Report Project 2

Data Analysis and Statistical Modeling

Prof Isabel Rodrigues

Uma imagem com captura de ecrã, design, círculo, texto

Descrição gerada automaticamente

**Grupo 1**

João Matos nº98949

Ana Pinto nº102949

Marina Nóbrega nº103880

Manuel Mota Dias nº96056

Maria Freitas nº96757

**Introduction**

For this project a datagram is given for us to analyse:

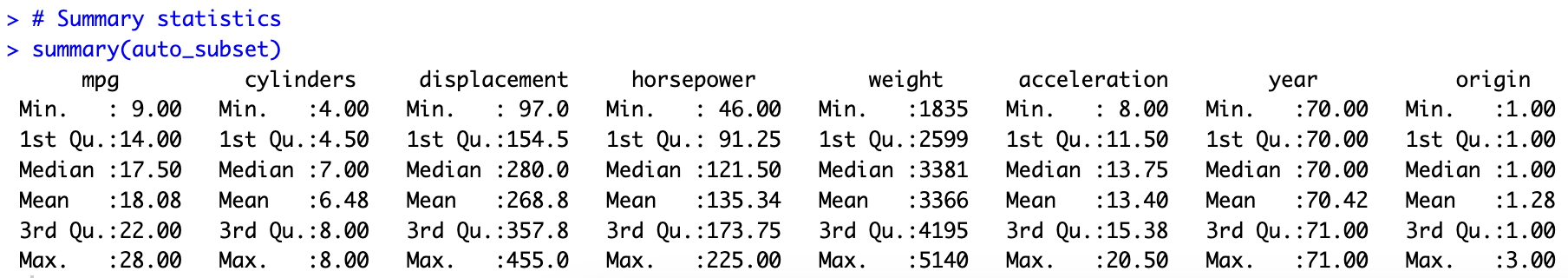
**Data frame**: Auto

**Subset**: observation 1 to 50

**Variables**: all except name:

* mpg – miles per gallon
* cylinders – number of cylinders between 4 and 8
* displacement – engine displacement (cu. Inches)
* horsepower – engine horsepower
* weight – vehicle weight (lbs.)
* acceleration – time to accelerate from 0 to 60 mph (sec.)
* year – model year (modulo 100)
* origin – (origin of the car (1. American, 2. European, 3. Japanese)

**1**

**Summary statistics**

Uma imagem com texto, Tipo de letra, branco, algebra

Descrição gerada automaticamente

Uma imagem com texto, captura de ecrã, Tipo de letra, branco

Descrição gerada automaticamente

**Total Variance**

Uma imagem com Tipo de letra, texto, branco, Gráficos

Descrição gerada automaticamente

**Summary Plots**





**2.A)**

To find the best subset of regressors, we applied the regression model, removing the predictor with the highest Pr(>|t|) until we got the ones that we considered useful:

Uma imagem com texto, Tipo de letra, captura de ecrã, branco

Descrição gerada automaticamenteUma imagem com texto, Tipo de letra, captura de ecrã, branco

Descrição gerada automaticamenteUma imagem com texto, Tipo de letra, branco, captura de ecrã

Descrição gerada automaticamenteUma imagem com texto, Tipo de letra, branco, captura de ecrã

Descrição gerada automaticamente

Uma imagem com texto, Tipo de letra, branco, algebra

Descrição gerada automaticamente

In the end, cylinders, weight, and acceleration are the selected ones. We chose

these because it would allow us to work with less predictors. The adjusted **r²** value of the last two iterations is similar and lower in the last iteration.

After fitting a regression model to explain the mpg variable using the predictors we just selected we get the values:

**r²** = 0.891092607413735

**r²adj** = 0.8839899513755

Given that our **r²** and **r²**adj values are relatively high (>0.8), it suggests that the current model explains a significant portion of the variability in the response variable.

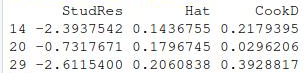
**2.B)**

For this regression we are using p = 3 predictors for n = 50 observations

Searching for possible influential/leverage observations we get this:

**influenceIndexPlot(reg\_mpg)**

**influencePlot(reg\_mpg)**

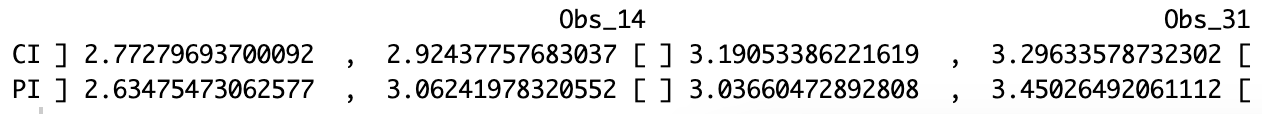




For mpg, the possible leverage observations are 29, 20 and 14. The two observations with highest cook’s distance are 14 and 29, so the more possible influential observations are 29 and 14.

**2.C)**

Calculating the 97.5% Confidence Interval and Prediction Interval for the expected values of the responses for observations **14** and **31** we get:



The prediction interval (PI) for Obs\_14 is noticeably wider than the confidence interval (CI). This wider width in the prediction interval reflects the additional uncertainty associated with predicting individual observations, considering both the uncertainty in estimating the mean and the variability of individual observations.

Similar to Obs\_14, the prediction interval (PI) for Obs\_31 is wider than the confidence interval (CI). This wider width suggests a higher level of uncertainty when predicting individual observations, considering both the variability in estimating the mean and the variability of individual data points.

In both cases, the prediction intervals are wider than the corresponding confidence intervals. This is a common characteristic, as prediction intervals need to account for the variability in individual observations, making them more conservative and wider. The confidence intervals, on the other hand, primarily focus on the uncertainty around estimating the mean.

In summary, the widths of the prediction intervals highlight the increased uncertainty when making predictions for individual observations compared to estimating the mean.

**Conclusion**

To finish, this analysis provided a comprehensive understanding of the relationships within the Auto dataset and helped us learn more about regression models and their uses.