Introduction to Data Science Homework 4

This question will explore various ways to expand the standard linear regression methods. In particular, we will use the method of subset selection, and also fit ridge and lasso to the data. The dataset has 3 inputs parameters ($x1, x2, x3$). The output y depends on different input parameters to different extents. (The code used to generate this data is provided.)

1. First, we will use the best subset selection method to find the best linear model to describe the data. Fit every possible linear model to the data, with every combination of input parameters (there should be 8 of them). For each model, calculate the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). From the list of models and their corresponding AICs and BICs, can you tell how strongly each input parameter affects the output?

2. Now, we will explore the ridge and lasso. Fit ridge and lasso to the data. Make sure the input parameters are normalized properly. Vary the value of λ in a sufficiently wide range to see the full range of behavior (λ should start from zero, up to large enough value that all regression coefficients are squeezed to zero). Plot the following quantities as a function of λ.

    - Coefficients β of each input parameter.
    - MSE, for both training and test datasets.