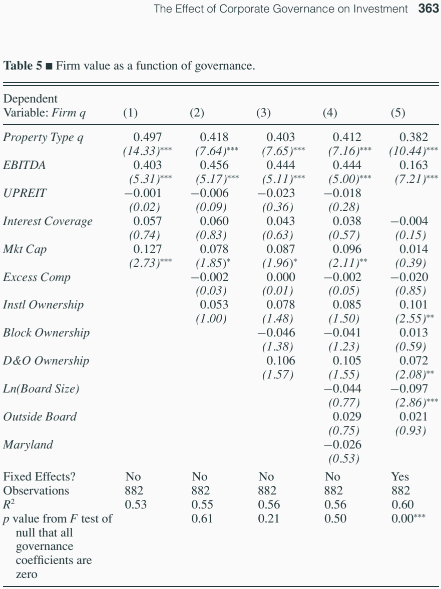
# Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects

**Carlos Carvalho** – The University of Texas at Austin.
Joint work with Jared Murray, P. Richard Hahn, David Yeager, et al.

April, 2019

# Policy Analytics (moving past the 1900's)



The Effect of Corporate Governance on Investment **363**

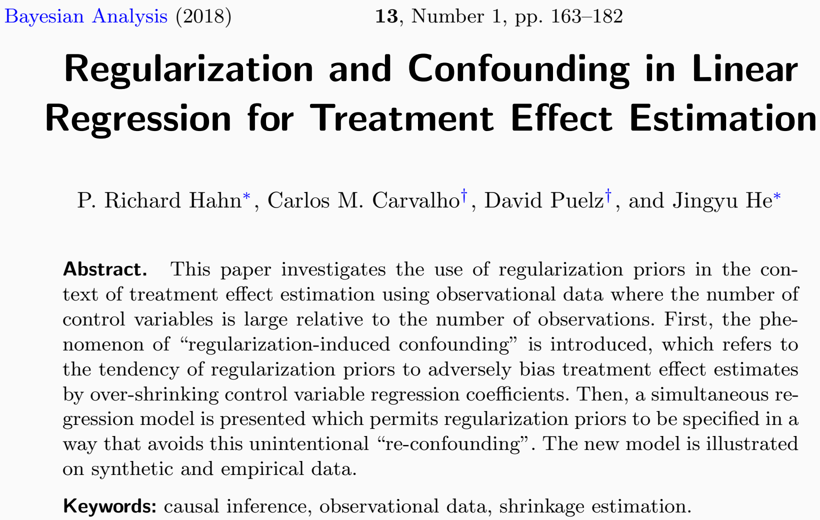
**Table 5** ■ Firm value as a function of governance.

| Dependent Variable: *Firm q* | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| *Property Type q* | 0.497 | 0.418 | 0.403 | 0.412 | 0.382 |
| | *(14.33)**** | *(7.64)**** | *(7.65)**** | *(7.16)**** | *(10.44)**** |
| *EBITDA* | 0.403 | 0.456 | 0.444 | 0.444 | 0.163 |
| | *(5.31)**** | *(5.17)**** | *(5.11)**** | *(5.00)**** | *(7.21)**** |
| *UPREIT* | −0.001 | −0.006 | −0.023 | −0.018 | |
| | *(0.02)* | *(0.09)* | *(0.36)* | *(0.28)* | |
| *Interest Coverage* | 0.057 | 0.060 | 0.043 | 0.038 | −0.004 |
| | *(0.74)* | *(0.83)* | *(0.63)* | *(0.57)* | *(0.15)* |
| *Mkt Cap* | 0.127 | 0.078 | 0.087 | 0.096 | 0.014 |
| | *(2.73)**** | *(1.85)** | *(1.96)** | *(2.11)*** | *(0.39)* |
| *Excess Comp* | | −0.002 | 0.000 | −0.002 | −0.020 |
| | | *(0.03)* | *(0.01)* | *(0.05)* | *(0.85)* |
| *Instl Ownership* | | 0.053 | 0.078 | 0.085 | 0.101 |
| | | *(1.00)* | *(1.48)* | *(1.50)* | *(2.55)*** |
| *Block Ownership* | | | −0.046 | −0.041 | 0.013 |
| | | | *(1.38)* | *(1.23)* | *(0.59)* |
| *D&O Ownership* | | | 0.106 | 0.105 | 0.072 |
| | | | *(1.57)* | *(1.55)* | *(2.08)*** |
| *Ln(Board Size)* | | | | −0.044 | −0.097 |
| | | | | *(0.77)* | *(2.86)**** |
| *Outside Board* | | | | 0.029 | 0.021 |
| | | | | *(0.75)* | *(0.93)* |
| *Maryland* | | | | −0.026 | |
| | | | | *(0.53)* | |
| Fixed Effects? | No | No | No | No | Yes |
| Observations | 882 | 882 | 882 | 882 | 882 |
| $R^2$ | 0.53 | 0.55 | 0.56 | 0.56 | 0.60 |
| *p* value from *F* test of null that all governance coefficients are zero | | 0.61 | 0.21 | 0.50 | 0.00*** |

# Moving from this...

$$y = \alpha Z + \mathbf{X}\beta + \epsilon$$

# Moving from this...



Bayesian Analysis (2018) **13**, Number 1, pp. 163–182

## Regularization and Confounding in Linear Regression for Treatment Effect Estimation

P. Richard Hahn[*], Carlos M. Carvalho[†], David Puelz[†], and Jingyu He[*]

**Abstract.** This paper investigates the use of regularization priors in the context of treatment effect estimation using observational data where the number of control variables is large relative to the number of observations. First, the phenomenon of "regularization-induced confounding" is introduced, which refers to the tendency of regularization priors to adversely bias treatment effect estimates by over-shrinking control variable regression coefficients. Then, a simultaneous regression model is presented which permits regularization priors to be specified in a way that avoids this unintentional "re-confounding". The new model is illustrated on synthetic and empirical data.

**Keywords:** causal inference, observational data, shrinkage estimation.

# ... to that!

$$y = f(Z, \mathbf{X}) + \epsilon$$

# ... to that!

$$y = f(Z, \mathbf{X}) + \epsilon$$

Today:

- a general, “default” framework
- a rich output that allows you to ask lots of different questions

# Our setting

We'll assume:

- **Observational and experimental data**
- **Conditional unconfoundedness/ignorability** (we've measured all the factors causally influencing treatment and response),
- **Covariate-dependent treatment effects** (individuals can have different responses to treatment according to their covariates)
- **Binary treatments**

# Our assumptions, more formally

*Strong ignorability*:

$$Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i \mid \mathbf{X}_i = \mathbf{x}_i,$$

*Positivity*:

$$0 < \Pr(Z_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) < 1$$

for all $i$. Therefore

$$\mathrm{E}(Y_i(z) \mid \mathbf{x}_i) = \mathrm{E}(Y_i \mid \mathbf{x}_i, Z_i = z),$$

so the conditional average treatment effect (CATE) is

$$\begin{aligned} \tau(\mathbf{x}_i) :&= \mathrm{E}(Y_i(1) - Y_i(0) \mid \mathbf{x}_i) \\ &= \mathrm{E}(Y_i \mid \mathbf{x}_i, Z_i = 1) - \mathrm{E}(Y_i \mid \mathbf{x}_i, Z_i = 0). \end{aligned}$$

# Modeling assumptions

We write

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

so that

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

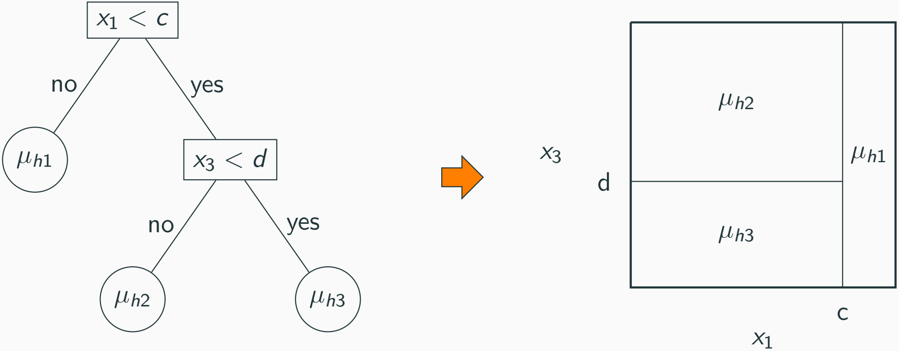
We assume iid Gaussian errors:

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

**nb:** Strong ignorability means $\epsilon_i \perp\!\!\!\perp Z_i \mid \mathbf{x}_i$.

How do we regularize estimates of $f$? (What prior on $f$?)

# Regression Trees



Tree $T_h$

$g(\mathbf{x}, T_h, M_h)$

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

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

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

# Bayesian Additive Regression Trees (BART)

Bayesian additive regression trees (Chipman, George, & McCulloch, 2008):

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

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

Hill (2011) proposes adopting Bayesian additive regression trees (BART) for causal inference.

# Making BART great for causal inference

BART is great as a prior over regression functions! It also works great for causal inference, but in some settings it can be problematic:

1. With strong confounding, estimates of individual/average treatment effects from BART can exhibit severe bias.
2. With homoegenous effects/moderate heterogeneity, BART's treatment effect estimates are highly variable.

**We can fix both of these!**

# Problem 1: Strong confouding can lead to high bias

Suppose that:

- $Y$: measure of heart distress,
- $Z$: took heart medication,
- $x_1$ and $x_2$ are blood pressure measurements.

Let's make this easy: $p = 2$, $n = 1,000$, with homogeneous treatment effects ($\tau = 1$).

# Problem 1: Strong confouding can lead to high bias

Assume the true model:

$$
\begin{aligned}
Y_i &= \mu(\mathbf{x}_i) - Z_i + \epsilon_i, \\
\mu(\mathbf{x}_i) &= 1 \text{ if } x_{i1} < x_{i2},\ -1 \text{ otherwise.} \\
\Pr(Z_i = 1 \mid x_{i1}, x_{i2}) &= \Phi(\mu(\mathbf{x}_i)), \\
\epsilon_i &\overset{iid}{\sim} \mathrm{N}(0, 0.7^2),\ \ x_{i1}, x_{i2} \overset{iid}{\sim} \mathrm{N}(0, 1).
\end{aligned}
$$

**This example demonstrates targeted selection into treatment**: Patients with $x_{i1} < x_{i2}$ are 5 times as likely to receive the new drug precisely because they are more likely to have higher levels of heart distress.

Despite low noise, low dimension, and homogeneous effects, BART has problems...

# BART is badly biased here

Across 250 simulated datasets with $n = 1000$, BART is badly biased:

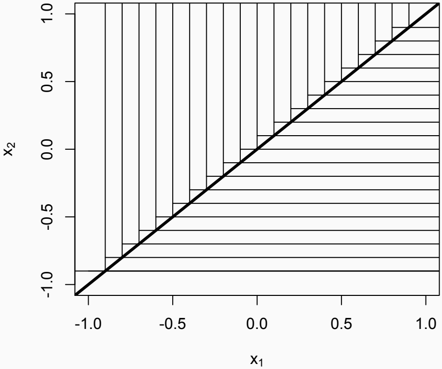
| Prior | Bias | Coverage | RMSE |
|---|---|---|---|
| BART | 0.14 | 31% | 0.15 |

This is due to a pheonomenon called **regularization induced confounding** (see Hahn, Carvalho, Puelz and He, 2018).

# Targeted selection induces regularization induced confounding

Why is BART biased in this example?

- $\pi(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$ is a noisy function of $\mu(\mathbf{x})$, so $\mu(\mathbf{x})$ "looks like" $\pi(\mathbf{x})$, and $\mu(\mathbf{x})$ is hard to approximate with trees



# Targeted selection induces regularization induced confounding

Why is BART biased in this example?

- $\pi(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$ is a noisy function of $\mu(\mathbf{x})$, so $\mu(\mathbf{x})$ “looks like” $\pi(\mathbf{x})$, and $\mu(\mathbf{x})$ is hard to approximate with trees
- Strong confounding means $Z \approx \pi(\mathbf{x})$, and targeted selection means $\mu(\mathbf{x})$ is a function of $\pi(\mathbf{x})$, so $\mu(\mathbf{x})$ is can be approximated by a tree that splits on $Z$
- The BART prior over $f$ penalizes the total number of splits, so to fit $\mu(\mathbf{x})$ BART would rather split on $Z$ once than $x_1$ and $x_2$ many times – confusing confounding for treatment effects: **regularization induced confounding** (Hahn et al (2016))

# A fix: Propensity Score BART

We can fix this by estimating $\pi(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$ (here using BART) and including $\hat{\pi}(\mathbf{x})$ as an extra predictor variable

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

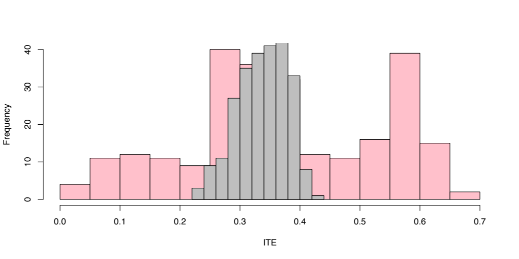
(With an ensemble estimate of $\hat{\pi}$, ps-BART$\approx$Oracle BART)

# Problem 2: Naive priors give high variance estimates

In the model

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

a BART prior on $f$ provides no direct mechanism to regularize the treatment effect function $\tau(\mathbf{x})$



ps-BART is in pink; our fix is in grey

# The fix: Bayesian causal forests

Reparameterize!

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

where $m$ and $\tau$ have independent BART priors.

Now the treatment effect is

$$\tau(\mathbf{x}_i)$$

and we can "shrink towards homogeneity" with stronger regularization on $\tau$, independent of regularization on $m$

# Tweaking priors in BCF

Several adjustments to the BART prior on $\tau$:

- Higher probability on smaller $\tau$ trees (than BART defaults)
- Higher probability on "stumps" (all stumps $=$ homogeneous effects)
- $N^{+}(0, v)$ Hyperprior on the scale of leaf parameters in $\tau$

# RIC in the wild: 2017 ACIC Data Analysis Challenge

Treatment-response pairs were simulated according to 32 distinct data generating processes (DGPs), given fixed covariates ($n = 4,302$, $p = 58$) from an empirical study.

We varied three parameters among two levels

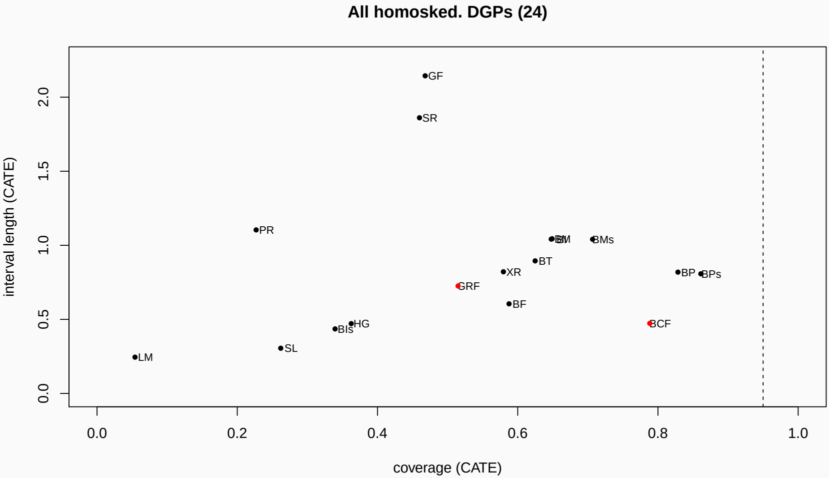
- **High** or **Low** *noise level*,
- **Strong** or **Weak** *confounding*,
- **Small** or **Large** *effect size*.

The error distributions were one of four types

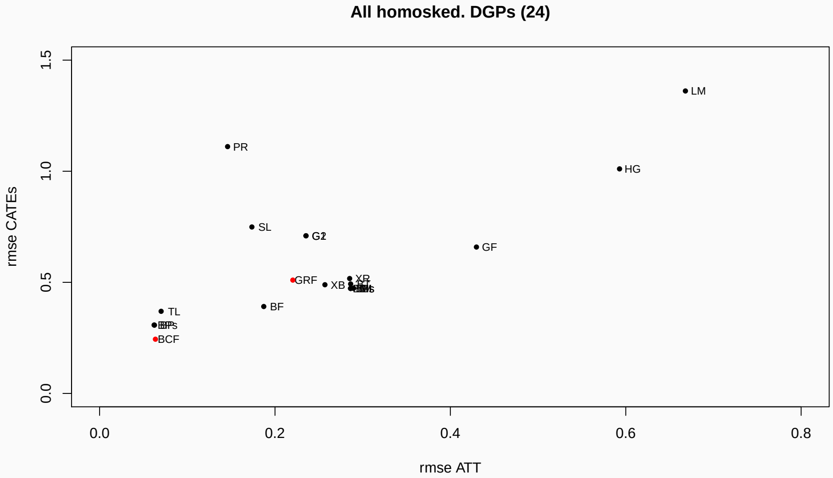
- Additive, homoskedastic, independent,
- Nonadditive, homoskedastic, independent,
- Additive, heteroskedastic, independent.

To assess coverage, 250 replicate data sets were generated for each DGP.

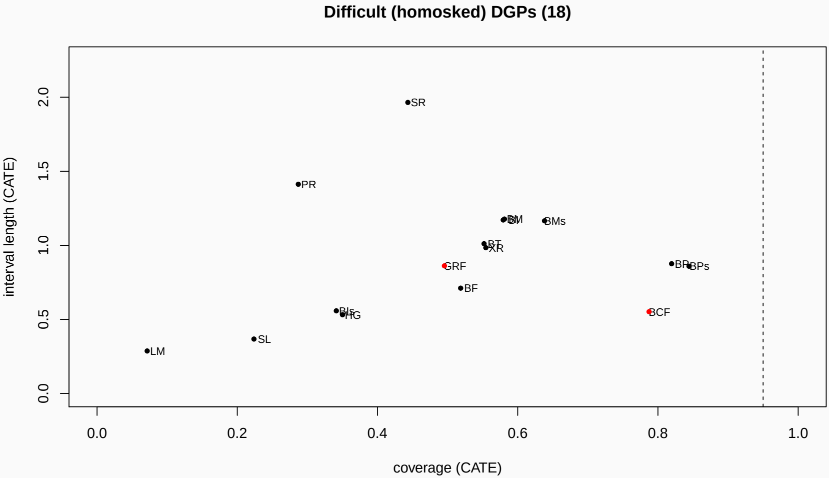
# Results: Inference for CATE on homoskedastic DGPs



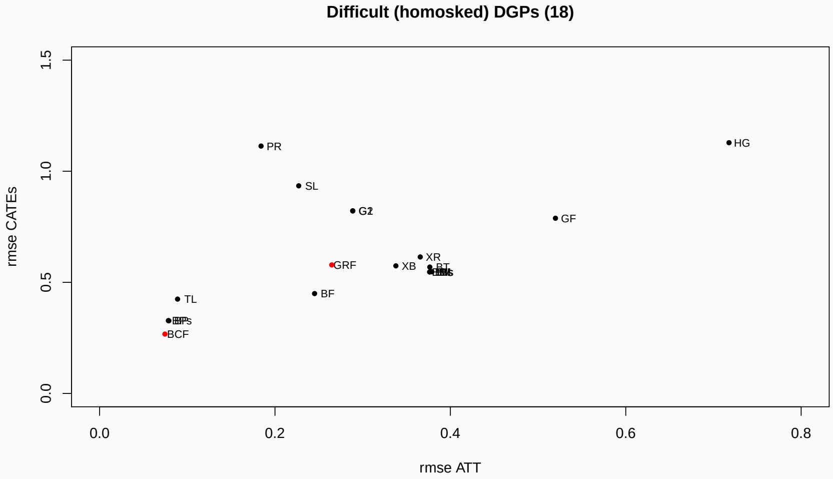
# Results: Estimation for homoskedastic DGPs



# Results: Inference for CATE on difficult homoskedastic DGPs

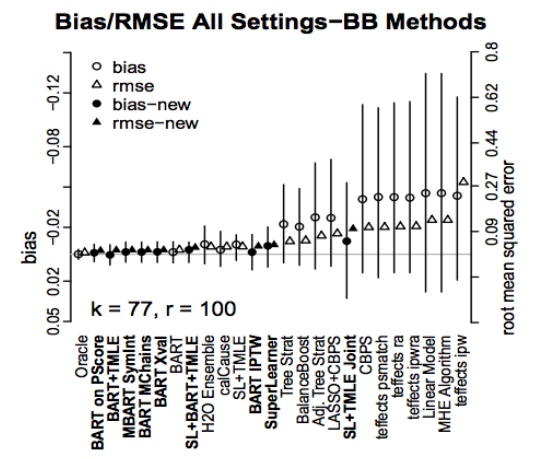


# Results: Estimation on difficult homoskedastic DGPs



# ACIC 2016 Redux

After our a preprint of our paper the ACIC 2016 challenge organizers ran ps-BART...



Large, highly variable treatment effects and no explicit targeted selection!

# ACIC 2016 Redux

Adding BCF and causal RF:

| | Cov | IL | Bias | (SD) | \|Bias\| | (SD) | PEHE | (SD) |
|---:|---:|---:|---:|---:|---:|---:|---:|---:|
| BCF | 0.82 | 0.026 | -0.0009 | (0.01) | 0.008 | 0.010 | 0.33 | 0.18 |
| ps-BART | 0.88 | 0.038 | -0.0011 | (0.01) | 0.010 | 0.011 | 0.34 | 0.16 |
| BART | 0.81 | 0.040 | -0.0016 | (0.02) | 0.012 | 0.013 | 0.36 | 0.19 |
| Causal RF | 0.58 | 0.055 | -0.0155 | (0.04) | 0.029 | 0.027 | 0.45 | 0.21 |

Average differences relative to BCF, pairwise permutation test p-value:

| | Diff Bias | p | Diff \|Bias\| | p | Diff PEHE | p |
|---:|---:|---:|---:|---:|---:|---:|
| ps-BART | -0.00020 | 0.146 | 0.0011 | $< 1e^{-6}$ | 0.010 | $< 1e^{-6}$ |
| BART | -0.00070 | $< 1e^{-6}$ | 0.0031 | $< 1e^{-6}$ | 0.037 | $< 1e^{-6}$ |
| Causal RF | -0.01453 | $< 1e^{-6}$ | 0.0204 | $< 1e^{-6}$ | 0.125 | $< 1e^{-6}$ |

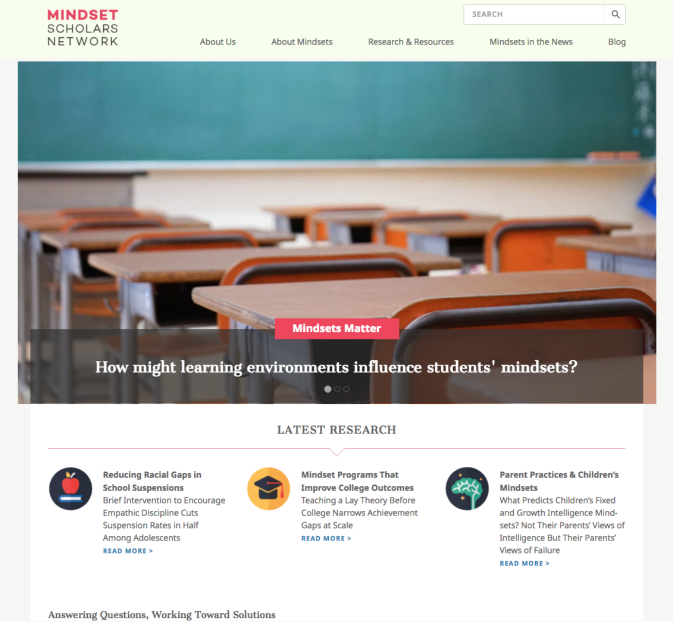
## Takeaways

In observational data regularization-induced confounding can adversely bias treatment effect estimates **from any method that uses regularization**. Explicitly modeling selection is necessary for robust inference.

BART is an impressive response surface method for causal inference; our new BCF models improve on “vanilla” BART in key respects:

- Propensity score estimates as covariates mitigate RIC
- Reparameterization allows regularization to be imposed robustly and directly on the estimand of interest.
- **It also facilitates extensions to multilevel models!**

# Multilevel BCF: The National Study of Learning Mindsets



How might learning environments influence students' mindsets?

LATEST RESEARCH

**Reducing Racial Gaps in School Suspensions**
Brief Intervention to Encourage Empathic Discipline Cuts Suspension Rates in Half Among Adolescents
READ MORE >

**Mindset Programs That Improve College Outcomes**
Teaching a Lay Theory Before College Narrows Achievement Gaps at Scale
READ MORE >

**Parent Practices & Children's Mindsets**
What Predicts Children's Fixed and Growth Intelligence Mindsets? Not Their Parents' Views of Intelligence But Their Parents' Views of Failure
READ MORE >

Answering Questions, Working Toward Solutions

# National Study of Learning Mindsets

- National Study of Learning Mindsets (Yeager et. al., 2017): Randomized controlled trial of a low-cost mindset intervention
- Probability sample of 76 schools ($\approx 14,000$ students)
- Specifically designed to assess treatment effect heterogeneity
- Preregistration plan included specific subgroups of interest, **and a blinded exploratory analysis of heterogeneity**

# National Study of Learning Mindsets

Desiderata for our analysis:

- Avoid model selection/specification search
- School-level effect heterogeneity explained and unexplained by covariates
- Interpretable model summaries for communicating results

# Multilevel Linear Models for Heterogeneous Treatment Effects

School-specific intercepts/fixed/random effects

School-specific “unexplained” heterogeneity

$$y_{ij} = \alpha_j + \sum_{h=1}^{p} \beta_h x_{ijh} + \left[ \sum_{\ell=1}^{k} \tau_\ell w_{ij\ell} + \gamma_j \right] z_{ij} + \epsilon_{ij}$$

Controls at the student and/or school level

Moderators at the student and/or school level

# Coloring outside the lines:
# Multilevel Bayesian Causal Forests

We replace linear terms with Bayesian additive regression trees (BART)

$$y_{ij} = \alpha_j + \beta(\mathbf{x}_{ij}) + \left[\tau(\mathbf{w}_{ij}) + \gamma_j\right] z_{ij} + \epsilon_{ij}$$

# Coloring outside the lines: Multilevel Bayesian Causal Forests

We replace linear terms with Bayesian additive regression trees (BART)

BART in causal inferece: Hill (2011), Green & Kern (2012), …

Parameterizing treatment effect heterogeneity with BART is due to Hahn, Murray and Carvalho (2017)

$$y_{ij} = \alpha_j + \beta(\mathbf{x}_{ij}) + \left[\tau(\mathbf{w}_{ij}) + \gamma_j\right] z_{ij} + \epsilon_{ij}$$

# Coloring outside the lines:
# Multilevel Bayesian Causal Forests

We replace linear terms with Bayesian additive regression trees (BART)

BART in causal inferece: Hill (2011), Green & Kern (2012), …

Parameterizing treatment effect heterogeneity with BART is due to Hahn, Murray and Carvalho (2017)

$$y_{ij} = \alpha_j + \beta(\mathbf{x}_{ij}) + \left[\tau(\mathbf{w}_{ij}) + \gamma_j\right] z_{ij} + \epsilon_{ij}$$

Allows for complicated functional forms (nonlinearity, interactions, etc) without pre-specification…

…while carefully regularizing estimates with prior distributions (shrinkage toward additive structure and discouraging implausibly large treatment effects)

# Analyzing data with ML BCF

- Obtain posterior samples for all the parameters, compute treatment effect estimates for each unit/school/etc.

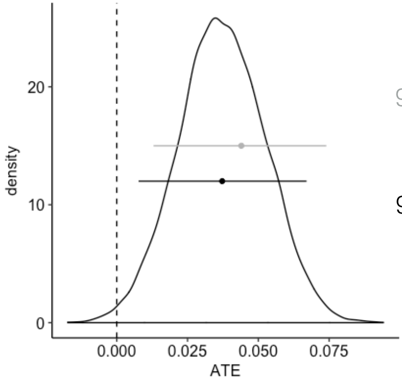
- The challenge: How do we summarize these complicated objects?

  - “Roll up” treatment effect estimates to ATE

  - Subgroup search

  - Counterfactual treatment effect predictions/“partial effects of moderators”

# Inference for the Average Treatment Effect

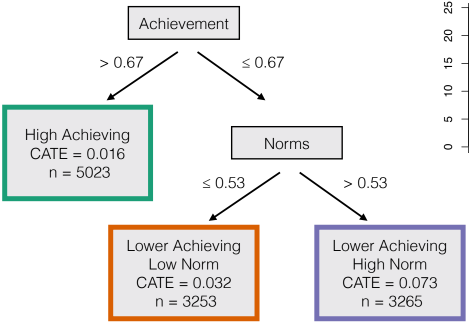


95% confidence interval
from ML Linear Model

95% uncertainty interval
from ML BCF

# Subgroup search

- Obtain posterior mean of treatment effects
- Use recursive partitioning (CART) **on the posterior mean** to find moderator-determined subgroups with high variation across subgroup ATE
    - Statistically kosher! We use the data once (prior -> posterior)
    - Can be formalized as the Bayes estimate under a particular loss function



Achievement

> 0.67

≤ 0.67

High Achieving
CATE = 0.016
n = 5023

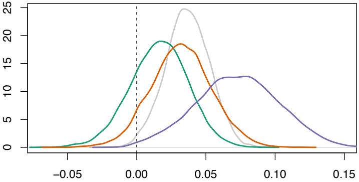
Norms

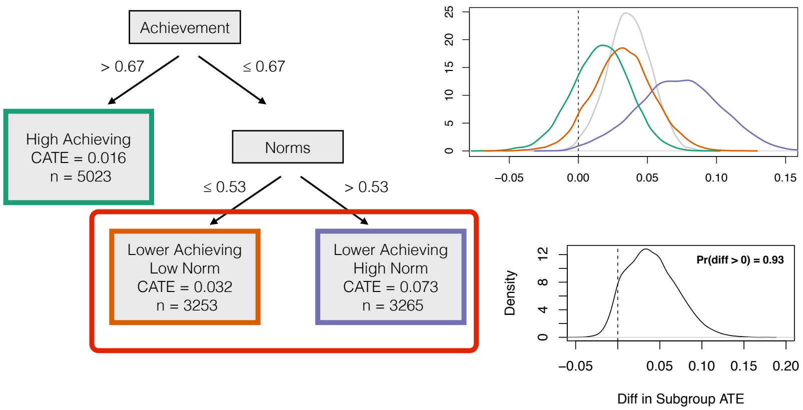
≤ 0.53

> 0.53

Lower Achieving
Low Norm
CATE = 0.032
n = 3253

Lower Achieving
High Norm
CATE = 0.073
n = 3265





Achievement

> 0.67

≤ 0.67

High Achieving
CATE = 0.016
n = 5023

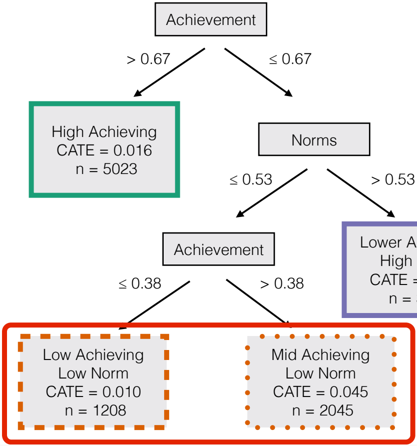
Norms

≤ 0.53

> 0.53

Lower Achieving
Low Norm
CATE = 0.032
n = 3253

Lower Achieving
High Norm
CATE = 0.073
n = 3265



Achievement

> 0.67

≤ 0.67

High Achieving
CATE = 0.016
n = 5023

Norms

≤ 0.53

> 0.53

Achievement

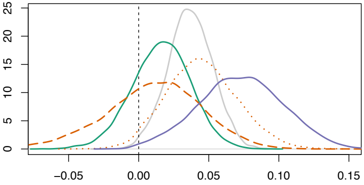
Lower Achieving
High Norm
CATE = 0.073
n = 3265

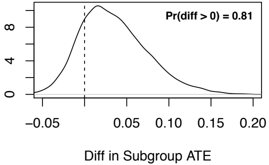
≤ 0.38

> 0.38

Low Achieving
Low Norm
CATE = 0.010
n = 1208

Mid Achieving
Low Norm
CATE = 0.045
n = 2045

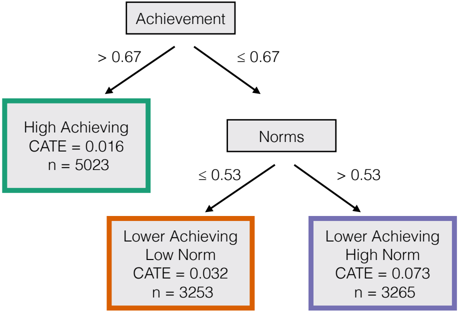




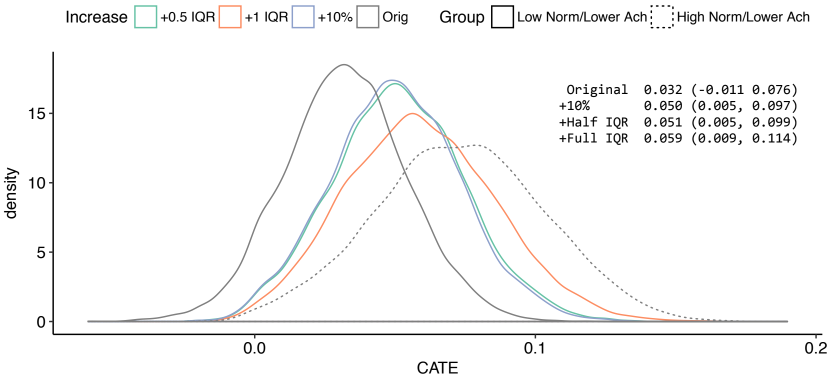
# Counterfactual treatment effect predictions

- How do estimated treatment effects change in lower achieving/low norm schools if norms increase, holding constant school minority comp & achievement?

- Not a formal causal mediation analysis (roughly, we would need “no unmeasured moderators correlated with norms”)



1 IQR = 0.6 extra problems
on worksheet task



# Conclusion

- Flexible models + careful regularization + posterior summarization is a winning combination
- Our approach takes the best parts of linear models with lots of researcher degrees of freedom and “black box” machine learning methods that only afford bankshot regularization and summarization
  - Many “degrees of freedom” in the summarization step, but these depend on the data only through the posterior
  - Unlike many ML methods, we can handle multilevel structure and prior knowledge with ease

Thank you!