Generation of Synthetic Schools to Improve Model Selection for Reading Interventions

Tim Lycurgus
Department of Statistics, University of Michigan

Motivation

- This project is motivated by a large scale IES funded randomized trial testing the efficacy of one of the featured products of a well-known vendor of educational services designed to improve reading abilities for students between kindergarten and third grade.
- Model selection to determine which method should be used in analysis of the overall treatment effect is crucial for this project along with any given experiment or study. Often, models are chosen such that they provide the largest and most significant treatment effect.
- Through this method of synthetic schools, we are able to calculate the power and size of various models by only looking at control data.
- By creation of synthetic schools, it is possible to select the ultimate model independent of whether it will provide the largest effect. In other words, model selection will not be influenced by the outcomes of those treated.

Method for Analysis

- We create 26 pairs of pseudo-control schools and pseudo-treatment schools through the synthetic school generation process described below
- The students are divided in a manner such that the students placed in the pseudo-control schools look similar to actual control students and the students placed in the pseudo-treatment schools look similar to students who actually received the treatment
- Since all students in pseudo-schools were originally in the overall control group, nobody received the treatment, meaning the average treatment effect should be zero. An effective model should find no effect
- We then impose an artificial treatment effect on the pseudo-treatment schools. Similar to the above scenario, an effective model should be able to parse out that imposed effect.
- Various models are tested both before and after the imposed effect to see their accuracy. Through bootstrapping, we are able to test these models multiple times

Generation of Synthetic Schools

- Within each pair of control/treatment schools, calculate propensity scores of a student attending a treatment school based on demographic covariates like age, race, gender, socio-economic status, etc.
- Within each pair, sample without replacement from the control school such that the odds of being selected into the pseudo-treatment group are proportional to the odds of the estimated propensity scores from the previous step.
- Sampling is performed in a manner such that the proportion of students within the pseudo-treatment school is roughly equivalent to the proportion of students within the actual treatment school for each pair through independent Bernoulli sampling
- The remaining students not selected into the pseudo-treatment school are placed in the pseudo-control school.

Models Tested

- Model 1: Regression of `Treatment` on `Outcome`
- Model 2: Regression of `Treatment` on `Outcome` controlling for `Pre-Test Scores`
- Model 3: Regression of `Treatment` on `Outcome` controlling for `Pre-Test Scores` and demographic covariates such as `Age, Gender, Race, Free-Lunch Status` , etc.
- Model 4: The model described in Model 3 but using a Peters-Belson technique
- Note: Models 1-3 use clustered sandwich estimators with a bias adjustment where the clusters are determined based on a student's baseline school
- Robust versions of the above models have been tested as well, but differences were negligible so they are not presented

Results of the Models

	No Effect			Imposed Effect		
	Bias	**RMSE**	**Size**	**Bias**	**RMSE**	**Power**
Model 1	-0.01	1.48	5%	0	1.64	50%
Model 2	-0.01	1.29	5%	-0.02	1.48	48%
Model 3	0	1.09	5%	0	1.31	94%
Model 4	0	1.16	4%	0	1.53	100%

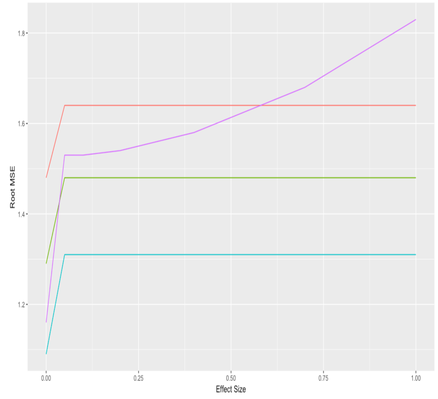
Table 1: Model Performance with $n = 1000$

- For the above results, the artificial effect was randomly imposed from a $N(0.10, 1.0)$ distribution
- Bias in the above table is calculated as the difference between the true effect of the treatment (either 0 or 0.10) and the effect of the treatment estimated by each model
- Note: Data is still in the processing stages and is periodically updated, explaining discrepancies in results from different versions

Discussion of Results

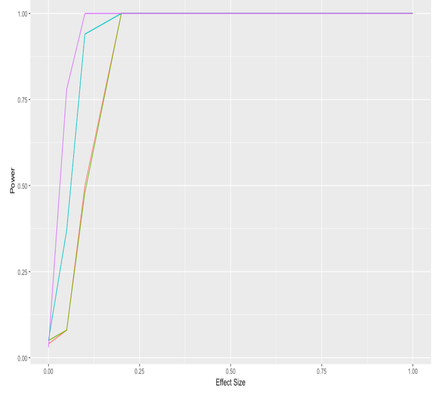
- All four models perform relatively similarly with regards to bias. In other words, each model accurately determines the treatment effect, although the simpler models do miss the target slightly
- Model 2 may be hindered with regards to bias by the fact that it is not a true pre-test in the sense that it isn't the same test. Instead, we can view it as more of a baseline indication of skill
- The real divergence in models occurs with regards to the RMSE of the various models. Clearly more complex models controlling for demographic variables perform better than their simpler counterparts.
- While the more complex models perform better with regards to RMSE, this only holds true for Peters-Belson if the effect of the intervention is small. Unlike the other three models, Peters-Belson sees its RMSE increase with an increase in the size of the effect

Effect Size vs. Root MSE



- Red: Method 1: Hypothesis Testing
- Green: Method 2: Pre-Test Model
- Cyan: Method 3: Full Demographic Model
- Purple: Method 4: Peters-Belson Approach

Effect Size vs. Power



- Red: Method 1: Hypothesis Testing
- Green: Method 2: Pre-Test Model
- Cyan: Method 3: Full Demographic Model
- Purple: Method 4: Peters-Belson Approach

Conclusions and Future Directions

- From these simulations, it is possible to select a potential model to be used in the ultimate analysis
- We were able to achieve this without looking at any actual results, meaning our choice of model was determined independent of outcome
- Ultimately, all four models capture the effect of the imposed intervention. That being said, from the above graph, Models 3 and 4 clearly achieve higher rates of power for smaller effects which is important since the effect of this reading intervention is expected to be small
- While Models 3 and 4 have similar RMSEs when there is no effect, that quickly diverges and Model 3 performs by far the best. Therefore, this is the preferred model as of this moment
- Robust versions of these models have been tested and have shown negligible differences. Going forward, hierarchical linear models will be tested as well, with random slopes at the school level and random intercepts at the student level.

For references, questions and comments please contact tlycurgu@umich.edu