R Notebook

This is the “Iris” dataset. Originally published at UCI Machine Learning Repository: Iris Data Set, this small dataset from 1936 is often used for testing out machine learning algorithms and visualizations (for example, Scatter Plot). Each row of the table represents an iris flower, including its species and dimensions of its botanical parts, sepal and petal, in centimeters.

Iris = readr::read\_csv(  
 "https://gist.githubusercontent.com/curran/a08a1080b88344b0c8a7/raw/0e7a9b0a5d22642a06d3d5b9bcbad9890c8ee534/iris.csv", show\_col\_types = FALSE)

Q23: Construct a one-way ANOVA model for the sepal length for each species of flower, using the species type as the predictor. Include the output showing the ANOVA table. Comment on what this output tells you about the sepal length across different spieces.

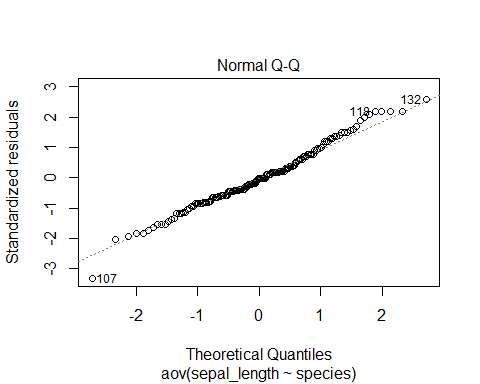
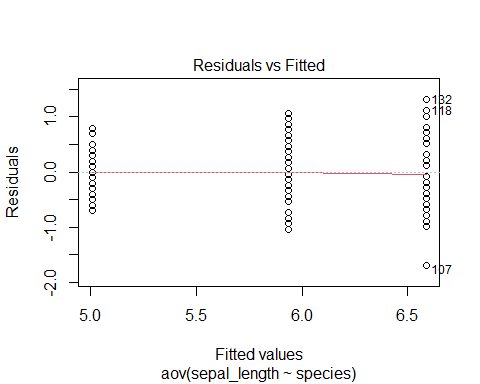
Iris23 = aov(sepal\_length~species, data = Iris)  
summary(Iris23)

## Df Sum Sq Mean Sq F value Pr(>F)   
## species 2 63.21 31.606 119.3 <2e-16 \*\*\*  
## Residuals 147 38.96 0.265   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

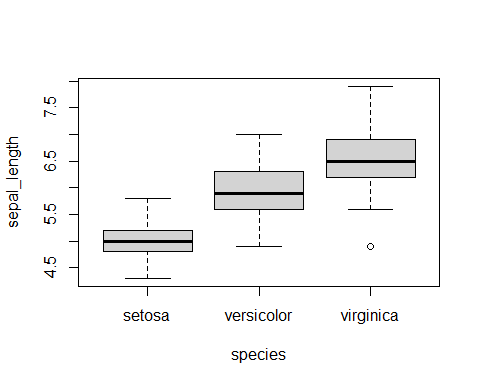
round(tapply(Iris$sepal\_length,Iris$species,sd),2)

## setosa versicolor virginica   
## 0.35 0.52 0.64

plot(Iris23, 1:2)



boxplot(sepal\_length~species, data = Iris)

 Since p-value is less than 0.05, we conclude that species is a significant predictor for sepal length. The spread of residuals is slightly increasing at each specie, however, there’s no specie that has a standard deviation of more than double of the other. From the qqplot, we can see that the points are generally following the fitted line as well. Therefore, there’s no big issue with constant variance and normalityï¼Œ which means species is an useful predictor in predicting sepal length.