TREATMENT EFFECT HETEROGENEITY

In most experiments the main quantity of interest is the average treatment effect (ATE)

The ATE only provides a partial answer to the causal question of interest when treatment effects vary across experimental units (Cox 1958)

Randomized experiments only identify the two marginal outcome distributions in the treatment and control groups, which are generally insufficient to identify the joint distribution of outcomes. We would need this joint distribution to identify aspects of the treatment effect distribution other than its mean

ESTIMATING CATEs

Parametrically modeling conditional CATEs often involves additional functional form assumptions not justified by randomization (Feller and Holmes 2009)

Danger of post hoc data dredging: Researchers may look at many treatment-covariate interactions but selectively report only “interesting” heterogeneity (Pocock 2002; Gabler et al. 2009)

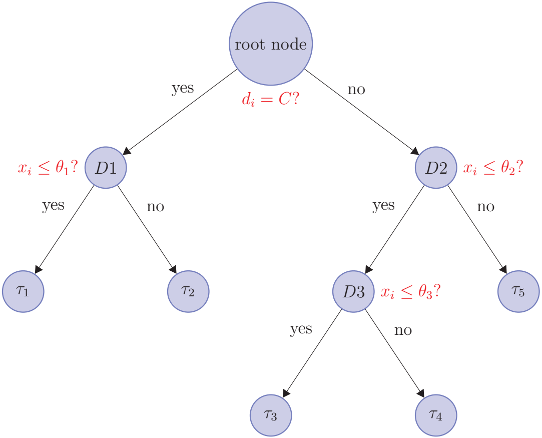
Multiple testing leads to measures of statistical uncertainty that are overly optimistic

OUR APPROACH

Estimate CATEs while addressing the post hoc data dredging, multiple comparisons, and model specification problems

We non-parametrically model CATEs using Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch 2007, Forthcoming; Hill 2010)

We clearly distinguish between an exploratory data analysis phase and a confirmatory phase (Tukey 1977) using a split-sample design


root node
$d_i = C?$
yes
no
$D1$
$x_i \leq \theta_1?$
$D2$
$x_i \leq \theta_2?$
yes
no
yes
no
τ_1
τ_2
$D3$
$x_i \leq \theta_3?$
τ_5
yes
no
τ_3
τ_4

NOTATION FOR SINGLE TREE MODEL

Let T denote a tree consisting of a set of (1) interior nodes and their associated decision rules and (2) a set of terminal nodes

Let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denote a set of parameter values associated with the b terminal nodes of T. μ_k represents the mean response of the subgroup of observations falling into terminal node k

NOTATION FOR SINGLE TREE MODEL

Let T denote a tree consisting of a set of (1) interior nodes and their associated decision rules and (2) a set of terminal nodes

Let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denote a set of parameter values associated with the b terminal nodes of T. μ_k represents the mean response of the subgroup of observations falling into terminal node k

For a given T and M, $g(x; T, M)$ denotes the function which assigns a $\mu_k \in M$ to an observation with covariate vector x. Thus,

$$Y = g(x; T, M) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{1}$$

NOTATION FOR SINGLE TREE MODEL

Let T denote a tree consisting of a set of (1) interior nodes and their associated decision rules and (2) a set of terminal nodes

Let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denote a set of parameter values associated with the b terminal nodes of T. μ_k represents the mean response of the subgroup of observations falling into terminal node k

For a given T and M, $g(x; T, M)$ denotes the function which assigns a $\mu_k \in M$ to an observation with covariate vector x. Thus,

$$Y = g(x; T, M) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{1}$$

SUM-OF-TREES MODEL

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

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 $\mu_{kj} \in M_j$ to an observation with covariate vector x

Each μ_{kj} 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_{kj} represents an interaction effect

SUM-OF-TREES MODEL

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

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 $\mu_{kj} \in M_j$ to an observation with covariate vector x

Each μ_{kj} 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_{kj} represents an interaction effect

A prior is put on the (T_j, M_j) and σ parameters, and the posterior is computed using MCMC. The prior favors small trees, restricting each tree to contribute only a small part to the overall fit

SUM-OF-TREES MODEL

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

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 $\mu_{kj} \in M_j$ to an observation with covariate vector x

Each μ_{kj} 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_{kj} represents an interaction effect

A prior is put on the (T_j, M_j) and σ parameters, and the posterior is computed using MCMC. The prior favors small trees, restricting each tree to contribute only a small part to the overall fit

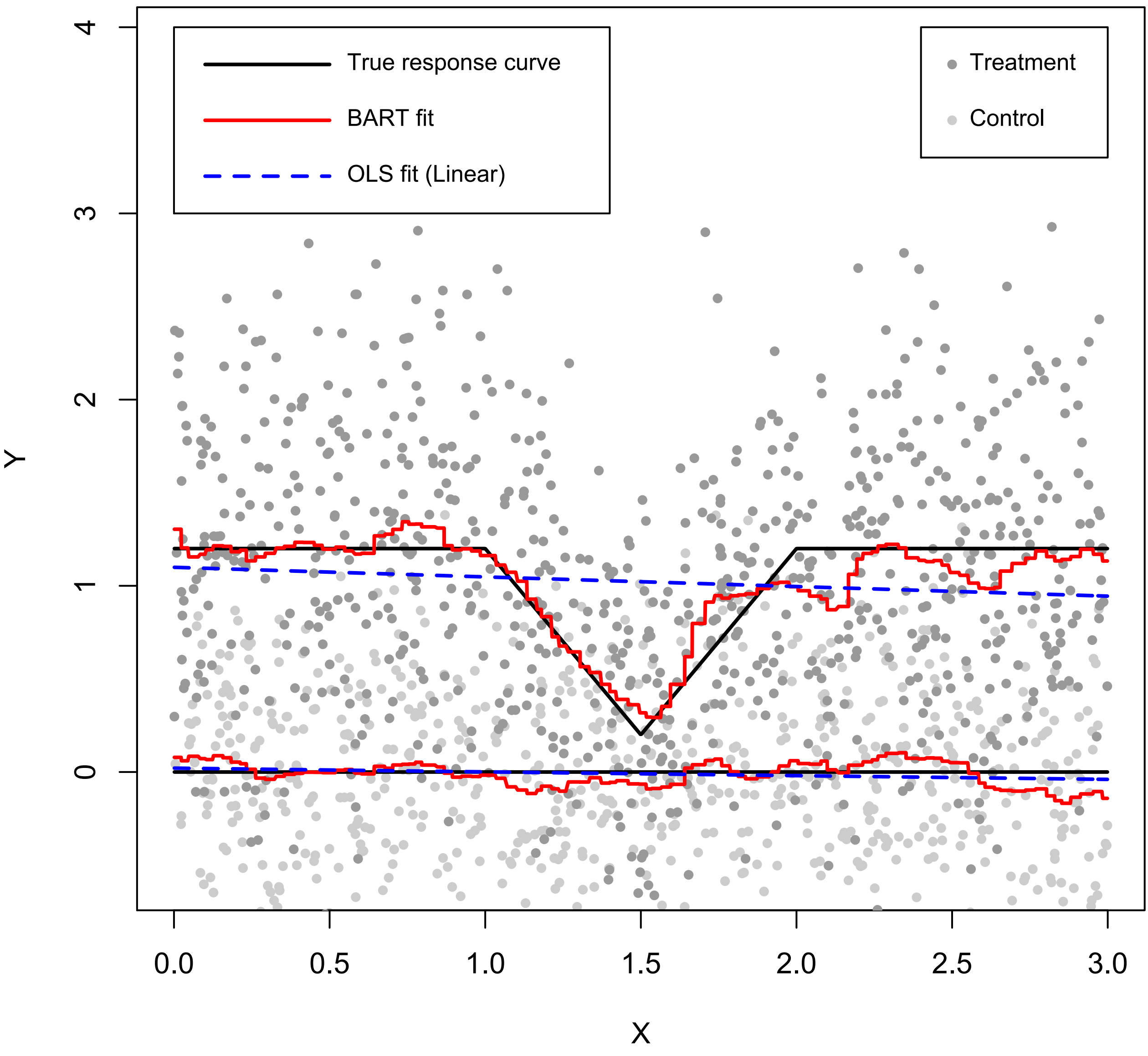
BART PROBIT

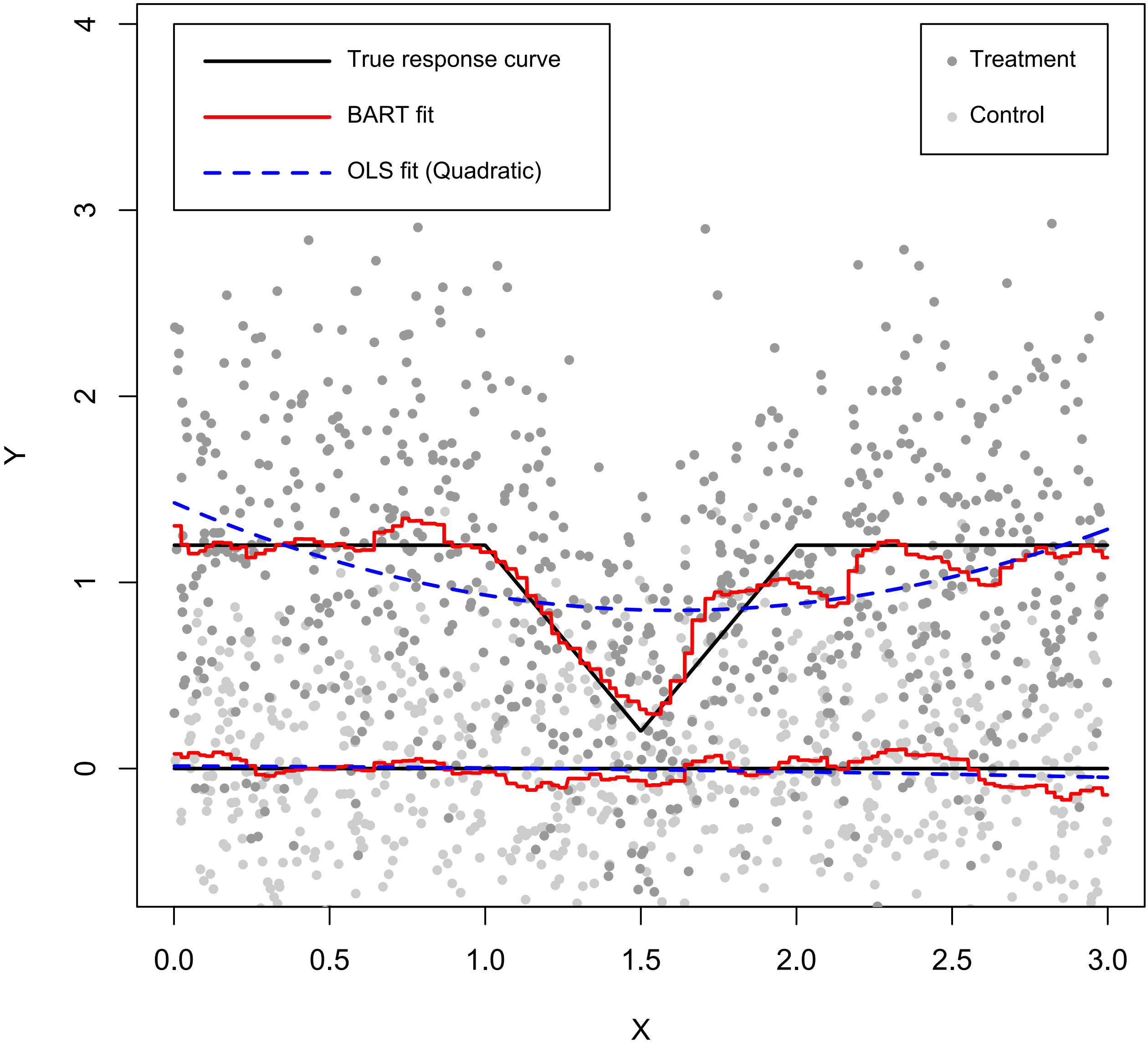
BART can also be used with binary outcomes. We rely on a probit version of BART:

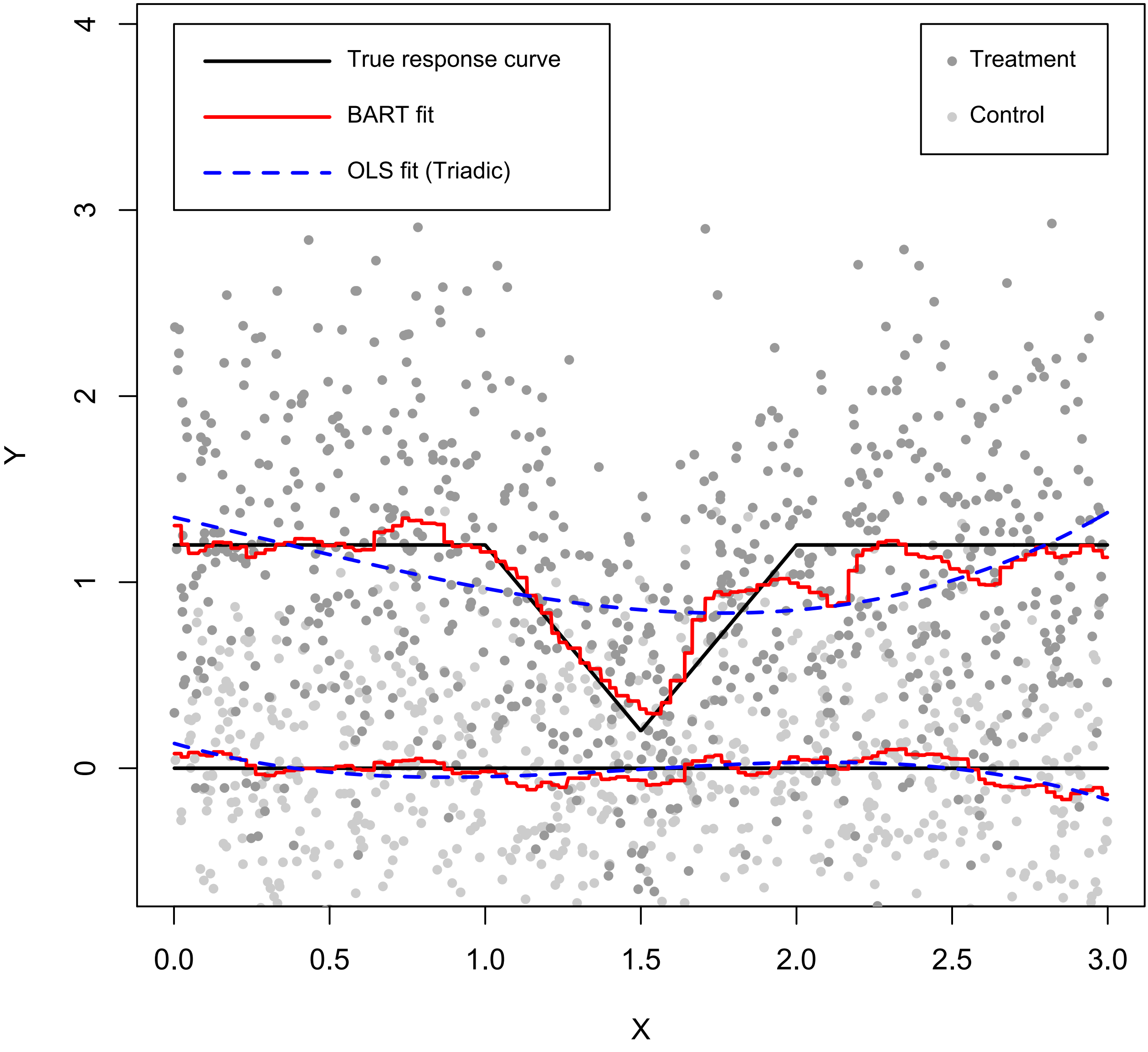
$$P(Y = 1 \mid x) = \Phi\left[\sum_{j=1}^{m} g(x; T_j, M_j)\right], \tag{3}$$

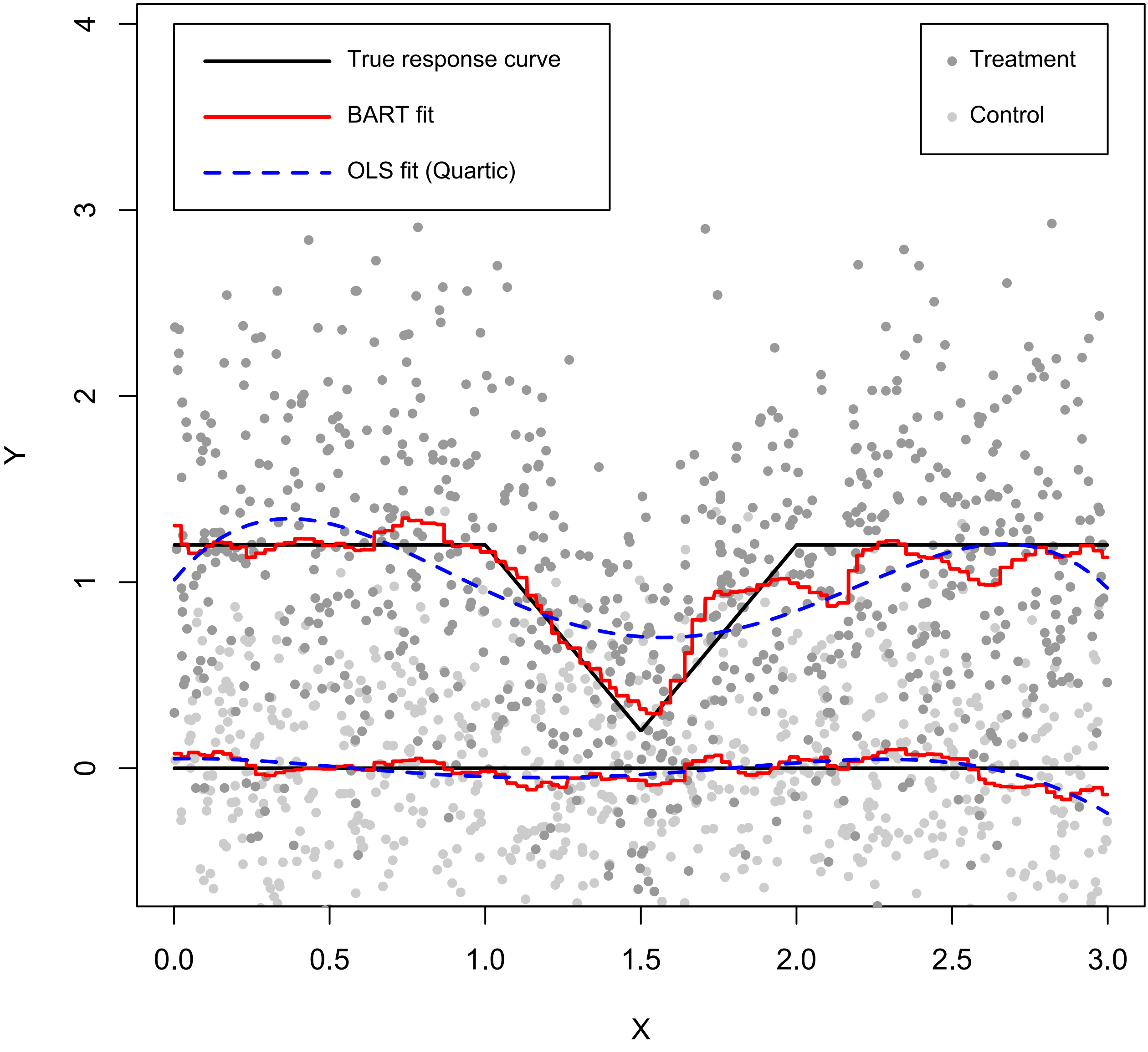
with $\Phi[\cdot]$ the standard normal cdf.

```r
### Artificial example to illustrate tree models
library(BayesTree)
rm(list = ls(all = TRUE))
set.seed(123456)
n <- 1000
Tr <- rbinom(n,1,.5)
x <- runif(n,0,3)
y0 <- rnorm(n,0,0.5)
y1 <- y0 + Tr*(x*(x<=1)+(x>1)*(x-2)^2) + rnorm(n,0,.5) + .2
y <- y0
y[Tr==1] <- y1[Tr==1]

dip <- function(x){
 (x<=1)*x + (x>1)*(x-2)^2
 }

# BART fit
temp.X <- data.frame(Tr,x)
temp.S <- data.frame(c(rep(1,n),rep(0,n)), c(sort(x),sort(x)))
colnames(temp.S) <- colnames(temp.X)

out.bart <- bart(x.train = temp.X, y.train = y, ndpost = 1000, nskip =
 1000, keepevery = 1, ntree = 200, usequants = TRUE, keeptrainfits =
 FALSE, x.test = temp.S)

cat(dim(out.bart$yhat.test), "\n")


out.bart0 <- out.bart
out.bart1 <- out.bart

out.bart1$yhat.test <-
 out.bart$yhat.test[,1:(ncol(out.bart$yhat.test)/2)]
out.bart0$yhat.test <-
 out.bart$yhat.test[,((ncol(out.bart$yhat.test)/2)+1):ncol(out.bart$yhat
 .test)]


out <- matrix(NA,n,3)
out[,1] <- sort(x)
colnames(out) <- c("x value", "Y0", "Y1")

for(i in 1:n)   {
    out[i,2] <- mean(out.bart0$yhat.test[,i])
    out[i,3] <- mean(out.bart1$yhat.test[,i])
                }
```


True response curve
BART fit
OLS fit (Linear)
● Treatment
● Control
Y
X
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5
3.0


True response curve
BART fit
OLS fit (Quadratic)
● Treatment
● Control
Y
X
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5
3.0


True response curve
BART fit
OLS fit (Triadic)
Treatment
Control
Y
X
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5
3.0


True response curve
BART fit
OLS fit (Quartic)
Treatment
Control
Y
X
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5
3.0

EMPIRICAL EXAMPLE

Public support for government spending on “Welfare” /“Assistance to the Poor” question wording experiment in General Social Survey (GSS)

Past literature has paid relatively little attention to treatment effect heterogeneity (Henry, Reyna, and Weiner 2004; Gilens 1999; Federico 2004; Kluegel and Smith 1986; Bullock, Williams, and Limbert 2003; Henry 2004; Federico 2004; Jacoby 2000)

N = 14,555, split randomly into equally-sized *training* and *test* datasets

Covariates: age, liberal-conservative scale, party identification, education, negative attitudes toward blacks scale, and survey wave

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 1987–2008. 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.

EMPIRICAL EXAMPLE

Public support for government spending on “Welfare” /“Assistance to the Poor” question wording experiment in General Social Survey (GSS)

Past literature has paid relatively little attention to treatment effect heterogeneity (Henry, Reyna, and Weiner 2004; Gilens 1999; Federico 2004; Kluegel and Smith 1986; Bullock, Williams, and Limbert 2003; Henry 2004; Federico 2004; Jacoby 2000)

N = 14,555, split randomly into equally-sized *training* and *test* datasets

EMPIRICAL EXAMPLE

Public support for government spending on “Welfare” /“Assistance to the Poor” question wording experiment in General Social Survey (GSS)

Past literature has paid relatively little attention to treatment effect heterogeneity (Henry, Reyna, and Weiner 2004; Gilens 1999; Federico 2004; Kluegel and Smith 1986; Bullock, Williams, and Limbert 2003; Henry 2004; Federico 2004; Jacoby 2000)

N = 14,555, split randomly into equally-sized *training* and *test* datasets

Covariates: age, liberal-conservative scale, party identification, education, negative attitudes toward blacks scale, and survey wave

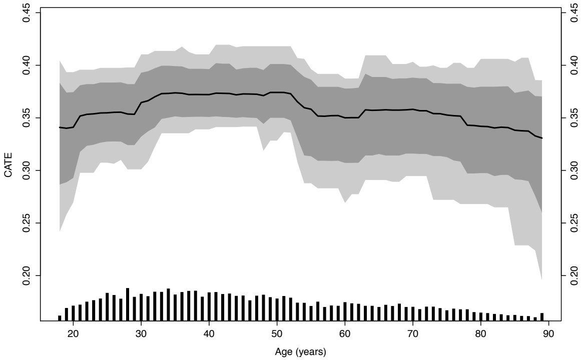
EMPIRICAL EXAMPLE

Public support for government spending on “Welfare” /“Assistance to the Poor” question wording experiment in General Social Survey (GSS)

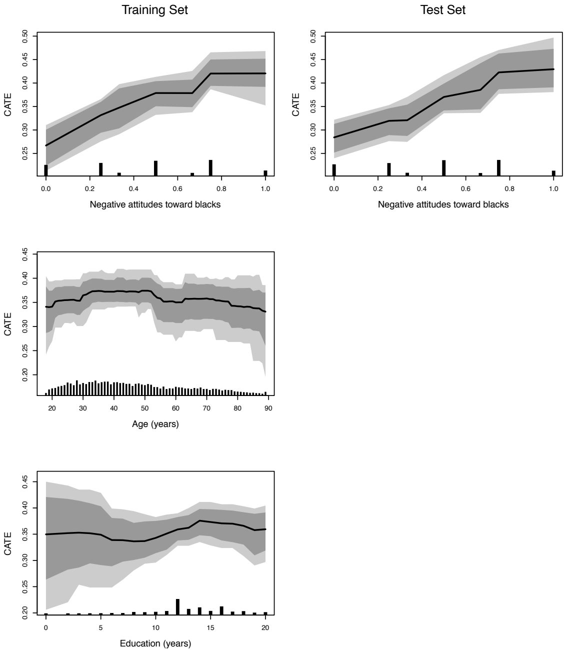
Past literature has paid relatively little attention to treatment effect heterogeneity (Henry, Reyna, and Weiner 2004; Gilens 1999; Federico 2004; Kluegel and Smith 1986; Bullock, Williams, and Limbert 2003; Henry 2004; Federico 2004; Jacoby 2000)

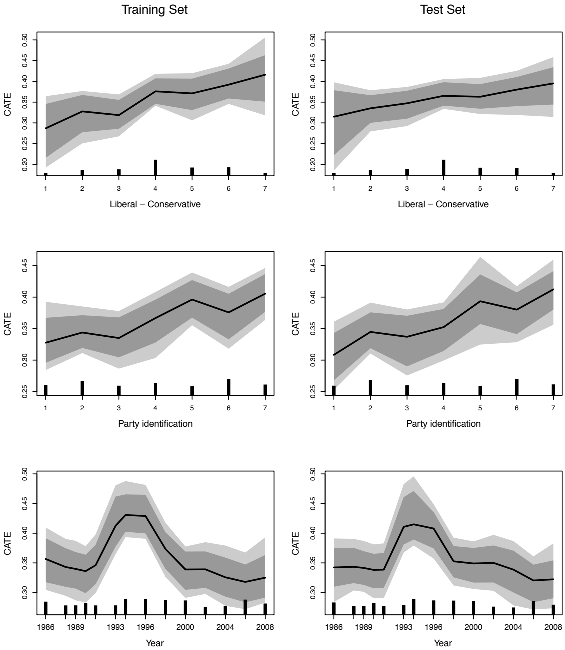
N = 14,555, split randomly into equally-sized *training* and *test* datasets

Covariates: age, liberal-conservative scale, party identification, education, negative attitudes toward blacks scale, and survey wave

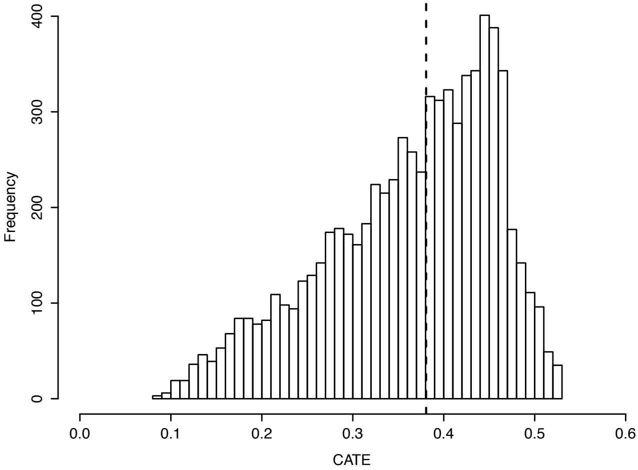


Note: The graph shows average treatment effects (the black curve) conditional on age for the training dataset. The dark grey area is a point-wise 95% posterior interval; the light grey area is a global 95% posterior interval. The marginal distribution of age is shown at the bottom of the graph.


Training Set
Test Set
CATE
0.25
0.30
0.35
0.40
0.45
0.50
0.0
0.2
0.4
0.6
0.8
1.0
Negative attitudes toward blacks
Negative attitudes toward blacks
CATE
0.20
0.25
0.30
0.35
0.40
0.45
20
30
40
50
60
70
80
90
Age (years)
CATE
0.20
0.25
0.30
0.35
0.40
0.45
0
5
10
15
20
Education (years)

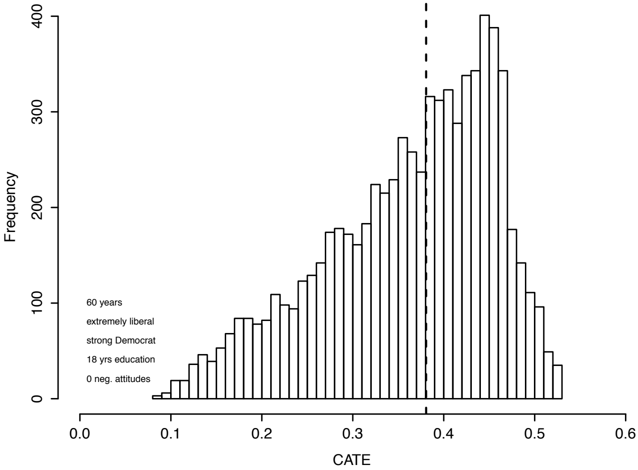

Training Set
Test Set
CATE
0.20
0.25
0.30
0.35
0.40
0.45
0.50
1
2
3
4
5
6
7
Liberal – Conservative
Party identification
Year
1986
1989
1993
1996
2000
2004
2008

Distribution of estimated CATEs



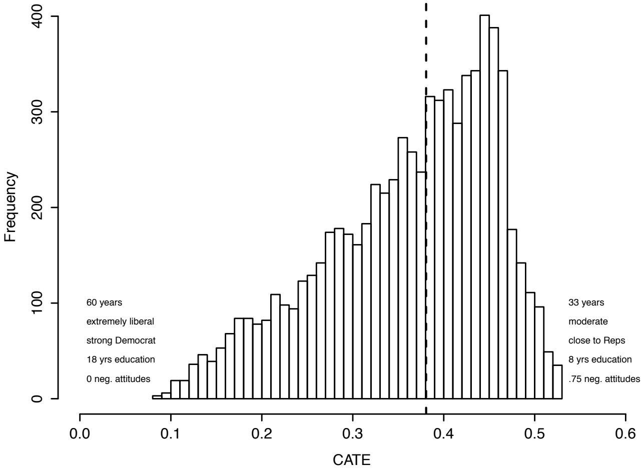
Source: GSS 1986–2008. The graph shows a histogram of conditional average treatment effects on the probability scale for the 7,278 individuals in the training set. The vertical dashed line denotes the median conditional average treatment effect.

Distribution of estimated CATEs



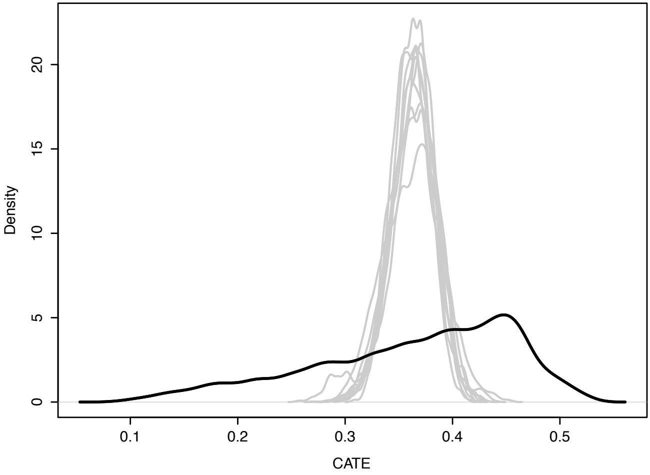
Source: GSS 1986–2008. The graph shows a histogram of conditional average treatment effects on the probability scale for the 7,278 individuals in the training set. The vertical dashed line denotes the median conditional average treatment effect.

Distribution of estimated CATEs



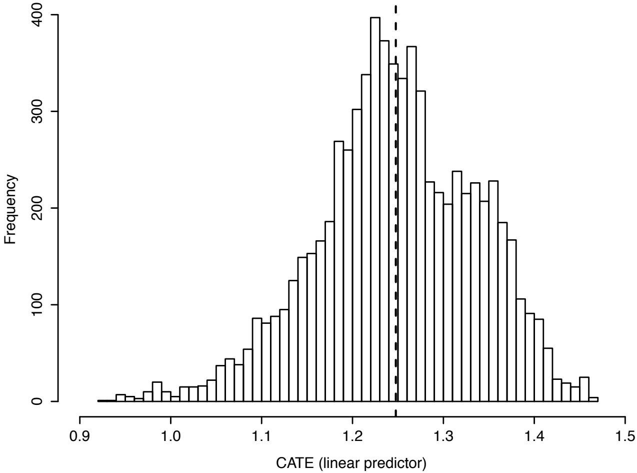
Source: GSS 1986–2008. The graph shows a histogram of conditional average treatment effects on the probability scale for the 7,278 individuals in the training set. The vertical dashed line denotes the median conditional average treatment effect.

Comparison with permutation baseline

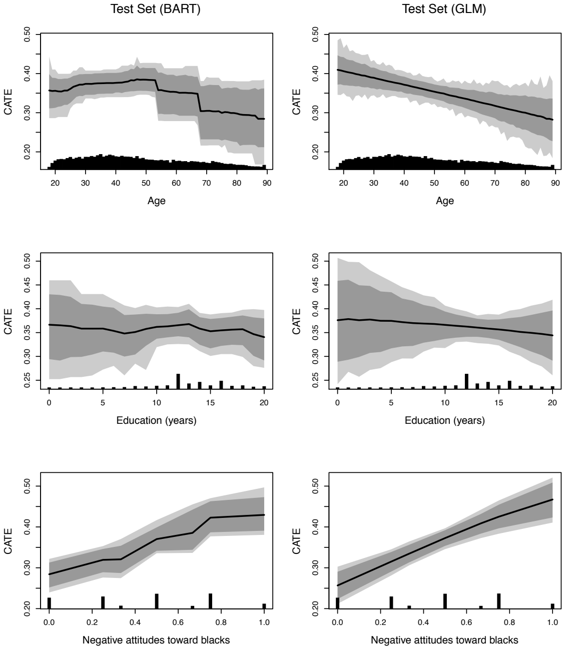


Source: GSS 1986–2008. The graph shows a kernel density plot of conditional average treatment effects on the probability scale for the 7,278 individuals in the training set (black curve) and 10 kernel density plots for the same individuals when covariate values are randomly permuted.

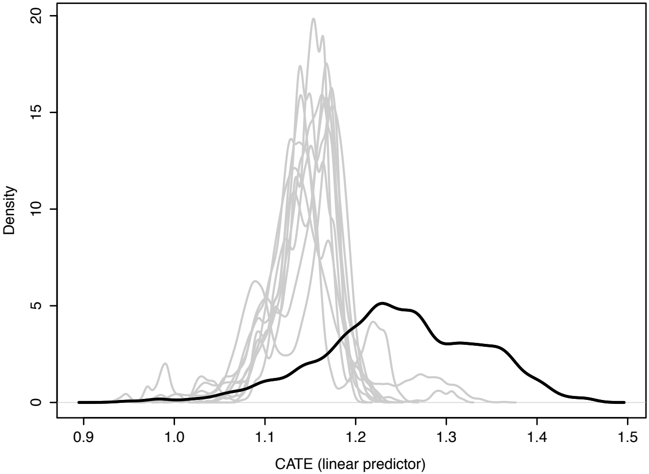
Distribution of estimated CATEs



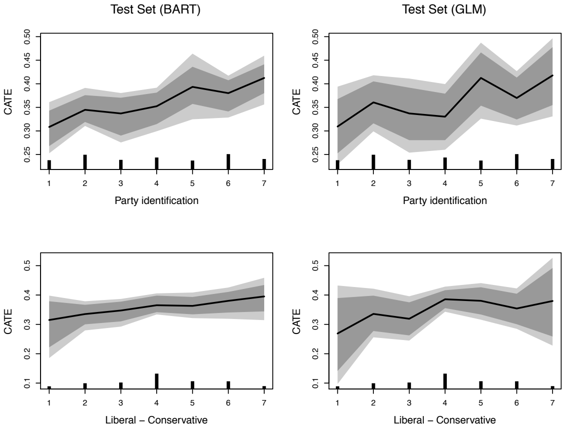
Source: GSS 1986–2008. The graph shows a histogram of conditional average treatment effects on the on the scale of the linear predictor for the 7,278 individuals in the training set. The vertical dashed line denotes the median conditional average treatment effect.


Test Set (BART)
Test Set (GLM)
CATE
0.20
0.30
0.40
0.50
20
30
40
50
60
70
80
90
Age
0.25
0.30
0.35
0.40
0.45
0.50
0
5
10
15
20
Education (years)
0.0
0.2
0.4
0.6
0.8
1.0
Negative attitudes toward blacks

Comparison with permutation baseline



Source: GSS 1986–2008. The graph shows a kernel density plot of conditional average treatment effects on the scale of the linear predictor 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.


Test Set (BART)
Test Set (GLM)
CATE
0.25
0.30
0.35
0.40
0.45
0.50
1
2
3
4
5
6
7
Party identification
CATE
0.1
0.2
0.3
0.4
0.5
Liberal – Conservative

CONCLUSION

Our approach provides a principled framework for the discovery of systematic treatment effect heterogeneity in large-scale experiments

BART largely automates the search for heterogeneity and allows researchers to flexibly model it nonparametrically

Our split sample design permits a relatively unstructured data-driven exploration but avoids charges of post hoc data dredging and mitigates multiple comparison problems

In the years ahead, we hope to see a fundamental change in the way that experimenters investigate and report systematic treatment effect heterogeneity