Quant II

Lab 9: Mediation and Moderation

Junlong Aaron Zhou

April 08, 2021

Outline

- Mediation Analysis
- Heterogeneous treatment effects

Revisit the difficulty of Mediation Analysis

- Mediator $M_i \in \{M_i(1), M_i(0)\}$
- Outcome $Y_i \in \{Y_i(D, M)\}$
- ACME: $\delta(t) = E(Y_i(t, M_i(1)) - Y_i(t, M_i(0)))$
- Mediation Effect is unnatural: we fix D_i, but manipulate the mediator M_i which should be a consequence of D_i.
- Even if we have random assignment of both D_i and M_i, we cannot identify ACME.

Imai et al. 2011

- What assumption do we need?
- **Sequential Ignorability**

$$\{Y_i(t', m), M_i(t)\} \perp D_i | X_i = x$$
$$Y_i(t', m) \perp M_i(t) | D_i = t, X_i = x$$

- First equation is guaranteed to hold in a standard experiment
- Second equation does not hold unless X_i includes all confounders

Application: Brader, Valentino, and Suhay (2008)

- Study the framing effect of media on attitude towards mmigration
- The authors identify two key factors that they hypothesize not only may alter opinions about immigration but also may spur people to political action.
- First, media messages that emphasize the costs of immigration on society should be expected to increase opposition, whereas stories that emphasize the benefits should reduce opposition.
- Second, given that immigration often has a racial component, whites will be more likely to oppose immigration when the immigrants being discussed in the media are nonwhite. Cues using nonwhite immigrants and messages emphasizing costs will have particularly negative effects on immigration attitudes.

Application: Brader, Valentino, and Suhay (2008)

- Treatment: First, the content of the news story was manipulated to emphasize the benefits or the costs of immigration. Second, the researchers varied whether the particular immigrant described and pictured was a white immigrant from Russia or a Hispanic immigrant from Mexico.
- Mediator: Anxiety
- They found negative immigration news story with a picture of a Hispanic immigrant elevated anxiety and eroded support for immigration.

Anxiety as Mediator

```
library(mediation)
data(framing)
med.fit <- lm(emo ~ treat + age + educ + gender
              + income, data = framing)
out.fit <- glm(cong_mesg ~ emo + treat + age +
                    educ + gender + income,
                data = framing, family = binomial("probit"))
# Estimation via quasi-Bayesian approximation
contcont <- mediate(med.fit, out.fit, sims=500,
                    treat="treat", mediator="emo")
```

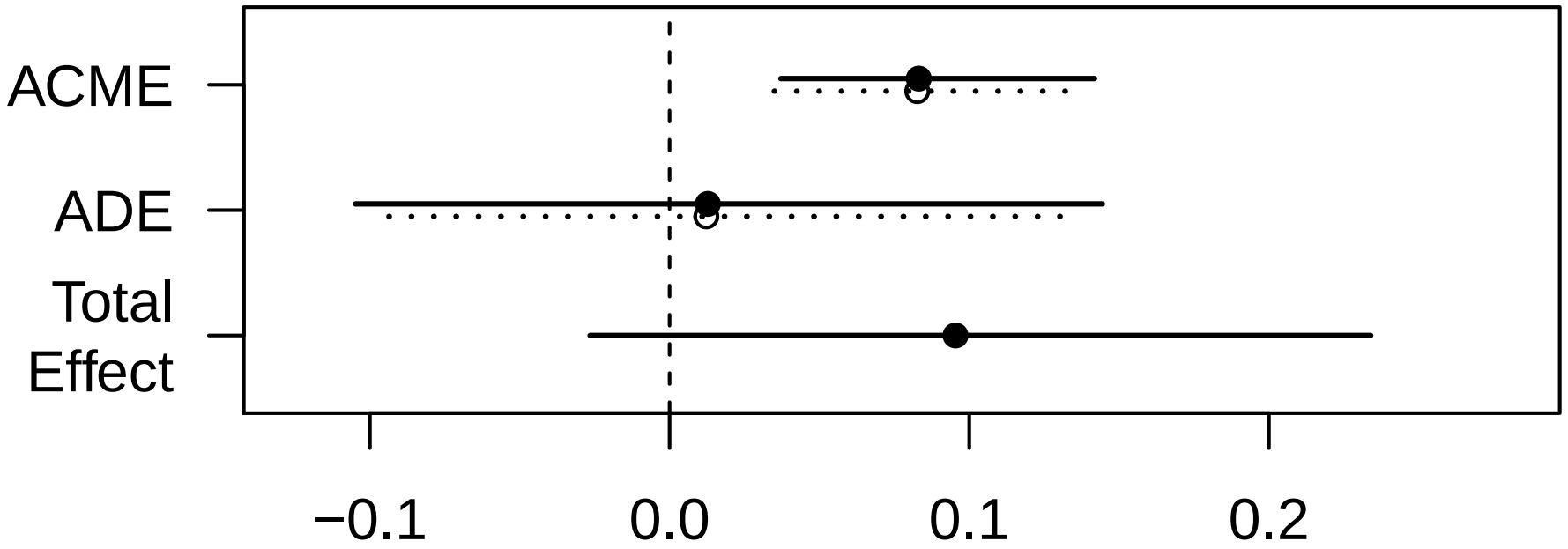
Anxiety as Mediator

```
summary(contcont)
```

```
##
## Causal Mediation Analysis
##
## Quasi-Bayesian Confidence Intervals
##
##                               Estimate 95% CI Lower 95% CI Upper
## ACME (control)                  0.0826       0.0349         0.14
## ACME (treated)                  0.0832       0.0371         0.14
## ADE (control)                   0.0123      -0.0937         0.14
## ADE (treated)                   0.0128      -0.1048         0.14
## Total Effect                    0.0954      -0.0265         0.23
## Prop. Mediated (control)        0.8331      -2.9782         7.84
## Prop. Mediated (treated)        0.8469      -2.6506         7.16
## ACME (average)                  0.0829       0.0364         0.14
## ADE (average)                   0.0125       0.0002         0.14
```

Anxiety as Mediator

```
plot(contcont)
```



Sensitivity Analysis

- The sequential ignorability is fairly strong assumption.
- Imai et. al propose a sensitivity analysis

Sensitivity Analysis

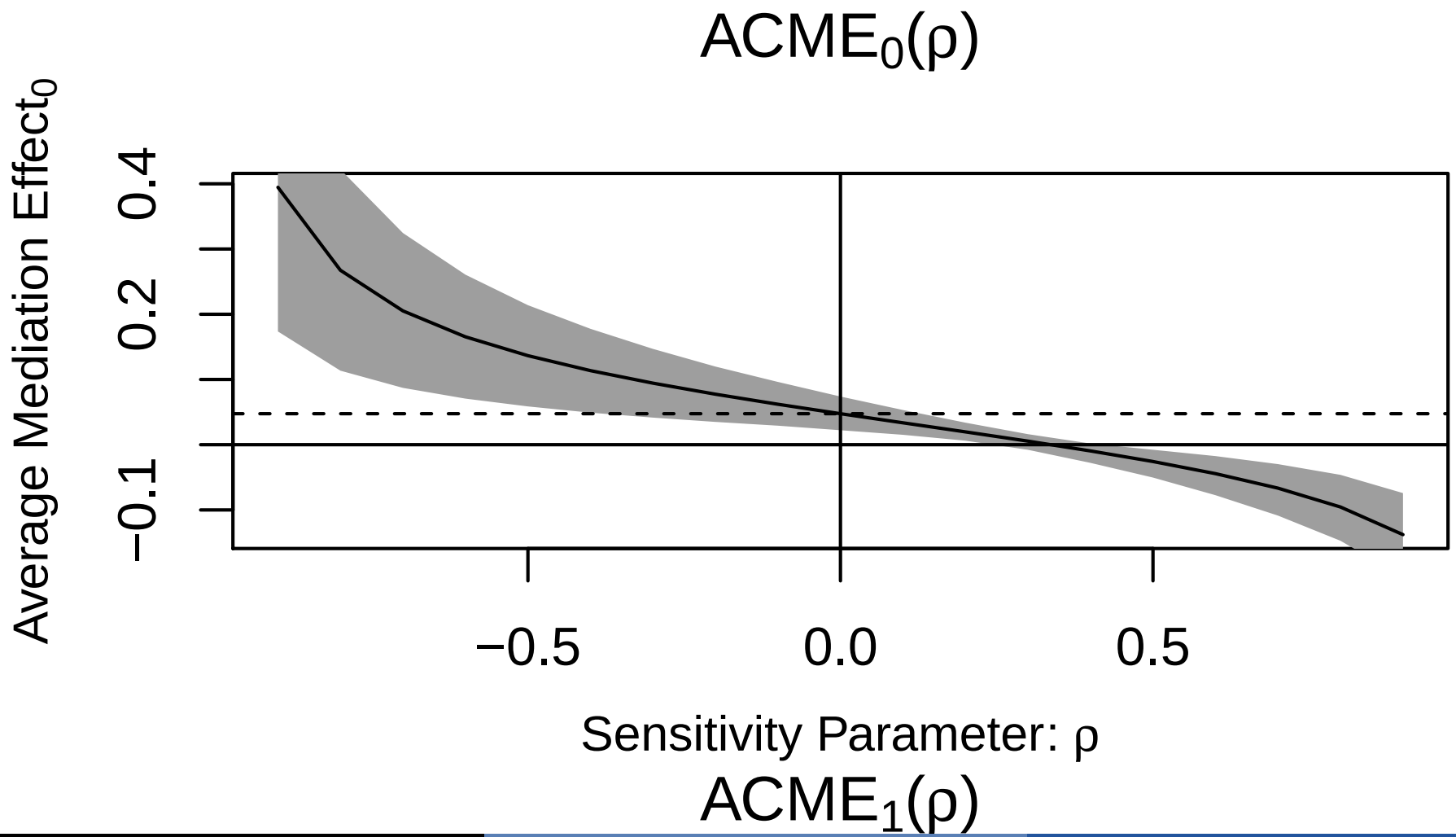
- The sequential ignorability is fairly strong assumption.
- Imai et. al propose a sensitivity analysis
- Recall this two-stage linear regression

$$M_i = \alpha_1 + \beta_1 D_i + \beta_2 X_i + \epsilon_{i1}$$
$$Y_i = \alpha_1 + \beta_3 D_i + \beta_5 M_i + \beta_4 X_i + \epsilon_{i2}$$

- Sequential ignorability implies $\rho = corr(\epsilon_{i1}, \epsilon_{i2}) = 0$.
- We use this parameter to conduct a sensitivity analysis.
- However, even if $\rho = 0$, we cannot justify this assumption.

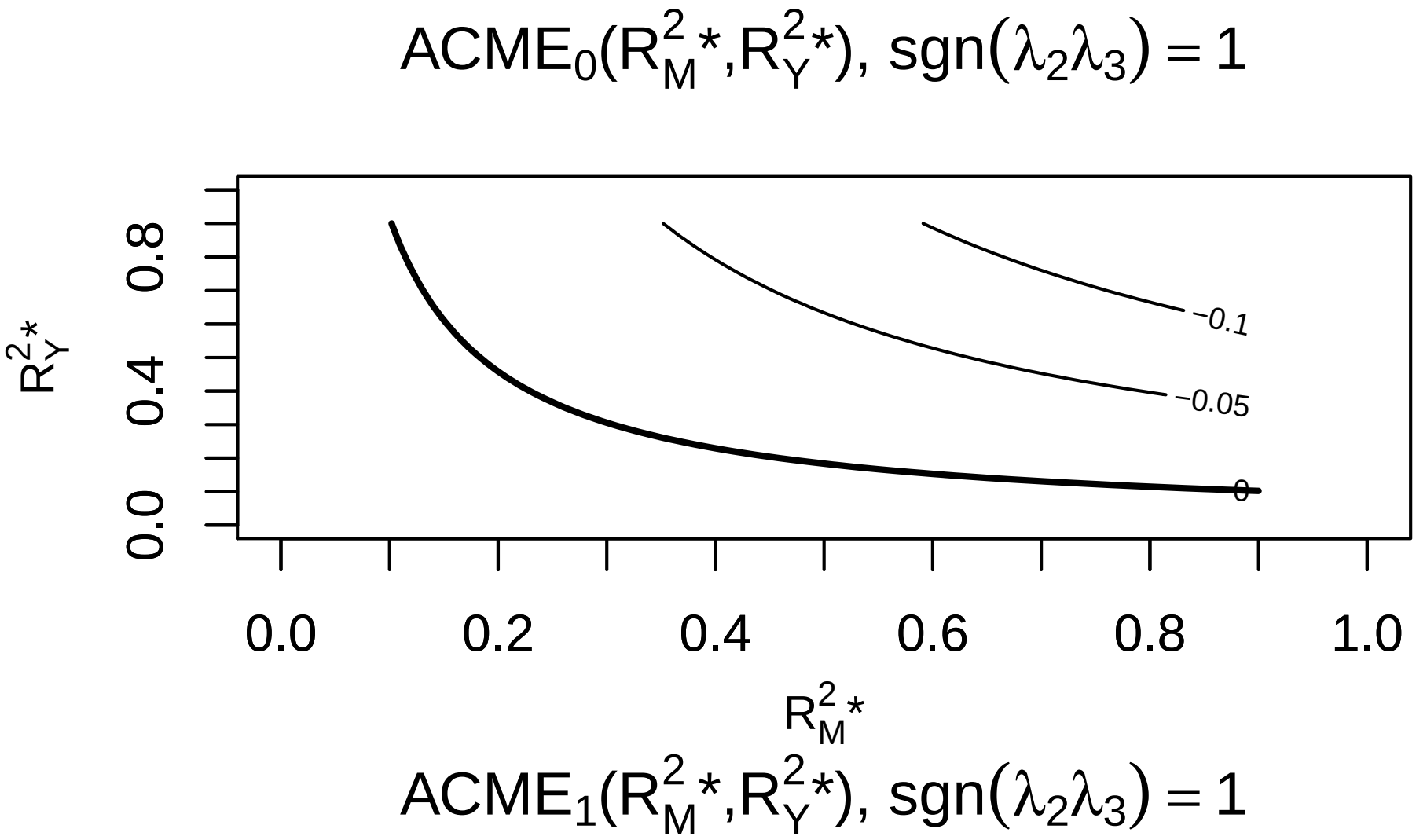
Sensitivity Analysis

```
s <- medsens(contcont)
plot(s)
```



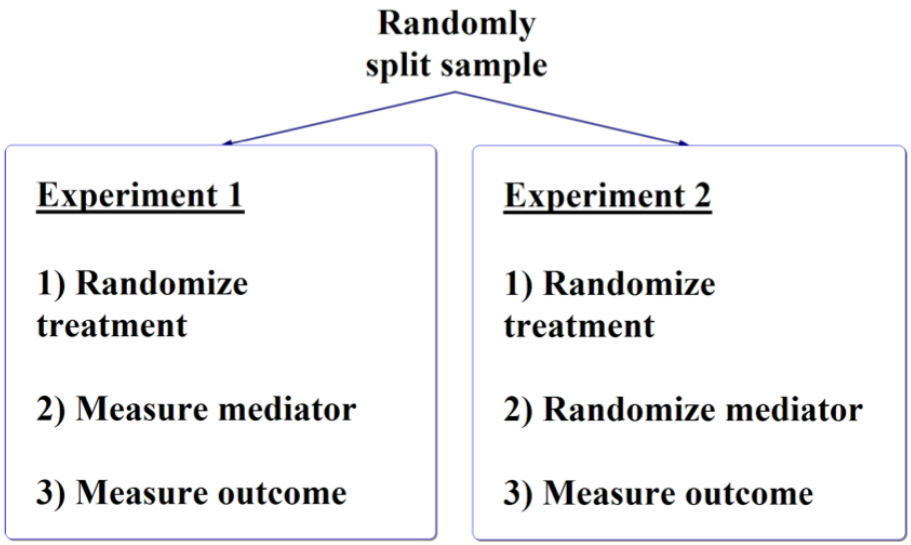
Sensitivity Analysis

```
plot(s,sens.par="R2")
```



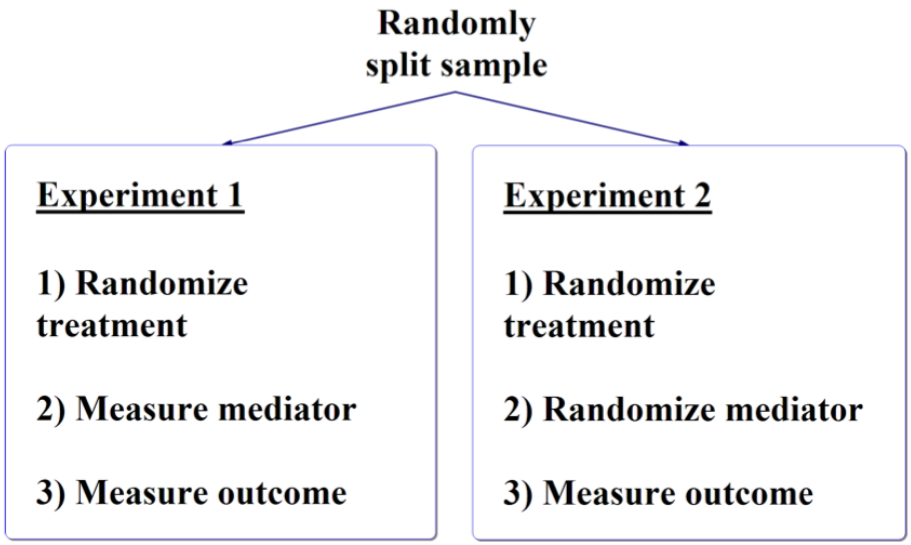
Experimental Design

- Allow us to relax sequential ignorability assumption.
- Parallel Design



Experimental Design

- Allow us to relax sequential ignorability assumption.
- Parallel Design



- Must assume no direct effect of manipulation on outcome
- More informative than standard single experiment
- If we assume no interaction between D and M, ACME is point identified

Parallel Design

- In the first experiment, we can estimate the effect of D on M.
- In the second experiment, we know $Y(D, M)$
- We plug-in the estimates from the first experiment to second experiment, we get $Y(D, \hat{M(d)})$.

Parallel Design

- In the first experiment, we can estimate the effect of D on M.
- In the second experiment, we know $Y(D, M)$
- We plug-in the estimates from the first experiment to second experiment, we get $Y(D, M\hat{(}d))$.
- Similarly, parallel encouragement design (encourage to change mediator).

Crossover Design

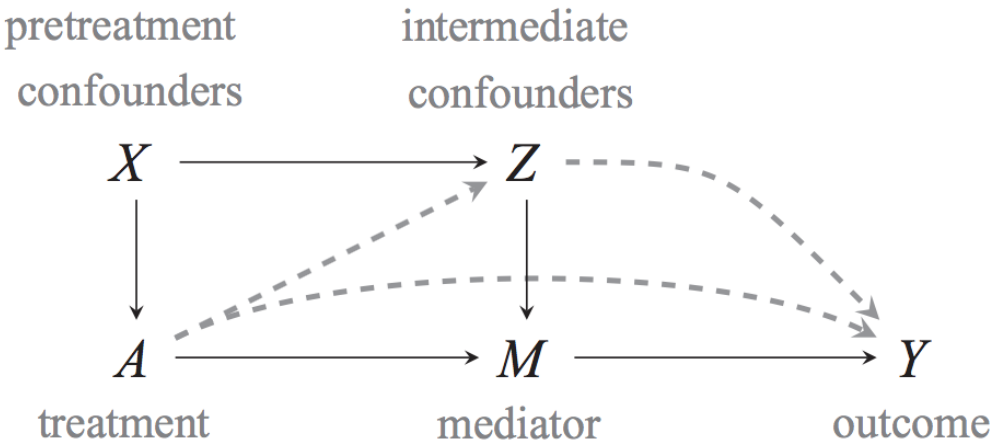
- Recall ACME can be identified if we observe $Y_i(t', M_i(t))$
- Get $M_i(t)$, then switch D_i to t' while holding $M_i = M_i(t)$

Crossover Design

- Recall ACME can be identified if we observe $Y_i(t', M_i(t))$
- Get $M_i(t)$, then switch D_i to t' while holding $M_i = M_i(t)$
- Crossover design:
  - Round 1: Conduct a standard experiment
  - Round 2: Change the treatment to the opposite status but fix the mediator to the value observed in the first round
- Very powerful: identifies mediation effects for each subject
- Must assume no carryover effect: Round 1 must not affect Round 2
- Can be made plausible by design

ACDE: average controlled direct effect

- Acharya et al. 2016 looks at ACDE: $E[Y_i(d,0) - Yi(0,0)|X_i = x]$.
- We control mediator at a fixed value (for example 0), and look at the direct effect of treatment D.
- This setting allows posttreatment confounder for mediator Z



Basic Idea of Estimating ACDE:

- Assumption 1: Sequential Unconfoundedness
  - First, no omitted variables for the effect of treatment (D_i) on the outcome (Yi), conditional on the pretreatment confounders (X_i). Second, no omitted variables for the effect of the mediator on the outcome, conditional on the treatment, pretreatment confounders, and intermediate confounders (Z_i).

Basic Idea of Estimating ACDE:

- Assumption 1: Sequential Unconfoundedness
  - First, no omitted variables for the effect of treatment (D_i) on the outcome (Yi), conditional on the pretreatment confounders (X_i). Second, no omitted variables for the effect of the mediator on the outcome, conditional on the treatment, pretreatment confounders, and intermediate confounders (Z_i).
- Assumption 2: No intermediate interactions
  - The effect of the mediator (M_i) on the outcome (Y_i) is independent of the intermediate confounders (Z_i).

Basic Idea of Estimating ACDE:

- Assumption 1: Sequential Unconfoundedness
  - First, no omitted variables for the effect of treatment (D_i) on the outcome (Yi), conditional on the pretreatment confounders (X_i). Second, no omitted variables for the effect of the mediator on the outcome, conditional on the treatment, pretreatment confounders, and intermediate confounders (Z_i).
- Assumption 2: No intermediate interactions
  - The effect of the mediator (M_i) on the outcome (Y_i) is independent of the intermediate confounders (Z_i).
- They adopts the idea of g-estimation (package: `DirectEffects`).
  - Estimate the controlled mediation effect.
    $\gamma(t, m, x) = E[Y_i(t, m) - Yi(t, 0)|X_i = x]$
  - Demediate that effect on the outcome: $\tilde{Y}_i = Y_i - \gamma(t, m, x)$
  - Estimate the effect of treatment: $\tilde{Y}_i = \beta D_i + \omega X_i$

Heterogeneous treatment effects

- We assume that all the observations have been properly weighted so no need to control for confounders.
- Group-mean-difference gives you an unbiased and consistent estimate of the ATE, $\tau = \frac{1}{N}\sum_{i=1}^{N}\tau_i$.

Heterogeneous treatment effects

- We assume that all the observations have been properly weighted so no need to control for confounders.
- Group-mean-difference gives you an unbiased and consistent estimate of the ATE, $\tau = \frac{1}{N}\sum_{i=1}^{N}\tau_i$.
- It is impossible to identify each τ_i, but we want to know more about their distribution.

Heterogeneous treatment effects

- We assume that all the observations have been properly weighted so no need to control for confounders.
- Group-mean-difference gives you an unbiased and consistent estimate of the ATE, $\tau = \frac{1}{N}\sum_{i=1}^{N}\tau_i$.
- It is impossible to identify each τ_i, but we want to know more about their distribution.
- There are many reasons for doing so:
  - Test your theory
  - Generalize the findings
  - Design better studies next time
- A common idea is assume $\tau_i = f(\mathbf{X}_i) + \varepsilon_i$ where f may be unknown.

Heterogeneous treatment effects

- What is not heterogeneous treatment effect?
- Recall $\tau_i = f(\mathbf{X}_i) + \varepsilon_i$, so we can get $E(\tau_i|X_i)$
- It is not $E(\tau_i(X_i))$ (or $E(\tau_i|do(X_i))$)
- We do not know the causal effect of X_i.
- Notice that it is a prediction problem that has nothing to do with causality.

Heterogeneous treatment effects

- Common approach is to add an interaction term.
- It suffers the potential problem of model misspecification.
- Simple interaction term assumes homogenous linear effect.
- It could be nonlinear (quardic, etc.) or heterogeneous (moderated by other variables). (Hainmueller et. al 2019; Blackwell and Olson, forthcoming)

Heterogeneous treatment effects

- Common approach is to add an interaction term.
- It suffers the potential problem of model misspecification.
- Simple interaction term assumes homogenous linear effect.
- It could be nonlinear (quardic, etc.) or heterogeneous (moderated by other variables). (Hainmueller et. al 2019; Blackwell and Olson, forthcoming)
- Since it's a prediction problem, machine learning can help!

A detour to the basic idea of machine learning

- Machine learning provides you with a variety of prediction tools.
- We assume that the form of f is unknown and try to minimize a loss function l.
- For example, the quadratic loss function $l = E[Y_i - f(\mathbf{X}_i)]^2$.
- Given the sample, what is the best prediction function f?

A detour to the basic idea of machine learning

- Machine learning provides you with a variety of prediction tools.
- We assume that the form of f is unknown and try to minimize a loss function l.
- For example, the quadratic loss function $l = E[Y_i - f(\mathbf{X}_i)]^2$.
- Given the sample, what is the best prediction function f?
- Just set $\hat{f}(\mathbf{X}_i) = Y_i$ and it is a perfect fit.
- Yet if a new sample is drawn from the same DGP, its prediction performance will be terrible.

A detour to the basic idea of machine learning

- Machine learning provides you with a variety of prediction tools.
- We assume that the form of f is unknown and try to minimize a loss function l.
- For example, the quadratic loss function $l = E[Y_i - f(\mathbf{X}_i)]^2$.
- Given the sample, what is the best prediction function f?
- Just set $\hat{f}(\mathbf{X}_i) = Y_i$ and it is a perfect fit.
- Yet if a new sample is drawn from the same DGP, its prediction performance will be terrible.
- You are fitting the noise ε_i rather than the relationship of interest.

A detour to the basic idea of machine learning

- A shift in perspective: we want to minimize the loss in prediction rather than in approximation.
- We want l to be as small as possible for a new random sample drawn from the same DGP.

A detour to the basic idea of machine learning

- A shift in perspective: we want to minimize the loss in prediction rather than in approximation.
- We want I to be as small as possible for a new random sample drawn from the same DGP.
- A natural idea is to split the current sample, using half to estimate the model and the other half to test its performance: training set vs. test set.
- We select the model that does the best job on the test set.
- How to split? How to test? How to select?

A simple example

- Consider the linear regression models with Y, D, and $\mathbf{X}$.
- We assume that strong ignorability holds, but are unsure what moderators to include in the regression.
- Should we control for all the higher order terms and interaction terms of the moderators?

A simple example

- Consider the linear regression models with Y, D, and $\mathbf{X}$.
- We assume that strong ignorability holds, but are unsure what moderators to include in the regression.
- Should we control for all the higher order terms and interaction terms of the moderators?
- We do not want the model to be too complicated.
- Therefore, we penalize the number of moderators in the regression and modify the loss function to be:

$$I = E[Y_i - \mathbf{X}_i\beta]^2 + \lambda \sum_{i=1}^{N} ||\mathbf{X}_i||$$

A simple example

- For each possible λ (from 1 to 100, for example), we split the sample into five folds.
- We use four folds to minimize the loss under the given λ, and apply the fitted model to calculate the value of I on the remaining fold.
- We then select the λ that gives us the smallest I on the test set.

A simple example

- For each possible λ (from 1 to 100, for example), we split the sample into five folds.
- We use four folds to minimize the loss under the given λ, and apply the fitted model to calculate the value of I on the remaining fold.
- We then select the λ that gives us the smallest I on the test set.
- It is called cross-validation.
- The idea is that the error term ε_i in the training set is independent to that in the test set.
- Hence, a model that fits the noise on the training set won't do well on the test set.
- The chosen $\hat{f}$ should be roughly uncorrelated to the noise.

A simple example

- Result from the algorithm above is called LASSO (Least Absolute Shrinkage and Selection Operator).
- It selects "features" that have the strongest prediction power into the model.
- Compared to the result of OLS, the coefficients of strong predictors are larger and the coefficients of weak predictors become zero.
- We will have a clearer picture of what moderators better explain the variation of the treatment effect.

From tree to forest

- LASSO still assumes linearity.

From tree to forest

- LASSO still assumes linearity.
- A more flexible method is called tree, or CART (Classification and Regression Tree).
- The tree algorithm automatically classifies observations into homogeneous blocks.

From tree to forest

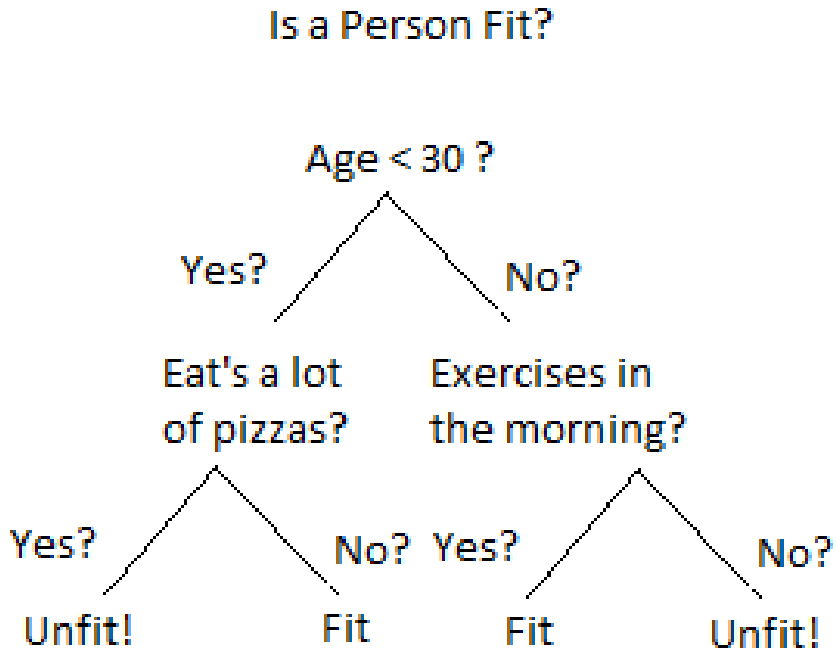
- LASSO still assumes linearity.
- A more flexible method is called tree, or CART (Classification and Regression Tree).
- The tree algorithm automatically classifies observations into homogeneous blocks.
- In the pure prediction case, we have only Y and $\mathbf{X}$.
- The estimate for each observation i is the average outcome in the block b it belongs to.
- Again, we want to balance fitness and complexity:

$$I = E[Y_i - \hat{Y}_i]^2 + \lambda|B| = E_b[E[Y_i - \bar{Y}_b]^2] + \lambda|B|$$

where B is the number of blocks.

From tree to forest

- We rely on a recursive algorithm to obtain the best partition.
- In each step, we find the "cut" that increases fitness the most.



- The algorithm proceeds on the training set until we have no more than 5 observations in each leave.
- Now we have maximized fitness.

From tree to forest

- To reduce complexity, we "prune" the tree reversely with the given λ.
- Again, cross-validation on the test set selects the λ that minimizes the loss.

From tree to forest

- To reduce complexity, we "prune" the tree reversely with the given λ.
- Again, cross-validation on the test set selects the λ that minimizes the loss.
- Recusive algorithm finds the optimum at each step, but the result may not be a global optimum as well.
- Consequently, tree is highly unstable.

From tree to forest

- To reduce complexity, we "prune" the tree reversely with the given λ.
- Again, cross-validation on the test set selects the λ that minimizes the loss.
- Recusive algorithm finds the optimum at each step, but the result may not be a global optimum as well.
- Consequently, tree is highly unstable.
- We may use the evolutionary algorithm to find the optimal tree partition.
- Or we average over a lot of trees— boosting or forest.

Causal tree

- To study HTE, we can divide observations into the blocks to optimize our prediction of the treatment effect.
- There is a problem: τ_i is not observable.

Causal tree

- To study HTE, we can divide observations into the blocks to optimize our prediction of the treatment effect.
- There is a problem: τ_i is not observable.
- Athey and Imbens (2016): Causal tree.
- Intuitively, we reward the heterogeneity across leaves and penalize the heterogeneity within leaves.

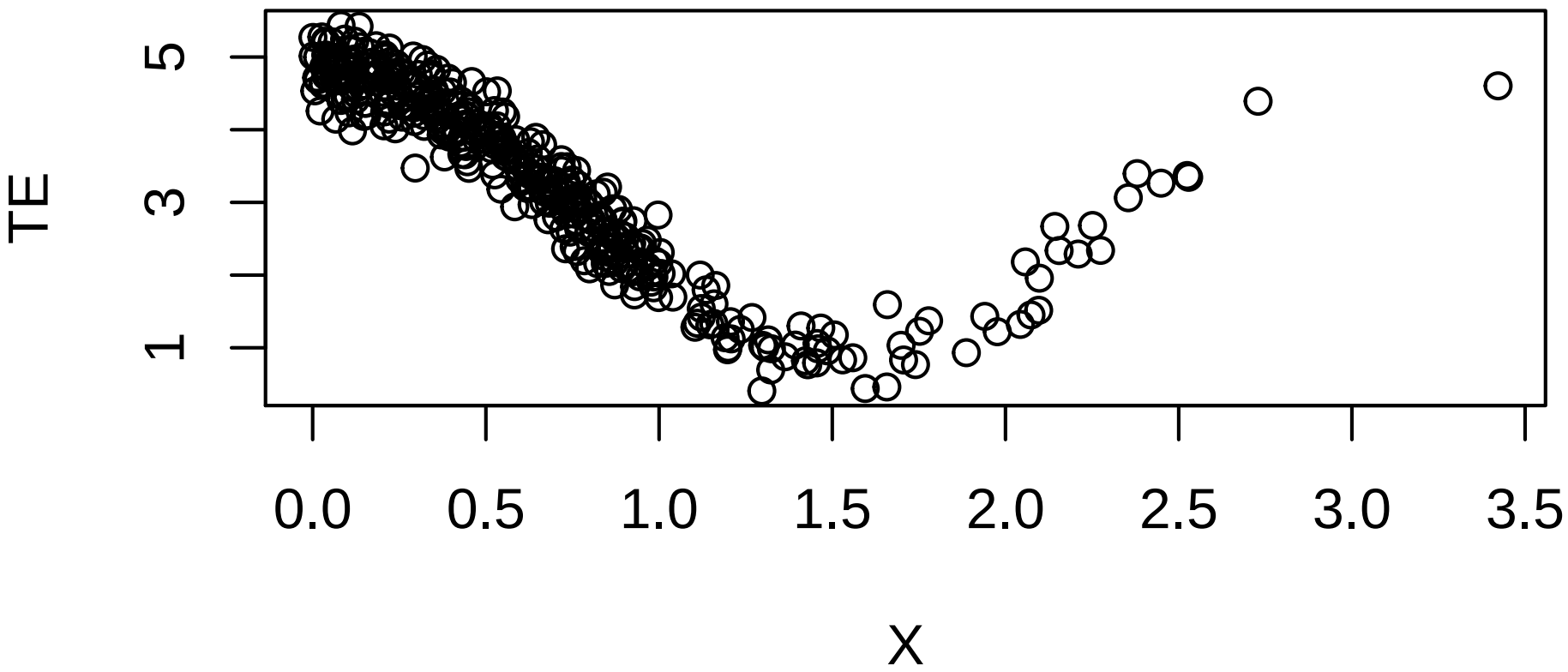
Causal tree

- To study HTE, we can divide observations into the blocks to optimize our prediction of the treatment effect.
- There is a problem: τ_i is not observable.
- Athey and Imbens (2016): Causal tree.
- Intuitively, we reward the heterogeneity across leaves and penalize the heterogeneity within leaves.
- They also suggest an "honest approach."
- Split the sample into three parts: the training set, the test set, and the estimation set.
- We use the first two sets to generate the optimal partition, and the last to estimate effects on each leave.
- It reduces bias and makes inference much easier.

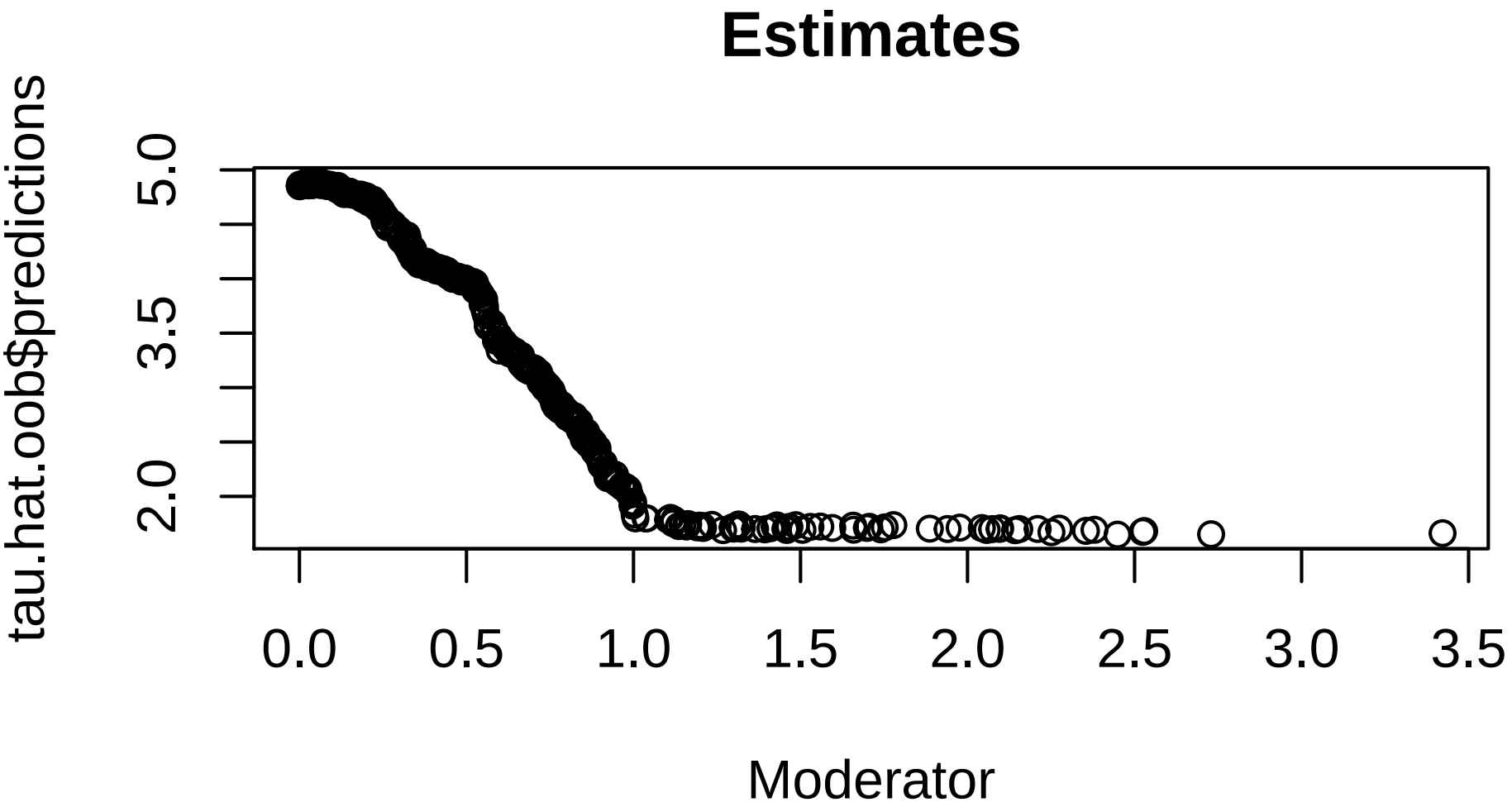
Causal forest

- We have the causal forest when combining results from various causal trees.
- Wager and Athey (2018) and Athey et al. (2019) establish the theory behind the algorithm.
- Forest also avoids lowering efficiency.
- The idea can be extended to estimate any local parameter (generalized causal forest).

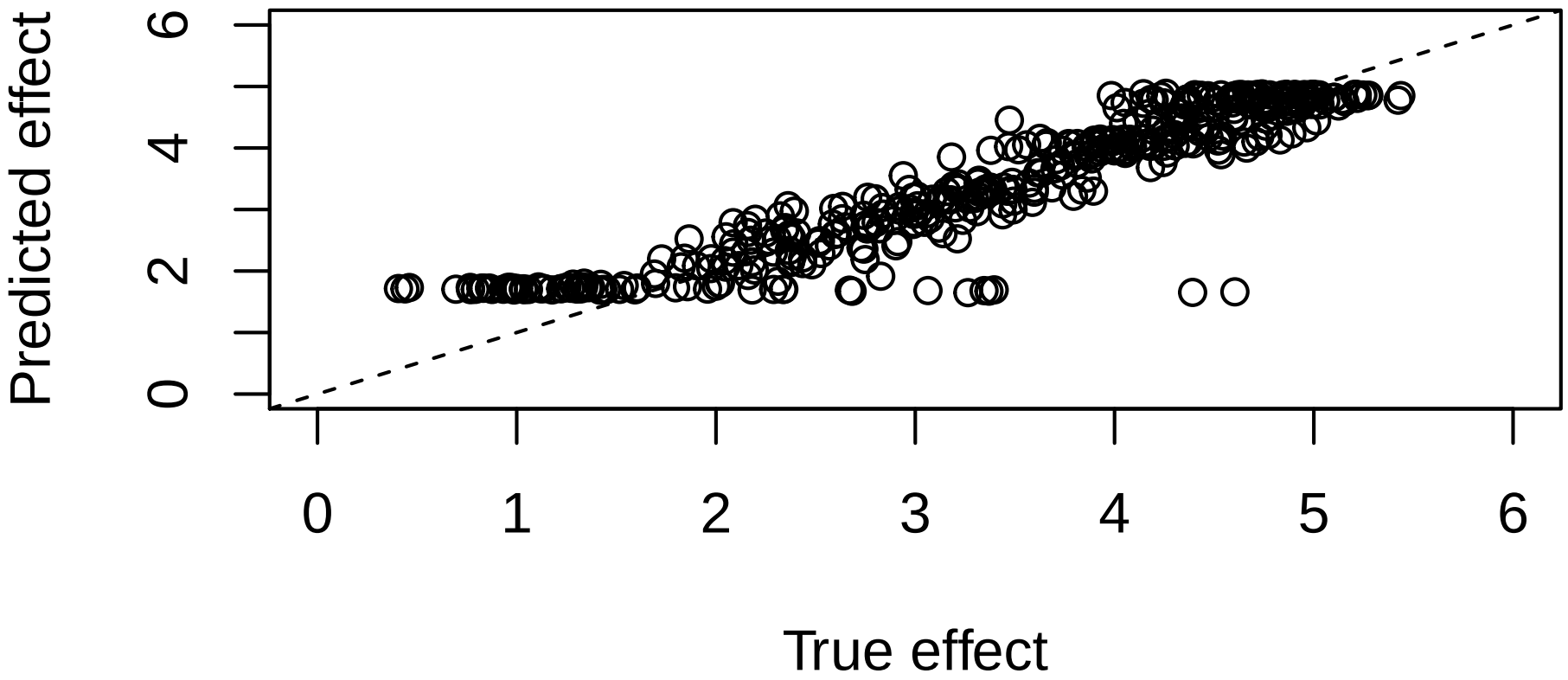
Causal forest



Causal forest



Causal forest



Ensemble methods

- Whether linear methods or tree-based methods perform better depends on the context.

Ensemble methods

- Whether linear methods or tree-based methods perform better depends on the context.
- A natural idea is to combine them together.
- Grimmer (2017):
  - Predict Y_i using M different methods and generate predicted values $(\hat{Y}_{i1}, \hat{Y}_{i2}, \ldots, \hat{Y}_{iM})$.
  - Regress Y_i on the M predictions to get their weights.
  - The ATE estimate equals to the weighted average of ATE estimates from each method.

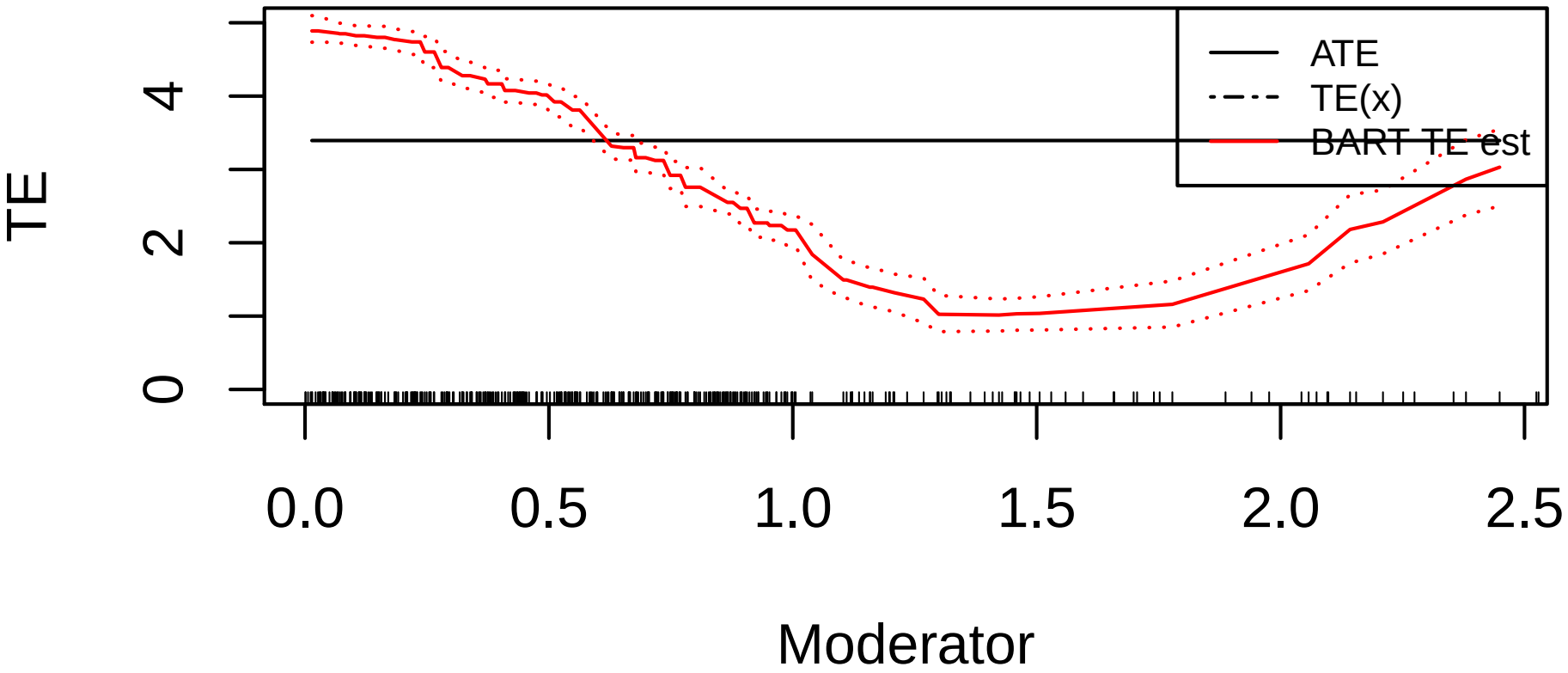
Ensemble methods

- More general frameworks:
  - X-learner in Kunzel et al. (2019);
  - R-learner in Nie and Wager (2019);
  - TMLE in Van der Laan and Ross (2011);
  - Double machine learning in Belloni et al. (2017).

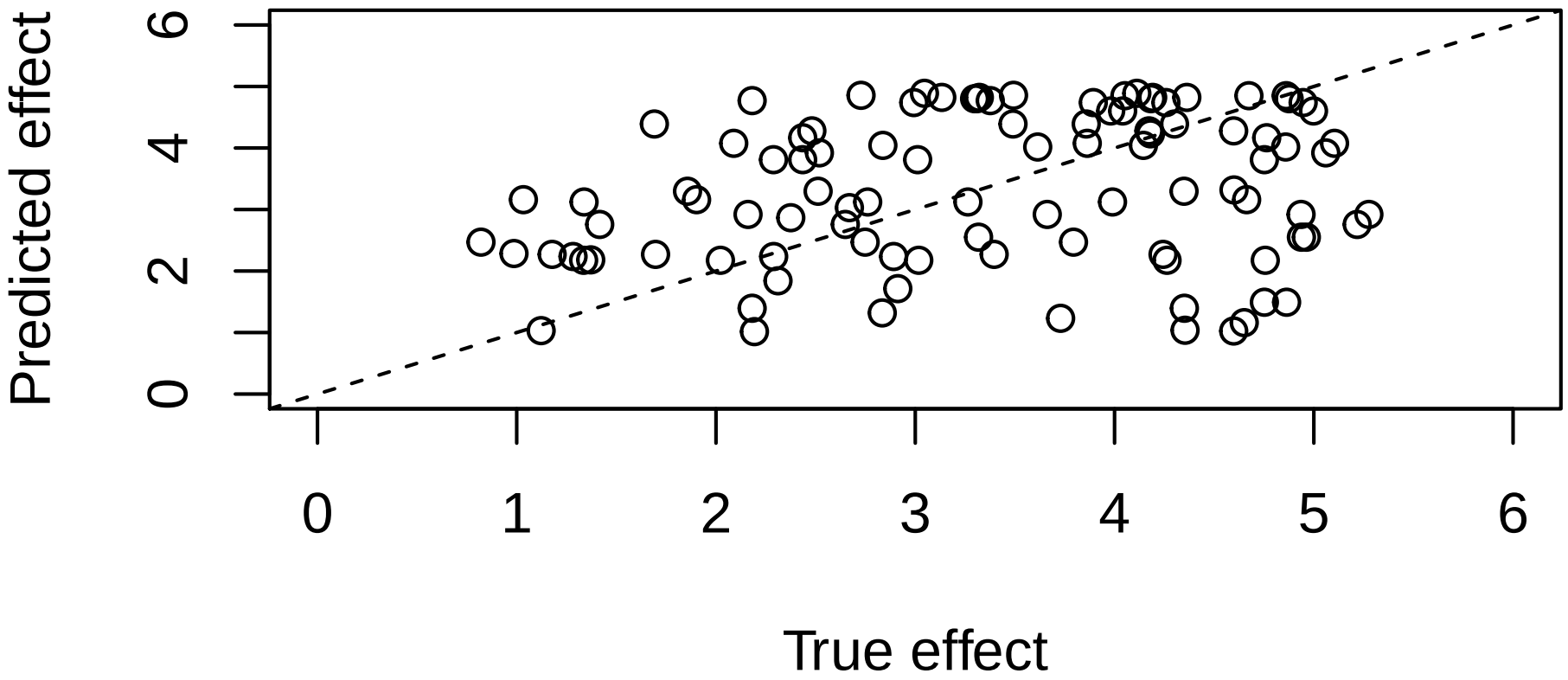
T-learner vs. S-learner

- All the previous methods run on the whole sample (S-learner).
- We can fit separate models for the treatment group and control group (T-learner).
- A popular approach is BART (Bayesian Additive Regression Trees, an S-learner).
- It is a Bayesian approach and the CIs are generated from the posterior.
- The T-learner is less stable as two response surfaces are fitted separately.

T-learner vs. S-learner



T-learner vs. S-learner



Structural estimation of the HTE

- Modern approaches remain agnostic about the source of heterogeneity.
- But we may add more "structures" into the model.
- For example, people may select into the program because they know its benefits.
- See the work of Heckman and Vytlacil, and the recent review by Xie and Zhou.
- Some simple structures may significantly enhance the model's performance.

Optimal Treatment Regime

- Once we know $E(\tau_i) = f(X_i)$, we may want to target the optimal policy.
- We define a policy $\pi : \mathbb{R}^p \to \{0, 1\}$: from your characteristics, I assign you to treatment group or control group.
- An intuitive idea: assign $D = 1$ if $\hat{\tau}_i > 0$: $\pi(x_i) = 1\{\hat{\tau}_i > 0\}$

Optimal Treatment Regime

- Once we know $E(\tau_i) = f(X_i)$, we may want to target the optimal policy.
- We define a policy $\pi : \mathbb{R}^p \to \{0, 1\}$: from your characteristics, I assign you to treatment group or control group.
- An intuitive idea: assign $D = 1$ if $\hat{\tau}_i > 0$: $\pi(x_i) = 1\{\hat{\tau}_i > 0\}$
- This idea is tempting but:
  - $\hat{\tau}_i$ contains error (or $f \in F$ is finite).
  - $\hat{\tau}_i$ comes from the training data.
  - The policy space $\pi \in \Pi$ is finite.
- Various literature works on the statistical property of optimal policy assignment. (Dehejia 2005, Kitagawa and Tetenov 2018, Athey and Wager,2021)

Optimal Treatment Regime

- Once we know $E(\tau_i) = f(X_i)$, we may want to target the optimal policy.
- We define a policy $\pi : \mathbb{R}^p \to \{0, 1\}$: from your characteristics, I assign you to treatment group or control group.
- An intuitive idea: assign $D = 1$ if $\hat{\tau}_i > 0$: $\pi(x_i) = 1\{\hat{\tau}_i > 0\}$
- This idea is tempting but:
  - $\hat{\tau}_i$ contains error (or $f \in F$ is finite).
  - $\hat{\tau}_i$ comes from the training data.
  - The policy space $\pi \in \Pi$ is finite.
- Various literature works on the statistical property of optimal policy assignment. (Dehejia 2005, Kitagawa and Tetenov 2018, Athey and Wager,2021)
- Furthermore, offline learning v.s. online learning.
- The latter refers to keep getting new information with new intervention.

Reference

- Mediation Analysis:
  - *A series of paper* by Kosuke Imai and his coauthors
  - Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." American Political Science Review 110 (3): 512–29.
- Heterogeneous Treatment Effect:
  - Green, Donald P., and Holger L. Kern. "Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees." Public opinion quarterly 76.3 (2012)
  - Athey, Susan, and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." Proceedings of the National Academy of Sciences 113.27 (2016)
  - Wager, Stefan, and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." Journal of the American Statistical Association 113.523 (2018)
  - Kitagawa, Toru, and Aleksey Tetenov. "Who should be treated? empirical welfare maximization methods for treatment choice." Econometrica 86.2 (2018)