Revisiting Reaction to Welfare: Exploration of Causal Forest Methods in Estimating Treatment Effect Heterogeneity and Parameter Robustness

ECMA 31330 Final Project

Spring 2021

Max Bronckers
Veronica Song
Dustin Zhang

Contents

1 Introduction **2**

2 Causal Forest Estimates of CATE in Welfare Dataset **3**
- 2.1 Discussion of data and motivation for research . . . 3
- 2.2 Causal forests and model assumptions . . . 3
- 2.3 Results and analysis . . . 4
- 2.4 Suggestions for further analysis . . . 6

3 Testing Parameter Robustness of Causal Forest Estimates on Synthetic DGP **6**
- 3.1 Baseline DGP specification . . . 6
- 3.2 Assessment of CF performance in Baseline DGP . . . 7
- 3.3 DGP parameter modification and expectations of CF performance . . . 8
  - 3.3.1 Sample size . . . 8
  - 3.3.2 Linearity in response surface . . . 9
  - 3.3.3 Propensity score . . . 9
  - 3.3.4 Overlap . . . 9
  - 3.3.5 Degree of treatment effect heterogeneity . . . 9
  - 3.3.6 Number of estimators . . . 9
- 3.4 Results and discussion . . . 10

4 Conclusion and Remarks **13**

A References from Green and Kern (2012) **20**

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 approach 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 the traditional OLS or single tree models popular in economics and is also resilient to problems with overfitting [7]. 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 an inference tool. Though recent work by Rockova 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 [10]. Causal Forest (CF) is an alternative method, proposed by Wager and Athey (2017). One of the advantages of CF is that its 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 it would allow decision makers to better assess policy impact [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 (GSS) as used by Green and Kern. To find the impact of the question phrasing on the responses, Green and Kern 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 benefits of CFs, we seek to evaluate CFs as an alternative method and apply it to the same empirical dataset and compare it to the authors' findings.

In section 2, we fit a CF model on the welfare data and compare our estimates against BART estimates of both the average treatment effect (ATE) and 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 qualitatively. In section 3, we remedy this shortcoming with an evaluation of the CF method applied to multiple 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 (as used 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' [9]. The treatment indicator T_i is 1 if the survey question presented to the respondent frames the government spending as for the 'poor' and 0 if the survey question frames it as 'welfare'. The outcome variable Y_i is 1 if the respondent's answer supports government spending in response to the survey question and 0 if the respondent's answer does not support it.

Using BART, Green and Kern seek to estimate the extent to which the reaction to the question wording of 'welfare" varies depending on the respondent's background characteristics, including years of education, race, or political alignment. BART has been a popular choice for heterogeneous treatment effect modeling as it assumes no specification of the treatment effect unlike parametric methods, and requires little parameter tuning. BART is able to learn complex, high-dimensional relationships from the data and detect interactions between covariates. Its probabilistic nature prevents overfitting as each individual tree in the forest has only a small effect on the model by assuming a prior distribution over the tree parameters [5]. The MCMC algorithm is used to sample tree parameters iteratively from the posterior distribution as the model is fit. We refer to [5] for more computational and theoretical details on BART. Though the prior distribution and back-fitting algorithm allow BART to be relatively invariant across its parameters, in the presence of confounding variables and treatment effect heterogeneity, such regularization may severely bias the treatment effect estimates [8]. 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 seek to investigate CFs as an alternative method to BART in treatment effect heterogeneity estimation. Just like BART, CF is able to model highly non-linear relationships and interactions between covariates. One advantage that CFs have over BART is that under weak assumptions, the estimates are asymptotically standard normal distributed with Gaussian confidence intervals [3]. In the context of economic policy making, the presence of an asymptotic theory allows for hypothesis testing on treatment effects–aiding policy decision making.

CFs 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 CFs uncover heterogeneous treatment effects with valid confidence intervals for statistical inference, they 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 consequently use the Double/Debiased Machine Learning method (DML) for causal forests, as proposed by Chernozhukov et al. (2016), which uses orthogonalized treatment on covariates to estimate the treatment effect [4].

For our evaluation, we use the *CausalForestDML* from the `econml` package. The method fits the estimators for fitting the response to the features and the treatment to the features in a first stage cross-fitting manner, using a `WeightedLassoCV` and `LogisticRegressionCV` respectively on discrete treatment. Afterwards, it fits a forest of trees to solve a local moment equation that involves the residualization of the treatment and outcome variables. We default to using 1000 honestly-trained trees in our CF and each tree splits to maximize the pure parameter heterogeneity score, which serves as approximation to the ideal heterogeneity score as described in [2]. We also maximize the number of samples in each subsample that is used to train every tree by setting *max_samples* to 0.5. We split our data in 80/20 train/test sets, train our tree using 5-fold cross validation (CV), and estimate treatment effects on the remaining test set. For inference, the implementation uses a bootstrap-of-little-bags to calculate the parameter vector covariance. All other parameters are left to their default values.

2.3 Results and analysis

We first estimate the ATE without conditioning on any covariates for possible effect heterogeneity. For the CATEs estimation, we follow Green and Kern's assessment. They estimate CATEs with respect to seven variables to condition the treatment effects on: *party identification, political views, age, education, negative attitude towards blacks*, and *survey year*. For a complete overview of their results, please refer to our appendix.

We find an estimated ATE of 0.336, with a standard deviation of 0.049 and a 95% confidence interval (CI) of (0.256, 0.416). This means that independent of the respondent's background, the framing of government spending as 'for the poor' gets estimated 33.6% greater approval visàvis framing it as public welfare spending. This is in line with Green and Kern's estimated ATE of 0.364. Since BART does not offer CIs and only posterior intervals, the presented ATE posterior interval by Green and Kern is not directly comparable to our confidence interval.

Figure (1) shows the estimated CATEs conditional on each of the seven variables obtained by our CF model. The blue areas represent the 95% CIs of our estimates.[1] The

[1]Note that the bands of the Green and Kern results in the appendix are % posterior intervals and *not* confidence intervals.

upper two graphs of figure 1 present CATEs as a function of *party identification* and self-identified *political alignment*. Controlling for all other covariates, we see an absolute 5% and 3% percentage points difference in the CATE when conditioned on *party identification* and *political alignment*, respectively. This means that there is a 5% increase in treatment effect for strong Republicans versus strong Democrats and 3% increase for conservatives versus liberals. In other words, we see a 5% and 3% difference in the effect of question wording on support for welfare spending between strong Republicans-Democrats and Conservative-Liberals, respectively. Conservatives and strong Republicans are more likely to be negatively affected (unsupportive) by the framing of the survey question as government spending for welfare. The CATE conditional on *age* is greatest for those in their 30 to 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. The CATE across time (*survey year*) was strongest during years 1993-1996. These estimates are all in agreement with the CATE estimates presented by Green and Kern using BART.

However, our CATE estimates conditional on *education* are different from those of Green and Kern. There seems to be an increasing effect of question wording as the respondent has more years of education, with the effect peaking at around 11-13 years (i.e. a college education). Whereas Green and Kern found no distinct moderation of treatment effects based on education, we find that there is around a 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".

On average, our CF-obtained CATE estimates are lower than those using BART. The higher estimates of the BART estimates can be attributed to the regularization-induced confounding (RIC) as identified by Hahn, Murray, and Carvalho (2019) [8]. The treatment effect in BART is obtained by taking the conditional expectation of the outcome conditional on treatment and certain covariates: $E[Y|x, Z = 1] - E[Y|x, Z = 0]$. In the presence of confounding in finite sample sizes, this may be dependent mostly on x rather than the treatment Z. RIC consequently states that the BART-obtained CATEs are heavily dependent on the regularization by the prior distribution specified in finite-size samples.

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'.

To conclude, our results largely agree with the results obtained through BART and illustrate that the effect of question phrasing on government spending is highly heterogeneous and depends on the respondent's background. We also find slightly lower CATE estimates using CF as estimator relative to BART, which could be because of BART's regularization-

induced confounding.

2.4 Suggestions for further analysis

We are also able to assess covariate importance by calculating SHAP values, as indicated in figure (3), using the CF implementation of `econml`. Aside from the covariates estimated by Green and Kern, we find that the covariates *work status* and *racial backgrounds* are also significant in our fitted CF model. Although our scope is limited to comparisons of CF-obtained CATE estimates with the Green and Kern's BART-obtained estimates, it may be of interest to estimate CATE conditional on those additional covariates with CF and BART in future lines of research.

3 Testing Parameter Robustness of Causal Forest Estimates on Synthetic DGP

The analysis on the welfare data in section 2 shows the real-life application of CF as estimator for heterogeneous treatment effects; however, since empirical data offers no ground truth treatment effect, it is hard to draw conclusions from the differences in estimates between CF and BART on the welfare data and evaluate performance objectively. To remedy this, we evaluate the performance of CF on simulated data that is somewhat similar to the empirical welfare dataset. Specifically, in this section we explore multiple synthetic DGPs and evaluate how CF performs under different data generating parameters. For every parameter evaluation, we ran $K = 30$ simulations on a server with 48 Intel Xeon Silver 4214 CPUs @ 2.20GHz and 126GB of RAM memory.

3.1 Baseline DGP specification

We specify a DGP that resembles our empirical data using a simplified model of N individuals and 120 covariates. The covariates are drawn from a multivariate Gaussian distribution with means and covariance matrix obtained from the empirical data. 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 $\omega = 3, i = 4$ (defined as *med*-degree linearity). T, our treatment vector, models binary treatment assignment via $T \sim Binomial(n, p)$ where p is the probability of receiving treatment (propensity score). We assume there is a 50% probability of treatment in the baseline model. 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 DGP

In our assessment of CF as treatment effect estimator applied to our baseline DGP, we evaluate the estimated ATE and CATEs conditioned on two covariates (one involved in treatment effect heterogeneity, the other not) against their ground truth values. In doing so, we assess whether the model accurately captures the heterogeneity, or possible lack thereof, when the covariate of interest is significant/insignificant with respect to the treatment effect.

We selected the covariate of interest involved in treatment effect heterogeneity to be *race* of the respondent, chosen from the important features based on our DGP specification. We chose the covariate not responsible for treatment effect heterogeneity in the sample to be years of *education* received. For each covariate, we estimate the CATE across 10 bins of the variable on a sample size of $N = 1000$ and $N = 30\,000$ to check for convergence of our estimates. These bins evenly divided the data based on the conditioned covariate value as the simulated covariates had no inherent meaning, and could only be divided numerically. Due to computational constraints, we only ran the estimation of CATEs on the simulated data once.

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 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 when the covariate on which CATE is conditioned is not actually a significant source of treatment effect heterogeneity. We also observe our overall interval lengths to be considerably greater, indicating reduced precision in our CATE estimates. Whereas the estimated 95% CI for CATE conditioned on race contains the true CATE for all bins, the CI for CATE on education is wider and misses the true CATE for some range of covariate values (bin 3 and 4). This suggests that CF yields relatively more biased and imprecise CATEs with wider CIs when the covariate of interest is not actually the source of treatment effect heterogeneity.

Table 1: CF Performance on Estimating CATEs for Covariates Responsible/Not Responsible for Treatment Effect Heterogeneity in Baseline Model (Sample Size 30000)

	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

3.3 DGP parameter modification and expectations of CF performance

Based on the synthetic DGP above, we test the CF model for robustness by modifying the parameters. We then estimate the model $K = 30$ 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 CF 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

The CF model 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 (6a) and Table (2a) 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 (6b) and Table (2b) 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 (6c) and table (2c) 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 (6d) and table (2d) 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, creating a 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. This is promising given that the assumption of overlap is made for other causal inference models that focus on treatment effect heterogeneity, meaning CF could have a relative advantage in scenarios where overlap is violated. Again, it is important to note that the violation of overlap was done by forcing half of the observations to have no treatment while the other half was assigned treatments probabilistically as per the baseline model. It would be interesting to extend this paper to test the performance and behavior of CFs under a greater degree of overlap violation.

Testing increasing degrees of treatment effect heterogeneity in the model yields promising results for the CF: performance significantly improves in models with some degree of heterogeneity compared to completely homogeneous models. As figure (6e) and table (2e) suggest, bias and RMSE increase only slightly with degrees of heterogeneity greater than 2 (i.e. there are > 2 features that drive treatment effect heterogeneity in the model). We

see that CF is best fit to assess heterogeneous models and is robust to increasing sources of heterogeneity.

Finally, we find that the CF produces stable estimates across variations in the tuning parameter for the number of trees to use. Figure (6f) and table (2f) show that the change in bias and RMSE of the estimates when increasing *n_estimator* is small, although the interval length decays. We observe that the baseline of 1000 trees in the CF in fact yields the result with the highest bias and RMSE but the smallest interval length among the parameter values, reflecting the bias-variance trade-off. In future analyses, it may be worth to explore whether this result changes with larger sample sizes.

In conclusion, we find that CF performs best in relatively linear response surfaces with some degree of treatment effect heterogeneity. Overall performance is rather stable across different tuning parameters as well as DGP characteristics, allowing CF estimates to be reliable in many economic applications. It is promising that the CF reaches asymptotic results and shows excellent coverage and interval widths even with smaller sample sizes, given the cost of acquiring large and thorough datasets. Moreover, we also note that the model parameters of the welfare GSS data in section 2 are in alignment with the condition under which CF yields stable results. For example, the welfare GSS data contained over 30,000 observations with only 120 covariates. This implies that the CATE estimates obtained in section 2 are likely to be reasonable estimates.

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

4 Conclusion and Remarks

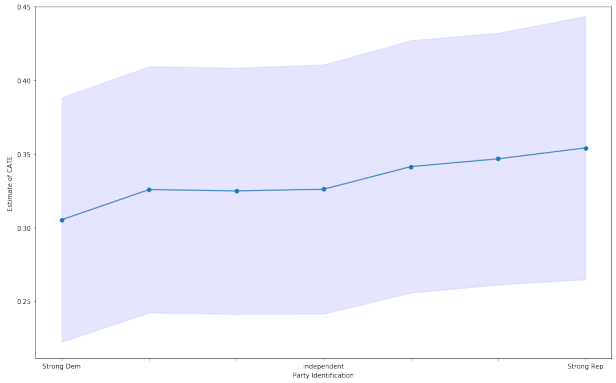
In order to investigate variation in treatment effects through interaction with individual background attributes in survey responses to welfare, Green and Kern settled on BART as their model of choice. However, the estimates provided through BART make policy implementation challenging given their theoretical shortcomings, including the lack of asymptotic theory of its estimates and adequate confidence intervals. In this research, we use DML Causal Forests (CF) to model treatment effect heterogeneity and test the robustness of the estimates across variations in tuning and DGP parameters.

Our analysis reveals that although the CF estimates are lower on average, the CF-obtained trends of the CATEs across different covariates are largely similar to those obtained by BART. We also find that on top of the six covariates identified as sources of systemic effect heterogeneity by Green and Kern, *work status* and *racial backgrounds* are also potentially significant drivers. The CATE estimates across different attributes are approximately normally distributed, confirming that we indeed have obtained asymptotic convergence in our sample and therefore a valid confidence interval for our estimates. When varying six different parameters of the DGP and the tree characteristics, we confirm that CF estimates are generally stable across diverse model specifications as well as relatively smaller sample sizes. CF clearly performs best when the covariates are largely linearly related and there is treatment effect heterogeneity present in the model. Our estimates also demonstrate excellent coverage across variations in model parameters. We notice that our model coverage significantly outperforms results suggested in existing literature [8]. Though our DGP is highly non-linear and high-dimensional, our outperformance may be due to the comparative simplicity of our DGP relative to other synthetic data (cf. the ACIC data analysis challenge) [6].

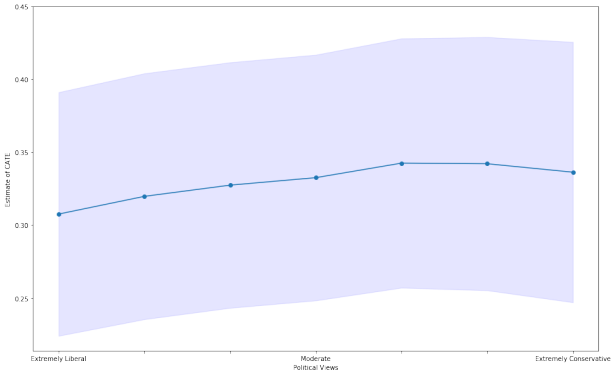
This paper provides promising results on the performance and stability of the CF method and highlights the applicability of CF as heterogeneous treatment effect estimator by evaluating the effect of question phrasing on support for welfare programs by conditioning on socio-economic factors. We still leave room for further research, such as identifying CATE based on influential features of the model previously unidentified by Green and Kern, exploring modifications to CF that may allow it to perform well in situations with extreme non-linearity, and evaluating CF's heterogeneous treatment effect estimation under highly sparse data. Moreover, comparisons between more recent versions of BART and DML CF on our synthetic DGP would also be of interest.

Figure 1: CATE Estimates by Covariate

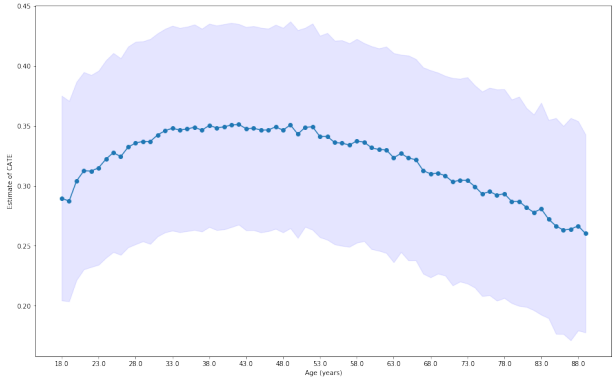
(a) Party Identification



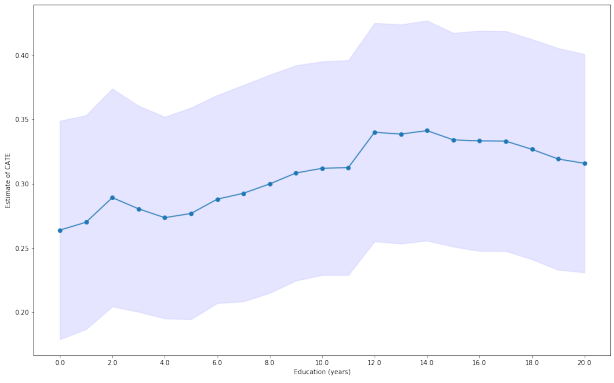
(b) Political Views



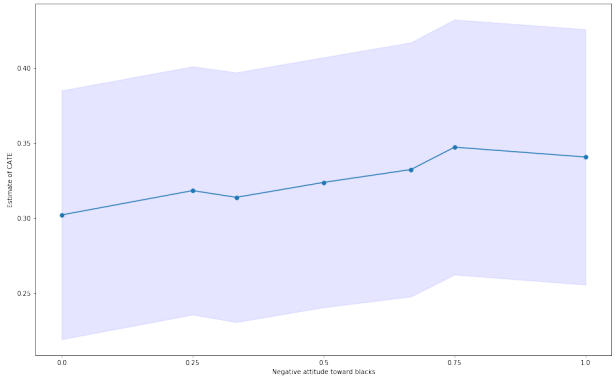
(c) Age



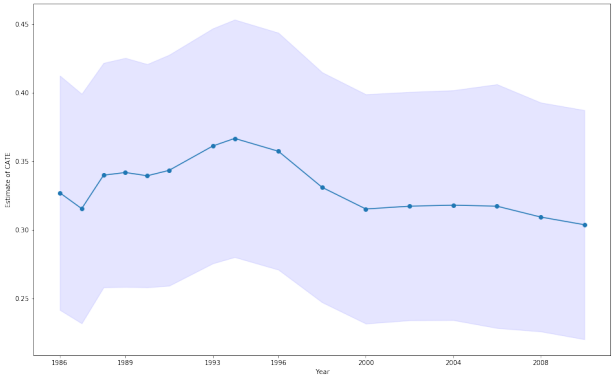
(d) Education Received



(e) Attitude Toward Blacks



(f) Survey Year



Notes: Each y-axis shows the estimated CATE conditioned on the covariate, each x-axis represents the range of covariate values for which the CATE was estimated.

Figure 2: Histogram of CATE Estimates

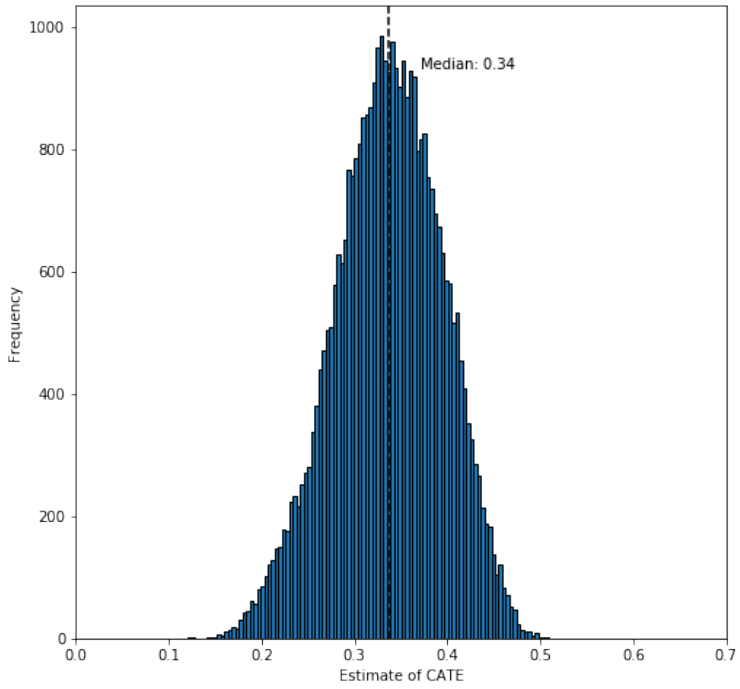


Figure 3: Summary of SHAP Values of Covariate Importance

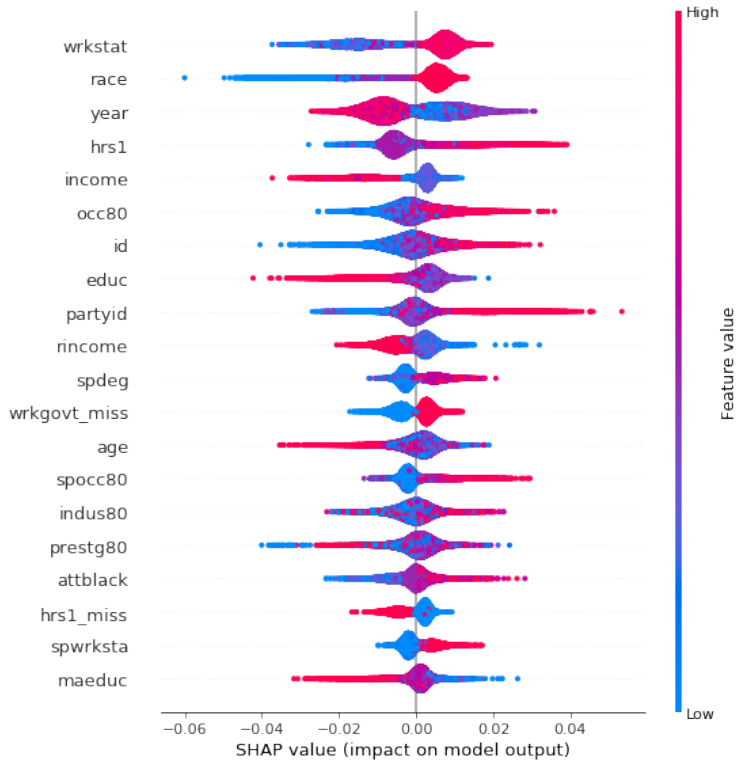
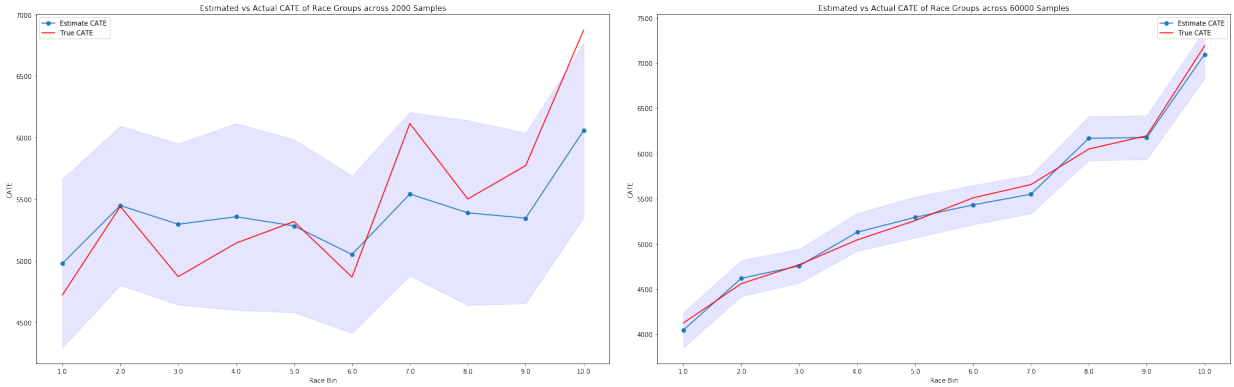


Figure 4: Estimated vs. Actual CATE across Race Groups

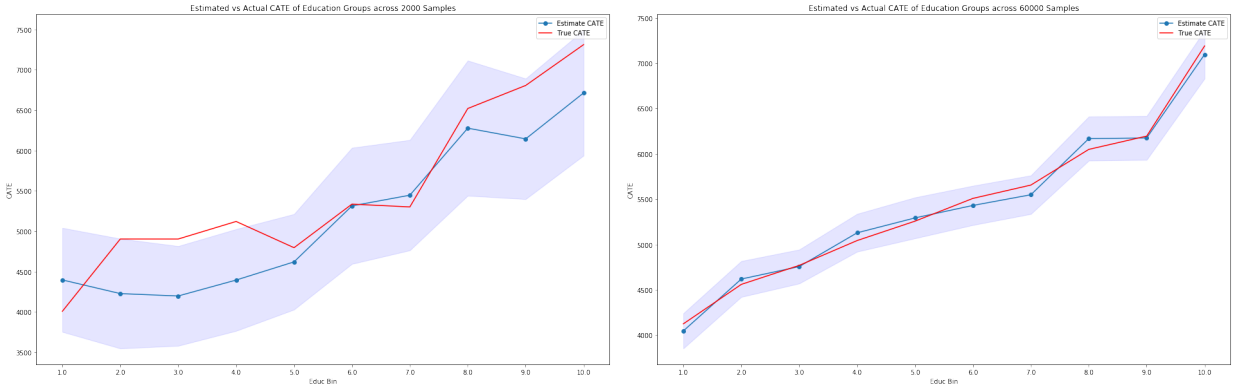
(a) 1000 Samples **(b) 30000 Samples**



Notes: y-axis is the CATE, x-axis the range of covariate values split into bins on which CATE is estimated. The red line is the true CATE and blue line is the estimate CATE with the shaded area representing the 95% CI.

Figure 5: Estimated vs. Actual CATE across Education Level

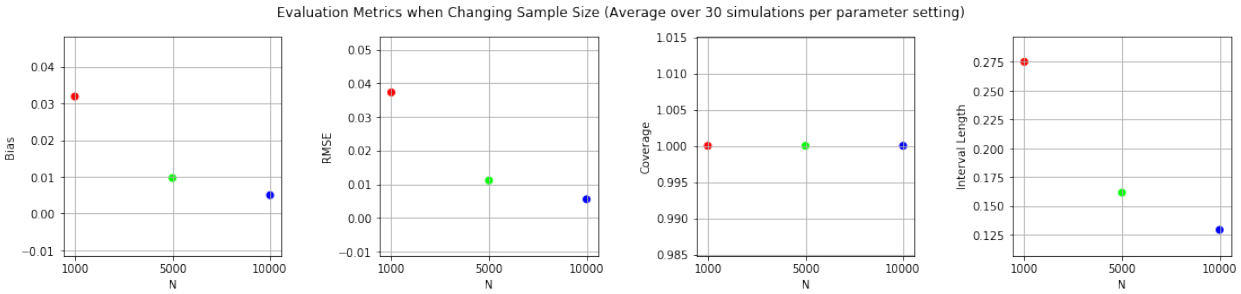
(a) 1000 Samples **(b) 30000 Samples**



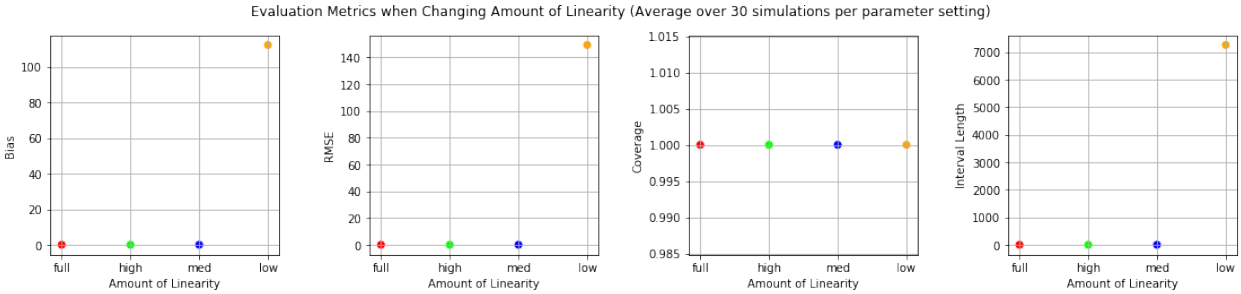
Notes: y-axis is the CATE, x-axis the range of covariate values split into bins on which CATE is estimated. The red line is the true CATE and blue line is the estimate CATE with the shaded area representing the 95% CI.

Figure 6: Relative Model Performance by Parameter Variation

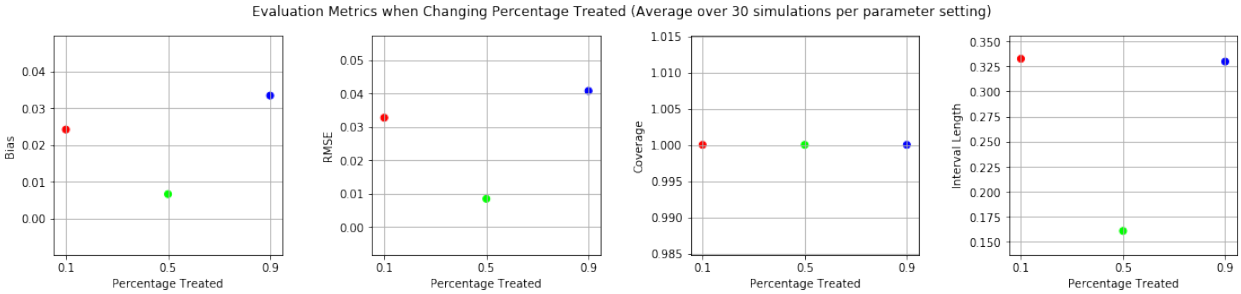
(a) Sample Size



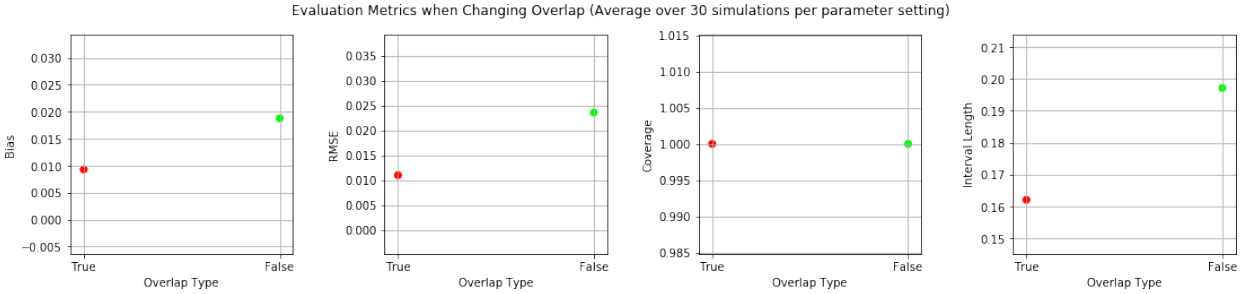
(b) Linearity



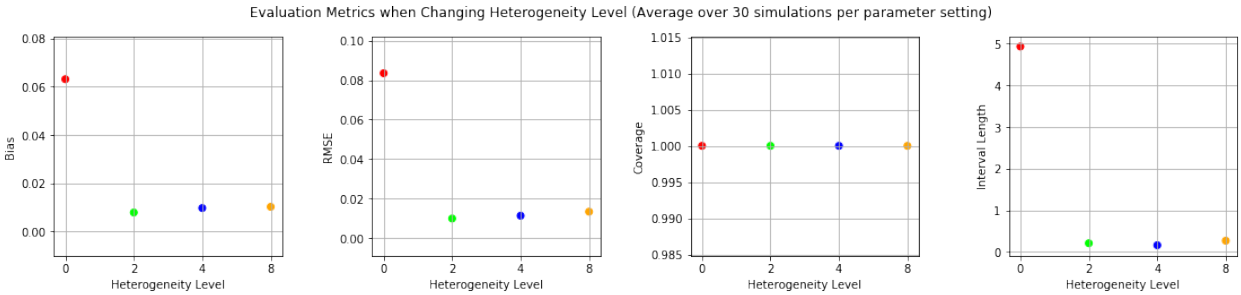
(c) Propensity Score



(d) Overlap



(e) Degree of Heterogeneity



(f) Number of Estimators

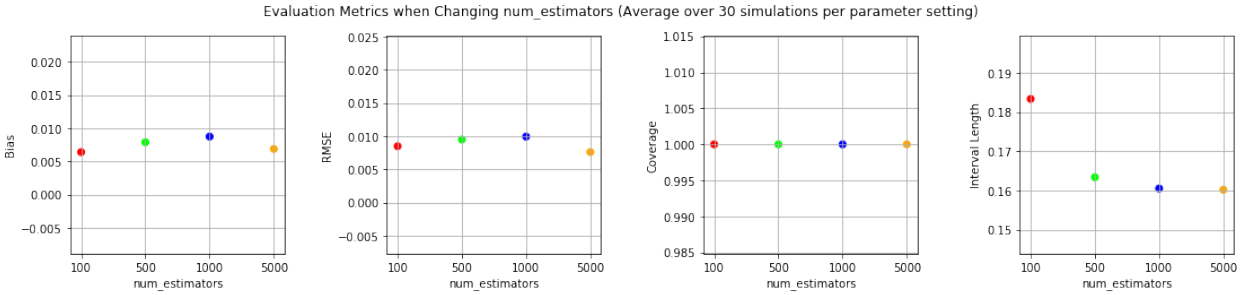
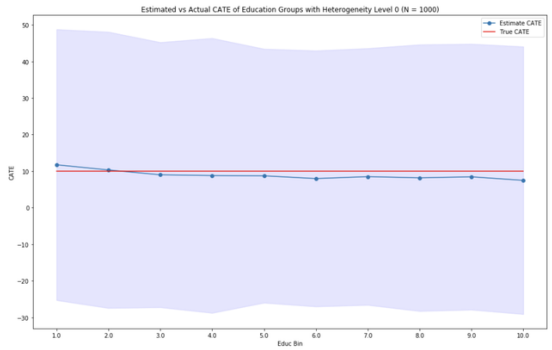
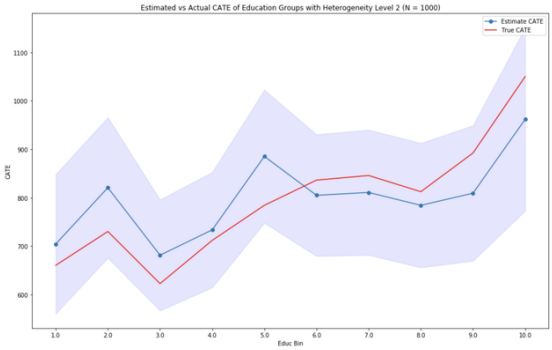


Figure 7: CATE Estimates by Variation in Degrees of Heterogeneity

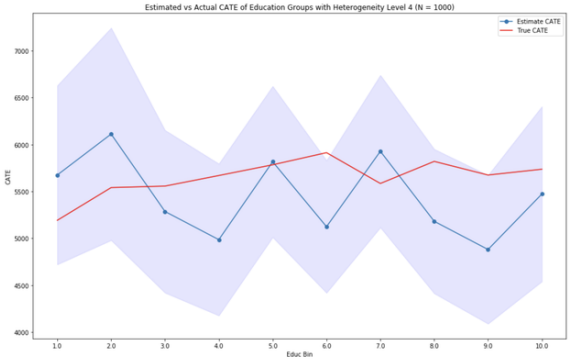
(a) $j = 0$ (Homogeneity)



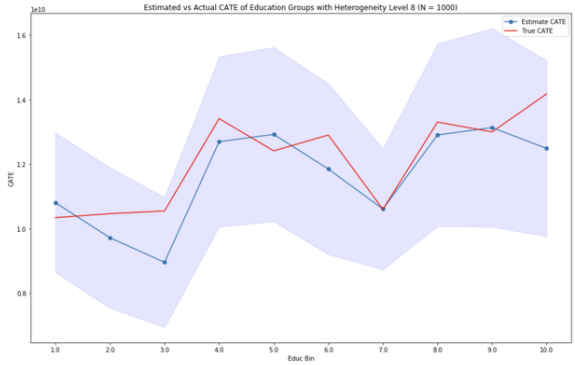
(b) $j = 2$



(c) $j = 4$



(d) $j = 8$



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] V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition, 2018.

[7] 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.

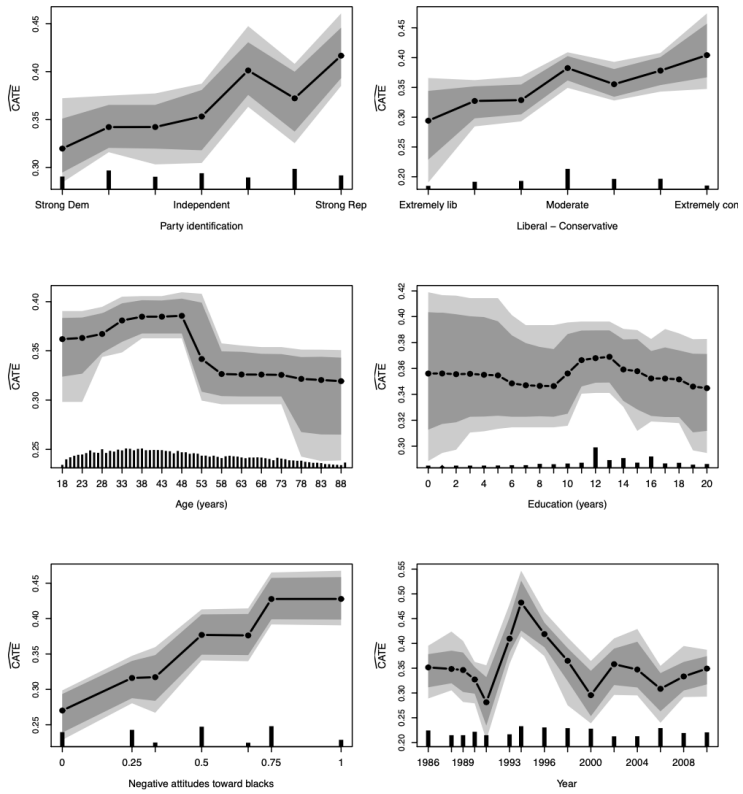
[8] P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965 – 1056, 2020.

[9] 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.

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

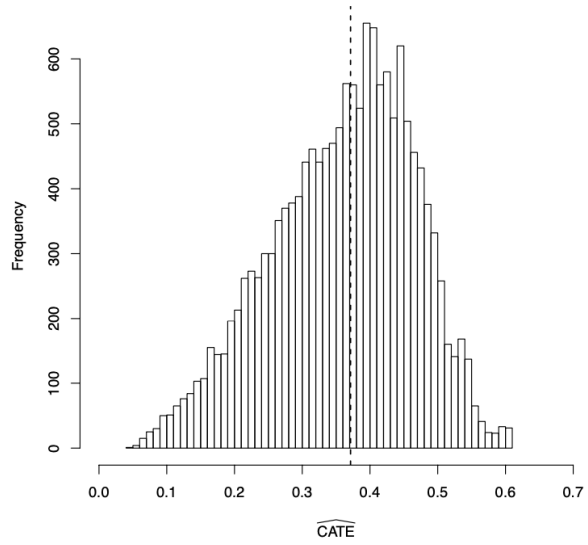
A References from Green and Kern (2012)

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.