Covariate Selection and Impact Estimation for Clustered RCTs Using Lasso and Design-Based Methods

September 2019

Peter Z. Schochet, Ph.D.
Senior Fellow
Mathematica Policy Research, Inc.
P.O. Box 2393
Princeton, NJ 08543-2393
Phone: (609) 936-2783
Fax: (609) 799-0005
pschochet@mathematica-mpr.com

Background / Context:
Description of prior research and its intellectual context.

In randomized control trials (RCTs) of education interventions, random assignment is often performed at the group level (such as a school or classroom) rather than at the student level. A common challenge for these clustered designs is to obtain sample sizes with sufficient statistical power to detect target average treatment effects within study resources. Power is a concern because standard errors of estimated impacts under clustered designs must be inflated to account for the correlation of outcomes between students in the same clusters (Donner and Klar, 1997; Hedges, 2004; Schochet, 2008).

Estimating impacts using regression models that control for baseline covariates is an effective and commonly-used approach for increasing precision under clustered designs (Raudenbush, 1997; Bloom et al., 1999; Schochet, 2008). The use of covariates—such as baseline values of the primary outcomes—can primarily increase precision for these designs by explaining the variation in mean outcomes between clusters. Accordingly, a critical design issue for clustered designs is the selection of covariates, which is especially complex for studies that collect and analyze data at the individual level. In these cases, there could be more candidate baseline covariates than clusters, which adds complexity because the degrees of freedom for hypothesis testing for clustered RCTs are based on the number of clusters, not individuals.

One way to address the covariate selection problem is to pre-specify covariates in RCT registries and design documents prior to data analysis, for example, using prior information about the strength of outcome-covariate relationships. This approach has been recommended by authors across a range of fields (see, for example, Raab et al., 2000; Senn, 1994; Pocock et al., 2002; Heinz et al., 2017; European Medicines Agency, 2013). Pre-specification has the advantage that it yields impact estimates with correct Type 1 errors across repeated samplings, is fully replicable, and avoids the criticism that covariates were selected to obtain “favorable” findings.

Major RCT registries across research areas, however, do not mandate that covariates be pre-specified, such as in medicine (ClinicalTrials.gov), education (sreereg.icpsr.umich.edu), and economics (socialscienceregistry.org). Similarly, major evidence review clearinghouses (such as the What Works Clearinghouse) do not require pre-specification of covariates for RCTs to meet evidence standards. Accordingly, many RCT evaluations do not pre-specify covariates, instead selecting covariates using study data once the outcomes have been measured.

An advantage of this data-driven selection approach is that it can improve precision by identifying strong outcome-covariate relationships that may have been unanticipated. However, there are drawbacks. First, this approach can suffer from the criticism that model covariates and their functional forms were selected to yield favorable findings. Second, standard error estimation is complicated by the need to account for the randomness of the covariate selection process across replications (Berk et al., 2013; Lee et al., 2014).

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

This paper discusses a data-driven approach for selecting amongst candidate covariates for clustered RCTs that maintains key advantages of the pre-specification approach. We adapt commonly-used Least Absolute Shrinkage and Selection Operator (Lasso) procedures for covariate selection (Tibshirani, 1996) to design-based estimators (Schochet, 2013, 2018) that we show are conductive to the Lasso framework. We focus on finite-population (FP) design-based estimators, where potential outcomes and covariates are assumed fixed for the study, rather than the super-population (SP) model, typically the focus of the covariate selection literature. As shown in the paper, the use of the FP model allows us to *fix* the covariate selection process across replications, thereby simplifying standard error estimation.

After presenting the theory, the paper presents results from simulations to quantify the extent to which the procedure can recover the "true" model covariates (those with nonzero coefficients). The simulations also examine Type 1 errors of the estimated impacts. The paper also presents results from an empirical analysis using clustered RCT data from the multi-site Social and Character Development (SACD) Evaluation (SACD Research Consortium, 2010) to demonstrate how the procedure can substantially improve precision without sacrificing technical rigor.

The paper contributes to the literature in several ways. First, it provides a framework for covariate selection for clustered RCTs using non-parametric design-based estimators rather than model-based HLM (random effects) estimators discussed in the literature (see, for example, Belloni et al., 2014; Bondell et al, 2010; Lin et al., 2013). Second, it provides consistent impact estimators for clustered RCTs that are likely to improve precision relative to the pre-specification approach, while decoupling covariate selection from impact estimation to improve replicability and transparency. The procedure is simple to apply because it is *sequential* where (1) covariates are selected first using Lasso adapted to clustered designs and (2) the selected covariates are then used to estimate impacts and standard errors using FP design-based regression estimators. Our focus differs from the literature on Lasso methods to simultaneously estimate treatment effects and covariate parameters (shrunk towards zero) and efficient standard errors (see, for example, Bloniarz et al., 2015 and Ertefaie et al., 2015 for non-clustered designs and the above references for HLM models).

The paper contributes to the conference theme, *Practical Significance and Meaningful Effects: Learning and Communicating What Matters*, because it presents easy-to-use, new methods to improve the precision of RCT impact findings that underlie any discussion of how to assess whether treatment effects matter. The methods are intended to increase the replicability and transparency of impact findings, thereby improving evaluation rigor and research practice.

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

To demonstrate our method, this section briefly summarizes design-based impact estimators for clustered RCT designs (following Schochet, 2013, 2018) and then discusses how this framework can be used to select covariates using Lasso.

Design-based estimators for clustered RCTs. Consider an RCT design where M clusters are randomized to a treatment or control group and where data are collected on individuals. Under this design, the data generating process for the observed outcome for individual i in cluster j (y_{ij}) can be expressed as follows:

$$y_{ij} = T_j Y_{ij}(1) + (1 - T_j) Y_{ij}(0), \tag{1}$$

where $Y_{ij}(1)$ is the potential outcome for individual *i* in the treatment condition, $Y_{ij}(0)$ is the potential outcome for the same individual in the control condition, and T_j is the treatment status indicator. Design-based theory is based on a rearrangement of (1) to produce a regression model with an error term defined by the randomization process. This model can be estimated using weighted least squares (WLS) with weights, w_{ij}, where y_{ij} is regressed on an intercept, T_j, and v baseline covariates, $\mathbf{X_{ij}}$. Importantly, in the FP framework, the only source of randomness is T_j, as $Y_{ij}(1)$, $Y_{ij}(0)$, and $\mathbf{X_{ij}}$ are assumed fixed. This means that study results are assumed to pertain to the study sample only (for example, volunteer schools) and not more broadly.

The resulting WLS estimator of the average treatment effect (ATE) is consistent and asymptotically normal as M approaches infinity. Further, the ATE variance estimator is based on regression residuals *averaged to the cluster level*, with separate additive terms for the treatment and control groups. Intuitively, the model is estimated using the individual data, but standard errors are calculated using residual sums of squares based on cluster-level residuals.

The above results hold for any choice of baseline covariates as long as the number of covariates (v) yield sufficient degrees of freedom for hypothesis testing $(M - v - 2)$. In practice, however, rigorous data-driven selection procedures are needed to select amongst the v^* candidate covariates (especially if v^* is large), while at the same time avoiding "favorability" bias and overfitting. Further, adopting methods that fix the selection process across replications can facilitate standard error estimation. Next, we discuss our approach to satisfy these goals.

Lasso for clustered designs. The paper uses Lasso (Tibshirani, 1996; Efron et al., 2004) for covariate selection because it aligns with the design-based framework. Lasso is a commonly-used penalized regression approach that selects covariates by shrinking some regression coefficients to zero. Lasso, which was developed for non-clustered designs, estimates coefficients by minimizing a least squares objective subject to the constraint that the sum of the absolute values of the model coefficients is bounded above by some positive number.

In our context, Lasso can be adapted to the design-based framework for clustered designs by minimizing the following penalized loss function using data *averaged to the cluster level*:

$$\hat{\boldsymbol{\gamma}}_{\mathbf{Lasso}} = \arg\min_{\gamma} \left\{ \frac{1}{2} \sum_{j=1}^{m} w_j^* (\bar{y}_{jW}^* - \bar{\mathbf{x}}_{\mathbf{jW}}^{*} \boldsymbol{\gamma})^2 + \lambda \sum_{k=1}^{v^*} |\, \gamma_k \,| \right\}, \tag{3}$$

where $\bar{y}^*_{jW}$ and $\bar{\mathbf{x}}^*_{\mathbf{jW}}$ are cluster-level outcomes and covariates, standardized to have mean 0 and variance 1; w^*_j are weights scaled to sum to 1; $\boldsymbol{\gamma}$ is the parameter vector; and λ controls the amount of regularization (shrinking). The weights can be set to 1 (to weight clusters equally), to cluster sample sizes (to weight individuals equally), or other values. As λ increases, more shrinking occurs and more parameter estimates are set to 0 (that is, omitted from the model).

A critical feature of our approach is that we *exclude* the treatment indicator, T_j, from the loss function in (3). Thus, the approach identifies predictive covariates that average across the two research groups. In the FP model, the only source of randomness is due to T_j (from reallocations of the sample to the treatment and control groups). Thus, for a given λ, our approach *fixes* the covariate selection process across replications, and decouples estimation of the treatment effect from covariate selection (building off a similar idea proposed by Tsiatis et al., 2000 in a different context who recommend separate models be estimated for the two research groups).

Relatedly, the objective function in (3) only selects covariates based on the strength of relationships with the outcome and not with T_j. This is because in RCTs, selecting covariates that are correlated with T_j but not with the outcome variable yields upwardly biased standard error estimates (Raab et al., 2000; Ertefaie et al., 2015; Koch et al., 2018).

Cross-validation (CV) can be used to select λ, for example, by partitioning the data into 5 random groups for training and validation (5-fold CV). A complication with this approach is that different random partitions of the sample could yield different covariate selections, which could lead to randomness in the Lasso procedure that we are trying to avoid. Thus, the paper presents a method to address this issue where Lasso is run a pre-specified number of times (detailed in the paper) and the most common selected covariate set across replications is identified to ensure the true model is identified with a high probability. Another approach is to use jack-knife (leave-one-out) CV methods that will not vary across replications, which could also be suitable for designs with a small number of clusters where 5-fold CV is not practical.

The paper also discusses a variant of Lasso—adaptive Lasso (Zou, 2006)—which penalizes coefficients differently based on first-stage coefficient estimates (for example, from ridge regression). This approach has been shown to have better asymptotic properties in recovering the true model for non-clustered designs. We conduct simulations using both approaches.

Finally, note that the Lasso model in (3) could be estimated using individual-level data instead of cluster-level data. However, while this approach could increase precision slightly, it runs the risk of identifying covariates that primarily explain outcome variation across individuals *within* clusters (as opposed to between clusters), which is a problem because within-cluster covariates do not enter the variance formulas for the design-based impact estimators, and thus, do not improve precision. Thus, we apply (3) using cluster-level data to avoid this possibility and to ensure we identify covariates that explain the variation in mean outcomes between clusters.

Impact estimation. After the Lasso procedure has been conducted, the selected covariates can be used to estimate impacts and standard errors using the design-based estimators discussed earlier. Because covariate selection under our FP framework is fixed (along with the outcomes and covariates), the approach avoids the need to adjust for randomness in the selection process. Fixing the selection process across replications for the SP model would be more difficult, because outcomes and covariates would also vary across replications.

Our approach does not use Lasso to simultaneously estimate treatment effects and covariate parameters. Rather, our sequential approach uses the fact that in the design-based framework, covariates do not enter the “true” data generating process underlying experiments (shown in (1)), but are ancillary. Stated differently, the design-based approach does not depend on the “true” relationship linking outcomes and covariates (unlike model-based estimators). Thus, our approach selects predictive covariates in the first stage that are independent of treatment status, and then uses these covariates to obtain consistent design-based estimators to improve precision relative to the pre-specification approach (but the estimators may not be fully efficient).

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

We believe that the methods discussed in the paper can improve the process of selecting covariates to estimate treatment effects for commonly-used clustered RCTs in the education field. The methods are easy to apply using existing software to estimate Lasso models (for example, using R, Stata, or SAS) and to then estimate impacts using design-based estimators (for example, using RCT-YES). Our hope is that education researchers will consider using these methods in the future, and conduct research to improve them.

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

Early simulation results suggest that the Lasso procedure using (3) is able to recover the true model covariates reasonably well based on standards found in the literature. The simulations suggest also that our sequential method for estimating impacts and their standard errors using the design-based estimators with Lasso-generated covariates yields Type 1 errors near nominal levels. Finally, early empirical analysis results using the SACD data show that our method yields considerable precision gains relative to an approach based on the pre-specification of a small number of pretest covariates only. Thus, initial results suggest that the approach shows promise.