RESIDUAL ANALYSIS

A.Mayuri(2348133)

LAB4- RESIDUAL ANALYSIS

A.Mayuri(2348133)

2023-11-24

Problem:

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.

Introduction:

Here we are interested in doing a residual analysis for the dependent variable(miles per gallon) which measures the eeficiency of the vehicle vs the independent variable(x) as engine displacement.

About the Variables:

1) data set on the gasoline mileage performance of 32 different automobiles,
2) Dependent Variable(y) is miles per gallon. ie, the efficiency of the automobile.
3) independent variable (X) is engine displacement.

Procedure:

Step1: Import Data set

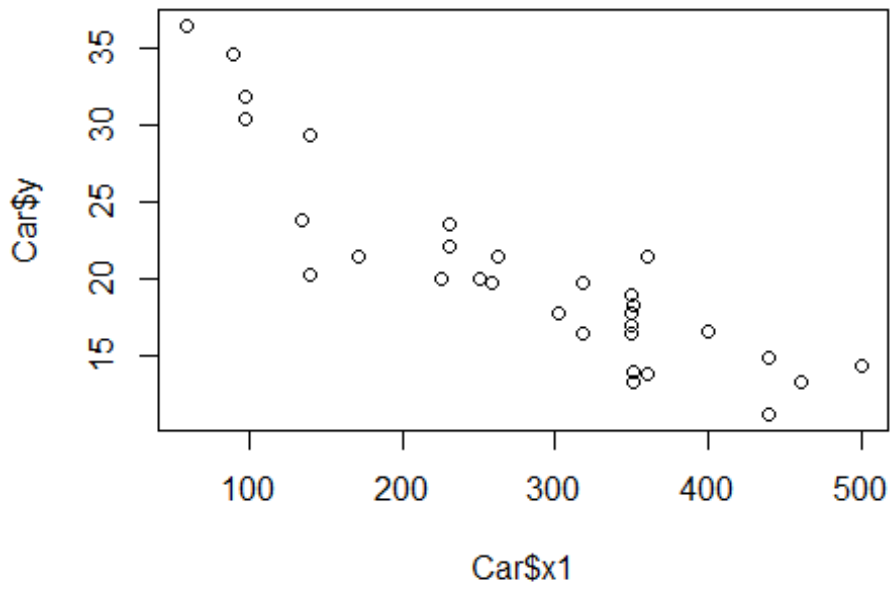
```
library(readxl)
Car <- read_excel("C:/Users/mayur/Desktop/Mstat/Semesters/Tri-sem2/Regression/Dataset/Car.xlsx")
View(Car)
attach(Car)
```

Step2: Preliminary Understanding of the Model

1) Visualization and Correlation

Let us plot the graph between the dependent and independent variable to visualize their relation also their correlation to understand the strength of the relation.

```
#Plotting the dataset
plot(Car$x1,Car$y)
```




```
# Finding the correlation
cor(Car$x1,Car$y)

## [1] -0.885896
```

Interpretation: we get a strong negative linear relationship between the independent and dependent variables. with a karl pearson correlation coefficient of -0.885896

2): Fit the model

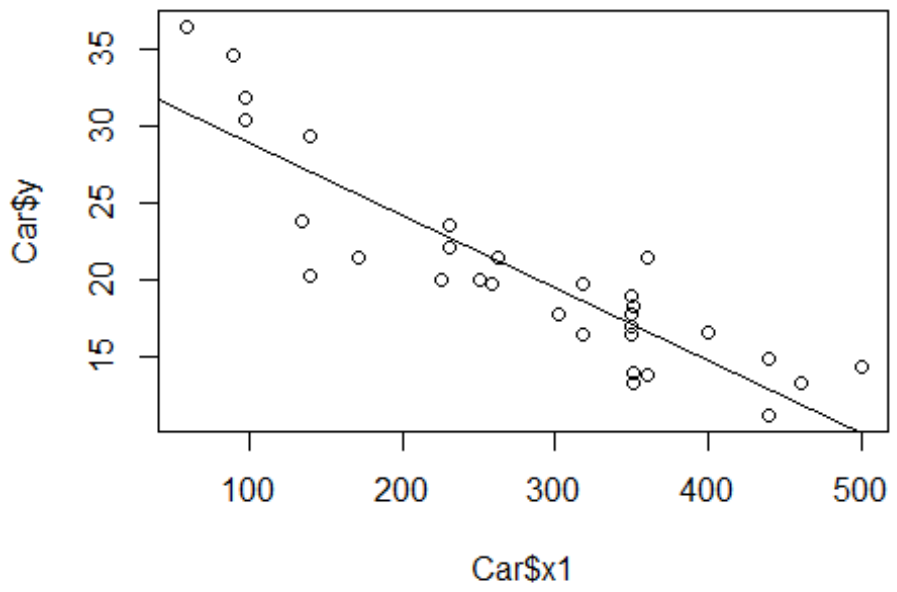
```
reg=lm(Car$y~Car$x1)
reg #fitting a sinple regression model
```

```
##
## Call:
## lm(formula = Car$y ~ Car$x1)
##
## Coefficients:
## (Intercept)      Car$x1
##    33.61922     -0.04719
```

we got the value of beta(0)=33.61622 and value of beta(1)=-0.04719 based on sample. since sign of beta 1 is negative there is a negative linear relationship between the variables. for each additional 1 unit of engine displacement the efficiency of the vehicle decreases by 0.04719.

3) Fitting the regression line

```
plot(Car$x1,Car$y)
abline(reg)
```



Hence the regression equation is given by y= 33.61922 -0.04719*X +E where Y is the depenent variable x-independent variable and E is the residual.

4) Overall Fit

the overall fit of the regression model can be obtained with summary command. which is required to devise the hypothesis and goodness of fit

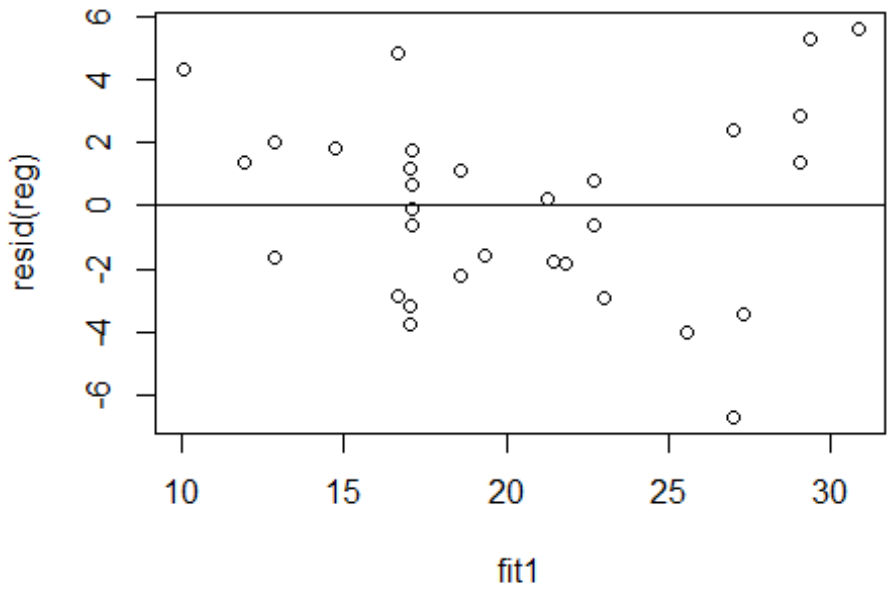
```
summary(reg)

##
## Call:
## lm(formula = Car$y ~ Car$x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7129 -1.9175  0.0553  1.8089  5.6318
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.619220   1.384763   24.28  < 2e-16 ***
## Car$x1      -0.047188   0.004511  -10.46 1.59e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.979 on 30 degrees of freedom
## Multiple R-squared:  0.7848, Adjusted R-squared:  0.7776
## F-statistic: 109.4 on 1 and 30 DF,  p-value: 1.589e-11
```

Hypothesis Testing::

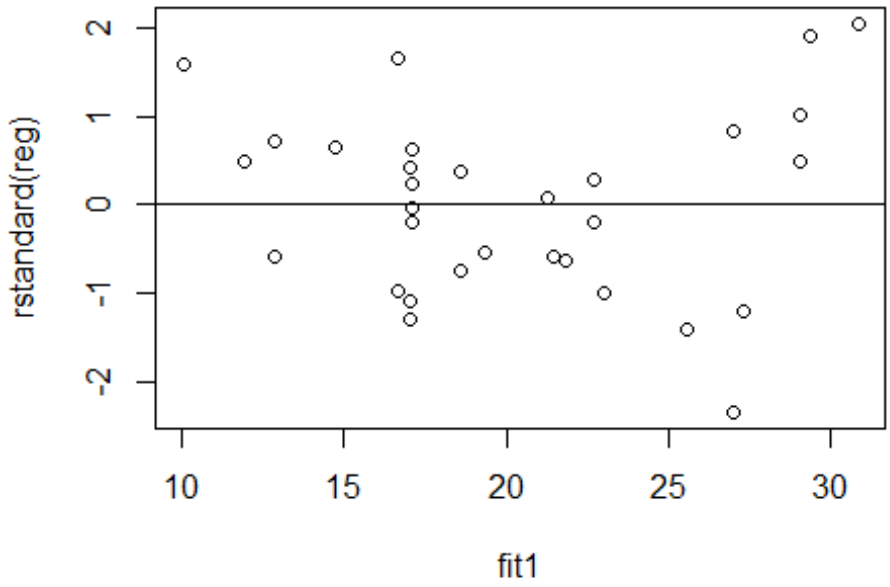
H0= Beta(1)=0 H1= Beta(1)=! 0 from the above table the p value in the line of units is 1.59e which is less than 0.05 level of significance.hence we reject null hypothesis and hence beta (1)=! 0 . thus these exist a significant linear relationship between the variables., ie there is a significant linear relationship between efficiency and misplacement of engine.

Residual Analysis

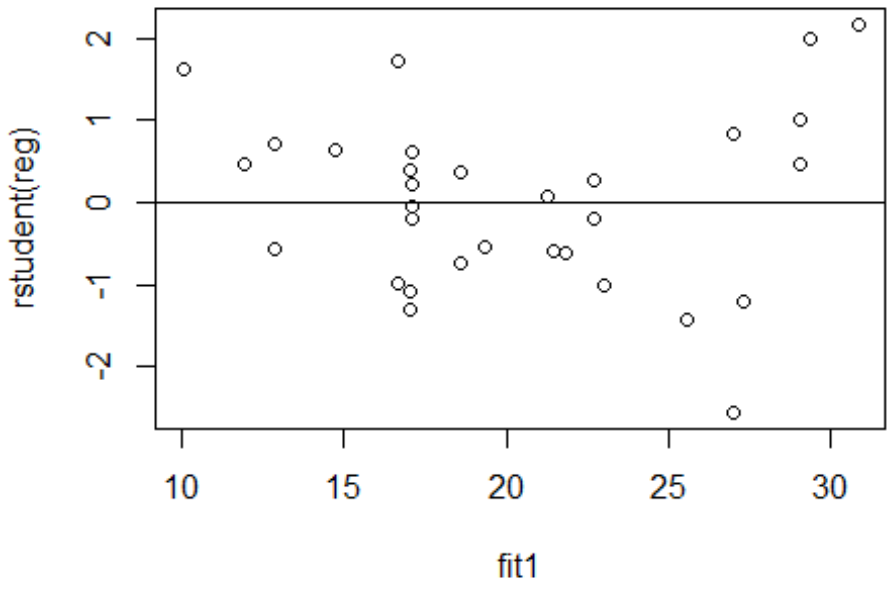
```
fit1=fitted.values(reg)
fit1

##       1       2       3       4       5       6       7       8
## 17.10341 17.10341 21.82221 17.05622 23.00191 12.85648 22.71878 21.25595
##       9      10      11      12      13      14      15      16
## 29.38645 29.04670 17.10341 30.86816 25.55006 21.44471 27.01289 19.36843
##      17      18      19      20      21      22      23      24
## 10.02520 12.85648 17.10341 18.61342 22.71878 16.63153 14.74400 29.04670
##      25      26      27      28      29      30      31      32
## 27.01289 11.91272 27.31490 18.61342 17.05622 17.05622 16.63153 17.10341

# Residual analysis
plot(fit1,resid(reg))
abline(0,0)
```



```
plot(fit1,rstandard(reg))#standardized residual model
abline(0,0)
```




```
plot(fit1,rstudent(reg))# studentized residual model
abline(0,0)
```

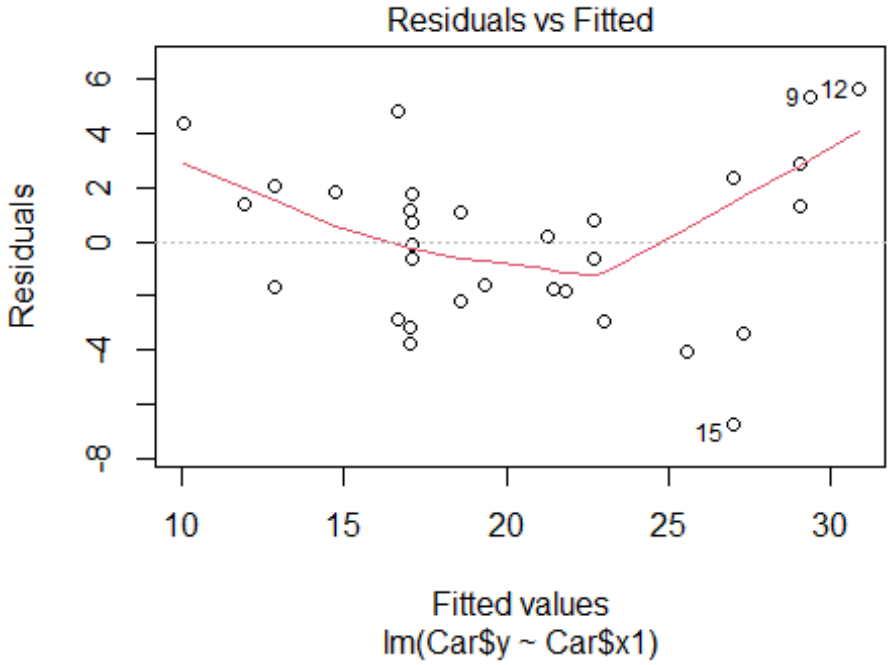


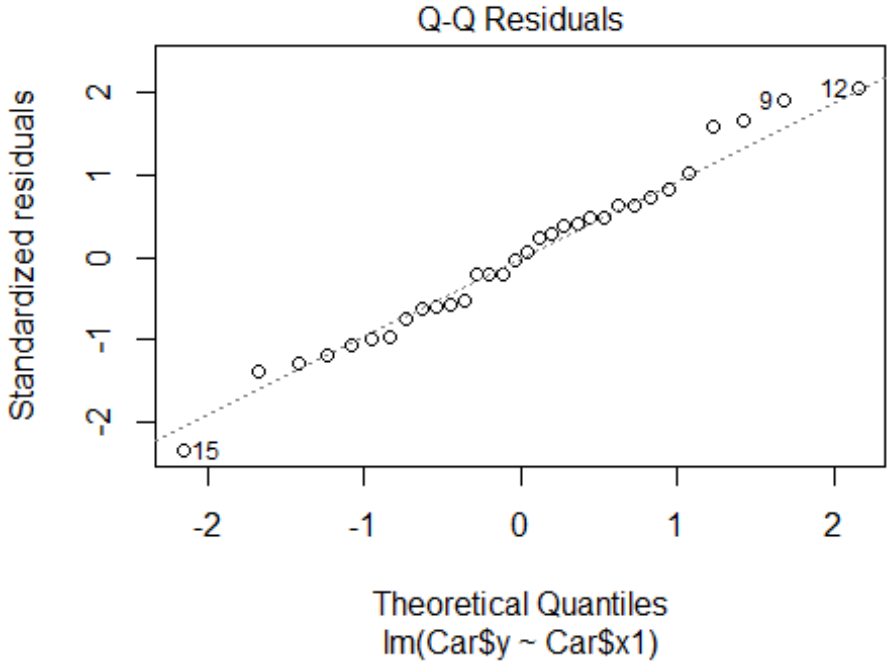
Using standard and studentized residual plot we observe that the residuals are within the bandwidth and hence the variables are linear and have a constant variance.

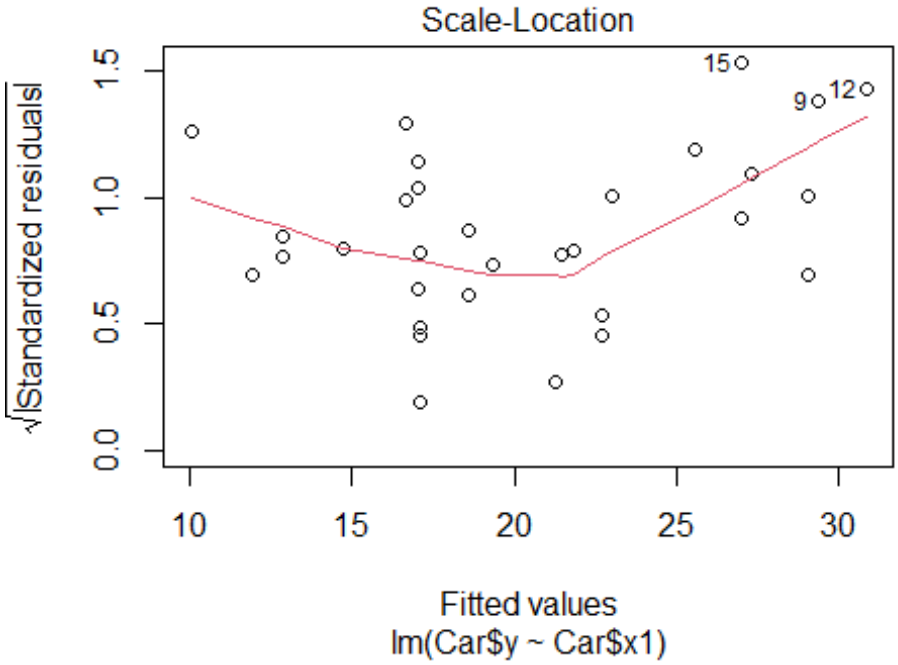
Normality

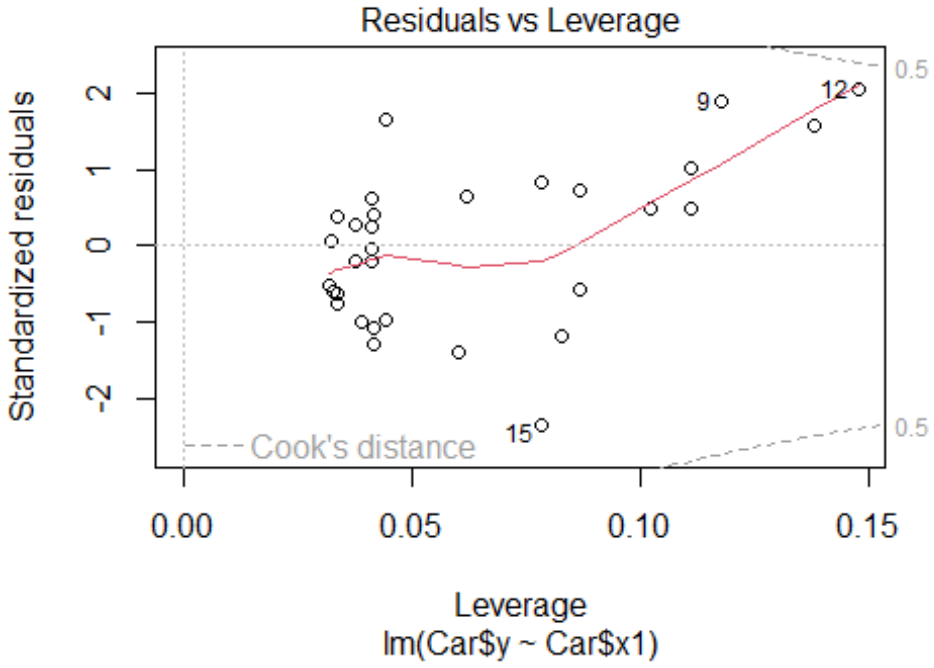
```
# to check for res vales (Normality check)
re1=rstandard(reg)
re1

##          1          2          3          4          5          6
##  0.61584738 -0.03544603 -0.62223509  0.40927708 -1.00393446 -0.58190488
##          7          8          9         10         11         12
## -0.20486911  0.07303318  1.89866445  0.48185358 -0.20683903  2.04777383
##         13         14         15         16         17         18
## -1.40241617 -0.59543172 -2.34738384 -0.53505625  1.57820808  0.71435363
##         19         20         21         22         23         24
##  0.23878278 -0.75242272  0.28097366  1.66138909  0.63978999  1.01594036
##         25         26         27         28         29         30
```

```
##  0.83472995  0.48082660 -1.19693832  0.38128762 -1.08207963 -1.29806922
##          31          32
## -0.98256314 -0.20683903

plot(reg)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Car$y ~ Car$x1)
9
12
15


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Car$y ~ Car$x1)
9
12
15






```
shapiro.test(re1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  re1
## W = 0.98283, p-value = 0.8763
```

Interpretation from the Graphs:

1 plot-residual vs fitted: flagged the possible outline as 9,12,15. assumption of linearity and constant variance.

2 plot- assumption of normality: the observation lie along the line thus the error terms follow a normal distribution.

the plot 3 shows the inverted u curve for std residual vs fitted values however it lies within the band range (but we observe an U shape). hence it may or may not have a constant variance.

However conducting specific hypothesis testing for each of these assumption would be a better and accurate form of conclusion. Further suggested tests are: Bp test -constant variance kolmogrov test for normality.

Kolmogrov Test:

Hypothesis testing for constant variance :

test for errors have constant variance through errors. use the fitted model.

H0: error have const variance

H1: error have not const variance

```
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

bptest(reg)

##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 4.9622, df = 1, p-value = 0.02591
```

the p value is 0.02591< 0.05 level of significance thus we accept null hypothesis and conclude that there is a constant variance for the errors.