Project 4: Causal Analysis

Presented by Group 5

Yue Liang yl4391

Xujie Ma xm2243

Yiqi Lei yl4353

Xinyi Zhang xz2862

Jiaqi Yuan jy3021

Dec. 3, 2020

Content

Project Overview

Content

Goal & Methodology

Project Overview

Goal

Compare the performance, both the runtime and accuracy of the three models on two datasets.

Datasets

Implement the three models on both Low Dimension Dataset and High Dimension Dataset

Causal Effect

Utilize the Propensity Scores and the variables to calculate the Average Treatment Effect (ATE)

Propensity Scores

The calculation of Propensity Scores is a middle step for us to calculate the ATE

/01

Doubly Robust Estimation + Boosted Stumps

Methodology and Implementation

Doubly Robust Estimation + Boosted Stumps

01

Boosted Stumps

- Boosting Ensemble of weak learners (eg. 1-depth decision tree)
- Predict the propensity score.
- Simulate a random sample in an observational setting.

02

Doubly Robust Estimation

- Having the smallest asymptotic variance
- Estimator remains consistent even one of the two models are correctly specifies but the other are not: Propensity score model **or** two regression models.

03

Implementation

- Estimate Propensity Scores based on Boosted Stumps
- Regress response Y on X in the two groups
- Combine the difference with regression models and difference based on propensity scores to calculate the ATE.

Result

Doubly Robust Estimation + Boosted Stumps

** Tuning Time not included.*

	ATE	True ATE	Difference	Training Time
High Dimension Data	-2.962	-3.0	0.038	0.477 secs
Low Dimension Data	2.519	2.5	0.019	0.029 secs

Conclusion:

The model has a pretty good prediction on the ATE. And it is reasonable that it takes longer on higher dimension data.

Suggestion:

The model can be used on both high or low dimension datasets.

Regression Adjustment
+ Boosted Stumps

/02

Methodology and Implementation

Regression Adjustment + Boosted Stumps

Boosted Stumps

- Boosting Ensemble of weak learners (eg. 1-depth decision tree)
- Predict the propensity score.
- Simulate a random sample in an observational setting.

Regression Adjustment

- Regress the outcome variable Y on treatment indicator variables A and propensity score pred_PS.
- Take the regressed coefficient of variables A as the ATE.

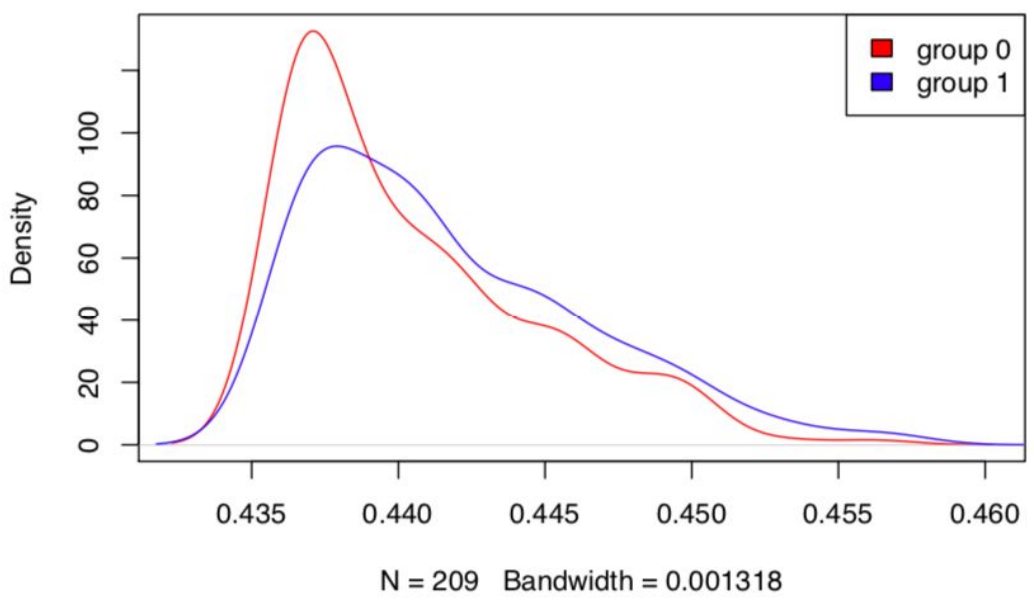
Implementation

- Learn the Boosted Stumps based on Generalized Boosted Regression Models **(bgm function in R)**
- Learn the ATE model using Linear Regression Model **(lm function in R)**

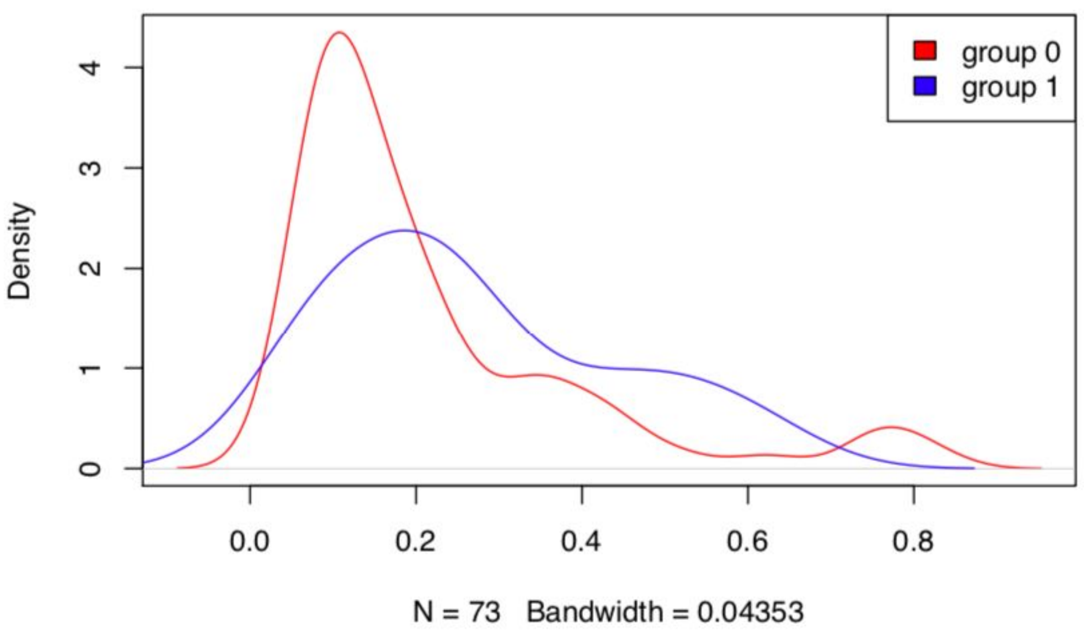
Visualization

Regression Adjustment + Boosted Stumps

High Dimension Data



Low Dimension Data



- The boosted stumps performs well on both of the datasets.
- Higher dimension data requires smaller learning rate and less number of trees.

Result

Regression Adjustment + Boosted Stumps

** Tuning Time not included.*

	ATE	True ATE	Difference	Training Time
High Dimension Data	-3.083	-3.0	0.083	0.099 secs
Low Dimension Data	2.527	2.5	0.027	0.085 secs

Conclusion:

The model has a pretty good prediction on the ATE due to the well-matched propensity scores estimated by Boosted Stumps.

Suggestion:

The model is better used on low dimensional datasets.

/03

Regression Estimate

Methodology and Implementation

Regression Estimate

Methodology

- Linear Regression on Two Groups
- Fit On The Whole Dataset
- Average The Difference

Straight Forward Model

- No Propensity Scores Required
- Easy Interpretation Of ATE Calculation
- As Well Performance As Other Mature Methodology

Implementation

- Computational Efficiency
- Use Linear Regression (lm function in R)

Result

Regression Estimate

** No Tuning Time needed.*

	ATE	True ATE	Difference	Training Time
High Dimension Data	-2.960	-3.0	0.040	0.166 secs
Low Dimension Data	2.527	2.5	0.027	0.015 secs

Conclusion:

The model has a pretty good prediction on the ATE and it runs the model pretty fast.

Suggestion:

The model is better used on lower dimension datasets.

/04

Comparison
Across Three Models

Accuracy Comparison

Across Three Models

Algorithm	ATE High	difference	ATE Low	difference
True ATE	-3	-	2.5	-
Doubly Robust Estimation + Boosted Stumps	-2.962	0.038	2.519	0.019
Regression Adjustment + Boosted Stumps	-3.083	0.083	2.527	0.027
Regression Estimate	-2.960	0.040	2.527	0.027

Doubly Robust Estimation

Best Performance
On Both Datasets

Regression Adjustment

Better Performed On Low
Dimension Dataset

Regression Estimate

Perform Equally Well
On Both Datasets

Training Time Comparison

Across Three Models

** Tuning Time not included.*

Algorithm	Training Time High	Training Time Low
Doubly Robust Estimation + Boosted Stumps	0.477	0.029
Regression Adjustment + Boosted Stumps	0.099	0.085
Regression Estimate	0.166	0.015

Doubly Robust Estimation

Higher Training Time On High Dimension Dataset

Regression Adjustment

Lower Runtime On High Dim Dataset Among models

Regression Estimate

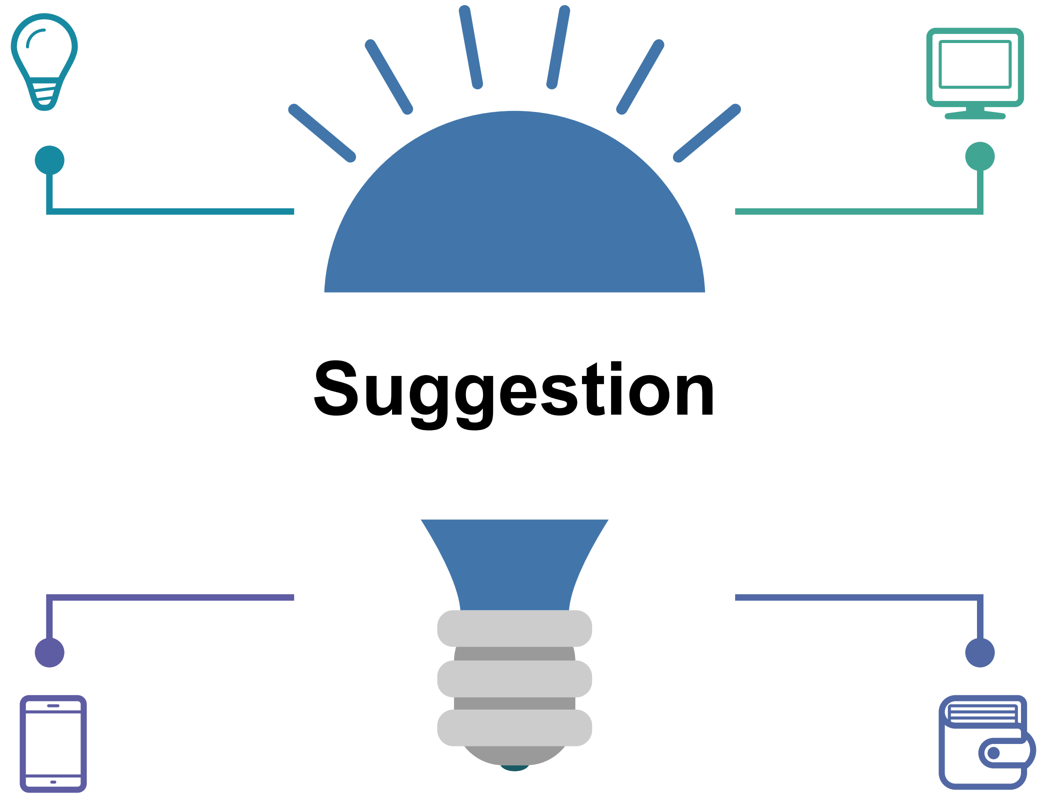
Lower Runtime On Low Dim Dataset Among models

Suggestion

Across Three Models

Higher Accuracy

Recommend Doubly Robust Estimation With Boosted Stumps



Computational Efficiency

Recommend Regression Estimate

Appendix

Algorithm of Three Models

Doubly Robust Estimation

$$ATE = E[E(Y|T=1,X) - E(Y|T=0,X)] + E[(\frac{I[T=1]}{propensity\ score} - \frac{I[T=0]}{(1-propensity\ score)})(Y - E(Y|T,X))]$$

in which $E(Y|T=t,X)$ is usually obtained by regressing the observed response Y on X in group t (where t = 0,1).

Regression Adjustment

$$Y = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 PS$$

$$ATE = \hat{\beta}_1$$

Y is the outcome variable

A is treatment indicator variables

PS is the estimated Propensity Scores

Regression Estimate

$$ATE = N^{-1}\sum_{i=1}^{N}(\hat{m}_1(X_i) - \hat{m}_0(X_i))$$

N is the number of samples in the dataset,

X_i is the datapoint in the dataset,

m_1 is the regression model learned from the treated groups,

m_0 is the regression model learned from the untreated groups,

$\hat{m}_1(X_i)$ is the prediction of the regression model m_1 on the datapoint X_i,

$\hat{m}_0(X_i)$ is the prediction of the regression model m_0 on the datapoint X_i.

THANKS
For Listening !

Contributors - Group 5

Xujie Ma, Yiqi Lei, Xinyi Zhang, Jiaqi Yuan, Yue Liang