The University of Texas at Austin
McCombs School of Business

Causal inference, an overview and where ML fits in

David Puelz

April 28, 2022

Outline

Causality?

Model-based causal inference (+ML)

Causality

A very broad field that is gaining momentum.

We have advanced models for prediction, why not figure out how to use them for (causal) inference?

Answer: That's a great idea! But inference is *much* harder than prediction. We have to be careful (and skeptical).

Notation for any causal discussion

- units of study are indexed by i (states, stores, customers, people, time, rows in dataframe, ...)

Notation for any causal discussion

- units of study are indexed by i (states, stores, customers, people, time, rows in dataframe, ...)
- treatment for unit i is given by z_i

$$z_i = \begin{cases} 0 & \text{not treated (control)} \\ 1 & \text{treated} \end{cases}$$

Notation for any causal discussion

- units of study are indexed by i (states, stores, customers, people, time, rows in dataframe, ...)
- treatment for unit i is given by z_i

$$z_i = \begin{cases} 0 & \text{not treated (control)} \\ 1 & \text{treated} \end{cases}$$

- denote a outcome of interest for each i as Y_i, but let's write it as a function of the treatment.

$$Y_i(z_i)$$

Notation for any causal discussion

- units of study are indexed by i (states, stores, customers, people, time, rows in dataframe, ...)
- treatment for unit i is given by z_i

$$z_i = \begin{cases} 0 & \text{not treated (control)} \\ 1 & \text{treated} \end{cases}$$

- denote a outcome of interest for each i as Y_i, but let's write it as a function of the treatment.

$$Y_i(z_i) \leftarrow \textbf{potential outcome} \text{ for unit } i$$

Note: The treatment can be continuous, too. Binary is useful in many contexts and defining potential outcomes are easy.

Example: COVID-19 lockdowns

How did state lockdowns at the beginning of 2020 affect the spread of the virus?

Example: COVID-19 lockdowns

How did state lockdowns at the beginning of 2020 affect the spread of the virus?

What are the potential outcomes?

Example: COVID-19 lockdowns

Let's set up the structure of this problem.

Let i denote a state, so

$$i \in \{\text{New York, California, Florida, Texas, South Dakota, ... }\}$$

First, we define what z_i is:

$$z_i = \begin{cases} 0 & \text{no lockdown} \\ 1 & \text{lockdown} \end{cases}$$

Example: COVID-19 lockdowns

Let's set up the structure of this problem.

Let i denote a state, so

$$i \in \{\text{New York, California, Florida, Texas, South Dakota, ... }\}$$

First, we define what z_i is:

$$z_i = \begin{cases} 0 & \text{no lockdown} \\ 1 & \text{lockdown} \end{cases}$$

Second, we define our outcome: Y_i : let's choose the cases per capita (in state i) after lockdown or no lockdown.

Organizing our data: The Science Table

i **(state)**	z_i **(lockdown)**	$Y_i(0)$	$Y_i(1)$
New York			
Florida			
California			
Texas			
South Dakota			
Illinois			
⋮	⋮	⋮	⋮

Organizing our data: The Science Table

i **(state)**	z_i **(lockdown)**	$Y_i(0)$	$Y_i(1)$
New York	1		
Florida	0		
California	1		
Texas	0		
South Dakota	0		
Illinois	1		
⋮	⋮	⋮	⋮

Organizing our data: The Science Table (CDC, cases/100k)

i **(state)**	z_i **(lockdown)**	$Y_i(0)$	$Y_i(1)$
New York	1		.0034
Florida	0	.007	
California	1		.0014
Texas	0	.004	
South Dakota	0	.0028	
Illinois	1		.002
⋮	⋮	⋮	⋮

Organizing our data: The Science Table (CDC, cases/100k)

i **(state)**	z_i **(lockdown)**	$Y_i(0)$	$Y_i(1)$
New York	1	??	.0034
Florida	0	.007	??
California	1	??	.0014
Texas	0	.004	??
South Dakota	0	.0028	??
Illinois	1	??	.002
⋮	⋮	⋮	⋮

What is the ideal scenario?

We are able to know both of the potential outcomes for each state!

Y_{NY}  and Y_{NY} 

Defining a causal effect for NY

We can define the **causal effect** of the "lockdown treatment" as the difference between the two potential outcomes.

$$\tau_{\text{NY}} = Y_{\text{NY}}\left(\text{  }\right) - Y_{\text{NY}}\left(\text{  }\right)$$

or written more generally:

$$\tau_i = Y_i(1) - Y_i(0)$$

The fundamental problem of causal inference

We only observe one of the two potential outcomes for New York and all other states. In general, we always only observe one of two potential outcomes.

The unknown outcomes are called the missing potential outcomes or counterfactuals. This is what makes causality a tough task ... it is a **missing data problem**.

Is all hope lost?

What do we do?

→ Model-based causal inference – write down a probability model for the potential outcomes, and use observed data to infer the causal effect.

- regression (linear & nonlinear)
- often used with observational data

→ Randomization-based causal inference – assume the potential outcomes are equal (Fisher null), and test for a causal effect.

- simple and robust
- used with randomized experimental data
- only probability distribution is experimental design

Regularized causal effect estimation (linear)

Kale

Suppose we're interested in the **treatment effect** of dietary kale intake.

And want to know how effective it is at lowering cholesterol, which is our **outcome variable**.

Unfortunately, we have only observational data (i.e., not a randomized study).

Kale intake predicts exercise

Our bad luck, only gym-rats seem to eat much kale. And exercise is known to lower cholesterol: the "direct" effect is **confounded**.

$$Y_i = \beta_0 + \alpha Z_i + \varepsilon_i,$$

Because $\text{cov}(Z_i, \varepsilon_i) \neq 0$, we can write

$$Y_i = \beta_0 + \alpha Z_i + \omega Z_i + \tilde{\varepsilon}_i.$$

Since $\text{cov}(Z_i, \tilde{\varepsilon}_i) = 0$, we mis-estimate α as $\alpha + \omega$.

We must "adjust" for weekly exercise

The good news is, we can **control** for weekly exercise, X_i, by including it in the regression:

$$Y_i = \beta_0 + \alpha Z_i + \beta X_i + \varepsilon_i.$$

This "clears out" the confounding: conditional on X_i, $\text{cov}(Z_i, \varepsilon_i) = 0$ and we're good to go.

But what if we don't know what we need to control for?

The problem with the kitchen sink

So what is wrong with just including anything we can think of in our regression?

The problem is that we typically have finite data. The more things we add, the more variable (untrustworthy) our estimator will be.

How do people deal with this in practice?

The "con" of econometrics (1983)

Let's Take the Con out of Econometrics

By EDWARD E. LEAMER*

Econometricians would like to project the image of agricultural experimenters who divide a farm into a set of smaller plots of land and who select randomly the level of fertilizer to be used on each plot. If some plots are assigned a certain amount of fertilizer while others are assigned none, then the difference between the mean yield of the fertilized plots and the mean yield of the unfertilized plots is a measure of the effect of fertilizer on agricultural yields. The econometrician's humble job is only to determine if that difference is large enough to suggest a real effect of fertilizer, or is so small that it is more likely due to random variation.

One should not jump to the conclusion that there is necessarily a substantive difference between drawing inferences from experimental as opposed to nonexperimental data. The images I have drawn are deliberately prejudicial. First, we had the experimental scientist with hair neatly combed, wide eyes peering out of horn-rimmed glasses, a white coat, and an electronic calculator for generating the random assignment of fertilizer treatment to plots of land. This seems to contrast sharply with the nonexperimental farmer with overalls, unkempt hair, and bird droppings on his boots. Another image, drawn by Orcutt, is even more damaging:

Enter ML and regularization

It is well-known that shrinkage priors (e.g., point-mass priors) allow us to "safely" include many covariates in a regression (even more than our sample size!)

Classical stats – LASSO, ridge, stepwise selection ...
Bayesian stats – spike-and-slab & horseshoe priors ...

We have lots of theory backing this up, too.

- bias-variance trade-off intuitions
- "bet on sparsity" ideas

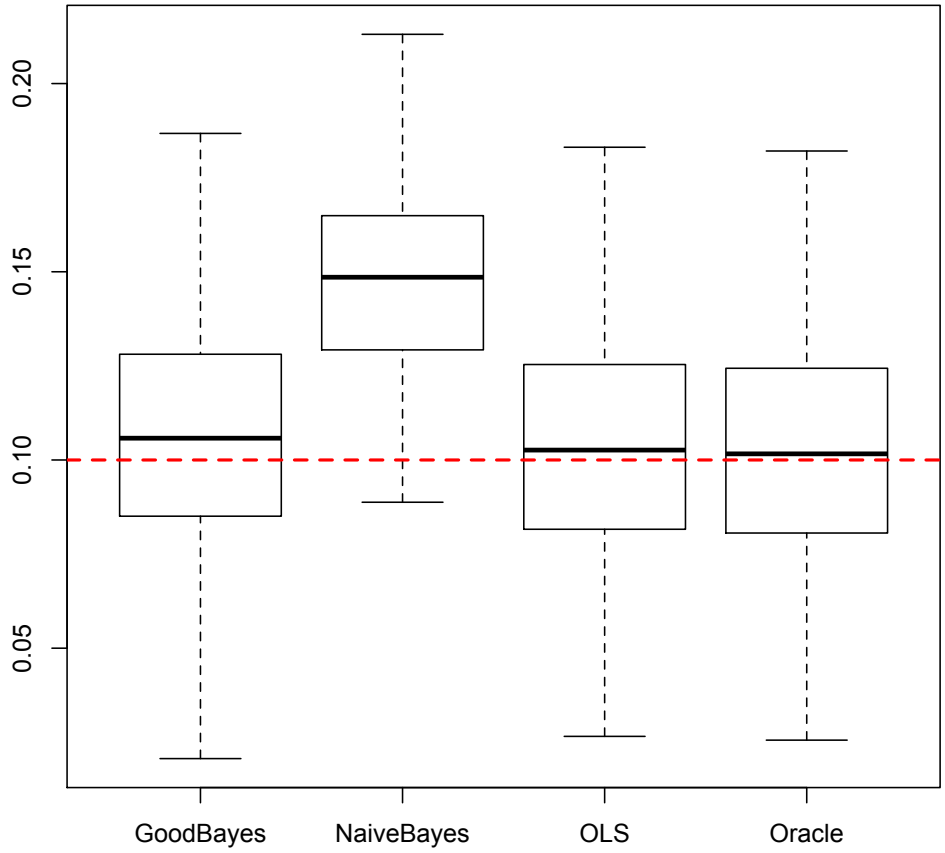
The obvious approach

$$Y_i = \beta_0 + \alpha Z_i + \beta X_i + \varepsilon_i.$$

- a flat prior on the treatment effect: $(\alpha, \sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon$,
- shrinkage prior on β (e.g., a horseshoe prior).

And we're off to the races!

oops



It turns out that this "obvious" approach is really bad at getting reasonable estimates of the treatment effect α.

Bad bias versus good bias

Assume that:

$$Z_i = \mathbf{X}_i^t \gamma + \epsilon_i.$$

Now substitute a shrunk estimate, $\beta - \Delta$, in place of the true (unknown) β vector:

$$Y_i = \alpha Z_i + \mathbf{X}_i^t(\beta - \Delta) + [\varepsilon_i + \mathbf{X}_i^t \Delta].$$

This implies that ε_i is taken to be $\varepsilon_i + \mathbf{X}_i^t \Delta$, which gives

$$\text{cov}(\mathbf{X}_i^t \gamma + \epsilon_i, \varepsilon_i + \mathbf{X}_i^t \Delta) \neq 0.$$

Biasing β towards zero biases $\text{cov}(Z, \varepsilon)$ away from zero!

Regularization-induced confounding (RIC)

This phenomenon biases inference in linear models, but it will also do the same for anything else that uses regularization:

- Random forests
- Neural nets
- LASSO and ridge
- all other ML methods!

Our solution for the linear model

Typical parameterization

$$\text{Selection Eq.:} \quad Z = \mathbf{X}^t\gamma + \epsilon, \qquad \epsilon \sim N(0, \sigma_\epsilon^2),$$

$$\text{Response Eq.:} \quad Y = \alpha Z + \mathbf{X}^t\beta + \nu, \qquad \nu \sim N(0, \sigma_\nu^2).$$

Our reparameterization: a latent error approach

We reparameterize as

$$\begin{pmatrix} \alpha \\ \beta + \alpha\gamma \\ \gamma \end{pmatrix} \rightarrow \begin{pmatrix} \alpha \\ \beta_d \\ \beta_c \end{pmatrix}.$$

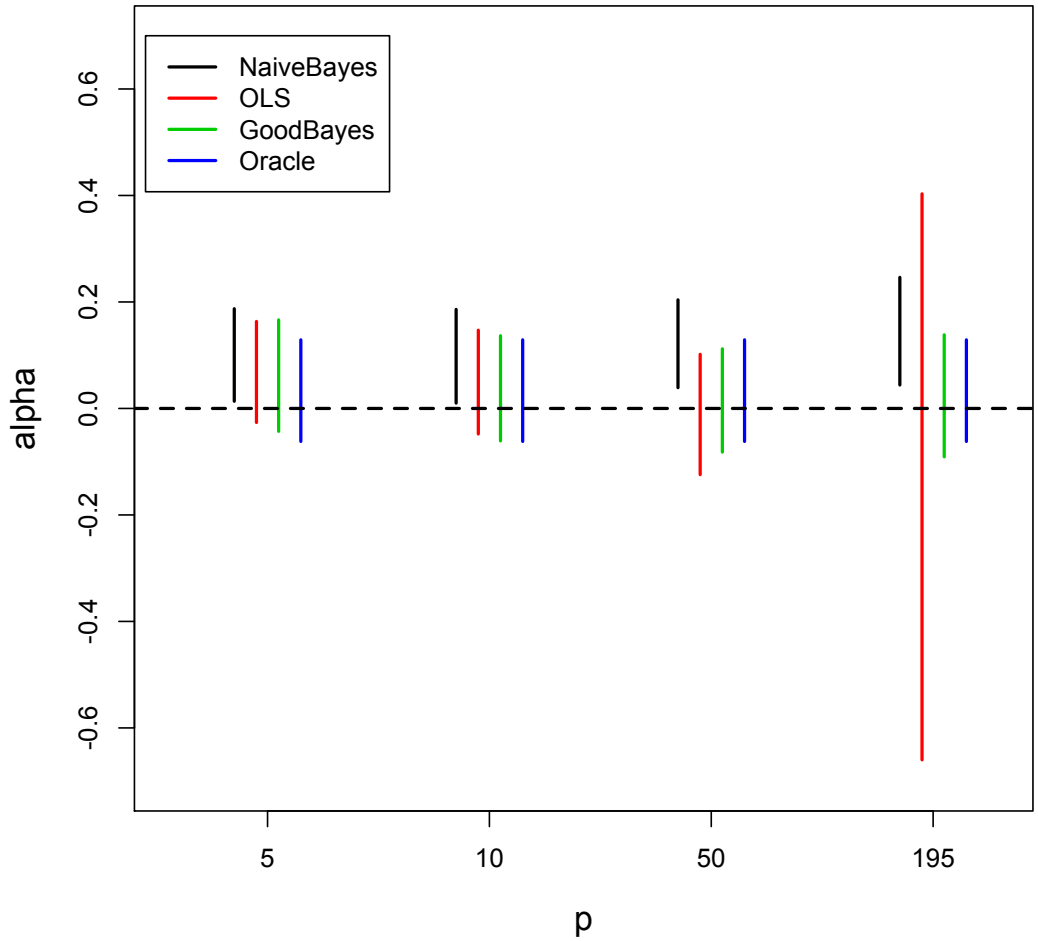
which gives the new equations

$$\begin{aligned} &\text{Selection Eq.:} && Z = \mathbf{X}^t\beta_c + \epsilon, && \epsilon \sim N(0, \sigma_\epsilon^2), \\ &\text{Response Eq.:} && Y = \alpha(Z - \mathbf{X}^t\beta_c) + \mathbf{X}^t\beta_d + \nu, && \nu \sim N(0, \sigma_\nu^2). \end{aligned}$$

We can now shrink β_d and β_c with impunity!

Note: This can also work by including $\hat{Z}$ as an unpenalized covariate with the original treatment Z.

Variance reduction with minimal effect biasing



Eventually OLS is no good (too variable). The obvious (naive) approach breaks way before that!

Regularized causal effect estimation (nonlinear)

Moving from this ...

$$Y = \beta_0 + \alpha Z + \beta X + \varepsilon$$

To this!

$$Y = f(X, Z) + \varepsilon$$

To this!

$$Y = f(X, Z) + \varepsilon$$

$\rightarrow$ default (supervised) framework for causal inference

$\rightarrow$ a rich output that allows you to ask many questions

Conditional average treatment effect

Estimand of interest:

$$\tau(\mathbf{x}_i) := E(Y_i|\mathbf{x}_i, Z_i = 1) - E(Y_i|\mathbf{x}_i, Z_i = 0)$$

So, treatment effect estimation is just response surface estimation!

Conditional average treatment effect

Assuming mean zero additive errors:

$$Y_i = f(\mathbf{x}_i, Z_i) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

such that

$$E(Y_i \mid \mathbf{x}_i, Z_i = z_i) = f(\mathbf{x}_i, z_i)$$

Then our quantity of interest is

$$\tau(\mathbf{x}_i) = f(\mathbf{x}_i, 1) - f(\mathbf{x}_i, 0)$$

So, how do we regularize (model) f?

BART

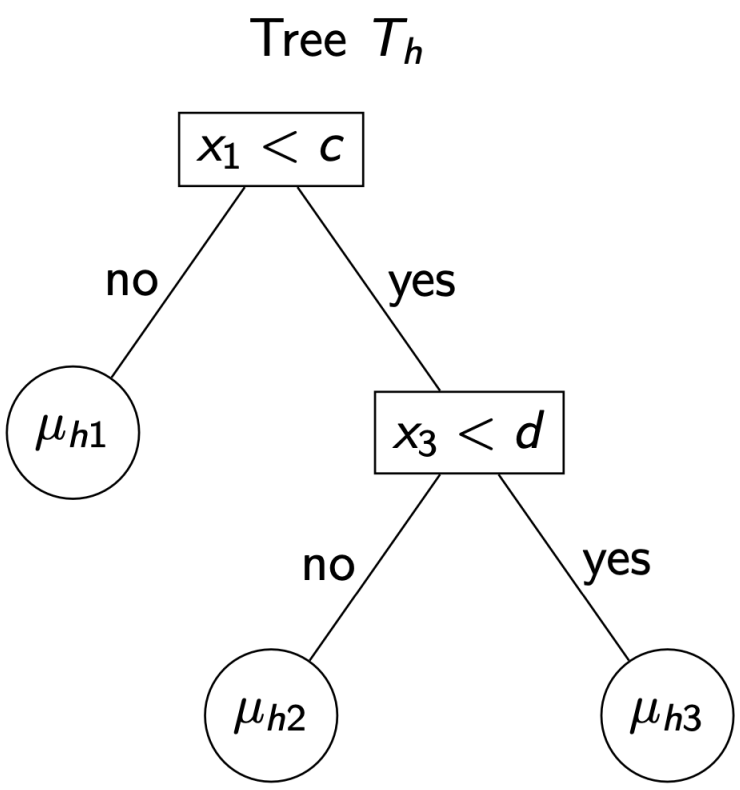
Bayesian Additive Regression Trees (BART) from Chipman, George, & McCulloch, (2008):

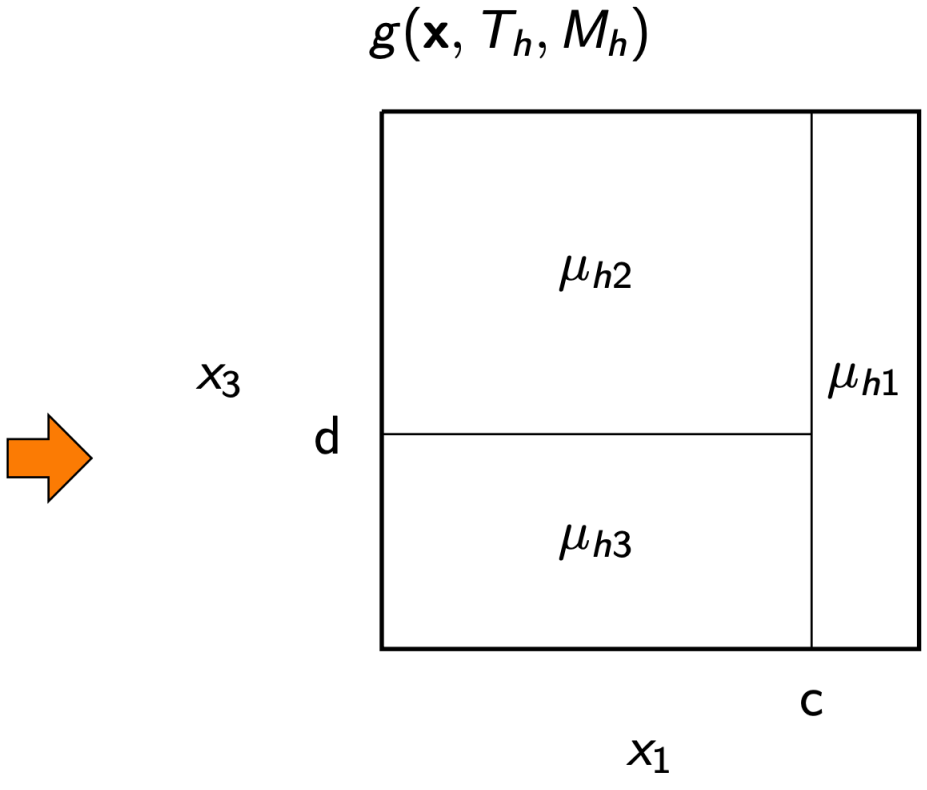
$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$f(\mathbf{x}) = \sum_{h=1}^{m} g(\mathbf{x}, T_h, M_h)$$

- Tree growth is probabilistic
- Results in ensembles of smaller, simpler trees (regularization!)

Regression trees



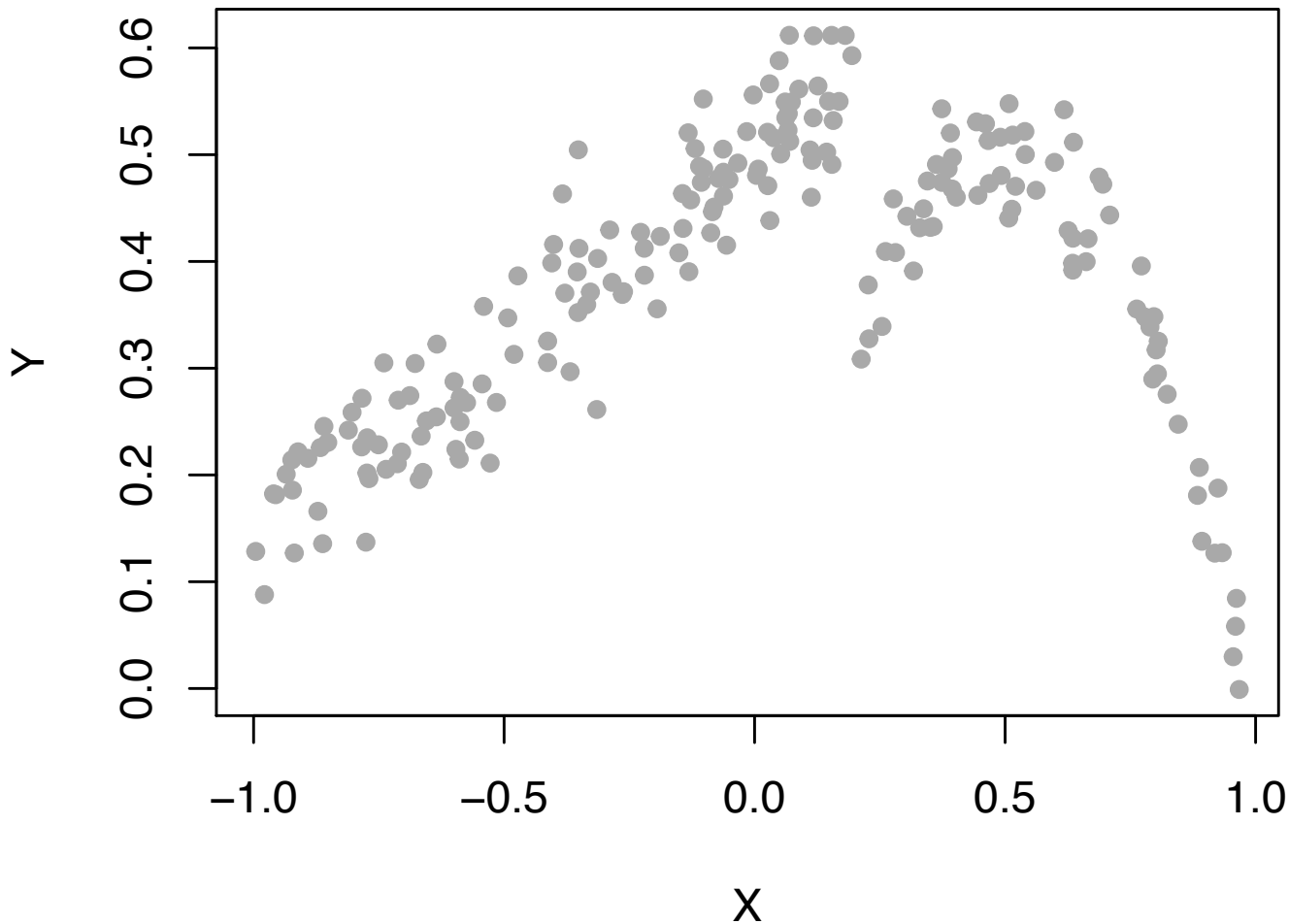


Leaf/End node parameters
$M_h = (\mu_{h1}, \mu_{h2}, \mu_{h3})$

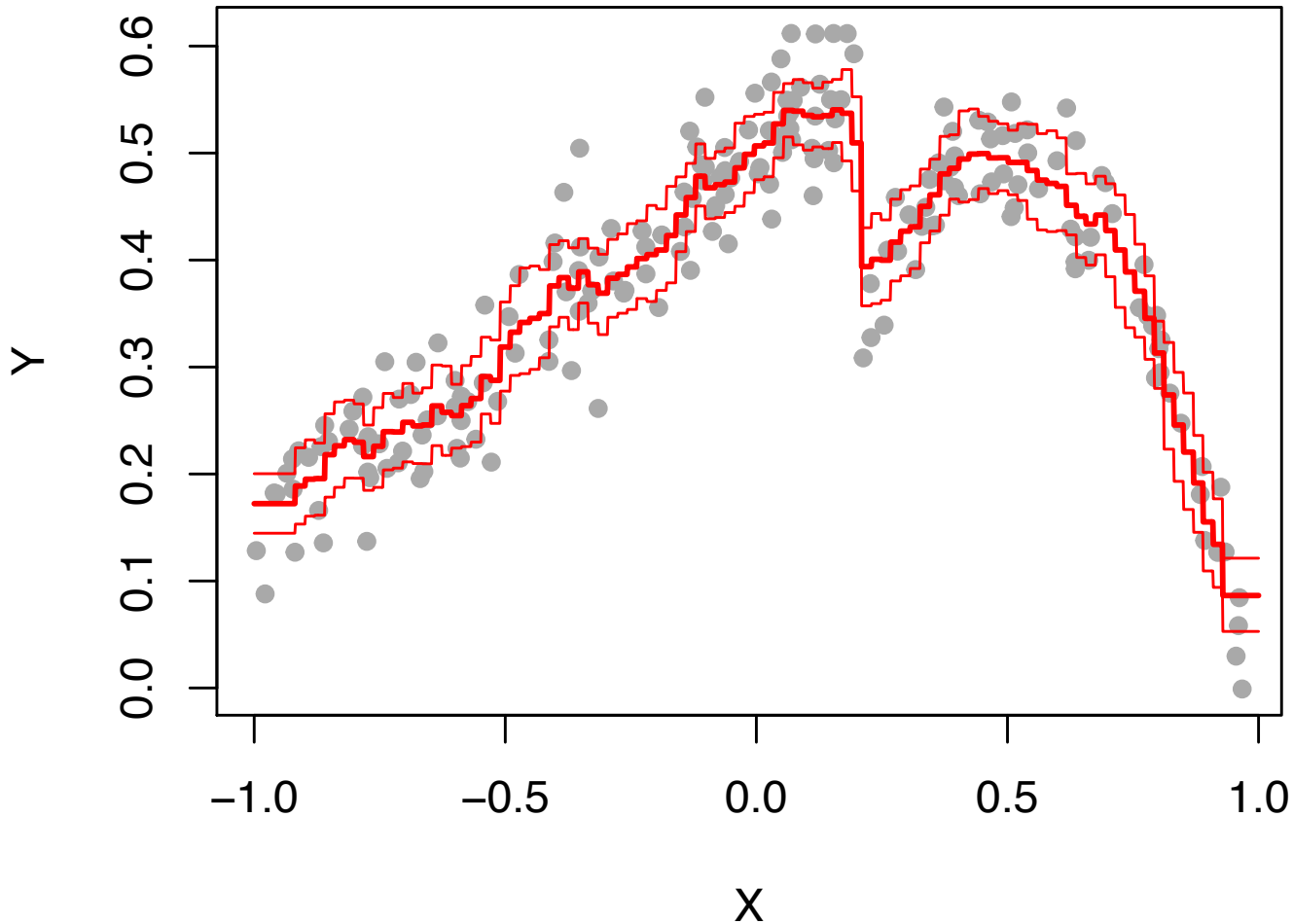
Partition
$\mathcal{A}_h = \{\mathcal{A}_{h1}, \mathcal{A}_{h2}, \mathcal{A}_{h3}\}$

$$g(\mathbf{x}, T_h, M_h) = \mu_{ht} \text{ if } \mathbf{x} \in \mathcal{A}_{ht} \text{ (for } 1 \leq t \leq b_h).$$

Example BART fit



Example BART fit



Add in a treatment variable and go?

Add treatment indicator as input variable (Hill, 2011):

$$y_i = f(\mathbf{x}_i, z_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$f(\mathbf{x}, z) = \sum_{h=1}^{m} g(\mathbf{x}, z, T_h, M_h)$$

Add in a treatment variable and go?

Add treatment indicator as input variable (Hill, 2011):

$$y_i = f(\mathbf{x}_i, z_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$f(\mathbf{x}, z) = \sum_{h=1}^{m} g(\mathbf{x}, z, T_h, M_h)$$

Prior	Bias	Coverage	RMSE
BART	0.14	31%	0.15

Regularization-induced confounding (RIC), again!

What do we do? (borrow experience from regularized linear models)

Fix #1 (ps-BART): Add in propensity $\hat{\pi}(\mathbf{x}) = P(Z = 1 \mid \mathbf{x})$ as a covariate.

$$y_i = f(\mathbf{x}_i, z_i, \hat{\pi}(\mathbf{x}_i)) + \epsilon_i$$

Fix #2 (BCF): Reparameterize to directly control regularization on prognostic and treatment effect functions.

$$\begin{aligned} y_i &= f(\mathbf{x}_i, z_i, \hat{\pi}(\mathbf{x}_i)) + \epsilon_i \\ &= \mu(\mathbf{x}_i, \hat{\pi}(\mathbf{x}_i)) + \tau(\mathbf{x}_i) z_i + \epsilon_i \end{aligned}$$

Put independent BART priors on μ and τ and you're good to go!

Propensity score BART

Prior	Bias	Coverage	RMSE
BART	0.14	31%	0.15
Oracle BART	0.00	98%	0.05
ps-BART	0.06	85%	0.08

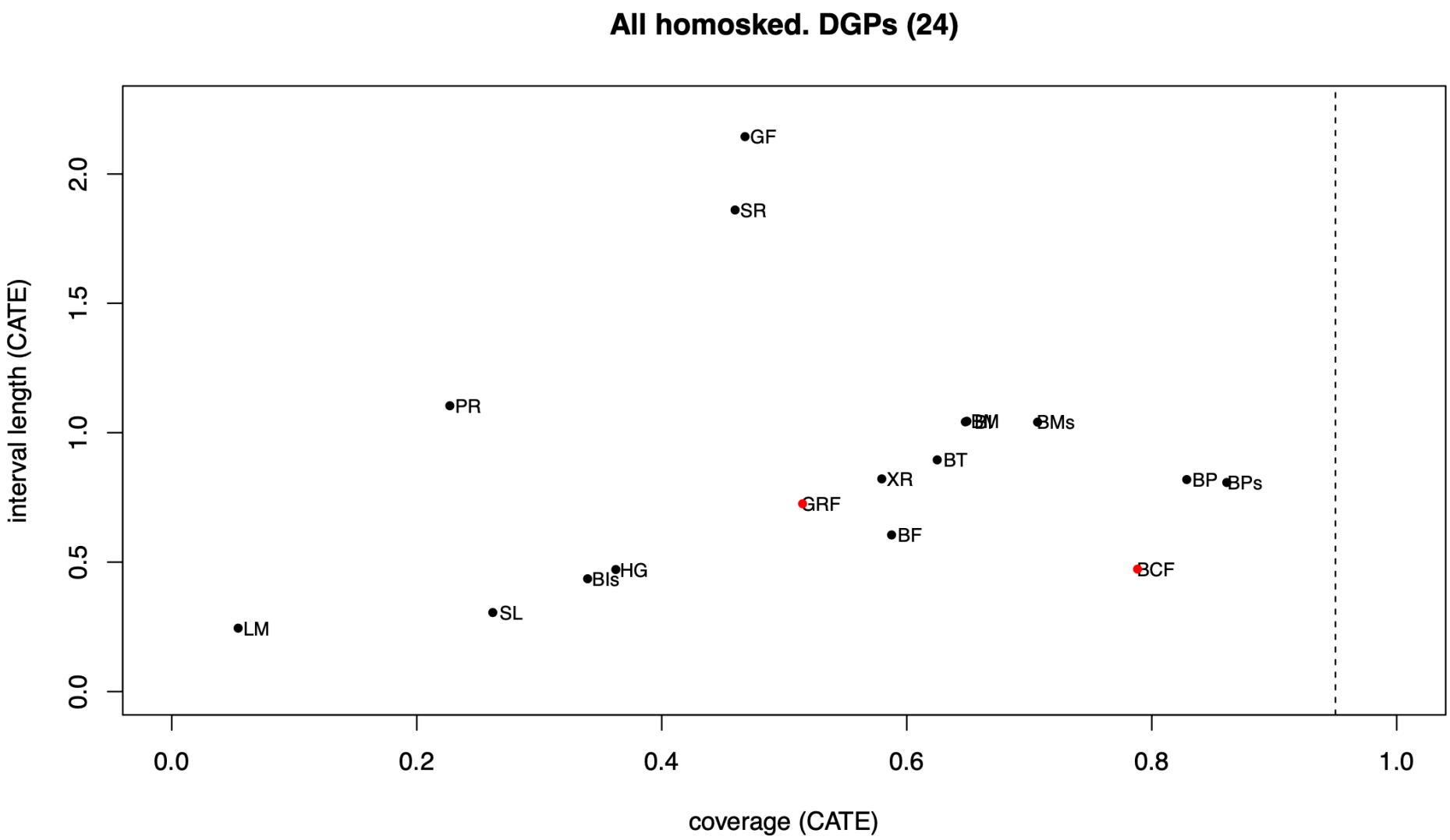
ACIC 2017 (data challenge for causal inference nerds)

with BCF and causal RF (Wager and Athey):

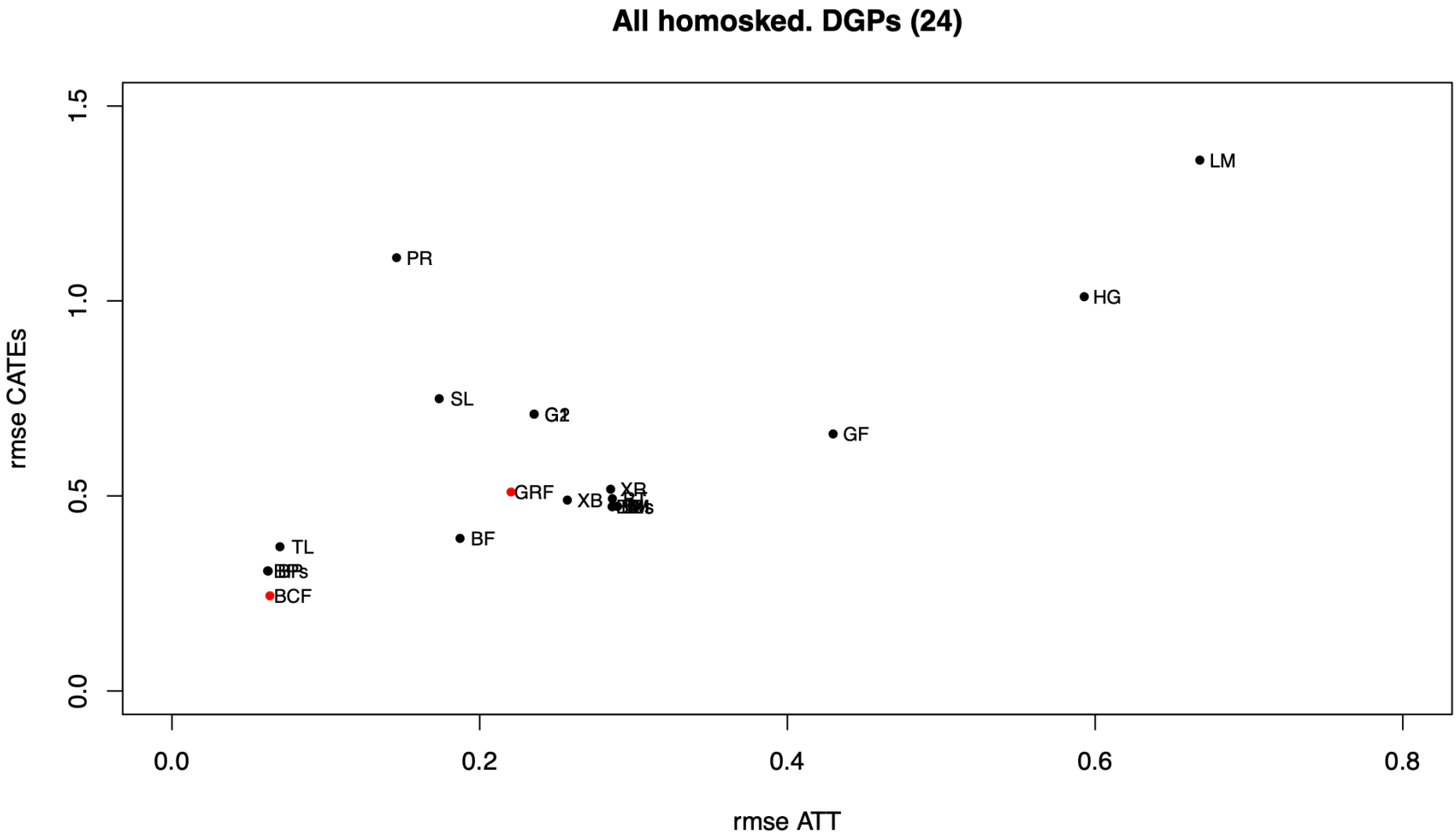
	Coverage	IL	Bias
BART	0.81	0.040	-0.0016
ps-BART	0.88	0.038	-0.0011
BCF	0.82	0.026	-0.0009
Causal RF	0.58	0.055	-0.0155

note that Wager and Athey's method also suffers from **RIC!**

ACIC 2017 (data challenge for causal inference nerds)



ACIC 2017 (data challenge for causal inference nerds)



Wrap up

Adjusting for confounding is fundamentally different than estimating the best predictive model.

Regularization helps "structure" a complex model, but it has to be deployed in the right way for causal inference (propensity score adjustment, reparameterizations, ...)

So much interesting work to do in this area, and it involves deeply understanding ML methods that were once only thought of as blackboxes.