Body Performance Analysis

PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR THE COURSE
STAT 250 – APPLIED STATISTICS

DEPARTMENT OF STATISTICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

2429041-Doğa Evirgen
2429132-Samet Kenar
2429330-Yusuf Turan
2290831-Zeynep Öz

7/04/2022

Abstract

In this project, we analyzed a dataset called BodyPerformance in short, which is the Body Performance of Korean athletes Data Set. In this dataset, various performance types such as grip force, sit and bend forward, sit up count, and broad jump given with age, gender, height, weight, body fat, diastolic and systolic values can be said that influence the performance of Korean athletes. These variables are analyzed to understand if there exists a pattern at all. These traits are then analyzed to find a statistical significance of body performance with different statistical metrics and perspectives, such as Comparison of means, proportions, linearity testing, the mean significance by z-test, linear regression, and variance analysis. Results indicate that there are some significant dependencies between body performance and body measurements.

Contents

Introduction

In our analysis, we wanted to conduct our study around the effects of different determinants on athletes' body performance[1] by considering several given features of these athletes. By doing that, we wanted to observe how various factors affect Athlete's body performance. To conduct this study, we utilized seven different methods and identified multiple factors' effects on Athlete's body performance. The data is taken from Kaggle as provider Korea Sports Promotion Foundation.

Data Description

In the data set, Body Performance Data [1] we have twelve features which are; age, gender, height, weight, body fat, diastolic blood pressure (min), systolic blood pressure (min), grip force, sit and bend forward, sit-ups counts, broad jump, and class. Let us have a look at their descriptions. In the appendices Table-1, there is a complete description of each data variable with definitions. To explain the feature in Table 1, since Body Performance Data did not include NA values, we did not tackle omitting or removing them from the data set. Our data set consists of ten numeric variables and two characteristic variables.

Research Questions

We created some questions that came to our minds to better understand the data set in the path we desired to follow. The questions of our project can be seen in Table-2 in the appendices.

By answering these questions, we planned on achieving the aim of the study that is being explained in detail in the next part.

Aim of the Study

The overall aim of this study conducted is to understand if there are significant relationships between variables such as body fat, gender, and age. By conducting this study, we, as group_3, wanted to learn if there is a statistical significance, then further studies on the case can be conducted to improve the performance of athletes further.

Methodology/Analysis

In our study of Body Performance analysis, we used seven different statistical methods to answer our research questions.

The first method is one-sample hypothesis testing to make inferences about the mean. We wanted to answer our first question given in table-2 which is testing whether the average age of athletes is thirty or not. Although we are not given the population standard deviation, as the central limit theorem (As the sample size increases, sampling distribution follows a normal distribution) suggests, we used the z-test to test our hypothesis. According to the critical value that comes from the z-table and test value, we can either reject or support the null hypothesis [2].

The second method is two-sample hypothesis testing to make comparisons of means. By using this method, we wanted to answer our second question in the table-2 which is testing whether the average diastolic blood pressure of females and males is different. As it was mentioned in the previous paragraph, although we are not given the population standard deviation, because our sample size is more than enough, we used the z-test to decide on the hypothesis. According to the critical value and test value, we can reject or support the null hypothesis [3].

The third method is one-sample hypothesis testing used to make inferences from proportions. This method is used to answer the third question in table-2. Because both sample size multiplied by sample proportion and sample size multiplied by sample proportion subtracted from one is greater than 10, the z-test is used to test the hypothesis. After the critical value and test value is found, we can either reject or support the null hypothesis [4].

The fourth method is two-sample hypothesis testing which is used to make comparisons of proportions. By using this method, we wanted to answer the fourth

question given in table-2. The necessary points to use two-sample hypothesis testing on two-sample hypothesis testing are; samples being randomly selected, observations in the samples being independent of each other, and lastly, both sample sizes multiplied by proportions and sample sizes multiplied by proportions subtracted from one are greater than five. The last argument means that our samples are large enough to use the normal distribution. Because our samples are appropriate in terms of these necessities, it was acceptable for us to use two-sample hypothesis testing on comparisons of proportions by considering our samples follow a normal distribution. By comparing test value and critical value, we can either reject or support the null hypothesis[4].

The fifth method is simple linear regression. By using this method we wanted to answer our fifth question in table-2 and decide whether there is a significant relationship between our given variables. We have also considered if our data met the assumptions of simple linear regression by using the QQ Plot, Residuals and Fitted plot, and Scale-Location plot. How our data met the assumptions will be explained in the results and findings part of this report. By comparing the p-value and level of significance we can either reject or support whether there is a significant relationship between the variables [5].

The sixth method is multiple linear regression which is used to answer the sixth question given in the table-2 which is whether there is a significant relationship between more than one independent variable. As we did when we used simple linear regression, we checked whether our data met the assumptions of multiple linear regression by using a QQ Plot, Residuals and Fitted plot, and Scale-Location plot. How our data behaves according to assumptions will be explained in the results and findings part of this report. [5]

The last method is one-way ANOVA to answer the last research question given in table-2 which is testing if there is a difference among the means of given one categorical and one numerical variable. One-way ANOVA has three assumptions that must be considered before applying and we have also considered these assumptions by using QQ plot [9] and testing equality of variances which will be explained in the results and findings part of this report.

Results and Findings

In our one-sample hypothesis testing to answer the first question (please refer to Table-2 in the appendices for questions), the null hypothesis is stated as the average age of athletes is thirty. As it can be seen from the table-3, we have observed that our test value is greater than the critical value that was obtained from the z-table having a significance level of 0.05. To conclude, by referring to the test value and the critical value, we can reject the null hypothesis and answer our question by stating that there is enough evidence to support the alternative hypothesis that the average age of athletes is different than thirty [2].

In the second question, alternative hypothesis is the claim that the average diastolic blood pressure of females and males are different. As it can be seen from the table-4, the test value is a lot smaller than the critical value obtained by assuming the level of significance is 0.05 and using the z-table. By referring to these values, it can be observed the null hypothesis must be rejected [3]. Finally, there is enough evidence to support the claim that the average diastolic blood pressure of females and males is different.

In the one-sample hypothesis testing that was used to make inferences about proportions to answer the third question, alternative hypothesis is the claim that more than fifty percent of athletes' grip force is above forty. As it can be seen in table-5, the test value is smaller than the critical value obtained from the z-table and has the assumption that the significance level is 0.05. Referring to the test value and the critical value, it can be concluded from the failure of rejection of the null hypothesis that there is not enough evidence to support the claim that more than fifty percent of athletes' grip force is above

forty [4].

For our fourth question, we wanted to observe whether the number of class A female athletes is smaller than the number of class A male athletes, which is the alternative hypothesis of our two-sample hypothesis testing using comparisons of proportions. As it can be observed from table-6, the test statistic is greater than the critical value obtained from the z-table by having an assumption that the significance level is 0.05. As a result of these values, the decision is a failure to rejection of the null hypothesis and it can be concluded that there is not enough evidence to support the claim that the proportions of the female class A athletes is smaller than the male class A athletes [4].

For our fifth question, it was aimed to test whether there is a significant relationship between the measure of sit and bend forward and body fat. Before testing the significance, the assumptions of linear regression was observed [5]. The first assumption which is independence was met by our model because we assumed that our dataset was collected using statistically valid sampling methods and there are no hidden relationships among observations. The second assumption which is the normality of data is met because as it can be seen in figure-2, QQ Plot [9] approximately follows the straight line which is the sign of normality [5]. The third assumption is met because as it can be seen from figure-1, residuals versus fitted plot's red line is approximately horizontal which is the sign of linearity between independent and dependent variables [5]. The last assumption was met by the scale-location plot in figure-3. The red line does not deviate much from being horizontal which is a sign of homoscedasticity [5]. As we have met all the assumptions of simple linear regression, we have created our model and the result can be seen in table-7. From the result, it can be observed that p-value is greater than our significance level. Thus, there is no significant relationship between athletes' body fat and their measurement of the sit and bend forward [5].

For our sixth question, multiple linear regression was used to examine whether there is a significant relationship between grip force and systolic blood pressure, weight and body fat. Before testing, assumptions were checked [5]. Firstly, according to figure-5, data follows a normal distribution as the QQ Plot [9] follows an approximately perfect straight line [5]. Secondly, by figure-4, because the red line is approximately horizontal,

the relationship between independent and dependent variables is linear [5]. Thirdly, as it can be seen from figure-6, red line does not deviate much from being horizontal which can be concluded as no clear sign of heteroscedasticity [5]. Lastly, it was assumed that there are no hidden relationships among observations. As all assumptions are met, we can comment on the result in table-8. Our significance level is 0.05, thus it can be observed that because all independent variables' p-value is smaller than 0.04, systolic blood pressure, weight, and body fat are significant for grip force [5].

For our last question, one-way ANOVA is used to test the claim that there is no difference among the means of body fat between the four (A, B, C, D) classes. Assumptions were checked [6]. Firstly, we assumed the samples making up the four groups are independent from one another [6]. Secondly, as we can observe from figure 7, because QQ Plot [9] follows an approximately straight line, our observations in each group are normally distributed [6]. Thirdly, by referring table-9, because the result is smaller than 2, we can assume the equality of variances [6]. It can be concluded that our we met all the assumptions which means we can construct ANOVA. Table-10 is the result of the one-way ANOVA. From this table, it can be said that because p-value is smaller than the usual threshold of 0.05, we can say that there is a statistical difference among the means of body fat between groups [7].

Discussion/Conclusion

As we can see from some of our linear models, because the multiple r-squared values are not that high, our models are actually not that much powerful on the contrary of expecting higher values. On the other hand, we have concluded some points to answer our main purpose of project, that is to find the factors to increase the body performance of Korean athletes. Firstly, it should be concluded that most of the physical properties of female and male athletes are different, which can be understood that they should be classified separately. Secondly, mean body fat differs between classes of athletes so when an athlete's class wanted to be changed, body fat is one of the factors that should be considered. Also, if an athlete's grip force is wanted to be increased, body fat is one of the factors that should not be considered because there is no significant relationship between. Lastly, it cannot be concluded class A has more male athletes. . If there existed more time, we could have researched more and had a better look at the dataset so that we can make our models more powerful. For example, we could use more detailed plots to see the relations better.

References

[1] *Body Performance data*. (n.d.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/kukuroo3/body-performance-data

[2] Çamlı, O. (2022, April 1). *Recitation 3: Classical Statistical Inference* [PDF]. Middle East Technical University.

[3] Çamlı, O. (2022, April 15). *Recitation 5: Hypothesis Tests Concerning Two Populations Means* [PDF]. Middle East Technical University.

[4] Çamlı, O. (2022, April 22). *Recitation 6: Hypothesis Tests Concerning Population Proportions* [PDF]. Middle East Technical University.

[5] Çamlı, O. (2022, May 13). *Recitation 7: Linear Regression Models* [PDF]. Middle East Technical University.

[6] Çamlı, O. (2022, May 27). *Recitation 9: Analysis of Variance* [PDF]. Middle East Technical University.

[7] ANOVA in R: A step-by-step guide. (2022, May 6). Scribbr. https://www.scribbr.com/statistics/anova-in-r/

[8] *Blood pressure chart & numbers (Normal range, systolic, diastolic).* (2008, November 25). WebMD. https://www.webmd.com/hypertension-high-blood-pressure/guide/diastolic-and-systolic-blood-pressure-know-your-numbers

[9] QQ-plots: Quantile-quantile plots – R base graphs – Easy guides – Wiki – STHDA. (n.d.). STHDA – Accueil. http://www.sthda.com/english/wiki/qq-plots-quantile-quantile-plots-r-base-graphs#infos

Appendix

Table-1

Features	Definitions
Age	Shows the ages of the athletes, a numerical variable
Gender	Shows the genders of the athletes, a categorical variable
Height	Shows the height of each athlete in cm, a numerical variable
Weight	Shows the weight of each athlete in kg, a numerical variable
Body Fat	Shows the body fat of each athlete in percentages, a numerical variable
Diastolic Blood Pressure	Shows the minimum pressure in the blood vessels when the heart rests between beats
Systolic Blood Pressure	Shows the minimum force at which a person's heart pumps blood around the body
Grip Force	Shows the grip force of athletes in kilograms that was measured by a special machine
Sit and Bend Forward	Shows how long can athletes stretch in centimeters when they sit and bend forward to their feet
Sit-Ups Counts	Shows how many sit-ups athletes can do in two minutes
Broad Jump	Show how long can athletes jump in centimeters
Class	Shows the grade of performance of athletes. The lowest grade is considered as D and the highest grade is considered as A

Table-2

Q#	Questions
Q1	Is there enough evidence to support the claim that the average age of athletes is 30 at a 0.05 significance level?
Q2	Is there enough evidence to support the claim that the average diastolic blood pressure of females and males is different at a 0.05 significance level?
Q3	A sample of 1000 athletes showed that 422 had grip force above 40. Is there enough evidence to support the claim that more than 50% of all athletes' grip force is above 40 at a 0.05 significance level?
Q4	383 of randomly selected female athletes, 115 of them are class A athletes. 617 randomly selected male athletes, 124 of them are class A athletes. At the 0.05 significance level, is there enough evidence to support the claim that proportion of the female class A athletes is smaller than the proportion of male class A athletes?
Q5	Is there a significant relationship between sitting and bending forward in cm (dependent variable) and body fat (independent variable) at a 0.05 significance level?
Q6	Is there a significant relationship between grip force in cm (dependent) and systolic blood pressure (independent 1), weight(independent 2), and body fat(independent 3)? at 0.05 significance level.
Q7	Four classes of athletes were observed to learn about their body fat. It is desired to test the claim that there is no difference between the means.

Table-3

```
sample_mean <- mean(data$age)
sample_sd <- sd(data$age)
n <- nrow(data)
round((sample_mean-30)/(sample_sd/sqrt(n)),2)
```

```
## [1] 14.62
```

Table 3: Test statistic

Table-4

```
female_diastolic <- as.data.frame(data %>% subset(gender=="F") %>% select(diastolic))
female_mean <- mean(female_diastolic$diastolic)
#n-female = 383
female_var <- var(female_diastolic$diastolic)
male_diastolic <- as.data.frame(data %>% subset(gender=="M") %>% select(diastolic))
male_mean <- mean(male_diastolic$diastolic)
#n-male = 617
male_var <- var(male_diastolic$diastolic)

round(((female_mean-male_mean)-0)/sqrt((female_var/383)+(male_var/617)),2)
```

```
## [1] -7.75
```

Table 4: Test statistic

Table-5

```
p_bar <- 422/1000
standard_error <- sqrt((0.50*0.50)/1000)

(p_bar-0.50)/standard_error
```

```
## [1] -4.933153
```

Table 5: Test statistic

Table-6

```
pf_bar <- 115/383
pm_bar <- 124/617
p_bar <- 239/1000
q_bar <- 1-p_bar
round((pf_bar-pm_bar)/(sqrt((p_bar*q_bar)*((1/383)+(1/617)))),2)
```

```
## [1] 3.58
```

Table 6: Test statistic

Figure-1

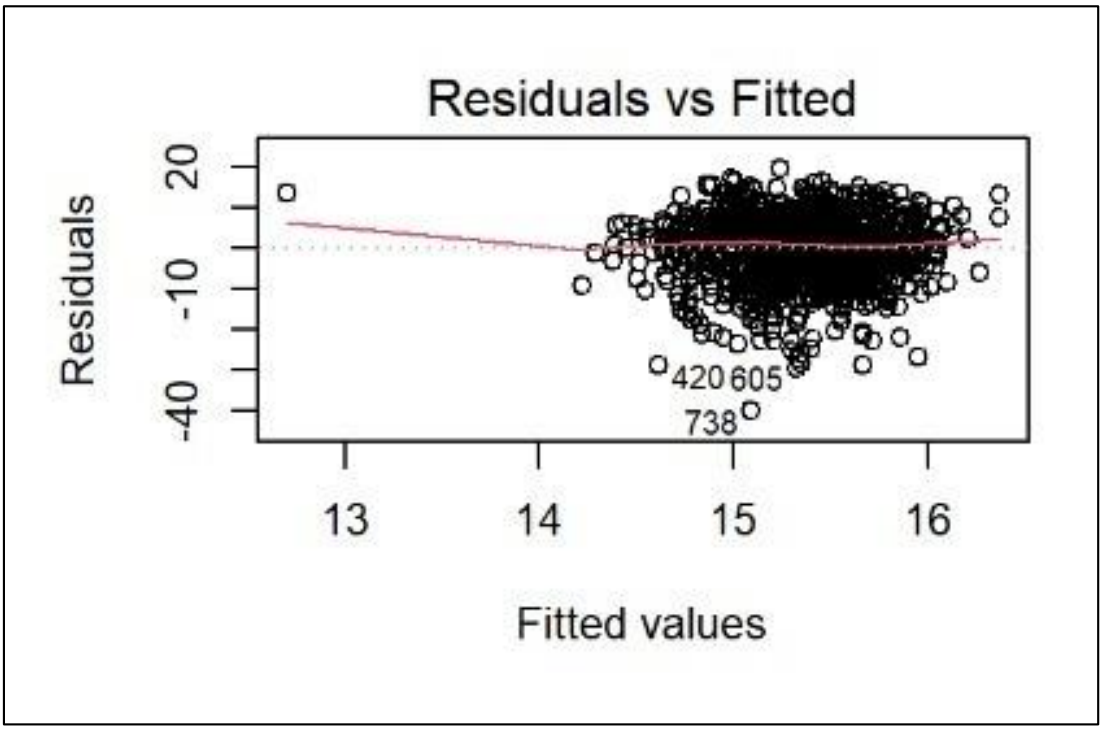

Figure 1: Representation of the linearity between the independent and dependent variable

Figure-2

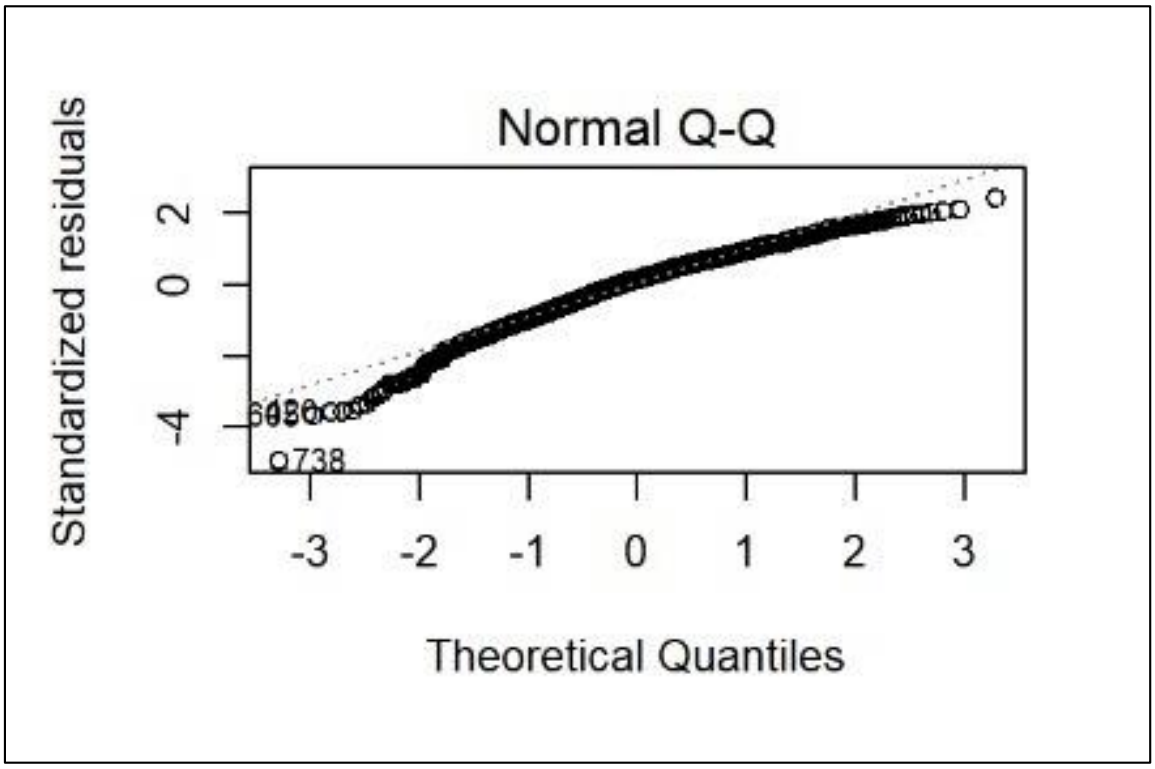

Figure 2: Normality test Q-Q plot

Figure-3

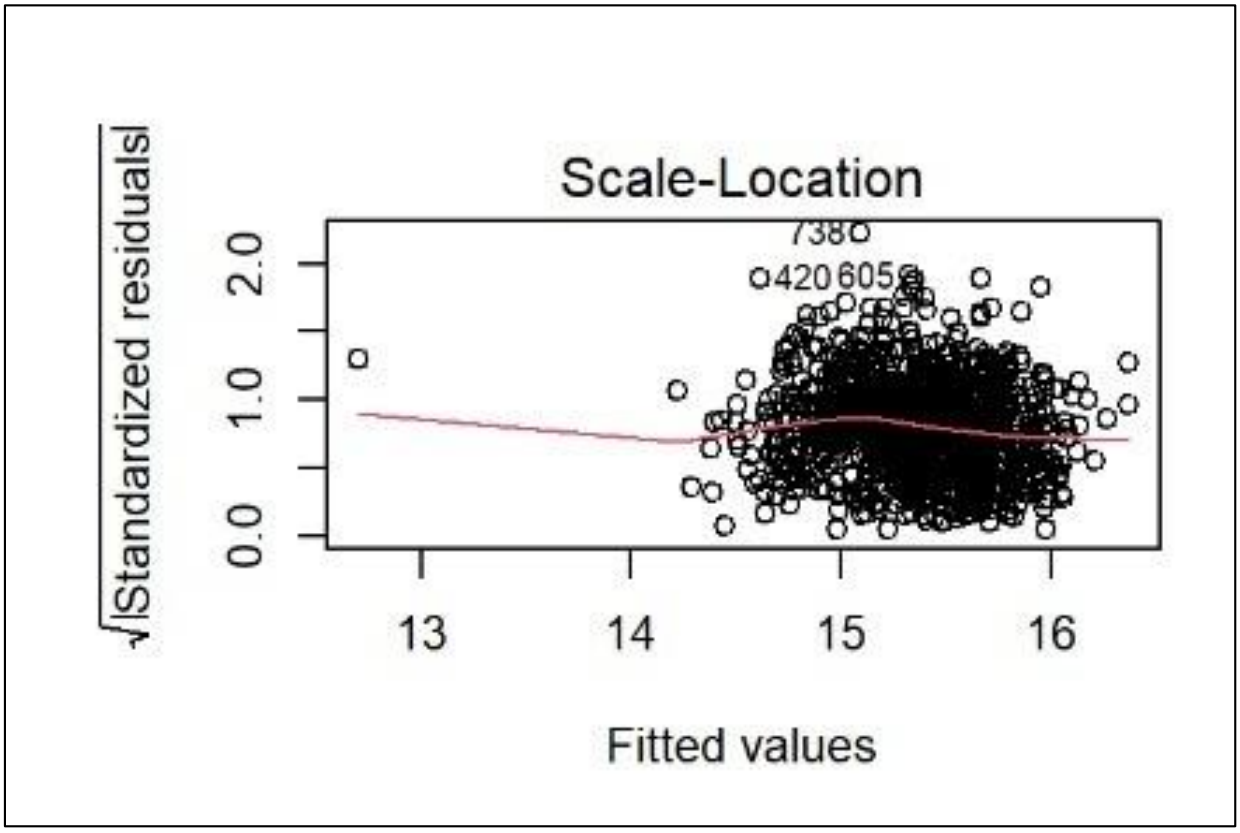

Figure 3: Representation of homoscedasticity of the data.

Table-7

```
##
## Call:
## lm(formula = sit.and.bend.forward_cm ~ body.fat_., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.093  -4.773   1.035   5.673  19.556
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.51372    0.85370  19.344   <2e-16 ***
## body.fat_.  -0.04865    0.03457  -1.407     0.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.112 on 998 degrees of freedom
## Multiple R-squared:  0.00198,    Adjusted R-squared:  0.0009803
## F-statistic:  1.98 on 1 and 998 DF,  p-value: 0.1597
```

Figure-4

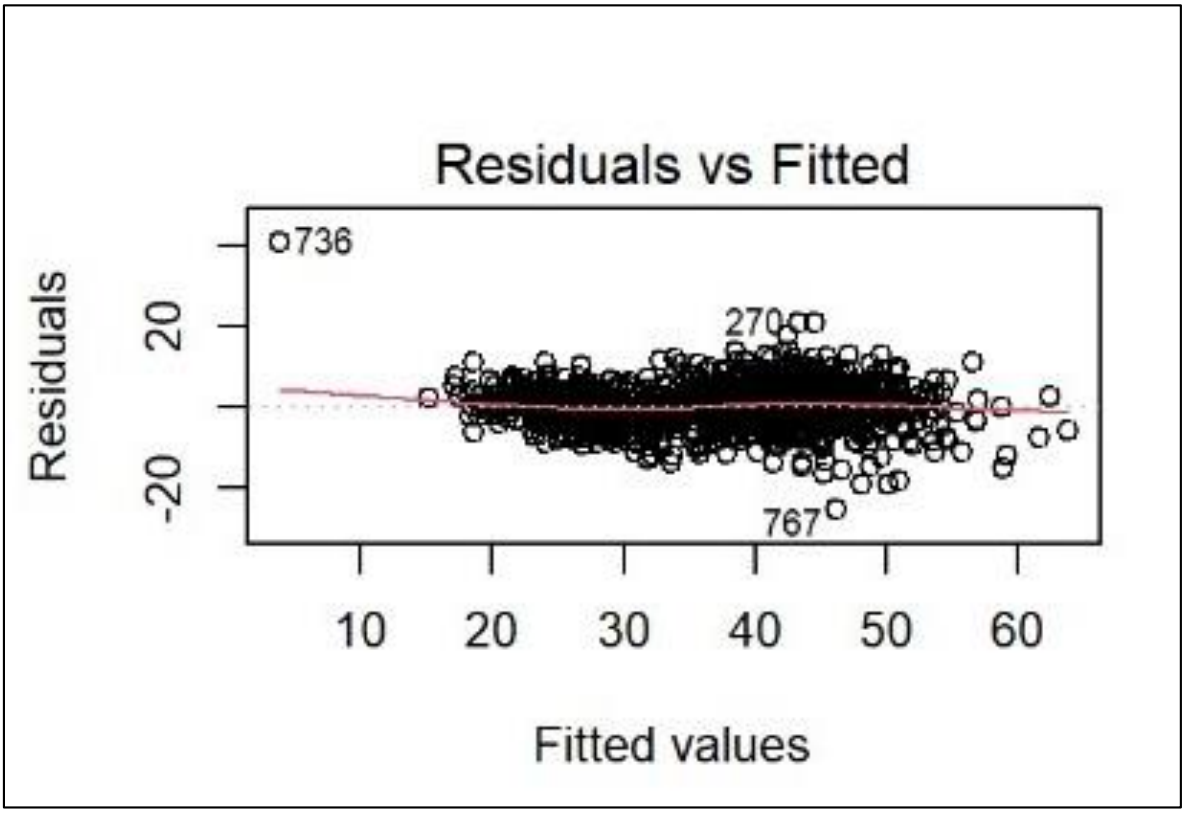

Figure 4: Representation of linearity between the independent and dependent variables.

Figure-5

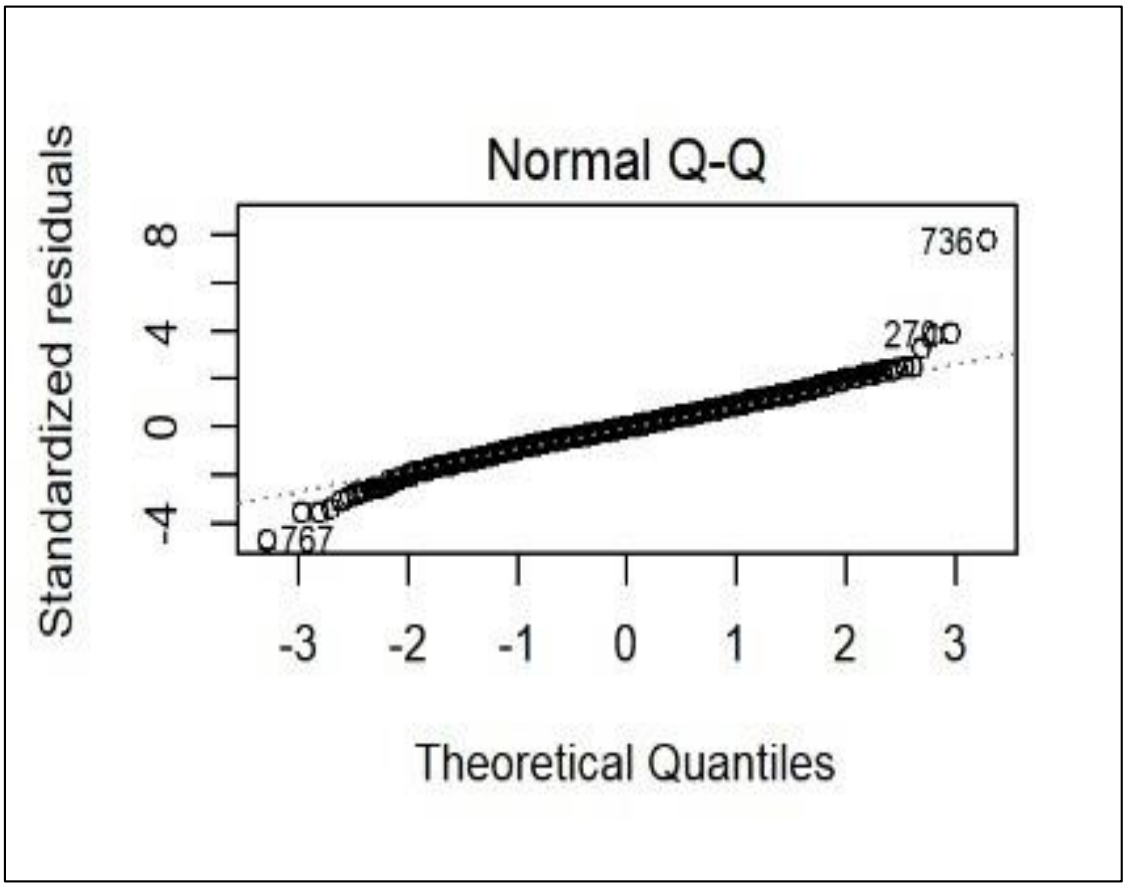

Figure 5: Normality test Q-Q Plot

Figure-6

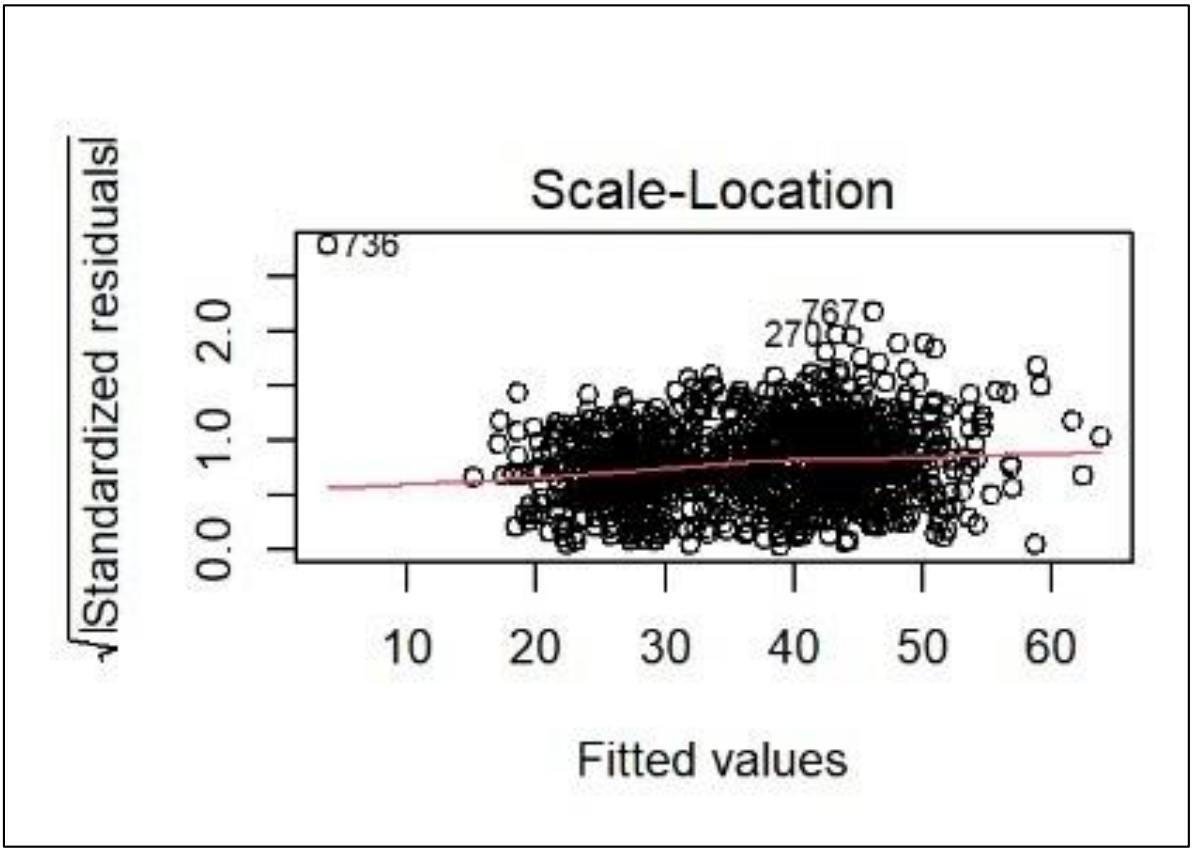

Figure 6: Representation of homoscedasticity of the data

Table-8

```
##
## Call:
## lm(formula = gripForce ~ systolic + weight_kg + body.fat_., data =
data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.641  -3.177  -0.016   3.307  41.105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.90402    1.70666   4.045 5.63e-05 ***
## systolic     0.05604    0.01244   4.506 7.38e-06 ***
## weight_kg    0.57011    0.01554  36.688  < 2e-16 ***
## body.fat_.  -0.66861    0.02350 -28.456  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.476 on 996 degrees of freedom
## Multiple R-squared:  0.7403, Adjusted R-squared:  0.7395
## F-statistic: 946.4 on 3 and 996 DF,  p-value: < 2.2e-16
```

Figure-7

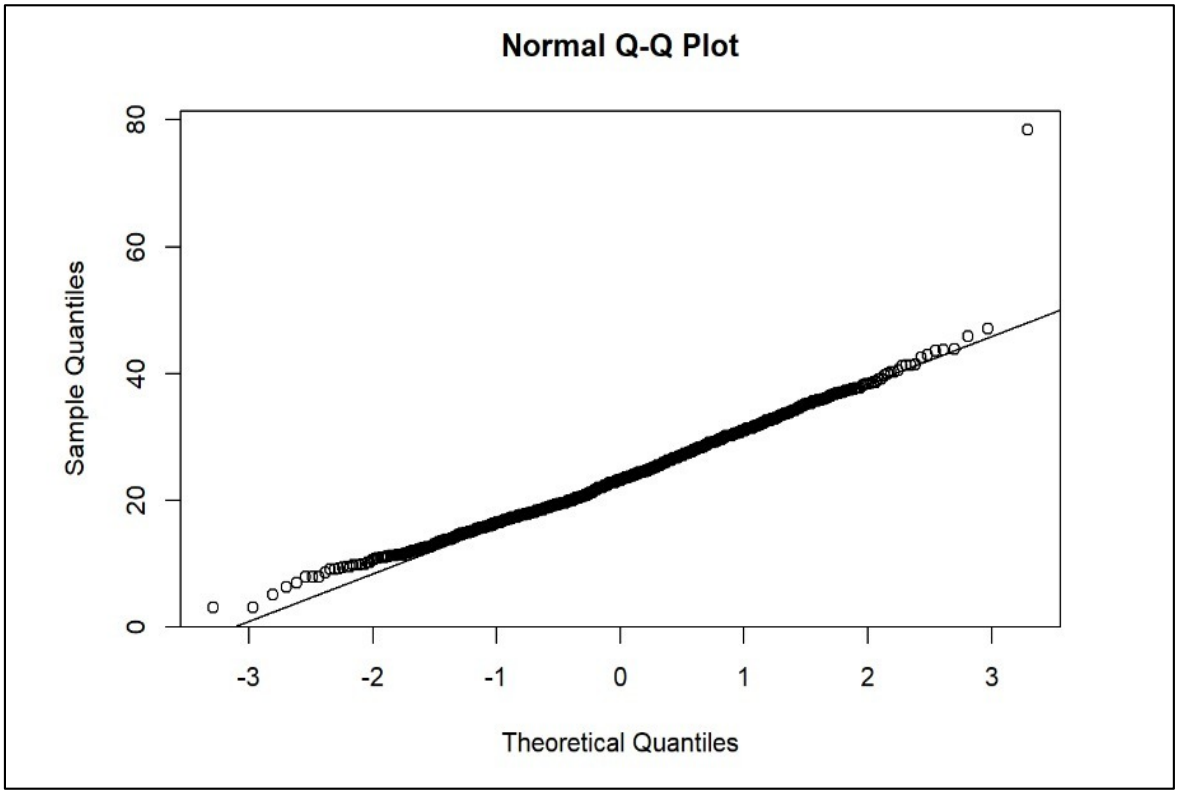

Figure 7: Normality test Q-Q Plot

Table-9

```
data.sds <- tapply(data$body.fat_.,INDEX=data$class,FUN=sd)
max(data.sds)/min(data.sds)
```

```
## [1] 1.207938
```

Table-10

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## class         3   5445  1814.8   36.43 <2e-16 ***
## Residuals   996  49624    49.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```