Machine Learning for Heterogenous Treatment Effects

Dylan Groves

October 10, 2018

Overview

1. Conceptual Goal, Challenge, and Strategy
2. Challenges with Heterogenous Treatment Effects
3. Bayesian Additive Regression Trees
4. Recursive Partitioning
5. Conclusion

Introduction

The Goal

- We are usually interested in the average treatment effect (ATE)
- But we might also be interested in how the ATE varies with other covariates: the conditional average treatment effect (CATE).

1. How does effect of "welfare priming" vary by respondents' baseline attitudes towards blacks?
2. How does effect of a village-level agriculture program vary with rainfall?

- Goal is prediction, not causal identification

The Obstacles

Modelling Conditional Average Treatment Effects (CATE) is challenging:

1. Ad-hockery and data dredging in search for "interesting" interactions (Pocock 2002; Gabler et al. 2009)
2. Non-statistical uncertainty (the interaction model might itself by biased)
3. Multiple comparisons problem

The historical record on heterogenous treatment effects in the social sciences is very weak

The Solution (Conceptually)

The solution is to automate the search for heterogeneity:

1. No ad-hockery: Let the data decide the key sources of heterogeneity
2. No functional form assumptions: non-parametric regression trees
3. No multiple hypothesis testing: identify the search parameters ex-ante

Notation

Formal Notation (no covariates)

No interaction and consistent treatment effect

$$Y_i = \beta_0 + \beta_{Di} D_i + \epsilon_i$$

Idiosyncratic treatment effect:

$$Y_i = B_0 + \beta_{Di} D_i + \underbrace{[B_{Di} - B_D]}_{\text{idiosyncratic effect of treatment}} D_i + \epsilon_i$$

Formal Notation (with covariates)

But both outcome and treatment effect may vary with covariates:

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

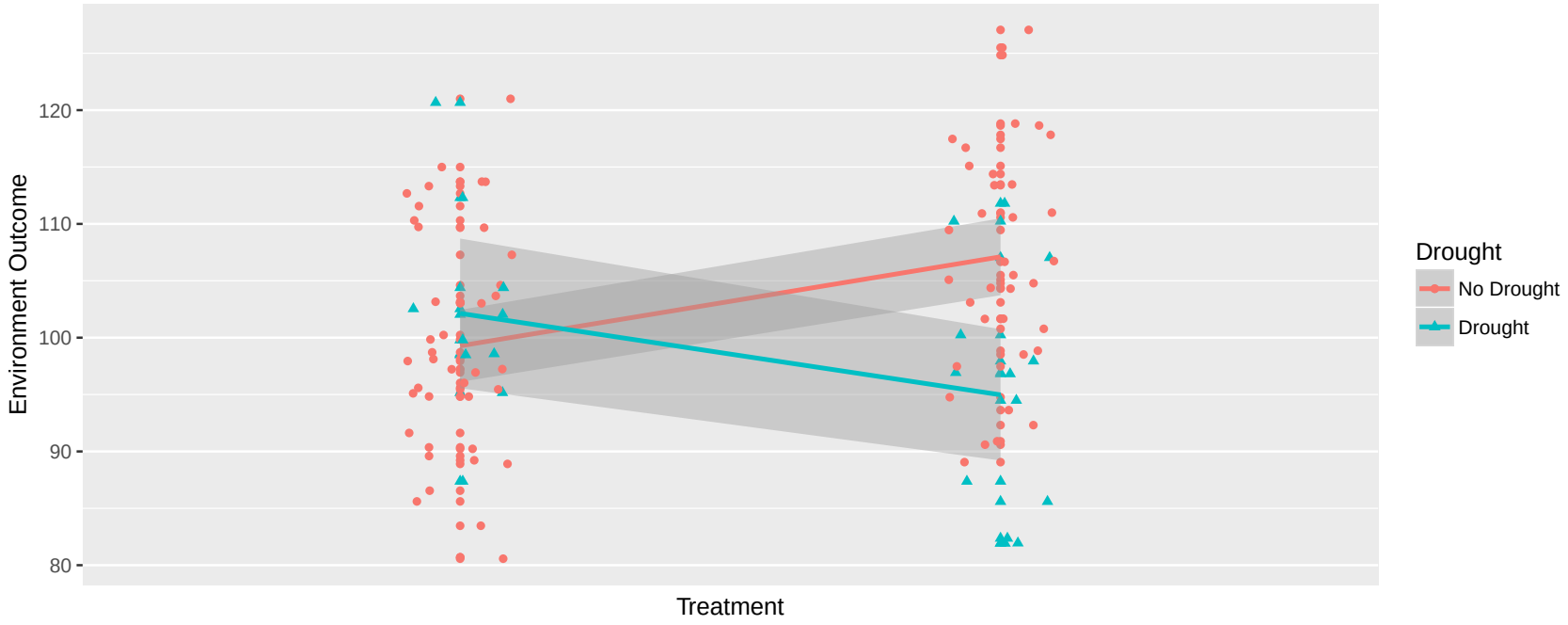
Systematic, not idiosyncratic, treatment effect heterogeneity (although the latter doesn't bias the CATE):

$$Y_i = \beta_0' + \beta_X X_i + \underbrace{(\beta_D + \beta_{DX} X_i) D_i}_{\text{How treatment effect varies with covariates}} + \epsilon_i'$$

Examples of Het FX

Example 1 - Binary Covariate

Consider a case where drought moderates the effect of treatment on a continous outcome.



Example 1 - Binary Covariate

Estimating the conditional average treatment effect for drought and non-drought conditions is straightforward:

$$\tau(x) = E[Y_i(0) - Y_i(1)|X_i]$$

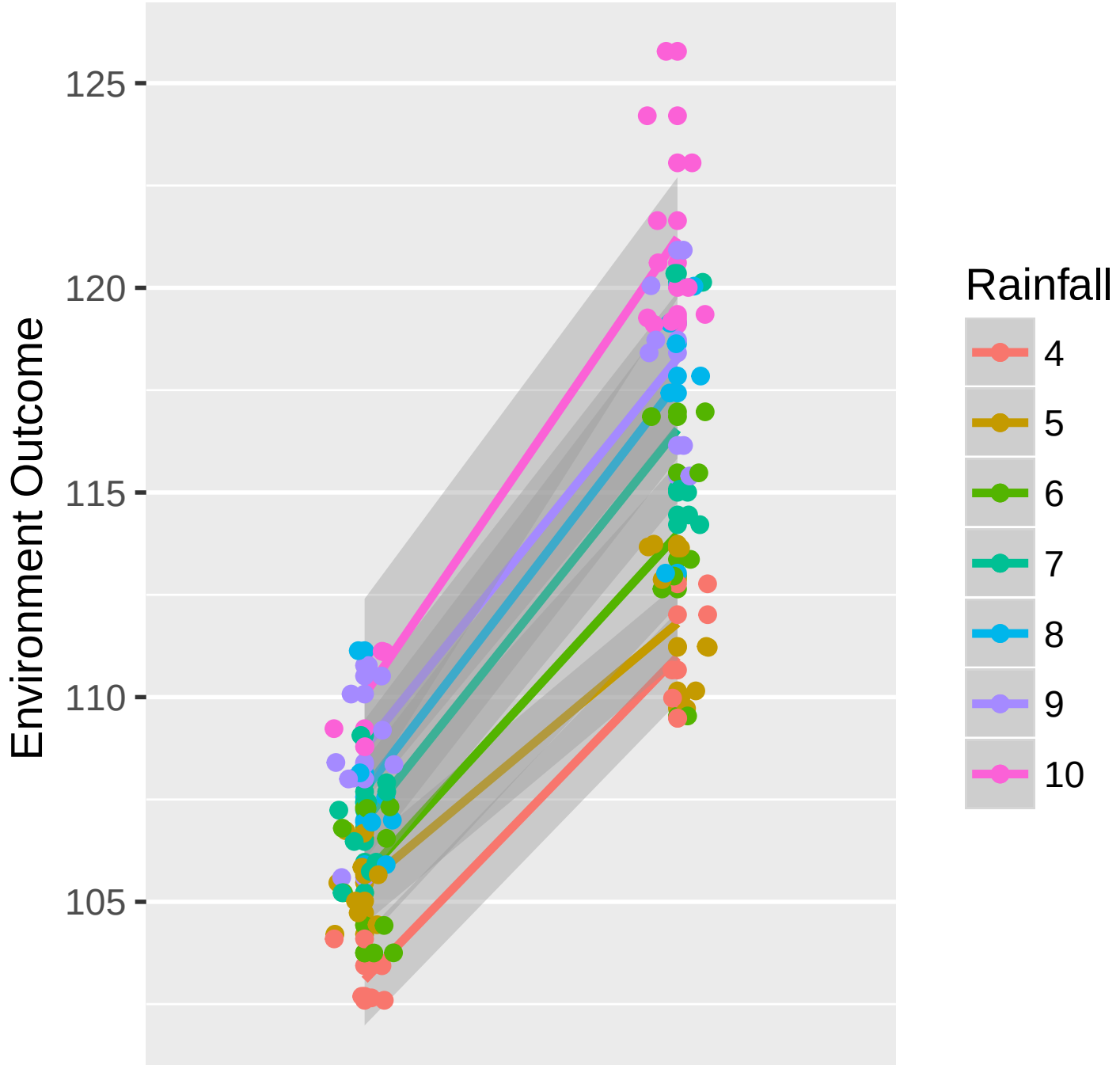
CATE (drought)	Pct Drought	CATE (no drought)	Pct No Drough
-7.16	0.23	7.83	0.77

We can then weight the CATE by the probability of each condition to estimate the ATE:

```
cate.drought*mean(drought) + cate.nodrought*(1-mean(drought
```

[1] 4.38243

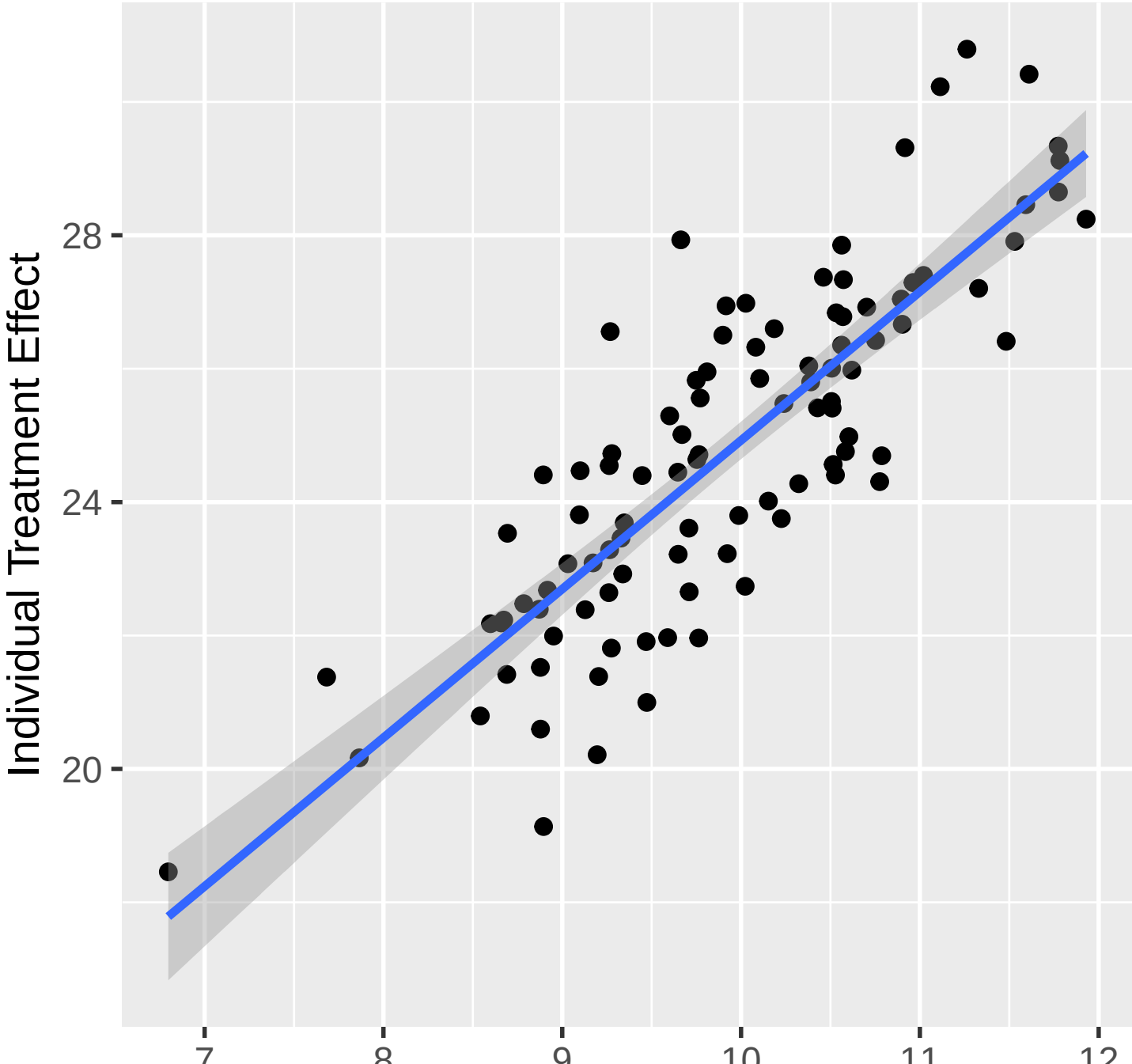
Example 2 - Categorical



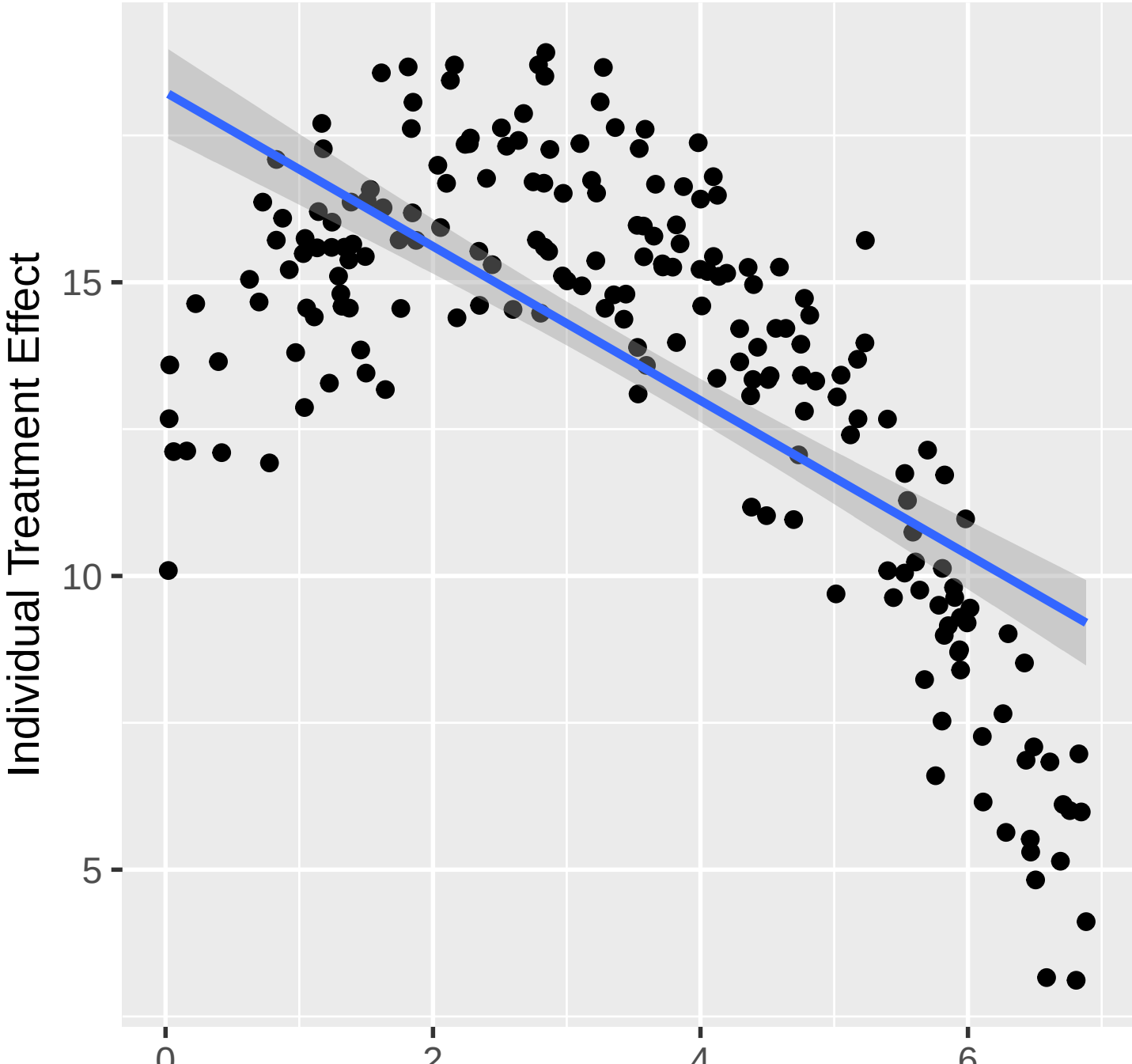
Example 2 - Categorical

Rainfall	Mean (Z=1)	Mean (Z=0)	Pct	Y(1)-Y(0)
4	110.98	103.09	0.01	7.89
5	111.81	105.42	0.02	6.39
6	113.97	105.64	0.02	8.34
7	116.54	106.97	0.02	9.57
8	117.69	107.73	0.01	9.96
9	118.28	108.77	0.02	9.51
10	121.22	110.06	0.01	11.16

Example 3 - Continous



Example 4 - Curvilinear



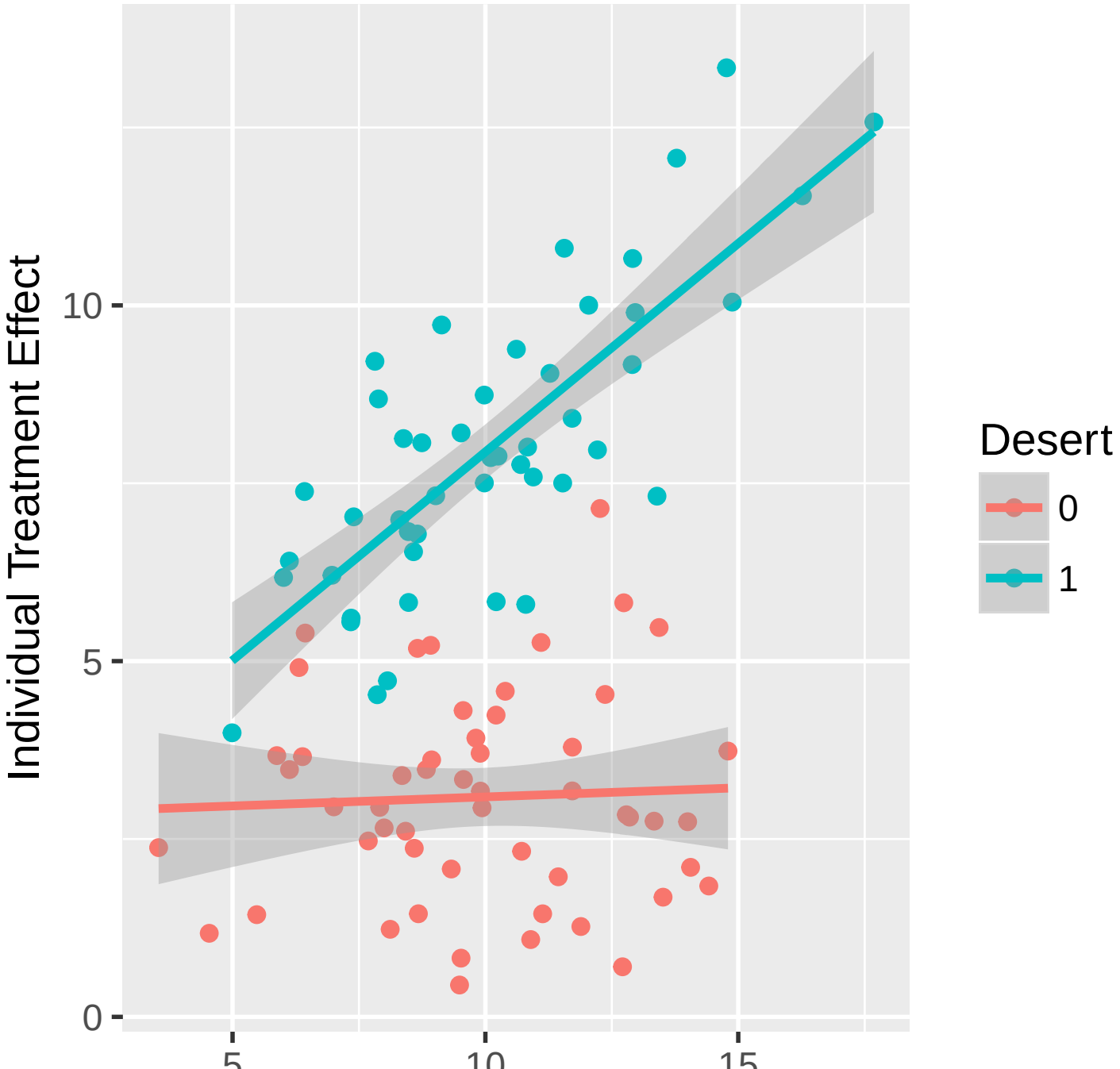
Example 5 - Multiple Interactions

Consider a case in which the the heterogenous effect is itself conditional on another covariate.

For example, perhaps rainfall only influences treatment effects in desertified areas.

$$Y_i = \beta_0 + \beta_1 * \mathit{Treat} + \beta_2 * \mathit{Rain} + \beta_3 * T * R * \mathit{Desert}$$

Example 5 - Complex



BART

Review of Challenges

1. Ad-hockery
2. Too many interactions (Imprecise estimates)
3. Sub-group analysis (Multiple comparisons problems)
4. Strong modelling assumptions (non-statistical uncertainty)

These problems expand with:

1. Continuous covariates (where do you make cuts)?
2. Large numbers of covariates (which covariates matter)?
3. Interacting and curvilinear effects (how do you know if you have specified the right model)?

The Solution in Principle

We want a solution that:

- Lets the data make the choices
- Identifies meaningful interactions (precise)
- Doesn't overfit the data (out-of-sample validity)

Bayesian Additive Regression Trees (BART)

Green and Holger 2012 propose Bayesian Additive Regression Trees (Chipman, George, and McCulloch 2010).

1. Eliminate ad-hoc data mining
2. Non-parametric estimation of CATEs
3. Distinguish exploration and confirmation by splitting samples into training and test groups

BART models an outcome Y as an unknown function f of a p-dimensional vector of predictors x and an i.i.d error term:

$$Y = f(x) + \epsilon$$

What is a Regression Tree?

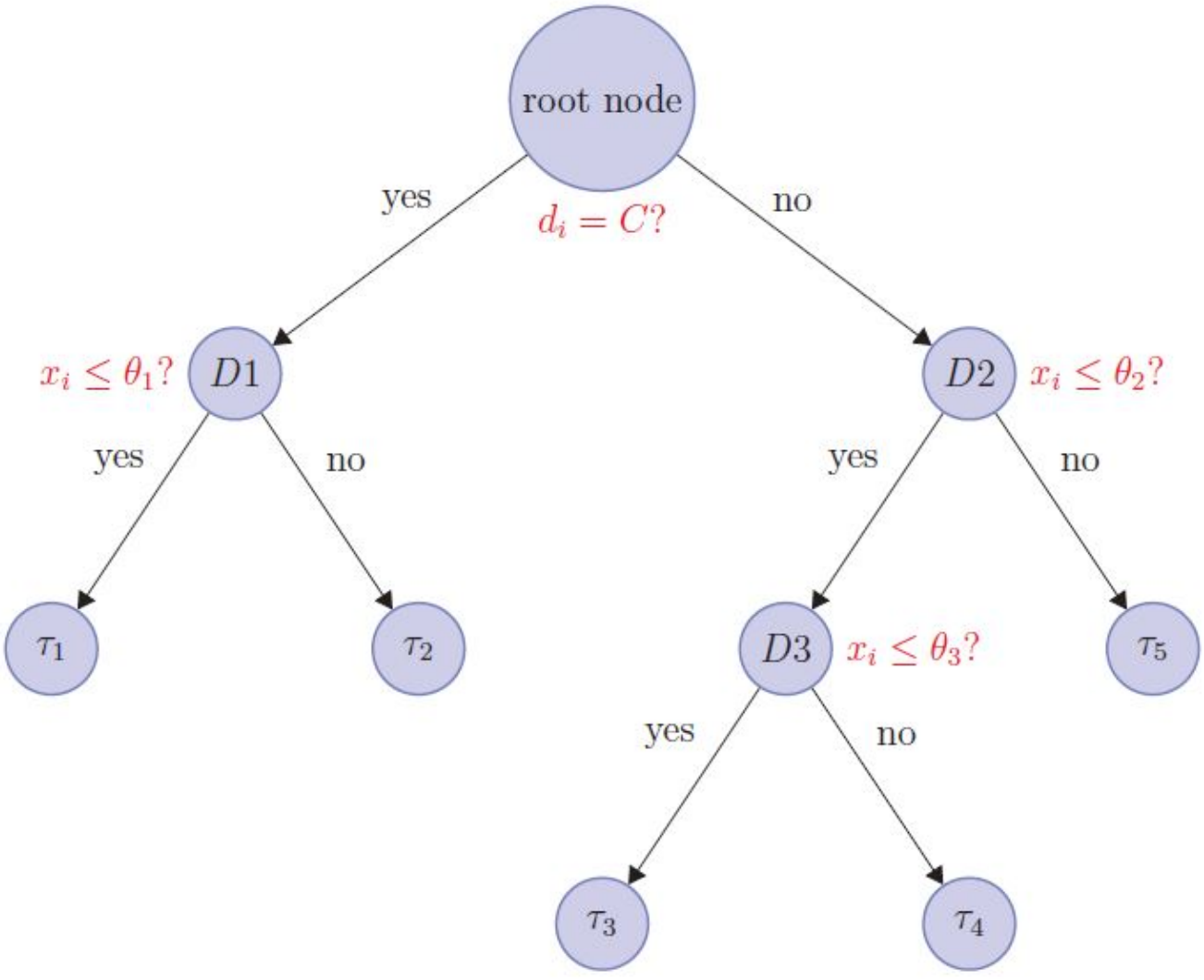
We call a tree T. It includes: (1) decision nodes and the decision rules at each node, (2) terminal nodes. Every observation fits in exactly one terminal node.

We call M the terminal node parameter values $\mu_1, \mu_2, ...\mu_b$. μ_k is the mean response of the subgroup of observations falling in terminal node k.

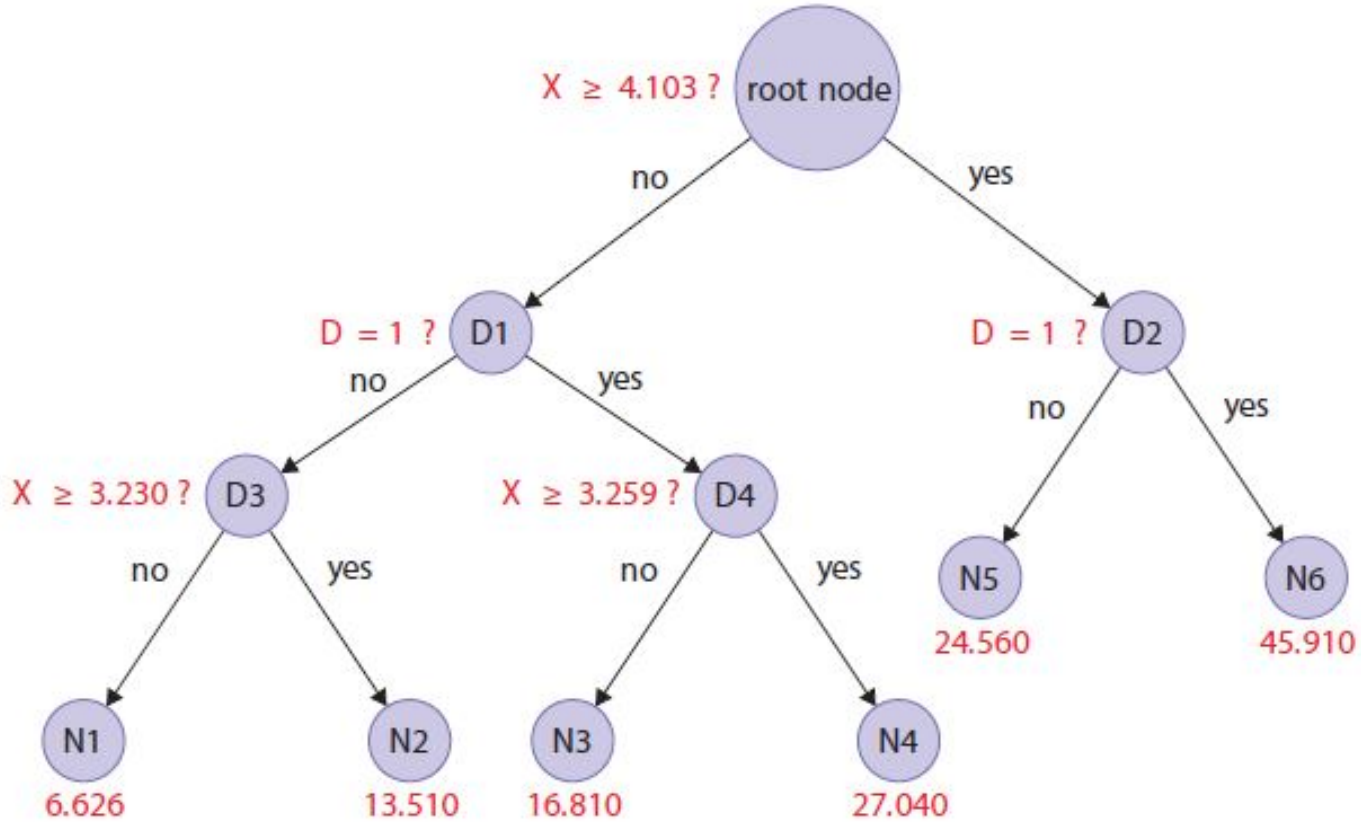
$g(x; T, M)$ generates a μ_{kj} for an observation with characteristics x by giving it the median of its terminal node. Y is therefore modelled as:

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

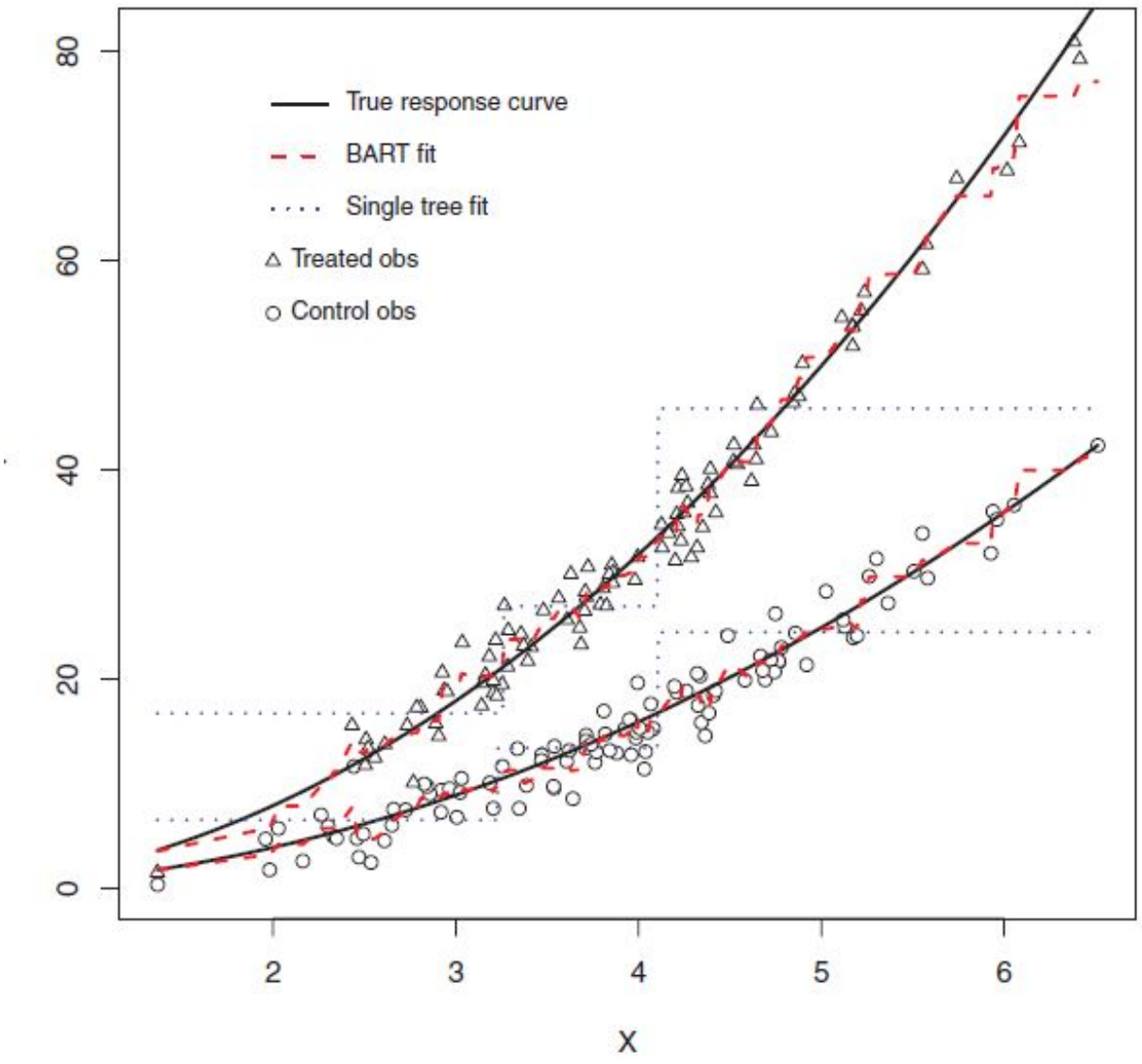
What is a Regression Tree?



Regression Tree Example



Single Tree vs BART fit



Aggregating Regression Trees

BART approximates $f(x) = E(Y|x)$ by a sum of regression trees:

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

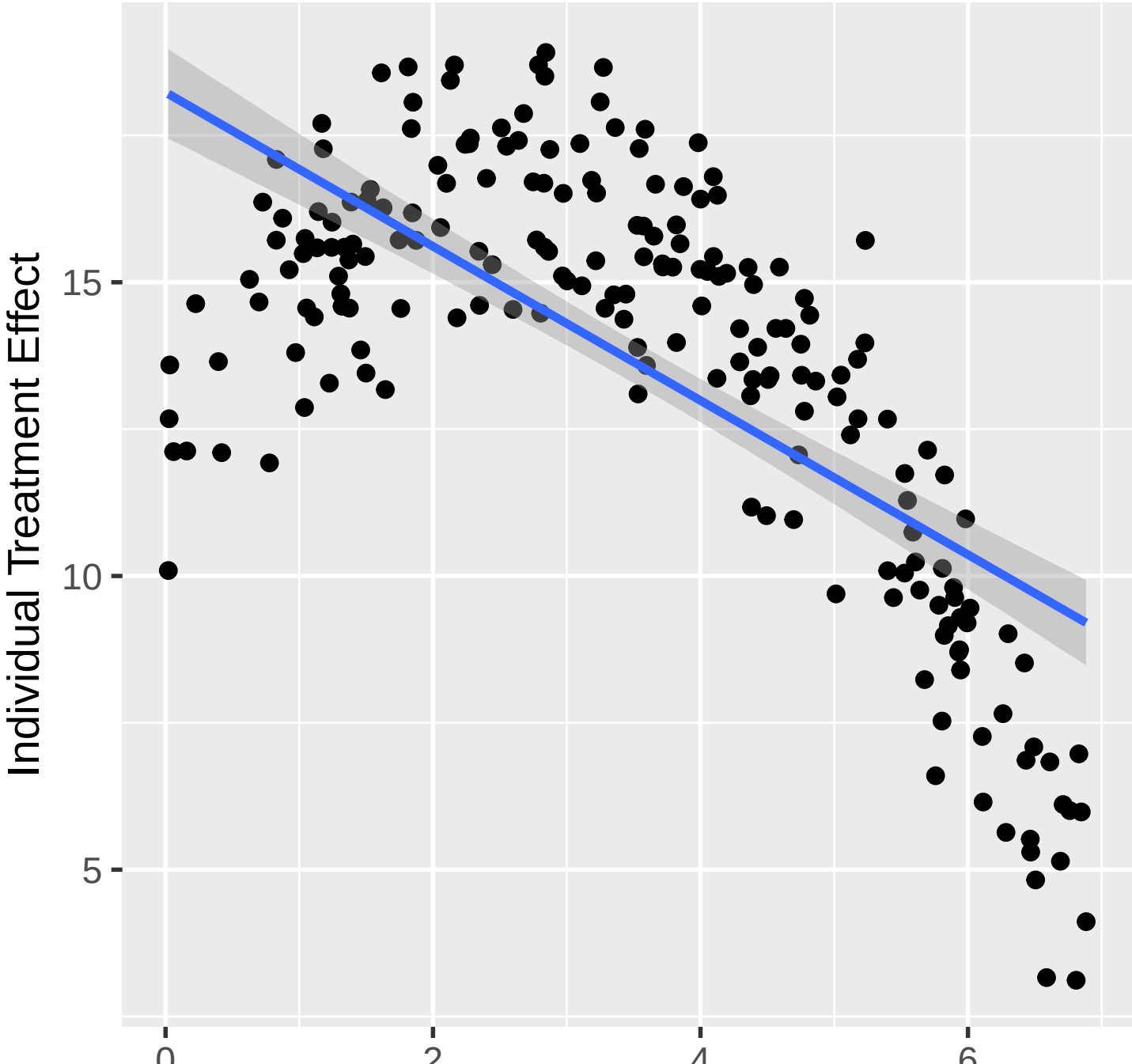
- The output of $g(x, T_j, M_j)$ is the value obtained by dropping an observation with chracteristics x down the tree until it hits a terminal node, and then reporting $\mu_{zj} \in M_j$.
- $E(Y|x)$ equals the sum of all terminal node parameters in $g(x, T_j, M_j)$ assigned to an observation with characteristics x

Aggregating Regression Trees

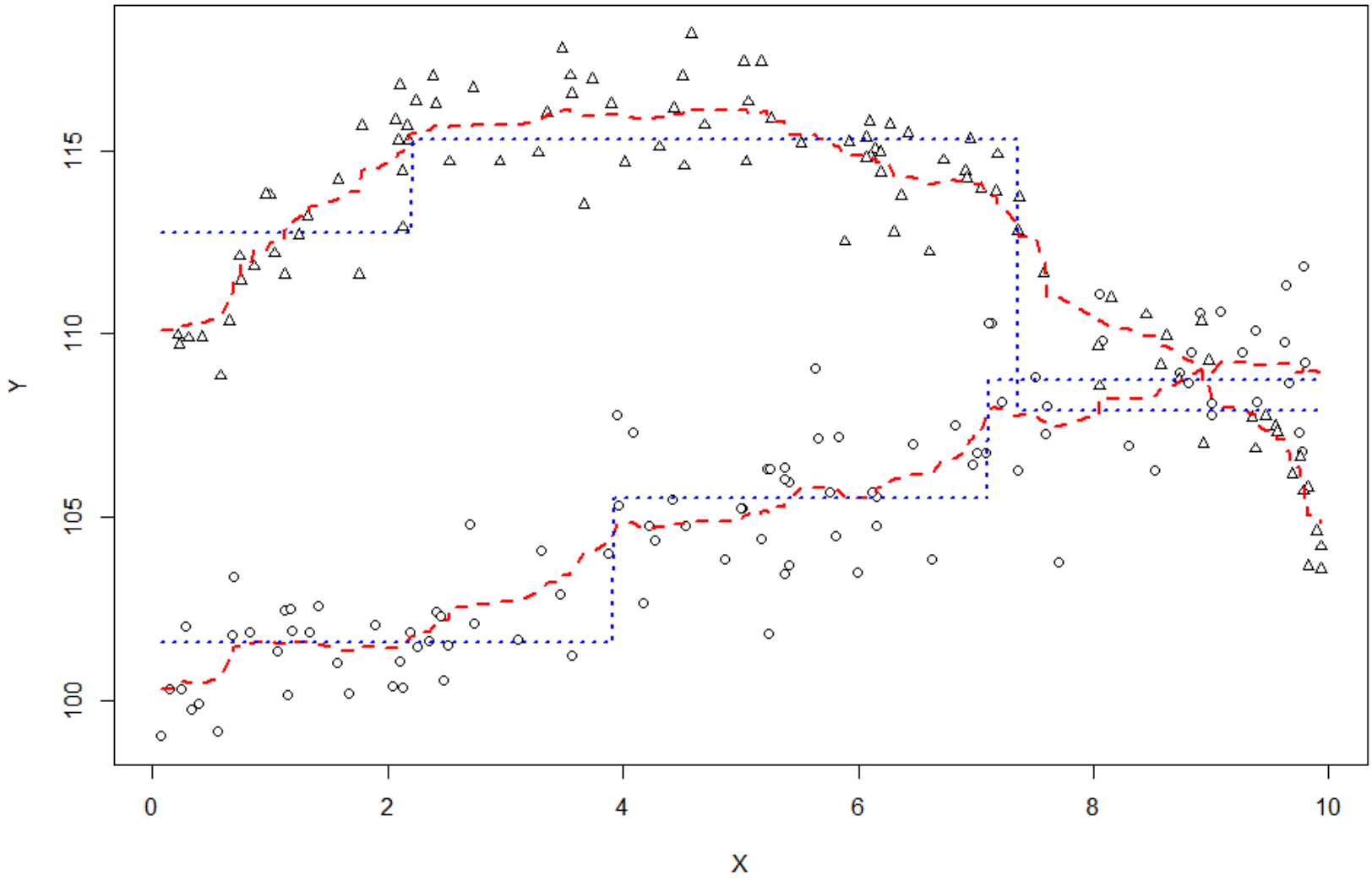
$$Y(\sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon, \quad \epsilon \; Normal(0, \sigma^2)$$

- μ_{zj} is a direct effect of x on Y when $g(x; T_j, M_j)$ includes only one component of x.
- μ_{zj} is a interaction effect when $g(x; T_j, M_j)$ depends on multiple x's
- A nice feature of BART is that many different trees can model many different interactions

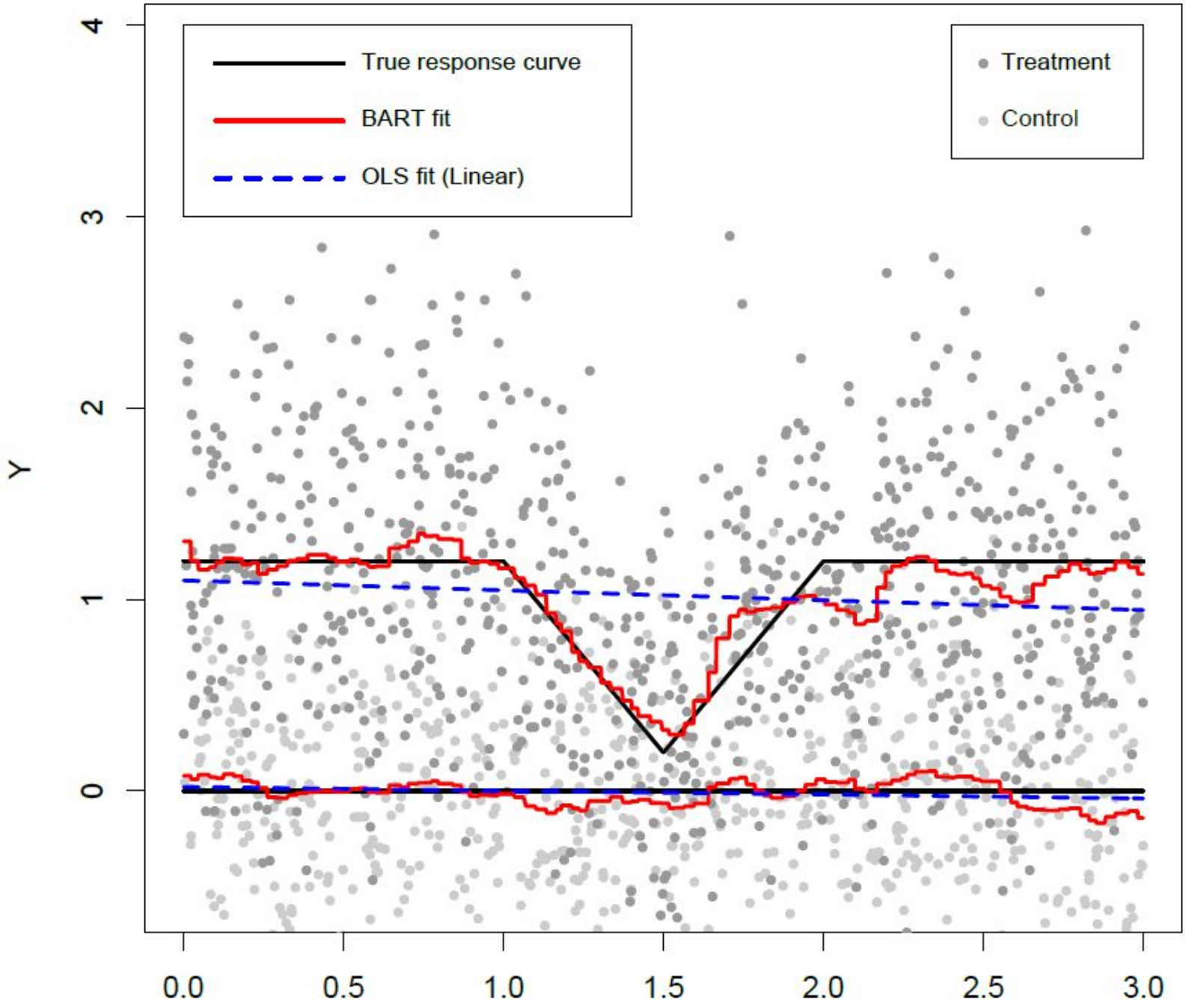
BART Fit (Ex 4)



BART Fit (Ex 4)



BART Fit


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

BART's Fitting Strategy

Need a way to decide (1) When to split, and (2) How to aggregate trees

- BART treats (T, M) and σ as parameters with priors in a statistical modelused Markov Chain Monte Carlo (MCMC)
- Posteriors are computed using Markov Chain Monte Carlo (MCMC). Redraw $T, M, and \sigma$ after each interation.
- Start with 1,000 burn-in draws and 1,000 draws from posterior

BART's Standard Priors

1. The T_j prior keeps the number of tree branches small by putting more weight on smaller trees (highest probability on trees with 2 or 3 terminal nodes)

- Doesn't mean large trees are impossible if data calls for it

2. The $\mu_i j | T_j$ prior shrinks tree parameters towards zero (as number of trees increases, contribution of each tree decreases.

3. m sets the number of trees for which BART cycles back through trees to refit.

Chipman, George, and McCullough (2010) propose a set of defaults and show that default priors work well across a variety of actual and simulated data sets.

Application

Empirical Example

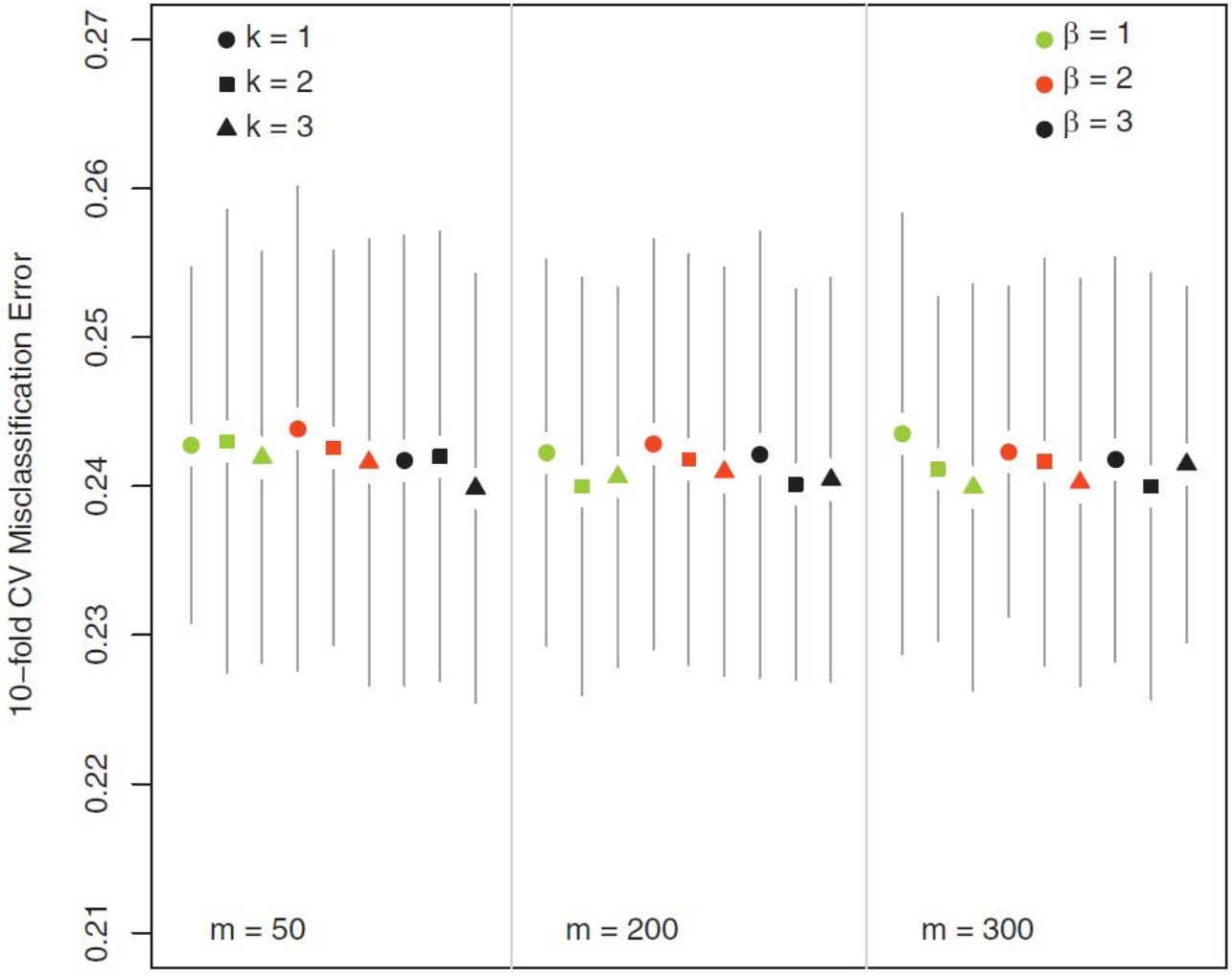
Green and Holger (2012) apply BART to classic survey experiment in the General Social Survey (GSS).

- Outcome: public support for government spending on "welfare"
- Randomized treatment: using "welfare" versus "assistance to the poor"
- Investigating treatment effect heterogeneity across year, age, education, party ID, negative attitudes towards blacks
- N = 14,555, split randomly into training and test datasets

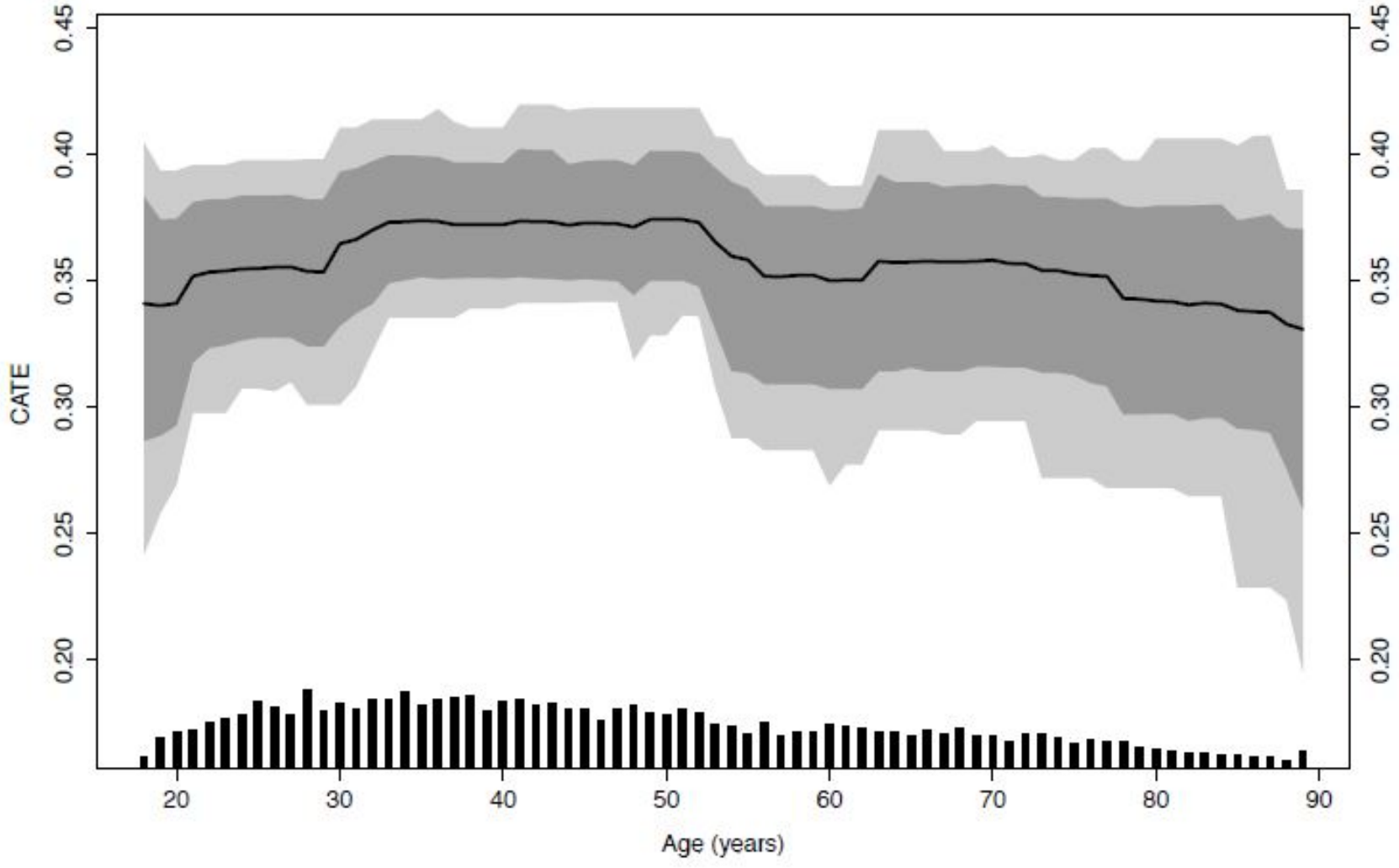
Overview of Data

Year	Sample Size		Means		ATE
	Assistance	Welfare	Assistance	Welfare	
1986	561	594	0.45	0.10	0.34
1988	359	404	0.45	0.08	0.37
1989	375	389	0.44	0.11	0.34
1990	510	536	0.42	0.09	0.34
1991	373	391	0.40	0.13	0.27
1993	418	418	0.60	0.15	0.45
1994	759	744	0.67	0.17	0.51
1996	700	704	0.64	0.22	0.42
1998	663	682	0.47	0.12	0.34
2000	664	635	0.41	0.13	0.28
2002	330	341	0.48	0.11	0.38
2004	338	341	0.48	0.07	0.41
2006	675	673	0.39	0.10	0.29
2008	456	487	0.41	0.09	0.32
2010	500	497	0.46	0.11	0.35

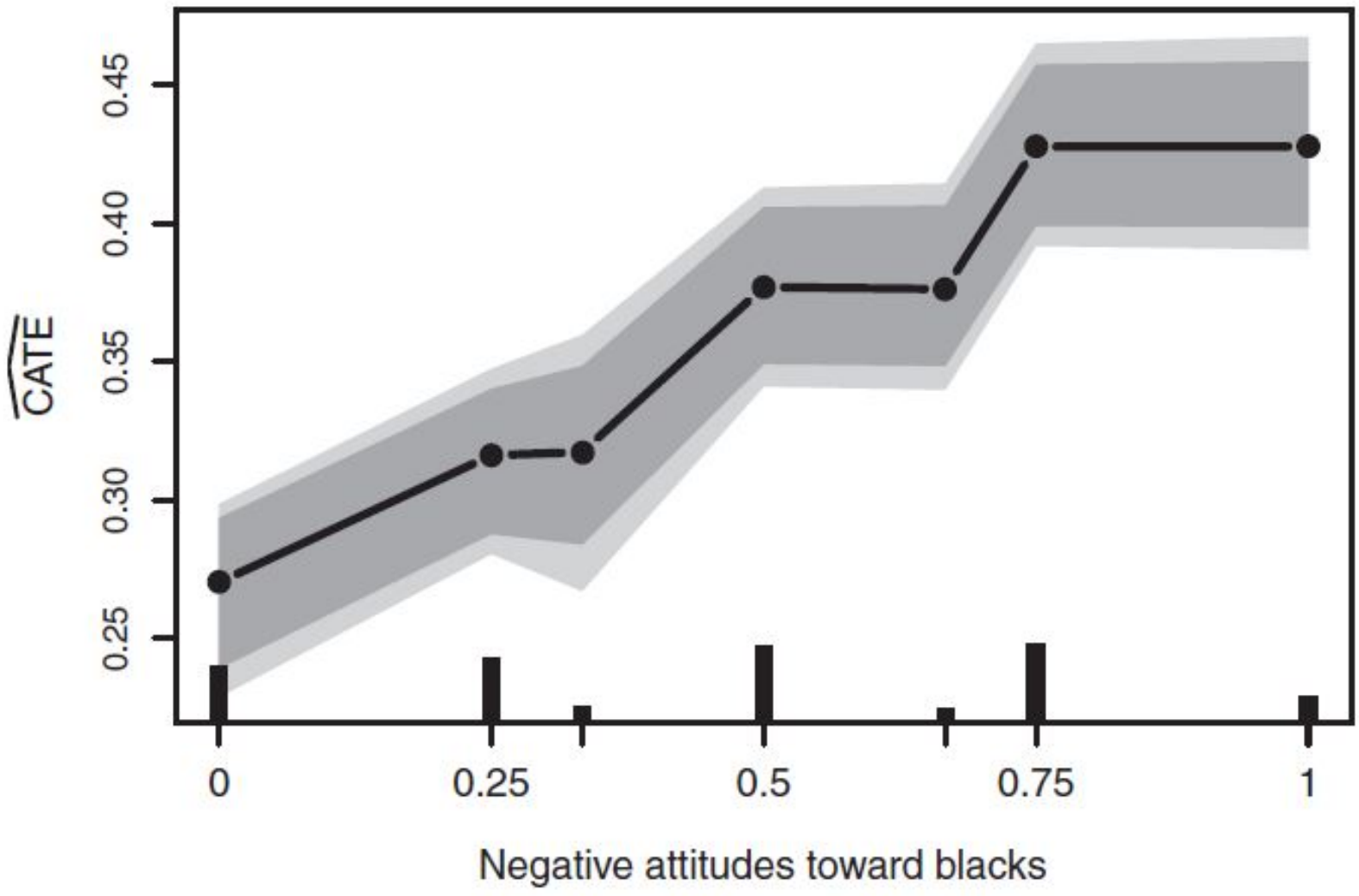
Misclassification Rates



CATE - Age



CATE - Attitudes Towards Blacks

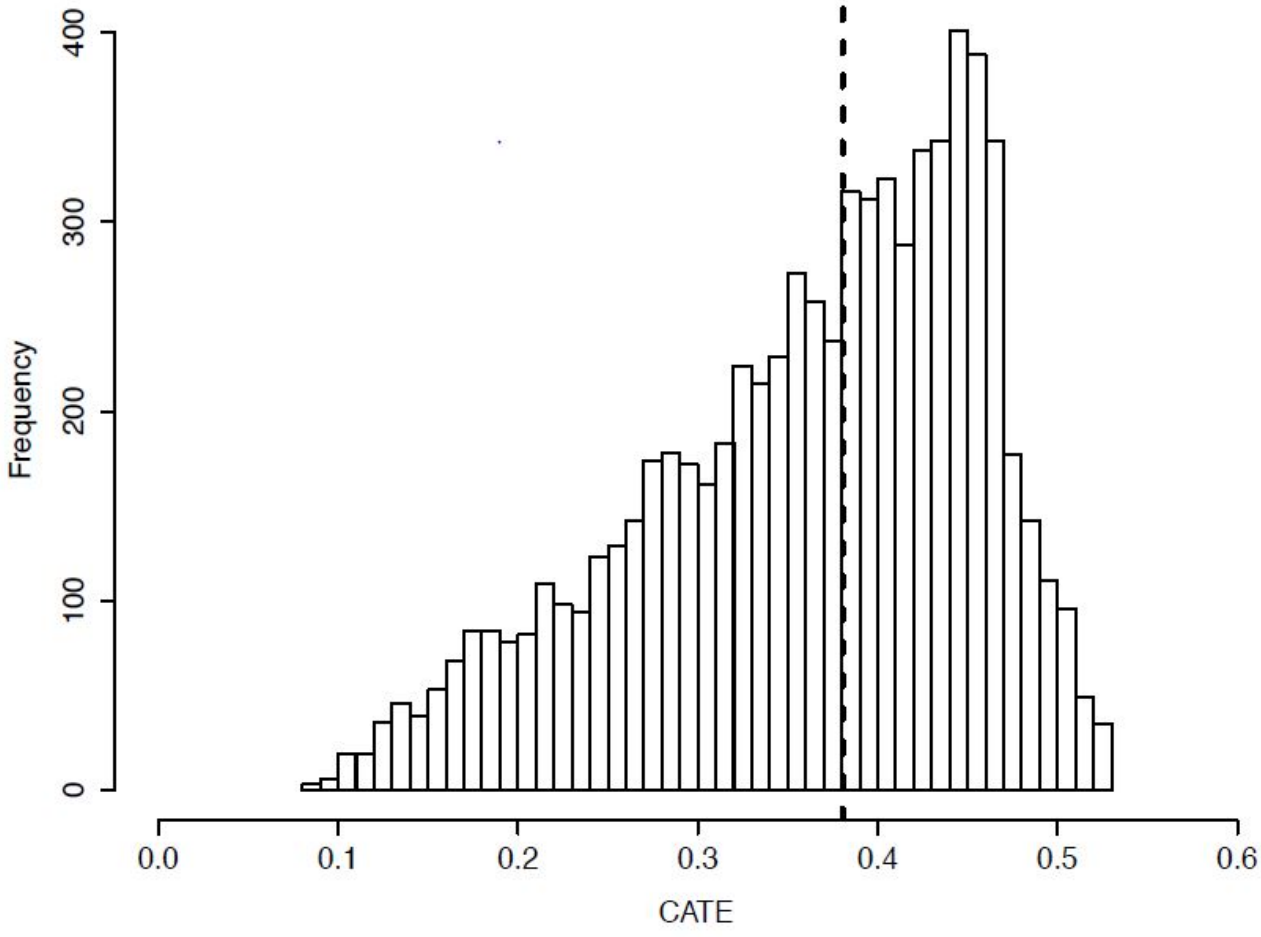


Conditional Average Treatment Effects

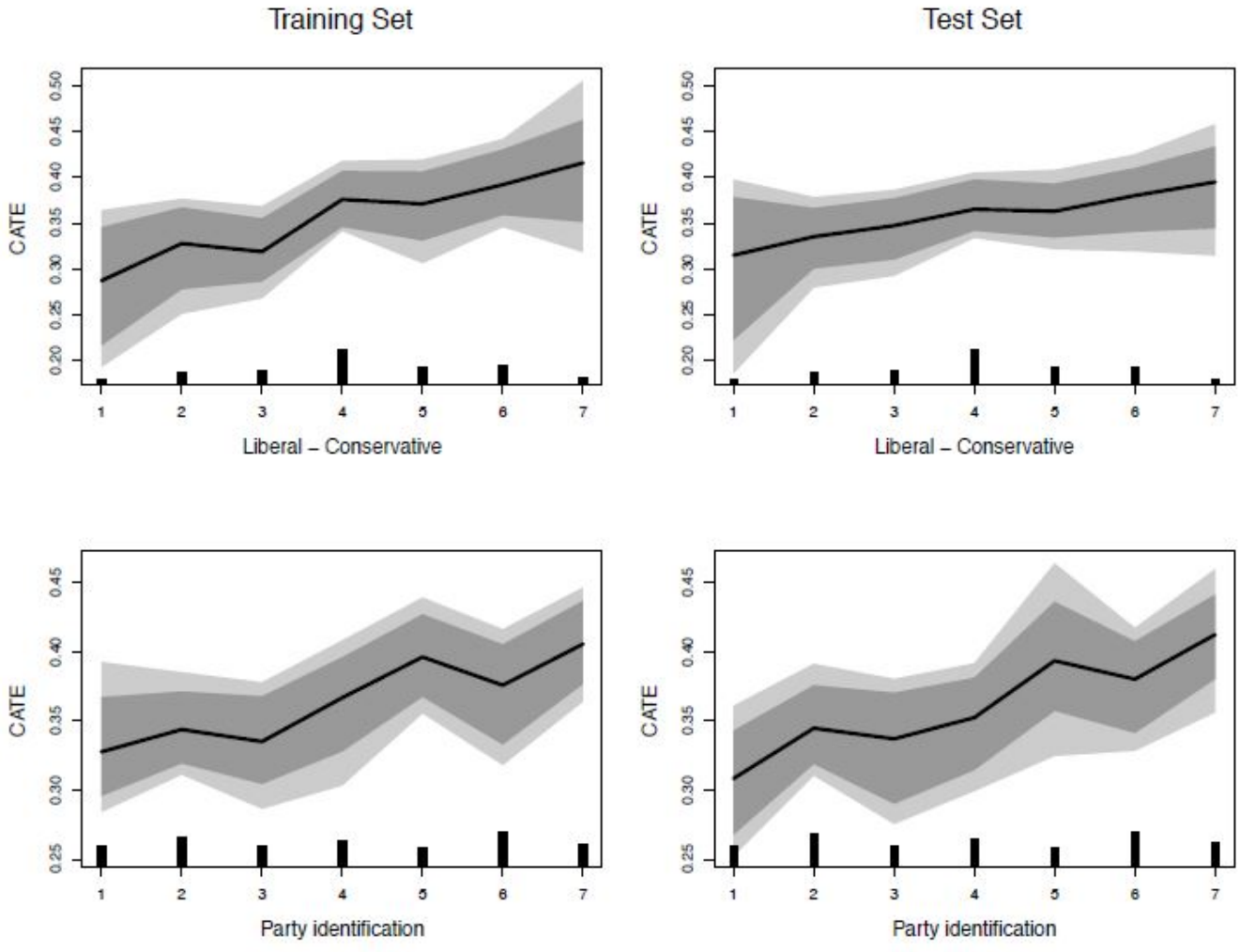
Test for treatment effect heterogeniety with two-sided Wald test (Cameron and Trivedi 2005)

- Party ID: Treatment effect increases with Republican Party ID and conservative ideology ($p = 0.029$)
- Education: Treatment effect not moderated by education ($p = 0.99$) towards blacks ($p < .001$)
- Variation across time (Start of Clinton presidency)

Overall Effect Heterogenity

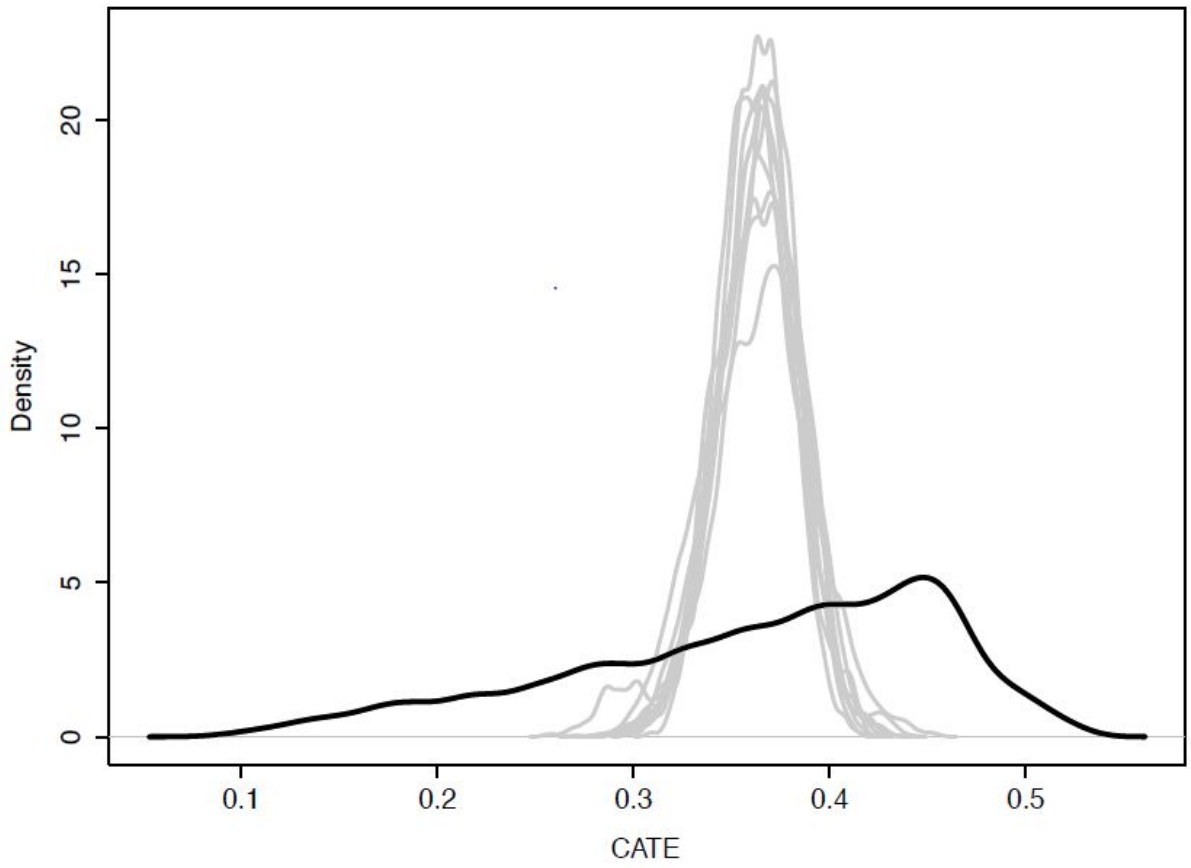


Cross Validation



Kernel-Density Plots of CATES

How much difference does allowing for systematic treatment effect heterogeneity make?



BART Conclusion

Approach provides a framework for investigation of systematic heterogeneity in large-scale experiments

BART is:

1. Automated (not ad-hoc)
2. Non-parametric (no functional form assumptions)
3. Cross-validated (split sample and multiple comparison problems)

Recursive Partioning

Alternative: "Honest" Recursive Partitioning

Athey and Imbens (2015) - Recursive Partitioning for Hterogeneous Causal Effects

Core proposal: Split the sample into two: 1. $S^t r$ to create the partition 2. $S^e st$ to estimate the conditional mean 3. "Honesty" in using difference information for model structure and model estimation

Recursive Partitioning

Tradeoffs: 1. Cost is sample size and precision: setting aside data for estimation leaves lest for training 2. Benefit is more valid confidence intervals and reduced sampling bias

Review of CART

We are interested in the conditional expectation

$$\mu(x) = E[Y_i|X_i = x]$$

CART takes place in two steps:

Tree building: CART recursively partions observations in training sample. - Identifies splits to achieve "in sampel goodness of fit" - Solve overfitting by estimating a penalty on tree depth

Cross-validation: select a complexity parameter for pruning - Estimate penalty by randomly selecting a cross-validation sample - Use cross-validation sample to choose a penalty parameter

CART for Prediction

In CART, we are interested in minimizing the Means Squared Error (MSE) of our partition:

$$Q^C(\pi) = -E_{S^{tes},S^{tr}}[MSE(S^{te}, S^{est}, \pi(S^{tr}))]$$

$$MSE(S^{te}, S^{est}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \mu^2(X_i; S^{tr}, \Pi)$$

RP for Prediction

In honest estimation, we are interested in the same measure over the test and estimation sample:

$$Q^H(\pi) = -E_{S^{tes},S^{est},S^{tr}}[MSE, S^{te}, S^{est}, \pi(S^{tr}))]$$

$$\hat{EMSE}(S^{tr}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \mu^2(X_i; S^{tr}, \Pi) - \frac{2}{N^{tr}} \cdot \sum_{\ell \in \Pi} S^2_{S^{tr}}(\ell)$$

Note: this doesn't make much of a difference for prediction, because gains in *MSE* and *EMSE* are proportional.

CART for Treatment Effects

CART estimator:

$$MSE(S^{te}, S^{est}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \tau^2(X_i; S^{tr}, \Pi)$$

RP for Treatment Effects

"Honest" estimator:

$$\hat{EMSE}(S^{tr}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \tau^2(X_i; S^{tr}, \Pi)$$

$$- \frac{2}{N^{tr}} \cdot \sum_{\ell \in \Pi} \frac{S^2_{S^{tr}_{Tr}}(\ell)}{p} + \frac{S^2_{S^{tr}_{CO}}(\ell)}{1-p}$$

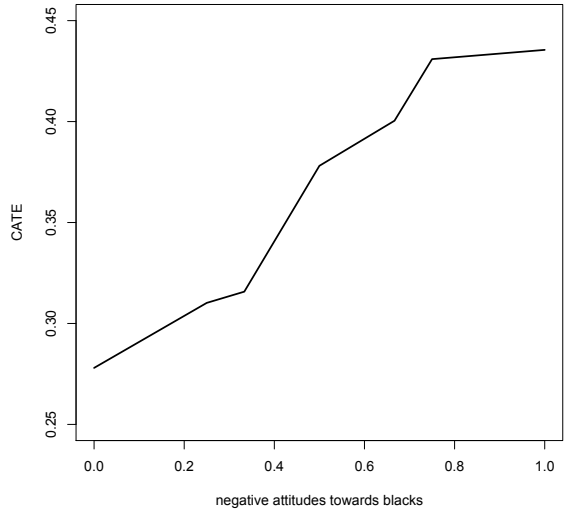
These are NOT proportional because covariates may effect the outcome but not the treatment effect, you can reduce the variance of a treatment effect estimator by introducing a split even if both child leaves have the same treatment effect.

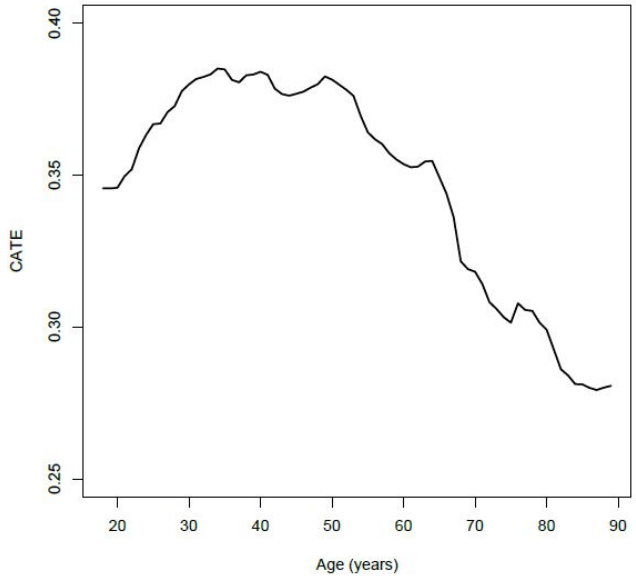
BART v RP Comparison - GSS

Does recursive partitioning (RP) make a difference in empirical settings? Lets return to the GSS analysis:

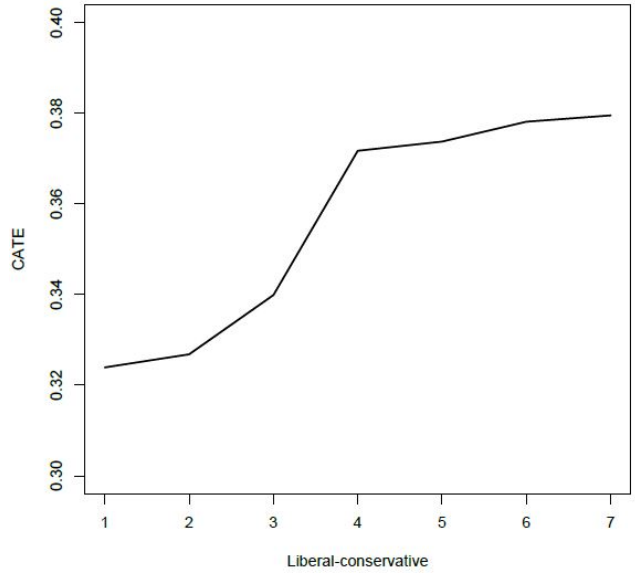
Thankfully, same ATE: 0.36

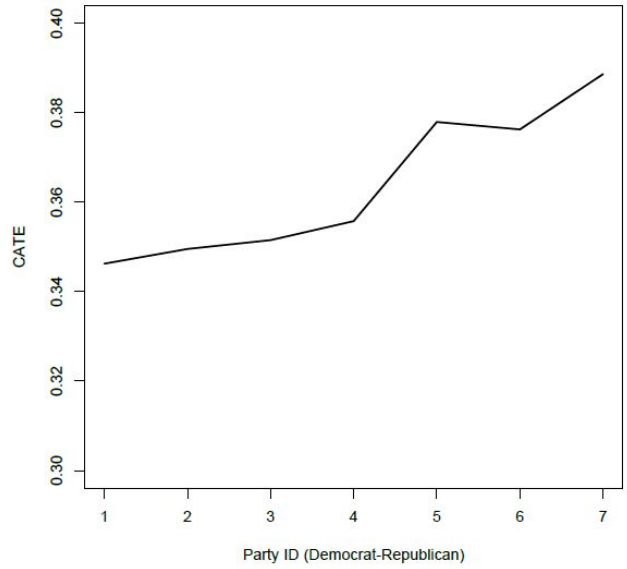
BART v RP Comparison





BART v RP Comparison





BART v RP Comparison

Table 1: Simulation Study

Design	1		2		3	
$N^{\text{tr}} = N^{\text{est}}$	500	1000	500	1000	500	1000
Estimator	Number of Leaves					
TOT	2.8	3.4	2.1	2.7	4.7	6.1
F-A	6.1	13.2	6.3	13.1	6.1	13.2
TS-A	4.0	5.6	2.5	3.3	4.4	8.9
CT-A	4.0	5.7	2.3	2.5	4.5	6.2
F-H	6.1	13.2	6.4	13.3	6.3	13.4
TS-H	4.4	7.7	5.3	11.0	6.0	12.3
CT-H	4.2	7.5	5.3	11.2	6.2	12.3
Infeasible MSE Divided by Infeasible MSE for CT-H*						
TOT-H	1.77	2.12	1.03	1.04	1.03	1.05
F-A	1.93	1.54	1.69	2.07	1.63	2.08
TS-H	1.01	1.02	1.06	0.99	1.24	1.38
CT-H	1.00	1.00	1.00	1.00	1.00	1.00
Ratio of Infeasible MSE: Honest to Adaptive**						
TOT-H/TOT-A		0.99		0.86		0.76
F-H/F-A		0.50		0.98		0.91
TS-H/TS-A		0.92		0.90		0.85
CT-H/CT-A		0.91		0.93		0.76
Coverage of 90% Confidence Intervals - Adaptive						
TOT-A	0.83	0.86	0.83	0.83	0.74	0.79
F-A	0.89	0.89	0.86	0.86	0.82	0.82
TS-A	0.85	0.85	0.80	0.83	0.77	0.80
CT-A	0.85	0.85	0.81	0.83	0.80	0.81
Coverage of 90% Confidence Intervals - Honest						
TOT-H	0.90	0.89	0.90	0.92	0.89	0.89
F-H	0.91	0.90	0.90	0.90	0.90	0.89
TS-H	0.89	0.90	0.90	0.90	0.90	0.90
CT-H	0.90	0.90	0.90	0.89	0.90	0.90

*$\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{Estimator}}(\mathcal{S}^{\text{tr}}))/\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{CT-H}}(\mathcal{S}^{\text{tr}}))$

**$\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}} \cup \mathcal{S}^{\text{tr}}, \pi^{\text{Estimator-A}}(\mathcal{S}^{\text{est}} \cup \mathcal{S}^{\text{tr}}))/ \quad \text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{Estimator-H}}(\mathcal{S}^{\text{tr}}))$

RP Conclusion

What does an "honest approach" mean in practice?

1. Divides data for tree building and parameter estimation
2. Performs similar cross-validation exercise
3. Sacrifices sample size for confidence interval coverage rates
4. In many situations, doesn't seem to matter much

Machine learning more broadly

BART and RP are two of MANY machine learning approaches to estimating heterogenous treatment effects:

1. Regression trees (Imai and Strauss, 2011)
2. BART (Holger and Green, 2012)
3. LASSO (Imai and Ratkovic, 2013)
4. Kernal Regularized Least Squares (Gelman et al 2008,, Hainmueller and Hzlett, 2014)
5. Elastic-Net (haste et al, 2001)
6. Ensemble of all methods (Grimmer et al, 2017)

Conclusion

Trying to avoid three problems in investigating CATEs

1. Ad-hockery and data dredging i
2. Non-statistical uncertainty
3. Multiple comparisons problem

The spirit of the exercise implies that data should be determining the methhods, with as little room for monkey business as possible.

MCMC

"Sculpting a complex figure by adding and subtracting small dabs of clay" (Chipman et al 2010)

- Start with m simple single node trees
- Iteratively increase (grow), decrease (prune) terminal nodes or change decision rule.
- Look for convergence

Predict Y after burn-in sample $f_1^*, ...f_K^*$ with the mean:

$$\frac{1}{K}\sum_{k=1}^{K} f_k^*(x)$$