Using Machine Learning and Auxilliary Data for Precise, Unibiased, Causal Inference

Adam C Sales Johann Gagnon-Bartsch Ed Wu
Neil T Heffernan Anthony Botelho Luke Maritrix
Thanporn March Patikorn

1 Background / Context

The ASSISTments online homework tool includes a platform, called the TestBed, on which education researchers can propose experimental modifications of specific modules within ASSISTments. Then, students whose teachers assign those modules are individually randomized between treatment conditions. In one RCT, 614 students working on a Pythagorean Theorem module were randomized to receive hints either as text or as videos. Researchers may estimate the effect of hint type on module completion rates by comparing outcomes between the randomized treatment groups, and may adjust those comparisons with provided covariates, such as prior problem correctness and completion rates.

However, much more "auxiliary" data is available for estimating causal effects. Hundreds of thousands of students have used ASSISTments—could we use data from the "remnant" from the experiment, i.e. students who were not randomized to either condition, to increase the precision of effect estimates? ASSISTments gathers anonymous computer log data for each of its students, comprising a rich, longitudinal dataset—how could we use this high-dimensional covariate data?

Including data from the remnant in the analysis by, say, pooling remnant students with the control group would undermine the entire randomization design. There is no way to guarantee that members of the remnant are statistically comparable to either randomized treatment group. Incorporating administrative covariates poses a number of serious challenges, mostly stemming from the fact that the number of available covariates is large compared to the number of students in the study. Further, it is hard to know, *a priori*, which variables are actually relevant, or how best to model their longitudinal structure.

That said, auxiliary data could be of immense value to experimental analyses, especially when analyzed with modern machine learning models. Administrative covariates could explain a large portion of the variation of experimental outcomes, thereby increasing the precision of educational RCT analyses. The remnant offers a large sample in which to model outcomes as a function of covariates.

2 Purpose / Objective / Research Question

This talk will introduce a method to integrate auxiliary data—high-dimensional administrative covariates and data from study non-participants—into designed-based experimental estimates. The method combines ideas from Sales, Botelho, Patikorn, and Heffernan (2018), and Wu and Gagnon-Bartsch (2017), which introduces LOOP, a method for non-parametric covariance adjustment in randomized experiments. We will describe and illustrate the method in an analysis of 22 ASSISTments TestBed experiments, and discuss avenues for future research.

3 Methodology

Our method proceeds in two overarching steps. First, use the remnant—study non-participants—to fit a model (say, $\hat{y}_C(\cdot)$) predicting outcomes Y as a function of administrative covariates $\boldsymbol{x}$. $\hat{y}_C(\cdot)$ may be a classical regression model or a modern machine learning model, and may involve human input. Most importantly, $\hat{y}_C(\cdot)$ needn't be correct or unbiased in any sense. Analysts may use remnant data to choose, modify, or tune $\hat{y}_C(\cdot)$ in an iterative process, based, perhaps, on its performance in cross-validation or other model assessments. After these steps, use $\hat{y}_C(\cdot)$ to predict outcomes for experiment participants as $\hat{y}_C(\boldsymbol{x})$.

Since $\hat{y}_C(\cdot)$ was fit to subjects in the remnant, and since it uses only covariate data from experimental participants, predictions $\hat{y}_C(\boldsymbol{x})$ are invariant to randomization; they are covariates.

In our analysis of the 22 ASSISTments experiments, we developed and applied a type of deep learning model known as a long short term memory network (Hochreiter and Schmidhuber, 1997), a variant of a recurrent neural network (Williams and Zipser, 1989) that is commonly applied to time series or panel data. We fit the model using log data from 686,590 assignments completed by 134,141 unique students, none of whom participated in any of the 22 A/B tests, and predicted module completion for all RCT participants.

The second step is to use $\hat{y}_C(\boldsymbol{x})$, in the LOOP algorithm, to estimate treatment effects. The LOOP estimator starts by predicting both treatment and control potential outcomes for each experimental subject. For subject i, say, the algorithm models Y as a function of covariates and treatment assignment, and fits the model using all experimental subjects *except for* i, and uses the fitted model to predict both of i's potential outcomes. In our example, we use either ordinary least squares (OLS) or random forests to model Y as a function of treatment assignment, $\hat{y}_C(\boldsymbol{x})$, and, perhaps, other covariates as well. Causal estimates using loop are design-based and exactly unbiased.

4 Findings / Results

Figure 1 displays estimated treatment effects and 95% confidence intervals for each of the 22 experiments run within ASSISTments. Three estimates are shown for each experiments:

estimates without covariance adjustment, with LOOP adjustment for standard covariates, and with LOOP adjustment for covariates and $\hat{y}_C(\boldsymbol{x})$. Notably, while standard covariates have little effect on the width of the confidence intervals, incorporating $\hat{y}_C(\boldsymbol{x})$ reduces many of the intervals considerably.

Figure 2 shows the amount of improvement in each experiment due to incorporating $\hat{y}_C(\boldsymbol{x})$. Incorporating $\hat{y}_C(\boldsymbol{x})$ decreased standard errors in all but three experiments, where its effect was trivial. In most experiments it reduced standard errors by over 30%, and in six, incorporating $\hat{y}_C(\boldsymbol{x})$ cut standard errors in half.

5 Conclusions

Randomized trials in education are always resource-constrained and often underpowered. On the other hand, a vast reservoir of rich auxiliary data are often available—administrative covariates and data from students who did not participate in the experiment.

The methods we present here, and their results, show how this untapped resource of auxiliary data could be used to improve the precision of causal estimates from educational experiments, sometimes substantially. They do so without invoking any additional assumptions beyond the classical experimental analysis framework.

Our illustration, 22 randomized tests run within a piece of educational technology, is not the typical education research field trial. Instead, it represents an attempt, by the ASSISTments team, to improve their product's effectiveness. Our analysis exemplifies how diverse demands for evidence, and their diverse solutions, often bring new challenges alongside new opportunities.

References

Anthony Botelho, Adam C Sales, Neil T Heffernan, and Thanaporn March Patikorn. The assistments testbed: Opportunities and challenges of online experimentation in intelligent tutors, 2018. URL `https://drive.google.com/file/d/15WCA14VqbFL`$_Z Irqakp0go3AIC1sGZTR.InSubmission$.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Adam C Sales, Anthony Botelho, Thanaporn M Patikorn, and Neil T Heffernan. Using big data to sharpen design-based inference in a/b tests. In *Proceedings of the 11th International Conference on Educational Data Mining. International Educational Data Mining Society*, pages 479–486, 2018.

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

Edward Wu and Johann Gagnon-Bartsch. The loop estimator: Adjusting for covariates in randomized experiments. *arXiv preprint arXiv:1708.01229*, 2017.

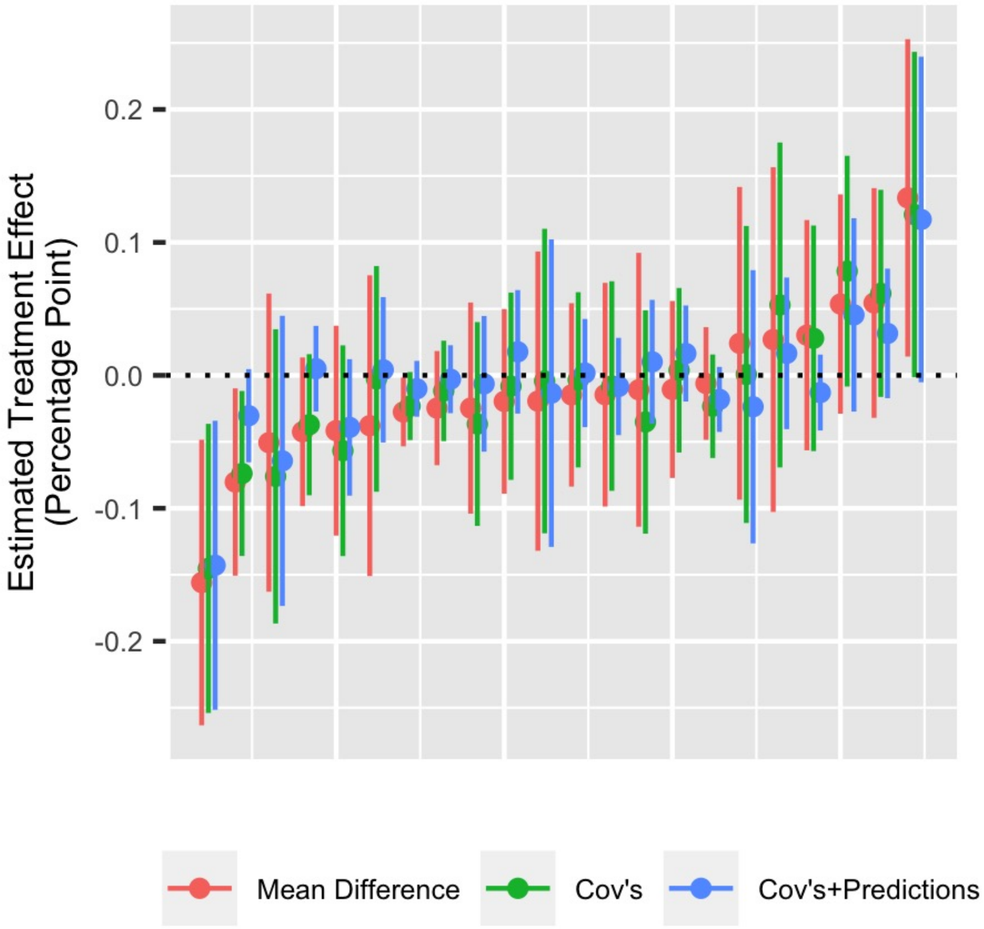

Figure 1: Estimated treatment effects and 95% confidence intervals for the difference between conditions in 22 experiments run within the ASSISTments platform. We calculated effects in three different ways: without covariance adjustment (Mean Difference), using LOOP to adjust only for standard aggregate covariates (Cov's), and using LOOP to adjust for covariates and $\hat{y}_C(\boldsymbol{x})$, (Cov's+Predictions)

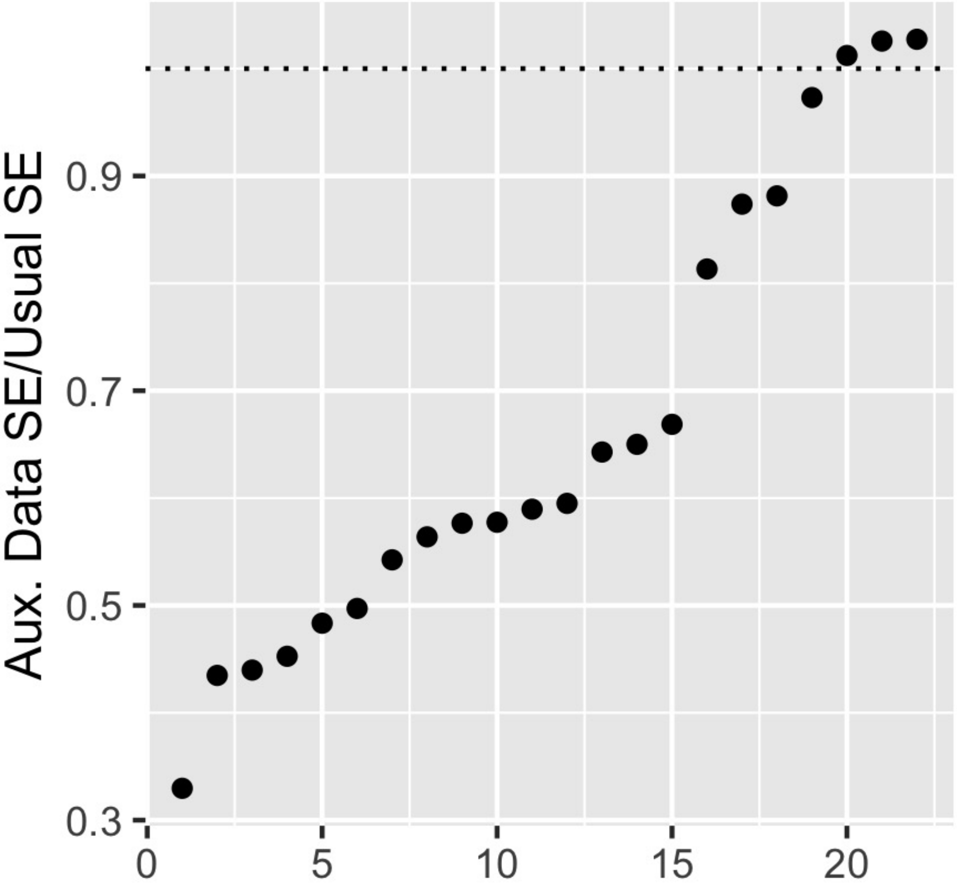

Figure 2: The ratios of the standard errors after using LOOP to adjust for covariates and $\hat{y}_C(\boldsymbol{x})$, to the standard errors without covariance adjustment, in 22 experiments run within ASSISTments