Homework 3

February 3, 2025

In this homework you will apply some of the methods we have covered in class to the problem of estimating heterogeous treatment effects in a simulated dataset.

You may use whatever sofware packages you like. For the generalized random forest in the last question you can use the grf package in R or the grf function in the EconML python package (see https://econml.azurewebsites.net/_autosummary/econml.grf.CausalFores
You may find the R examples here helpful https://grf-labs.github.io/grf/.

Download the dataset from the course Moodle page. The data is from an (imaginary) trial of a novel training course aimed at improving year 6 students' mathematical intuition. Students in the sample were randomly assigned to treatment and control groups. However, some students in the treatment group opted out of the training course. The treatment indicator in the data only measures whether the student actually attended the course, not whether they were initially assigned to the treatment group.

Variable Name	Description
math_outcome	Math test score collected one year post-experiment
reading_outcome	Reading test score collected one year post-experiment
treat	A binary indicator equal to 1 if a student was treated and 0 otherwise
age	Age of the student at the start of the experiment
gender	Student gender (0 if male, 1 if female)
math_score	Math score prior to the experiment
reading_score	Reading test score prior to the experiment
socioecon	Measure of socioeconomic advantage (0="low", 1="medium", 2="high")
gpa	Student grade point average prior to the start of the experiment
free_meals	Whether the student qualifies for free school meals
single_parent	Whether the student belongs to a single parent household
postal_code	A integer value that indicates the postal area in which the student lives

i. Drop all variables other than the treatments and outcomes. Check whether any observations have missing values and if so, drop them from your sample. Replace the variable postal_code with a matrix of binary indicator variables. More precisely, for each number x that appears in postal_code create a new variable that is equal to 1 if an observation has postal_code equal to x and 0 otherwise. Why might this be helpful?

ii. Split the data at random into a training sample that consists of roughly 80% of observations and a test sample that contains the remaining 20%. Train a model that predicts the outcomes from the pre-treatment covariates using OLS and LASSO. For OLS do not include the postal_code binary indicator variables in your model. Choose the penalty parameter for LASSO using cross-validation with 5 folds. Do this separately for the treated and untreated students. Use only data in your training sample to train these models.

iii. Are either of the coefficient vectors you estimated using LASSO sparse? Assess the performance of the OLS and LASSO estimates on the test sample. Describe your findings.

iv. Using the models you trained using LASSO estimate the average treatment effect for individuals in the training sample.

v. Train the causal forests algorithm of Athey, Tibshrani and Wager (Annals of Statistics, 2019). Use the trained model to predict individual treatment effects in the test sample.