Propensity Score-Matching Methods for Observational Studies: An Application to Stat 216 Data

Allison Theobold

Advisor: Laura Hildreth

April 2016

Abstract

Many fields of science are faced with the inability to perform randomized experiments, but wish to have the ability to estimate a treatment effect and make causal inference. Propensity score matching is a method that can be used in observational studies to obtain unbiased estimates of the treatment effect. In this paper we consider the theory behind utilizing propensity score matching to obtain these such estimates, as well as explain how to implement propensity score matching in `R` using the `Matching` package for data from Montana State University's Introductory Statistics curriculum.

Contents

1 Introduction

Under the general framework of statistical inference, causal inferences and unbiased estimates of treatment effects require randomization. An important problem in statistical inference is how to obtain these estimates and have the ability to make causal inferences when the treatment assignment is not random, and is instead observed. One method commonly used to obtain unbiased estimates of the treatment effect and allow for causal inferences in observational studies is propensity score matching.

Previously, propensity scores have been used in a variety of fields, such as, economics, and criminal justice. In these fields observational data are common as it is either unethical, impractical or impossible to randomly assign treatment groups. These analyses often cannot directly compare the treatment and the control groups as there are issues of self-selection bias or systematic differences between the groups, possibly due to a decision by the researcher in deciding what units were assigned to the treatment. A direct comparison of the treatment and control groups will not, therefore, yield an unbiased estimate of the true treatment effect.

In this paper we will provide a tutorial of the theory and implementation of propensity score matching. First, we will discuss the data motivating the method, followed by a section outlining the theory behind propensity score matching and how it allows us to obtain unbiased estimates of a treatment effect. Third, we will demonstrate how to implement propensity score matching using the `Matching` package in `R`, connecting the theory as we go. Lastly, we will discuss conclusions from implementing propensity score matching on these data, as well as further extensions that a researcher could consider.

Example

To illustrate propensity score matching, we use data from Montana State University's Stat 216 course. The data are from spring of 2015, where two different curricula were administered to students, the simulation based curriculum and the "traditional" curriculum. The simulation based curriculum is a hands-on simulation based course, taught in a interactive TEAL (Technology Enhanced Active Learning) classroom, with daily simulation based activities from a course pack, and very little lecture. The "traditional" curriculum is a lecture style course, what we will consider the control. This course was given in a typical lecture style classroom, using the Lock

[8] textbook, with daily lectures, as well as some weekly activities similar to the TEAL course.

For the semester in question our goal is to utilize propensity scores to construct treatment and control groups that are as similar as possible under all pre-treatment covariates and to estimate the overall effect of a TEAL course on a student's grade.

2 Theory and Definitions

Using the notation of Rosenbaum and Rubin (1983), consider a population of N units, and let i, $i = 1, ..., N$, denote the i^{th} observation of the N units under consideration. For each of the N units under consideration let y_1 be the response that would have resulted if the unit was assigned to the treatment ($z = 1$), and let y_0 be the response that would have resulted if the unit was assigned to the control ($z = 0$), where z is a treatment indicator. Under this formulation, the **average treatment effect** is the expected difference between y_1 and y_0 for each unit, or

$$E(y_1) - E(y_0), \tag{1}$$

where $E(\cdot)$ denotes expectation. In many observational studies, we instead only observe a response y_1 if the unit was assigned to the treatment and only observe a response y_0 if the unit was assigned to the control. Therefore, what we in fact observe is the following,

$$E(y_1|z = 1) - E(y_0|z = 0), \tag{2}$$

which does not equal (1) in general [1] because our samples are from the conditional distribution of y, given z, not the marginal distribution. The estimated treatment effect from the conditional distribution can be biased do to selection bias, and does not, in general, provide an unbiased estimate of the ATE. Our overall goal is to estimate the average treatment effect (1), using what we observe in our study (2), by utilizing propensity score matching. What follows are the general definitions and theorems that will allow us to estimate (1) from (2) using the propensity score.

In the introductory statistics data, the overall goal is to estimate the average effect of a TEAL

[1] Unless $y \perp z$, but in that case we don't need to consider propensity score matching, as our responses are independent of our treatment assignment.

curriculum, over a traditional curriculum, on each student's course grade. Each student is not randomly assigned into a specific course, but instead is observed to enroll in either one course or the other, leaving curriculum as observed. Additionally, it is possible that there are systematic differences between the students enrolled in the course, i.e. student registering for a TEAL (non-traditional) course could be more adapt learners, more likely to excel, or more invested in their education.

Generally, there are four steps associated with implementing propensity score matching to obtain unbiased estimates of the treatment effect. First, the researcher chooses what measured pre-treatment covariates she/he believes impacts the response. Then, using these covariates, the researcher models the propensity (probability) of an individual to have received the treatment based upon their observed covariates. Next, these propensities are then used to construct pairs or groups of individuals, using a variety of methods, all of which must pair at least one treated individual to one control individual, where their propensity scores are the "closest" that the algorithm can locate. Finally, the matches are then used to provide unbiased estimates of the treatment effect.

Step 1

We will begin our discussion first with selecting the pre-treatment covariates that are believed to impact the response. A pre-treatment covariate is measured by the researcher prior to administering the treatment. For example, in the introductory statistics data, some of the observed pre-treatment variables on each student are as follows:

- Previous semester's GPA (cumulative)
- Score on the math portion of the ACT
- Score on the math portion of the SAT
- Score on the math placement exam (MPLEX) administered at MSU

None of these variables are measured while the student is taking the course, as those would be post-treatment variables, and should not be controlled for. "It is generally not a good idea to control for variables measured after the treatment...controlling for a post-treatment variable messes up the estimate of total treatment effect, and also the difficulty of using regression on

"mediators" or "intermediate outcomes" [3]. Each student's vector of observed pre-treatment covariates can be any subset of the above list.

We will denote our vector of observed pre-treatment covariates by $\mathbf{x}$.

Definition 2.1. *Strong Ignorability*

Treatment assignment is strongly ignorable, given a vector of covariates $\boldsymbol{x}$*, if*

$$(y_1, y_0) \perp z|\boldsymbol{x},\ 0 < pr(z = 1|\boldsymbol{x}) < 1, \tag{3}$$

for all $\boldsymbol{x}$*, where* $\perp$ *represents independence.*

This definition says that the distribution of potential outcomes, (y_0, y_1), is the same across the levels of the treatment variable, $\mathbf{z}$, once we condition on the covariates, $\mathbf{x}$. In a designed study, we have randomly assigned our treatment and therefore, on average, the confounding covariates of the two will be the same. In an observational study, we are not guaranteed this nicety, instead we need to ensure that this "balance" of the pre-treament covariates is met. We can do this by adding all of the pre-treatment covariates that we believe impact the response into our model.

For example, in the introductory statistics students, we cannot control what students enroll in one curriculum over the other. We add all the pre-treatment covariates that we believe impact a student's course grade into our model, so that two students with the same values of pre-treatment covariates, $\mathbf{x}$, should ideally have the same probability of being in a TEAL classroom.

Step 2

Once a researcher has selected the pre-treatment covariates that impact the response, then we begin to consider constructing matches of treatment and control units. Matching each treatment observation to its nearest control observation on all covariates proves to be a difficult task. In step 3 we will discuss how we can simplify this matching problem, but first we need to define what a propensity score is.

Definition 2.2. *Propensity Score*

The conditional probability of assignment to the treatment ($z = 1$)*, given the covariates, be*

denoted by

$$e(\boldsymbol{x}) = pr(z = 1|\boldsymbol{x}),$$

The function $e(\boldsymbol{x})$ *is called the* ***propensity score****, or the propensity of unit* i *to be given the treatment, conditional on the observed covariates* $\boldsymbol{x}$*.*

We will outline in section 3 how propensity scores are obtained in practice.

The following theorems demonstrate the motivation for considering the propensity score and its ability to allow us to estimate the average treatment effect and make causal inferences.

Theorem 2.3. *Rosenbaum & Rubin (1983, Theorem 1)*
Treatment assignment and the observed covariates are conditionally independent given the propensity score, that is

$$\boldsymbol{x} \perp \boldsymbol{z}|e(\boldsymbol{x}).$$

This theorem implies that if a subclass of units or a matched pair of units are homogeneous in $e(\mathbf{x})$ then the subclass or pair of units will have the same distribution of $\mathbf{x}$. For example, if we find students in the TEAL and traditional courses with the same propensity score, in theory, the distribution of the pre-treatment covariates for these students should be the same. Or, to break it down further, if two students are found to have the same propensity score (probability of enrolling in a TEAL course), then their previous semester GPA's, math SAT scores, etc. are assumed to come from the same distribution.

Theorem 2.4. *Rosenbaum & Rubin (1983, Theorem 2)*
If treatment assignment is strongly ignorable given $\boldsymbol{x}$*, then it is strongly ignorable given the propensity score* $e(\boldsymbol{x})$*; in other words,*

$$(y_1, y_0) \perp \boldsymbol{z}|\boldsymbol{x},\ 0 < pr(\boldsymbol{z} = 1|\boldsymbol{x}) < 1,$$

for all $\boldsymbol{x}$ *implies*

$$(y_1, y_0) \perp \boldsymbol{z}|e(\boldsymbol{x})$$

and

$$0 < pr(\boldsymbol{z} = 1|e(\boldsymbol{x})) < 1$$

for all $e(\boldsymbol{x})$.

The outline of this proof can be found in [1], however Theorem 2.4 follows directly from Theorem 2.3 and the definition of strong ignorability. The previous discussions of our example are echoed here, piecing the definition of strong ignorability and conditional independence of treatment and covariates together. Now we have a result stating that the distribution of the potential outcomes, (y_0, y_1), is the same across the levels of the treatment variable, $\mathbf{z}$, once we condition on the propensity score, $e(\mathbf{x})$.

Previously, we discussed that under strongly ignorable treatment assignment, two students with the same values of pre-treatment covariates $\mathbf{x}$ will have the same probability of enrolling in a TEAL class. Theorem 2.4 gives us that if two students have the same values of propensity scores $e(\mathbf{x})$ they will have the same probability of enrolling in a TEAL class. Matching students on every pre-treatment covariate is difficult, but matching students on a propensity score can be much simpler. These ideas lead us to the theory behind matching on a propensity score, which we will discuss in the next step.

Step 3

Now, suppose a specific value of $\mathbf{x}$, say $\mathbf{x}_0$, is sampled from the entire population of units, both treatment and control, and then a treatment and a control unit are found to both have this value of $\mathbf{x}$. In this two-step sampling process, the expected difference in the responses is

$$E_{\mathbf{x}_0}\{E[y_1|\mathbf{x}_0, z = 1] - E[y_0|\mathbf{x}_0, z = 0]\}, \tag{4}$$

where $E_{\mathbf{x}_0}$ denotes the expectation with respect to the distribution of $\mathbf{x}_0$ for the entire population of units. Under strongly ignorable treatment assignment, then (4) can be written as

$$E_{\mathbf{x}_0}\{E[y_1|\mathbf{x}_0] - E[y_0|\mathbf{x}_0]\},$$

as $(y_0, y_1) \perp \mathbf{z}|\mathbf{x}$. By the law of iterated expectation, the above result leads to an estimation of

the average treatment effect (1).

Suppose that we sample a specific vector of previous semester GPA, math ACT, and grade in last math/stat course from the population of all students. Then, we find both a student from a TEAL course and a student from a traditional course that have this specific vector of covariates. Under strongly ignorable treatment assignment, the differences in the course grades of the TEAL and traditional students is an unbiased estimate of the true effect of the TEAL curriculum for that vector of covariate values, $\mathbf{x}$. However, as mentioned previously, matching on p covariates can prove very difficult. Theorem 2.4 suggests that if we are willing to assume strongly ignorable treatment assignment matching on $\mathbf{x}$ is equivalent to matching on the propensity score $e(\mathbf{x})$, a far simpler method.

Now suppose instead that a specific value of a propensity score, $e(\mathbf{x})$, say $e(\mathbf{x}_0)$, is sampled randomly from the entire population of units, and then both a treatment and a control unit are randomly sampled from all units having that propensity score. The propensity score for the sampled treatment and control units will be equal, but they will perhaps have differing values of pretreatment covariates, $\mathbf{x}$.

If, once again, we have strongly ignorable treatment assignment, then by Theorem 2.4 it follows that,

$$E[y_1|e(\mathbf{x}_0), z = 1] - E[y_0|e(\mathbf{x}_0), z = 0] = E[y_1|e(\mathbf{x}_0)] - E[y_0|e(\mathbf{x}_0)]. \tag{5}$$

Conditioning on the population distribution of $e(\mathbf{x}_0)$ we have

$$\begin{aligned} E_{e(\mathbf{x}_0)}[y_1|e(\mathbf{x}_0), z = 1] - E[y_0|e(\mathbf{x}_0), z = 0] &= E_{e(\mathbf{x}_0)}[y_1|e(\mathbf{x}_0)] - E[y_0|e(\mathbf{x}_0)] \\ &= E[y_1] - E[y_0]. \end{aligned}$$

Therefore, under strongly ignorable treatment assignment, units with the same propensity score but different values of covariates can act as counter-factuals for each other, relating what would have happened had the other been assigned a different treatment. In other words, the

expected differences between these treatment and the control units is an unbiased estimate of the average treatment effect.

Similar to the previous discussion, in this process we draw a random sample of a propensity score, $e(\mathbf{x})$, from the population of propensity scores, and both a TEAL student and a traditional student are randomly sampled with that specific propensity score. Then, once again, under strongly ignorable treatment assignment the the differences in the course grades of the TEAL and traditional students is an unbiased estimate of the true effect of the TEAL curriculum for that value of $e(\mathbf{x})$.

Step 4

The above discussion of estimability of the average treatment effect for each of the two scenarios was motivated by the following theorems.

Theorem 2.5. *Rosenbaum & Rubin (1983, Theorem 4)*
Suppose treatment assignment is strongly ignorable and $e(\boldsymbol{x})$ is a propensity score. Then the expected difference in observed responses to the two treatments at $e(\boldsymbol{x})$ is equal to the average treatment effect at $e(\boldsymbol{x})$, that is,

$$E[y_1|e(\boldsymbol{x}), z = 1] - E[y_0|e(\boldsymbol{x}), z = 0] = E[(y_1 - y_0)|e(\boldsymbol{x})].$$

Corollary 2.6. *Rosenbaum & Rubin (1983, Corollary 4.1)*
Suppose treatment assignment is strongly ignorable. Further suppose that a value of a propensity score $e(\boldsymbol{x})$ is randomly sampled from the population of units, and then one treated ($z = 1$) and one control ($z = 0$) unit are sampled with this value of $e(\boldsymbol{x})$. Then the expected difference in response to the two treatments for the units in the matched pair equals the average treatment effect at $e(\boldsymbol{x})$. Moreover, the mean of the matched pair differences obtained by this two-step sampling process is unbiased for the average treatment effect (1).

Corollary 2.7. *Rosenbaum & Rubin (1983, Corollary 4.2)*
Suppose treatment assignment is strongly ignorable. Suppose further that a group of units is sampled using $e(\boldsymbol{x})$ such that:

i) $e(\boldsymbol{x})$ is constant for all units in the group, and

ii) at least one unit in the group received each treatment.

Then, for these values, the expected difference in treatment means equals the average treatment effect at that value of $e(\boldsymbol{x})$. Moreover, the weighted average of such differences, that is, the directly adjusted difference, is unbiased for the treatment effect (1), when the weights equal the fraction of the population at $e(\boldsymbol{x})$.

Conclusions

The above theorems and definitions allow us to take our preliminary discussion of what we observe in our observational study framework (2) and estimate the average treatment effect (1). These theorems and definitions demonstrate the theory behind propensity score matching (PSM), but do not address how we can implement PSM to obtain unbiased estimates of the average effect of a TEAL curriculum on a student's course grade. In the next section we address implementing PSM, the steps and methods that a researcher needs to know in order to obtain an unbiased estimate of the average treatment effect.

3 Implementing Propensity Score Matching

Introduction

In this section we illustrate the use of the `R` package `Matching` [7] to lead our discussion of how to implement propensity score matching in practice, and to provide a variety of examples of different matching methods. The steps necessary to obtain an unbiased estimate of the average treatment effect are as follows:

Step 1: Choose the covariates that the researcher believes impact the response

Step 2: Use these covariates to estimate propensity scores for each student

Step 3: Use the propensity scores to construct matches of TEAL students and traditional students

Step 4: Using these matches, estimate the average treatment effect of the TEAL curriculum on course grade

In the following sections we will discuss each of these steps, why they are necessary and how to implement them in `R`.

Selecting Covariates

To begin this process, we first need to select the pre-treatment covariates that we believe impact the response variable. As previously mentioned, these data consist of 4 measured pre-treatment covariates. For simplicity and to illustrate how to implement PSM we consider just a single covariate, previous semester's cumulative GPA, and later generalize to multiple covariates.

We will use a data set from the spring 2015 semester of Stat 216, named `stat216_s15`. The `Matching` package in `R` requires that we store our data as follows:

- A vector of responses (each student's Stat 216 course grade)

```
Y <- stat216_s15$GPAgrade
```

- A vector of treatments (which curriculum each student took), which must be converted to binary

```
Tr <- stat216_s15$curric
Tr <- ifelse(Tr == "MSU.0", 1, 0)
## converting to binary: simulation = 1, traditional = 0
```

Estimating the Propensity Score

To obtain the propensity scores for each of the students in our dataset, we need to model the probability that they were given the treatment (TEAL curriculum) given their vector of covariates. A propensity score is typically modeled using a regression with a logit link, although any model that returns a value between 0 and 1 could be used, such as a probit or cumulative log-log link. Using the logistic regression, we have the following model model of the probability of each individual to enroll in a TEAL course,

$$\begin{aligned} logit(e(\mathbf{x}_i)) &= \beta_0 + \beta_1 \text{Previous Semester GPA}_i \quad \text{or} \\ log\left[\frac{e(\mathbf{x})}{1-e(\mathbf{x})}\right] &= \beta_0 + \beta_1 \text{Previous Semester GPA}_i \end{aligned}$$

where previous semester's GPA is our only predictor. These propensity scores are estimated using the `glm` function in `R` as follows.

```
glm1 <- glm(Tr ~ PREVGPA , family = "binomial", data = stat216_s15)
```

Matching with Propensity Scores

Taken literally, Corollary 2.6 and 2.7 require *exact* matching on the propensity score. This can be difficult in practice, which leads to the divide between the theory of propensity score matching and the practice. The objective that we instead consider, is to construct matched treatment and control units whose propensity scores are as a close as possible, given the data, so that the assumptions of Corollary 2.14 and 2.15 have been approximately met.

When implementing matching, the goal is to pair each TEAL student with the nearest traditional student, on every covariate. In constructing these matches of traditional and TEAL students, we wish to "even out" the distributions of each of the pre-treatment covariates. This matching problem is not something that needs to be considered in randomized experiments, as randomization, on average, "evens out" the distribution of each of the pre-treatment covariates. The benefits of Theorem 2.4 become more clear as we consider tackling this matching problem. To "even out" the distributions of each covariate for the traditional and TEAL students we can match the students on **each** covariate, or, using Theorem 2.4, equivalently we can match the students on propensity scores.

In order to perform matching using our estimated propensity scores, there are two main questions to consider in choosing a matching algorithm: to match with or without replacement, and how many control units should be matched with each treatment unit. Below we describe the common matching algorithms used and how they are implemented in `R`.

One-to-one Matching without Replacement

First we consider matching without replacement and one-to-one matching. In matching without replacement, each treated unit is matched with the nearest control unit sequentially through the data, i.e. the first TEAL student in the data set is paired with the traditional student with the nearest propensity score, within a specified tolerance. If the algorithm cannot locate a traditional student within the tolerance, that TEAL student is omitted from the analysis, and it moves on

to the next TEAL student in the data set. This process continues until every TEAL student has been considered. Under matching without replacement, once a traditional student has been paired with a TEAL student they are unable to be matched with a different TEAL student. The tolerance allowed for matching is specified by the researcher, and if there are no traditional students with propensity scores within that tolerance of a TEAL student, that TEAL student is discarded and not used in our analysis.

Issues can arise using this method. If there are few traditional students similar to the TEAL students, we may pair students that have quite different propensity scores. These matches would be required to be within the specified tolerance, but the distances between the propensity scores may be further than what would have occurred if all traditional students were available to match with (instead of the smaller pool of traditional students). Additionally, in matching without replacement, there is the possibility that the results are sensitive to the order in which the matches are made. The matches are made sequentially through the data set, so if we sort the TEAL propensity scores in descending order we may obtain different matches that what would have occurred if they were sorted in ascending order. This problem arises due to the removal of the traditional students after matching, where the TEAL students at the end of the data set have fewer traditional students to choose from in constructing matches. Depending on the overlap of the TEAL and traditional student's propensity scores, different orderings of the data set can make substantial differences in the matches constructed. This topic is discussed further in section 4.

One-to-one matching without replacement in the `Matching` package is implemented by declaring the following arguments,

- `replace` - if matching should be done with/without replacement (T/F)
- `ties` - determines if ties will be allowed in the matched data set, if true, the matched dataset will include the multiple matched control observations and the matched data will be weighted to reflect the multiple matches, while if false, the ties will be randomly broken
- `M` - the number of matches that should be found for each treated unit

Below we have a display of the `Match` function, which takes in responses, treatments, and propensity scores, along with the above mentioned matching arguments.

```
match1 <- Match(Y = Y, Tr = Tr, X = glm1$fitted.values,
                estimand = "ATE", ties = FALSE,
                replace = FALSE, M = 1)
```

The `Match` function provides for us the matches of traditional and TEAL students, from which we can estimate the average treatment effect of a TEAL curriculum on a student's grade. The function also allows for us to check the "balance" of each of the covariates that we used to model the propensity scores. Checking the "balance" of our pre-treatment covariates provides a visual assessment of our strong ignorability assumption. In assuming that treatment assignment is strongly ignorable, we assume that the distribution of each covariate is the same between our treatment and control, Definition 2.1. For our data, this consists of the distribution of previous semester's GPA being the same for the TEAL and traditional students.

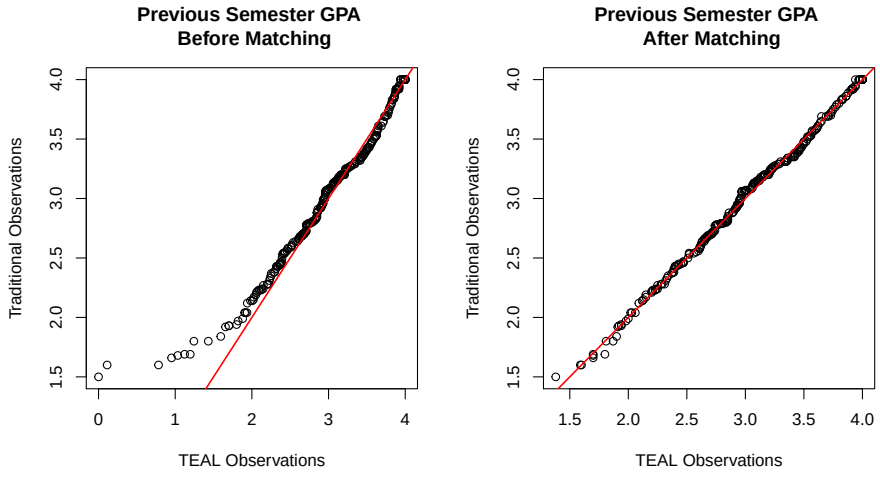
```
##
## ***** (V1) PREVGPA *****
##                           Before Matching     After Matching
## mean treatment........     3.0117          3.0117
## mean control..........     2.9566          3.0037
## std mean diff.........     9.1953          1.3339
##
## mean raw eQQ diff.....   0.090227       0.052197
## med  raw eQQ diff.....       0.06           0.03
## max  raw eQQ diff.....        1.6            1.5
##
## mean eCDF diff........   0.026066       0.017443
## med  eCDF diff........   0.026552       0.015152
## max  eCDF diff........   0.056793       0.056818
##
## var ratio (Tr/Co).....    0.70109        0.81031
## T-test p-value........      0.318        0.28571
## KS Bootstrap p-value..      0.661          0.317
## KS Naive p-value......    0.75137        0.36153
## KS Statistic..........   0.056793       0.056818
```

The above output displays the mean of the TEAL students previous semester's GPA (treatment) verses the mean of the traditional students previous semester's GPA (control), followed

by the mean of the raw differences (standardized differences) [2], as well as the mean of the differences on the GPA scale. It separates the output into two columns, before matching, which is the observed differences in the raw data, and after matching, the difference on the matched data, for which some TEAL students may no longer be included.

To satisfy the strong ignorability assumption, ideally the ratio of variances between treatment and control should be close to 1 and the difference in the means should be close to 0. The balance output provides summary measures of the distributions of previous semester's GPA for the traditional and TEAL students. Under these distributional assumptions the t-test for a difference in means should return a "large" p-value (test for a difference in means of previous semester's GPA), and the KS (Kolmogorov-Smirnov) statistic should be close to 0 with also a "large" p-value (testing for a difference in distributions of previous semester's GPA between traditional and TEAL). [3]

We see in the summary output, small gains in the proximity of the means of traditional and TEAL previous semester's GPA, as well as the ratio of the variances. Prior to matching, the means already were fairly close with a variance ratio near 0.7, indicating that we may not have had a serious violation of our strong ignorability assumption prior to matching.

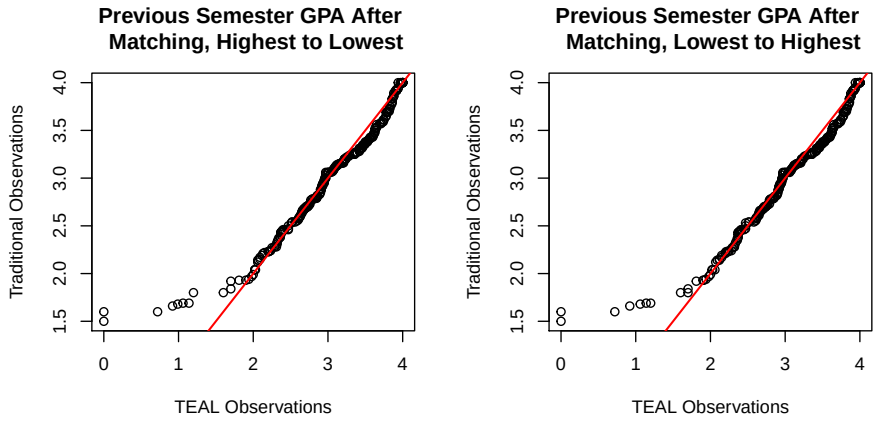


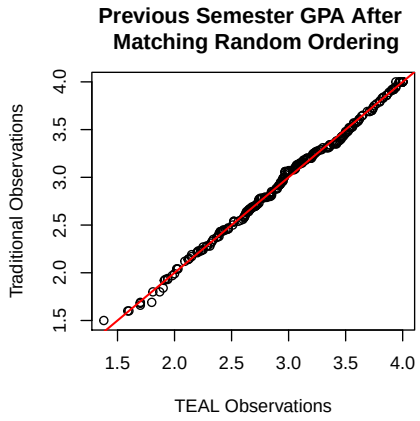
[2] It appears that the package is not subtracting the mean of the control from the mean of the treatment, as the documentation describes. I have contacted the authors of the package to inform them of this error.

[3] To the best of my knowledge the Kolmogorov-Smirnov test is bootstrapping means for both the treatment and the control groups, using the specified number of bootstraps, and then comparing the centers of the two bootstrap distributions. The KS statistic is a comparison of cumulative distribution functions of the treatment and control, and the test statistic is the maximum difference in value. Exact p-values are obtained as described in Marsaglia, Tsang & Wang (2003)

The plots displayed above plot the quantiles of the two data sets, TEAL students previous semester's GPA and traditional students previous semester's GPA, against each other. Prior to matching, large differences are seen in the lower levels of previous GPA between TEAL and traditional students in the original data. This can be attributed to 10 of the traditional students having lower previous semester's GPA's than any of the TEAL students. After matching, there appears to be a large improvement in the balance of previous GPA between TEAL and traditional, with large gains in the lower levels. For the visual assessment of our matching algorithm, we wish to see matches falling on or near the QQ-line, indicating the traditional students matched are similar to the TEAL students.

Now let us consider the order in which we match the propensity scores. As we are matching without replacement, it is possible that propensity scores matched earlier will be less biased. For the plots below, propensity scores were ordered in increasing or decreasing order before the matching occurred.





The plots do not show large differences in the matches when the propensity scores are ordered

least to greatest or visa verse. But, there is a substantial difference in the matches achieved by random ordering, a similar result to that found by Dehejia and Wahba [5].

One-to-one Matching with Replacement

The second option is to match the treated units with the control units, while allowing for replacement. Each TEAL student will, once again, be matched with the traditional student with the closest propensity score, however, now we will allow for the possibility that a traditional student can be paired with more than one TEAL student. This method can potentially minimize the distance in the propensity scores of the matches, and has the potential to reduce the bias of the pre-treatment covariates far more than matching without replacement, as we are allowing for a larger pool of traditional students to be paired with each TEAL student. It is for this reason that the default algorithm in the majority of propensity score packages for `R, SAS,` and `STATA` use matching with replacement.

Similar to the method of matching without replacement, using the `Matching` package, the only argument modifications are to the `ties` and `replace` pieces. The `replace` argument is now set to true, as we are matching with replacement, and the `ties` argument is now set to true, as we are allowing for the treatment observations to be matched twice (or more) if there are ties in propensity scores. The matching is still one-to-one, so the `M` argument has not changed. Below is the code for matching with replacement solely on previous semester's GPA.

```
match2 <- Match(Y = Y, Tr = Tr, X = glm1$fitted.values, estimand = "ATE", ties = TRUE,
                replace = TRUE, M = 1)
```

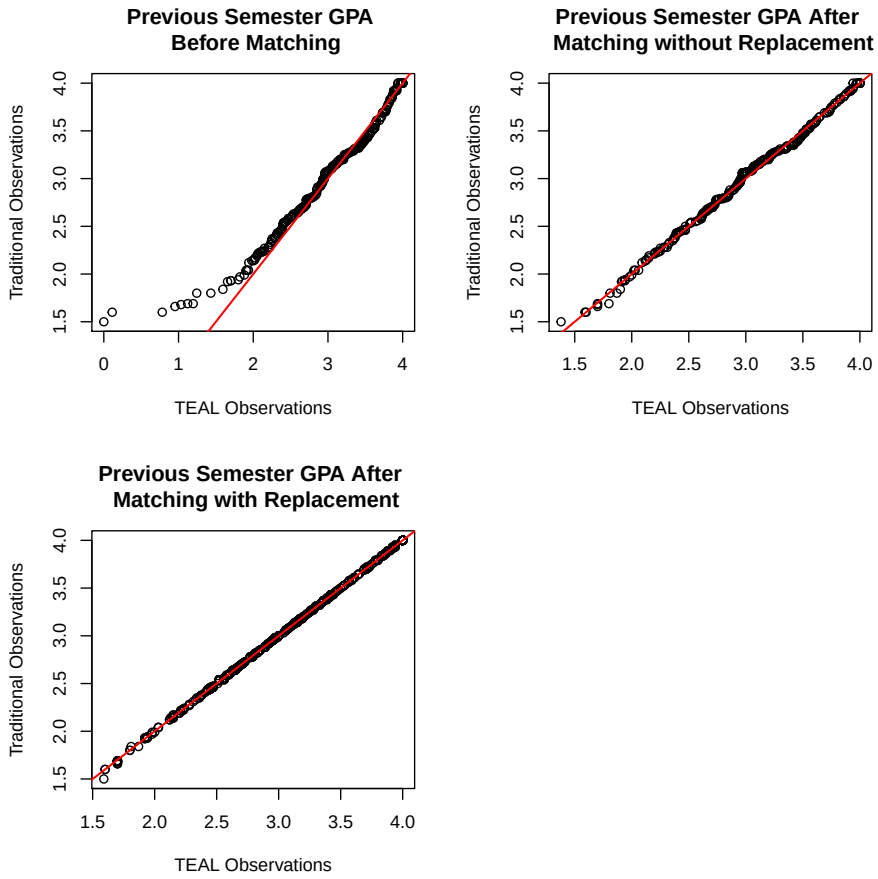
Once again, now that we have performed the matching, we can check the balance of previous semester's GPA between the two groups, using the `MatchBalance` function.

```
##
## ***** (V1) PREVGPA *****
##                            Before Matching     After Matching
## mean treatment........     3.0117          2.9933
## mean control..........     2.9566          2.9825
## std mean diff.........     9.1953          1.7133
```

```
##
## mean raw eQQ diff.....   0.090227      0.0078243
## med  raw eQQ diff.....       0.06              0
## max  raw eQQ diff.....        1.6            1.5
##
## mean eCDF diff........   0.026066      0.0017202
## med  eCDF diff........   0.026552      0.0013514
## max  eCDF diff........   0.056793       0.012162
##
## var ratio (Tr/Co).....    0.70109        0.90214
## T-test p-value........      0.318       0.013624
## KS Bootstrap p-value..      0.667          0.998
## KS Naive p-value......    0.75137         0.9999
## KS Statistic..........   0.056793       0.012162
```

Similar to before, matching has made improvements in the balance of the previous semester's GPA of the TEAL and traditional students. Again, the observed difference in mean previous semester's GPA is not large, but allowing for matching with replacement the difference in means is nearly negligible and the ratio of the variances is nearly 1.

In the plots below, we see a comparison of the observed TEAL verses traditional previous GPA's, after matching without replacement, and after matching with replacement. We see slight differences in the plot of the matches done with replacement, notably the lowest observed previous GPA's for TEAL students have been matched with traditional students with more similar GPA's. Due to the ability to sample with replacement, these lower values of previous GPA have a larger selection of traditional students to match with, decreasing the bias.



Next, we consider the question of how many control units should be paired with each treatment unit. The two choices we will consider for many-to-one matching are the nearest-neighbor and caliper matching methods.

Caliper Matching

In caliper matching, all the control units within a predetermined "caliper" or radius of each treated unit's propensity score will be matched with that treated unit. Similar to before, if there are no traditional students within the specified caliper of each TEAL student, then no matches are made and that TEAL student is omitted from the analysis. The main advantage of caliper matching is that it does not require there to be a specific number of matches, allowing for more traditional students to be paired with a TEAL students when there are many close by, and fewer traditional students being paired with a TEAL student when there are few close by.

In choosing a caliper value, careful inspection of the range of propensity scores and the pair-wise distances between the propensity scores should be used. In these data, the range of

propensity scores is minimal, nearly 0.25, with the majority of the propensity scores only 0.0005 from their nearest counterpart. However, in some data the range of propensity scores could be near 1, with large gaps between each propensity score and its nearest counterpart. Choosing a small caliper in this case could potentially lead to few matches being made, placing a large constraint on our scope of inference.

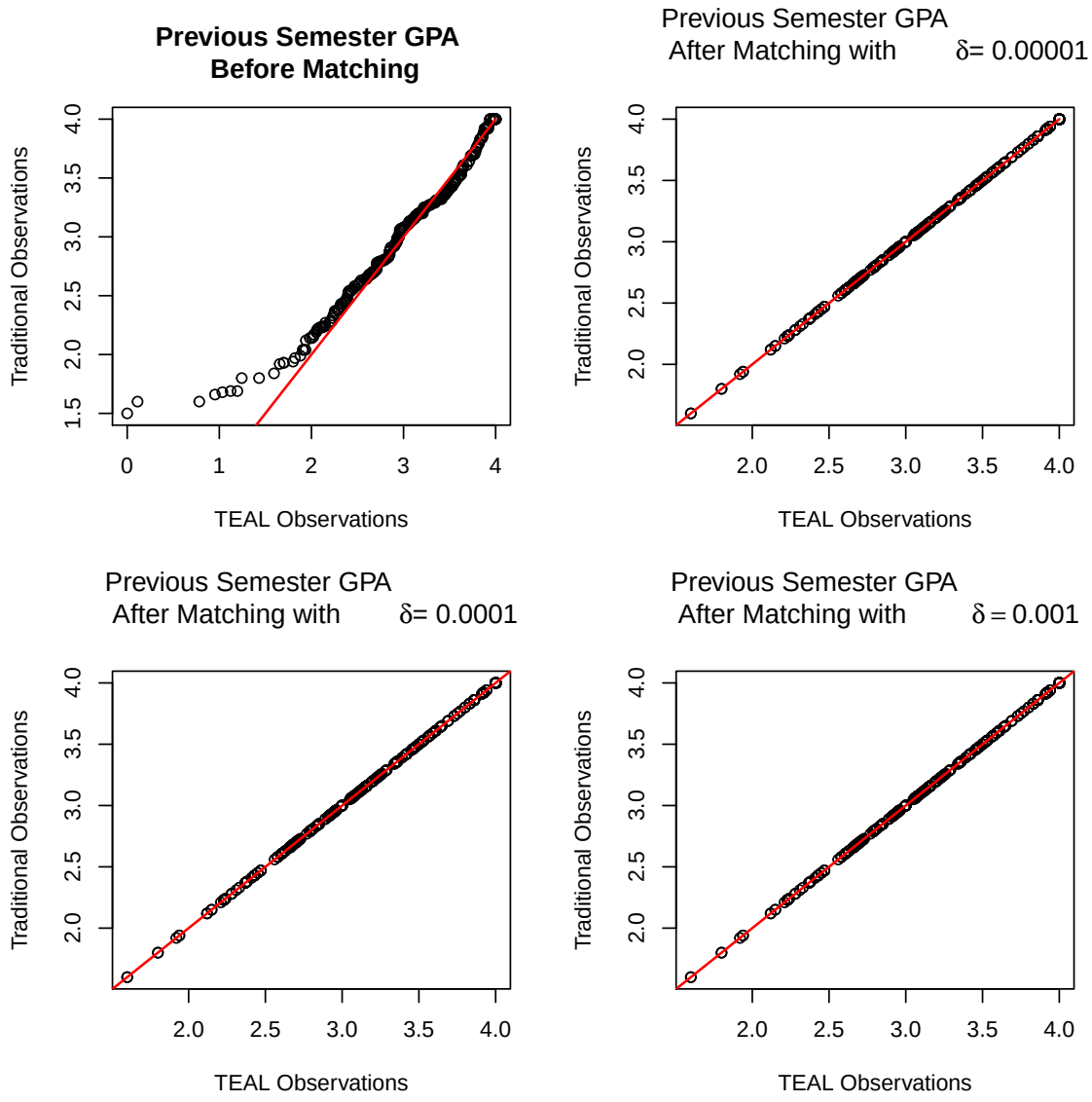
In using `Matching`, the only argument modification is to add the `caliper` argument. We are utilizing sampling with replacement to construct our caliper matches, allowing for ties of matches. We experiment with three caliper choices, similar to the calipers used in [5], of $\delta = 0.00001, \delta = 0.0001$, and $\delta = 0.001$.

```
match6 <- Match(Y = Y, Tr = Tr, X = glm1$fitted.values, estimand = "ATE", ties = TRUE,
                replace = TRUE, M = 1, caliper = 0.00001)

match7 <- Match(Y = Y, Tr = Tr, X = glm1$fitted.values, estimand = "ATE", ties = TRUE,
                replace = TRUE, M = 1, caliper = 0.0001)

match8 <- Match(Y = Y, Tr = Tr, X = glm1$fitted.values, estimand = "ATE", ties = TRUE,
                replace = TRUE, M = 1, caliper = 0.001)
```

The plots show almost no noticeable difference between the three caliper values. It does appear from the plots that the caliper matching has in fact perfectly balanced the TEAL and traditional values of previous GPA.



The below summary output checking the balance of the values of previous GPA between TEAL and traditional students verifies, as we suspected, that our matching method has constructed matches so that the mean of the two groups is the same and the ratio of the variances is 1.

```
##
## ***** (V1) PREVGPA *****
##                         Before Matching     After Matching
## mean treatment........     3.0117          3.0972
## mean control..........     2.9566          3.0972
## std mean diff.........     9.1953               0
```

```
##
## mean raw eQQ diff.....   0.090227             0
## med  raw eQQ diff.....       0.06             0
## max  raw eQQ diff.....        1.6             0
##
## mean eCDF diff........   0.026066             0
## med  eCDF diff........   0.026552             0
## max  eCDF diff........   0.056793             0
##
## var ratio (Tr/Co).....    0.70109             1
## T-test p-value........      0.318             1
## KS Bootstrap p-value..      0.677             1
## KS Naive p-value......    0.75137             1
## KS Statistic..........   0.056793             0
```

Nearest Neighbor Matching

The nearest-neighbor method selects the M control units whose propensity scores are within the specified tolerance of the treatment unit in question. For each TEAL student, this method requires that they be paired with exactly M traditional students, all within a specified tolerance. This method has the potential to increase the bias of the pre-treatment covariates, as the M control units could have large differences in their pre-treatment covariates, and averaging over them could construct a poorer overall match. Similar to before, under the nearest-neighbor method, if there are not exactly M traditional students within the specified tolerance of the TEAL students, then that TEAL student is omitted from the analysis.

Using the `Matching` package, the only argument modification is to the `M` portion. We are utilizing matching with replacement to construct our nearest-neighbor matches, allowing for ties of matches. We will experiment with three arbitrary choices for M, 2, 4, and 6.

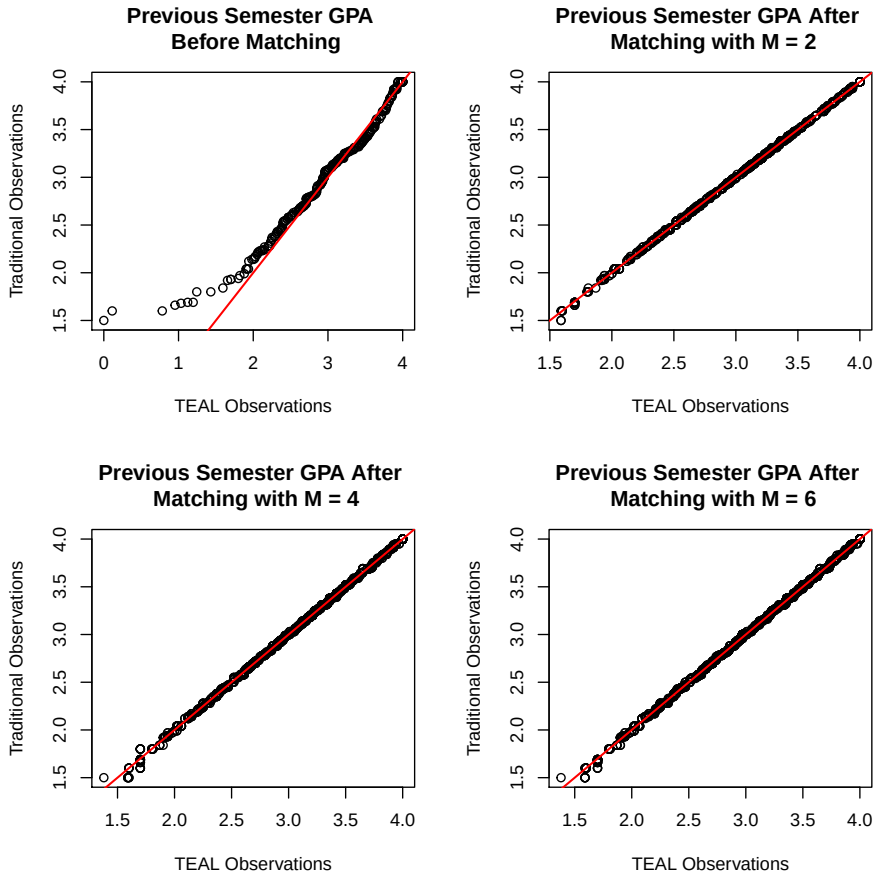
```
match3 <- Match(Y = Y, Tr = Tr, X = glm1$fitted.values, estimand = "ATE", ties = TRUE,
                replace = TRUE, M = 2)

match4 <- Match(Y = Y, Tr = Tr, X = glm1$fitted.values, estimand = "ATE", ties = TRUE,
```

```
                replace = TRUE, M = 4)

match5 <- Match(Y = Y, Tr = Tr, X = glm1$fitted.values, estimand = "ATE", ties = TRUE,
                replace = TRUE, M = 6)
```

In the below plots, there appear to be small differences in the matches made between the three values of M. The main differences, once again, are in the matches at the lower levels of previous semester's GPA. With $M = 2$ it appears that all but one of the matches lie on the one-to-one line, while for $M = 4$ and $M = 6$ there appears to be slightly more bias in the matches in the lower values of previous semester's GPA. This makes sense, as we are forcing there to be larger numbers of traditional observations per TEAL student, and taking the average of their previous semester's GPA's could lead to poorer balance in the matches. The overall matches between $M = 4$ and $M = 6$ appear nearly identical, less a few ties.



The below table gives a brief overview of the changes in the summary measures after matching under the three values of M.

Summary Measure	$M = 2$	$M = 4$	$M = 6$
Mean Treatment	3.0117	3.0117	3.0117
Mean Control	3.0125	3.012	3.0124
Std Mean Difference	-0.13822	-0.046849	-0.11808
Var Ratio (Tr/Co)	1.005	1.0042	1.0059
T-test p-value	0.22247	0.81932	0.64253

If we base our decision solely on the summary measures, the matches made with a nearest-neighbor value of 4 give the smallest difference of mean previous semester's GPA between the TEAL and traditional students (standard mean diff = -0.047, variance ratio = 1.004), however, the graphical display of the matches seems to imply that the matches made with $M = 2$ are the best. This sight decrease in the overall difference in the means could potentially be attributed to the larger sample size when using a nearest-neighbor value of 4 or 6.

Estimating the Average Treatment Effect

Lastly, using our matches we can calculate the estimated effect of the TEAL curriculum on a student's course grade. Each student's course grade has been converted to the GPA scale (A = 4.0, B = 3.0, etc.) so as to provide a rough numerical estimate of the ATE of the TEAL curriculum.

Using the `Matching` package, we now use the `summary` operator on the matches made. This will provide for us the ATE estimates and their standard errors. Below, we compare the estimated ATE for matching with replacement, matching without replacement, using a caliper of $\delta = 0.001$, and using a nearest neighbor of $M = 4$.

```
summary(Match1)
summary(Match2)
summary(Match4)
summary(Match8)
```

We see that the estimated treatment effect is very similar across the four matching methods, roughly 0.33, the difference between a B and a B+ in the course. It is worth noting that all

Matching Method	Estimated ATE	95% CI
Without Replacement	0.36515	(0.1771, 0.5532)
With Replacement	0.40869	(0.1831, 0.6342)
Caliper, $\delta = 0.001$	0.34771	(0.1689, 0.5265)
Nearest Neighbor, $M = 4$	0.37916	(0.1842, 0.5741)

of the intervals are greater than 0, implying that the TEAL curriculum has an overall positive impact on student's course grades, with estimated effects for some methods reaching nearly 0.6, the difference between a B and an A-.

4 Discussion and Conclusions

We have presented the theory of propensity score matching and how to implement PSM in `R`. We discussed the main methods to consider when performing propensity score matching, as well as presenting a method that is able to yield matches which eliminate all bias. Additionally, we demonstrated how to obtain estimates of the average treatment effect using these assorted matching methods.

The difference in mean previous GPA between TEAL and traditional students in the observed data is not large ($\bar{x}_{TEAL} - \bar{x}_{traditional} = 0.05514755$), with the largest differences occurring in lowest values of previous GPA (5^{th} percentile: TEAL = 1.9445, traditional = 1.802). The data have a slight lack of complete overlap, as there are 10 students in the traditional group with previous GPA's below that of the lowest TEAL student. Overall, the caliper matching with a δ of at least 0.001 performs the best, however, in implementation there could be large differences in sample sizes as δ increases.

For data that have a more prominent lack of complete overlap, we would recommend to stay away from any matching that involves the use of more than one control for each treatment, as this has shown to lead to more bias in the matches for data with only a slight lack of overlap. Additionally, we recommend allowing for matching with replacement, as this allows for the control observations closest to the threshold where the lack of overlap begins to be paired more than once, leading to less biased matches. Caution it to be heeded in choosing a matching algorithm, as for some data one algorithm may work better than for others.

Generalizations to Multiple Covariates

A topic associated with propensity score matching, not discussed previously in this paper is the matter of matching when multiple covariates are present. In all of the examples given in this section, a single covariate was considered, which is rarely the case in any analysis. Referring back to the theorems given in section 2, a key assumption of the process of propensity score matching is, conditional on the propensity score, the treatment assignment is *strongly ignorable*. This assumption is almost surely not met when an analysis only includes a single covariate, as responses, such as student success, naturally depend on multiple factors.

In the dataset for the spring 2015 introductory statistics course, many additional variables were recorded. Each of these covariates should be included in the model, as they are all pre-treatment covariates that impact a student's success in college courses, and potential exclusion from the model could bias the estimate of the overall treatment effect.

Additionally, by Theorem 2.7, given a propensity score, it is assumed that the distribution of $\mathbf{x}$ is the same for the treatment and the control. Therefore, balancing the observed covariates between the treatment and the control is a fundamental concept of propensity score matching. However, assessing balance on multiple covariates can become cumbersome and difficult. The methods described by Rosenbaum and Rubin [2] are intended to produce a model for the propensity scores, where each covariate is balanced. They fit a step-wise logistic regression with all main effects, n-way interactions, as well as quadratic and cubic terms. This saturated model is not used for statistical inference, only to insure that balance is reached in each of the pre-treatment covariates.

Lastly, when estimating the propensity scores for each student, a mixed effects model could be considered. Typically in the Stat 216 course there are multiple instructors for each curriculum, with each instructor usually only teaching a single course. If we believe that different instructors could potentially have different impacts on student's success in the course, it would be reasonable to include instructor as a random effect in our logistic regression, by which we obtain our

propensity scores.

Further Extensions

We have presented methods of balancing pre-treatment covariates, so as to provide unbiased estimates of the treatment effect, but have only estimated the effect by converting each student's course grade to the GPA scale. We have not discussed the possible models to fit to estimate this effect as a categorical variable, ranging from A to F and possibly including withdrawals (W) and incompletes (I). As the outcome, is ordered categorical (grade received) many different models could be fit. The most familiar model would be a proportional odds logit model, or a cumulative logit model. This would model the probability of a student receiving a grade of C or higher given their observed pre-treatment covariates. Other approaches could include reclassifying course grade as pass or fail and using a logistic regression to model the probability of passing for TEAL verses traditional students. Another potential model is an adjacent category model, which models the probability of achieving a grade of a B, conditional on them having achieved a C.

Additional extensions could include modeling more than one treatment verses a control, which also has the ability to utilize propensity score matching to achieve balance between all n groups. For example, in the spring of 2014 Montana State University was administering three different introductory statistics curricula.

References

[1] 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.

[2] Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Sub-classification the Propensity Score. *Journal of the American Statistical Association*, **79(387)**, 516-524.

[3] Gelman, A., Hill, J.(2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

[4] Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A Propensity Score Matching Analysis of the Effects of Special Education Services. *The Journal of Special Education*, **43(4)**, 236-254.

[5] Dehejia, R. H., Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics*, **84(1)**, 151-161.

[6] Kim, R. H. & Clark, D. (2013). The effect of prison-based college education programs on recidivism: Propensity Score Matching approach. *The Journal of Criminal Justice*, **41**, 196-204.

[7] Jasjeet S. Sekhon. 2011. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. *Journal of Statistical Software*, **42(7)**, 1-52.

[8] Lock R., Lock P. F., Lock E. F., Lock D. F., Morgan K. L. (2012). Unlocking the Power of Data. Wiley.