Residual Analysis

ANGEL LAL

INTRODUCTION

The objective of the project is to do the residual analysis on the available data. Residual analysis points in particular to 1. Plotting appropriate residual plot to check the assumption of Normality, homoscedasticity and comment on it. Also, testing for homoscedasticity using a suitable statistical test and checking whether the residuals are normally distributed 2. Finding out the remedial measures to be taken if the constant variance assumption is violated. 3. Finding out whether the residuals are uncorrelated using a suitable test procedure. 4. Including an unusual y observation to the data set and examining whether it is an influential point.

PROCEDURE AND ANALYSIS

1. Accessing the data set

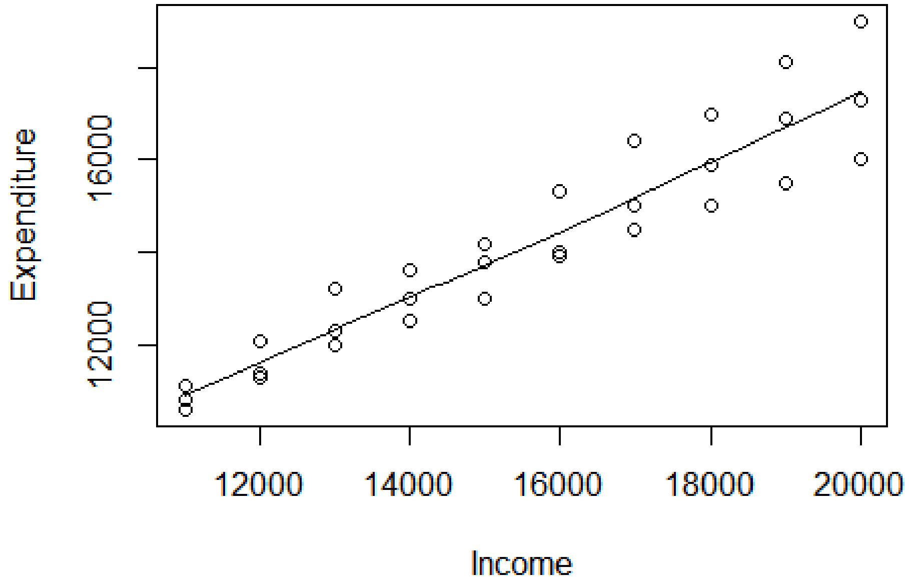
```
library(readxl)

## Warning: package 'readxl' was built under R version 4.1.3

data = read_xlsx("D:/MSTAT/SEM2/REGRESSION/LAB/Lab 6 data.xlsx")
head(data)

## # A tibble: 6 x 3
##    S.No. Expenditure Income
##    <dbl>       <dbl>  <dbl>
## 1      1       10600  11000
## 2      2       11400  12000
## 3      3       12300  13000
## 4      4       13000  14000
## 5      5       13800  15000
## 6      6       13900  16000

attach(data)
```

Interpretation

The loaded data set shows the annual consumption and disposable income for 30 households in India.

2. Plotting the matrix of scatter diagram and finding the matrix of coefficient of correlation

```
scatter.smooth(Expenditure~Income)
```



##

INTERPRETATION The scatter plot between Expenditure and Income is obtained. A linear relationship between Expenditure and Income variables can be observed from the plot. Also we can identify the dependent and independent variables. Y = Expenditure X = Income

3. Simple Linear Regression model using Expenditure and Income variables

```
model = lm(Expenditure~Income)
summary(model)

## 
## Call:
## lm(formula = Expenditure ~ Income)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1453.94  -473.48   -73.94   483.33  1546.06 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.799e+03  7.564e+02   3.701 0.000931 ***
## Income      7.327e-01  4.798e-02  15.271 4.16e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 754.8 on 28 degrees of freedom
## Multiple R-squared:  0.8928, Adjusted R-squared:  0.889
## F-statistic: 233.2 on 1 and 28 DF,  p-value: 4.165e-15
```

INTERPRETATION

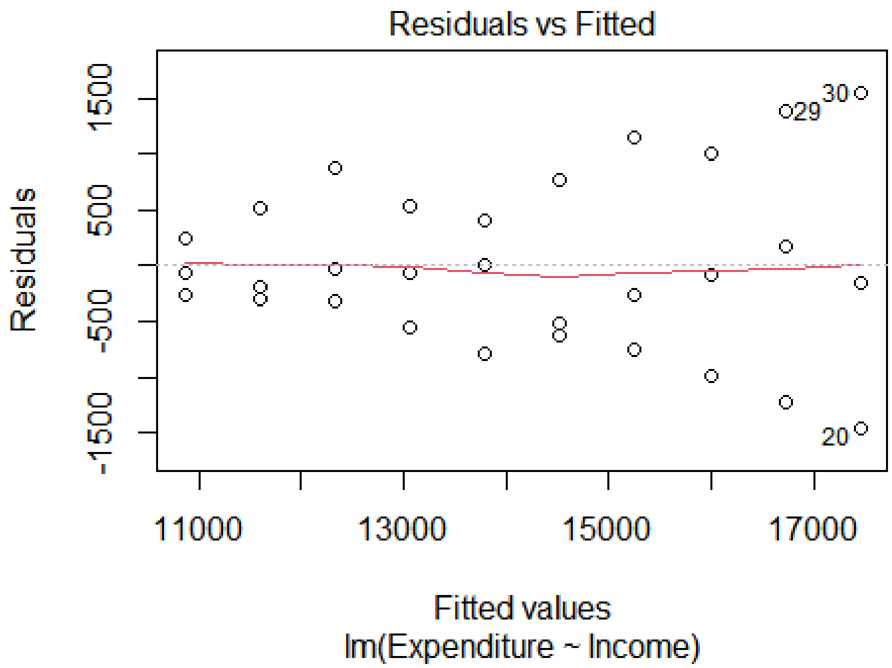
The simple linear regression model is obtained using the function lm() and the model obtained is:

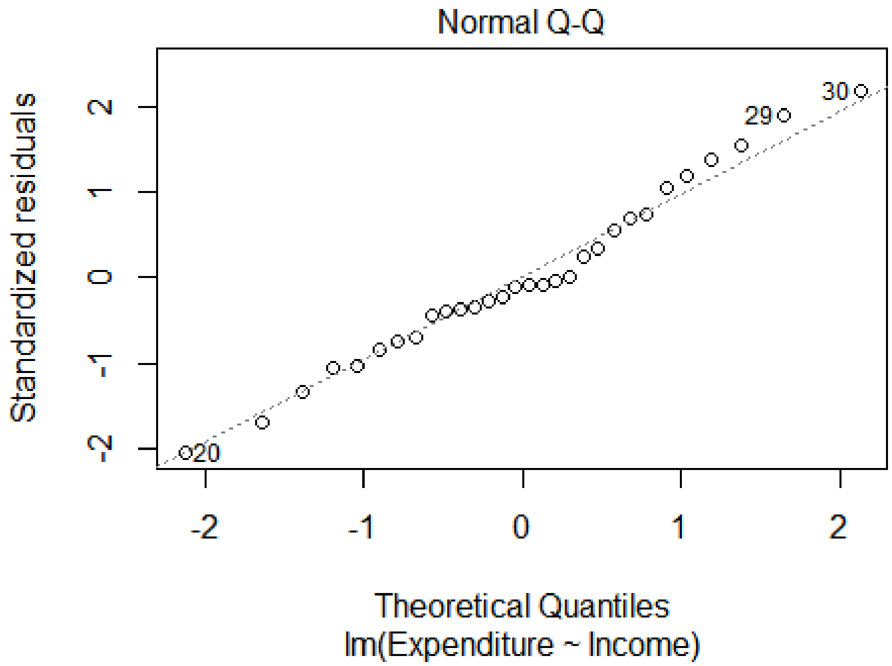
Expenditure_hat = 2.799e^(03) + 7.327e^-(01) * Income

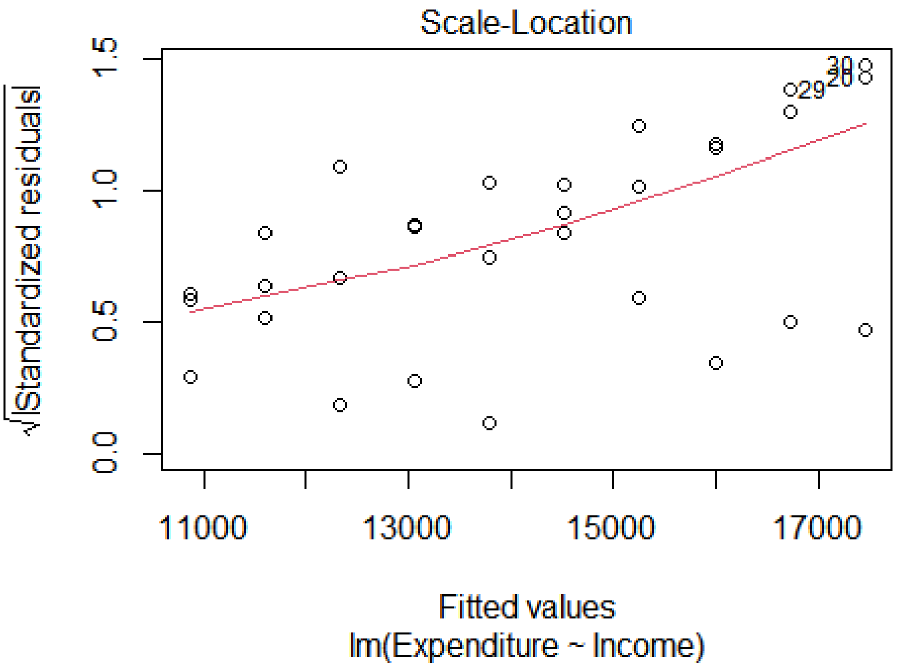
One unit change in Income will result in 7.327e^-01 unit change in expenditure. R^2 value obtained is 0.89 and hence we can conclude that 89% of variability in expenditure is explained by the variable income.

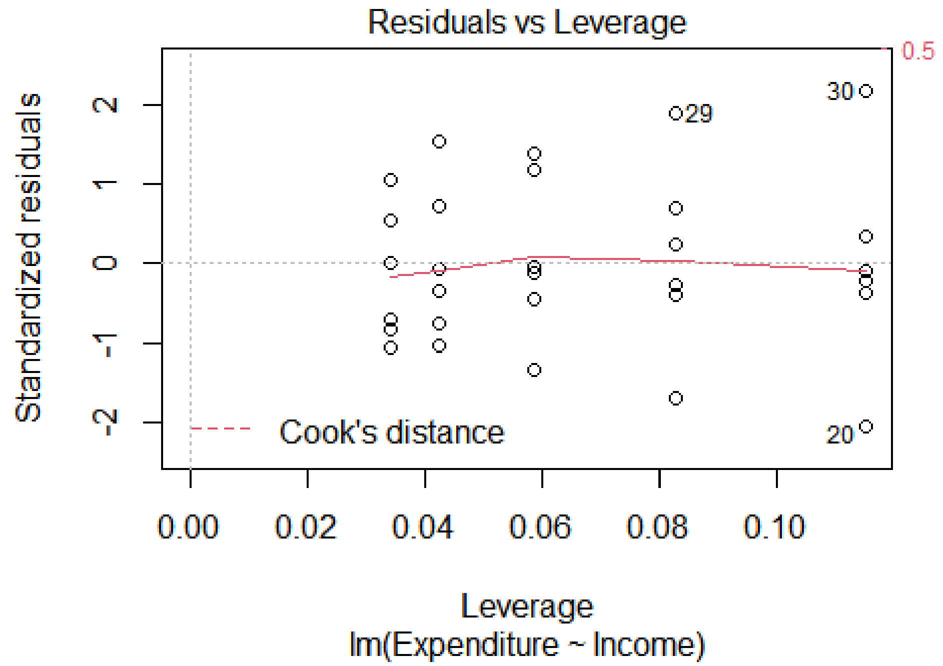
##4. Normality and Homoscedasticity of the model

```
plot(model)
```


Residuals vs Fitted
Residuals
1500
500
-500
-1500
11000
13000
15000
17000
Fitted values
lm(Expenditure ~ Income)
29
30
20


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-2
-1
0
1
2
Theoretical Quantiles
lm(Expenditure ~ Income)
30
29
20





## Interpretation

Scatter plots of residual vs fitted, Noramal Q-Q plot, Scale location plot and residual vs leverage graphs are obtrained using the plot command. From the plots of residual vs fitted we can observe a funnel shape in the graph which tells that the constant variance assumption is violated.i.e residual is the increasing sequence in Y pointing towards a defective model Also 20,29 and 30th observations are marked and hence this could be an outlier. We should be suspicious about these observations being outliers. From the Normal Q-Q plot it is observed that the residulas follows a normal distribution. From the leverage vs residual graoh also it is obtained that teh spread of standardised residuals shoudnt change as a function of leverage. But here it appears to change indicating hetroscedasticity. This means that the model adequacy is not met.

5. Homoscedasticity using Breusch - Pagan test

H0 : The residual values are having constant variance. (Homoscedastic)
H1 : The residual values are not having constant variance (Hetroscedastic)

```
library(lmtest)

## Warning: package 'lmtest' was built under R version 4.1.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.1.3

##
## Attaching package: 'zoo'

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

bptest(model)

##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 11.373, df = 1, p-value = 0.000745
```

INTERPRETATION

Bp-test value is obtained. Here pvalue = 0.000745 which is <0.05 and hence we are rejecting the null hypothesis. Hence we can tell that the residuals are not having constant variance. They are hetroscedastic.

6. Shapiro-wilk test for normality

H0 : Residual are coming from a normal distribution
H1 : Residuals does not follow a normal distribution

```
rstd = rstandard(model)

shapiro.test(rstd)

##
##  Shapiro-Wilk normality test
```

```
##
## data:  rstd
## W = 0.98167, p-value = 0.868
```

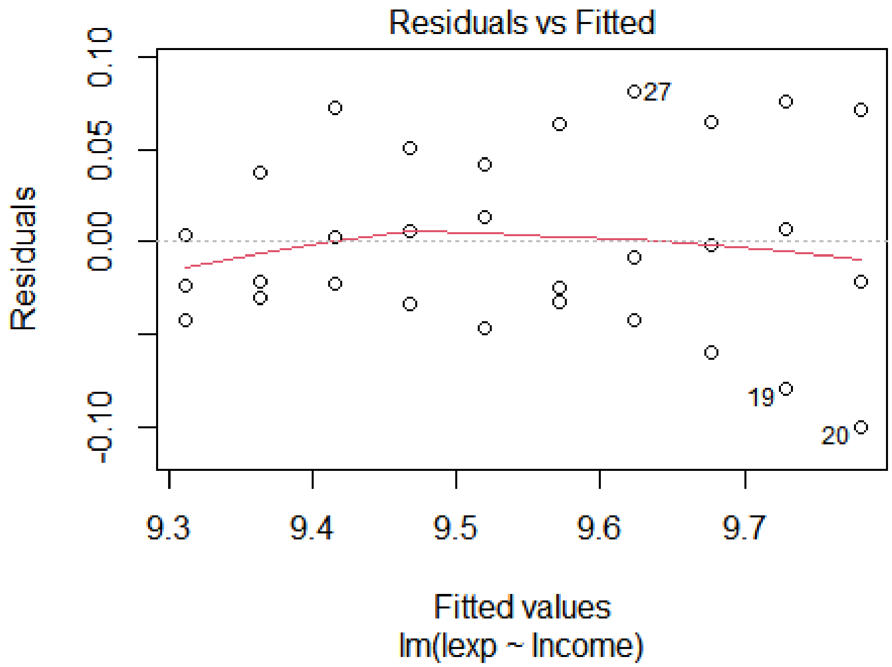
INTERPRETATION

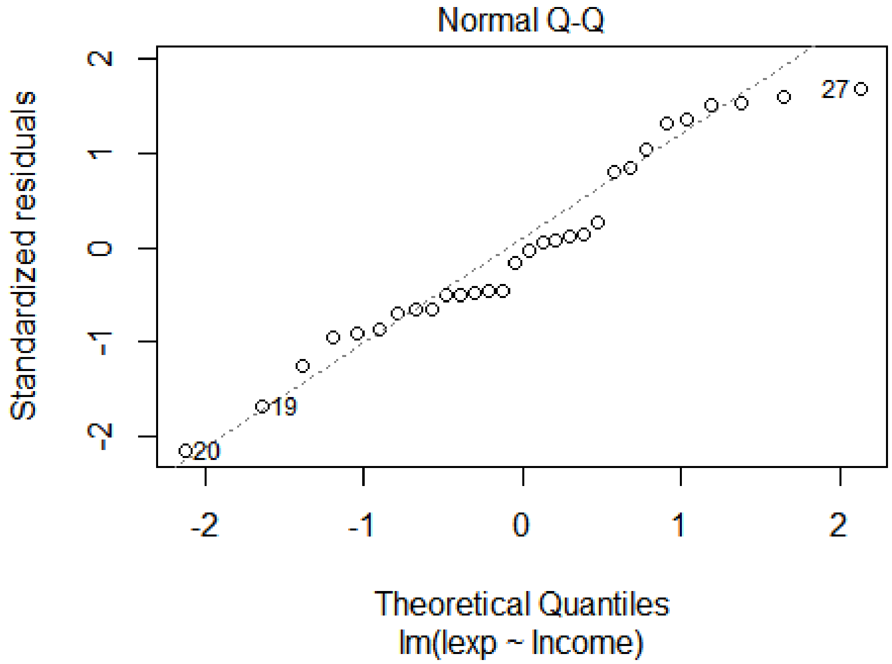
shapiro wilk test is run for the residuals. Here p-value = 0.868 =>0.05. Hence there is no evidence for rejecting the null hypothesis. So we can conclude that the residuals follow a normal distribution

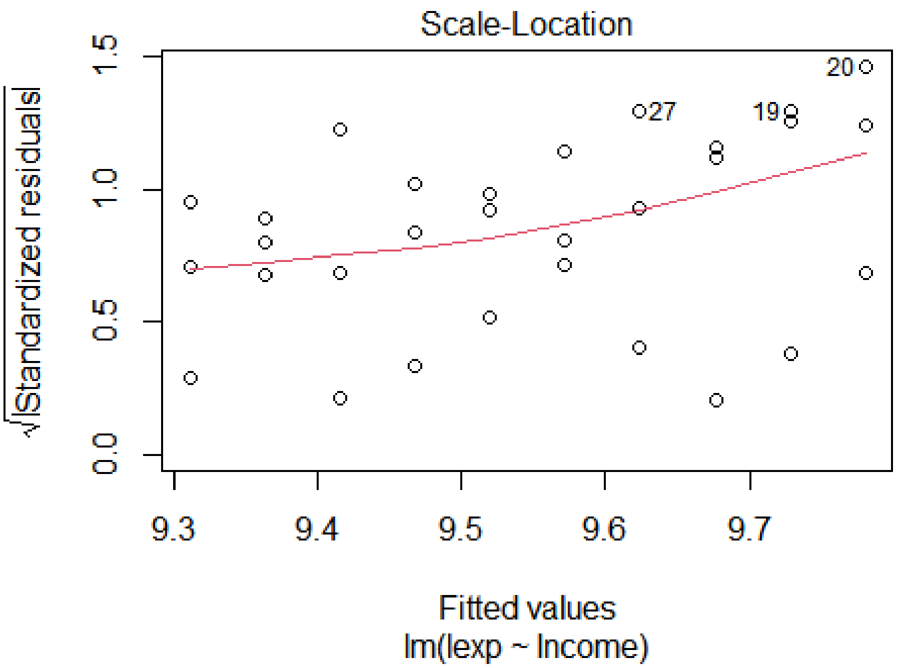
6.Transformation of residual since constant variance assupmtion is violated

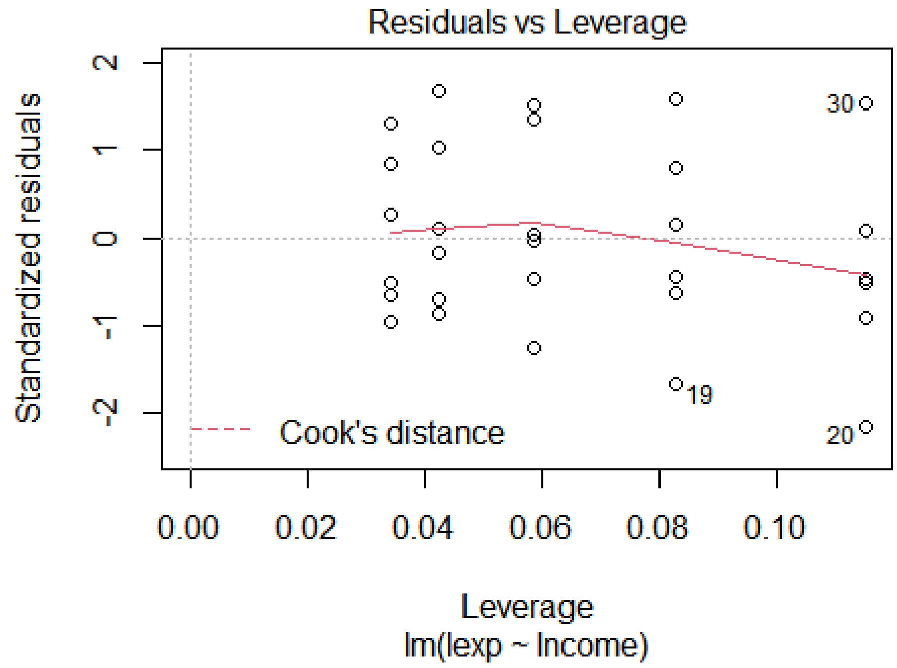
H0 : The residual values are having constant variance. (Homoscedastic)
H1 : The residual values are not having constant variance (Hetroscedastic)

```
# since constant variance is not true we are transforming the data by taking
log of Y

lexp = log(Expenditure)
reg1 = lm(lexp~Income)
plot(reg1)
```


Residuals vs Fitted
Residuals
0.10
0.05
0.00
-0.10
27
19
20
9.3
9.4
9.5
9.6
9.7
Fitted values
lm(lexp ~ Income)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
27
19
20
-2
-1
0
1
2
Theoretical Quantiles
lm(lexp ~ Income)

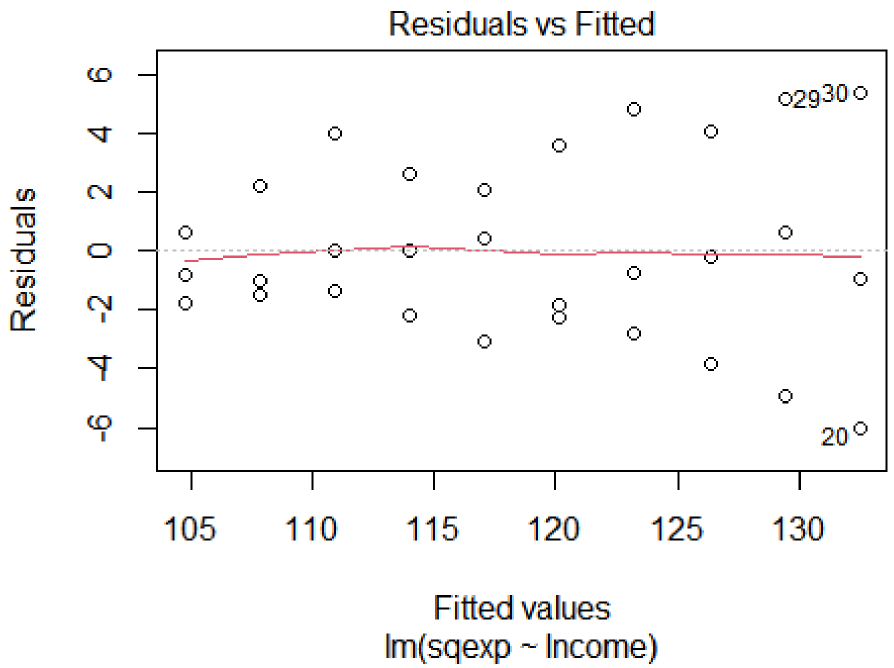


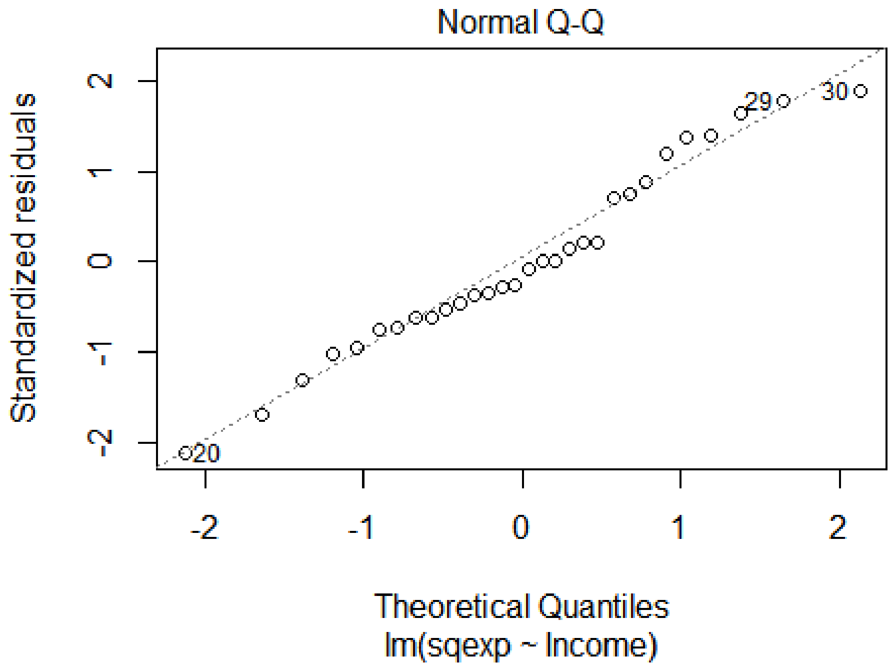


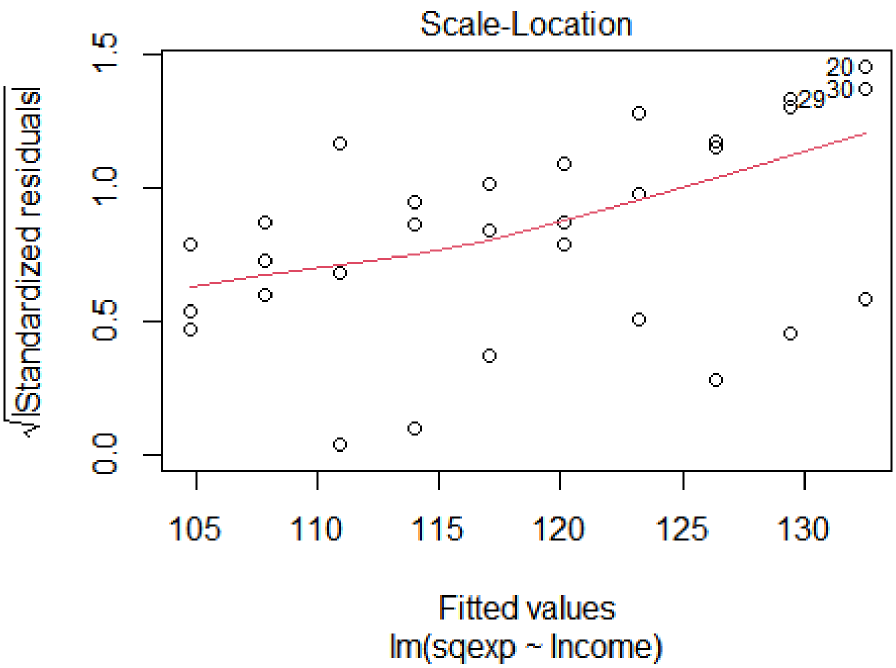
```
bptest(reg1)

##
##  studentized Breusch-Pagan test
```

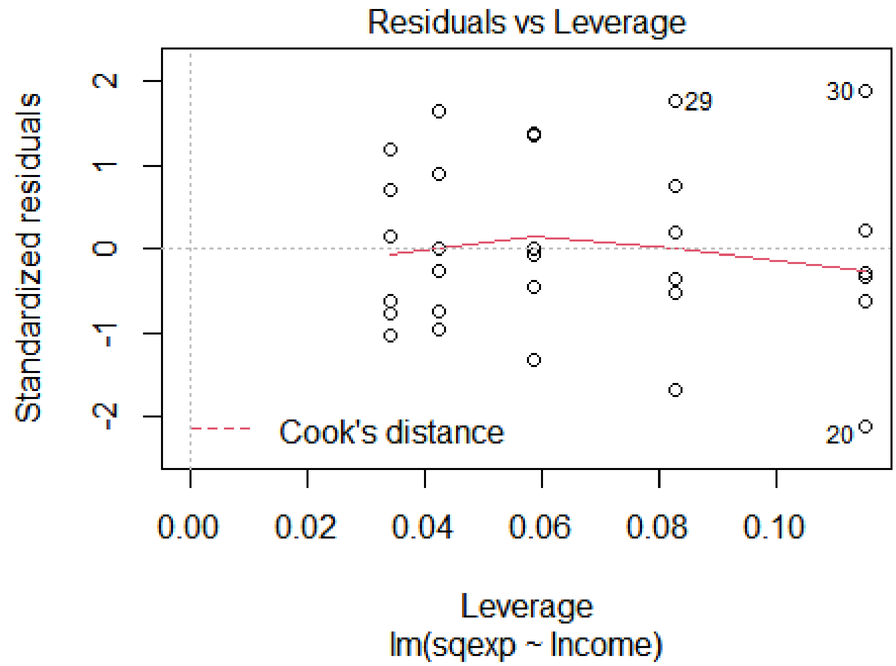
```
##
## data:  reg1
## BP = 7.0783, df = 1, p-value = 0.007802
```

```
#SQRT(Y)
sqexp = sqrt(Expenditure)
reg2 = lm(sqexp~Income)
plot(reg2)
```


Residuals vs Fitted
Residuals
6
4
2
0
-2
-4
-6
29
30
20
105
110
115
120
125
130
Fitted values
lm(sqexp ~ Income)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
29
30
20
-2
-1
0
1
2
Theoretical Quantiles
lm(sqexp ~ Income)

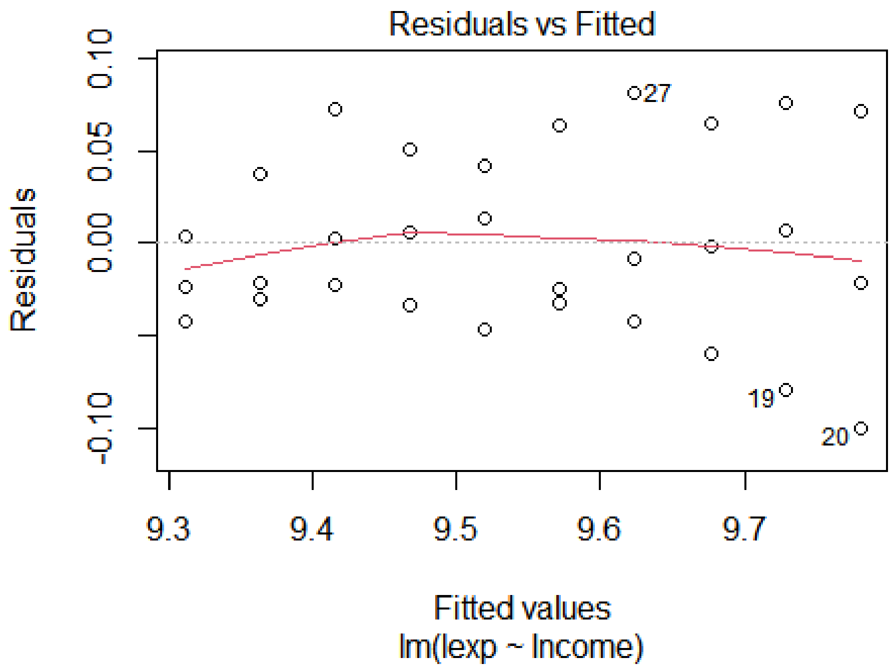


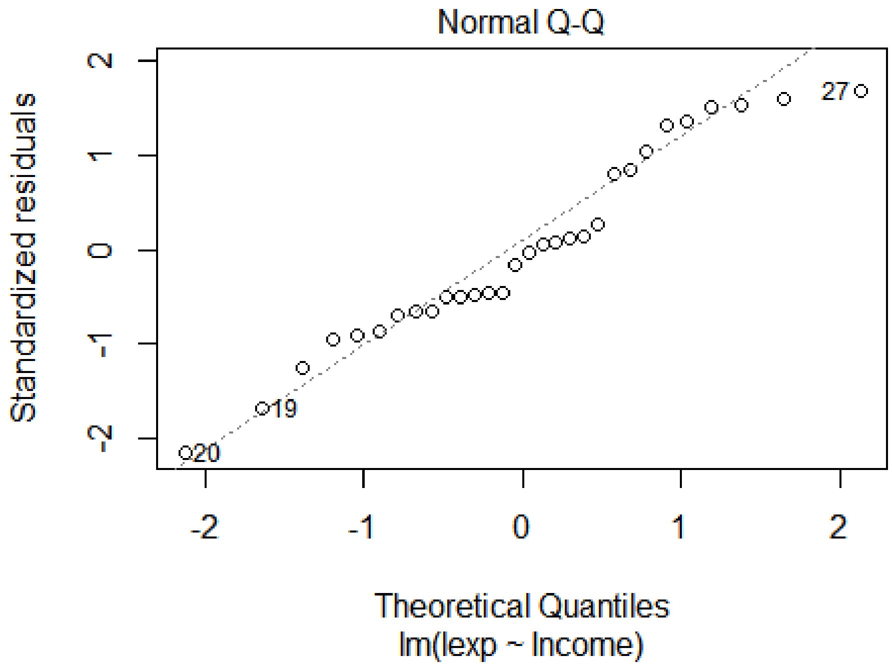


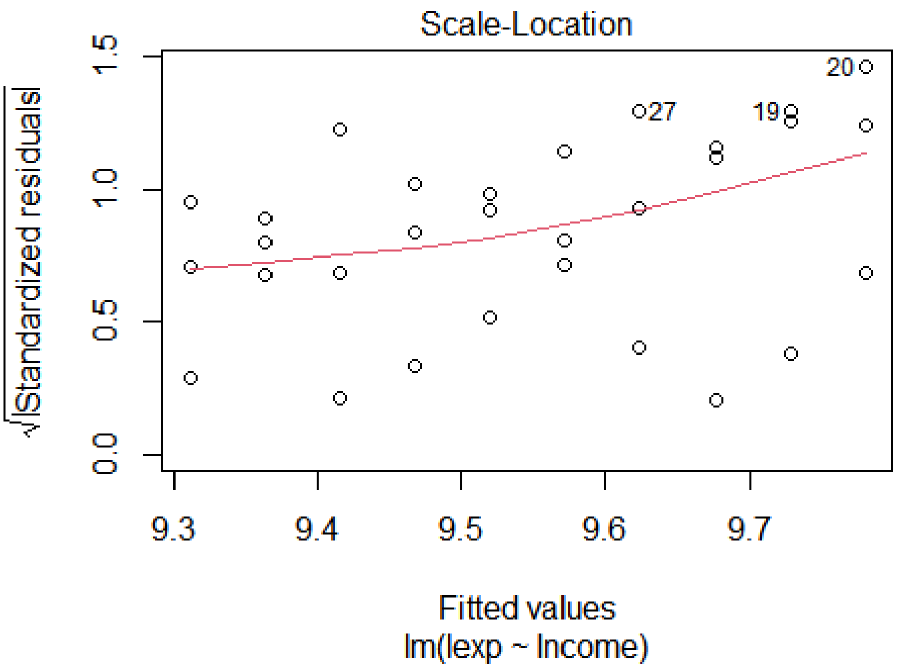
```
bptest(reg2)

## 
##  studentized Breusch-Pagan test
```

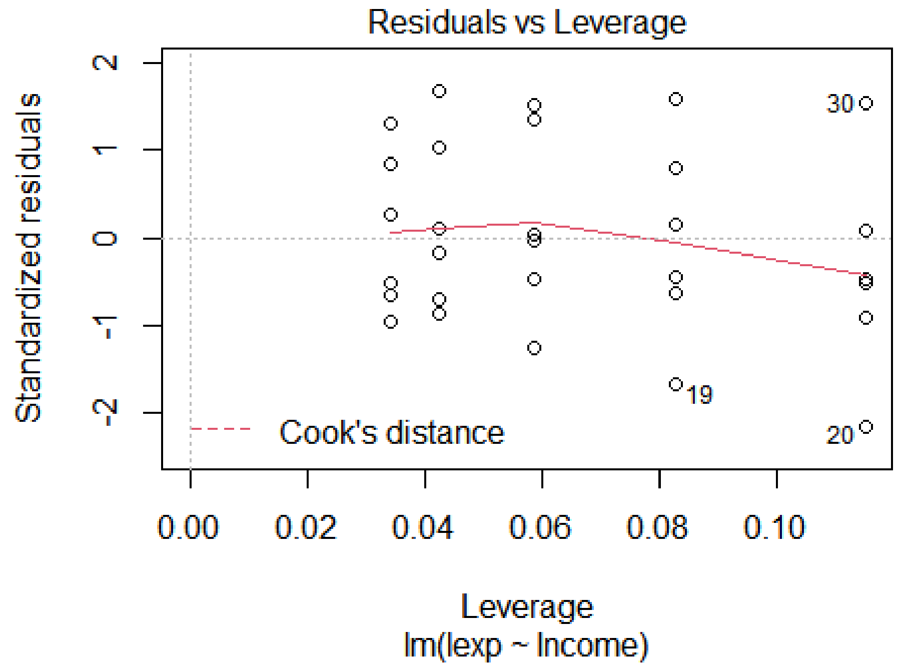
```
##
## data:  reg2
## BP = 9.9329, df = 1, p-value = 0.001624
```

```
#1/Y TRANSFORMATION
bexp = 1/Expenditure
reg3 = lm(bexp~Income)
plot(reg1)
```


Residuals vs Fitted
Residuals
0.10
0.05
0.00
-0.10
27
19
20
9.3
9.4
9.5
9.6
9.7
Fitted values
lm(lexp ~ Income)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
27
19
20
-2
-1
0
1
2
Theoretical Quantiles
lm(lexp ~ Income)





```
bptest(reg3)

##
##  studentized Breusch-Pagan test
```

```
##
## data:  reg3
## BP = 1.0701, df = 1, p-value = 0.3009
```

INTERPRETATION

Since the constant variance assumption is not satisfied we are going for transformation y variable. First we tried taking the log of y variable and hence doing bptest on this log y variable. But still the pvalue <0.05 and hence we are rejecting the null hypothesis. Then we go for Sqrt transformation but resulted in the same result. Then we went for Inverse transformation,by finding out the inverse of y variable and found out the bptest value, which we obtained p value = 0.3009 >0.05 hence we can't reject the null hypothesis. So using the inverse transformation we found out that the residuals are homoscedastic.

7.Autocorrelation of residuals

DURBIN WATSON TEST H0 - The correlation of popln errors is 0, i.e there doesn't exist an autocorrelation H1 - There is an autocorelation

```
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 0.30717, p-value = 1.553e-10
## alternative hypothesis: true autocorrelation is greater than 0
```

INTERPRETATION

Durbin watson test is used for checking autocorrelation of residuals. The p-value = 1.553e^-10 <0.05. Hence we reject the null hypothesis. I.e there is an autocorrelation between the residuals. Alos from durbin watson table we obtained the DL= 1.134 and DU=1.264 as the lower and upper limits. D = 0.30717 and is less than D lower limit. hence we reject the null hypothesis.

8.Including an unusual y observation to the data set and examining whether it is an influential point.

```
new_data = rbind(data,c(31,60000,21000))
reg_new = lm(new_data$Expenditure~new_data$Income)
summary(reg_new)
```

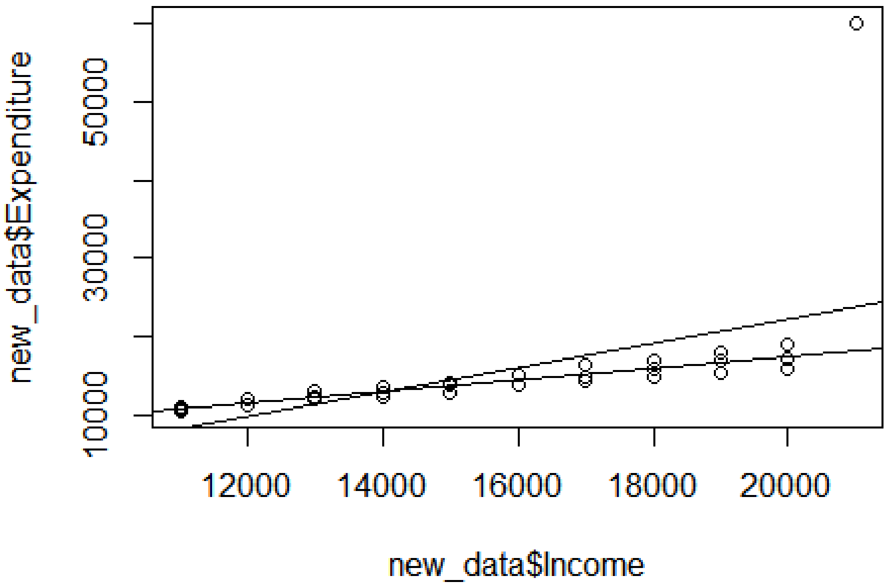
```
##
## Call:
## lm(formula = new_data$Expenditure ~ new_data$Income)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6279  -2918   -831   1048  36185
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -8458.0420  6965.5913  -1.214  0.23444
## new_data$Income     1.5368     0.4365   3.521  0.00144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7261 on 29 degrees of freedom
## Multiple R-squared:  0.2995, Adjusted R-squared:  0.2753
## F-statistic:  12.4 on 1 and 29 DF,  p-value: 0.001442
```

```
summary(model) #old data
```

```
##
## Call:
## lm(formula = Expenditure ~ Income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1453.94  -473.48   -73.94   483.33  1546.06
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.799e+03  7.564e+02   3.701 0.000931 ***
## Income      7.327e-01  4.798e-02  15.271 4.16e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 754.8 on 28 degrees of freedom
## Multiple R-squared:  0.8928, Adjusted R-squared:  0.889
## F-statistic: 233.2 on 1 and 28 DF,  p-value: 4.165e-15
```

```
plot(new_data$Expenditure~new_data$Income)
abline(reg_new)
abline(model)
```



##

INTERPRETATION

A new data value with an extreme y value is added to the data using rbind() command. new model is also obtained using the new data. Plot of the new model is obtained with the abline of new data along with the abline of the old data. A drastic change in the slope of the regression line is observed from the plot. Hence we can conclude that the point added is an influential point. Since the new data added was having an extreme y value it should be an outlier. This we can confirm from the summary of the model. The old model is having an R^2 value of0.89 while the new model is having an R^2 of 0.29.R^2 value reduced drastically and hence we can tell that the data value that we added is an outlier. By comparing the p-value of both models we can observe that there is an overall decrease in the significance of the model.

CONCLUSION

1. The scatter plot between Expenditure and Income shows a linear relationship between them Also we can identify the dependent and independent variables. Y = Expenditure X = Income

2. The simple linear regression model obtained is:

   Expenditure_hat = 2.799e^(03) + 7.327e^-(01) * Income

3. One unit change in Income will result in 7.327e^-01 unit change in expenditure. R^2 value obtained is 0.89 and hence we can conclude that 89% of variability in expenditure is explained by the variable income.

4. From the plots of residual vs fitted we can observe a funnel shape in the graph which tells that the constant variance assumption is violated.i.e residual is the increasing sequence in Y pointing towards a defective model Also 20,29 and 30th observations are marked and hence this could be an outlier.
5. From the Normal Q-Q plot it is observed that the residuals follows a normal distribution.
6. From the leverage vs residual graph it appears to change indicating hetroscedasticity. This means that the model adequacy is not met. 7.Residuals are hetroscedastic. 8.Residuals follow a normal distribution
7. So using the inverse transformation we found out that the residuals are homoscedastic.
8. There is an autocorrelation between the residuals
9. After adding a new data value with an extreme y value (60000,21000), A drastic change in the slope of the regression line is observed from the plot. Hence concluded that the point added is an influential point.