Simulation Options for CATE Estimation in Target Sample

Carly Lupton Brantner

Stuart Lab: 1/31/2023

Overview

Goal: Estimate conditional average treatment effect (CATE) for target group of individuals after fitting a multi-study CATE model.

1. Background
2. Notation
3. Simulation Setup
4. Open Questions

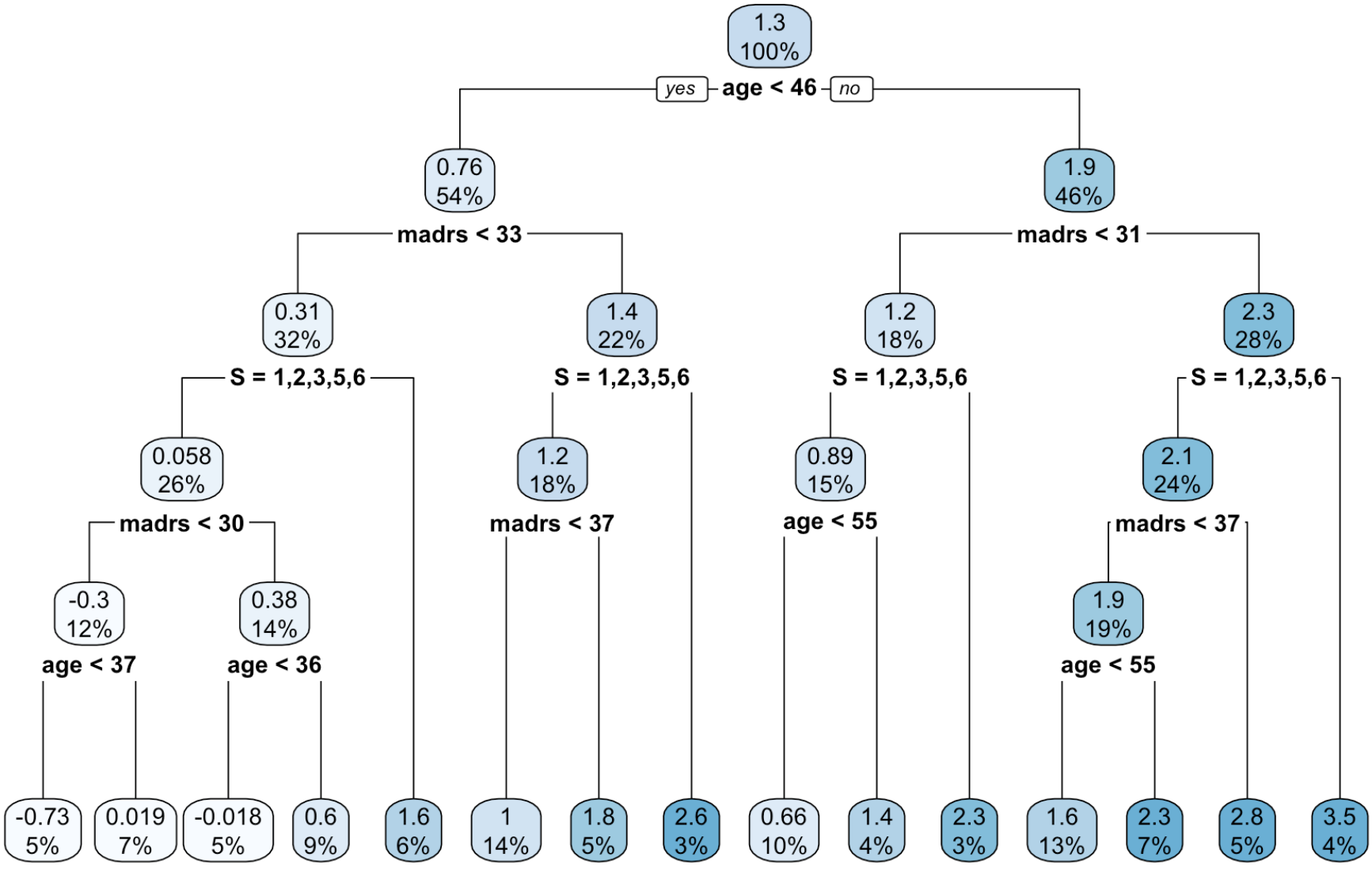
Background

- Finishing up a project where I estimate the conditional average treatment effect by combining data from multiple randomized controlled trials
- Several methods yield study-specific functions where study matters at varying levels
- Goal is either a universal treatment effect function or a way to apply these functions to individuals who are not coming from the original sample
- Idea: For the target sample, treat study as *missing* and compare different ways of imputing it.
  - Bootstrapping framework for imputation
  - Creating a simulation design to compare methods for imputing study and estimating a mean and confidence interval for the CATE estimates of target individual

Background continued

- More on the idea: Can we do a sort of bootstrapping/posterior sampling procedure where we repeat the individual observation N times, and for each iteration we randomly assign study and then predict the CATE
- Then we can aggregate CATE estimates
- We want to see which option for doing this bootstrapping works the best
- Options: completely random, based on study membership probability, causal forest default, imputing within causal forest

Background continued



Notation

- $A \in \{0, 1\}$ indicates treatment status
- $\boldsymbol{X}$ are covariates (continuous)
- Y is a continuous outcome
  - $Y(1)$ is the potential outcome under treatment
  - $Y(0)$ is the potential outcome under control
- $S \in \{1, ..., K\}$ is a study indicator

Estimand

The estimand is the study-specific conditional average treatment effect:

$$\tau_s(\boldsymbol{X}) = E(Y(1)|\boldsymbol{X}, S = s) - E(Y(0)|\boldsymbol{X}, S = s)$$

Simulation Steps: Training Data

Training Data: Mimic MDD dataset: RCTs comparing Duloxetine versus Vortioxetine for reduction in depressive symptoms, measured by MADRS score

- 6 studies with 200 or 500 people in each
- Probability of treatment is 1/2
- 5 covariates per person: sex, smoking status, weight, baseline MADRS (depression scale), age such that:
  1. Same covariate distribution
  2. MADRS varies
  3. MADRS and age vary
  4. Age completely distinguishable across studies

Simulation Steps: Target Data

- Simulate target data using one of three options
  1. Random sample of 100 individuals from the training data
  2. Random sample of 100 individuals from the training data, where individuals from studies 3 and 5 have three times the chance of getting selected
  3. Different sample of 100 individuals who are younger and have less severe depression according to their MADRS score
- In all three options, study is assigned to be missing

Simulation Steps: Outcome and Treatment Effect

- Define m and τ function in one of two ways:
  - **Simple:** $m = -0.02 * age - 0.7 * madrs - 0.15 * sex$ and $\tau = -8.5 + 0.07 * age + 0.20 * madrs + \epsilon_{study}$, where $\epsilon_{study} \sim N(0, \epsilon_{sd}^2)$ in the training data and $\epsilon_{study} = 0$ in the target data
  - **Linear:** randomly sample study coefficients from normal distributions with different parameters and small SDs: $m = 10.7 - S_{main} - .02 * age - 0.87 * madrs - 0.15 * sex + S_{inter} * madrs$ and $\tau = -10.5 + .07 * age + .2 * madrs + S_\tau$.
- Define $Y = m + A * \tau + \epsilon$ where $\epsilon \sim N(0, 0.05^2)$
- Fit model to training data using causal forest with pooling with trial indicator to estimate CATE

Simulation Steps: Method Comparison

- Predict on target data and create confidence intervals for each individual using each imputation method
  1. **Completely random:** Repeat each observation N times and for each repetition, randomly sample S with equal probability for each value. Predict and then calculate a confidence interval using the mean and standard deviation of the N predictions.
  2. **Study membership model:** Same as completely random but when assigning study, do it according to a distribution defined by the probabilities of a study membership multinomial logistic regression model.
  3. **Within-forest default:** Follow the default process of the causal forest, meaning split such that all missing values go in the same direction. This won't be a bootstrap anymore.
  4. **Within-forest random sampling:** assign study based on the probabilities in each split of the tree.

Simulation Steps: Evaluation Metrics

- Calculate measures of accuracy
  1. MSE in training data
  2. MSE in target data
  3. Confidence interval coverage in target data
  4. Confidence interval length in target data

Simulation Parameters

Overall, the simulation will vary depending on the following parameters:

1. $N = 100$, where N is the number of repetitions in the bootstrap
2. $K = 6$, where K is the number of RCTs
3. $n_k \in \{200, 500\}$ where n_k is the sample size for study k
4. m and τ scenario: Simple with $\epsilon_{sd} \in \{0.01, 0.05, 1, 3\}$ or Linear
5. Training data distributional shift: same, varying MADRS, varying MADRS and age, distinguishable age
6. Target data setup: random sample, random with upweighted studies, different distribution

Open Questions

1. Any thoughts on the simulation setup? Does anyone have experience / know of any helpful papers that simulate based on a real dataset?
2. Running into issues with the last idea - does anyone have any coding suggestions for predictions from trees?
3. Any other ideas for how to do the "imputation" / sampling?