

Master BeNeFri in Computer Science

Course: Statistical Learning Methods
Spring 2016

Exercise #4: Linear regression with R

1. In the folder Exercise#4, we have the file EducationBis.txt containing a new version of the data of Exercise #1 (mainly without outliers). Moreover, the variable `Gender` is coded as “male” or “female”. We still want to explain the variable `Wage`. Build a linear model for men and another for women. Explain the results of the output provided by the function `lm()` for both models.

Download from the ILIAS website the dataset `Computer` dataset (filename: `ComputerData.txt`). This dataset is composed by various variables that can be used to predict the performance of a computer system. The description of the data is given in the pdf file “`ComputerDescription.pdf`”. The performance of the system (response) is indicated either by the variable `ERP` or `PRP`. We will use the target variable `PRP`.

2. Check the different variables (predictors) you have to predict the system performance (response). Which variables do you think cannot be used to “explain” the system performance?

3. You’re allowed to use only a single variable (predictor) to predict the value of `PRP`. Which one do you use? Does your model explain something? What is the confidence interval around the slope?

4. Visualize graphically the (linear) relationship you have found.

Download from the ILIAS website the dataset `Cars2` dataset (filename: `Cars2Data.txt`). This dataset is composed by various variables that can be used to predict the performance of the car. The description of the data is given in the pdf file “`Cars2Description.pdf`”. The performance of the system (response) is indicated by the variable `mpg`.

5. Check the different variables (predictors) you have to predict the system performance (response). Which variables do you think cannot be used to “explain” the system performance?

6. You’re allowed to use only a single variable (predictor) to predict the value of `mpg`. Which one do you use? Does your model explain something? What is the confidence interval around the slope?

7. Visualize graphically the (linear) relationship you have found.