Problem Set 8

Ridge Regression

Consider the following data generating process with $n = 500$ observations and K covariates. Initially set the number of predictors $K = 10$ and $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. $\mathbf{\Sigma}$ is (initially) diagonal and contains values ranging from 1 to 10 in some sequence.

The true coefficients range from $\boldsymbol{\beta} = 0 - 0.3$ (you can sample values from that range or use equi-spaced values on that interval) and the errors are drawn from a normal distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$.

Exercise 1: Initial set-up and implementation

The aim of this exercise: compare the predictions generated by OLS, ridge regression and the Lasso.

- Implement the ridge regression estimator and the lasso using glmnet for a grid of different penalty parameters.
- Draw a test data set and plot the test error for the different values of λ for both ridge regression and Lasso.
- Choose the optimal λ using the build in cross-validation function from glmnet and calculate the mean squared test error drawing a new test data set and compare ridge regression and OLS.

```
###For calculating the ridge regression coefficients you need to install
###the glmnet package and call the library#############################
library(glmnet)
#####For calculating the prediction error you can use the predict() function##
```

Exercise 2: Simulation Study

- Set at least one coefficient in $\boldsymbol{\beta}$ equal to zero (unless this was already the case). For 100 simulation runs, record how many times the Lasso selects the correct model, i.e. where $\hat{\beta}_L$ is set to zero correctly.
- What part of the data generating process could you change that would make the lasso perform worse (in terms of the prediction error) than ridge regression? Confirm your intuition in a simulation study.