Journal of MultiDisciplinary Evaluation
Volume 11, Issue 25, 2015

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Propensity Scores: A Practical Introduction Using R

Antonio Olmos
University of Denver

Priyalatha Govindasamy
University of Denver

Background: This paper provides an introduction to propensity scores for evaluation practitioners.

Purpose: The purpose of this paper is to provide the reader with a conceptual and practical introduction to propensity scores, matching using propensity scores, and its implementation using statistical R program/software.

Setting: Not applicable

Intervention: Not applicable

Research Design: Not applicable

Data Collection and Analysis: Not applicable

Findings: In this demonstration paper, we describe the context in which propensity scores are used, including the conditions under which the use of propensity scores is recommended, as well as the basic assumptions needed for a correct implementation of the technique. Next, we describe some of the more common techniques used to conduct propensity score matching. We conclude with a description of the recommended steps associated with the implementation of propensity score matching using several packages developed in R, including syntax and brief interpretations of the output associated with every step.

Keywords: *propensity score analysis; propensity score matching; R.*

Introduction

The aim of this paper is to provide the reader with a conceptual and practical introduction to propensity scores, matching using propensity scores, and its implementation using a statistics program. We start with a description of the context in which propensity scores have been used, the basic assumptions needed to use propensity scores, and a brief description of some of the most useful techniques for propensity score matching. We then provide a detailed description of how to estimate propensity scores, matching using propensity scores, and brief examples of the results of implementing propensity scores matching using several packages developed in R.

Context for Propensity Scores

We live in a period with serious social problems, such as low academic achievement, obesity, homelessness, and drug addiction, to name a few. This era is also characterized by accountability. Social programs intended to address these social problems need to demonstrate their effectiveness (Weiss, 1998). It is in this environment of social responsibility and social accountability that program evaluation plays a crucial role. Evaluation is defined as: "*the process of determining the merit, worth and value of things*" (Scriven, 1991, p. 1), and the aim of program evaluation is to "*systematically assess the merit or worth of something*" (Guskey, 1999, p. 37). Program evaluation is used to assess results and help improve outcomes intended for programs.

The stress on accountability has led to the development of interventions that are considered evidence-based practices. For example, in mental health, evidence-based practices are defined as "interventions for which there is consistent scientific evidence showing that they improve client outcomes" (Drake et al., 2001, page 180). A centerpiece of evidence-based practices is proof of causality. That is, evidence-based practices require a demonstration that the improvement/changes observed in individuals is due to the intervention.

According to Guo and Fraser, (2015), causal inferences have four requirements: (1) there is a statistical relationship between the treatment and the outcome, (2) the presumed cause happens before the effect, (3) the researchers are able to rule-out alternative explanations for the observed change, and (4) there is a reasonable counterfactual.

The first three requirements are straightforward. For example, regarding a statistical relationship, we should be able to detect it using some statistical method, often correlation. Similarly for precedence: careful observation should be helpful in establishing whether the cause precedes in time the effect/outcome. With regard to ruling out alternative explanations, there are multiple strategies that can be implemented. Cook and Campbell (1979), and more recently, Shadish, Cook, and Campbell (2002) have described multiple strategies to rule out alternative explanations, random assignment being one strategy. Although more involved, the counterfactual is still straightforward. In every study, each subject can be assigned to one of two (or more) treatment alternatives. For example, if the study is a comparison between an online versus a traditional course, the alternatives are whether a student is assigned to online or traditional. Although the assigned treatment will affect the outcome of the study (e.g., the final grade), we know that each student has two potential outcomes (one for each treatment option), even though we can only observe the outcome for one of them at any point in time.

A counterfactual is defined as "*knowledge of what would have happened to those same people if they simultaneously had not received treatment*" (Shadish et al., 2002, p. 5). Thus the counterfactual is a thought experiment. The estimate of the effect is the difference between what happen (the real outcome) and what would have happened (the potential outcome) if the assignment had been reversed. This is what is known as the Neyman-Rubin (Guo & Fraser, 2015) framework: Individuals selected into a treatment/control condition have potential outcomes in both states, or $Y_i = W_i Y_{1i} + (1 - W_i) Y_{0i}$

However, this is a thought experiment. In practice, one of the outcomes proposed by the counterfactual is not observed (Holland, 1986; Morgan & Winship, 2012) which is the fundamental problem of causal inference. However, the Neyman-Rubin counterfactual framework holds that we can estimate the counterfactual by examining: $\bar{x}_{tx} - \bar{x}_{control}$. Since both outcomes are observable, we can then define the treatment effect as a mean difference.

A way to guarantee that the counterfactual works as planned is to assure that the only difference between the two groups is the treatment. That means that all extraneous variables are controlled/eliminated. And the best way to control for the effect of extraneous variables is by random assignment.

Given the importance of causality, and the requisites needed to assign it with confidence, evidence-based programs tend to rely on the use of experimental approaches. The experimental approach has two characteristics: 1) it manipulates the independent variable, that is, whether an individual receives (or not) the intervention under scrutiny. 2) Individuals are randomly assigned to the independent variable. The first characteristic does not define the experimental approach: most of the so-called quasi-experiments (Shadish et al., 2002) also manipulate the independent variable. What defines the experimental method is the use of random assignment. In particular, the use of random assignment helps to prove causality by improving the chances that we have ruled out alternative explanations. Another way to think about the importance of random assignment is that it increases the chances that groups are probabilistically balanced on some variables that otherwise may affect the final outcome (D'Agostino & D'Agostino, 2007; Shadish et al., 2002), and, therefore, the Neyman-Rubin counterfactual framework holds. For example, in an obesity-reduction program, there may be several reasons for weight loss (such as peer support, level of motivation), that are not associated with the intervention, and that may affect weight reduction. Balancing through random assignment becomes important because then we can determine with a high degree of certainty that the reason we observe the weight change in this obesity reduction program is because of the intervention, and not because of some other reason (Bonell et al., 2009).

However, if one of these two conditions (manipulating the independent variable, or random assignment to rule out alternative explanations) is not met, our confidence about the causal relationship between independent and dependent variable is substantially reduced. There are several reasons why we may not be able to meet these two assumptions: 1) Despite the use of random assignment, equivalent groups are not achieved. 2) Due to ethical or logistical reasons random assignment is not possible (Bonell et al., 2009).

The first reason is known as randomization failure (Bonell et al., 2009), and sometimes can go undetected. Usual reasons why randomization can fail are associated with missing data which happened in a systematic way. In the obesity reduction example, some individuals in the control group may drop out because they are not losing weight. Or individuals in the treatment group may drop out because they lost the weight they had as a goal, and therefore are not motivated to continue with the program. In both instances, an analysis based only on the outcomes of the individuals who stayed until the end of the program may produce biased results (i.e., the average weight loss observed across groups is either under-estimated or over-estimated with respect to the real weight loss, had all the individuals in the original sample stayed until the end of the study).

Sometimes random assignment cannot be accomplished because of ethical or logistical reasons. For example, it will be unethical to randomly assign individuals to either a control or treatment condition if those assigned to the control condition were to lose access to some important resource (e.g., a drug that may save or prolong their lives), or those in the treatment group could risk to lose some benefit they might already have (e.g., Medicare or Medicaid benefits). Logistically, there are multiple treatment conditions that are attributes (i.e., intrinsic to the individual such as gender, ethnicity, socio-economic status, disability (Gliner, Morgan & Leech, 2009) which cannot be manipulated by the evaluator (i.e., individuals cannot be assigned to a different gender). In both cases, individuals are not assigned to a treatment condition at random, thus confidence regarding causality is compromised and the study becomes a quasi-experiment.

There are some steps we can take to improve quasi-experiments:

1. We can try to rule-out alternative explanations by adding elements to the research design; for example, we can add: (a) observations (pretest and posttests), (b) comparison groups (control, placebo, other treatments); (c) other factors that may be related to outcomes, and (d) other outcome variables that should not be affected by the intervention.
2. We can use statistical adjustments to try to control alternative explanations. For example: (a) matching, stratification, weighting, using covariates with ANCOVA or regression models; (b) single, multiple, or aggregate covariates; or (c) propensity scores

As a consequence of randomization failure, or because of the logistical or ethical reasons just described, in a very large number of real-world interventions, experimental approaches are impossible or very difficult to implement. If we are still interested in demonstrating the causal link between our intervention and the observed change, our options become limited. Some options include regression discontinuity designs (Trochim, 1984; Shadish et. al., 2002) which can strengthen

our confidence about causality by selecting individuals to either the control or treatment condition based on a cutoff score. Another alternative when random assignment fails, or when we cannot randomly assign people to treatment conditions because of ethical or logistical reasons, is propensity scores.

Propensity Scores

Propensity scores is a statistical technique that has proven useful to evaluate treatment effects when using quasi-experimental or observational data (Austin, 2011; Rubin, 1983). Some of the benefits associated with propensity scores are: (a) Creating adequate counterfactuals when random assignment is infeasible or unethical, or when we are interested in assessing treatment effects from survey, census administrative, or other types of data, where we cannot assign individuals to treatment conditions (Austin, 2011). (b) The development and use of propensity scores reduces the number of covariates needed to control for external variables (thus reducing its dimensionality) and increasing the chances of a match for every individual in the treatment group. (c) The development of a propensity score is associated with the selection model, not with the outcomes model, therefore the adjustments are independent of the outcome.

Propensity scores are defined as the conditional probability of assigning a unit to a particular treatment condition (i.e., likelihood of receiving treatment), given a set of observed covariates: $(z=i|\mathrm{X})$

Where z = treatment, i = treatment condition, and X = covariates. In a two-group (treatment, control) experiment with random assignment, the probability of each individual in the sample to be assigned to the treatment condition is: $(z=i \mid \mathrm{X})=0.5$. In a quasi-experiment, the probability $(z=i \mid \mathrm{X})$ is unknown, but it can be estimated from the data using a logistic regression model, where treatment assignment is regressed on the set of observed covariates (the so-called *selection model*). The propensity score then allows matching of individuals in the control and treatment conditions with the same likelihood of receiving treatment. Thus, a pair of participants (one in the treatment, one in the control group) sharing a similar propensity score are seen as equal, even though they may differ on the specific values of the covariates (Holmes, 2014).

Why not Use Covariates?

Conventional matching using covariates can work well; however, as the number of covariates increases, it becomes difficult to find good matches for subjects in the treatment group. Thus matching using covariates can result in dropping cases and, when there are many covariates or lots of variation, matching may be impossible. As described earlier, propensity scores provide an advantage in this case because they reduce the dimensionality by summarizing many covariates into a single score. Rosenbaum and Rubin (1983) explain that propensity scores summarize many fine scores into a single coarse score. They have also shown that a coarse score can balance differences observed in the fine scores between treated and control participants.

Endogeneity and the Ignorable Treatment Assignment Assumption (ITAA)

There are two assumptions associated with causality that we need to understand before we can use propensity scores:

Endogeneity

In order to work, regression models need to meet some assumptions (Draper & Smith, 1998). One of them calls for independence between the independent variables in the model and the error term. Violations of this assumption are usually associated with omitted variables. That is, there is some other variable that is not included in the model which is correlated with both the dependent and the independent(s) variables. Omitted variables are one of the major problems in non-experimental (observational/quasi-experimental) studies, because if we do not take them into account, they will create a biased estimate of the effect. That is, our interpretation of the regression model will either under-estimate or over-estimate the relationship between the independent and dependent variables. Omitted variables represent a form of endogeneity which affects our ability to establish accurate causal relationships.

Ignorable Treatment Assignment Assumption

One of the four requirements needed to demonstrate causality is the counterfactual, which is supported by the Neyman-Rubin counterfactual

model (Rubin, 2005). This framework relies in one important assumption known as the Ignorable Treatment Assignment Assumption, which states that conditional on covariates, the assignment of study participants to treatment conditions is independent of the outcome: $(Y_i0, Y_i1) \perp W|X$. Under random assignment, this assumption holds, but it does not necessarily hold under quasi-experiments, where it may be important to investigate how participants were assigned to the treatment conditions. Although some of the processes by which individuals select/are assigned to specific treatment conditions can be examined empirically, a full account of treatment selection is sometimes impossible (e.g., when subjects are motivated to select one treatment condition, and the researcher does not have a valid measure/is not aware of their motivation).

If we cannot determine all the reasons why a participant is assigned to a treatment, then we will have an endogeneity problem (Morgan & Winship, 2012). Thus it is important to make sure that we can identify all of the reasons why participants are in the treatment or control conditions.

Conventional Matching in R

Conventional forms of matching allow researchers to create two groups of individuals (control, treatment groups) that are matched in variables that are believed to be critical in the selection process, thus creating counterfactuals for the individuals in the opposite group. Below, we briefly describe two of the conventional ways of matching, as well as R code to conduct it. However, a more thorough description of the interpretation will be included in the section: "Implementing a propensity score analysis with R." The following code and examples use the dataset "Lalonde," included and described in the packages MatchIt (Ho, Imai, King & Stuart, 2011) and matching (Sekhon, 2011).

Mahalanobis Metric Matching

Mahalanobis metric distance matching is based on the Mahalanobis distance, which calculates distances in a multidimensional space (Guo & Fraser, 2015). In the context of matching, once participants in the control group are selected at random, we calculate distances between treated and control participants: $d(i,j) = (u-v)^T C^{-1}(u-v)$ where u (treatment) and v (control) are the matrices with covariates, and C^{-1} is the inverse variance-covariance matrix for the control subjects. The control subject with the minimum distance d(i,j) is chosen as the match for a treated subject. Both subjects are removed from the pool, and the process repeats until we match all treated subjects. Notice that this approach does not lose participants, because we are selecting participants from the control group who have a minimum distance to participants in the treatment group. However, it is possible that an individual from the control group who is selected as the match for a participant in the treatment group, is not close in a multidimensional space. In fact, as the number of covariates increases, the average Mahalanobis distance between observations also increases (Gu & Rosenbaum, 1993; Stuart, 2010; Zhao, 2004).

Mahalanobis Metric Matching with Calipers

A proposed solution to this distance problem is the use of Mahalanobis matching but where the distance is estimated based on a caliper. In this context, the selection of the closest match is determined by $\|d(i,j)\| < \varepsilon$ where epsilon is a pre-specified tolerance for matching (a "caliper"). Cochran and Rubin (1973) suggested using a caliper size of one-fourth of a standard deviation of the sample estimated propensity scores (i.e., $\varepsilon < .25\ \sigma_p$, where σ_p denotes standard deviation of the estimated propensity scores of the sample). Figure 1 below presents the commands used to conduct matching using Mahalanobis distance in the package Matchit (Ho, Imai, King & Stuart, 2011).

```
m.mahal <- matchit(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, mahvar
s = c("age", "educ", "nodegree", "re74", "re75"), caliper = 0.25, replace = FALSE, di
stance= "mahalanobis")
summary(m.mahal)
```

Figure 1. Conventional matching using Mahalanobis distance with the package MatchIt

In this figure it can be observed that a caliper is included (i.e., caliper = 0.25). As recommended by Rosenbaum and Rubin, 1985, the default is 0.25 $\varepsilon_{p.}$

Types of Matching Using Propensity Scores

Conventional matching techniques work well in many conditions. However, as described earlier, the use of propensity scores for matching has some advantages such as coarse score balancing (Rosenbaum and Rubin, 1983). Once propensity scores have been calculated, we proceed to find participants in the control group that will have similar propensity scores to those in the treatment group. We use a typology similar to that proposed by Bai and Clark (2012) to differentiate between greedy matching and more complex forms of matching.

Greedy Matching

This type of matching is called greedy because the match for a participant in the treatment group is based on the first case of the control group that meets the criteria for matching. Even if that participant in the control group would serve as a better match for a subsequent participant in the treatment group, the match will still be based on the first case. Most algorithms will select participants from both the control and treatment group at random; thus, the match from one run to the next will render different groups with different degrees of matching. The most common greedy matching algorithms and their code in R are described below:

Near Neighbor Matching

The near neighbor matching procedure matches participants from the control group to participants from the treatment group based on closeness. A participant (j) with propensity score P_j in the control sample (I_0) is a match for a participant (i) with propensity score P_i in the treatment group, if the absolute difference between their propensity scores is the smallest $C(P_i) = \min_j \|P_i - P_j\|, j \in I_0$. The most traditional matching is of one participant in the control to one participant in the treatment. In those cases, we speak about 1-to-1 (1:1) matching. However, it is possible to have more than one participant from the control group to be matched with a participant in the treatment group. In those cases, we speak about an m-to-1 (m:1) matching. Having more individuals from the control group matched to every individual in the treatment group means better estimates for the counterfactual in the control group. However, this approach requires a sample size for the control group several times larger than the number of individuals in the treatment group. Figure 2 presents the commands to estimate near neighbor using the default (1:1) using the package MatchIt. Figure 3 presents the commands to estimate near neighbor using a 2:1 ratio:

```
#---Match using near-neighbor
m.nn <- matchit(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, method= "
nearest", ratio = 1)
summary(m.nn)
```

Figure 2. Propensity score matching using near neighbor for a 1:1 ratio with the package MatchIt

```
#---match using near-neighbor with ratio = 2
m.nn2 <- matchit(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, method=
"nearest", ratio = 2)
summary(m.nn2)
```

Figure 3. Propensity score matching using near neighbor for a 2:1 ratio with the package MatchIt

The keyword to change the ratio from the default (1:1) to a (2:1) ratio is the subcommand: ratio = 2

Near Neighbor with Caliper Matching

Similarly to the case defined for Mahalanobis distance, in near neighbor matching, sometimes individuals who are not close in terms of their propensity scores, can be matched. Thus, in this case, near neighbor-match can be considered only if the absolute distance between treatment and control participants meets the condition: $\|P_i - P_j\| < \varepsilon,\ j \in I_0$ where P_i and P_j are propensity scores for treatment and control, ϵ is the caliper. As described earlier, Rosenbaum and Rubin (1985) have suggested that $\epsilon < 0.25\ \sigma_p$, where σ_p = standard deviation of the estimated propensity scores of the sample. This approach is popular because multivariate analyses using the matched sample can be undertaken, if the sample is sufficiently large. Figure 4 presents the commands to estimate near neighbor using a caliper matching of $0.1\ \sigma_p$ using MatchIt:

```
#---Match using near neighbor with calipers
m.nnc <- matchit(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, method=
"nearest", caliper =.1)
summary(m.nnc)
```

Figure 4. Propensity score matching using near neighbor with a caliper of 0.1 with the package MatchIt

The default caliper in MatchIt is 0.25.

Combining Propensity Scores and Mahalanobis Matching

This strategy combines propensity scores and the Mahalanobis distance for matching. After propensity scores have been calculated in all the participants, the estimates are added to the datafile. Next, participants in the treatment group are randomly ordered. Afterwards, the Mahalanobis distances for the participants in the control and treatment groups are calculated using the combination of variables (x) and the propensity score $\hat{e}(x)$ (Guo & Fraser, 2015). Figure 5 presents the commands to estimate the propensity scores, and the Mahalanobis distance plus propensity scores using MatchIt. The steps are recounted below:

```
#---Compute Propensity score
ps<- glm(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, family = binomia
l())
summary(ps)

#---Attach the predicted propensity score to the datafile
lalonde$psvalue <- predict(ps, type = "response")


#---match using Mahalanobis distance including Propensity score
m.mahalp <- matchit(treat ~ age + educ + nodegree + re74 + re75 + psvalue, data =lalo
nde, mahvars = c("age", "educ", "nodegree", "re74", "re75", "psvalue"), caliper = 0.2
5, replace = FALSE, distance= "mahalanobis")
summary(m.mahalp)
```

Figure 5. Propensity scores plus Mahalanobis distance using MatchIt

Optimal Matching

This is a more complex approach to propensity score matching, which is possible because of fast computer processing speed that can make the implementation of these algorithms possible. The main goal of this approach is to find the matched samples with the smallest average absolute propensity score distance across all the matched pairs. To accomplish this goal, matched sets of treatment and control participants are identified so that the matching minimizes the total propensity score distance (Δ) for a given data set according to the following equation:

$$\Delta = \sum_{s=1}^{S} \omega\,(|A_s|, |B_s|)\delta(A_s, B_s)$$

where ω is the weight of the number of subjects in the stratum, $|A_s|$ or $|B_s|$ represent the number of elements in the stratum that belongs to A/B, and δ is the distance between elements in the stratum. The optimal matching approach identifies sets of participants with the aim of minimizing the total distance between their propensity scores. A very important characteristic of this approach is that the match among participants can change, if it is found that a different match will in fact minimize the total distance even further. Figure 6 presents the commands to estimate Optimal distance using MatchIt.

```
#---Optimal matching with 1:1 ratio
m.om <- matchit(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, method= "
optimal", ratio = 1)
summary(m.om)
```

Figure 6. Optimal matching using MatchIt.

Be aware that MatchIt calls the package optmatch (Hansen, Fredrickson, Bertsekas & Tseng, 2013). Thus, it may be best to install the package optmatch in advance. Figure 6 also shows that different matching ratios can be set using the subcommand ratio.

Full Matching

Full matching is a form of optimal matching where participants in either the control or treatment groups will be matched to one or more individuals in the opposite group (Hansen, 2004; Stuart, 2010). This matching is accomplished by creating subclasses automatically, and then assigning at least one individual from either the control or treatment group, and as many individuals of the opposite group. Similarly to optimal matching this approach is intended to minimize the average of the distances between treatment and control individuals within each set (Stuart, 2010). This approach is ideal for researchers who would rather not discard any participant in either sample. Figure 7 presents the commands to estimate Full distance using MatchIt.

```
#---Full matching
m.fl <- matchit(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, method= "
full", min.controls = 1, max.controls = 10, discard = "both")
summary(m.fl)
```

Figure 7. Full matching using MatchIt.

Just like for optimal matching, MatchIt calls the package optmatch (Hansen et. al., 2013). Figure 7 also shows that a minimum number and a maximum number of control cases can be specified.

Implementing a Propensity Score Analysis with R

In this section, we provide some suggestions for the implementation of a propensity score analysis. We use the statistical program R (R Core Team, 2014), because it has multiple packages intended to calculate propensity scores, interpret the results using both statistical and graphical procedures, and it can also estimate post-matching outcomes analysis.

The analysis of quasi-experimental or observational data using propensity scores involves the development of two models: 1) the so-called selection model (which is intended to balance the groups using variables that affect only the selection process), and 2) the outcomes model (which will include variables that are associated with the outcomes only). The result of the selection model affects the outcome model through the propensity scores. Thus several of the steps described below are aimed exclusively to the

selection process, and some others are intended for the outcomes model.

Steps Suggested for Conducting a Propensity Score Analysis

When conducting a propensity score analysis, the authors follow these steps:

1. Preliminary analysis
2. Estimation of Propensity scores
3. Propensity Score matching
4. Outcome analysis
5. Sensitivity analysis

1. Preliminary Analysis

Before propensity scores are calculated, it is a good practice to determine if the two groups are balanced. The best practice to determine the covariates that influence group assignment is based on theoretical evidence. In addition, statistical tests can also be used to determine if the covariates are imbalanced across groups. Traditional statistical approaches include the estimation of a normalized difference (Imbens & Wooldridge, 2009), which calculates the difference between the control and treatment group for every variable included in the selection model. Hansen and Bowers (2008) suggested the equivalent of an omnibus test that checks if there is at least one variable in the selection model for which the two groups are different. The package RItools (Bowers, Fredrickson & Hansen, 2014) includes the routine "XBalance" that estimates a chi-square test to perform this omnibus test. Figure 8 includes examples of the code and the output generated by R when estimating the standardized/normalized difference and the Chi-square test using XBalance:

```
#---Computing indices of covariate imbalance before matching
### 1. Standardized difference
treated <- (lalonde$treat==1)
cov <- lalonde[,2:9]
std.diff <- apply(cov,2,function(x) 100*(mean(x[treated])- mean(x[!treated]))/(sqrt(0
.5*(var(x[treated]) + var(x[!treated])))))
abs(std.diff)

##        age       educ      black     hispan    married   nodegree
##  24.190362   4.475509 166.771881  27.693960  71.949196  23.504820
##       re74       re75
##  59.575159  28.700211

### 2. chi-square test
library("RItools")

xBalance(treat ~ age + educ + nodegree + re74 + re75, data = lalonde, report = c("chi
square.test"))

## ---Overall Test---
##         chisquare df  p.value
## unstrat      50.3  5 1.19e-09
## ---
## Signif. codes:  0 '***' 0.001 '** ' 0.01 '*  ' 0.05 '.  ' 0.1 '   ' 1
```

Figure 8. Testing imbalance before matching

For the standardized difference, absolute scores higher than 25% are considered suspect, and may indicate an imbalance for that specific variable (Stuart & Rubin, 2008). A statistically significant chi-square will indicate that at least one of the variables included in the model is creating an imbalance between the two groups. Variables that create imbalance should be included in the selection model.Also as part of the preliminary analysis it is a good practice to assess the effects of the treatment on the outcome variable by running the outcome model. This assessment can be based on the treatment variable only (using a t-test), or include covariates (using a regression model) Figure 9 includes examples of the code and the output of a regression analysis.

```
#--- Outcome model using Regression analysis
reg <-lm(re78 ~ treat + age + educ + nodegree + re74 + re75 + married + black + hispa
n, data = lalonde)
summary(reg)

##
## Call:
## lm(formula = re78 ~ treat + age + educ + nodegree + re74 + re75 +
##     married + black + hispan, data = lalonde)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13595  -4894  -1662   3929  54570
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.651e+01  2.437e+03   0.027   0.9782
## treat        1.548e+03  7.813e+02   1.982   0.0480 *
## age          1.298e+01  3.249e+01   0.399   0.6897
## educ         4.039e+02  1.589e+02   2.542   0.0113 *
## nodegree     2.598e+02  8.474e+02   0.307   0.7593
## re74         2.964e-01  5.827e-02   5.086 4.89e-07 ***
## re75         2.315e-01  1.046e-01   2.213   0.0273 *
## married      4.066e+02  6.955e+02   0.585   0.5590
## black       -1.241e+03  7.688e+02  -1.614   0.1071
## hispan       4.989e+02  9.419e+02   0.530   0.5966
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6948 on 604 degrees of freedom
## Multiple R-squared:  0.1478, Adjusted R-squared:  0.1351
## F-statistic: 11.64 on 9 and 604 DF,  p-value: < 2.2e-16
```

Figure 9. Outcome model using a Regression analysis

2. Estimation of the Propensity Scores

In this step, the propensity score is estimated. Although propensity scores can be estimated using models such as discriminant analysis, probit regression, boosted regression (McCaffrey, Ridgeway & Morral, 2004), and even genetic algorithms (Sekhon, 2011), logistic regression is widely used. Packages such as MatchIt (Ho, Imai, King & Stuart, 2011) and Matching (Sekhon, 2011) estimate propensity scores using logistic regression as the default option. However, when estimating propensity scores using the default option, the fit of the model cannot be assessed. Therefore, it is recommended that a logistic regression is run to determine the model fit. Figure 10 includes the estimation of the propensity scores using logistic regression

```
#---Calculates the propensity score
ps <- glm(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, family = binomi
al())
summary(ps)

##
## Call:
## glm(formula = treat ~ age + educ + nodegree + re74 + re75, family = binomial(),
##     data = lalonde)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2559  -0.9053  -0.6053   1.2060   2.9809
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.694e+00  7.989e-01  -3.372 0.000746 ***
## age          2.464e-03  1.025e-02   0.240 0.810019
## educ         1.569e-01  5.299e-02   2.962 0.003059 **
## nodegree     8.502e-01  2.813e-01   3.023 0.002503 **
## re74        -1.225e-04  2.576e-05  -4.756 1.98e-06 ***
## re75         2.574e-05  3.955e-05   0.651 0.515252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.49  on 613  degrees of freedom
## Residual deviance: 692.88  on 608  degrees of freedom
## AIC: 704.88
##
## Number of Fisher Scoring iterations: 5
```

Figure 10. Propensity score estimation using logistic regression

Statistically significant estimates are identified by low (i.e., < 0.05) p-values. There are no clear suggestions as to whether to include in the final model all the variables (even non-significant). Some authors (Austin, Grootendorst & Anderson, 2007; Caliendo & Kopeinig, 2008) suggest that the final model should include not only statistically significant variables, but also variables known to be associated with selection.

Once the propensity scores have been calculated, a graphical approach can be used to assess the distributional similarity between score distributions. This graphical approach uses back to back histograms such as those created through the package Hmisc (Harrell, 2015). Back to back histograms cannot be used with Mahalanobis distance, because it is a multidimensional technique. Figure 11 presents the commands and the histograms Hmisc generates.

```
#---Attach the predicted propensity score to the datafile
lalonde$psvalue <- predict(ps, type = "response")
#---Back to back histogram
histbackback(split(lalonde$psvalue, lalonde$treat), main= "Propensity score before ma
tching", xlab=c("control", "treatment"))
```

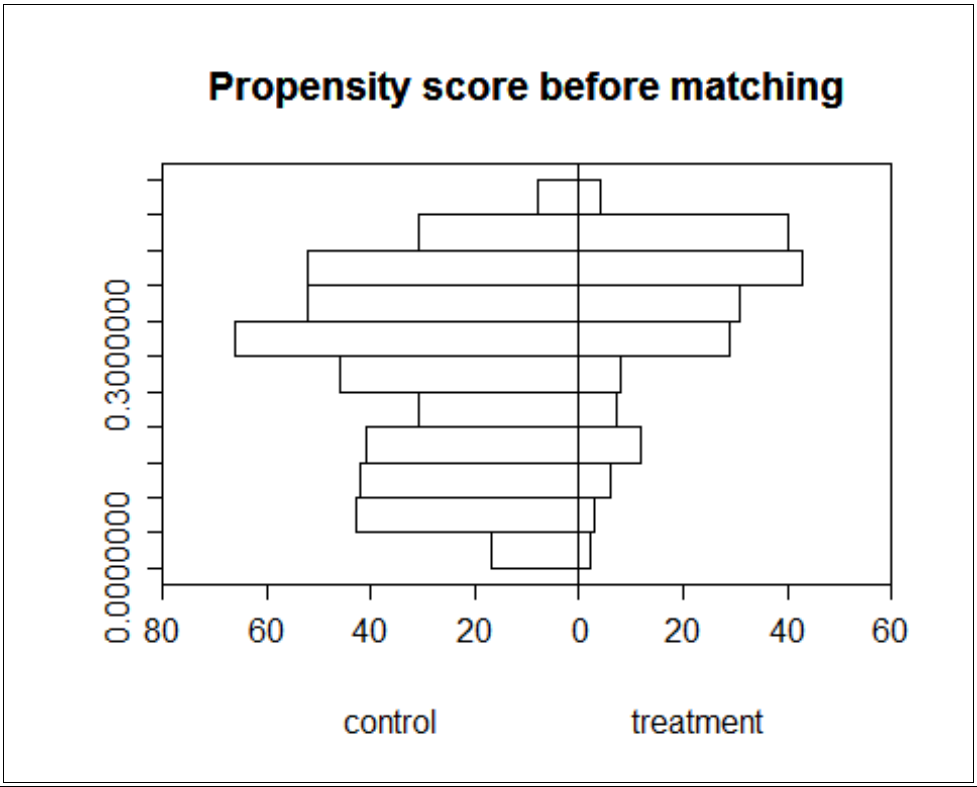


Figure 11. Back to back histogram using Hmisc

Important parameters to determine the fit are not only the shape, but also degree of overlap between the two distributions (known as the common support region (Lehner, 2008)). Matching is best when there is a common support region.

3. Propensity Score Matching

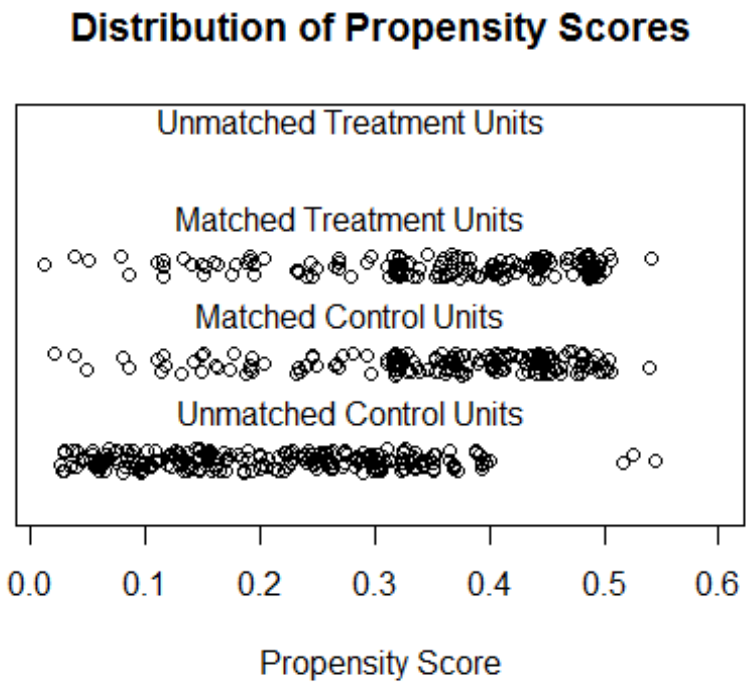
The code for running propensity score matching was provided earlier, therefore, in this section, the aim is to describe some key issues in reviewing the output of the matching algorithms. For example, it is important to check how well matching worked. Packages such as MatchIt provide summary tables that include means and standard deviations for the two groups both before and after the matching was completed. It also includes percent improvement, and finally, it provides a summary of the number of individuals included in the final sample, and cases that were not matched. The number of matched and unmatched cases is usually dependent on the match ratio imposed by the user, and the number of cases in the treatment group. For example, if the match ratio was 1:1, and there were 500 individuals in the control group and 250 in the treatment group, then 250 individuals from the control group will not be matched. Figure 12 is an output summary for near neighbor matching.

```
#---Match using near-neighbor
m.nn <- matchit(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, method= "
nearest", ratio = 1)
summary(m.nn)

##
## Call:
## matchit(formula = treat ~ age + educ + nodegree + re74 + re75,
##     data = lalonde, method = "nearest", ratio = 1)
##
## Summary of balance for matched data:
##          Means Treated Means Control SD Control Mean Diff  eQQ Med
## distance        0.3650        0.3603     0.1092    0.0047   0.0016
## age            25.8162       24.7838     9.6480    1.0324   3.0000
## educ           10.3459       10.1676     2.6166    0.1784   0.0000
## nodegree        0.7081        0.7459     0.4365   -0.0378   0.0000
## re74         2095.5737     2218.4725  4371.6213 -122.8988 104.5930
## re75         1532.0553     1428.9774  2297.0371  103.0779 172.5310
##          eQQ Mean    eQQ Max
## distance   0.0052     0.0303
## age        2.9568     8.0000
## educ       0.5351     4.0000
## nodegree   0.0378     1.0000
## re74     445.2718  9177.7500
## re75     409.0697 13737.8900
##
## Percent Balance Improvement:
##          Mean Diff.   eQQ Med eQQ Mean   eQQ Max
## distance    94.8731   98.2359  94.3852   82.4986
## age         53.3698 -200.0000   9.4371   20.0000
## educ       -61.4069  100.0000  23.8462    0.0000
## nodegree    66.0256    0.0000  66.6667    0.0000
## re74        96.5122   95.6879  87.7028    0.4204
## re75        88.9689   82.4145  61.4325 -102.1762
##
## Sample sizes:
##           Control Treated
## All           429     185
## Matched       185     185
## Unmatched     244       0
## Discarded       0       0

match.data = match.data(m.nn)
```

Figure 12. Near neighbor matching using MatchIt

Graphical approaches (such as the jitter type in the package MatchIt), will help the user get some idea of whether the individuals not matched are in some specific part of the propensity-score continuum. Figure 13 shows an example of such a plot:

```
plot(m.nn, type = "jitter")
```




```
## [1] "To identify the units, use first mouse button; to stop, use second."

## integer(0)
```

Figure 13. Jitter-type plot, package MatchIt

As can be observed in this figure, the section labeled "Unmatched control Units" shows that most of the non-matched individuals were in the lower (0.0 to 0.4) part of the propensity scores. However, there were a few cases in a higher range (0.5-0.6).

It is important to determine that the groups are balanced, thus eliminating (or substantially reducing) the initial selection bias. In step 1 in this section (preliminary analysis) it was mentioned that there are both statistical as well as graphical approaches that can be used to determine the degree of imbalance. After the match has been conducted, both techniques are used again to determine that all the critical variables have been balanced. Figure 14 shows the output after the original match.

```
#---Computing indices of covariate imbalance after matching

### 1. Standardized difference
treated1 <- (match.data$treat==1)
cov1 <- match.data[,2:9]
std.diff1 <- apply(cov1,2,function(x) 100*(mean(x[treated1])- mean(x[!treated1]))/(sq
rt(0.5*(var(x[treated1]) + var(x[!treated1])))))
abs(std.diff1)

##        age       educ      black     hispan    married   nodegree
##  12.155616   7.644579 154.578773  34.492122  38.194233   8.478250
##       re74       re75
##   2.650808   3.686064

### 2. chi-square test
xBalance(treat ~ age + educ + nodegree + re74 + re75, data = match.data, report = c("
chisquare.test"))

## ---Overall Test---
##         chisquare df p.value
## unstrat      2.64  5   0.755
## ---
## Signif. codes:  0 '***' 0.001 '** ' 0.01 '*  ' 0.05 '.  ' 0.1 '   ' 1
```

Figure 14. Post-match analysis using standard difference and the overall chi-square test

As can be observed in this figure, although the chi-square test indicates no significance, thus suggesting equivalence between the groups, the standardized difference test shows that there are some variables with a large difference (i.e., black, hispan, married) that can still be improved. Potential suggestions might include the use of interactions terms, or polynomial terms to try to reduce their imbalance. McCaffrey, Ridgeway, and Morral (2004) has suggested the use of Generalized Boosted Models (GBM) to improve the fit. This approach uses decision trees that combine simpler models into a more powerful model. Similarly, Sekhon (2011) proposes the use of a genetic algorithm to develop the best possible model. Readers are directed to the original sources for more information about these techniques.

Other methodologists (Austin, Grootendorst, & Anderson, 2007) suggest that when balance in the selection model cannot be achieved on all the variables, those variables where balance was not achieved, and that may also be associated with the dependent variable could be included in the outcome model as covariates.

Another graphical approach that can be used to determine the match between groups is the back to back histograms. Figure 15shows a back to back histogram after the match:

```
histbackback(split(match.data$psvalue, match.data$treat), main= "Propensity score aft
er matching", xlab=c("control", "treatment"))
```

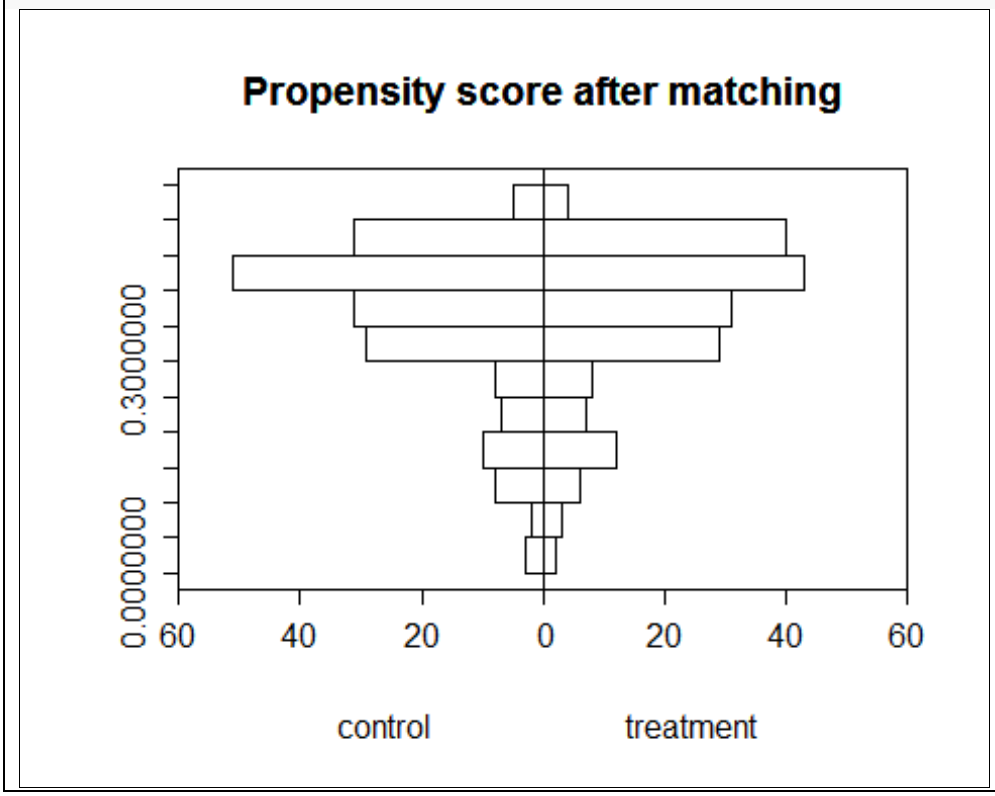


Figure 15. Back to back histogram after match

As can be observed in this figure, there is a remarkable improvement in the match between the two distributions of propensity scores after the match (compared to Figure 11, which shows the histograms for the same data before the match). This match suggests that the two groups are much more similar in terms of their propensity scores, and thus, the selection bias has been reduced substantially.

4. Outcomes Analysis

Once the researcher is satisfied with the propensity score matching, it is time to proceed with the outcome model. Several of the more frequently used techniques such as near neighbor, and Mahalanobis distances, can be used with analytic techniques such as linear regression models, ANCOVA, or even matched t-tests. However, the selection of any analytic approach to estimate the treatment effect and statistical significance should take into account the fact that the propensity score creates matched samples (Austin, 2008). Figure 16 shows the outcome analysis using paired t-test.

```
#---Outcome analysis using paired t-test
# this command saves the data matched
matches <- data.frame(m.nn$match.matrix)
#these commands find the matches. one for group 1 one for group 2
group1 <- match(row.names(matches), row.names(match.data))
group2 <- match(matches$X1, row.names(match.data))
# these commands extract the outcome value for the matches
yT  <- match.data$re78[group1]
yC  <- match.data$re78[group2]
# binding
matched.cases <- cbind(matches, yT, yC)
#Paired t-test
t.test(matched.cases$yT, matched.cases$yC, paired = TRUE)

##  Paired t-test
##
## data:  matched.cases$yT and matched.cases$yC
## t = 0.4342, df = 184, p-value = 0.6647
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1156.468  1809.111
## sample estimates:
## mean of the differences
##                326.3214
```

Figure16. Outcome analysis using a paired t-test

The simplicity of these outcomes analysis make them easy to complete and their implications are usually easy to understand. For optimal and full matching however, the outcome analyses are more complex. Under optimal and full matching, it is possible to have the same individual matched against more than one individual from the other group. And for full matching, this can be the case not only for individuals in the control group, but also for individuals in the treatment group. Given that individuals can be used more than once in both the treatment and control groups, multivariate techniques should not be used.

For these two techniques, it is recommended to use the Hodges Lehman (Hodges & Lehmann, 1962) and the difference in differences (Rubin, 1979). To these authors' knowledge, there are no packages within R that can be used to compute the outcomes analysis.

An added complexity in the outcomes analysis is the possibility to compute different types of treatment effects such as the Average Treatment Effect (ATE), which is the type of treatment effect evaluators are more familiar with. Another treatment effect that can be estimated is the Average Treatment for the Treated (ATT) where the main focus is to identify individuals that can be matched to those in the treatment group. And finally the Average Treatment for the Control (ATC) where the main emphasis is to find matches for individuals in the control group. Readers are directed to Morgan and Winship (2012), for more information about the differences among the different types of treatment effects.

5. Sensitivity Analysis

A question that any evaluator who uses propensity score matching should ask herself is: how sensitive are these results to hidden bias? Rosenbaum (2002, 2005) recommends that researchers try to answer this question by conducting a sensitivity analysis. The idea is to determine how susceptible the results presented might be to the presence of biases not identified by the researcher or removed by the matching. Rosenbaum (2002) developed methods to determine bias through several non-parametric tests such as McNemar's and Wilcoxon's signed rank test. Keele (2015) developed the package rbounds which estimates the sensitivity of the results to hidden bias. rbounds can compute sensitivity analysis straight from the package matching (Sekhon, 2011). For other propensity scores packages, some file reformatting needs to be completed before it can be submitted to rbounds. Figure 17 shows the

sensitivity analysis using the Wilcoxon's rank sign test.

```
library("Matching")
attach(lalonde)
Y <- lalonde$re78
Tr <-lalonde$treat
ps <- glm(treat ~ age + educ + nodegree + re74 + re75 + married + black + hispan, dat
a =lalonde, family = binomial())

#---Match - without replacement
Match <- Match(Y=Y, Tr=Tr, X=ps$fitted, replace=FALSE)

#---Runs the sensitivity test based on the matched sample using Wilcoxon's rank sign
test

psens(Match, Gamma = 2, GammaInc = 0.1)

##
##  Rosenbaum Sensitivity Test for Wilcoxon Signed Rank P-Value
##
## Unconfounded estimate ....  0.2858
##
##  Gamma Lower bound Upper bound
##    1.0      0.2858      0.2858
##    1.1      0.1338      0.4904
##    1.2      0.0541      0.6809
##    1.3      0.0194      0.8227
##    1.4      0.0063      0.9113
##    1.5      0.0019      0.9595
##    1.6      0.0005      0.9828
##    1.7      0.0001      0.9932
##    1.8      0.0000      0.9975
##    1.9      0.0000      0.9991
##    2.0      0.0000      0.9997
##
##  Note: Gamma is Odds of Differential Assignment To
##  Treatment Due to Unobserved Factors
##
```

Figure 17. Sensitivity analysis using Wilcoxon's rank sign test

In Figure 17, the value of Gamma is interpreted as the odds of treatment assignment hidden bias. A change in the odds lower/upper bounds from significant to non-significant (or vice-versa) indicates by how much the odds need to change before the statistical significance of the outcome shifts. For example, in Figure 17, the lower bound estimate changes from non-significant (0.0541) to significant (0.0194) when gamma is 1.3. That is, a change of 0.3 in the odds will produce a change in the significance value. Rosenbaum (2002) defines a study as sensitive if values of Gamma close to 1 lead to changes in significance compared to those that could be obtained if the study is free of bias. Thus results will be more robust to hidden bias, if a very large change in the odds is needed before a change in statistical significance happens.

Limitations

In spite of the benefits described above, propensity scores have important limitations. Endogeneity problems are not controlled. That is, researchers still need to be able to identify and measure many if not all the variables associated with treatment

assignment/selection. There is a possibility that one can use proxies to address some of these variables; for example, using age as a proxy for general state of health (Gagne, 2010). However, it is not clear to what extent proxies can alleviate this problem. Rosenbaum (2005) and Guo and Fraser (2015) suggest the use of sensitivity analyses to explore the extent to which the results can be trusted as identification of all associated variables is unlikely.

Also important is the fact that there needs to be a strong overlap of the distributions of propensity scores between the two groups (the so-called *common support region*). If the overlap is small, that there may not be enough participants in the control group to match all the participants in the treatment group, then propensity score matching will be no better than any standard form of matching.

Given that lack of overlap can be most times be associated with a similar (or smaller) sample for the control group, it is recommended that the sample size for the control group be at least 3-4 times larger than for the treatment group to assure matches in the common support region. A larger sample for the control group also increases the number of matches for every treatment participant.

Randomized Control vs. Propensity Scores

The question of whether propensity scores can remove selection bias, and thus represent a viable option for Randomized Control Trials is still unresolved. To date, there is no clear indication of whether propensity scores can remove the selection bias that jeopardizes quasi-experiments. Using a set of studies in the medical field intended to measure the effectiveness of on-pump versus off-pump coronary artery bypass grafting, Olmos & Govindasamy (2014) conducted a meta-analysis that showed no differences in the estimated treatment effect size between randomized control trials and studies using propensity score matching. This finding seems to support the claims that propensity score matching produces similar results to those from randomized control studies. This is an area where more research is clearly needed.

Conclusion

Propensity scores can provide an alternative that can strengthen quasi-experiments and observational studies in their quest to demonstrate causality. In particular, they are intended to identify the probabilities associated with assignment to treatment conditions, and match participants based on those probabilities. This matching in particular helps directly with one of the four requirements associated with causal inference (the counterfactual), and indirectly with another (ruling out alternative explanations). However, the use of propensity scores requires a deep understanding and measurement of all the variables that can affect selection into groups. Furthermore, if any variable that can be critical for the selection into treatment is not included in the propensity scores, then the propensity scores will not be able to eliminate selection bias. Finally, a sensitivity analysis is always recommended as a way to determine how robust the results are.

References

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, *26*, 734-753.

Austin, P.C. (2008). A critical appraisal of propensity score matching in the medical literature between 19996-2003. *Statistics in Medicine*, *27*, 2037-2049. doi: 10.1002/sim.3150

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*, 399-424. doi: 10.1080/00273171.2011.568786

Bai, H., & Clark, M. H. (2012, October). *Propensity score matching: Theories and Applications*. Workshop presented at the American Evaluation Association, Minneapolis, MN.

Bowers, J., Fredrickson, M., & Hansen, B. (2014). RItools: Randomization Inference Tools. R package version 0.1-12.

Bonell, C. P., Hargreaves, J., Cousens, S., Ross, D., Hayes, R., Petticrew, M., & Kirkwood, B. R. (2009). *Journal of Epidemiology Community Health*, 1-6. doi: 10.1136/jech.2008.082602

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of*

Economic Surveys, *22*(1), 31-72. doi: 10.1111/j.1467-6419.2007.00527.x

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. United States of America: Houghton Mifflin Company.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Indian Journal of Statistics Series*, *35*(4), 417-446.

Cook, T. D., & Campbell, D. T. (1979). Quasi - experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin Company.

D'Agostino, R. B., & D'Agostino, R. B. (2007). Estimating treatment effects using observational data. *Journal of American Medical Association*, *297*(3), 314-316.

Drake, R. E., Goldman, H. H., Leff, H. S., Lehman, A. F., Dixon, L., Mueser, K. T., & Torrey, W. C. (2001). Implementing evidence-based practices in routine mental health service settings. *Psychiatric Service, 52*(2), 197-182.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. (3rd ed.). United States of America: John Wiley & Sons, Inc.

Gagne, J. J. (2010). High-dimensional propensity scores for comparative effectiveness research. Presentation at the Lewin Summit, June 15, 2010

Gliner, J. A., Morgan, G. A., & Leech, N. L. (2009). *Research methods in applied settings* (2nd. Ed). Mahwah, NJ: Lawrence Erlbaum.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4), 405-420.

Guo, X. S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Guskey. T. (1999). The age of our accountability. *Journal of Staff Development*, 19(4), 36-44.

Hansen, B. B., Fredrickson, M., Bertsekas, D., & Tseng, P., (2013) Package optmatch. R package version 0.8-1

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American Statistical Association, 99*(467). doi: 10.1198/016214504000000647

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science, 23*(2), 219-236. doi:10.1214/08-STS254

Harrell, F. E. (2015). Hmisc: Harrell Miscellaneous. R package version 3.15-0

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1-28.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Holmes, W. M. (2014). *Using propensity scores in quasi-experimental design*. United States of America: Sage Publication, Inc.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(10), 5-86. doi: 10.1257/jel.47.1.5

Keele, L.J. (2015). Rbounds: An R Package For Sensitivity Analysis with Matched Data. R. package version 2.1

Lechner, M. *(*2008*).* A note on the common support problem in applied evaluation studies. *Econometric Evaluation of Public Policies: Methods and Applications, 91/92, 217-235.*

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403-425. doi:10.1037/1082-989X.9.4.403

Morgan S. L., & Winship, C. (2012). *Counterfactuals and causal inference: Methods and principles for social research*. New York: Cambridge University Press.

Olmos, A. & Govindasamy, P. (2014). *Randomized experiments vs. Propensity scores matching: A Meta-analysis*. Paper presented at the American Evaluation Association, Denver, CO.

R Core Team (2014). R: A language and environment for statistical computing. (3.0.3) [Computer software]. Vienna, Austria: Foundation for Statistical Computing.

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

Rosenbaum, P.R., & Rubin, D.B. (1985).Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician,* 39, 33–38.

Rosenbaum, P. R. (2002). *Observational studies*. NY: Springer

Rosenbaum, P. R. (2005). Observational Study. In Everitt, B. S., & Howell, D. C. (3rd ed.), *Encyclopedia of Statistics in Behavioral Science* (pp. 1451-1462). Chichester: John Wiley & Sons.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*(366), 318-328.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322-331.

Scriven, M. (1991). *Evaluation Thesaurus*. Thousand Oaks, CA: Sage

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software, 42*(7), 1-52.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston: Houghton Mifflin Company.

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental design: Matching methods for causal inference. In Osborne, J. *Best Practices in Quantitative Methods* (pp. 155-177). Thousand Oaks, CA:Sage.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1-21.

Trochim, W. M. K. (1984). *Research design for program evaluation*. Thousand Oaks, CA: Sage.

Weiss, C. H. (1998). *Evaluation: Methods for Studying Programs and Policies*. Upper Saddle NJ: Prentice Hall

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics*, *86*(1), 91-107.