

Exercise

Lasso Regression

Load the data `Hitters` from the package `ISLR`. This means that you first need to install the package with

```
install.packages("ISLR")
```

and then load the data with

```
data(Hitters, package="ISLR")
```

Look at `?Hitters` or at `str(Hitters)` for more detailed information. Remove all observations which contain missings by using the command `na.omit()`.

Our goal is to find a model which allows to predict the variable `Salary`, i.e. the salary of the players on the basis of various statistics associated with performance in the previous years.

For the following tasks, split the data randomly into training and test data (about equal halves), build the model with the training data, and evaluate for the test data (using the MSE as a criterion).

Look first at your data. Is any preprocessing necessary or useful?

1. *Multiple linear regression model:* Estimate the full regression model and interpret the results.
2. *Lasso Regression:*
 - (a) Use the function `glmnet()` from the `library(glmnet)` and apply it to the training data. Plot the result object. How can you interpret the plot? Which default parameters are used for `lambda`? What is the meaning of the parameter `alpha`?
 - (b) Use the function `cv.glmnet()` and apply it to the training data. Visualize and interpret the results. How do you obtain the optimal tuning parameter and the regression coefficients?
 - (c) Use the optimal model to predict the Salary for the test data. Compare the predictions with the reported values graphically and with the MSE. Compare with the results from Ridge regression.

Compare the obtained models by calculating the MSE for the test data. Which model is preferable?