DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

Project 2

Data Analysis and Statistical Modelling

(ADME, 2023/2024 (P1))

Handed out on 15 Jan 2023.

To be handed back on 5 Jan 2024.

Consider for **Auto** data set, available in R **library(ISLR)**, all variables except **name** and select the subset from observation **1** to **50**.

1. Make a exploratory analysis, using plots and summary statistics, to describe the data.
2. One researcher has rudimentary knowledge about multiple linear regression analysis and wants your help to find a way to explain the variable **mpg** with some predictors variables.
   (a) Fit a regression model to this data set. Test for significance of the regression. Is there any evidence that a subset of the original variables should be excluded from the model? Proceed in order to find the best subset of regressors. Evaluate the results taking in account the p-values and coefficients of multiple determination.
   (b) Check model adequacy, investigate possible influential/leverage observations and outliers.
   (c) Calculate 97.5% confidence interval (CI) for the expected value of the responses for observation **14** and for observation **31**. For the same values of the regressors, and the same confidence level, calculate the prediction interval (PI). Compare and discuss the obtained results.

About the report:

The report should not exceed 10 pages. Do not forget to include: introduction, the dataset in study, objectives of study, decisions, conclusions and bibliography. The R code and the report must be upload in fenix.