Probability Project

# Data Analysis

**Group Members:**

Rahim Muhammad Syed 18K-0122 CS-2H

Syed Abdullah Muzaffar 18K-0169 CS-2H

Muhammad Ahmed Khan 18K-1103 CS-2H

**Introduction:**

The Project consists of the following dataset:

“iris”

The aforementioned dataset can be found in the base R package (and is included in the Report as “iris.csv”).

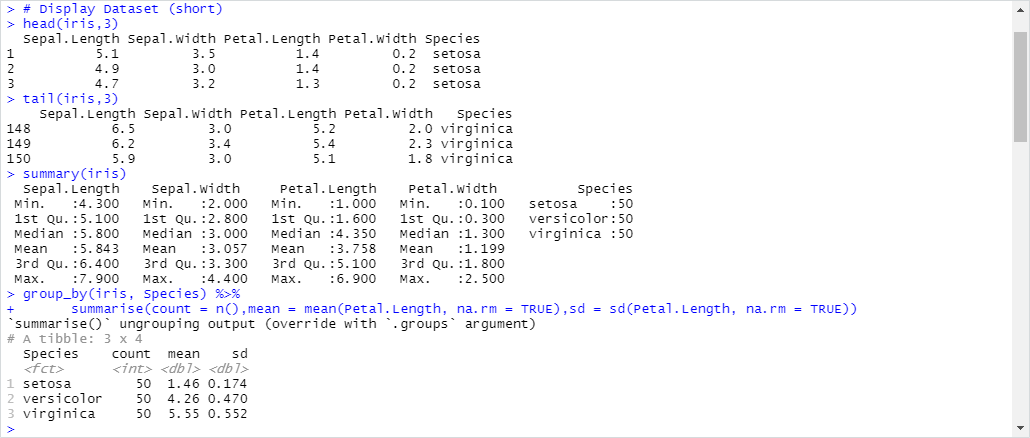
The data consists of 5 Columns of observations on flowers:

"Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"

where Species is categorical, and the rest are numerical observations.

Analysis of the dataset was performed using R Language.

**Data Summary:**

****

**Methods:**

Hypothesis Testing:

Null Hypothesis: The mean “Sepal Length” of the two species:

* 1. “Versicolor”
  2. “Setosa”

should be equal.

The “t.test(data1,data2)” function was used to calculate the p.value for the Hypothesis test.

ANOVA:

We ran ANOVA Hypothesis with y=Petal.Length against the Species as the x variable using the function “aov(formula = y ~ x)”.

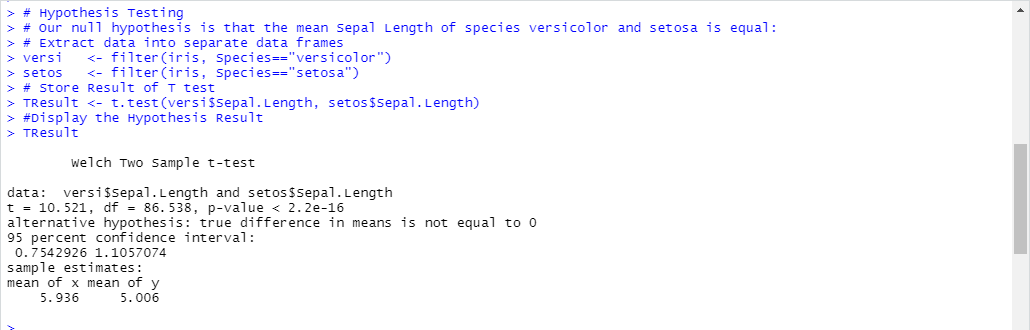
Linear Regression:

We assumed y=Sepal.Length, and x1…xn as the remaining numerical observations, to find the Linear Regression Line for “Sepal.Length” against “Sepal.Width”, “Petal.Length”, “Petal.Width”.

The “lm(formula=y~x1+x2.. , dataset)” function was used to calculate the linear regression model.

**Result:**

Hypothesis Testing:

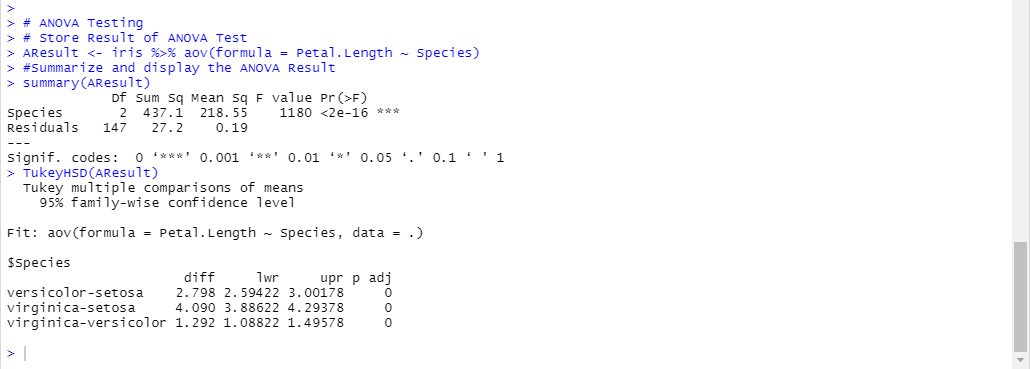


Function: t.test(versi$Sepal.Length, setos$Sepal.Length)

p-value <2.2e-16 indicates that the alternative hypothesis (the difference in means !=0 ) is likely, with the means of “Sepal.Length” differing largely between the two species/groups (“versicolor”=5.936 and “setosa”=5.006).

Our null hypothesis is thus untrue.

ANOVA:



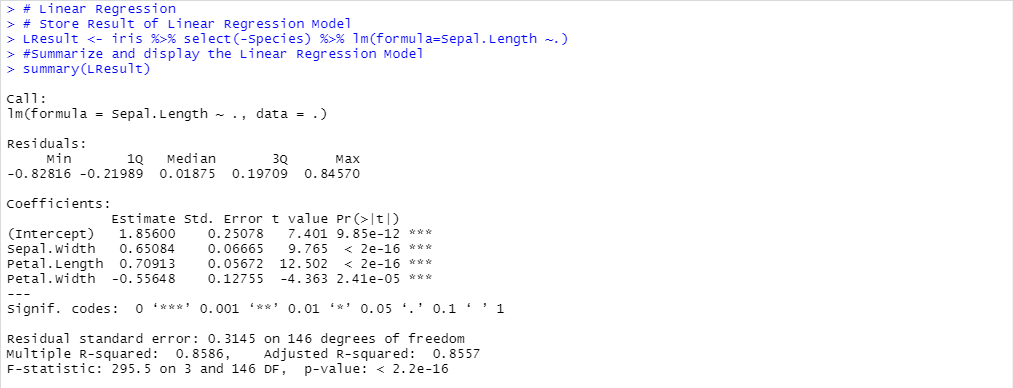
Function: aov(formula = Petal.Length ~ Species)

The results of the Tukey's honest significance test are also shown.

With an F Value of 1180, there is a large variation in mean of the Petal.Length, which is to be expected since we are comparing 3 different species of flowers.

This proves the alternative Hypothesis that at least 1 result is different.

Linear Regression:



Function: iris %>% select(-Species) %>% lm(formula=Sepal.Length ~.)

The Equation for ”Sepal.Length” is thus:

Sepal.Length =

0.65084\*Sepal.Width +

0.70913\*Petal.Length +

-0.55648\*Petal.Width +

1.85600

The R script file for the methods and results shown is included in the folder.