matching and optimal matching**

Matching algorithms include two types: nearest neighbor matching (greedy matching) and optimal matching. In nearest neighbor matching, a patient is randomly selected from the treatment group and subsequently paired with a patient in the control group with the closest propensity score [13].

Nearest neighbor matching within a caliper is generally used and the propensity scores of the matched sample lie within a specified width of calipers. A narrower caliper will result in more similarly matched pairs but a decreased number of matched patients. A previous study suggested that using a caliper width of 0.25 of the standard deviation of the propensity score logit removed 98% of the bias due to measured covariates [14]. Another study recommended using a caliper width equal to 0.2 of the standard deviation of the propensity score logit [15].

Optimal matching aims to minimize the sum of within-pair differences in the propensity scores. However, optimal matching may not necessarily create more balanced distribution than nearest neighbour matching.

Overall, I recommend researchers to use nearest neighbor matching within a caliper width equal to 0.2 or 0.25 of the standard deviation of the propensity score logit without replacement.

Propensity score matching can be conducted using the Matching or MatchIt package in R Software, or the PSMATCH2 module in Stata.

**Outcome comparisons after propensity score matching**

It is generally recommended to compare outcomes after propensity score matching as if the data are paired. A paired $t$ test or the Wilcoxon signed-rank test can be used to test the inter-group differences for continuous outcomes, and the McNemar’s test for binary outcome.

Conditional logistic regression or generalized estimating equations for logistic regression can be used to estimate the odds ratio of the treatment for the outcome. With binary outcomes, the effect of treatment can also be described using the relative risk or the number needed to treat.

#### 4-2. Covariate Adjustment Using the Propensity Score

Propensity scores themselves can be used as a covariate in the regression model, where the outcome variable is regressed against the treatment assignment and the estimated propensity scores. This approach is useful, because
it can incorporate many covariates [16].

A linear regression model can be used for continuous outcomes, while logistic regression model can be selected for binary outcomes. The treatment effect is estimated using the regression coefficient of the treatment variable.

This approach assumes that the relationship between the propensity score and the outcome is correctly modeled. However, this approach should be carefully used because bias may increase when the variance in the treatment and control groups are very different [17].

4. Propensity Score Weighting

Propensity score weighting is estimated for binary outcomes. However, this approach should be carefully used because bias may increase when the variance in the treatment and control groups are very different [17].

4.3. Stratification

Stratification divides the sample population into mutually exclusive subgroups based on their propensity scores. A previous study showed that using the quintiles (five approximately equal-size strata) reduced at least 90% of the confounding [18]. With a large sample size, we can use 10 or 20 strata.

Propensity score matching can eliminate a greater proportion of the differences between groups, compared with stratification [19]. Studies using stratification are relatively rare, and thus this approach is not strongly recommended.

4.4. Propensity Score Weighting

Inverse probability of treatment weighting (IPTW) using the propensity score creates a synthetic sample where the distribution of measured covariates is independent of treatment assignment.

The weight for ith patient is defined as:

\[
\frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i}
\]

where \(Z_i\) denotes whether the ith subject was treated (=1) or not (=0); \(e_i\) denotes the propensity score; and \(p\) denotes the marginal probability of treatment in the overall sample.

IPTW should be performed with caution. IPTW can obtain unbiased estimates of ATE. However, these estimates are only valid without residual systematic differences in measured covariates between treated and control patients [20]. Most problematically, a very low propensity score can result in a very large weight. IPTW may have poor performance when the weights for a small number of patients are extremely large [21].

Alternatively, trimmed or truncated weights are used to address the issues related to very large weights. The thresholds are often set as the 1st and 99th percentiles [22].

Another alternative is to use stabilized IPTW [23]. The weight for ith patient is defined as:

\[
\frac{pZ_i}{\bar{e}_i} + \frac{(1 - p)(1 - Z_i)}{1 - \bar{e}_i}
\]

where \(Z_i\) denotes whether the ith subject was treated (=1) or not (=0); \(e_i\) denotes the propensity score; and \(p\) denotes the marginal probability of treatment in the overall sample.

Using stabilized IPTW preserves the sample size of the original data, and produces appropriate estimation of the variance of treatment effects [24].

5. Comparison with Conventional Methods

A literature review showed 87% of studies using propensity scores did not have substantially different estimates compared with conventional multivariable methods [25]. On the other hand, a study using a simulated population showed estimation of a general treatment effect with PS methods was closer to the true marginal treatment effect than a logistic regression model [26]. In studies with small numbers of events, propensity score analysis yielded less biased, more robust, and more precise estimates than a regression model [27].

REFERENCES


