Statistical Data Analysis
Linear Regression and Feature Selection Tutorial

Introduction The goal of this tutorial is to get familiar with advanced methods of linear regression: Ridge regression and the LASSO (least absolute shrinkage and selection operator). You should learn to understand what can make them preferrable to ordinary least-squares linear regression and how they should be used properly.

The chapter on Linear Regression in "An Introduction to Statistical Learning" book is relevant to this tutorial and this assignment is inspired by the Lab section in that chapter. If you feel lost, reading that chapter may be helpful.

Data We will analyze a dataset of death rates depending on various factors (variables) in 60 different regions. In the end we should obtain a model predicting death rate based on the relevant variables and discover which variables are in fact irrelevant. The dataset is in the file `data.csv`. If you are interested, you can look up the meanings of the variables in the file `data_labels.txt`.

Assignment You are provided with the R script `assignment.r`. You are expected to complete the 4 tasks that are written in the comments of the script and submit a modified version of the script along with a PDF report with written asnwers to the questions. First of all, open the assignment script and go through the code to make sure that you understand how the data are loaded and split to training and testing sets and how the sample LASSO model is generated.

Task 1 (1 Point) There is a methodological error in the block of code which generates the lasso model. Find it and correct it. Hint: The error causes the variable `lasso.coefficients` contain values of lesser precision than what we could get from the data. Print the coefficients onto the output.

Task 2 (1 Point) Implement analogous fitting method for Ridge regression. Compute the Mean Squared Error for Ridge regression, LS and LASSO and compare them. Print the MSE values onto the output.

Task 3 (2 Points) Assume we want LASSO to select exactly 2 variables while still minimizing MSE. What is then the desired parameter λ (with 1e-1 precission)? What are the variables? What is the MSE? Print the values onto the output. Implement an algorithm to determine that. Check if the selected variables are the same as the ones exhaustive subset search would select. You may use the `regsubsets` function from the `leaps` library to do this or implement the search yourself for subsets of size 2.