Modeling heterogeneous treatment effects in survey experiments with Bayesian Additive Regression Trees

Donald P. Green[†]
Holger L. Kern[‡]

Word count: 6,360

The authors are grateful to Peter Aronow, Kate Cowles, Alan Gerber, Jennifer Hill, Dan Hopkins, Kosuke Imai, Gary King, Scott Long, Mary McGrath, Jas Sekhon, and Rocío Titiunik for their comments, Arjun Shenoy for his help in reviewing the literature on public support for welfare, and the facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center for their support.
[†]Department of Political Science, Columbia University.
[‡]Department of Political Science, University of South Carolina.

Abstract: Survey experimenters routinely test for systematically varying treatment effects by using interaction terms between the treatment indicator and the covariates. Parametric models such as linear or logistic regression currently used to search for systematic treatment effect heterogeneity suffer from several shortcomings, in particular the potential for bias due to functional form misspecification and the large amount of discretion they introduce into the analysis of experimental data. Here we explicate what we believe to be a better approach. Drawing on recent advances in statistical learning, we recommend the use of Bayesian Additive Regression Trees (BART) as a method for analyzing treatment effect heterogeneity in survey experiments. BART has a number of important advantages over parametric modeling strategies, including its ability to *automatically* detect and model non-linear relationships and interactions, which removes researchers' discretion from the data analysis, and its insensitivity to the choice of tuning parameters. These features make BART an appealing "off-the-shelf" tool for survey experimenters who want to model systematic treatment effect heterogeneity in a flexible and robust manner. In order to illustrate how BART can be used to detect and model heterogeneous treatment effects, we reanalyze a well-known survey experiment on welfare attitudes from the General Social Survey.

1. Introduction

Experiments on the effects of question wording, order, and context date back to the dawn of survey research (Schuman and Presser 1981) but became ubiquitous as new technologies (such as computer-assisted telephone interviewing (Sniderman and Grob 1996) and web-based surveys (Birnbaum 2004; Iyengar 2011)) and new institutions (such as Time-sharing Experiments for the Social Sciences (Sniderman 2011), the Cooperative Congressional Election Study (Vavreck and Rivers 2008), and Amazon's Mechanical Turk (Paolacci, Chandler, and Ipeirotis 2010; Mason and Suri 2011)) dramatically reduced the cost of experimentation. Survey experiments now suffuse every substantive domain. To cite but a few examples, experimental variations in visual imagery are used to prime social attitudes (Greenwald and Banaji 1995); experimental variations in question format are used to detect opinions or information that respondents might otherwise conceal, such as racial prejudice or illegal conduct (Droitcour et al. 1991; Kuklinski, Cobb, and Gilens 1997); and experimental variations in hypothetical scenarios are used to study the role of domestic audience costs in foreign policy decision-making (Tomz 2007).

The rapid development of experimental data collection has in some ways outpaced the technical sophistication of experimental data analysis. One area in which the gap is particularly evident is the study of treatment effect heterogeneity. Experimental researchers routinely search for interactions between their randomly-assigned treatments and the rich set of background attributes that surveys often provide. Although the study of interactions has the potential to reveal important insights about when, why, and for whom a treatment works, in practice the investigation of treatment effect heterogeneity often seems ad hoc and unstructured.

The standard approach for investigating treatment effect heterogeneity is to estimate conditional average treatment effects (CATEs), or average treatment effects among subgroups defined by baseline covariates. For example, a researcher might split her sample by gender and then estimate CATEs separately for men and women. While such subgroup analysis seems unobjectionable, it introduces some subtle statistical concerns. One problem

is that the credibility of CATE estimates diminishes when researchers get to choose how to divide their data into subgroups. A related problem is that researchers conducting many subgroup analyses rarely adjust their standard errors for such multiple testing, ending up with downwardly biased standard errors. Finally, the curse of dimensionality makes the systematic exploration of subgroup-specific treatment effects impossible when the number of covariates is large or covariates are continuous. In this case researchers tend to rely on strong modeling assumptions, which are subject to uncertainty. This uncertainty, however, is seldom reflected in the standard errors that researchers report.

The aim of our paper is to explicate what we believe to be a better approach. Drawing on recent advances in statistical learning, we recommend the use of Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch 2007, 2010) as a method for analyzing treatment effect heterogeneity in survey experiments. BART is well suited to survey experiments, which typically comprise large samples for which extensive covariate information is available. As we explain fully below, BART has several important advantages over both parametric regression models and competing statistical learning techniques such as random forests, boosting, neural nets, and support vector machines (Izenman 2008; Hastie, Tibshirani, and Friedman 2009). These advantages include BART's ability to *automatically* detect and model non-linear relationships and interactions, which together with BART's insensitivity to the choice of tuning parameters removes researchers' discretion from the data analysis. The next section of our paper discusses the estimation of CATEs using the potential outcomes framework. We then introduce BART and discuss some of its properties. Finally, we present an empirical example—a survey experiment on welfare attitudes from the General Social Survey (GSS)—to illustrate how BART can be used to detect and model heterogeneous treatment effects in a principled and robust manner.[1]

[1]We conducted all computations using a slightly customized version of the easy-to-use BayesTree `R` package (Chipman and McCulloch 2009). Unlike the original BayesTree package, our version allows posterior simulation for datasets with tens of thousands of observations. The original BayesTree package is available from the Comprehensive R Archive Network; the source code and Windows binaries for our customized version as well as detailed example code can be downloaded from our websites.

1.1. Potential outcomes framework

Our approach to the estimation of CATEs is best explicated using the potential outcomes framework (Holland 1986). In the case of a binary treatment, subject i has two potential outcomes, denoted $Y_i(1)$ for the outcome under treatment and $Y_i(0)$ for the outcome under control. We observe only one of these two potential outcomes for each subject, depending on whether the subject receives the treatment or the control. We define an indicator for the randomly assigned treatment, D_i, which takes the value 1 if subject i receives the treatment and 0 if i receives the control. The observed outcome is written as $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$. The effect of the treatment on subject i is defined as $\beta_{Di} = Y_i(1) - Y_i(0)$. Experimenters typically focus on the average treatment effect (ATE), $\mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(\beta_{Di})$.

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

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

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 covariates to the model allows both the outcome under control and the treatment effect to vary with these covariates, so that

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

where X_i is a vector of covariates, i.e., variables that are unaffected by treatment assignment. This formulation distinguishes between two components that together comprise the causal effect for subject i. The first component $(\beta_D + \beta_{DX} X_i)$ is a *systematic* component that describes how the treatment effect varies with covariates. 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. The regression model in (2) then 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 covariates but are invariant within strata defined by covariate values (Imbens 2004). The regression model in (3) then becomes

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

The model in (5) is significantly more flexible than equation (4), especially when an extensive set of covariates is used. In what follows we estimate models similar to (5), allowing for systematic but not idiosyncratic treatment effect heterogeneity.[2] When there is systematic treatment effect heterogeneity we will say that the covariates "moderate" the treatment effect.[3]

Equation (5) suggests that we can estimate CATEs by means of a regression model that includes a binary treatment indicator (D_i), covariates (X_i), and interaction terms between the treatment indicator and the covariates. However, this approach is problematic for several reasons. To begin with, parametric models such as linear or logistic regression rest on functional form assumptions about the relationship between the outcome and the covariates. The incorrect modeling of these functional forms can lead to biased CATE estimates (Royston and Sauerbrei 2004; Feller and Holmes 2009; Imai and Strauss 2011). The inclusion of a large number of interaction terms can also lead to multicollinearity problems and imprecise inferences. The usual regression diagnostics and goodness-of-fit tests can aid

[2]Note that the existence of idiosyncratic heterogeneity does not imply that CATE estimates are biased. Similar to ATE estimates, which are unbiased (but potentially uninformative) in the presence of treatment effect heterogeneity, CATE estimates are unbiased even when there is idiosyncratic treatment effect heterogeneity. In this case CATEs simply fail to provide a complete picture of treatment effect heterogeneity.

[3]This is a descriptive and not a causal statement. We are not claiming that the covariates have a causal effect on the outcome. We are merely describing how the causal effect of the randomly assigned treatment varies with covariate values. Since covariates (in contrast to the treatment) are not randomly assigned, the *substantive* interpretation of the finding that a given covariate moderates the treatment effect assumes the absence of unobserved moderating covariates correlated with the covariate of interest.

survey experimenters in finding a better fitting model, but since such "tinkering" is not blind to the outcomes, experimenters run the risk of unconsciously biasing their inferences. Experimenters might refrain from examining additional model specifications as soon as they find one that meets their prior expectations or bypass specifications that produce what they deem to be implausible results. Specification searches, while understandable, introduce an undesirable amount of discretion into the analysis of experimental data. Moreover, since standard errors are usually not adjusted for specification searches or multiple comparisons, they will understate the uncertainty of the resulting CATE estimates (Pocock et al. 2002).

In order to avoid these problems, we recommend that survey experimenters take advantage of recent advances in the statistical learning literature and use BART to model systematic treatment effect heterogeneity in randomized survey experiments (Chipman, George, and McCulloch 2007, 2010). As we explain in detail in the next section of the paper, BART automatically detects and models the (possibly highly non-linear) relationship between the outcome variable and the predictors, including interactions between predictor variables. BART is thus able to learn interesting features from complex, high-dimensional data that parametric regression models have little hope of discovering. Since BART models the outcome variable in an extremely flexible manner, it also reduces the risk of biased CATE estimates due to model misspecification. Another appealing aspect of BART is that it minimizes the role of discretion in the analysis of experimental data. When using BART, survey experimenters do not need to make any of the specification choices that are required when fitting parametric methods. The only input required by BART is a vector of outcomes and a matrix of covariates. Moreover, in contrast to many other statistical learning techniques, BART's performance is largely unaffected by the choice of tuning parameters, making it an attractive "off-the-shelf" tool for experimenters interested in flexibly modeling systematic treatment effect heterogeneity.[4]

[4]Chipman, George, and McCulloch (2010), Zhang and Härdle (2010), and Bonato et al. (2011) show that BART performs well compared to other statistical learning techniques such as neural nets, boosting, random forests, and the lasso.

2. Bayesian Additive Regression Trees (BART)

We now explain how BART works, focusing on the intuition behind it. Our exposition closely follows Chipman, George, and McCulloch (2010) and Chipman et al. (2010), which provide additional technical details. Hill (2011) is an excellent introduction to BART and demonstrates its usefulness for nonparametric causal inference in observational studies with ignorable treatment assignment.

BART builds upon regression and classification tree models (Breiman et al. 1984; Izenman 2008; Hastie, Tibshirani, and Friedman 2009). Tree models explain variation in an outcome variable by repeatedly splitting the sample into ever more homogeneous subgroups. 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 always split into two daughter nodes. These splits are based on Boolean (i.e., "yes"/"no") questions about a single predictor; for example, is $X_i \leq \theta_j$?, where X_i is the value of a predictor variable for observation i and θ_j is a threshold value. Each observation in the node is assigned to one of the two daughter nodes depending on whether the answer is "yes" or "no" for that observation.

Figures 1 and 2 provide an illustration by means of a simple hypothetical example featuring a randomly assigned treatment variable and one continuous covariate. There are 200 observations, generated as follows: $D_i \sim \text{Bernoulli}(.5)$, $X_i \sim \text{Normal}(4, 1)$, $Y_i(0) \sim \text{Normal}(X_i^2, 4)$, and $Y_i(1) \sim Y_i(0) + X_i^2 + \text{Normal}(0, 1)$. The observed outcomes are equal to $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$. These simulated data are shown in Figure 1, with circles denoting control observations and triangles denoting treated observations. The two solid black curves represent the true response curves $\mathbb{E}(Y(0) \mid X)$ and $\mathbb{E}(Y(1) \mid X)$. The vertical distance between these two response curves at any X value represents the CATE. It is apparent that larger X values are associated with larger CATEs. The long-dashed curves show the fit of a single tree model to these data; the dashed curves show the BART fit (they are somewhat hard to see since they closely hug the true response curves).

Figure 2 displays the tree that generated the single tree fit in Figure 1 (the two long-dashed curves). At the top of Figure 2 is the root node, which contains all 200 observations. These observations are first queried about their X value. Observations with $X < 4.103$ drop down the tree to the left daughter node D1; all other observations drop down the tree to the right daughter node D2. Observations in D1 are then queried about their treatment status. Control observations drop down to daughter node D3; treated observations drop down to daughter node D4. Next, observations in D3 are again queried about their X value, with observations having an X value of less than 3.230 dropping down into terminal node N1 and the rest dropping down into terminal node N2. The same thing happens in daughter nodes D2 and D4, which assign the observations they contain to terminal nodes N3 through N6.

The tree in Figure 2 has partitioned the data into 6 terminal nodes, N1 through N6. Each observation has been assigned to one, and only one, of these terminal nodes. N1, for example, contains all observations that satisfy $X_i < 3.230$ and $D_i = 0$. Fitted values are assigned to each of the terminal nodes based on the mean outcome among observations falling into that terminal node. For example, the fitted value for observations in terminal node N1 is equal to 6.626. In Figure 1, this fitted value is represented by the first horizontal stretch of the lower long-dashed curve on the left side of the graph.

One drawback of single tree models is that their piecewise-constant fit leads to a lack of smoothness of the fitted 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 tree models can model complicated interactions, they are not well suited for modeling simple additive structures (Hastie, Tibshirani, and Friedman 2009). From Figure 1, it is obvious that BART achieved a much smoother fit to the true response curves than the single tree model. We now explain how BART is able to achieve such good fits by generating a large number of trees and then combining their predictions.

BART can be used to model binary or continuous outcomes. We start by describing how BART handles the latter case. BART models a continuous outcome Y as an unknown

function f of a p-dimensional vector of predictors $x = (x_1, \ldots, x_p)$ plus an iid error term:

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

We start by introducing notation for a single tree. Let T denote a tree consisting of a set of interior nodes, a set of terminal nodes, and the decision rules connecting these nodes. In other words, T captures all the information necessary to draw a tree such as the one shown in Figure 2. Let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denote a set of parameters associated with the b terminal nodes of T. In a single tree model, these b parameters represent the fitted values for the b terminal nodes. Given T and M, we define the output of the function $g(x; T, M)$ as the value obtained by first dropping an observation with characteristics x down the tree until it hits a terminal node and then reporting the μ_z associated with that terminal node:

$$Y = g(x; T, M) + \epsilon, \quad \epsilon \sim \text{Normal}(0, \sigma^2). \tag{7}$$

Under (7), $\mathbb{E}(Y \mid x)$ equals the terminal node parameter μ_z assigned by $g(x; T, M)$.

Instead of merely fitting one tree, BART fits an ensemble of m trees, with m typically in the hundreds. This sum-of-trees model can be written as

$$Y = \left(\sum_{j=1}^{m} g(x; T_j, M_j) \right) + \epsilon, \quad \epsilon \sim \text{Normal}(0, \sigma^2). \tag{8}$$

For each tree T_j and its associated set of terminal node parameters M_j, the output of $g(x; T_j, M_j)$ is the value obtained by dropping an observation with characteristics x down the tree until it hits a terminal node and then reporting the appropriate terminal node parameter $\mu_{zj} \in M_j$. Under (8), $\mathbb{E}(Y \mid x)$ equals the sum of all terminal node parameters assigned to an observation with characteristics x by $g(x; T_j, M_j)$. Note that each μ_{zj} represents a main effect when $g(x; T_j, M_j)$ depends on only one component of x. When $g(x; T_j, M_j)$ depends on more than one component of x, μ_{zj} represents an interaction effect. In this way, BART naturally incorporates both main and interaction effects since some of the trees can represent main effects while others represent interactions. And since the trees in (8) are allowed to be of varying sizes, the interactions may be of varying orders. The ensemble of trees in (8) is capable of producing extremely flexible fits since each of the

m trees can specialize in fitting one particular aspect of the data (such as a complicated interaction).

In contrast to algorithmic approaches to statistical learning (Breiman 2001), BART treats (T_j, M_j) and σ as parameters in a formal statistical model. A prior is put on the parameters and the posterior is computed using Markov Chain Monte Carlo (MCMC). At each iteration of the MCMC algorithm (see Chipman, George, and McCulloch 2010 for computational details), (T_j, M_j) and σ are redrawn. For large m, there are hundreds of parameters of which only σ is identified. (To see this, note that swapping (T_1, M_1) with (T_2, M_2) in (8) produces a different parametrization but leads to an identical fit.) This lack of identification poses no difficulty since the individual tree parameters are of no intrinsic interest.[5]

In order to rein in the huge number of free parameters so as to avoid overfitting the data, a regularization prior is put on σ as well as each tree T_j and its terminal node parameters M_j. The goal is to keep the contribution of each tree small, so that many different trees have a chance to contribute to the fit, preserving the flexibility of (8). The complexity of the prior formulation problem is reduced by letting T_j be iid, μ_{zj} be iid conditional on T_j, and σ be independent of all T and μ. It then only becomes necessary to specify marginal priors for a single T, a single μ, and σ. The prior on T puts larger weight on small trees while the prior on μ shrinks the fit of each terminal node towards zero. The amount of shrinkage increases with m, so that the contribution of each tree decreases as the number of trees increases. Chipman, George, and McCulloch (2010) provide default settings for these priors that they show to be highly effective in producing a very flexible fit that simultaneously avoids overfitting the data.[6] We have found these default priors to

[5]It also helps the MCMC algorithm to mix well so that convergence is usually not a problem (Chipman, George, and McCulloch 2010). We have found that in various datasets a burn-in period of $1,000$ draws followed by $1,000$ draws from the posterior to compute CATEs provided good results.

[6]For the prior on T, Chipman, George, and McCulloch (2010) use the specification given in Chipman, George, and McCulloch (1998) for a Bayesian single tree model. There, the probability that a node is non-terminal is $\alpha(1+d)^{-\beta}$, where d is the depth of the node. The default prior sets $\alpha = .95$ and $\beta = 2$, so that the prior distribution of the number of terminal nodes has probabilities of 0.05, 0.55, 0.28, 0.09, and 0.03 for trees of size 1, 2, 3, 4, and ≥ 5. However, even with the prior putting larger weight on small trees, large trees can be grown if required by the data. At any non-terminal node, the prior on the associated decision rule puts equal probability on each predictor and each possible split given the predictor. To place a prior on μ,

work well in a wide variety of datasets. As we will illustrate in the next section of the paper, there is normally little to be gained from using cross-validation to choose the number of trees and the regularization prior. The fact that BART's performance is largely unaffected by the choice of these tuning parameters means it can be easily used "off-the-shelf," greatly enhancing its appeal.

Up to this point we have discussed the use of BART with continuous outcomes. BART can also be used when outcomes are binary, as in the empirical examples we present in this paper. The probit version of BART can be written as

$$p(x) \equiv P(Y = 1 \mid x) = \Phi\left[\sum_{j=1}^{m} g(x; T_j, M_j)\right], \tag{9}$$

where $\Phi[\cdot]$ denotes the standard normal cdf. The regularization prior for probit BART is very similar to the one discussed above, except that the model sets σ to 1 so that only priors on (T_j, M_j) are needed.[7]

In contrast to parametric regression models, a BART fit cannot be summarized by a small number of regression coefficients. Instead, one uses simulation to generate CATE estimates.[8] Let N be the number of observations and K the number of predictors including the treatment indicator. We run BART by supplying it with an $N-$vector of outcomes and an $N \times K$ matrix of covariates that also includes the treatment indicator. In order to estimate CATEs, we ask BART to generate posterior draws for synthetic observations that we will describe next. We construct two new data matrices identical to the original

the outcome variable is standardized so that $\mathbb{E}(Y \mid x)$ falls in the interval $(-0.5, 0.5)$ with high probability and let $\mu \sim \text{Normal}(0, \sigma_\mu^2)$. Conditional on the set of T_j, $\mathbb{E}(Y \mid x) = \sum_{j=1}^{m} g(x; T_j, M_j) = \sum_{j=1}^{m} u_j$, where each u_j corresponds to the terminal node parameter from one of the m trees. The standard deviation of this sum equals $\sqrt{m}\sigma_\mu$. Setting $\sigma_\mu = \frac{0.5}{k\sqrt{m}}$, $\pm.5$ is within k standard deviations of zero. Chipman, George, and McCulloch's (2010) default choice for k is $k = 2$, so that there is a 95% prior probability that $\mathbb{E}(Y \mid x)$ is in $(-0.5, 0.5)$. For the prior on σ, let $\sigma^2 \sim \frac{\nu\lambda}{\chi_\nu^2}$, where χ_ν^2 denotes a chi-squared random variable with ν degrees of freedom. Chipman, George, and McCulloch (2010) choose ν and λ so that $P(\sigma < \hat{\sigma}) = q$. The default sets $\hat{\sigma}$ equal to the linear regression estimate, $q = 0.9$, and $\nu = 3$, and chooses λ so that there is 90% prior probability that σ is smaller than the linear regression estimate. See Chipman, George, and McCulloch (2010) for a detailed discussion of these priors.

[7] The prior on T is the same as above, and the prior on μ takes the form $\mu \sim \text{Normal}(0, \sigma_\mu^2)$, where $\sigma_\mu = \frac{3.0}{k\sqrt{m}}$. The value of k is chosen so that $p(x)$ will with high probability be in the interval $(\Phi[-3.0], \Phi[3.0])$. Chipman, George, and McCulloch (2010) recommend $k = 2$ as default choice.

[8] See King, Tomz, and Wittenberg (2000) and Gelman and Hill (2007) for a discussion of the simulation of quantities of interest.

$N \times K$ data matrix, except that the value of the covariate of interest is set to one of the 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 $1,000$ iterations, we take $1,000$ posterior draws of the predicted value 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 (observations) of each matrix, which results in two vectors of average predicted values with $1,000$ elements each. Subtracting these vectors from each other produces $1,000$ posterior draws of the CATE at the specified covariate value. The mean of these posterior draws is the estimated CATE; the .025 and .975 quantiles represent 95% posterior uncertainty bounds. We create such pairs of matrices for each unique covariate value to estimate CATEs at all observed values of the covariate.[9] We repeat this procedure for all covariates for which we want to estimate CATEs. It is straightforward to investigate higher-order interactions between the treatment and several covariates by varying multiple covariates at the same time.

3. Empirical example

A well-known survey experiment from the GSS illustrates how BART can be used to model systematic treatment effect heterogeneity. The ATE estimate in this experiment is sizable and well-established by replication studies. The sample size is large enough to allow us to investigate systematic treatment effect heterogeneity with ample statistical power. Moreover, the experiment has prompted other researchers to investigate treatment-covariate interactions.

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 subject of sustained experimental inquiry in the mid-1980s, when the GSS included a question wording experiment in its national survey of adults. Respon-

[9] For continuous covariates, we use a number of more or less equally spaced observed values covering the observed range of the covariate.

dents 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" and in the other they are asked about "assistance to the poor."[10]

This seemingly innocuous variation in question wording has a profound effect on support for government spending in this domain. Using GSS surveys from 1986–2010, Table 1 compares the proportion of respondents stating that "too much" is being spent on either welfare or assistance to the poor.[11] In each survey wave, the question wording experiment generates a huge and highly statistically significant ATE estimate. Estimates range from 27.4 percentage points to 50.7 percentage points.

The magnitude and robustness of this question wording effect have 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 systematic 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 such covariates moderate the question wording effect.

[10] "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 . . . ?"

[11] 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 characteristics (Smith and Peterson 1986). Also note that the GSS is not conducted every year.

3.1. Results

Using the sample difference-in-means, we estimate the ATE to be equal to 0.364 with a 95% confidence interval of $(0.351, 0.377)$. Because the outcome variable is binary, we use a probit BART model with BART's default tuning parameter settings to search for systematic treatment effect heterogeneity. Cross-validation confirms that the BART fit is insensitive to the specific choice of tuning parameters (see Figure 3).[12] In this example and many others we have examined, the cross-validation misclassification error rate is essentially constant over a wide range of tuning parameter settings, justifying the use of BART's default settings.[13]

Figure 4 displays CATE estimates. The dark grey areas are point-wise 95% posterior bands; the light grey areas are global 95% posterior bands that simultaneously account for uncertainty in all CATE estimates (Mandel and Betensky 2008). Marginal covariate distributions are shown at the bottom of each graph. The model includes the following covariates: 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.[14] The BART fit reveals substantive treatment

[12]Cross-validation randomly partitions a dataset into 5 or 10 subsets, fits the model to all subsets except one, and then evaluates the model fit by comparing the model's out-of-sample predictions with the observed outcomes in the omitted subset (Izenman 2008). This procedure is repeated for each subset and the misclassification errors are averaged. We varied three tuning parameters: the number of trees (m), the prior on the complexity of the trees (β), and the prior on the contribution of an individual tree to the overall fit (k).

[13]We did find that even though point estimates are generally unaffected by the choice of tuning parameters, a smaller number of trees sometimes leads to more precise inferences. If we use 50 instead of the default 200 trees in our analysis of the GSS data, the posterior intervals reported below shrink somewhat, but our substantive conclusions are unaffected. We also compared BART's fit to the fit from eight parametric probit models using various plausible combinations of squared and interaction terms. BART had a slight edge over all of these models in terms of cross-validated area under the receiver operating characteristic (ROC) curve (see King and Zeng 2001 for a discussion of ROC curves). More importantly, the use of BART removed all discretion from the data analysis since it did not require us to commit to one particular model specification.

[14]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

effect heterogeneity. The graphs at the top of Figure 4 display CATE estimates as a function of party identification and liberal-conservative self-placement. Both graphs show strong moderation of the treatment effect: as respondents become more Republican or more conservative, the question wording effect increases. The estimated difference in treatment effects between respondents at either end of the scale is about 10 percentage points for party identification and 11 percentage points for liberal-conservative self-placement. (Note that these CATE estimates represent the moderating effect of party identification or liberal-conservative self-placement while controlling for all other covariates in the model.) We can formally test for treatment effect heterogeneity by conducting a Wald test based on the CATE estimates and their estimated variance-covariance matrix (Cameron and Trivedi 2005). For both party identification and liberal-conservative self-placement, we decisively reject the null hypothesis that the CATEs are identical ($p < .001$). We also find some moderation of the treatment effect by age, with effect sizes that are somewhat smaller among older respondents ($p = 0.029$). Interestingly, there is no evidence that education moderates the treatment effect conditional on the other covariates ($p = 0.99$). Treatment effects are most strongly moderated by respondents' negative attitudes toward blacks, with CATE estimates increasing monotonically with increases in the negative attitudes scale. The CATE estimate is .26 for respondents who score a zero on the negative attitudes scale and .42 for respondents who score a one, for an estimated difference of 16 percentage points between respondents at either end of the scale ($p < .001$). Finally, the bottom right graph of Figure 4 shows how estimated CATEs vary with time. We find that the treatment effect varies significantly across years ($p < .001$), with particularly large effect sizes in 1993–96, coinciding more or less with the first Clinton administration. These results confirm the longstanding hypothesis that imagery associated with welfare has racial connotations but call into question some of the complex interactions that others have reported. For example, we find little evidence that education moderates the treatment effect, either alone or in combination with negative attitudes toward blacks.[15]

all responses. When an individual's responses were partially missing we used the remaining responses to construct the index for this individual.

[15]The graphs in Figure 4 display CATE estimates one covariate at a time. Treatment effects could

The individual graphs in Figure 4 help us visualize how the effect of the question wording change varies with each covariate, but they do not allow us to judge the overall amount of systematic treatment effect heterogeneity in the data. From the histogram of CATE estimates in Figure 5, we can see that the treatment effect varies enormously across respondents. Estimates range from 5 percentage points to 61 percentage points, with the median estimated CATE equal to 37 percentage points. Substantively, it is interesting to note that even the smallest CATE estimates are positive, which means that every subgroup defined by our covariates responds more favorably to "assistance to the poor" than "welfare." Although CATE estimates vary dramatically, it appears as if everyone in the sample is moving in the same direction in response to the treatment.

How much of a difference does allowing for systematic treatment effect heterogeneity make? We can judge the importance of systematic treatment effect heterogeneity in these data by randomly permuting individuals' covariate vectors (leaving the outcome and treatment indicator unchanged) and re-running BART. This permutation scheme destroys any systematic relationship between the covariates and the outcome (and therefore also any systematic treatment effect heterogeneity) but leaves the relationships between the covariates intact.

Figure 6 shows a kernel density plot of CATE estimates (black curve). Additionally, it shows 10 kernel density plots of CATE estimates when individuals' 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 in any of the 10 permuted datasets. Clearly, our covariates contain valuable information about systematic treatment effect heterogeneity that we would fail to exploit if we were to solely focus on the ATE.[16]

potentially be a more complex function of several covariates, leading to three-way or even higher order interactions between the treatment indicator and the covariates. BART automatically incorporates such higher-order interactions in the model if they improve the fit. We examined all three-way interactions between the treatment indicator and the covariates but could never reject the null hypothesis that covariates do not jointly moderate the treatment effect.

[16]In the non-linear probit BART model, CATEs necessarily vary with an individual's covariate values, even if these covariates are not interacted with the treatment indicator. This "compression" effect can increase or decrease the treatment effect heterogeneity beyond that visible on the scale of the linear predictor

4. Conclusion

Survey experiments are highly valued for their ability to generate unbiased treatment effect estimates. In an attempt to uncover systematic variation in treatment effects, survey experimenters routinely search for interactions between their randomly-assigned treatments and the rich set of background attributes that surveys often provide. However, parametric approaches currently used to search for systematic treatment effect heterogeneity suffer from a variety of shortcomings. Perhaps most importantly, they introduce an undesirable amount of discretion into the analysis of experimental data. In an effort to address these shortcomings, we recommend that survey experimenters draw on recent advances in the statistical learning literature and use Bayesian Additive Regression Trees (BART) to model systematic treatment effect heterogeneity. BART has a number of important advantages over parametric modeling strategies such as its ability to automatically detect and model non-linear relationships, including interactions between predictor variables, and its insensitivity to the choice of tuning parameters. These features make BART an appealing tool for survey experimenters who do not have the experience to confidently employ more difficult-to-use statistical learning techniques such as neural networks but still want to model treatment effect heterogeneity in a flexible and robust manner.

Recent years have seen a dramatic increase in the number and scale of survey experiments conducted in a wide variety of fields, and there are indications that researchers are becoming increasingly sensitive to the challenges involved in drawing robust causal inferences from such experiments (e.g., Gaines and Kuklinski 2007). Both the supply of high-quality survey experiments and the demand for safeguards against data dredging recommend the approach described here. Advances in computing power and the availability of easy-to-use public domain software now make computationally intensive estimators like

(Berry, DeMeritt, and Esarey 2010). Although we think that the probability scale is appropriate for reporting CATE estimates when outcomes are binary, we can also look at treatment effect heterogeneity on the scale of the underlying linear predictor, which is unaffected by compression (results not shown but available upon request). Even on this scale, we find that the range of CATE estimates in our original analysis is much larger than the range of CATE estimates in any of the permuted datasets. In other words, the systematic treatment effect heterogeneity visible in Figure 4 is not purely due to the fact that probabilities are bounded between zero and one.

BART readily available to survey experimenters. In the years ahead, as estimators like BART come into currency, we are likely to see a fundamental change in the way in which survey experimenters investigate and report systematic treatment effect heterogeneity in their survey experiments.

5. References

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.

Birnbaum, Michael H. 2004. "Human research and data collection via the internet." *Annual Review of Psychology* 55: 803–32.

Bonato, Vinicius, Veerabhadran Baladandayuthapani, Bradley M. Broom, Erik P. Sulman, Kenneth D. Aldape, and Kim-Anh Do. 2011. "Bayesian ensemble methods for survival prediction in gene expression data." *Bioinformatics* 27 (3): 359–367.

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

Breiman, Leo. 2001. "Statistical modeling: The two cultures (with discussion)." *Statistical Science* 16 (3): 199–215.

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. 2007. "Bayesian ensemble learning." In Bernhard Schölkopf, John Platt, and Thomas Hofmann (eds.). *Advances in neural information processing systems 19*. Cambridge, MA: MIT Press.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian additive regression trees." *Annals of Applied Statistics* 4 (1): 266–298.

Chipman, Hugh, Edward George, Jason Lemp, and Robert McCulloch. 2010. "Bayesian flexible modeling of trip durations." *Transportation Research Part B* 44 (4): 686–698.

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

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, Wendy Visscher, and Trena M. Ezzati. 1991. "The item count technique as a method of indirect questioning: A review of its development and a case study application." In

Paul P. Biemer et al. (eds.). *Measurement errors in surveys.* New York: John Wiley & Sons, pp. 185–210.

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."

Gaines, Brian J. and James H. Kuklinski. 2007. "The logic of the survey experiment reexamined." *Political Analysis* 15 (1): 1–20.

Gelman, Andrew and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

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

Greenwald, Anthony G. and Mahzarin R. Banaji. 1995. "Implicit social cognition: Attitudes, self-esteem, and stereotypes." *Psychological Review* 102 (1): 4–27.

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

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. 2011. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics* 20 (1): 217–240.

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

Imai, Kosuke and Aaron Strauss. 2011. "Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign." *Political Analysis* 19 (1): 1–19.

Imbens, Guido W. 2004. "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and Statistics* 86 (1): 4–29.

Iyengar, Shanto. 2010. "Experimental designs for political communication research." In Eric P. Bucy and R. Lance Holbert (eds.). *Sourcebook for political communication research: Methods, measures, and analytical techniques.* New York: Routledge, pp. 129–148.

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.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the most of statistical analyses: Improving interpretation and presentation." *American Journal of Political Science* 44 (2): 341–355.

King, Gary and Langche Zeng. 2001. "Improving forecasts of state failure." *World Politics* 53 (4): 623–58.

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.

Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. "Racial attitudes and the 'New South'." *Journal of Politics* 59 (2): 323–349.

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

Mason, Winter and Siddharth Suri. 2011. "Conducting behavioral research on Amazon's Mechanical Turk." SSRN Working Paper.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. "Running experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5 (5): 411–419.

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.

Schuman, Howard and Stanley Presser. 1981. Q*uestions and answers in attitude surveys: Experiments on question form, wording, and context.* New York: Academic Press.

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.

Sniderman, Paul M. and Douglas Grob. 1996. "Innovations in experimental design in attitude surveys." *Annual Review of Sociology* 22: 377–399.

Sniderman, Paul. 2011. "The logic and design of the survey experiment: An autobiography of a methodological innovation." In James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia (eds.). *Cambridge handbook of experimental political science.* Cambridge University Press.

Tomz, Michael. "Domestic audience costs in international relations: An experimental approach." *International Organization* 61 (4): 821–40.

Vavreck, Lynn and Douglas Rivers. 2008. "The 2006 Cooperative Congressional Election Study." *Journal of Elections, Public Opinion & Parties* 18 (4): 355–366.

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

Zhang, Junni L. and Wolfgang K. Härdle. 2010. "The Bayesian Additive Classification Tree applied to credit risk modelling." *Computational Statistics & Data Analysis* 54 (5): 1197–1205.

6. Tables

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

Year	Sample Size		Mean		$\widehat{ATE}$
	Assistance	Welfare	Assistance	Welfare	
1986	594	561	0.104	0.447	0.343
1988	404	359	0.079	0.451	0.372
1989	389	375	0.105	0.443	0.337
1990	536	510	0.086	0.424	0.338
1991	391	373	0.125	0.399	0.274
1993	418	418	0.148	0.598	0.450
1994	744	759	0.168	0.675	0.507
1996	704	700	0.217	0.639	0.421
1998	682	663	0.125	0.469	0.344
2000	635	664	0.131	0.413	0.282
2002	341	330	0.109	0.485	0.376
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
2010	497	500	0.111	0.456	0.345
Total/Mean	7,836	7,681	0.123	0.487	0.364

Source: General Social Survey 1986–2010. 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). In every year, average treatment effect estimates are statistically significant at the .001 level or better.

7. Figures

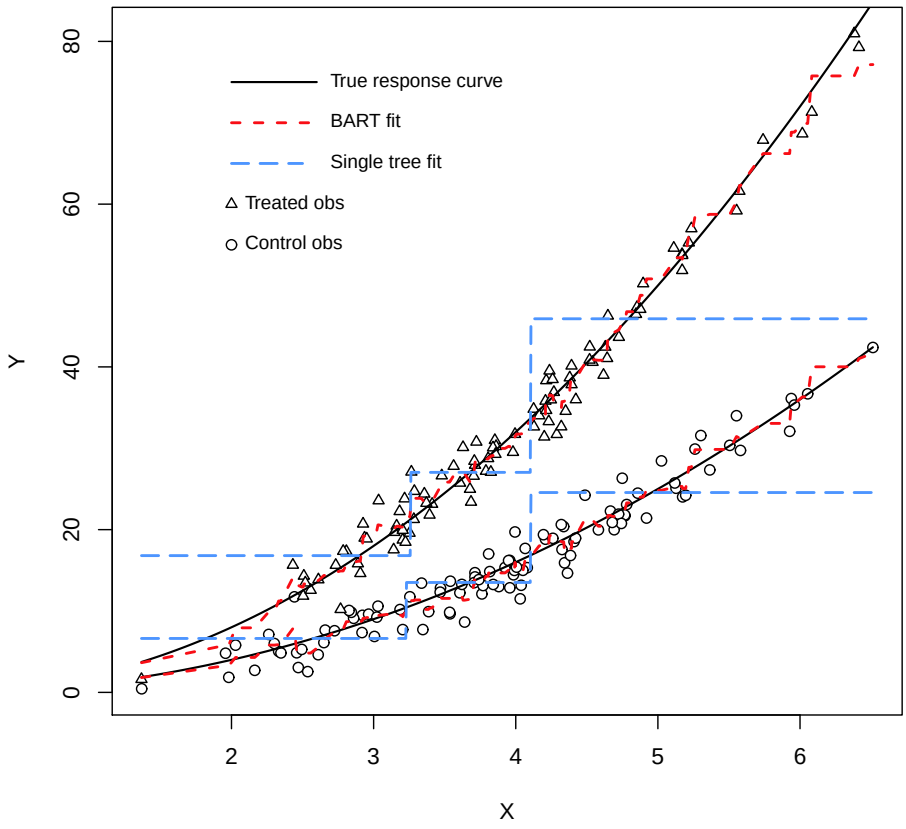

Figure 1:

The graph shows the fit from a single tree model and a BART fit to the simulated data described in the text.

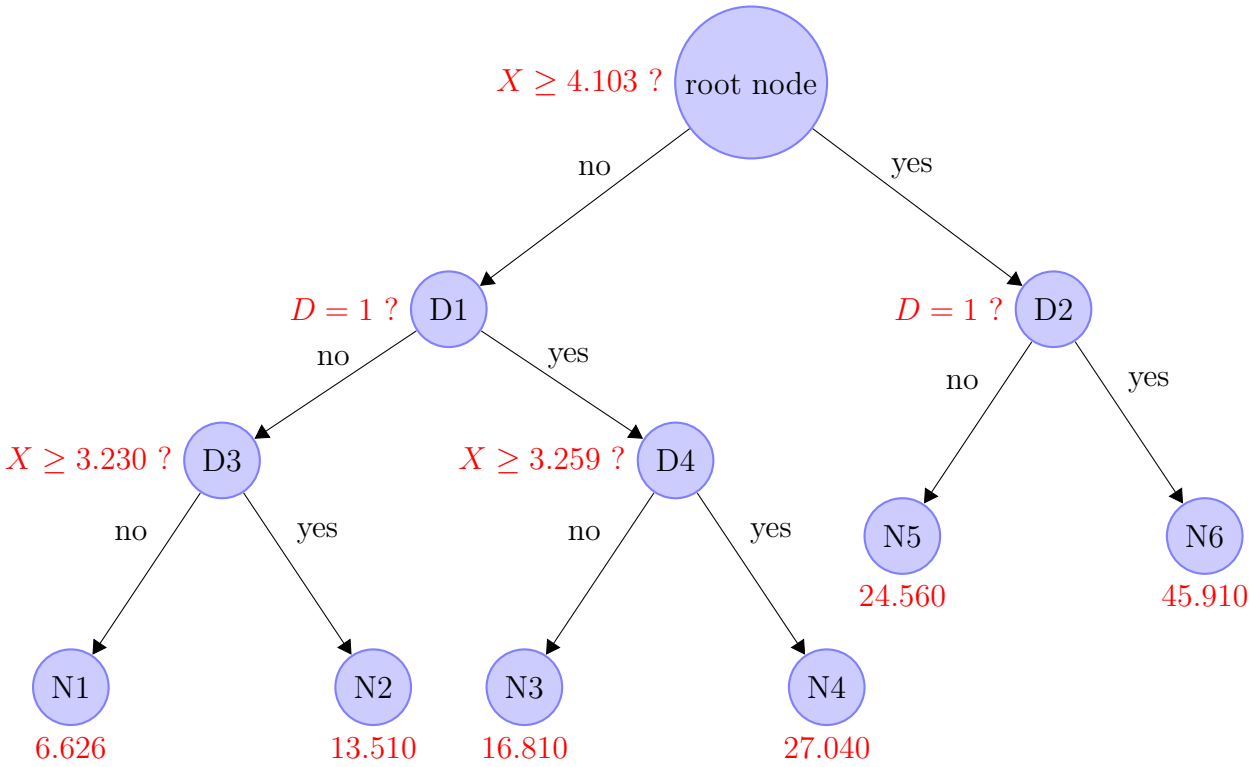

Figure 2:

The graph shows the tree corresponding to the single tree fit shown in Figure 1. Displayed next to the root node and the daughter nodes D1 through D4 are the associated decision rules. Displayed below the terminal nodes N1 through N6 are the fitted values associated with each terminal node.

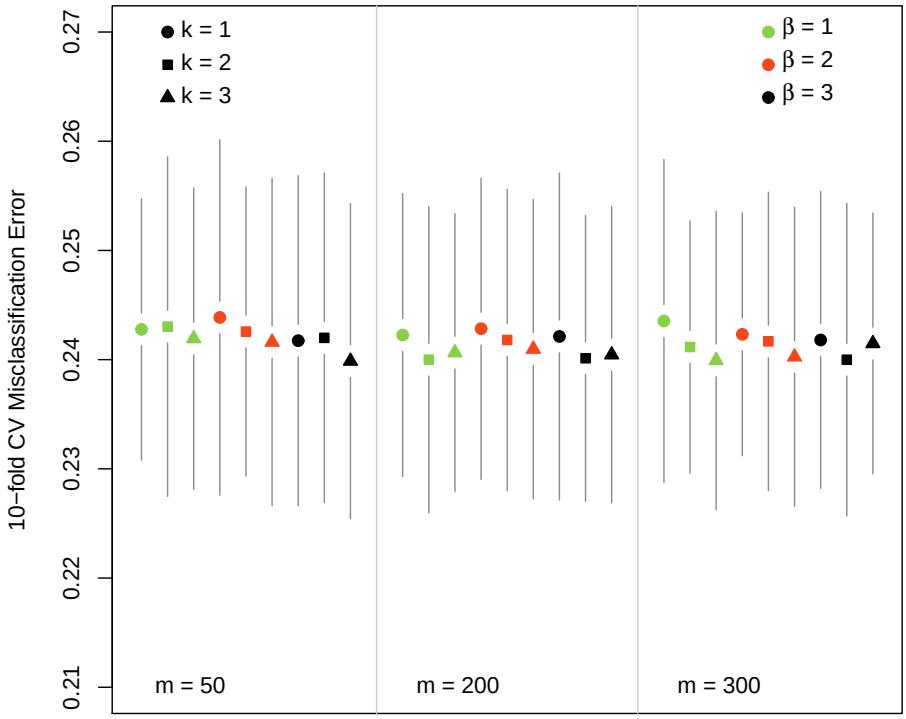

Figure 3:

Tenfold cross-validation misclassification error rates are shown for 27 different combinations of BART tuning parameters ($k \in \{1, 2, 3\}$, $\beta \in \{1, 2, 3\}$, and $m \in \{50, 200, 300\}$). The standard errors of the misclassification error rates are displayed as bars.

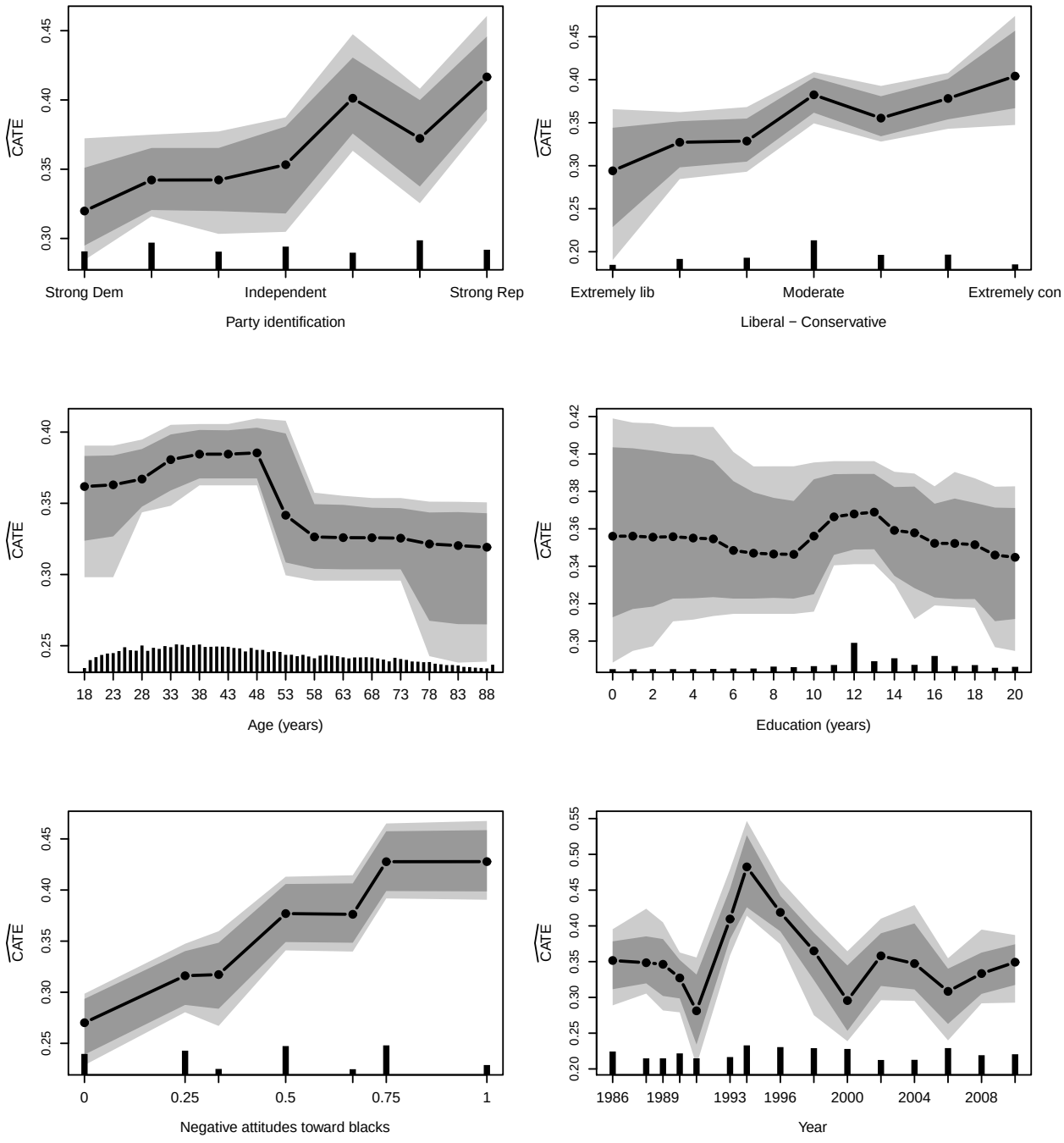

Figure 4:

Source: 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.

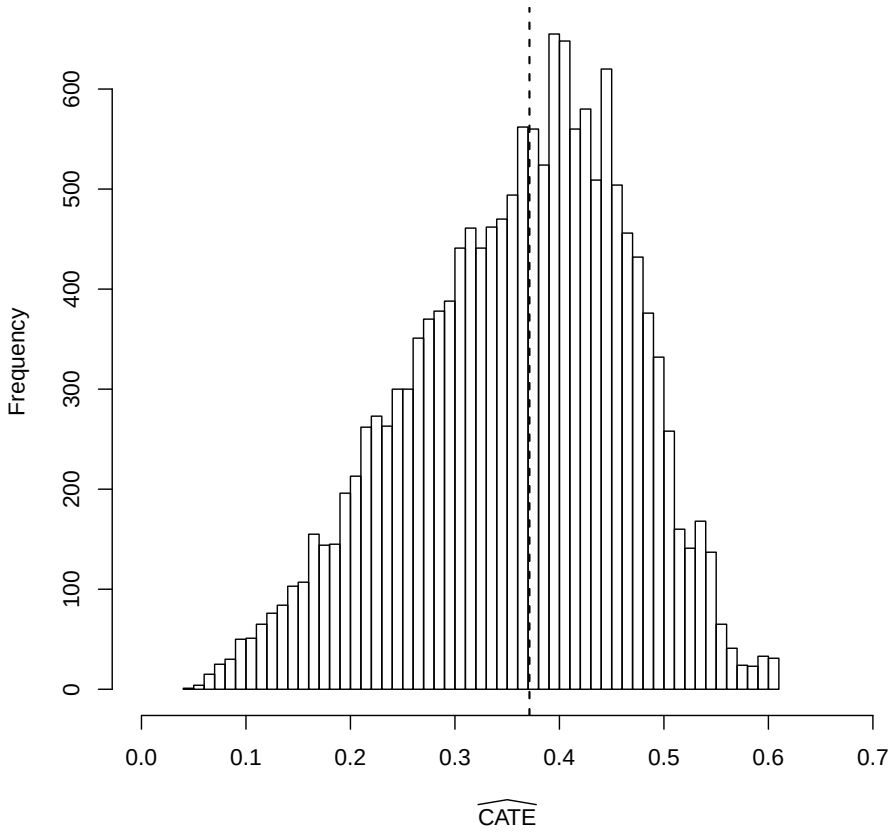

Figure 5:

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

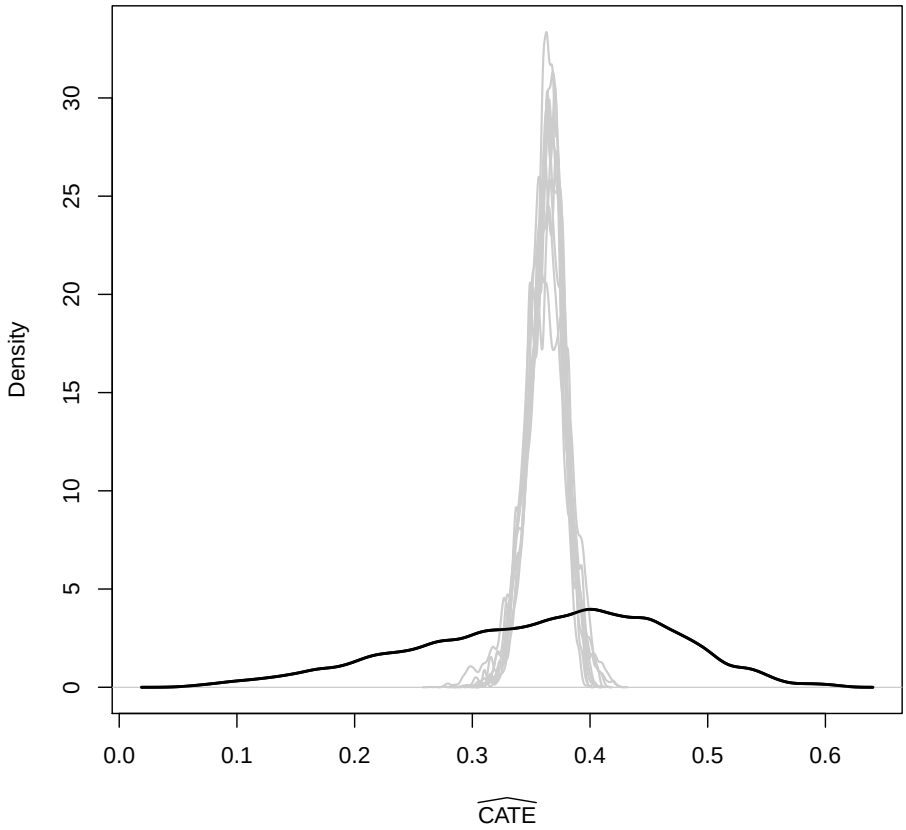

Figure 6:

Source: GSS 1986–2010. The graph displays a kernel density plot of CATE estimates (on the probability scale) and 10 kernel density plots when covariate vectors are randomly permuted (see text for details).