Causal Inference at the Intersection of Statistics and Machine Learning

Jennifer Hill

presenting joint work with **Vincent Dorie** and

Nicole Carnegie (Montana State University), Dan Cervone (L.A. Dodgers), Masataka Harada (National Graduate Institute for Policy Studies, Tokyo), Marc Scott (NYU), Uri Shalit (Technion), Yu Sung Su (Tsinghua University)

March 8, 2018

Acknowledgement: This work was supported by IES Grants [ID: R305D110037 and R305B120017].

Most research questions are causal questions

What Drives Success?

Does exposing preschoolers to music make them smarter?

Can we alter genes to repel HIV?

Is obesity contagious?

Grief Can Cause a Heart Attack

Did the introduction of CitiBike make New Yorkers healthier?

Does the death penalty reduce crime?

What Happens When the Poor Receive a Stipend?

The New York Times

Causal Inference is Important

- Misunderstanding the evidence presented by data can lead to lost time, money and lives
- So why do we get it wrong so often?
- Consider some examples
  - Salk Vaccine
  - Internet ads and purchasing behavior
  - Hormone Replacement Therapy: Nurses Health Study versus Women's Health Initiative

*See great work by Hernan and Robins (2008) for subtleties in this one

Salk vaccine

- Observational studies did not support the effectiveness of the Salk vaccine at preventing Polio
- Randomized experiments showed it was effective
- Lives saved!

Lesson learned about thinking carefully about causality, right….?

Flash forward 50 years: Hubris

WIRED MAGAZINE: 16.07

Science : Discoveries

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

The Petabyte Age:

"There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show."

Cautionary Tale: Search Engine Marketing

- $31.7 billion spent in the U.S. in 2011 on internet advertising
- Common wisdom based on naive "data science":
  - internet advertising is highly effective
  - impact easy to measure because we can track info on those who click on ads (including "Did they buy or find site?")
- Prediction models suggest that clicking on the ads strongly increases probability of success (e.g., buying product/finding site)
- What if shoppers would have bought the product *anyway*?
- Results of quasi-experiments at eBay showed just that: 99.5% of click traffic was simply redirected through "natural" (unpaid) search traffic. i.e. almost everyone found the site anyway

From Blake, T., Nosko, C., and S. Tadelis (2013) "Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment"

Hormone Replacement Therapy

- Nurses Health Study (massive long-term observational study) shows benefits of HRT for coronary heart disease
- Women's Health Initiative (randomized experiment) shows the opposite results
- Hernan and Robins (*Epidemiology*, 2008) use thoughtful statistical analyses and careful causal thinking to reconcile results (a win for statistical causal inference!)

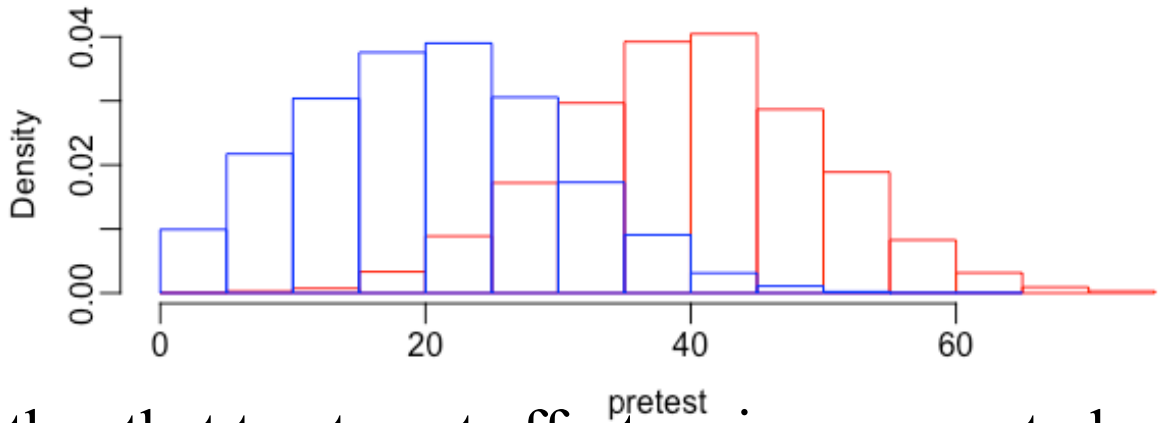
How can we make good choices when we don’t have a randomized experiment?

Quick review of Causal Inference

Consider a simple example

- Effect of an enrichment program on subsequent test scores
- Suppose that exposure to the program is
  - determined based on one pre-test score, and
  - is probabilistic, as in:

red for treated
blue for controls



- Suppose further that treatment effect varies across students as a function of pre-test scores (next slide).

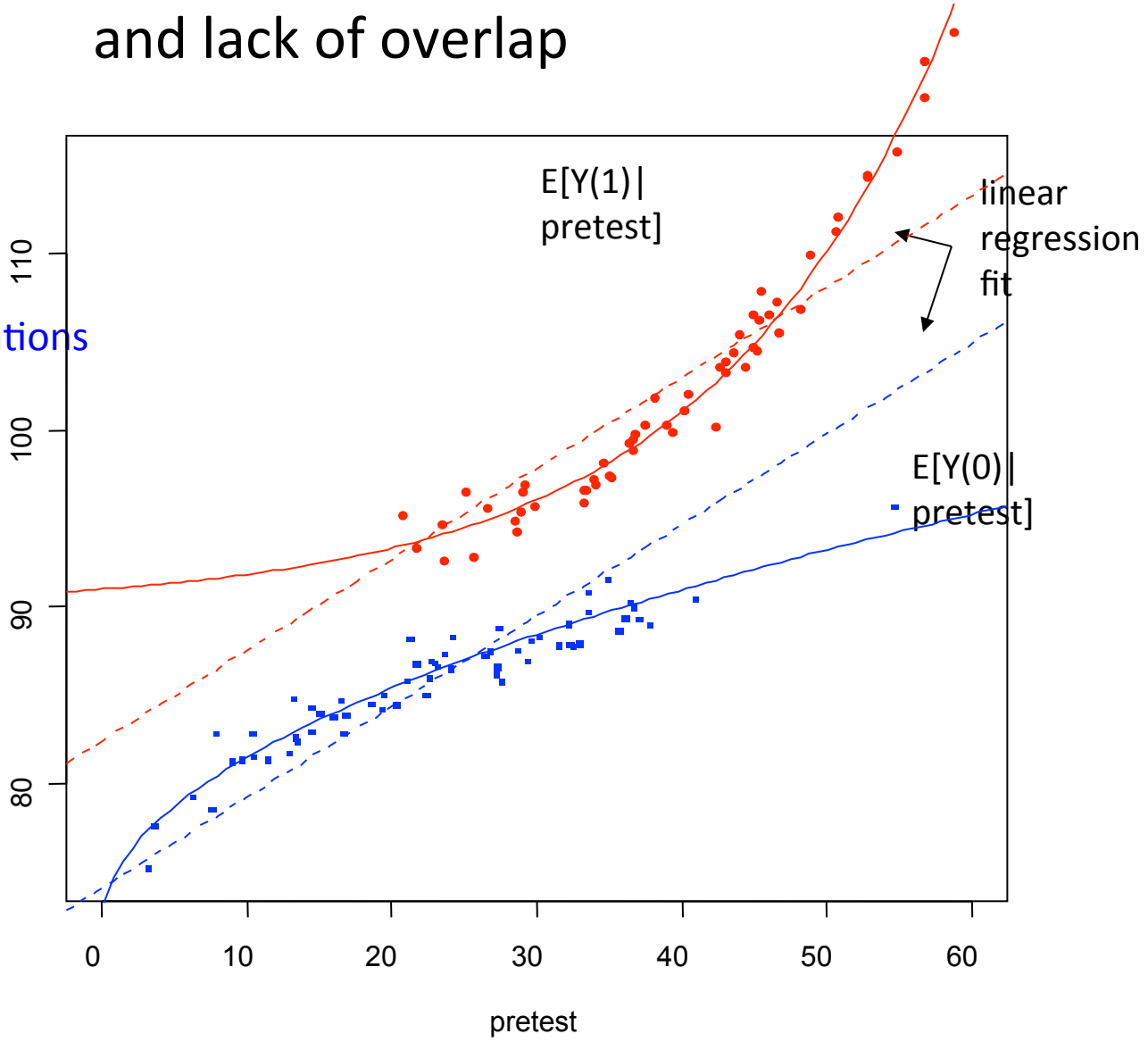
Parametric assumptions: implications of non-linearity and lack of overlap

red for treatment observations and response surface

blue for control observations and response surface

Linear regression (dotted lines) is not an appropriate model here

Lack of overlap in pretest scores exacerbates the problem by forcing model extrapolation



This is tricky even though we've assumed only one confounder!

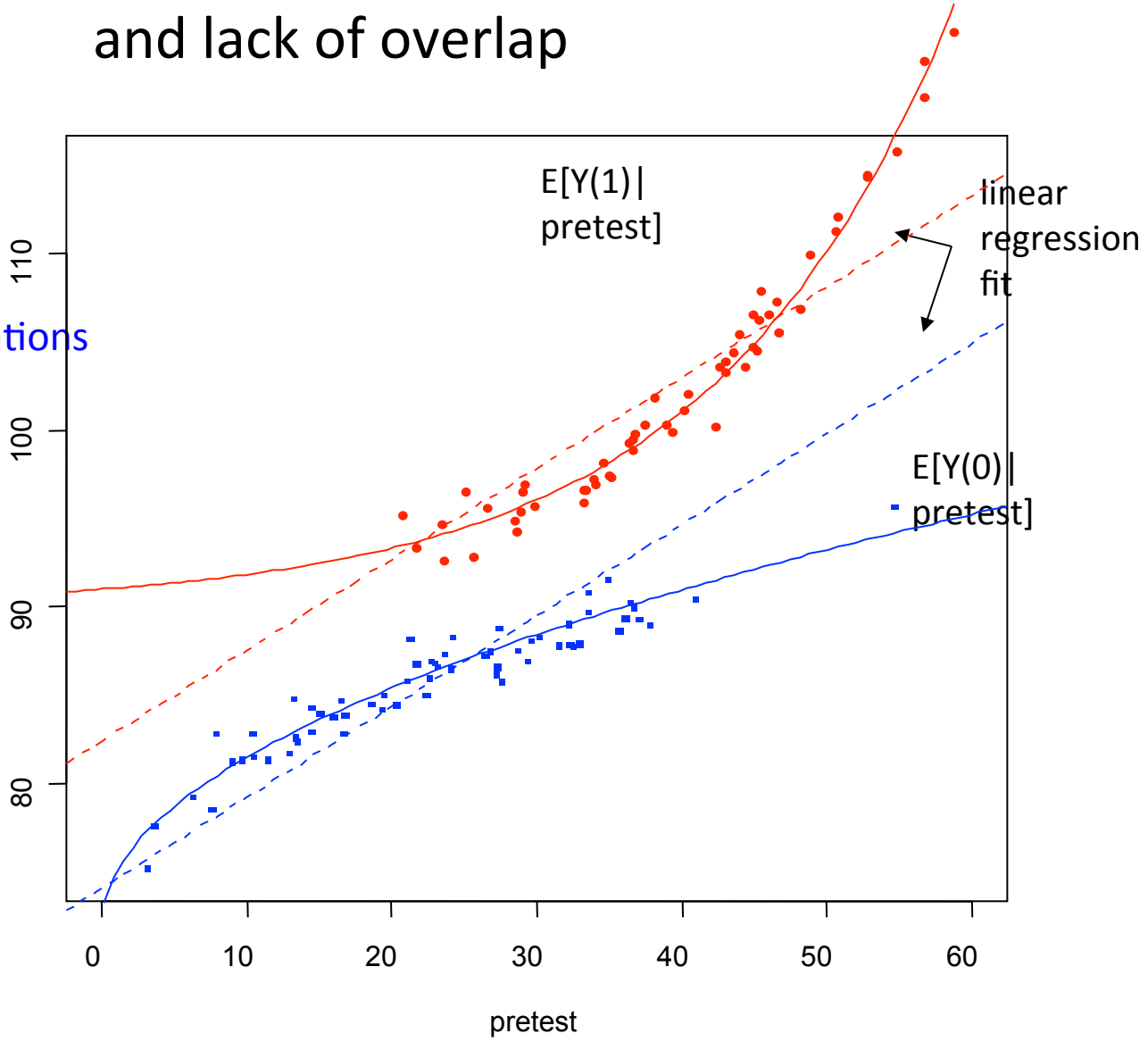
Parametric assumptions: implications of non-linearity and lack of overlap

red for treatment observations and response surface

blue for control observations and response surface

Linear regression (dotted lines) is not an appropriate model here

Lack of overlap in pretest scores exacerbates the problem by forcing model extrapolation



Causal inference is hard.

- For most interesting causal research questions we typically cannot perform experiments. Appropriate natural experiments are hard to find.
- Observational studies require strong assumptions
  - structural: all confounders measured
  - parametric: for the model used to adjust for all these confounders…

Causal inference is hard.

- For most interesting causal research questions we typically cannot perform experiments. Appropriate natural experiments are hard to find.
- Observational studies require strong assumptions
  - structural: all confounders measured (this was assumed in our simple example)
  - parametric: for the model used to adjust for all these confounders… (there was only 1 confounder in our simple example)

Notation/Estimands

Let

- X be a (vector of) observed covariates
- Z be a binary treatment variable
- $Y(0)$, $Y(1)$ are potential outcomes
- Y is the observed outcome

Individual level causal effects compare potential outcomes, e.g.

$$Y_i(1) - Y_i(0)$$

The goal is to estimate something like

$$E[Y(1) - Y(0)]$$

or

$$E[Y(1) - Y(0) \mid Z = 1]$$

Structural Assumptions

- The key assumption in most observational studies is that **all confounders** have been **measured** (ignorability, selection on observables, conditional independence, …) Formally this implies

$$Y(0), Y(1) \perp Z \mid X$$

This assumption is **untestable** and **difficult to satisfy**

- Stable Unit Treatment Value Assumption (no interference, consistency, etc)

Also untestable. Can design to help satisfy.

Parametric Assumptions

- We can tie our potential outcomes to X through a model. For instance if we assumed a linear model

$$E[Y(0) \mid X] = X\beta^y$$

$$E[Y(1) \mid X] = X\beta^y + \tau$$

- The more covariates we include (e.g. to satisfy "all confounders measured") the more we have to worry about parametric assumptions
- Most of the time we don't believe a linear model is appropriate
- The massive literature on propensity score matching is primarily aimed at reducing our reliance on these assumptions

Roadmap

- What role can Bayesian additive regression trees (BART) play in addressing issues in causal inference?
  - Parametric assumptions in causal inference
    - BART to fitting the response surface, e.g. E[Y | Z, X]
    - Use BART automatic uncertainty quantification to understand when don't have sufficient common support
    - Heterogeneity, generalizability
    - bartCause
  - Structural Assumptions
    - Sensitivity analysis to explore violations to the assumption that all confounders measured
    - treatSens
- Why BART? What about other machine learning approaches?

BART

(Chipman, George, and McCulloch, 2007, 2010)

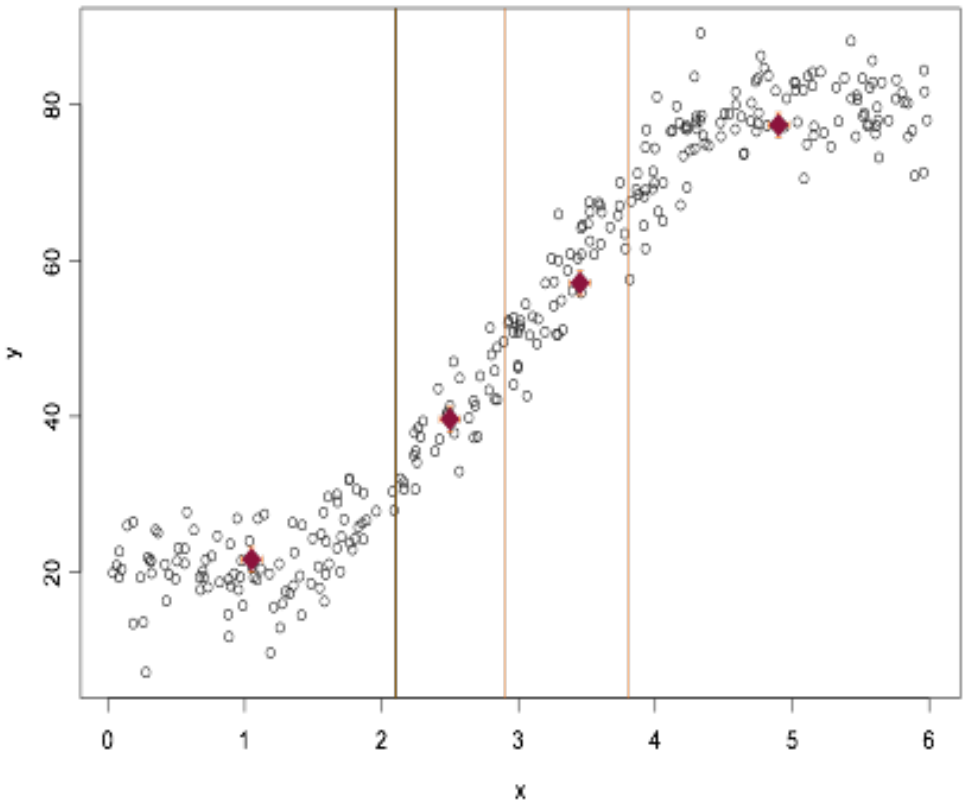
Understanding how BART works

BART: Bayesian Additive Regression Trees (BART, Chipman, George, and McCulloch, 2007, 2010) can be informally conceived of as a Bayesian form of boosted regression trees. So to understand better we'll first briefly discuss

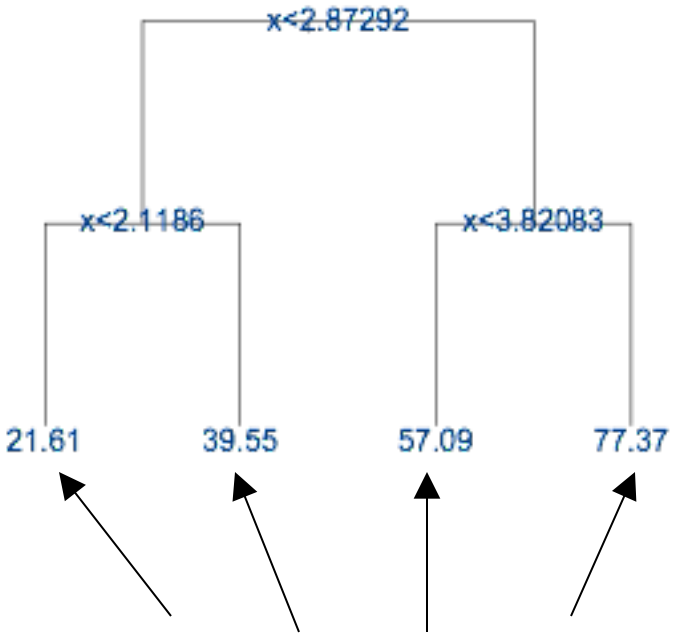
- Regression trees
- Boosted regression trees
- Bayesian inference/MCMC

Will find interactions, non-linearities. Not the best for additive models.

Regression trees



Progressively splits the data into more and more homogenous subsets. Within each of these subsets the mean of y can be calculated



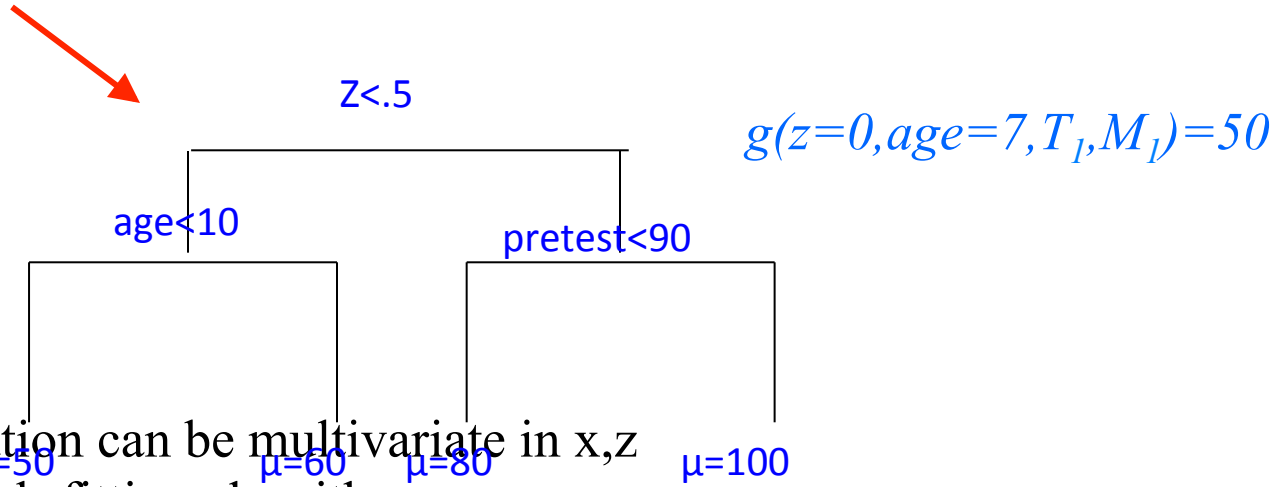
"terminal nodes"

Boosting of regression trees

Builds on the idea of a treed model to create a "sum-of-trees" model

- Each tree is small – a "weak learner"– but we may include many (e.g. 200) trees
- Let $\{T_j, M_j\}$ j=1,…,m, be a set of tree models
  T_j denotes the j^{th} tree,
  M_j denotes the means from the terminal nodes from the j^{th} tree,

$$f(z,x) = g(z,x,T_1,M_1) + g(z,x,T_2,M_2) + \ldots + g(z,x,T_m,M_m)$$



- Each contribution can be multivariate in x,z
- Fit using a back-fitting algorithm.

Boosting: Pros/Cons

- Boosting is great for prediction but …
  - Requires *ad-hoc choice of tuning parameters* (# trees, depths of trees, shrinkage for the fit of each tree)
  - *How estimate uncertainty*? Generally, people use bootstrapping which can be cumbersome and time-consuming

Bayesian Additive Regression Trees (CGM, 2007, 2011) (similar to boosting with important differences)

BART can be thought of loosely as a stochastic alternative to boosting algorithms for fitting a sum-of-trees model:

$$f(z,x) = g(z,x,T_1,M_1) + g(z,x,T_2,M_2) + \ldots + g(z,x,T_m M_m) \text{ and } \sigma^2$$

- It differs because:
  - $f(x,z)$ is a random variable sampled using MCMC
    - (1) $\sigma \mid \{T_j\},\{M_j\}$
    - (2) $T_j, M_j \mid \{T_i\}_{i\neq j},\{M_i\}_{i\neq j}, \sigma$
  - Trees are exchangeable
  - *Avoids overfitting by the prior specification* that shrinks towards a simple fit:
    - Priors tend towards small trees ("weak learners")
    - Fitted values from each tree are shrunk using priors

BART in R

- There are a few BART packages available, three of which are in R.
- We prefer `dbarts` (Dorie et al.)
  - drop in replacement for BayesTree
  - C++ with efficient data structures (fast!)
  - natively parallelized within and across chains
  - sampler state can be updated/can embed in larger model
  - parallel cross-validation
  - can use model fit to predict for another dataset

BART to address parametric assumptions in causal inference (Hill, JCGS, 2011)

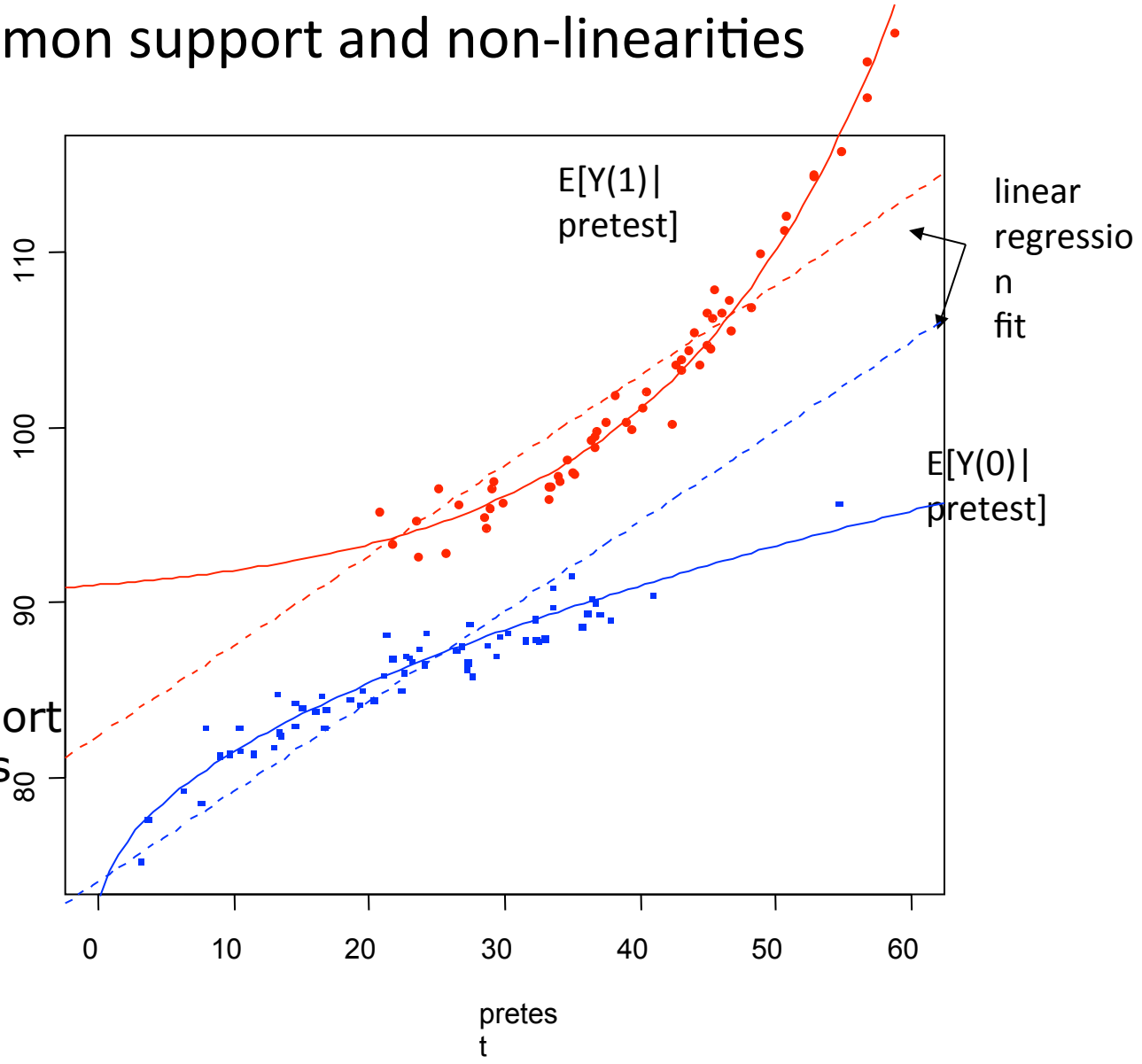
Parametric assumptions: implications of lack of common support and non-linearities

red for treatment observations and response surface

blue for control observations and response surface

Linearity is not an appropriate model here

Lack of common support in pretest exacerbates the problem by forcing model extrapolation

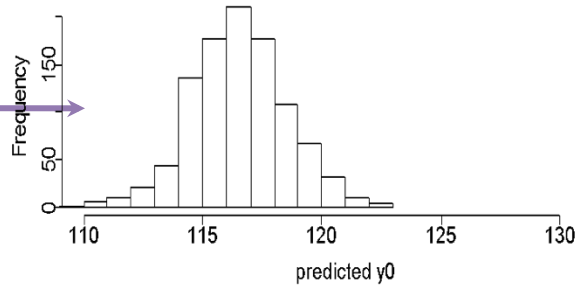


Distribution for the treatment effect created by differencing the (posterior predictive) distributions for Y(1) and Y(0)

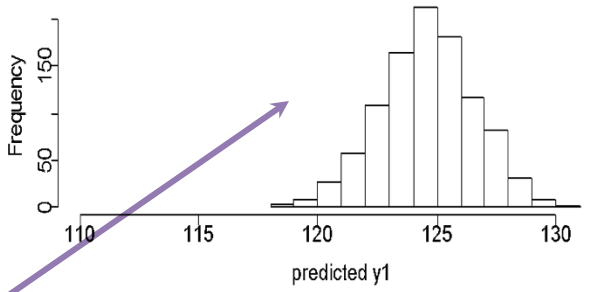
mom age	mom hs grad.	mom work	mom race	mom marital status	Z	child test score
19	1	1	B	1	0	114
22	0	1	W	0	0	92
27	0	1	B	0	0	80
23	1	0	H	1	0	98
20	1	0	H	0	1	110
25	1	1	W	1	1	82
24	0	1	B	0	1	102
⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	0	1	H	1	1	89

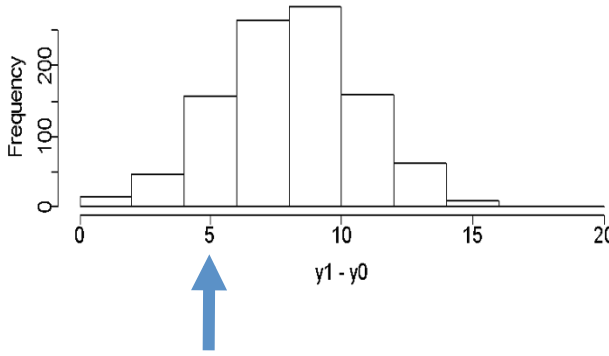
test score = f(age, hs, work, race, marital, Z) + error

mom age	mom hs grad.	mom work	mom race	mom marital status	Z	Predicted child test score
19	1	1	B	1	0	116.6
22	0	1	W	0	0	90.5
27	0	1	B	0	0	79.0
23	1	0	H	1	0	96.2
20	1	0	H	0	0	107.1
25	1	1	W	1	0	74.8
24	0	1	B	0	0	98.4
⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	0	1	H	1	0	83.2



mom age	mom hs grad.	mom work	mom race	mom marital status	Z	Predicted child test score
19	1	1	B	1	1	124.6
22	0	1	W	0	1	94.5
27	0	1	B	0	1	86.0
23	1	0	H	1	1	101.2
20	1	0	H	0	1	110.1
25	1	1	W	1	1	80.8
24	0	1	B	0	1	104.4
⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	0	1	H	1	1	88.2

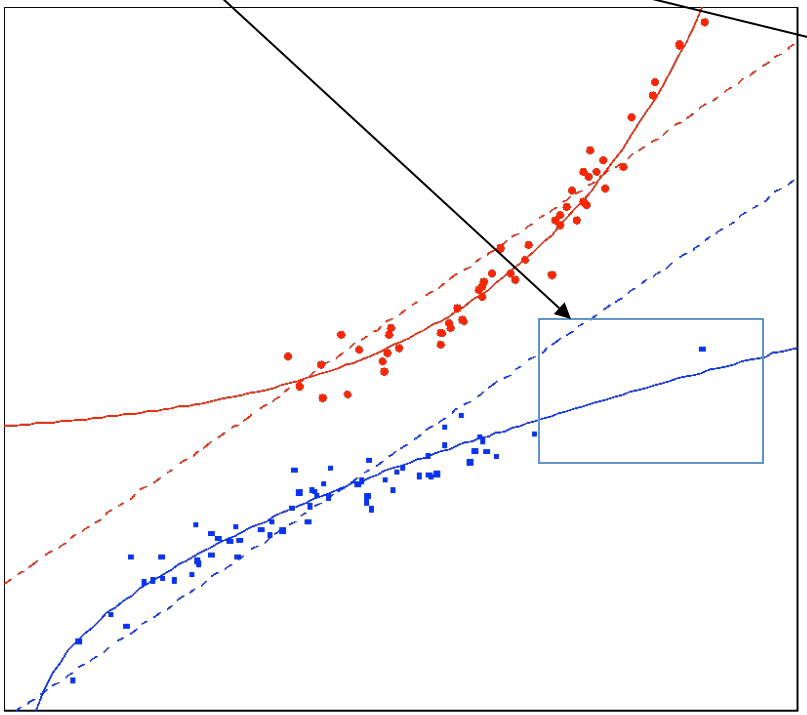




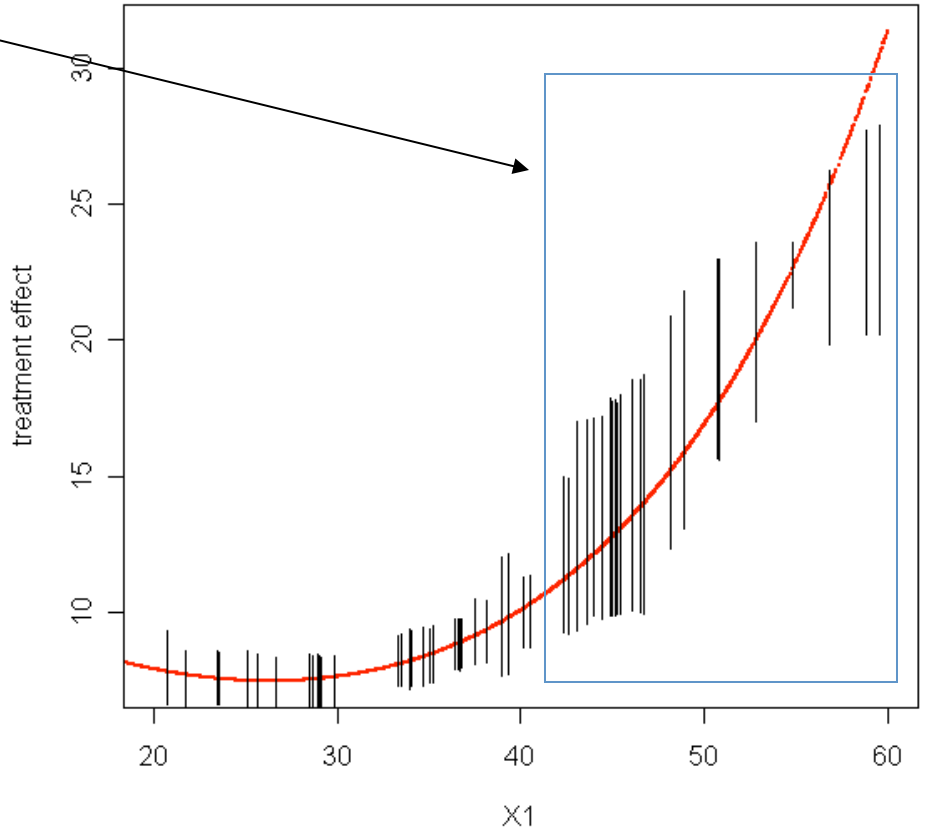
posterior dist. of individual level treatment effect estimate

BART performance when we lose common support – uncertainty reflected in posterior intervals!

Notice the point at which we lose empirical counterfactuals for the treatment group… this is where we see uncertainty increasing in BART estimates



Here lines show uncertainty intervals around BART point estimates for the treatment effect at values of X1 observed in the data



Evidence of relative performance in the absence of any discarding

- In previous studies BART has outperformed propensity score matching and IPTW using propensity scores (combined with regression adjustment)
  - Hill, Journal of Computational and Graphical Statistics, 2011
  - Hill, Multivariate Behavioral Research, 2012
- When no discarding is performed the advantage starts to fade as lack of overlap increases

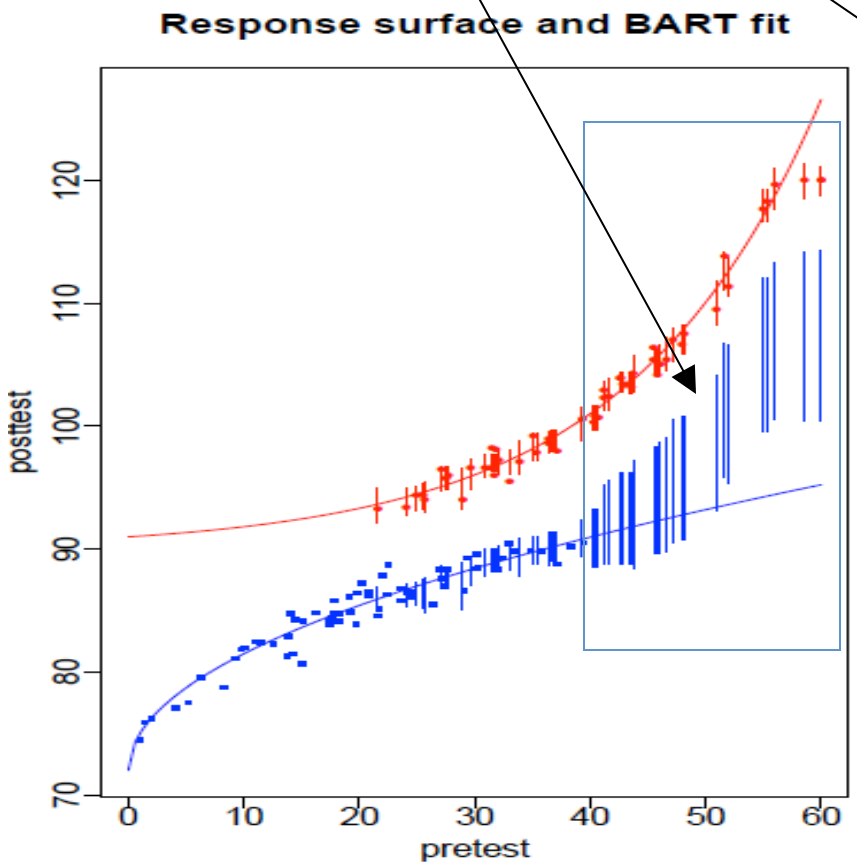
Overlap/Common Support

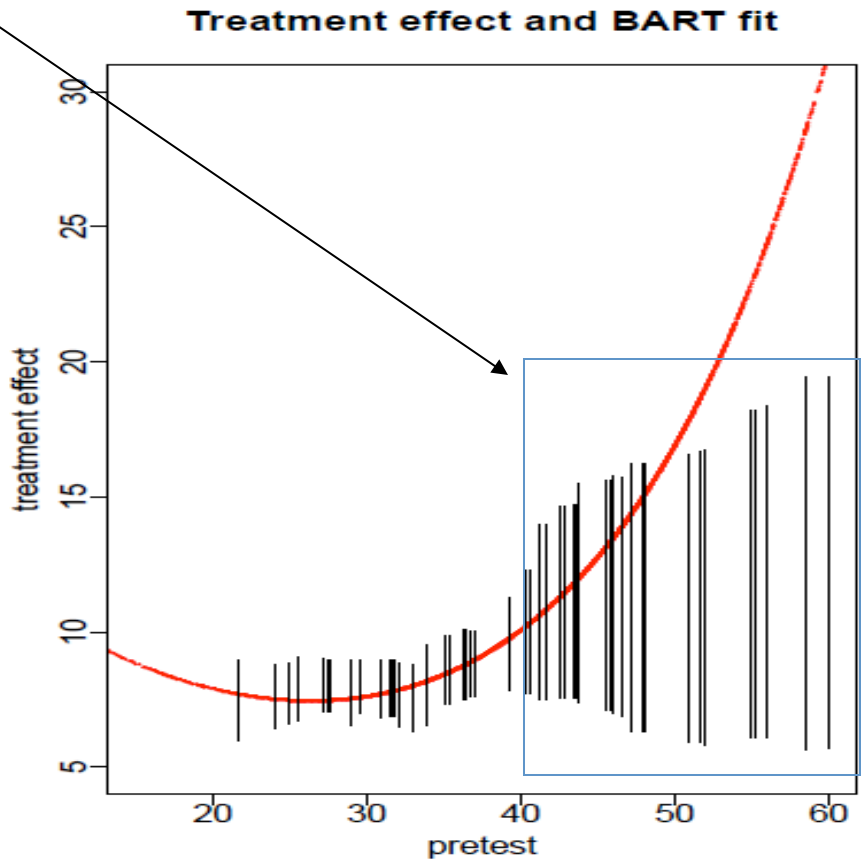
BART uncertainty increases when we lack common support

Note that propensity score matching would not attempt inference in the light blue square!

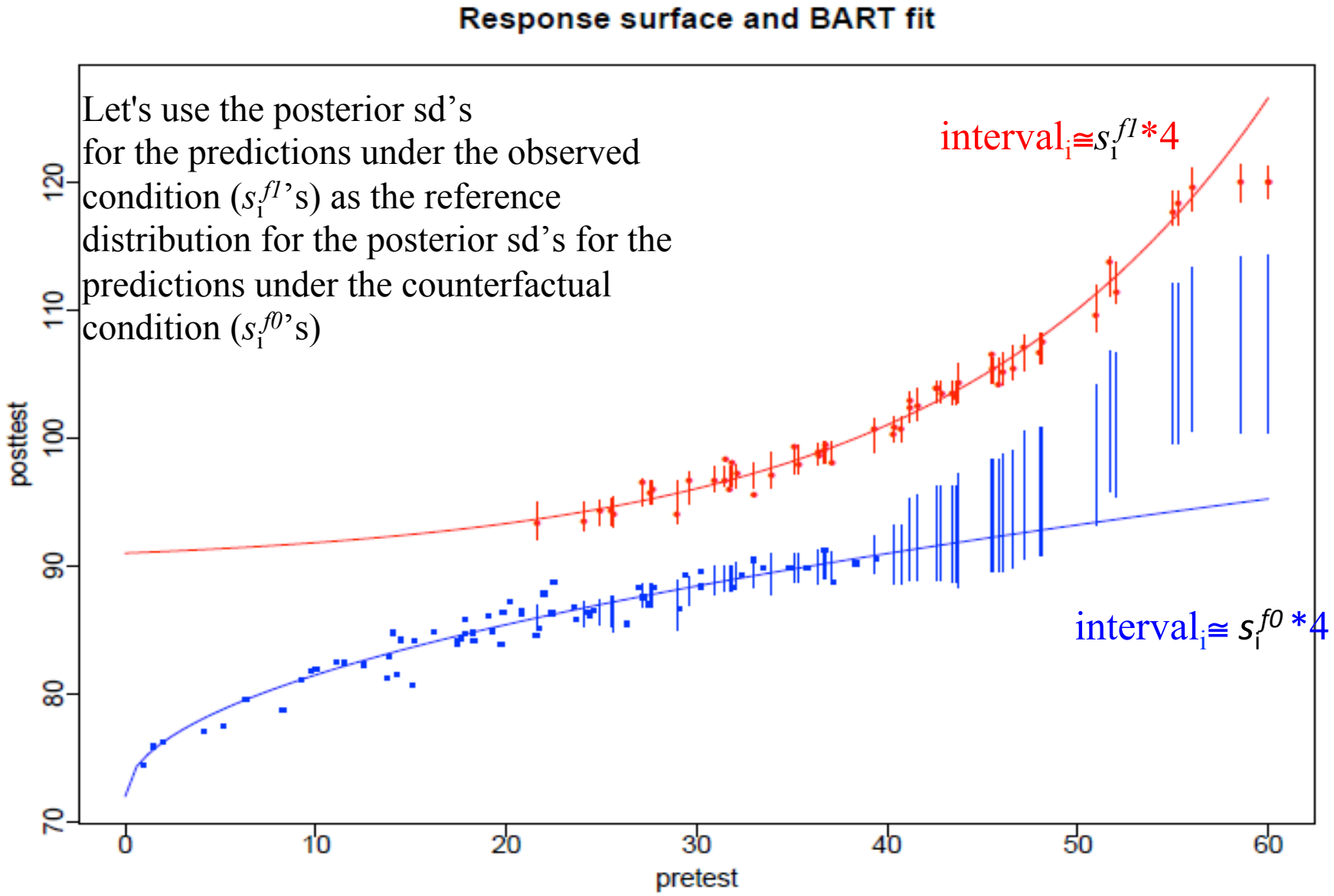
Notice the point at which we lose empirical counterfactuals for the treatment group…this is where we see uncertainty increasing in BART estimates

Here lines show uncertainty intervals around BART point estimates for the treatment effect at values of X1 observed in the data

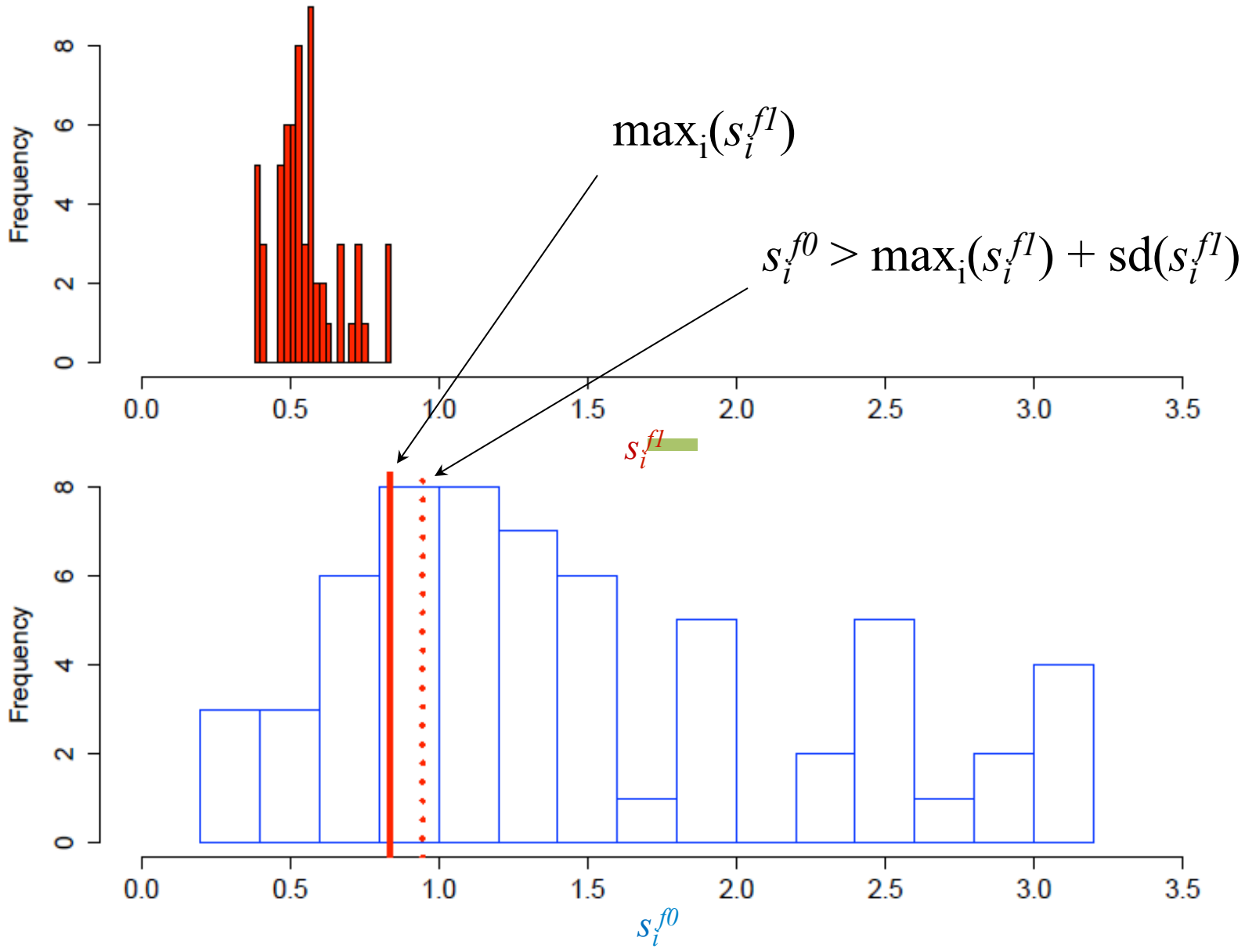




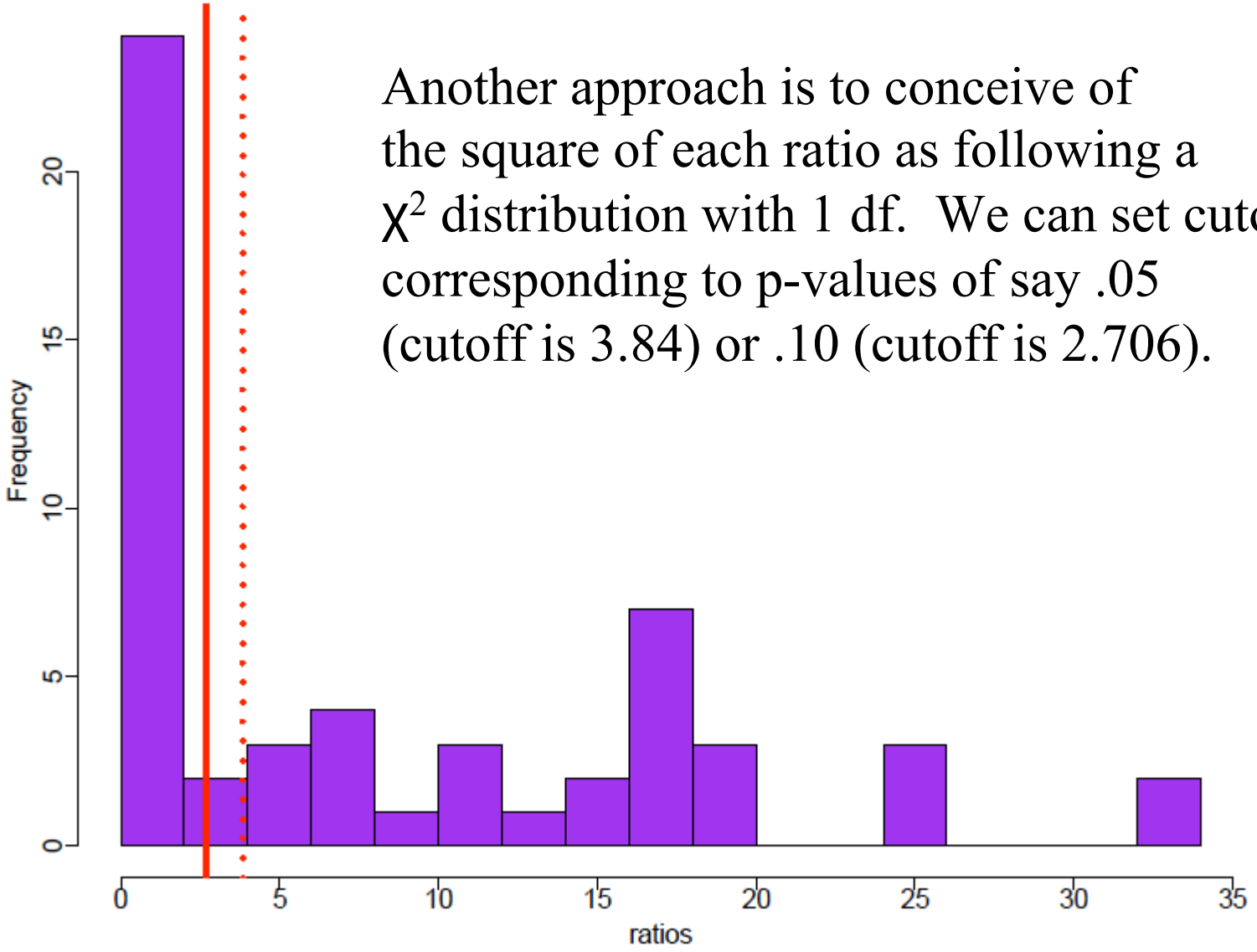
How do we decide when there is too much uncertainty?



sd discard rule



$(s_i^{f0}/s_i^{f1})^2$ ratio rules



Another approach is to conceive of the square of each ratio as following a χ^2 distribution with 1 df. We can set cutoffs corresponding to p-values of say .05 (cutoff is 3.84) or .10 (cutoff is 2.706).

Proposed BART discard rules

Rule 1: **1 sd rule**

discard for unit i (i: z_i=1) if

$s_i^{f0} > \max_i(s_i^{f1}) + \mathrm{sd}(s_i^{f1})$

Rule 2: **.90 rule**

discard for unit i (i: z_i=1) if

$(s_i^{f0}/s_i^{f1})^2 > 2.706$

Rule 3: **.95 rule**

discard for unit i (i: z_i=1) if

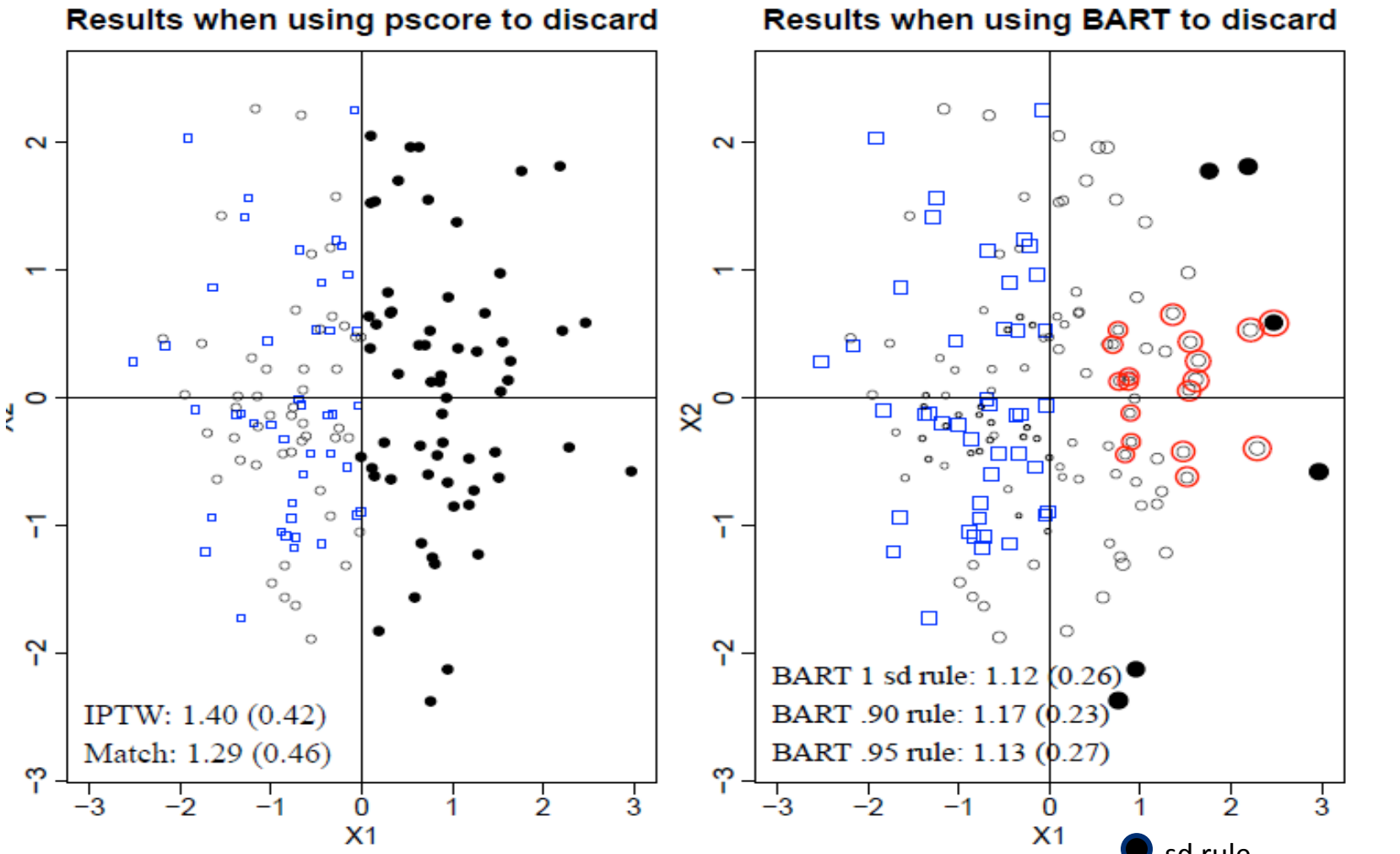
$(s_i^{f0}/s_i^{f1})^2 > 3.84$

sd rule

ratio rules

Example: X_1 predicts treatment, X_2 predicts response

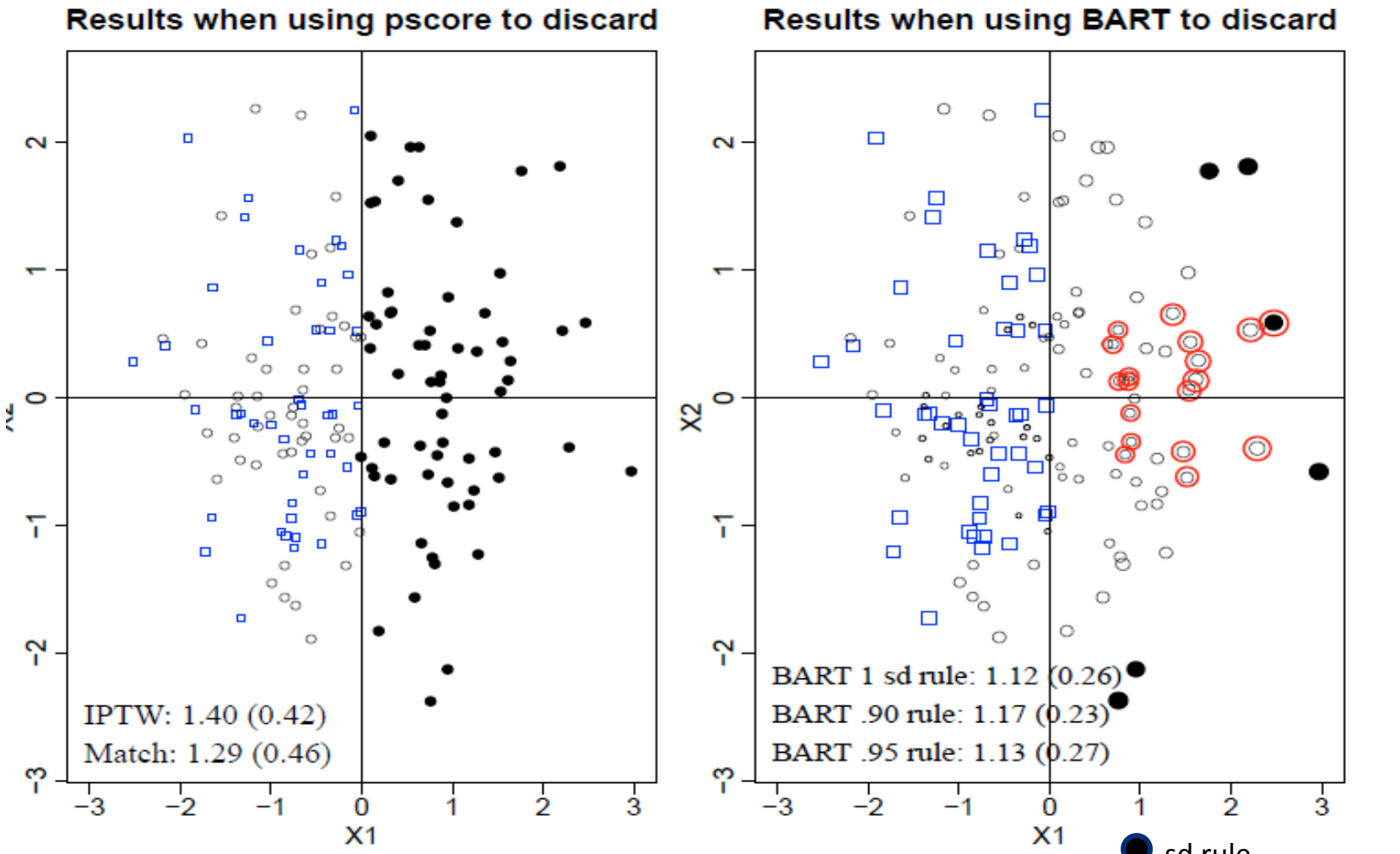
$$E[Y \mid Z, X_1, X_2] = Z + X_2 + X_2^2$$



That means there are NO confounders!!!

Example: X_1 predicts treatment, X_2 predicts response

$$E[Y \mid Z, X_1, X_2] = Z + X_2 + X_2^2$$



true treatment effect =1

Overlap and BART

- Simulation evidence that BART better identifies neighborhoods that lack common support and thus achieves better treatment effect estimates in settings that lack common support compared to methods that ignore outcome info (like pscore methods)
- Including the outcome variable is crucial for understanding common *causal* support
- See full paper for more details (Hill and Su, 2012, Annals of Applied Statistics)

Treatment effect heterogeneity and Generalization

Generalizability/Heterogeneous treatment effects

- When treatment effects are heterogeneous (vary over observations) then it can be tricky to generalize average treatment effects to different/broader populations
- BART has been shown to outperform traditional propensity score approaches at
  - **estimating heterogeneous treatment effects** (Hill, 2011; Green & Kern, 2012) and
  - then **generalizing experimental treatment effects** to broader populations (Kern, Stuart, Hill & Green, 2016)
- Even better versions of BART recognize and correct for bias that can be incurred by the built-in regularization (see Hahn, Murray, Carvalho, 2017 available at https://arxiv.org/abs/1706.09523)

bartCause

A new R package

bartCause

Features…….

- causal effects under the assumption of all confounders measured (average and individual level)
- overlap – tells you what to drop, makes a plot
- IPTW/weights (helps with generalizability though can also just make predictions for a new dataset) [doubly robust]
- TMLE correction option
- uncertainty estimates (with options for different types of confidence intervals)

•Still beta testing HELP US! Find the code at:

https://github.com/vdorie/BartCause

bartCause: The Basics

```
library(bartCause)

fit <- bartc(y, z, x)
fit <- bartc(response, treatment, confounders)
summary(fit)

## summary output
Call: bartc(response = y, treatment = z, confounders = x)

Causal inference model fit by:
  model.rsp: bart
  model.trt: none

Treatment effect:
    estimate     sd ci.lower ci.upper
ate    0.609 0.5208  -0.4118     1.63

95% credible interval calculated by: normal approximation
Result based on 400 posterior samples across 4 chains
```

bartc Methods

```
bartc(y, z, x,
      method.rsp = c("bart", "pweight", "tmle"),
      method.trt = c("none", "glm", "bart", "bart.xval"))
```

bart - just response model, naive bart

pweight - take bart fit, use to compute propensity score weighted average

tmle - like pweight, but use values for TMLE correction

glm - naive fit with confounders added linearly

bart - naive BART fit on treatment model

bart.xval - crossvalidate the prior node sensitivity on treatment model

more bartCause….

```
## common support rules

>bartc(y, z, x,
     commonSup.rule = c("none", "sd", "chisq"),
     commonSup.cut  = c(NA_real_, 1, 0.05))
```

(implements rules discussed above)

```
## visualization
```

`plot_sigma(x)` – trace plot of residual variance, convergence diagnostic
`plot_est(x)` – trace plot of estimate, convergence diagnostic
`plot_indiv(x)` – histogram of individual treatment effect estimates
`plot_support(x)` – common support plot

What about the hard part?

RECALL: **Structural Assumptions**

- The key assumption in most observational studies is that all confounders have been measured (ignorability, selection on observables, conditional independence, …) Formally this implies

$$Y(0), Y(1) \perp Z \mid X$$

This assumption is **untestable** and **difficult to satisfy**

Can we loosen the assumption that we've measured all confounders?

Sensitivity Analysis

Goal of our sensitivity analysis work

- The goal is to provide a way of sensitivity of our inferences to violations of the assumption that all confounders have been measured while
  - Allowing for less strict parametric assumption
  - Maintaining interpretability of parameters

Classic example: smoking and cancer

- In the 1950's the empirical *association* between smoking and lung cancer was clear
- However those arguing against a causal interpretation posited that a third variable might be (at least partially) responsible both for the desire to smoke and the probability of getting lung cancer

Cornfield's sensitivity analysis

- Cornfield responded by focusing on this potential omitted variable
- What would that variable have to "look like" in order for it to explain away the association between the two? In particular how *strongly* would it have to be associated with both smoking and lung cancer?
- Turns out, it would have to be *very strong*
  - it would have to be an almost perfect predictor of lung cancer
  - it would have to be about 9 times more prevalent in smokers than non-smokers
- This level of association was implausible for any factor that anyone could conceive of

Sensitivity to an unobserved confounder

- Thought experiment: What if our identification strategy fails to control for one important confounder, ***U*** **?**
- That is. what if we are missing one important confounder, ***U***:

$$Y(0), Y(1) \perp Z \mid X, \boldsymbol{U}$$

- The problem is that we do not know what ***U*** "looks like" (i.e. what's it's relationship is with Y and Z)
- Thus we will need to explore the impact of this potential ***U*** over a range of plausible options

Extension of previous work on 2-parameter sensitivity approaches

This work builds on similar approaches to sensitivity analysis that have 2 parameters

- Imbens (2003) likelihood-based approach using standard models for response surface and assignment mechanism.
- Harada (2012) Simulation based approach based on linear models
- Carnegie, Harada and Hill (2015). Likelihood-based approach that uses simulation to explore the range of possible values of the treatment effect corresponding to a set of sensitivity parameters. Operationalized in R package `treatSens` available on CRAN.

Data Generating Process

The full model simulated by our semiparametric SA is:

$$
\begin{aligned}
Y \mid U, \mu_{\mathrm{xz}}, \sigma_{\mathrm{y}}^2 &\sim \mathrm{N}\left(\mu_{\mathrm{xz}} + \zeta^{y} U, \sigma_{\mathrm{y}}^2\right), \\
Z \mid X, U, \beta^{z} &\sim \mathrm{Bernoulli}(\Phi(X\beta^{z} + \zeta^{z} U)), \\
\mu_{\mathrm{xz}}, \sigma_{\mathrm{y}}^2 \mid X, Z &\sim \mathrm{BART}(X, Z), \\
U &\sim \mathrm{Bernoulli}(\pi^{u}), \\
\beta^{z} &\sim p(\beta^{z}),
\end{aligned}
$$

where $p(\beta^z)$ is a flat, normal, or t-distribution and π^u is a hyperparameter

Data Generating Process

These are the sensitivity parameters

The full model simulated by our semiparametric SA is:

$$
\begin{aligned}
Y \mid U, \mu_{xz}, \sigma_y^2 &\sim \mathrm{N}\left(\mu_{xz} + \zeta^y U, \sigma_y^2\right), \\
Z \mid X, U, \beta^z &\sim \mathrm{Bernoulli}(\Phi(X\beta^z + \zeta^z U)), \\
\mu_{xz}, \sigma_y^2 \mid X, Z &\sim \mathrm{BART}(X, Z), \\
U &\sim \mathrm{Bernoulli}(\pi^u), \\
\beta^z &\sim p(\beta^z),
\end{aligned}
$$

where $p(\beta^z)$ is a flat, normal, or t-distribution and π^u is a hyperparameter

Data Generating Process

The full model simulated by our semiparametric SA is:

$$
\begin{aligned}
Y \mid U, \mu_{xz}, \sigma_y^2 &\sim \mathrm{N}\left(\mu_{xz} + \zeta^y U, \sigma_y^2\right), \\
Z \mid X, U, \beta^z &\sim \mathrm{Bernoulli}(\Phi(X\beta^z + \zeta^z U)), \\
\mu_{xz}, \sigma_y^2 \mid X, Z &\sim \mathrm{BART}(X, Z), \\
U &\sim \mathrm{Bernoulli}(\pi^u), \\
\beta^z &\sim p(\beta^z),
\end{aligned}
$$

This is the "machine learning" piece

where $p(\beta^z)$ is a flat, normal, or *t*-distribution and π^u is a hyperparameter

Overview of MCMC algorithm

- We start by specifying a pair if sensitivity parameters (ζ^y, ζ^z)
- If we know U then it's easy to fit a model for the treatment assignment and a model for the response surface and get a treatment effect estimate.
- If we know all the other parameters in the model it's "easy" to draw U
- We iterate between these states until convergence; this yields a posterior distribution of the treatment effect conditional on ζ^y and ζ^z
- We repeat this process across a grid of (ζ^y, ζ^z) values

Example: effect of medication on high blood pressure

Evidence of nonlinearities

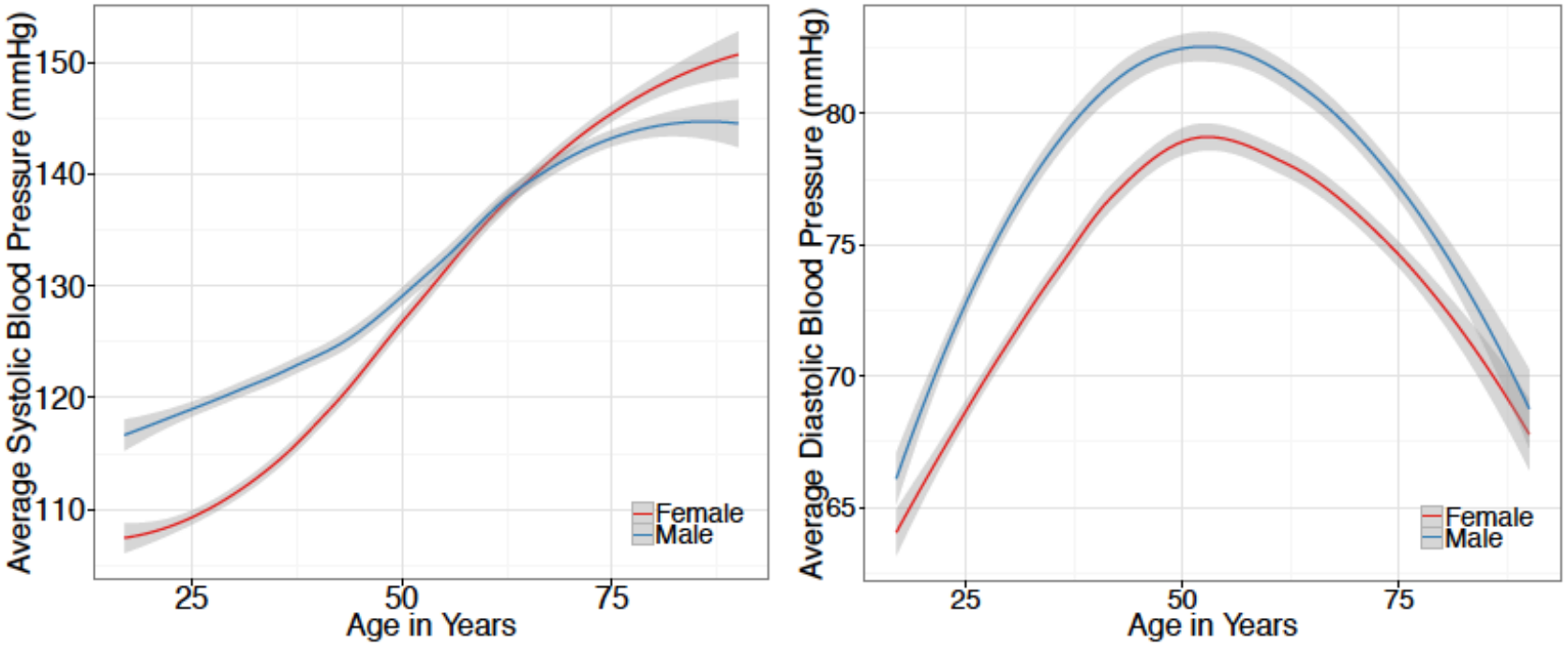
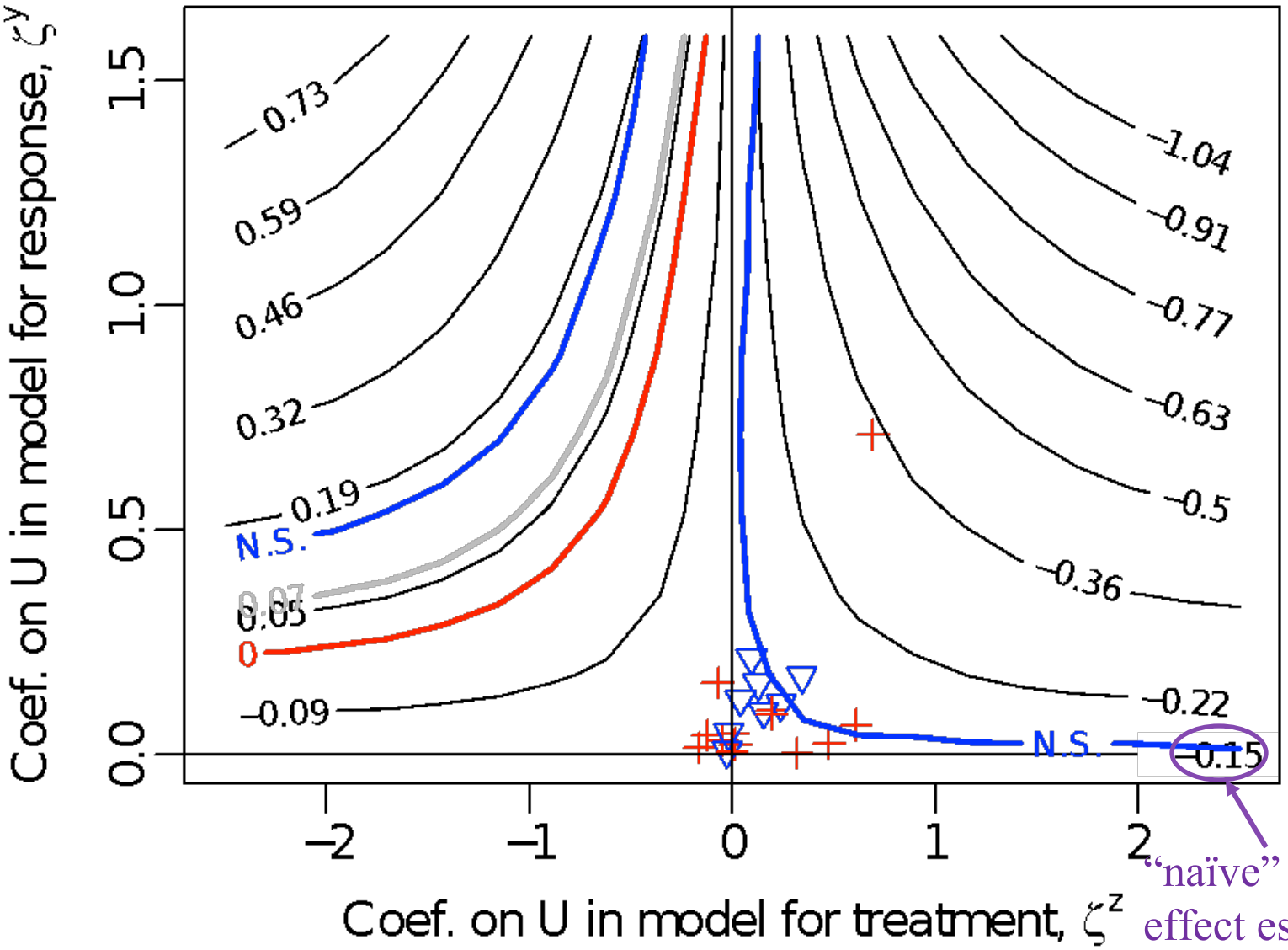


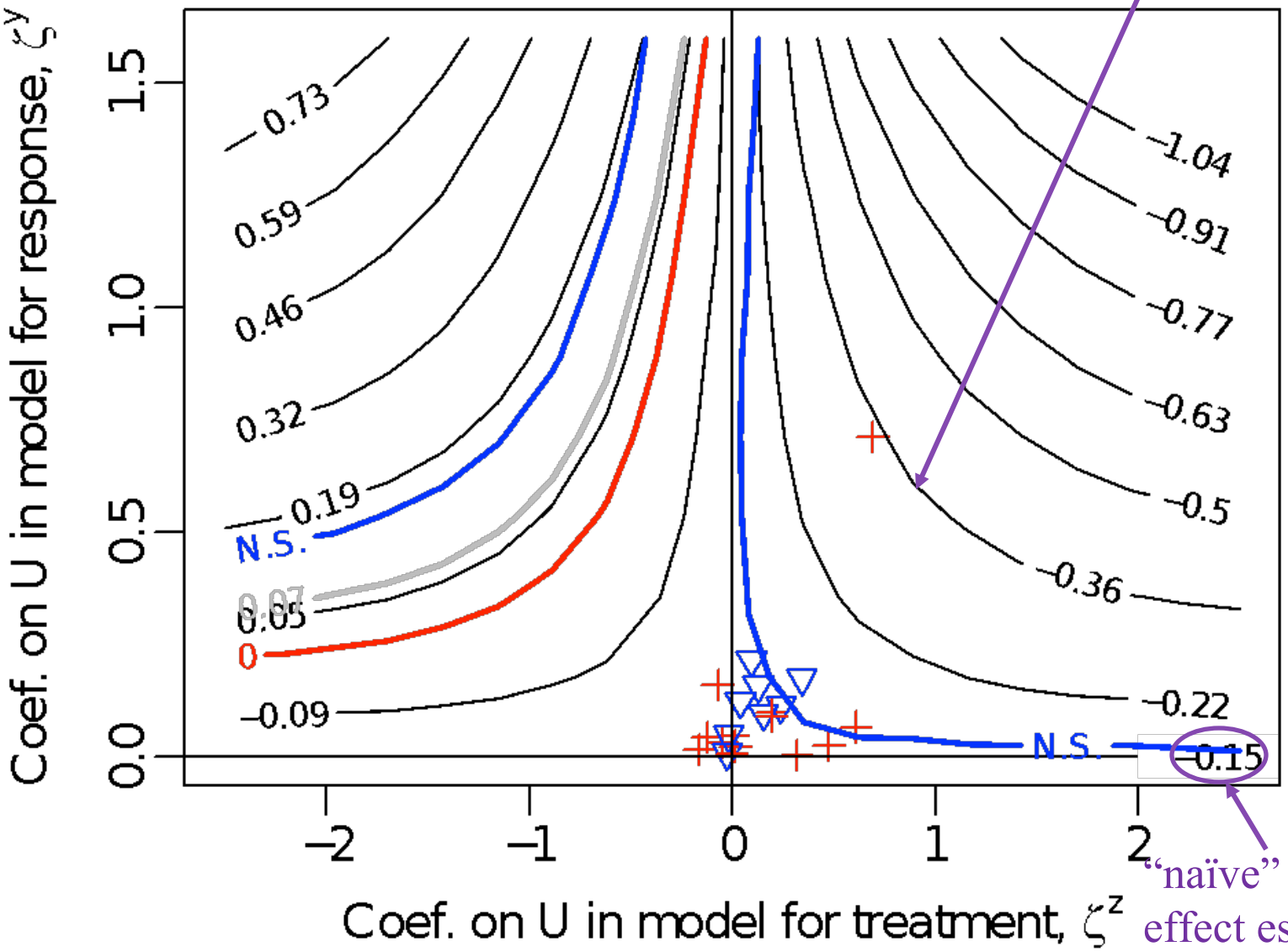
Figure: LOESS curves of average systolic (left) and dystolic (right) blood pressure for all individuals in the NHANES III dataset, plotted against age and separated by sex. The shading show point-wise 95% confidence intervals for the mean.

Sensitivity: Effect of beta blockers/diuretics on systolic b.p.



combinations of ζ^z and ζ^y that would lead to a standardized treatment effect estimate of -.36

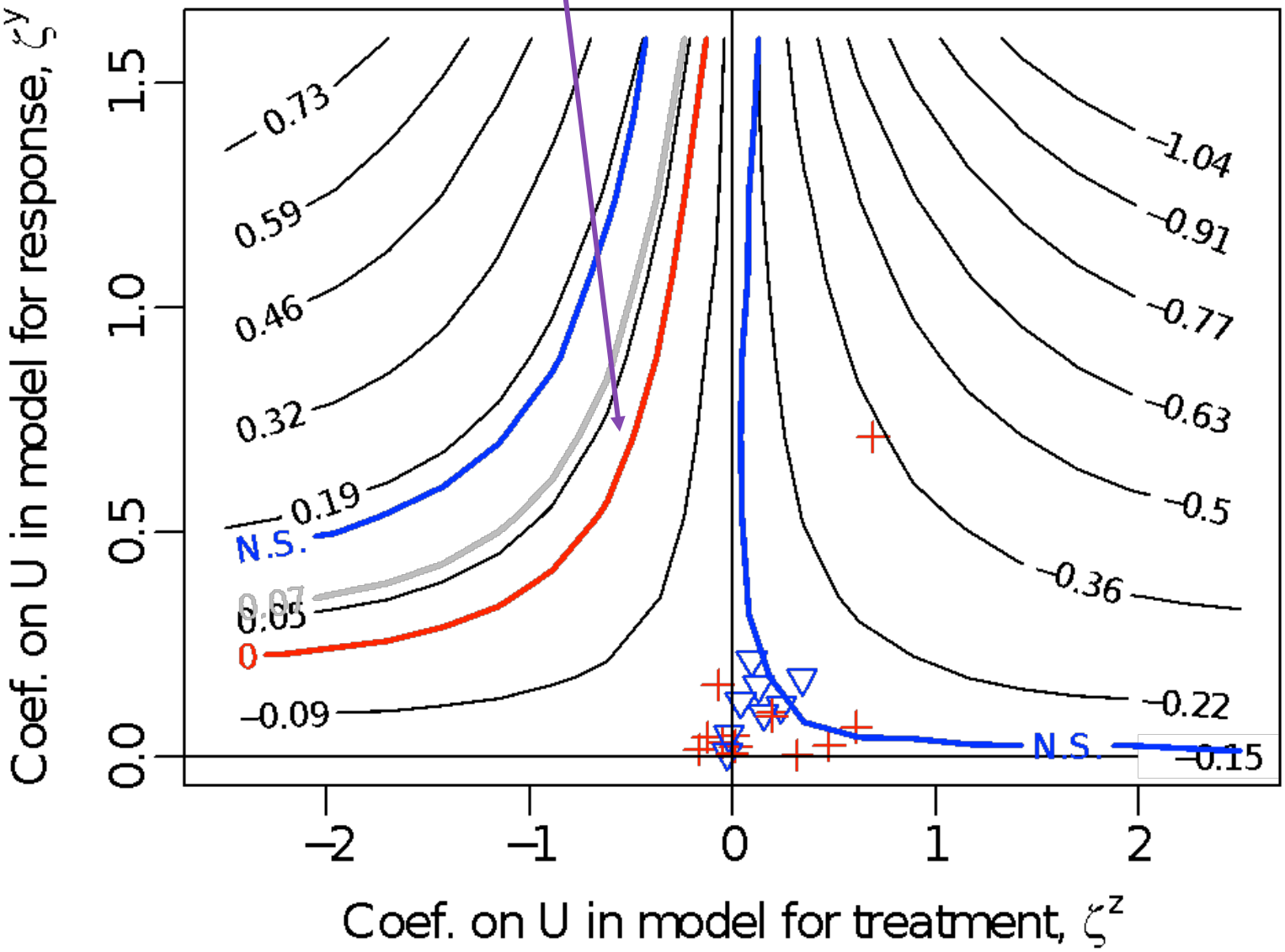
Sensitivity: Effect of beta blockers/diuretics on systolic b.p.



"naïve" treatment effect estimate

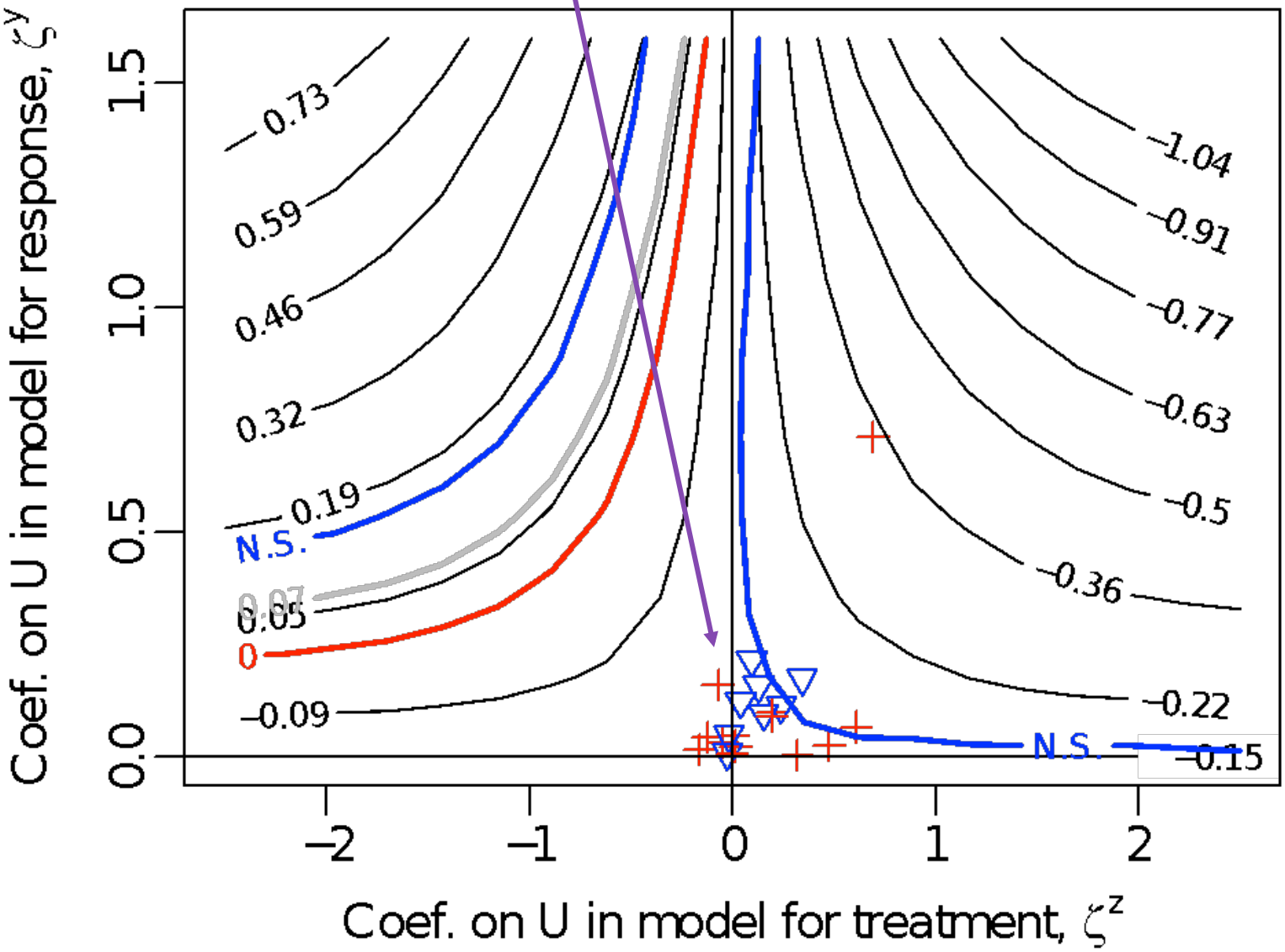
combinations of ζ^z and ζ^y that would lead to a treatment effect estimate of 0

Sensitivity: Effect of beta blockers/diuretics on systolic b.p.



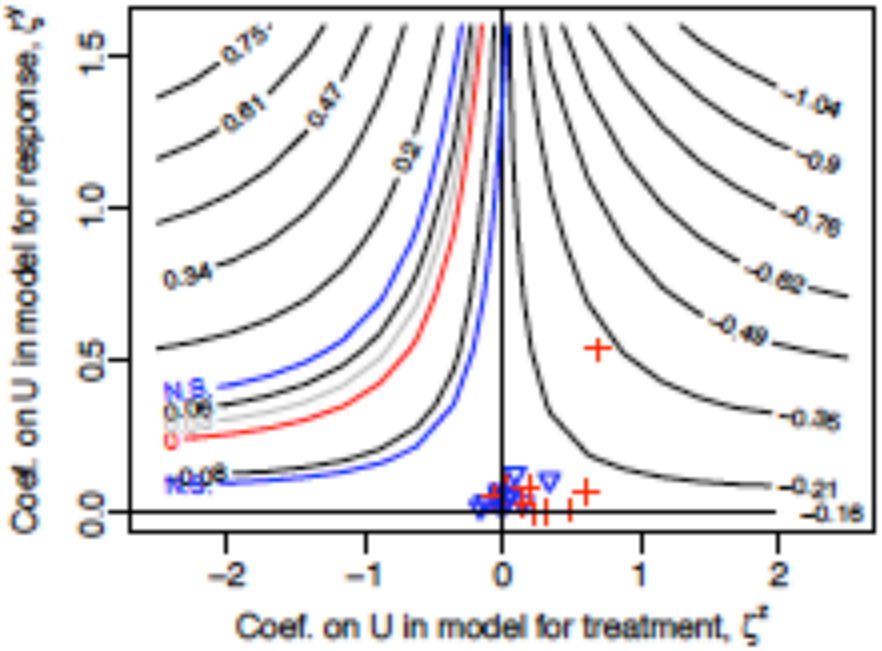
(standardized) coefficients from a regression of the outcome on the treatm
and observed covariates (▽ are for variables whose sign we reversed)

Sensitivity: Effect of beta blockers/diuretics on systolic b.p.

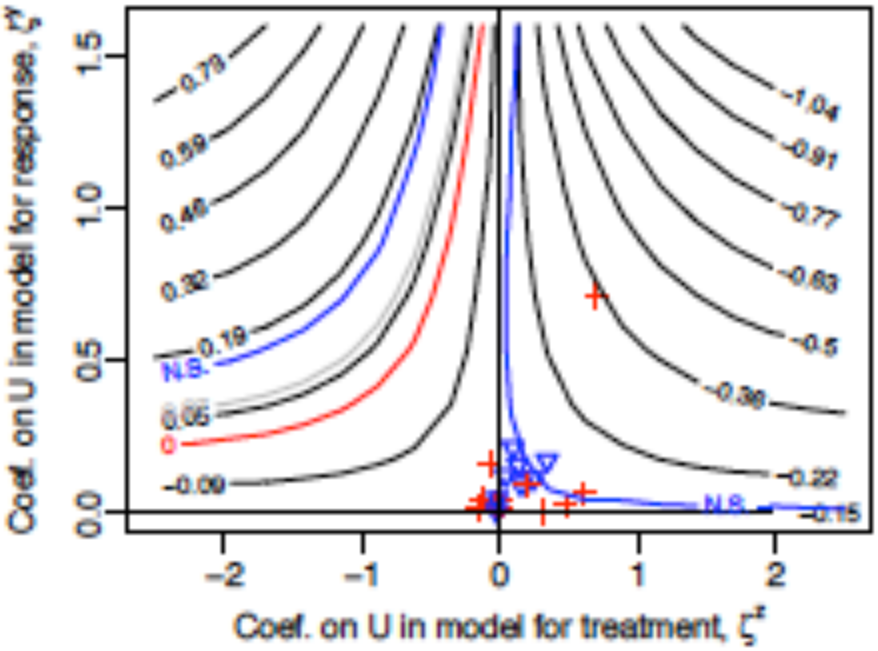


Results: Effects of beta blockers/diuretics on systolic

Linear Fit



BART Fit



Sensitivity analysis with a linear fit to the covariates would have told a very different story.

paper/code

Dorie et al., *Statistics in Medicine*, 2016

treatSens (CRAN) -- use BART option

Why BART?

Is BART the key?

- What's so special about BART? Could we use any machine learning tool here?
- Key features
  - Flexible fit to response surface (lots of competition)
  - Coherent uncertainty intervals (some competition)
  - Incorporating the outcome for understanding what covariates are important/understanding common support (mixed)
  - Easy!!!!!!
- What about alternatives?
- Some being developed by currently don't seem to have this breadth of features – see results of the recent Causal Inference Data Analysis Challenge (https://arxiv.org/abs/1707.02641)

Conclusions/Discussion

Overall Conclusions: BART for Causal Inference

- BART is useful for causal inference along several dimensions including
  - loosening parametric assumptions
  - examining overlap of true confounders
  - quantifying uncertainty
- Not the only game in town in terms of flexible model fitting but is easy and seems to work well across a variety of settings
- Need more development... more testing... more features…. (binary BART, grouped data structures, other languages, ……)

R packages

R packages

- dbarts (on CRAN)
- bartCause (beta testing before putting on CRAN)
  https://github.com/vdorie/BartCause
- treatSens (should be on CRAN, otherwise https://github.com/vdorie/treatSens/tree/v2.1)
- Package to create data from 2016 Causal Inference Data Analysis Challenge
  https://github.com/vdorie/aciccomp/tree/master/2016
  (on CRAN soon)