Causal Inference with Observational Multilevel Data in Educational Research

This is work done together with Jee-Seon Kim and Courtney Hall. The research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D100033.
What is Special in Educational Research?

- Data typically show a *multilevel structure*
  - students : classrooms (teacher) : schools :
    : school districts : states : countries
  - time (longitudinal data)

- *Interventions* can be implemented at different levels: student- or classroom-, or school-level
  - Long interventions / time-varying interventions

- *SUTVA* is more likely violated
  - within classrooms/schools students are not independent
  - also school might not be independent
What is Special in Educational Research?

- **Randomization** is not always possible
  - Rely on alternative designs like RDD / ITS / non-equivalent control group designs (observational studies)

- In observational studies *selection processes* can be very complex because they might
  - involve *many stakeholders* (students, peers, parents, teachers, school management, parent teacher association)
  - take place at *multiple levels* and
  - *differ* across clusters, e.g., from classroom to classroom, school to school, district to district, etc.
  - introduce *bias in opposite directions* at different levels
What is Special in Educational Research?

Observational studies in educational research frequently have a rich set of covariates in order to control for selection bias

- Large number of covariates measured at multiple points in time and different levels, particularly, direct or proxy pretest measures on the outcome of interest are frequently available (e.g. achievement scores)

- Specifying the PS model (model selection) and achieving balance on all observed covariates is more challenging
Multilevel Structure & Treatment & Matching

- To simplify matters, we focus only on a two-level structure
  - Students are nested within school

- Treatment might be implemented either at the
  - school level (school level treatment) or
  - student level (student level treatment)

- Basic matching strategy for observational studies depends on the treatment level
  - if treatment is implemented at the school level match intact schools
  - if treatment is implemented at the student level match students
School Level Treatment

- **Treatment at the *school level*:**
  - Schools select themselves or are assigned to the treatment or control condition
  - All students within a school receive the same treatment

- **RCT equivalent (school level):**
  - Cluster randomized controlled trial (*schools* are randomly assigned to the control or treatment condition)

- **Matching strategies (school level):**
  - *Matching of intact schools*
    - Standard PS model with school level covariates, incl. aggregated student level covariates
School Level Treatment

Matching strategies (cont.):

- **Local matching** using geographic information:
  - match schools that are locally very close (in the same neighborhood or school district)—the hope is that local matching controls for unobserved neighborhood characteristics
  - PS model includes geographic information (e.g., distance measure based on the longitude and latitude) as covariates with/without caliper

- Matching of schools followed by an *additional matching of students* within matched (pairs of) of schools—if matched schools show residual imbalance on student level covariates
Student Level Treatment

- Treatment at the *student level*:
  - students select themselves or are assigned to the treatment or control condition within schools (student level selection)
  - moreover, schools might choose to participate or not (school level selection)

- RCT equivalent (student level):
  - multisite RCT / randomized block design (randomization of *students* within schools)
  - overall sample of schools might be randomly or deliberately selected
Student Level Treatment

Matching strategy:

- Match students within schools
  
  - local matching: students in the treatment and control condition share the same learning, social, and geographic environment
  
  - Separate PS model for each school using student level covariates
  
  - School-specific estimates are then pooled/averaged across school
Matching strategy (cont.):

- *Match students within & between schools*
  
  - if sample sizes within schools are very small or lack of overlap within schools $\rightarrow$ need to “borrow” students from other schools
  
  - PS model is a two-level model with student and school level covariates
  
  - violates the idea of a randomized block design and of local matching
Causal Estimand & Assumption

- Simple two-level model:
  - students are nested within schools
  - treatment at student level

- Given the two-level structure, we can begin by defining a set of potential outcomes for each student $i$ in school $j$ that
  - does not need SUTVA but instead
  - depends on the assignment of peers to treatment and of students to school (Hong & Raudenbush, 2006)
Potential Outcomes

The potential outcomes $Y_{ij}(Z_{ij}, Z_{-ij}, S)$ depend on

- student i’s treatment: $Z_{ij} = 1$ if student $i$ is treated and $Z_{ij} = 0$ if student $i$ is not treated

- the peers’ treatment: $Z_{-ij}$ (treatment indicator vector)

- The assignment of students to schools: $S$

This general formulation generates a huge set of potential outcomes

- Allows for many different causal estimands (differences between average potential outcomes for specific assignments)
Causal Estimand

- **Average (causal) treatment effect** is then given by
  \[ E[Y_{ij}(Z_{ij} = 1, Z'_{-ij}, S) - Y_i(Z_{ij} = 0, Z'_{-ij}, S')] \]

- with \( Z'_{-ij}, S' \) being alternative assignments

- estimable if assignments to schools and treatment are *ignorable* (e.g. if students were randomly assigned to schools and students were randomly assigned to treatments within schools)

- **ATE generalizes** across schools and students

- In practice, it is barely estimable due to social selection and geographic segregation and the resulting lack of overlap between groups
Restricted Causal Estimands

In order to obtain causal estimates in practice we might *restrict generalizability*

- to the *observed school assignment* $S = s^*$ such that we get the average treatment effect for the observed composition of students in existing schools

\[
\tau_{s^*} = E[Y_{ij}(Z_{ij}, Z_{-ij}, s^*) - Y_{ij}(Z'_{ij}, Z'_{-ij}, s^*) | S = s^*]
\]

- or even more restrictive, to *observed selection of peers* $Z_{-ij} = z_{-ij}^*$ and *schools* $S = s^*$

\[
\tau_{z_{-ij}^*, s^*} = E[Y_{ij}(Z_{ij}, z_{-ij}^*, s^*) - Y_{ij}(Z'_{ij}, z_{-ij}^*, s^*) | Z_{-ij} = z_{-ij}^*, S = s^*]
\]
Potential Outcomes (two-level model)

- A reduction of potential outcomes is also achieved by SUTVA
  
  □ **SUTVA for all students**
  \[
  \tau = E[Y_i(Z_i,s^*) - Y_i(Z'_i,s^*) | S = s^*]
  \]

  □ **SUTVA within groups of peers** \( v = f(Z) \) (e.g. high- vs. low-retention schools, or high- vs. low-ability classes)
  \[
  \tau_v = E[Y_i(Z_i,v,s^*) - Y_i(Z'_i,v',s^*) | S = s^*]
  \]

- SUTVA for schools (seems to be less problematic)

- Restrictions of potential outcomes and plausibility of SUTVA depends on data and research question at hand
Strong Ignorability (SI)

- Average treatment effects are identified if selection is strongly ignorable.
- Depending on the matching strategy, we need set of level-1 and level-2 covariates, $X$ and $W$, (or a PS based on these covariates) such that the expectations of potential outcomes do not depend on treatment assignment:

$$(Y(0), Y(1)) \perp Z \mid X, W$$
Strong Ignorability (SI)

- Alternative formulation:
  - for $\tau_Z$: $E[Y_i(Z_i, Z_{-i}, s^*) | Z_i = z_i, Z_{-i} = z_{-i}, X = x, W = w]$
    $= E[Y_i(Z_i, Z_{-i}, s^*) | X = x, W = w]$
  - for $\tau_{restr}$: $E[Y_i(Z_i, z_{-1}^*, s^*) | Z_i = z_i, X = x, W = w]$
    $= E[Y_i(Z_i, z_{-1}^*, s^*) | X = x, W = w]$

- SI with PS: $(Y(0), Y(1)) \perp Z \mid e(X, W)$

Note: While SI with $X$ and $W$ typically implies a within-cluster matching, SI with $e(X, W)$ implies a between-cluster matching. → Why?
Simulation Study

We investigated the *matching strategies for multilevel data* in a simulation study that focuses on

- selection within clusters
- within-cluster and between-cluster matching strategies
- different PS models: single-level models, fixed effects models, random effects models
- different PS techniques: inverse-propensity weighting, PS stratification, optimal full matching
Simulation Study

- Used four different selection and outcome models to generate *16 populations* (from which we then draw random samples)
- Two level-1 covariates and two level-2 covariates
- Each population has
  - 5000 clusters
  - 300 (SD = 50) units per cluster
  - ICC of approximately 0.3
  - Homogeneous treatment effects
Data-Generating Selection Model

- 4 different selection models
Data-Generating Outcome Model

- 4 different outcome models

---

**Parallel**

**Monotonic**

**Simpson's Paradox**

**Mixture**
16 Populations

Combining the different selection and outcome models results in 16 different populations

<table>
<thead>
<tr>
<th>Selection 1</th>
<th>Outcome 1</th>
<th>Outcome 2</th>
<th>Outcome 3</th>
<th>Outcome 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>parallel</td>
<td>parallel, parallel</td>
<td>parallel, monotonic</td>
<td>parallel, Simpson</td>
<td>parallel, mixture</td>
</tr>
<tr>
<td>monotonic</td>
<td>monotonic, parallel</td>
<td>monotonic, monotonic</td>
<td>monotonic, Simpson</td>
<td>monotonic, mixture</td>
</tr>
<tr>
<td>Simpson</td>
<td>Simpson, parallel</td>
<td>Simpson, monotonic</td>
<td>Simpson, Simpson</td>
<td>Simpson, mixture</td>
</tr>
<tr>
<td>mixture</td>
<td>mixture, parallel</td>
<td>mixture, monotonic</td>
<td>mixture, Simpson</td>
<td>mixture, mixture</td>
</tr>
</tbody>
</table>
Simulation & Evaluation

Simulation

- From each population we repeatedly drew 40 clusters and 40% of the units within each cluster.
- 1000 iterations (i.e., repeated draws from a population).

Evaluation of matching strategies and analytic methods:

- remaining bias (in %): mean & 90% probability interval.
PS Estimation & ATE Estimation

- **Within-cluster matching**: estimated a separate PS model for each cluster (do not need level-2 covariates)

- **Between-cluster matching**: estimated several models
  - single level models (with and without level-2 covariates)
  - different (misspecified) fixed effects models
  - different (misspecified) random effects models

- **Doubly robust estimation**: combined the different PS methods with an additional regression/covariance adjustment using different fixed effects and random effects models
Results: PS Estimators

different PS techniques and PS model specifications

<table>
<thead>
<tr>
<th>Selection Model</th>
<th>Outcome Model</th>
<th>PS Estimators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel</td>
<td>Parallel</td>
<td>S_0 S_1 S_2 F_1 F_3 R_1 W_1 W_2</td>
</tr>
<tr>
<td>Monotonic</td>
<td>S_0 S_1 S_2 F_1 F_3 R_1 W_1 W_2</td>
<td></td>
</tr>
<tr>
<td>Simpson's</td>
<td>S_0 S_1 S_2 F_1 F_3 R_1 W_1 W_2</td>
<td></td>
</tr>
<tr>
<td>Mixture</td>
<td>S_0 S_1 S_2 F_1 F_3 R_1 W_1 W_2</td>
<td></td>
</tr>
</tbody>
</table>

ATE Estimates for PS only (no covariate adjustment)

Error Bars show 5th & 95th percentiles

○ Weighting  □ Stratification  △ Matching
Results: Doubly Robust Estimators

Correctly specified PS model

Outcome Model
- Parallel
- Monotonic
- Simpson’s
- Mixture

Selection Model
- Parallel
- Monotonic
- Simpson’s
- Mixture

PS Specification for Between-Cluster Matching = Fixed Effects: Clusters each have own slope and intercept
PS Specification for Within-Cluster Matching = Level-1 main effects

Error Bars show 5th & 95th percentiles

○ Weighting   □ Stratification   △ Matching
# Results: Doubly Robust Estimators

## Misspecified PS model

<table>
<thead>
<tr>
<th>Selection Model</th>
<th>Outcome Model</th>
<th>Parallel</th>
<th>Monotonic</th>
<th>Simpson's</th>
<th>Mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel</td>
<td>Parallel</td>
<td>E</td>
<td>S₀, S₁, R₂, F₁, F₃, W₁, W₃</td>
<td>E</td>
<td>S₀, S₁, R₂, F₁, F₃, W₁, W₃</td>
</tr>
<tr>
<td>Parallel</td>
<td>Monotonic</td>
<td>E</td>
<td>S₀, S₁, R₂, F₁, F₃, W₁, W₃</td>
<td>E</td>
<td>S₀, S₁, R₂, F₁, F₃, W₁, W₃</td>
</tr>
<tr>
<td>Parallel</td>
<td>Simpson's</td>
<td>E</td>
<td>S₀, S₁, R₂, F₁, F₃, W₁, W₃</td>
<td>E</td>
<td>S₀, S₁, R₂, F₁, F₃, W₁, W₃</td>
</tr>
<tr>
<td>Parallel</td>
<td>Mixture</td>
<td>E</td>
<td>S₀, S₁, R₂, F₁, F₃, W₁, W₃</td>
<td>E</td>
<td>S₀, S₁, R₂, F₁, F₃, W₁, W₃</td>
</tr>
</tbody>
</table>

PS Specification for Between-Cluster Matching = Level-1 and Level-2 main effects
PS Specification for Within-Cluster Matching = Level-1 main effects
Error Bars show 5th & 95th percentiles

- ○ Weighting
- □ Stratification
- △ Matching
Preliminary Conclusions

**Within-cluster estimators**
- are approximately unbiased
- do not require level-2 covariates (i.e., confounding level-2 covariates do not need to be measured)
- Might be problematic with small cluster sizes and a lack of overlap within clusters

**Between-cluster estimators**
- are surprisingly robust to model misspecifications thought they match rather heterogeneous units across clusters—except when the selection and outcome model have clusters with different slopes (a mixture negative and positive slopes)
Across-Cluster Matching within Homogenous Classes of Clusters

Matching *across cluster* but *within homogeneous classes of clusters*

Define homogeneity of clusters according to

- selection process
- outcome model
- or both

Classes membership of clusters may be manifest or latent

- Use finite mixture modeling for identifying latent classes
Across-Cluster Matching within Homogenous Classes of Clusters

Advantages

- DEALS with lack of overlap within clusters
- Strong ignorability might be more likely met (level-2 covariates are less important)
- Investigate treatment effect heterogeneity

Disadvantages

- Class memberships might need to be estimated
- Need large enough sample sizes (at level-1 and level-2)

(Kim, Steiner & Lim, 2015; Hall, Steiner & Kim, 2015)
Preliminary Conclusions

The choice of a specific *PS technique* is not very important, particularly if one uses a doubly robust estimator (inverse-propensity weighting occasionally performs less well)

Whether one uses *fixed-effects* or *random-effects models* for estimating the PS does not make a big difference either—it is presumably more important for the outcome model (i.e., estimating the treatment effect) → standard errors!

Note: These are only preliminary results and conclusions! Thus, use them with caution.