**Task 2 (Individual Task) — Exploratory Data Analysis and Linear Models**

**Question 1:**

Use R to build a scatter matrix use ggpairs(). Identify and discuss any particularly strong or weak correlations. Look further into the relationship between X1 = total length and X3 = length of beak and head and discuss whether these two variables might be sufficient to identify survived and died birds.

**Code:**

library(tidyverse)

bird = read\_csv("birds.csv") #Read the input file

colnames(bird) #print coloumn names

head(bird) #print first 6 rows

# Arrange rows by died/survival and then total length

bird = bird %>%

group\_by(GROUP) %>%

arrange(X1) %>%

ungroup()

bird #print first 10 rows

print('/n Individual Task -- part (1) scatter matrix')

library(GGally)

ggpairs(bird,aes(color=bird$GROUP))

print('/n individual Task -- part (1) X1 and X3 ')

ggplot(bird,aes(x=X1,y=X3))+

geom\_point(aes(colour=GROUP))

ggplot(bird,aes(x=X3-X1,colour=GROUP))+

geom\_boxplot()

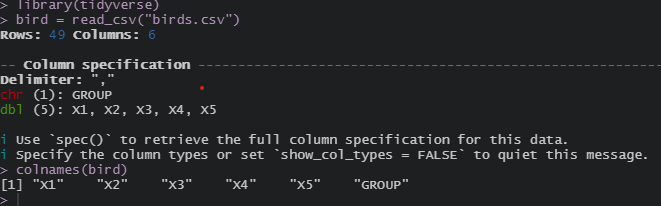
ggplot(bird,aes(x=X3,colour=GROUP))+

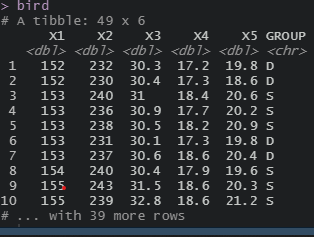
geom\_boxplot()

ggplot(bird,aes(x=X1,colour=GROUP))+

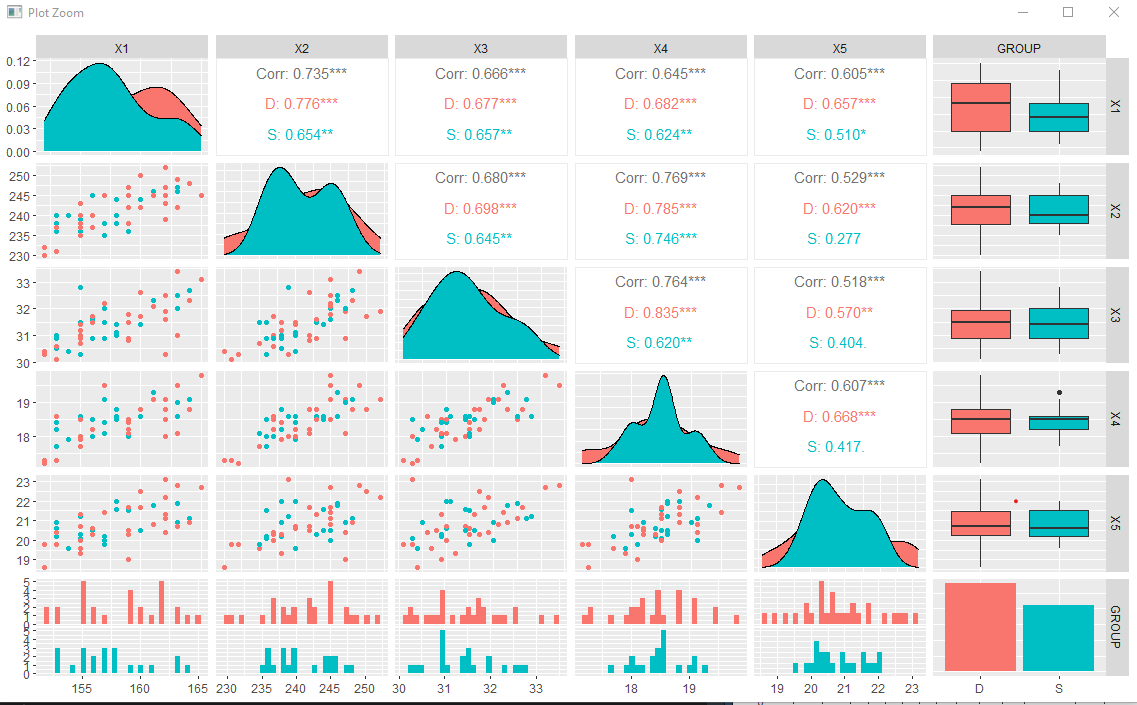
geom\_boxplot()

**Output:**





**Plot:**



The diagonal elements are curve fitting for each variables. Red curve is plotted for Died sparrows and the green curve is plotted for Survival. Lower diagonal elements are scatter plot of pair of the data. Upper diagonal elements are correlation between the corresponding variables.

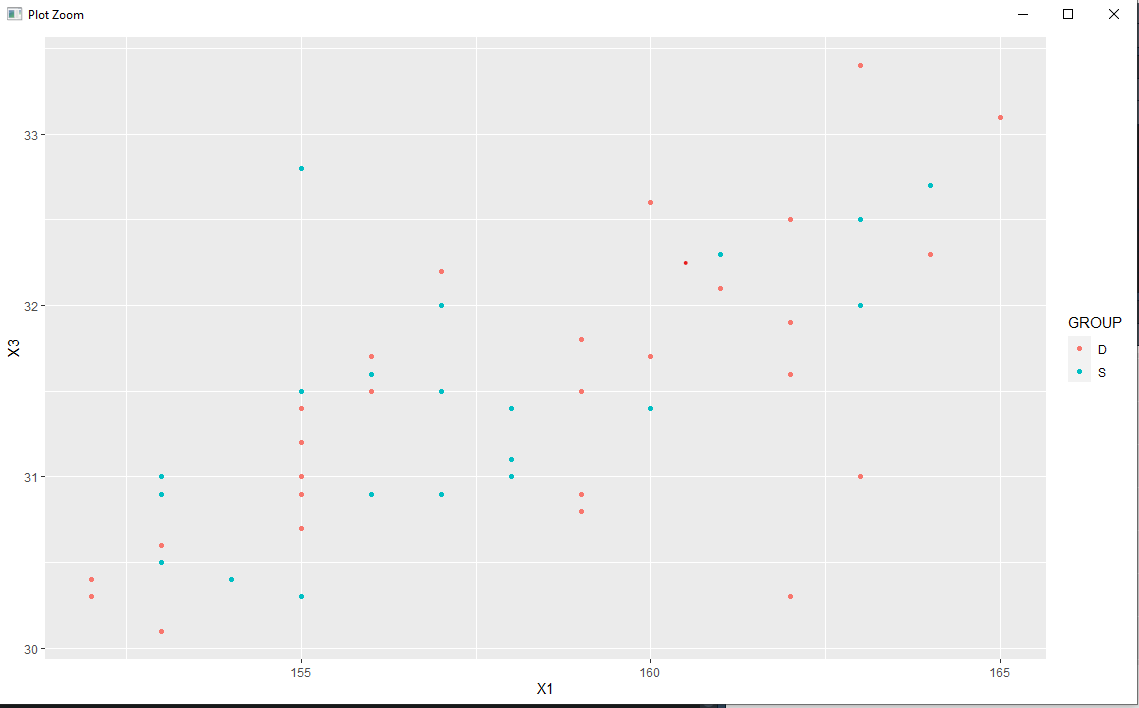
By using ggpairs plot,we can easily identify the strong and weak correlation.

* With respect to died and survival: The strongest correlation is 0.835 between X4 and X3 i.e; for died sparrows and it has more significance . it means there is a linear relationship between X4 and X3. The weakest correlation is 0.277 between X5 and X2 i.e; for survival sparrows and it is not significant.
* With respect to the 5 bird measurements : The strongest correlation is 0.769 between X4 and X2 with more significance. The weakest correlation is 0.518 between X5 and X3 with more significance

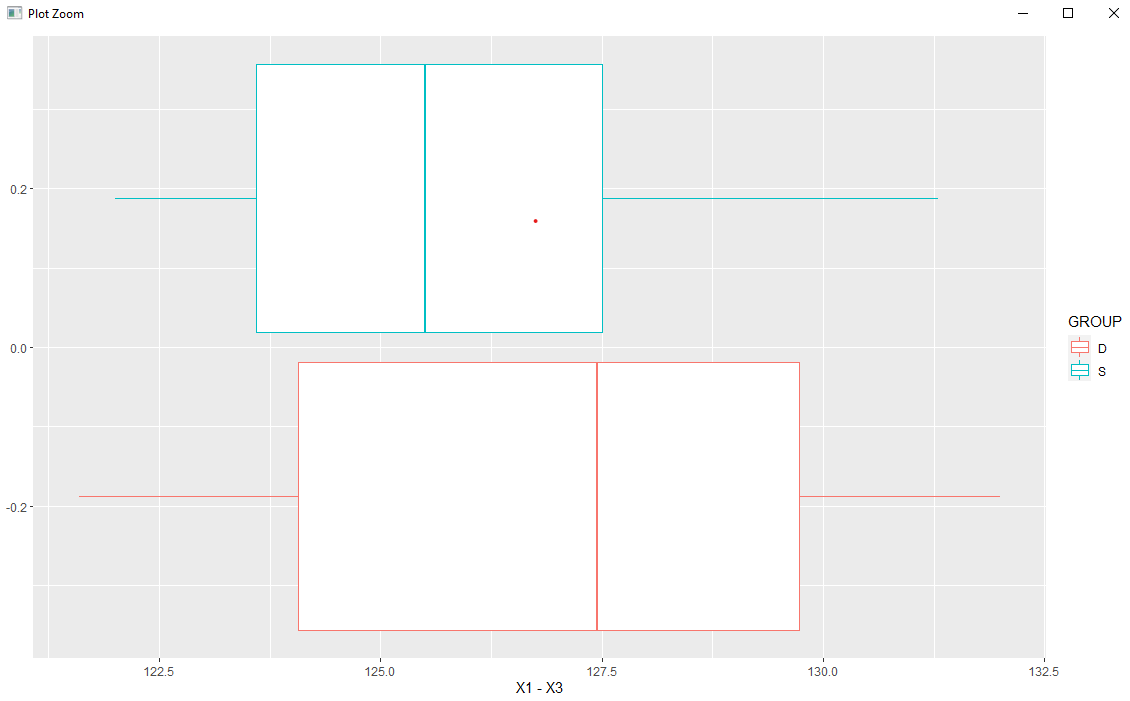
**Output**:

[1] "/n individual Task -- part (1) X1 and X3 "

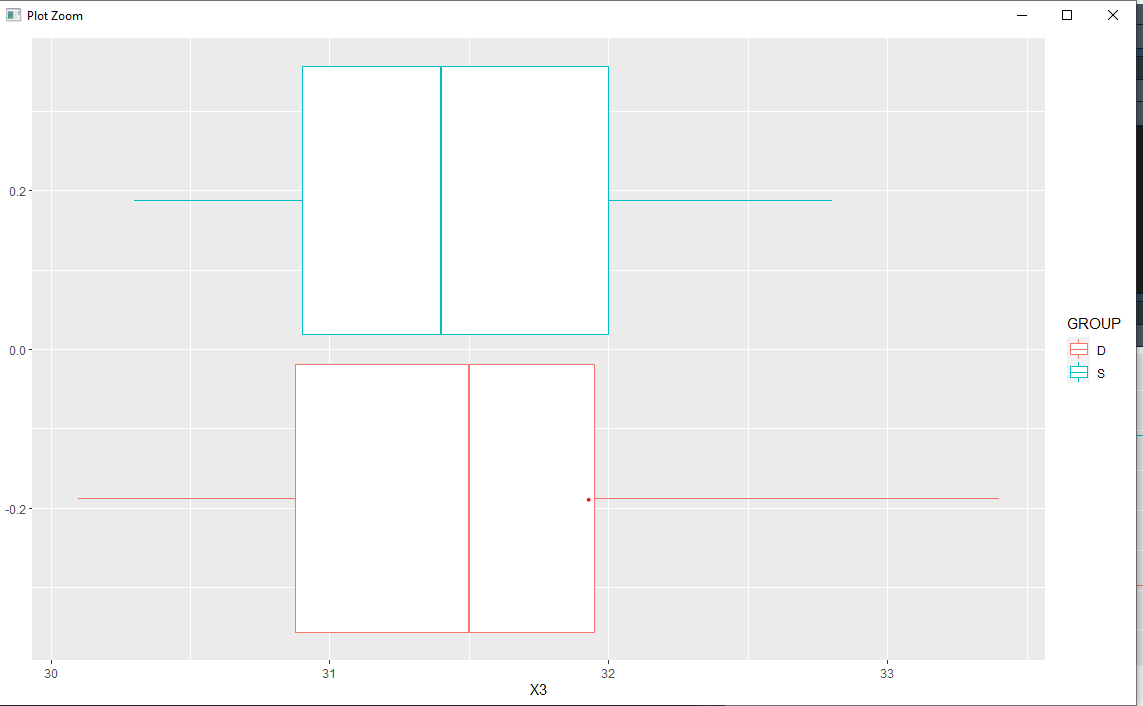
* Scatter Plot between X1 and X3 with respect to Group ( Died / Survival)



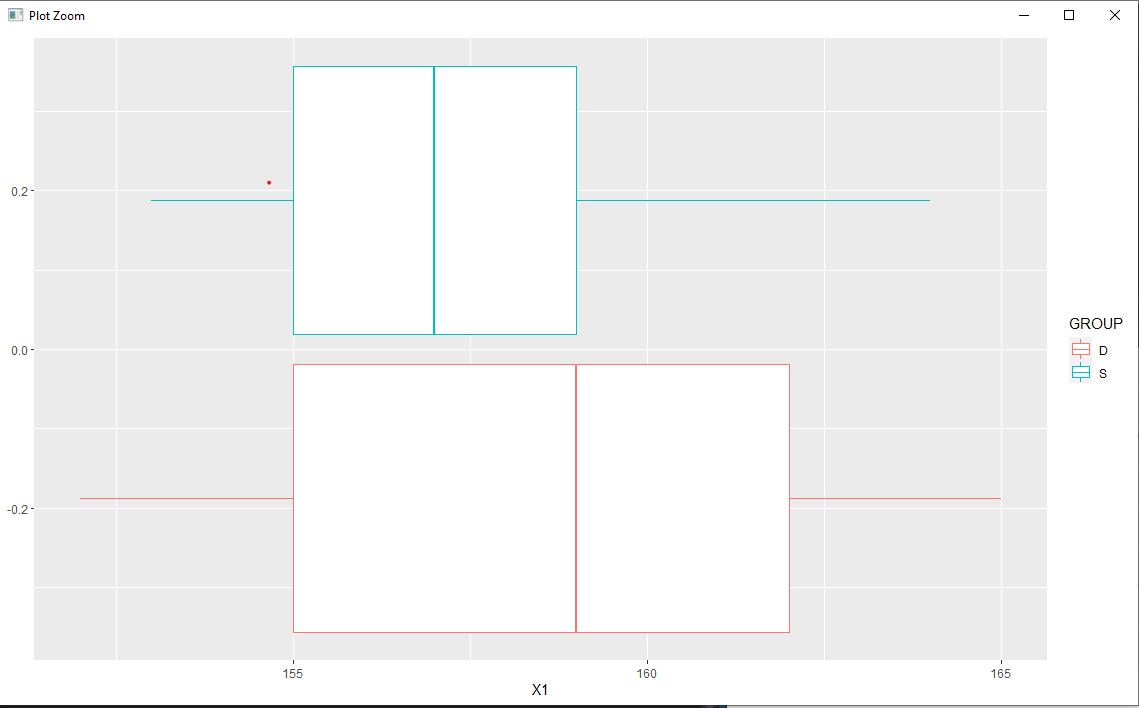
* Box Plot for X1-X3 with respect to Group ( Died / survival)



* Box Plot X3 with respect to Group ( Died / survival)



* Box Plot X1 with respect to Group ( Died / survival)



* By observing the 3 box plots ,
* X1-X3 , X1 and X3 – the range of values for died are more compared to survived.
* So with this I can conclude that X1 and X3 are sufficient to identify the died and survived birds.

**Question 2:**

Which single-predictor linear model “best” predicts total length and which single-predictor linear model “best” predicts length of beak? Justify your answers and critically assess any linear models you recommend.

**Code:**

library(stats)

library(olsrr)

library(ggfortify)

library(tidyverse)

bird = read\_csv("birds.csv")

colnames(bird)

#Best one-predictor model of X1 :

#Univariate linear regression

#predictor is X2,response variable is X1

model = lm(X1~X2, data=bird)

#distPred <- predict(model, bird)

summary(model) #R2 = 0.54 (correlation coefficient of y and yhat)

AIC(model) #AIC = 232.97

autoplot(model)

#smaller AIC model is good

model = lm(X1~X3, data=bird)

#distPred <- predict(model, bird)

summary(model) #R2 = 0.44 (correlation coefficient of y and yhat)

AIC(model) #AIC = 242.27

autoplot(model)

model = lm(X1~X4, data=bird)

#distPred <- predict(model, bird)

summary(model) #R2 = 0.41 (correlation coefficient of y and yhat)

AIC(model) #AIC = 244.65

autoplot(model)

model = lm(X1~X5, data=bird)

#distPred <- predict(model, bird)

summary(model) #R2 = 0.366 (correlation coefficient of y and yhat)

AIC(model) #AIC = 248.70

autoplot(model)

#best one-predictor model of X3:

model = lm(X3~X1, data=bird)

summary(model) #R2 = 0.44

AIC(model) #AIC = 92.73

autoplot(model)

model = lm(X3~X2, data=bird)

summary(model) #R2 = 0.46

AIC(model) #AIC = 91.14

autoplot(model)

model = lm(X3~X4, data=bird)

summary(model) #R2 = 0.583

AIC(model) #AIC = 78.54

autoplot(model)

model = lm(X3~X5, data=bird)

summary(model) #R2 = 0.583

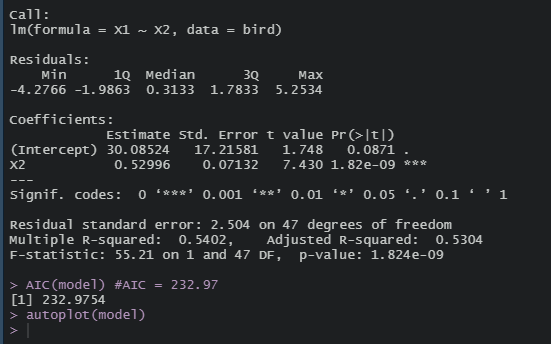
AIC(model) #AIC = 106.21

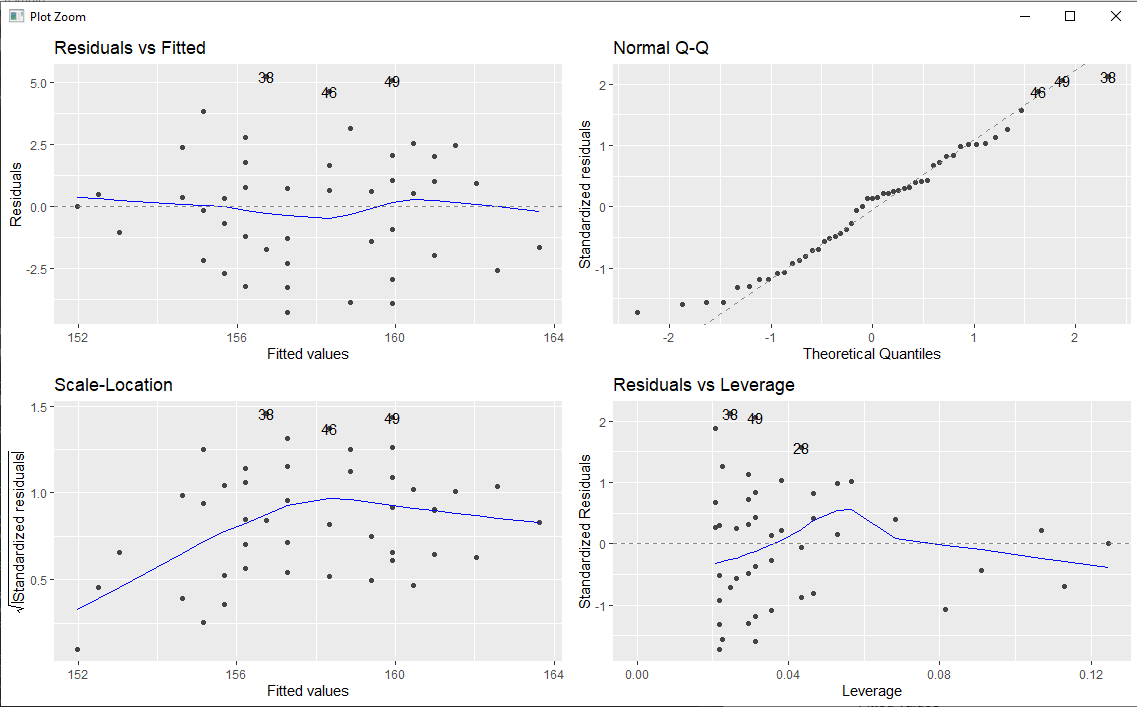
autoplot(model)

**Output & Plots**

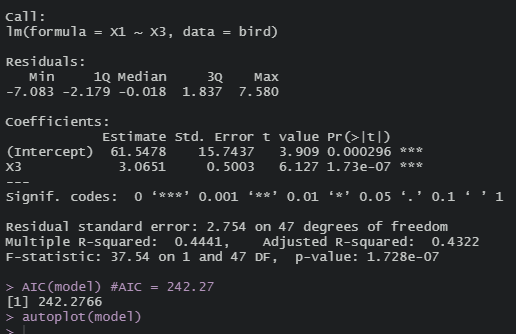
Single-predictor linear model to predict Total Length -X1

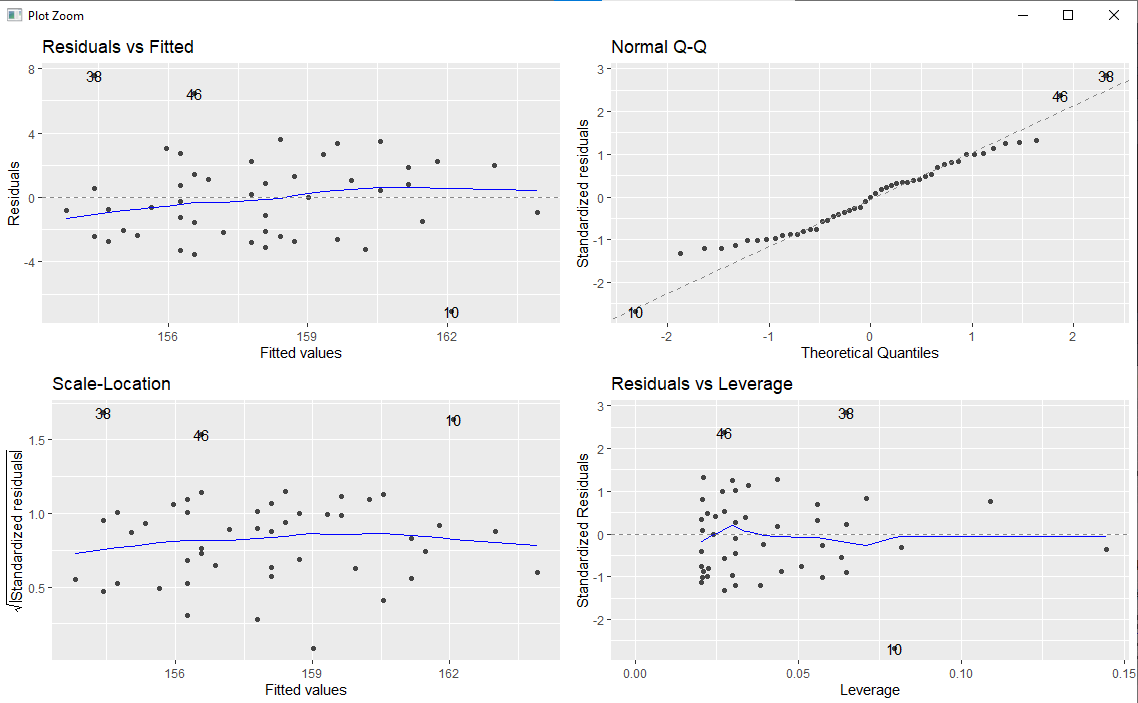
X1~X2



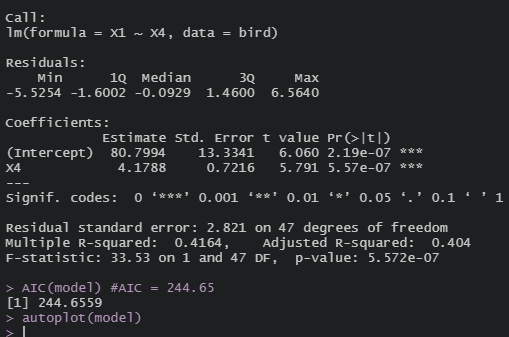


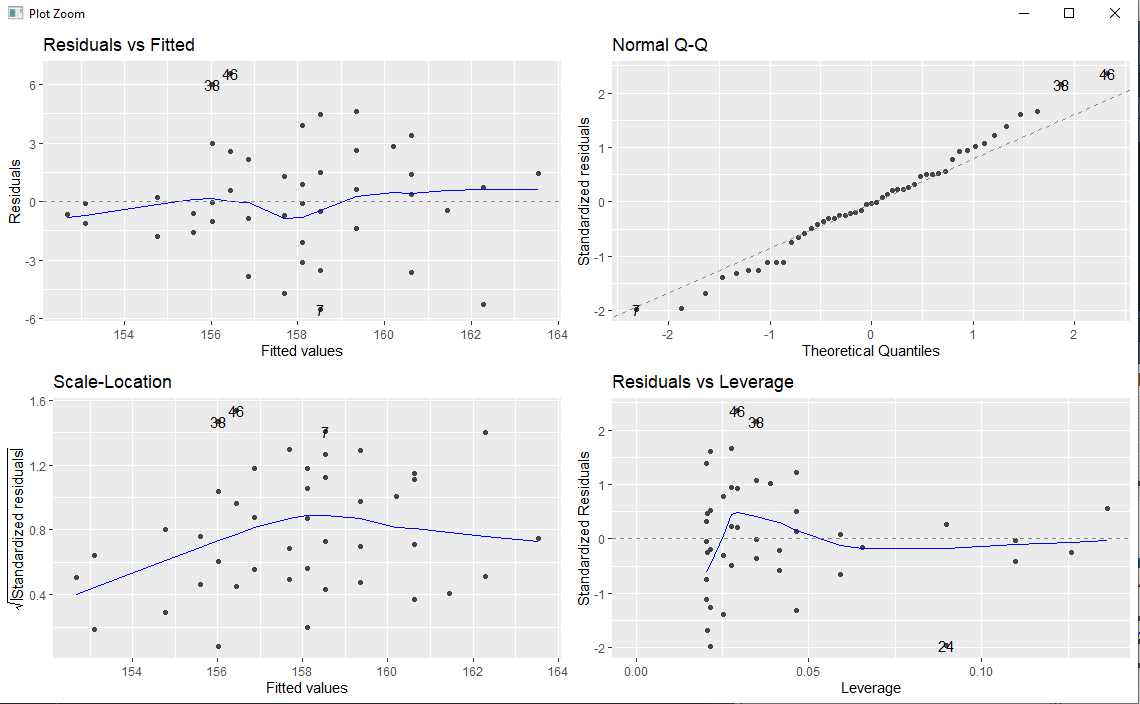
X1~X3



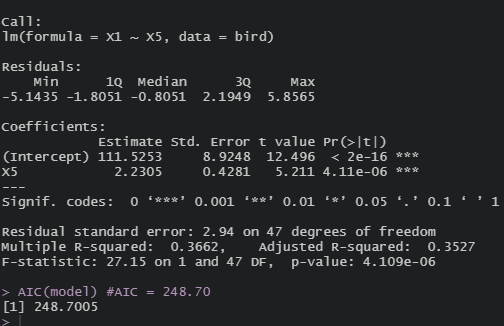


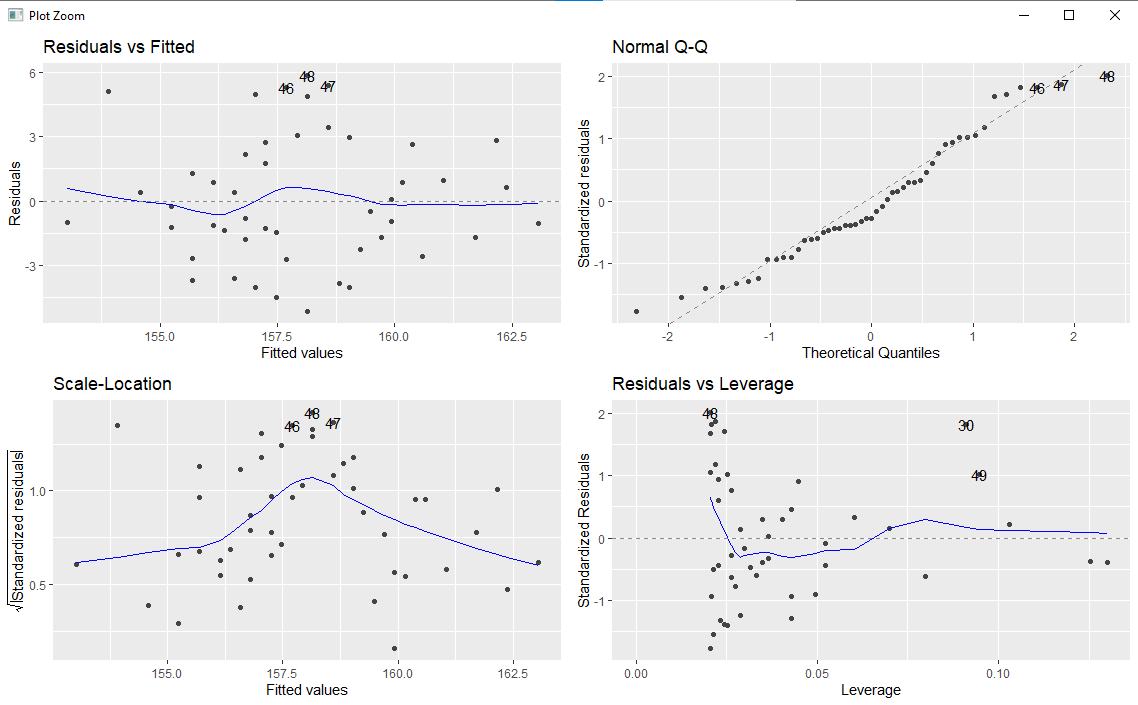
X1~X4





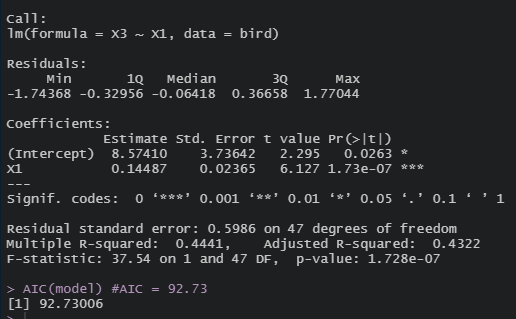
X1~X5

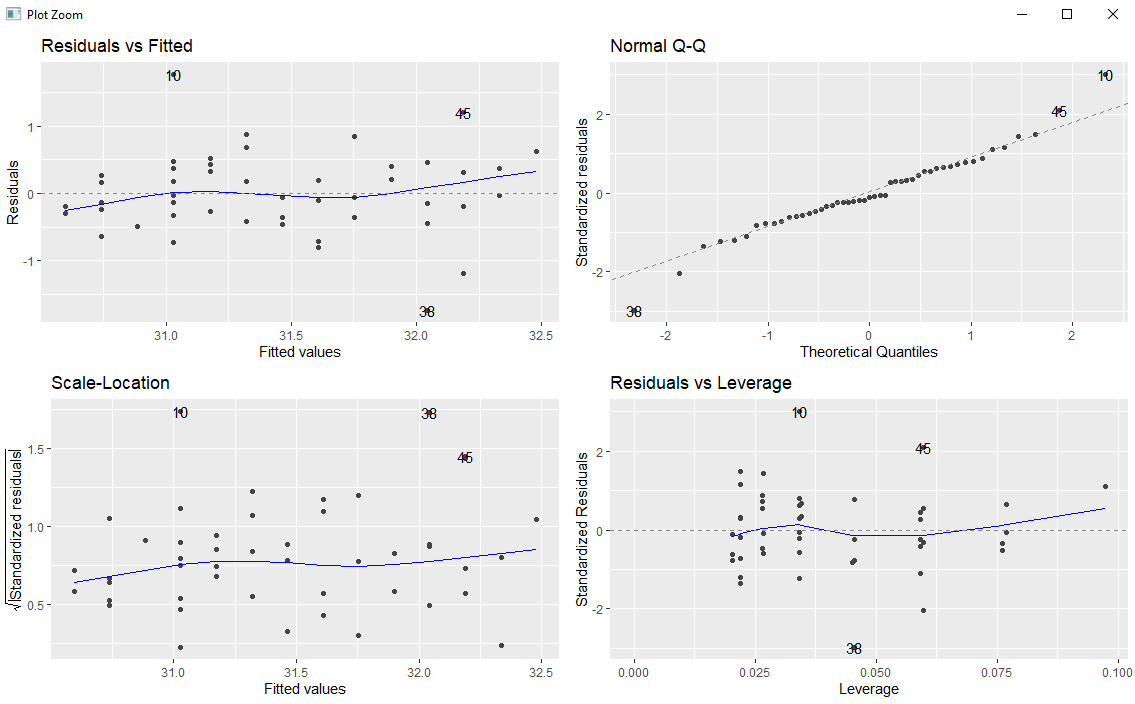




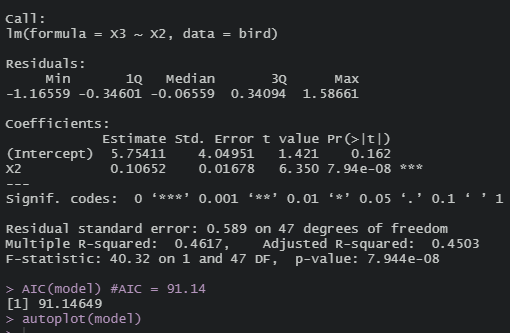
Single-predictor linear model to predict Length of beak -X3

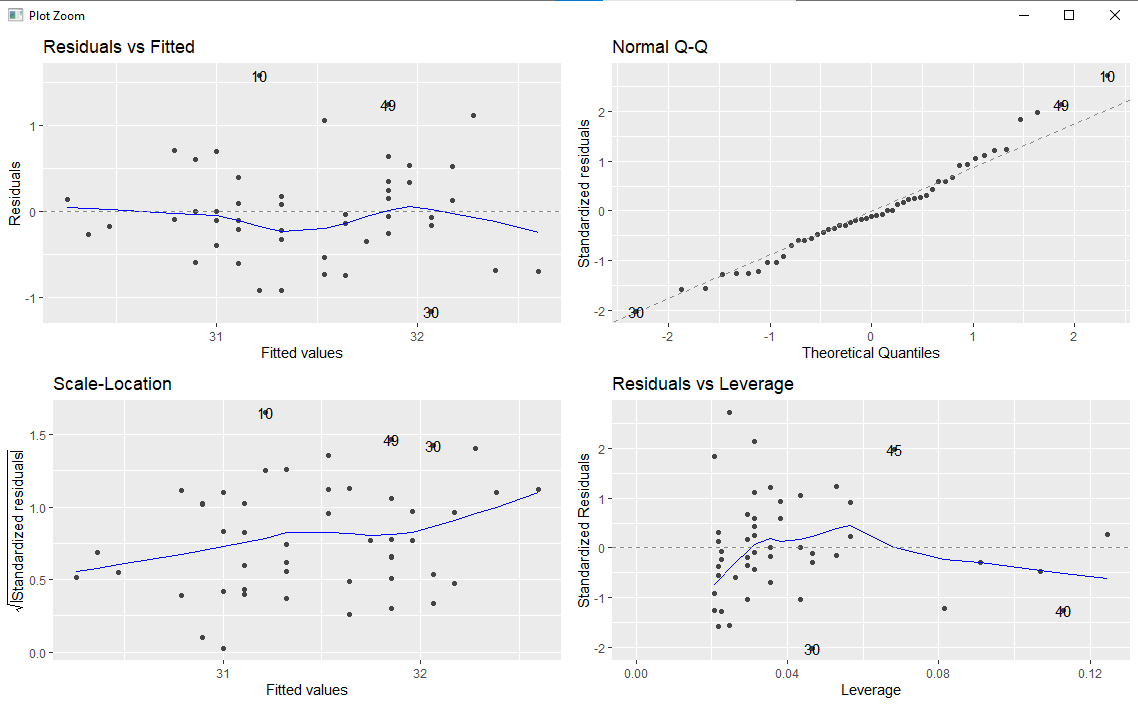
X3~X1



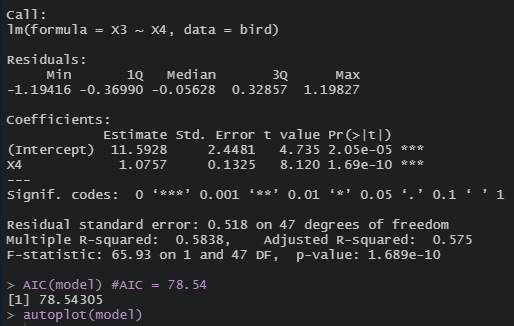


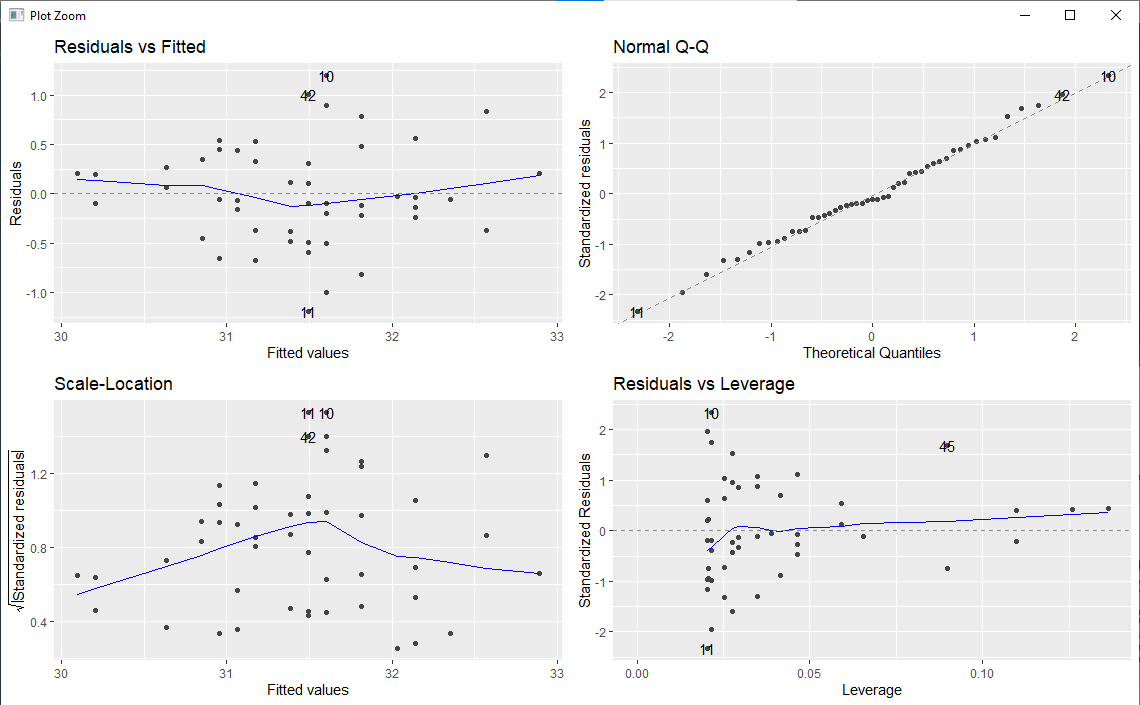
X3~X2



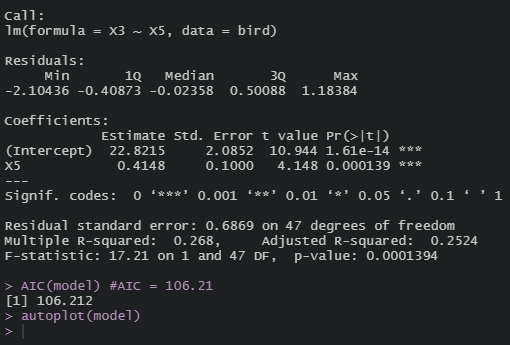


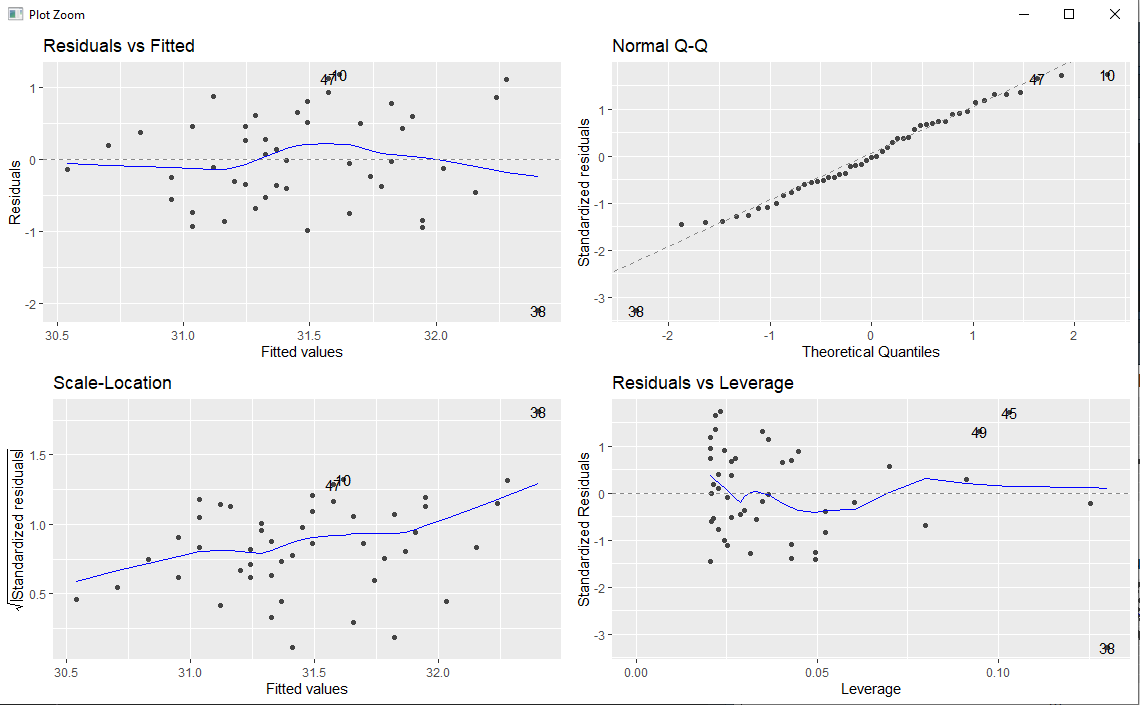
X3~X4





X3~X5





* Single Predictor model for X1~X2 : Std Error = 0.07, R\_ Squared = 0.54 , AIC = 232.97 ,Significance \*\*\*
* Single Predictor model for X1~X3 : Std Error = 0.50, R\_ Squared = 0.44 , AIC = 242.27 ,Significance \*\*\*
* Single Predictor model for X1~X4 : Std Error = 0.72, R\_ Squared = 0.41 , AIC = 244.65 ,Significance \*\*\*
* Single Predictor model for X1~X5 : Std Error = 0.42, R\_ Squared = 0.36 , AIC = 248.70 ,Significance \*\*\*
* Single Predictor model for X3~X1 : Std Error = 0.02, R\_ Squared = 0.44 , AIC = 92.73 ,Significance \*\*\*
* Single Predictor model for X3~X2 : Std Error = 0.01, R\_ Squared = 0.46 , AIC = 91.14 ,Significance \*\*\*
* Single Predictor model for X3~X4 : Std Error = 0.13, R\_ Squared = 0.58 , AIC = 78.54 ,Significance \*\*\*
* Single Predictor model for X3~X5 : Std Error = 0.1, R\_ Squared = 0.26 , AIC = 106.21 ,Significance \*\*\*

To know the model is best fit for my data, there are some most common metrics to look at while selecting the model.

* R\_squared – Higher the better
* Std. Error – close to zero
* AIC – Lower the better

My conclusion

* X2 single-predictor linear model “best” predicts total length X1 and
* X4 single-predictor linear model “best” predicts length of beak X3.

**Question 3:**

Briefly explain the acronym “AIC” used to compare models. Considering only the five measurements, what two-predictor linear models would you recommend predicting total length? Compare your best two-predictor models with the linear best models including any of the available predictors. Summarise the results from your best linear models in a small table. Discuss whether any of the sparrows in the dataset need further investigation. Justify your answers and critically assess any linear models you consider.

AIC: The Akaike information criterion (AIC) is a metric that is used to compare the fit of several regression models.

It is calculated as: AIC = 2K – 2ln(L)

K – no of model parameters.

Ln(L) – log-likelihood of the model

The AIC is designed to find the model that explains the most variation in the data, while penalizing for models that use an excessive number of parameters. Once you’ve fit several regression models, you can compare the AIC value of each model. The lower the AIC, the better the model fit.

**Code: Two predictor models**

print('/n Individual Task -- part(3) Linear Models(AIC and more predictors)')

library(olsrr)

library(ggfortify)

library(tidyverse)

bird = read\_csv("birds.csv")

bird\_sub = select(bird,-GROUP)

colnames(bird\_sub)

head(bird\_sub)

model = lm(X1~X2+X3,data=bird\_sub)

summary(model) #R-squared = 0.59

AIC(model) # AIC = 229.12

autoplot(model)

model = lm(X1~X2+X4,data=bird\_sub)

summary(model) #R-squared = 0.556

AIC(model) # AIC = 233.26

autoplot(model)

model = lm(X1~X2+X5,data=bird\_sub)

summary(model) #R-squared = 0.605

AIC(model) # AIC = 227.51

autoplot(model)

model = lm(X1~X3+X4,data=bird\_sub)

summary(model) #R-squared = 0.488

AIC(model) # AIC = 240.18

autoplot(model)

model = lm(X1~X3+X5,data=bird\_sub)

summary(model) #R-squared = 0.536

AIC(model) # AIC = 235.36

autoplot(model)

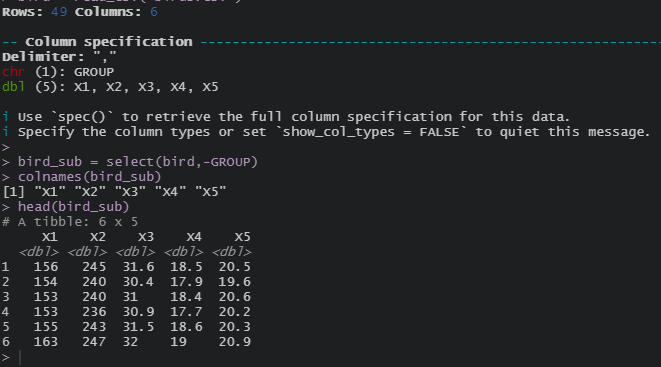
model = lm(X1~X4+X5,data=bird\_sub)

summary(model) #R-squared = 0.488

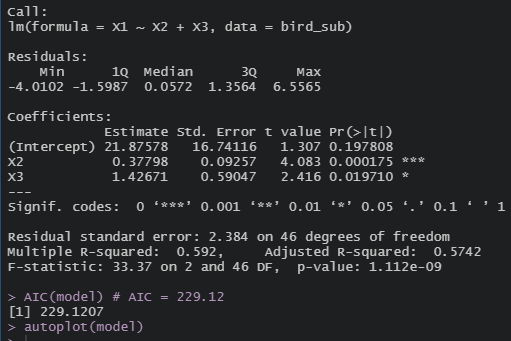
AIC(model) # AIC = 240.18

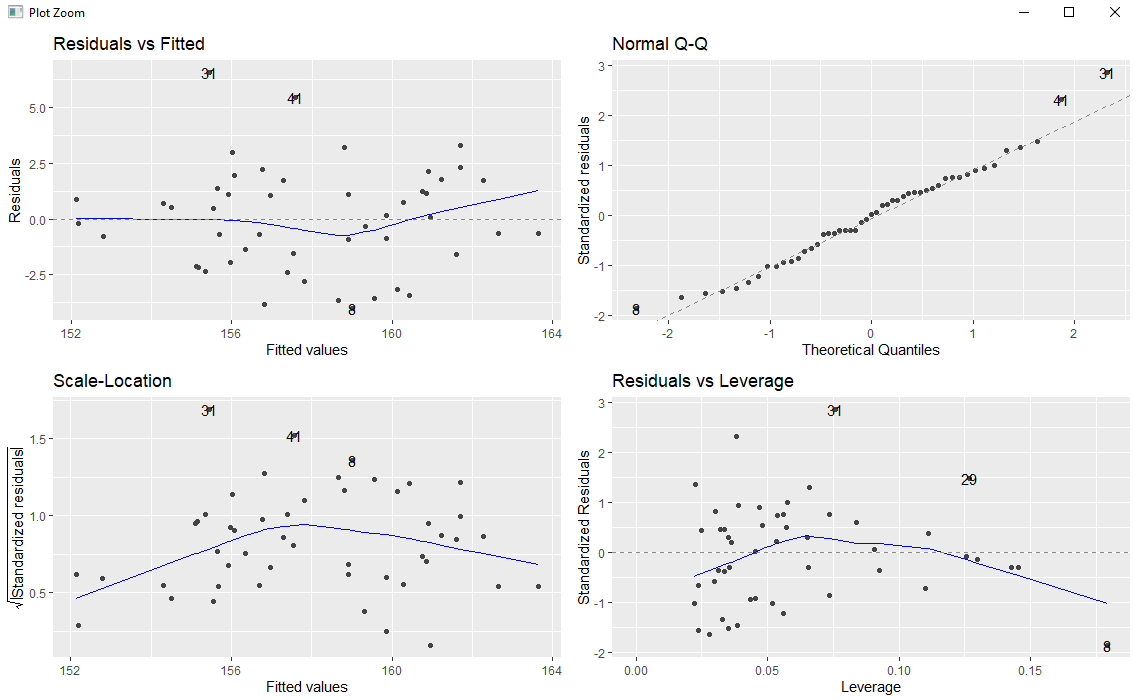
autoplot(model)

Output:

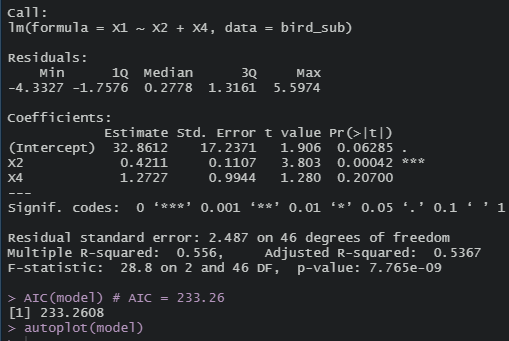


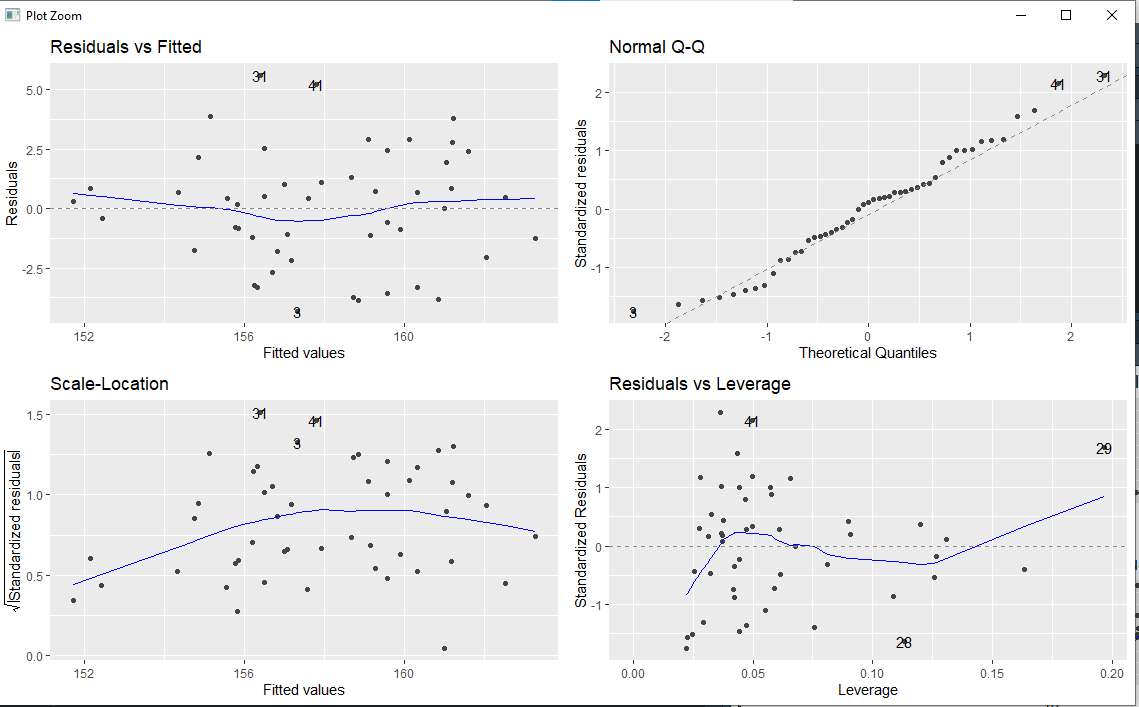
X1 ~ X2+X3



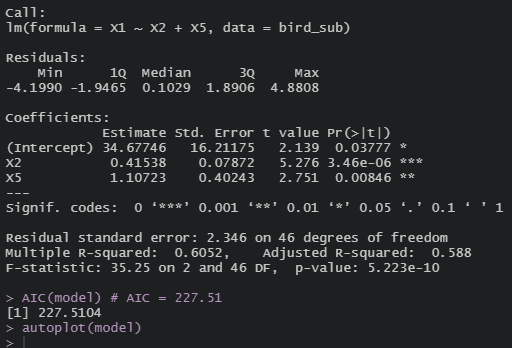


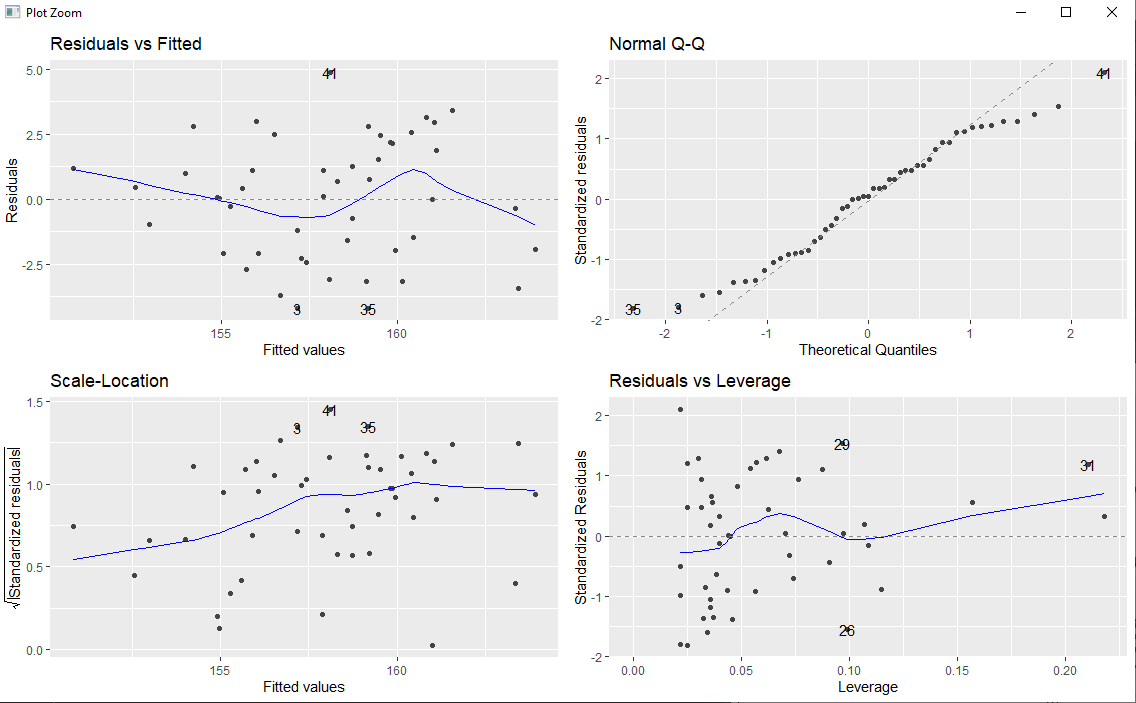
X1 ~ X2+X4



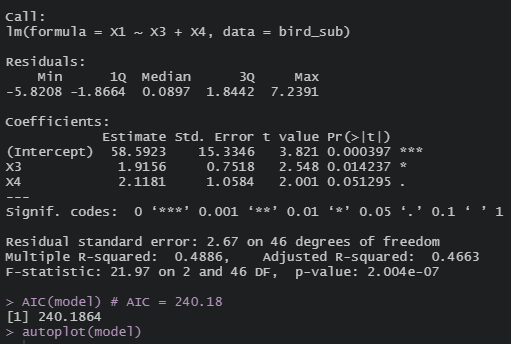


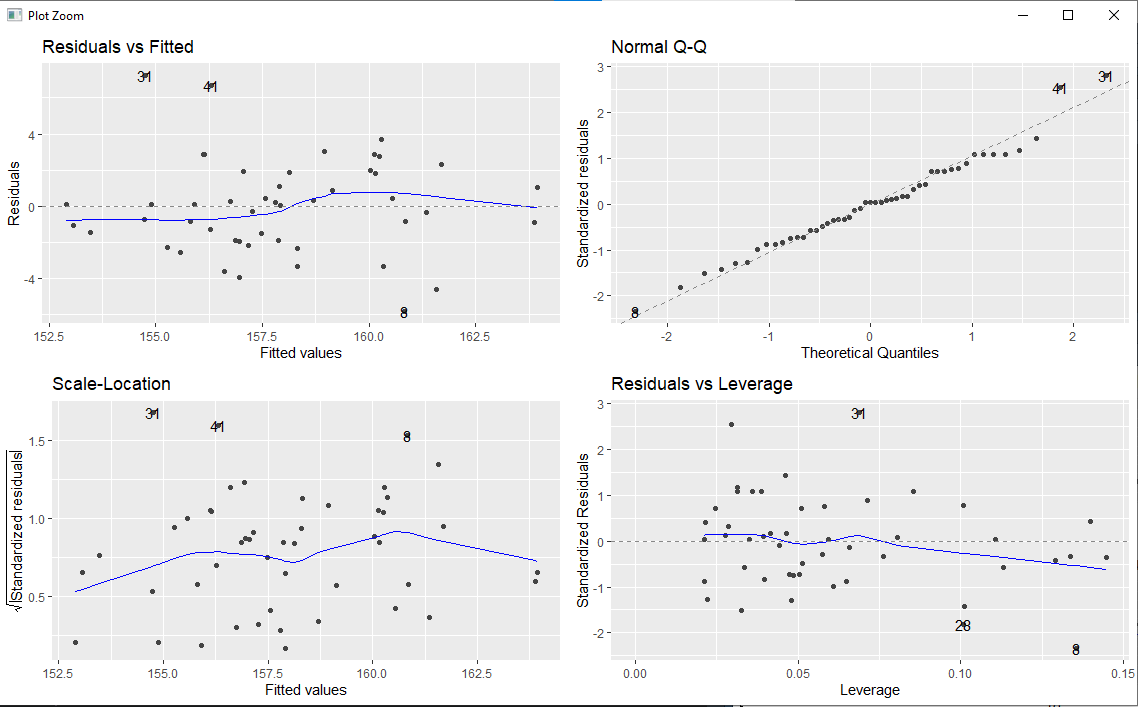
X1 ~ X2+X5



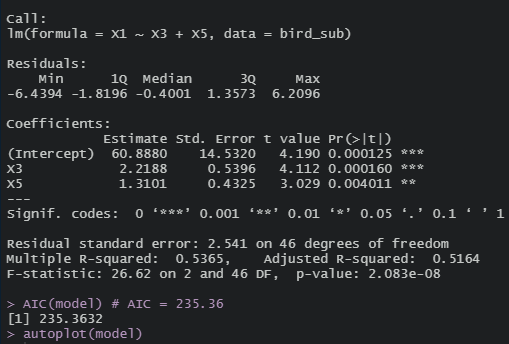


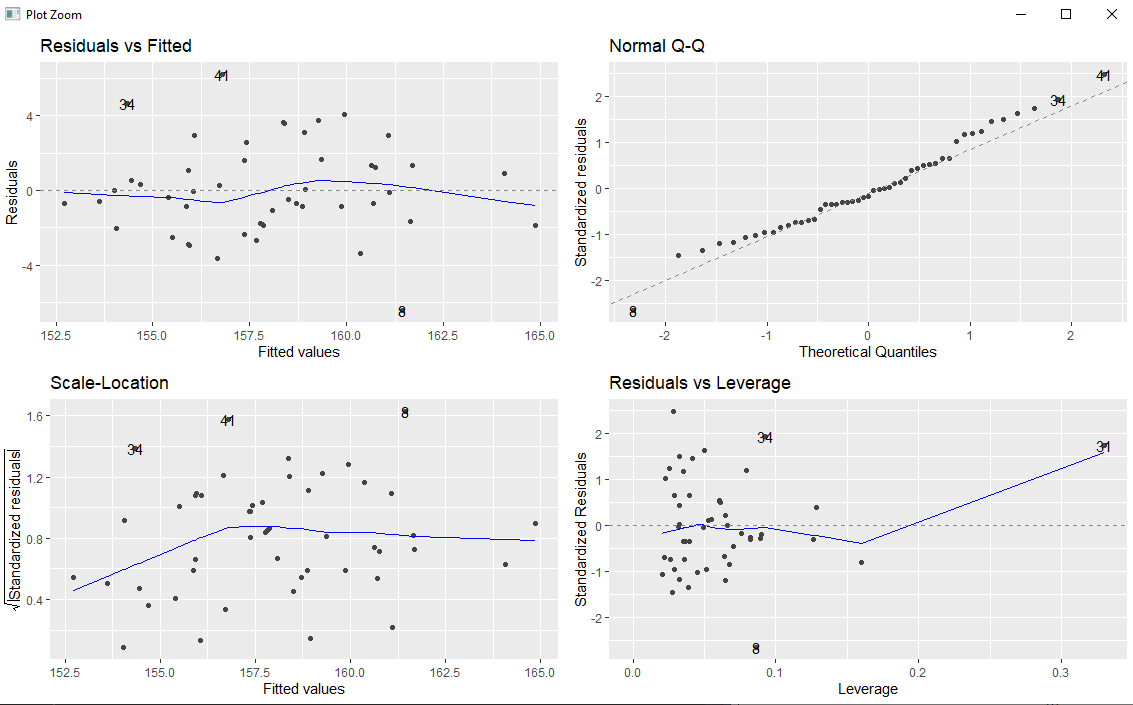
X1 ~ X3+X4



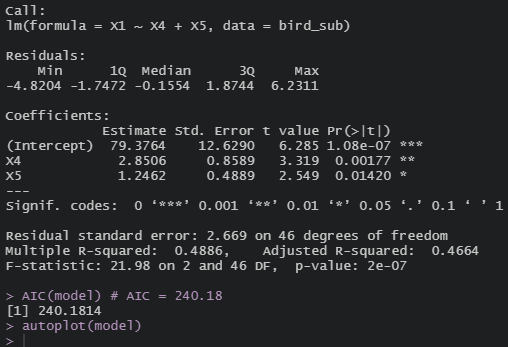


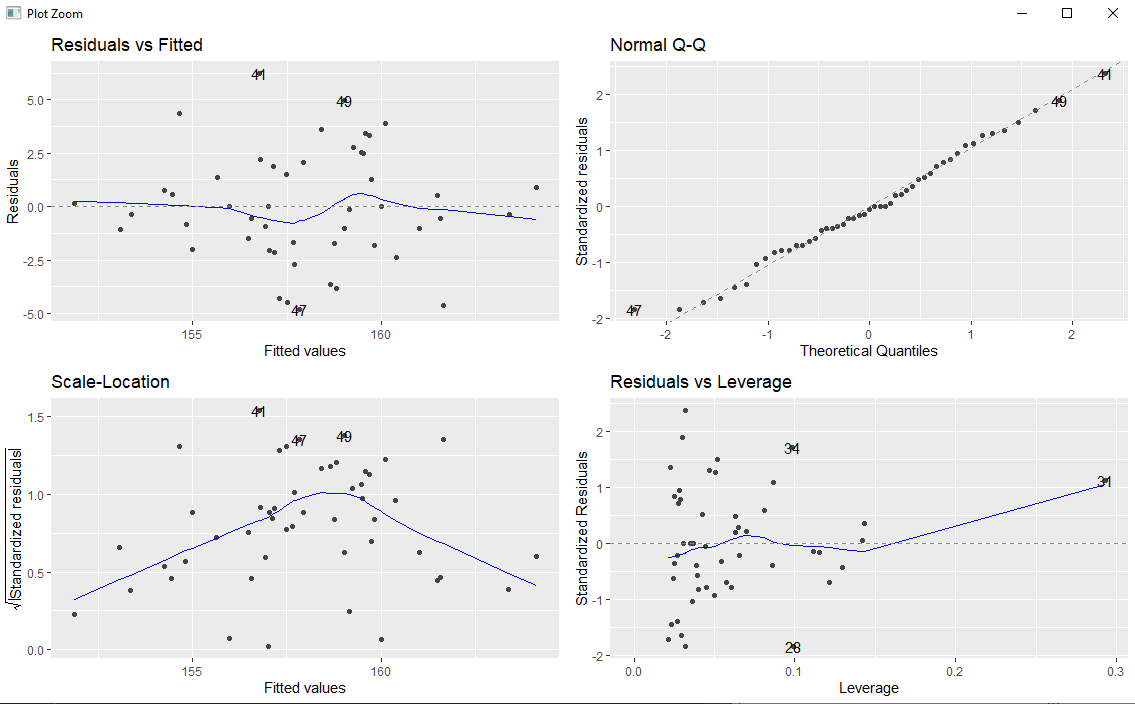
X1 ~ X3+X5





X1 ~ X4+ X5





* Two Predictor model for X1~X2+X3 : R\_ Squared = 0.59, AIC = 229.12
* Two Predictor model for X1~X2+X4 : R\_ Squared = 0.55 , AIC = 233.26
* Two Predictor model for X1~X2+X5 : R\_ Squared = 0.60 , AIC = 227.51
* Two Predictor model for X1~X3+X4 : R\_ Squared = 0.48 , AIC = 240.18
* Two Predictor model for X1~X3+X5 : R\_ Squared = 0.53 , AIC = 235.36
* Two Predictor model for X1~X4+X5 : R\_ Squared = 0.48 , AIC = 240.18

So as per the above values, X2+X5 two-predictor linear model “best” predicts total length X1 (low AIC).

**Comparing the best two-predictor models with the linear best models:**

Best two-predictor model to predict X1 is X2+X5 with AIC = 227.51 ,R\_squared = 0.06

Best linear best model to predict X1 is X2 with AIC = 232.97 , R\_ Squared = 0.54

**Summarizing Results of linear models for X1 in a small table :**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | X1~X2 | X1~X3 | X1~X4 | X1~X5 | X3~X1 | X3~X2 | X3~X4 | X3~X5 |
| R-SQUARED | 0.54 | 0.44 | 0.41 | 0.36 | 0.44 | 0.46 | 0.58 | 0.26 |
| Adj R-SQUARED | 0.53 | 0.43 | 0.40 | 0.35 | 0.43 | 0.45 | 0.57 | 0.25 |
| F-Statistic | 55.21 | 37.54 | 33.53 | 27.15 | 37.54 | 40.32 | 65.93 | 17.21 |
| Std. Error | 0.07 | 0.50 | 0.72 | 0.42 | 0.02 | 0.016 | 0.132 | 0.100 |
| t-statistic | 7.430 | 6.127 | 5.791 | 5.211 | 6.127 | 6.350 | 8.120 | 4.148 |
| P-value | 1.824e-09 | 1.728e-07 | 5.57e-07 | 4.11e-06 | 1.73e-07 | 7.94e-08 | 1.69e-10 | 0.000139 |
| AIC | 232.97 | 242.27 | 244.65 | 248.70 | 92.73 | 91.14 | 78.54 | 106.212 |

**Discussion whether any of the sparrows in the dataset need further investigation:** No sparrows in the dataset need further investigation, because with the above investigation we got all the results of single predictor linear model and two-predictor linear models of all the 49 sparrows using five measurements separately.

**Justifying the outputs from the two predictor and single predictor linear models:**

**Code to check the best predictor models:**

**print**('/n Individual Task -- part(3) Linear Models(AIC and more predictors)')

library(olsrr)

library(ggfortify)

library(tidyverse)

bird = read\_csv("birds.csv")

bird\_sub = select(bird,-GROUP)

colnames(bird\_sub)

head(bird\_sub)

# Best two-predictor model of X1 : X2+X5

model = lm(X1~.,data=bird\_sub)

summary(model)

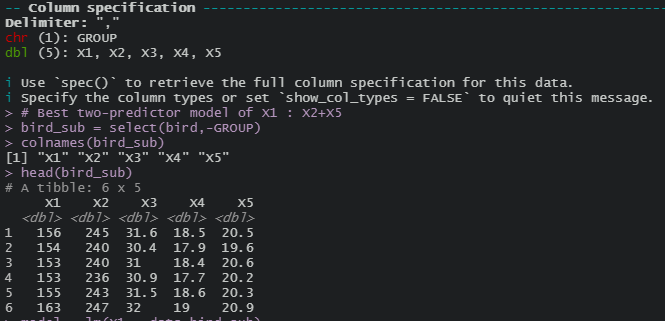
ols\_step\_best\_subset(model)

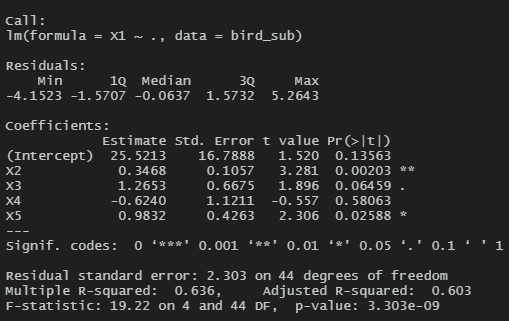
# greater coefficient are the best two predictor models to predict X1

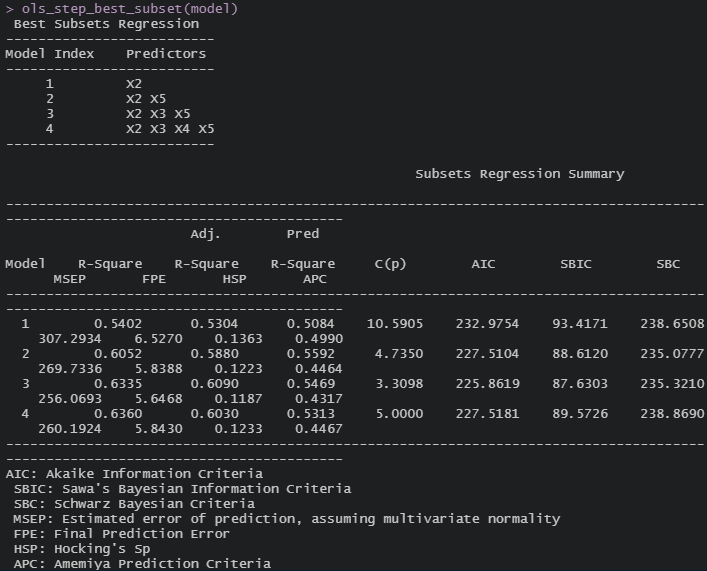
# X2 and X5 have greater coefficient

#X2 and X5 are significant which means that we can predict X1 based on X2 and X5

Outputs:







In the library OLSRR , we have an inbuild function ols\_step\_best\_subset() to find the best predictor models for predicting any particular values. With the above results we can directly find the X2 is the best single predictor linear model to predict X1 and X2+X5 is the best two predictor model to predict X1.

**Question 4:**

Consider the linear model using X2 = alar extent, X3 = length of beak and head to predict total length (call this model A) and the linear model using the other four sparrow measurements to predict total length (call this model B). Compare the residuals from these two models for each individual in the dataset. Develop a customised plot where the residuals from the two models is given on the vertical axis, total length is given on the horizontal axis, and residuals from each individual are somehow linked in the plot. Comment on what you discover.

Code:

library(ggfortify)

library(tidyverse)

bird = read\_csv('birds.csv')

bird\_sub=select(bird,-GROUP)

colnames(bird\_sub)

#compare on special plots

model\_A = lm(X1~X2+X3, data=bird\_sub)

model\_B = lm(X1~X2+X3+X4+X5, data=bird\_sub)

#compare residuals

ggplot(bird,aes(x=model\_A$residuals, y=model\_B$residuals,colour=GROUP)) +

geom\_point()

cor(model\_A$residuals, model\_B$residuals)

#customised plot

ggplot(bird,aes(x=X1))+

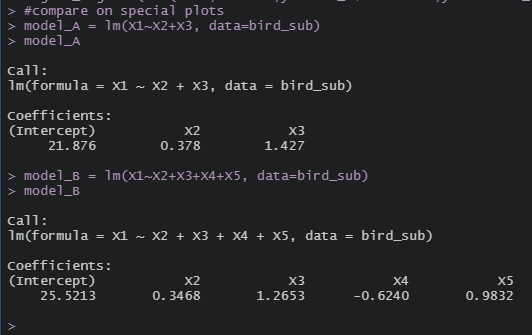
geom\_point(aes(y=model\_A$residual),shape=**2**,size=**5**)+

geom\_point(aes(y=model\_B$residuals),shape=**16**,size=**5**)+

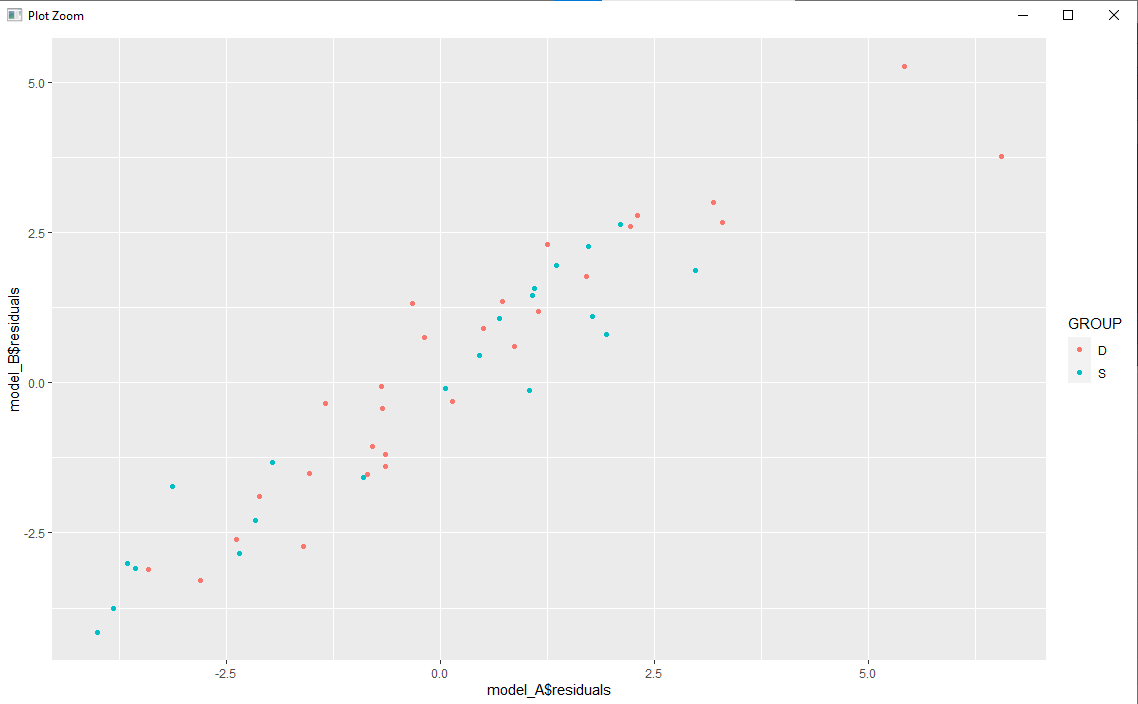
geom\_segment(aes(x=X1,xend=X1,y=model\_A$residuals,yend=model\_B$residuals))

model\_A – linear model using X2 = alar extent, X3 = length of beak and head to predict total length.

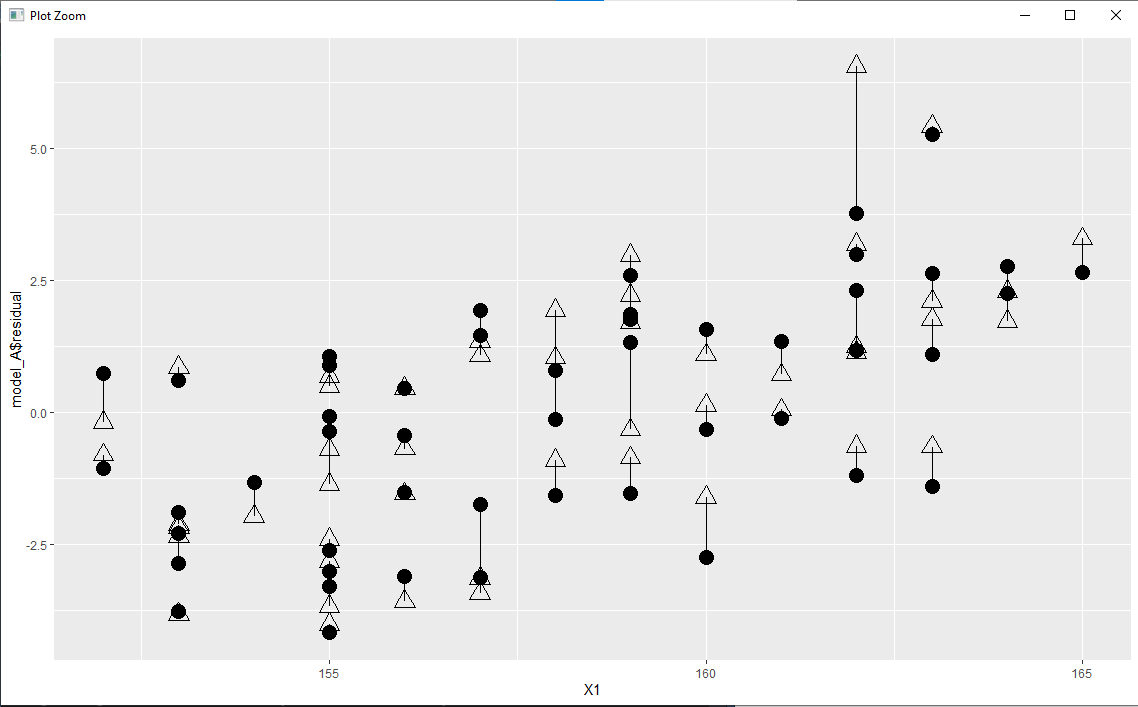
model\_B-- linear model using the other four sparrow measurements to predict total length.



Plot between model\_A and model\_B



Developed a customized plot where the residuals from the two models is given on the vertical axis, total length is given on the horizontal axis.



The residuals from each individuals are linked in the plot with respect to died /survival of the sparrow. According to my understanding ,I can see two types of representation on the plot. The first representation is vertical line which is having dot on the top and triangle on the bottom which represents died and the second representation is vertical line which is having triangle on the top and dot on the bottom which represents survival. The length of the vertical line between the dot and triangle represents the total length with respect to the two models.

**Question 5:**

Critically assess your conclusions from fitting linear models and draw comparisons with the results from PCA and Cluster Analysis. What advice would you provide to the organisers of any future study using these measurements to predict length of sparrow.

PCA-Cluster Analysis-Linear models:

* Principal components are the linear combination of original variables.PC1 produce 73% of the original information.
* Cluster analysis is different from PCA. Cluster analysis groups observations while PCA groups variables rather than observations. PCA can be used as a final method (by adding rotation to perform factor analysis) or to reduce the number of variables to conduct another analysis, such as regression or other data mining (classifying etc.) techniques.
* PCA is reducing variables of experiment based on its correlation towards observation (called PC), then the observation is grouped based on the PCs. Cluster analysis is a grouping of observations (or variables) based on their similarity.
* Both of them are exploratory methodologies, and give different approaches to the same objective: Studying the variability in the dataset.. The above analysis use both PCA and HCA to study the covariance structure in resonant inelastic x ray scattering spectroscopy. It's not the same technique, but the idea is basically the same. You may see what information we get from each of these exploratory methodologies.
* Both two methods are termed as unsupervised machine learning  because modeling problem related to this techniques need no prior groups.PCA: extracts of large related features measuring a certain variable of interest into a lower dimension without losing the original information. Sometimes is referred as direction analysis of variables to the component to maximize variations. Thus, the lower dimensions explain the variations more significantly and visibly unlike the original p-variable. Yet you can eve go beyond PCA, by fitting the modified model which used sparse matrix for the purpose of having stable features extracted.Hierarchical clustering is the natural assignment of the n-data set  following the nested trees.
* PCA seeks to represent all nn data vectors as linear combinations of a small number of eigenvectors, and does it to minimize the mean-squared reconstruction error. In contrast, K-means seeks to represent all nn data vectors via small number of cluster centroids, i.e. to represent them as linear combinations of a small number of cluster centroid vectors where linear combination weights must be all zero except for the single 11. This is also done to minimize the mean-squared reconstruction error.

References :

<https://towardsdatascience.com/7-steps-to-run-a-linear-regression-analysis-using-r-d5897a66b835>

<https://www.scribbr.com/statistics/linear-regression-in-r/>

<https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/>

<https://r-statistics.co/Linear-Regression.html#Predicting%20Linear%20Models>

<https://www.statology.org/aic-in-r/>

<https://www.researchgate.net/post/The_difference_between_principal_component_analysis_PCA_and_HCA_hierarchical_cluster_analysis>