SMDE - 1st Assignment

Oriol Martínez Acón

1 First Question

Download the data set decathlon from the web site of Kaggle given in the link below and read it in R. `https://www.kaggle.com/drisskaouthar/decathlon#decathlon.csv`

1.1 Analyze the distribution of "X100m" according to the type of competition by using boxplot. Write your conclusion.

```
decathlon <- read.csv("decathlon.csv") #Import the decathlon data
boxplot(X100m ~ Competition, data=decathlon, main="Decathlon Data")
```

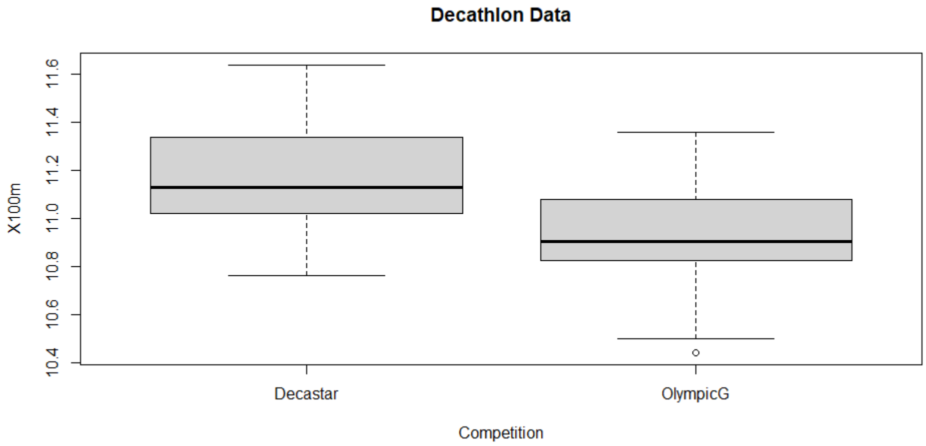
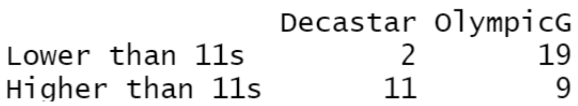


Figure 1: Boxplot of X100M distribution by Decastar and OlympicG competition.

The boxplot shows us that for 100 Meters course in the Decastar competition the mean time of the atheletes to complete it is around 11.1 seconds, and for the OlympicG the mean time of the atheletes to complete it is around 10.0 seconds. Also the standard deviation for the OlympicG athletes is less than the Decastar, so it means that if we had another athlete in OlympicG it would be have a similar time with the mean, with a well assurance. To sum up, we could say that the OlympicG athletes are faster than the Decastar ones because of the means of each Competitions.

1.2 Create a new categorical variable with two categories from the variable “X100m” by using 11 seconds as the cut-off point. Make a cross table from the new categorical variable and the “Competition”. Are these two variables independent? Write your conclusion by checking marginal probabilities and test the independency of two variables by using Chi-Square test.

```
decathlon$new_category<-cut(decathlon$X100m, c(0,11,20)) #Cutting on 11
levels(decathlon$new_category)<-c("Lower than 11s","Higher than 11s") #Categorizing

tab<-table(decathlon$new_category, decathlon$Competition) #New table with the new category
tab
margin.table(tab)
```

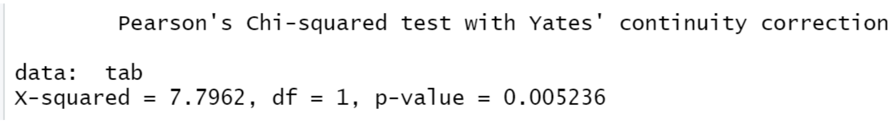


```
                Decastar OlympicG
Lower than 11s         2       19
Higher than 11s       11        9
```

Figure 2: Cross table from the new categorical variable and the “Competition”.

The Chi-Square test of independence is used to analyze the contigency table formed by two categorical variables. Chi-Square test is the method that determines if the two categorical variables have significant correlation between them. The null hypothesis of Chi-Square (H0): says that the row and the column variables of the contingency table are independent.

```
chisq.test(tab)
```



```
	Pearson's Chi-squared test with Yates' continuity correction

data:  tab
X-squared = 7.7962, df = 1, p-value = 0.005236
```

Figure 3: Output of Chi-Square Test.

As we can see the $p - value$ of the XSquared (0.005236) is < 0.05 (significance level), we can reject the null hypothesis, that means that completing the **100 Meters course** in less or more than 11 seconds is **not correlated** to the kind of **competition**.

1.3 Visualize the distribution of quantitative variables by using proper graph. Which of these variables follow a Normal distribution?

```r
plot(density(decathlon$X100m),main="Density function of X100m")
plot(density(decathlon$Long.jump),main="Density function of Long Jump")
plot(density(decathlon$Shot.put),main="Density function of Shot Put")
plot(density(decathlon$High.jump),main="Density function of High Jump")
plot(density(decathlon$X400m),main="Density function of X400m")
plot(density(decathlon$X110m.hurdle),main="Density function of X110m")
plot(density(decathlon$Discus),main="Density function of Discus")
plot(density(decathlon$Pole.vault),main="Density function of Pole")
plot(density(decathlon$Javeline),main="Density function of Javeline")
plot(density(decathlon$X1500m),main="Density function of X1500m")
```

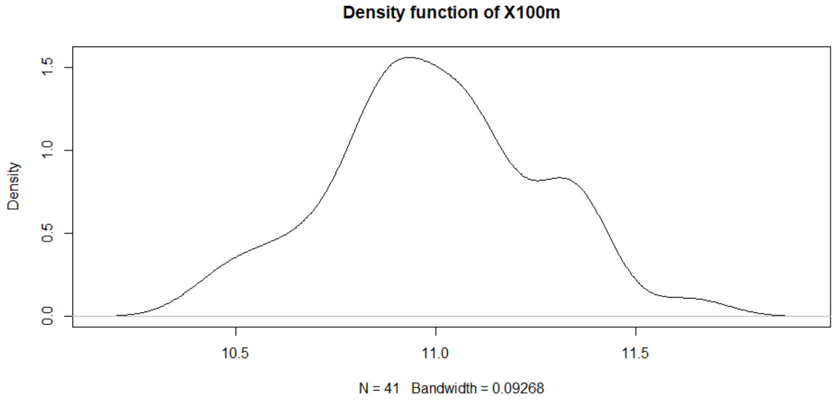


Figure 4: Density plot of X100M.

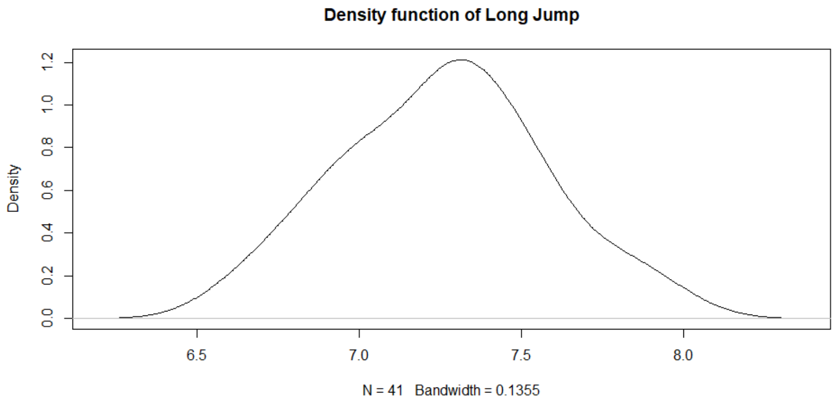


Figure 5: Density plot of Long Jump.

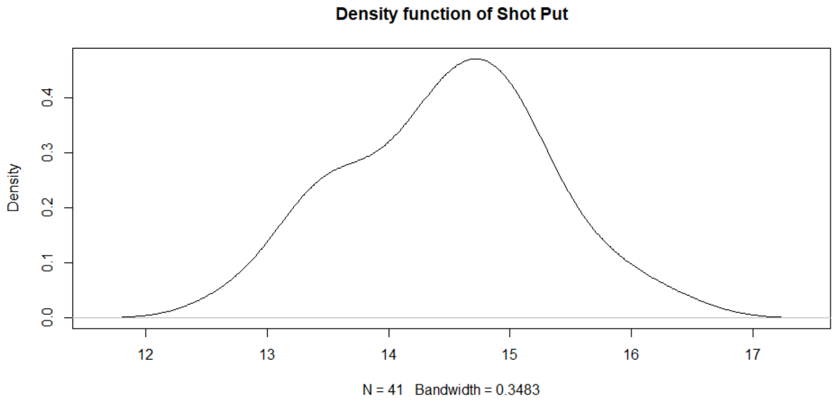


Figure 6: Density plot of Shot Put.

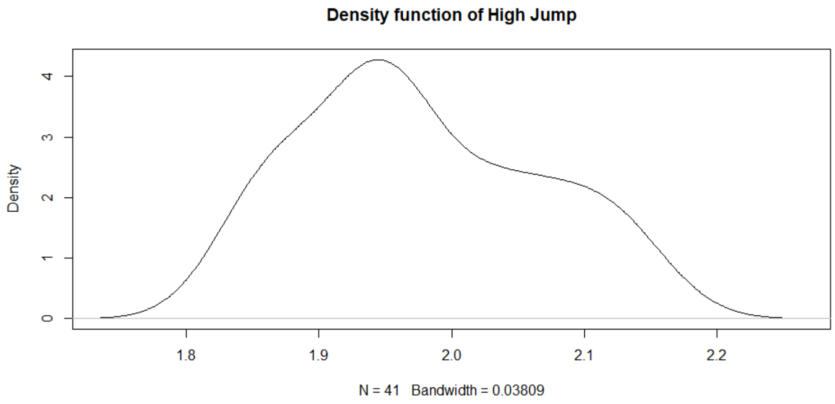


Figure 7: Density plot of High Jump.

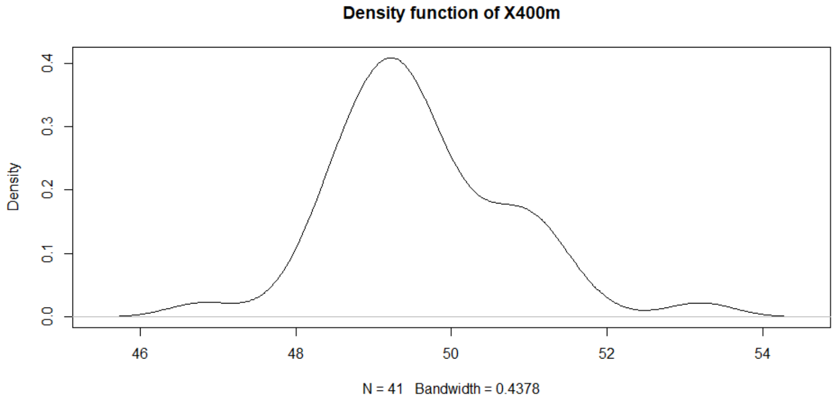


Figure 8: Density plot of X400M.

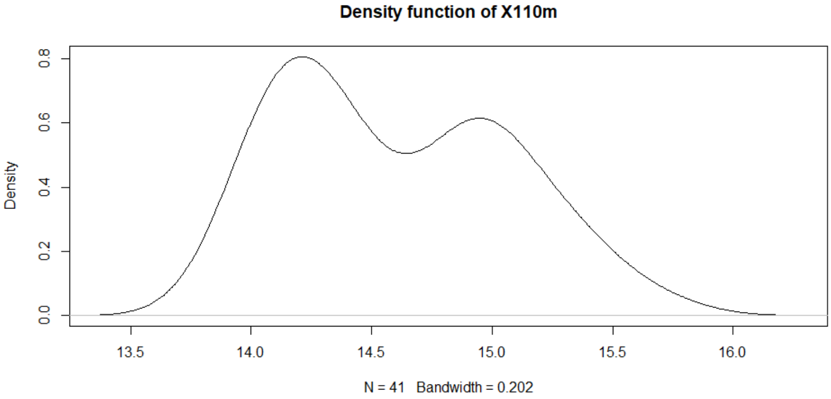


Figure 9: Density plot of X110M.

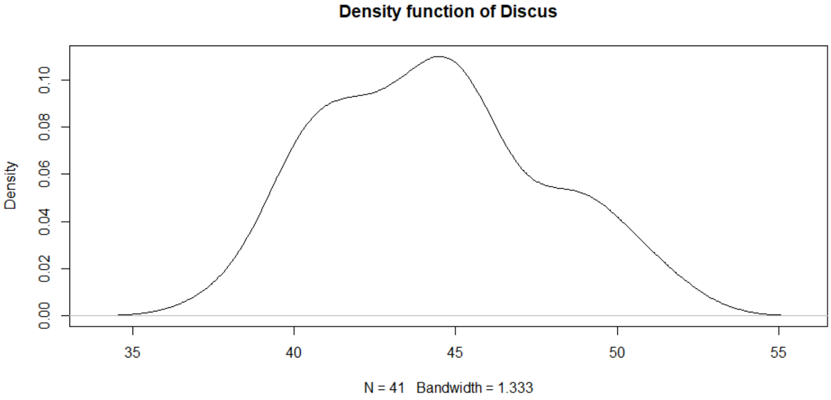


Figure 10: Density plot of Discus.

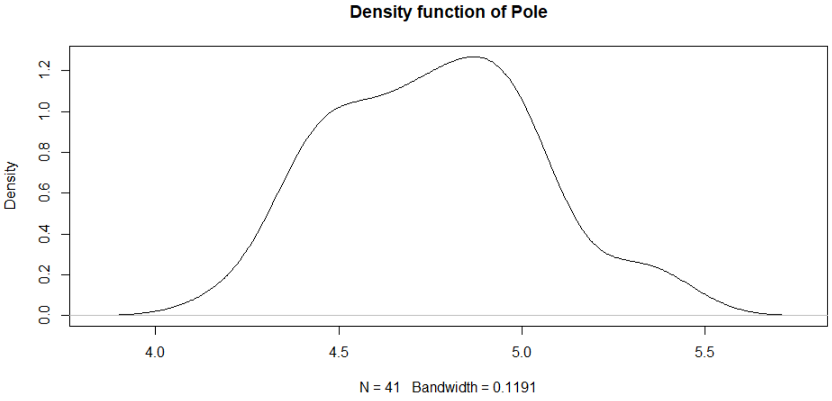


Figure 11: Density plot of Pole.

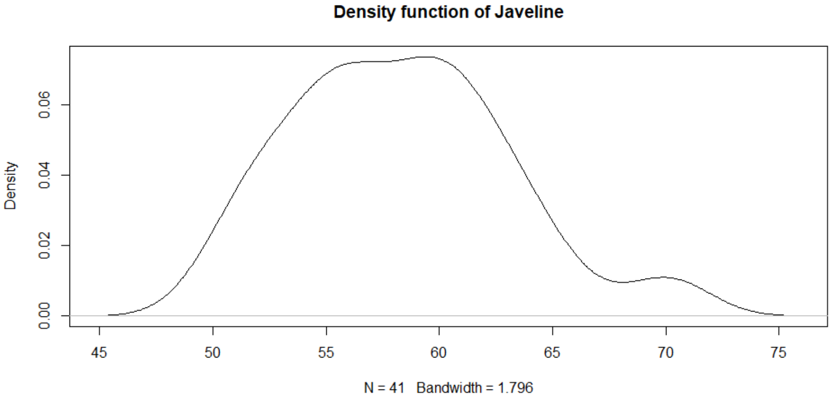


Figure 12: Density plot of Javeline.

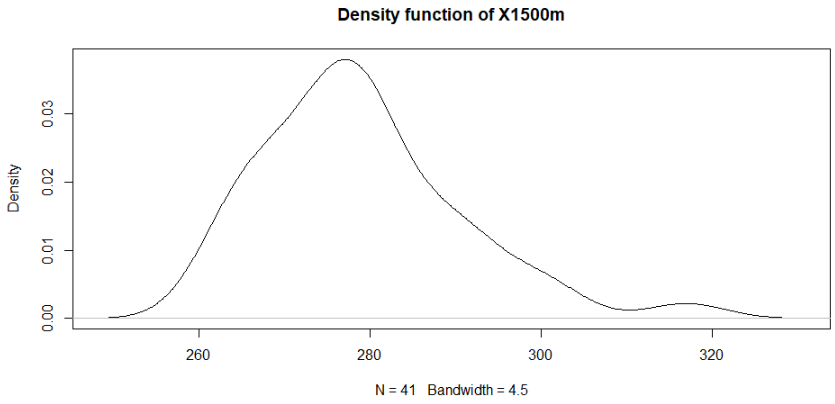


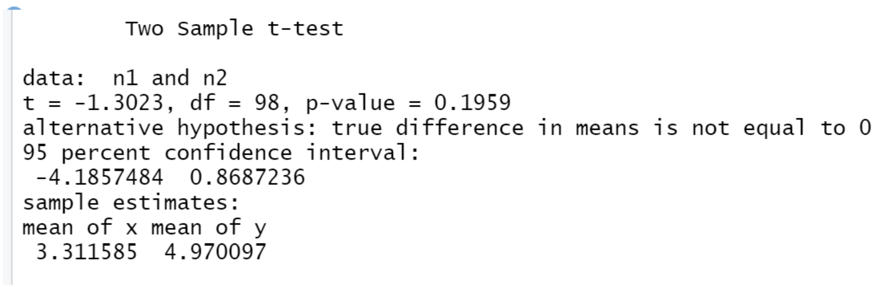
Figure 13: Density plot of X1500M.

1.4 Generate three Normally distributed random variables of length 50. Two of them should have the same mean, different standard deviations while the third one has a different mean but the same standard deviation with the first distribution. Use t test to compare mean differences between three variables.

```
n1<-rnorm(50, mean=5, sd=7)
n2<-rnorm(50, mean=5, sd=3)
n3<-rnorm(50, mean=8, sd=7)
```

T-Tests is used to determine if the means of two groups are equal to each other. There is one assumption for the test, both groups are sampled from normal distribution with same standard deviation.

```
t.test(n1, n2, var.equal = TRUE)
```

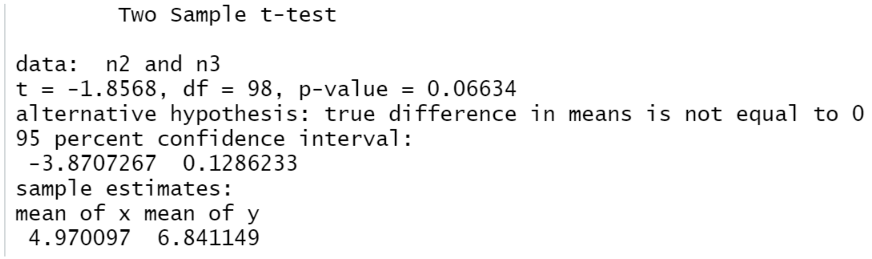


```
        Two Sample t-test

data:  n1 and n2
t = -1.3023, df = 98, p-value = 0.1959
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.1857484  0.8687236
sample estimates:
mean of x mean of y
 3.311585  4.970097
```

Figure 14: Output of T-Test of n1 and n2.

T-test shows has a $p - value > 0.05$. That means we can't refuse the null hypothesis that the two means are equal.

```
t.test(n2, n3, var.equal = TRUE)
```

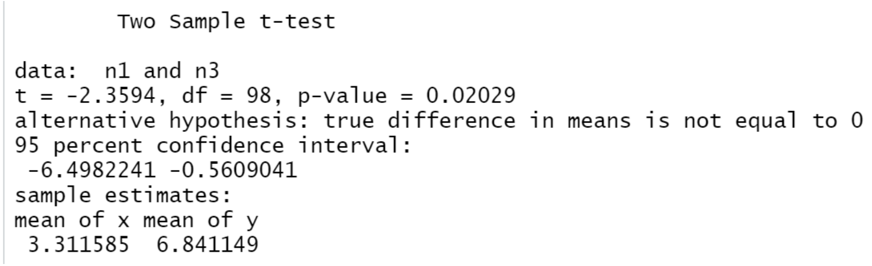


```
        Two Sample t-test

data:  n2 and n3
t = -1.8568, df = 98, p-value = 0.06634
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.8707267  0.1286233
sample estimates:
mean of x mean of y 
 4.970097  6.841149 
```

Figure 15: Output of T-Test of n2 and n3.

T-test shows has a $p - value > 0.05$. That means we can't refuse the null hypothesis that the two means are equal.

```
t.test(n1, n3, var.equal = TRUE)
```



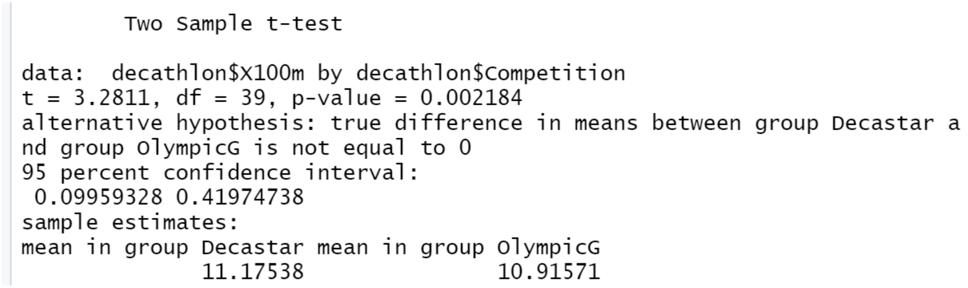
```
        Two Sample t-test

data:  n1 and n3
t = -2.3594, df = 98, p-value = 0.02029
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.4982241 -0.5609041
sample estimates:
mean of x mean of y 
 3.311585  6.841149 
```

Figure 16: Output of T-Test of n1 and n3.

T-test shows has a $p - value < 0.05$. That means we can refuse the null hypothesis that the two means are equal.

1.5 Test if there is a difference between two type of competitions according to the variables "X100m" and "X400m" by using t test.

```
t.test(decathlon$X100m ~ decathlon$Competition, var.equal=TRUE)
```

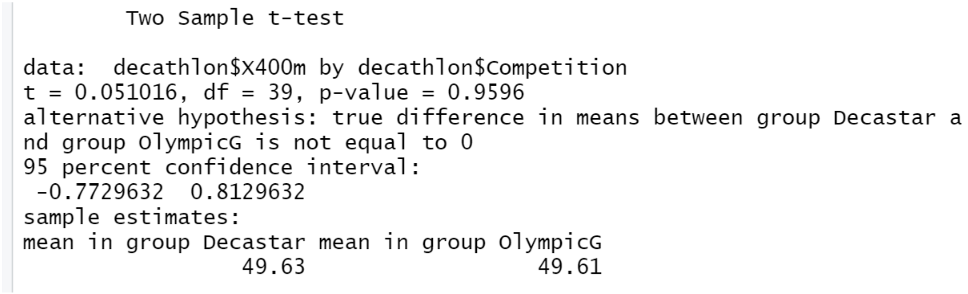


```
	Two Sample t-test

data:  decathlon$X100m by decathlon$Competition
t = 3.2811, df = 39, p-value = 0.002184
alternative hypothesis: true difference in means between group Decastar a
nd group OlympicG is not equal to 0
95 percent confidence interval:
 0.09959328 0.41974738
sample estimates:
mean in group Decastar mean in group OlympicG 
              11.17538               10.91571 
```

Figure 17: Output of T-Test of 100M and Competition.

T-test shows has a $p-value < 0.05$. That means we can refuse the null hypothesis. The kind of the competition doesn't influence an athlete in a course of 100 meters.

```r
t.test(decathlon$X400m ~ decathlon$Competition, var.equal=TRUE)
```



```
	Two Sample t-test

data:  decathlon$X400m by decathlon$Competition
t = 0.051016, df = 39, p-value = 0.9596
alternative hypothesis: true difference in means between group Decastar a
nd group OlympicG is not equal to 0
95 percent confidence interval:
 -0.7729632  0.8129632
sample estimates:
mean in group Decastar mean in group OlympicG 
                 49.63                  49.61 
```

Figure 18: Output of T-Test of 400M and Competition.

T-test shows has a $p-value > 0.05$. That means we can't refuse the null hypothesis that the kind of competition is doesn't influence an athlete in a course of 400 meters.

2 Second Question

2.1 Generate three populations that follow a normal distribution, using your own algorithm. As an example, the first is a population that follows a normal distribution with a parameter mean=10, the second with mean=40, and the third with mean=10. Select the SAME variance for the three distributions at your convenience (a value > 0).

```r
#Creation of the three population
n1<-rnorm(100, mean=10, sd=5)
n2<-rnorm(100, mean=15, sd=5)
n3<-rnorm(100, mean=10, sd=5)

plot(density(n2),main="Three Populations")
lines(density(n1),col=2)
lines(density(n3),col=3)
```

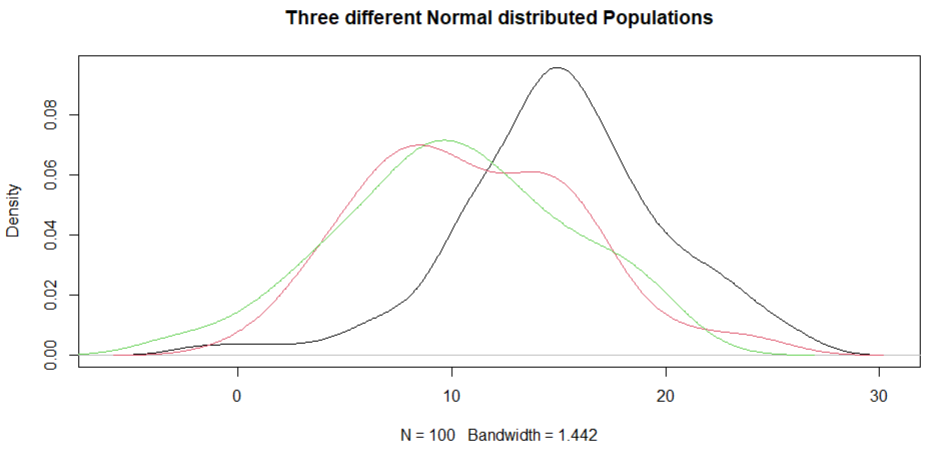


Figure 19: Plot of the three populations.

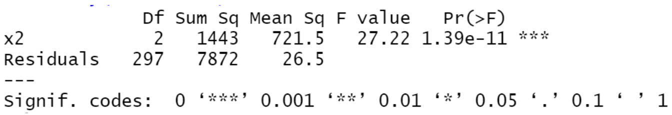
As we can see in the plot, the three population follows a normal distribution. ANOVA will help us to check if the mean of each distribution is close to the other ones and then see by the mean distance we'll know if there's a correlation between them.

2.2 We want to analyze using an ANOVA if these three populations are different (or not) depending on the parameter selected.

Remember to test the ANOVA assumptions. What do you expect on the assumptions?

```r
#We prepare the data for the ANOVA
v1n=data.frame(x1=n1, x2="n1")
v2n=data.frame(x1=n2, x2="n2")
v3n=data.frame(x1=n3, x2="n3")

```

```
library(RcmdrMisc)
#New dataframe
data=mergeRows(v1n, v2n, common.only=FALSE)
data=mergeRows(as.data.frame(data), v3n, common.only=FALSE)

AnovaModel <- aov(x1 ~ x2, data=data)
summary(AnovaModel)
```



```
             Df Sum Sq Mean Sq F value   Pr(>F)    
x2            2   1443   721.5   27.22 1.39e-11 ***
Residuals   297   7872    26.5                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 20: P-values for populations.

We can see that after using ANOVA, the Pr(¿F) is significant because is below the minimum significant code. We can reject the null hypothesis that there is no significance difference between the means of each normal distribution.

```
Boxplot(x1~x2,data=data,id=FALSE)
```

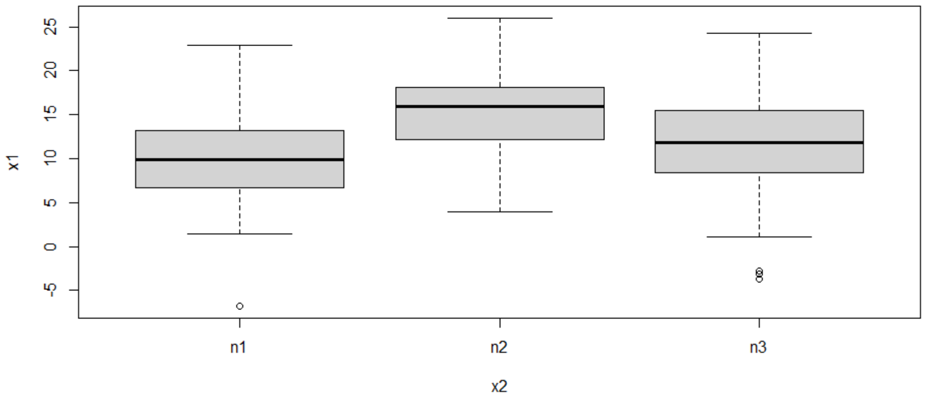


Figure 21: Boxplot of the three populations.

Once you finish the analysis and you are familiar with ANOVA test. Download the diabetes data set from Kaggle link below: `https://www.kaggle.com/uciml/pima-indians-diabetes-database`

2.3 We want to analyze if both Age and diabetes affects the risk factors. First categorize age in three groups: <=30 (young), 31-50 (middle age) and 50+ (old). Once you complete preprocessing steps, please answer the following questions:

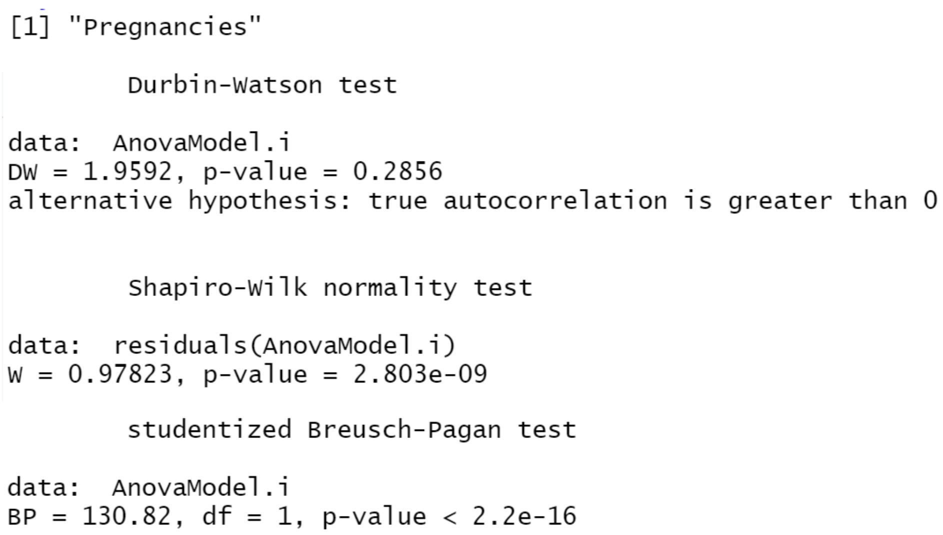
```
#Importing the diabetes database
diabetes <- read.csv("diabetes.csv")
head(diabetes)
#Categorize the Ages into groups (Young, Middle Age, Old)
diabetes$Age_Groups<-cut(diabetes$Age, c(0, 30, 50, 100))
levels(diabetes$Age_Groups)<-c("Young","Middle Age", "Old")
```

It is important to verify that the assumptions of ANOVA are fullfiled for each numerical variable, in other case the results of ANOVA might be invalid. To do it we will use Shapiro-Wilk normality, studentized Breusch-Pagan and Durbin-Watson tests.

2.3.1 How does age influence on the risk factors associated with diabetes?

```
library(lmtest) #Library for the tests
for (i in 1:7){
  print((colnames(diabetes)[i])) #Pregnancies, Glucose,
    BloodPressure, Skin Thickness, Insulin, BMI,
    DiabetesPedigreeFunction
  AnovaModel.i<-aov(diabetes[,i]~Age, data=diabetes)
  print(dwtest(AnovaModel.i))  #Independency
  print(shapiro.test(residuals(AnovaModel.i))) #Normality
  print(bptest(AnovaModel.i)) #Homogeneity of variances
  print(summary(AnovaModel.i))
}
```

The results of the ANOVA tests (Durbin-Watson, Shapiro-Wilk normality test, studentized Breusch-Pagan test) for the category of Pregnancies with Age are:



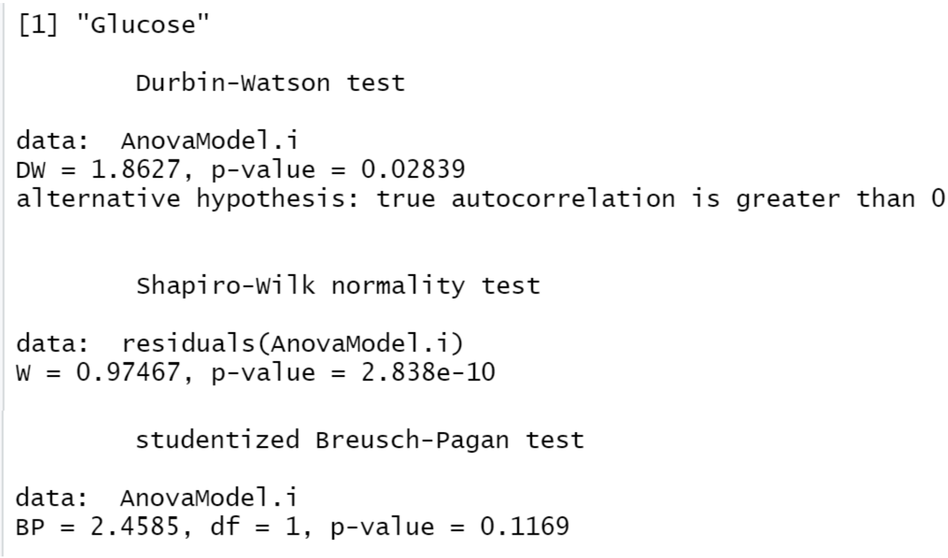
```
[1] "Pregnancies"

        Durbin-Watson test

data:  AnovaModel.i
DW = 1.9592, p-value = 0.2856
alternative hypothesis: true autocorrelation is greater than 0


        Shapiro-Wilk normality test

data:  residuals(AnovaModel.i)
W = 0.97823, p-value = 2.803e-09

        studentized Breusch-Pagan test

data:  AnovaModel.i
BP = 130.82, df = 1, p-value < 2.2e-16
```

Figure 22: Output of the different test for the category Pregnancies with Age.

- In the Durbin-Watson test the $p - value > 0.05$, then the variables are independent.
- In the Shapiro test $p - value < 0.05$. So the sample doesn't follow a normal distribution.
- In the Breusch-Pagan test the $p - value < 0.05$. That means that in the variances there's no homogeneity.

Two of the ANOVA tests (Breusch-Pagan and Shapiro) don't succed so the ANOVA tests might be invalid as we said in the beggining of the exercise.

The results of the ANOVA tests (Durbin-Watson, Shapiro-Wilk normality test, studentized Breusch-Pagan test) for the category of Glucose with Age are:



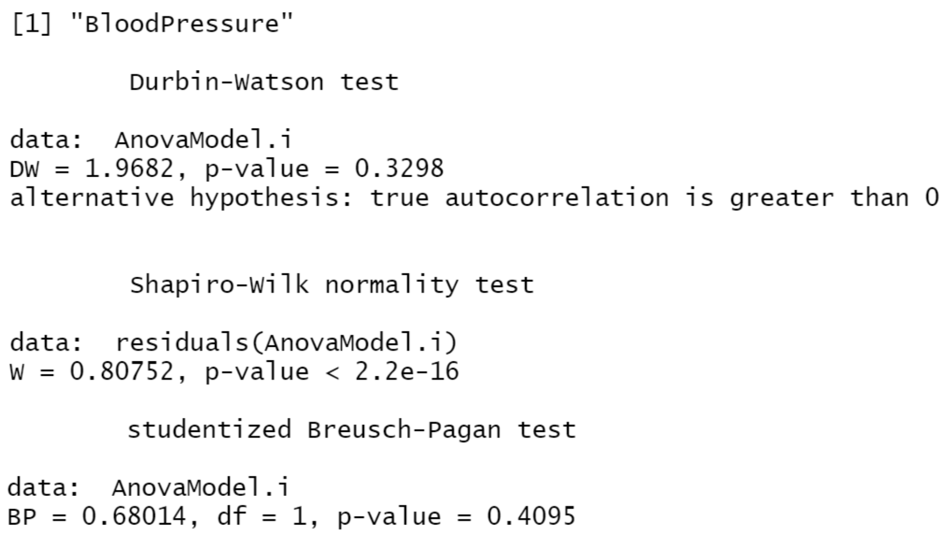
```
[1] "Glucose"

        Durbin-Watson test

data:  AnovaModel.i
DW = 1.8627, p-value = 0.02839
alternative hypothesis: true autocorrelation is greater than 0


        Shapiro-Wilk normality test

data:  residuals(AnovaModel.i)
W = 0.97467, p-value = 2.838e-10

        studentized Breusch-Pagan test

data:  AnovaModel.i
BP = 2.4585, df = 1, p-value = 0.1169
```

Figure 23: Output of the different test for the category Glucose with Age.

- In the Durbin-Watson test the $p-value < 0.05$, then the variables aren't independent.
- In the Shapiro test $p-value < 0.05$. So the sample doesn't follow a normal distribution.
- In the Breusch-Pagan test the $p-value > 0.05$. That means that in the variances there's homogeneity.

Two of the ANOVA tests (Shapiro and Durbin-Watson) don't succed so the ANOVA tests might be invalid as we said in the beggining of the exercise.

The results of the ANOVA tests (Durbin-Watson, Shapiro-Wilk normality test, studentized Breusch-Pagan test) for the category of BloodPressure with Age are:



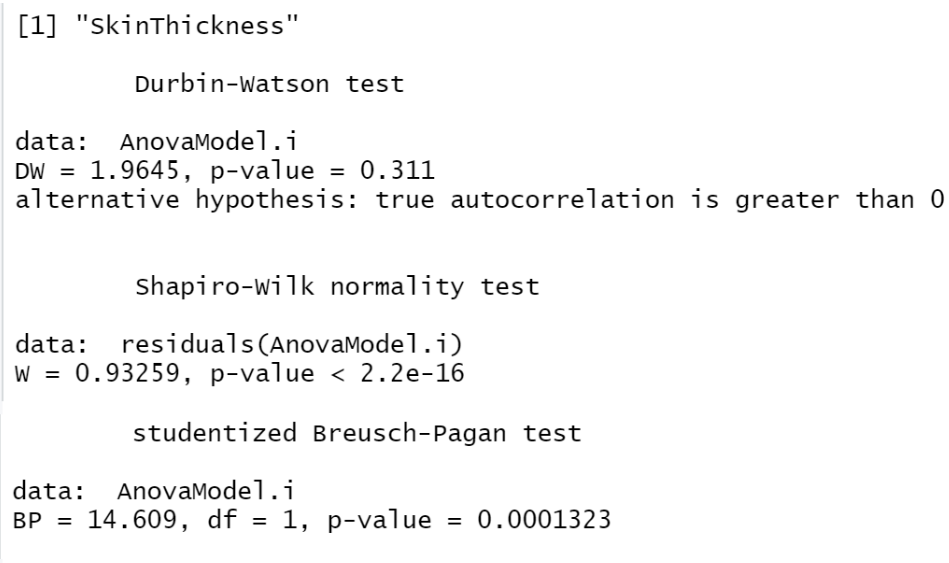
```
[1] "BloodPressure"

        Durbin-Watson test

data:  AnovaModel.i
DW = 1.9682, p-value = 0.3298
alternative hypothesis: true autocorrelation is greater than 0


        Shapiro-Wilk normality test

data:  residuals(AnovaModel.i)
W = 0.80752, p-value < 2.2e-16

        studentized Breusch-Pagan test

data:  AnovaModel.i
BP = 0.68014, df = 1, p-value = 0.4095
```

Figure 24: Output of the different test for the category BloodPressure with Age.

- In the Durbin-Watson test the $p - value > 0.05$, then the variables are independent.
- In the Shapiro test $p - value < 0.05$. So the sample doesn't follow a normal distribution.
- In the Breusch-Pagan test the $p - value > 0.05$. That means that in the variances there's homogeneity.

One of the ANOVA tests (Shapiro) don't succeed so the ANOVA tests might be invalid as we said in the beginning of the exercise.

The results of the ANOVA tests (Durbin-Watson, Shapiro-Wilk normality test, studentized Breusch-Pagan test) for the category of SkinThickness with Age are:



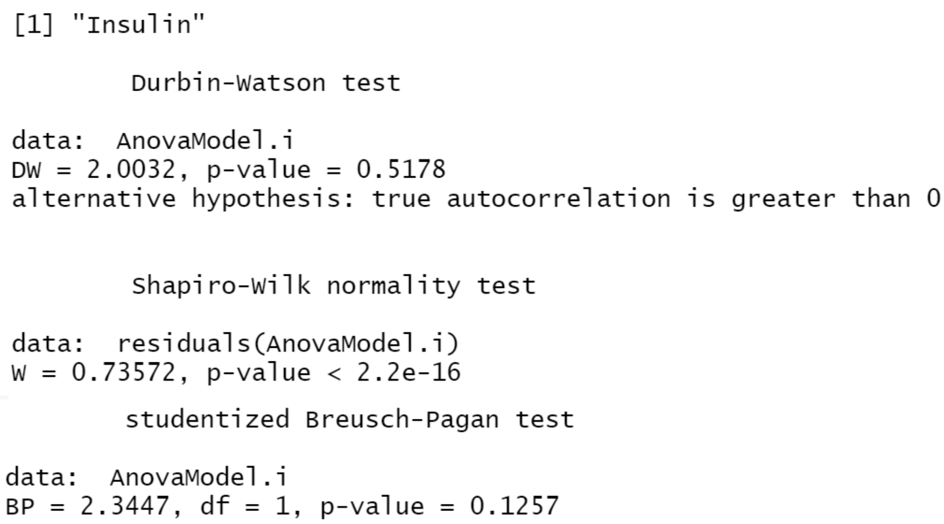
```
[1] "SkinThickness"

        Durbin-Watson test

data:  AnovaModel.i
DW = 1.9645, p-value = 0.311
alternative hypothesis: true autocorrelation is greater than 0


        Shapiro-Wilk normality test

data:  residuals(AnovaModel.i)
W = 0.93259, p-value < 2.2e-16

        studentized Breusch-Pagan test

data:  AnovaModel.i
BP = 14.609, df = 1, p-value = 0.0001323
```

Figure 25: Output of the different test for the category SkinThickness with Age.

- In the Durbin-Watson test the $p - value > 0.05$, then the variables aren independent.
- In the Shapiro test $p - value < 0.05$. So the sample doesn't follow a normal distribution.
- In the Breusch-Pagan test the $p - value < 0.05$. That means that in the variances there's no homogeneity.

Two of the ANOVA tests (Shapiro and Breusch-Pagan) don't succed so the ANOVA tests might be invalid as we said in the beggining of the exercise.

The results of the ANOVA tests (Durbin-Watson, Shapiro-Wilk normality test, studentized Breusch-Pagan test) for the category of Insulin with Age are:



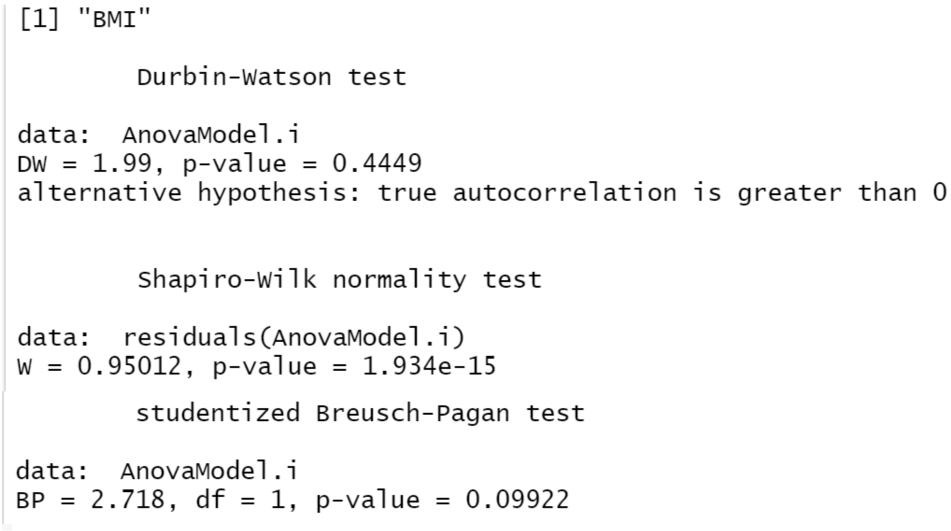
```
[1] "Insulin"

        Durbin-Watson test

data:  AnovaModel.i
DW = 2.0032, p-value = 0.5178
alternative hypothesis: true autocorrelation is greater than 0


        Shapiro-Wilk normality test

data:  residuals(AnovaModel.i)
W = 0.73572, p-value < 2.2e-16

        studentized Breusch-Pagan test

data:  AnovaModel.i
BP = 2.3447, df = 1, p-value = 0.1257
```

Figure 26: Output of the different test for the category Insulin with Age.

- In the Durbin-Watson test the $p - value > 0.05$, then the variables are independent.
- In the Shapiro test $p - value < 0.05$. So the sample doesn't follow a normal distribution.
- In the Breusch-Pagan test the $p - value > 0.05$. That means that in the variances there's homogeneity.

One of the ANOVA tests (Shapiro) don't succed so the ANOVA tests might be invalid as we said in the beggining of the exercise.

The results of the ANOVA tests (Durbin-Watson, Shapiro-Wilk normality test, studentized Breusch-Pagan test) for the category of BMI with Age are:



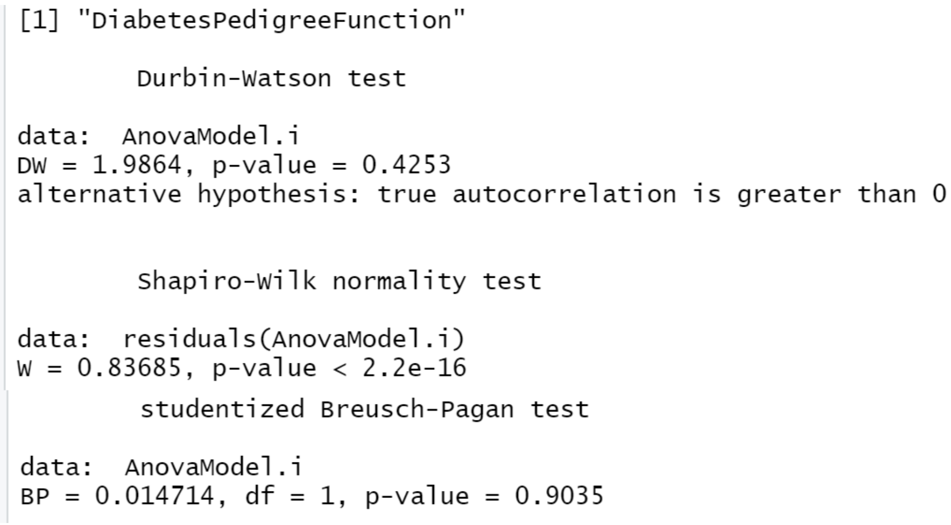
```
[1] "BMI"

        Durbin-Watson test

data:  AnovaModel.i
DW = 1.99, p-value = 0.4449
alternative hypothesis: true autocorrelation is greater than 0


        Shapiro-Wilk normality test

data:  residuals(AnovaModel.i)
W = 0.95012, p-value = 1.934e-15
        studentized Breusch-Pagan test

data:  AnovaModel.i
BP = 2.718, df = 1, p-value = 0.09922
```

Figure 27: Output of the different test for the category BMI with Age.

- In the Durbin-Watson test the $p - value > 0.05$, then the variables are independent.
- In the Shapiro test $p - value < 0.05$. So the sample doesn't follow a normal distribution.
- In the Breusch-Pagan test the $p - value < 0.05$. That means that in the variances there's no homogeneity.

Two of the ANOVA tests (Shapiro and Breusch-Pagan) don't succed so the ANOVA tests might be invalid as we said in the beggining of the exercise.

The results of the ANOVA tests (Durbin-Watson, Shapiro-Wilk normality test, studentized Breusch-Pagan test) for the category of DiabetesPedigreeFunction with Age are:

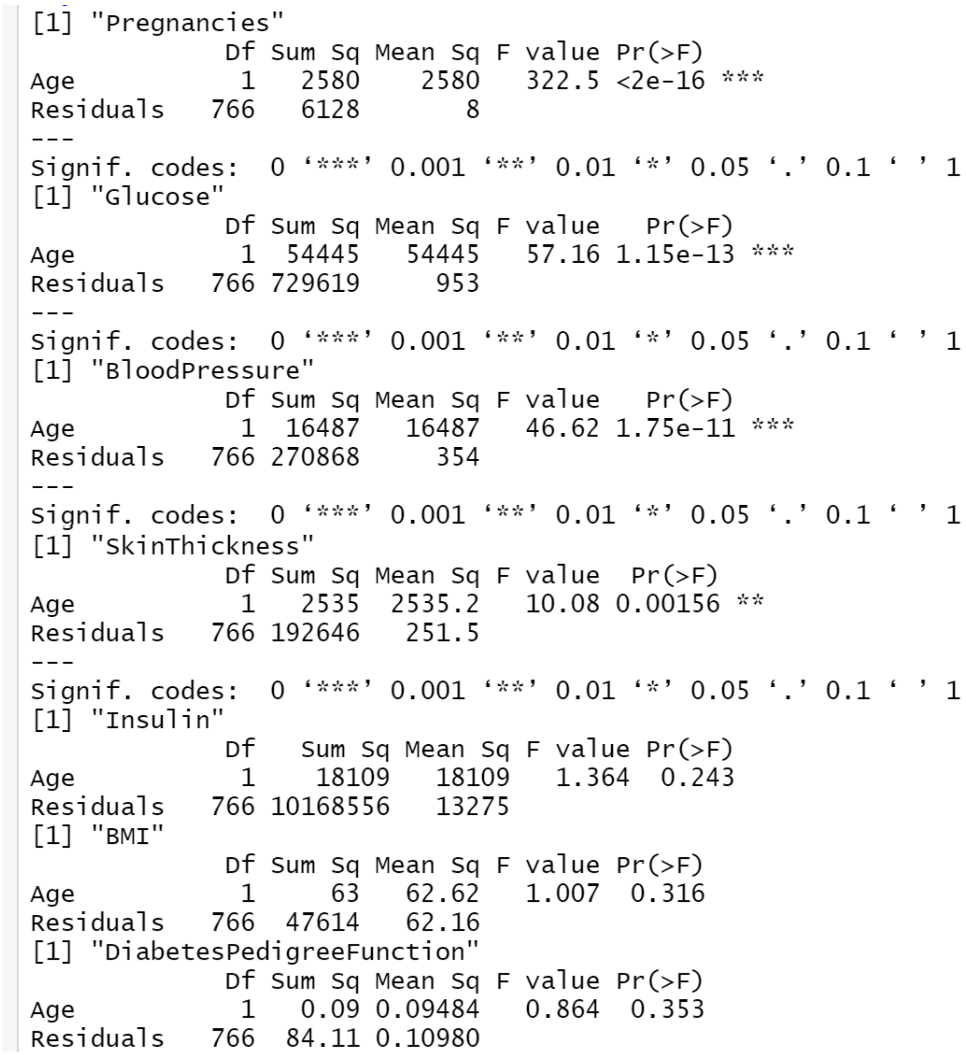


```
[1] "DiabetesPedigreeFunction"

        Durbin-Watson test

data:  AnovaModel.i
DW = 1.9864, p-value = 0.4253
alternative hypothesis: true autocorrelation is greater than 0


        Shapiro-Wilk normality test

data:  residuals(AnovaModel.i)
W = 0.83685, p-value < 2.2e-16
        studentized Breusch-Pagan test

data:  AnovaModel.i
BP = 0.014714, df = 1, p-value = 0.9035
```

Figure 28: Output of the different test for the category DiabetesPedigreeFunction with Age.

- In the Durbin-Watson test the $p - value > 0.05$, then the variables are independent.
- In the Shapiro test $p - value < 0.05$. So the sample doesn't follow a normal distribution.
- In the Breusch-Pagan test the $p - value > 0.05$. That means that in the variances there's homogeneity.

One of the ANOVA tests (Shapiro) don't succed so the ANOVA tests might be invalid as we said in the beggining of the exercise.



```
[1] "Pregnancies"
             Df Sum Sq Mean Sq F value Pr(>F)    
Age           1   2580    2580   322.5 <2e-16 ***
Residuals   766   6128       8                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Glucose"
             Df Sum Sq Mean Sq F value   Pr(>F)    
Age           1  54445   54445   57.16 1.15e-13 ***
Residuals   766 729619     953                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "BloodPressure"
             Df Sum Sq Mean Sq F value   Pr(>F)    
Age           1  16487   16487   46.62 1.75e-11 ***
Residuals   766 270868     354                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "SkinThickness"
             Df Sum Sq Mean Sq F value  Pr(>F)   
Age           1   2535  2535.2   10.08 0.00156 **
Residuals   766 192646   251.5                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Insulin"
             Df   Sum Sq Mean Sq F value Pr(>F)
Age           1    18109   18109   1.364  0.243
Residuals   766 10168556   13275               
[1] "BMI"
             Df Sum Sq Mean Sq F value Pr(>F)
Age           1     63   62.62   1.007  0.316
Residuals   766  47614   62.16               
[1] "DiabetesPedigreeFunction"
             Df Sum Sq Mean Sq F value Pr(>F)
Age           1   0.09 0.09484   0.864  0.353
Residuals   766  84.11 0.10980               
```

Figure 29: Output of Anova models (Age with categories)

In the ANOVA test of the risk factors: Pregnancies, Glucose, BloodPressure and SkinThickness, the $p-values$ (Pr(>F)) are below threshold of 0.05, so we can say that these categories are directly correlated with the Age category. However the categories like: Insulin, BMI and DiabetesPedigreeFunction are above the threshold, so we can say that they aren't directly correlated with the Age category.

To see the amount of influence of the age corresponding to the directly correlated categories we will use boxplots.

```
Boxplot(Pregnancies~Age,data=diabetes,id=FALSE)
Boxplot(Glucose~Age,data=diabetes,id=FALSE)
Boxplot(BloodPressure~Age,data=diabetes,id=FALSE)
Boxplot(SkinThickness~Age,data=diabetes,id=FALSE)
```

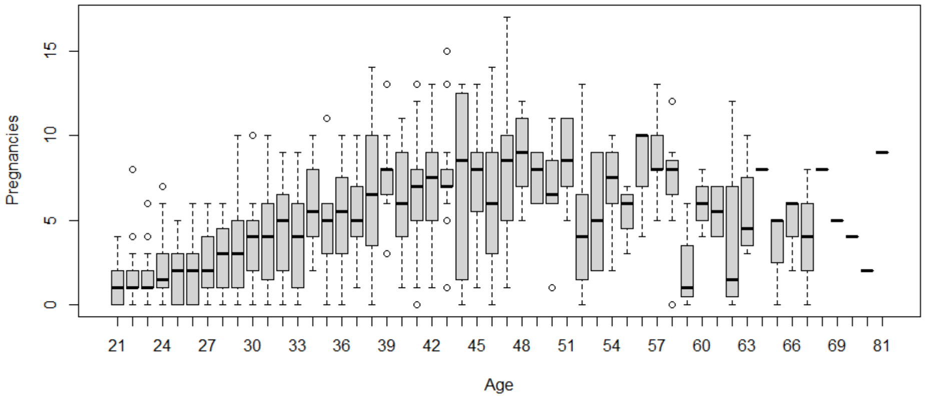


Figure 30: Boxplot Pregnancies respect Age.

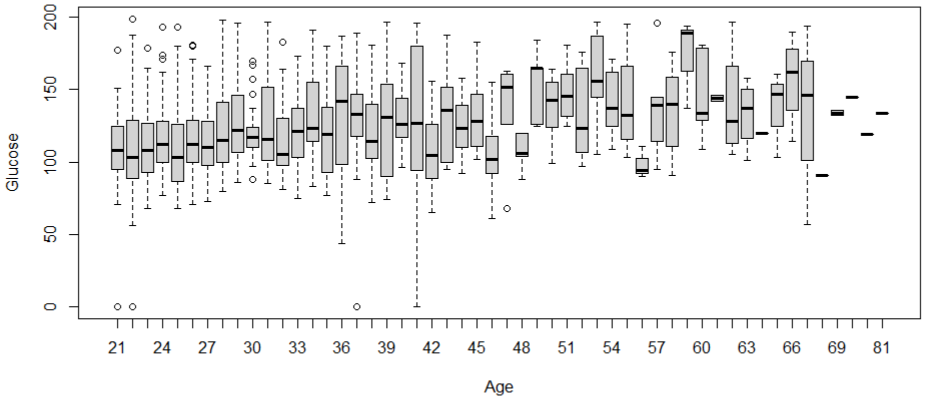


Figure 31: Boxplot Glucose respect Age.

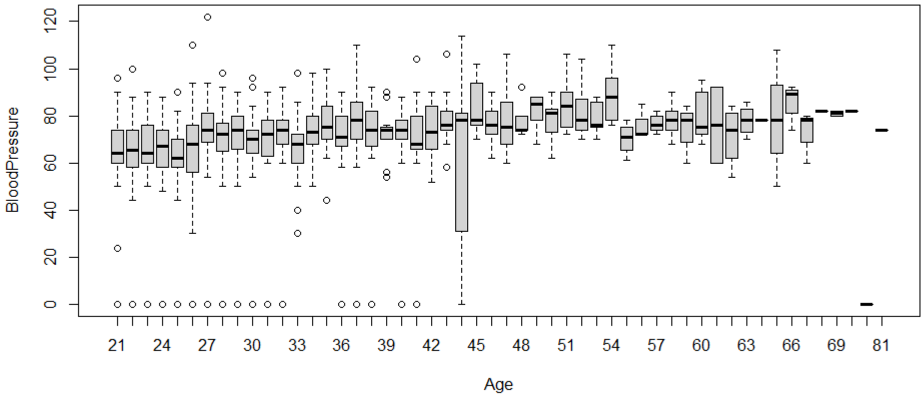


Figure 32: Boxplot BloodPressure respect Age.

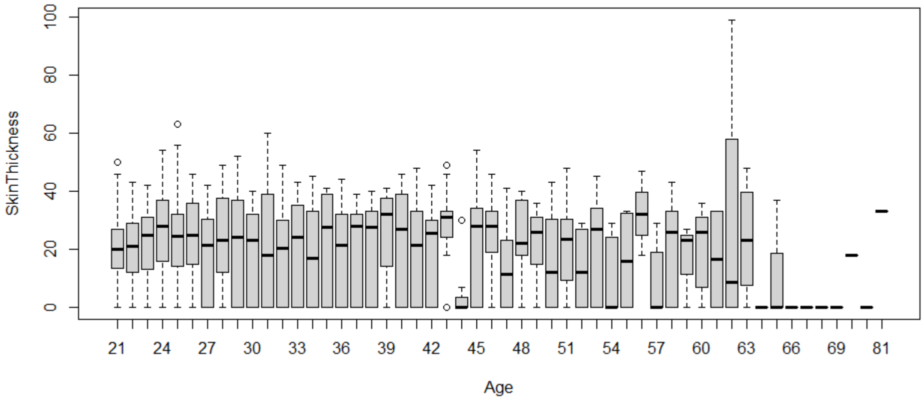
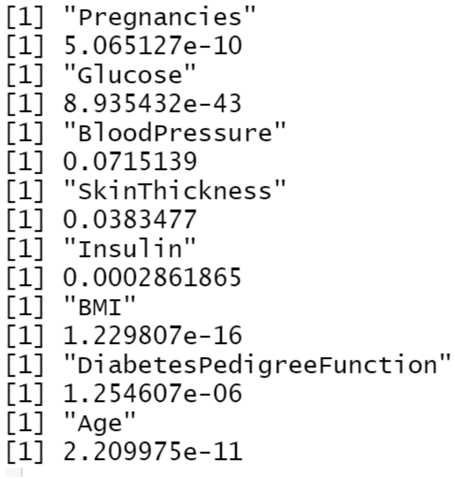


Figure 33: Boxplot SkinThickness respect Age.

2.3.2 Which of the risk factors are related with diabetes?

The fact of having diabetes is explicit in the outcome category, where 1 equals to have diabetes and 0 not having. To see which risk factors are related with diabetes we will do new models of ANOVA comparing each category with the Output.

```
for (i in 1:8){
  print((colnames(diabetes)[i]))
  AnovaModel.i<-aov(diabetes[,i]~Outcome, data=diabetes)
  print(summary(AnovaModel.i)[[1]][1,5]) #Extract the Pr(>F), p-value
}
```



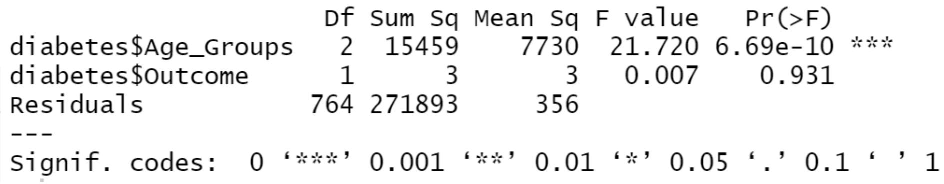
```
[1] "Pregnancies"
[1] 5.065127e-10
[1] "Glucose"
[1] 8.935432e-43
[1] "BloodPressure"
[1] 0.0715139
[1] "SkinThickness"
[1] 0.0383477
[1] "Insulin"
[1] 0.0002861865
[1] "BMI"
[1] 1.229807e-16
[1] "DiabetesPedigreeFunction"
[1] 1.254607e-06
[1] "Age"
[1] 2.209975e-11
```

Figure 34: P-values of the risk factors related with Diabetes (outcome).

As we can see the Two-Way Anova for the categories, Pregnancies, Glucose, SkinThickness, Insulin and DiabetesPedigreeFunction and age, shows us there are a correlation between them and diabetes parameter (outcome). However, the closer correlated category is Glucose because it has the smallest p-value in comparisson with the others, 8.93542e-43.

2.3.3 Detail the results of Two-Way ANOVA considering “Blood Pressure” as dependent variable, and the age groups and the indicator of diabetes as independent variables. Analyze the interaction term of two factors.

```
model<-aov(diabetes$BloodPressure~diabetes$Age_Groups+diabetes$Outcome)
summary(model)
shapiro.test(diabetes$BloodPressure) #Normality
dwtest(model) #Independency
bptest(model) #Homogeneity of variances
```



```
                     Df Sum Sq Mean Sq F value   Pr(>F)    
diabetes$Age_Groups   2  15459    7730  21.720 6.69e-10 ***
diabetes$Outcome      1      3       3   0.007    0.931    
Residuals           764 271893     356                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 35: Twoway ANOVA with BloodPressure as dependant by Age_Groups and Outcome.

The results show that only Age_Groups is very significant. This means that only the rank age of the people is related to the BloodPressure, so being in a specific range of the age will impact in the mean of the BloodPressure.

2.3.4 Analyze the distribution of Insulin variable. What would you recommend to fit an ANOVA model on Insulin levels?

3 Third Question

To start: load the package RCmdrPlugin.FactoMinerR.

Load the data “decathlon” located in the package.

The data represents a data frame with observations for different athletes.

```
install.packages("FactoMineR")
library(FactoMineR)
data(decathlon)
```

3.1 What is the linear expression that better predicts the behavior of an athlete for 1500m? Explore different expressions describing the power and the features of each one of them. Justify your answers

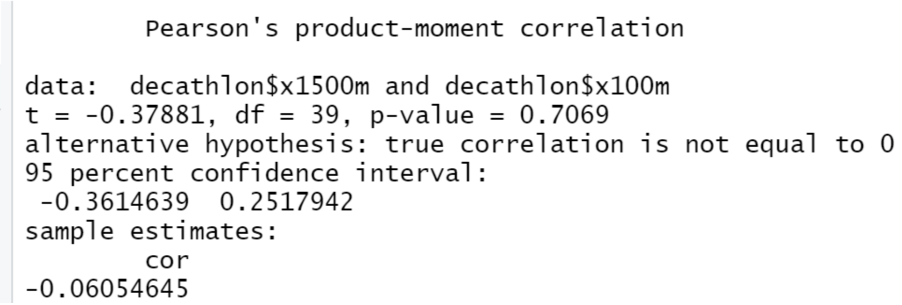
We have to add the character ”x” to each of the categories that we want to analyze, meter categories, because in other case, the function cor.test() doesn’t recognize correctly the parameters.

```
colnames(decathlon)[c(1,5,6,10)]<-c("x100m", "x400m", "x110m.hurdle", "x1500m")
```

First we will for each category (x100m, x400m, x1500m) the Pearson's correlation test. The test shows the correlation between two variables compared, the output values are between -1 and 1. -1 indicates that the correlation is totally inversely proportional and, 1 indicates that the correlation is totally directly proportional. Then the scatter plot with the regression line will show if there's linearity, i.e. positive correlation. After, we will create the simple linear regression model for each category and we'll see more parameters to check which of the categories fit better for the linear regression.

Category x100m

```
cor.test(decathlon$x1500m,decathlon$x100m)
```



```
        Pearson's product-moment correlation

data:  decathlon$x1500m and decathlon$x100m
t = -0.37881, df = 39, p-value = 0.7069
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3614639  0.2517942
sample estimates:
        cor
-0.06054645
```

Figure 36: Output Pearson's correlation test with x100m and x1500m.

The correlation between x100m and x1500m is -0.06054645, that means there's almost no correlation between this two categories.

```
scatterplot(x1500m~x100m, smooth=FALSE, data=decathlon)
```

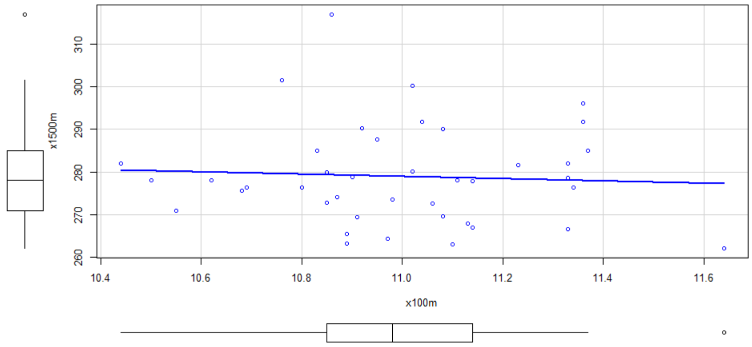
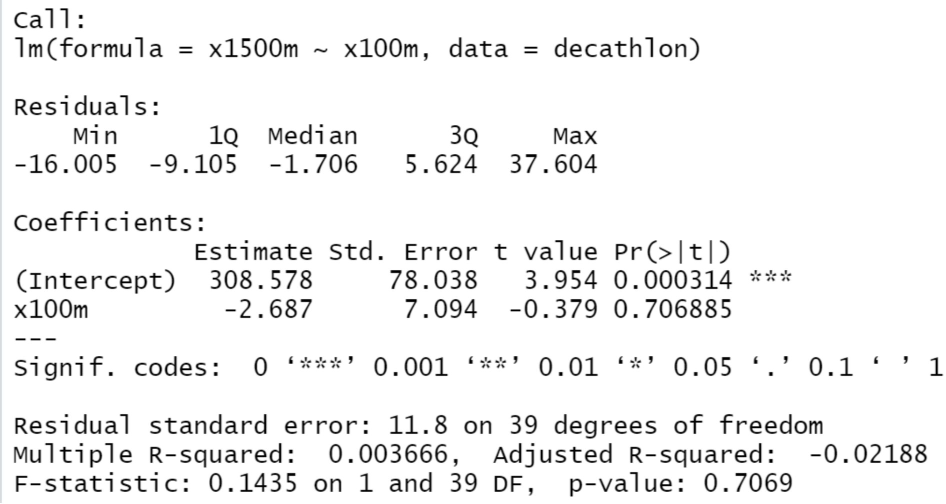


Figure 37: Scatter plot of x1500m and x100m.

The Scatter Plot of the x100m with x1500m doesn't show a good correlation as we saw with the Pearson's correlation test.

```
RegModel1 <-lm(x1500m ~ x100m, data=decathlon)
summary(RegModel1)
```



```
Call:
lm(formula = x1500m ~ x100m, data = decathlon)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.005  -9.105  -1.706   5.624  37.604 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  308.578     78.038   3.954 0.000314 ***
x100m         -2.687      7.094  -0.379 0.706885    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.8 on 39 degrees of freedom
Multiple R-squared:  0.003666,	Adjusted R-squared:  -0.02188 
F-statistic: 0.1435 on 1 and 39 DF,  p-value: 0.7069
```

Figure 38: Regression model of x1500m and x100m.

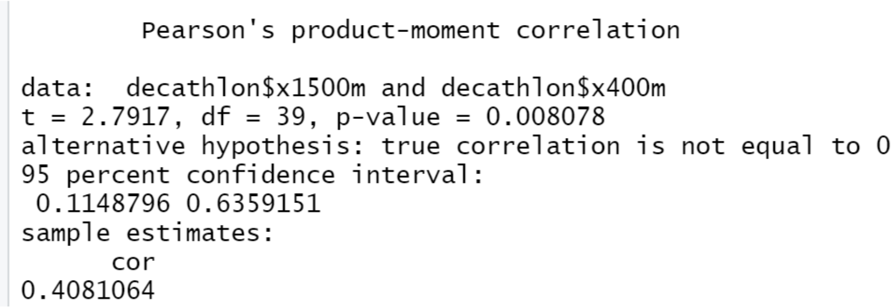
As we can see the $Pr(>|t|) = 0.000314$ of the category x100m shows a really very significant variable because it's almost close to the 0.

The residual standard error for x110m.hurdle is 11.8.

As we can see for the x100m the Multiple R-squared = 0.36 %. This means the regression model will not fit perfectly with the observations.

Category x400m

```
cor.test(decathlon$x1500m,decathlon$x400m)
```



```
        Pearson's product-moment correlation

data:  decathlon$x1500m and decathlon$x400m
t = 2.7917, df = 39, p-value = 0.008078
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1148796 0.6359151
sample estimates:
      cor
0.4081064
```

Figure 39: Output Pearson's correlation test with x400m and x1500m.

The correlation between x400m and x1500m is 0.4081064, that means there's a good proportional correlation between the categories.

```
scatterplot(x1500m~x400m, smooth=FALSE, data=decathlon)
```

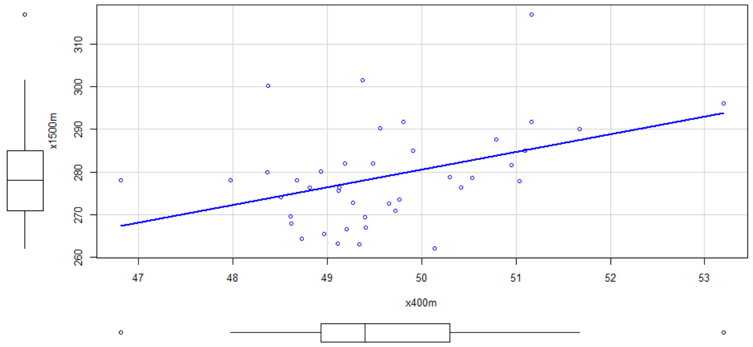
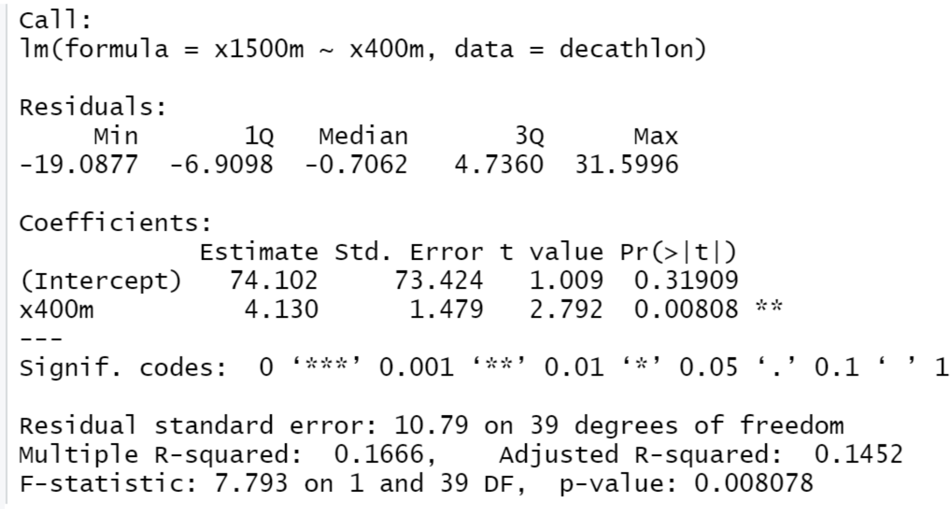


Figure 40: Scatter plot of x1500m and x400m.

The Scatter Plot of the x400m with x1500m shows a good correlation as we saw with the Pearson's correlation test.

```
RegModel2 <-lm(x1500m ~ x400m, data=decathlon)
summary(RegModel2)
```



```
Call:
lm(formula = x1500m ~ x400m, data = decathlon)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.0877  -6.9098  -0.7062   4.7360  31.5996 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   74.102     73.424   1.009  0.31909   
x400m          4.130      1.479   2.792  0.00808 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.79 on 39 degrees of freedom
Multiple R-squared:  0.1666,	Adjusted R-squared:  0.1452 
F-statistic: 7.793 on 1 and 39 DF,  p-value: 0.008078
```

Figure 41: Regression model of x1500m and x400m.

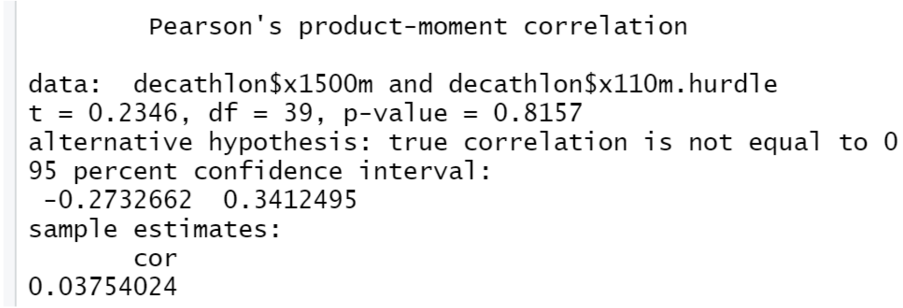
As we can see the $Pr(> |t|) = 0.00808$ of the category x400m shows a very signficant variable.

The residual standard error for x400m.hurdle is 10.79.

As we can see for the x400m the Multiple R-squared = 16.66 %. This means the regression model could fit well with the observations.

Category x110m.hurdle

```
cor.test(decathlon$x1500m,decathlon$x110m.hurdle)
```



```
	Pearson's product-moment correlation

data:  decathlon$x1500m and decathlon$x110m.hurdle
t = 0.2346, df = 39, p-value = 0.8157
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2732662  0.3412495
sample estimates:
       cor 
0.03754024 
```

Figure 42: Output Pearson's correlation test with x110m.hurdle and x1500m.

The correlation between x400m and x1500m is 0.03754024, that means there's almost no correlation between this two categories.

```
scatterplot(x1500m~x110m.hurdle, smooth=FALSE, data=decathlon)
```

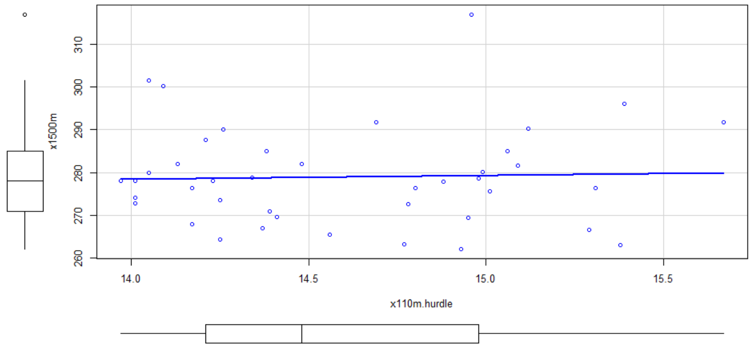
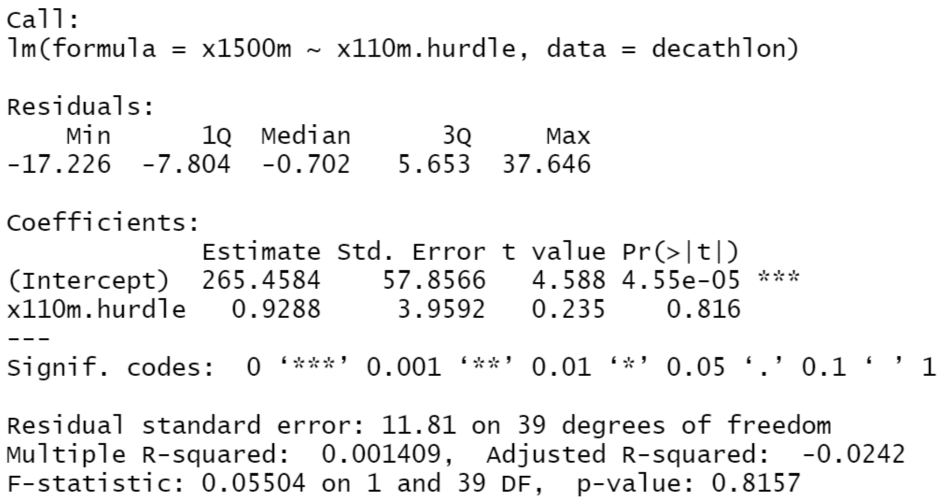


Figure 43: Scatter plot of x1500m and x110m.hurdle.

The Scatter Plot of the x110m.hurdle with x1500m doesn't show a good correlation as we saw with the Pearson's correlation test.

```
RegModel3 <-lm(x1500m ~ x110m.hurdle, data=decathlon)
summary(RegModel3)
```



```
Call:
lm(formula = x1500m ~ x110m.hurdle, data = decathlon)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.226  -7.804  -0.702   5.653  37.646 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  265.4584    57.8566   4.588 4.55e-05 ***
x110m.hurdle   0.9288     3.9592   0.235    0.816    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.81 on 39 degrees of freedom
Multiple R-squared:  0.001409,	Adjusted R-squared:  -0.0242 
F-statistic: 0.05504 on 1 and 39 DF,  p-value: 0.8157
```

Figure 44: Regression model of x1500m and x110m.hurdle.

As we can see the $Pr(>|t|) = 4.55e-05$ of the category x110m.hurdle shows a really very significant variable because it's almost close to the 0.

The residual standard error for x110m.hurdle is 11.81.

As we can see for the x110.mhurdle the Multiple R-squared = 0.1 %. This means the regression model will not fit perfectly with the observations.

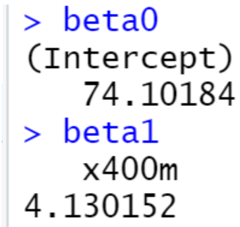
The model that we will use is the **model x1500m and x400m** because it has the smallest residual standard error, bigger Multiple R-Squared, good correlation value and positive correlation in the scatter plot. This model will have the best prediction.

We know that the regression line that helps us to predict the behaviour of given categories, with a certain confidence, follows the next definition:

$$\boxed{y = \beta_0 + \beta_1 x + \epsilon}$$

To get the values of β_0 and β_1 we can use the following function in R.

```
beta0<-RegModel2$coefficients[1]
beta0
beta1<-RegModel2$coefficients[2]
beta1
```



```
> beta0
(Intercept)
   74.10184
> beta1
   x400m
4.130152
```

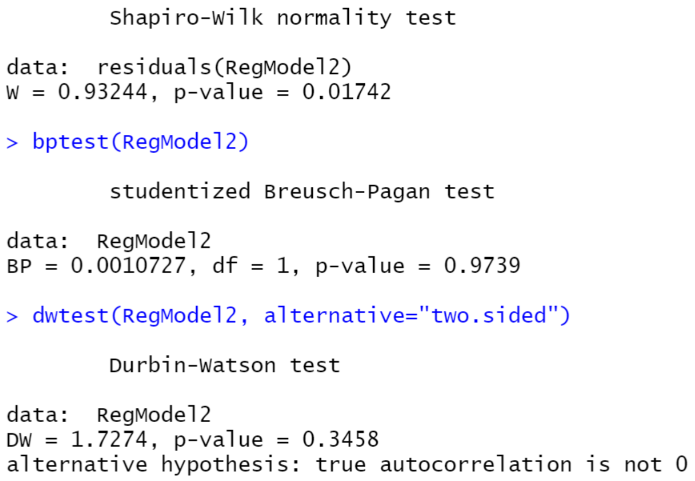
Figure 45: Beta values of the regression model.

So the function of the line is:

$$\boxed{y = 74.102 + 4.130x}$$

Finally we have to check for the assumptions, by doing the Shapiro-Wilk Normality, Breusch-Pagan and Durbin-Watson tests.

```
library(lmtest)
shapiro.test(residuals(RegModel2))
bptest(RegModel2)
dwtest(RegModel2, alternative="two.sided")
```



```
        Shapiro-Wilk normality test

data:  residuals(RegModel2)
W = 0.93244, p-value = 0.01742

> bptest(RegModel2)

        studentized Breusch-Pagan test

data:  RegModel2
BP = 0.0010727, df = 1, p-value = 0.9739

> dwtest(RegModel2, alternative="two.sided")

        Durbin-Watson test

data:  RegModel2
DW = 1.7274, p-value = 0.3458
alternative hypothesis: true autocorrelation is not 0
```

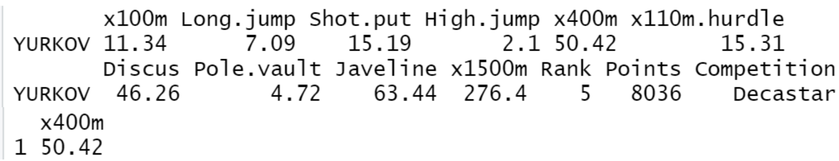
Figure 46: Output of the assumptions.

- In the Durbin-Watson test the $p - value > 0.05$, then the variables aren independent.
- In the Shapiro test $p - value < 0.05$. So the sample doesn't follow a normal distribution.
- In the Breusch-Pagan test the $p - value > 0.05$. That means that in the variances there's homogeneity.

3.2 Now use the expression to predict the behavior for a specific athlete. Analyze and explain the results obtained. Is the model accurate? What do you expect? Justify your answers.

To predict the behaviour of an specific athlete we will take a random athelete from all the decathlon dataset and his value for the category of x400m.

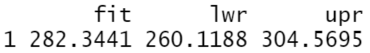
```
random_athlete<-data.frame(decathlon[runif(1, 1, 41),]) #random athlete of 41
random_athlete
random_athletex400m<-data.frame(x400m=random_athlete$x400m) #x400m of the random athlete
random_athletex400m
```



```
       x100m Long.jump Shot.put High.jump x400m x110m.hurdle
YURKOV 11.34      7.09    15.19       2.1 50.42        15.31
       Discus Pole.vault Javeline x1500m Rank Points Competition
YURKOV  46.26       4.72    63.44  276.4    5   8036    Decastar
  x400m
1 50.42
```

Figure 47: Data from the random athlete.

```
predict.lm(RegModel2, newdata=random_athletex400m, interval="prediction") #prediction with the regression model
```



```
       fit      lwr      upr
1 282.3441 260.1188 304.5695
```

Figure 48: Output of the prediction.

The random athlete runs the course of 1500 meters in 276.4 seconds. As we can see the prediction give us a value of 282.3441, this value is not equal to the real value but give an approximate value.

4 Fourth Question

4.1 Define a PCA for the decathlon dataset and discuss why the model you own works well (or not) looking the variables chart. We recommend using FactoMineR package to do this analysis.

```
library(FactoMineR)
data(decathlon)
competition <- which(colnames(decathlon) == "Competition")
plot(decathlon[,-c(competition)])
pca<-PCA(decathlon[,-c(competition)])
```

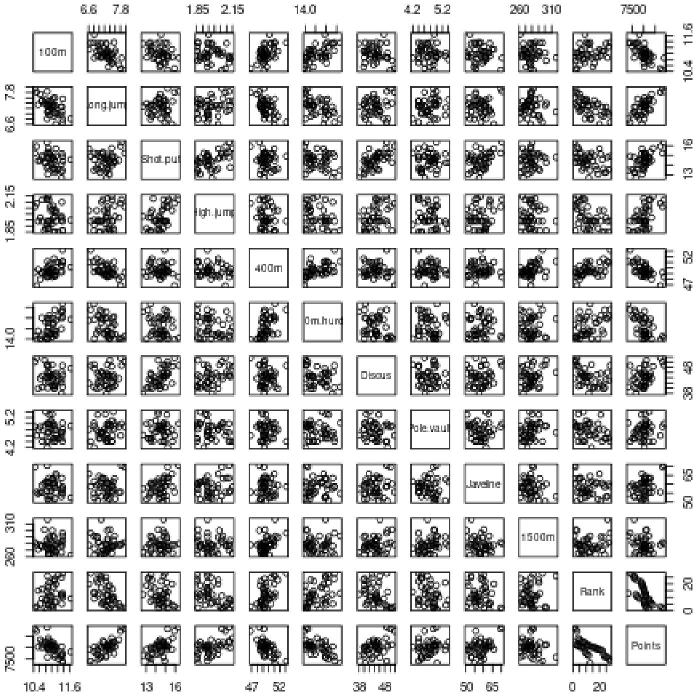


Figure 49: Correlation matrix.

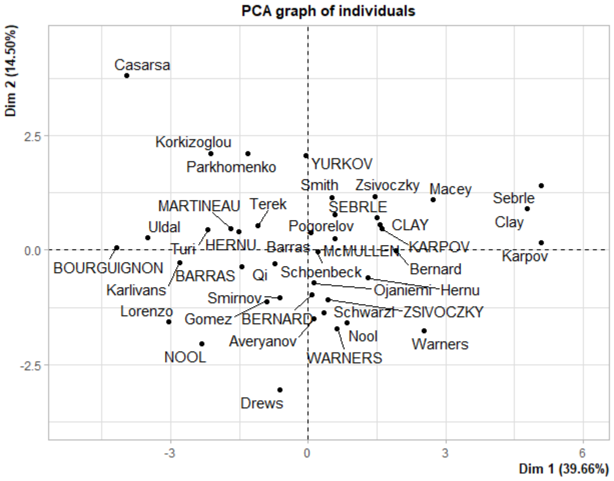


Figure 50: PCA graph of indivduals.

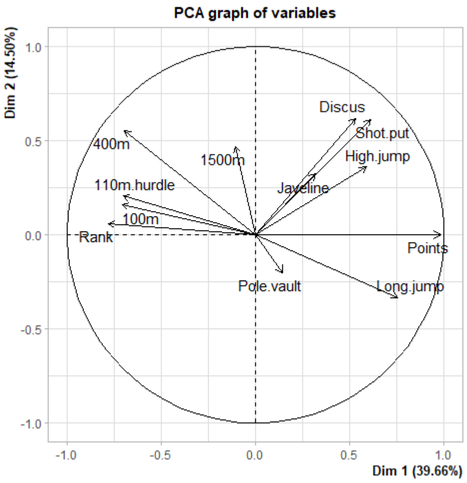


Figure 51: PCA graph of variables.

As we can see in the PCA graph of variables the more points you get the lower rank you get, where the rank implies the better position.

4.2 Now we are going to continue with principal component regression. We want to construct a linear regression model to predict the points of each athlete. To do so you must first decide the number of principal components to be included in the regression as independent variables. Justify your answer. Check the assumptions of the regression model. Is the prediction accurate enough?

To have an accumulative variance $\geq \frac{2}{3}$ we need to take at least 4 components.

```
competition <- which(colnames(decathlon) == "Competition")
pca$eig
plot(pca$eig[,1], type="o", main="Scree Plot")
```

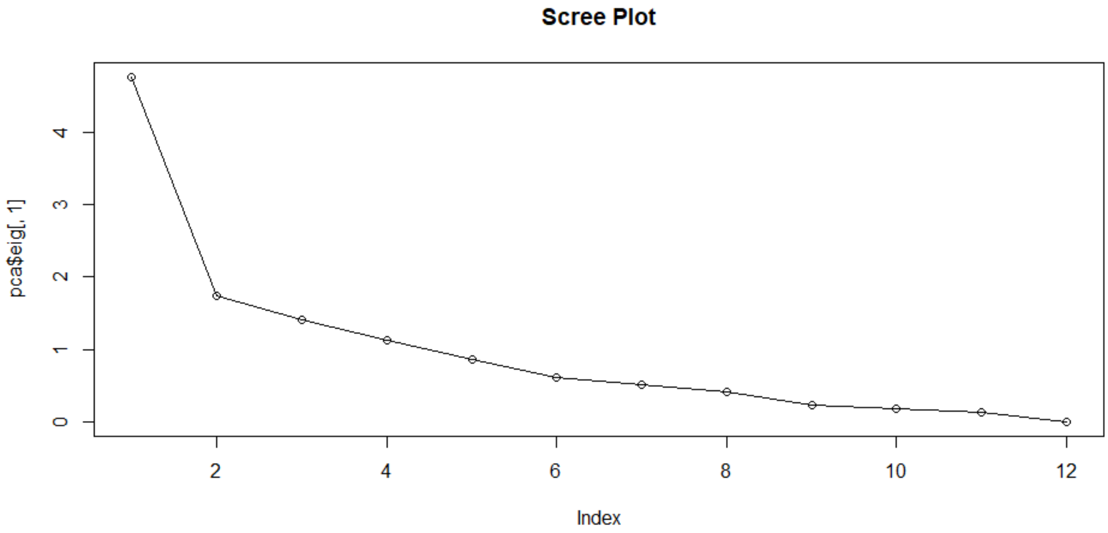
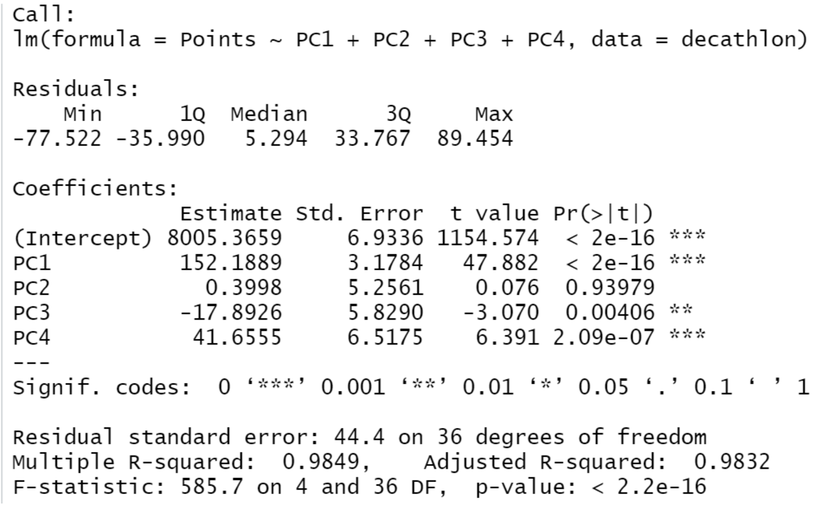


Figure 52: Scree Plot.

```r
summary(pca)
decathlon$PC1<-pca$ind$coord[,1]
decathlon$PC2<-pca$ind$coord[,2]
decathlon$PC3<-pca$ind$coord[,3]
decathlon$PC4<-pca$ind$coord[,4]

pc<-lm(Points~PC1 + PC2 + PC3 + PC4, data=decathlon)
summary(pc)
```

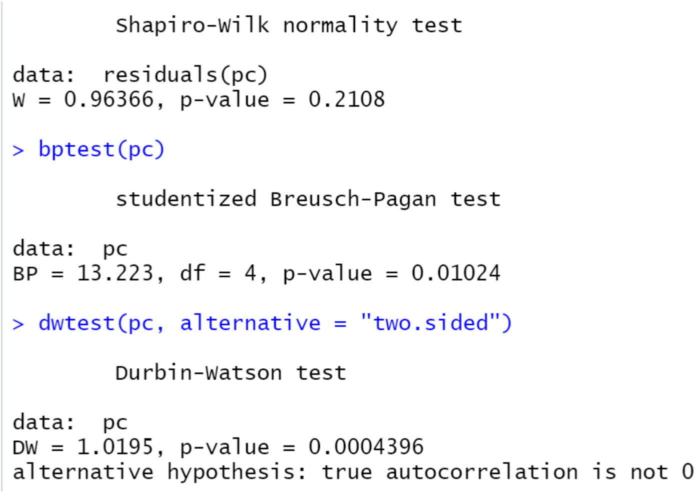


```
Call:
lm(formula = Points ~ PC1 + PC2 + PC3 + PC4, data = decathlon)

Residuals:
    Min      1Q  Median      3Q     Max 
-77.522 -35.990   5.294  33.767  89.454 

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)    
(Intercept) 8005.3659     6.9336 1154.574  < 2e-16 ***
PC1          152.1889     3.1784   47.882  < 2e-16 ***
PC2            0.3998     5.2561    0.076  0.93979    
PC3          -17.8926     5.8290   -3.070  0.00406 ** 
PC4           41.6555     6.5175    6.391 2.09e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.4 on 36 degrees of freedom
Multiple R-squared:  0.9849,	Adjusted R-squared:  0.9832 
F-statistic: 585.7 on 4 and 36 DF,  p-value: < 2.2e-16
```

Figure 53: Output linear model.

As we can see the categories as Intercept, PC1, PC3 and PC4 have significant values, $Pr(>|t|)$. Now we'll do the test for the linear model (pc).

```
shapiro.test(residuals(pc))
bptest(pc)
dwtest(pc, alternative = "two.sided")
```



```

        Shapiro-Wilk normality test

data:  residuals(pc)
W = 0.96366, p-value = 0.2108

> bptest(pc)

        studentized Breusch-Pagan test

data:  pc
BP = 13.223, df = 4, p-value = 0.01024

> dwtest(pc, alternative = "two.sided")

        Durbin-Watson test

data:  pc
DW = 1.0195, p-value = 0.0004396
alternative hypothesis: true autocorrelation is not 0
```

Figure 54: Output linear model.

- In the Durbin-Watson test the $p - value < 0.05$, then the variables aren't independent.
- In the Shapiro test $p - value > 0.05$. So the sample follows a normal distribution.
- In the Breusch-Pagan test the $p - value < 0.05$. That means that in the variances there's no homogeneity.

```
plot(residuals(pc))
```

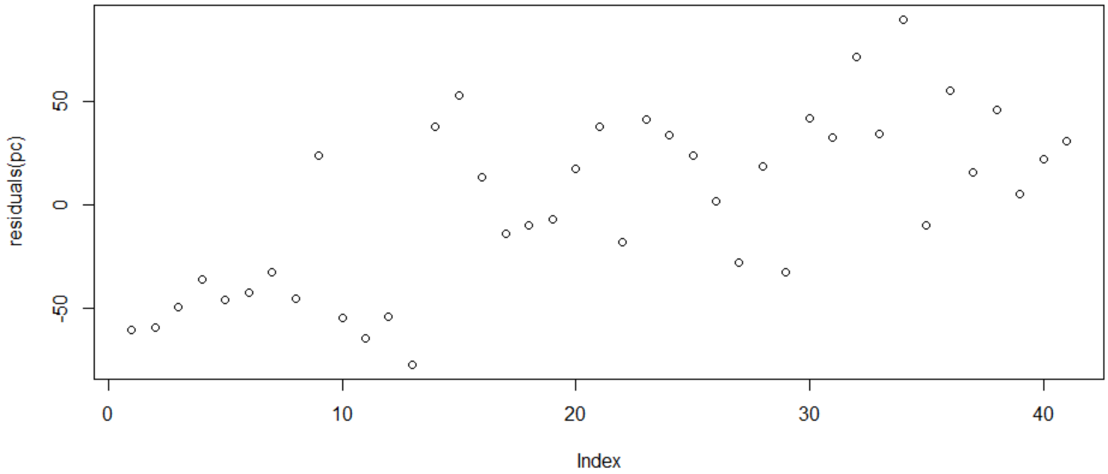
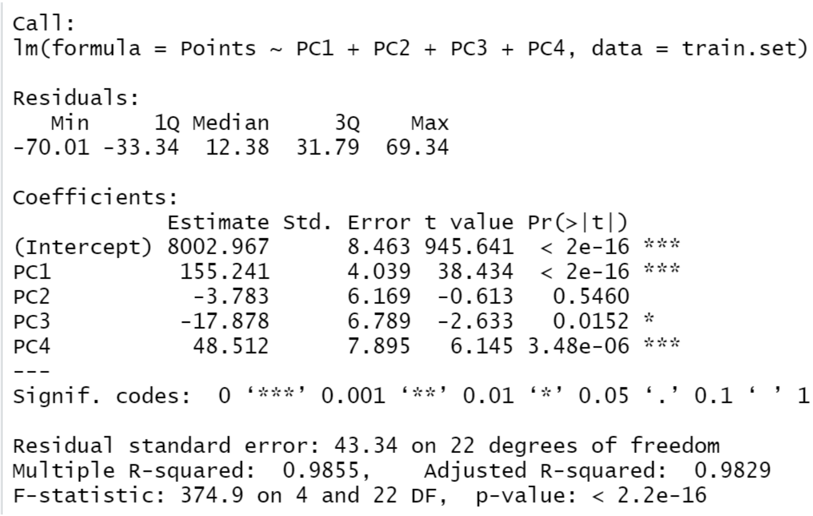


Figure 55: Plot linear model.

Finally we want to predict the points of the athlete by using the regression model. The RMSE is small to the points scale what it means that the model regression model generated is accurated with the real values.

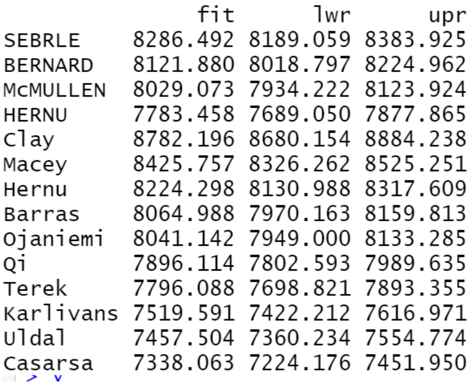
```
n <- nrow(decathlon)
train.sample <- sample(1:n, round(0.67*n))

train.set <- decathlon[train.sample, ]
test.set <- decathlon[-train.sample, ]

train.model <- lm(Points ~ PC1+PC2+PC3+PC4 , data = train.set)
summary(train.model)
```



```
Call:
lm(formula = Points ~ PC1 + PC2 + PC3 + PC4, data = train.set)

Residuals:
   Min     1Q Median     3Q    Max 
-70.01 -33.34  12.38  31.79  69.34 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8002.967      8.463 945.641  < 2e-16 ***
PC1          155.241      4.039  38.434  < 2e-16 ***
PC2           -3.783      6.169  -0.613   0.5460    
PC3          -17.878      6.789  -2.633   0.0152 *  
PC4           48.512      7.895   6.145 3.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.34 on 22 degrees of freedom
Multiple R-squared:  0.9855,	Adjusted R-squared:  0.9829 
F-statistic: 374.9 on 4 and 22 DF,  p-value: < 2.2e-16
```

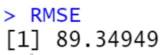
Figure 56: Output Train Model.

```
y<-predict(train.model, test.set, interval="prediction")
y
yt<-test.set$score
error<-cbind(y[,1,drop=FALSE],yt,(y[,1]-y)^2)
sqr_err<-error[,1]
sse<-sum(sqr_err)
RMSE<-sqrt(sse/(nrow(test.set))) #raiz(SSE/N)
RMSE
```



```
                fit      lwr      upr
SEBRLE     8286.492 8189.059 8383.925
BERNARD    8121.880 8018.797 8224.962
McMULLEN   8029.073 7934.222 8123.924
HERNU      7783.458 7689.050 7877.865
Clay       8782.196 8680.154 8884.238
Macey      8425.757 8326.262 8525.251
Hernu      8224.298 8130.988 8317.609
Barras     8064.988 7970.163 8159.813
Ojaniemi   8041.142 7949.000 8133.285
Qi         7896.114 7802.593 7989.635
Terek      7796.088 7698.821 7893.355
Karlivans  7519.591 7422.212 7616.971
Uldal      7457.504 7360.234 7554.774
Casarsa    7338.063 7224.176 7451.950
```

Figure 57: Predict output.



```
> RMSE
[1] 89.34949
```

Figure 58: MSE value for the regression model.