Overview

Description of the method

1. Assumptions:

Unconfoundedness: $[W_i \perp (Y_i(0), Y_i(1)) \mid X_i]$

2. Outcomes:

ATE: $ATE = \mathbb{E}[\beta] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$

CATE: $\tau(x) = \mathbb{E}[\, Y_i(1) - Y_i(0) \mid X_i = x\,]$

Research Motivation

I generate the data referring to the essay, and its *DGP* comes from the essay:

- Huseyin Guleny, Candace E. Jensz, T. Beau Pagex, An application of causal forest in corporate finance: How does financing affect investment? ,2020.04.23.

Theoretical references:

- Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113, 7353-7360.
- GuidoWImbens & Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.
- Stefan Wager & Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association, 0162-1459.

My Project:

This project aims to compare how well OLS and causal forest work in different situations. Here I will change several sample sizes and models to show the difference.

This project compares the linear and non-linear specification estimations respectively, along three dimensions: bias, precision, and coverage. I measure bias as the difference between the estimated ATE and the true value of ATE of each model, with confidence interval under 97.5% confidence level. Precision is the root mean squared error (RMSE), measured as the standard deviation of the difference between the estimate and the true value. Coverage is the fraction of trials in which a t-test of the estimate compared to the true value rejects at the 5% level.

I visualize bias by line/point/box plots, and present precision (RMSE) and coverage in tables. All else equal, we prefer an estimator with zero bias, low RMSE, and test coverage of 5%.

DGP

y is the dependent variable, the three *x* variables are additional covariates, *D* is a binary variable equal to one if the forcing variable *d* is greater than zero. I simulate the *x* and *d* variables using a multivariate normal distribution. For the vector (x1, x2, x3, d), I set the mean to (1.40, 5.50, 0.10, 0.20), standard deviation to (1.00, 1.50, 0.30, 0.30), and use the correlation matrix:

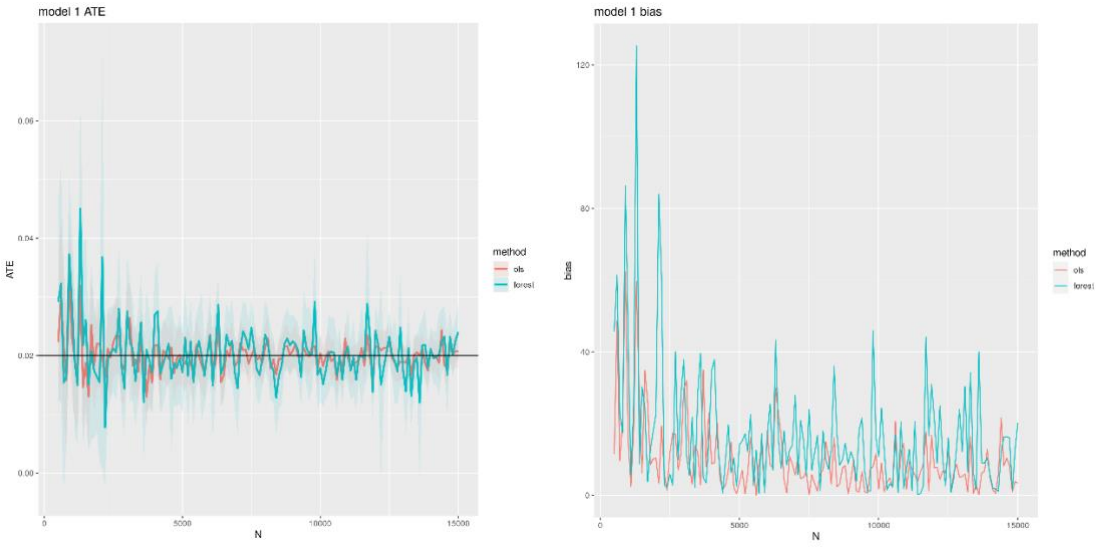
$$\begin{bmatrix} 1.00 & -0.05 & -0.30 & 0.15 \\ -0.05 & 1.00 & 0.20 & 0.35 \\ -0.30 & 0.20 & 1.00 & 0.50 \\ 0.15 & 0.35 & 0.50 & 1.00 \end{bmatrix}$$

The ε term has a mean of zero and a standard deviation of 0.065.

Model 1: linear model

$$y = 0.05x_1 - 0.005x_2 + 0.01x_3 + 0.02D + \varepsilon$$

Simulations for the linear base case demonstrate the following: we can see from plots and table that OLS is the best estimator for sample size of 500 to 5000, since causal forest has higher variance and performs more unstable than OLS.



Line plots 1

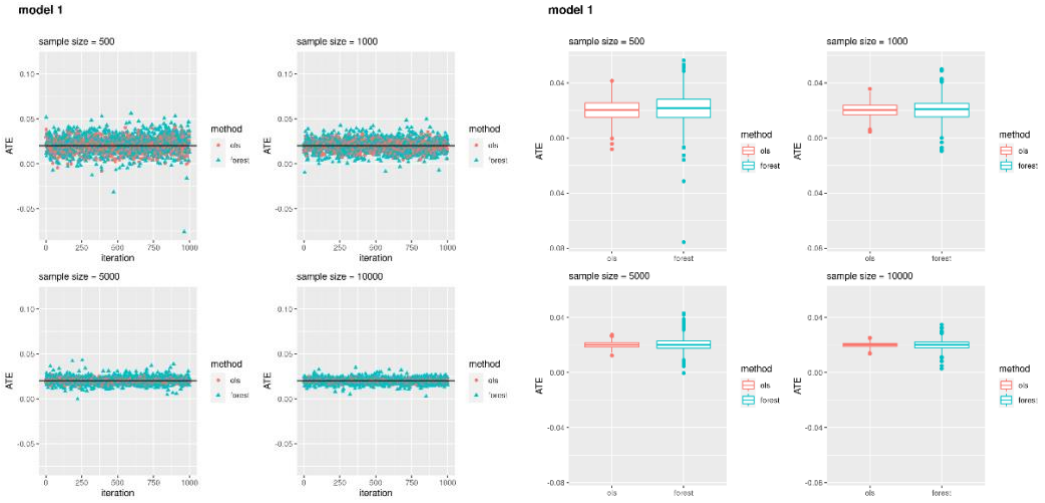
(OLS is the red line, causal forest is the blue line)

As sample size increases, RMSE decreases for both OLS and causal forest, and causal forest is the dominant estimator in large sample size. Here in the table 1 you can see the plot data in sheets (which is the original data I used for simulation 1 and 2 of line plot and box respectively) and calculate the mean value of them to get an overall view about it. Here is the result:

Method	N	ATE	RMSE	Bias	Coverage	Model
forest	500	0.021424	0.066894	40.91071	0.030196	model 1
forest	1000	0.020374	0.066424	30.08157	0.003989	model 1
forest	5000	0.020306	0.065945	16.20785	6.45E-11	model 1
forest	10000	0.020011	0.065871	12.71641	4.52E-17	model 1
ols	500	0.020235	0.064968	30.34241	0.06012	model 1
ols	1000	0.020046	0.06494	20.76277	0.007693	model 1
ols	5000	0.020027	0.064843	9.231935	2.68E-10	model 1
ols	10000	0.020029	0.064686	6.660551	1.75E-18	model 1

Table 1

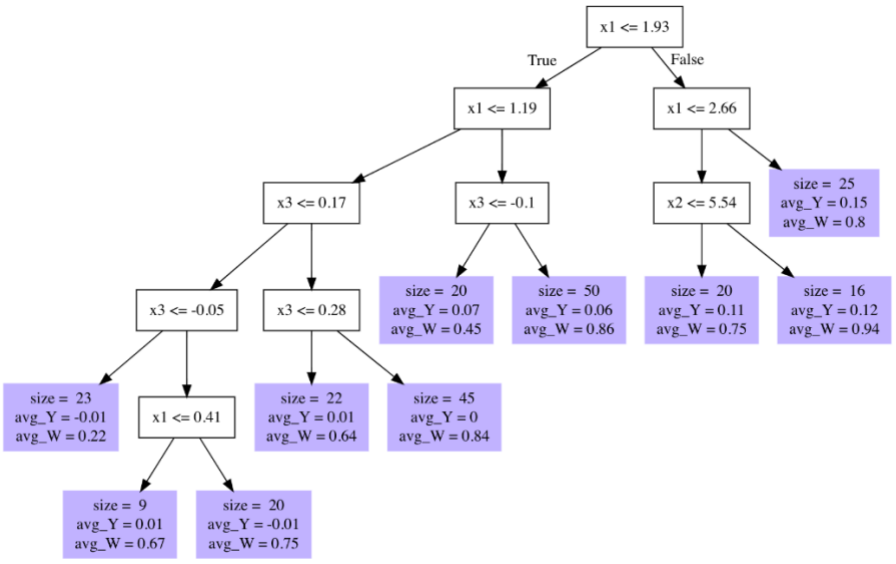
Although causal forest and OLS have comparable RMS for all sample sizes, causal forest deceases its bias as sample size increases and it has lower bias than OLS at sample size of 10000. This is because causal forest is data hungry and thus needs large sample size to work well.



Point & box plots 1

(OLS is the red line, causal forest is the blue line)

To better understand the tree method, I draw a tree plot to show how causal forest works in this process with a sample size of 1000. It shows the sample size as well as the average value of W (ATE) in each branch.

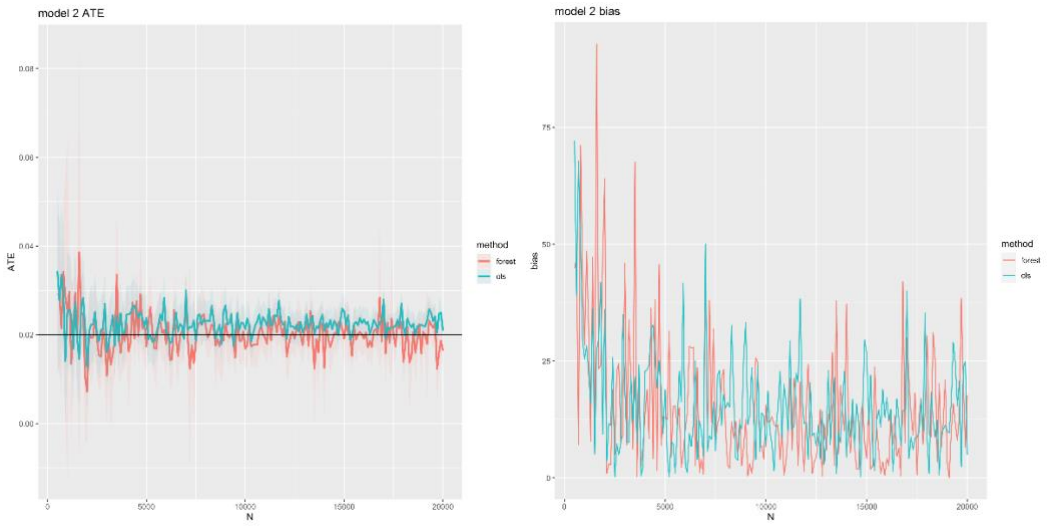


Tree 1

Model 2: polynomial specification

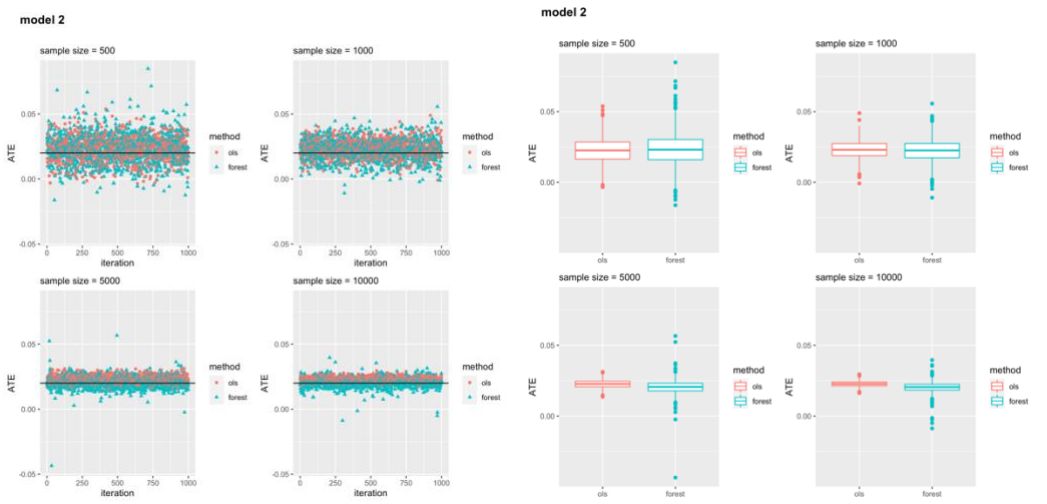
$$y = 0.05x_1 - 0.005x_2 + 0.01x_3 + 0.025x_1^2 - 0.01x_2^2 + 0.015x_3^2 + 0.02D + \varepsilon$$

In the following models we will have misspecification problem. Therefore, the treatment effect is biased for OLS model. However, causal forest works pretty well in heterogenous designed model and its ATE value is closer to the true value. We can see from the plots that ATE in OLS (green line) is always waving above the true value, while causal forest (red line) has a fluctuation up and down the true value line.



Line plots 2

As sample size increases, causal forest works better than small sample size case and both OLS and causal forest decrease their variance (improve precision) in this process.



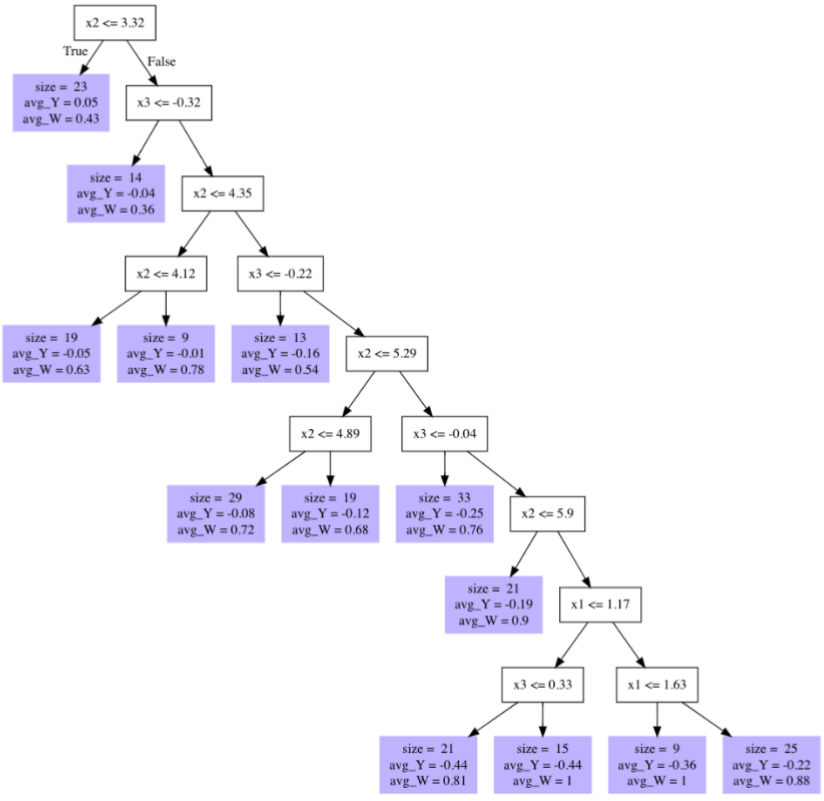
Point & box plots 2

Bias is higher for causal forest in small sample size, then it performs better as sample size increases. The reason is listed above that causal forest needs large data to perform well. As sample size increases, both OLS and causal forest work better in variance and coverage.

Method	N	ATE	RMSE	Bias	Coverage	Model
forest	500	0.022861	0.081339	47.0289	0.03996	model 2
forest	1000	0.021653	0.074952	34.27837	0.007238	model 2
forest	5000	0.020746	0.068529	17.07244	1.76E-09	model 2
forest	10000	0.020299	0.067438	12.64895	8.28E-21	model 2
ols	500	0.022877	0.080555	39.0469	0.084918	model 2
ols	1000	0.022571	0.080572	29.38204	0.01897	model 2
ols	5000	0.02271	0.080214	16.52549	6.14E-08	model 2
ols	10000	0.022647	0.080066	14.16817	5.26E-18	model 2

Table 2

The tree plot works as follows:

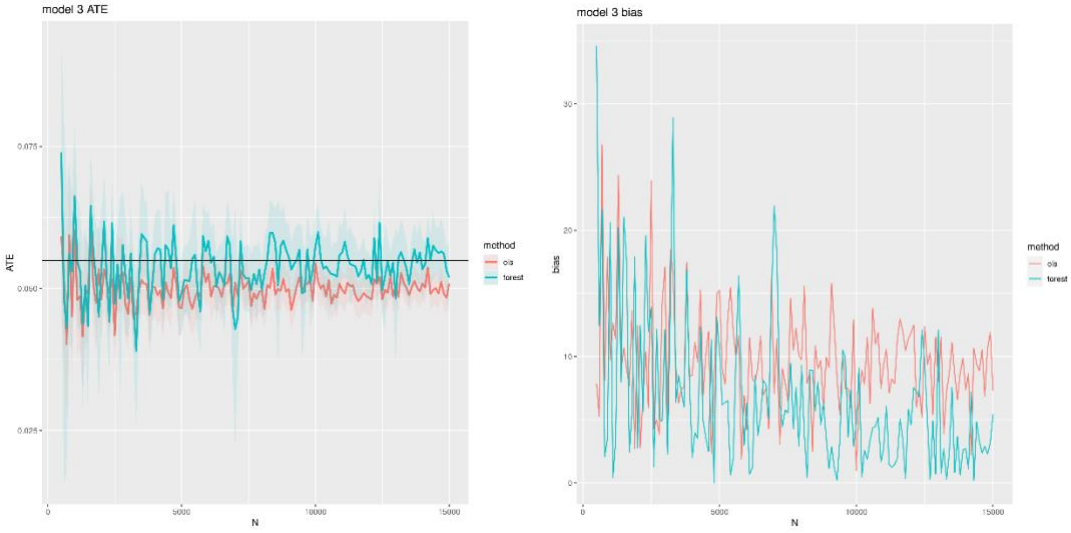


Tree 2

Model 3: with interaction terms

$$y = 0.05x_1 - 0.005x_2 + 0.01x_3 + 0.02D + 0.01x_1D + 0.001x_2D + 0.002x_1x_2D + \varepsilon$$

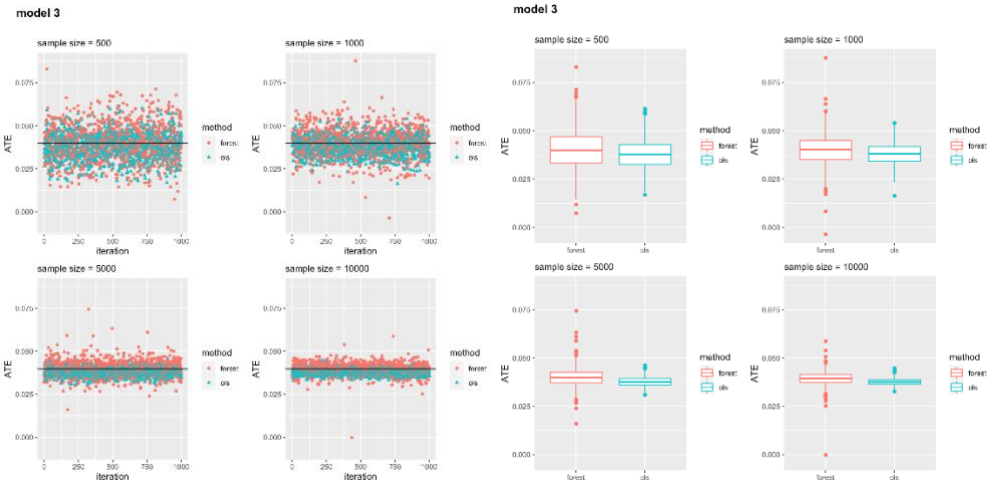
In model 3 we have omitted variable problem. Because of the heterogeneous design, causal forest should perform better than OLS. As can be seen from results, OLS has high bias in every specification and, as sample size increases, worse coverage.

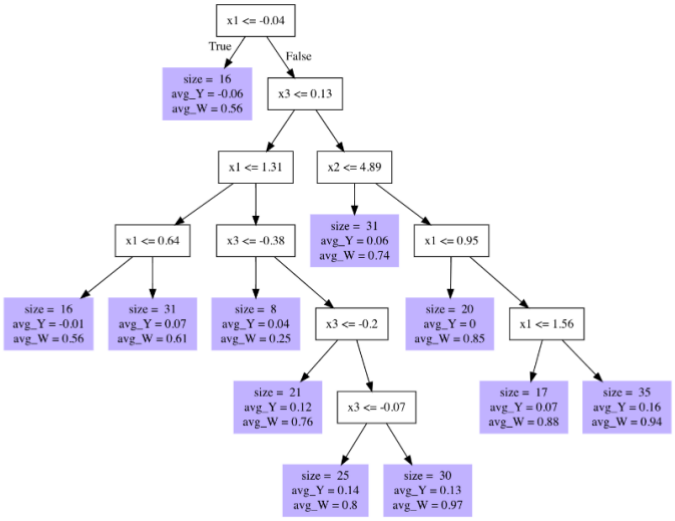


Line plots 3

Method	N	ATE	RMSE	Bias	Coverage	Model
forest	500	0.054335	0.07004	15.17324	2.52E-06	model 3
forest	1000	0.054238	0.06957	11.28336	2.61E-11	model 3
forest	5000	0.054079	0.068853	6.228131	2.16E-71	model 3
forest	10000	0.054412	0.068705	4.734073	1.55E-154	model 3
ols	500	0.049736	0.06569	13.69813	5.16E-06	model 3
ols	1000	0.049922	0.065689	11.02836	2.16E-11	model 3
ols	5000	0.049963	0.065614	8.856372	1.12E-67	model 3
ols	10000	0.049961	0.065292	8.755199	1.12E-141	model 3

Table 3

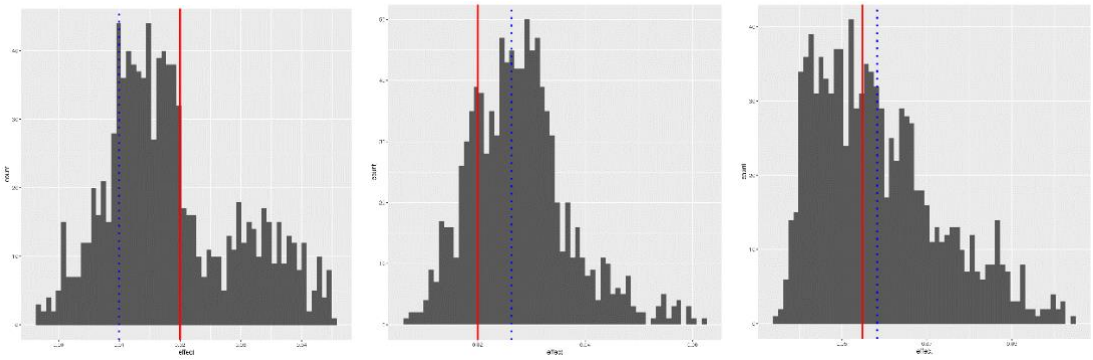




Tree Plot 3

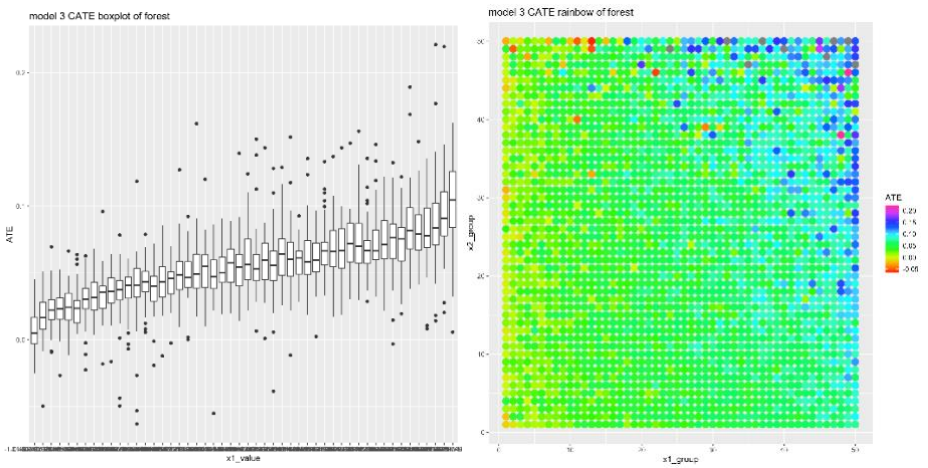
Comparison of the tree plots:

We can see from the tree plots that tree branches increase from model 1 to 3, and meanwhile bias decreases. From left to right are plots for Model 1 to model 3 respectively, we can see that bias of causal forest becomes smaller in this process. (True ATE is the red line, estimated ATE is the blue line).

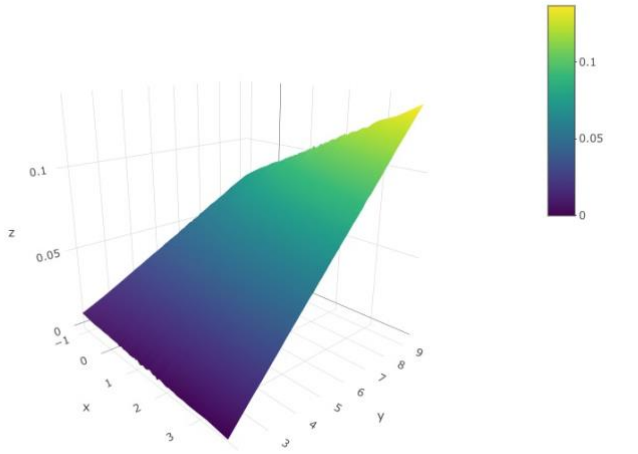


Histogram 1

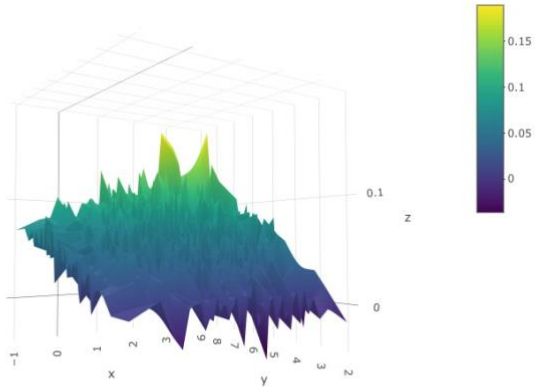
Next step we will calculate CATE value for x1 ad x2, and present it in rainbow plot and 3D plot. I equally divide the simulation into 50 partitions with sample size of 500000, but there are missing data due to the Gaussian distribution of X. The plots are presented as follows:



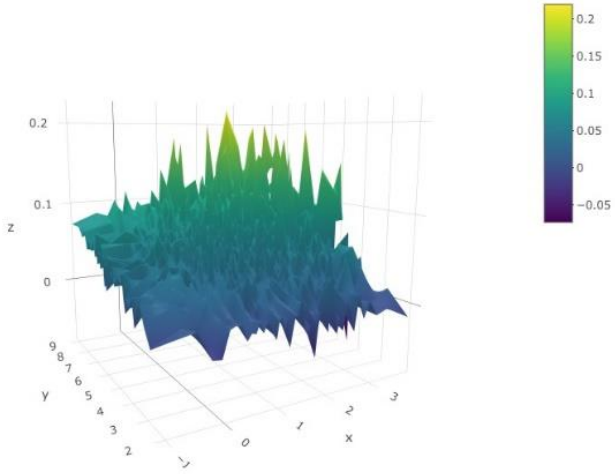
See more details on the websites included in *Yuxin_ Project_ Pics & Tables* File.



Beta true for model 3



OLS estimates for model 3

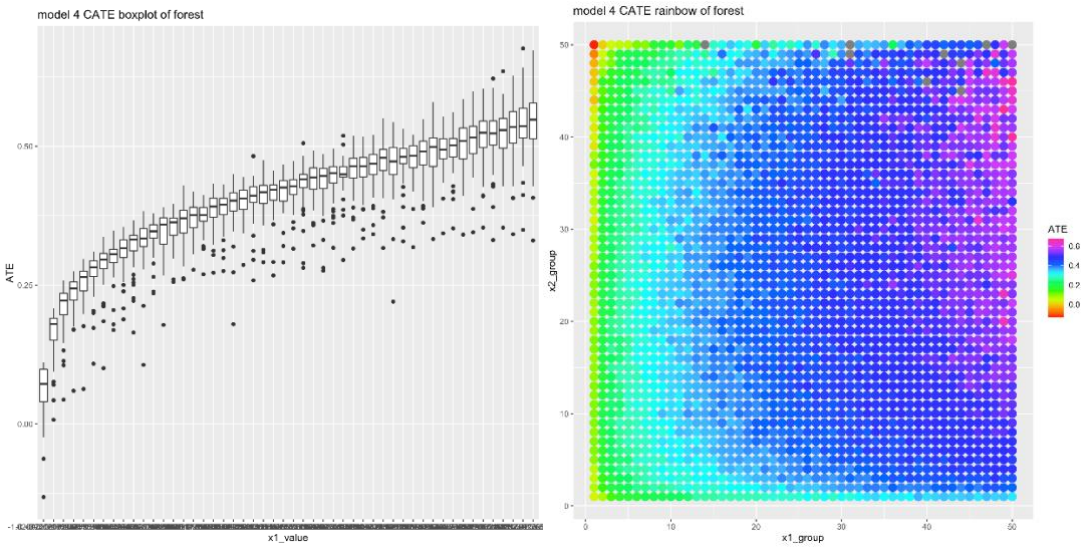


Causal forest estimates for model 3

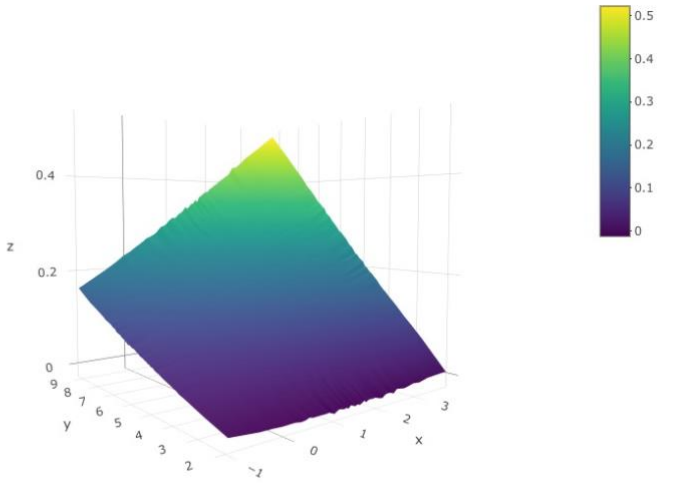
Model 4: with polynomials and interaction terms

$$y = 0.05x_1 - 0.005x_2 + 0.01x_3 + 0.02D + 0.01x_1D + 0.001x_2D - 0.02x_1^2D - 0.01x_2^2D + 0.01x_1x_2D + \varepsilon$$

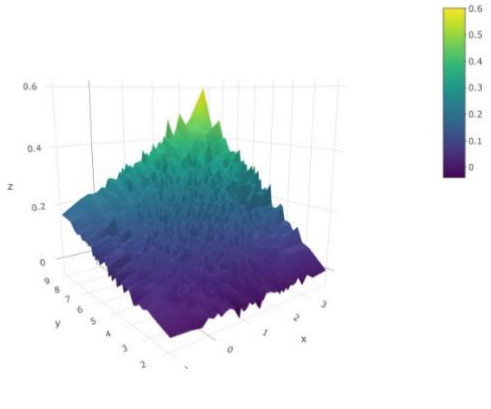
Here is the boxplot and rainbow plot for CATE of X1 and X2:



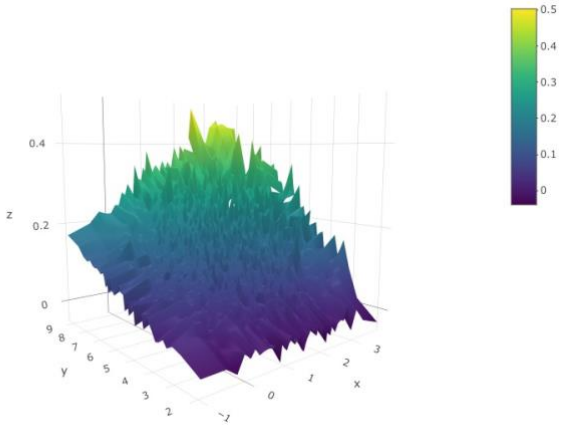
For 3D plots, see more details on the websites included in *Yuxin_ Project_ Pics & Tables* File.



Beta true for model 4

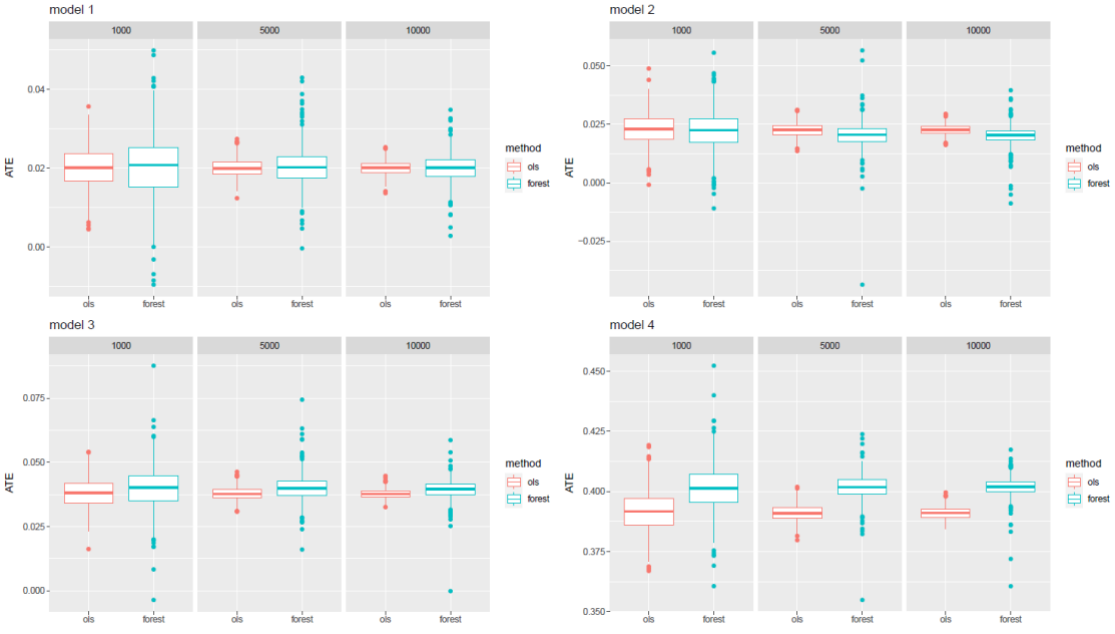


OLS estimates for model 4



Causal forest estimates for model 4

Finally, by the aggerated box plot for 4 models, we can have a clear look about it.



Aggregated Box Plot

Empirical Application:

As we can see from the project results, causal forest treatment has great extendibility and it has lots of empirical applications. For example, the essay I refer to is about an application of causal forest in corporate finance. Here I list two reasons for its goodness as an estimating tool:

First, many questions in economics or finance involve evaluating the treatment effect of a binary status or decision, in which case we need to deal with treatment effect heterogeneity. This is the type of question causal forest is designed to address.

Second, plenty of macroeconomic studies tend to have large data to deal with, thus providing a sufficient sample size. Our simulations show the importance of sample size in a causal forest estimation to recovering unbiased treatment effects, and confirm sample sizes as small as 5,000 to 15,000 are sufficient for a causal forest estimation with low error.