Leveraging Large Language Models To Improve Precision in Randomized Controlled Trials

Jaylin Lowe[1], Adam Sales[2], and Johann Gagnon-Bartsch[1]

[1]University of Michigan [2]Worcester Polytechnic Institute

Motivation

Randomized controlled trials (RCTs) are often considered the gold standard in causal inference, but they can be small and expensive, leading to estimates with high variance. Covariate adjustment, which models small chance in imbalances in characteristics across groups, can help. Precision can also be improved by integrating RCT data with larger external datasets. Large external datasets can be used to produce predicted outcomes, which can be included in the model as an additional univariate covariate.

What if we used a large language model (LLM) in place of an external dataset?

Setup

Under the potential outcomes framework [4, 6] we have:

- N observations, indexed from $i = 1, ...N$
- Z, the vector of treatment assignments. Each observation is randomly assigned to treatment ($Z_i = 1$) or control ($Z_i = 0$).
- p, the probability of being assigned to treatment. We assume a Bernoulli experiment.
- y_i^t and y_i^c, the potential outcomes under treatment and control, respectively. These represent the outcome value we would observe if observation i was assigned treatment or control, respectively.

We want to estimate the average treatment effect.

Using External Data For Covariate Adjustment

We use the method outlined in Gagnon-Bartsch et. al, 2023 [2]. This method is designed based and makes no assumptions about the quality of the predictions. The method adapted to use LLM predictions, is:

(1) Obtain a vector of predictions $\hat{y}_i^{LLM}$ from an LLM. These predictions with be used as a covariate in the future, so let $x_i^{LLM} \equiv \hat{y}_i^{LLM}$.

(2) Augment the vector of RCT covariates, $\mathbf{x_i}$ with this additional covariate. We call the resulting vector $\tilde{\mathbf{x}}_\mathbf{i}$, where $\tilde{\mathbf{x}}_\mathbf{i} = [x_i, x_i^{LLM}]$.

(3) Obtain imputations $\hat{y}^t(\tilde{\mathbf{x}}_\mathbf{i})$ and $\hat{y}^c(\tilde{\mathbf{x}}_\mathbf{i})$ for each observation i in the RCT. Crucially, the imputations for observation i must be independent of observation i's treatment assignment. These could be leave-one-out predictions froma linear mdoel, out-of-bag predictions from a random forest, or something similar.

(4) Calculate

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N} Z_i \cdot \frac{Y_i - \hat{m}_i}{p} - \frac{1}{N}\sum_{i=1}^{N}(1 - Z_i) \cdot \frac{Y_i - \hat{m}_i}{1-p} \quad (1)$$

where $\hat{m}_i = p\hat{y}^c(\tilde{\mathbf{x}}_\mathbf{i}) + (1-p)\hat{y}^t(\tilde{\mathbf{x}}_\mathbf{i})$. The estimated variance of $\hat{\tau}$ is:

$$\widehat{\mathrm{Var}}(\hat{\tau}) = \frac{1}{N}\left[\frac{p}{1-p}\hat{E}_c^2 + \frac{1-p}{p}\hat{E}_t^2 + 2\sqrt{\hat{E}_c^2\hat{E}_t^2}\right] \quad (2)$$

where $\hat{E}_c^2 = \frac{1}{n_c}\sum_{i=1}^{N}(1 - Z_i) * [y_i^c - \hat{y}_{-i}^c(\tilde{\mathbf{x}}_i)]^2$ estimates the mean squared prediction error in the control group and $\hat{E}_t^2$ is defined analogously. Since the LLM prediction augments the data already collected in the RCT, a $\hat{y}(\tilde{\mathbf{x}}_i)$ obtained using $\tilde{\mathbf{x}}_i$ may contain more information than a version based on $\mathbf{x}_i$ alone. If this prediction is highly correlated with the RCT outcome, then significant gains in precision can be made.

Case Studies

Case Study 1: Sentencing of Defendants and Recidivism [3]
Goal: Investigate effects of incarceration and probation on recidivism.
Treatment: Defendants are assigned randomly to judges.
Data: 1,003 defendants arrested in DC on drug related charges. Dataset included demographic information, prior history, type of drug, and type of charge.
Outcome: Recidivism, a binary variable capturing whether the defendant was arrested again within 4 years.

Case Study 2: Cognitive Tutor Algebra [5]
Goal: Evaluate the efficacy of a new algebra curriculum.
Treatment: Schools are randomized to continue with their typical curriculum (control) or switch (treatment)
Data: 19,053 middle and high school students. Dataset includes demographic information plus a score on an algebra readiness exam administered before the experiment.
Outcome: Score on algebra proficiency exam.

Case Study 3: Open Access Paper Citations [1]
Goal: Investigate imapact of open-access journal papers on paper citations
Treatment: Research articles are randomized to open-access upon publication (treatment) or available only to subscribers, subject to the journals typical policy (control)
Data: 1,248 papers from five journals with basic covariates such as number of pages, number of authors, review article or not, and self-archived or not. We also obtain abstracts from PubMed for every paper.
Outcome: Citation counts 3 years after publication.

Our Method

Goal: Obtain an additional covariate from the LLM that is predictive of the RCT outcome.

Step 1: Pair Observations

- Most straightforward way is to ask the LLM to predict to each observation's outcome. Doesn't work!
- Instead, pair observations and ask the LLM to compare each pair and predict which observation is more likely to exhibit a specific quality.
- May use only a subset of all of the possible pairs for gains in precision or computational time. Subsets can be chosen by stratifying on a particular covariate.

Step 2: Question Formulation

- All questions follow a similar format. We give all covariates for both observations in each pair and ask the LLM to predict which one will exhibit a specific quality.
- This quality may be the outcome variable or another related quality.
- We will use the selection frequency as an additional covariate.

Step 3: Modeling and Evaluation

- Check significance of LLM generated covariate in a regression model fit to the original RCT covariates plus the new covariate.
- If significant, proceed with calculating the variance of the ATE using the earlier equation.

Results

Case Study 1: Sentencing of Defendants and Recidivism

- Observations are stratified into 10 total groups of approximately equal size, based on out of bag predictions from a random forest on the RCT covariates. We consider only pairs where both observations come from the same stratum.
- Extracted LLM covariate is not statistically significant.
- LLM prioritizes prior history, but age is the most predictive covariate here and is still weak overall.

Case Study 2: Cognitive Tutor Algebra

- Observations are stratified into groups of size 10 based on out of bag predictions again.
- Extracted LLM covariate is statistically significant, which is particularly impressive because this RCT already has many predictive covariates.
- This does not translate into a practical difference in the estimator. The calculated standard errors are 0.009577 and 0.009571 for the estimators with and without LLM predictions, respectively.

Case Study 3: Open Access Paper Citations

- Observations are stratified by journal.
- In addition to asking the LLM to predict which paper will get more citations, we also ask it to predict which paper best exhibits each of 10 qualities.
- We compare four models: one with just the base covariates, one with base covariates plus the LLM citation based score, one with the base covariates plus the 10 LLM based quality scores, and one with all of these covariates.

Table 1. Estimator Standard Errors by Journal

Journal	Base	Citation	10 Qualities	All
Science	0.1227	0.1122	0.1073	0.0977
Neurophysiology	0.1300	0.1102	0.1147	0.1080
Genetics	0.1110	0.1060	0.1019	0.0999
Applied Physiology	0.1707	0.1814	0.1720	0.1680
FASEB	0.1066	0.0978	0.0971	0.0912

References

[1] P. M. Davis. Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *FASEB Journal*, 25(7):2129–2134, July 2011.

[2] J. A. Gagnon-Bartsch, A. C. Sales, E. Wu, A. F. Botelho, J. A. Erickson, L. W. Miratrix, and N. T. Heffernan. Precise unbiased estimation in randomized experiments using auxiliary observational data. *Journal of Causal Inference*, 11(1):20220011, Jan. 2023.

[3] D. P. Green and D. Winik. Using Random Judge Assignments to Estimate the Effects of Incarceration and Probation on Recidivism Among Drug Offenders. *Criminology*, 48(2):357–387, 2010.

[4] J. Neyman. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472, 1923.

[5] J. F. Pane, B. A. Griffin, D. F. McCaffrey, and R. Karam. Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis*, 36(2):127–144, 2014.

[6] D. B. Rubin. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Poster created using template by David Gamba, University of Michigan. Available at `https://www.overleaf.com/latex/templates/university-of-michigan-umich-poster-template/xpnqzzxwbjzc`.