Causal Inference and Data Science: Why They Need Each Other

Jennifer Hill

presenting work that is joint with

Nicole Carnegie (Harvard University), Masataka Harada (NYU), Yu-Sung Su (Tsinghua University), Chris Weiss (Langer Research Assoc.), and Fuhua Zhai (Stony Brook), Vincent Dorie (NYU)

March, 2010

Acknowledgement: This work was supported by IES Grant [ID: R305D110037].

Roadmap

1. Data Science Needs Causal Inference
2. Causal Inference
3. Causal Inference Needs Data Science
4. Example of work at the intersection

Part I:
Data Science Needs Causal Inference

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?

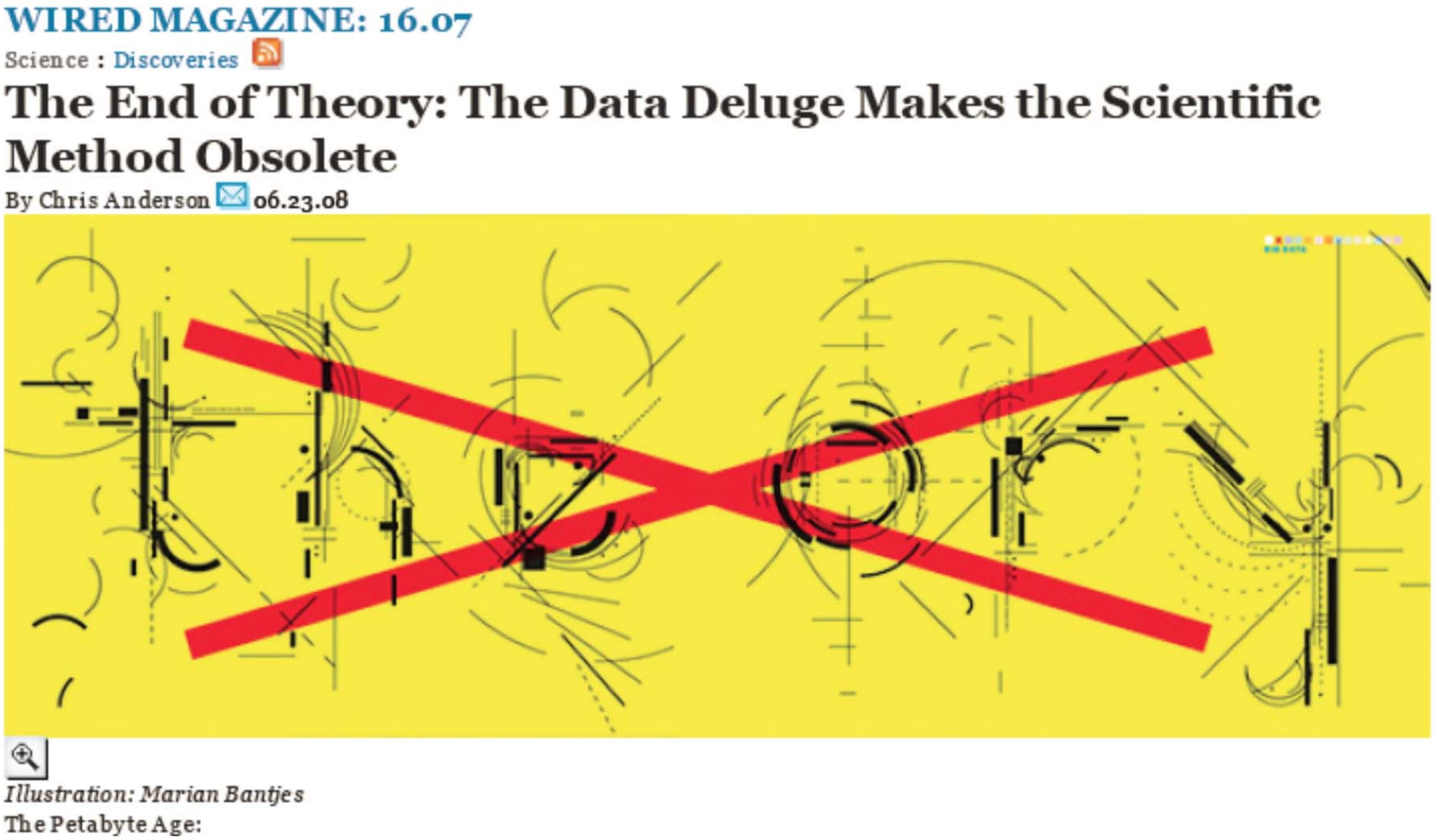
The New York Times

Social/behavioral/health sciences and causal inference

- The social, behavioral, and health sciences are littered with cautionary tales of the dangers of inappropriately inferring causality
  - Salk vaccine
  - Reading is Fundamental
  - Nurses Health Study versus WHI*
- While there is work to be done in educating researchers (and policy makers) to better understand these, the basic understanding of the issues is there for most researchers
- Data Science on the other hand is vulnerable to the dangers of not knowing what it doesn’t know……

*However see great work by Hernan and Robins (2008) has used domain knowledge and clever design/methods to reconcile these…

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 was spent in the U.S. in 2011 on internet advertising
- Common wisdom in the marketing world is that internet advertising is highly effective (can position ads based on type of browsing activity to better target interested shoppers)
- Impact of this type of advertising has been considered easier to measure because can track info on those who click on ads (including "Did they buy?")
- Prediction models suggest that clicking on the ads strongly increase probability of success (reaching website, buying product)
- However what is difficult to measure by just observing the (unmanipulated) system is whether shoppers would have reached the relevant website or bought the product *anyway*

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

Cautionary Tale: SEM (cont)

- Researchers now have conducted quasi-experiments to investigate
- Question: What are the returns of brand keyword search advertising? (e.g. “ebay shoes”)
- Study design: In March 2012 eBay
  - halted advertising of its brand related terms on Yahoo! and MSN, but
  - continued advertising on Google.
- Results: 99.5% of click traffic was simply redirected through “natural” (unpaid) search traffic. i.e. almost everyone found the site anyway
- Results from other studies showed that non-brand keyword ads have a positive effect on new/infrequent users but no effect for the more frequent users (who represent most of the cost), thus average returns are *negative*.
- From Blake, T., Nosko, C., and S. Tadelis (2013) “Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment”

Summary

- Data Science needs Causal Inference to prevent data scientists from saying silly things and making recommendations that could be problematic
- Big Data
  - has the potential to be helpful (particularly if we can use it to measure more and better)
  - can actually exacerbate the problem. If we have a poor design/use inappropriate methods/don’t think hard about the assumptions, more data will simply *make us more confident about our incorrect inference*

    (see work by Paul Rosenbaum on why less can be more)

Part II:
What is Causal Inference?

Why are causal questions so tricky to answer??

Because implicitly we are trying to figure out **what** would have happened to people **if** they had taken a different path

Causal Inference, Notation, Assumptions

Counterfactuals

- The most important concept in causal inference is that of the **counterfactual**
- Most causal inference statisticians define causal effects as comparisons between what *would happen* in two or more different states (one of which will be *factual*, the other(s) *counterfactual*)
- Examples
  - headache status one hour after taking ibuprofin versus not
  - contracting Polio if vaccinated versus if not
  - reading ability at age 6 if Illinois doesn't send children books versus if it sends a book a month until Kindergarten
  - coronary heart disease if receiving HRT versus if no HRT
  - decision to buy a product on ebay if eBay is the top sponsored choice versus if eBay were not included in this list at all

Potential outcome notation

- We operationalize counterfactuals by using potential outcome notation (Rubin, 1978)
- For a given treatment and control condition, each person can be thought of as having two potential outcomes, Y(0) and Y(1)
  - Y(Z=0) =Y(0) is the outcome if treatment is not received
  - Y(Z=1) =Y(1) is the outcome if treatment is received

Defining Causal Effects

- We use these potential outcomes (Y(0), Y(1)) to define a causal effect for subject *i* as a comparison between them, such as,

  $$\tau_\iota = Y_i(1) - Y_i(0)$$

- **Fundamental Problem of Causal Inference:** we can never observe both $Y_i(0)$ and $Y_i(1)$
- (One can think of this as a missing data problem)
- So how do we proceed?
  - focus on average causal effects
  - make some assumptions

Causal Inference Notation

We will use the following notation. Let

- X be a (vector of) observed covariates
- Z be a binary treatment variable (can be extended beyond binary)
- $Y(0)$, $Y(1)$ be the set of potential outcomes corresponding to a treatment variable
- Y be the observed outcome

$$Y = Y(0)*(1-Z) + Y(1)*Z$$

Structural assumptions: Ignorability

- For binary Z we assume

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

- This assumption is referred as ignorability in the Statistics literature
- It is referred to by different names in different fields (all confounders measured, selection on observables, randomization, CIA, exchangeability,…).
- Colloquially, it means that we have conditioned on all confounders (pre-treatment variables that predict both treatment and outcome)

Ignorability special case: randomized experiment

- A special case occurs when we have a completely randomized experiment

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

- When we have a randomized block experiment we know that

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

  (where W denotes the blocking variables)
- In these cases we know the assumptions hold
- Otherwise it is a leap of faith because the assumption is untestable

Implications of ignorability

- When ignorability holds we know that
- E[Y | X, Z=1] = E[Y(1) | X], and
- E[Y | X, Z=0] = E[Y(0) | X]

  thus we can estimate E[Y(1) – Y(0) | X], without bias by estimating the relevant conditional expectations)
- Then we can average over this the distribution of X to estimate E[Y(1) – Y(0)]
- In most settings ignorability is a heroic assumption. Thus in general we try to
  - avoid it by using a randomized experiment
  - weaken it by using a strong quasi-experimental design
  - test how problematic it is using sensitivity analysis

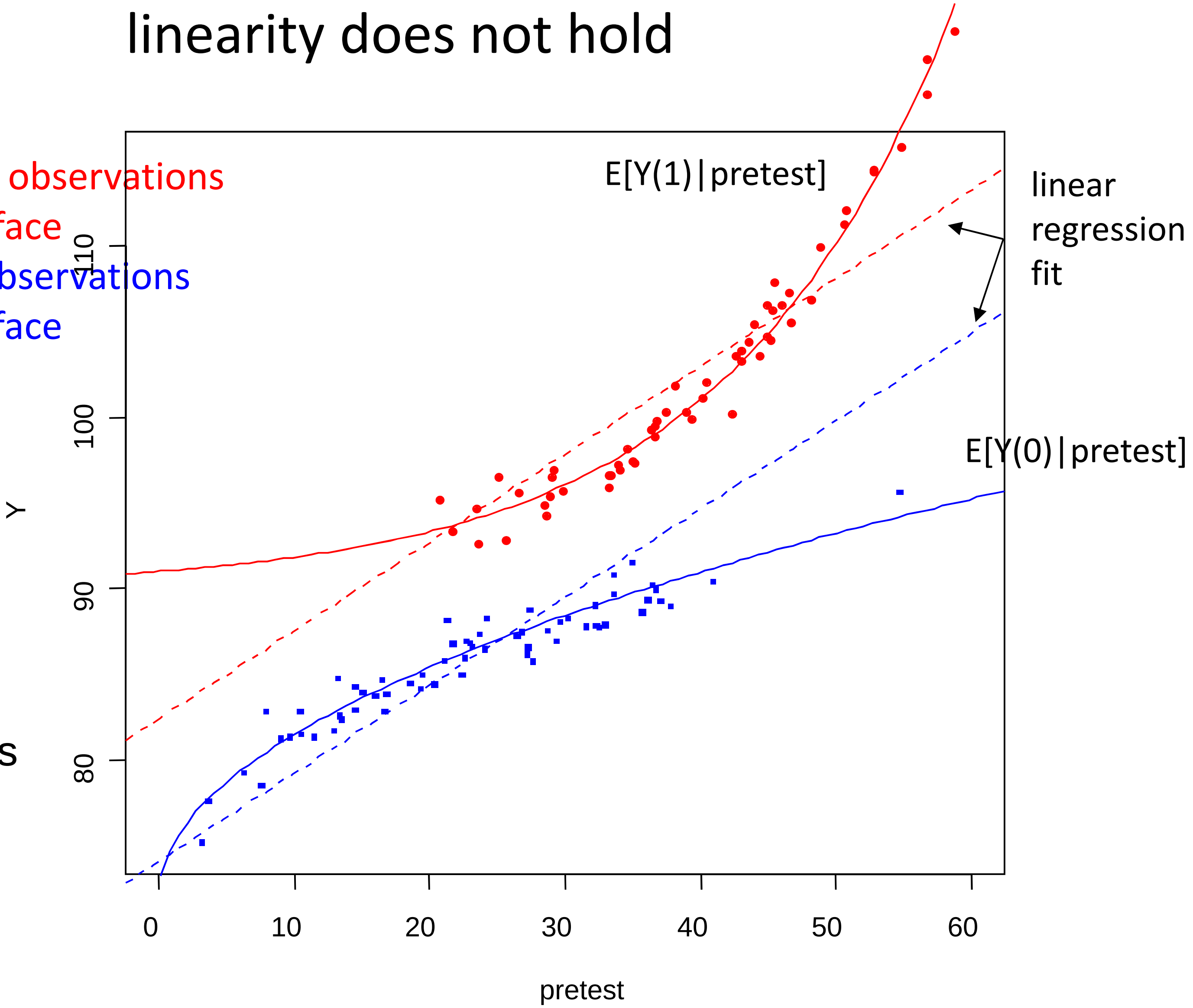
Parametric Assumptions

- We need models or fitting algorithms to estimate the requisite conditional expectations
- The easiest solution is to fit using linear models

$$E[Y(Z = z) \mid X = x] = \beta x + \tau z$$

Observational studies: one confounding covariate, linearity does not hold

red for treatment observations
and response surface
blue for control observations
and response surface



Linearity is not an appropriate model here

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

Parametric Assumptions

- We need models or fitting algorithms to estimate the requisite conditional expectations
- The easiest solution is to fit using linear models

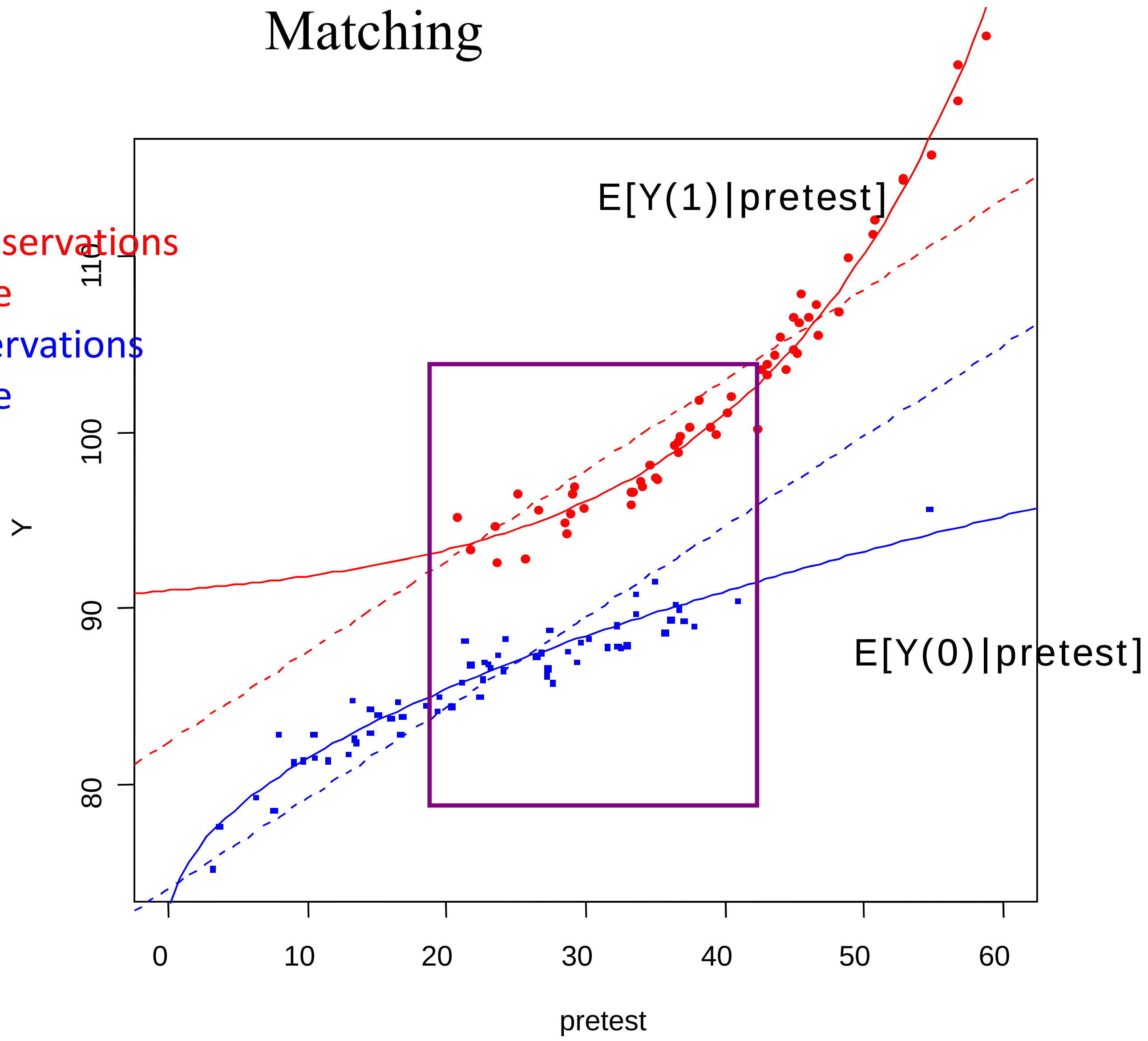
  $$E[Y(Z = z) \mid X = x] = \beta x + \tau z$$

  typically not a good choice
- How to address?
  - Randomized experiments obviate the need for fitting such models and if models are used (e.g. to increase efficiency) they are more robust to model misspecification
  - A huge amount of work in causal inference in the past two decades has focused on relaxing these assumptions

Matching

red for treatment observations
and response surface

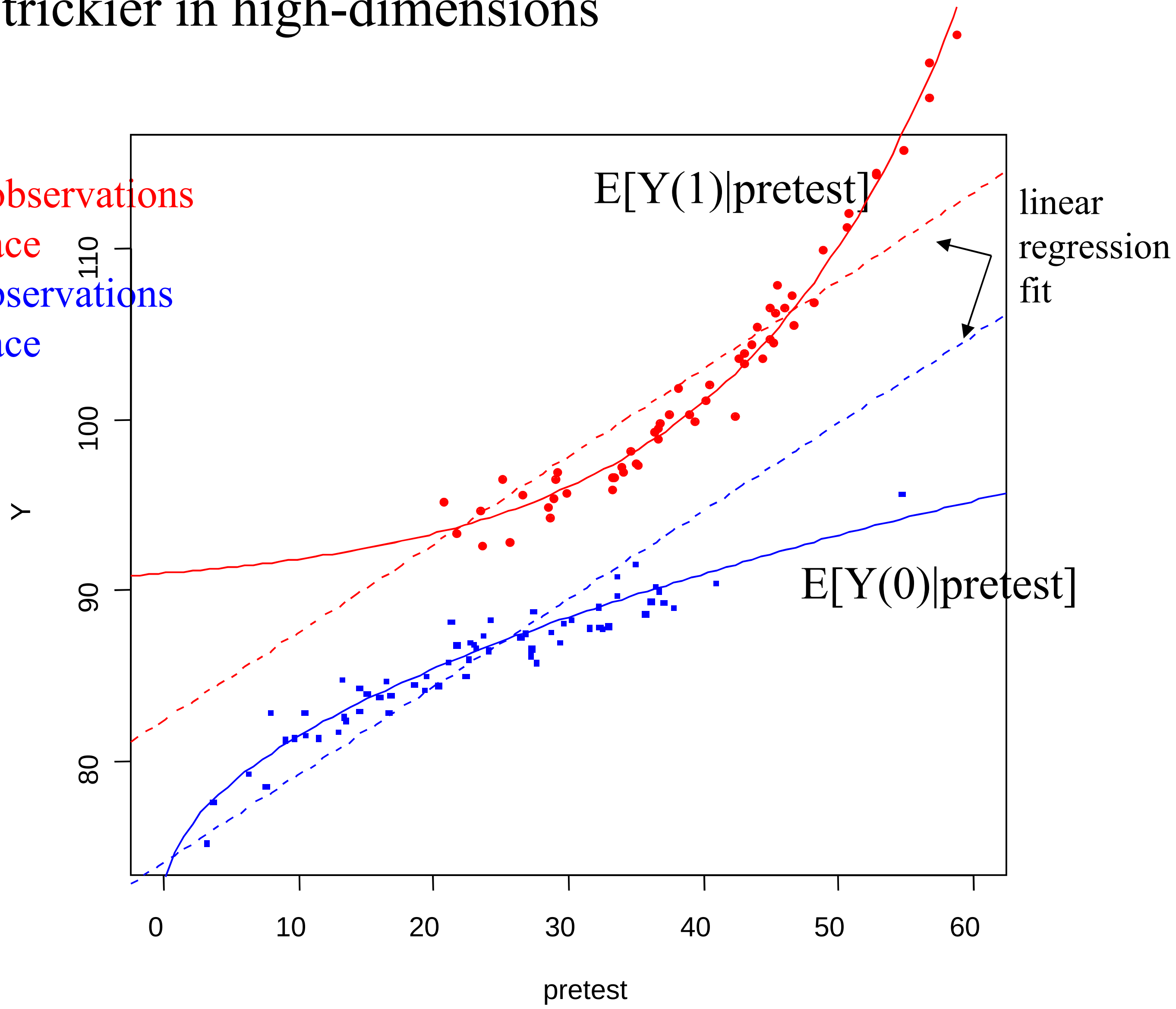
blue for control observations
and response surface



Matching would constrain inference to areas of common support (overlap) – purple box

The other alternative is to fit the curve directly – this gets trickier in high-dimensions

red for treatment observations
and response surface
blue for control observations
and response surface



Part III:
Causal Inference Needs Data Science

Why Causal Inference Needs Data Science

- First need to define Data Science.
- And since this is the inaugural lecture in this series, it turns out I get to decide.
- :)
- I'm going to define Data Science as the optimal intersection of
  - Data engineering
  - Data visualization
  - Machine Learning
  - Statistics
  - Domain-specific knowledge

What I won’t be talking about

To my knowledge there has been less work on the ways that

- Data engineering
- Data visualization

could benefit causal inference though these could be fertile areas for further exploration

What will be implicit?
We need domain expertise to do causal inference.

Seems obvious but….

- Figure out the interesting research questions
- Understand when the assumptions are plausible
- Inform measurement and choice of confounders
- Inform study design
- Inform choice of estimands
- Help understand the “why” of an effect
- ….

Why do we need Statistics to do causal inference?

Statistics* has been the primary (first?) field to

- Formalize the assumptions underpinning causal inference
- Formalize specific causal inference estimands
- Develop **study designs** that allow researchers to satisfy or weaken these assumptions
- Develop **methods** for data analysis to satisfy or weaken the assumptions and allow researcher to estimate a variety of different estimands
- Understand our **uncertainty** about our inferences. Account for multilevel or otherwise dependent data structures.

*(I'll include econometrics, psychometrics, etc in the definition of Statistics because my goal here is to distinguish this work from the work that goes on in the other Data Science fields)

Why do we need Machine Learning to do causal inference?

Machine learning can help by

- development of nonparametric prediction algorithms that can be used to fit response surfaces using fewer assumptions
- can be used to automate exploration of competing models for causal processes (causal exploration of DAG's)
- more…

Part IV:
Example of work at the intersection

Illustration of my work at the intersection

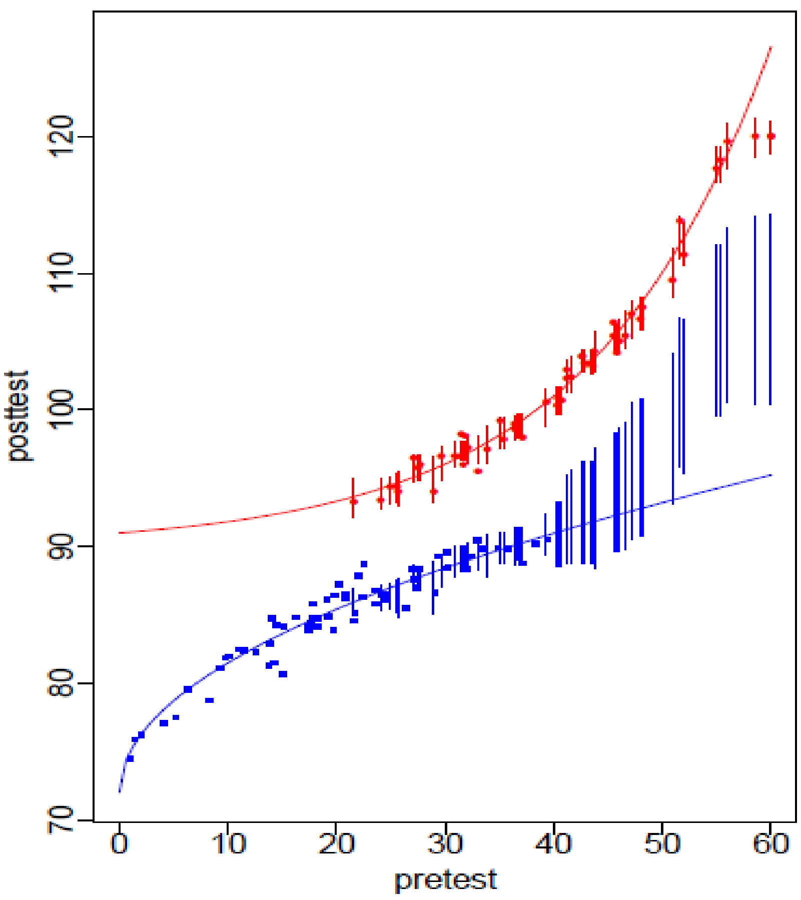
- Causal inference
- Statistics
- Machine Learning

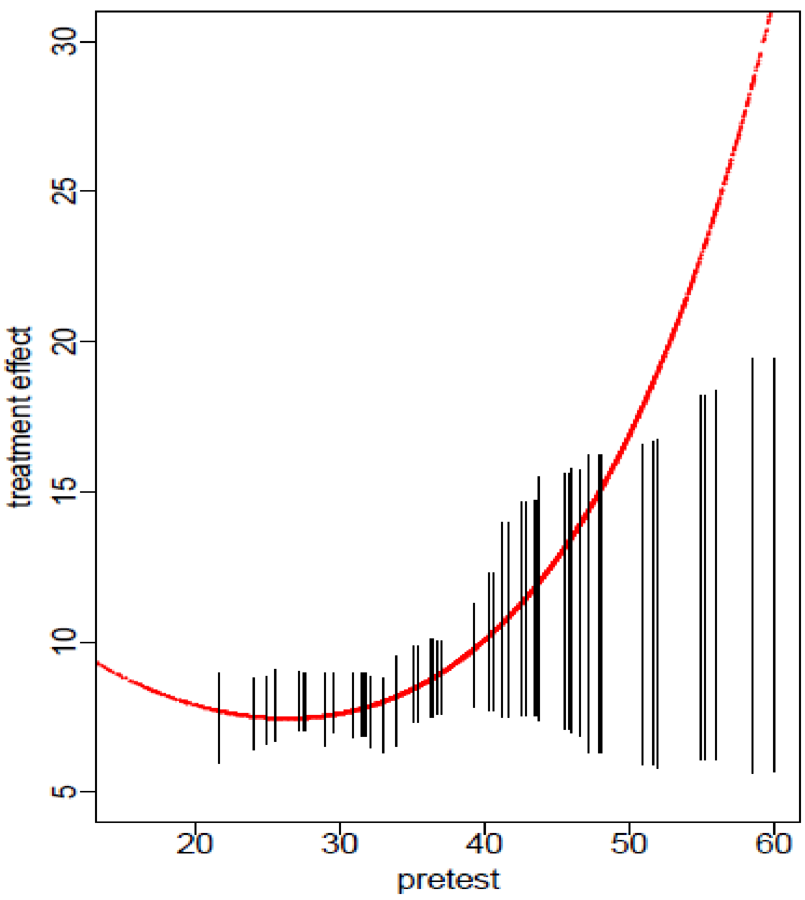
1. BART for causal inference as a strategy for relaxing *parametric* assumptions
2. Integrating BART with more traditional sensitivity analysis as a strategy for relaxing *structural* assumptions

Motivation: relaxing parametric assumptions

- In non-experimental studies we typically have to estimate conditional expectations such as, E[Y | Z, X].
- The major challenge in this area is to avoid strong parametric assumptions.
- Strategies that have been used
  - matching
  - semi-parametric modeling (theory/math driven, can be difficult in high dimensions)
  - Bayesian nonparametric algorithms (Hill, Tokdar, Karabatsos working in this area**)**
- They may also have to account for more complicated independence assumptions (e.g. using multilevel models)

Causal inference using BART: treatment on treated


Response surface and BART fit


Treatment effect and BART fit

Bayesian Additive Regression Trees
BART

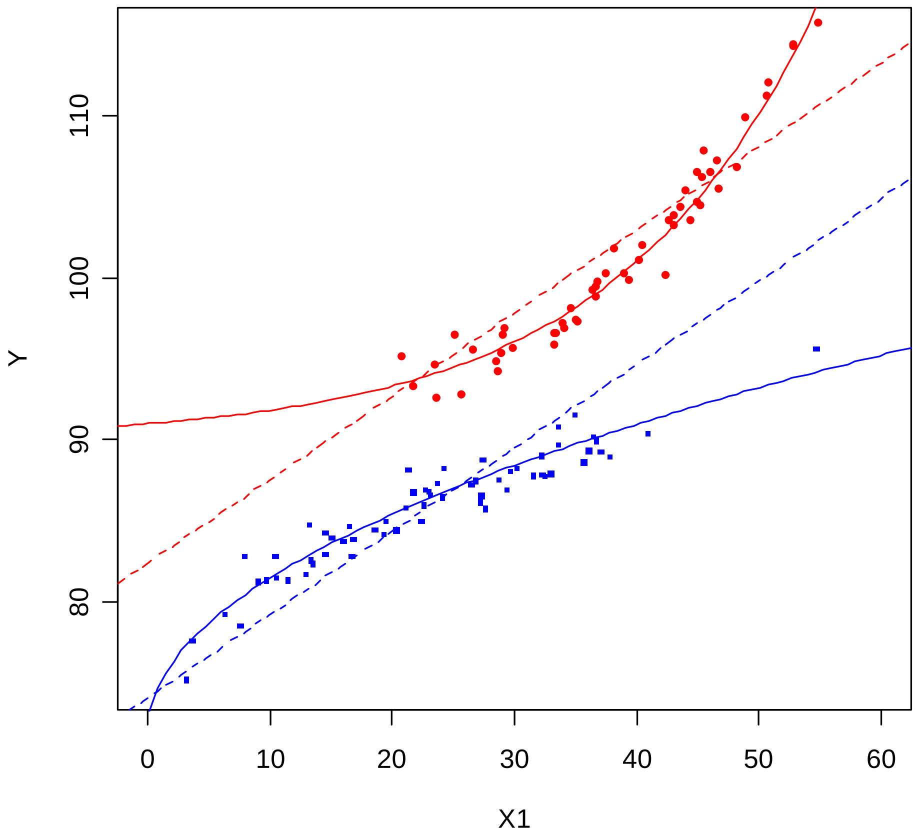
BART (Chipman, George, McCulloch, 2007, 2010)

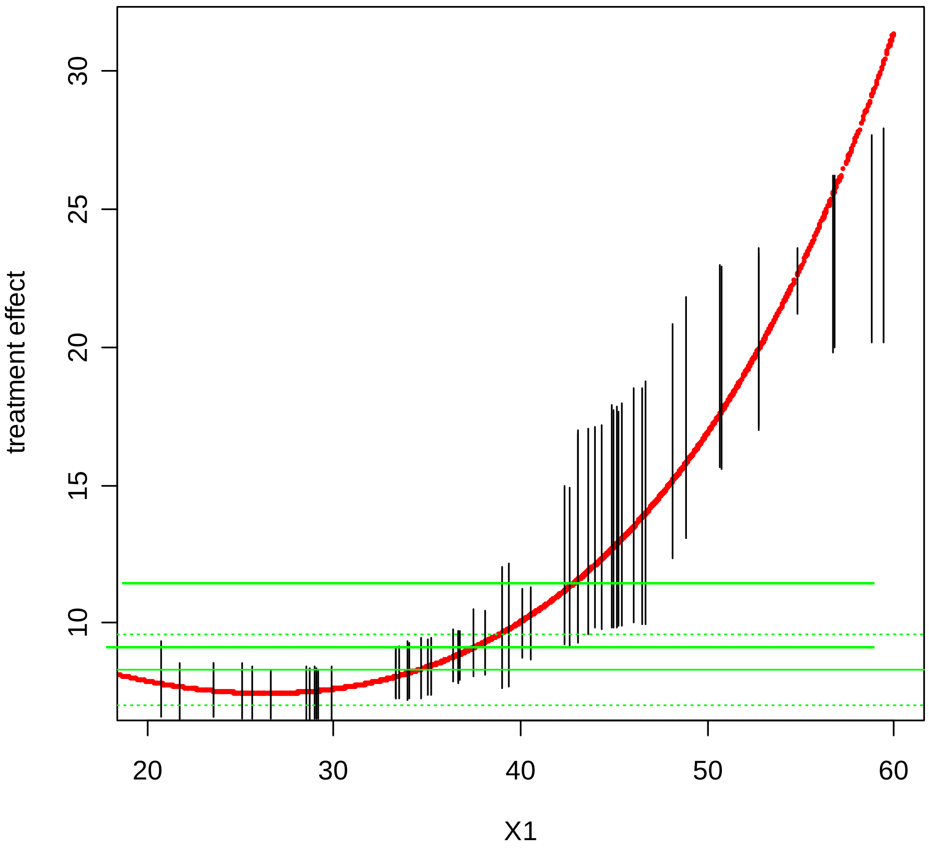
- ❖ overview
- ❖ review of previous studies using BART for causal inference
- ❖ BART model and algorithm
- ❖ BART for causal inference
- ❖ BART for sensitivity analysis

Motivation for BART for causal inference: flexibility, honesty, simplicity

- FLEXIBILITY:
  - BART can flexibly fit even highly non-linear response surfaces even with a large number of predictors (with great out-of-sample prediction properties)
  - BART produces coherent uncertainty intervals
- HONESTY:
  - BART does not require the researcher to specify which variables are important or the functional form of the relationship between predictors and the outcome – it *finds* interactions and non-linearities! This helps to keep the researcher honest! (no playing around with the functional form until you get the answer you want)
- SIMPLICITY
  - Software freely available (on CRAN as BayesTree) and easy to use!

BART versus linear regression





Green shows regression estimate of treatment effect and uncertainty bounds

Motivation for BART for causal inference: flexibility, honesty, simplicity

- FLEXIBILITY
  - BART can flexibly fit even highly non-linear response surfaces even with a large number of predictors (with great out-of-sample prediction properties)
  - BART produces coherent uncertainty intervals
- HONESTY
  - BART does not require the researcher to specify which variables are important or the functional form of the relationship between predictors and the outcome – it *finds* interactions and non-linearities! This helps to keep the researcher honest! (no playing around with the functional form until you get the answer you want)
- SIMPLICITY
  - Software freely available (on CRAN as BayesTree) and easy to use!

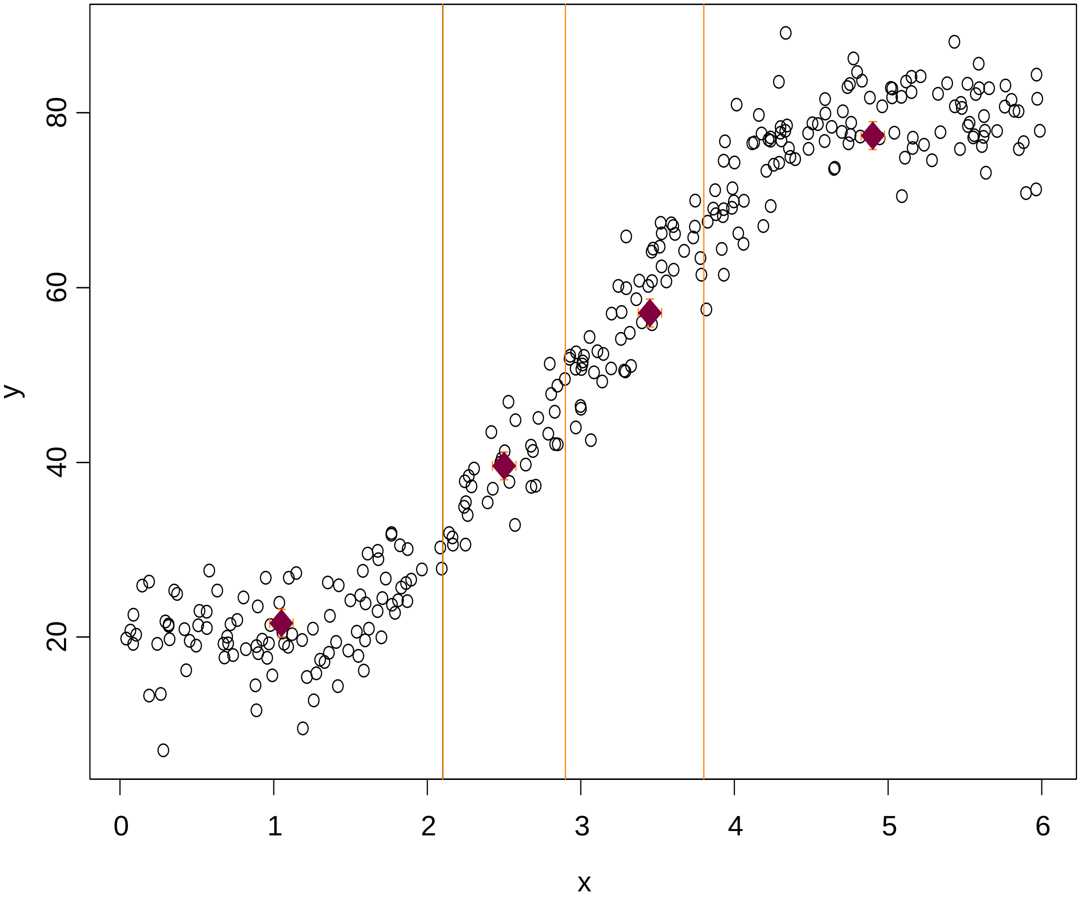
Understanding how BART works

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

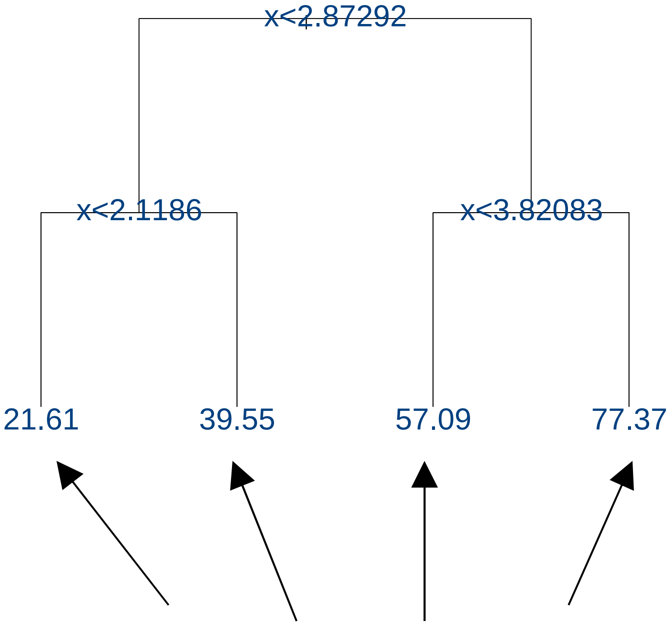
- regression trees
- boosted regression trees
- Bayesian inference/MCMC

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

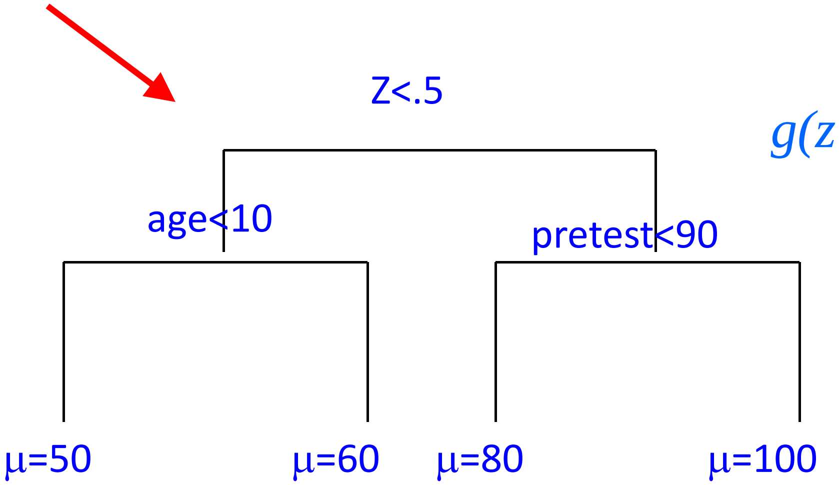
Builds on the idea of a treed model to create a “sum-of-trees” model

- 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 jth tree,

  M_j denotes the means from the terminal nodes from the jth tree,

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



$g(z=0,age=7,T_1,M_1)=50$

- So 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* to force trees to be weak learners (shrink each mean towards zero) -- these can be chosen by cross-validation (time-consuming)
- *How estimate uncertainty*? Generally, people use bootstrapping which can be cumbersome and time-consuming

How BART differs from boosting

- BART can be thought of as a stochastic alternative to boosting algorithms. It differs because:
  - $f(x,z)$ is a random variable
  - Using an MCMC algorithm, we sample *f(x,z)* it from a posterior distribution (i.e. allows for uncertainty in our model and our data). At each iteration we get a new draw of

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

  - 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

Prior distribution

T_j denotes the jth tree,
M_j denotes the means from terminal nodes of j^{th} tree

- Simplified by many independence assumptions

$$p((T_1, M_1), (T_2, M_2), \ldots, (T_m, M_m), \sigma)$$

$$= p(T_1, T_2, \ldots, T_m)\, p(M_1, M_2, \ldots, M_m \,|\, T_1, T_2, \ldots, T_m)\, p(\sigma)$$

$$p(T_1, T_2, \ldots, T_m) = \prod p(T_j),$$

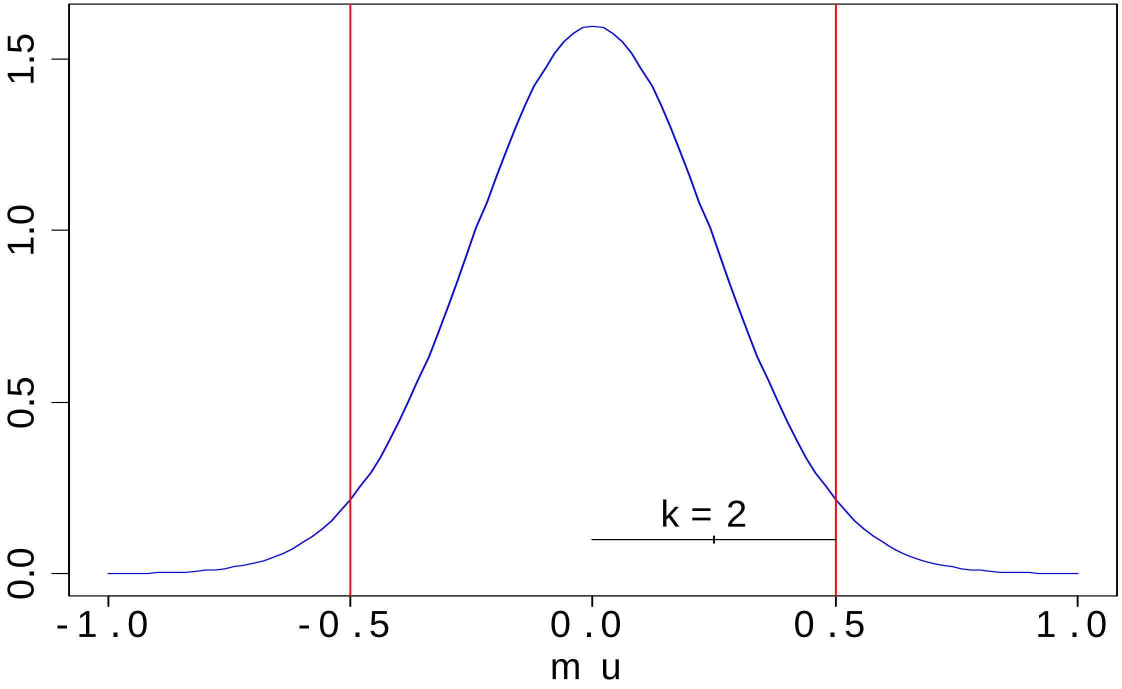
$$p(M_1, M_2, \ldots, M_m \,|\, T_1, T_2, \ldots, T_m) = \prod p(M_j \,|\, T_j),$$

$$p(M_j \,|\, T_j) = \prod p(\mu_{i,j} \,|\, T_j),$$

Prior on μ

m=number of trees

Here σ_μ^2=.25 ⟶



(i.i.d.)

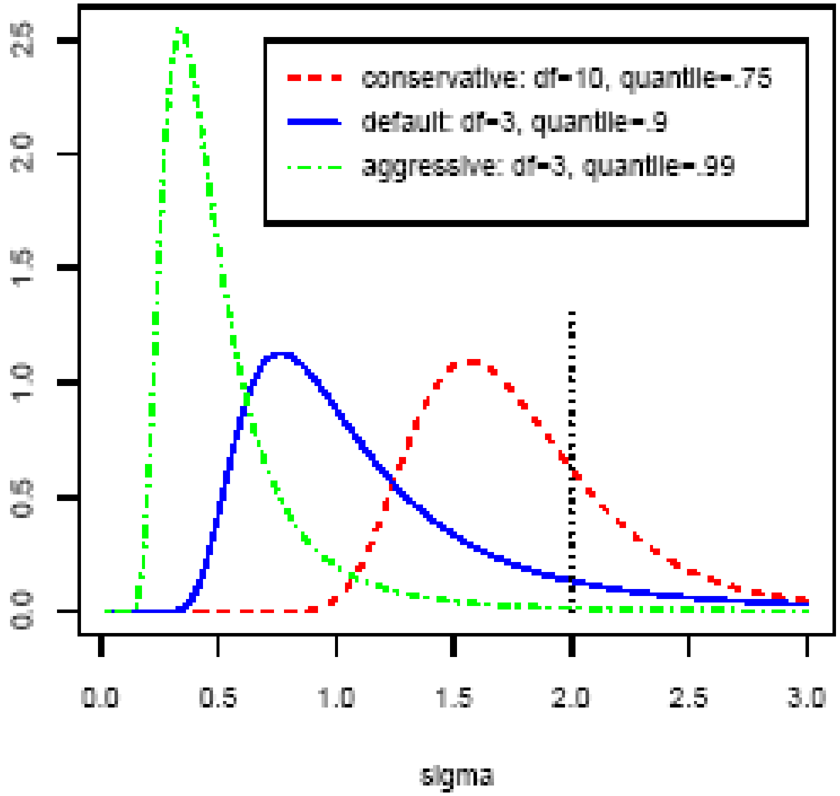
- Standardize y so that E(y | x) is in [-.5,.5] with high probability (transform all data points to lie in [-.5,.5])
- For each bottom node μ, let μ ~ N(0,σ_μ^2)
- E(y | x) is the sum of *m* independent μ's (a priori), so the prior standard deviation of E(y | x) is $\sqrt{m}\sigma_\mu$
- Thus, we choose σ_μ so that $k\sqrt{m}\sigma_\mu = .5 \Rightarrow \sigma_\mu = \frac{.5}{k\sqrt{m}}$

where k is the number of standard deviations of E(y | x) from the mean of 0 to the interval boundary of .5 (default is 2)

Prior on σ

$\hat{\sigma}$ is a rough guess at σ, eg. least squares estimate

Then choose *df* for the χ^2 distribution, and choose the quantile that $\hat{\sigma}$ represents for that distribution

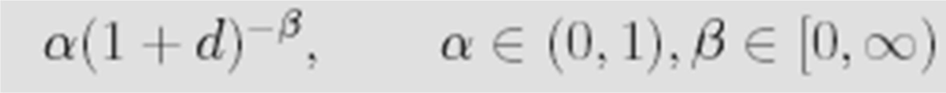


In this example

$\hat{\sigma} = 2$

Prior on the tree

- Must specify
  - Probability a node at depth d is non- terminal



$$\alpha(1+d)^{-\beta}, \qquad \alpha \in (0,1), \beta \in [0,\infty)$$

default: α=.95, β=2, encourages smaller trees:

tree size	**1**	**2**	**3**	**4**	**5**
probability	**.05**	**.55**	**.28**	**.09**	**.03**

- Distribution on the splitting variable in each non-terminal node (default: uniform) and
- Distribution on the splitting rule in each non-terminal node, conditional on the splitting variable (default: uniform)

BART: model fitting

Let $\{T_j,M_j\}$ be a set of tree models
T_j denotes the jth tree,
M_j denotes the means from the terminal nodes from the jth tree

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

The model is fit using a "simple" gibbs sampler:
(1) $\sigma \mid \{T_j\},\{M_j\}$
(2) $T_j,M_j \mid \{T_i\}_{i\neq j},\{M_i\}_{i\neq j}, \sigma$

Thousands of parameters, only σ identified
Key is prior information to keep trees and M small

Causal Inference for BART

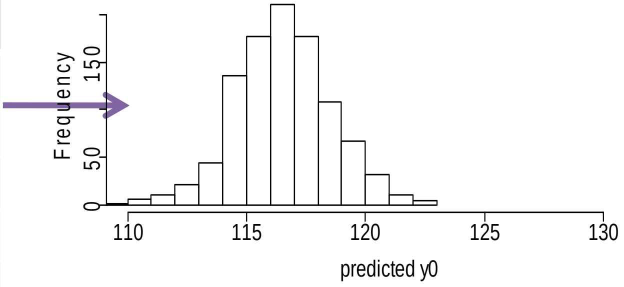
Distribution for the treatment effect created by differencing the (posterior predictive) distributions for Y(1) and Y(0)

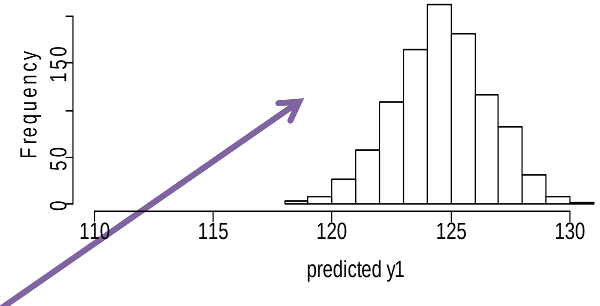
mom age	mom hs grad.	mom work	mom race	mom marital status	Z	child test score
19	1	1	B	1	0	114
22	0	1	W	0	0	92
27	0	1	B	0	0	80
23	1	0	H	1	0	98
20	1	0	H	0	1	110
25	1	1	W	1	1	82
24	0	1	B	0	1	102
⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	0	1	H	1	1	89

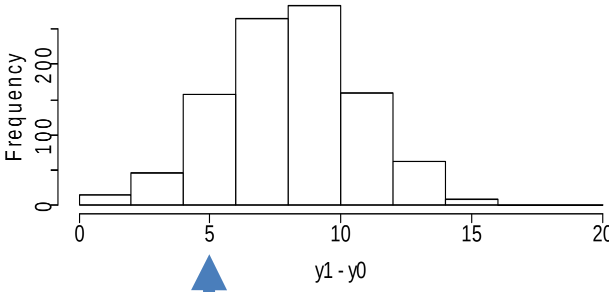
test score = f(age, hs, work, race, marital, Z) + error

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

Causal inference using 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
- It can we used to estimate average causal effects such as

$$\mathrm{E}[Y(1)|\mathrm{X=x}]- \mathrm{E}[Y(0)|\mathrm{X=x}] = f(1,x) - f(0,x)$$

- Each iteration of the BART Markov Chain generates a new draw of f from the posterior distribution. Let f^{r} denote the r^{th} draw of f. We then compute

$$d_i^{\mathrm{r}}=f^{\mathrm{r}}(1,x_i) - f^{\mathrm{r}}(0,x_i), \text{ for } i = 1, \ldots, n$$

  Averaging the d_i^{r} values over i with r fixed, the resulting values are a Monte Carlo approximation to the posterior distribution of the average treatment effect for the associated population

- For example, we average over $\{i : z_i=1\}$ if we want to estimate the effect of the treatment on the treated

$$\mathrm{E}[Y(1)|\mathrm{Z}=1]- \mathrm{E}[Y(0)|\mathrm{Z}=1]$$

Causal inference using BART

BART can be used to estimate average causal effects such as

$$E[Y(1)|X=x]- E[Y(0)|X=x] = f(1,x) - f(0,x)$$

- Fit BART (using an MCMC algorithm) to the full sample.
- Get posterior predictions for each treated unit at

  (a) the observed treatment condition and

  (b) the counterfactual condition.
- The differences between these predictions create posterior distributions for individual level treatment effects.
- Average over these to get posterior distributions for subpopulations of interest (for instance, the treated).

Evidence re BART for causal inference performance

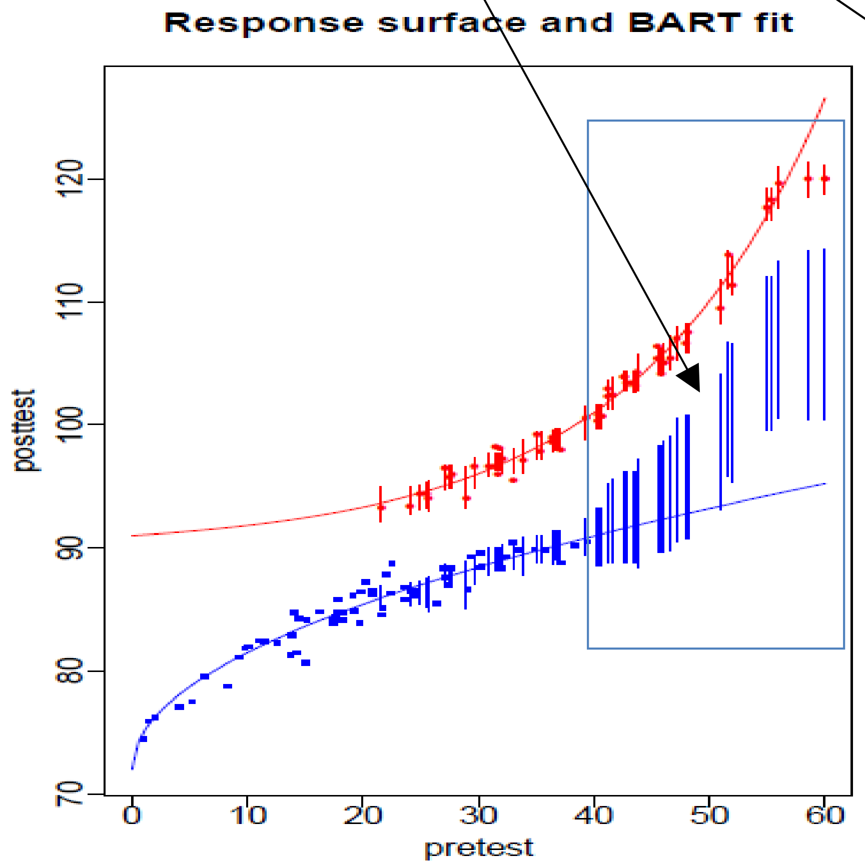
- BART has been shown to outperform propensity score approaches
  - in scenarios with nonlinear response surfaces (Hill, *JCGS, 2011)*
  - in a scenario with a many covariates (over 100) and high control/treated ratio (Hill et al., *MBR*, 2012)
- BART has been shown to outperform propensity score approaches to identifying observations that lack common causal support, thus leading to more accurate treatment effect estimates (Hill and Su, *AOAS*, 2013)
- BART can yield superior estimates of heterogeneous treatment effect estimates (Hill, *JCGS*, 2011; Green & Kern, *POQ*, 2012) and can outperform propensity score approaches to generalizing experimental treatment effects when ignorability assumptions are satisfied (Kern, Stuart, Hill, Green, *in review*, 2014)

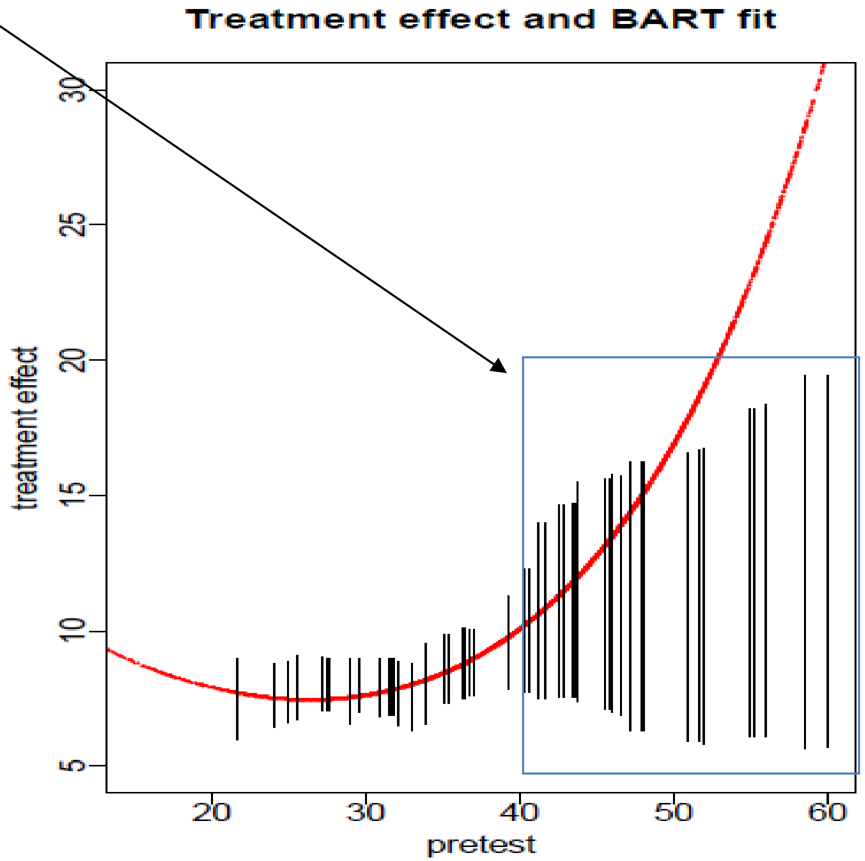
BART uncertainty increases when we lack common *causal* support

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





BART and sensitivity analysis

Sensitivity to an unobserved confounder

- What if our identification strategy (for instance propensity score matching conditioning on all observed covariates) fails to control for one important confounder, ***U***
- That is what if ignorability holds once we condition on ***U***:

$$\{Y(z)\}_{z \in \mathcal{Z}} \perp Z \mid X, \boldsymbol{U}$$

- The problem is of course that we don't know what ***U*** looks like, so we'll need to explore the impact of this potential ***U*** over a range of plausible options

Data Generating Process: Binary Z

- We factor the complete-data likelihood as

  $p(Y, Z, U, X \mid \theta) = p_1(U \mid \theta_1)\, p_2(Z \mid X, U, \theta_2)\, p_3(Y \mid X, U, Z, \theta_3)$

- However in this scenario we'd prefer to fit a more appropriate model for Z

- We proceed using the following

$$U \sim \mathrm{B}ernoulli(\pi)$$

$$Z \mid X, U \sim \mathrm{B}ernoulli\left(\Phi^{-1}(\beta^z x + \zeta^z u)\right)$$

$$\mathrm{Y} \mid X, U, Z \sim \mathrm{N}(f(z, x) + \zeta^y u, \sigma^2_{y.xuz})$$

Data Generating Process: Binary Z

- We factor the complete-data likelihood as

$$p(Y, Z, U, X \mid \theta) = p_1(U \mid \theta_1)\, p_2(Z \mid X, U, \theta_2)\, p_3(Y \mid X, U, Z, \theta_3)$$

- However in this scenario we'd prefer to fit a more appropriate model for Z

- We proceed using the following

$$U \sim Bernoulli(\pi)$$

$$Z \mid X, U \sim Bernoulli\left(\Phi^{-1}(\beta^z x + \zeta^z u)\right)$$

$$\mathrm{Y} \mid X, U, Z \sim \mathrm{N}(\boldsymbol{f}(\boldsymbol{z}, \boldsymbol{x}) + \zeta^y u, \sigma^2_{y.xuz})$$

fit using BART

How best to fit

- Incorporate BART into sensitivity analysis algorithm (poor man's Bayes)
- Incorporate latent U into BART?

Incorporate latent U into the algorithm

The model is fit using an approximate Gibbs Sampler:

(1) $\sigma \mid \{T_j\},\{M_j\}, U$

(2) $T_j, M_j \mid \{T_i\}_{i \neq j}, \{M_i\}_{i \neq j}, \sigma, U$

(3) $U \mid \{T_j\},\{M_j\}, \sigma$

Part 3 requires some parameters from the assignment mechanism as well which is fit using Maximum Likelihood, this can be generalized in the future

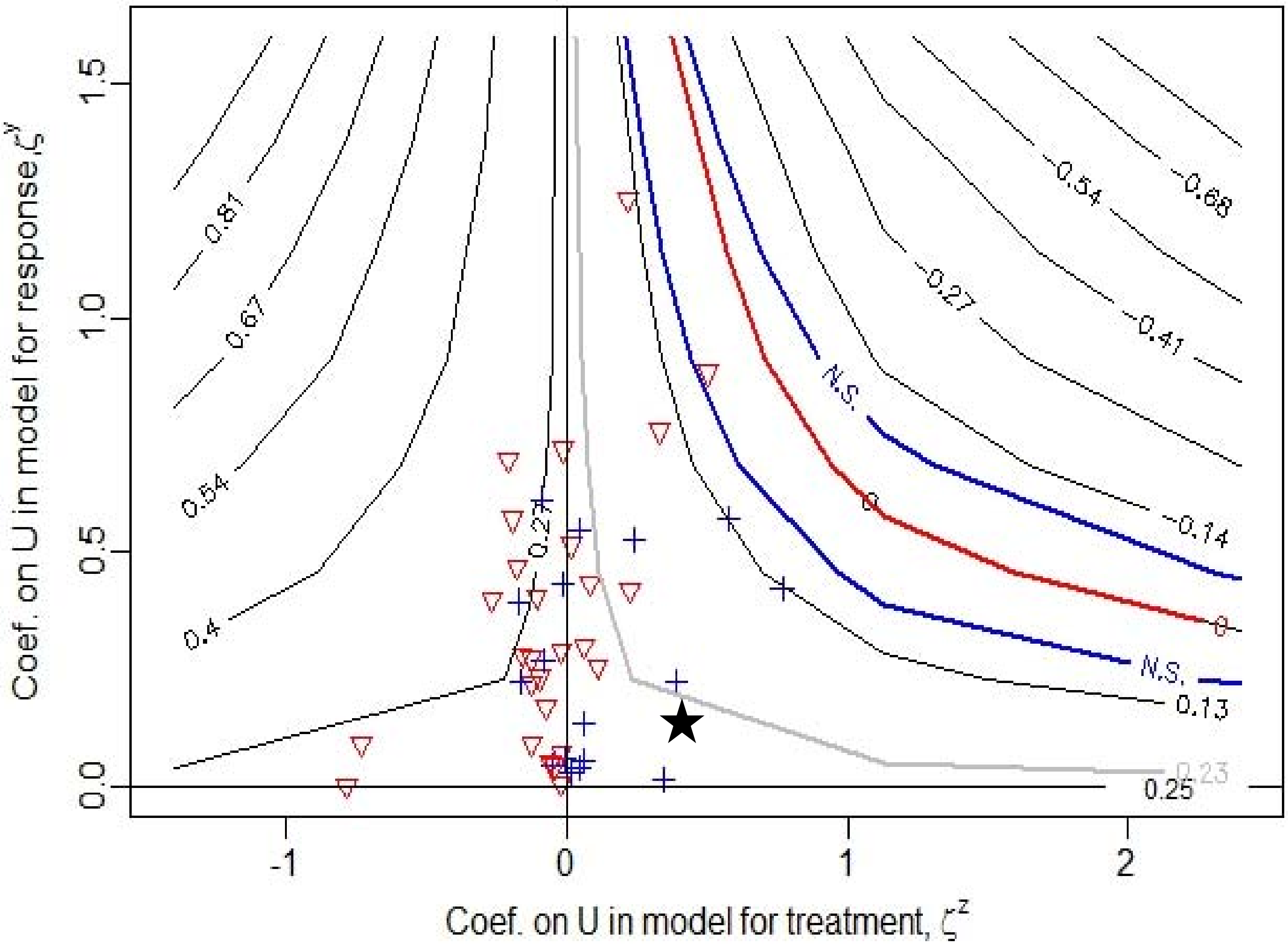
Bernoulli draw of U details

We derive Pr(U=1 | Y, Z, X) using Bayes rule:
Pr(Y, Z, U=1 | X)/Pr(Y, Z | X)

$$\begin{aligned}
\pi^{y,z,x,u=1} = {} & \left(2\pi\sigma^2_{\mathrm{y\cdot xuz}}\right)^{-1/2} \exp\left(-\frac{(Y-\zeta^y-f(X,Z))^2}{2\sigma^2_{\mathrm{y\cdot xuz}}}\right) \\
& \cdot \left(1-\Phi^{-1}(X\beta^z+\zeta^z)\right)^{(1-Z)} \left(\Phi^{-1}(X\beta^z+\zeta^z)\right)^{Z} \pi^u
\end{aligned}$$

$$\begin{aligned}
\pi^{y,z,x} = {} & \left(2\pi\sigma^2_{\mathrm{y\cdot xuz}}\right)^{-1/2} \exp\left(-\frac{(Y-f(X,Z))^2}{2\sigma^2_{\mathrm{y\cdot xuz}}}\right) \\
& \cdot \left(1-\Phi^{-1}(X\beta^z)\right)^{(1-Z)} \left(\Phi^{-1}(X\beta^z)\right)^{Z} (1-\pi^u) + \\
& \left(2\pi\sigma^2_{\mathrm{y\cdot xuz}}\right)^{-1/2} \exp\left(-\frac{(Y-\zeta^y-f(X,Z))^2}{2\sigma^2_{\mathrm{y\cdot xuz}}}\right) \\
& \cdot \left(1-\Phi^{-1}(X\beta^z+\zeta^z)\right)^{(1-Z)} \left(\Phi^{-1}(X\beta^z+\zeta^z)\right)^{Z} \pi^u
\end{aligned}$$

Sensitivity plot for effect of breastfeeding on intelligence



Joint work with John Protzko and Josh Aronson

Final thoughts

- The premise of this talk highlights the problem: Causal Inference should already, by definition, be a part of Data Science. Why hasn't this happened?
- I've illustrated some work that builds on ideas and tools across two of the core Data Science disciplines. This is one of many possible intersections.
- Imagine how much more powerful work could be that brings in additional Data Science strengths, e.g.
  - could expand to handle larger, more complex, less structured data
  - could provide better visualization tools for understanding the results (my SA plots need help!)