Regression Estimate

Regression Estimate

Understanding

Regression Estimate is a really simple estimation model to the calculate ATE, which do not require Propensity Scores calculation. This makes it a straight forward model and a computational efficient model. By implementing the linear regression on treated groups and untreated groups, we could regress on different groups to get the two different sets of paramaters and then by predicting the models on the whole dataset, substracting the prediction we can get the difference between the two regression models. In the end, we can calculate the ATE(Average Treatment Effect) by taking the average of the difference.

$$ATE = N^{-1} \sum_{i=1}^{N} (\hat{m_1}(X_i) - \hat{m_0}(X_i))$$

Denote that

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.

Implementation

Read the data and split the data into two groups – Treated Group and Untreated Group

```
high_data <-read.csv('../data/highDim_dataset.csv')
low_data <-read.csv('../data/lowDim_dataset.csv')

N_high <- dim(high_data)[1]
N_low <- dim(low_data)[1]

high_data_X <- high_data[,3:dim(high_data)[2]]
low_data_X <- low_data[,3:dim(low_data)[2]]

high_treated <- high_data[high_data$A==1,-2]
high_untreated <- high_data[high_data$A==0,-2]

N_high_treated <- dim(high_treated)[1]
N_high_untreated <- dim(high_untreated)[1]

low_treated <- low_data[low_data$A==1,-2]
low_untreated <- low_data[low_data$A==0,-2]

N_low_treated <- dim(low_treated)[1]
N_low_untreated <- dim(low_untreated)[1]
```

Train the data and record the training time of two datasets

```
time<- system.time({
  high_treated_lm <- lm(Y~.,data = high_treated);
  high_untreated_lm <- lm(Y~.,data = high_untreated);
  high_treated_predict_all <- predict(high_treated_lm,newdata = high_data_X);
  high_untreated_predict_all <- predict(high_untreated_lm,newdata = high_data_X)})
train_time_high <- time[1]
train_time_high
```

```
## user.self
##     0.144
```

```
time<- system.time({
  low_treated_lm <- lm(Y~.,data = low_treated);
  low_untreated_lm <- lm(Y~.,data = low_untreated);
  low_treated_predict_all <- predict(low_treated_lm,newdata = low_data_X);
  low_untreated_predict_all <- predict(low_untreated_lm,newdata = low_data_X)})
train_time_low <- time[1]
train_time_low
```

```
## user.self
##      0.01
```

Calculate the ATE

```
reg_est_ATE_high<-sum(high_treated_predict_all - high_untreated_predict_all)/N_high
reg_est_ATE_low<-sum(low_treated_predict_all - low_untreated_predict_all)/N_low
reg_est_ATE_high
```

```
## [1] -2.95978
```

```
reg_est_ATE_low
```

```
## [1] 2.526944
```

Compare the ATE with the true ATE

```
# True ATE:
true_ATE_high <- -3
true_ATE_low <- 2.5

# Comparison:
abs(true_ATE_high - reg_est_ATE_high) /abs(true_ATE_high)
```

```
## [1] 0.01340679
```

```
abs(true_ATE_low - reg_est_ATE_low) /abs(true_ATE_low)
```

```
## [1] 0.01077759
```

Conclustions

Comparision between the two dataset

We can conclude that the model is more fit to the low dimension dataset. With higher dimension, the ATE has higher bias rate(1.34% vs 1.08%).

Comparision among the three models

The table shows the result of the three algorithm's ATE in the two different datasets.

Algorithm	High ATE	Low ATE	High Train Time	Low Train Time
True ATE	-3	2.5	-	-
Doubly Robust Estimation + Boosted Stumps	-2.9626	2.5187	1.2180	0.0230
Regression Estimate	-2.9598	2.5269	0.2270	0.0190
Regression Adjustment + Boosted Stumps	-3.0830	2.5271	0.5060	0.1287

From the table above, we can clearly conclude that the Regression Estimate's accuracy is relatively high, but slightly lower than the Doubly Robust Estimation + Boosted Stumps model. However, the training time of Doubly Robust Estimation + Boosted Stumps model is higher than the Regression Estimate model for both high dimension dataset and low dimension dataset. We can conclude that the Regression Estimate is more computational efficient but slightly less accuracy than the Doubly Robust Estimation + Boosted Stumps model.