Abstract Title Page
Not included in page count.

Title: Using Local Matching to Improve Estimates of Program Impact: Evidence from Project STAR.

Author(s):

Nathan Jones
Northwestern University
Peter Steiner
University of Wisconsin--Madison
Tom Cook
Northwestern University

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Experimental designs are widely considered the preferable method of validly determining a program's causal impacts, primarily because, through the use of random assignment, experiments produce a reliable counterfactual. There will likely be cases where experiments are either not feasible or not desirable, and in such situations, alternatives to experiments will always be needed.

To determine the kinds of observational studies that are most likely to lead to unbiased results, researchers have increasingly conducted within-study comparisons, in which non-experimental designs are evaluated by comparing their causal estimates of program impacts to the estimates drawn from an experiment (e.g., LaLonde, 1986; Heckman, Ichimura, & Todd, 1997; 1998; Heckman et al., 1998; Smith & Todd, 2005). As reviews of the within-study literature in job training have shown, although observational studies frequently produce causal estimates that are different than those produced by experiments, there are characteristics of observational studies than can improve estimates; these include having a pretest measure of the outcome variable, having a local comparison group, and similar measures across the experimental and counterfactual groups (Bloom, Michalapoulos, & Hill, 2005; Glazerman, Levy, & Myers, 2003).

More recently, Cook, Shadish, and Wong (2008) compared three conditions under which observational studies may produce comparable causal estimates – regression discontinuity (RD) designs, intact matching from maximally similar groups, and using statistical procedures (e.g., propensity scores, OLS regression, instrumental variables) to attempt to make groups equivalent. Of the three kinds of studies, RD designs and intact groups appeared to reduce the most bias; and, in cases where the selection process was completely known, methods relying on statistical procedures also performed well. As Cook et al. suggest, it is in cases where researchers rely on "off-the-shelf" covariates that these studies typically fail to reduce bias.

In the educational context, a recent prominent example of a within-study comparison of experimental and quasi-experimental designs is Wilde and Hollister's (2007) analysis of data from the Tennessee class size experiment (Project STAR), in which kindergarten classrooms were randomly assigned to either regular or small class sizes to determine whether small classes affects student achievement. The authors evaluated the impact of estimates derived from propensity score matching relative to the study's experimental results (i.e., students from treatment classrooms were matched to students in control classrooms in the other 79 schools in the sample), concluding that the propensity scores do not reliably produce estimates of the "true" impact of small classes.

Given what we know from previous within-study comparisons, these results are not entirely surprising. In the absence of pre-test information on students in the control and treatment classes, Wilde and Hollister were limited to making comparison groups based on extant data on student, teacher, and school characteristics, leaving open the possibility that there were one or several unmeasured covariates that were correlated to selection and to the outcome. Further, rather than

matching based on intact local comparison groups, the authors matched at the student-level. As Cook et al. show, intact group matching will almost always reduce initial selection differences more than individual matching based on covariates.

Purpose / Objective / Research Question / Focus of Study:
Description of the focus of the research.

In this study we test whether matching using intact local groups improves causal estimates over those produced using propensity score matching at the student level. Like the recent analysis of Wilde and Hollister (2007), we draw on data from Project STAR to estimate the effect of small class sizes on student achievement. We propose a strategy for intact group matching in which we match treatment cases to control cases in other schools.

A secondary goal of this analysis is to determine whether the use of geographic covariates (including latitude and longitude, as well as Census variables) improve the quality of matches over the use of student, teacher, and school covariates alone. We hypothesize that by incorporating physical distance into our matching, we will increase our likelihood of finding maximally similar comparison classrooms.

Setting/Population
Description of the research location.

The setting for this study is a sample of 79 kindergarten classrooms in the state of Tennessee. The Project STAR data was collected beginning during the 1985 school year, although longitudinal data collection continued on students through fourth grade. In total, over 7,000 students participated in the experiment.

For the purpose of comparison with Wilde and Hollister's (2007) study, we restrict our treatment sample to the 11Project STAR schools with over 100 kindergartners, which ensures that there are multiple treatment classrooms per school and allows for 11 separate experimental and observational causal estimates; a pooled analysis is also conducted to determine the average treatment effect across the sample. In constructing the counterfactual groups, treatment schools were matched not just to the 11 comparison classrooms, but to all other 78 schools in the sample.

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.
(May not be applicable for Methods submissions)

Students in participating schools were randomly assigned to one of three settings: the treatment condition (small classrooms with 13-17 students) or one of two control conditions (regular size classrooms or regular classrooms with an instructional aide). Past studies have shown that students in the small classes generally outperform their peers in large classes (Krueger, 1999; Word et al., 1990).

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

We provide a method of matching that is likely to reduce the bias associated with previous propensity score matching techniques. For one, our preliminary findings show that matching at the school level (while maintaining intact classrooms) is a potentially useful method for obtaining less biased resulted from an observational study. And, further, our results suggest that, in the absence of pretest information, the inclusion of geographic information as a covariate can improve the quality of matches obtained.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

We match schools using Mahalanobis distance metric matching, as developed by Rubin and Cochran (Cochran & Rubin, 1973; Rubin, 1976, 1979). We restrict the covariance matrix to treated schools, given that the schools that serve as controls change depending on the covariates used in calculating Mahalanobis distances. Similarly, when the analyses are conducted at the classroom level, we use the covariance matrix of only treated classrooms within the 11 schools with over 100 students.

To find matches, schools are first randomly sorted, and the distances are calculated between each "treatment" school and all available control schools. In finding matches that minimize the Mahalanobis distance, we allowing for replacement; i.e., control schools are allowed to be used more than once.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

As described in the section on the setting/population used in our research, we demonstrate the usefulness of the method using the Project STAR data. We have chosen this dataset for a number of reasons. For one, our results expand on the within-study comparison conducted by Wilde and Hollister (2007). Additionally, Project STAR is well-known within the educational research community as being a strong example of a large-scale experiment that has shown its intervention (class size) to be effective. Lastly, the fact that all schools are drawn from a common state is advantageous for us, given that we use geographic distance as a matching strategy.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)

As suggested above, in the application of our matching, we introduce a series of models that incorporate additional covariates at each step. We begin by using our distance measures alone, i.e., latitude and longitude, which provides a match that is geographically closest to the "treatment" school. We then conduct the same analyses using only the STAR covariates: % free and reduced lunch, % minority, % special education, % repeats, and % pullouts (all of which are aggregated to the school level). Next, we run a model with both distance and the STAR covariates. Finally, we incorporate a series of Census variables to the existing model to determine whether they improve further on our matches. These variables include: median

income, percent vacant homes, percent families, percent manufacturing/professional positions, percent high school graduates, and percent with graduate/professional degrees.[*]

Based on these various matching strategies, we estimated the treatment effect by calculating mean differences between students in treatment classrooms and their matched comparison cases in each of the 11 schools. Additionally, when conducting the pooled analysis we used hierarchical linear modeling, treating students as nested within schools.

Findings / Conclusions

Initial results suggest that matching while maintaining intact groups offers advantages over previous methods for matching at the student level. We can also reduce bias further when using geographic distance to supplement the "off-the-shelf" covariates that were collected at the time of the Project STAR study.

Our results should be put into context however. There are other strategies that remain preferable to intact local matching when attempting to reduce bias in observational studies. Where are method is useful is in situations where the data available is similar to that of Project STAR, i.e., information is missing on students' pretest information and access to covariates is limited. In these cases, we believe we have provided a potentially useful improvement over student-level matching.

[*] The Census variables are taken from the 1980 Census. In 1980, only households in urban areas (>50,000) have census tracts, block groups, and blocks attached to them; it wasn't until 1990 that all areas were assigned blocks. Because we did not want to lose data on non-urban areas, our Census variables are aggregated at the city level (which means that schools within the same FIPS-MCD have the same aggregate values for each of the housing and community variables).

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Bloom, H. S., Michalopoulos, C., & Hill, C. J. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In H. S. Bloom (Ed.), Learning more from social experiments (pp. 173–235). New York: Russell Sage Foundation.

Cook, T. Shadish, W., & Wong, V. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New nindings from within-study comparisons. *Journal of Policy Analysis and Management, 27(*4), 724-750.

Heckman, J., Ichimura, H., Smith, J. C., & Todd, P. (1998). Characterizing selection bias. *Econometrica*, 66, 1017–1098.

Heckman, J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies*, 64, 605–654.

Heckman, J., Ichimura, H., & Todd, P. E. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65, 261–294.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy*, 589, 63–93.

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Testing nonexperimental estimates of impact against experimental estimates of impact with education test scores as outcomes. *Journal of Policy Analysis and Management*, 26, 455–477.