NERA Conference Proceedings 2018

Northeastern Educational Research Association (NERA) Annual Conference

2018

A New Heuristic for Propensity Score Matching in Observational Studies

Logan Rome
Curriculum Associates, lrome@cainc.com

Elizabeth Patton
Curriculum Associates, bpatton@cainc.com

A New Heuristic for Propensity Score Matching in Observational Studies

Logan Rome & Bess Patton

Curriculum Associates

Abstract

Propensity score matching (PSM) is a popular technique for selecting a sample in observational research that mimics the desirable qualities of a randomized controlled trial. This paper introduces a new algorithm for propensity score matching that iteratively selects only the mutual best matching treatment-control pairs. The new approach, referred to here as iterative matching, is compared to the popular nearest neighbor with caliper method. The utility and importance of the new algorithm is demonstrated in an applied example through an ANCOVA examining the efficacy of *i-Ready Diagnostic* and *Instruction* in improving scores on the Florida Standards Assessment (FSA). Results show that the iterative matching algorithm results in fewer matched pairs than nearest neighbor with caliper; however, when the treatment-to-control ratio is balanced in the sampling pool, iterative matching tends to result in slightly higher quality matches. In the applied example, the effect of *i-Ready* on FSA scores, controlling for prior year FSA scores, is statistically significant for a sample constructed using iterative matching, but not for the nearest neighbor with caliper-matched sample or the unmatched sample. Overall, this study demonstrates the importance of PSM and the choice of PSM method while also providing efficacy evidence for *i-Ready Diagnostic* and *Instruction*.

A New Heuristic for Propensity Score Matching in Observational Studies

In observational studies, where subjects cannot be randomly assigned to treatment and control groups, propensity score matching (PSM) can be used to ensure the equivalence of groups. This equivalence allows researchers to conclude that the results of statistical analyses conducted after matching are due primarily to the effect of the treatment and not to preexisting differences between the treatment and control groups. Thus, PSM attempts to mimic the best characteristics of randomized controlled trials using observational data (Austin, 2011a).

This study outlines a new heuristic approach for using propensity scores to match treatment and control units. The new approach, called iterative matching, selects only the mutual best matching treatment and control combinations iteratively in order to minimize the differences between the two groups on the specified matching variables. This paper begins by introducing propensity score matching and the nearest neighbor with caliper approach. Next, the iterative matching approach is introduced and the two approaches to PSM are compared in a series of real-data examples. Finally, the utility of PSM is demonstrated through an applied data analysis on the matched and unmatched samples that examines the efficacy of Curriculum Associates' *i-Ready Diagnostic* and *Instruction* products.

Propensity Score Matching

While randomized controlled trials are often referred to as the "gold standard" in research, these studies are often not feasible, particularly in education research. Observational studies, on the other hand, are often much more manageable. The key disadvantage, however, to observational research is the fact that the researcher has no control over which individuals are assigned to the treatment or control groups. Therefore, it's possible, and oftentimes likely, that individuals in the treatment group differ systematically from those in the control group on a number of important characteristics. These differences can often be confounded with the outcome of the study, making it difficult to make causal inferences from observational research.

Propensity score matching attempts to select a sample from a pool of treatment and control units, such that the two groups are similar on a set of key matching variables. The propensity score, as defined by Rosenbaum and Rubin (1983), is the probability of assignment to the treatment, conditional on a set of covariates. This probability can be estimated using logistic regression, where the outcome is group assignment (treatment or control) and the covariates are predictors in the model (Austin, 2011a). After calculating the propensity score for each unit in the pool, a matching method is used to select the final sample of treatment and control units.

Two classes of matching algorithms are commonly employed: optimal and greedy matching. Optimal matching is based on network-flow theory and attempts to minimize the difference in propensity scores between the treatment and control group. Greedy matching, on the other hand, selects the best matching control group for each treatment group, one at a time. Greedy approaches are called “greedy” because the order of selection matters; the algorithm selects the best matching control unit for each treatment unit sequentially, regardless of whether a future treatment group would result in a better match. While optimal matching is theoretically better than greedy matching, it is difficult to implement, and the improvement over greedy approaches in practice is often minimal (Gu & Rosenbaum, 1993). A commonly used greedy matching method, nearest neighbor matching, will serve as a comparison in this study.

Nearest Neighbor

Nearest neighbor matching is one of the more popular algorithms for PSM in education. Nearest neighbor matching can be thought of as a family of matching methods, with specific constraints available which will dictate the results of the matching process. This study makes use of caliper-based nearest neighbor matching without replacement, which will be described subsequently.

The method begins by obtaining the propensity scores from a logistic regression model. Next, treatment units are randomly ordered, and, for each treatment unit, the best matching control unit is selected. Matches are limited to those treatment and control pairs whose propensity scores fall within a predefined caliper width (Austin, 2014). Both the successfully matched treatment and control units are removed from the pool of available sampling units. The process then repeats for the next treatment unit and ends either when all treatment or control groups have been successfully matched or when no more matches within the desired caliper width remain in the sampling pool.

Iterative Matching

This study introduces a new PSM algorithm, referred to as iterative matching, that may improve upon the nearest neighbor with caliper approach and is easier to implement than optimal matching. In the new approach, treatment-control pairs are selected iteratively, with only the mutual best matches selected in each iteration. Specifically, the approach can be described in the following steps:

1. Obtain the propensity scores from the logistic regression model.
2. For each treatment unit, select the best matching control unit (the unit with the closest propensity score) within the specified caliper width.
3. For each control unit chosen in step 2, select the best matching treatment unit. The resulting treatment-control pairs are mutual best matches for one another.
4. Remove the successfully matched pairs from the pool of available sampling units.
5. Repeat steps 2 through 4 until no possible pairs remain within the specified caliper width.

By selecting all the mutual best matches at once, iterative matching, unlike nearest neighbor with caliper, is not dependent on the order of selection. Additionally, iterative matching can be implemented simply in the programming language of choice and does not require a specialized software package like optimal matching. These characteristics make iterative matching an attractive option for applied researchers conducting observational research.

Method

This study has two goals: (1) to compare the performance of nearest neighbor with caliper and iterative matching on several sampling pools with varying treatment-to-control ratios; and (2) to demonstrate the utility of propensity score matching in observational research.

Propensity Score Matching

Twelve sampling pools were created using publicly available state test data from the Florida Standards Assessment (FSA) for the 2015–2016 and 2016–2017 school years; with one sampling pool for each subject (ELA and Math) and grade (3 through 8) combination. The FSA data was merged with data from Curriculum Associates' *i-Ready Diagnostic* and *Instruction*, an online adaptive assessment and instruction suite. Schools that used *i-Ready* with fidelity (defined as at least 75% of students in a given grade taking *i-Ready Diagnostic* three times and having at least one *i-Ready Instruction* lesson) during the 2016–2017 school year were defined as the treatment group, while schools that did not use *i-Ready* at all were labeled as the control group. Table 1 summarizes the 12 sampling pools.

Table 1 Sampling pools used for propensity score matching

Subject	Grade	Treatment Groups	Control Groups	$\frac{\textbf{Treatment}}{\textbf{Control}}$
ELA	3	831	782	1.06
	4	803	797	1.01
	5	760	803	0.95
	6	180	664	0.27
	7	159	629	0.25

	8	155	633	0.24
Math	3	879	850	1.03
	4	899	820	1.10
	5	844	822	1.03
	6	211	684	0.31
	7	193	644	0.30
	8	158	613	0.26

To compare nearest neighbor with caliper and iterative matching, PSM was conducted using both methods on each of the samples from Table 1. To ensure only quality matches were selected, both approaches used a caliper width of 0.2 standard deviations of the logit of the propensity score and units were matched based on the logit of the propensity score, rather than the score itself. This approach to PSM, and the chosen caliper width, are widely used and were first recommended by Austin (2011b). Four school-level variables and two district-level variables were included in each logistic regression model used to obtain propensity scores. The school-level variables were locale (Locale; categorical with 12 levels[1]), total population (Total), percentage of students with free or reduced-price lunch (FRPL), and percentage of nonwhite students (Nonwhite). The district-level variables were percentage of students who are English Language Learners (ELL), and percentage of students in special education (Spec. Ed.). These variables came from the school- and district-level demographic data released annually by the National Center for Education Statistics (U.S. Department of Education). Iterative matching was conducted using a macro written in SAS, while nearest neighbor with caliper matching was done using the MatchIt program version 3.0.2 in R (Imai, 2018).

The PSM methods are evaluated on two criteria: (1) the number of matches obtained; and (2) the quality of the matching. For the first criterion, the percentage of possible matches was calculated as the total number of matched pairs divided by the minimum of the total control or

[1] City-Large, City-Midsize, City-Small, Suburban-Large, Suburban-Midsize, Suburban-Small, Town-Fringe, Town-Distant, Town-Remote, Rural-Fringe, Rural-Distant, Rural-Remote

treatment groups in the sampling pool. The second criterion was evaluated in several ways. First, plots of the distributions of the treatment and control groups on each variable were compared prior to and after matching. Next, matches were compared using descriptive statistics and significance tests. Specifically, standardized mean differences and *t*-tests between the treatment and control groups were calculated prior to and after each matching. Groups were considered significantly different if the *p*-value associated with the *t*-test was less than 0.05 or if the standardized mean difference had an absolute value greater than 0.1 (Nyugen et al., 2017; Yang & Dalton, 2012).

Applied Analyses

To demonstrate the utility of PSM and how results may differ for matched and unmatched samples, an analysis was conducted on all three samples (unmatched and matched using the two PSM methods) to investigate the efficacy of *i-Ready* through its impact on school-level FSA results. Specifically, an Analysis of Covariance (ANCOVA) was conducted to look at differences in 2016–2017 FSA scores between the treatment and control groups in grade 4 ELA, controlling for 2015–2016 FSA scores. The purpose of this applied example is twofold: (1) to compare the results of applied analyses with samples that are unmatched and samples that are matched using two different matching methods; and (2) to demonstrate the efficacy of *i-Ready Diagnostic* and *Instruction*.

Results

Propensity Score Matching

Results of the 12 PSMs using both nearest neighbor with caliper and iterative matching are evaluated first in terms of the percentage of possible matches from the entire pool. These results are shown in Figure 1, with sampling pools ordered by ascending treatment-to-control ratio (r). Nearest neighbor with caliper resulted in more matches across all sampling pools. However, this difference was small; ranging from 1 to 5 percentage points. The advantage of nearest neighbor with caliper was greater when the number of treatment and control units was balanced, while the two methods performed more similarly when the treatment-to-control ratio was low.

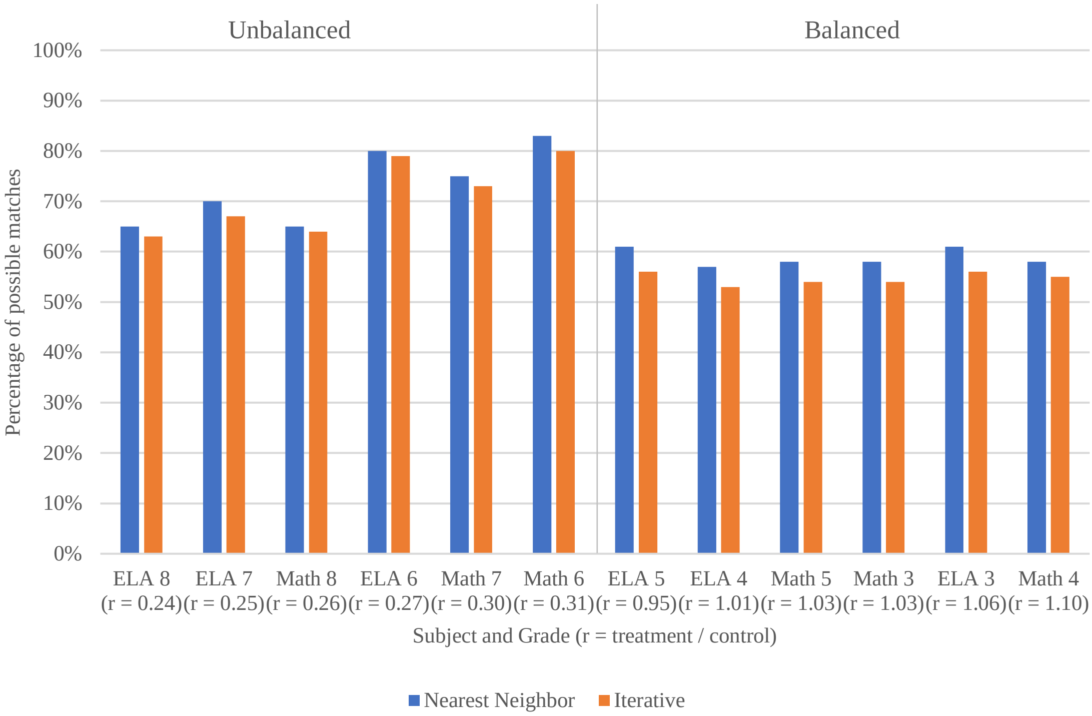

Figure 1. Percentage of possible matches for nearest neighbor with caliper and iterative matching across sampling pools, ordered by treatment to control group ratio.

While achieving a large sample size from PSM is ideal, the quality of the matches is most important, as the purpose of PSM is to minimize group difference on the matching variables. *t*-tests comparing means for the treatment and control groups on all five numeric variables prior to matching revealed significant differences on FRPL, Nonwhite, ELL, and Spec. Ed. in each sampling pool. For the samples obtained after nearest neighbor with caliper matching, none of the variables had significant mean differences between groups. Only one variable showed a significant mean difference after iterative matching (Spec. Ed. in grade 4 Math). Table 2 shows the percentage of variables that had a standardized mean difference of greater than 0.1 after nearest neighbor with caliper and iterative matching, separated for balanced (treatment-to-control ratios near 1) and unbalanced (treatment-to-control ratios less than 0.5) pools. Grades 3, 4, and 5 for both ELA and Math were considered to have balanced ratios, while the middle school grades were considered unbalanced.

Table 2 Percentage of variables with meaningful standardized mean differences, by matching method and treatment to control group ratio

Matching Method	Balanced (grades 3–5)	Unbalanced (grades 6–8)
Nearest Neighbor with Caliper	22.2%	30.6%
Iterative Matching	19.4%	38.9%

In general, the quality of matching was lower when the numbers of treatment and control units were unbalanced in the sampling pool. When the groups were balanced, iterative matching showed a higher quality of matching than nearest neighbor with caliper. When the groups were unbalanced, nearest neighbor with caliper showed an advantage. Comparing the matching methods within each sampling pool, in terms of the number of variables with standardized mean differences greater than 0.1, nearest neighbor with caliper showed an advantage in five of the 12 matchings (two balanced; three unbalanced), while iterative matching showed an advantage in three matchings (two balanced; one unbalanced); neither method showed an advantage in the remaining four pools.

Applied Example

The applied example focuses on grade 4 ELA, as the results demonstrate how the results of an analysis can differ depending on whether or not samples are matched and what matching method is used. Prior to matching, the treatment and control groups differed significantly on mean FRPL, Nonwhite, ELL, and Spec. Ed.; the standardized mean difference was greater than 0.1 on all matching variables, with the exception of Total. After nearest neighbor with caliper matching, no variables had significant mean differences, but Locale and Nonwhite both had standardized mean differences greater than 0.1. No variables had significant mean differences or standardized mean differences greater than 0.1 after iterative matching. Figure 2 shows the standardized mean differences for each matching variable, by matching method for grade 4 ELA.

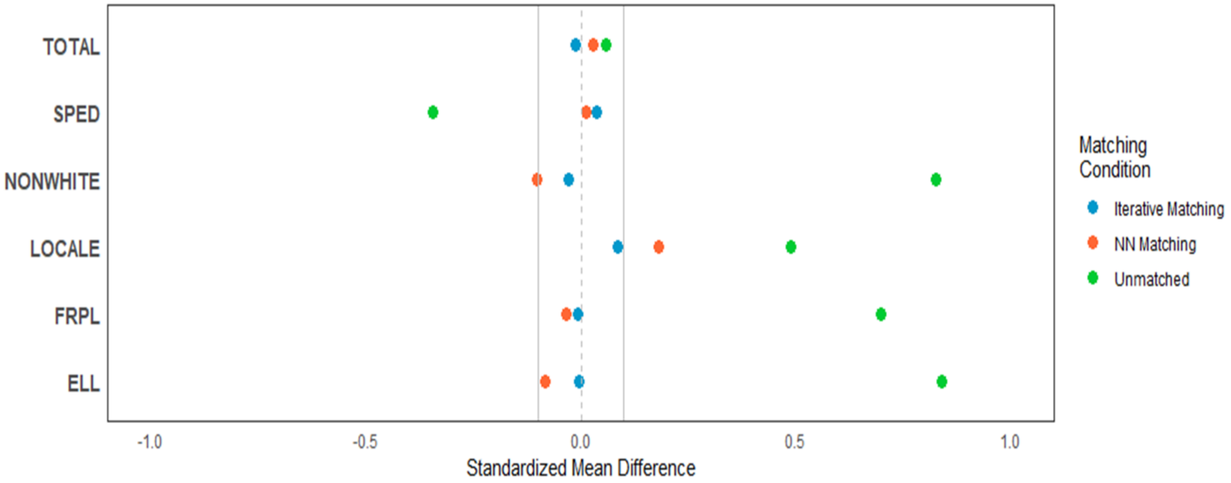

Figure 2. Standardized mean differences for each matching variable prior to matching and after both nearest neighbor with caliper and iterative matching.

ANCOVAs examining differences in 2016–2017 grade 4 ELA FSA scores between *i-Ready* users and non-*i-Ready* users were conducted on all three samples: unmatched, matched with nearest neighbor with caliper, and matched with iterative matching. The effect of treatment on 2016–2017 FSA scores, controlling for 2015–2016 scores, was not significant on either the unmatched sample (t = 0.33, p = 0.742) or the sample created using nearest neighbor with caliper (t = 1.87, p = 0.062). After matching with the iterative method, however, the group effect was significant (t = 3.15, p = 0.002). The least squares mean 2016–2017 FSA scores, controlling for 2015–2016 FSA scores, were 312.45 for the treatment group and 311.53 for the control group; the *Cohen's D* effect size was 0.22. These results demonstrate the importance of quality PSM. While there was a significant treatment effect for *i-Ready* in the sample created using iterative matching, which had the highest quality matches, this effect was masked by differences in covariates for the unmatched and nearest neighbor with caliper-matched samples.

Discussion

This paper introduced a new approach to propensity score matching that iteratively selects only the mutual best matches for the treatment and control groups. The new approach was compared to the widely used nearest neighbor with caliper matching method across 12 different sampling pools with varying treatment to control group ratios. Finally, an ANCOVA was conducted on an unmatched sample, and samples matched using the two PSM methods. The analysis demonstrated how results may differ between matched and unmatched samples, as well as between matched samples created using different matching approaches. The results also provided some evidence of the efficacy of Curriculum Associates' *i-Ready Diagnostic* and *Instruction* products in improving school-level state test scores in grade 4 ELA.

While the iterative method makes sense intuitively, comparisons with the nearest neighbor with caliper approach show mixed results. Nearest neighbor with caliper consistently results in more matched pairs and, overall, shows a slightly higher quality of matches, as measured by *t*-tests and standardized mean differences. However, iterative matching appears to result in slightly higher quality matches when the treatment and control groups are balanced in the sampling pool. This is shown through a lower percentage of instances where matched treatment and control groups have standardized mean differences greater than 0.1. At the very least, these results show that researchers conducting PSM should try several methods and compare both the quantity and quality of the matches prior to deciding on a final sample.

In the applied example, the effect of *i-Ready* on 2016–2017 FSA scores, controlling for 2015–2016 FSA scores, for grade 3 ELA is not significant when the ANCOVA is conducted on the unmatched and nearest neighbor with caliper matched samples. However, a significant effect is found when the analysis is conducted on the sample created from iterative matching. The ANCOVA demonstrated how applied analyses can be affected by the PSM method. The example also demonstrated how PSM can be applied to observational research; two matching methods were conducted and the sample with the higher quality matches, in this case iterative matching, resulted in a significant ANCOVA. In applied analyses, it is recommended to try multiple PSM approaches and select the sample with the better matches for the subsequent analyses.

Limitation and Future Directions

There are some notable limitations to this study. First, while the matching techniques were replicated on 12 sampling pools, there is some clear overlap in these pools, as all the data comes from one state's worth of schools. That is, the samples that differed only by the subject (e.g., grade 3 ELA vs. grade 3 Math) are bound to have many overlapping schools, making the sets of PSMs more of a replication than an application to new data. Additionally, since student-level data was not collected for this study, results of the applied analyses should be interpreted with caution. The main goal of this paper was to showcase the iterative matching method; thus, more research on *i-Ready* is needed to yield inferences regarding its impact on state test performance. Future research on the efficacy of *i-Ready* could leverage the iterative matching approach to more fairly examine the impact of the program.

This study provides an overview of a new PSM method and demonstrates the method on several samples; however, more work is needed to evaluate the utility of iterative matching. The results in this study could be due to any number of factors, such as: the school-level nature of the data, the treatment to control group ratios examined, and the variables included in the matching.

We hope that more researchers will examine iterative matching and its performance relative to other methods.

References

Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46,* 399–424.

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics, 10,* 150–161.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine, 33,* 1,057–1,069.

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

Imai, K. (2018). *Package 'MatchIt'*. Retrieved from https://cran.r-project.org/web/packages/MatchIt/MatchIt.pdf.

Nyugen, T. L., Collins, G. S., Spence, J., Daures, J. P., Devereaux, P. J., Landais, P., & Le Manach, Y. (2017). Double-adjustment in propensity score matching analysis: Choosing a threshold for considering residual imbalance. *BMC Medical Research Methodology 17*(78).

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.

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

Yang, D. & Dalton, J. E. (2012). A unified approach to measuring the effect size between two groups using SAS. Paper presented at the SAS Global Forum.