Enhancement of ATE Estimation in A/B Testing Using MLRATE

■ What types of problems are you solving?

In online experiments and A/B testing, estimating the treatment effect between the treatment and control groups is a crucial part. Knowing how much the treatment causes the outcomes to change is important for businesses to come up with strategies. However, there is variability in estimating the treatment effect, and we need to construct confidence intervals to quantify this uncertainty. Therefore, potential reduction in this variability can make our estimation more meaningful.

In our previous assignment, we explored the estimation of the average treatment effect (ATE) and its confidence interval. With this project, we aim to reduce the variance in estimating ATE by using the machine learning regression-adjusted treatment effect estimator (MLRATE), proposed by Machine Learning for Variance Reduction in Online Experiments. Guo Y., et al., 2021. This allows us to increase the statistical power of our estimation and derive more precise predictive intervals with the given sample size.

■ What is the practical application or significance of it?

In a business sense, a single point estimation without large numbers of simulations can be misleading. The uncertainty and extreme cases can cause huge losses to the company if being overconfident with the point estimation. Constructing a confidence interval with a reasonable range can address the variability of the estimation. Reducing the variability results in a narrower confidence interval, thus providing a more reliable result. This can help companies to be more confident with their estimations.

■ What are the challenges?

As we are using simulated A/A tests to imply similar performance in A/B tests, the simulated data can never truly resemble real life data. The purpose of using simulated data is to test the first stage performance of the estimator when the i.i.d assumption holds true. The other challenge is that using more complex estimators can be prone to overfitting. The stratified K-fold splitting method helps address this challenge.

Executive Summary

Compared to the unadjusted difference-in-means estimator, MLRATE was able to reduce the variance in estimating the treatment effect (from 3.785 to 2.453). The average confidence interval widths after 200 bootstraps decreased by 35.2%, with a higher coverage rate. MLRATE performed well in reducing the variability of ATE estimation, while being able to cover more true ATE cases.

Methodology

The design of the MLRATE method is to first randomly split the data into K-folds. Here we used stratified K-fold splitting with K=2, as the paper suggested (train, test set). Then we use the train set to fit on a machine learning model (can be any choice) and predict the Y values in the test set. One improvement here is that since there's no accuracy testing involved, we could test the model onto the test set or an additionally splitted validation set. This way, we can tune the hyperparameters in the ML model for potential better performance through cross-validation. Once we have the predicted Y (Yhat), we use an OLS estimator with Y_hat being the components to estimate the treatment effect. The OLS formula is as follow:

$$Y_i = a_0 + a_1 T_i + a_2\, Yhat_i + a_3 T_i (Yhat_i - mean(Yhat)) + e_I$$

where a1 is the estimated treatment effect, Ti is the treatment group (1,0), and Yi is the true Y outcomes.

Data Analysis

■ Method Replication

In this project, we applied the MLRATE method and compared against a simple estimator like the standard difference-in-means, as suggested by the referenced paper. The detailed process will be in the Jupyter Notebook.

■ Data Utilization

Simulated data was generated in accordance with the paper's specifications, ensuring an accurate

representation of the experimental conditions for MLRATE application, with 200 bootstrap samples. The X covariates are randomly drawn from a multivariate normal distribution $Xi \sim N(0,$ 1) with 10 X variables and 1000 observations (original data in paper was 100 X variables and 10000 observations, reduced for faster run time). The treatment and control groups are randomly assigned with 0.5 probability $T_i \sim Bernoulli(0.5)$. The treatment effect here is non-constant and generated by natural log and exponential functions of X covariates. Note that since the treatment effect is not a constant, no distribution plot of ATE estimates was created. Rather, we calculated the coverage rate of our resulting confidence intervals (the rate of our ATE CIs covering the true treatment effects). Error terms are generated from N(0, 25^2). Lastly, the Y outcomes are generated by non-linear functions of X, including the Friedman function.

As the paper stated, this data generating process will ensure the i.i.d assumption of X covariates, and the independence among treatment groups, X covariates, and the error term.

■ Analysis Procedure

Instructed by the paper, we first split the data into two sets using the stratified K-fold split (stratification here is the treatment/control group). Then we fit a machine learning model to the train data. The model we chose here is XGBoost, and the intuition is to learn the non-linear features in Y and perform well. After predicting the Y outcomes using XGBoost, we construct our OLS estimator with the components mentioned in the Methodology section. Finally, we obtain our OLS estimate of the treatment effect and the corresponding standard error.

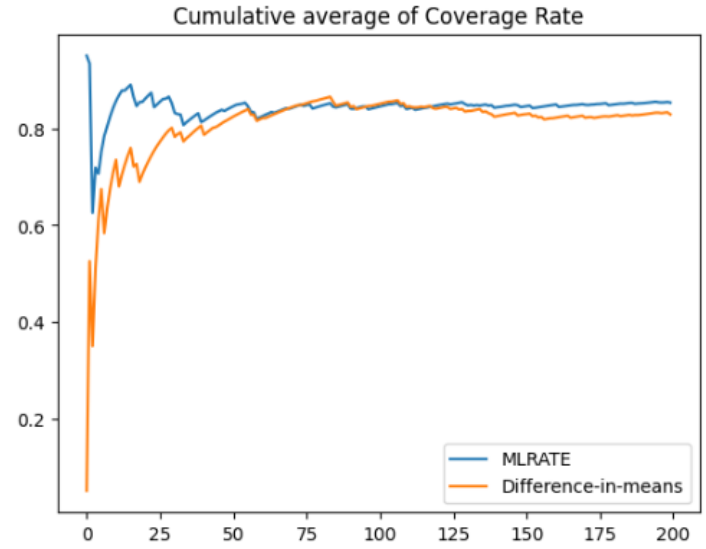
In the process of constructing intervals and calculating coverage rate, we utilized two hundred bootstrap samples to compare the average performance of MLRATE against the standard difference in means estimator. The analysis focused on two metrics: coverage rate and confidence interval width.

■ Observations:

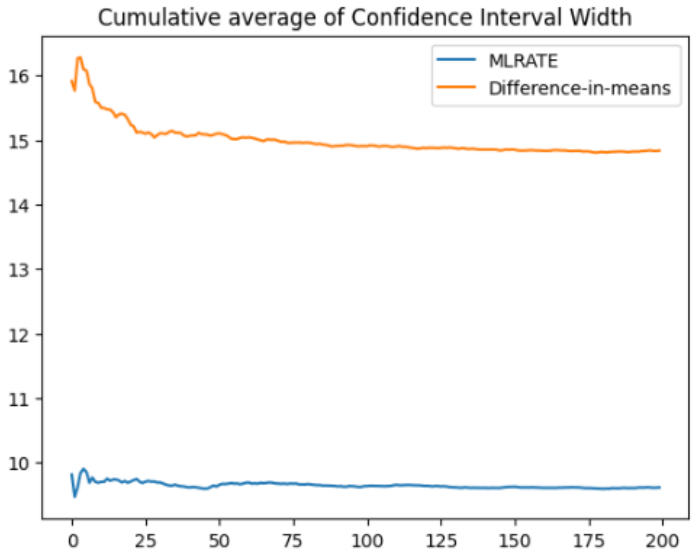
From the below graphs, it can be observed that the cumulative average of coverage rate of MLRATE is 0.853, and for Difference-in-means it is 0.829, increasing by 2.9%.

From the other graph, it can be seen that the average confidence interval width of MLRATE is 9.62, as opposed to the 14.84 CI width of Difference-in-means, decreasing by 35.2%.

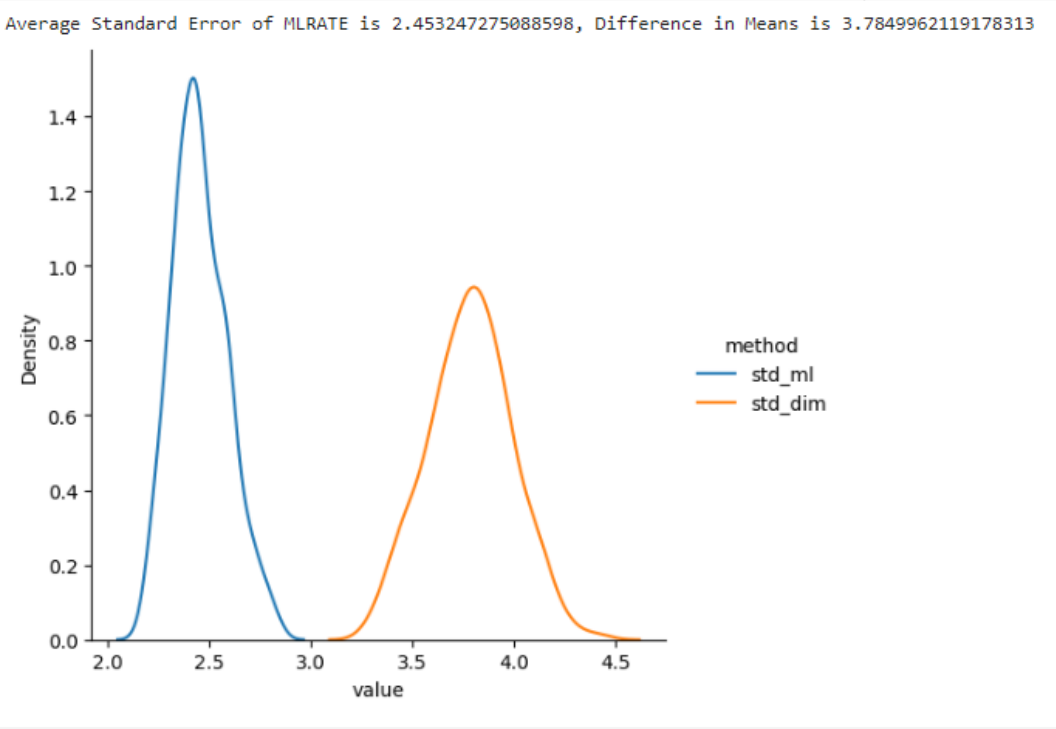
In summary, MLRATE achieved a higher coverage rate and a narrower confidence interval width, enhancing the reliability and precision of the ATE estimates.



Average Coverage Rate of MLRATE is 0.8532150000000001, Difference in Means is 0.828995



Average CI Width of MLRATE is 9.616552608689357, Difference in Means is 14.83691251395897



From the above graph, the standard error for ATE with MLRATE was significantly lower (2.45) than that from the difference in means (3.78), indicating improved efficiency. The distribution of standard error for two estimators showed great differences without any overlap, meaning that MLRATE significantly reduced the variance of ATE estimates.

Conclusion

The MLRATE method demonstrated indeed reduced the variance of treatment effect estimates over the simple difference-in-means approach, certified by large numbers of bootstrap samples. The analysis confirms the effectiveness of MLRATE in providing tighter confidence intervals and a higher coverage rate, consistent with literature claims. These findings support MLRATE's utility in online experimental settings, particularly where variance reduction is critical.

Citation

Machine Learning for Variance Reduction in Online Experiments. Guo Y., et al., 2021. Variance Reduction in Experiments, Part 2 - Covariance Adjustment Methods. Unal M.

Code reference:

https://github.com/muratunalphd/blog-posts