Big Data and Economics

Causal Forests

Kyle Coombs
Bates College | ECON/DCS 368

Table of contents

Prologue

Prologue

- Most causal inference will estimate a single treatment effect and maybe a few interactions
- If an RCT shows that a financial education intervention increases earnings by 5K, does that mean everyone who experiences a 5K increase in earnings?

Prologue

- Most causal inference will estimate a single treatment effect and maybe a few interactions
- If an RCT shows that a financial education intervention increases earnings by 5K, does that mean everyone who experiences a 5K increase in earnings?
- In reality, there are many treatment effects that vary across the population
- If you know who benefits the most from a treatment, you can target the treatment to those people
- If you get it right, you're maximizing the gain from each policy dollar spent on a treatment

Remember this RCT example?

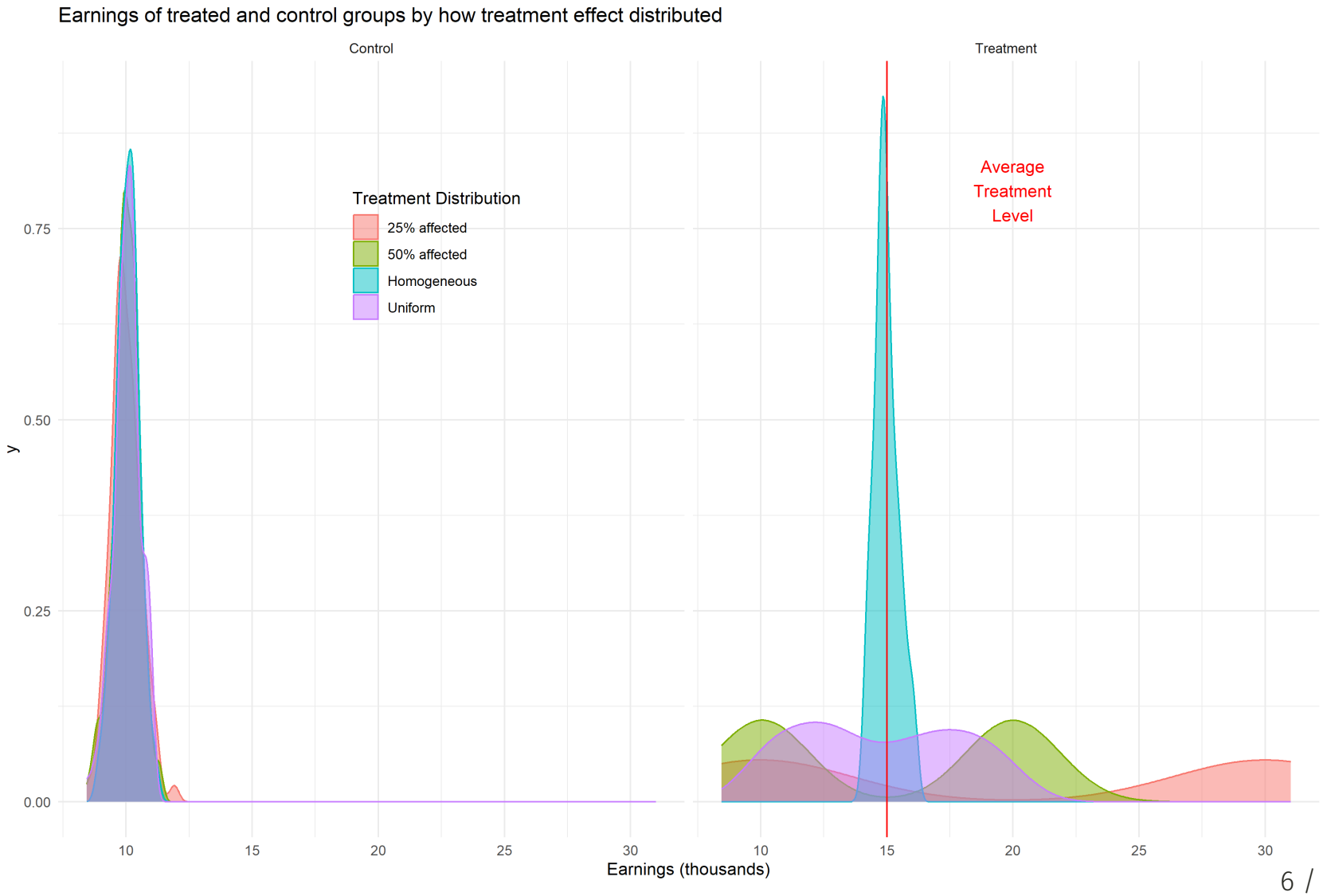
- Imagine you are evaluating who to give a financial education intervention to
- An RCT shows an educational intervention increases earnings by 5K on average, that is the treatment effect τ:

$$\tau = \mathbb{E}[y|\text{Treated}] - \mathbb{E}[y|\text{Control}] = \$5K$$

- Which of the following can you rule out?

1. Every treated student experienced a 5K increase in earnings
2. Half of treated students received a 10K increase in earnings, half experienced nothing
3. 25\% of treated students experienced a 20K increase in earnings, 90\% experienced nothing
4. Earnings increases uniformally distributed between 0 and 10K for the treated

Visualizing treatment effects



Who do we target?

- Imagine a policymaker believes the RCT (a miracle!)
  - **Problem**: There's a limited budget to select students to receive the intervention in the future
  - **Question**: How do we select students to maximize the impact of the intervention later?

Who do we target?

- Imagine a policymaker believes the RCT (a miracle!)
  - **Problem**: There's a limited budget to select students to receive the intervention in the future
  - **Question**: How do we select students to maximize the impact of the intervention later?
- Students with high baseline scores might benefit more/less than students with low baseline scores
- Alternatively, students of color may benefit more/less than white students
- Or students from low-income families may benefit more/less than students from high-income families
- It could be a combination of all of them! Or something unobserved!

Conditional Average Treatment Effects

- The ATE is the average treatment effect for everyone in the sample
- But what if we want the ATE for a specific group?
- For example, what if we want the ATE conditional on low baseline test scores?

$$CATE = \mathbb{E}[y|\text{Treated}, \text{Low Baseline}] - \mathbb{E}[y|\text{Control}, \text{Low Baseline}]$$

- How might we typically see how a treatment differs by a covariate?

Hint: an ATE is typically just a β in a regression -- how do we see estimate changes to β from a new variable?

Conditional Average Treatment Effects

- The ATE is the average treatment effect for everyone in the sample
- But what if we want the ATE for a specific group?
- For example, what if we want the ATE conditional on low baseline test scores?

$$CATE = \mathbb{E}[y|\text{Treated, Low Baseline}] - \mathbb{E}[y|\text{Control, Low Baseline}]$$

- How might we typically see how a treatment differs by a covariate?

Hint: an ATE is typically just a β in a regression -- how do we see estimate changes to β from a new variable?

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 + \epsilon$$

- Interactions work for just a few variables, but what if we have dozens of potential interactions?
  - You quickly lose statistical power as you add more interactions
  - Also, you can quickly descending into p-hacking if you try interactions until one shows a significant effect
  - Why does p-hacking lead to bad policy?

Conditional Average Treatment Effects

- The ATE is the average treatment effect for everyone in the sample
- But what if we want the ATE for a specific group?
- For example, what if we want the ATE conditional on low baseline test scores?

$$CATE = \mathbb{E}[y|\text{Treated, Low Baseline}] - \mathbb{E}[y|\text{Control, Low Baseline}]$$

- How might we typically see how a treatment differs by a covariate?

Hint: an ATE is typically just a β in a regression -- how do we see estimate changes to β from a new variable?

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 + \epsilon$$

- Interactions work for just a few variables, but what if we have dozens of potential interactions?
  - You quickly lose statistical power as you add more interactions
  - Also, you can quickly descending into p-hacking if you try interactions until one shows a significant effect
  - Why does p-hacking lead to bad policy?
- There are bound to be spurious correlations in any dataset

Causal Forests and CATEs

- Causal forests provide a way to estimate a CATE as a function of covariates without having to specify interactions
- Basically, it maps a person to a CATE based on their observable characteristics
- This is a very powerful tool for policy
- It is also really tricky to implement correctly
- And it often does not work as well in practice as it does in theory

Decision Tree Review

What is a decision tree?

What is a decision tree?

- A decision tree organizes variables into tree like structure
    - It is essentially, a really fancy flowchart
- At each node, pick the variable that best meets a decision rule
- At node 1, the algorithm cycles through each X variable and finds the split in the data that best meets the decision rule
    - It picks the best X variable
    - It follows the branch down and creates nodes by looking at the remaining X's that best meet the decision rule
- When making a decision about an observation, follow the tree down the branches

Types of decision trees

Regression trees

- The decision rule is what variable X best predicts y when split at some cutoff point $\bar{X}$
    - Typically the predicted $\hat{y}$ is the average of y conditional on X being less than or greater than $\bar{X}$
    - Alternatively, it could be the mode
- At the terminal node, the prediction $\hat{y}$ is the average of y for all observations in that node
- The decision rule is whatever split minimizes the sum of squared errors (SSE) between the predicted $\hat{y}$ and the actual y

Causal Trees

- Instead of splitting based on prediction of y, split to maximize the difference in the average treatment effect (ATE) between the two branches
- At each node, the X covariate that maximizes the difference in the ATE is selected
- Why?
- The goal is to see how varied the treatment effect is across different subgroups of the population

Random Forests

Many trees make a forest

- Decision trees are fairly easily to interpret once you make one
- But one drawback is that they are very sensitive to the data
    - Too many nodes and you could overfit
    - Too few nodes and you'll just have noise
- So what if we made many trees and averaged the predictions?
    - Technically this is just called "bagging" (bootstrap aggregating)
    - Random forests also randomize the variables available to split the nodes
    - See more at Introduction to Statistical Learning, Chapter 8.2
- But won't we just repeat the same tree over and over?

Pull yourself up by your bootstraps

- How could we use bootstrapping?

Pull yourself up by your bootstraps

- How could we use bootstrapping?
- If you bootstrap the data B times, you create B new samples of the data indexed b

1. For each bootstrap sample b, create a decision tree T_b using the bootstrap sample b
2. For each observation i in the original sample, predict the outcome y_i using all B trees
3. Average the predictions as $\hat{y}_i = \frac{1}{B} \sum_{b=1}^{B} T_b(X_i)$

- This is called bagging (bootstrap aggregating)
- **Intuition**: With many trees, you can average out the noise and get a better prediction
- Random forests add a twist to bagging by randomly selecting a subset of X variables to split the nodes in the tree
    - This ensures the trees are uncorrelated with each other
    - Minimizes variance

Intuition: By randomizing the X variables available to a tree, they are less likely to only use the same variables to split the nodes in the tree. As a result, the algorithm evaluates other variables in the data.

Use cases of random forests

- Random forests are a very popular machine learning technique
- They are used for prediction, classification, and causal inference
- Kleinberg et al. (2018) use random forests to predict the judicial bail decisions in NYC

Causal Forests

Causal Forests

- Causal forests are a type of random forest that estimate the conditional average treatment effect (CATE) for each observation
- Causal forests are just a bunch of causal trees
    - Each node in the tree maximizes the difference in the ATE between the two branches
    - The result is a tree-specific CATE for each observation
- The average of the tree-specific CATEs is the CATE for each observation

Limitations of causal forests

- Causal forests cannot resolve the fundamental problem of causal inference: unobserved confounders
- Causal forests **cannot, will not, and never** will be able to create a causal effect where not exists

Separate Causal Trees

- A single causal tree is created by finding the cutoff in each variable that maximizes the treatment effect variance across groups
- What is "treatment effect variance across groups"?
    - Roughly it corresponds to the difference in ATE between the two groups
    - There are different ways to write this out, but they should discount any variation in the treatment effects within the groups
- The result is a tree-specific CATE for each observation

Honest Causal Forests

- One issue that can arise is if you use the **same** sample to split the data as you use to estimate the treatment effects in a sample
- Why? This leads to issues with the standard errors.
- **Intuition**: Once you split the data based on the treatment effects, any estimated treatment effects are no longer truly random.
    - In general, you never want to use input data to an algorithm to evaluate its performance
- "Honest causal forests" provide a workaround
- Split the data to make a causal tree in a "splitting sample" and "estimation sample"
    1. The "splitting sample" is used to pick the splitting variables/cutoffs
    2. The "estimation sample" is grouped based on the splitting rules, then the treatment effects are calculated
- The goal is to maintain the randomness that generated the treatment assignment, while using the same groups

Causal Forest algorithm

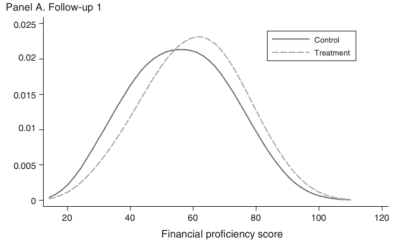
Here is an algorithm for causal forests taken from Davis and Heller (2016)

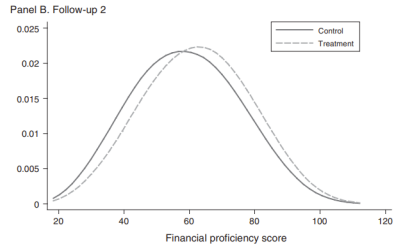
1. Draw subsample b without replacement of $n_b < N$ observations from the original sample of size N.
2. Randomly split the n_b observations to form a training sample tr and an estimation sample e so $n_t r = n_e = \frac{n_b}{2}$.
3. For each value of each $X_j = x$, form candidate splits of the training sample into two groups based on whether $X_j \leq x$.
    - Choose the split that maximizes treatment effect variance across the two subgroups.
    - If the split increases variance relative to no split, split. If no split increases the variance, this is a terminal node.
4. Once no more splits possible, group the n_e observations in this tree based on Xs.
5. With the estimation sample, calculate $\tau^l = y_T - y_C$ within each terminal node. (Makes it honest!)
6. In full sample, assign $\tau_b^l = \tau^l$ to each observation whose Xs would place it in node l. Save τ_b^l.
7. Repeat steps 1-6 B times to create B trees
8. Define each i's CATE as $\tau_{CF}^i(x) = \frac{1}{B} \sum_{b=1}^{b} \tau_b^l$, the average prediction for that individual across trees.

Financial education in Brazil

- Bruhn et al. (2016) evaluate the effect of a financial education intervention on student achievement in Brazil
    - 892 (~25K students) schools across six states were randomly assigned to treatment or control
    - Treatment: three semesters of financial education during 11th and 12th grade by trained students with free textbooks
    - Sub-treatment: Parental workshop on financial education
- Average treatment effect that student financial proficiency increased by 7% initially, dropping to 5% in second follow-up

Distributions of financial proficiency





Causal forests and Bruhn et al. (2016)

- The makers of the R package **grf** for generalized random forest, create a tutorial of how to use causal forests to estimate the conditional average treatment effects in Bruhn et al. (2016)
- Find CATE vary a decent amount across different variables
- Specifically, those with lower financial autonomy benefited more than average from the program
- The application then shows how to use the package **policytree** to estimate the optimal policy implementation

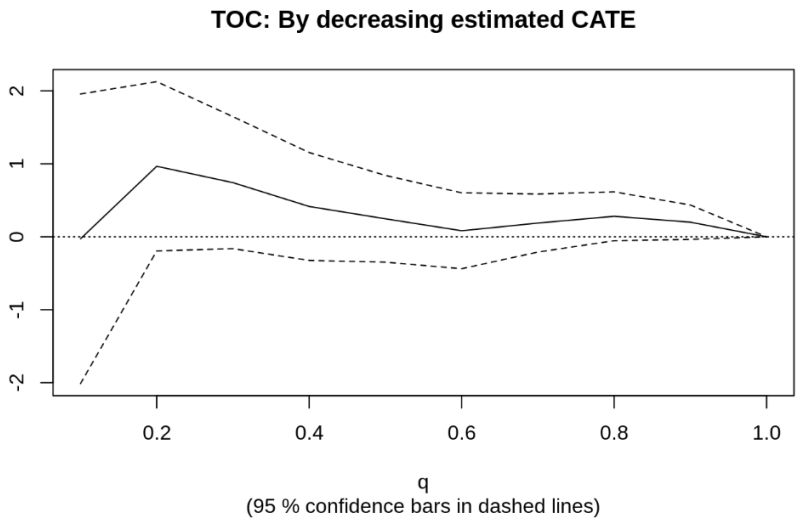
Best Linaar Projection

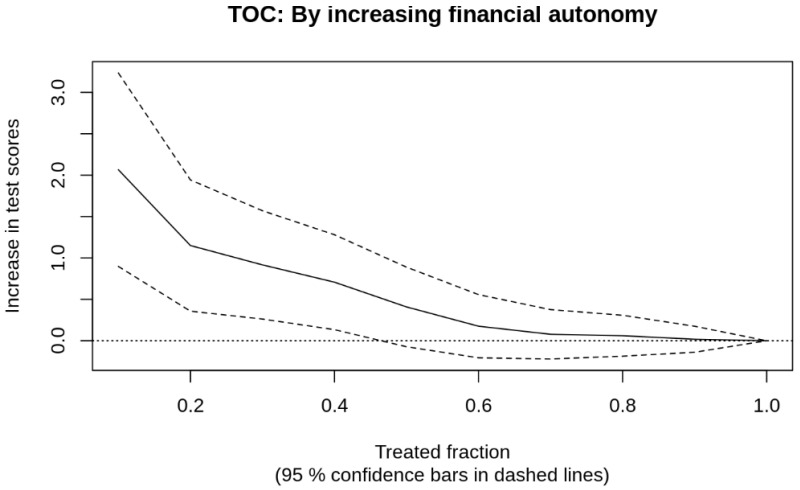
```
best_linear_projection(cf, X[ranked.vars[1:5]])
#>
#> Best linear projection of the conditional average treatment effect.
#> Confidence intervals are cluster- and heteroskedasticity-robust (HC3):
#>
#>                                 Estimate Std. Error t value Pr(>|t|)
#> (Intercept)                       7.3364     1.0807    6.79  1.2e-11 ***
#> financial.autonomy.index         -0.0244     0.0140   -1.74    0.082 .
#> intention.to.save.index          -0.0124     0.0156   -0.80    0.426
#> family.receives.cash.transfer    -0.8700     0.6363   -1.37    0.172
#> has.computer.with.internet.at.home  -0.8212     0.6264   -1.31    0.190
#> is.female                        -0.8019     0.5356   -1.50    0.134
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ranked Average Treatment Effect (RATE)

- Once the CATE is known, you can start seeing how the treatment effect varies across the population
- Specifically, you can rank the population by their CATE and then calculate the CATE on a separate sample
    - "Train sample" to train the forest
    - "Test sample" to predict the CATE
    - "Evaluation sample" to see how well the CATE predicts the treatment effect
- Alternatively, you can do the same thing, but rank instead by a covariate of interest (e.g. financial autonomy) that seems to drive the CATE differences

Ranked Average Treatment Effect





Other applications

- Davis and Heller (2017) show how to use causal forests to estimate the effect of a job training program for at-risk youth on employment and criminal activity
    - Find minimal evidence of heterogeneity in treatment effect on crime, some on employment
- Athey and Wager (2018) look at the effect of a growth mindset intervention on student achievement and how that varies across the population
    - Finds evidence of heterogeneity in treatment effect (unless they account for school-level clustering of treatment)
- Mark White finds somewhat heterogeneous treatment effects in work by Broockman and Kalla (2016) on reducing transphobia through canvassing
- Farbmacher et al. (2021) find heterogeneous treatment effects of the effect of payday on cognitive test performance
    - Suggests low-income young and elderly people most inattentive when payday is far away

What next?

- Get your hands dirty!
- Navigate to the Generalized Random Forest vignette

```
#install.packages('grf')
library(grf)
```

- This will walk you through how to use the **grf** package to estimate causal forests
- Once you finish, try the **grf** guided tour
    - I recommend you try the application to school program evaluation example
- This package is full of vignettes that you could use for the problem set

Next lecture: Least Absolute Shrinkage and Selection Operator (LASSO)