# ASSIGNMENT 3 :

**Introduction:**

This study investigates few questions about the relationship of the gender and higher education demand, the difference of final grade in higher education. The study also investigates the linear regression between first grade and final grades. The data is obtained from kaggle, and the data is based on a survey of students of math and Portuguese language courses in secondary school. The data contains information about sex, gender, age, address, mother and fathers jobs etc.

We will use Chi-Square test of association, Hypothesis test for independent samples, linear regression to solve the investigation. Moreover, some important statistical and descriptive stats and methods will help us show and analyse the results visually, such as boxplot, histogram and so on.

**Library Importing and loading Packages:**

```
# installing packages
Install.packages(“dplyr”)
Install.packages(“readr”)
Install.packages(“magrittr”)
Install.packages(“ggplot2”)
Install.packages(“Car”)
# dplyr included
library(dplyr)
library(readr)
library(magrittr)
library(lattice)
library(ggplot2)
library(car)
```

We have imported the required packages for the exploration of dataset.

**Data:**

When the data set is initially loaded into R, we find that the sex category is denoted by M’s and F’s.

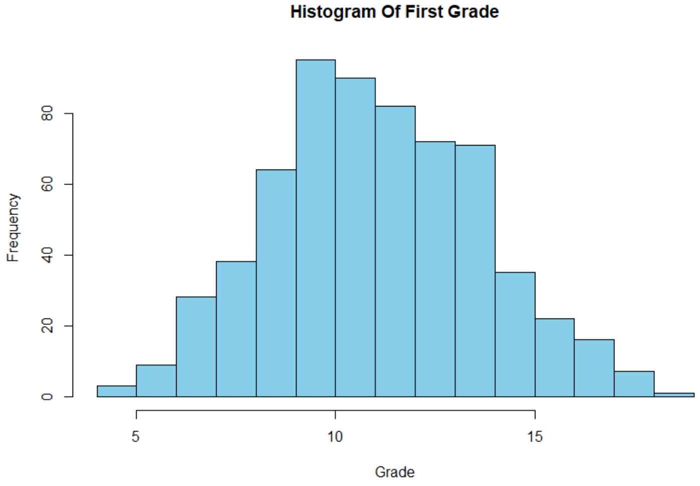
```
# importing the dataset
student_por<-studentpor
por <- select(student_por, "sex", "higher", "G1", "G3")
por$sex <- factor(por$sex, levels = c("M", "F"), labels = c("female", "male"))
# We need to filter test = 0 data
por1 <- por %>% filter(G3 > 0 & G1 > 0)
# Delete the 0 records.
por1 %>% head()
```

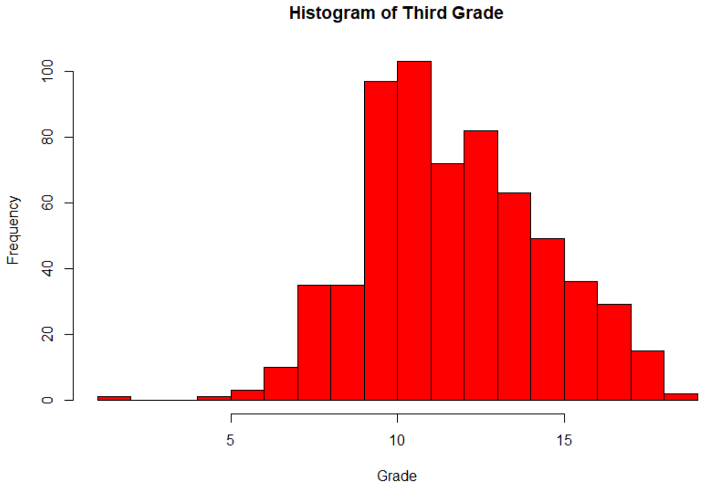
```
# A tibble: 6 x 4
  sex    higher    G1    G3
  <fct>  <chr>  <dbl> <dbl>
1 male   yes        9    11
2 male   yes       12    12
3 male   yes       14    14
4 male   yes       11    13
5 female yes       12    13
6 female yes       13    13
```

```
# Visualisation

por1$G1 %>% hist(main = "Histogram Of First Grade", breaks = 20, col = "skyblue", xlab = "Grad
e")

por1$G3 %>% hist(main = "Histogram Of Third Grade", breaks = 20, col = "red", xlab = "Grade")
```





## Summary Statistics:

By using group by function, we have descriptive stats such as mean, median, standard deviation, first and third quartile, inter quartile range etc;

```
# Summary Statistics generated by summarise.
```

```
por1 %>% group_by(sex)%>% summarise(Min = signif(min(G3,na.rm = TRUE),3),
                 Q1 = signif(quantile(G3,probs = .25,na.rm = TRUE),3),
                 Median = signif(median(G3, na.rm = TRUE),3),
                 Q3 = signif(quantile(G3,probs = .75,na.rm = TRUE),3),
                 Max = signif(max(G3,na.rm = TRUE),3),
                 Mean =signif( mean(G3, na.rm = TRUE),4),
                 SD = signif(sd(G3, na.rm = TRUE),4),
                 IQR = signif(IQR(G3, na.rm = TRUE),3),
                 n = n(),
```

```
                 Missing = sum(is.na(G3)))
```

```
sex       Min    Q1 Median    Q3   Max  Mean    SD   IQR     n Missing
  <fct>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
1 female     1    10     11  13.8    19  11.8  2.68  3.75   258       0
2 male       7    10     12  14      19  12.5  2.66  4      375       0
```

```
por1 %>% group_by(sex)%>% summarise(Min = signif(min(G1,na.rm = TRUE),3),
                 Q1 = signif(quantile(G1,probs = .25,na.rm = TRUE),3),
                 Median = signif(median(G1, na.rm = TRUE),3),
                 Q3 = signif(quantile(G1,probs = .75,na.rm = TRUE),3),
                 Max = signif(max(G1,na.rm = TRUE),3),
                 Mean =signif( mean(G1, na.rm = TRUE),4),
                 SD = signif(sd(G1, na.rm = TRUE),4),
                 IQR = signif(IQR(G1, na.rm = TRUE),3),
                 n = n(),
```

```
                 Missing = sum(is.na(G1)))
```

```
sex       Min    Q1 Median    Q3   Max  Mean    SD   IQR     n Missing
  <fct>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
1 female     4     9     11    13    18  11.2  2.56     4   258       0
2 male       5    10     12    14    19  11.8  2.68     4   375       0
```

Investigating the relationship between genders and desires of higher education with Chi-square. Hypothesis of Chi-Square test is an association where $H_0$: There is no association between gender and desire of higher education in the population of student. $H_A$: There is an association between gender and the desire of higher education in the population of student.

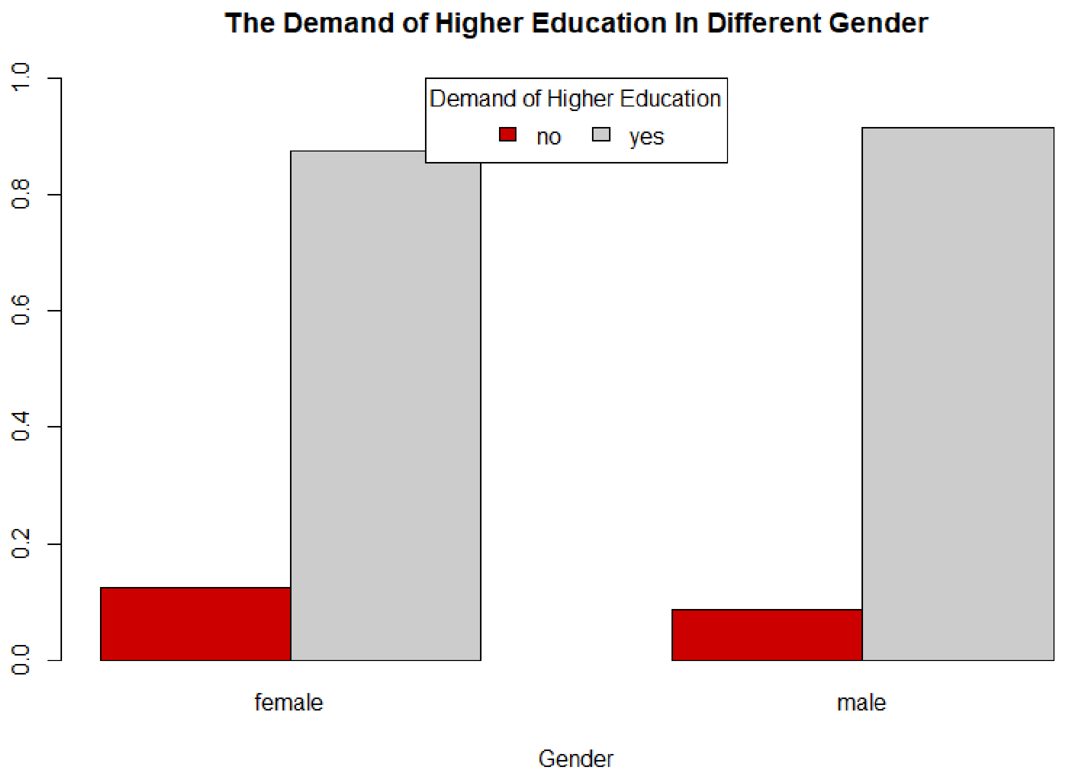
```
# # Try to find if the higher education demand is same in both male and female
tbl1 <- table(por1$higher, por1$sex)
knitr::kable(tbl1)
```

```
|    | female| male|
|:---|------:|----:|
|no  |     32|   32|
|yes |    226|  343|
```

```r
tb2 <- table(por1$higher, por1$sex) %>% prop.table(margin = 2)
knitr::kable(tb2)
```

```
|    |    female|      male|
|:---|---------:|---------:|
|no  | 0.124031| 0.0853333|
|yes | 0.875969| 0.9146667|
```

```r
# Provide a crosstabulation about gender and demand of higher education.
tb2 %>% barplot(main = "The Demand of Higher Education In Different Gender",
                ylim = c(0, 1),
                legend = rownames(tb1),
                beside = TRUE,
                args.legend = c(x = "top", horiz = TRUE, title = "Demand of Higher 
Education"),
                xlab="Gender",col = c("red3", "gray80"))
```



```r
# The proportion of higher education demand in different gender.
chi2 <- chisq.test(table(por1$higher,por1$sex))
chi2
```

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

data:  table(por1$higher, por1$sex)
```

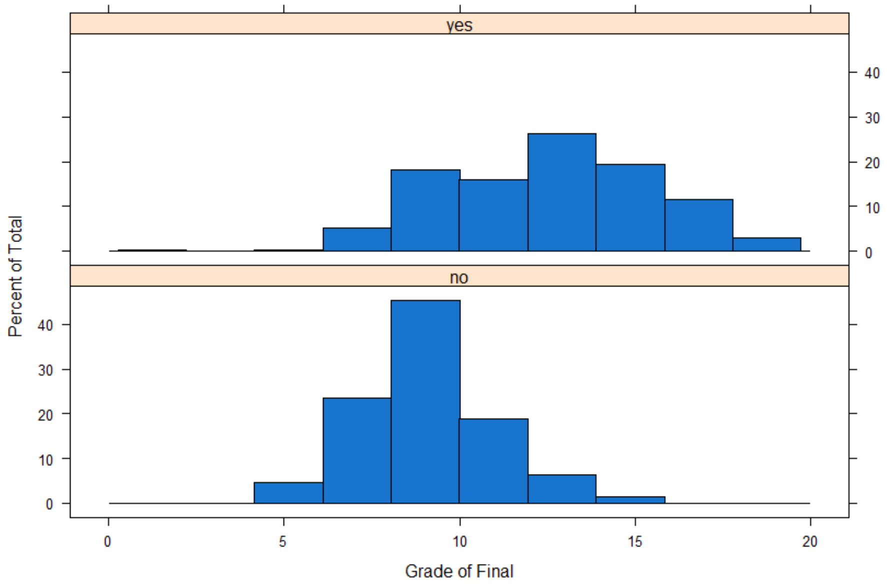
```
X-squared = 2.1106, df = 1, p-value = 0.1463
```

Investigating the difference of final grade in the relationship of higher education. We use boxplot to check whether we need to remove the outliers. As the difference is not wide, there remains no need to remove the outliers. We also use qqPlot to check the normality of the difference. Since we use of a large sample (n>30 for both groups), we do not worry about the normality.

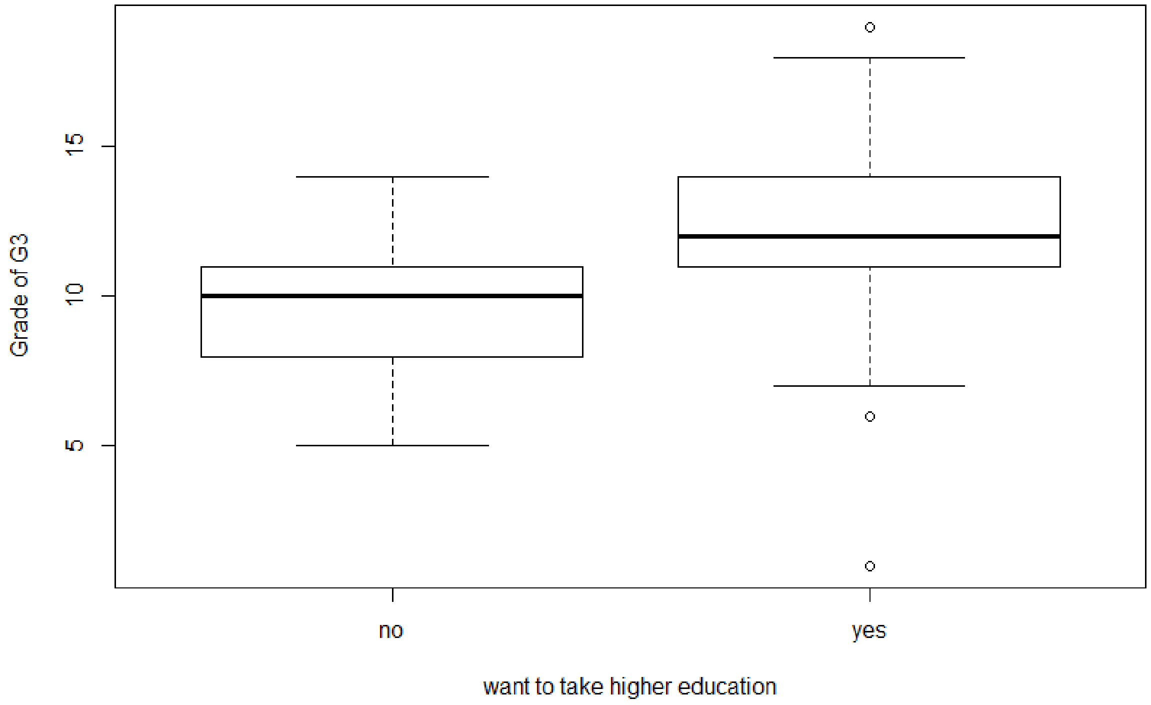
```
# summary of data
por1 %>% group_by(higher) %>% summarise(Mean = mean(G3,na.rm = TRUE),
                                        Median = median(G3, na.rm = TRUE),
                                        SD = sd(G3, na.rm = TRUE),
                                        Q1 = quantile(G3, probs = .25, na.rm = TRUE),
                                        Q3 = quantile(G3, probs = .75, na.rm = TRUE),
                                        Min = min(G3, na.rm = TRUE),
                                        Max = max(G3, na.rm = TRUE),
                                        IQR = IQR(G3, na.rm=TRUE),
                                        n = n()) -> table1
knitr::kable(table1)
```

```
|higher |      Mean| Median|       SD| Q1| Q3| Min| Max| IQR|   n|
|:------|---------:|------:|--------:|--:|--:|---:|---:|---:|---:|
|no     |  9.484375|     10| 1.708914|  8| 11|   5|  14|   3|  64|
|yes    | 12.493849|     12| 2.613692| 11| 14|   1|  19|   3| 569|
```

```
# Provide hisrogram of final grade divided by the desire of higher education.
por1 %>% histogram(~ G3|higher, col = "dodgerblue3",
                   layout = c(1, 2), data = ., xlab = "Grade of Final")
```

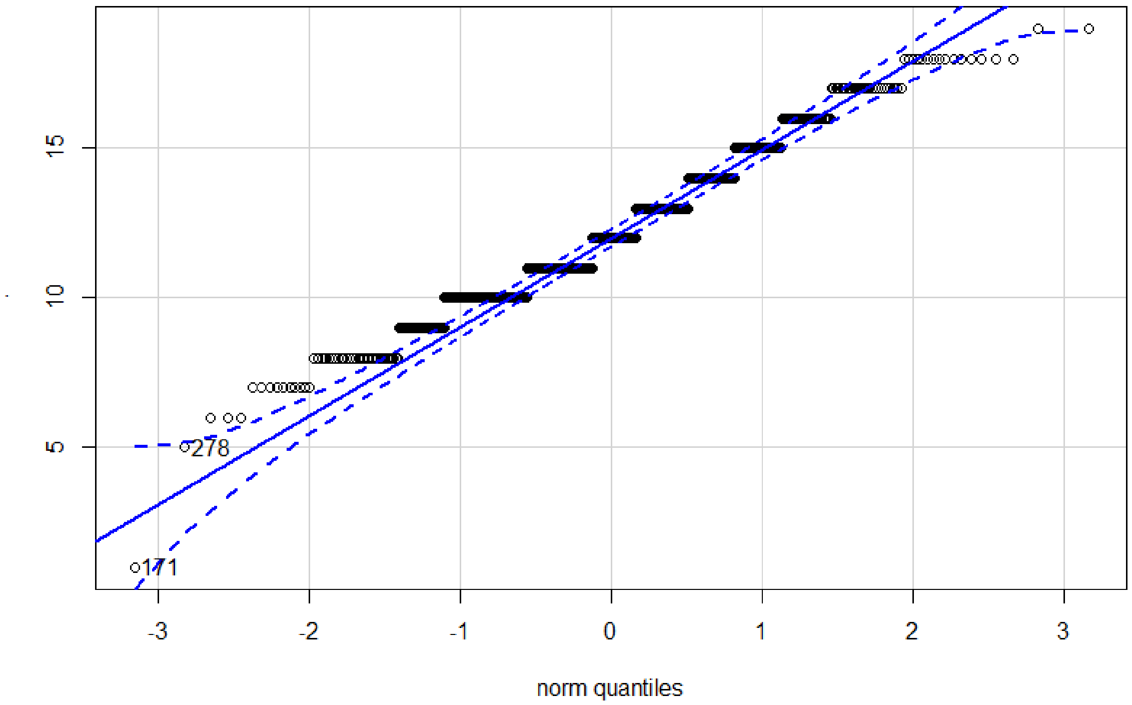


```
# boxplot
boxplot(
  por1$G3 ~ por1$higher,
  ylab = "Grade of G3",
  xlab = "want to take higher education")
```



## qqPlot

```
# qqplot
# Because the sample size greater than 30, it is acceptable if we ignore this step.(not necessary)
por1$G3 %>% qqPlot(dist="norm")
```



Hypothesis Testing: Hypothesis for two-sample t-test. The significance level is 0.05.

```
t.test(
  G3 ~ higher,
  data = por1,
  var.equal = FALSE,
  alternative = "less”)
```

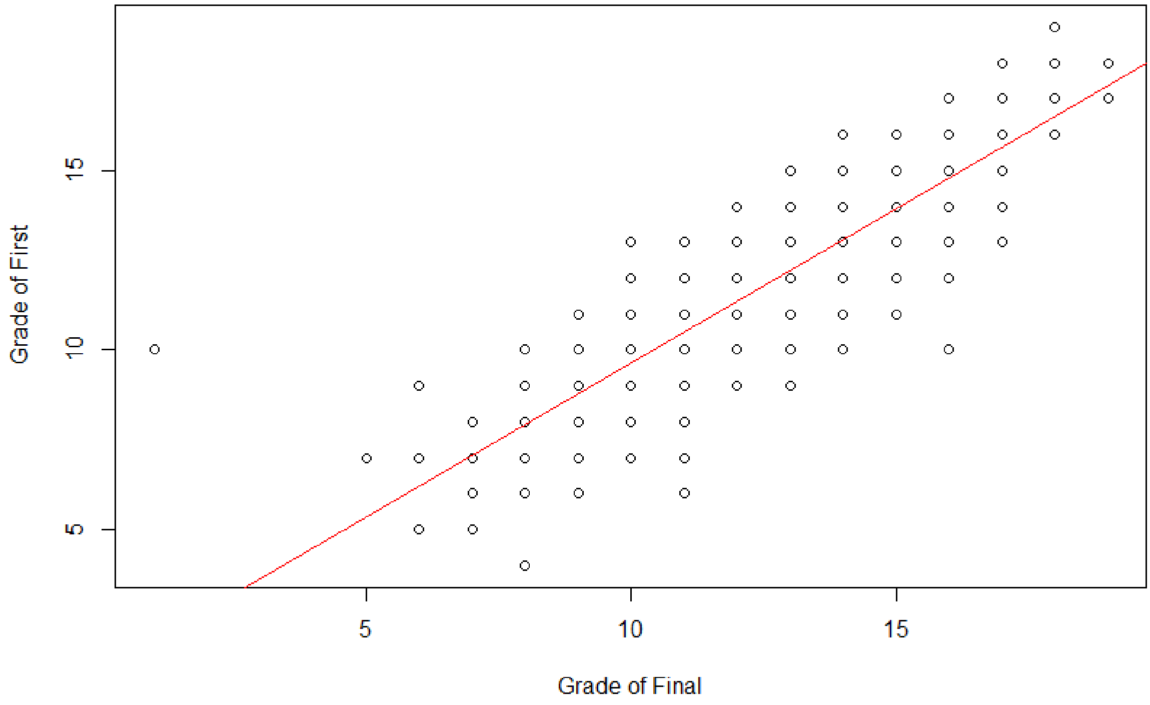
```
                    Welch Two Sample t-test

                     data:  G3 by higher
            t = -12.535, df = 99.747, p-value < 2.2e-16
  alternative hypothesis: true difference in means is less than 0
                  95 percent confidence interval:
                          -Inf -2.61088
                        sample estimates:
              mean in group no mean in group yes
                  9.484375          12.493849
```

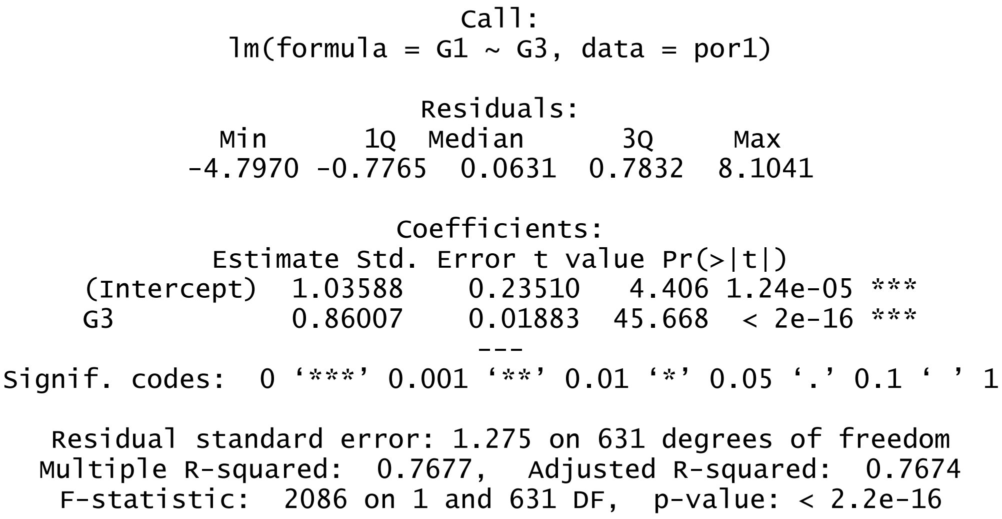
Investigating the correlation G1 and G3. Hypothesis for overall model:

$H_0$: The data does not fit the linear regression model. $H_A$: The data fit the linear regression model.

```
plot(G1 ~ G3, data = por1, xlab = "Grade of Final", ylab = "Grade of First")
R1 <- lm(G1 ~ G3, data = por1)
abline(R1, col = "red")
```



```
R1 %>% summary()
```



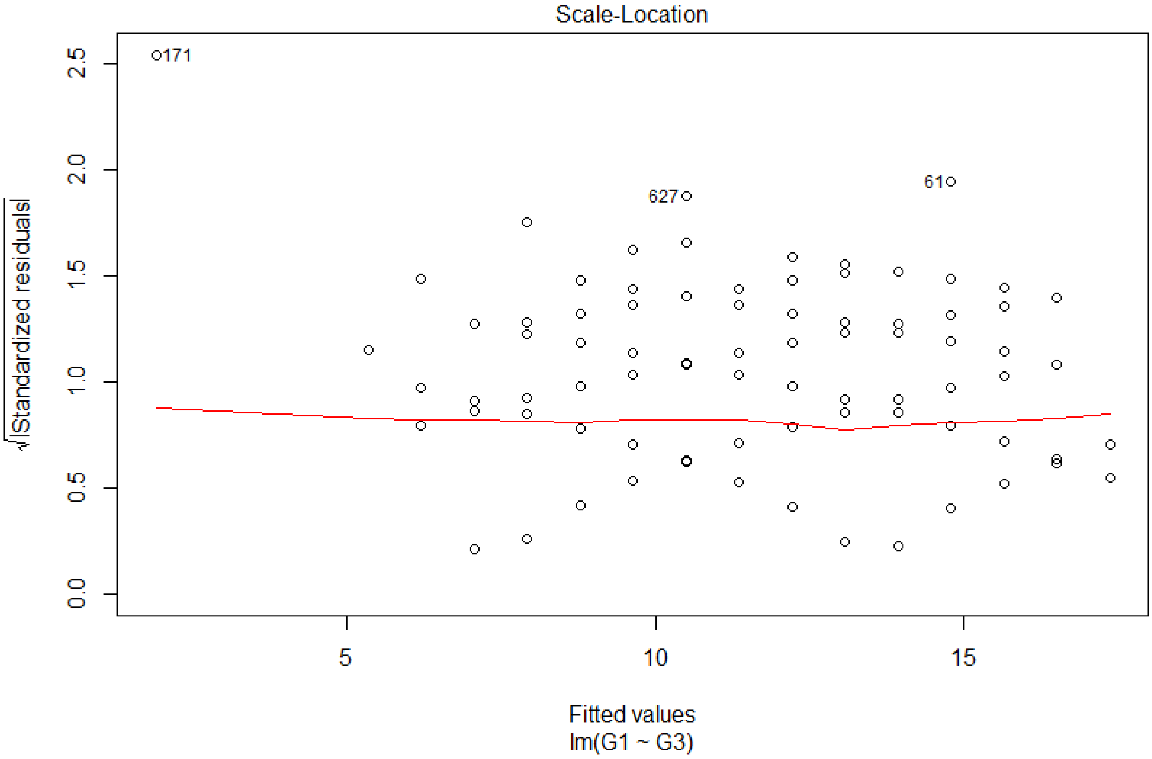
```
                         Call:
           lm(formula = G1 ~ G3, data = por1)

                       Residuals:
          Min       1Q  Median       3Q      Max
      -4.7970 -0.7765  0.0631  0.7832  8.1041

                      Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   (Intercept)  1.03588    0.23510   4.406 1.24e-05 ***
   G3           0.86007    0.01883  45.668  < 2e-16 ***
                           ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Residual standard error: 1.275 on 631 degrees of freedom
   Multiple R-squared:  0.7677,   Adjusted R-squared:  0.7674
     F-statistic:  2086 on 1 and 631 DF,  p-value: < 2.2e-16
```

```
Plot(R1)
```



**INTERPRETATION:**

For the first investigation, due to p-value = 0.1463 > 0.05 (significance level), we fail to reject $H_0$. The results of the test discovered that there are no statistical association between gender and higher education. The test has no significant difference compared to gender. Welch's two sample t-test states because p-value is less than 0.05 (significant level), we reject $H_0$. The results found a statistically significant mean difference between final grade and the desired education. For the further investigation the p-value is greater than 0.05 (significant level), the test found a statistically positive linear relationship between the first period grade and the final period grade. The advantage of the study is that we delete zero grade which minimise basis and grade. The limitation is only that we analyse students who study Portuguese language. The result is only suitable for students from Portuguese language course in the secondary school.