Using Lasso Regression to Predict Impacts Locally: Reanalysis of Data on Teach for America

Elizabeth Petraglia, Rob Olsen, and Elizabeth Stuart

Background/Context: National studies are often used to inform local policy decisions. Evidence-based policy assumes that multi-site impact evaluations can be used to predict the impact of an intervention in sites that did not participate in the evaluation. But impacts of educational interventions may vary across sites (e.g., Weiss et al., 2017), and modeling this variation can be challenging—suggesting that it may be difficult to use statistical models to make accurate out-of-sample predictions of impact. Findings published in the Journal of Policy Analysis and Management (Orr et al., 2018) cast doubt on how well such models applied to national impact studies can inform local policy decisions.

Purpose/Objective/Research Question: The goal of this research is to test whether building a predictive model using Lasso regression substantially improves researchers' ability to predict the impacts in a single site using data from a multi-site impact evaluation, as compared to using standard model-building techniques such as stepwise regression.

Setting: An evaluation for the Institute of Education Sciences conducted between 2008 and 2013 that used an experimental design to estimate the impact of Teach for America on math achievement in middle and high school (Clark et al., 2013).

Population/Participants/Subjects: The evaluation included 4,573 students in middle and high school, 136 math teachers, and 45 schools spanning 11 school districts. Participating schools were similar to other schools with TFA teachers in their demographic characteristics but were larger, more heavily African American, and more likely to be eligible for Title I than secondary schools nationwide. Data from this evaluation are available from the National Center for Education Statistics' Restricted Access Data Program.

Intervention/Program/Practice: TFA selects high-achieving college graduates—typically from selective colleges—and places them in classrooms without requiring them to complete all of the usual requirements. As a result, TFA teachers tend to be younger and less qualified on traditional metrics, but substantially better at mathematics based on a test of math content knowledge, scoring almost a full standard deviation higher than their colleagues teaching the same math classes in the same school (Clark et al, 2013).

Research Design: The design of the TFA evaluation was based on random assignment: Participating students were randomly assigned to a math classroom taught by a TFA teacher or to a math classroom taught by a non-TFA teacher. The impact of attending TFA was estimated by comparing the average math achievement of students assigned to TFA teachers to the average math achievement of students assigned to non-TFA teachers. The research design for our reanalysis of TFA data exploits the experimental design of the original study and the rich data collected.

Data Collection and Analysis: Our reanalysis of TFA data will estimate the impact of TFA separately for each of the 45 schools participating in the study, using the same analysis methods as the authors of the original study. In addition, we will identify school-level variables from the TFA data that may predict the magnitude of the impact in a particular school; identify teacher- and student-level variables that may affect the impact of being assigned to a TFA teacher; and create teacher- and school-level aggregates of these variables to include as possible predictors of the impact of TFA in individual schools.

This process will result in a very large number of possible predictor variables—far more than could be included in a statistical model for predicting the impact of TFA in individual schools. The predictive performance of the final model is likely to be sensitive to the subset of these variables included in the prediction model: Failing to include important variables will result in biased predictions, while including variables that are not strong predictors of the impact will increase variance without improving predictive performance.

The analysis will compare two different approaches to deciding which school-level variables to include in statistical models for making out-of-sample predictions of the impacts of educational interventions. In particular, we will compare two different approaches to building a model to make out-of-sample predictions: (1) Lasso regression, which uses regularization to shrink weak predictors towards zero and is typically fit by cross-validation, both of which tend to improve the model's accuracy in making out-of-sample predictions; and (2) stepwise regression, which is more broadly known but relies on in-sample metrics such as p-values or AIC, and may not result in a model that performs well out-of-sample.

In reporting the findings, we will use root mean square prediction error (RMSPE) as the primary metric. We will estimate RMSPE for each site as in Orr et al. (2018), adjusting for sampling error, and then use a nonparametric test such as the Mann-Whitney U test to determine whether the median RMSPE is significantly different between the two methods. We will also explore whether prediction errors were made in the "correct" direction. For example, if the true impact is positive and moderately large and the predicted value is positive and large, this prediction error is not egregious because a local policymaker would likely still decide (correctly) to implement the intervention. However, if the true impact is small and positive but the predicted value is small and negative, this is potentially more problematic even if the value of the RMSPE is the same in both situations. Visualizations of selected results will be produced for this poster.