**Linear Models**

**Assignment – 1**

**Name:** Patil Urvi Milind

**PRN:** 22060641044

**Batch:** 2022 – 24

**Multiple Linear Regression**

**Aim:**

To fit multiple linear regression model and check whether the Weight of the fish is dependent on various factors like Vertical Length (Length1), Diagonal Length (Length2), Cross Length (Length3), Height and Diagonal Width.

**Dataset:**

* The dataset is taken from the open-source platform Kaggle.
* The dataset is a record of 7 different fish species in fish market sales.
* The data contains 159 observations for 7 variables out of which 1 is categorical variable and other 6 are numerical variables.
* There are five independent (input) variables and one dependent (output) variable.
* The dependent variable is Weight and independent variables are Length1, Length2, Length3, Height & Width.

**Methodology:**

* The first step is importing the dataset in R.
* The following step is to determine the data type of each variable, which is done in R by using the str () command.
* Then the descriptive statistics is found using summary () command. It helps us to give the minimum value, 1st quartile, median, mean. 3rd quartile & maximum value.
* The variance and standard deviation of the data is found using var () and sd () command.
* The cor () command gives the correlation coefficient value between two variables. Here, it is used it to find correlation between the dependent variable and all independent variables.
* Then, we try to fit a multiple linear regression model with Weight as the dependent variable & Length1, Length2, Length3, Height & Width as the independent variables.
* The summary () command gives the Adjusted R-square value and the p-value, which helps us in determining whether the model is a good fit for the data or not & whether the data is significant or not.
* The next step involves plotting the model with the help of plot () command and checking for the assumptions of linear regression.
* We finally conclude whether the model is a good fit or not.

**Interpretation:**

* Correlation between all variables is found out. The variables Length1, Length2 and Length3 are highly correlated (>0.9) and multicollinearity is present in the data. Hence, Ridge Regression has been performed later.
* The multiple linear regression model has been fitted.

The Adjusted R-squared value is 0.88, which is > 0.5. This suggests that the model is good fit for data. 88% error in the dataset can be explained by the model.

The p-value is <2.2e-16, which is <0.5 and hence data is significant.

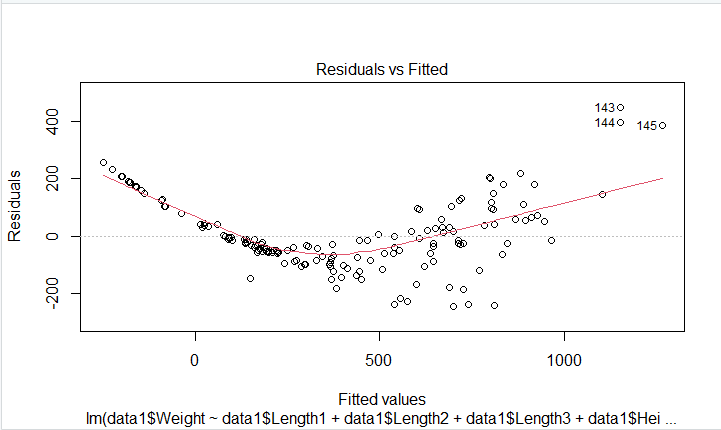
The regression equation obtained is as follows:

Weight = - 499.587 + 62.355\*Length1 - 6.527\*Length2 - 29.026\*Length3 + 28.297\*Height + 22.473\*Width

* The assumptions of residual analysis are checked by using the plot () command.

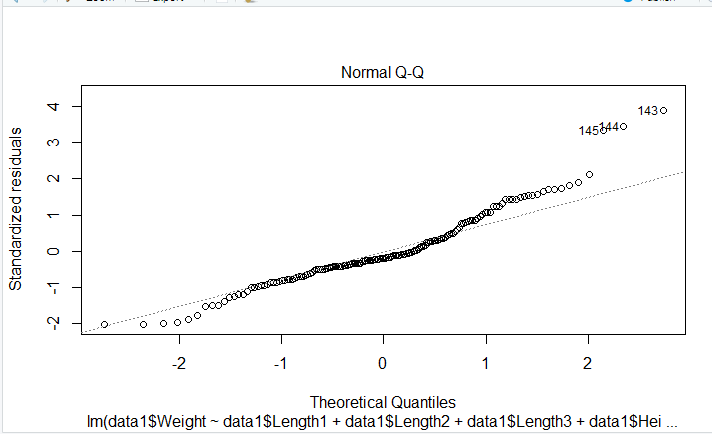
1. **Linearity:**

The Residuals vs Fitted plot is used to check linearity. From Residual VS Fitted plot, we do not find equally spread residuals around horizontal line. We have non-linear relationship. Hence the linearity assumption is not satisfied.

****

1. **Normality:**

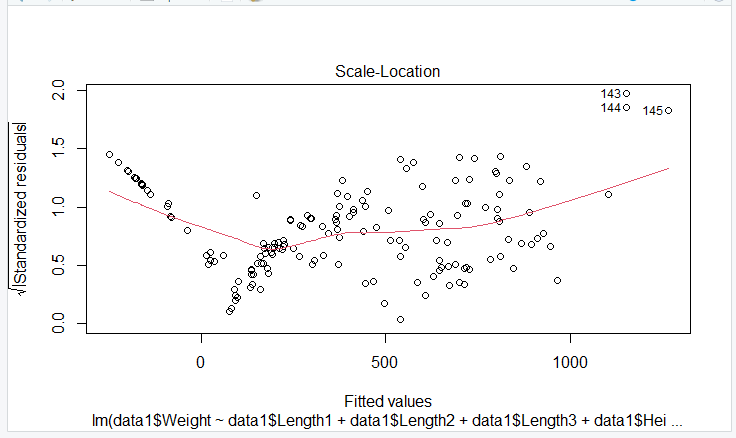
Normal QQ plot is used to check the condition of normality. From Normal QQ plot, we observe that the observations are deviating at the end and hence the residuals are not normally distributed. Hence normality assumption is not satisfied.

****

1. **Homoscedasticity:**

Scale-Location plot is used to determine homoscedasticity. From Scale Location plot, we are unable to locate a horizontal line. Therefore, homoscedasticity is present. Hence the assumption of homoscedasticity is satisfied.

The conclusions regarding presence of homoscedasticity can also be seen using BP Test.

****

1. **Outlier detection:**

From Residual VS Leverage plot, we tend to look for outliners. From the graph we can see that there are no outliers present since none of the values lie outside the Cook”s distance. Thus, the data has no outliers.

****

* Autocorrelation is checked by using Durbin-Watson test. The test statistic is lower than one (DW = 0.42372) and the p-value (< 2.2e-16) is significant. Hence, we reject the null hypothesis and conclude that the residuals are autocorrelated.

**Conclusion:**

* The assumptions that are not satisfied are linearity and normality.
* The assumptions that are satisfied are homoscedasticity and the absence of outliers.
* Multicollinearity is present in the dataset. Hence, we have used Ridge Regression.
* Autocorrelation is present.

**R-Code and Output:**

#Importing dataset in R

library(readxl)

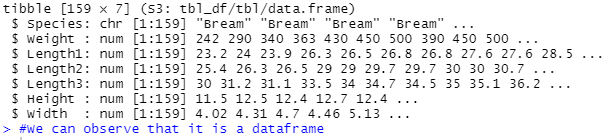
data1 <- read\_excel("C:/Users/urvip/OneDrive/Desktop/data1.xlsx")

View(data1)

#We observe that there are 159 obs. of 7 variables

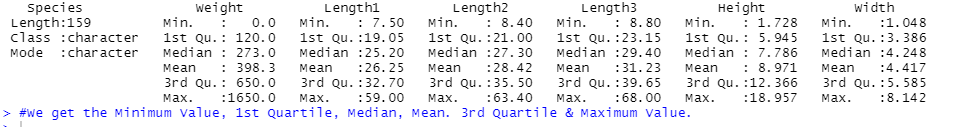
#Finding Structure of Dataset

str(data1)



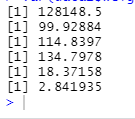
#Finding Descriptive Statistics

summary(data1)



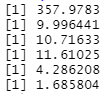
#Variance

var(data1$Weight);var(data1$Length1);var(data1$Length2);var(data1$Length3);var(data1$Height);var(data1$Width)



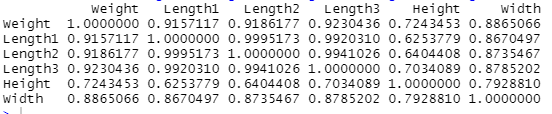
#Standard Deviation

sd(data1$Weight);sd(data1$Length1);sd(data1$Length2);sd(data1$Length3);sd(data1$Height);sd(data1$Width)



#Correlation

cor(data1[,c("Weight","Length1","Length2","Length3","Height","Width")])

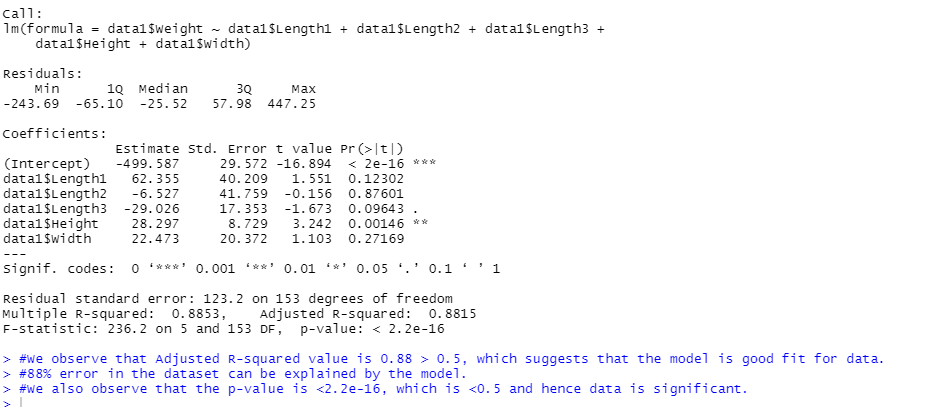


#Fitting the model

#Multiple Linear Regression

model1<-lm(data1$Weight~data1$Length1+data1$Length2+data1$Length3+data1$Height+data1$Width)

summary(model1)



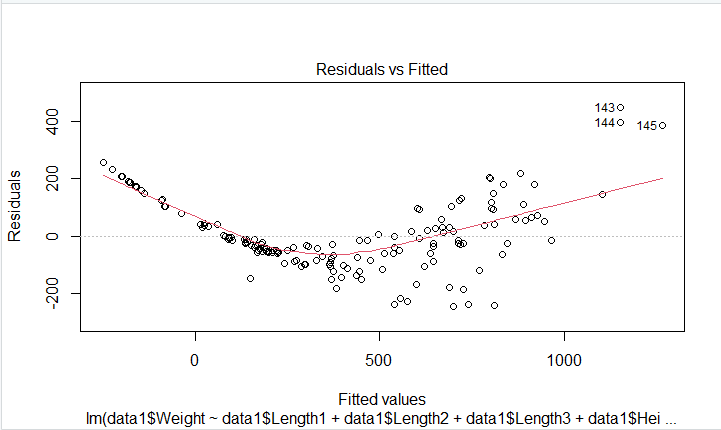
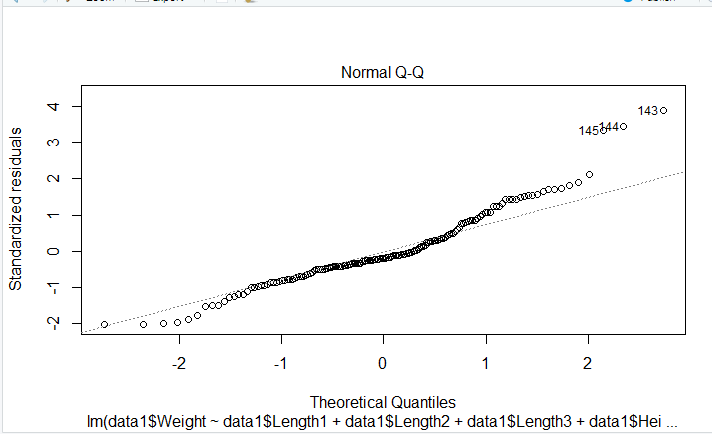
plot(model1)

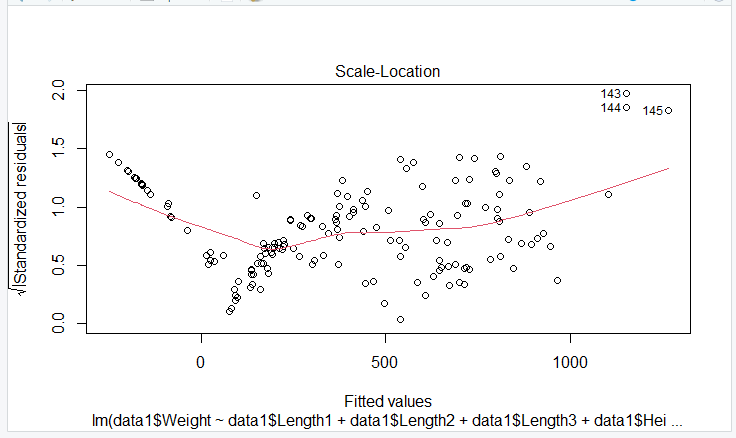
#Residual vs Fitted : Since we do not find equally spread residuals around horizontal line. We have non-linear relationship.

#Normal Q-Q plot : Since the observations are deviating at the end, the residuals are not normally distributed.

#Scale-Location : Since we are unable to locate a horizontal line, homoscedasticity is present.

#Residual vs Leverage : We tend to look for outliners. There are no outliers present since none of the values lie outside the Cook”s distance.

**** ****

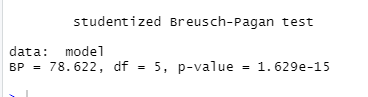
**** ****

#Homoscedasticity

#BP Test

library(lmtest)

lmtest::bptest(model)



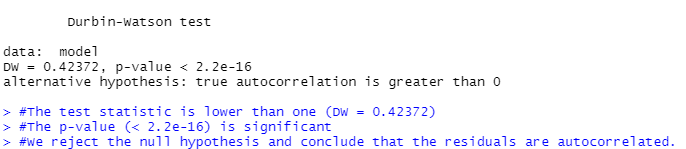
#Autocorrelation

#Durbin-Watson Test

library(lmtest)

model<-lm(data1$Weight~data1$Length1+data1$Length2+data1$Length3+data1$Height+data1$Width)

lmtest::dwtest(model)



#Ridge Regression

