Abstract Title Page
Not included in page count.

Title: Matching Strategies for Observational Data with Multilevel Structure

Author(s): Peter M. Steiner

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

In educational research, causal inference from observational studies has gained in importance during the last decade—particularly propensity score (PS) techniques like PS matching or PS stratification are now more frequently used in estimating the effects of educational interventions. In comparison to other fields of research, observational studies in educational research typically face and additional challenge: the data usually show a multilevel structure: students are nested within classrooms and schools, or schools are nested within districts or states, for instance. Complications arise for several reasons: (i) Units within clusters are typically not independent; (ii) Interventions may be implemented at different levels (e.g., student-, classroom-, or school-level); (iii) Selection processes may simultaneously take place at different levels and involve many stakeholders (students, peers, parents, teachers, school management, parent teacher association), differ from school to school or district to district, and might introduce selection biases of different directions at different levels. Therefore, the implementation of matching techniques for removing selection bias is more challenging than for data structures with a single level only.

In this study we consider a two-level structure where students are nested within schools. Thus, treatment assignment or selection might take place either at the school level or the student level. Treatment selection at the school level implies that the treatment status only varies between schools but is constant for all students within schools (all students of a school are either assigned to the treatment or control condition). On the other hand, if treatment selection takes place at the student level students might be assigned to or self-select into the treatment or control condition within each single schools. Depending on the level of treatment selection, two main matching strategies are possible. First, if treatment selection is at the school level comparable treatment and control schools need to be matched—matching of individual students is not necessarily required. This type of matching mimics a cluster randomized controlled trial where schools are randomly assigned to treatment. Local and focal matching approaches that match geographically neighboring treatment and control schools of the same type is a promising strategy for obtaining unbiased school level treatment effects (Cook, Shadish & Wong, 2008). If school level matching results in matched schools that considerably differ on observed student level covariates an additional matching of students within matched schools might further increase the comparability of matched schools. Second, whenever treatment selection is at the student level students need to be matched within schools, thereby mimicking a randomized block design where students are randomly assigned to the treatment condition within schools (blocks). However, if extreme selection processes take place we might be confronted with a lack of comparable students within schools and, thus, be forced to look for matches from other schools (i.e., match between schools).

Though the popularity of matching approaches has considerably increased, only a few methodological papers on matching in the context of multilevel data are available (several of them unpublished): Hong & Raudenbush (2006) extend the Rubin Causal Model (Rubin, 1974) to the multilevel case and give an example on the effect of retaining students in Kindergartens. However, in estimating the retention effect on reading and math achievement scores, they matched retained and promoted students between schools and made no attempt to match students

within schools; similarly Hong, in press). Arpino & Mealli (2008), Kim & Seltzer (2007), and Thoemmes & West (2010) focused on student level matching and conducted simulation studies using different multilevel models for estimating propensity scores. Also these studies did not consider matching strategies that match students within schools—their suggested matching strategies allow for matches of students between schools.

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

Given the different possibilities of matching in the context of multilevel data and the lack of research on corresponding matching strategies, we investigate two main research questions. The first research question investigates the advantages and disadvantages of different matching strategies that can be pursued with multilevel data structures. The goal is first to outline possible matching strategies and then to identify an optimal matching strategy for different treatment selection scenarios (here, optimal refers to design aspects rather than technical aspects of a matching algorithm). In following Hong & Raudenbush (2006), theoretical foundations are discussed within the Rubin Causal Model framework and its potential outcomes notation (Rubin, 1974).
The second research question focuses on the matching of students (when treatment is implemented at the student level) in more detail. As outline above, one can either match students within schools or match students between schools. Matching within school is time-consuming and might fail due to a lack of comparable treatment and control students within schools. Matching between schools, on the other hand, can be more conveniently implemented by estimating an overall PS model for all students together. This can be done using hierarchical linear modeling. Students can then be matched within and between schools. Thus, treatment and control students might be successfully matched even if no close matches are available within schools. The question then is whether and under which conditions we can get unbiased effect from such a matching strategy.

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

This study systematically investigates and compares matching strategies in the context of hierarchical data structures. In particular, it demonstrates that some of the matching strategies suggested by methodological papers may result in biased estimates. Another novel contribution of that study is that it discusses matching strategies for different scenarios of treatment selection (i.e., selection at different levels) and that it investigates the conditions under which less optimal strategies might lead to unbiased effect estimates.

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

We investigate the second research question (i.e., strategies for student-level matching) by conduction simulation studies and analyzing a real multilevel dataset (with a two level structure: students nested within schools). First, using a small simulated example we demonstrate that

matching between schools (instead of within schools) may result in considerably biased estimates of the treatment effect. Then, a more extensive simulation study that varies sample sizes, intraclass correlations, the complexity of both the selection process and data generating outcome model (which was not done in any of the simulation studies mentioned above), degree of group overlap, and the extend of initial covariate imbalance at each level is used to simulate more realistic scenarios for educational research. In the case of matching within schools a logistic regression model is estimated for each school in order to get the estimated propensity score. When we allow for matches between schools we estimate an overall PS model using hierarchical linear models.
The simulation study is complemented by a re-analysis of Hong & Raudenbush's study on the effect of kindergarten retention on student achievement scores. While Hong and Raudenbush (2006) analyzed the data using a multilevel PS and allowed for matches between schools we test whether we get a different retention effect if we only allow for matches within schools.

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

The findings of the study help researches in identifying an optimal matching strategy for their data at hand. The results of the study also show that choosing a less than optimal strategy—a strategy that does not properly reflect the selection process—may not be able to reduce all the selection bias from the treatment effect of interest.

Findings / Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

For hierarchical data structures, theoretical considerations and preliminary simulation results clearly indicate that matching approaches for causal inference need to reflect the (multilevel) selection process. If matching does not reflect the selection process that actually took place biased treatment effect may result. If selection takes place at the student level (within schools) one should match students within school. Matching students between schools may result in biased treatment effects. However, if treatment and control students cannot be matched within schools, matching students between schools might still be considered but stronger assumptions are required. If treatment selection takes place at the school level an individual-student matching is not necessarily required.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Arpino, B., & Mealli, F. (2008). The specification of the propensity score in multilevel observational studies. *Dondena working paper, 6*. Available at www.dondena.unibocconi.it/wp6.

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.

Hong, G. (in press). Marginal mean weighting trough stratification: adjustment for selection bias in multi-level data. *Journal of Educational and Behavioral Statistics*.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*, 901-910.

Kim, J. & Seltzer, M. (2007). Causal inference in multilevel settings in which selection process cary across schools. *Working Paper 708*, Center for the Study of Evaluation (CSE), UCLA: Los Angeles.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* , 688-701.

Thommes, F., & West, S. (2010). The use of propensity scores in multisite non-randomized trials. Unpublished manuscript.