Estimating Causal Effect Heterogeneity Using Machine Learning Techniques with Social Science Applications

Ye Wang
New York University

PolyU, June 27 2018

Outlines

- What is treatment effect heterogeneity
- Why is regression wrong
- How can machine learning help
- Linear methods and applications
- Tree-based methods and applications

The potential outcome framework

Suppose there is an outcome of interest, Y_i. We assume that

$$Y_i = \begin{cases} Y_i(1) \text{ if } W_i = 1 \\ Y_i(0) \text{ if } W_i = 0 \end{cases}$$

where W_i is the "treatment" assigned to subject i.

The potential outcome framework

Suppose there is an outcome of interest, Y_i. We assume that

$$Y_i = \begin{cases} Y_i(1) \text{ if } W_i = 1 \\ Y_i(0) \text{ if } W_i = 0 \end{cases}$$

where W_i is the "treatment" assigned to subject i.

For each subject, the idiosyncratic treatment effect is defined as

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

The potential outcome framework

Suppose there is an outcome of interest, Y_i. We assume that

$$Y_i = \begin{cases} Y_i(1) \text{ if } W_i = 1 \\ Y_i(0) \text{ if } W_i = 0 \end{cases}$$

where W_i is the "treatment" assigned to subject i.

For each subject, the idiosyncratic treatment effect is defined as

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

Traditionally, we care about the average treatment effect (ATE): $\tau_{ATE} = \mathrm{E}\left[Y_i(1) - Y_i(0)\right]$, or the average treatment effect on the treated (ATT): $\tau_{ATT} = \mathrm{E}\left[Y_i(1) - Y_i(0) \mid W_i = 1\right]$.

Treatment effect heterogeneity

Today's empirical researchers want to know not only the average of τ_i, but also its distribution

Treatment effect heterogeneity

Today's empirical researchers want to know not only the average of τ_i, but also its distribution

Treatment effect heterogeneity: How does τ_i vary with a subject's characteristics

Treatment effect heterogeneity

Today's empirical researchers want to know not only the average of τ_i, but also its distribution

Treatment effect heterogeneity: How does τ_i vary with a subject's characteristics

Real world examples:

- To which patient should this new medicine be assigned?
- Which subgroup should be targeted on if a candidate wants to run a campaign?
- On which part of Hong Kong the Umbrella Movement generated the largest impact?

Treatment effect heterogeneity

Today's empirical researchers want to know not only the average of τ_i, but also its distribution

Treatment effect heterogeneity: How does τ_i vary with a subject's characteristics

Real world examples:

- To which patient should this new medicine be assigned?
- Which subgroup should be targeted on if a candidate wants to run a campaign?
- On which part of Hong Kong the Umbrella Movement generated the largest impact?

The knowledge on treatment effect heterogeneity allows us to design more efficient experiments and generalize our findings more easily

The fundamental problem of causal inference

In practice, we try to estimate *CATE*:

$$\tau_{CATE}(x_i) = \mathrm{E}\left[Y_i(1) - Y_i(0) \mid X_i = x_i\right]$$

The fundamental problem of causal inference

In practice, we try to estimate *CATE*:

$$\tau_{CATE}(x_i) = \mathrm{E}\left[Y_i(1) - Y_i(0) \mid X_i = x_i\right]$$

But...

The fundamental problem of causal inference

In practice, we try to estimate *CATE*:

$$\tau_{CATE}(x_i) = \mathrm{E}\left[Y_i(1) - Y_i(0) \mid X_i = x_i\right]$$

But...

The “fundamental problem of causal inference”: We cannot observe $Y_i(0)$ and $Y_i(1)$ at the same time (Holland, 1986).

The fundamental problem of causal inference

In practice, we try to estimate *CATE*:

$$\tau_{CATE}(x_i) = \mathrm{E}\left[Y_i(1) - Y_i(0) \mid X_i = x_i\right]$$

But...

The “fundamental problem of causal inference”: We cannot observe $Y_i(0)$ and $Y_i(1)$ at the same time (Holland, 1986).

“Causal inference is a missing data problem.”- Donald Rubin

The fundamental problem of causal inference

In practice, we try to estimate *CATE*:

$$\tau_{CATE}(x_i) = \mathrm{E}\left[Y_i(1) - Y_i(0) \mid X_i = x_i\right]$$

But...

The “fundamental problem of causal inference”: We cannot observe $Y_i(0)$ and $Y_i(1)$ at the same time (Holland, 1986).

“Causal inference is a missing data problem.”- Donald Rubin

All we need to do is to estimate/predict $Y_i(0)$ ($Y_i(1)$) for treated (control) subjects

Various solutions

The estimation/prediction of $Y_i(0)$ ($Y_i(1)$) can be proceeded under various structures (assumptions):

1. Classic solution: $Y_i(0) = \alpha + \beta X_i + \varepsilon_i$, $Y_i(1) = Y_i(0) + \tau$, $E[\varepsilon_i \mid X_i, W_i] = 0$ (parametric model and constant treatment effect)
2. Complete experiments: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i$ (the assignment is at random)
3. Blocking experiments or selection on observables: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid X_i^{'}$ (different from the X_i in CATE!)

What is wrong with regression?

Regression works well when:

- All the necessary variables are included
- The linear model is correct
- The effect is constant across all observations

What is wrong with regression?

Regression works well when:

- All the necessary variables are included
- The linear model is correct
- The effect is constant across all observations

None is likely in practice...

- How to argue that you have controlled for “enough” variables?
- How can one event affect everyone to the same extent?

What is wrong with regression?

Regression works well when:

- All the necessary variables are included
- The linear model is correct
- The effect is constant across all observations

None is likely in practice...

- How to argue that you have controlled for "enough" variables?
- How can one event affect everyone to the same extent?

Aronow and Samii (2016): The estimation of W's coefficient is biased even when we have the right model

What is wrong with regression?

Regression works well when:

- All the necessary variables are included
- The linear model is correct
- The effect is constant across all observations

None is likely in practice...

- How to argue that you have controlled for "enough" variables?
- How can one event affect everyone to the same extent?

Aronow and Samii (2016): The estimation of W's coefficient is biased even when we have the right model

$$\hat{\beta} \xrightarrow{p} \frac{E[\omega_i \tau_i]}{E[\omega_i]} \neq E[\tau_i], \quad \omega_i = (W_i - E[W_i \mid X_i])^2$$

Some (imperfect) solutions

- Matching and weighting (IPW): $\omega_i = \frac{W_i}{P(W_i|X_i)} + \frac{1-W_i}{1-P(W_i|X_i)}$
- Non-parametric regression: kernel and splines
- Use Lin (2013)'s approach, and estimate the following equation instead:

$$Y_i = \alpha + \tau W_i + \beta X_i + \gamma W_i * (X_i - \bar{X}_i) + \varepsilon_i$$

Some (imperfect) solutions

- Matching and weighting (IPW): $\omega_i = \frac{W_i}{P(W_i|X_i)} + \frac{1-W_i}{1-P(W_i|X_i)}$
- Non-parametric regression: kernel and splines
- Use Lin (2013)'s approach, and estimate the following equation instead:

$$Y_i = \alpha + \tau W_i + \beta X_i + \gamma W_i * (X_i - \bar{X}_i) + \varepsilon_i$$

- Again we are not sure about the "right collection" of covariates

How can machine learning help

Let's first focus on the treated group

How can machine learning help

Let's first focus on the treated group

For these observations, $Y_i(1)$ is known, and what we need is $\widehat{Y_i(0)}$

How can machine learning help

Let's first focus on the treated group

For these observations, $Y_i(1)$ is known, and what we need is $\widehat{Y_i(0)}$

With the linear model, $\widehat{Y_i(0)} = \hat{\alpha} + \hat{\beta} X_i$, and $\tau_i = Y_i(1) - Y_i(0)$

How can machine learning help

Let's first focus on the treated group

For these observations, $Y_i(1)$ is known, and what we need is $\widehat{Y_i(0)}$

With the linear model, $\widehat{Y_i(0)} = \hat{\alpha} + \hat{\beta} X_i$, and $\tau_i = Y_i(1) - Y_i(0)$

Without any assumption on model specification, $Y_i(0) = f(X_i)$, and we need to estimate f using information from the control group

How can machine learning help

Let's first focus on the treated group

For these observations, $Y_i(1)$ is known, and what we need is $\widehat{Y_i(0)}$

With the linear model, $\widehat{Y_i(0)} = \hat{\alpha} + \hat{\beta} X_i$, and $\tau_i = Y_i(1) - Y_i(0)$

Without any assumption on model specification, $Y_i(0) = f(X_i)$, and we need to estimate f using information from the control group

Now it becomes a machine learning problem

What is (supervised) machine learning?

With a dataset $\{(X_i, Y_i)\}_{i=1}^n$, how to optimally predict Y_i using X_i?

What is (supervised) machine learning?

With a dataset $\{(X_i, Y_i)\}_{i=1}^n$, how to optimally predict Y_i using X_i?

Require no causality here, just accuracy

What is (supervised) machine learning?

With a dataset $\{(X_i, Y_i)\}_{i=1}^n$, how to optimally predict Y_i using X_i?

Require no causality here, just accuracy

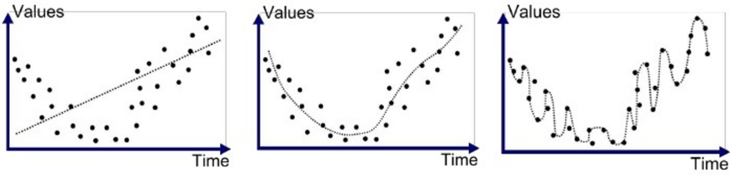
Prediction is not explanation, and we need to capture the underlying pattern rather than noises in the data (avoid overfitting)

What is (supervised) machine learning?

With a dataset $\{(X_i, Y_i)\}_{i=1}^n$, how to optimally predict Y_i using X_i?

Require no causality here, just accuracy

Prediction is not explanation, and we need to capture the underlying pattern rather than noises in the data (avoid overfitting)



Underfitted

Good Fit/Robust

Overfitted

What is (supervised) machine learning?

General idea:

- Split the data into two parts, one training set and one test set
- Fit the model on the training set
- Tune the model based on its performance on the test set and keep the best one

Two sets of basic methods:

- Linear methods (Ridge, LASSO, kernel, SVM)
- Tree-based methods (CART, BART, bagging, random forest)
- It becomes more and more popular to use neural networks and ensemble methods

Linear Methods

A motivating example...

Linear Methods

A motivating example...

Wang and Wong (2018): How does the Umbrella Movement affect the 2016 Legco election?

Linear Methods

A motivating example...

Wang and Wong (2018): How does the Umbrella Movement affect the 2016 Legco election?

Y: the vote share change for the opposition; W: each DCC's distance to the protest site

Linear Methods

A motivating example...

Wang and Wong (2018): How does the Umbrella Movement affect the 2016 Legco election?

Y: the vote share change for the opposition; W: each DCC's distance to the protest site

Traditional approach: Assume that selection on observables holds
($\{Y_i(w)\}_{\forall w} \perp\!\!\!\perp W_i \mid X_i$)
and run regression with covariates

Linear Methods

A motivating example...

Wang and Wong (2018): How does the Umbrella Movement affect the 2016 Legco election?

Y: the vote share change for the opposition; W: each DCC's distance to the protest site

Traditional approach: Assume that selection on observables holds
$(\{Y_i(w)\}_{\forall w} \perp\!\!\!\perp W_i \mid X_i)$
and run regression with covariates

But what is the "right collection" of covariates (X) in this case?

Linear Methods

We replace selection on observables with sparsity: The “right collection” can be approximated by the transformation and combination of variables in a given set

Linear Methods

We replace selection on observables with sparsity: The “right collection” can be approximated by the transformation and combination of variables in a given set

Suppose we have ten covariates for all the DCCs

Linear Methods

We replace selection on observables with sparsity: The "right collection" can be approximated by the transformation and combination of variables in a given set

Suppose we have ten covariates for all the DCCs

We create a pool of features ($\tilde{X}$) by generating the higher order terms of all the covariates and the interactions between them

Linear Methods

We replace selection on observables with sparsity: The "right collection" can be approximated by the transformation and combination of variables in a given set

Suppose we have ten covariates for all the DCCs

We create a pool of features ($\tilde{X}$) by generating the higher order terms of all the covariates and the interactions between them

Then the number of variables may be larger than that of observations

Linear Methods

We replace selection on observables with sparsity: The "right collection" can be approximated by the transformation and combination of variables in a given set

Suppose we have ten covariates for all the DCCs

We create a pool of features ($\tilde{X}$) by generating the higher order terms of all the covariates and the interactions between them

Then the number of variables may be larger than that of observations

We have to use LASSO to select the collection of features that fits the data the best

Ridge, LASSO, and Elastic Net

All can be seen as "penalized regression":

$$Loss = (Y - X\beta)'(Y - X\beta) + \lambda * Penalty$$

Ridge, LASSO, and Elastic Net

All can be seen as "penalized regression":

$$Loss = (Y - X\beta)'(Y - X\beta) + \lambda * Penalty$$

When *Penaly* equals to:

Ridge, LASSO, and Elastic Net

All can be seen as “penalized regression”:

$$Loss = (Y - X\beta)'(Y - X\beta) + \lambda * Penalty$$

When *Penaly* equals to:

- $\sum_{i=1}^{K} \beta_k^2$, Ridge
- $\sum_{i=1}^{K} |\beta_k|$, LASSO
- $\begin{pmatrix} \sum_{i=1}^{K} \beta_k^2 \\ \sum_{i=1}^{K} |\beta_k| \end{pmatrix}$, Elastic Net

Ridge, LASSO, and Elastic Net

All can be seen as "penalized regression":

$$Loss = (Y - X\beta)'(Y - X\beta) + \lambda * Penalty$$

When *Penaly* equals to:

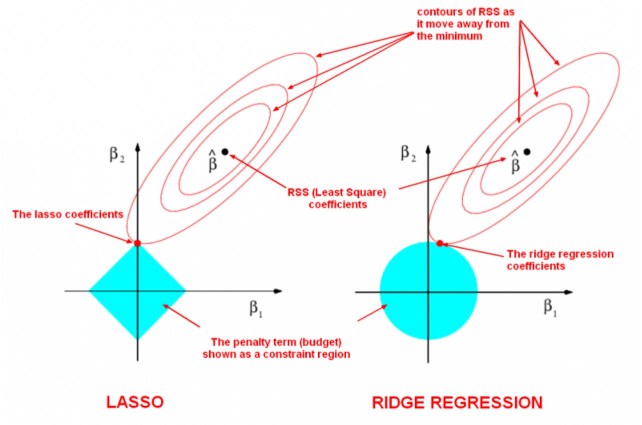
- $\sum_{i=1}^{K} \beta_k^2$, Ridge
- $\sum_{i=1}^{K} |\beta_k|$, LASSO
- $\begin{pmatrix} \sum_{i=1}^{K} \beta_k^2 \\ \sum_{i=1}^{K} |\beta_k| \end{pmatrix}$, Elastic Net

The value of λ is decided by cross-validation:

- Split the sample into training and test set (4:1)
- Estimate β on the training set for a given λ
- Try different values of λ, and find the one that minimizes the loss on the test set

Ridge, LASSO, and Elastic Net

Notice that the loss function looks like the equation for Lagrangian multiplier



Clearly, LASSO can select variables while Ridge cannot

Ridge, LASSO, and Elastic Net

For our question, we need “double selection” (Belloni et al., 2013)

Ridge, LASSO, and Elastic Net

For our question, we need "double selection" (Belloni et al., 2013)

Select features that are correlated with Y and features that are correlated with W

- Run LASSO on the relationship between Y and $\tilde{X}$, and select a subset $\tilde{X}_1$
- Run LASSO on the relationship between Y and $\tilde{X}$, and select a subset $\tilde{X}_2$
- Regress Y on W and the union of $\tilde{X}_1$ and $\tilde{X}_2$ (follow Lin's suggestion!)

Ridge, LASSO, and Elastic Net

For our question, we need "double selection" (Belloni et al., 2013)

Select features that are correlated with Y and features that are correlated with W

- Run LASSO on the relationship between Y and $\tilde{X}$, and select a subset $\tilde{X}_1$
- Run LASSO on the relationship between Y and $\tilde{X}$, and select a subset $\tilde{X}_2$
- Regress Y on W and the union of $\tilde{X}_1$ and $\tilde{X}_2$ (follow Lin's suggestion!)

An improved version: Double machine learning (first split and then double select)

Takeaway

- Open your R and load your dataset
- install.packages("glmnet")
- Generate your features using covariates (you can simply write a loop)
- output $<-$ cv.glmnet(Features, Outcome, family = "gaussian", alpha = 1)
  coefs $<-$ coef(output)

Other linear methods

- Ridge is rare, but elastic net has some good properties
- When the outcome is binary, we use Support Vector Machine, which can also select variables
- If you want to use all the features, kernalize them

Trees

A motivating example

Trees

A motivating example

- A politician wants to run a campaign and she has limited fund
- She knows the demographics of her constituents and there are several tools available (door-to-door, phone, mail..)
- How to choose the most effective tool for a given group (e.g. uneducated Hispanic blue-collar workers who are younger than 30)

We can conduct an experiment on a small sample, and randomly assign tools to different households

Trees

A motivating example

- A politician wants to run a campaign and she has limited fund
- She knows the demographics of her constituents and there are several tools available (door-to-door, phone, mail..)
- How to choose the most effective tool for a given group (e.g. uneducated Hispanic blue-collar workers who are younger than 30)

We can conduct an experiment on a small sample, and randomly assign tools to different households

And then?

Trees

Suppose we know the idiosyncratic treatment effect τ_i for each household under each of the tools (e.g. mail)

Trees

Suppose we know the idiosyncratic treatment effect τ_i for each household under each of the tools (e.g. mail)

Which covariates predict the effect the best?

Trees

Suppose we know the idiosyncratic treatment effect τ_i for each household under each of the tools (e.g. mail)

Which covariates predict the effect the best?

Knowing that, the politician can target the subgroup on whom the campaign is the most effective

Trees

Suppose we know the idiosyncratic treatment effect τ_i for each household under each of the tools (e.g. mail)

Which covariates predict the effect the best?

Knowing that, the politician can target the subgroup on whom the campaign is the most effective

In essence, it is a classification problem, and tree-based methods deal with that very well

Trees

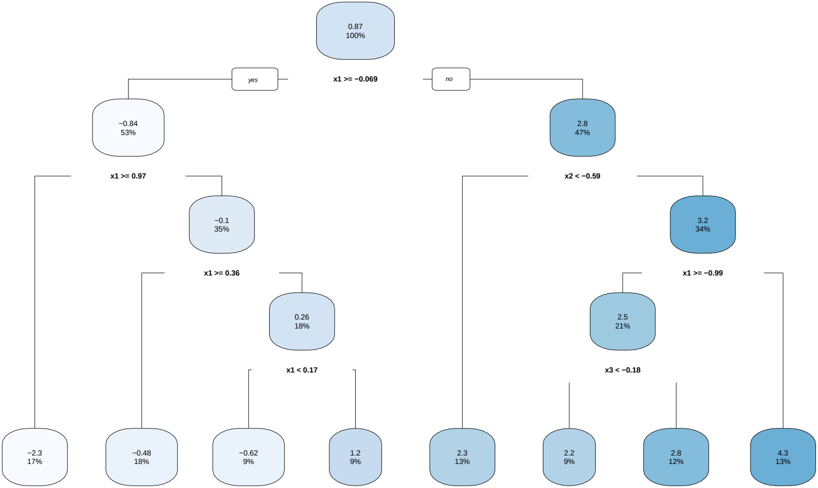
The most popular tree-based method is Classification and Regression Tree (CART)

Trees

The most popular tree-based method is Classification and Regression Tree (CART)

It classifies observations into homogeneous blocks (leaves) based on the outcome (within each block the outcome is as stable as possible)

Trees


0.87
100%
yes
x1 >= −0.069
no
−0.84
53%
2.8
47%
x1 >= 0.97
x2 < −0.59
−0.1
35%
3.2
34%
x1 >= 0.36
x1 >= −0.99
0.26
18%
2.5
21%
x1 < 0.17
x3 < −0.18
−2.3
17%
−0.48
18%
−0.62
9%
1.2
9%
2.3
13%
2.2
9%
2.8
12%
4.3
13%

Trees

- It predicts the outcome within each “leave” with the sample average and minimizes:

$$Loss = \sum_{i}^{N} (Y_i - \bar{Y}_{l(i \in l)})^2 + \lambda |L|$$

Trees

- It predicts the outcome within each "leave" with the sample average and minimizes:

$$Loss = \sum_{i}^{N} (Y_i - \bar{Y}_{l(i \in l)})^2 + \lambda |L|$$

- CART is based on greedy algorithm (find the best partition step by step) thus the result may not be a global optimum

Trees

- It predicts the outcome within each "leave" with the sample average and minimizes:

$$Loss = \sum_{i}^{N} (Y_i - \bar{Y}_{l(i \in l)})^2 + \lambda |L|$$

- CART is based on greedy algorithm (find the best partition step by step) thus the result may not be a global optimum
- Obviously the final solution will contain many leaves withouth penalty

Trees

- It predicts the outcome within each "leave" with the sample average and minimizes:

$$Loss = \sum_{i}^{N} (Y_i - \bar{Y}_{l(i \in l)})^2 + \lambda |L|$$

- CART is based on greedy algorithm (find the best partition step by step) thus the result may not be a global optimum
- Obviously the final solution will contain many leaves withouth penalty
- So we need to prune leaves from a grown-up tree using cross-validation

Trees

- It predicts the outcome within each "leave" with the sample average and minimizes:

$$Loss = \sum_{i}^{N} (Y_i - \bar{Y}_{l(i \in l)})^2 + \lambda |L|$$

- CART is based on greedy algorithm (find the best partition step by step) thus the result may not be a global optimum
- Obviously the final solution will contain many leaves withouth penalty
- So we need to prune leaves from a grown-up tree using cross-validation
- We can combine multiple trees to increase accuracy (bagging and random forest)

Pros and Cons for CART

CART is popular because:

Pros and Cons for CART

CART is popular because:

- It is very straightforward
- Very fast
- No need for model
- It performs well when the relationship is highly non-linear

Pros and Cons for CART

CART is popular because:

- It is very straightforward
- Very fast
- No need for model
- It performs well when the relationship is highly non-linear

But

- It is not causal yet (τ_i is assumed to be known)
- Overfitting is a big problem even with pruning

Causal Trees

Imai and Strauss (2011) shows how to use CART to analyze experimental results

Causal Trees

Imai and Strauss (2011) shows how to use CART to analyze experimental results

Covariates may affect both the level of the outcome (Y) and the magnitude of the effect (τ_i)

Causal Trees

Imai and Strauss (2011) shows how to use CART to analyze experimental results

Covariates may affect both the level of the outcome (Y) and the magnitude of the effect (τ_i)

1. Select covariates that explain the outcome well
2. With the selected variable controlled, estimate the contribution of each covariate to the treatment effect, and rank the covariates based on that
3. Choose the number of covariates (j), fit the model using the best j covariates
4. Pick out the j that gives the model the strongest prediction power

General idea: Fit separated trees for the treated and control group

Causal Trees

Athey, Imbens and Wager have a series of papers on how to directly apply trees to causal inference

Causal Trees

Athey, Imbens and Wager have a series of papers on how to directly apply trees to causal inference

- Split the sample into three parts: training set, estimate set, and test set
- Use the first one to fit the tree, the second to estimate the effect, and the third to decide the number of leaves
- Modify the loss function to replace the average outcome with the estimated effect on each leave and reduce overfitting
- A more advanced version: Causal Forest

Takeaway

- install.packages("causalTree")
- honestTree < − honest.causalTree(formula, data = train_data, treatment = train_data$treatment, est_data = est_data, est_treatment = est_data$treatment, split.Rule = "CT", split.Honest = T, HonestSampleSize = nrow(est_data), split.Bucket = T, cv.option = "CT")
- opcp < − honestTree$cptable[,1][which.min(honestTree$cptable[,4])]
- opTree < − prune(honestTree, opcp)
- rpart.plot(opTree)
- est_data$leaves < − as.vector(rpart.predict.leaves(opTree, est_data, type = "where"))

Tree-based shape test

What if we are interested in not only an estimate, but also some shape-related hypothesis (i.e. the moderator effect is U-shaped)?

Tree-based shape test

What if we are interested in not only an estimate, but also some shape-related hypothesis (i.e. the moderator effect is U-shaped)?

An intuitive solution: Split the moderator's support into several bins using CART and test whether the bin-specific effect first drops and then rises

Tree-based shape test

What if we are interested in not only an estimate, but also some shape-related hypothesis (i.e. the moderator effect is U-shaped)?

An intuitive solution: Split the moderator's support into several bins using CART and test whether the bin-specific effect first drops and then rises

However, CART relies on greedy algorithm, so the splitting varies a lot across similar datasets

Tree-based shape test

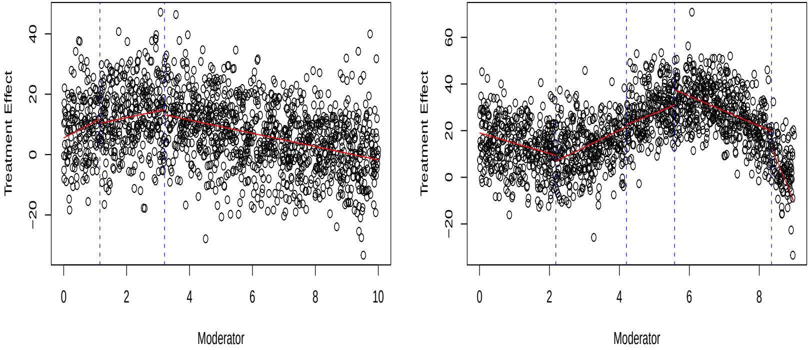
What if we are interested in not only an estimate, but also some shape-related hypothesis (i.e. the moderator effect is U-shaped)?

An intuitive solution: Split the moderator's support into several bins using CART and test whether the bin-specific effect first drops and then rises

However, CART relies on greedy algorithm, so the splitting varies a lot across similar datasets

- Samii and Wang (2018): Evolution tree algorithm can solve the problem perfectly
- It uses an iteration method to find the globally optimal splitting
- The approach is also honest: One half of the sample is used for splitting, and the other is used for testing

Tree-based shape test



- One dataset with a U-shaped moderator effect, one with a more
- Blue lines: partition generated by the tree algorithm
- Red lines: estimated moderator effect in each leave

Summary

- Machine learning is more and more popular in social sciences
- Now we can predict the counterfactual with high accuracy
- These algorithms shed light on subtlties in the dataset
- However, they are not designed for causal identification
- More modifications are required to meet the demand of social scientists