Generation of Synthetic Schools to Improve Model Selection for Reading Interventions

Tim Lycurgus

January 15, 2018

I Introduction

We use a novel data-generation scheme to simulate paired pseudo-control and pseudo-treatment schools (Wyss et al., 2017) in service of improving the analysis of expensively collected data about a widely used education intervention. The creation of these pseudo schools in turn allows us to test various specifications of student-level regression models in order to estimate treatment effects. This also produces information about the MSE of different treatment effect estimators along with other information relevant to these models such as the size and power of the tests. Furthermore, we utilize modern cluster-robust standard error calculations as described in Pustejovsky and Tipton (2016) which in turn built off the bias-reduced linearization variance estimator proposed by Bell and McCaffrey (2002), in order to achieve correct Type I error rates.

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.

II Objectives

The primary objective of this simulation is to determine which method of analysis will provide the most power for examination of the effect of the reading intervention. In addition, we will examine whether accompanying our regression analysis with a modern cluster-robust standard error calculation will lead to t-statistics for regression coefficients with correct Type I error rates.

III Method

As outlined in Wyss et al. (2017) and justified in Hansen (2006), we created 26 pairs of pseudo-control and pseudo-treatment schools, to match the 26 pairs of actual schools in our study. The generation of these schools was performed as follows:

- Within each pair of control/treatment schools, calculate propensity scores of a student attending a treatment school based on demographic covariates (Rosenbaum and Rubin, 1983).
- 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. To perform this, independent Bernoulli sampling was used.
- The remaining students not selected into the pseudo-treatment school are placed in the pseudo-control school.

By performing these steps, we have essentially divided each control school into a pseudo-treatment and a pseudo-control school where the students placed in the pseudo-treatment school look roughly similar to the students in the actual treatment school. This division into pseudo schools provides multiple benefits. Primarily, since all of these students were actually in the control group, none received the treatment. That means without adjustment, an accurate model should find a treatment effect of roughly 0. Similarly, we can artificially impose a treatment effect on the pseudo-treatment students and see how accurately various models diagnose that effect. By bootstrapping the students who are placed into the pseudo-treatment and pseudo-control schools, we are then able to calculate the power and size of the various models. Another benefit of running our analysis on these synthetic schools rather than on the actual data is the model we select will not be influenced by the outcomes of those actually treated. In other words, the ultimate model will be selected independently of whether it will provide the largest average treatment effect.

Since the primary objective of this simulation study is to select the model with the most power, we then test various models on the synthetic schools. The first model tested is a basic hypothesis test on the effect of treatment on the outcome of interest (post-treatment test scores). The next models use linear models paired with Huber-White sandwich estimators (Huber, 1967; White, 1980). In particular, we use clustered sandwich estimators with a bias adjustment as described in Bell and McCaffrey (2002) where the clusters are determined based on a student's baseline school. The first of these examines the effect of the treatment, controlling for pre-test scores. The idea behind only using the pre-test is that in theory, all demographic differences are soaked up into the pre-test and we would not need additional covariates. The next model includes those demographic covariates (Age/Race/Gender/SES/etc.). We then extend this demographic model by applying the modern cluster-robust standard error correction as described in Pustejovsky and Tipton (2016) to see how this would affect the Type I error rates. Finally, we test this same demographic model but using a Peters-Belson technique (Peters, 1941; Belson, 1956).

IV Results

Preliminary results for the five models can be found in the appendix. The table below shows the accuracy of the models when there is no treatment effect along with the accuracy of the models when a treatment effect has been artificially imposed upon the students in the pseudo-treatment schools. In both situations, 1000 bootstrap simulations were run. The artificial treatment effect was randomly generated from a normal distribution with mean 1 and standard deviation 0.20.

As can be seen, all five 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. However, a clear difference between the first two models and the remaining three arises when looking at the RMSE. Clearly the simple hypothesis test and just controlling for pre-test scores miss additional information that can be used to make accurate predictions. Therefore, when selecting a model for analysis of the reading intervention, demographic variables should be included.

V Tables and Figures

	No Effect		Imposed Effect	
	Bias	**RMSE**	**Bias**	**RMSE**
Model 1	-0.1	1.49	-0.1	1.50
Model 2	-0.1	1.26	-0.1	1.27
Model 3	0	1.14	0	1.15
Model 4	0	1.14	0	1.15
Model 5	0	1.13	0	1.18

Table 1: Model Performance with $n = 1000$

References

Bell, R. M. and McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–182.

Belson, W. A. (1956). A technique for studying the effects of a television broadcast. *Applied Statistics*, pages 195–202.

Hansen, B. B. (2006). Bias reduction in observational studies via prognosis scores. Technical report, Technical Report 441, University of Michigan, Statistics Department.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA.

Peters, C. C. (1941). A method of matching groups for experiment with no loss of population. *The Journal of Educational Research*, 34(8):606–612.

Pustejovsky, J. E. and Tipton, E. (2016). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

Wyss, R., Hansen, B. B., Ellis, A. R., Gagne, J. J., Desai, R. J., Glynn, R. J., and Stürmer, T. (2017). The dry-run analysis: A method for evaluating risk scores for confounding control. *American journal of epidemiology*, 185(9):842–852.