A Comparison of Matching and Weighting Algorithms for Treatment Effect Estimation

By: Sophie Sommer

Motivation.

Matching (and/or weighting) is a commonly used method for estimating causal effects in the absence of a randomized experiment. Effect estimates based on observational data can be biased because of the existence of confounders: variables that are related to an individual's chance of receiving treatment and that individual's outcome of interest. Matching allows the researcher to restrict and/or weight a control group in order to more closely match a treatment group, and thus control for observed confounders by creating balance across the two groups (Olmos & Govindasamy, 2015). This theoretically allows the researcher to use observational data to obtain more accurate estimates of treatment effects. Previous research comparing estimates from randomized experiments to estimates from constructed observational studies with the same treated group have shown that this can be an effective method for reducing bias (Hill, Reiter, & Zanutto, 2004) (Dehejia & Wahba, 1999).

Matching can be nonetheless problematic, as it is sensitive to a number of researcher decisions. There are two main categories of methods by which matching is often done: propensity score-based matching (Rosenbaum & Ruben, 1983) and distance-based methods, like Mahalanobis distance (Rubin, 1980). Within each of these categories, there are many algorithms that can be used for matching, leading to different effect estimates. Beyond the choice of matching algorithm, propensity score methods also require specification of a model by which to estimate the probability of receiving treatment. Similarly, distance-based methods require the researcher to choose potential confounders to use for distance calculations.

All of these choices can impact estimates. The purpose of this research is therefore to compare the bias and efficiency of three different commonly-used matching algorithms, and investigate the sensitivity of these algorithms to misspecified propensity score models and/or missing confounders. Inverse probability of treatment weights (IPTW) and optimal matching are two common propensity-score based approaches. Meanwhile, genetic matching is a relatively new method, which uses a more computationally-intensive search algorithm to calculate distances between each pair of control- and treatment-group observations, which are then used to create matched pairs that optimize balance across treatment groups with respect to a chosen set of variables (Sekhon, 2011).

Assumptions.

Causal inference based on any of these methods relies on ignorability, sufficient overlap, correct specification of the propensity score model (or balance), and the stable unit treatment value assumption (SUTVA). The ignorability assumption would require that the researcher knows and has data for all confounders. Sufficient overlap requires that the treatment and control groups are enough alike that there is a sufficiently large enough group of observations for whom inference can be made. Correct specification of the propensity score model is a very strong assumption in real life research; however, the researcher can at least perform checks on the observed covariates (i.e., post-matching balance across groups) to see if any particular model is successful at balancing the control and treatment groups. For genetic matching, a propensity score model is not required (although it can help), but balance is still a necessary assumption. Finally, SUTVA implies that any individual's outcome is independent of other individuals' treatment assignments. This is also often a strong assumption in real research. For example, the present study considers the hypothetical effect of a tutoring program on final exam grades. If a student attends a tutoring program, then they may share information and methods with friends who do not receive tutoring, thus potentially altering their friends' outcomes. This would be a violation of SUTVA.

For all of these methods, a treatment effect is eventually estimated. This is usually done via a regression with weights based on matching. By including weights in the regression, the researcher can essentially restrict the region of inference to an area with sufficient overlap, which relaxes some of the assumptions that would be necessary without matching (i.e., the model only needs to be correctly specified in the region of overlap). However, the usual assumptions of linear regression still apply within the region of inference: linear functional form, homoscedasticity, independence across observations, and normality of the error terms.

Estimators.

IPTW: First, using the treatment assignment and all measured pre-treatment covariates (i.e., possible confounders), a propensity score model is estimated, often using either logistic or probit regression. The model is then used to estimate propensity scores for each observation (i.e., the estimated probability of receiving the treatment). Finally, (assuming that the estimand of interest is the average treatment effect on the treated [ATT]), the treated observations get a weight of 1, while the controls are given weights of p/(1-p) (where p is the estimated propensity score for each control unit). In other words, controls that have higher estimated probability of being treated, and therefore are more similar to members of the treated group, get larger weights. Finally, these weights can be used in a regression to estimate the ATT (or a different estimand depending on how the weights were defined).

One advantage of this method is that no control observations are thrown out completely. Controls that are very dissimilar to members of the treated population simply get very small weights when estimating effect sizes.

Optimal 1:1 matching: Similar to the above method, optimal propensity score matching requires estimation of propensity scores using some researcher-determined model. Next, assuming that the estimand of interest is again the ATT, every treatment group observation is assigned a control observation. Pairings are optimized such that the sum of the differences between propensity scores of matched units is minimized. Control units that are not matched are given weights of 0, while all matched units are given weights of 1. These weights are then used in a regression to estimate the ATT. For this simulation, optimal matching was implemented using the optmatch package in R (Hansen & Klopfer, 2006).

Optimal matching to estimate the ATT therefore requires that there are more control units than treated units. This is most useful when the researcher has a large pool of control observations and wants to choose the best subset of those potential controls to use as counterfactuals for a treated group. If the pool of control units is significantly larger than the treated group, this method could also be generalized to match k units to each treated observation. This would allow the researcher to keep more data; however, in some instances control units that are very different from the treatment group may get matched by necessity (unless some limit is set).

Genetic matching: All matching methods (including those described above) are based on finding control units that are similar to treated units, with the ultimate goal of balancing the treated and control groups on the basis of all confounders. As mentioned above, similarity is usually operationalized in one of two ways: a) as the difference between propensity scores or b) Mahalanobis distance. However, matching based on these methods can sometimes make balance worse with respect to some confounders (Sekhon, 2011). Genetic matching is an algorithm that allows the researcher to search over the space of potential distance metrics and choose the one that results in optimal balance across a set of covariates chosen by the researcher (Sekhon, 2011). The researcher also has the option to include estimated propensity scores (generally estimated using a logistic or probit model, as above). If estimated propensity scores are provided, the genetic search algorithm will use them if they are the best method by which to optimize balance; however, this algorithm has the flexibility to look at other potential metrics (i.e., Mahalanobis distance or a weighted version of Mahalanobis distance) (Sekhon, 2011).

The result is matched pairs minimizing the distance between paired units (almost identical to the above) based on the chosen distance metric. For comparison's sake, I used 1:1 optimal matching; however, 1:k matching can also be easily implemented. One small difference is that the GenMatch function in the Matching package (Sekhon, 2011) allows the researcher to specify rules for "ties" (i.e., instances when two control observations are equally good matches for a treated observation, within some tolerance). For the purposes of this analysis, I allowed ties within the default distance (0.00001 units), which very rarely resulted in some treated units being matched to multiple controls with equal weighting.

Comparison: All of these methods are similar in that they use some operationalized distance measure to specify an optimized counterfactual for a treatment group. IPTW and propensity-score-based optimal matching both require specification of a propensity score model. In practice, a researcher would often need to try a variety of different models in order to choose one that optimizes balance with respect to some set of covariates. Genetic matching takes some of the guess-work out by searching for a metric that optimizes balance, without the researcher needing to specify any models. This certainly reduces the amount of work that the researcher needs to do in order to find an appropriate propensity score model. It also reduces the researcher degrees of freedom and is more reproducible. Finally, given that the ultimate goal is to achieve balance with respect to observed confounders, genetic matching is sometimes the fastest and most effective way to get optimal balance.

On the other hand, one benefit of IPTW as compared to the other methods is that no data is thrown out. Also, in a situation where there are a similar number of controls and treated units, optimal and genetic matching may not improve balance sufficiently and may therefore lead to biased estimates. For example, consider the limiting case where there are an equal number of treated and control units. In this case, both optimal and genetic matching would result in every control unit being matched to a treated unit with a weight of 1. Therefore, even if the control and treated units are paired optimally, the ATT estimate will be no different than if matching had not been used at all.

Simulation set-up.

Features: For this analysis I am interested in understanding the sensitivity of these matching methods to missing confounders. In the first simulation, I am interested in confounders that are completely missing, while in the second simulation I am interested in sensitivity to missing interaction effects and higher order polynomial terms. If confounding variables are missing, then even if the researcher is able to achieve balance on observed confounders, ignorability will not hold.

Estimand: The estimand of interest is the average treatment effect on the treated, or ATT. The ATT is the average difference in expected outcomes for individuals who receive the treatment. For example, in this analysis, I am estimating the effect of a hypothetical tutoring program on final course grades. The ATT is the average difference in final grades for the students who completed the tutoring program compared to what their grades would have been if they had not completed the tutoring program. Because these students all received the treatment (i.e., I cannot observe their grades under a no-tutoring control condition), I am choosing control units that are similar based on observed covariates as a counterfactual. In this hypothetical example, some students may never consider or be offered tutoring; for example, a school probably does not want to provide extra resources to students who are already getting high grades in the course. These students might have a treatment effect of zero, which could lead to a small average treatment effect for the entire population, even if the treatment effect for struggling students (who are most likely to receive tutoring) is large. In this type of situation, it is probably not interesting or useful to estimate treatment effects for people who would never receive the treatment anyways; the ATT is more actionable.

Data generating process: For both data generating processes, I am using covariates from a real dataset (Cortez & Silva, 2008), which has information for 349 high school math students from a single school. I am using the following variables from this data: sex, age, traveltime (ordered categories: home to school travel time), failures (number of past class failures), romantic (binary indicator of whether the student reports being in a romantic relationship), famrel (ordered categories: reported quality of family relationships), goout (ordered categories: reported frequency of going out with friends), Walc (ordered categories: reported weekend alcohol consumption), absences (number of school absences), health (ordered categories: current health status), and G1 (first semester grade in Math class, on a scale from 0-20). The simulated outcomes are hypothetical final course grades, while the hypothetical treatment is an in-school tutoring program. Both simulations look at sampling distributions (as opposed to randomization distributions) of ATT estimates, compared to the population average treatment effect on the treated (PATT). Thus, new outcomes and treatment assignments are sampled in each iteration.

Set-up A: In the first simulation study, all of these covariates are confounders. The logit of any individual's probability of treatment is a linear function of all covariates with normally distributed errors. Potential outcomes (i.e., expected final grades for individuals if they do or do not receive tutoring) are also additive linear functions of all covariates, with normally distributed errors. The treatment effect for all individuals is 3. On average, 38% of the population was assigned to the treatment group. While this set-up is highly improbable in real life, this simulation was used to obtain a baseline for how well these methods would perform (or fail) in a simple world.

Two simulations were run based on data from this data generating process. In the first, all covariates were used to estimate propensity scores, and all covariates were included in the regression analysis. In the second, six of the confounders were dropped. This simulation demonstrates the performance of each of these matching methods in a perfect world, then demonstrates potential bias that can occur when confounders are not included in an analysis, violating ignorability.

Set-up B: In the second simulation study, all of these covariates are confounders, plus there are six additional confounders in the form of squared terms and interaction effects. The logit of any individual's probability of treatment is a linear function of all covariates, squared terms, and interaction effects, with normally distributed errors. Potential outcomes (i.e., expected final grades for individuals if they do or do not receive tutoring) are similarly defined. The treatment effect is a function of pretest score, such that the treatment effect for any individual is 4-.1*pre-test. Thus, students who already have high pre-test scores will have smaller treatment effects. Again, about 38% of the population was assigned to the treatment group on average. Based on 100,000 simulations, the true ATT for this population is 2.84.

Two simulations were run based on data from this data generating process. In the first, all confounders were used to estimate propensity scores, and all confounders were included in the regression analysis. In the second, all of the squared terms and interaction effects were ignored. This simulation demonstrates the performance of each of these matching methods in a world where all confounders are known, then demonstrates potential bias that can occur when all variables are accounted for, but more complex relationships between variables are ignored. This again violates ignorability.

Simulation results.

Set-up A: All three methods performed similarly with respect to bias and efficiency under both simulation conditions (i.e., when the model was correctly vs. incorrectly specified) (Table 1; Figure 1). In the condition where all confounders were used for balance and treatment effect estimates, all three matching methods led to unbiased ATT estimates; however, optimal matching produced ATT estimates with the smallest root mean squared error (RMSE). When confounders were omitted, IPTW was least biased, but optimal matching still led to smaller RMSE by a hair. Meanwhile, genetic matching was more biased than the other two methods in the misspecified condition and had the largest RMSE in both conditions. Based on mean absolute standardized differences between groups after matching (Hill, Weiss, & Zhai, 2011), optimal and genetic matching did a slightly better job than IPTW of balancing treatment and control groups with respect to “observed” confounders (“observed” refers to variables that were used to obtain ATT estimates). In the setting where confounders were omitted, all three methods performed similarly with respect to balance on “unobserved” confounders.

Set-up B: All three methods performed similarly with respect to bias and efficiency under both simulation conditions (i.e., when the model was correctly vs. incorrectly specified) (Table 2; Figure 2). In the condition where all confounders were used for balance and treatment effect estimates, IPTW and genetic matching led to unbiased ATT estimates, while optimal matching led to very slight bias. However, in the condition where the model was mis-specified, optimal matching led to the least biased ATT estimates. Also, in both conditions, optimal matching produced the smallest RMSE, suggesting that it was more efficient than the other two methods. Meanwhile, genetic matching was more biased than the other two methods when confounders were omitted and had the largest RMSE in both conditions. Based on mean absolute standardized differences between groups after matching, optimal matching did a slightly better job than IPTW or genetic matching of balancing treatment and control groups with respect to “observed” confounders. In the setting where confounders were omitted, all three methods performed similarly with respect to balance on “unobserved” confounders.

Table 1: Simulation Results, Set-Up A

	Bias	RMSE	Mean absolute standardized diff. after matching: “observed” confounders	Mean absolute standardized diff. after matching: “unobserved” confounders
IPTW (correct model)	0.01	0.16	0.24	—
Optimal Matching (correct model)	0.01	0.13	0.22	—
Genetic Matching (correct model)	0.01	0.18	0.22	—
IPTW (wrong model)	0.39	0.49	0.38	0.16
Optimal Matching (wrong model)	0.42	0.48	0.35	0.15
Genetic Matching (wrong model)	0.46	0.57	0.35	0.16

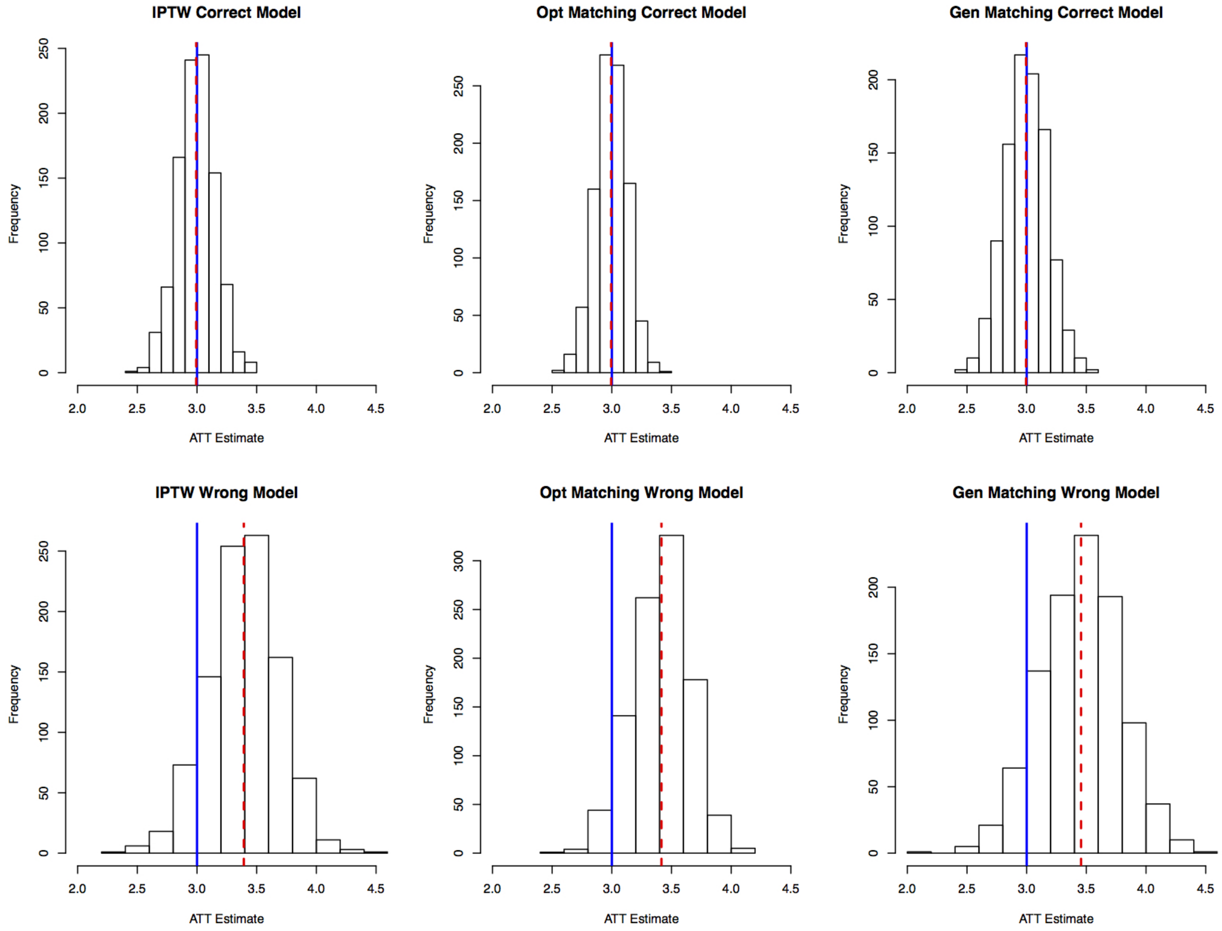

Figure 1: Sampling distributions of the ATT (Set-Up A). Red dotted lines indicate the mean of the sampling distribution; blue lines indicate the PATT

Table 2: Simulation Results, Set-Up B

	Bias	RMSE	Mean absolute standardized diff. after matching: “observed” confounders	Mean absolute standardized diff. after matching: “unobserved” confounders
IPTW (correct model)	0.00	0.15	0.20	—
Optimal Matching (correct model)	0.04	0.13	0.17	—
Genetic Matching (correct model)	0.01	0.18	0.19	—
IPTW (wrong model)	0.15	0.24	0.15	0.25
Optimal Matching (wrong model)	0.12	0.21	0.13	0.23
Genetic Matching (wrong model)	0.20	0.29	0.16	0.25

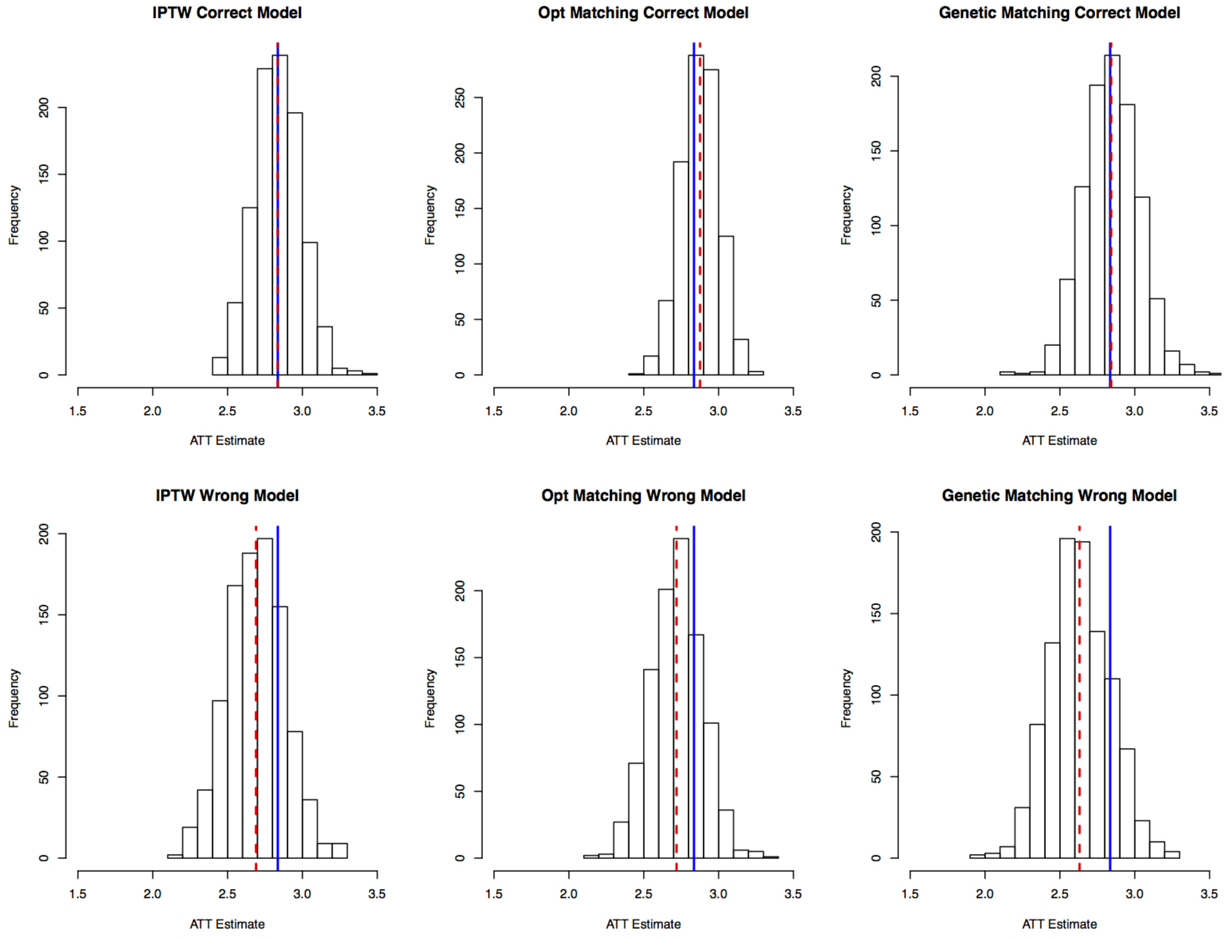

Figure 2: Sampling distributions of the ATT (Set-Up B). Red dotted lines indicate the mean of the sampling distribution; blue lines indicate the PATT

Discussion.

In both sets of simulations, these three matching methods performed similarly in terms of bias and efficiency. Based on RMSE, optimal matching outperformed the other methods in all four simulation settings. Meanwhile, genetic matching was the most biased of the three methods when confounders were omitted. One possible explanation for this is that I did not provide the Genetic Matching algorithm with a propensity score model, and also chose to keep the population size parameter at the default value of 100 so that I could run 1000 iterations in a reasonable amount of time. The function information for GenMatch explains that good solutions to the optimization problem are asymptotic to population size (Sekhon, 2011); therefore, given the similarities between genetic matching and optimal matching after a distance metric is chosen, I would imagine that genetic matching should perform similarly to optimal matching if this parameter is increased. A future study could test this.

It is important to note that, in an earlier version of this simulation, both genetic and optimal matching led to biased ATT estimates, even when the model was correctly specified. In contrast, 1-1 greedy matching with replacement was unbiased, leading me to believe that the issue was related to an insufficient number of comparable control observations to match with treatment observations. This problem was accidentally fixed by a small change in the data generating process; however, it still made me wary that it is important to look at overlap carefully before choosing a matching method. In general, it seems like 1-1 matching without replacement (i.e., optimal and genetic matching) are most useful when the pool of controls is much larger than the treatment population, and there is plenty of overlap between the treatment and control groups.

A huge pro of the genetic matching algorithm is that it performed similarly to the other two methods, despite the lack of propensity score model. However, genetic matching does not completely free the researcher from model specification. By default, the GenMatch function produces a treatment effect estimate that is based on a difference in group means, rather than a regression with all other confounders. These effect estimates were biased, even when all confounders were used for balancing. Therefore, even after genetic matching, I would suggest that treatment effects still be estimated using a regression model with available variables as controls.

In conclusion, this simulation would lead me to choose different methods depending on my data. If I had similarly sized treatment and control groups, or if it seemed like a large portion of the control group was very different from the treated group, I would choose IPTW. On the other hand, if I had a large enough control group with sufficient overlap, genetic matching seems like an efficient way to achieve balance without needing to iterate on different propensity score models. Regardless of matching method, I will estimate the ATT using a regression with all variables that I balanced on.

References

Dehejia, R. H., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. ***Journal of the American Statistical Association, 94***(448), 1053-1062. doi:10.1080/01621459.1999.10473858

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. ***Journal of Computational and Graphical Statistics***, ***15***(3), 609–627.

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A Comparison of Experimental and Observational Data Analyses. ***Applied Bayesian Modeling and Causal Inference from Incomplete−Data Perspectives***, 49-60. doi:10.1002/0470090456.ch5

Hill, J., Weiss, C., & Zhai, F. (2011). Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative. ***Multivariate Behavioral Research, 46***(3), 477-513. doi:10.1080/00273171.2011.570161

Olmos, A., & Govindasamy, P. (2015). Propensity Scores: A Practical Introduction Using R. ***Journal Of MultiDisciplinary Evaluation, 11***(25), 68-88. Retrieved from http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/431

P. Cortez and A. Silva. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

R Core Team. (2019). ***R: A Language and Environment for Statistical Computing***. Vienna, Austria: R Foundation for Statistical Computing.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. ***Biometrika***, ***70***(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. (1980). Bias Reduction Using Mahalanobis-Metric Matching. ***Biometrics, 36***(2), 293-298. doi:10.2307/2529981

Sekhon, J. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. ***Journal of Statistical Software, 42***(7), 1 - 52. doi:http://dx.doi.org/10.18637/jss.v042.i07