More Powerful A/B Testing using Auxiliary Data and Deep Learning ⋆

Adam C. Sales[1], Ethan Prihar[1], Johann Gagnon-Bartsch[2], Ashish Gurung[1], and Neil T. Heffernan[1]

[1] Worcester Polytechnic Institute, Worcester, MA 01609, USA
[2] University of Michigan, Ann Arbor, MA 48109, USA

Presented at: Artificial Intelligence in Education
Durham, UK, July 27-31, 2022

Abstract. Randomized A/B tests allow causal estimation without confounding but are often under-powered. This paper uses a new dataset, including over 250 randomized comparisons conducted in an online learning platform, to illustrate a method combining data from A/B tests with log data from users who were not in the experiment. Inference remains exact and unbiased without additional assumptions, regardless of the deep-learning model's quality. In this dataset, incorporating auxiliary data improves precision consistently and, in some cases, substantially.

Introduction In randomized A/B tests on an online learning platform, students are randomized between different educational conditions and their subsequent outcomes are compared. Estimates from A/B tests are unbiased, but may be imprecise due to small sample sizes. An observational study can often boast a larger sample size but is subject to confounding so conventional analysis of A/B tests discards data from the "remnant" of the experiment—students who were not randomized, but for whom covariate and outcome data are available.

However, data from the remnant an can play a valuable role in causal estimation. [2] suggests first using the remnant data to train a model using covariates to predict outcomes; then, using that fitted model to predict (or impute) outcomes for participants in the experiment. Finally, use those imputations as a covariate in a causal effect estimator. This method builds on recent work in design-based covariate adjustment, e.g. [5], and in particular, using the remnant to improve precision [e.g.] [1]. Unfortunately, [2] provides only limited evidence of the method's success in practice.

This paper reviews two of the causal estimators of [2], and applies them to an new dataset: a collection of 84 multi-armed A/B tests run on the ASSISTments TestBed [3], which together include 377 different two-way comparisons, and 41,226 students. Alongside this experimental data, we collected log data for an additional 193,218 students who worked on similar skill builders in ASSISTments but did not participate in any of the 84 experiments—the remnant. We used these datasets to estimate the causal effects of each of the conditions on

⋆ The final authenticated publication is available online at https://doi.org/10.1007/978-3-031-11647-6 Published July 25, 2022.

assignment completion. Our interest here is not on the treatment effects themselves, but on the extent to which these methods reduce standard errors. Our results give a much clearer picture of the potential impacts of using remnant data in design-based causal inference: incorporating remnant data consistently improves statistical precision, sometimes substantially.

Method For each subject i in a randomized experiment, let $Z_i = 1$ if i is randomized to the treatment condition and $Z_i = 0$ if i is randomized to control, and let Y_i be the outcome of interest. Following [4], define y_i^c and y_i^t as the outcomes i would have exhibited had i been assigned to control or treatment, respectively. Then, assuming no spillover effects, $Y_i = Z_i y_i^t + (1 - Z_i) y_i^c$, and the treatment effect for student i is $\tau_i \equiv y_i^t - y_i^c$.

Let $\boldsymbol{x}_i$ be a $k \times 1$ vector of baseline covariates for subject i, and let $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$ be functions from $\mathbb{R}^k \rightarrow \mathbb{R}^1$ that impute y_i^c and y_i^t, respectively, as a function of $\boldsymbol{x}_i$. Finally, if $Pr(Z = 1) = 1/2$, let $m_i = 1/2(y_i^c + y_i^t)$, subject i's expected counterfactual potential outcome, and let $\hat{m}_i = 1/2(\hat{y}^c(\boldsymbol{x}_i) + \hat{y}^t(\boldsymbol{x}_i))$ be it's estimate. Then, if $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$ are constructed such that $\{\hat{y}^c(\boldsymbol{x}_i), \hat{y}^t(\boldsymbol{x}_i)\} \perp\!\!\!\perp Z_i$, then

$$\hat{\tau} = \frac{1}{n} \sum_{i \in \mathcal{T}} \frac{Y_i - \hat{m}_i}{p} - \frac{1}{n} \sum_{i \in \mathcal{C}} \frac{Y_i - \hat{m}_i}{1 - p} \tag{1}$$

is an unbiased estimate for $\bar{\tau}$. In fact, this unbiasedness holds regardless of $\hat{y}^c(\cdot)$ or $\hat{y}^t(\cdot)$'s other properties—they need not be unbiased, or consistent, or correct in any sense for $\hat{\tau}$ to be unbiased.

[2] combines two approaches to ensuring that $\{\hat{y}^c(\boldsymbol{x}_i), \hat{y}^t(\boldsymbol{x}_i)\} \perp\!\!\!\perp Z_i$: the first uses a leave-one-out algorithm using observations other than i to train models $\hat{y}^c_{-i}(\cdot)$ and $\hat{y}^t_{-i}(\cdot)$ that will in-turn give rise to imputations $\hat{y}^c(\boldsymbol{x}_i)$ and $\hat{y}^t(\boldsymbol{x}_i)$ and finally m_i. As long as $Z_i \perp\!\!\!\perp Z_j$ for $i \neq j$, then $\{\hat{y}^c(\boldsymbol{x}_i), \hat{y}^t(\boldsymbol{x}_i)\} \perp\!\!\!\perp Z_i$ will hold.

The second approach uses the remnant to train a different model, $\hat{y}^r(\cdot)$, producing imputations $x^r \equiv \hat{y}^r(\boldsymbol{x}_i)$. Importantly, x^r is a baseline covariate, unaffected by treatment assignment, since it is a function of baseline covariates $\boldsymbol{x}$ and a model fit to a separate sample. Therefore, it can be incorporated into an estimator such as (1), perhaps alongside other covariates. If $\hat{y}^r(\cdot)$ performs well in the experimental sample, so that $|x^r - y_i^c|$ tends to be small, then doing so can drastically improve precision; in the limit, if $x^r = y_i^c$ for all i, then the standard error of $\hat{\tau}$ would be due only to treatment effect heterogeneity, and the average effect on treated subjects would be known exactly. On the other hand, if $\hat{y}^r(\cdot)$ does not perform well it will not threaten the validity of the inference, and in large samples it will not harm precision.

Here, we include two specific versions of $\hat{\tau}$: first, $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ uses x^r as the only covariate and uses ordinary least squares linear regression (OLS) for leave-one-out imputation models $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$. We expect that when $\hat{y}^r(\cdot)$ performs well, OLS will be optimal since the relationship between Y and x^r will be approximately linear. Second, $\hat{\tau}^{\mathrm{SS}}[\tilde{\boldsymbol{x}}, \mathrm{EN}]$ uses x^r alongside a vector of other covariates $\boldsymbol{x}$; leave-one-out imputation models $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$ are ensembles of OLS regression of Y on x^r and a random forest imputing Y from both $\boldsymbol{x}$ and x^r.

Application We gathered a set of 84 A/B tests run on the TestBed with assignment completion as a binary outcome. We also gathered standard student-level aggregated predictors. Several experiments included multiple conditions; in those cases we estimated each pairwise contrast separately, as long as the p-value testing $Pr(Z = 1/2)$ was greater than 0.1.

We used remnant data to train a deep learning model $\hat{y}^r(\cdot)$ imputing completion from covariates. Three different sets of data were collected for each sample in the datasets: prior student statistics, prior assignment statistics, and prior daily actions. The full dataset used in this work can be found at https://osf.io/k8ph9/?view_only=ca7495965ba047e5a9a478aaf4f3779e. Each of the three types of data in the remnant dataset were used to predict both skill builder completion and number of problems completed for mastery. a fourth neural network was trained using a combination of the previous three models. The details and code can be found at https://github.com/adamSales/reloop377abTests. We used this fourth model, $\hat{y}^r(\cdot)$, to construct imputations x^r for each subject i in each experiment.

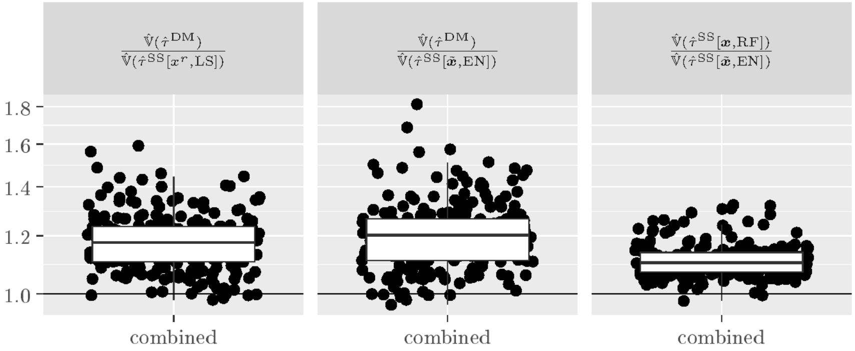

Fig. 1. Boxplots and jittered scatter plots of the ratios of estimated sampling variances of $\hat{\tau}^{\text{DM}}$, $\hat{\tau}^{\text{SS}}[x^r; OLS]$, $\hat{\tau}^{\text{SS}}[\boldsymbol{x}; RF]$, and $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}; EN]$

Figure 1 gives boxplots of ratios of estimated sampling variances $\hat{\mathbb{V}}(\cdot)$ for causal estimates: $\hat{\tau}^{\text{DM}}$, the Welch two-sample t-test, $\hat{\tau}^{\text{SS}}[\boldsymbol{x}, \text{RF}]$, the leave-one-out estimator using student-level covariates but no information from the remnant, and the two new estimators, $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ and $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$. The left and middle panels including remnant-based imputations is equivalent to increasing the sample size, relative to a t-test, by a factor of about 10-25% in about half of all cases, but up to 50%-70% in the most extreme cases.[3] The right panel

[3] Since sampling variance is typically $\propto 1/n$, ratios of sampling variances can be interpreted as ratios of effective sample sizes.

shows that compared to $\hat{\tau}^{\mathrm{SS}}[\boldsymbol{x}; RF]$, including remnant based imputations was equivalent to increasing the sample size by roughly 8-12% in half of all cases, but as much as 30% in others.

Discussion The approach illustrated here shows that data that do not meet an assumption—randomization—can still be used to help learn connections between covariates and outcomes. Its causal estimates will be unbiased, and inference correct, regardless of the data quality or model properties in the remnant. However, better data and better model fit will lead to better precision. The results in the ASSISTments A/B tests show that it sometimes improves precision greatly, and sometimes barely at all. Future research will explain this variance, as well as formulate suitable defaults and recommendations for when and how it should be used.

Acknowledgements This work was supported by IES grant #R305D210031.

References

1. Deng, A., Xu, Y., Kohavi, R., Walker, T.: Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 123–132 (2013)
2. Gagnon-Bartsch, J.A., Sales, A.C., Wu, E., Botelho, A.F., Erickson, J.A., Miratrix, L.W., Heffernan, N.T.: Precise unbiased estimation in randomized experiments using auxiliary observational data. arXiv preprint arXiv:2105.03529 (2021)
3. Ostrow, K.S., Selent, D., Wang, Y., Van Inwegen, E.G., Heffernan, N.T., Williams, J.J.: The assessment of learning infrastructure (ali): the theory, practice, and scalability of automated assessment. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. pp. 279–288. ACM (2016)
4. Rubin, D.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology; Journal of Educational Psychology **66**(5), 688 (1974)
5. Wu, E., Gagnon-Bartsch, J.A.: The loop estimator: Adjusting for covariates in randomized experiments. Evaluation review **42**(4), 458–488 (2018)