Residuals for the Simple regression model

Nithya S

2023-11-24

Objectice

Take a suitable data set for the Simple linear regression model and analyze it by establishing the linear relationship between the variables and hence examine the various residual plots to comment on the adequacy of the model.

URL for the dataset:

https://drive.google.com/file/d/1XkzjIz9JFUg20ynbQeTLG6h4nhogr8WN/view?usp=sharing

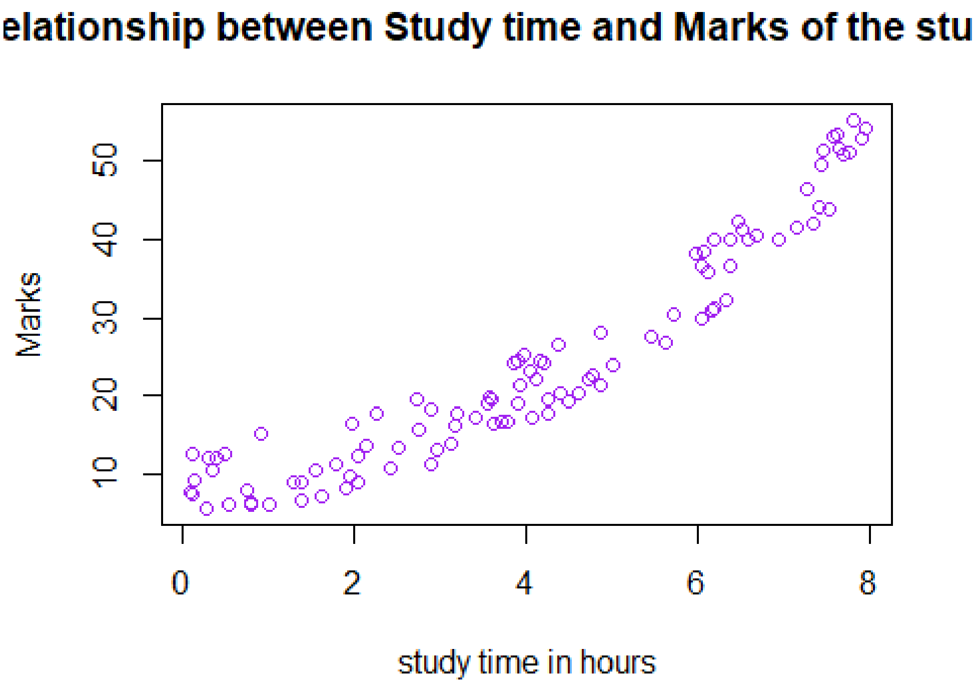
Procedure and Analysis

We import the dataset as follows:

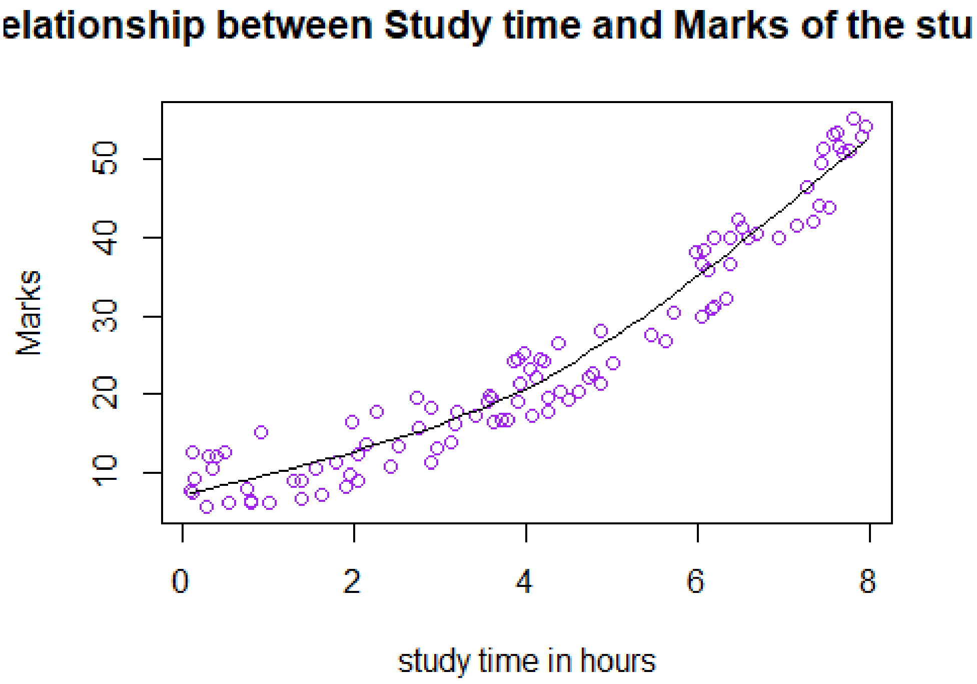
```
library(readr)
Student_Marks_dataset <- read_csv("G:/My Drive/Linear
Regression/Datasets/Student_Marks_dataset.csv")

## Rows: 100 Columns: 3
## ── Column specification
────────────────────────────────────────────────────────

## Delimiter: ","
## dbl (3): number_courses, time_study, Marks
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

View(Student_Marks_dataset)
```

We plot the variables [marks and study time]on the graph to get the idea about the relationship

```
plot(Student_Marks_dataset$time_study,Student_Marks_dataset$Marks,col="purple
",main="Relationship between Study time and Marks of the
students",xlab="study time in hours",ylab="Marks")
```



```
scatter.smooth(Student_Marks_dataset$time_study,Student_Marks_dataset$Marks,c
ol="purple",main="Relationship between Study time and Marks of the
students",xlab="study time in hours",ylab="Marks")
```



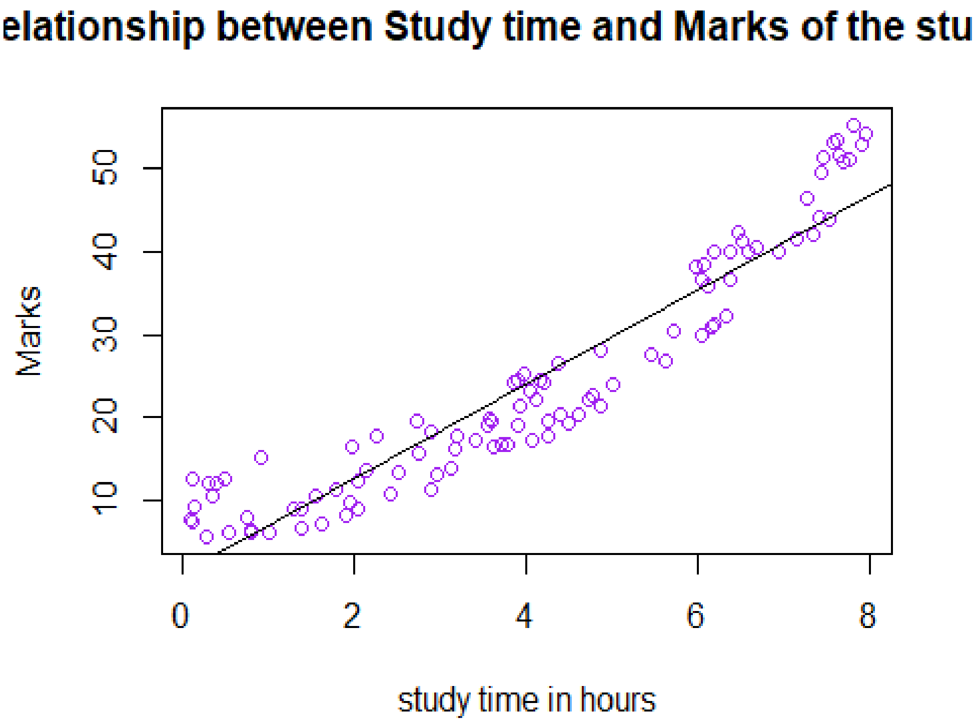
In order to get a better understanding between the variables, we obtain the correlation coefficient as follows:

```
cor(Student_Marks_dataset$time_study,Student_Marks_dataset$Marks)

## [1] 0.9422539
```

From the correlation coefficient, we say that the relationship between the study time and marks of students is 0.9422539 indicating a strong positive relationship.

Now we build our model as follows:

```
regmodel=lm(Student_Marks_dataset$Marks~Student_Marks_dataset$time_study)
regmodel

## 
## Call:
## lm(formula = Student_Marks_dataset$Marks ~ 
Student_Marks_dataset$time_study)
## 
## Coefficients:
##                      (Intercept)  Student_Marks_dataset$time_study  
##                            1.224                             5.689

summary(regmodel)

## 
## Call:
## lm(formula = Student_Marks_dataset$Marks ~ 
Student_Marks_dataset$time_study)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -7.866 -4.034 -0.384  2.979 10.628 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                        1.2239     0.9623   1.272    0.206    
## Student_Marks_dataset$time_study   5.6888     0.2042  27.853   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.822 on 98 degrees of freedom
## Multiple R-squared:  0.8878, Adjusted R-squared:  0.8867 
## F-statistic: 775.8 on 1 and 98 DF,  p-value: < 2.2e-16

plot(Student_Marks_dataset$time_study,Student_Marks_dataset$Marks,col="purple
",main="Relationship between Study time and Marks of the
students",xlab="study time in hours",ylab="Marks")

abline(regmodel, main="Relationship between Study time and Marks of the
students",xlab="study time in hours",ylab="Marks")
```



The model built is

$$Y = B_0 + B_1X + E$$

where,

Y is marks of the student which is an independent variable

X is the study time which is the dependent variable

B_0 and B_1 are the intercept and slope regression coefficients respectively

E is the error associated in the model

The estimated model after building is

$$Y^{hat} = 1.224 + 5.689X$$

We draw conclusions from the above fitted model as If study time(x) is zero say, then the average increase in marks is 1.224 If intercept is zero then, unit increase in study time will increase the marks by 5.689

To validate the model:

We see if the regression coefficients are significant. We check the same by looking at the p-values of the same. Here the p-value of study time is <2e-16 which is almost zero. So we say that the slope intercept B_1 is significant.

Whereas the intercept B_0 is not significant as the p-value [0.206] is greater than the significant level. So, the intercept B_0 is not significant.

After checking for the coefficient's significance, we see for the overall model as follows:

We see for the R^2 value. For our model, the R^2 values is 0.8878, which says that 88.78% of the data is explained by the study time for the marks of the student.

Now to check if the assumptions are satisfied by the residuals

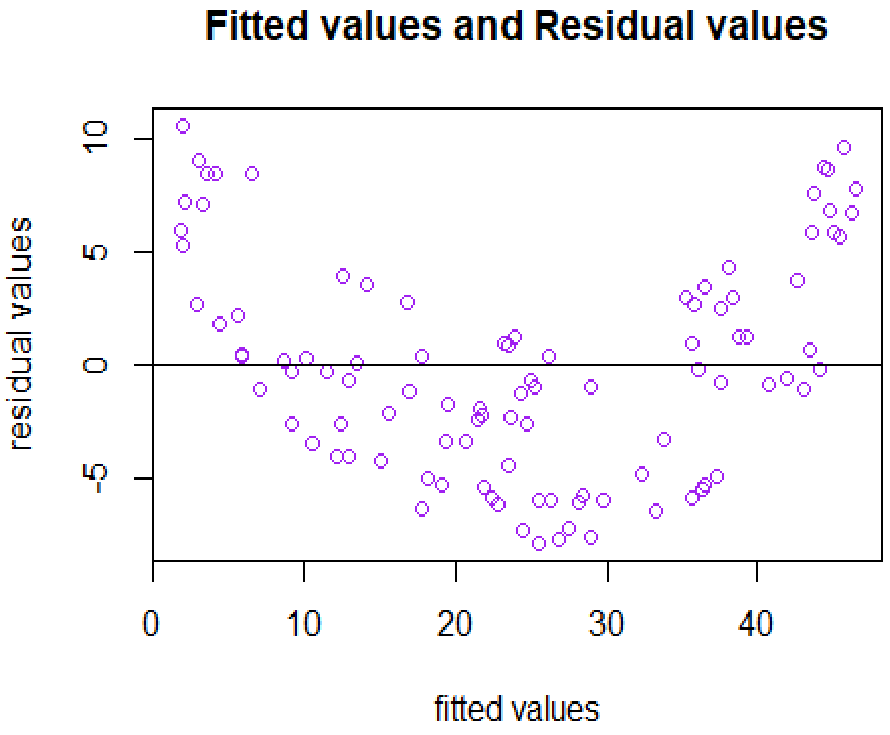
Residual analysis

```
fit=fitted.values(regmodel)
fit

##         1         2         3         4         5         6         7 
8 
## 26.868745  1.769978 19.046713 46.216185 45.658688 19.490436 35.714752 
20.639563 
##         9        10        11        12        13        14        15 
16 
## 26.311247 36.340514 43.053240  3.630200 25.219007 25.537577 17.766744 
25.457935 
##        17        18        19        20        21        22        23 
24 
## 33.757822 35.811461 45.089813 23.848019 28.148714 36.073143 12.891485 
28.956516 
##        25        26        27        28        29        30        31 
32 
## 21.902466  9.227930  4.113743 26.129207  2.111303  8.613545 23.205190 
12.163325 
##        33        34        35        36        37        38        39 
40 
##  6.525774 38.735478 24.451026 44.134103 18.096692 42.655027 38.388465 
45.453893 
##        41        42        43        44        45        46        47 
48 
##  2.020283 16.890677 21.652161 10.081243 12.339676 12.948373 22.824043 
28.410396 
##        49        50        51        52        53        54        55 
56 
## 33.279967 23.483939 39.355552 24.718397  5.609885 35.635109 44.407163 
17.795188 
##        57        58        59        60        61        62        63 
64 
## 44.691600 44.737110 36.482733 43.707446  3.362828 15.093032 21.737492 
28.922384 
##        65        66        67        68        69        70        71 
72 
##  1.963396 13.409162 32.358389  4.352671  9.159665 23.683045 22.477030 
15.548132
```

```
##        73        74        75        76        77        78        79
80
## 27.579839 10.490833 40.783429  5.791925 37.512397 35.271029 43.610737
5.803302
##        81        82        83        84        85        86        87
88
## 46.489245 14.091812 43.377499 19.410793 12.498961 36.499799 24.360006
7.100337
##        89        90        91        92        93        94        95
96
## 11.480675 37.495331 25.014212 16.754147 29.821206 38.035762 23.518071
21.481498
##        97        98        99       100
##  2.936172 41.972377  2.981682 37.262092
```

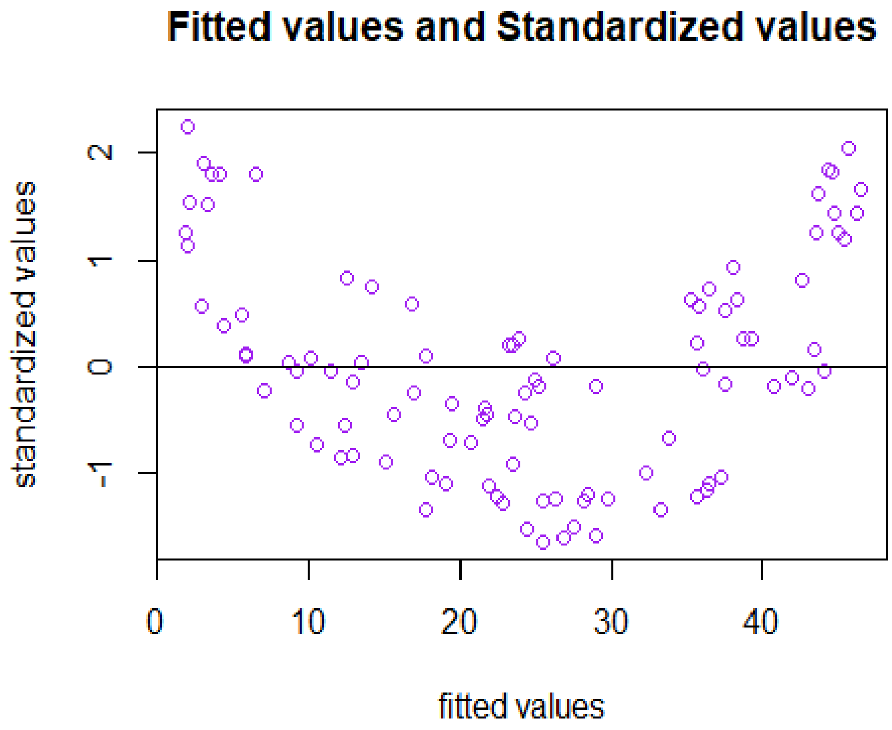
We plot the fitted values and residual values of the model as:

```
plot(fit,resid(regmodel),col="purple",main="Fitted values and Residual
values",xlab="fitted values",ylab="residual values")
abline(0,0)
```



To check if there are any outliers and scale out the residual values we use standardized residuals.

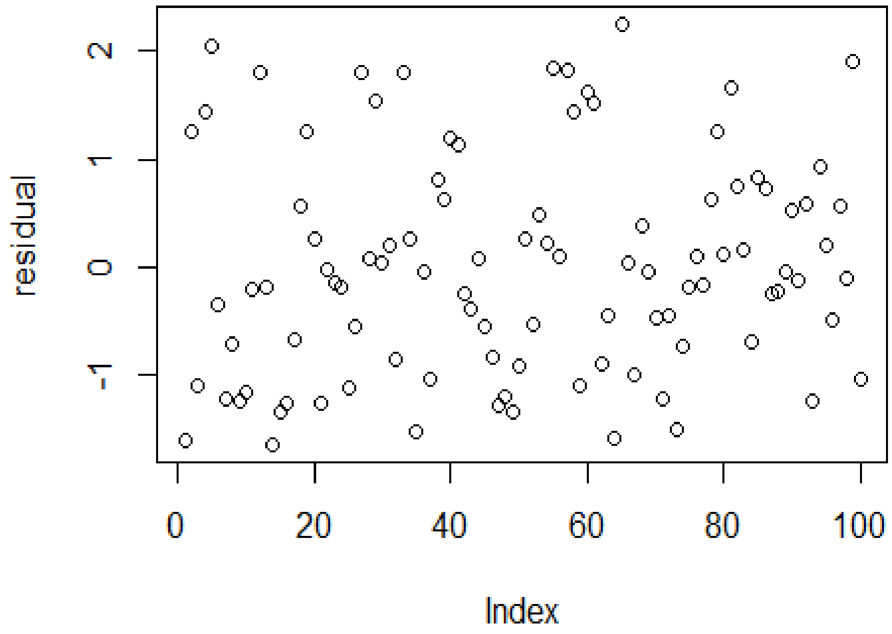
```
plot(fit,rstandard(regmodel),col="purple",main="Fitted values and
Standardized values",xlab="fitted values",ylab="standardized values")
abline(0,0)
```



From the above plot we say that no values lie outside the range of modulus 3. So there are no outliers in the dataset.

To be very precise we can even check in the residual values for outliers as follows:

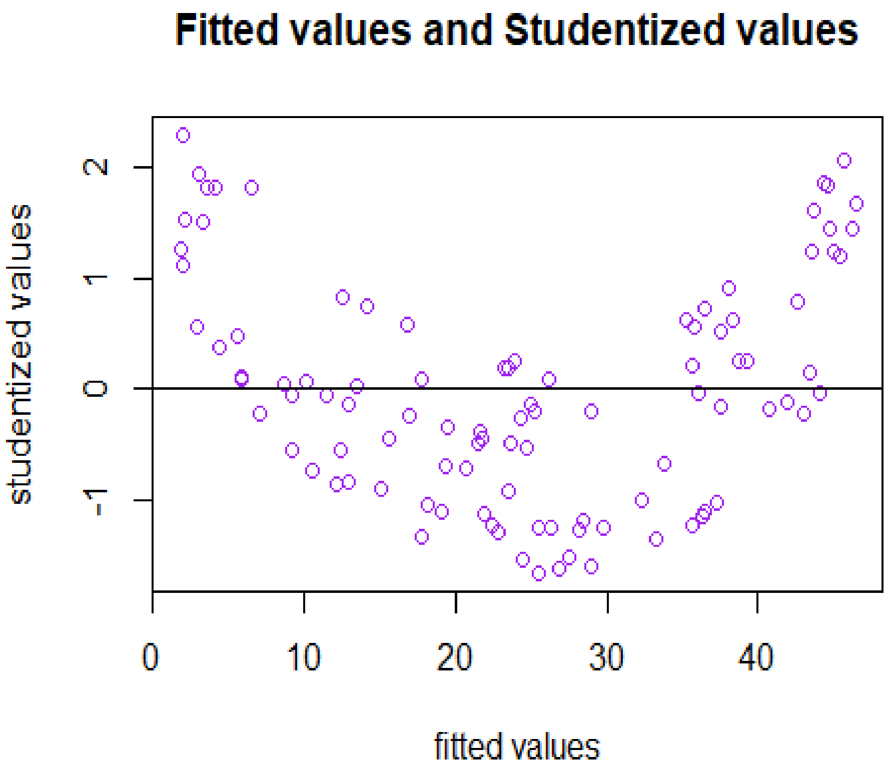
```
residual=rstandard(regmodel)
residual

##           1           2           3           4           5           6 
## -1.59814650  1.26124499 -1.09209241  1.43685536  2.03507232 -0.34796645 
##           7           8           9          10          11          12 
## -1.21854565 -0.70380507 -1.24296513 -1.14638538 -0.21410165  1.79375047 
##          13          14          15          16          17          18 
## -0.18778836 -1.63937523 -1.32920822 -1.24885712 -0.67062035  0.56029238 
##          19          20          21          22          23          24 
##  1.24383510  0.26781403 -1.26677401 -0.02806462 -0.14277334 -0.19050198 
##          25          26          27          28          29          30 
## -1.12261987 -0.54645222  1.79936178  0.08395558  1.52653754  0.04690089 
##          31          32          33          34          35          36 
##  0.20150770 -0.85044782  1.79020691  0.25773631 -1.51727904 -0.03289434 
##          37          38          39          40          41          42 
## -1.03859652  0.79903618  0.62231066  1.20046238  1.12377915 -0.24333277
##          43          44          45          46          47          48
## -0.39214931  0.07289916 -0.54362389 -0.84185257 -1.27581838 -1.19046497
##          49          50          51          52          53          54
## -1.33638483 -0.91245776  0.26141878 -0.52821206  0.48041238  0.21289671
##          55          56          57          58          59          60
##  1.84457217  0.09240292  1.82758401  1.44358187 -1.09799061  1.60821838
##          61          62          63          64          65          66
##  1.51096130 -0.88773276 -0.44766326 -1.57306109  2.24691390  0.03196272
##          67          68          69          70          71          72
## -0.99995546  0.38626661 -0.05027896 -0.47583184 -1.22374966 -0.44535379
##          73          74          75          76          77          78
## -1.49723433 -0.72859712 -0.17459775  0.08946599 -0.16050230  0.62877993
##          79          80          81          82          83          84
##  1.24954482  0.11485213  1.65499625  0.75530683  0.15190950 -0.68925801
##          85          86          87          88          89          90
##  0.82905879  0.72350122 -0.25239391 -0.22013866 -0.04767521  0.52955855
##          91          92          93          94          95          96
## -0.12926390  0.58658561 -1.23175093  0.91978687  0.19444239 -0.49062679
##          97          98          99         100
##  0.56440126 -0.11108523  1.90992547 -1.02705787

plot(residual)
```



Now we have numerically seen that there are no outliers in the dataset. The maximum value is 2.24691390

We can also try for studentized residual

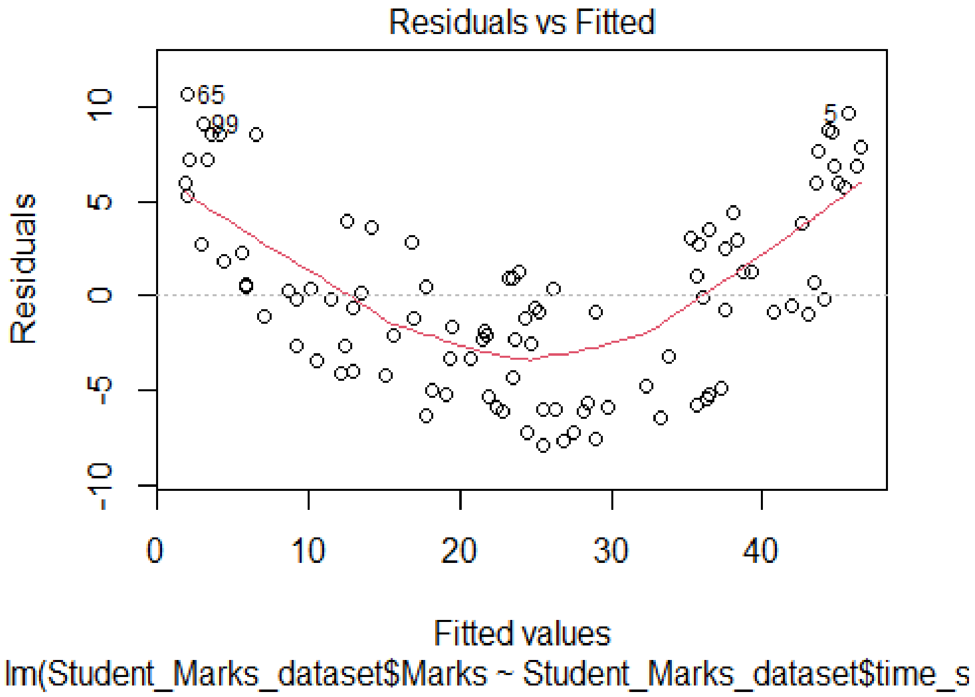
```
plot(fit,rstudent(regmodel),col="purple",main="Fitted values and Studentized
values",xlab="fitted values",ylab="studentized values")
abline(0,0)
```

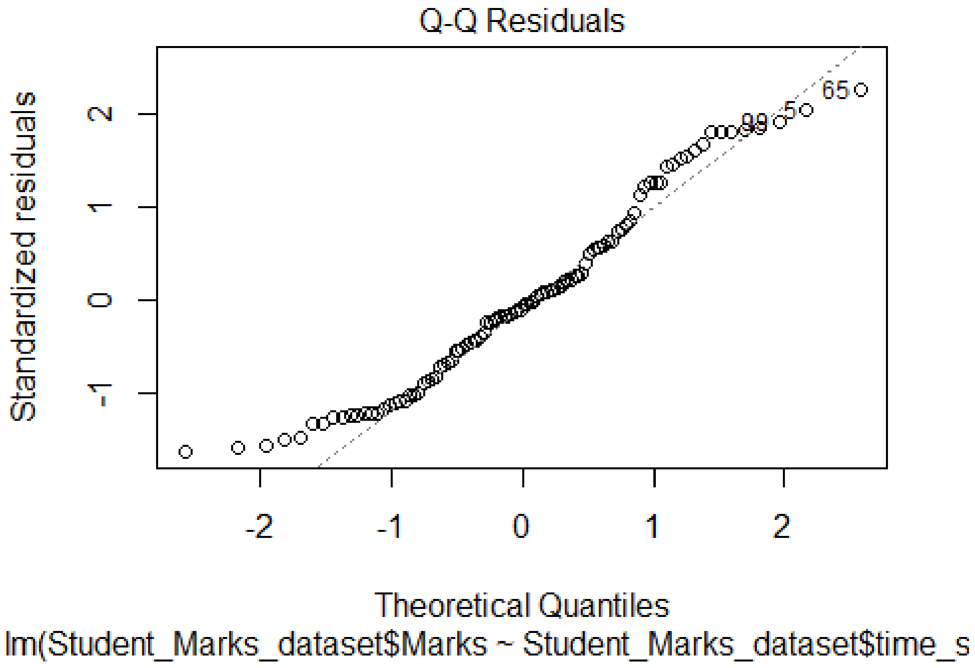


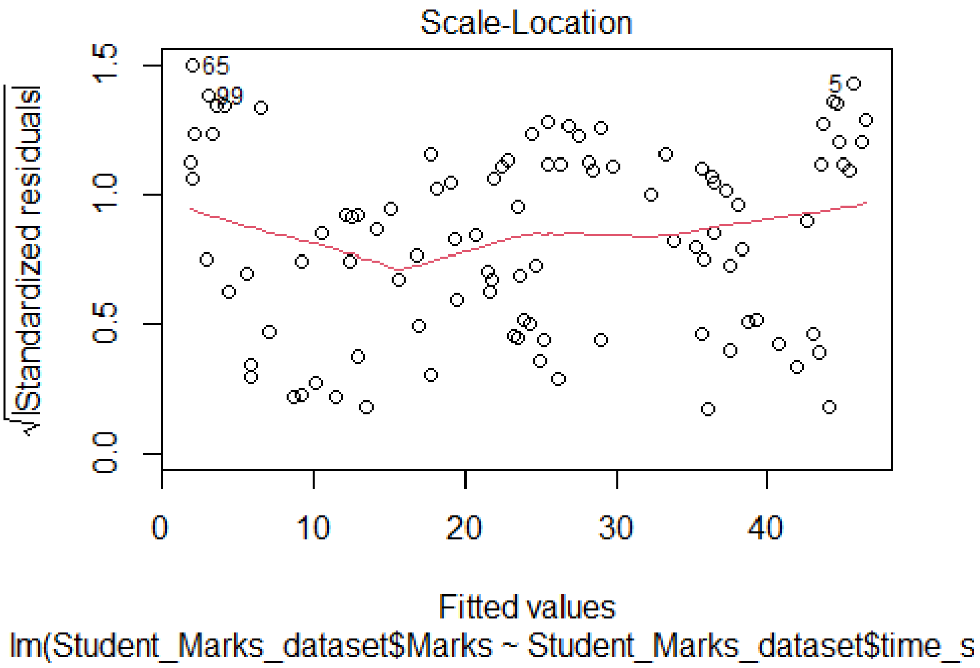
The plot resembles a u-shaped upward open funnel. The plot throws light on the linearity of the variables. Here we say that X and Y variables are not linear. This graph does not infer anything about the variance of the error term.

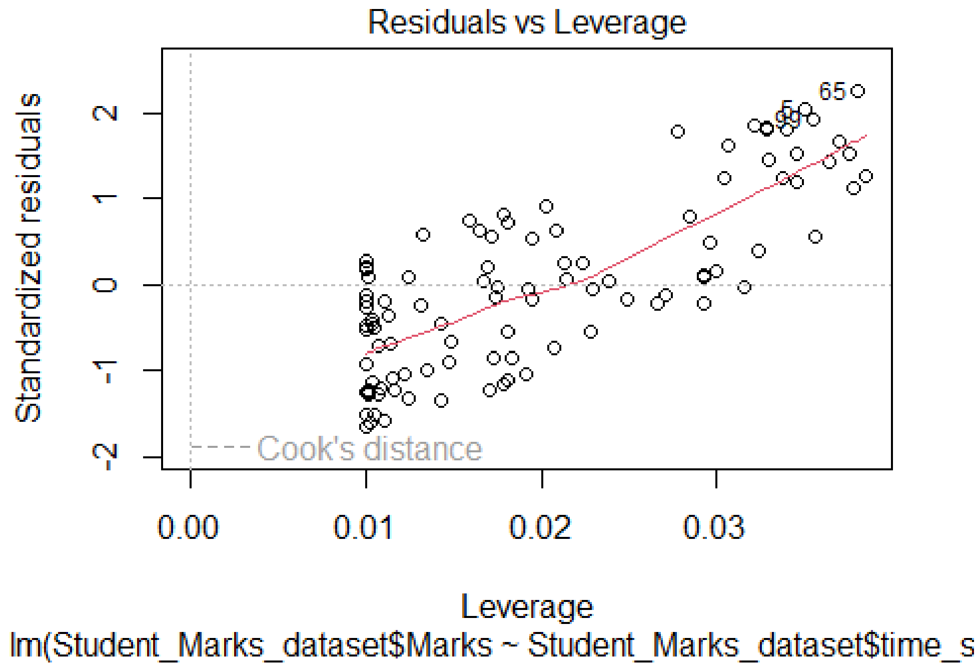
All the plots of the model can be obtained:

```
plot(regmodel)
```


Residuals vs Fitted
Residuals
10
5
0
-5
-10
65
99
5
0
10
20
30
40
Fitted values
lm(Student_Marks_dataset$Marks ~ Student_Marks_dataset$time_s


Q-Q Residuals
Standardized residuals
2
1
0
-1
65
99
5
-2
-1
0
1
2
Theoretical Quantiles
lm(Student_Marks_dataset$Marks ~ Student_Marks_dataset$time_s





Here the inference is drawn only from the first graph.

Conclusion

The dataset considered here was the marks of the students and the study time invested in hours. We checked for the correlation for the variables. We built the model and got the estimated model as

$$Y^{hat} = 1.224 + 5.689X$$

We draw conclusions from the above fitted model as If study time(x) is zero say, then the average increase in marks is 1.224 If intercept is zero then, unit increase in study time will increase the marks by 5.689

After the model building, we started the residual analysis by getting the fitted values of the model.Then we plotted the same with the residual values. We used standardized and studentized residuals in order to check for the outliers in the dataset. Both the residuals gave the same results and there were no outliers in the dataset. To check the numerical values, we got the residual values and found that the maximum value is 2.24691390

After that we plotted the residual graph and inferred, the plot takes a u-shaped upward opening funnel. The plot throws light on the linearity of the variables. Here we say that X[study time] and Y[Marks] variables are not linear. This graph does not infer anything about the variance of the error term.

So we say that, for the dataset considered, the variables are non-linear and have no outliers present. The variance of the error term cannot be inferred.