Market Forecasting Assignment: Multiple Linear Regression

Teacher: Dr. Stefan Groesser, Dr. Christoph Imboden

Autors: Benedech Rodolpho, Buehler Pascal, Geeler Ken 04.12.2021

Short Theory Recap Multiple Linear Regression (MLR)

The model equation for MLR has the form $y = X\Theta + \epsilon$,with y being the target vector, X the explanatory Matrix,Θ the parameter vector and the residuals with the assumption that they are normally distributed $\epsilon \sim N(0, \sigma^2)$. The estimation of Θ is done by solving the equation $\Theta = (X^T X)^{-1} X^T y$.

Dataset

The data from 7 common fish species measured on a fishmarket is provided from Kaggle.

https://www.kaggle.com/aungpyaeap/fish-market

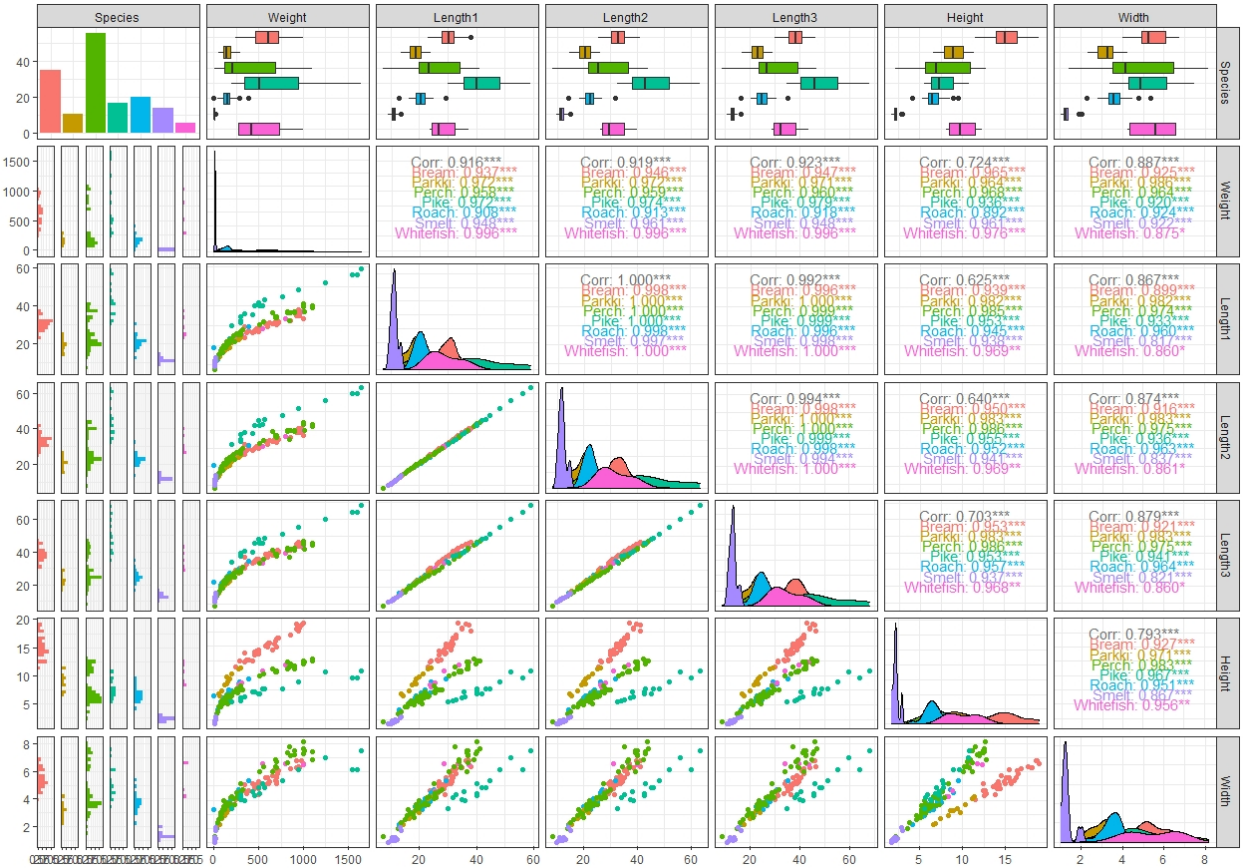

Figure 1: pairwise plot of the fish dataset

Data exploration

Information of the Dataset is gained with the following commands. *Note: the output is not shown to reduce space in the document*

```
fish_dat=readr::read_csv("Fish.csv")                          #  reading as tibble
GGally::ggpairs(fish_dat,aes(color = Species)) + theme_bw() #  pairwise plot
summary(fish_dat);str(fish_dat)                               #  statistics and structure
```

The pairwise plot Figure 1 provides a visual overview of the dataset. It consists of 159 observations measured for 6 numerical variables which measure properties of the fish and one categorical variable to describe the Species. The coloring in the pairwise plot allows a distinction of the Species, one can immediately observe that the species Perch has the most observations.

From the pairwise scatterplots a nonlinear relation between weight and the other numerical variables is observed, therefore the correlation *cannot* be correctly interpreted with these variablepairs. Further the other numerical variables are linearly correlated (the correlation statistic can be interpreted), interestingly the 3 length measurements have a very strong correlation to each other.

The density plots helps see how the variables themselves are distributed. For example the width of Smelt tends to have a bidmodal distribution, altough this has to be taken with a grain of salt due to the sparse data. The distribution of the target variables is actually very important since MLR assumes multivariate normality of the data, for other distributions we might choose another model such as a gamma link to improve the model validity, but for this analysis we stick to MLR.

Also the pairwise boxplots gives a good overview for comparison between the different variables. For example it can be seen that the variance of the species vary amongst the levels.

Building a Model

The dataset provides many different options to do linear regression, nevertheless for MLR there is a problem with multicollinearity since a lot of the varibles are strongly correlated, meaning that if multiple independet variables would be in the modell it wouldnt be clear which one explains the effect the best for the target variable.