Presenting groups: 17 and 18, Date: 14.07.2021

The set-up is retrieved from ”Estimating Treatment Effects with Causal Forests: An Application” (Link to paper), by Susan Athey and Stefan Wager (2019). The goal of the problem set is to simulate a dataset according to the specifications below and compare the performance of causal random forests with the performance of alternative treatment effect estimators. You can use pre-installed packages such as **grf** for causal forests.

Consider the following data generating process given in the paper:

- *S3: Student's self-reported expectations for success in the future, a proxy for prior achievement, measured prior to random assignment*
- *C1: Categorical variable for student race/ethnicity*
- *C2: Categorical variable for student identified gender*
- *C3: Categorical variable for student first-generation status, i.e. first in family to go to college*
- XC : *School-level categorical variable for urbanicity of the school, i.e. rural, suburban, etc.*
- X1 : *School-level mean of students' fixed mindsets, reported prior to random assignment*
- *X2: School achievement level, as measured by test scores and college preparation for the previous 4 cohorts of students*
- *X3: School racial/ethnic minority composition, i.e., percentage of student body that is Black, Latino, or Native American*
- X4 : *School poverty concentration, i.e., percentage of students who are from families whose incomes fall below the federal poverty line*
- *X5: School size, i.e., total number of students in all four grade levels in the school Post-treatment outcome, a continuous measure of achievement*
- *W: Treatment, i.e., receipt of the intervention*

  Our analysis is based on data from $n = 10,391$ children from a probability sample of $J = 76$ schools. For each child $i = 1, \ldots, n$, we observe a binary treatment indicator W_i, a real-valued outcome Y_i, as well as 10 categorical or real-valued covariates described in Table 1. *We expanded out categorical random variables via one-hot encoding, thus resulting in covariates $X_i \in \mathbb{R}^p$ with $p = 28$ Given this data, the workshop organizers expressed particular interest in the three following questions:*

  1. *Was the mindset intervention effective in improving student achievement?*
  2. *Was the effect of the intervention moderated by school level achievement (X2) or pre-existing mindset norms* (X1)? *In particular there are two competing hypotheses about how* X2 *moderates the effect of the intervention: Either it is largest in middle-achieving schools (a "Goldilocks effect") or is decreasing in school-level achievement.*
  3. *Do other covariates moderate treatment effects?*

Exercise 1:

1. Simulate the data set as described above. The outcome is determined solely by the included covariates and the treatment W. You may choose the scale and range of all variables, as well as the shares represented in the categorical variables.

2. Compare the results from a linear regression estimate of the treatment effect and the causal (random) forest to answer question 1 from above.

Exercise 2:

Induce moderation of the treatment effect and evaluate the relative performance of causal (random) forests and a linear model with simple interactions for four different sample sizes.

1. Sequentially introduce higher level interactions in the DGP to reproduce a scenario outlined in question 2.

2. Introduce other interactions to evaluate the scenario in question 3.