# Bayesian nonparametric modeling for causal inference

## Jennifer L. Hill (2011)

2022120096 이재용

# Table of contents

# Background and motivation

# Background and motivation

**Previous strategies for identifying causal effects in nonexperimental settings**

Many studies assume ignorability of the treatment assigment mechanism and fit two models:

(1) one for the assignment mechanism

(2) one for the response surface (model for the outcome conditional on the treatment and confounding covariates)

Then the causal effects are estimated by the conditional expectations given the treatment assignment and confounding covariates.

# Background and motivation

**Previous strategies for identifying causal effects in nonexperimental settings**

Problem 1: the estimation of the conditional expectations is difficult if the distribution of X is different across treatment groups.

Problem 2: the difficulty in estimating the conditional expectations required for causal inference is exacerbated when, more plausibly, there are many confounding covariates or uncertainty about which predictors are needed to satisfy ignorability.

There have been many studies to tackle these problems, however they require a higher level of researcher sophistication to understand and implement.

# Background and motivation

## Study objective (proposed strategy)

A robust but simpler modeling approach by focusing on accurately modeling the response surface using a Bayesian nonparametric model called Bayesian additive regression trees (BART). BART has following advantages:

(1) Simple to implement

(2) BART can detect interactions and nonlinearities in the response surface, and identify heterogenous treatment effects

(3) Natural quantification of uncertainty measures by posterior intervals

(4) Accurate treatment effect point estimates in the nonlinear settings, yet very capable for linear settings as well.

Thus, BART is a simple method that is both robust and accurate in the estimation of causal effects.

# Causal effect

# Causal effect

## Recap

For, binary treatment covariate Z, where Z = 1: assignment to treatment, Z = 0: assignment to control, potential outcomes for individual $i$ is defined as

$$Y_i(1) = Y_i^{Z=1},\ Y_i(0) = Y_i^{Z=0}.$$

Since we cannot observe both of these values, researchers generally focus on estimating average treatment effects defined over sample or population or the conditional average treatment effects.

Sample average treatment effect (SATE), sample average effect of the treatment on the treated (SATT), PATE, PATT, CATE, and CATT.

# Causal effect

## Recap

In observational studies potential outcomes are typically not independent of treatment assignments. Thus we identify the average causal effects under assumptions such as strong ignorability of treatment assignment.

Strong ignorability = unconfoundedness + common support,

unconfoundedness: $Y(0), Y(1) \perp Z \,|\, X$, (conditional independence) where $X$ is a vector of confounding covariates,

common support: $0 < P(Z = 1 \,|\, X) < 1$.

# Causal effect

## Recap

Then the estimation of the causal effects are done by the conditional expectations,

$$E[Y(1)\,|\,X] = E[Y\,|\,X, Z = 1],\ E[Y(0)\,|\,X] = E[Y\,|\,X, Z = 0].$$

However, the estimation is difficult if $Y(0)$ or $Y(1)$ are not linearly related to $X$ and if the distribution of $X$ is different across treatment groups.

Also it also becomes difficult when there are many confounding covariates or uncertainty about which predictors are needed to satisfy ignorability.

# Causal effect

## Previous studies to tackle the problem and the article's proposal

A host of new methods have been proposed in the past three decades to address this estimation problem. One of them is if the response surface is correctly specified, we do not have to worry about correctly specifying the assignment mechanism, i.e., focuses solely on **precise estimation of the response surface**.

Nonparametric and semiparametric versions of these methods are more robust but require a higher level of researcher sophistication to understand and implement.

This article proposes that the benefits of the Bayesian additive regression trees (BART) strategy in terms of simplicity, precision, robustness, and lack of required researcher interference.

# Bayesian additive regression trees (BART)

# BART

## How it works

Suppose we model an outcome variable $Y$ as

$$Y = f(z, x) + \epsilon, \epsilon \sim N(0, \sigma^2).$$

BART estimates $f(\cdot)$ nonparametrically as a sum of trees.

$$=> \quad Y = g(z, x; T_1, M_1) + g(z, x; T_2, M_2) + \cdots + g(z, x; T_m, M_m) + \epsilon,$$

Here, $g(\cdot)$: value obtained by a single tree,

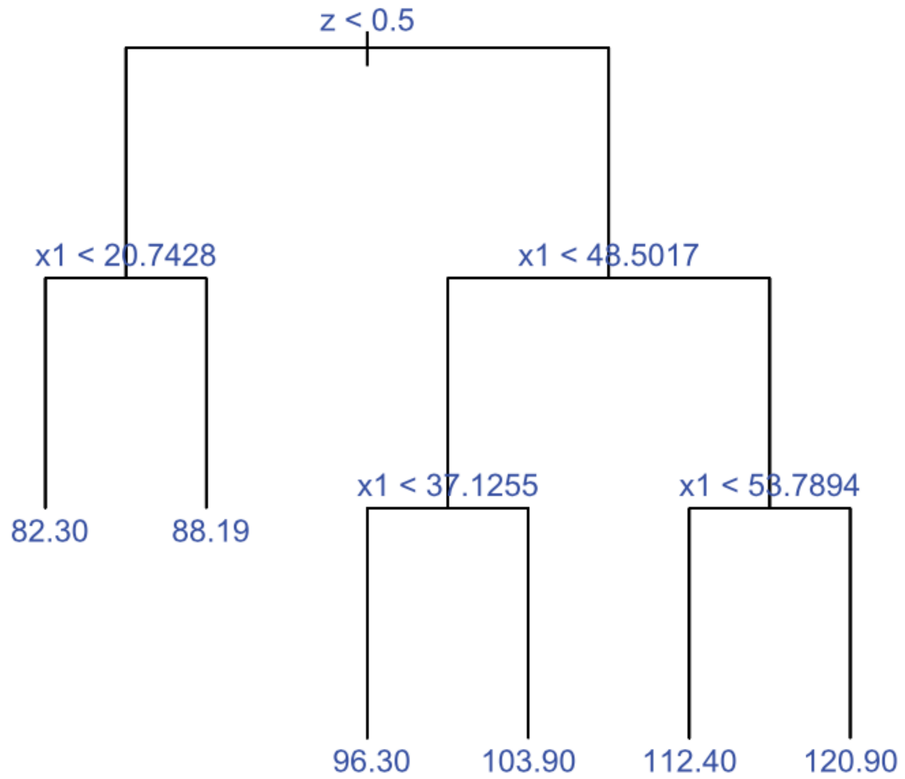
$T_j$: $j$th tree structure,

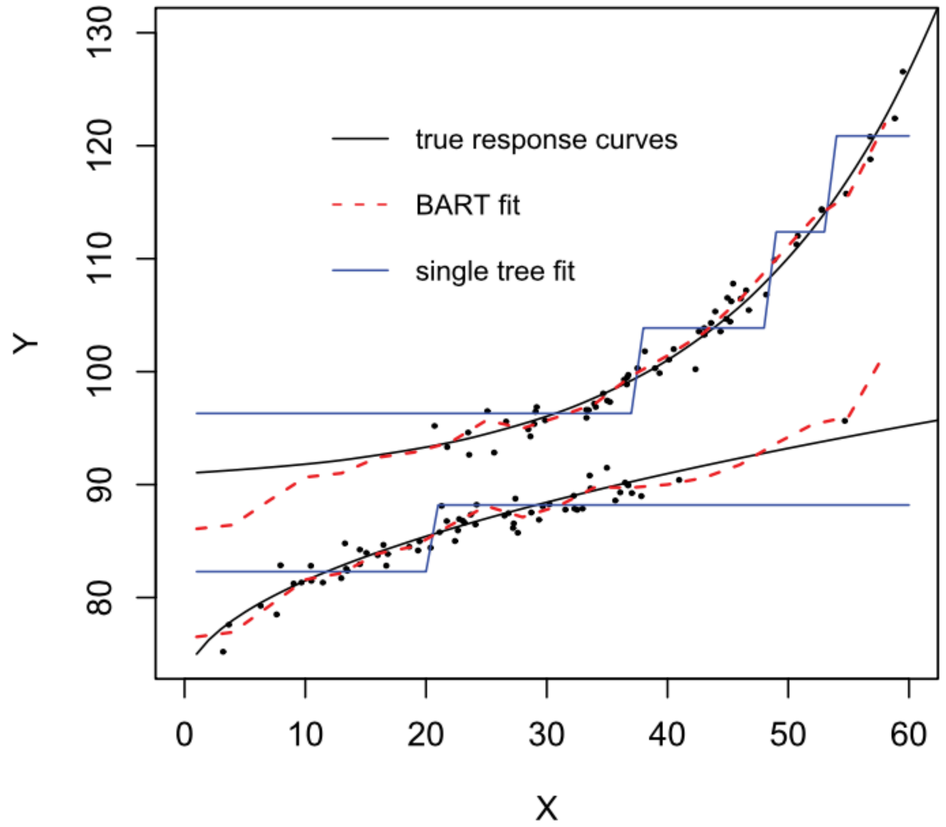
$M_j$: set of terminal nodes in $T_j$.

# BART

## How it works

$$Y = g(z, x; T_1, M_1) + g(z, x; T_2, M_2) + \cdots + g(z, x; T_m, M_m) + \epsilon,$$



# BART

## What set BART apart from other models

$(T_j, M_j)$, $\sigma$ are the parameters that are to be estimated under Bayesian statistics framework (use of prior and inference on posterior distributions).

Use of MCMC algorithm.

Use regularization prior to stop a tree from growing too large: boosting (combination of weak models outperform a single strong model).

# BART

## Advantages

BART's performance with default settings of priors is highly competitive in terms of predictions (no need to use cross validation to tune hyperparameters).

Natural uncertainty quantification by Bayesian posterior measures from MCMC samples.

BART captures both nonlinearities and interactions without parameter specification by researcher.

BART can handle very large number of predictors and can identify which variables are useful (by how much they get used in the trees).

BART = flexibility of machine learning + formality of Bayesian modeling framework.

# BART

## Estimating causal effects using BART

We want to estimate conditional average treatment effect for the treated (CATT),

$$\frac{1}{n}\sum_{i:Z_i=1}\left\{E\left[Y_i(1)\,|\,X_i\right]-E\left[Y_i(0)\,|\,X_i\right]\right\}=\frac{1}{n}\sum_{i:Z_i=1}\left\{f(1,x_i)-f(0,x_i)\right\},$$

and conditional average treatment effect for the controls (CATC),

$$\frac{1}{n}\sum_{i:Z_i=0}\left\{E\left[Y_i(1)\,|\,X_i\right]-E\left[Y_i(0)\,|\,X_i\right]\right\}=\frac{1}{n}\sum_{i:Z_i=0}\left\{f(1,x_i)-f(0,x_i)\right\}.$$

# Simulations based on real data

# Simulations based on real data

## Data source

Experimental data from Infant Health and Development Program (IHDP)

A randomized experiment began in 1985, targeted low-birth-weight, premature infants, and provided the treatment group with both intensive high-quality child care and home visits from a trained provider. The program was successful at raising cognitive test scores of the treated children relative to the controls at the end of the intervention.

The author will generate the outcome only.

# Simulations based on real data

## Covariates used

Pretreatment covariates:

Measurements on the child (birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex, twin status),

Behaviors engaged in during the pregnancy (smoking, drinking, drugs),

Measurements on the mother at the time she gave birth (age, marital status, educational attainment, whether she worked during pregnancy, received prenatal care),

The site in which the family resided at the start of the intervention.

=> 6 continuous covariates, 19 binary covariates.

# Simulations based on real data

## Study design

Starting with the experimental data, an observational data is created by throwing away a nonrandom portion of the treatment group: all children with nonwhite mothers.

This leaves treatment group with 139 children and the control group with 608 children.

Thus the treatment and control groups are no longer balanced and simple comparisons of outcome would lead to bias estimates of the treatment effect.

This design ensures that the overlap assumption is satisfied for the treatment group (i.e., the support of $X$ for the treated is a subset of the support of $X$ for the controls), however overlap is not ensured for the control group.

Therefore, the conditional average treatment effect on the treated (CATT) is referred to as the "overlap" setting, the conditional average treatment effect on the controls (CATC), is referred to as the "incomplete overlap" setting.

# Simulations based on real data

## Outcome generation

$p = 25$ confounding covariates are used to generate two different response surfaces.

Response surface A:

$$Y(0) \sim N(X\beta_A, 1),\ Y(1) \sim N(X\beta_A + 4, 1),$$

where the coefficients $\beta_A$ are randomly sampled values (0, 1, 2, 3, 4) with probabilities (0.5, 0.2, 0.15, 0.1, 0.05).

This response surface is linear and parallel across treatment groups and all the estimates such as CATT, CATC equal 4. Also there is no heterogeneous treatment effect (linear regression should have a good performance).

# Simulations based on real data

## Outcome generation

Response surface B:

$$Y(0) \sim N(e^{(X+W)\beta_B}, 1),\ Y(1) \sim N(X\beta_B - \omega_B^s, 1),$$

where $W$ is an offset matrix of the same dimension as $X$ with every value equal to 0.5, the coefficients $\beta_A$ are randomly sampled values (0, 0.1, 0.2, 0.3, 0.4) with probabilities (0.6, 0.1, 0.1, 0.1, 0.1). For the $s$th simulation, $\omega_B^s$ was chosen in the overlap setting, where we estimate the effect of the treatment on the treated such that CATT equals 4. Similarly, in the incomplete overlap setting, it was chosen so that CATC equals 4.

# Simulations based on real data

## Methods compared

Inference for CATT and CATC using BART is compared with estimates from

Linear regression

Propensity score matching

Propensity-score-based weighting estimator

Random forest (but the performance was so poor it was omitted)

# Simulation results

# Simulation results

## Evaluation metrics used

RMSE: Root mean squared error **(lower the better)**.,

Coverage: Coverage of the 95% intervals **(higher the better)**,

Missed: Percentage of times the method did not detect a significant treatment effect (Type II error) at a 5% significant level **(lower the better)**.,

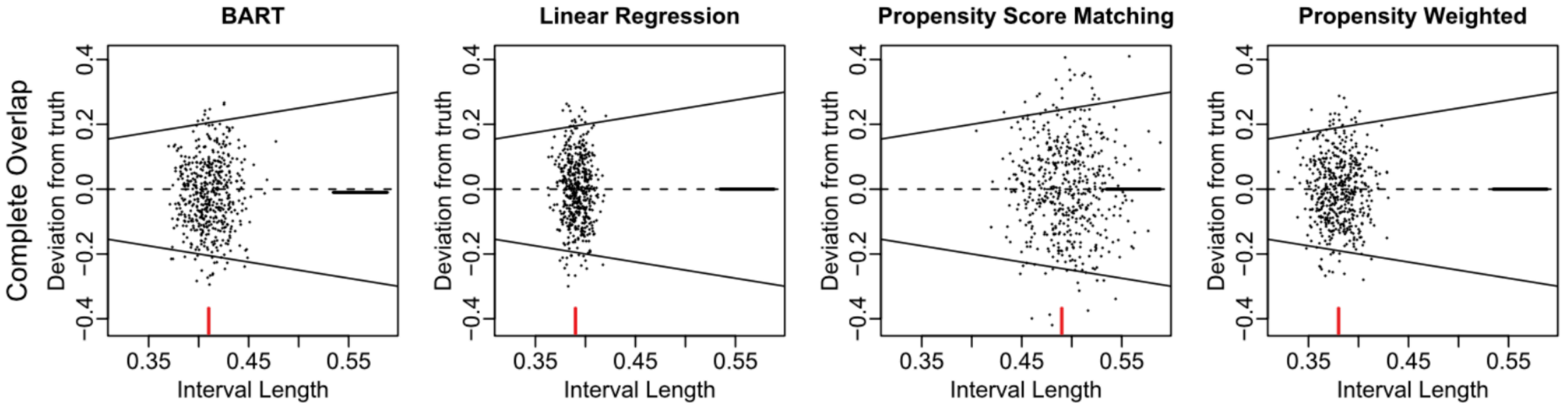
PEHE: Precision in estimation of heterogeneous effects which evaluates the ability of each method to capture treatment effect heterogeneity **(lower the better)**.

These summaries are calculated over 1000 simulations.

# Simulation results

## Results from 1000 simulation runs for response surface A

Estimating CATT




RMSE 0.1
Missed 0.00
Coverage 0.95
PEHE 0.2

RMSE 0.1
Missed 0.00
Coverage 0.94
PEHE 0.1

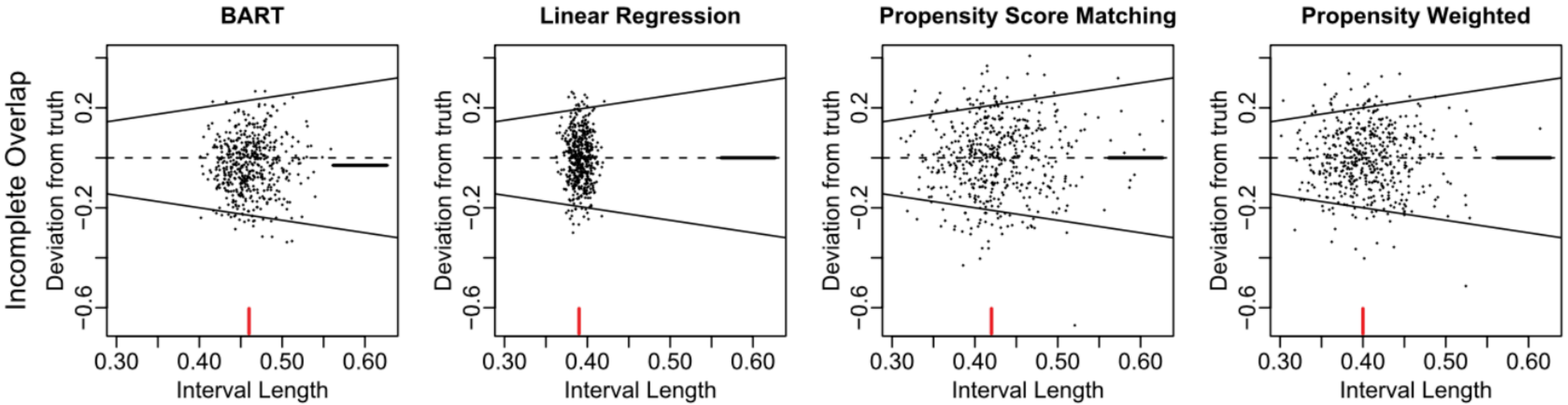
RMSE 0.1
Missed 0.00
Coverage 0.93
PEHE 0.1

RMSE 0.1
Missed 0.00
Coverage 0.93
PEHE 0.1

# Simulation results

## Results from 1000 simulation runs for response surface A

Estimating CATC




RMSE 0.1
Missed 0.00
Coverage 0.95
PEHE 0.2

RMSE 0.1
Missed 0.00
Coverage 0.94
PEHE 0.1

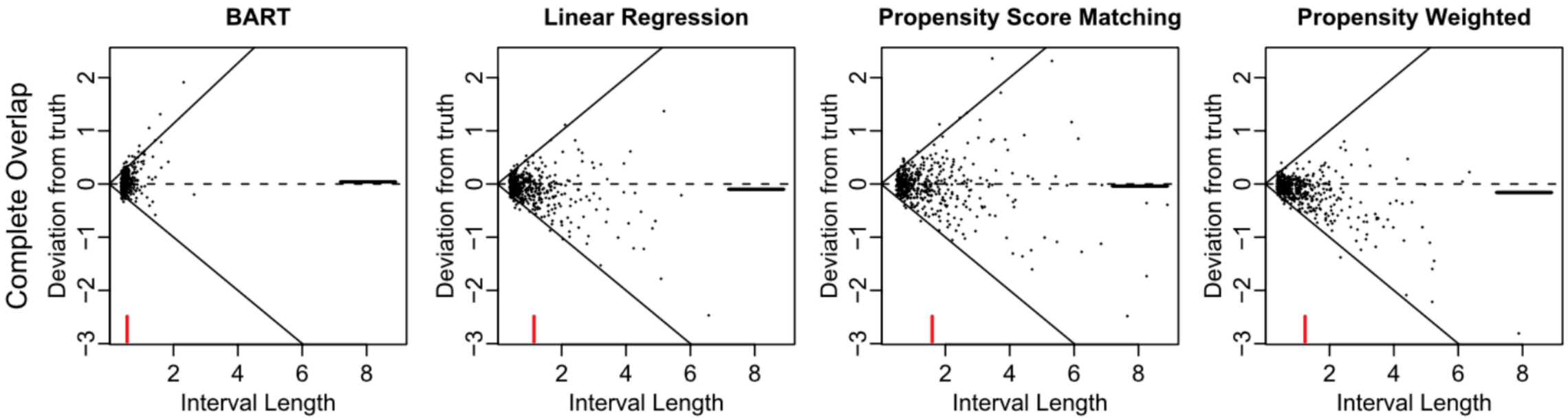
RMSE 0.2
Missed 0.00
Coverage 0.85
PEHE 0.1

RMSE 0.1
Missed 0.00
Coverage 0.89
PEHE 0.1

# Simulation results

## Results from 1000 simulation runs for response surface B

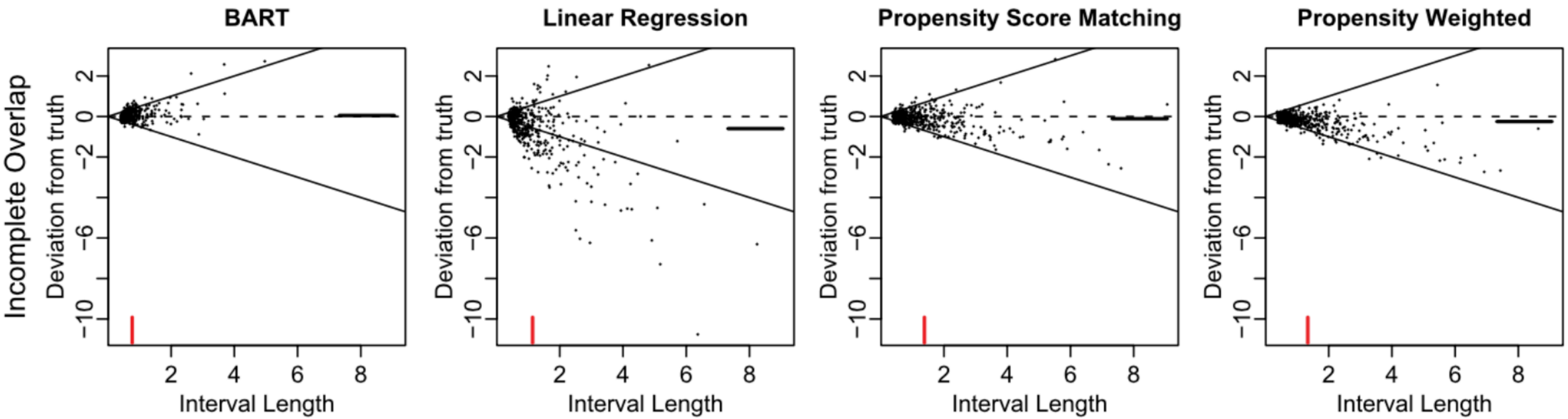
Estimating CATT



# Simulation results

## Results from 1000 simulation runs for response surface B

Estimating CATC




RMSE 0.3
Missed 0.00
Coverage 0.89
PEHE 1.9

RMSE 1.4
Missed 0.05
Coverage 0.49
PEHE 5.4

RMSE 0.5
Missed 0.01
Coverage 0.92
PEHE 5.3

RMSE 0.5
Missed 0.02
Coverage 0.90
PEHE 5.4

# Discussion

# Discussion

## Summary and possible improvements

The article proposes BART as a very capable model for precise estimation of response surfaces.

BART has several advantages, mainly its simple implementation without requiring technical sophistication from a researcher (compared to previous methodologies).

BART performed very well in simulation with linear specification and outperformed competitive models in nonlinear specification. BART was also able to identify heterogeneous treatment effects well.

The article only tested on few simulation scenarios so more work can be done in broader range of settings.

# Q & A