CAUSAL TEAMWORK: REGRESSION AND MATCHING AS SUPPLEMENTS IN THE ESTIMATION OF CAUSAL EFFECTS

The estimation of causal effects is guided by the potential outcomes framework. Regression analysis is often used to estimate causal effects from observational data, and matching methods are also gaining prominence. However, both methods are likely to produce biased and inconsistent estimators due to the violation of strong assumptions associated with approximating the potential outcomes framework using observational data. It is possible to use regression and matching methodologies in conjunction in order to make the estimation "doubly robust". This paper examines estimation of causal effects using propensity score matching methods as a supplement to regression. The paper reviews regression and matching methods in a potential outcomes framework. We then conduct two simulation studies that assess the performance of regression and matching methods as supplements in terms of reducing bias. Both simulation studies indicate that: (1) Regression alone performs best when one of the covariates is unobserved, (2) Regression with inverse propensity score weighting performs best by a margin when all covariates are observed, (3) Regression on a sample balanced by matching produces low bias when one or all covariates are observed.

1. INTRODUCTION

The potential outcomes or counterfactual framework is at the heart of causal effect estimation. The framework involves comparing hypothetical potential outcomes, in which the same individual receives all levels of treatment. Experimental methods can easily be used to compute treatment effects since the random assignment to treatment makes individuals in all treatment groups comparable. However, using observational data to approximate the potential outcomes framework often leads to violation of assumptions needed to produce unbiased and consistent estimators of the causal effect (Morgan and Winship 2012). Thus, estimating causal effects using observational data requires careful thought about estimation strategies and methods.

Regression analysis is a useful tool for predicting the outcomes given a set of covariates, but in order to obtain the least biased causal estimate, there needs to be more thought about the distribution of covariates in the data and the characteristics of individuals in different treatment groups. The ideal usage of matching methodologies involves repeatedly assessing the distribution of covariates and estimating the probability that an individual receives a level of treatment. The strong link to the potential outcomes framework of matching methods and the strong predictive power of regression analysis make a case for using both methods together to estimate causal effects. However, as is the case in any observational data analysis, the estimate of the effect must not be perceived as the true effect size and must undergo sensitivity analysis (Stuart 2010). The focus of this paper is on identifying a method that produces the least bias in the estimate of the causal effect, while taking the variance of the estimator into consideration.

The paper begins with a review of existing research on causal effect estimation using observational data. Section 3 discusses the potential outcomes framework, and its connections to regression and matching methods. It explains the link between regression and matching methods, and then explains different ways to combine regression and matching to estimate causal effects. Section 4 outlines two simulation studies that we conduct in order to compare the estimation strategies discussed. Each simulated dataset contains a problem that is likely to occur during observational data analysis and compares how each estimation strategy fares. Section 5 discusses limitations of this research project and future areas of work.

2. LITERATURE REVIEW

There has been a lot of interest in causality in social science, and it has always been a statistical challenge due to its strong assumptions. The potential outcomes model attempts to capture the heart of causal effect estimation and is discussed in great detail by Morgan and Winship (2012). They describe the potential outcomes framework as the fundamental concept for causal analysis, and discuss the estimation of effects by approximating parts of the framework. They also describe regression analysis and matching methods in this framework, and explore ways to combine regression and matching.

There has been considerable research in and formalization of regression analysis, but matching methods are still emerging. Stuart (2010) provides and overview of theoretical advancements of matching methodology and explains the process involved in matching, which is later described in this paper. She also provides advice on how to structure estimation using matching methods.

Many statistics researchers focus on developing additions or improvements to matching methods. For instance, Abadie and Imbens (2011) developed a bias correction for a previously biased estimate produced by nearest neighbor matching. They also tested out the

methods using simulation studies and supported their theory. Diamond and Sekhon (2012) also developed a useful matching method called genetic matching, and tested their method using simulation studies. Rosenbaum (1987) pioneered important work in direct adjustment methods, a concept useful in weighting when using matching methods.

There are also important studies that test the effectiveness of matching methods, either by using simulation or using real datasets in which the causal effect is known. Smith and Todd (2005) examined difference in difference matching using the National Supported Work (NSW) data which had been used by LaLonde (LaLonde 1986). The true causal effect in this dataset is known, since it was carried out as an experiment. Smith and Todd concluded that propensity score matching methods were useful, but did not solve the overall econometric evaluation problem. Kurth et al (2006) evaluated regression and matching estimates when there were non-uniform treatment effects in the data. They concluded that the methods produced sensitive estimates, and advised tailoring the method used according to the population being analyzed.

Some social science researchers have begun to use matching methods to estimate causal effects, particularly in the field of labor economics. Imbens (2015) provided a guide for empirical researchers to use matching methods, and listed some useful applications of these methods. For instance, Imbens et al. (2001) use propensity score methods to estimate the effect of winning a large lottery prize on labor earnings. Another example is Dehejia and Wahba (1999), who estimate the effect of an experimental job training program on subsequent earnings.

Matching methods for causal inference are relatively new as compared to regression. There is a growing body of research on matching methods and their applications, but it is scattered across disciplines. This paper attempts to explain matching and regression in the potential outcomes framework and explore links between them in a manner that is accessible and comprehensive. This paper also adds to the small, but growing body of research on the combination of regression and matching methodologies by testing these methods on specific types of observational data. Most matching research is related to theoretical improvements, and this paper attempts to compare different methods that have been developed, but are not often tested together. Morgan and Winship (2012) and Stuart (2010) provide a useful summary of causal effect estimation using matching and regression methods and suggest best practices for using matching. This paper applies some of those concepts as a first step in advancing matching research, and testing already existing methods in the context of specific problems that may arise during the analysis of observational data.

3. CONCEPTS

3.1. POTENTIAL OUTCOMES FRAMEWORK

The potential outcomes framework is the underlying framework for understanding causality and its estimation. This section explains the potential outcomes framework based on the theory presented by Morgan and Winship (2012).

The framework assumes that there exist well defined causal states that have potential outcomes associated with them. Thus, the individual level causal effect (in the binary treatment case) is the difference between the potential outcome if the person receives the treatment and the potential outcome if the person does not receive the treatment. The individual level causal effect can be represented by,

$$\delta_i = y_i^1 - y_i^0$$

where i is an individual, δ_i is the individual level treatment effect, y_i^1 is the outcome if individual i receives the treatment and y_i^0 is the outcome if individual i does not receive the treatment. Since this is the counterfactual framework, we assume that both y_i^1 and y_i^0 can be observed for the same individual.

However, in practice, it is not possible for an individual to be in the treatment as well as control group. Thus, estimation strategies try to approximate this difference in potential outcomes by comparing individuals who are almost indistinguishable from each other. This makes it possible to estimate the true average treatment effect for the population, which is given by,

$$\delta = E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$$

where δ is the treatment effect for the population, Y^1 is the distribution of the potential outcome of receiving treated, and Y^0 is the distribution of the potential outcome of not receiving treatment.

It is also possible to estimate conditional treatment effects, which are the treatment effects for certain subsets of the population. For instance, the average treatment effect on the treated is given by,

$$\delta|(D = 1) = E[Y^1|D = 1] - E[Y^0|D = 1]$$

where D is the treatment. In this paper, D is treated as a binary variable. Thus it is the effect of the treatment on those who were treated, which is obtained by comparing the outcome when an individual in the treatment group is treated and the hypothetical outcome when an individual in the treatment group is not treated.

Another conditional treatment effect is the average treatment effect on the untreated, which is given by,

$$\delta|(D = 0) = E[Y^1|D = 0] - E[Y^0|D = 0]$$

Even though these definitions require impossible to observe, hypothetical outcomes, it is possible to use observed outcomes to find the average treatment effect or conditional average treatment effect if some assumptions are fulfilled. The naïve estimator based on observed values is represented as,

$$\delta_{NAIVE} = E[Y|D = 1] - E[Y|D = 0]$$

where Y is the distribution of the observed outcome, that is, $Y = Y^1$ if $D = 1$ and $Y = Y^0$ if $D = 0$.

This is simply the difference in sample means across the treatment and control (no treatment) groups. When the sample of observations being analyzed is a random sample, the naïve estimator is asymptotically equal to,

$$\delta_{NAIVE} = E[Y^1|D = 1] - E[Y^0|D = 0]$$

Which can be manipulated to yield,

$$\begin{gathered} E[Y^1|D = 1] - E[Y^0|D = 0] = E[\delta] \\ +\{E[Y^0|D = 1] - E[Y^0|D = 0]\} \\ +(1 - \pi)\{E[\delta|D = 1] - E[\delta|D = 0]\} \end{gathered}$$

where $\pi = E[D]$ is the proportion of the population that receives treatment, $E[Y^0|D = 1] - E[Y^0|D = 0]$ is baseline bias, that is the difference between those in the treatment and control groups in the absence of treatment, $E[\delta|D = 1]$ is the average treatment effect on the treated, $E[\delta|D = 0]$ is the average treatment effect on the untreated, and the difference $E[\delta|D = 1] - E[\delta|D = 0]$ is the differential treatment effect bias, that is, if the treatment affects those in the treatment group differently than in the control, then it is the size of that differential effect.

where
- $\pi = E[D]$ is the proportion of the population that receives treatment
- $E[Y^0|D = 1] - E[Y^0|D = 0]$ is baseline bias, that is the difference between those in the treatment and control groups in the absence of treatment
- $E[\delta|D = 1]$ is the average treatment effect on the treated
- $E[\delta|D = 0]$ is the average treatment effect on the untreated
- $E[\delta|D = 1] - E[\delta|D = 0]$ is the differential treatment effect bias, that is, if the treatment affects those in the treatment group differently than in the control, then it is the size of that differential effect.

Both the baseline bias and the differential treatment effect bias are eliminated if a fundamental assumption, called the Stable Unit Treatment Value Assumption (SUTVA) is fulfilled. This assumption is defined by Rubin (1986) as, "SUTVA is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive." This can be formalized as,

$$(Y^0, Y^1)||D$$

Where || indicated joint independence. Thus treatment assignment must be independent of both Y^0 and Y^1 and their functions. The formalization can alternatively be broken down as,

Assumption 1: $E[Y^1|D = 1] = E[Y^1|D = 0]$
Assumption 2: $E[Y^0|D = 1] = E[Y^0|D = 0]$

If both Assumption 1 and Assumption 2 are fulfilled, it can be seen from the naïve estimator expression that the biases cancel out. Since these are strong assumptions, it is important to consider the case in which both assumptions are not fulfilled. If only assumption 1 is fulfilled, then the naïve estimator estimates the average treatment effect on the untreated. If only assumption 2 is fulfilled, then the naïve estimator estimates the average treatment effect on the treated. Assumption 2 is just the assertion that baseline bias does not exist, which is often easier to assert than Assumption 1. As a result, social science research often focusses on estimating the average treatment effect on the treated, since it is an effect with useful implications, and a variety of research designs can be used to minimize baseline bias.

3.2. ORDINARY LEAST SQUARES REGRESSION

Ordinary least squares (OLS) regression finds the best fitting association between the outcome and the explanatory variables, given by a conditional expectation function, which is the outcome conditional on treatment status (Goldberger 1991). The conditional expectation function can be represented as,

$$E[Y|D] = \alpha + \delta D + S$$

Where β is the effect of treatment D on outcome Y, and S is a vector of other predictors or covariates. Regression analysis obtains the best-fitting linear approximation to the conditional expectation function and obtains an estimate $\hat{\delta}$ of δ. This estimate is obtained by minimizing the average squared differences between the predicted values of the outcome from the linear approximation and the true values from the $E[Y|D]$ function (Morgan and Winship 2012).

A generic regression equation is written as,

$$Y = \hat{\alpha} + \hat{\delta} D + \epsilon$$

Where ϵ is the error term, which accounts for the other covariates S, and also random variation in individual outcomes. The OLS estimator for a sample is obtained as,

$$\delta_{OLS} = \frac{Cov(y_i, d_i)}{Var(d_i)}$$

where y_i is the observed outcome for each individual in the sample, and d_i is the treatment status of each individual in the sample.

Causal effect estimation using regression analysis is subject to some assumptions. If these assumptions are fulfilled, the estimated effect, δ_{OLS} can be considered an estimate of the causal effect, rather than just a predictor. One of the most important assumptions is that the causal effect should not vary with the other covariates. It is important to note that the causal effect should not only be independent of the other covariates on average, but should also be fully independent of the covariates at the individual level. This important distinction is often overlooked while estimating causal effects (Morgan and Winship 2012).

Another important assumption, one that is often violated and renders regression estimates biased, is that of individual level homogeneity of treatment effects. This is the assumption that the treatment affects all individuals in the same way, that is, the difference between potential outcomes is the same for all individuals. However, they may have different baseline values for the outcomes. In social sciences applications, this assumption is often violated, since different individuals tend to behave differently. The problem can usually solved by controlling or partialling out the covariates, such that the difference between only similar individuals is calculated within strata, and then then a weighted average of the within-strata differences is calculated for the population. However, if the effect of the treatment depends on the individual's propensity to receive treatment, which can plausibly be the case in most social science applications, then the regression estimator will be biased (Xie et al. 2012).

In the presence of heterogeneous treatment effects, the regression estimator can only be interpreted as a conditional variance weighted estimator, and may give a biased estimate of average treatment effect. Consider a simple example in which there is one covariate S that takes on values 1 and 2, and a binary treatment variable D. The OLS estimator is calculated as,

$$\frac{E[Var[D|S=1]]}{\sum_S E[Var[D|S=s]]}\{E[Y|D=1,S=1] - E[Y|D=0,S=1]\} +$$
$$\frac{E[Var[D|S=2]]}{\sum_S E[Var[D|S=s]]}\{E[Y|D=1,S=2] - E[Y|D=0,S=2]\}$$

The reason why regression implicitly invokes conditional variance weighting is because it is a minimum variance estimator, and hence assigns a higher weight to stratum specific effects with lower variance (Morgan and Winship 2012). This property of regression can bias its estimate in the presence of individual level heterogeneity of treatment effect. This paper later examines if this problem can be solved by replacing conditional variance weights with inverse propensity score weights.

3.3. MATCHING METHODS

Using matching procedures pares down the data such that the observations in the treatment groups are comparable across covariates, by discarding information that is unrelated to variation in the treatment (Morgan and Winship 2012). Stuart (2010) outlines matching procedures, and states that the first step is choosing a distance measure. A distance measure quantifies how different two observations are from each other based on their covariates. The distance can be calculated based on exact equality of covariates, or other metrics that capture the characteristics of the covariates such as the propensity score or the Mahalanobis metric. The distance estimation methods can also be used in combination with each other. This paper utilizes propensity scores as a measure of distance.

The propensity score was introduced by Rosenbaum and Rubin (1983). It is the within-stratum probability of an individual receiving treatment. In other words, it quantifies the probability of receiving treatment, given the covariates, that is, $\Pr(D = 1|S)$. The estimated propensity score is the estimated probability of receiving the treatment as a function of variables that predict treatment assignment (Morgan and Winship, 2012). Propensity scores are useful because they summarize all the covariates into one value, that is, the probability of being treated. As explained by Stuart (2010), propensity scores have two important properties: First, they balance out the covariates. For each propensity score, the distribution of covariates across treatment groups is same. So, grouping by propensity score is similar to recreating a random experiment, albeit based on only observed characteristics. Second, if treatment assignment is ignorable given covariates, then it is ignorable given propensity scores. This justifies matching based on propensity scores rather than exact matching (Abadie and Imbens 2011).

Once the distance measure has been selected, the matching method needs to be chosen. There is a general framework that matching methods follow in estimating treatment effects. Smith and Todd (2005) provide a formalization of this framework for the average treatment effect on the treated, such that all matching estimators of this effect can be represented as,

$$\hat{\delta}_{TT} = \frac{1}{n^1} \sum_i \left[(y_i | d_i = 1) - \sum_j \omega_{i,j} (y_j | d_j = 0) \right]$$

Where n^1 is the number of individuals receiving treatment, i is the index for treatment cases, j is the index for control cases, and $\omega_{i,j}$ is a set of weights that incorporate the distance measure between each control case and a given treatment case (. Thus, control cases that are more similar to the treatment cases are weighted more. As different matching methods are applied, the weights in this formula are specified differently. Similar generalized formulas can be obtained for the unconditional treatment effect and other conditional treatment effects.

Nearest neighbor matching and interval matching are some important and commonly used matching methods. In nearest neighbor matching, individuals are matched by creating pairs of treatment and control individuals that have the least distance. This is a flexible

method, and one can also set a maximum distance, above which a pair is discarded from analysis. There is also the option of matching with replacement, in which an individual may be counted more than once in order to create a matched pair. However, it is important to be cautious that one individual is not repeated too many times for forming matched pairs (Stuart 2010). Interval matching, also known as stratification matching is performed by sorting treatment and control individuals into segments according to a metric (such as propensity score). The population is then divided into strata having the same metric value, and the within-strata difference in outcomes between treatment and control units is calculated. Each within strata difference is then weighted according to the joint distribution of covariates and summed together to find the treatment effect (Rubin 1977). Interval matching is similar to the idea of controlling or partialling out of other covariates, in which similar subsets of the population are compared to each other by partialling out variation in covariates. Similar to interval matching, the within subset difference is then combined as a weighted average for the population.

Matching concepts can also be used for pre-processing the data, and then can be combined with other methods to find causal effects. Matching can be used to achieve optimal balance in the sample, that is, to ensure that the sample contains only comparable treatment and control units. In the case of estimating the average treatment effect on the treated, control units that did not match with any treatment effects are discarded. Thus, it is important to compare the number of treated and control units in the sample. If there are too few treatment units as compared to control units, discarding a large subset of the sample is likely to bias the estimate of causal effect.

If used on a sample with a similar number of treatment and control units, matching prepares the data and allows the researcher to ensure that there is sufficient overlap between the covariates. Overlap between covariates simply means that the distribution of the covariates across treatment and control units is similar. Checking for covariate overlap is a fundamental part of using matching for achieving balance, and can be verified using numerical or graphical methods (Stuart 2010). The most popular numerical method proposed by Rosenbaum and Rubin (1985) involved calculating the standardized difference in means for each covariate across the treatment and control groups. Graphical methods involve creating boxplots or quantile-quantile plots of each of the covariates before and after the sample has been matched and trimmed (Stuart 2010). An important point to keep in mind is that it is not only acceptable, but in fact advised, to check for overlap after each alteration of the dataset and modify the matching method if overlap is not satisfactory. This is not considered as data mining, since modifying the matching method does not alter a prediction of the outcome. Selecting the matching method is similar to designing an experiment, and changes in the matching method are aimed at balancing treatment assignment.

Matching methods are also useful for developing weighting methods. Propensity scores are an important subset of the matching procedure that can be used for weighting and comparing similar individuals. An example of using matching concepts for weighting is the Inverse Probability of Treatment Weighting (IPTW) method, in which each individual is assigned a weight depending on their propensity of receiving the treatment that they were assigned. These weights can be applied to interval matching, or regression analysis. The average treatment effect is then calculated as the sum of these weighted units (Lunceford and Davidian 2004).

3.4. MATCHING AND REGRESSION AS SUPPLEMENTS

An important perspective when examining matching methods is that matching is not designed to compete with regression and can effectively be used in conjunction with it. Matching

forces the researcher to look at the joint distributions of covariates and assess overlap, a process that is not required for regression and can lead to biased estimates. On the other hand, matching estimators can have large standard errors, which the variance minimizing approach of regression could help to counter (Morgan and Winship 2012).

Morgan and Winship (2012) succinctly describe the use of matching methods as a supplement, "In the methodological literature, matching is usually introduced in one of two ways: (1) as a method to form quasi-experimental contrasts by sampling comparable treatment and control cases from among two larger pools of such cases or (2) as a nonparametric method of adjustment for treatment assignment patterns when it is feared that ostensibly simple parametric regression estimators cannot be trusted." Thus, this begins to suggest that the problems associated with regression could be solved by combining with matching methods, a hypothesis that I test using simulations later in the paper.

I test the first method of forming quasi-experimental contrasts by using matching methods to balance the data, and then use regression analysis to estimate the effect on the processed and balanced data. I also test if a combination of matching and treatment effects can produce an unbiased estimate of the treatment effect when there is individual level heterogeneity of treatment effects. I use the IPTW with regression analysis, which is a method formally known as Doubly Robust estimation. That is, if either the propensity score estimation or the regression model is correctly specified, the estimate will be unbiased (Lunceford and Davidian 2004). Thus, inverse propensity scores are used as weights in the regression model. The weights are as follows,

$$w_i = \frac{d_i}{\mathrm{p_i}} + \frac{(1 - d_i)}{(1 - \mathrm{p_i})}$$

where $\mathrm{p_i}$ is the estimated propensity of receiving treatment. Thus for an individual with $d_i = 1$ in the treatment group, the outcome is weighted by the inverse of the propensity of being treated, and for those in the control group, the outcome is weighted by the inverse of the propensity of not receiving treatment. Individuals in the treatment who are more similar to those in the control group, that is, those who have a higher propensity of not receiving treatment are weighted more, and vice versa for those in the control group. This ensures that the effect is being calculated for comparable individuals.

4. SIMULATION STUDIES

4.1. METHODOLOGY

We use two Monte Carlo simulations with 1000 iterations to test for two different problems that arise during the analysis of observational data. In Simulation 1, the the estimated propensity score is misspecified, since in observational data, it is unlikely that the true propensity score will be known. In Simulation 2, we model individual level heterogeneity of treatment effects, such that the treatment outcome depends on the propensity of receiving treatment. In each simulation, we test two uses of matching: balancing and weighting, and combine it with regression for causal effect estimation.

The general framework of both simulations is similar. We generate a sample consisting of 10,000 observations with two covariates A and B ($100A \times 100B = 10{,}000$). A and B are both stratified as $(.01, .02, \ldots, 1)$. Thus, the 10,000 observations contain all the possible combinations of A and B. There is a binary treatment variable, D. The propensity of receiving treatment is defined as,

$$\Pr(D = 1|S) = \frac{1}{1 + \frac{1}{\exp(S\phi)}}$$

Where $S\phi$ is the index function of a logit, defined differently for Simulations 1 and 2. Based on the propensity score function, we generate a matrix of propensity scores for each of the 10,000 combinations of A and B.

The above information is then combined into a dataset with each row representing an individual having covariates A and B, and a propensity of receiving treatment. Each individual is then assigned two potential outcomes, one for treatment and one for control according to function which is varied in both the simulations. Each individual is also assigned a treatment state which has a Bernoulli distribution with probability as the true propensity of receiving treatment. Based on the treatment state, the individual is assigned an observed outcome which is the potential outcome for the assigned treatment. Thus, if an individual was assigned to the treatment group, the observed outcome is the potential outcome of receiving treatment.

Simulation 1

Simulation 1 tests the performance of propensity score matching supplements when matching is expected to be at a disadvantage, that is, when the propensity score estimate is misspecified.

The potential outcomes are assigned as,

$$y_0 = 100 + 3A + 2B + \epsilon$$
$$y_1 = 102 + 6A + 4B + \epsilon$$

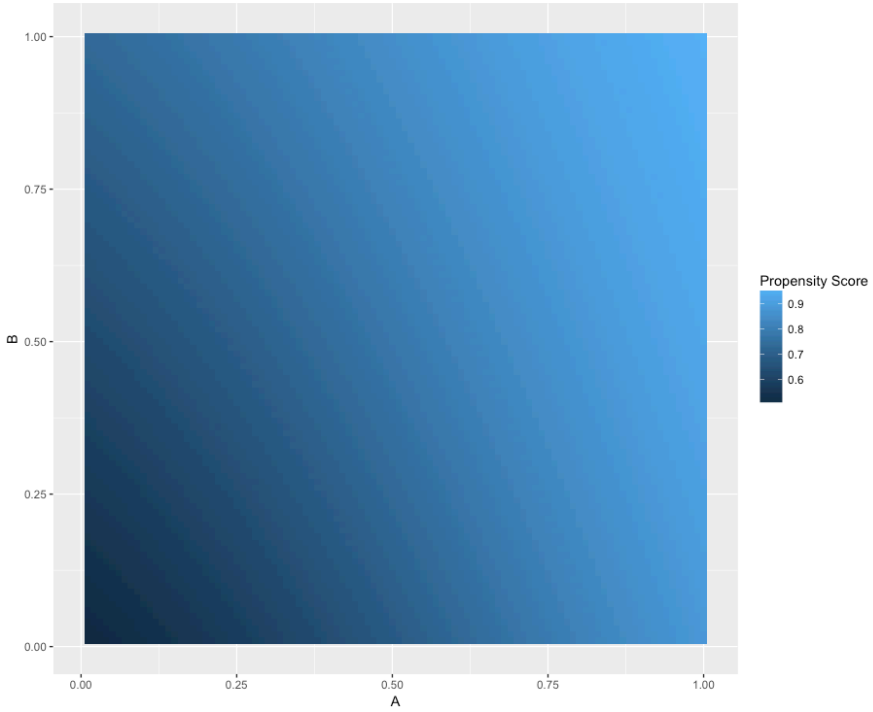
where ϵ is random error term drawn from a normal distribution, $\sim N(0,5)$. This term is added to account for individual level random variation in outcomes.

The true propensity score index function is,

$$S\phi = \alpha + \beta_1 A + \beta_2 A^2 + \gamma B$$

Thus the propensity score assignment is quadratic in A, while the estimated propensity score treats A linearly within the index of the logit function.

Figure 1: Propensity Score Function for Simulation 1



Simulation 2

Simulation 2 tests the performance of propensity score matching supplements when regression is supposed to produce biased estimates, that is, when the data contain individual level heterogeneity of treatment effects.

The potential outcomes are assigned as,

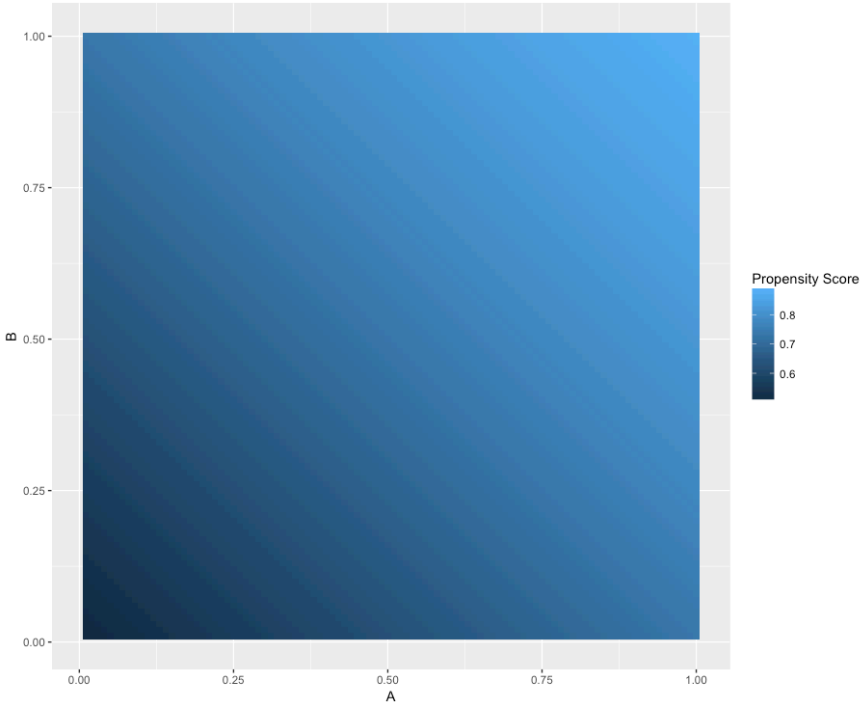
$$y_0 = 100 + 3A + 2B + \epsilon$$
$$y_1 = 102 + 6A + 4B + 5PScore + \epsilon$$

where $PScore$ is the propensity of receiving treatment. Thus the potential outcome of receiving treatment depends on the propensity of the individual to receive treatment.

The propensity score function index is now linear in A and is defined as,

$$S\phi = \alpha + \beta A + +\gamma B$$

Figure 2: Propensity Score Function for Simulation 2



4.2. ESTIMATION METHODS

We use three estimation methods: simple regression, regression on a sample balanced by nearest neighbor matching, and regression with inverse propensity of treatment weights. In each case, we first assume that A is not observed and is a source of omitted variable bias. Then it is assumed that all the covariates, that is, both A and B are observed.

In estimation method 1, the regression is specified as a linear regression, which in its fully specified form can be represented as,

$$Y = \alpha + \beta A + \gamma B + \epsilon$$

In estimation method 2, the sample is balanced by matching treatment and control units using the nearest neighbor matching method. At first, we used the method without replacement and verified covariate balance using boxplots, but it did not achieve covariate balance. We then used nearest neighbor with replacement and the covariates seemed adequately balanced. However, when nearest neighbor with replacement is used, there is a risk that one control observation will be used multiple times to match with a treatment observation. Thus, the distance measure is also weighted by the frequency at which the control observation is used, so that the relative frequency of treatment and control units is not unreasonably skewed. Following the calculation of distances, treatment or control units that are not matched are removed and the matched sample is generated. On this sample, we run a regression similar to estimation method 1.

In estimation method 3, we estimate the propensity score using a logistic regression with index function,

$$S\phi = \alpha + \beta A + \gamma B$$

As mentioned previously, this estimated propensity score has a different index than the true propensity score. Even though it is usually impossible to identify the true propensity score

function in observational data, a logistic regression on the covariates often provides a useful estimate of the true propensity score with low variance (Morgan and Winship 2012). We use the estimated propensity scores as inverse probability weights in the regression, replacing the conditional variance weights.

4.3. RESULTS

In the first simulation, propensity score matching methods combined with regression outperform simple regression by a large margin when both covariates are observed. Despite the case of propensity score estimation being misspecified, regression with inverse propensity score weights produces a very lightly biased estimate, which is much smaller than the estimates produced by simple regression and regression on a balanced sample. This suggests that it is not too harmful to misspecify propensity score estimates after carefully assessing balance. It is interesting to note that when the misspecified quadratic covariate A is treated as unobserved, all the methods perform better relative to when A is included in the specification. This suggests that it is better to not specify a covariate than to misspecify it. When both A and B are observed, the standard errors are similar across the different estimation methods, indicating that there was not a large variance penalty for using matching and regression methods together.
However, when A is not observed, the simple regression estimate produces the least bias, although this bias is not much smaller than the bias observed using regression on a matched sample. Regression using propensity score weights produces about double the bias as compared to simple regression and regression with a matched sample. This is not unexpected, since the the former relies the most directly on the estimated propensity score, which is missing the input of covariate A. Since the true propensity score is quadratic in A and influenced by A more than B, we would expect the estimated propensity score to provide extremely biased estimates of the the true propensity score.

Table 1: Results from Simulation 1

	ATE Estimate	Standard Error	Bias
True ATE:	5.8707		
Only Regression:			
Unadjusted	6.5316	0.0446	0.6609
A is unobserved	6.0055	0.0389	0.1348
A and B are observed	5.5651	0.0103	-0.3056
Regression on balanced sample:			
A is unobserved	6.0306	0.0404	0.1599
A and B are observed	5.6278	0.0103	-0.2430
Regression with inverse propensity score weights:			
A is unobserved	6.1668	0.0278	0.2961
A and B are observed	5.8638	0.0104	-0.0069

In the second simulation, propensity score matching methods again outperform simple regression when both A and B are observed. In this case, it was expected that matching methods would perform better than simple regression, since the data contained individual level heterogeneity of treatment effects. Similar to simulation 1, regression with inverse propensity score weights produced the least biased estimate, less than half the size of the bias

produced by simple regression and regression on a matched sample. However, it is worth noting that inverse propensity score weighting outperformed simple regression by a smaller magnitude than in simulation 1. This indicates that heterogeneity of treatment effects are a cause of concern for matching methods as well, and can increase bias. The standard errors of all three methods are comparable when all the covariates are observed.
In the case of A being unobserved, simple regression again outperforms regression with matching supplements, although the bias obtained using simple regression and regression on a matched sample is similar. The estimate obtained using regression with inverse propensity score weights indicates biased values of of the estimated propensity score, though not as heavily biased as those in simulation 1. This is expected since the missing covariate A is linear in simulation 2.

Table 2: Results from Simulation 2

	ATE Estimate	Standard Error	Bias
True ATE:	8.2955		
Only Regression:			
Unadjusted	8.8556	0.0480	0.5601
A is unobserved	8.5262	0.0406	0.2307
A and B are observed	8.0410	0.0142	-0.2545
Regression on balanced sample:			
A is unobserved	8.5581	0.0424	0.2626
A and B are observed	8.1191	0.0144	-0.1764
Regression with inverse propensity score weights:			
A is unobserved	8.6329	0.0319	0.3375
A and B are observed	8.2204	0.0142	-0.0751

Both simulation studies indicated in general that when all the covariates are observed, regression with inverse propensity score weights outperform regression on a matched sample and simple regression. However, when there is a risk of some covariates being unobserved, inverse propensity score weighted regression does not fare as well. Simple regression produces the least bias when one covariate is unobserved. Regression on a sample balanced by nearest neighbor matching provides a middle ground in terms of bias reduction. When both covariates are observed, it produces less bias than simple regression, and when one covariate is unobserved, it produces bias that is comparable to simple regression.

5. DISCUSSION AND LIMITATIONS

The results from the simulation study generated some expected results, such as regression and matching as supplements providing less biased estimates than simple regression. However, some results, such as inverse probability weighting producing an estimate with very small bias with misspecified propensity scores were unexpected. An avenue for further research would be to examine the level of propensity score misspecification that renders matching methods useless. This would be useful information if a researcher does not have theory to inform the estimation of propensity scores, and runs the risk of high misspecification of the propensity score.

There are some caveats about the two simulation studies that are worth mentioning. We included only two covariates in order to make misspecification problems easy to

simulate. However, further simulation studies should use more covariates that are dependent on each other, unlike the independent covariates used in this study. This study also does not include enough variations in the simulations. There should be more simulations with different types of covariates, some observed and some unobserved, and different misspecifications of the regression equation and the propensity score estimation should be tested.

Although this paper attempts to use suggestions for best practices for using propensity score methods, there are still some procedures that remain. When using observational data, sensitivity analysis of the estimate is important for testing causality. Stuart (2010) suggests some methods for sensitivity analysis, such as estimating the minimum magnitude of a confounding variable required to nullify the causal effect estimate obtained. It is also important to acknowledge that simulation studies have low generalizability, and cannot be used to make definitive claims about causal effect estimation. The purpose of this paper was to make the combination of regression and matching methods accessible, and then test if these methods perform well in the face of problems that commonly arise in observational data. The next step in this research project would be test out these estimation methods on real datasets, in which the causal effect is known, either by combining treatment units from an experiment with non-experimental control units from another dataset (for example, the LaLonde dataset), or by using doubly robust methods in studies in which the treatment effect has been estimated using instrumental variables, or difference in difference research designs.

Bibliography

A. Smith, Jeffrey, and Petra E. Todd. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125.1 (2005): 305–353. *ScienceDirect*. Web. Experimental and Non-Experimental Evaluation of Economic Policy and Models.

Abadie, Alberto, and Guido W. Imbens. "Bias-Corrected Matching Estimators for Average Treatment Effects." *Journal of Business & Economic Statistics* 29.1 (2011): 1–11. *amstat.tandfonline.com (Atypon)*. Web.

Dehejia, Rajeev H., and Sadek Wahba. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94.448 (1999): 1053–1062. *amstat.tandfonline.com (Atypon)*. Web.

Diamond, Alexis, and Jasjeet S. Sekhon. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *The Review of Economics and Statistics* 95.3 (2012): 932–945. *MIT Press Journals*. Web.

Goldberger, Arthur Stanley. *A Course in Econometrics*. Harvard University Press, 1991. Print.

Imbens, Guido W. "Matching Methods in Practice: Three Examples." *Journal of Human Resources* 50.2 (2015): 373–419. *jhr.uwpress.org*. Web.

Imbens, Guido W., Donald B. Rubin, and Bruce I. Sacerdote. "Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players." *The American Economic Review* 91.4 (2001): 778–794. Print.

Kurth, Tobias et al. "Results of Multivariable Logistic Regression, Propensity Matching, Propensity Adjustment, and Propensity-Based Weighting under Conditions of Nonuniform Effect." *American Journal of Epidemiology* 163.3 (2006): 262–270. *academic.oup.com*. Web.

LaLonde, Robert J. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review* 76.4 (1986): 604–620. *JSTOR*. Web.

Lunceford, Jared K., and Marie Davidian. “Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study.” *Statistics in Medicine* 23.19 (2004): 2937–2960. *Wiley Online Library*. Web.

---. “Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study.” *Statistics in Medicine* 23.19 (2004): 2937–2960. *Wiley Online Library*. Web.

Morgan, Stephen L., and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 1 edition. New York: Cambridge University Press, 2007. Print.

Rosenbaum, Paul R. “Model-Based Direct Adjustment.” *Journal of the American Statistical Association* 82.398 (1987): 387–394. *JSTOR*. Web.

Rosenbaum, Paul R., and Donald B. Rubin. “Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score.” *The American Statistician* 39.1 (1985): 33–38. *amstat.tandfonline.com (Atypon)*. Web.

Stuart, Elizabeth A. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical science : a review journal of the Institute of Mathematical Statistics* 25.1 (2010): 1–21. *PubMed Central*. Web.

Xie, Yu, Jennie E. Brand, and Ben Jann. “Estimating Heterogeneous Treatment Effects with Observational Data.” *Sociological methodology* 42.1 (2012): 314–347. Print.