Modeling heterogeneous treatment effects in large-scale experiments using Bayesian Additive Regression Trees

ARTICLE · JANUARY 2010

CITATIONS
5

READS
52

2 AUTHORS, INCLUDING:

Donald P. Green
Columbia University
156 PUBLICATIONS **7,715** CITATIONS

SEE PROFILE

Modeling heterogeneous treatment effects in large-scale experiments using Bayesian Additive Regression Trees

Donald P. Green[†]
Holger L. Kern[‡]

March 31, 2010

The authors are grateful to Alan Gerber, Gary King, and Scott Long for their comments, Jingjing Song for her help in reviewing the experimental literature in political science, Arjun Shenoy for his help in reviewing the literature on public support for welfare, Lynn Vavreck for sharing data, and the facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center for their support.
[†]Institution for Social and Policy Studies and Department of Political Science, Yale University.
[‡]Institution for Social and Policy Studies, Yale University.

Abstract: We present a methodology that largely automates the search for systematic treatment effect heterogeneity in large-scale experiments. We introduce a nonparametric estimator developed in statistical learning, Bayesian Additive Regression Trees (BART), to model treatment effects that vary as a function of covariates. BART has several advantages over commonly employed parametric modeling strategies, in particular its ability to automatically detect and model relevant treatment-covariate interactions in a flexible manner. To increase the reliability and credibility of the resulting conditional treatment effect estimates, we suggest the use of a split sample design. The data are randomly divided into two equally-sized parts, with the first part used to explore treatment effect heterogeneity and the second part used to confirm the results. This approach permits a relatively unstructured data-driven exploration of treatment effect heterogeneity while avoiding charges of data dredging and mitigating multiple comparison problems. We illustrate the value of our approach by offering two empirical examples, a survey experiment on Americans support for social welfare spending and a voter mobilization field experiment. In both applications, BART provides robust insights into the nature of systematic treatment effect heterogeneity.

1. Introduction

Over the last decade, increased attention to issues of identification has led to rapid growth in the number of randomized field and survey experiments conducted in political science (Druckman et al. 2006). In many of these randomized experiments the main quantity of interest is the average treatment effect (ATE). The ATE is easy to estimate: one simply compares the mean outcomes in the treatment and control groups. However, it only provides a partial answer to the causal question of interest when treatment effects vary across experimental units. When treatment effects are heterogeneous, the ATE could be positive even when the treatment has negative effects on some experimental units. In fact, because of the mean's well-known vulnerability to gross outliers, the ATE could be positive even when the treatment has negative effects on all units but one.

Clearly, the usefulness of the ATE (and other common estimands such as the average treatment effect among the treated, ATT) as a summary of the distribution of treatment effects depends on the extent and form of treatment effect heterogeneity in any given experiment (Cox 1958; see also Gelman 2004 and Horiuchi, Imai, and Taniguchi 2007). The challenge is to devise a workable method for gauging treatment effect heterogeneity. Randomized experiments only identify the two marginal outcome distributions in the treatment and control groups, which are generally insufficient to estimate the joint distribution of outcomes. One would need this joint distribution to identify aspects of the treatment effect distribution other than its mean.

To date, political scientists have paid little attention to the methodological challenges posed by treatment effect heterogeneity (for exceptions, see Horiuchi, Imai, and Taniguchi 2007; Feller and Holmes 2009, and Imai and Strauss 2009). This stands in marked contrast to the extensive methodological literature on treatment effect heterogeneity in statistics (e.g., Byar 1985; Dixon and Simon 1991; Follman and Proschan 1999; Pocock et al. 2002; Royston and Sauerbrei 2004; Rothwell 2005) and econometrics (e.g., Heckman, Smith, and Clements 1997; Abadie 2003; Poirier and Tobias 2003; Angrist 2004; Dehejia 2005; Bitler, Gelbach, and Hoynes 2006, 2009; Abbring and Heckman 2007; Crump et al. 2008; Djeb-

bari and Smith 2008). Lack of interest in treatment effect heterogeneity among political scientists cannot explain this lacuna. Between 2005–09, the *American Political Science Review*, the *American Journal of Political Science*, *International Organization*, and *World Politics* published 21 articles that reported causal effect estimates from randomized laboratory, field, or survey experiments. Fully two thirds (14) of these articles tested at least one hypothesis related to treatment effect heterogeneity.

The search for treatment effect heterogeneity raises a number of important methodological issues such as the use and misuse of conditional average treatment effects (CATEs), or average treatment effects among subgroups defined by baseline covariates such as gender or age. When experiments are not explicitly designed to estimate pre-specified CATEs, reported treatment-covariate interactions generally have limited credibility (Pocock 2002; Gabler et al. 2009). The concern is that after an exhaustive search for treatment-covariate interactions, researchers might only report the "most interesting" findings, leading to measures of statistical uncertainty that are overly optimistic. Rothwell (2005: 181) for example likens such selective reporting of statistically significant treatment-covariate interactions to "placing a bet on a horse after watching the race."

The econometric literature has proposed several other ways of dealing with treatment effect heterogeneity besides CATEs. Under strong and rather arbitrary assumptions such as perfect positive dependence between the two outcome distributions or independence between the distribution of untreated outcomes and the distribution of treatment effects, the whole distribution of treatment effects can be identified (Heckman, Smith, and Clements 1997; Abbring and Heckman 2007; Djebbari and Smith 2008). We do not pursue this solution here because the required assumptions often lack theoretical motivation. It is also possible to use the marginal outcome distributions in the treatment and control groups to place non-parametric bounds on their joint distribution. These bounds are generally wide but sometimes informative enough to rule out constant treatment effects (Heckman, Smith, and Clements 1997; Abbring and Heckman 2007; Djebbari and Smith 2008).

The aim of our paper is to deal with treatment effect heterogeneity by estimating

conditional average treatment effects. Our approach differs in two respects from current practice. First, we non-parametrically model CATEs using Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch 2008). BART has several advantages over commonly employed modeling strategies, in particular its ability to automatically detect and model relevant treatment effect heterogeneity. Second, we clearly distinguish between an exploratory data analysis phase and a confirmatory phase. We randomly select half of our observations for exploration and reserve the other half for confirmation, which allows us to circumvent the dangers of data dredging and mitigates multiple comparison problems. Two empirical examples illustrate the utility of our approach. We apply BART to two large-scale experiments, a survey experiment and a field experiment, and demonstrate how it can be used to detect and model treatment effect heterogeneity in a principled and flexible manner.

1.1. Potential outcomes framework

We rely on the potential outcomes framework for causal inference (Rubin 1974, 1978; Holland 1986). In the case of a binary treatment, unit i has two potential outcomes, denoted $Y_i(1)$ for the outcome under treatment and $Y_i(0)$ for the outcome under control. We only observe one of these two potential outcomes for each unit, depending on whether the unit receives the treatment or the control. The other potential outcome is counterfactual, which creates a missing data problem that Holland (1986) has characterized as the "fundamental problem of causal inference." We define a treatment indicator D_i which takes the value 1 if unit i receives the treatment and the value 0 if i receives the control. This allows us to write the observed outcome as $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. The effect of the treatment on unit i is defined as $\beta_{Di} = Y_i(1) - Y_i(0)$. The literature generally focuses on two different parameters of interest, the average treatment effect (ATE), $\mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(\beta_{Di})$, and the average treatment effect on the treated (ATT), $\mathbb{E}(Y_i(1) - Y_i(0) \mid D_i = 1) = \mathbb{E}(\beta_{Di} \mid D_i = 1)$.

As in Djebbari and Smith (2008), we can write the observed outcome in form of a

regression model,

$$\begin{aligned} Y_i &= \beta_0 + \beta_{Di} D_i + \epsilon_i \quad (1) \\ &= \beta_0 + \beta_D D_i + [(\beta_{Di} - \beta_D) D_i + \epsilon_i], \quad (2) \end{aligned}$$

where $\beta_0 = \mathbb{E}(Y(0))$ and $\beta_D = \mathbb{E}(\beta_{Di} \mid D_i = 1)$. In (2), the composite error term in square brackets contains the idiosyncratic effect of the treatment (for treated units) and the idiosyncratic component of the outcome under control. Adding baseline covariates to the model while allowing both the outcome under control and the causal effect of the treatment to vary with these covariates yields

$$Y_i = \beta_0 + \beta_X X_i + (\beta_D + \beta_{DX} X_i) D_i + [(\beta_{Di} - \beta_D - \beta_{DX} X_i) D_i + \epsilon_i], \quad (3)$$

where X_i is a vector of baseline covariates. This formulation distinguishes between two components that together comprise the causal effect for unit i. The first component $(\beta_D + \beta_{DX} X_i)$ is a *systematic* component that describes how the causal effect varies with baseline covariate values. The second component $(\beta_{Di} - \beta_D - \beta_{DX} X_i)$ is an *idiosyncratic* component not explainable in terms of observed covariates.

The assumption that the treatment effect is the same for all units, $\beta_{Di} = \beta_D$, is very restrictive. In this case the regression model in (3) simplifies to

$$Y_i = \beta_0 + \beta_D D_i + \epsilon_i. \quad (4)$$

A more general regression model assumes $\beta_{Di} = \beta_D + \beta_{DX} X_i$, i.e., treatment effects potentially vary across covariate values but are invariant within strata formed by these covariate values. The regression model in (4) then becomes

$$Y_i = \beta_0 + \beta_{DX} X_i + (\beta_D + \beta_{DX} X_i) D_i + \epsilon_i. \quad (5)$$

The model in (5) is significantly more flexible than (4), especially when a rich set of baseline covariates is available. Assuming (5), we allow for systematic heterogeneity but not idiosyncratic heterogeneity in treatment effects. Note that the existence of idiosyncratic

heterogeneity does not imply that CATE estimates are biased. Just as the ATE is still unbiased in the presence of treatment effect heterogeneity, CATE estimates are still unbiased when there is idiosyncratic heterogeneity in treatment effects. In this case CATE estimates simply fail to provide a complete picture of treatment effect heterogeneity.

Equation (5) suggests that we can estimate CATEs by estimating a regression model that includes a dummy variable for the treatment (D_i), unit-level covariates (X_i), and multiplicative interaction terms between the treatment indicator and the covariates. This strategy, however, while straightforward in principle, encounters several obstacles in practice. First, a parametric regression model imposes additional functional form assumptions about the relationship between the covariates and the outcome that are not justified by randomization. Getting these functional forms wrong may lead to biased CATE estimates (Royston and Sauerbrei 2004; Feller and Holmes 2009). Second, the inclusion of a large number of interaction terms can lead to multicollinearity problems and imprecise inferences. Third, the search for systematic variation in treatment effects can easily degenerate into data dredging when researchers search for treatment-covariate interactions until they discover "interesting" heterogeneity for some subset of experimental units. Since measures of statistical uncertainty are usually not adjusted for this specification search, standard errors and confidence intervals will understate the uncertainty of the resulting CATE estimates (Pocock et al. 2002; Gabler et al. 2009).

In order to avoid these problems we build on the recent literature on statistical learning and propose the use of Bayesian Additive Regression Trees (BART) to model systematic treatment effect heterogeneity in large-scale randomized experiments. BART is a nonparametric procedure in the sense that it does not require us to specify the functional forms linking covariates and outcomes. It also does not require us to specify the interaction terms included in the model. Instead, BART uses the data to detect relevant interactions. Finally, in order to account for the testing of a potentially large number of candidate treatment-covariate interactions, we clearly distinguish between an exploratory data analysis stage and a confirmatory stage (Tukey 1977). Before looking at the data, we randomly divide our

dataset into two equal-sized parts: a training set and a test set. We then use the training set to search for treatment-covariate interactions without having to worry about how this preliminary analysis of the data will affect our final inferences. If we do find systematic treatment effect heterogeneity in the training set we then move to the test set and attempt to confirm our preliminary results with fresh data.

The idea to use split samples to adjust for specification searches is not new (see, e.g., Cox 1975). It has even been used in previous political science research (Luskin 1990). Split samples are widely used in the statistical learning literature to address the problem of overfitting (Izenman 2008; Hastie, Tibshirani, and Friedman 2009). Heller, Rosenbaum, and Small (2009) randomly split the data from an observational study into a smaller planning sample and a larger analysis sample, using the planning sample to determine tuning parameters and the analysis sample for actual estimation. Our approach is similar to theirs, except that we split our dataset in half, using 50% of the data for exploration and the other 50% for actual estimation.

Studying treatment effects as a function of observable characteristics allows us to go beyond simple mean impacts. Randomized field or survey experiments are often quite expensive, with the largest of them costing millions of dollars. Restricting ourselves to estimating the ATE potentially leaves valuable information on the table. CATEs also provide a more detailed picture of the effects of a treatment and evidence for or against theories that predict (or fail to predict) such heterogeneity. They therefore offer additional opportunities for theory testing and theory development. Understanding treatment effect heterogeneity also allows us to be more confident about the generalizability of our results, since we can adjust predictions for new populations for observable differences in baseline characteristics between our sample and the target population. And finally, CATEs enable us to better target interventions to maximize their impact and cost-effectiveness (Imai and Strauss 2009).

2. Bayesian Additive Regression Trees (BART)

BART builds upon regression and classification tree models developed in the statistical learning literature (Morgan and Sonquist 1963; Breiman et al. 1984; Berk 2008; Izenman 2008; Hastie, Tibshirani, and Friedman 2009). Since tree models are rarely used in political science we discuss them here in some detail. In a nutshell, tree models explain variation in a dependent variable by repeatedly splitting the data into even more homogeneous groups, using combinations of independent variables that may be categorical and/or numeric. More formally, tree models are based on an algorithm known as recursive partitioning. Recursive partitioning is a step-by-step process by which a tree is constructed by either splitting or not splitting each node on the tree into two daughter nodes. A tree is the result of a sequence of ordered Boolean ("yes"/"no") questions (e.g., is $X_i \leq \theta_j$?, where θ_j is a threshold value). For classification problems, the outcome variable is binary. For regression problems, the outcome variable is continuous.

The starting point in constructing a tree is the root node, which consists of the entire sample. A node is a subset of the sample; it can be terminal (without daughter nodes) or non-terminal (with daughter nodes). Non-terminal nodes, also called parent nodes, always split into two daughter nodes. These splits are based on Boolean questions about a single variable; depending on the answer given, an observation in the parent node is then assigned to either one of the two daughter nodes.

Figure 1 provides an example with a continuous covariate X_1 and a nominal covariate X_2 with 5 categories, $X_2 \in \{a, b, c, d, e\}$. At the top is the root node, which contains the entire sample. The observations in the root node are first queried about their X_2 values. Observations with $X_2 \in \{a, d\}$ drop down the tree to the left daughter node $D1$; observations with $X_2 \in \{b, c, e\}$ drop down the tree to the right daughter node $D2$. Observations in $D1$ are then queried about their X_1 values. Observations with $X_1 \leq \theta_1$ drop down to the left daughter node, denoted τ_1 in Figure 1 because it is a terminal node; observations with $X_1 > \theta_1$ drop down to the right daughter node, denoted τ_2. The same happens on the right side of the tree, where observations that originally landed in the right

daughter node below the root node ($D2$) are split according to their X_1 values. Some of them are then split again based on their X_2 values.

In Figure 1, the final tree has partitioned the data into five terminal nodes, τ_1 to τ_5. Each observation is assigned to one, and only one, of these terminal nodes. For example, τ_1, the first terminal node in Figure 1, contains all observations that satisfy $\{X_1 \leq \theta_2, X_2 \in \{a, d\}\}$. Observations in these terminal nodes are then assigned fitted values, which usually consist of a majority "vote" in the case of binary outcomes and the average outcome in the case of continuous outcomes. For example, if five observations landed in τ_1, three of which had the outcome 0 and two of which had the outcome 1, all five observations would be assigned the fitted value 0. If the five outcomes were continuous (e.g., 24.2, 12.7, 32.7, 8.2, and 32.0), the fitted value assigned to each of them would be their mean (21.96). Other, more complicated rules for assigning fitted values to terminal nodes are discussed in Berk (2008).

The tree-growing algorithm has to answer two questions: First, at each nonterminal node, how should it choose the variable to split on and its threshold value (in the case of a continuous covariate) or subset (in the case of a nominal covariate)? For example, why was the first split in Figure 1 on X_2 instead of X_1, and why did the algorithm use the subset $X_2 \in \{a, d\}$ to split the tree and not some other subset? Second, how should the tree-growing algorithm decide when a daughter node becomes a terminal node (i.e., is not split any further)?

The first question is answered by evaluating all possible splits for all covariates by a goodness-of-split criterion.[1] Recall that the tree-growing algorithm attempts to create terminal nodes that are as homogeneous as possible in terms of their outcomes. In the case of binary outcomes, the goodness-of-split criterion measures the decrease in impurity (compared to the parent node) that would result from a given split into two daughter nodes, weighted by the number of observations in each daughter node.[2] With continuous outcomes,

[1]The number of possible splits for a continuous variable is the number of its unique values -1; the number of possible splits for a nominal variable is $2^{M-1}-1$, where M is the number of categories (Izenman 2008). One simply sums the number of splits for each variable to get the total number of possible splits at a node. For example, if X_1 had $1,000$ unique values, the total number of possible splits at the root node in Figure 1 would have been $(1000-1)+(2^{5-1}-1) = 999+15 = 1,014$.

[2]Several different measures of impurity exist; a commonly used one is the Gini index $I(n) = 2p(1-p)$,

goodness-of-split can be evaluated by improvements in the weighted mean squared error (MSE). After the tree-growing algorithm has evaluated all possible splits on all possible covariates, it chooses the one that leads to the largest reduction in impurity or MSE. In Figure 1, for example, the first split on X_2, using the subset $X_2 \in \{a, d\}$, was chosen because it lead to a larger reduction in impurity than any other possible split on X_1 or X_2.

The second question has a number of possible answers. The tree-growing algorithm could stop growing a tree whenever the maximum reduction in impurity or MSE falls below some threshold, so that splitting the tree was no longer worthwhile. Alternatively, one could set some minimum number of observations below which a node is considered a terminal node. A superior alternative is to grow large trees and to then selectively prune them back. Breiman et al. (1984), Izenman (2008), and Hastie, Tibshirani, and Friedman (2009) discuss pruning in detail.

While tree models avoid parametric assumptions and excel at the modeling of complex structures, they also suffer from a number of drawbacks. Ensemble methods such as boosting (Freud and Schapire 1997), random forests (Breiman 2001), or BART (discussed below) overcome these drawbacks by combining the predictions from multiple trees. One drawback of single tree models is that their piecewise-constant fit leads to a lack of smoothness of the response surface. Single tree models can also have high variability in the sense that small changes in the data can lead the algorithm to grow a radically different tree. Finally, while single trees can model complicated interactions, they are not well suited for modeling simple additive structures (Hastie, Tibshirani, and Friedman 2009). Ensemble methods that combine predictions from a large number of trees provide a more powerful alternative. We now summarize the BART approach. Our exposition closely follows Chipman, George, and McCulloch (2008). Hill and McCulloch (Forthcoming) demonstrate BART's usefulness for nonparametric causal inference in observational studies with ignorable treatment assignment.

where $I(n)$ is the value of the Gini index for node n and p is the proportion of 1s in node n. As is easily seen, $I(n)$ reaches its minimum when all observations in a node are either 1s or 0s; it reaches its maximum value of .5 when exactly half the observations are 1s ($p = .5$). Other measures of impurity are discussed in Izenman (2008).

Consider the problem of making inferences about an unknown function f that predicts a continuous outcome Y using a p-dimensional vector of predictors $x = (x_1, \ldots, x_p)$ with

$$Y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \tag{6}$$

BART approximates $f(x) = \mathbb{E}(Y \mid x)$ by a sum of m regression trees: $f(x) \equiv h(x) \approx \sum_{j=1}^{m} g_j(x)$, where each g_j denotes a regression tree. Thus (6) is approximated by a sum-of-trees model

$$Y = h(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \tag{7}$$

which is essentially an additive model with multivariate components (the individual trees), allowing it to naturally incorporate interaction effects. Each of the individual trees predicts a small and different portion of f, allowing BART to flexibly model Y as a possibly complex function of the predictors.

We start by introducing notation for a single tree: Let T denote a tree consisting of a set of interior nodes and a set of terminal nodes, and let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denote a set of parameter values associated with the b terminal nodes of T. For a given T and M, $g(x; T, M)$ denotes the function which assigns a $\mu_i \in M$ to an observation with covariate vector x. Thus,

$$Y = g(x; T, M) + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \tag{8}$$

Under (8), $\mathbb{E}(Y \mid x)$ equals the terminal node parameter μ_i assigned by $g(x; T, M)$. The sum-of-trees model can be written as

$$Y = \left(\sum_{j=1}^{m} g(x; T_j, M_j) \right) + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \tag{9}$$

where for each tree T_j and its associated terminal node parameters M_j, $g(x; T_j, M_j)$ is the function that assigns $\mu_{ij} \in M_j$ to covariate vector x. Under (9), $\mathbb{E}(Y \mid x)$ equals the sum of all the terminal nodes μ_{ij} assigned to x by $g(x; T_j, M_j)$. Note that each μ_{ij} represents a main effect when $g(x; T_j, M_j)$ depends on only one component of x, i.e., when the tree only splits on a single variable. When $g(x; T_j, M_j)$ depends on more than one component

of x, μ_{ij} represents an interaction effect. Therefore, and in contrast to single tree models, BART can naturally incorporate both main and interaction effects since some of the trees can represent main effects while others represent interactions. And since the trees in (9) are normally of varying sizes, the interactions may be of varying orders.

Up to this point we have discussed the use of BART with continuous outcomes. BART however can also be used when outcomes are binary. There are various standard binary outcome models that one could use, among which logit and probit are the most common. Here we rely on a probit version of BART; this choice is unlikely to appreciably affect our inferences.

$$p(x) \equiv P(Y = 1 \mid x) = \Phi[G(x)], \tag{10}$$

where

$$G(x) \equiv \sum_{j=1}^{m} g(x; T_j, M_j) \tag{11}$$

and $\Phi[\cdot]$ is the standard normal cdf.

BART's flexibility in modeling Y comes at a computational cost. For a fixed number of trees, m, the sum-of-trees model is determined by the terminal node parameters, $(M_1, \ldots, M_m)$, the parameters describing the tree structures, $(T_1, \ldots, T_m)$, and σ (for continuous outcomes only). These parameters are estimated using a Gibbs sampler. Priors for $(T_1, M_1, \ldots, T_m, M_m)$ and σ (for continuous outcomes only) are chosen so that the contribution of each individual tree is relatively small. Chipman, George, and McCulloch (1998, 2008) discuss the choice of priors in detail and show that the default priors they suggest are remarkably effective in a wide variety of applications. We use these default priors in our empirical analyses; our results do not change appreciably when we explore different priors.[3] Chipman, George, and McCulloch (2008) recommend to set the number of trees to 200; fitting 100 or 500 trees instead has no substantial effects on our results, confirming

[3]Results are available upon request.

Chipman, George, and McCulloch's (2008) observation that BART is quite insensitive to overfitting.

We estimate conditional average treatment effects using posterior simulation. We construct two new data matrices of size $N \times K$, where N is the number of observations and K is the number of covariates +1 (for the treatment indicator). Both of these matrices are identical to the original data matrix, except that the value of the covariate of interest is set to one of its sample values for all observations. All other covariates remain at their observed values. For all observations, the treatment indicator is set to 0 in the first matrix and to 1 in the second matrix. After a burn-in phase of 300 iterations, we take 1,000 posterior draws for each observation in each of the two matrices, resulting in two $N \times 1,000$ matrices of predicted values. We then average over the rows of each matrix, which results in two vectors of predicted values (averaged over all observations) with 1,000 elements each. Subtracting the first vector from the second produces 1,000 posterior draws of the conditional average treatment effect at the specified covariate value. We create such pairs of matrices for each unique covariate value to estimate CATEs at all observed values of the covariate. We repeat this procedure for all covariates for which we want to estimate CATEs. All computations were conducted using a slightly customized version of the BayesTree R package (Chipman and McCulloch 2009).[4]

3. Empirical examples

In order to illustrate how BART may be used to detect and model systematic treatment effect heterogeneity, we offer two examples, a survey experiment and a field experiment. The estimated average treatment effects in both experiments are sizable and well-established by replication studies. Both datasets are large enough to allow us to investigate treatment effect heterogeneity with ample statistical power. And both experiments have prompted researchers to investigate treatment-covariate interactions (e.g., Federico 2004; Imai and

[4]Specifically, we modified the original C^{++} code of the BayesTree package to reduce memory requirements when estimating CATEs from large datasets. In our version, predicted values are averaged over all observations internally, which dramatically reduces the amount of memory needed. The original BayesTree package is available from CRAN; our modified version can be found on our websites.

Strauss 2009; Feller and Holmes 2009). We offer two examples rather than one in order to emphasize the fact that BART provides an appropriately cautious approach to the estimation of CATEs. In one experiment, BART detects strong treatment-covariate interactions in the training set that are for the most part confirmed in the test set. In the other experiment, BART finds very limited evidence of systematic treatment effect heterogeneity. In sum, BART offers a principled and robust estimation framework for experimentalists interested in exploring systematic treatment effect heterogeneity.

3.1. A Survey Experiment on Support for Public Assistance

For decades, scholars studying Americans' support for social welfare spending have noted the special disdain that Americans harbor for programs labeled "welfare" (Williamson 1974; Kluegel and Smith 1986; Smith 1987; Rasinski 1989; Shaw and Shapiro 2002). This phenomenon became the object of sustained experimental inquiry in the mid-1980s, when the General Social Survey (GSS) included a question wording experiment in its national survey of adults. Respondents in each survey are randomly assigned to one of two questions about public spending. Both questions have the same introduction and the same response options, but in one experimental condition respondents are asked about "welfare," while respondents in the other condition are asked about "assistance to the poor."[5]

This seemingly innocuous variation in question wording has a profound effect on support for government spending in this domain. Using GSS surveys from 1986–2008, Table 1 compares the proportion of respondents stating that "too much" is being spent on either welfare or assistance to the poor.[6] In each survey wave, this question wording experiment generates a large and statistically significant ATE ranging from 27.9 percentage points to 50.6 percentage points, implying that between one-quarter and one-half of the population

[5] "We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount. Are we spending too much, too little, or about the right amount on . . . ?"

[6] We dropped the 1984 and 1985 GSS surveys from the analysis because they relied on a flawed question wording randomization procedure that created imbalances in a number of socio-demographic predictors (Smith and Peterson 1986). Also note that the GSS was not conducted every year.

believes that government is spending too much on welfare but not on assistance to the poor. We randomly split this sample into training (7,278 respondents) and test (7,277 respondents) sets.

The magnitude and robustness of this average treatment effect has attracted a fair amount of scholarly attention. The effect has been attributed to the contrasting stereotypes associated with welfare recipients and poor people (Henry, Reyna, and Weiner 2004), particularly racial stereotypes (Gilens 1999; Federico 2004), and to political orientations such as individualism and conservatism (Kluegel and Smith 1986; Bullock, Williams, and Limbert 2003). Relatively little attention, however, has been devoted to the question of treatment effect heterogeneity. Henry (2004) considers the interaction between the treatment and attributions, while Federico (2004) examines a complicated three-way interaction between the treatment, education, and racial perceptions. Jacoby (2000) suggests that party and ideology may make some respondents especially receptive to the more specific program content of "assistance to the poor." In the next section we use BART to investigate the extent to which covariates such as these moderate the question wording effect.

3.1.1. Results

The ATE in the training set is 0.366 with a .95 confidence interval of $(0.347, 0.385)$. Because our outcome variable is binary, we rely on a BART probit model to search for systematic treatment effect heterogeneity. Figures 2 and 3 display conditional average treatment effects. We included the following covariates in the model: a dummy variable for each survey wave, a scale that measures negative attitudes toward blacks, age (in years), education (in years), a 7-point liberal-conservative scale, and a 7-point party identification scale.[7]

[7]The negative attitudes toward blacks scale is based on 4 "yes"/"no" responses to the following survey question: "On average Blacks have worse jobs, income, and housing than white people. Do you think these differences are . . . ", where respondents were presented with 4 possibilities: "Mainly due to discrimination?" ("yes" = 0; "no" = 1) ; "Because most Blacks have less in-born ability to learn?" ("yes" = 1; "no" = 0); "Because most Blacks don't have the chance for education that it takes to rise out of poverty?" ("yes" = 0; "no" = 1); "Because most Blacks just don't have the motivation or will power to pull themselves up out of poverty?" ("yes" = 1; "no" = 0). We coded each response as either 0 or 1 and took the average over all responses. When some responses were missing we used the remaining responses to construct the index. Party identification ranges from 1 (strong Democrat) to 7 (strong Republican).

Conditional average treatment effects are estimated using posterior simulation, as explained above. The left column of Figures 2 and 3 shows results for the training set. If we detect significant treatment effect heterogeneity conditional on a given covariate in the training set, we replicate the analysis using the test set (right column). The dark grey areas are point-wise .95 percentile uncertainty bands; the light grey areas are global .95 percentile uncertainty bands (Mandel and Betensky 2008). Marginal covariate distributions are shown at the bottom of the graphs.

From the topleft graph in Figure 2 we can see that the effect of changing the question wording from "assistance to the poor" to "welfare" is strongly moderated by respondents' attitudes toward blacks. The probability of stating that "too much" money is being spent increases by about 27 percentage points for respondents with a score of 0 on the anti-black attitudes scale. This conditional average treatment effect increases monotonically with increases in the attitudes scale; the treatment effect for respondents with a score of 1 is about 42 percentage points, for a difference of 15 percentage points between respondents at either end of the scale. This represents the moderating effect of negative attitudes toward blacks controlling for all other covariates in the model. We can formally test whether this treatment effect heterogeneity is statistically significant by conducting a Wald test based on the estimated CATEs and their variance-covariance matrix (Cameron and Trivedi 2005), which decisively rejects the null hypothesis that the CATEs are identical ($p < .0001$).

The two graphs below the topleft graph display treatment effects as a function of age and education, respectively. Both CATE curves appear to be more or less flat. We cannot reject the null hypotheses that CATEs are identical across all values of age or education.

The left column of Figure 3 displays CATEs as a function of liberal-conservative self-placement and party identification, respectively. Both graphs show strong effect moderation: as respondents become more conservative or more Republican, the treatment effect increases. This effect is especially pronounced for the liberal-conservative scale, where the difference in treatment effects between respondents at either end of the scale is about 13 percentage points. We strongly reject the null hypotheses that CATEs are identical across

the liberal-conservative scale ($p < .0001$) or levels of party identification ($p < .0001$).

The right columns of Figures 2 and 3 replicate our analysis using the test set. We do not present results for age and education because we failed to discover meaningful treatment effect heterogeneity conditional on these two covariates in the training set. Results in the test set mostly replicate the earlier results: negative attitudes toward blacks and party identification moderate the treatment effect (both $p < .0001$). Results for the liberal-conservative scale, however, are less clear-cut than in the training set. Although we still find that estimated treatment effects vary across the liberal-conservative scale, its moderating effect is no longer as strong; we cannot reject the null hypothesis that CATEs across all values of the liberal-conservative scale are identical ($p = .23$).

The graphs shown in Figures 2 and 3 display conditional average treatment effects one covariate at a time. Treatment effects of course could potentially be a more complicated function of several covariates, leading to three-way or even higher order interactions between the treatment indicator and covariates. BART automatically incorporates such higher-order interactions in the model if they improve model fit. We examined all three-way interactions between the treatment and the covariates for which we found evidence of treatment effect heterogeneity in the training set: negative attitudes toward blacks, the liberal-conservative scale, and party identification. For none of these three-way interactions could we reject the null hypothesis that covariates do not jointly moderate the treatment effect.[8]

The individual graphs in Figures 2 and 3 visualize how the effect of the treatment varies with each covariate, but they do not allow us to judge the overall amount of systematic treatment effect heterogeneity in our question wording experiment. The top graph in Figure 4 displays a histogram of estimated conditional average treatment effects for the 7,278 respondents in the training set. We can see that estimated CATEs vary enormously across respondents, depending on individuals' baseline characteristics. Estimated CATEs

[8]Treatment, negative attitudes toward blacks, and liberal-conservative scale ($p = .26$); treatment, negative attitudes toward blacks, and party identification ($p = .69$); and treatment, party identification, and liberal-conservative scale ($p = .99$).

range from 9 percentage points to 53 percentage points, with the median CATE equal to 38 percentage points. Interestingly, all of the respondents appear to be positively affected by the treatment to some degree.

How much of a difference does allowing for systematic treatment effect heterogeneity make? If there were no systematic treatment effect heterogeneity in the data, the range of CATEs shown in Figure 4 would probably shrink dramatically.[9] We can judge the importance of systematic treatment effect heterogeneity in our data by randomly permuting individuals' covariate vectors and re-running BART. This permutation destroys any systematic relationship between the covariates and the outcome, and therefore also any systematic treatment effect heterogeneity. Note that we permute covariate vectors instead of single covariate values; the correlation between covariates is therefore unaffected by our permutation scheme.

The top graph in Figure 5 shows a kernel density plot of CATEs for the individuals in the training set (black curve). It also shows 10 kernel density plots for the same individuals when covariate vectors are randomly permuted (grey curves). It is readily apparent that the range of CATE estimates in the original analysis is much larger than the range of CATE estimates for any permuted dataset. Clearly, our covariates contain valuable information about systematic treatment effect heterogeneity that we would fail to exploit if we were to only estimate the ATE.

CATEs necessarily vary with an individual's baseline probability of success, which is determined by the values of *all* covariates in the BART probit model, even if these covariates are not interacted with the treatment indicator. This "compression" effect can induce additional treatment effect heterogeneity beyond that visible on the scale of the linear predictor. It can also dampen it (Berry, DeMeritt, and Esarey 2010). Since our CATE estimates are reported on the probability scale, they are affected by both sources of heterogeneity. Even though we think that the probability scale is appropriate for reporting conditional average treatment effects when outcomes are binary, we can also look at

[9]The CATEs would not collapse to a single number, the ATE, because of uncertainty in our posterior inferences.

treatment effect heterogeneity on the scale of the linear predictor ($G(x)$ in (10)), which is unaffected by compression. The bottom graphs of Figures 4 and 5 show CATEs on this scale. Even without compression, we find that the range of estimated CATEs in our original dataset is noticeably larger than the range of CATE estimates for any permuted dataset, which provides evidence for the existence of systematic treatment effect heterogeneity on the scale of the linear predictor. The systematic treatment effect heterogeneity visible in Figures 2 and 3 is not purely due to the fact that probabilities are bounded between zero and one.

3.2. A Field Experiment on Voter Mobilization

Our second empirical example involves a voter turnout experiment in which direct mail was used to mobilize voters. The original experiment, reported in Gerber, Green, and Larimer (2008), involved four different kinds of mailings; here, we focus on just one of the experimental mailings, which encouraged recipients to vote by showing an official-looking record of whether they and other members of their household had cast ballots in the previous two elections. This "Self" mailing has been the subject of several large replication studies (Abrajano and Panagopoulos 2009; Gerber, Green, and Larimer 2010; Mann 2010; McConnell, Sinclair, and Green 2010), and there is strong reason to believe that its average treatment effect is substantial and reproducible.

The setting for the experiment considered here was a 2006 statewide primary election in Michigan in which the only contested races involved Republican candidates. The target population was a subset of the registered electorate that was thought to have a non-negligible rate of participation in this low-salience election: those who had voted in the prior presidential election and those who were unlikely to be Democrats. The sample also excluded those with a high probability of voting by mail (i.e., people who might have cast their ballots before receiving the mailing). The net effect of these restrictions was to make the sample somewhat more affluent and less urban than the average registered voter in Michigan. In this experiment, households were randomly assigned to treatment or control conditions. In order to sidestep issues of within-household clustering, we randomly

selected one individual from each multi-person household. Overall, our sample comprises 17,214 individuals in the treatment group and 17,214 individuals in the control group, randomly split into equal-sized training and test sets. The combination of a large sample and a robust average treatment effect makes exploration of treatment effect heterogeneity both feasible and interesting.

3.2.1. Results

The ATE in the training set is 0.053 with a .95 confidence interval of $(0.039, 0.067)$: on average, the "Self" treatment increases turnout by about 5 percentage points. We again rely on a BART probit model to search for systematic treatment effect heterogeneity. We included the following covariates: age (in years), the predicted probability of turning out in primary and general elections, the predicted probability of voting Democratic, the predicted probability of voting by mail, block-level turnout in past primary and general elections, the number of votes cast in previous elections, and indicator variables for gender and turnout in the 2000 and 2002 general elections and 2000, 2002, and 2004 primary elections.

Figures 6 to 8 visualize CATE estimates for the training set. The amount of systematic treatment effect heterogeneity is generally small. The individual graphs show modest heterogeneity in terms of age, the predicted probability of voting Democratic, the number of votes cast in previous elections, and turnout in the 2000 primary election. However, we cannot reject the null hypothesis of identical CATEs for any of the covariates.

This finding of minimal systematic treatment effect heterogeneity is confirmed by Figures 9 and 10. Figure 9 displays a histogram of estimated conditional average treatment effects for the 17,214 respondents in the training set. We can see that estimated CATEs vary somewhat across respondents, ranging from 1 percentage point to 8 percentage points, with the median CATE equal to 5 percentage points. Figure 10 compares this range of CATEs to its expected range if covariates were not informative about treatment effect heterogeneity at all. The black curve in Figure 10 is a kernel density plot of CATEs for the individuals in the training set. The 10 grey curves are kernel density plots for the same individuals when covariate vectors are randomly permuted. In contrast to the GSS data,

where the range of CATE estimates in the original analysis was much larger than the range of CATE estimates for any of the permuted datasets, here the range of CATE estimates in the original analysis looks very similar to the range of CATE estimates based on the permuted datasets. This finding confirms the results from the individual graphs shown in Figures 6 to 8: there is very little systematic treatment effect heterogeneity apparent in the data. Given this absence of systematic treatment effect heterogeneity, we can have greater trust in the ATE to provide a reliable summary measure of the unobservable distribution of treatment effects.

4. Conclusion

Throughout this paper we have highlighted the importance of going beyond the average treatment effect when analyzing randomized experiments. Assessing treatment effect heterogeneity is crucial when applying the results of an experiment to target populations whose observable baseline characteristics differ from the experimental sample. A better understanding of treatment effect heterogeneity may also shed light on causal mechanisms (Imai, Tingley, and Yamamoto 2010). In the GSS question wording experiment, for example, we saw that racial hostility intensifies the effect of a question wording change, suggesting that at least part of this effect is due to the racial connotations of welfare, as suggested previously (e.g., Federico 2004). Whether one's aim is to develop effective treatments or deeper theoretical understanding about the conditions under which they operate, investigation of systematic treatment effect heterogeneity plays a central role in experimental analysis.

The methodology presented here provides a principled framework for exploring such heterogeneity. In an effort to minimize the role of discretion in the analysis of experimental data, we offer a method that largely automates the search for treatment effect heterogeneity. We use Bayesian Additive Regression Trees (BART) to guide the search for heterogeneous treatment effects. BART has the virtue of modeling heterogeneity in a flexible nonparametric manner. In order to ensure that this flexible estimation approach generates reliable insights, we split the data into two equally-sized parts, one to explore treatment effect heterogeneity and the other to confirm the robustness of the preliminary

results. This approach permits a relatively unstructured data-driven exploration while at the same time guarding against the kinds of unreliable inferences that might otherwise arise due to data dredging and multiple comparisons.

The two substantive examples presented above illustrate the value of this conservative approach. In the question wording experiment, we found strong evidence of treatment effect heterogeneity in the training set. The specific sources of heterogeneity that we identified during our exploration of the training data set were confirmed in our analysis of the test set. Our findings confirm the longstanding hypothesis that imagery associated with welfare has racial connotations but call into question the two- and three-way interactions that others have reported. For example, we find little evidence of an interaction with education or a more complex interaction involving education and negative attitudes toward blacks. Indeed, we find no evidence of any higher-order interactions. By narrowing the range of possible interactions to just the robust interactions involving racial attitudes and party identification, our approach provides useful guidance for those seeking to understand the cognitive underpinnings of this question wording effect.

As interesting as it is to discover systematic variation in treatment effects, a reliable method for detecting heterogeneity must also be able to discover the *absence* of heterogeneity. Applying BART to the voter mobilization field experiment revealed no evidence of systematic treatment effect heterogeneity. In this experiment, calling attention to the voter's record of participation in past elections had similar effects, regardless of attributes such as age, voting propensity, or the voting propensity of one's neighbors. Indeed, when we applied BART to another large experiment using an identical treatment and a similar population of voters, this pattern of homogeneous treatment effects was essentially confirmed.[10] BART appears to be a reliable method for detecting heterogeneous treatment effects when, and only when, meaningful heterogeneity exists in the data. Again, this type of finding can be of enormous importance to those seeking to understand the basis of this

[10]We replicated our analysis using another large-scale voter mobilization experiment conducted in the context of 2007 Michigan municipal elections in selected towns. Unlike the 2006 study, the 2007 study did not place partisan or turnout propensity restrictions on the sample. For details see Gerber, Green, and Larimer 2010.

experimental effect. An adequate theory must explain why the effects of social pressure are so uniform across different segments of the electorate.

Recent years have seen a dramatic increase in the number and scale of experiments conducted in political science (Druckman et al. 2006), and there are indications that researchers are becoming increasingly sensitive to the challenges involved in drawing robust causal inferences. Both the supply of large-scale experiments and the demand for safeguards against analyst discretion recommend the methodology described here. Advances in computing power and the availability of public domain software mean that computationally intensive nonparametric methods like BART are now readily available to social scientists. In the years ahead, as methods like BART come into currency, we are likely to see a fundamental change in the way that experimenters investigate and report systematic heterogeneity in treatment effects.

5. References

Abrajano, Marisa, and Costas Panagopoulos. 2009. "Does language matter? The impact of Spanish versus English-language GOTV efforts on latino turnout." Unpublished manuscript.

Abadie, Alberto 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics* 113 (2): 231–263.

Abbring, Jaap H. and James J. Heckman. 2007. Econometric evaluations of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In James J. Heckman and Edward E. Leamer (eds.). *Handbook of Econometrics*, volume VI, chapter 72.

Angrist, Joshua D. 2004. "Treatment effect heterogeneity in theory and practice." *Economic Journal* 114: C52–C83.

Berk, Richard A. 2008. *Statistical learning from a regression perspective.* Springer.

Berry, William D., Jacqueline H. R. DeMerrit, and Justin Esarey. 2010. "Testing for interaction in binary logit and probit models: Is a product term essential?" *American Journal of Political Science* 54 (1): 248–266.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes. 2006. "What mean impacts miss: Distributional effects of welfare reform experiments." *American Economic Review* 96 (4): 988–1012.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes. 2009. "Can constant treatment effects within subgroup explain heterogeneity in welfare reform effects?" Working paper.

Byar, David P. 1985. "Assessing apparent treatment-covariate interactions in randomized clinical trials." *Statistics in Medicine* 4: 255–263.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R.A. Olshen. 1984. *Classification and Regression Trees.* Chapman & Hall.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–42.

Bullock, Heather E., Wendy R. Williams, and Wendy M. Limbert. 2003. "Predicting support for welfare policies: The impact of attributions and beliefs about inequality." *Journal of Poverty* 7 (3): 35–56.

Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics.* Cambridge University Press.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 1998. "Bayesian CART model search." *Journal of the American Statistical Association* 94 (443): 935–948.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2008. "BART: Bayesian additive regression trees." Working paper.

Chipman, Hugh and Robert McCulloch. 2009. BayesTree: "Bayesian methods for tree based models." R package version 0.3–1.

Cox, David R. 1958. *Planning of experiments.* John Wiley & Sons.

Cox, David R. 1975. "A note on data-splitting for the evaluation of significance levels." *Biometrika* 62 (2): 441–444.

Crump, Richard K, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2008. "Nonparametric tests for treatment effect heterogeneity." *Review of Economics and Statistics* 90 (3): 389–405.

Dehejia, Rajeev H. 2005. "Program evaluation as a decision problem." *Journal of Econometrics* 125 (1–2): 141–173.

Dixon, Dennis O. and Richard Simon. 1991. "Bayesian subset analysis." *Biometrics* 47 (3): 871–881.

Djebbari, Habiba and Jeffrey Smith. 2008. "Heterogeneous impacts of PROGRESA." *Journal of Econometrics* 145: 64–80.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The growth and development of experimental research in political science." *American Political Science Review* 100 (4): 627–636.

Federico, Christopher M. 2004. "When do welfare attitudes become racialized? The paradoxical effects of education." *American Journal of Political Science* 48 (2): 374–391.

Feller, Avi and Chris C. Holmes. 2009. "Beyond toplines: Heterogeneous treatment effects in randomized experiments."

Freud, Yoav and Robert E. Schapire. 1997. "A decision-theoretic generalization of online learning and an application to boosting." *Journal of Computer and System Sciences* 55 (1): 119–139.

Gabler, Nicole B., Naihua Duan, Diana Liao, Joann G. Elmore, Theodore G. Ganiats, and Richard L. Kravitz. 2009. "Dealing with heterogeneity of treatment effects: Is the literature up to the challenge?" *Trials* 10: 43.

Gelman, Andrew. 2004. Treatment effects in before-after data. In Andrew Gelman and Xiao-Li Meng (eds.). *Applied Bayesian modeling and causal inference from incomplete-data perspectives.* John Wiley & Sons.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American Political Science Review* 102 (1): 33–48.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2010. "An experiment testing the relative effectiveness of encouraging voter participation by inducing feelings of pride or shame." *Political Behavior*.

Gilens, Martin. 1999. *Why Americans hate welfare: Race, media and the politics of anti-poverty policy.* Chicago: University of Chicago Press.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning.* Second edition. Springer.

Heckman, James J., Jeffrey Smith, and Nancy Clements. 1997. "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts." *Review of Economic Studies* 64 (4): 487–535.

Heller, Ruth, Paul R. Rosenbaum, and Dylan S. Small. 2009. "Split samples and design sensitivity in observational studies." *Journal of the American Statistical Association* 104 (487): 1090–1101.

Henry, P. J., Christine Reyna, and Bernard Weiner. 2004. "Hate welfare but help the poor: How the attributional content of stereotypes explains the paradox of reactions to the destitute in America." *Journal of Applied Social Psychology* 34 (1): 34–58.

Hill, Jennifer L. and Robert E. McCulloch. Forthcoming. "Bayesian nonparametric modeling for causal inference." *Journal of the American Statistical Association.*

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 81 (396): 945–960.

Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. "Designing and analyzing randomized experiments: Application to a Japanese election survey experiment." *American Journal of Political Science* 51 (3): 669–687.

Imai, Kosuke and Aaron Strauss. 2009. "Planning the optimal get-out-the-vote campaign using randomized field experiments." Working paper.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto. 2010. "Experimental identification of causal mechanisms." Working paper.

Izenman, Alan Julian. 2008. *Modern multivariate statistical techniques: Regression, Classification, and manifold learning.* Springer.

Jacoby, William G. 2000. "Issue framing and public opinion on government spending. *American Journal of Political Science* 44 (4): 750–767.

Kluegel, James R. and Eliot R. Smith. 1986. *Beliefs about inequality: Americans' views of what is and what ought to be.* New York: Aldine de Gruyter.

Luskin, Robert C. 1990. "Explaining political sophistication." *Political Behavior* 12 (4): 331–361.

Mandel, Micha and Rebecca A. Betensky. 2008. "Simultaneous confidence intervals based on the percentile bootstrap approach." *Computational Statistics & Data Analysis* 52 (4): 2158–2165.

Mann, Christopher B. 2010. "Is there backlash to social pressure? A large-scale field experiment on voter mobilization." *Political Behavior*.

McConnell, Margaret, Betsy Sinclair, Donald P. Green. 2010. "Detecting social networks: Design and analysis of multilevel experiments." Paper presented at the Third Annual New York University Center for Experimental Social Sciences Conference on Experimental Political Science.

Morgan, James N. and John A. Sonquist. 1963. "Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association* 58 (302): 415–434.

Poirier, Dale J. and Justin L. Tobias. 2003. "On the predictive distributions of outcome gains in the presence of an unidentified parameter." *Journal of Business & Economic Statistics* 21 (2): 258–268.

Royston, Patrick and Willi Sauerbrei. 2004. "A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials." *Statistics in Medicine* 23: 2509–2525.

Rothwell, Peter M. 2005. "Subgroup analysis in randomized controlled trials: Importance, indications, and interpretation." *Lancet* 365: 176–186.

Pocock, Stuart J., Susan E. Assmann, Laura E. Enos, and Linda E. Kasten. 2002. "Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems." *Statistics in Medicine* 21: 2917–2930.

Rasinski, Kenneth A. 1989. "The effect of question wording on public support for government spending." *Public Opinion Quarterly* 53 (3): 388–394.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66: 688–701.

Rubin, Donald B. 1978. "Bayesian inference for causal effects: The role of randomization." *Annals of Statistics* 6 (1): 1–26.

Shaw, Greg M. and Robert Y. Shapiro. 2002. "Trends: Poverty and public assistance." *Public Opinion Quarterly* 66 (1): 105–128.

Smith, Tom W. 1987. "That which we call welfare by any other name would smell sweeter." *Public Opinion Quarterly* 51 (1): 75–83.

Smith, Tom W. and Bruce L. Peterson. 1986. "Problems in form randomization on the General Social Survey." GSS Methodological Report No. 36.

Tukey, John W. 1977. *Exploratory data analysis*. Addison-Wesley.

Williamson, John B. 1974. "Beliefs about the motivation of the poor and attitudes toward poverty policy." *Social Problems* 21 (5): 634–648.

6. Tables

Table 1: Public Support for Government Spending on Welfare/Assistance to the Poor

year	sample size		mean		ATE
	Assistance	Welfare	Assistance	Welfare	
1986	598	561	0.104	0.447	0.344
1988	404	359	0.079	0.451	0.372
1989	393	375	0.104	0.443	0.338
1990	536	511	0.086	0.423	0.337
1991	394	379	0.127	0.406	0.279
1993	418	418	0.148	0.598	0.450
1994	744	761	0.168	0.674	0.506
1996	705	700	0.217	0.639	0.422
1998	683	665	0.124	0.468	0.343
2000	639	666	0.131	0.413	0.281
2002	344	332	0.110	0.482	0.371
2004	341	338	0.070	0.476	0.406
2006	673	675	0.098	0.393	0.295
2008	487	456	0.092	0.414	0.322
total/mean	7,359	7,196	0.124	0.489	0.365

Source: General Social Survey 1986–2008. The table displays the proportion of respondents stating that "too much" money is spent on Assistance to the Poor (the control condition) or Welfare (the treatment condition). All average treatment effect estimates are statistically significant at the .01 level or better.

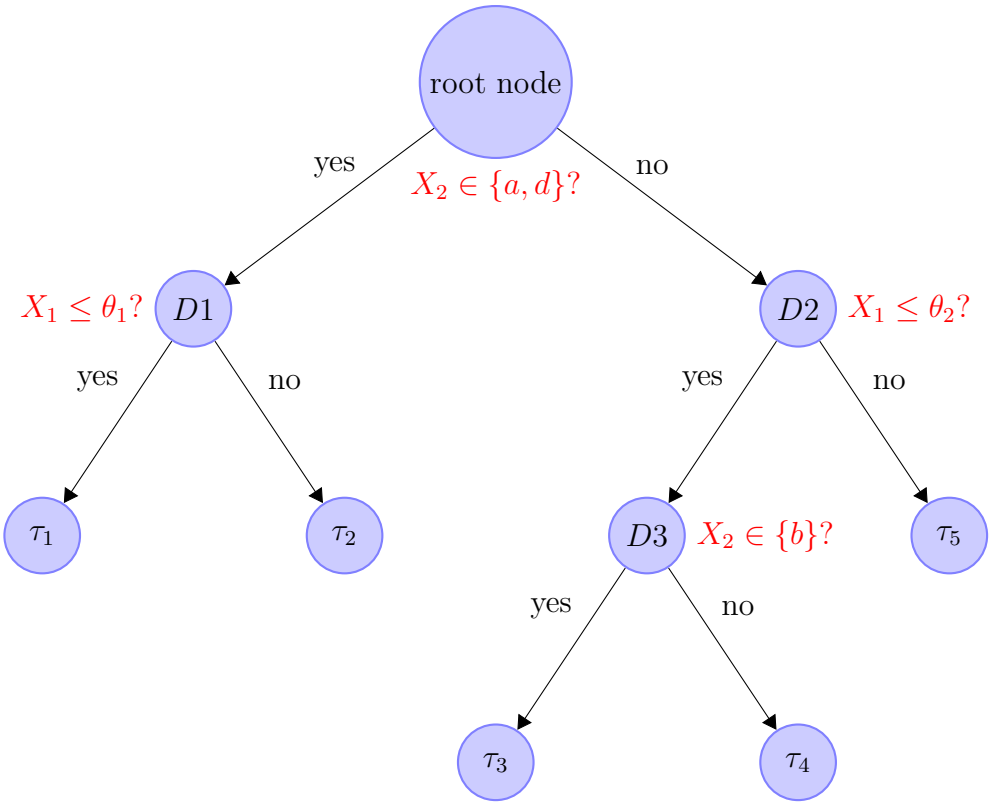

Figure 1: Example of a regression or classification tree

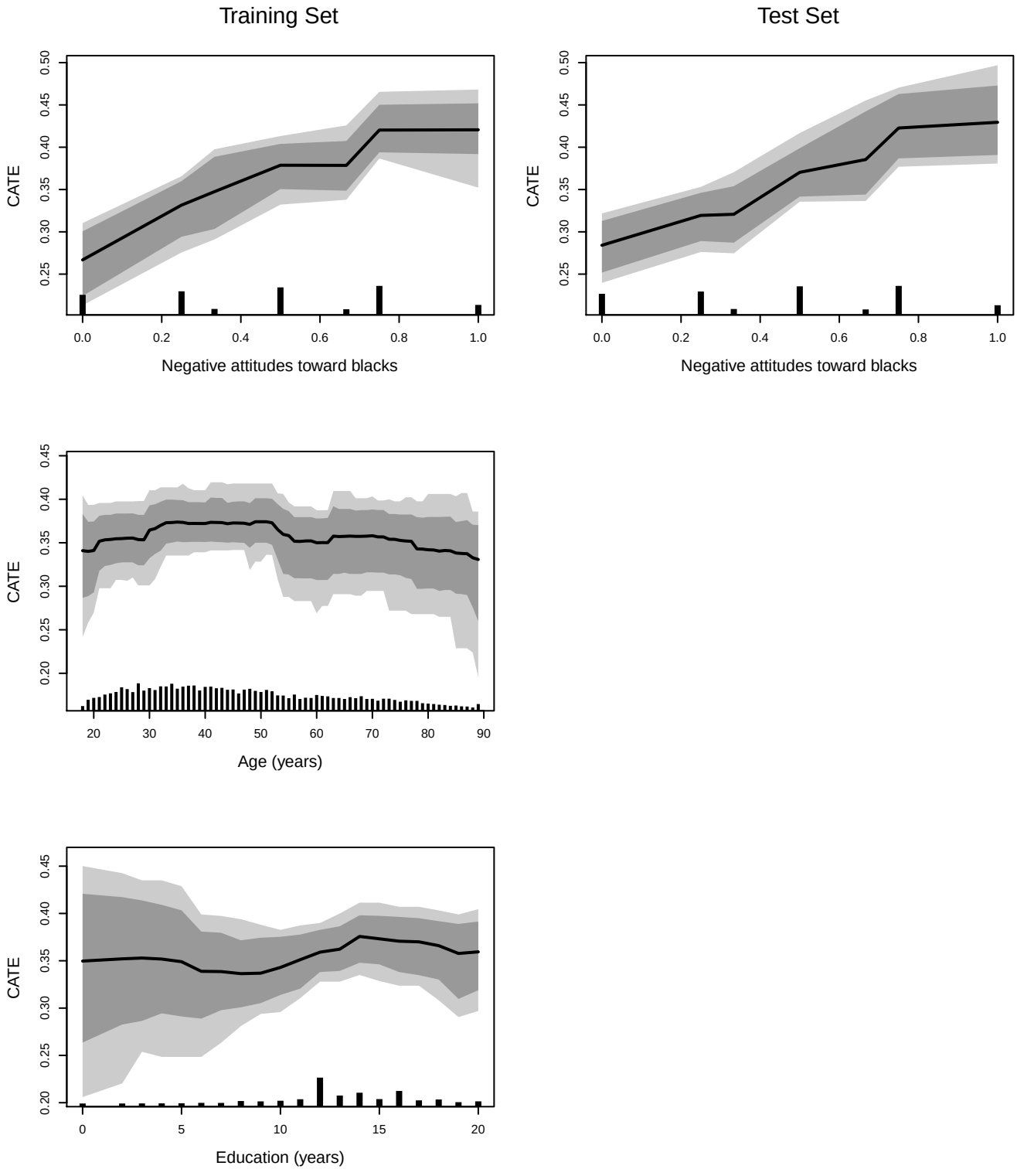

Figure 2:

Source: GSS 1986–2008. The graphs show average treatment effects (the black curves) conditional on covariates for the training set (left column) and test set (right column), where appropriate. The dark grey areas are point-wise .95 percentile uncertainty bands; the light grey areas are global .95 percentile uncertainty bands. Marginal covariate distributions are shown at the bottom of the graphs.

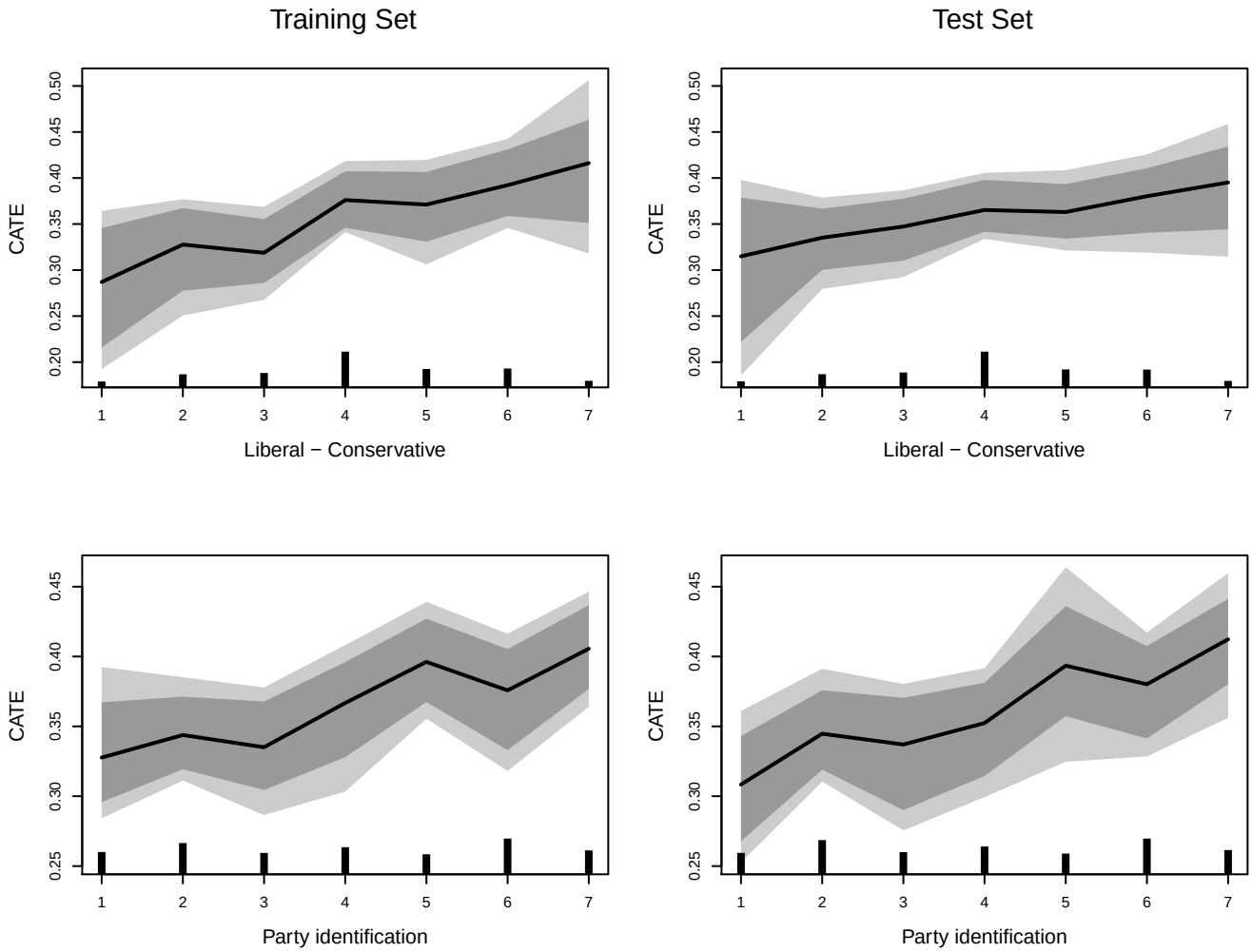

Figure 3:

Source: GSS 1986–2008. The graphs show average treatment effects (the black curves) conditional on covariates for the training set (left column) and test set (right column). The dark grey areas are point-wise .95 percentile uncertainty bands; the light grey areas are global .95 percentile uncertainty bands. Marginal covariate distributions are shown at the bottom of the graphs.

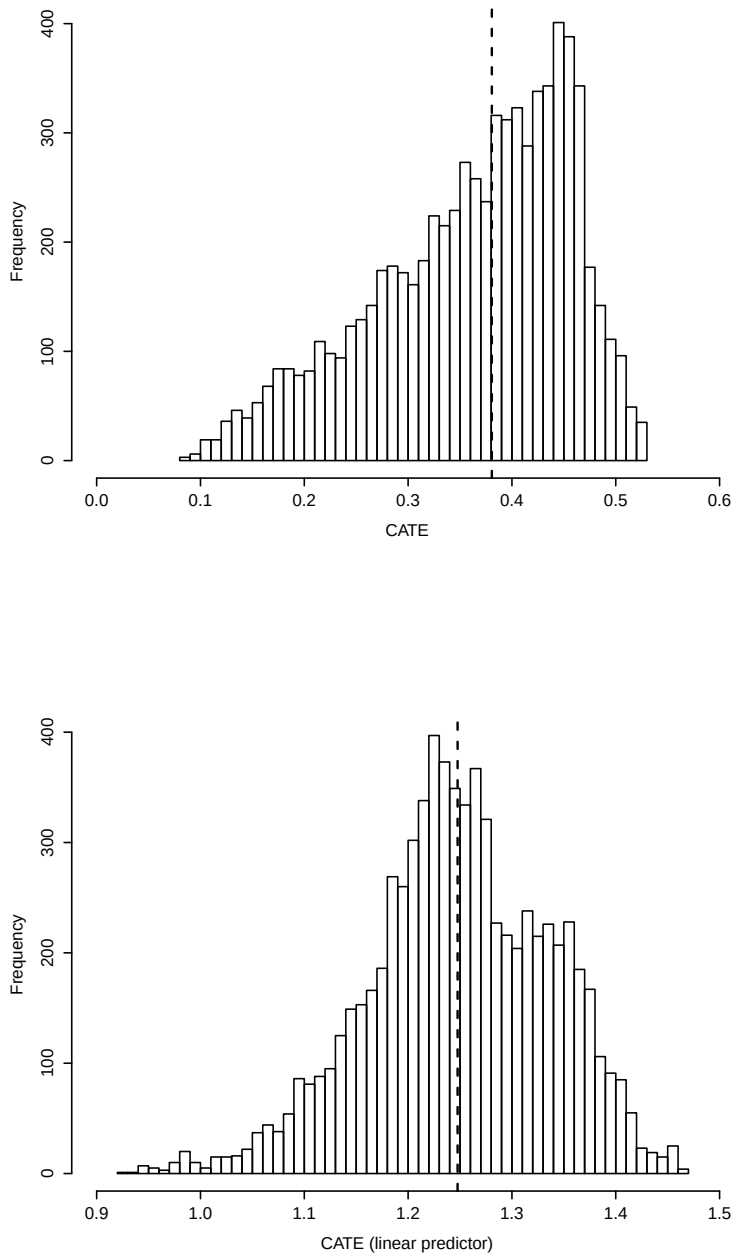

Figure 4:

Source: GSS 1986–2008. The top graph shows a histogram of conditional average treatment effects on the probability scale for the 7,278 individuals in the training set. The bottom graph shows a histogram of conditional average treatment effects on the scale of the linear predictor for the same individuals. The vertical dashed lines denotes the median conditional average treatment effects.

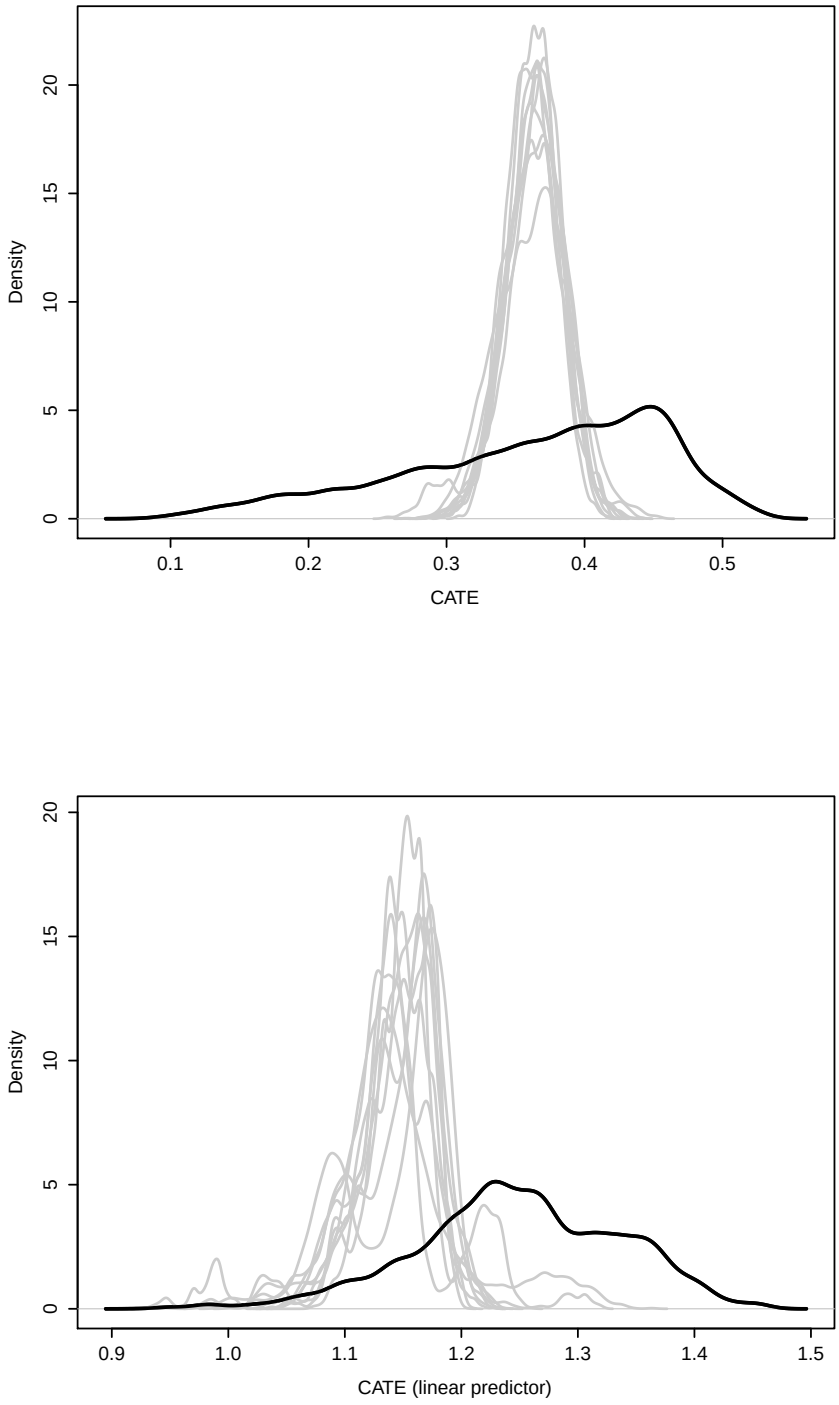

Figure 5:

Source: GSS 1986–2008. The top graph shows a kernel density plot of conditional average treatment effects on the probability scale for the 7,278 individuals in the training set (black curve) and 10 kernel density plots for the same individuals when covariate values are randomly permuted (see text for details). The bottom graph shows the same plots for conditional average treatment effects on the scale of the linear predictor.

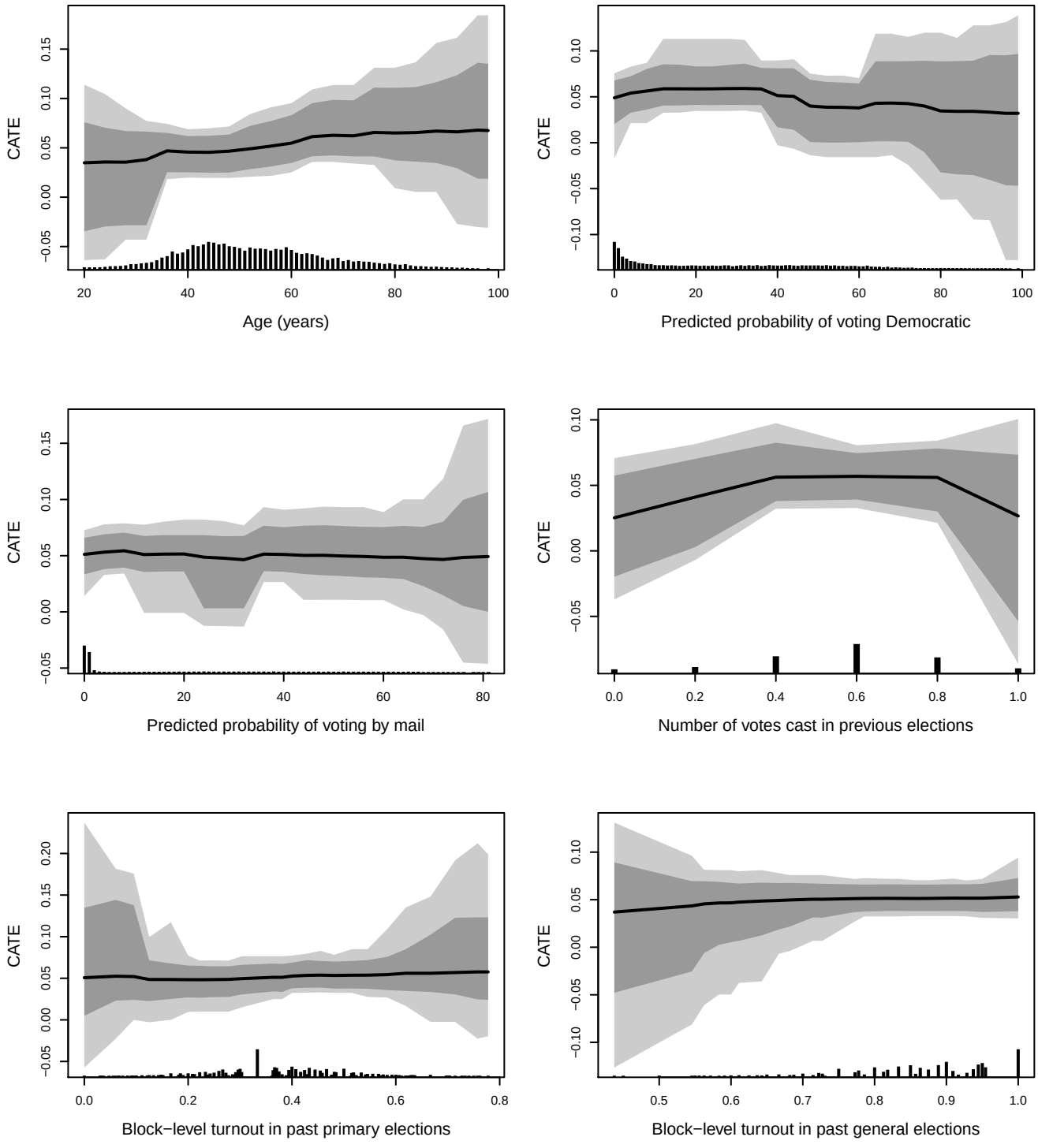

Figure 6:

Source: Michigan 2006 voter mobilization experiment (Gerber, Green, and Larimer 2008). The graphs show average treatment effects (the black curves) conditional on covariates for the training set. The dark grey areas are point-wise .95 percentile uncertainty bands; the light grey areas are global .95 percentile uncertainty bands. Marginal covariate distributions are shown at the bottom of the graphs.

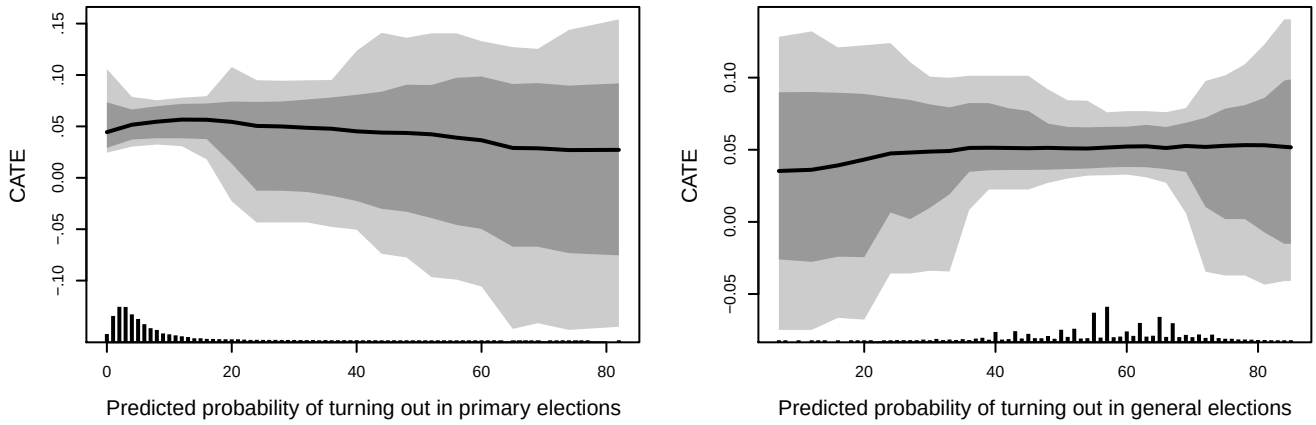

Figure 7:

Source: Michigan 2006 voter mobilization experiment (Gerber, Green, and Larimer 2008). The graphs show average treatment effects (the black curves) conditional on covariates for the training set. The dark grey areas are point-wise .95 percentile uncertainty bands; the light grey areas are global .95 percentile uncertainty bands. Marginal covariate distributions are shown at the bottom of the graphs.

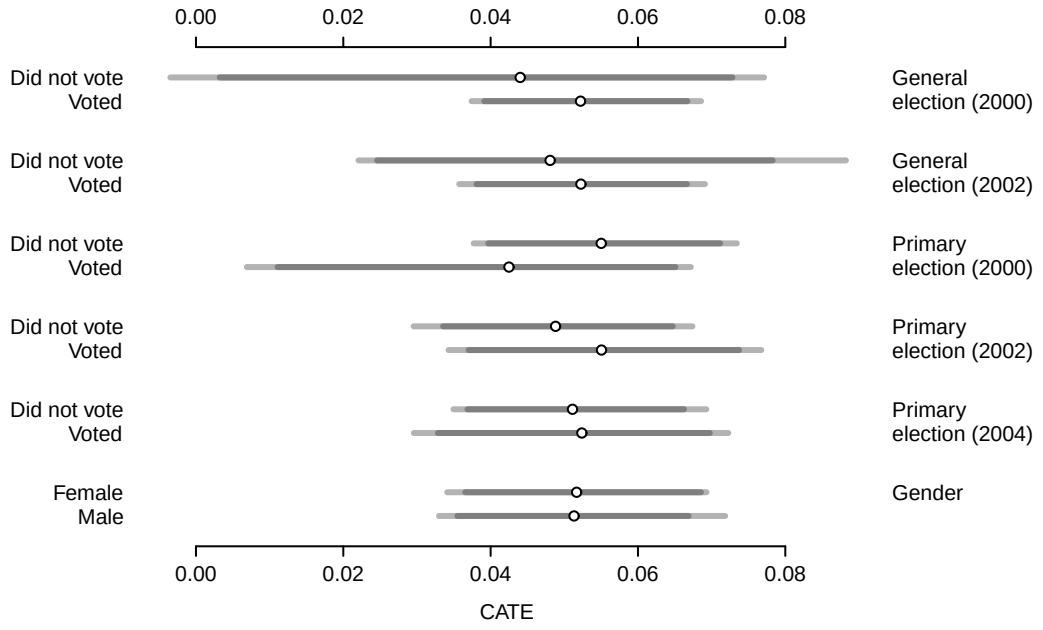

Figure 8:

Source: Michigan 2006 voter mobilization experiment (Gerber, Green, and Larimer 2008). The graph shows average treatment effects (circles) conditional on past turnout in general elections (2000 and 2002) and primary elections (2000, 2002, and 2004) and gender for the training set. The dark grey bars are point-wise .95 percentile uncertainty intervals; the light grey bars are global .95 percentile uncertainty intervals.

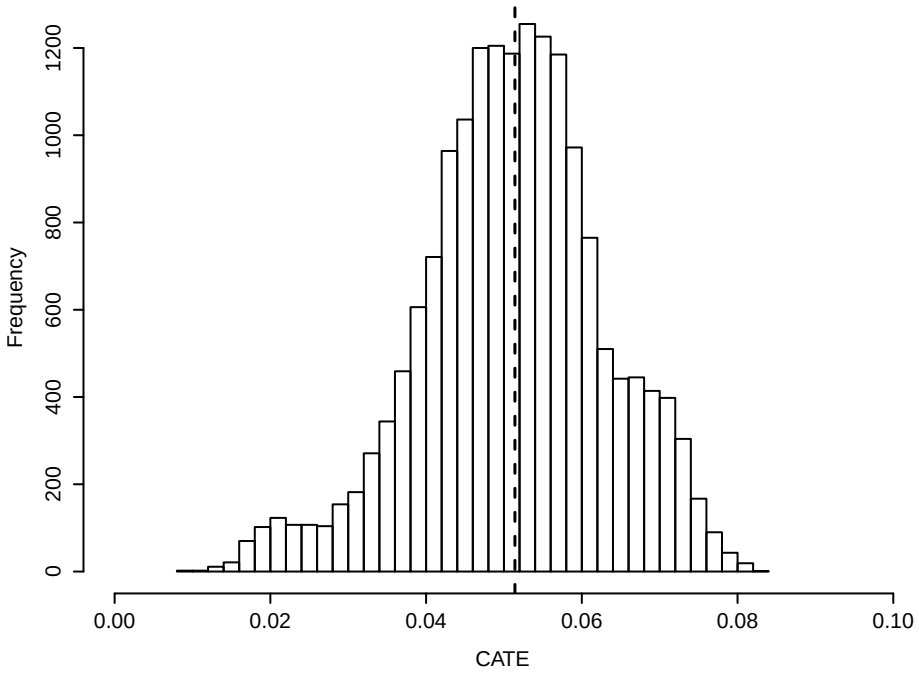

Figure 9:

Source: Michigan 2006 voter mobilization experiment (Gerber, Green, and Larimer 2008). The graph shows a histogram of conditional average treatment effects for the 17,214 individuals in the training set. The vertical dashed line denotes the median conditional average treatment effect.

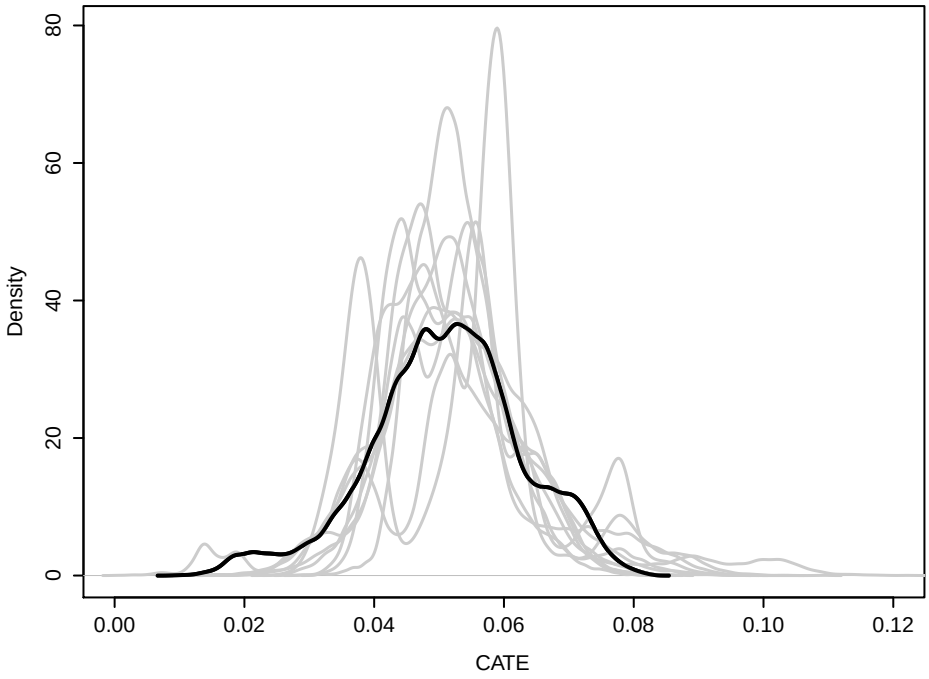

Figure 10:

Source: Michigan 2006 voter mobilization experiment (Gerber, Green, and Larimer 2008). The graph shows a kernel density plot of conditional average treatment effects for the 17,214 individuals in the training set (black curve) and 10 kernel density plots for the same individuals when covariate values are randomly permuted (see text for details).