PAPER TITLE Blah Blah Estimating Reaction to Welfare Using Causal Forests

ECMA 31330 Final Project

Spring 2021

Max Bronckers
Veronica Song
Dustin Zhang

1 Introduction

There has been an increasing adoption of Machine Learning (ML) methods in economics for causal inference. While initially ML methods were avoided due to an uncertainty in their consistency, normality, and efficiency, major developments in methodology have allowed a stable large-sample confidence interval to be constructed around treatment effect estimates conditional on multiple covariates–leading to the broader adoption of these methods. [1]

One such method is Bayesian Additive Regression Trees (BART), a non-parametric method that models heterogeneous treatment effects flexibly by building on the concept of ensembles of trees. Using a Markov Chain Monte Carlo (MCMC) algorithm that derives effects from the posterior mean and interval instead of pre-specified tree parameters, BART has a much smoother and adaptive structure than traditional OLS or single tree models popular in economics, and is also resilient to problems with overfitting. [6] However, despite its excellent predictive capacity, BART lacks an asymptotic explanation for its estimates and is not guaranteed to converge in polynomial time [3], weakening its effectiveness as inference tool. Though recent work by Ročková and Saha (2018) suggests modifications to BART that may allow asypmtotic concentration of the posterior mean around the true mean, the construction of an asymptotic theory of BART estimates is still an ongoing effort. [9] An alternative method is the Causal Forest (CF), proposed by Wager and Athey (2017). One of the advantages of CF is that their estimates are asymptotically gaussian and unbiased, allowing proper confidence intervals to be constructed around the treatment effect. The construction of adequate confidence intervals is especially relevant in policy applications, as consistent estimates can be produced for the treatment group. [3]

In this paper, we evaluate Causal Forests as a method for heterogeneous treatment effect estimation on both empirical and simulated datasets. Specifically, we compare the CF method to BART in the estimation of heterogenous treatment effects in a survey experiment on welfare opinions from the General Social Survey as used by Green and Kern. Green and Kern use the GSS survey dataset with individual perception on public spending. To find the impact of the question phrasing on the responses, they use BART to estimate heterogeneous treatment effects by conditioning on a suite of socioeconomic backgrounds of the respondent. Given the theoretical shortcomings of BART and the theoretical benefits of CFs, we apply CFs to same problem and empirical dataset and compare it to the authors' findings.

In section 2, we fit a CF model on the welfare data and compares our estimates against BART estimates of conditional average treatment effect (CATE) obtained by Green and Kern. Since empirical data offers no ground truth CATEs, we are confined to comparing the methods on interval length and RMSE. In section 3, we remedy this shortcoming with an evaluation of the CF method applied to DGPs that attempt to represent the empirical dataset. Using the ground truth ATE and CATEs of our DGPs, we assess the CF's performance under a variety of different DGP parameters and assess the conditions under which CFs perform well or poorly.

2 Causal Forest Estimates of CATE in Welfare Dataset

2.1 Discussion of data and motivation for research

In our analysis below, we use a survey experiment from GSS which is also utilized by Green and Kern to investigate interactions between treatment and covariates that may lead to treatment effect heterogeneity. The experiment was conducted in the mid-1980s by GSS to study the negative sentiment Americans carry toward government programs labeled as "welfare". Due to associations with racial connotations and poorly managed welfare programs, respondents were found much more likely to endorse government spending for "the poor" than for public "welfare". [8]

Using BART, Green and Kern illustrates the extent to which such reaction to the question wording as "welfare" varies based on the respondents' background characteristics such as years of education, race, or political alignment. BART has been a popular choice for heterogeneous treatment effect modeling, as its estimates require little parameter tuning, allow accurate detection of interactions between covariates, and are much smoother than those of single tree models. Each individual tree in the forest has only a small effect on the model, by assuming a prior distribution over the tree parameters. [5] Such lack of any individual influential trees allow regular model fit of the BART set up. Then, a MCMC algorithm is used to sample tree parameters iteratively from the posterior distribution as the model is fit. Though this assumption of prior distribution and back-fitting algorithm allows BART to be relatively invariant across, in the presence of confounding variables and treatment effect heterogeneity, such regularization may severely bias the treatment effect estimates [7]. Moreover, BART estimates still lack theoretical explanation on its asymptotic concentration, making the construction of adequate confidence intervals challenging.

2.2 Causal forests and model assumptions

For these reasons, we choose to investigate causal forests to estimate the the treatment effect of question wording on welfare program support. Just like BART, causal forests is fit to model non-linear relationships and interactions between covariates. One advantage that causal forests carry over BART is that under weak assumptions, the estimate are asymptotically standard normal distributed with Gaussian confidence intervals. [3]. Put in context of economic policy, the presence of an asymptotic theory allows hypothesis testing regarding treatment results and thus aids policy decisions to be made.

Causal forests are a specific form of generalized random forests (GRF), which uses adaptive sample splitting criterion taking into account the MSE. To avoid overfitting and reduce bias in the estimates, we also ensure that the tree is 'honest' - the subsample with which we grow a tree is disparate from the subsample with which we drop down and obtain predictions. Additionally, we also note that although causal forests uncover heterogeneous treatment effects with valid confidence intervals for statistical inference, it does not necessarily address the affect of confounding due to the regularization on our trees. The terminal leaves of our tree are not homogenous across covariates for the sake of lowering variance and thus increasing precision, but with confounding within the leaves, we cannot guarantee that our treatment effect estimates will be unbiased. We thus use the Double/Debiased Machine Learning method (DML) for causal forests proposed by Chernozhukov et al. (2016) that uses orthogonalized treatment on covariates to estimate the treatment effect. [4] We achieve this through the *CausalForestDML* function. In our model, we use 80% of the welfare data to build our tree and estimate treatment effects on the rest with a 5-fold CV for parameters of the model.

2.3 Results of analysis and explanation

We first investigate the average treatment effects (ATE) without conditioning on any covariates for possible effect heterogeneity. We obtain an ATE of 0.336, with a standard deviation of 0.049 and a 95% confidence interval of (0.256, 0.416). Green and Kern (2012) identified seven variables to condition the treatment effects on: *party identification, political views, age, education, negative attitude towards Blacks*, and *survey year*. Figure 1 displays the CATE estimates obtained by the causal forest model conditional on each of the seven variables. The blue areas represent the 95% confidence intervals of our estimates.

The top two graphs represents CATE as a function of party identification and self-identified political alignment from liberal to conservative. We see around 5 and 3 percentage points difference in the effect of question wording on support for welfare spending between strong Republicans-Democrats and Conservative-Liberals, respectively, controlling for all other covariates. Conservatives and strong Republicans are more likely to be affected by the framing of the survey question as for welfare. The treatment effect conditional on age, on the other hand, is greatest for those in their 30-40s, and diminishes past that age group. The negative bias toward Blacks has a less pronounced moderation on the treatment effect than the BART estimates. We note that the trends in CATE estimates obtained via causal forest are generally similar to those obtained using BART for most of the covariate groups except education. There seems to be an increasing effect of question wording as the responded receives greater years of education, with the effect peaking at around 11-13 years (with college education). Whereas Green and Kern found no distinct moderation of treatment effects based on education, we find that there is around 5 percentage point difference in treatment effect estimates between those who received no education and those who received post-Graduate education. This indicates that more educated individuals respond more favorably to the question worded as 'assistance to the poor'. Lastly, we examine that change in CATE through time (survey year) and find that the treatment effect was strongest during years 1993-1996. Our results once again largely agree with the results obtained through BART, and illustrate that response to "welfare" is highly associated with the respondent's personal background characteristics.

We note that our CATE estimates from causal forest are on average lower than those of BART. The higher estimates of the BART estimates can be attributed to the regularization-induced confounding (RIC) identified by Hahn, Murray, and Carvalho (2019).[7] The treatment effect in BART is obtained by taking the conditional expectation of the outcome conditional on treatment and certain covariates: $E[Y|x, Z = 1] - E[Y|x, Z = 0]$, which in the presence of confounding in finite sample size, outcome may be dependent mostly on x rather than the treatment Z. RIC thus states that the obtained CATEs are heavily dependent on the regularization by the prior distribution in samples of defined sizes.

Figure 2 shows the overall presence of treatment effect heterogeneity in the sample. The histogram shows that treatment effect ranges from 10 percentage points to 52 percentage points, with the median estimated CATE of 34 percentage points. We note that compared to the BART median of 37 percentage points, we again obtain lower estimates on average. All estimates of CATE were positive, indicating that although the degrees of the response varies based on personal background covariates, all respondents react more favorably to a question framed for 'the poor' than for 'welfare'.

Add Interpretation of results, Comparison to BART (Bias, RMSE, Coverage, Int length)?? TBD

2.4 Suggestions for further analysis

The causal forests approach also allows us to gauge covariate importance by calculating SHAP values, as indicated in figure 3. We note that aside from the covariates estimated by Green and Kern, we find that variables as work status and racial backgrounds are also significant in the model. Although in this paper we limit the scope of our research to comparisons with the authors' estimates through BART, it may be meaningful in future research to estimate CATE conditional on those variables.

3 Testing Parameter Robustness of Causal Forest Estimates on Synthetic DGP

The analysis using the welfare data in the section above gives us an interesting real-life application of causal forest method; however, given the lack of ground truth in the empirical situation, we are unable to test the performance of our model in terms of metrics as bias and MSE. In this section, we explore multiple synthetic DGPs to estimate how causal forest performs under different data parameters.

3.1 Baseline DGP specification

We specify a DGP that resembles our empirical data using a simplified model of N individuals and 120 covariates. Each variable is drawn from a gaussian normal with mean and covariance obtained from the empirical data. We then take each randomly generated covariate and convert to its corresponding form. We assume that the outcome data is generated as:

$$Y = B\Gamma + T \cdot \Theta(X) \tag{1}$$

where $\Gamma = X + \sum_{i=0}^{I} X_i^{\omega}$ is the full list of covariates including their higher-order interactions with the i most important features in the model. This parameter models the degree of linearity in the response surface: i features are selected based on SHAP values obtained in the empirical analysis from section 1, and raised to the order of ω to model non-linear interactions and increased importance of these variables. In our baseline model, we sample N=5000 individuals with parameters $\omega = 3, i = 4$ for 'med'-degree linearity. T, our treatment vector, models both binary and continuous treatment assignment; in the binary case, $T \sim Binomial(n, p)$ where p is the probability of receiving treatment (propensity score). We assume there is a 50% probability of treatment in the baseline model. In the continuous case, we sample treatment from $T \sim Unif(0, 1)$. We model heterogeneous or homogeneous treatment effects through $\Theta(X)$ as defined in equation (2).

$$\Theta(X) = \begin{cases} \delta_H & \{H = 0\} \\ \delta_H \cdot \left(\sum_{j=0}^{J} X_j + (\sum_{j=0}^{J} X_j)^{\omega} \right) & \{H = 1\} \end{cases} \tag{2}$$

Variable H indicates whether heterogenous treatment effects are present in the model ($H = 1$), and δ_H is the size of the effect. We take $\delta_0 = 10, \delta_1 = 2$ in the base model for each homogeneous and heterogeneous cases. X_j parameters are again the j most influential features in the empirical data as obtained by SHAP values.

3.2 Assessment of CF performance in Baseline Model

In the following analysis on parameter robustness of the model, we omit CATE estimation due to computational constraints. However, we are able to test the accuracy of the causal forest CATE estimate based on the ground truth of the DGP. We pick two covariates that are each responsible and not responsible for treatment effect heterogeneity in the baseline model, and see if the model accurately predicts the CATE. By doing this, we aim to see that the model accurately captures heterogeneity or lack thereof when the covariate of interest is significant/insignificant.

The covariate driving treatment effect heterogeneity was selected to be race of the respondent, chosen from the important features based on our DGP specification. On the other hand, the covariate less responsible for treatment effect heterogeneity in the sample was selected to be years of education received. For each covariate, we estimate the CATE across 10 bins of the variable on a sample size of 2000 and 60000 to check for convergence of our estimates.

Figures (4) and (5) demonstrate the convergence of estimates to the true CATE as sample size increases. It is clear that the CATE estimates conditional on race and education seemingly converge and the estimates are smoothed out asymptotically. Table (1) details our results, and here it is clear that the CF estimates are more biased with when the covariate on which CATE is estimated on is not actually a significant source of treatment effect heterogeneity. We also observe our overall interval lengths to be larger, indicating reduced precision in our CATE estimates. This suggests that CF may yield relatively more biased

Table 1: CF Performance on Estimating CATEs for Covariates Responsible/Not Responsible for Treatment Effect Heterogeneity in Baseline Model (Sample Size 60000)

	Bins	True ATE	Estimated ATE	Absolute Bias	Relative Bias	Int. Length	Relative Int. Length	Coverage
Race	**1**	4926.555939	5009.312198	82.756259	0.016798	469.196415	0.095238	1
	2	5142.082670	5179.515633	37.432963	0.007280	468.244818	0.091061	1
	3	5360.951560	5329.162879	31.788681	0.005930	421.547511	0.078633	1
	4	5353.864625	5357.187781	3.323156	0.000621	462.124527	0.086316	1
	5	5366.344329	5414.664407	48.320078	0.009004	434.799978	0.081023	1
	6	5480.692735	5488.121219	7.428484	0.001355	449.521619	0.082019	1
	7	5420.216131	5447.816893	27.600762	0.005092	379.287211	0.069976	1
	8	5662.928104	5635.069118	27.858986	0.004920	433.594357	0.076567	1
	9	5785.191579	5693.767873	91.423706	0.015803	428.865523	0.074132	1
	10	6059.622143	5897.952609	161.669534	0.026680	469.932142	0.077551	1
Educ	**1**	4008.051589	4396.320602	-388.269013	-0.096872	1288.243497	0.321414	1
	2	4903.381084	4226.946009	676.435075	0.137953	1359.802736	0.277319	1
	3	4903.101215	4197.434466	705.666749	0.143923	1234.965926	0.251874	0
	4	5120.469566	4394.805763	725.663803	0.141718	1260.452777	0.246160	0
	5	4795.588912	4619.913145	175.675767	0.036633	1183.145838	0.246715	1
	6	5335.190920	5313.454736	21.736183	0.004074	1439.000123	0.269719	1
	7	5300.574686	5446.060638	-145.485952	-0.027447	1367.322118	0.257957	1
	8	6519.749205	6276.318752	243.430452	0.037337	1673.917805	0.256746	1
	9	6805.388626	6143.575069	661.813557	0.097248	1493.331095	0.219434	1
	10	7310.980441	6712.932172	598.048269	0.081801	1553.484575	0.212486	1

Relative values calculated as percentage of estimated value compared to true population ATE

and imprecise results when the covariate of interest is not actually the source of treatment effect heterogeneity.

3.3 DGP parameter modification and expectations of CF performance

Based on the synthetic DGP above, we test the causal forest model for robustness by modifying the parameters. We then estimate the model $K = \mathit{50}$ times for each parameter value in a range of different parameters. We consider the following dimensions for parameter tweaking: *1) sample size, 2) non-linearity in covariates, 3) propensity score, 4) overlap, 5) degree of treatment effect heterogeneity,* and *6) number of estimators.*

3.3.1 Sample size

In order to test the asymptotic theory of the causal forest estimates, we test the model performance on multiple sample sizes, $N = \{1000, 5000, 10000\}$. We expect from standard statistics for both our bias and variance to decrease with increasing sample size, given our estimate converges to the true CATE value. Accordingly, interval length would also decrease.

3.3.2 Linearity in response surface

We tweak the degree of linearity in our covariates to check if the causal forests model is able to handle higher degree relationships and complex interactions between variables. 4 degrees of linearity are explored, specified by i in our definition of Γ in equation (1) - which corresponds to *full (i = 0), high (i=2), med (i=4), low (i=8).* As causal forests are expected to be apt at detecting non-linear relationships than traditional treatment effect estimation methods as OLS, we would expect the model to be robust to the introduction of a non-linear response surface.

3.3.3 Propensity score

In this scenario, the data has an imbalanced treatment and control group sizes. Since treatment assignment is random with the propensity score $p = \pi(X)$ being constant across all individuals we do not expect this change to introduce any selection bias. However, as the ATE estimates are obtained by taking differences in means at the terminal node, the deviation in group sizes from a 50/50 split will likely harm the power of our model and increase the variance in our estimates. We estimate treatment probabilities of $p = \{0.1, 0.5, 0.9\}$ for extreme cases of imbalance.

3.3.4 Overlap

We explore situations in which the propensity score p satisfies or fails the overlap condition: $0 < p = \pi(X) < 1$. Given that in traditional statistical settings, overlap ensures there are observations on which we can estimate credible counterfactuals, we expect our causal forest estimate variances to increase without the condition. When overlap does not hold, we randomly sample half of the observations and force $T = 0$ to ensure that half the sample always has a 0% probability of receiving treatment.

3.3.5 Degree of treatment effect heterogeneity

Causal forest optimizes sample-split by preferring leaves with heterogeneity in a key parameter and penalizing those with greater variance [2]. The model is expected to show stable performance across the presence of complex heterogeneous effects. For this reason, we decide to alter the degree of treatment effect heterogeneity in the model, $\Theta(X)$, to check if the causal forest estimates are robust under such changes. We specify 4 different parameter values of $j = \{0, 2, 4, 8\}$ for j in equation (2).

3.3.6 Number of estimators

We vary the number of trees we fit in our causal forest estimator to see if our estimates are sensitive to tuning parameters. The baseline model had 1000 trees in each forest, and we expect with increasing number of trees we will be able to reduce overfitting and the importance of each tree in the estimate. Naturally, we believe the increase in this parameter will also show an increase in bias for CATE as our estimates are averaged out over multiple trees. However, this may not be as pronounced in the ATE estimate, which takes the average across the entire sample. We take the *number of trees* = {100, 500, 1000, 5000} in each forest.

3.4 Results and discussion

Figure (6-a) and Table (2-a) illustrate the parameter resilience of causal forest ATE estimates for our specified DGP across the six parameters we have varied. We confirm that both bias and variance of our CF estimates improve with increasing sample size. The concave trend of the bias and variance also suggest that increasing sample size does improve CF performance, yet the improvement in the model marginally diminishes. Note coverage is 1 for all values of N we chose, indicating that the CF obtains asymptotic results pretty quickly even with relatively smaller sample sizes.

Our test of CF robustness across varying degrees of linearity in the response surface, on the other hand, reveals that the CF is relatively stable across non-linear situations, but breaks down in situations with extremely low linearity. Figure (6-b) and Table (2-b) demonstrates this, and we observe the bias and variance of the estimates dramatically increase when the degree of linearity is 'low' ($i = 8$). In this case, we have a total of 156 high-dimensional interaction terms added to the Γ given the baseline order of $\omega = 3$. With the presence of such extremely high degrees of non-linearity, the bias and variance of the estimate exponentially increases, contrary to our initial expectations.

The analysis of model resilience across imbalanced treatment and control group size behaves as initially predicted: figure (6-c) and table (2-c) show bias and RMSE is minimized at a 50/50 split between control and treatment. An interesting result is that with significant differences in group size, the CF performs better when we have a 10/90 split treatment-control than when there is a 90/10 split. Observing very few treatment cases yields more favorable results than observing very few control group cases, perhaps as it yields more conservatives estimates of the treatment effect. We also note that in situations with such extreme imbalance in group sizes, it may be challenging to estimate heterogeneous treatment effects for the lack of data points we can estimate adequate counterfactuals on. Figure (6-d) and table (2-d) also shows similar results as modifying the propensity score, since the failure of the overlap condition forces some individuals to have 0% change of receiving treatment, and thus creates and mismatch in the treatment-control group size. As expected, the violation of the overlap assumption does not significantly worsen CF performance, although there is some increase in bias, RMSE, and the interval length can be noted.

Testing increasing degrees of treatment effect heterogeneity in the model yields promising results for the CF, as performance is significantly improved with some degree of heterogeneity compared to complete homogeneity.

4 Conclusion

WRITESOME SHIT

Table 2: Model Performance with Parameter Variation

(a) Sample Size

N	True ATE	Estimated ATE	Absolute Bias	Relative Bias	RMSE	Relative RMSE	Int. Length	Relative Int. Length	Coverage
1000	5439.607849	5284.487310	173.520082	0.031899	202.689775	0.037262	1496.219129	0.275060	1.0
5000	5480.217592	5428.476930	53.144836	0.009698	60.950723	0.011122	885.568983	0.161594	1.0
10000	5464.342620	5437.754914	27.367210	0.005008	30.277228	0.005541	706.760847	0.129341	1.0

Relative values calculated as percentage of estimated value compared to true population ATE

(b) Linearity

i	True ATE	Estimated ATE	Absolute Bias	Relative Bias	RMSE	Relative RMSE	Relative Int. Length	Int. Length	Coverage
full	5468.773915	5440.039188	36.353349	0.006647	45.801340	0.008375	1.358807e+03	0.248466	1.0
high	5453.578001	5417.485903	38.292994	0.007022	47.719423	0.008750	1.289210e+03	0.236397	1.0
med	5535.015621	5489.769929	45.867383	0.008287	56.011837	0.010120	8.908016e+02	0.160939	1.0
low	5455.065609	11248.379766	612400.414644	112.262704	813537.503514	149.134321	3.958072e+07	7255.772887	1.0

Relative values calculated as percentage of estimated value compared to true population ATE

(c) Propensity Score

p	True ATE	Estimated ATE	Absolute Bias	Relative Bias	RMSE	Relative RMSE	Relative Int. Length	Int. Length	Coverage
0.1	5490.340503	5427.326138	132.827121	0.024193	179.524301	0.032698	1825.641746	0.332519	1.0
0.5	5481.328925	5449.166690	36.238026	0.006611	45.833821	0.008362	881.562651	0.160830	1.0
0.9	5469.355627	5387.042317	182.914155	0.033443	222.873837	0.040750	1803.281973	0.329706	1.0

Relative values calculated as percentage of estimated value compared to true population ATE

(d) Overlap

overlap	True ATE	Estimated ATE	Absolute Bias	Relative Bias	RMSE	Relative RMSE	Relative Int. Length	Int. Length	Coverage
True	5453.466708	5403.129840	50.684381	0.009294	60.060401	0.011013	883.900621	0.162081	1.0
False	5456.466230	5404.707229	102.429344	0.018772	128.709960	0.023589	1075.963858	0.197191	1.0

Relative values calculated as percentage of estimated value compared to true population ATE

(e) Degree of Heterogeneity

j	True ATE	Estimated ATE	Absolute Bias	Relative Bias	RMSE	Relative RMSE	Relative Int. Length	Int. Length	Coverage
0	1.000000e+01	9.826431e+00	6.307467e-01	0.063075	8.345605e-01	0.083456	4.932942e+01	4.932942	1.0
2	7.931915e+02	7.874031e+02	6.198745e+00	0.007815	7.912705e+00	0.009976	1.647809e+02	0.207744	1.0
4	5.458236e+03	5.407529e+03	5.270965e+01	0.009657	6.209724e+01	0.011377	8.760214e+02	0.160495	1.0
8	1.262573e+10	1.256825e+10	1.283857e+08	0.010169	1.695761e+08	0.013431	3.384565e+09	0.268069	1.0

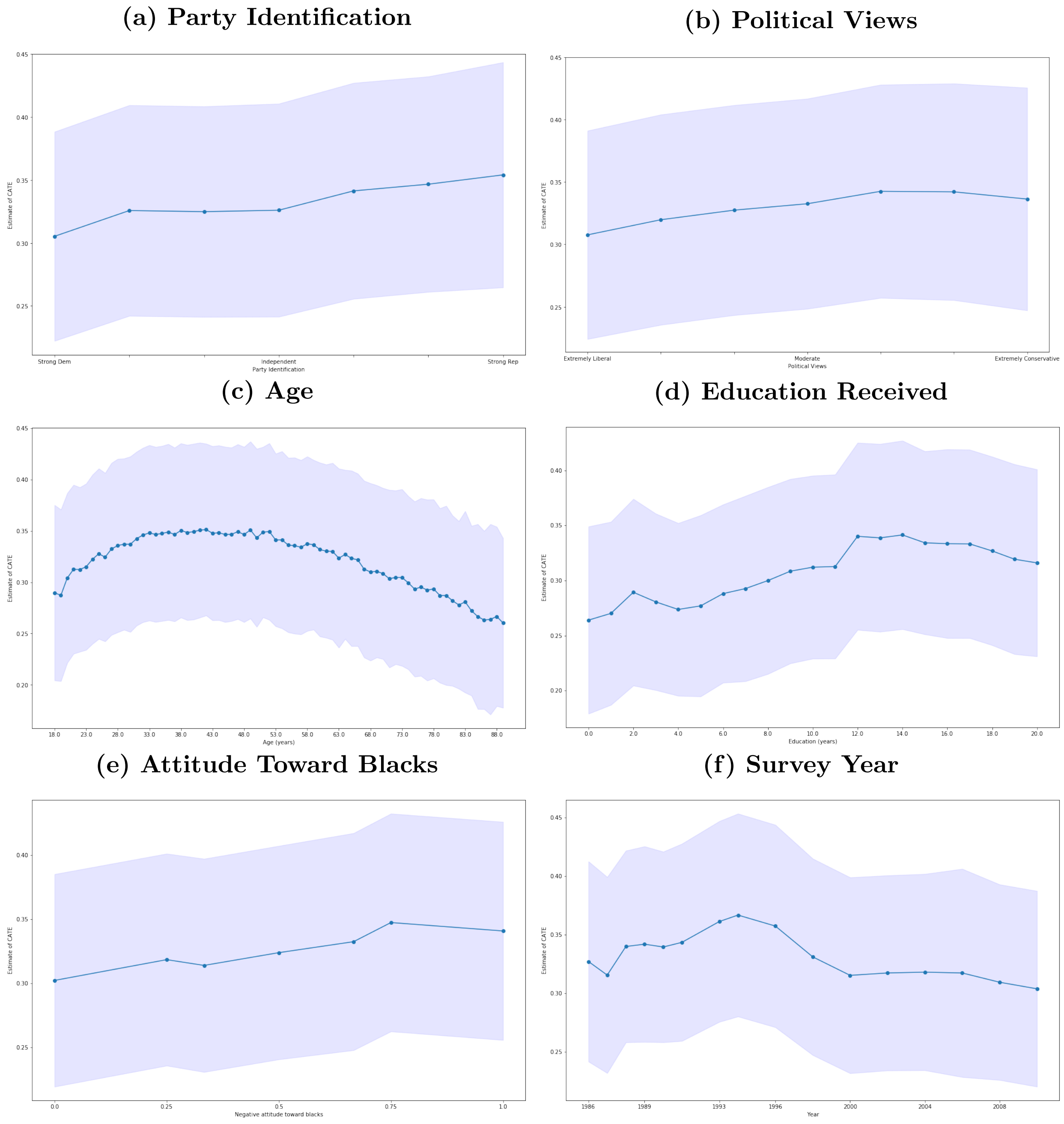
Relative values calculated as percentage of estimated value compared to true population ATE

(f) Number of Estimators

Num_Estimator	True ATE	Estimated ATE	Absolute Bias	Relative Bias	RMSE	Relative RMSE	Relative Int. Length	Int. Length	Coverage
100	5565.936961	5532.114094	33.822867	0.006077	33.822867	0.006077	1042.320012	0.187268	1.0
500	5523.942920	5449.390716	74.552204	0.013496	74.552204	0.013496	830.519095	0.150349	1.0
1000	5474.819893	5450.791857	24.028036	0.004389	24.028036	0.004389	819.989987	0.149775	1.0
5000	5361.625270	5287.516977	74.108293	0.013822	74.108293	0.013822	840.772017	0.156813	1.0

Relative values calculated as percentage of estimated value compared to true population ATE

Figure 1: CATE Estimates by Covariate



(a) Party Identification

(b) Political Views

(c) Age

(d) Education Received

(e) Attitude Toward Blacks

(f) Survey Year

Figure 2: Histogram of CATE Estimates

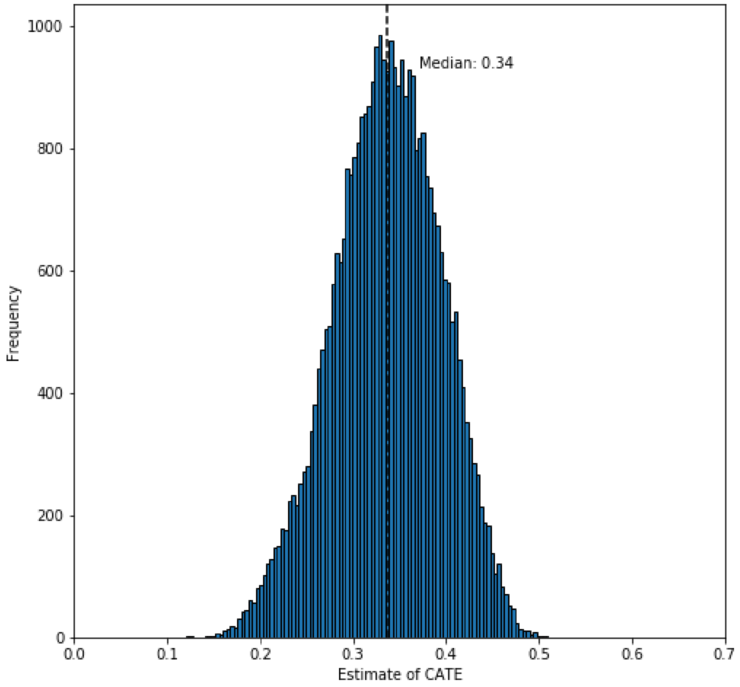


Figure 3: Summary of SHAP Values of Covariate Importance

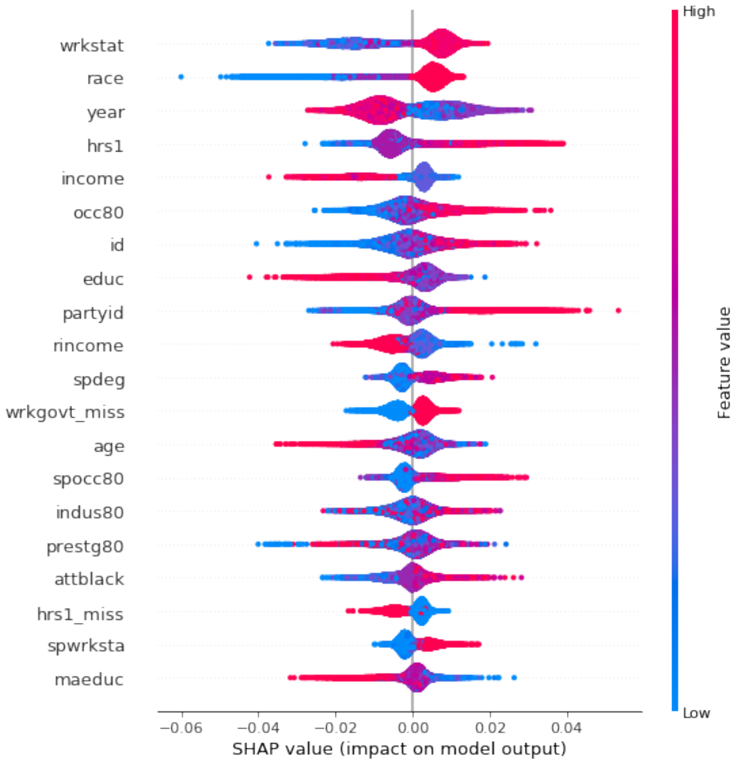
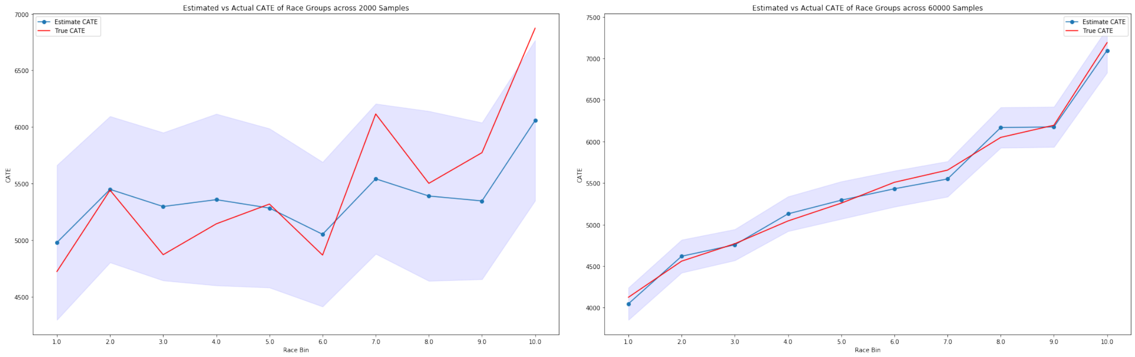


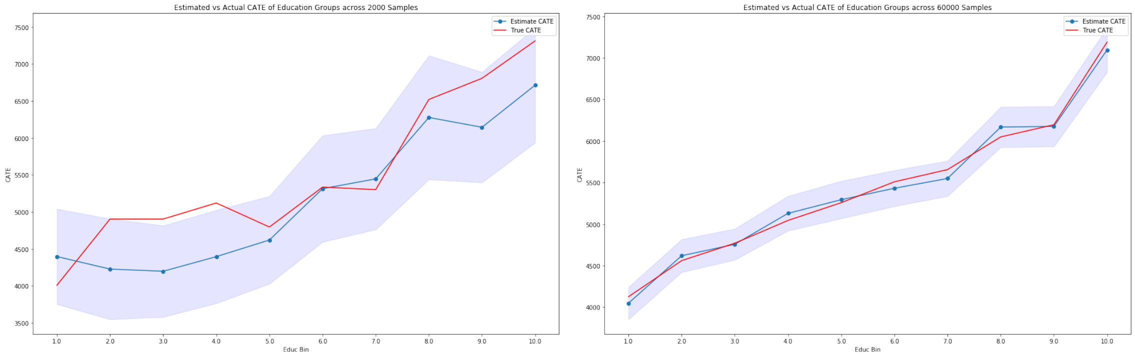
Figure 4: Estimated vs. Actual CATE across Race Groups



(a) 2000 Samples

(b) 60000 Samples

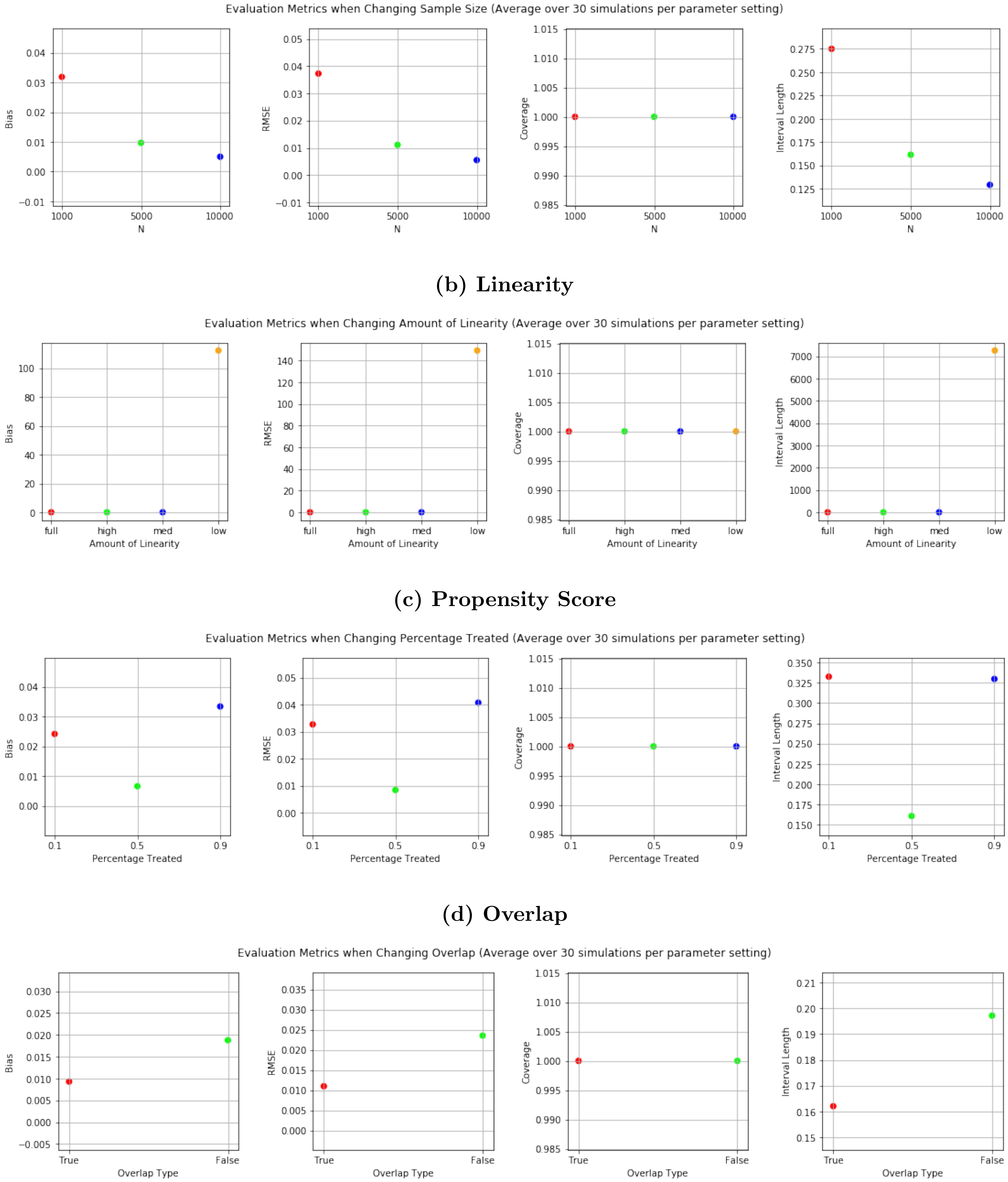
Figure 5: Estimated vs. Actual CATE across Education Level



(a) 2000 Samples

(b) 60000 Samples

Figure 6: Relative Model Performance by Parameter Variation



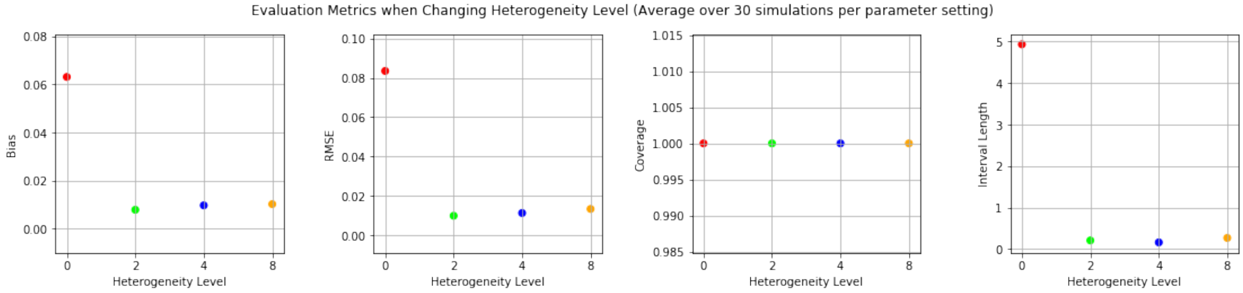
(a) Sample Size

(b) Linearity

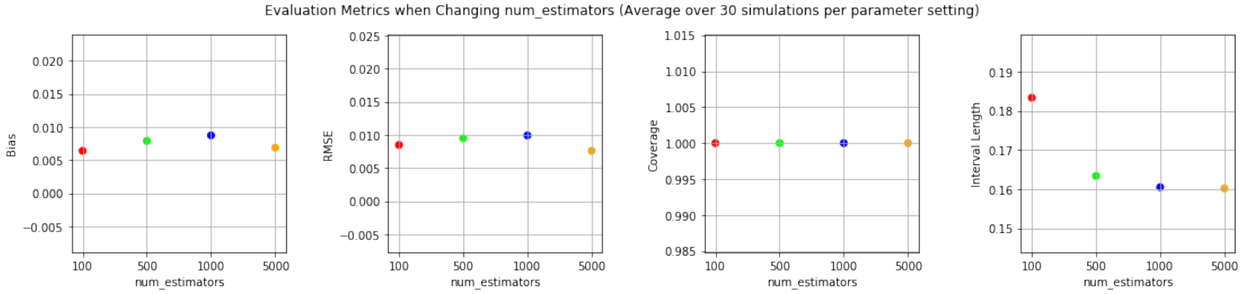
(c) Propensity Score

(d) Overlap

(e) Degree of Heterogeneity



(f) Number of Estimators



References

[1] S. Athey and G. Imbens. Machine learning methods economists should know about, 2019.

[2] S. Athey, J. Tibshirani, and S. Wager. Generalized random forests, 2018.

[3] S. Athey and S. Wager. Estimating treatment effects with causal forests: An application, 2019.

[4] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2014.

[5] H. A. Chipman, E. I. George, and R. E. McCulloch. Bart: Bayesian additive regression trees. 4(1), Mar 2010.

[6] D. P. Green and H. L. Kern. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *The Public Opinion Quarterly*, 76(3):491–511, 2012.

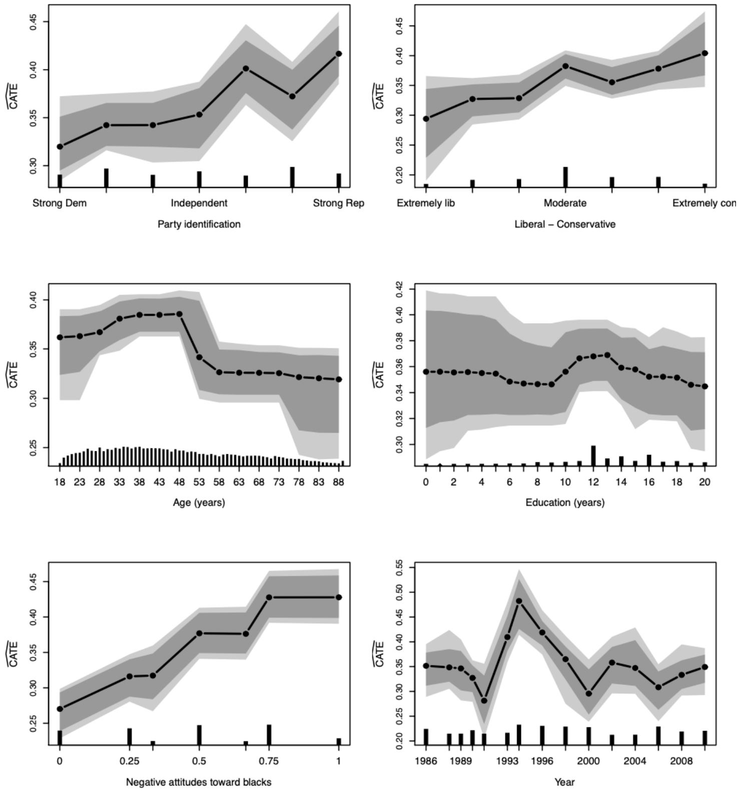
[7] P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965 – 1056, 2020.

[8] K. A. Rasinski. The effect of question wording on public support for government spending. 53(3):388–394. _eprint: https://academic.oup.com/poq/article-pdf/53/3/388/5301247/53-3-388.pdf.

[9] V. Rockova and E. Saha. On theory for bart, 2018.

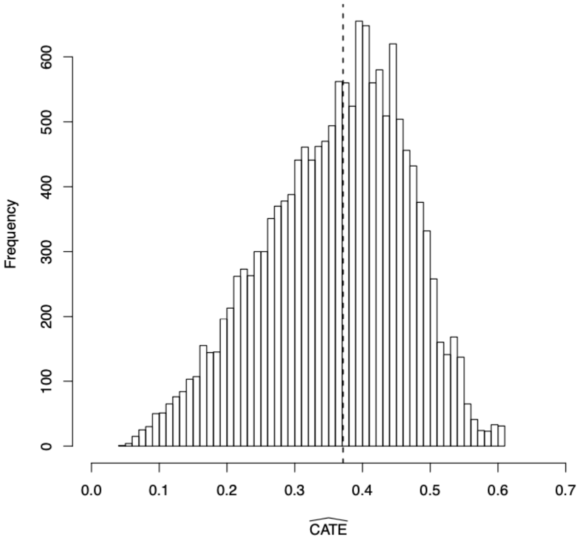
A References from Green and Kern (2012)

Figure A1: CATE Estimates Obtained by BART



Notes: Source: Green and Kern (2012). GSS 1986–2010. CATE estimates (on the probability scale) are shown. The dark grey areas are point-wise 95% posterior bands; the light grey areas are global 95% posterior bands. Marginal covariate distributions are displayed at the bottom of the graphs.

Figure A2: Histogram of CATE Estimates (BART)



Notes: Source: Green and Kern (2012). GSS 1986–2010. The graph displays a histogram of CATE estimates (on the probability scale). The vertical dashed line denotes the median CATE estimate.