The aim of this exercise: Compare prediction properties of the linear regression model with a regression tree.

Exercise 1:

Consider the following data generating process in which we have $n = 100$ observations and two covariates $X_1 \sim \mathcal{N}(0, 4)$ and $X_2 \sim \mathcal{N}(0, 4)$. y_i is generated by some nonlinear function of $\mathbf{X}$ of your choice.

- Generate the data according to the dgp described above and fit a regression tree.
- Use a newly generated test data set to calculate the mean squared error using a naive linear regression model and compare with a full tree and an optimally pruned tree.

Exercise 2 (Simulation Study)**:**

The goal here is to think about how a regression tree makes its predictions and consequently when a regression tree might yield a better result than other methods.

- Propose a dgp that will be well suited for analysis using regression trees and evaluate relevant properties in a small simulation study.
- Propose a dgp that will be well suited for analysis using the linear regression model (i.e. where the linear regression model is more likely to "beat" the regression tree method) and evaluate.
- Show how pruning reduces the variance of the regression tree prediction.