**REGRESSION AND MATCHING AS SUPPLEMENTS IN THE ESTIMATION OF CAUSAL EFFECTS IN SOCIAL SCIENCE RESEARCH**

*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 at the same time. 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 the most commonly used tool in social science research for predicting the outcomes given a set of covariates. However, in order to obtain the least biased causal estimate, the distribution of the covariates across different treatment groups must be examined. The use of matching methods involves 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 comparing the performance of propensity score methods with ordinary least squares regression in various scenarios that may occur in observational data analysis across disciplines.

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.

1. **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 forms the theoretical basis of causal effect estimation and is discussed in great detail by Morgan and Winship (2012). They describe the potential outcomes framework and discuss the estimation of treatment effects by approximating parts of the framework. They also describe regression analysis and matching methods in the context of 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 strategies 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 an 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 explains 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 in situations that occur in observational research. 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.

1. **CONCEPTS**
   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,

where is an individual, is the individual level treatment effect, is the outcome if individual receives the treatment and is the outcome if individual does not receive the treatment. Since this is the counterfactual framework, we assume that both and 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,

where is the treatment effect for the population, is the distribution of the potential outcome of receiving treated, and 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,

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,

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,

where is the observed outcome for individuals who receive the treatment and is the observed outcome for those in the control group.

This is simply the difference in sample means across the treatment and control (no treatment) groups. It can be manipulated to yield,

where is the proportion of the population that receives treatment,

is baseline bias, that is the difference between those in the treatment and control groups in the absence of treatment, is the average treatment effect on the treated, is the average treatment effect on the untreated, and the difference 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,

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

Assumption 1:

Assumption 2:

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.

* 1. **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,

Where is the effect of treatment on outcome , and 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 of . 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 function (Morgan and Winship 2012).

A generic regression equation is written as,

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

Causal effect estimation using regression analysis is subject to some assumptions. If these assumptions are fulfilled, the estimated effect, 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 that takes on values and , and a binary treatment variable . The OLS estimator is calculated as,

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.

* 1. **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, . 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,

Where is the number of individuals receiving treatment, is the index for treatment cases, is the index for control cases, is a set of weights that incorporate the distance measure between each control case and a given treatment case, and . Thus, each treatment case has control cases assigned it to, and control cases that are more similar to the treatment cases are weighted more and vice versa. 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. In this case, where is the number of matched control units for each treatment unit . 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). In this case, the weights are chosen in such a way that treatment and control units within each interval are given equal weights. 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. 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).

* 1. **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).

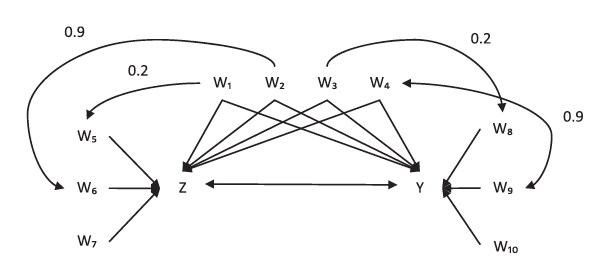
Morgan and Winship (2012) succinctly describe the use of matching methods as a supplement, "In the methodological literature, matching is usually introduced in one of two ways: (1) as a method to form quasi-experimental contrasts by sampling comparable treatment and control cases from among two larger 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 we test using simulations later in the paper.

We 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. We 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. We 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,

where is the estimated propensity of receiving treatment. Thus for an individual with 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.

1. **SIMULATION STUDIES**
   1. **METHODOLOGY**

We conducted Monte Carlo simulations to compare the performance of regression, regression on a sample balanced by nearest neighbor matching, and regression with inverse propensity of treatment weighting. The simulation study is structured according to the simulations in Leacy and Stuart (2014), Setoguchi et al (2008), and Lee et al (2009). Ten covariates whose distribution is the standard normal distribution were created. The covariance structure was defined such that , and . Only determined the treatment assignment process, determined the outcome, and were confounders. The simulation structure is further elucidated by Figure 1 from Leacy and Stuart (2014). Each dataset contained 1000 simulated observations.



**Figure 1: Simulated dataset structure**

The performance of the three methods was assessed across (5x2)=10 scenarios. We defined five propensity score models with varying degrees of linearity and additivity. As explained in Leacy and Stuart (2014), the five different propensity score models were as follows:

1. additivity and linearity
2. moderate non-additivity (with ten two-way interactions among covariates)
3. mild non-additivity (three two way interactions) and non-linearity (one quadratic term)
4. moderate non-linearity (three quadratic terms)
5. moderate non-additivity (ten two-way interaction terms) and non-linearity (three quadratic terms)

Another degree of variation is the individual-level heterogeneous treatment effects. We generate two scenarios to vary individual-level heterogeneity of treatment effects. In one scenario, the only source of individual level heterogeneity is the error term included in the potential outcome generating model. In the second scenario, the potential outcomes also depend on the propensity of receiving treatment. In observational data, treatment effects may vary systematically with an individual’s propensity score (Xie et al, 2012). Modelling this scenario allows us to test the performance of estimation methods in the presence of a higher degree of individual level heterogeneity.

The exact functional forms of the treatment assignment and potential outcome generating mechanisms are provided in the Appendix.

* 1. **ESTIMATION METHODS**

As mentioned before, we use three estimation methods: simple regression, regression on a sample balanced by nearest neighbor matching, and regression with inverse propensity of treatment weights.

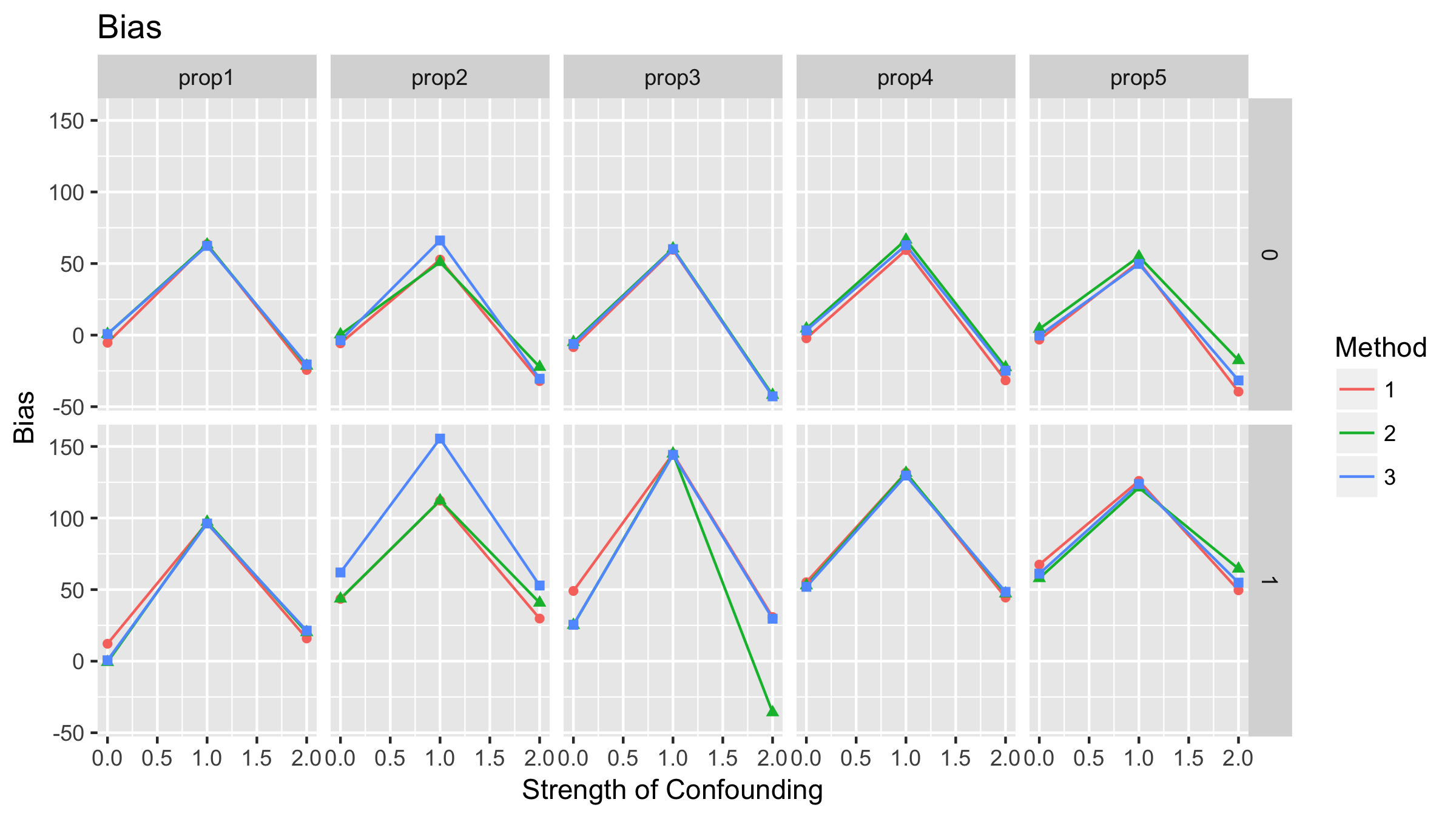
In the first estimation method, the regression is specified as a linear regression. In the second estimation method, the sample is balanced by matching treatment and control units using the nearest neighbor matching method with replacement. However, when nearest neighbor with replacement is used, there 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. In the third estimation method, we estimate the propensity score using a logistic regression. We then use the estimated propensity scores as inverse probability weights in the regression, replacing the conditional variance weights.

In each case, both the propensity score and treatment effect models are assumed to be linear. Thus, the propensity score is mis-specified in the scenarios in which it is not generated linearly, and the individual level heterogeneity is not captured. The two main degrees of variation (mis-specification of propensity score and presence of individual level heterogeneity) represent scenarios that commonly arise in observational datasets in social science. Additionally, each estimation method is subjected to three degrees of confounding, that is the number of covariates that are known. In case 1, it is assumed that all covariates are observed. In case 2, it is assumed that and are not observed. In case 3, we assume that are not observed.

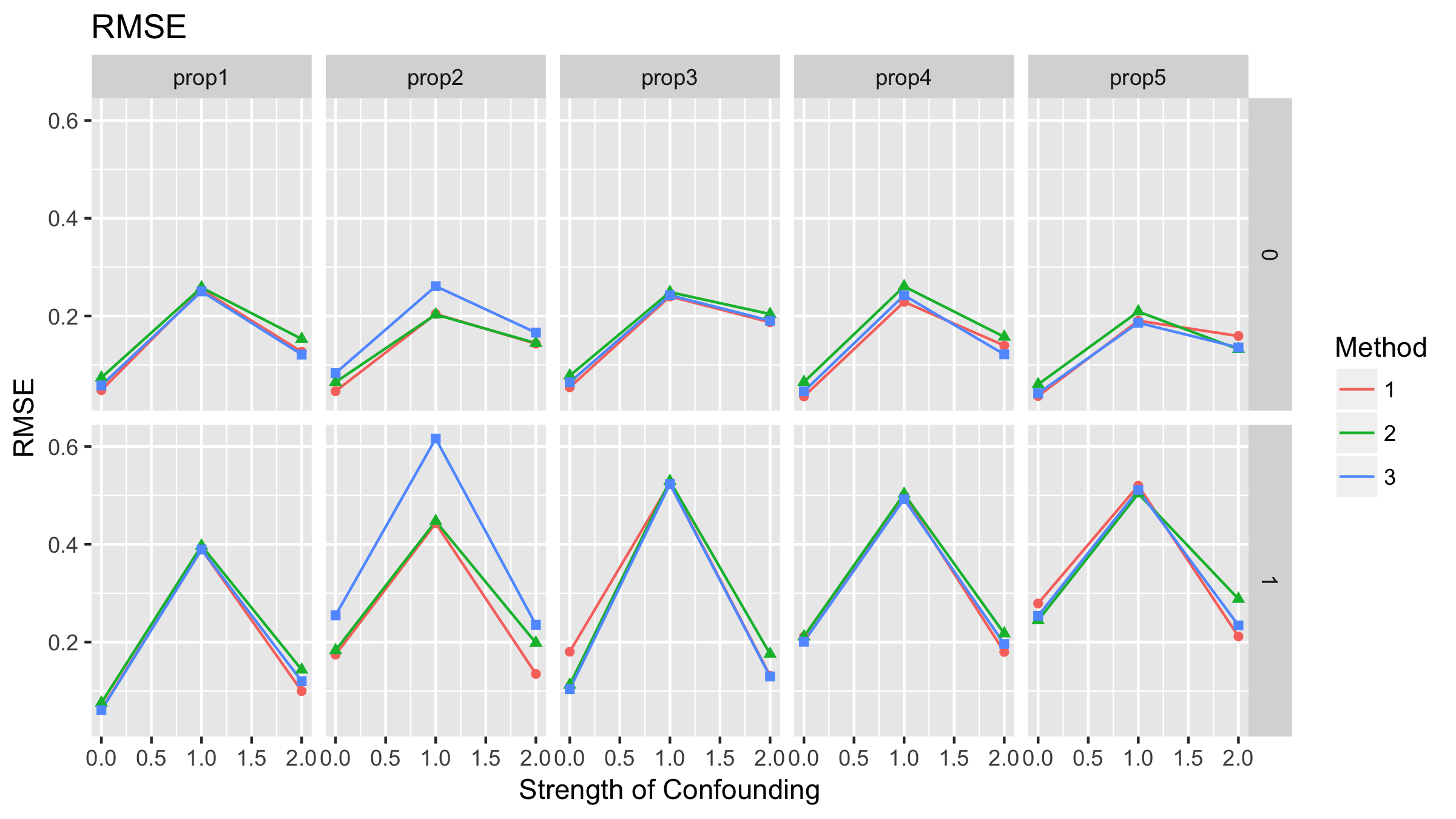
* 1. **RESULTS**

The performance of these (3 estimation methods x 3 degrees of confounding) = 9 models in (5 propensity score models x 2 degrees of heterogeneity) = 10 scenarios is assessed on the basis of relative bias, root mean squared error, and model-generated standard error.

* + 1. **Relative Bias**

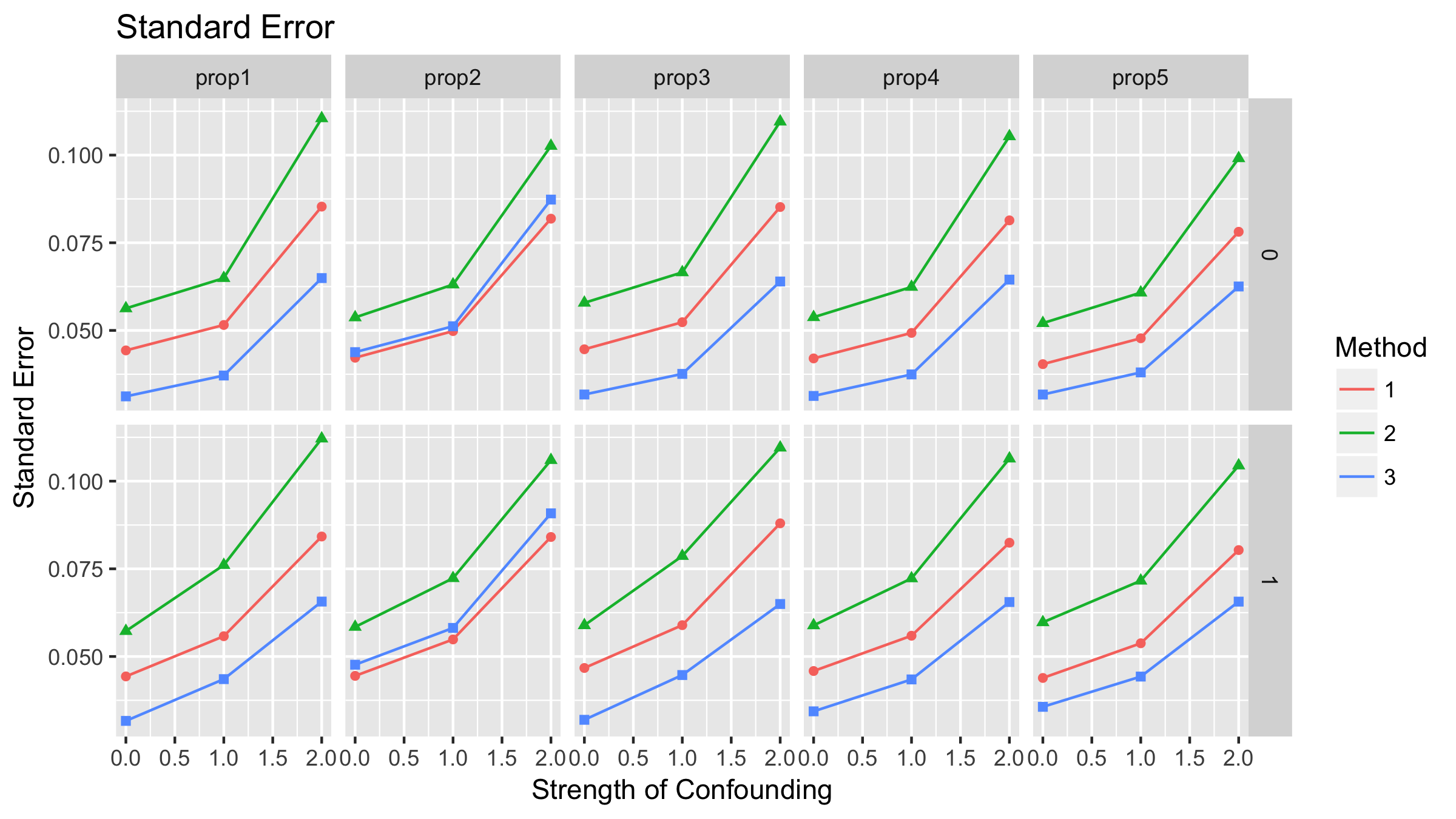
The relative bias in this case is defined as . From Figure 2, we can see that all three methods are comparable in some scenarios such as linear true propensity score, mild non-additivity and non-linearity of propensity score and less individual-level heterogeneity, and moderate non-additivity of the propensity score. However, in the case of moderate non-additivity of the propensity score, inverse propensity of treatment weighting performs better than the other two methods in the presence of mild confounding, especially when there is high individual level heterogeneity. Additionally, inverse propensity of treatment weighting performs best in the case of moderate non-additivity and non-linearity with moderate confounding, in both cases of heterogeneity. It is also important to note that regression on a matched sample greatly underestimates the estimate in the case of mild non-additivity and non-linearity of the propensity score and high heterogeneity with moderate confounding.

* + 1. **RMSE**

****

As seen from Figure 3, inverse propensity of score weighting has the highest RMSE in the case of moderate non-additivity of the propensity score. In other scenarios, the performance of the three methods in similar.

* + 1. **Standard Error**



As seen from Figure 4, inverse propensity of treatment weighting consistently has the highest standard error, followed by simple regression, and regression on a matched sample (except for moderate on-additivity of the propensity score, in which regression on a matched sample has a high standard error than regression alone). It is important to note that the standard error is bounded by 0 and 0.125 and thus the magnitude of the standard error remains small.

1. **SUMMARY AND DISCUSSION**

In summary, in terms of bias, inverse propensity of treatment weighting performs either the best or approximately as well as other methods in the estimation of the average treatment effect on the treated. However, in terms of the root mean squared error and standard error, this method performs worse than other methods, but not by a large margin. Regression on a sample balanced by nearest neighbor matching performed the worst among the methods in terms of bias.

The results from the simulation study generated some expected results, such as inverse propensity of treatment weighting providing less biased estimates than simple regression in the case of heterogeneity of treatment effects. 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.

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*, vol. 125, no. 1, Mar. 2005, pp. 305–53. *ScienceDirect*, doi:10.1016/j.jeconom.2004.04.011.

Abadie, Alberto, and Guido W. Imbens. “Bias-Corrected Matching Estimators for Average Treatment Effects.” *Journal of Business & Economic Statistics*, vol. 29, no. 1, Jan. 2011, pp. 1–11. *amstat.tandfonline.com (Atypon)*, doi:10.1198/jbes.2009.07333.

Dehejia, Rajeev H., and Sadek Wahba. “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association*, vol. 94, no. 448, Dec. 1999, pp. 1053–62. *amstat.tandfonline.com (Atypon)*, doi:10.1080/01621459.1999.10473858.

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*, vol. 95, no. 3, Oct. 2012, pp. 932–45. *MIT Press Journals*, doi:10.1162/REST\_a\_00318.

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

Imbens, Guido W., et al. “Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players.” *The American Economic Review*, vol. 91, no. 4, 2001, pp. 778–94.

---. “Matching Methods in Practice: Three Examples.” *Journal of Human Resources*, vol. 50, no. 2, Mar. 2015, pp. 373–419. *jhr.uwpress.org*, doi:10.3368/jhr.50.2.373.

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*, vol. 163, no. 3, Feb. 2006, pp. 262–70. *academic.oup.com*, doi:10.1093/aje/kwj047.

LaLonde, Robert J. “Evaluating the Econometric Evaluations of Training Programs with Experimental Data.” *The American Economic Review*, vol. 76, no. 4, 1986, pp. 604–20. *JSTOR*, doi:10.2307/1806062.

Leacy, Finbarr P., and Elizabeth A. Stuart. “On the Joint Use of Propensity and Prognostic Scores in Estimation of the Average Treatment Effect on the Treated: A Simulation Study.” *Statistics in Medicine*, vol. 33, no. 20, Sept. 2014, pp. 3488–508. *Wiley Online Library*, doi:10.1002/sim.6030.

Lee, Brian K., et al. “Improving Propensity Score Weighting Using Machine Learning.” *Statistics in Medicine*, vol. 29, no. 3, Feb. 2010, pp. 337–46. *Wiley Online Library*, doi:10.1002/sim.3782.

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*, vol. 23, no. 19, Oct. 2004, pp. 2937–60. *Wiley Online Library*, doi:10.1002/sim.1903.

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

Rosenbaum, Paul R. “Model-Based Direct Adjustment.” *Journal of the American Statistical Association*, vol. 82, no. 398, 1987, pp. 387–94. *JSTOR*, doi:10.2307/2289440.

Rosenbaum, Paul R., and Donald B. Rubin. “Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score.” *The American Statistician*, vol. 39, no. 1, Feb. 1985, pp. 33–38. *amstat.tandfonline.com (Atypon)*, doi:10.1080/00031305.1985.10479383.

Setoguchi, Soko, et al. “Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study.” *Pharmacoepidemiology and Drug Safety*, vol. 17, no. 6, June 2008, pp. 546–55. *Wiley Online Library*, doi:10.1002/pds.1555.

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*, vol. 25, no. 1, Feb. 2010, pp. 1–21. *PubMed Central*, doi:10.1214/09-STS313.

Xie, Yu, et al. “Estimating Heterogeneous Treatment Effects with Observational Data.” *Sociological Methodology*, vol. 42, no. 1, Aug. 2012, pp. 314–47.